Universidad Internacional de La Rioja

Facultad de Economía y Empresa

Máster Universitario en Inteligencia de Negocio

Proyecto de Análisis de Datos y elaboración de dashboard para la multinacional tabacalera Altadis.

Trabajo fin de estudio presentado por:	González Cuerva, Sara; Ferreño Pico, Lucia; Vicente Lasa, Francisco
Tipo de trabajo:	Proyecto de Inteligencia de Negocio
Modalidad:	Grupal
Director/a:	Merillas Santos, Iván
Fecha:	04/07/2025

Organización de trabajo en equipo y mecanismos de coordinación

Dado que el presente Trabajo de Fin de Máster "Proyecto de Análisis de Datos y elaboración de dashboard para la multinacional tabacalera Altadis" se ha elaborado en la modalidad grupal, era necesario implementar una planificación metódica y establecer una eficaz coordinación entre los miembros del equipo. En cada fase del proyecto y principalmente antes de alcanzar cada hito, el equipo acordó la asignación de tareas y los criterios de decisión, lo que permitió mantener una dinámica de trabajo estructurada y coherente. El compromiso personal de cada miembro del equipo y la constante colaboración demostrada por cada uno de ellos han sido fundamentales en el cumplimiento de los objetivos del proyecto, respetando tanto los plazos acordados como los criterios académicos exigidos.

Para asegurar una coordinación efectiva a lo largo del desarrollo del trabajo, el equipo estableció dinámicas de seguimiento y comunicación regulares. Se mantuvieron reuniones virtuales que permitieron alinear criterios, resolver dudas y supervisar el avance de las tareas. Asimismo, se llevaron a cabo encuentros y comunicaciones periódicas con el tutor, orientadas a validar el progreso del proyecto y recibir orientación sobre los siguientes pasos.

En cuanto a las herramientas empleadas, se utilizó Microsoft Teams para las reuniones virtuales de seguimiento y como repositorio de documentos compartidos, lo que facilitó el trabajo colaborativo y sin duplicidades; la plataforma Canvas facilitada por UNIR, se utilizó como canal principal para la comunicación formal con el tutor y el envío de documentación relevante; y canales complementarios como correo electrónico y grupo de WhatsApp, que se utilizaron para coordinar reuniones y resolver cuestiones puntuales de forma ágil.

A continuación, se muestra un cuadro resumen con la organización grupal del proyecto, en el que se detallan las principales tareas y el grado de participación de los miembros en las distintas fases del trabajo.

Epígrafe	Alumno Responsable	Objetivos perseguidos
Introducción: planteamiento, objetivos y elementos innovadores	Trabajo colaborativo.	Describir y justificar el proyecto, definir objetivos alineados con la estrategia de la empresa y especificar los aportes innovadores aplicables.
Alcance y planificación	Trabajo colaborativo.	Analizar el contexto actual de la empresa y diseñar un proyecto de Inteligencia de Negocio ajustado a sus necesidades, capacidades y metas estratégicas.
Análisis y definición	Trabajo colaborativo (Lucía Ferreño Pico y Sara González Cuerva).	Entender y preparar los datos para extraer patrones útiles mediante análisis estadístico y técnicas de machine learning.
Construcción, prueba, implementación y despliegue	Trabajo colaborativo (Lucía Ferreño Pico y Sara González Cuerva).	Construir y validar modelos analíticos útiles, demostrando su valor para apoyar decisiones comerciales reales mediante ejemplos y visualizaciones representativas.
Conclusiones	Trabajo colaborativo.	Recopilar los hallazgos clave del proyecto y destacar su utilidad para la mejora de la estrategia y la toma de decisiones empresariales.
Limitaciones y prospectiva	Trabajo colaborativo.	Reconocer las limitaciones del estudio y proponer líneas de mejora y desarrollo futuro del proyecto.

Resumen

El presente trabajo constituye un ejercicio de simulación de consultoría empresarial para la tabaquera española Altadis, con el objetivo de transformar los datos en conocimiento útil para mejorar la toma de decisiones. La empresa facilitó seis conjuntos de datos anonimizados sobre variables comerciales que fueron la base sobre la cual se diseñó y se desarrolló el proyecto de inteligencia de negocio.

Para alcanzar los resultados se utilizó la metodología CRISP-DM y una ejecución híbrida tipo waterfall-agile, concatenando los procesos de comprensión de los datos, preparación, modelado, evaluación y reporte. El análisis comenzó con el enriquecimiento geográfico, seguido de técnicas de aprendizaje no supervisado y análisis de series temporales, para finalmente elaborar un cuadro de mando interactivo.

A pesar de las limitaciones en la calidad de los datos, se han podido determinar factores relevantes para explicar patrones en las ventas, como son ubicación fronteriza-costera, turismo, la estacionalidad o la concentración demográfica.

Palabras clave: inteligencia de negocio, análisis de datos, CRISP-DM, aprendizaje automático, Altadis

Abstract

This paper presents a business consulting simulation exercise for the Spanish tobacco company Altadis, focused on transforming data into actionable insights to enhance decision-making processes. The company provided six anonymized datasets on commercial variables, which formed the basis for designing and developing the business intelligence project.

To achieve the results, the CRISP-DM methodology was used along with a hybrid waterfall-agile execution approach, linking the processes of data understanding, preparation, modelling, evaluation, and reporting. The analysis began with geographic enrichment, followed by unsupervised learning techniques and time series analysis, and concluded with the development of an interactive dashboard.

Despite limitations in data quality, it was possible to identify relevant factors explaining sales patterns, such as border-coastal location, tourism, seasonality, and demographic concentration.

Keywords: business intelligence, data analysis, CRISP-DM, machine learning, Altadis

Índice de contenidos

Organización de trabajo en equipo y mecanismos de coordinación 2

Introducción 13

1.1. Planteamiento general: descripción y justificación del proyecto 15

1.2. Objetivos del TFE 17

1.2.1. Objetivo General 17

1.2.2. Objetivos Específicos 18

1.3. Elementos innovadores del proyecto 18

2. Alcance y planificación 21

2.1. Fase de descubrimiento: evaluación del entorno actual 22

2.1.1. Información deseada 27

2.1.2. Información actual: deficiencias y soluciones alternativas. 28

2.1.3. Habilidades analíticas actuales. 31

2.2. Fase de análisis: identificación de gaps 32

2.2.1. Capacidad de los informes actuales 32

2.2.2. Proveedores de tecnología necesarias 33

2.2.3. Diferencia entre los informes actuales y la información deseada 34

2.2.4. Cronología, costes y recursos humanos implicados 34

2.3. Fase de recomendaciones: alcance, prioridades y presupuesto 37

2.3.1. Promoción del proyecto en la organización 41

2.3.2. Demostraciones 42

3. Análisis y definición 44

3.1. Análisis de los datos a utilizar 44

3.1.1. Análisis datos a utilizar en análisis de segmentos y clubes de convergencia .. 44

3.1.2. Análisis datos a utilizar en análisis de temporalidad de las ventas 46

3.1.3. Análisis datos a utilizar en la elaboración del dashboard 47

3.2. Análisis histórico y/o limpieza de datos 49

3.2.1. Análisis histórico y/o limpieza de datos: Análisis Segmentación y Clubes de Convergencia 49

3.2.1.1. Depuración de registros incompletos o inválidos y estandarización de los datos estructurales. 49

3.2.1.2. Conversión y codificación de variables 53

3.2.1.3. Estructuración de datos y agregación 54

3.2.1.4. Transformaciones estadísticas 54

3.2.1.5. Normalización y escalado 55

3.2.2. Análisis histórico y/o limpieza de datos: Temporalidad de las ventas 55

3.2.2.1. Análisis exploratorio y visualizaciones 56

3.2.3. Limpieza de datos para el desarrollo del dashboard 62

3.3. Modelado propuesto 62

3.3.1. Segmentación estructural de puntos de venta 63

3.3.2. Segmentación por rendimiento comercial 73

3.3.3. Agrupación territorial por clubes de convergencia 79

3.3.4. Temporalidad de las ventas 80

3.3.5. Dashboard 81

4. Construcción, prueba, implementación y despliegue 83

4.1. Medidas de bondad del modelado 83

4.1.1. Segmentación de puntos de venta (estructural y por rendimiento comercial) 83

4.1.1.1. Método de Elbow 84

4.1.1.2. Índices de Silhouette, Davies-Bouldin y Calinsky-Harabasz 85

4.1.2. Clubes de convergencia territorial 86

4.1.3. Temporalidad de las ventas: ARIMA y ETS 87

4.1.4. Dashboard 88

4.2. Ejemplo de aplicación 90

4.2.1. Segmentación estructural de puntos de venta 90

4.2.2. Segmentación por rendimiento comercial 91

4.2.3. Agrupación territorial por clubes de convergencia 93

4.2.4. Temporalidad de las ventas: ARIMA y ETS 93

4.2.5. Dashboard 93

5. Conclusiones 95

6. Limitaciones y prospectiva 99

6.1. Limitaciones 99

6.2. Prospectiva 100

Referencias bibliográficas 102

Anexo A. Código Clubes de Convergencia 105

Anexo B. Código Temporalidad de las ventas 110

Anexo C. Proceso de cálculo del número óptimo de clústeres en la segmentación de puntos de venta por rendimiento comercial. 123

Anexo D. Gráficos Clubes de Convergencia 127

Anexo E. Gráficos de Temporalidad de las ventas 132

Anexo F. Elaboración Dashboard 153

Índice de figuras

Figura 1. Metodología Crisp-DM 21

Figura 2. Análisis DAFO sobre el sector del tabaco y los productos de nueva generación (NGP). 27

Figura 3. Esquema de relación de las tablas disponibles..... 30

Figura 4. Cronología y fases en las que se desarrolla el proyecto. 35

Figura 5. Equipo dedicado..... 36

Figura 6. Tabla resumen presupuesto dedicado al proyecto en base a la realización por un equipo interno. 41

Figura 7. Tabla resumen del dataset estructural puntos_venta_enriquecido2_sinNA.xlsx.... 45

Figura 8. Distribución de los puntos de venta según ubicación. 50

Figura 9. Distribución de los puntos de venta según el tamaño del local. 50

Figura 10. Distribución territorial de las provincias con mayor número de puntos de venta. 51

Figura 11. Distribución de las ventas medias en función del tipo de ubicación 51

Figura 12. Distribución de las ventas medias en función del tamaño del establecimiento. ... 52

Figura 13. Distribución de las ventas totales por provincia..... 52

Figura 14. Mapa de calor de la matriz de correlación entre Location, Tam_m2 y Engage. 53

Figura 15. Tendencia y estacionalidad por provincia (g1). 57

Figura 16. Ventas por día de la semana y Media diaria por mes (g2 + g2_normalizado). 59

Figura 17. Ventas totales según tipo de día (g3), Evolución de ventas por tipo de día(g4) y Ventas por día de la semana y tipo de día (g5). 60

Figura 18. Ranking ventas por producto (g6), Ranking Ventas Totales por provincia (g7) y Top 20 puntos venta por ventas totales (g8). 61

Figura 19. Código DAX. 62

Figura 20. Nivel de implicación comercial (Engage) medio en función de la ubicación del punto de venta (Location). 63

Figura 21. Proporción de los tipos de ubicación en las provincias con mayor número de puntos de venta. 64

Figura 22. Representación gráfica del método Elbow 65

Figura 23. Representación gráfica de la matriz de correlación. 66

Figura 24. Mapa de calor de las medias de las variables por clúster, siendo k=4. 67

Figura 25. Distribución de los puntos de venta agrupados con cluster con etiquetas descriptivas de perfil ya definidas en la segmentación estructural. 68

Figura 26. Importancia de las variables en la predicción de clústeres. 69

Figura 27. Distribución de los clústeres en las provincias con mayor número de puntos de venta. 70

Figura 28. Distribución territorial de los segmentos. 71

Figura 29. Distribución de los grupos por tipo de ubicación 72

Figura 30. Distribución de los grupos por tamaño del local. 72

Figura 31. Distribución de las ventas medias por ubicación y nivel de implicación comercial. 73

Figura 32. Distribución de las ventas medias por provincia y tamaño del local. 74

Figura 33. Representación gráfica del método Elbow. 75

Figura 34. Distribución de las ventas medias por perfil estratégico definido en la segmentación por rendimiento comercial. 76

Figura 35. Distribución de los perfiles estratégicos en las provincias con mayor número de puntos de venta. 76

Figura 36. Distribución de ventas per cápita por provincia. 77

Figura 37. Distribución población y ventas por segmento. 78

Figura 38. Modelo Estrella. 81

Figura 39. Eje calendario. 82

Figura 40. Eje geográfico 82

Figura 41. Método de Elbow en segmentación estructural. 84

Figura 42. Método de Elbow en segmentación por rendimiento de ventas 84

Figura 43. Comparativa de resultados de las métricas de validación interna de clústeres en la segmentación estructural. 85

Figura 44. Comparativa de resultados de las métricas de validación interna de clústeres en la segmentación por rendimiento económico. 86

Figura 45. Modelo ARIMA diario España. 88

Figura 46. Menú de filtros con selecciones. 89

Figura 47. Informe de ventas con selecciones. 89

Figura 48. Distribución de los clústeres en las provincias con mayor número de puntos de venta. 91

Figura 49. Distribución de la eficiencia de ventas por m2 con relación a los perfiles estratégicos obtenidos en la segmentación por rendimiento comercial. 92

Figura 50. Detalle de Ventas con selecciones. 94

Figura 51. Ejemplo de búsqueda de patrones en los prefijos de los productos: 99

Índice de tablas

Tabla 1. Régimen Fiscal aplicable a productos del tabaco y productos de nueva generación (NGP). 25

Tabla 2. *Tabla Calendario (parte 1).* 48

Tabla 3. *Tabla Calendario (parte 2).* 48

Tabla 4. *Tabla Calendario (parte 3).* 49

Tabla 5. *Tabla resumen MAPE.* 87

Introducción

En el contexto empresarial actual, marcado por la ultra competencia global y la transformación digital, las organizaciones enfrentan un desafío crucial: gestionar y aprovechar el potencial de los datos como un recurso estratégico. Las empresas generan y acumulan grandes volúmenes de información diariamente, provenientes de diversas fuentes primarias tales como transacciones comerciales, interacciones con clientes, redes sociales, sistemas operativos internos y datos externos de los mercados. Sin embargo, en muchos casos, esta información es gestionada de manera fragmentada e incompleta, lo que limita su capacidad para influir en la toma de decisiones y, por ende, permanece subutilizada su capacidad para impactar en los resultados de la empresa.

En el informe "Uso de inteligencia artificial y big data en las empresas españolas. 2023" realizado por el Observatorio Nacional de Tecnología y Sociedad del Ministerio de Asuntos Económicos y Transformación Digital se plasma que desde el 2019 el incremento en España en uso de la IA y de la Analítica de datos crece a una velocidad extraordinaria, año tras año y su uso está liderado por las grandes empresas. (Observatorio Nacional de Tecnología y Sociedad, 2023).

Existe una necesidad constante de optimizar procesos y recursos para diferenciarse de los competidores. Para lograr esta diferenciación, las empresas con una cultura data-driven desarrollada pueden hacer uso de datos relacionados con clientes, proveedores, competidores y empleados para mejorar su estrategia y competitividad en su sector. Para ello, estas deberán apoyarse en la Inteligencia de Negocios ya que les permitirá transformar los datos en información y ésta, a su vez, en conocimiento, de tal forma que puedan usarlo como base para la toma de decisiones tanto estratégicas como operativas.

No obstante, una gran cantidad de datos que la empresa puede compilar y gestionar puede que no lleguen a explotarse apropiadamente debido, entre otros factores, a que la empresa no cuente con una tecnología adecuada que permita procesar la totalidad de los datos disponibles o a una falta de personal capacitado para procesar y extraer valor de la información.

El uso y la interconexión de dispositivos digitales ha aumentado exponencialmente y con ella lo ha hecho el volumen, la velocidad y la variedad de datos disponibles, creando una masividad compleja y difícil de gestionar para las organizaciones que no se adaptan y mantienen los métodos tradicionales. Las empresas que no logren incorporarse a este nuevo paradigma basado en datos enfrentan riesgos significativos, a saber: pérdida de cuota de mercado, incapacidad para ajustar su oferta a las preferencias de los usuarios, desaprovechar oportunidades y finalmente desembocar en una caída de la rentabilidad o a medio plazo poner el jaque la existencia de la organización.

Hay que tener en cuenta que el panorama competitivo actual se caracteriza por márgenes cada vez más reducidos, ciclos de innovación acelerados y consumidores altamente informados que exigen experiencias personalizadas y respuestas inmediatas. En este escenario, las decisiones basadas en intuición o en análisis superficiales resultan insuficientes y potencialmente peligrosas.

Varios autores han tratado el fenómeno de la "commoditización" refiriéndose a la tendencia a la homogeneización de los productos y servicios, un contexto en el cual sólo el uso de la información acabaría diferenciando a unas empresas de otras. Y es que los datos deben ser considerados un activo productivo para la empresa; una gestión empresarial data-driven supone crear ventajas competitivas a partir de los datos y ser capaces de extraer conocimiento accionable de las diferentes fuentes de información (Porter, 1985) Porter, M. E. (1985). *Competitive Advantage: Creating and Sustaining Superior Performance*. Free Press.

Las organizaciones enfrentan problemas recurrentes como la falta de claridad en sus procesos de ventas, dificultades para identificar segmentos de clientes prioritarios, incapacidad para prever fluctuaciones en la demanda o roturas de stock y una limitada capacidad para evaluar el impacto de las campañas promocionales. Estos desafíos son comunes a muchas empresas que no cuentan con un sistema integrado para gestionar y analizar sus datos. En este escenario, surge la necesidad de implementar un proyecto de *Business Intelligence* como una solución estratégica para transformar los datos en activos productivos que impulsen el rendimiento empresarial.

«Muchas empresas se enfrentan a retos derivados de las limitaciones de tiempo, la falta de conocimientos especializados y las habilidades necesarias para integrar eficazmente paneles

de control y visualizaciones de datos en sus operaciones. El panorama tecnológico en constante evolución agrava estos desafíos, dificultando aún más el dominio de la gestión eficaz de datos empresariales.» (Ross, 2023)

La relevancia de un proyecto de Inteligencia de Negocio radica en su capacidad para abordar problemas críticos que afectan la competitividad y sostenibilidad de la empresa. Los datos no deben ser vistos únicamente como registros históricos; deben convertirse en un recurso estratégico capaz de generar valor tangible para la empresa. Un sistema de BI efectivo permite monetizar este activo intangible al facilitar su transformación en *insights* aplicables y decisiones fundamentadas que impactan directamente en los resultados financieros. En un mercado saturado, contar con información precisa y oportuna permite a las empresas tomar decisiones fundamentadas que les otorguen una ventaja sobre sus competidores. La capacidad para identificar tendencias emergentes, responder ágilmente a cambios en el comportamiento del consumidor o adaptar estrategias comerciales en tiempo real constituye un diferenciador crítico en entornos altamente competitivos.

1.1. Planteamiento general: descripción y justificación del proyecto

Este trabajo de fin de máster se realiza con los datos proporcionados a la Universidad Internacional de la Rioja por la empresa multinacional, Altadis, dedicada a la producción y comercialización de productos del tabaco y otros servicios relacionados. Se formó en 1999 tras la fusión de la española Tabacalera y la francesa SEITA (Société d'Exploitation Industrielle des Tabacs et des Allumettes), consolidándose como una de las principales compañías del sector en Europa. En 2008, Altadis fue adquirida por *Imperial Brands*, una de las mayores tabacaleras del mundo, lo que reforzó su presencia global.

La compañía tiene una fuerte presencia en mercados clave, con operaciones que abarcan desde la fabricación de cigarrillos y puros hasta la distribución de tabaco en diferentes países, tal y como se explica en su página web corporativa. (Altadis, S.A., s.f.)

En España, Altadis ha desempeñado un papel fundamental en la evolución del sector tabacalero, adaptándose a nuevas regulaciones y tendencias de consumo. Además de su actividad en la industria del tabaco, la empresa ha participado en la logística y distribución de productos de conveniencia. (Altadis, S.A., s.f.)

De igual forma, en el próximo apartado 2.1 Fase de descubrimiento: evaluación del entorno actual, se desarrolla una contextualización de la empresa con información más detallada de la misma.

Bajo el planteamiento de un rol de un departamento de Inteligencia de Negocio dentro la empresa, o bien el de una consultora externa especializada en el análisis de datos, se llevará a cabo un proyecto *Business Intelligence* a partir de los datos disponibles proporcionados por la empresa y su enriquecimiento con fuentes externas.

El análisis detallado de los datos permitirá identificar ineficiencias operativas y mejorar procesos clave como la gestión del inventario o la cadena logística. Mediante modelos predictivos de aprendizaje automático, basados en datos históricos, es posible anticipar tendencias futuras, así como fluctuaciones en las ventas o patrones de comportamiento por parte de los clientes.

El análisis de datos mediante algoritmos de aprendizaje no supervisado facilitará la identificación precisa de segmentos específicos dentro de la base de clientes, permitiendo personalizar estrategias comerciales y mejorar la experiencia del cliente, lo que posibilita campañas de marketing mejor ajustadas con mayor retorno sobre la inversión y programas de fidelización adaptados a las necesidades específicas de cada grupo. Los análisis geográficos y temporales revelarán patrones valiosos que pueden ser utilizados para optimizar campañas promocionales o expandir operaciones a nuevas regiones. Se usarán técnicas de minería de datos permitiendo identificar correlaciones no evidentes entre variables de negocio tales como nuevos nichos de mercado que permanecerían invisibles bajo metodologías analíticas tradicionales.

En síntesis, la implementación de un proyecto de *Business Intelligence* no representa simplemente una actualización tecnológica, sino una transformación estratégica en la forma en que la organización utiliza sus datos para crear valor, competir eficazmente y asegurar su sostenibilidad en un entorno de negocios cada vez más complejo y desbordado de datos e información.

1.2. Objetivos del TFE

La estrategia empresarial de Altadis a medio y largo plazo está diseñada con un enfoque Customer Center y entre los objetivos destacan el Análisis de la calidad y consistencia de los datos, la implementación de un sistema de dashboard de control y generación de insights empresariales con los que producir información de valor basada en los datos proporcionados. Por lo que, con el deseo de alinear el presente trabajo y sus objetivos, con los objetivos estratégicos de la multinacional Altadis, se establece el siguiente objetivo general y con él los específicos.

1.2.1. Objetivo General

El objetivo general de este proyecto es desarrollar un proyecto de inteligencia de negocio para la empresa tabacalera Altadis, S.A. con el objetivo de extraer *insights* que respalden la toma de decisiones estratégicas de la empresa y así mejorar sus resultados operativos, financieros y de conocimiento de los consumidores.

Para ello se alineará este trabajo con el propósito y visión de la compañía, que tal y como explican en su página web: "…la visión del Grupo es "desarrollar una empresa fuerte y competitiva impulsada por la responsabilidad y la capacidad de enfoque y elección." (Proposito y Vision , s.f.)

Por ende, los insights de este proyecto de inteligencia de negocio están centrados en cumplir los 5 valores de Altadis, como por ejemplo situar al cliente en el centro.

"Start with the Consumer"

"Conocimiento profundo y granular de los consumidores.

- Todo empieza por nuestros consumidores. Estamos aquí por y para ellos
- Conocerlos es la clave para alcanzar el éxito
- Las buenas decisiones se toman combinando datos, experiencia y conocimiento" (Proposito y Vision , s.f.)

1.2.2. Objetivos Específicos

Con el fin de alcanzar el objetivo general se establecen los siguientes objetivos específicos:

- Realizar un estudio de comprensión del negocio, incluyendo revisión de información interna proporcionada por la empresa, así como de fuentes externas, tales como la industria tabaquera en la que opera, o tendencias concretas del mercado español.
- Llevar a cabo un análisis exploratorio y de procesamiento de los datos. Esto incluye un proceso de ETL (Extracción-Transformación-Carga) así como el análisis exploratorio de los datos (EDA) usando herramientas de estadística descriptiva.
- Análisis avanzado de los datos usando algoritmos de aprendizaje automático no-supervisado y técnicas de análisis de series temporales, para generar modelos de segmentación de puntos de venta, detección de patrones en los datos y modelar la temporalidad de las ventas.
- Creación de un sistema de cuadro de mandos incluyendo la definición de Indicadores Clave de Rendimiento (Key Performance Indicators, KPIs) y su representación visual para contextualizar, interpretar y aplicar la información obtenida en el proyecto.

1.3. Elementos innovadores del proyecto

En la actualidad, la innovación se ha consolidado como un pilar fundamental en la estrategia de las organizaciones que desean mantenerse competitivas y adaptarse a entornos cambiantes. Se puede definir la innovación como la introducción de mejoras significativas en productos, procesos, estructuras organizativas o modelos de negocio. La innovación puede ser incremental, optimizando lo ya existente; disruptiva, transformando mercados o sectores; o radical, que implica la creación de soluciones completamente nuevas. También existe la innovación organizativa o estratégica, que reconfigura los sistemas de gestión, la toma de decisiones y cultura corporativa. Este proyecto integra elementos de estas categorías, combinando la adopción tecnológica con nuevos enfoques analíticos y transformaciones en el modelo de gestión empresarial:

Integración de la Inteligencia de Negocio como núcleo de la estrategia organizativa: en una empresa como Altadis, con una estructura consolidada en un sector tradicional, la incorporación de la Inteligencia de Negocio supone una ruptura con prácticas basadas en la

intuición o en informes fragmentados. Este enfoque orientado por datos permite generar conocimiento estratégico, detectar oportunidades y gestionar el negocio de forma más eficaz. Supone una innovación organizativa y cultural, al promover una toma de decisiones fundamentada en evidencias analíticas, y representa una transformación radical en el uso de la información como activo productivo.

Hibridación metodológica entre Agile y Waterfall: su incorporación constituye un elemento innovador a la hora de la implementar un proyecto de BI en la empresa. Esta metodología híbrida combina la rigurosidad estructurada del enfoque secuencial con la adaptabilidad del enfoque iterativo, permitiendo una respuesta ágil a nuevas necesidades de análisis surgidas durante el desarrollo.

Enriquecimiento de los datos mediante fuentes externas: como bases de datos geográficas (GeoNames), estadísticas oficiales del sector (Ministerio de Hacienda) y datos de códigos postales, ha permitido superar las limitaciones de la información interna, especialmente en lo que respecta a la geolocalización de puntos de venta y la categorización de productos. Este enfoque de enriquecimiento mejora la profundidad del análisis y amplía la aplicabilidad de los resultados.

Aplicación de modelos de aprendizaje automático no supervisado: técnicas de segmentación como algoritmo K-means se han empleado para realizar clusterización avanzada de los puntos de venta, permitiendo segmentar la red comercial según criterios estructurales, comportamiento de ventas y ubicación geográfica. Esta metodología innovadora aporta una visión más precisa de la heterogeneidad del mercado y facilita la personalización de las acciones comerciales.

Uso de modelos de predicción temporal, concretamente ARIMA y ETS: si bien estos modelos no aportaron resultados concluyentes, su utilización es relevante para entender las propias limitaciones de la empresa en cuanto a la recopilación y gestión de sus datos. En este sentido el uso de estos métodos para prever el comportamiento de las ventas en diferentes regiones del país constituye otra aportación novedosa.

Utilización de herramientas tecnológicas avanzadas: R, Python (con librerías especializadas tipo pandas, numpy, ggplot2, etc.), y Power BI, ha permitido realizar un procesamiento y

análisis de datos profundo y desarrollar visualizaciones de gráficos e interactivas. El uso de estas tecnologías supone una innovación incremental clara para la empresa.

Análisis geográfico y territorial mediante clubes de convergencia: técnica poco común en proyectos empresariales que permite identificar provincias o regiones con comportamientos similares de ventas, lo que facilita la asignación de recursos y la adaptación regional de estrategias comerciales. Esta innovación técnica aporta una dimensión territorial estratégica al análisis de datos.

Análisis temporal enriquecido: con variables derivadas como festivos nacionales y autonómicos, días de la semana y períodos vacacionales, aporta una visión más compleja y contextualizada del comportamiento de las ventas. Esta incorporación de información del calendario en el análisis de la serie temporal permite detectar estacionalidades y ajustar mejor las previsiones.

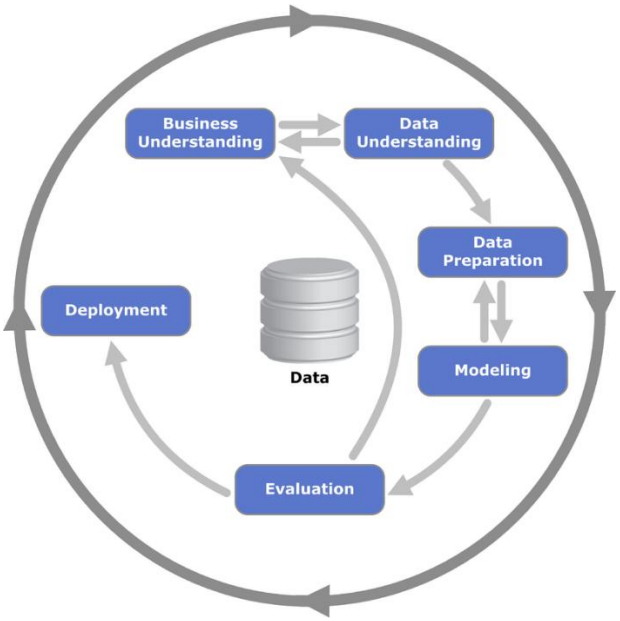
Implementación de un cuadro de mando visual: la creación de un *dashboard* dinámico e interactivo facilita la toma de decisiones por parte de perfiles no técnicos dentro de la organización. Este sistema permite una monitorización eficaz del negocio, tanto a nivel global como desagregado por producto, punto de venta o región, lo cual representa un avance significativo respecto a los informes estáticos tradicionales.

En conjunto, estos elementos innovadores no solo mejoran la capacidad analítica y predictiva de la organización, sino que suponen un salto cualitativo en la cultura de gestión de la información. El enfoque integral adoptado en este proyecto, que combina tecnologías modernas con la adopción de estrategias corporativas de integración de datos, demuestra que la innovación en el ámbito empresarial no se limita a la creación de nuevos productos, sino que puede —y debe— materializarse también en los procesos de toma de decisiones, análisis estratégico y diseño de soluciones adaptadas al entorno.

2. Alcance y planificación

El desarrollo de este trabajo se lleva cabo mediante el método Crisp-DM (Cross-Industry Standard Process for Data Mining) que es una metodología de trabajo para la organización de proyectos de análisis de datos y en consecuencia de Inteligencia de Negocio. En la siguiente figura se muestra de manera gráfica el proceso de la metodología Crisp-DM.

Figura 1. Metodología Crisp-DM



Fuente: www.healthdataminer.com (Vallalta Rueda, s.f.)

La primera fase es el entendimiento del negocio con lo que, en este apartado de alcance y planificación del Proyecto de Inteligencia de Negocio, se aborda este primer step de entendimiento. Donde se va a recolectar información específica sobre la empresa y el mercado en el que opera. Así mismo, se desarrollará el alcance del proyecto, los gaps existentes, la planificación de este, los recursos humanos necesarios y los costes.

Obteniendo así de este análisis, información relevante, con la que conseguir los objetivos estratégicos planteados por la organización a medio y largo plazo, así como los de este proyecto en concreto.

2.1. Fase de descubrimiento: evaluación del entorno actual

En esta fase de descubrimiento se ha de analizar los requisitos comerciales y de sistemas de información, las capacitaciones y formación de los miembros de la organización que intervienen y la arquitectura de datos y tecnología. (Sherman, 2014). Y así obtener información acerca de:

- Necesidades de herramientas y datos de la organización. Análisis e información deseados.
- Prioridades y necesidades de conocimiento en Altadis, para conocer qué es lo más importante para ellos para la toma de decisiones.
- Carencias actuales y deficiencias en los datos.
- Habilidades analíticas de los recursos humanos generales y miembros directivos.

En la primera fase de la metodología Crisp-DM, se expone la necesidad del entendimiento del negocio, la empresa y de los datos recibidos, por lo que se ha investigado la compañía y los datos recibidos de la misma, que a continuación se detallan.

Altadis fue adquirida por Imperial Brands, una compañía de Origen Británico que cuenta en su cartera con 160 marcas distintas, 25.000 empleados, que operan en 120 mercados y cuenta con 27 plantas de fabricación (una de ellas en Cantabria, España). (Altadis, S.A., s.f.)

Imperial Brands y por consecuente Altadis declaran en su “Annual Report and Accounts 2024” su enfoque en el consumidor, la puesta de este en el centro de su negocio y su sólido conocimiento de este, que guía su toma de decisiones. Poseen marcas tradicionales con origen local e internacionales con un atractivo geográfico más amplio. Actualmente se dirigen principalmente a sus cinco mercados clave prioritarios que son España, Reino Unido, Estados Unidos, Alemania y Australia. Declaran contar con alianzas que le permiten una respuesta ágil a la dinámica cambiante del mercado y un rápido desarrollo de productos. (Imperial Brands PLC). En España tienen su sede principal en Madrid y además cuentan con una oficina en Gran Canaria, disponen de más de 400 profesionales en toda España (Imperial Brands Altadis, s.f.)

A continuación, se va a aportar información sobre los factores que dan forma al comportamiento actual del mercado, los agentes o fuerzas externas al ámbito empresarial y las tendencias emergentes del sector:

Para ello se ha realizado una exploración, del entorno sectorial del mercado de tabaco y de la posición de la empresa Altadis dentro de su sector, donde se ha tomado como fuente principal el informe anual 2024 de Imperial Brands PLC (Imperial Brands, s.f.) por ser una herramienta de gran valor que nos permite observar el posicionamiento estratégico de la empresa y la evolución del mercado en el contexto europeo y español.

- **Transformación del sector: productos de nueva generación (NGP)**

Actualmente el mercado de tabaco está sufriendo una transformación estructural, fomentada por la expansión de los productos de nueva generación (Next Generation Products, NGP), que se constituyen como alternativa al cigarrillo tradicional y tratan de reducir los efectos nocivos del consumo de nicotina, dando respuesta a la creciente búsqueda de opciones menos perjudiciales para la salud de los consumidores. Estos productos se pueden categorizar como:

- Dispositivos de vapeo, son unos aparatos a batería que calientan líquidos con nicotina hasta transformarlos en vapor para su inhalación.
- Tabaco calentado, similares a los anteriores, pero contienen tabaco.
- Nicotina oral: según el Ministerio de Sanidad (2023), las bolsitas de nicotina oral "contienen sales de nicotina junto con otras sustancias, entre las que se encuentran la celulosa microcristalina, carbonato de sodio y otras sales de carbónico, ácido cítrico y diversos aromas. En ningún caso contienen tabaco y por ello, y por su origen químico, a veces los definen como productos completamente "blancos"" (p. 5) (Ministerio de Sanidad. Gobierno de España, 2023).

Estos dispositivos están diseñados para atraer a consumidores adultos que quieren o necesitan reducir los riesgos para la salud o para cambiar su experiencia de consumo. El crecimiento de estos productos ha diversificado el mercado, desarrollando un esquema más amplio de productos enfocado hacia nuevos estándares y preferencias de los consumidores.

- **Mercado global**

Según el informe Perspectiva del mercado de tabaco (Informesdeexpertos.com, s.f.) el mercado mundial de tabaco alcanzó un valor de más de 877 billones de dólares en 2024 y se estima que para 2034 el mercado crezca hasta los 1.057 billones de dólares. La Organización Mundial de la Salud (OMS) calcula que aproximadamente el 20% de la población adulta

mundial son consumidores de tabaco, lo que equivaldría a unos 1.250 millones de personas (Organización Mundial de la Salud, 2023).

El mercado mundial de tabaco se enfrenta a una gran presión y complejidad normativa y regulatoria, al crecimiento del comercio ilícito y a una gran carga fiscal, sobre todo en Europa.

- **Mercado español**

Para Imperial Brands, el mercado español supuso un 5% de su cifra de negocio global en ingresos por tabaco y NGP del grupo en 2024, por lo que España es uno de los cinco mercados más importantes para este grupo empresarial. Desde Imperial Brands se señala una leve recuperación de la cuota de mercado, de alrededor de 5 puntos básicos, debido a:

- La introducción de productos NGP.
- La revalorización de marcas locales impulsadas por una mayor sensibilidad del consumidor al precio.
- La adaptación de la oferta comercial a nuevos patrones de consumo.

No obstante, el mercado español encuentra las siguientes limitaciones:

- Elevada presión fiscal sobre el producto, que veremos de forma más detallada a continuación, así como la previsión de endurecimiento del marco regulador, con medidas como la posible implantación del empaquetado genérico. (Cadenas, 2024).
- Proliferación de la actividad comercial ilegal.
- Transición del consumidor hacia alternativas no combustibles.

- **Marco impositivo aplicado en España a productos del tabaco y productos de nueva generación (NGP)**

El régimen tributario sobre los productos de tabaco en España se articula en la Ley 38/1992, de 28 de diciembre, de Impuestos Especiales (Ministerio de Hacienda, 1992) junto con sus modificaciones posteriores, en especial las que afectan a productos emergentes como los cigarrillos electrónicos y los líquidos de vapeo. A través de esta ley, se regula la fiscalidad aplicable a cada categoría de producto con relación a elementos como:

- Si el producto lleva tabaco y cuál es la cantidad de tabaco contenida.
- Si se inhala mediante combustión o no.

- Si se trata de productos tradicionales, como los cigarrillos, o bien de productos de nueva generación.
- Y si el producto contiene nicotina o está exento de ella.

A continuación, se muestra una tabla comparativa de régimen fiscal aplicable a productos de tabaco y de nueva generación.

Tabla 1. Régimen Fiscal aplicable a productos del tabaco y productos de nueva generación (NGP).

Régimen fiscal aplicable a productos del tabaco y productos de nueva generación (NGP)

PRODUCTO	TIPO DE FISCALIDAD APLICABLE	OBSERVACIONES
CIGARRILLOS	- Impuesto específico: fijo por cada 1.000 unidades (59,1 €) - Impuesto ad valorem: 51% sobre el PVP - Mínimo obligatorio: 136,8 € por 1.000 unidades	Se grava tanto el volumen como el valor, se garantiza una tributación mínima. Es el producto con mayor carga impositiva del sector.
TABACO DE LIAR	- Impuesto específico: calculado por peso (€/kg de tabaco) - Impuesto ad valorem: % sobre el PVP - Mínimo obligatorio: aplicable según categoría	La carga fiscal es inferior que en los cigarrillos, aunque también sigue un modelo mixto de gravámenes, por lo que es más asequible para el consumidor.
TABACO CALENTADO	- Impuesto específico: por peso neto del tabaco contenido - No se aplica ad valorem - IVA general (21%)	Se diferencia fiscalmente del cigarrillo. Se le aplica una carga tributaria intermedia.
LÍQUIDO PARA CIGARRILLOS ELECTRÓNICOS CON NICOTINA	- Impuesto especial: 0,13 €/ml (desde 2021) - IVA general: 21%	La tributación de este tipo de prodcutos se recoge en los Presupuestos Generales del Estado de 2021. Son productos con carga fiscal moderada.
LÍQUIDO PARA CIGARRILLOS ELECTRÓNICOS SIN NICOTINA	- Solo IVA (21%) - No se aplica impuesto especial	Actualmente no está gravado con impuestos especiales, pero el Ministerio de Hacienda está estudiando su futura inclusión si se generalizara el consumo
DISPOSITIVOS ELECTRÓNICOS (HARDWARE)	- Solo IVA (21%) - No se aplica impuesto especial	El uso de este producto no genera un impacto negativo inmediato por sí mismo sobre la salud pública que justifique una tributación específica adicional, por lo que no se le grava con ningún impuesto especial

Fuente: elaboración propia en base a datos de la Ley 38/1992, de 28 de diciembre, de Impuestos Especiales

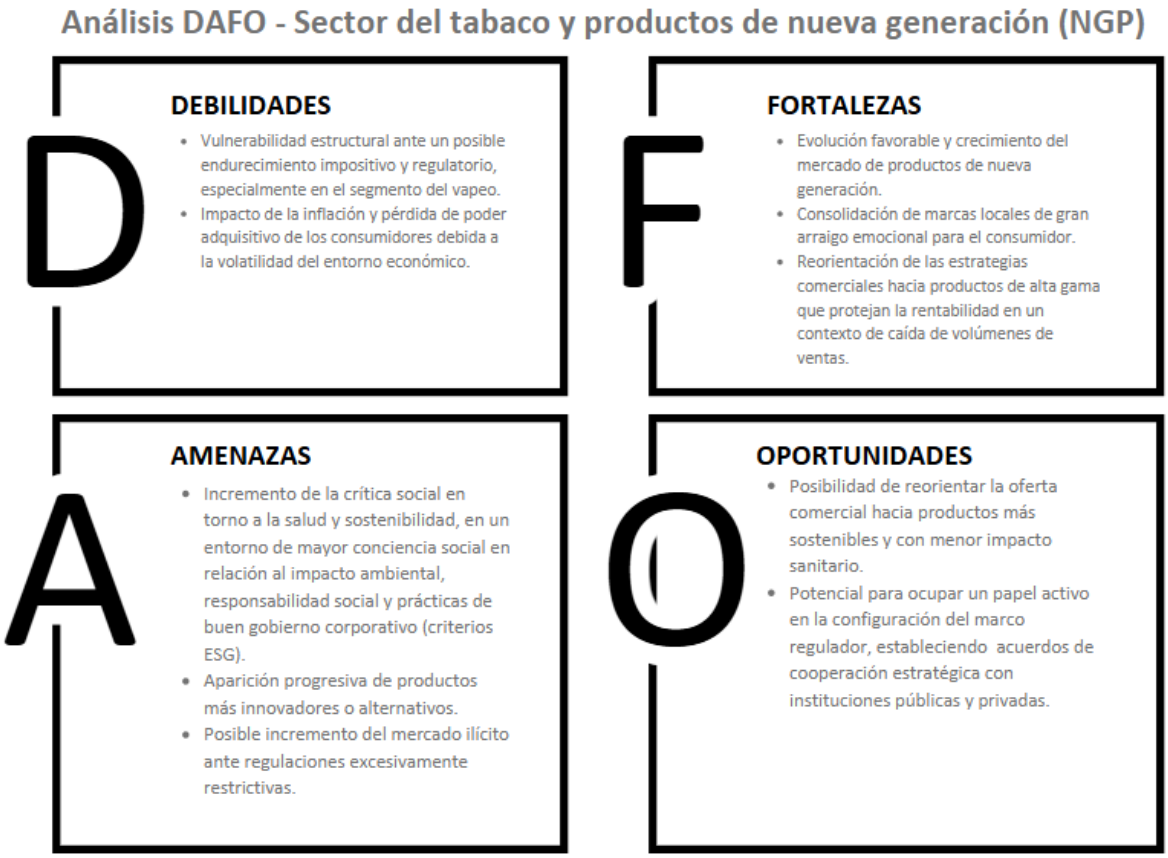
Conociendo estos datos, se estima que aproximadamente el 75% del precio de venta al público del paquete de cigarrillos corresponde a impuestos, soportando el consumidor final un coste muy por encima del coste de producción y distribución, por lo que cabe deducir que la presión impositiva influye en el comportamiento del consumidor, promoviendo que:

- El consumidor busque alternativas más económicas como, por ejemplo, el tabaco de liar que tiene una fiscalidad más baja que el cigarrillo tradicional.
- El producto se obtenga en canales no regulados o ilícitos, que no aplican impuestos, por lo que el precio de venta al público es sustancialmente inferior.
- El consumo tienda a dirigirse hacia productos de nueva generación ya que actualmente tienen una menor carga fiscal, por lo que resultan más atractivos económicamente.

- **Conclusión: Análisis de Debilidades, Amenazas, Fortalezas y Oportunidades del sector del tabaco y productos de nueva generación (NGP)**

Tras analizar el contexto del mercado global y español y la transición que está viviendo el mercado de tabaco, se ha desarrollado el siguiente análisis DAFO con los factores internos y externo que afectan al sector del tabaco:

Figura 2. Análisis DAFO sobre el sector del tabaco y los productos de nueva generación (NGP).



Fuente: elaboración propia a partir de la información del Informe Annual 2024 de Imperial Brands

2.1.1. Información deseada

La información que se desea generar en este proyecto de BI es:

- Hábitos de consumo y conocimiento profundo del comportamiento de los consumidores, la temporalidad en las ventas y la previsión de la evolución futura.
- Conocimiento de los clientes de la compañía en España, logrando una segmentación de clientes detallada y detección de patrones de comportamiento en los datos.
- Se desea conocer el comportamiento geográfico de las ventas que permita incentivar las mismas, la estrategia comercial, la distribución de los productos y la toma de decisiones regionales.

2.1.2. Información actual: deficiencias y soluciones alternativas.

Con el fin de averiguar las posibles deficiencias y soluciones alternativas, se ve necesario conocer cómo se estructura su Inteligencia de Negocio. En este contexto, con la información facilitada por José Luis Gómez, Chief Data Officer de Altadis a la Universidad Internacional de la Rioja, mediante el briefing sobre los Principales Conceptos del Negocio. Se sabe que la compañía extrae datos del flujo comercial de la empresa "Route to market" esto es desde la fabricación hasta el consumo.

Los elementos de "Route to market" son **Producto – Distribuidor - Cliente – Actividades en el punto de Venta - Consumidor** (Imperial Brands.Principales Conceptos de Negocio., s.f.)

- En Producto, el origen de los datos siempre es interno y si, se tiene en cuenta las características del producto, es marketing quien aporta esa información.
- El Distribuidor, es el que se encarga de proporcionar los datos de sell-in y sell-out.
- En Cliente según la modalidad de venta se diferencian dos elementos:
  - Estancos que proporcionan información de sell-out de las ventas a través de los TPVs.
  - Maquinas Vending que proporcionan información de las ventas gracias al proyecto MiVending que colecta y recoge datos a través de los modem instalados en las propias máquinas.
- Actividades en el punto de venta:
  - CRM datos introducidos por la fuerza de ventas, usuarios de la oficina y logística.
  - Empresas Colaboradoras son fuentes de datos externas como por ejemplo empresas de azafatas o compañías de instalación de mobiliario y máquinas de vending.
- Consumidores research información obtenida de encuestas periódicas realizadas a los consumidores.

En la actualidad, Imperial Tobacco Group, mantiene un programa de Inteligencia Empresarial que cuenta con:

- Un Data Warehouse, el lugar centralizado donde almacenan todos los datos extraídos.

- El proceso de extracción, transformación y carga (ETL) mediante el cual convierten los datos obtenidos de las fuentes anteriormente mencionadas en información útil y entendible.

- Microstrategy una de las herramientas de inteligencia de negocio que emplean, y es un software de inteligencia empresarial, que se usa para facilitar las consultas, el análisis y la elaboración de informes de datos empresariales para todos los usuarios (Imperial Brands.Principales Conceptos de Negocio., s.f.)

Para abordar este proyecto de BI, se proporcionan por parte de la compañía seis conjuntos de datos anonimizados en archivos Excel:

- Affiliated_Outlets: contiene información sobre establecimientos afiliados (clientes de la empresa), tales como el código de cliente, el código postal, el tipo de compromiso, segmentación, ubicación, tamaño en metros cuadrados del local.

- Product: contiene el detalle de los productos comercializados a los clientes, tales como el código de producto, el tamaño y el formato.

- OoSDay: comprende el registro diario de incidencias por rotura de stock e incluye la fecha, el código de cliente y el código de producto.

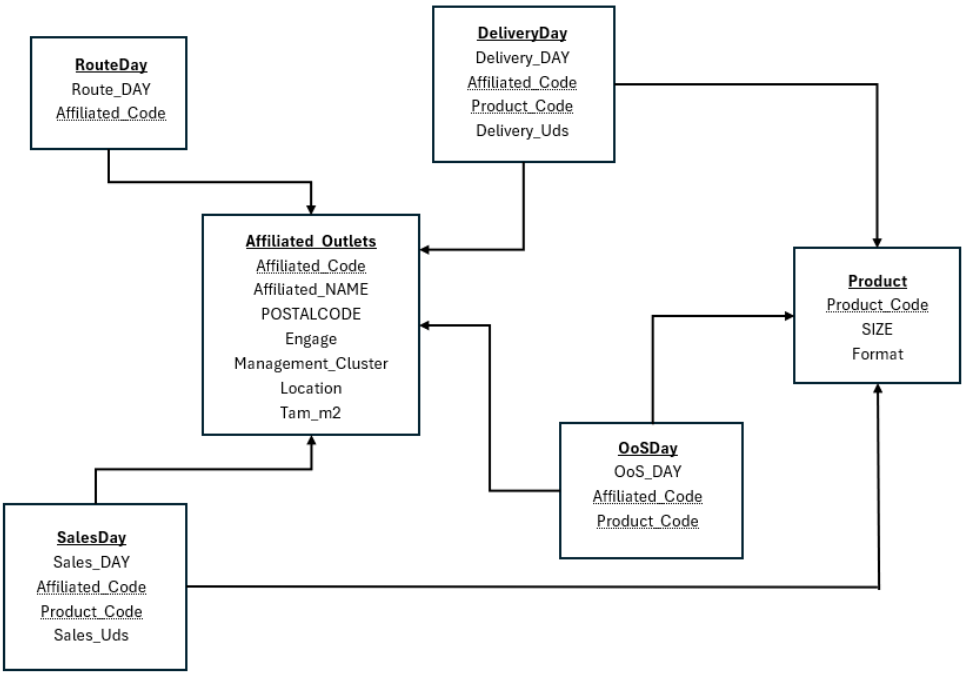
- DeliveryDay: incluye los datos de entregas realizadas o reposición de mercancía, representando las compras realizadas por los estancos por fecha. Se recoge el código de cliente, el código de producto y las unidades entregadas.

- RouteDay: recoge información sobre las rutas de entregas realizadas, incluyendo la fecha y el código de cliente.

- SalesDay: incorpora las ventas diarias registradas en cada establecimiento, detallando la fecha de venta, el código de cliente, el código de producto y las unidades vendidas.

A continuación, se muestra el esquema de relación entre los conjuntos de datos proporcionados por Altadis.

Figura 3. Esquema de relación de las tablas disponibles.



Fuente: elaboración propia

Todos estos datos tienen carencias y necesitan ser enriquecidos de diferentes formas.

Es por ello qué, se empleará el enriquecimiento geográfico en la parte de clientes y ventas, ya que únicamente se dispone de los códigos postales, estos se asociarán a una región concreta del territorio español para así proceder a proporcionar conocimiento útil de sus clientes y de los consumidores de tabaco, buscando patrones de comportamiento geográfico.

Debido a la extrema anonimización de los datos que la compañía ha proporcionado, se ve un problema el abordar el conocimiento de las bases de datos y proporcionar información de valor. Se tratará de solventar esta cuestión con enriquecimiento de segmentos, descripciones y precios en la parte de producto mediante el empleo de fuentes externas abiertas o especializadas del sector tabacalero como es el Comisionado para el Mercado de Tabacos (Ministerio de Hacienda, s.f.). Así como mediante el estudio pormenorizado de los datos disponibles para hacer posibles suposiciones y/o averiguaciones.

Otra deficiencia importante que solventar en este trabajo es el escaso periodo de tiempo de los datos, lo que complicará el realizar predicciones, puesto que:

- Los datos disponibles en el archivo SalesDay son tan sólo desde el 09/03/2015 al 06/09/2015, esto es apenas 6 meses.
- Los datos disponibles en el archivo RouteDay son tan sólo desde el 09/03/2015 al 11/12/2015, esto es poco más de 9 meses.
- Los datos disponibles en el archivo DeliveryDay son tan sólo desde el 09/03/2015 al 16/10/2015, esto es poco más de 7 meses.

Pero para conseguir mitigar y abordar esta deficiencia se centrará el análisis en periodos más cortos de tiempo centrándose en establecer patrones semanales y/o mensuales intentando encontrar micro-tendencias que puedan ser de utilidad en la planificación operativa de la empresa, así mismo también se usarán técnicas de predicción que no necesiten largos periodos de tiempo como por ejemplo el modelo autorregresivo ARIMA sin estacionalidad y si es viable se podrán generar datos sintéticos para pruebas de modelo.

Debido a que los datos disponibles son de hace 10 años y tanto la compañía como el mercado ha cambiado y evolucionado en este periodo de tiempo, se ve la necesidad de un enriquecimiento temporal de los datos, actualizándolos, es decir analizar el comportamiento del consumidor en la actualidad mediante, por ejemplo, comparativas de métricas claves o simulaciones de nuevos escenarios.

2.1.3. Habilidades analíticas actuales.

Altadis es una multinacional que cuenta con departamento de Inteligencia y planificación, su organigrama muestra un departamento llamado “Insights & Intelligence” que como se sabe los Insights son la averiguaciones e información obtenida de los datos y la inteligencia es el proceso de análisis y generación de valor de los datos e información obtenida. (The Official Board, 2024).

Gracias a la formación de los miembros del departamento anteriormente mencionado y a la de los componentes del equipo de elaboración de este proyecto que cuentan con conocimientos técnicos de lenguaje de programación para el análisis de datos y elaboración de informes profesionales en las herramientas necesarias, para la elaboración de cuadros de mando como por ejemplo Power BI, se parte de que el entendimiento entre las partes es fluido y puede contener información técnica alta.

2.2. Fase de análisis: identificación de gaps

En esta fase de análisis, se ha evaluado toda la documentación aportada por Altadis a la UNIR, así como los materiales e información recabada en la fase de descubrimiento. El análisis implica examinar la arquitectura de los datos, la organización y los requisitos funcionales con el fin de identificar donde se encuentra hoy la empresa y a dónde le gustaría llegar con el proyecto de Inteligencia de Negocio.

Se trata de acercar la realidad de la empresa con los objetivos establecidos, las mejores prácticas del sector y las expectativas del mercado. Basándonos en datos, herramientas y metodologías propias de la inteligencia de negocio, analizaremos las brechas tecnológicas y operativas que puedan limitar el rendimiento de la empresa, logrando un mayor conocimiento de sus oportunidades y que se tomarán como base para el futuro diseño de estrategias de optimización.

2.2.1. Capacidad de los informes actuales

Actualmente se cuenta con informes no provenientes directamente de la compañía y si disponibles y obtenidos del entorno web, de ellos se puede extraer gran cantidad de información que facilite el desarrollo de este proyecto de Inteligencia de Negocio, en el contexto del conocimiento del consumidor español y del mercado del país, a continuación, se detallan los mismos:

- Informe titulado “El mercado de tabaco español desde un análisis de su demanda”. Informe que estudia la elasticidad de la demanda al aumento de precio abordando los efectos renta y sustitución. Elaborado por Rocío Ingelmo Heras para Altadis Imperial Brands. (Ingelmo Heras)
- Estadísticas Mercado de tabacos. Donde encontramos gran cantidad de información sobre la evolución de las ventas del mercado de tabacos, por tipo de producto y distribución geográfica que aporta un fuerte punto de partida para este proyecto. (Ministerio de Hacienda, s.f.)
- Informe sobre el posible impacto en el mercado de tabaco la introducción del empaquetado genérico. Que aporta información relevante sobre la introducción de nuevas medidas disuasorias del consumo. (Analistas Financieros Internacionales, S.A.)

Se sabe que la compañía dispone de distintos informes y emplea el uso de cuadros de mando elaborados en MicroStrategy (fuerte plataforma de inteligencia empresarial con acceso a la gestión de datos y la creación de informes y análisis), pero actualmente se carece de acceso a ellos para elaborar este trabajo.

Por otro lado, en la información actual de los datos aportados por Altadis para la elaboración de este proyecto:

- En relación con las ventas registradas, existen códigos postales, aunque no están claramente asociados a una jerarquía geográfica completa por lo que no se aprovecha completamente la localización para su análisis.
- En relación con los archivos relativos a ventas, entregas y rutas son registros planos sin la posibilidad de visualizar en tiempo real la información, cruzar KPIs o encontrar tendencias fácilmente.
- La información sobre los puntos de venta contiene datos del tamaño de los locales, su ubicación y el tipo, pero no se dispone de una clusterización exhaustiva y formal.
- Se registran datos de rutas y entregas, pero no existe análisis integrado con los que optimizar y evaluar la eficiencia de las entregas.

2.2.2. Proveedores de tecnología necesarias

Para realizar el análisis de los datos y la elaboración del dashboard se emplearán los siguientes proveedores:

- Herramientas de visualización: dado que los ficheros facilitados por la empresa se encuentran en Excel, se empleará Power BI por su capacidad para conectarse directamente a Excel, su rápida implementación y capacidad para crear visualizaciones interactivas.
- Herramientas de procesamiento de datos: Power BI para proceso ETL y automatizar las visualizaciones.
- Herramientas de procesamiento y análisis avanzado: se empleará R Studio para el análisis de los datos, elaboración de clúster, enriquecimiento geográfico, etc. Y, además, se usarán las librerías NumPy y pandas de Python que nos permitirán realizar un preprocesamiento de los datos, donde podremos limpiar y transformar las bases

de datos con las que vamos a trabajar y realizaremos análisis más avanzados. Con Python también se podrán realizar análisis avanzados con machine learning.

Los resultados que se extraigan de este estudio se implementarán en Microestrategy ya que, en la actualidad, la empresa usa licencias en esta plataforma que es más potente en inteligencia de negocio empresarial al manejar mejor grandes volúmenes de datos, ofrecer mayor seguridad y escalabilidad.

2.2.3. Diferencia entre los informes actuales y la información deseada

El conocimiento del sector que aportan los informes actuales facilitará la comparativa del mercado actual con el análisis propio de los datos aportados por la misma para llevar a cabo el cuadro de mando de este proyecto. No obstante, al carecer de acceso a los dashboards elaborados por Altadis, se parte de cero en este aspecto ya que se desconoce su capacidad de aportar valor al negocio.

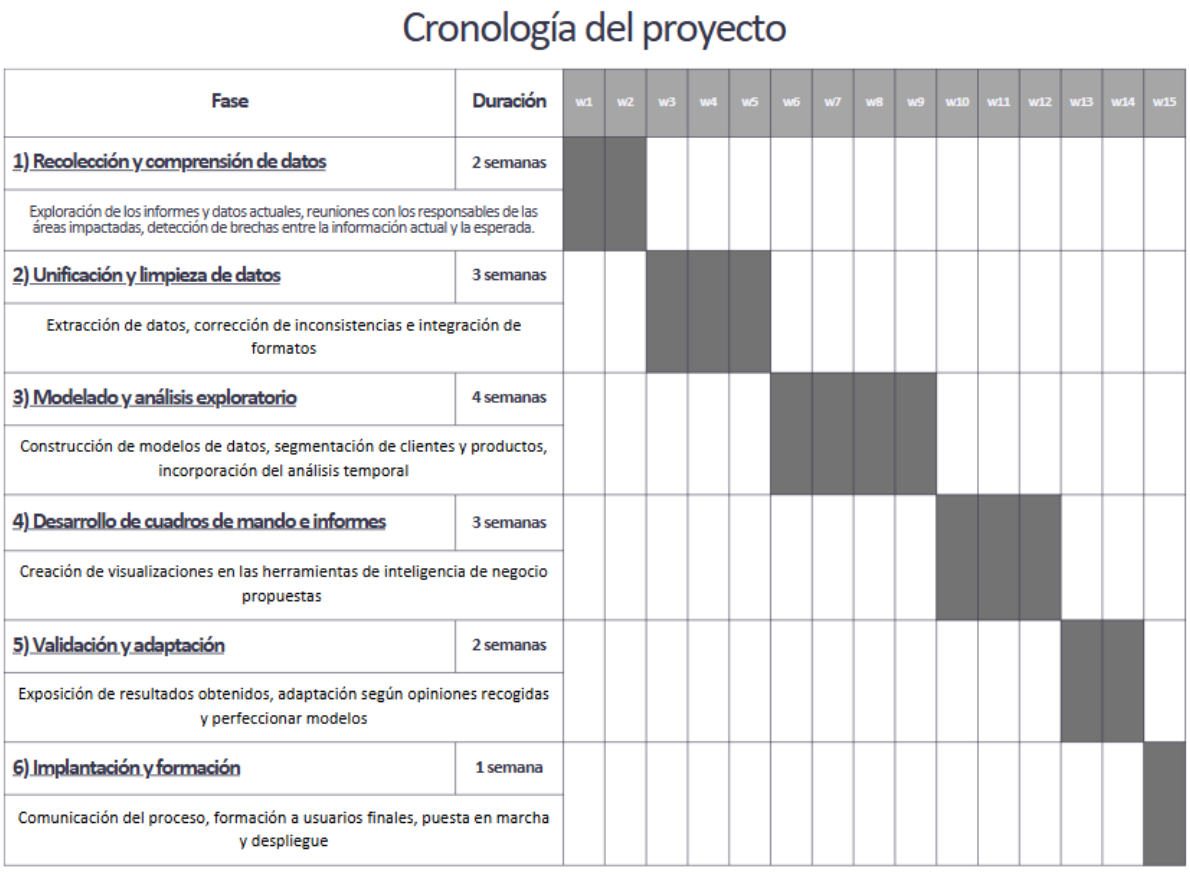
Los Gaps encontrados entre la información que se desea encontrar expuesta en el apartado anterior de 2.1.1. y la información actual, se resumen a continuación:

- Falta de análisis geográfico detallado.
- Carencia de visualización dinámica y analítica de los informes.
- Ausencia de cuadros de mando interactivos que faciliten la toma de decisiones en tiempo real, basada en datos.
- Falta de capacidad predictiva.
- Segmentación avanzada de clientes.

2.2.4. Cronología, costes y recursos humanos implicados

Dividiremos el proyecto de análisis en 6 fases que se desarrollarán en un tiempo estimado de entre 3 y 4 meses.

Figura 4. Cronología y fases en las que se desarrolla el proyecto.



Fuente: elaboración propia

Se estiman unos costes de entre 36.000€ y 150.000€, dependiendo de si se emplea un equipo interno o una consultoría externa. En estos costes se incluyen las licencias para las herramientas de inteligencia de negocio y bases de datos externas, los salarios de los equipos empleados, la inversión en infraestructura tecnológica y costes en formación y documentación, con el siguiente desglose:

a) Licencias:
   - Programas de inteligencia de negocio: para realizar el estudio y análisis objeto de este proyecto se decide comprar 2 licencias de Power BI Pro con un coste total por proyecto de 80€ y se utilizan las licencias gratuitas de Python
   - Compra de la base de datos oficial de códigos postales de Correos con un coste total por proyecto de 282€

b) Salarios: según el Informe de Tendencias Salariales (Randstad Research, 2025) los costes aplicables al proyecto serían un total de 31.900€. Se realiza un detalle de los perfiles y dedicaciones a continuación.
c) Infraestructura tecnológica: se decide trabajar con infraestructura en la nube por lo que los costes de los servidores y seguridad para el total del proyecto serían unos 240€.
d) Formación: se realizará una formación a los usuarios finales de los reportes y visualizaciones que asciende a unos 2.500€.

El equipo humano que se empleará en este proyecto estará compuesto por analista de inteligencia de negocio, ingeniero de datos, jefe de proyecto, consultor de inteligencia de negocio, consultor de negocio y consultor técnico de inteligencia de negocio. Si se necesitara un equipo reducido, se debería contar con un perfil mixto de analista de datos con conocimiento en ETL e inteligencia de negocio, que podría asumir parte de las funciones del ingeniero de datos.

Figura 5. Equipo dedicado.

Equipo dedicado

	ROL	DEDICACIÓN ESTIMADA	COSTE ESTIMADO PROYECTO
ANALISTA IN	Análisis, desarrollo de visualizaciones, establece KPI's	100%	14.000€
INGENIERO DE DATOS	Extracción, transformación y carga de datos	50%	5.750€
JEFE DE PROYECTO	Coordinación del proyecto	25%	3.950€
CONSULTOR IN	Coordinación del proyecto, comunicación con interesados	50%	3.100€
CONSULTOR NEGOCIO	Define los objetivos y valida los reportes	25%	3.700€
CONSULTOR TÉCNICO IN	Configura herramientas IN	10%	1.400€
		Total proyecto	**31.900€**

Fuente: elaboración propia

2.3. Fase de recomendaciones: alcance, prioridades y presupuesto

Tras haber identificado las diferencias entre la situación actual y la situación esperada explicadas en los puntos anteriores, en esta fase se definirán las acciones a acometer para lograr los objetivos propuestos, donde se dispondrá de un plan estructurado que recoja el alcance del proyecto, se definirán las prioridades y se estimarán los recursos y presupuesto necesario. Asimismo, incluirá la estrategia para la promoción interna del proyecto, con el fin de consolidar su aceptación en la empresa, planteándose presentaciones prácticas que validen su utilidad y beneficios, antes de su implantación.

Una vez identificadas las diferencias entre el entorno actual y el escenario deseado, el siguiente paso será estructurar recomendaciones estratégicas para impulsar la evolución en la Inteligencia de Negocio en Altadis, que responderán directamente a los gaps detectados y a las necesidades estratégicas identificadas, proponiendo mecanismos de promoción interna y demostraciones prácticas que ayuden a aceptar y validar los beneficios del proyecto antes de su implantación completa.

Es por esto por lo que se ha definido un plan de acción donde se recoge de manera organizada el alcance del proyecto, la priorización de las áreas críticas de intervención y una estimación realista de los recursos y presupuesto necesario. Esta hoja de ruta establece los fundamentos que ayudarán a garantizar una implementación efectiva y flexible, adaptada a la realidad operativa de la empresa.

Partiendo de nuestro análisis en las fases de descubrimiento y evaluación, presentamos las siguientes recomendaciones para el proyecto de Inteligencia de Negocio. Estas recomendaciones responden directamente a las necesidades identificadas, las deficiencias actuales y los objetivos estratégicos de la organización.

Recomendación 1: definir una metodología de trabajo como estrategia de implementación del proyecto de BI.

El enfoque Waterfall implica una planificación secuencial y lineal, donde cada fase del proyecto (diseño, requisitos, desarrollo, pruebas e implementación) debe completarse antes de iniciar la siguiente. Esta metodología, al ofrecer una definición completa al inicio permite un diseño integral de la solución, mejor planificación y seguimiento más preciso del progreso.

Sin embargo, en la práctica, los usuarios de negocio suelen tener dificultad para definir completamente sus necesidades analíticas desde el principio, lo que genera cambios frecuentes en los requisitos durante el desarrollo.

Por otro lado, la metodología Agile representa un enfoque más moderno y flexible, basado en iteraciones cortas (sprints). Esta metodología prioriza la colaboración y las iteraciones constantes entre los usuarios involucrados en el proyecto, permitiendo un desarrollo más rápido y adaptable. Esta mayor participación facilita la evolución de los requisitos y una mejor adecuación a las necesidades reales. Sin embargo, también tiene sus limitaciones, ya que en un escenario práctico podría generarse falta de coherencia arquitectónica y comprometer la calidad global de la solución.

La recomendación para este proyecto es una metodología híbrida que tome lo mejor de las anteriores, es decir desarrollo iterativo e incremental del proyecto de Bi. Se recomienda utilizar principios incrementales para definir el programa general y sus lanzamientos, incorporando aspectos de Waterfall para establecer un marco arquitectónico consistente y gestionar interdependencias entre componentes, y aplicar desarrollo iterativo Agile para optimizar los resultados.

La metodología híbrida resulta particularmente adecuada para proyectos de BI porque reconoce una realidad fundamental: independientemente del esfuerzo dedicado a la recopilación inicial de requisitos, es inherente a estos proyectos que la comprensión de las necesidades analíticas evolucione durante el desarrollo, hasta llegar a la mejor solución posible.

Recomendación 2: enriquecer la información con fuentes externas.

Aunque contenido en el plan de Arquitectura de la Información y Datos, se quiere destacar aquí la posibilidad de incluir en las fuentes de información elementos cuantitativos y cualitativos de otras fuentes. Opciones de elementos cuantitativos podrían ser bases de datos públicas como las disponibles en el INE o Eurostat, mientras que ejemplos de aportes cualitativos podrían ser informes o estudios académicos sectoriales, como los analizados en la tesis doctoral del Profesor Juan Manuel Martin Álvarez, que relaciona ventas de cajetillas de tabaco con ciclos de expansión/ contracción económica (cita)

Una línea de investigación muy interesante es el análisis de sentimiento en internet, recogiendo y procesando datos mediante técnicas de text-mining desde redes sociales, mediante el cual se pueden extraer insights valiosos sobre tendencias o percepciones subjetivas frente a nuevos productos en el mercado tales como vaporizadores u otros.

Recomendación 3: adoptar un enfoque integral y estructurado, alineado con la estrategia de negocio, para implementar una solución de Inteligencia de Negocio, mediante la elaboración articulada de cuatro planes arquitectónicos.

Inspirados en las directrices entregadas por Sherman (2014), debemos definir clara y detalladamente los siguientes aspectos:

Plan de Arquitectura de Información: Define qué solución de Inteligencia de Negocio se debe construir en el contexto del negocio conteniendo de esta manera una noción del alcance del proyecto. Este plan debe detallar los procesos comerciales impactados, los departamentos y responsables que utilizarán la solución, y cómo esta se alineará con las prioridades establecidas por la alta dirección (ya sea aumentar ventas, reducir roturas de stock, cambiar la imagen corporativa o implementar automatismos en los sistemas de análisis).

Plan de Arquitectura de Datos: Establece a qué datos se accederá y cómo se procesarán. Incluye las aplicaciones comerciales y fuentes externas de información, definir los conjuntos de datos tales como tablas, columnas, campos, variables, así como requisitos de datos históricos, integración y transformación de datos, parámetros de calidad y limpieza de datos necesarios para cumplir con los objetivos prioritarios.

Plan de Arquitectura Tecnológica: Enumera qué tecnologías se necesitan para soportar las arquitecturas de información y datos. Se refiere a los sistemas y herramientas tecnológicas para el desarrollo del proyecto, que pueden ser elementos de hardware, software, licencias, o incluso requerimientos de know-how para su implementación.

Plan de Arquitectura del Proyecto: Define la metodología de implementación escogida (Híbrida) y cronogramas tales como Carta Gantt. El plan debe incluir la gobernanza del proyecto, definición de roles y responsabilidades (se recomiendo el uso de RACI Matrix). También mecanismos de comunicación, iteración y escalamiento, hitos de control para evaluar avances, criterios de aceptación para los entregables, y procedimientos de gestión de

cambios que permitan evaluar el impacto de modificaciones sobre el alcance, presupuesto y cronograma del proyecto.

Alcance

El proyecto comprenderá los siguientes aspectos:

- Integración de las actuales bases de datos de Altadis en un sistema de análisis agrupado.
- Modelado de datos que permitirá la segmentación de establecimientos, la predicción de ventas futuras y el análisis temporal y geográfico de la demanda.
- Diseño y despliegue del cuadro de mando dinámico que facilite la toma de decisiones en tiempo real.
- Formación del personal en el uso de herramientas y aplicaciones de Inteligencia de Negocio.

La puesta en marcha del proyecto se realizará de forma iterativa y se priorizarán las áreas de venta y distribución, según las necesidades indicadas por la empresa.

Prioridades

En relación con el análisis realizado en los puntos anteriores, se detectan como prioridades para la fase inicial del proyecto las siguientes:

1. Fortalecimiento de los datos de ventas y clientes.
2. Segmentación avanzada mediante técnicas de clustering de los establecimientos.
3. Optimización de las rutas de entrega y distribución de los productos.
4. Predicción de ventas por producto y provincia.
5. Desarrollo de cuadros de mando comprensibles y de fácil uso para los usuarios no técnicos.

Presupuesto estimado

El presupuesto estimado detallado se encuentra en el punto 2.2.4 del presente documento, pero en este punto se resume en el siguiente gráfico:

Figura 6. Tabla resumen presupuesto dedicado al proyecto en base a la realización por un equipo interno.

Presupuesto (equipo interno)

	Coste imputado al proyecto
Licencias Software (Power BI Pro, SQL Server Standard, base de datos Correos)	1.000 €
Salarios	31.900 €
Infraestructura tecnológica	240 €
Formación	2.500 €
Total	**35.640 €**

Fuente: elaboración propia

En el caso del proyecto de Inteligencia de Negocio que ocupa este estudio, se recomienda aplicar un modelo de uso de Software como Servicio (Software as a Service, SaaS) que permitirá una rápida puesta en marcha, una mayor escalabilidad en función del número de usuarios finales, así como la reducción de costes operativos, ya que no se requerirá infraestructura propia ni mantenimiento técnico exclusivo.

2.3.1. Promoción del proyecto en la organización

Para garantizar que el proyecto de Inteligencia de Negocio propuesto en este estudio se implante con éxito en Altadis, es primordial realizar una promoción interna ajustada a la realidad de la empresa, ya que el objetivo de dicha promoción es que tanto los perfiles técnicos como los usuarios de negocio apoyen y participen activamente en la implantación del proyecto.

Se recomienda que la promoción interna del proyecto se fundamente en las siguientes acciones:

- Desarrollar una campaña de comunicación interna, compuesta por presentaciones, sesiones informativas y talleres prácticos donde se muestren los beneficios que aportará la implantación del proyecto. Asimismo, se difundirán y presentarán los resultados preliminares obtenidos tanto en los portales web internos, como en boletines informativos o redes sociales propias que utilice la empresa.
- Reforzar la percepción positiva del impacto del proyecto de Inteligencia de Negocio a través de charlas y testimonios de promotores internos que participen en el proyecto desde sus fases iniciales.
- Establecer sesiones informales de aprendizaje donde se expliquen los avances conseguidos y se obtengan ideas y sugerencias de los participantes.

Con estas acciones se implicará a los usuarios, evitando resistencias al cambio y se favorecerá la aceptación de la nueva solución.

2.3.2. Demostraciones

Se propone realizar demostraciones prácticas antes del realizar el despliegue final de proyecto con los siguientes objetivos:

- Validar que las hipótesis de negocio son correctas, confirmando así que los cuadros de mando e indicadores de negocio propuestos satisfacen las necesidades de información de las distintas áreas.
- Generar confianza en el proyecto, ya que al mostrar ejemplos concretos de los nuevos informes de Inteligencia de Negocio se expondrán las distintas mejoras en el proceso de toma de decisiones.

Para ello, se proponen las siguientes demostraciones iniciales:

- Cuadro de mando de ventas regionales donde se muestre una visualización por provincia, evolución temporal y comparativa entre los distintos productos de la oferta comercial.
- Mapa de rutas de distribución, que servirán para optimizar las entregas y los tiempos de reparto, y para ello se integrarán los datos de "RouteDay.csv" y "DeliveryDay.csv".

- Análisis de segmentos de clientes, donde se usarán los clústeres obtenidos para los establecimientos incluidos en el conjunto de datos "Affiliated_Outlets.csv" y que servirán para personalizar y optimizar campañas comerciales.

Para realizar estas demostraciones se utilizarán una muestra representativa de los datos históricos recogidos en "SalesDay.csv", alimentados con información de productos de "Product.csv".

3. Análisis y definición

En este apartado se recogen los principales análisis que se realizaron a partir de los datos proporcionados por Altadis, con el objetivo de identificar patrones relevantes en el comportamiento de los establecimientos. Para ello, se combinaron herramientas estadísticas, visualización de datos y técnicas de aprendizaje no supervisado, con el objetivo de extraer información útil a partir de las características estructurales de los puntos de venta, sus niveles de venta y su localización. De esta forma, se ofrecerá a la dirección de Altadis una base analítica que le facilite la toma de decisiones estratégicas diferenciadas según el tipo de establecimiento y su ubicación geográfica.

3.1. Análisis de los datos a utilizar

Para el desarrollo de este trabajo se tomaron como base varios conjuntos de datos proporcionados por Altadis. Estos archivos fueron posteriormente enriquecidos y tratados para construir una base analítica robusta y adecuada para los distintos modelos aplicados. A continuación, se describen las principales fuentes de información utilizadas para llevar a cabo el análisis de segmentos y realización de clubes de convergencia, así como los empleados en el estudio de la temporalidad de las ventas.

3.1.1. Análisis datos a utilizar en análisis de segmentos y clubes de convergencia

1. Ventas diarias por punto de venta (SalesDay.csv)

Este archivo contiene el histórico de ventas realizadas por establecimiento, incluyendo los productos comercializados en cada operación. Incluye las siguientes variables:

- Sales_Day: fecha en la que se realizó la venta.
- Affiliated_Code: código identificador del establecimiento.
- Product_Code: código del producto vendido.
- Sales_Uds: unidades vendidas en cada transacción.

2. Información estructural y geográfica de los puntos de venta (puntos_venta_enriquecido2_sinNA.xlsx)

Este conjunto de datos fue construido a partir del dataset original facilitado por Altadis, Affiliated_Outlets.csv. A partir del código postal, se incorporó información geográfica

adicional – municipio y provincia – mediante fuentes externas como Geonames (geonames.com). Incluye las siguientes variables:

- Affiliated_Code: código identificador del establecimiento.
- Affiliated_Name: nombre codificado del establecimiento.
- CP: código postal del municipio.
- Engage: aunque la empresa no especifico la definición exacta de esta variable, se interpretó como el nivel de implicación comercial del establecimiento con la marca.
- Managemente_Cluster: tampoco se definió explícitamente por Altadis, por lo que se interpretó como una categoría que refleja la necesidad de atención o gestión comercial específica.
- Location: aunque Altadis no proporciona la definición de esta variable se interpretó como una clasificación de la ubicación del punto de venta (zona rural, de ocio, urbana, etc.).
- Tam_m2: tamaño del local, agrupado por intervalos de superficie.
- Municipio y provincia: datos geográficos derivados del código postal.

Figura 7. Tabla resumen del dataset estructural puntos_venta_enriquecido2_sinNA.xlsx.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3583 entries, 0 to 3582
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   Affiliated_Code     3583 non-null   object
 1   Affiliated_NAME     3583 non-null   object
 2   CP                  3583 non-null   int64
 3   Engage              3583 non-null   int64
 4   Management_Cluster  3583 non-null   int64
 5   Location            3583 non-null   object
 6   Tam_m2              3583 non-null   object
 7   municipio           3583 non-null   object
 8   provincia           3583 non-null   object
dtypes: int64(3), object(6)
memory usage: 252.1+ KB
```

Fuente: elaboración propia

3. Información de productos (Product.csv)

Aunque finalmente este archivo no fue incluido en los análisis cuantitativos, ya que el estudio se centró en el comportamiento agregado de ventas por establecimiento, sirvió de referencia para entender la estructura de la cartera comercial.

- Product_Code: código interno de producto.
- SIZE: tamaño del producto.
- Format: formato de presentación del producto.

4. Integración de las fuentes

A partir de la unión de los conjuntos de datos SalesDay.csv y puntos_venta_enriquecido2_sinNA.xlsx, se generó una tabla maestra que permitió agregar y analizar la información de ventas por tienda y provincia. Gracias a esta integración, se habilitaron tres enfoques analíticos, en nuestra opinión, complementarios entre sí: la segmentación estructural de puntos de venta, la segmentación según rendimiento comercial y el análisis territorial mediante clubes de convergencia. De esta forma, se logró una visión más completa y estratégica del comportamiento comercial en la red de establecimientos.

3.1.2. Análisis datos a utilizar en análisis de temporalidad de las ventas

El estudio de la temporalidad de las ventas se ha llevado a cabo con el dataset llamado "ventas_completo_2" que se ha realizado con la integración de los siguientes datasets:

1.- "Product" y "Sales Day": Dos de los proporcionados por Altadis y de los cuales ya se han explicado sus variables anteriormente.

2.- "puntos_venta_enriquecido2_sinNA": elaborado en este trabajo para el desarrollo del estudio de la temporalidad, de los clústeres y de los clubes de convergencia.

Y este se ha creado mediante la unión de:

- "Affiliated_Outlets" proporcionado por Altadis y del cual también se han explicado sus variables anteriormente.

-Y "codigos_postales_2025_españa_geonames.org.xlsx" el cual ha sido elaborado a raíz del documento de texto descargado de la web www.geonames.org donde se han obtenido los

municipios y provincias los códigos postales presentes en el data set aportado originalmente. Obteniendo así el buscado enriquecimiento geográfico. (GeoNames, s.f.)

Se obtiene un data set con 2878182 observaciones de 14 variables:

- Variables claves utilizadas en el análisis de temporalidad: "Sales_DAY", "Sales_Uds", "Product_Code", "Affiliated_Code", "CP" y variables geográficas como provincia y municipio.
- Dimensión temporal: La información de ventas comprende un periodo corto de menos de un año, concretamente desde el 9 de marzo al 4 de octubre de 2015, esto es información de las ventas de tan sólo 210 días.
- Estadísticos descriptivos: La variable "Sales_Uds" tiene un valor medio diario de 1958 unidades con una alta desviación ya que su máximo es 351 unidades y su mínimo es –57 esto es devoluciones o perdidas. y la variable SIZE tiene un mínimo de 43 un y un máximo de 1415.
- Localización de los datos: El data set incluye las ventas en 48 provincias, todas las peninsulares españolas y la insular Baleares.

Al cual para elaborar diferentes gráficos y análisis exploratorios se le ha implementado enriquecimiento temporal con nuevas columnas, variables derivadas, del tipo tiempo, como días festivos nacionales y de las CCAA y/o generando nuevos data sets y vectores de fechas mínima, máxima y pertenencia de cada provincia a CCAA.

3.1.3. Análisis datos a utilizar en la elaboración del dashboard

La elaboración del cuadro de mandos de se ha llevado a cabo con los data set anteriormente expuestos y explicados: "Product", "Sales Day", "puntos_venta_enriquecido2_sinNA", y con la creación de una nueva tabla llamada "Calendario" con código Dax.

Al dataset "puntos_venta_enriquecido2_sinNA", se le ha implementado enriquecimiento geográfico con una nueva columna, con la pertenencia de cada provincia a su correspondiente CCAA.

La tabla “Calendario” ha sido creada en Power Query con código M, con la que puede realizar visualizaciones de forma directa y sencilla ya que tiene todo tipo de información de calendario como, día de la semana, trimestre, comienzo de mes, fin de mes, etcétera.

A continuación, se muestran imágenes de los encabezados de la tabla anteriormente mencionada.

Tabla 2. ***Tabla Calendario (parte 1).***

1 B01 Calendario =

Date	Año	Inicio año	Fin año	Número mes	Inicio mes	Fin de mes	Días mes	Año mes número	Año mes	Día mes número	Día semana	Día semana abr
01/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	1	jueves	ju.
02/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	2	viernes	vi.
03/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	3	sábado	sá.
04/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	4	domingo	do.
05/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	5	lunes	lu.
06/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	6	martes	ma.
07/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	7	miércoles	mi.
08/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	8	jueves	ju.
09/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	9	viernes	vi.
10/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	10	sábado	sá.
11/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	11	domingo	do.
12/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	12	lunes	lu.

Fuente: elaboración propia

Tabla 3. ***Tabla Calendario (parte 2).***

1 B01 Calendario =

Día semana número	Día año número	Mes	Mes abreviado	Trimestre número	Trimestre	Trimestre año número	Trimestre año	Fecha inicio trimestre	Fecha fin trimestre	Número
5	1	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
6	2	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
7	3	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
1	4	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
2	5	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
3	6	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
4	7	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
5	8	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
6	9	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
7	10	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
1	11	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
2	12	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
3	13	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
4	14	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
5	15	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
6	16	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
7	17	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	

Fuente: elaboración propia

Tabla 4. ***Tabla Calendario (parte 3).***

	Trimestre año	Fecha inicio trimestre	Fecha fin trimestre	Número semana año	Fecha inicio semana	Fecha fin semana
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	1	28/12/2014 0:00:00	03/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	1	28/12/2014 0:00:00	03/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	1	28/12/2014 0:00:00	03/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	3	11/01/2015 0:00:00	17/01/2015 0:00:00

Fuente: elaboración propia

3.2. Análisis histórico y/o limpieza de datos

En los siguientes epígrafes se detallan las tareas que se realizaron antes de aplicar los modelos analíticos, agrupadas según su naturaleza metodológica. Estas tareas comprendieron un proceso riguroso de depuración, validación y estructuración de los datos, lo que ayudó a garantizar que la información estuviese en condiciones óptimas para el análisis. Gracias a ellas, se redujeron posibles sesgos y errores que pudieran distorsionar los resultados.

3.2.1. Análisis histórico y/o limpieza de datos: Análisis Segmentación y Clubes de Convergencia

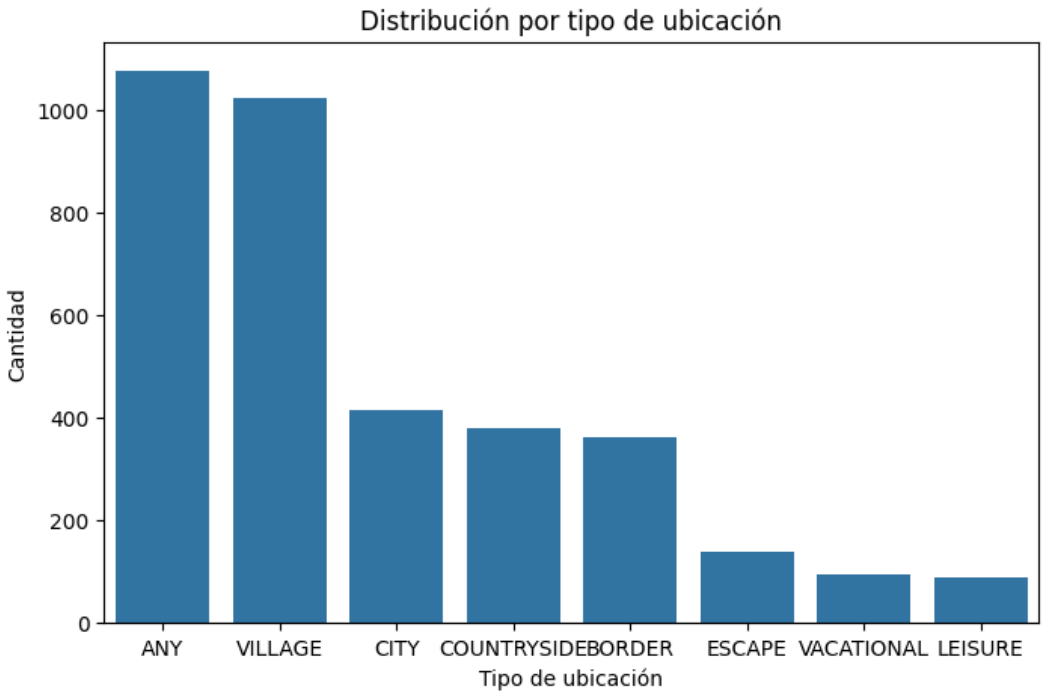
A continuación, se desarrolla en epígrafes diferenciados para la facilidad de lectura lo realizado en los análisis mencionados.

3.2.1.1. Depuración de registros incompletos o inválidos y estandarización de los datos estructurales.

Como primer paso, se revisó la calidad y completitud de los datos, donde se detectaron y gestionaron registros incompletos de forma diferenciada según el enfoque analítico:

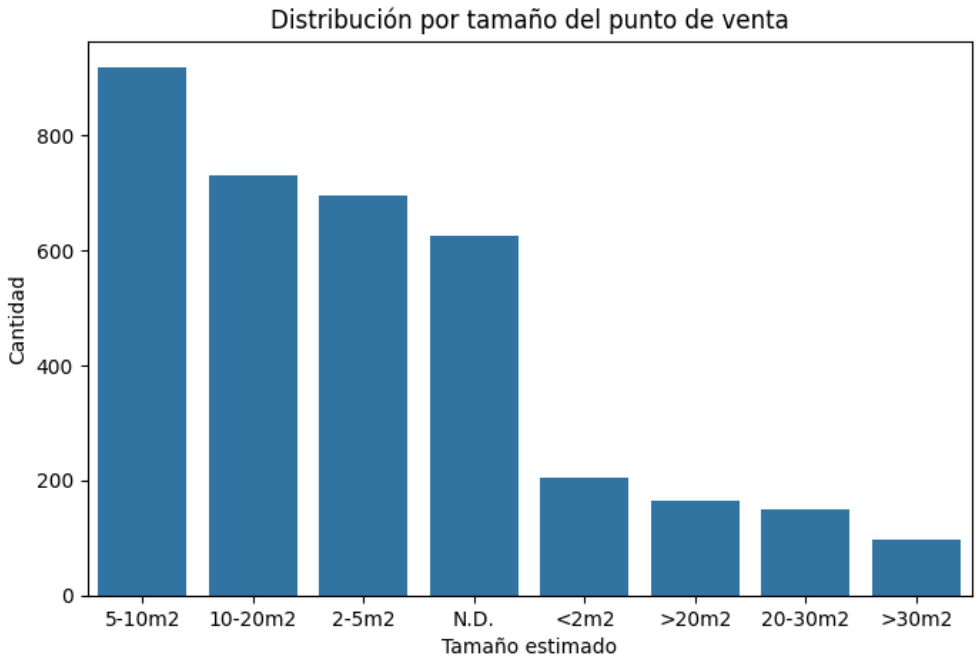
- En los modelos de clusterización - estructural y comercial – se comprobó que no hubiera valores nulos en las variables (Location, Tam_m2, Engage y provincia). Como estas observaciones estaban completas, no fue necesario aplicar funciones de eliminación directa, como dropna. Se generaron distintos gráficos con la información disponible que permitieron describir un contexto del entorno comercial de la empresa.

Figura 8. Distribución de los puntos de venta según ubicación.



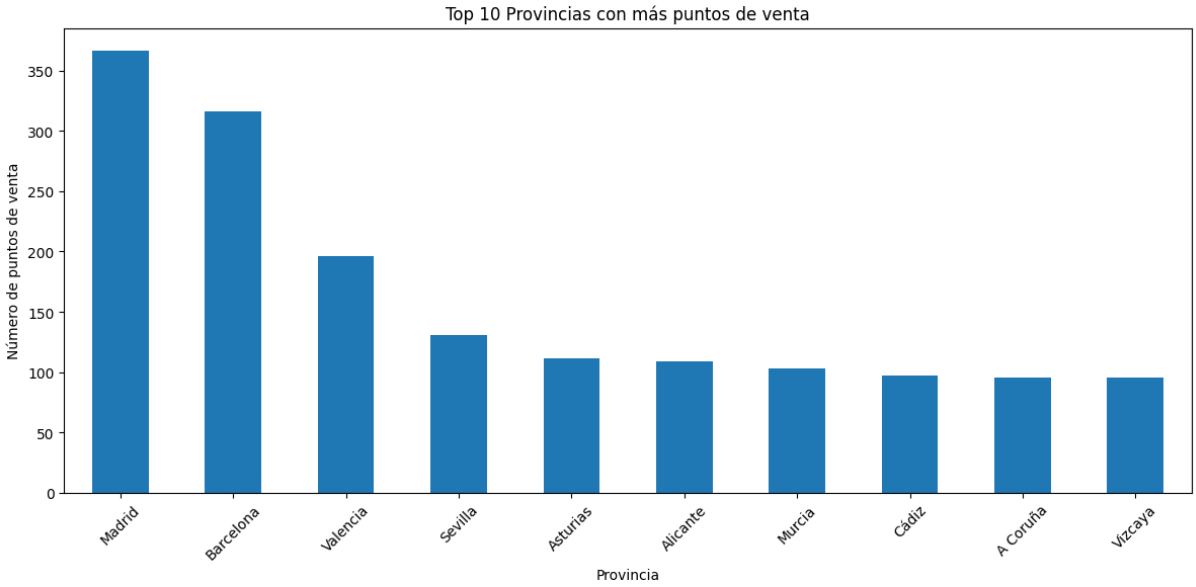
Fuente: elaboración propia

Figura 9. Distribución de los puntos de venta según el tamaño del local.



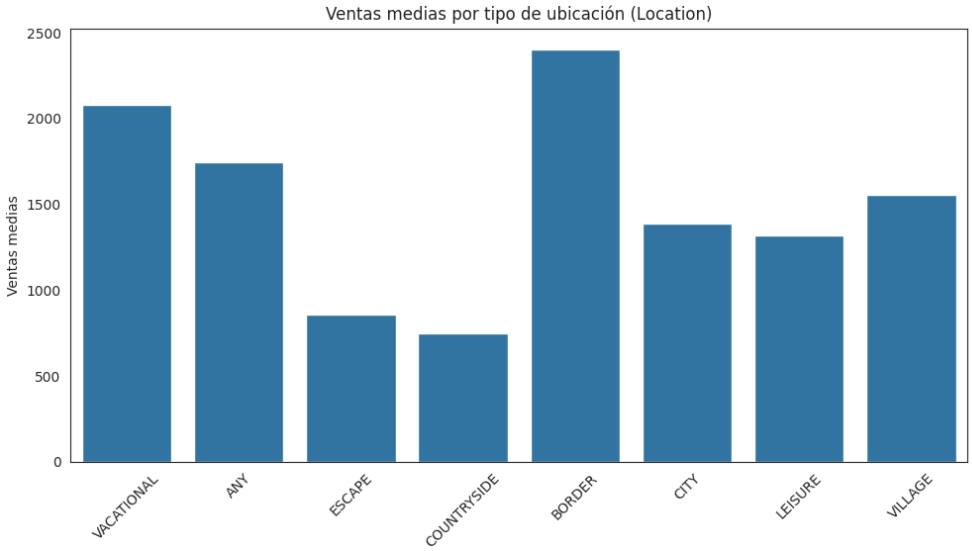
Fuente: elaboración propia

Figura 10. Distribución territorial de las provincias con mayor número de puntos de venta



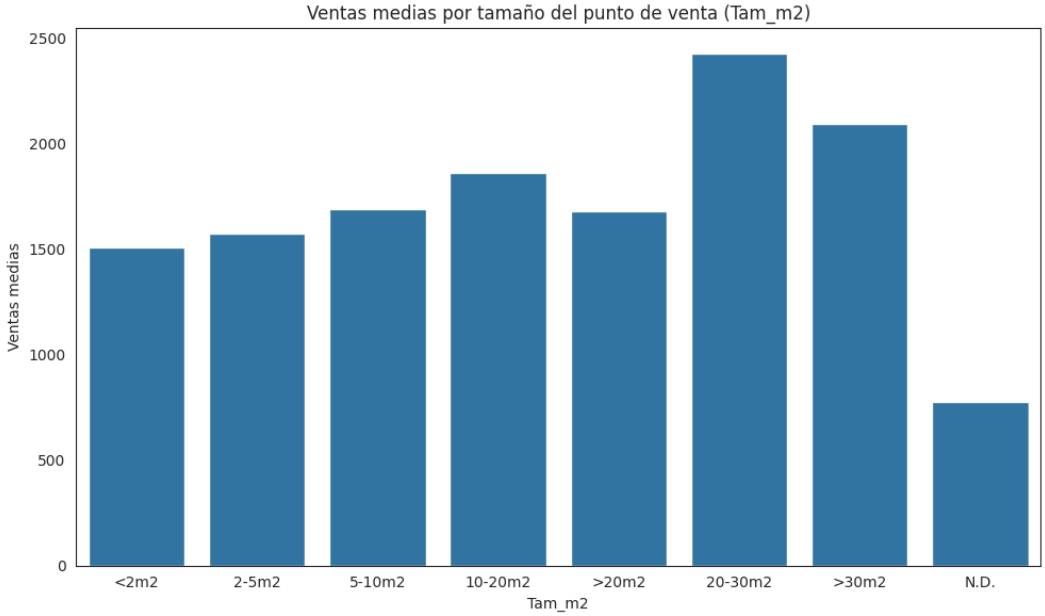
Fuente: elaboración propia

Figura 11. Distribución de las ventas medias en función del tipo de ubicación



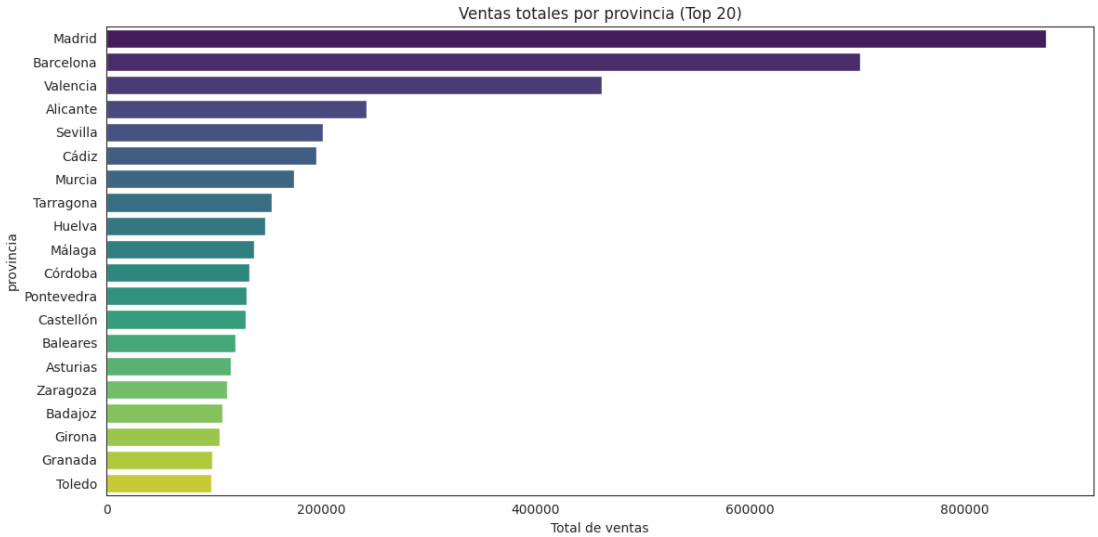
Fuente: elaboración propia

Figura 12. Distribución de las ventas medias en función del tamaño del establecimiento.



Fuente: elaboración propia

Figura 13. Distribución de las ventas totales por provincia.



Fuente: elaboración propia

- En el análisis territorial, se aplicaron los siguientes filtros específicos:
  - Se eliminaron fechas no válidas, utilizando (is.na (Sales_DAY)).
  - Se prescindió de los puntos de venta que no tenían una provincia asignada, usando la expresión (is.na(provincia)).

- Se excluyeron las provincias cuya serie temporal incluía valores infinitos o constantes, para asegurar la consistencia estadística. Para ello se usó la expresión (is.finite()).

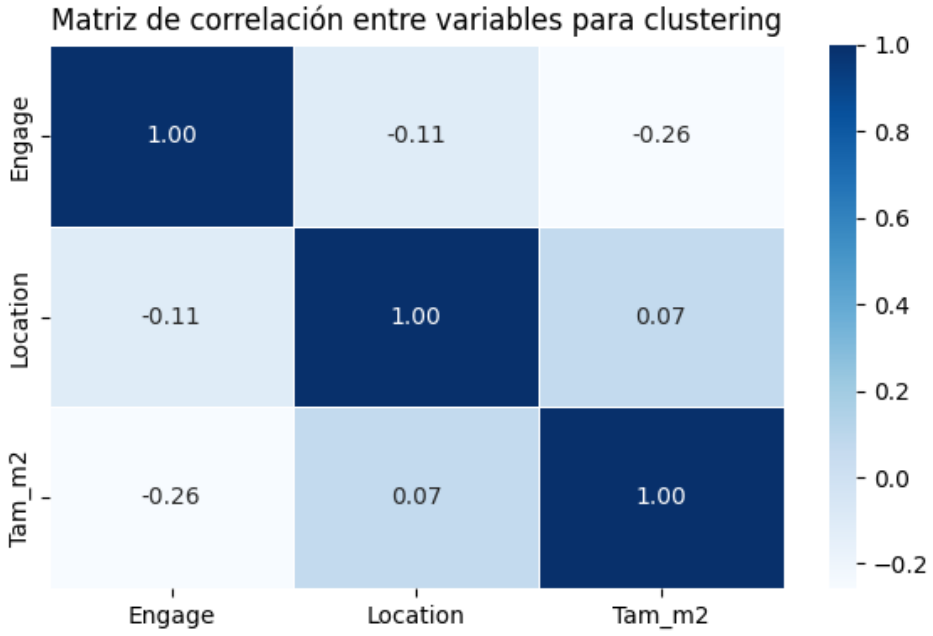
Este trabajo de depuración fue fundamental para evitar distorsiones en el análisis temporal, especialmente al aplicar transformaciones como logaritmos o normalizaciones.

3.2.1.2. Conversión y codificación de variables

Para adaptar las variables a los modelos estadísticos, se requirieron ciertas modificaciones iniciales. Como caso ilustrativo:

- Las variables categóricas Localización y Tam_m2 fueron codificadas como número, lo que permitió su aplicación en los algoritmos de clustering como K-means.

Figura 14. Mapa de calor de la matriz de correlación entre Location, Tam_m2 y Engage.



Fuente: elaboración propia

- Dentro del estudio territorial, la variable Sales_Uds (unidades vendidas) se transformó a formato numérico, utilizando la expresión as.numeric, permitiendo a partir de entonces realizar sumas y agregaciones sin errores.

- Las fechas (Sales_DAY) se convirtieron al formato Date, utilizando la función específica ymd() en R Studio, lo que sirvió para simplificar el estudio de la evolución de las ventas en cada provincia a lo largo del tiempo.

Estas adaptaciones técnicas fueron necesarias para integrar los datos en los modelos, asegurando un funcionamiento correcto de los algoritmos utilizados.

3.2.1.3. Estructuración de datos y agregación

Una vez corregidos y codificados los datos, se reorganizó su estructura para facilitar los diferentes enfoques de análisis:

- En los modelos de clusterización, se consolidó un conjunto de datos organizado por punto de venta, que incluía tanto sus características estructurales como el volumen agregado de ventas, de tal forma que se obtuvo una tabla con una fila por tienda.
- En el estudio territorial de convergencia, se diseñó una tabla en formato "ancho" ("wide format"), con una fila por provincia y una columna por cada día. Las ventas se agregaron por fecha y provincia, y los valores vacíos se completaron con ceros (values_fill = 0).

Esta disposición hizo posible que se pudieran comparar fácilmente los patrones de evolución entre provincias, así como aplicar transformaciones temporales homogéneas.

3.2.1.4. Transformaciones estadísticas

Para consolidar la varianza y estabilizar el comportamiento de las series, se implementaron diversas transformaciones adicionales:

- Para el estudio del territorio, se empleó el logaritmo log (x + 1) en las ventas por provincia, buscando reducir el peso de valores extremos.
- También se eliminaron columnas con fechas que tenían varianza nula o no finita, ya que no añadían valor al análisis de la evolución.

Esta estrategia mejoró la solidez del modelo de convergencia, evitando que días atípicos o con poca información influyeran en los patrones globales.

3.2.1.5. Normalización y escalado

Para analizar la convergencia de ventas entre provincias, fue necesario normalizar cada serie temporal dividiendo los valores diarios por la media de cada provincia: esta técnica es común en estudios de convergencia, ya que permite focalizar el análisis en el comportamiento relativo, más que en el tamaño de cada territorio. Esto permitió analizar la evolución relativa, independientemente del volumen absoluto de ventas.

Gracias a todas estas tareas de preparación, depuración y transformación, se consolidaron unas bases de datos analíticas bien organizadas, íntegras y con lógica estadística. Esto hizo posible aplicar con confianza los modelos, facilitando la extracción de insights valiosos sobre el comportamiento comercial de la red de puntos de venta.

3.2.2. Análisis histórico y/o limpieza de datos: Temporalidad de las ventas

Con el fin de estudiar la temporalidad de las ventas se realiza un tratamiento y limpieza de los datos de la siguiente forma:

- Comprobación de la no existencia de valores ausentes.

-Cumplimentación de municipios faltantes (25 códigos postales de municipios y 7 provincias) de forma manual, mediante la búsqueda de estos.

- Búsqueda de duplicados: En el data set “Product” se comprueba que existe un producto duplicado el llamado Natu122 y se elimina.

- La variable “Sales_ DAY” se convierte a formato fecha.

- Fusión de los data set mediante las claves “Product_Code” y “Affiliated_Code”.

-Creación de variables nuevas como: ventas normalizadas (Sales_Uds_1), agrupación temporal por semana y por mes, búsqueda de días festivos y laborables, identificación de CCAA de cada provincia.

-Se identifican los distintos valores únicos para el conocimiento de los datos:

- La variable “Management_Cluster” tiene los valores únicos: 0, 1 ,2,3,4
- Existen 48 provincias todas las de la península ibérica y baleares
- Hay 54 productos diferentes

- 8 categorías en “Location”: CITY, ANY, VILLAGE, BORDER, COUNTRYSIDE, VACATIONAL, ESCAPE y LEISURE.
- 23 tamaños de productos
- 3 formatos de producto: "ASL" "ATA" "ETO"

3.2.2.1. Análisis exploratorio y visualizaciones

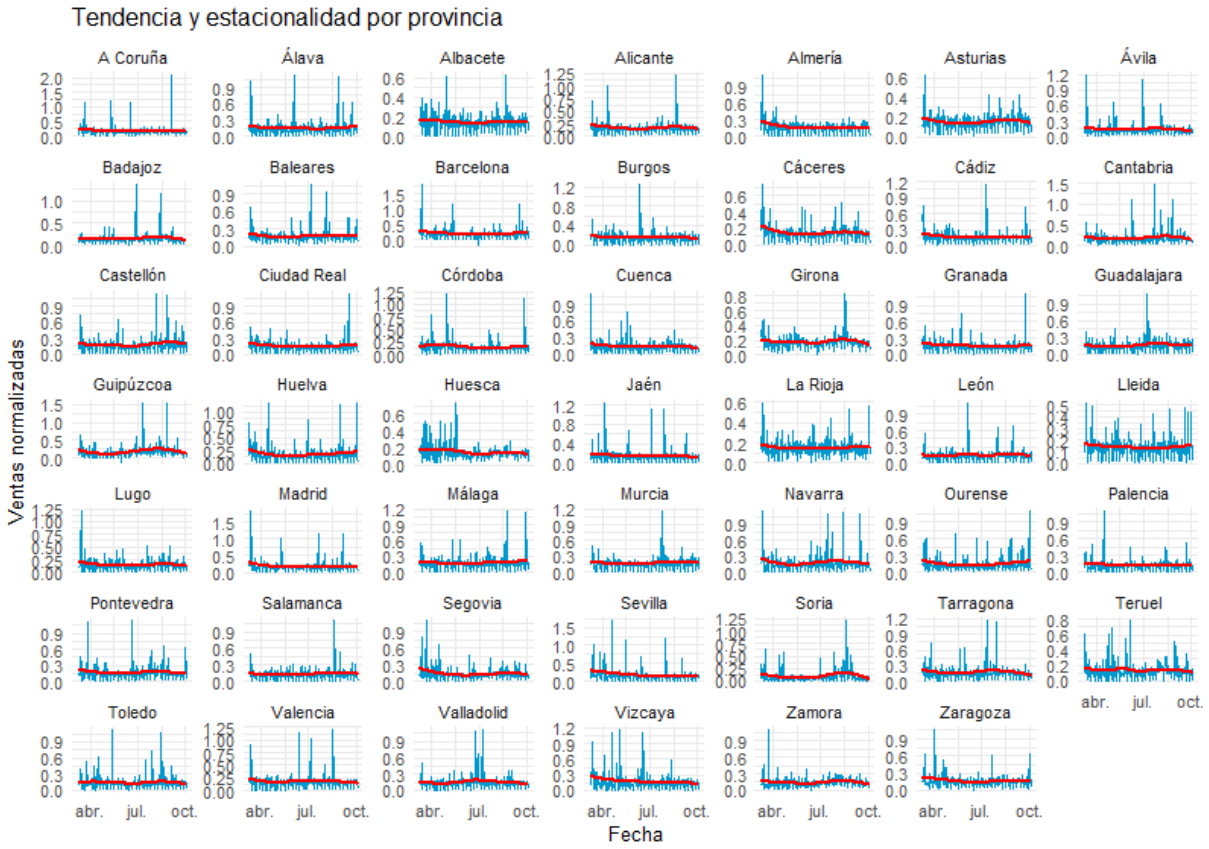
Con el principal objetivo de buscar conocimiento y con ello insights relevantes, se ha realizado el análisis exploratorio de los datos en busca de patrones estacionales e información sobre el patrón temporal de las ventas de Altadis en el periodo de marzo a octubre de 2015. Centrándose en la segregación del comportamiento por provincia, día de la semana y tipo de día del calendario (festivo, laboral y fin de semana).

A. ***Comportamiento de la serie temporal de las ventas en cada una de las provincias***

Después de normalizar las ventas y mediante la librería ggplot2, se realiza un gráfico donde se observa la evolución diaria de las ventas normalizadas por provincias. Para ello se tuvo crear un data frame llamado “ventas_diarias_completas” en el que se tiene en cuenta las fechas faltantes para cada provincia y se añaden, así como se cubren esos días con el valor cero. Con esto y la normalización de las ventas (creación de la variable Sales_DAY_1 ventas en rango 0-1, se consigue evitar así saltos de líneas en los gráficos y posibles sesgos.

El gráfico final obtenido g0 permite visualizar los patrones de cada provincia, asimismo se elaboró gráfico g1 en el que se añade “geom_smooth” para tratar de visualizar con mayor facilidad la tendencia de los datos.

Figura 15. Tendencia y estacionalidad por provincia (g1).



Fuente: elaboración propia

Tras la inspección visual de los gráficos se obtienen los siguientes hallazgos de interés:

- Posible estacionalidad en las provincias de Costa: se visualiza un aumento de las ventas durante el verano en esas áreas concretamente en las provincias de Alicante, Málaga y Valencia, Posiblemente por la llegada de turismo y aumento en la población flotante.
- Se aprecia menor inestabilidad en provincias del interior como Lugo, León o Cuenca, donde la demanda parece menos sujeta a influencias externas.
- En general se ven caídas pronunciadas que coinciden con días de baja actividad comercial como pueden ser domingos o festivos nacionales autonómicos, esto se investiga a continuación con mayor detalle.

*B. **Tendencia por días de la semana***

Para llevar a cabo la búsqueda de patrones en las ventas durante los días de la semana, se ha realizado un gráfico(g2) para todo el país distinguiendo las ventas agregadas para cada día y se ha averiguado que:

- Los lunes y los viernes son los días con mayor volumen de venta.
- Los domingos es el día donde las ventas caen considerablemente, esto era previsible, aunque sea verificado gracias a este análisis.
- Este patrón visto muestra la actividad del canal minorista, ya que los fines de semana es el periodo en el que disminuye la distribución y las compras.

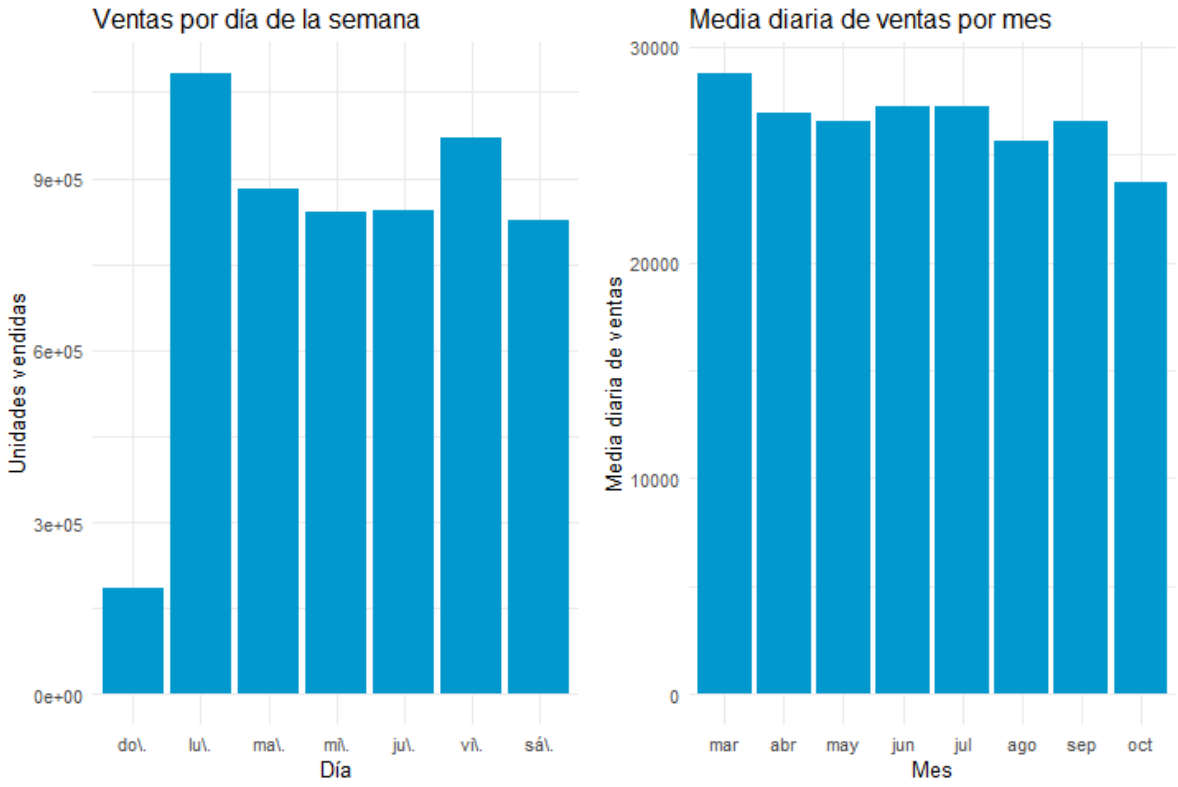
*C. **Tendencia mensual***

Con el fin de conocer cuál es la tendencia mensual, se ha elaborado un gráfico de barras con la media diaria de las ventas totales para ello se han tenido que normalizar las ventas de octubre y de marzo ya que no disponemos de los datos completos de los mencionados meses.

En el gráfico realizado (g2_normalizado) se encuentra un comportamiento estacional creciente desde abril hasta agosto, alcanzando su cumbre en verano. Esta propensión se entiende por:

- El aumento de población en los meses estivales debido al turismo.
- Mayor interacción social con lo que aumenta el consumo por una posible vinculación del producto a hábitos de consumo relajados.
- Desde septiembre se encuentra una caída, que coincide con el final de verano.
- La estacionalidad intra anual se podría haber observado de una mejor forma si se hubiese contado con los datos de un año entero.

Figura 16. Ventas por día de la semana y Media diaria por mes (g2 + g2_normalizado).



Fuente: elaboración propia

D. Influencia de los festivos nacionales y autonómicos

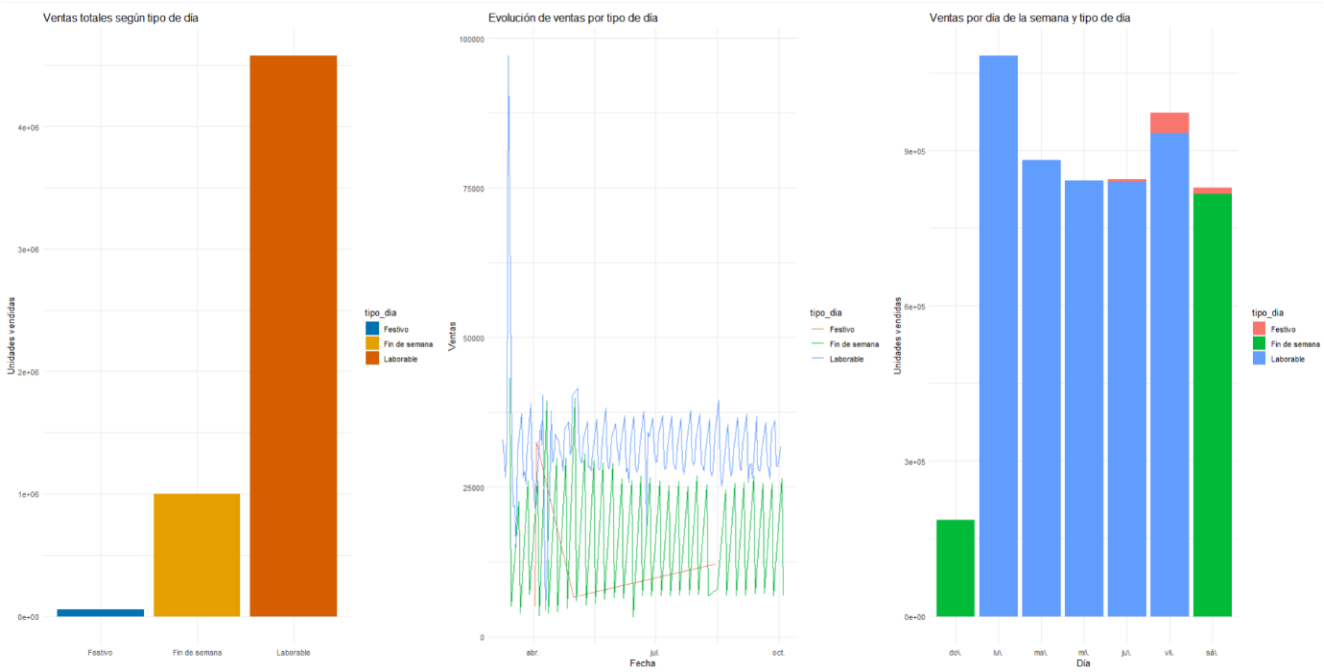
Considerando esencial en el análisis temporal la vinculación de los datos de ventas al calendario laboral y de días festivos. Se implementaron cuatro variables nuevas, mediante la creación de un vector de los festivos nacionales y una lista de los festivos de CCAA y un vector con la pertenencia de cada provincia a una comunidad autónoma:

- La variable “dia_semana” distingue el día de la semana.
- La variable “ccaa” pone la comunidad autónoma de pertenencia de la provincia.
- La variable “festivo_especifico” true o false si pertenece a festivo o no.
- La variable “tipo_dia” clasifica cada fecha en festivo, fin de semana o laborable.

A partir de este enriquecimiento temporal de los datos, se realizaron los siguientes gráficos:

- Ventas totales según tipo de día (g3): se confirma que los días laborables tienen el volumen más alto de ventas, siendo los fines de semana los segundos en volumen y por último los festivos debido al cierre de gran parte de los puntos de venta.
- Evolución de ventas por tipo de día(g4): caídas abruptas durante fines de semana y festivos y los días laborables cuentan con la gran parte de las ventas.
- Ventas por día de la semana y tipo de día (g5): se cruzan tipo de día y día de la semana, y se observan comportamientos atípicos, los días festivos jueves o lunes, con descensos más pronunciados en las ventas. Además, se ve que los sábados tienen una tendencia mixta, con volúmenes similares a los días intermedios de la semana.

Figura 17. Ventas totales según tipo de día (g3), Evolución de ventas por tipo de día(g4) y Ventas por día de la semana y tipo de día (g5).



Fuente: elaboración propia

*E. **Rankings de productos, provincias y puntos de ventas***

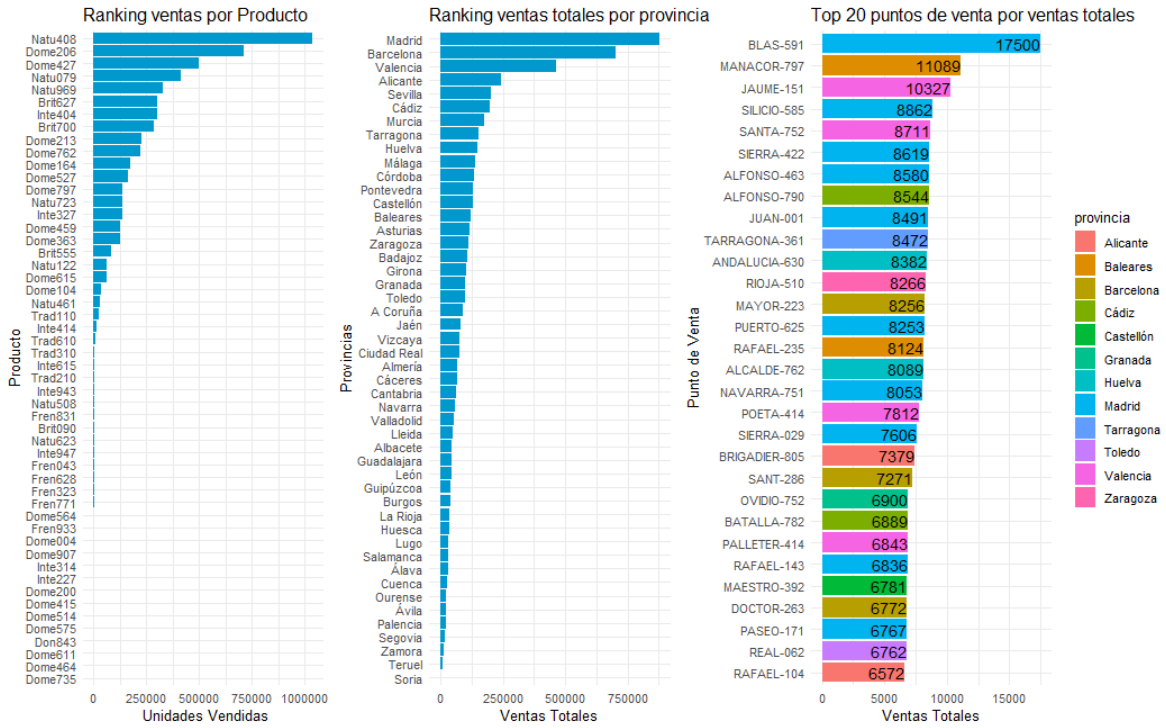
Para conocer cuantitativamente los lideres de ventas se ha realizado una jerarquía agregada para:

- <u>Producto (g6):</u> se localiza que Natu408 es el producto más vendido, pasando las 1.000.000 Uds. vendidas seguido de Dome206 y Dome 427 aunque este último no

alcanza las 50.000 Uds. vendidas. Se observa que las ventas se concentran en unos pocos productos y los hay sin ventas lo que sugiere que hay productos que se podrían valorar para su retirada del mercado, centrado los esfuerzos de fabricación y comercialización en los que funcionan.

- Provincia (g7): tal y como se puede observar en el gráfico, las provincias lideres en volumen de ventas son Madrid, Barcelona, Valencia y Alicante mostrando la densidad de población existente como un factor determinante para las ventas, así como su mayor volumen de puntos de venta.
- Punto de venta (g8): se han localizado los 30 establecimientos con las ventas más altas, la dispersión geográfica de estos es alta, con presencia de locales en varias provincias, el establecimiento líder está en Madrid.

Figura 18. Ranking ventas por producto (g6), Ranking Ventas Totales por provincia (g7) y Top 20 puntos venta por ventas totales (g8).



Fuente: elaboración propia

3.2.3. Limpieza de datos para el desarrollo del dashboard

Se ha realizado la comprobación de la no existencia de valores ausentes, duplicados y creación de variables como ha sido la variable CCAA para la comunidad autónoma de pertenencia de cada provincia, creando una columna condicional.

A continuación, se observa un extracto del código M empleado.

Figura 19. Código DAX.

```
= Table.AddColumn(#"Columnas con nombre cambiado1", "CCAA", each if [Provincia] = "A Coruña" then "Galicia" else
if [Provincia] = "Albacete" then "Castilla-La Mancha" else
if [Provincia] = "Alicante" then "Comunidad Valenciana" else
if [Provincia] = "Almería" then "Andalucía" else
if [Provincia] = "Asturias" then "Principado de Asturias" else
```

Fuente: elaboración propia

3.3. Modelado propuesto

Durante esta fase del proyecto, se aplicaron varias técnicas de modelado a los datos previamente estructurados y preparados.

Con el objetivo de segmentar los puntos de venta según su estructura física, rendimiento comercial y ubicación geográfica, se recurrió a técnicas de aprendizaje no supervisado, en concreto al algoritmo K-means. Por otro lado, se aplicaron métodos econométricos de análisis de convergencia entre provincias, que permitieron extraer tipologías de comportamiento que proporcionaron estrategias diferenciadas, adaptadas a las características y al entorno de cada punto de venta.

La aplicación de estos modelos ayudó a obtener una visión clara y estructurada del negocio, facilitando la comprensión de la realidad operativa de la actividad de la empresa.

En ambos casos de segmentación, se optó por aplicar el algoritmo K-means, una técnica de aprendizaje no supervisado ampliamente utilizada por su eficiencia y por su capacidad para generar agrupaciones que resultan fácilmente interpretables desde el punto de vista comercial. Fue especialmente útil dado que las variables con las que se trabajó eran numéricas o fácilmente codificables, facilitando la aplicación del modelo.

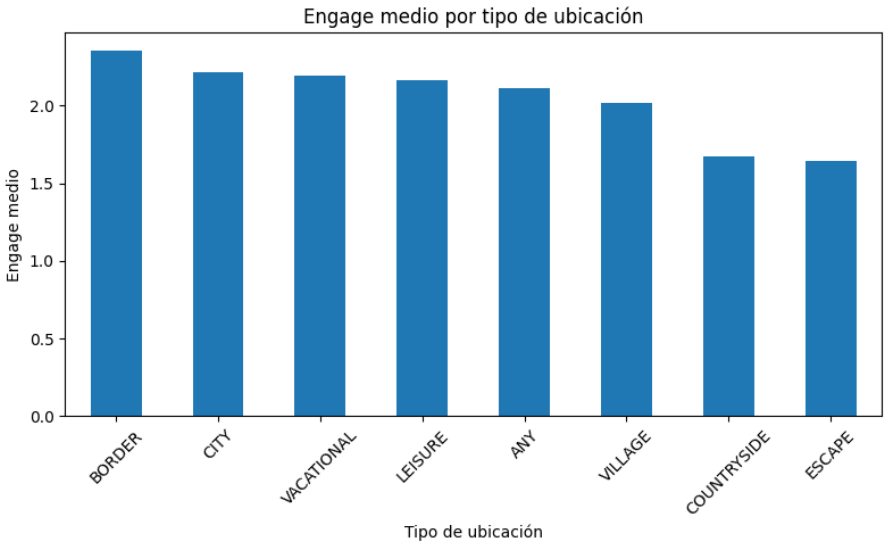
Además de lo anterior y con el fin de generar un modelo empírico que funcione con la variable de interés volumen de ventas a lo largo del tiempo y con los componentes de estacionalidad y ruido presentantes encontrados anteriormente, se decide emplear los modelos ARIMA (Autoregressive Integrated Moving Average) y ETS (Error-Trend-Seasonality). Modelos autorregresivos de machine learning bien establecidos en el análisis temporal y que funcionan bien cuando existe tendencia o estacionalidad y ruido.

A continuación, se detallan los modelos que se usaron y los objetivos específicos de cada uno.

3.3.1. Segmentación estructural de puntos de venta

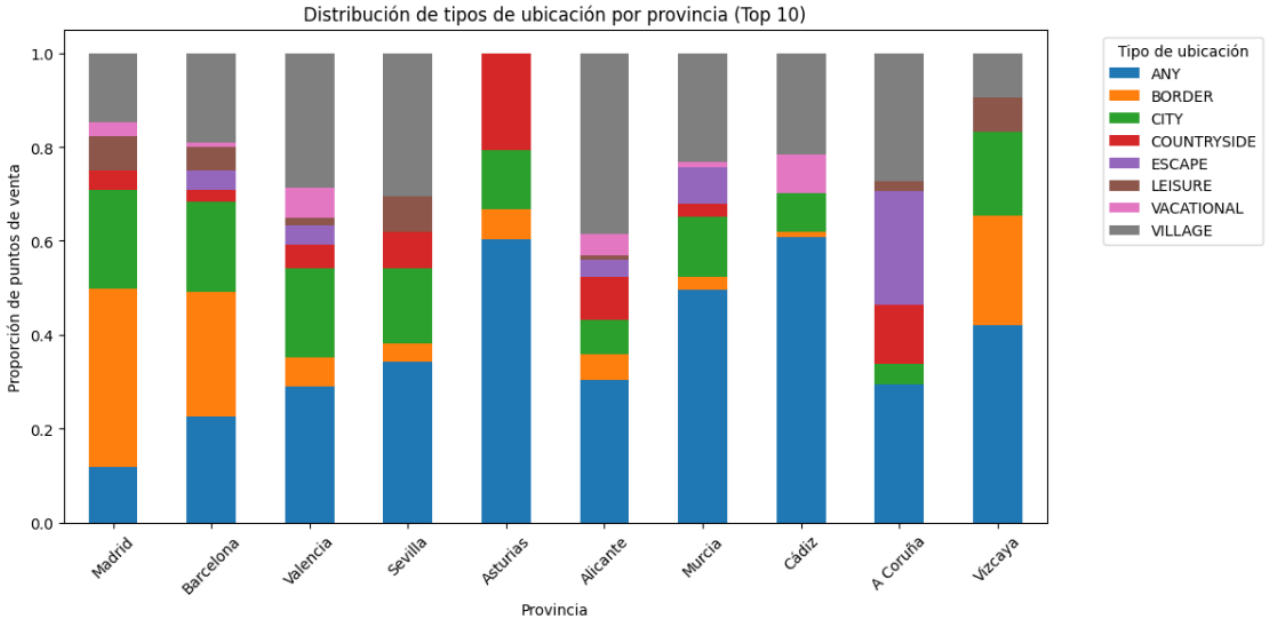
En esta primera aproximación se buscó identificar grupos de puntos de venta con características físicas y operativas similares. Inicialmente se visualizaron en forma de gráficos las relaciones entre diversas variables lo que permitió identificar, entre otras conclusiones, provincias con un alto número de ubicaciones inespecíficas, lo que podría indicar que la red comercial es dispersa o con escasa diferenciación geográfica. Esto permitió inferir que la distribución de los tipos de ubicación (Location) varía en cada provincia, por lo que estas diferencias se deberían tener en cuenta al aplicar el modelo de segmentación, ya que la ubicación afecta no sólo a la implicación comercial (Engage) sino también a la viabilidad de las acciones promocionales, logísticas o de fidelización de los puntos de venta.

Figura 20. Nivel de implicación comercial (Engage) medio en función de la ubicación del punto de venta (Location).



Fuente: elaboración propia

Figura 21. Proporción de los tipos de ubicación en las provincias con mayor número de puntos de venta.



Fuente: elaboración propia

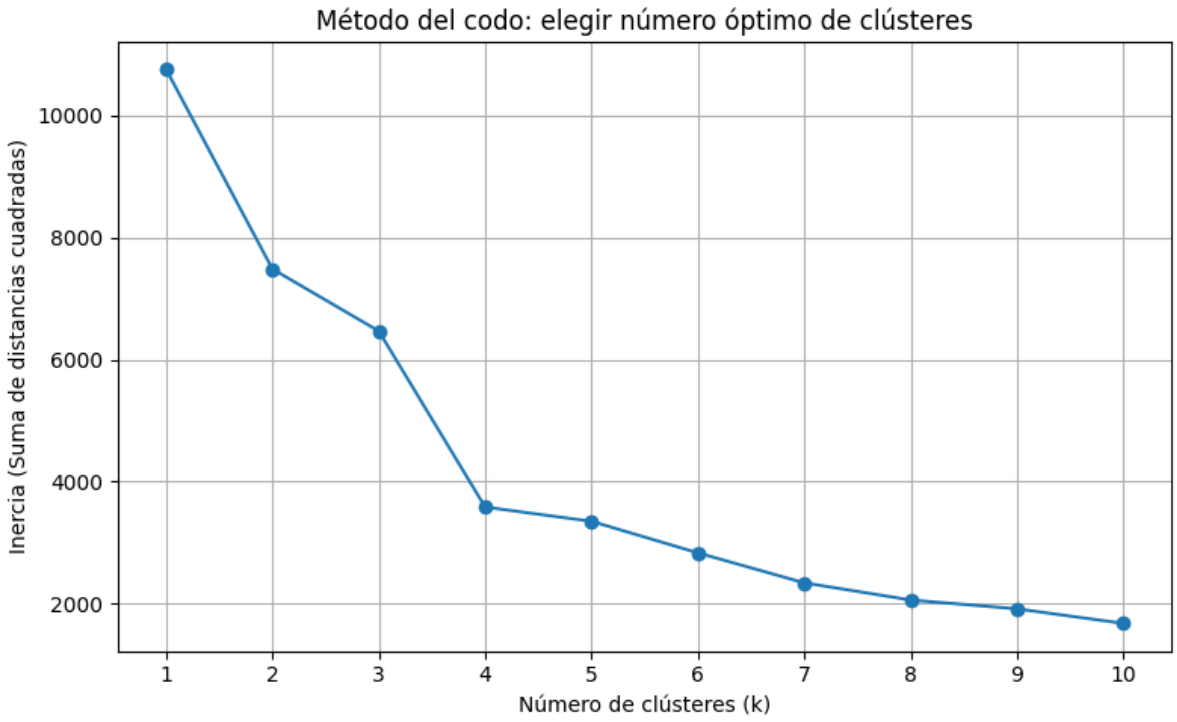
Posteriormente, se aplicó el algoritmo de segmentación K-means, con el objetivo de agrupar los puntos de venta según sus características estructurales. Se definió que estas características quedaban recogidas en las variables clave previamente codificadas Location_Num, Tam_m2_Num y Engage: estas variables recogen la ubicación - zonas fronterizas o de paso, urbanas, rurales, de segunda residencia, de ocio, turísticas, pueblos con estructura comercial y zonas genéricas o sin clasificar-, el tamaño del local y el nivel de implicación comercial. Se descartaron para la agrupación las variables código postal, municipio y provincia ya que, aunque aportaban información geográfica detallada, se estimó que podían llegar a introducir demasiado ruido en el análisis. Se descartó, de igual forma, la variable Management_Cluster ya que, al desconocer su contenido específico, podría estar identificando una segmentación previa que no debía influir en esta clusterización.

Se escalaron las variables con StandardScaler() para evitar sesgos que pudieran condicionar toda la segmentación.

Para determinar el número óptimo de grupos, se aplicaron métricas de validación interna como Elbow, Silhouette, Davies-Bouldin y Calinski-Harabasz. En base a estas métricas, se

consideró que el valor óptimo de grupos debía ser cuatro, lo que proporcionó una segmentación equilibrada.

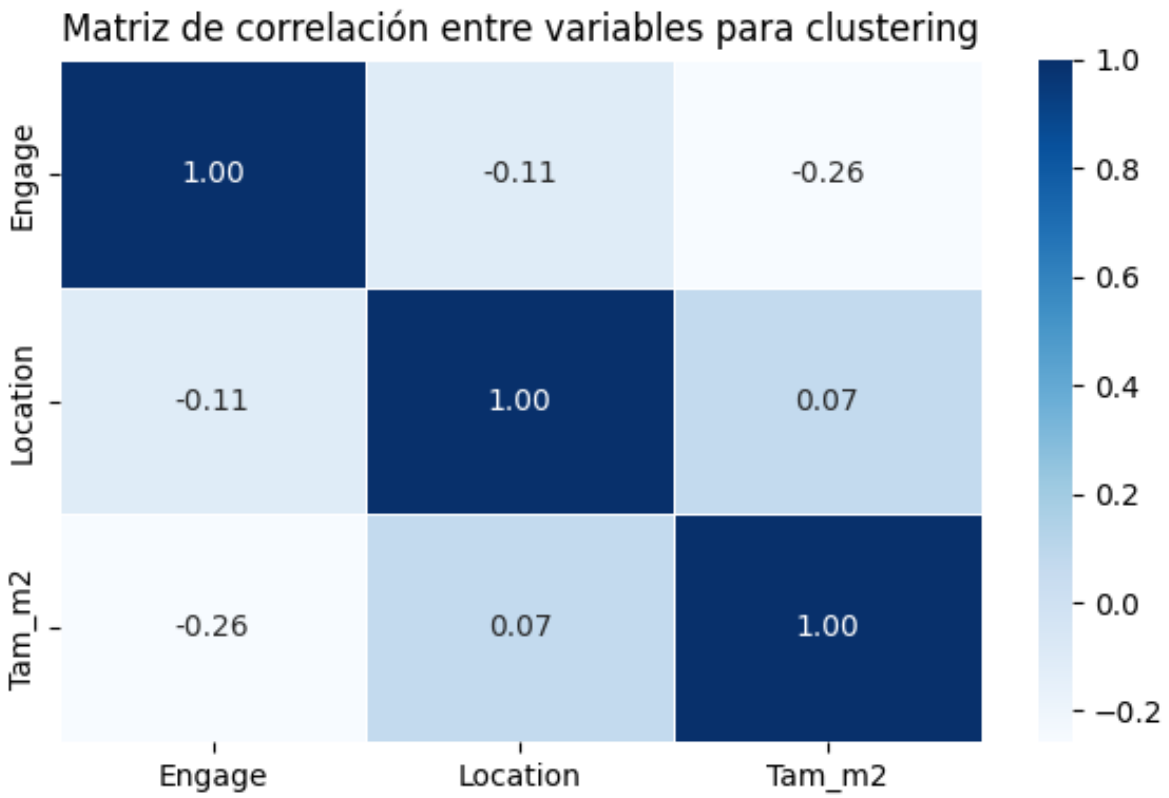
Figura 22. Representación gráfica del método Elbow



Fuente: elaboración propia

Se realizó una matriz de correlación de las variables implicadas en la segmentación que sugirió que las tres variables aportan información complementaria y no redundante, justificando su uso conjunto en el modelo de clusterización. Se apreció también que, aunque las variables no están fuertemente correlacionadas, sí se revelaban ligeras tensiones, como por ejemplo entre el tamaño de las tiendas y el nivel de compromiso comercial, lo que ayudaba a que la segmentación de los puntos de venta tuviera mayor riqueza interpretativa.

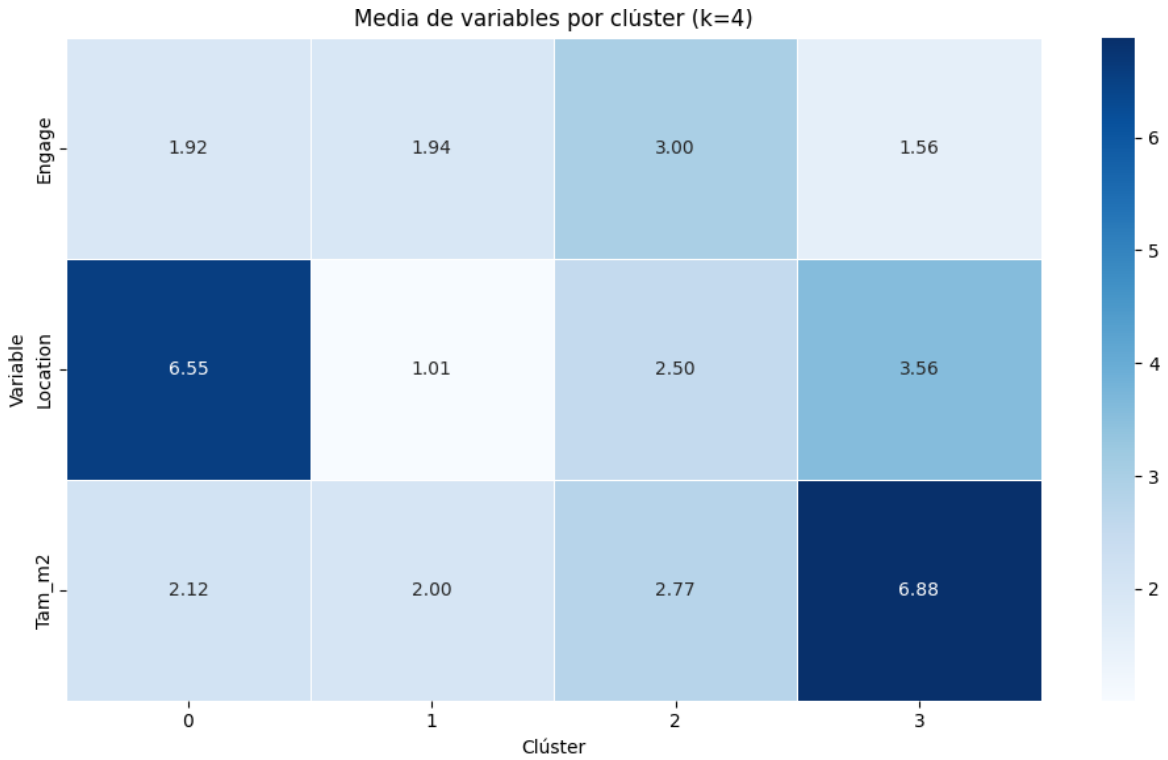
Figura 23. Representación gráfica de la matriz de correlación.



Fuente: elaboración propia

Se representaron las medias de cada variable por grupo, de tal manera que como, por ejemplo, si nos fijamos en la variable Engage, podemos observar que el segmento 2 tiene mayor Engage (nivel de implicación comercial) de media o cómo en el segmento 3, los locales tienen un tamaño superior medio que en el resto de los segmentos.

Figura 24. Mapa de calor de las medias de las variables por clúster, siendo k=4.

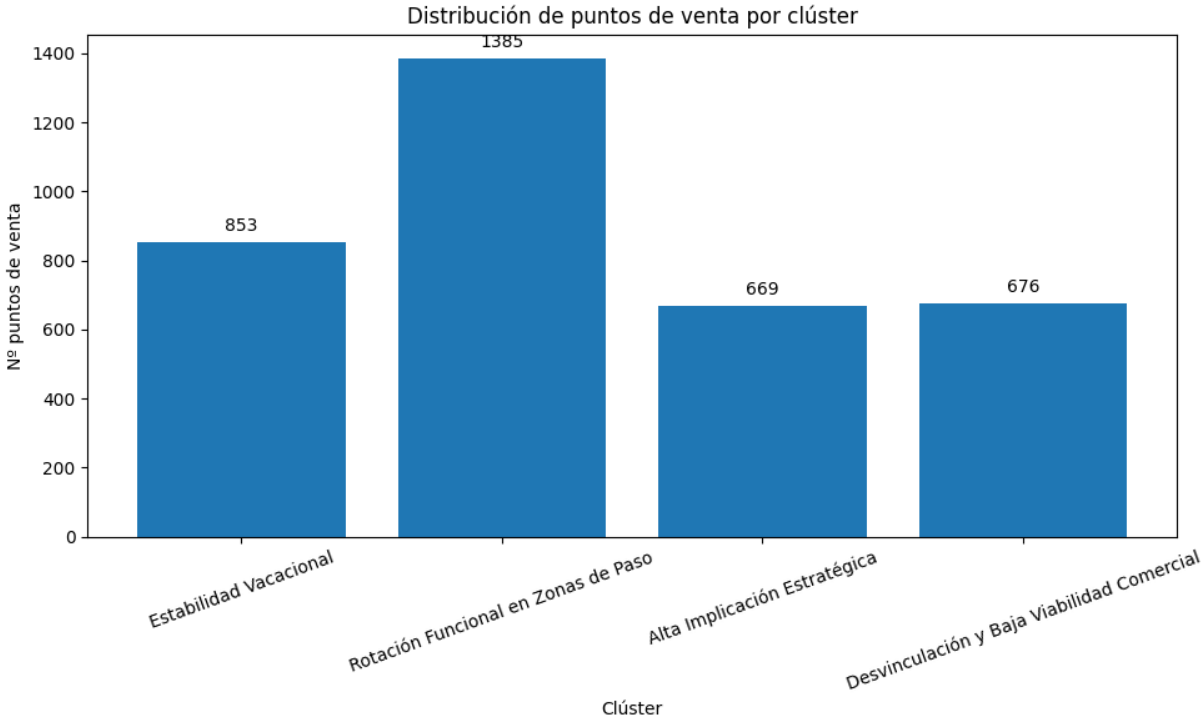


Fuente: elaboración propia

Con estas medidas, podemos obtener un perfil de tiendas según sus características estructurales, cuyas características principales se definieron en las siguientes etiquetas descriptivas:

- Grupo 0 “Estabilidad Vacacional”: en él se agrupan los establecimientos ubicados en zonas rurales o vacacionales, con implicación comercial media y tamaño pequeño.
- Grupo 1 “Rotación Funcional en Zonas de Paso”: las tiendas aquí agrupadas se ubican en zonas de paso, con implicación comercial media y tamaño pequeño.
- Grupo 2 “Alta Implicación Estratégica”: recoge las tiendas con mayor compromiso comercial, ubicadas en zonas urbanas o rurales, con locales de tamaño mediano.
- Grupo 3 “Desvinculación y Baja Viabilidad Comercial”: agrupa las tiendas que menor implicación comercial tienen de toda la muestra, se ubican en zonas rurales o turísticas. En este grupo se encuentran los locales más grandes o de los que se desconoce su tamaño (N.D. en el dataset).

Figura 25. Distribución de los puntos de venta agrupados con cluster con etiquetas descriptivas de perfil ya definidas en la segmentación estructural.

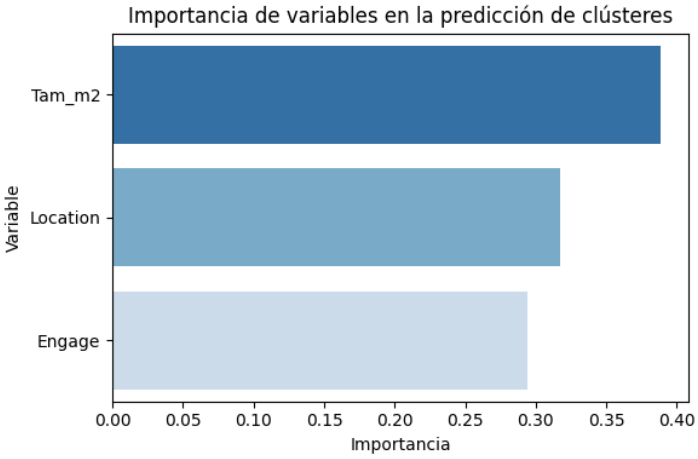


Fuente: elaboración propia

Con la representación gráfica de la distribución de los puntos de venta agrupados por segmento, podemos observar que la red comercial de Altadis no está equilibrada, ya que un gran número de tiendas se agrupa en el grupo 1 “Rotación Funcional en Zonas de Paso” que debe ser gestionado eficientemente.

Después de obtener los resultados de la segmentación, se decidió aplicar el método de aprendizaje no supervisado de los árboles de decisión, con el objetivo de tratar la asignación de cada observación a un segmento como una etiqueta y utilizar técnicas de clasificación para entender qué variables explican mejor la segmentación realizada.

Figura 26. Importancia de las variables en la predicción de clústeres.

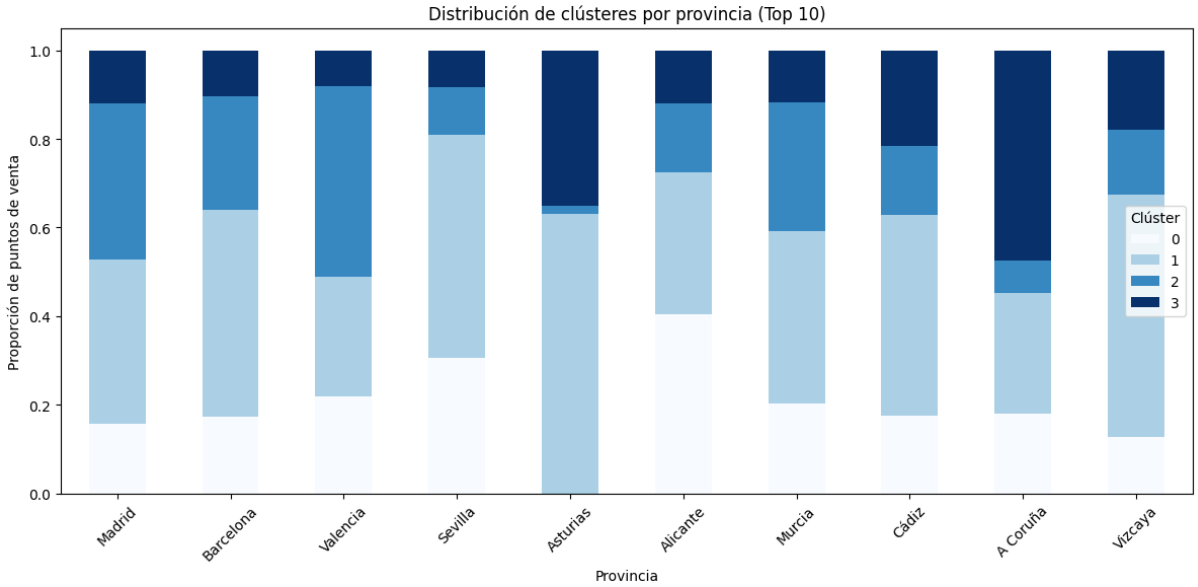


Fuente: elaboración propia

A través de la expresión "feature_importances_" aplicada al conjunto de entrenamiento del árbol de decisión, se observó que la variable que más influye para la agrupación es la relacionada con el tamaño de los locales, seguida de la ubicación y el nivel de implicación comercial. Aun así, se observa que las tres variables están bastante equilibradas, por lo que podemos confirmar que la segmentación de los puntos de venta responde a una combinación armónica de características estructurales, tales como el tamaño del local y la ubicación, como de características comerciales.

A continuación, se realizó un análisis territorial para comprender la distribución geográfica de los grupos detectados, para lo cual se incorporó al estudio la variable provincia.

Figura 27. Distribución de los clústeres en las provincias con mayor número de puntos de venta.

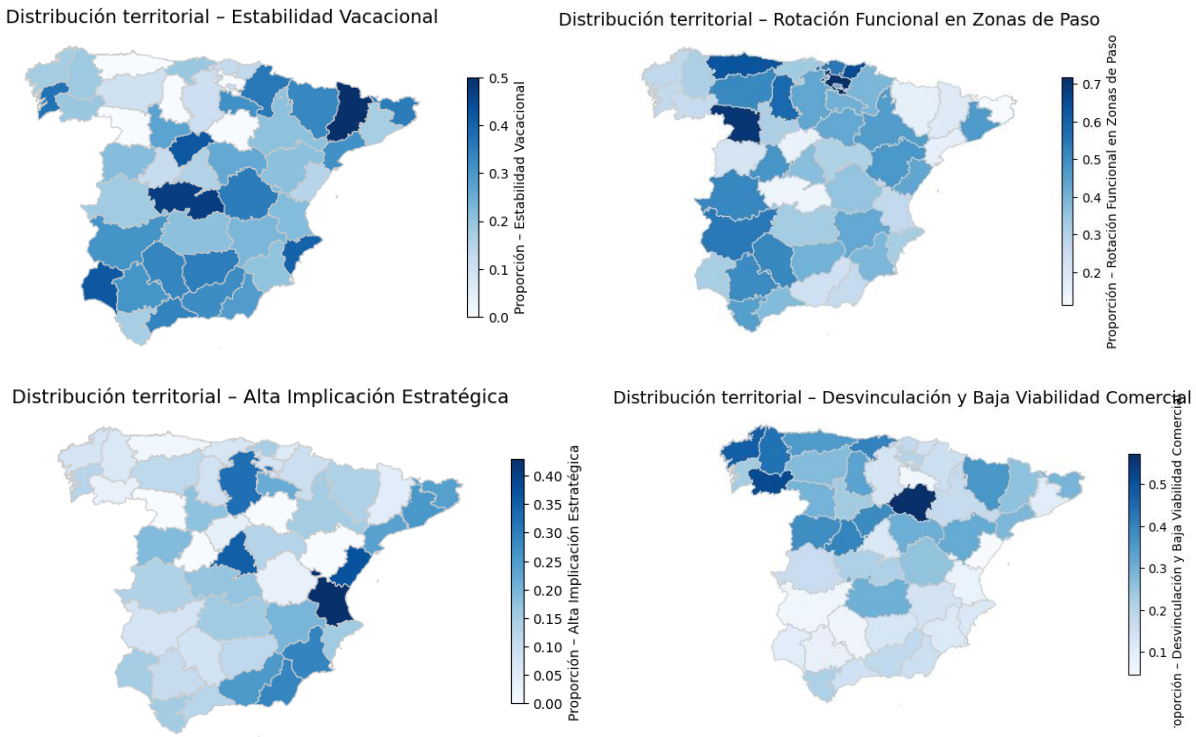


Fuente: elaboración propia

De esta forma, se puede ver cómo hay provincias en las que hay un fuerte predominio de algunos segmentos, como en el caso de Asturias o A Coruña, por lo que se pudo llegar a la conclusión de que cada provincia tiene una composición distinta de perfiles comerciales, por lo que se debería adaptar la estrategia comercial a cada zona y que, gracias a la segmentación, se pueden priorizar recursos comerciales y facilitar la toma de decisiones.

Para facilitar la toma de decisiones y completar el análisis territorial se generaron gráficos con la distribución de los segmentos por provincia, de tal forma que se puede observar qué grupos predominan en cada provincia.

Figura 28. Distribución territorial de los segmentos.

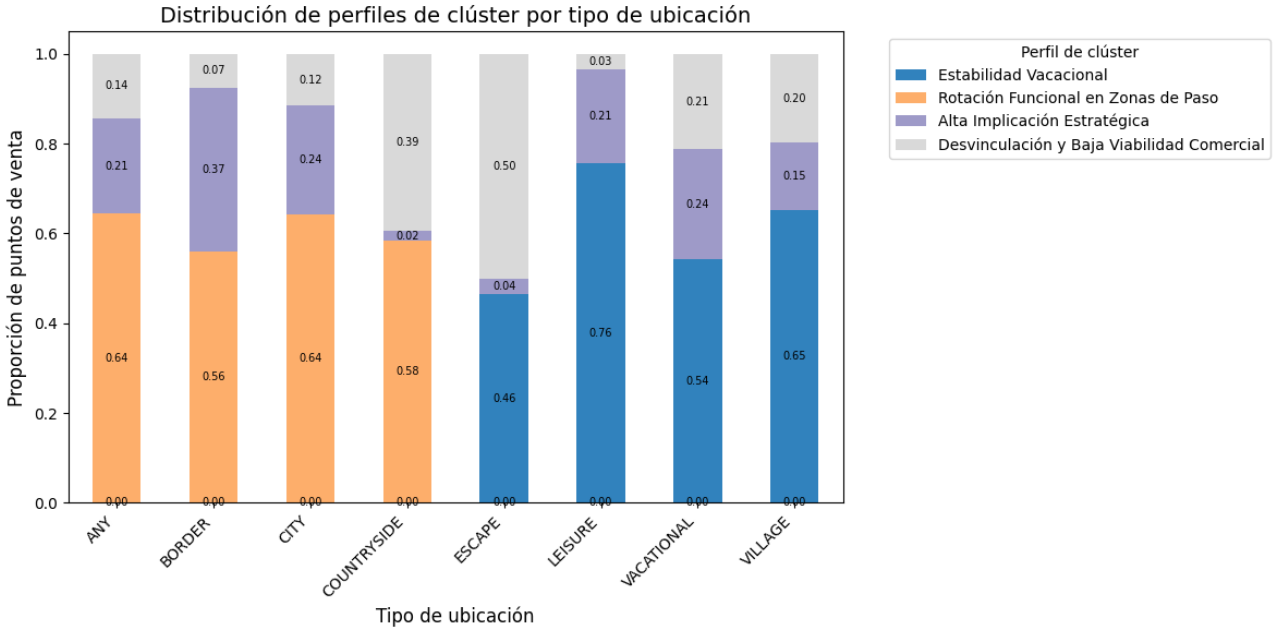


Fuente: elaboración propia

Esta visualización permitió establecer una composición comercial del territorio, lo que proporcionó una poderosa herramienta para la toma de decisiones en función de cada región, por lo que, a partir de esta visualización, la empresa puede:

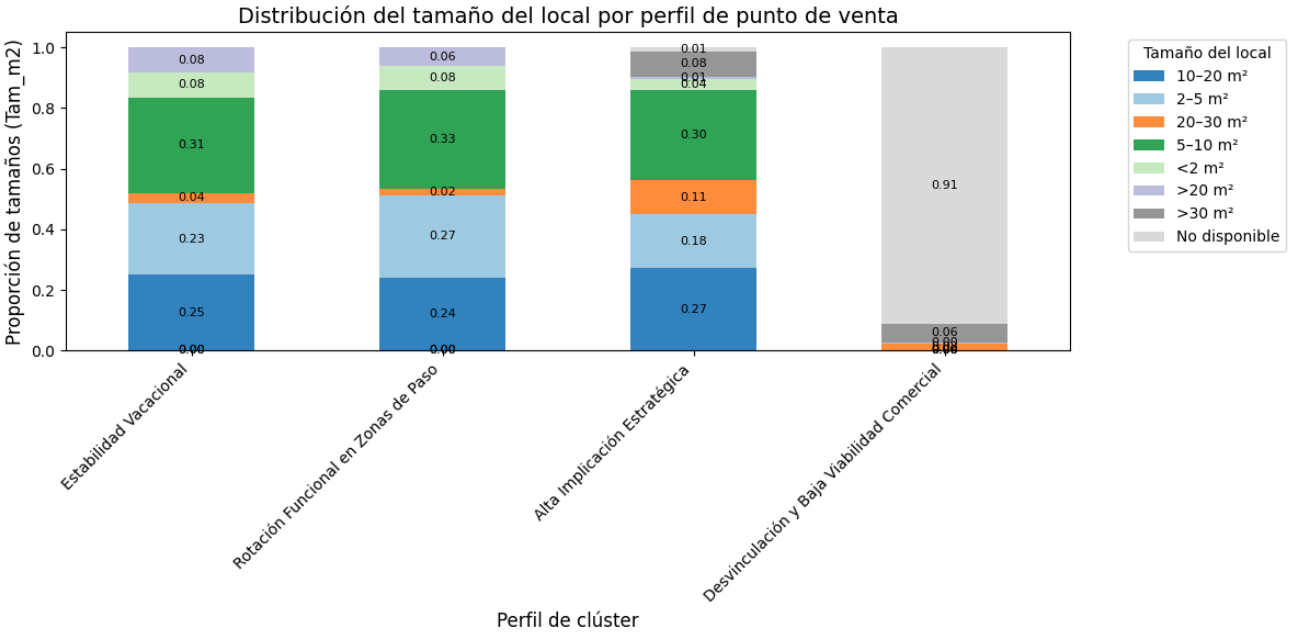
- focalizar la inversión donde pueda tener retorno, como en las provincias donde predomina el grupo 2,
- optimizar la eficiencia comercial en las provincias donde predomina el grupo 1,
- mantener la presencia comercial con bajo esfuerzo presupuestario en las provincias, con mayor proporción de tiendas agrupadas en el grupo 0, o
- auditar y desvincularse comercialmente de las zonas ineficientes, es decir, las zonas donde predomina el grupo 3.

Figura 29. Distribución de los grupos por tipo de ubicación



. Fuente: elaboración propia

Figura 30. Distribución de los grupos por tamaño del local.



Fuente: elaboración propia

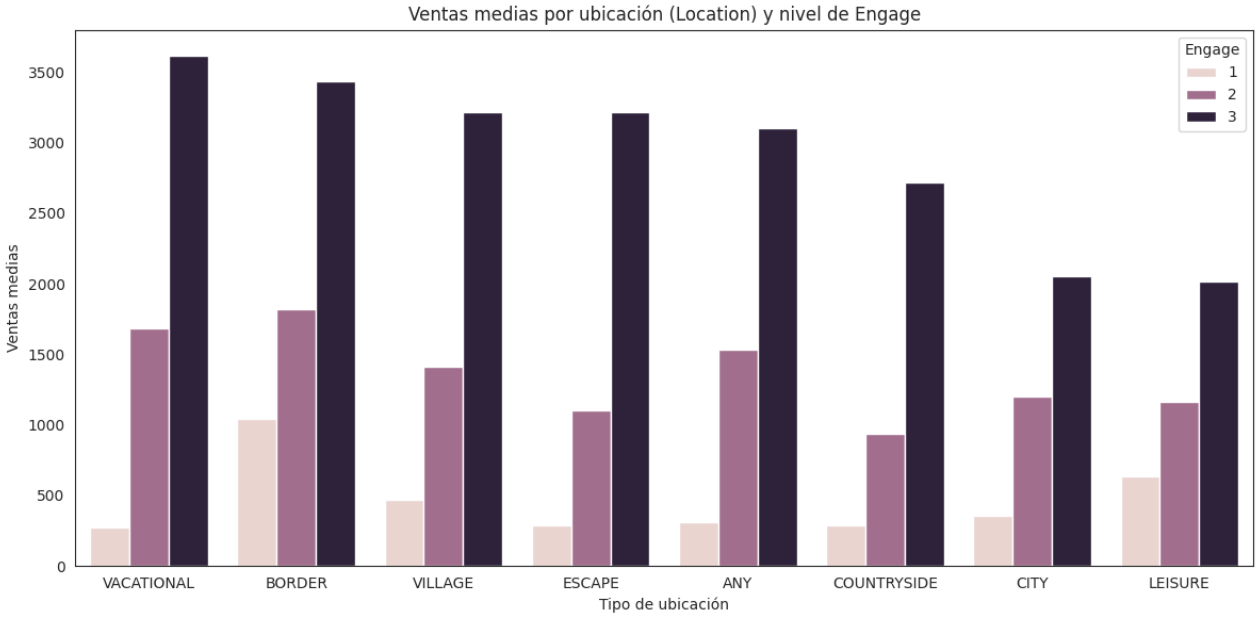
Se realizaron otra serie de visualizaciones para observar la distribución de los segmentos en función del tipo de ubicación y del tamaño del local, lo que proporcionó la información para observar que, por ejemplo, hay una alta proporción de tiendas ubicadas en zonas “ESCAPE”

(zonas de turismo de fin de semana) que se encuentran dentro del segmento "Desvinculación y baja viabilidad comercial" o que en ese mismo segmento hay una gran mayoría de locales de los que no se dispone de información relativa al tamaño, lo que podría indicar que no hay una relación comercial fluida.

3.3.2. Segmentación por rendimiento comercial

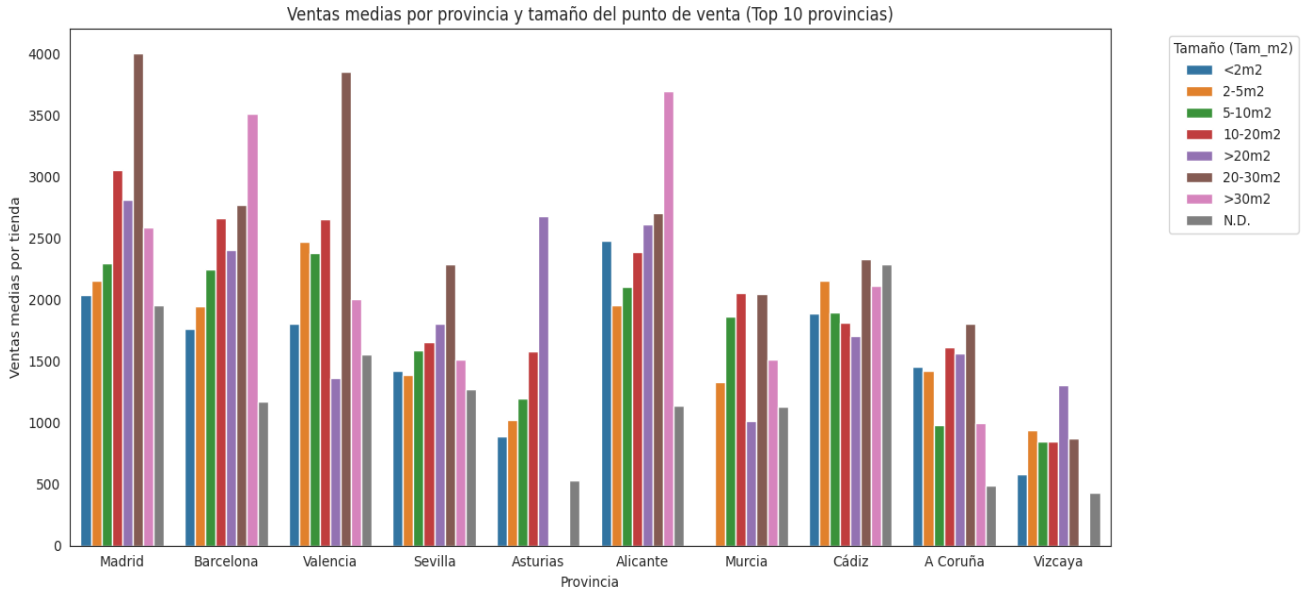
Con el objetivo de contextualizar la información comercial, se generaron distintas gráficas para comparar las variables estructurales anteriormente estudiadas (Location, Engage y Tam_m2) con los rendimientos comerciales.

Figura 31. Distribución de las ventas medias por ubicación y nivel de implicación comercial.



Fuente: elaboración propia

Figura 32. Distribución de las ventas medias por provincia y tamaño del local.



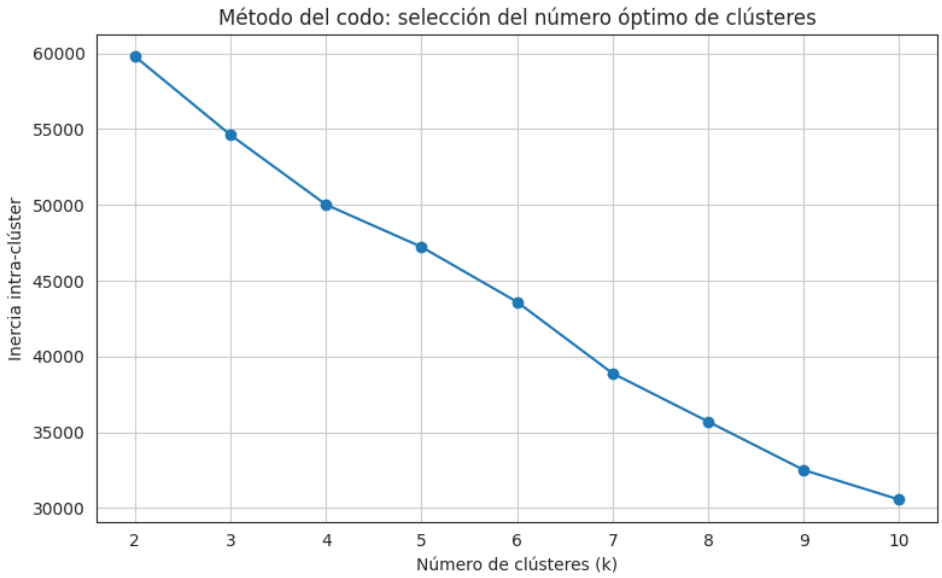
Fuente: elaboración propia

Gracias a estas visualizaciones, se pudieron confirmar las ideas ya inferidas previamente, ya que se observó que se dan mayores ventas medias en locales ubicados en zonas turísticas o fronterizas, en locales de tamaño grande y con el nivel de implicación comercial máximo.

A continuación, se implementó de nuevo un modelo K-means pero que ahora integraba el volumen global de ventas por tienda (Total_Ventas) a los factores estructurales ya estudiados. Con este análisis, se esperaba originar una agrupación más estratégica, al integrar el desempeño comercial.

Se emplearon de nuevo las métricas de validación interna Elbow, Silhouette, Davies-Bouldin y Calinski-Harabasz con varios valores posibles de k y de nuevo estas métricas revelaron que el valor óptimo de grupos debía ser cuatro, para evitar el sobreajuste y poder proporcionar una agrupación equilibrada.

Figura 33. Representación gráfica del método Elbow.

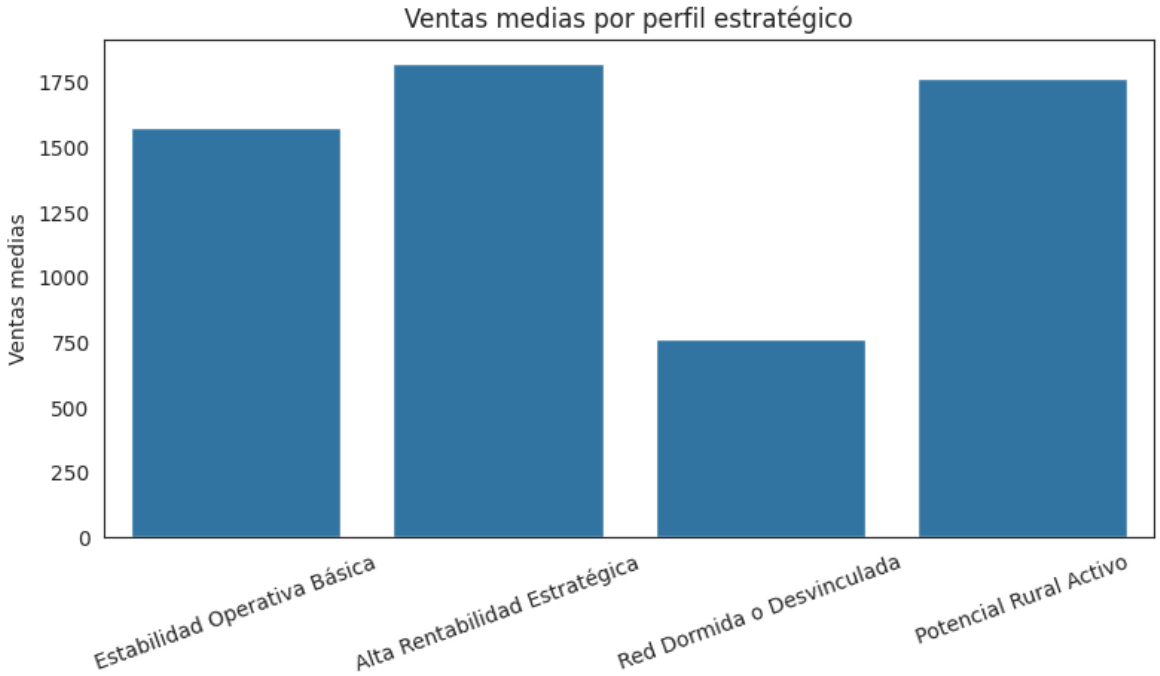


Fuente: elaboración propia

A partir del resultado de la agrupación se pudieron extraer nuevas etiquetas descriptivas para facilitar la interpretación de los segmentos, cuyo resumen es el siguiente:

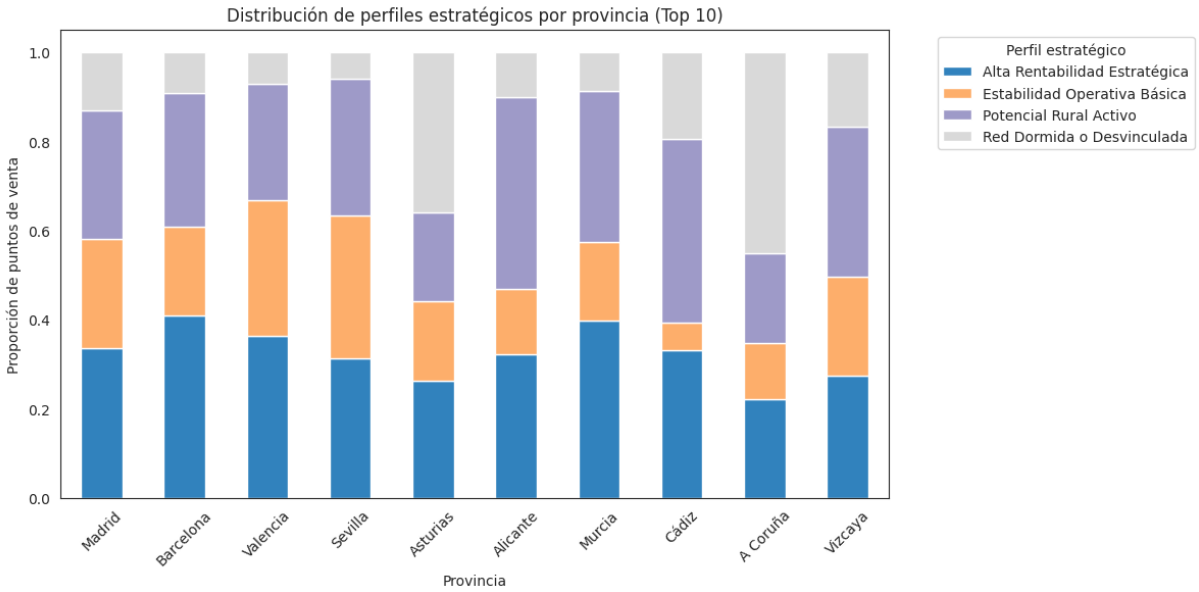
- Grupo 0 "Estabilidad operativa básica": incluye tiendas ubicadas en zonas calificadas como "ANY" o zonas rurales, en locales pequeños y con un nivel de implicación comercial medio que no destacan en ventas, pero mantienen un desempeño aceptable.
- Grupo 1 "Alta rentabilidad estratégica": en este clúster destacan las ventas medias, ya que son las más altas de todas las agrupaciones, con un nivel de implicación comercial máximo, con locales principalmente grandes y ubicados en distintas zonas.
- Grupo 2 "Red dormida o desvinculada": agrupa las tiendas con menor nivel de implicación comercial y menores ventas medias. Se ubican principalmente en zonas rurales y en su práctica totalidad no hay tamaño descrito para los locales.
- Grupo 3 "Potencial rural activo": incluye tiendas con un nivel comercial medio-alto y un buen nivel de compromiso comercial. Se ubican principalmente en pueblos medianos y alguna ciudad, en locales de tamaño medio.

Figura 34. Distribución de las ventas medias por perfil estratégico definido en la segmentación por rendimiento comercial.



Fuente: elaboración propia

Figura 35. Distribución de los perfiles estratégicos en las provincias con mayor número de puntos de venta.

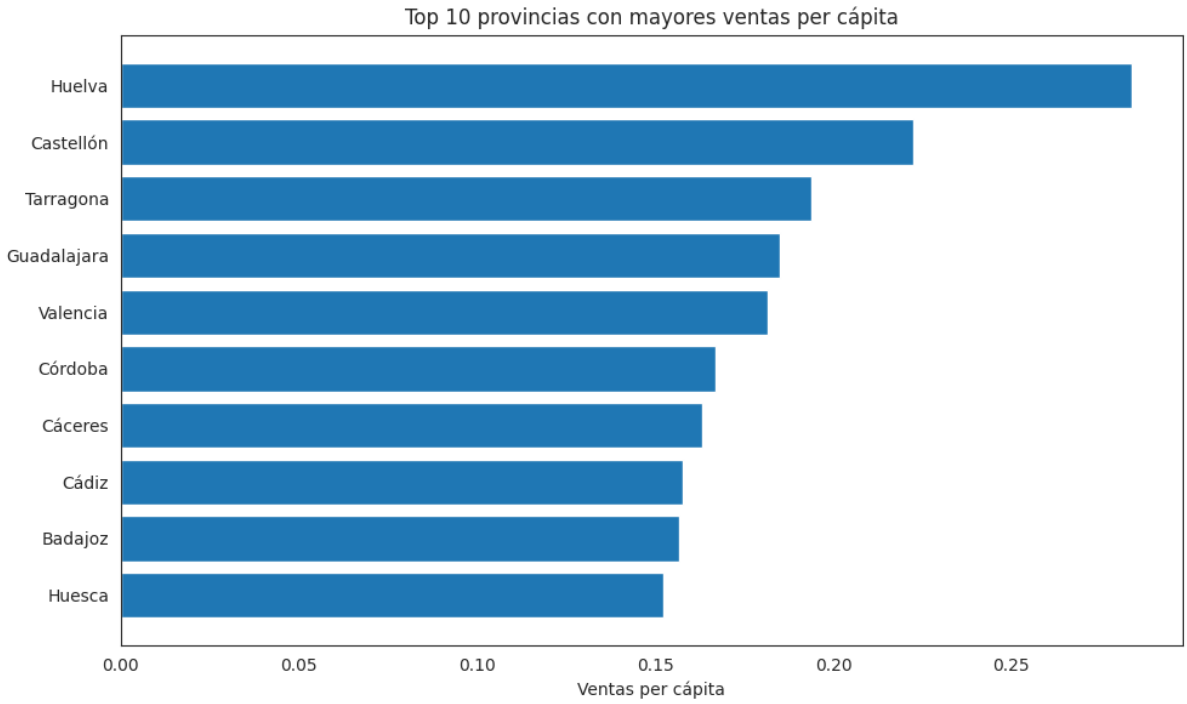


Fuente: elaboración propia

Para completar tanto el análisis de los perfiles estratégicos como su distribución territorial, así como para ayudar a contextualizar el rendimiento de la red comercial, se añadió la métrica de las ventas *per cápita*. Esta medida se construyó a partir de las ventas totales registradas en cada provincia y su población correspondiente al año 2015, según los datos recogidos en el INE (Instituto Nacional de Estadística). Esta métrica se implementó con el objetivo de observar en qué territorios se vende más en términos absolutos e identificar dónde se consume más por habitante para obtener una medida de la eficiencia del mercado en relación con la población.

Este análisis reveló una distribución dispar en el rendimiento comercial en las provincias, como es el caso de Huelva que despunta en el indicador de las ventas *per cápita*, aunque no lidera el ranking de ventas totales. Debido a esto, se pudieron identificar territorios con una elevada eficiencia comercial, que puede deberse a una gran penetración del canal, menor competencia o unas condiciones demográficas que impulsan el consumo.

Figura 36. Distribución de ventas per cápita por provincia.

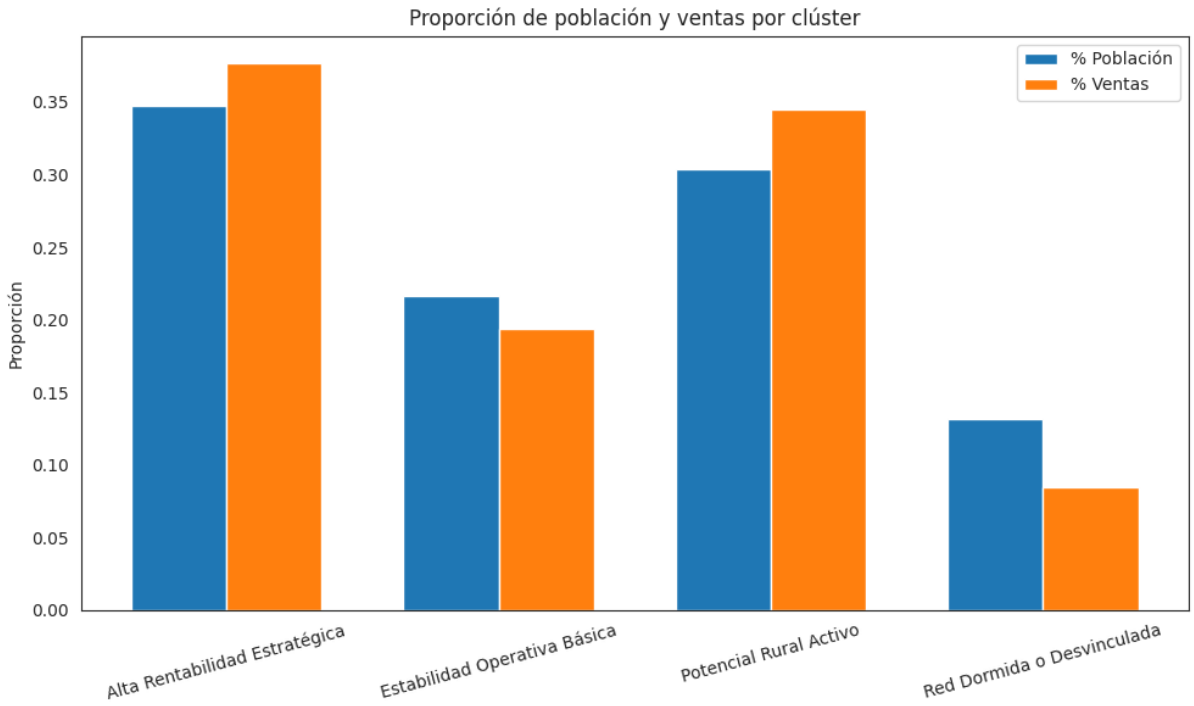


Fuente: elaboración propia

Este gráfico revela estos casos y plantea oportunidades relevantes para el desarrollo comercial, ya que estas provincias pueden actuar como referencia o benchmark interno frente a otras zonas con características estructurales similares, pero con menor rendimiento.

Tomando como base este análisis, se construyó un segundo nivel de segmentación por clúster, basado en el rendimiento comercial agregado: se calculó la proporción que representa cada segmento tanto sobre la población total nacional como sobre el total de ventas, con el objetivo de evaluar la eficiencia relativa de cada grupo. De esta forma, se pueden detectar desequilibrios entre el potencial demográfico y el rendimiento comercial efectivo.

Figura 37. Distribución población y ventas por segmento.



Fuente: elaboración propia

Esta comparativa demuestra diferencias significativas y aporta una visión estratégica complementaria en la toma de decisiones, ya que se constata el volumen de actividad comercial y, además, su eficiencia relativa:

- El grupo “Alta Rentabilidad Estratégica” presenta un rendimiento comercial alineado con su peso demográfico.
- El segmento “Estabilidad Operativa Básica” podría estar infrautilizado comercialmente ya que su contribución a las ventas es algo inferior a su peso demográfico.
- El segmento “Potencial Rural Activo” genera un porcentaje de ventas superior al de población, lo que indica un alto rendimiento comercial relativo. Esto puede ser debido a la fidelidad de las tiendas, a la existencia de poca competencia o a una distribución eficiente.

- En relación con el segmento “Red Dormida o Desvinculada” es el que muestra el mayor desequilibrio negativo en la relación entre ventas y población. Esto refuerza su clasificación como un perfil de bajo dinamismo y estos resultados pueden justificar estrategias de revisión, reactivación o incluso retirada selectiva.

Este enfoque analítico libera nuevas líneas de investigación, ya que se ha demostrado que la incorporación de otras variables externas de carácter sociodemográfico, como podría ser el nivel de renta, permitiría potenciar la interpretación y ayudaría a desarrollar hipótesis de mayor peso que pudieran explicar qué factores afectan en el desarrollo comercial relativo en cada zona del país.

3.3.3. Agrupación territorial por clubes de convergencia

En la tercera fase de modelado, se adoptó una perspectiva territorial, utilizando el algoritmo de los "clubes de convergencia" ideado por Phillips y Sul en 2007, que se aplicó a las series temporales de las ventas totales por provincia. Con este método, se pudo identificar patrones evolutivos similares entre distintos territorios, lo que aportó una visión más equilibrada del comportamiento comercial.

Antes de aplicar el modelo, los datos se acondicionaron mediante transformaciones logarítmicas, estandarizaciones y depuraciones estadísticas. Más tarde, se utilizó la función findClubs() para detectar comportamientos convergentes. El modelo distinguió inicialmente 11 clubes, que posteriormente se redujeron a 7 usando técnicas de fusión (mergeClubs()), lo que aportó claridad al estudio sin perder solidez estadística.
Los resultados evidenciaron que no existe una convergencia total entre las provincias españolas en lo que respecta a las ventas, ya que se observan dinámicas diferenciadas, por tanto, se hace necesario implementar estrategias de marketing específicas para cada grupo territorial. De igual modo, los resultados se representaron en mapas temáticos, lo que facilitó su comprensión y reforzó su uso como herramienta base para la toma de decisiones en campañas de marketing regionales.

En resumen, los tres modelos desarrollados ofrecieron una perspectiva detallada y estructurada del mercado desde la doble perspectiva del rendimiento comercial y la evolución territorial. Este planteamiento global contribuyó a obtener conclusiones relevantes y

operativas para orientar las decisiones comerciales y estratégicas de Altadis de manera informada y diferenciada.

En el **Anexo. A** se puede leer el código empleado para realizar los modelos.

3.3.4. Temporalidad de las ventas

El análisis exploratorio y visual de las ventas realizado ha permitido conocer su comportamiento en función de la temporalidad existente, así como determinar dinámicas concretas en función de provincia, día de la semana, mes y producto.

Este conocimiento adquirido y los patrones validados evidencian que el desarrollo de modelos predictivos ayudará a negocio a anticipar la demanda futura, con la consecuente posibilidad en la mejora y optimización de procesos de fabricación, gestión de inventarios, planificación logística y desarrollo de campañas comerciales u de marketing.

Se han aplicado y llevado a cabo los modelos ARIMA y ETS en las 4 provincias donde existe la concentración de las ventas, esto es como ya se ha dicho anteriormente: Madrid, Barcelona, Valencia y Alicante y en el conjunto de los datos de todo el país.

- Modelo ARIMA diario: se ha aplicado el modelo con frecuencia diaria en el total de los datos del país y en las 4 provincias, mediante la agregación de las ventas diarias "Sales_Uds" y su agrupación por día "Sales_DAY" para cada una, con la intención de captar cualquier fluctuación a muy corto plazo y ya que es un buen modelo para autocorrelación. Tiene una aplicabilidad útil para la gestión de stocks y planificación operativa.
- Modelo ARIMA semanal: se ha aplicado el modelo con frecuencia semanal en todo el país y en las mismas 4 provincias, mediante la agregación de las ventas de manera semanal, de esta manera se reduce el ruido causado por la variabilidad brusca de los fines de semana y festivos, y se pueden detectar patrones estacionales y tendencias con un enfoque más general, lo que es muy útil a nivel de planificación comercial.
- Modelo ETS diario: se ha aplicado el modelo en los datos de todo el territorio y en las 4 provincias en la suma de las ventas diarias para cada una, ya que el modelo es menos susceptible que ARIMA al ruido y detecta mejor estacionalidad y tendencias.

En el **Anexo B.** se puede leer el código empleado para realizar los modelos.

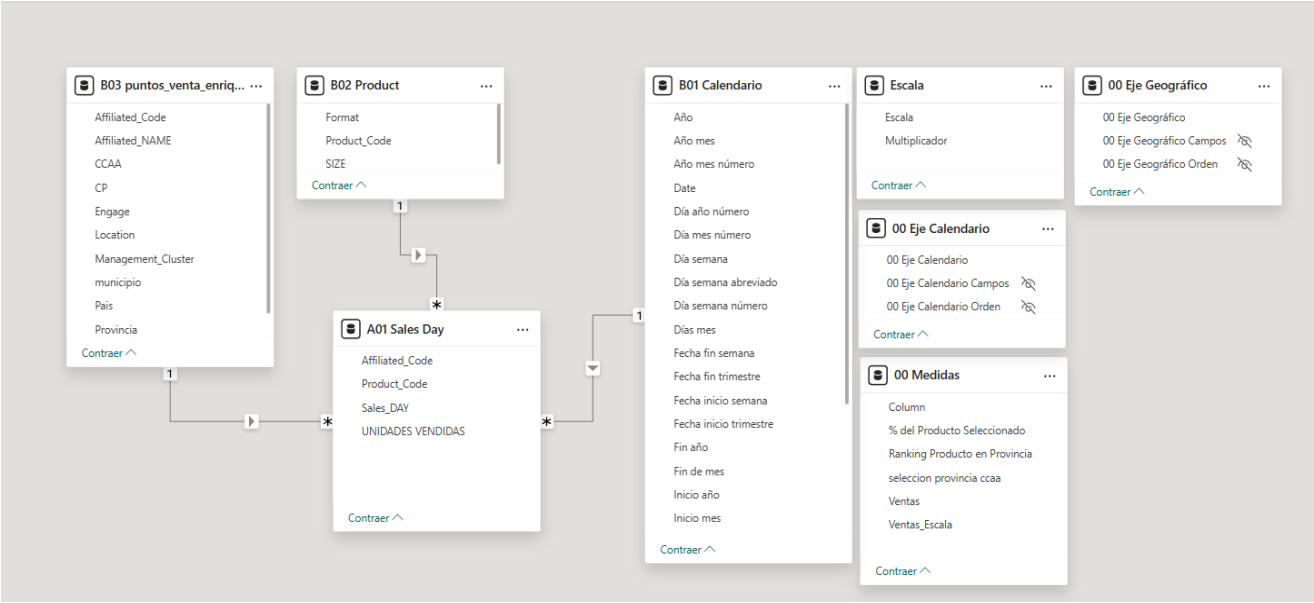
3.3.5. Dashboard

El cuadro de mandos implementado ha sido elaborado mediante un modelo estrella diseñado con la siguiente estructura, una tabla de hechos y tres de dimensiones:

- A00 Sales Day: Tabla de Hechos, ya que tiene las variables cuantitativas.
- B01 Calendario: que permite en el cuadro de mandos la segmentación por año, trimestre, mes y día, así como la visulaizacion de los diferentes días de la semana.
- B02 Product: con atributos como Product_Code y Format.
- B03 puntos_venta_enriquecido2_sinNA: compuesta por CCAA, Provincia y Affiliated Name.

A continuación, se puede ver el modelo estrella empleado:

Figura 38. Modelo Estrella.



Fuente: elaboración propia

Este modelo dimensional ha sido realizado con el fin de llevar a cabo un análisis OLAP y así analizar los datos desde múltiples perspectivas con las que conseguir información de valor para Altadis, de una forma interactiva, pudiendo consultar tantas agregaciones temporales como se desee, en cada momento de estudio de los datos y comparaciones geográficas, por producto, por punto de venta o por formato de producto.

Han sido creadas distintas medidas DAX para calcular las ventas y los porcentajes, y se han incluido parámetros geográficos y de calendario. A continuación, se muestran estos parámetros.

Figura 39. Eje calendario.

```
00 Eje Calendario = {
    ("Año", NAMEOF('B01 Calendario'[Año]), 0),
    ("Trimestre", NAMEOF('B01 Calendario'[Trimestre año]), 1),
    ("Mes", NAMEOF('B01 Calendario'[Año mes]), 2),
    ("Día", NAMEOF('B01 Calendario'[Date]), 3)
}
```

00 Eje Calendario	00 Eje Calendario Campos	00 Eje Calendario Orden
Año	'B01 Calendario'[Año]	0
Trimestre	'B01 Calendario'[Trimestre año]	1
Mes	'B01 Calendario'[Año mes]	2
Día	'B01 Calendario'[Date]	3

Fuente: elaboración propia

Figura 40. Eje geográfico

```
[00 Eje Geográfico ] = {
    ("Provincia", NAMEOF('B03 puntos_venta_enriquecido2_sinNA'[Provincia]), 0),
    ("CCAA", NAMEOF('B03 puntos_venta_enriquecido2_sinNA'[CCAA]), 1)
}
```

00 Eje Geográfico	00 Eje Geográfico Campos	00 Eje Geográfico Orden
Provincia	'B03 puntos_venta_enriquecido2_sinNA'[Provincia]	0
CCAA	'B03 puntos_venta_enriquecido2_sinNA'[CCAA]	1

. Fuente: elaboración propia

4. Construcción, prueba, implementación y despliegue

Esta etapa del estudio recoge el proceso de desarrollo de los modelos y su puesta en marcha. Se trató de validar los resultados obtenidos, entender su uso y aplicación en la realidad comercial y validar su potencial para integrarlo en el proceso de toma de decisiones de Altadis. Se incorporaron técnicas de análisis exploratorio, aprendizaje no supervisado y análisis territorial basado en series temporales. Durante la construcción de los modelos, se realizaron pruebas iterativas, cuantitativas y visuales, con el fin de comprobar que los grupos tuvieran coherencia interna y que tuvieran utilidad desde el punto de vista comercial. El trabajo se desarrolló en dos entornos principales, Python (Jupyter Notebooks) y R (RStudio), empleándose cada herramienta según las necesidades metodológicas de cada etapa.

4.1. Medidas de bondad del modelado

Para asegurar que los modelos desarrollados en este estudio fueran relevantes, válidos y consistentes se evaluaron mediante métricas cuantitativas y herramientas visuales, adaptadas a la naturaleza de cada técnica empleada.

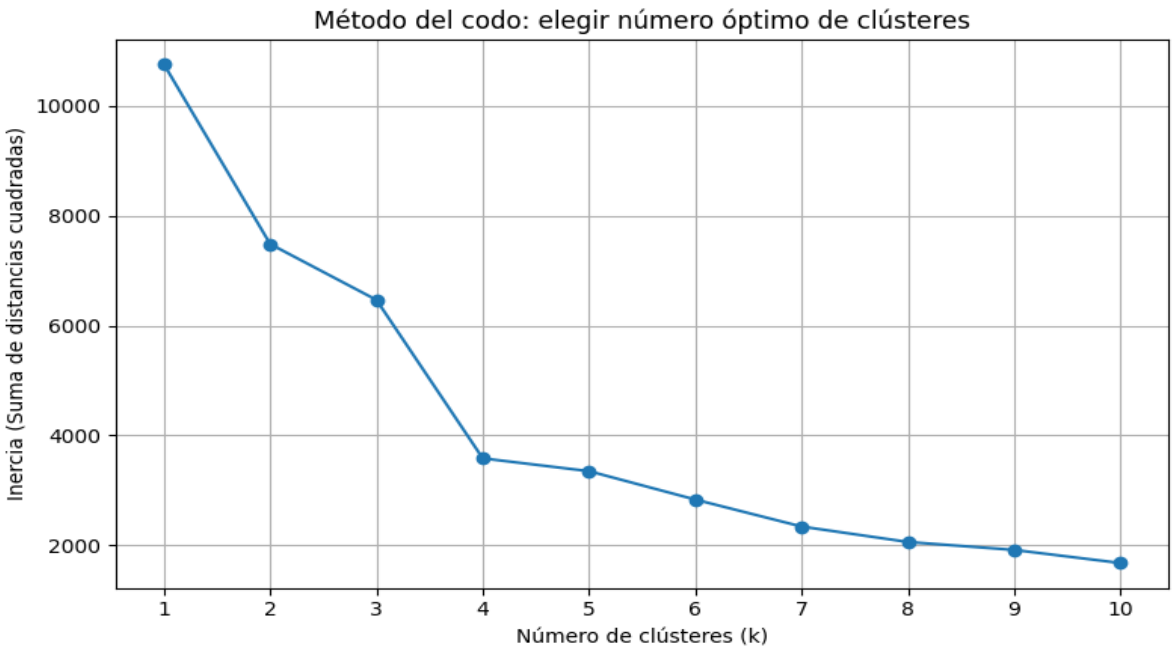
4.1.1. Segmentación de puntos de venta (estructural y por rendimiento comercial)

En los ejercicios de segmentación, tanto estructural como por rendimiento comercial, se aplicó el algoritmo K-means tras la estandarización de las variables. La selección del número óptimo de clústeres se basó en los criterios detallados a continuación.

4.1.1.1. Método de Elbow

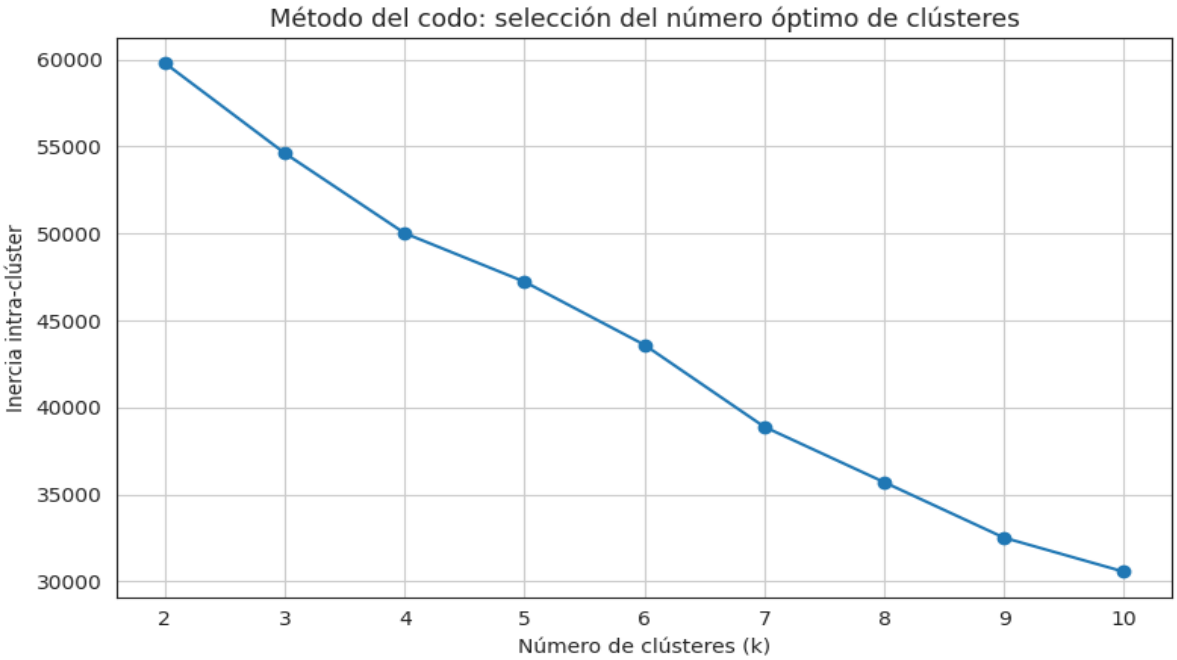
En este método se representa la inercia intra-clúster para valores de k entre 1 y 10

Figura 41. ***Método de Elbow en segmentación estructural.***



Fuente: elaboración propia

Figura 42*. Método de Elbow en segmentación por rendimiento de ventas*



. Fuente: elaboración propia

En ambos casos, se decidió complementar la validación visual que proporciona el método de Elbow con las métricas de Silhouette, Davies-Bouldin y Calinsky-Harabasz, ya que el método Elbow no ofrecía conclusiones definitorias para la selección del número óptimo de clústeres.

4.1.1.2. Índices de Silhouette, Davies-Bouldin y Calinsky-Harabasz

- El índice de Silhouette ayudó a verificar la calidad de la asignación de cada punto a su clúster. Para este coeficiente, los valores oscilan entre –1 y 1, considerándose una buena medida cuanto más se acerque a 1.
- El índice de Davies-Bouldin se utilizó para estudiar la compacidad y separación de los grupos. Para este índice, cuanto menor es el valor, mejor es la calidad de la agrupación.
- El índice de Calinski-Harabasz se utilizó para comparar la dispersión entre grupos respecto a la dispersión interna. En este índice, un valor elevado sugiere una estructura de clústeres más definida.

En el análisis, se calcularon estas métricas para distintos valores de k, por tanto, se utilizaron como criterio de selección del número óptimo de clústeres.

Figura 43. Comparativa de resultados de las métricas de validación interna de clústeres en la segmentación estructural.

Métodos	k=2	k=3	k=4
Silhouette Score	0.3788	0.2974	**0.4634**
Davies-Bouldin Index	1.1490	1.3882	**0.8420**
Calinski-Harabasz Score	1564.57	1186.20	**2383.98**

Fuente: elaboración propia

En el modelo de segmentación estructural las distintas métricas dieron como conclusión evidente que el número óptimo de clústeres era 4.

En el caso del modelo de segmentación por rendimiento comercial, se calcularon las métricas para k=2 hasta k=10, pero comparando con los resultados visuales que ofrecía el método Elbow, parecía apuntar a que el codo se mostraba a partir de k=4.

Figura 44. Comparativa de resultados de las métricas de validación interna de clústeres en la segmentación por rendimiento económico.

Métodos	k=3	k=4	k=5
Silhouette Score	0.121	0.178	0.197
Davies-Bouldin Index	2.629	2.332	2.060
Calinski-Harabasz Score	323.708	345.265	327.031

Fuente: elaboración propia

Se decidió finalmente optar por 4 grupos justificando la decisión en estos puntos:

- K=4 ofrecía equilibrio entre calidad y simplicidad ya que, con valores mayores de k, las métricas de validación interna mejoran notablemente, pero es a partir de k=4 donde se alcanzó una segmentación estructurada sin aumentar la complejidad del modelo
- Utilizando k=4 se obtuvieron grupos claramente diferenciados e interpretables, lo que facilitaría su aplicación en estrategias de marketing territorial
- Con k=4 se mantenía un buen rendimiento y se evitaba el sobreajuste.

De esta forma, los valores obtenidos apoyaron la elección final del modelo, mostrando una estructura clara y segmentaciones bien definidas y relevantes desde el punto de vista comercial. No obstante, se recoge el detalle del proceso de cálculo y las métricas obtenidas para cada valor de k en el **Anexo C.**

4.1.2. Clubes de convergencia territorial

El modelo se evaluó en varias etapas:

- Para comprobar la existencia de convergencia absoluta en las series temporales de las ventar normalizadas por provincia se aplicó el índice H.
- Para descubrir grupos de provincias con comportamientos de ventas similares en el tiempo se aplicó la función findClubs()
- Se simplificó el modelo sin perder rigor estadístico tras ejecutar la función mergeClubs() del método de Philips y Sul, ya que inicialmente se determinó la existencia de 11 clubes que, posteriormente, pasaron a ser 7.

Los resultados mostraron que no todas las provincias tienen comportamientos similares, es decir, no existe convergencia absoluta, sino que hay grupos con patrones de comportamiento distintos. Se generaron representaciones gráficas, como transiciones temporales, tendencias promedio por club y mapas, que permitieron verificar visualmente la solidez de modelo. Se confirmó de esta forma que el modelo se puede aplicar en la construcción de estrategias comerciales diferenciadas por región.

En el **Anexo D.** se puede observar todos los gráficos y mapas realizados.

4.1.3. Temporalidad de las ventas: ARIMA y ETS

A continuación, se presenta una tabla resumen donde se muestran los porcentajes medios de error MAPE (porcentaje de error absoluto medio) de cada modelo y su nivel de precisión predictiva en función de este, sabiendo que un MAPE con un porcentaje de <10% es excelente y varia hasta un MAPE de >50% que indica un mal ajuste del modelo.

Tabla 5. ***Tabla resumen MAPE.***

Provincia	Modelo	MAPE (%)	Precisión
Madrid	ARIMA diario	58.55	Mala
	ARIMA semanal	69.86	Mala
	ETS diario	71.13	Mala
Barcelona	ARIMA diario	83.15	Mala
	ARIMA semanal	99.85	Mala
	ETS diario	99.36	Mala
Valencia	ARIMA diario	98.66	Mala
	ARIMA semanal	126.37	Muy mala
	ETS diario	132.44	Muy mala
Alicante	ARIMA diario	**57.52**	Mala
	ARIMA semanal	89.03	Mala
	ETS diario	75.14	Mala
Todas las provincias	ARIMA diario	57.04	Mala
	ARIMA semanal	85.49	Mala
	ETS diario	72.67	Mala

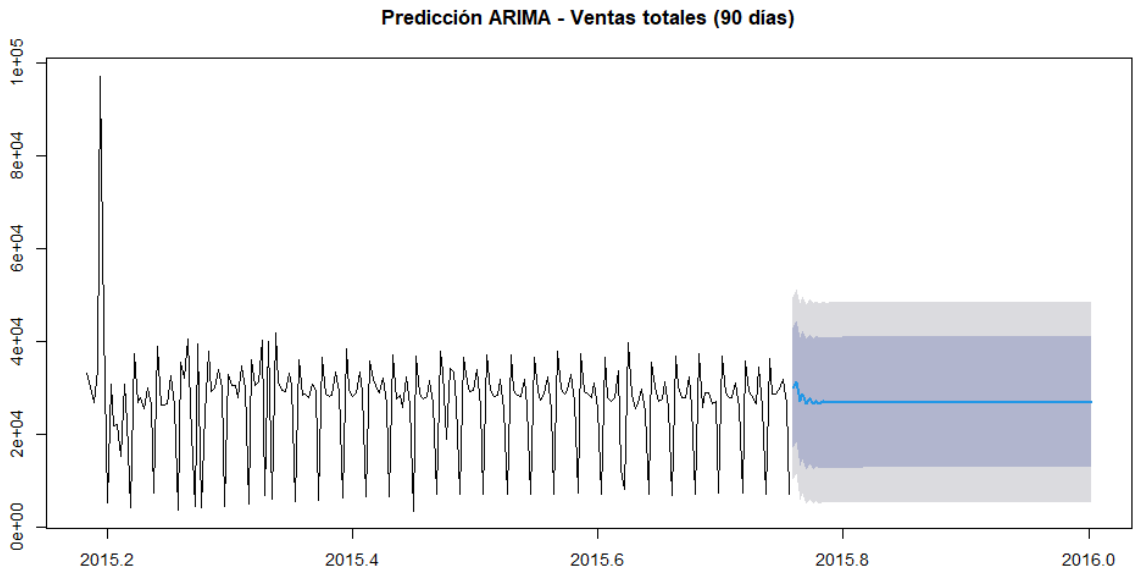
Fuente: elaboración propia

El modelo ARIMA no ha encontrado patrones estacionales ni tendencias estables ni en la serie diaria ni en la serie semanal de las ventas, aunque funciona mejor que el modelo ETS. La mala

precisión de los modelos puede ser debida al fuerte ruido en los datos la falta de ciclos estacionales detectables y un período temporal muy corto de menos de 1 año.

El modelo ARIMA diario para todo el territorio es claramente el más idóneo para un pronóstico de ventas totales y en contra de lo que se podría pensar e intuir el semanal eleva considerablemente su MAPE indicando que la precisión del pronóstico no mejora con la agregación semanal, dificultando así la observación de tendencias comerciales.

Figura 45. ***Modelo ARIMA diario España.***



Fuente: elaboración propia

En el **Anexo E.** se puede observar todos los gráficos y predicciones desarrolladas.

4.1.4. Dashboard

La evaluación de modelo se ha conseguido mediante la comprobación de:

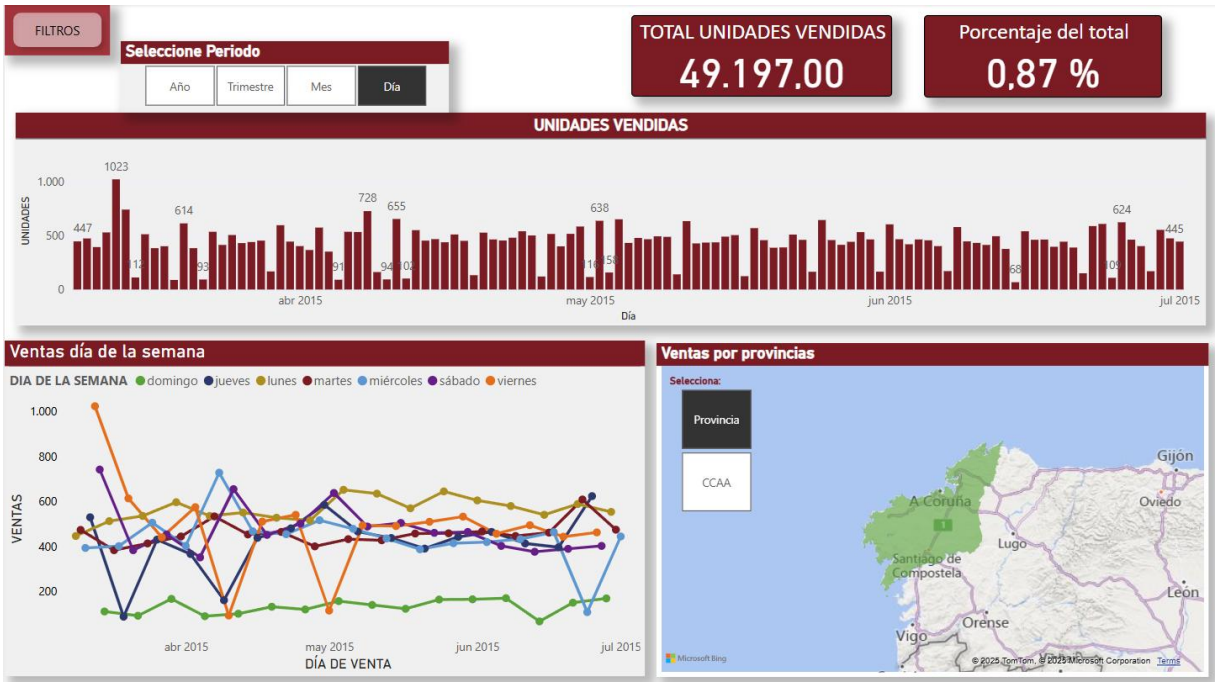
- Funcionalidad de las medidas, comprobación entre los filtros aplicados y los cálculos realizados.
- Facilidad de uso, se han verificado los tiempos de carga con los distintos filtros seleccionados.
- Correcto funcionamiento de los diferentes filtros y sincronización de las visualizaciones.

Estas comprobaciones garantizan la fiabilidad del modelo construido. Se muestran diferentes comprobaciones de usabilidad.

Figura 46. ***Menú de filtros con selecciones.***

Fuente: elaboración propia

Figura 47. ***Informe de ventas con selecciones.***



Fuente: elaboración propia

4.2. Ejemplo de aplicación

Para demostrar la utilidad de los modelos generados, se identificaron diversos ejemplos de aplicación a partir de los resultados obtenidos. Estos casos prácticos evidenciaron cómo los análisis permiten tomar decisiones estratégicas diferenciadas, a partir de las características estructurales de cada tienda, del comportamiento territorial de las ventas o del factor tiempo.

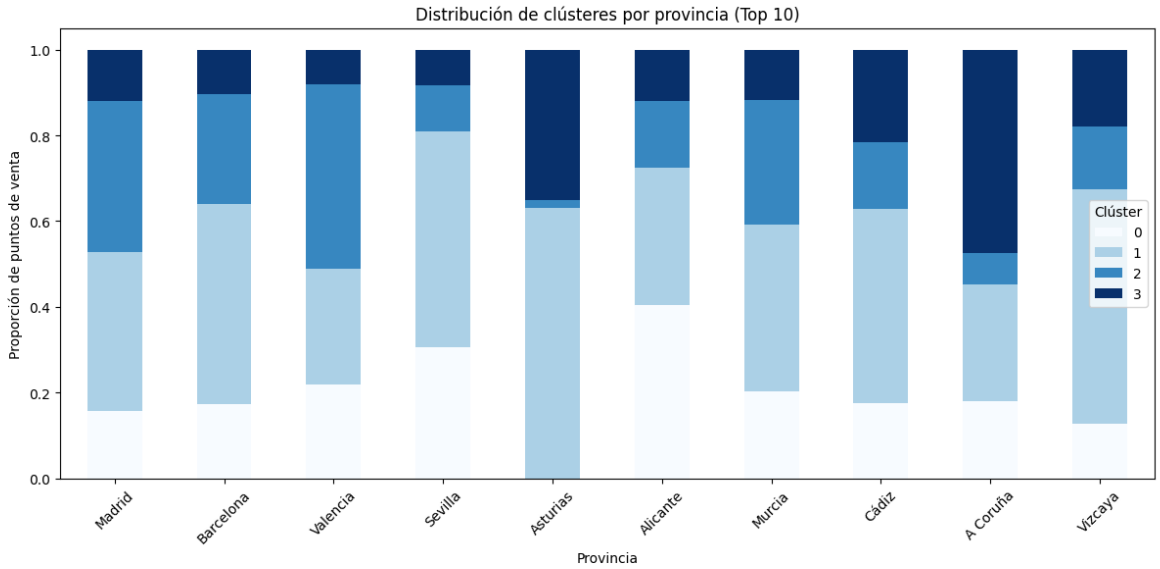
4.2.1. Segmentación estructural de puntos de venta

A partir del modelo de clusterización estructural, se pudo asignar a cada punto de venta un perfil definido según sus características, identificándose 4 tipos principales de establecimientos:

- Grupo 0 “Estabilidad Vacacional”: establecimientos ubicados en zonas rurales o vacacionales, con implicación comercial media y tamaño pequeño.
- Grupo 1 “Rotación Funcional en Zonas de Paso”: tiendas ubicadas en zonas de paso, con implicación comercial media y tamaño pequeño.
- Grupo 2 “Alta Implicación Estratégica”: tiendas con alto compromiso comercial, ubicadas en zonas urbanas o rurales en locales de tamaño mediano.
- Grupo 3 “Desvinculación y Baja Viabilidad Comercial”: tiendas con muy bajo nivel de implicación comercial, localizadas en zonas rurales o turísticas. Locales muy grandes o con tamaño desconocido (N.D. en el dataset).

Esta segmentación permitió agrupar cada tienda según unas características estructurales concretas que tenían relación con la implicación comercial de las tiendas en la estrategia comercial de la empresa.

Figura 48. ***Distribución de los clústeres en las provincias con mayor número de puntos de venta.***



Fuente: elaboración propia

4.2.2. Segmentación por rendimiento comercial

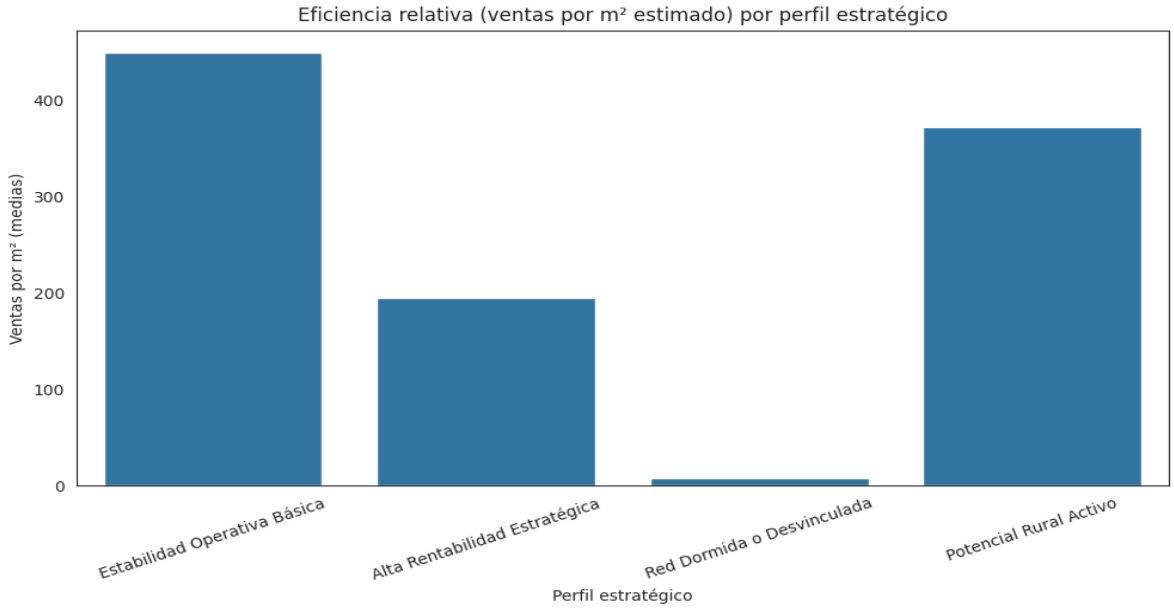
Al integrar el volumen total de ventas en el esquema de segmentación, se consiguió distinguir categorías de alto y bajo desempeño comercial y se pudo definir una serie de perfiles comerciales bien diferenciados:

- Grupo 0 “Estabilidad operativa básica”: incluye tiendas ubicadas en zonas inespecíficas o rurales, en locales pequeños, con nivel de implicación comercial medio y desempeño comercial aceptable.
- Grupo 1 “Alta rentabilidad estratégica”: ventas muy altas, nivel de implicación comercial máximo. Tiendas grandes ubicadas en distintas zonas.
- Grupo 2 “Red dormida o desvinculada”: tiendas con bajo nivel de implicación comercial y bajas ventas. Ubicadas en zonas rurales y con tamaño desconocido.
- Grupo 3 “Potencial rural activo”: ventas medias-altas y buen nivel de compromiso comercial. Tiendas ubicadas en pueblos medianos y alguna ciudad, en locales de tamaño medio.

El establecimiento de estos segmentos permite la fijación, a su vez, de estándares de referencia (benchmark) internos, que ofrecen a la dirección la posibilidad de establecer

estrategias de impulso o expansión comercial hacia los puntos de venta que compartan características estructurales similares.

Figura 49. ***Distribución de la eficiencia de ventas por m2 con relación a los perfiles estratégicos obtenidos en la segmentación por rendimiento comercial.***



Fuente: elaboración propia

Por otra parte, en el análisis de las ventas *per cápita* de cada perfil comercial, se evidenció el potencial comercial de cada segmento frente a su rendimiento real. Por poner un ejemplo, el segmento "Potencial Rural Activo" agrupa alrededor del 30% de la población censada en el INE en 2015, pero este segmento produce un volumen de negocio superior a esa cifra, lo que puede indicar que se está realizado un buen aprovechamiento del canal en estos territorios y podría dar pie a implementar estrategias de refuerzo comercial o consolidación de las inversiones. Por el contrario, el segmento "Red Dormida o Desvinculada" engloba un amplio sector de la población, pero su aporte a ventas es sensiblemente menor. Este desequilibro podría justificar un estudio de la eficiencia de la red comercial o el lanzamiento de acciones concretas para reactivar el interés comercial de las tiendas. De igual modo, el segmento "Alta Rentabilidad Estratégica" presenta equilibrio entre población y facturación por lo que Altadis podría utilizar este segmento como nivel de referencia interno a imitar por los segmentos menos eficientes.

Con este tipo de análisis se puede pasar de una segmentación puramente descriptiva al desarrollo de una estrategia comercial basada en datos objetivos, donde se puedan optimizar los recursos comerciales adaptándolos al contexto real de cada grupo de tiendas.

4.2.3. Agrupación territorial por clubes de convergencia

Tras analizar las ventas por provincia desde el enfoque de la convergencia territorial, se observó que la demanda no se comporta de forma homogénea a lo largo del tiempo.

Se identificó que, por ejemplo, Madrid, Valencia, Cataluña y Baleares se incluyeron en un mismo grupo, ya que seguían una dinámica de ventas convergente. Con esta información, se pueden desarrollar estrategias que abarquen regiones enteras, diseñadas a la medida del club y no sólo para cada provincia por separado.

De acuerdo con los mapas y gráficos diseñados, se permitió interpretar de forma clara y sencilla los patrones temporales y geográficos, lo que posibilita establecer acciones diferenciadas según la región y adaptadas a la realidad del mercado.

4.2.4. Temporalidad de las ventas: ARIMA y ETS

Ya que los modelos han salido con una calidad baja, no pueden ser aplicados para la planificación de stocks, realización de estrategias tácticas-operativas y/o comerciales. Se necesitarían mayor volumen de datos para poder entrenar al modelo correctamente y que sirvan para predecir la demanda de tabaco en los diferentes periodos del año.

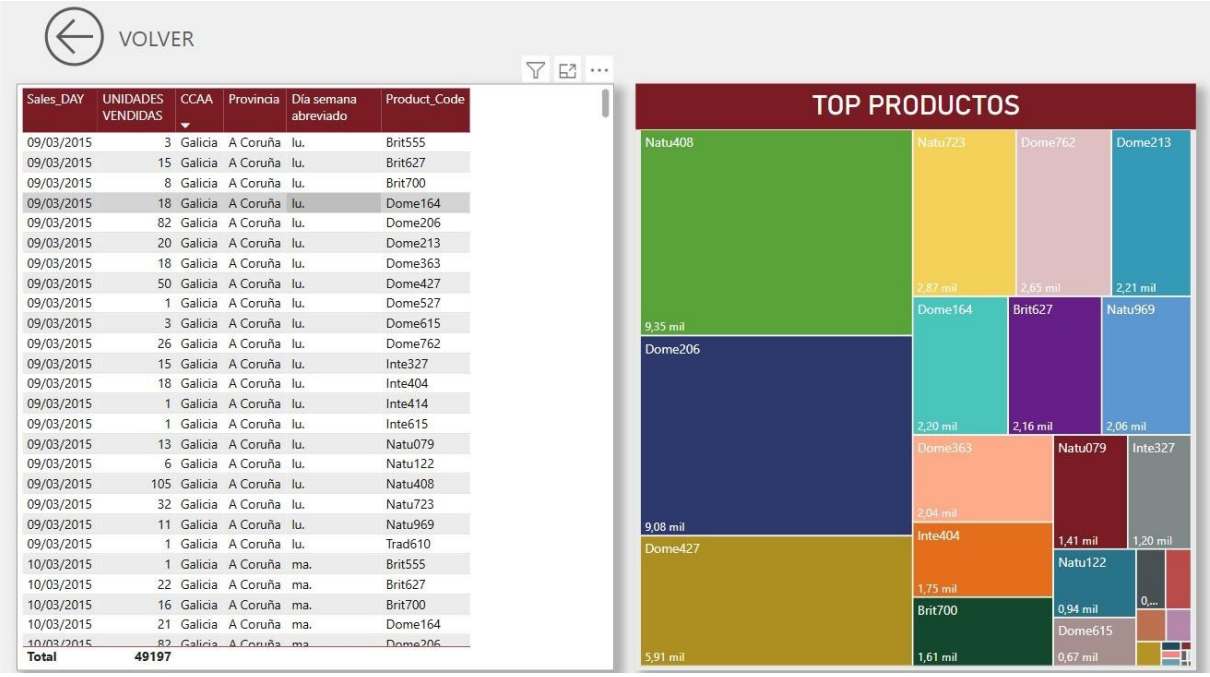
4.2.5. Dashboard

El cuadro de mando es muy útil para el análisis de las ventas, y a modo de ejemplo se analizan para la provincia de A Coruña entre marzo y julio de 2015. Aplicando los filtros, se averigua que:

- Volumen total de ventas: 4197 unidades, lo que representa el 0,87 % del total de las ventas.
- Productos más vendidos: destacan los códigos Natu408 y Dome206 como se observa en el tooltip al pasar por encima del grafico del mapa.
- Distribución temporal: se aprecian como los días de ventas más elevadas los viernes y lunes, por la influencia del fin de semana.

- Y si se desea se puede ampliar la información en el detalle de ventas.

Figura 50. ***Detalle de Ventas con selecciones.***



Sales_DAY	UNIDADES VENDIDAS	CCAA	Provincia	Día semana abreviado	Product_Code
09/03/2015	3	Galicia	A Coruña	lu.	Brit555
09/03/2015	15	Galicia	A Coruña	lu.	Brit627
09/03/2015	8	Galicia	A Coruña	lu.	Brit700
09/03/2015	18	Galicia	A Coruña	lu.	Dome164
09/03/2015	82	Galicia	A Coruña	lu.	Dome206
09/03/2015	20	Galicia	A Coruña	lu.	Dome213
09/03/2015	18	Galicia	A Coruña	lu.	Dome363
09/03/2015	50	Galicia	A Coruña	lu.	Dome427
09/03/2015	1	Galicia	A Coruña	lu.	Dome527
09/03/2015	3	Galicia	A Coruña	lu.	Dome615
09/03/2015	26	Galicia	A Coruña	lu.	Dome762
09/03/2015	15	Galicia	A Coruña	lu.	Inte327
09/03/2015	18	Galicia	A Coruña	lu.	Inte404
09/03/2015	1	Galicia	A Coruña	lu.	Inte414
09/03/2015	1	Galicia	A Coruña	lu.	Inte615
09/03/2015	13	Galicia	A Coruña	lu.	Natu079
09/03/2015	6	Galicia	A Coruña	lu.	Natu122
09/03/2015	105	Galicia	A Coruña	lu.	Natu408
09/03/2015	32	Galicia	A Coruña	lu.	Natu723
09/03/2015	11	Galicia	A Coruña	lu.	Natu969
09/03/2015	1	Galicia	A Coruña	lu.	Trad610
10/03/2015	1	Galicia	A Coruña	ma.	Brit555
10/03/2015	22	Galicia	A Coruña	ma.	Brit627
10/03/2015	16	Galicia	A Coruña	ma.	Brit700
10/03/2015	21	Galicia	A Coruña	ma.	Dome164
10/03/2015	82	Galicia	A Coruña	ma.	Dome206
Total	**49197**				

Fuente: elaboración propia

En el siguiente enlace y en el código QR se puede acceder al informe si se está autorizado para su visualización y si se dispone de Power BI pro, ya que se elaboró bajo una licencia pro:

https://app.powerbi.com/Redirect?action=OpenReport&appId=9c313ac0-6763-4828-b1e7-08c712d58ddc&reportObjectId=1f74add9-9b67-41f2-b149-85929f421c93&ctid=899789dc-202f-44b4-8472-a6d40f9eb440&reportPage=dac09aff027e04d4a7b7&pbi_source=appShareLink&portalSessionId=df9afbe5-a4e7-43ba-8e87-c85a55869a59

Power BI

Además, se adjunta **Anexo F.** con imágenes donde se comprueba su funcionalidad.

5. Conclusiones

El presente trabajo ha consistido en el diseño e implementación de un proyecto de inteligencia de negocio para la empresa Altadis S.A., simulando un servicio de consultoría con el fin de extraer información relevante que respalde la toma de decisiones estratégicas de la organización. Se obtienen las siguientes conclusiones en sintonía con los desafíos planteados:

Objetivo general: *Desarrollar un proyecto de inteligencia de negocio para Altadis con el objetivo de extraer insights que respalden la toma de decisiones estratégicas.*

A lo largo del proyecto se ha demostrado que incluso con datos parciales y anonimizados, es posible construir modelos analíticos que arrojan hipótesis valiosas sobre el comportamiento del mercado. En particular, se ha identificado que la estructura comercial de Altadis presenta patrones coherentes con tres factores estructurales principales: el factor fronterizo, el factor turístico y la densidad de población. Estas dimensiones explican de manera relevante tanto los volúmenes de venta como la lógica detrás de los clústeres de establecimientos identificados en el análisis.

Objetivo 1: *Realizar un estudio de comprensión del negocio, incluyendo revisión de fuentes internas y externas.*

Se realizó una revisión sistemática y profunda tanto de la información interna proporcionada por Altadis como de fuentes externas relevantes: informes sectoriales, análisis de mercado, estudios regulatorios y tendencias de consumo en el sector tabacalero español y europeo. Este proceso permitió alcanzar un conocimiento cabal del contexto operativo de la compañía, identificando los principales retos y oportunidades a los que se enfrenta en un entorno marcado por la transformación digital, la presión regulatoria y la evolución de los hábitos de consumo. La integración de perspectivas externas e internas fue clave para contextualizar los datos y orientar el análisis hacia los factores estructurales que inciden en el desempeño comercial de Altadis, como la influencia de la proximidad fronteriza-costera, la estacionalidad turística y la concentración demográfica en grandes urbes.

Objetivo 2: *Llevar a cabo un análisis exploratorio y procesamiento de los datos disponibles mediante técnicas de ETL y estadística descriptiva.*

El trabajo de análisis de datos comenzó con una exhaustiva fase de limpieza y depuración: se eliminaron duplicados, se corrigieron inconsistencias y se gestionaron errores derivados de la exportación y transcripción de los datos originales. Posteriormente, se enriquecieron las tablas mediante la incorporación de información de bases de datos de códigos postales, lo que permitió añadir variables territoriales y sociodemográficas de gran valor analítico. Se realizaron uniones y cruces de tablas para consolidar una base de datos robusta, apta para el análisis exploratorio. La estadística descriptiva aplicada facilitó la detección de patrones temporales y geográficos, como la concentración de ventas en determinados días de la semana (especialmente los lunes, debido al cierre dominical de los puntos de venta) y la marcada estacionalidad en las zonas costeras mediterráneas durante los meses de verano. Este proceso sentó las bases para la posterior segmentación y permitió una primera aproximación y guía para encontrar variables comerciales determinantes.

Objetivo 3: *Aplicar algoritmos de segmentación para identificar patrones y grupos relevantes entre los puntos de venta*s *y estudiar la temporalidad de las ventas con el objetivo de predecir las ventas futuras.*

La aplicación de técnicas de clustering y de clubes de convergencia permitió segmentar la red de puntos de venta en grupos con comportamientos claramente diferenciados, así como encontrar patrones geográficos. Se segmentó según la estructura del establecimiento, identificando grupos de afinidad como por ejemplo “Rotación Funcional en Zonas de Paso”, definidas como puntos de venta ubicados en zonas de paso, con implicación comercial media y tamaño pequeño. De esta forma se observó que variable estadísticamente más significativa para una agrupación estructural es la relacionada con el tamaño de los locales, seguida de la ubicación y el nivel de implicación comercial. También se aplicó la técnica de clusterización de aprendizaje no supervisado para encontrar grupos por similitudes en rendimientos comerciales, encontrando segmentos como por ejemplo “Alta rentabilidad estratégica”, destacando por las ventas medias más altas de todas las agrupaciones, con un nivel de implicación comercial máximo y con locales principalmente grandes.

Se encontraron convergencias en patrones caracterizados por picos estacionales asociados al turismo vacacional en provincias mediterráneas; zonas limítrofes como Girona o Badajoz, donde las ventas pueden estar vinculadas al turismo de compra y a dinámicas transfronterizas; o grupos que engloban áreas urbanas densas como Madrid o Barcelona, con una estructura comercial heterogénea y coexistencia de establecimientos de alta y baja rotación. Además, se detectaron clústeres de 'Potencial Rural Activo" así como una "Red dormida o desvinculada" en zonas rurales poco densas. Estos resultados avalan la hipótesis de que la estructura comercial de Altadis no es homogénea, sino que responde a la interacción de factores geográficos y sociodemográficos bien definidos como los factores estructurales señalados. Por otro lado, los modelos de predicción de ventas ARIMA y ETS no ha resultado ser de un buen valor predictivo principalmente debido a que los datos para el aprendizaje proporcionados se limitan a un muy corto periodo de tiempo.

Objetivo 4: *Diseñar e implementar u cuadro de mandos con KPIs representativos para facilitar la toma de decisiones.*

Se diseñó una propuesta de cuadro de mando que permite visualizar los principales indicadores de negocio necesarios para la gestión, valoración y análisis de las ventas con un profundo detalle. Para ello, se definieron indicadores clave de rendimiento (KPIs) adaptados al contexto comercial de Altadis y se estructuraron visualizaciones dinámicas con filtros interactivos geográficos, de fecha y de producto, que facilitan la toma de decisiones. Aun no habiéndose desplegado en el entorno real de MicroStrategy en el que trabaja la empresa, esta propuesta está preparada para integrarse fácilmente en esta plataforma, sirviendo como herramienta de soporte para los equipos de Altadis dedicados a la estrategia comercial y planificación.

En definitiva, los resultados del análisis apuntan a que la estructura comercial de Altadis no es homogénea y que existen factores estructurales que explican las diferencias entre provincias y grupos de establecimientos. Sin embargo, también es importante destacar que los *insights* obtenidos deben tomarse con precaución, ya que los datos analizados cubren un periodo corto, están anonimizados y no incluyen todas las dimensiones relevantes del negocio, tales como precios o caracterización del consumidor final. La metodología empleada —enriquecimiento geográfico, clustering no supervisado y análisis temporal— ha sido efectiva

para el objetivo del trabajo, aunque su potencia analítica se ha visto limitada por la calidad y antigüedad de los datos disponibles. A pesar de las limitaciones, se han logrado hipótesis operativas que podrían ser de interés para una segunda fase del proyecto, centrada en la validación de los factores detectados mediante datos más actuales, validaciones cualitativas y una integración transversal con otros departamentos de la compañía. Así, se sientan las bases para una inteligencia de negocio más sofisticada y alineada con los retos del sector y la visión estratégica de Altadis.

6. Limitaciones y prospectiva

Una vez finalizada la exposición y conclusiones del trabajo, es fundamental poner de manifiesto tanto las restricciones que han limitado el alcance del proyecto como las oportunidades de estudio descubiertas en el mismo, que se detallan en los siguientes apartados.

6.1. Limitaciones

Datos anonimizados: la codificación atributos relevantes como nombres de productos y clientes supone un hándicap e impide personalizar completamente los análisis. Se realizaron intentos por lograr la desanonimización de los datos, buscando patrones en los propios datos como apelando a fuentes externas. Finalmente se desestimó esta vía en aras de priorizar los plazos.

Figura 51. ***Ejemplo de búsqueda de patrones en los prefijos de los productos:***

Patrones Observables en la Codificación - Altadis 2015

Prefijo	N° Productos	Rango Tamaños	Formatos Únicos	Formato Dominante
Dome	24	29 - 538	3	ASL
Inte	11	43 - 1132	3	ASL
Natu	9	85 - 283	2	ASL
Fren	6	85 - 1415	3	ASL
Brit	4	85 - 142	1	ASL
Trad	4	142 - 142	1	ASL
Don1	1	142 - 142	1	ASL
Don8	1	425 - 425	1	ATA

Fuente: elaboración propia

Corto periodo temporal de datos: La cobertura temporal es insuficiente (solo algunos meses de 2015). Esto imposibilita análisis de estacionalidad anual y restringe la posibilidad de utilizar con efectividad modelos de predicción de series temporales como los modelos ARIMA.

Obsolescencia de los datos: Los datos tienen casi una década de antigüedad. Esto impide extrapolar los hallazgos al contexto actual del mercado tabacalero sin un proceso adicional de actualización o enriquecimiento temporal.

Variables clave ausentes o mal definidas: elementos como el precio de los diferentes productos no se conoce, tampoco los costes, lo cual limita la posibilidad de realizar análisis de tipo económico – financiero.

Falta de acceso a datos de consumo final: solo se tuvo acceso a datos de distribución y venta, pero no a datos directos de consumidor, lo cual limita la comprensión de patrones de comportamiento y preferencias, en consecuencia, la capacidad de proponer acciones de maximización de ventas o de fidelización.

Carencia de acceso a herramientas internas: No se ha podido acceder a los cuadros de mando o informes internos actuales de Altadis, lo que impide realizar benchmarking interno o validar hipótesis. En ese sentido, se supone, pero no se conoce con certeza, el significado de variables como *Cluster_management* o *Engage.*

Tiempo y recursos limitados: Al tratarse de un trabajo académico con plazos cerrados, no fue posible explorar otras áreas de alcance como aspectos logísticos, stock o cadena de suministro.

Rol externo ficticio: Al no ser una consultoría real integrada en la empresa, existen barreras informativas que impiden realizar un análisis organizacional profundo (resistencia al cambio, cultura de datos, capacidades internas reales, etc.).

6.2. Prospectiva

Tras el trabajo llevado a cabo, se detectan diferentes líneas de ampliación y mejora con capacidad para mejorar el análisis de inteligencia de negocio:

Actualización y ampliación de los datos: para mejorar la estacionalidad, los modelos predictivos y adaptarse a las dinámicas de mercado actuales, se deberían añadir registros más recientes, así como ampliar el periodo de análisis.

Análisis de sentimiento: Utilización de técnicas de text mining aplicadas a redes sociales, reseñas o foros sobre percepción de marcas y nuevos productos (NGP).

Integración de datos sociodemográficos y turísticos: Incorporar fuentes externas (INE, Eurostat, Google Mobility, etc.) para enriquecer la segmentación geográfica y anticipar picos de demanda en regiones específicas.

Modelos predictivos territoriales: Desarrollar predicciones de ventas per cápita o por canal, incorporando variables exógenas como fiscalidad, demografía, turismo o campañas comerciales.

Simulación de escenarios regulatorios: Analizar impacto de aumentos fiscales o prohibiciones sobre el canal vending, mediante modelos de elasticidad y simulación.

Análisis temporal de largo plazo: estudio de la evolución de ventas y cambios en comportamiento de cliente usando series temporales actualizadas.

Incorporación de datos digitales y de CRM: Integrar variables como frecuencia de compra, engagement en canales digitales o historial de promociones para enriquecer la personalización comercial. Caracterización del perfil sociodemográfico del cliente a través de medios digitales o mediante programas de fidelización.

Optimización logística avanzada: optimización de rutas y distribución, considerando factores como ventas históricas, geolocalización, tráfico o estacionalidad, activación de ordenes de reposición automáticas, integración de stocks centralizada.

Auditoría y gobierno del dato: Implementar un departamento de IS o Master Data, estableciendo marcos de calidad y trazabilidad del dato para evitar inconsistencias como las detectadas en este proyecto (duplicados, errores de ubicación, variables mal definidas).

Ampliación del alcance a otros departamentos: En futuras fases, el modelo de análisis podría extenderse a áreas como control de costes, evaluación de campañas promocionales o satisfacción del cliente. Estas medidas podrían dar soporte a estudios sobre el impacto de los marcos regulatorios o a potenciales estrategias innovadoras en la industria como *pricing* dinámico en el mercado de tabaco.

Referencias bibliográficas

Altadis, S.A. (s.f.). *Imperial Brands Altadis*. Obtenido de https://www.altadis.com/sobre-altatis/quienes-somos/#:~:text=Desde%20el%20a%C3%B1o%202008%20que,Estados%20Unidos%2C%20Alemania%20y%20Australia.

Analistas Financieros Internacionales, S.A. (s.f.). *El impacto socioeconmico del empaquetado generico.* mesa del tabaco. Obtenido de https://www.mesadeltabaco.com/public/Attachment/2024/10/241030_InformeAfi_Elimpactosocioeconmicodelempaquetadogenrico.pdf

Cadenas, J. F. (5 de Abril de 2024). *Newtral*. Obtenido de Newtral: https://www.newtral.es/empaquetado-neutro-tabaco/20240405/

GeoNames. (s.f.). Obtenido de https://download.geonames.org/export/zip/

Imperial Brands Altadis. (s.f.). Obtenido de https://www.altadis.com/sobre-altatis/quienes-somos/

Imperial Brands. (s.f.). *iseazy.com*. Obtenido de https://iseazy.com/dl/c9398f1ce9c44c958b61c3252a305f4d#/slide/VZ0qDyvXQF

Imperial Brands PLC. (s.f.). *Annual Reports and Account 2024.*

Imperial Brands.Principales Conceptos de Negocio. (s.f.). Obtenido de https://iseazy.com/dl/c9398f1ce9c44c958b61c3252a305f4d#/slide/VZ0qDyvXQF

Informesdeexpertos.com. (s.f.). Obtenido de https://www.informesdeexpertos.com/informes/mercado-de-tabaco

Ingelmo Heras, R. (s.f.). *El Mercado de tabaco español desde un analisis de su demanda.* Altadis Imperial Brands. Obtenido de https://www.altadis.com/uploads/data/altadis-elmercadodetabacoespan%CC%83oldesdeunana%CC%81lisisdesudemanda-045710.pdf

Instituto Nacional de Estadística. (s.f.). *Instituto Nacional de Estadística.* Obtenido de ine.es: https://www.ine.es/jaxiT3/Datos.htm?t=2852#_tabs-tabla

Ministerio de Hacienda. (s.f.). Obtenido de Vicepresidencia Primera del Gobierno: https://www.hacienda.gob.es/es-ES/Areas%20Tematicas/CMTabacos/Paginas/EstadisticasCMT2024.aspx

Ministerio de Hacienda. (1992). *Ley 38/1992, de 28 de diciembre, de Impuestos Especiales.* Obtenido de Boletín Oficial del Estado: https://www.boe.es/buscar/act.php?id=BOE-A-1992-28741

Ministerio de Hacienda. (s.f.). *Estadísticas Mercado de Tabacos 2024.* Obtenido de https://www.hacienda.gob.es/ca-ES/Areas%20Tematicas/CMTabacos/Paginas/EstadisticasCMT2024.aspx

Ministerio de Sanidad. Gobierno de España. (2023). *Informe sobre los "nicotine pouches" (bolsas de nicotina): situación y posicionamiento.*

Ministerio de Sanidad. Gobierno de España. (2023). *Ministerio de Sanidad.* Obtenido de https://www.sanidad.gob.es/areas/promocionPrevencion/tabaco/profesionales/docs/NicotinePouches_informe.pdf

Observatorio Nacional de Tecnología y Sociedad. (2023). *Uso de inteligencia artificial y big data en las empresas españolas.* Ministerio de Asuntos Económicos y Transformación Digital.

Organización Mundial de la Salud. (2023). *WHO report on the global tobacco epidemic 2023: Protect people from tobacco smoke.* Obtenido de https://iris.who.int/handle/10665/375711

Proposito y Vision . (s.f.). Obtenido de Imperial Brands Altadis: https://www.altadis.com/sobre-altatis/nuestra-vision/

Randstad Research. (2025). *Informe de Tendencias Salariales.*

Ross, P. (29 de Agosto de 2023). *The Power Of Business Intelligence—And How To Put It Into Practice*. Obtenido de Forbes.com: https://www.forbes.com/councils/forbesagencycouncil/2023/08/29/the-power-of-business-intelligence-and-how-to-put-it-into-practice/

Sherman, R. (2014). *Business intelligence guidebook: From data integration to analytics.* Morgan Kaufman.

The Official Board. (octubre de 2024). Obtenido de https://www.theofficialboard.es/organigrama/altadis

Vallalta Rueda, J. F. (s.f.). *www.healthdataminer.com*. Obtenido de Health Data Miner: https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

Anexo A. Código Clubes de Convergencia

Código de CLUBES DE CONVERGENCIA

```
#Comprovacion de la existencia de convergencia absoluta (las ventas por provincia se comportan
de manera homogénea)


# Cargar librerías necesarias
library(readr)
library(dplyr)
library(tidyr)
library(lubridate)
library(readxl)
library(ConvergenceClubs)
library(ggplot2)


ventas<-read_delim("C:/unir/MASTER/2 CUATRIMESTRE/TFM/empresa/Data set Red Proyectum -
Altadis/SalesDay.csv",delim = ";",
                    col_names = c("Sales_DAY", "Affiliated_Code", "Product_Code",
"Sales_Uds"),col_types = cols(.default = "c"))%>%
                      mutate(Sales_DAY = ymd(Sales_DAY),Sales_Uds = as.numeric(Sales_Uds))%>%
                      filter(!is.na(Sales_DAY))

puntos<-read_excel("C:/unir/MASTER/2 CUATRIMESTRE/TFM/empresa/Data set Red Proyectum -
Altadis/nuevos dataset creados/puntos_venta_enriquecido2_sinNA.xlsx")

#Se crea un suma por provincia y fecha

ventas_prov<-ventas%>%
                left_join(puntos, by = "Affiliated_Code")%>%
                filter(!is.na(provincia))%>%
                group_by(provincia, Sales_DAY)%>%
                summarise(Unidades = sum(Sales_Uds, na.rm = TRUE), .groups = "drop")

# Se hace el dataframe tipo Wide

prov_wide<-ventas_prov%>%
              pivot_wider(names_from = Sales_DAY, values_from = Unidades, values_fill = 0)

#Se aplica logaritmos

log_mat<-log(prov_wide[,-1] + 1)

prov_wide<-data.frame(Provincias=prov_wide[[1]],log_mat)

#Se eliminan fechas (columnas) con varianza cero o NA

var_cols<-apply(prov_wide[,-1], 2, var, na.rm = TRUE)

cols_validas<-which(var_cols > 0 & is.finite(var_cols))

prov_wide<-data.frame(Provincias = prov_wide[[1]],
                        prov_wide[ ,cols_validas + 1])

#Se eliminan provincias (filas) con Inf, NaN o NA

rows_validas<-apply(prov_wide[,-1], 1, function(x) all(is.finite(x)))

prov_wide<-prov_wide[rows_validas, ]

#Se normaliza cada serie por su media si hay mucha dispersión

prov_wide_norm<-prov_wide

prov_wide_norm[,-1]<-t(apply(prov_wide[,-1], 1, function(x) x / mean(x, na.rm = TRUE)))

#Convergencia absoluta
H <- computeH(prov_wide_norm[,-1], quantity = "H")
estimate <- estimateMod(H, time_trim = 1/3, HACmethod = "FQSB")
```

```
print(round(estimate))

#   beta std.err  tvalue  pvalue
#####-1       0      -3       0

clubs_provincias<-findClubs(prov_wide_norm,
                                dataCol = 2:ncol(prov_wide_norm),
                                unit_names = 1,
                                refCol = ncol(prov_wide_norm),
                                time_trim = 1/3,
                                cstar = 0,
                                HACmethod = "FQSB")

# Resultados
summary(clubs_provincias)

#Sale un resultados de 11 cluves y 0 divergentes
#Number of convergence clubs: 11
#Number of divergent units: 0

| # of units  | beta      | std.err  | tvalue   | cstar
  --------- ------------- ---------- ---------- ---------- --------
#  club1   |   5         |   0.155  |   0.398  |   0.389  |   0
#club2   |   6         |  -0.63   |   0.411  |  -1.534  |   0
#club3   |   14        |  -0.54   |   0.348  |  -1.55   |   0
#club4   |   5         |  -0.477  |   0.383  |  -1.244  |   0
#club5   |   2         |  -0.849  |   0.664  |  -1.278  |   0
#club6   |   2         |  -0.465  |   0.593  |  -0.784  |   0
#club7   |   2         |  -0.123  |   0.779  |  -0.158  |   0
#club8   |   2         |  -1.059  |   0.77   |  -1.376  |   0
#club9   |   6         |  -0.598  |   0.409  |  -1.461  |   0
#club10  |   2         |  -0.875  |   0.747  |  -1.172  |   0
#club11  |   2         |  -0.447  |   0.902  |  -0.496  |   0


print(clubs_provincias)

#Existen 11 clubes que son provincias donde las ventas se comportan de manera homogenea y
#Se pueden llevar a cabo acciones diferenciadas para cada club, y no quedando unidades de
#provincias divergentes, que no se puedan agrupar.


#club 1 Baleares, Madrid, Zaragoza, Girona, Albacete
#club 2  Asturias, A Coruña, Zamora, Tarragona, Córdoba, Huesca
#club 3  Cádiz, Barcelona, Pontevedra, Alicante, La Rioja, Ourense, Valencia, Granada, Toledo,
Murcia, Ciudad Real,
#Huelva, Valladolid, Lugo
#club 4  Badajoz, Castellón, Málaga, Sevilla, Cáceres
#club 5  Ávila, León
#club 6  Cantabria, Álava
#club 7  Guipúzcoa, Vizcaya
#club 8  Lleida, Navarra
#club 9  Almería, Guadalajara, Cuenca, Salamanca, Jaén, Segovia
#club 10  Soria, Teruel
#club 11  Burgos, Palencia


# Dibujamos las transiciones de las marcas

plot(clubs_provincias)

#Para intentar ver mejor los graficos anteriores, quitamos la media de
#la transicion("average transition paths"):

plot(clubs_provincias,avgTP = FALSE , nrows = 3,ncols = 4, plot_args = list (type='l'))

#Ahora hacemos un grafico para mostrar un conjunto de gráficos (uno por club),
# en los que se enseñe cómo es el comportamiento promedio de ventas de tabaco a lo largo
#del tiempo en las provincias que pertenecen a ese club
```

```
plot(clubs_provincias,clubs=NULL,avgTP = TRUE , legend= TRUE,, plot_args = list (type='o'))

df_clubs_11 <- data.frame(
  provincia = c("Baleares", "Madrid", "Zaragoza", "Girona", "Albacete",
                "Asturias", "A Coruña", "Zamora", "Tarragona", "Córdoba", "Huesca",
                "Cádiz", "Barcelona", "Pontevedra", "Alicante", "La Rioja", "Ourense",
"Valencia","Granada","Toledo","Murcia","Ciudad Real","Huelva", "Valladolid","Lugo",
                "Badajoz", "Castellón", "Málaga", "Sevilla","Cáceres",
                "Ávila", "León",
                "Cantabria","Álava",
                "Guipúzcoa", "Vizcaya",
                "Lleida", "Navarra",
                "Almería", "Guadalajara","Cuenca", "Salamanca", "Jaén", "Segovia",
                "Soria", "Teruel",
                "Burgos", "Palencia"),
  club = c(1,1,1,1,1,
           2,2,2,2,2,2,
           3,3,3,3,3,3,3,3,3,3,3,3,3,3,
           4,4,4,4,4,
           5,5,
           6,6,
           7,7,
           8,8,
           9,9,9,9,9,9,
           10,10,
           11,11))
df_clubs_11$provincia <-
recode(df_clubs_11$provincia,"Valencia/València"="Valencia","Alicante/Alacant"="Alicante",
                              "A Coruña" = "La Coruña",
                              "Balears (Illes)" = "Baleares","Girona"="Gerona",
                              "Castellón/Castelló" = "Castellón","Rioja (La)"="La
Rioja","Ourense"="Orense","Lleida"="Lérida")
library(mapSpain)

provincias_sf<-esp_get_prov()

library(dplyr)

# Unir el mapa con tu tabla de clubes por nombre de provincia

provincias_clubs_11 <- provincias_sf %>%
  left_join(df_clubs_11, by = c("prov.shortname.es" = "provincia"))

ggplot(provincias_clubs_11) +
  geom_sf(aes(fill = factor(club)), color = "white") +
  labs(title = "Distribución geográfica de los clubes de provincias",
       fill = "Club") +
  theme_minimal() +
  theme(legend.position = "bottom")


#6º Vamos a intentar juntar los clubes similares para así reducir el numero de clubes
#lo hacemos con Mergeclusters usando Phillips and soul method (2009)

mergeclubprovincias<-mergeClubs(clubs_provincias,mergeMethod =
'PS',threshold=-1.65,mergeDivergent=FALSE)

summary(mergeclubprovincias)
#El resultado son 7 clubes y 0 divergentes
#Number of convergence clubs: 7
#Number of divergent units: 0

# merged clubs          | # of units  | beta     | std.err  | tvalue
  -------- ---------------------- ------------- ---------- ---------- ----------
#club1  | clubs: 1, 2, 3, 4, 5  |   32        |  -0.486  |   0.331  |  -1.47
#club2  | clubs: 6              |   2         |  -0.465  |   0.593  |  -0.784
#club3  | clubs: 7              |   2         |  -0.123  |   0.779  |  -0.158
#club4  | clubs: 8              |   2         |  -1.059  |   0.77   |  -1.376
```

```
#club5  | clubs: 9              |   6         |  -0.598  |   0.409  |  -1.461
#club6  | clubs: 10             |   2         |  -0.875  |   0.747  |  -1.172
#club7  | clubs: 11             |   2         |  -0.447  |   0.902  |  -0.496


print(mergeclubprovincias)
plot(mergeclubprovincias)
plot(mergeclubprovincias,clubs=NULL,avgTP = TRUE , legend= TRUE,, plot_args = list (type='o'))

#se consigue reducir a 7 clubes
#club 1  Baleares, Madrid, Zaragoza, Girona, Albacete, Asturias, A Coruña, Zamora, Tarragona,
Córdoba,
#Huesca, Cádiz, Barcelona, Pontevedra, Alicante, La Rioja, Ourense, Valencia, Granada, Toledo,
#Murcia, Ciudad Real, Huelva, Valladolid, Lugo, Badajoz, Castellón, Málaga, Sevilla, Cáceres,
#Ávila, León
#club 2 Cantabria, Álava
#club 3 Guipúzcoa, Vizcaya
#club 4 Lleida, Navarra
#club 5 Almería, Guadalajara, Cuenca, Salamanca, Jaén, Segovia
#club 6 Soria, Teruel
#club 7 Burgos, Palencia


#Apartado 4b)¿Existe algún patrón de comportamiento geográfico y/o relacionado con el ranking
#elaborado en el segundo apartado?

#4.1 Comportamiento Geográfico-->Visualizamos en un mapa de españa los clubes para poder ver si
existe
#algun patron norte sur este y oeste de colocacion de los clubes


df_clubs_7 <- data.frame(
  provincia = c("Baleares", "Madrid", "Zaragoza", "Girona", "Albacete", "Asturias", "A Coruña",
"Zamora", "Tarragona", "Córdoba",
                "Huesca", "Cádiz", "Barcelona", "Pontevedra", "Alicante", "La Rioja",
"Ourense", "Valencia","Granada","Toledo",
                "Murcia","Ciudad Real", "Huelva", "Valladolid", "Lugo", "Badajoz", "Castellón",
"Málaga", "Sevilla","Cáceres",
                "Ávila", "León",
                "Cantabria","Álava",
                "Guipúzcoa", "Vizcaya",
                "Lleida", "Navarra",
               "Almería", "Guadalajara","Cuenca", "Salamanca", "Jaén", "Segovia",
               "Soria", "Teruel",
               "Burgos", "Palencia"),
              club = c(1,1,1,1,1,1,1,1,1,1,
                       1,1,1,1,1,1,1,1,1,1,
                       1,1,1,1,1,1,1,1,1,1,
                       1,1,
                       2,2,
                       3,3,
                       4,4,
                       5,5,5,5,5,5,
                       6,6,
                       7,7))

df_clubs_7$provincia <-
recode(df_clubs_7$provincia,"Valencia/València"="Valencia","Alicante/Alacant"="Alicante",
                            "A Coruña" = "La Coruña",
                            "Balears (Illes)" = "Baleares","Girona"="Gerona",
                            "Castellón/Castelló" = "Castellón","Rioja (La)"="La
Rioja","Ourense"="Orense","Lleida"="Lérida")
library(mapSpain)

provincias_sf <- esp_get_prov()

library(dplyr)

# Unir el mapa con tu tabla de clubes por nombre de provincia
```

```
provincias_clubs_7<-provincias_sf %>%
                      left_join(df_clubs_7, by = c("prov.shortname.es" = "provincia"))

ggplot(provincias_clubs_7) +
            geom_sf(aes(fill = factor(club)), color = "white") +
            labs(title = "Distribución geográfica de los clubes de provincias",
                 fill = "Club") +
            theme_minimal() +
            theme(legend.position = "bottom")
```

Anexo B. Código Temporalidad de las ventas

Scrip R Temporalidad de las ventas

```
#leemos los tres datasets

library(readxl)
Product <- read_excel("C:/unir/MASTER/2 CUATRIMESTRE/TFM/empresa/Data set Red Proyectum -
Altadis/nuevos dataset creados/Product.xlsx")
View(Product)

library(readxl)
puntos_venta_enriquecido2_sinNA <- read_excel("C:/unir/MASTER/2 CUATRIMESTRE/TFM/empresa/Data
set Red Proyectum - Altadis/nuevos dataset creados/puntos_venta_enriquecido2_sinNA.xlsx")
View(puntos_venta_enriquecido2_sinNA)

library(readr)

SalesDay <- read_delim("C:/unir/MASTER/2 CUATRIMESTRE/TFM/empresa/Data set Red Proyectum -
Altadis/SalesDay.csv",delim = ";")
View(SalesDay)

library(dplyr)
library(ggplot2)
library(forecast)
library(caTools)
library(rpart)
library(rpart.plot)
library(caret)
library(tidyverse)
library(patchwork)
library(caret)
library(cluster)
library(tidyr)
library(ConvergenceClubs)
library(lubridate)
library(rlang)
#primero se comprueva si hay duplicados en product

any(duplicated(Product$Product_Code))



#como si que hay, se mira cuales.

Product$Product_Code[duplicated(Product$Product_Code)]



#solo hay un producto duplicado que es Natu122 con lo que se elimina se ha duplicado por el
formato es ATA y ASL
#se elimina y se crea un nuevo data set sin esa duplicidad quedandose con el primero que
aparece que es ATA ya que Product_Code
#debe ser clave primaria UNICA

Product_sin_duplicados <- Product %>%
  distinct(Product_Code, .keep_all = TRUE)



#ahora si podemos unir SalesDay con Product (por Product_Code)

ventas_productos <- SalesDay %>%
  left_join(Product_sin_duplicados, by = "Product_Code")

#se mantienen el numero de observaciones tras el join (2878182)

#Ahora se une con puntos_venta_enriquecido2_sinNA (por Affiliated_Code)
ventas_completo_2 <- ventas_productos %>%
  left_join(puntos_venta_enriquecido2_sinNA, by = "Affiliated_Code")

#se siguen mantiendo el numero de observaciones tras el join (2878182)
```

```
#ahora tenemos un dataset de 2878182obs y 14 variables se extrae como excel y
#se procede a su limpieza

#creamos un excell llamado para el archivo

library(readr)

write_csv(ventas_completo_2, "C:/unir/MASTER/2 CUATRIMESTRE/TFM/empresa/Data set Red Proyectum
- Altadis/ventas_completo_2.csv")


#ahora procedemos a trabajar con nuestro nuevo dataset ventas_completo


ventas_completo_2

dim(ventas_completo_2)
str(ventas_completo_2)
summary(ventas_completo_2)
unique(ventas_completo_2$Management_Cluster)
unique(ventas_completo_2$Product_Code)
unique(ventas_completo_2$Location)
unique(ventas_completo_2$SIZE)
unique(ventas_completo_2$Format)
unique(ventas_completo_2$Engage)
unique(ventas_completo_2$provincia)


is.na(ventas_completo_2)
which(is.na(ventas_completo_2))
colSums(is.na(ventas_completo_2))
colSums(is.na(puntos_venta_enriquecido2_sinNA))


#No hay celdas vacias

#se le dice a r que la variable Sales_DAY es fecha
ventas_completo_2<-ventas_completo_2%>%
            mutate(Sales_DAY = as.Date(as.character(Sales_DAY),
                                       format = "%Y%m%d"))
#Se comprueva y esta como fecha
str(ventas_completo_2)

#Se realiza una representación gráfica que permita conocer el comportamiento
#de la serie temporal de las ventas en cada una de las provincias.


#Se hacen los graficos y se observa que no los puedo comparar

ventas_completo_2 %>%
  ggplot(aes(x=Sales_DAY,y =Sales_Uds,color="Ventas"))+
  geom_line()+
  facet_wrap(~provincia)+
  scale_color_manual(values = c("#0098cd"))+
  labs(title = "Evolución diaria de ventas por provincia",
                          x = "Fecha", y = "Ventas diarias") +
  theme_minimal()

#Se escala para comparar (normalizacion 0-1) y se quitan na's:

#se calcula una columna de ventas totales en cada provincia

ventas_completo_2_escaladas <- ventas_completo_2 %>%
  drop_na()%>%
  group_by(provincia,Product_Code)%>%
  mutate (Ventas_Totales=sum(Sales_Uds))

#columna de cociente 0-1
```

```
ventas_completo_2_escaladas$Sales_Uds_1 <-
ventas_completo_2_escaladas$Sales_Uds/ventas_completo_2_escaladas$Ventas_Totales

# table para saber si todas las provincias estan presentes todas las veces

table(ventas_completo_2$provincia)

#como se puede ver no todas las provincias tienen ventas todos los días con lo
#que para Visualizar el comportamiento diario de las ventas (normalizadas entre 0-1)
#para cada provincia, y analizar tendencias o estacionalidad hay que crear un nuevo
#dataframe donde aparaezcan todos los dias aunq no haya ventas.

#hay que crear una tabla de fechas completas incluso para los días sin ventas y así evitar
#saltos de linea en el grafico

# Hacemos los vectores de fecha minima y maximas del dataset
fecha_min <- min(ventas_completo_2_escaladas$Sales_DAY, na.rm = TRUE)



fecha_max <- max(ventas_completo_2_escaladas$Sales_DAY, na.rm = TRUE)



# Crear secuencia de fechas para cada provincia, primero extraemos un vector donde
#ponga cada provincia distinta, vemos que son 48 y generamos todas las combinaciones posibles
entre ambas (provincias y fechas)
provincias <- unique(ventas_completo_2_escaladas$provincia)

fechas_provincias<-expand.grid(
  Sales_DAY = seq.Date(from = fecha_min, to = fecha_max, by = "day"),
  provincia = provincias)

#Se unen las tablas ventas_completo_2_escaladas y el nuevo dataframe de
fechas_provincias(combinaciones posibles entre provincias y fechas)
#y además se rellenan los días sin ventas poniendo CERO ventas y Así se mantienen todos los
días (aunque no haya ventas en todos):
#se observa que se pasa de 2878182 obs a 2878191 obs


ventas_diarias_completas<-fechas_provincias %>%
  left_join(ventas_completo_2_escaladas, by = c("Sales_DAY", "provincia")) %>%
  replace_na(list(Sales_Uds = 0, Sales_Uds_1 = 0))


#Ahora se resumen las ventas escaladas diarias por provincia


ventas_provincia_dia<- ventas_diarias_completas %>%
  group_by(provincia, Sales_DAY) %>%
  summarise(Ventas_Totales_Diarias = sum(Sales_Uds_1, na.rm = TRUE), .groups = "drop")

#hago un table para saber si todas las provincias estan presentes todas las veces y se observa
que si (210)

table(ventas_provincia_dia$provincia)

#4º)Visualizar la serie temporal



g0<-ggplot(ventas_provincia_dia, aes(x = Sales_DAY, y = Ventas_Totales_Diarias)) +
  geom_line(color = "#0098cd") +
  facet_wrap(~ provincia, scales = "free_y") +
  labs(title = "Evolución diaria de ventas normalizadas por provincia",
       x = "Fecha", y = "Ventas normalizadas (0-1)") +
  theme_minimal()

g0
```

```
g1<-ggplot(ventas_provincia_dia, aes(x = Sales_DAY, y = Ventas_Totales_Diarias)) +
  geom_line(color = "#0098cd") +
  geom_smooth(method = "loess", se = FALSE, color = "red") +
  facet_wrap(~provincia, scales = "free_y") +
  labs(title = "Tendencia y estacionalidad por provincia", x = "Fecha", y = "Ventas
normalizadas") +
  theme_minimal()

g1

g0+g1

#Los graficos ya son comparables y hemos obtenido el comportamiento diario por provincia de las
ventas
#(aunque no haya ventas todos los días, se visualizarán como cero).


#ahora se estudian las ventas por día de la semana y mes (normalizado para que tengamos tambien
la vision de octubre)

ventas_por_dia_de_semana<-ventas_completo_2 %>%
  mutate(dia_semana = wday(Sales_DAY, label = TRUE)) %>%
  group_by(dia_semana) %>%
  summarise(Ventas = sum(Sales_Uds))

g2<-ggplot(ventas_por_dia_de_semana, aes(x = dia_semana, y = Ventas)) +
  geom_col(fill = "#0098cd") +
  labs(title = "Ventas por día de la semana", x = "Día", y = "Unidades vendidas") +
  theme_minimal()

g2


ventas_diarias_mes_normalizado <- ventas_completo_2 %>%
  mutate(mes = month(Sales_DAY, label = TRUE)) %>%
  group_by(mes, Sales_DAY) %>%
  summarise(Ventas_dia = sum(Sales_Uds), .groups = "drop") %>%
  group_by(mes) %>%
  summarise(Media_diaria = mean(Ventas_dia), .groups = "drop")

g2_normalizado <- ggplot(ventas_diarias_mes_normalizado, aes(x = mes, y = Media_diaria)) +
  geom_col(fill = "#0098cd") +
  labs(title = "Media diaria de ventas por mes", x = "Mes", y = "Media diaria de ventas") +
  theme_minimal()

g2_normalizado
g2+g2_normalizado

#Se va a generar un dataset que contenga dia de la semana ccaa y festividades para graficar y
conocer comportamiento
# Se crea un vector de festivos nacionales en el 2015

festivos_nacionales<-as.Date(c("2015-01-01", "2015-01-06", "2015-04-03", "2015-05-01",
                               "2015-08-15", "2015-10-12", "2015-11-01","2015-12-06", "2015-12-
08", "2015-12-25"))

# generamos una lista de festivos autonomicos en el 2015 extraidos de
# https://www.ccoo-servicios.es/archivos/tecnocom/20150120_Calendario_Laboral_2015.pdf

festivos_ccaa<-list("Andalucía" = as.Date(c("2015-02-28","2015-04-02","2015-11-02","2015-12-
07"),
  "Aragón" = as.Date(c("2015-04-23","2015-04-02","2015-11-02","2015-12-07")),
  "Asturias" = as.Date(c("2015-09-08","2015-04-02","2015-11-02","2015-12-07")),
  "Islas Baleares" = as.Date(c("2015-04-06","2015-04-02","2015-11-02","2015-12-07")),
  "Canarias" = as.Date(c("2015-05-30","2015-04-02","2015-11-02")),
  "Cantabria" = as.Date(c("2015-04-02","2015-11-02","2015-04-06","2015-09-15")),
  "Castilla-La Mancha" = as.Date(c("2015-04-02","2015-04-06","2015-06-04","2015-12-07")),
  "Castilla y León" = as.Date(c("2015-04-23","2015-04-02",,"2015-12-07","2015-11-02")),
```

```
  "Cataluña" = as.Date(c("2015-04-06", "2015-06-24", "2015-09-11","2015-12-26")),
  "Comunidad Valenciana" = as.Date(c("2015-03-19", "2015-10-09","2015-04-06","2015-12-07")),
  "Extremadura" = as.Date(c("2015-09-08","2015-04-02","2015-11-02","2015-12-07")),
  "Galicia" = as.Date(c("2015-07-25", "2015-04-02","2015-03-20","2015-11-02")),
  "Madrid" = as.Date(c("2015-03-19", "2015-04-02", "2015-05-02","2015-06-04")),
  "Murcia" = as.Date(c("2015-06-09","2015-03-19", "2015-04-02","2015-12-07")),
  "Navarra" = as.Date(c("2015-03-19","2015-04-02","2015-06-04","2015-07-25")),
  "País Vasco" = as.Date(c("2015-03-19","2015-04-02","2015-06-04","2015-07-25")),
  "La Rioja" = as.Date(c("2015-06-09","2015-04-02","2015-06-04","2015-12-07")),
  "Ceuta" = as.Date(c("2015-09-25","2015-12-07","2015-04-02")),
  "Melilla" = as.Date(c("2015-09-25","2015-12-07","2015-03-19","2015-04-02"))))

# vector de provincias a que ccaa pertenece

provincia_a_ccaa<-c("Álava" = "País Vasco","Albacete" = "Castilla-La Mancha","Alicante" =
"Comunidad Valenciana",
                    "Almería" = "Andalucía","Asturias" = "Asturias","Ávila" = "Castilla y
León","Badajoz" = "Extremadura",
                    "Barcelona" = "Cataluña","Burgos" = "Castilla y León","Cáceres" =
"Extremadura","Cádiz" = "Andalucía","Cantabria" = "Cantabria",
                      "Castellón" = "Comunidad Valenciana","Ciudad Real" = "Castilla-La
Mancha","Córdoba" = "Andalucía",
                    "Cuenca" = "Castilla-La Mancha","Girona" = "Cataluña","Granada" =
"Andalucía","Guadalajara" = "Castilla-La Mancha",
                    "Guipúzcoa" = "País Vasco","Huelva" = "Andalucía","Huesca" = "Aragón","Illes
Balears" = "Islas Baleares","Jaén" = "Andalucía",
                      "A Coruña" = "Galicia","La Rioja" = "La Rioja","Las Palmas" =
"Canarias","León" = "Castilla y León","Lleida" = "Cataluña","Lugo" = "Galicia",
                     "Madrid" = "Madrid","Málaga" = "Andalucía","Murcia" = "Región de
Murcia","Navarra" = "Navarra",
                     "Ourense" = "Galicia","Palencia" = "Castilla y León","Pontevedra" =
"Galicia","Salamanca" = "Castilla y León",
                      "Santa Cruz de Tenerife" = "Canarias","Segovia" = "Castilla y
León","Sevilla" = "Andalucía",
                      "Soria" = "Castilla y León","Tarragona" = "Cataluña","Teruel" =
"Aragón","Toledo" = "Castilla-La Mancha","Valencia" = "Comunidad Valenciana",
                      "Valladolid" = "Castilla y León","Vizcaya" = "País Vasco","Zamora" =
"Castilla y León","Zaragoza" = "Aragón","Ceuta" = "Ceuta","Melilla" = "Melilla")

#añadir al dataset el día de la semana y el tipo de día
#(laborable, festivo o fin de semana), utilizando festivos nacionales y autonómicos

ventas_completo_2_festivos<-ventas_completo_2 %>%
  mutate(dia_semana = wday(Sales_DAY, label = TRUE),
         ccaa = recode(provincia, !!!provincia_a_ccaa),
         festivo_especifico = case_when(Sales_DAY %in% festivos_nacionales ~ TRUE,
                                                        !is.na(ccaa) & map2_lgl(ccaa,
Sales_DAY,
                                                        ~ .y %in% festivos_ccaa[[.x]]) ~ TRUE,
                                                        TRUE ~ FALSE),
         tipo_dia = case_when(festivo_especifico ~ "Festivo",
           dia_semana %in% c("sá\\.", "do\\.") ~ "Fin de semana",
           TRUE ~ "Laborable"))

unique(ventas_completo_2_festivos$tipo_dia)

ventas_completo_2_festivos %>% count(tipo_dia)

ventas_completo_2_festivos %>% count(dia_semana)

ventas_completo_2_festivos %>%
  count(dia_semana) %>%
  arrange(dia_semana)

str(ventas_completo_2$Sales_DAY)

#Ver si se vende más en laborables, fines de semana o festivos.

g3<-ventas_completo_2_festivos %>%
```

```
  group_by(tipo_dia) %>%
  summarise(Ventas = sum(Sales_Uds)) %>%
  ggplot(aes(x = tipo_dia, y = Ventas, fill = tipo_dia)) +
  geom_col() +
  labs(title = "Ventas totales según tipo de día", x = "", y = "Unidades vendidas") +
  scale_fill_manual(values = c("#0072B2", "#E69F00", "#D55E00")) +
  theme_minimal()

g3

#Ver la evolución temporal con una línea por tipo de día.

g4<-ventas_completo_2_festivos %>%
  group_by(Sales_DAY, tipo_dia) %>%
  summarise(Ventas = sum(Sales_Uds), .groups = "drop") %>%
  ggplot(aes(x = Sales_DAY, y = Ventas, color = tipo_dia)) +
  geom_line() +
  labs(title = "Evolución de ventas por tipo de día", x = "Fecha", y = "Ventas") +
  theme_minimal()

g4

#Cruzar los días (lun-dom) con el tipo de día para detectar patrones.

g5<-ventas_completo_2_festivos %>%
  group_by(dia_semana, tipo_dia) %>%
  summarise(Ventas = sum(Sales_Uds), .groups = "drop") %>%
  ggplot(aes(x = dia_semana, y = Ventas, fill = tipo_dia)) +
  geom_col(position = "stack") +
  labs(title = "Ventas por día de la semana y tipo de día", x = "Día", y = "Unidades vendidas")
+
  theme_minimal()

g5

g3+g4+g5

#Ahora se van a dibujar diferentes graficos para observar el comportamiento de las ventas
generales


#se calculan las ventas por producto y se dibujan

ventas_productos <- ventas_completo_2 %>%
  group_by(Product_Code, SIZE, Format) %>%
  summarise(Unidades = sum(Sales_Uds), .groups = "drop") %>%
  arrange(desc(Unidades))

g6<-ggplot (ventas_productos)+
  geom_col(aes(x= Unidades, y= reorder(Product_Code, Unidades)),fill="#0098cd")+
  labs(title = "Ranking ventas por Producto", x=" Unidades Vendidas", y= "Producto")+
  theme_minimal()

g6

#se calculan las ventas totales por provincia y se dibujan , se trata de un ranking total
#por provincia (suma de las ventas en todos los puntos de venta de cada provincia)

ventas_totales_provincia<-ventas_completo_2 %>%
  group_by(provincia) %>%
  summarise(Ventas_Totales = sum(Sales_Uds), .groups = "drop") %>%
  arrange(desc(Ventas_Totales))

g7<-ggplot(ventas_totales_provincia) +
                geom_col(aes(x = Ventas_Totales, y = reorder(provincia, Ventas_Totales)), fill
= "#0098cd") +
                labs(title = "Ranking ventas totales por provincia",
                     x = "Ventas Totales",
                     y = "Provincias") +
```

```
                theme_minimal()
g7

# Ahora se muestran los Top 20 puntos de venta por ventas totales


# Top 20 puntos de venta con más ventas

top_puntos_venta <- ventas_completo_2 %>%
  group_by(Affiliated_Code, Affiliated_NAME, provincia) %>%
  summarise(Ventas_Totales = sum(Sales_Uds), .groups = "drop") %>%
  arrange(desc(Ventas_Totales)) %>%
  slice_max(order_by = Ventas_Totales, n = 30)

g8<-ggplot(top_puntos_venta) +
                geom_col(aes(x = Ventas_Totales, y = reorder(Affiliated_NAME, Ventas_Totales),
fill = provincia)) +
                geom_text (aes(x = Ventas_Totales, y = reorder(Affiliated_NAME,
Ventas_Totales), label = round (Ventas_Totales, 0)),
                           hjust = 1, size = 4)+
                labs(title = "Top 20 puntos de venta por ventas totales",
                     x = "Ventas Totales",
                     y = "Punto de Venta") +
                theme_minimal()

g8

g6+g7+g8


#Modelos Propuestos: ARIMA y ETS
#Se hace una predicción a 3 meses para las 4 provincias en las que más se vende
#como se vio anteriormente


library(forecast)

#Se hace un dataset de ventas diarias totales que se usara para despues discriminar las
#filas de cada provincia

ventas_provincia_dia_totales<- ventas_completo_2%>%
  group_by(provincia, Sales_DAY)%>%
  summarise(Ventas_Totales_Diarias = sum(Sales_Uds), .groups = "drop")

#Se Predice Madrid

#se separan las filas de madrid

madrid<-ventas_provincia_dia_totales%>%
  filter(provincia == "Madrid") %>%
  arrange(Sales_DAY)


#Se le da formato de serie temporal y graficamos para verlo mejor, no se puede
#descomponer debido a que no tenemos un año completo

serie_diaria_madrid<-ts(madrid$Ventas_Totales_Diarias,
                           start = c(2015, yday(min(madrid$Sales_DAY))),
                           frequency = 365)

plot(serie_diaria_madrid,main = "Serie temporal diaria de ventas en Madrid (2015)")


#Se hace la prediccion diaria modelo ARIMA

modelo_diario_madrid<- auto.arima(serie_diaria_madrid, seasonal = TRUE)

prediccion_diaria_madrid<-forecast(modelo_diario_madrid,h=90)
```

```
plot(prediccion_diaria_madrid,
     main = "Predicción de ventas diarias en Madrid (90 días)",
     ylab = "Ventas diarias",
     xlab = "Días futuros")

#Se calcula como de buena es la prediccion
accuracy(modelo_diario_madrid)

#MAPE >50% indica un modelo con una precision mala

#El modelo ARIMA no encuentra patrones estacionales ni tendencias estables en la serie diaria
de ventas.

#Esto puede deberse a:
#Fuerte ruido diario en los datos
#Falta de ciclos estacionales detectables (por ejemplo, no se repiten patrones semanales o
mensuales).
#Período temporal muy corto (menos de un año) y sin años anteriores para detectar
estacionalidad real.

#se comprueba el modelo arima en periodo semanal

madrid_semanal <- madrid %>%
  mutate(semana = floor_date(Sales_DAY, "week")) %>%
  group_by(semana) %>%
  summarise(Ventas_Semanales = sum(Ventas_Totales_Diarias), .groups = "drop")

serie_semanal_madrid <- ts(madrid_semanal$Ventas_Semanales, start = c(2015, 1), frequency = 52)
plot(serie_semanal_madrid,main = "Serie temporal semanal de ventas en Madrid (2015)")

#Se hace prediccion semanal modelo ARIMA

modelo_semanal_madrid<- auto.arima(serie_semanal_madrid, seasonal = TRUE)

prediccion_semanal_madrid<-forecast(modelo_semanal_madrid,h=12)

plot(prediccion_semanal_madrid,
     main = "Predicción de ventas semanal en Madrid (12 semanas)",
     ylab = "Ventas diarias",
     xlab = "Semanas futuras")

#Se calcula como de buena es la prediccion

accuracy(modelo_semanal_madrid)

#ME     RMSE      MAE       MPE     MAPE MASE      ACF1
#Training set 328.8017 6598.653 4190.321 -55.41987 69.85865  NaN 0.1410504

#A nivel semanal, el modelo capta algo de estructura, pero la variabilidad sigue siendo
elevada. La predicción mejora visualmente y
#es más estable que la diaria, pero debe tomarse con cautela por el alto error relativo (MAPE).

#Se prueva con el modelo ETS que funciona bien con ruido y datos que fluctuan mucho

modelo_ets_madrid<-ets(serie_diaria_madrid)

prediccion_ets_madrid<-forecast(modelo_ets_madrid, h = 90)

plot(prediccion_ets_madrid,
     main = "Predicción ETS (Madrid - 90 días)",
     ylab = "Ventas diarias",
     xlab = "Días futuros")

accuracy(modelo_ets_madrid)

#MAPE muy alto lo que dice que el modelo tiene una mala precision predictiva

#Se prediceBarcelona
```

```
#Filas de Barcelona

Barcelona<-ventas_provincia_dia_totales%>%
  filter(provincia == "Barcelona") %>%
  arrange(Sales_DAY)


#Formato de serie temporal y se grafica para verlo mejor, no se puede
#descomponer debido a que no tenemos un año completo

serie_diaria_Barcelona<-ts(Barcelona$Ventas_Totales_Diarias,
                        start = c(2015, yday(min(Barcelona$Sales_DAY))),
                        frequency = 365)

plot(serie_diaria_Barcelona,main = "Serie temporal diaria de ventas en Barcelona (2015)")


#Hacemos la prediccion diaria modelo ARIMA

modelo_diario_Barcelona<- auto.arima(serie_diaria_Barcelona, seasonal = TRUE)

prediccion_diaria_Barcelona<-forecast(modelo_diario_Barcelona,h=90)

plot(prediccion_diaria_Barcelona,
     main = "Predicción de ventas diarias en Barcelona (90 días)",
     ylab = "Ventas diarias",
     xlab = "Días futuros")

#Calculamos como de buena es la prediccion

accuracy(modelo_diario_Barcelona)

#MAPE >50% indica un modelo con una precision mala

#El modelo ARIMA no encuentra patrones estacionales ni tendencias estables en la serie diaria
de ventas.

#Esto puede deberse a:
#Fuerte ruido diario en los datos
#Falta de ciclos estacionales detectables (por ejemplo, no se repiten patrones semanales o
mensuales).
#Período temporal muy corto (menos de un año) y sin años anteriores para detectar
estacionalidad real.

#se comprueba el modelo arima en periodo semanal

Barcelona_semanal<-Barcelona%>%
  mutate(semana = floor_date(Sales_DAY, "week")) %>%
  group_by(semana) %>%
  summarise(Ventas_Semanales = sum(Ventas_Totales_Diarias), .groups = "drop")

serie_semanal_Barcelona<-ts(Barcelona_semanal$Ventas_Semanales, start = c(2015, 1), frequency =
52)

plot(serie_semanal_Barcelona,main = "Serie temporal semanal de ventas en Barcelona (2015)")

#Hacemos la prediccion semanal modelo ARIMA

modelo_semanal_Barcelona<-auto.arima(serie_semanal_Barcelona, seasonal = TRUE)

prediccion_semanal_Barcelona<-forecast(modelo_semanal_Barcelona,h=12)

plot(prediccion_semanal_Barcelona,
     main = "Predicción de ventas semanal en Barcelona(12 semanas)",
     ylab = "Ventas diarias",
     xlab = "Semanas futuras")

#Calculamos como de buena es la prediccion
```

```
accuracy(modelo_semanal_Barcelona)

#ME     RMSE      MAE       MPE     MAPE MASE       ACF1
#Training set 165.4978 4784.412 2787.293 -88.44689 99.84792  NaN 0.07760957

#A nivel semanal, el modelo resulta no valido.

#Se prueva con el modelo ETS que funciona bien con ruido y datos que fluctuan mucho

modelo_ets_Barcelona<-ets(serie_diaria_Barcelona)
prediccion_ets_Barcelona<- forecast(modelo_ets_Barcelona, h = 90)

plot(prediccion_ets_Barcelona,
     main = "Predicción ETS (Barcelona - 90 días)",
     ylab = "Ventas diarias",
     xlab = "Días futuros")
accuracy(modelo_ets_Barcelona)

#MAPE muy alto lo que dice que el modelo no tiene precision predictiva


#Predecimos Valencia

#Filas de Valencia

Valencia<-ventas_provincia_dia_totales%>%
  filter(provincia == "Valencia") %>%
  arrange(Sales_DAY)


#Le damos formato de serie temporal y graficamos para verlo mejor, no se puede
#descomponer debido a que no tenemos un año completo

serie_diaria_Valencia<-ts(Valencia$Ventas_Totales_Diarias,
                          start = c(2015, yday(min(Valencia$Sales_DAY))),
                          frequency = 365)

plot(serie_diaria_Valencia,main = "Serie temporal diaria de ventas en Valencia (2015)")


#Hacemos la prediccion diaria modelo ARIMA

modelo_diario_Valencia<-auto.arima(serie_diaria_Valencia, seasonal = TRUE)

prediccion_diaria_Valencia<-forecast(modelo_diario_Valencia,h=90)

plot(prediccion_diaria_Valencia,
     main = "Predicción de ventas diarias en Valencia (90 días)",
     ylab = "Ventas diarias",
     xlab = "Días futuros")

#Calculamos como de buena es la prediccion

accuracy(modelo_diario_Valencia)

#MAPE tan alto indica un modelo con una precision mala

#El modelo ARIMA no encuentra patrones estacionales ni tendencias estables en la serie diaria
de ventas.

#Esto puede deberse a:

#Fuerte ruido diario en los datos
#Falta de ciclos estacionales detectables (por ejemplo, no se repiten patrones semanales o
mensuales).
#Período temporal muy corto (menos de un año) y sin años anteriores para detectar
estacionalidad real.

#se comprueba el modelo arima en periodo semanal
```

```
Valencia_semanal<-Valencia%>%
  mutate(semana = floor_date(Sales_DAY, "week")) %>%
  group_by(semana) %>%
  summarise(Ventas_Semanales = sum(Ventas_Totales_Diarias), .groups = "drop")

serie_semanal_Valencia<-ts(Valencia_semanal$Ventas_Semanales, start = c(2015, 1), frequency =
52)
plot(serie_semanal_Valencia,main = "Serie temporal semanal de ventas en Valencia (2015)")

#Hacemos la prediccion semanal modelo ARIMA

modelo_semanal_Valencia<-auto.arima(serie_semanal_Valencia, seasonal = TRUE)

prediccion_semanal_Valencia<-forecast(modelo_semanal_Valencia,h=12)

plot(prediccion_semanal_Valencia,
     main = "Predicción de ventas semanal en Valencia(12 semanas)",
     ylab = "Ventas diarias",
     xlab = "Semanas futuras")

#Calculamos como de buena es la prediccion

accuracy(modelo_semanal_Valencia)

#                     ME     RMSE      MAE       MPE     MAPE MASE       ACF1
#Training set -788.9627 3229.224 1387.007 -122.5167 126.3663  NaN 0.05397777


#Se prueva con el modelo ETS que funciona bien con ruido y datos que fluctuan mucho

modelo_ets_Valencia<-ets(serie_diaria_Valencia)
prediccion_ets_Valencia<- forecast(modelo_ets_Valencia, h = 90)

plot(prediccion_ets_Valencia,
     main = "Predicción ETS (Valencia - 90 días)",
     ylab = "Ventas diarias",
     xlab = "Días futuros")
accuracy(modelo_ets_Valencia)

#MAPE muy alto lo que dice que el modelo no tiene precision predictiva

#Predecimos Alicante

#Separamos las filas de Alicante

Alicante<-ventas_provincia_dia_totales%>%
  filter(provincia == "Alicante") %>%
  arrange(Sales_DAY)


#Le damos formato de serie temporal y graficamos para verlo mejor, no se puede
#descomponer debido a que no tenemos un año completo

serie_diaria_Alicante<-ts(Alicante$Ventas_Totales_Diarias,
                          start = c(2015, yday(min(Alicante$Sales_DAY))),
                          frequency = 365)

plot(serie_diaria_Alicante,main = "Serie temporal diaria de ventas en Alicante (2015)")


#Hacemos la prediccion diaria modelo ARIMA

modelo_diario_Alicante<- auto.arima(serie_diaria_Alicante, seasonal = TRUE)

prediccion_diaria_Alicante<-forecast(modelo_diario_Alicante,h=90)

plot(prediccion_diaria_Alicante,
     main = "Predicción de ventas diarias en Alicante (90 días)",
```

```
     ylab = "Ventas diarias",
     xlab = "Días futuros")

#Calculamos como de buena es la prediccion
accuracy(modelo_diario_Alicante)

#MAPE del 57.51% indica un modelo con una precision mala aunque es el mejor de todas las
provincias

#El modelo ARIMA no encuentra patrones estacionales ni tendencias estables en la serie diaria
de ventas.

#Esto puede deberse a:

#Fuerte ruido diario en los datos
#Falta de ciclos estacionales detectables (por ejemplo, no se repiten patrones semanales o
mensuales).
#Período temporal muy corto (menos de un año) y sin años anteriores para detectar
estacionalidad real.

#se comprueba el modelo arima en periodo semanal

Alicante_semanal<-Alicante%>%
  mutate(semana = floor_date(Sales_DAY, "week")) %>%
  group_by(semana) %>%
  summarise(Ventas_Semanales = sum(Ventas_Totales_Diarias), .groups = "drop")

serie_semanal_Alicante<-ts(Alicante_semanal$Ventas_Semanales, start = c(2015, 1), frequency =
52)
plot(serie_semanal_Alicante,main = "Serie temporal semanal de ventas en Alicante (2015)")

#Hacemos la prediccion semanal modelo ARIMA

modelo_semanal_Alicante<-auto.arima(serie_semanal_Alicante, seasonal = TRUE)

prediccion_semanal_Alicante<-forecast(modelo_semanal_Alicante,h=12)

plot(prediccion_semanal_Alicante,
     main = "Predicción de ventas semanal en Alicante(12 semanas)",
     ylab = "Ventas diarias",
     xlab = "Semanas futuras")

#Calculamos como de buena es la prediccion

accuracy(modelo_semanal_Alicante)

#                     ME     RMSE      MAE       MPE     MAPE MASE       ACF1
#Training set 37.53617 1467.122 629.2236 -81.26976 89.03175  NaN 0.02041297


#Se prueva con el modelo ETS que funciona bien con ruido y datos que fluctuan mucho

modelo_ets_Alicante<-ets(serie_diaria_Alicante)
prediccion_ets_Alicante<- forecast(modelo_ets_Alicante, h = 90)

plot(prediccion_ets_Alicante,
     main = "Predicción ETS (Alicante - 90 días)",
     ylab = "Ventas diarias",
     xlab = "Días futuros")
accuracy(modelo_ets_Alicante)

#MAPE muy alto lo que dice que el modelo tiene precision predictiva mala

#TODO EL PAIS

# Se agregan las ventas totales diarias de todo el país

ventas_diarias_total_pais<-ventas_completo_2 %>%
                              group_by(Sales_DAY) %>%
```

```
                              summarise(Ventas_Totales = sum(Sales_Uds), .groups = "drop") %>%
                              arrange(Sales_DAY)

# Se agregan las ventas totales semanales de todo el país

ventas_semanales_total_pais<-ventas_completo_2 %>%
                              mutate(semana = floor_date(Sales_DAY, "week"))%>%
                              group_by(semana)%>%
                              summarise(Ventas_Semanales = sum(Sales_Uds), .groups = "drop")%>%
                              arrange(semana)

#Se genera la serie temporal diaria

serie_total_pais<-ts(ventas_diarias_total_pais$Ventas_Totales,
                       start = c(2015, yday(min(ventas_diarias_total_pais$Sales_DAY))),
                       frequency = 365)

#Se genera la serie temporal semanal
serie_semanal_total_pais <- ts(ventas_semanales_total_pais$Ventas_Semanales,
                               start = c(2015, 1),
                               frequency = 52)

#Aplicamos los Modelos ARIMA y ETS

modelo_arima_pais<-auto.arima(serie_total_pais, seasonal = TRUE)
modelo_arima_semanal_pais<-auto.arima(serie_semanal_total_pais, seasonal = TRUE)
modelo_ets_pais<-ets(serie_total_pais)


#Predicciones

pred_arima_pais<-forecast(modelo_arima_pais, h = 90)
pred_arima_semanal_pais<-forecast(modelo_arima_semanal_pais, h = 12)
pred_ets_pais<-forecast(modelo_ets_pais, h = 90)

#Precision de los modelos

accuracy(modelo_arima_pais)
accuracy(modelo_arima_semanal_pais)
accuracy(modelo_ets_pais)

#Visualización de las prediciones
plot(pred_arima_pais, main = "Predicción ARIMA - Ventas totales (90 días)")
plot(pred_arima_semanal_pais, main = "Predicción ARIMA - Ventas semanales totales (12
semanas)")
plot(pred_ets_pais, main = "Predicción ETS - Ventas totales (90 días)")
```

Anexo C. Proceso de cálculo del número óptimo de clústeres en la segmentación de puntos de venta por rendimiento comercial.

En el siguiente fragmento de código se detalla el método que se desarrolló para calcular el número óptimo de clústeres (k) en la segmentación de puntos de venta por rendimiento comercial. Para ello, se aplicó el algoritmo K-Means sobre el conjunto de variables que incluía la ubicación de la tienda (Location), su tamaño en metros cuadrados (Tam_m2), el nivel de implicación comercial (Engage) y las ventas totales (Total_Ventas). Primero se transformaron las variables categóricas con la expresión OneHotEncoder y luego se ajustaron a una misma escala utilizando StandardScaler.

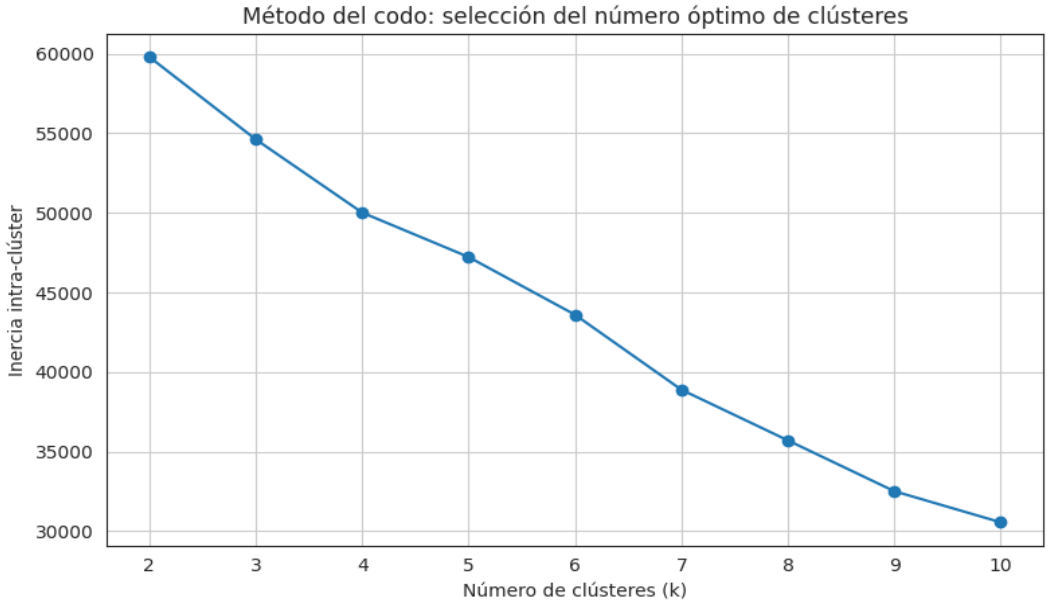
Para encontrar el valor de k óptimo, se calcularon varias métricas de validación interna para un rango de valores de k comprendido entre 2 y 10. De esta forma, se pudo evaluar qué valor de k ofrecía una mejor segmentación:

- Índice de Silhouette: evalúa la cohesión interna de los grupos formados.
- Índice de Davies-Bouldin: mide cuánto se solapan los clústeres.
- Índice de Calinski-Harabasz: valora la separación y compacidad entre los grupos.

A continuación, se presenta el código completo utilizado junto a los resultados obtenidos para cada métrica.

```python
# Carga de librerías
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score, davies_bouldin_score, calinski_harabasz_score


# Unión del dataset con las columnas seleccionadas para el clustering de tiendas con las ventas
clustering_df = tiendas_df[['Affiliated_Code', 'Location', 'Tam_m2', 'Engage']].copy()
ventas_totales = ventas_df.groupby('Affiliated_Code')['Sales_Uds'].sum().reset_index()
ventas_totales.rename(columns={'Sales_Uds': 'Total_Ventas'}, inplace=True)
clustering_df = pd.merge(clustering_df, ventas_totales, on='Affiliated_Code', how='inner')

# Codificación y escalado
clustering_encoded = pd.get_dummies(clustering_df.drop(columns='Affiliated_Code'), columns=['Location', 'Tam_m2'])
scaler = StandardScaler()
X_scaled = scaler.fit_transform(clustering_encoded)

# Cálculo de inercia para la selección del número óptimo de k
inertia = []
silhouette_scores = []
davies_scores = []
calinski_scores = []
k_range = range(2, 11)

for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42, n_init=10)
    labels = kmeans.fit_predict(X_scaled)

    inertia.append(kmeans.inertia_)
    silhouette_scores.append(silhouette_score(X_scaled, labels))
    davies_scores.append(davies_bouldin_score(X_scaled, labels))
    calinski_scores.append(calinski_harabasz_score(X_scaled, labels))
```

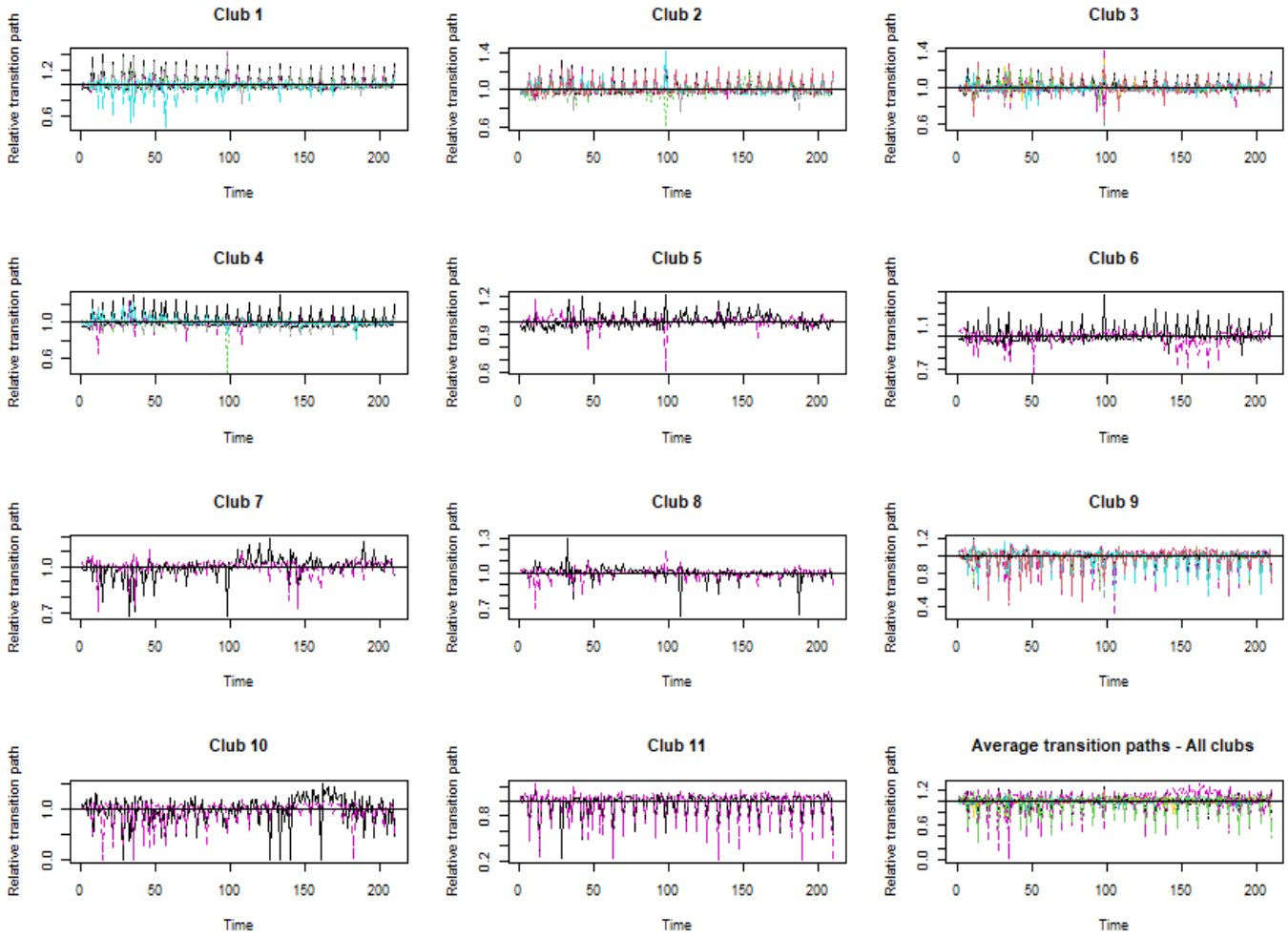
```python
# Gráfico del Método del codo (Elbow)
plt.figure(figsize=(8, 5))
plt.plot(k_range, inertia, marker='o')
plt.title('Método del codo: selección del número óptimo de clústeres')
plt.xlabel('Número de clústeres (k)')
plt.ylabel('Inercia intra-clúster')
plt.grid(True)
plt.tight_layout()
plt.show()

# Comparativa de métricas
metricas_df = pd.DataFrame({
    'k': list(k_range),
    'Silhouette': silhouette_scores,
    'Davies-Bouldin': davies_scores,
    'Calinski-Harabasz': calinski_scores
})

metricas_df.round(3)
```



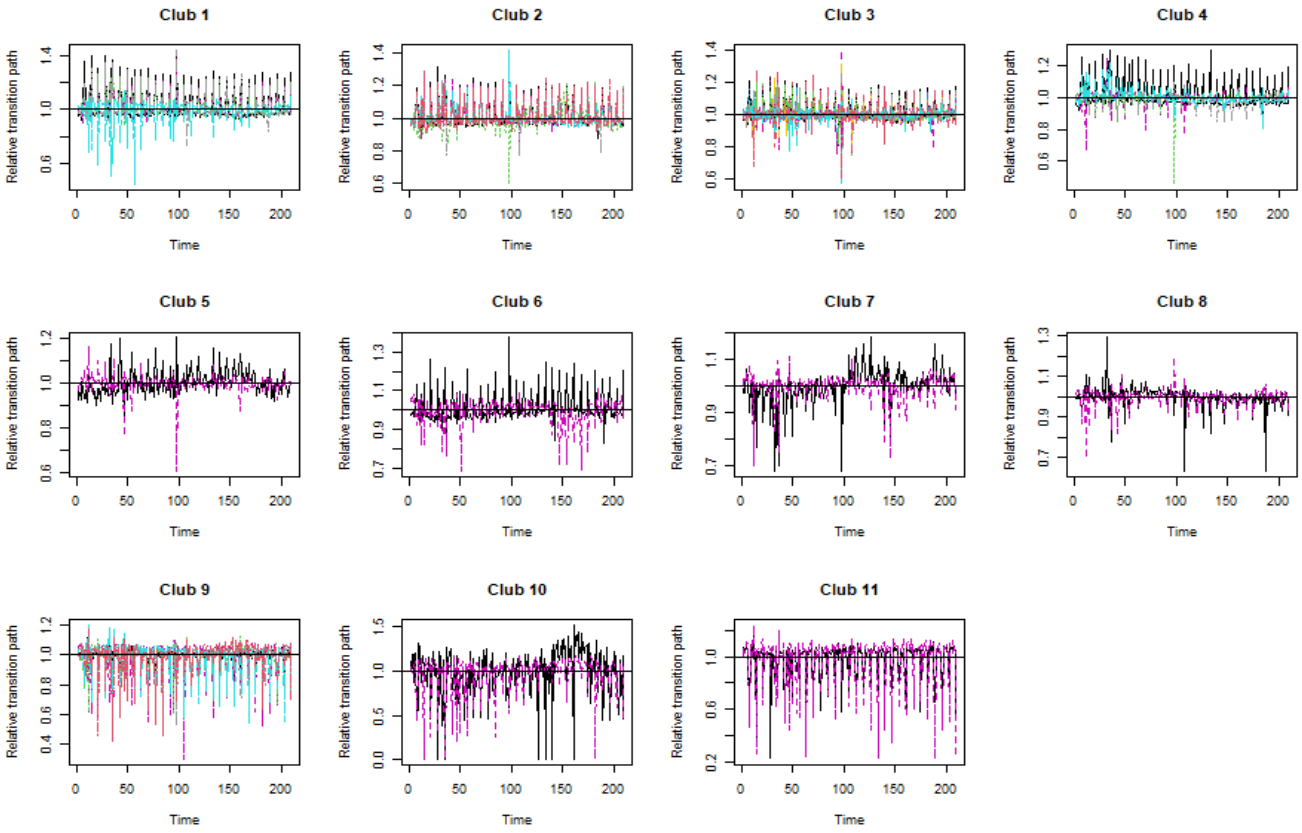
	k	Silhouette	Davies-Bouldin	Calinski-Harabasz
0	2	0.096	3.368	281.538
1	3	0.121	2.629	323.708
2	4	0.178	2.332	345.265
3	5	0.197	2.060	327.031
4	6	0.213	1.724	342.738
5	7	0.263	1.631	392.316
6	8	0.274	1.476	411.647
7	9	0.300	1.341	439.272
8	10	0.302	1.375	440.728

Anexo D. Gráficos Clubes de Convergencia

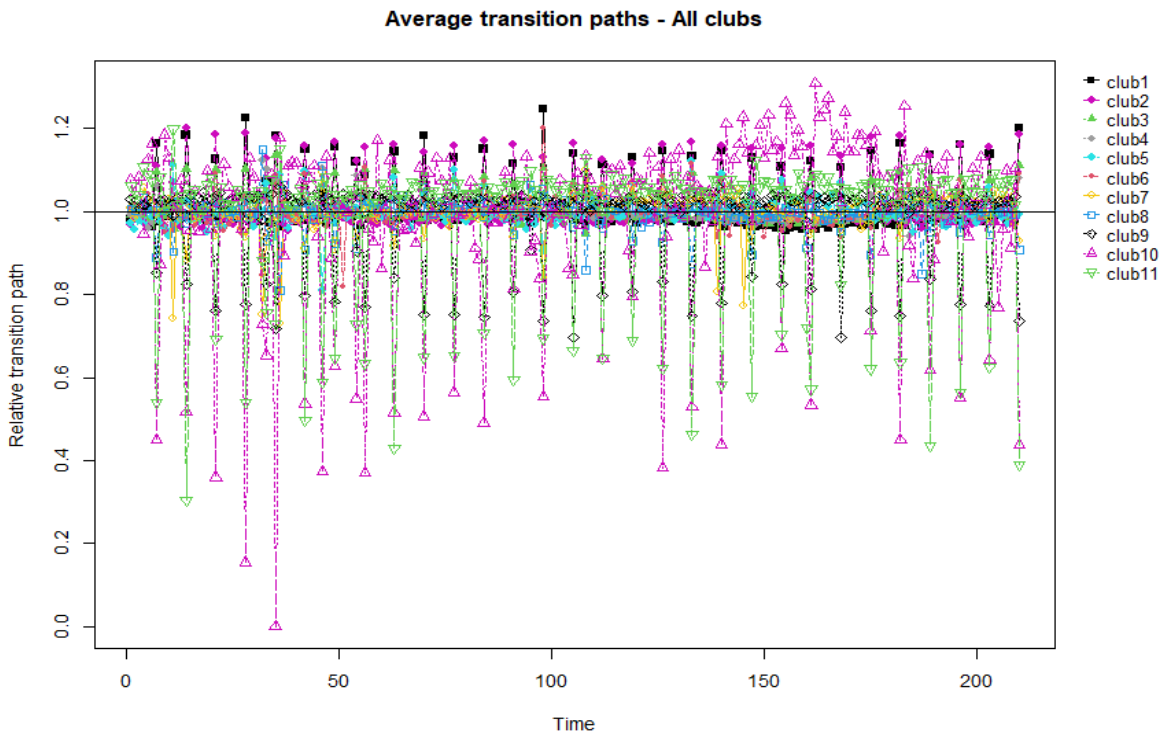
RPLOT 11 CLUBES



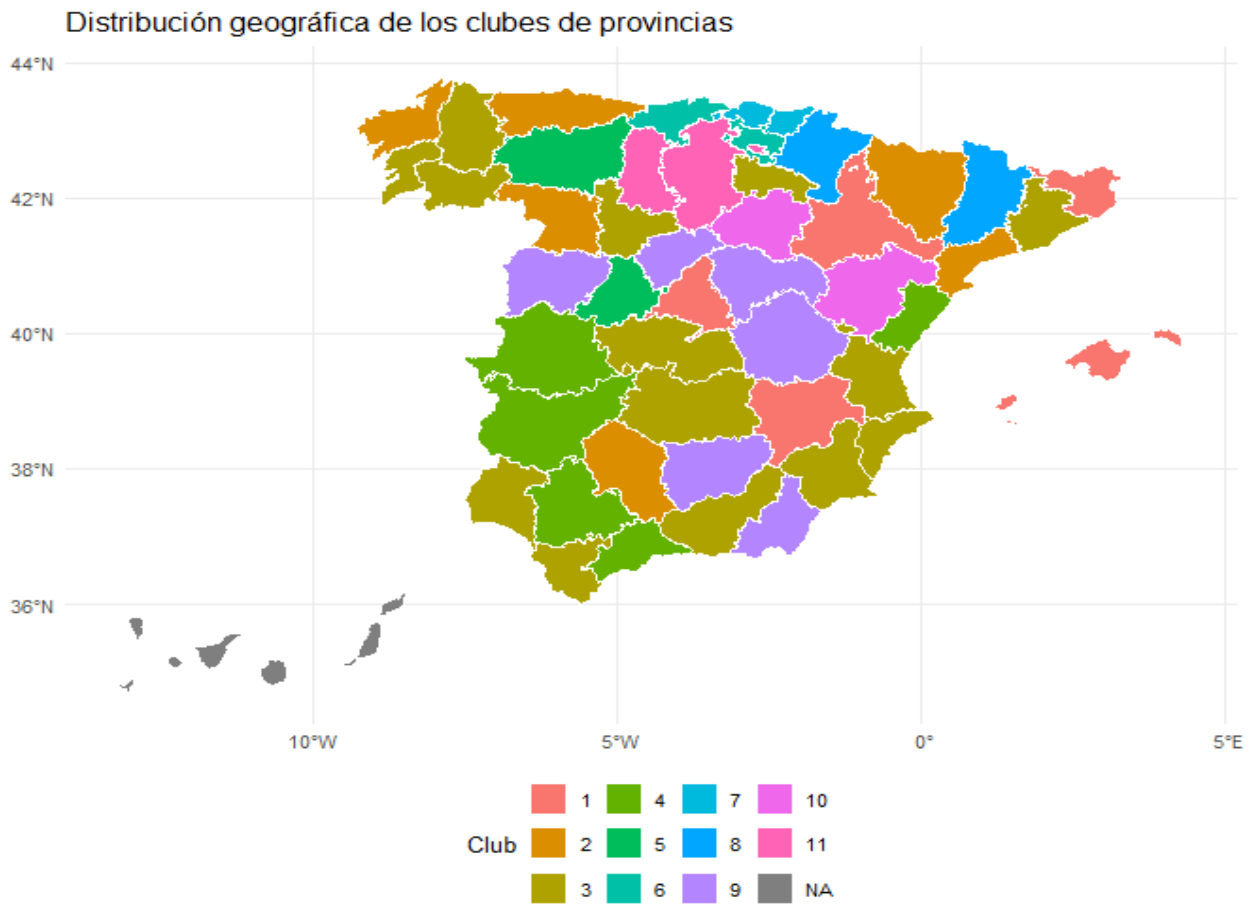
RPLOT 11 CLUBES SIN TRANSICION



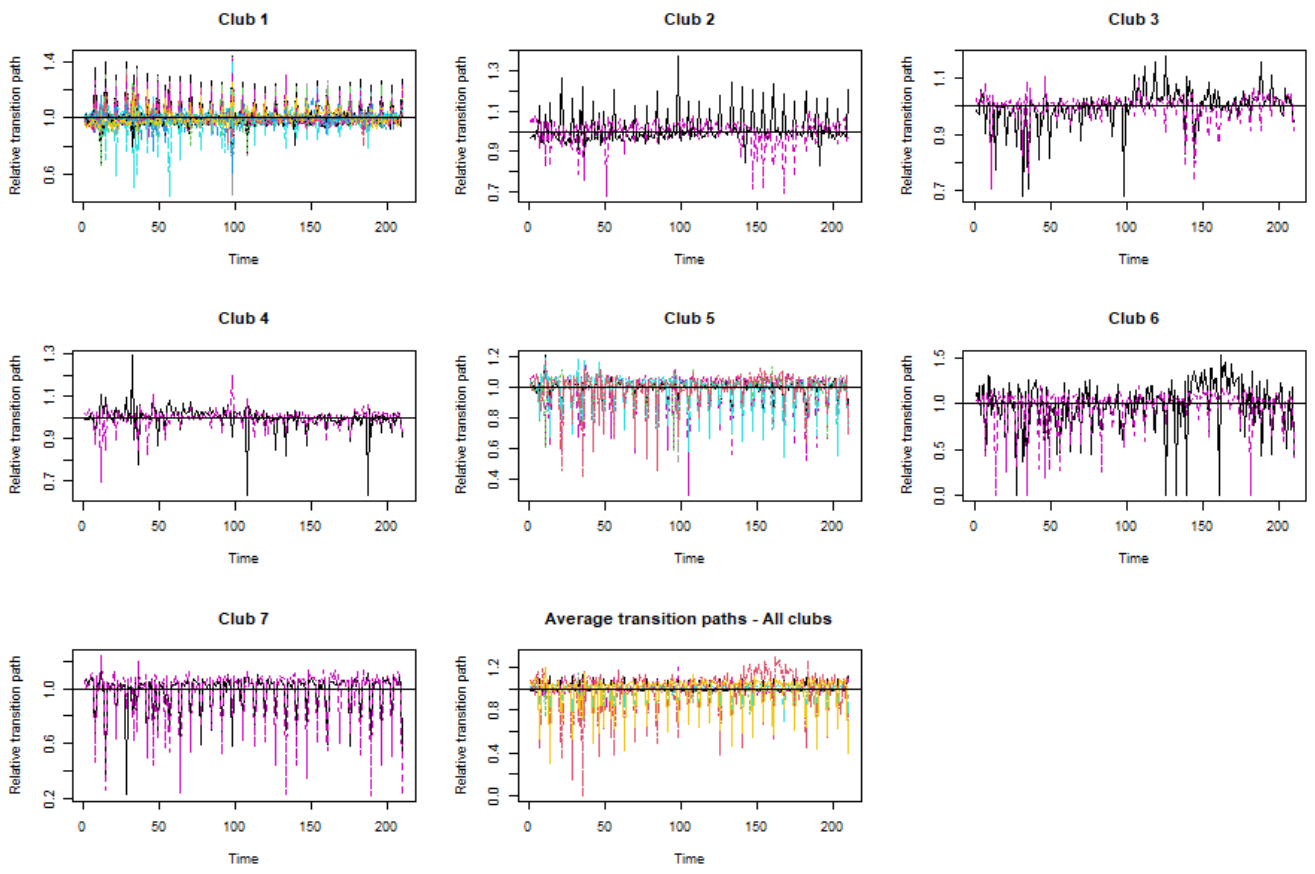
RPLOT COMPORTAMIENTO PROMEDIO DE LAS VENTAS EN CADA CLUB EN EL TIEMPO



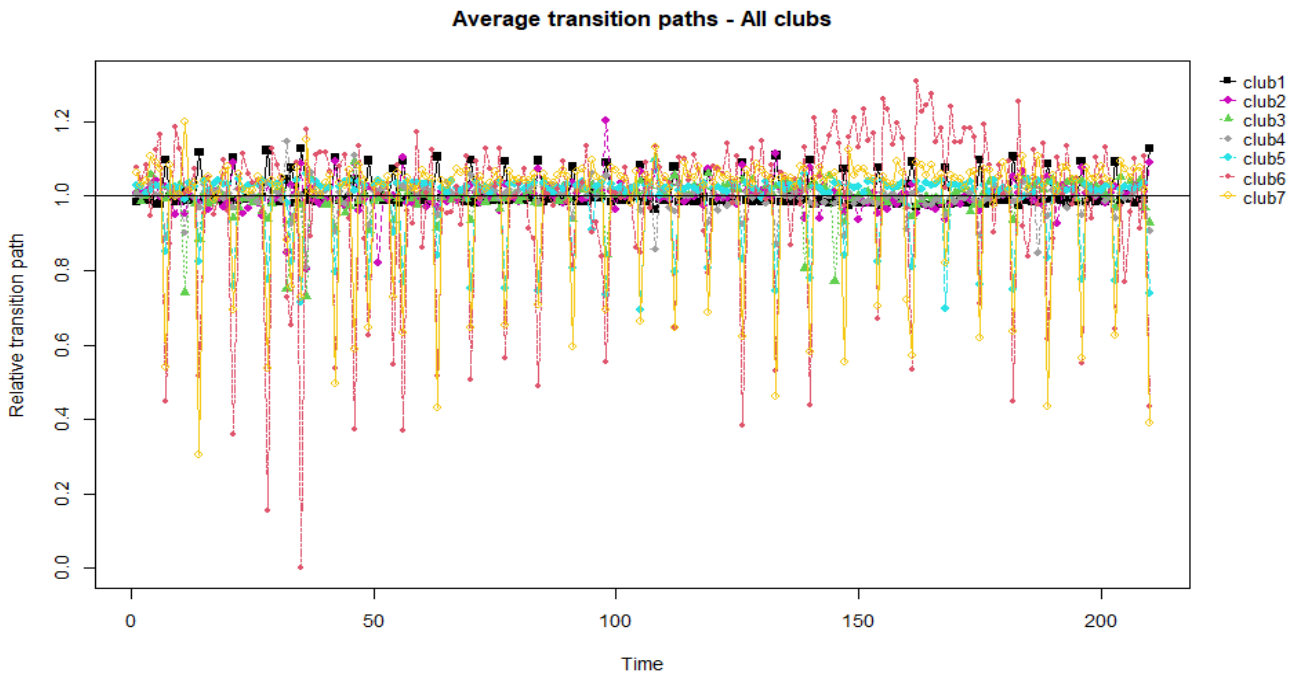
MAPA DE SITUACION GEOGRAFICA DE CADA CLUB



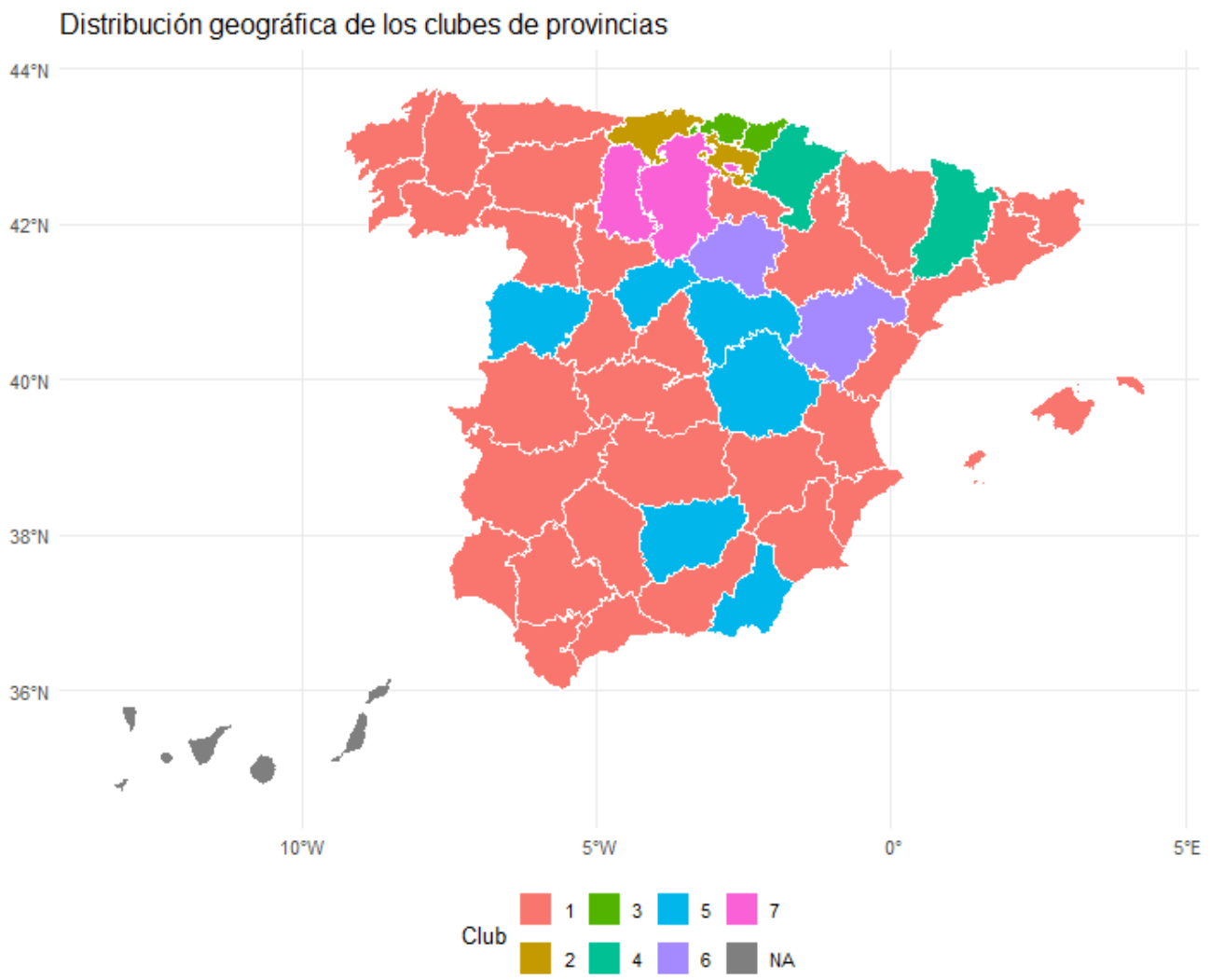
REDUCCION A 7 CLUBES



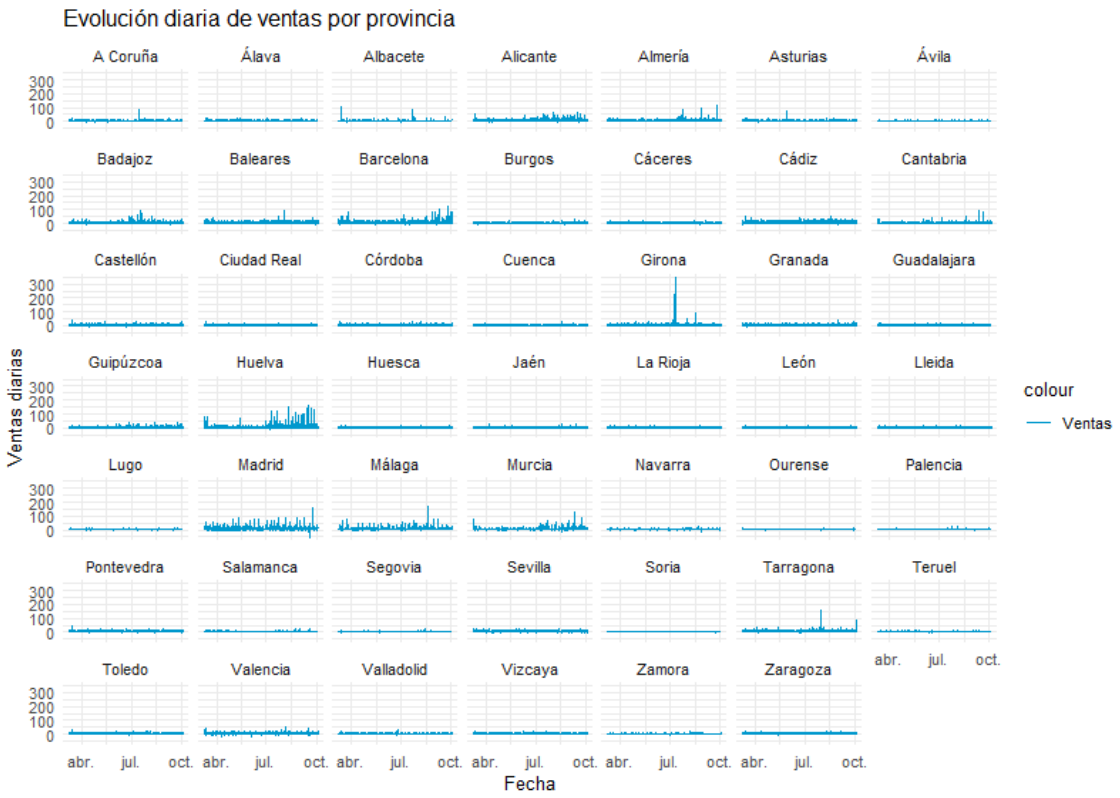
RPLOT COMPORTAMIENTO PROMEDIO DE LAS VENTAS EN LOS 7 CLUBES EN EL TIEMPO

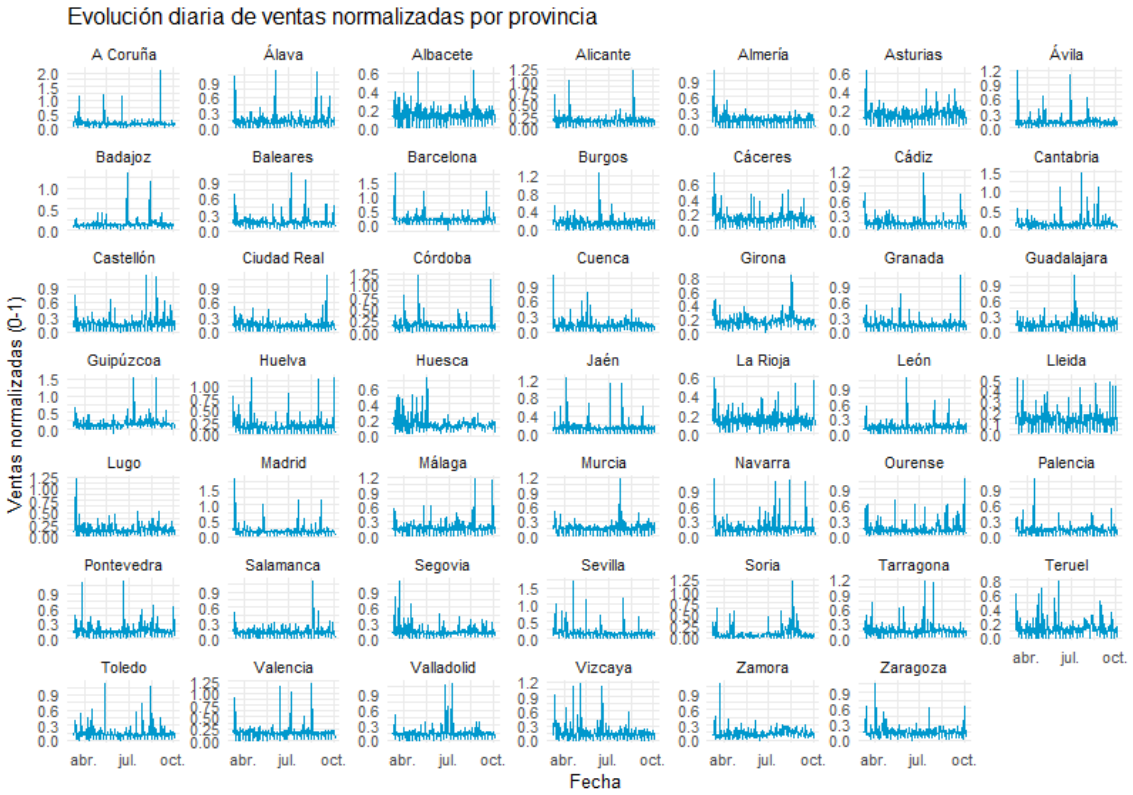


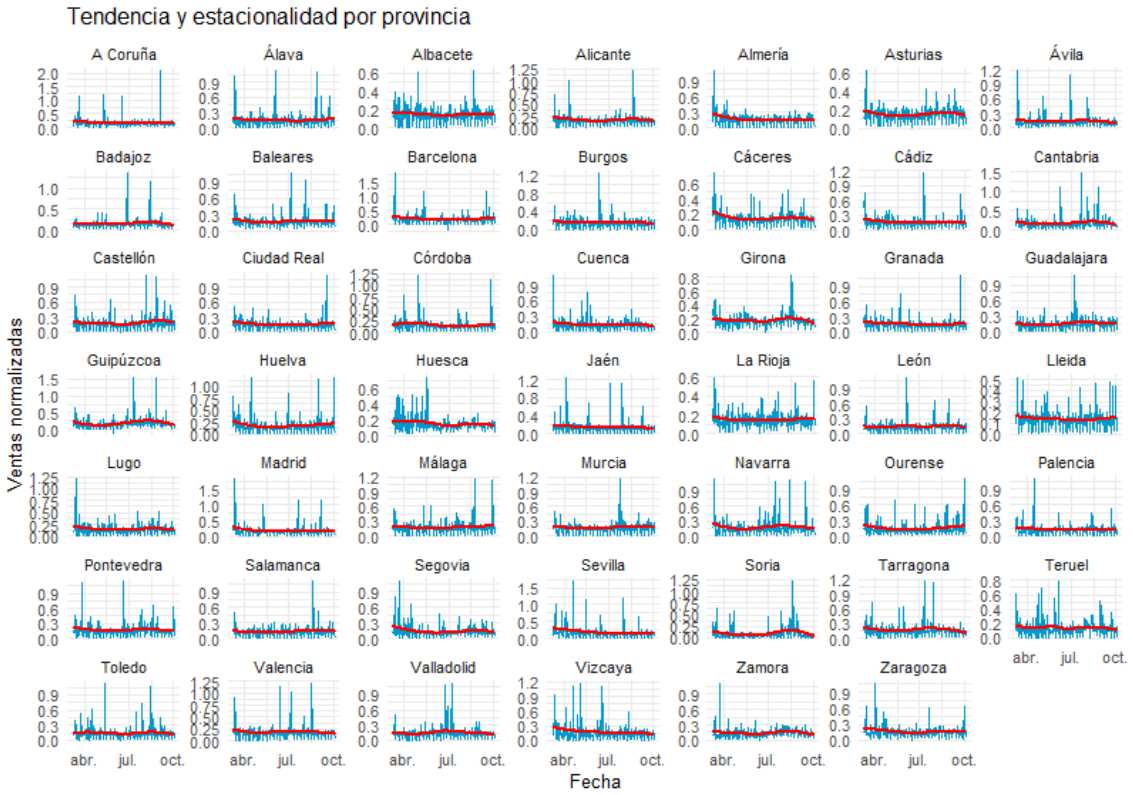
MAPA DE SITUACION GEOGRAFICA DE CADA UNO DE LOS 7 CLUBES

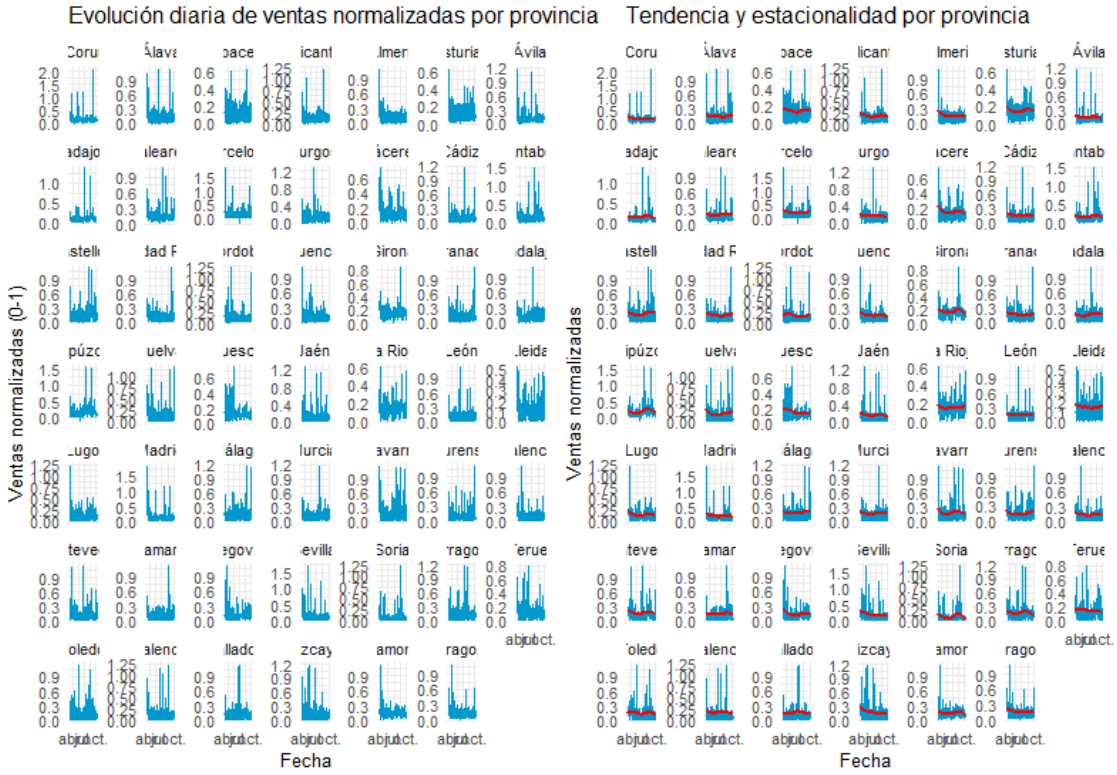


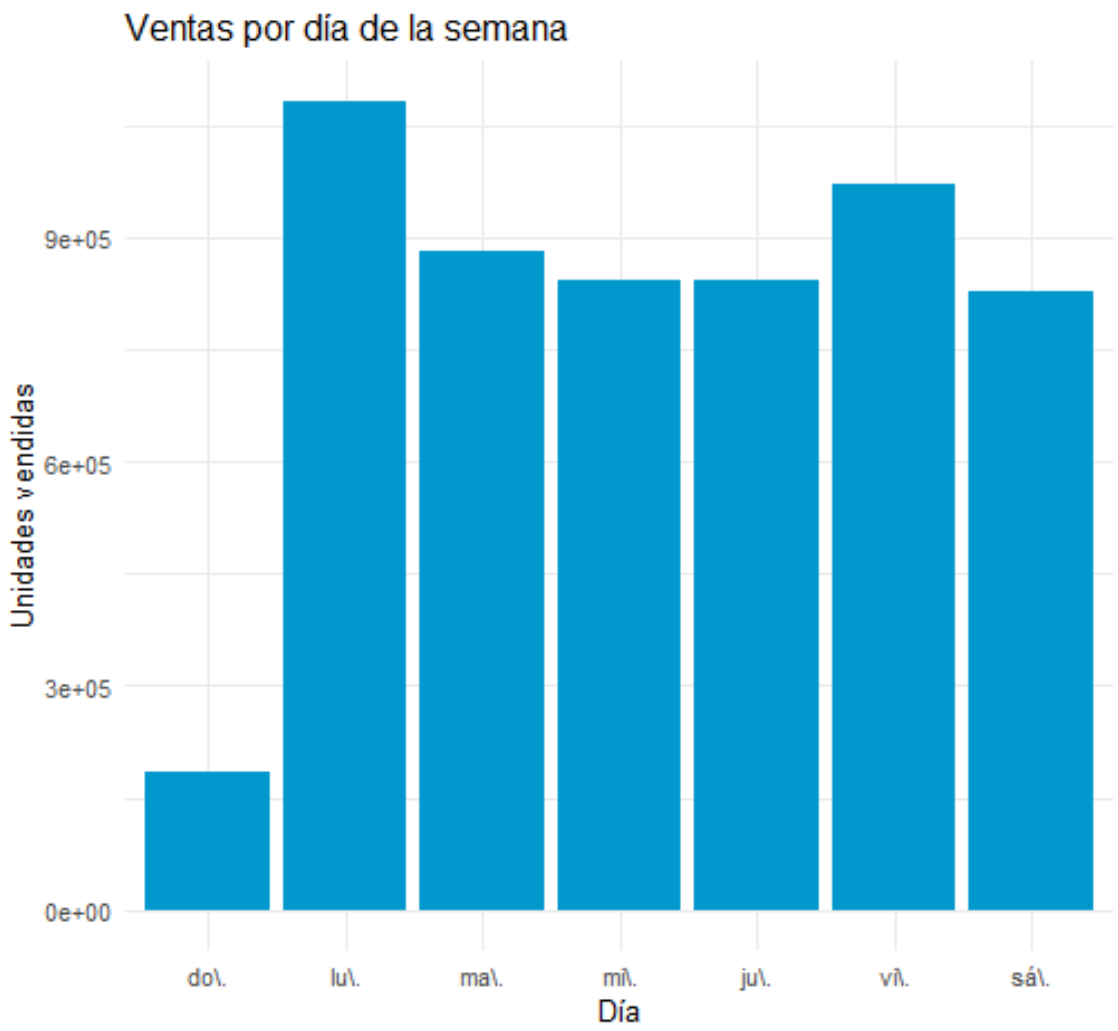
Anexo E. Gráficos de Temporalidad de las ventas

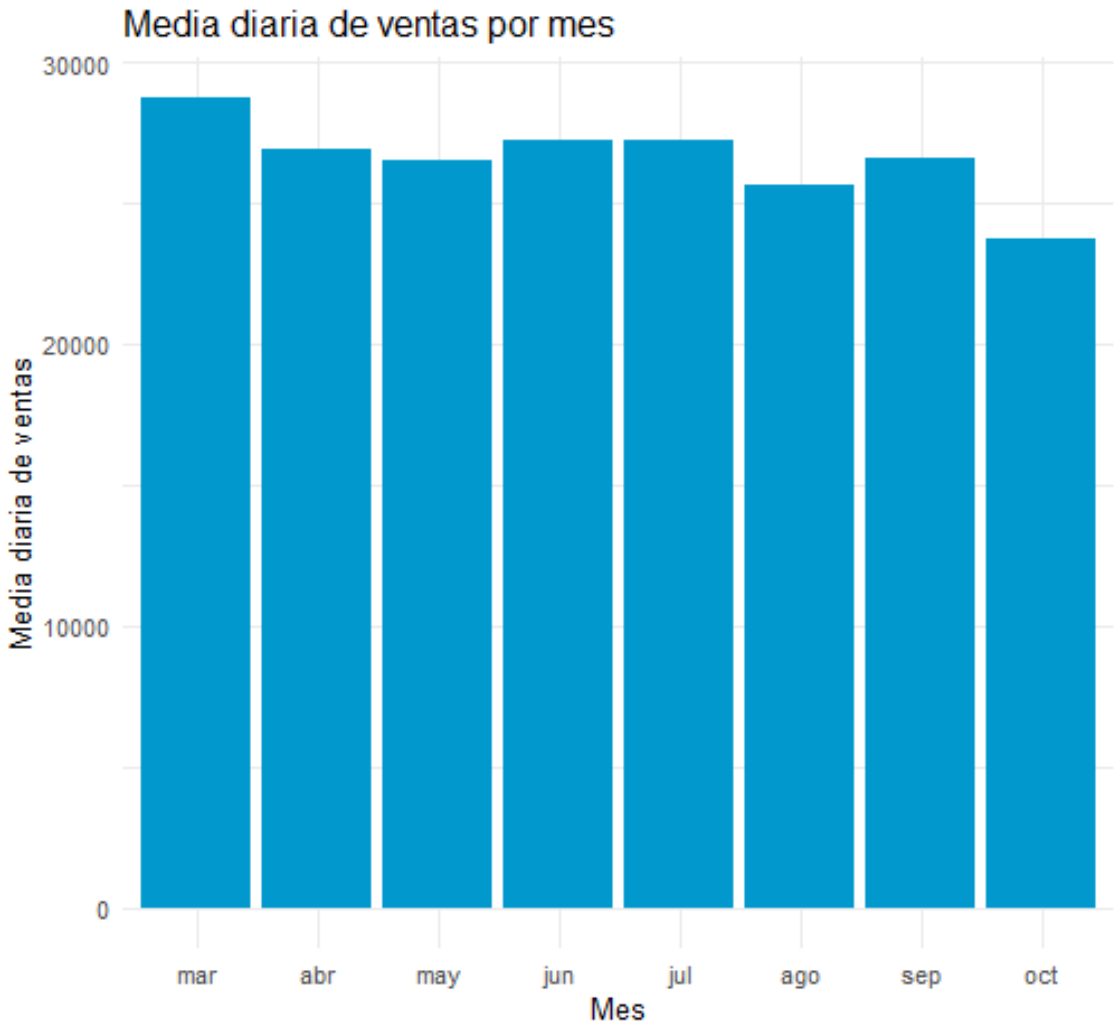


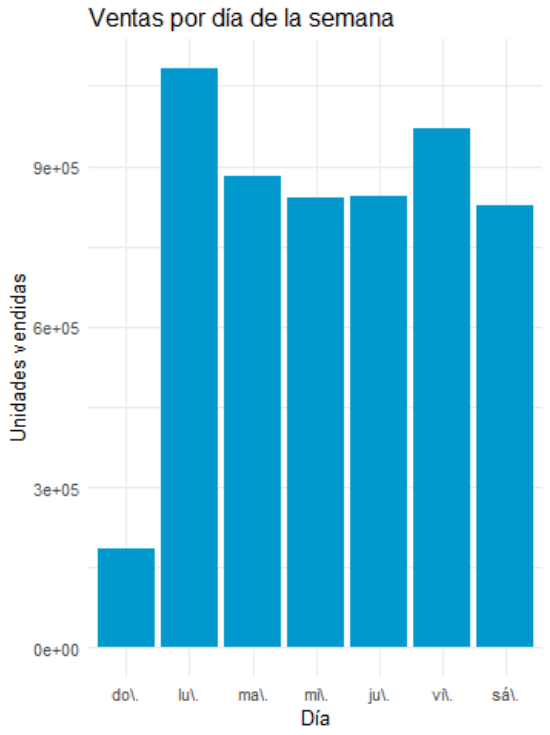

Evolución diaria de ventas normalizadas por provincia
A Coruña
Álava
Albacete
Alicante
Almería
Asturias
Ávila
Badajoz
Baleares
Barcelona
Burgos
Cáceres
Cádiz
Cantabria
Castellón
Ciudad Real
Córdoba
Cuenca
Girona
Granada
Guadalajara
Guipúzcoa
Huelva
Huesca
Jaén
La Rioja
León
Lleida
Lugo
Madrid
Málaga
Murcia
Navarra
Ourense
Palencia
Pontevedra
Salamanca
Segovia
Sevilla
Soria
Tarragona
Teruel
Toledo
Valencia
Valladolid
Vizcaya
Zamora
Zaragoza
Ventas normalizadas (0-1)
abr.
jul.
oct.
Fecha

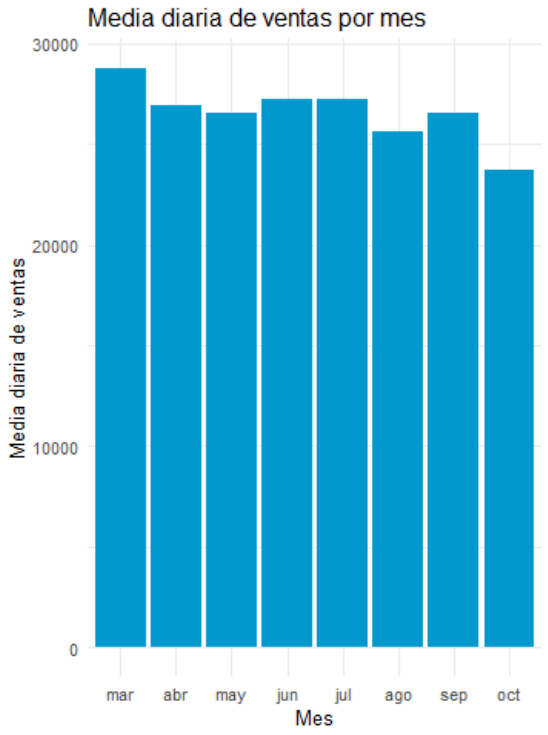

Tendencia y estacionalidad por provincia
A Coruña
Álava
Albacete
Alicante
Almería
Asturias
Ávila
Badajoz
Baleares
Barcelona
Burgos
Cáceres
Cádiz
Cantabria
Castellón
Ciudad Real
Córdoba
Cuenca
Girona
Granada
Guadalajara
Guipúzcoa
Huelva
Huesca
Jaén
La Rioja
León
Lleida
Lugo
Madrid
Málaga
Murcia
Navarra
Ourense
Palencia
Pontevedra
Salamanca
Segovia
Sevilla
Soria
Tarragona
Teruel
Toledo
Valencia
Valladolid
Vizcaya
Zamora
Zaragoza
Ventas normalizadas
abr.
jul.
oct.
Fecha

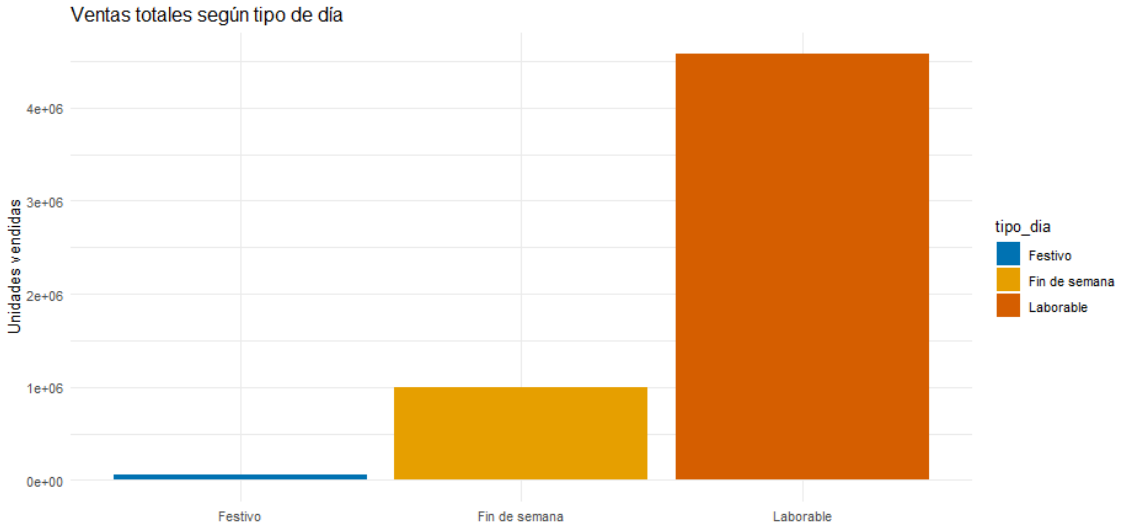

Evolución diaria de ventas normalizadas por provincia
Tendencia y estacionalidad por provincia
Ventas normalizadas (0-1)
Ventas normalizadas
Fecha
Fecha

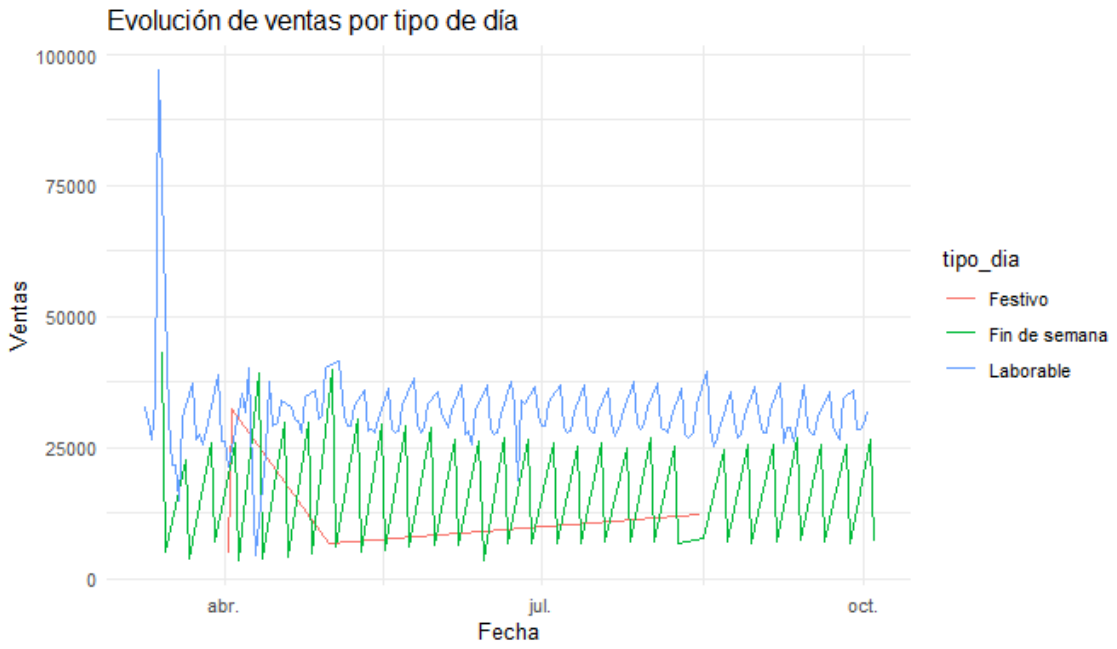

Ventas por día de la semana
Unidades vendidas
9e+05
6e+05
3e+05
0e+00
do\.
lu\.
ma\.
mi\.
ju\.
vi\.
sá\.
Día

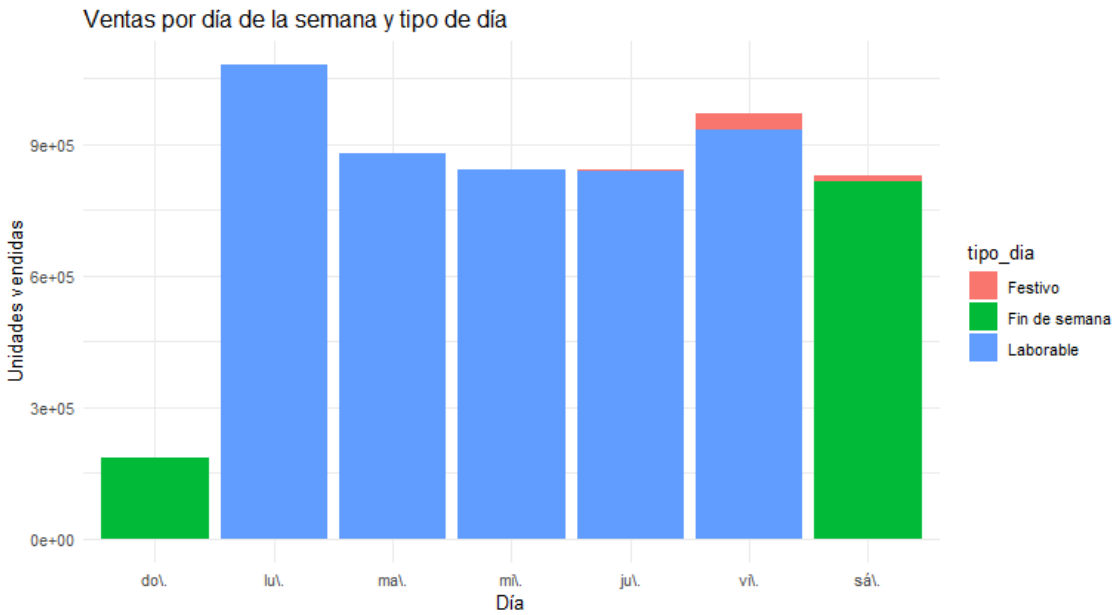

Media diaria de ventas por mes
30000
20000
10000
0
Media diaria de ventas
mar
abr
may
jun
jul
ago
sep
oct
Mes

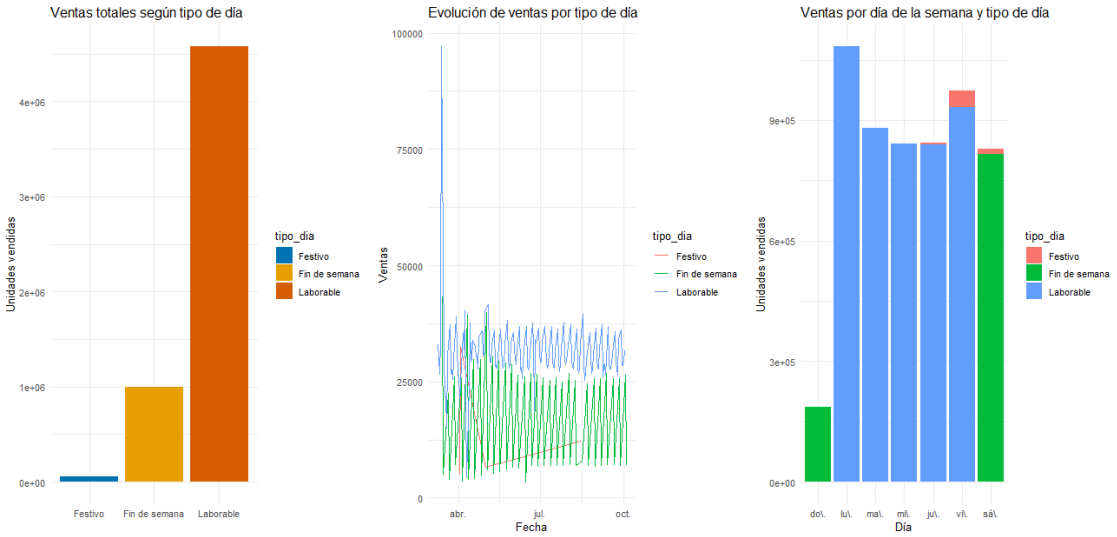

Ventas por día de la semana
Unidades vendidas
0e+00
3e+05
6e+05
9e+05
do\.
lu\.
ma\.
mi\.
ju\.
vi\.
sá\.
Día

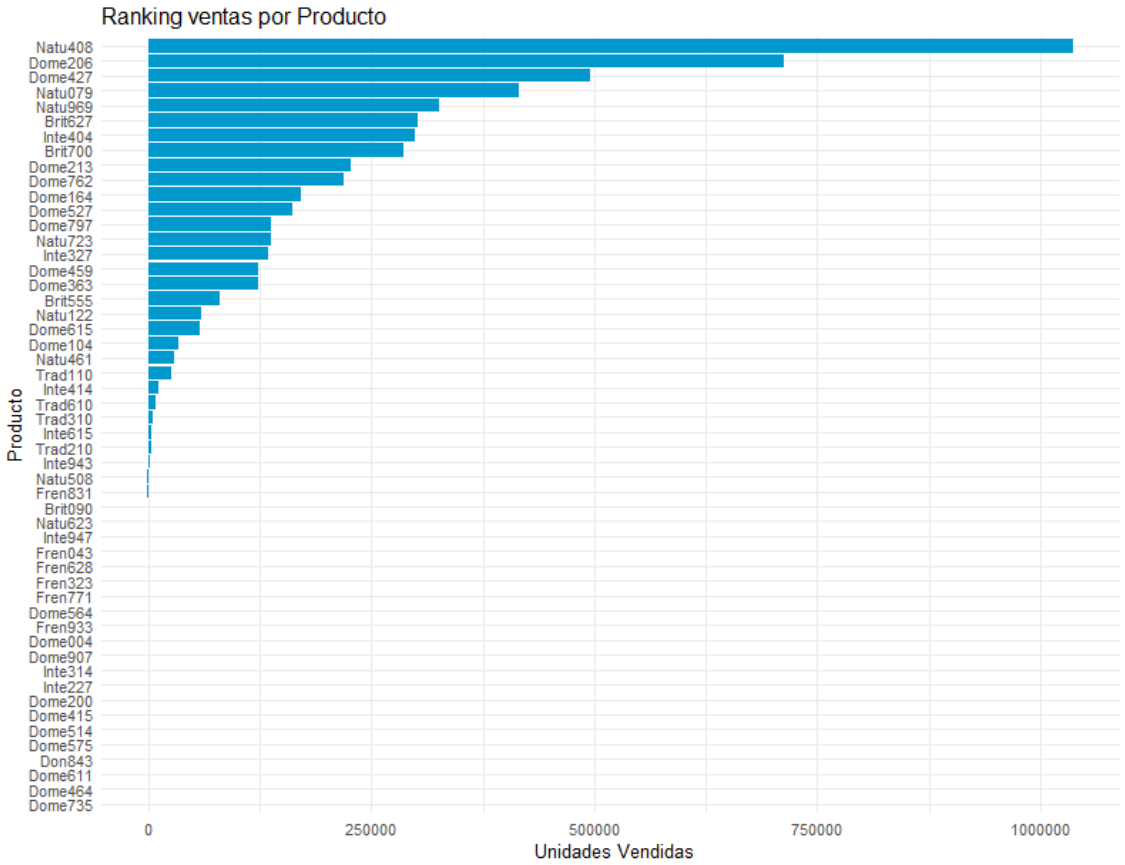

Media diaria de ventas por mes
Media diaria de ventas
0
10000
20000
30000
mar
abr
may
jun
jul
ago
sep
oct
Mes

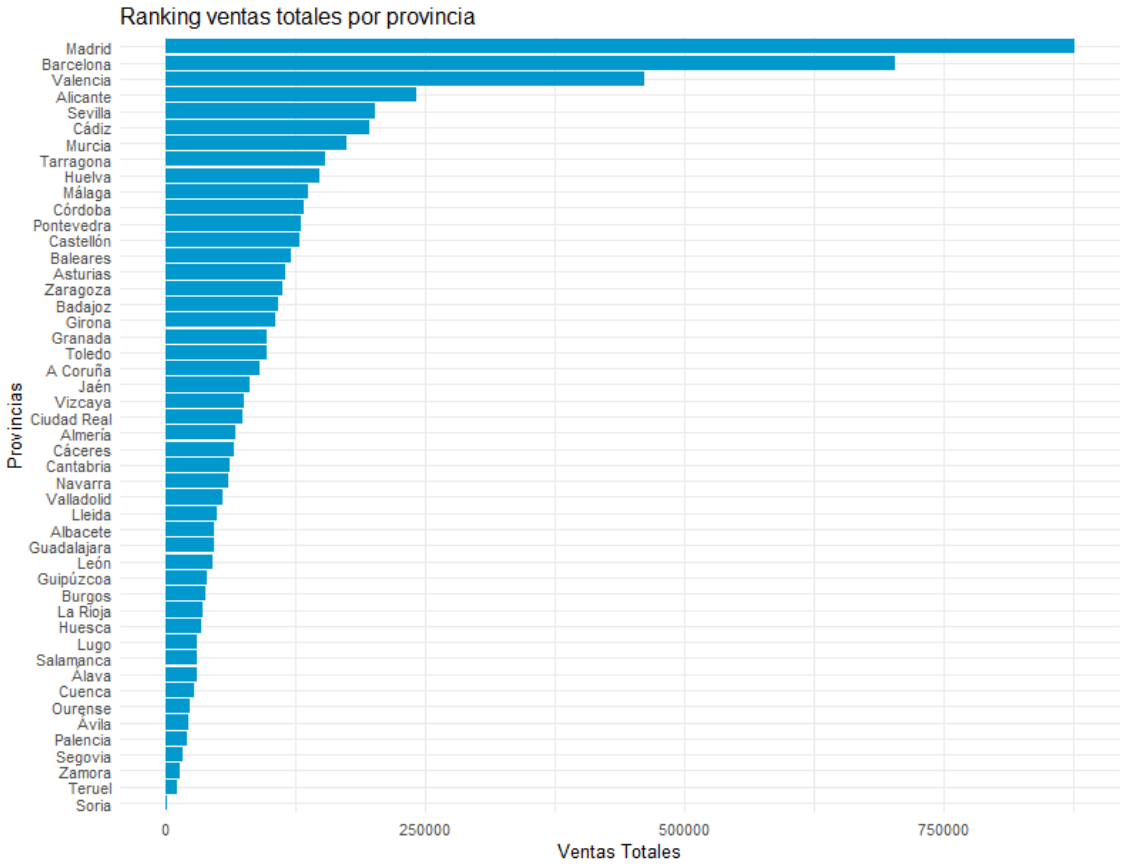

Ventas totales según tipo de día
Unidades vendidas
0e+00
1e+06
2e+06
3e+06
4e+06
Festivo
Fin de semana
Laborable
tipo_dia
Festivo
Fin de semana
Laborable

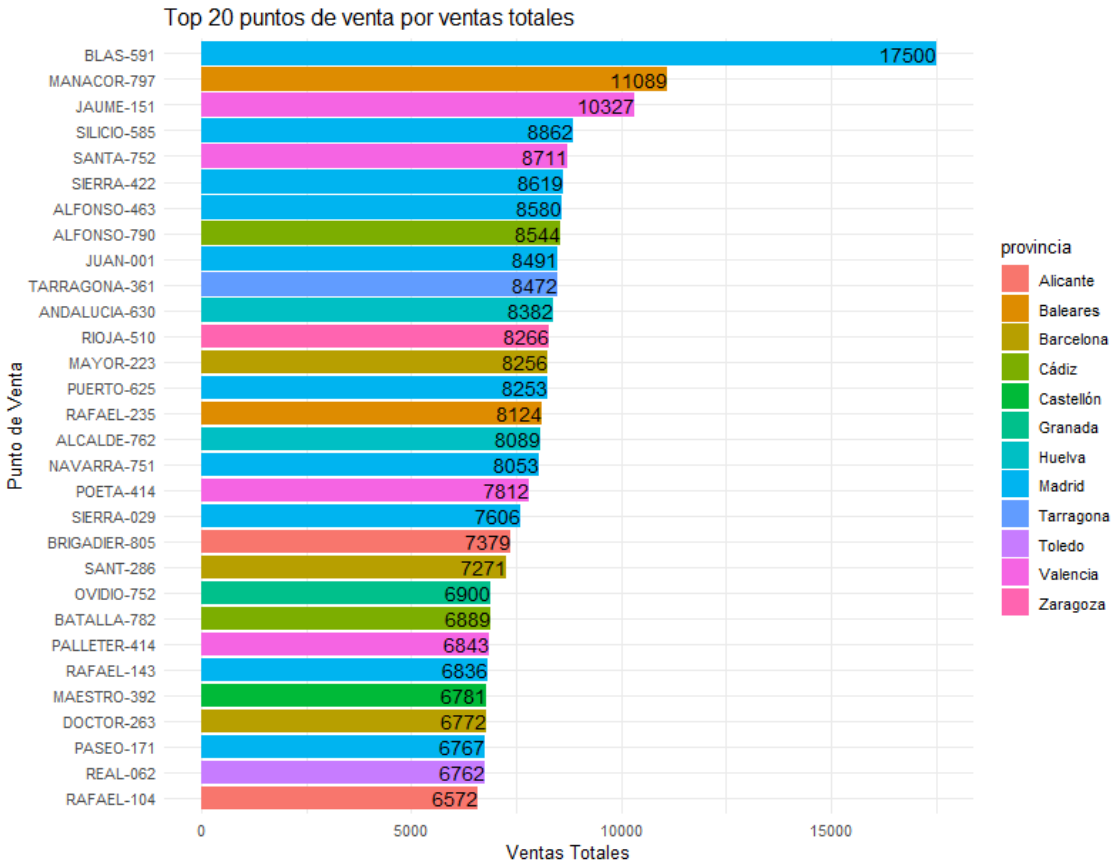

Evolución de ventas por tipo de día
100000
75000
50000
25000
0
Ventas
abr.
jul.
oct.
Fecha
tipo_dia
Festivo
Fin de semana
Laborable

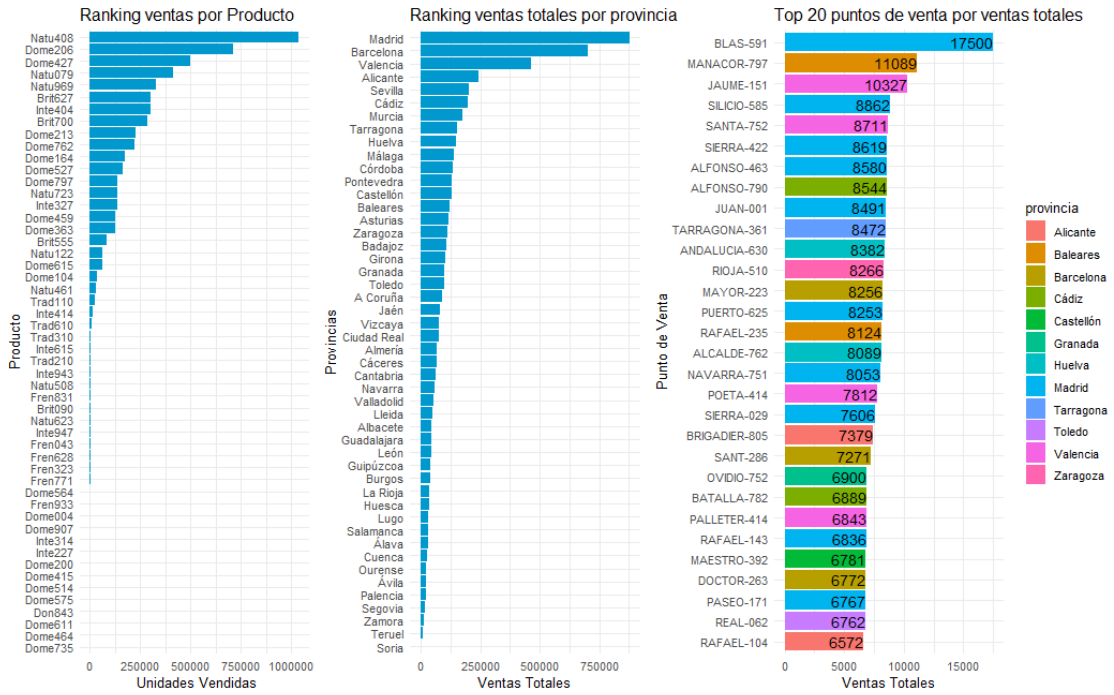

Ventas por día de la semana y tipo de día
9e+05
6e+05
3e+05
0e+00
Unidades vendidas
do\.
lu\.
ma\.
mi\.
ju\.
vi\.
sá\.
Día
tipo_dia
Festivo
Fin de semana
Laborable

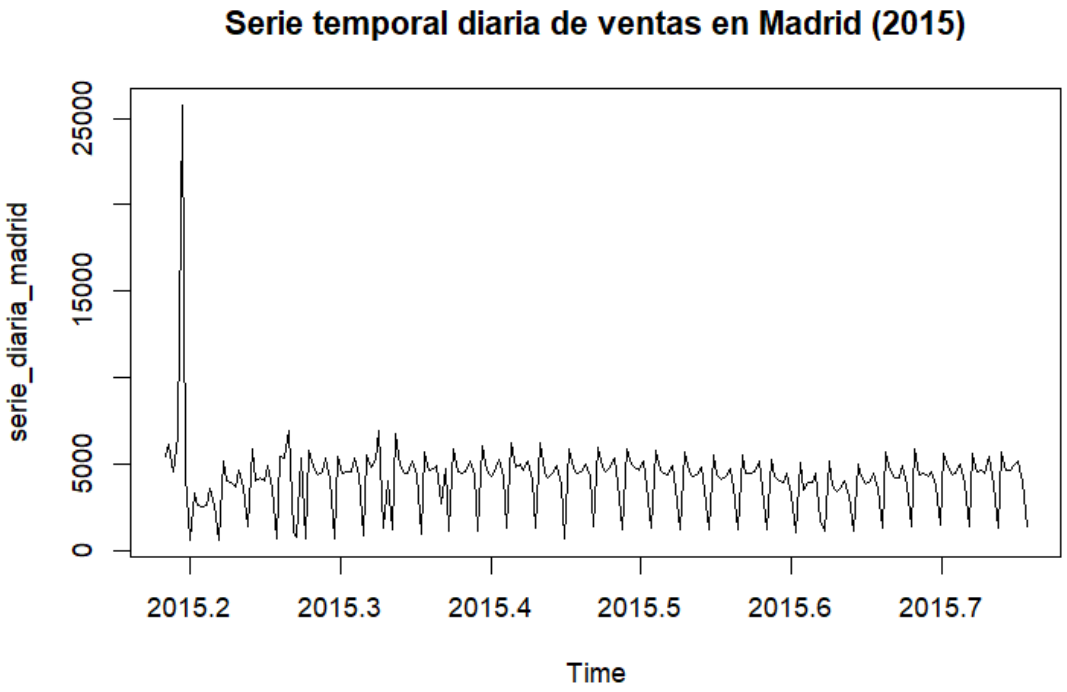

Ventas totales según tipo de día
Unidades vendidas
4e+06
3e+06
2e+06
1e+06
0e+00
Festivo
Fin de semana
Laborable
tipo_dia
Festivo
Fin de semana
Laborable
Evolución de ventas por tipo de día
Ventas
100000
75000
50000
25000
0
abr.
jul.
oct.
Fecha
tipo_dia
Festivo
Fin de semana
Laborable
Ventas por día de la semana y tipo de día
Unidades vendidas
9e+05
6e+05
3e+05
0e+00
do\.
lu\.
ma\.
mi\.
ju\.
vi\.
sá\.
Día
tipo_dia
Festivo
Fin de semana
Laborable

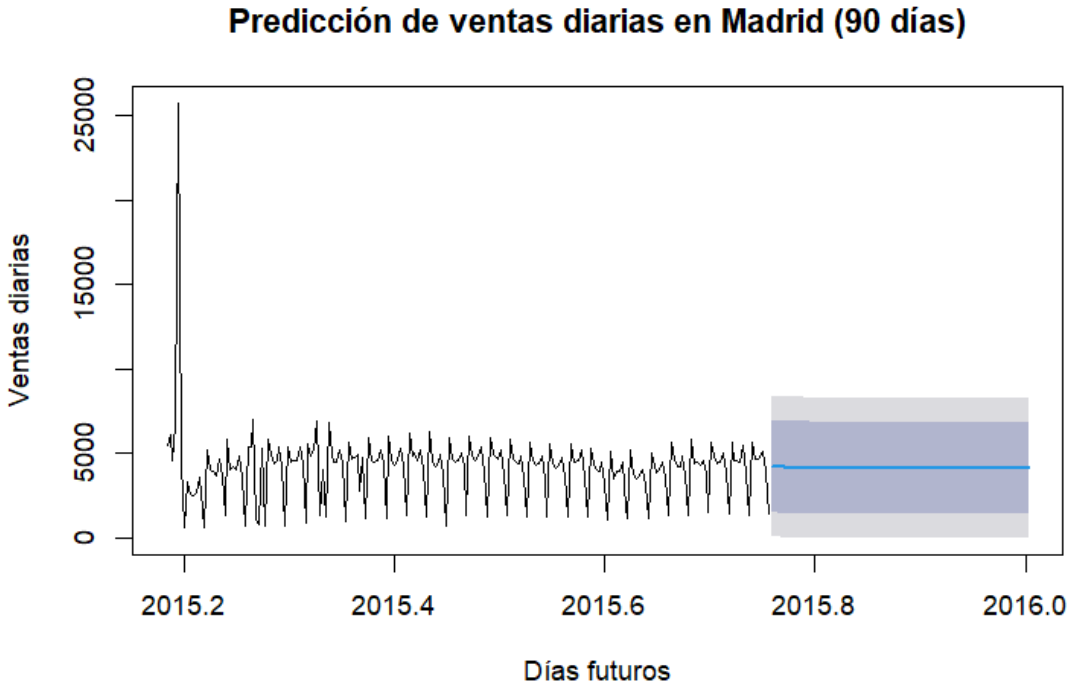

Ranking ventas por Producto
Producto
Natu408
Dome206
Dome427
Natu079
Natu969
Brit627
Inte404
Brit700
Dome213
Dome762
Dome164
Dome527
Dome797
Natu723
Inte327
Dome459
Dome363
Brit555
Natu122
Dome615
Dome104
Natu461
Trad110
Inte414
Trad610
Trad310
Inte615
Trad210
Inte943
Natu508
Fren831
Brit090
Natu623
Inte947
Fren043
Fren628
Fren323
Fren771
Dome564
Fren933
Dome004
Dome907
Inte314
Inte227
Dome200
Dome415
Dome514
Dome575
Don843
Dome611
Dome464
Dome735
0
250000
500000
750000
1000000
Unidades Vendidas

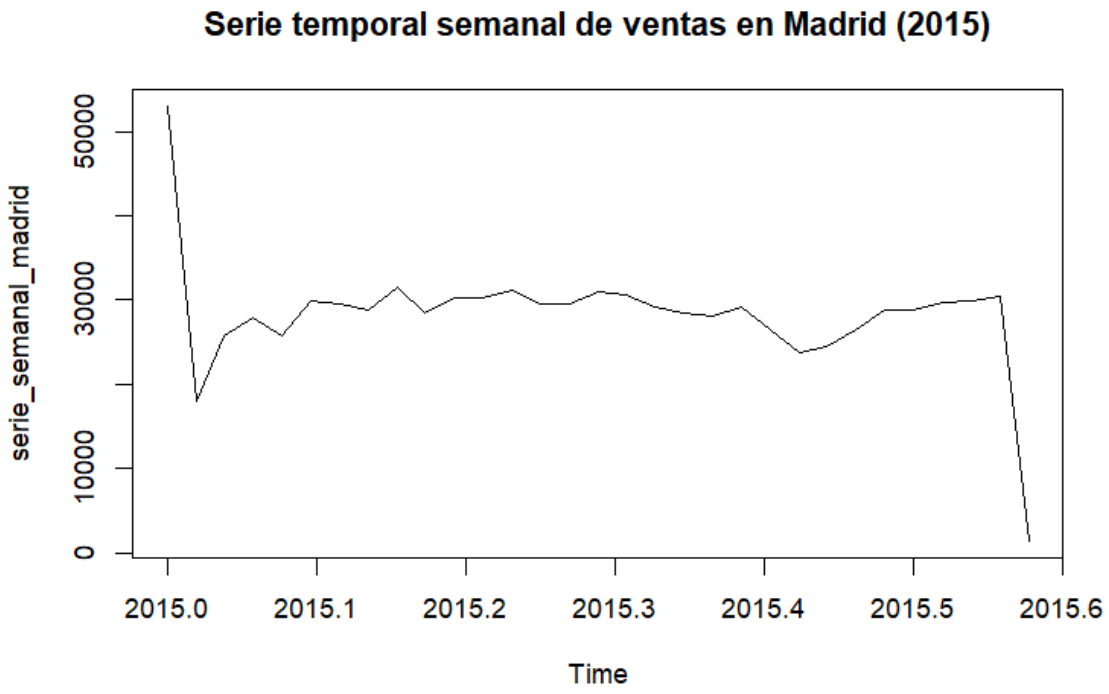

Ranking ventas totales por provincia
Madrid
Barcelona
Valencia
Alicante
Sevilla
Cádiz
Murcia
Tarragona
Huelva
Málaga
Córdoba
Pontevedra
Castellón
Baleares
Asturias
Zaragoza
Badajoz
Girona
Granada
Toledo
A Coruña
Jaén
Vizcaya
Ciudad Real
Almería
Cáceres
Cantabria
Navarra
Valladolid
Lleida
Albacete
Guadalajara
León
Guipúzcoa
Burgos
La Rioja
Huesca
Lugo
Salamanca
Álava
Cuenca
Ourense
Ávila
Palencia
Segovia
Zamora
Teruel
Soria
Provincias
0
250000
500000
750000
Ventas Totales

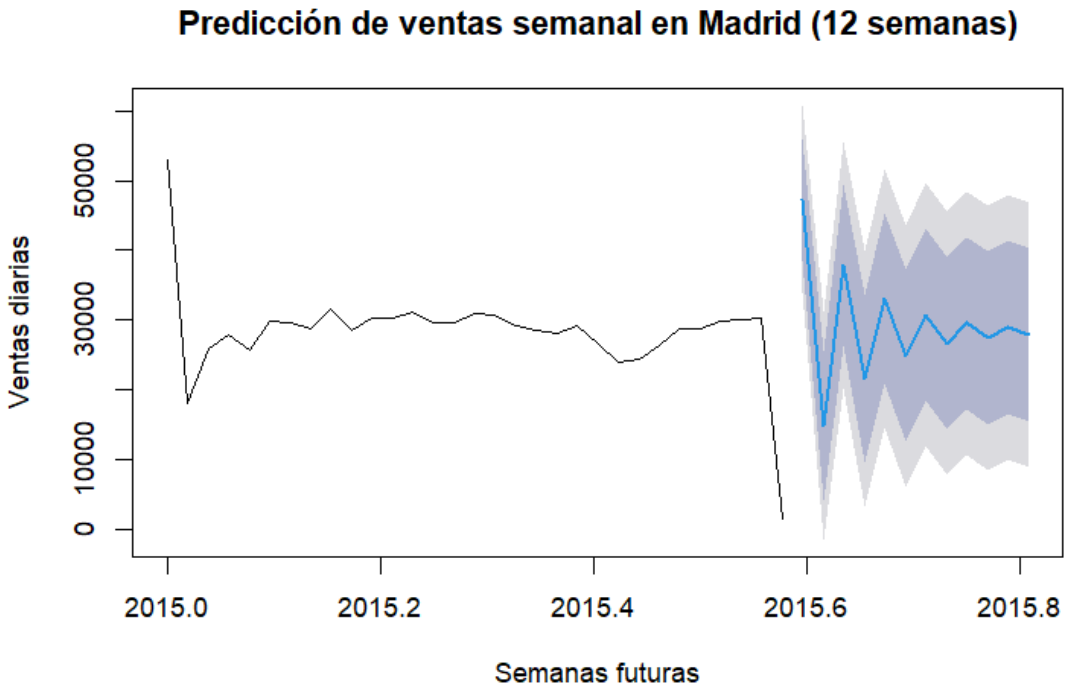

Top 20 puntos de venta por ventas totales
BLAS-591 17500
MANACOR-797 11089
JAUME-151 10327
SILICIO-585 8862
SANTA-752 8711
SIERRA-422 8619
ALFONSO-463 8580
ALFONSO-790 8544
JUAN-001 8491
TARRAGONA-361 8472
ANDALUCIA-630 8382
RIOJA-510 8266
MAYOR-223 8256
PUERTO-625 8253
RAFAEL-235 8124
ALCALDE-762 8089
NAVARRA-751 8053
POETA-414 7812
SIERRA-029 7606
BRIGADIER-805 7379
SANT-286 7271
OVIDIO-752 6900
BATALLA-782 6889
PALLETER-414 6843
RAFAEL-143 6836
MAESTRO-392 6781
DOCTOR-263 6772
PASEO-171 6767
REAL-062 6762
RAFAEL-104 6572
Punto de Venta
0
5000
10000
15000
Ventas Totales
provincia
Alicante
Baleares
Barcelona
Cádiz
Castellón
Granada
Huelva
Madrid
Tarragona
Toledo
Valencia
Zaragoza

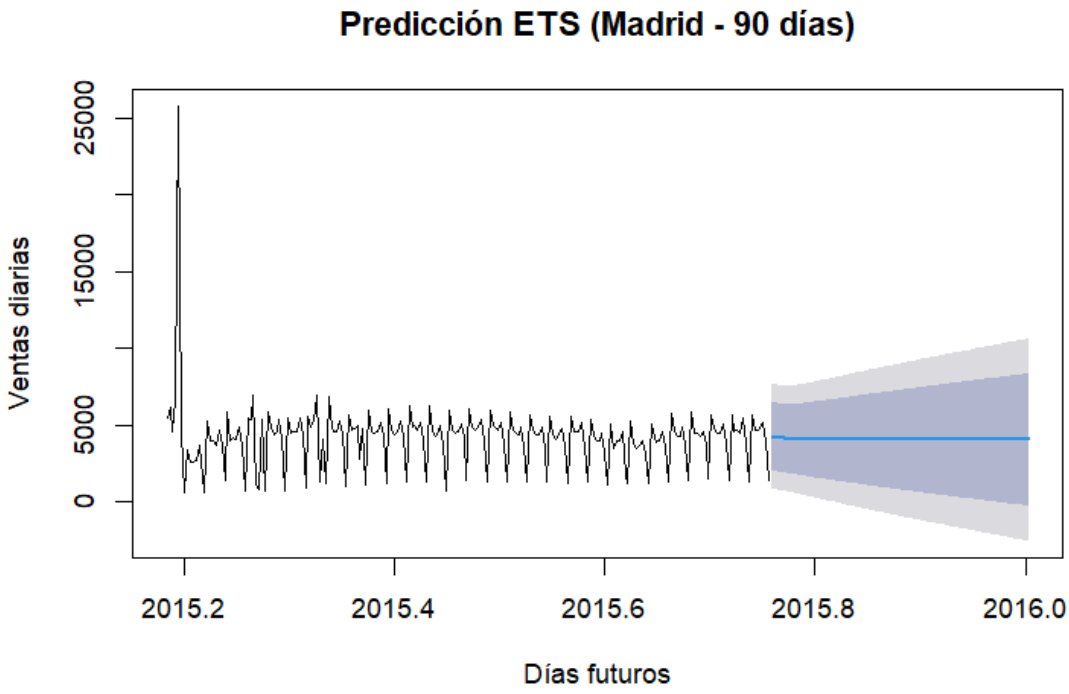

Ranking ventas por Producto
Producto
Unidades Vendidas
Ranking ventas totales por provincia
Provincias
Ventas Totales
Top 20 puntos de venta por ventas totales
Punto de Venta
Ventas Totales
BLAS-591 17500
MANACOR-797 11089
JAUME-151 10327
SILICIO-585 8862
SANTA-752 8711
SIERRA-422 8619
ALFONSO-463 8580
ALFONSO-790 8544
JUAN-001 8491
TARRAGONA-361 8472
ANDALUCIA-630 8382
RIOJA-510 8266
MAYOR-223 8256
PUERTO-625 8253
RAFAEL-235 8124
ALCALDE-762 8089
NAVARRA-751 8053
POETA-414 7812
SIERRA-029 7606
BRIGADIER-805 7379
SANT-286 7271
OVIDIO-752 6900
BATALLA-782 6889
PALLETER-414 6843
RAFAEL-143 6836
MAESTRO-392 6781
DOCTOR-263 6772
PASEO-171 6767
REAL-062 6762
RAFAEL-104 6572
provincia
Alicante
Baleares
Barcelona
Cádiz
Castellón
Granada
Huelva
Madrid
Tarragona
Toledo
Valencia
Zaragoza

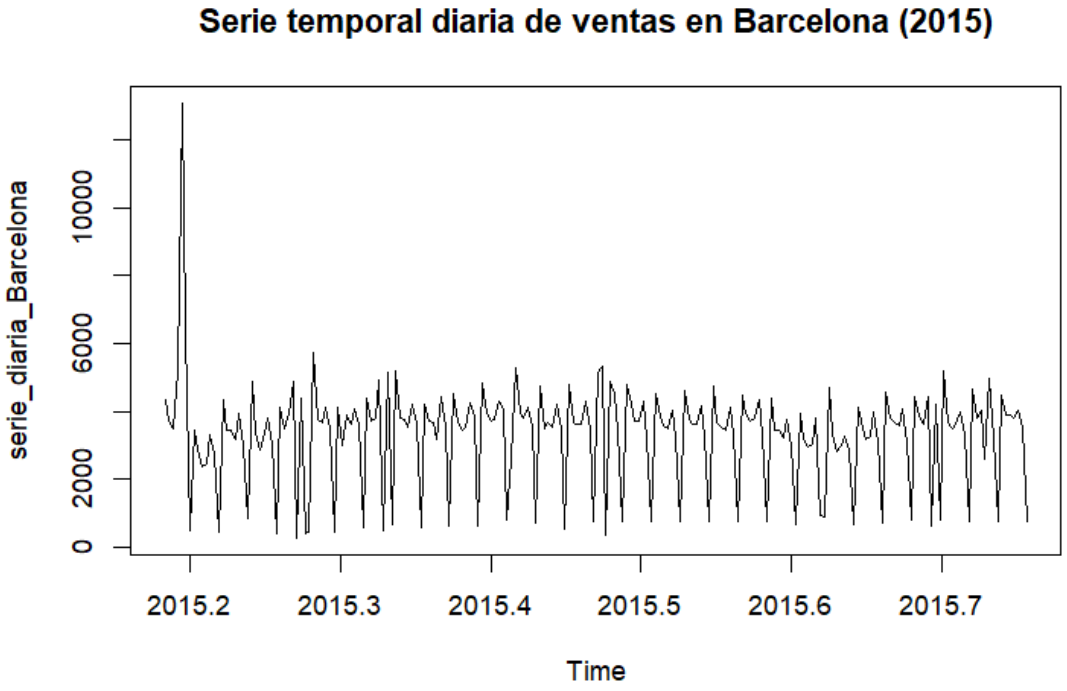

Serie temporal diaria de ventas en Madrid (2015)
serie_diaria_madrid
0
5000
15000
25000
2015.2
2015.3
2015.4
2015.5
2015.6
2015.7
Time

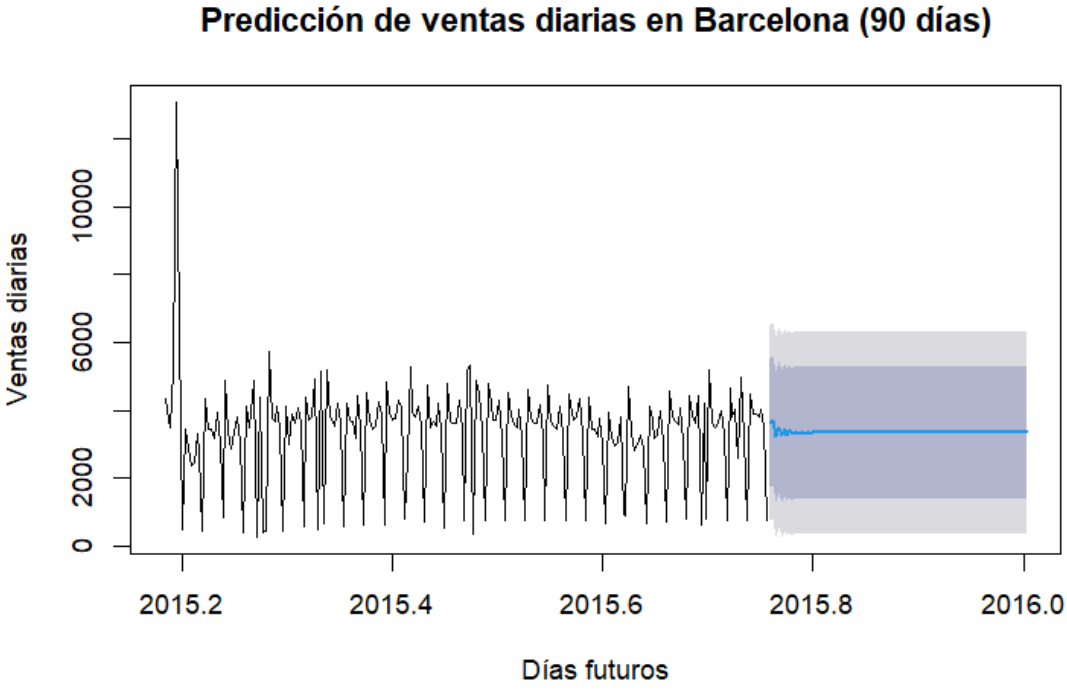

Predicción de ventas diarias en Madrid (90 días)
Ventas diarias
0
5000
15000
25000
2015.2
2015.4
2015.6
2015.8
2016.0
Días futuros

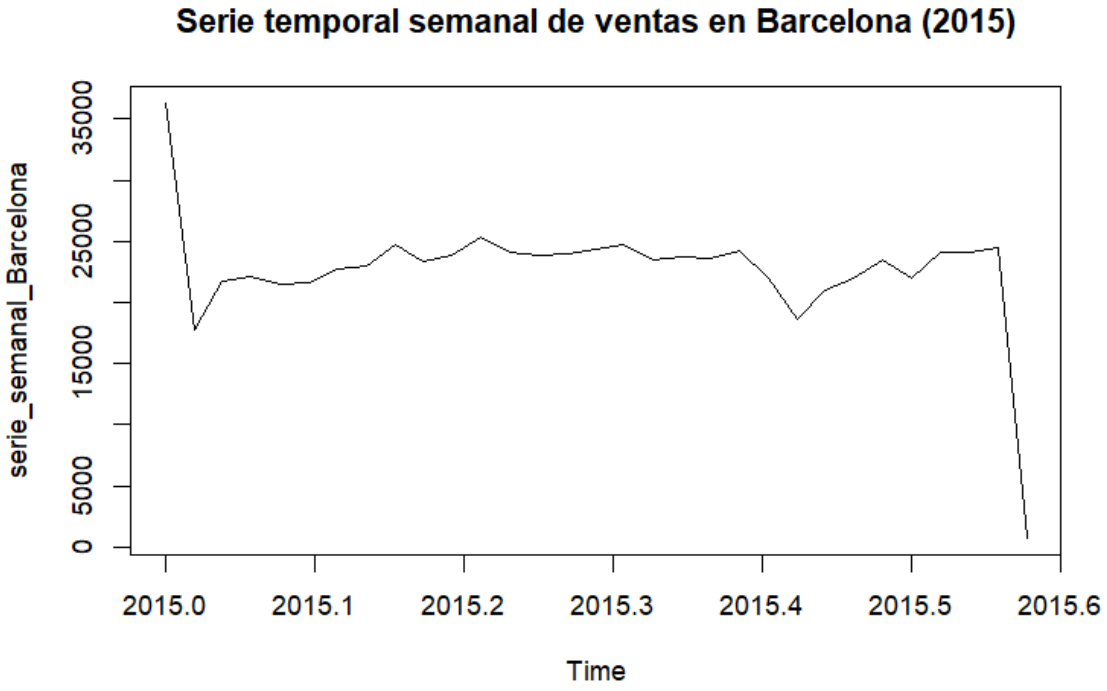

Serie temporal semanal de ventas en Madrid (2015)
serie_semanal_madrid
0
10000
30000
50000
2015.0
2015.1
2015.2
2015.3
2015.4
2015.5
2015.6
Time

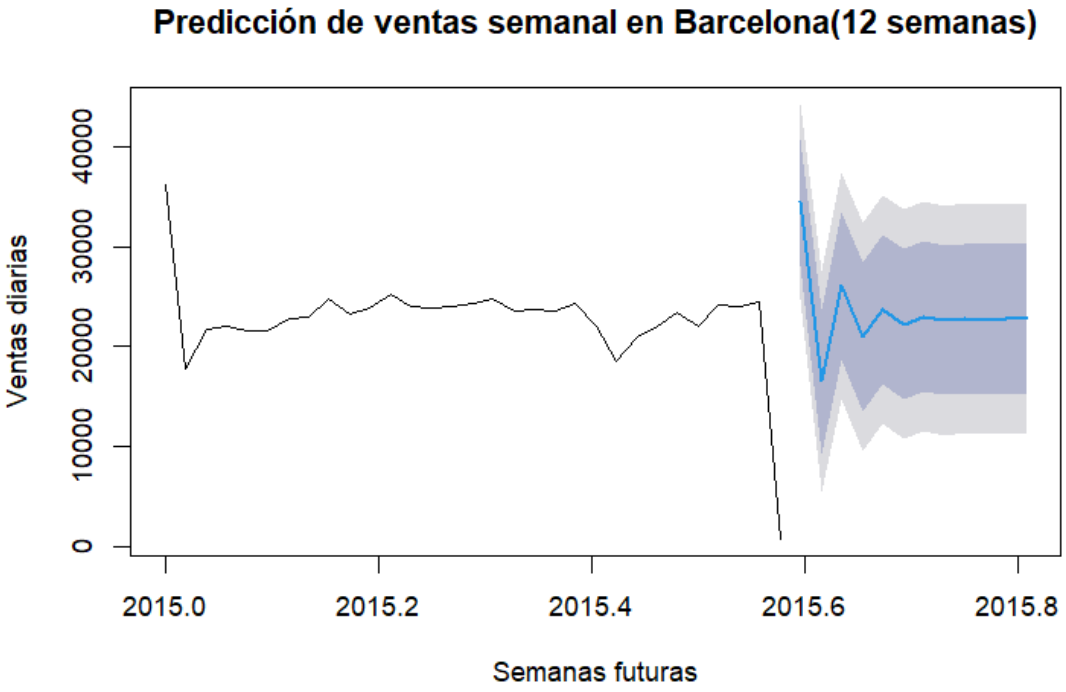

Predicción de ventas semanal en Madrid (12 semanas)
Ventas diarias
0
10000
30000
50000
2015.0
2015.2
2015.4
2015.6
2015.8
Semanas futuras

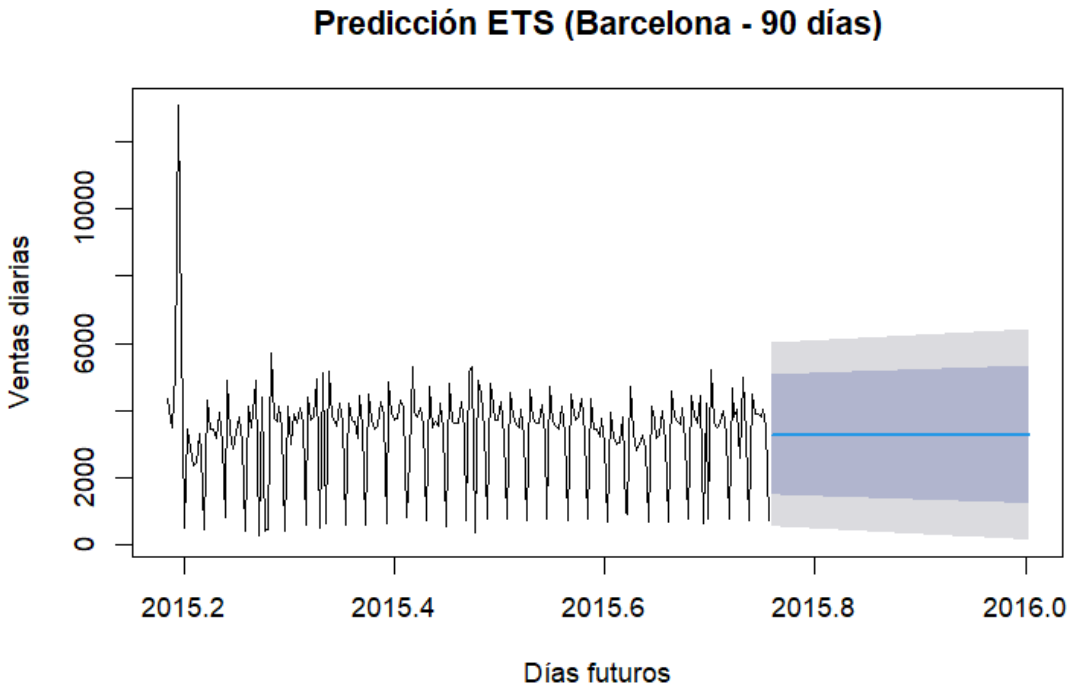

Predicción ETS (Madrid - 90 días)
Ventas diarias
0
5000
15000
25000
2015.2
2015.4
2015.6
2015.8
2016.0
Días futuros

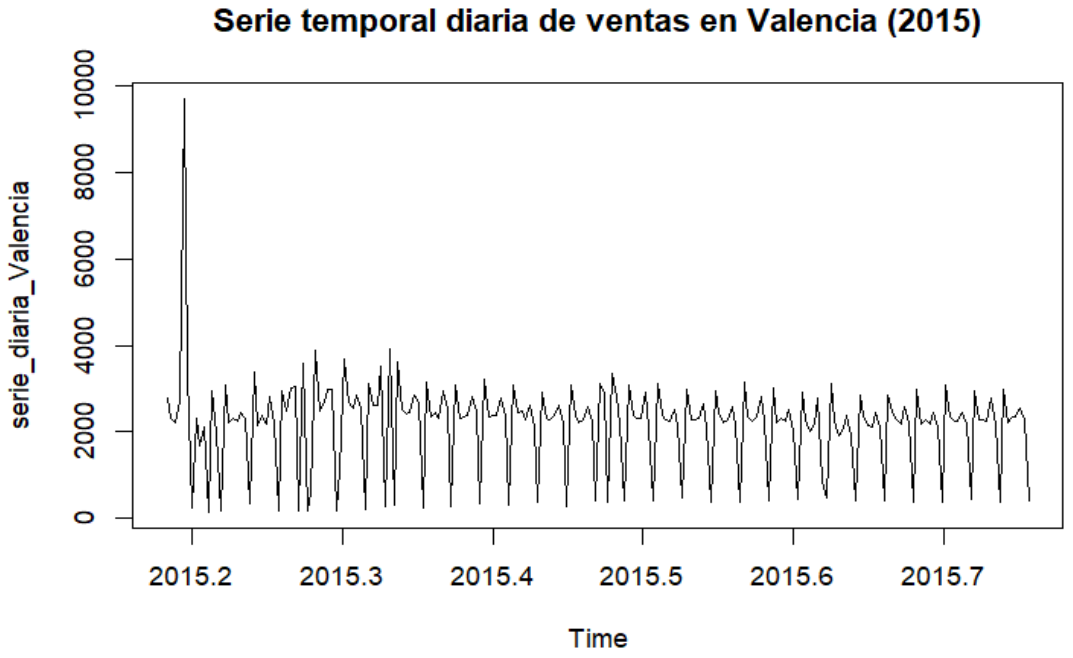

Serie temporal diaria de ventas en Barcelona (2015)
serie_diaria_Barcelona
0
2000
6000
10000
2015.2
2015.3
2015.4
2015.5
2015.6
2015.7
Time

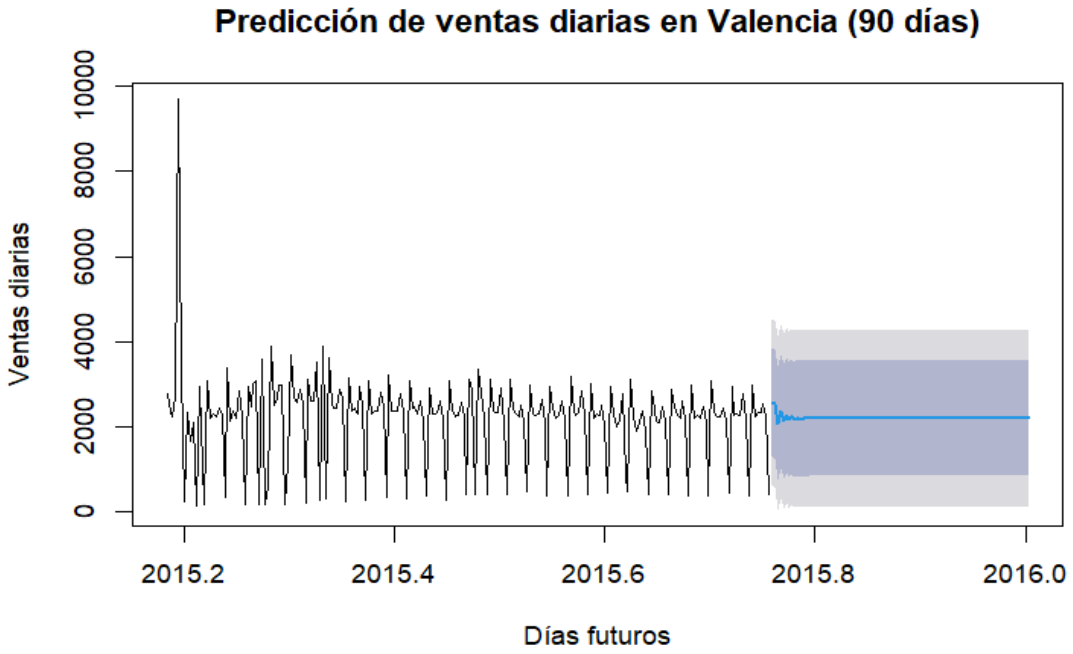

Predicción de ventas diarias en Barcelona (90 días)
Ventas diarias
0
2000
6000
10000
2015.2
2015.4
2015.6
2015.8
2016.0
Días futuros

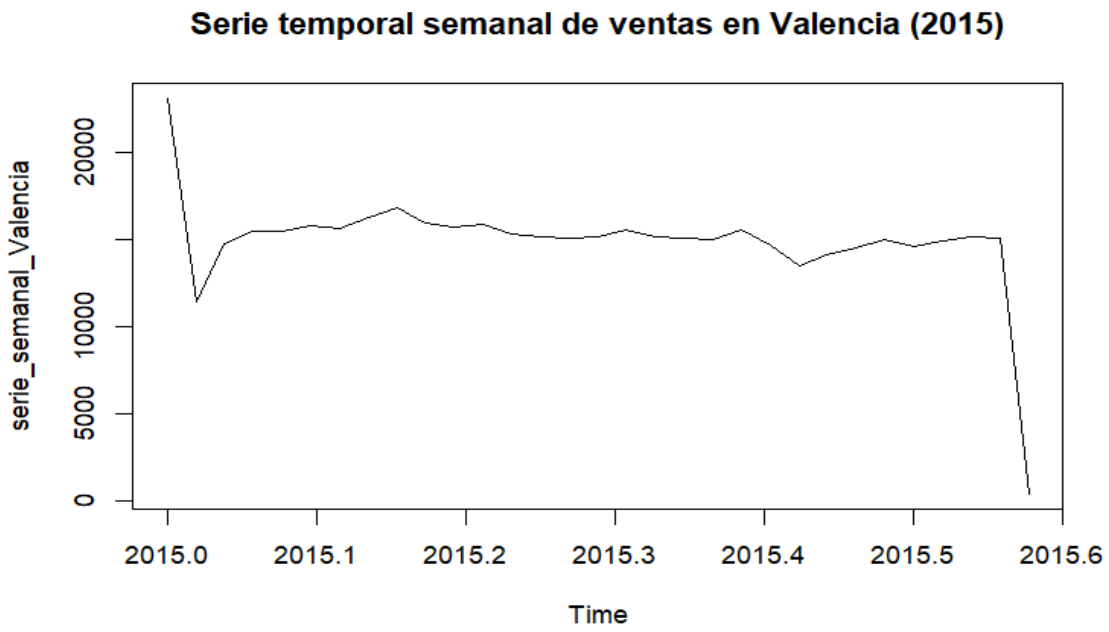

Serie temporal semanal de ventas en Barcelona (2015)
serie_semanal_Barcelona
0
5000
15000
25000
35000
2015.0
2015.1
2015.2
2015.3
2015.4
2015.5
2015.6
Time

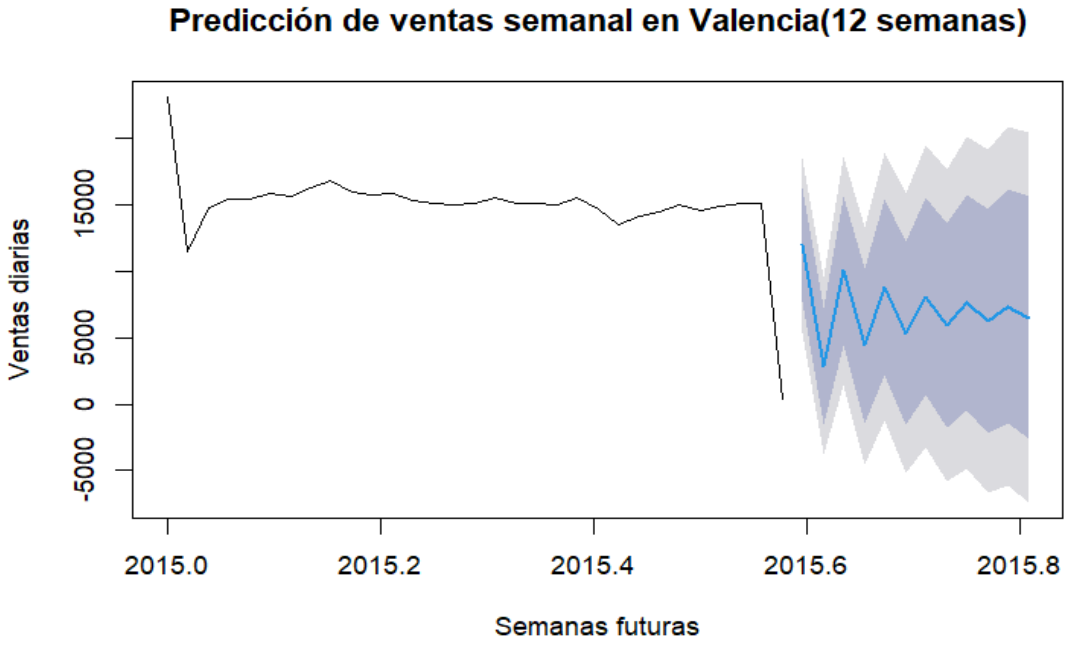

Predicción de ventas semanal en Barcelona(12 semanas)
Ventas diarias
0
10000
20000
30000
40000
2015.0
2015.2
2015.4
2015.6
2015.8
Semanas futuras

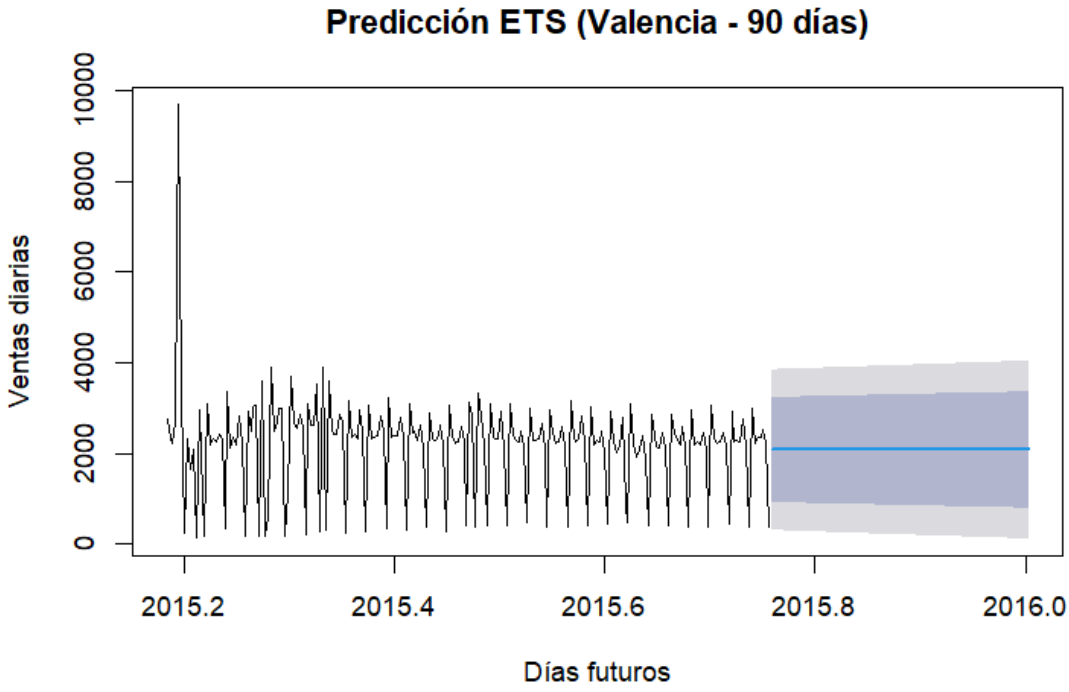

Predicción ETS (Barcelona - 90 días)
Ventas diarias
0
2000
6000
10000
2015.2
2015.4
2015.6
2015.8
2016.0
Días futuros

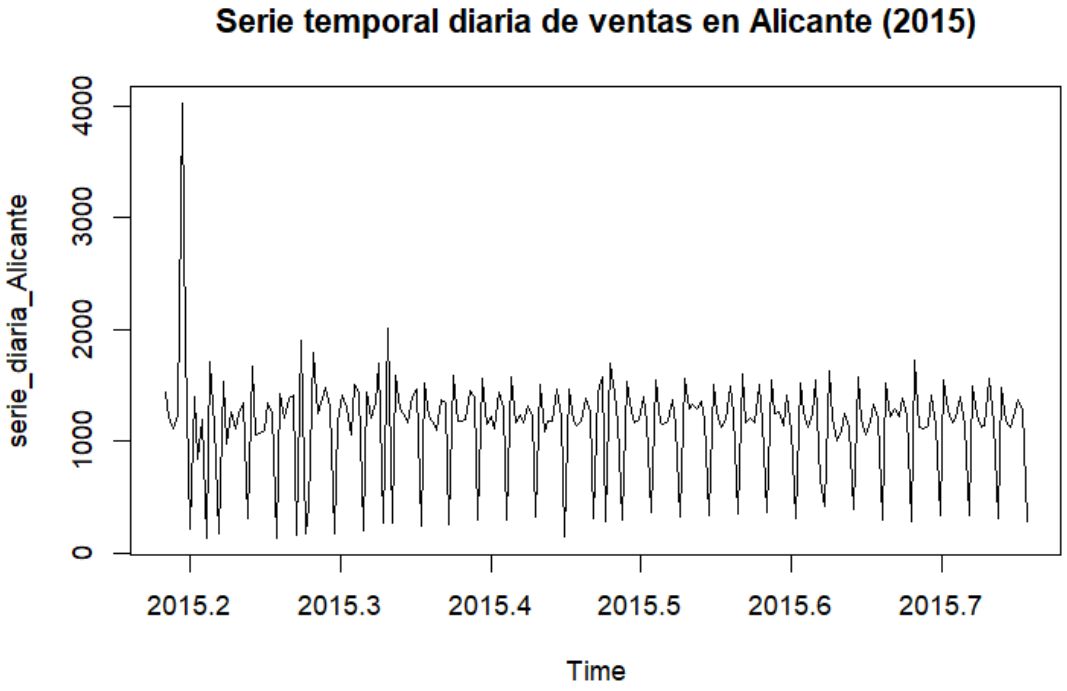

Serie temporal diaria de ventas en Valencia (2015)
serie_diaria_Valencia
0
2000
4000
6000
8000
10000
2015.2
2015.3
2015.4
2015.5
2015.6
2015.7
Time

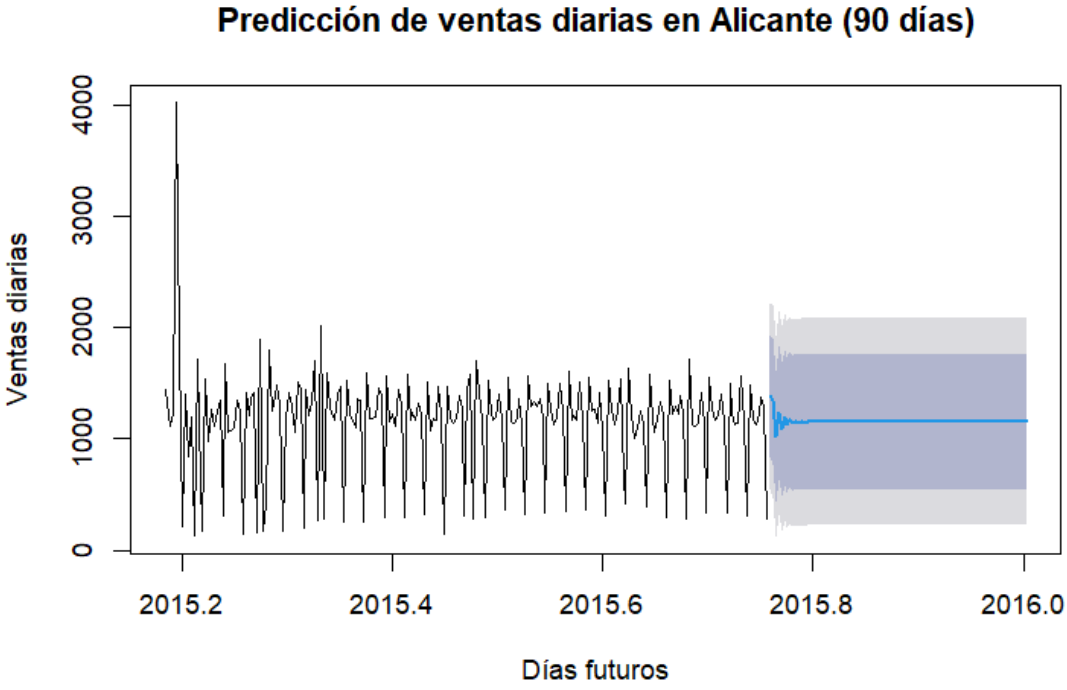

Predicción de ventas diarias en Valencia (90 días)
Ventas diarias
0
2000
4000
6000
8000
10000
2015.2
2015.4
2015.6
2015.8
2016.0
Días futuros

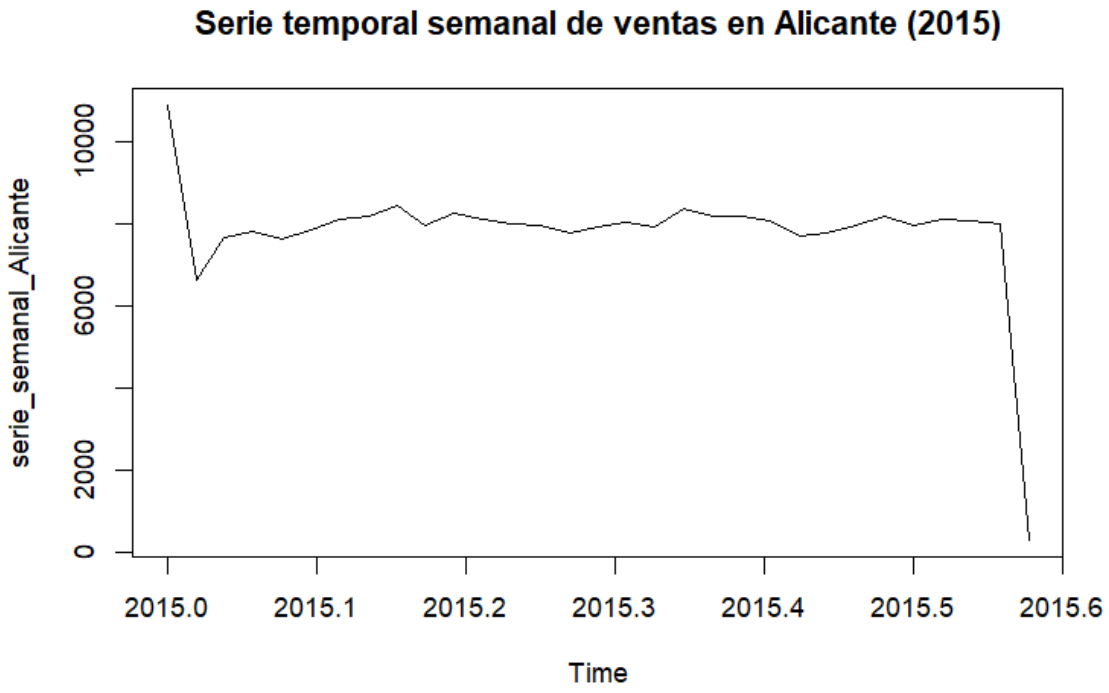

Serie temporal semanal de ventas en Valencia (2015)
serie_semanal_Valencia
0
5000
10000
20000
2015.0
2015.1
2015.2
2015.3
2015.4
2015.5
2015.6
Time

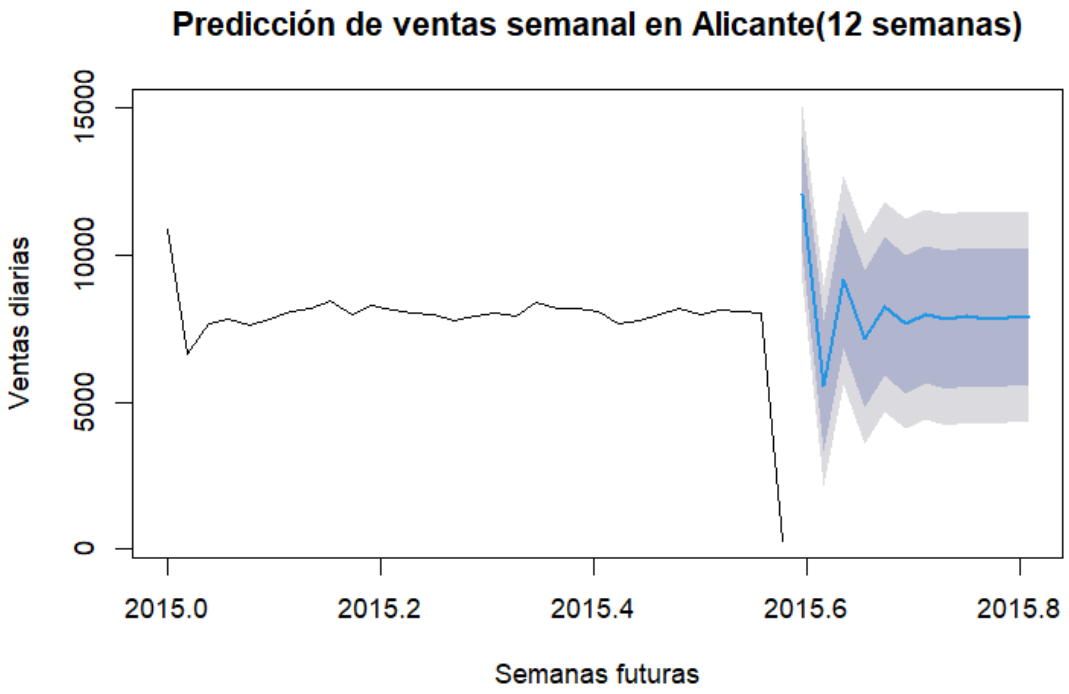

Predicción de ventas semanal en Valencia(12 semanas)
Ventas diarias
-5000
0
5000
15000
2015.0
2015.2
2015.4
2015.6
2015.8
Semanas futuras

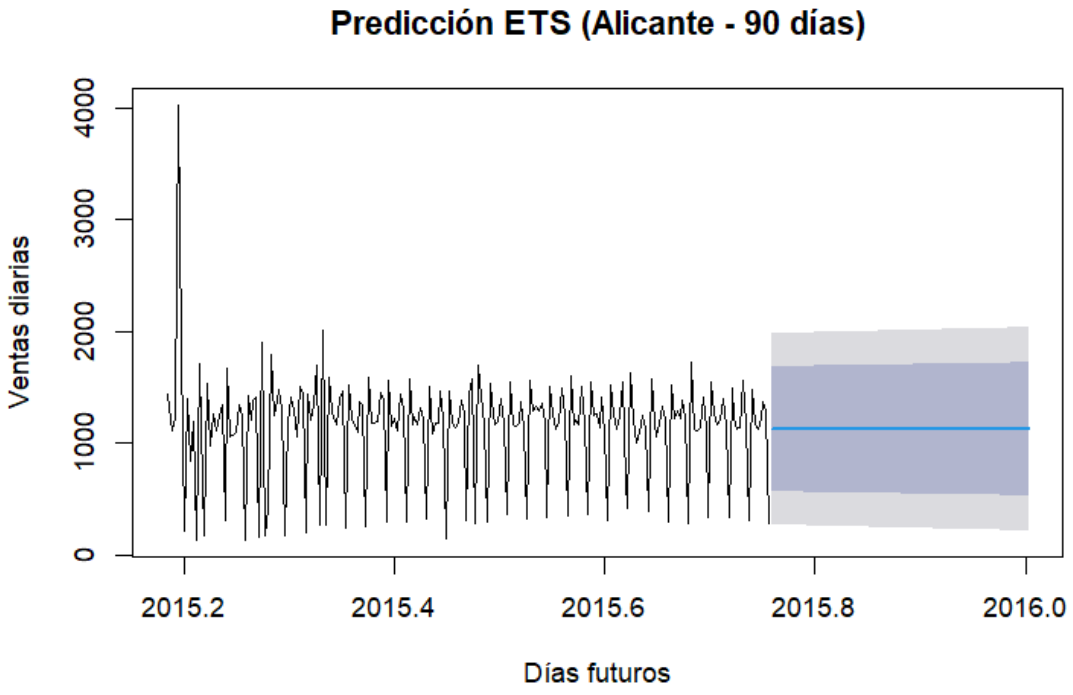

Predicción ETS (Valencia - 90 días)
Ventas diarias
0
2000
4000
6000
8000
10000
2015.2
2015.4
2015.6
2015.8
2016.0
Días futuros

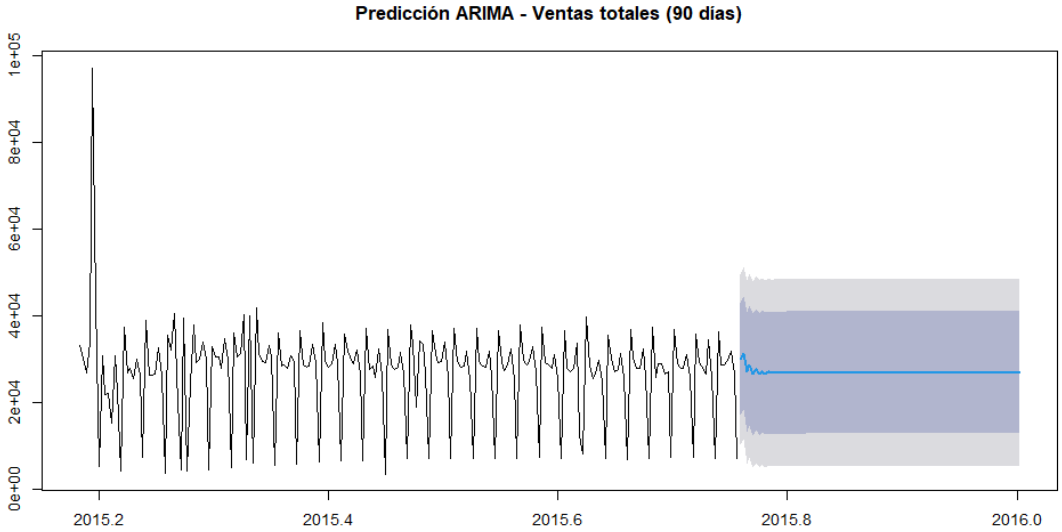

Serie temporal diaria de ventas en Alicante (2015)
serie_diaria_Alicante
0
1000
2000
3000
4000
2015.2
2015.3
2015.4
2015.5
2015.6
2015.7
Time

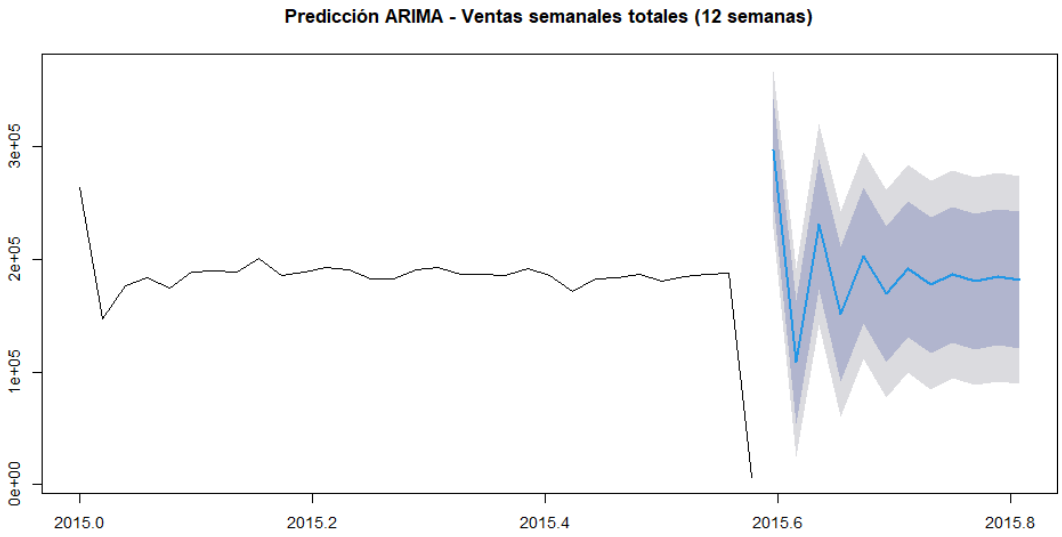

Predicción de ventas diarias en Alicante (90 días)
Ventas diarias
0
1000
2000
3000
4000
2015.2
2015.4
2015.6
2015.8
2016.0
Días futuros

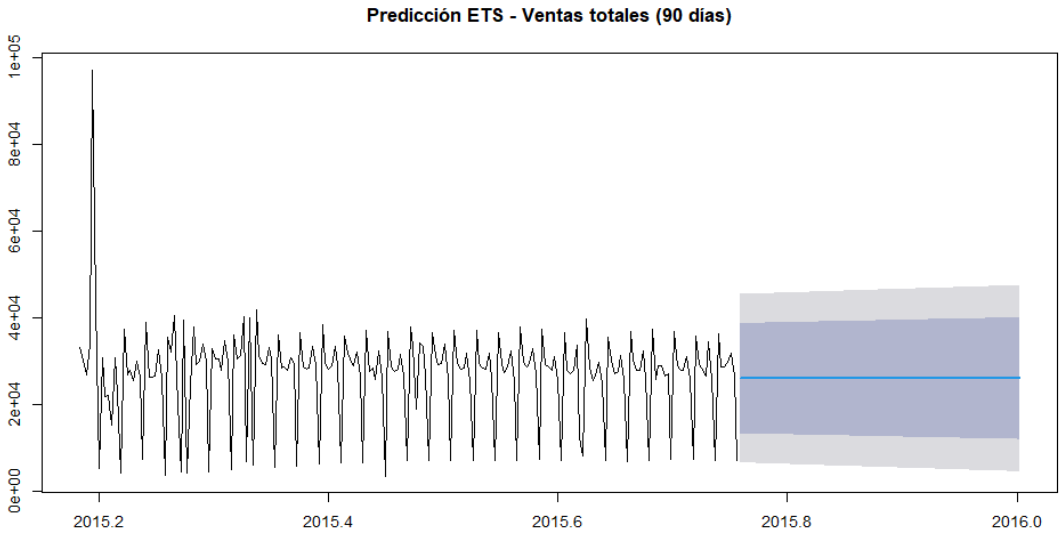

Serie temporal semanal de ventas en Alicante (2015)
serie_semanal_Alicante
0
2000
6000
10000
2015.0
2015.1
2015.2
2015.3
2015.4
2015.5
2015.6
Time


Predicción de ventas semanal en Alicante(12 semanas)
Ventas diarias
0
5000
10000
15000
2015.0
2015.2
2015.4
2015.6
2015.8
Semanas futuras


Predicción ETS (Alicante - 90 días)
Ventas diarias
0
1000
2000
3000
4000
2015.2
2015.4
2015.6
2015.8
2016.0
Días futuros


Predicción ARIMA - Ventas totales (90 días)
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
2015.2
2015.4
2015.6
2015.8
2016.0


Predicción ARIMA - Ventas semanales totales (12 semanas)
0e+00
1e+05
2e+05
3e+05
2015.0
2015.2
2015.4
2015.6
2015.8


Predicción ETS - Ventas totales (90 días)
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
2015.2
2015.4
2015.6
2015.8
2016.0

Anexo F. Elaboración Dashboard

MODELO ESTRELLA

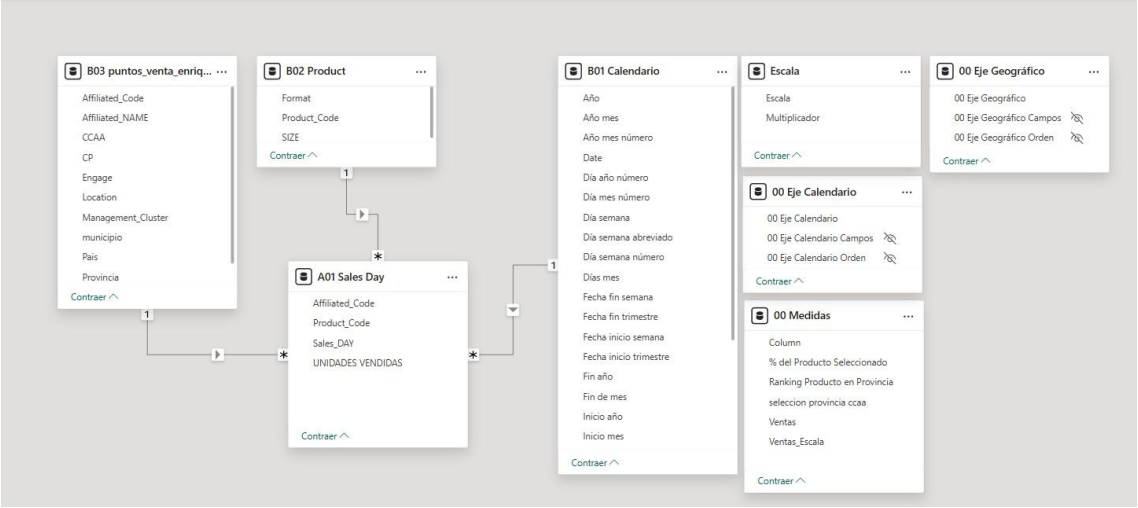


TABLA CALENDARIO

1 B01 Calendario =

Date	Año	Inicio año	Fin año	Número mes	Inicio mes	Fin de mes	Días mes	Año mes número	Año mes	Día mes número	Día semana	Día semana abr
01/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	1	jueves	ju.
02/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	2	viernes	vi.
03/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	3	sábado	sá.
04/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	4	domingo	do.
05/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	5	lunes	lu.
06/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	6	martes	ma.
07/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	7	miércoles	mi.
08/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	8	jueves	ju.
09/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	9	viernes	vi.
10/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	10	sábado	sá.
11/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	11	domingo	do.
12/01/15	2015	01/01/2015 0:00:00	31/12/2015 0:00:00	1	01/01/2015 0:00:00	31/01/2015 0:00:00	31	201501	ene-15	12	lunes	lu.

1 B01 Calendario =

Día semana número	Día año número	Mes	Mes abreviado	Trimestre número	Trimestre	Trimestre año número	Trimestre año	Fecha inicio trimestre	Fecha fin trimestre	Número
5	1	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
6	2	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
7	3	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
1	4	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
2	5	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
3	6	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
4	7	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
5	8	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
6	9	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
7	10	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
1	11	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
2	12	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
3	13	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
4	14	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
5	15	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
6	16	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	
7	17	enero	ene	1	1T	20151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	

	Trimestre año	Fecha inicio trimestre	Fecha fin trimestre	Número semana año	Fecha inicio semana	Fecha fin semana
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	1	28/12/2014 0:00:00	03/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	1	28/12/2014 0:00:00	03/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	1	28/12/2014 0:00:00	03/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	2	04/01/2015 0:00:00	10/01/2015 0:00:00
)151	2015-1T	01/01/2015 0:00:00	31/03/2015 0:00:00	3	11/01/2015 0:00:00	17/01/2015 0:00:00

PARAMETROS

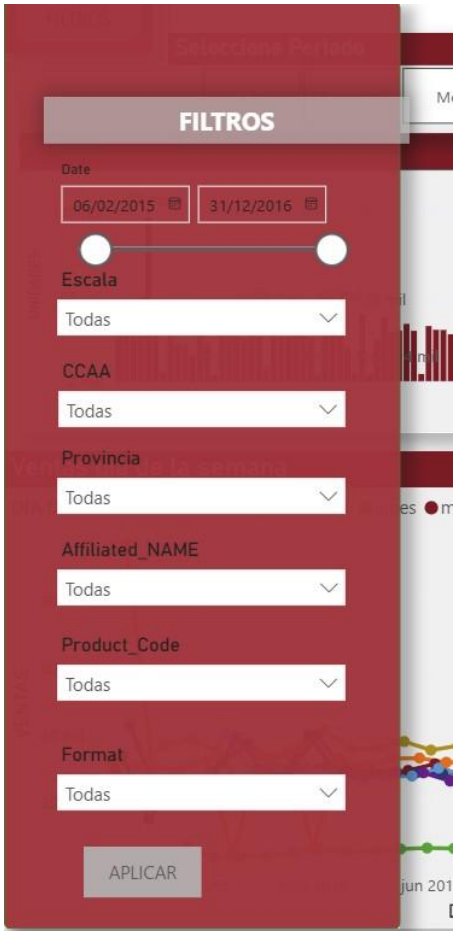
```
00 Eje Calendario = {
    ("Año", NAMEOF('B01 Calendario'[Año]), 0),
    ("Trimestre", NAMEOF('B01 Calendario'[Trimestre año]), 1),
    ("Mes", NAMEOF('B01 Calendario'[Año mes]), 2),
    ("Día", NAMEOF('B01 Calendario'[Date]), 3)
}
```

00 Eje Calendario	00 Eje Calendario Campos	00 Eje Calendario Orden
Año	'B01 Calendario'[Año]	0
Trimestre	'B01 Calendario'[Trimestre año]	1
Mes	'B01 Calendario'[Año mes]	2
Día	'B01 Calendario'[Date]	3

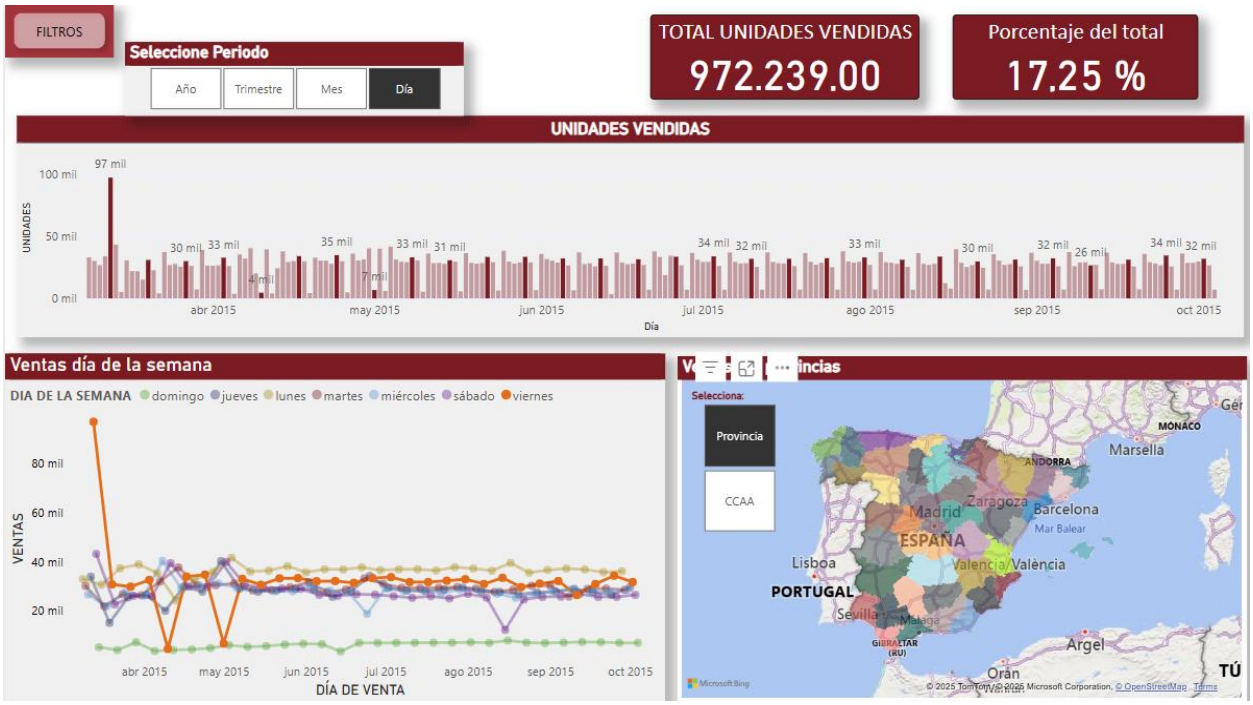
```
[00 Eje Geográfico ] = {
    ("Provincia", NAMEOF('B03 puntos_venta_enriquecido2_sinNA'[Provincia]), 0),
    ("CCAA", NAMEOF('B03 puntos_venta_enriquecido2_sinNA'[CCAA]), 1)
}
```

00 Eje Geográfico	00 Eje Geográfico Campos	00 Eje Geográfico Orden
Provincia	'B03 puntos_venta_enriquecido2_sinNA'[Provincia]	0
CCAA	'B03 puntos_venta_enriquecido2_sinNA'[CCAA]	1

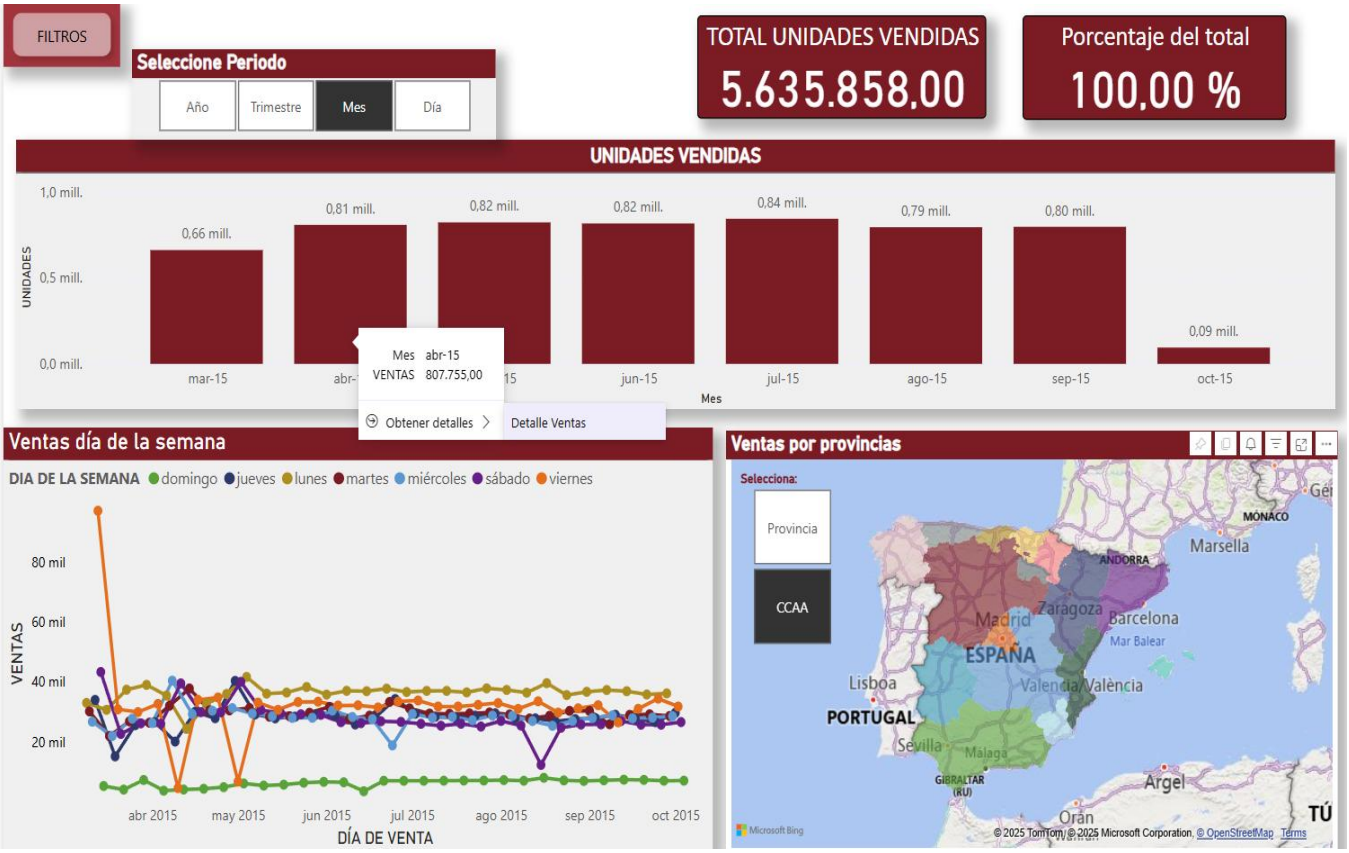
MENÚ DE FILTROS

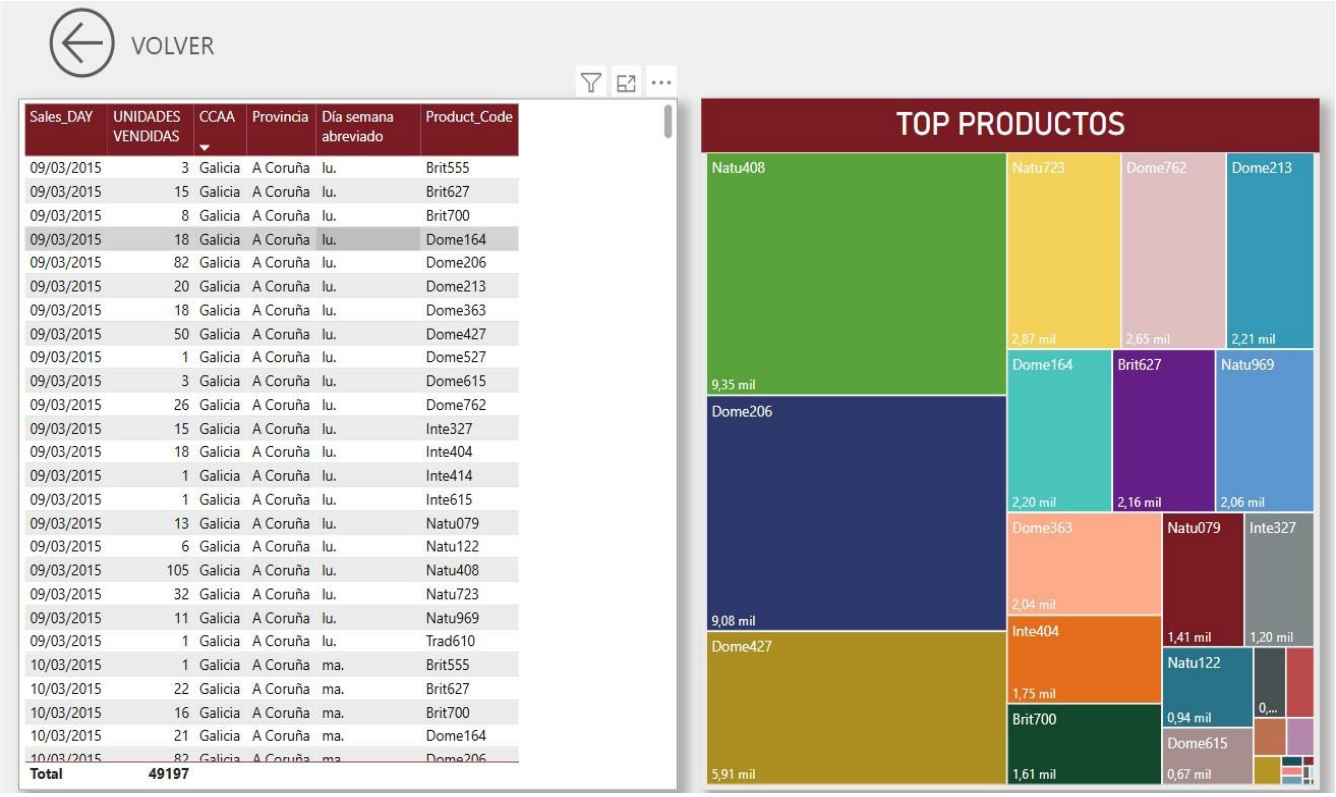


PANEL ANÁLISIS VENTAS INTERACTIVO



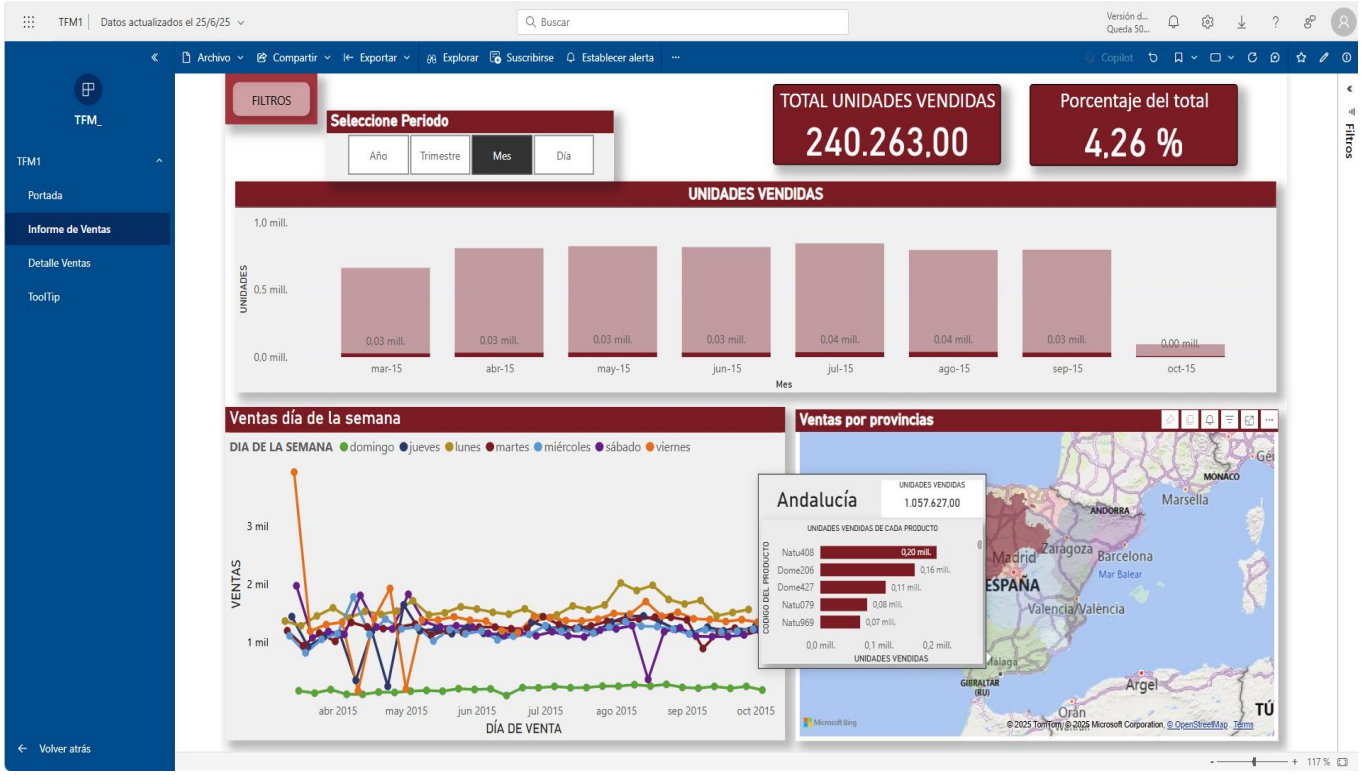
PAGINA DETALLE VENTAS (donde te dirige pinchando sobre detalle de ventas)

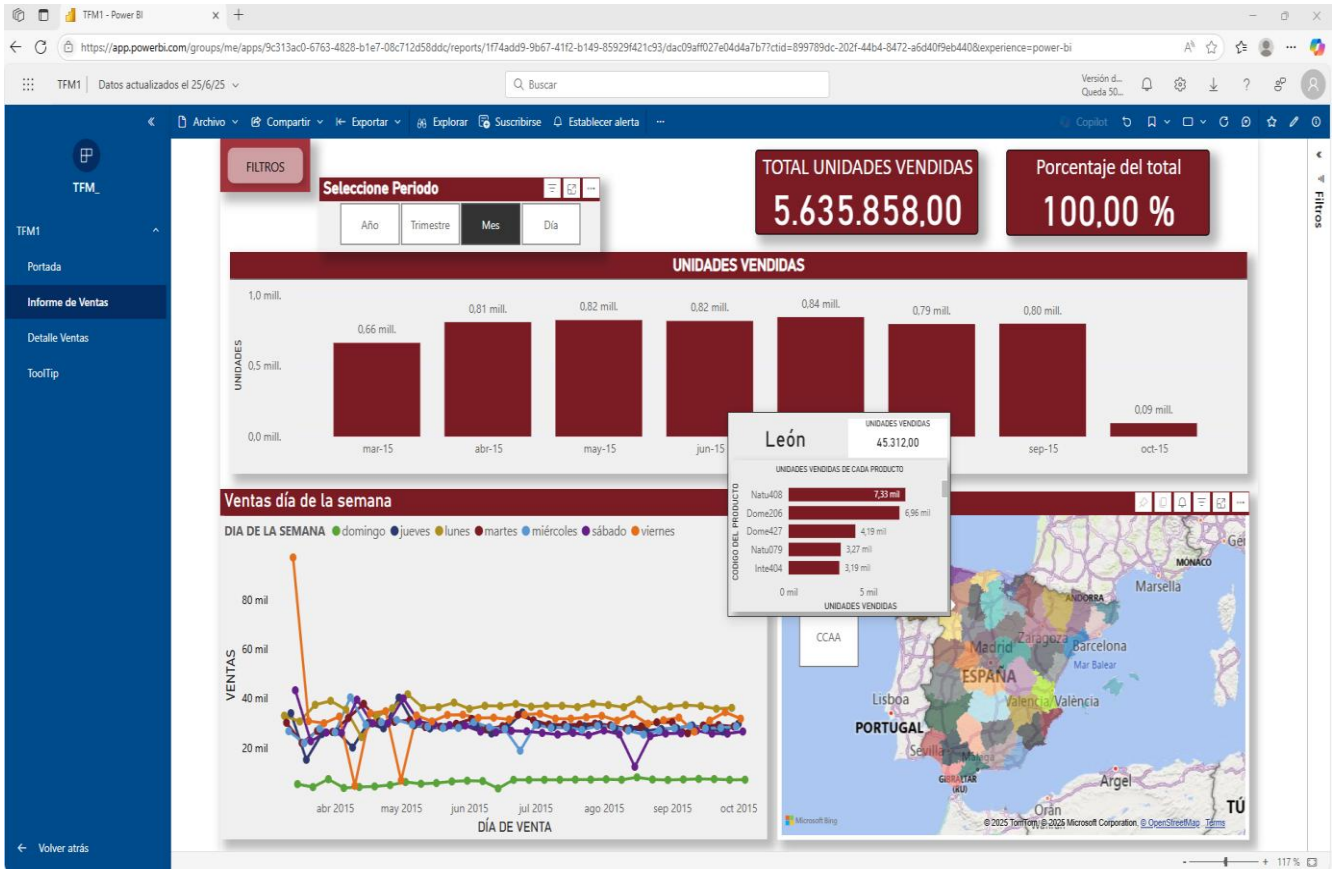




Sales_DAY	UNIDADES VENDIDAS	CCAA	Provincia	Día semana abreviado	Product_Code
09/03/2015	3	Galicia	A Coruña	lu.	Brit555
09/03/2015	15	Galicia	A Coruña	lu.	Brit627
09/03/2015	8	Galicia	A Coruña	lu.	Brit700
09/03/2015	18	Galicia	A Coruña	lu.	Dome164
09/03/2015	82	Galicia	A Coruña	lu.	Dome206
09/03/2015	20	Galicia	A Coruña	lu.	Dome213
09/03/2015	18	Galicia	A Coruña	lu.	Dome363
09/03/2015	50	Galicia	A Coruña	lu.	Dome427
09/03/2015	1	Galicia	A Coruña	lu.	Dome527
09/03/2015	3	Galicia	A Coruña	lu.	Dome615
09/03/2015	26	Galicia	A Coruña	lu.	Dome762
09/03/2015	15	Galicia	A Coruña	lu.	Inte327
09/03/2015	18	Galicia	A Coruña	lu.	Inte404
09/03/2015	1	Galicia	A Coruña	lu.	Inte414
09/03/2015	1	Galicia	A Coruña	lu.	Inte615
09/03/2015	13	Galicia	A Coruña	lu.	Natu079
09/03/2015	6	Galicia	A Coruña	lu.	Natu122
09/03/2015	105	Galicia	A Coruña	lu.	Natu408
09/03/2015	32	Galicia	A Coruña	lu.	Natu723
09/03/2015	11	Galicia	A Coruña	lu.	Natu969
09/03/2015	1	Galicia	A Coruña	lu.	Trad610
10/03/2015	1	Galicia	A Coruña	ma.	Brit555
10/03/2015	22	Galicia	A Coruña	ma.	Brit627
10/03/2015	16	Galicia	A Coruña	ma.	Brit700
10/03/2015	21	Galicia	A Coruña	ma.	Dome164
10/03/2015	82	Galicia	A Coruña	ma.	Dome206
Total	**49197**				

TOOLTIP DEL MAPA (que va cambiando según nos movemos por el mapa y seleccionamos CCAA o Provincia)





A Coruña
UNIDADES VENDIDAS
5.635.858,00
UNIDADES VENDIDAS DE CADA PRODUCTO
CODIGO DEL PRODUCTO
Natu408
1,04 mill.
Dome206
0,71 mill.
Dome427
0,50 mill.
Natu079
0,42 mill.
Natu969
0,33 mill.
0,0 mill.
0,5 mill.
1,0 mill.
UNIDADES VENDIDAS