Universidad Internacional de La Rioja
Facultad de Empresa, Comunicación y Marketing

Máster Universitario en Inteligencia de Negocio

Inteligencia de Negocio aplicada a la Segmentación de clientes: Modelo RFM y Análisis de Clúster

Trabajo fin de estudio presentado por:	Gregorio Manuel Intriago Solorzano
Tipo de trabajo:	Proyecto de Inteligencia de Negocio
Modalidad:	Individual
Director/a:	Juan Manuel Martin Álvarez
Fecha:	Julio 2022

Resumen

La segmentación de mercados tiene una serie de beneficios para las empresas comerciales y ofrece la oportunidad de pensar y repensar sobre las mejores estrategias enfocadas a sus clientes, y conduce a nuevos conocimientos y perspectivas fundamentales de su negocio basada en datos.

Por otro lado, los datos de transacciones de facturación que se registran en cualquier compañía comercial cuentan con la información básica y necesaria para establecer las principales características que describen el comportamiento de los consumidores, por ejemplo, determinar cuán reciente es la transacción de un cliente, la frecuencia de compras en un período de tiempo y su magnitud financiera de las transacciones.

En este sentido, el presente trabajo tiene como objetivo principal generar una segmentación de clientes de una empresa de comercialización y distribución de productos de consumo masivo en el Ecuador, a través de la aplicación del análisis descriptivo de datos, el método RFM (Recency, Frecuency y Monetary) y el análisis no supervisado de clúster mediante el modelo de partición K-means, al conjunto de datos que corresponden a los pedidos de productos de la compañía en un período de referencia.

Finalmente, se creará una solución de inteligencia de negocio dinámica, actualizada e integrada, que viabilice la visualización y el análisis de datos de los clientes para la generación de información y conocimiento del negocio para la toma de decisiones basada en datos.

Palabras clave: Segmentación de Clientes, Modelo RFM, Análisis de Clúster.

Abstract

Market segmentation has a number of benefits for trading companies and offers the opportunity to think and rethink about the best strategies to focus on their customers, and leads to fundamental new data-driven insights into their business.

On the other hand, the billing transaction data that is recorded in any commercial company has the basic and necessary information to establish the main characteristics that describe consumer behavior, for example, determining how recent a customer's transaction is, the frequency of purchases over a period of time and their financial magnitude of transactions.

In this sense, the main objective of this work is to generate a customer segmentation of a marketing and distribution company of mass consumption products in Ecuador, through the application of descriptive data analysis, the RFM (Recency, Frequency and Monetary) method and unsupervised cluster analysis using the K-means partition model, to the data set corresponding to the company's product orders in a period of time.

Finally, a dynamic, up-to-date and integrated business intelligence solution will be created to enable the visualization and analysis of customer data for the generation of business information and knowledge for data-driven decision making.

Keywords: Customer Segmentation, RFM Model and Cluster Analysis

Índice de contenidos

Índice de figuras

Índice de tablas

1. Introducción

En el entorno empresarial actual, la innovación y el desarrollo de la tecnología avanzan a pasos gigantescos influyendo de forma directa sobre las necesidades y deseos de los clientes, de tal manera, que las empresas deben ser flexibles para adaptarse a los nuevos requerimientos de los consumidores y la forma de satisfacerlos, con el claro objetivo de incrementar clientes fieles a las marcas de una empresa.

Es de vital importancia para las empresas aprovechar la información para realizar el análisis de negocio, permitiendo la toma de decisiones basadas en los datos y tener una visión integral de la organización, implementar la inteligencia de negocio en la organización y analizar su información, conlleva a las compañías ser más competitivas, incrementar sus ganancias y de fidelizar a sus clientes, al contar con un sistema de apoyo posibilitará mitigar las decisiones basadas en la opinión.

La segmentación del mercado tiene una serie de beneficios. En el nivel más general, la segmentación del mercado obliga a las organizaciones a hacer un balance de dónde se encuentran y dónde quieren estar en el futuro. Al hacerlo, obliga a las organizaciones a reflexionar sobre aquello en lo que son particularmente buenas en comparación con la competencia, y hacer un esfuerzo por obtener información sobre lo que quieren los consumidores. La segmentación del mercado ofrece la oportunidad de pensar y repensar, y conduce a nuevos conocimientos y perspectivas fundamentales (Dolnicar, Grun, & Leisch, 2018, p. 8).

Las diferentes transacciones de facturación que se registran en el tiempo tienen los datos necesarios para establecer las principales variables descriptoras del comportamiento de los clientes. Basados en dichos datos de compras es posible investigar y determinar cuán reciente es la transacción de un cliente, así como la frecuencia y la magnitud de sus transacciones en un período de referencia.

La combinación de las tres variables: recencia, frecuencia y magnitud, permite conocer en cualquier momento, el tipo de comportamiento de un cliente, y también si esta conducta tiene una tendencia positiva o negativa, dentro de la migración esperada para el cliente a través de las fases del ciclo de vida. El resultado de dicha composición de variables conforma el modelo RFM (Recencia o Actualidad, Frecuencia y Magnitud o Valor Monetario).

El modelo RFM es una metodología con la que podemos segmentar clientes según características similares entre sí, en base a cuándo fue su última compra, con qué frecuencia han comprado en el pasado y cuánto han gastado en total, con el objetivo de incrementar las ventas en la empresa. (Glotzer, Simla, 2022).

Por otro lado, el análisis de clúster con el método de "K-means" o "K-medias" es un método no jerárquico o no supervisado, con un enfoque muy popular para la clasificación debido a su simplicidad de implementación y rápida ejecución y ha sido ampliamente utilizado en la segmentación del mercado, el reconocimiento de patrones, la recuperación de información, entre otros usos en el ámbito empresarial (Wu, Chang, & Lo, 2009) (Palakshappa & Patil, 2022).

En este sentido, el presente Trabajo Final de Máster plantea la realización de una segmentación de clientes de una empresa de comercialización y distribución de productos de consumo masivo en el Ecuador, basada en la combinación de métodos RFM y K-means. El análisis del modelo RFM permitirá agruparlos por sus hábitos de compra y apreciar cuales son los mejores clientes, lo más rentables, en riesgo de pérdida; mientras que con el análisis de clúster con el método de "K-medias" se definirá diferentes nuevos subsegmentos de clientes que permitirán distinguir cuántos tipos de clientes tiene el negocio y cómo son con respecto a su hábito de compra.

1.1. Planteamiento general

1.1.1. Justificación

Según Cuadros (2017), en su artículo científico dedicado al análisis multivariado para la segmentación de clientes basado en RFM, establece un lineamiento para construir una exitosa relación con los clientes, menciona que las empresas deben identificar el verdadero valor de los clientes, ya que esto proporciona información básica para implementar estrategias de marketing dirigidas.

El entorno empresarial es cada vez más competitivo e identificar a los clientes más rentables, frecuentes o nuevos, sin duda es algo que las empresas deben tener muy claro para poder enfocar sus esfuerzos entendiendo que no todos los clientes necesitan la misma importancia y estrategias de marketing.

La definición de este nivel de importancia se evalúa con técnicas de segmentación de clientes e identificación de los clientes más rentables, frecuentes, nuevos, VIP, en riesgo de pérdida, los que no realizan compras, entre otros grupos, esta información permitirá focalizar los esfuerzos y recursos para maximizar su valor.

El análisis RFM es un punto inicial para identificar el comportamiento de clientes para la toma de decisiones de las empresas comerciales, el mismo que puede ser fortalecido con técnicas multivariantes como es el análisis de clúster con el método de k-medias.

La empresa comercializadora y distribuidora de productos de consumo masivo en el Ecuador con el afán de dar cumplimiento a sus objetivos estratégicos de incrementar las ventas y fidelizar a sus clientes, considera oportuno y necesario en realizar e identificar a sus clientes por medio de un modelo de segmentación aplicando el análisis RFM y una clusterización de los puntajes obtenidos en el modelo RFM mediante el método no supervisado de k-medias.

1.1.2. Planteamiento del problema

La compañía en estudio actualmente realiza análisis de información descriptiva mediante la generación de consultas periódicas a la base de datos y de reportes en hojas de cálculo Excel en base a la necesidad de las diferentes áreas de negocio de la empresa, considerando el uso de una o dos variables de interés como máximo.

El generar las consultas y los reportes estáticos les conlleva un tiempo significativo a la compañía, perdiendo así dar respuesta inmediata y tener control del comportamiento de sus clientes, para reformular o fortalecer las estrategias de mercadeo a los distintos segmentos de compradores.

La empresa tiene una idea general de cómo es el comportamiento de los clientes considerando los tabulados univariantes y bivariantes que son de forma estática e inoportuna, sin embargo, la toma de decisiones es basada en la opinión de los jefes antiguos y personal de experiencia de mandos medio, por ello nace la necesidad de obtener información útil y actualizada a través del uso de los datos con la aplicación de modelos de segmentación multivariante para identificar, visualizar y entender el comportamiento de sus clientes que permitan canalizar los esfuerzos de la compañía para maximizar su ganancia y fidelizar a sus consumidores.

1.2. Objetivos del TFE

1.2.1. Objetivo General

Analizar y visualizar la información comercial de una empresa comercializadora y distribuidora del Ecuador, mediante la aplicación de modelos de segmentación de clientes, que aporte información útil sobre el valor y status de los clientes, para la toma de decisiones para la compañía.

1.2.2. Objetivos específicos

- Generar un diagnóstico estadístico descriptivo y comercial de los datos de facturación de los clientes proporcionada por la empresa.
- Realizar un modelo de segmentación de clientes que permita conocer cuáles son los clientes más y menos rentables para la compañía, segmentándolos por características similares de hábitos de compra mediante el análisis de Recencia, Frecuencia, Valor monetario (RFM) y el modelo de K-medias.
- Implementar una solución de Inteligencia de Negocio para la generación de informes interactivos, que permita la visualización y estudio de los segmentos de clientes de la compañía para la toma de decisiones y dote a la empresa de una ventaja competitiva.

1.3. Elementos innovadores del proyecto

El contar con modelos de segmentación de clientes combinando los análisis de Recencia, Frecuencia, Valor monetario (RFM) y de Clúster a través del modelo de K-medias permitirá contar con información técnica para la toma de decisiones, las cuales, actualmente, se toman de acuerdo con el criterio y perfiles diseñados por la gerencia, los cuales son el resultado de años de trabajo y al conocimiento adquirido en todo ese tiempo. Sin embargo, el comportamiento del cliente cambia de manera exponencial, lo cual evidencia vulnerabilidades en el modelo experto vigente y crea la necesidad de trabajar con un modelo empírico.

Actualmente, todos los reportes de datos se generan en la herramienta Excel de forma estática y tienen un alto componente de manejo humano, lo cual genera un riesgo operativo que puede afectar al resultado de la decisión. La intención es minimizar también este riesgo, automatizando el proceso para que de esta manera sea más eficiente.

La herramienta de inteligencia de negocio funcionará en línea, es decir, el resultado será obtenible en el mismo momento de la consulta. Para esto, se incorporarán consultas en lenguaje SQL de los datos de pedidos y facturación, el modelo RFM y K-medias en R Studio y un panel de control en Power BI que permita visualizar el resultado.

2. Alcance y planificación

Según lo señala Sherman (2014), la práctica recomendable para empezar un proyecto de Inteligencia de Negocios implica la evaluación de este, lo cual permitirá dar respuestas a las preguntas: ¿Cuál es el alcance? ¿Qué es la mejor manera de proceder? ¿Cuánto tiempo tardará su ejecución? ¿Cuánto costará? ¿Quién necesita estar involucrado? ¿Qué productos necesitamos?, y así mantener una línea de trabajo que minimice los imprevistos.

Esta fase implica definir el alcance del proyecto en términos de requisitos generales, plazos, recursos y costes. Incluye crear el presupuesto del proyecto y obtener los recursos para llevarlo a cabo. Una decisión clave será elegir quién llevará a cabo el proyecto, un equipo interno, uno externo, o un mixto de los dos.

El resultado de esta evaluación es establecer una hoja de ruta para alcanzar los objetivos estratégicos de la organización, de una manera que sea lo más eficiente posible en cuanto a costos y recursos. Así mismo, ayudará a desarrollar el alcance y el diseño.

Con información detallada sobre comentarios y prioridades, puede refinar aún más el cronograma del proyecto, los planes de recursos y los entregables, mientras mitiga el riesgo. Utilizará esta información para evaluar el estado de la empresa en su programa de BI, dónde debería estar y cuáles son sus capacidades actuales. Estas observaciones ayudarán a refinar los objetivos en términos de proyecto, personas, procesos y tecnología (Sherman, 2014).

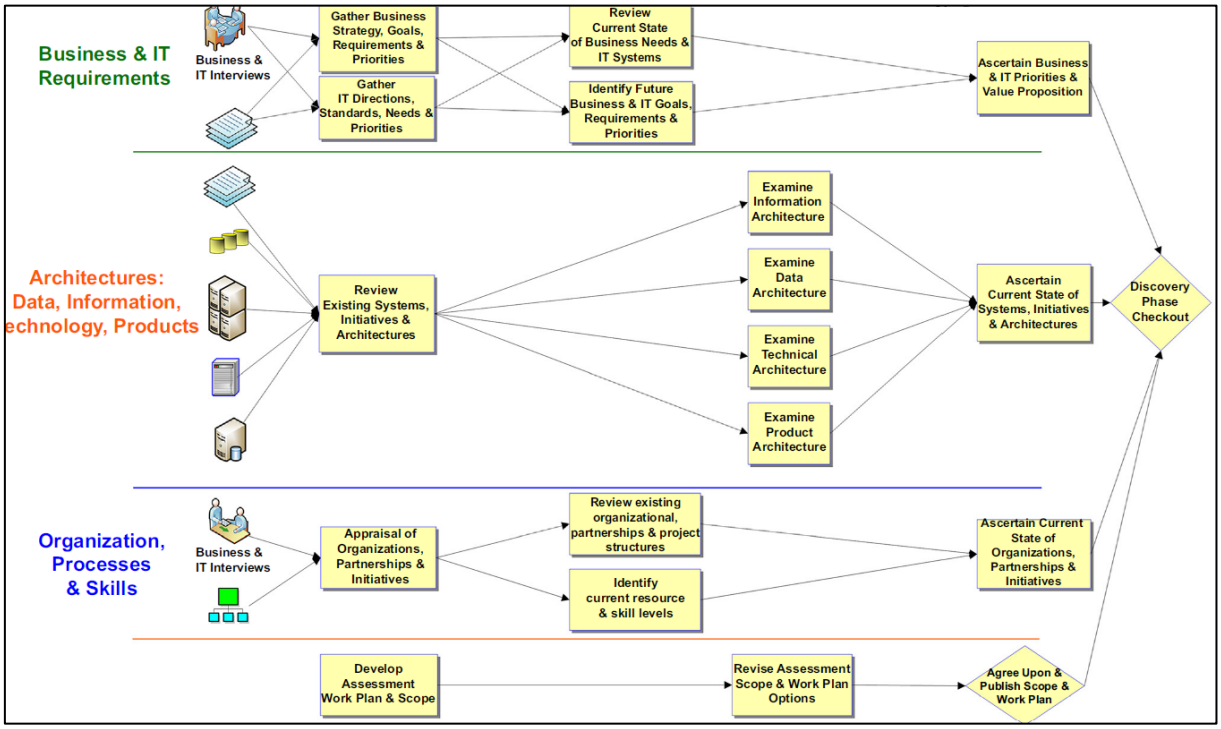
Las fases que se deben seguir en la evaluación previa son tres: evaluación, análisis y recomendaciones. Estas tres fases han de servir de ayuda para desarrollar el diseño y alcance del proyecto, analizar la situación actual de la organización y establecer cuál es el potencial de ésta.

2.1. Fase de descubrimiento: evaluación del entorno actual

La fase de descubrimiento consiste en analizar los requisitos comerciales y de los sistemas de información, las skills de los miembros de la organización que intervienen y la arquitectura de

datos y tecnología (Ver *Figura 1*), mediante la cual, permita identificar los informes y análisis deseados; los informes actuales, las deficiencias de datos y soluciones alternativas; prioridades de la compañía; y las habilidades analíticas con las que cuentan (Sherman, 2014).

Figura 1. *Evaluación de inteligencia de negocios: Fase de descubrimiento*



Fuente: Business intelligence guidebook : from data integration to analytics

2.1.1. Información deseada

En esta sección, se establecerá que herramientas de decisión se requieren en la compañía y que datos se precisan conocer. En este caso, se ha detectado que la compañía de logística y distribución ecuatoriana utiliza herramientas de análisis inadecuado para el tamaño e importancia de la empresa, actualmente realiza reportes univariantes y bivariantes en hojas de cálculo Excel y el uso básico de la herramienta Qlik View.

El análisis de información no está basado en sus objetivos estratégicos, y se les dificulta el cruzar la información, en razón que no poseen la data integrada y unificada, cada departamento trabaja y administra sus datos de forma individual según es de su conveniencia y no existe gobernanza en la misma.

Después de realizar el análisis en la organización y apalancado en el objetivo estratégico planteado para el año 2022 el cual establece como proyecto primordial el crecimiento en ventas y la fidelización de clientes; se ha determinado como necesidad conocer el

comportamiento de sus clientes y crear segmentos de consumo por medio del método RFM que evaluará la **Recencia:** proximidad de la última compra, **Frecuencia:** frecuencia de compra y **Monetización**: valor monetario de las compras (Birant, 2011) (Chen, Sain, & Guo, 2012).

Con el análisis de RFM permitirá realizar una segmentación de clientes y encontrar respuestas a las siguientes interrogantes:

- ¿Quiénes son mis mejores clientes?
- ¿Qué clientes están en riesgo de perderse?
- ¿Quién tiene el potencial para convertirse en clientes más rentables?
- ¿Quiénes son los clientes perdidos a los que no necesitas prestar mucha atención?
- ¿Qué clientes deben retenerse?
- ¿Quiénes son los clientes leales?
- ¿Qué grupo de clientes es más probable que responda a su campaña actual?

Adicionalmente, se pretende llevar a cabo un análisis clúster mediante el modelo de k-medias que permita formar una nueva segmentación de un grupo de clientes en base a los tres parámetros establecidos (Recencia, Frecuencia y Monetización). Con ello, se pretende extraer información más detallada sobre el comportamiento de los clientes que resulte útil para la toma de decisiones, principalmente aquellas relacionadas con el marketing relacional.

Al realizar estos análisis se pretende conocer de manera más real el comportamiento de compra de los clientes para así apalancar en estrategias comerciales y de marketing, para alcanzar el objetivo trazado por la compañía.

2.1.2. Información actual: deficiencias y soluciones alternativas.

La empresa posee una gran cantidad de información de sus clientes y negocio, sin embargo, este gran volumen de información se ha vuelto una problemática por la falta de un buen manejo de los datos, su extracción, procesado y presentación.

Por medio del análisis y entrevistas con los lideres de negocios se ha detectado varias deficiencias en la información que posee la compañía, entre las cuales resaltan:

- El análisis de información se enfoca sólo en realizar reportes y no brindan información que permitan tomar decisiones estratégicas. El asunto de la reportería no es de prioridad para la empresa, en razón que tiene una duración de un mes y cuando ha

sido elaborada la información que se presenta, la misma es desactualizada y no es relevante para la toma de decisiones.

- La mayoría de los reportes son procesados en Excel, no cuentan con personal especializado para realizar análisis en herramientas especializadas para el tratamiento y análisis de datos como Python o R.
- Cada departamento maneja sus datos como mejor le parezca, no poseen una base de datos centralizada o una tabla máster.
- No existe un tratamiento adecuado de los datos ni políticas de calidad en los mismos, por lo general la tabla de ventas y clientes muestra registros duplicados, inexistentes o con datos faltantes.
- Falta de conexión de datos entre herramientas para el tratamiento y análisis de datos como Python o R con la herramienta de BI que posee la empresa para construir e implementar una herramienta de informes interactivos que permitan disponer de información actualizada y que sirva para detectar desviaciones

Para el presente proyecto, se plantea hacer un proceso de ETL, consolidar la información en una tabla única, realizar la segmentación mediante el análisis RFM y de Clúster, y generar la conexión de información útil entre las herramientas R y Power BI que viabilice la generación de reportes dinámicos e interactivos, con el fin de dar cumplimiento al objetivo estratégico de incrementar las ventas y fidelizar clientes.

2.1.3. Habilidades analíticas actuales.

Las habilidades analíticas con las que cuenta la empresa para el registro, almacenamiento y consulta de los datos son las necesarias para el funcionamiento de la compañía, sin embargo, no se evidencia el uso de herramientas para el tratamiento y análisis masivo de datos[1], así como la falta de personal especializado y capacitado para el uso de las nuevas tecnologías de inteligencia empresarial como Power BI. Esto hace que el tiempo invertido en cálculo de indicadores y generación de reportes sea más extenso lo cual limita hasta cierto punto el proceso de toma de decisiones data - driven.

[1] La empresa cuenta con la herramienta Excel licenciada para la generación de reportes de información estáticos y desactualizados.

La empresa cuenta con una base de datos transaccional que permite almacenar los datos del negocio y la consulta de estos mediante el gestor de base de datos con la respectiva licencia, pero a pesar de aquello, se evidencia la falta de políticas de respaldo de la información o la creación de soluciones idóneas para la extracción de datos, tales como vistas materializadas o cubos de información para el uso oportuno de todos los departamentos estratégicos de la organización.

Por este motivo, la empresa se ha comprometido en crear un departamento de BI con la incorporación de personal altamente especializado en el tratamiento, análisis, modelado y visualización de datos, que permita fortalecer las relaciones de los diferentes departamentos estratégicos con la generación y presentación de información útil para la toma de decisiones de la compañía.

2.2. Fase de análisis: identificación de *gaps*

Como resultado de las conversaciones, discusiones y entrevistas junto con los requisitos de los Sistemas de Información y los materiales reunidos en la fase de descubrimiento, se ha examinado la arquitectura, la organización y los requisitos para determinar en donde se encuentra ahora la organización y a dónde le gustaría llegar con el proyecto planteado.

Para avanzar esta fase, se deben analizar los siguientes aspectos:

- Capacidad de los informes actuales.
- Diferencia entre los informes actuales y la información deseada Tecnología necesaria.
- Proveedores de tecnologías necesarias
- Costes y Recursos Humanos Implicados

2.2.1. Capacidad de los informes actuales

La capacidad de los informes actuales muestra un corte trimestral posterior a ello, el tiempo de demora en elaborar los informes es de un mes, con base a esto se estaría tomando una tardanza total de cuatro (4) meses desde la recolección de los datos hasta desarrollar los análisis de estos para su posterior socialización a la alta gerencia.

Así mismo, los informes actuales se muestra una deficiencia de calidad en la información, esto provocado por que existen campos vacíos, datos duplicados, entre otros aspectos de los datos,

así como una mayor precisión de los análisis por la falta de personal especializado y herramientas encaminadas al tratamiento, análisis y modelado de datos.

La compañía tampoco posee una estrategia correcta de Storytelling, cada departamento de la empresa maneja su información independientemente y generan informes según sus necesidades a través de la herramienta Excel. Una vez obtenidos los informes, indicadores y representaciones graficas necesarias para tomar decisiones, se crean reportes en PowerPoint donde se presentan estos resultados, lo cual es un proceso ineficiente, en razón que el resultado es un reporte estático que no permite la interacción con los datos.

2.2.2. Diferencia entre los informes actuales y la información deseada

La principal diferencia entre los informes actuales y la información deseada radica que el proyecto se está basando en cumplir un objetivo global y no un reporte individual de uno o varios departamentos.

Apalancado en un objetivo estratégico de la compañía para el año 2022 que define el incremento en ventas y fidelización de clientes en este proyecto de Inteligencia de Negocio se pretende conocer el comportamiento y segmentos de clientes por medio de un análisis RFM y modelo k-medias, para posteriormente establecer campañas y estrategias comerciales.

Por ello, los nuevos informes que se pretenden generar deben aportar la información necesaria para la realización del proyecto; en este caso es la referida a la recencia, frecuencia y valor monetario de las compras de los clientes. El poder realizar una segmentación de los clientes y agruparlos en función de distintas características, va a permitir ofrecer un servicio más personalizado y también el diseño de ofertas individualizadas.

2.2.3. Proveedores de tecnologías necesarias

La compañía cuenta con la tecnología necesaria, con sus licencias al día de sistemas operativas, motores de base de datos y licencias de uso de sus usuarios clientes de los servidores, tanto para el sistema empresarial como para su almacén de datos.

Donde se debe invertir en nuevas tecnologías es en la incorporación de dos herramientas básicas para el proyecto de Inteligencia de Negocio: 1) Herramienta de visualización, donde se recomienda reemplazar el actual Qlick View por Power BI, por su flexibilización, integración con los diferentes sistemas operativos (Windows, Linux, IOS, Ubuntu, entre otros), R y Python, su personalización de informes, innovación entre otras ventajas; y 2) Herramienta de analítica

de datos como R (lenguaje de programación gratuito) y su entorno de desarrollo integrado (IDE) como RStudio Desktop Pro que facilita el aprendizaje y uso de R para complementar y personalizar algunos de los análisis. También se puede incorporar provecho del software estadístico gratuito Gretl.

En principio, se trata de aplicaciones del estilo "open source", donde se puede partir de uso limitado sin costo o de muy bajo costo, pero que para poder aprovecharlas de manera completa o para uso colaborativo, corporativo y seguro deba accederse a versiones de pago.

Asimismo, desde el punto de vista de hardware se deberán incorporar algunas máquinas, preferentemente portátiles work station, para permitir el trabajo remoto y colaborativo; de entorno Microsoft, por compatibilidad; y con buenas capacidades de procesamiento, capacidad gráfica y almacenamiento; pues si bien se va a fomentar el uso "Cloud", la disciplina involucra el manejo masivo de archivos muy grandes y se recomienda contar con capacidad adecuada.

Análisis de datos en R:

R es un lenguaje de programación, con diversas librerías o paquetes de análisis estadístico bastante potentes que pueden suplir el campo de aplicación, R está pensado para explotar su potencial que es la "estadística". Este lenguaje nos permite una primera toma de contacto con los datos debido a su flexibilidad por la exploración, limpieza y análisis a diferentes fuentes de datos, así como aplicar modelos y algoritmos predictivos (Laude, 2017).

Visualización de datos en Power Bi:

Power BI es una herramienta de análisis de datos y visualización de datos, basada en un servicio en la nube y una aplicación de escritorio. Mediante el uso de Power BI podremos visualizar, analizar y compartir información de los datos analizados.

2.2.4. Cronología, costes y recursos humanos implicados

El presente proyecto de inteligencia de negocio tendrá una duración estimada de 12 semanas, iniciando desde el 24 de marzo de 2022 con el análisis y exploración de requerimientos del proyecto, hasta la creación de dashboard que contienen los resultados de la segmentación de clientes (Ver *Tabla 1*).

Tabla 1. *Cronología del proyecto*

Actividad	Sem. 1	Sem. 2	Sem. 3	Sem. 4	Sem. 5	Sem. 6	Sem. 7	Sem. 8	Sem. 9	Sem. 10	Sem. 11	Sem. 12
Análisis y Exploración de requerimientos	X	X										
Mapeo de fuentes e identificación de variables		X	X									
Extracción y preparado de datos				X	X	X						
Construcción y exploración de los modelos de segmentación						X	X					
Evaluación del modelo							X	X	X			
Análisis de los Resultados del Modelo								X	X	X		
Creación del Dashboard											X	X

Fuente: Elaboración propia

El proyecto no implica recursos económicos a menos que la empresa aplique las recomendaciones referentes a nuevas tecnologías y decida obtener las licencias de las herramientas premium y de Power BI y Rstudio, así como la creación del departamento de BI para que la compañía pueda adquirir sus propias competencias y un líder de BI con experiencia para acompañar a los demás directivos del negocio en la toma de decisiones basadas en información y que pueda suplir el gap faltante. Sin embargo, para alcanzar los objetivos planteados en el presente trabajo se requerirá el recurso humano de un analista de business intelligence[2].

2.3. Fase de recomendaciones: alcance, prioridades y presupuesto

2.3.1. Alcance del proyecto

La empresa dispone de la información de facturación desde inicios del año 2019, de la cual, para elaborar el proyecto, la misma que facilita la entrega de información hasta principios del año 2022 para ser procesada y analizada con las respectivas medidas de confidencialidad de sus clientes y en archivos planos.

La segmentación de clientes RFM se aplicará exclusivamente a los clientes que consumieron productos de la empresa de logística y distribución en el año 2021, mientras que el modelo de cluster K-Medias se aplicará a un grupo de clientes de importancia e identificados mediante el análisis de Recencia, Frecuencia y Monetario.

[2] El analista de BI con el apoyo de personal de los departamentos de TI y comerciales de la empresa ecuatoriana abordaron las fases de Introducción, Alcance y planificación en sesiones ejecutivas de trabajo.

2.3.2. Promoción del proyecto en la organización

El proyecto está promocionado y coordinado por la gerencia general de la empresa, apalancado con la gerencia comercial y TI, apalancado en dar cumplimiento al objetivo estratégico de la organización.

Por ello se ha dispuesto en conocer cuál es el comportamiento de sus clientes por medio de una segmentación y clasificación, mediante un análisis RFM y de Clúster, con el propósito de efectuar estrategias comerciales que permita incrementar las ventas y la fidelización de clientes para el año 2022.

3. Análisis y definición

En este punto del proyecto se efectuará el análisis exploratorio de los datos disponibles, partiendo inicialmente con la descripción de la fuente de información a utilizar, su conformación y estructura, el tratamiento inicial de los datos, así como la contextualización del modelo estadístico empleado para la segmentación de clientes aplicando previamente el análisis RFM.

3.1. Análisis de los datos a utilizar

Para el presente estudio, la compañía facilitó la réplica de las dos principales tablas de datos que conforman la base de datos transaccional de la corporación con la respectiva confidencialidad, de los requerimientos de sus consumidores, para el período enero de 2019, fecha desde la cual se encuentra vigente la solución informática para la generación de pedidos y facturación, hasta el mes de enero de 2022.

Para contar con las principales características en ser consideradas en el análisis de RFM y la segmentación, se procedió a la obtención del conjunto de datos en formato texto, el mismo que cuenta con las variables relacionadas con el detalle del pedido que efectuaron los clientes a la compañía con un total de 2.395.897 registros.

Los ficheros de datos constan de las siguientes características:

Detalle de los datos de pedido

- Código único del pedido - idpedido
- Valor total del pedido - total

- Subtotal del pedido (Sin IVA) – subtotal.
- Descuento del pedido correspondiente al plan comercial. (El plan comercial contiene descuentos dirigidos a distintos segmentos de productos y clientes) – plancomercial.
- Descuento del pedido correspondiente a las promociones aplicadas. (Estas bonificaciones son financiadas por los proveedores para promocionar sus productos) – promociones.
- Este valor corresponde al descuento que pueden aplicar los vendedores en la negociación con los clientes – descuentomanual.
- Código único del cliente – codigocliente.
- Sucursal desde la cual se va a despachar el pedido – sucursal.
- Fecha en la cual se tomó el pedido (DD/MM/YYYY) – date.
- Estado del pedido (0-Cancelado, 1-Activo, 9-Cotización) – estado.
- Provincia de ubicación del cliente – provincia.
- Ciudad de ubicación del cliente – ciudad.
- Latitud de ubicación del cliente – latitud.
- Longitud de ubicación del cliente – longitud.

3.2. Preparación de los datos

En esta etapa, se van a realizar diversas tareas para adaptar los datos en bruto y aproximarlos a un conjunto de datos final, lo que implica llevar a cabo una limpieza de los datos para mejorar su calidad, añadir nuevas variables a partir de las ya existentes y asignar un formato adecuado a los datos para que puedan ser procesados correctamente por la herramienta de modelado.

3.2.1. Información general de los datos

Previo a iniciar con la limpieza de los datos, resulta conveniente contar con una descripción general de las variables o características que componen el set de datos que servirán de insumo en el presente estudio. En la *Figura 2*, se representa la información general de cada una de las variables, donde se observa el número de registros de cada una de ellas, así como su tipología.

Figura 2. *Información general de las variables del dataset*

```
  idpedido             total              subtotal         plancomercial        promociones         descuentomanual
Length:2395879     Min.   :     0.0   Min.   :0.000e+00   Min.   :-59889.98   Min.   :-1.396e+09   Min.   :   0.00
Class :character   1st Qu.:    10.9   1st Qu.:1.100e+01   1st Qu.:    -0.38   1st Qu.:-4.000e+00   1st Qu.:   0.00
Mode  :character   Median :    19.1   Median :1.900e+01   Median :    -0.07   Median :-1.000e+00   Median :   0.00
                   Mean   :   168.5   Mean   :7.690e+02   Mean   :    -7.69   Mean   :-6.040e+02   Mean   :   1.08
                   3rd Qu.:    43.5   3rd Qu.:4.500e+01   3rd Qu.:     0.00   3rd Qu.: 0.000e+00   3rd Qu.:   0.00
                   Max.   :391000.0   Max.   :1.396e+09   Max.   :   473.46   Max.   : 3.924e+04   Max.   :3463.56
codigocliente        sucursal             fecha                estado   provincia            ciudad             latitud
Length:2395879     Length:2395879     Min.   :2015-12-31   Min.   :1   Length:2395879     Length:2395879     Length:2395879
Class :character   Class :character   1st Qu.:2019-09-12   1st Qu.:1   Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Median :2020-07-20   Median :1   Mode  :character   Mode  :character   Mode  :character
                                      Mean   :2020-07-10   Mean   :1
                                      3rd Qu.:2021-04-30   3rd Qu.:1
                                      Max.   :2022-01-30   Max.   :1
  longitud          aniopedido         mespedido
Length:2395879     Length:2395879     Length:2395879
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character
```

Fuente: Datos de la empresa de estudio, extraído al Lenguaje de programación de R

Es importante destacar que se efectuaron modificaciones en la tipología de variables de acuerdo con su propia naturaleza, así como la creación de los campos de año y mes del pedido y la modificación de nombre al lenguaje español para el caso de la variable fecha del pedido, para facilitar el tratamiento y descripción de los datos.

3.2.2. Limpieza de los datos

La limpieza de datos es un proceso que tiene por objetivo detectar, corregir o eliminar los registros inexactos presentes en la base de datos a analizar. Para ello, se van a utilizar herramientas que permitan detectar y limpiar:

- Variables no relevantes
- Valores faltantes, duplicados o incorrectos
- Datos atípicos.

3.2.2.1. Eliminación de variables no relevantes

La información facilitada por la compañía incorpora una serie de campos o características sobre las promociones o descuentos aplicados a los productos que distribuye a sus consumidores, las mismas que son de gran utilidad para un análisis de impacto de las estrategias comerciales que impulsa la empresa, pero que sobresale del alcance del presente trabajo.

Adicionalmente, se ha constado que las variables de ubicación geográfica de latitud y longitud presentan códigos que no corresponden al territorio ecuatoriano, así como, el ingreso de valores nulos o de ceros, por lo cual se decide considerar como variable para georreferenciar a los clientes de la compañía la provincia de localización.

Por lo anterior, se procede a la supresión de las siguientes variables del conjunto de datos: subtotal, plancomercial, promociones, descuentomanual, estado, longitud y latitud.

3.2.2.2. Valores faltantes, duplicados o incorrectos

En el conjunto de datos se observan 30 transacciones realizadas previo al año 2019, las mismas que son consideradas como datos no válidos y por tanto se proceden a su eliminación, así mismo, se cuenta con registros del mes de enero 2022, que no serán empleados en el presente trabajo en razón que la segmentación de clientes propuesta se aplicará a todos los pedidos ejecutados y facturados durante el año 2021.

Tabla 2. *Número y monto de los Pedidos 2019 - 2021*

Año Pedido	Monto Pedidos	Total de Pedidos
2019	$ 125,699,216	826,253
2020	$ 137,607,336	724,496
2021	$ 129,668,099	782,947
TOTAL	$ 392,974,651	2,333,696

Lo referente a la presencia de datos duplicados, se observa que cada pedido está identificado de manera única mediante su ID de control en la herramienta informática para la generación de los requerimientos de productos, por tanto, no existen registros duplicados en el conjunto de datos, tal como se contrasta con los resultados de la *Figura 3*.

Figura 3. *Verificación de datos duplicados del dataset*

```
count(distinct(pedidos, idpedido, .keep_all = TRUE))
      n
2333696
```

En lo concerniente con la falta de datos en el dataset, se comprueba que no existen datos omitidos en la generación y despacho de los pedidos por parte de los promotores de la empresa a los consumidores (Ver *Figura 4*), es decir, que la herramienta tecnológica obliga al personal al llenado de todos los campos referente a los distintos requerimientos de productos en territorio.

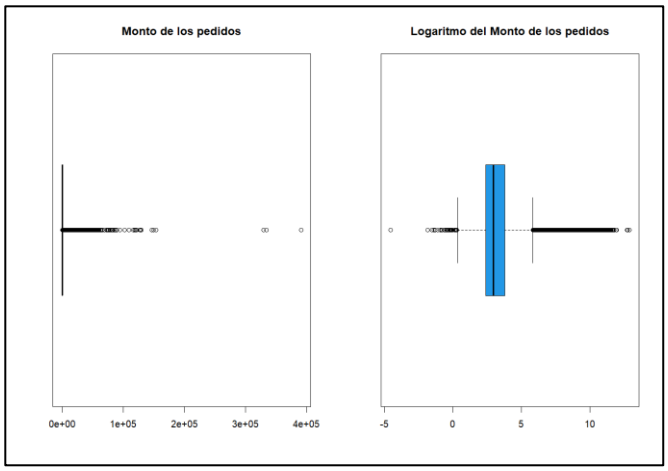
Figura 4. *Verificación de datos faltantes del dataset*

```
pedidos %>% summarise_all(funs(sum(is.na(.))))
idpedido total codigocliente sucursal fecha provincia ciudad aniopedido mespedido
       0     0             0        0     0         0      0          0         0
```

3.2.2.3. Datos atípicos

Del conjunto de datos establecido para el presente estudio, se analiza la distribución de la variable del monto de cada pedido, de la cual se identifica valores extremos en su comportamiento (Ver *Figura 5*), es decir, que existen clientes que efectúan requerimientos masivos de productos con un alto valor económico que pueden darse por la creación de una nueva tienda o local por una única vez, por efecto que no es recurrente un monto elevado para los siguientes requerimientos. Asimismo, existen consumidores que efectuaron compras de artículos con un monto bajo del pedido realizado, por tal razón, en el análisis RFM se deberá considerar este comportamiento en la variable monetaria para precisar los segmentos de consumidores y la clusterización mediante el modelo de K-medias.

Figura 5. *Verificación de datos atípicos en el Monto del Pedido*



3.2.3. Construcción de los datos

El objetivo del presente estudio es la segmentación de los clientes de la empresa en base al comportamiento de los pedidos realizados en el año 2021 y empleando el análisis RFM, para lo cual es imprescindible generar una serie de variables que aporten información sobre ese comportamiento, como verificar el número de días desde la última compra realizada (Recencia), el número de veces que se compró (Frecuencia) y el gasto realizado en las compras (Monetización o Valor Monetario).

Para conocer los valores de Recencia y Frecuencia se ha considerado la variable Fecha en la cual se tomó el pedido (fecha) y el número de pedidos que efectúa cada cliente en el año 2021. En relación con el Valor Monetario, se ha tomado la variable valor total del pedido (total).

Utilizando las tres variables y realizando una agrupación por cliente o código único de cliente se va a generar el set de datos correspondiente a la Recencia, Frecuencia y Valor Monetario.

Cabe recordar que las variables hacen referencia a Recencia que es el periodo transcurrido desde última compra realizada por un cliente. Por ello, se establece dicho espacio de tiempo, y es necesario realizar una transformación de la variable “Fecha en la cual se tomó el pedido” en la que consiste establecer una fecha de referencia para establecer la cantidad de días por cliente en el periodo comprendido entre la fecha final de referencia (31 de diciembre de 2021) y la fecha de del último pedido, siendo este valor los días transcurridos desde el último pedido o compra realizada.

Por otro lado, la variable Frecuencia indica cuantas veces ha comprado el cliente o ha realizado el pedido por lo que para generar esta variable se debe realizar un conteo del número de compras efectuadas por cada cliente.

La última variable del Valor Monetario hace referencia al ¨Valor del pedido” por parte del cliente, la misma que se obtendrá con el promedio del monto de cada pedido a nivel de cada consumidor. No se considera la sumatoria del monto de cada pedido, en razón que existen registros de requerimientos con un alto valor económico por la creación o instauración de un nuevo local o tienda, sin embargo, en los restantes pedidos del año su valor es muy inferior.

En la *Figura 6*, se detallan las diferentes variables que se han consolidado a nivel de cada cliente de la empresa, que incluye la creación de las métricas de recency, frecuency y monetary para el respectivo análisis RFM y de clúster.

Figura 6. *Detalle del dataset de clientes con las variables RFM*

RStudio Source Editor

clientes

Filter

	codigocliente	sucursal	provincia	ciudad	monto_pedido	numero_pedidos	primer_pedido	ultimo_pedido	dias_compra	recency	frecuency	monetary
1	10012	IBARRA	IMBABURA	PIMAMPIRO	2139.7084	18	2021-02-08	2021-12-15	310	17	18	118.872689
2	10013	QUITO	AZUAY	AMALUZA	366.6501	26	2021-02-10	2021-12-16	309	16	26	14.101927
3	10021	QUITO	PICHINCHA	QUITO	351.8641	25	2021-02-02	2021-12-21	322	11	25	14.074564
4	10025	GUAYAQUIL	GUAYAS	GUAYAQUIL	484.0242	35	2021-01-07	2021-12-30	357	2	35	13.829263
5	10044	LAGO AGRIO	ORELLANA	LA JOYA DE LOS SACHAS	141.7600	18	2021-01-05	2021-09-07	245	116	18	7.875556
6	10046	QUITO	PICHINCHA	QUITO	827.6865	26	2021-02-16	2021-12-29	316	3	26	31.834096
7	10055	QUITO	PICHINCHA	ALANGASI	513.9682	48	2021-01-06	2021-12-22	350	10	48	10.707671
8	10058	QUITO	PICHINCHA	QUITO	840.8416	37	2021-01-06	2021-12-28	356	4	37	22.725449
9	10060	QUITO	PICHINCHA	ALANGASI	1120.3546	27	2021-01-07	2021-11-19	316	43	27	41.494615
10	10062	LAGO AGRIO	ORELLANA	ORELLANA	1047.6044	43	2021-01-05	2021-12-28	357	4	43	24.362893
11	10077	GUAYAQUIL	GUAYAS	ALFREDO BAQUERIZO MORENO	514.1400	41	2021-01-06	2021-12-17	345	15	41	12.540000
12	10086	GUAYAQUIL	GUAYAS	ALFREDO BAQUERIZO MORENO	1128.7695	36	2021-01-11	2021-12-27	350	5	36	31.354708
13	10093	QUITO	PICHINCHA	ALANGASI	468.7343	28	2021-01-05	2021-11-08	307	54	28	16.740511
14	101	QUITO	PICHINCHA	QUITO	497.2563	36	2021-01-12	2021-12-01	323	31	36	13.812675
15	10108	SANTO DOMINGO	SANTO DOMINGO DE LOS TSACHILAS	LAS VILLEGAS	2208.4072	20	2021-01-05	2021-12-21	350	11	20	110.420360
16	10123	GUAYAQUIL	GUAYAS	ALFREDO BAQUERIZO MORENO	381.3600	14	2021-02-12	2021-08-20	189	134	14	27.240000
17	10124	MANTA	MANABÍ	PORTOVIEJO	179055.3600	75	2021-01-06	2021-04-29	113	247	75	2387.404800
18	1013	GUAYAQUIL	GUAYAS	ALFREDO BAQUERIZO MORENO	289.1243	24	2021-01-18	2021-12-20	336	12	24	12.046846
19	10134	IBARRA	IMBABURA	IBARRA	1969.8925	24	2021-01-12	2021-11-30	322	32	24	82.078854
20	10136	IBARRA	IMBABURA	IBARRA	2241.5797	22	2021-01-20	2021-11-30	314	32	22	101.889986
21	10141	QUITO	AZUAY	AMALUZA	5735.7886	48	2021-01-12	2021-12-28	350	4	48	119.495596
22	10149	IBARRA	ESMERALDAS	SAN LORENZO	24993.9464	39	2021-01-05	2021-12-29	358	3	39	640.870421
23	10170	QUITO	PICHINCHA	ALANGASI	254.4405	25	2021-01-12	2021-11-29	321	33	25	10.177620
24	10172	QUITO	AZUAY	AMALUZA	262.9013	15	2021-01-08	2021-12-23	349	9	15	17.526753

Showing 1 to 24 of 24.885 entries. 12 total columns

3.3. Modelado propuesto

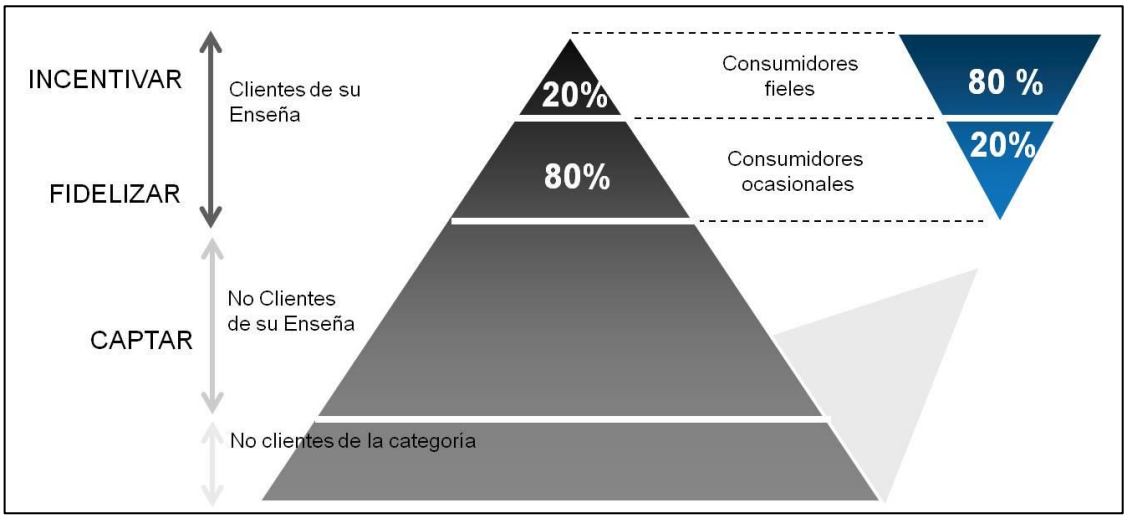
Posterior a la creación de las métricas de recency, frecuency y monetary a nivel de cada cliente y en función al objetivo del proyecto, es necesario emplear el análisis RFM al conjunto de datos definidos que permita identificar grupos o segmentos de clientes para obtener información importante para la toma de decisiones de la compañía, acompañada con una técnica de segmentación o clusterización que permita ampliar la estructura de uno o varios grupos de interés que se obtuvieron a través del modelo RFM, consiguiendo que los subgrupos de individuos tengan características similares o proxis entre sí, y que entre los grupos existan comportamientos diferentes entre los ellos.

3.3.1. Análisis RFM

El método RFM ha sido la base de la mayoría de las segmentaciones de marketing directo durante décadas (Miglautsch, 2002). RFM tiene una trayectoria de décadas. No es una moda ni un truco del marketing. Es un proceso científicamente probado. Se basa en el principio de Pareto, comúnmente conocido como la regla 80-20.

La regla de Pareto indica que el 80% de los resultados provienen del 20% de las causas. Del mismo modo, el 20% de los clientes contribuyen al 80% de sus ingresos totales (Ver *Figura 7*). El principio de Pareto es el núcleo del modelo RFM. Es probable que centrar sus esfuerzos en segmentos críticos de clientes le proporcione un retorno de la inversión mucho mayor (Anish, 2022).

Figura 7. *Principio de Pareto: Regla del 80-20*



Fuente: https://www.unica360.com/analisis-rfm-en-retail-empezando-a-segmentar-clientes-i

Según Kumar & Reinartz (2018), la idea general del análisis RFM es clasificar o segmentar a los clientes según su medida de Recency, Frecuency y Monetary. Los grupos de clientes resultantes están asociados con el comportamiento de compra. El método RFM es similar al enfoque de matriz de transición en el sentido de que también rastrea el comportamiento del cliente a lo largo del tiempo en lo que se denomina un espacio de estado. Es decir, los clientes se mueven en el tiempo a través del espacio con ciertos estados de actividad definidos.

La agrupación en clústeres basada en atributos de RFM proporciona más conocimiento del comportamiento de los niveles reales de marketing de los clientes que otros análisis de clústeres. Las reglas de clasificación descubiertas a partir de las variables demográficas de los clientes y las variables RFM brindan conocimientos útiles para que los gerentes predigan el comportamiento futuro de los clientes, por ejemplo, qué tan recientemente comprará el cliente, con qué frecuencia comprará y cuál será el valor de sus compras. Las reglas de asociación basada en medidas RFM analiza las relaciones de las propiedades del producto y las contribuciones/lealtades de los clientes para proporcionar una mejor recomendación para satisfacer las necesidades de los clientes (Birant, 2011).

El proceso para cuantificar el comportamiento del cliente a través del modelo RFM es el siguiente. Primero, ordene la base de datos por cada dimensión de RFM y luego divida la lista de clientes en cinco segmentos iguales. Se sabe que el método tiene un tamaño exactamente igual. Los diferentes quintiles de RFM tienen diferentes tasas de respuesta (Wei, Lin, & Wu, 2010).

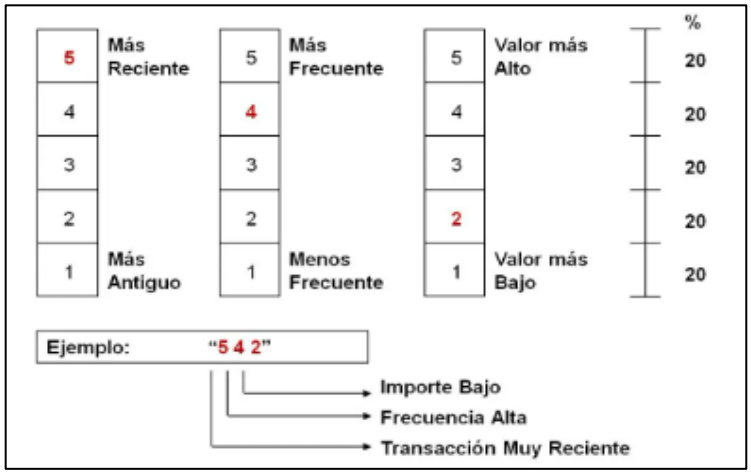
Para la medida de recency, los clientes se ordenan considerando la fecha de compra en orden descendente y luego asignar valores numéricos del 1 al 5 al conjunto de datos. El conjunto de datos original con la fecha de compra se transforma en un valor de 1 a 5 según la fecha de compra más reciente. Por lo tanto, el valor de 5 se asigna al 20 % superior del conjunto de datos en términos de la fecha de compra más reciente. El valor de 4 se le da al próximo 20% del conjunto de datos y así sucesivamente.

Para la frecuencia, la técnica es ordenar el número de transacciones en un período de tiempo determinado, como el número de transacciones por mes, en orden descendente. El 20% superior del conjunto de datos viene dado por el valor de 5. El valor de 4 se asigna al siguiente 20% del conjunto de datos y así sucesivamente. El conjunto de datos original con el número de transacciones se transforma en un valor de 1 a 5.

Finalmente, para el valor monetario o monetary, la técnica es ordenar la cantidad de dinero gastado como promedio por mes, año o transacción en orden descendente. Asigne el valor de 5 al 20% superior del conjunto de datos. Asigne el valor de 4 al próximo 20% de los datos y así sucesivamente. El conjunto de datos original con la cantidad de dinero gastado se transforma en un valor de 1 a 5 (Jun, et al., 2020) (Wei, Lin, & Wu, 2010).

El análisis RFM asigna puntajes de valor a cada cliente sobre la base de su comportamiento anterior. Utilizando el sistema de quintiles explicado anteriormente, se pueden asignar como máximo 125 puntuaciones diferentes (5x5x5). Estas agrupaciones difieren en tamaño entre sí (Ver *Figura 8*).

Figura 8. *Segmentos de Clientes con el modelo RFM*



Fuente: https://www.unica360.com/analisis-rfm-en-retail-empezando-a-segmentar-clientes-i

La puntuación de un cliente puede oscilar entre 555, que es la más alta calificación a un cliente, y 111, que es la más baja. Los mejores clientes están en el quintil 5 de cada factor (555) que han comprado más recientemente, con mayor frecuencia y que han gastado más dinero.

Varios estudios de investigación muestran que cuanto mayor sea el valor de Recency o Frecuency, mayor será la probabilidad de que el cliente correspondiente realice una nueva transacción con el vendedor. Además, cuanto mayor sea Monetary, más probable es que el cliente correspondiente vuelva a comprar productos o servicios del vendedor (Jun, et al., 2020).

Si bien existen innumerables formas de realizar la segmentación descriptiva de clientes, el análisis RFM es popular por tres razones:

- Utiliza escalas numéricas objetivas que producen una descripción concisa e informativa de alto nivel de los clientes.
- Es simple: los especialistas en marketing pueden usarlo de manera efectiva sin la necesidad de científicos de datos o software sofisticado.
- Es intuitivo: el resultado de este método de segmentación es fácil de entender e interpretar.

3.3.2. Análisis de Conglomerados

Los métodos de análisis y clasificación de grupos agrupan de casos o elementos, en base a criterios cualitativos o cuantitativos (distancias o similaridades). A veces, en lugar de los casos, se forman grupos con las variables. Entre los métodos estadísticos de clasificación, que tratan de analizar la pertenencia de casos a diversos grupos, se puede distinguir fundamentalmente tres: Análisis discriminante, análisis de segmentaciones y análisis de conglomerados (Álvarez, 1995).

En el análisis discriminante, las agrupaciones o grupos son conocidos y establecidos a priori, y esta técnica explica la pertenencia de un elemento a uno u otro grupo, en base a los valores de un grupo de casos.

El análisis de segmentaciones también pretende definir grupos a partir de variables consideradas como relevantes. La diferencia con respecto al análisis de conglomerados es que, una de las variables indica el criterio, y el resto definen los grupos. Este tipo de análisis es utilizado con mayor frecuencia en el campo de la salud.

En el análisis de conglomerados, el investigador no tiene conocimiento de la existencia de los subgrupos o conglomerados, del número de grupos a formar, ni mucho menos de las características que los definen. Es una técnica, por lo tanto, eminentemente exploratoria y descriptiva sin variables dependientes. Los grupos predefinidos, se definen mediante el cálculo de distancias o similaridades, a partir de los valores de las variables consideradas como adecuadas para el agrupamiento en función de los aspectos teóricos y prácticos del fenómeno en estudio (Álvarez, 1995).

El análisis de clúster o de conglomerados es una técnica de clasificación que sirve para poder detectar, describir y analizar grupos de sujetos o variables homogéneas en función de los valores observados dentro de un conjunto aparentemente heterogéneo. Se fundamenta en el

estudio de las distancias entre ellos, permitiendo en el análisis, cuantificar el grado de similitud, en el caso de las proximidades, y el grado de diferencia, en el caso de las distancias. Como resultado aparecen agrupaciones homogéneas.

El objetivo principal de esta técnica multivariante y de interdependencia es la configuración de grupos homogéneos y heterogéneos, es decir, que los individuos u objetos de cada una de las agrupaciones posean similares características, y que los grupos sean disímiles entre ellos.

El análisis de conglomerados es conocido también como análisis **Q**, construcción de tipología, análisis de clasificación y taxonomía numérica, debido a su amplio uso o aplicación en diversas disciplinas tales como psicología, sociología, economía, ingeniería y negocios. El análisis de conglomerados es comparable al análisis de componentes principales en su objetivo de evaluar la estructura del conjunto de datos, pero el análisis de conglomerados difiere del Análisis de Componentes Principales, ACP, en que este primero agrupa objetos o individuos mientras que el de componentes principales se enfoca en la creación de combinaciones lineales con la agrupación de variables (Hair, Anderson, Tatham, & Black, 1999).

3.3.2.1. Procedimientos del análisis de conglomerados

Dentro del análisis de conglomerados existen dos procedimientos que permiten clasificar individuos, estos son los procedimientos jerárquicos y no jerárquicos.

La diferencia entre estos grupos radica en que el primer procedimiento forma los grupos en pasos sucesivos y pueden analizar en cada paso las distancias entre los grupos formados, mientras que el segundo se basa en la asignación de los individuos en los conglomerados, que estos a su vez, ya se encuentran definidos previamente (Hair, Anderson, Tatham, & Black, 1999).

3.3.2.2. Procedimientos jerárquicos

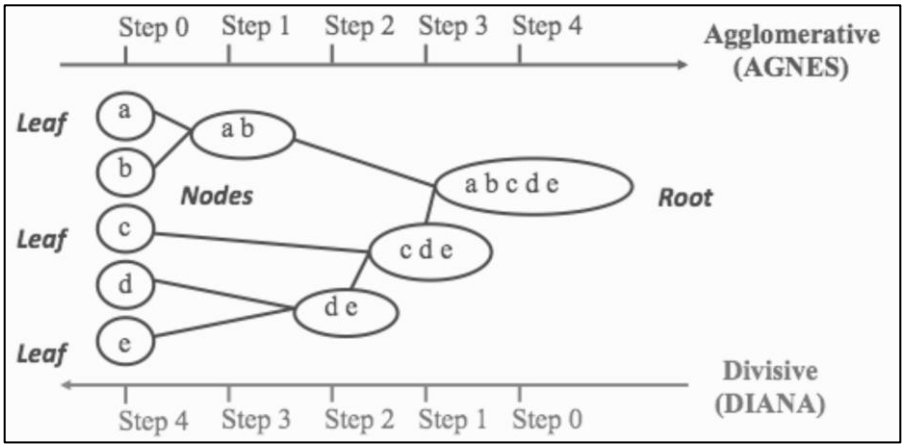
Los procedimientos jerárquicos consisten en la generación de una estructura en forma de árbol (Hair, Anderson, Tatham, & Black, 1999). El agrupamiento se lo realiza mediante un proceso que conlleva una seria de fases de agrupación o des agrupación sucesiva.

Los métodos de agrupación jerárquica son la forma más intuitiva de agrupar datos porque imitan cómo un humano abordaría la tarea de dividir un conjunto de n observaciones (consumidores) en k grupos (segmentos). Si el objetivo es tener un gran segmento de mercado (k = 1), la única solución posible es un gran segmento de mercado que contenga a todos los

consumidores en los datos X. En el otro extremo, si el objetivo es tener tantos segmentos de mercado como haya consumidores en el conjunto de datos (k = n), el número de segmentos de mercado debe ser n, y cada segmento debe contener exactamente un consumidor. Cada consumidor representa su propio grupo. El análisis de segmentación del mercado se produce entre esos dos extremos (Dolnicar, Grun, & Leisch, 2018).

El resultado final de los procedimientos jerárquicos según se observa en la *Figura 9*, es la obtención de una jerarquía de unión completa en la que cada grupo se une o separa en una determinada fase. Existen dos tipos de procedimientos para la obtención de conglomerados jerárquico: Aglomerativos y divisivos.

Figura 9. *Procedimientos de Conglomerados Jerarquicos: Aglomerativos y divisivos*



Fuente: (Practical Guide To Cluster Analysis in R: Unsupervised Machine Learning, 2017)

Los procedimientos de agrupamiento jerárquico divisivo comienzan con el conjunto completo de datos (k = 1) y lo dividen en dos segmentos de mercado en un primer paso. Luego, cada uno de los segmentos se divide nuevamente en dos segmentos. Este proceso continúa hasta que cada consumidor tiene su propio segmento de mercado (k = n).

El procedimiento jerárquico aglomerativo aborda la tarea desde el otro extremo del método de divisivos. El punto de partida es que cada consumidor represente su propio segmento de mercado (k = n). Paso a paso, los dos segmentos de mercado más cercanos entre sí se fusionan hasta que el conjunto completo de datos forma un gran segmento de individuos (k = 1).

Ambos enfoques dan como resultado una secuencia de particiones anidadas. Una partición es una agrupación de observaciones tal que cada observación está contenida exactamente en un grupo. La secuencia de particiones varía desde particiones que contienen solo un grupo

(segmento) hasta n grupos (segmentos). Están anidados porque la partición con k + 1 grupos (segmentos) se obtiene de la partición con k grupos al dividir uno de los grupos (Dolnicar, Grun, & Leisch, 2018).

Según lo señala Shmueli, Bruce, Yahav, Patel, & Kenneth C. Lichtendahl (2018), el agrupamiento jerárquico es muy atractivo en su aplicación porque no requiere la especificación de la cantidad de clústeres y, en ese sentido, está puramente basado en datos. La capacidad de representar el proceso de agrupación y los resultados a través de dendogramas también es una ventaja de este método, en razón que es más fácil de entender e interpretar. Sin embargo, existen algunas limitaciones que se deben tener en cuenta:

i. El agrupamiento jerárquico requiere el cálculo y almacenamiento de una matriz de distancia n x n. Para conjuntos de datos muy grandes, esto puede ser costoso y lento a nivel de recursos de memoria del computador.
ii. El algoritmo jerárquico hace solo un paso a través de los datos. Esto significa que los registros que se asignan incorrectamente al principio del proceso no se pueden reasignar posteriormente.
iii. La agrupación jerárquica también tiende a tener poca estabilidad. Las acciones de reordenar datos o eliminar algunos registros puede conducir a una solución diferente.
iv. Con respecto a la elección de la distancia entre los conglomerados, el enlace único y completo es resistente a los cambios en la métrica de distancia, siempre que se mantenga el orden relativo. Por el contrario, la vinculación promedio está más influenciada por la elección de la métrica de distancia y puede conducir a grupos completamente diferentes cuando se cambia la métrica.
v. La agrupación jerárquica es sensible a los valores atípicos.

3.3.2.3. Procedimientos no jerárquicos

Los procedimientos no jerárquicos son convenientes utilizarlos cuando los individuos a clasificar son muchos y/o para refinar la agrupación realizada a través de un método jerárquico. En este método se supone que el número de agrupaciones son conocidas a priori (Anderberg, 1973).

En los procedimientos no jerárquicos no se realizan gráficos de árbol o dendograma, en su lugar, se asignan los individuos a grupos una vez que el número de grupos a formar son

determinados previamente, de tal manera que la varianza dentro de cada grupo sea mínima y entre estos sea la máxima. Para conjuntos de datos más grandes, los dendrogramas son difíciles de leer y la matriz de distancias por pares generalmente no cabe en la memoria de la computadora.

Para conjuntos de datos que contienen más de 1000 observaciones (clientes), los métodos de agrupamiento que crean una sola partición son más adecuados que una secuencia anidada de particiones. Esto significa que, en lugar de calcular todas las distancias entre todos los pares de observaciones en el conjunto de datos al comienzo de un análisis de conglomerados de partición jerárquica utilizando una implementación estándar, solo se calculan las distancias entre cada consumidor en el conjunto de datos y el centro de los segmentos (Dolnicar, Grun, & Leisch, 2018).

Un enfoque no jerárquico para formar buenos conglomerados es especificar previamente un número deseado de conglomerados, k, y asignar cada caso a uno de los k grupos para minimizar una medida de dispersión dentro de los conglomerados. En otras palabras, el objetivo es dividir la muestra en un número predeterminado k de conglomerados que no se superponen para que los conglomerados sean lo más homogéneos posible con respecto a las medidas utilizadas (Shmueli, Bruce, Yahav, Patel, & Kenneth C. Lichtendahl, 2018).

Una medida común de la dispersión dentro de un grupo es la suma de las distancias (o la suma de las distancias euclidianas al cuadrado) de los registros desde el centroide de su grupo. El problema se puede configurar como un problema de optimización que involucra programación entera, pero debido a que resolver programas enteros con una gran cantidad de variables lleva mucho tiempo, los clústeres a menudo se calculan utilizando un método heurístico rápido que produce soluciones buenas (aunque no necesariamente óptimas). El algoritmo k-medias es uno de esos métodos.

El método de k-medias comienza con una partición inicial de los registros en k grupos. Los pasos subsiguientes modifican la partición para reducir la suma de las distancias de cada registro desde su centroide de clúster. La modificación consiste en asignar cada registro al más cercano de los k centroides de la partición anterior. Esto conduce a una nueva partición en la que la suma de las distancias es menor que antes. Se calculan las medias de los nuevos conglomerados y se repite el paso de mejora hasta que la mejora sea muy pequeña (Anderberg, 1973) (Shmueli, Bruce, Yahav, Patel, & Kenneth C. Lichtendahl, 2018).

El procedimiento K-medias, normalmente utiliza una de las siguientes tres aproximaciones para asignar las observaciones individuales de uno de los grupos (Hair, Anderson, Tatham, & Black, 1999):

- Umbral Secuencial
- Umbral Paralelo
- Optimización

Con el transcurso del tiempo, los procedimientos no jerárquicos han ganado una creciente aceptación y se aplican cada vez más en la creación de grupos (Hair, Anderson, Tatham, & Black, 1999). Su uso, sin embargo, depende de la capacidad del investigador para seleccionar los puntos de semilla de acuerdo con bases teóricas y prácticas, por lo que es el investigador quien define el número de aglomeraciones a realizar y no el método como sucede en los jerárquicos, por lo que podría ser una desventaja de este procedimiento.

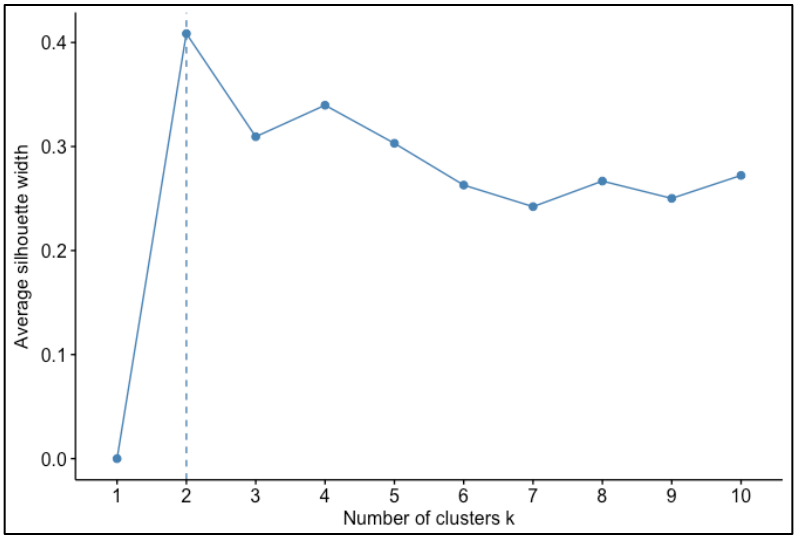
Una solución para evitar la limitación que presentan los procedimientos no jerárquicos, es la combinación de estos con los procedimientos jerárquicos, es decir, establecer el número de conglomerados a formar, los perfiles de los centros y la identificación de atípicos mediante los procedimientos jerárquicos, mientras que la agrupación de los individuos con los no jerárquicos (Hair, Anderson, Tatham, & Black, 1999). Sin embargo, el alto costo computacional para una muestra de más de 1000 registros o consumidores condiciona este tipo de procedimiento.

Otra solución para la determinación óptima del número de conglomerados es la aplicación de métodos que permiten observar la consistencia de los clústeres que se desean formar a través del procedimiento k-medias (Charrad, Ghazzali, Boiteau, & Niknafs, 2014) (Everitt, Landau, Leese, & Stahl, 2011), entre los cuales, los más usados son el método de la silueta “Average Silhouette Method” (Laude, 2017) y método del Codo “elbow method”, (Dolnicar, Grun, & Leisch, 2018) (Kassambara, 2017).

El método de la silueta mide la calidad del agrupamiento o clustering. Mide la distancia de separación entre los clústers. Es decir, determina qué tan bien se encuentra cada objeto dentro de su grupo. Un ancho de silueta promedio alto indica una buena agrupación. El método de la silueta promedio calcula la silueta promedio de las observaciones para

diferentes valores de k grupos. El número óptimo de conglomerados k es el que maximiza la silueta promedio sobre un rango de valores posibles para k (Ver *Figura 10*).

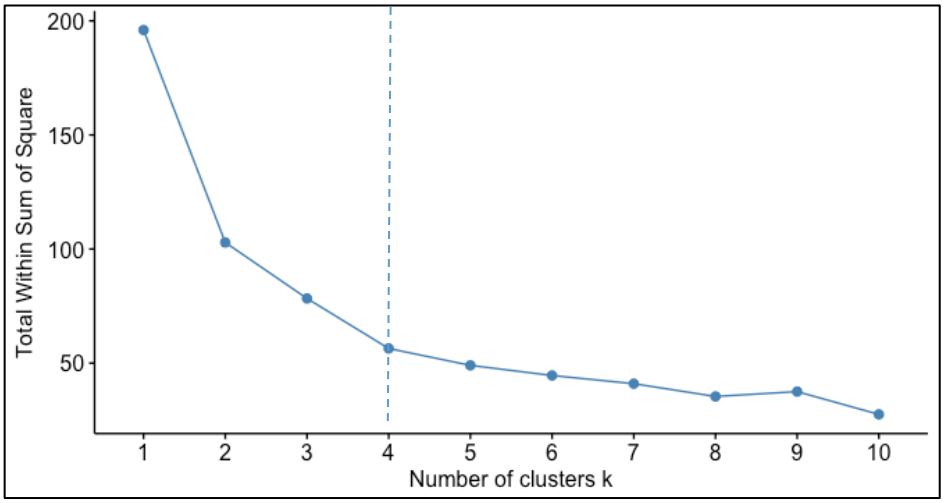
Figura 10. *Número óptimo de clusters - Método de la Silueta*



Fuente: https://uc-r.github.io/kmeans_clustering

Por otro lado, el método del codo emplea la distancia media de las observaciones a su centroide. Es decir, se fija en las distancias intra-cluster. Cuanto más grande es el número de clusters k, la varianza intra-cluster tiende a disminuir. Cuanto menor es la distancia intra-cluster, los clústers son más compactos y por ende se maximiza la varianza inter-cluster. El método del codo busca el valor k que satisfaga que un incremento de k, no mejore sustancialmente la distancia media intra-cluster (Ver *Figura 11*).

Figura 11. *Número óptimo de clusters - Método del Codo*



Fuente: (Practical Guide To Cluster Analysis in R: Unsupervised Machine Learning, 2017)

4. Construcción, prueba, implementación y despliegue

Una vez realizada la fase de preparación de los datos y la definición de los modelos a emplearse para cumplir con los objetivos propuestos en el presente estudio, se procede a la construcción e implementación de la segmentación de clientes mediante la aplicación del modelo RFM y el análisis de clúster mediante el modelo de K-medias a un grupo de clientes, para crear información útil para la compañía y su posterior despliegue de los resultados en una herramienta de BI.

4.1. Construcción e implementación de la segmentación de Clientes - Modelo RFM

Considerando el dataset de la *Figura 6*, que incorpora las variables de Recencia, Frecuencia y Valor Monetario, se dispone de la información insumo para la construcción e implementación del modelo RFM y la posterior aplicación del análisis de clúster.

Según el resumen de las variables de recencia, frecuencia y valor monetario detallados en la *Figura 12*, se observa que el tiempo promedio de la última compra hasta la fecha de referencia (31-diciembre-2021) es de 30.2 días, mientras que el 75% de los consumidores han realizado una compra en los últimos 29 días, y unos cuantos han decidido en no adquirir productos en los últimos 312 días, es decir que son clientes que pueden catalogarse como PERDIDOS para la empresa. En relación con la frecuencia, se identifica que durante el año 2021 los consumidores han realizado en promedio 26.3 compras, mientras que el 75% ha efectuado 32 requerimientos. Finalmente, el monto medio per cápita es de USD 125.68, sin embargo, se verifica que en el tercer cuartil de la distribución el gasto es de apenas USD 41.05.

Figura 12. *Resumen de las variables RFM*

```
> summary(clientes[,10:12])
    recency         frecuency         monetary
 Min.   :  1.00   Min.   : 12.00   Min.   :    5.509
 1st Qu.:  5.00   1st Qu.: 16.00   1st Qu.:   13.733
 Median : 12.00   Median : 22.00   Median :   20.168
 Mean   : 30.22   Mean   : 26.35   Mean   :  125.681
 3rd Qu.: 29.00   3rd Qu.: 32.00   3rd Qu.:   41.051
 Max.   :312.00   Max.   :681.00   Max.   :20871.023
```

4.1.1. Construcción de los segmentos RFM

Considerando la señalado por Miglautsch (2002) y Wei, Lin, & Wu (2010) para cuantificar y segmentar el comportamiento del cliente a través del modelo RFM, se debe ordene la base de datos por cada dimensión de RFM y luego dividirla la lista de clientes en cinco segmentos iguales. Es conocido que el método tiene un tamaño exactamente igual, sin embargo, los diferentes quintiles de RFM tienen diferentes tasas de respuesta.

En la *Figura 13* se detallan los 5 segmentos de clientes para cada una de las variables RFM, etiquetados como recency_puntaje, frecuency_puntaje y monetary_puntaje, y enumerados del 1 al 5.

Figura 13. *Dataset con las Variables RFM y quintiles RFM*

RStudio Source Editor

clientes_rfm

Filter

	codigocliente	recency	frecuency	monetary	recency_puntaje	frecuency_puntaje	monetary_puntaje
1	10012	17	18	118.872689	2	2	5
2	10013	16	26	14.101927	3	4	2
3	10021	11	25	14.074564	3	4	2
4	10025	2	35	13.829263	5	5	2
5	10044	116	18	7.875556	1	2	1
6	10046	3	26	31.834096	5	4	4
7	10055	10	48	10.707671	3	5	1
8	10058	4	37	22.725449	5	5	3
9	10060	43	27	41.494615	1	4	4
10	10062	4	43	24.362893	5	5	3
11	10077	15	41	12.540000	3	5	1
12	10086	5	36	31.354708	4	5	4
13	10093	54	28	16.740511	1	4	2
14	101	31	36	13.812675	2	5	2
15	10108	11	20	110.420360	3	3	5
16	10123	134	14	27.240000	1	1	4
17	10124	247	75	2387.404800	1	5	5
18	1013	12	24	12.046846	3	3	1
19	10134	32	24	82.078854	2	3	5

Showing 1 to 20 of 24,885 entries, 7 total columns

4.1.2. Composición y análisis de los segmentos RFM

Una vez construidos los segmentos de clientes para las variables de recencia, frecuencia y valor monetario, se observa que los diferentes grupos presentan similares números de casos y rangos diferenciados entre sí, como resultado a la segmentación por quintiles (Ver *Tabla 3*). Así mismo, es importante conocer cómo se han establecido los cortes de los quintiles en el

dataset para saber entre qué números se asigna cada quintil. Por ello, verificamos los valores mínimos y máximos para cada una de las variables RFM y así sabremos en torno a qué números significa una Recencia igual a 5 o a 1; al igual que para la Frecuencia y para el Valor Monetario.

Tabla 3. *Descripción de los clientes por quintiles de las variables RFM*

RFM	Segmentos	Promedio	Mínimo	Máximo	No. Clientes
recency	1	101.7	38	312	5,015
	2	24.1	17	37	5,170
	3	12.6	10	16	5,888
	4	7.0	5	9	4,031
	5	3.2	1	4	4,781
frecuency	1	13.0	12	14	4,269
	2	16.4	15	18	4,705
	3	21.3	19	24	5,414
	4	28.8	25	34	5,288
	5	49.0	35	681	5,209
monetary	1	10.6	5.5	12.8	4,977
	2	14.8	12.8	17.2	4,977
	3	20.4	17.2	24.7	4,977
	4	35.2	24.7	55.6	4,977
	5	547.3	55.6	20,871.0	4,977

De los segmentos construidos, se muestra que para la recencia el quinto grupo oscila entre 1 día y 4 días que compraron por última vez, es decir son compradores muy recientes de productos. En la frecuencia, el quintil 5 contempla el mayor número de compras para la compañía entre 35 y 681. En el Valor Monetario el quintil 5 oscila entre USD 55.6 y USD 20871.0, por el mismo motivo, muy pocos clientes se han gastado tanto dinero.

Por el lado contrario, se observa que en el primer quintil de la recencia los clientes han realizado su última compra hace 38 días o más, por tal efecto, se puede denominar a este grupo como compradores que olvidaron a la empresa. En cuanto a la frecuencia, el quintil 1 está conformado por consumidores que compraron como máximo 14 veces durante el año, lo cual se puede corroborar con la estructura del primer grupo de valor monetario, que han realizado compras de hasta USD 12.80 en el período de análisis.

Con relación a la distribución de los segmentos por cada una de las variables RFM detallada en el **¡Error! No se encuentra el origen de la referencia.**, se confirma la alta distribución que e xiste en el quinto quintil de las métricas de frecuencia y valor monetario, así como para el

primer grupo de la métrica de recencia, lo que conlleva a concluir que estos grupos de clientes deben ser analizados de manera multidimensional mediante la generación de las combinaciones entre éstos que permite obtener el modelo RFM.

En la *Tabla 4* se detalla el resultado de las distintas combinaciones de los grupos por las tres métricas, en esta tabla podemos evidenciar que la empresa tiene 402 clientes con (R=1) (F=1) (M=1), y 361 clientes con (R=5) (F=5) (M=5), por lo que se puede concluir que la compañía tiene más clientes "malos" que "buenos" desgraciadamente.

Tabla 4. *Combinación de segmentos de Clientes Modelo RFM*

Recency	Frecuency	Monetary				
		1	2	3	4	5
1	1	402	333	295	279	274
	2	321	275	281	268	200
	3	214	241	216	233	173
	4	141	154	178	167	111
	5	17	24	63	86	69
2	1	287	206	170	150	256
	2	310	253	209	175	287
	3	261	270	229	226	291
	4	184	195	198	196	239
	5	53	95	110	129	191
3	1	193	142	127	116	188
	2	236	218	175	166	226
	3	263	308	253	261	272
	4	233	256	329	317	304
	5	146	220	268	318	353
4	1	115	73	63	75	71
	2	112	132	119	103	98
	3	162	189	169	149	148
	4	204	209	227	208	162
	5	155	188	274	333	293
5	1	124	100	89	71	70
	2	158	130	82	90	81
	3	226	190	203	147	120
	4	230	252	231	224	139
	5	230	324	419	490	361

Además, de forma general se pueden concluir más situaciones como:

- No existen un importante número de clientes (259) que tengan una Recencia muy baja (R=1), es decir, que hayan comprado hace mucho tiempo, con una Frecuencia muy alta (F=5), es decir, que hayan comprado muchas veces.

- Tampoco existen muchos clientes (124) con valores altos de Recencia y Frecuencia (R=5) y (F=5) y que se gasten poco dinero (M=1), es decir que, aunque no se gasten mucho dinero, lo bueno es que repiten los requerimientos.
- Existen 1,007 clientes que la última compra la efectuaron durante los últimos 16 días (R=1,2,3) y sus montos de compras y frecuencia corresponden al rango de los valores superiores (F=5) (M=5).

4.1.3. Segmentación de Clientes con el Modelo RFM

Como vemos, en base al análisis RFM se puede obtener hasta 125 segmentos (5*5*5) de clientes que nos permiten saber cuántos existen de cada grupo y las características propias de estos para después accionar sobre ellos. Sin embargo, y en base a la segmentación recomendada por el proveedor en línea de servicios analíticos Putler (Anish, 2022), en el presente trabajo se considera para un análisis más efectivo y resumido de clientes 11 segmentos (Kabasakal, 2020), que se derivan de la estructura detallada en la *Tabla 5*.

Tabla 5. *Clasificación de segmentos efectivos de clientes RFM*

Segmento de clientes	Rango de puntuación de recencia	Rango de puntaje combinado de frecuencia y monetario
Campeones	4-5	4-5
Clientes leales	2-5	3-5
Lealista potencial	3-5	1-3
Clientes recientes	4-5	0-1
Prometedor	3-4	0-1
Necesitan atención	2-3	2-3
A punto de dormir	2-3	0-2
En riesgo	0-2	2-5
No puedo perderlos	0-1	4-5
Hibernado	1-2	1-2
Perdidos	0-2	0-2

Fuente: https://www.putler.com/rfm-analysis/

Al aplicar los criterios de clasificación de 11 segmentos de clientes a nuestro conjunto de datos, se identifica que el 52.5% de los consumidores son catalogados como “LEALES” o “POTENCIALES LEALES” para la compañía, mientras que el 12.7% son definidos como clientes “CAMPEONES” o “ESTRELLAS”, en razón que son los que mayor número de veces compraron,

su última compra fue realizada hace muy pocos días y generaron un mayor ingreso a la empresa (Ver *Tabla 6*).

Además, se observa que el 7.4% de los consumidores se encuentran en un estado de hibernación o totalmente perdidos para la organización, es decir, que son consumidores que han realizado muy pocas compras durante el año 2021, su último requerimiento de productos lo efectuó hace mucho tiempo, y con bajos montos de gasto.

Del mismo modo, se puede identificar la conformación de dos grupos de individuos importantes en cuanto a tamaño y cualidades, que deben ser de interés para la compañía para aumentar sus ventas y fidelización de clientes, estos grupos son los catalogados "EN RIESGO" y "NECESITAN ATENCIÓN". Los referidos segmentos cuentan con una representación del 10.1% y 9.7%, respectivamente.

Tabla 6. *Resultado de la segmentación efectiva de clientes RFM*

Segmento de clientes	No. Clientes
Campeones	3,171
Clientes leales	6,275
Lealista potencial	6,784
Clientes recientes	239
Prometedor	193
Clientes que necesitan atención	2,405
A punto de dormir	803
En riesgo	2,506
No puedo perderlos	669
Hibernado	1,438
Perdidos	402

De acuerdo al resumen de las métricas RFM detallado en la *Tabla 7*, los clientes en riesgo se caracterizan por presentar una recencia promedio de 102 días, sin embargo el 50% de estos realizaron su última adquisición en un tiempo máximo de 38 días, es decir un poco más de un mes; también efectuaron cerca de 20 compras durante el año 2021 y generaron en promedio un monto per cápita por pedidos de USD 112.40 a la compañía, pero existe un 5% de los clientes que generan un gasto por encima de USD 374.40.

Por otro lado, la empresa debe prestar atención a los consumidores que durante el año 2021 compraron productos por 19 ocasiones en promedio y generaron un ingreso medio per cápita por pedidos de USD 67.8, en razón que su última compra fue realizada hace 24 días, es decir cerca de un mes este grupo de consumidores no realizaron pedidos a la organización.

Tabla 7. *Resumen de las métricas RFM por segmentos de clientes*

Segmento de clientes	Recency			Frecuency			Monetary		
	Media	Mediana	Perc. 95	Media	Mediana	Perc. 95	Media	Mediana	Perc. 95
Campeones	4.9	1	9	45.3	19	81	272.3	17.2	1,409.7
Clientes leales	14.2	2	30	32.0	15	57	208.0	12.8	1,026.2
Lealista potencial	8.2	2	16	22.0	12	39	38.7	5.9	53.8
Clientes recientes	5.2	2	9	13.0	12	14	10.5	5.9	12.5
Prometedor	12.5	10	16	12.9	12	14	10.5	6.9	12.6
Necesitan atención	24.4	17	36	19.6	12	31	67.8	5.6	155.5
A punto de dormir	24.6	17	36	14.3	12	18	11.5	5.5	15.7
En riesgo	102.2	38	225	19.6	12	31	112.4	6.3	374.4
No puedo perderlos	90.9	38	179	31.7	19	50	317.7	17.2	1,610.8
Hibernado	104.7	38	212	15.6	12	22	14.4	5.9	22.3
Perdidos	105.7	38	222	12.9	12	14	10.4	5.5	12.5

Finalmente, en base a los diferentes segmentos de clientes obtenidos mediante la aplicación del método RFM, la empresa cuenta con información útil, ejecutiva y fácil de utilizar en los diferentes departamentos de la compañía para la toma de decisiones basada en sus propios datos, y establecer las distintas estrategias comerciales y de marketing, que son recomendadas por el proveedor en línea de servicios analíticos, Putler (Anish, 2022) en la siguiente tabla.

Tabla 8. *Detalle de las actividades y estrategias por segmentos de clientes*

Segmento de clientes	Actividad	Estrategia
Campeones	Compra recientemente, compra a menudo y gasta más	Recompensarlos. Pueden ser los primeros en adoptar nuevos productos. Promoverá su marca.
Clientes leales	Gasta buen dinero con la empresa. Sensible a las promociones.	Vender productos de mayor valor. Pide reseñas. Involúcrelos.
Lealista potencial	Clientes recientes, pero gastaron una buena cantidad y compraron más de una vez.	Ofrecer programa de membresía/fidelidad, recomendar otros productos.
Clientes recientes	Compras más recientemente, pero no a menudo.	Proporcionar apoyo de incorporación, brindarle éxito temprano, comience a construir una relación.
Prometedor	Compradores recientes, pero no han gastado mucho.	Crear conciencia de marca, ofrezca pruebas gratuitas
Necesitan atención	Por encima del promedio de actualidad, frecuencia y valores monetarios. Sin embargo, puede que no haya comprado muy recientemente.	Hacer ofertas por tiempo limitado, recomiende en base a compras anteriores. Reactivarlos.
A punto de dormir	Por debajo del promedio de actualidad, frecuencia y valores monetarios. Los perderá si no se reactiva.	Comparta recursos valiosos, recomiende productos populares/renovaciones con descuento, vuelva a conectarse con ellos.
En riesgo	Gastó mucho dinero y compró a menudo. Pero hace mucho tiempo. ¡Necesito traerlos de vuelta!	Envíe correos electrónicos personalizados para volver a conectarlos con la empresa, ofrezca renovaciones y brinde recursos útiles.
No puedo perderlos	Hizo compras más grandes, y con frecuencia. Pero hace mucho tiempo que no regresa.	Recupérelos a través de renovaciones o productos más nuevos, no los pierda ante la competencia, hable con ellos.

Hibernado	La última compra fue hace mucho tiempo, con bajos gastos y bajo número de pedidos.	Ofrecer otros productos relevantes y descuentos especiales. Recrear valor de marca.
Perdidos	Las puntuaciones monetarias, de frecuencia y de actualidad más bajas.	Reaviva el interés con una campaña de divulgación; de lo contrario, ignóralo.

4.2. Aplicación del análisis de Clúster a un segmento de clientes RFM

El análisis de clúster o de conglomerados es un amplio conjunto de técnicas multivariantes que permiten encontrar subgrupos de observaciones dentro de un conjunto de datos. Cuando agrupamos las observaciones o individuos, se busca que las observaciones en el mismo grupo sean similares entre estas y que los individuos en diferentes grupos sean diferentes.

La idea principal es caracterizar los conglomerados de manera que sean útiles para los objetivos del análisis o estudio de interés. Esta idea se ha aplicado en muchas áreas, incluida la astronomía, la arqueología, la medicina, la química, la educación, la psicología, la lingüística y la sociología (Shmueli, Bruce, Yahav, Patel, & Kenneth C. Lichtendahl, 2018).

En el ámbito del marketing, el análisis de clúster a tenido una amplia aplicación en la segmentación de mercados, donde los clientes se segmentan en función de la información demográfica y del historial de sus transacciones, y se adapta la estrategia comercial y de marketing para cada segmento, como es el caso en el presente trabajo.

Debido a que no hay una variable de respuesta, este es un método no supervisado, lo que implica que busca encontrar relaciones entre las observaciones sin ser entrenado por una variable de respuesta.

Tal como se señaló anteriormente, dentro del análisis de clúster existen dos procedimientos que permiten clasificar individuos, estos son los procedimientos jerárquicos y no jerárquicos. Su principal diferencia entre estos métodos radica en que el primero forma los grupos en pasos sucesivos y pueden analizar en cada paso las distancias entre los grupos formados, mientras que el segundo se basa en la asignación de los individuos en los conglomerados, que deben ser definidos previamente.

Los métodos de agrupamiento que crean una sola partición son más adecuados que una secuencia anidada de particiones. Esto significa que, en lugar de calcular todas las distancias entre todos los pares de observaciones en el conjunto de datos al comienzo de un análisis de

conglomerados jerárquico utilizando una implementación estándar, solo se calculan las distancias entre cada consumidor en el conjunto de datos y el centro de los segmentos.

De lo anterior y considerando las limitaciones del análisis de clúster jerárquico expuestas en el numeral 3.3.2.2, así como del coste de tiempo computacional, en el presente trabajo se aplicará el análisis de conglomerados no jerárquico mediante el método de k-medias, que es el algoritmo de aprendizaje automático no supervisado más utilizado para dividir un conjunto de datos dado en un conjunto de k grupos, donde k representa la cantidad de clúster preespecificados por el analista.

4.2.1. Análisis y tratamiento de los datos

Con la aplicación del modelo RFM se obtuvieron 11 segmentos de clientes con diferentes características y tamaño según lo detalla la *Tabla 7*, sin embargo, para mejorar el direccionamiento de las distintas estrategias comerciales y de marketing, es de vital importancia para la empresa obtener un mayor detalle de estos grupos de consumidores, de ahí que, la aplicación del análisis de conglomerados no jerárquicos mediante el algoritmo de k-medias, permitirá identificar y conocer subgrupos de clientes en cada segmento RFM.

4.2.1.1. Análisis descriptivo de los datos

Considerando las características, el tamaño y las estrategias a considerarse en cada uno de los segmentos obtenidos con las métricas de recencia, frecuencia y valor monetario, en este proyecto se aplicará una segmentación de los clientes catalogados “EN RIESGO” a través del análisis de clúster con el método de k-medias

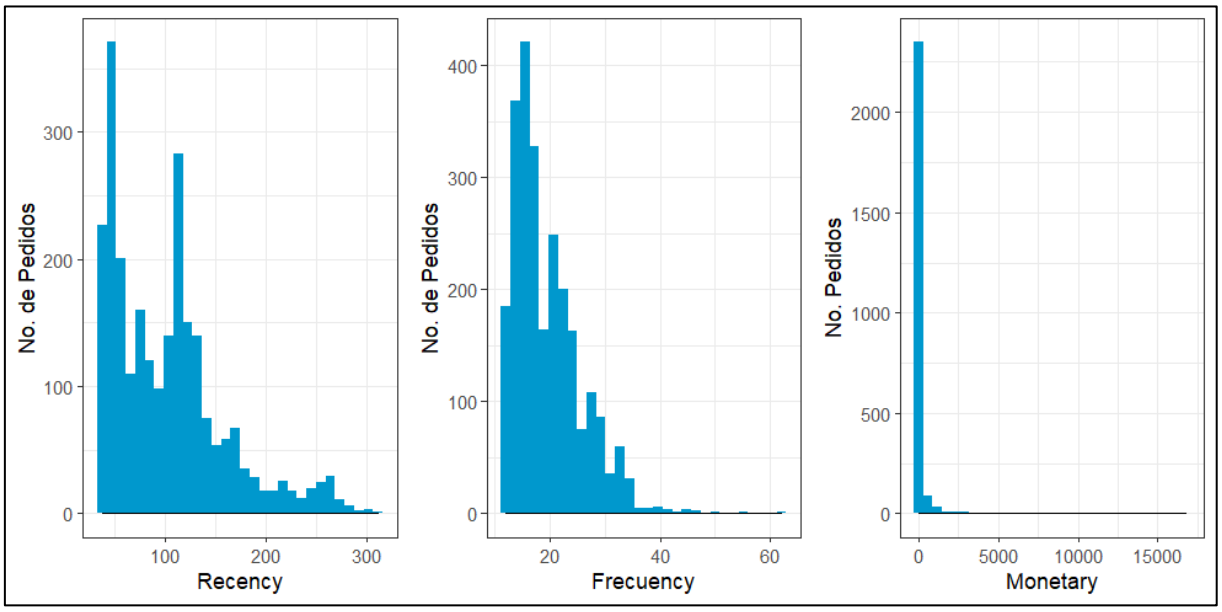
El grupo de clientes en riesgo para la organización está compuesto por 2,506 individuos que fueron caracterizados por presentar una recencia promedio de 102 días, con alrededor de 20 compras en promedio durante el año, las cuales generaron un monto medio per cápita por pedido de USD 112.40 a la compañía (Ver *Figura 14*).

Figura 14. *Resumen estadístico de las métricas RFM de los clientes en Riesgo*

```
> summary(clientes_rfm_riesgo[,c(2:4)])
    recency         frecuency        monetary       
 Min.   : 38.0   Min.   :12.00   Min.   :    6.293  
 1st Qu.: 52.0   1st Qu.:15.00   1st Qu.:   17.562  
 Median : 95.5   Median :18.00   Median :   24.722  
 Mean   :102.2   Mean   :19.58   Mean   :  112.422  
 3rd Qu.:129.0   3rd Qu.:23.00   3rd Qu.:   42.647  
 Max.   :312.0   Max.   :62.00   Max.   :16782.153  
```

Según los histogramas de frecuencias de las variables Recencia, Frecuencia y Valor Monetario, que se muestran en la *Figura 15*, se observa que en todas las métricas RFM existe una alta dispersión de los datos, pero destacando una mayor desviación del monto per cápita por pedido, en razón que el 90% de la distribución se concentra en un valor de hasta USD 135.70 (decil 9 de monetary = 135.70) mientras que el último decil (decil 10) se amplía hasta el valor máximo de 16,782.15.

Figura 15. *Histogramas de las métricas RFM de los Clientes en Riesgos*



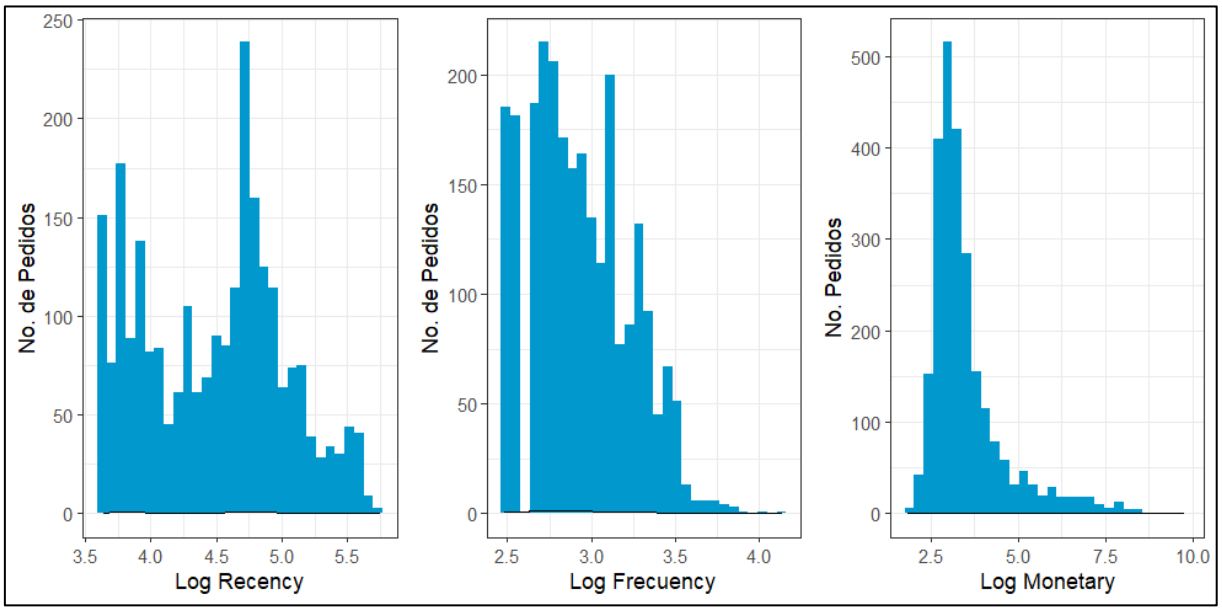
Asimismo, se observa una clara asimetría hacia el lado derecho en las tres métricas, resultando la presencia de un sesgo positivo de la distribución, el cual se corrobora con los coeficientes de asimetría obtenidos (recency = 1.08, frecuency = 1.20 y monetary = 17.70), lo cual conlleva a una evidente presencia de valores atípicos para las referidas variables que pueden afectar en gran medida en los resultados que se obtengan con el modelo de k-medias.

4.2.1.1. Tratamiento de los datos

Según lo señalado por Hair, Anderson, Tatham, & Black (1999, p. 69), existen varias soluciones a las problemáticas de alta dispersión, asimetría positiva y por ende la no normalidad de los datos, entre las cuales destacan la transformación raíz cuadrática, logarítmica o incluso la inversa de la variable, sin embargo, la conversión a logaritmo funciona de mejor manera para las variables con sesgo positivo.

De lo anterior, en el presente estudio se procede a transformar las métricas RFM por el logaritmo de estas, resultado de la cual se puede evidenciar que las tres variables mejoraron su distribución y la disminución del sesgo positivo que presentaban previamente (Ver *Figura 16*), pero las variables continúan mostrando la presencia de valores atípicos, los cuales pueden distorsionar la verdadera estructura de los grupos y hacen que los mismos no sean representativos de la verdadera estructura de la población.

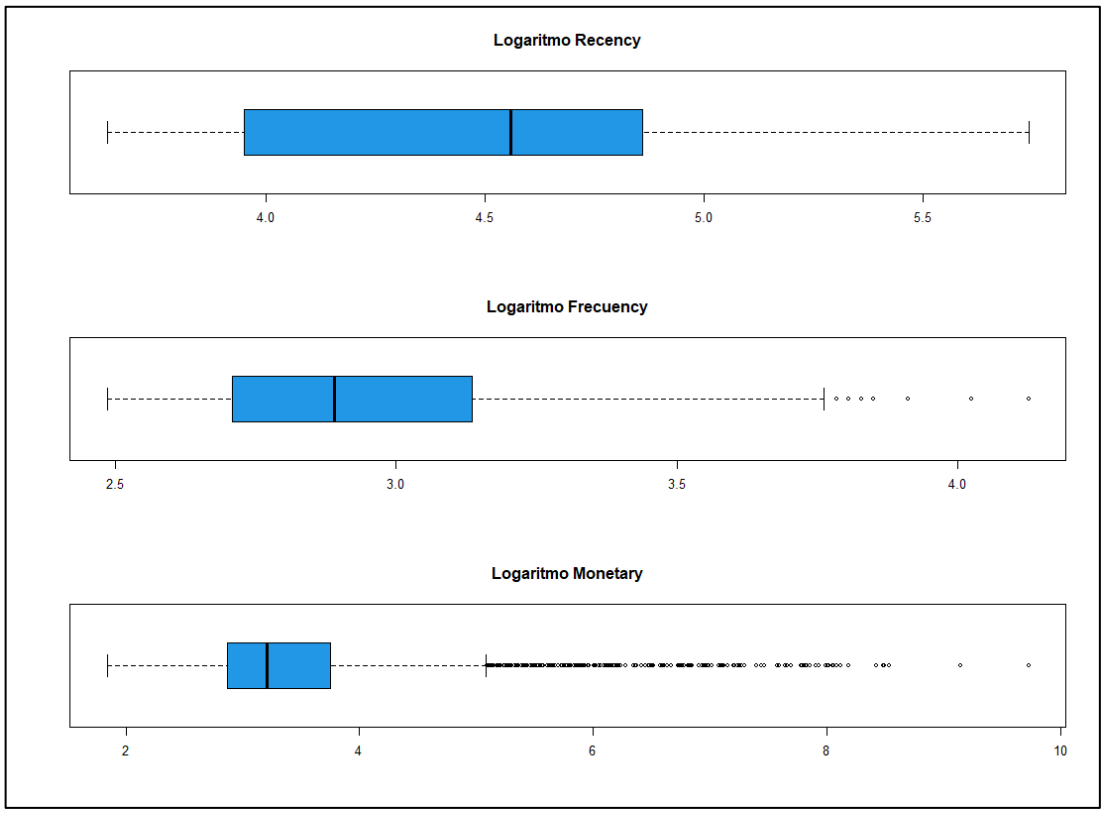
Figura 16. *Histogramas de las métricas RFM transformadas en Logaritmos*



A pesar de contar con valores atípicos en las variables transformadas, se deduce que las observaciones extraordinarias son valores válidos del comportamiento real de los clientes catalogados como "CLIENTES EN RIESGO" según el análisis RFM (Ver *Figura 17* 65), por tal razón, se decide en no aplicar ningún procedimiento de eliminación o reemplazo de estos, y por ende se mantienen para el análisis de clúster.

Posterior al análisis descriptivo de los datos de las variables recencia, frecuencia y valor monetario, es necesario que las métricas a considerarse por el modelo de K-medias, muestren una distribución homogénea, es decir que el valor promedio y desviación estándar sean similares. A este respecto, una vez corregido el sesgo positivo de las características, se puede volver a realizar una transformación de los datos utilizando uno de los mecanismos recomendados por Anderberg (1973, p. 103).

Figura 17. *Diagrama de cajas de las métricas RFM transformadas a Logaritmos*



Considerando las medidas de resumen de las métricas RFM transformadas a logaritmos, expuestas en la *Figura 1Figura 18*, se confirma que las tres variables expresadas en logaritmo no cuentan con una media y desviación estándar similar entre las mismas, por tal motivo, y con el objeto de darle una importancia por igual a la recencia, frecuencia y valor monetario en el análisis de clúster por k-medias, se estandarizan los valores a una distribución normal con media = 0 y desviación = 1 (Ver *Figura 19*).

Figura 18. *Resumen de las métricas RFM transformadas a Logaritmos*

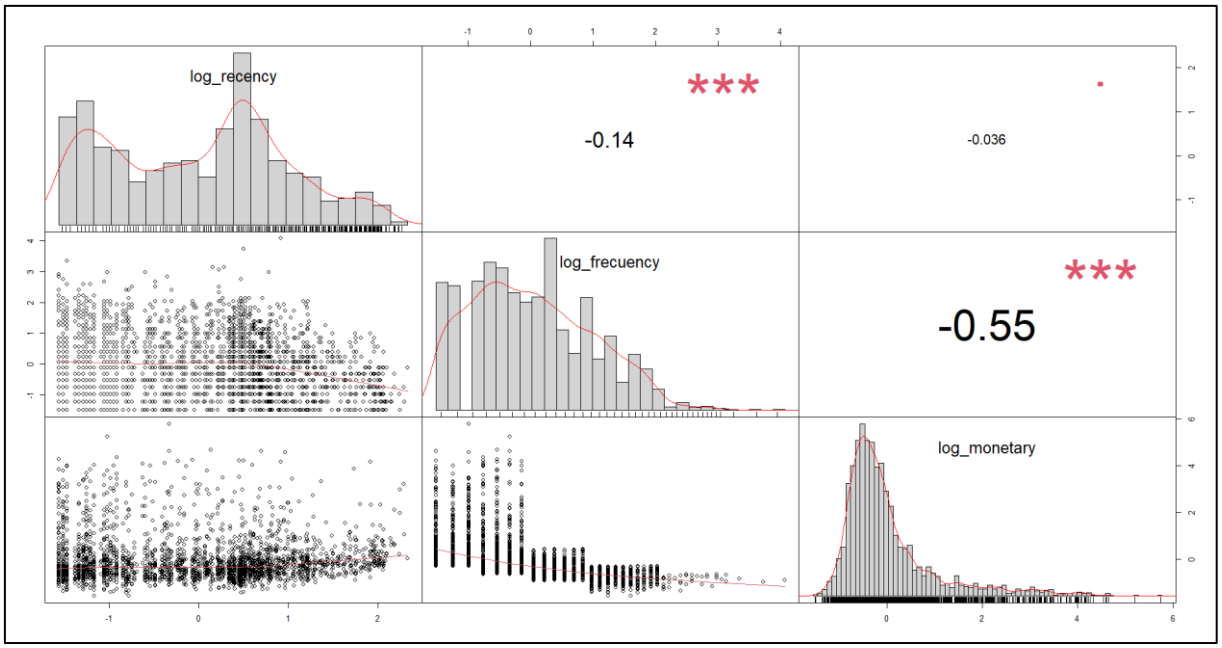
```
> describe(data_cluster)
             vars    n mean   sd median trimmed  mad  min  max range skew kurtosis   se
log_recency     1 2506 4.48 0.54   4.56    4.47 0.61 3.64 5.74  2.11 0.08    -0.98 0.01
log_frecuency   2 2506 2.93 0.29   2.89    2.91 0.30 2.48 4.13  1.64 0.44    -0.43 0.01
log_monetary    3 2506 3.52 1.08   3.21    3.32 0.59 1.84 9.73  7.89 2.02     4.64 0.02
```

Figura 19. *Resumen de las métricas RFM estandarizadas*

```
> describe(data_cluster)
             vars    n mean sd median trimmed  mad   min  max range skew kurtosis   se
log_recency     1 2506    0  1   0.14   -0.03 1.13 -1.57 2.34  3.90 0.08    -0.98 0.02
log_frecuency   2 2506    0  1  -0.13   -0.05 1.01 -1.51 4.08  5.59 0.44    -0.43 0.02
log_monetary    3 2506    0  1  -0.29   -0.18 0.55 -1.56 5.77  7.33 2.02     4.64 0.02
```

Con respecto a la relación que presentan las tres variables estandarizadas, observamos que no existen correlaciones significativas entre las diferentes métricas RFM, a pesar del grado de relación entre frecuency y monetary que oscila en -0.55, sin embargo, en el gráfico de dispersión de las dos variables no muestran un comportamiento que tienda a un problema de multicolinealidad (Ver *Figura 20*).

Figura 20. *Diagrama de correlación de las métricas RFM*



4.2.2. Determinación del número óptimo de clúster

Una vez finalizada la etapa de la descripción y preparación de los datos, conlleva a dar respuesta a la interrogante de ¿Cuántos grupos deben formarse en el conjunto de datos con el método de k-medias?, por tal razón y siguiendo lo aplicado por Dolnicar, Grun, & Leisch (2018, pp. 97-98), Kassambara (2017, p. 40) y Laude (2017, pp. 392-394) para la determinación del número k de grupos se consideran, los criterios del Codo y de Silueta, y en el caso de existir diferencias entre los valores de k obtenidos entre los dos procedimientos, se tomará como decisión el resultado de las frecuencias de los distintos criterios de validación señalados por Charrad, Ghazzali, Boiteau, & Niknafs, (2014, p. 19).

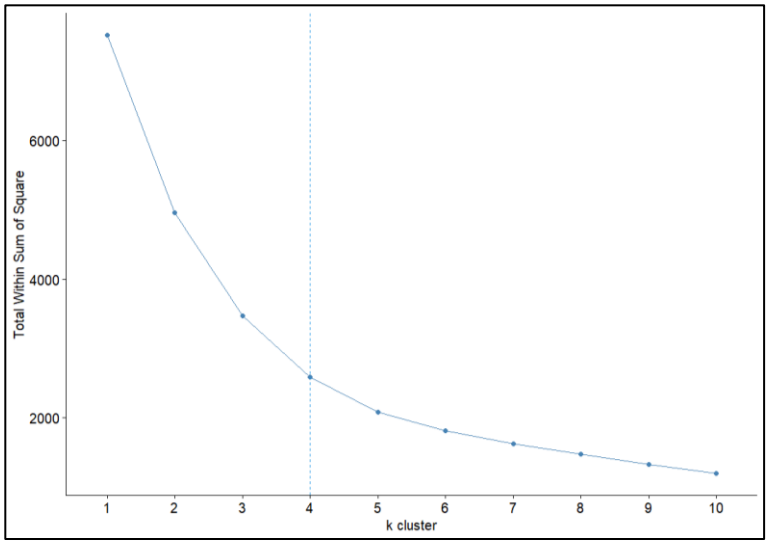
Como se indicó anteriormente, el método del codo emplea la distancia media de las observaciones a su centroide, es decir, considera las distancias intra-cluster. Por tanto, cuanto menor es la distancia intra-cluster, los clústers son más compactos y por ende se maximiza la

varianza inter-cluster. Este método examina el valor k que satisfaga que un incremento de k, no mejore sustancialmente la distancia media intra-cluster.

4.2.2.1. Criterio de Elbow Method o método de codo

En la *Figura 21* se muestra el comportamiento de la suma de cuadrados dentro de los grupos para cada valor de k clúster, resultando de lo cual se distingue que del pasar de una segmentación con 4 grupos a 5, existe una mínima reducción de la suma de cuadrados, lo cual implica que no existe una mejora significativa en la distancia media dentro de los grupos (intra-cluster), razón por la cual, con la aplicación del método de codo se considera pertinente incorporar un valor de k = 4.

Figura 21. *Número óptimo de clusters de Clientes - Método del Codo*

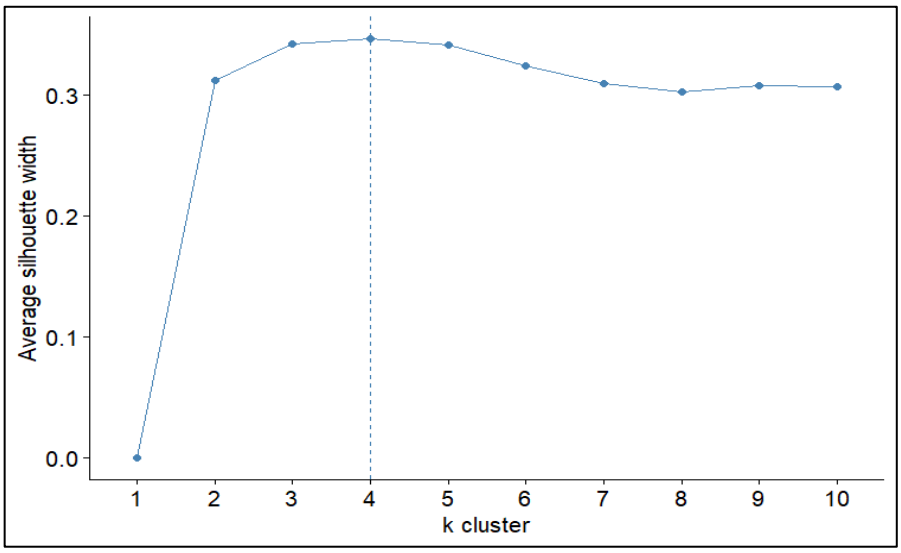


4.2.2.1. Criterio de Average Silhouette Method o método de la silueta promedio

Por su parte, el método de la silueta promedio mide la distancia de separación entre los clústers. Es decir, determina qué tan bien se encuentra cada objeto dentro de su grupo. Este método calcula la silueta promedio de las observaciones para diferentes valores de k grupos. El número óptimo de conglomerados k es el que maximiza la silueta promedio sobre un rango de valores posibles para k.

Según los resultados obtenidos en la *Figura 22*, el número de k grupos que maximiza la silueta promedio entre grupos, es el valor de k = 4, de tal forma que se corrobora con el número de grupos óptimos encontrados mediante la aplicación del criterio de Elbow Method al conjunto de datos en estudio.

Figura 22. *Número óptimo de clusters de Clientes - Método de la Silueta*



4.2.3. Implementación del análisis de Clúster con el método de K-medias

Dado que los dos criterios sugieren 4 como el número de conglomerados óptimos, se procede a implementar el modelo de k-medias para particionar y etiquetar los datos, y así posteriormente observar cómo se agrupan los clientes en función de las variables de recencia, frecuencia y valor monetario.

En la *Tabla 9*, se describen los valores de la media, mediana y el percentil95 de cada una de las variables originales RFM y el número de clientes para cada uno de los clústeres obtenidos mediante el modelo k-medias.

Tabla 9. *Resumen de las métricas RFM por segmentos de clientes*

Métricas RFM	Medidas	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Recency	Media	78.9	54.6	152.9	84.9
	Mediana	38	38	88	38
	Perc. 95	154	81	260	138
Frecuency	Media	14.2	16.7	16.7	27.3
	Mediana	12	12	12	19
	Perc. 95	18	21	22	35
Monetary	Media	867.3	36.2	37.9	17.0
	Mediana	107	13	13	6
	Perc. 95	2,995	87	93	27
Número de Casos		245	596	925	740

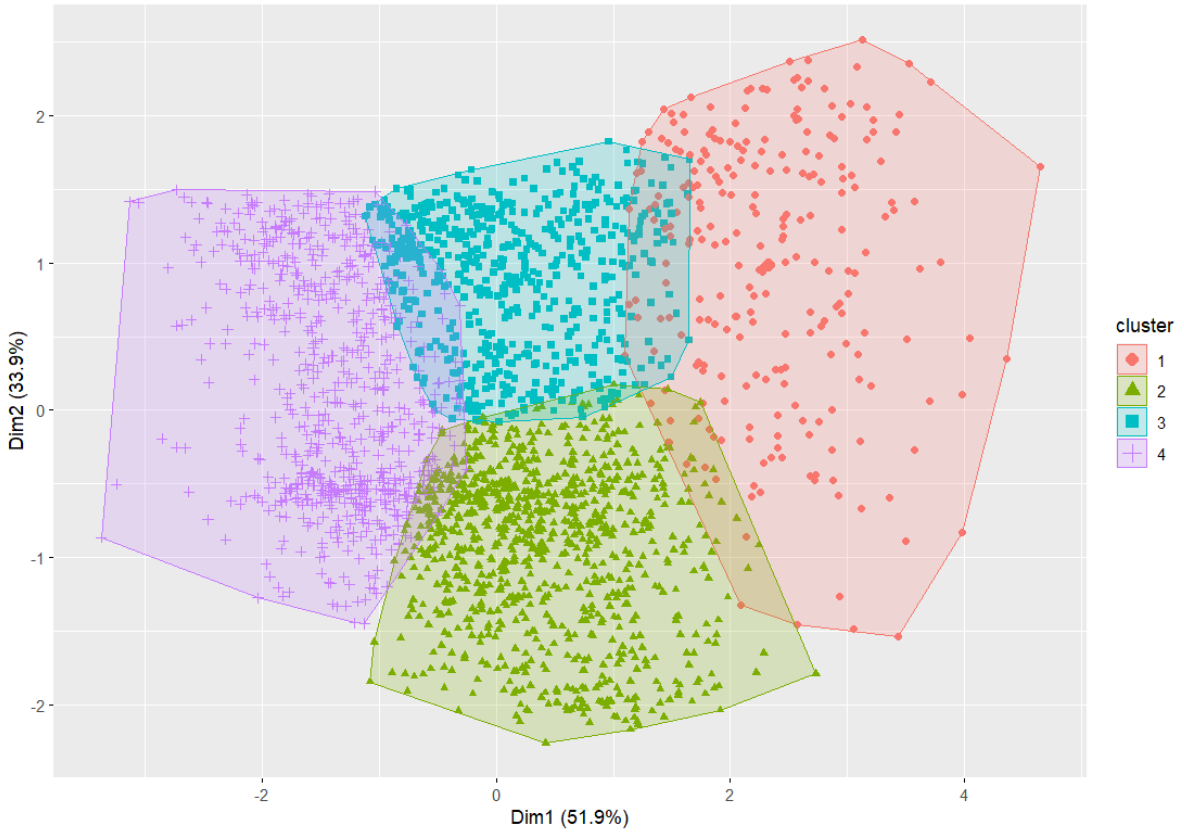
Al revisar las medidas de resumen de los 4 segmentos obtenidos, se puede observar que el clúster 1 presenta diferencias importantes en el valor medio percapita de pedidos, sin embargo, su frecuencia de compra es inferior a los demás grupos y tiene un tiempo promedio

de 79 días desde que realizó la última compra en la compañía. Por el otro lado, se encuentran los clientes del clúster 4, quienes realizaron un gasto mínimo a la empresa, a pesar de haber realizado mayores requerimientos en promedio que los demás segmentos y de tener un valor próximo de recencia que el grupo 1.

Asimismo, se examina que los grupos 2 y 3 presentan similares puntuaciones media de monetary y frecuency, pero sus valores promedio de recency son totalmente opuestos, los clientes del cluster 2 son quienes realizaron compras a la empresa en los últimos 54 días en promedio, mientras lo contrario sucede con el segmento 3 donde la tasa media de recencia que presentan es la de mayor número y diferencia con relación a todas las agrupaciones.

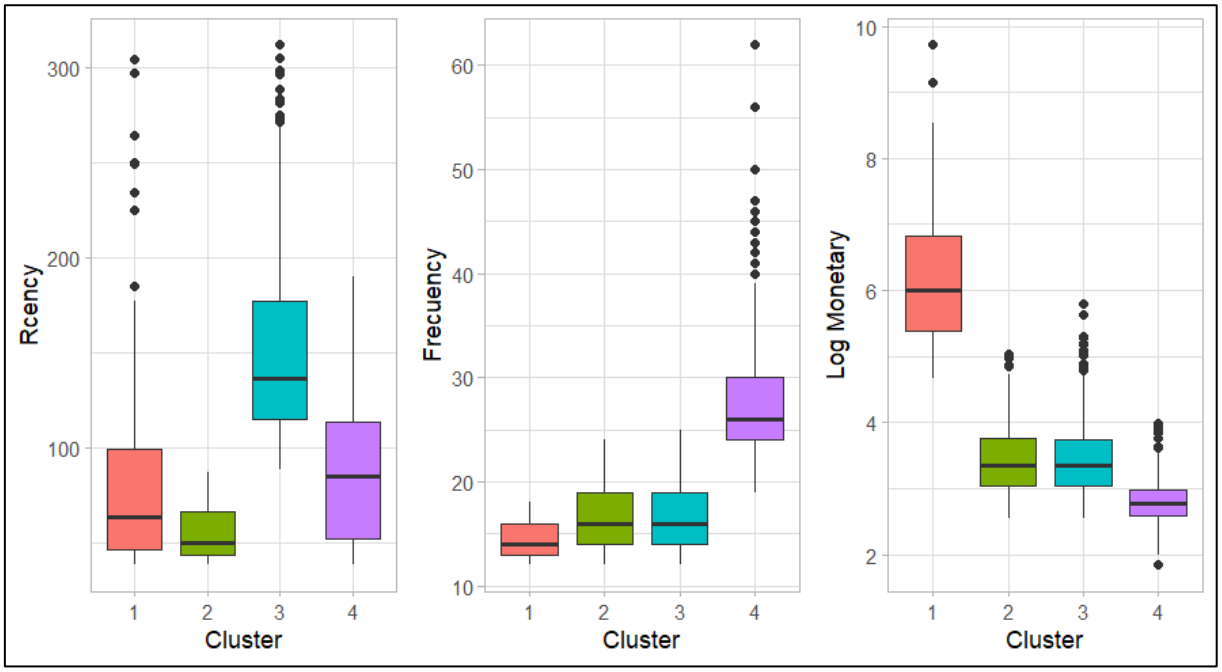
En la *Figura 23*, se delimitan de forma visual la segmentación de clientes obtenida mediante el análisis de conglomerados con el modelo de k-medias, donde se observan los contrastes que presentan los consumidores del grupo 1 vs el grupo 2 en la primera componente del gráfico, así como, la discrepancia que existen entre los individuos de los clústeres 2 y 3 con relación a la segunda componente.

Figura 23. *Segmentación de clientes con 4 Clúster*



Adicional a los resultados mostrados, es importante profundizar en el análisis de los clústeres con el uso de otras herramientas visuales que son utilizadas de forma amplia para mostrar los resultados de la segmentación en el marketing, como lo son los gráficos de densidad sobre las distintas variables empleadas en el análisis.

Figura 24. *Distribución de las Métricas RFM por Clúster*



Considerando los diagramas de caja y bigote mostrados en la *Figura 24*, se observan las disimilitudes que existen entre los grupos de clientes catalogados en riesgo para la empresa y obtenidos mediante el análisis de clúster, eso conlleva que la compañía permita reconfigurar las estrategias establecidas de forma general para el segmento 8 determinado con el análisis RFM, y de ser el caso aplicar una priorización a ciertos compradores que facilite el incremento de las ventas y fidelización, sin efectuar una campaña total a los 2,506 consumidores.

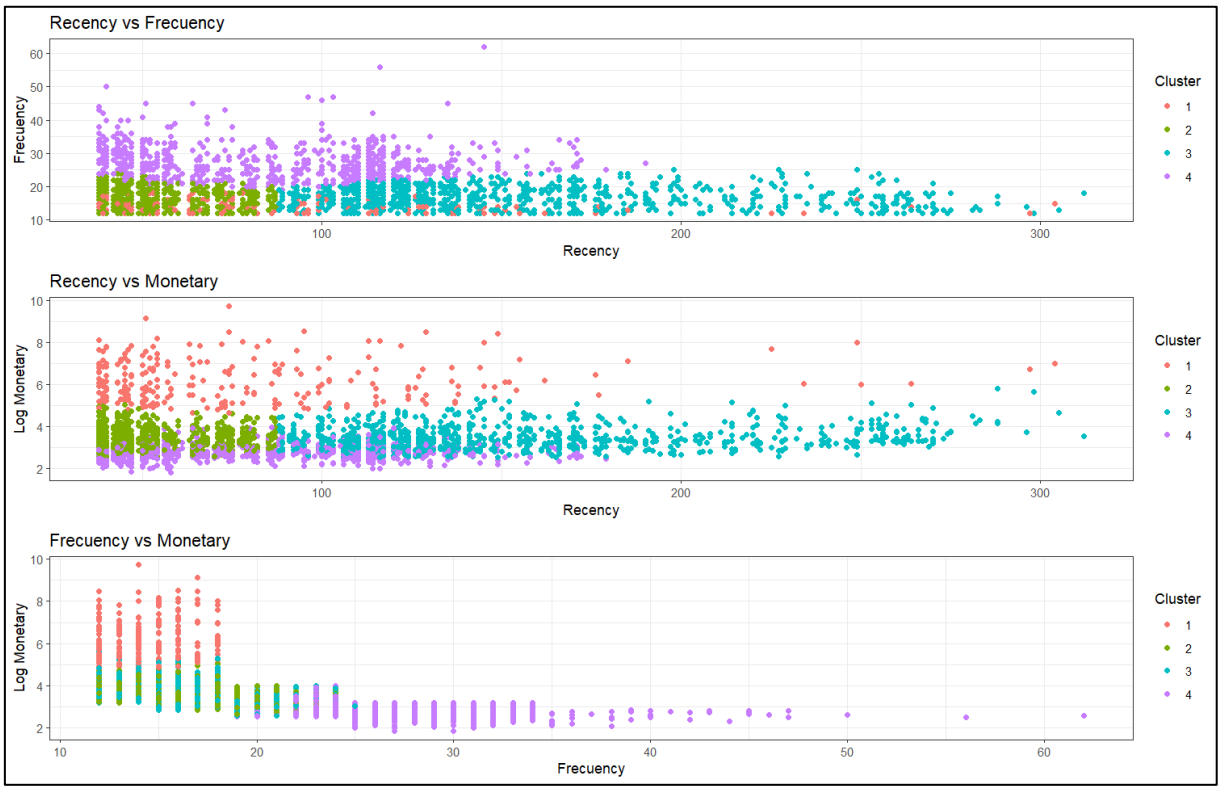
De igual forma, la diferencia existente de los grupos a nivel univariante se logra asemejar en los comportamientos de los individuos entre pares de las métricas RFM (Ver *Figura 25*), es así, que para el caso de los grupos 1, 3 y 4 se encuentran influidos fuertemente por las variables monetary, recency y frecuency, respectivamente. Exceptuando al cluster 2 que presenta un proceder análogo al cluster 3 para las métricas de frecuencia y valor monetario, pero variado claramente en las bajas tasas de recencia.

Siguiendo con el análisis de los cuatro segmentos de clientes que previamente fuesen clasificados por el modelo RFM y catalogados como CLIENTES EN RIESGOS, es de apremio

efectuar una descripción más detallada de los mismos, fusionando los resultados entre las dos metodologías de segmentación aplicadas en el presente trabajo, los cuales se detallan en el *Anexo F*, y que a breve rasgos precisan lo siguiente:

- El clúster 1 está representado por 245 clientes que gastan en promedio USD 867 por cada pedido efectuado a la compañía, y un 50% de éstos generan ingresos para la compañía en al menos 404.92 por cada requerimiento realizado. Además, se podría mencionar que a pesar de presentar un valor importante de monetary, este segmento no realiza muchas compras en el año, apenas una media de 14.2, pero lo preocupante para la empresa, es que exhiben una tasa media de recencia de 78.9, es decir que son clientes que dejaron de comprar productos hace más de dos meses.
- El clúster 4 representa a 742 consumidores con el menor valor monetario de todos los segmentos, USD 17 en promedio por pedido, pero a pesar de aquello, muestran el mayor número promedio de requerimientos que han realizado todos los clientes catalogados en riesgo, 27 pedidos, pero al igual que el grupo 1, llevan cerca de tres meses que no han realizado adquisición de bienes a la compañía.
- El clúster 3 está compuesto por 923 compradores con el mayor número promedio de días sin adquirir productos a la empresa en relación con los demás segmentos, más de cinco meses (152.9). El valor monetario en promedio por pedido es de 37.90, alcanzando hasta los USD 93.00 en el 95% de los clientes que lo conforman (percentil 95), mientras que el número promedio de pedidos efectuados en el año 2021 a la empresa es de alrededor de 17. Adicionalmente, se identifica que existen un subgrupo de 143 consumidores que gastan más de USD 90 en promedio por cada requerimiento, pero que alcanzan una tasa de recencia superior a los 150 días.
- En cuanto al clúster 2, se podría indicar que está compuesto por clientes con similares comportamientos en las variables de frecuencia y valor monetario que lo evidenciado en el grupo 3, (frecuency media = 16.7 y monetary media = USD 36.2); pero a pesar de aquello, muestran una evidente disimilitud en términos de la recencia con el segmento en comparación, alcanza el valor medio de días más bajo de todos los clúster, cerca de 55 días, lo cual representa a una tercera parte del tiempo medio que evidencia el cluster 3.

Figura 25. *Distribución Bivariantes de los Clientes por las métricas RFM y Clúster*



Como resultado de la aplicación de los métodos RFM y k-medias, se puede sintetizar que las agrupaciones obtenidas, resultantes de la combinación de la segmentación por el análisis de clúster y los grupos de clientes establecidas en función de la cuantificación y agrupación de los puntajes de recencia, frecuencia y valor monetario, brindan una enorme cantidad de información útil para la toma de decisiones de la compañía, la misma, que se hace imprescindible de ser transmitida de manera práctica e iterativa, por aquello resulta necesario la implementación y despliegue de una herramienta de Inteligencia de Negocio, BI, que facilite la visualización y el análisis de toda la información conseguida.

4.3. Implementación y despliegue de la Herramienta de Inteligencia de Negocio

Con el propósito de dar cumplimiento a uno de los objetivos planteados en el presente proyecto, así como de disponer de información útil para la toma de decisiones, se ha utilizado la herramienta de Inteligencia de Negocio “Power BI” para sintetizar y reflejar de manera clara y práctica los resultados obtenidos de la segmentación de clientes.

El procedimiento de implementación y despliegue de la herramienta BI consistió en primer lugar, en integrar las tablas de datos empleadas y preparadas durante la etapa de análisis y definición, así como de la ejecución de los modelos de segmentación, entre las herramientas Power BI y R de forma fácil y directa, sin la necesidad de generar, almacenar e importar datos de archivos planos.

Tal como se señaló en la etapa de análisis y definición, la tabla de pedidos contiene todos los datos derivados de las transacciones diarias que realiza la empresa y sus clientes, las tablas de clientes y clientes RFM reflejan toda la información relacionada con cada uno de los consumidores y los resultados de las segmentaciones mediante los criterios de Recencia, Frecuencia y Valor Monetario, combinados con el análisis de clúster, tal como se comprueba en el Anexo G.

Ejecutado el proceso de integración de datos entre las herramientas analíticas empleadas para el efecto, se han generado diferentes visualizaciones para reflejar en primera instancia el comportamiento de los requerimientos que realizan los clientes a la compañía con diferentes niveles de desagregación, como segundo apartado, la conducta y composición mediante las segmentaciones obtenidas con el análisis RFM y por último el comportamiento de los consumidores catalogados como CLIENTES EN RIESGO.

La información que se muestra en los siguientes cuadros de mando implementados en la herramienta de Inteligencia de Negocio Power Bi, será de cuantiosa utilidad para sustentar la toma de decisiones de la compañía en todos sus Departamentos, que permitan cumplir con el objetivo establecido por la empresa en el año 2022, que es el de incrementar las ventas y la fidelización de clientes.

Figura 26. *Cuadro de mando de la Situación de los Clientes de la empresa durante el período 2019 - 2021*

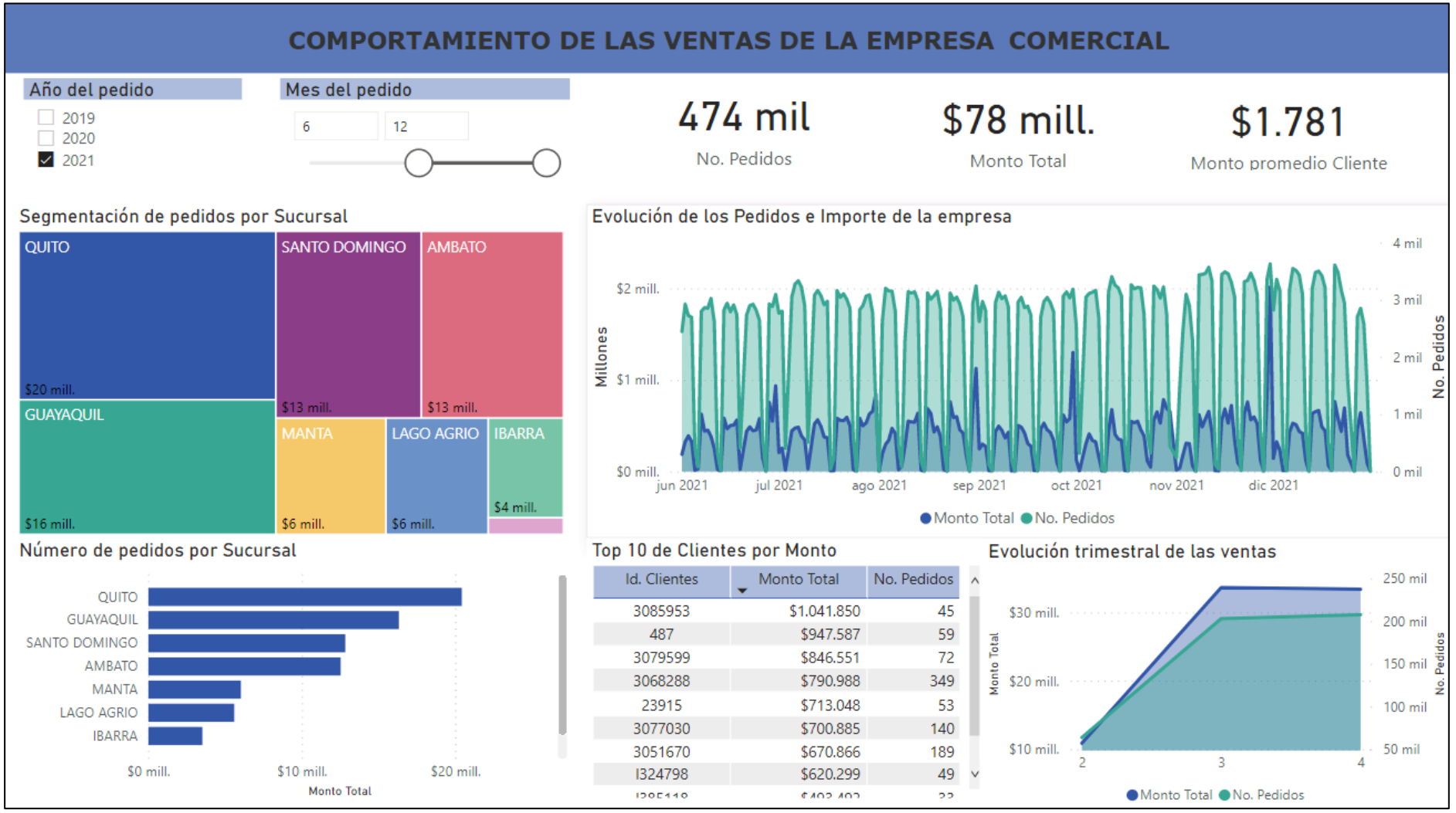
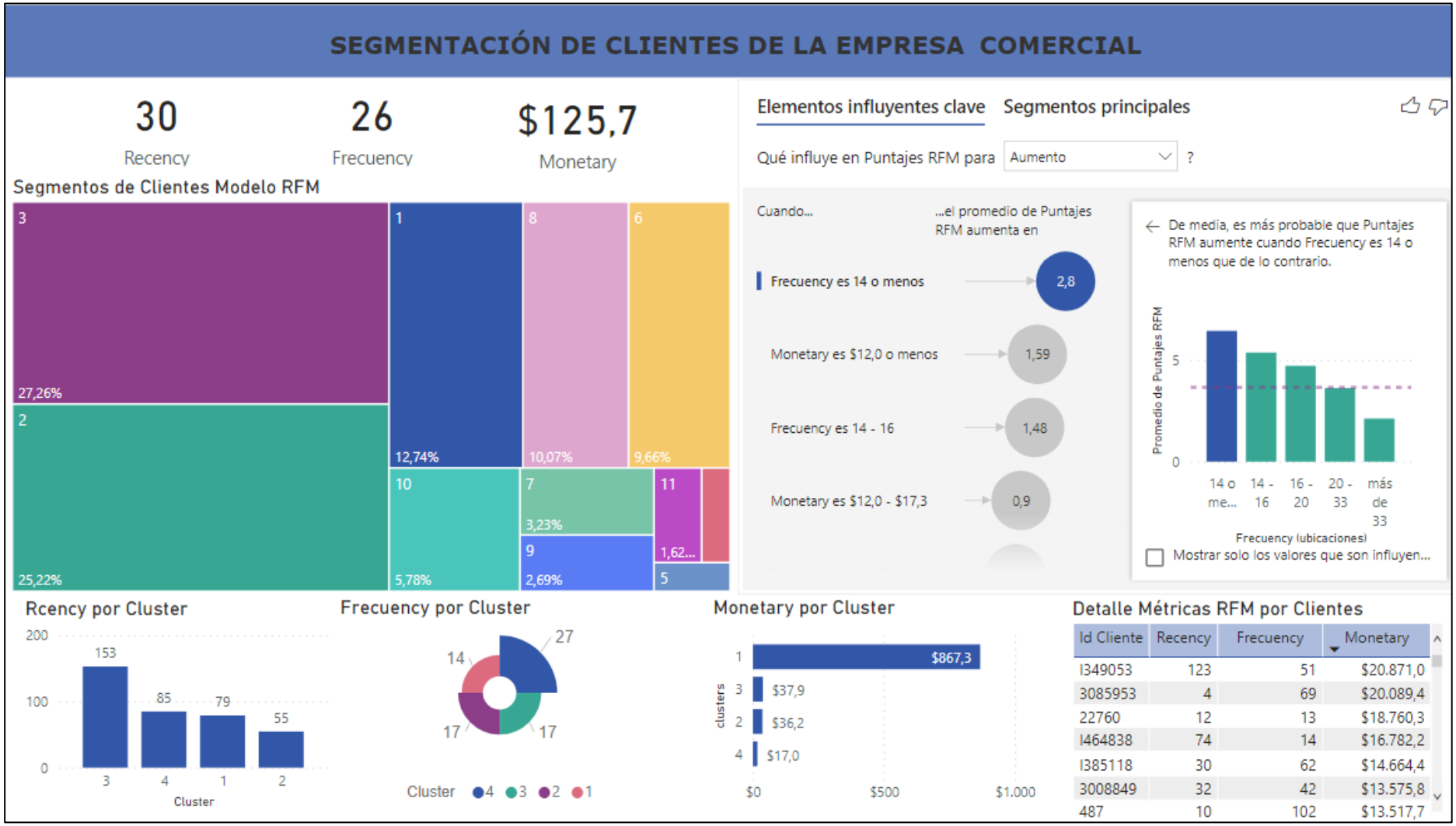


Figura 27. *Cuadro de mando de la Segmentación de Clientes mediante el análisis RFM y modelo K-medias para el año 2021*



5. Cronograma del proyecto

Durante la etapa de alcance y planificación del presente proyecto de inteligencia de negocio, se consideró 12 semanas como tiempo de duración del mismo, iniciando desde el 24 de marzo de 2022 con el análisis y exploración de requerimientos del proyecto, hasta la creación de los dashboard que contienen los resultados de la segmentación de clientes (Ver *Tabla 1*).

Considerando las etapas de un proyecto de Inteligencia de Negocio y el cronograma establecido en el presente trabajo, se detalla la secuencia de tareas que se han ejecutado para los distintos swim lane del proyecto:

Detalle del Swimlane	Principales Actividades	Mar. 22 Sem.4	Abr.22 Sem.1	Abr.22 Sem.2	Abr.22 Sem.3	Abr.22 Sem.4	May.22 Sem.1	May.22 Sem.2	May.22 Sem.3	May.22 Sem.4	Jun.22 Sem.1	Jun.22 Sem.2	Jun.22 Sem.3
Inteligencia de Negocio	Prioridades y Presupuesto												
	Requisitos de negocio y datos												
	Diseño de soluciones de BI preliminares.												
	Recopilar comentarios de los usuarios y realizar pruebas unitarias.												
	Refinar y publicar entregables.												
	Realizar sesiones de feedback con usuarios seleccionados.												
Datos y bases de datos	Examinación de las fuentes de datos actuales e históricas.												
	Modelado de los datos												
	Almacenamiento de datos												
	Prueba de datos												
	Análisis y presentación de datos												
Integración de Datos	Análisis de sistemas fuente.												
	Procesos de integración de datos.												
Infraestructura	Tecnología necesaria												
	Adquisición de tecnología												
	Instalación y uso de la tecnología												
Gestión de proyecto	Justificación de negocio.												
	Gestión del cambio												
	Programación y planificación de recursos.												
	Informe de estado.												
	Coordinación entre el equipo del proyecto y otras personas												
	Gestión de riesgos y planificación de contingencias.												

6. Conclusiones

1. La segmentación de clientes mediante la aplicación combinada del modelo RFM y el análisis de clúster con el método de k-medias, permite contar con información sintetizada, oportuna e intuitiva a las empresas comerciales y de servicios, sobre la conformación y descripción de grupos de clientes con similares características dentro de cada segmento y heterogéneos entre éstos, con el objetivo de viabilizar y priorizar las campañas de marketing y fidelización sustentada en datos.
2. Con la aplicación del modelo RFM a la base de datos de pedidos facilitada por una empresa comercial y de distribución de Ecuador, se logró obtener una segmentación inicial de 125 segmentos o grupos de clientes, por la combinación de los quintiles de cada una las métricas de Recency, Frecuency y Monetary, no obstante, se generó una conformación adicional y resumida de 11 segmentos, según los parámetros definidos

por el proveedor en línea de servicios analíticos Putler, que facilita a la compañía la selección y análisis de los clientes de interés para la misma.

3. De los 11 grupos de clientes obtenidos mediante la combinación de las variables de recency, frecuency y monetary, se logró identificar que cerca del 13% de los consumidores son catalogados como CAMPEONES, es decir son nuestros clientes estrellas, por el otro lado, se pudo constatar que aproximadamente el 2% de los compradores se consideran como PERDIDOS, en razón que los mismos han realizado muy pocas compras, con valores económicos muy bajos y hace mucho tiempo su última adquisición. Por último, resaltar a los clientes clasificados como EN RIESGO, quienes representan el 10% de la población de estudio y son importantes para la compañía por efecto a sus puntajes de frecuencia y valor monetario, pero que han realizado su última compra hace un tiempo significativo.
4. Con la aplicación del análisis de clúster o conglomerados mediante el método de k-medias se logró obtener una mayor especificidad del grupo de clientes denominados EN RIESGO, resultando la identificación de 4 subgrupos con diferentes comportamientos entre los mismos, lo que permitirá a la empresa en direccionar de mejor manera sus estrategias comerciales y de marketing para no perder a este importante segmento de consumidores, en especial a los compradores del subgrupo 1 que presentan el mayor valor monetario en relación a los restantes 3 subgrupos.
5. La creación de informes interactivos e integrados con la herramienta de análisis de datos implementada en el presente trabajo ha facilitado la visualización y el estudio de los segmentos de clientes de la compañía para la toma de decisiones.
6. Con la implementación de la metodología para el desarrollo de los proyectos de Inteligencia de Negocio, en el presente trabajo se ha cumplido con los objetivos planteados inicialmente, consecuente por la obtención de los resultados del análisis descriptivo del conjunto de datos correspondiente a los pedidos y clientes de la compañía, la creación de segmentos de clientes mediante el modelo RFM y su ampliación de la segmentación con el análisis estadístico de clúster mediante el método de k-medias, y la creación de una solución integrada y dinámica de visualización de información útil para la toma de decisiones en la empresa.

7. Limitaciones y prospectiva

- En el presente trabajo se generó una segmentación de clientes utilizando únicamente las métricas de Recencia, Frecuencia y Valor Monetario con la información general de los pedidos, sin embargo, la compañía debe incorporar nuevas variables sociodemográficas para fortalecer la segmentación y extraer mayor conocimiento de sus compradores.
- La recolección de los datos de ubicación geográfica de los clientes de la empresa comercial y distribución presenta una serie de errores y precisiones en sus coordenadas, lo que implica una imposibilidad de fortalecer el análisis de los segmentos de clientes a nivel de territorio, razón por la cual, la compañía debe realizar de manera urgente el levantamiento e integración de la información de latitud y posición de cada uno de los clientes mediante el uso de herramientas SIG, priorizando a los segmentos de consumidores que se encuentran en una situación de Riesgo de pérdida.
- La empresa cuenta con una herramienta de inteligencia de negocio que contiene varios reportes de información que se pueden considerar como básicos, desactualizados y no dinámicos para la toma de decisiones, es por aquello, que es imprescindible la conformación de un Departamento de BI en la compañía que trabaje articulada con los demás departamentos e implemente soluciones de negocios trasversales, dinámicas y actualizadas para generar conocimiento sobre los clientes y negocio.
- La segmentación de clientes mediante la aplicación combinada del modelo RFM y análisis de clúster ha permitido identificar y conocer comportamientos de los clientes de la empresa únicamente del último año, es por aquello que se debe continuar con la ampliación del análisis para los años anteriores, así como de los meses más recientes, con el único objetivo de proveer información útil, actualizada y resumida del estado del negocio a todos los mandos de la organización para la toma de decisiones basada en datos.

Referencias bibliográficas

Álvarez, R. (1995). *Estadística multivariante y no paramétrica con SPSS.* Lavel, S.A.

Anderberg, M. R. (1973). *Cluster analysis for applications.* New York, San Francisco, London: Academic Press Inc.

Anish, N. (2022, Enero 4). *Putler.* Retrieved from Putler - ecommerce analytics software: https://www.putler.com/rfm-analysis/

Birant, D. (2011). Data Mining Using RFM Analysis. In K. Funatsu, & K. Hasegawa, *Knowledge-Oriented Applications in Data Mining* (pp. 91-108). Rijeka, Croacia: IntechOpen. doi:10.5772/13683

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014, Octubre). NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software, 61*(6), 36. Retrieved from http://www.jstatsoft.org/

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Database Marketing & Customer Strategy Management, 19*(3), 197-208. Retrieved from www.palgrave-journals.com/dbm

Cuadros López, Á. J., Gonzales Caicedo, C., & Jiménez Oviedo, P. C. (2017). Análisis multivariado para segmentación de clientes basada en RFM. *Tecnura*(21(54)), 41-51. doi:https://doi.org/10.14483/22487638.12957

Dolnicar, S., Grun, B., & Leisch, F. (2018). *Market Segmentation Analysis.* Brisbane, Australia; Linz, Austria; Bettina Grün: Springer Open. Retrieved 05 29, 2022, from http://www.springer.com/series/10101

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis 5th ed.* Noida: John Wiley & Sons, Ltd.

Glotzer, S. (2022, Abril 29). *Simla*. Retrieved from Simla.com: https://www.simla.com/blog/para-que-sirve-un-analisis-rfm-y-como-utilizar-los-resultados

Hair, J., Anderson, R., Tatham, R., & Black, W. (1999). *Análisis Multivariante* (5.a ed. ed.). Prentice Hall Iberia.

Jun, W., Li, S., Wen-Pin, L., Sang-Bing, T., Yuanyuan, L., Liping, Y., & Guangshu, X. (2020). An Empirical Study on Customer Segmentation by Purchase Behaviors Using a RFM Model and K-Means Algorithm. *Mathematical Problems in Engineering, 2020*, 1-7. doi:10.1155/2020/8884227

Kabasakal, İ. (2020, 01). Customer Segmentation Based On Recency Frequency Monetary Model: A Case Study in E-Retailing. *Bilişim Teknolojileri Dergisi, 13*, 47-56. doi:10.17671/gazibtd.570866

Kassambara, A. (2017). *Practical Guide To Cluster Analysis in R: Unsupervised Machine Learning.* STHDA.

Kumar, V., & Reinartz, W. (2018). *Customer Relationship Management: Concept, Strategy, and Tools - Third Edition.* Berlin: John Wiley & Sons.

Laude, H. (2017). *Data Scientist y lenguaje R: Guía de autoformación para el uso de Big Dta.* Barcelona: Editions ENI.

Miglautsch, J. (2002). Application of RFM principles: What to do with 1–1–1 customers? *Journal of Database Marketing, 9*(4), 319-324.

Palakshappa, A., & Patil, M. M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University - Computer and Information Sciences, 34*(5), 1785-1792. Retrieved from https://doi.org/10.1016/j.jksuci.2019.12.011.

Sherman, R. (2014). *Business intelligence guidebook : from data integration to analytics.* Waltham: Elsevier Science; Morgan Kaufman.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Kenneth C. Lichtendahl, J. (2018). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R.* Hoboken: John Wiley & Sons.

Wei, J. T., Lin, S. Y., & Wu, H. H. (2010). A review of the application of RFM model. *African Journal of Business Management, 4*(19), 4199-420. Retrieved from http://www.academicjournals.org/AJBM

Wu, H.-H., Chang, E.-C., & Lo, C.-F. (2009). Applying RFM Model and K-Means Method in Customer Value Analysis of an Outfitter. In S.-Y. Chou, A. Trappey, J. Pokojski, & S. Smith, *Global Perspective for Competitive Enterprise, Economy and Ecology* (pp. 665-672). London: Springer London.

Anexos

Anexo A. Detalle del conjunto de datos de pedidos

RStudio Source Editor

pedidos

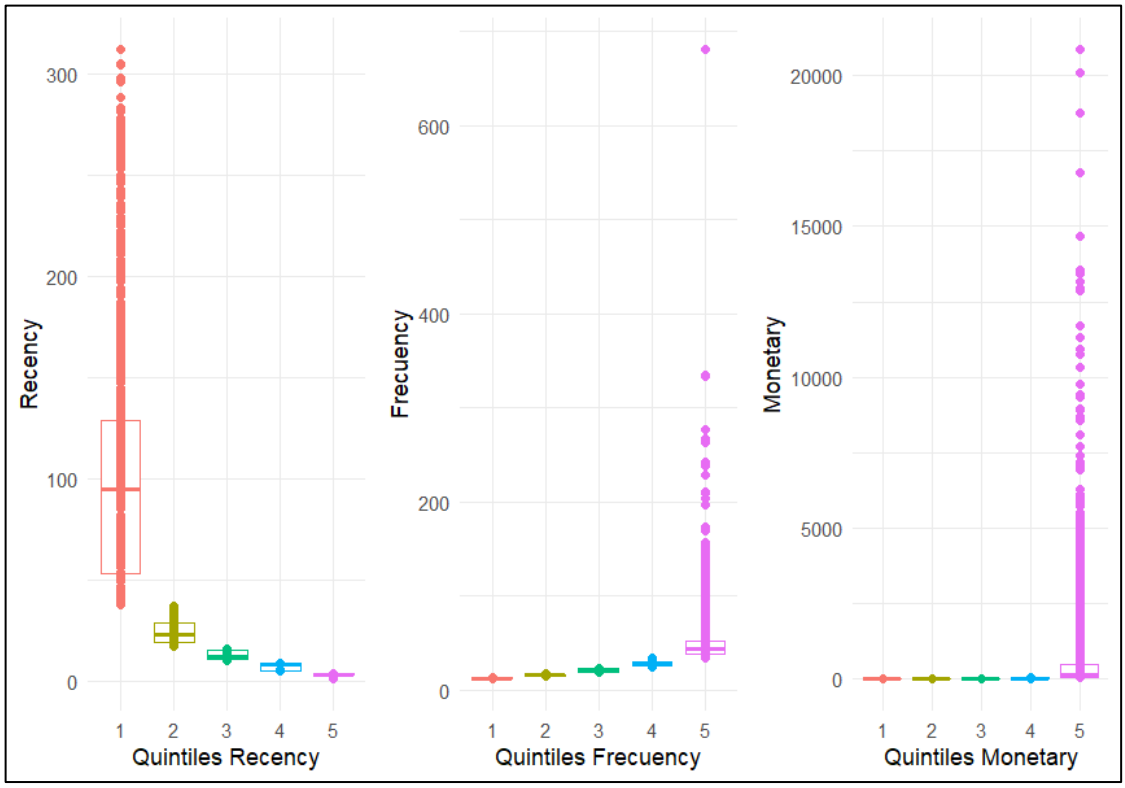
Filter

	idpedido	total	subtotal	plancomercial	promociones	descuentomanual	codigocliente	sucursal	date	estado	provincia	ciudad	latitud	longitud
1	10030113198	39.74	36.88	0.00	-1.40	0	I3095	QUITO	02/01/2019	1	PICHINCHA	ALANGASI	-0.141937910000000	-78.502674589999998
2	10030113229	66.91	59.74	0.00	0.00	0	3069338	QUITO	02/01/2019	1	PICHINCHA	COTOCOLLAO	-0.115387980000000	-78.500886520000000
3	10030113233	5.39	4.81	0.00	0.00	0	I299989	QUITO	02/01/2019	1	PICHINCHA	QUITO	-0.118921770000000	-78.499255969999993
4	10030113257	7.90	7.05	0.00	0.00	0	3068627	QUITO	03/01/2019	1	PICHINCHA	QUITO	-0.093978400000000	-78.503998899999999
5	10030113270	25.72	22.97	0.00	0.00	0	3076083	QUITO	03/01/2019	1	PICHINCHA	ALANGASI	-0.093462662771344	-78.500197073444724
6	10030113273	11.49	10.26	0.00	0.00	0	16345	QUITO	03/01/2019	1	PICHINCHA	QUITO	-0.091963349841535	-78.502491787075996
7	10030113277	11.49	10.26	0.00	0.00	0	3014833	QUITO	03/01/2019	1	PICHINCHA	QUITO	-0.094532947987318	-78.502808455377817
8	10030113295	28.60	25.54	0.00	0.00	0	3065135	QUITO	03/01/2019	1	AZUAY	AMALUZA	-0.103101110000000	-78.510303570000000
9	10030113298	23.01	20.54	0.00	0.00	0	I3971	QUITO	04/01/2019	1	PICHINCHA	QUITO	-0.008932600000000	-78.445770899999999
10	10030113300	6.90	6.16	0.00	0.00	0	3016508	QUITO	04/01/2019	1	PICHINCHA	QUITO	-0.009728300000000	-78.444571699999997
11	10030113302	16.99	15.17	0.00	0.00	0	I3394	QUITO	04/01/2019	1	PICHINCHA	ALANGASI	-0.007618300000000	-78.447781699999993
12	10030113307	14.61	13.05	0.00	0.00	0	15064	QUITO	04/01/2019	1	PICHINCHA	QUITO	-0.009030500000000	-78.445575300000002
13	10030113308	5.50	4.91	0.00	0.00	0	3087915	QUITO	04/01/2019	1	PICHINCHA	QUITO	-0.002203640000000	-78.453957349999996
14	10030113315	13.62	12.16	0.00	0.00	0	3050241	QUITO	04/01/2019	1	PICHINCHA	QUITO	-0.008145700000000	-78.446875800000001
15	10030113332	27.25	24.33	0.00	0.00	0	I349252	QUITO	04/01/2019	1	PICHINCHA	ALANGASI	-0.011905200000000	-78.446122299999999
16	10030113334	22.04	19.68	0.00	0.00	0	12655	QUITO	05/01/2019	1	PICHINCHA	QUITO	-0.121731330000000	-78.483934039999994
17	10030113337	32.68	29.18	0.00	0.00	0	3575	QUITO	05/01/2019	1	PICHINCHA	QUITO	-0.119528321525660	-78.481721129228400
18	10030113345	29.54	26.37	0.00	0.00	0	I343463	QUITO	05/01/2019	1	PICHINCHA	QUITO	-0.123294610530138	-78.488856358453631
19	10030113349	18.95	16.92	0.00	0.00	0	I278640	QUITO	05/01/2019	1	PICHINCHA	QUITO	-0.128216980000000	-78.487752310000005
20	10030113351	28.49	25.43	0.00	0.00	0	3039473	QUITO	05/01/2019	1	PICHINCHA	QUITO	-0.126289510000000	-78.486728029999995
21	10030113358	12.91	11.53	0.00	0.00	0	I292914	QUITO	07/01/2019	1	PICHINCHA	QUITO	-0.153199280000000	-78.482789949999997
22	10030113368	5.50	4.91	0.00	0.00	0	I277348	QUITO	07/01/2019	1	PICHINCHA	KENNEDY	-0.152403570000000	-78.481116470000003
23	10030113370	30.03	26.81	0.00	0.00	0	I291203	QUITO	07/01/2019	1	PICHINCHA	QUITO	-0.153436018154025	-78.482590885832906
24	10030113376	6.02	5.37	0.00	0.00	0	4077	QUITO	07/01/2019	1	PICHINCHA	QUITO	-0.146372043527663	-78.478819448500872
25	10030113378	5.60	5.00	0.00	0.00	0	3038226	QUITO	07/01/2019	1	PICHINCHA	QUITO	0.000000000000000	0.000000000000000
26	10030113385	16.36	14.61	0.00	0.00	0	I4060	QUITO	07/01/2019	1	PICHINCHA	QUITO	-0.144032560000000	-78.478820380000002
27	10030113387	19.77	17.65	0.00	0.00	0	15577	QUITO	08/01/2019	1	PICHINCHA	QUITO	-0.153899945421564	-78.472687575834500

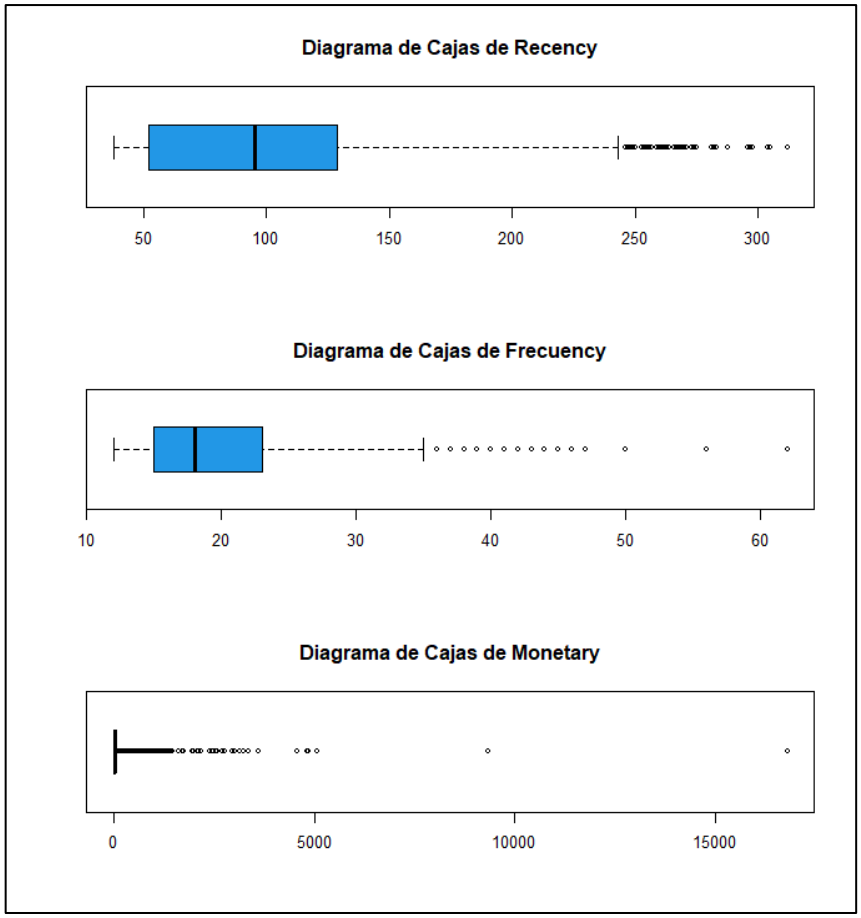
Showing 1 to 28 of 2,395,879 entries, 14 total columns

Fuente: Datos de la empresa de estudio extraído al Lenguaje de programación de R

Anexo B. Distribución de los quintiles RFM



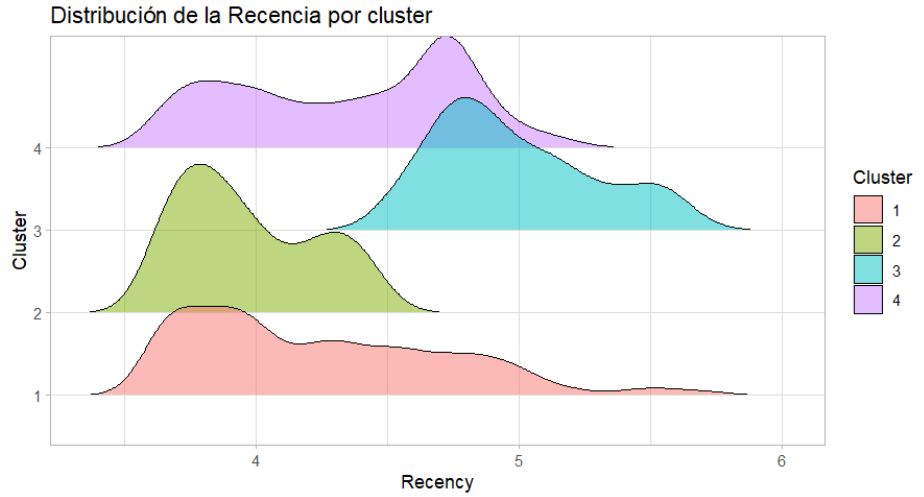
Anexo C. Descripción de los Clientes en Riesgo

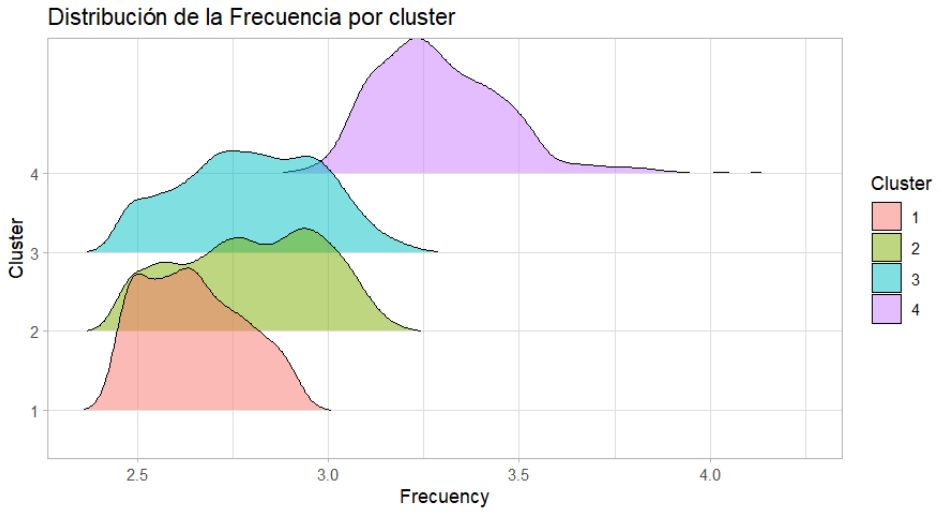


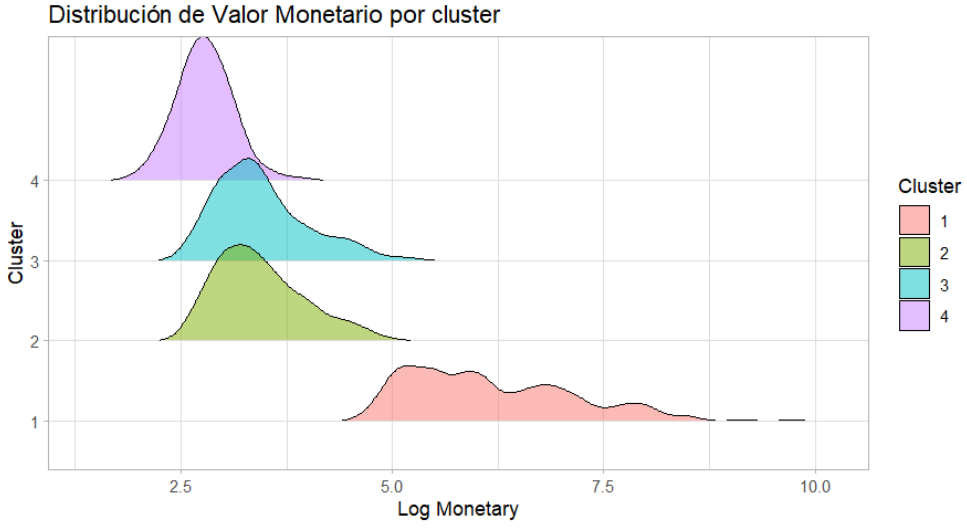
Anexo D. Estadística de Resumen de los Clúster de Clientes

```
 Descriptive statistics by group
clusters: 1
          vars   n   mean      sd median trimmed    mad    min      max    range skew kurtosis    se
recency      1 245  78.91   46.55  63.00   70.99  32.62  38.00   304.00   266.00 2.04     5.39  2.97
frecuency    2 245  14.24    1.85  14.00   14.09   1.48  12.00    18.00     6.00 0.49    -0.83  0.12
monetary     3 245 867.31 1469.22 404.92  574.76 354.30 106.77 16782.15 16675.39 6.45    58.88 93.86
------------------------------------------------------------------------------------------------
clusters: 2
          vars   n  mean    sd median trimmed   mad   min    max  range skew kurtosis   se
recency      1 596 54.59 14.49  50.00   53.19 13.34 38.00  87.00  49.00 0.71    -0.80 0.59
frecuency    2 596 16.68  3.00  16.00   16.62  4.45 12.00  24.00  12.00 0.13    -0.99 0.12
monetary     3 596 36.21 22.93  28.29   31.91 13.56 12.82 152.99 140.17 1.95     4.29 0.94
------------------------------------------------------------------------------------------------
clusters: 3
          vars   n   mean    sd median trimmed   mad   min   max  range skew kurtosis   se
recency      1 925 152.94 50.17 136.00  146.50 38.55 88.00 312.0 224.00 1.01     0.07 1.65
frecuency    2 925  16.70  2.98  16.00   16.60  2.97 12.00  25.0  13.00 0.29    -0.65 0.10
monetary     3 925  37.89 29.54  28.54   31.99 12.97 12.81 331.3 318.49 3.52    20.03 0.97
------------------------------------------------------------------------------------------------
clusters: 4
          vars   n  mean    sd median trimmed   mad   min    max  range skew kurtosis   se
recency      1 740 84.91 34.13  85.00   83.07 43.00 38.00 190.00 152.00 0.33    -0.82 1.25
frecuency    2 740 27.29  4.94  26.00   26.75  4.45 19.00  62.00  43.00 1.73     6.11 0.18
monetary     3 740 17.04  6.15  16.02   16.38  4.75  6.29  53.54  47.25 1.99     7.34 0.23
```

Anexo E. Distribución de las métricas RFM por los Clúster de Clientes







Anexo F. Segmentación de Clientes en Riesgo combinando los modelos RFM y Clúster

Cluster	Grupos RFM	Recency	Frecuency	Monetary	No. Clientes
1	115	85.4	13.0	728.6	149
	125	68.8	16.2	1,082.7	96
2	114	57.0	13.0	35.6	113
	115	52.5	13.0	78.3	52
	123	57.8	16.4	20.7	109
	124	58.0	16.5	35.6	100
	125	51.1	16.2	88.4	34
	132	48.7	19.9	15.1	58
	133	49.4	20.1	20.6	56
	134	53.1	20.6	35.6	74
3	114	165.7	13.1	34.6	166
	115	171.7	13.0	91.8	73
	123	135.2	16.2	20.4	172
	124	157.8	16.4	35.4	168
	125	159.9	16.5	92.6	70
	132	141.9	19.9	14.9	75
	133	139.3	20.2	20.6	82
	134	152.2	20.6	33.0	115
	142	212.5	25.0	13.8	2
4	143	238.0	25.0	19.2	2
	132	81.3	22.4	14.8	108
	133	89.0	22.8	20.2	78
	134	84.0	23.0	34.0	44
	141	84.2	28.3	10.7	141
	142	85.6	28.0	14.9	152
	143	88.6	28.8	20.6	176
	151	71.9	39.2	10.6	17
	152	71.5	41.5	15.1	24

Anexo G. Modelo de datos de la Segmentación de Clientes integrados al Power BI desde la herramienta R