MASTER DE ESTADÍSTICA APLICADA

CON R SOFTWARE

DESCUBRIMIENTO DE PATRONES DE CLIENTES EN LA CATEGORÍA ATUNES EN SECTOR RETAIL A TRAVES DE ANALISIS DE CONGLOMERADOS: CLUSTERING

AUTOR: NATALIA RAFFO LÓPEZ

DIRECTOR: Juan Luis López Garrancho

Fecha: 17/01/21

Entidad Colaboradora:

TABLA DE CONTENIDO

LISTADO DE TABLAS
LISTADO DE FIGURAS
LISTADO DE ANEXOS

RESUMEN

INTRODUCCIÓN

1. PLANTEAMIENTO DEL PROBLEMA Y JUSTIFICACIÓN 1,6

2. OBJETIVOS 7
   2.1 OBJETIVO GENERAL 7
   2.2 OBJETIVOS ESPECIFICOS 7

3. MARCO CONCEPTUAL 8
   3.1 ORIGEN DE LA INDUSTRIA EN COLOMBIA 8
   3.2 ENCUESTA ANUAL MANUFACTURERA 9
   3.3 INDICADORES ECONOMICOS 14
   3.4 SECTORES INDUSTRIALES 15
4. ESTADO DEL ARTE 19

5. MARCO TEÓRICO 21
   5.1 ANÁLISIS FACTORIAL MÚLTIPLE (AFM) 21

6. METODOLOGÍA 27

7. RESULTADOS 31
   7.1 ANALISIS DESCRIPTIVO 31
   7.2 RESULTADOS AFM 34

8. CONCLUSIONES 52

9. RECOMENDACIONES 54

BIBLIOGRAFÍA

ANEXOS

RESUMEN

En este trabajo se utilizaron los principales indicadores de tendencia central, posición a través de un análisis descriptivo para tener una visión general de la información, así como el uso de la técnica estadística multivariada: análisis de conglomerados a través del Análisis clustering, para descubrir patrones de comportamiento de los hábitos de consumo en clientes, con el fin de identificar los clientes oro, plata y bronce de clientes de la categoría atunes de una compañía del sector retail en Colombia en los distritos: Occidente, Bogotá, Cafetero, Costa en el periodo 2019-2020 en una marca específica, permitiendo obtener información que caracteriza los clientes y analiza el comportamiento de las principales variables demográficas, transaccionales, durante el periodo investigado y establecer estrategias focalizadas de retención. Esta técnica estadística de clustering, permite conocer el comportamiento de los clientes en la historia, para establecer estrategias según los hábitos de consumo de los clientes.

***Palabras clave**: Análisis Clustering, Regresión Logística, Sector Retail, Descubrimiento de patrones, Estrategias focalizadas.*

AGRADECIMIENTOS

En este trabajo doy especial agradecimiento a MAXIMA FORMACIÓN como entidad colaboradora, la cual tuvo la comprensión y apoyo constante en hacer realidad este trabajo, así como al directo Juan Luis López, quien estuvo en constante retroalimentación y orientación en este trabajo, también agradezco a la entidad del sector Retail, la cual puso en constante disposición al personal experto del negocio para despejar inquietudes en todos los momentos necesarios del trabajo.

INTRODUCCIÓN

El sector Retail en la economía de un país juega un importante papel. Su importancia ha radicado en disponer de productos alimenticios y de otros tipos para abastecer el país. Según diferentes estudios realizados por el DANE, los almacenes grandes e hipermercados hacen parte de las empresas, donde las ventas son realizadas en gran medida, teniendo diferencias marcadas por características principalmente referentes a: tamaños, tipo de productos a vender. Normalmente los supermercados tienen tamaños superiores a 2500 metros cuadrados y vender elementos como perfumerías, perecederos, abarrotes, limpieza bebidas, entre otros, mientras las tiendas tienen un tamaño más reducido y su venta es de menor cantidad de categorías de productos.

De acuerdo con la información presentada por el DANE, los grupos de mercancía con más alta participación en las ventas son: alimentos y bebidas no alcohólicas, muebles, electrodomésticos, productos textiles y prendas de vestir. El sector de alimentos y bebidas en Colombia han venido registrando alta expansión en los últimos años. Según estimaciones realizadas tendrá ventas anuales por más de USD 26.500 millones en 2024, mientras la demanda de la industria tenderá a crecer un 4% anual en los próximos 5 años. De acuerdo a los datos estadísticos presentados por el DANE, es la razón por la cual en este trabajo se quiere abordar este sector siendo de gran importancia para el país.

Dentro de este sector de alimentos se encuentra la categoría atunes, la cual hace referencia a atunes enlatados y de conserva en diferentes referencias, marcas y tipos, algunos son en agua, otros son en aceite, y también tienen otras características en tamaños o acompañamientos como verduras, entre otros. El atún hace parte de una importante fuente proteínica para la población mundial, es la razón por la cual motivo a estudiar esta categoría, este pez se encuentra en los principales océanos del mundo y se caracteriza por su carácter permanentemente migratorio a velocidades de crucero entre 3 y 7 km/hora, logrando llegar hasta 70 km/hora, haciendo travesías transatlánticas en menos de 60 días, reproduciéndose durante todas las épocas del año y tiene diferentes especies, siendo las más relevantes el barrilete, el aleta amarilla y el patudo, con base en el carácter migratorio, es presente en aguas internacionales como en las jurisdicciones marinas de los distintos países[1]. En Colombia se tienen las costas en los océanos: Atlántico y Pacifico, siendo predominante su pesca en el Pacifico.

Según estudios Nielsen en Colombia, al analizar el carrito de compras saludables: alimentos, sobresale la categoría liderada por los aceites, seguida de atunes, galletas, margarinas y carnes frías, siendo el canal mas fuerte para realizar las compras el canal moderno, con el 68%, el 16% de los alimentos saludables es comprado en el canal tradicional y el 16% restante en otros, en esencia mayoristas.

Este trabajo considera un gran aporte al sector retail, principalmente a la categoría atunes, en diferentes compañías de distribuidores en Colombia, dado que desde hace un par de años se ha venido evidenciando perdida de clientes casi en un 40%, quienes han abandonado la categoría y han sido identificados de forma tardía, sin tener conocimiento del porque se han ido de la categoría, también se ha identificado que estos clientes han sido rentables para la categoría y esto ha generado la necesidad de buscar técnicas estadísticas robustas que permitan analizara altos volúmenes de datos, procesar y analizar información venta marcada de consumidores finales de la categoría a través de su información transaccional , la cual permite tener un análisis de la trazabilidad de hábitos de compra y consumo de la categoría para identificar porque se han ido estos clientes rentables, así como aplicar estrategias de retención, también se ha identificado la necesidad de conocer el comportamiento de los clientes, sus preferencias y necesidades para lograr entender su patrón y así llegar con estrategias en segmentos de forma focalizada, es por esto que surge este gran interés de este trabajo, logrando hallazgos relevantes para descubrir patrones de comportamiento de los clientes y predecir la fuga de abandono de clientes de esta categoría.

Debido a la importancia de esta categoría en la evolución de la economía y las necesidades identificadas para generar herramientas solidas en tomar decisiones que permitan conocer los clientes y evitar la fuga de la categoría se realizó un modelo de descubrimiento de patrones para conocer los comportamientos de consumo e los clientes y un modelo de regresión logística para predecir la fuga de clientes de la categoría atunes en una compañía del sector retail en Colombia durante el periodo 2019 – 2020, a través de técnicas estadísticas robustas. El trabajo se enfoco en las principales variables demográficas y transaccionales de la venta marcada, las cuales permitieron tener una visión del comportamiento de compra de la categoría atunes.

En los diferentes trabajos que presentan descubrimiento de patrones así como predicción de fuga en el sector retail y esta categoría en el país mostrados en la literatura colombiana son escasos en la implementación de técnicas estadísticas que permitan llegar a este fin, los análisis son descriptivos, sin usar análisis multivariados como análisis de conglomerados y los que utilizan regresión logística en la predicción de fuga van más a asociados a otros sectores.

En este trabajo se utilizó el análisis de conglomerados a través de clustering, principalmente mediante el uso de los algoritmos: clara, k-means para formar grupos de clientes, los cuales fueron perfilados y caracterizados con el fin de identificar los clientes oro, plata y bronce y lograr activar estrategias en cada segmento según el comportamiento de cada uno. En el caso del modelo predictivo de fuga de clientes de la categoría atunes en una marca específica, se utilizó el modelo de regresión logística, el cual obtuvo el mejor ajuste a los datos con la mas alta precisión cercana al 85%, esto con el fin de aplicar estrategias de retención en el grupo de clientes con mas alto riesgo de abandono o fuga de la categoría.

Las variables incluidas en los análisis corresponden a información transaccional de los clientes, que realizaron durante el periodo de un año histórico compras en diferentes supermercados de los distritos Occidente, Cafetero, Costa y Bogotá, en total se analizaron (10 variables).

Se pretende mostrar mediante este trabajo de una forma eficaz y practica la importancia del uso de estas técnicas estadísticas como herramientas para tomar decisiones, produciendo resultados de gran ayuda y complemento al avance en materia de conocimiento.

Para realizar este estudio se revisó la información disponible en la red acerca del sector retail, sector de alimentos, categoría atunes y técnicas estadísticas de análisis de conglomerados como clustering, así como análisis de regresión logística, en diferentes artículos, consulta de las paginas del Departamento Nacional de Estadística (DANE).

En la primera sección se presenta el planteamiento del problema, seguido de los objetivos general y específicos, marco conceptual , marco teórico, metodología y resultados descriptivos y de descubrimiento de patrones, conclusiones, lo cual permitirá tener claridad del objeto de interés en este trabajo del Master.

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

1. PLANTEAMIENTO DEL PROBLEMA

La economía colombiana presenta cambios importantes en el consumo de la categoría atunes, los estudios realizados hasta el momento se han hecho mediante estadística descriptiva haciendo uso de análisis univariados a través de indicadores de tendencia central: promedio, mediana, mínimo, máximo, moda, indicadores de dispersión como: coeficiente de variación, desviación estándar, varianza e indicadores de posición: cuartiles, percentiles, análisis de tendencias, variaciones entre los más usados, análisis bivariados: análisis de correlaciones, razón por la cual resulta importante realizar análisis a través del uso de técnicas estadísticas robustas que muestran en forma precisa los cambios y entendimiento en el patrón de comportamiento de compra de clientes de la categoría atunes.

Mediante el uso de las técnicas como clustering, las cuales permitieron formar grupos de clientes e identificar el comportamiento de los clientes oro, plata y bronce, así como aplicar estrategias de retención en cada segmento según el patrón de comportamiento en la categoría atunes en una marca especifica en los distritos Occidente, Cafetero, Bogotá y Occidente en Colombia en un periodo de 2019 a 2020.

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

2. OBJETIVOS

2.1 OBJETIVO GENERAL

Descubrir patrones de comportamiento de los clientes a través del análisis de conglomerados: clustering para conocer y entender el patrón de comportamiento de los clientes a través de formar grupos de clientes e identificar los clientes oro, plata y bronce y aplicar estrategias sobre segmentos de forma focalizada en la categoría atunes en una marca especifica en una compañía del sector retail en los distritos Occidente, Bogotá, Costa y Cafetero en Colombia en el periodo 2019 – 2020, mediante las principales variables transaccionales que reflejan el comportamiento de compra de los clientes.

2.2 OBJETIVOS ESPECÍFICOS

- Analizar descriptivamente los datos, para tener una visión general de los datos.
- Descubrir patrones de clientes de la categoría atunes a través de un análisis de conglomerados: clustering.
- Identificar los clientes oro, plata y bronce de la categoría atunes.
- Realizar un perfilamiento y caracterización de los grupos resultantes del análisis de conglomerados: clustering.
- Establecer estrategias sobre los clientes oro, plata y bronce de la categoría atunes.

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

3. MARCO CONCEPTUAL

3.1 Categoría Atunes en Colombia

Las principales etapas de la cadena de atún en Colombia están dadas como se presenta a continuación:

El proceso del atun en Colombia

Se produce principalmente atún en lomos precocidos y congelados, atún en latas de conserva y harina de pescado, siendo la materia prima fundamental el atún capturado por barcos nacionales y extranjeros. El proceso de eviscerado y conversión en lomos para ser enlatados, tiene un uso para atún enlatado de aproximadamente el 50% del peso y el 50% restante se usa para harina de pescado, llevada p a la industria de alimentos balanceados para consumo animal. El atún como materia prima del producto enlatado tiene la mayor participación en el costo (entre 50% y 70%) según el tipo de producto. El procesamiento consiste en cortar el pescado, eliminarle la piel, espinas y vísceras, realizado principalmente por mujeres cabeza de familia, este atún es precocido, enfriado y enlatado (en agua, aceite u otros componentes vegetales) o empacado en bolsas al vacío y congelado: lomos destinados a exportación, los demás insumos utilizados como materia prima en la industria de atún enlatado son los aceites utilizados que pueden ser de oliva o de soya (entre 8% y el 23% del costo), los envases de hojalata con sus tapas (12% promedio de costo), cajas de cartón corrugado y etiquetas de papel (2% promedio del costo).

Las principales empresas que producen atún enlatado y atún precocido en lomos empacados al vacío y congelados son Atunec S.A., Gralco S.A. y Seatech International.

Producción y consumo de la categoría atunes

La fuente principal de información sobre el procesamiento del atún es el DANE que tiene estadísticas correspondientes a la Encuesta Anual Manufactura (EAM), asi como estudios Nielsen que muestran la participación de la categoría respecto

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

a las demás. Por razones de reserva estadística, la información de la encuesta del DANE se procesa para el sector de "transformación y conservación de pescado y de derivados del pescado" 27/, a través del modelo de equilibrio general de Fedesarrollo que incluye las estadísticas de la Encuesta Anual Manufacturera, la participación del sector de atún en el total de la producción del sector de procesamiento de pescado corresponde a un nivel del 80.82%. La producción bruta, el consumo intermedio y el valor agregado del sector de transformación de pescado muestran a precios constantes una tendencia creciente en los 5 años más recientes.

Para analizar el comportamiento de los clientes de la categoría atunes y predecir la fuga de la categoría se recurrió a la información de las principales variables transaccionales suministradas por una compañía distribuidora de atún en Cali – Colombia en el periodo 2019- 2020, mediante las siguientes variables incluidas en los análisis:

3.2 DISEÑO DE VARIABLES

VARIABLES DEMOGRAFICAS

ID_CLIENTE

Hace referencia a un código identificante de los clientes tenidos en cuenta en este análisis, es una variable cuantitativa con un numero que representa al cliente, a través de un periodo histórico de un año, se tuvieron en cuenta los clientes de la categoría atunes en una marca específica, aquellos que compraron al menos una vez en el periodo de análisis, los cuales realizaron compras.

DISTRITO

Corresponde al distrito Occidente, Cafetero, Bogotá y Costa, estos distritos agrupan todos los departamentos de Colombia, donde fue comprado el producto de atunes, .

ANTIGÜEDAD DEL CLIENTE

Variable cuantitativa, muestra la antigüedad del cliente en comprar atunes, en meses, entre más cercano a 12 este significa más antiguo entre más cercano a 1 esta, significa menos antiguo.

EDAD DEL CLIENTE

Variable cuantitativa que representa la edad de la persona que compra atunes, en años.

DIRECCIÓN DEL CLIENTE

Variable de localización que representa la la dirección de residencia del cliente que compra atunes.

---

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

CONTACTO DEL CLIENTE

Variable cuantitativa que representa el número de celular donde se puede contactar al cliente comprador final de atunes.

CORREO ELECTRONICO DEL CLIENTE

Variable que representa el correo electrónico del cliente que compra atunes.

PROFESION DEL CLIENTE

Variable cualitativa que presenta la profesión del cliente que compra atunes.

FECHA DE COMPRA

Variable fecha: día, mes, año y ahora que presenta las veces que el cliente ha ido a comprar atunes.

PUNTO DE VENTA

Variable cualitativa que presenta donde fue comprado el producto de atunes.

VARIABLES TRANSACCIONALES

VALOR TOTAL

Variable cuantitativa que presenta el valor total comprado en el periodo histórico de la categoría atunes.

DESEMBOLSO PROMEDIO

Variable cuantitativa que presenta el valor promedio comprado en el periodo histórico de la categoría atunes.

CANTIDAD DE PRODUCTOS

Variable cuantitativa que presenta la cantidad de latas de atún comprado en el periodo histórico de la categoría atunes.

FRECUENCIA DE COMPRA

Variable cuantitativa que presenta la cantidad de veces que el cliente ha comprado atún en el periodo histórico de la categoría atunes, tantas veces como el cliente ha ido a comprar atunes.

RECENCIA DE COMPRA

Variable cuantitativa que presenta el tiempo transcurrido desde la última compra de atunes en la categoría.

---

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

VARIABLES DE PRODUCTOS

EAN

Variable cuantitativa que presenta un numero de referencia del producto que ha comprado el cliente.

SEGMENTO DE PRODUCTO

Variable cualitativa que representa si el producto es en agua o aceite.

ACOMPAÑAMIENTO

Variable cualitativa que representa si el producto es acompañado de verduras u otro componente.

VOLUMEN

Variable cuantitativa que presenta el volumen del contenido de la lata de producto.

PROMOCION

Variable binaria que presenta si el producto es de promoción o no.

CANTIDAD

Variable cuantitativa que presenta la cantidad de latas de atún compradas.

3.4 INDICADORES GENERADOS

TotalVentas: Suma de las ventas en cada grupo resultante de los clusters.

Clientes: Suma de clientes en cada grupo resultante de los clusters.

Total Visitas: suma de visitas de todos los clientes en cada grupo resultante de los clusters.

Frecuencia (días): TotalVisitas/Conteo clientes

Transacción promedio : TotalVentas/TotalVisitas

ComprasPromedio: Transaccion promedio *Frecuencia

Recencia: mediana recencia de los clientes en el grupo cluster resultante.

4. MARCO TEÓRICO

5.1 ANALISIS CONGLOMERADOS: CLUSTERING

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

El análisis clúster es una técnica multivariante donde el objetivo principal es clasificar objetos formando grupos/conglomerados llamados clusters que sean lo más homogéneos posible dentro de sí mismos y heterogéneos entre sí.
Este análisis es resultado de las necesidades identificadas de diseñar e implementar una estrategia que permita formar grupos de objetos homogéneos, la forma de realizar las agrupaciones se basa en la medida de distancia o similitud entre las observaciones y la obtención de dichos clusters depende del criterio o medida de distancia considerada.

El análisis clúster es forma de lograr obtener una clasificación bajo un argumento estadístico lógico tomando las variables necesarias para obtener grupos de forma más precisa que permitan evidenciar comportamientos diferentes y generar caracterizaciones para tomar decisiones ,asertivas siendo un apoyo y aporte fundamental en empresas de diferentes sectores para clasificar grupos de consumidores respecto a sus preferencias en nuevos productos o clasificar las entidades bancarias donde sería más rentable invertir, clasificar las estrellas del cosmos en función de su luminosidad Entre muchas otras aplicaciones que se han venido usando en los últimos años.

Existen diferentes algoritmos basados en técnicas estadísticas que permiten realizar la clasificación, a continuación, se mencionan los utilizados en este trabajo:

- ✓ K-MEANS
- ✓ KMEDOIDS:PAM
- ✓ CLARA

Al considerar una muestra X formada por n individuos sobre los que se miden p variables, *X1,…,Xp* (p variables numéricas observadas en n objetos). Sea xij el valor de la variable Xj en el i -ésimo objeto i = 1,…,n; j = 1,…,p.
Este conjunto X de valores numéricos se pueden ordenar en una matriz como se presenta a continuación:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

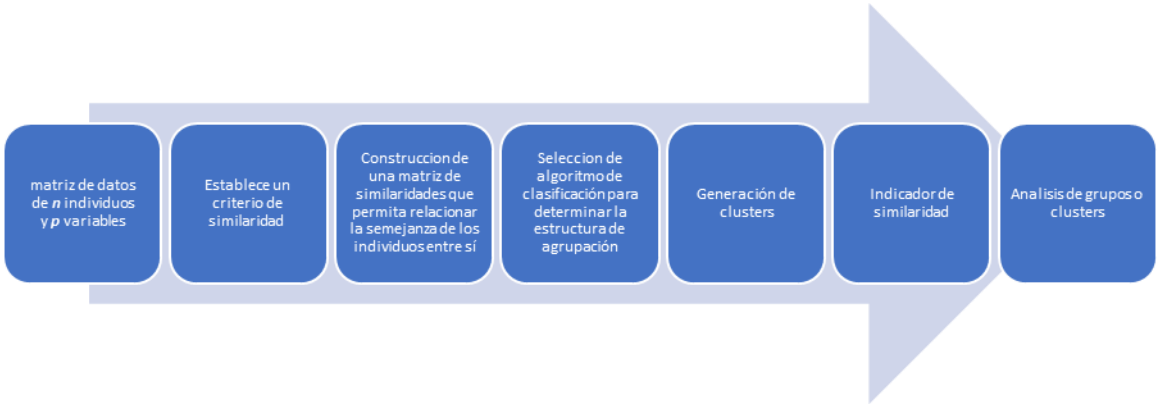
De acuerdo a la matriz anterior de datos se tiene; La i-ésima fila de la matriz X contiene los valores de cada variable para el i-ésimo individuo, mientras que la j-ésima columna muestra los valores pertenecientes a la j-ésima variable a lo largo de todos los individuos de la muestra.

Se trata de resolver el siguiente problema: Con un conjunto de n individuos caracterizados por la información de p variables Xj, (j = 1,2,…, p), se plantea clasificarlos de tal forma que los individuos pertenecientes a un grupo (clúster) (y siempre con respecto a la información disponible de las variables) sean lo más similares posibles entre sí y los distintos grupos sean entre ellos tan disimilares como sea posible.

El proceso general se presenta a continuación:

---

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en:
https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y



Despues de seleccionadas las variables a clasificar, es necesario seleccionar una medida de proximidad o de distancia entre ellos que permita medir el grado de similaridad entre cada par de objetos.

- Las medidas de proximidad, similitud o semejanza miden el grado de semejanza entre dos objetos de forma que, cuanto mayor (menor) es su valor, mayor (menor) es el grado de similaridad existente entre ellos y mayor (menor) la probabilidad de que los métodos los asignen en el mismo grupo.

- Las medidas de disimilitud, desemejanza o distancia miden la distancia entre dos objetos de forma que, cuanto mayor (menor) sea su valor, más (menos) diferentes son los objetos y menor (mayor) la probabilidad de que los métodos de clasificación los asignen en el mismo grupo.

Métodos de clasificación

Se tienen dos métodos de clusters: Métodos jerárquicos y Métodos no-jerárquicos, a continuación se presenta cada uno:

- **Métodos Jerárquicos:** En cada paso del algoritmo sólo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo. La clasificación resultante tiene un número creciente de clases anidadas.
- **Métodos No jerárquico o Repartició**n: Comienzan con una solución inicial, un número de grupos g fijado de antemano y agrupa los objetos para obtener los g grupos.

Los métodos jerárquicos se subdividen a su vez en aglomerativos y **divisivos**:

- Los métodos jerárquicos aglomerativos comienzan con tantos clusters como objetos tengamos que clasificar y en cada paso se recalculan las distancias entre los grupos existentes y se unen los dos grupos más similares o menos disimilares. El algoritmo acaba con un clúster conteniendo todos los elementos.
- Los métodos jerárquicos divisivos comienzan con un clúster que engloba a todos los elementos y en cada paso se divide el grupo más heterogéneo. El algoritmo acaba con tantos clusters (de un elemento cada uno) como objetos se hayan clasificado.

Existen diferentes criterios para ir formando los clusters; todos estos criterios se basan en una matriz de distancias o similitudes. A continuación se presentan algunos ejemplos de los métodos:

Jerárquicos aglomerativos:

- Método del Linkage Simple, Enlace Simple o Vecino más próximo

---

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

- Método del Linkage Completo, Enlace Completo o Vecino más alejado
- Método del Promedio entre grupos
- Método del Centroide
- Método del la Mediana
- Método de Ward

Jerárquicos divisivos

- Método del Linkage Simple
- Método del Linkage Completo
- Método del Promedio entre grupos
- Método del Centroide
- Método del la Mediana
- Análisis de Asociación

A continuación se presentan algunos algoritmos utilizados para el análisis de conglomerados clustering utilizados en este análisis:

ALGORITMO K-MEANS

Este algoritmo hace posible procesar un numero ilimitado de casos permitiendo utilizar un método de aglomeración incluyendo previamente el numero de clusters requeridos, este método comienza agrupando los k casos mas distantes entre si , calcula centroides, asigna cada caso al centro mas próximo y actualiza el valor de los centros a medida que va incorporando nuevos casos, al tener todos los casos asignados a uno de los k conglomerados inicia un proceso iterativo para realizar el calculo de los centroides finales de esos k conglomerados.
Este es muy útil al tener un numero elevado de casos.

Dado unas observaciones x1, x2, x3, …., xn , en donde cada observación es un vector de d dimensiones, este método consiste en una partición de las observaciones en k conjuntos, con el fin de minimizar la suma de los cuadrados dentro de cada grupo, donde cada grupo se representa por S ={S1, S2, …Sn}, como se presenta a continuación:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2$$

De acuerdo a lo anterior se tiene: µi es la media de puntos en Si.

A continuación se presentan los principales pasos de este algoritmo:
En forma inicial:

1. Se escogen k centroides de forma aleatoria.
2. Se forman k grupos donde se realiza la asignación de cada punto al centroide más cercano.

En el proceso de forma iterativa:

3. Se calculan las distancias de todos los puntos a los k centroides.
4. Se forman k grupos asignando cada punto al centroide mas cerca.
5. Se calculan de nuevo los nuevos centroides.

Para recalcular los centroides se utiliza la siguiente formula:

---

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d^2(m_i, x)$$

Si las distancias utilizadas son principalmente la euclídea, entonces SSE se minimiza de tal manera que se usa la media por cada atributo, mientras si por el contrario se usa la distancia manhattan SSE se minimiza usando la mediana.

ALGORITMO PAM (Partitioning around medoids)

De forma similar al algoritmo K-medias o k-means, el algoritmo de partitioning around medoids (PAM) genera una partición inicial y un número preespecificado de grupos, este algoritmo busca los K "individuos representativos" (o medoids) entre el conjunto de datos (mientras que K-medias se enfoca en los promedios del grupo), que minimizan la suma de las disimilaridades al resto de los registros. Sin embargo PAM es mas robusto que K-medias y requiere como la matriz de disimilaridades entre observaciones y no los datos originales como entrada. También se considera mas intensivo computacionalmente, principalmente ya que debe realizar la búsqueda de medoids (Izenman, 2008).

El resultado de usar estas técnicas de particionamiento dependen de la elección del número de clusters o grupos así como de realizar un desarrollo inicial.

A continuación se presentan los pasos principales que realiza:

1. **Inicialización:** Selección al azar de k de los n puntos de datos como los candidatos a medoides.
2. **fase de construcción,** donde asigna cada observación al cluster con el medoide mas próximo, dependiendo de la distancia elegida (euclidiana, Manhattan o Minkowski). Luego se encuentra un mínimo local para la función objetivo, es decir, una solución de tal manera el cambio de observación con un medoide haga que la función objetivo decrezca (esto se denomina la fase de intercambio).
3. Se repiten los pasos anteriores hasta que los medoides queden estables (es decir que haya cambios en los medoides).

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

ALGORITMO CLARA (Clustering Large Aplications)

Este algoritmo es una versión del algoritmo PAM, diseñado para mayor cantidad de datos, su funcionamiento esta enfocado en seleccionar subgrupos aleatorios dentro del grupo completo de datos y aplicar sobre estos el algoritmo PAM, los grupos de menor error cuadrático son seleccionados como clusters, es muy eficiente en la medida que se incrementa el número de datos.

Este algoritmo es bastante en robusto en altos volúmenes de datos, considera subconjuntos de datos de tamaño fijo (sampsize) de modo que los requisitos de tiempo y almacenamiento se vuelvan lineales en n en lugar de cuadráticos.

Cada subconjunto de datos se divide en k grupos utilizando el mismo algoritmo que en pam. Una vez k que se han seleccionado objetos representativos del subconjunto de datos, cada observación de todo el conjunto de datos se asigna al medoide más cercano. La media (equivalente a la suma) de las diferencias de las observaciones con su medoide más cercano se usa como una medida de la calidad de la agrupación. Se retiene el subconjunto de datos para el cual la media (o suma) es mínima. Un análisis adicional se lleva a cabo en la partición final.

Cada subconjunto de datos se ve obligado a contener los medoides obtenidos del mejor subconjunto de datos hasta entonces. Las observaciones dibujadas al azar se agregan a este conjunto hasta que sampsizese haya alcanzado.

Pasos principales del algoritmo CLARA:

1. Para i = 1 a n, sigue repetidamente estos pasos:

2. Elige la muestra con k objetos aleatoriamente del conjunto total de datos, llama al algoritmo PAM para encontrar c medoids en la muestra.

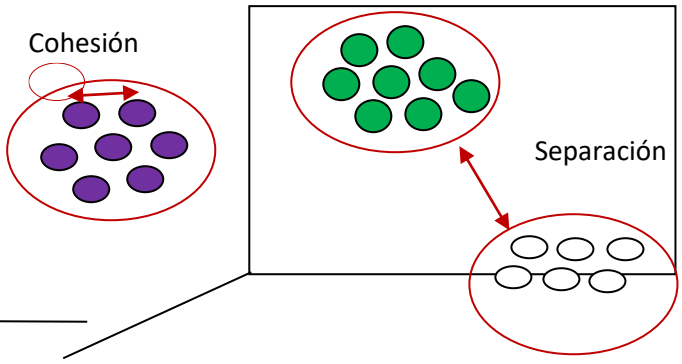
3. Para cada uno de los objetos oj del conjunto total de datos, identifica cuál de los c medoids es el más similar a oj.

4. Realiza el calculo de la disimilaridad promedio del agrupamiento obtenido en el paso anterior, en el caso de ser menor el valor al mínimo actual, se utiliza este valor como el mínimo actual y retener los c medoids encontrados en el paso (2) como el mejor conjunto de medoids generados.

5. Se debe regresar nuevamente al paso (1) para realizar la próxima iteración.

Evaluación de Resultados de los Clusters para ambos algoritmos: K-means y Clara

- ✓ Se debe validar si con estos clusters o grupos generados efectivamente se minimizo la distancia intra-cluster, a través de un indicador de cohesión.
- ✓ También es necesario revisar si con estos clusters o grupos generados efectivamente se logró maximizar la distancia inter-cluster a través de indicadores que permiten medir la separación de los grupos garantizando que sean totalmente heterogéneos.



[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

Por lo tanto una buena medida en el caso de usar la medida de distancia euclídea es a través de la siguiente fórmula:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d^2(m_i, x)$$

Otra medida importante para validar resultados es la cohesión de silueta.

5. METODOLOGÍA

RECOLECCIÓN DE LA INFORMACIÓN

Para la construcción de la base de datos se tuvo acceso a las bases de datos de un distribuidor en la categoría atunes referente a la información demográfica, transaccional del año 2019 hasta 2020 en Colombia en un supermercado en Cali Colombia.

Se utilizó la metodología Crisp Data Mining para el abordaje del trabajo, siguiendo las siguientes etapas como se presenta a continuación:

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

A continuación se describe cada una de las etapas seguidas:

Recolección de datos

En esta etapa se llevo a cabo la recolección de información necesaria para el trabajo, se tomaron datos demográficos, transaccionales suministrados por la compañía distribuidora de atún, correspondiente a un año de información histórica en el periodo 2019 a 2020, en la categoría atunes de un supermercado en la ciudad de Cali- Colombia.

En total se tuvo una base de datos inicial con un total de 1.453.113 registros, con la información referente a los diferentes distritos donde se tiene presencia de venta de la categoría atunes: total costa, Bogotá, Cafetero y Occidente, referente a una marca especifica. En total se analizaron 513.313 clientes.

Entendimiento de Negocio

Se realizaron sesiones con el negocio para tener un contexto de negocio, conocer las necesidades actuales, dolores o problemáticas actuales, metas, oportunidades de mejora y sueños a través de dinámicas de visión de negocio, las cuales permitieron conocer el contexto de negocio.

También identificaron y levantaron las principales fuentes de datos actuales, como una revisión de los diccionarios de variables para entender el contenido de las fuentes de datos.

Las fuentes de datos principales identificadas se presentan a continuación:

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

Preparación Datos

Se realiza la integración de fuentes para generar una fuente consolidada y lograr tener una base de datos única de la información, se realiza la construcción del data set para el análisis de conglomerados: clustering, generando una tabla resumen con la siguiente estructura, en la cual el cliente aparece de forma única, con el resumen de las principales variables, como se presenta a continuación:

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

Distrito	Departamento	Ciudad	Punto de venta	Segmento	EAN	Cod_cliente	Ventas	Cantidad
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 407 ROSALES	Lomo Aceite	7702367000015	106	40000	1080
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 404 SANTA A	Lomo Aceite	7702367000015	122	13350	184
Distrito Bogota	CUNDINAMARCA	BOGOTA	SAO 405 PLAZA D	Lomo Agua	7702367000022	144	4340	184
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 402 CALLE 10	Lomo Aceite	7702367000015	172	8300	368
Distrito Bogota	META	VILLAVIC	STO 434 RECREO	Lomo Aceite	7702367000015	175	8100	184
Distrito Occider	CAUCA	POPAYAN	SUPERTIENDA PC	Lomo Agua	7702367000022	176	4340	184
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 406 SANTA I	Lomo Agua	7702367000985	183	52950	1368
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 418 COLINA	Lomo Aceite	7702367000404	189	59470	704
Distrito Bogota	CUNDINAMARCA	BOGOTA	DROGUERIA AVE	Lomo Agua	7702367000985	208	7550	240
Distrito Bogota	CUNDINAMARCA	BOGOTA	HACIENDA SANT	Lomo Aceite	7702367000398	209	17500	480
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 412 CENTRO	Lomo Aceite	7702367001326	245	31050	800
Distrito total co	ANTIOQUIA	RIONEGR	RIONEGRO	Lomo Aceite	7702367000404	304	9400	240
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 418 COLINA	Lomo Aceite	7702367000015	348	9200	184
Distrito Bogota	CUNDINAMARCA	BOGOTA	HACIENDA SANT	Lomo Aceite	7702367000398	396	35800	640
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 402 CALLE 10	Lomo Aceite	7702367000015	400	13350	184
Distrito Occider	VALLE	CALI	SUPERTIENDA PA	Lomo Aceite	7702367000404	414	17900	240
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 428 MAZURE	Lomo Aceite	7702367000015	443	4050	184

Después de tener el Data set inicial se construyen las siguientes variables nuevas con base en las existentes, las cuales generaron mayor información para conocer el patrón de comportamiento de los clientes, las cuales se presentan a continuación:

- ✓ **Frecuencia:** Corresponde al numero de veces que el cliente a comprado el producto atunes, se calcula como el numero de visitas realizadas por el cliente, esta variable es de vital importancia, ya que permite identificar que tan frecuente es el cliente en su historia.
- ✓ **Recencia:** El tiempo transcurrido desde la última compra del cliente, permite identificar que tan reciente es el cliente en sus compras, se encuentra en días, por lo tanto un cliente que no compra hace 6 meses podría pensarse que es un cliente ocasional, mientras un cliente el cual su última compra fue hace una semana es un cliente muy reciente en sus compras.

Distrito	Departamento	Ciudad	Punto de venta	Segmento	Perfiles	EAN	Cod_cliente	Recencia	Frecuencia	Ventas	Cantidad
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 407 ROSALES	Lomo Aceite	Cliente Espejo	7702367000015	106	152	6	40000	1080
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 404 SANTA A	Lomo Aceite	Cross Sell	7702367000015	122	20	1	13350	184
Distrito Bogota	CUNDINAMARCA	BOGOTA	SAO 405 PLAZA D	Lomo Agua	Cliente Ocasional	7702367000022	144	103	1	4340	184
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 402 CALLE 10	Lomo Aceite	Up Sell	7702367000015	172	155	2	8300	368
Distrito Bogota	META	VILLAVIC	STO 434 RECREO	Lomo Aceite	Cliente Ocasional	7702367000015	175	303	1	8100	184
Distrito Occider	CAUCA	POPAYAN	SUPERTIENDA PC	Lomo Agua	Cliente Ocasional	7702367000022	176	51	1	4340	184
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 406 SANTA I	Lomo Agua	Cliente Espejo	7702367000985	183	48	6	52950	1368
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 418 COLINA	Lomo Aceite	Cliente Espejo	7702367000404	189	32	3	59470	704
Distrito Bogota	CUNDINAMARCA	BOGOTA	DROGUERIA AVE	Lomo Agua	Cliente Ocasional	7702367000985	208	332	1	7550	240
Distrito Bogota	CUNDINAMARCA	BOGOTA	HACIENDA SANT	Lomo Aceite	Cliente Espejo	7702367000398	209	235	2	17500	480
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 412 CENTRO	Lomo Aceite	Cliente Espejo	7702367001326	245	47	4	31050	800
Distrito total co	ANTIOQUIA	RIONEGR	RIONEGRO	Lomo Aceite	Cliente Ocasional	7702367000404	304	214	1	9400	240
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 418 COLINA	Lomo Aceite	Cliente Ocasional	7702367000015	348	104	1	9200	184
Distrito Bogota	CUNDINAMARCA	BOGOTA	HACIENDA SANT	Lomo Aceite	Cliente Espejo	7702367000398	396	50	2	35800	640
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 402 CALLE 10	Lomo Aceite	Cross Sell	7702367000015	400	40	1	13350	184
Distrito Occider	VALLE	CALI	SUPERTIENDA PA	Lomo Aceite	Cross Sell	7702367000404	414	351	1	17900	240
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 428 MAZURE	Lomo Aceite	Cliente Ocasional	7702367000015	443	373	1	4050	184
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 420 CENTRO	Lomo Aceite	Cliente Ocasional	7702367000015	467	125	1	4250	184
Distrito Bogota	CUNDINAMARCA	BOGOTA	HACIENDA SANT	Lomo Agua	Cliente Espejo	7702367000022	468	139	2	25000	368
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 402 CALLE 10	Lomo Aceite	Cliente Espejo	7702367002712	469	60	3	22000	800
Distrito Bogota	CUNDINAMARCA	BOGOTA	STO 409 AVENIDA	Lomo Agua	Cliente Ocasional	7702367000022	494	142	1	4150	184

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

Modelación

Se construyeron los modelos de descubrimiento de patrones.

Entrenamiento y validación

Se validaron los modelos a través de indicadores estadísticos para identificar el ajuste de los modelos a los datos analizados.

Análisis Resultados

Se analizaron los resultados de los modelos.

Implementación

Se realizaron los ajustes pertinentes a los modelos.

6. RESULTADOS

6.1 ANÁLISIS DESCRIPTIVO RESUMEN PRINCIPALES VARIABLES

Inicialmente se revisa la base de datos con el primer acercamiento para conocer las variables cargadas, el tipo de variables, cantidad de registros y variables, opciones de respuesta top 5, a través de la siguiente función:

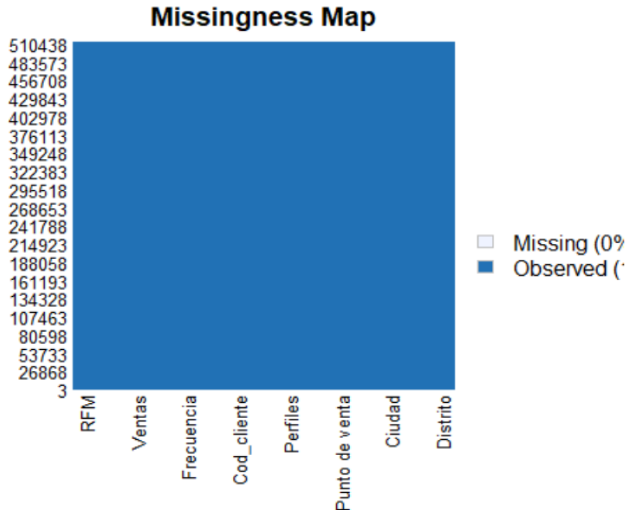
```
> str(datos)
tibble [513,304 x 14] (S3: tbl_df/tbl/data.frame)
 $ Distrito      : chr [1:513304] "Distrito Bogota" "Distrito Bogota" "Distrito
 Bogota" "Distrito Bogota" ...
 $ Departamento  : chr [1:513304] "CUNDINAMARCA" "CUNDINAMARCA" "CUNDINAMARCA"
 "CUNDINAMARCA" ...
 $ Ciudad        : chr [1:513304] "BOGOTA" "BOGOTA" "BOGOTA" "BOGOTA" ...
 $ Edad          : num [1:513304] 34 24 33 26 27 33 26 26 32 24 ...
 $ Punto de venta: chr [1:513304] "STO 407 ROSALES" "STO 404 SANTA ANA" "SAO 40
5 PLAZA DE LAS AMERICAS" "STO 402 CALLE 100" ...
 $ Segmento      : chr [1:513304] "Lomo Aceite" "Lomo Aceite" "Lomo Agua" "Lomo
 Aceite" ...
 $ Perfiles      : chr [1:513304] "Cliente Espejo" "Cross Sell" "Cliente Ocasio
nal" "Up Sell" ...
 $ EAN           : num [1:513304] 7.7e+12 7.7e+12 7.7e+12 7.7e+12 7.7e+12 ...
 $ Cod_cliente   : num [1:513304] 106 122 144 172 175 176 183 189 208 209 ...
 $ Recencia      : num [1:513304] 152 20 103 155 303 51 48 32 332 235 ...
 $ Frecuencia    : num [1:513304] 6 1 1 2 1 1 6 3 1 2 ...

$ Ventas        : num [1:513304] 40000 13350 4340 8300 8100 ...
$ Cantidad      : num [1:513304] 1080 184 184 368 184 ...
```

ANALISIS DE CALIDAD GENERAL

A través de la función missmap(datos), se tiene una aproximación inicial de los ~~valores atípicos~~, sin embargo genera un 0%, lo cual conviene ahondar mas detallado en el tema a través del resumen descriptivo por variables, para identificar inconsistencias que no sean captadas en este acercamiento.

---

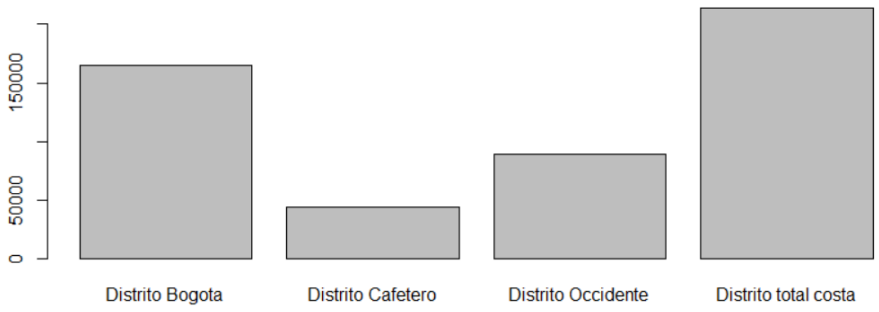
[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y



También se utilizó la función describe(), para tener un Resumen de los datos, mostrando los valores únicos de las variables e información resumida como los menores valores y los mas altos en variables cuantitativas y en las variables cualitativas, la frecuencia correspondiente a la cantidad de registros, la cual corresponde a la cantidad de clientes por cada categoría de cada variable y el porcentaje, permitiendo un acercamiento inicial de los datos con una visión global, como se presenta a continuación:

RESUMEN VARIABLE DISTRITO

```
---------------------------------------------------------------------------
Distrito
       n  missing distinct
  513304        0        4

Value          Distrito Bogota    Distrito Cafetero   Distrito Occidente
Frequency                165278                44848                89516
Proportion                0.322                0.087                0.174

Value      Distrito total costa
Frequency                213662
Proportion                0.416
```



De acuerdo a lo anterior se tiene información de 4 distritos: Costa, Cafetero, Bogotá y Occidental, la mayoría de clientes de la categoría atunes se encuentran en el distrito total costa con 213.662 registros, correspondiente al 42%, seguido del distrito Bogotá con 165.278, es decir el 32%, mientras la menor cantidad de registros se presenta en el distrito Cafetero con 44.848 clientes, es decir el 8% de clientes.

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

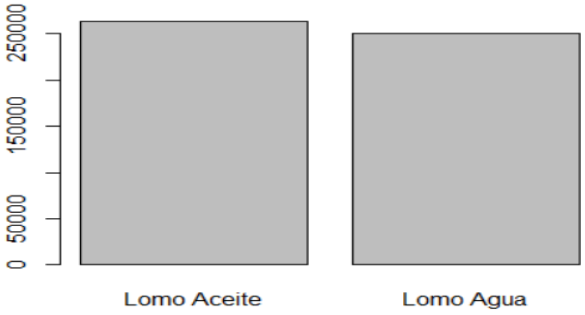
RESUMEN VARIABLE DEPARTAMENTO

```
Departamento
       n  missing distinct
  513304        0       19

lowest : ANTIOQUIA ATLANTICO BOLIVAR   BOYACA    CALDAS
highest: QUINDIO   RISARALDA SUCRE     TOLIMA    VALLE

ANTIOQUIA (17315, 0.034), ATLANTICO (68102, 0.133), BOLIVAR (54739, 0.107)
BOYACA (2885, 0.006), CALDAS (9584, 0.019), CASANARE (3205, 0.006), CAUCA
(13728, 0.027), CESAR (13391, 0.026), CORDOBA (21678, 0.042), CUNDINAMARCA
(133608, 0.260), GUAJIRA (4340, 0.008), HUILA (14765, 0.029), MAGDALENA
(21436, 0.042), META (4468, 0.009), QUINDIO (15299, 0.030), RISARALDA
(11140, 0.022), SUCRE (12661, 0.025), TOLIMA (6347, 0.012), VALLE (84613,
0.165)
```

De acuerdo a lo anterior se tiene información de 19 departamentos: la mayoría de clientes de la categoría atunes se encuentran en el departamento Valle con 84.613 registros, correspondiente al 17%, seguido de Atlántico con 68.102, es decir el 13%, mientras la menor cantidad de registros se presenta en Casanare y Boyacá con 3.205 y 2885 clientes, es decir el 0.6% de clientes respectivamente.

RESUMEN VARIABLE SEGMENTO

```
Segmento
       n  missing distinct
  513304        0        2

Value      Lomo Aceite   Lomo Agua
Frequency       263083      250221
Proportion       0.513       0.487
```

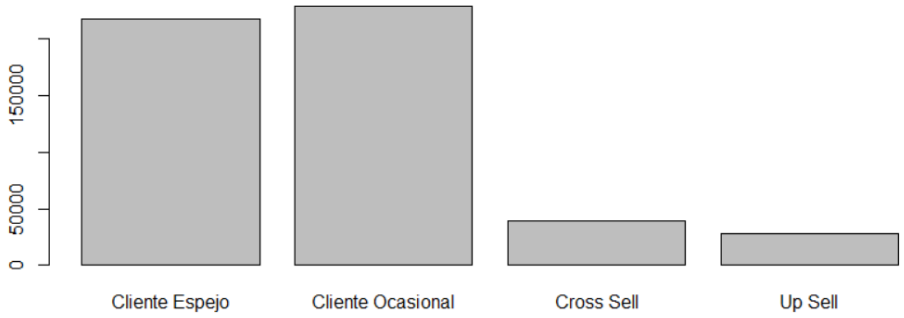


De acuerdo a lo anterior se tiene información de 2 segmentos: la mayoría de clientes se encuentran en el segmento Lomo Aceite con 263.083 registros, correspondiente al 51%, seguido del segmento Lomo Agua con 250.221 clientes, es decir el 49%.

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

RESUMEN VARIABLE PERFILES

```
Perfiles
       n  missing distinct
  513304        0        4

Value         Cliente Espejo Cliente Ocasional        Cross Sell
Frequency             217522             228640             38974
Proportion             0.424              0.445             0.076

Value                Up Sell
Frequency              28168
Proportion             0.055
```



De acuerdo a lo anterior se tiene información de 4 perfiles: la mayoría de clientes son clientes ocasionales con 228.640 registros, correspondiente al 45%, seguido de los clientes espejo con 217.522 clientes, es decir el 42%, mientras la menor cantidad de clientes son Up sell con 28.166.

RESUMEN VARIABLE COD_CLIENTE

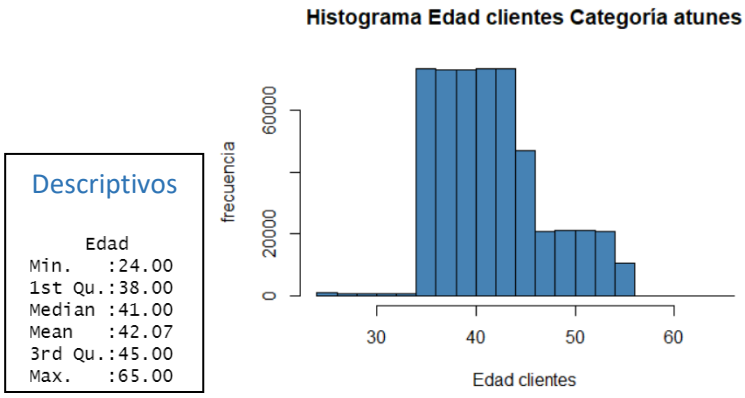
Descriptivos

```
 Cod_cliente
Min.   :    106
1st Qu.:1671573
Median :3834990
Mean   :3940854
3rd Qu.:6369222
Max.   :7792440
```

De acuerdo a lo anterior se tiene información de códigos con 3 dígitos lo cual podría ser una alerta para ver si este código es correcto o hace parte de un cliente con tipo diferente, se recomienda revisar estos códigos de 3 dígitos, ya que el resto tiene en promedio 7 dígitos.

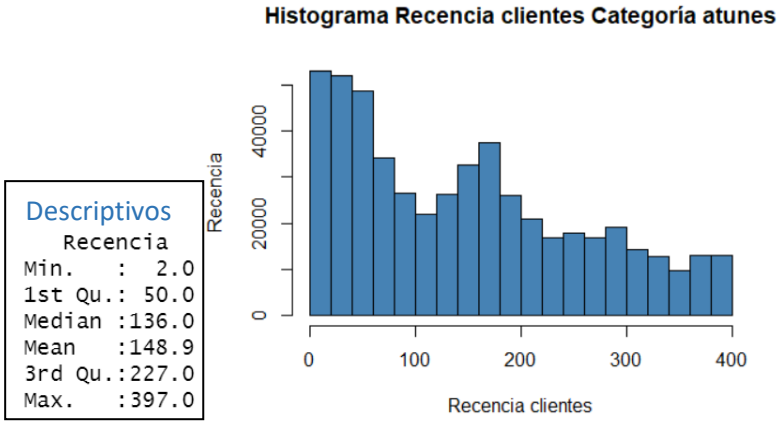
RESUMEN VARIABLE EDAD

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y



De acuerdo a lo anterior se tiene, el valor mínimo de edad es 24 años, el promedio d ellos clientes tienen una edad de 42 años cercano al valor de la mediana, la cual es 41 años, mientras el valor máximo es 65 años, el 25% de los clientes de atunes tienes menos de 38 años, mientras el 75% menos de 45 años.

RESUMEN VARIABLE RECENCIA



De acuerdo a lo anterior se tiene: se tiene un cliente, donde el tiempo transcurrido desde la ultima compra son 2 días como valor mínimo, mientras el promedio de los clientes es cercano a la mediana con un valor de 148 y 136 días transcurridos desde la ultima compra respectivamente, lo que corresponde a 5 meses, un tiempo bastante amplio en la compra mas reciente de la categoría atunes.

RESUMEN VARIABLE FRECUENCIA

Esta variable registra un valor muy atípico, por lo tanto se toma la decisión de separarlo del análisis para validar si es correcto.

```
   Frecuencia
Min.    :   1.000
1st Qu.:   1.000
Median :   1.000
Mean   :   2.828
3rd Qu.:   3.000
Max.   :2840.000
```

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

Después se vuelve a montar la base y se obtiene el siguiente gráfico:

De acuerdo a lo anterior se tiene: se tienen algunos clientes, donde la frecuencia de compra de la categoría atunes en el periodo histórico es 1 vez , mientras el promedio de los clientes compran en promedio 3 veces en el periodo, sin embargo se registra un valor máximo de 2840 veces en el periodo es decir, compran 8 veces al día, lo cual genera una alerta de un dato atípico.

RESUMEN VARIABLE VENTAS

```
     Ventas
Min.   :    3220
1st Qu.:    7370
Median :   12930
Mean   :   27282
3rd Qu.:   27330
Max.   :26882470
```

De acuerdo a lo anterior se tiene: se tienen algunos clientes, quienes registran un valor mínimo de $3.220 referente al valor de una unidad de atún, mientras el promedio compraron en el periodo histórico en total $27.282, mientras la mediana es igual a $12.930, sin embargo se evidencia un valor atípico demasiado alto respecto a los demás en el valor máximo, generando una alerta, se debe revisar para saber si es correcto y en caso contrario se sugiera separarlo del análisis. Se toma decisión de separar este valor, ya que afectaría el comportamiento de los datos.

RESUMEN VARIABLE CANTIDAD

```
    Cantidad
Min.   :   120.0
1st Qu.:   184.0
Median :   368.0
Mean   :   674.5
3rd Qu.:   736.0
Max.   :616048.0
```

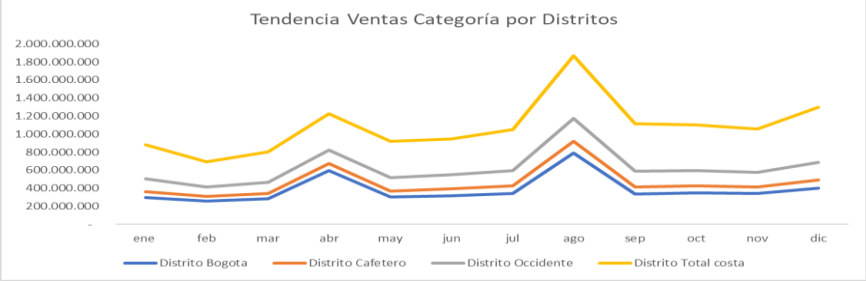
De acuerdo a lo anterior se tiene: se tienen algunos clientes, quienes registran un valor mínimo de 120 cantidades de producto en el periodo histórico, mientras el promedio compraron en el periodo histórico en total 675 cantidades, mientras la mediana es igual a 368, sin embargo se evidencia un valor atípico demasiado alto respecto a los demás en el valor máximo, generando una alerta, se debe revisar para saber si es correcto y en caso contrario se sugiera separarlo del análisis. El primer cuartil es 184, es decir que el 25% de los datos es menor a este valor, mientras el 75% de los datos es 736 , es decir superior a este valor. Se identificó un valor muy atípico respecto a los demás de 616048, por lo tanto es separado del análisis.

A continuación se presenta resumen general:

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

Diagnóstico de Variables		
Total Variables	14	
Total registros	513.304	
Variables	**Tipo**	**Cantidad**
Distrito	Cualitativa	4: Distrito Bogota, Distrito Occidente, Distrito total costa, Distrito Cafetero
Departamento	Cualitativa	19: Antioquia, Atlantico, Bolivar, Boyaca, Caldas, entre otros.
Ciudad	Cualitativa	80: Aguachica, Arjona, Armenia, Barano, Valledupar entre otras.
Edad	Cuantitativa	lowest : 24 25 26 27 28, highest: 61 62 63 64 65
Punto de venta	Cualitativa	277: 13 de junio, 6 de marzo, 7 de agosto, Altavista, Villa campestre entre otros.
Segmento	Cualitativa	2: Lomo Aceite, Lomo Agua
Perfiles	Cualitativa	4: Cliente espejo, Cross sell, Cliente ocasional, Up sell
EAN	Cuantitativa	28
Cod_cliente	Cuantitativa	513.304
Recencia	Cuantitativa	lowest : 2 3 4 5 6, highest: 393 394 395 396 397
Frecuencia	Cuantitativa	lowest : 1 2 3 4 5, highest: 235 339 468 519 2840
Ventas	Cuantitativa	lowest : 3220 3290 3300 3320 3340 ; highest: 7518000 7608970 8324640 8327230 26882470
Cantidad	Cuantitativa	lowest : 120 160 184 240 280, highest: 51540 74420 101596 118068 616048

Distribución ventas en el Categoría a nivel de Distritos



De acuerdo a lo anterior se tiene:

- ✓ El distrito total costa registra mayores ventas respecto a los demás distritos.
- ✓ La mediana de la recencia es igual a 1, lo que significa que el comportamiento normal de los clientes realizaron tan solo 1 visita en el periodo de análisis

ANALISIS BIVARIADO

VARIABLE DISTRITO POR SEGMENTO

```
                    Lomo Aceite Lomo Agua       Sum
Distrito Bogota       0.5409915 0.4590085 1.0000000
Distrito Cafetero     0.6041741 0.3958259 1.0000000
Distrito Occidente    0.6367577 0.3632423 1.0000000
Distrito total costa  0.4192276 0.5807724 1.0000000
```

De acuerdo a lo anterior se tiene, La mayoría de clientes del Distrito Bogotá comparan atún del segmento lomo aceite (54%), mientras el restante consumen atún lomo en agua (46%), este mismo comportamiento se percibe en los distritos Cafetero y Occidente, sin embargo sobresale Distrito total costa dado que tiene un comportamiento diferente, la mayoría consumen atún en lomo agua con un 58%, el restante consumen en lomo aceite.

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

VARIABLE DISTRITO POR PERFILES

	Cliente Espejo	Cliente Ocasional	Cross Sell	Up Sell	Sum
Distrito Bogota	0.43854596	0.42718329	0.09018744	0.04408330	1.00000000
Distrito Cafetero	0.38209062	0.47591866	0.06472975	0.07726097	1.00000000
Distrito Occidente	0.39417534	0.47566915	0.05543143	0.07472407	1.00000000
Distrito total costa	0.43348373	0.44047140	0.07583473	0.05021014	1.00000000

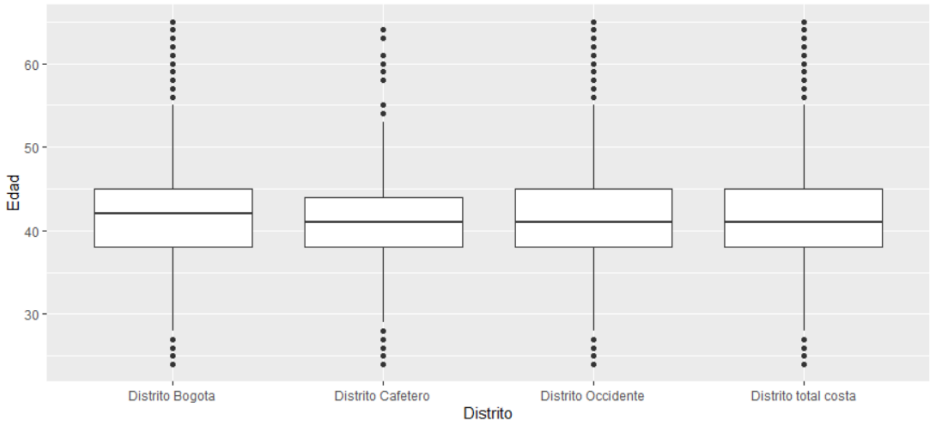
De acuerdo a lo anterior se tiene, la mayoría de clientes del Distrito Bogotá son clientes espejo (44%), lo que es una alerta para aplicar estrategias de fidelización en estos clientes, mientras en el Distrito Cafetero, la mayoría de clientes son clientes son clientes ocasionales, es decir no son frecuentes en sus compras, su desembolso promedio es baja y se demoran en comprar atunes, el mismo patrón de comportamiento ocurre en los segmentos restantes.

VARIABLE PERFILES POR SEGMENTO

	Lomo Aceite	Lomo Agua	Sum
Cliente Espejo	0.5334127	0.4665873	1.0000000
Cliente Ocasional	0.4894682	0.5105318	1.0000000
Cross Sell	0.4897111	0.5102889	1.0000000
Up Sell	0.5700085	0.4299915	1.0000000

De acuerdo a lo anterior se tiene, la mayoría de clientes Espejo consumen atún del segmento lomo aceite (53%), el restante 47% consumos atún del segmento lomo agua, un comportamiento similar se presenta en los clientes Up Sell, mientras por el contrario en los clientes ocasionales y cross sell, la mayoría consumen atún del segmento lomo agua con un 51% el restante 49% consumen atún del segmento lomo aceite.

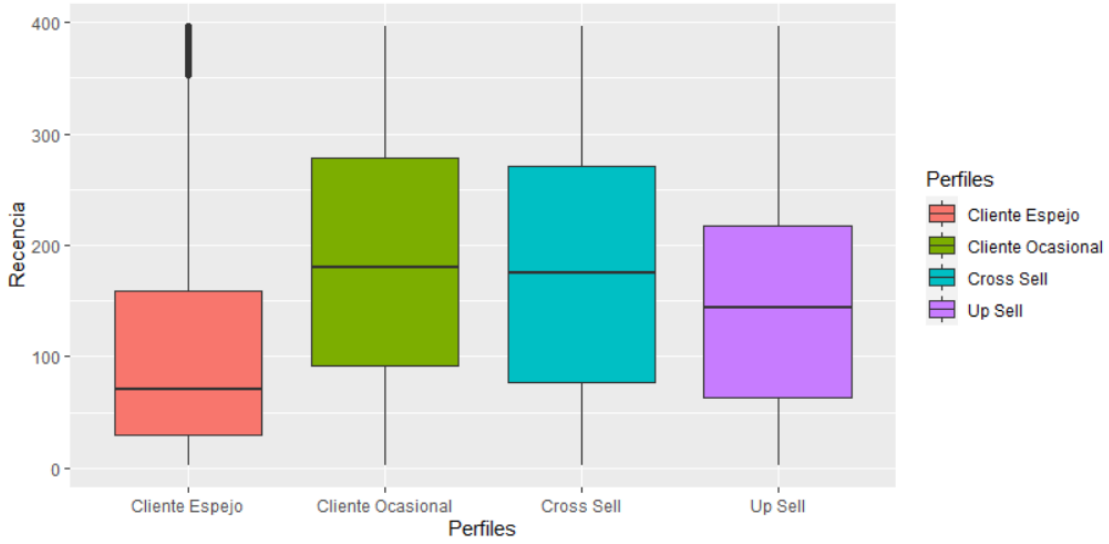
VARIABLE EDAD POR DISTRITO- BOXPLOT



De acuerdo a lo anterior se tiene, la mayoría de clientes en todos los distritos la mediana es similar cercana a 43 años, sin embargo en todos los distritos sobresalen valores atípicos de clientes con un comportamiento diferente al normal, al presenta edades superiores a 54 año en todos los distritos y menores a 26 años.

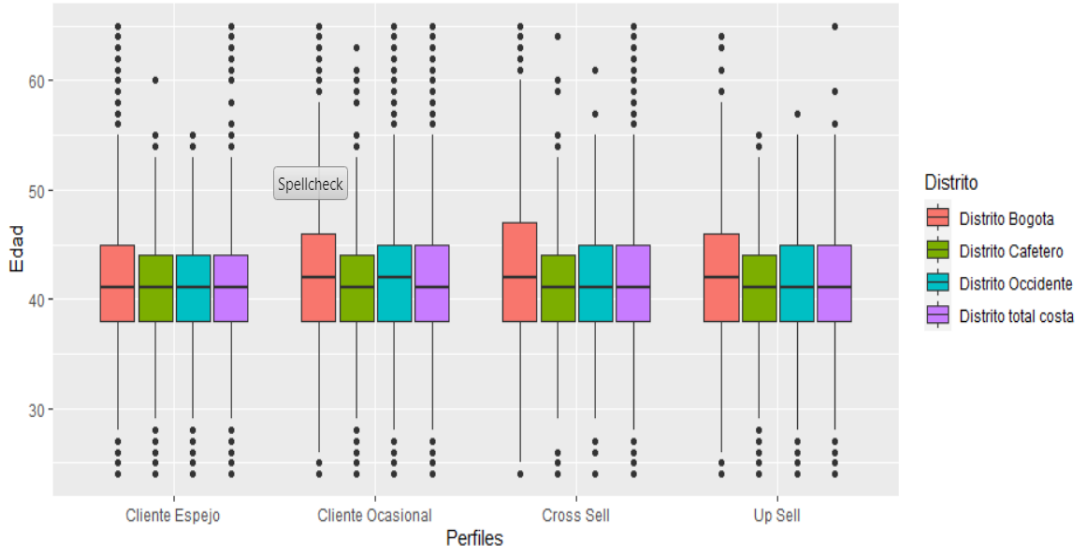
[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

Recencia vs Perfiles



De acuerdo a lo anterior se tiene un mayor valor de la mediana en recencia en los clientes ocasionales y cross cell cercano a 200 días desde la ultima compra, mientras en los clientes espejo se observa muy marcada una baja recencia es decir que sus compras son recientes cercana la mediana a 70 días. Loa clientes up sell presentan un valor de la mediana de la recencia cercana a 80 días.

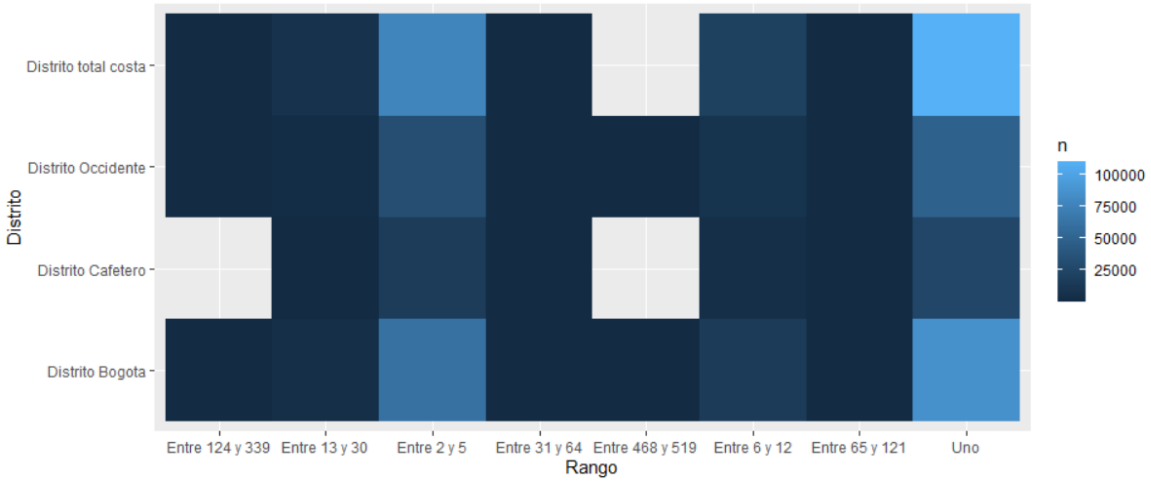
Edad vs Distrito vs Perfiles



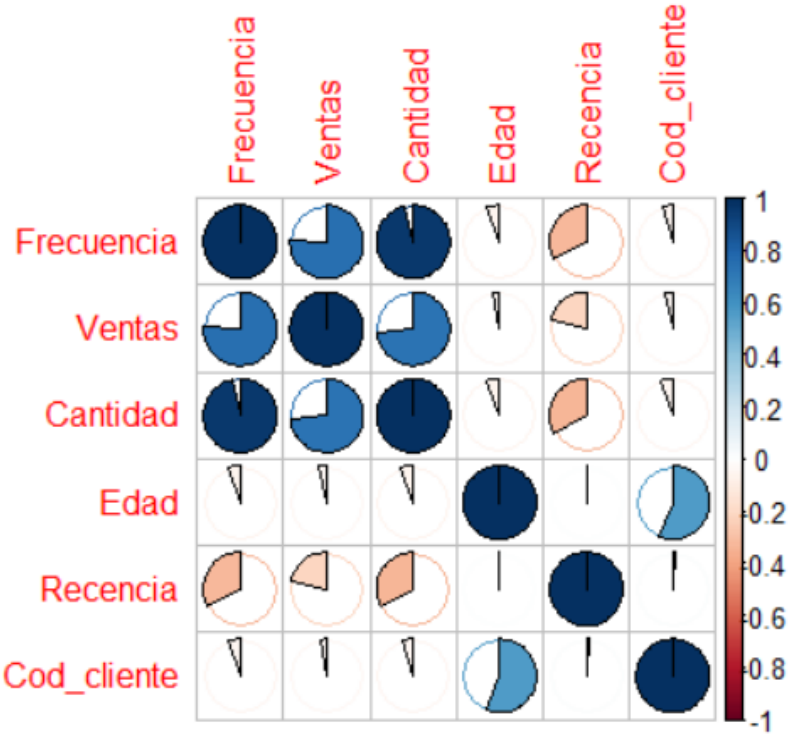
[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

De acuerdo a lo anterior se tiene que la mediana de la edad es similar en todos los distritos por perfiles con valores de edad cercana a 40 años.

MAPA DE CALOR DISTRITO POR RANGOS DE FRECUENCIA



ANALISIS DE CORRELACIONES



[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

	Frecuencia	Ventas	Cantidad	Edad	Recencia	Cod_cliente
Frecuencia	1.00	0.76	0.96	-0.06	-0.33	-0.05
Ventas	0.76	1.00	0.73	-0.04	-0.21	-0.03
Cantidad	0.96	0.73	1.00	-0.06	-0.33	-0.05
Edad	-0.06	-0.04	-0.06	1.00	0.01	0.56
Recencia	-0.33	-0.21	-0.33	0.01	1.00	0.01
Cod_cliente	-0.05	-0.03	-0.05	0.56	0.01	1.00

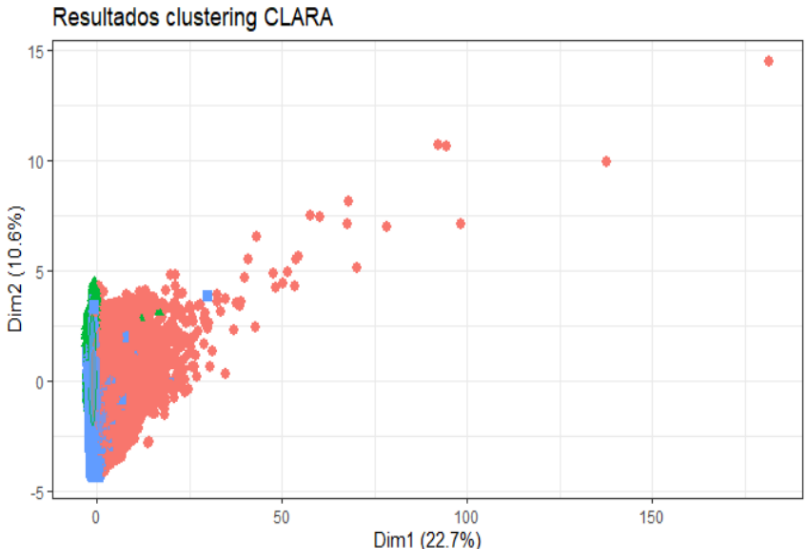
De acuerdo a lo anterior se tiene que sobresalen las correlaciones entre las variables por ser las altas respecto a las demás con un valor de 0.96 entre Cantidad y frecuencia, seguido de la correlación entre las variables ventas y cantidad, mientras la correlación mas baja se da entre las variables recencia y edad.

6.2 Resultados Modelo de Análisis de Conglomerados: Clustering

Se utilizó un análisis de conglomerados clustering, a través del algoritmo PAM, el cual no se pudo realizar dado el tamaño del vector así estuviese normalizado, hierarchical clustering, tampoco obtuvo un buen resultado ya que no fue posible calcular el vector, kmeans y clara, donde se seleccionaron 3 clusters después de revisar el grafico ilustrativo a seleccionar los grupos óptimos, para la generación de grupos se utilizó la base de datos resumida, con los clientes únicos con las variables resumen en el periodo histórico de análisis como edad, ventas, recencia, frecuencia, cantidad, este descubrimiento de patrones es de gran importancia para conocer los clientes de la categoría, perfilar y caracterizar cada grupo a través de las variables que marcan la diferencia de un grupo respecto a otro, permitiendo identificar el grupo de clientes oro, plata y bronce, para aplicar acciones focalizadas en cada grupo según el comportamiento de cada uno.

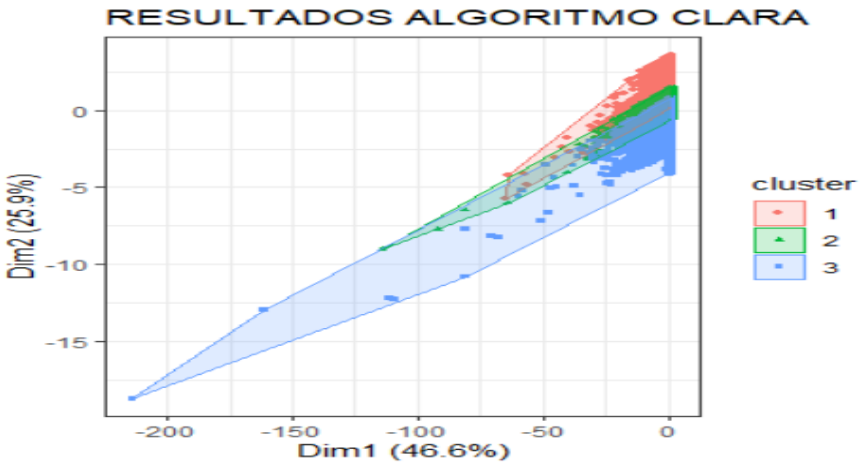
ALGORITMO CLARA

A continuación: se presenta el resultado del algoritmo CLARA:



[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

De acuerdo a lo anterior se observa el grupo rosado con mayores valores atípicos respecto a los demás grupos. También se utilizó otro gráfico y de nuevo se observa el grupo en azul es el mismo del grupo anterior es color azul como más valores atípicos respecto al resto de grupos.



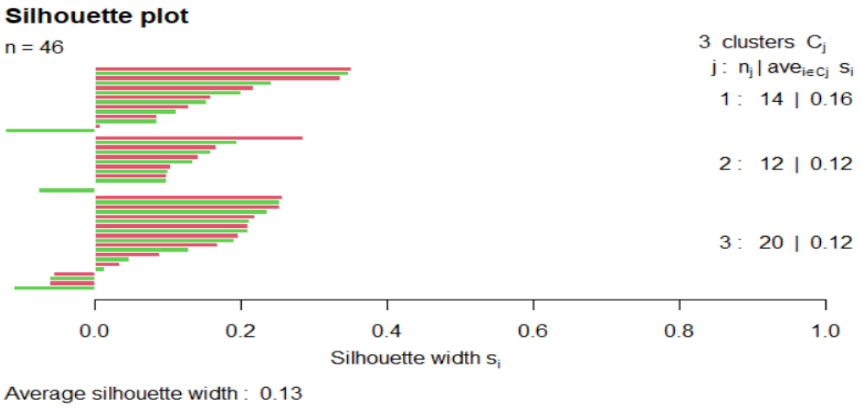
Al generar los resultados del clustering con la información resumen se tiene lo siguiente:

- ✓ Los 3 grupos dieron muy proporcionados, el primer grupo agrupa a 195.913 clientes, mientras el grupo 2 tiene 167.749 clientes, y el grupo 3 a 149.641 clientes. Según los indicadores de validación es posible decir que se conserva que un grupo es totalmente heterogéneo respecto al otro, así como que los individuos dentro de cada grupo son homogéneos.

```
> clara_clusters$clusinfo
       size max_diss   av_diss isolation
[1,] 195913 804.0269 13.578177 63.213262
[2,] 167749  93.2895  9.963722  8.873985
[3,] 149641 158.8189 10.618564 15.107342

 $clus.avg.widths
 [1] 0.1648491 0.1172194 0.1214271

 $avg.width
 [1] 0.1335448
```

MAPA DE SILUETA

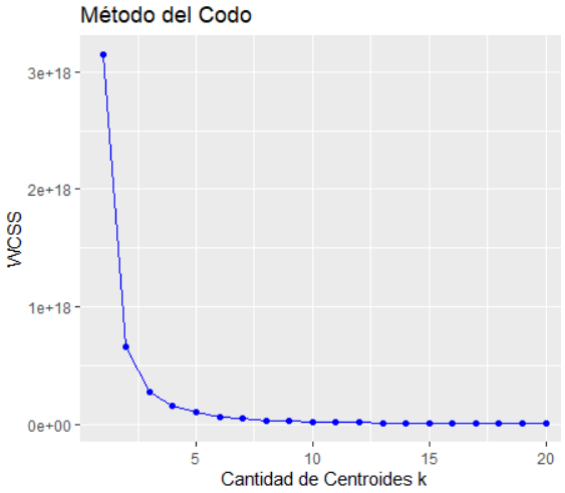
[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y



De acuerdo a lo anterior se tiene un indicador bajo, dado que es un indicador entre -1 y 1, donde un valor alto indica que el objeto esta bien emparejado con su propio cumulo y mal emparejado con los cúmulos vecinos.

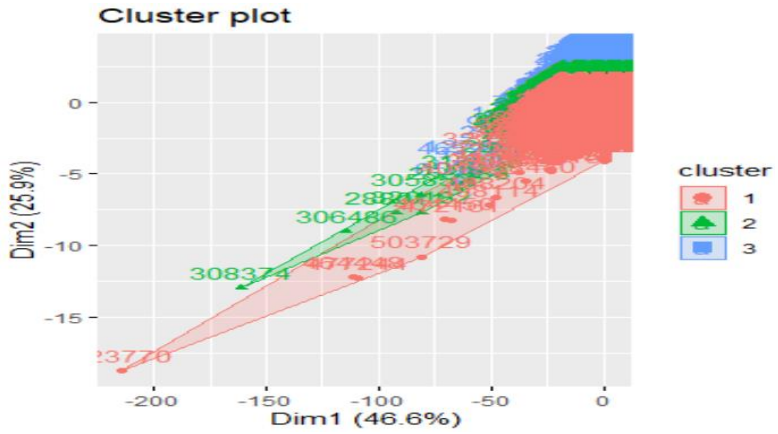
ALGORITMO K-MEANS

Al utilizar este algoritmo se realizo el grafico para tener una visión y seleccionar el numero optimo clusters, como se presenta a continuación:



Según la curva obtenida se puede ver a medida que aumenta la cantidad de centroides, el valor de WCSS disminuye de tal forma que la gráfica adopta una forma de codo, para seleccionar el valor óptimo de k, se selecciona el punto en donde ya no se dejan de producir variaciones importantes del valor de WCSS al aumentar k, en este caso, vemos que esto se produce a partir de k >= 5, por lo que evaluaremos los resultados del agrupamiento, de acuerdo a lo anterior se tomó la decisión de tomar 3 clusters, también por la estrategia de la compañía después de revisar los resultados en conjunto con los expertos de dominio para aplicar estrategias focalizadas según las preferencias o hábitos de consumo en cada grupo suele ser mas eficiente en 3 grupos que en 5, donde podría empezar complicarse, cada que sea ejecutado el modelo.

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y



A continuación se presentan los promedios resumen obtenido de cada variable en los grupos obtenidos:

```
K-means clustering with 3 clusters of sizes 147328, 175527,
 190448

Cluster means:
  Frecuencia   Ventas Cantidad     Edad Recencia
1   2.860176 26984.42 682.5451 40.01240 151.6510
2   2.986578 29008.61 712.1418 39.69127 146.2663
3   2.643063 25779.02 630.2535 45.85840 149.2128

Within cluster sum of squares by cluster:
[1] 9.475634e+16 8.912711e+16 8.889238e+16
 (between_SS / total_SS =  91.3 %)
```

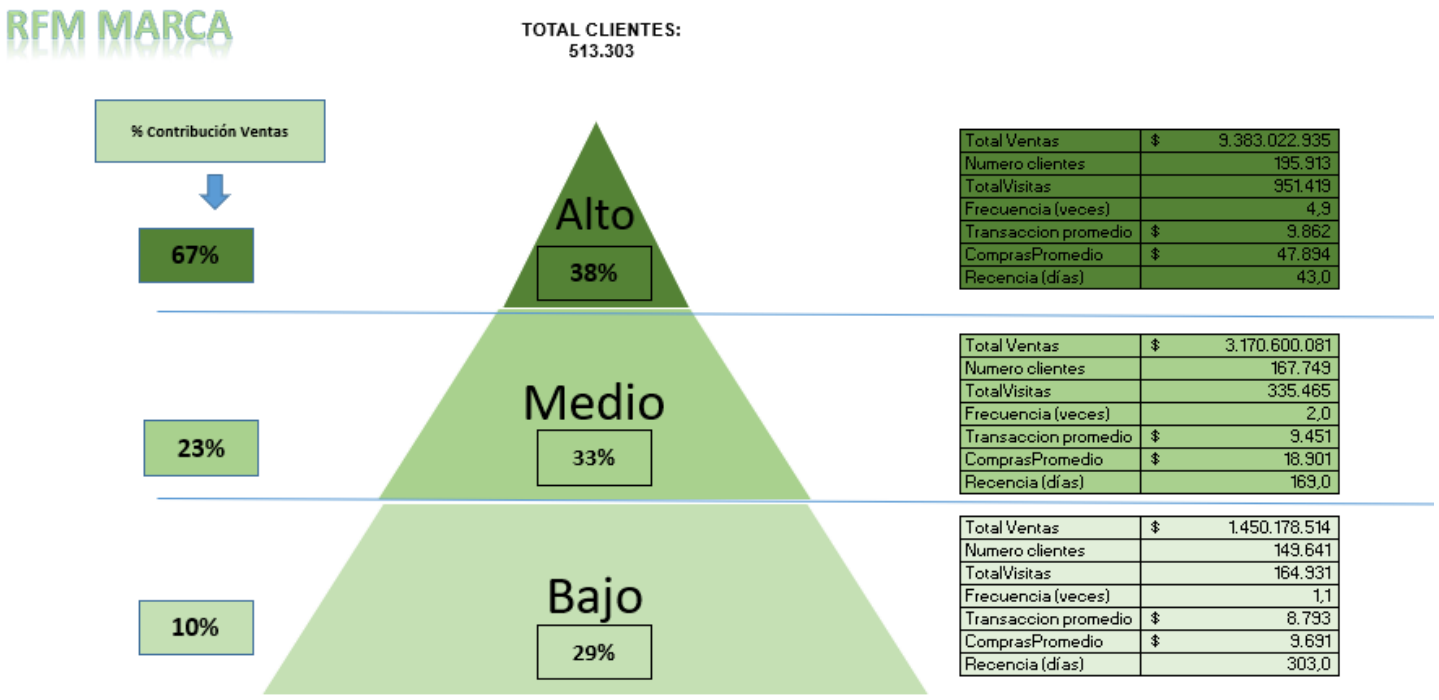
De acuerdo a lo anterior se obtuvieron 3 grupos con las siguientes cantidades de clientes:

Grupo 1= 147.328 clientes, Grupo 2= 175.527 clientes y grupo 3=190.448 clientes, el grupo 2, presenta los clientes con mayor frecuencia= 3 veces en el periodo histórico, con mayores ventas y edad promedio de 39 años.

El algoritmo PAM fue utilizado pero el resultado no fue asertivo por lo tanto no se tomo en cuenta.

Después de revisar los resultados de los clusters en conjunto con los expertos de negocio se construyó la siguiente presentación de resultados, con 3 grupos principales, seleccionado los del algoritmo clara:

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y



De acuerdo a lo anterior se obtuvieron hallazgos muy relevantes para los clientes, lo cual les permite focalizar estrategias en cada grupo fue posible identificar los perfiles principales de los clientes a través de 3 grupos principales: Clientes Oro: altos, Clientes Plata: medios y Clientes Bronce: bajos en cada segmento se perfilaron a continuación se presentan los perfiles:

- ✓ **Clientes Oro:** altos**,** este grupo se caracteriza ´por ser los clientes con mas altas ventas respecto a los demás grupos contribuyendo en un 67% de las ventas, corresponden al 38% del total de clientes de la categoría, son los mas frecuentes en sus compras en el periodo con una frecuencia igual a 5 veces en el mes, sus compras son las mas recientes en el periodo, presentan mas altas compras promedio respecto al resto, por lo cual se deben aplicar estrategias focalizadas para mantenerlos en este segmento, fidelizándolos con estrategias según su patrón de comportamiento de consumo haciendo combos promocionales del atunes con otros productos frecuentes.
- ✓ **Clientes Plata: medios,** este grupo se caracteriza ´por ser los clientes con ventas ni tan altas ni tan bajas, cercanas al promedio, respecto a los demás grupos contribuyendo en un 23% de las ventas, corresponden al 33% del total de clientes de la categoría, son poco frecuentes en sus compras en el periodo con una frecuencia igual a 2 veces en el mes, sus compras son no son recientes en el periodo con una recencia de 169 días transcurridos desde su última compra, presentan compras promedio respecto al resto, por lo cual se deben aplicar estrategias focalizadas para subirlos al segmento oro y evitar que pasen el segmento bajo.
- ✓ **Clientes Bronce: bajos ,** este grupo se caracteriza ´por ser los clientes con menores, respecto a los demás grupos contribuyendo en un 10% de las ventas, corresponden al 29% del total de clientes de la categoría, son muy poco frecuentes en sus compras en el periodo con una frecuencia igual a 1 vez en el mes, sus compras son no son recientes en el periodo con una recencia muy alta desde su última compra, presentan

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

compras promedio muy bajas respecto al resto, por lo cual se deben aplicar estrategias focalizadas para subirlos al segmento plata.

7. CONCLUSIONES

A partir del análisis clustering se concluye;

- ✓ La mayoría de los clientes de la categoría atunes corresponden a clientes del segmento alto, con mas altas ventas respecto a los demás grupos contribuyendo en un 67% de las ventas, corresponden al 38% del total de clientes de la categoría, son los mas frecuentes en sus compras en el periodo con una frecuencia igual a 5 veces en el mes, sus compras son las mas recientes en el periodo, presentan mas altas compras promedio respecto al resto, por lo cual se deben aplicar estrategias focalizadas para mantenerlos en este segmento, fidelizándolos con estrategias según su patrón de comportamiento de consumo haciendo combos promocionales del atunes con otros productos frecuentes.

- ✓ Con base en los algoritmos probados como k-means, PAM y CLARA, el de mejor ajuste a los datos fue el CLARA, después de validar con los expertos de dominio los resultados de los grupos formados por los diferentes algoritmos, y diferentes cantidades de grupos 3 y 5, se tomó la decisión de usar el modelo con 3 grupos por facilidad y simplicidad en tener 3 grupos, para tomar decisiones y aplicar estrategias focalizadas en cada segmento.

- ✓ El algoritmo PAM, no obtuvo buen resultado en los grupos seleccionando 5 o 3, dado que no fue posible generar los grupos por la dimensión de la matriz en el cálculo, por lo tanto no fue tenida en cuenta en los análisis.

- ✓ Las estrategias a implementar en cada grupo son estrategias negocio, en el grupo de clientes altos u oro, se propone utilizar otros modelos de venta cruzada que permitan analizar los productos que llevan en conjunto los clientes para aplicar regalos sorpresas y fidelizarlos, mientras en los clientes medios o plata se pretende también establecer bonos descuentos de atún y productos que llevan en conjunto para incentivarlos a comprar mas atunes e incrementar su compra promedio para empezar a subirlos al segmento oro, en los clientes bronce por ahora por ser tan ocasionales no se desgastaran en ellos.

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

8. RECOMENDACION

Posterior a la realización del análisis presentado en este trabajo, el cual se obtuvo a partir de las bases suministradas por el cliente del sector retail, se recomienda lo siguiente;

- Utilizar un modelo de regresión logística para predecir la fuga de clientes de la categoría y aplicar estrategias de retención y fidelización mas robustas, en el anexo se presenta una aproximación de avance inicial.
- Utilizar el método Tucker 3 en R, así como realizar un acercamiento con el método multivariado: AFM: Análisis Factorial Múltiple, el cual es más robusto para analizar diferentes dimensiones de los datos a través de individuos, en años y distritos, ambos métodos presentan elementos comunes, sin embargo el método Tucker tiene una ventaja en relación a las trayectorias para mostrar cambios marcados en la evolución del tiempo, las cuales favorecen la interpretación y son más diversas, permitiendo una interpretación más profunda y clara del estudio.
- Debido al enfoque en el sector retail de la categoría atunes, resulta de gran interés, considerar la posibilidad de construir otros indicadores a partir de las principales variables transaccionales para dar información adicional acerca del comportamiento de los clientes en la categoría, lo cual permite enriquecer y complementar el análisis para su interpretación. Esto podría hacerse con la ayuda de expertos en la construcción de estos indicadores.

BIBLIOGRAFIA

LEBART, L, et al. *Statistique Exploratoire Multidimensionnelle*. Dunod, Francia, 1995.

AMAT RODRIGO J, *CLUSTERING Y HEATMAPS: APRENDIZAJE NO SUPERVISADO*, [Consulta en línea]: https://rpubs.com/Joaquin_AR/310338, 2019.

AMAT RODRIGO J, *REGRESION LOGISTICA SIMPLE Y MULTIPLE*, [Consulta en línea]: https://rpubs.com/Joaquin_AR/229736, 2016.

HIDALGO, K, *Cluster,* [Consulta en línea]: https://rpubs.com/mcelestemc/550188, 2020

GOMEZ, SAMCHEZ, J, *Análisis comparativo de diferentes métodos de agrupación para el tratamiento de datos de expresión genética*, Universidad de Cataluña, 2018.

LRomero, MRamirez, JRojas, EDarghan*, Analisis Cluster,* [CONSULTA EN LÍNEA]: HTTPS://RPUBS.COM/LHROMEROJ/ANALISISDECLUSTERR

PEÑA, D. *Análisis de datos multivariantes.* McGraw Hill, Madrid – España, 2002.

PARDO, Campo Elías & CABARCAS, G. *Métodos Estadísticos Multivariados en Investigación Social*. Simposio de Estadística, Colombia, 2001, p. 53-71.

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

ESCOFIER, B. & PAGÉS, J. *Análisis factoriales simples y múltiples*. Servicio Editorial Universidad País Vasco, París – Francia, 1992, p. 179-229.

GONZÁLEZ ROJAS, Víctor Manuel. *Notas de Clase del Curso Estadística Aplicada IV*. Programa de Estadística, Universidad del Valle, Colombia, 2006.

ABASCAL FERNÁNDEZ, E.; LANDALUCE CALVO, M.I. (2002) "Análisis Factorial Múltiple como técnica de estudio de la estabilidad de los resultados de un Análisis Componentes Principales", *Questiió*, 26, 1-2, pp. 109-122.

LORA, E. Técnicas de medición económica. Editorial Alfaomega.

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

ANEXOS

ANEXO 1. Distribución de ventas mensuales categoría atunes

RESUMEN VENTAS MENSUAL CATEGORÍA ATUNES

DISTRITO	ene	feb	mar	abr	may	jun	jul	ago	sep	oct	nov	dic
Distrito Bogota	296.529.792	257.522.297	280.892.019	593.351.787	301.764.562	317.712.026	345.255.870	789.925.150	338.255.200	350.929.960	340.031.250	402.299.610
Distrito Cafetero	67.198.480	53.467.670	58.567.386	81.037.124	68.529.620	73.150.530	82.430.370	127.166.000	77.685.560	74.566.050	71.793.000	92.020.980
Distrito Occidente	138.288.056	102.578.558	125.352.470	149.621.090	149.178.020	157.997.960	170.963.775	255.235.750	175.370.784	171.891.750	164.379.170	193.496.790
Distrito Total costa	382.327.062	278.471.496	338.891.368	401.551.240	400.437.380	399.132.530	453.434.530	697.513.080	524.944.710	504.356.159	481.106.892	606.298.007
Total	**884.343.390**	**692.040.021**	**803.703.243**	**1.225.561.241**	**919.909.582**	**947.993.046**	**1.052.084.545**	**1.869.839.980**	**1.116.256.254**	**1.101.743.919**	**1.057.310.312**	**1.294.115.387**

Fuente: Propia

CODIGO ANEXO.

```
datos<-read_excel("C:/Users/user/Desktop/Tesis Maestria Estadistica/BADE DE DATOS.xlsx")
library(readxl)
library(gmodels)
library(Hmisc)
library(ggthemes)
library(dplyr)
library(ggplot2)
library(scales)
library(reshape2)
library(car)
library(tidyverse)
library(datos)
library(PerformanceAnalytics)
library(corrplot)
library(Amelia)
#Analisis Descriptivo
#Resumen General
```

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

```
head(datos)
glimpse(datos)
str(datos)
describe(datos)
summary(datos)
ggplot(datos, aes(x = Perfiles, y = Edad,
fill = Distrito)) + geom_boxplot()
ggplot(datos, aes(x = Perfiles, y = Recencia,
fill = Perfiles)) + geom_boxplot()
#Filtro variables cuantitativas
datos2 <- datos[, c("Frecuencia", "Ventas", "Cantidad", "Edad", "Recencia","Cod_cliente")]
head(datos2)
corrplot(cor(datos2),method = c("pie"))
pie(table(datos2$Distrito))
ggplot(datos2, aes(x = Distrito, y = Ventas,
fill = Distrito)) + geom_boxplot()
# Calidad Datos
missmap(datos)
#Tablas Frecuencias
table(datos$Distrito, datos$Segmento)
#tabla en porcentajes Distrito vs segmento
prop.table(table(datos$Distrito, datos$Segmento))
# tabla filas Distrito vs segmento
tablafilas<-prop.table(table(datos$Distrito, datos$Segmento),1)
tablafilas
#tabla en porcentajes Distrito vs perfil
prop.table(table(datos$Distrito, datos$Perfiles))
# tabla filas Distrito vs perfil
tablafilas<-prop.table(table(datos$Distrito, datos$Perfiles),1)
```

---

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

```
tablafilas
addmargins(tablafilas)
# tabla filas Perfiles vs segmento
tablafilas<-prop.table(table(datos$Perfiles, datos$Segmento),1)
tablafilas
addmargins(tablafilas)
#tabla columnas
tablacolumnas<-prop.table(table(datos$Distrito,datos$Segmento),2)
tablacolumnas
addmargins(tablafilas, margin=1)

tablacolumnas<-CrossTable(datos$Distrito, datos$Segmento)
CrossTable(datos$Distrito, datos$Segmento, prop.t=F, prop.chisq = F)
CrossTable(datos$Segmento)

# Graficos
#Histograma Edad
hist(datos$Edad, main = "Histograma Edad clientes Categoría atunes",
xlab = "Edad clientes", ylab = "frecuencia",
col = "steelblue")
#Histograma Recencia
hist(datos$Recencia, main = "Histograma Recencia clientes Categoría atunes",
xlab = "Recencia clientes", ylab = "Recencia",
col = "steelblue")
#Histograma clientes por distrito
barplot(table(datos$Distrito))
#Histograma clientes por Perfiles
barplot(table(datos$Perfiles))
#Histograma clientes por segmento
```

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

```
barplot(table(datos$Segmento))
#Mapa de calor rango de frecuencia por distrito
datos %>%
count(Rango, Distrito) %>%
ggplot(mapping = aes(x = Rango, y = Distrito)) +
geom_tile(mapping = aes(fill = n))
#Mapa de calor rango de frecuencia por segmento
datos %>%
count(Rango, Segmento) %>%
ggplot(mapping = aes(x = Rango, y = Segmento)) +
geom_tile(mapping = aes(fill = n))

#Boxplot edad por distrito
qplot(Distrito, Edad, data=datos, geom="boxplot")

#Filtro variables cuantitativas
datos2 <- datos[, c("Frecuencia", "Ventas", "Cantidad", "Edad", "Recencia","Cod_cliente")]
head(datos2)

#MATRIZ DE CORRELACIONES
correlacion<-round(cor(datos2), 2)
correlacion
correlacion<-round(cor(datos2), 1)
corrplot(correlacion, method="number", type="upper")
chart.Correlation(datos2, histogram = F, pch = 19)
#Modelo clustering algoritmo clara
library(factoextra)
library(cluster)
library(clv)
```

---

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

```
library(FactoMineR)

#metodo de conglomeracion
clara_clusters <- clara(x = datos, k = 3, metric = "manhattan", stand = TRUE,
samples = 50, pamLike = TRUE)
clara_clusters
d1<-cclust(clara_clusters, 3, dist="euclidean")
shadow(d1)
#Grafico clusters CLARA
fviz_cluster(object = clara_clusters, ellipse.type = "t", geom = "point",
pointsize = 2.5) +
theme_bw() +
labs(title = "Resultados clustering CLARA") +
theme(legend.position = "none")
#Indicadores cluster

clara_clusters$silinfo #medida de silueta
clara_clusters$objective
clara_clusters$diss
clara_clusters$data
clara_clusters$clustering #vector del clustering
clara_clusters$clusinfo
clara_clusters$medoids
clara_clusters$call
plot(silhouette(clara_clusters),  col = 2:3, main = "Silhouette plot")
fviz_cluster(clara_clusters)
#Variable cluster
cl<- data.frame(clara_clusters$clustering)
cl
```

---

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

```
#Exportar
write.csv(c1,"C:/Users/vhmb1218/Desktop/cluster.csv")
dataexpr<-data.frame(datos2,cl)
#Grafico CLARA
clara_cluster<-clara(datos2,3)
fviz_cluster(clara_cluster,stand = T,geom = "point",pointsize = 1)+
theme_bw() +
labs(title = "RESULTADOS ALGORITMO CLARA")

#Algoritmo k-means
#Determinación óptima del número de cluster

km.res <- kmeans(datos2, 3, nstart = 25)
km.res

#seleccion de clusters
set.seed(1234)
wcss <- vector()
for(i in 1:20){
wcss[i] <- sum(kmeans(datos2, i)$withinss)
}
#Grafico clusters
ggplot() + geom_point(aes(x = 1:20, y = wcss), color = 'blue') +
geom_line(aes(x = 1:20, y = wcss), color = 'blue') +
ggtitle("Método del Codo") +
xlab('Cantidad de Centroides k') +
ylab('WCSS')

set.seed(1234)
```

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

```
kmeans <- kmeans(datos2, 3, iter.max = 1000, nstart = 10)
kmeans

km <- eclust(datos2,FUNcluster="kmeans", k=3,hc_metric = "euclidean")

#indicadores cluster k-means
kmeans$cluster #vector cluster
kmeans$centers
kmeans$withinss
kmeans$totss



#Algoritmo PAM

pam.res <- pam(datos2, 3)
cliente1_scale<- scale(datos2)
fviz_nbclust(x = cliente1_scale,FUNcluster = pam, method="wss" , k.max = 15,
diss = dist(cliente1_scale,method = "manhattan"))

set.seed(111)
pam_cluster <- pam(x = datos2,k = 3,metric = "manhattan")
pam_cluster

fviz_nbclust(x = datos2, FUNcluster = pam, method = "wss", k.max = 15,
diss = dist(datos2, method = "manhattan"))

pm <- eclust(datos2,FUNcluster="pam", k=3,hc_metric = "euclidean")
#Algoritmo hierarchical clustering
sus_arrests<-scale(datos2)
```

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y

```
head(datos2)
dist_data<-dist(sus_arrests, method = 'euclidean')


library(Factoclass)
FC.col <-FactoClass(datos2, dudi.pca)
```

[1]Fedesarrollo: El Mercado de atún en Colombia, Luis Alberto Zuleta, Alejandro Becerra, Mayo 2013, [Consulta en linea].Disponible en: https://www.repository.fedesarrollo.org.co/bitstream/handle/11445/205/El%20mercado%20del%20atun%20en%20Colombia%20.pdf?sequence=2&isAllowed=y