CARRERA: ESPECIALIZACIÓN EN CIENCIA DE DATOS

TRABAJO FINAL INTEGRADOR

SEGMENTACIÓN DE MERCADO – AGROQUÍMICOS

Nombre y Apellido del Alumno: Juan Manuel Domínguez

Título de grado o posgrado (último): Ingeniero Industrial (U.B.A.)

Tutor:

Mario Rossi

Lugar y Fecha – Buenos Aires, julio de 2022

Índice

1. Introducción

El mercado local de Agroquímicos está compuesto por una gama amplia de clientes. Entenderlo de forma agregada es clave para que las empresas productoras y comercializadoras de productos agroquímicos establezcan sus pilares de competitividad.

La Empresa sobre la cual se desarrolló el presente Trabajo Final Integrador (TFI) – *una empresa productora y distribuidora local de agroquímicos y proveedora de servicios integrales para el campo* – entiende a su mercado como un único grupo o segmento de clientes.

Esta situación impacta desfavorablemente en dos instancias del negocio: la operativa – los resultados a corto plazo – y la estratégica – la calidad de las decisiones a largo plazo.

En la instancia operativa, hay evidencias de un bajo desempeño debido a una falta de foco: *desvíos en costos, bajo nivel de servicio, entre otros*.

En la instancia estratégica, la falta de una segmentación clara deja a la compañía mal posicionada al momento de la toma de decisiones:

- Crecimiento rentable – ¿en qué segmentos de clientes la compañía puede (y debe) maximizar su margen?
- Despliegue de recursos logísticos y de producción – ¿de qué manera servir a los segmentos objetivo al mínimo costo?

Por otro lado, La Empresa no cuenta con la capacidad necesarias para resolver la segmentación de clientes. Las razones:

- falta de conocimiento acerca de enfoques de solución
- poca práctica en temas de acondicionamiento y tratamiento de datos

En este contexto, es necesario que La Empresa alcance un entendimiento objetivo y más preciso acerca del mercado: cantidad de segmentos y similitudes entre clientes de acuerdo con un conjunto de características relevantes para el negocio.

El camino para resolver este problema es dotar a La Empresa de un proceso formal de segmentación de clientes asistido por una solución basada en algoritmos de *Machine Learning* (ML).

La segmentación, o Clustering, es una técnica de aprendizaje automático no supervisada que permite descubrir de manera automática la agrupación natural en los datos. A diferencia del aprendizaje supervisado (como el modelado predictivo), los algoritmos de segmentación interpretan datos de entrada sin "etiquetar", *no hay datos input que previamente correspondan a un determinado segmento o output*. Estos algoritmos encuentran grupos de datos con valores similares en sus principales características, los cuales tienen menos o ninguna similitud con los datos de otro grupo.

2. Objetivos

El presente TFI tiene como objetivo presentar los elementos de una solución formal de segmentación que permita a La Empresa establecer la cantidad óptima de segmentos de clientes y, al mismo tiempo, identificar objetivamente similitudes entre clientes de acuerdo con un conjunto de características relevantes para el negocio.

Para ello, se trabajó en definir y crear un *dataset* de clientes con atributos y registros definidos y en desarrollar un modelo de segmentación de clientes apoyado por algoritmos de *ML*.

Objetivo general:

- Presentar los elementos de una solución de segmentación de clientes basada en algoritmos de *ML*.

Objetivos específicos:

- Preparar el *dataset* – supone definir una tabla simple de clientes y sus atributos, alimentada por un proceso de extracción y transformación de datos con origen en distintas tablas
- Desarrollar el modelo – requiere seleccionar el algoritmo de *machine learning* y el lenguaje de programación, las variables y la cantidad de *clusters* (segmentos) a considerar a priori
- Evaluar el modelo – se establecen las funciones (o métricas) de evaluación para medir la confiabilidad del modelo
- Usar el modelo – para definición de segmentos de clientes atractivos para compañía

3. Metodología

El proyecto se organizó a través de metodologías ágiles – *AGILE*. Se utilizó el enfoque *SCRUM* con períodos de entrega de valor (*Sprints*) de 1 mes. En cada *Sprint* fueron priorizadas las tareas pendientes (*Product Backlog*) conforme a las necesidades y requerimientos del presente trabajo.

El *Product backlog* tuvo como norte los objetivos específicos presentados en el capítulo anterior:

- Preparar el *Dataset*
- Desarrollar el modelo
- Evaluar el modelo
- Definir el mejor modelo

Técnicas

Fueron aplicadas técnicas de *ML* a través de algoritmos no supervisados disponibles en las librerías de *Python* y R – sigue a continuación un listado de algoritmos utilizados:

- *K-Means*
- *MiniBatchKMeans*
- GaussianMixture
- *AgglomerativeClustering*
- *SpectralClustering*
- *OPTICS*
- *MeanShift*
- *DBSCAN*

Sobre los algoritmos

K-Means – el algoritmo K-Means no requieren ningún conocimiento previo del *dataset*, dado que será capaz de identificar grupos o segmentos de clientes mediante comparaciones repetitivas de datos.

Mini-Batch K-Means - es una versión del algoritmo estándar K-Means. Utiliza lotes de datos pequeños, aleatorios y de tamaño fijo para almacenarlos en la memoria y luego, con cada iteración, se recopila una muestra aleatoria de los datos y se usa para actualizar los clústeres.

Gaussian Mixture – es un modelo probabilístico en el que se considera que las observaciones siguen una distribución probabilística formada por la combinación de múltiples distribuciones normales (componentes). Puede entenderse como una generalización de K-Means con la que, en lugar de asignar cada observación a un único cluster, se obtiene una distribución probabilidad de pertenencia a cada uno.

Agglomerative Clustering – es el tipo más común de agrupación jerárquica utilizada para agrupar objetos en clústeres en función de su similitud. También se conoce como AGNES (*Aglomerative Nesting*). El algoritmo comienza tratando cada objeto como un grupo único. A continuación, los pares de grupos se fusionan sucesivamente hasta que todos los grupos se han fusionado en un gran grupo que contiene todos los objetos. El resultado es una representación en forma de árbol de los objetos, denominada dendrograma.

Spectral Clustering – es un algoritmo de agrupamiento que trata cada punto de datos como un nodo gráfico y, por lo tanto, transforma el problema de agrupamiento en un problema de partición de gráficos.

DBSCAN – es un algoritmo de agrupamiento basado en la densidad que funciona asumiendo que los agrupamientos son regiones densas en el espacio separadas por regiones de menor densidad. Agrupa puntos de datos 'densamente agrupados' en un solo grupo. Puede identificar grupos en grandes conjuntos de datos espaciales observando la densidad local de los puntos de datos.

OPTICS – estrechamente relacionado con DBSCAN, encuentra una muestra central de alta densidad y expande los agrupamientos a partir de ellos. A diferencia de DBSCAN, mantiene la jerarquía de clústeres para un radio de vecindad variable. Más adecuado para su uso en grandes conjuntos de datos.

MeanShift – tiene como objetivo descubrir "manchas" en una densidad uniforme de muestras. Es un algoritmo basado en centroides, que funciona mediante la actualización de candidatos para que los centroides sean la media de los puntos dentro de una región determinada.

Herramientas

Se utilizaron dos lenguajes de programación – *Python y R* – junto con algunas de sus bibliotecas específicas para *ML, gráficos y manipulación de datos,* como por ejemplo:

Para Python:

- *Scikit-learn,*
- *Matplotlib*,
- entre otras.

Para R:

- *FactoMineR,*
- *Factoextra*,
- *ggplot2*,
- entre otras.

Por otro lado, el proyecto se desarrolló de manera "stand-alone", o sea, no integrado con los sistemas de La Empresa. Por lo tanto, los procesos de captura y movimiento de datos (ETL) fueron realizados de forma local a través de archivos planos mediante comandos de *Python y R.*

4. Hipótesis

Con la utilización de algoritmos de *ML* fue posible resolver la segmentación de clientes de acuerdo con los valores de sus principales características.

Variables de la hipótesis:

- Segmentos (de clientes)
- Las principales características (de los clientes)

Relación entre las variables:

- Causa (valores de las principales características de los clientes) – Efecto (segmento de cliente)

Tipo de variables:

- Variable dependiente: segmento
- Variable independiente: principales características de los clientes
- Variable contextual: clientes de una empresa de agroquímicos

Definición nominal:

- Segmento: grupo de clientes con valores similares en sus principales características
- Principales características de los clientes: son los elementos de definen a los clientes en términos de preferencias y nivel de consumo, localización geográfica, entre otros
- Clientes de una empresa de agroquímicos: son personas físicas o empresas que compran productos agroquímicos a la empresa

Definición operacional:

- Segmento: agrupación de clientes con valores similares en sus principales características
- Valores de las principales características de los clientes: resultado de la medición clasificatoria, de rango o cuantitativa de los elementos que definen a un cliente
- Clientes de una empresa de agroquímicos: es un conjunto de caracteres únicos que identifica a un cliente que compra productos agroquímicos a la empresa

Medición de las variables:

- Categóricas:
  - Rango de Precios pagados por los clientes,
  - Zonas PAS de cada cliente, zonas del Panorama Agrícola Semanal

- Cuantitativas:
  - Gama de productos por cliente,
  - Facturación por cliente,
  - Margen Bruto por cliente,
  - Facturación promedio por producto por cliente,
  - Margen Bruto sobre Facturación por cliente y
  - Frecuencia de compra por cliente.

5. *Dataset*

El *Dataset* tuvo origen en una tabla de datos con casi 84 mil registros (83.988 filas) y más de 50 variables (columnas) asociados a ventas del año 2020. Dicha tabla fue obtenida a partir de una consulta y extracción del sistema transaccional de La Empresa.

Cada registro es una transacción de venta que involucra a un solo cliente, un solo producto, una fecha, un monto de venta y margen bruto – principalmente.

En el año 2020 La Empresa ofreció una gama de 435 productos finales. De acuerdo con los registros, un total de 5.853 clientes compraron al menos 1 producto de dicha gama.

Para obtener un *Dataset* de trabajo de calidad fue necesario realizar una seria de "preprocesamientos", entre ellos: agrupaciones convenientes de datos y la generación de nuevas variables categóricas y numéricas.

Producto de una agrupación de datos surgieron las primeras 5 variables del *Dataset* asociadas a cada cliente:

- cantidad total de productos adquiridos (gama),
- monto de venta total (facturación),
- frecuencia de compra
- zona de operación (zona PAS) y
- margen bruto total.

Por otro lado, fueron creadas 3 variables nuevas asociadas a cada cliente, cuyo objetivo fue aportar *"dimensiones"* adicionales al proceso de segmentación:

- margen bruto sobre facturación (rentabilidad),
- facturación sobre cantidad de productos y
- rango de precio.

Por lo tanto, a partir de la tabla de consulta original de casi 84 mil registros se obtuvo un *Dataset* de trabajo de 5.853 registros, uno por cada cliente.

A continuación, se describirá cada una de las variables que contiene el *Dataset* de trabajo:

i. Cantidad de productos comprados por cliente. Es la cantidad total de productos diferentes que compró el cliente en el año 2020 – a partir de una gama de 435 productos.

   El nombre de esta variable en el *Dataset* es **Gama_Productos**. Es una variable cuantitativa discreta.

ii. Rango de precio pagado por el cliente. Es una variable categórica, dado que es un rango asociado al precio que pagan los clientes por los productos comprados en el año 2020.

En este negocio, el precio final pagado por un producto dado está abierto a la negociación en el momento la operación.

Fueron establecidos 5 valores para esta variable categórica a partir de 4 rangos de precio:

- 1: BAJO – el precio de todos los productos que compró el cliente está entre el precio mínimo observado y el primer cuartil de precio pagado por los clientes.
- 2: MEDIO-BAJO – el precio de todos los productos que compró el cliente entre el primer y la mediana del precio pagado por los clientes.
- 3: MEDIO-ALTO – el precio de todos los productos que compró el cliente está entre la mediana y tercer cuartil de precio pagado por todos los clientes.
- 4: ALTO – el precio de todos los productos que compró el cliente está entre el tercer cuartil y el precio máximo pagado por los clientes.
- 5: MIX – los productos comprados pertenecen a distintos rangos de precio.

El nombre de esta variable en el *Dataset* es **rango_precio**. Es una variable categórica.

iii. Venta total. Es la facturación total obtenida por cliente en 2020.

Es una variable cuantitativa continua. El nombre de esta variable en el *Dataset* es **Facturacion**.

iv. Margen bruto. Es la diferencia entre la facturación y el costo por cantidad de los productos vendidos a cada cliente. Es una variable cuantitativa continua.

El nombre de esta variable en el *Dataset* es **Margen_Bruto**.

v. Venta promedio por producto. Es la venta total dividida la cantidad de productos comprados por cliente.

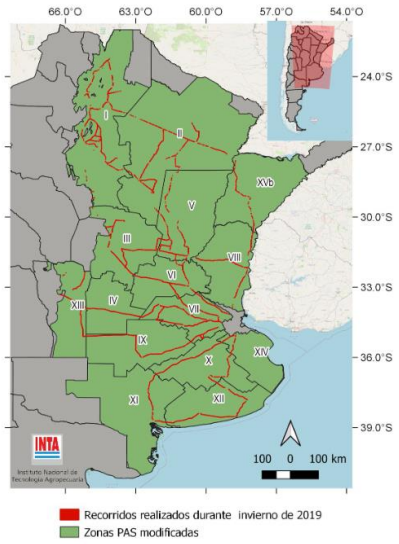
Es una variable cuantitativa continua. El nombre de esta variable en el *Dataset* es **Facturación_s_#Producto**.

vi. Porcentaje de margen sobre facturación. Es el margen bruto dividido por la venta total por cliente – o “rentabilidad” por cliente.

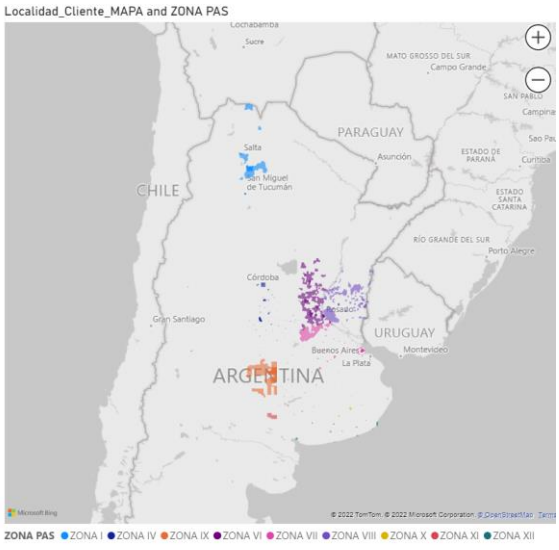
El nombre de esta variable en el *Dataset* es **Margen_s_Facturacion**. Es una variable cuantitativa continua.

vii. Zona. Para cada cliente se consideró una zonificación que incluye las principales áreas agrícolas del país basada en las zonas del Panorama Agrícola Semanal (PAS) de la

Bolsa de Cereales de Buenos Aires (2019)[1] – ver mapa 1. En el mapa 2 se localizaron los 5.853 clientes por zonas (en 10 zonas PAS). El nombre de esta variable en el *Dataset* es **ZONA PAS**. Es una variable categórica.



Mapa 1



Mapa 2

Datos sobre las zonas PAS

ZONA_PAS	Count - #cliente	% - #cliente	Sum - Facturacion	% - Facturacion	Sum - Margen_Bruto	% - Margen_Bruto	Sum - Frecuencia_Compra	% - Frecuencia_Compra	Margen_s_Facturacion
ZONA_I	93	1,59 %	960.435	1,15 %	178.823	1,02 %	428	1,09 %	25
ZONA_IV	898	15,34 %	11.290.920	13,55 %	2.205.466	12,60 %	5.346	13,62 %	26
ZONA_IX	233	3,98 %	3.817.498	4,58 %	721.908	4,13 %	1.648	4,20 %	21
ZONA_VI	886	15,14 %	12.176.390	14,61 %	2.420.860	13,83 %	5.466	13,93 %	25
ZONA_VII	2.393	40,89 %	33.856.452	40,62 %	7.298.563	41,71 %	16.907	43,09 %	25
ZONA_VIII	783	13,38 %	11.857.080	14,22 %	2.502.079	14,30 %	5.419	13,81 %	26
ZONA_X	154	2,63 %	2.523.827	3,03 %	656.378	3,75 %	990	2,52 %	28
ZONA_XI	89	1,52 %	1.269.511	1,52 %	218.372	1,25 %	746	1,90 %	21
ZONA_XII	293	5,01 %	4.181.787	5,02 %	1.056.735	6,04 %	2.067	5,27 %	26
ZONA_XIII	31	0,53 %	1.424.009	1,71 %	240.601	1,37 %	220	0,56 %	22
Total Result	**5.853**	**100,00 %**	**83.357.909**	**100,00 %**	**17.499.785**	**100,00 %**	**39.237**	**100,00 %**	**25**

viii. Frecuencia de Compra. Es la cantidad de contactos de compra que realizó un cliente durante el año, medido en cantidad de contactos por año.

El nombre de esta variable en el *Dataset* es **Frecuencia_Compra**. Es una variable cuantitativa discreta.

[1] https://inta.gob.ar/sites/default/files/mapa_nacional_de_cultivos_2019_2020_v1.pdf

5.1 Análisis del Dataset

La Tabla 1 muestra una vista de la consola de *Python* con el resultado de la importación del *Dataset* – primeros y últimos 5 registros.

#cliente	Gama_Productos	rango_precio	Facturacion	Margen_Bruto	Facturacion_s_#Producto	Margen_s_Facturacion	ZONA_PAS	Frecuencia_Compra
C_1	3	5	2302	243	767	11	ZONA_VII	1
C_10	3	4	2700	214	900	8	ZONA_VIII	1
C_100	23	3	126873	73973	5516	58	ZONA_VII	35
C_1000	17	1	55172	11428	3245	21	ZONA_IV	19
C_1001	17	4	12893	2660	758	21	ZONA_VIII	12
...	...	...	...	...	...	...	...	...
C_995	3	3	1675	396	558	24	ZONA_VII	5
C_996	28	5	13196	3026	471	23	ZONA_VIII	17
C_997	2	5	29500	11977	14750	41	ZONA_VI	15
C_998	10	1	8935	702	894	8	ZONA_VI	8
C_999	8	1	4229	979	529	23	ZONA_VI	5

[5853 rows x 8 columns]

Tabla 1

Principales características de los datos de la muestra

La tabla 2 presenta datos de las principales características de posición y dispersión de las cinco (5) variables numéricas del *Dataset*: Gama_Productos, Facturacion, Margen_Bruto, Facturacion_s_#Producto, Margen_s_Facturacion y Frecuencia_Compra.

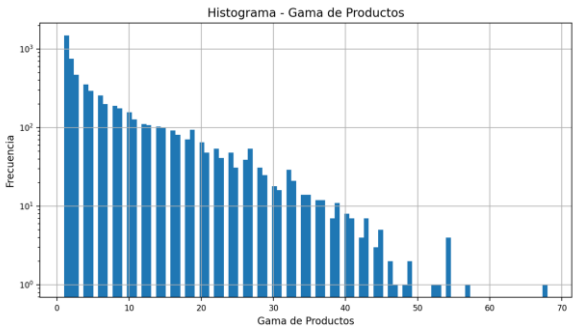
	Gama_Productos	rango_precio	Facturacion	Margen_Bruto	Facturacion_s_#Producto	Margen_s_Facturacion	Frecuencia_Compra
count	5853.0	5853.0	5853.0	5853.0	5853.0	5853.0	5853.0
mean	8.0	3.0	14242.0	2990.0	1907.0	25.0	7.0
std	9.0	2.0	31539.0	7310.0	5632.0	15.0	8.0
min	1.0	1.0	4.0	1.0	4.0	1.0	1.0
25%	1.0	1.0	1027.0	229.0	400.0	15.0	1.0
50%	4.0	3.0	3945.0	833.0	861.0	22.0	4.0
75%	11.0	4.0	13371.0	2746.0	1840.0	32.0	9.0
max	68.0	5.0	572583.0	151862.0	258156.0	125.0	101.0

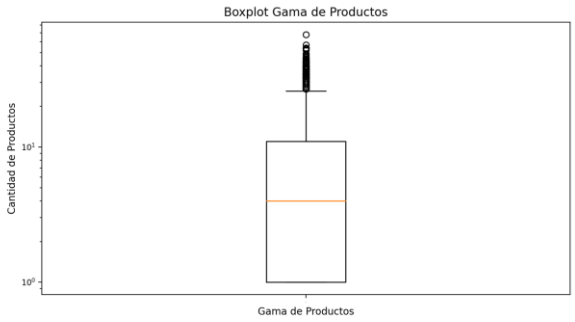
Tabla 2

Para la variable **Gama_Productos**, la muestra presenta un rango entre 1 y 68, o sea, los clientes compraron entre 1 y 68 productos, de 435 disponibles. Adicionalmente, el 25% de los clientes compró sólo 1 producto, el 50% menos de 5 y el 75% menos de 12 productos, casi el 3% de la gama.

Podemos ver en el histograma una fuerte asimetría a derecha, indicando una concentración de clientes comprando pocos productos. El gráfico *boxplot* confirma también la asimetría de la muestra con varios *outliers* por encima del límite superior.

Por lo tanto, y de acuerdo con los datos, la gran mayoría de los clientes compran muy pocos productos de la gama ofrecida – esta información es relevante para La Empresa dado que una de sus fortalezas competitiva es la GAMA de productos.



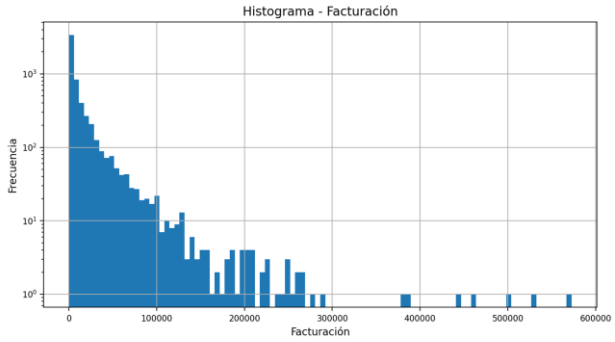


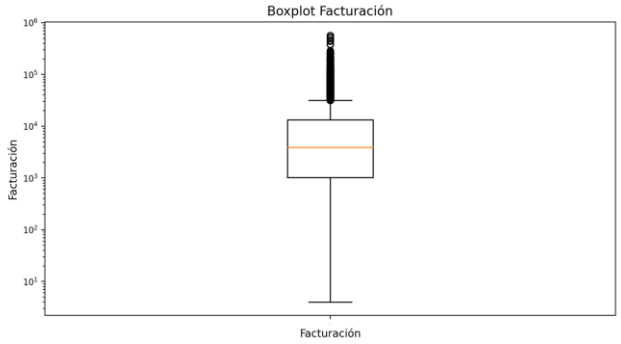
En cuanto a la variable **Facturacion**, la muestra presenta un rango amplio entre 4 y 572.583 dólares por cliente al año en compras de fertilizantes a La Empresa. Un 25% compra menos de 1.024 dólares al año (125 litros de fertilizante aprox.). El 50% de los clientes compran entre 1.024 dólares y 13.387 dólares y, en promedio, un cliente compra 14.235 dólares al año (menos de 1.800 litros de fertilizante al año).

Podemos ver en el histograma de la variable Facturacion una fuerte asimetría a derecha, indicando una concentración de clientes comprando por montos bastante menores al promedio. El gráfico *boxplot* confirma también la asimetría de la muestra con varios *outliers* por encima del límite superior. Son los clientes del percentil 75% hacia arriba, el 25% que más compra, quienes mueven fuertemente hacia arriba el promedio de facturación por cliente.

De acuerdo con estadísticas[2], un productor pequeño consume aproximadamente 1.200 litros de fertilizantes al año y uno mediano alrededor de 2.500 litros. Por lo tanto, una parte significativa de los clientes no estarían comprando toda su necesidad de fertilizante a La Empresa – La Empresa ofrece y se destaca en el mercado por su amplia gama de productos, la cual fue diseñada estratégicamente para cubrir toda la necesidad de los clientes punta a punta.

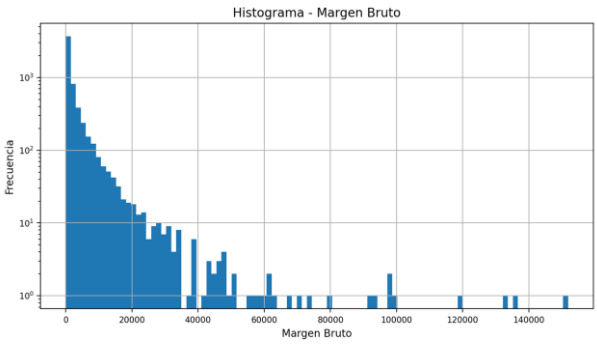
[2]Bolsa de Cereales de Buenos Aires.

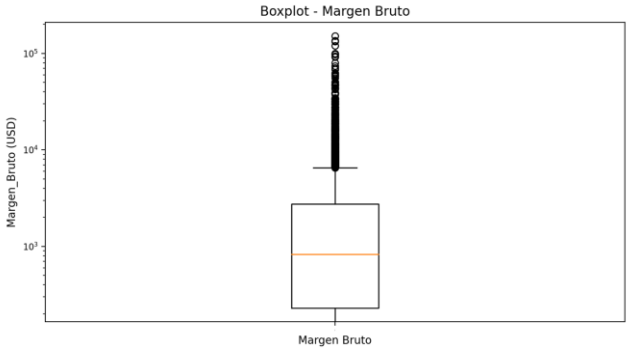




La variable **Margen_Bruto** presenta un rango amplio, entre 1 y 151.862 dólares, con una su media de 2.988. El 50% de los clientes aportan un margen bruto menor 823 dólares al año, bastante menor al promedio, lo cual anticipa una fuerte asimetría de la muestra y que, nuevamente, vemos un perfil parecido a la facturación: el 25% de clientes que aportan los márgenes mayores son los que mueven fuertemente el promedio del margen hacia arriba.

A partir de estos gráficos, histograma y boxplot, se confirma la fuerte asimetría a derecha de la variable Margen_Bruto y que el promedio se ve impactado por una minoría de clientes que aportan los márgenes brutos mayores.

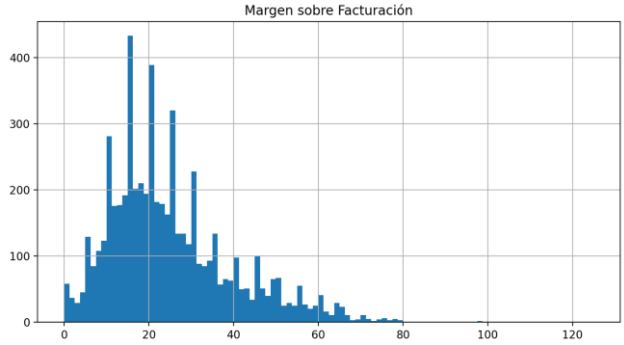


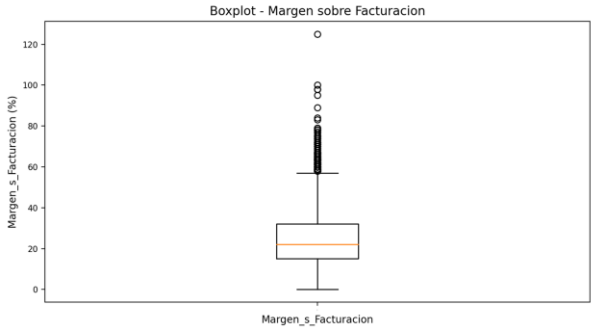


Respecto a la variable **Margen_s_Facturacion** (rentabilidad), el 50% de los clientes muestran márgenes sobre facturación entre el 15% y 32%. El porcentaje promedio es 25% y 22% el valor central (mediana).

De todas las variables, Margen_s_Facturacion es la más simétrica y con menor dispersión. Se identificó un *outlier* con valor 125%.

En el histograma y boxplot de la variable Margen_s_Facturacion se aprecia la simetría.



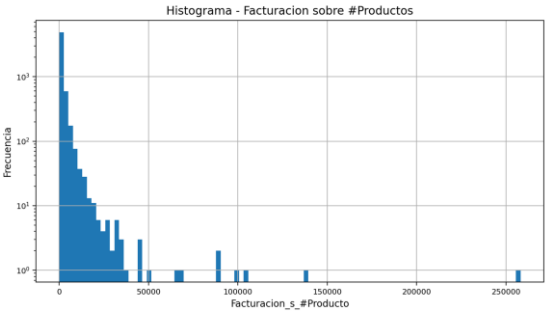


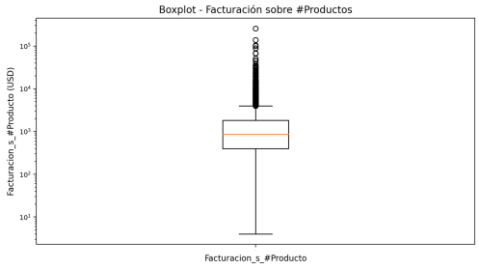
En particular, la variable margen sobre facturación es el cociente de otras dos variables: margen_bruto y facturacion. Esta relación es de importancia y relevancia para el negocio porque habla de cuán "rentable" es un cliente para la compañía.

Tener segmentos explicados por rentabilidad permitirá a la compañía enfocar su atención y aplicar tiempo y recursos en acciones que permitan mantener y también maximizar la cantidad de clientes en estos segmentos.

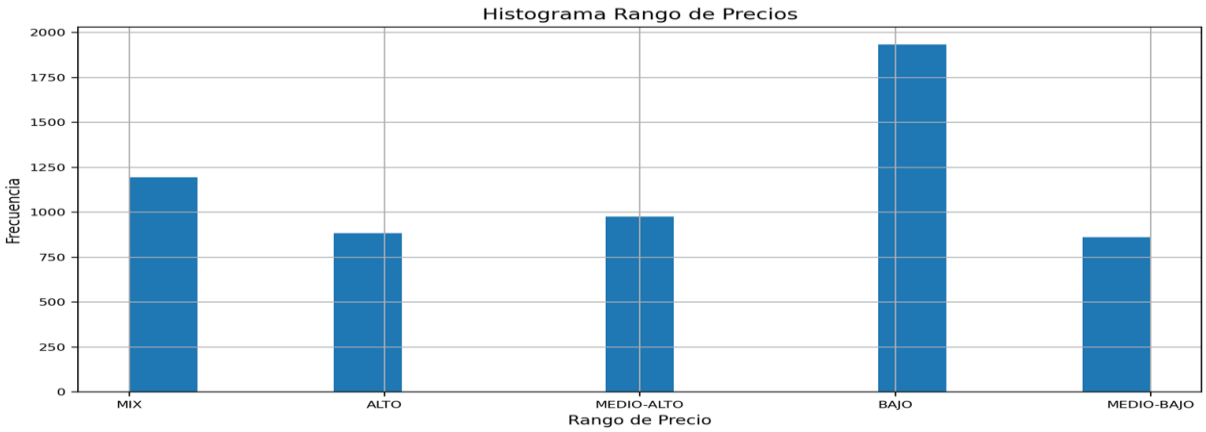
Respecto a la variable **Facturacion_s_#Producto**, el 50% de los clientes compran por producto entre 400 y 1.840 dólares al año, lo equivalente a 50 litros y 250 litros aproximadamente. Hay un 25% de los clientes que compra menos de 400 dólares por producto por año (50 litros). El promedio es 1.908 dólares (250 litros aprox.) y 860 dólares el valor central (100 litros aprox.).

A partir de estos gráficos, histograma y boxplot, se confirma la fuerte asimetría a derecha de la variable Facturacion_s_#Producto y que el promedio se ve impactado por una minoría de clientes empujan hacia arriba el promedio.

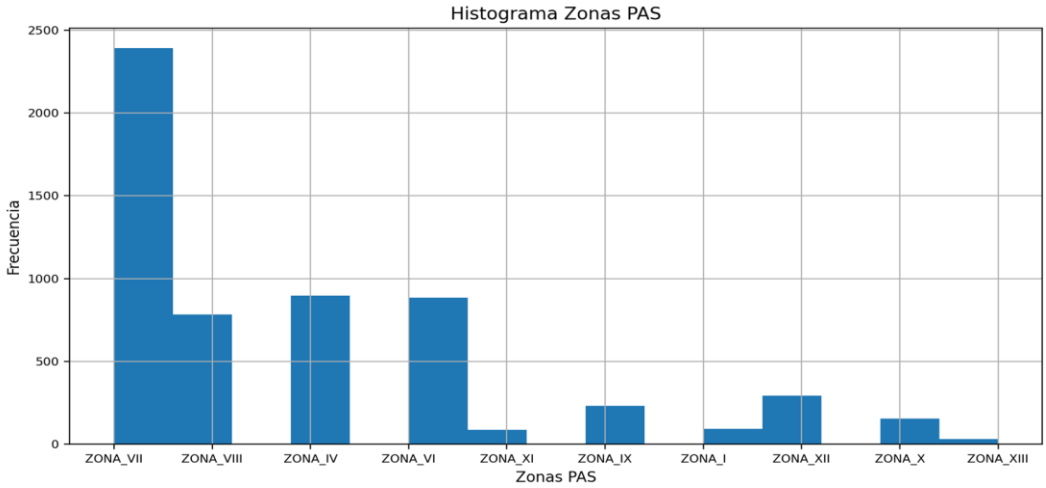




Respecto a la variable categórica **rango_precio**, el rango de precio "BAJO" es el más significativo, con el 30% de los clientes aprox. El resto se mantienen en valores similares, con el 15% de los clientes aprox.

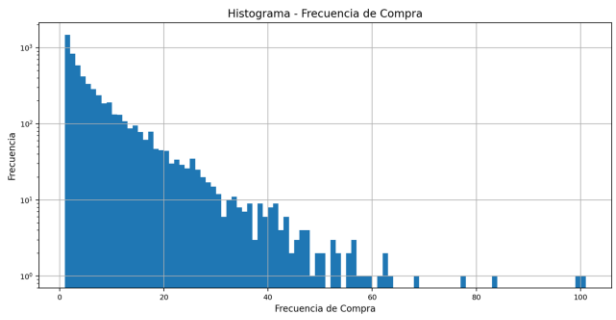


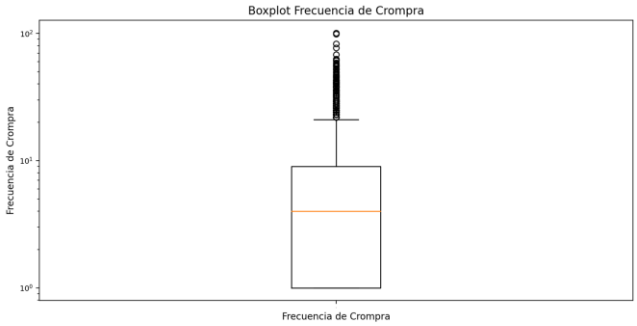
Respecto a la variable categórica **zonas_pas**, la zona "ZONA_VII" es el más significativa, con cerca de 2.500 clientes (el 40% de los clientes). Las zonas "ZONA_VIII", "ZONA_IV" y "ZONA_VI" se llevan el otro 40% y el resto de las zonas el 20% restante de los clientes.



En cuanto a la variable **Frecuencia_Compra**, la muestra presenta un rango amplio entre 1 y 101 contactos de compra en 2020. El 50% de los clientes tuvieron una frecuencia de compra menor a 4 compras por año y el 75% menos de 9 en el año. Por otro lado, el promedio está en 7 contactos en el año.

Podemos ver en el histograma una fuerte asimetría a derecha, indicando una concentración de clientes comprando en pocos momentos al año. El gráfico *boxplot* confirma también la asimetría de la muestra con varios *outliers* por encima del límite superior.



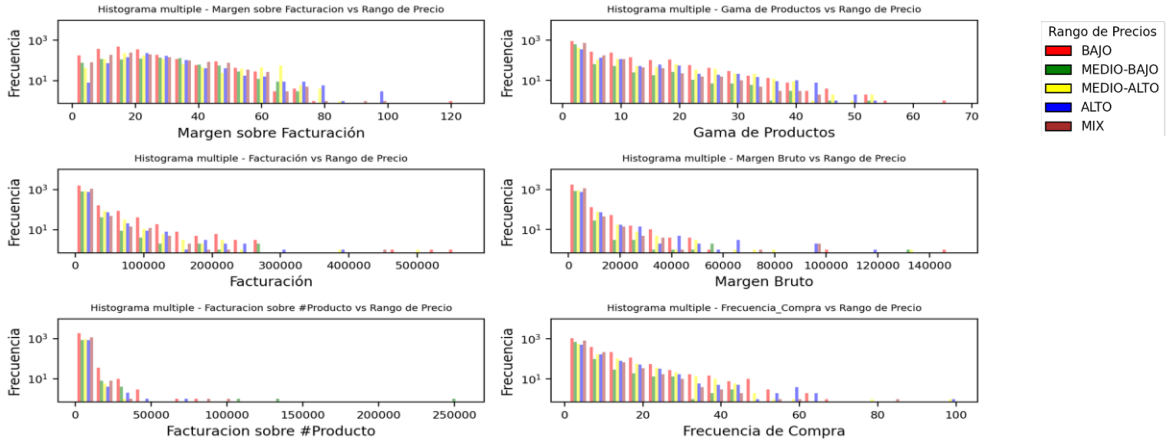


Comportamiento de los datos de acuerdo con las variables categóricas

Comenzaremos describiendo el comportamiento de las variables numéricas de acuerdo con la variable categórica **rango_precio**.

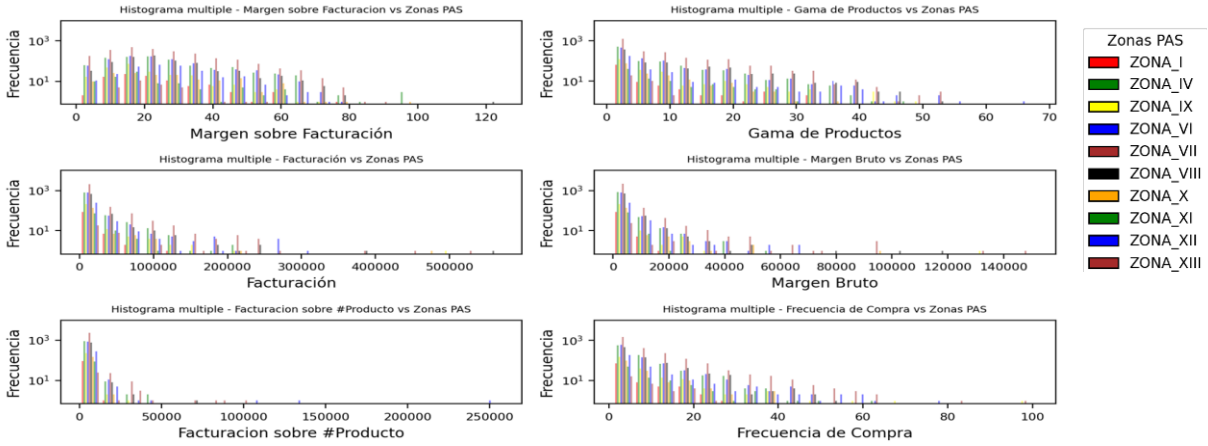
Los siguientes cuadros muestran las características de posición de todas las variables numéricas en todos los rangos de precio.

En general, la posición de todas las variables numéricas es similar para todos los valores de la variable categórica rango_precio. Así se observa en el siguiente gráfico.



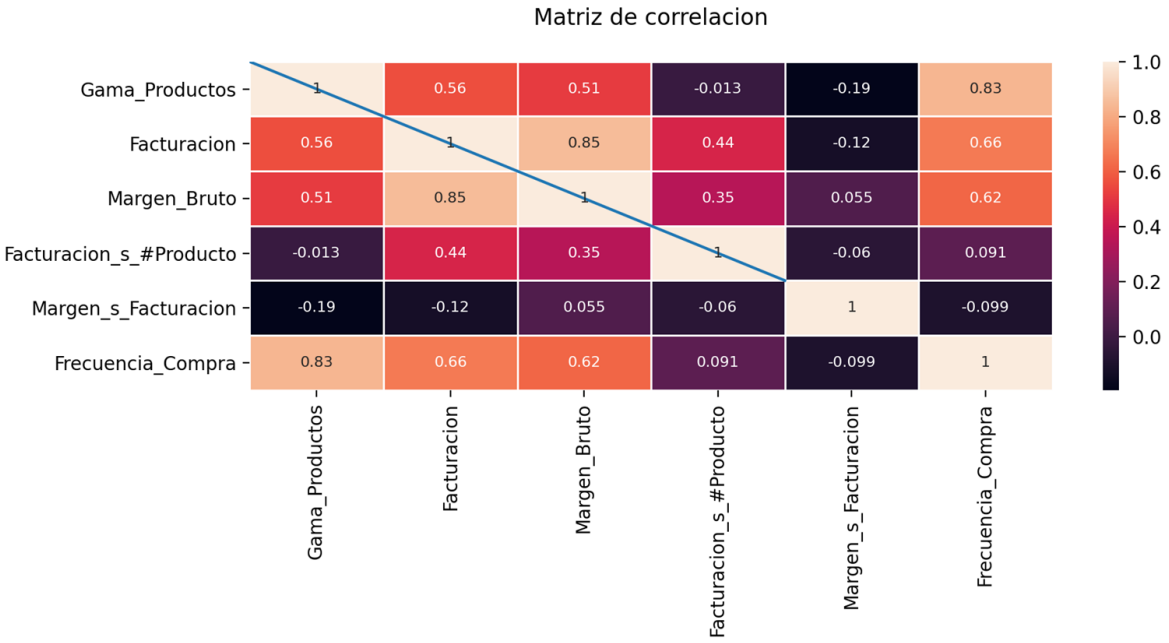
Respecto a variable categórica **ZONA_PAS**, los siguientes cuadros muestran las características de posición de todas las variables numéricas en todas las zonas geográficas.

En general, la posición de todas las variables numéricas es similar para todos los valores de la variable categórica ZONA_PAS. Así se observa en el siguiente gráfico.



Correlación entre las variables

Para analizar la relación entre las variables cuantitativas aplicaré la matriz de correlación – *si la correlación entre variables es menor a 0.2 entonces su correlación es despreciable o muy baja y, por lo tanto, son variables independientes* – también utilizaré el gráfico Plot Pairs para apoyar visualmente este análisis.



De acuerdo con los valores de correlación entre las variables, podremos decir que hay **independencia** – o una correlación despreciable o muy baja – entre las siguientes variables:

Margen sobre Facturación (rentabilidad) y

i. Gama de Productos
ii. Facturación
iii. Margen Bruto
iv. Facturación sobre # productos
v. Frecuencia de Compra

Facturación sobre # productos y

i. Gama de Productos

En el otro extremo, se observa una alta correlación entre las siguientes variables:

Frecuencia de Compra y

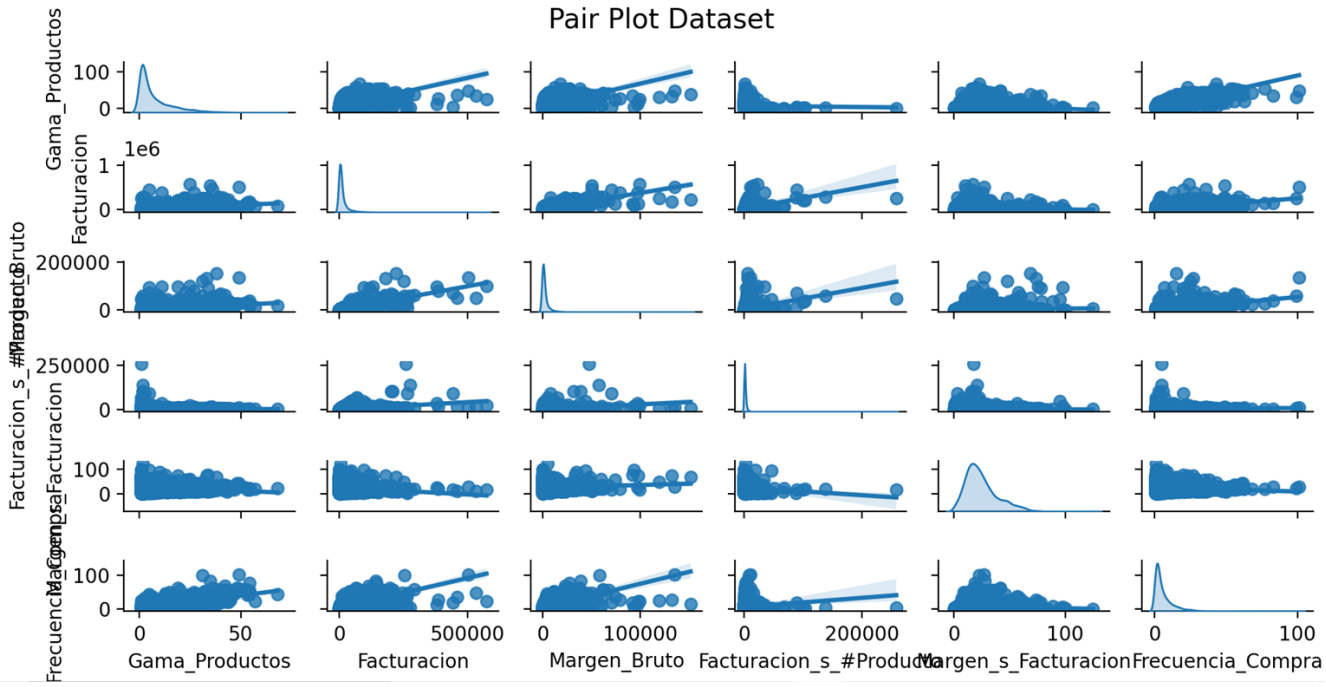
i. Gama de Productos
ii. Facturación
iii. Margen Bruto

Facturación

i. Margen Bruto

Por lo tanto, a priori, la variable margen sobre facturación aportará variabilidad al Dataset, lo cual favorecerá la calidad del clustering.

En el gráfico *Plot Pairs*, observamos la buena correlación de la variable Frecuencia de compra con Gama de Productos, Facturación y Margen Bruto (última fila). Por otro lado, se observa la baja correlación de la variable Margen sobre Facturación respecto al resto de las variables (fila 5).



5.2 Resumen de lo observado

a) la mayoría de las variables numéricas son fuertemente asimétricas – excepto Margen_s_Facturacion.

a.1. Gama de Productos: la gran mayoría de los clientes compran muy pocos productos de la gama ofrecida,

a.2. Margen Bruto: el 25% de clientes con los márgenes mayores son los que mueven fuertemente hacia arriba el margen promedio

a.3. Facturación: se cumple la regla del 80/20 – una gran parte de la facturación se genera gracias una porción pequeña de clientes,

a.4. Frecuencia de compra: el 75% de los clientes tuvieron una frecuencia de compra menor a 9 contactos en el año

b) la zona "ZONA_VII" es el más significativa, con cerca de 2.500 clientes (40% de todos los clientes)

c) el rango de precio "BAJO" es el más significativo, con el 30% de los clientes aprox.

d) la posición de todas las variables numéricas es similar para todos los valores de las variables categóricas.

e) La variable Margen sobre Facturación (rentabilidad) tiene una baja correlación con la mayoría de las variables

f) Se observa una fuerte correlación entre las variables Frecuencia de Compra, Gama de Productos, Facturación y Margen Bruto

6. Estrategias de Trabajo

Para poder identificar el mejor enfoque de segmentación sobre el *Dataset* se decidió establecer estrategias de trabajo que nos permitan:

- *aplicar ordenadamente los distintos algoritmos de segmentación de acuerdo al tipo de variable: cuantitativas o categóricas*
- *analizar los resultados obtenidos en cada estrategia*
- *comparar los resultados entre estrategias*
- *seleccionar el mejor enfoque o estrategia de clustering*

En este contexto, siguen las estrategias de trabajo definidas:

1. **Clustering sobre variables cuantitativas – análisis visual**

   Objetivo: trabajar con componentes principales para analizar visualmente los datos, generar clusters (con K-Means) y evaluar con estadísticos la tendencia de los datos al clustering. Luego, realizar clustering sobre las 6 variables cuantitativas y finalmente comparar resultados.

2. **Clustering sobre variables cuantitativas – aplicación de varios algoritmos**

   Objetivo: aplicar distintos algoritmos para entender los clusters resultantes, comparar dichos algoritmos de clustering a través de métricas y finalmente identificar el mejor algoritmo para esta estrategia

3. **Clustering sobre variables categóricas – Kmodes & MCA**

   Objetivo: aplicar distintos enfoques para resolver el clustering sobre variables categóricas e identificar el mejor enfoque para esta estrategia

4. **Clustering sobre todo el Dataset – 2 etapas & aplicación de varios algoritmos**

   Objetivo: aplicar distintos algoritmos sobre variables categóricas (previo proceso MCA) y variables numéricas, comparar algoritmos de clustering a través de métricas y finalmente identificar el mejor algoritmo para esta estrategia

5. **Clustering sobre todo el Datset – 1 etapa & aplicación de varios algoritmos**

   Objetivo: aplicar distintos algoritmos sobre variables categóricas y numéricas procesadas en conjunto con FAMD, comparar algoritmos de clustering a través de métricas y finalmente identificar el mejor algoritmo para la estrategia

7. Resultados

A continuación, evaluaremos cada una de las estrategias de trabajo mencionadas en el capítulo anterior.

Para ello, en cada estrategia analizaremos los siguientes aspectos:

- tendencia de los datos al clustering
- inspección visual sobre los datos
- identificación de clusters
- aplicación de algoritmos
- evaluación del desempeño de cada algoritmo

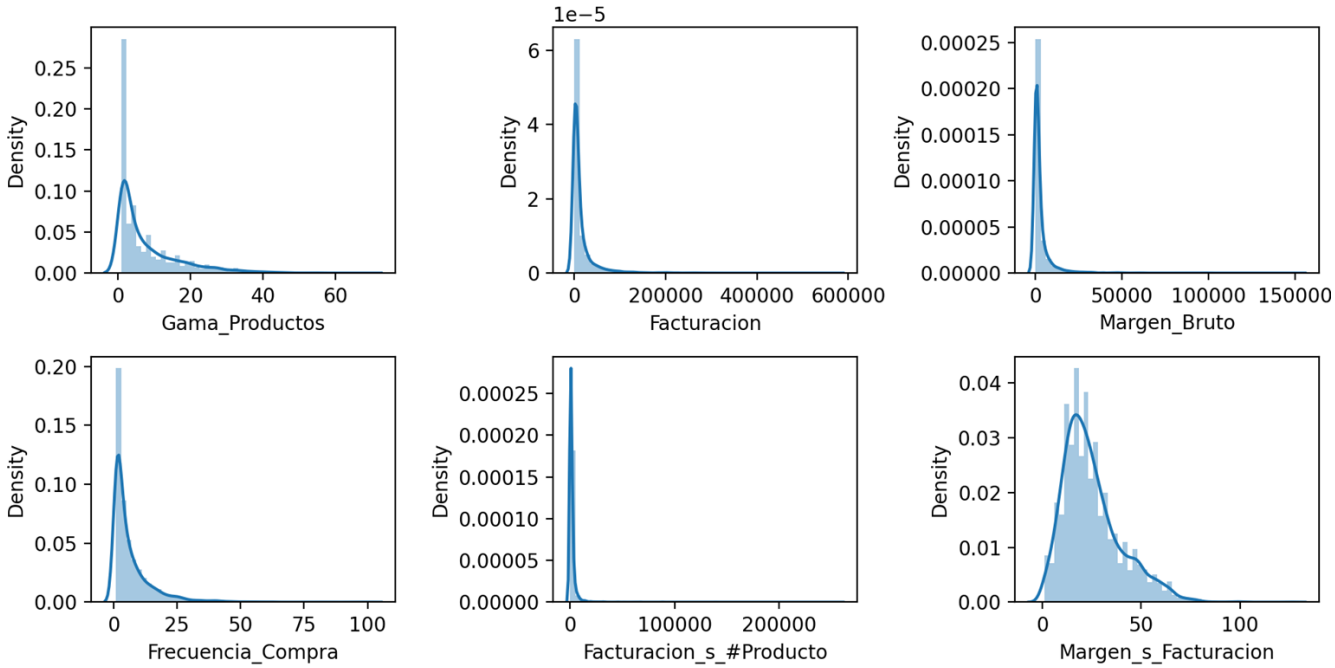
7.1. Clustering sobre variables cuantitativas – análisis visual

El objetivo de esta primera estrategia es trabajar con componentes principales para analizar visualmente los datos, generar clusters (K-Means) y evaluar con estadísticos la tendencia de los datos al clustering. Luego, realizar clustering sobre las 6 variables cuantitativas y finalmente comparar resultados.

Como primer paso, antes de avanzar con el análisis visual, es necesario **gestionar la asimetría y normalización** de las 6 variables cuantitativas.

Gestión de asimetría

Situación inicial: datos asimétricos.



Para que los datos tengan una forma más simétrica (y favorezcan los procesos de clustering) se les aplicarán transformaciones. Para ello, hay distintos métodos que podemos utilizar:

a. transformación log

b. transformación de raíz cuadrada

c. transformación box-cox

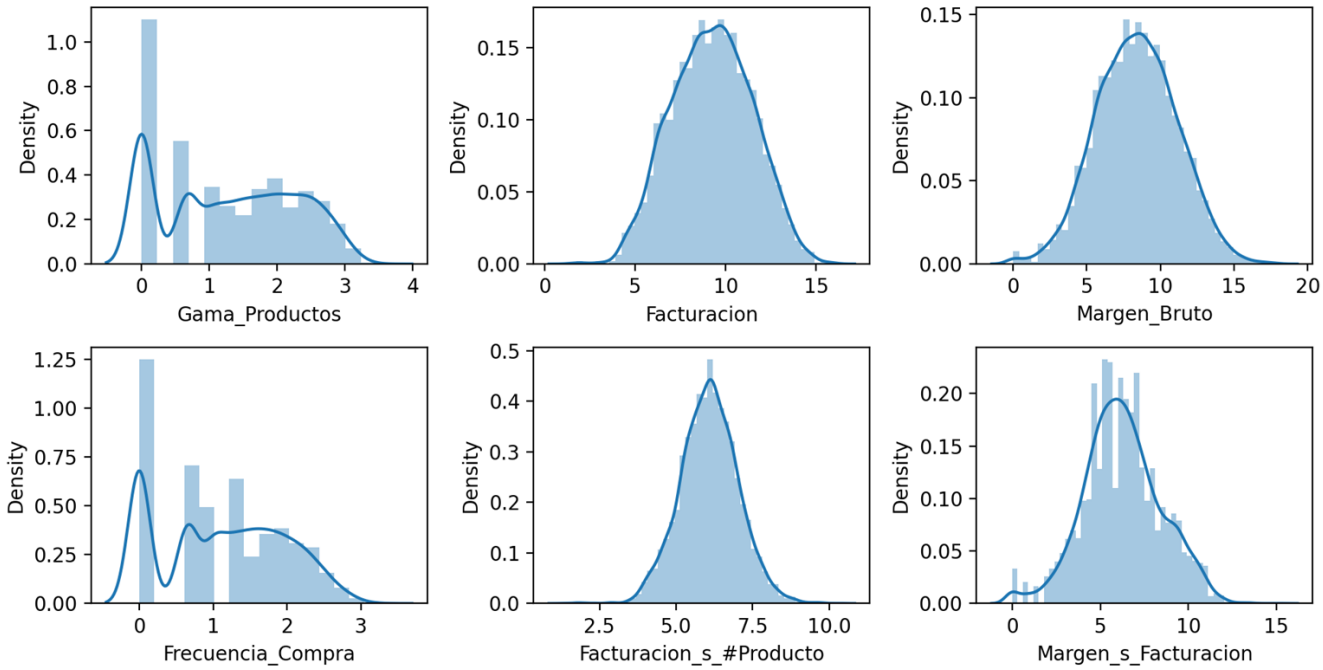
Se calculó el valor de asimetría "Skewness" para los distintos métodos sobre las variables cuantitativas. Si el valor de asimetría se acerca a 0, la variable tiende a tener forma simétrica con ese método y, por lo tanto, se aplica el método.

Resultados de test de asimetrías

Variables cuantitativas	*No transformacion*	*Transformacion Log*	*Transformacion Raiz cuadrada*	*Transformacion Box-Cox*
Gama_Productos	1.81	0.17	0.91	+0.05
Facturacion	6.63	-0.12	2.18	-0.01
Margen_Bruto	8.48	-0.35	2.5	-0.00
Facturacion_s_#Producto	23.28	0.13	4.21	-0.01
Margen_s_Facturacion	1.07	-1.17	0.21	+0.02
Frecuencia_Compra	3.03	0.31	1.33	+0.07

Por lo tanto, aplicaremos el método Box-Cox a todas las variables.

Datos simétricos



Gestión normalización (Estandarización)

Se ejecuta en Python el siguiente código sobre las variables cuantitativas a normalizar:

```
scaler = StandardScaler()
scaler.fit(datafile)
datafile_normalized = scaler.transform(datafile)
```

Dataset simétrico y normalizado

#cliente	rango_precio	ZONA_PAS	Gama_Productos	Facturacion	Margen_Bruto	Facturacion_s_#Producto	Margen_s_Facturacion	Frecuencia_Compra
C_1	MIX	ZONA_VII	-0.255743	-0.289148	-0.666508	-0.090234	-1.001484	-1.332928
C_10	ALTO	ZONA_VIII	-0.255743	-0.199929	-0.731850	0.046116	-1.361575	-1.332928
C_100	MEDIO-ALTO	ZONA_VII	1.435807	2.075223	2.892588	1.545367	1.860283	1.869684
C_1000	BAJO	ZONA_IV	1.205319	1.562803	1.585469	1.115388	-0.110198	1.442772
C_1001	ALTO	ZONA_VIII	1.205319	0.695475	0.667641	-0.100325	-0.110198	1.092391
...	...	...	...	...	...	...	...	...
C_995	MEDIO-ALTO	ZONA_VII	-0.255743	-0.465891	-0.410452	-0.363525	0.104838	0.347572
C_996	MIX	ZONA_VIII	1.582221	0.709067	0.745429	-0.510249	0.035032	1.360361
C_997	MIX	ZONA_VI	-0.634053	1.185097	1.616446	2.323935	1.094455	1.265815
C_998	BAJO	ZONA_VI	0.784073	0.482037	-0.099980	0.040426	-1.361575	0.760638
C_999	BAJO	ZONA_VI	0.600374	0.053188	0.085635	-0.409640	0.035032	0.347572

Luego, **se evaluará la tendencia de los datos al clustering**. Para ello, utilizaremos 2 métodos:

a) Métodos estadísticos: estadístico **Hopkins**

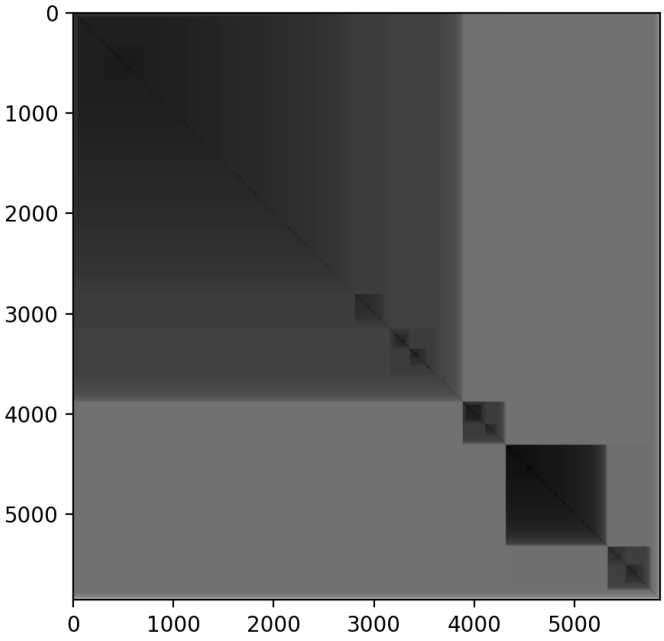
b) Métodos visuales: algoritmo **VAT** (*Visual Assessment of cluster Tendency*)

Valor Hopkins obtenido: (H) = 0.061

La literatura es precisa respecto al valor esperado de H. Por ejemplo, Lawson & Jurs (1990) y Banerjee & Dave (2004) explican que se pueden esperar 3 resultados diferentes en el valor del estadístico Hopkins (H-1):
1) H = 0.5, los datos muestran que no poseen una estructura clusterizable (los datos son aleatorios).
2) H cercano a 1.0, hay evidencia significativa que los datos poseen una estructura ordenada / cluterizable.
3) H cercano a 0, en este caso el test es indeciso (sin embargo, los datos no son aleatorios)

Por lo tanto, el resultado del test es: indeciso (sin embargo, los datos no son aleatorios).

VAT (*Visual Assessment of cluster Tendency*)



El gráfico muestra cuadrados negros que representan clusters. En este caso, podemos observar 4 cuadrados pricipales, indicando la presencia de 4 grupos o clusters en los datos.

Antes de aplicar un método de *Clustering* a los datos es conveniente evaluar si hay indicios de que realmente existe algún tipo de agrupación en dichos datos. Por lo tanto, se realizará una **inspección visual** 2D sobre los datos.

Para poder graficar resultados en 2 D necesitamos reducir las dimensiones de los datos, de 5 a 2. Para ello realizaremos el proceso **PCA – Análisis de Componentes Principales**.

Variación Explicada por cada componente principal (son 6 componentes principales): [6.35244229e-01 1.82003879e-01 1.53795330e-01 2.74386679e-02 1.30313843e-03 2.14756963e-04]

La variación explicada por los primeros 2 componentes principales es de aproximadamente **82%**.

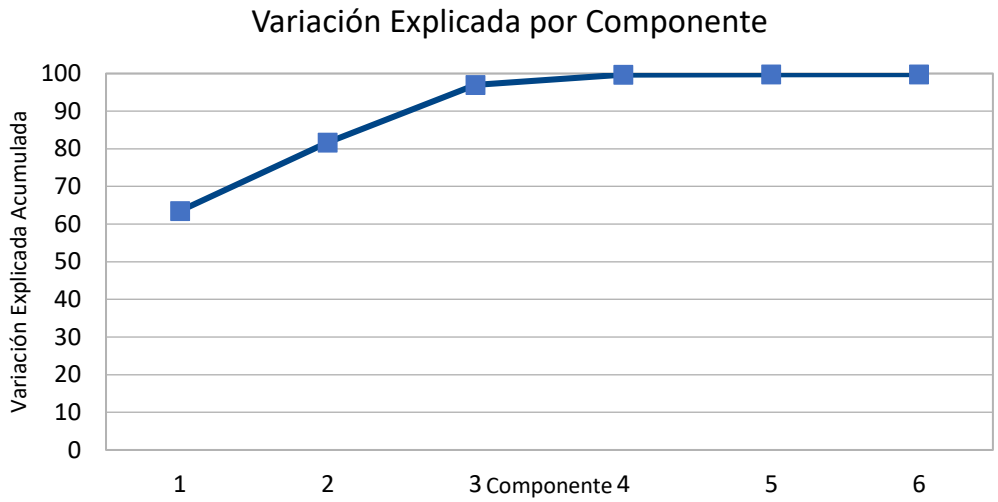
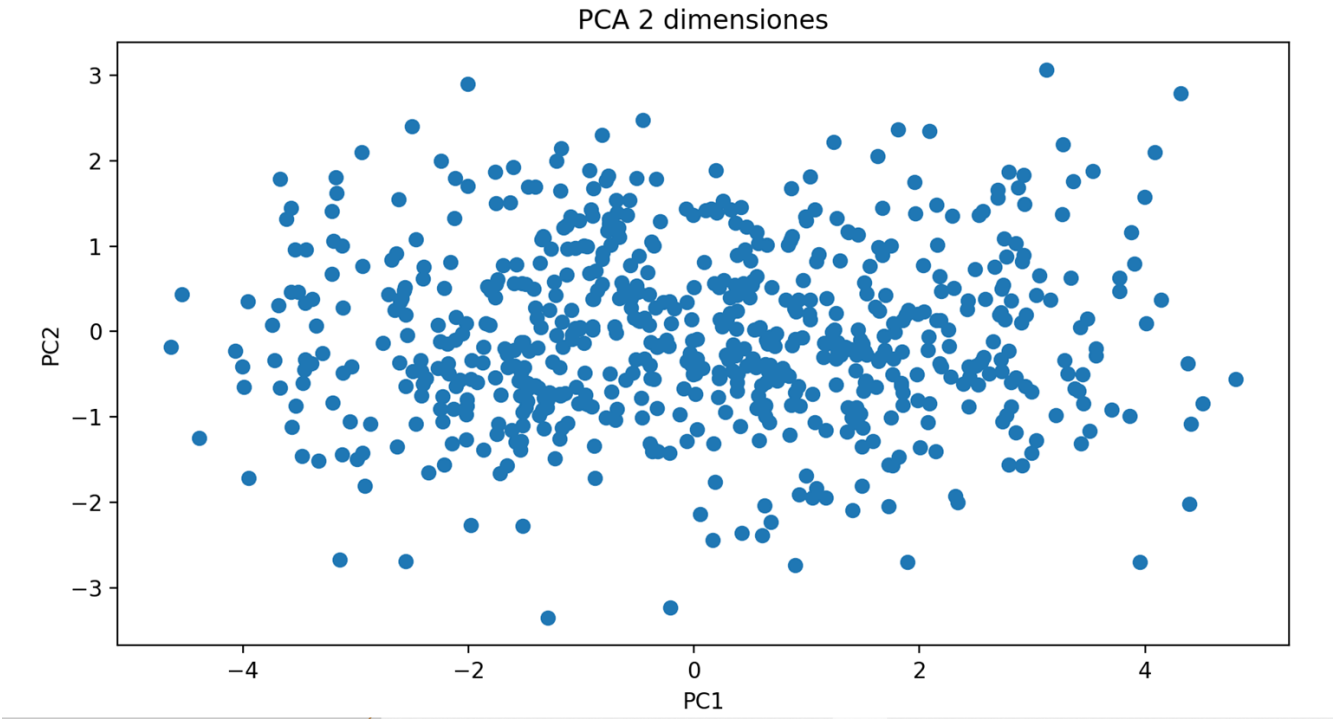


Gráfico PCA sobre componentes principales PC1 y PC2 – cada punto es un cliente (muestra reducida)



Del gráfico se observa que no se trata de puntos concentrados en un sector. Seguramente, los algoritmos de clustering nos ayudarán a identificar los segmentos que en una primera instancia no se identifican claramente.

A continuación, sigue información aportada por el análisis PCA

Matriz de rotación de las componentes principales

	PC 1	PC 2	PC 3	PC 4	PC 5	PC 6
Gama_Productos	-0.416489	-0.207844	-0.527356	0.450744	0.430752	-0.341345
Facturacion	-0.503903	-0.054804	0.149753	0.210492	0.096606	0.816708
Margen_Bruto	-0.485535	0.280104	0.042025	0.243044	-0.745538	-0.262934
Facturacion_s_#Producto	-0.356575	0.111998	0.735098	-0.096220	0.415430	-0.371619
Margen_s_Facturacion	0.062606	0.928834	-0.213341	0.044526	0.276884	0.095847
Frecuencia_Compra	-0.453661	-0.007945	-0.334395	-0.825961	0.006842	-0.007055

Variables cuantitativas (6)

X1: Gama_Productos

X2: Facturacion

X3: Margen Bruto

X4: Facturacion_s_#Productos

X5: Margen_s_Facturacion

X6: Frecuencia_Compra

Interpretación del PCA por matriz de rotación: variables de componentes principales PC1 y PC2 en función de las variables cuantitativas

PC1= – (0.503903 * **X2** + 0.485535 * **X3** + 0.453661 * **X6** + 0.416489 * **X1** + 0.356575 * X4) + (0.062606 *X5)

→ Entonces la componente PC1 diferencia a los clientes que destacan por **Facturación**, **Margen Bruto**, **Frecuencia de Compra** y **Gama de Productos** del resto.

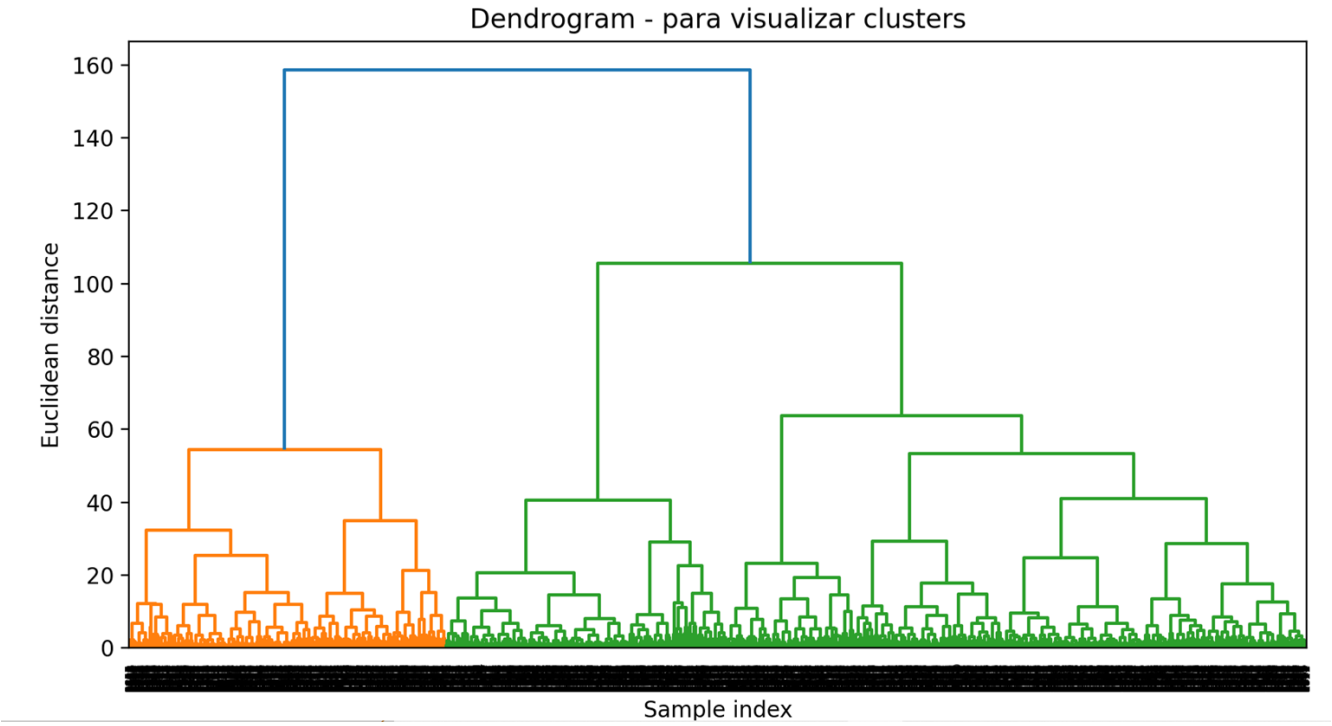
PC2 = + (0.928834 * **X5** + 0.280104 * X3 + 0.111998 * X4) – (0.207844 * X1 + 0.054804 * X2 + 0.007945 * X6)

→ Por otro lado, la componente PC2 diferencia a los clientes que destacan por **Margen sobre Facturación** (rentabilidad) del resto.

Una vez reducida la cantidad de dimensiones a través de PCA e interpretado sus resultados, se analizarán los datos para tratar de **identificar clusters valiosos**. Para ello, se utilizará el algoritmo **clustering jerárquico** y **K-Means**.

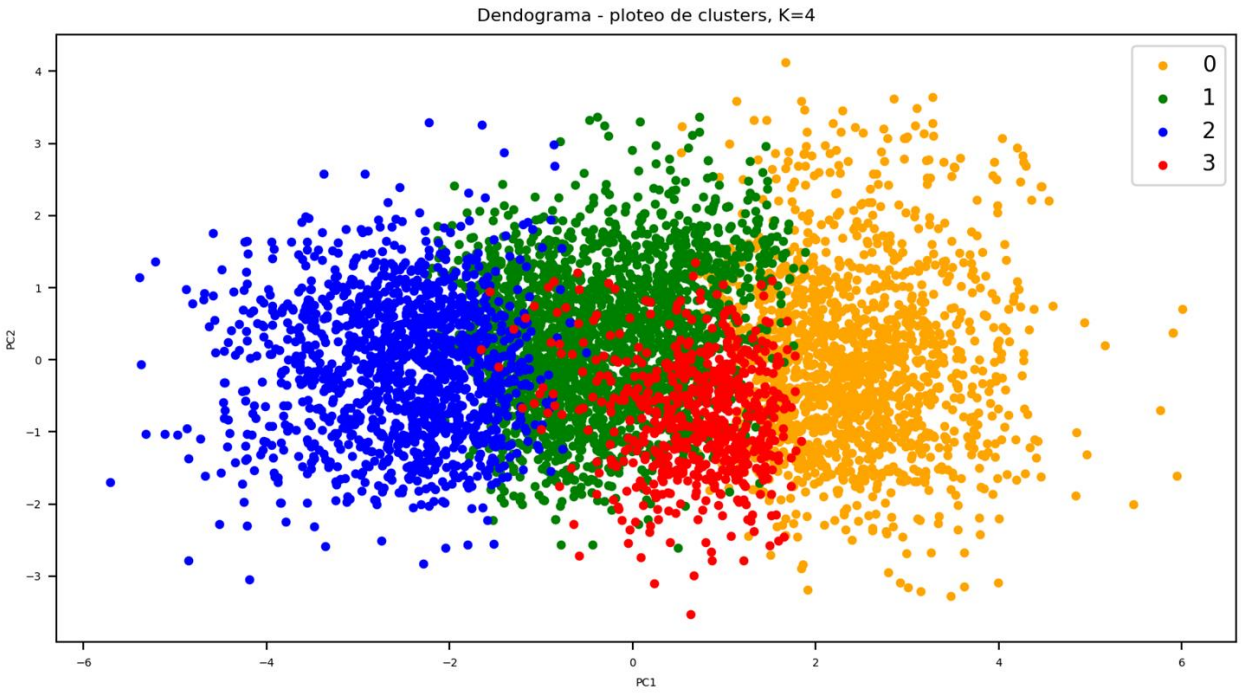
Clustering jerárquico

Dendograma



El **Dendograma** también muestra una clusterización de los datos. De acuerdo con la distancia Eucleidiana, eje vertical, la cantidad de clusters para este Dataset estaría definida entre 3 y 4.

Se graficarán los 4 clusters sugeridos por el dendograma sobre los ejes de las 2 componentes principales.

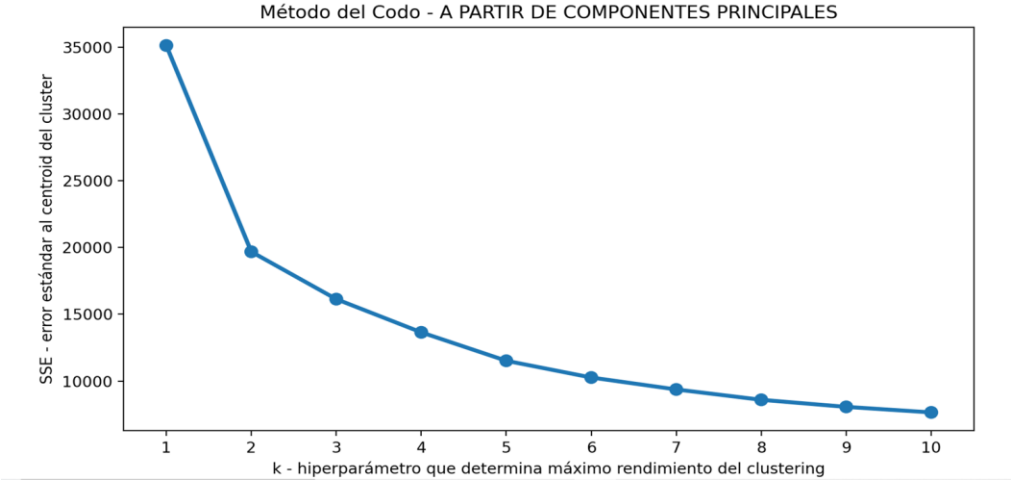


Del gráfico se puede observar una primera propuesta de segmentos o clusters, aunque la separación entre ellos no sea del todo clara. No obstante, se vuelve a confirmar visualmente la tendencia de los datos al clustering.

Seguiremos el proceso de clustering o segmentación con otro algo más complejo: K-Means

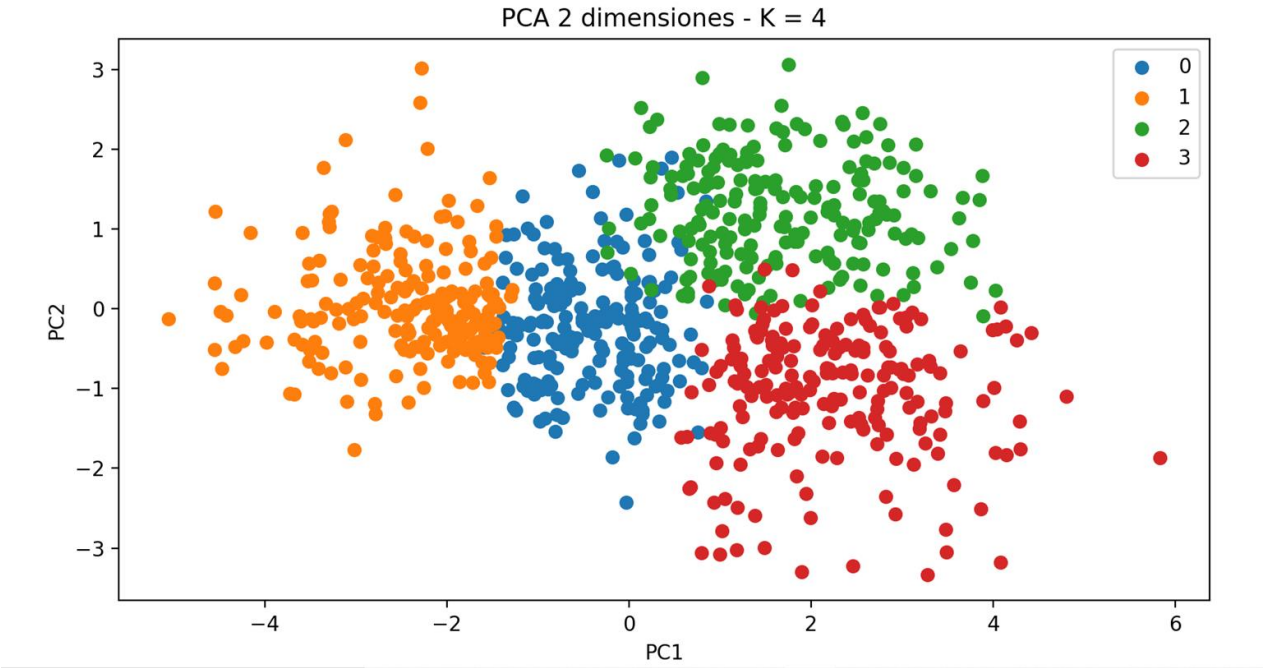
Antes de correr el algoritmo de clustering K-Means estimemos a través del método del codo (por medio del error estándar al centro del clúster - SSE) la cantidad "ideal" de clúster.

Gráfico del Codo



Según el gráfico de codo, a partir del 4to cluster ya no se obtienen saltos del SSE significativos, por lo tanto, **la cantidad de clusters a probar será 4**.

Clusters a partir de **K-Means** sobre las componentes principales graficados en ejes PCA



A continuación, sigue una tabla con las principales características de los 4 clusters generados por K-Means.

Clusters K-Means	Cantidad de Clientes	Cantidad de Clientes (%)	% Facturación	Facturación Promedio	Rentabilidad Promedio (%)	Frecuencia de Compra (prom.)	Gama de Productos (prom.)
0	2.090	36 %	17 %	6.664	21 %	5	7
1	1.442	25 %	80 %	46.467	22 %	17	18
2	1.330	23 %	2 %	1.112	43 %	2	2
3	991	17 %	1 %	955	13 %	1	2
	5.853						

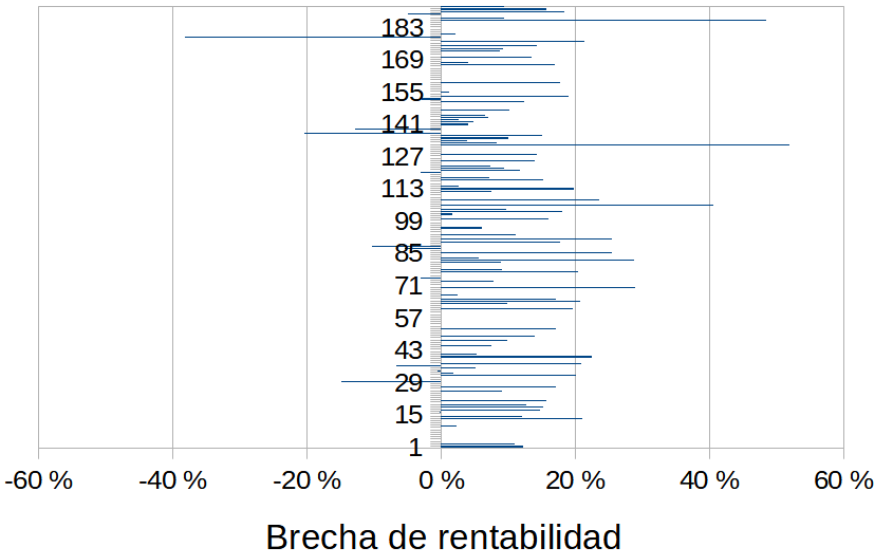
Desde el punto de vista del dominio, surgen agrupaciones interesantes, especialmente considerando que la estrategia comercial actual de La Empresa es mantener un trato y esfuerzo de venta similar con todos los clientes.

Analizaremos los clusters propuestos por K-Means en función de las variables cuantitativas utilizadas.

Cluster 0 – es el cluster con mayor **oportunidad** de negocio por su gran cantidad de clientes (el 36% de la cartera). La venta promedio de los clientes en este cluster equivale a 800 litros de fertilizante al año, una cantidad significativa para un productor mediano o chico – por lo tanto, hay presencia de la marca en este segmento. La Empresa debería lograr un entendimiento más profundo de este segmento que permita el desarrollo de acciones comerciales que saquen provecho de la presencia de la marca en los clientes, buscando un impacto positivo en gama y rentabilidad (*up-selling*).

Cluster 1 – es el cluster con mayor **facturación**, con el 80% de la venta de La Empresa. Se observa la correlación positiva entre facturación, la gama de productos y la frecuencia de compra. Dentro de este segmento hay clientes que compran grandes cantidades, es factible pensar para este segmento en otro tipo de abordaje comercial, por ejemplo: la figura del *Key Account*.

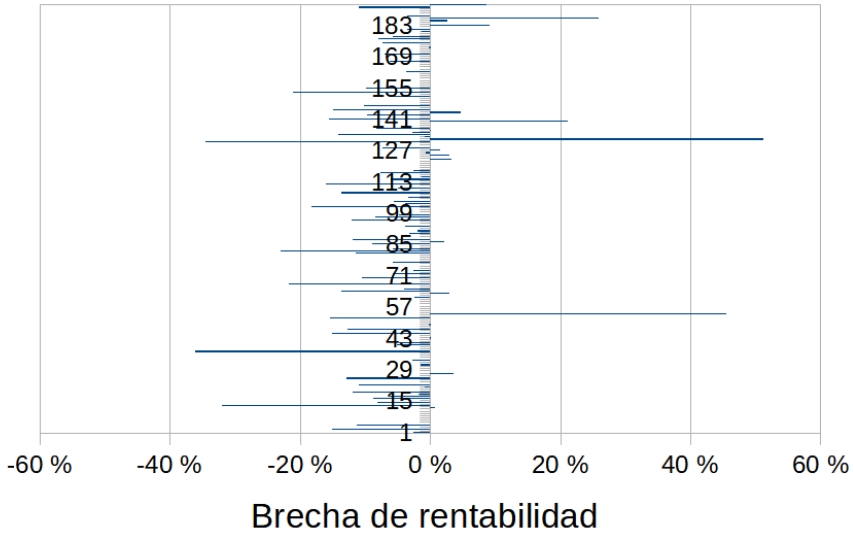
Cluster 2 – es el cluster con mayor **rentabilidad**, 43% en promedio, casi el doble respecto a los otros segmentos.
El siguiente gráfico muestra una comparación de rentabilidad entre clientes del cluster 1 respecto a clientes del cluster 2 sobre mismos productos. Las barras azules horizontales representan la brecha de rentabilidad entre segmentos.



Se observa que las brechas de rentabilidad son significativamente positivas en casi todos los productos a favor del cluster 2. Esto muestra el potencial del cluster 2 y abre la

posibilidad de pensar en invertir en crecimiento de marca a través de gama de productos y también crecer en servicio comercial con mayor frecuencia de compra (contacto).

Cluster 3 – es el cluster menos atractivo en lo comercial, merece una **revisión**. Se debe buscar un enfoque comercial diferente – por ejemplo, un nuevo canal que no consuma tantos recursos (fuerza de venta). La Empresa debe evaluar si el esfuerzo comercial que aplica para llegar a casi 1.000 clientes a través de su Fuerza de Venta (vendedores) justifica los valore de venta y rentabilidad de este segmento.
El siguiente gráfico muestra una comparación de rentabilidad entre clientes del cluster 1 respecto a clientes del cluster 3 sobre mismos productos. Las barras azules horizontales representan la brecha de rentabilidad entre segmentos.



Se observa que la rentabilidad del cluster 3 están por debajo en casi todos los productos. Lo cual muestra el contexto de poca atractividad para el negocio. El esfuerzo comercial debe ser revisado.

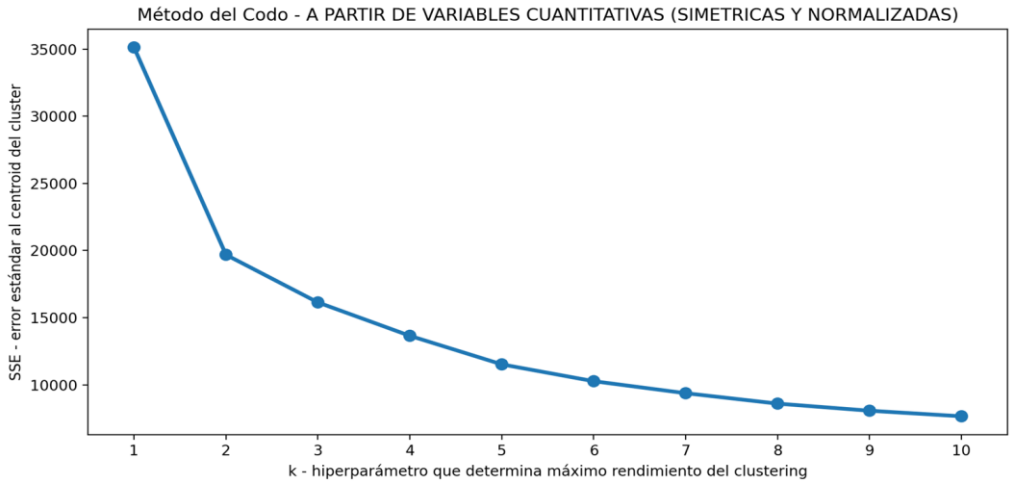
Primera conclusión hasta aquí, los clusters obtenidos por el algoritmo K-Means sobre las 2 componentes principales han aportado información valiosa para La Empresa.

Clustering sobre las 6 variables cuantitativas

Previamente ejecutamos K-Means sobre las 2 componentes principales CP1 y CP2. Ahora se ejecutará dicho algoritmo directamente sobre las 6 variables cuantitativas (sin PCA) y luego se analizarán los resultados obtenidos.

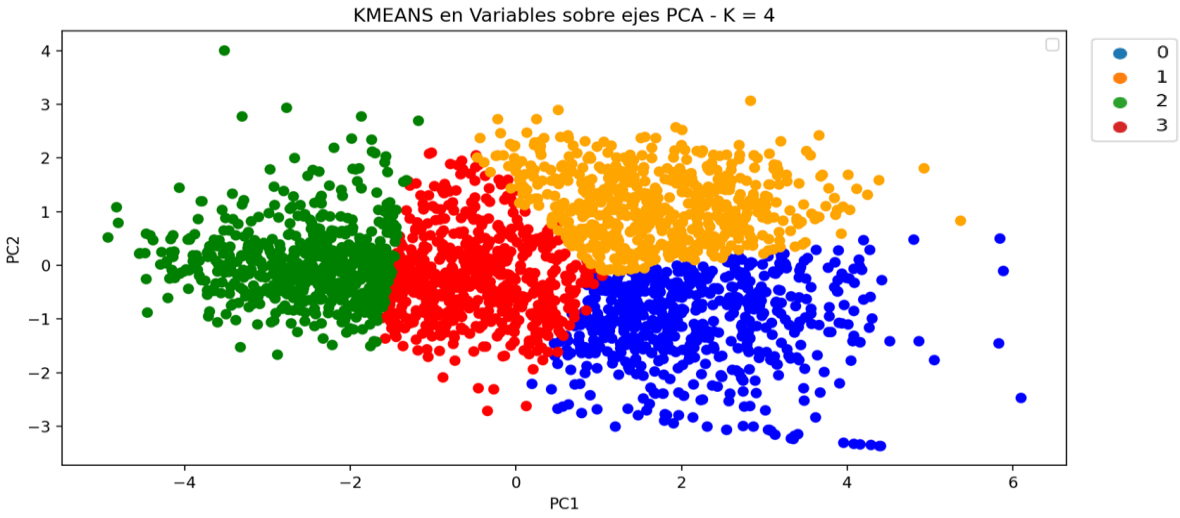
Nuevamente, antes de ejecutar el algoritmo de clustering K-Means se estimará la cantidad estimada de clúster a través del método del codo (por medio del error estándar al centro del clúster - SSE).

Gráfico del Codo



Según el gráfico de codo, a partir del 4to cluster ya no se obtienen saltos del SSE significativos, por lo tanto, **la cantidad de clusters a probar será 4**.

Clusters a partir de **K-Means** sobre variables cuantitativas graficados en ejes PCA



El gráfico muestra clusters parecidos a los obtenidos a partir de las 2 componentes principales. Más aún, comparado cliente por cliente, casi el 98% de los clientes coincidieron en los clusters generados por K-Means utilizando las 2 componentes principales. Este resultado es debido a la alta variabilidad explicada por las componentes principales 1 y 2 (82%).

7.2. Clustering sobre variables cuantitativas – aplicación de varios algoritmos

El objetivo de esta segunda estrategia es aplicar distintos algoritmos a las 6 variables numéricas del Dataset para entender los clusters resultantes, comparar dichos algoritmos a través de métricas y finalmente identificar el mejor algoritmo para esta estrategia.

Algoritmos que se utilizarán:

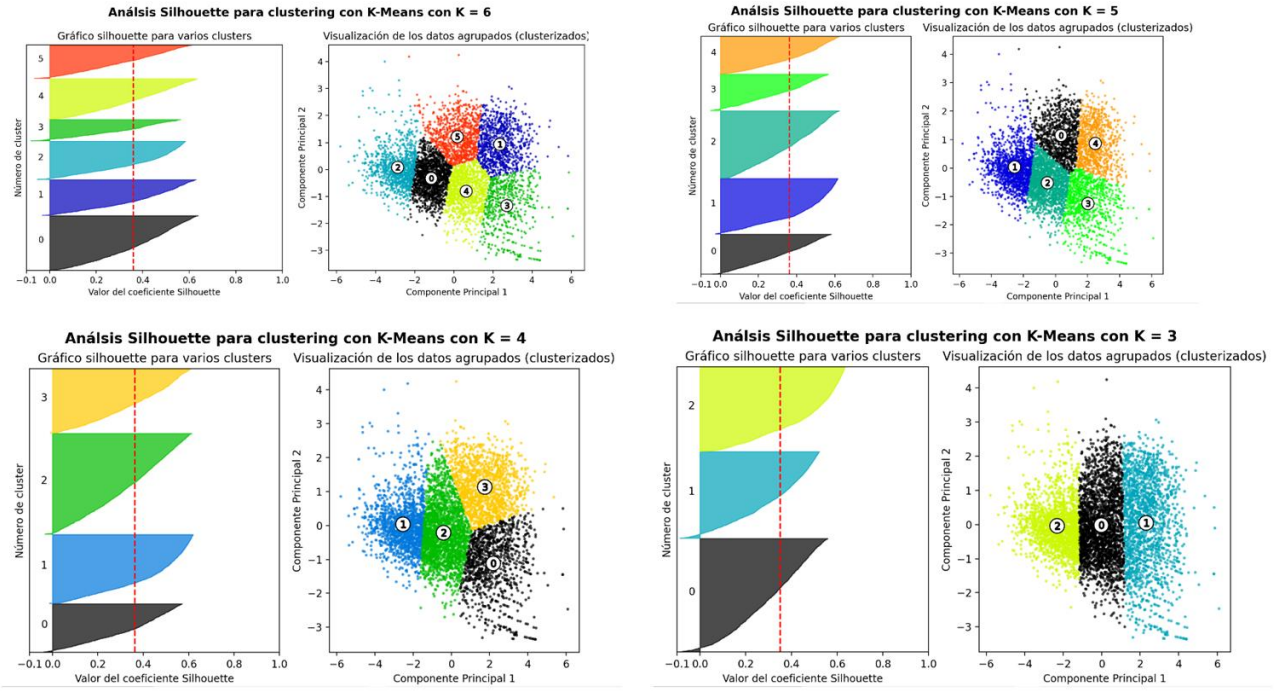
- *K-Means*
- *MiniBatchKMeans*
- GaussianMixture
- *AgglomerativeClustering*
- *SpectralClustering*
- *OPTICS*
- *MeanShift*
- *DBSCAN*

Para que un algoritmo de clustering alcance su máximo rendimiento, se debe determinar qué valor de K (hiperparámetro – cantidad de clusters) se ajusta mejor a los datos.

A continuación, se determinará hiperparámetro K para cada algoritmo mediante las gráficas de la **métrica Silhoutte**.

Algoritmo K-Means

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]



Se observa que los gráficos (áreas de colores) están balanceados para los distintos valores de K – clusters homogéneos. Esto muestra que K-Means está funcionando bien para el Dataset. Se revisarán los valores del coeficiente solhouette para cada K.

Valores del **coeficiente silhouette** para cada K – algoritmo K-Means

silhouette(K=6): 0.3589

silhouette(K=5): 0.3616

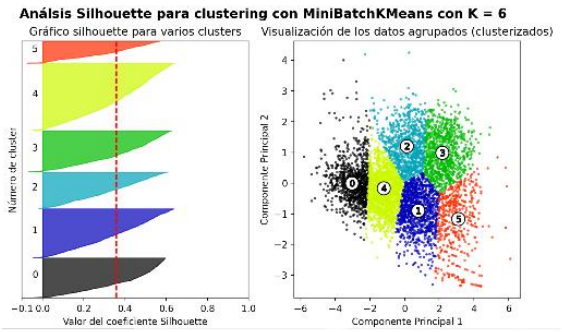
silhouette(K=4): 0.3641

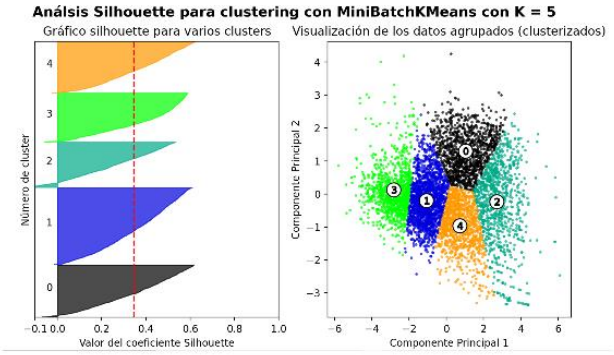
silhouette(K=3): 0.3511

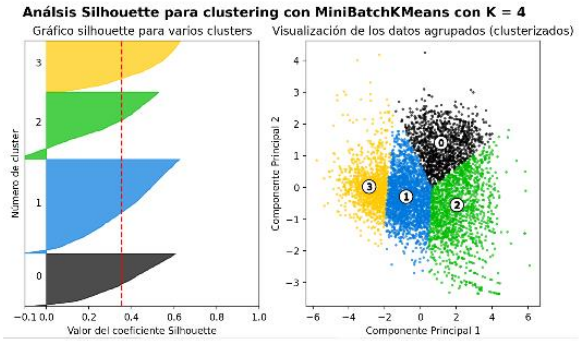
Por lo tanto, **el K que maximiza el rendimiento de K-Means es 4**. Desde el punto de vista del dominio, un K igual a 3 no aportaría mucho valor al análisis dado que agrupa básicamente por la componente 1. Los valores de K igual a 5 y 6 permitirían una mayor apertura pero, a la vez, nos alejan de una primera segmentación práctica a los efectos de toma de decisión sobre la cartera de cliente.

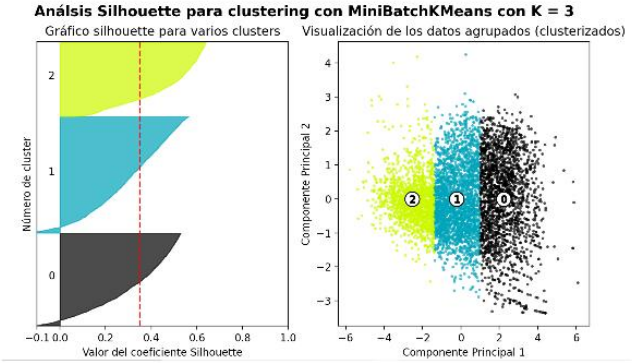
Algoritmo MiniBatchKMeans

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Valores del **coeficiente silhouette** para cada K – algoritmo MiniBatchKMeans

silhouette(K=6): 0.3590

silhouette(K=5): 0.3459

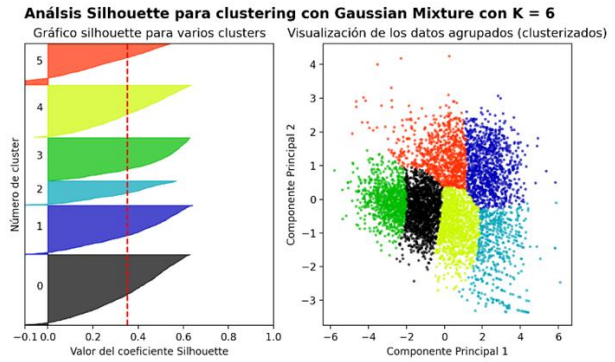
silhouette(K=4): 0.3529

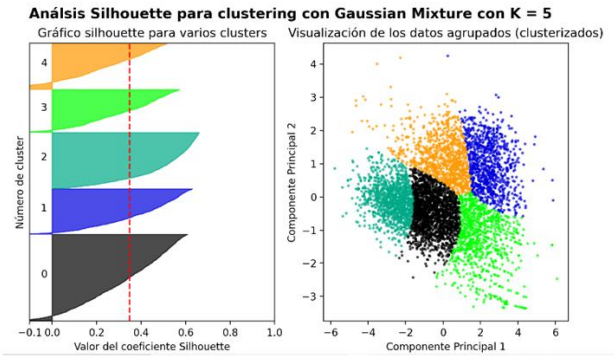
silhouette(K=3): 0.3504

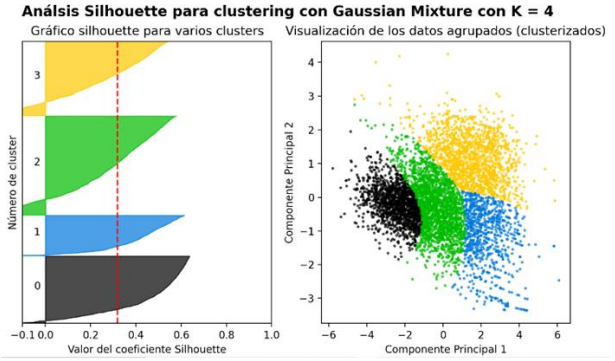
Por lo visto, los resultados son similares a los obtenidos con K-Means. En este caso, **el K que maximiza el rendimiento de MiniBatchKMeans es 6.** Nuevamente, desde el punto de vista del dominio, K=6 permitirían una mayor apertura pero a la vez nos aleja de una primera segmentación práctica para efectos de toma de decisión sobre la cartera de cliente.

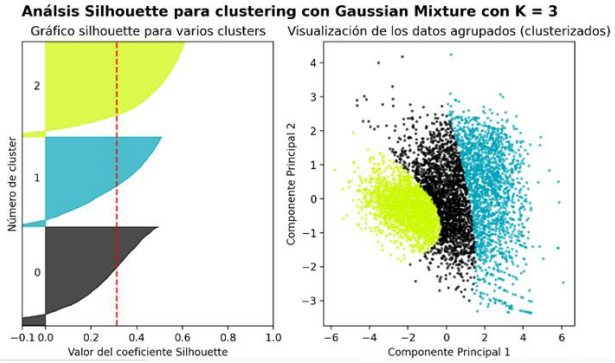
Algoritmo Gaussian Mixture

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Se observa que los gráficos están menos balanceados para los distintos valores de K – por lo tanto, los clusters no son tan homogéneos como con K-Means. Se revisarán los valores del coeficiente silhouette para cada K.

Valores del **coeficiente silhouette** para cada K – algoritmo Gaussian Mixture

silhouette(K=6): 0.3514

silhouette(K=5): 0.3493

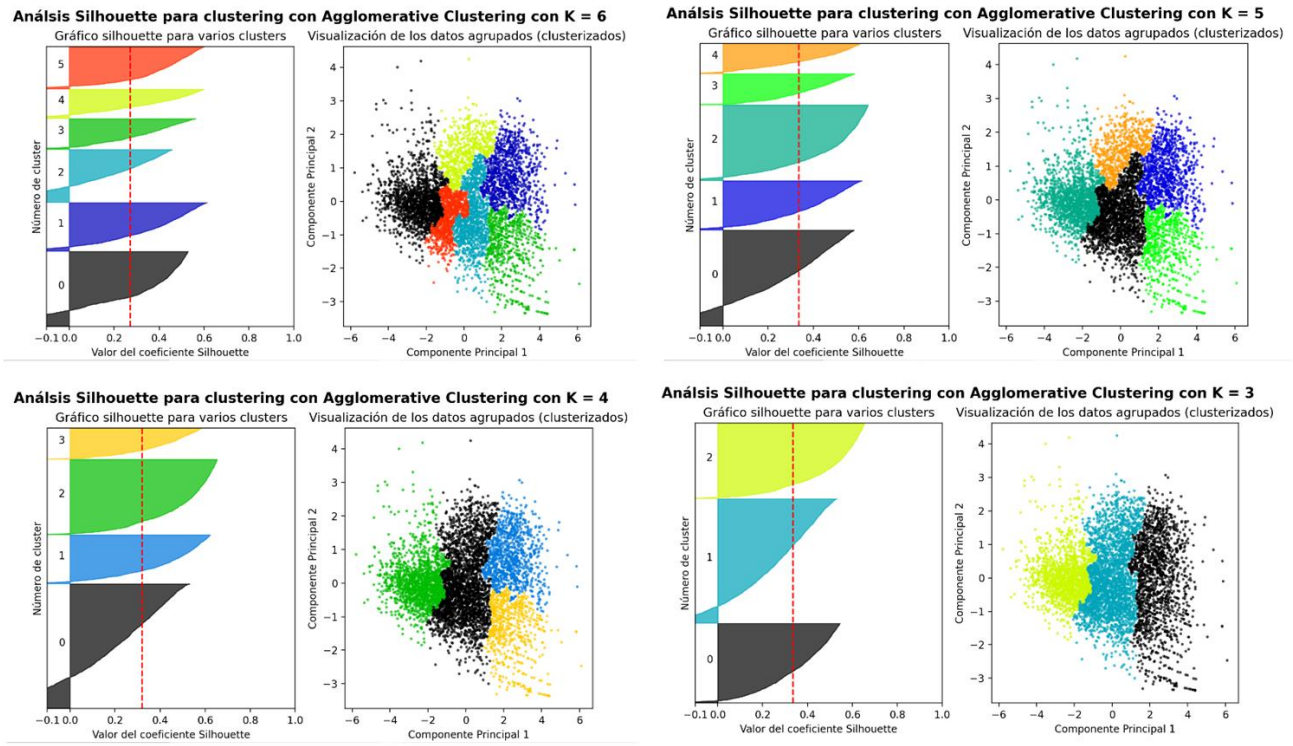
silhouette(K=4): 0.3186

silhouette(K=3): 0.3129

Por lo tanto, **el K que maximiza el rendimiento de Gaussian Mixture es 6.**

Algoritmo Agglomerative Clustering

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]



Se observa que los gráficos están menos balanceados para los distintos valores de K – por lo tanto, los clusters no son tan homogéneos. Se revisarán los valores del coeficiente silhouette para cada K.

Valores del **coeficiente silhouette** para cada K – algoritmo Agglomerative Clustering

silhouette(K=6): 0.2708

silhouette(K=5): 0.3363

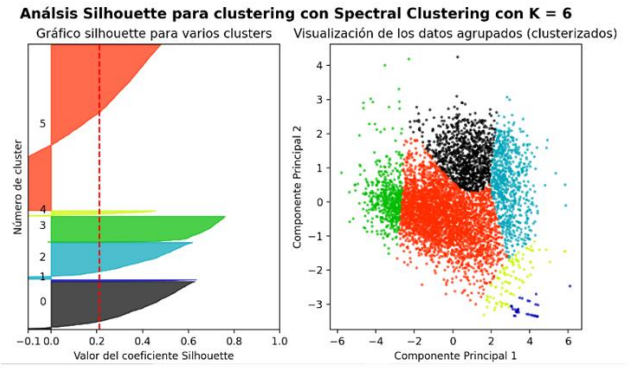
silhouette(K=4): 0.3192

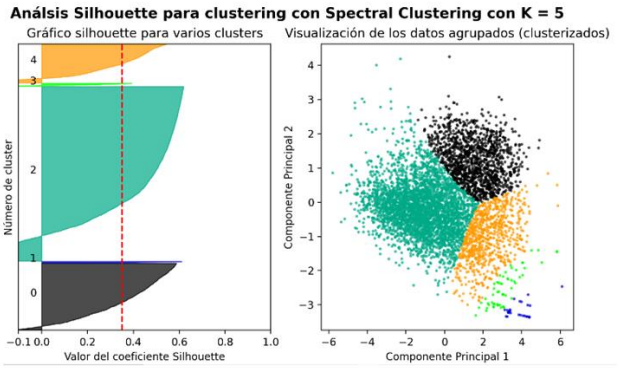
silhouette(K=3): 0.3361

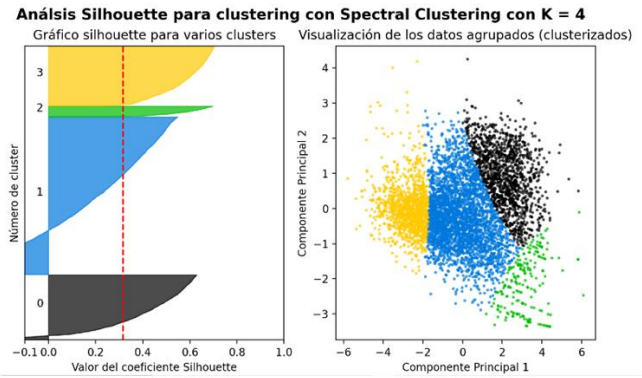
Por lo tanto, **el K que maximiza el rendimiento de Agglomerative Clustering es 5.**

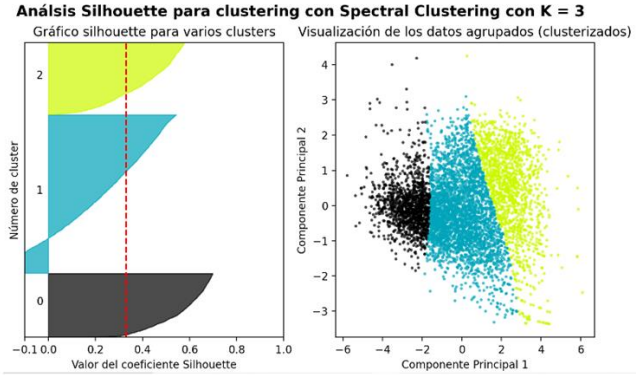
Algoritmo Spectral Clustering

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Se observa que los gráficos están cada vez menos balanceados para los distintos valores de K. Se revisarán los valores del coeficiente silhouette para cada K.

Valores del **coeficiente silhouette** para cada K – algoritmo Spectral Clustering

silhouette(K=6): 0.2103

silhouette(K=5): 0.3499

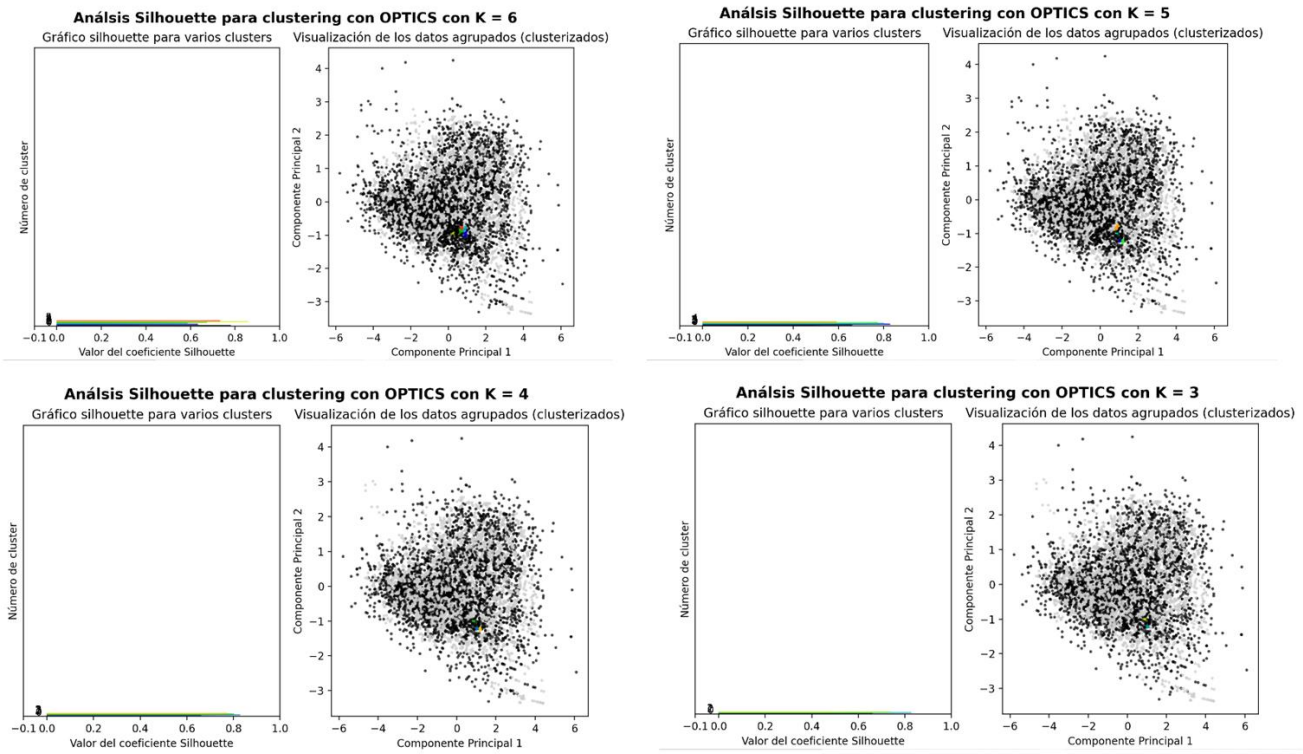
silhouette(K=4): 0.3159

silhouette(K=3): 0.3301

Por lo tanto, **el K que maximiza el rendimiento de Spectral Clustering es 5**. En el gráfico para K=5 se observa una agrupación menor (en azul) que podría ser de interés para el dominio. Luego de un análisis, se observa que dicha agrupación está conformada por 26 clientes y con un atractivo comercial bajo: rentabilidad 1% promedio, gama de 1 producto promedio y frecuencia de compra igual a 1 por año promedio.

Algoritmo OPTICS

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]



A través de los gráficos se observa que el desempeño de este algoritmo no es bueno. Se revisarán los valores del coeficiente silhouette para cada K.

Valores del **coeficiente silhouette** para cada K – algoritmo OPTICS

silhouette(K=6): -0.2340

silhouette(K=5): -0.2026

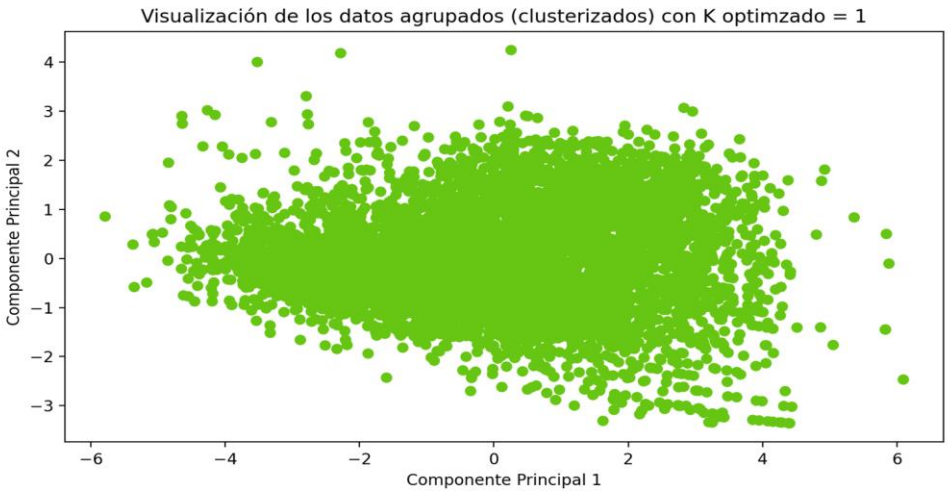
silhouette(K=4): -0.1645

silhouette(K=3): -0.1410

Por lo tanto, **el K que maximiza el rendimiento de OPTICS es 3.**

Algoritmo *Mean Shift*

Gráfico con la visualización de los clusters obtenidos con Mean Shift.

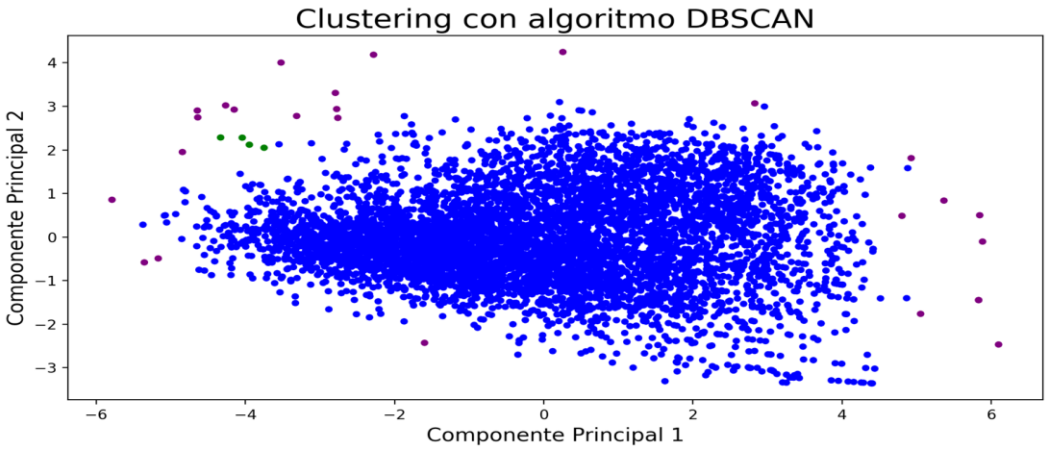


A través del gráfico se observa que el desempeño de este algoritmo no es bueno.

En este caso, **el K que maximiza el rendimiento de Mean Shift es 1.**

Algoritmo DBSCAN

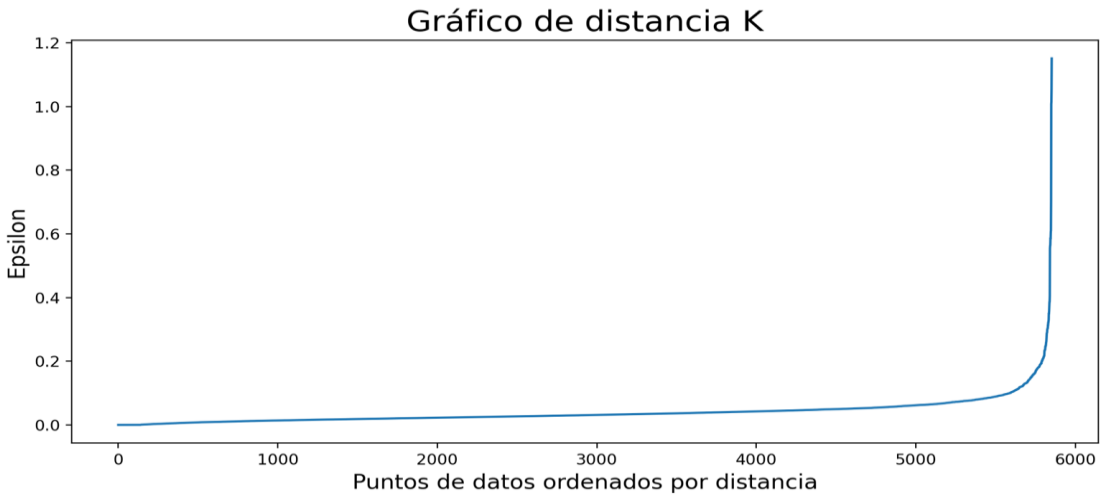
Gráfico con la visualización de los clusters obtenidos con DBSCAN sin optimizar ninguno de sus parámetros (epsilon & minPoints).



De acuerdo a la lógica de este algoritmo, todos puntos de datos fueron considerados como "ruido" dado que sus parámetros no fueron optimizados.

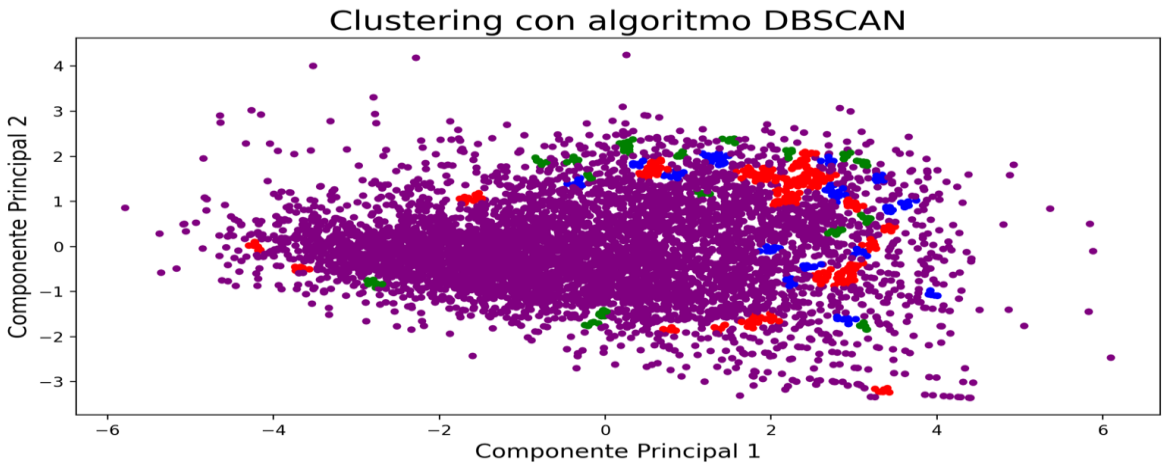
Para optimizar el parámetro epsilon se utiliza el gráfico de "distancia K".

Gráfico de "distancia K"



El valor óptimo de epsilon es el punto de máxima curvatura en el gráfico de distancia K, que en este caso sería 0,1. Respecto al parámetro minPoints, seleccionaremos el valor 6.

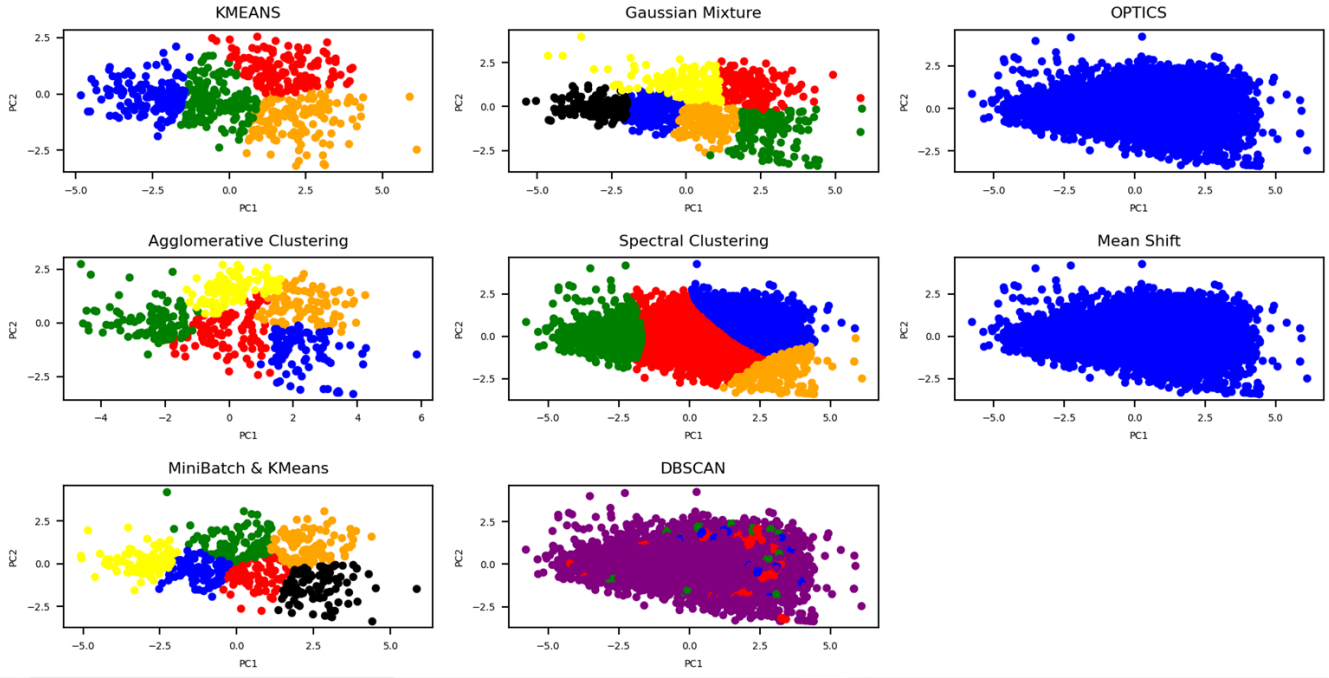
Gráfico con la visualización de los clusters obtenidos con DBSCAN.



A pesar de haber optimizado el parámetro epsilon, se observa que el desempeño de este algoritmo no es bueno.

Finalmente, se mostrarán un gráficos con todos los algoritmos evaluados y con su valor de K optimizado.

Gráfica de todos los algoritmos utilizados – cada uno con su valor de K optimizado



El gráfico muestra los algoritmos de bajo desempeño: Mean Shift, Optics y DBSAN. Por otro lado, se observa que el resto aportan agrupaciones similares, más allá de la cantidad de clusters, especialmente Gaussian Mixture con MinibatchKMeans y K-Means con Spectral Clustering.

Otro aspecto observado, los algoritmos Gaussian Mixture, Spectral y Agglomerativel muestran gráficas de Silhouette desbalancedas. K-Means y MinibatckKMeans demuestran ser los algoritmos más atractivos para resolver la segmentación.

Finalmente, de acuerdo con las observaciones y resultados y obtenidos hasta aquí, se concluye que los clusters que se asignarían a los clientes tendrían origen en los segmentos generados por K-Means para un K = 4.

7.3. Clustering sobre variables categóricas – K-Modes & MCA

El objetivo de esta tercera estrategia es aplicar distintos algoritmos y enfoques de segmentación para resolver el clustering sobre las variables categóricas del Dataset: rango de precio y zona PAS.

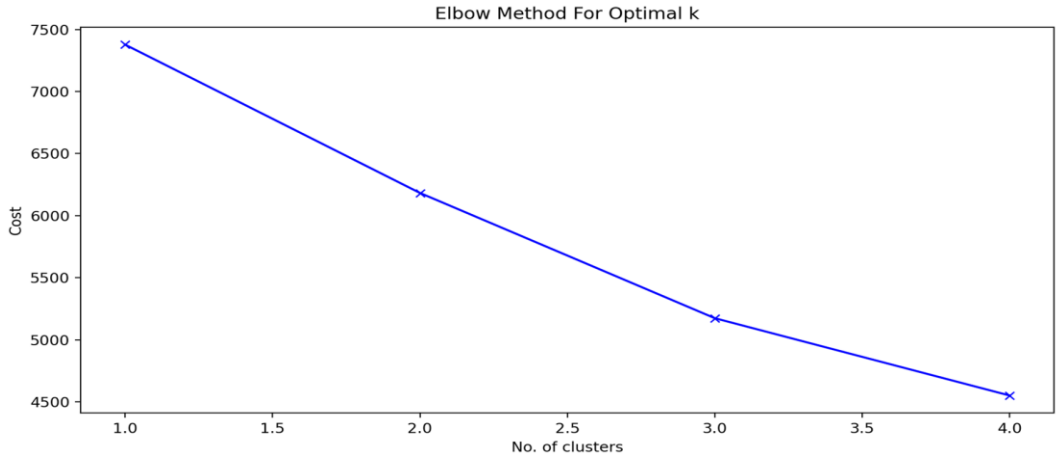
Algoritmos y enfoques de clustering que se aplicarán en esta estrategia:
- Algoritmo **K-Modes**
- Análisis de Correspondencia Múltiple (**MCA**)

K-Modes es un algoritmo utilizado para segmentación cuando las variables del dataset son categóricas. Para dicho tipo de variables no es posible calcular distancias entre puntos (como lo hace el algoritmo K-Means). K-Modes utiliza diferencias (diferencias totales) entre datos de los registros del Dataset. A menor cantidad de diferencias mayor similitud entre los datos de los registros.

A diferencia del clustering jerárquico, K-Modes requiere que especifiquemos un valor para K (cantidad de clusters).

Por lo tanto, aplicaremos el concepto de gráfico de codo para optimizar el hiperparámetro K.

Gráfico del Codo



Para K-Modes, se grafica el costo para diferentes valores de K. El costo es la suma de todas las diferencias entre clusters. Se selecciona el K que indique el quiebre (el codo) de la curva de costo.

Observamos el K que optimiza el desempeño del algoritmo K-Modes es K=3.

La columna clusters de la siguiente tabla contiene los segmentos generados por el algoritmo K-Modes

#cliente	principal component 1	principal component 2	clusters
C_1	1.150004	-1.047418	0
C_10	1.065610	-1.389805	2
C_100	-4.330941	2.284186	1
C_1000	-3.118454	0.119032	0
C_1001	-1.643319	-0.223896	2
...	...	...	...
C_995	0.519075	0.017620	1
C_996	-1.811219	-0.194334	2
C_997	-2.452329	1.786395	2
C_998	-0.965644	-1.483581	0
C_999	-0.327849	-0.119814	0

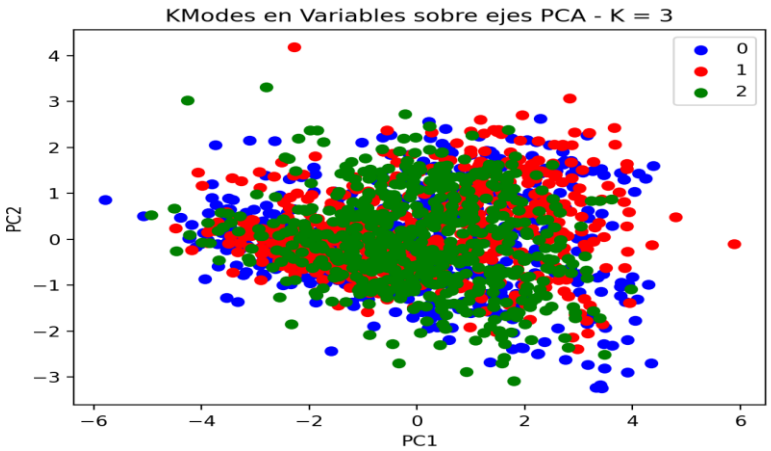
Para entender mejor las agrupaciones realizadas por K-Modes graficaremos el resultado en 2D. Para ello, tomaremos 3 criterios:

- Graficar los clusters sobre el eje de las componentes principales CP1 y CP2 obtenidas en la estrategia anterior

- Graficar los clusters sobre 2 variables cuantitativas: Facturación y Margen sobre Facturación

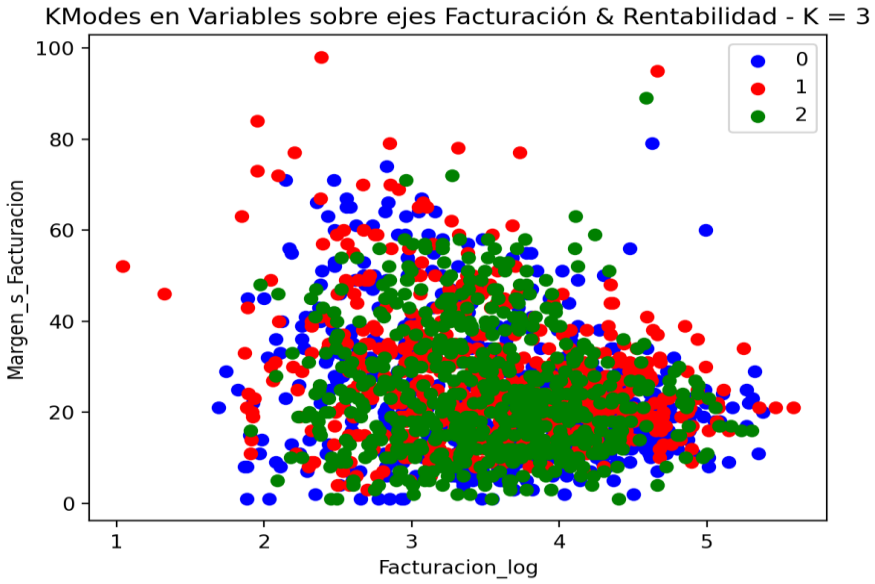
- Graficar los clusters sobre las variables categóricas: rango de precio y zona PAS.

Gráfico sobre el eje de las componentes principales CP1 y CP2



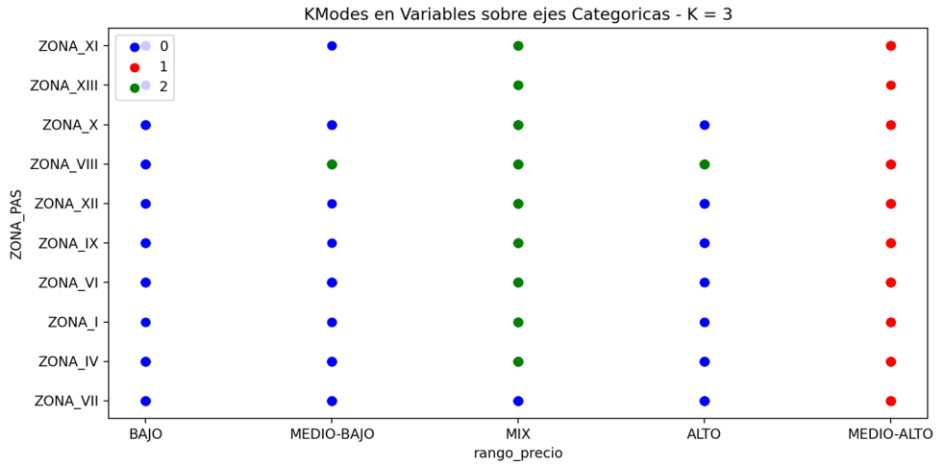
De acuerdo con el gráfico, no se observan clusters definidos con claridad en el plano PC1-PC2.

Gráfico sobre 2 variables cuantitativas: Facturación y Margen sobre Facturación



De acuerdo al gráfico, no se observan clusters definidos con claridad en el plano definido por los ejes "Facturacion_log" y "Margen_s_Facturacion".

Gráfico sobre las variables categóricas: rango de precio y zona PAS



De acuerdo con este gráfico, K-Modes agrupó a los clientes de la siguiente manera:

Cluster 1 – clientes del rango de precio **MEDIO-ALTO** con **todas las zonas PAS**.

Cluster 2 – clientes pertenecientes al rango de precio **MIX** y a casi **todas las zonas PAS**

Resto de los clientes - resto de los rangos de precios y todas las zonas PAS.

Por lo tanto, podemos decir que el algoritmo K-Modes diferenció principalmente por 2 rangos de precio – MEDIO-ALTO y MIX – del resto de clientes.

Finalmente, desde el punto de vista del negocio los clusters no aportan mucho valor. Sin embargo, tendremos en cuenta estas agrupaciones al momento de analizar las próximas estrategias con todas las variables en juego.

MCA permite analizar el patrón de relaciones de varias variables dependientes categóricas. Como tal, también puede verse como una generalización del análisis de componentes principales (PCA) cuando las variables a analizar son categóricas en lugar de cuantitativas.

Para obtener una agrupación o clusters se deberá correr el proceso MCA sobre las variables categóricas. De esta forma, se obtendrán las componentes principales (coordenadas numéricas) para poder aplicar sobre ellas un algoritmo de clustering, como por ejemplo: K-Means.

Se ejecuta MCA sobre las variables categóricas del Dataset y se obtienen las coordenadas numéricas para sus 2 componentes:

	0	1
0	-0.052766	0.225560
1	1.810530	-0.008681
2	0.063869	0.464130
3	-0.697059	-0.277407
4	1.810530	-0.008681
...	...	...
5848	0.063869	0.464130
5849	0.729332	0.591356
5850	-0.645252	0.175318
5851	-1.074461	-0.600132
5852	-1.074461	-0.600132

Nuevamente, antes de aplicar a los datos cualquier método de *Clustering* es conveniente **evaluar la tendencia de los datos al clustering**. Para ello, utilizaremos una vez más los siguientes métodos:

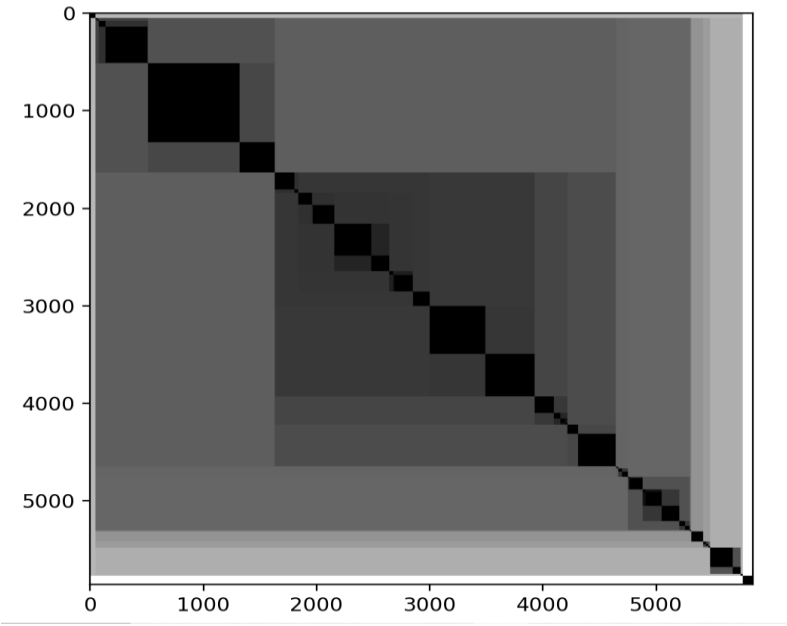
c) Métodos estadísticos: estadístico **Hopkins**

d) Métodos visuales: algoritmo **VAT** (*Visual Assessment of cluster Tendency*)

Valor Hopkins obtenido: (H) = 0.0001

En este caso el test es indeciso (sin embargo, los datos no son aleatorios)

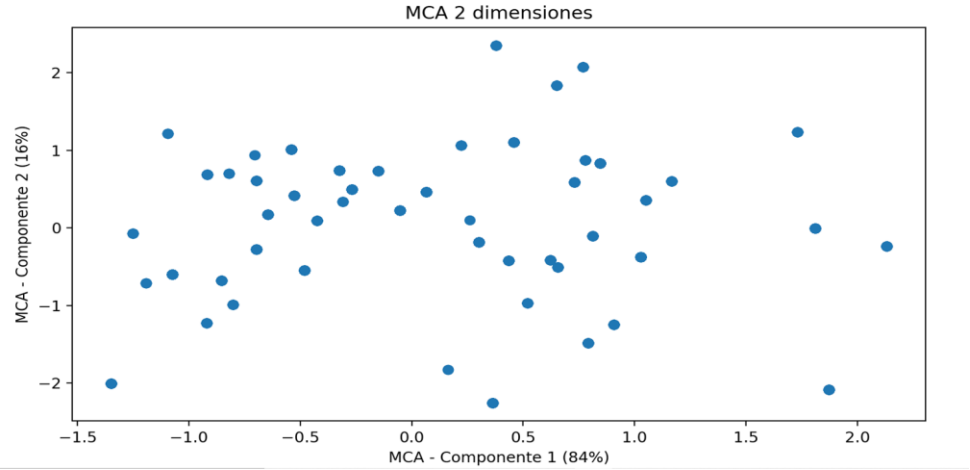
VAT (*Visual Assessment of cluster Tendency*)



El gráfico muestra poca evidencia sobre la presencia de clusters definidos. A pesar de ello, esta información es valiosa porque nos permite entender cuáles son las variables del dataset que aportarían información para una segmentación. En este contexto, las 2 variables categóricas analizadas de forma aislada no estarían aportando – a priori – información para poder obtener agrupaciones valiosas de clientes.

De todas formas, como siguiente paso, se realizará una **inspección visual** 2D sobre los datos para evaluar si hay indicios de posibles agrupaciones con las variables categóricas. Para ello, se utilizará el gráfico MCA sobre las coordenadas numéricas y se correrá el algoritmo K-Means.

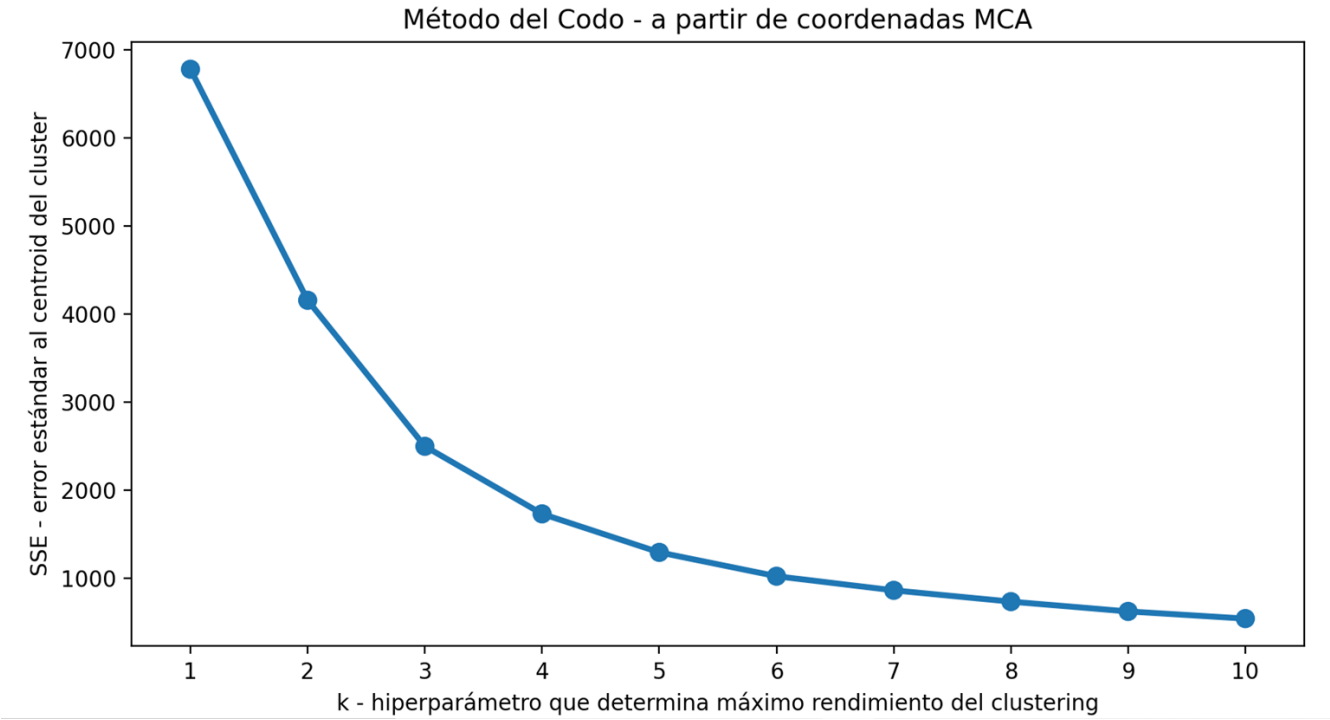
Gráfico MCA sobre sus 2 componentes



KMEANS para definir clusters

Nuevamente, antes de correr el algoritmo de clustering K-Means se estimará a través del método del codo (por medio del error estándar al centro del clúster - SSE) la cantidad estimada de clúster.

Gráfico del Codo

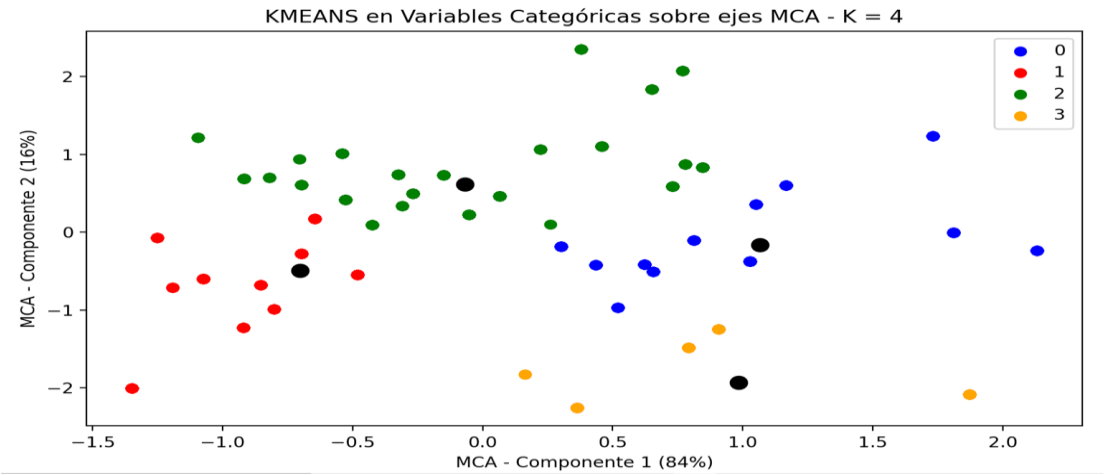


Según el gráfico de codo, a partir del 4to cluster ya no se obtienen saltos del SSE significativos, por lo tanto, **la cantidad de clusters a probar será 4**.

Clusters a partir de K-Means sobre coordenadas numéricas MCA

	y_predict_dataframe
0	2
1	0
2	2
3	1
4	0
...	...
5848	2
5849	2
5850	1
5851	1
5852	1

Gráfico 2D con los segmentos propuestos por K-Means



Análisis de los segmentos propuestos por K-Means:

cluster 0 – el 70% de los clientes de este cluster están en el **rango de precio ALTO**, destacándose la **zona VII**. (total cluster 1.111 clientes)

cluster 1 – el 88% de los clientes de este cluster están el **rango de precio BAJO**, destacándose nuevamente y principalmente la **zona VII**. (total cluster 1.818 clientes)

cluster 2 – el 47% de los clientes de este cluster pertenecen a la **zona PAS VII**, con rangos de precios MEDIO-ALTO, MEDIO-BAJO y MIX. (total cluster 2.664 clientes)

cluster 3 – el 98% de los clientes de este cluster pertenecen a la **zona PAS XII**, con rangos de precios ALTO principalmente y BAJO. (total cluster 280 clientes)

Por lo tanto, la segmentación propuesta divide a la zona PAS VII, la mayor en cantidad de clientes, en función de los rangos de precios ALTO y BAJO, segmento 0 y 1 respectivamente. Por otro lado, los segmentos 2 y 3 aportan poca información.

La segmentación no es de gran aporte para el negocio. A pesar de ello, tendremos en cuenta estas agrupaciones al momento de analizar las próximas estrategias con todas las variables en juego.

7.4. Clustering sobre todo el Dataset – 2 etapas & aplicación de varios algoritmos

El objetivo de esta cuarta estrategia es aplicar distintos algoritmos sobre todas las variables del Dataset original, utilizando MCA sobre las variables categóricas. Luego, se utilizarán diversos algoritmos de clustering y se compararán a través de métricas. Finalmente, se identificará el mejor algoritmo para esta estrategia.

Algoritmos que se utilizarán:

- *K-Means*
- *MiniBatchKMeans*
- GaussianMixture
- *AgglomerativeClustering*
- *SpectralClustering*
- *OPTICS*
- *MeanShift*
- *DBSCAN*

Como se mencionó anteriormente, se aplicará MCA con el objetivo de obtener las coordenadas numéricas de las 2 variables categóricas, lo cual permitirá sumarlas con resto de las 6 variables cuantitativas para finalmente completar el Dataset con las 8 variables. Por otro lado, todas las variables serán ajustadas para asegurar simetría y normalización.

Dataset simétrico y normalizado con las 8 variables

#cliente	rango_precio	ZONA_PAS	Gama_Productos	Facturacion	Margen_Bruto	Facturacion_s_#Producto	Margen_s_Facturacion	Frecuencia_Compra
C_1	-0.067532	0.304573	-0.255743	-0.289148	-0.666508	-0.090234	-1.001484	-1.332928
C_10	2.317173	-0.011722	-0.255743	-0.199929	-0.731850	0.046116	-1.361575	-1.332928
C_100	0.081741	0.626713	1.435807	2.075223	2.892588	1.545367	1.860283	1.869684
C_1000	-0.892118	-0.374581	1.205319	1.562803	1.585469	1.115388	-0.110198	1.442772
C_1001	2.317173	-0.011722	1.205319	0.695475	0.667641	-0.100325	-0.110198	1.092391
...	...	...	...	...	...	...	...	...
C_995	0.081741	0.626713	-0.255743	-0.465891	-0.410452	-0.363525	0.104838	0.347572
C_996	0.933422	0.798506	1.582221	0.709067	0.745429	-0.510249	0.035032	1.360361
C_997	-0.825813	0.236732	-0.634053	1.185097	1.616446	2.323935	1.094455	1.265815
C_998	-1.375129	-0.810357	0.784073	0.482037	-0.099980	0.040426	-1.361575	0.760638
C_999	-1.375129	-0.810357	0.600374	0.053188	0.085635	-0.409640	0.035032	0.347572

Nuevamente, antes de aplicar a los datos cualquier método de *Clustering* es conveniente **evaluar la tendencia de los datos al clustering**. Para ello, utilizaremos una vez más los siguientes métodos:

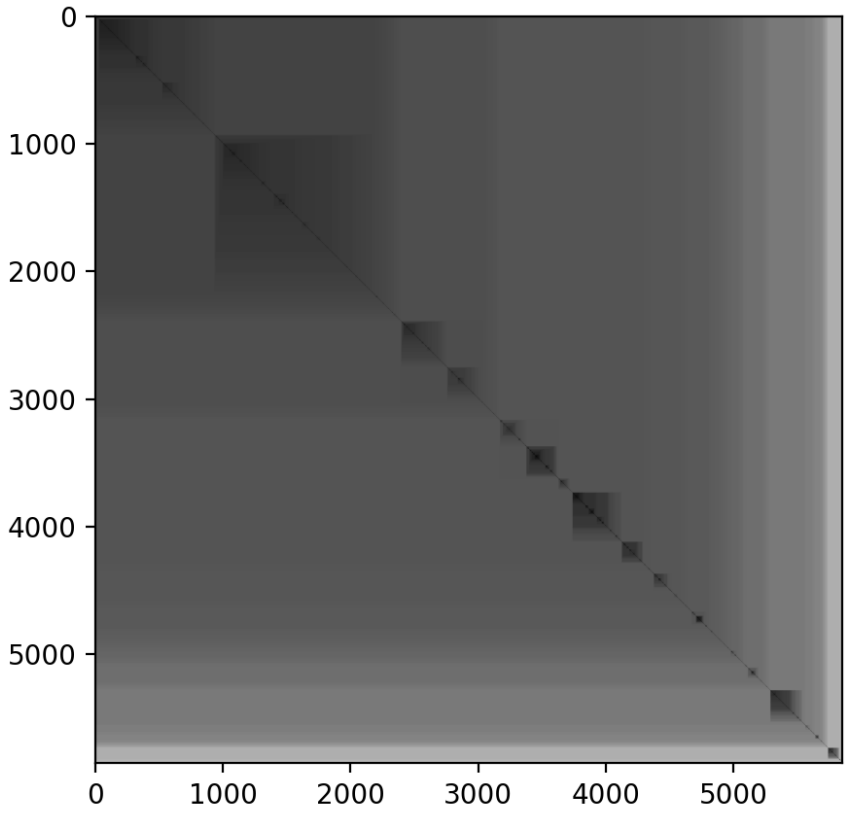
e) Métodos estadísticos: estadístico **Hopkins**

f) Métodos visuales: algoritmo **VAT** (*Visual Assessment of cluster Tendency*)

Valor Hopkins obtenido: (H) = 0.1173

En este caso el test es indeciso (sin embargo, los datos no son aleatorios).

VAT (*Visual Assessment of cluster Tendency*)



El gráfico muestra cuadrados negros que representan clusters. En este caso, podemos observar varios cuadrados pequeños, indicando poca definición de grupos o clusters en los datos.

Como siguiente paso, evaluaremos si hay indicios de algún tipo de agrupación en los datos. Por lo tanto, se realizará una **inspección visual** 2D sobre los mismos.

Para poder graficar resultados en 2 D necesitamos reducir las dimensiones de los datos, de 8 a 2. Para ello realizaremos el proceso **PCA – Análisis de Componentes Principales**.

Variación Explicada por cada componente principal:

[4.79500671e-01; 1.46926621e-01; 1.29611246e-01; 1.20210544e-01;

1.02239813e-01; 2.03739081e-02; 9.76506033e-04; 1.60691321e-04]

En este caso, la variación explicada por las primeras 2 componentes principales es de aproximadamente **63%**.

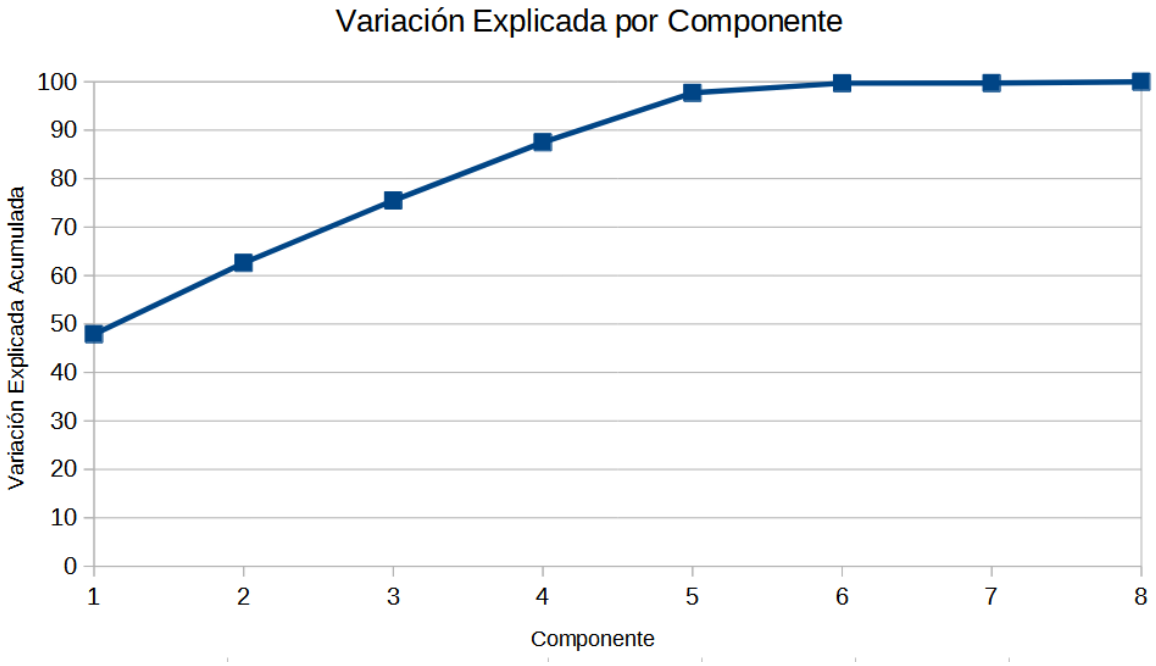
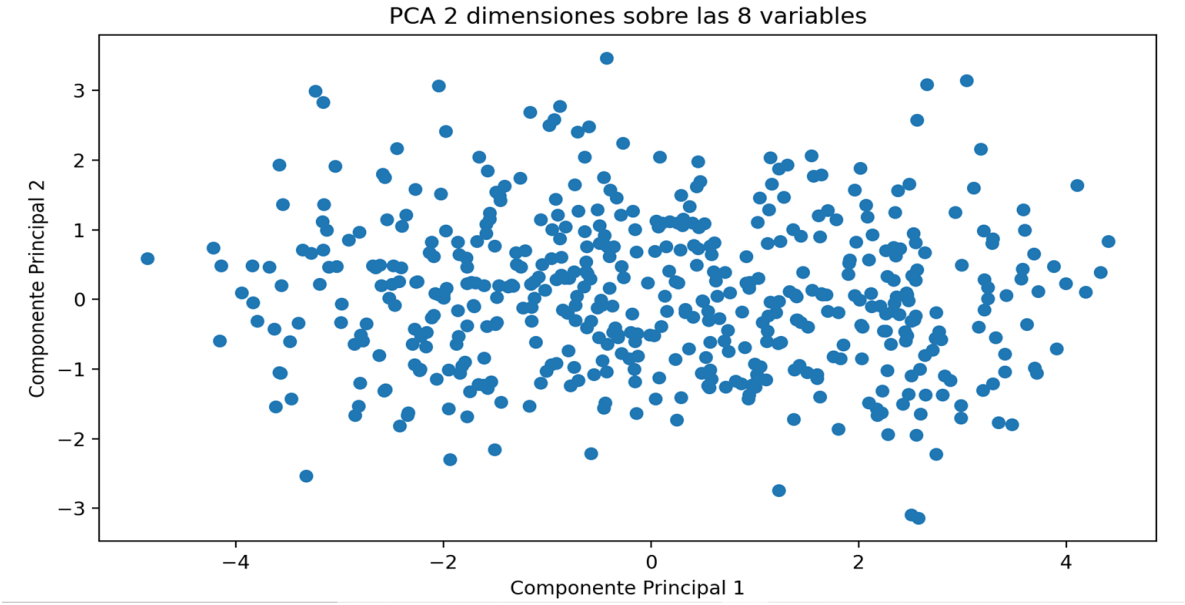


Gráfico PCA sobre componentes principales PC1 y PC2 – cada punto es un cliente (muestra reducida)



Del gráfico se observa que no se trata de puntos concentrados en un sector. Seguramente, los algoritmos de clustering nos ayudarán a identificar los segmentos que en una primera instancia no se identifican.

Matriz de rotación – variables X1 a X8 (filas) y componentes principales Y1 y Y2 (columnas)

```
[[-0.01431442 -0.57611418]
 [ 0.09151763 -0.22113084]
 [-0.41512043 -0.0211787 ]
 [-0.50181461  0.07461938]
 [-0.48316362 -0.20316757]
 [-0.35481356  0.13687557]
 [ 0.0634858  -0.73898432]
 [-0.45156014 -0.08406666]]
```

X1: rango_precio,
X2: ZONA_PAS,
X3: Gama_Productos,
X4: Facturacion,
X5: Margen_Bruto,
X6: Facturacion_s_#Producto,
X7: Margen_s_Facturacion,
X8: Frecuencia_Compra

Interpretación del PCA por matriz de rotación: variables de componentes principales PC1 y PC2 en función de las 8 variables del dataset:

PC1 = – (0.50181461 * X4 + 0.48316362 * X5 + 0.45156014 * X8 + 0.41512043 * X3 + 0.35481356 * X6) + (0.09151763 *X2)

→ Entonces PC1 diferencia a los clientes que destacan por **Facturación**, **Margen Bruto**, **Frecuencia de Compra**, **Gama de Productos** y **Facturacion_s_#Producto** del resto.

PC2 = - (0.73898432 * X7 + 0.57611418 * X1 + 0.22113084 * X2) + (0.13687557 * X6 + 0.07461938 * X4)

→ Por otro lado, PC2 diferencia a los clientes que destacan por **Margen sobre Facturación** (rentabilidad), **Rango de Precio** y **Zona PAS** del resto.

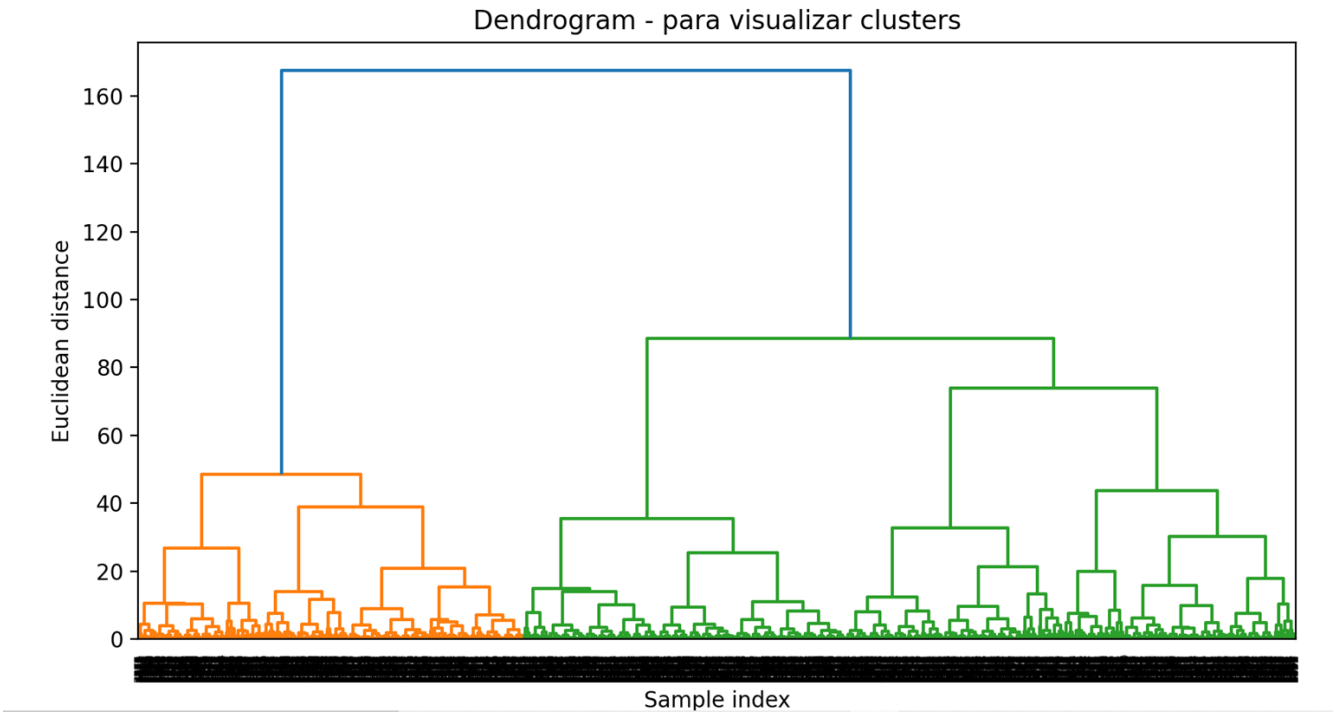
Un primer análisis de esta información muestra – a priori – que no hubo un aporte significativo de las variables categóricas (zona_PAS y rango_precio) al proceso de clustering. Por otro lado, la variable Margen sobre Facturación sigue teniendo un peso fuerte sobre la componente 2 y, en segunda instancia, la variable Rango de Precio.

No obstante, la información obtenida hasta aquí (a pesar del poco aporte de las variables categóricas) sigue siendo valiosa para La Empresa, porque ayudaría al área comercial a orientar el diseño de sus campañas de ventas – por ejemplo: no lanzando mensajes o iniciativas comerciales segmentadas geográficamente, sino por otro tipo de variables que definan el parecido entre clientes (ejemplo: rentabilidad).

Siguiendo con los objetivos de la presente estrategia, se avanzará en el análisis del Dataset para tratar de **identificar clusters valiosos**. Para ello, utilizaremos el algoritmo **clustering jerárquico** y **K-Means**.

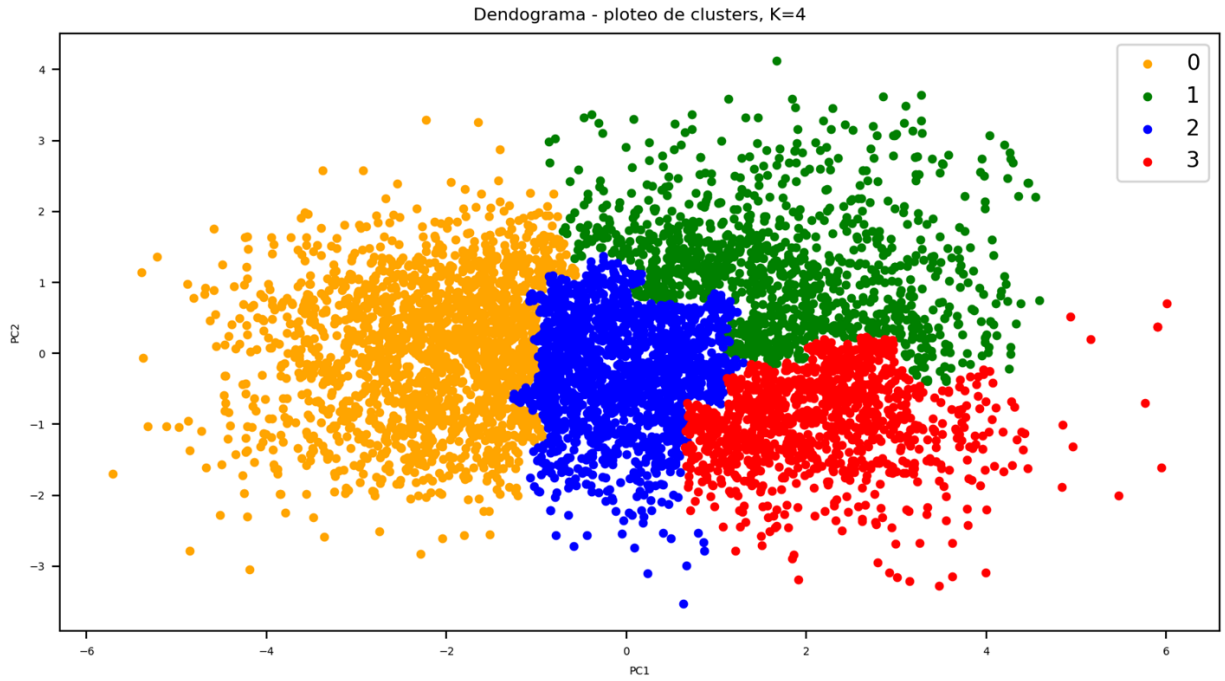
Clustering jerárquico

Dendograma



El **Dendograma** muestra una clusterización de los datos. De acuerdo con la distancia Eucleidiana, eje vertical, la cantidad de clusters razonables para este Dataset estaría definida entre 3 y 4.

Se graficarán los 4 clusters sugeridos por el dendograma sobre los ejes de las 2 componentes principales.



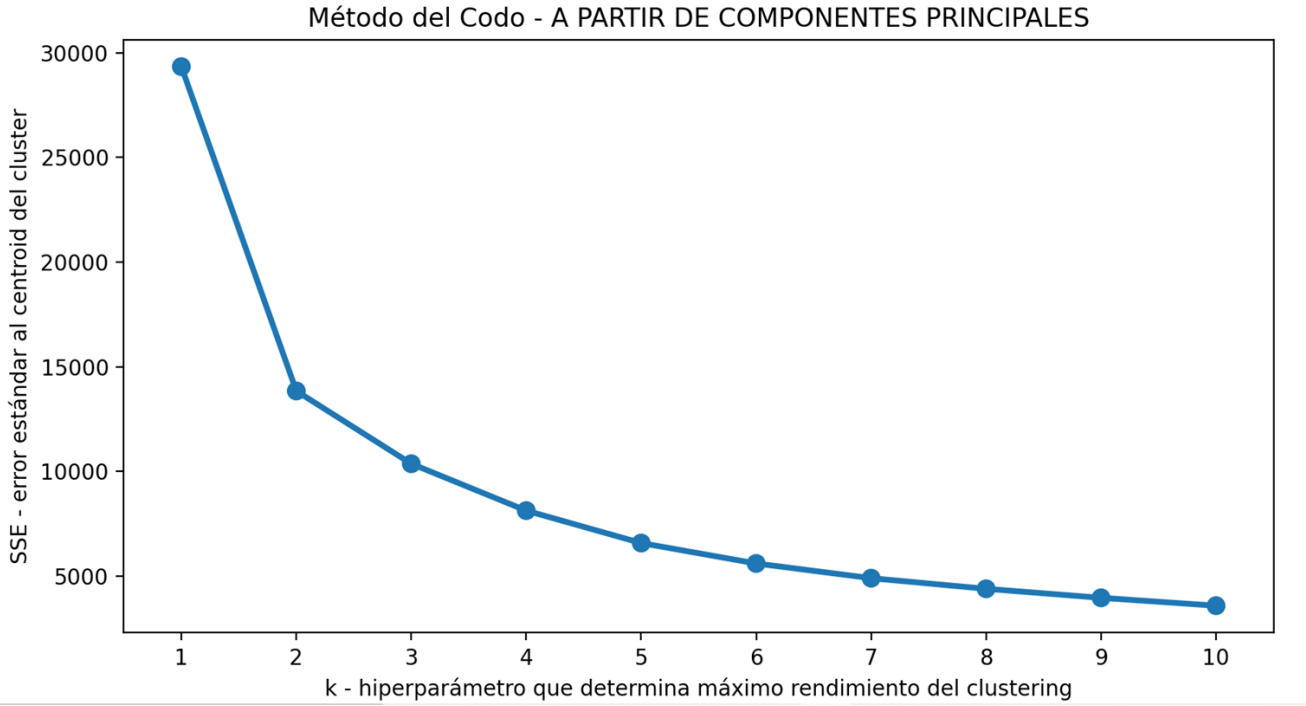
Del gráfico se observa una primera propuesta de segmentos o clusters.

Seguiremos el proceso de clustering o segmentación con otro algoritmo más complejo: K-Means.

Clustering K-Means sobre las 8 variables cuantitativas

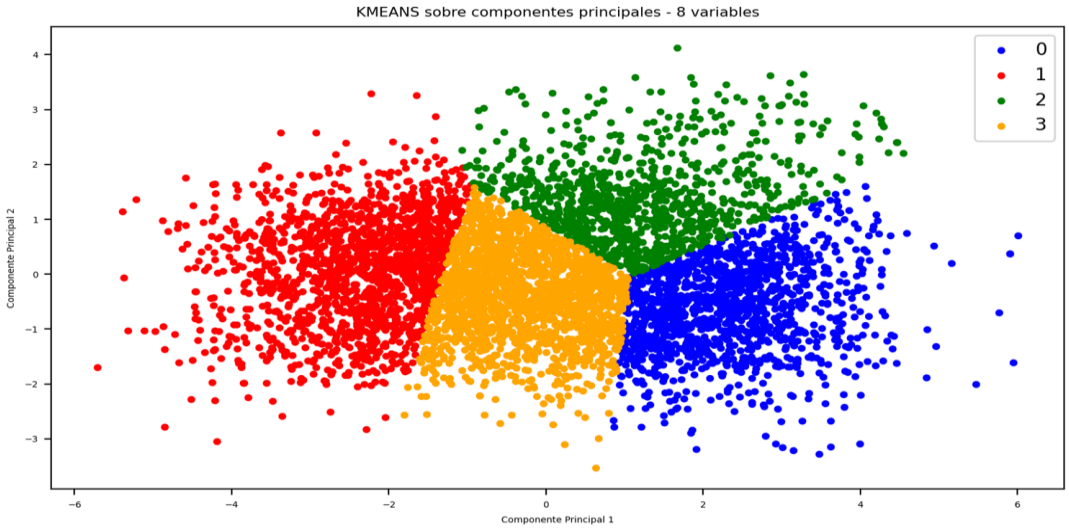
Nuevamente, antes de correr el algoritmo de clustering K-Means se estimará la cantidad estimada de clúster a través del método del codo (por medio del error estándar al centro del clúster - SSE).

Gráfico del Codo



Según el gráfico de codo, a partir del 4to cluster ya no se obtienen saltos del SSE significativos, por lo tanto, **la cantidad de clusters a probar será 4**.

Clusters a partir de **K-Means** sobre variables cuantitativas graficados en ejes PCA



Del gráfico se observan los 4 clusters generados por K-Means. Los cuales, luego de un análisis, muestran que mantienen grandes similitudes respecto a los clusters obtenidos en las estrategias 1 y 2 a partir de las 6 variables cuantitativas.

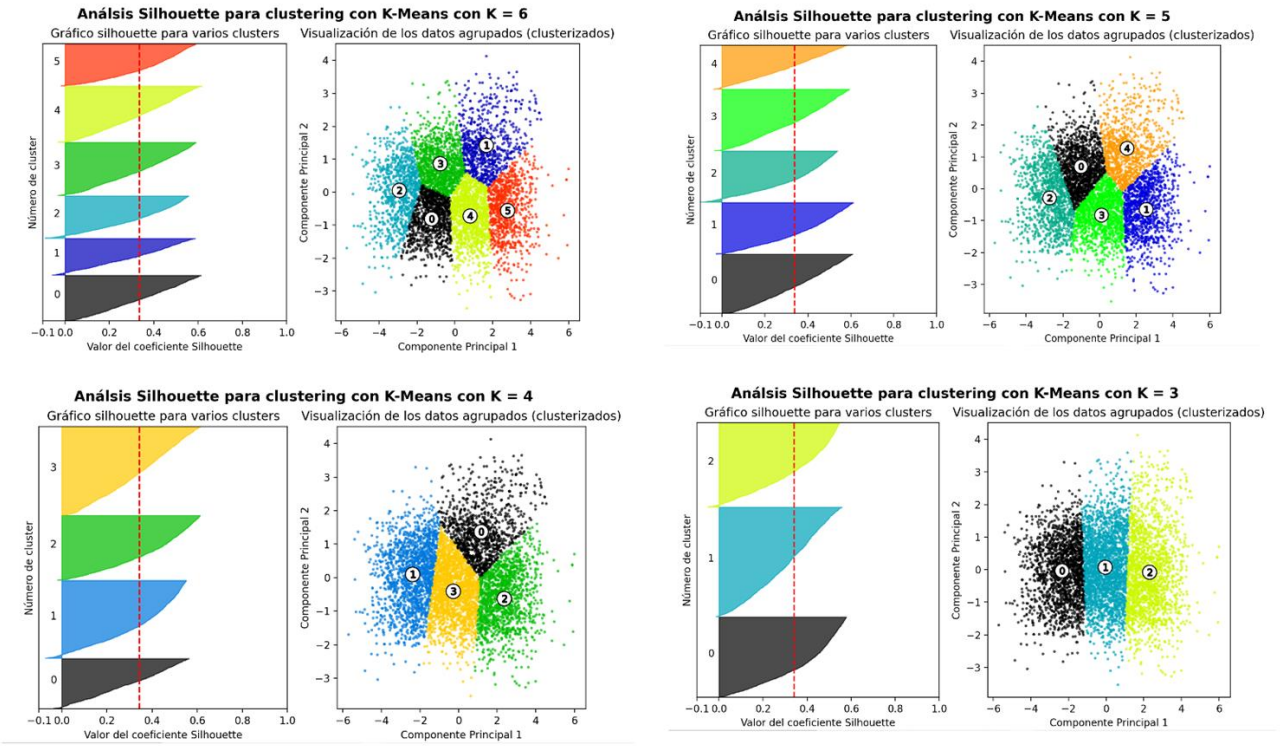
A continuación, se determinará hiperparámetro K para cada algoritmo mediante las gráficas de la **métrica Silhoutte**.

Algoritmos que se utilizarán:

- *K-Means*
- *MiniBatchKMeans*
- GaussianMixture
- *AgglomerativeClustering*
- *SpectralClustering*
- *OPTICS*
- *MeanShift*
- *DBSCAN*

Algoritmo K-Means

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]



Se observa que los gráficos están balanceados para los distintos valores de K – clusters homogéneos. Esto muestra que K-Means está funcionando bien para nuestro Dataset. Se revisarán los valores del coeficiente silhouette para cada K.

Valores del **coeficiente silhouette** para cada K – algoritmo K-Means

silhouette(K=6): 0.3352

silhouette(K=5): 0.3391

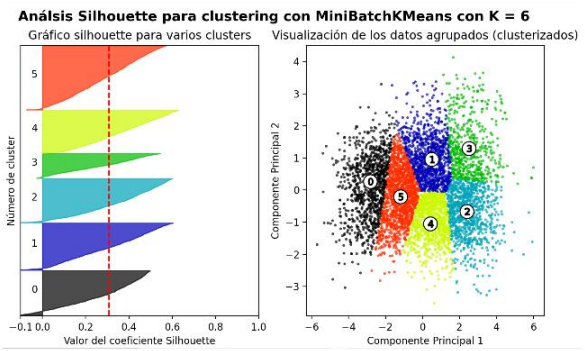
silhouette(K=4): 0.3442

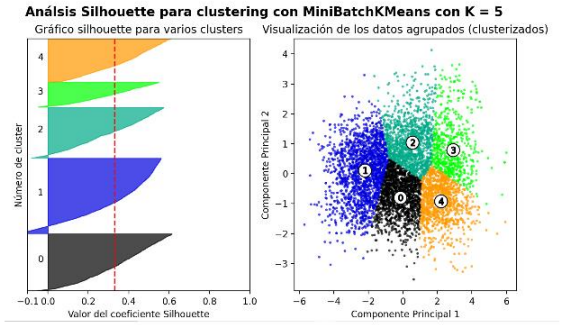
silhouette(K=3): 0.3398

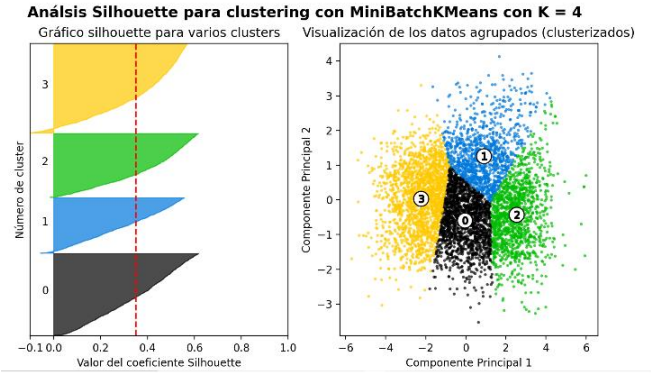
Por lo tanto, **el K que maximiza el rendimiento de K-Means es 4.** Nuevamente, desde el punto de vista del dominio, un K igual a 3 no aportaría mucho valor al análisis dado que agrupa básicamente por la componente 1.

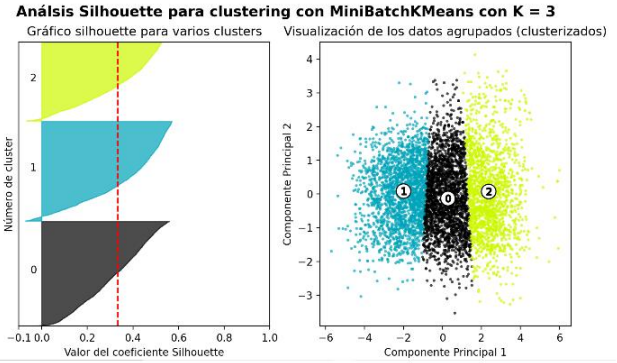
Algoritmo MiniBatchKMeans

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Valores del **coeficiente silhouette** para cada K – algoritmo MiniBatchKMeans

silhouette(K=6): 0.3086

silhouette(K=5): 0.3310

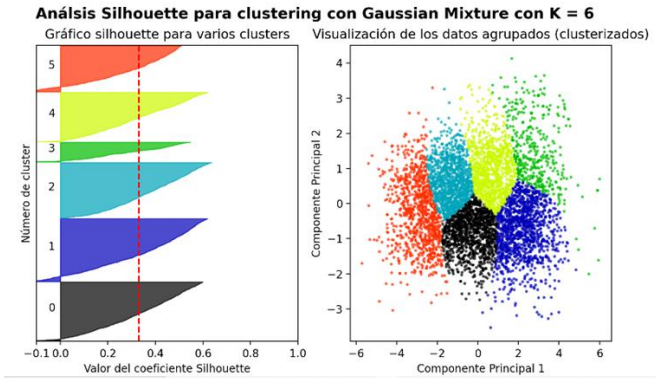
silhouette(K=4): 0.3502

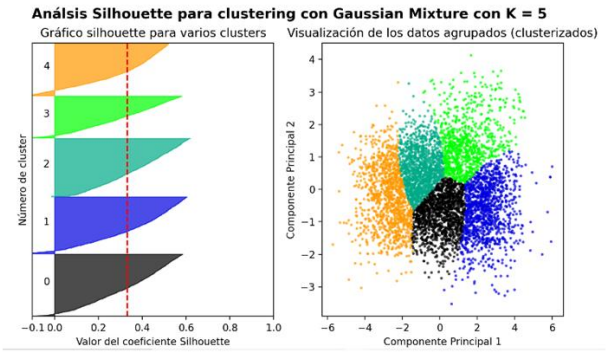
silhouette(K=3): 0.3344

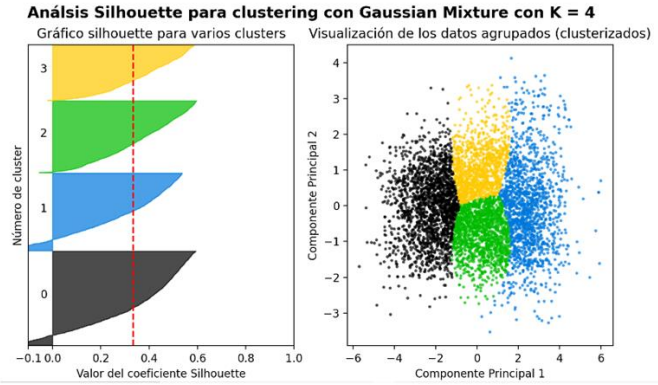
De acuerdo a los gráficos, los resultados son similares a los obtenidos con K-Means. Según los valores silhouette, **el K que maximiza el rendimiento de MiniBatchKMeans es 4.**

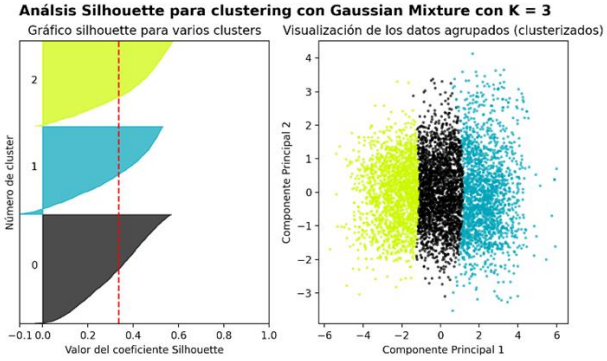
Algoritmo Gaussian Mixture

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Se observa que los gráficos están menos balanceados para los distintos valores de K – por lo tanto, los clusters no son tan homogéneos como con K-Means.

Valores del **coeficiente silhouette** para cada K – algoritmo Gaussian Mixture

silhouette(K=6): 0.3304

silhouette(K=5): 0.3328

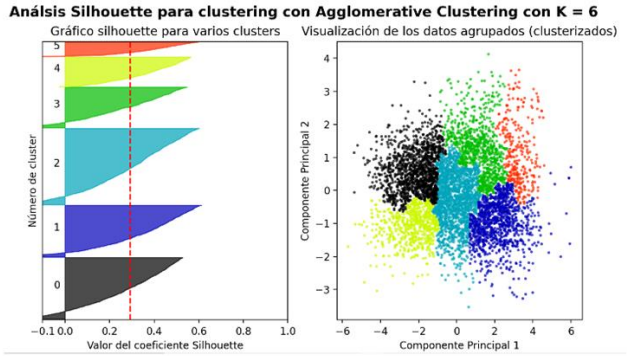
silhouette(K=4): 0.3337

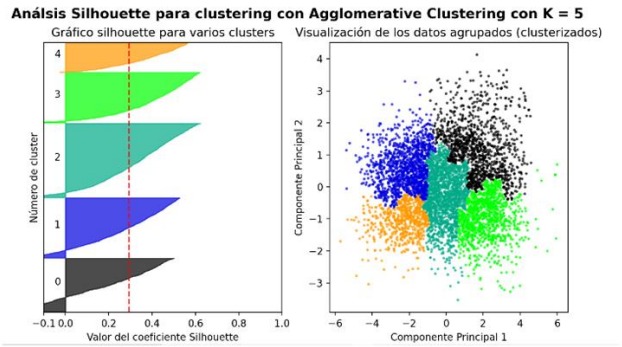
silhouette(K=3): 0.3362

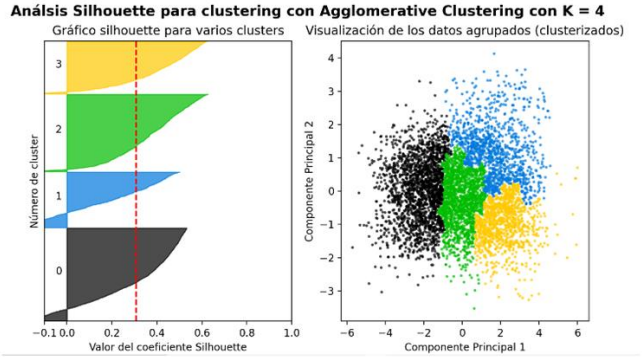
En este caso, **el K que maximiza el rendimiento de Gaussian Mixture es 3.** Aquí también, desde el punto de vista del dominio, un K = 3 no aporta valor significativo al clustering dado que la agrupación está dada principalmente por la componente 1.

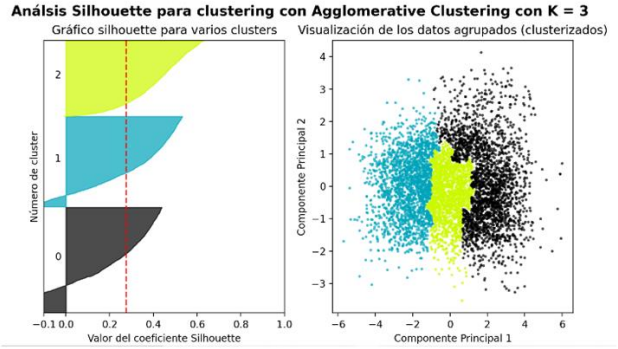
Algoritmo Agglomerative Clustering

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Se observa que los gráficos están menos balanceados para los distintos valores de K – por lo tanto, los clusters no son tan homogéneos como con K-Means.

Valores del **coeficiente silhouette** para cada K – algoritmo Agglomerative Clustering

silhouette(K=6): 0.2925

silhouette(K=5): 0.2942

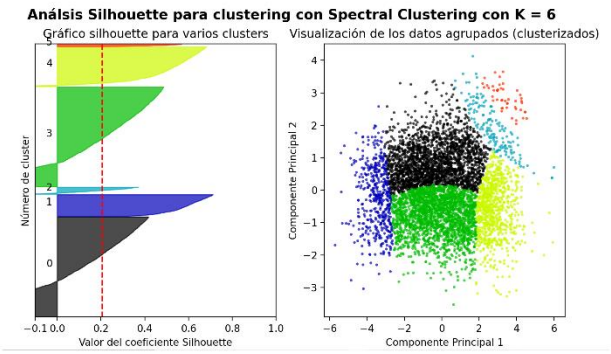
silhouette(K=4): 0.3075

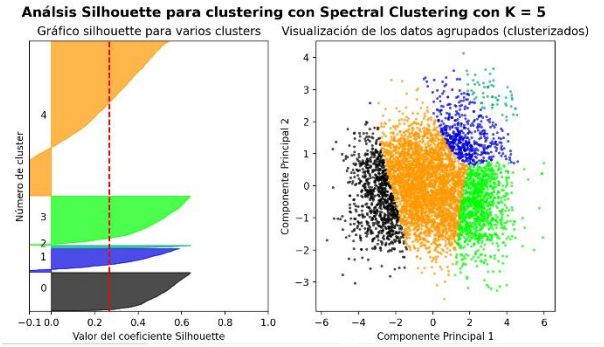
silhouette(K=3): 0.2759

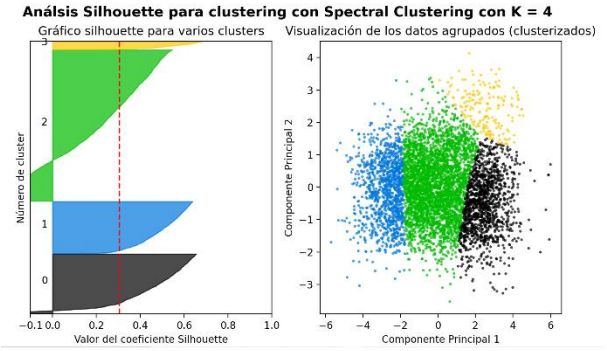
Por lo tanto, **el K que maximiza el rendimiento de Agglomerative Clustering es 4.**

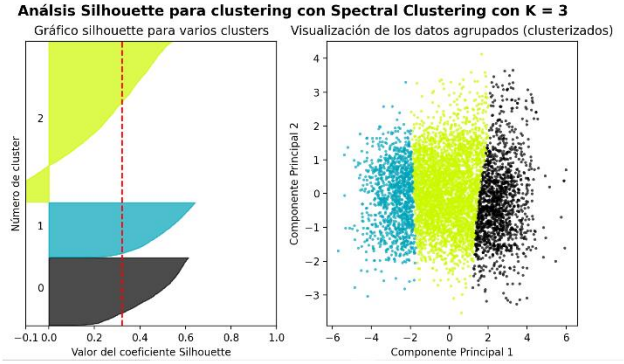
Algoritmo Spectral Clustering

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Se observa que los gráficos están menos balanceados para los distintos valores de K – por lo tanto, los clusters no son tan homogéneos.

Valores del **coeficiente silhouette** para cada K – algoritmo Spectral Clustering

silhouette(K=6): 0.2066

silhouette(K=5): 0.2671

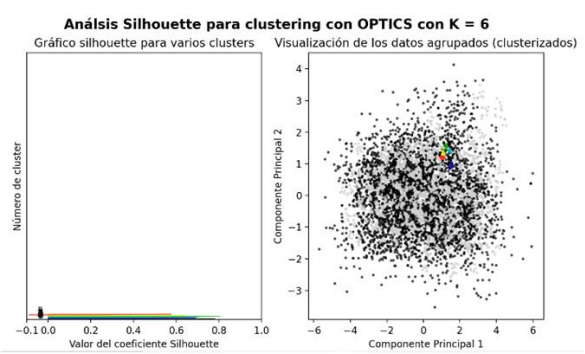
silhouette(K=4): 0.3052

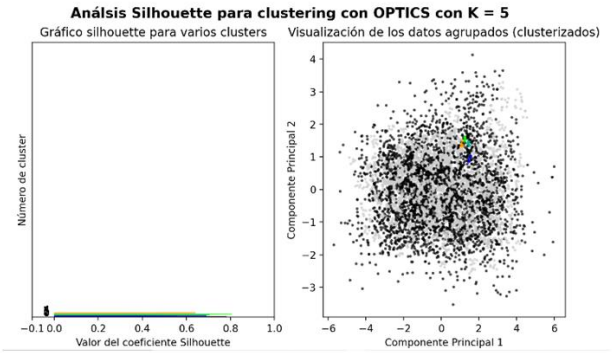
silhouette(K=3): 0.3212

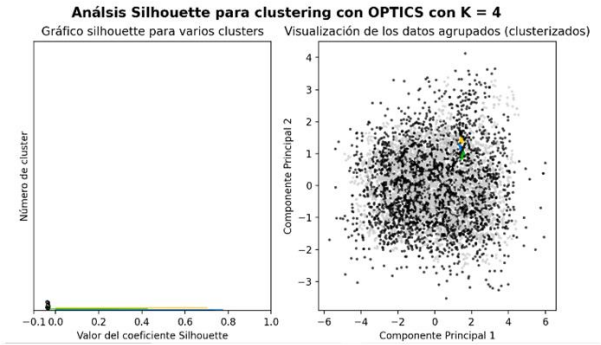
Por lo tanto, **el K que maximiza el rendimiento de Spectral Clustering es 3**. Aquí también, desde el punto de vista del dominio, un K = 3 no aporta valor significativo al clustering dado que la agrupación está dada principalmente por la componente 1.

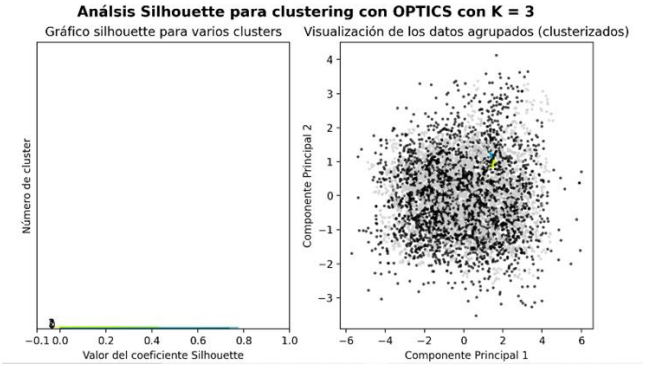
Algoritmo OPTICS

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Los gráficos muestran que el desempeño de este algoritmo no es bueno. Se revisarán los valores del coeficiente silhouette para cada K.

Valores del **coeficiente silhouette** para cada K – algoritmo OPTICS

silhouette(K=6): -0.2221

silhouette(K=5): -0.1810

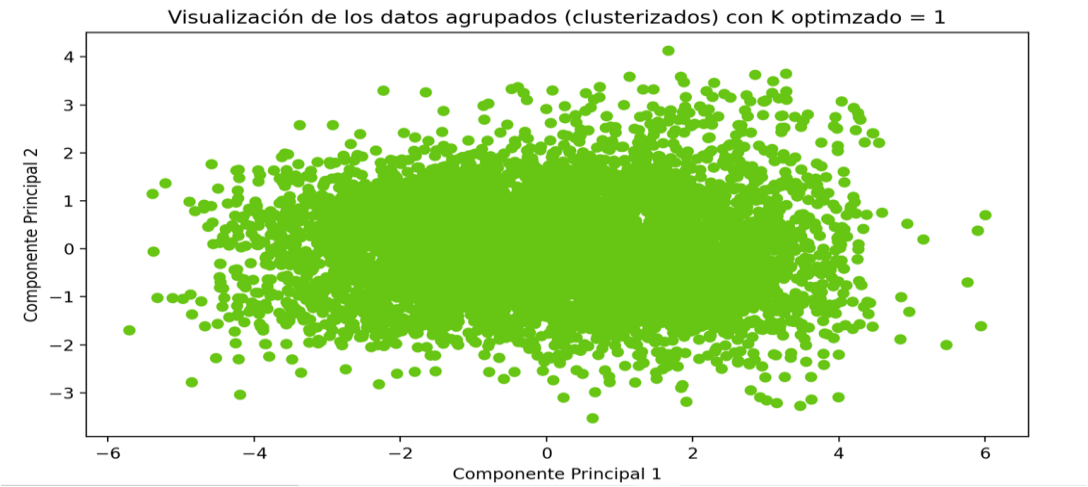
silhouette(K=4): -0.1598

silhouette(K=3): -0.1460

Por lo tanto, **el K que maximiza el rendimiento de OPTICS es 3.**

Algoritmo *Mean Shift*

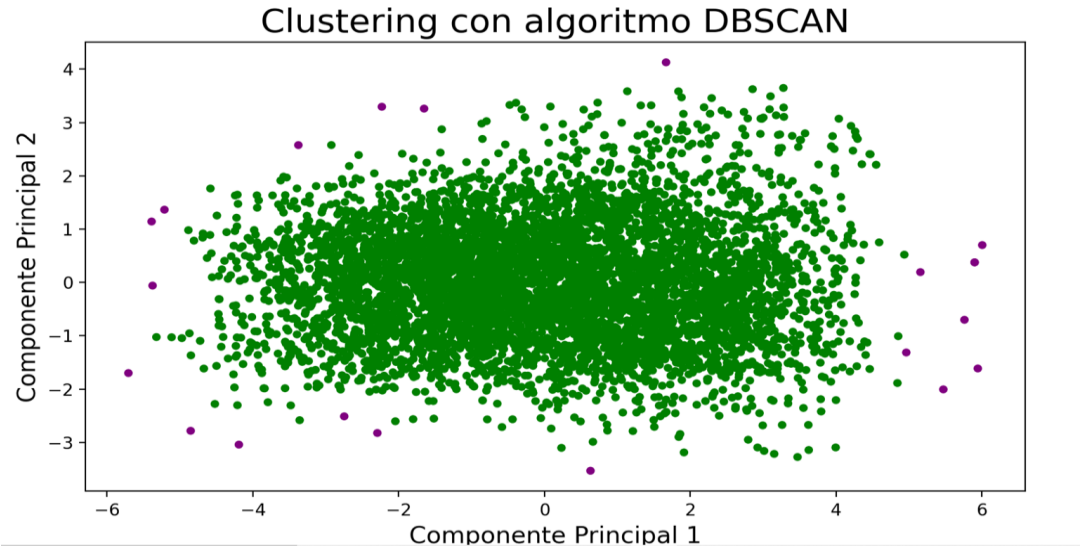
Gráfico con la visualización de los clusters obtenidos con Mean Shift.



En este caso, **el K que maximiza el rendimiento de Mean Shift es 1.**

Algoritmo DBSCAN

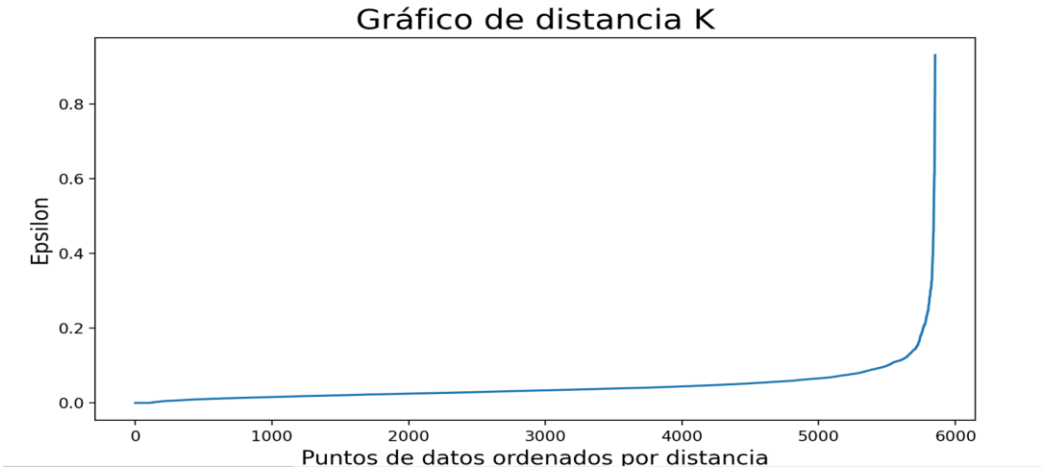
Gráfico con la visualización de los clusters obtenidos con DBSCAN sin optimizar ninguno de sus parámetros (epsilon & minPoints).



De acuerdo a la lógica de este algoritmo, todos puntos de datos fueron considerados como "ruido" porque sus parámetros no fueron optimizados.

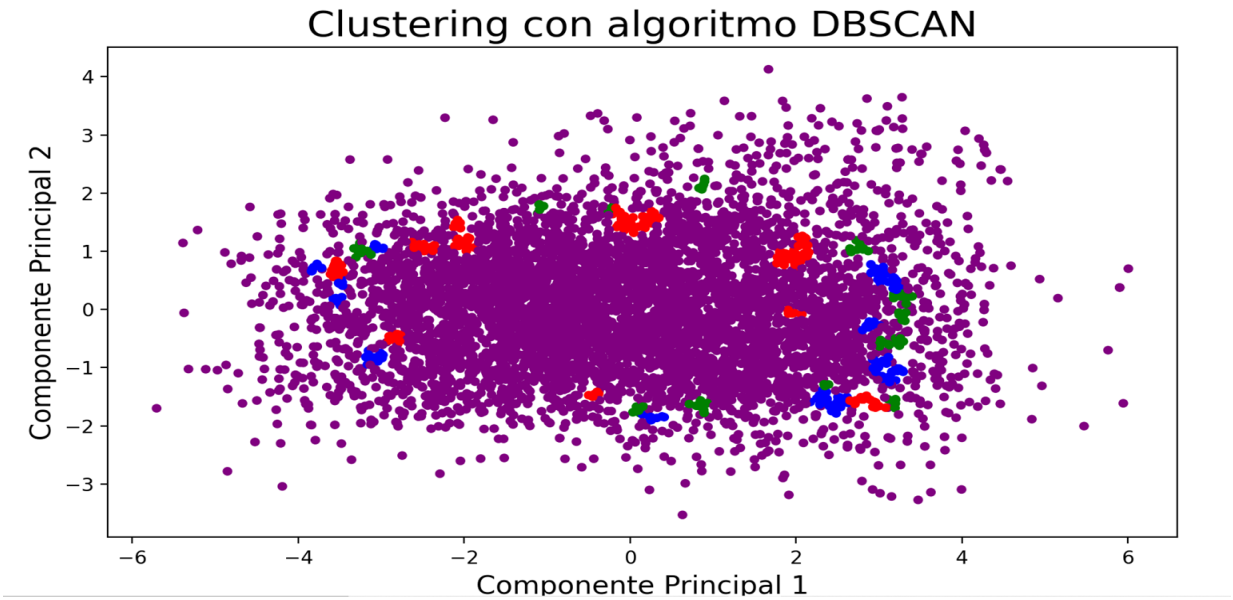
Para optimizar el parámetro epsilon se utiliza el gráfico de "distancia K".

Gráfico de "distancia K"



El valor óptimo de epsilon es el punto de máxima curvatura en el gráfico de distancia K, que en este caso sería 0,1. Respecto al parámetro minPoints, seleccionaremos el valor 6.

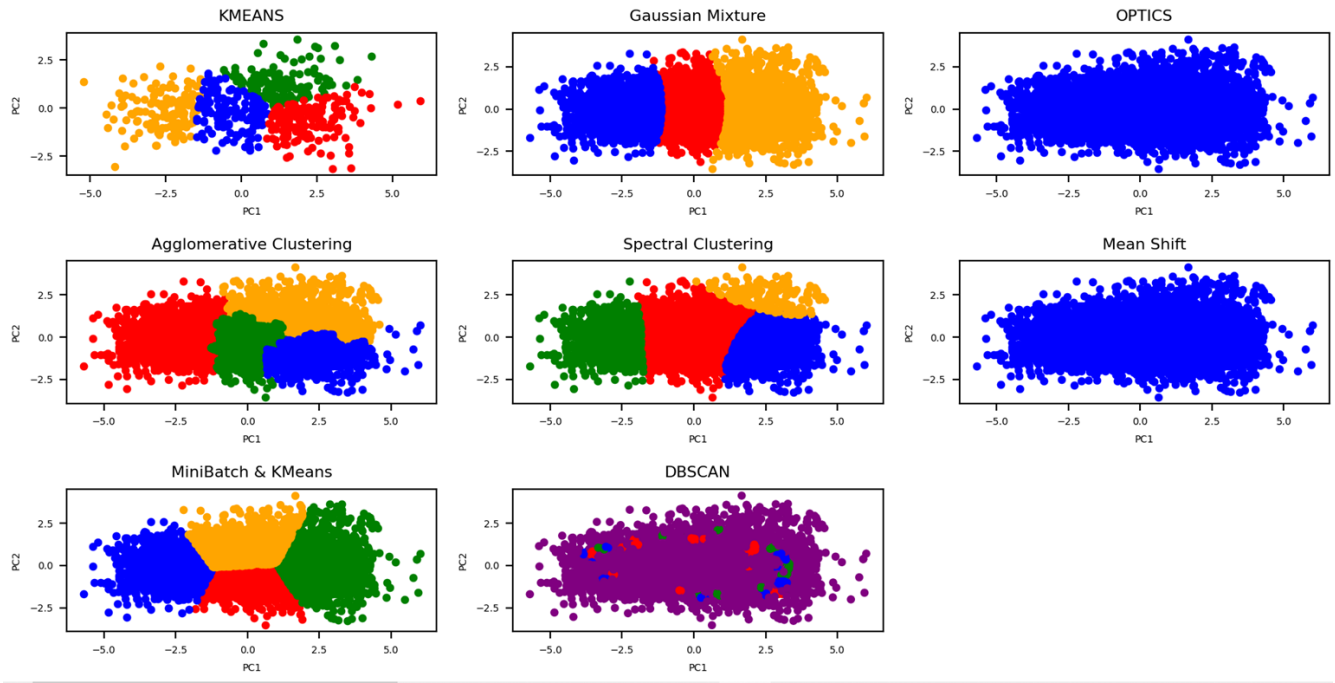
Gráfico con la visualización de los clusters obtenidos con DBSCAN.



A pesar de haber optimizado el parámetro epsilon, se observa que el desempeño de este algoritmo no es bueno.

Resultados de los algoritmos utilizados

Gráfica de todos los algoritmos utilizados – cada uno con su valor de K optimizado



Nuevamente, se destacan los algoritmos de bajo desempeño: Mean Shift, Optics y DBSAN.

Al mismo tiempo, los algoritmos Gaussian Mixture, Spectral y Agglomerative mostraron gráficas de Silhouette desbalancedas o con mucho peso sobre una de las componentes.

K-Means y MinibatchKMeans siguen siendo los algoritmos de mejor desempeño, sin embargo, la tendencia de los datos al clutestering para las 8 variables y los valores del coeficiente Silhouette en general no fueron mejores respecto a las primeras 2 estrategias.

Nuevamente, de acuerdo con las observaciones y resultados y obtenidos hasta aquí, se concluye que los clusters que se asignarían a los clientes tendrían origen en los segmentos generados por K-Means para un K = 4 en las estrategias 1 y 2.

7.5. Clustering sobre todo el Dataset – 1 etapa & aplicación de varios algoritmos

El objetivo de esta quinta estrategia es aplicar distintos algoritmos pero directamente sobre todo el Dataset, con sus variables categóricas y numéricas. En este caso, todas las variables serán previamente procesadas con FAMD (Análsis Factorial para Datos Mixtos – *Factor Analysis for Mixed Data*). Luego se correrán y compararán algoritmos de clustering a través de métricas. Finalmente, se identificará el mejor algoritmo para la estrategia.

FAMD es una técnica para convertir dataset con datos categóricos y continuos mixtos en componentes continuos. Permite la reducción de dimensiones, transformando datos complejos en subespacios de menor dimensión mientras se preservan características importantes del Datset original. Esta técnica es generalmente útil para reducir la complejidad de grandes sets de datos y apoyar la toma de decisiones.

Dataset origina con las 8 variables, 2 categóricas y 6 cuantitativas

	rango_precio	ZONA_PAS	Gama_Productos	Facturacion	Margen_Bruto	Facturacion_s_#Producto	Margen_s_Facturacion	Frecuencia_Compra
0	MIX	ZONA_VII	-0.255743	-0.289148	-0.666508	-0.090234	-1.001484	-1.332928
1	ALTO	ZONA_VIII	-0.255743	-0.199929	-0.731850	0.046116	-1.361575	-1.332928
2	MEDIO-ALTO	ZONA_VII	1.435807	2.075223	2.892588	1.545367	1.860283	1.869684
3	BAJO	ZONA_IV	1.205319	1.562803	1.585469	1.115388	-0.110198	1.442772
4	ALTO	ZONA_VIII	1.205319	0.695475	0.667641	-0.100325	-0.110198	1.092391
...	...	...	...	...	...	...	...	...
5848	MEDIO-ALTO	ZONA_VII	-0.255743	-0.465891	-0.410452	-0.363525	0.104838	0.347572
5849	MIX	ZONA_VIII	1.582221	0.709067	0.745429	-0.510249	0.035032	1.360361
5850	MIX	ZONA_VI	-0.634053	1.185097	1.616446	2.323935	1.094455	1.265815
5851	BAJO	ZONA_VI	0.784073	0.482037	-0.099980	0.040426	-1.361575	0.760638
5852	BAJO	ZONA_VI	0.600374	0.053188	0.085635	-0.409640	0.035032	0.347572

El siguiente cuadro muestra el eigenvalue, la variabilidad explicada por cada dimensión y la acumulada

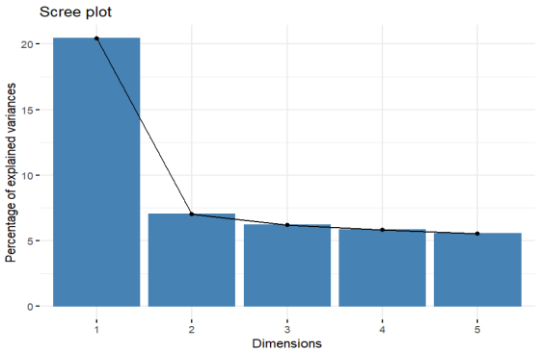
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	**3.878971**	20.415638	20.41564
Dim.2	**1.338748**	7.046043	27.46168
Dim.3	**1.176214**	6.190603	33.65228
Dim.4	**1.110393**	5.844172	39.49646
Dim.5	**1.053690**	5.545738	45.04219

Un eingenvalue mayor a 1 indica que el componente principal (Dim) representa más varianza que la explicada por las variables originales en los datos.

En este caso, sólo las primeras cinco componentes representan más varianza que cada una de las variables originales. Por otro lado, juntas representan solo el 45,04 % de la varianza total en el conjunto de datos.

Esto sugiere que los patrones y relaciones entre las variables probablemente no sean lineales y también sean complejos.

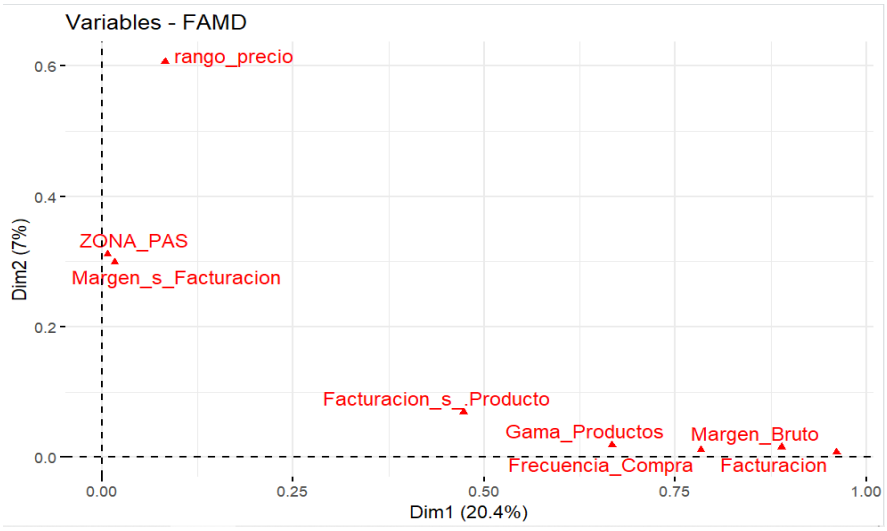
Varianza explicada por las 5 componentes



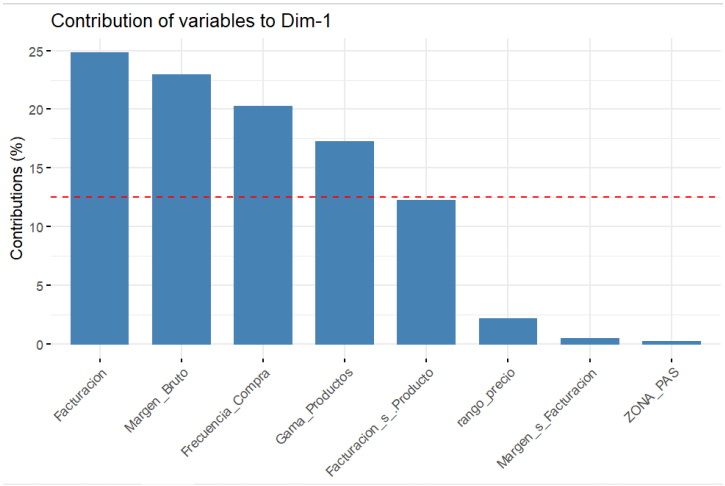
Matriz de rotación

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Gama_Productos	**0.66723607**	0.018337436	0.117180275	0.0043240390	0.000182168
Facturacion	**0.96097687**	0.007455114	0.003419111	0.0008934674	0.003890580
Margen_Bruto	**0.88902640**	0.015861314	0.056651534	0.0049830270	0.000964815
Facturacion_s_#Producto	**0.47288006**	0.068824519	0.168092295	0.0123301134	0.012236540
Margen_s_Facturacion	0.01603537	**0.299479441**	**0.273319571**	**0.079015933**	**0.06459336**
Frecuencia_Compra	**0.78321598**	0.011299545	0.013511615	0.0107195004	0.000727523

Contribución (proyección) de las variables del Dataset sobre las primeras 2 componentes



Contribución de las variables sobre la primera componente



Contribución de las variables sobre la segunda componente

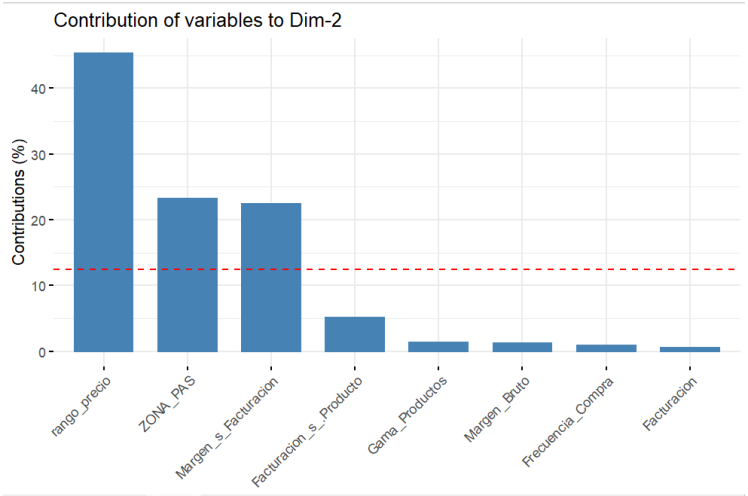


Gráfico de variables cuantitativas

De acuerdo a: (a) matriz de rotación y (b) contribución a las componentes principales

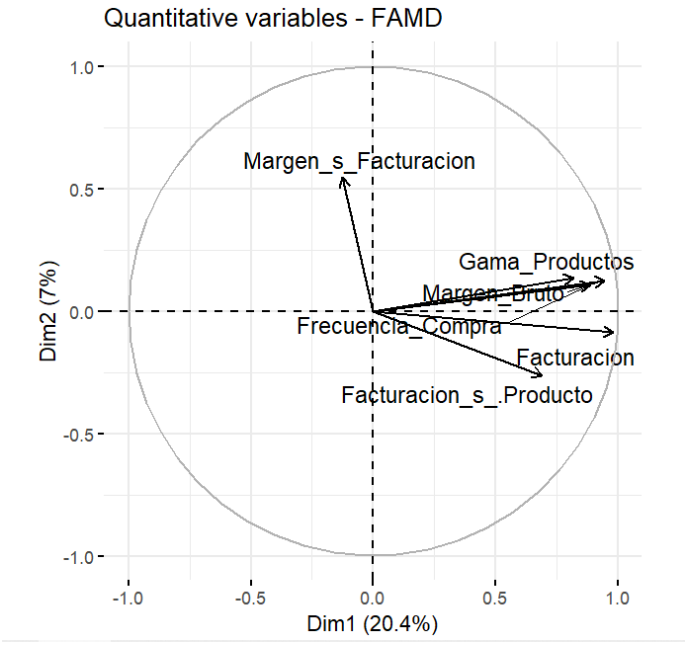
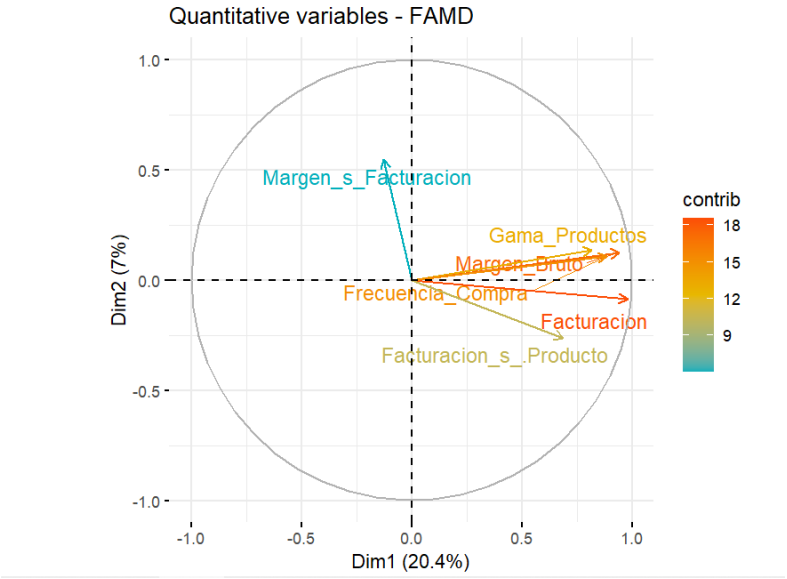


Gráfico de variables (círculo de correlación)

Muestra la relación entre variables, la proyección sobre las componentes principales y la correlación variables y la componente



Al igual que las variables numéricas, siguen los **resultados de las variables cualitativas**

Gráfico de variables cualitativas

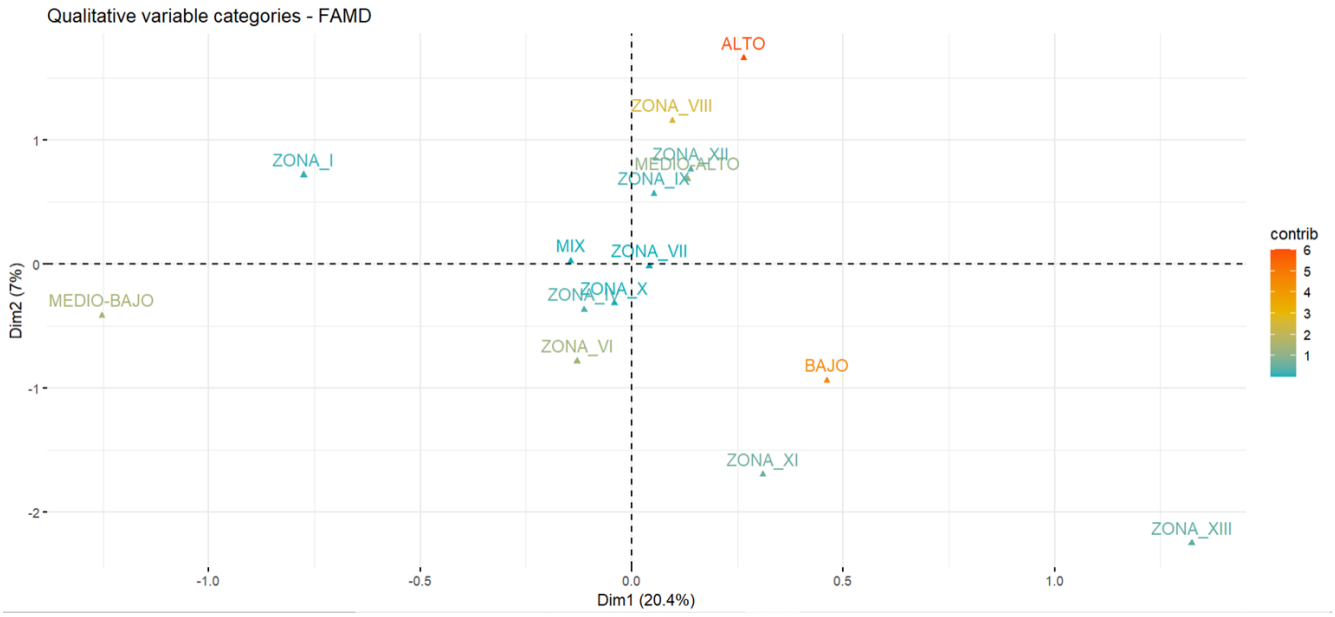
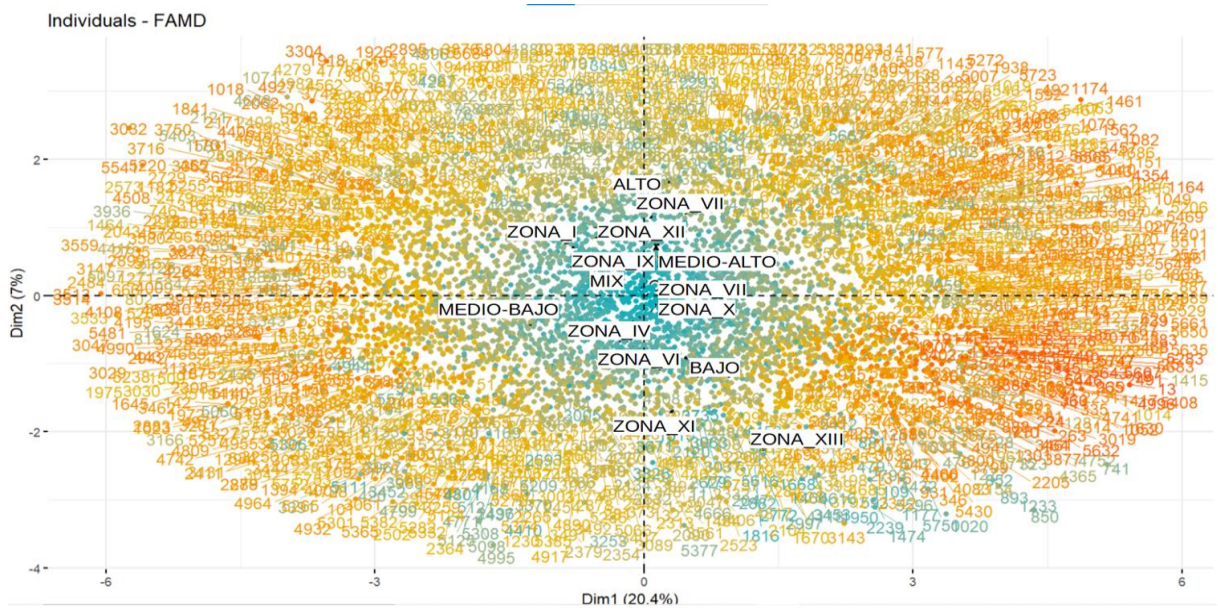
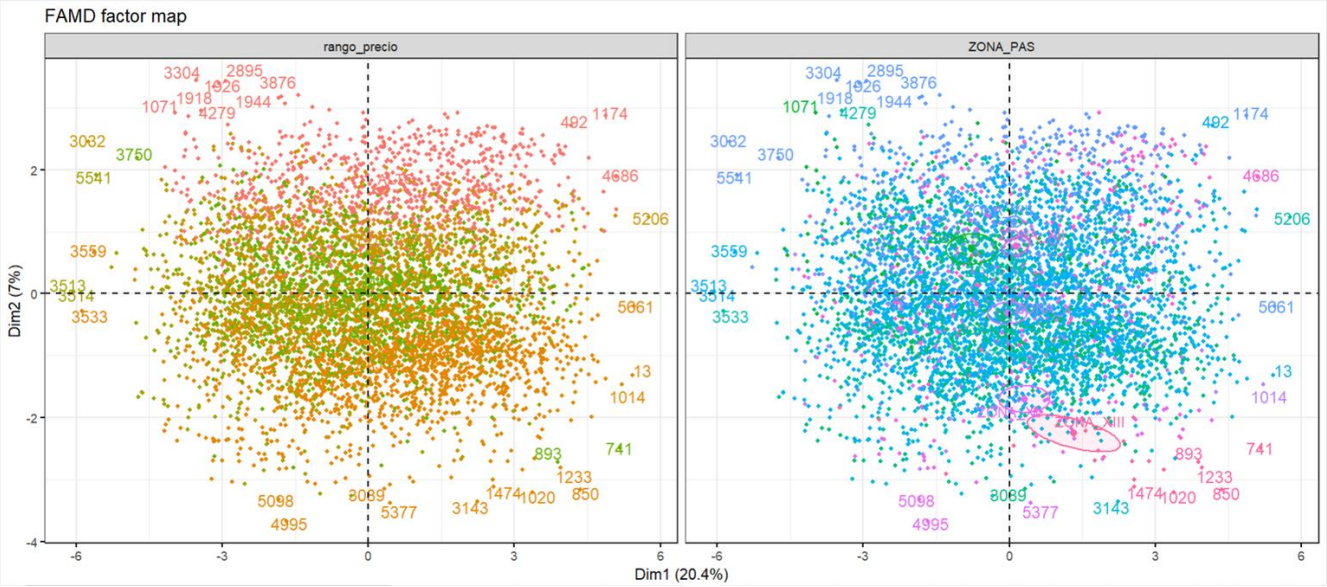


Gráfico por registro (cliente)



Mapa de factor

Muestra la cercanía entre cliente similares



Antes de correr diversos algoritmos, **se evaluará la tendencia de los datos al clustering**. Para ello, utilizaremos 2 métodos:

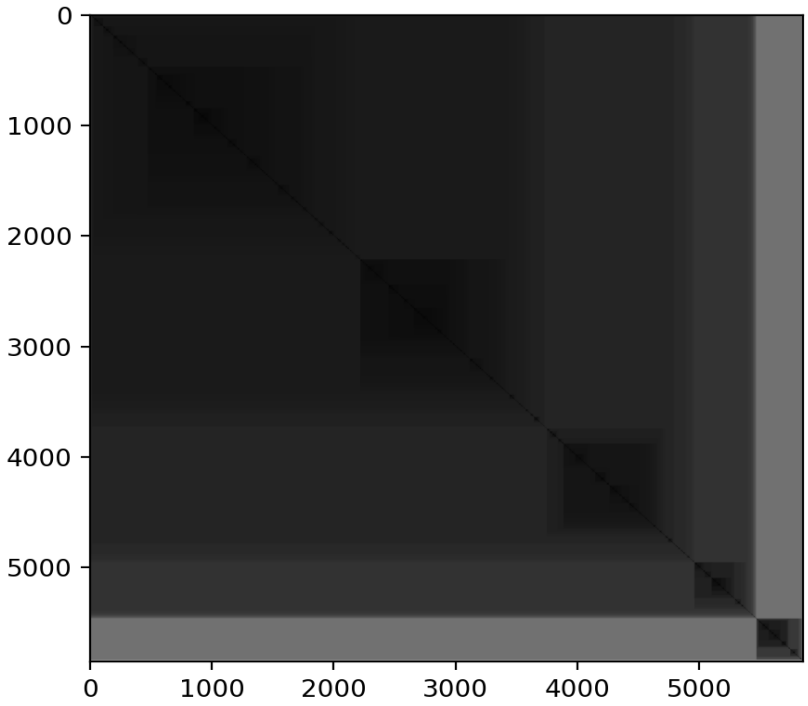
g) Métodos estadísticos: estadístico **Hopkins**

h) Métodos visuales: algoritmo **VAT** (*Visual Assessment of cluster Tendency*)

Valor Hopkins obtenido: (H) = 0.1663

En este caso el test es indeciso (sin embargo, los datos no son aleatorios)

VAT (*Visual Assessment of cluster Tendency*)

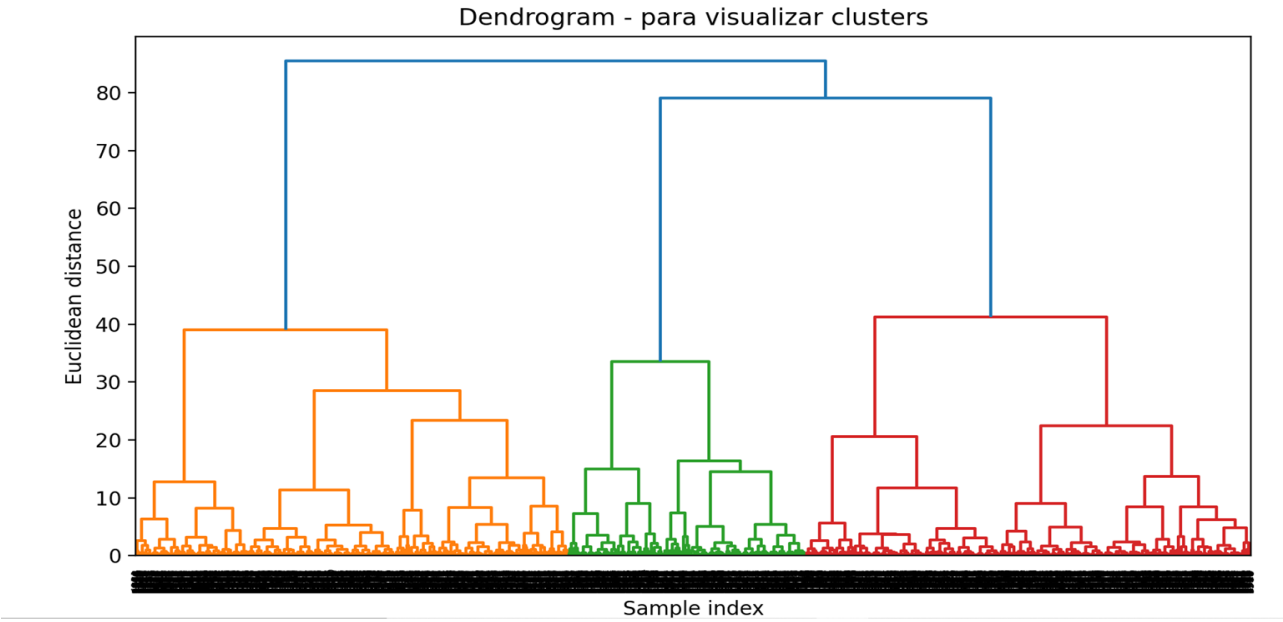


El gráfico muestra cuadrados negros que representan clusters. En este caso, podemos observar varios cuadrados, indicando la presencia poco clara de grupos o clusters en los datos.

Una vez reducida la cantidad de dimensiones a través de FAMD e interpretado sus resultados, sigamos analizado los datos para tratar de **identificar clusters valiosos**. Para ello, utilizaremos el algoritmo **clustering jerárquico** y **K-Means**.

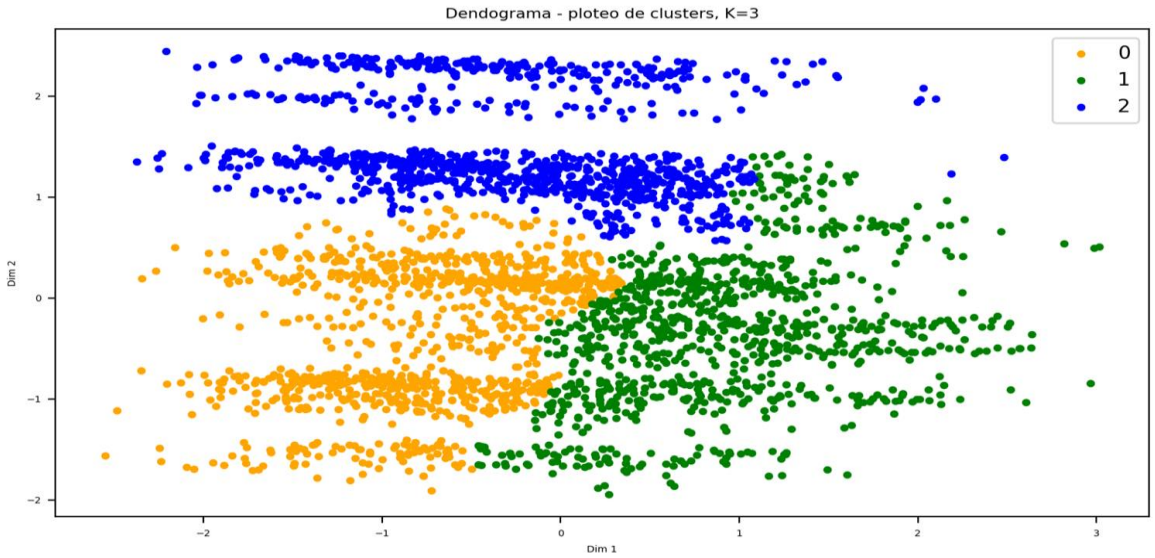
Clustering jerárquico

Dendograma



El **Dendograma** también muestra una clusterización de los datos. La cantidad de clusters que indica el gráfico es 3, en función de la altura de la mayor línea vertical – distancia euclidiana en el eje vertical.

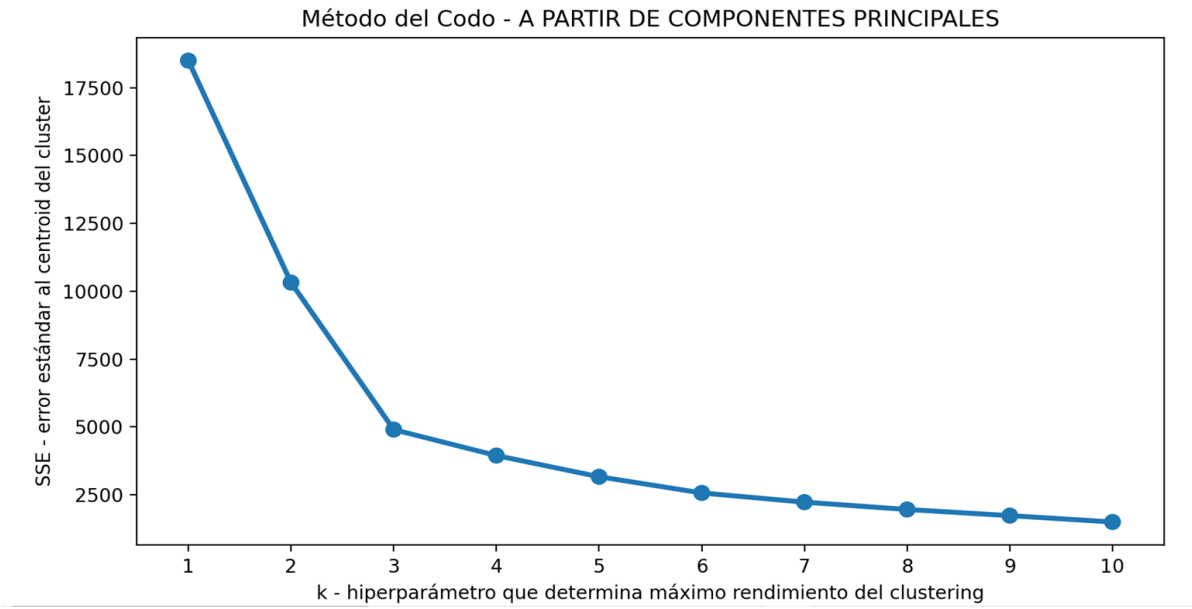
Se graficarán los clusters sugeridos por el dendograma sobre las primeras 2 Dimensiones.



Seguimos con otro algoritmo más complejo: K-Means

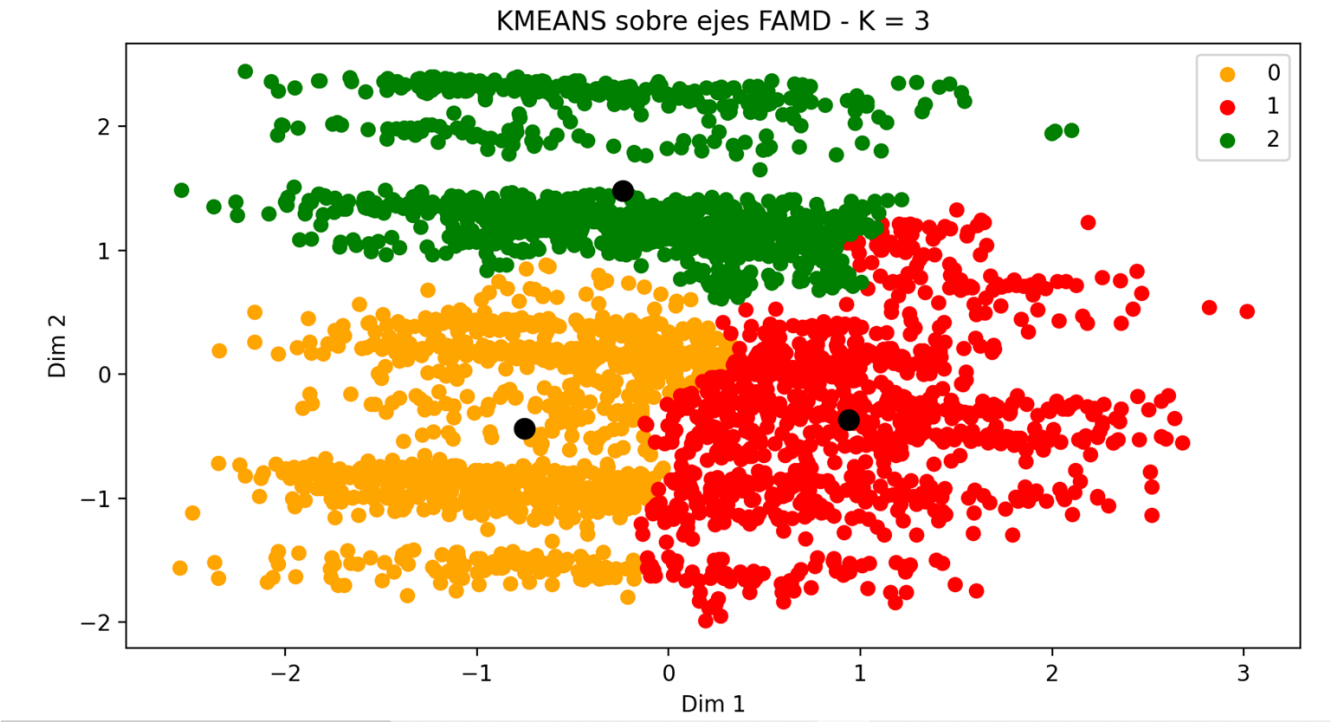
Antes de correr el algoritmo de clustering K-Means estimemos a través del método del codo (por medio del error estándar al centro del clúster - SSE) la cantidad "ideal" de clúster.

Gráfico del Codo



Según el gráfico de codo, a partir del 4to cluster ya no se obtienen saltos del SSE significativos, por lo tanto, **la cantidad de clusters a probar será 3**.

Clusters a partir de **K-Means** sobre Dimensiones FAMD



A continuación, sigue una tabla con las principales características **cuantitativas** de los 3 clusters generados por K-Means.

Clusters K-Means	Cantidad de Clientes	Cantidad de Clientes (%)	% Facturación	Facturación Promedio	Rentabilidad Promedio (%)	Frecuencia de Compra (prom.)	Gama de Productos (prom.)
0	1.809	31 %	67 %	30.775	21,2 %	12	14
1	2.708	46 %	8 %	2.486	26,7 %	3	3
2	1.336	23 %	25 %	15.684	27,1 %	8	10
	5.853						

De la tabla se observa que el cluester 0 se lleva casi el 70% de la facturación de La Empresa – similar a lo propuesto en la segmentación de la estrategia 2. Por oro lado, no habría una diferencia de rentabilidad significativa entre los segmentos. El resto de las variables gama de producto y frecuencia respetan la correlación con la variable facturación, de acuerdo al gráfico de variables.

Principales características **cualitativas** de los 3 clusters generados por K-Means.

Clusters vs. Rango de Precio

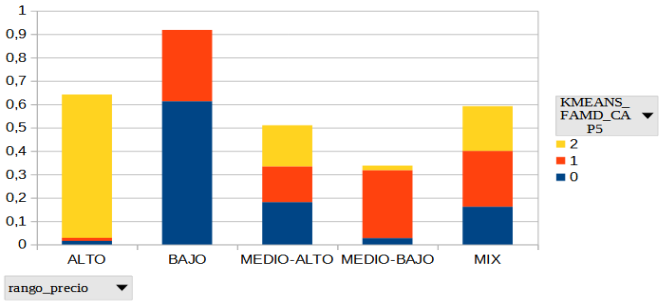


gráfico a

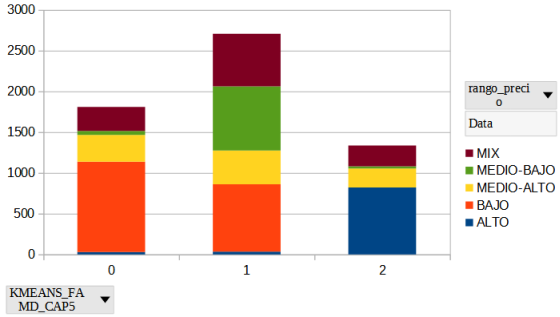
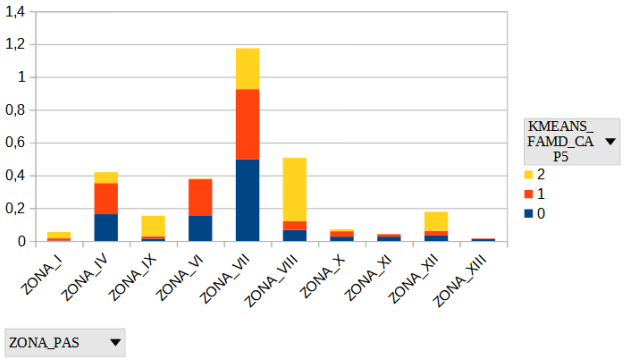


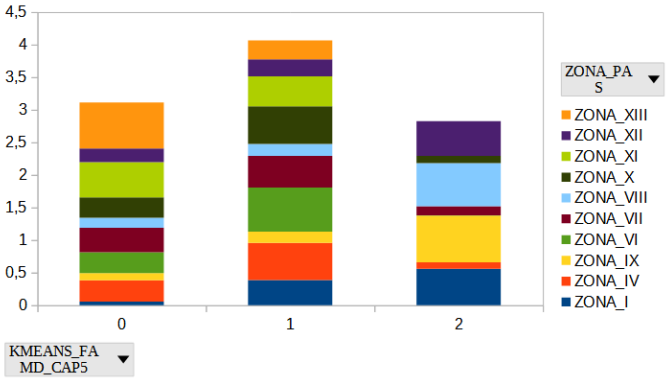
gráfico b

Count - KMEANS_FAMD	KMEANS_FAM	CAP5		
rango_precio	0	1	2	Total Result
ALTO	1,55 %	1,29 %	61,38 %	**15,09 %**
BAJO	61,30 %	30,50 %		**33,06 %**
MEDIO-ALTO	18,19 %	15,29 %	17,51 %	**16,69 %**
MEDIO-BAJO	2,76 %	29,03 %	2,02 %	**14,74 %**
MIX	16,20 %	23,89 %	19,09 %	**20,42 %**
Total Result	**100,00 %**	**100,00 %**	**100,00 %**	**100,00 %**

Se observa en el gráfico b que el segmento 2 absorbe a casi todos los clientes del rango de precio ALTO – más del 60% de los clientes del cluster 2 están en dicho rango de precios. La presencia de clientes de rango de precio BAJO es significativo en el cluster 0. En general, los segmentos contienen clientes de todos los rangos de precios.

Clusters vs. Zonas PAS





Count - KMEANS_FAMD	KMEANS_FAM	CAP5		
ZONA_PAS	0	1	2	Total Result
ZONA_I	0 %	1 %	4 %	**2 %**
ZONA_IV	16 %	19 %	7 %	**15 %**
ZONA_IX	1 %	1 %	13 %	**4 %**
ZONA_VI	16 %	22 %	0 %	**15 %**
ZONA_VII	50 %	43 %	25 %	**41 %**
ZONA_VIII	7 %	5 %	39 %	**13 %**
ZONA_X	3 %	3 %	1 %	**3 %**
ZONA_XI	3 %	2 %		**2 %**
ZONA_XII	3 %	3 %	12 %	**5 %**
ZONA_XIII	1 %	0 %		**1 %**
Total Result	**100 %**	**100 %**	**100 %**	**100 %**

Se observa que los segmentos 0 y 1 contienen clientes en todas las zonas PAS.

Analizaremos los clusters propuestos por K-Means en función de **todas las variables** utilizadas: cuantitativas y cualitativas.

Cluster 0 – es el cluster con mayor **facturación**, con el 67% de la venta de La Empresa. Por otro lado, es el grupo de clientes que más productos compra (gama de productos) y con mayor frecuencia de compra. Desde el punto de vista de negocio, este segmento puede ser valioso para trabajar el crecimiento de la gama de productos y rentabilidad en cada cliente. Nuevamente, es factible pensar para este segmento en otro tipo de abordaje comercial, por ejemplo: la figura del *Key Account*.

Cluster 1 – es el cluster con mayor cantidad de clientes, el 46% de la cartera. Se observa que la facturación promedio por cliente (compra por cliente) indica que la marca no está bien posicionada en los clientes del segmento. Por otro lado, las variables zona PAS y rango de precio tampoco aporta información valiosa para el desarrollo de estrategias comerciales. A prior, este segmento no aportaría información valiosa para La Empresa y se requeriría un **análisis** mucho más profundo para identificar aspectos en común entre clientes.

Cluster 2 – es el cluster con mayor rango de **precio**, con un 61% de clientes con precios en dicho rango de precios. La facturación promedio por cliente (compra por cliente) indica que la marca está bien posicionada en los clientes del segmento. Por lo tanto, es un segmento con potencial para invertir en un fuerte crecimiento de la marca a través de la gama de productos y también con servicio comercial fomentando la frecuencia de compra (contacto).

Como conclusiones hasta aquí, se observan sólo 3 cluster: 2 de ellos, sin nuevos aportes de información – *prácticamente se desprenden las mismas conclusiones mencionadas en las primeras dos estrategias* – y el cluster 1 no aporta valor.

A pesar de ello y sabiendo que la tendencia al clustering de los datos no fue la mejor respecto a otras estrategias, completaremos el análisis propuesto en el circuito de trabajo original.

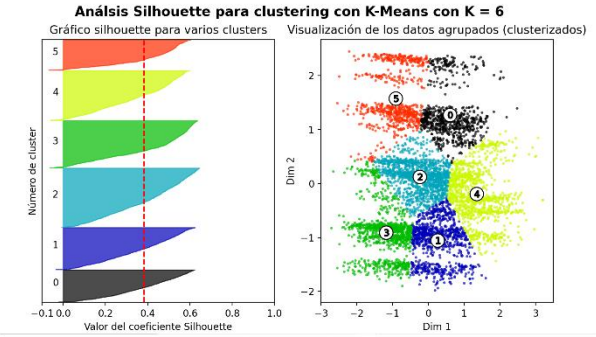
Para ello, previamente determinaremos el valor del hiperparámetro K para cada algoritmo mediante las gráficas de la **métrica Silhoutte**.

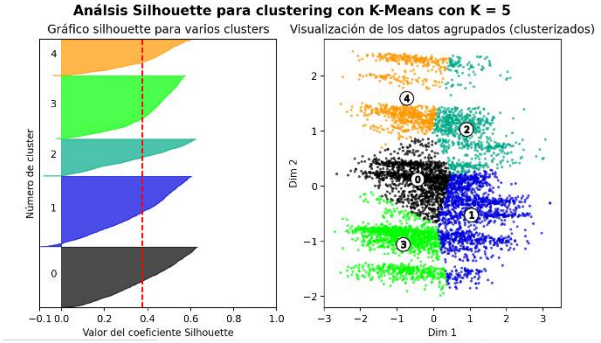
Algoritmos que se utilizarán:

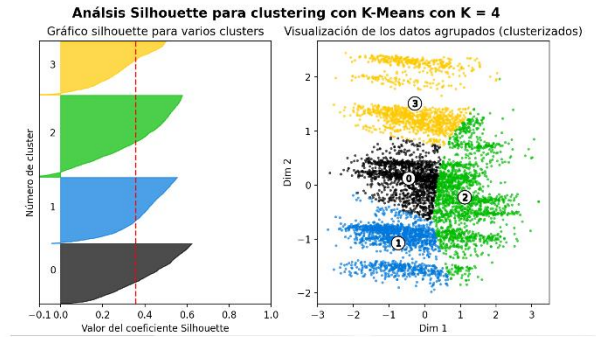
- *K-Means*
- *MiniBatchKMeans*
- GaussianMixture
- *AgglomerativeClustering*
- *SpectralClustering*
- *OPTICS*
- *MeanShift*
- *DBSCAN*

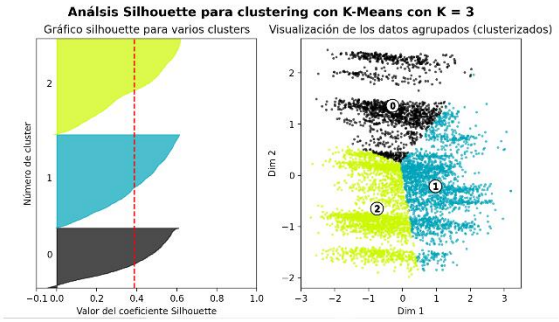
Algoritmo K-Means

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Se observa que todos los segmentos están bien balanceados de acuerdo con el gráfico de análisis silhouette.

Valores del **coeficiente silhouette** para cada K – algoritmo K-Means

silhouette(K=6): 0.3831

silhouette(K=5): 0.3755

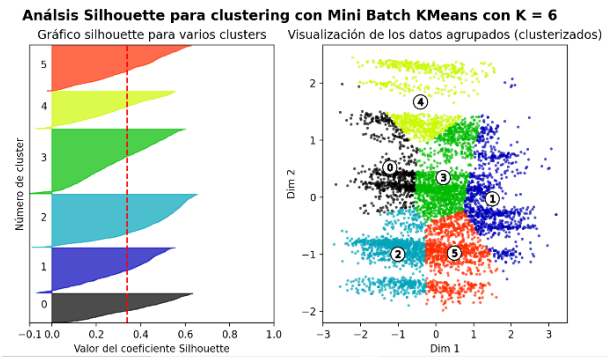
silhouette(K=4): 0.3576

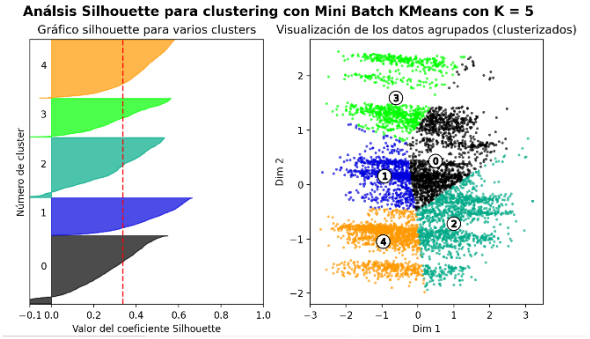
silhouette(K=3): 0.3875

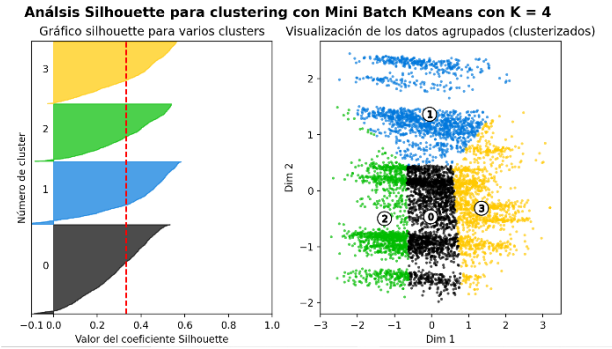
Por lo tanto, **el K que maximiza el rendimiento de K-Means es 3**.

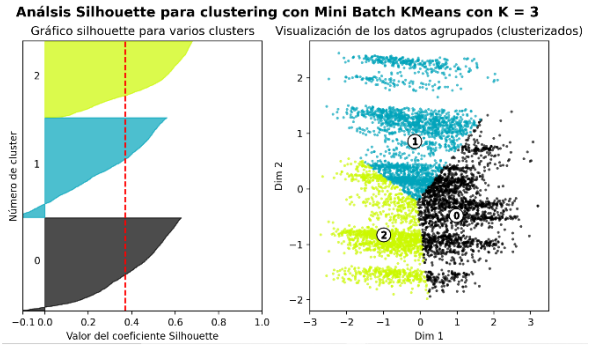
Algoritmo MiniBatchKMeans

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Valores del **coeficiente silhouette** para cada K – algoritmo MiniBatchKMeans

silhouette(K=6): 0.3390

silhouette(K=5): 0.3366

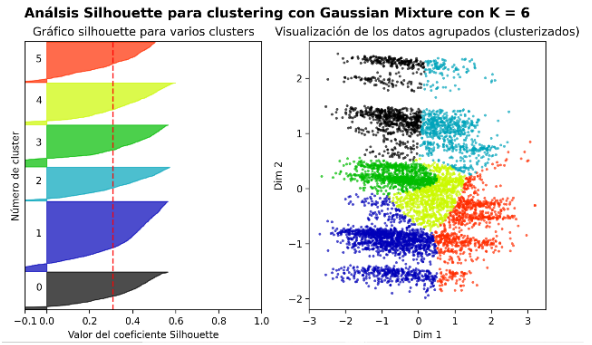
silhouette(K=4): 0.3320

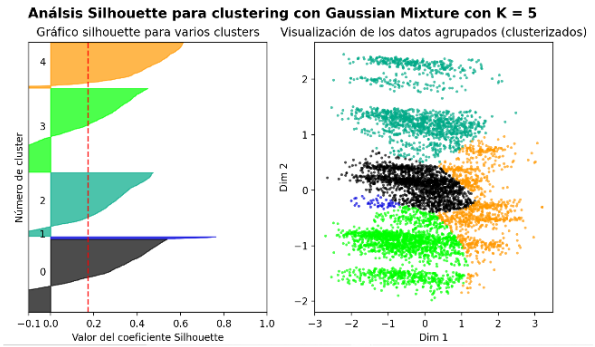
silhouette(K=3): 0.3705

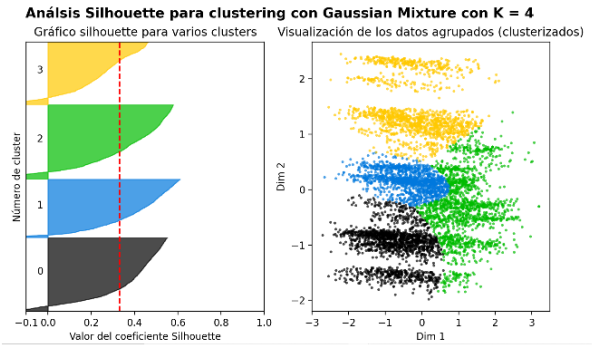
Por lo tanto, **el K que maximiza el rendimiento de MiniBatchKMeans es 3.**

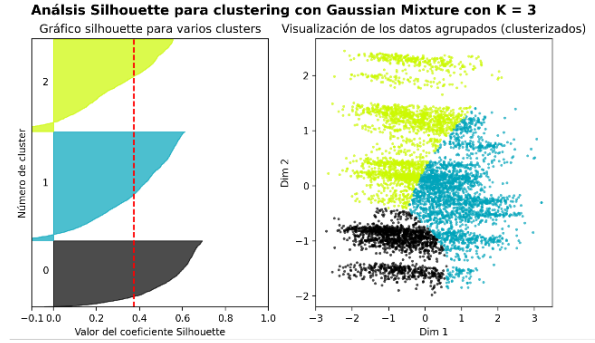
Algoritmo Gaussian Mixture

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Valores del **coeficiente silhouette** para cada K – algoritmo Gaussian Mixture

silhouette(K=6): 0.3073

silhouette(K=5): 0.1732

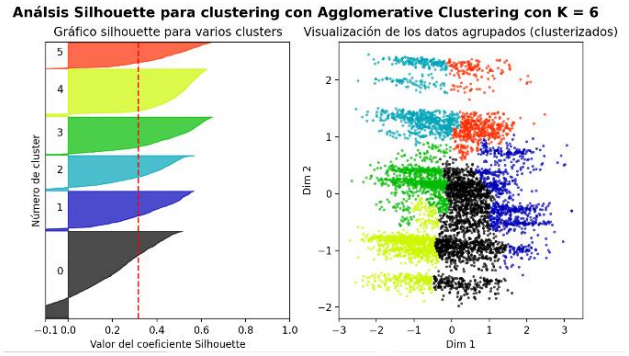
silhouette(K=4): 0.3308

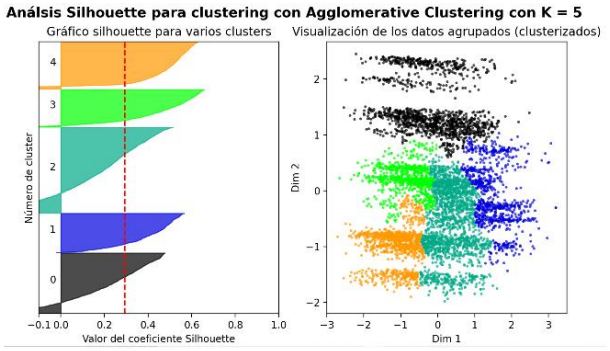
silhouette(K=3): 0.3743

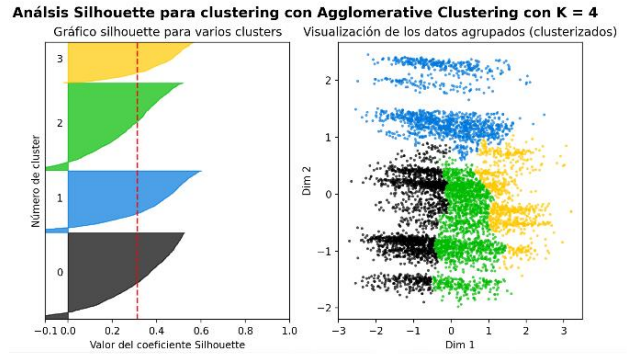
Por lo tanto, **el K que maximiza el rendimiento de Gaussian Mixture es 3.**

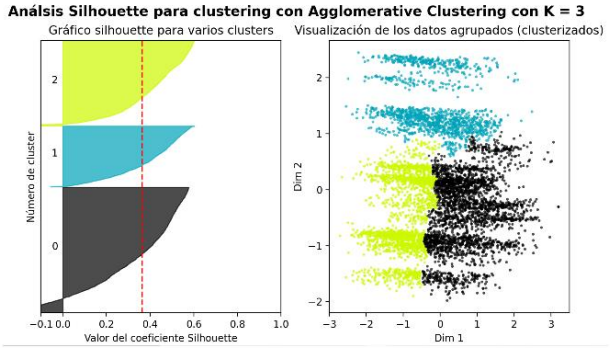
Algoritmo Agglomerative Clustering

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Valores del **coeficiente silhouette** para cada K – algoritmo Agglomerative Clustering

silhouette(K=6): 0.3168

silhouette(K=5): 0.2938

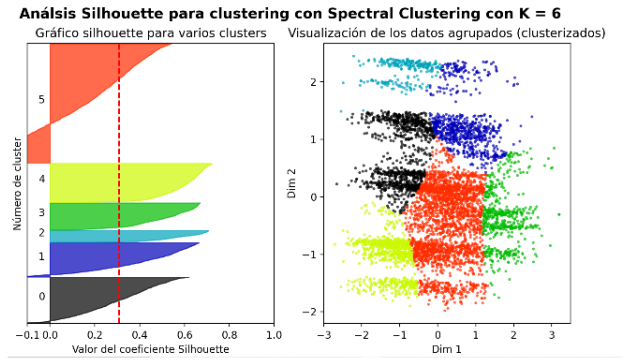
silhouette(K=4): 0.3123

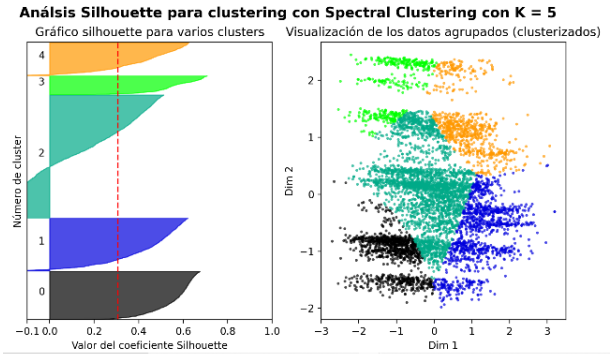
silhouette(K=3): 0.3635

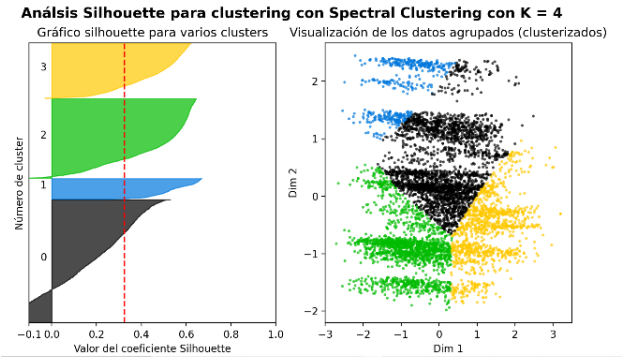
Por lo tanto, **el K que maximiza el rendimiento de Agglomerative Clustering es 3.**

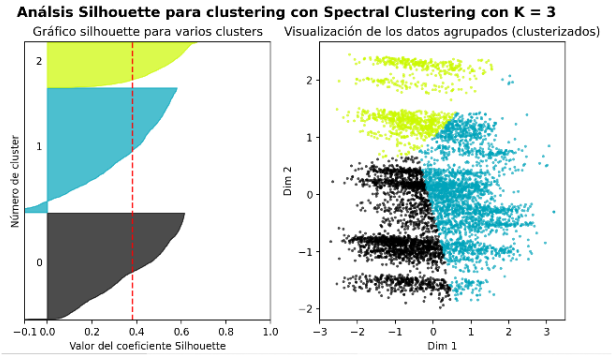
Algoritmo Spectral Clustering

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]









Valores del **coeficiente silhouette** para cada K – algoritmo Spectral Clustering

silhouette(K=6): 0.3083

silhouette(K=5): 0.3063

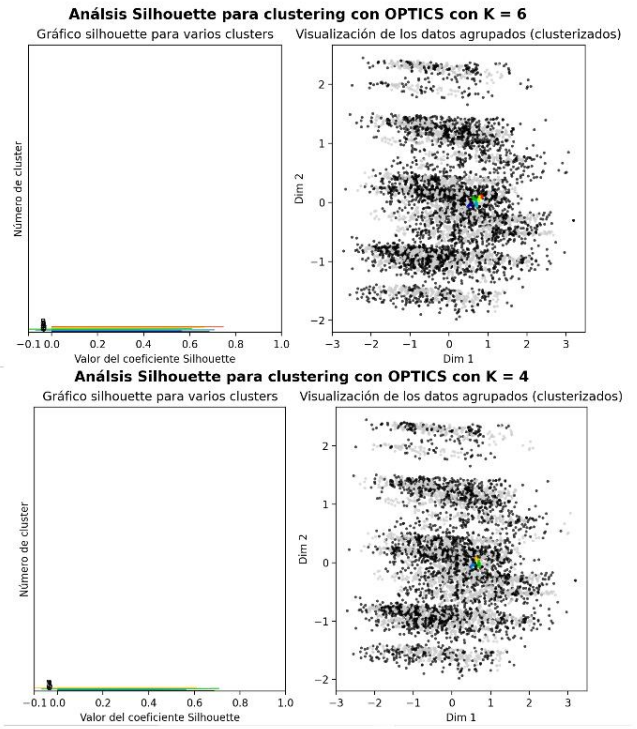
silhouette(K=4): 0.3250

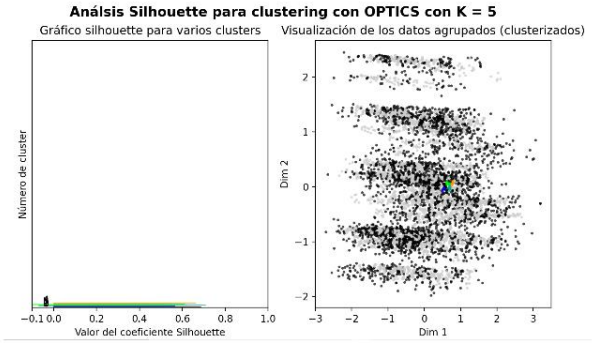
silhouette(K=3): 0.3811

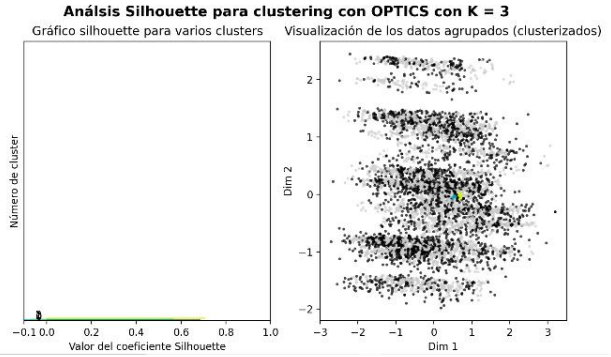
Por lo tanto, **el K que maximiza el rendimiento de Spectral Clustering es 3.**

Algoritmo OPTICS

Gráficos silhouette junto con la visualización de los clusters para cada valor de K [3,4,5,6]







Valores del **coeficiente silhouette** para cada K – algoritmo OPTICS

silhouette(K=6): -0.2151

silhouette(K=5): -0.1939

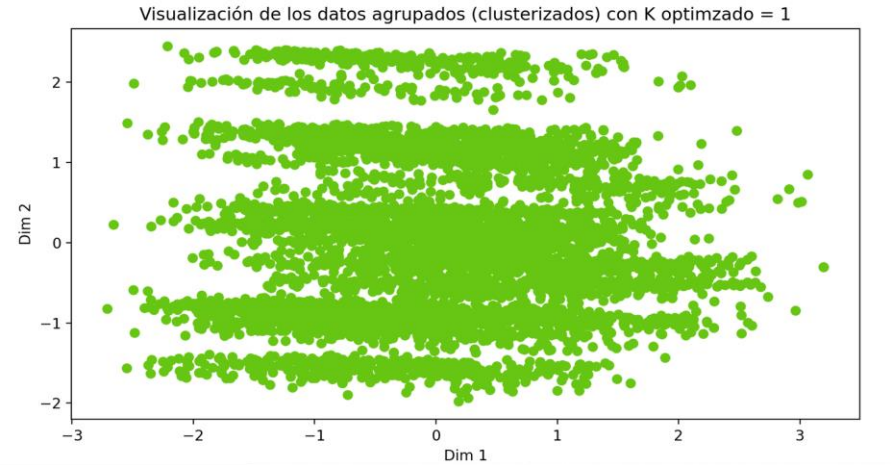
silhouette(K=4): -0.1608

silhouette(K=3): -0.1422

Por lo tanto, **el K que maximiza el rendimiento de OPTICS es 6.**

Algoritmo *Mean Shift*

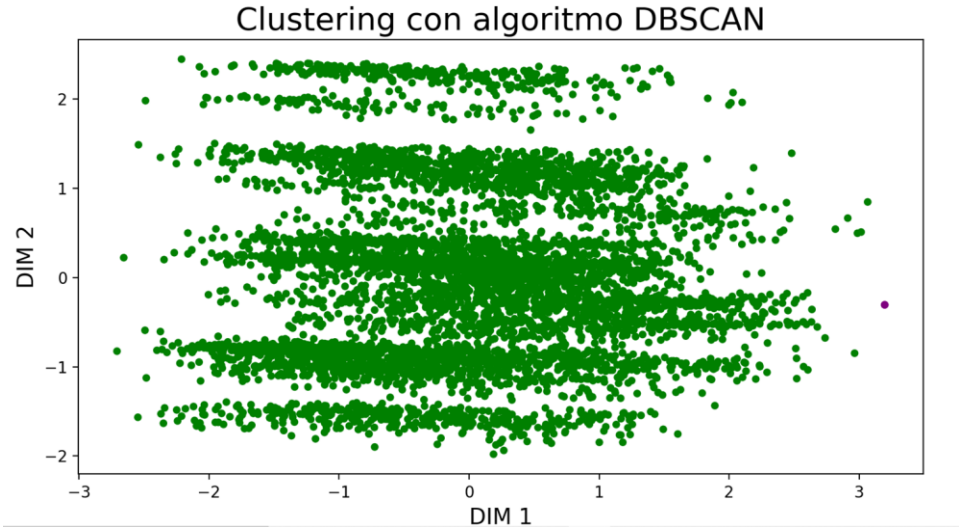
Gráfico con la visualización de los clusters obtenidos con Mean Shift.



En este caso, **el K que maximiza el rendimiento de Mean Shift es 1.**

Algoritmo DBSCAN

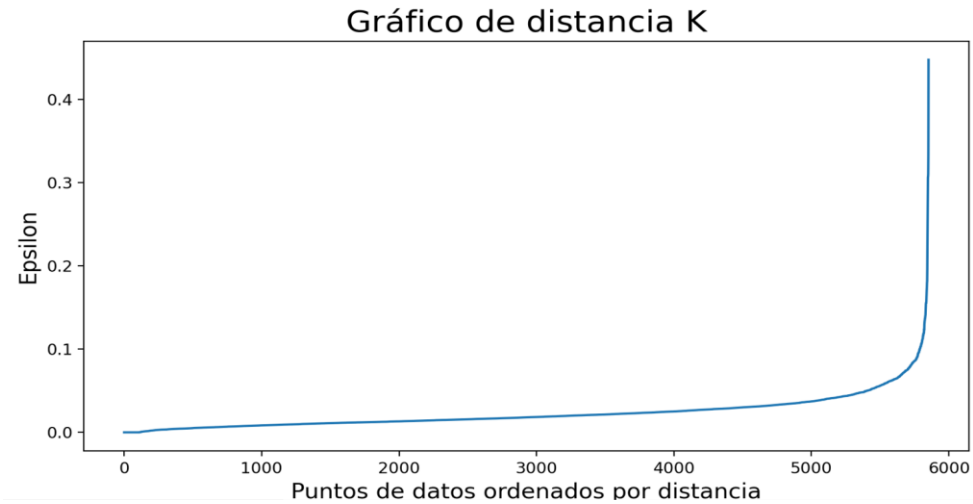
Gráfico con la visualización de los clusters obtenidos con DBSCAN sin optimizar ninguno de sus parámetros (epsilon & minPoints).



De acuerdo a la lógica de este algoritmo, todos puntos de datos fueron considerados como "ruido" por que sus parámetros no fueron optimizados.

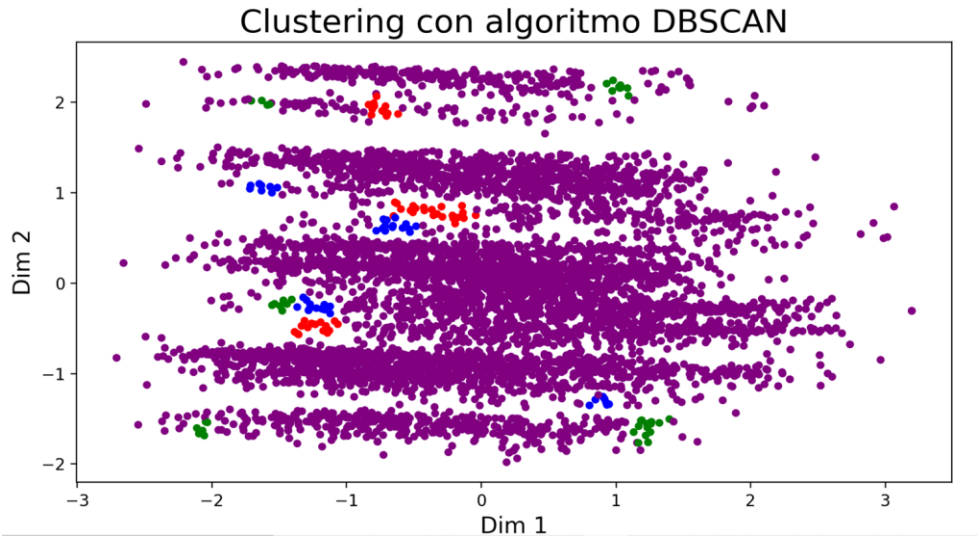
Para optimizar el parámetro epsilon se utiliza el gráfico de "distancia K".

Gráfico de "distancia K"



El valor óptimo de epsilon es el punto de máxima curvatura en el gráfico de distancia K, que en este caso sería 0,1. Respecto al parámetro minPoints, seleccionaremos el valor 6.

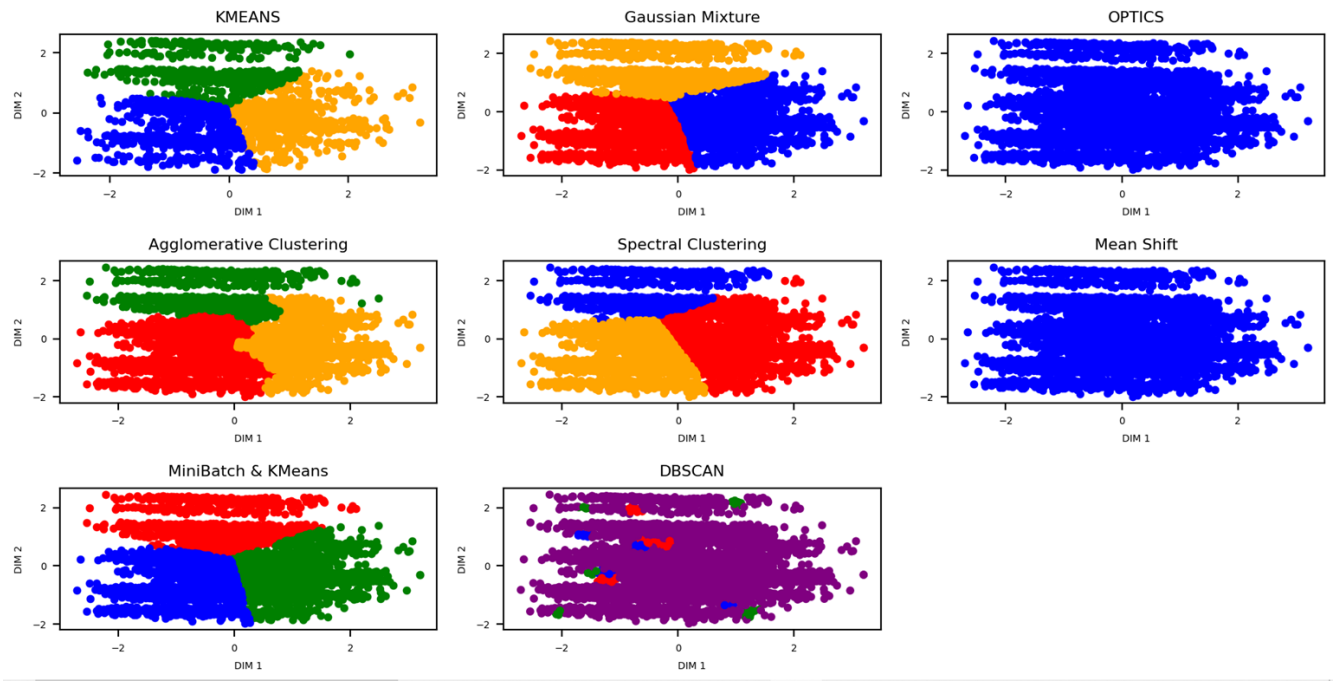
Gráfico con la visualización de los clusters obtenidos con DBSCAN.



A pesar de haber optimizado el parámetro epsilon, se observa que el desempeño de este algoritmo no es bueno.

Resultados de los algoritmos utilizados hasta aquí

Gráfica de todos los algoritmos utilizados – cada uno con su valor de K optimizado



Una vez más, se destacan los algoritmos de bajo desempeño: Mean Shift, Optics y DBSAN.

Los algoritmos Gaussian Mixture, Spectral, Agglomerative e inclusive MiniBatchKMeans mostraron gráficas de Silhouette desbalancedas o con mucho peso sobre una de las componentes.

Por otro lado, K-Means sigue siendo el algoritmo de mejor desempeño. Sin embargo, como se mencionó anteriormente, los clusters provistos no han aportado información nueva para La Empresa. Más aún, uno de los 3 clusters no aportó ningún valor.

Finalmente, la tendencia al clustering de los datos no fue la mejor respecto a otras estrategias.

8. Conclusiones Finales

El propósito de este trabajo fue acercar a La Empresa un proceso formal de segmentación de clientes asistido por una solución basada en algoritmos de *Machine Learning* (ML).

Para cumplir con el objetivo establecido inicialmente fueron planteadas 5 estrategias de trabajo que permitieron un desarrollo ordenado de las tareas de análisis, comparación de resultados y selección del mejor enfoque o estrategia de segmentación. Producto de trabajar dichas estrategias surgen las siguientes conclusiones:

- Los clusters obtenidos en la primera estrategia fueron los que aportaron la información más valiosa para La Empresa: segmentos **facturación**, **oportunidad**, **rentabilidad** y **revisión** – a partir de las 6 variables numéricas y con el uso del algoritmo K-Means sobre 2 componentes principales.

- Las variables categóricas tuvieron un aporte secundario. No obstante, el hecho de no tener un peso fuerte en la segmentación -especialmente la variable categórica Zona PAS- permitirá desafiar o cuestionar algunas de las prácticas vigentes, las cuales se apoyan en iniciativas estructuradas por zonas geográficas y no por criterios o variables que realmente establezcan una similitud entre clientes.

- No hay un único enfoque o algoritmo que resuelva una problemática de este estilo – no todos los algoritmos sirven para todos los Datasets. Mean Shift, Optics y DBSAN mostraron muy bajos desempeños. Por otro lado, K-Means y MiniBatchKMeans ofrecieron segmentaciones muy parecidas en algunos casos.

- La utilización de técnicas como PCA, MCA, FAMD fueron elementos claves para reducir la complejidad de los datos y facilitar la visualización, procesamiento de datos y comparación y validación de resultados.

Finalmente, remarcar que el punto de partida es muy relevante. Las variables del Dataset fueron seleccionadas y generadas para maximizar la calidad de los resultados. Sin embargo, siempre habrá oportunidad para incorporar nuevas variables que mejoren el proceso de segmentación, por ejemplo: área disponible para siembra, el vendedor asignado al cliente ,antigüedad de los clientes, entre otras.

9. Referencias-Bibliografía

1) Chinedu Pascal Ezenkwu, Simeon Ozuomba, Constance kalu (2015). Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services. IJARAI, International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.10.

2) Cormac Dullaghan and Eleni Rozaki (2017). Integration of machine learning technics to evaluate dynamic customer segmentation analysis for mobile customers. IJDKP, International Journal of Data Mining & Knowledge Management Process,Vol.7, No.1.

3) Kishana R. Kashwan, Member, IACSIT, and C. M.Velu (2013). Customer Segmentation Using Clustering and Data Mining Techniques – International Journal of Computer Theory and Engineering. Vol. 5, No. 6.

4) Diego Zumarraga Mera (2020). Machine Learning. Segmentación de clientes usando ML.NET. Acelera.Tech. Recuperado de https://acelera.tech/2020/01/07/machine-learning-segmentacion-de-clientes-usando-ml-net/

5) PREDICTLAND (enero 2018). Big Data: Segmentación de Clientes. Recuperado de https://www.predictland.com/big_data_segmentacion_clientes/