Proyecto Integrador - Maestria en Ciencia de Datos y Analitica
Reporte Técnico y de Modelos

Título del proyecto

Análisis de datos de tiendas de comercio electrónico para segmentación de clientes

Repositorio GitHub

https://github.com/jarcil13/PI-EAFIT-DS2020

Integrantes:

Juan David Arcila Moreno, jarcil13@eafit.edu.co

Diego Alejandro Cifuentes García, dcifuen3@eafit.edu.co

Sebastián Patiño Barrientos, spatino6@eafit.edu.co

Universidad EAFIT

Medellín

2020

Descripción del problema:

En la industria del comercio electrónico es importante conocer las características de los clientes que adquieren productos o servicios con la compañía, esta información se vuelve el insumo principal para diseñar y ejecutar estrategias de fidelización para clientes mediante campañas de marketing dirigidas. El problema radica en la capacidad de realizar procesos de analitica efectivos sobre la gran cantidad de datos disponibles y que de sus resultados surja información significativa para realizar análisis que aporten valor a la compañía.

Para esto, decidimos realizar un proceso de analítica para el sector de comercio electrónico partiendo de conjuntos de datos públicos que permitieran la experimentación con diversas técnicas estadísticas que luego pudiesen ser aplicadas a los conjuntos de datos de una compañía grande como el Éxito.

Metodología: CRISP-DM

1. **Entendimiento del negocio:**

   En el contexto de las empresas de comercio electrónico el objetivo principal que deseamos desarrollar en este proyecto es mejorar la retención de los clientes actuales y tener información sobre clientes que se puedan atraer al negocio.

   Para el desarrollo de este objetivo la pregunta principal que queremos resolver en este contexto es “¿Qué tipos de clientes es posible encontrar en nuestro negocio?” junto a otras como “¿Cuál es el comportamiento de compras que tienen nuestros clientes?”. Basándonos en estas preguntas tomaremos los siguientes pasos como parte del plan para este proyecto: primero se realizará un análisis superficial de los datos para tener conocimiento de su naturaleza y principales características previo a realizar análisis más profundos. Luego de esto haremos los procesos de transformación de los datos necesarios para obtener tablas de datos limpias que puedan ser utilizadas después en un análisis de correlaciones y modelos estadísticos. Una vez se tengan resultados aceptables el código obtenido será estructurado siguiendo buenas prácticas de programación y estilo para después exponerlo como un modelo que puede ser usado posteriormente con datos de otras empresas de comercio electrónico.

   Cómo criterio de éxito se espera obtener un análisis que permita conocer qué tipos de clientes se tiene en la empresa de comercio electrónico y cómo se puede clasificar su comportamiento de compras, información que es de

utilidad para diseñar campañas de marketing que posiblemente pueda atraer una empresa de comercio electrónico.

2. Entendimiento de los datos:

Dada la naturaleza académica de este proyecto utilizaremos las siguientes fuentes de datos relacionadas con la problemática que quisimos resolver:

2.1. Ecommerce: https://www.kaggle.com/carrie1/ecommerce-data

2.2. Instacart: https://www.kaggle.com/c/instacart-market-basket-analysis/data

El primer conjunto de datos pertenece a una tienda electrónica localizada en el Reino Unido. En este se presentan las ventas realizadas entre el 1/12/2009 y el 09/12/2011. Esta compañía se enfoca en ventas de regalos para toda ocasión.

El segundo conjunto de datos pertenece a otra tienda electrónica localizada en los Estados Unidos, enfocada en la venta de productos de la canasta familiar al menudeo en línea.

Para el objetivo de resolver la pregunta de negocio planteada en el paso anterior vamos a asumir que los conjuntos de datos que tenemos provienen de la misma empresa de comercio digital y que para efectos prácticos realizaremos dos análisis con enfoques diferentes, usando los conjuntos de datos *ecommerce-data* e *instacart*: en el primero nos enfocaremos en los precios de los artículos adquiridos por los clientes y se busca entender cómo su patrón de gastos es similar o diferente a los otros clientes; en el segundo se buscar hacer énfasis en los artículos, donde se busca conocer la relación que tiene artículos que se adquirieron en cada transacción y cuales características comparten estos artículos en el historial de transacciones de un cliente.

Idealmente estos dos conjunto de datos serían mucho más grandes y provendrían de una sola organización, en caso se vuelve necesario el uso de almacenamiento distribuido y frameworks para el procesamiento de grandes volúmenes de datos, sin embargo, debido a que estos conjuntos de prueba no los son, no es necesario utilizar dichas técnicas para el desarrollo de este proyecto en específico. De ese modo, es importante resaltar que se espera que las transformaciones y análisis desarrollados para este proyecto sean igualmente válidos para un conjunto de datos del caso ideal, en ese caso

sería necesario adaptar el código resultante a los frameworks que se integren con la fuente de datos que se va a usar.

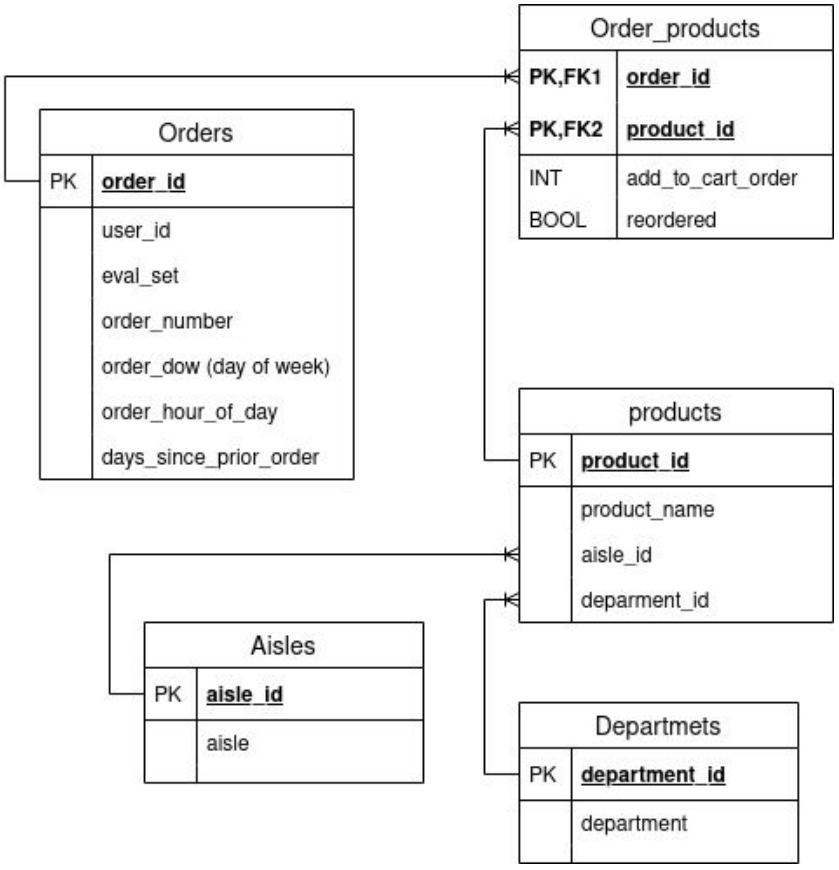
La siguiente es la explicación de las variables de cada conjunto de datos:

Ecommerce:

- InvoiceNo: Número de 6 que indica la factura a la que pertenece el registro. Si el código inicia con ‘c’ indica una cancelación.
- StockCode: Número de 5 dígitos que es asignado de forma única a cada producto.
- Description: Nombre del producto referenciado.
- Quantity: Cantidad de un producto por factura.
- InvoiceDate: Fecha en la que se realizó el pedido.
- UnitPrice: Precio por unidad en libras esterlinas.
- CustomerID: Número de 5 dígitos que identifica cada cliente único.
- Country: Nombre del país en el que reside cada cliente.

Este dataset contiene 541909 registros.

Instacart:



Orders (3.4M filas, 206k usuarios):

- order_id: Número que identifica cada orden
- user_id: Número que identifica cada usuario
- eval_set: Conjunto de evaluación al que pertenece el registro
- order_number: Número en la secuencia de órdenes por cada cliente asociado
- order_dow: Día de la semana en que se efectuó la order
- order_hour_of_day: Hora de la semana en que se efectuó la orden
- days_since_prior: Días desde la última orden, limitado a 30 días. Para las primeras órdenes se asigna N/A.

Products (50k filas):

- product_id: Número que identifica cada producto
- product_name: Nombre del producto
- aisle_id: Llave foránea con aisles
- department_id: Llave foránea con departments

Aisles (134 filas):

- aisle_id: Identificador del pasillo
- aisle: Nombre del pasillo

Departments (21 filas):

- department_id: Identificador del departamento
- department: Nombre del departamento

Order_products__SET (30m+ filas):

- order_id: Llave foránea con orders
- product_id: Llave foránea con products
- add_to_cart_order: Orden en el que el producto fue agregado al carrito
- reordered: 1 si el producto fue ordenado alguna vez por el usuario en el pasado, 0 en caso contrario.

donde SET es uno de los siguientes conjuntos (equivalente a eval_set en orders):

"prior": Lista de órdenes de cada cliente exceptuando la última realizada (~3.2M ordenes)

"train": Última orden realizada por cada cliente separada para probar modelos para el reto de Kaggle (~131k ordenes)

"test": Última orden realizada por cada cliente separada para evaluar modelos para el reto de Kaggle (~75k ordenes)

3. **Preparación de los datos:** Para el proyecto utilizamos Python con las herramientas de Pandas y Numpy como base para el procesamiento de los datos y los cálculos numéricos, también se utilizaron varias rutinas de la biblioteca de Scikit-learn para realizar los procesos de clusterización y predicción.

   A continuación se describirán con más detalles los procedimientos realizados sobre cada uno de los datasets:

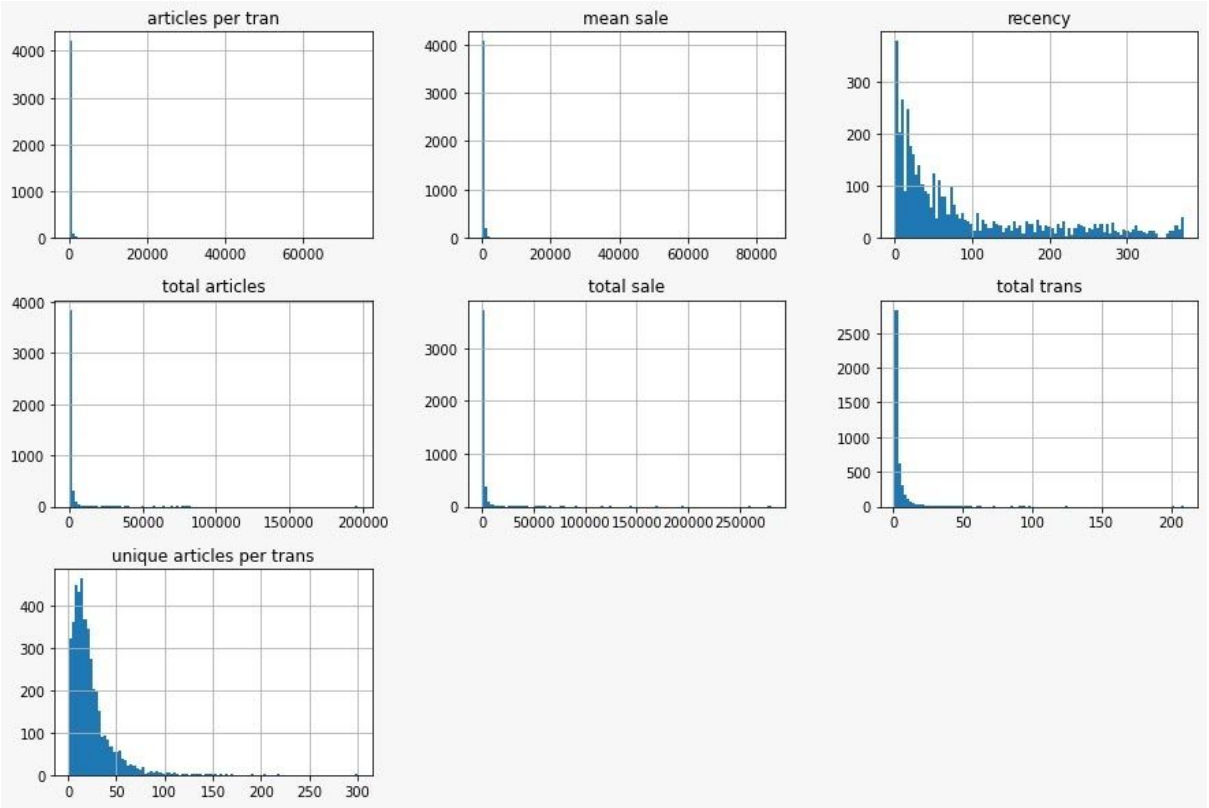
   ### 3.1. Dataset Ecommerce:

   Para el primer dataset se generaron varias variables adicionales que consideramos podrían ser útiles para los modelos que se planeó entrenar en la siguiente etapa. Las variables generadas son la siguientes:

   - articles_per_tran: Cantidad de artículos totales por cada factura
   - mean_sale: Media de los precios de los artículos de las transacciones del cliente
   - recency: Tiempo en días desde la última compra
   - total_articles: Total de artículos que ha adquirido
   - total_sale: Costo total de los artículos que ha adquirido
   - total_trans: Número de transacciones que ha hecho el cliente
   - unique_articles_per_trans: Cantidad de artículos únicos comprados por cada transacción.

   Luego de realizar un proceso de evaluación de varios modelos al final se tomó la decisión de solamente utilizar 3 variables que consideramos describen en lo mejor posible a los clientes: mean_sale, recency y total_trans.

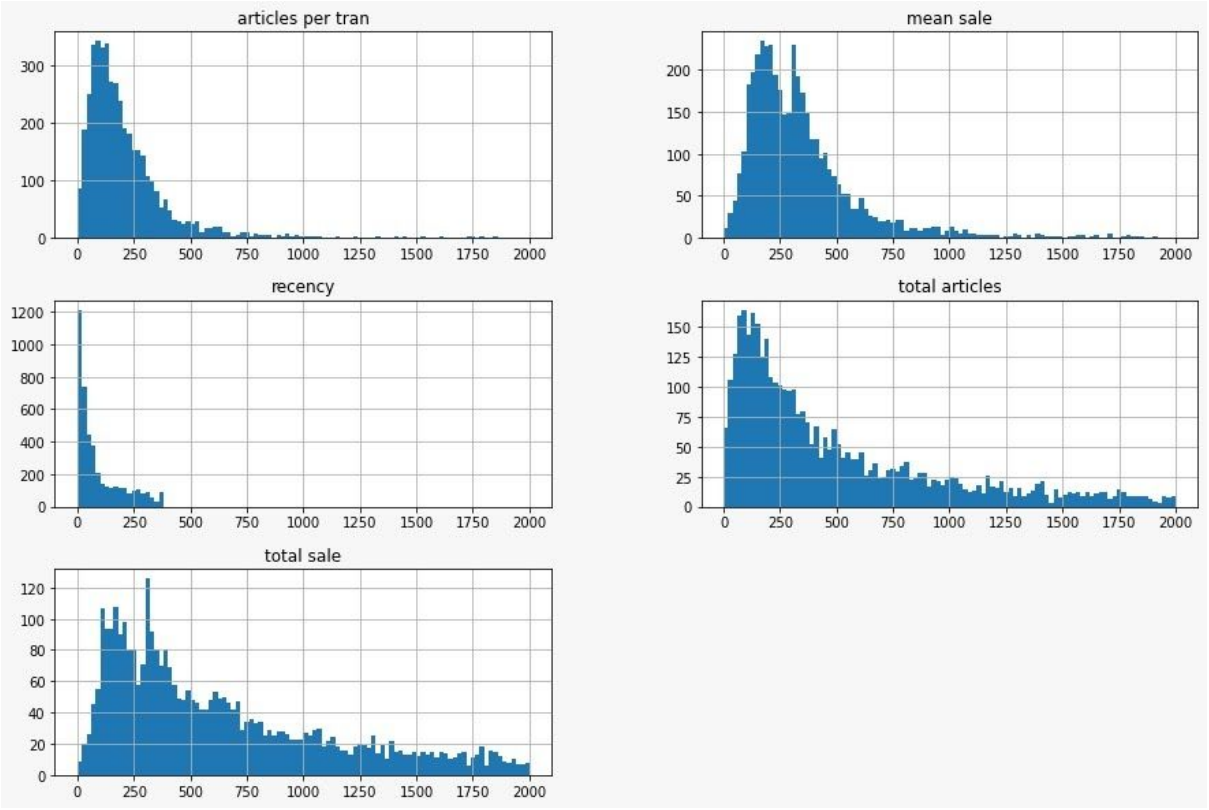
   Las variables generadas y las distribuciones obtenidas se pueden visualizar en la siguiente gráfica.

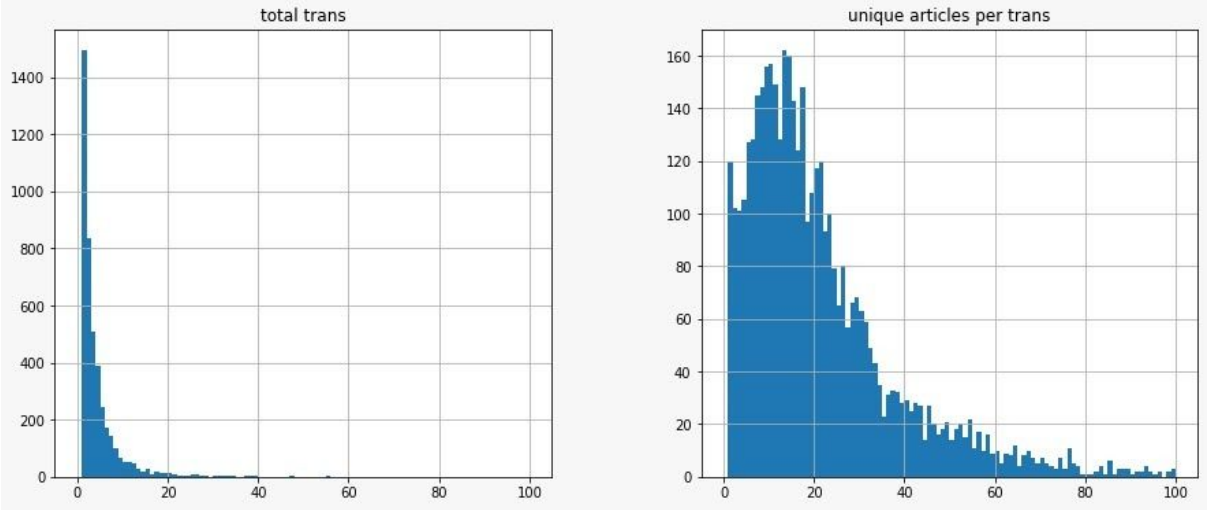


El proceso de detección de outliers se realizó antes de entrenar el modelo de clusterización del dataset. Se detectó que una gran cantidad de los clientes solo han realizado una sola transacción, sin embargo teniendo la cantidad de datos disponibles y que estos clientes no aportan información relevante en su comportamiento de compra para el negocio, se procedió a removerlos.

Una vez removidos los clientes con una sola transacción, se calculó el Z-Score de cada uno, el cuál es la cantidad de desviaciones estándar que está alejado de la media, y se removieron todos aquellos clientes que tuvieron al menos una variable que estaba alejada más de 3 desviaciones estándar de su media respectiva.

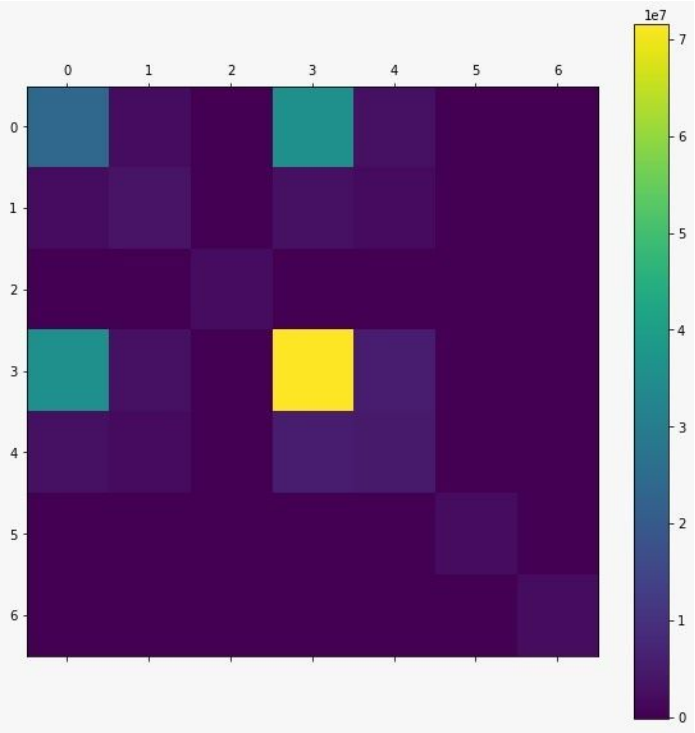
Luego de realizar este proceso en el dataset de E-commerce obtenemos las siguientes visualizaciones de las columnas generadas:



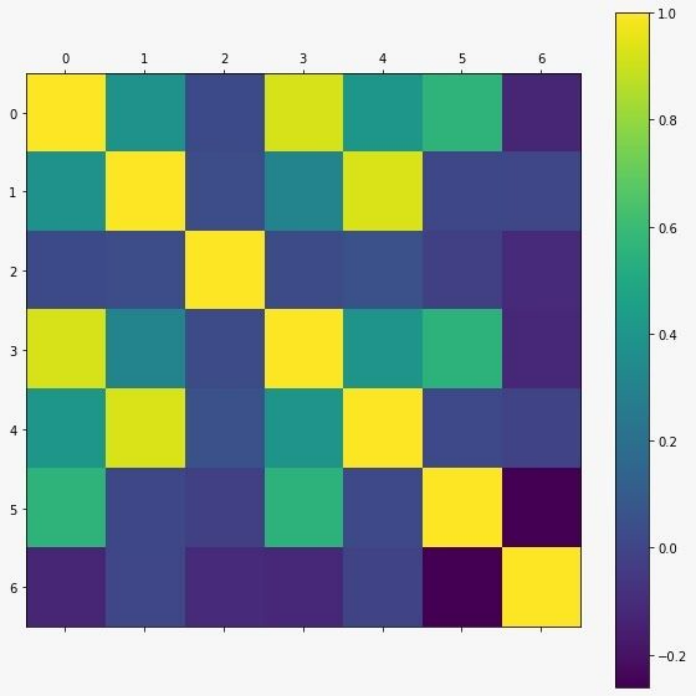


También se realizó un proceso de análisis de dependencias y correlaciones entre los datos. El resultado de este proceso fue el siguiente:

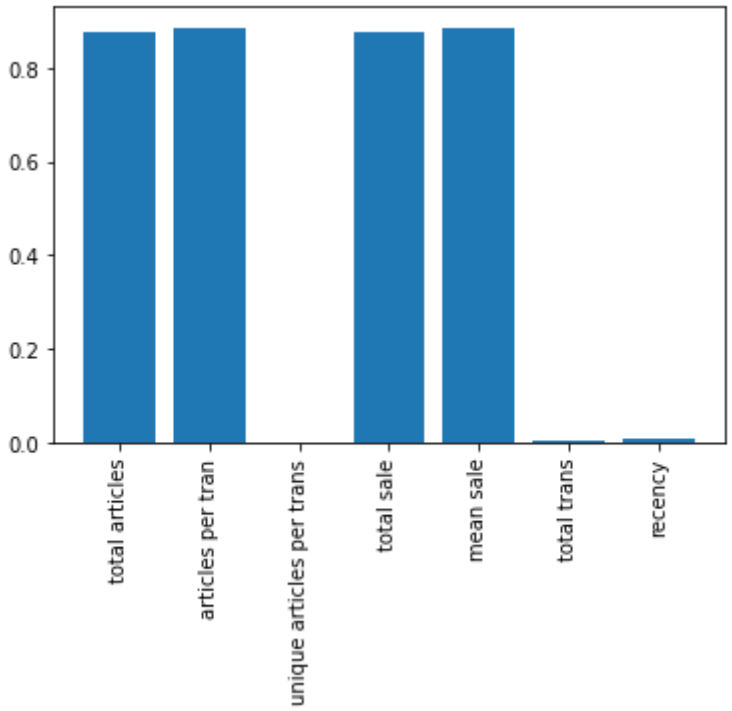
Covarianza:



Correlaciones:



También se realizó un análisis de dependencias entre cada una de las variables y se obtuvieron los siguiente valores de R^2 para cada variable:

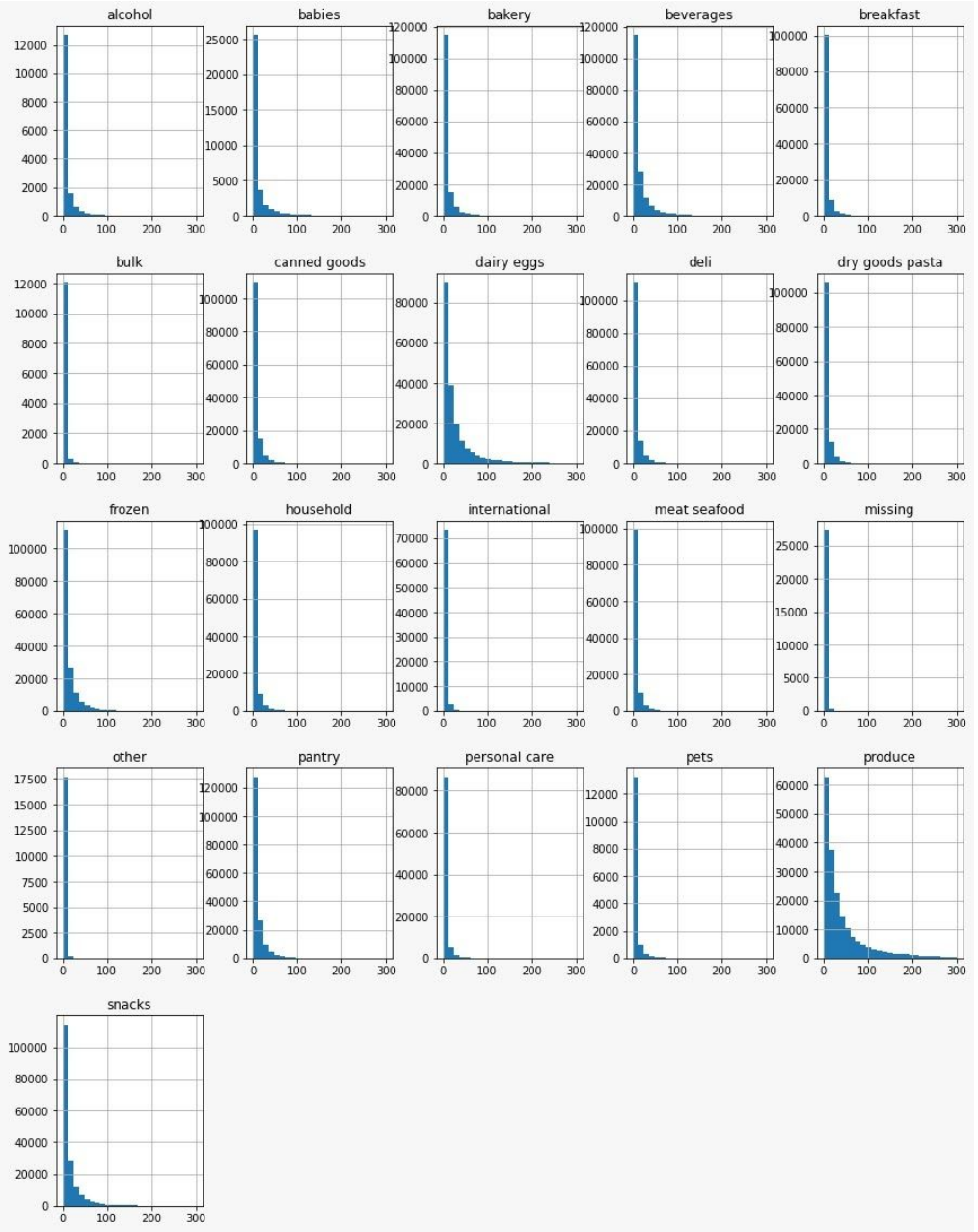


Luego de este análisis se optó por utilizar solamente las columnas de mean_sale, recency y total_trans.

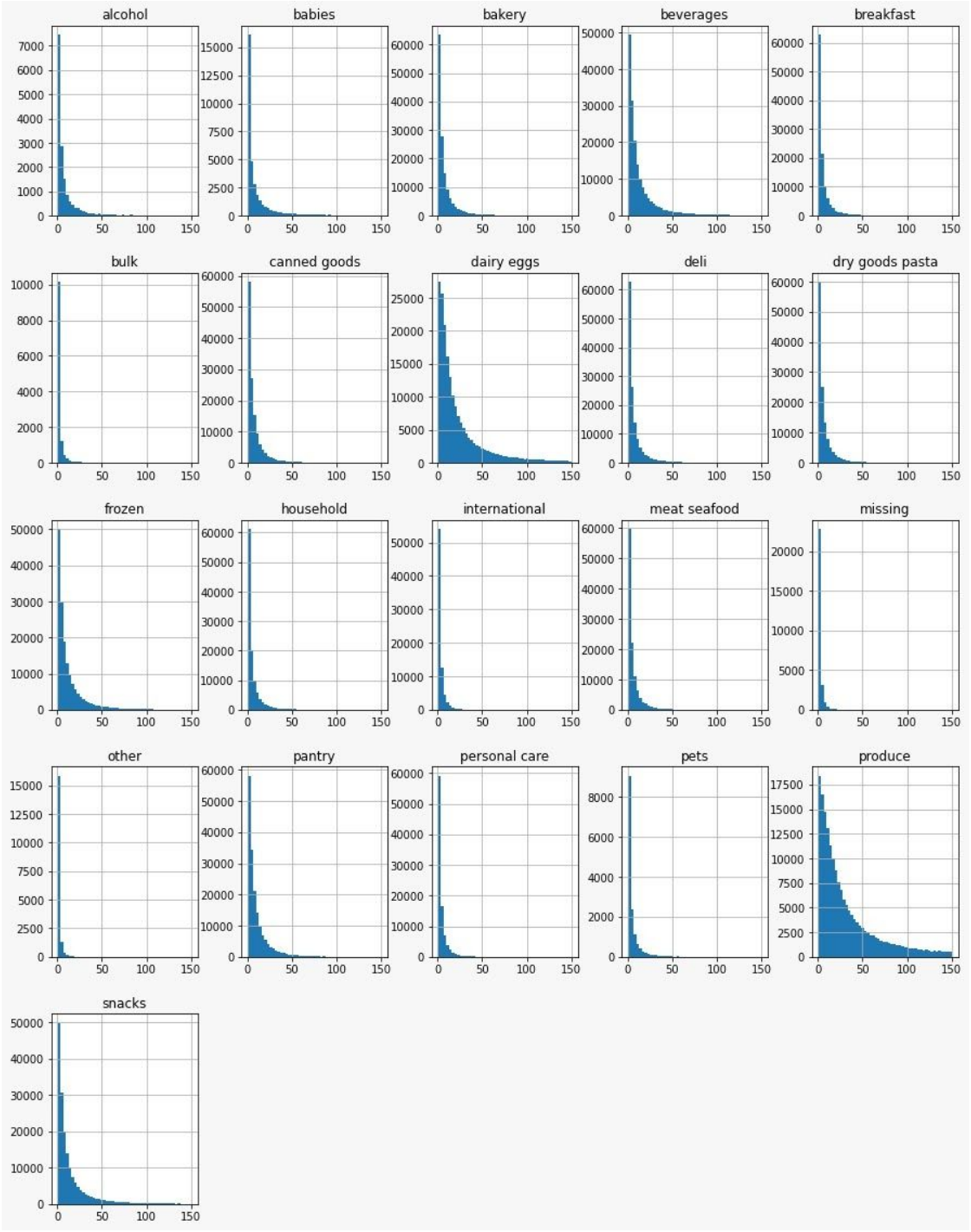
3.2. Dataset Instacart:

Para el segundo dataset se comenzó viendo los datos como parte de una base de datos relacional y mediante ***AWS Glue*** se realizó un proceso ETL para obtener una sola tabla que nos permitiera describir a las órdenes y a los clientes que luego se emplearán para alimentar los modelos estadísticos con los que se experimenta en el siguiente paso. Estas descripciones están basadas en los departamentos de los productos que son comprados en cada orden y en total por cada cliente.

La siguiente es la distribución de los datos por cada departamento en el resultado final de esta transformación.



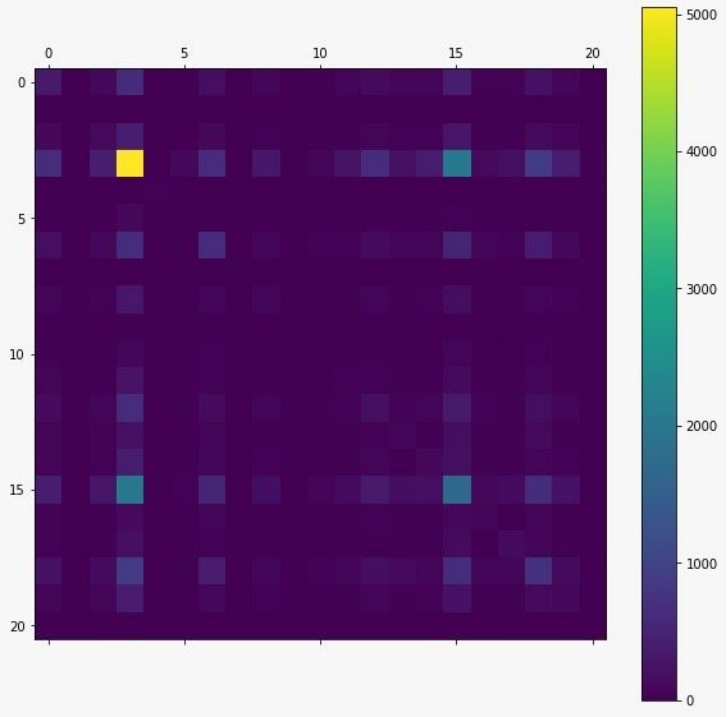
Si removemos los datos en 0 podemos apreciar mejor la distribución de los datos sobre las personas que compraron al menos un producto de cada departamento.



Adicionalmente se describieron los clientes utilizando los pasillos en lugar de los departamentos, pero dada la gran cantidad de categorías consideramos que dibujar todas las distribuciones ocupa demasiado espacio para incluirlo en este documento.

Posteriormente se realizó un proceso de análisis de dependencias y correlaciones entre los datos. Para los datos utilizando los departamentos se obtuvo lo siguiente:

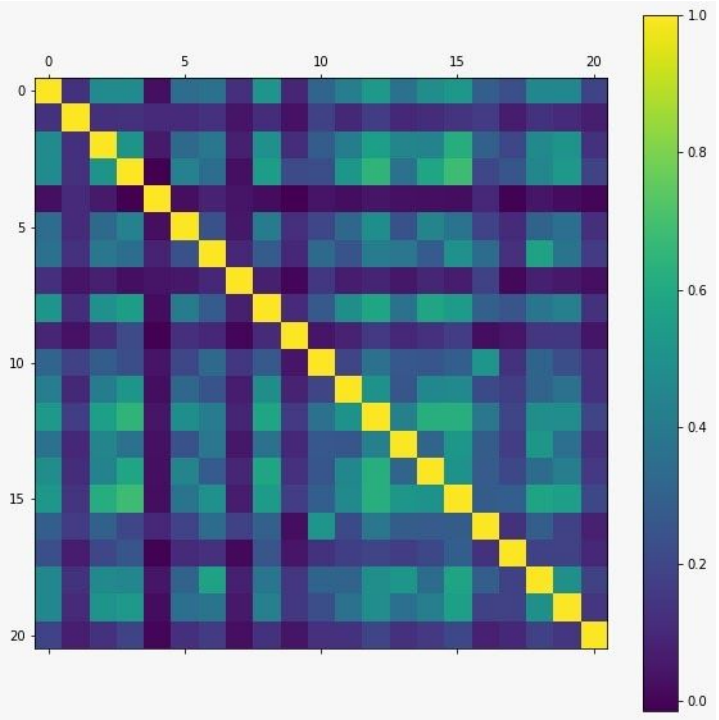
Covarianza:



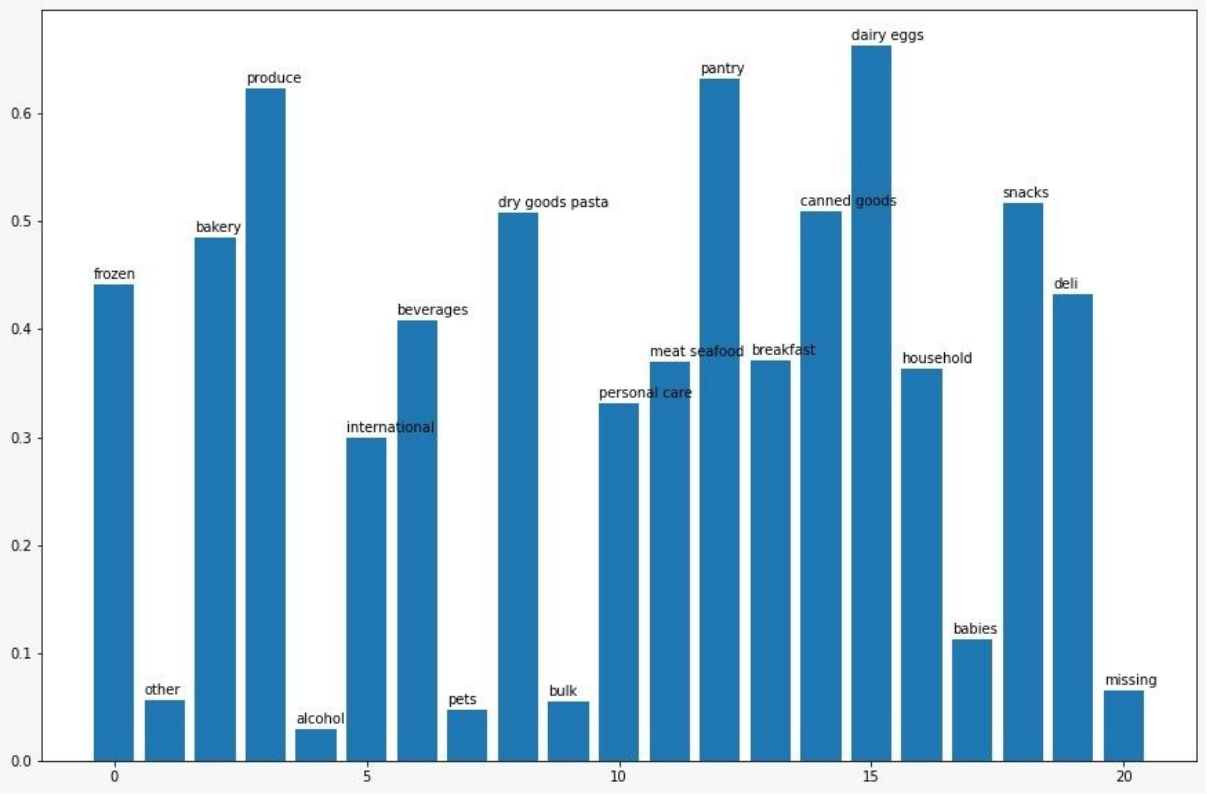
La cual tiene un número condición de 6744.65 luego de realizarle el encogimiento de Ledoit y Wolf.

Para tener una mejor idea de estos datos calculamos la varianza global, la cual nos dio un valor de 3.64x10^35. También calculamos el coeficiente de dependencia el cual nos dio 0.31. Con este se concluye que globalmente, la dependencia lineal explica 31% de la variabilidad de este conjunto de datos.

Correlaciones:

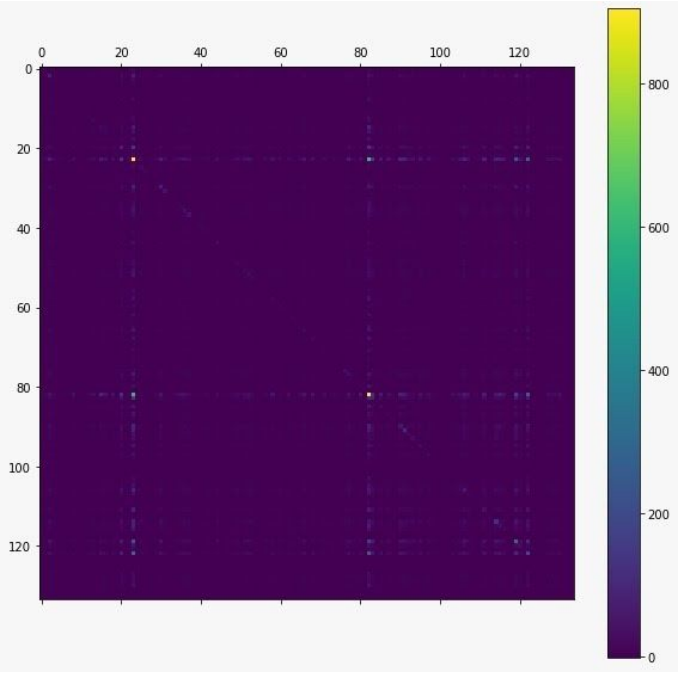


Dependencia entre cada variable y el resto:



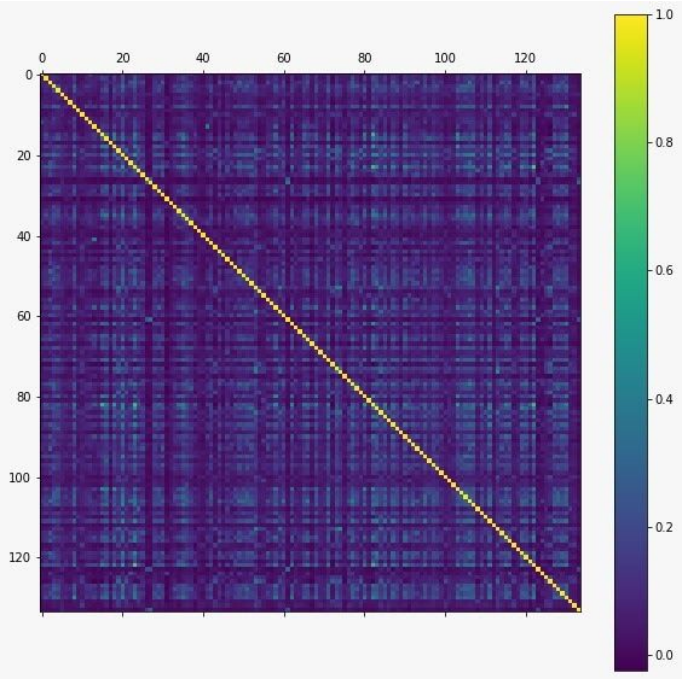
Y para los datos utilizando los pasillos se obtuvo lo siguiente:

Covarianza:



La cual tiene un número condición grande de 29246.30. Para los cálculos de las siguientes métricas que utilizan la covarianza intentamos utilizar el encogimiento de Ledoit y Wolf el cual bajó el número condición a 25135.52.

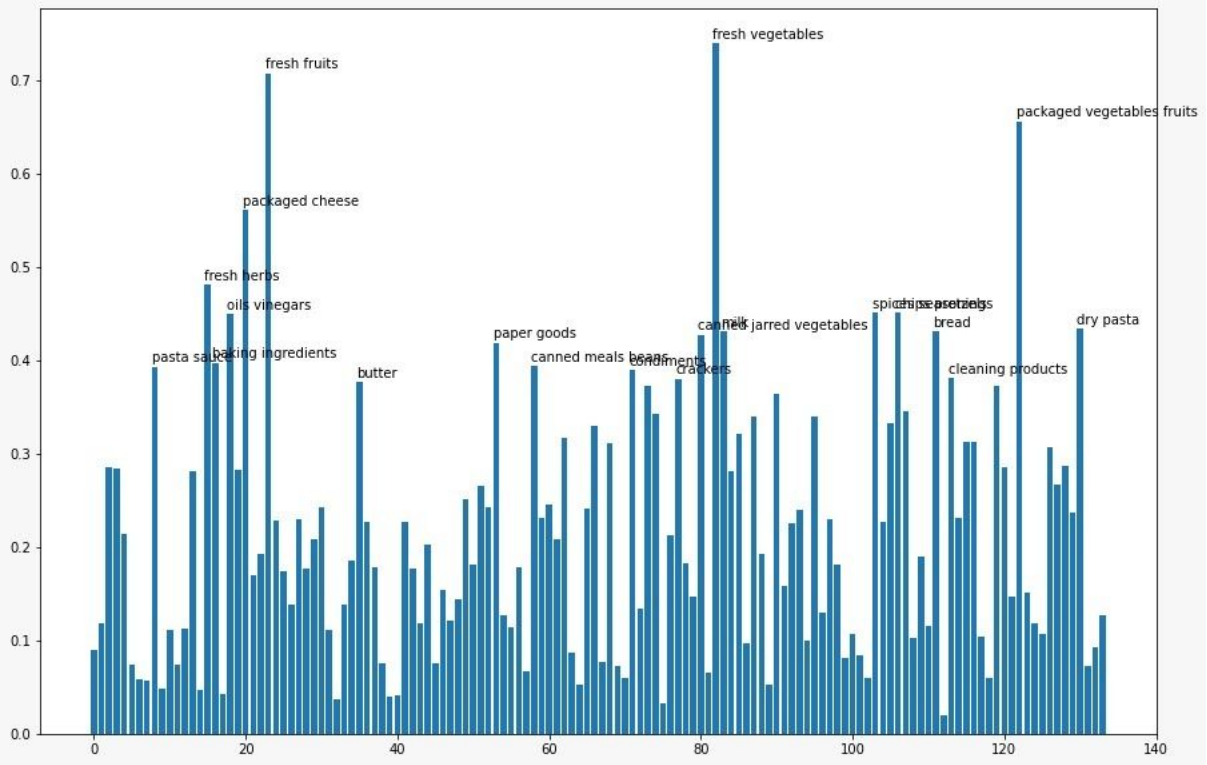
Correlaciones:



Para tener una mejor idea de estos datos calculamos la varianza global, la cual nos dio un valor de 3.73x10^73. También calculamos el coeficiente de

dependencia el cual nos dio 0.18. Con este se concluye que globalmente, la dependencia lineal explica 18% de la variabilidad de este conjunto de datos.

Dependencia entre cada variable y el resto:



4. Modelado:

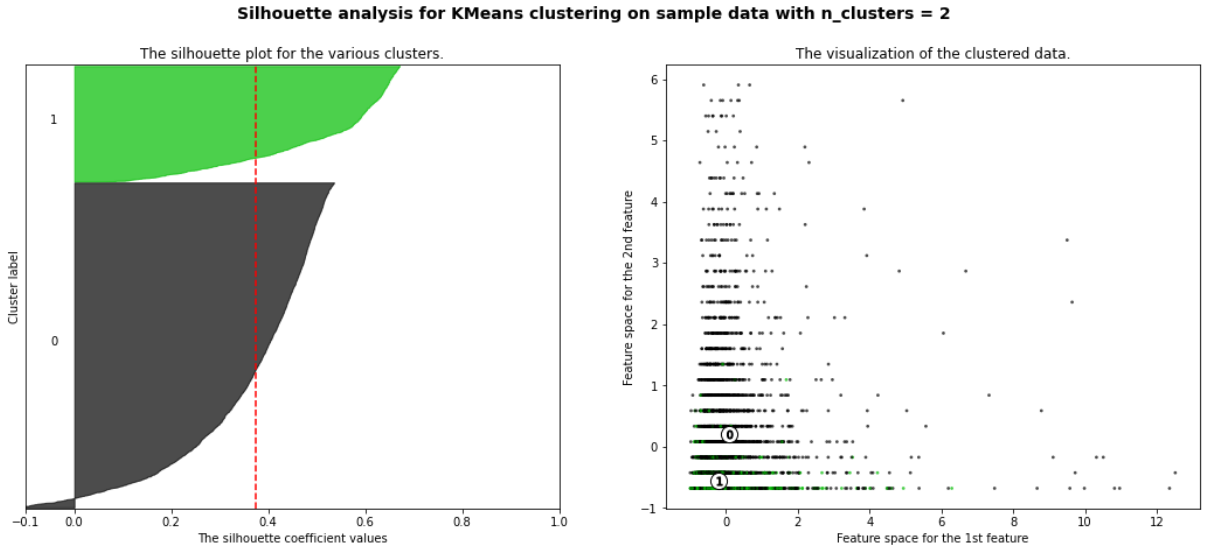
Con el fin de cumplir con los objetivos de negocio seleccionados se decidió utilizar un modelo de clustering k-means que nos ofreciera la oportunidad de segmentar los clientes existentes basados en varios parámetros de los datos utilizados. Nuestra propuesta es que las características compartidas de los clientes en el mismo cluster aumentan las probabilidades de éxito de una campaña de marketing diseñada para los segmento de clientes específicos. Para la evaluación de estos modelos hicimos un *análisis de siluetas* usando la métrica de *mahalanobis* y la *métrica euclidiana* para calcular la distancia para esta métrica, probando con varios valores de k para el algoritmo de clustering *k-means*.

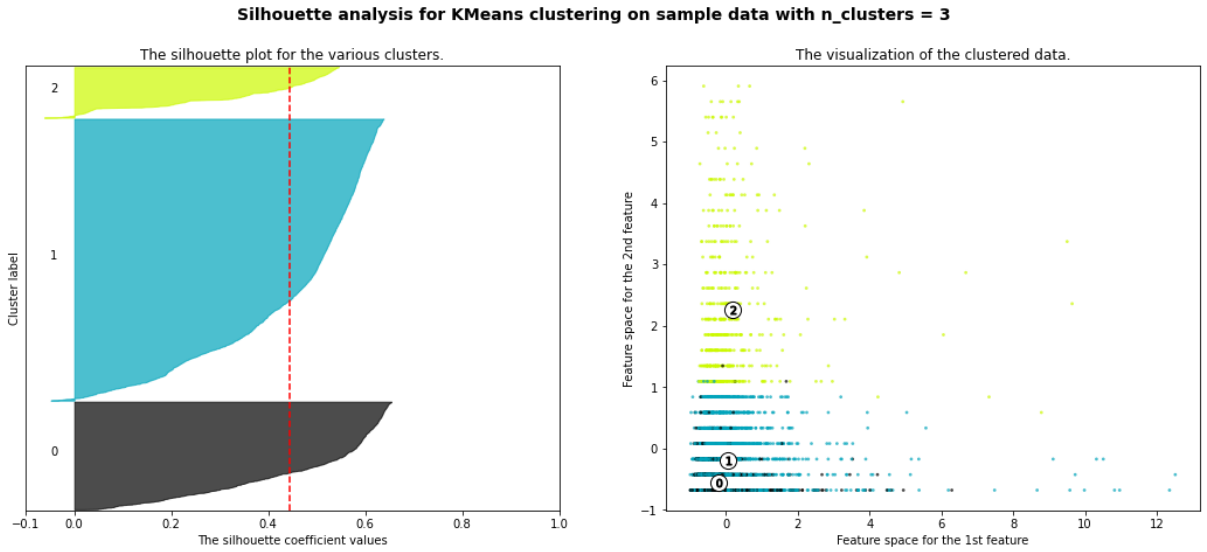
Para el dataset de Instacart, se decidió comprobar el resultado de efectuar la regresión lineal del dato con mayor R^2 entre los pasillos y verificar si con alguno de los tipos de regularización era posible obtener un resultado similar con menos parámetros.

Lo siguiente son los resultados de los modelos probados para cada dataset:

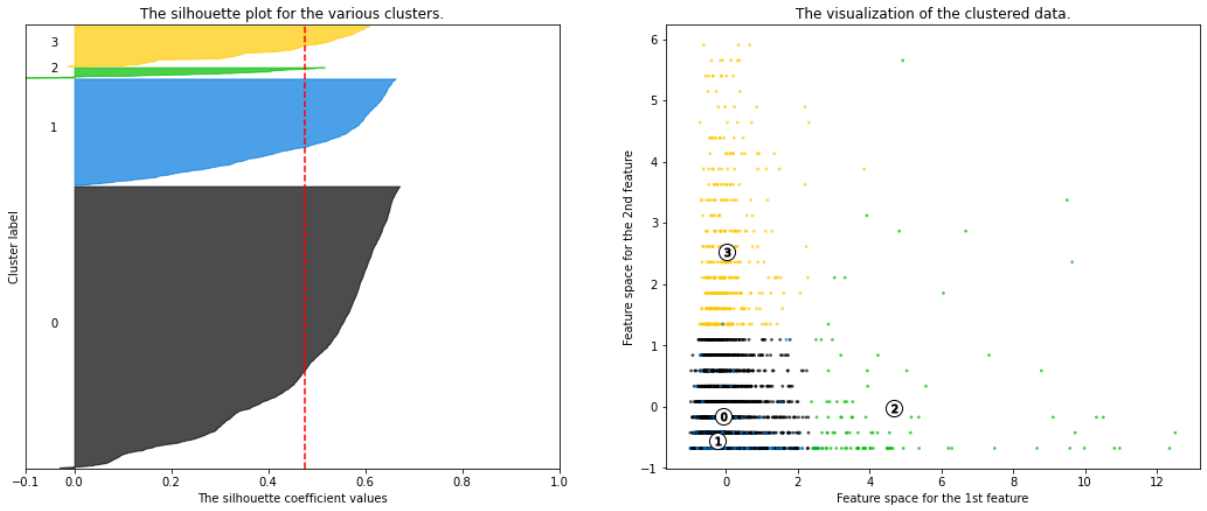
4.1 Dataset Ecommerce:

Para n_clusters = 2 la silhouette_score promedio es: 0.37
Para n_clusters = 3 la silhouette_score promedio es: 0.44
Para n_clusters = 4 la silhouette_score es: 0.47
Para n_clusters = 5 la silhouette_score es: 0.48
Para n_clusters = 6 la silhouette_score es: 0.43

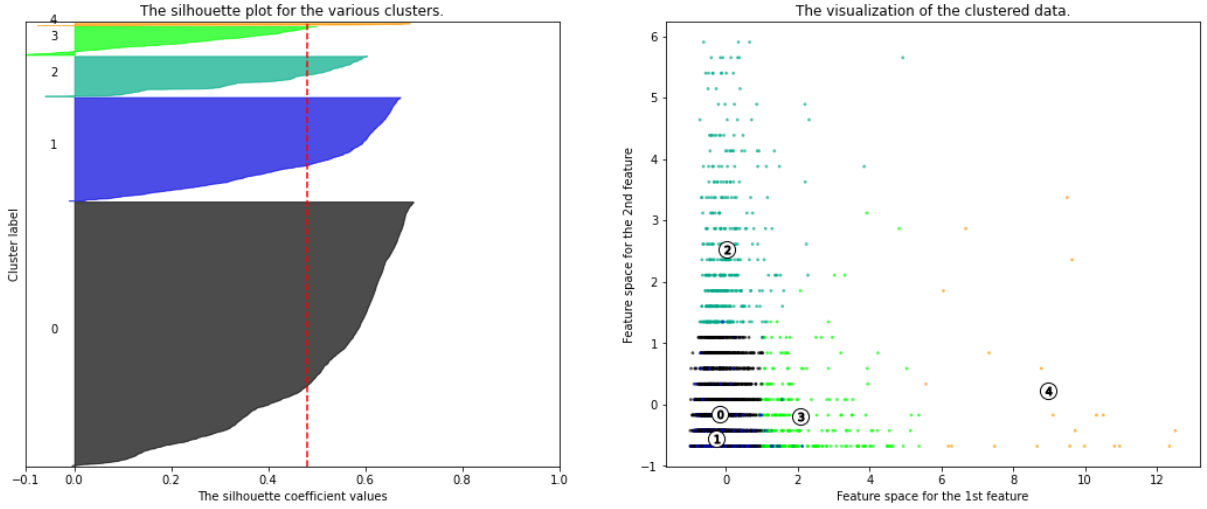




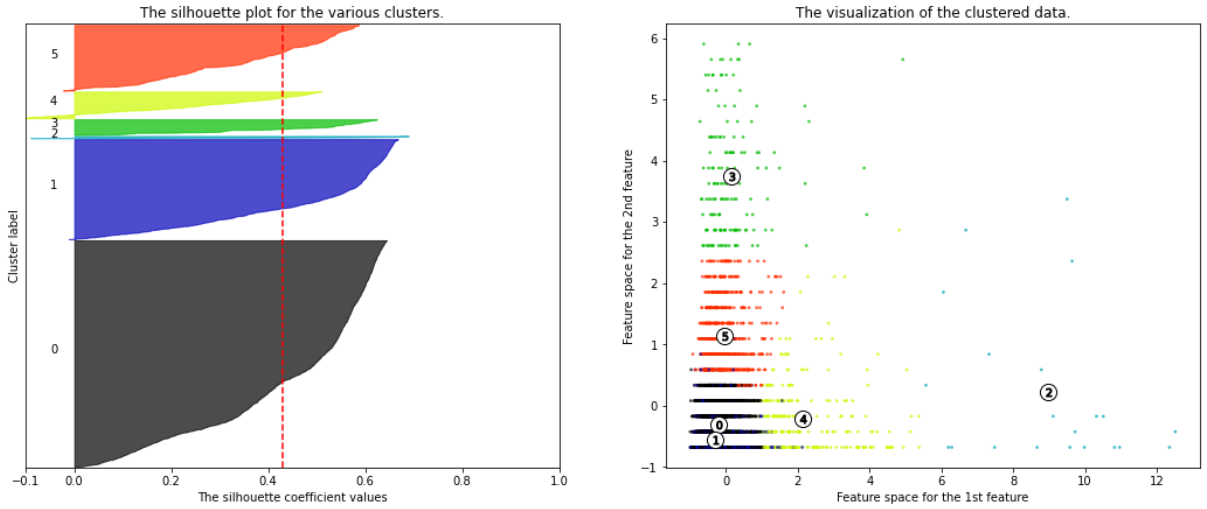
Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



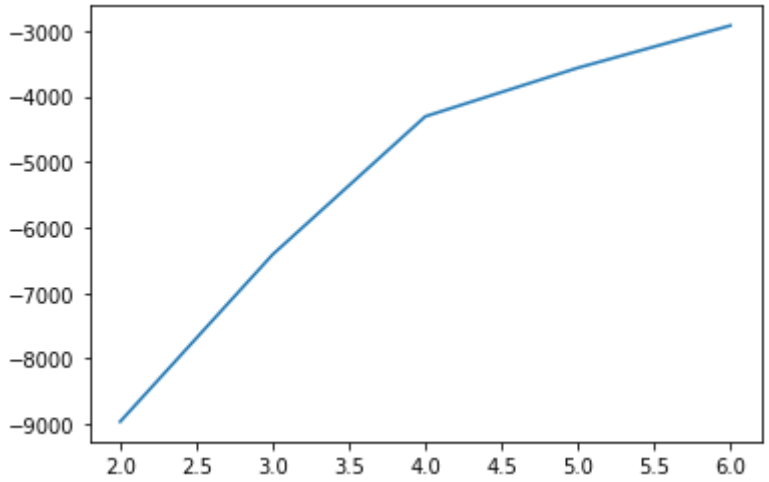
Silhouette analysis for KMeans clustering on sample data with n_clusters = 5



Silhouette analysis for KMeans clustering on sample data with n_clusters = 6



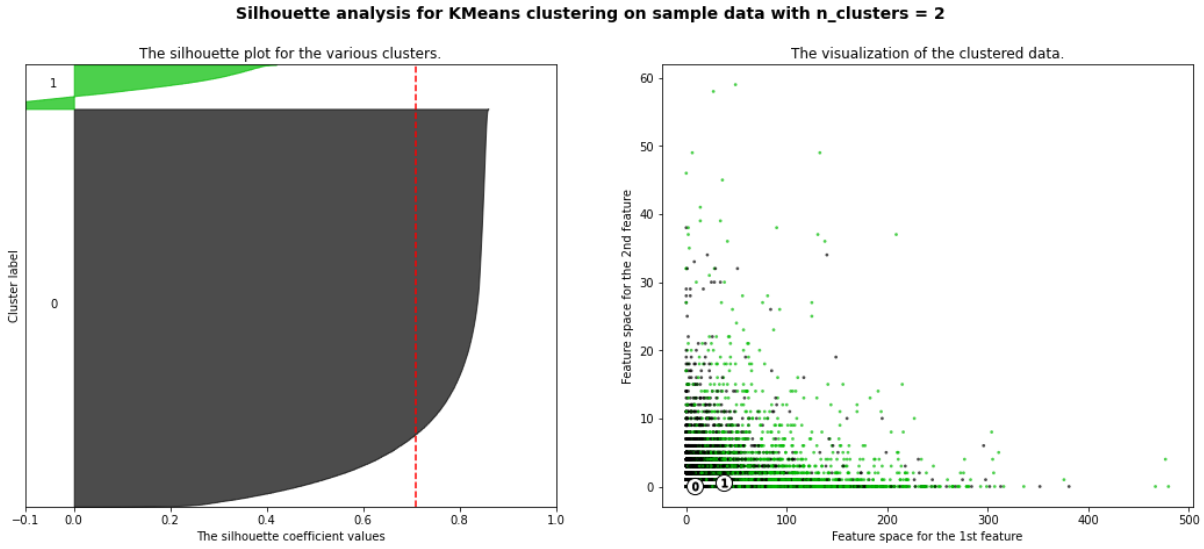
Adicionalmente también calculamos la suma de todas las distancias para cada k en este dataset y obtuvimos lo siguiente:



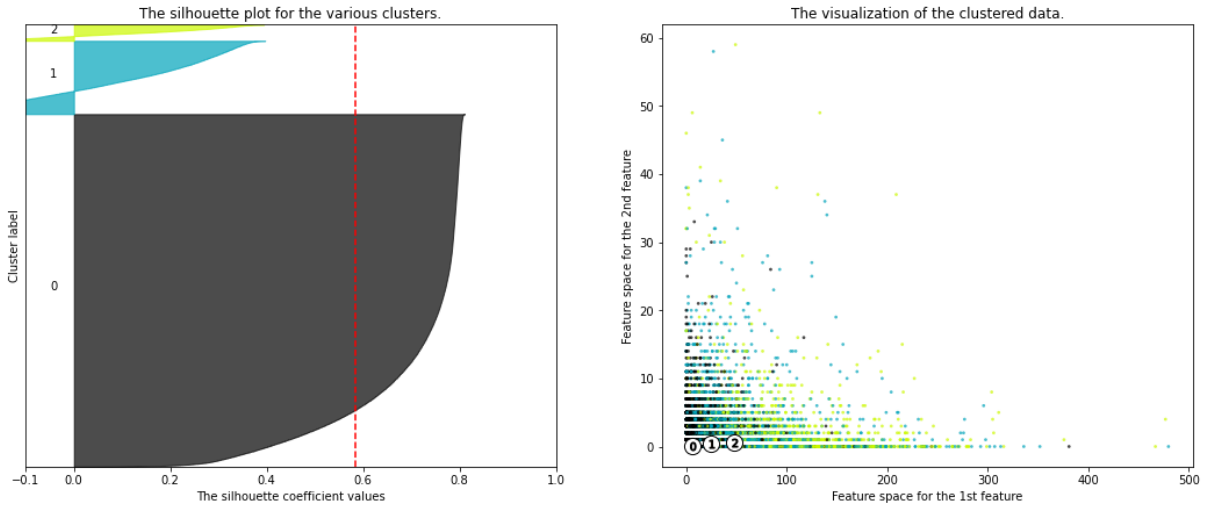
Realizando la ley del codo con esta gráfica y verificando el análisis de siluetas consideramos que el clustering óptimo es con k=4.

4.2 Dataset Instacar:

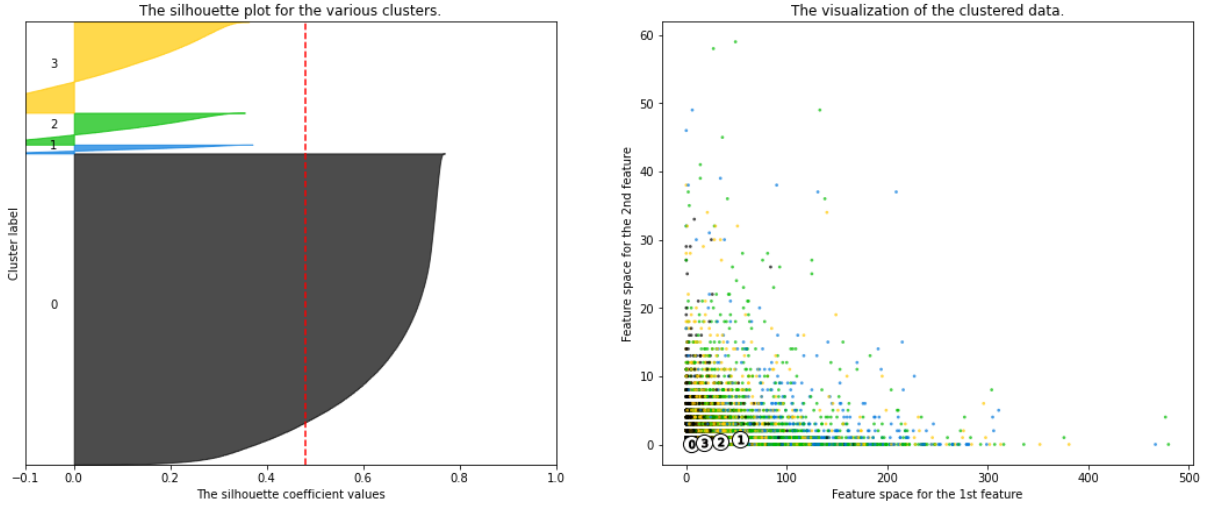
Para n_clusters = 2 la silhouette_score promedio es: 0.71
Para n_clusters = 3 la silhouette_score promedio es: 0.58
Para n_clusters = 4 la silhouette_score promedio es: 0.48
Para n_clusters = 5 la silhouette_score promedio es: 0.47
Para n_clusters = 6 la silhouette_score promedio es: 0.41

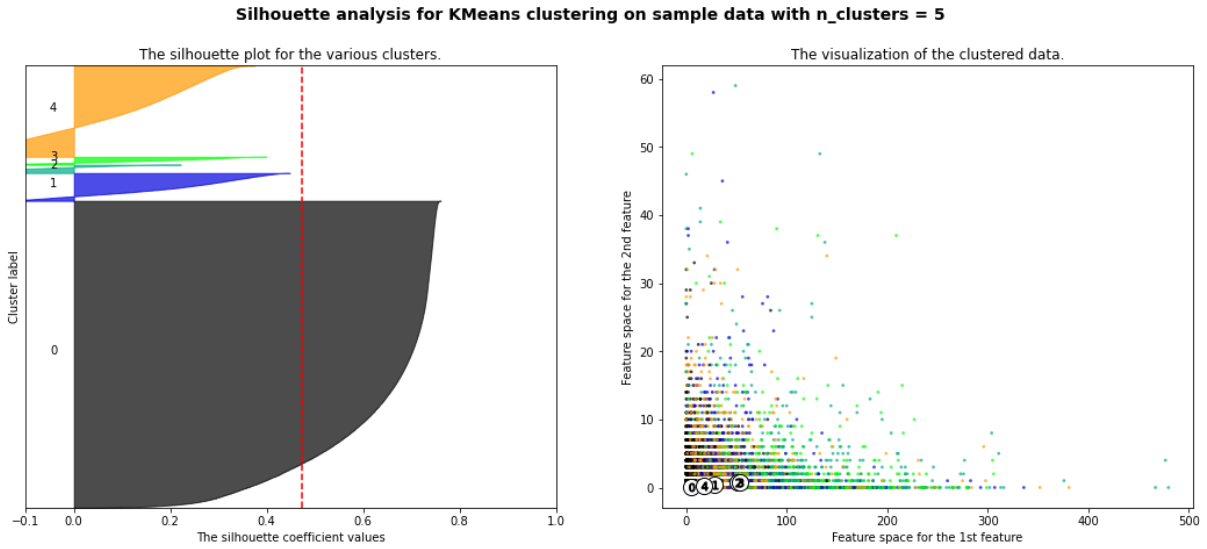


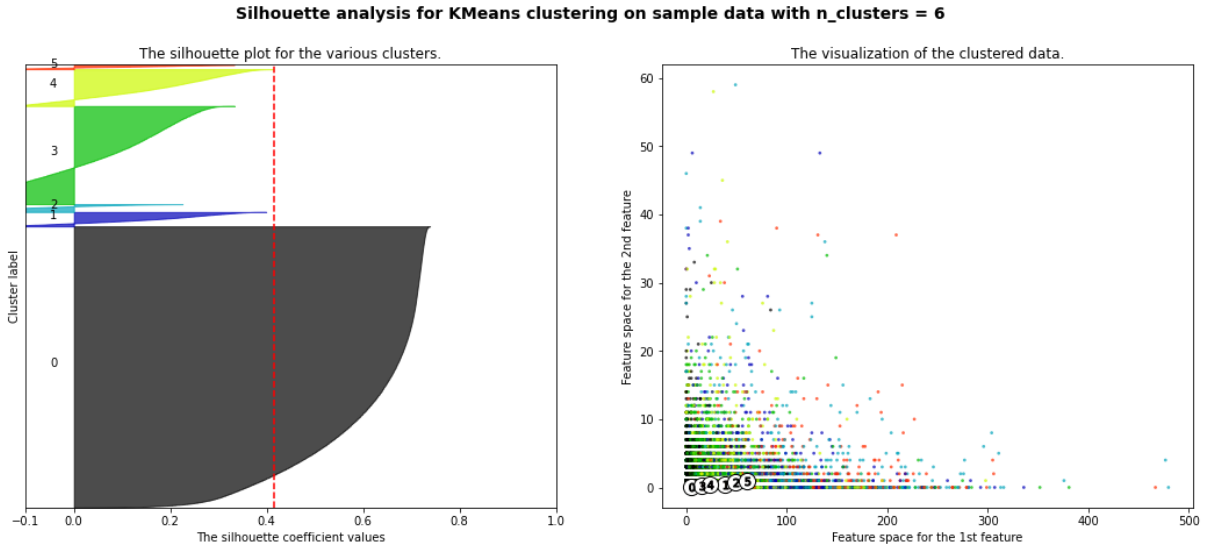
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3



Silhouette analysis for KMeans clustering on sample data with n_clusters = 4







Los resultados de la regresión lineal para varios tipos de modelos fueron los siguientes:

Modelo	**Train (R^2)**	**Test (R^2)**	**# Parámetros**
Lineal	0.740932	0.731464	133
Ridge	0.740932	0.731464	133
Lasso	0.733220	0.723949	54
Elastic	0.736440	0.727093	71

5. Evaluación:

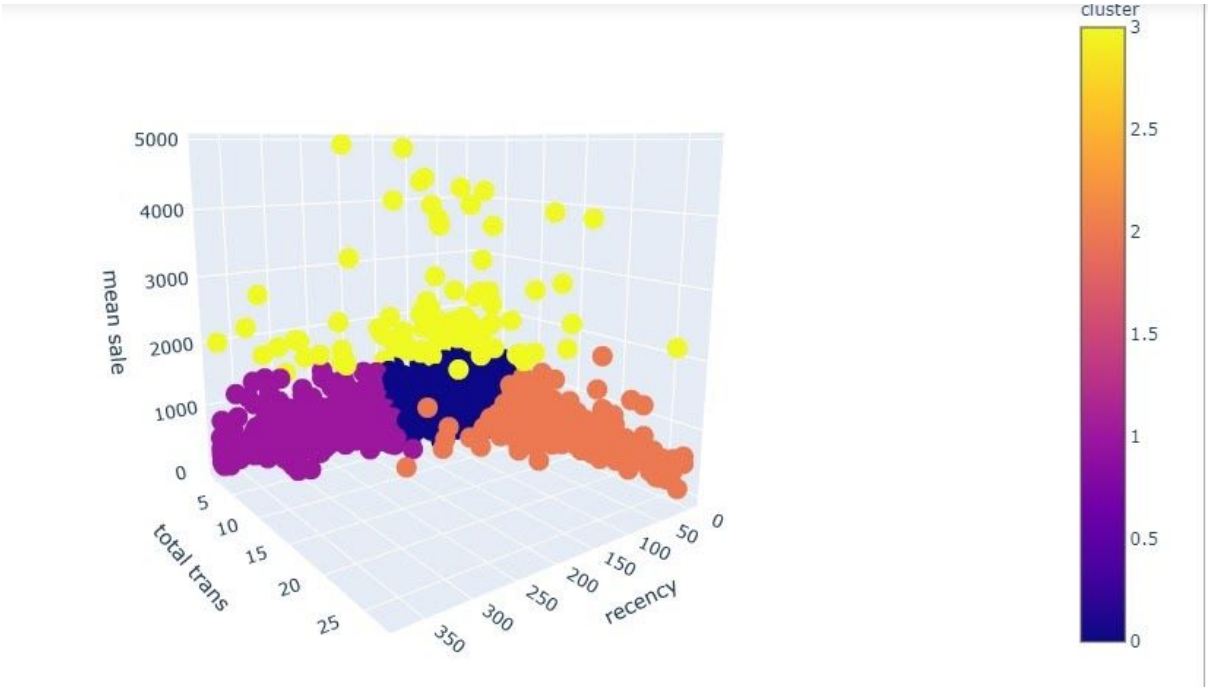
Al ver los resultados anteriores consideramos que se han encontrado varias aproximaciones aceptables para identificar y segmentar los clientes de un comercio electrónico basándonos en diferentes características que tengamos disponibles por cada cliente.

5.1 E-commerce:

Partiendo del primer conjunto de datos logramos encontrar una clasificación con 4 categorías. Para facilitar su explicación graficamos los datos en 2 y 3 dimensiones para ver si era posible observar los clusters en baja dimensionalidad. Para graficar en 2 dimensiones se utilizó el algoritmo de PCA para reducir la dimensionalidad de los datos. Con este algoritmo obtuvimos que 2 dimensiones explicaban 79% de la varianza de los datos. Explicando la primera dimensión el 47% y la segunda el 32%.

2D:

3D:



Según estas gráficas y los clusters encontrados es posible deducir lo siguiente. Los clientes morados son los que tienen una recencia alta, pocas transacciones y poca venta promedio, es decir, es el grupo de los clientes que ya no ha vuelto a comprar, que prácticamente ya abandonaron. Los clientes anaranjados son los que tienen una recencia pequeña y alta cantidad de transacciones. Esos son el grupo de los clientes activos o fieles de la compañía. Los clientes azules son los que tienen pocas transacciones y niveles medianamente altos y bajos de recencia. Estos se podrían considerar como los clientes nuevos de la compañía. Finalmente los amarillos son los clientes con una venta promedio alta, estos son los que, sin importar si siguen viniendo o no, generan muchas ganancias cada vez compran.

5.2 Instacart:

Al revisar los resultados obtenidos con el análisis realizado a este conjunto de datos consideramos que las características seleccionadas para clasificar a los clientes no ofrecen un análisis tan claro de los comportamientos de los usuarios de este servicio. Esto lo podemos concluir del resultado del análisis de siluetas que se realizó al k-mean que fue entrenado describiendo cada cliente mediante los departamentos de los productos que compra. En todas las gráficas de siluetas se puede apreciar una gran cantidad de outliers en la

mayoría de categorías identificadas. Debido a esto decidimos que este modelo no ofrece un resultado satisfactorio para la identificación y clasificación de los clientes de este dataset.

Por otro lado, la regresión lineal realizada para el elemento con mayor dependencia con el resto de variables, que en este caso es “fresh vegetables”, nos ofrece una oportunidad para identificar el comportamiento de aquellos clientes que utilizan el servicio de Instacart para la compra habitual de vegetales. Consideramos que este modelo puede ser utilizado como base de conocimiento para identificar clientes habituales que utilizan el servicio de Instacart para abastecerse de víveres de manera recurrente en lugar de tiendas físicas de retail y, excluyendo estos clientes, observar que comportamiento suelen tener los clientes que probablemente utilicen este servicio de forma más ocasional para luego incentivarlos a que usen más la plataforma de Instacart para realizar el mercado en lugar de recurrir a tiendas físicas de retail.