Prueba Data Scientist.

Juan David García Castro

juancastro97125@gmail.com

A continuación, se presentarán los resultados y el análisis de cada una de las tareas de la prueba.

1) **Tarea 1: Análisis descriptivo de los datos.**

a) **Líneas de producto.**

Se llevó a cabo un análisis exploratorio de los datos en donde los principales hallazgos fueron los siguientes.

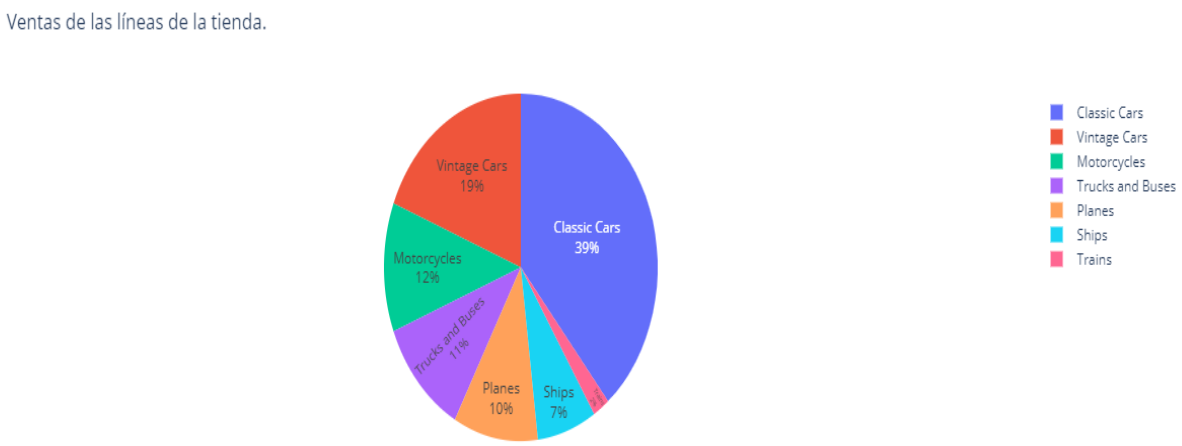

Diagrama 1: Distribución de las ventas por línea de producto.

Con base en la gráfica anterior se encontró que la línea más vendida es Classic Cars con un total de ventas de $3919615.66 y 33992 unidades pedidas.

Por otro lado, la línea menos vendida es Trains con un total de ventas de $226243.47 y 2712 unidades pedidas.

b) **Producto.**

El código del producto más vendido es S_18_3232 con un total de ventas de $288245.42 y 1774 unidades pedidas.

El código del producto menos vendido es S_24_3269 con un total de ventas de $33181.66 y 745 unidades pedidas.

c) Clientes.

El cliente que más compró es Euro Shopping Channel con un total de compras de $912294.11 y 9327 unidades pedidas.

El cliente que menos compró es Boards & Toys Co. con un total de compras de $9129.35 y 102 unidades pedidas.

d) Paises.

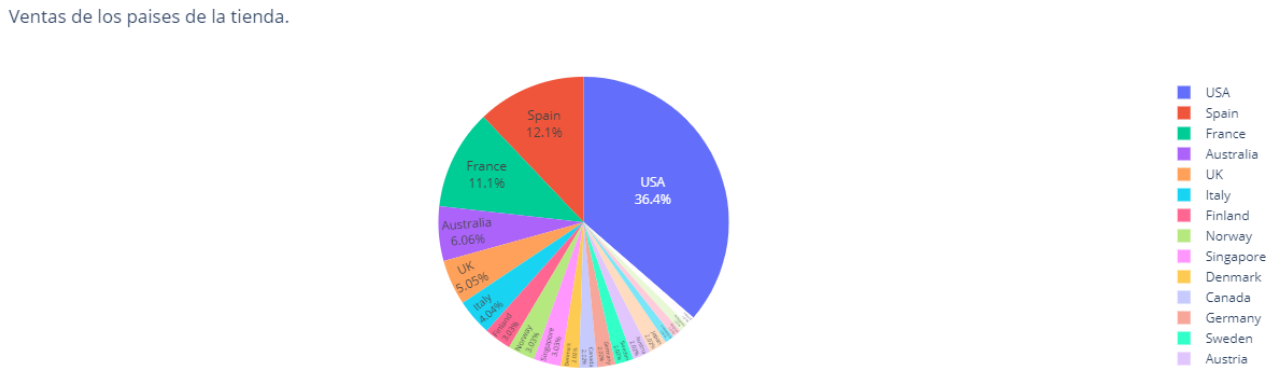

Diagrama 2: Distribución de las ventas por paises.

El país que más compró es USA con un total de compras de $3627982.83 y 35659 unidades pedidas.

El país que menos compró es Irlanda con un total de compras de $57756.43 y 490 unidades pedidas.

e) Meses del año.

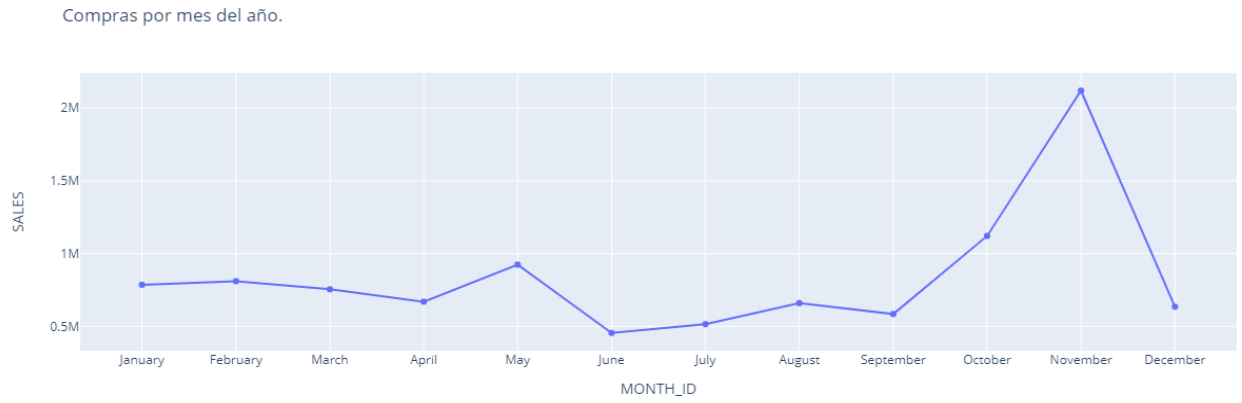

Diagrama 3: Ventas a lo largo de los meses del año.

En esta gráfica se puede observar que, en los últimos meses del año, principalmente en noviembre es cuando se presentan la mayor cantidad de las ventas. Por otro lado, a mitad de año, principalmente en junio es cuando se presentan una menor cantidad de ventas.

Es importante resaltar que en esta gráfica se visualizan las ventas haciendo un acumulado de las ventas durante cada mes teniendo en cuenta los tres años de los cuales se tienen registro. Es decir, desde el año 2003 hasta el año 2005.

2) Tarea 2: Segmentación de los clientes haciendo uso de la tabla RFM.

Tabla 1 Clasificación de las empresas según el RFM.

Segmentos	**R**	**F**	**M**	**Descripción.**
Potenciales.	La Rochelle Gifts Euro Shopping Channel	Euro Shopping Channel Mini Gifts Distributors	Euro Shopping Channel Mini Gifts Distributors	Son los clientes que menos días tardan en comprar, los que más a menudo lo realizan

				y son los que más gastan
Perdidos	Norway Gifts by Mail, Co. Mean 'R' US Retailers, Ltd. Double Decker Gift Stores, Ltd.	Boards & Toys Co. Atelier graphique Auto-Moto Classics Inc.	Boards & Toys Co. Atelier graphique Auto-Moto Classics Inc.	Son clientes que compraron hace mucho tiempo, poca cantidad y poco gasto.
Derrochadores			Euro Shopping Channel	Los que más gastan
Leales		Euro Shopping Channel Mini Gifts Distributors		Son los que más frecuente van a comprar
Nuevos	La Rochelle Gifts Euro Shopping Channel			Clientes que han comprado hace poco pero no a menudo

Según los resultados presentados en la tabla anterior, es evidente que aquellos clientes que piden encargos con mayor frecuencia son aquellos que proporcionan un mayor valor monetario a la tienda. Igualmente, aquellos clientes que compran con mayor frecuencia son aquellos que también proporcionan un mayor valor monetario a la empresa.

Esto puede sugerir que hay cierto segmento de clientes que está bastante satisfecho con los productos que ofrece la empresa y son estos en los cuales deberían concentrarse todos los esfuerzos por mantenerlos.

3) Segmentación por clustering.

Al llevar a cabo la segmentación por métodos de clustering se hizo uso del método de K-Means. Con este método se segmentan los puntos con base a la distancia al centroide de cada cluster.

Con el fin de conocer cuál es el número de clusters optimo se llevó a cabo el método del 'codo' con el cual se evalúa el punto donde el cambio en el WCSS (within cluster sum of squares) es más evidente, es decir, donde se evidencia en la gráfica una flexión de la recta.

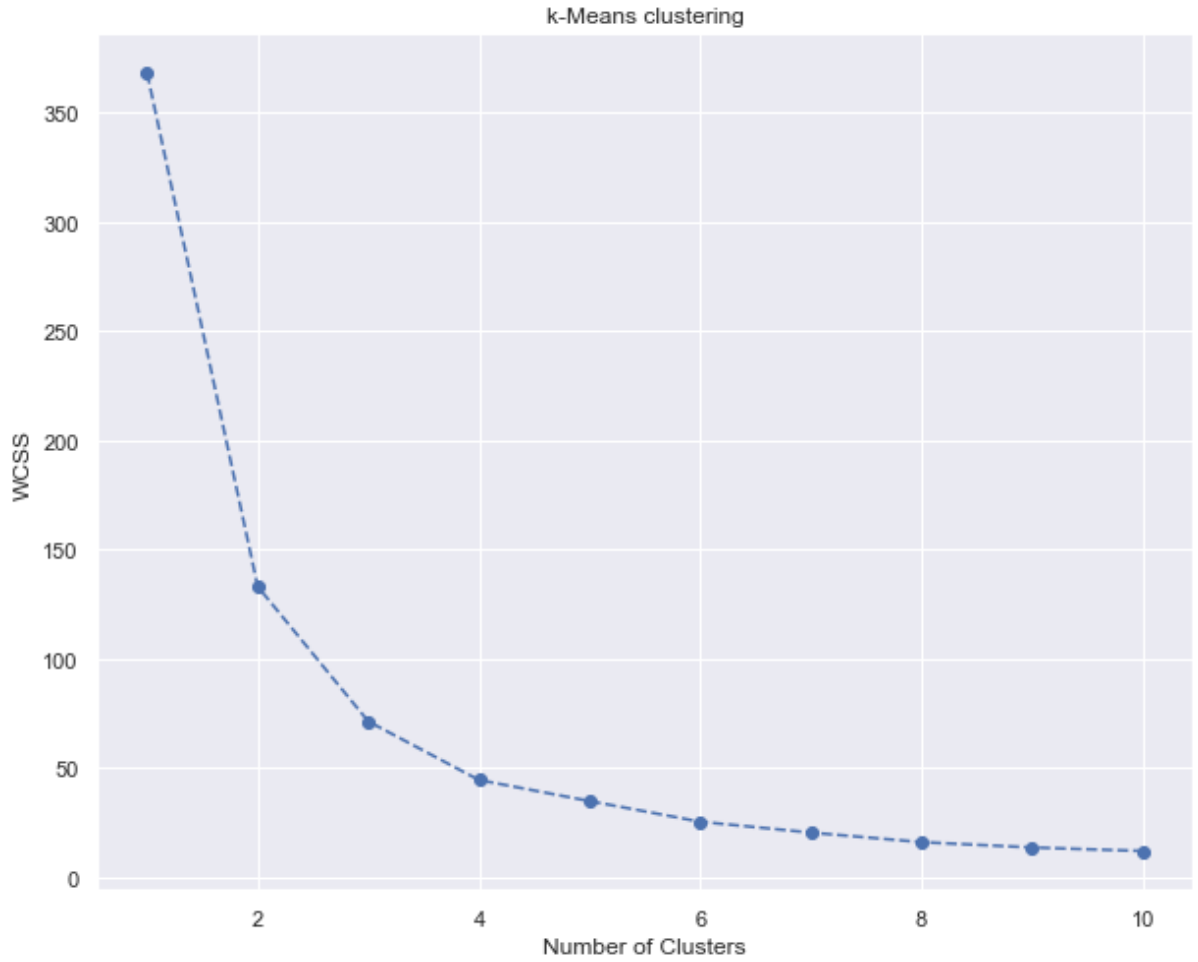

Diagrama 3: WCSS vs número de clusters.

A partir de la gráfica anterior se determinó que el mejor número de clusters era 3.

Posterior a esto se obtuvo los 3 clusters principales de la tienda en donde los clasifique de la siguiente manera.

Tabla 2 Segmentación de las empresas haciendo uso de métodos de clustering (K-means).

Segmentos	Descripción	Cantidad de empresas.
Perdidos	Está conformado por personas que en promedio su última compra fue hace 460.2 días, con un número promedio de compras de 18.7, cuyo valor promedio es de $66365.6 y 661.4 unidades ordenadas en promedio.	13
Promedio	Está conformado por personas que en promedio su última compra fue hace 143.7 días, con un número promedio de compras de 27.8, cuyo valor promedio es de $98736.7 y 971.1 unidades ordenadas en promedio.	77
Potenciales	Está conformado por personas que en promedio su última compra fue hace 1.0 días, con un número promedio de compras de 219.5, cuyo valor promedio es de $783576.1 y 7846.5 unidades ordenadas en promedio.	2

Como se puede observar, para esta segmentación se tienen menos grupos siendo el grupo de clientes potenciales conformado únicamente por 2 empresas, el grupo de clientes promedio conformado por 77 empresas y el grupo de clientes perdidos conformado por 13 empresas.

El resultado es tan drástico debido a que el algoritmo le da mucho peso a la frecuencia con la cual el cliente solicita pedidos, siendo la diferencia de frecuencia de pedidos entre el grupo de potenciales y promedio un número bastante significativo (459 días en promedio).

Realice el mismo procedimiento realizando un PCA antes de la segmentación con el fin de evaluar si obtenía una mejor segmentación haciendo uso de los componentes que

describieran mejor al conjunto de datos. Sin embargo, el resultado fue igual por lo que a nivel de producción se podría utilizar cualquier modelo.

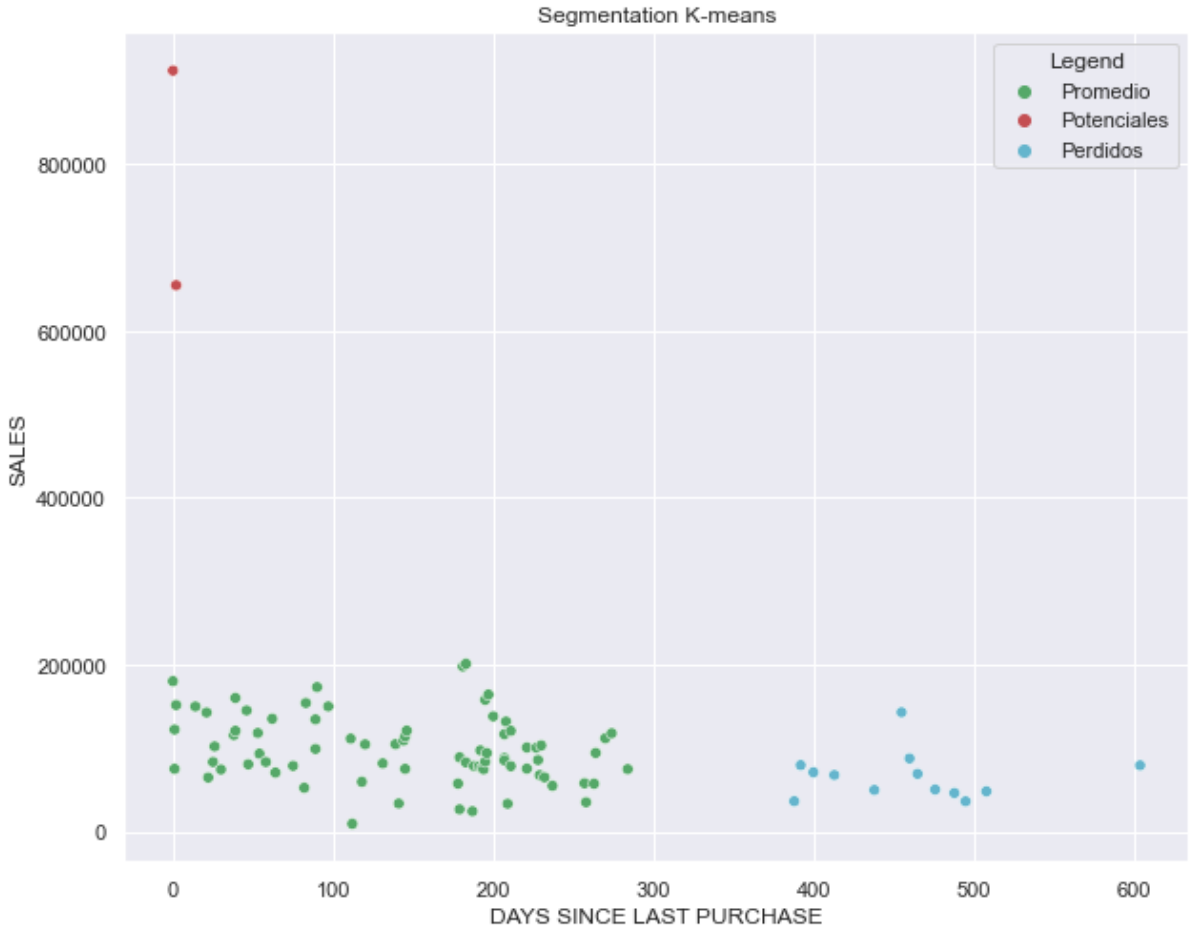

Diagrama 4: Segmentación de los clientes haciendo uso de K-means.

4) Tarea 4: Modelo de recomendación de productos.

Finalmente se elaboró un modelo de recomendación de productos el cual tenía en cuenta la orden del cliente junto con los respectivos productos y sus cantidades.

Una de las formas más prácticas de hacer el sistema de recomendación es calculando la similitud del coseno.

Para ello se tuvo en cuenta el código del producto, la orden a la que pertenece y la línea del producto.

Esta similitud nos puede resultar útil ya que se espera que productos que hayan sido comprados dentro de la misma orden compartan ciertas características como por ejemplo la línea del producto.

Al final se recomendaron 3 combos tanto para el producto más comprado por el cliente en dicha orden como para el producto menos comprado.

CONCLUSIONES.

El trabajo presentado cumple de manera suficiente los requerimientos mínimos de un modelo base de segmentación de clientes y recomendación de productos.

Se puede evidenciar que los principales clientes de la tienda son de Estados Unidos y la línea más vendida es la línea de carros clásicos. Con base en esto, se pueden optar estrategias para evaluar el precio de los productos con el fin de incrementar las ventas y reducir la cantidad de unidades producidas de aquellos productos que no son altamente comprados.

Se pueden llevar a cabo campañas de promoción de productos en ciertas épocas del año, lo cual, según mi experiencia en segmentación de mercados, puede influir en el comportamiento de los clientes.

Considero que la cantidad de clientes perdidos, haciendo uso del modelo de K-means, es un número bastante significativo por lo que la empresa debe evaluar medidas con el fin de incentivar la recurrencia en la compra de productos por parte de este grupo de empresas.

Con respecto al modelo de recomendación se podrían llevar a cabo modelos con pesos en donde las cantidades ordenadas y las ventas de cada producto sean el criterio mediante el cual se recomienden los productos. Esto sería un modelo muy similar a aquellos usados con películas en servicios como el de Netflix, solo que en este caso los pesos se calculan con base en el rating de la película.