PRUEBA DE SELECCIÓN – Claro Insurance
INFORME

INTRODUCCIÓN

El presente informe documenta el análisis de los datos correspondientes al conjunto de datos originalmente de Penthaho DI Kettle, de una tienda minorista de venta de autos y motocicletas. Inicialmente se realizó una exploración de los datos que se describe en la sección siguiente, en seguida se realizó la segmentación de clientes bajo los criterios que se describen más adelante, usando 2 técnicas de segmentación diferentes: segmentación RFM y segmentación por Clustering.

Por último, se propuso un modelo de sistema de recomendación con el que se sugieren productos a los clientes con base en sus compras previas. El periodo de estudio es entre el 6 de Enero de 2003 y el 31 de mayo de 2005.

1. Análisis descriptivo

Se leen los datos desde una copia de estos en un repositorio personal en GitHub, el cual permanecerá público únicamente mientras se realiza la calificación de la presente prueba para garantizar el acceso desde el notebook adjunto. Usando este método de adquisición de datos no se requiere ninguna autenticación al ejecutar el Notebook.

Se leyeron los datos mediante la función *pd.read_csv,* almacenándolos inicialmente en 4 variables correspondientes a cada una de las tablas que componen el caso de estudio.

En seguida se exploraron las 5 primeras entradas de cada tabla para dar un vistazo en la información, obteniendo los siguientes resultados.

Tabla1. Primeras 5 entradas de la tabla “Cliente”.

	ID_Cliente	CUSTOMERNAME	PHONE	ADDRESSLINE1	ADDRESSLINE2	CITY	STATE	POSTALCODE	COUNTRY	TERRITORY	CONTACTLASTNAME	CONTACTFIRSTNAME
0	C001	Alpha Cognac	61.77.6555	1 rue Alsace-Lorraine	NaN	Toulouse	NaN	31000	France	EMEA	Roulet	Annette
1	C002	Amica Models & Co.	011-4988555	Via Monte Bianco 34	NaN	Torino	NaN	10100	Italy	EMEA	Accorti	Paolo
2	C003	Anna's Decorations, Ltd	299368555	201 Miller Street	Level 15	North Sydney	NSW	2060	Australia	APAC	O'Hara	Anna
3	C004	Atelier graphique	40.32.2555	54, rue Royale	NaN	Nantes	NaN	44000	France	EMEA	Schmitt	Carine
4	C005	Australian Collectables, Ltd	61-9-3844-6555	7 Allen Street	NaN	Glen Waverly	Victoria	3150	Australia	APAC	Connery	Sean

Tabla2. Primeras 5 entradas de la tabla “DetalleOrden”.

	ORDERNUMBER	ID_Cliente	QUANTITYORDERED	PRICEEACH	ORDERLINENUMBER	SALES	PRODUCTCODE	DEALSIZE
0	10107	C046	30	95.70	2	2871.00	S10_1678	Small
1	10121	C068	34	81.35	5	2765.90	S10_1678	Small
2	10134	C048	41	94.74	2	3884.34	S10_1678	Medium
3	10145	C087	45	83.26	6	3746.70	S10_1678	Medium
4	10159	C024	49	100.00	14	5205.27	S10_1678	Medium

PRUEBA DE SELECCIÓN – *Claro Insurance*. Informe
Andrés Amaya Chaves

Tabla3. Primeras 5 entradas de la tabla “Orden”.

	ORDERNUMBER	ORDERDATE	DAY_ID	QTR_ID	MONTH_ID	YEAR_ID	STATUS
0	10100	1/06/2003 0:00	6	1	1	2003	Shipped
1	10101	1/09/2003 0:00	9	1	1	2003	Shipped
2	10102	1/10/2003 0:00	10	1	1	2003	Shipped
3	10103	1/29/2003 0:00	29	1	1	2003	Shipped
4	10104	1/31/2003 0:00	31	1	1	2003	Shipped

Tabla4. Primeras 5 entradas de la tabla “Producto”.

	PRODUCTCODE	PRODUCTLINE	MSRP
0	S10_1678	Motorcycles	95
1	S10_1949	Classic Cars	214
2	S10_2016	Motorcycles	118
3	S10_4698	Motorcycles	193
4	S10_4757	Classic Cars	136

Tabla5. Extensión de cada una de las tablas anteriores.

```
{'LenCliente': 92,
 'LenDetalleOrden': 2823,
 'LenOrden': 307,
 'LenProducto': 109}
```

Para realizar el análisis se planteó determinar los siguientes valores (más adelante se encuentra la descripción detallada y gráficas de cada caso):

- **(1A)** Productos más vendidos. Productos menos vendidos.
- **(1B)** Histograma volumen de compras por cada cliente, cantidad total ordenada durante to do el periodo de estudio. Media, desviación estándar y varianza (μ, σ, σ^2)
- **(1C)** Distribución de las ventas en el tiempo. Mes con máximo en ventas, mes con mínimo en ventas.
- **(1D)** Distribución de los clientes según ciudad. Ciudad donde se ubica la mayor cantidad de clientes.
- **(1E)** Distribución de los clientes según estado. Estado donde se ubica la mayor cantidad de clientes.

(1A) Productos más vendidos. Productos menos vendidos.

PRUEBA DE SELECCIÓN – *Claro Insurance*. Informe
Andrés Amaya Chaves

Se realizó una Exploración de la proporción de valores vacíos en cada una de las tablas, cuyi resultado se encuentra en la Tabla 6.

Tabla 6. Proporción valores vacíos en cada tabla

Orden →

ORDERNUMBER	0.0
ORDERDATE	0.0
DAY_ID	0.0
QTR_ID	0.0
MONTH_ID	0.0
YEAR_ID	0.0
STATUS	0.0

Cliente →

ID_Cliente	0.000000
CUSTOMERNAME	0.000000
PHONE	0.000000
ADDRESSLINE1	0.000000
ADDRESSLINE2	0.902174
CITY	0.000000
STATE	0.500000
POSTALCODE	0.032609
COUNTRY	0.000000
TERRITORY	0.413043
CONTACTLASTNAME	0.000000
CONTACTFIRSTNAME	0.000000

DetalleOrden →

ORDERNUMBER	0.0
ID_Cliente	0.0
QUANTITYORDERED	0.0
PRICEEACH	0.0
ORDERLINENUMBER	0.0
SALES	0.0
PRODUCTCODE	0.0
DEALSIZE	0.0

Producto →

PRODUCTCODE	0.0
PRODUCTLINE	0.0
MSRP	0.0

Se construyó una tabla donde se muestra la cantidad de ítems vendidos de cada uno de los productos, mostrando además la línea de producto a la cual corresponde.

Tabla 7. Volumen de ventas por cada producto por cada línea.(head(10)).

	PRODUCTCODE	PRODUCTLINE	MSRP	QUANTITYORDERED
39	S18_3232	Classic Cars	169	1774
76	S24_3856	Classic Cars	140	1052
50	S18_4600	Trucks and Buses	121	1031
106	S700_4002	Planes	74	1029
14	S12_4473	Trucks and Buses	118	1024
77	S24_3949	Planes	68	1008
91	S50_1341	Vintage Cars	43	999
16	S18_1097	Trucks and Buses	116	999
29	S18_2432	Trucks and Buses	60	998
18	S18_1342	Vintage Cars	102	997

Tabla 8. Volumen de ventas por cada producto por cada línea (tail(10)).

	PRODUCTCODE	PRODUCTLINE	MSRP	QUANTITYORDERED
13	S12_3990	Classic Cars	79	800
32	S18_2795	Vintage Cars	168	789
58	S24_1785	Planes	109	784
71	S24_3191	Classic Cars	85	779
48	S18_4409	Vintage Cars	92	750
30	S18_2581	Planes	84	746
78	S24_3969	Vintage Cars	41	745
26	S18_2248	Vintage Cars	60	743
54	S24_1046	Classic Cars	73	724
53	S18_4933	Classic Cars	71	714

De las Tablas 7 y 8 se deduce:

```
----------------------------------------
De los  109 productos disponibles, aquel con la mayor cantidad de unidades vendidas corresponde al siguiente:
   PRODUCTCODE   PRODUCTLINE  MSRP  QUANTITYORDERED
39    S18_3232  Classic Cars   169             1774
----------------------------------------
De los  109 productos disponibles, aquel con la menor cantidad de unidades vendidas corresponde al siguiente:
   PRODUCTCODE   PRODUCTLINE  MSRP  QUANTITYORDERED
53    S18_4933  Classic Cars    71              714
----------------------------------------
```

(1B) Histograma volumen de compras por cada cliente.

Cantidad total de ítems ordenados durante todo el periodo de estudio.
Media, desviación estándar y varianza (μ, σ, σ^2)

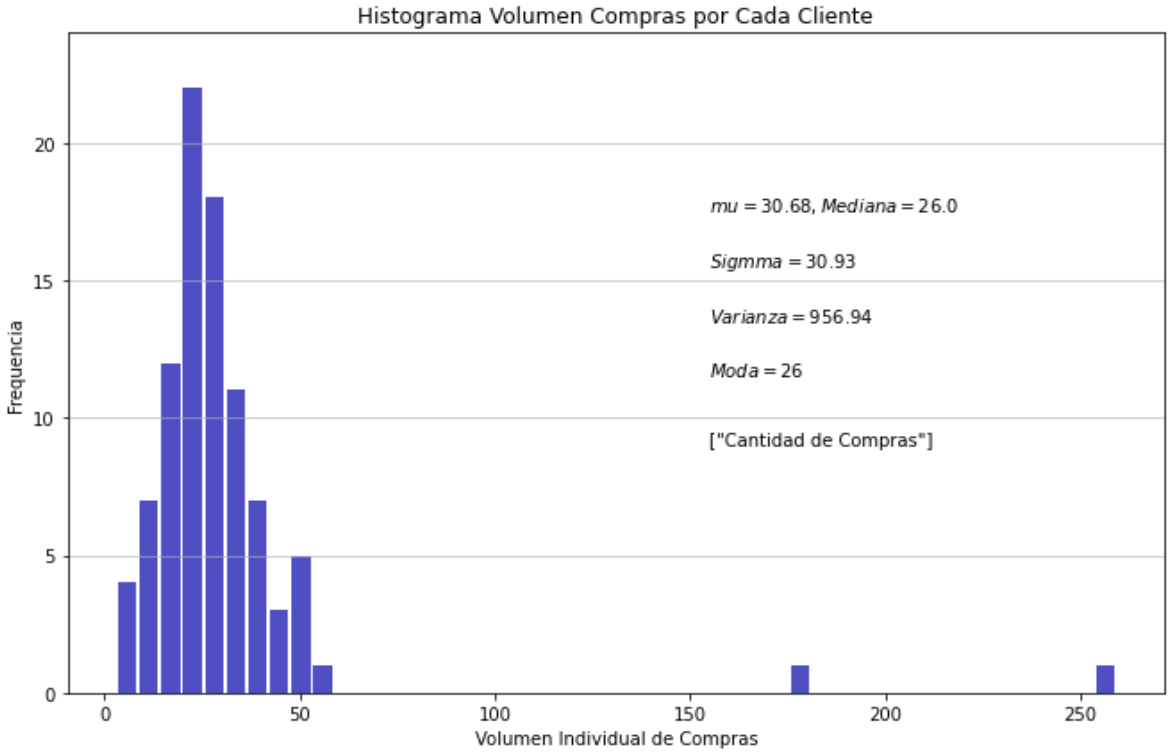


Figura 1. Histograma Volumen de compras por cada cliente.

En la figura 1 se observan las siguientes medidas de tendencia central para el volumen de compras por cliente:

```
{'mu': 30.68, 'Mediana': 26.0, 'Sigma': 30.93, 'Varianza': 956.94, 'Moda': 26}
```

Se observan dos datos bastante excéntricos por encima de 150 compras.

Estos dos valores pueden provenir de datos erróneos al registrar en la base de datos, los cuales causan *outliers* en el conjunto de datos. Tratándose de información agregada, es posible que existan uno o varios datos mal ingresados en los registros de compra.

(1C) Identificación de los clientes con mayor volumen de compras.

Se filtraron los datos para cantidades superiores a 150 compras, obteniendo únicamente 2 clientes.

Tabla 9. Clientes con datos excéntricos en volumen de compras.

	ID_Cliente	QUANTITYORDERED	CUSTOMERNAME	PHONE	ADDRESSLINE1	ADDRESSLINE2	CITY	STATE	POSTALCODE	COUNTRY	TERRITORY	CONTACTLASTNAME	CONTACTFIRSTNAME
0	C034	259	Euro Shopping Channel	(91) 555 94 44	C/ Moralzarzal, 86	NaN	Madrid	NaN	28034	Spain	EMEA	Freyre	Diego
1	C057	180	Mini Gifts Distributors Ltd.	4155551450	5677 Strong St.	NaN	San Rafael	CA	97562	USA	NaN	Nelson	Valarie

Se reitera que estos dos datos pueden corresponder a *outliers* dentro del conjunto de datos.

(1D) Distribución del monto de ventas en el tiempo.

Mes con máximo en ventas, mes con mínimo en ventas.

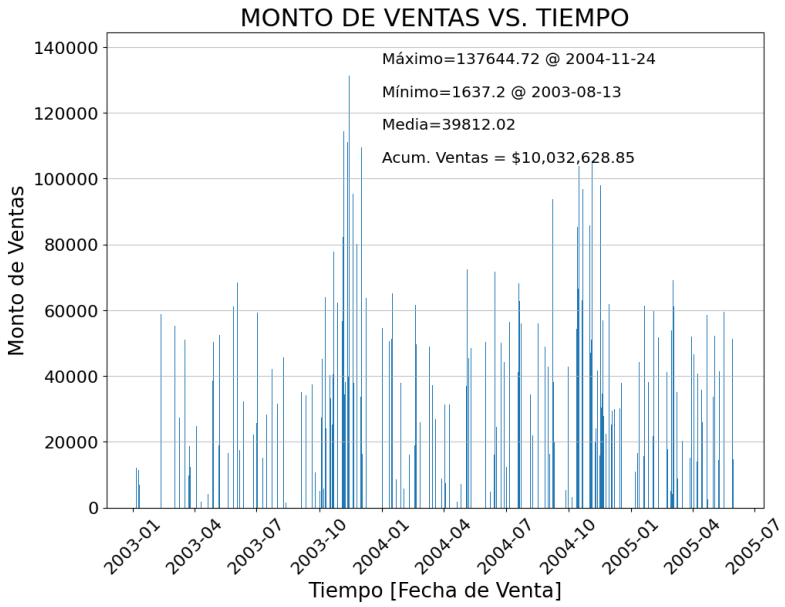


Figura 2. Monto de Ventas vs. Tiempo.

En la figura 2 se observa la distribución del mondo de ventas en el tiempo calculado en una base diaria. Allí se reporta que en el periodo de estudio:

- La fecha con **mayores ventas** de la empresa fue el 24 de noviembre de 2004, con un valor de 137644.72.
- La fecha con **menores ventas** de la empresa fue el 13 de agosto de 2003, con un valor de1637.2.
- Durante el periodo de estudio en total se acumularon 10’032.628.85 en ventas.

A continuación, en la Figura 3 se observa la distribución de las ventas agregadas en una base mensual, correspondiente a la Tabla 10.

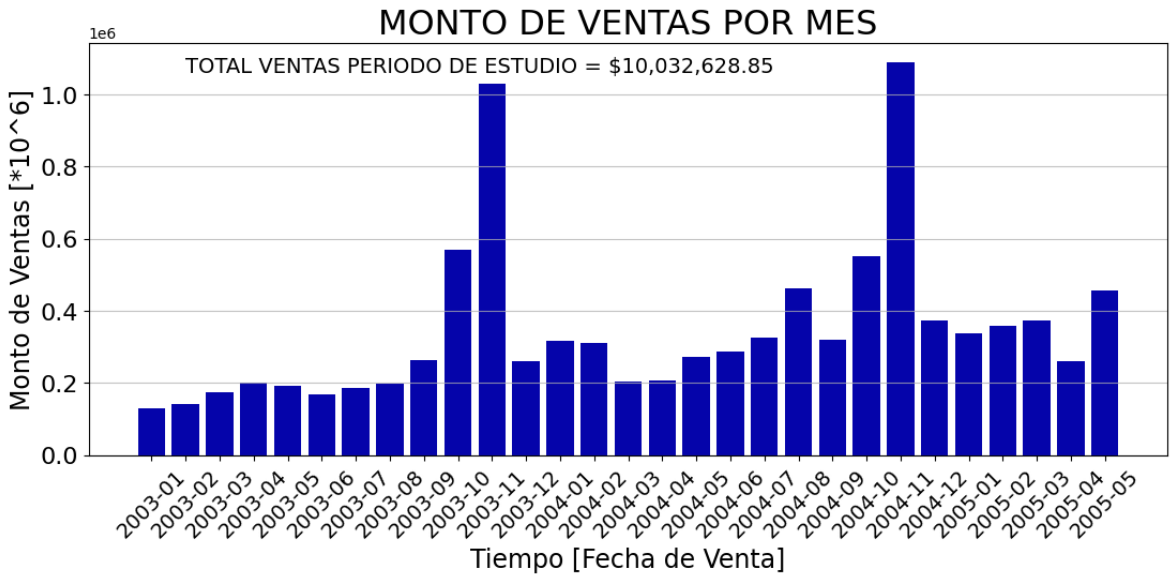


Figura 3. Volumen de ventas en base mensual.

Tabla 10. Tabla volumen de ventas en base mensual (head(5)).

	SALES
YearMonth	
2003-01-01	129753.60
2003-02-01	140836.19
2003-03-01	174504.90
2003-04-01	201609.55
2003-05-01	192673.11

Consistentemente en ambos años completos que abarca el periodo de estudio (i.e. 2003 y 2004) se observa un gran aumento en las ventas en los meses de octubre y noviembre, especialmente en este último, lo cual causa que el segundo semestre del año tenga ventas especialmente altas, se observa un aumento en ventas en el año 2004 respecto al año 2003, y de igual manera, los 5 primeros meses del 2005, de los que se tiene registro, muestran mayores ventas que en esos mismos meses de los 2 años anteriores.

Se señala que la suma de las ventas puede estar distorsionada en cierta medida dado que la información base no presenta valores absolutos, pues según se observa en los rangos correspondientes a cada línea de producto, la columna ‘PRICEEACH de la tabla ‘’DetalleOrden’

muestra valores regularizados entre 0 y 100, con lo cual se deduce que (1) pueden estar representando un porcentaje de un único valor, o (2) pueden estar representando un porcentaje en referencia a varios valores distintos.

Resolviendo el punto 4 del presente informe, la dicotomía anterior se resuelve al observar que en todas las líneas de producto el valor máximo observado es igual a 100, de manera que se puede deducir que **los precios en la columna 'PRICEEACH' de la tabla "DetalleOrden" representan un porcentaje del valor máximo de la línea a la cual el producto pertenece.** Entonces:

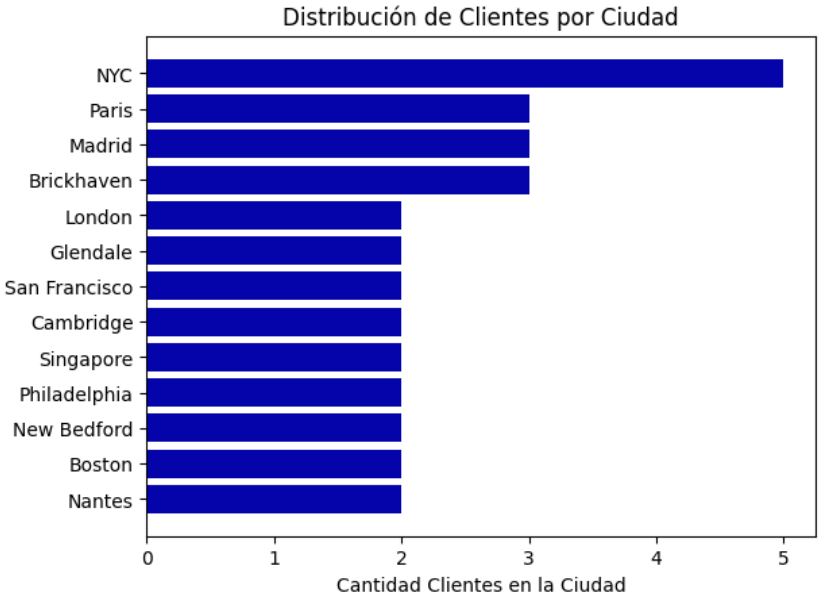
$$'PRICEEACH' = \frac{PrecioItem}{\max(LíneaProd_{Item})} * 100$$

Es importante notar esta transformación en los datos, pues si el valor máximo en cada línea de producto es distinto, las sumatorias que se realizaron para poder acumular los valores de ventas no representan directamente un valor monetario real.

Dicho valor podría ser calculado desagregando los valores correspondientes a cada línea de producto; y para lograrlo sería necesario conocer los máximos valores monetarios reales correspondientes a cada línea de producto.

(1E) Distribución de los clientes según ciudad. Ciudad donde se ubica la mayor cantidad de client es.

Figura 4. Distribución de Clientes por ciudad



NYC es la ciudad con mayor cantidad de clientes en el periodo de estudio. Allí se ubican 5 clientes. Las ciudades donde se ubican al menos 2 clientes son 13 y se muestran en la figura 4; existen al menos 2 clientes dentro de estas ciudades. Las otras 60 ciudades registran un solo cliente cada una.

(1F) Distribución de los clientes según estado. Estado donde se ubica la mayor cantidad de client es.

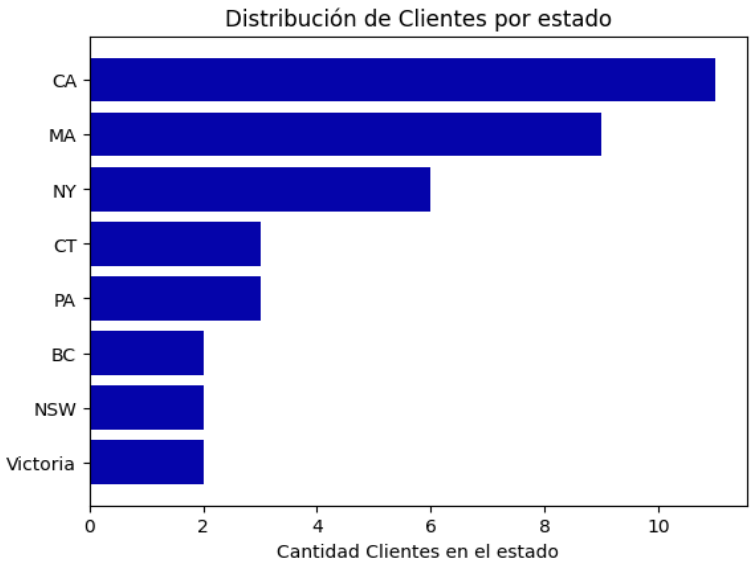


Figura 5. Distribución de Clientes por estado.

California es el estado con mayor cantidad de clientes en el periodo de estudio. Allí se ubican 11 clientes. Los estados donde se ubican al menos 2 clientes son 8 y se muestran en la figura; existen al menos 2 clientes dentro de estos estados. Los otros 8 estados registran un solo cliente cada uno.

2. Segmentación de Clientes RFM

Se usaron las siguientes variables para describir las caracteristicas RFM, en todos los casos se trata de i nformación agregada de acuerdo a cada uno de los clientes:

- Recency: ORDERDATE. Tiempo desde la última compra hasta la fecha de referencia.
- Freceuncy: ORDERNUMBER. Cantidad de ordenes generadas en el periodo de estudio.
- Money: SALES. Cantidad de dinero en compras en el periodo de estudio.

Inicialmente se calcularon los valores de Recencia, frecuencia y valor monetario, para cada uno de los clientes de la empresa.

Tabla 11. Tabla RFM (head(5)).

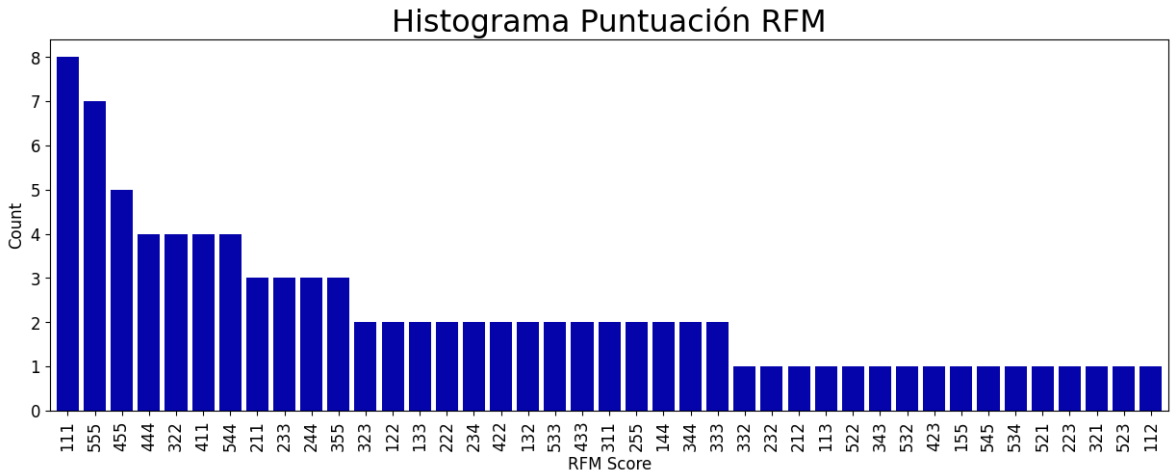
	ID_Cliente	Recency	Frequency	Monetary
0	C001	64	20	70488.44
1	C002	264	26	94117.26
2	C003	83	46	153996.13
3	C004	187	7	24179.96
4	C005	22	23	64591.46

En seguida se calcularon los quintiles para cada una de las variables RFM, resultados que se muestran en la Tabla 12.

Tabla 12. Quintiles RFM.

	Q_Recency	Q_Frequency	Q_Monetary
min	0.0	3.0	9129.350
first_part (Q1)	54.8	18.0	64640.032
second_part (Q2)	144.4	23.0	79323.366
third_part (Q3)	196.6	27.0	100479.962
forth_part(Q4)	257.8	36.0	133744.524
max	508.0	259.0	912294.110

(a)



(b)

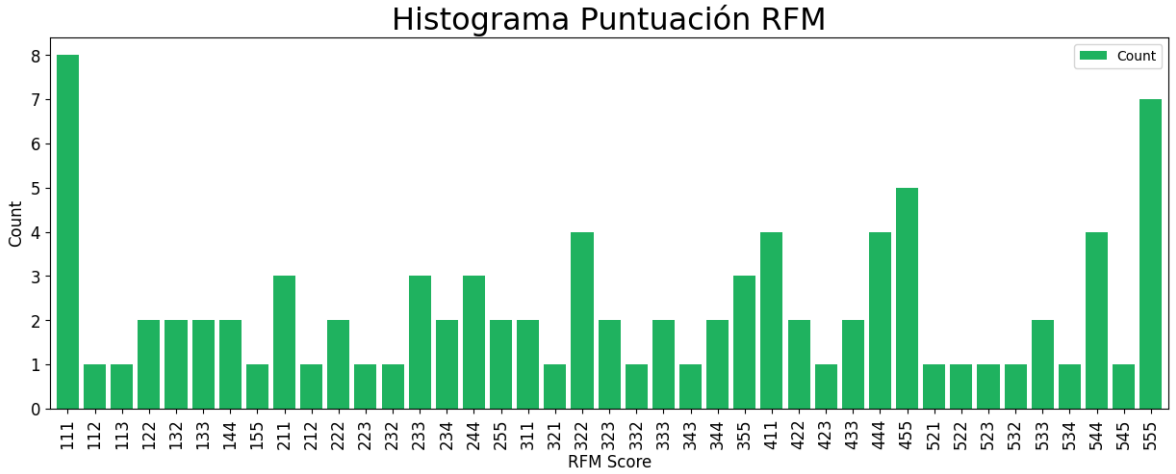


Figura 6. Histograma RFM indvidual. (a) Ordenado según conteo de forma descendiente. (b) Ordenado según consecutivo de la clasificación.

En la Figura 6 se observa el histograma de las clasificaciones individuales según los criterios de la Figura 8. Después, se realizó el histograma de clasificación RFM consolidado, obteniendo la Figura 7.

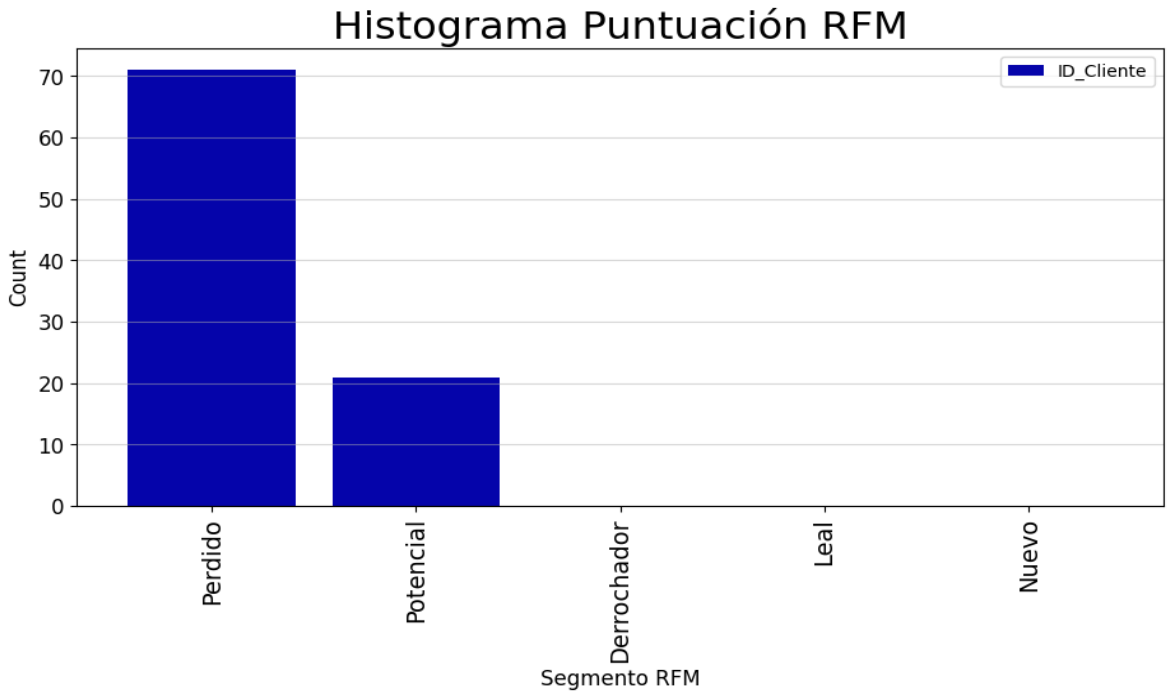


Figura 7. Histograma RFM consolidado.

Tabla 1 Clasificación de las empresas según el RFM

Recencia	Frecuencia	Valor Monetario
5, más reciente	5, Más frecuente	5, valor más alto
4	4	4
3	3	3
2	2	2
1 más antiguo	1, menos frecuente	1, valor más bajo

Fuente: Elaboración propia.

Teniendo en cuenta estos resultados, clasifican las empresas en 5 grupos (Tabla 5).

Tabla 5. Clasificación de los grupos según los resultados del RFM

Segmentos	R	F	M	Descripción
Potenciales	4,5	4,5	4,5	Son los clientes que menos días tardan en comprar, los que más a menudo lo realizan y son los que más gastan.
Perdidos	1,2, 3	1,2, 3	1,2, 3	Son clientes que compraron hace mucho tiempo, poca cantidad y poco gasto.
Derrochadores			5	Los que más gastan
Leales		4,5		Son los que más frecuente van a comprar
Nuevos	4,5			Clientes que han comprado hace poco pero no a menudo.

Figura 8. Criterios de clasificación RFM.

Se cuenta con 71 clientes clasificados como “perdidos”, y 21 clasificados como “potenciales”, no se registra ningún cliente clasificado como “derrochador,”, “leal”, ni “nuevo”.

3. Segmentación por Clustering

Se utilizó método *kmeans* para la segmentación RFM.

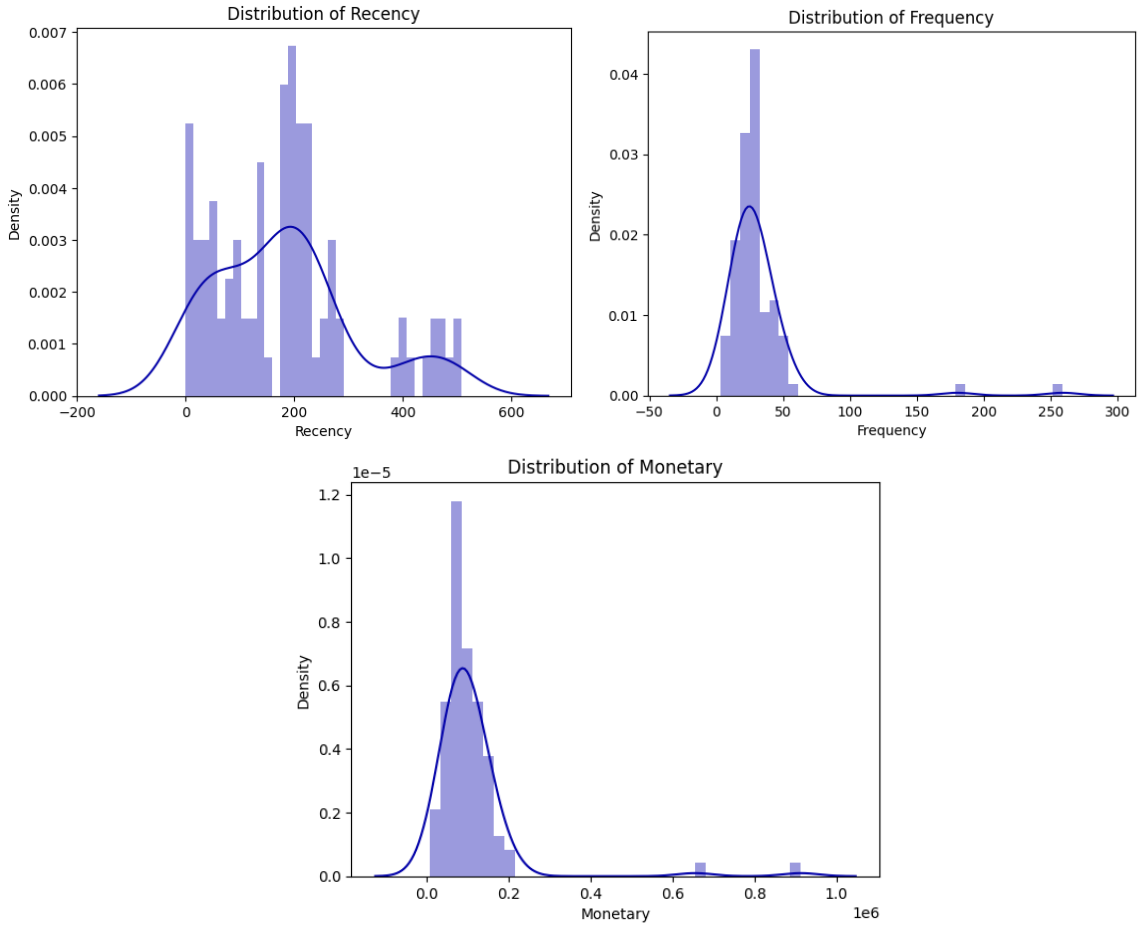


Figura 8. Distribución de la información.

Para comenzar, tomando las mismas variables de Recencia, Frecuencia y Valor Monetario ya construidas en la sección anterior, se trazaron las distribuciones de cada una de ellas con el fin de verificar su grado de asimetría (*skewness*), el resultado fue la Figura 8; como complemento de esta, a continuación se encuentran los valores numéricos correspondientes (incluyendo estadístico y p-value).

```
Recency's: Skew: 0.7673634065510251, : SkewtestResult(statistic=2.9211180198423654, pvalue=0.003487776773306202)
Frequency's: Skew: 5.653296151215707, : SkewtestResult(statistic=9.653805927007603, pvalue=4.736448574059028e-22)
Monetary's: Skew: 5.570979953309047, : SkewtestResult(statistic=9.598373148084844, pvalue=8.121615124734495e-22)
```

Las tres variables se encuentran sesgadas positivamente, la Recencia es bastante simétrica, si sesgo es muy bajo. Tanto la variable Frecuencia como Valor Monetario, están bastante

sesgadas hacia la derecha, lo cual se explica por los 2 valores excéntricos de los que se habló en la primera sección (y se calificaron como posibles '*outliers*').

Se procedió a realizar transformaciones logarítmicas a los datos para reducir la asimetría, el resultado se encuentra en la Figura 9.

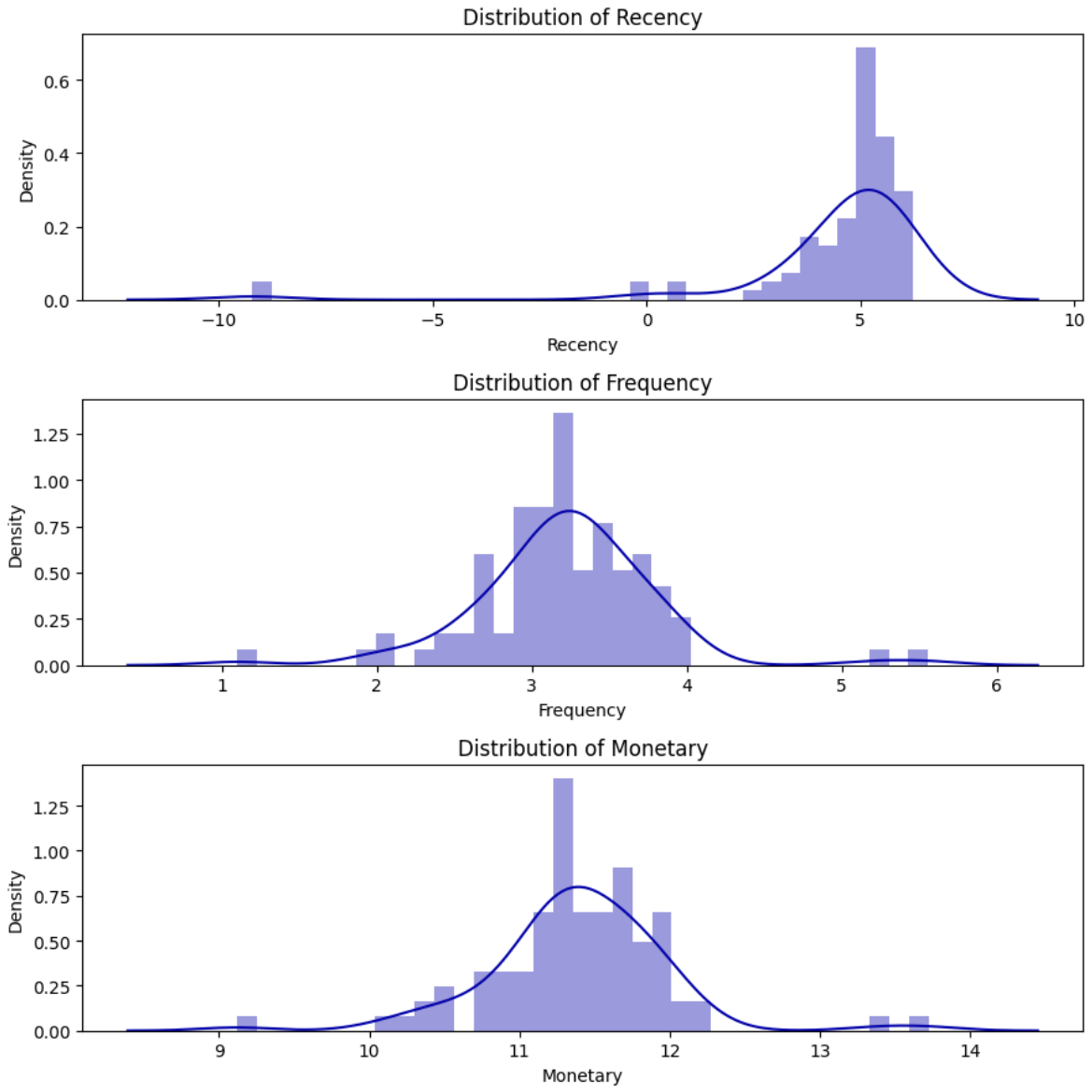


Figura 9. Distribuciones tras las transformaciones tipo 'log'.

Adicionalmente se realizó un proceso de escalado de los datos, para poder hacer comparables los datos entre un cliente y otro, y de esa manera hacer posible el agrupamiento (*clustering*) de ellos.

En seguida, se procedió a buscar el número (k) de clusters más conveniente según los datos. Para ello, se utilizó el "método del codo" (*the elbow method*).

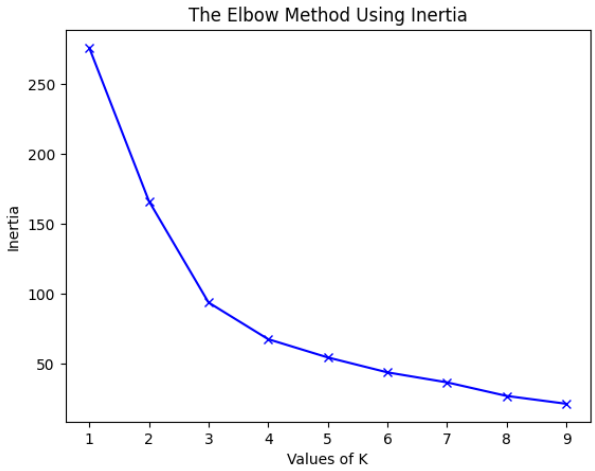


Figura 10. Resultado del "método del codo".

Con este método se observan las deflexiones de la curva de inercia vs. Valores de k, y será más conveniente aquel valor de k al que le corresponda la mayor "curvatura" de la función.

En este caso, K = 3 parece un valor apropiado, pues allí es donde la gráfica presenta mayor curvatura. Sin embargo, a continuación se exploran las gráficas 'aplanadas' que resultan de usar K=3, K=4 y K=5.

Estas son gráficas que ubican los datos en sus respectivos *clusters,* son el resultado del modelo, y corresponden a una tridimensionalidad, pero se 'aplanan' para poder ser representadas en 2 dimensiones. Ver resultado en la Figura 11.

Se efectuaron los 'gráficos de serpiente' (*snake plots*) correspondientes a cada valor de K (3, 4,5), los cuales se pueden observar en la Figura 12.

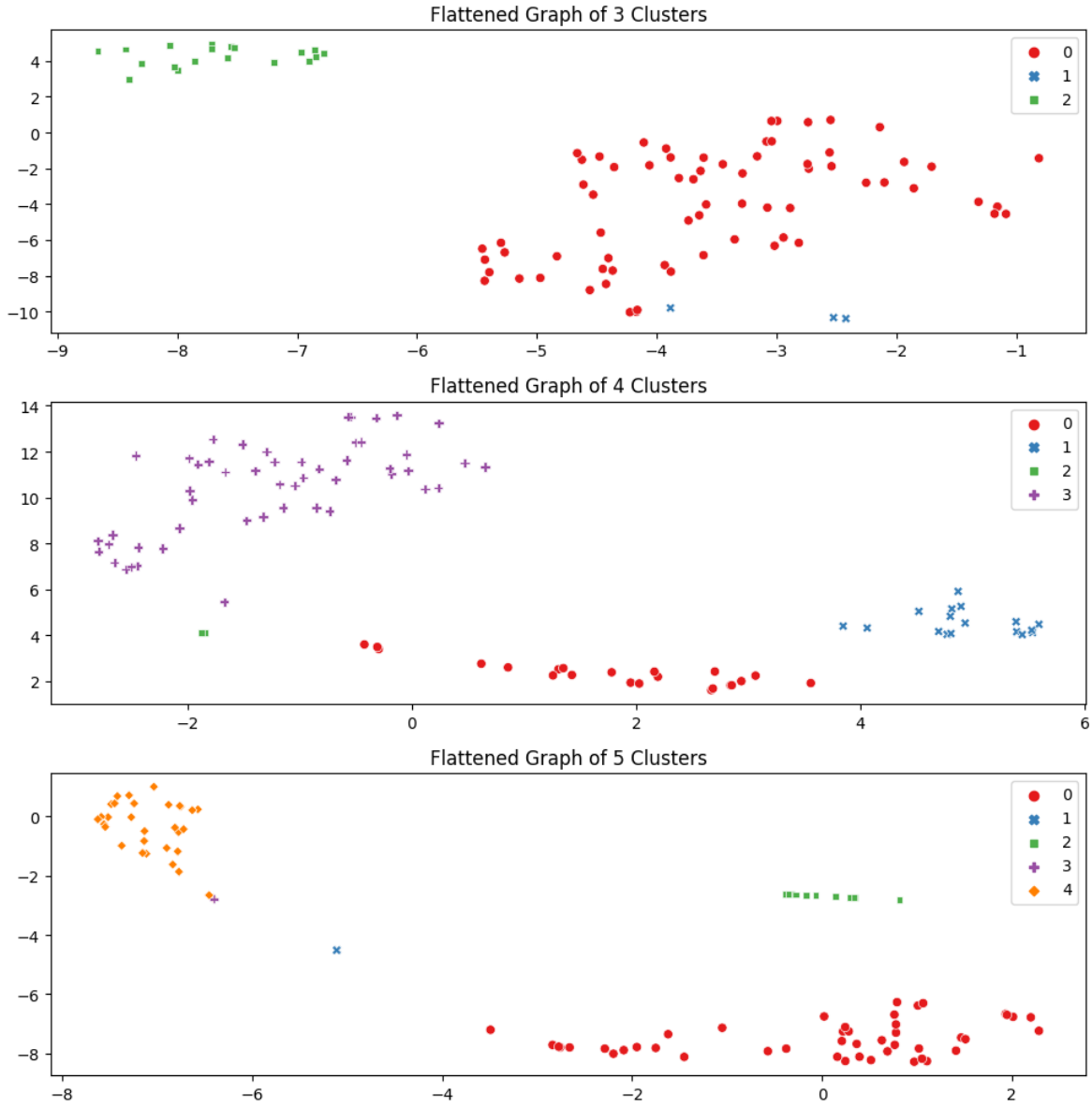
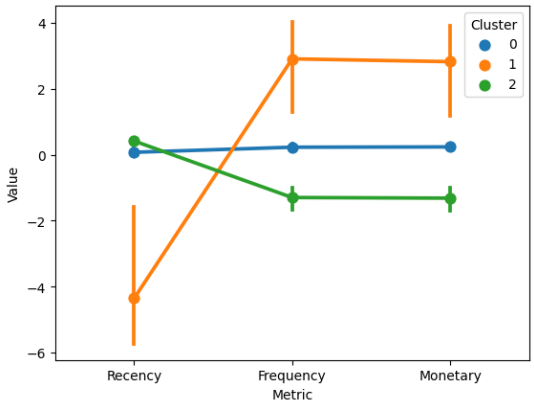
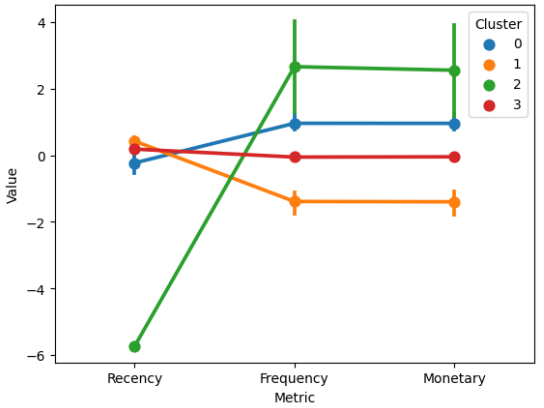


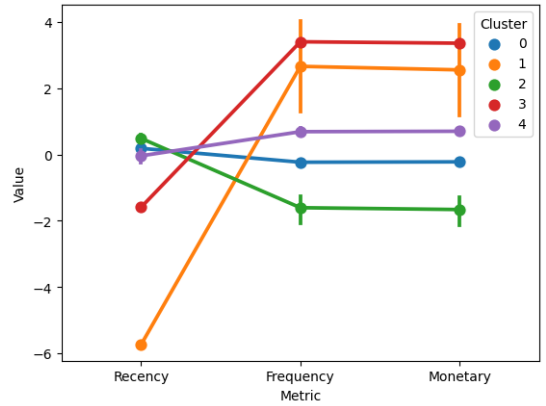
Figura 11. Gráficos de *clusters*, aplanados



	Recency	Frequency	Monetary	
	mean	mean	mean	count
Cluster				
0	5.0	3.0	12.0	70
1	-6.0	5.0	13.0	3
2	5.0	2.0	11.0	19



Cluster	Recency mean	Frequency mean	Monetary mean	count
0	4.0	4.0	12.0	22
1	6.0	2.0	11.0	17
2	-9.0	5.0	13.0	2
3	5.0	3.0	11.0	51



Cluster				
0	5.0	3.0	11.0	46
1	-9.0	5.0	13.0	2
2	6.0	2.0	10.0	12
3	1.0	5.0	13.0	1
4	4.0	4.0	12.0	31

Figura 12. 'Gráficos de serpiente' para k = [3,4,5].

Comparación de técnicas (2) y (3).

Al comparar la técnica (2) con el modelo de 3 *clusters*, se observa bastante similitud, en cuanto a que el de *k-means* tiene un *cluster* casi vacío, y solamente hay 2 de ellos que tienen valores significativos, y coinciden con el tipo de segmento y (muy cerca también) con las cantidades que se hallaron con el primer método.

Dado que el "método del codo" muestra que la razón de cambio entre la inercia del sistema y el valor de k varía más fuertemente donde k=3, entonces se deduce que usar 3 *clusters* es la mejor opción para los datos del problema; dada la similitud de este modelo con el desarrollado en la primera sección, se puede concluir que el primer modelo es muy adecuado para la segmentación de los clientes de este problema.

4. Modelo de Sistema de Recomendación

Para crear el modelo de sistema de recomendación se siguieron los principios del método *Matrix Factorization for Recommendation Systems*, adaptados a este problema específico; en primera medida se observó que los productos están agrupados por 'líneas de producto'; también se prestó atención a que el registro de la tabla *'DetalleOrden'* permite calcular cuáles son los productos más comprados por cada cliente, así como la cantidad comprada por cliente por producto.

El método *Matrix Factorization* usa *ratings* para basar la recomendación, en este caso, lo lógico es utilizar la cantidad de compras de un mismo producto como el *rating* que cada cliente da a cada producto. De acuerdo con lo anterior, se definió lo siguiente:

> *"El sistema recomendará 3 productos a cada cliente, los cuales deberán ser:*
>
> *(a) De aquellos que se encuentren dentro de la línea del producto más comprado por cada cliente, los 2 productos más cercanos hacia arriba en precio.*
> *(b) De aquellos que se encuentren dentro de la línea del segundo producto más comprado por cada cliente, el producto más cercano hacia arriba en precio.*
> *(c) Si los 2 productos de (a) son de la misma línea, entonces la tercera recomendación será el tercer producto más cercano hacia arriba en precio, dentro de esa línea.*
> *(d) Para el caso en que el 'productoA' más comprado corresponda a alguno de los 3 productos de más alto precio de la línea, se procederá así:*
> *a. De la misma línea del 'productoA' (aquel con la mayor frecuencia de compra) se recomendarán los 3 productos más cercanos hacia abajo en precio al 'productoA'.*

*Se toman en cuenta todas las compras realizadas por cada cliente d*urante todo el periodo de estudio.

A continuación, se calcularon (a manera de *rating*) las veces que un cliente incluyó cada uno de los productos en una orden de compra, obteniendo la tabla 13.

Tabla 13. Cantidad de órdenes por cada cliente por cada producto.

	ID_Cliente	PRODUCTCODE	Quantity_of_Orders
1	C001	S72_1253	12
2	C001	S72_3212	5
0	C001	S18_3232	3
3	C002	S24_4258	17
4	C002	S72_3212	9

Ahora se creó un DataFrame que almacena los 3 productos más cercanos hacia arriba en precio, y demás condiciones descritas al principio de esta sección. Para ello, inicialmente se construyeron 7 variables de tipo DataFrame, una por cada línea de producto, en las cuales se ordenaron los productos de cada línea de acuerdo con su precio, en orden ascendente. El resultado se explora en la Tabla 14.

Los últimos valores de las 3 últimas columnas siguen las disposiciones descritas para los productos con más alto precio.

Tabla 14. 3 productos más cercanos hacia arriba en precio a cada producto de cada línea.

	PRODUCETLINE	PRODUCTCODE	PRICEEACH	ProdRec1	ProdRec2	ProdRec3
0	Classic Cars	S24_2840	44.010357	S24_2972	S24_1628	S24_1444
1	Classic Cars	S24_2972	45.498519	S24_1628	S24_1444	S24_3371
2	Classic Cars	S24_1628	52.849167	S24_1444	S24_3371	S24_1046
3	Classic Cars	S24_1444	61.533571	S24_3371	S24_1046	S18_4933
4	Classic Cars	S24_3371	62.048846	S24_1046	S18_4933	S12_3990
...	...	...	...	...	...	...
104	Vintage Cars	S18_3320	94.165385	S18_2795	S18_3140	S18_2325
105	Vintage Cars	S18_2795	95.105769	S18_3140	S18_2325	S18_1749
106	Vintage Cars	S18_3140	97.225600	S18_2325	S18_1749	S18_2325
107	Vintage Cars	S18_2325	98.547200	S18_1749	S18_2325	S18_3140
108	Vintage Cars	S18_1749	98.820000	S18_2325	S18_3140	S18_2795

109 rows × 6 columns

Para satisfacer la necesidad de asignar un producto recomendado, en el caso de los productos con más alto precio, se presentó la siguiente situación:

> *Como se observa en la Tabla 14, la línea "Trains" solamente tiene 3 productos asociados, entonces al tratar de efectuar la tercera recomendación para el producto más costoso, se requerían los 3 productos cuyo precio fuera menor al máximo de ellos, lo cual implica que hacer la recomendación con ese criterio sería necesario que la línea contara con al menos 4 productos, y no es el caso.*

Tabla 14. Cantidad de productos que conforman cada línea.

	PRODUCTLINE	QUANTITY_PRODUCTS
0	Classic Cars	37
1	Motorcycles	13
2	Planes	12
3	Ships	9
4	Trains	3
5	Trucks and Buses	11
6	Vintage Cars	24

Para resolver esa situación, se procedió a recomendar el cuarto producto más caro de la línea con promedio de precio inmediatamente siguiente a la línea del producto en cuestión, únicamente para las líneas con menos de 4 productos (para este caso, únicamente "Trains").

Para efectuar tal procedimiento, fue necesario calcular los promedios de los precios de cada línea, lo cual se muestra en la Tabla 15.

Tabla 15. Promedio de precios de cada línea.

PRODUCTLINE	PRODUCTLINE	PRICEEACH
Trains	Trains	75.654675
Vintage Cars	Vintage Cars	78.148204
Planes	Planes	81.740915
Motorcycles	Motorcycles	82.997553
Ships	Ships	83.855470
Classic Cars	Classic Cars	87.335781
Trucks and Buses	Trucks and Buses	87.527940

Se construyó una nueva tabla (ver Tabla 15) en la que aparece cada cliente con sus 2 productos más comprados, y a cada producto se le relacionó con 3 productos recomendados "provisionales"; es decir que a cada cliente se le asignaron 6 productos recomendados provisionales (PrdRec1_1, PrdRec1_2, PrdRec2_1, PrdRec2_2, PrdRec3_1, PrdRec3_2)[1].

Tabla 15. Tabla con 6 recomendaciones por cliente.

	index	ID_Cliente	PRODUCTCODE	Quantity_of_Orders	PRODUCTLINE	PRODUCETLINE	PRICEEACH	ProdRec1	ProdRec2	ProdRec3
0	1	C001	S72_1253	12	Planes	Planes	55.897037	S24_2841	S24_3949	S700_4002
1	2	C001	S72_3212	5	Ships	Ships	63.865769	S700_1138	S700_2610	S18_3029
2	3	C002	S24_4258	17	Vintage Cars	Vintage Cars	92.874231	S18_2949	S18_3856	S18_3320
3	4	C002	S72_3212	9	Ships	Ships	63.865769	S700_1138	S700_2610	S18_3029
4	5	C003	S24_3432	14	Classic Cars	Classic Cars	93.649565	S12_1108	S18_1984	S10_4757
...	...	...	...	...	...	...	...	...	...	...
178	299	C090	S50_4713	9	Motorcycles	Motorcycles	84.011200	S10_1678	S32_1374	S10_2016
179	303	C091	S700_1138	6	Ships	Ships	70.566923	S700_2610	S18_3029	S700_1938
180	304	C091	S24_2300	6	Trucks and Buses	Trucks and Buses	100.000000	S12_1666	S18_4600	S50_1392
181	306	C092	S24_4258	10	Vintage Cars	Vintage Cars	92.874231	S18_2949	S18_3856	S18_3320
182	305	C092	S700_3167	3	Planes	Planes	82.059600	S18_2581	S24_1785	S700_1691

183 rows × 10 columns

A partir de las recomendaciones provisionales se formó finalmente un conjunto de 3 productos recomendados por cliente siguiendo los criterios ya descritos arriba; el resultado se puede apreciar en la Tabla 15, la cual era el objetivo de este modelo de sistema de recomendación.

Tabla 15. Resultado general de la recomendación.

[1] Se denomina con "_1" al producto en relación al producto más comprado por cada cliente, y con "_2" al segundo producto más comprado por cada uno.

PRUEBA DE SELECCIÓN – *Claro Insurance*. Informe

Andrés Amaya Chaves

	ID_Cliente	Recomendadcion1	Recomendadcion2	Recomendacion3
0	C001	S24_2841	S24_3949	S700_1138
1	C002	S18_2949	S18_3856	S700_1138
2	C003	S12_1108	S18_1984	S10_1678
3	C004	S24_1444	S24_3371	S18_3782
4	C005	S18_2949	S18_3856	S24_2841
5	C006	S18_2581	S24_1785	S18_2581
6	C007	S700_1691	S700_2466	S18_2248
7	C008	S24_2887	S24_3432	S24_2300
8	C009	S10_1678	S32_1374	S10_1678
9	C010	S24_2841	S24_3949	S18_3320
10	C011	S50_1341	S24_2022	S700_1138
11	C012	S24_2887	S24_3432	S18_3259
12	C013	S24_2841	S24_3949	S24_2841
13	C014	S18_3259	S50_1514	S18_3029
14	C015	S18_3029	S700_1938	S700_1138
15	C016	S700_1138	S700_2610	S18_2949
16	C017	S18_2949	S18_3856	S24_2887
17	C018	S24_2887	S24_3432	S24_2887
18	C019	S24_2887	S24_3432	S700_1138
19	C020	S700_1138	S700_2610	S18_3259
20	C021	S18_3259	S50_1514	S700_1138
21	C022	S700_1138	S700_2610	S24_2766
22	C023	S24_2766	S700_2824	S10_1678
23	C024	S10_1678	S32_1374	S24_2887
24	C025	S24_2887	S24_3432	S10_1678
25	C026	S10_1678	S32_1374	S10_1678
26	C027	S10_1678	S32_1374	S700_1138
27	C028	S700_1138	S700_2610	S24_2887
28	C029	S24_2887	S24_3432	S18_4600
29	C030	S18_4600	S12_1666	S24_2841
30	C031	S24_2841	S24_3949	S18_3259
31	C032	S18_3259	S50_1514	S700_1138
32	C033	S700_1138	S700_2610	S700_1138
33	C034	S700_1138	S700_2610	S18_3232
34	C035	S18_3232	S18_1129	S18_1662
35	C036	S18_1662	S700_2834	S18_2581
36	C037	S18_2581	S24_1785	S700_1138
37	C038	S700_1138	S700_2610	S24_2841
38	C039	S24_2841	S24_3949	S50_1341
39	C040	S50_1341	S24_2022	S700_3962
40	C041	S700_3962	S700_2047	S18_3232
41	C042	S18_3232	S18_1129	S18_4600
42	C043	S18_4600	S12_1666	S24_2887
43	C044	S24_2887	S24_3432	S24_4278
44	C045	S24_4278	S700_3167	S24_2841
45	C046	S24_2841	S24_3949	S18_3232
46	C047	S18_3232	S18_1129	S10_1678
47	C048	S10_1678	S32_1374	S18_2625
48	C049	S18_2625	S18_3782	S18_4600
49	C050	S18_4600	S12_1666	S24_2841
50	C051	S24_2841	S24_3949	S10_1678
51	C052	S10_1678	S32_1374	S24_2887
52	C053	S24_2887	S24_3432	S12_1108
53	C054	S12_1108	S18_1984	S24_4278
54	C055	S24_4278	S700_3167	S24_2841
55	C056	S24_2841	S24_3949	S10_1678
56	C057	S10_1678	S32_1374	S12_4473
57	C058	S12_4473	S18_1097	S10_1678
58	C059	S10_1678	S32_1374	S50_1341
59	C060	S50_1341	S24_2022	S24_4278
60	C061	S24_4278	S700_3167	S18_3232
61	C062	S18_3232	S18_1129	S10_1678
62	C063	S10_1678	S32_1374	S24_4278
63	C064	S24_4278	S700_3167	S700_1138
64	C065	S700_1138	S700_2610	S700_1138
65	C066	S700_1138	S700_2610	S700_1138
66	C067	S700_1138	S700_2610	S24_2841
67	C068	S24_2841	S24_3949	S24_2841
68	C069	S24_2841	S24_3949	S700_1138
69	C070	S700_1138	S700_2610	S18_3320
70	C071	S18_3320	S18_2795	S24_2841
71	C072	S24_2841	S24_3949	S700_1138
72	C073	S700_1138	S700_2610	S700_1138
73	C074	S700_1138	S700_2610	S18_2949
74	C075	S18_2949	S18_3856	S50_1341
75	C076	S50_1341	S24_2022	S24_2841
76	C077	S24_2841	S24_3949	S24_2887
77	C078	S24_2887	S24_3432	S10_2016
78	C079	S10_2016	S32_4485	S24_2887
79	C080	S24_2887	S24_3432	S24_4278
80	C081	S24_4278	S700_3167	S24_4278
81	C082	S24_4278	S700_3167	S24_4278
82	C083	S24_4278	S700_3167	S24_2841
83	C084	S24_2841	S24_3949	S18_2625
84	C085	S18_2625	S18_3782	S24_2841
85	C086	S24_2841	S24_3949	S18_1662
86	C087	S18_1662	S700_2834	S18_3232
87	C088	S18_3232	S18_1129	S24_2766
88	C089	S24_2766	S700_2824	S24_2887
89	C090	S24_2887	S24_3432	S700_2610
90	C091	S700_2610	S18_3029	S18_2949
91	C092	S18_2949	S18_3856	S18_2949

5. CONCLUSIONES Y RECOMENDACIONES

- De acuerdo con la exploración de los datos, se encontró que el volumen de compras por cliente presenta las siguientes medidas de tendencia central:

```
{'mu': 30.68, 'Mediana': 26.0, 'Sigma': 30.93, 'Varianza': 956.94, 'Moda': 26}
```

  Con lo cual se observa una desviación estándar bastante alta, pues se encontró un sesgo positivo muy significativo.

- Lo anterior, aunado a las verificaciones del sesgo, se puede explicar porque se detectaron 2 datos bastante excéntricos (superiores a 150 compras por cliente) que hacen pensar que corresponden a *outliers,* que pudieron ser causados por el ingreso erróneo de datos en alguna(s) de las órdenes de compra de los dos clientes en cuestión.

- Estos datos aparentemente erróneos reflejan su efecto en todas las gráficas que tienen que ver con valor monetario, a lo largo de todos los métodos.

- **Se recomienda** generar los conjuntos de datos pertinentes, que excluyan los dos datos mencionados, para utilizarlos en los modelos aquí desarrollados, obtener resultados acordes a la suposición de que se trata de 2 datos *outliers*, para poder comparar entre los dos escenarios y poder decidir finalmente si esos datos son errados o no.

- Se señala que la suma de las ventas puede estar distorsionada en cierta medida dado que la información base no presenta valores absolutos, se concluye que **los precios en la columna 'PRICEEACH' de la tabla "DetalleOrden" representan un porcentaje del valor máximo de la línea a la cual el producto pertenece.**

- De un total de 73 ciudades, ***New York City*** es la ciudad con mayor cantidad de clientes en el periodo de estudio. Allí se ubican 5 clientes, de un total de 92 clientes, para un **5.4%**.

- De los 16 estados registrados en la base de datos, **California** es el estado con mayor cantidad de clientes en el periodo de estudio, allí se ubican 11 clientes, de un total de 92 clientes, para un **12%**.

- Según el modelo de segmentación RFM, se puede concluir que:

  "Se cuenta con 71 clientes clasificados como "perdidos", y 21 clasificados como "potenciales", no se registra ningún cliente clasificado como "derrochador,", "leal", ni "nuevo"."

- Dado que el "método del codo" muestra que la razón de cambio entre la inercia del sistema y el valor de k varía más fuertemente donde k=3, entonces se deduce que usar 3 *clusters* es la mejor opción para los datos del problema; dada la similitud de este modelo con el desarrollado en la primera sección, se puede concluir que el primer modelo es muy adecuado para la segmentación de los clientes de este problema.

- El modelo de recomendación arrojó los resultados de la Tabla 15.

Para poder probar del nivel de certeza (*accuracy*) del **modelo de sistema de recomendación,** se aconseja realizar partición a los datos (en *Training Set*, *Test Set* y *Validation Set*), y de esta manera tener una idea de qué tan bueno puede ser el modelo en sus predicciones, al compararlas con los datos reales del periodo correspondiente al *test set*.

Tratándose de información de compra, y teniendo en cuenta que hay una variación en el tiempo que tiene que ver con el comportamiento de compra como un continuo, se recomienda particionar de manera **continua**, es decir, no tomando valores aleatorios dentro todo el *dataset,* sino dividiendo en ciertas fechas la información, dejando el {*training set + test set}* antes de una fecha, el *test set* después de esta, y el *Validation Set* a partir de otra fecha.

6. ANEXO

Diagrama de clases del problema