ROSARIO
MONTALBAN SARDI

PROYECTO DE MACHINE LEARNING

SEGMENTACIÓN DE CLIENTES

ANÁLISIS SOBRE LA RECENCIA, FRECUENCIA Y VALOR MONETARIO DE LOS CLIENTES PARA SU SEGMENTACIÓN

THE BRIDGE TECH SCHOOL

ANÁLISIS RFM

El análisis RFM es una técnica de segmentación de clientes basada en los datos que tenemos sobre su recencia, frecuencia y valor monetario.

Los clientes serán segmentados en base a la última vez que compraron, cuántas veces han frecuentado y el dinero que gastaron en total en sus visitas.

Según las puntuaciones, los clientes pertenecerán a un clúster con el fin de elaborar una estrategia de marketing lo más ajustada posible a los perfiles.

LAS FASES DEL PROYECTO

Fase 1

Recogida de datos

Selección de un buen dataset que contenga datos de negocio e IDs únicos de clientes para poder hacer un buen análisis RFM.

Fase 2

Objetivo

Descripción de los datos y de las tendencias de venta.
Selección del objetivo al que aplicarle el análisis. En este caso, Reino Unido.

Fase 3

Limpieza y perfilado de datos

Ubicación y eliminación de valores nulos y negativos.
Creación de nuevas columnas basadas en la información del dataset crudo.

Fase 4

Creación de dataset

Creación de un nuevo dataset agrupado por los clientes únicos en los que las features a tener en cuenta se correspondan con la Recencia, Frecuencia y Valor monetario.

Fase 5

Transformación de los datos

Detección y extracción de (algunos) outliers. Comprobación de la distribución de los datos.
Transformación del dataset.

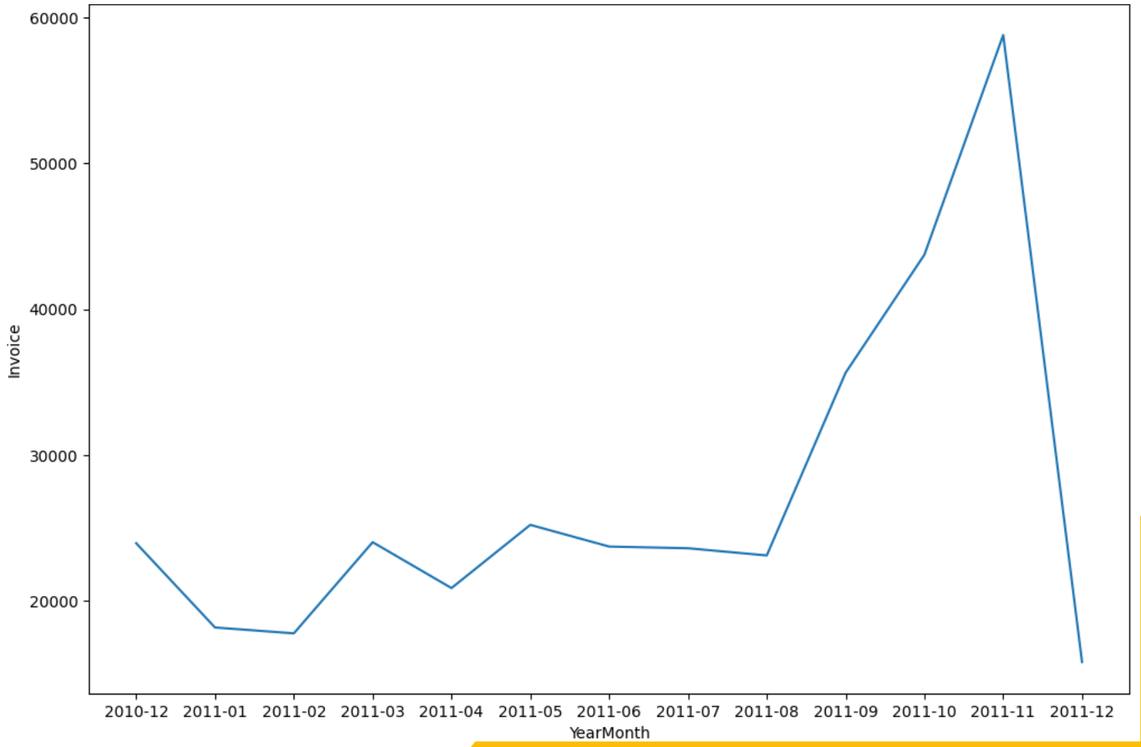
Fase 6

Aplicación de los modelos

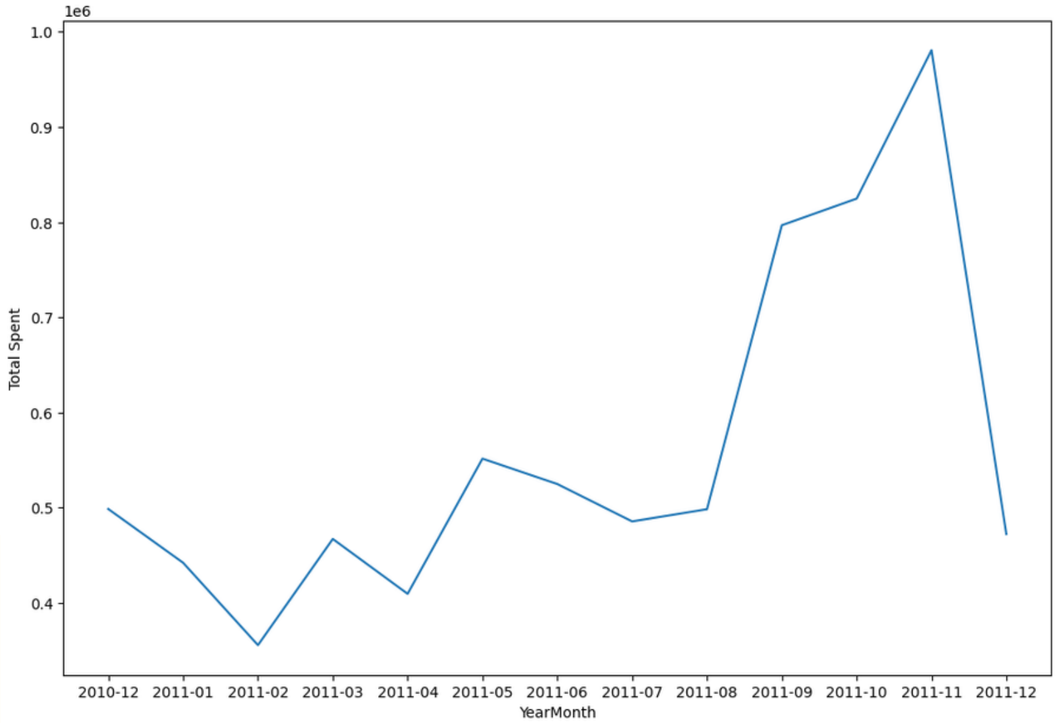
K-Means para la segmentación de los clientes.
Decision Tree, Random Forest y Logistic Regression para la predicción de las clases.

LAS VENTAS

COMPRAS AL MES



VENTAS AL MES



EL DATASET PRE RFM

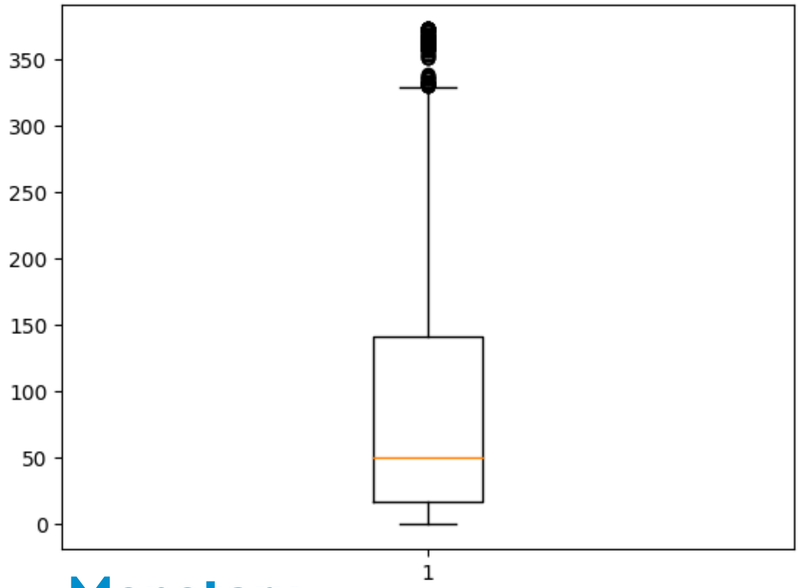
	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country	Total Spent	YearMonth
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom	15.30	2010-12
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34	2010-12
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom	22.00	2010-12
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34	2010-12
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	20.34	2010-12

	Quantity	Price	Customer ID	Total Spent
count	354345.000000	354345.000000	354345.000000	354345.000000
mean	12.048913	2.963793	15552.436219	20.625073
std	190.428127	17.862067	1594.546025	326.033014
min	1.000000	0.000000	12346.000000	0.000000
25%	2.000000	1.250000	14194.000000	4.160000
50%	4.000000	1.950000	15522.000000	10.200000
75%	12.000000	3.750000	16931.000000	17.700000
max	80995.000000	8142.750000	18287.000000	168469.600000

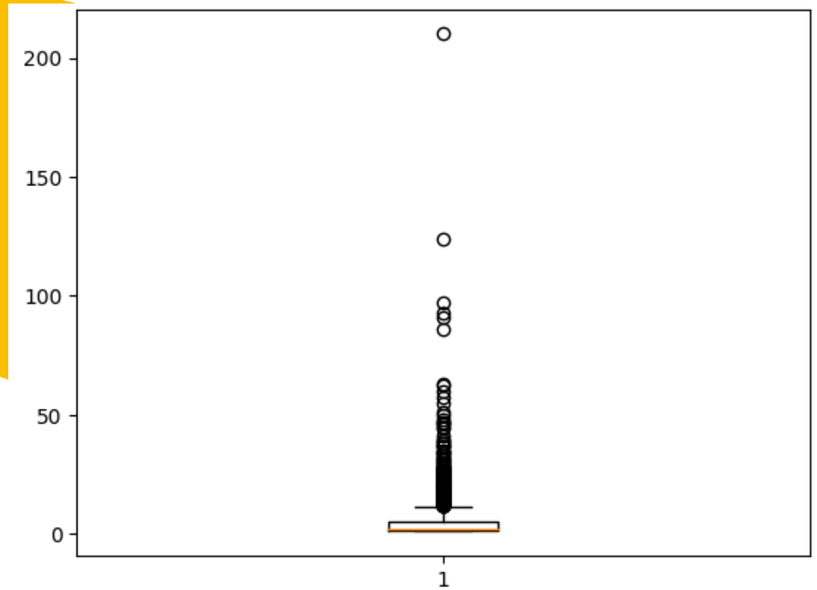
EL DATASET RFM

Customer ID	Recency	Frequency	Monetary
12346	325	1	77183.60
12747	2	11	4196.01
12748	0	210	33719.73
12749	3	5	4090.88
12820	3	4	942.34

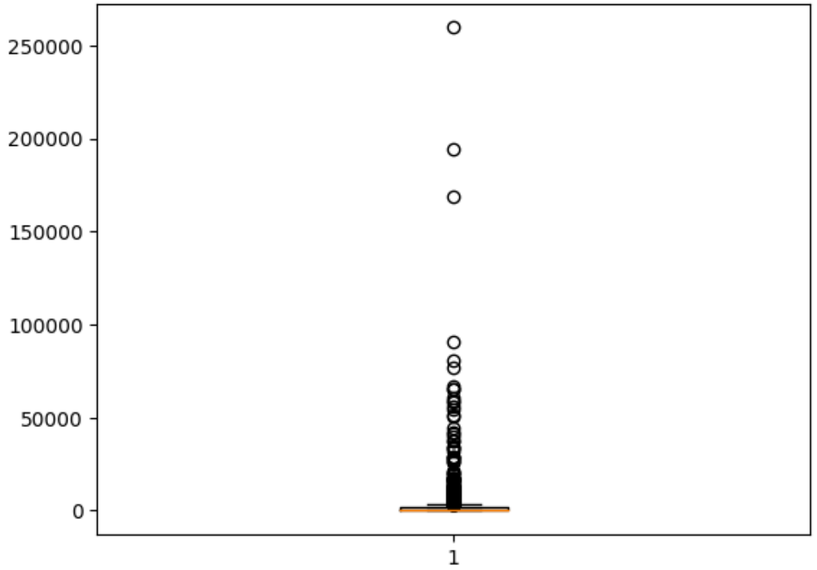
Recency

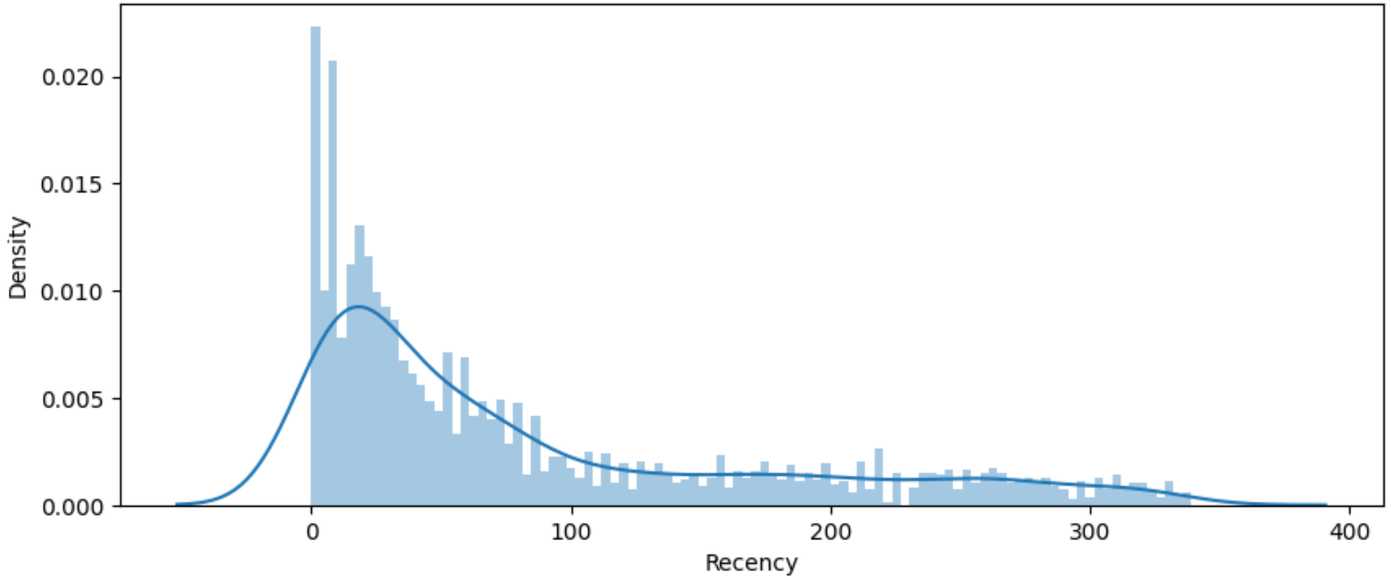


Frequency



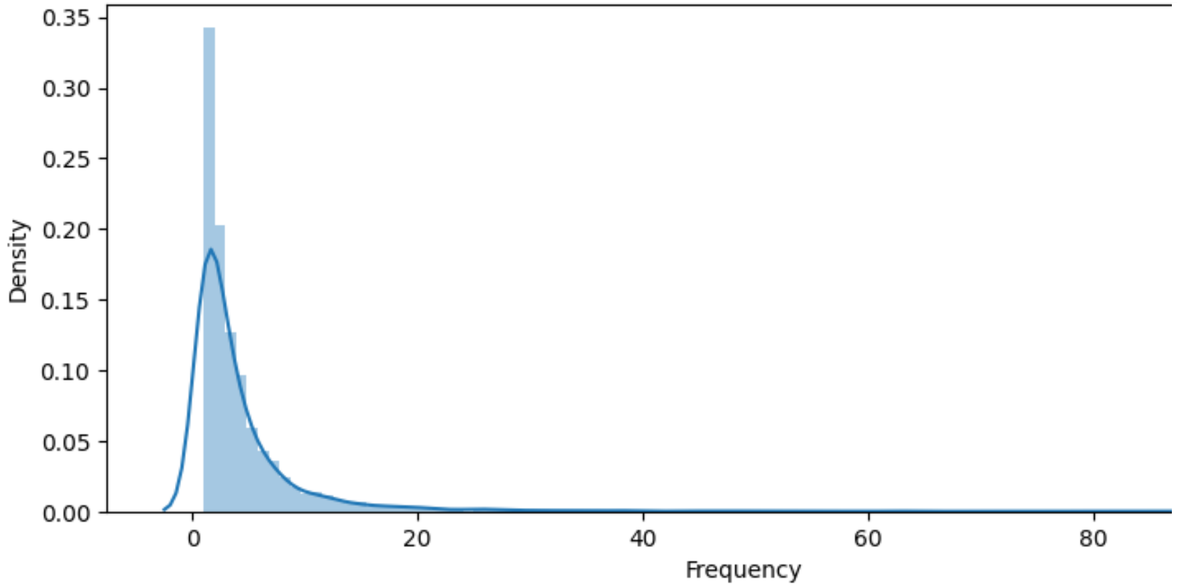
Monetary

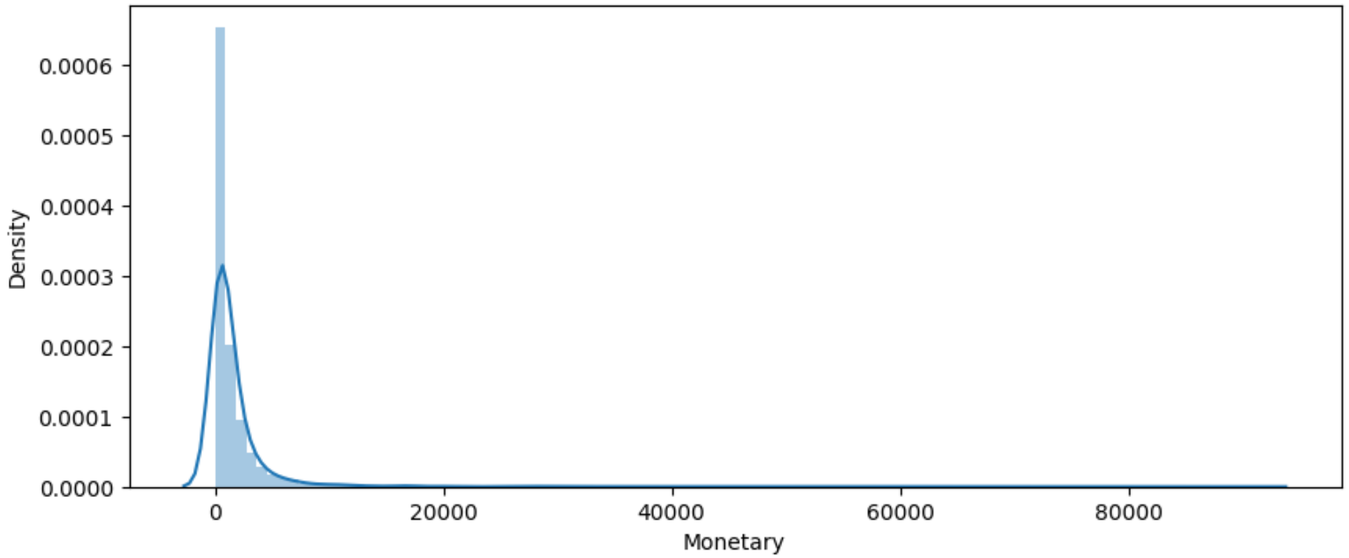




LOS OUTLIERS

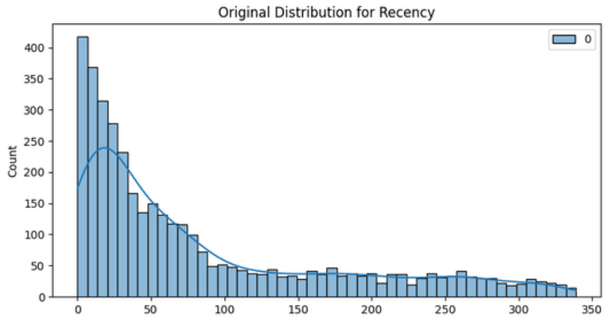
Como los outliers son una grandísima parte de nuestros datos, solo eliminamos en **Recency** aquellos que son mayores a **340**, en **Frequency** aquellos que son mayores a **100**, y en **Monetary**, aquellos que son mayores a **100.000**

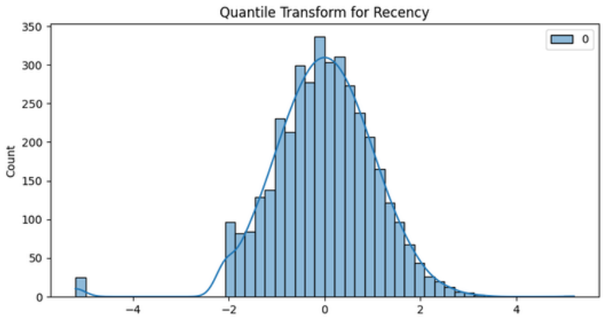


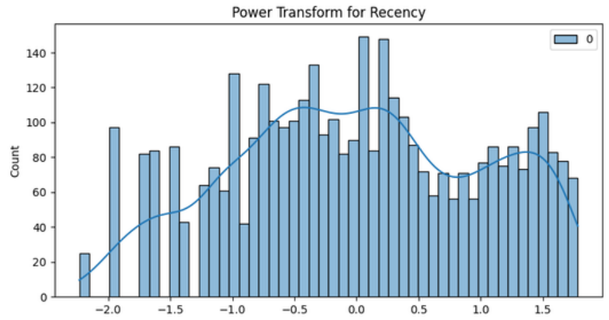


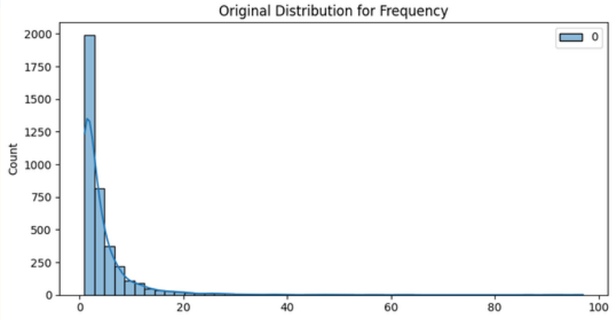
LA NORMALIZACIÓN

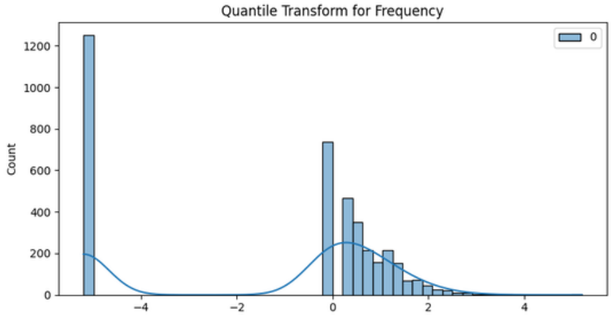
Requisito para el K-Means, ya que es sensible a las asimetrías en la distribución

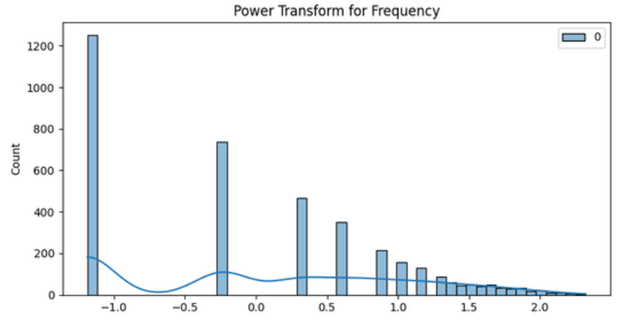


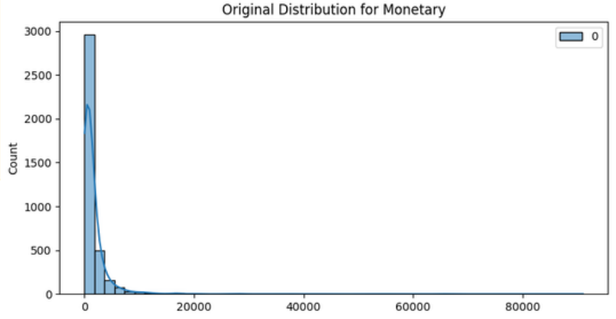


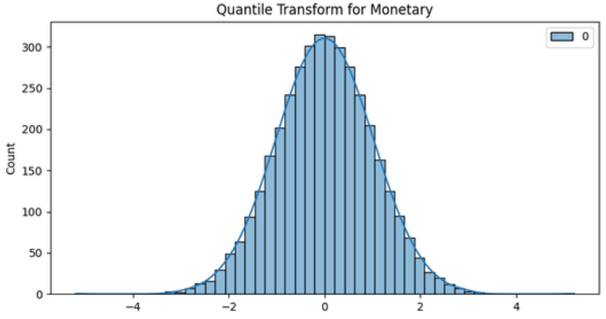


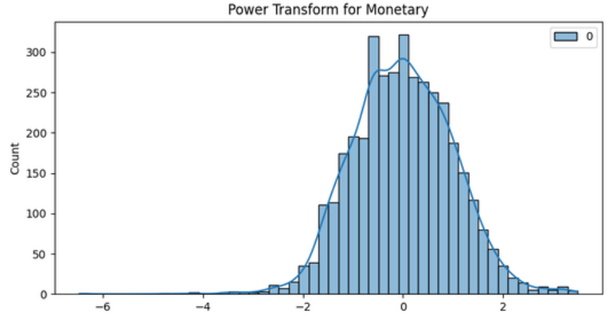












EL ESTANDARIZADO

El otro requisito para el K-Means, ya que convierte los datos a las mismas medidas

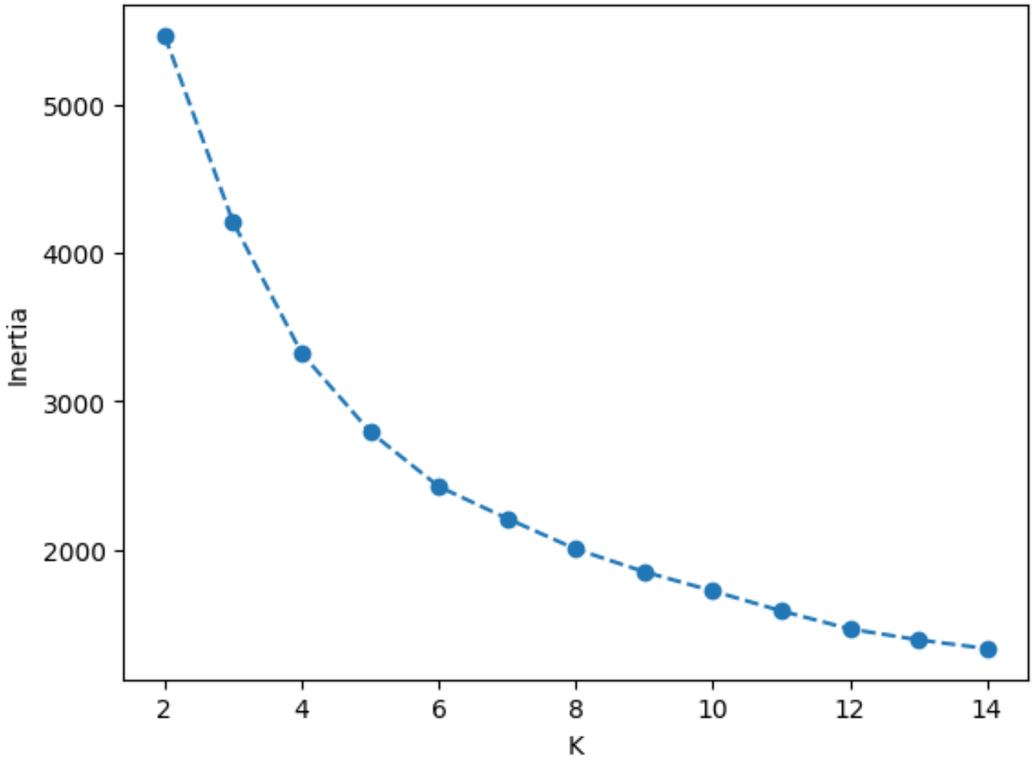
Customer ID	Recency	Frequency	Monetary
12346	1.735536	-1.186682	3.394483
12747	-1.739483	1.527384	1.411879
12749	-1.594947	0.863203	1.393125
12820	-1.594947	0.625468	0.261059
12821	1.306956	-1.186682	-1.723005

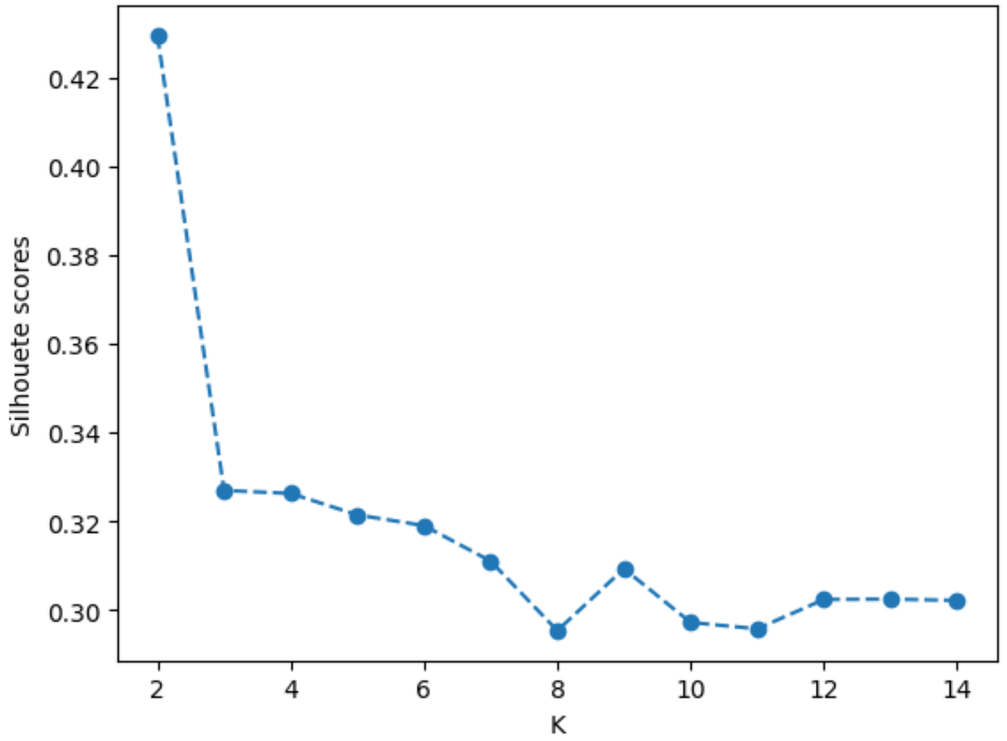
Con esto nuestros datos están listos para el K-means

EL MODELO - SELECCIÓN DE K

INERTIA VS. SILHOUETTE SCORE

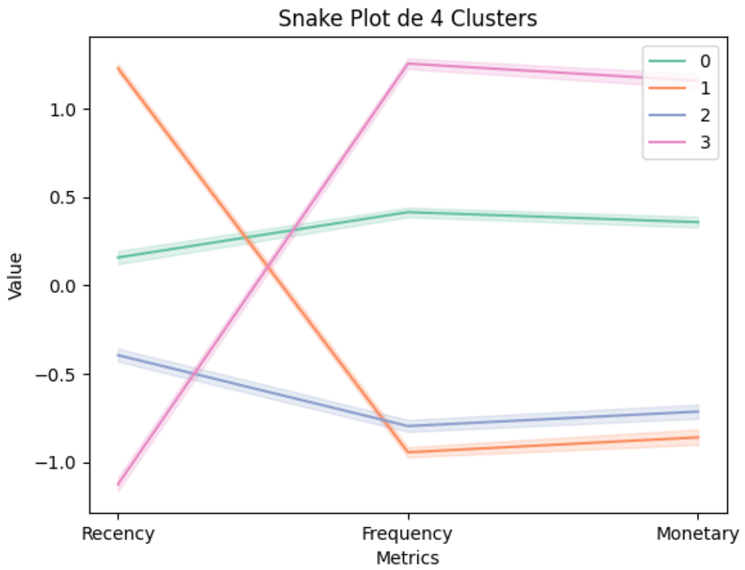
Método del codo

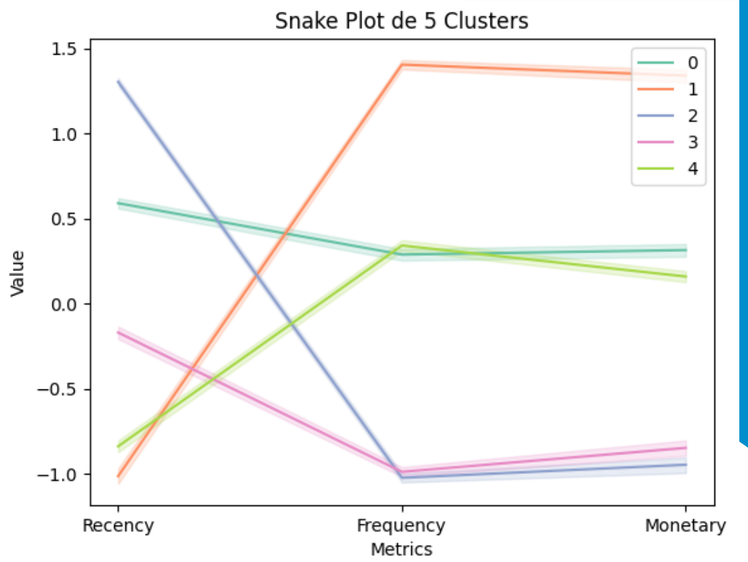


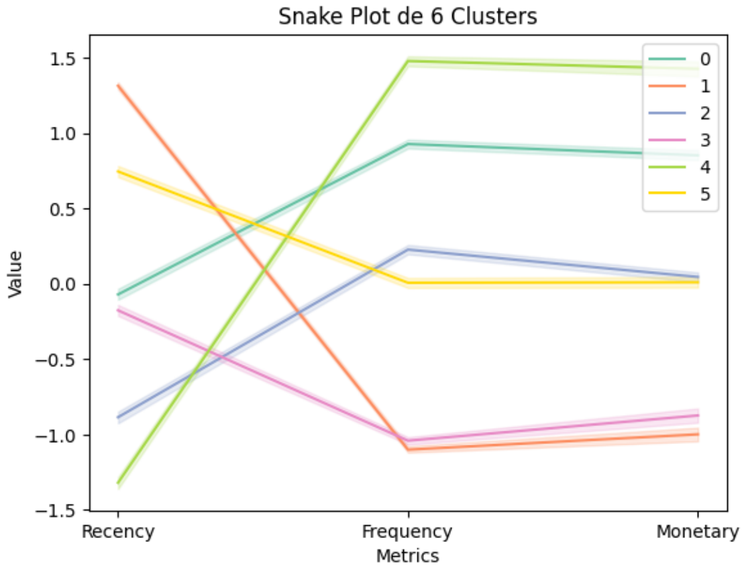


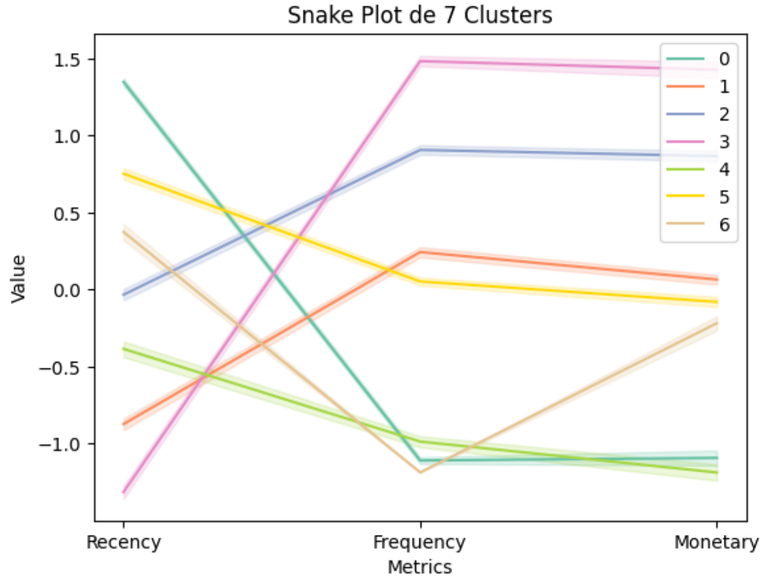
EL MODELO - SELECCIÓN DE K

COMPARACIÓN DE K EN VISUALIZACIÓN

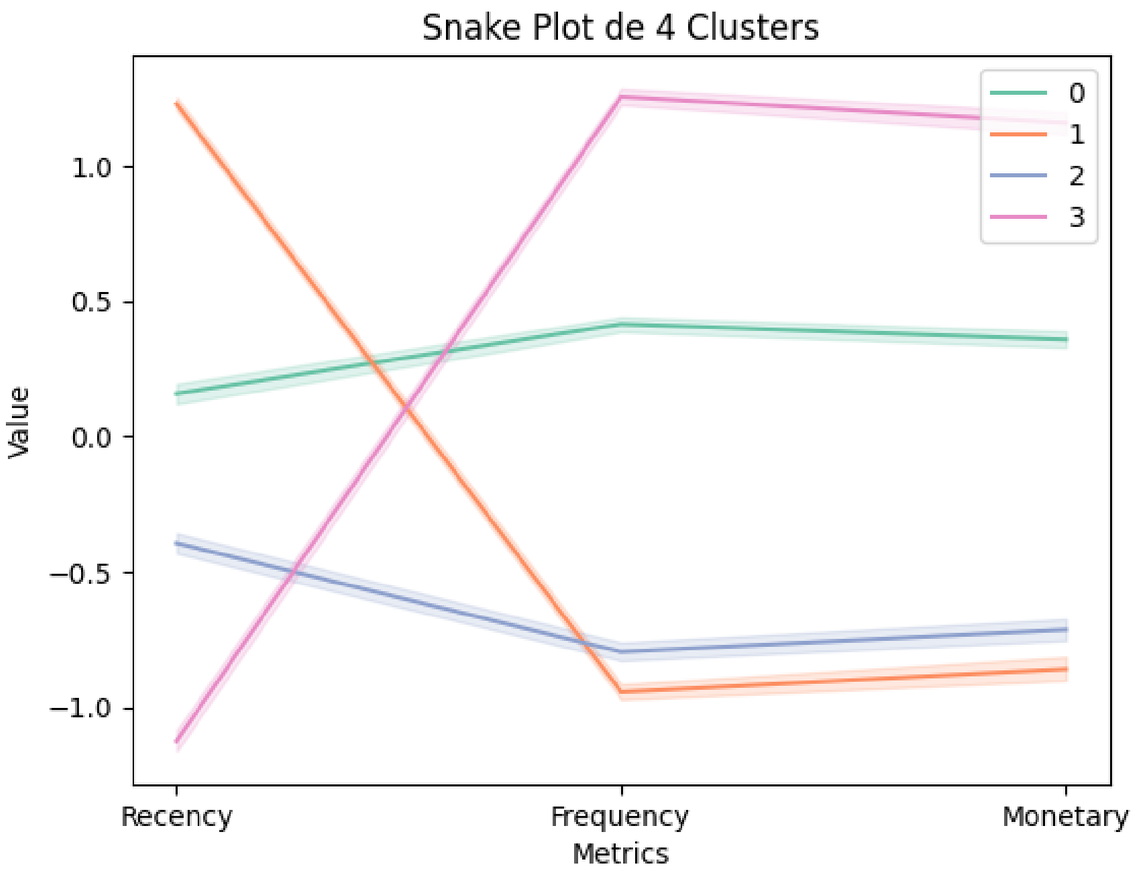








LOS 4 SEGMENTOS



CLUSTER 0 - SILVER

Son los que hace algun tiempo que no vienen, pero tienen una buena relacion frecuency-monetary.

CLUSTER 1 - BRASS

Son los peores clientes, aquellos que no vienen hace mucho, frecuentan poco y gastan poco.

CLUSTER 2 - BRONZE

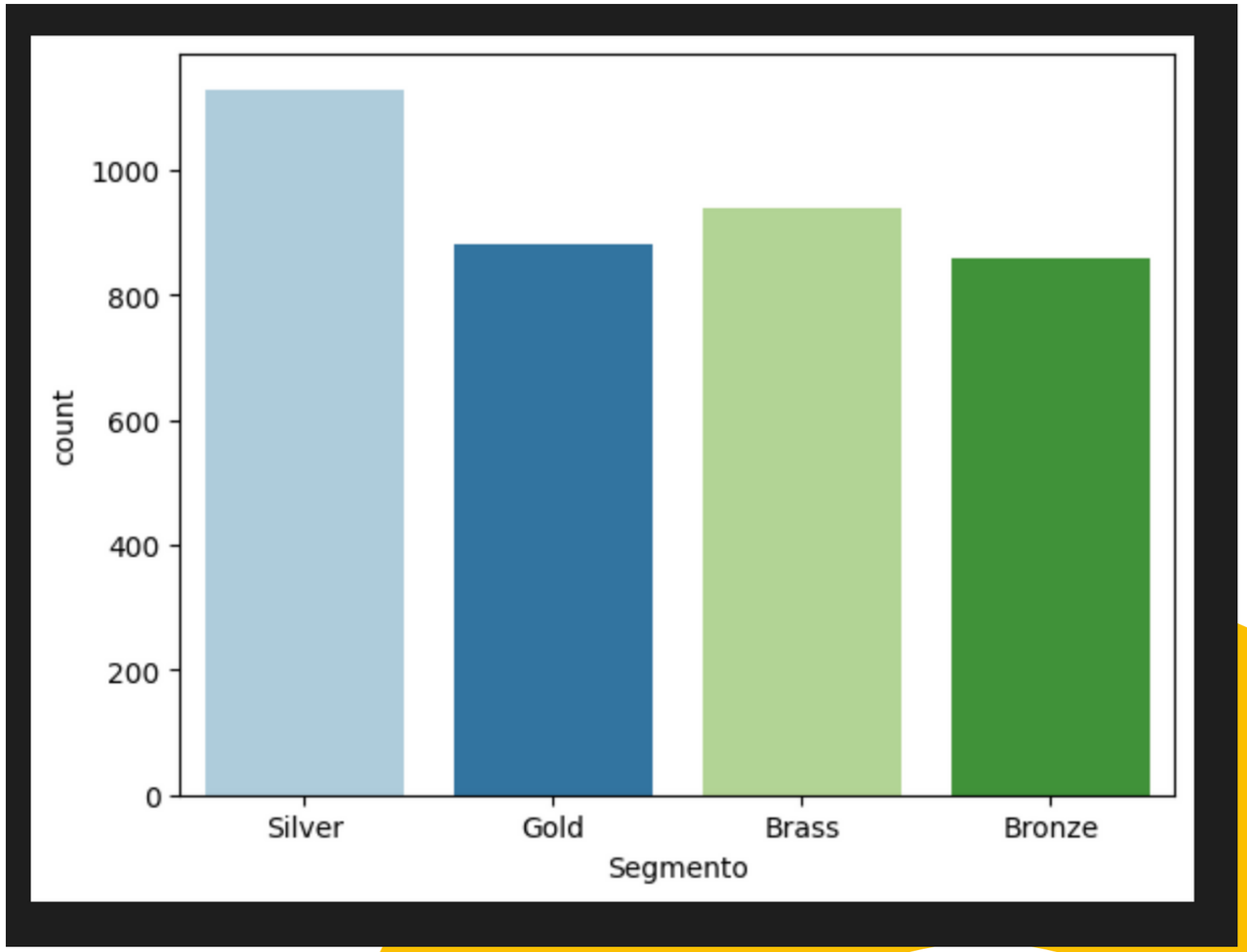
Son los que vinieron hace poco pero no frecuentan mucho ni gastan demasiado.

CLUSTER 3 - GOLD

Son los mejores clientes, pues hace muy poco que compraron, frecuentan mucho el negocio y son los que mas gastan.

CLASIFICACIÓN USANDO LAS ETIQUETAS DE K-MEANS

DISTRIBUCIÓN DE LAS 4 CLASES



	Recency	Frequency	Monetary	
	mean	mean	mean	count
Segmento				
Brass	209.0	1.0	326.0	940
Bronze	35.0	1.0	368.0	859
Gold	12.0	11.0	4958.0	881
Silver	74.0	4.0	1404.0	1129

```
Porcentaje de cada segmento en la muestra:

Silver    29.64
Brass     24.68
Gold      23.13
Bronze    22.55
Name: Segmento, dtype: float64
```

CLASIFICACIÓN USANDO LAS ETIQUETAS DE K-MEANS

Customer ID	Recency	Frequency	Monetary	Cluster_4
12346	1.735536	-1.186682	3.394483	0
12747	-1.739483	1.527384	1.411879	3
12749	-1.594947	0.863203	1.393125	3
12820	-1.594947	0.625468	0.261059	3
12821	1.306956	-1.186682	-1.723005	1

RESULTADOS DE GRIDCV

	Grid	Best score
2	Log_Regression	0.988183
1	Random_Forest	0.967834
0	Decision_Tree	0.959630

```
modelos_grid['Log_Regression'].best_params_
✓ 0.2s
{'C': 2.6, 'class_weight': 'balanced', 'penalty': 'l2'}
```

```
modelos_grid['Random_Forest'].best_params_
✓ 0.3s
{'class_weight': 'balanced',
 'max_depth': 9,
 'max_features': 'sqrt',
 'min_samples_leaf': 10,
 'n_estimators': 100}
```

RESULTADOS EN TEST

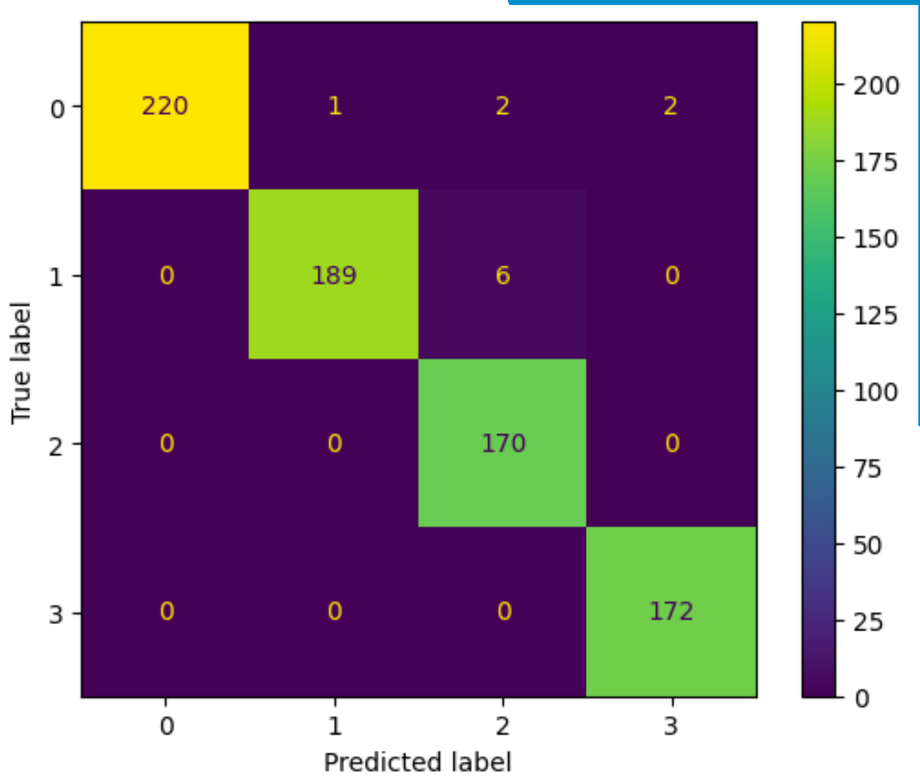
```
modelos_grid['Log_Regression'].best_estimator_.score(X_test, y_test)
✓ 0.3s
0.9855643044619422
```

```
modelos_grid['Random_Forest'].best_estimator_.score(X_test, y_test)
✓ 0.2s
0.9540682414698163
```

LOS RESULTADOS

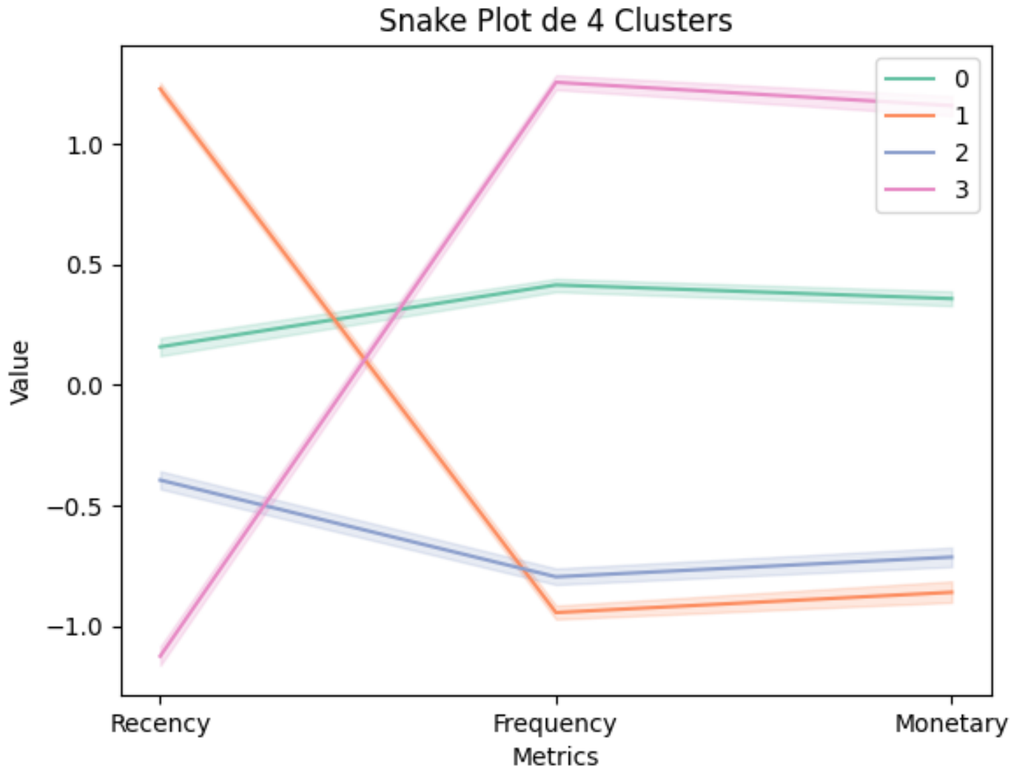
CLASSIFICATION REPORT Y CONFUSION MATRIX

	precision	recall	f1-score	support
0	1.00	0.98	0.99	225
1	0.99	0.97	0.98	195
2	0.96	1.00	0.98	170
3	0.99	1.00	0.99	172
accuracy			0.99	762
macro avg	0.98	0.99	0.99	762
weighted avg	0.99	0.99	0.99	762



¿DEMASIADO BONITO?

Con estos valores en las métricas de test podríamos pensar que hemos caído en overfitting, sin embargo, nuestros datos apuntan a que los clusters generados por K-Means estuvieran muy definidos. Al tratarse de un dataset de solo tres features es normal que nuestro modelo de regresión logística sea muy preciso.



A DESTACAR

1

LOS SEGMENTOS YA NOS PERMITEN DEFINIR ESTRATEGIAS DE MERCADO PARA LOS CLIENTES

2

LOS VALORES EN LAS MÉTRICAS 'DEMASIADO PERFECTAS' DE NUESTRO MODELO PUEDEN DEBERSE MÁS A LA CLASIFICACIÓN INICIAL QUE A OVERFITTING

GRACIAS

ROSARIO MONTALBAN SARDI

DATA SCIENCE BOOTCAMP
THE BRIDGE TECH