Customer Segmentation

With K-means clustering based on RFM metrics (Recency, Frequency and MonetaryValue)

Diego Beteta

Customer Segmentation

Contenido

1. Retención de clientes (%)
2. Segmentación RFM (Recency, Frequency and MonetaryValue)
3. Segmentación K-means Clustering
4. Análisis de K-means Clustering
5. Conclusiones

1. Retención de clientes (%)

Importar, limpiar, filtrar y organizar datos

dataset.csv

```
,InvoiceNo,StockCode,Description,Quantity,InvoiceDate,UnitPrice,CustomerID,Country
416792,572558,22745,POPPY'S PLAYHOUSE BEDROOM ,6,2011-10-25 08:26:00,2.1,14286,United Kingdom
482904,577485,23196,VINTAGE LEAF MAGNETIC NOTEPAD,1,2011-11-20 11:56:00,1.45,16360,United Kingdom
263743,560034,23299,FOOD COVER WITH BEADS SET 2 ,6,2011-07-14 13:35:00,3.75,13933,United Kingdom
495549,578307,72349B,SET/6 PURPLE BUTTERFLY T-LIGHTS,1,2011-11-23 15:53:00,2.1,17290,United Kingdom
204384,554656,21756,BATH BUILDING BLOCK WORD,3,2011-05-25 13:36:00,5.95,17663,United Kingdom
125615,547051,22028,PENNY FARTHING BIRTHDAY CARD,12,2011-03-20 12:06:00,0.42,12902,United Kingdom
483123,577493,20724,RED RETROSPOT CHARLOTTE BAG,10,2011-11-20 12:13:00,0.85,17323,United Kingdom
449888,575143,23343,JUMBO BAG VINTAGE CHRISTMAS ,10,2011-11-08 15:37:00,2.08,13643,United Kingdom
127438,547223,22934,BAKING MOULD EASTER EGG WHITE CHOC,2,2011-03-21 15:10:00,2.95,12867,United Kingdom
```

	Unnamed: 0	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	416792	572558	22745	POPPY'S PLAYHOUSE BEDROOM	6	2011-10-25 08:26:00	2.10	14286	United Kingdom
1	482904	577485	23196	VINTAGE LEAF MAGNETIC NOTEPAD	1	2011-11-20 11:56:00	1.45	16360	United Kingdom
2	263743	560034	23299	FOOD COVER WITH BEADS SET 2	6	2011-07-14 13:35:00	3.75	13933	United Kingdom
3	495549	578307	72349B	SET/6 PURPLE BUTTERFLY T-LIGHTS	1	2011-11-23 15:53:00	2.10	17290	United Kingdom
4	204384	554656	21756	BATH BUILDING BLOCK WORD	3	2011-05-25 13:36:00	5.95	17663	United Kingdom

	Unnamed: 0	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	InvoiceMonth	CohortMonth
0	416792	572558	22745	POPPY'S PLAYHOUSE BEDROOM	6	2011-10-25 08:26:00	2.10	14286	United Kingdom	2011-10-01	2011-04-01
1	482904	577485	23196	VINTAGE LEAF MAGNETIC NOTEPAD	1	2011-11-20 11:56:00	1.45	16360	United Kingdom	2011-11-01	2011-09-01
2	263743	560034	23299	FOOD COVER WITH BEADS SET 2	6	2011-07-14 13:35:00	3.75	13933	United Kingdom	2011-07-01	2011-07-01
3	495549	578307	72349B	SET/6 PURPLE BUTTERFLY T-LIGHTS	1	2011-11-23 15:53:00	2.10	17290	United Kingdom	2011-11-01	2011-11-01
4	204384	554656	21756	BATH BUILDING BLOCK WORD	3	2011-05-25 13:36:00	5.95	17663	United Kingdom	2011-05-01	2011-02-01

	Unnamed: 0	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	InvoiceMonth	CohortMonth	CohortIndex
0	416792	572558	22745	POPPY'S PLAYHOUSE BEDROOM	6	2011-10-25 08:26:00	2.10	14286	United Kingdom	2011-10-01	2011-04-01	7
1	482904	577485	23196	VINTAGE LEAF MAGNETIC NOTEPAD	1	2011-11-20 11:56:00	1.45	16360	United Kingdom	2011-11-01	2011-09-01	3
2	263743	560034	23299	FOOD COVER WITH BEADS SET 2	6	2011-07-14 13:35:00	3.75	13933	United Kingdom	2011-07-01	2011-07-01	1
3	495549	578307	72349B	SET/6 PURPLE BUTTERFLY T-LIGHTS	1	2011-11-23 15:53:00	2.10	17290	United Kingdom	2011-11-01	2011-11-01	1
4	204384	554656	21756	BATH BUILDING BLOCK WORD	3	2011-05-25 13:36:00	5.95	17663	United Kingdom	2011-05-01	2011-02-01	4

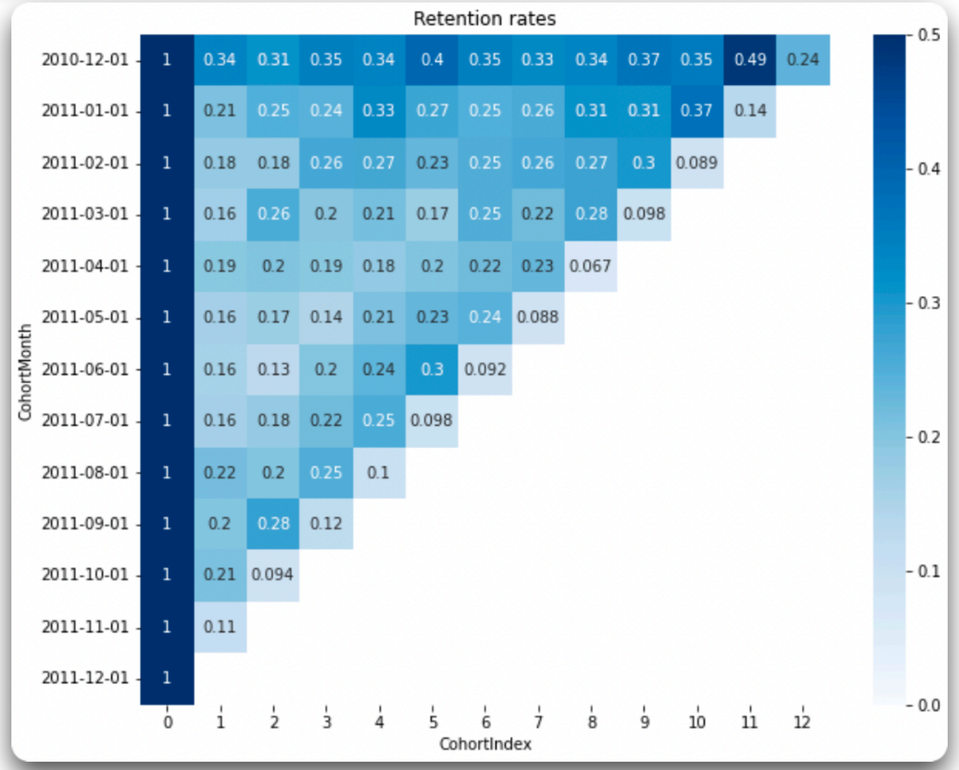
1. Retención de clientes (%)

Transformar datos en tablas pivotes

	Unnamed: 0	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	InvoiceMonth	CohortMonth	CohortIndex
0	416792	572558	22745	POPPY'S PLAYHOUSE BEDROOM	6	2011-10-25 08:26:00	2.10	14286	United Kingdom	2011-10-01	2011-04-01	7
1	482904	577485	23196	VINTAGE LEAF MAGNETIC NOTEPAD	1	2011-11-20 11:56:00	1.45	16360	United Kingdom	2011-11-01	2011-09-01	3
2	263743	560034	23299	FOOD COVER WITH BEADS SET 2	6	2011-07-14 13:35:00	3.75	13933	United Kingdom	2011-07-01	2011-07-01	1
3	495549	578307	72349B	SET/6 PURPLE BUTTERFLY T-LIGHTS	1	2011-11-23 15:53:00	2.10	17290	United Kingdom	2011-11-01	2011-11-01	1
4	204384	554656	21756	BATH BUILDING BLOCK WORD	3	2011-05-25 13:36:00	5.95	17663	United Kingdom	2011-05-01	2011-02-01	4

CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
CohortMonth													
2010-12-01	716.0	246.0	221.0	251.0	245.0	285.0	249.0	236.0	240.0	265.0	254.0	348.0	172.0
2011-01-01	332.0	69.0	82.0	81.0	110.0	90.0	82.0	86.0	104.0	102.0	124.0	45.0	NaN
2011-02-01	316.0	58.0	57.0	83.0	85.0	74.0	80.0	83.0	86.0	95.0	28.0	NaN	NaN
2011-03-01	388.0	63.0	100.0	76.0	83.0	67.0	98.0	85.0	107.0	38.0	NaN	NaN	NaN
2011-04-01	255.0	49.0	52.0	49.0	47.0	52.0	56.0	59.0	17.0	NaN	NaN	NaN	NaN
2011-05-01	249.0	40.0	43.0	36.0	52.0	58.0	61.0	22.0	NaN	NaN	NaN	NaN	NaN
2011-06-01	207.0	33.0	26.0	41.0	49.0	62.0	19.0	NaN	NaN	NaN	NaN	NaN	NaN
2011-07-01	173.0	28.0	31.0	38.0	44.0	17.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08-01	139.0	30.0	28.0	35.0	14.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09-01	279.0	56.0	78.0	34.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10-01	318.0	67.0	30.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-11-01	291.0	32.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-12-01	38.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

CohortIndex	1	2	3	4	5	6	7	8	9	10	11	12	13
CohortMonth													
2010-12-01	100.0	34.4	30.9	35.1	34.2	39.8	34.8	33.0	33.5	37.0	35.5	48.6	24.0
2011-01-01	100.0	20.8	24.7	24.4	33.1	27.1	24.7	25.9	31.3	30.7	37.3	13.6	NaN
2011-02-01	100.0	18.4	18.0	26.3	26.9	23.4	25.3	26.3	27.2	30.1	8.9	NaN	NaN
2011-03-01	100.0	16.2	25.8	19.6	21.4	17.3	25.3	21.9	27.6	9.8	NaN	NaN	NaN
2011-04-01	100.0	19.2	20.4	19.2	18.4	20.4	22.0	23.1	6.7	NaN	NaN	NaN	NaN
2011-05-01	100.0	16.1	17.3	14.5	20.9	23.3	24.5	8.8	NaN	NaN	NaN	NaN	NaN
2011-06-01	100.0	15.9	12.6	19.8	23.7	30.0	9.2	NaN	NaN	NaN	NaN	NaN	NaN
2011-07-01	100.0	16.2	17.9	22.0	25.4	9.8	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-08-01	100.0	21.6	20.1	25.2	10.1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-09-01	100.0	20.1	28.0	12.2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-10-01	100.0	21.1	9.4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-11-01	100.0	11.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2011-12-01	100.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN



Interpretación de HeatMap:

"Sólo el 22% de los clientes que compraron (por primera vez) nuestros productos en julio 2011, volvieron a comprar 03 meses después".

"El 49% de los clientes que compraron (por primera vez) nuestros productos en diciembre 2010, volvieron a comprar 11 meses después. Convirtiendo a noviembre 2011 como el mes con mayor % de retención de clientes".

2. Segmentación RFM

Definición de métricas (Recency, Frequency and MonetaryValue)

```
Min:2010-12-10; Max:2011-12-09
```

	Unnamed: 0	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	416792	572558	22745	POPPY'S PLAYHOUSE BEDROOM	6	2011-10-25	2.10	14286	United Kingdom
1	482904	577485	23196	VINTAGE LEAF MAGNETIC NOTEPAD	1	2011-11-20	1.45	16360	United Kingdom
2	263743	560034	23299	FOOD COVER WITH BEADS SET 2	6	2011-07-14	3.75	13933	United Kingdom
3	495549	578307	72349B	SET/6 PURPLE BUTTERFLY T-LIGHTS	1	2011-11-23	2.10	17290	United Kingdom
4	204384	554656	21756	BATH BUILDING BLOCK WORD	3	2011-05-25	5.95	17663	United Kingdom

	Unnamed: 0	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	TotalSum
0	416792	572558	22745	POPPY'S PLAYHOUSE BEDROOM	6	2011-10-25	2.10	14286	United Kingdom	12.60
1	482904	577485	23196	VINTAGE LEAF MAGNETIC NOTEPAD	1	2011-11-20	1.45	16360	United Kingdom	1.45
2	263743	560034	23299	FOOD COVER WITH BEADS SET 2	6	2011-07-14	3.75	13933	United Kingdom	22.50
3	495549	578307	72349B	SET/6 PURPLE BUTTERFLY T-LIGHTS	1	2011-11-23	2.10	17290	United Kingdom	2.10
4	204384	554656	21756	BATH BUILDING BLOCK WORD	3	2011-05-25	5.95	17663	United Kingdom	17.85

CustomerID	Recency	Frequency	MonetaryValue
12747	3	25	948.70
12748	1	888	7046.16
12749	4	37	813.45
12820	4	17	268.02
12822	71	9	146.15

Interpretación RFM

- **Recency:**
Mide la cantidad de días que han pasado desde la última compra que hizo el cliente durante los últimos 12 meses.

- **Frequency:**
Mide la cantidad acumulada de las veces que el cliente compró durante los últimos 12 meses.

- **MonetaryValue:**
Mide la cantidad acumulada de dinero que el cliente ha gasto en nuestros productos en los últimos 12 meses.

El análisis podría extenderse a 24 o 36 meses, sin embargo, por razones prácticas del proyecto y por darle mayor importancia al comportamiento más recientes de los clientes filtré el dataset de los últimos 12 meses.

2. Segmentación RFM

Análisis de métricas (Recency, Frequency and MonetaryValue)

CustomerID	Recency	Frequency	MonetaryValue	R	F	M	RFM_Segment	RFM_Score	General_Segment
12747	3	25	948.70	4	4	4	444	12	Gold
12748	1	888	7046.16	4	4	4	444	12	Gold
12749	4	37	813.45	4	4	4	444	12	Gold
12820	4	17	268.02	4	3	3	433	10	Gold
12822	71	9	146.15	2	2	3	223	7	Silver

	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
RFM_Segment				
111	246.9	2.1	28.4	345
112	234.5	2.9	82.4	105
113	254.1	2.3	202.6	42
114	225.9	2.2	1434.6	16
121	246.5	6.5	38.4	63
...	...	...	...	...
433	9.2	14.5	229.2	113
434	10.5	16.7	776.4	71
442	9.4	27.1	101.3	18
443	10.3	38.6	231.1	67
444	8.0	75.6	1653.9	372

62 rows × 4 columns

	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
RFM_Score				
3	246.9	2.1	28.4	345
4	162.2	3.1	47.8	337
5	138.9	4.3	78.2	393
6	101.0	6.3	146.3	444
7	78.0	8.5	160.2	382
8	62.6	12.8	196.3	376
9	46.8	16.7	330.3	345
10	31.9	24.0	443.1	355
11	21.8	38.9	705.3	294
12	8.0	75.6	1653.9	372

	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
General_Segment				
Bronze	180.8	3.2	52.7	1075
Gold	20.3	47.1	959.7	1021
Silver	73.9	10.7	202.9	1547

Análisis RFM

Es posible segmentar los clientes de tres maneras:

- **RFM Segment:**
Es la concatenación de las columnas RFM. Dependerá de la cantidad grupos de percentiles de igual tamaño que asignes inicialmente (en este caso 04). La ventaja es que se tiene una gran variedad de combinaciones para una segmentación más específica (exploración de nichos de mercado).

- **RFM Score:**
Es la suma de las columnas RFM. Permite tener un panorama más generalizado en cuanto a la segmentación de clientes.

- **General Segment:**
Son etiquetas comerciales personalizadas en función a la agrupación de valores de RFM_Score.

3. Segmentación K-means Clustering

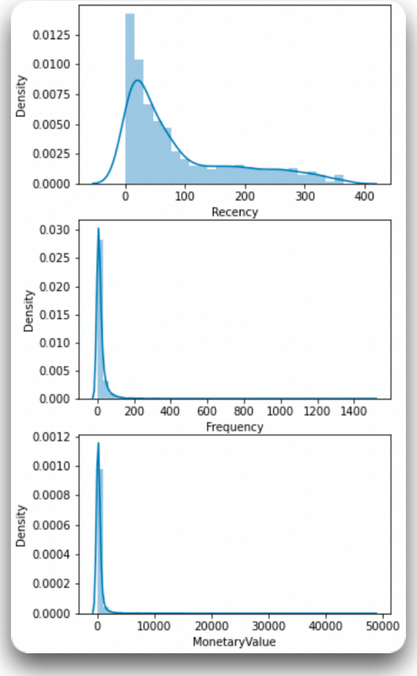
Transformación logarítmica y estandarización de variables RFM

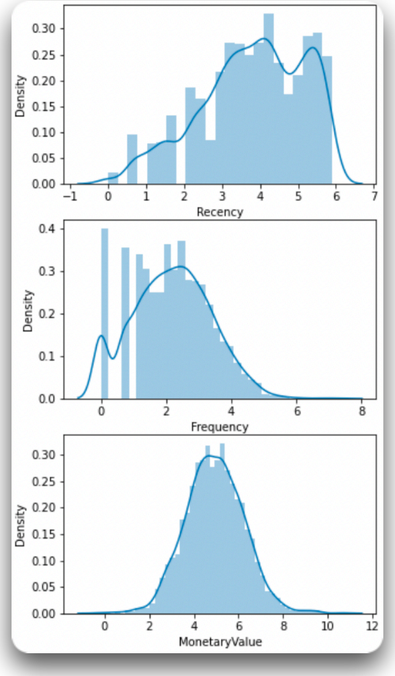
CustomerID	Recency	Frequency	MonetaryValue
12747	3	25	948.70
12748	1	888	7046.16
12749	4	37	813.45
12820	4	17	268.02
12822	71	9	146.15

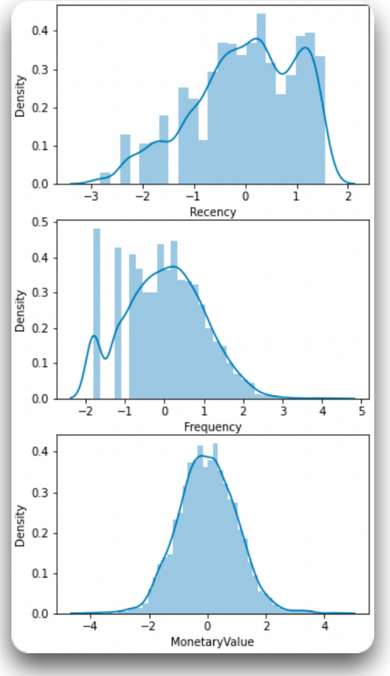
	Recency	Frequency	MonetaryValue
count	3643.00000	3643.000000	3643.000000
mean	90.43563	18.714247	370.694387
std	94.44651	43.754468	1347.443451
min	1.00000	1.000000	0.650000
25%	19.00000	4.000000	58.705000
50%	51.00000	9.000000	136.370000
75%	139.00000	21.000000	334.350000
max	365.00000	1497.000000	48060.350000

	Recency	Frequency	MonetaryValue
count	3643.000000	3643.000000	3643.000000
mean	3.806481	2.171902	4.934900
std	1.352631	1.210321	1.310945
min	0.000000	0.000000	-0.430783
25%	2.944439	1.386294	4.072524
50%	3.931826	2.197225	4.915372
75%	4.934474	3.044522	5.812188
max	5.899897	7.311218	10.780213

	Recency	Frequency	MonetaryValue
count	3643.00	3643.00	3643.00
mean	-0.00	0.00	0.00
std	1.00	1.00	1.00
min	-2.81	-1.79	-4.09
25%	-0.64	-0.65	-0.66
50%	0.09	0.02	-0.01
75%	0.83	0.72	0.67
max	1.55	4.25	4.46







Nota importante:

La desventaja de estos métodos anteriores es que no hay manera de saber si una cantidad de segmentos es mejor que otra y está sujeta a la subjetividad del equipo.

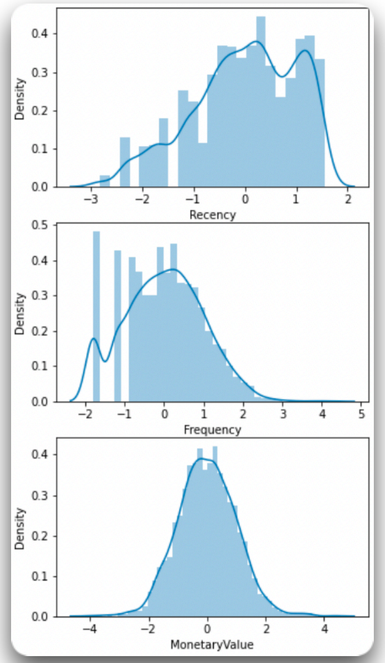
K-Means busca un balance entre tener una adecuada segmentación comercial de clientes y reducir en lo posible el error de pronóstico (SSE) mediante “The elbow method”.

K-means trabaja mucho mejor siempre y cuando la distribución de las variables estén estandarizadas, es decir, media 0 y desviación estándar 1.

3. Segmentación K-means Clustering

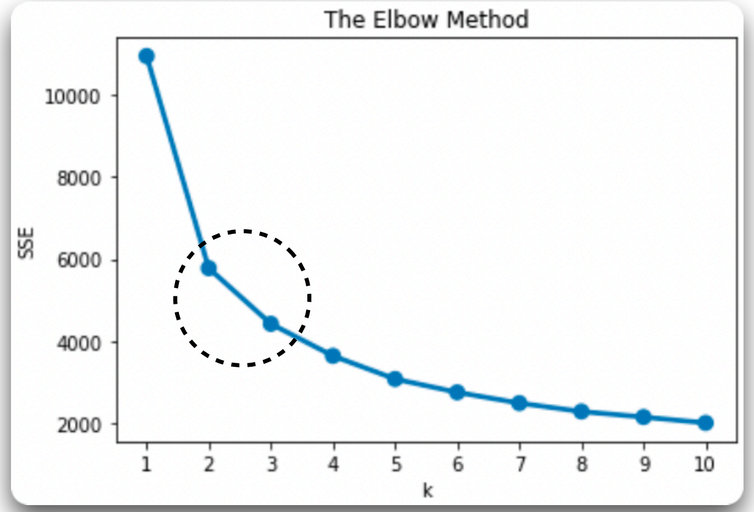
Identificación de valor óptimo de K

	Recency	Frequency	MonetaryValue
count	3643.00	3643.00	3643.00
mean	-0.00	0.00	0.00
std	1.00	1.00	1.00
min	-2.81	-1.79	-4.09
25%	-0.64	-0.65	-0.66
50%	0.09	0.02	-0.01
75%	0.83	0.72	0.67
max	1.55	4.25	4.46




```python
# Importar librerías
from sklearn.cluster import KMeans
import seaborn as sns
from matplotlib import pyplot as plt

# Entrenar el KMeans y calcular el SSE para cada *k*
sse = {}
for k in range(1, 11):
    kmeans = KMeans(n_clusters=k, random_state=1)
    kmeans.fit(datamart_rfm_normalized)
    sse[k] = kmeans.inertia_ # sum of squared distances to closest cluster center

# Graficar SSE para cada *k*
plt.title('The Elbow Method')
plt.xlabel('k')
plt.ylabel('SSE')
sns.pointplot(x=list(sse.keys()), y=list(sse.values()))
plt.show()
```

✓ 3.3s




```python
# El gráfico The Elbow Method nos sugiere que
# los números óptimos de clustering son 2 y 3.

# En esta ocasión, escogeré k-means=3 para
# una mejor flexibilidad en las estrategias comerciales
# y para incrementar la información en el resumen estadístico.
from sklearn.cluster import KMeans
kmeans = KMeans(n_clusters=3, random_state=1)

# Calcular el k-means clustering sobre la data pre-procesada
kmeans.fit(datamart_rfm_normalized)

# Extraer etiquetas de cluster desde el atributo labels_
cluster_labels = kmeans.labels_
```

✓ 0.2s

```python
# Crear una columna de etiquetas de clúster en datamart_rfm
datamart_rfm_k3 = datamart_rfm.assign(Cluster = cluster_labels)

# Calcular el promedio de los valores RFM
# y la cantidad de clientes por cada cluster
datamart_rfm_k3.groupby(['Cluster']).agg({
    'Recency': 'mean',
    'Frequency': 'mean',
    'MonetaryValue': ['mean', 'count']
}).round(0)
```

✓ 0.7s

	Recency	Frequency	MonetaryValue
	mean	mean	mean	count
Cluster				
0	16.0	50.0	1051.0	901
1	167.0	3.0	53.0	1156
2	77.0	12.0	216.0	1586

4. Análisis de K-means Clustering

Con Snake Plot e Importance Relevance HeatMap

```python
# Preparar la data para el snake plot
# Transformar datamart_normalized como DataFrame y agregar una columna 'Cluster'
datamart_rfm_normalized = pd.DataFrame(datamart_rfm_normalized,
                                       index=datamart_rfm.index,
                                       columns=datamart_rfm.columns)
datamart_rfm_normalized['Cluster'] = datamart_rfm_k3['Cluster']

# Moldear los datos en un formato largo para que los valores RFM
#  y los nombres de las métricas se almacenan en una sóla columna
datamart_melt = pd.melt(datamart_rfm_normalized.reset_index(),
                        # Asignar CustomerID y Cluster como variables ID
                        id_vars=['CustomerID', 'Cluster'],
                        # Asignar RFM_Values como value_vars
                        value_vars=['Recency', 'Frequency', 'MonetaryValue'],
                        var_name='Attribute',
                        value_name='Value')
datamart_melt
```

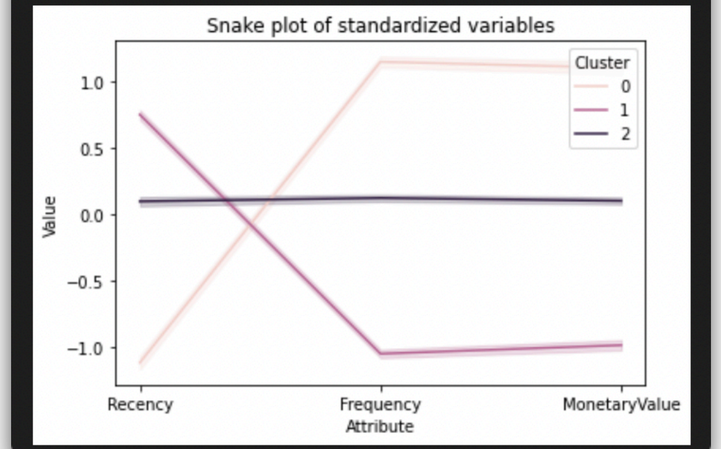
✓ 0.7s

	CustomerID	Cluster	Attribute	Value
0	12747	0	Recency	-2.002202
1	12748	0	Recency	-2.814518
2	12749	0	Recency	-1.789490
3	12820	0	Recency	-1.789490
4	12822	2	Recency	0.337315
...	...	...	...	...
10924	18280	1	MonetaryValue	-0.975812
10925	18281	1	MonetaryValue	-1.125628
10926	18282	1	MonetaryValue	-1.152485
10927	18283	0	MonetaryValue	0.866422
10928	18287	2	MonetaryValue	0.797937

10929 rows × 4 columns

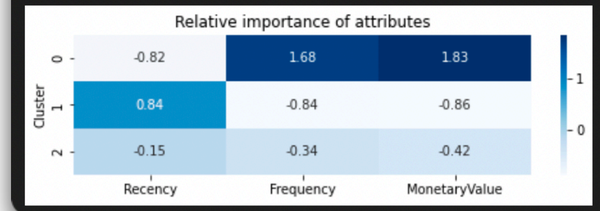
```python
# Graficar el snake plot
plt.title('Snake plot of standardized variables')
sns.lineplot(x='Attribute', y='Value', hue='Cluster',
             data=datamart_melt)

# Presentar gráfico
plt.show()
```

✓ 0.2s




```python
# Calcular los valores promedio de RFM para cada Cluster de clientes
cluster_avg = datamart_rfm_k3.groupby(['Cluster']).mean()

# Calculate los valores promedio de RFM para la población total de clientes
population_avg = datamart_rfm.mean()

# Calcular la importancia relativa del valor del atributo del cluster
# en comparación con la población
relative_imp = cluster_avg / population_avg - 1

# Imprimir puntajes de importancia relativa redondeados a 2 decimales
relative_imp.round(2)

# Graficar el heatmap
plt.figure(figsize=(8, 2))
plt.title('Relative importance of attributes')
sns.heatmap(data=relative_imp, annot=True, fmt='.2f', cmap='Blues')

# Presentar el gráfico
plt.show()
```

✓ 0.1s



5. Conclusiones

Puntos claves

- La segmentación RFM permite a las empresas identificar tendencias sobre el comportamiento de sus clientes: los más recientes, los más frecuentes y los que más gastan en nuestros productos.
- El objetivo es identificar a los clientes que gastan más y adquieren nuestros productos más a menudo. Crear estrategias de fidelización para asegurar nuestras ventas y, por ende, mayor rotación de inventarios.
- K-means es una solución frente a la subjetividad que está sujeta el RFM_Segment y RFM_Score, buscando un equilibrio entre disminuir la suma de errores al cuadrado de su clasificación y una cantidad interpretable para estrategias comerciales.

Ideas

- La segmentación no debe limitarse sólo a clientes sino también a productos con el objetivo de crear estrategias comerciales:
  - Aumentar la rotación de inventarios
  - Exploración de nichos de mercado
  - Incrementar capital de trabajo en productos más comerciales
  - Ofrecer al mercado productos estacionales mediante el análisis del ‘Relative Importance of Attributes’ plot para identificar los meses más comerciales para una cierta cantidad de productos.