Customer Segmentation

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

Diego Beteta

Customer Segmentation

Content

1. Customer retention (%)
2. RFM (Recency, Frequency and MonetaryValue) segmentation
3. Segmentation K-means Clustering
4. K-means Clustering Analysis
5. Conclusions

1. Customer retention (%)

Import, clean, filter and organize data

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

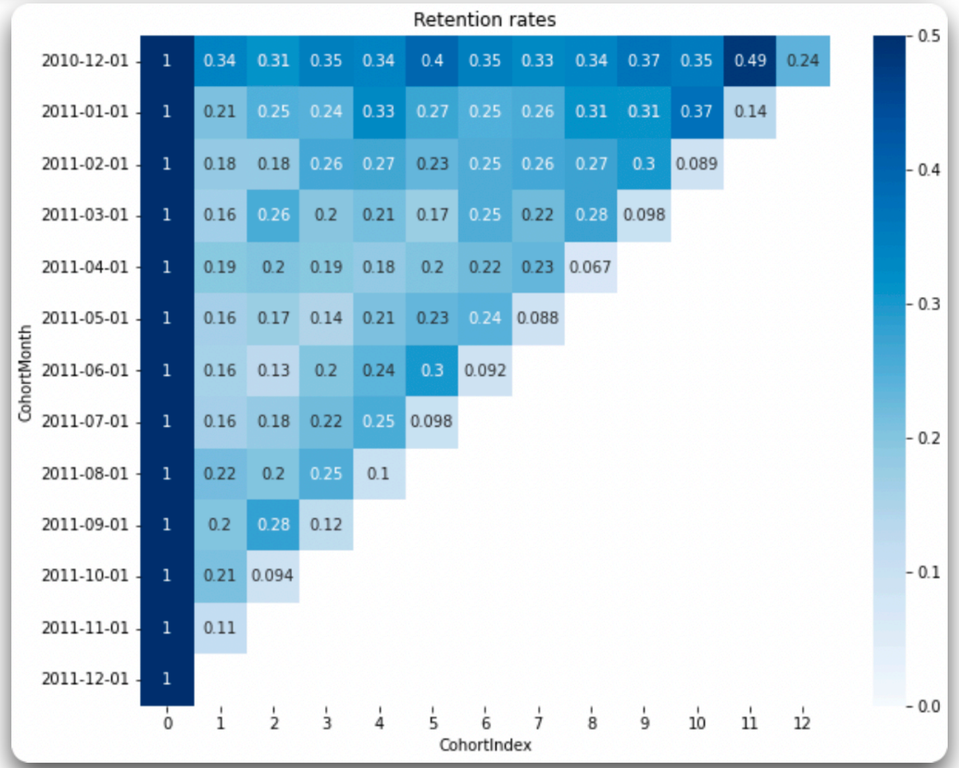
1. Customer retention (%)

Transform data in pivot tables

	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



HeatMap interpretation:

"Only 22% of customers who bought our products (for the first time) in July 2011, bought again 03 months later."

“49% of customers who bought our products (for the first time) in December 2010, bought again 11 months later. Making November 2011 the month with the highest percentage of customer retention”.

2. RFM segmentation

Definition of metrics (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

RFM interpretation

- **Recency:**
Measures the number of days that have passed since the customer's last purchase in the last 12 months.

- **Frequency:**
It measures the cumulative number of times the customer purchased during the last 12 months.

- **MonetaryValue:**
It measures the cumulative amount of money that the customer has spent on our products in the last 12 months.

The analysis could be extended to 24 or 36 months, however, for practical reasons of the project and to give greater importance to the most recent behavior of the clients, I filtered the dataset of the last 12 months.

2. RFM segmentation

Metric analysis (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

RFM Analysis

Customers can be segmented in three ways:

- **RFMS Segment:**

It is the concatenation of the RFM columns. It will depend on the number of percentile groups of equal size that you initially assign (in this case 04). The advantage is that there is a wide variety of combinations for a more specific segmentation (exploration of market niches).

- **RFM Score:**

It is the sum of the RFM columns. It allows to have a more generalized panorama in terms of customer segmentation.

- **General Segment:**

They are personalized commercial labels based on the grouping of RFM_Score values.

3. Segmentation K-means Clustering

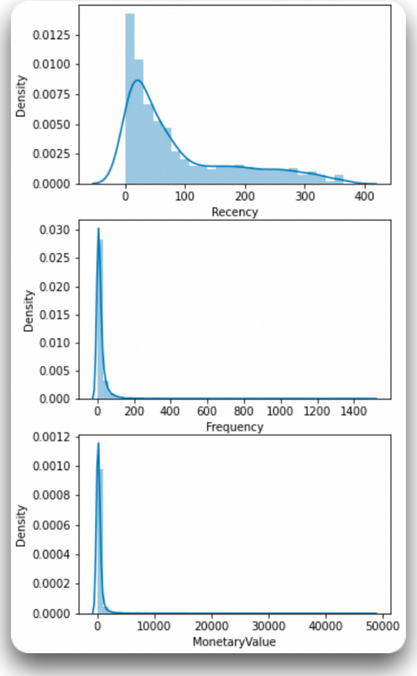
Logarithmic transformation and standardization of RFM variables

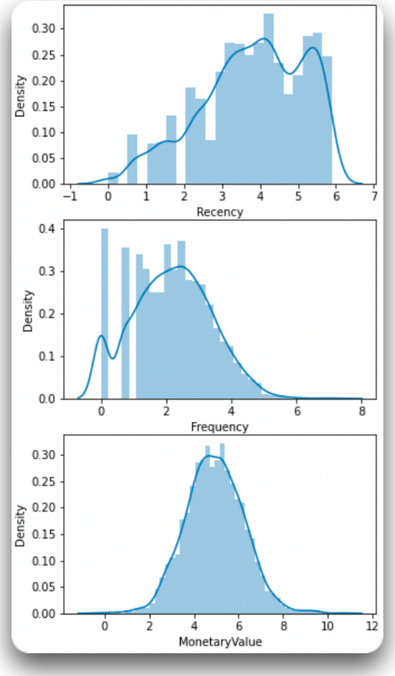
CustomerID	Recency	Frequency	MonetaryValue
12747	3	25	948.70
12748	1	888	7046.16
12749	4	37	813.45
12820	4	17	268.02
12822	71	9	146.15

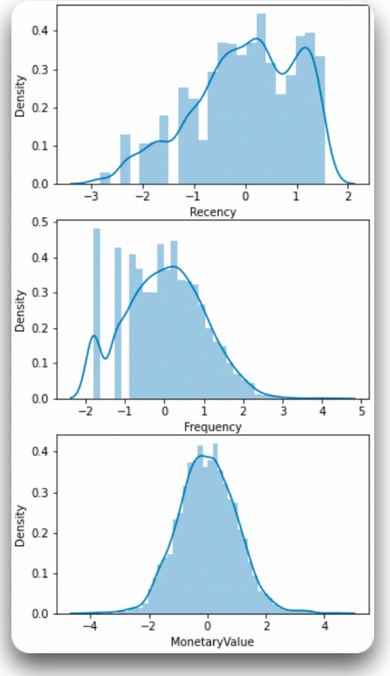
	Recency	Frequency	MonetaryValue
count	3643.00000	3643.000000	3643.000000
mean	90.43563	18.714247	370.694387
std	94.44651	43.754468	1347.443451
min	1.00000	1.000000	0.650000
25%	19.00000	4.000000	58.705000
50%	51.00000	9.000000	136.370000
75%	139.00000	21.000000	334.350000
max	365.00000	1497.000000	48060.350000

	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







Important note:

The downside of these older methods is that there is no way to know if one number of segments is better than another and is subject to team subjectivity.

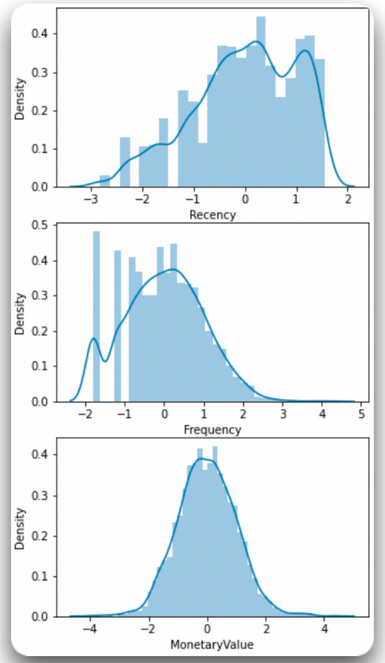
K-Means seeks a balance between having an adequate commercial segmentation of clients and reducing forecast error (SSE) as much as possible through "The elbow method".

K-means works much better as long as the distribution of the variables is standardized, that is, mean 0 and standard deviation 1.

3. Segmentation K-means Clustering

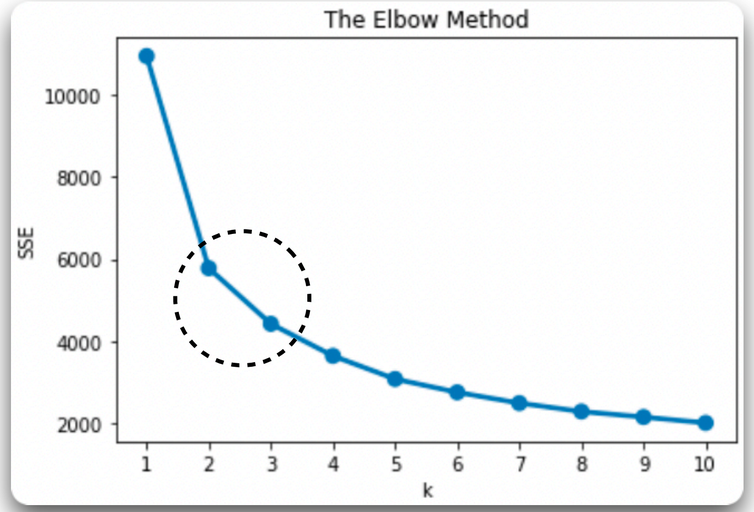
Identification of optimal value of K

	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. K-means Clustering Analysis

With Snake Plot and 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 varialbles ID
                        id_vars=['CustomerID', 'Cluster'],
                        # Asignar RFM_Values como value_vars
                        value_vars=['Recency', 'Frequency', 'MonetaryValue'],
                        var_name='Attribute',
                        value_name='Value')
datamart_melt
```

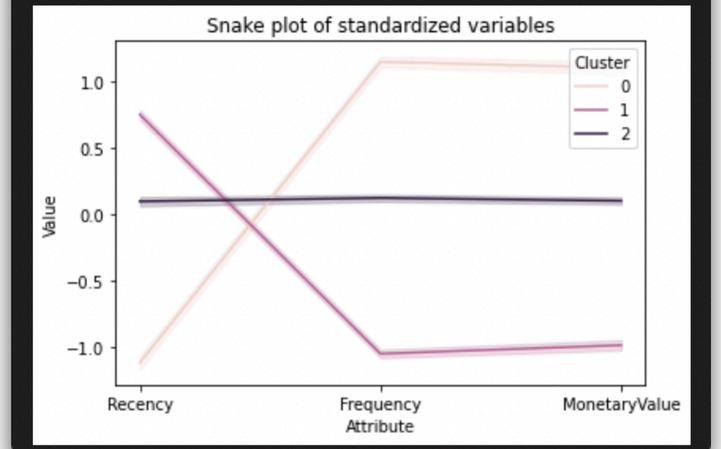
✓ 0.7s

	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

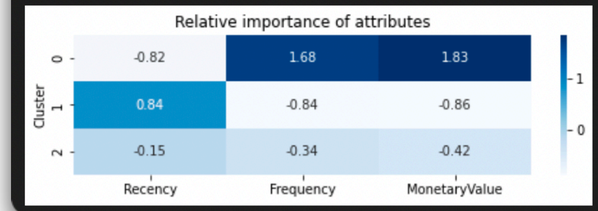
```python
# Graficar el snake plot
plt.title('Snake plot of standardized variables')
sns.lineplot(x='Attribute', y='Value', hue='Cluster',
             data=datamart_melt)

# 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. Conclusions

Key points

- RFM segmentation allows companies to identify trends in the behavior of their customers: the most recent, the most frequent and those who spend the most on our products.
- The goal is to identify customers who spend more and buy our products more often. Create loyalty strategies to ensure our sales and, therefore, greater inventory turnover.
- K-means is a solution against the subjectivity that is subject to the RFM_Segment and RFM_Score, looking for a balance between reducing the sum of squared errors of its classification and an interpretable quantity for commercial strategies.

Ideas

- Segmentation should not be limited only to customers but also to products with the aim of creating business strategies:
  - Increase inventory turnover
  - Exploration of market niches
  - Increase working capital in more commercial products
  - Offer seasonal products to the market by analyzing the 'Relative Importance of Attributes' plot to identify the most commercial months for a certain number of products.