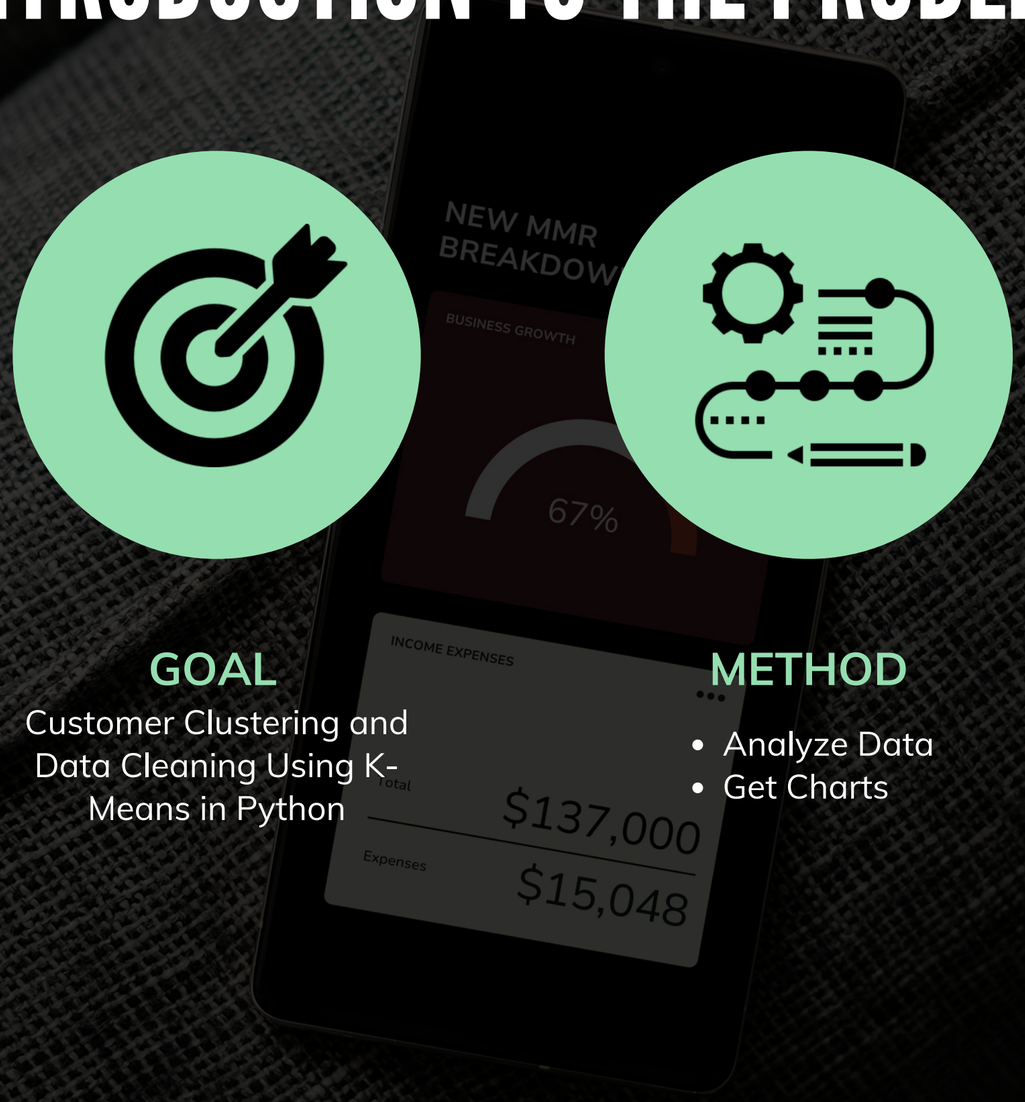
CUSTOMER CLUSTERING AND DATA CLEANING USING K-MEANS IN PYTHON

MARÍA FERNANDA RUBÍ EGUEZ
19/08/2025

TOOLS AND LIBRARIES:
PANDAS, NUMPY, MATPLOTLIB, SEABORN, SKLEARN

INTRODUCTION TO THE PROBLEM



GOAL

Customer Clustering and Data Cleaning Using K-Means in Python

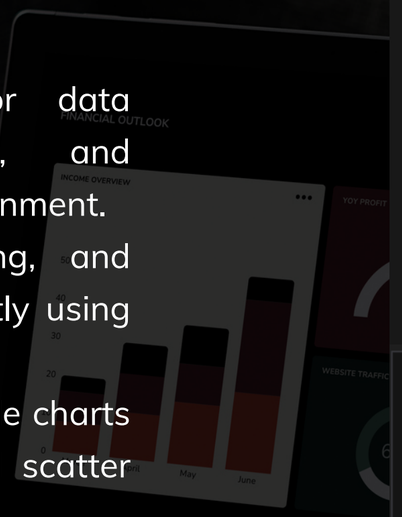
METHOD

- Analyze Data
- Get Charts

TOOLS USED

DATA SOURCE:
VENTAS_2025.XLSX

- Python (Jupyter Notebook): For data analysis, predictive modeling, and visualization in an interactive environment.
- Pandas: For structuring, cleaning, and manipulating tabular data efficiently using DataFrames.
- Matplotlib: For building customizable charts such as line, bar, histogram, and scatter plots.
- Seaborn: For creating professional and statistical visualizations (heatmaps, boxplots, regression plots) with an elegant style.
- Scikit-learn (sklearn): For machine learning tasks such as classification, regression, clustering (e.g., K-Means), dimensionality reduction, and model evaluation.



```
#Import all the neccesary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
```

```
#Import the Dataframe
df = pd.read_excel("Ventas_2025.xlsx")
df.head(5)
```

	Edad	Frecuencia_Compra	Monto_Promedio	Compras_Anuales
0	56	7	242	16
1	46	2	73	71
2	32	3	77	69
3	60	17	342	37
4	25	5	379	44

DATA EXPLORATION AND STRUCTURE

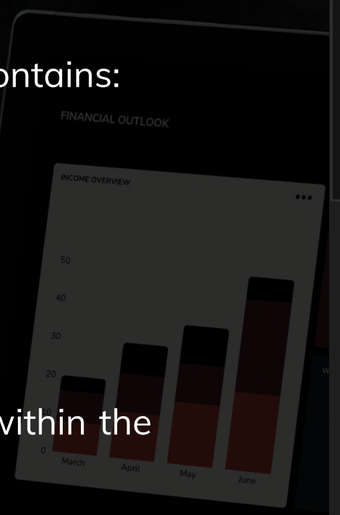
READING HISTORICAL DATA:

Historical data is loaded, the dataset contains:

- Edad
- Frecuencia_Compra
- Monto_Promedio
- Compras_Anuales

There are no null or duplicate values within the table.

The names of the columns were transcribed from Spanish to English




```python
#Import the Dataframe
df = pd.read_excel("Ventas_2025.xlsx")
df.head(5)
```

	Edad	Frecuencia_Compra	Monto_Promedio	Compras_Anuales
0	56	7	242	16
1	46	2	73	71
2	32	3	77	69
3	60	17	342	37
4	25	5	379	44

```python
#Renaming the columns from spanish to english
df.rename(columns={"Edad":"Age"}, inplace=True)
df.rename(columns={"Frecuencia_Compra":"Purchase_freq"}, inplace=True)
df.rename(columns={"Monto_Promedio":"Avg_amount"}, inplace=True)
df.rename(columns={"Compras_Anuales":"Annual_purchases"}, inplace=True)
```

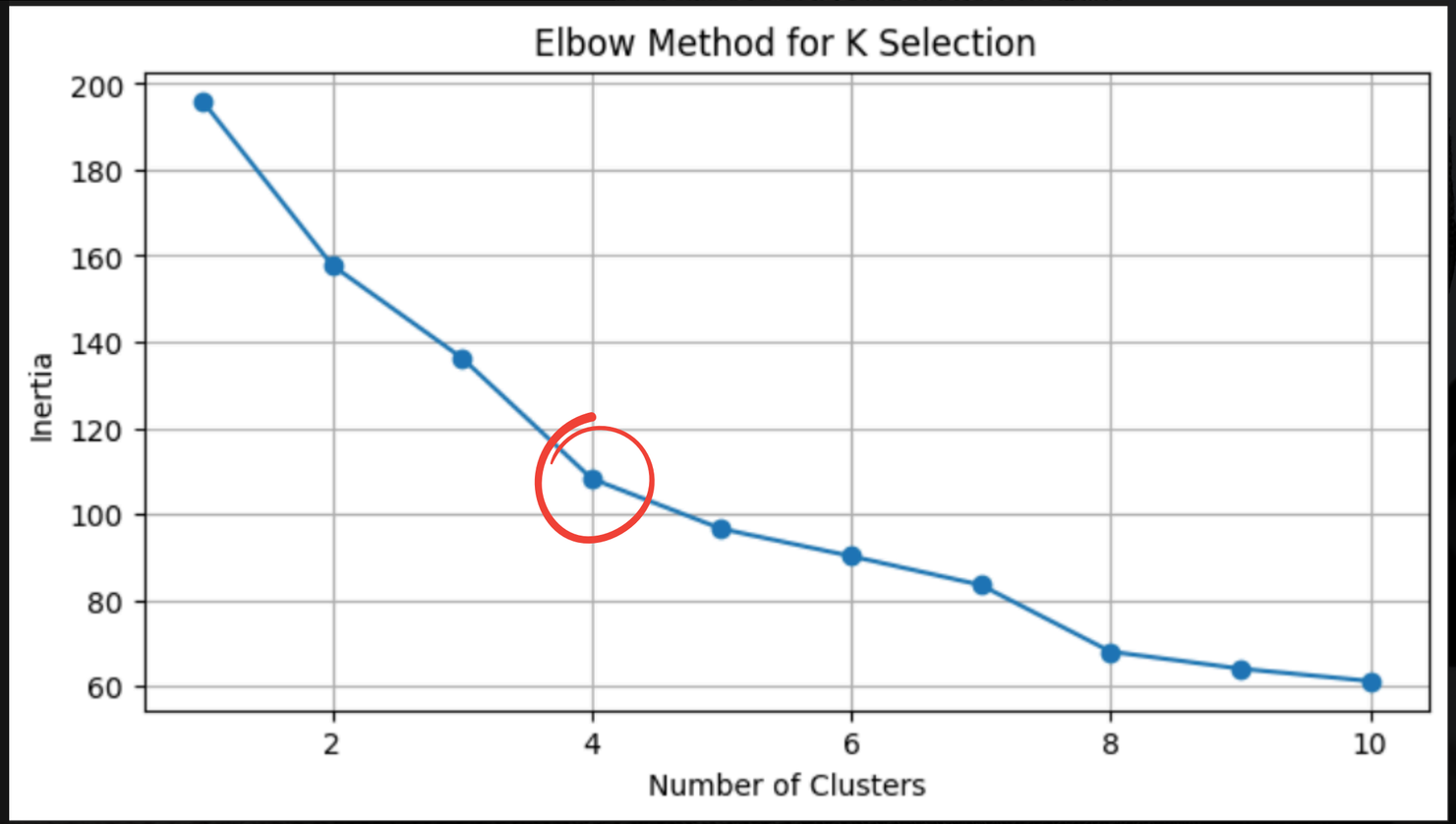
MODEL DEVELOPMENT

To perform the elbow method analysis, a range of clusters from 1 to 11 was considered. This allows dividing the data into up to 10 possible cluster solutions in order to identify the point where the elbow occurs
The curve shows a sharp drop between K=1 and K=4. From K=4 onwards, the decrease in inertia becomes more gradual and does not provide as much gain.

```
#Scaling numeric data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df)

#Elbow Method
inertia = []
k_range = range(1, 11)
for k in k_range:
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(X_scaled)
    inertia.append(kmeans.inertia_)

plt.figure(figsize=(8, 4))
plt.plot(k_range, inertia, marker='o')
plt.title('Elbow Method for K Selection')
plt.xlabel('Number of Clusters')
plt.ylabel('Inertia')
plt.grid(True)
plt.show()
scaler = StandardScaler()
X_scaled = scaler.fit_transform(df)
```

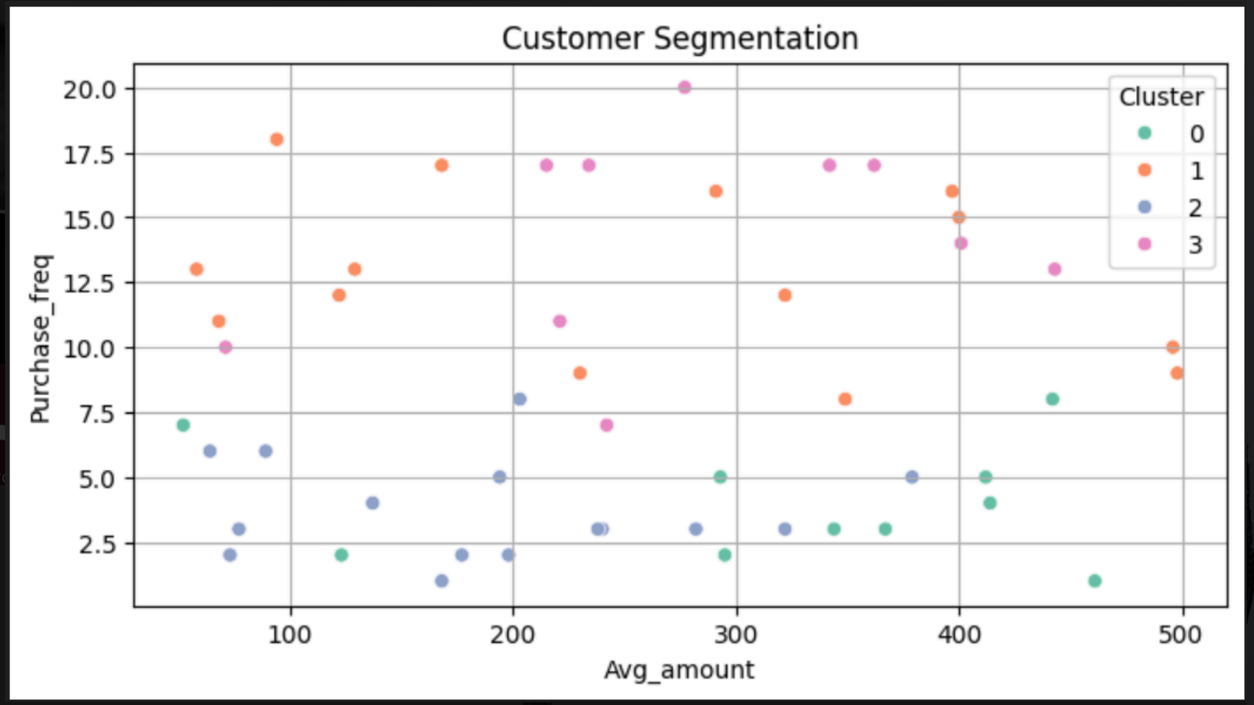


MODEL EVALUATION

Using the elbow method, the optimal number of clusters was determined to be 4. This means that the dataset of customers can be grouped into four distinct segments. Each cluster represents a group of customers that share similar characteristics. The scatter plot shows the distribution of customers across these four clusters, providing a visual representation of the segmentation.

```python
#Train model with optimal K (you can adjust here)
k_optimo = 4
modelo_kmeans = KMeans(n_clusters=k_optimo, random_state=42)
clusters = modelo_kmeans.fit_predict(X_scaled)
df['Cluster'] = clusters

#Cluster visualization (we use 2 main variables if there are many)
plt.figure(figsize=(8, 4))
sns.scatterplot(data=df, x=df.columns[2], y=df.columns[1], hue='Cluster', palette='Set2')
plt.title('Customer Segmentation')
plt.xlabel(df.columns[2])
plt.ylabel(df.columns[1])
plt.grid(True)
plt.legend(title='Cluster')
plt.show()
```

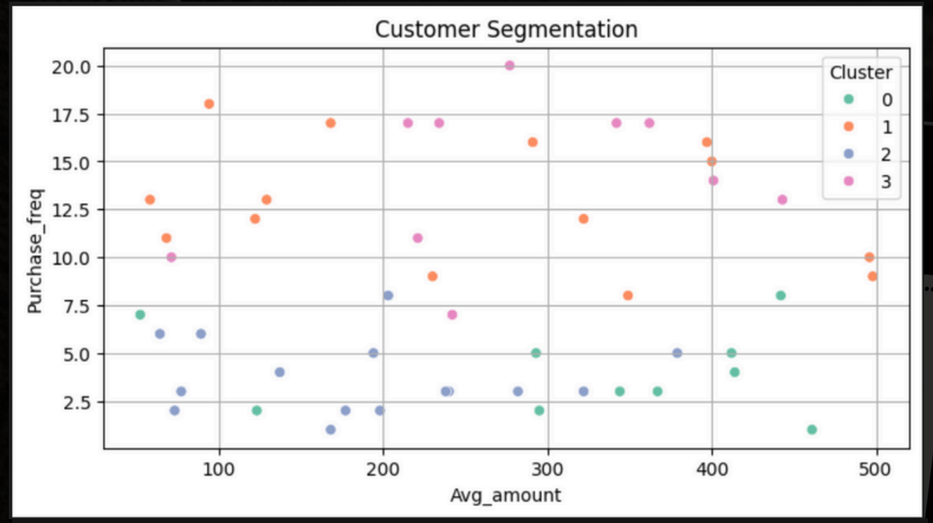



```python
#Averages per cluster
print("📊 Average profile per cluster:")
display(df.groupby('Cluster').mean().round(2))
```

📊 Average profile per cluster:

Cluster	Age	Purchase_freq	Avg_amount	Annual_purchases
0	56.30	4.00	320.30	57.30
1	36.14	12.79	258.71	64.14
2	37.67	3.73	189.40	40.60
3	52.60	14.30	280.80	25.30

CONCLUSIONS




```
#Averages per cluster
print("📊 Average profile per cluster:")
display(df.groupby('Cluster').mean().round(2))
```

```
📊 Average profile per cluster:
```

Cluster	Age	Purchase_freq	Avg_amount	Annual_purchases
0	56.30	4.00	320.30	57.30
1	36.14	12.79	258.71	64.14
2	37.67	3.73	189.40	40.60
3	52.60	14.30	280.80	25.30

Cluster 0 (56.30 years old, 4.00 purchases, $320.30 average, $57.30 annually): Older customers with low purchase frequency, but high average spending and moderate annual purchases. They could be occasional high-value customers.

Cluster 1 (36.14 years old, 12.79 purchases, $258.71 average, $64.14 annually): Younger customers with high purchase frequency and high annual spending, although average spending per purchase is moderate. They represent an active and potentially loyal group.

Cluster 2 (37.67 years old, 3.73 purchases, $189.40 average, $40.60 annually): Middle-aged customers with low purchase frequency, low average spending, and moderate annual purchases. They could be occasional customers with less commitment.

Cluster 3 (52.60 years old, 14.30 purchases, average $280.80, annual $25.30): Older customers with high purchase frequency, high average spending, but low annual volume. They could be frequent buyers with smaller transactions.

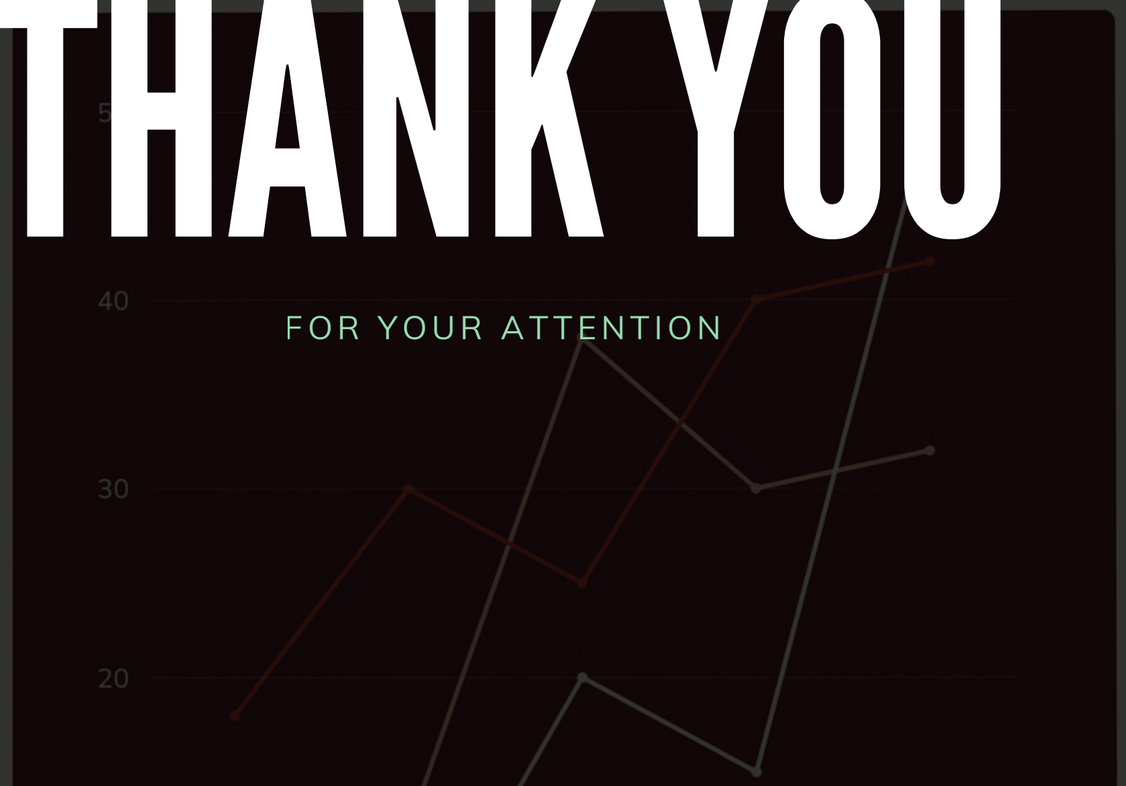
Some marketing strategies that I would recommend would be:

Cluster 0: Have special discounts on premium products to make the high expense worthwhile.
Cluster 1: Focus on a customer retention campaign, loyalty or to become premium customers.
Cluster 2: Encourage them to buy more
Cluster 3: Maintain customer loyalty.

I WANT TO SAY

THANK YOU

FOR YOUR ATTENTION



GROUND

HE INDUSTRY'S HISTORY