NTNU

Norwegian University of Science and Technology

Customer segmentation with clustering

- Aduragbemi Samuel Adetunji
- Azimil Gani Alam
- Muhammad Tsaqif Wismadi
- Preston William Buscay
- Giacomo Melloni
- Chinmay Anirudha Patwardhan
- Johan Nicolas Suarez Lojan

Content

- Purpose
- Method
- Example
- Dataset
- Step-by-step
- Results

Purpose

- Key technique in business and marketing analysis.
- **Perform a cluster analysis** to divide your **consumers into groups** and customize your marketing strategy for each group.
- You wouldn’t want your company targeting vegan meal plans to barbecue enthusiasts or promoting hiking gear to people living in urban areas with no access to nature.

Method

Understand the Business Problem
- How to segment the customers

Apply **a clustering algorithm**

Interpret the Cluster Result

Method

Segmentation models

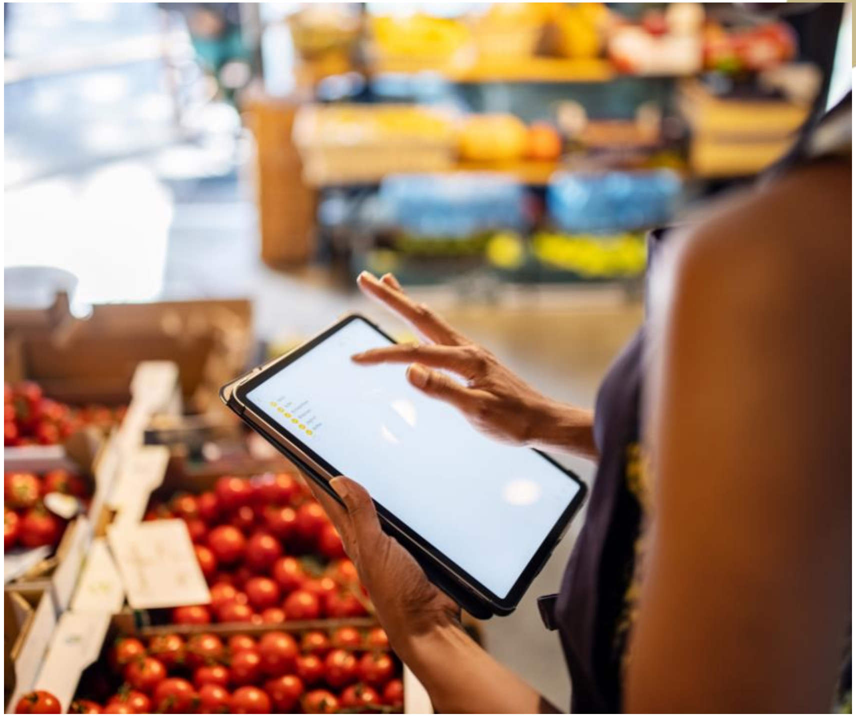
- **Demographic Segmentation** : Age, gender, income, education, marital status
- **Geographic Segmentation** : Country, state, city, town
- **Psychographic Segmentation** : Personality, attitude, values, and interest
- **Technographic Segmentation** : Mobile use, desktop use, apps, and software
- **Behavioral Segmentation** : Tendencies and frequent actions, feature or product use, habits
- **Needs-Based Segmentation** : Product or service must-haves and needs of specific customer groups
- **Values-Based Segmentation** : Economic value of specific customer groups on the business

Example

- A convenience store have a comprehensive data of their customer.

- They have a total of :
  2,240 data points : **29 features**.

- **Objective**: Group customers into 4 classes of similiar traits for a more *effective advertisement campaign*.



Dataset: Demographic

Feature	Description
ID	A unique identifier for each customer.
Year_Birth	Customer's birth year (used to calculate age).
Education	Customer's education level (e.g., graduation, PhD, etc.).
Marital_Status	Customer's marital status (e.g., single, married, etc.).
Income	Customer's yearly household income.
Kidhome	Number of children living in the customer's household.
Teenhome	Number of teenagers living in the customer's household.
Dt_Customer	The date when the customer enrolled with the company.
Recency	Number of days since the customer’s last purchase.
Complain	Indicator if the customer has complained in the last two years (1 = Yes, 0 = No).

Dataset: Products

Feature	Description
MntWines	Amount spent on wine in the last 2 years.
MntFruits	Amount spent on fruits in the last 2 years.
MntMeatProducts	Amount spent on meat products in the last 2 years.
MntFishProducts	Amount spent on fish products in the last 2 years.
MntSweetProducts	Amount spent on sweet products in the last 2 years.
MntGoldProds	Amount spent on gold products in the last 2 years.

Dataset: Promotion

Feature	Description
NumDealsPurchases	Number of purchases made with a discount.
AcceptedCmp1-5	Indicator if the customer accepted offers from campaigns 1 through 5 (1 = Yes, 0 = No).
Response	Whether the customer accepted the offer in the last campaign (1 = Yes, 0 = No).

Dataset: Place

Feature	Description
NumWebPurchases	Number of purchases made through the company’s website.
NumCatalogPurchases	Number of purchases made using the catalog.
NumStorePurchases	Number of purchases made directly in stores.
NumWebVisitsMonth	Number of visits to the company’s website in the last month.

Step-by-step

1) Load the dataset
2) Handle missing values
3) Encode categorical variables
4) Standardize the data
5) Apply K-Means
6) Visualise clusters
7) Display cluster traits

Step-by-step

0. Import libraries

[1] 2s

```
# Importing necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.cluster import KMeans
```

1. Load dataset

[2] 0s

```
# Load the dataset
df = pd.read_csv('marketing_campaign.csv', delimiter='\t')
```

2. Handle missing values

[3] 0s

```
# Handle missing values
df.fillna(df.median(numeric_only=True), inplace=True)
```

Step-by-step

3. Encode categorical variables

```
[4] label_cols = ['Education', 'Marital_Status']
    label_encoders = {col: LabelEncoder() for col in label_cols}

    for col in label_cols:
        df[col] = label_encoders[col].fit_transform(df[col])
```

```
[5] # Convert 'Dt_Customer' column to datetime format and extract year
    df['Dt_Customer'] = pd.to_datetime(df['Dt_Customer'], format='%d-%m-%Y')
    df['Customer_Year'] = df['Dt_Customer'].dt.year
```

```
[6] # Drop 'Dt_Customer' and 'ID' columns as they are not needed for clustering
    df.drop(columns=['Dt_Customer', 'ID'], inplace=True)
```

Step-by-step

4. Standardize the data

✓ 0s [7]

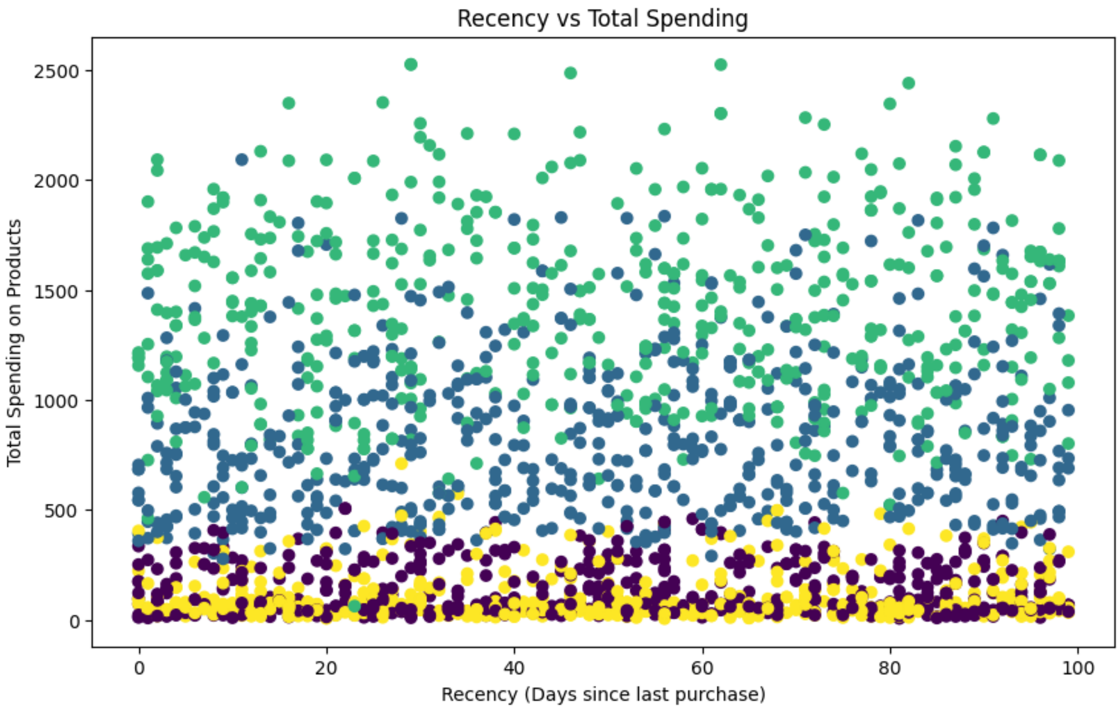
```
# Standardizing the data
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df)
```

5. Apply K-means clustering

✓ 0s [8]

```
# Apply KMeans clustering to classify into 4 clusters
kmeans = KMeans(n_clusters=4, random_state=42)
df['Cluster'] = kmeans.fit_predict(df_scaled)
```

```python
# Plotting Recency vs Total Spending on products
total_spending = df['MntWines'] + df['MntFruits'] + df['MntMeatProducts'] + df['MntFishProducts'] + df['MntSweetProducts'] + df['MntGoldProds']
plt.figure(figsize=(10, 6))
plt.scatter(df['Recency'], total_spending, c=df['Cluster'], cmap='viridis')
plt.title('Recency vs Total Spending')
plt.xlabel('Recency (Days since last purchase)')
plt.ylabel('Total Spending on Products')
plt.show()
```



7. Display cluster traits

```
[12] # Step 7: Displaying key traits of each cluster
     cluster_traits = df.groupby('Cluster').mean()
     print(cluster_traits)
```

```
          Year_Birth  Education  Marital_Status        Income   Kidhome  \
Cluster
0        1962.891743   2.596330        3.746789  43838.816514  0.653211
1        1965.003401   2.610544        3.709184  59736.526361  0.188776
2        1968.765385   2.350000        3.775000  78209.469231  0.040385
3        1978.141397   2.027257        3.695060  29517.731687  0.863714

         Teenhome    Recency    MntWines  MntFruits  MntMeatProducts  ...  \
Cluster                                                               ...
0        0.955963  50.253211   78.355963   4.623853        28.185321  ...
1        0.903061  48.403061  506.901361  29.894558       158.442177  ...
2        0.105769  49.440385  617.201923  67.632692       480.342308  ...
3        0.045997  48.461670   32.553663   6.218058        26.686542  ...

         AcceptedCmp3  AcceptedCmp4  AcceptedCmp5  AcceptedCmp1  AcceptedCmp2  \
Cluster
0            0.042202      0.025688      0.000000      0.005505      0.003670
1            0.066327      0.137755      0.015306      0.028912      0.015306
2            0.086538      0.136538      0.296154      0.236538      0.036538
3            0.095400      0.001704      0.000000      0.001704      0.000000

         Complain  Z_CostContact  Z_Revenue  Response  Customer_Year
Cluster
0        0.009174            3.0       11.0  0.047706    2013.231193
1        0.010204            3.0       11.0  0.117347    2012.818027
2        0.003846            3.0       11.0  0.317308    2013.053846
3        0.013629            3.0       11.0  0.126065    2013.027257

[4 rows x 28 columns]
```

Results: Clusters of customer based on their similiar traits

Cluster	Age	Income	Family	Product Purchases	Campaign Responses	Complaints
Cluster 0	1964 (Age around 59-60)	Moderate income, around $57,066	0.25 children, 0.94 teenagers	Moderate spending on wine ($451), lower spending on meat and other products	Low acceptance rates	Very few customers complained
Cluster 1	1968 (Age around 55-56)	Higher income, around $73,871	Smaller household (0.05 children, 0.23 teenagers)	High spending on wine ($488), higher spending on meat products ($429)	Moderate acceptance rates	Similar to Cluster 0
Cluster 2	1972 (Age around 51-52)	Lower income, around $34,826	More children (0.8) and fewer teenagers (0.43)	Very low spending on all products (wine spending is $39)	Lowest acceptance rates across all clusters	Slightly higher complaints
Cluster 3	1969 (Age around 55)	Highest income, around $81,747	Very few children or teenagers	Highest spending on wine ($875), meat products ($469), and overall product categories	Highest acceptance rates (94%)	Very low complaint rate

References

- K. Baker, "Customer segmentation: How to segment users & clients effectively," *HubSpot*, Apr. 18, 2024. [Online]. Available: https://blog.hubspot.com/service/customer-segmentation. [Accessed: Sep. 30, 2024].
- A. Nair, "Customer segmentation with clustering," *Towards Data Science*, Nov. 8, 2021. [Online]. Available: https://towardsdatascience.com/customer-segmentation-with-clustering-933caa4c9ea3. [Accessed: Sep. 30, 2024].