Customer segmentation: Clustering

# Customer personality analysis is an in-depth understanding of company’s customers. This analysis helps the company to understand their customer base better and also, they can offer more relevant products to each customer based on the cluster (Customer segment) they fall in. Customer personality analysis helps the company to make a better product. This is a win-win situation for both customer and company because, customers get better recommendations and experience this in turn increases the revenue for the company.

# This is an unsupervised machine learning project where we have to identify customer segments using clustering technique. We have to find out hidden insights of customer’s personal traits based on various clusters.

## The dataset is from a marketing campaign of a company.

## Dataset source: <https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis>

## There are 4 kinds of details provided:

1. Customer’s personal details: Birth year, Education, Marital status, Income, Kidhome, teenhome, The date they became customer for the company, Recency, Complain by customer
2. Product information: Amount spent by customer on Wine, Fruit, Meat, Fish, Sweet, Gold since last 2 years
3. Promotions details: Number of deals purchased by customer, Promotion acceptance rate for 1st to 5th promotion
4. Place details: Count of web based purchases, catalog based purchases, purchase made in store, web visits per month

## Customer’s personal details and product details are used in the clustering.

## Steps in the project and its details are given below:

## Importing the required libraries

* In this stage of project various libraries required for the project are imported, such as PANDAS, NUMPY, MATPLOTLIB, SEABORN, SCIKITLEARN.

## Loading the data

* This is a very simple stage of project, in which theS csv file is read.
* Although it’s simple file to read, after loading it in notebook it was found that the file is a tab separated one and not a comma separated.
* Just by specifying sep = “\t” the file was correctly loaded into the Jupyter notebook.

## Data cleaning and feature engineering

* This was a very crucial stage of the project as it involved creation of new features from existing features and removing redundant features.
* Converted birth year to age column
* Instead of multiple degree converted them all into either undergraduate, graduate or postgraduate
* Converted Married, Together, Divorced, Widow, Single into either Partner or Alone
* Using teenhome and kidhome created a new column as children count
* Created a column family size using living\_with column and children count
* Date time conversion of dt\_customer column
* All the spending on wine, fruit, gold etc. are summed up and put into new column total\_spending
* A new column is created for customer duration by subtracting each date from newest date.
* Removed outliers from Age (few people were having age > 120 years) and income (few people were having income > 600000)

## Data preprocessing

* Label encoding for Education column and living\_with column
* Removing redundant columns
* Scaling the data before clustering (Standard scaler is used)

## EDA on data

* In this part of the project two kinds of plots are used to identify relation of each feature with total\_spending target:

1. All Numeric columns are plotted against total\_spending on scatter plot
2. All categorical columns are plotted against total\_spending on bar plot

## Dimensionality reduction

* Even after removing redundant columns, still there are many columns.
* To cluster the data and visualization purpose PCA is used.
* 9 PCA components are kept for dimensionality reduction (Explained variance ratio = 0.8)

## Clustering of data

* Before clustering, number of clusters needs to be identified so for that elbow plot and silhouette score visualizer is used.
* 4 no. of clusters are selected.
* K-means clustering is used for clustering the data.

## Customer personality analysis based on clusters found

* After clustering is done 4 kinds of clusters are found

1. High-income high spending group
2. Medium-income low spending group
3. Low-income low spending group
4. Medium-income medium-high spending group

## Conclusion

* Low-income low spending group:

Comparatively younger

0 or 1 child

Majority living with partner

Majority is parent

Family size 1 to 3

* High income high spending group:

No children

All age range

More people living with partner than alone

Definitely not a parent

Family size 1 to 2

* Medium income medium-high spending group:

Majority have 1 teen at home

Comparatively high aged

1 - 2 children

Definitely a parent

Family size 2 to 4

* Medium income low spending group

Comparatively high aged

Majority living with partner

1 to 3 children

Family size 3 to 5