Customer Segmentation for a Retail Store

Name: Pushkar Varshney

Start Date: 14-july-2024

End Date: 19-july-2024

Submitted to: Abhishek Sir

Institute: Cipher School

Objectives: To Segment customers into distinct groups based on their purchasing behaviour.

Scope: Data cleaning, EDA, Customer segmentation using "K-Means", Visualization using Matplotlib and Power BI.

Used Materials: Google colab, PowerBI, python libraries, dataset file.

Introduction:

The main goal of this project is to learn about customer segmentation. This involves understanding customers and grouping them based on their age, annual Income, Spending Score. Once we have divided the customers into different groups, we can do some strategical steps for profits.

Data Information & Dependencies:

This dataset is composed by the following five features:
Customer ID: Unique ID assigned to the customer,
Gender: Gender of the customer
Age: Age of the customer,
Annual Income (k$): Annual Income of the customer
Spending Score (1-100): Score assigned by the mall based on customer behavior and spending nature.
In this particular dataset we have 200 samples to study.

Here are the dependencies I used:
pandas: Will help us treat and explore the data, and execute vector and matrix operations.
matplotlib|seaborn: Will help us plot the information so we can visualize it in different ways and have a better understanding of it.
plotly: Will also help us plotting data in a fancy way.
sklearn: Will provide all necessary tools to train our models and test them afterwards.

Data Collection & Cleaning:

We collect the data from the "Mall_Customer.csv" file which contain 200 records of dataset(customer Id, gender, age, annual income, spending score). The dates set also contain the missing and wrong headings. So, to correct this we first "clean the data" by putting the mean values in missing cells and rename the wrong headings in correct ones.

To check missing values: "df.isna().sum()" and rename function to rename the headings.

Exploratory Data Analysis (EDA):

Initial exploration revealed a seasonal trend in sales, with peaks during holiday seasons. The sales amount has a positive correlation with marketing spend. We visualize the data in the form of graphs for better understanding. To do so, we are only going to consider the following features: Annual_income, Spending_score and Age. Gender will only be used to make data sepparation so we can differentiate values for men and women.To begin with, we are plotting the histograms for each of the three.

After plotting the graph I find that distribution of these values resembles a Gaussian distribution, where the vast majority of the values lay in the middle and for gender I use pie chart to find the percentage of male and female.

Analysis and Modeling:

This stage involves applying various analytical techniques and models to understand the data better and derive actionable insights. For this we use clustering (KMeans) and to do clustering first we have toselect the “Number of clusters” and to do so we use “Elbow” technique.

Elbow technique: The elbow method is used to determine the optimal number of clusters in k-means clustering. In the clustering process, we used the elbow method to determine the optimal number of clusters. This method involves plotting the cost function, specifically the inertia (sum of squared distances to the nearest cluster center), against different values of k. The optimal number of clusters is found where adding another cluster does not significantly improve the model. In our case, the elbow point is at k=5, so we chose to divide our data into five clusters.

We decided not to consider the gender factor in the clustering process. The primary reason is that the difference in spending between men and women is not significant in this dataset, so differentiating by gender would not provide additional useful information. Additionally, most stores do not target specific genders anymore, as they offer products for both men and women. By excluding gender, we avoid interfering with the unsupervised learning process, allowing the algorithm to naturally identify clusters based on other factors.

After running the K-means algorithm and plotting the results on a 3D graphic, our task is to identify and describe the five clusters created. This approach helps us extract meaningful insights and understand the different customer segments.

- **Green Cluster** – The green cluster groups young people with moderate to low annual income who actually spend a lot.
- **Blue Cluster** – The blue cluster groups reasonably young people with pretty decent salaries who spend a lot.
- **Purple Cluster** – The purple cluster basically groups people of all ages whose salary isn't pretty high and their spending score is moderate.
- **Yellow Cluster** – The yellow cluster groups people who actually have pretty good salaries and barely spend money, their age usually lays between thirty and sixty years.
- **Dark green Cluster** – The dark green cluster groups whose salary is pretty low and don't spend much money in stores, they are people of all ages.

Conclusion:

- **KMeans Clustering**: Effective for customer segmentation.
- **Customer Segmentation**: Helps understand customer behavior and plan marketing strategies.
- **Spending Scores**: Similar between men and women.

- **Young Shoppers**: Spend the most, making them key marketing targets.
- **Middle Class**: Largest cluster, important to consider in strategies.
- **Discounts**: Could encourage higher spending from less active shoppers.

The owner next profitable step based on cluster data, tailor strategies for each group: “Target young spenders in the Green Cluster with engaging promotions and products. For the Blue Cluster, offer premium deals to attract high earners. The Purple Cluster benefits from value-oriented promotions, while the Yellow Cluster should receive exclusive luxury offers to encourage higher spending. For the Dark Green Cluster, focus on budget-friendly options and loyalty programs to increase their engagement”. This approach ensures each cluster’s specific needs and behaviors are addressed, driving overall sales.