Mall Customer Segmentation

Abhinav Duvvuru   
 Artificial Intelligence and Machine Learning  
Lambton CollegeToronto, Canada  
abhinavduvvuru@gmail.com  
  
Sandeep Reddy Alla  
Artificial Intelligence and Machine LearningLambton CollegeToronto, Canada  
itsmealla05@gmail.com  
  
SivaPriya TatiReddy  
Artificial Intelligence and Machine LearningLambton CollegeToronto, Canada  
sivapriya721@gmail.com   
  
Uma Maheswararao Gutta Artificial Intelligence and Machine LearningLambton CollegeToronto, Canada umamaheshgutta@gmail.com

*Abstract*— In a study, we analyze the mall’s customer data and group or cluster the customers into different groups based on their spending score and annual income. Shopping complexes use their customers’ data and develop machine learning models to target their audience to amplify sales. We analyze the customer data and perform exploratory data analysis and visualize the trends/patterns in different graphs, charts, and plots. Finally, we apply the K-means clustering algorithm to cluster the customers.

# Introduction

Customer segmentation plays an important role in analyzing the behavior of the customers. Retail stores and companies often divide the customers based on various factors like their age, gender, annual income, and spending score etc. to create targeted marketing strategies and services and to cross-sell and upsell specific products based on customer purchase behavior. Finally, to establish better customer relationships and to improve services.

# Dataset description

Mall Customer data is an interesting dataset that has hypothetical customer data. The dataset used in this project is the data obtained from Kaggle, which has Customer ID, Customer Gender, Customer Age, Annual Income of the customer and Spending score of the customer. The dataset has total 5 attributes. Below are the list of attributes and their type:

1. Customer ID (numerical) - ID of the Customer.

2. Customer Gender(categorical) -Gender of the customer.

3. Customer Age – Age of the customer.

4. Annual Income – Annual income of the customer

5. Spending score - Score assigned by the mall based on customer behavior and spending nature

## Dataset loading

The “Mall\_Customers.csv” dataset can be loaded using Python Pandas and stored as a dataframe. The below figure 1 depicts the data in the dataset.

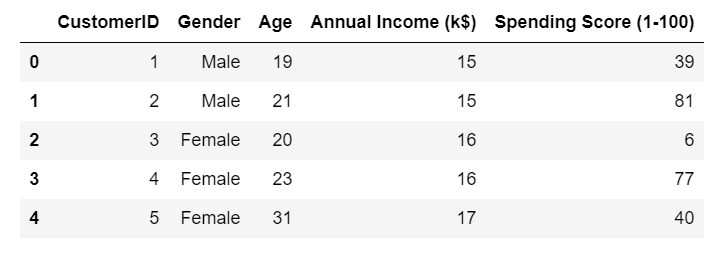


Fig. 1. Overview of dataset

## Data preprocessing

The column names of the dataset are renamed for further analysis as shown in figure 2 and from the info() method it is evident that the dataset has both categorical and numerical data and there are no missing values that needs to be handled or imputed.

Table

Description automatically generated

Fig. 2. Renamed columns of the dataset

# Exploratory data analysis

We applied different visualization libraries like matplotlib and seaborn to analyze the data in various formats like graphs, charts, and plots.

## Analyzing each attribute

Chart, bar chart

Description automatically generated

Fig. 2. Plot of gender with x-axis as gender category and y-axis as count

Chart, pie chart

Description automatically generated

Fig. 3. Pie chart of Gender attribute

From the above figures 2 and 3, there are more female customers when compared to male customers.

Chart, histogram

Description automatically generated

Fig. 4. Distribution plot of Age attribute

From figure 4, the Age attribute has values that are normally distributed and the distplot looks like a bell curve, which means that majority of data points are relatively similar.

The histogram (figure 5) and density plot (figure 6) of annual income attribute shows that the average income of all the customers is 60.56 and the distribution is normal

Chart, histogram

Description automatically generated

Fig. 5. Histogram of Annual Income attribute

Chart, histogram

Description automatically generated

Fig. 6. Density plot of Annual Income attribute

## Handling outliers

The outliers if any in a dataset can be detected using boxplot. The below boxplot of annual income attribute shows that there are some outliers which needs to be handled.

Chart, box and whisker chart

Description automatically generated

Fig. 7. Boxplot of Annual Income attribute before removing outliers

Outliers in the data can be detected and handled using interquartile range method or standard deviation method. We used IQR method to handle the outliers and below figure 8 represents the data without outliers of annual income attribute.

Chart, box and whisker chart

Description automatically generated

Fig. 8. Boxplot of Annual Income attribute after removing outliers

## Correlation between attributes

The corr() method finds correlation between numeric attributes and relationship among them, both positively and negatively. Figure 9 depicts the relation between attributes.

Chart

Description automatically generated

Fig. 9. Heatmap of numeric attributes

The above heatmap depicts that Age is negatively correlated with Score, AIncome is positively correlated with CustomerId and then followed by Age.

## Converting categorical data into numeric type

The gender attribute has values as male and female, which can be converted into numeric types as 0 and 1 for male and female respectively, as shown in figure 10.

Table

Description automatically generated

Fig. 10. After converting categorical data into numeric data

## Scatter plot between attributes

The scatter plot between Age and Score shows that spending among young is more compared to adults.

Chart, scatter chart

Description automatically generated

Fig. 11. Scatter plot between Age and Score

Chart, scatter chart

Description automatically generated

Fig. 12. Scatter plot between Annual income and Age

The above plot shows that people between age group of 30-50 years earns highest annual income when compared to others

Chart, scatter chart

Description automatically generated

Fig. 13. Scatter plot between Annual income and spending score

## Normalization of features

Before apply the model, all the features are standardized using standardscaler() method to get them around the same scale.

# model building

After converting and successfully processing the data we applied K-means clustering algorithm with number of clusters(K) as 3 as shown in figure 14. To find the most appropriate value of K, we applied elbow method.

Chart, scatter chart

Description automatically generated

Fig. 14. K-means with number of clusters as 3

Elbow method and silhouette score methods are used to find the number of clusters (k value) from the below applied elbow method, the K value can be assumed as 5.

Chart, line chart

Description automatically generated

Fig. 15. Elbow method to find number of clusters

Chart, scatter chart

Description automatically generated

Fig. 16. Clustering with number of clusters as 5

From the above final scatter plot, we can perceive that the customers present in our dataset could be clustered into 5 distinct groups based on their annual income and spending score.

1. Red: low annual income, low spending score
2. Orange: low annual income, high spending score
3. Violet: intermediate annual income, intermediate spending score
4. Blue: high annual income, high spending score
5. Green: high annual income, low spending score

# Visualizing The results

Here we try to visualize the data based on wide variety of ages and correlations between the income and spending scores. Visualize the patterns and customers into groups based on the k-value.

Chart, bar chart

Description automatically generated

Fig. 17. Bar graph between clusters predicted and age feature

Chart, bar chart

Description automatically generated

Fig. 18. Bar graph between clusters predicted and annual income feature

## Interactive plots between clusters predicted and other features

Below are various interactive plots between clusters predicted and other attributes like spending score, customer id and annual income.

Chart, bar chart

Description automatically generated

Fig. 19. Interactive plot between clusters and spending score

Chart, bar chart

Description automatically generated

Fig. 19. Interactive plot between clusters and customer id

Chart, bar chart

Description automatically generated

Fig. 20. Interactive plot between clusters and Annual Income

# Conclusion

Customer segmentation is a separation of a market into multiple distinct groups of consumers who share the similar characteristics. Segmentation of market is an effective way to define and meet customer needs.

Blue cluster (figure 16)- high annual income, high spending score The average age is 32 years; predominantly female; Average Annual Income is 85k in dollars; Average Spending Score is 82

Violet cluster-intermediate annual income, intermediate spending score The average age is 43 years; predominantly female; Average Annual Income is 55k in dollars; Average Spending Score is 49

Red cluster- low annual income, low spending score The average age is 45 years; predominantly female; Average Annual Income is 26k in dollars; Average Spending Score is 21

Green cluster- high annual income, low spending score The average age is 41 years; there is an almost equal proportion of males and females; Average Annual Income is 86k in dollars; Average Spending Score is 17

##### References

1. “Customer segmentation Dataset”. Kaggle [online dataset], November 2021.Available:Kaggle,https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python [Accessed: November 16, 2021]
2. “Who’s who: Understanding your business with customer segmentation”. Intercom [online document], Decemeber 2021. Available: Intercom, <https://www.intercom.com/blog/customer-segmentation/> [Accessed: December 7, 2021]
3. Chris Piech, “K means”. Standford [online document], December 2021. Available: Stanford.edu <https://stanford.edu/~cpiech/cs221/handouts/kmeans.html>