LOVELY PROFESSIONAL UNIVERSITY

CAPSTONE PROJECT REPORT

(Project Term ~ January - May 2023)

ON

MALL CUSTOMER SEGMENTATION

Submitted by

ABHAY SAHANI

REGISTRATION NUMBER : 12115826

SECTION - K21HC ROLL NO. - 68

Under The Guidance Of

Mr. VED PRAKASH CHAUBEY SIR

63892

SUBJECT - PYTHON PROJECTS (INT216)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

LOVELY PROFESSIONAL UNIVERSITY

PHAGWARA , PUNJAB

Declaration

We hereby declare that the project work entitled " **Mall Customer Segmentation** " is an authentic record of my own work carried out as requirements of Capstone Project for the award of B.Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara, under the guidance of **Mr. Ved Prakash Chaubey**, during January to May 2023. All the information furnished in this capstone project report is based on my own intensive work and is genuine.

Name of Student~ **Abhay Sahani**

Registration No. ~ **12115826**

(Signature of Student)

Date : 5th May , 2023

Certificate

This is to certify that the declaration statement made by this group of students is correct to the best of my knowledge and belief. They have completed this Capstone Project under my guidance and supervision. The present work is the result of their original investigation, effort and study. No part of the work has ever been submitted for any other degree at any University. The Capstone Project is fit for the submission and partial fulfillment of the conditions for the award of B.Tech degree in Computer Science with Specializations in Data Science (AI and ML) from Lovely Professional University, Phagwara.

Signature :

Designation :

School of Computer Science & Engineering

Lovely Professional University

Phagwara, Punjab

Date : 05 - 05 - 2023

Acknowledgement

I would like to acknowledge the inspiration and support that helped me complete this project. **Ved Prakash Chaubey Sir**, helped me to get equipped with the necessary resources to complete this project. His constructive advice became a very integral part of the successful completion of the project.

I am grateful to the online data science community for sharing their knowledge and expertise, which helped me learn and develop my skills. I also appreciate the support of my family and friends, who encouraged and motivated me throughout the project.

Table Of Content

Introduction

Customer segmentation is a powerful technique that allows businesses to better understand their customers and tailor their marketing strategies accordingly. In this project, we will use clustering algorithms to segment customers of a shopping mall based on their spending behavior. The goal is to identify different customer groups based on their spending habits and demographics, so that the mall can create targeted marketing campaigns and improve its overall customer experience.

Overall customer segmentation is a very crucial tool becoming a must for every organization to use in this modern era of data. It further helps the organizations to analyze their strategies better & present the best of the possible options to the customer in the most effective way.

Data :

The dataset used for this project contains information on customers of a shopping mall. It includes the following variables:

- CustomerID: A unique identifier for each customer.
- Gender: The gender of the customer.
- Age : The age of the customer.

- Annual Income(k$): The annual income of the customer is in thousands of dollars.
- Spending Score(1-100): A score assigned by the mall based on the customer's spending behavior.

Data Preprocessing :

First, we import the necessary libraries and load the dataset into a pandas dataframe. We then perform some initial data exploration to understand the structure of the data. We check for missing values, data types, and descriptive statistics.

Next, we perform some data visualization to gain insights into the data. We use histograms to plot the distribution of age, income, and spending score. We also use scatter plots to visualize the relationships between different variables.

Clustering :

We use the K-means clustering algorithm to segment the customers based on their spending score and annual income. We first normalize the data using StandardScaler, then fit the K-means model with a range of cluster numbers from 1 to 10. We use the elbow method to determine the optimal number of clusters. The elbow method suggests that the optimal number of clusters is 5.

Objective of the Product

The objective of Mall customer segmentation is to group customers with similar characteristics and behaviors into different segments. Following are the objectives needs to cater :

- Personalization
- Customer retention
- Cross-selling and upselling
- Risk management
- Resource allocation

Overall, the objective of this project is to perform customer segmentation analysis for a shopping mall. The goal is to identify different groups of customers based on their spending behavior and demographic characteristics. This analysis will help the mall to better understand its customers and develop targeted marketing strategies.

Description of the Project

The project involves analyzing a dataset containing information on customers of a shopping mall. The dataset includes variables such as age, gender, annual income, and spending score. The data will be preprocessed, visualized, and then clustered using the K-means clustering algorithm. The optimal number of clusters will be determined using the elbow method. The results will be visualized using scatterplots with different colors for each cluster. The characteristics of each cluster will be analyzed to gain insights into customer behavior.

We will be using NUMPY, PANDAS, MATPLOTLIB, SKLEARN libraries to properly segment data into distinct attributes.

Source Code

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import plotly.express as px
import plotly.graph_objects as go

#Read CSV with Pandas
customer_df = pd.read_csv("Mall_Customers.csv")
customer_df.head()
# shape of the dataset
print(customer_df.shape)
# my personal reusable function for detecting missing data
```

```
def missing_value_describe(data):
    # check missing values in the data
    missing_value_stats = (data.isnull().sum() / len(data)*100)
    missing_value_col_count = sum(missing_value_stats > 0)
    missing_value_stats =
missing_value_stats.sort_values(ascending=False)[:missing_value_col_count]
    print("Number of columns with missing values:", missing_value_col_count)
    if missing_value_col_count != 0:
        # print out column names with missing value percentage
        print("\nMissing percentage (desceding):")
        print(missing_value_stats)
    else:
        print("No missing data!!!")
missing_value_describe(customer_df)
px.histogram(customer_df, y="Spending Score (1-100)",
             marginal="rug", title="Spending score distribution")
customer_df.Gender.value_counts()
px.box(customer_df, x="Spending Score (1-100)", y="Gender",
          color="Gender", points='all',
          title="Distribution of spending score by gender")
# male customer spending score statistic summary
customer_df.query("Gender == 'Male'")["Spending Score (1-100)"].describe()
# female customer spending score statistic summary
customer_df.query("Gender == 'Female'")["Spending Score (1-100)"].describe()
px.histogram(customer_df, x="Age", title="Customer age distribution", nbins=10)
customer_df.Age.describe()
# create a new column of age group with increment of 10 units
def bin_age(age):
    if age < 20: return "10-19"
```

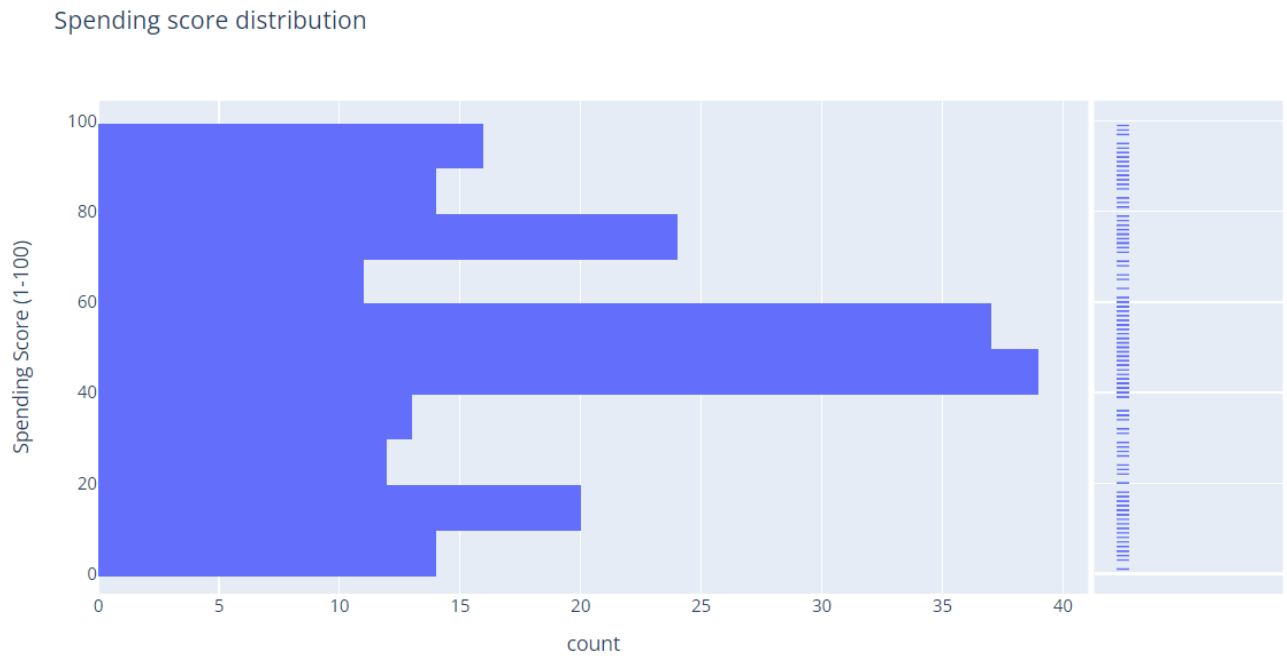
```
    elif age < 30: return "20-29"
    elif age < 40: return "30-39"
    elif age < 50: return "40-49"
    elif age < 60: return "50-49"
    elif age < 70: return "60-69"
    else: return "70-79"
customer_df["age_group"] = customer_df.Age.apply(bin_age)
customer_df["age_group"].unique()
age_group_order = ['10-19', '20-29', '30-39', '40-49', '50-49', '60-69', '70-79']
px.box(customer_df, y="Spending Score (1-100)", x="age_group",
       title="Distribution of spending score by age group and gender group",
       color="Gender",
       category_orders={"age_group": age_group_order},
       facet_row="Gender")
# What are the annual income distributions of different age groups?
age_group_order = ['10-19', '20-29', '30-39', '40-49', '50-49', '60-69', '70-79']
px.box(customer_df, y="Annual Income (k$)", x="age_group",
       title="Distribution of annual income by age group and gender group",
       color="Gender",
       category_orders={"age_group": age_group_order},
       facet_row="Gender")
customer_df.drop(["CustomerID"], axis=1).corr()
px.scatter(customer_df,
       x="Annual Income (k$)", y="Spending Score (1-100)",
       color="Gender",
       hover_name="Spending Score (1-100)",
       title="Non-linear relationship between the annual income and spending score")
px.scatter(customer_df,
       x="Age", y="Spending Score (1-100)",
```

```
    color="Gender",
    hover_name="Spending Score (1-100)",
    title="Non-linear relationship between the age and spending score")
# K-mean modeling for supermarket member segmentation
from sklearn.cluster import KMeans
X = customer_df[["Age", "Annual Income (k$)", "Spending Score (1-100)"]]
model = KMeans()
# fit kmean model with k=4
kmeans = KMeans(n_clusters=4, random_state=2)
kmeans.fit(X)
X["cluster"] = kmeans.predict(X)
X.head()
# cluster centers
kmeans.cluster_centers_
#Visualize the customer segmentations
px.scatter_3d(X, x="Annual Income (k$)", y="Spending Score (1-100)", z="Age",
                color = 'cluster', title="Supermarket member segmentation")
#Let me break it down to 2D plots for further analysis.
px.box(X, x="cluster", y="Annual Income (k$)",points='all', color="cluster",
       title="Distribution of annual income by cluster")
px.box(X, x="cluster", y="Age",points='all', color="cluster",
       title="Distribution of customer age by cluster")

px.box(X, x="cluster", y="Spending Score (1-100)",points='all', color="cluster",
       title="Distribution of spending score by cluster")
# number of customers in each cluster
pd.DataFrame(X.value_counts("cluster").reset_index(drop=True), columns=["customer count"])
```

Graph Analysis

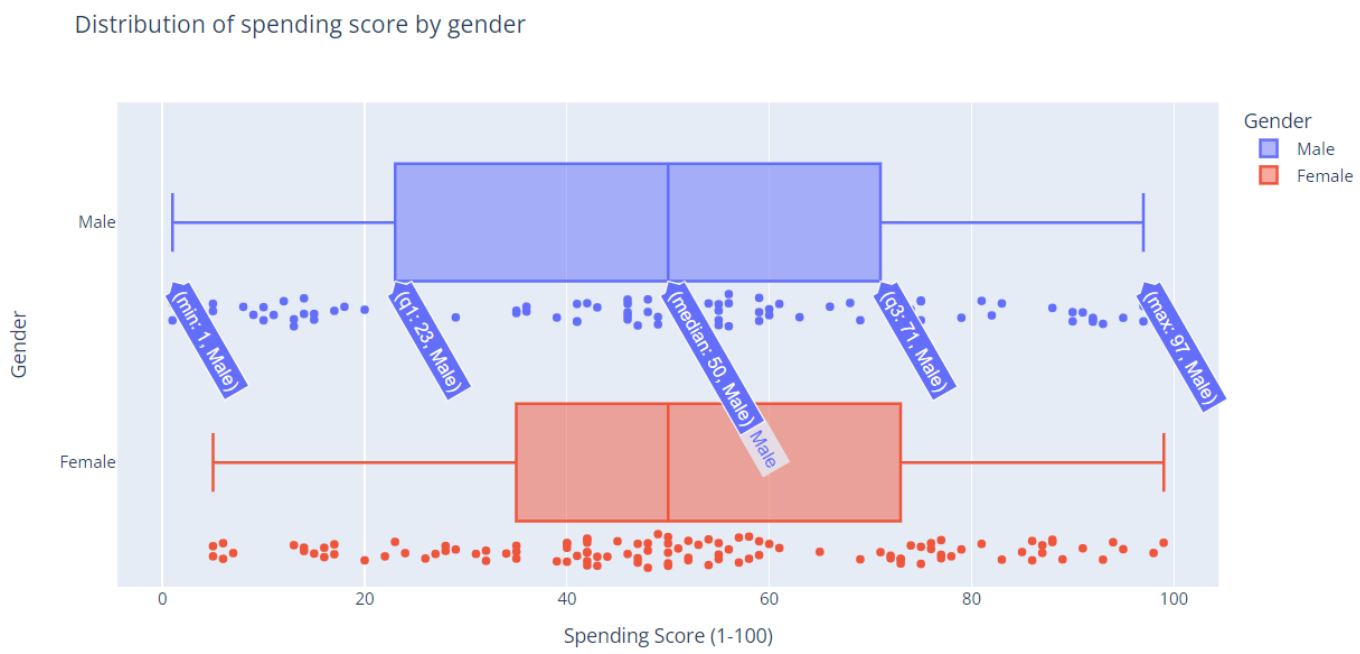
Plot A : Spending Score Distribution



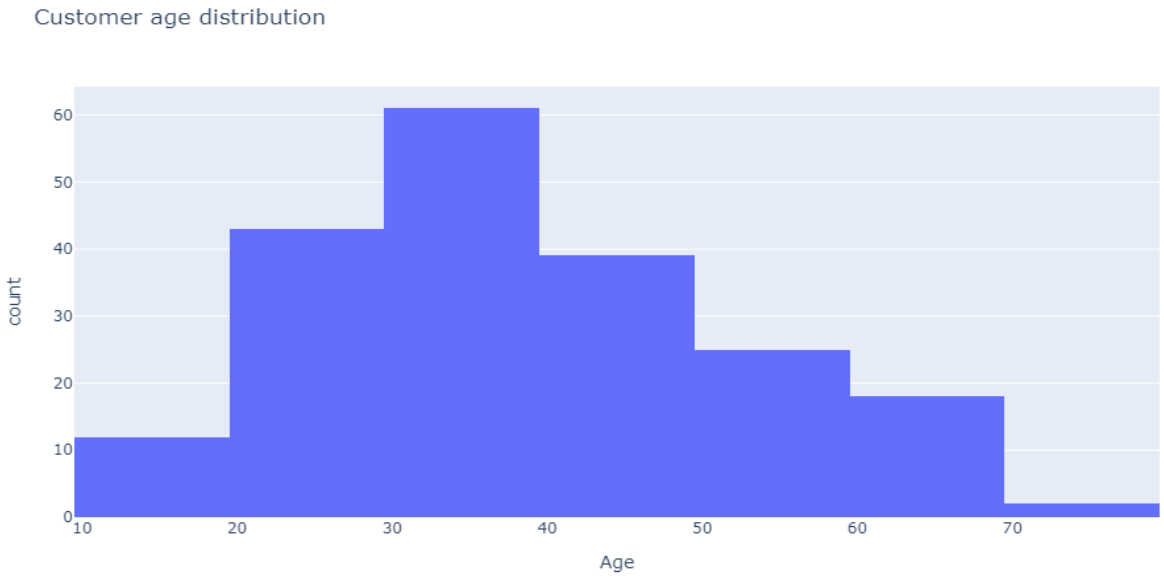
From the spending score distribution histogram binned by increment of 10, we can see there are 3 spikes around spending score of 10-19, 40-49, and 70-79 with max count of 39 customers fall in to the spending score range of 40-49. The data is centered around the spending score range of 40-49.

Plot B : Distribution of spending score by gender

From the visualizations, we can observe that there is no clear distinction between the medians and the top quantiles of the customer spending scores compared for the female and male customers. The 1st quantile of the male spending score is 10.5 unit lower than the 1st quantile of the female spending score. The max (99) of the spending scores for female is higher than the max (97) of the male spending scores. The min (5) of the spending scores for female is higher than the min (1) of the male spending scores.

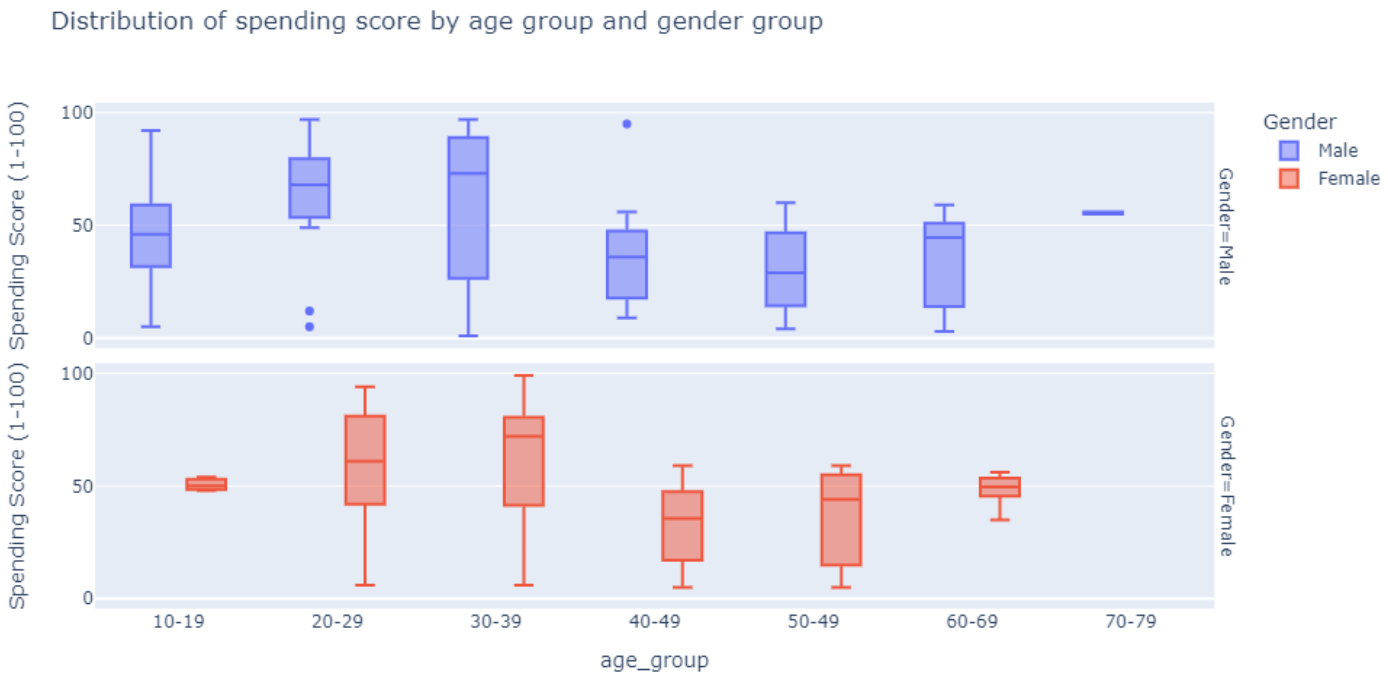


Plot C : Customer Age distribution



The distribution of the customer age data is relatively normal centered around the age range of 30-39 with total count of 61 customers.

Plot D：Distribution of spending score by age group and gender group



- Most of the customers are in the 20-40 age group.
- Spending score is high for the customers in the age group of 20-40.
- The distributions of the two gender groups share a similar pattern for the two age groups of 20-29 and 30-39.

Plot E : Non-Linear Relationships (Scatter Graphs)

Scope of the Project

The scope of this project includes collecting customer data such as age, gender, annual income, and spending behavior. Then, different types of analysis will be performed on the data to identify different customer segments. The analysis will include spending score distribution, distribution of spending score by gender, customer age distribution, distribution of spending score by age group and gender group, non-linear relationship between annual income and spending score, non-linear relationship between age and spending score, and supermarket member segmentation. These analyses will help to identify different types of customer segments and their spending behavior.

Development

The project was divided into several phases, including data collection and preprocessing, exploratory data analysis, customer segmentation, and recommendation development.

During the data collection phase, transactional data was collected from the mall's point-of-sale system and preprocessed to ensure data quality and consistency. Exploratory data analysis was conducted to understand the distribution of key variables and identify any data outliers or anomalies.

Customer segmentation was carried out using several machine learning techniques, including k-means clustering, hierarchical clustering, and decision tree analysis. These techniques were used to identify distinct customer segments based on spending behavior, age, gender, and membership status.

Once the customer segments were identified, recommendations were developed for each segment, including marketing and customer engagement strategies to improve retention and drive additional sales.

Conclusion

In conclusion, the Mall Customer Segmentation Project was successfully completed, and different types of customer segments were identified based on their spending behavior, age, gender, and other characteristics. The project provided valuable insights into customer behavior and can be used by mall owners to tailor their marketing strategies and product offerings to meet the needs of each customer segment. By doing so, mall owners can improve customer satisfaction and increase sales, which can ultimately lead to the growth and success of their business.