SUMMER TRAINING PROJECT REPORT

**TITLE OF THE PROJECT – MALL CUSTOMERS SEGMENTATION**

Submitted by :

DIVESH THAKUR

Registration No. : 12214492

DOMAIN : DATA SCIENCE

COURSE : Python , Data Science & Machine Learning Integrated - Hybrid

Under the Guidance of

Mr. ABHISHEK RAJ

Lovely School of Computer Science Enginnering

Lovely Professional University, Phagwara

****

INDEX

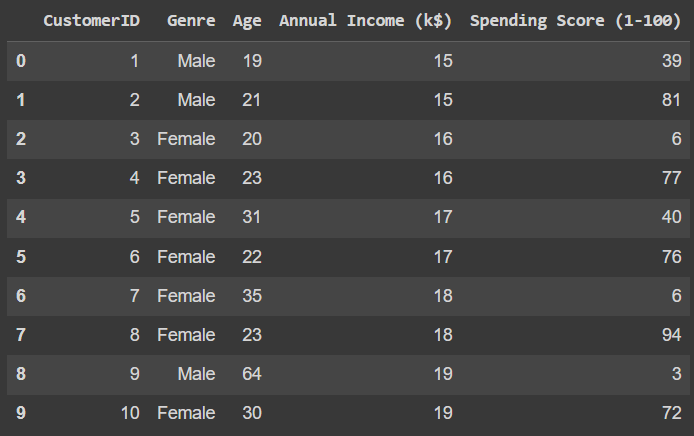
|  |  |
| --- | --- |
| SNO. | TITLE |
| 1. | INTRODUCTION |
| 2. | SOURCE OF DATASET |
| 3. | DATA CLEANING & PREPROCESSING |
| 4. | EDA |
| 5. | DATA VISUALIZATION |
| 6. | VISUALIZING CLUSTERS |
| 7. | CLUSTERS ANALYSIS |
| 8. | SUMMARY |
| 9. | BIBLIOGRAPHY |

INTRODUCTION

**Objective and Use Case**

The objective of this project is to segment customers of a mall based on their demographic characteristics and spending patterns. This segmentation will help the mall management understand different customer groups, allowing for tailored marketing strategies, personalized customer experiences, and optimized product offerings. The use case is to identify distinct clusters of customers who share similar traits, which can be leveraged to improve customer satisfaction and drive sales growth.

**Overview of the Dataset**

****

The dataset consists of information about 200 customers, including:

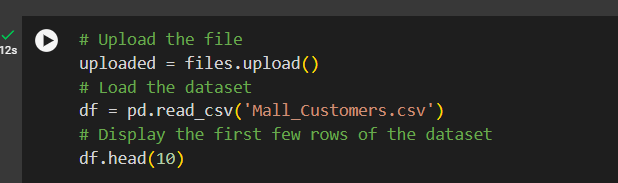
* **CustomerID**: A unique identifier for each customer.
* **Gender**: The gender of the customer (Male/Female).
* **Age**: The age of the customer.
* **Annual Income**: The annual income of the customer in thousands.
* **Spending Score**: A score assigned by the mall based on customer behavior and spending nature (0-100).

SOURCE OF DATASET

**SOURCE** : Mall\_Customers.csv dataset is taken from [www.kaggle.com](http://www.kaggle.com)

**Importing the Dataset**

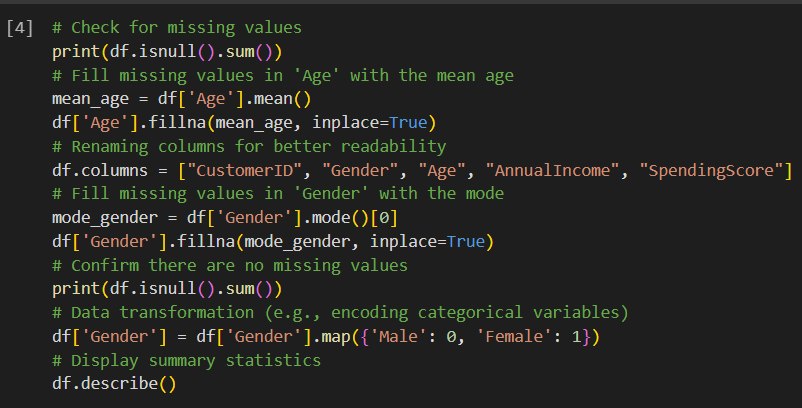
The dataset Mall\_Customers.csv was imported into Google Colab for analysis using the pandas library.



DATA CLEANING & PREPROCESSING

**Handling Missing Values & Transformation**

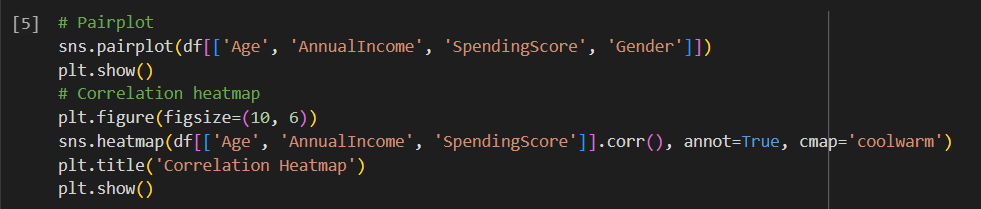
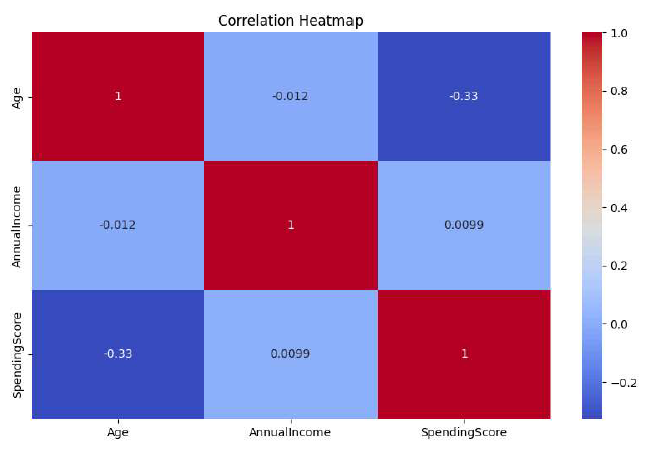
The dataset was checked for missing values. Any missing values in the 'Age' column were filled with the mean age to prevent bias from removing rows. The 'Gender' column was transformed into numerical values (0 for Male and 1 for Female) to facilitate clustering.

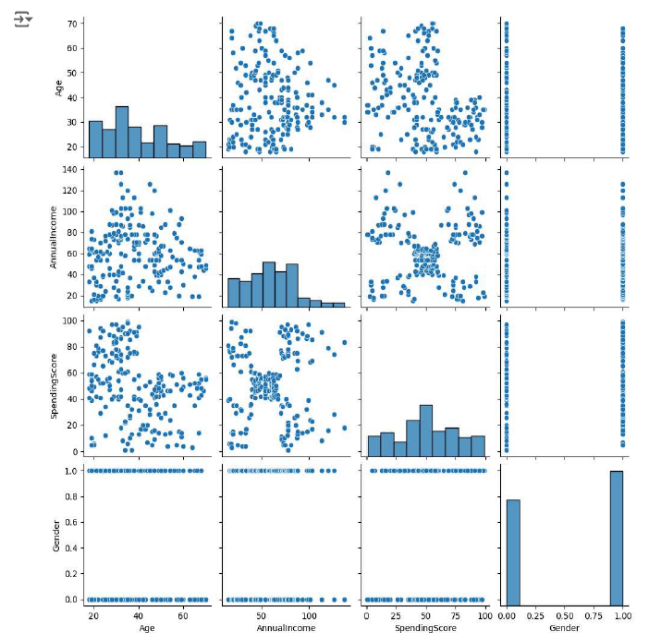
****

EXPLORATORY DATA ANALYSIS (EDA)

EDA involves various techniques and tools to visualize and summarize the data, providing insights that inform the subsequent steps of data analysis, modeling, and decision-making.

Key Components :

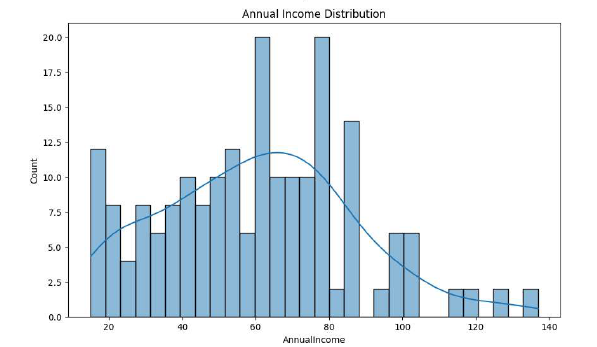
* Descriptive Statistics
* Data Visualization
* Data Cleaning
* Various Analysis



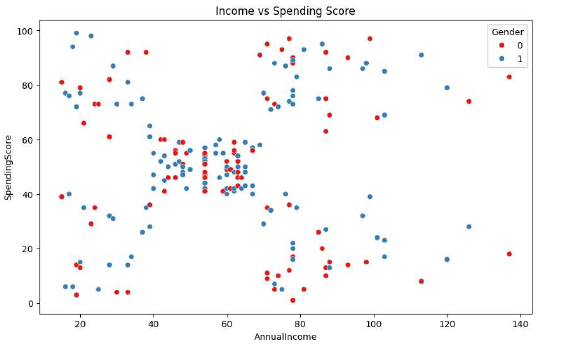
DATA VISUALIZATION

Data visualization is the process of representing data through graphical and visual formats to help understand patterns, trends, and insights more easily. In the context of this mall customer segmentation project, data visualization serves several important purposes.

* **ANNUAL INCOME DISTRIBUTION**

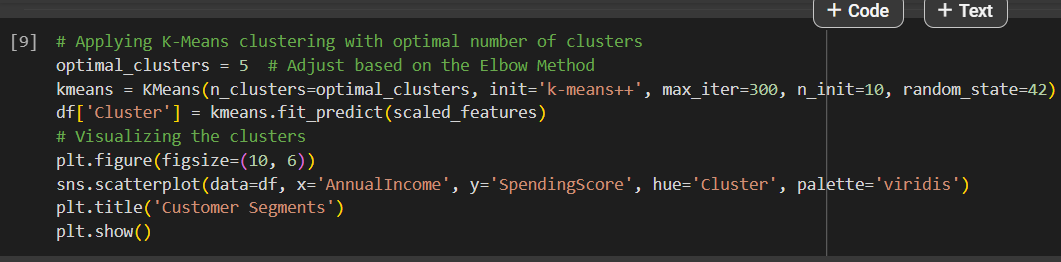
****

* **INCOME VS SPENDING SCORE**

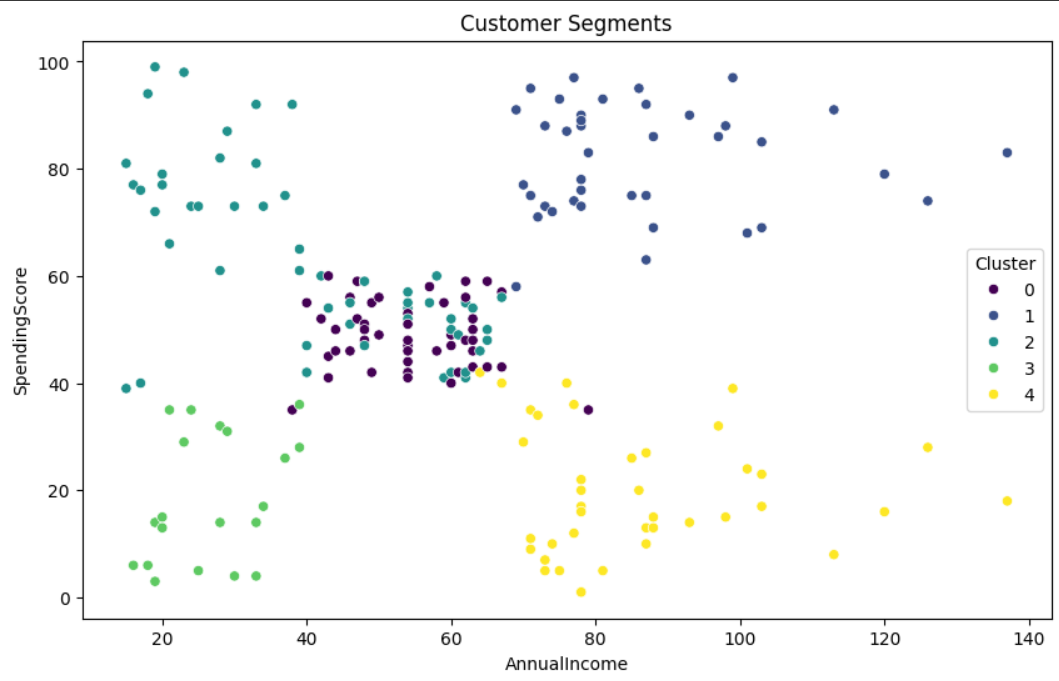
****

VISUALIZING CLUSTERS

Clusters were visualized using scatter plots to show the relationship between annual income, spending score, and cluster assignment.

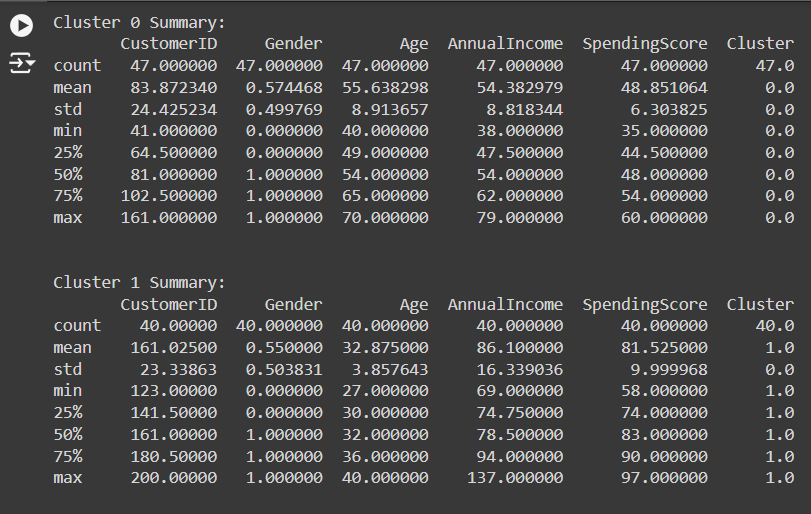


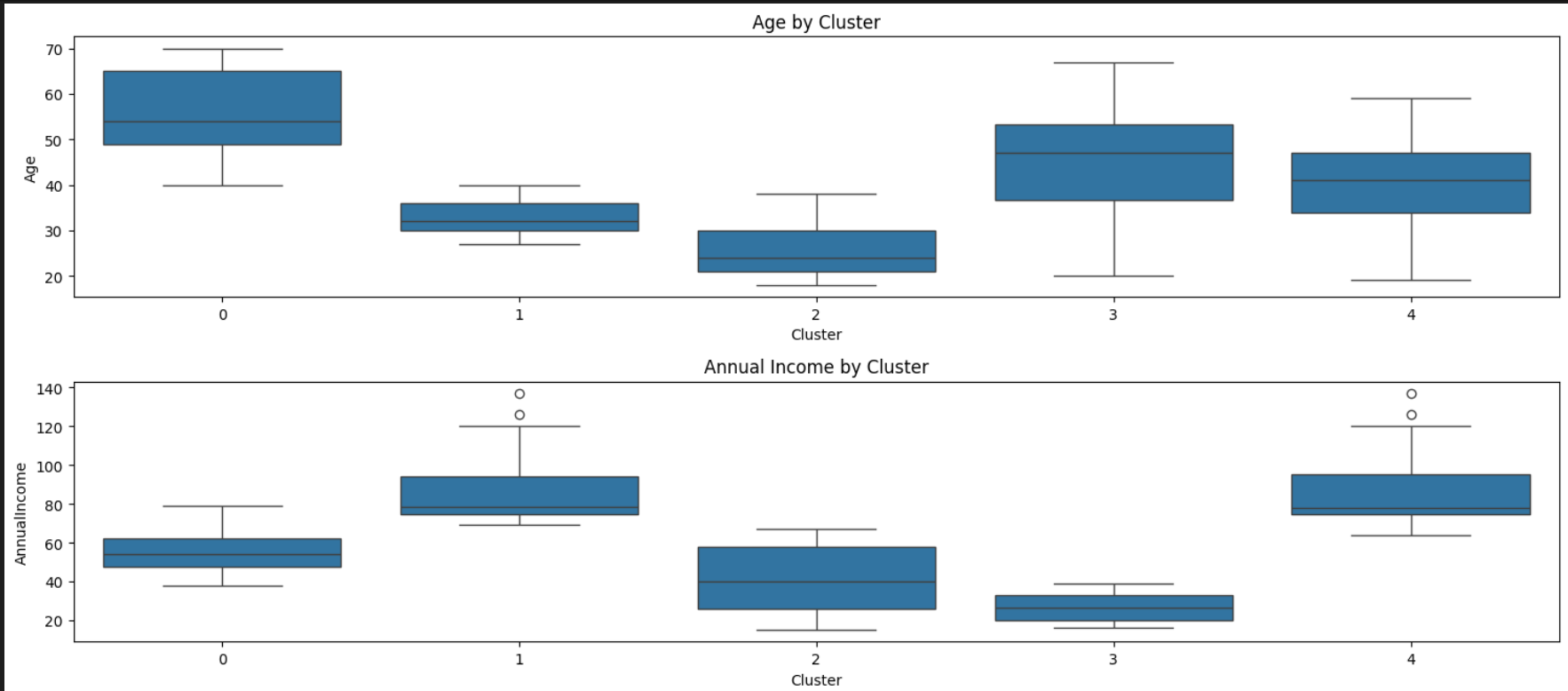
* **CUSTOMER SEGMENTS SCATTER PLOT**



CLUSTERS ANALYSIS

Box plots were used to analyze the distribution of age, annual income, and spending score within each cluster.



* **BOX PLOT OF AGE & INCOME**

SUMMARY

* Customers can be segmented into distinct groups based on their age, income, and spending behavior.
* Each cluster shows unique characteristics that can be targeted with specific marketing strategies.
* High-spending young customers were identified as a key segment.
* There is a diverse range of annual incomes and spending scores.

BIBLIOGRAPHY

* [www.kaggle.com](http://www.kaggle.com)
* www.chatgpt.com
* [www.goggle.com](http://www.goggle.com)
* [www.cipherschools.com](http://www.cipherschools.com)
* Lecture PPTs