JHC ANALYTICS

Shopping Customer Segmentation

USING KMEANS CLUSTERING TO IDENTIFY KEY CUSTOMER SEGMENTS

BY: JOSEPH CHOI

Table of Contents

- Introduction
- Exploratory Data Analysis (EDA)
- Model Development (KMeans Clustering)
- Key Findings
- Marketing Strategy Recommendations

Introduction

Problem Statement: The marketing team needs to better understand the target customers to develop effective marketing strategies. Identifying distinct customer groups based on demographics and purchasing behavior is crucial for developing marketing efforts and optimizing customer engagement.

Project Objective: Perform customer segmentation using an unsupervised machine learning technique, KMeans clustering, to divide the customer base into meaningful segments based on demographic and behavioral characteristics. Identify target groups for marketing to improve strategic decisions and customer satisfaction/retention.

About the Dataset: The dataset consists of entries representing shopping customers at a mall. It contains the following features: Customer ID, Gender, Age, Annual Income, and Spending Score (1-100) assigned by the mall based on customer spending behavior and purchasing patterns.

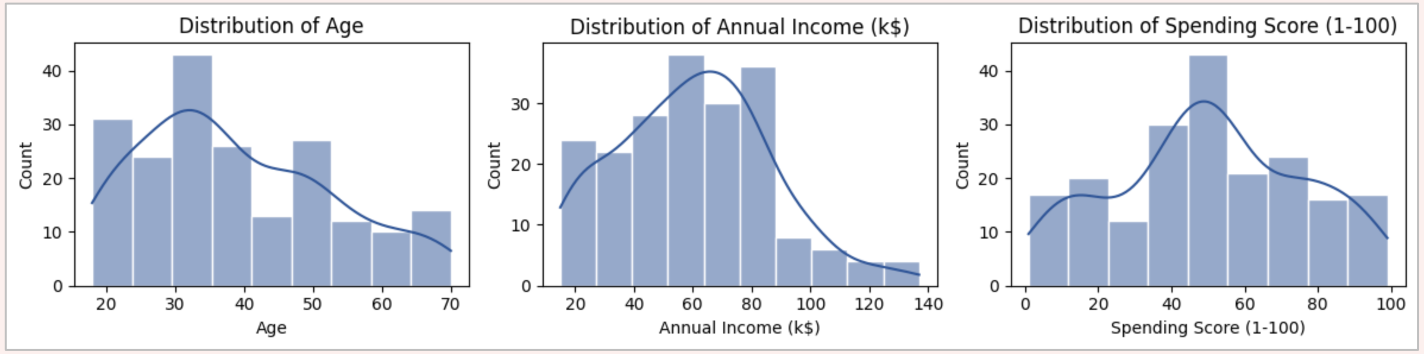
Exploratory Data Analysis

Univariate Analysis

JHC ANALYTICS

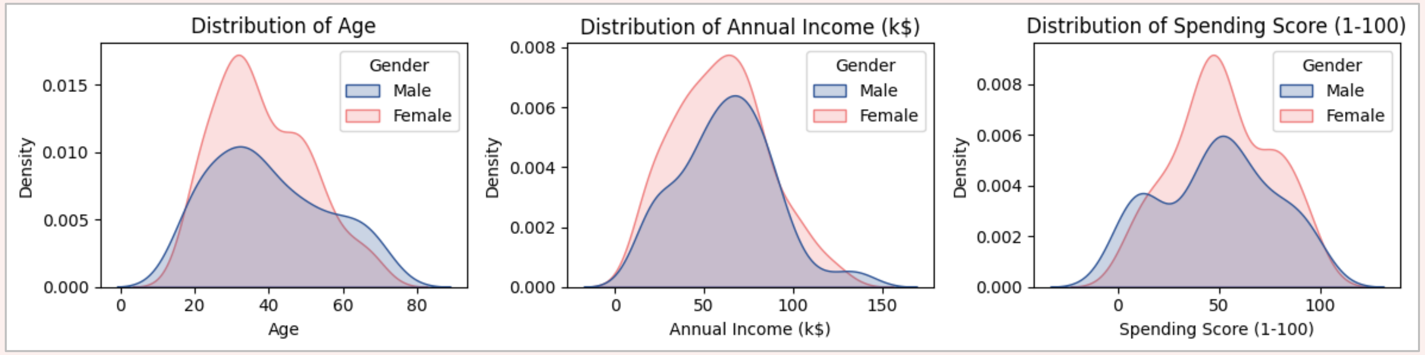
Histograms:

- **Age:** Peaking in the 30s, dropping after 50
- **Annual Income:** Right-skewed, most earn 40k-80k, few above 100k
- **Spending Scores:** Normal distribution; most are moderate, with few extremes



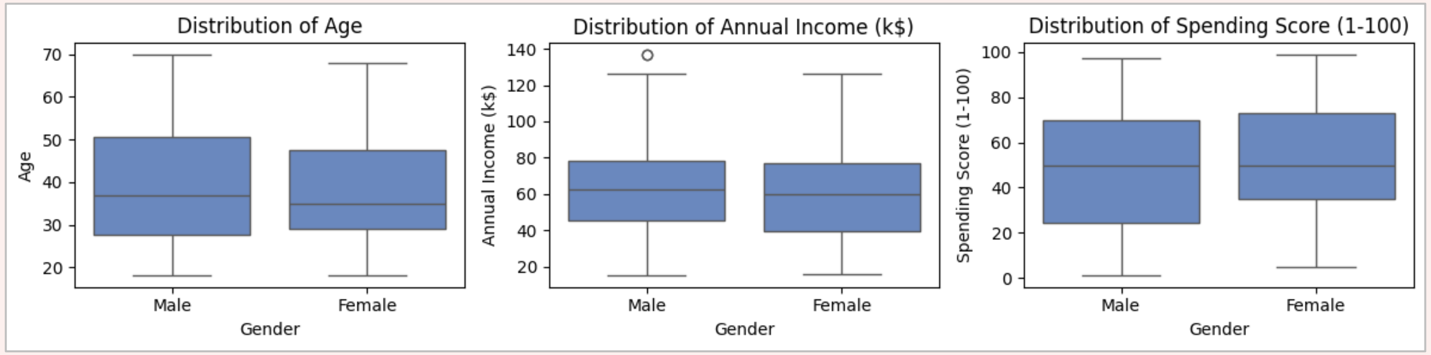
KDE Plots:

- Higher density peaks for females
- Females cluster to specific values more than males
- 56% females, 44% males
- High-income outliers among males



Box Plots:

- **Males:** Slightly higher median age and income, more variability, high-income outliers
- **Females:** Slightly higher median spending scores, more overall spending

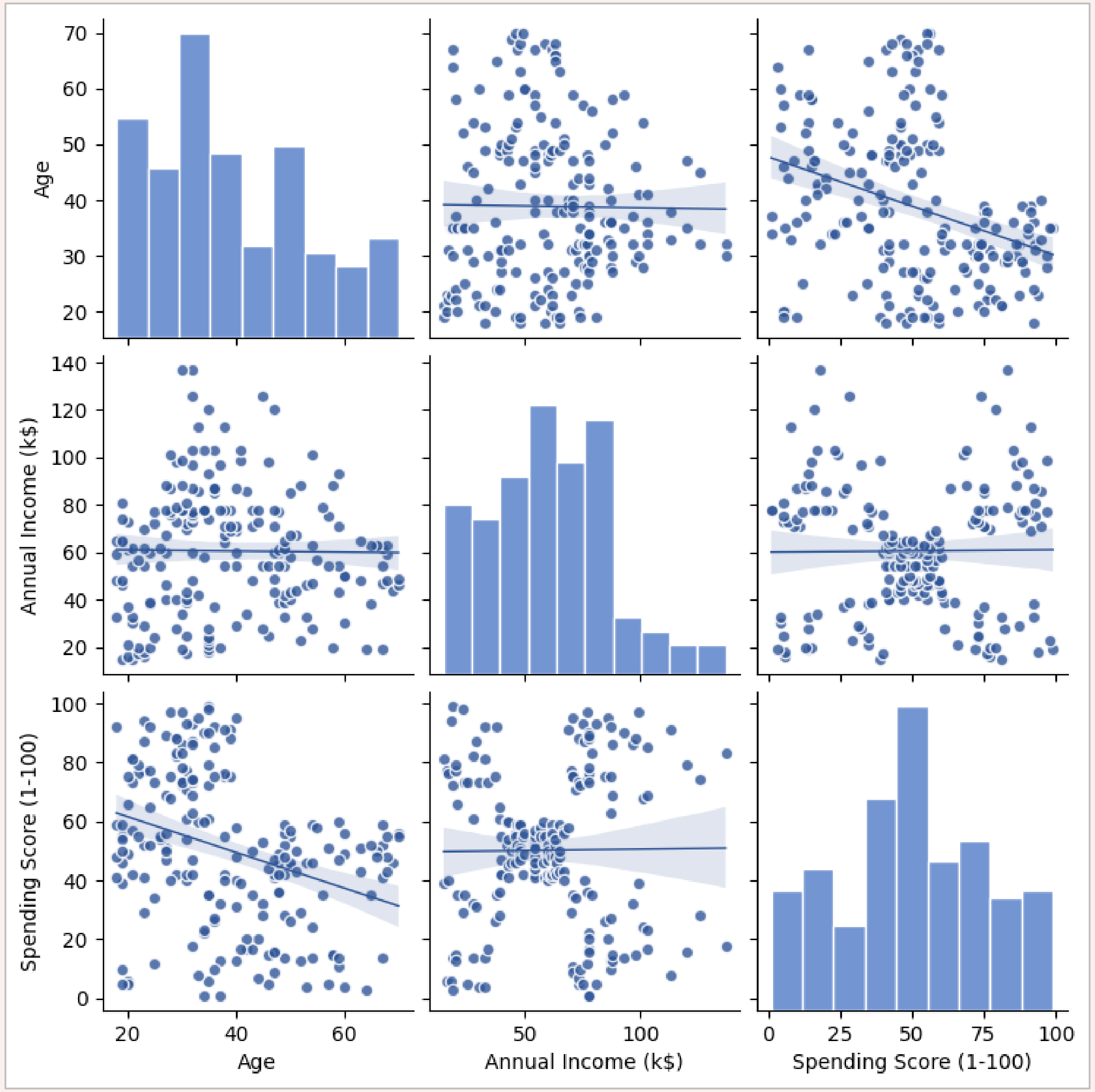


Exploratory Data Analysis

Bivariate Analysis - Part 1

Pair Plot:

- No strong relationship between age and income, suggesting minimal impact of age on annual income
- Negative relationship between age and spending score, indicating customers spend less as they get older
- Noticeable grouping pattern between spending and income, suggesting five distinct clusters
- These groupings indicate natural segments within the data, useful for customer segmentation and target marketing strategies

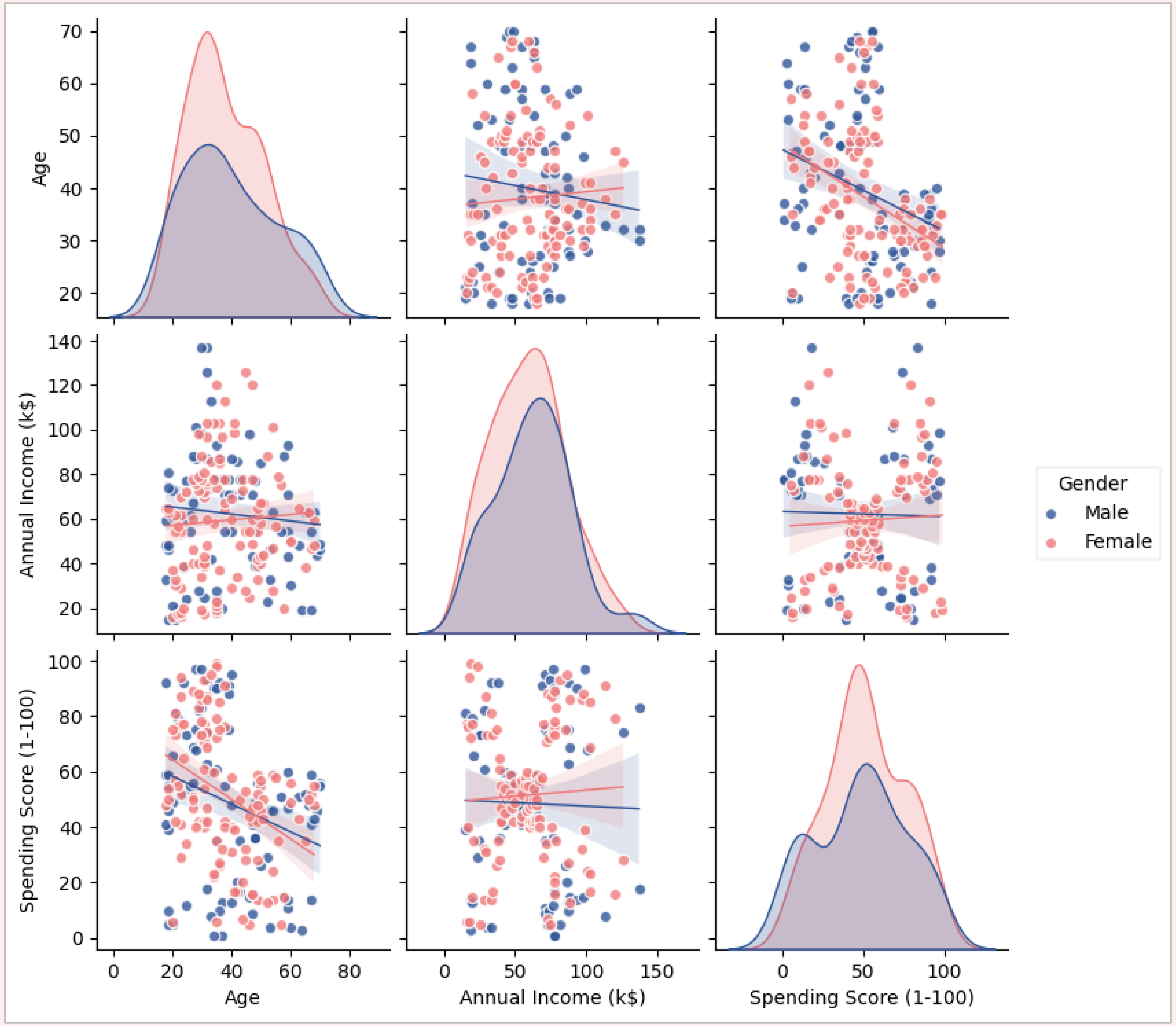


Exploratory Data Analysis

Bivariate Analysis - Part 2

Pair Plot w/ Gender:

- Regression line for males shows a negative slope for age and income, indicating income decreases with age for males
- Regression line for females shows a positive slope for age and income, indicating income slightly increases with age for females
- Both regression lines for age and spending score have a negative slope, with females showing a steeper decline
- Female spending decreases more sharply with age compared to males
- The five segments seen in the previous pair plot are still present for income and spending score
- These segments highlight the need to explore these groupings for better customer segmentation

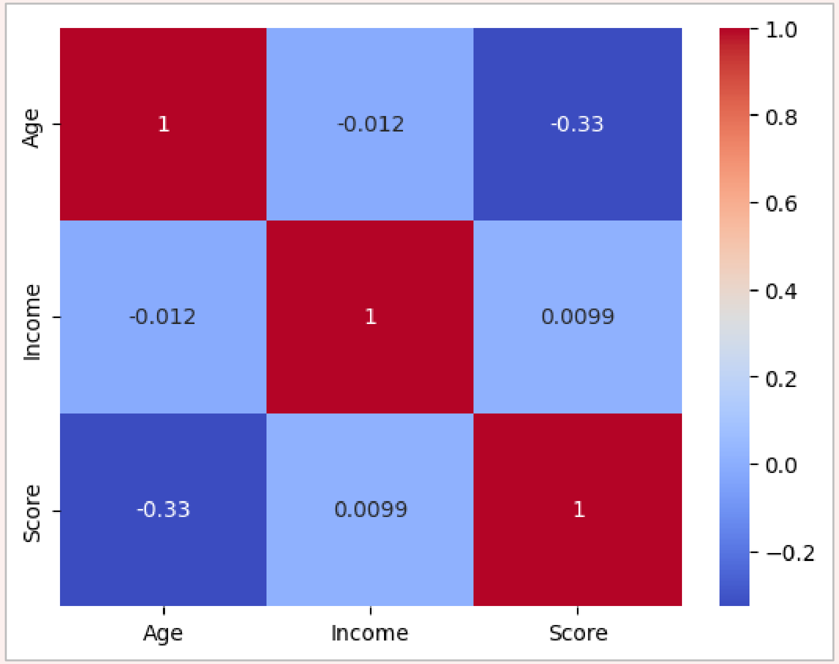


Exploratory Data Analysis

Bivariate Analysis - Part 3

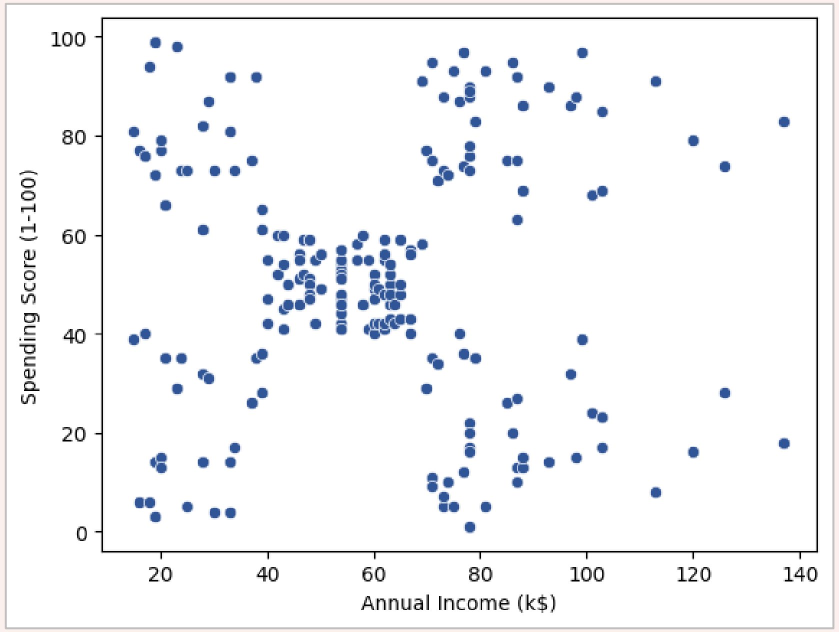
Correlation Heat Map:

- Aligns with trends from the pair plot's regression lines
- Correlation coefficients show relatively low numbers
- Indicates weak relationships between features



Scatter Plot:

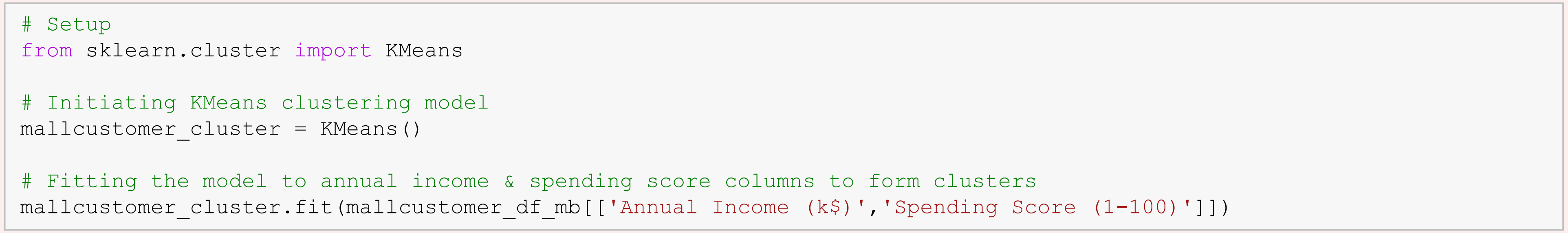
- Plotted to explore the five segments observed in the spending score and annual income plot
- Indicates distinct clusters of customers with varying income and spending behaviors
- Description of each observed segment:
  - **Top Left**: Low Income, High Spending
  - **Top Right**: High Income, High Spending
  - **Middle**: Moderate Income, Moderate Spending
  - **Bottom Left**: Low Income, Low Spending
  - **Bottom Right**: High Income, Low Spending



Model Development Part 1

Initialization and Fitting:

- Introduced and fitted the KMeans clustering model to the dataset
- Utilized features: Annual Income and Spending Score

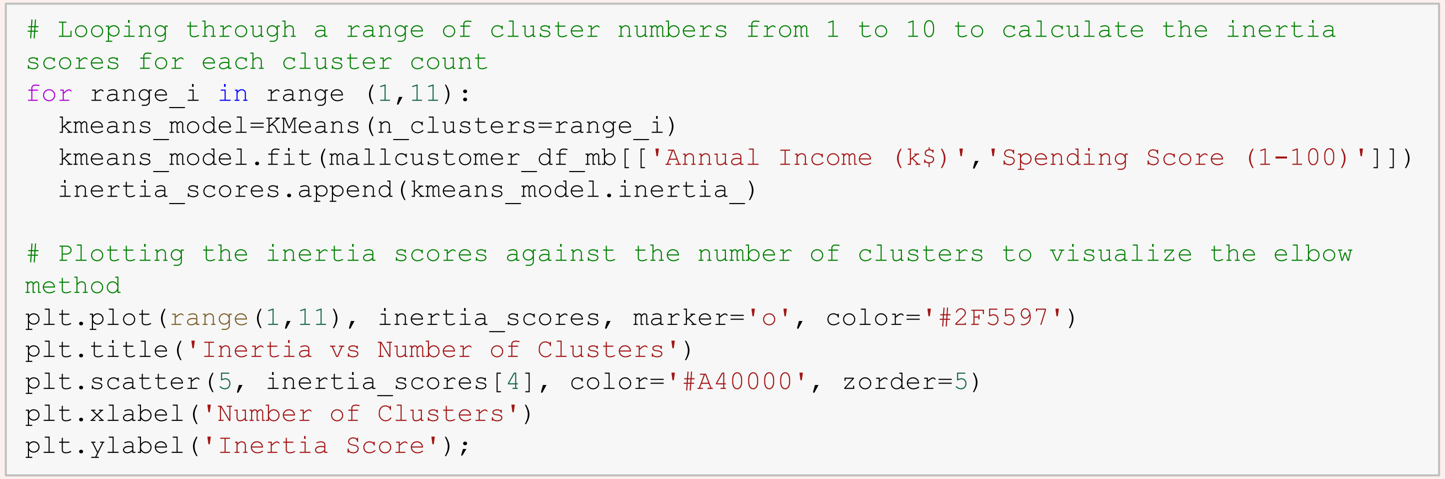


```
# Setup
from sklearn.cluster import KMeans

# Initiating KMeans clustering model
mallcustomer_cluster = KMeans()

# Fitting the model to annual income & spending score columns to form clusters
mallcustomer_cluster.fit(mallcustomer_df_mb[['Annual Income (k$)','Spending Score (1-100)']])
```

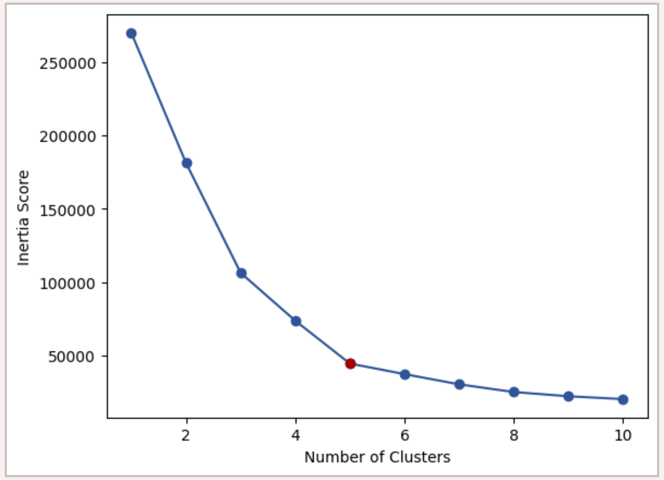
Model Development Part 2

Inertia and Elbow Method:

- Applied the inertia and elbow method to determine the optimal # of clusters
- Selected 5 clusters based on the elbow plot



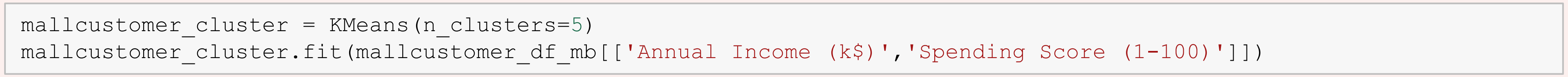
```
# Looping through a range of cluster numbers from 1 to 10 to calculate the inertia
scores for each cluster count
for range_i in range (1,11):
  kmeans_model=KMeans(n_clusters=range_i)
  kmeans_model.fit(mallcustomer_df_mb[['Annual Income (k$)','Spending Score (1-100)']])
  inertia_scores.append(kmeans_model.inertia_)

# Plotting the inertia scores against the number of clusters to visualize the elbow
method
plt.plot(range(1,11), inertia_scores, marker='o', color='#2F5597')
plt.title('Inertia vs Number of Clusters')
plt.scatter(5, inertia_scores[4], color='#A40000', zorder=5)
plt.xlabel('Number of Clusters')
plt.ylabel('Inertia Score');
```



Model Development Part 3

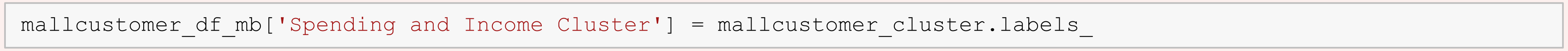
Cluster Labeling:

- Re-initializing and fitting the model with the optimal number of clusters



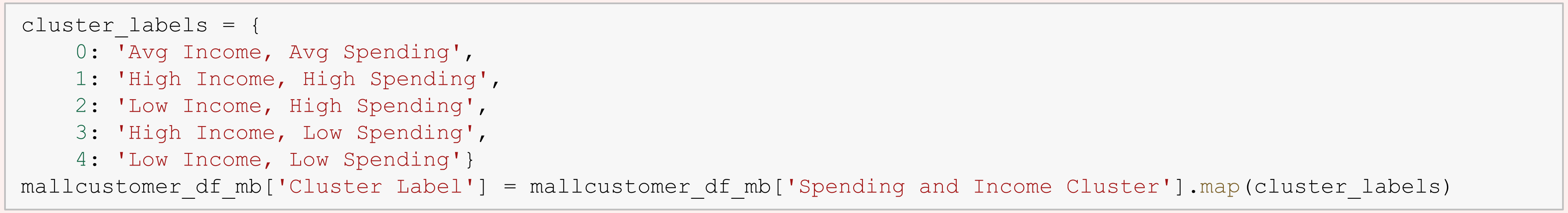
```
mallcustomer_cluster = KMeans(n_clusters=5)
mallcustomer_cluster.fit(mallcustomer_df_mb[['Annual Income (k$)','Spending Score (1-100)']])
```

- Assigned numeric cluster labels



```
mallcustomer_df_mb['Spending and Income Cluster'] = mallcustomer_cluster.labels_
```

- Mapped numeric labels to descriptive cluster names for better interpretation



```
cluster_labels = {
    0: 'Avg Income, Avg Spending',
    1: 'High Income, High Spending',
    2: 'Low Income, High Spending',
    3: 'High Income, Low Spending',
    4: 'Low Income, Low Spending'}
mallcustomer_df_mb['Cluster Label'] = mallcustomer_df_mb['Spending and Income Cluster'].map(cluster_labels)
```

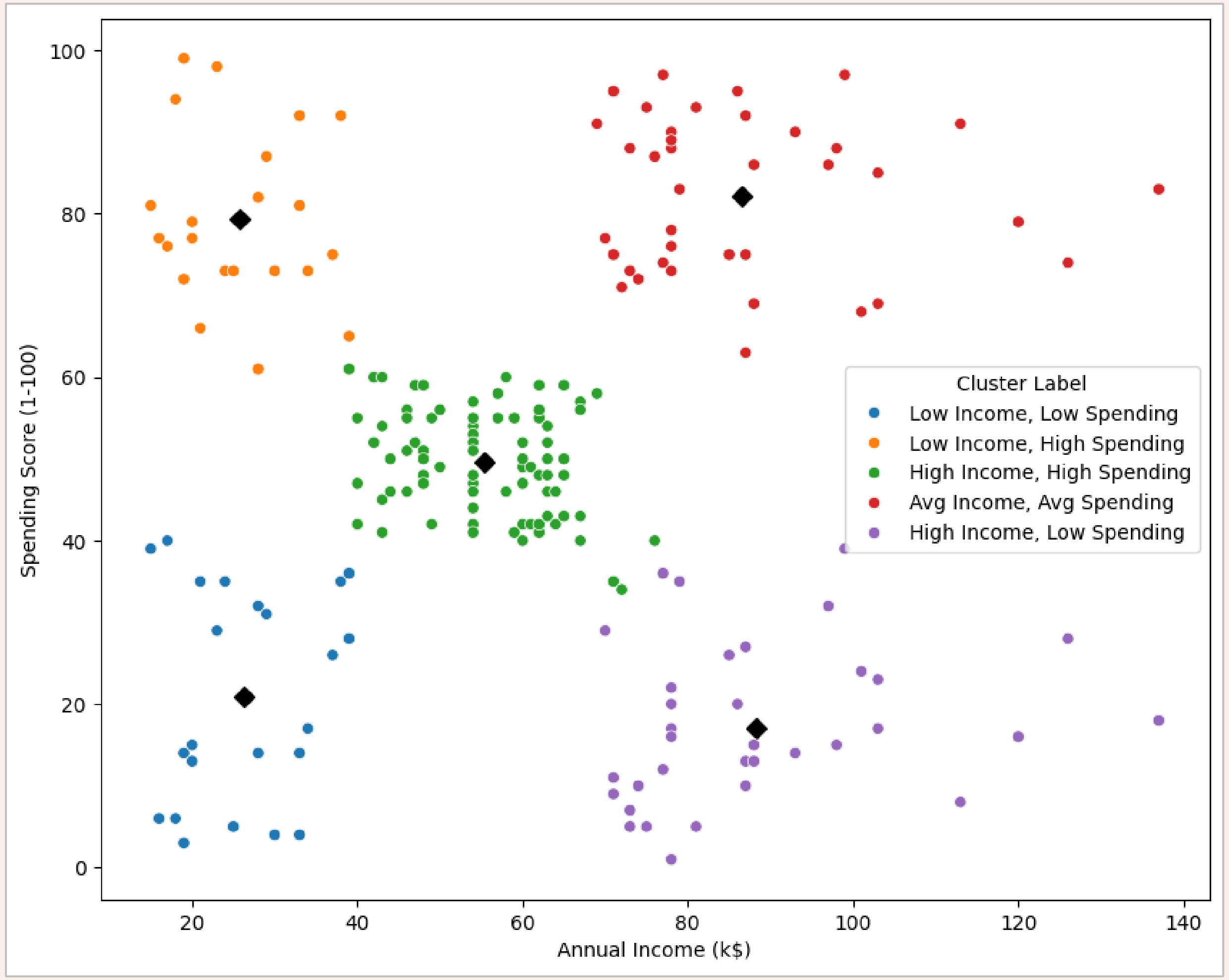
Key Findings

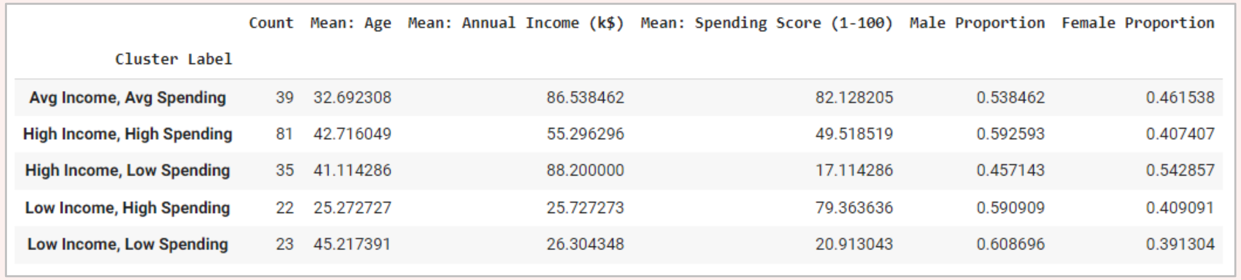
From the KMeans Clustering, I identified five distinct clusters:

- **Avg Income, Avg Spending**: Consists of customers with moderate income and spending habits, forming the largest group
- **High Income, High Spending**: Includes high-income, high-spending customers
- **Low Income, High Spending**: Features customers with low income but high spending scores, suggesting they prioritize shopping despite lower earnings
- **High Income, Low Spending**: Made up of high-income, low-spending customers, possibly indicating a focus on saving
- **Low Income, Low Spending**: Includes those with both low income and low spending scores

Age vs. Spending Scores:

- Younger customers tend to spend more:
  - **Low Income, High Spending** (avg. age 25)
- Older customers tend to spend less:
  - **High Income, Low Spending** (avg. age 41)
  - **Low Income, Low Spending** (avg. age 45)





Cluster Label	Count	Mean: Age	Mean: Annual Income (k$)	Mean: Spending Score (1-100)	Male Proportion	Female Proportion
Avg Income, Avg Spending	39	32.692308	86.538462	82.128205	0.538462	0.461538
High Income, High Spending	81	42.716049	55.296296	49.518519	0.592593	0.407407
High Income, Low Spending	35	41.114286	88.200000	17.114286	0.457143	0.542857
Low Income, High Spending	22	25.272727	25.727273	79.363636	0.590909	0.409091
Low Income, Low Spending	23	45.217391	26.304348	20.913043	0.608696	0.391304

Marketing Strategy Recommendations

To effectively target each customer cluster, employ specific strategies:

- **High Income, High Spending**: Premium loyalty programs and exclusive deals
- **Low Income, High Spending**: Discounts, installment plans, or cashback offers
- **High Income, Low Spending**: Promote savings, investments, and high-end products
- **Low Income, Low Spending**: Value-for-money products and loyalty rewards for frequent small purchases
- **Avg Income, Avg Spending**: Maintain loyalty with consistent value and personalized recommendations
- **Younger Customers**: Trendy, personalized marketing through social media and mobile-friendly shopping