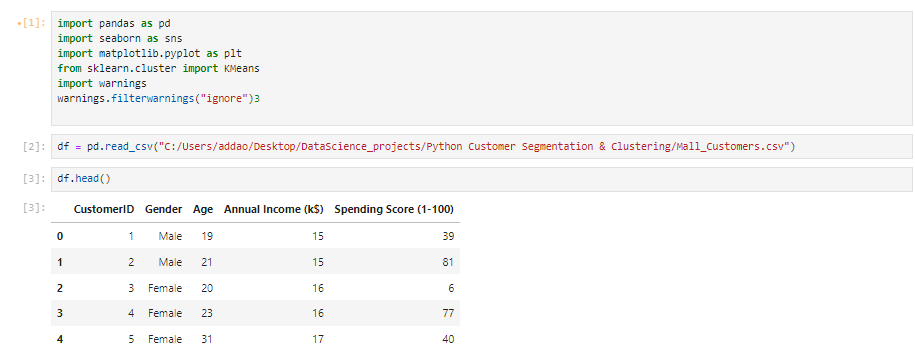
**Customer Segmentation and Clustering Analysis Report**

**Introduction**

This report details the steps and findings from a data analysis project focused on customer segmentation and clustering. The dataset used in this analysis was obtained from a retail shopping mall and contains the following columns: "CustomerID", "Gender", "Age", "Annual Income (k$)", and "Spending Score (1-100)". Various univariate, bivariate, and multivariate analyses were conducted to gain insights into the data and to perform clustering.

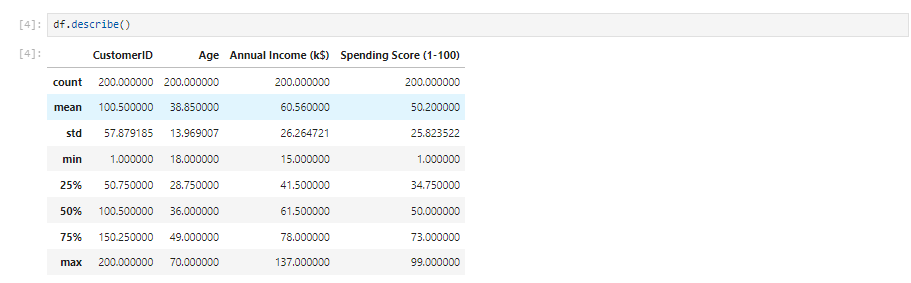
**Data Import and Initial Exploration**

The dataset was imported using the pandas library:



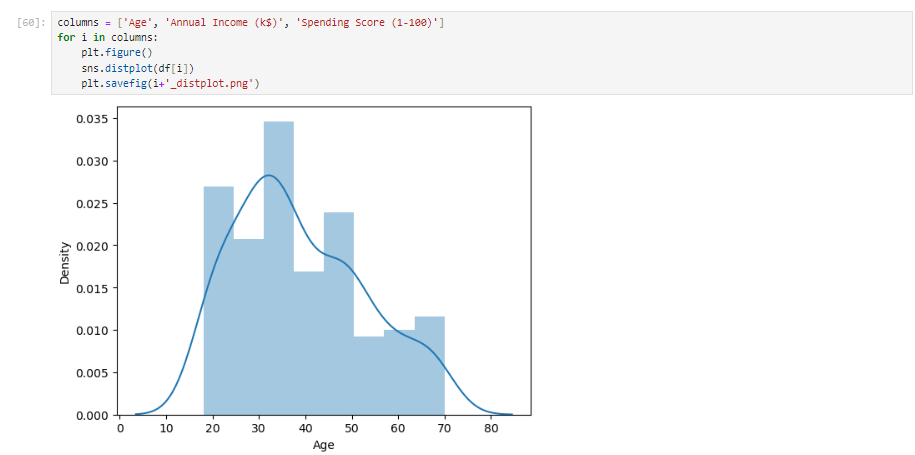
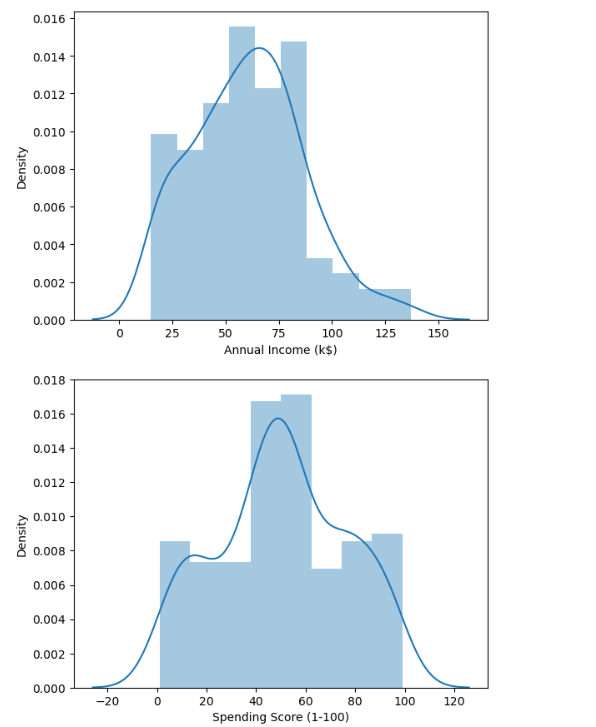
**Descriptive Statistics**

A summary of the dataset's descriptive statistics was obtained using 'df.describe()'



**Distribution Analysis**

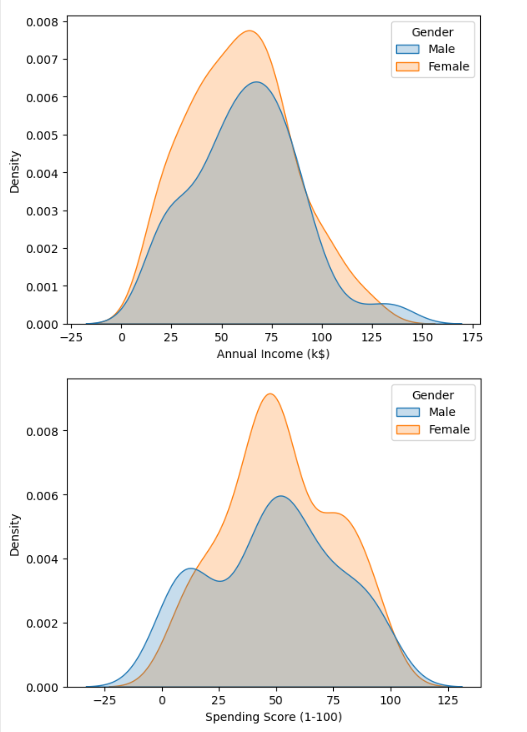
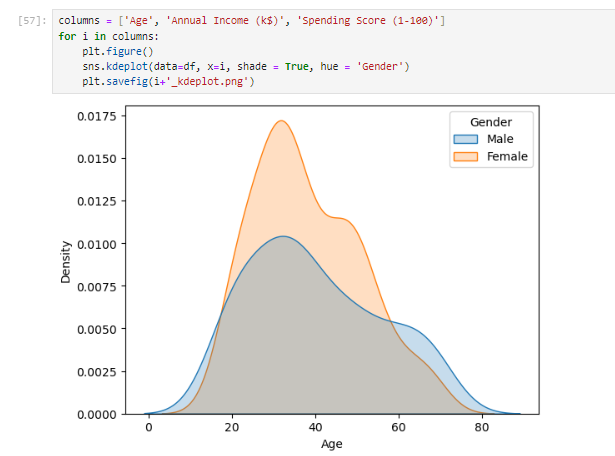
The distributions of "Age", "Annual Income (k$)", and "Spending Score (1-100)" were visualized using seaborn's distplot':

* Age: The majority of customers are between 20 and 48 years old.
* Annual Income: The annual income density peaks between 25k and 80k.
* Spending Score: The spending score shows high density around 40 to 60, with additional peaks around 10 and 90.

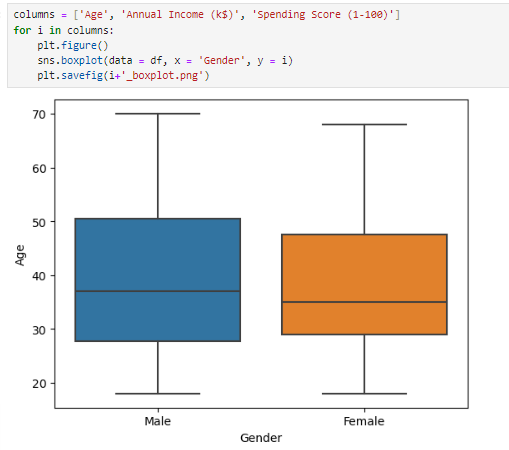
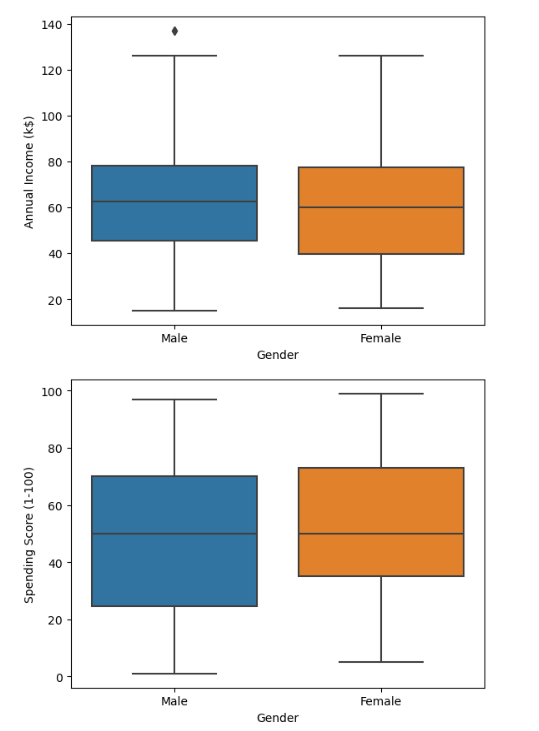
**Kernel Density Estimates by Gender**

Kernel Density Estimates (KDE) were plotted to explore the distribution of variables by gender:



**Boxplots by Gender**

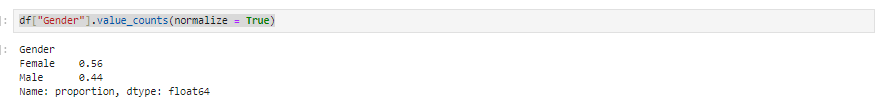
Boxplots were used to compare the distributions of Age, Annual Income, and Spending Score by gender:



* Age: Most men are aged between 28 and 50, while most women are aged between 30 and 47.
* Annual Income: Men predominantly earn between 43k and 80k, with outliers around 140k. Women predominantly earn between 40k and 79k.
* Spending Score: Men typically have a spending score between 23 and 70, while women are between 37 and 72.

**Gender Distribution**

The gender distribution was examined:

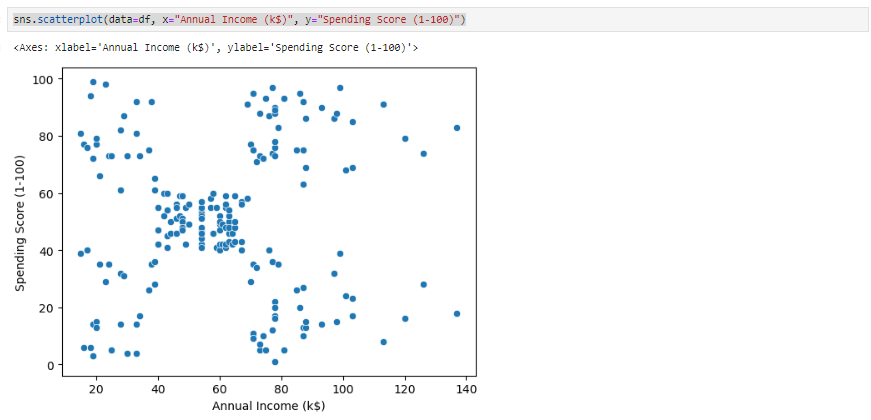


As expected, there are more female customers than male customers in the dataset.

**Bivariate Analysis**

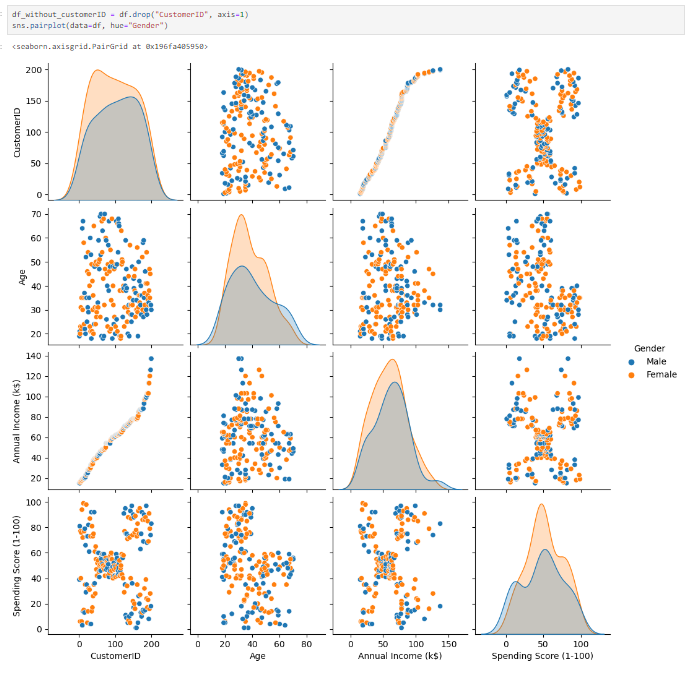
Scatter Plot of Annual Income vs. Spending Score

A scatter plot was created to examine the relationship between "Annual Income (k$)" and "Spending Score (1-100)":



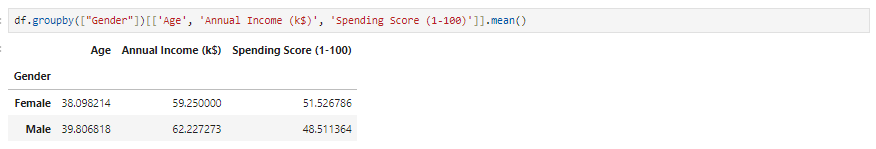
**Pairplot by Gender**

A pairplot was generated to visualize the relationships between numerical variables, colored by gender:



**Mean Values by Gender**

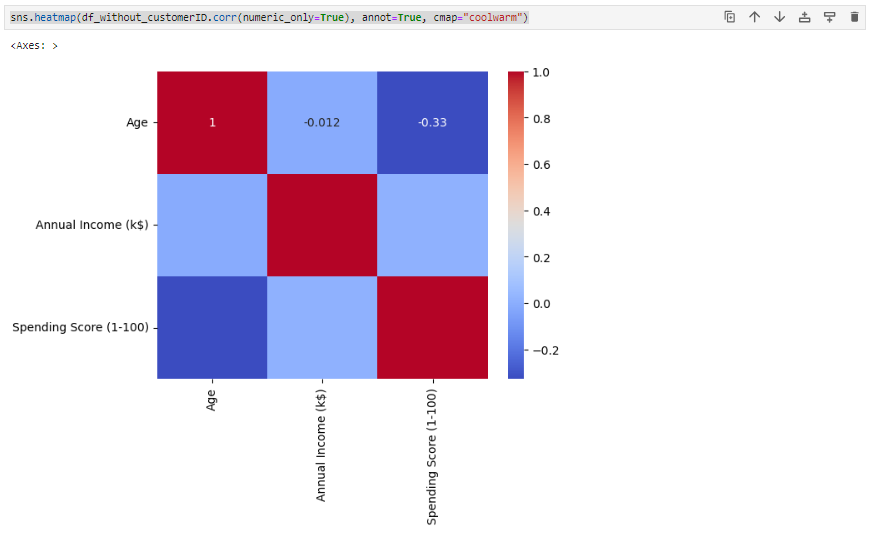
Mean values of "Age", "Annual Income", and "Spending Score" were calculated for each gender:



The average age for both men and women is 38. The average annual income for women is 59k and for men is 62k. The spending score is almost the same for both genders, around 50.

**Correlation Analysis**

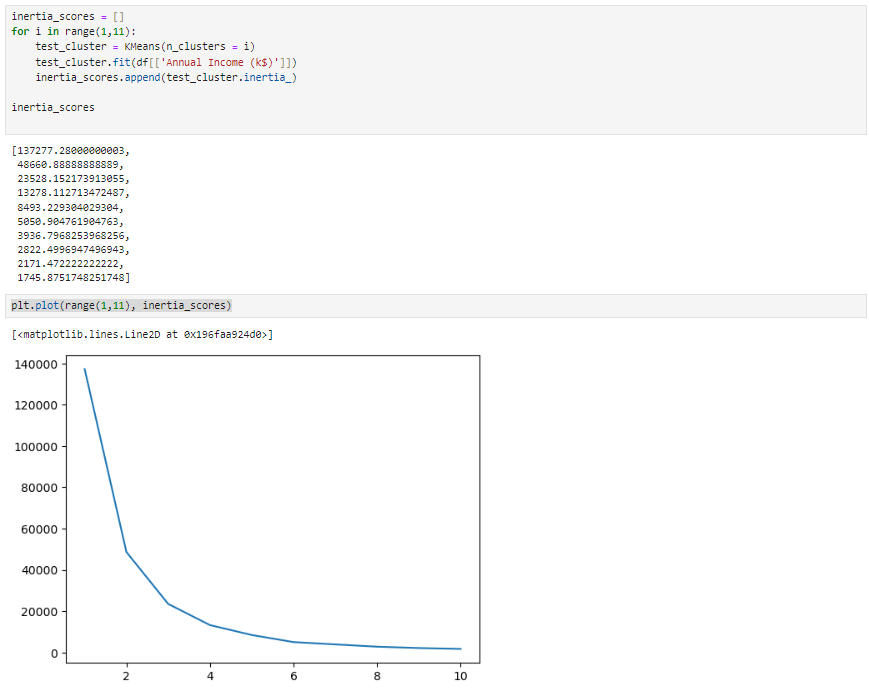
The correlation matrix was computed and visualized using a heatmap:



**Clustering Analysis**

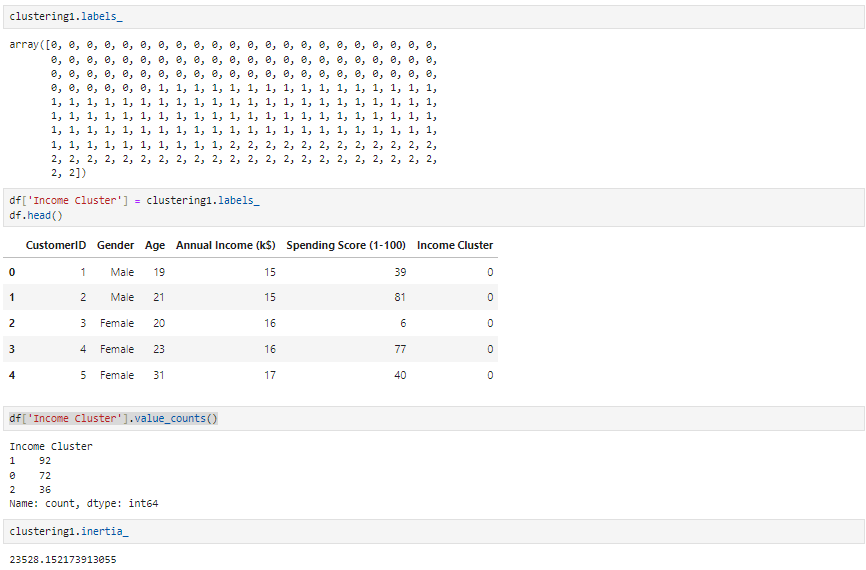
Univariate Clustering

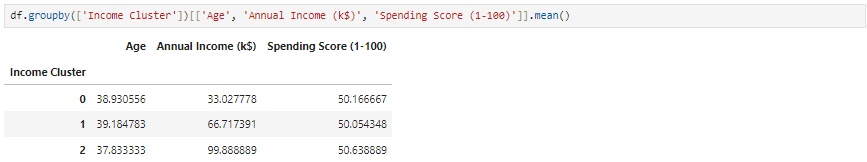
KMeans clustering was performed on "Annual Income (k$)":



Based on the elbow method, 3 clusters were chosen:

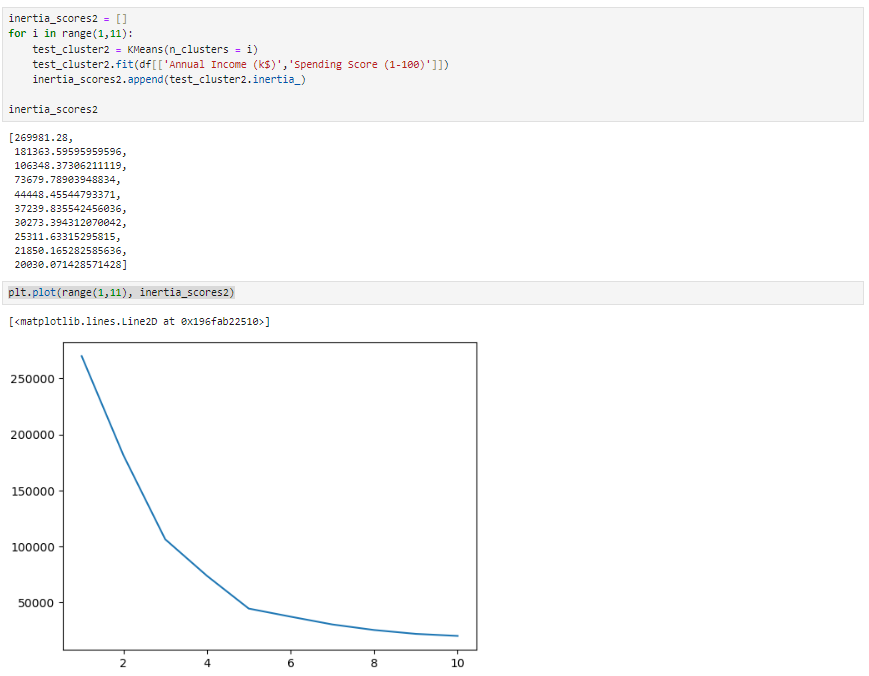




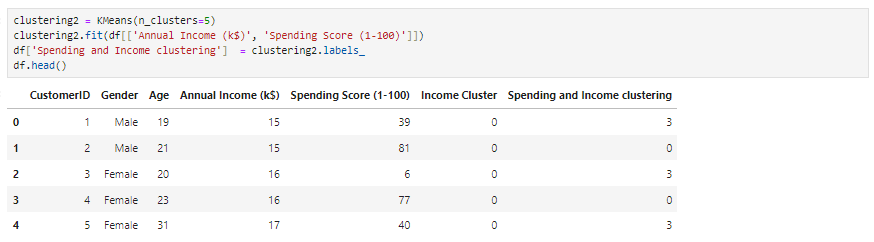


**Bivariate Clustering**

KMeans clustering was performed on "Annual Income (k$)" and "Spending Score (1-100)":

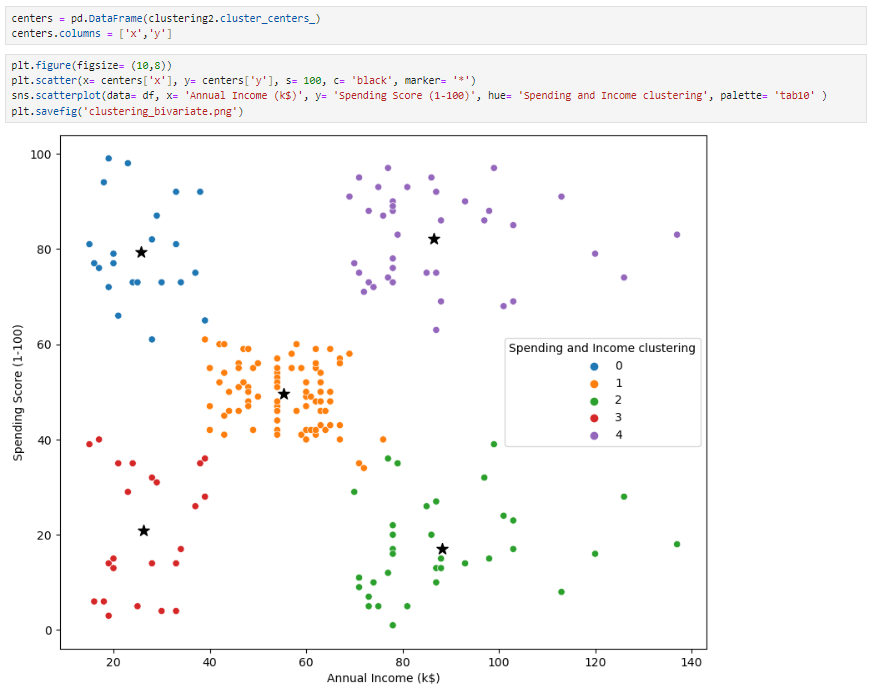


Based on the elbow method, 5 clusters were chosen:



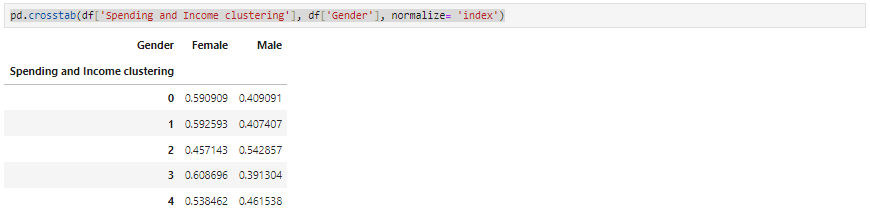
**Cluster Centers and Visualization**

Cluster centers were visualized using a scatter plot:



**Cluster Gender Distribution**

**The gender distribution within each cluster was examined:**

****

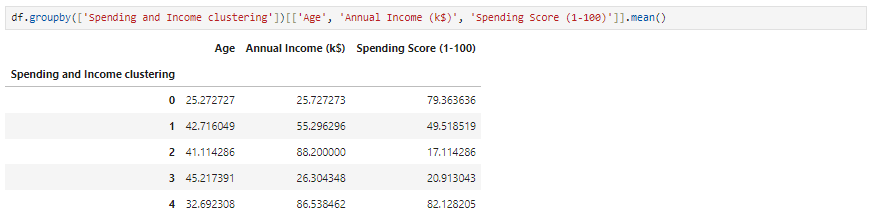
- target group would be cluster 1 which has a high spending score and high income[¶](http://localhost:8888/notebooks/Desktop/DataScience_projects/Python%20Customer%20Segmentation%20&%20Clustering/customer.ipynb#--target-group-would-be-cluster-1-which-has-a-high-spending-score-and-high-income)

- from below we can see that 59% of cluster 1 shoppers are woman. We should look for ways to attract these customers using a marketing campaign targeting popular items in this cluster.

- Cluster 2 presents an interesting opportunity to market to the customers for sales event on popular items.

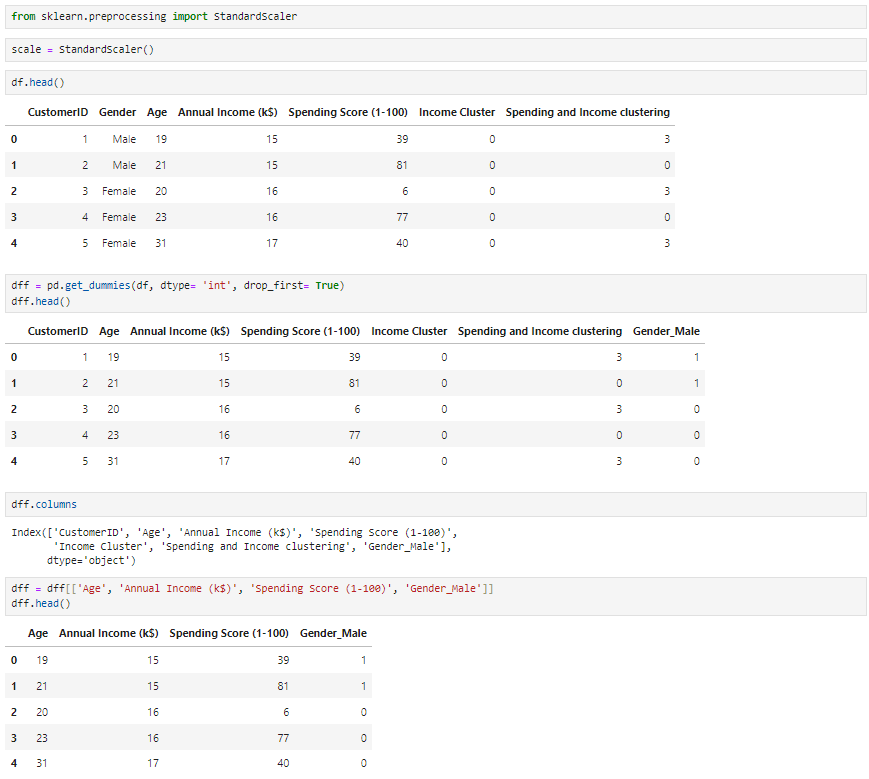
**Cluster Characteristics**

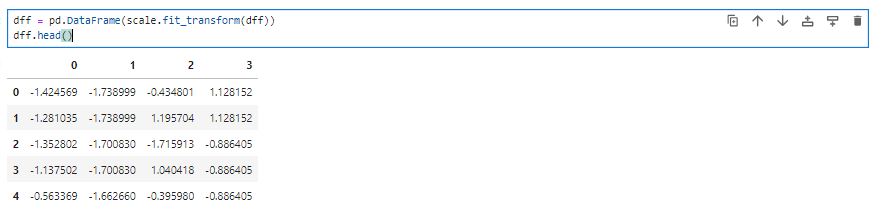
The characteristics of each cluster were analyzed:



**Multivariate Clustering**

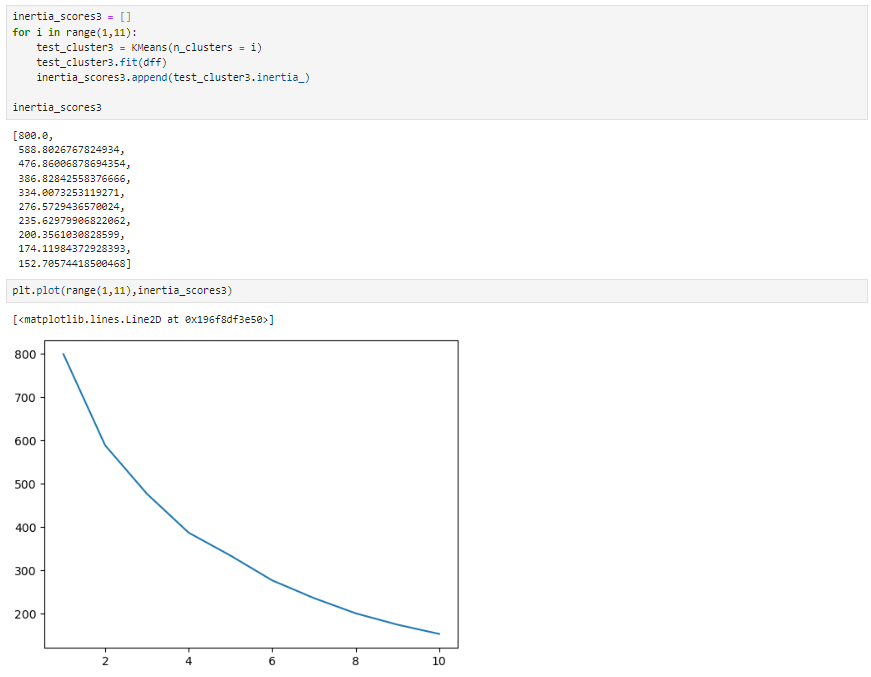
**Data Preparation**

Data was pre-processed for multivariate clustering:

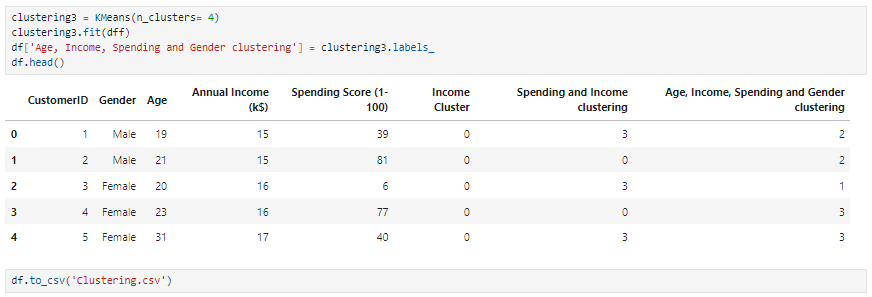


**KMeans Clustering**

KMeans clustering was performed on the scaled data:



Based on the elbow method, 4 clusters were chosen:



**Conclusion**

The analysis provided detailed insights into the customer segmentation based on age, gender, annual income, and spending score. KMeans clustering identified distinct groups of customers that can be targeted for specific marketing campaigns.

This comprehensive analysis enables better understanding of customer behaviour and aids in developing targeted marketing strategies.