# **Data Analyst Project #1**

# **Customer Segmentation & Clustering**

**Problem Statement**

1. Understand the target customers for the marketing team to plan the marketing strategy
2. Identify the purchase pattern based on income, age and shopping score

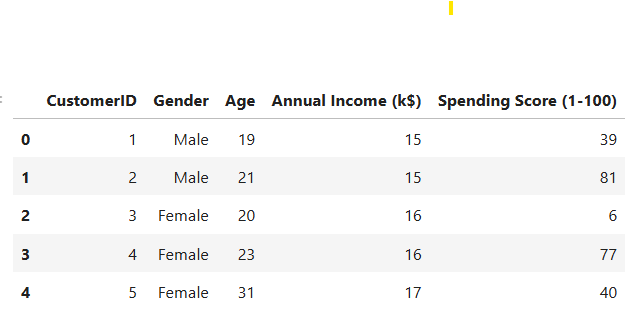
**Objective**

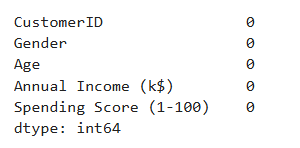
Divide the target market into groups. Create subsets of a market based on demographic behavioral criteria to better understand the target for marketing activities

**Approaches**

1. Clean the Data
2. Perform EDA
3. Use the K-Means algorithm to create segments
4. Summary
5. Visualize

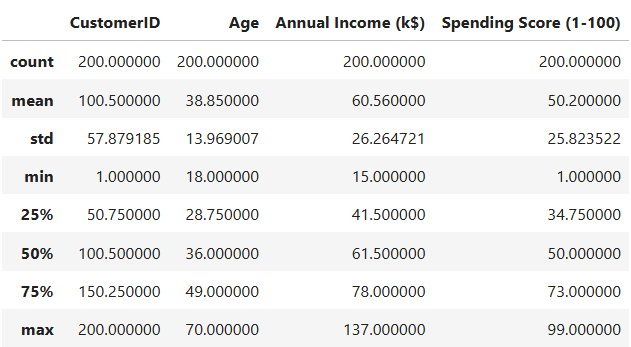
**Data Report**





**Table 1 - DataSet**

The dataset includes **Customer ID**, **Gender**, **Age**, **Annual Income (in thousands of dollars)**, and **Spending Score (a measure from 1 to 100)**. There are **no missing values** in this customer dataset.



**Table 2 - Data\_Description**

**CustomerID**

There are **200 unique customer IDs** in the dataset. The IDs range from 1 to 200, with a mean of 100.5 and a standard deviation of approximately 57.88.

**Age**

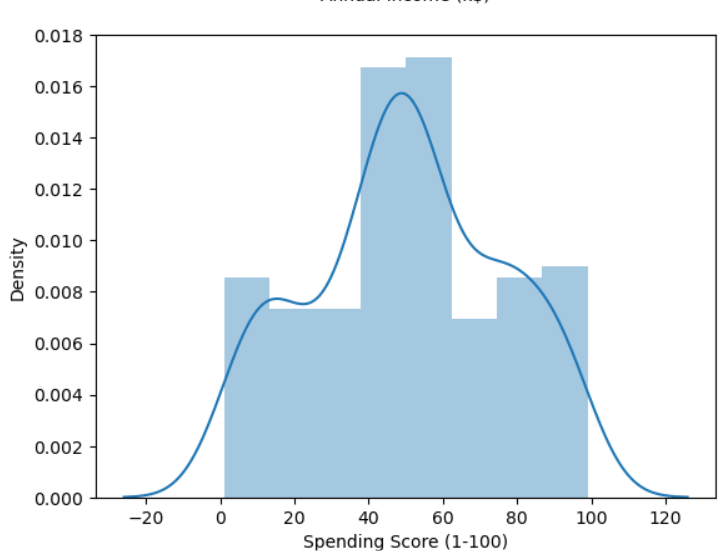
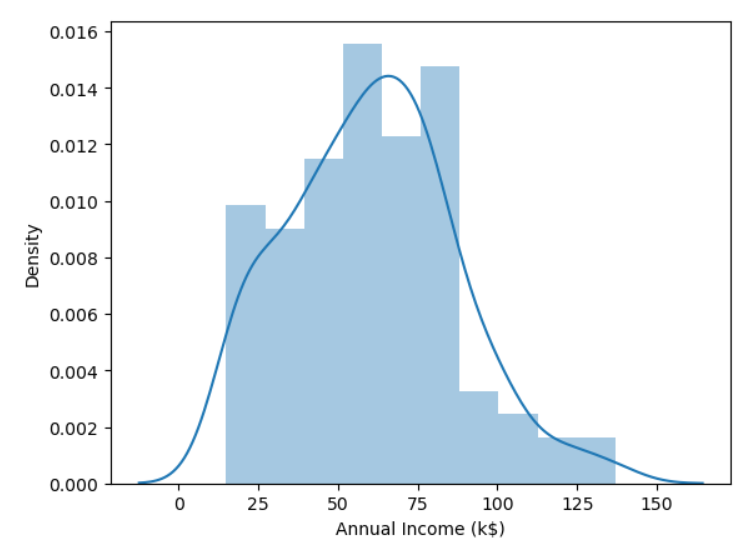
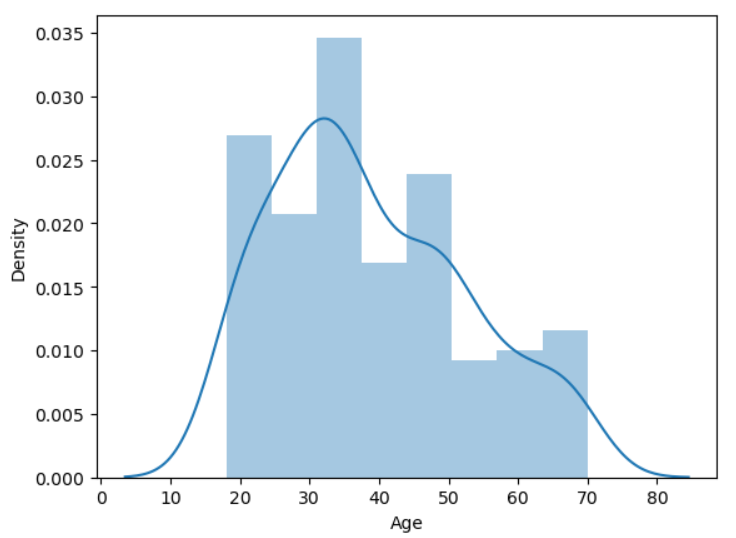
The age of the customers ranges from **18 to 70 years old**, with an average age of **38.85 years**. The median age is **36 years**. The standard deviation is about 13.97, indicating a relatively wide spread of ages.

**Annual Income (k$) 💰**

The annual income, in thousands of dollars, for the 200 customers ranges from a minimum of **$15k to a maximum of $137k**. The average annual income is **$60.56k**, with a median of **$61.5k**. The standard deviation is approximately 26.26.

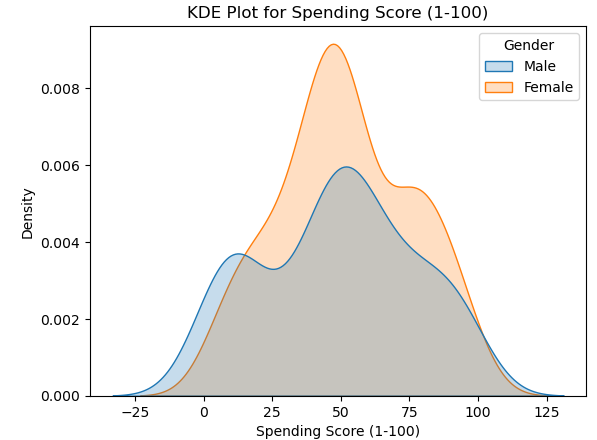
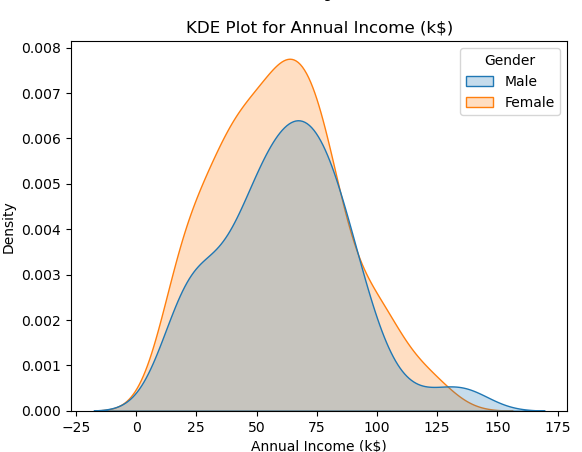
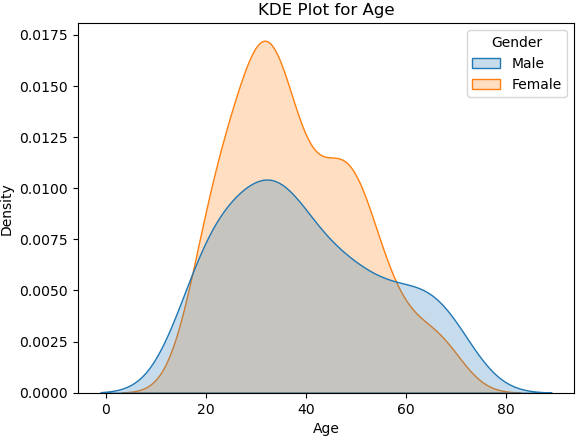
**Spending Score (1-100) 🛍️**

The spending score, which is a value from 1 to 100, is also provided for the **200 customers**. The scores range from a minimum of **1 to a maximum of 99**. The mean spending score is **50.2**, with a median of **50**. The standard deviation is about 25.82. The median and mean values are very close, suggesting a fairly symmetrical distribution of spending scores.



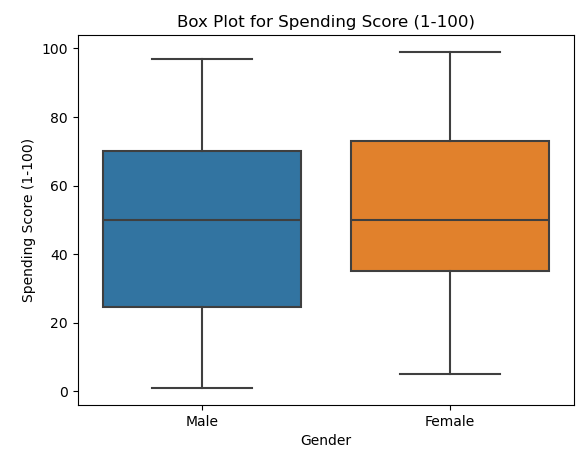
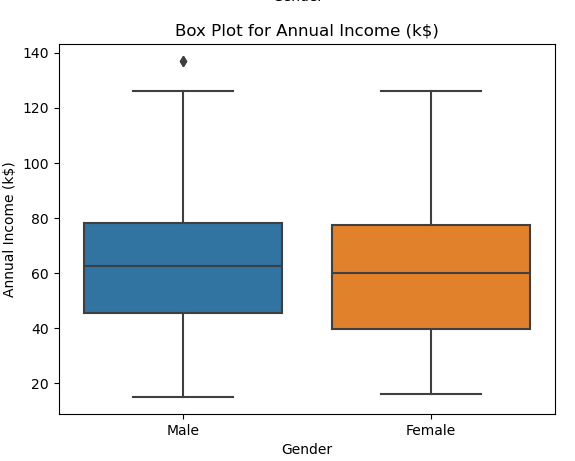
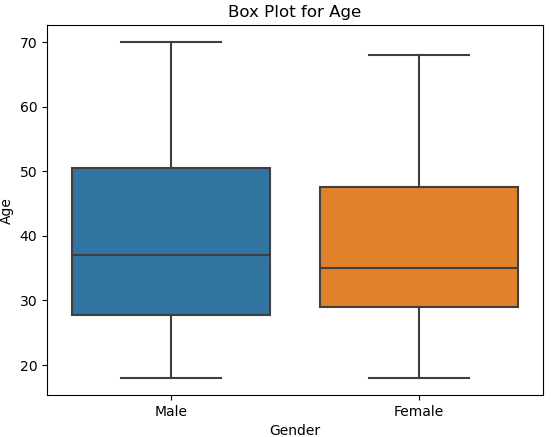
**Figure 1 - Univariate Analysis of Age, Annual Income, Spending Score**

Based on the data, the distributions of **Age** and **Annual Income** appear to be approximately normal. The majority of customers have a **Spending Score** that falls within the **40-60** range.

****

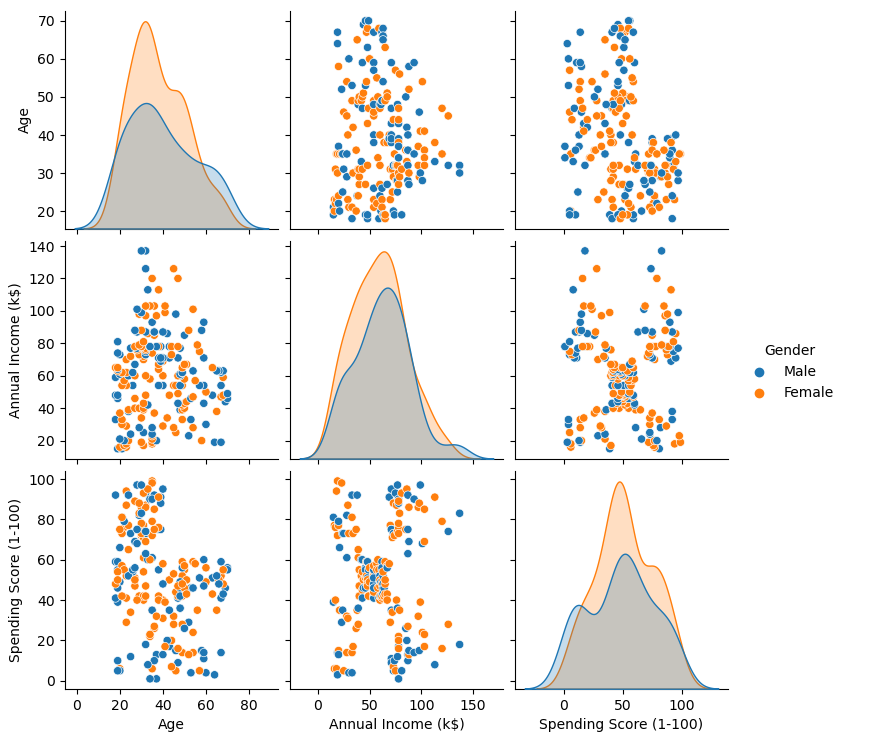
**Figure 2 - KDE plot - Age, Annual Income, Spending Score**

* The highest frequency of female customers is in the **20-40 age range**.
* Female customers appear more frequently in the dataset overall, with some outliers observed in the **150-175 range**



**Figure 3** - **Box plot - Age, Annual Income, Spending Score**

* The median age for males is higher than for females.
* The income data contains outliers.
* While the median spending scores for males and females are similar, there is a wide range of spending scores for both genders.

****

**Figure 4** - **Pair plot - Age, Annual Income, Spending Score**

**Age:** The plot shows that **females tend to be younger on average**, with their age distribution peaking earlier than males. The male distribution is slightly more spread out toward higher ages.

**Annual Income (k$)**: The income distributions for males and females look very similar. Both are approximately normal and centered around the same range, suggesting there isn't a significant difference in average annual income between the genders in this dataset.

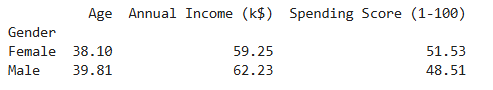
**Spending Score (1-100)**: The spending score distributions are also quite similar. Both genders have a spending score that is roughly normally distributed, peaking in the 40-60 range.

**Relationships Between Variables by Gender**

**Age vs. Spending Score:** There is no strong correlation between age and spending score for either gender. The points are scattered, indicating that a person's age doesn't consistently predict their spending score, regardless of whether they are male or female

**Annual Income vs. Spending Score**: The distinct clusters we saw in the previous pairplot are still visible, and they are composed of both male and female customers. This suggests that while income and spending are related, the specific clusters (e.g., high-income/high-spending or low-income/high-spending) are not exclusively dominated by one gender. Both males and females exist in all the major clusters, making this a useful variable for general segmentation but not for gender-specific targeting.

**Age vs. Annual Income**: There is **no clear relationship** between age and annual income for either gender. The data points for both males and females are widely scattered, indicating that an individual's income is not strongly dependent on their age in this dataset.



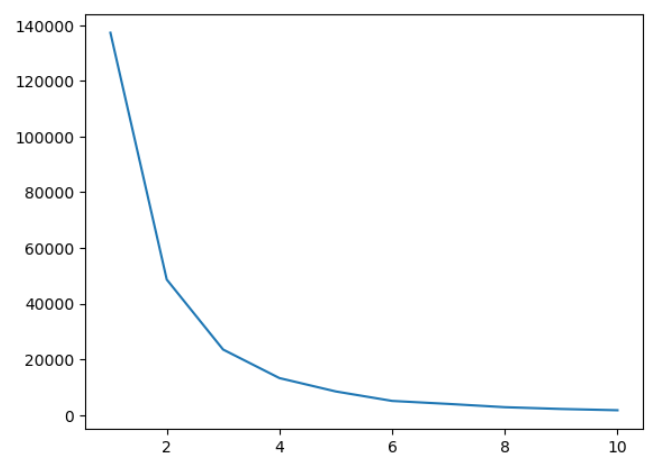
**Table 3: Grouped by - Gender**

**Age**: The average age of **male** customers (39.81) is slightly higher than that of **female** customers (38.10).

**Annual Income**: On average, **males** have a slightly higher annual income ($62.23k) compared to **females** ($59.25k).

**Spending Score**: The average spending score for **females** (51.53) is higher than the average for **males** (48.51). This suggests that female customers, on average, have a higher spending score despite having a slightly lower average income.

**Clustering**

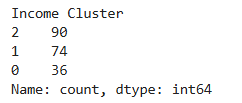
****

**Figure 5 - Elbow plot - Income Cluster**

Based on the inertia plot generated for various numbers of clusters (the elbow method), we observe a **distinct elbow at k=3**. When moving from k=1 to k=3, inertia decreases rapidly, indicating that more clusters significantly improve the separation of the data. However, beyond k=3, the rate of decrease slows greatly: additional clusters provide only marginal benefit in explaining variance.

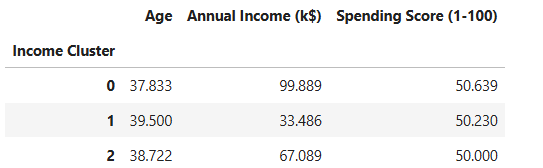
Therefore, the "elbow" suggests that the optimal number of clusters in this dataset is 3—this is the point where we strike a balance between maximizing within-cluster similarity (low inertia) and avoiding unnecessary complexity (too many clusters).

Selecting k=3 is consistent with the standard interpretation of the elbow method and ensures that the clusters discovered by KMeans are both meaningful and practical for analysis



**Table 4 - Income Cluster Value Count**

The KMeans clustering analysis grouped the dataset into three clusters, containing 90, 74, and 36 data points, respectively. This segmentation reflects meaningful, data-driven partitions of the population based on their similarities in age, income, and spending habits. The cluster sizes are somewhat uneven, suggesting that one group **(cluster 2) considerably dominates the dataset**, while the other clusters are smaller and **may represent niche or outlier segments**. This **clustering enables further analysis for targeted marketing, personalized recommendations, or customer profiling.**

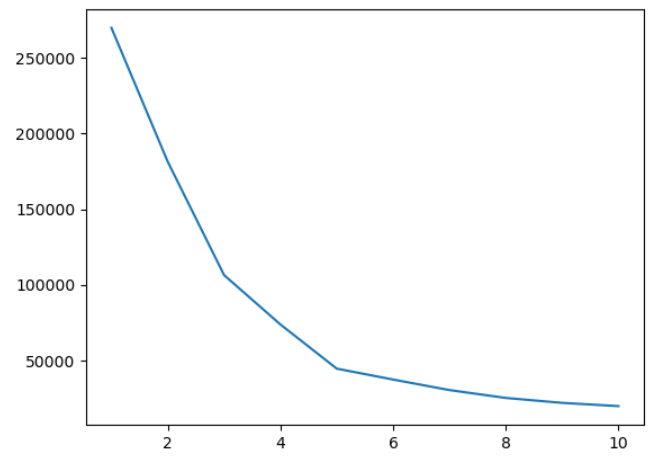


**Table 5 - Income Cluster Groupby Age, Annual Income, Spending Score**

Customer **income level** is the key variable for **differentiation** in this dataset, while **age and spending score remain almost constant** across groups.

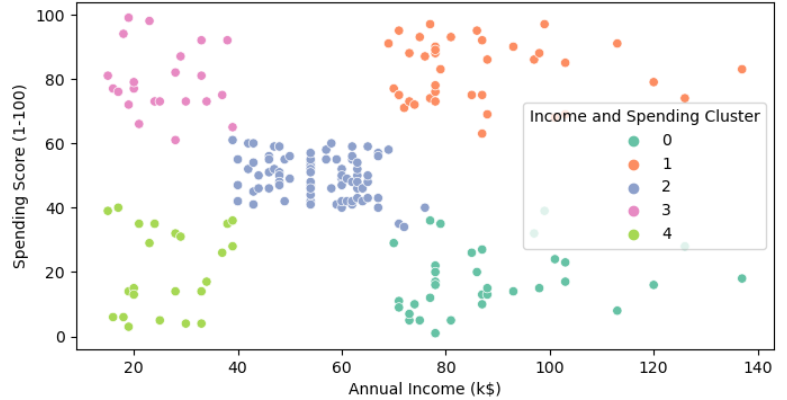
Such segmentation can be directly used for **personalized marketing strategies**, pricing decisions, and inventory planning.

**Bivariate Analysis - Clustering**



**Figure 6 - Elbow plot - Annual Income & Spending Score**

Based on the inertia plot generated for various numbers of clusters (the elbow method), we observe a **distinct elbow at k=5**

****

**Figure 7 - Clusters - Annual Income & Spending Score**

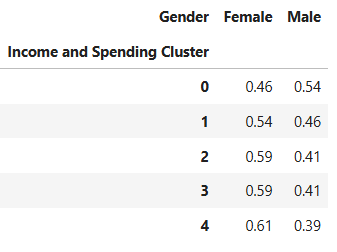
**High-Value Customers (Cluster 1):** This group has **high income and high spending**. They are our most profitable segment and should be the focus of retention and loyalty programs.

**High-Income Savers (Cluster 0)**: These customers have **high income but low spending**. They represent a significant opportunity for growth through targeted marketing and upselling campaigns.

**Budget-Conscious Spenders (Cluster 3)**: This group has **low income but high spending**. They are likely very responsive to promotions and discounts, making them an ideal target for sales-driven campaigns.

**Average Customers (Cluster 2)**: This is the largest segment, with **average income and average spending**. They form the core customer base and can be targeted with general campaigns to encourage increased engagement.

**Low-Engagement Customers (Cluster 4)**: This group has **low income and low spending**. They are the least engaged segment and a lower priority for high-value marketing efforts.



**Table 6 - Income & Spending Cluster of Female & Male**

**Cluster 0 (High Income, Low Spending):** This cluster is slightly more **male-dominated**, with 54% male customers and 46% female customers.

**Cluster 1 (High Income, High Spending)**: This is the opposite of Cluster 0, with a slight **majority of female customers** at 54% and 46% male customers.

**Cluster 2 (Average Income, Average Spending)**: This is the **core customer** base, and it is predominantly **female**, with 59% female customers and 41% male customers.

**Cluster 3 (Low Income, High Spending)**: Similar to Cluster 2, this group is also predominantly female, with 59% female customers and 41% male customers.

**Cluster 4 (Low Income, Low Spending)**: This cluster has the highest proportion of females, with 61% female customers and 39% male customers

## **Analysis**

## Based on our analysis, we should target Cluster 1, as it shows the highest average income and spending per customer. This cluster presents the most significant opportunity for growth.

In Cluster 1, approximately 54% of customers are female. We should develop targeted marketing and advertising campaigns specifically designed to attract and engage this demographic, which represents a key segment of our most profitable customer base.