



**Project Tittle:** Mall Customer Segmentation using K-Means Clustering.

**Date of Submission:** 01/05/23

**Course Title:** Introduction to Data Science

**Section:** [D] **Semester:** Spring 2022-2023

**Course Teacher:** Tohedul Islam

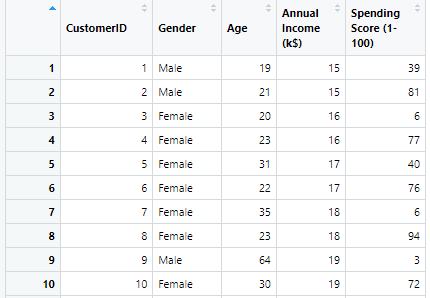
**Submitted By:**

**NAME:** Tanjim, Samiul Arif

**ID:** 20-42694-1

**Dataset Description:** The dataset contains information about Mall Customer. It has 5 attributes and 200 instances.

1. CustomerID: A unique ID issued to the customer
2. Gender: Gender of the customer
3. Age: Age of the customer
4. Annual Income (k$): Annual Income of the customer
5. Spending Score (1-100): Score assigned by the mall based on customer behavior and spending habits.



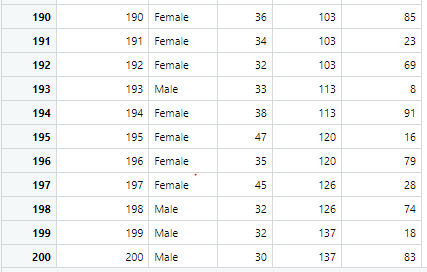


Fig: Mall Customer Dataset

**Dataset Preparation**

**Find Duplicate Values:** In this process, we try to find if there is any duplicate instance in the dataset.





It is found that there are no duplicate values in the dataset.

**Find Missing Values:** Here, we try to find if there are any missing value in the dataset.





It shows that there are no missing values in the dataset.

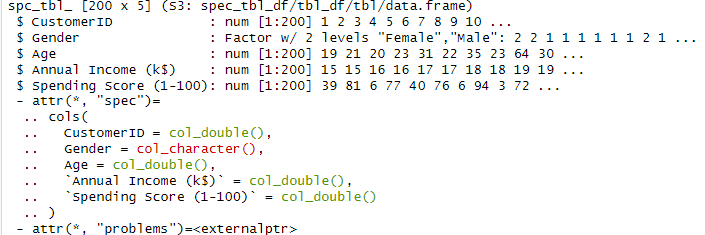
**Factoring Gender:** In the dataset, Gender column contains categorical value. So we have to convert the character vector to factor using **as.factor()** function.





After doing factorization two levels has been found. One is female another one is male.

**Dimension/size of dataset:** 200 \* 5

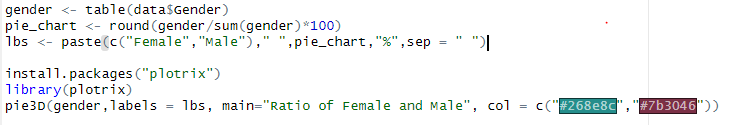
**Structure of dataset:** 

**Univariate Exploration**

**Univariate Exploration of Gender:**







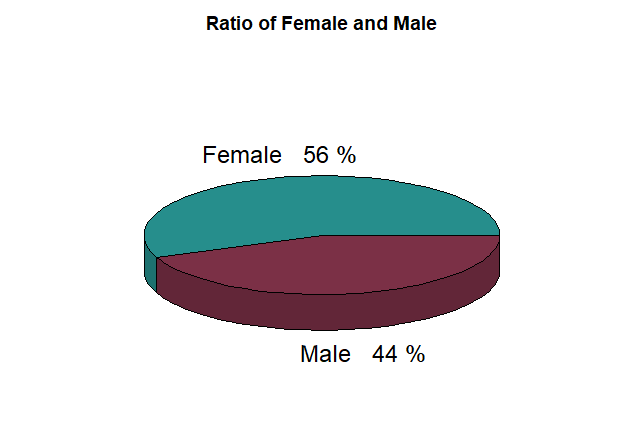
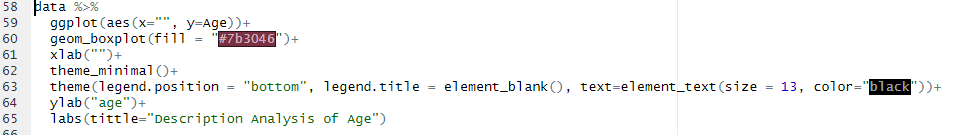


Fig: Pie Chart of Gender

**Univariate Exploration of Age:**







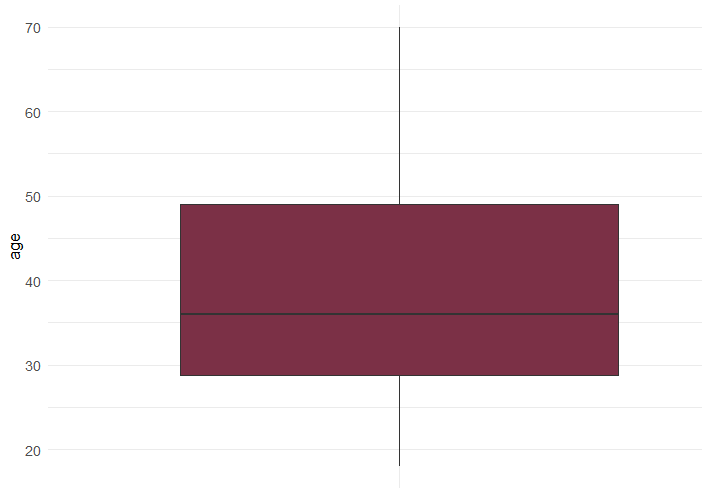
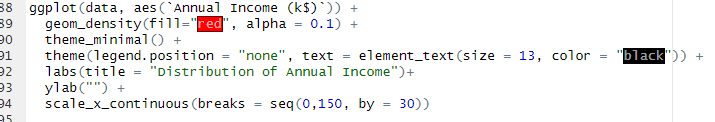


Fig: Box Plot of Age

**Univariate Exploration of Annual Income:**







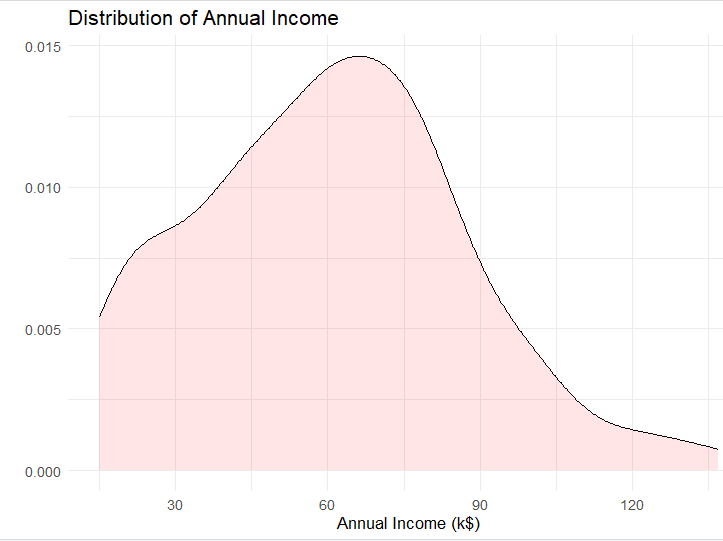
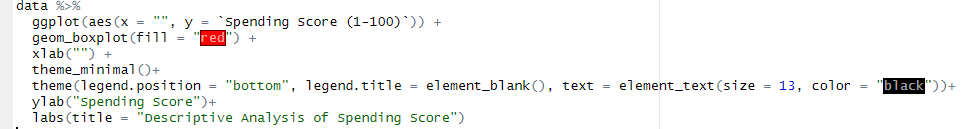


Fig: Density Distribution of Annual Income

**Univariate Exoloration of Spending Score:**







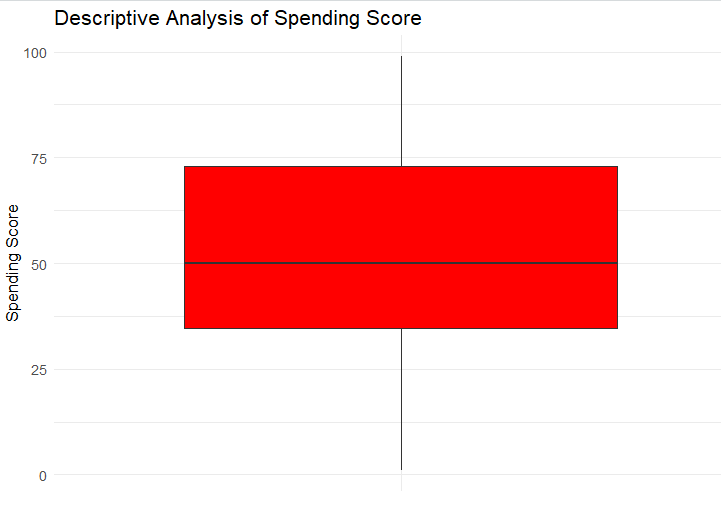
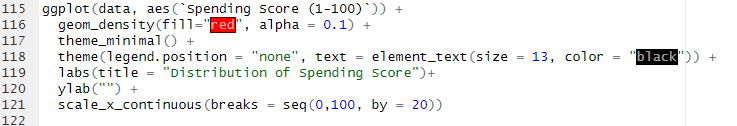


Fig: Box Plot of Spending score



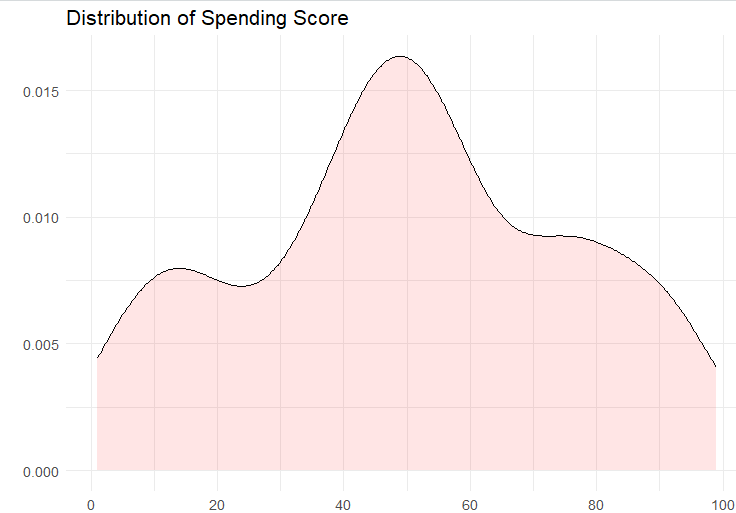
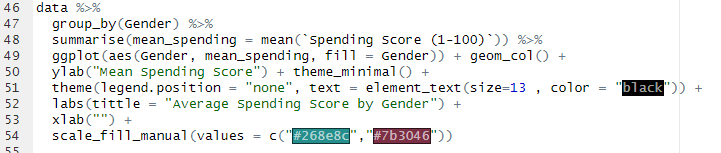


Fig: Density Distribution of Spending Score

**Multivariate Exploration**

**Multivariate Exploration of Spending Score & Gender:**



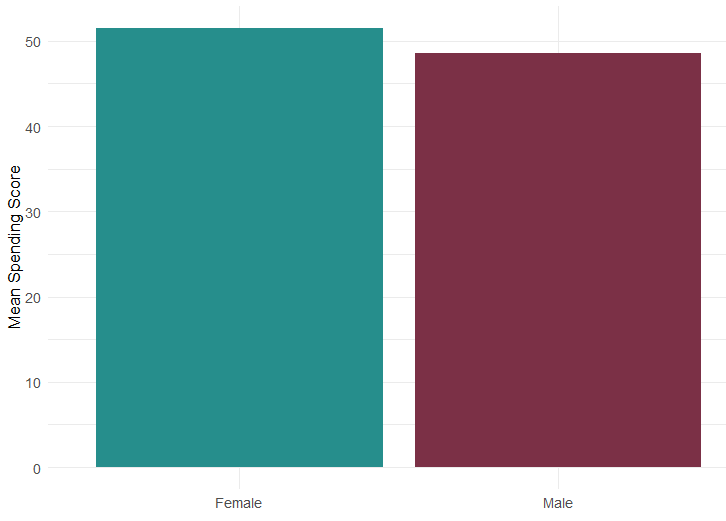
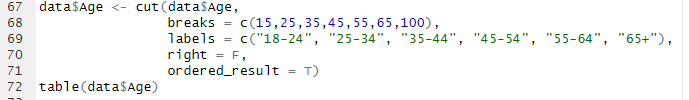
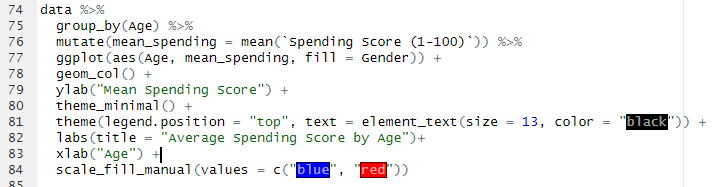


Fig: Plot of Average Spending Score by Gender

**Multivariate Exploration of Spending Score & Age:**







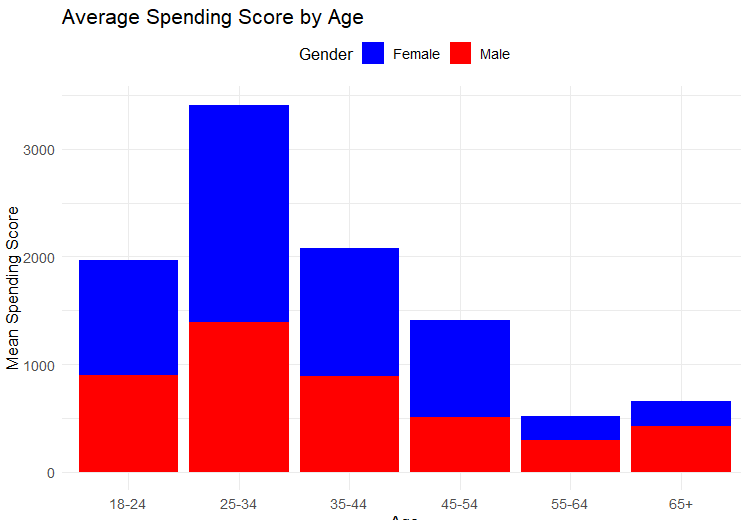


Fig: Historgram of Average Spending Score by Age (Female & Male)

**Multivariate Exploration of Spending Score & Annual Income:**

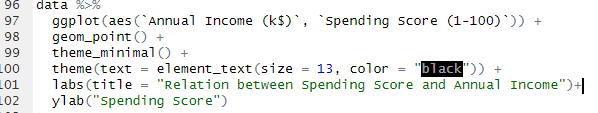


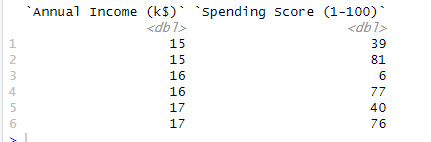


Fig: Scatter Plot of Spending Score by Annual Income

**K-Means Clustering**

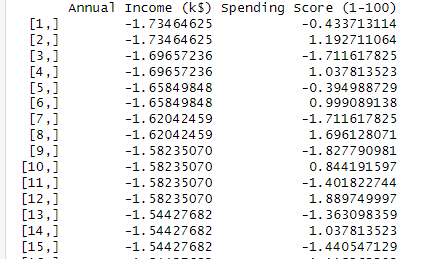
We will select two attribute for applying k-means alogirthm. K-Means Clustering will applied on the annual Income and the spending score.



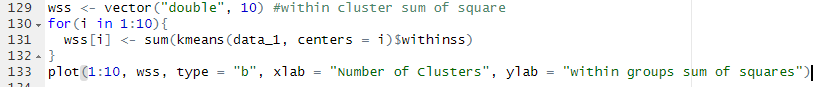


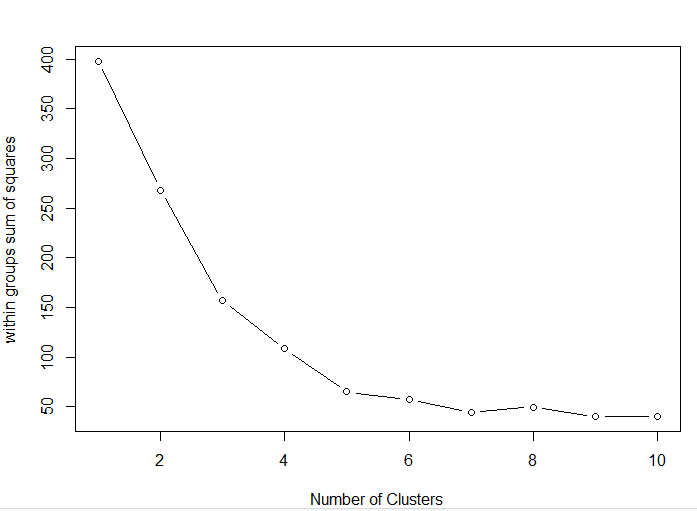
We have to normalize these values before applying k-means algorithm for better precision.





Finding the optimal number of cluster using **Elbow method**. Within-cluster sum of square function (WSS). WSS is total distance of data points from their respective cluster centroids.

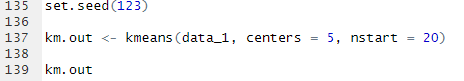


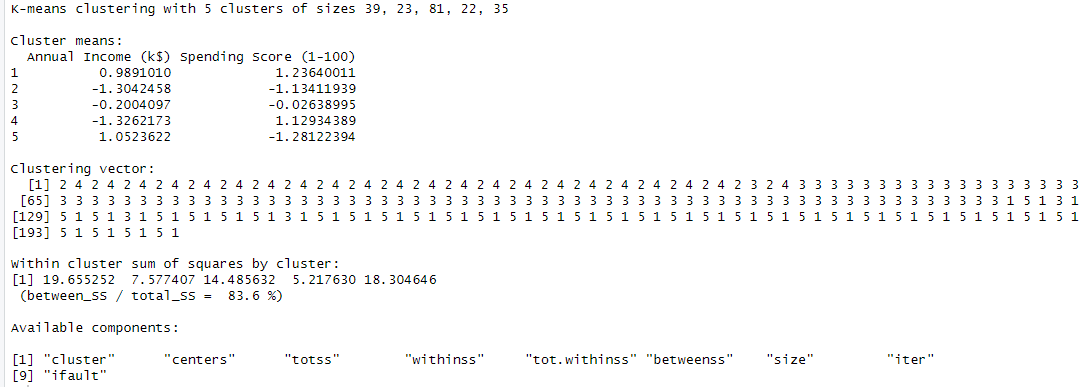


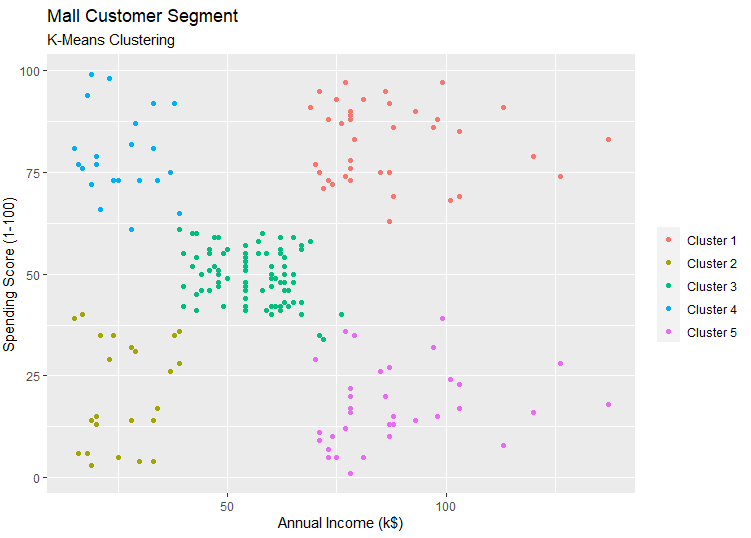
The scree plot shows how the overall within-cluster sum of squares reduces as the number of clusters increases. The criterion for determining the number of clusters is to discover an elbow where WSS declines considerably more slowly after adding another cluster.

Based on this plot, we can conclude that setting the number of clusters to k=5 is the optimal option. After 5, it appears to gradually degrade performance.

**As a result, the cluster number will be , K=5.**

****

****

****

**Fig: K-Means Algorithm (Clusters Of Customers)**

The generated "Clusters of Customers" plot indicates the distribution of the five clusters. A reasonable explanation for the mall customer segments is:

Cluster 1: Customers with a high annual income and a high annual spend

Cluster 2: Customers with a low annual income and a low annual spend

Cluster 3: Customers with a medium annual income and a medium annual spend

Cluster 4: Customers with low annual income but significant annual spending

Cluster 5: Customers with high annual income but low annual spending

Dataset Link: <https://github.com/kennedykwangari/Mall-Customer-Segmentation-Data/blob/master/Mall_Customers.csv>