```r
library(ggplot2)  # Load the ggplot2 package for data visualization

# Read the dataset from the specified file path
data <- read.csv("C://Users//Minusha Attygala//OneDrive//Documents//Big Data Practicals//Customer_Data.csv")
# Display the first few rows of the dataset
head(data)
# Display the structure of the dataset (data types of each column)
str(data)

# Plot Annual Income against Spending Score to visualize data distribution
ggplot(data, aes(x= AnnualIncome, y= SpendingScore, color="red"))+
  geom_point(size=3, alpha=0.7)+  # Add points with size and transparency
  labs(title = "Annual Income against Spending Score Distribution",  # Add title and axis labels
       x = "Annual Income",
       y = "Spending Score")+
  theme_minimal()  # Use a minimal theme for better aesthetics

# Implementation of K-Means clustering
# Select relevant columns (Annual Income and Spending Score) for clustering
model_data <- data[,c("AnnualIncome", "SpendingScore")]
# Display the first few rows of the selected data
head(model_data)

# Apply the K-Means clustering algorithm with 3 clusters
set.seed(100)  # Set seed for reproducibility
results <- kmeans(model_data, centers = 3)  # Perform clustering with 3 clusters

# Add cluster labels to the original dataset
data$Cluster <- as.factor(results$cluster)
```

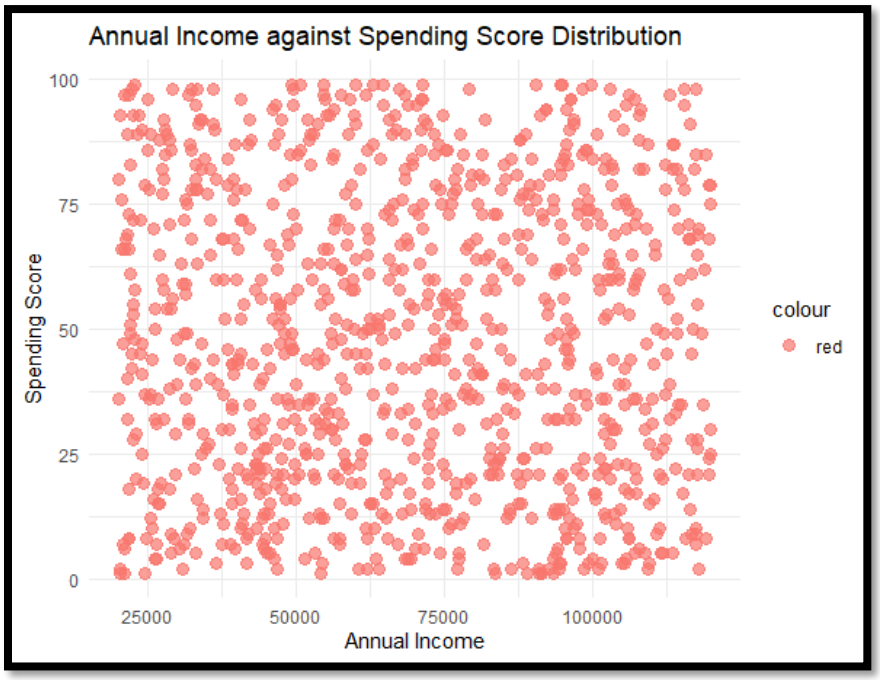
```r
# Plot Annual Income against Spending Score, colored by assigned clusters
ggplot(data, aes(x= AnnualIncome, y= SpendingScore, colour = Cluster))+
  stat_ellipse(aes(fill = Cluster), geom = "polygon", alpha=0.4)+  # Draw cluster boundaries
  geom_point(size=3, alpha=0.7)+  # Plot data points
  labs(title = "Annual Income against Spending Score Distribution of Clusters",  # Add title and axis label
       x = "Annual Income",
       y = "Spending Score")+
  theme_minimal()  # Use a minimal theme for better aesthetics

# Finding the optimal number of clusters using the Elbow Method
optimal_number <- sapply(1:10, function(k) kmeans(model_data, centers = k)$tot.withinss)

# Plot the Elbow Method to determine the optimal number of clusters
plot(1:10, optimal_number,  # X-axis: Number of clusters, Y-axis: Total within-cluster sum of squares
     type = "b",  # "b" type means both points and lines will be plotted
     pch = 19,  # Use solid circle points for better visibility
     col = "red",  # Set color to red for clear distinction
     xlab = "Number of Clusters",  # Label for X-axis
     ylab = "Total Within-Cluster Sum of Squares",  # Label for Y-axis
     main = "Elbow Method for Optimal Clusters")  # Title for the plot
```

This code performs K-Means clustering on customer data, specifically analyzing the relationship between Annual Income and Spending Score to identify meaningful customer segments. It uses the ggplot2 library for visualization and applies K-Means clustering to group customers based on their spending behavior.

First, the dataset is loaded from the specified file path and its structure is inspected using **head()** and **str()**. A scatter plot is then created using **ggplot2** to visualize the distribution of Annual Income vs. Spending Score, helping to understand how customers are spread in terms of spending habits.

The K-Means clustering algorithm is then applied to two selected features: Annual Income and Spending Score. The number of clusters is initially set to 3, and the clustering results are stored in the results variable. The assigned cluster labels are added to the original dataset, and another scatter plot is generated, now color-coded by clusters. Additionally, **stat_ellipse()** is used to draw boundaries around each cluster, helping to visualize the grouping.

Finally, the Elbow Method is implemented to determine the optimal number of clusters. It calculates the total within-cluster sum of squares for different values of **k** (from 1 to 10). The results are plotted, allowing the user to identify the optimal number of clusters by looking for the elbow point the point where adding more clusters does not significantly reduce the sum of squared distances within clusters. This approach ensures that the chosen number of clusters provides the best balance between simplicity and accuracy.

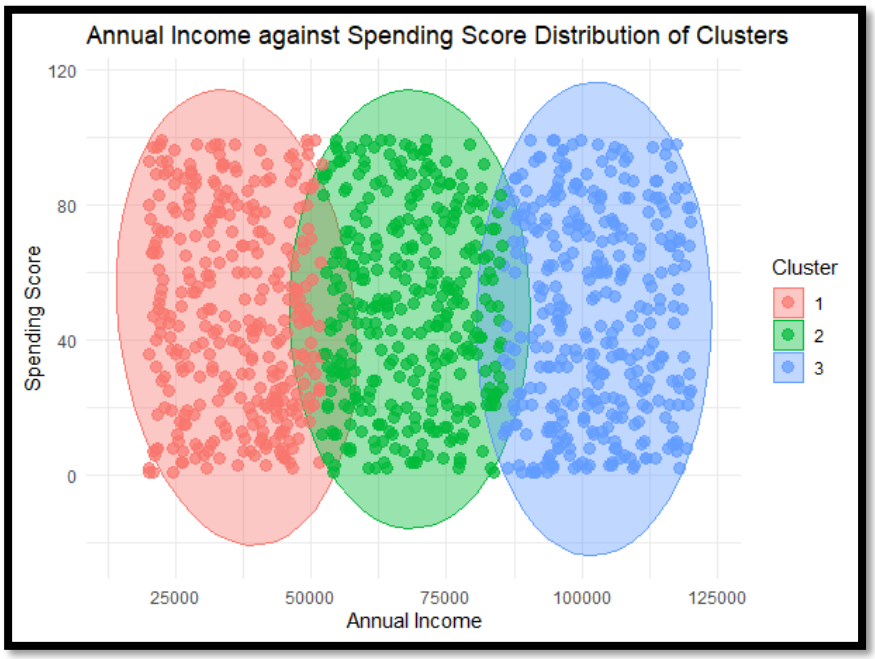


- The scatter plot represents the relationship between **Annual Income** and **Spending Score** of customers, with each red dot indicating an individual data point.

- The distribution appears to be widespread, showing no clear clustering pattern, which suggests that income and spending habits vary significantly among customers.

- The color legend indicates that all points are assigned the same color (red).

- To improve visualization, proper clustering or categorization should be applied to distinguish different customer groups based on their spending behavior.

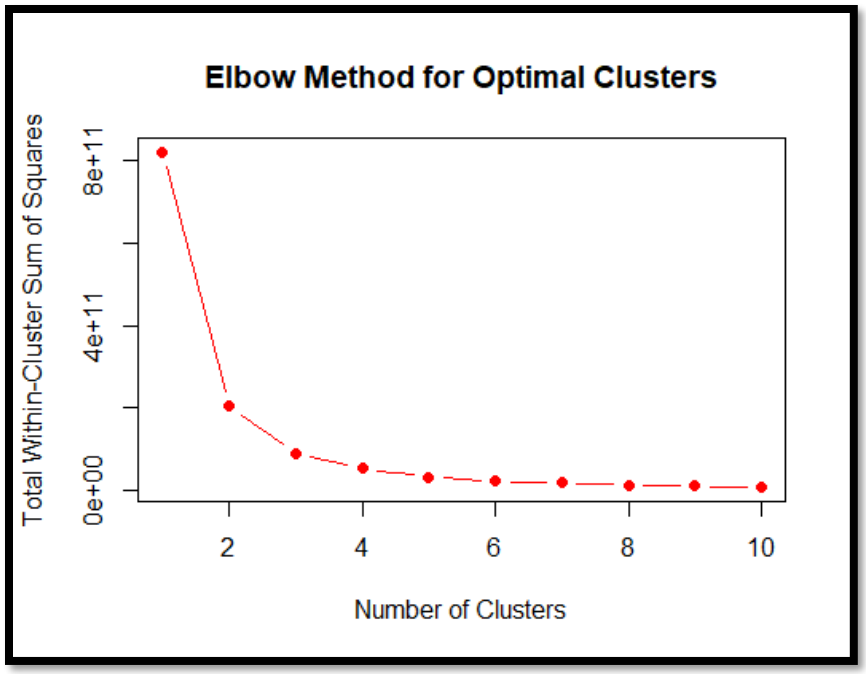


- The scatter plot represents customer segmentation based on **Annual Income** and **Spending Score**, with 3 distinct clusters visualized using different colors (red, green, and blue).

- Each cluster is enclosed within an ellipse, indicating the **distribution and density** of data points within that group.

- The segmentation appears to be primarily based on **Annual Income**, with lower-income customers (Cluster 1) on the left, mid-income customers (Cluster 2) in the center, and higher-income customers (Cluster 3) on the right.

- The clustering approach effectively groups customers with **similar spending behaviors**, which can be useful for targeted marketing and business decision-making.



- The plot represents the **Elbow Method** for determining the optimal number of clusters in a K-Means clustering model.

- The **y-axis** shows the **total within-cluster sum of squares (WCSS)**, which measures how compact the clusters are, while the **x-axis** represents the number of clusters.

- The plot exhibits a sharp decline in WCSS up to **3 clusters**, after which the decrease slows, forming an "elbow" at **k = 3** indicating that 3 is likely the optimal number of clusters.

- Adding more clusters beyond this point does not significantly reduce WCSS, meaning additional clusters provide diminishing returns in terms of improved segmentation.

The **Elbow Method** suggests that the optimal number of clusters for segmenting the customers based on **Annual Income and Spending Score** is **3**, as the WCSS significantly drops at this point before stabilizing.

The **K-Means clustering visualization** confirms this, showing 3 well-separated clusters that effectively categorize customers into distinct groups based on their spending behavior and income levels.

This segmentation can help businesses tailor their marketing strategies, offering personalized promotions or services to different customer groups based on their spending patterns and financial capacity.