Exploratory Data Analysis

- Exploratory Data Analysis (EDA) is an important step in understanding and visualizing your data.
- EDA in R is essential for gaining a comprehensive understanding of the dataset, detecting issues or anomalies, exploring relationships, and generating insights that can drive further analysis, modeling, and decision-making. It serves as the foundation for subsequent data analysis tasks and helps ensure the accuracy, reliability, and validity of your results.

- Exploratory Data Analysis (EDA) is a crucial step in the data analysis process, and it is particularly useful in R for several reasons:
- Data Understanding
- Data Quality Check
- Insights and Patterns
- Variable Selection
- Assumptions Checking
- Communication and Visualization

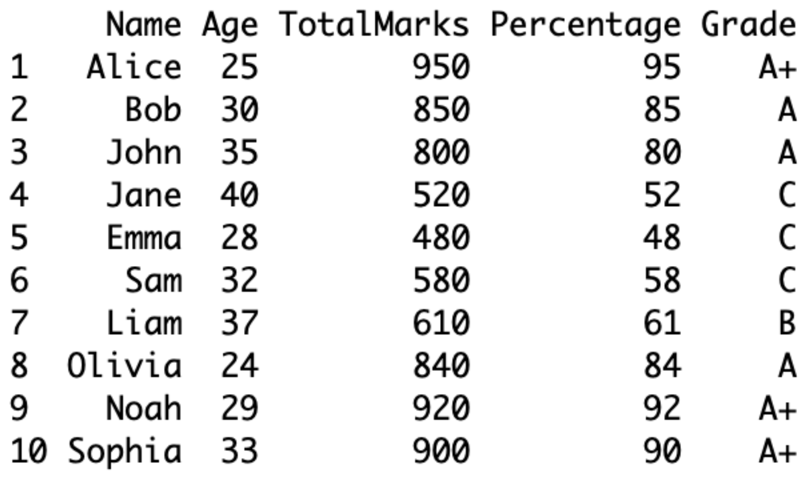
Exploratory Data Analysis (EDA) can be done through

- Box Plot

Box Plot

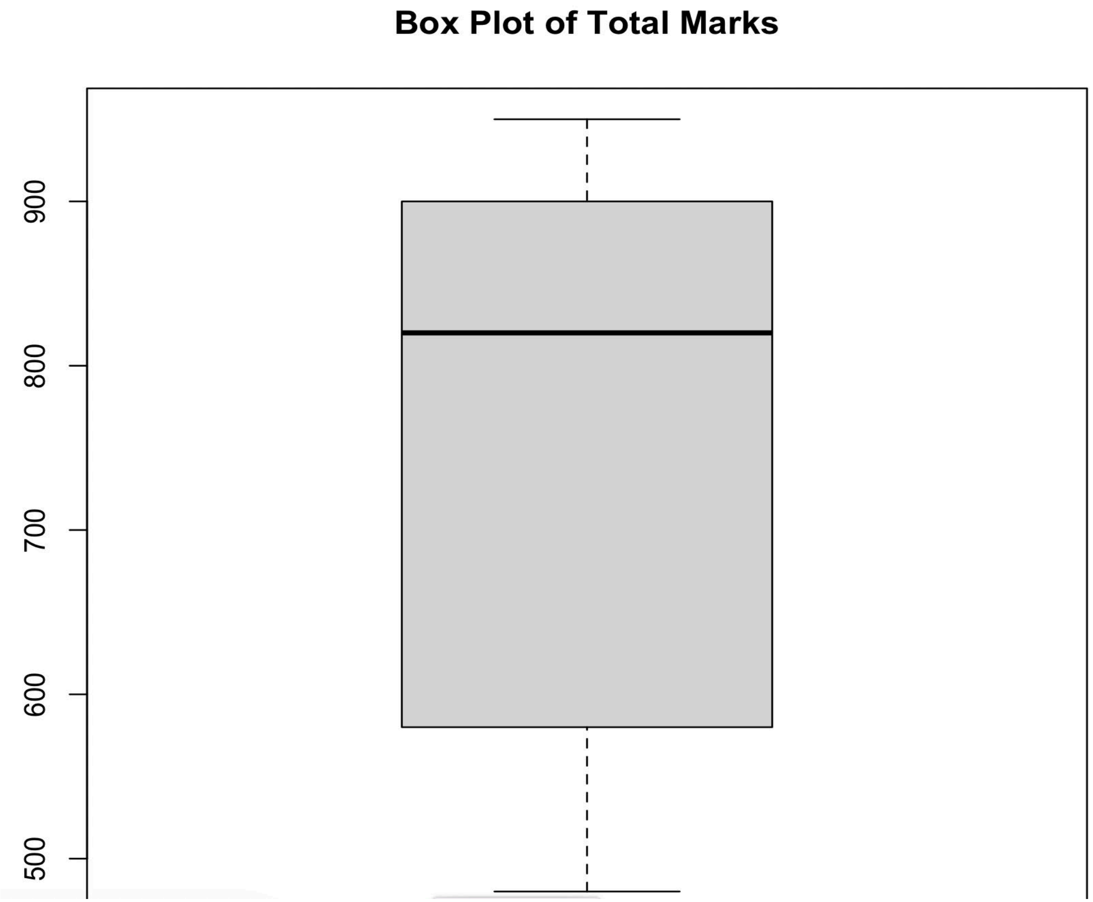
- A box plot, also known as a box-and-whisker plot, provides a graphical representation of the distribution of a numerical variable.
- It displays the median, quartiles, and possible outliers in the data.
- It helps identify skewness, variability, and potential outliers in the dataset.

```
df <- data.frame( Name = c("Alice", "Bob", "John", "Jane", "Emma", "Sam",
"Liam", "Olivia", "Noah", "Sophia"),Age = c(25, 30, 35, 40, 28, 32, 37, 24,
29, 33),TotalMarks = c(950, 850, 800, 520, 480, 580, 610, 840, 920, 900),
Percentage = c(95, 85, 80, 52, 48, 58, 61, 84, 92,90), Grade = c("A+", "A",
"A", "C", "C", "C","B", "A", "A+", "A+") )
```



```
     Name Age TotalMarks Percentage Grade
1   Alice  25        950         95    A+
2     Bob  30        850         85     A
3    John  35        800         80     A
4    Jane  40        520         52     C
5    Emma  28        480         48     C
6     Sam  32        580         58     C
7    Liam  37        610         61     B
8  Olivia  24        840         84     A
9    Noah  29        920         92    A+
10 Sophia  33        900         90    A+
```

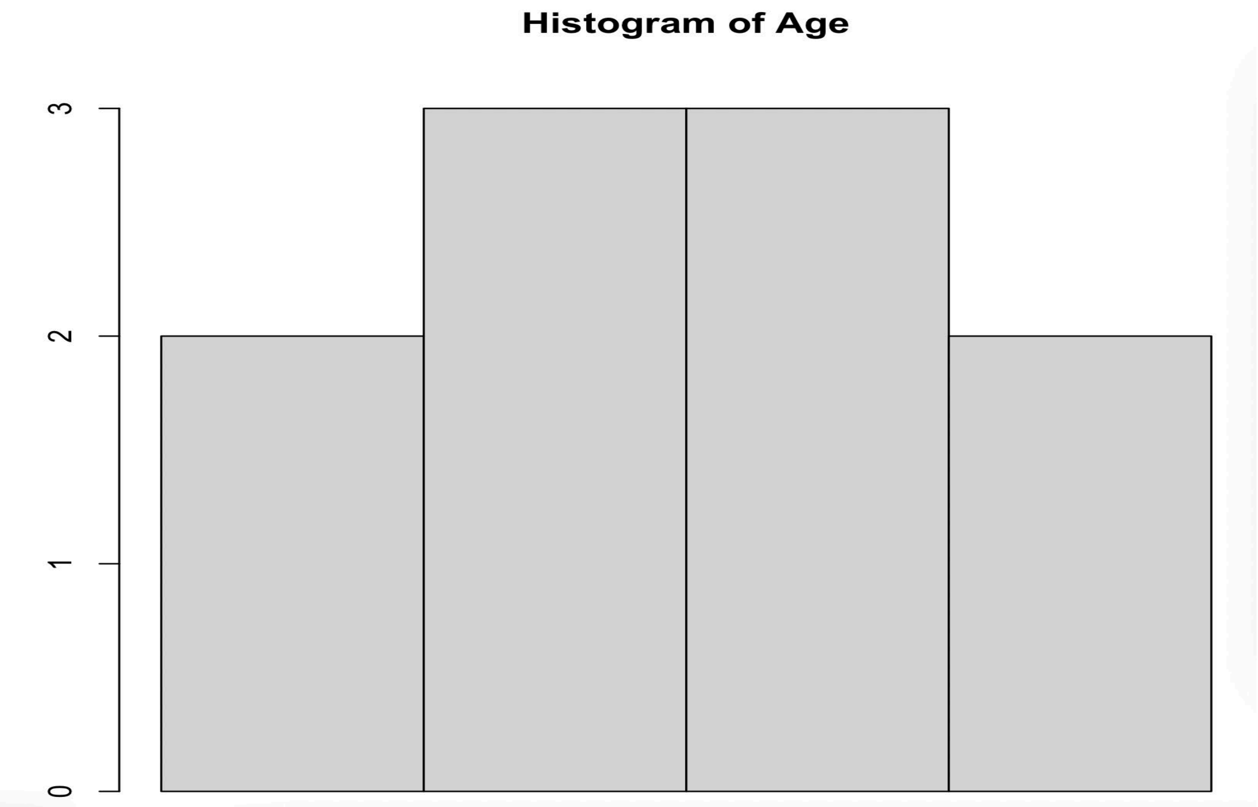
```
boxplot(df$TotalMarks, main = "Box
Plot of Total Marks")
```



Histogram:

- A histogram is used to visualize the distribution of a numerical variable.
- It divides the data into bins and displays the frequency or count of observations falling into each bin.
- It helps identify patterns, skewness, and central tendency in the data.

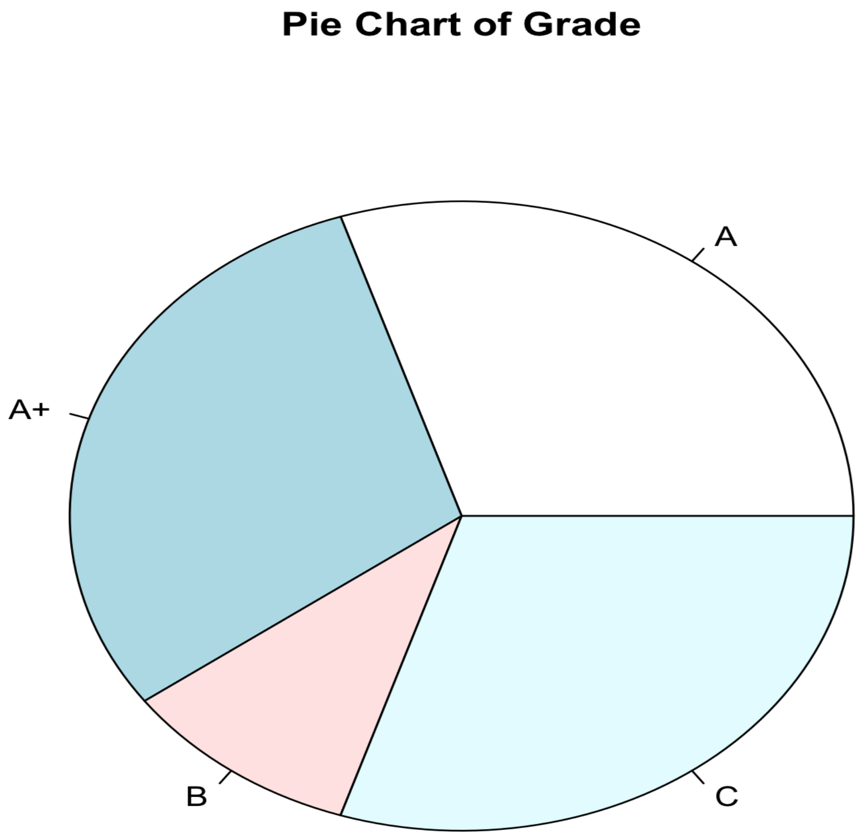
- hist(df$Age, main = "Histogram of Age", xlab = "Age")



Pie Chart:

- A pie chart is used to represent categorical data as a proportion of a whole.
- It displays the distribution of categories as slices of a pie, where each slice represents the proportion of each category.
- It helps understand the relative frequencies or proportions of different categories in the dataset.

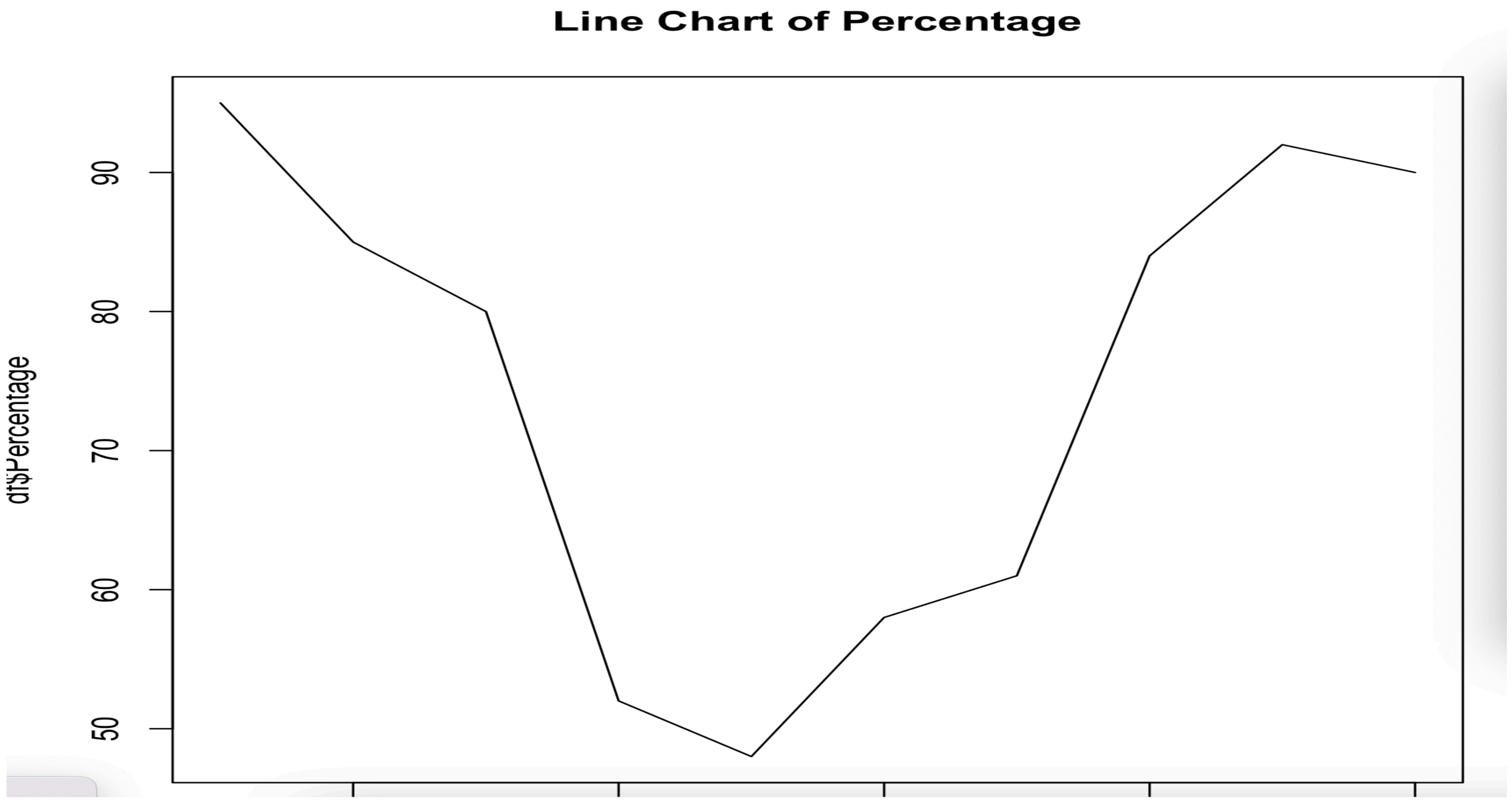
```
pie(table(df$Grade), main = "Pie Chart of Grade")
```



Line Chart:

- A line chart, or line plot, is used to display the relationship between two numerical variables over a continuous scale.
- It connects data points with straight lines to show trends or patterns in the data.
- It helps visualize the changes in one variable with respect to the other over time or any other continuous scale.

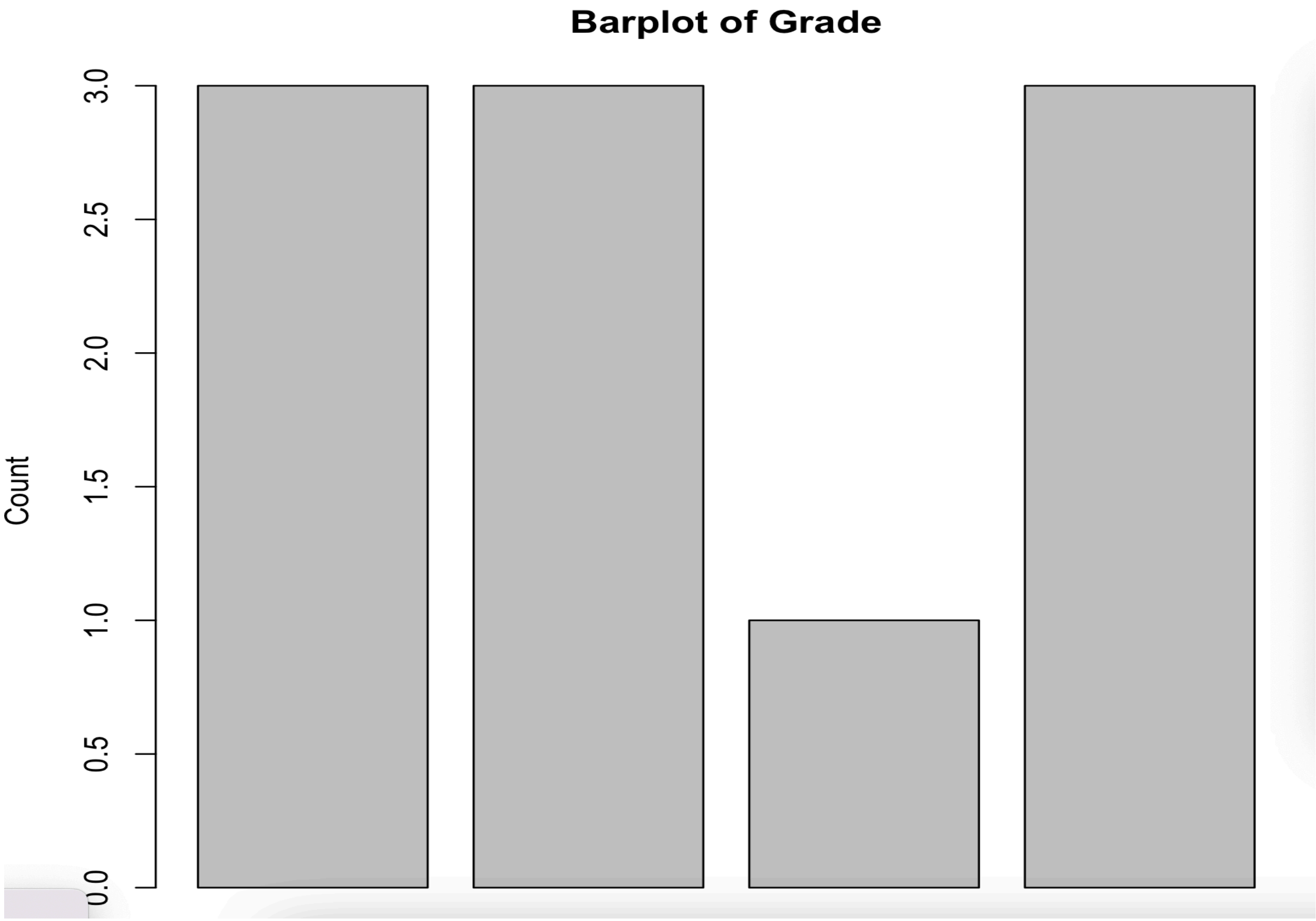
```
plot(df$Percentage, type = "l", main = "Line Chart of Percentage")
```



Barplot

- A barplot is used to represent the distribution or comparison of categorical data.
- It displays bars of different heights or lengths, where each bar represents a category and its height represents the frequency, count, or any other measure associated with that category.
- It helps compare the values or frequencies of different categories.

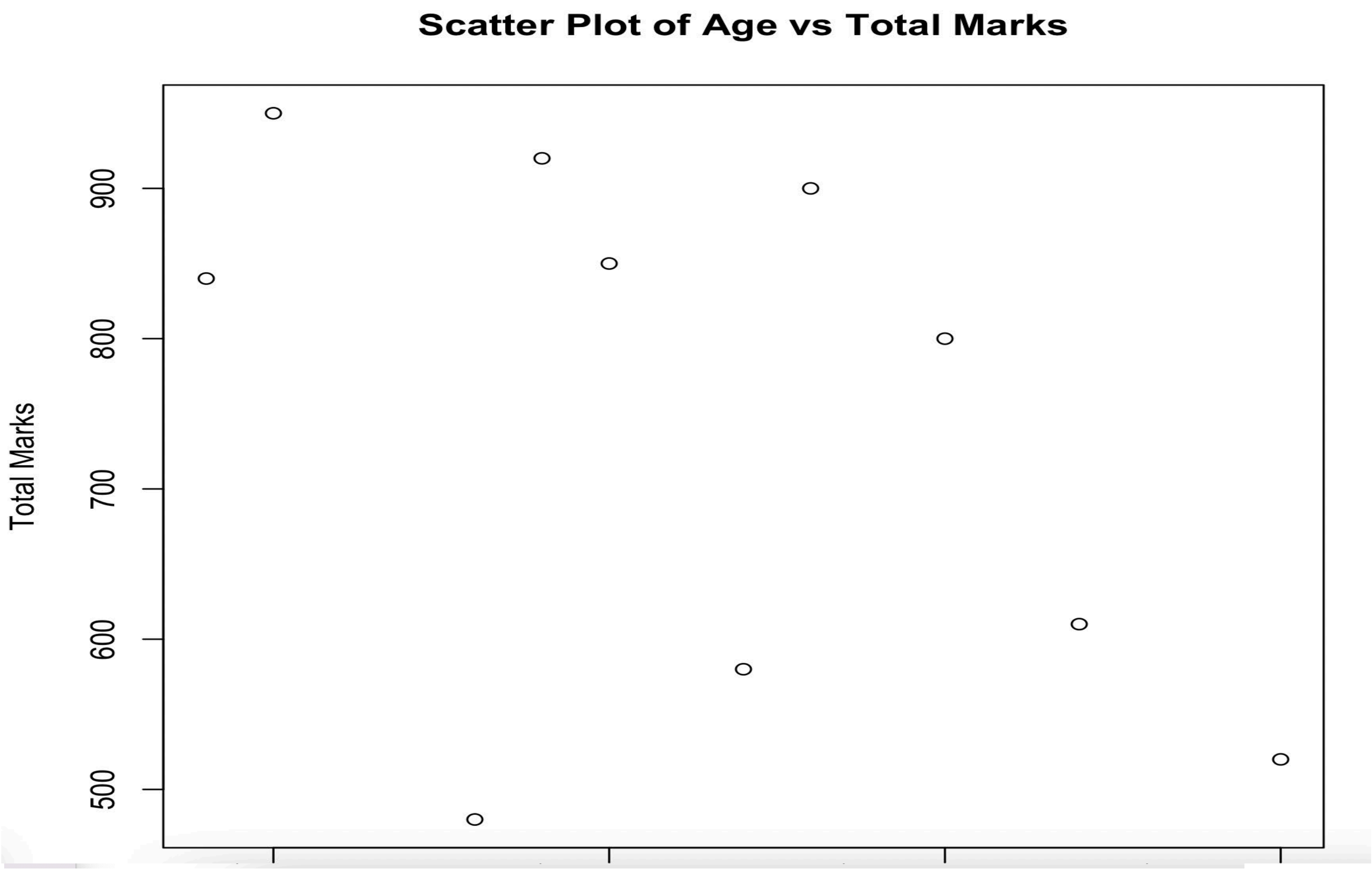
- barplot(table(df$Grade), main = "Barplot of Grade", xlab = "Grade", ylab = "Count")



Scatter Plot

- A scatter plot is used to visualize the relationship between two numerical variables.
- It displays data points as individual dots on a two-dimensional plot, where each dot represents a combination of values for the two variables.
- It helps identify patterns, correlations, clusters, or outliers in the data.

-
  ```
  plot(df$Age, df$TotalMarks, main = "Scatter Plot of Age vs Total Marks", xlab = "Age", ylab = "Total Marks")
  ```


Scatter Plot of Age vs Total Marks
Total Marks