

**AMERICAN INTERNATIONAL UNIVERSITY–BANGLADESH (AIUB)**

**FACULTY OF SCIENCE AND TECHNOLOGY**

**MIDTERM PROJECT**

**INTRODUCTION TO DATA SCIENCE**

**Summer 2022-2023 Section: B**

**Submitted By**

Name: Md. Nasifur Rahman

ID: 20-43651-2

**Submitted To**

Touhedul Islam

Assistant Professor

Department of Computer Science

**Date of Submission:** July 17,2023

**Introduction:**

The Titanic dataset is a well-known and widely used dataset in the field of data analysis and machine learning. It provides information about the passengers who were aboard the Titanic during its maiden voyage in 1912. The Titanic dataset typically includes variables such as gender, age, siblings of the passenger (sibsp), parents / children aboard the Titanic (parch), passenger fare, port of embarkation, ticket class, categories to passengers(who), he was alone in ship or no (alone) and survival status. The dataset is to explore the factors that influenced the survival of passengers. By examining patterns and correlations within the dataset, researchers and data scientists can gain insights into the dynamics of the tragedy and potentially identify variables that played a significant role in determining survival.

**Import Dataset:**

**A screenshot of a computer

Description automatically generated**

**Output:**

**A screenshot of a computer

Description automatically generated**

**Explanation:**

* library(dplyr) loads the dplyr package, which provides a set of functions for data manipulation and transformation.
* read.csv() function is used to read the CSV file into R and create a data frame, which is assigned to the variable dataframe.
* header = TRUE argument indicates that the first row of the CSV file contains column names.
* sep=',' specifies that the columns are separated by commas.
* summary() function to provide a summary of the data stored in the dataframe object.
* str() function provides a concise overview of the data frame's.

**Data Pre-processing:**

**Gender Section**

* **Detecting null values’ row number:**

****

**Output:**

****

**Explanation:** It is observed that “gender” section contains 4 null values.

* **Measure of Center Tendency:**

**A close up of text

Description automatically generated**

**Output:**

****

**Explanation:**

* + table() provides the number of frequencies for a certain Gender.
  + which.max() gives the max frequency value and associated Gender.
  + names() extracts the maximum frequency.
  + paste() function mainly combines the string and numeric data type for the output.
* **Replacing null values with mode:**

****

**Output:**

**A close up of text

Description automatically generated**

**Explanation:**

* ifelse() is to allow conditional operations on dataframe .
* mutate() if for modifying existing dataframe information with the median .
* summary() calculates a summary of the modified "Gender" column in the dataframe.
* **Histogram Plot:**

****

* **Output:**

**A graph of a person and person

Description automatically generated**

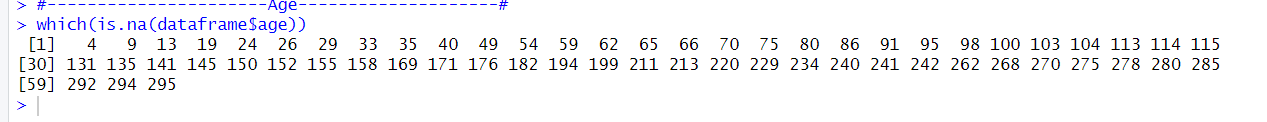
* **Explanation:**

The histogram illustrates two bars representing "0" and "1," with the bar corresponding to "0" having double the frequency compared to the bar representing "1." This pattern suggests an imbalance or disparity in the occurrence of the two values within the dataset.

**Age Section**

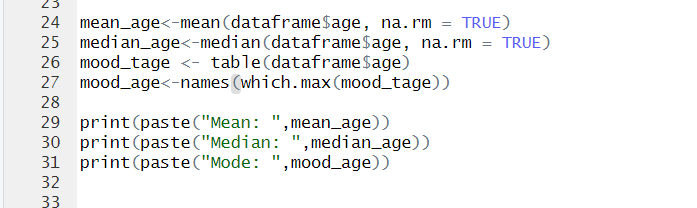
* **Detecting null values’ row number:**

**Output:**

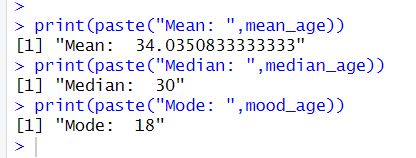
****

**Explanation:** It is observed that “Age” section contains 4 null values.

* **Measure of Center Tendency:**

****

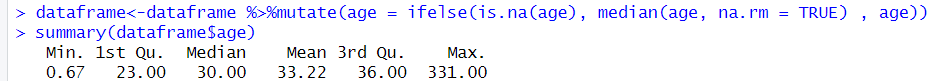
**Output:**

****

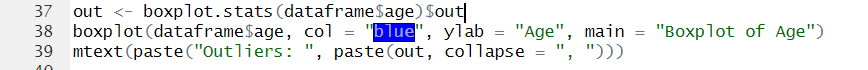
**Explanation:**

* + na.rm == TRUE is used to skip the null (N/A) values.
  + table() provides the number of frequencies for a certain Age.
  + which.max() gives the max frequency value and associated Age .
  + names() extracts the maximum frequency .
  + paste() function mainly combines the string and numeric data type for the output.
* **Replacing null values with Median:**

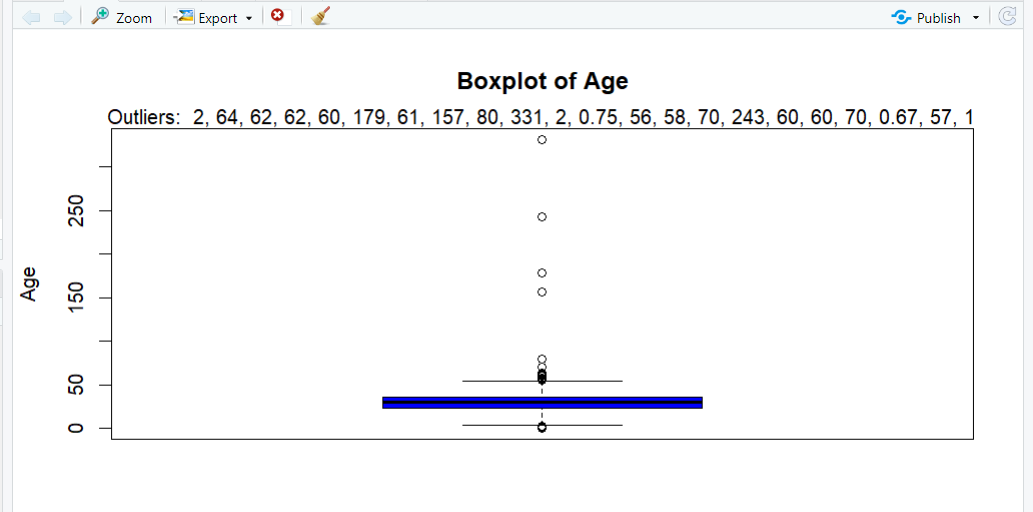
**Output:**

**Explanation:**

* ifelse() is to allow conditional operations on dataframe .
* mutate() if for modifying existing dataframe information with the median .
* summary() calculates a summary of the modified "age" column in the dataframe.
* **Outliers Detection:**

****

**Output:**

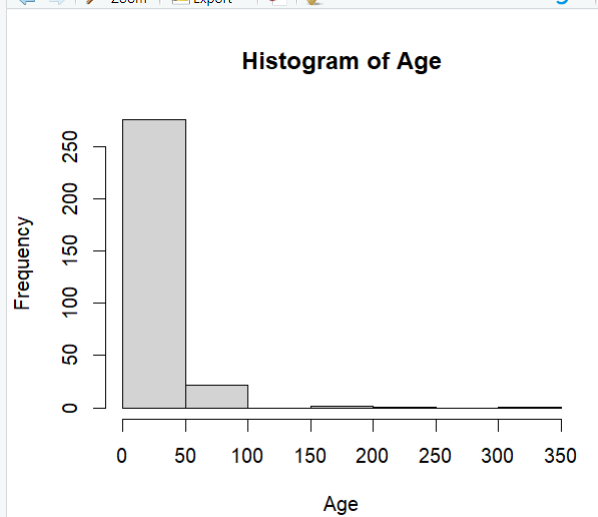
****

**Explanation:**

* boxplot.stats() function calculates the outliers.
* boxplot() – takes a list and shows it’s boxplot. col = "blue" argument sets the color of the boxplot to blue, ylab argument sets the label for the y-axis and main used to set the title.
* mtext() function is used to place text on the margin of the plot.
* **Graph Plot:**

****

**Output:**

****

**Explanation:**

In this histogram, the data exhibits a skewed distribution towards lower values. The mean value of 34 suggests that, on average, the data tends to be relatively higher. However, the median value of 30 indicates that half of the data falls below this point, implying a concentration of values towards the lower end. Additionally, the mode of 18 represents the most frequent value, suggesting a peak or high concentration of data points around this specific value.

**Sibsp Section**

* **Detecting null values’ row number:**

****

**Output:**

****

**Explanation:** It is observed that “Sibsp” has no null values.

* **Measure of Center Tendency:**

**A close-up of a computer code

Description automatically generated**

**Output:**

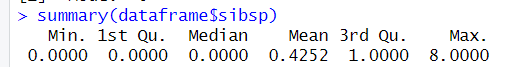
****

**Explanation:** “Sibsp” is categorical, so we calculate mood only.

* **Summary:**

****

**Output:**

****

* **Histogram Plot:**

**Output:**

****

**Explanation:**

In this specific histogram, the data appears to be concentrated around the value of 0. This suggests that most data points fall into this category, resulting in a prominent peak at 0 on the histogram. The distribution may indicate a significant prevalence or emphasis on values at or near zero in the dataset.

**Parch Section**

* **Detecting null values’ row number:**

****

**Output:**

****

**Explanation:** “Parch” has no null values.

* **Measure of Center Tendency:**

**A close up of text

Description automatically generated**

**Output:**

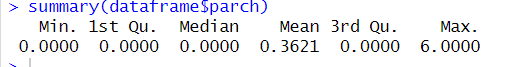
****

**Explanation:** “Sibsp” is categorical, so we calculate mood only.

* **Summary:**

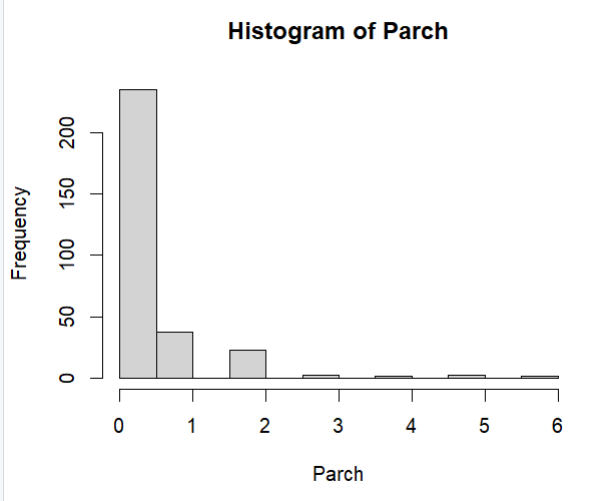
****

**Output:**

****

* **Histogram Plot:**

**Output:**

****

**Explanation:**

There is a distinct peak at a value of 0, indicating a high concentration of data points at that specific value. Additionally, the height of this peak is notably higher compared to the other values on the histogram, implying that there is a significant difference in frequency between the value of 0 and the other values present. This suggests that the data is skewed towards 0, with a much higher occurrence of that value compared to the rest of the data points.

**Fare Section**

* **Detecting null values’ row number:**

**A black screen with white text

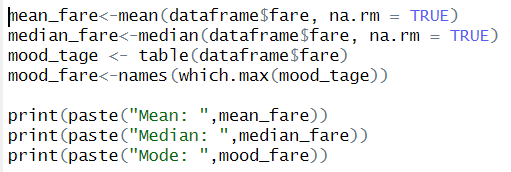
Description automatically generated**

**Output:**

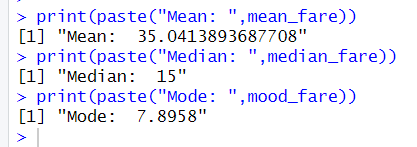
****

**Explanation:** “Fare” has no null values.

* **Measure of Center Tendency:**

****

**Output:**

****

* **Summary:**

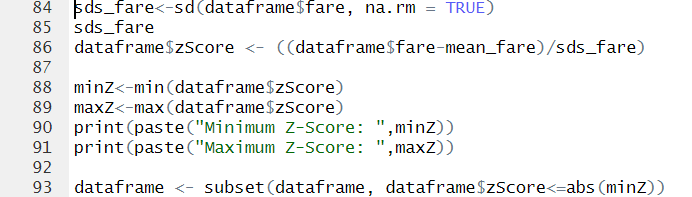
****

**Output:**

**A close up of a number

Description automatically generated**

* **Outliers Detection:**

****

**Output:**

**A number of numbers and letters

Description automatically generated**

**Explanation:**

* sd() used to calculate standard deviation.
* na.rm = TRUE argument specifies that any missing values should be ignored during the calculation.
* min() finds the minimum value.
* max() finds the maximum value.
* **Histogram:**

**A graph of a number of gray bars

Description automatically generated**

* **Explanation:**

The histogram shows variations in the frequencies of the different categories, indicating that the data is not evenly distributed. The shape of the histogram suggests that there might be some values that are more prominent or occur more frequently than others.

**Embarked Section**

* **Converting to numeric:**

****

* **Detecting null values’ row number:**

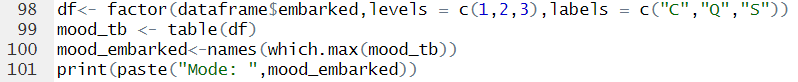
****

**Output:**

****

**Explanation:** “Embarked” has no null values.

* **Measure of Center Tendency:**

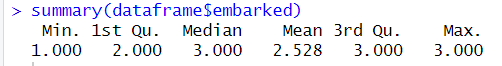
**Output:**

**A close up of a computer screen

Description automatically generated**

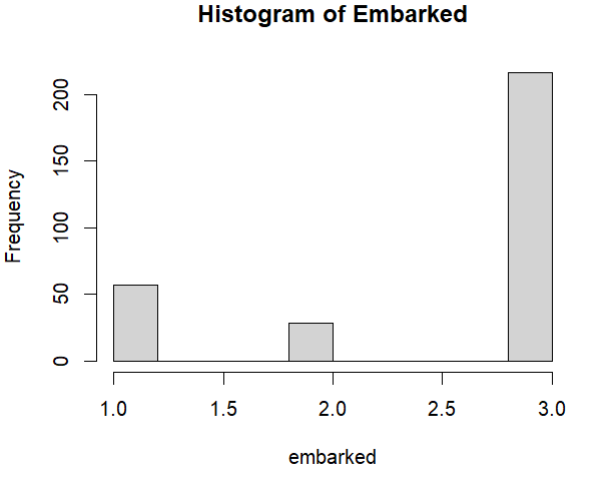
* **Replacing null values with Mood:**

**Output:**

****

**Explanation:** Embarked is categorical so we used mood to replace null values.

* **Graph:**

****

**Explanation:**

The histogram exhibits a notable peak in the frequency of the value "S" indicating its dominance over other values.

**Class Section**

* **Converting to numeric:**

****

* **Detecting null values’ row number:**

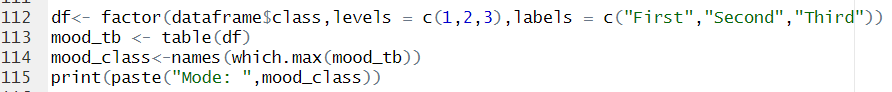
****

**Output:**

****

**Explanation:** It is observed, “Class” section contains 4 null values.

* **Measure of Center Tendency:**

**Output:**

****

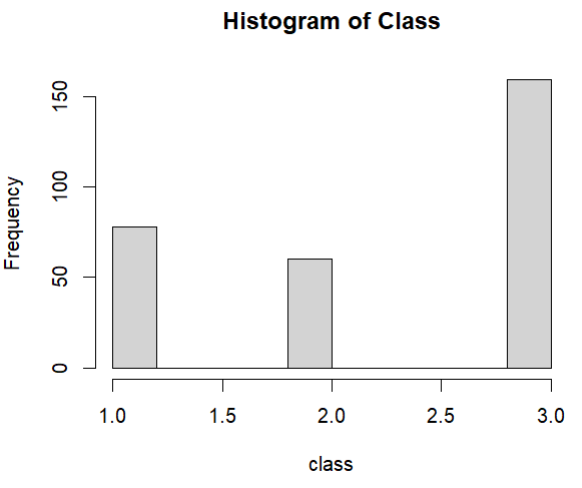
* **Replacing null values with Mood:**

**Output:**

****

**Explanation:**

* ifelse() is to allow conditional operations on dataframe .
* mutate() if for modifying existing dataframe information with the median .
* as.numeric() converts column to numeric format.
* **Graph:**

****

**Explanation:**

The histogram exhibits a notable peak in the frequency of Third indicating its dominance over other values.

**Who Section**

* **Converting to numeric:**

****

**Explanation:**

* substr() function is used to extract the first character of each value in the "who" column.
* ifelse() function is applied to check if the extracted first character is equal to "m".
* **Detecting null values’ row number:**

****

**Output:**

****

**Explanation:** There is no null values.

* **Measure of Center Tendency:**

****

**Output:**

****

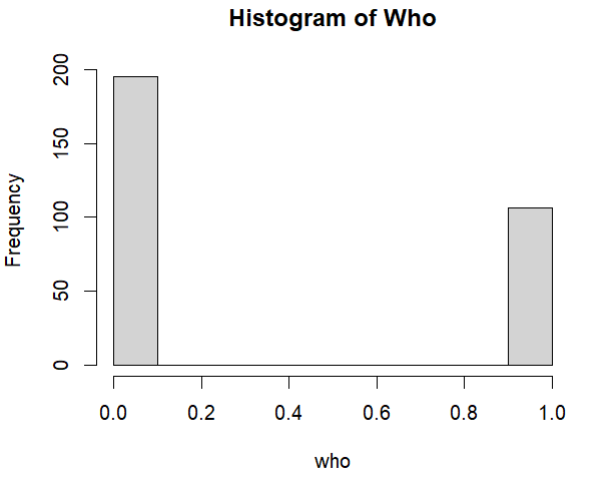
* **Summary:**

****

**Output:**

****

* **Graph:**

****

**Explanation:**

The histogram reveals a higher frequency of male travelers compared to female travelers on the Titanic.

**Alone Section**

* **Converting to numeric:**

****

**Explanation:**

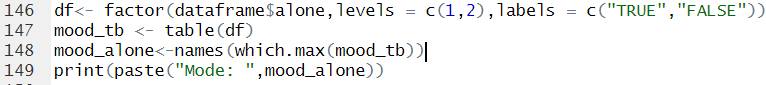
* factor() allows explicitly represent categorical data with ordered levels and subsequently encode them as numeric values if needed.
* **Detecting null values’ row number:**

****

**Output:**

****

* **Measure of Center Tendency:**

**Output:**

****

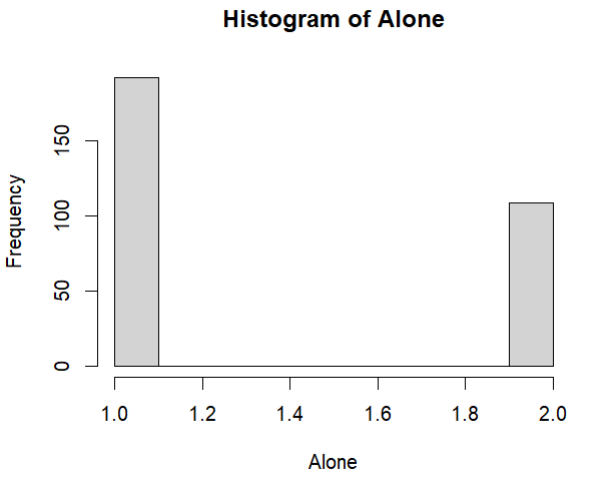
* **Summary:**

****

**Output:**

****

* **Graph:**

****

**Explanation:**

The histogram demonstrates that the frequency of "TRUE" is approximately twice as high as that of "FALSE."

**Survived Section**

* **Detecting null values’ row number:**

****

**Output:**

****

**Explanation:** Survived section has no null values.

* **Measure of Center Tendency:**

**A close-up of a computer code

Description automatically generated**

**Output:**

****

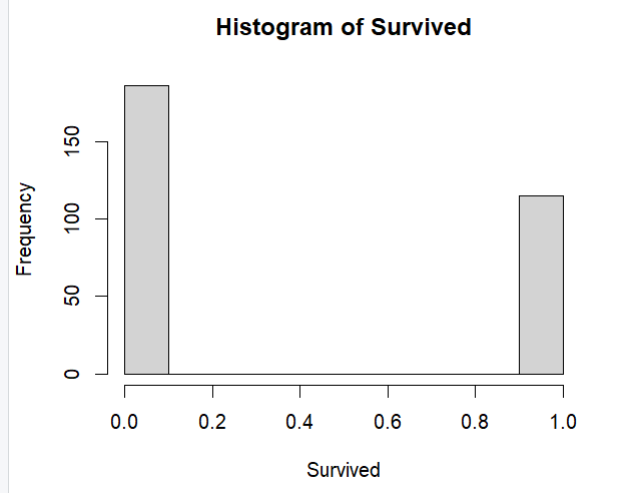
* **Summary:**

****

**Output:**

****

* **Graph:**

****

* **Explanation:**

The histogram illustrates that the frequency of the value "0" is higher than that of "1" that represents maximum people can’t survive.