**Step 1: Loading the Titanic Dataset**

library(tidyverse) # for data manipulation and visualization

> library(mice) # for imputation

#Load the titanic dataset

titanic.dataset <- read.csv("C:/Users/swamn/Downloads/titanic dataset.csv")

> View(titanic.dataset)

> # Display the first few rows of the dataset

> head(titanic.dataset)

PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket

1 892 0 3 Kelly, Mr. James male 34.5 0 0 330911

2 893 1 3 Wilkes, Mrs. James (Ellen Needs) female 47.0 1 0 363272

3 894 0 2 Myles, Mr. Thomas Francis male 62.0 0 0 240276

4 895 0 3 Wirz, Mr. Albert male 27.0 0 0 315154

5 896 1 3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0 1 1 3101298

6 897 0 3 Svensson, Mr. Johan Cervin male 14.0 0 0 7538

Fare Cabin Embarked

1 7.8292 Q

2 7.0000 S

3 9.6875 Q

4 8.6625 S

5 12.2875 S

6 9.2250 S

**Step 2: Data Cleaning**

> # Check for missing values

> colSums(is.na(titanic.dataset))

PassengerId Survived Pclass Name Sex Age SibSp Parch Ticket

0 0 0 0 0 86 0 0 0

Fare Cabin Embarked

1 0 0

**Step 3: Exploratory Data Analysis (EDA)**

*Summary Statistics:*

# Summary statistics for numerical variables

> summary(titanic.dataset[c("Age", "Fare")])

Age Fare

Min. : 0.17 Min. : 0.000

1st Qu.:23.00 1st Qu.: 7.896

Median :30.27 Median : 14.454

Mean :30.27 Mean : 35.627

3rd Qu.:35.75 3rd Qu.: 31.500

Max. :76.00 Max. :512.329

NA's :1

> # Summary statistics for categorical variables

> table(titanic.dataset$Sex)

female male

152 266

> table(titanic.dataset$Pclass)

1 2 3

107 93 218

> table(titanic.dataset$Survived)

0 1

266 152

> table(titanic.dataset$Embarked)

C Q S

102 46 270

*Visualizations:*

***Visualizing Survival by Gender*:**

|  |
| --- |
| # Bar plot of survival by gender using base R graphics |
|  |
| |  | | --- | |  | |

barplot(table(titanic.dataset$Survived, titanic.dataset$Sex),

+ beside = TRUE,

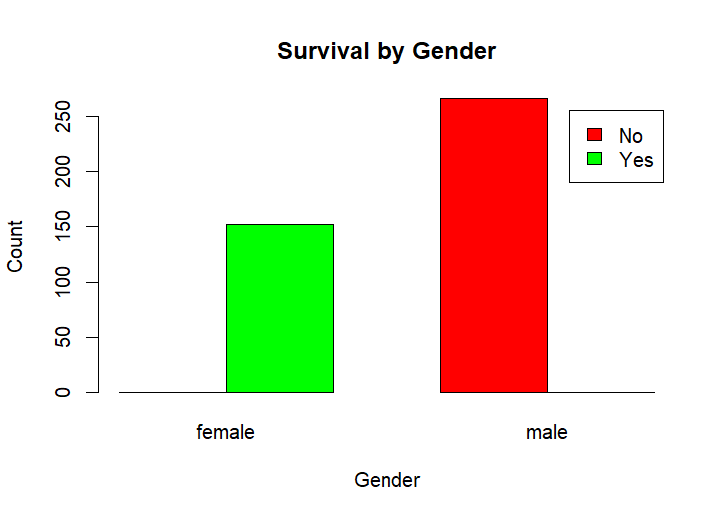
+ col = c("red", "green"),

+ legend = c("No", "Yes"),

+ main = "Survival by Gender",

+ xlab = "Gender",

+ ylab = "Count")



***Visualizing Survival by Passenger Class:***

# Bar plot of survival by passenger class using base R graphics

> barplot(table(titanic.dataset$Survived, titanic.dataset$Pclass),

+ beside = TRUE,

+ col = c("red", "green"),

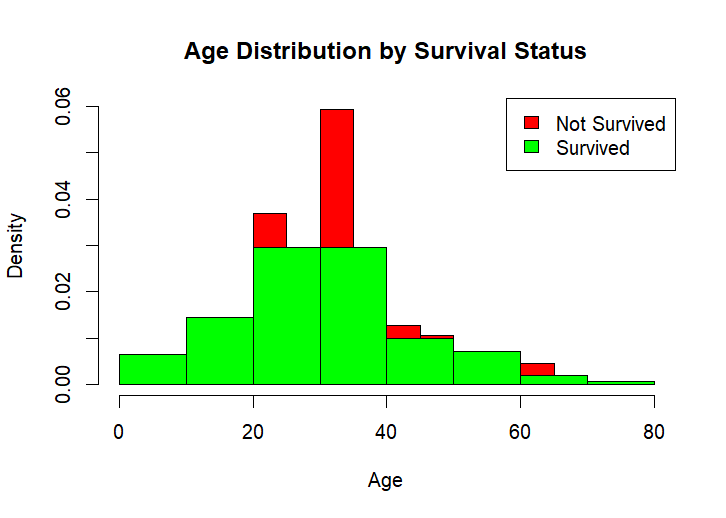
+ legend = c("No", "Yes"),

+ main = "Survival by Passenger Class",

+ xlab = "Passenger Class",

+ ylab = "Count")

|  |  |  |
| --- | --- | --- |
| # Density plot of age d *Visualizing Age Distribution by Survival Status:* #Age distribution by survival status using base R graphics  > par(mfrow = c(1, 1)) # reset the layout  > hist(titanic.dataset$Age[titanic.dataset$Survived == 0],  + col = "red",  + xlim = c(0, 80),  + main = "Age Distribution by Survival Status",  + xlab = "Age",  + ylab = "Density",  + freq = FALSE)  > hist(titanic.dataset$Age[titanic.dataset$Survived == 1],  + col = "green",  + add = TRUE,  + freq = FALSE)  > legend("topright", legend = c("Not Survived", "Survived"), fill = c("red", "green"))   |  |  | | --- | --- | | |  | | --- | |  | | |
|  |
|  |



### ***Step 4: Insights and Conclusion:***

Based on the visualizations and summary statistics:

* **Survival by Gender**: Females had a higher survival rate compared to males.
* **Survival by Passenger Class**: Passengers in higher classes (1st class) had a higher survival rate.
* **Age Distribution by Survival Status**: There is a peak in survival among children and young adults.
* **Survival by Embarked Port**: Passengers who embarked from port C had a higher survival rate compared to others.

These insights provide a preliminary understanding of the Titanic dataset and help identify trends and patterns in the data. Further analysis could involve deeper statistical tests or machine learning models to explore relationships more rigorously.