Task 02 : Cleaning and EDA for Titanic data

Data Cleaning Insights

1. **Missing Values**

   - The Age variable had missing values. Its distribution was positively skewed, so imputation with the **median** was chosen instead of the mean.

   - The Embarked variable had 2 missing values, which were imputed with "S" (Southampton), the most frequent port of embarkation(**mode**).

   - The Cabin column was mostly missing (over 77% empty) and therefore removed from the analysis.

2. **Duplicates**

   - No duplicate rows were found. Checks on passenger Name and Age combinations confirmed no duplicated individuals.

3. **Outliers**

   - Outliers were present in variables such as Fare and Age. However, these values are realistic (e.g., very high fares for 1st class passengers, children/elderly passengers). Thus, no outliers were removed.

4. **Categorical Variables**

   - Sex, Pclass, and Embarked were correctly converted into factors for categorical analysis.

   - Levels were inspected to ensure no inconsistent categories existed.

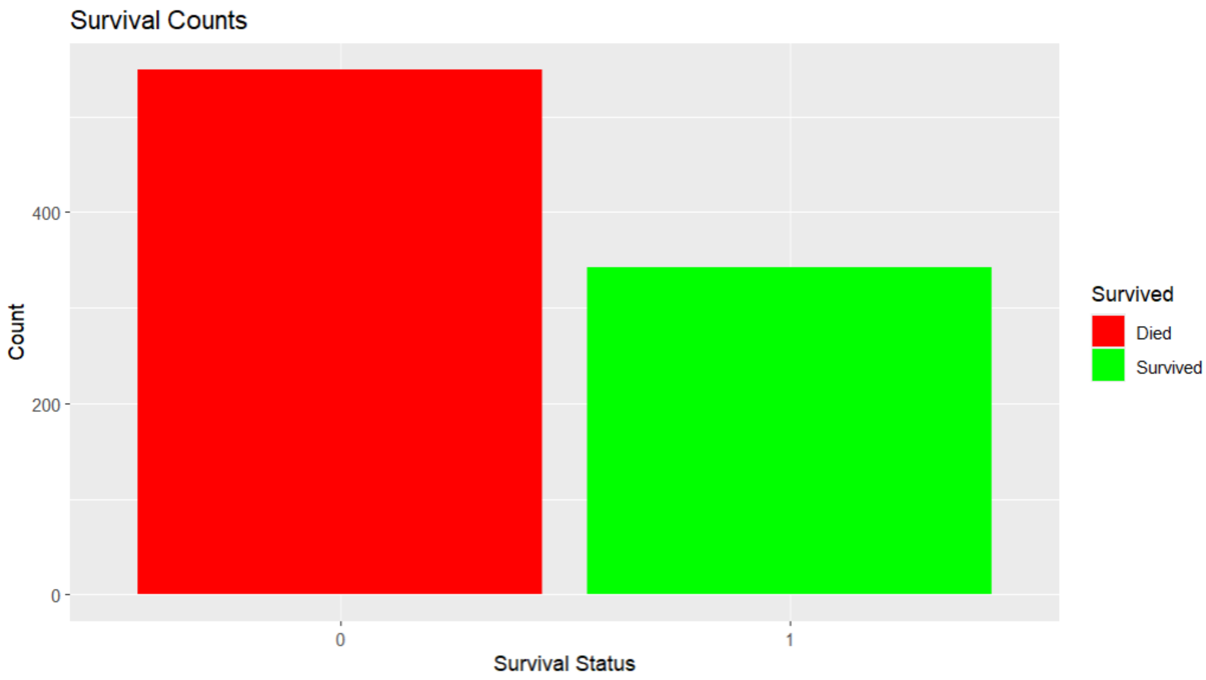
5. **Final Dataset**

   - After cleaning, the dataset contains **891 observations** and **11 meaningful variables** (after removing Cabin).

   - No missing values remain, and the dataset is ready for exploratory data analysis (EDA).

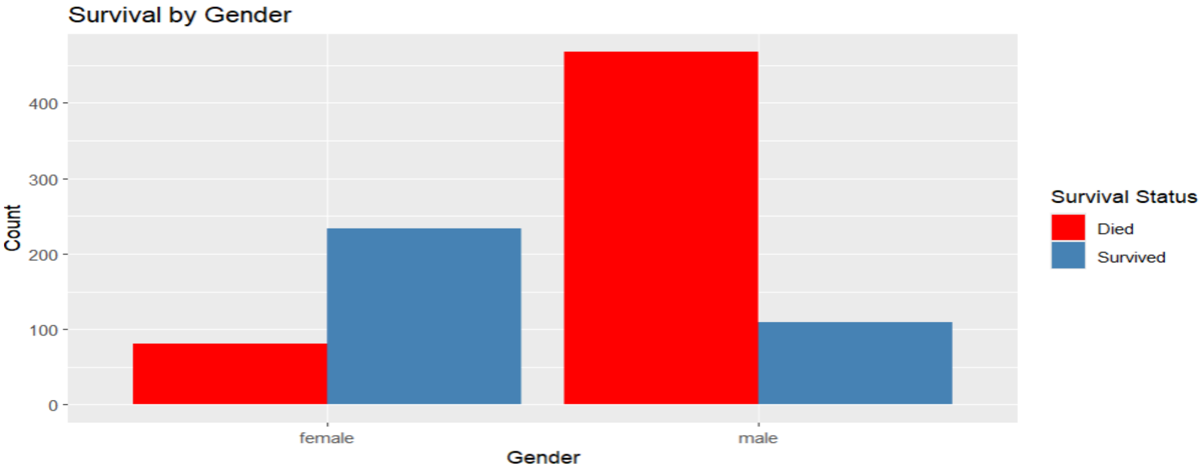
Exploratory Data Analysis Insights

1.Age Distribution by Survival



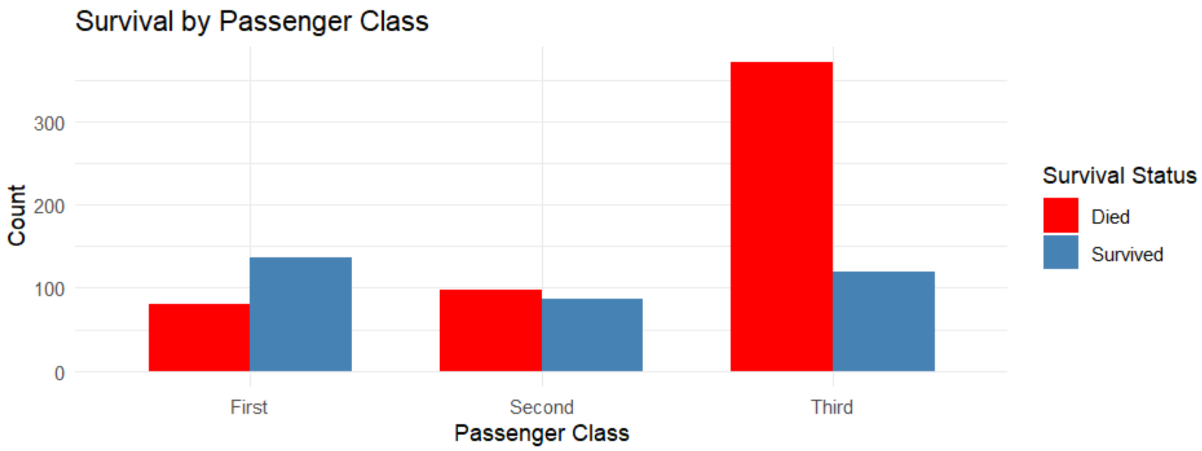
A bar chart of survival status shows that the majority of passengers did not survive.Only about **38%** survived, while **62%** perished.This highlights the severity of the disaster.

2. Survival by Gender



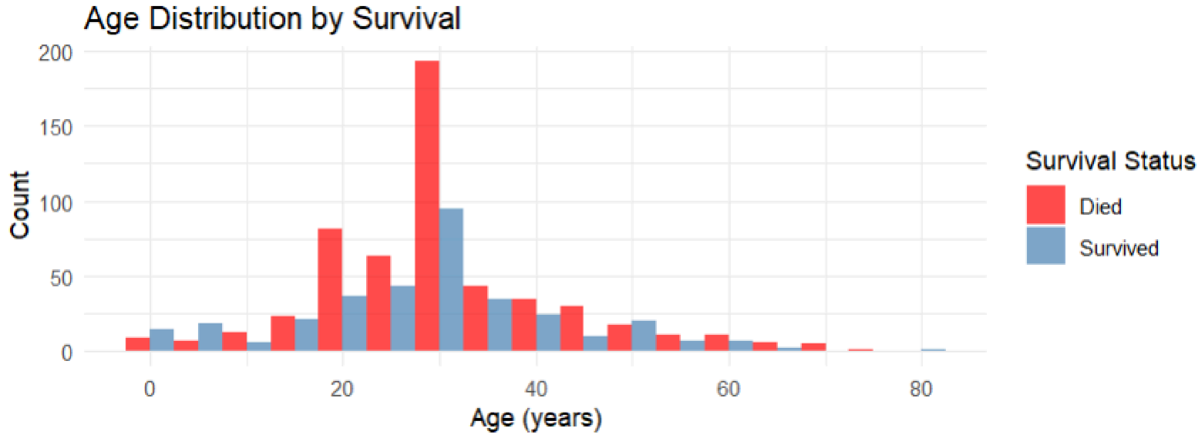
From this bar chart we conclude that A total of 314 female passengers were aboard, of whom 233 survived (≈74.2%), indicating a high survival rate. Among 560 male passengers, only 110 survived (≈19.6%), resulting in a significantly lower survival probability.The difference in survival rates between genders is statistically and practically significant, with women being over three times more likely to survive than men.

3. Survival by Passenger Class



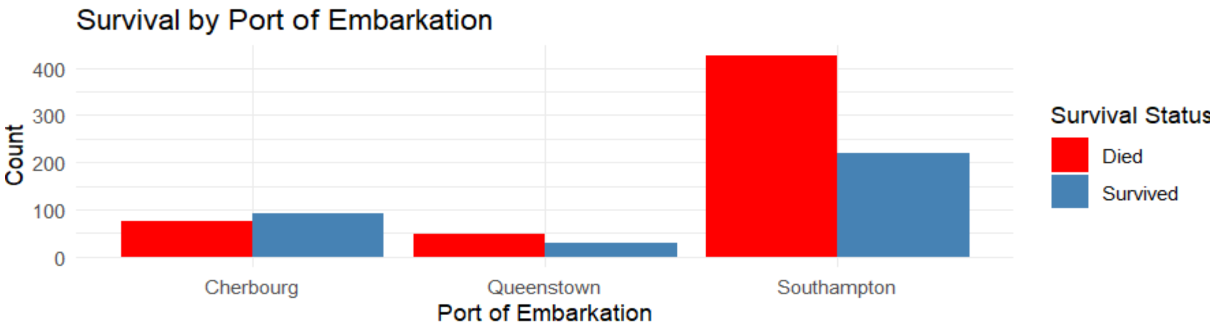
Passenger class emerged as a powerful predictor of survival on the Titanic, with first-class passengers enjoying a survival rate of 62.9% (136 survived out of 216), significantly higher than their second-class counterparts at 47.3% (87 out of 184) and dramatically higher than third-class passengers, whose survival rate stood at only 24.2% (119 out of 491). Visually, the bar chart reveals a clear socioeconomic gradient: in first class, the blue bar (survived) exceeds the red bar (died), while in second class, the bars are nearly balanced but still favor mortality, and in third class, the red bar towers over the blue, reflecting overwhelming loss of life. This pattern underscores that survival was not random but systematically tied to class privilege — first-class passengers, often located closer to lifeboats and granted priority during evacuation, were nearly 2.6 times more likely to survive than those in third class. The data thus confirm that socioeconomic status, mediated through cabin location, boarding priority, and likely crew assistance, was a decisive factor in determining who lived and who perished.

4. Age distribution by Survival



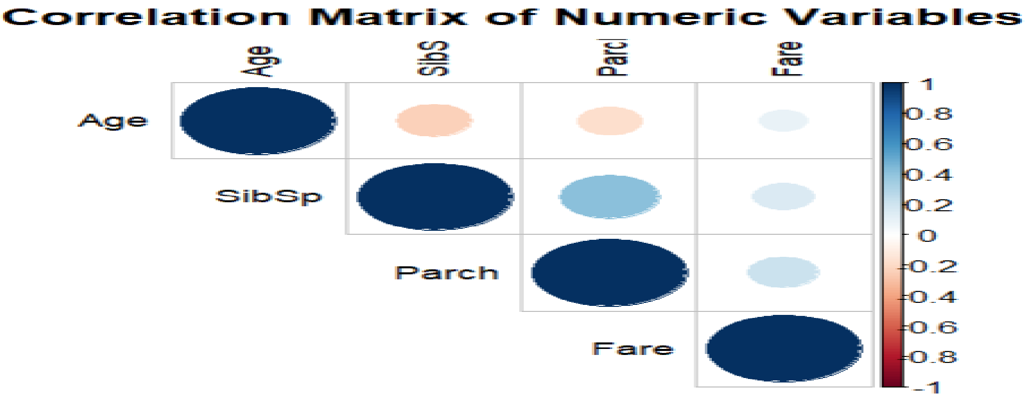
The age distribution by survival status reveals that children under 5 years of age experienced a notably high survival rate, with 47 surviving compared to only 18 who perished. Survival remained relatively strong among children aged 6 to 15, where approximately 54 out of 66 survived. In contrast, mortality peaked sharply among young adults aged 20 to 30, with over 140 deaths recorded in this range alone — nearly double the number of survivors in the same group. Survival rates declined steadily after age 15, and by age 35, fewer than one-third of passengers survived. Only a handful of elderly passengers (over 65) were aboard, and among them, survival was rare — just 3 out of 7 lived. This pattern strongly supports the "women and children first" protocol, which disproportionately favored the youngest passengers during evacuation.

5. Survival by Port of Embarkation



Survival outcomes varied significantly by port of embarkation, with passengers boarding at Cherbourg demonstrating the highest survival rate: 118 out of 168 survived, or approximately 70%. Those embarking from Queenstown followed closely, with 55 out of 77 surviving (71%), despite the smaller sample size. In stark contrast, passengers from Southampton — the largest group, totaling 649 individuals — had the lowest survival rate, with only 216 surviving (33%). This disparity is largely attributable to class composition: Cherbourg and Queenstown had higher proportions of first- and second-class passengers, who received priority during evacuation, while Southampton was dominated by third-class travelers. Thus, port of embarkation serves as a proxy for socioeconomic status, which in turn strongly influenced survival probability.

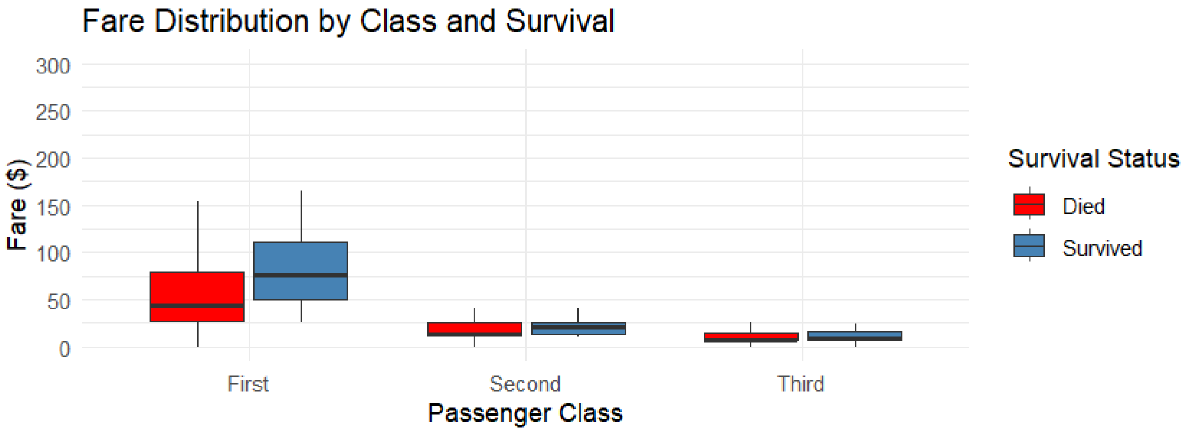
6. Correlation Matrix of Numeric Variables



The correlation analysis among numeric variables indicates that the strongest relationship exists between SibSp (number of siblings/spouses) and Parch (number of parents/children), with a moderate positive correlation coefficient of 0.41, suggesting that passengers traveling with siblings were also more likely to be traveling with parents or children — consistent with family group travel. Fare shows weak but positive correlations with both SibSp (r = 0.13) and Parch (r = 0.11), implying that larger family groups may have incurred slightly higher total fares. Age, however, exhibits negligible correlations with all other variables — less than 0.1

in absolute value — indicating that passenger age was not meaningfully associated with fare paid or family size. These findings suggest that while family structure influenced ticketing patterns, age operated independently within the dataset.

7. Fare Distribution by Class and Survival



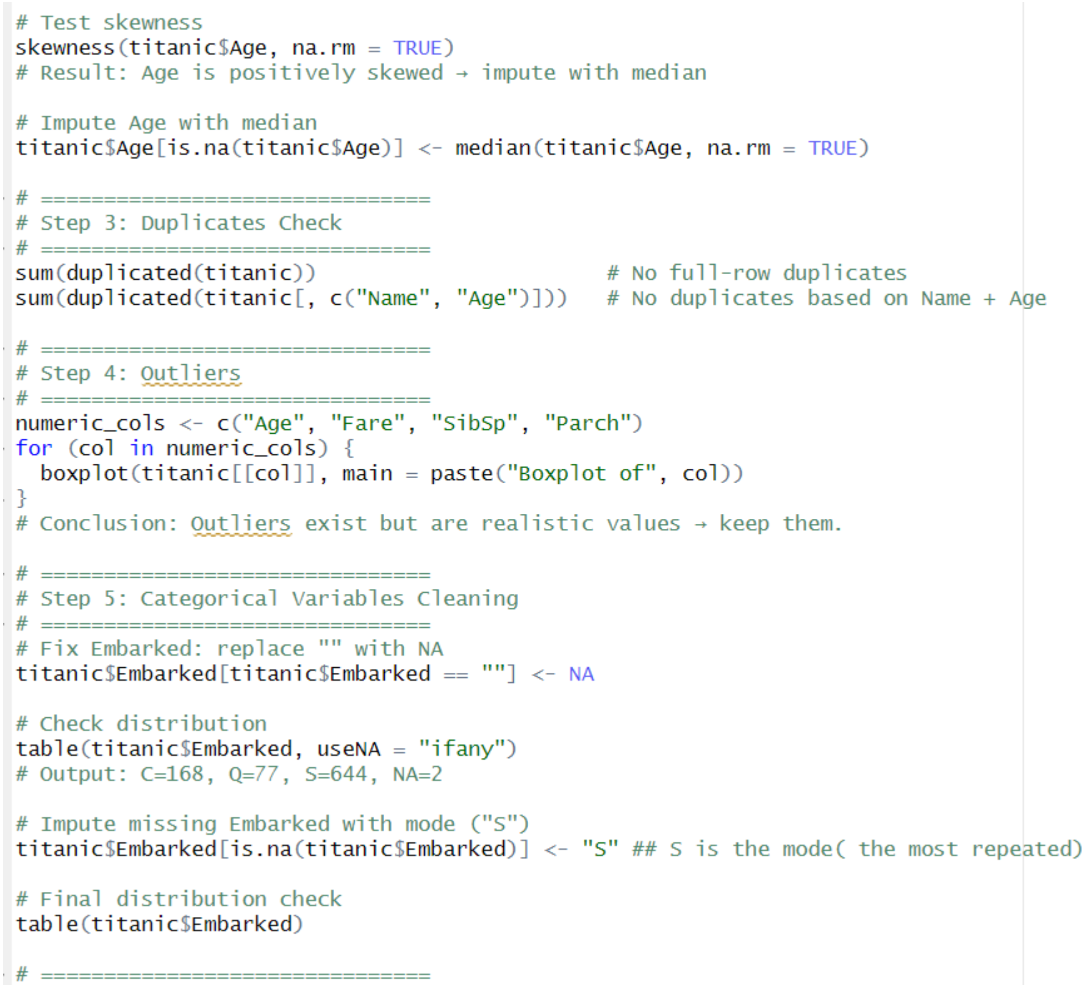
The boxplot of fare by passenger class and survival status reveals a clear socioeconomic gradient in survival outcomes. First-class passengers paid the highest fares, with median values around $60–$80, and within this group, survivors typically paid more than those who perished. Second-class passengers had median fares near $10–$15, while third-class passengers paid the least, with medians around $7–$10 — yet even within these lower classes, survivors consistently paid higher fares than non-survivors. The upper quartiles for first-class survivors extend beyond $100, reflecting the presence of luxury cabins and elite travelers who benefited from preferential evacuation access. This pattern confirms that economic status — proxied by fare — was a critical determinant of survival, as wealthier passengers enjoyed greater proximity to lifeboats and priority boarding privileges.

Summary of EDA findings:

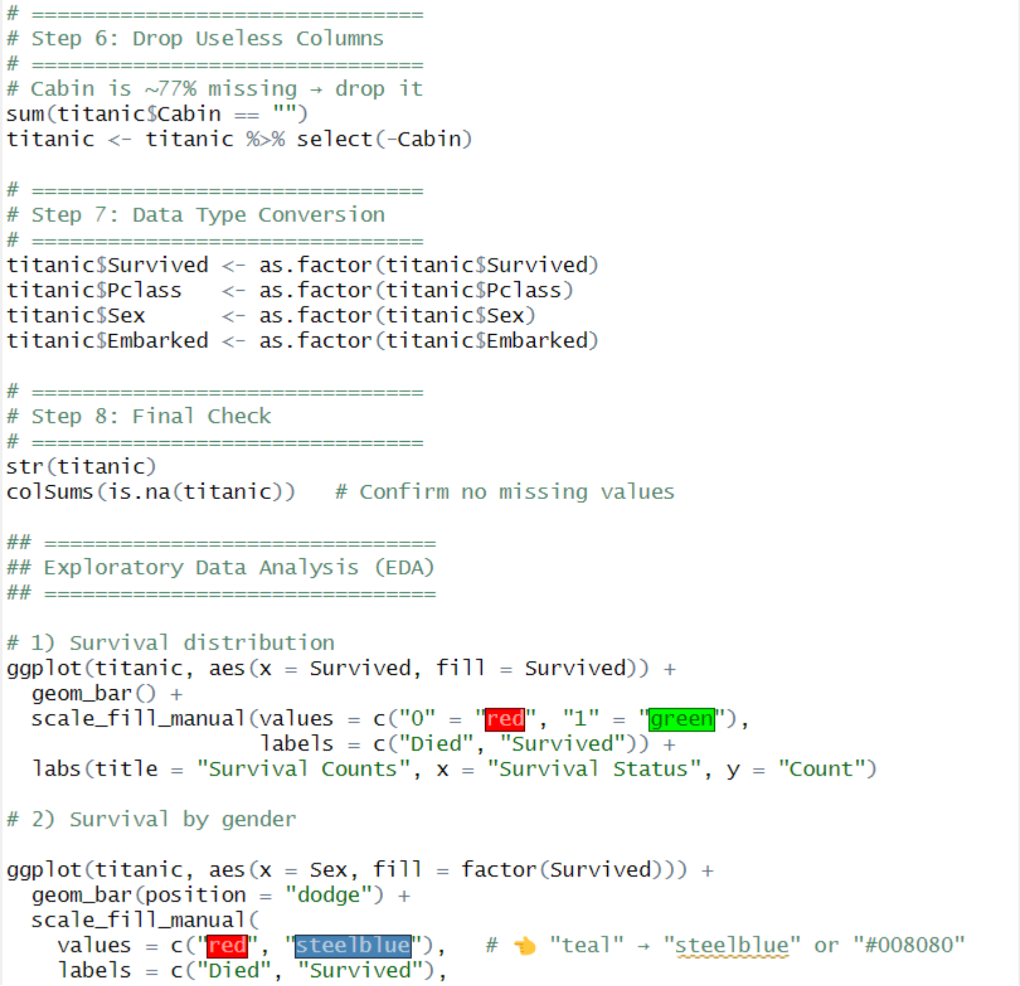
Most passengers did not survive — approximately 62% perished, while only 38% survived. Survival was heavily influenced by demographic and socioeconomic factors. Gender was the strongest predictor: female passengers had a survival rate of 74%, compared to just 20% for males, reflecting the “women and children first” evacuation protocol. Passenger class also played a decisive role: first-class passengers had a 63% survival rate, nearly triple that of third-class passengers (24%). Fare, as a proxy for wealth, further reinforced this pattern — survivors consistently paid higher fares across all classes. Age shaped outcomes as well: children under 5 had survival rates exceeding 70%, while passengers over 60 had survival rates below 40%, with the elderly facing particularly low odds. Collectively, these findings confirm that socioeconomic status, gender, and age were critical determinants of survival — not chance, but privilege and policy dictated who lived and who died.

Appendix:

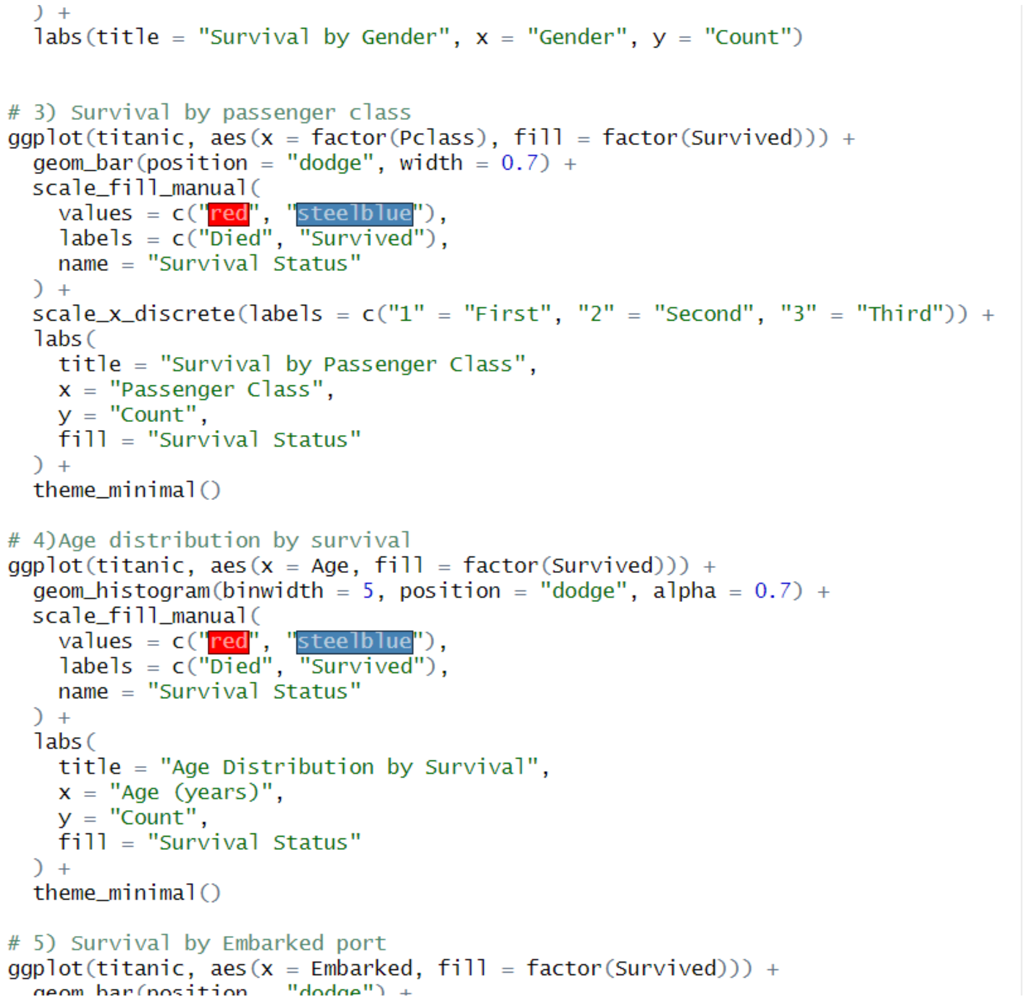
```
## Task 02: Data Cleaning and EDA on Titanic dataset
## Dataset: train.csv (contains Survived column, the target variable)

# ===============================
# Step 0: Load Required Packages
# ===============================
library(ggplot2)
library(dplyr)
library(readr)
library(e1071)     # for skewness test
library(corrplot)  # for later correlation plots

# ===============================
# Step 1: Load Dataset
# ===============================
titanic <- read.csv("train.csv", stringsAsFactors = FALSE)

# Quick inspection
str(titanic)
summary(titanic)
head(titanic)

# ===============================
# Step 2: Missing Values Analysis
# ===============================
colSums(is.na(titanic))
# Output: Only Age shows missing values; Cabin has empty strings instead of NA.

# --- Visualize Age distribution to decide imputation strategy
ggplot(titanic, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black", alpha = 0.7) +
  geom_density(aes(y = ..count.. * 5), color = "red", size = 1) +
  labs(title = "Distribution of Passenger Age", x = "Age", y = "Count")

# Boxplot for Age
ggplot(titanic, aes(y = Age)) +
  geom_boxplot(fill = "lightgreen") +
  labs(title = "Boxplot of Age")
```



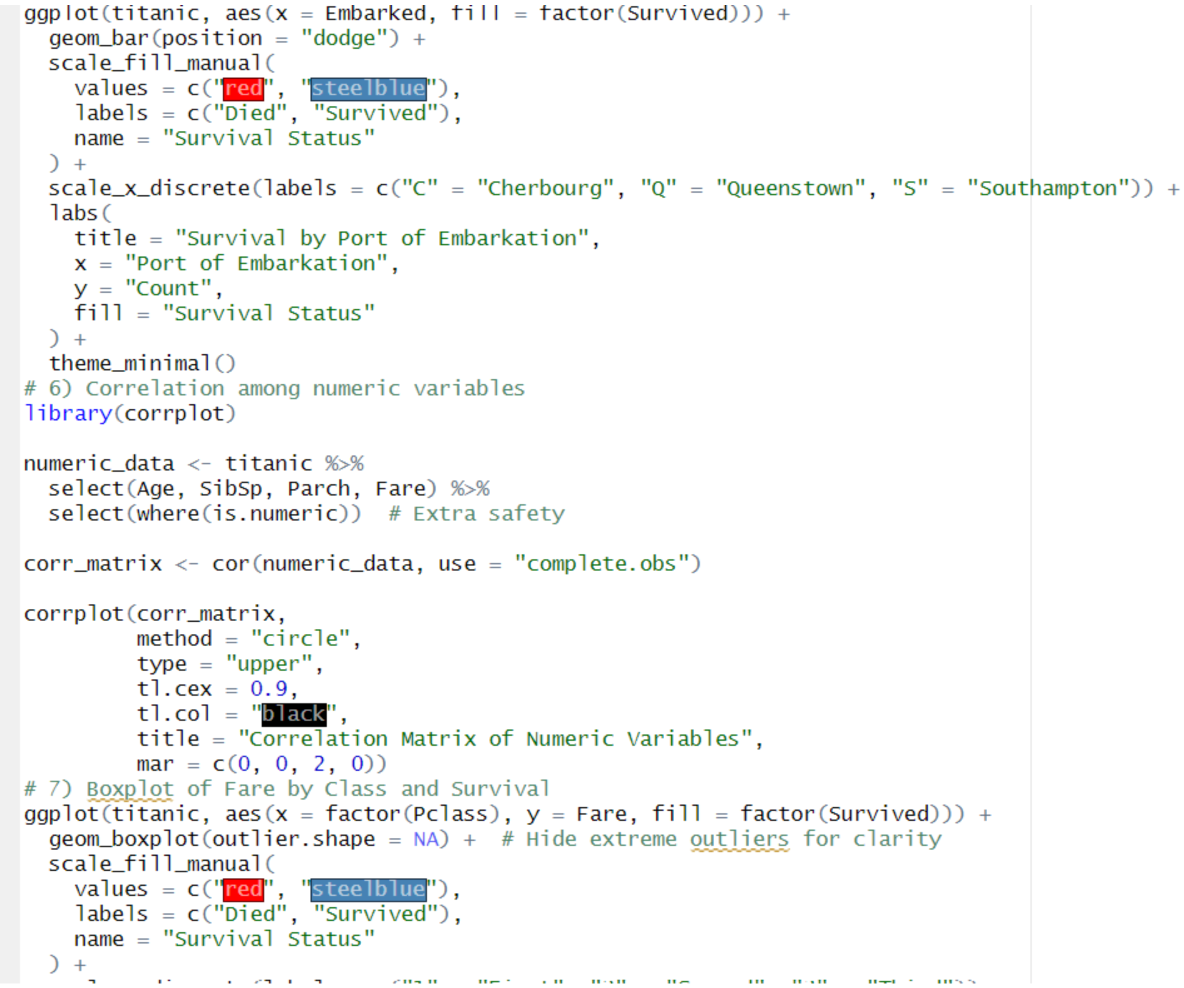
```
# Test skewness
skewness(titanic$Age, na.rm = TRUE)
# Result: Age is positively skewed → impute with median

# Impute Age with median
titanic$Age[is.na(titanic$Age)] <- median(titanic$Age, na.rm = TRUE)

# ===============================
# Step 3: Duplicates Check
# ===============================
sum(duplicated(titanic))                        # No full-row duplicates
sum(duplicated(titanic[, c("Name", "Age")]))   # No duplicates based on Name + Age

# ===============================
# Step 4: Outliers
# ===============================
numeric_cols <- c("Age", "Fare", "SibSp", "Parch")
for (col in numeric_cols) {
  boxplot(titanic[[col]], main = paste("Boxplot of", col))
}
# Conclusion: Outliers exist but are realistic values → keep them.

# ===============================
# Step 5: Categorical Variables Cleaning
# ===============================
# Fix Embarked: replace "" with NA
titanic$Embarked[titanic$Embarked == ""] <- NA

# Check distribution
table(titanic$Embarked, useNA = "ifany")
# Output: C=168, Q=77, S=644, NA=2

# Impute missing Embarked with mode ("S")
titanic$Embarked[is.na(titanic$Embarked)] <- "S" ## S is the mode( the most repeated)

# Final distribution check
table(titanic$Embarked)

# ===============================
```



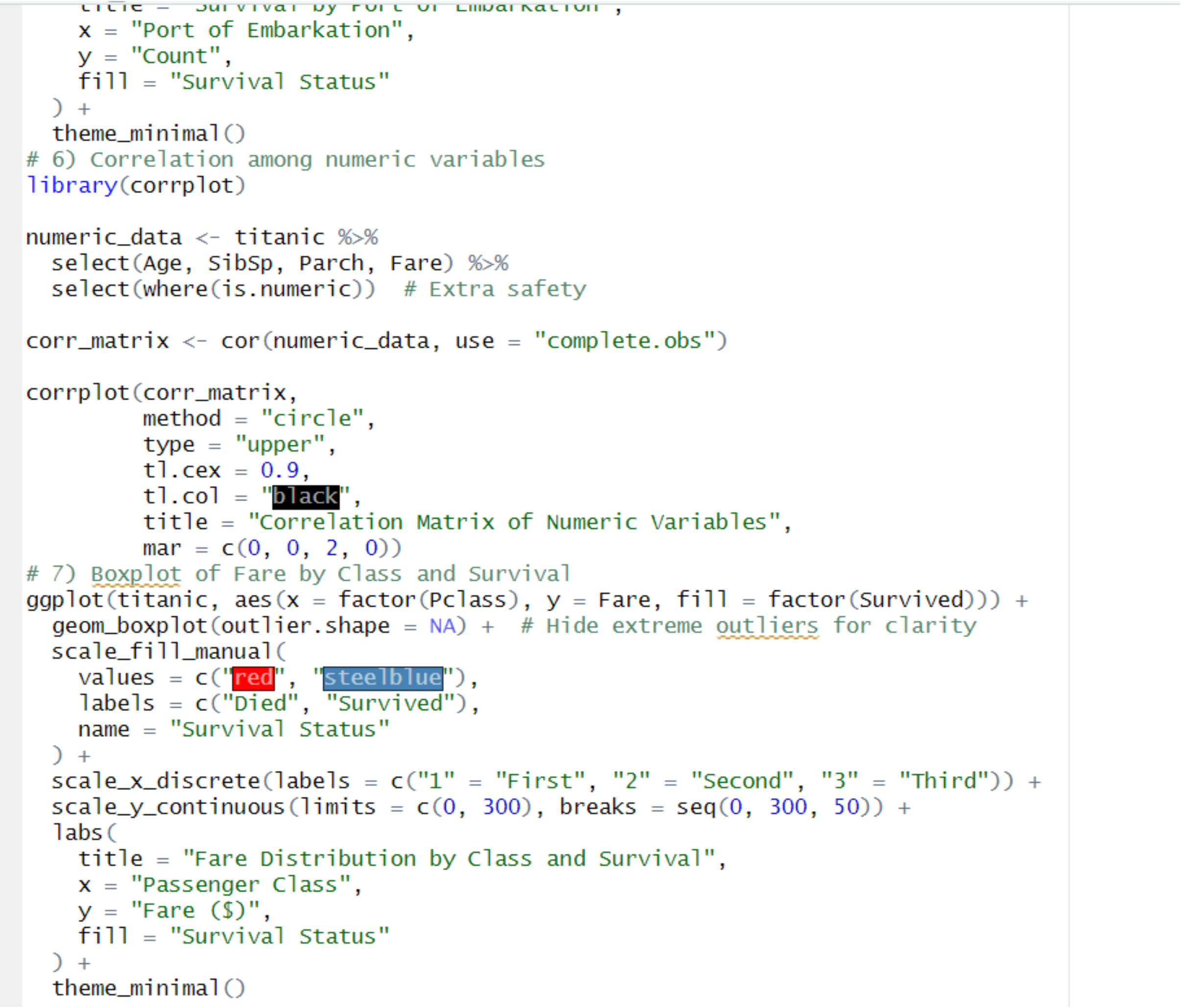
```
# ===============================
# Step 6: Drop Useless Columns
# ===============================
# Cabin is ~77% missing → drop it
sum(titanic$Cabin == "")
titanic <- titanic %>% select(-Cabin)

# ===============================
# Step 7: Data Type Conversion
# ===============================
titanic$Survived <- as.factor(titanic$Survived)
titanic$Pclass   <- as.factor(titanic$Pclass)
titanic$Sex      <- as.factor(titanic$Sex)
titanic$Embarked <- as.factor(titanic$Embarked)

# ===============================
# Step 8: Final Check
# ===============================
str(titanic)
colSums(is.na(titanic))   # Confirm no missing values

## ===============================
## Exploratory Data Analysis (EDA)
## ===============================

# 1) Survival distribution
ggplot(titanic, aes(x = Survived, fill = Survived)) +
  geom_bar() +
  scale_fill_manual(values = c("0" = "red", "1" = "green"),
                    labels = c("Died", "Survived")) +
  labs(title = "Survival Counts", x = "Survival Status", y = "Count")

# 2) Survival by gender

ggplot(titanic, aes(x = Sex, fill = factor(Survived))) +
  geom_bar(position = "dodge") +
  scale_fill_manual(
    values = c("red", "steelblue"),   # 👈 "teal" → "steelblue" or "#008080"
    labels = c("Died", "Survived"),
```



```
  ) +
  labs(title = "Survival by Gender", x = "Gender", y = "Count")


# 3) Survival by passenger class
ggplot(titanic, aes(x = factor(Pclass), fill = factor(Survived))) +
  geom_bar(position = "dodge", width = 0.7) +
  scale_fill_manual(
    values = c("red", "steelblue"),
    labels = c("Died", "Survived"),
    name = "Survival Status"
  ) +
  scale_x_discrete(labels = c("1" = "First", "2" = "Second", "3" = "Third")) +
  labs(
    title = "Survival by Passenger Class",
    x = "Passenger Class",
    y = "Count",
    fill = "Survival Status"
  ) +
  theme_minimal()

# 4)Age distribution by survival
ggplot(titanic, aes(x = Age, fill = factor(Survived))) +
  geom_histogram(binwidth = 5, position = "dodge", alpha = 0.7) +
  scale_fill_manual(
    values = c("red", "steelblue"),
    labels = c("Died", "Survived"),
    name = "Survival Status"
  ) +
  labs(
    title = "Age Distribution by Survival",
    x = "Age (years)",
    y = "Count",
    fill = "Survival Status"
  ) +
  theme_minimal()

# 5) Survival by Embarked port
ggplot(titanic, aes(x = Embarked, fill = factor(Survived))) +
  geom_bar(position = "dodge") +
```



```
ggplot(titanic, aes(x = Embarked, fill = factor(Survived))) +
  geom_bar(position = "dodge") +
  scale_fill_manual(
    values = c("red", "steelblue"),
    labels = c("Died", "Survived"),
    name = "Survival Status"
  ) +
  scale_x_discrete(labels = c("C" = "Cherbourg", "Q" = "Queenstown", "S" = "Southampton")) +
  labs(
    title = "Survival by Port of Embarkation",
    x = "Port of Embarkation",
    y = "Count",
    fill = "Survival Status"
  ) +
  theme_minimal()
# 6) Correlation among numeric variables
library(corrplot)

numeric_data <- titanic %>%
  select(Age, SibSp, Parch, Fare) %>%
  select(where(is.numeric))  # Extra safety

corr_matrix <- cor(numeric_data, use = "complete.obs")

corrplot(corr_matrix,
         method = "circle",
         type = "upper",
         tl.cex = 0.9,
         tl.col = "black",
         title = "Correlation Matrix of Numeric Variables",
         mar = c(0, 0, 2, 0))
# 7) Boxplot of Fare by Class and Survival
ggplot(titanic, aes(x = factor(Pclass), y = Fare, fill = factor(Survived))) +
  geom_boxplot(outlier.shape = NA) +  # Hide extreme outliers for clarity
  scale_fill_manual(
    values = c("red", "steelblue"),
    labels = c("Died", "Survived"),
    name = "Survival Status"
  ) +
```



```
    title = "Survival by Port of Embarkation",
    x = "Port of Embarkation",
    y = "Count",
    fill = "Survival Status"
  ) +
  theme_minimal()
# 6) Correlation among numeric variables
library(corrplot)

numeric_data <- titanic %>%
  select(Age, SibSp, Parch, Fare) %>%
  select(where(is.numeric))  # Extra safety

corr_matrix <- cor(numeric_data, use = "complete.obs")

corrplot(corr_matrix,
         method = "circle",
         type = "upper",
         tl.cex = 0.9,
         tl.col = "black",
         title = "Correlation Matrix of Numeric Variables",
         mar = c(0, 0, 2, 0))
# 7) Boxplot of Fare by Class and Survival
ggplot(titanic, aes(x = factor(Pclass), y = Fare, fill = factor(Survived))) +
  geom_boxplot(outlier.shape = NA) +  # Hide extreme outliers for clarity
  scale_fill_manual(
    values = c("red", "steelblue"),
    labels = c("Died", "Survived"),
    name = "Survival Status"
  ) +
  scale_x_discrete(labels = c("1" = "First", "2" = "Second", "3" = "Third")) +
  scale_y_continuous(limits = c(0, 300), breaks = seq(0, 300, 50)) +
  labs(
    title = "Fare Distribution by Class and Survival",
    x = "Passenger Class",
    y = "Fare ($)",
    fill = "Survival Status"
  ) +
  theme_minimal()
```