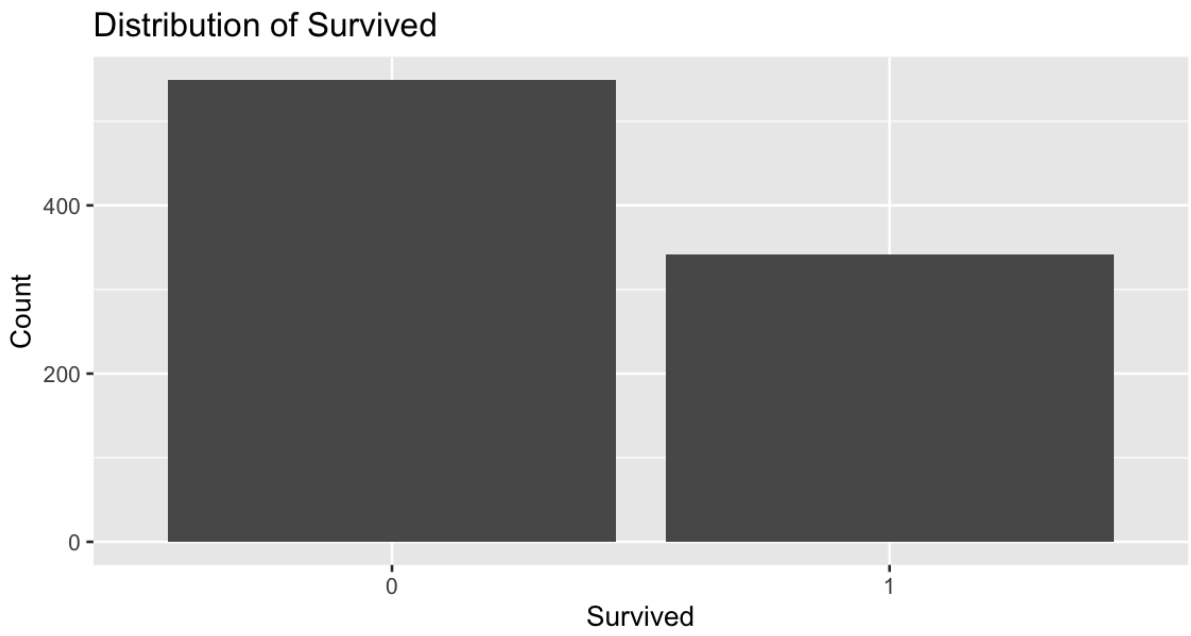
DATA CLEANING

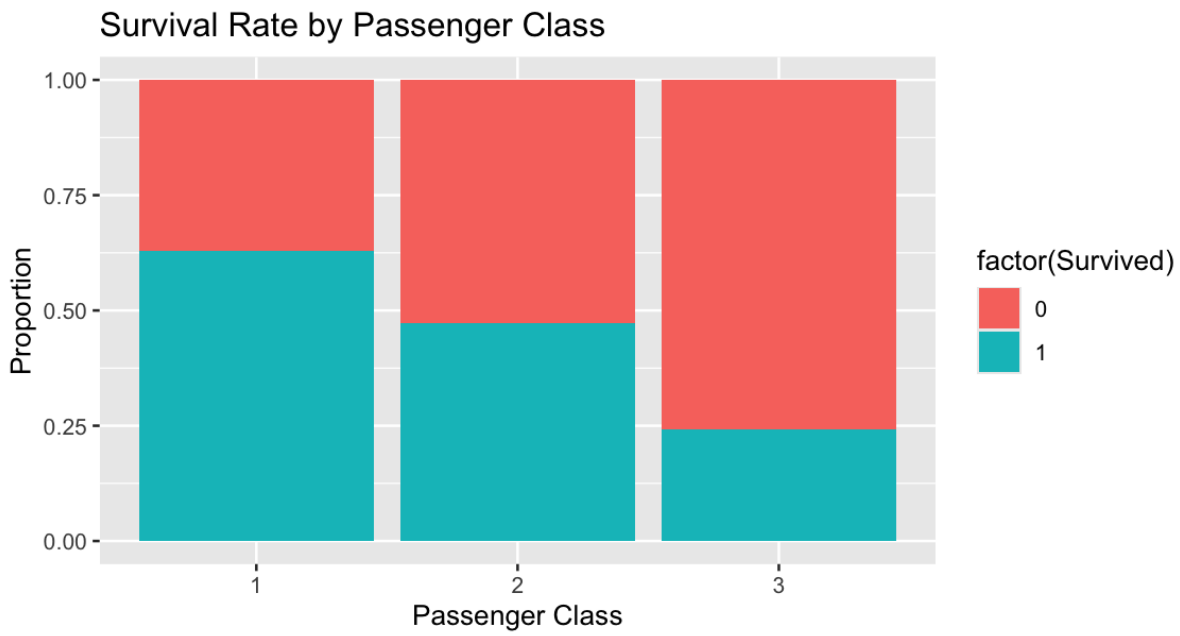
```
> train <- read.csv("~/Desktop/INTERNSHIP TASK/TASK 2/train.csv", stringsAsFactors=TRUE)
> View(train)
> df <- train
> head(df)
  PassengerId Survived Pclass                                                Name    Sex Age SibSp
1           1        0      3                             Braund, Mr. Owen Harris   male  22     1
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
3           3        1      3                              Heikkinen, Miss. Laina female  26     0
4           4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
5           5        0      3                            Allen, Mr. William Henry   male  35     0
6           6        0      3                                    Moran, Mr. James   male  NA     0
  Parch           Ticket    Fare Cabin Embarked
1     0        A/5 21171  7.2500              S
2     0         PC 17599 71.2833   C85        C
3     0 STON/O2. 3101282  7.9250              S
4     0           113803 53.1000  C123        S
5     0           373450  8.0500              S
6     0           330877  8.4583              Q
> #HANDLING MISSING VALUES
> colSums(is.na(df))
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp       Parch
          0           0           0           0           0         177           0           0
     Ticket        Fare       Cabin    Embarked
          0           0           0           0
> df$Age[is.na(df$Age)] <- median(df$Age, na.rm = TRUE)
> df <- df[, !names(df) %in% c("Cabin")]
```

DATA VISUALIZATION

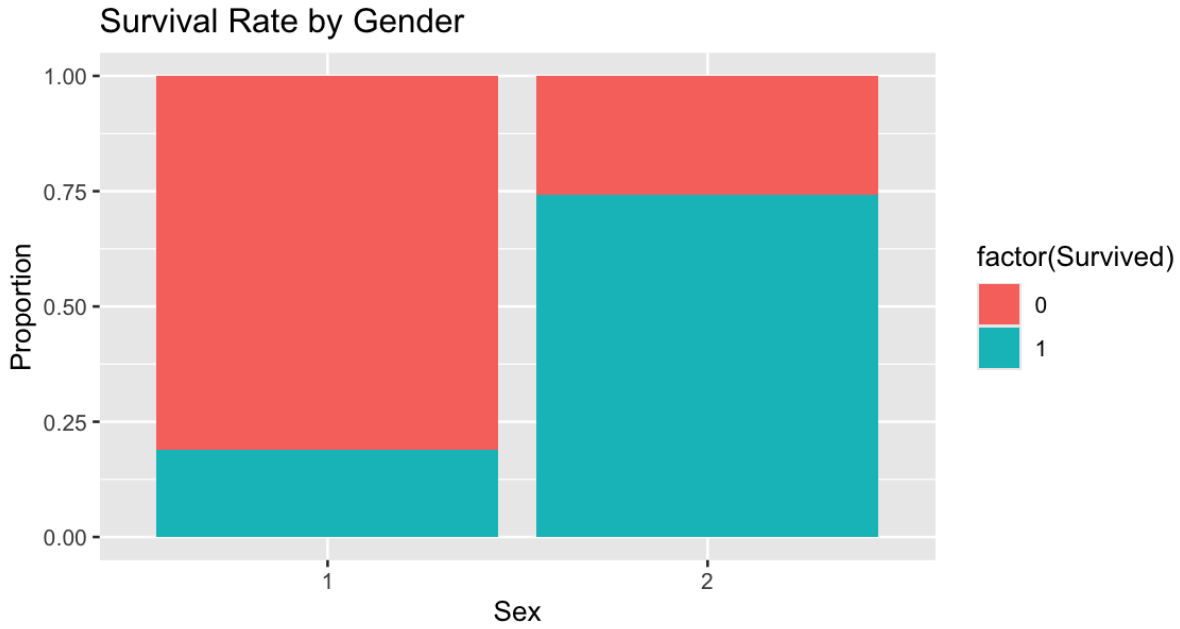
```
> library(ggplot2)
> ggplot(df, aes(x = factor(Survived))) +
+   geom_bar() +
+   labs(title = "Distribution of Survived", x = "Survived", y = "Count")
```



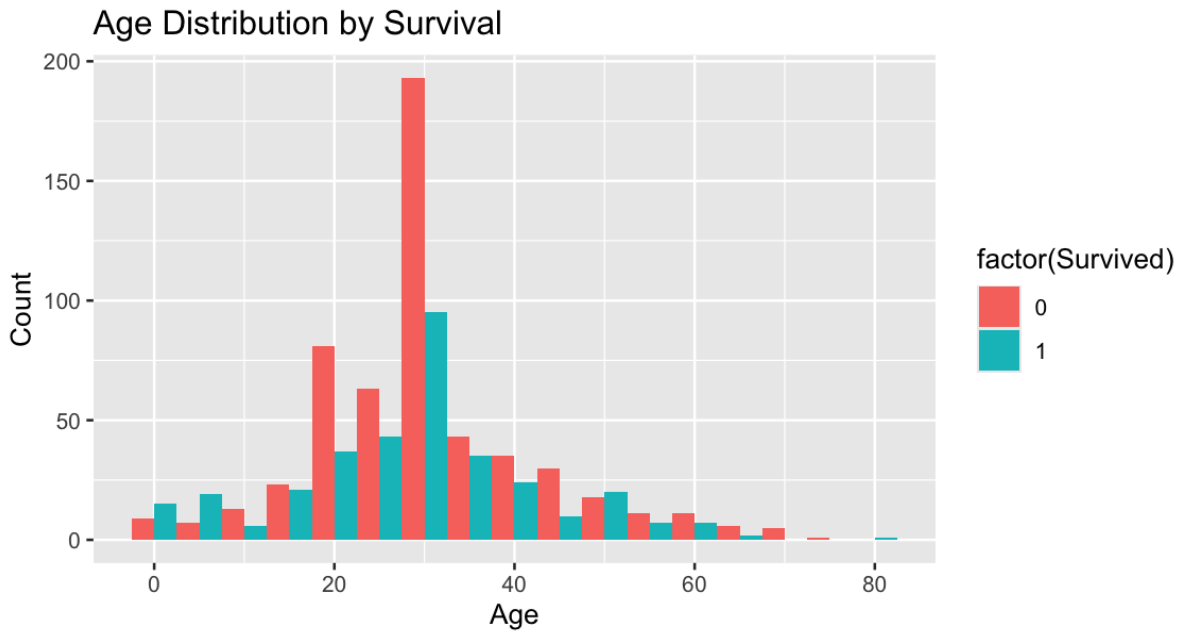
```
> ggplot(df, aes(x = factor(Pclass), fill = factor(Survived))) +
+   geom_bar(position = "fill") +
+   labs(title = "Survival Rate by Passenger Class", x = "Passenger Class", y = "Proportion")
```




```
> ggplot(df, aes(x = factor(Sex), fill = factor(Survived))) +
+   geom_bar(position = "fill") +
+   labs(title = "Survival Rate by Gender", x = "Sex", y = "Proportion")
```



```
> ggplot(df, aes(x = Age, fill = factor(Survived))) +
+   geom_histogram(binwidth = 5, position = "dodge") +
+   labs(title = "Age Distribution by Survival", x = "Age", y = "Count")
```



CORRELATION ANALYSIS

```
> correlation_matrix <- cor(df[, sapply(df, is.numeric)], use = "complete.obs")
> library(corrplot)
corrplot 0.92 loaded
> corrplot(correlation_matrix, method = "color", tl.cex = 0.8)
> ggplot(df, aes(x = Age, fill = factor(Survived))) +
```