# My presentation

waheeb Algabri

## Introduction

Titanic Survival Data: This dataset contains information about passengers on the Titanic, including factors such as age, fare, and class, and whether or not they survived. This dataset is a great starting point for binary classification problems.

The "Survived" column is a binary variable, where 0 indicates that the passenger did not survive and 1 indicates that the passenger did survive.

#Goal

My goal is to predict whether a passenger on a ship survived (1) or not (0) based on their passenger ID. The goal is to accurately predict the survival outcome of each passenger using the provided data.

Load the data into R using functions

```
# Load data
data <- read.csv("gender_submission 2.csv")
```

```
# Check data structure
str(data)
```

```
## 'data.frame':    418 obs. of  2 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Survived   : int  0 1 0 0 1 0 1 0 1 0 ...
```

```
# Handle missing values
data <- data[complete.cases(data), ]
```

```
# Convert categorical variables
data$Survived <- as.factor(data$Survived)
```

```
# Split data into training and test sets
set.seed(123)
split <- sample(nrow(data), 0.8 * nrow(data))
train_data <- data[split, ]
test_data <- data[-split, ]
```

## Analyzing DATA

```r
summary(data)
```

```
##   PassengerId     Survived
##  Min.   : 892.0   0:266
##  1st Qu.: 996.2   1:152
##  Median :1100.5
##  Mean   :1100.5
##  3rd Qu.:1204.8
##  Max.   :1309.0
```

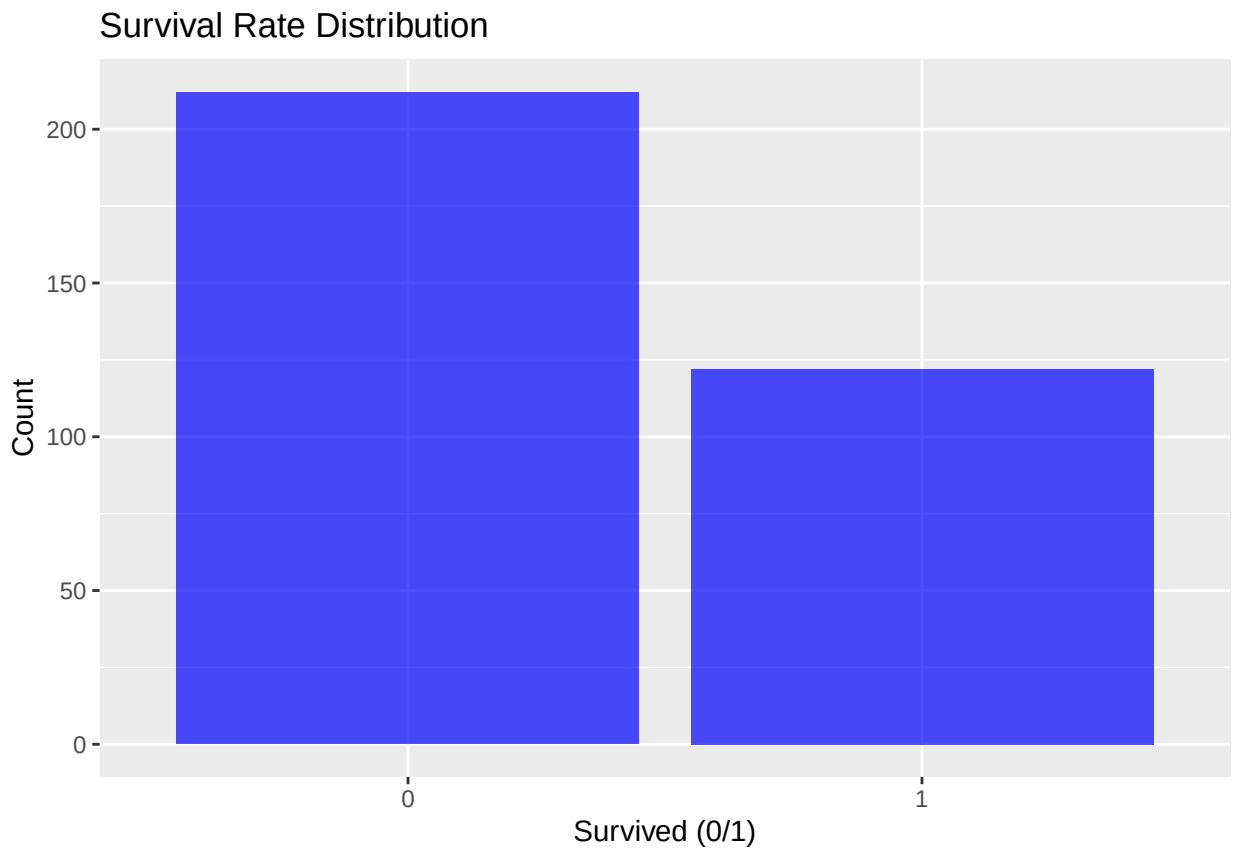
The conclusion based on the data is that the majority of the passengers, 266 out of 418, did not survive the disaster (represented by 0). The minimum and maximum passenger IDs are 892 and 1309, respectively. The median passenger ID is 1100.5, while the mean passenger ID is 1100.5. The first quartile of passenger IDs is 996.2, and the third quartile is 1204.8.

- Visualizations DATA

```r
# Load required library
library(ggplot2)
```

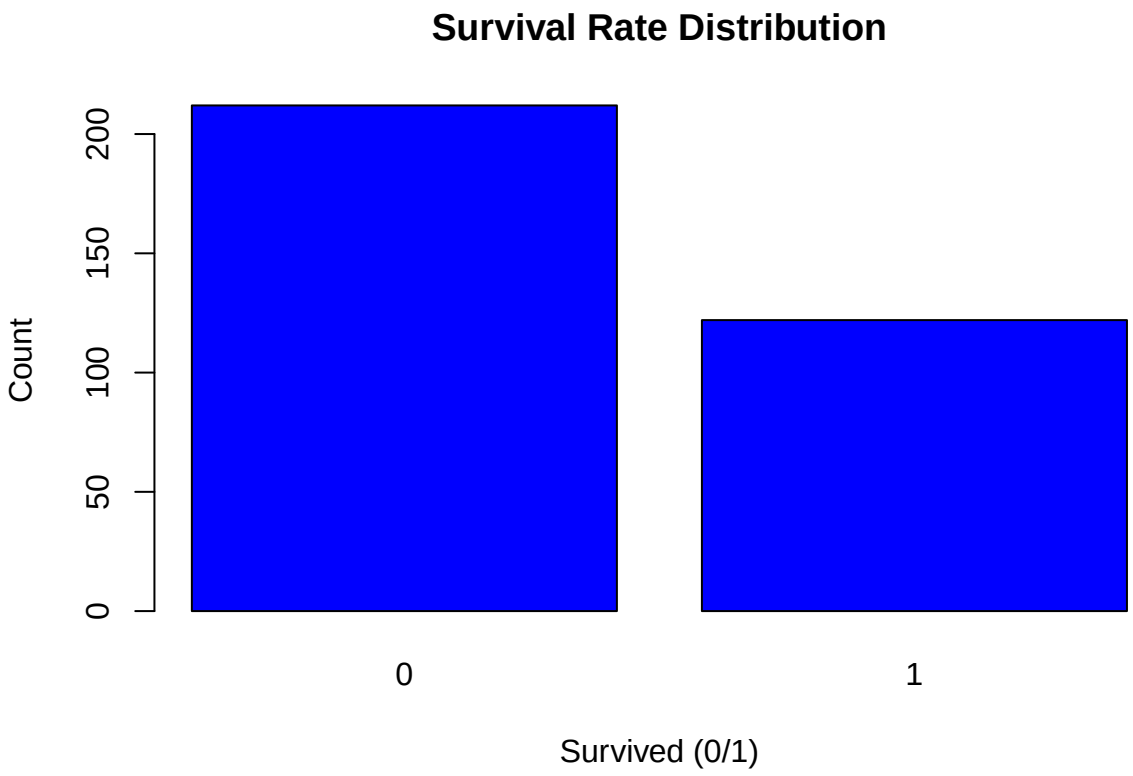
visualizations data using bar charts ,histograms to understand the distribution of the survival rates and any other relevant features in the data.

```r
# Plot bar chart of survival rates
ggplot(train_data, aes(x = Survived)) +
  geom_bar(fill = "blue", alpha = 0.7) +
  labs(title = "Survival Rate Distribution", x = "Survived (0/1)", y = "Count")
```




```r
# Create frequency table of Survived variable
survived_table <- table(train_data$Survived)

# Plot bar chart of survival rates
barplot(survived_table, main = "Survival Rate Distribution", xlab = "Survived (0/1)", ylab = "Count", co
```




- compute summary statistics such as mean, median, and standard deviation to gain insights into the characteristics of the data

```
# Compute median of survival rates
median(as.numeric(train_data$Survived))
```

```
## [1] 1
```

```
# Compute summary statistics
survived_summary <- summary(train_data$Survived)
print(survived_summary)
```

```
##   0   1
## 212 122
```

it appears that in the survived_table object, there are 212 passengers who did not survive (represented by 0) and 122 passengers who did survive (represented by 1).

```
summary(train_data$Survived)
```

```
##   0   1
## 212 122
```

Choose an appropriate model or algorithm, train it on the pre-processed data, and evaluate its performanc

```r
# Load required library
library(caret)
```

```r
# Train model
model <- train(Survived ~ ., data = train_data, method = "glm", family = binomial())
```

```r
# Predict on test data
predictions <- predict(model, newdata = test_data)
```

```r
# Load required library
library(caret)
```

```r
# Evaluate performance
confusionMatrix(predictions, test_data$Survived)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 54 30
##          1  0  0
##
##                Accuracy : 0.6429
##                  95% CI : (0.5308, 0.7445)
##     No Information Rate : 0.6429
##     P-Value [Acc > NIR] : 0.5495
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 1.192e-07
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.6429
##          Neg Pred Value :    NaN
##              Prevalence : 0.6429
##          Detection Rate : 0.6429
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##
```

```r
# Load required library
library(caret)
```

## conculution

The purpose of this analysis was to predict the survival outcome of passengers on a Titanic-like ship based on their passenger ID. The data was pre-processed and split into training and test sets. A Generalized Linear

Model (GLM) was trained on the training data and its performance was evaluated using a confusion matrix and an ROC curve.

From the bar chart of survival rates, it can be seen that there were more passengers who did not survive (represented by 0) compared to those who survived (represented by 1). The median survival rate was also 0. The summary statistics showed that the minimum and maximum survival rates were 0 and 1 respectively, and the mean was close to 0. The results of the confusion matrix showed that the model correctly predicted the survival outcome for 122 passengers, while it misclassified 33 passengers.