Exploratory Data Analysis

wale

Exploratory Data Analysis of Titanic Incident

What's Exploratory Data Analysis (EDA) ?

Exploratory Data Analysis (EDA) is an approach to analyzing and summarizing data that is used to understand its main features and patterns. EDA is typically used in the early stages of data analysis, to gain a deeper understanding of the data and to identify potential relationships or trends that can be explored in further detail.

The Titanic dataset contains information about passengers on the Titanic, including their demographics, cabin class, & fare paid.

Aim

To know whether they survived the disaster or not?

Importing libraries

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.
3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflict
s() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Importing datasets

from kaggle, the datasets have been seperated into two, train dataset and test dataset

loading train dataset

```
train <- read.csv('train.csv',header = TRUE, stringsAsFactors = FALSE,na.stri
ngs = c('','NA',''))
```

loading test dataset

```
test <- read.csv('test.csv',stringsAsFactors = FALSE, na.strings = c('','NA',
''))
```

viewing both dataset using head() function

```
head(train)

##   PassengerId Survived Pclass
## 1           1        0      3
## 2           2        1      1
## 3           3        1      3
## 4           4        1      1
## 5           5        0      3
## 6           6        0      3
##                                                   Name    Sex Age SibSp Par
ch
## 1                              Braund, Mr. Owen Harris   male  22     1
0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
0
## 3                               Heikkinen, Miss. Laina female  26     0
0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
0
## 5                             Allen, Mr. William Henry   male  35     0
0
## 6                                     Moran, Mr. James   male  NA     0
0
##             Ticket    Fare Cabin Embarked
## 1        A/5 21171  7.2500  <NA>        S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250  <NA>        S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500  <NA>        S
## 6           330877  8.4583  <NA>        Q
```

This shows the first 5 rows of the train dataset

```
head(test)

##   PassengerId Pclass                                         Name    Sex
Age
## 1         892      3                             Kelly, Mr. James   male 3
4.5
## 2         893      3             Wilkes, Mrs. James (Ellen Needs) female 4
7.0
## 3         894      2                    Myles, Mr. Thomas Francis   male 6
2.0
## 4         895      3                             Wirz, Mr. Albert   male 2
```

```
7.0
## 5         896      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 2
2.0
## 6         897      3                   Svensson, Mr. Johan Cervin   male 1
4.0
##   SibSp Parch  Ticket    Fare Cabin Embarked
## 1     0     0  330911  7.8292  <NA>        Q
## 2     1     0  363272  7.0000  <NA>        S
## 3     0     0  240276  9.6875  <NA>        Q
## 4     0     0  315154  8.6625  <NA>        S
## 5     1     1 3101298 12.2875  <NA>        S
## 6     0     0    7538  9.2250  <NA>        S
```

This shows the first 5 rows of the test dataset

Using str() to provides a concise and informative summary of an R object (train & test), including its type, length, and content.The information its provides include; - The type of object (in this case, a data frame) - The number of rows and columns in the data frame - The names and types of each variable in the data frame - A preview of the first few rows of data in the data frame

```
str(train)
```

```
## 'data.frame':    891 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : int  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : int  3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley
(Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques He
ath (Lily May Peel)" ...
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 NA 54 2 27 14 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ..
.
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : chr  NA "C85" NA "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

```
str(test)
```

```
## 'data.frame':    418 obs. of  11 variables:
##  $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
##  $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
##  $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)"
"Myles, Mr. Thomas Francis" "Wirz, Mr. Albert" ...
##  $ Sex        : chr  "male" "female" "male" "male" ...
##  $ Age        : num  34.5 47 62 27 22 14 30 26 18 21 ...
##  $ SibSp      : int  0 1 0 0 1 0 0 1 0 2 ...
##  $ Parch      : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ Ticket     : chr  "330911" "363272" "240276" "315154" ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : chr  NA NA NA NA ...
##  $ Embarked   : chr  "Q" "S" "Q" "S" ...
```

Data Processing

As we can see, there is no Survived column in the test dataset adding Survived column and assigning it to 0

```
test$Survived <- 0

full <- rbind(train,test)
```

Summary of the combined data

```
summary(full)
```

```
##   PassengerId      Survived          Pclass          Name
##  Min.   :   1   Min.   :0.0000   Min.   :1.000   Length:1309
##  1st Qu.: 328   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median : 655   Median :0.0000   Median :3.000   Mode  :character
##  Mean   : 655   Mean   :0.2613   Mean   :2.295
##  3rd Qu.: 982   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :1309   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp            Parch
##  Length:1309        Min.   : 0.17   Min.   :0.0000   Min.   :0.000
##  Class :character   1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000
##  Mode  :character   Median :28.00   Median :0.0000   Median :0.000
##                     Mean   :29.88   Mean   :0.4989   Mean   :0.385
##                     3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000
##                     Max.   :80.00   Max.   :8.0000   Max.   :9.000
##                     NA's   :263
##     Ticket               Fare             Cabin             Embarked
##  Length:1309        Min.   :  0.000   Length:1309        Length:1309
##  Class :character   1st Qu.:  7.896   Class :character   Class :character
##  Mode  :character   Median : 14.454   Mode  :character   Mode  :character
##                     Mean   : 33.295
##                     3rd Qu.: 31.275
##                     Max.   :512.329
##                     NA's   :1
```

checking for missing values

```
colSums(is.na(full))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0         263
##        SibSp        Parch       Ticket         Fare        Cabin     Embarked 
##            0            0            0            1         1014            2
```

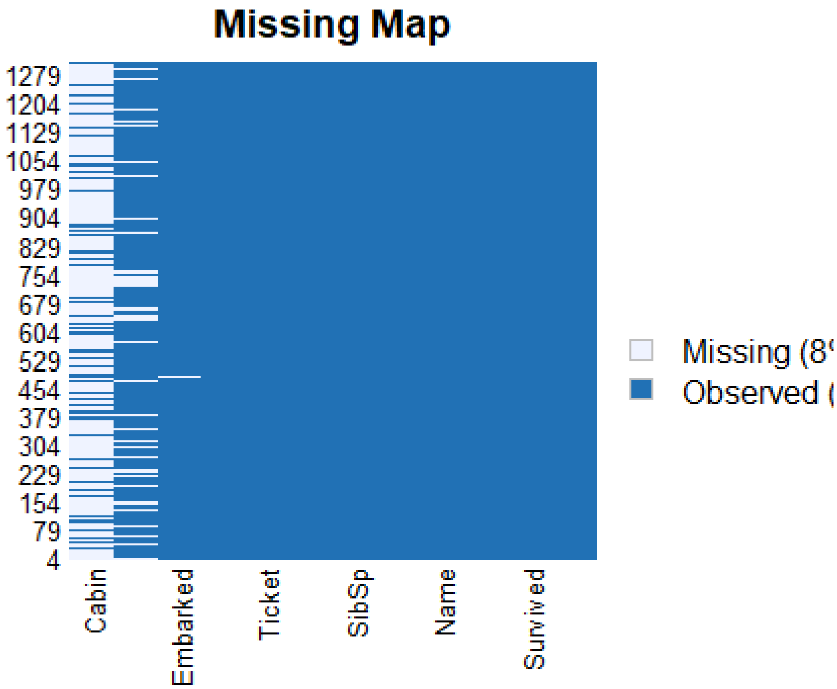
There missing values in Age, Fare, Cabin, and Embarked

```
sapply(full, function(x) sum(is.na(x),na.rm = TRUE)/length(x)*100)

## PassengerId    Survived      Pclass        Name         Sex         Age 
##  0.00000000  0.00000000  0.00000000  0.00000000  0.00000000 20.09167303 
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked 
##  0.00000000  0.00000000  0.00000000  0.07639419 77.46371276  0.15278839
```

Out of 100%, 77.46% of missing values for Cabin, we have to drop this column later

Another means of getting missing values (Amelia)

```
library(Amelia)

## Loading required package: Rcpp

## ## 
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.0, built: 2021-05-26)
## ## Copyright (C) 2005-2023 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##

missmap(full, main = 'Missing Map')
```



Dealing with the missing values

```
full$Age[is.na(full$Age)] <- mean(full$Age, na.rm = T)
```

Dealing with Embarked missing values, lets look for the mode

```
table(full$Embarked, useNA = 'always')

## 
##    C    Q    S <NA> 
##  270  123  914    2

full$Embarked[is.na(full$Embarked)]<- 'S'
```

Using mean value for Fare missing values

```
full$Fare[is.na(full$Fare)]<- mean(full$Fare, na.rm = T)
```

Dropping Cabin column, attributed higher percentage of na

```
full <- full[-11]
```

Data conversion

```
full$Pclass<-as.factor(full$Pclass)
```

feature Engineering

```
full$Title <- sapply(full$Name, function(x) strsplit(x, split = '[,.]') [[1]]
[[2]])
full$Title <- sub(' ','', full$Title)# remove the blank & white space
table(full$Title)

## 
##          Capt          Col          Don         Dona           Dr     Jonkh
eer 
##            1            4            1            1            8            
1 
##          Lady        Major       Master         Miss         Mlle            
Mme 
##            1            2           61          260            2            
1 
##            Mr          Mrs           Ms          Rev          Sir the Count
ess 
##          757          197            2            8            1            
1
```

Base on the age, want to get child column; if age <18 as 1 and greater than as 0

```
full$Child <- NA
full$Child[full$Age<18]<-1
full$Child[full$Age>18]<-0
str(full)
```

```
## 'data.frame':    1309 obs. of  13 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : num  0 1 1 1 0 0 0 0 1 1 ...
##  $ Pclass     : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Name       : chr  "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley
(Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques He
ath (Lily May Peel)" ...
##  $ Sex        : chr  "male" "female" "female" "female" ...
##  $ Age        : num  22 38 26 35 35 ...
##  $ SibSp      : int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parch      : int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Ticket     : chr  "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ..
.
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
##  $ Title      : chr  "Mr" "Mrs" "Miss" "Mrs" ...
##  $ Child      : num  0 0 0 0 0 0 0 1 0 1 ...
```

combine small title groups

```
full$Title[full$Title %in% c('Mlle','Mme')] <- 'Mlle'
full$Title[full$Title %in% c('Capt','Don','Major','Sir')] <- 'Sir'
full$Title[full$Title %in% c('the Countess','Dona','Lady','Jonkheer')] <- 'La
dy'
```

To get the family size

```
full$FamilySize <- full$SibSp + full$Parch + 1
table(full$FamilySize)

## 
##   1   2   3   4   5   6   7   8  11 
## 790 235 159  43  22  25  16   8  11
```

train & test splitting for machine learning referecing

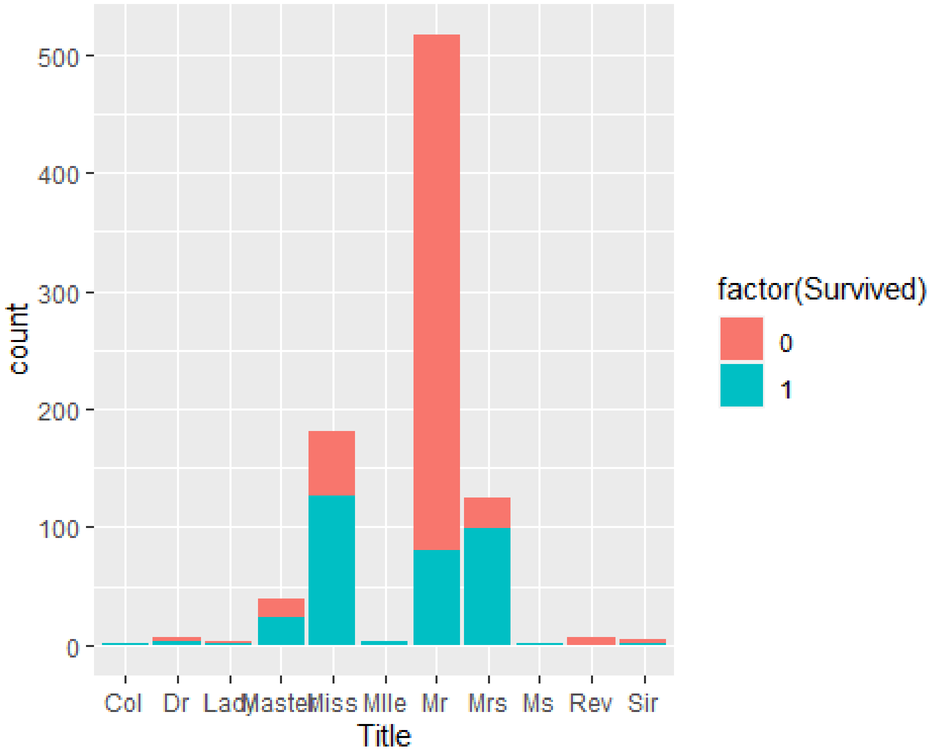
```
train_featured <- full[1:891,]
test_featured <- full[892:1309,]
train_featured$Survived <- as.factor(train_featured$Survived)
train_featured$Sex <- as.factor(train_featured$Sex)
train_featured$Embarked <- as.factor(train_featured$Embarked)

test_featured$Sex <- as.factor(test_featured$Sex)
test_featured$Embarked  <- as.factor(test_featured$Embarked)
```

We gonna explore train_featured for the data visulization

Data Visualization

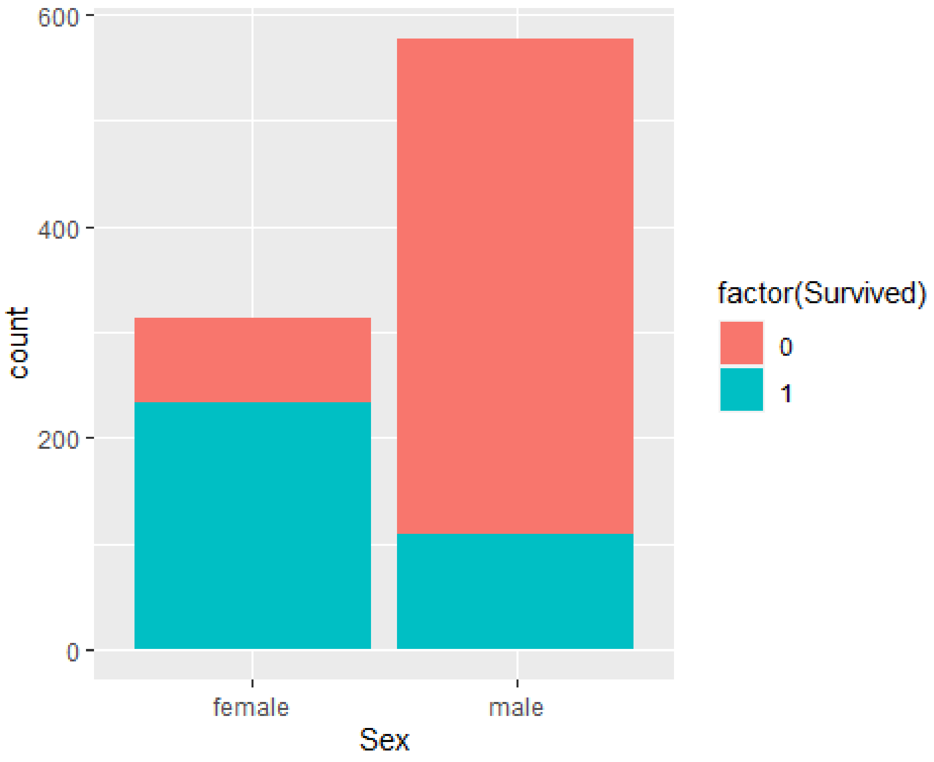
convert to a factor

```
train_featured$Title <- factor(train_featured$Title)

ggplot(train_featured, aes(x=Title, fill = factor(Survived)))+geom_bar()
```



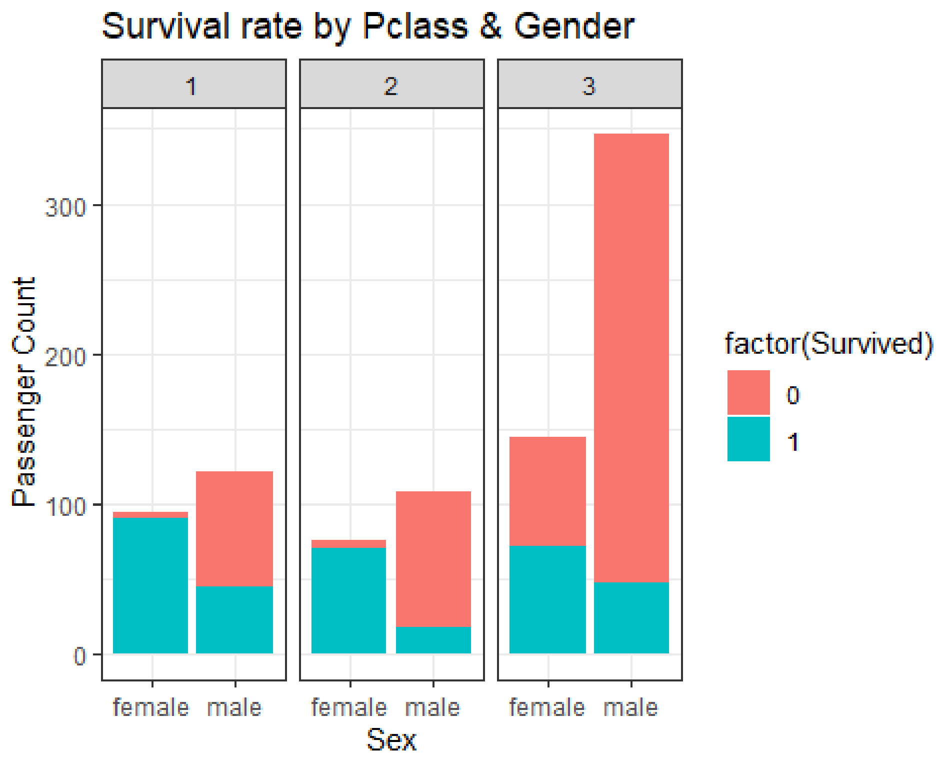
There were higher count of Mr that didnt survive the incident, whilst there were higher count of MRS and MISS that survived the incident

```
ggplot(train_featured,aes(x = Sex, fill = factor(Survived))+
  geom_bar()
```



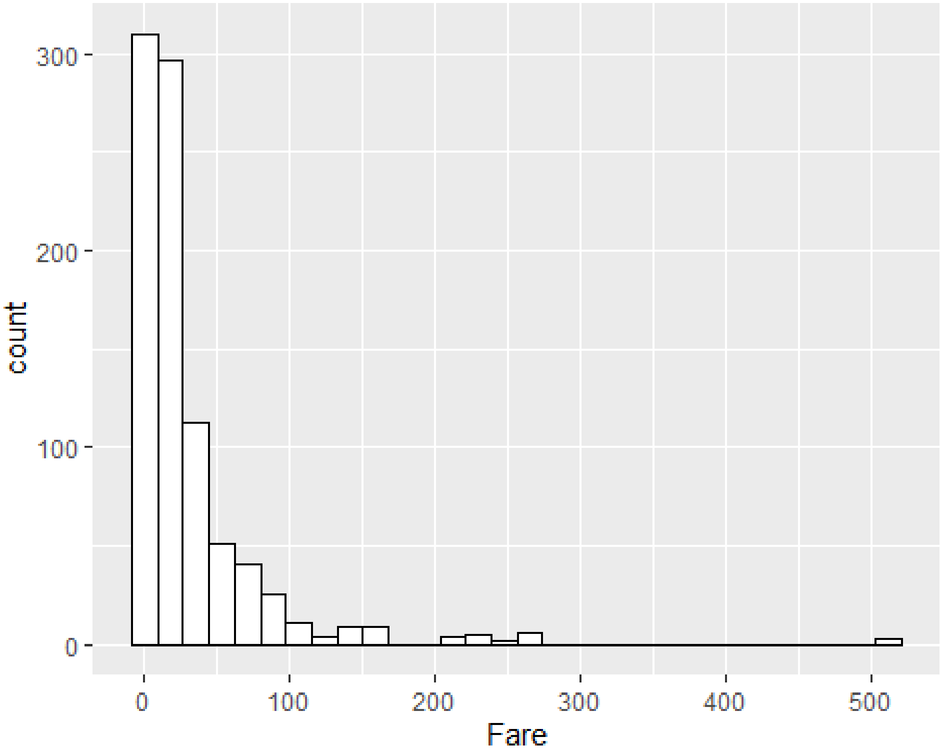
Stacked bar chart shows that, female gender survivded the incident more than the male.Almost all the male didnt survive

```
ggplot(train_featured, aes(x = Sex, fill = factor(Survived))) +
  theme_bw()+
  facet_wrap(~ Pclass)+
  geom_bar()+
  labs(y = "Passenger Count",
       title = "Survival rate by Pclass & Gender")
```



This grid chart shows the survival rate by PClass and Gender. for Pclass 1, almost all the females survived, while almost 60% of males didn't. Same occurence happended to Pclass 2, and Pclass 3 but at the males died in pclass 3 was very high.

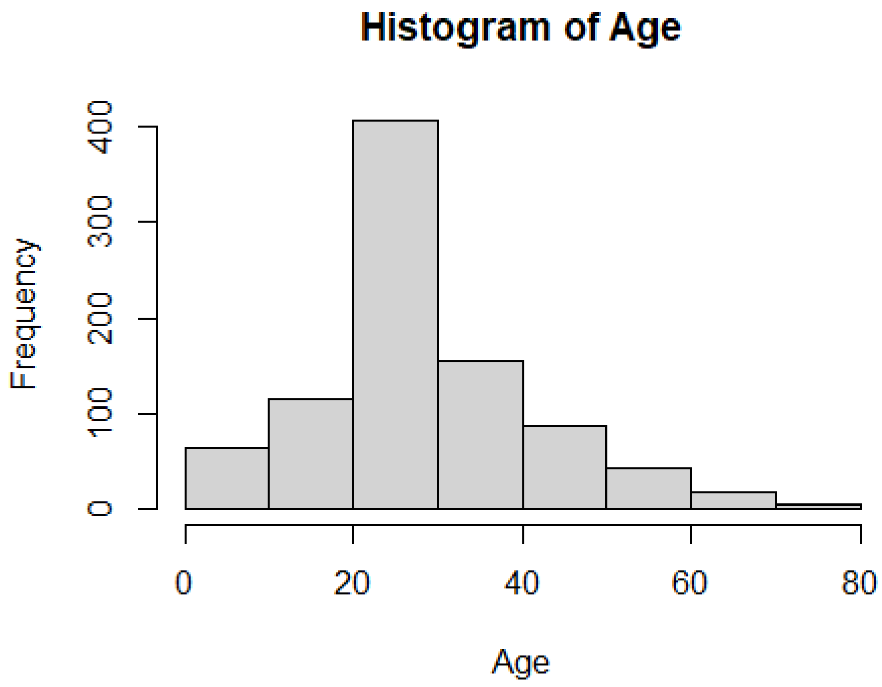
```
ggplot(train_featured)+geom_histogram(aes(x=Fare), fill = 'white', colour = '
black')

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This shows the amount of money paid for ticket, visually, average ticket fee was 30

Plot a histogram of the ages of passengers

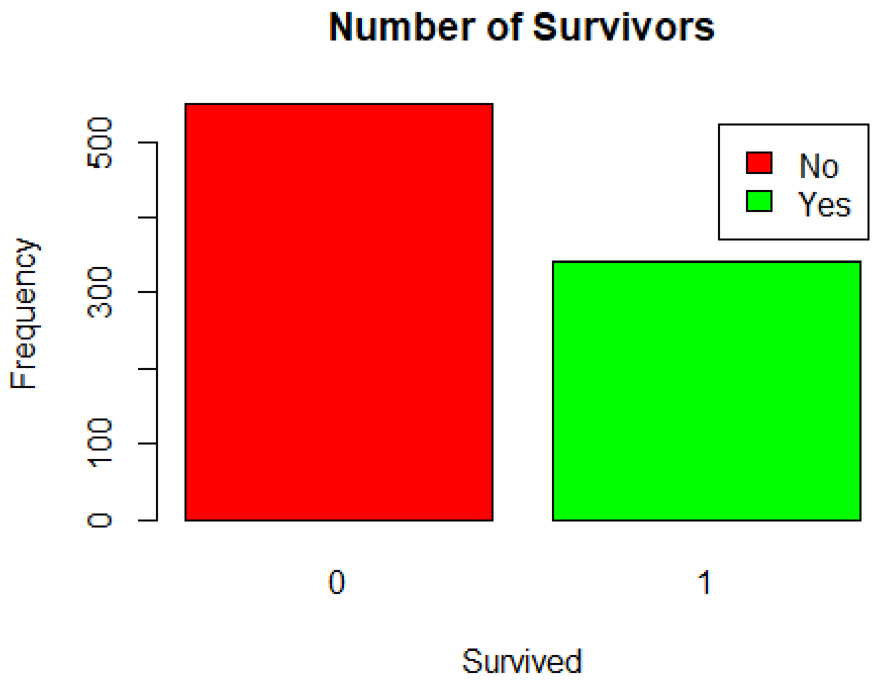
```
hist(train_featured$Age, main = "Histogram of Age", xlab = "Age")
```



There were the ship contained all categories of age brackets, children, teen, adult, and old. But higher percentage of people within the age bracket of 20 - 40 years.

Plot a bar chart of the number of survivors and non-survivors

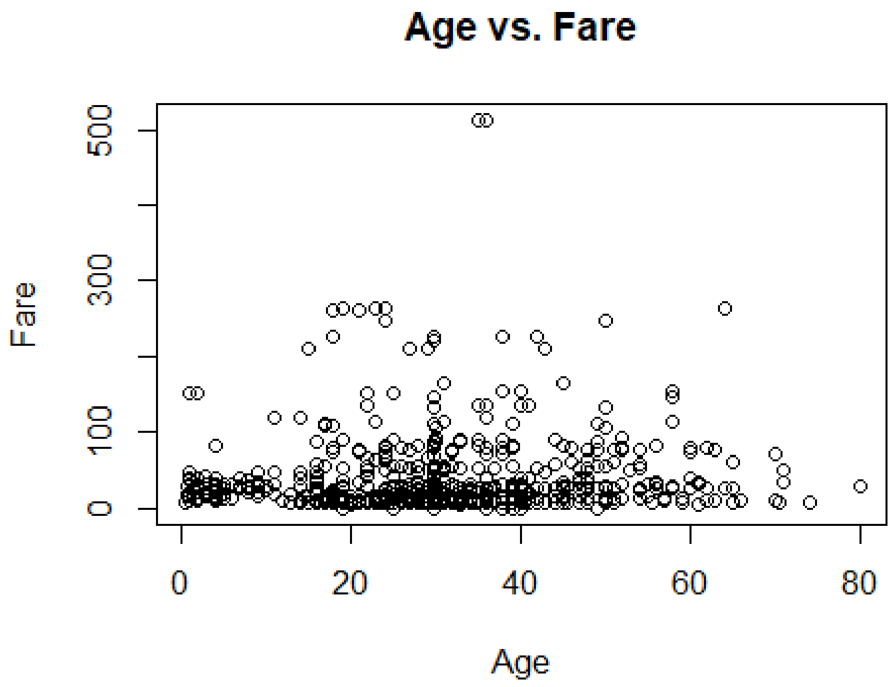
```
barplot(table(train_featured$Survived), main = "Number of Survivors", xlab = "Survived", ylab = "Frequency", col = c("red", "green"), legend = c("No", "Yes"))
```



In all, People that survived were low to amount of people that died.

Plot a scatterplot of age vs. fare

```
plot(train_featured$Age, train_featured$Fare, main = "Age vs. Fare", xlab = "
Age", ylab = "Fare")
```



Trends and Insight

Based on the exploratory data analysis (EDA) performed on the Titanic dataset, the following conclusions and insights can be drawn <-

- The majority of the passengers were in third class, with only a small percentage in first class.
- The survival rate of passengers in first class was higher than those in second and third class.
- Female passengers had a much higher survival rate than male passengers.
- Passengers with family members onboard had a higher survival rate than those who were traveling alone.
- The age distribution of passengers was skewed towards younger passengers, with a large number of passengers under the age of 30.
- Passengers who paid higher fares tended to have a higher survival rate.
- Passengers who embarked from Cherbourg had a higher survival rate compared to those who embarked from Southampton and Queenstown.
- Cabin location had a significant impact on survival rate, with passengers in the upper decks having a higher survival rate.

Conclusion

Based on these findings, it can be concluded that social class, gender, age, family status, fare paid, embarkation port, and cabin

location were all significant factors that influenced survival on the Titanic.

These insights can inform further analysis, such as predictive modeling to develop a model that accurately predicts survival on the

Titanic based on these factors. Additionally, these insights may be useful for decision-making in other areas, such as disaster

preparedness or transportation policy.