Final Project - Titanic dataset

Group 2: Christine Otruba & Shweta Kalbhor

04/27/2023 @ 4:30pm - 5:00 pm

Instructions for Final Project

The final project is to investigate a dataset using all the tools we learn in class. Your project should include the following three parts.

1. Data Preparation: Show the information of the dataset. E.g.# of observations, # of attributes, data types, missing values, etc.
2. Data Exploration(EDA): Data Visualization, at least one histogram, one boxplot, and one overlay histogram with your conclusion
3. Data Analysis Hypothesis Testing: Construct a hypothesis testing with null and alternative hypotheses. Use the appropriate test to get the conclusion. Build a linear regression model with subset selection. Please indicate all significant attributes, assess your model, and predict in a test dataset.

Read and View the `titanic_original` dataset which was obtained from https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html (https://urldefense.com/v3/__https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/problem12.html__;!!B24N9PvjPQId!fnwPWmS_yqt-7Y_YM4b1w9nn4pSBkC9qcYx_c3shHuDOziqCPhB81nCUOlsxtap2n8h1IyN7WJXETjIEkTVutcKMbvtIG2Eanw$)

```
# import packages
library(MASS)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   1.0.1
## ✓ tibble  3.1.8      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.5.0
## ✓ readr   2.1.3      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ✕ dplyr::select() masks MASS::select()
```

```
library(rcompanion)
library(MLmetrics)
```

```
##
## Attaching package: 'MLmetrics'
##
## The following object is masked from 'package:base':
##
##     Recall
```

```
#Load data
titanic_original <- read.csv("~/Documents/IIT/Spring2023/ITMD514/FinalProject/titanic.csv", header = TRUE, sep=",")
str(titanic_original)
```

```
## 'data.frame':    887 obs. of  8 variables:
##  $ 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  "Mr. Owen Harris Braund" "Mrs. John Bradley (Florence Briggs Thayer) Cumings" "Miss.
Laina Heikkinen" "Mrs. Jacques Heath (Lily May Peel) Futrelle" ...
##  $ Sex                    : chr  "male" "female" "female" "female" ...
##  $ Age                    : num  22 38 26 35 35 27 54 2 27 14 ...
##  $ Siblings.Spouses.Aboard: int  1 1 0 1 0 0 0 3 0 1 ...
##  $ Parents.Children.Aboard: int  0 0 0 0 0 0 0 1 2 0 ...
##  $ Fare                   : num  7.25 71.28 7.92 53.1 8.05 ...
```

```
head(titanic_original)
```

```
##   Survived Pclass                                            Name    Sex Age
## 1        0      3                         Mr. Owen Harris Braund   male  22
## 2        1      1 Mrs. John Bradley (Florence Briggs Thayer) Cumings female  38
## 3        1      3                          Miss. Laina Heikkinen female  26
## 4        1      1        Mrs. Jacques Heath (Lily May Peel) Futrelle female  35
## 5        0      3                        Mr. William Henry Allen   male  35
## 6        0      3                                Mr. James Moran   male  27
##   Siblings.Spouses.Aboard Parents.Children.Aboard    Fare
## 1                       1                       0  7.2500
## 2                       1                       0 71.2833
## 3                       0                       0  7.9250
## 4                       1                       0 53.1000
## 5                       0                       0  8.0500
## 6                       0                       0  8.4583
```

Background Information:

The RMS Titanic, a luxury steamship, sank in the early hours of April 15, 1912, off the coast of Newfoundland in the North Atlantic after sideswiping an iceberg during its maiden voyage. Of the 2,240 passengers and crew on board, more than 1,500 lost their lives in the disaster. The dataset selected is comprised of fictional names and related data but follows realistic values of the passengers on the Titanic. This presentation will attempt to find correlation regarding the survival rate and price paid for fare/Class (1st, 2nd, 3rd), age, and/or sex.

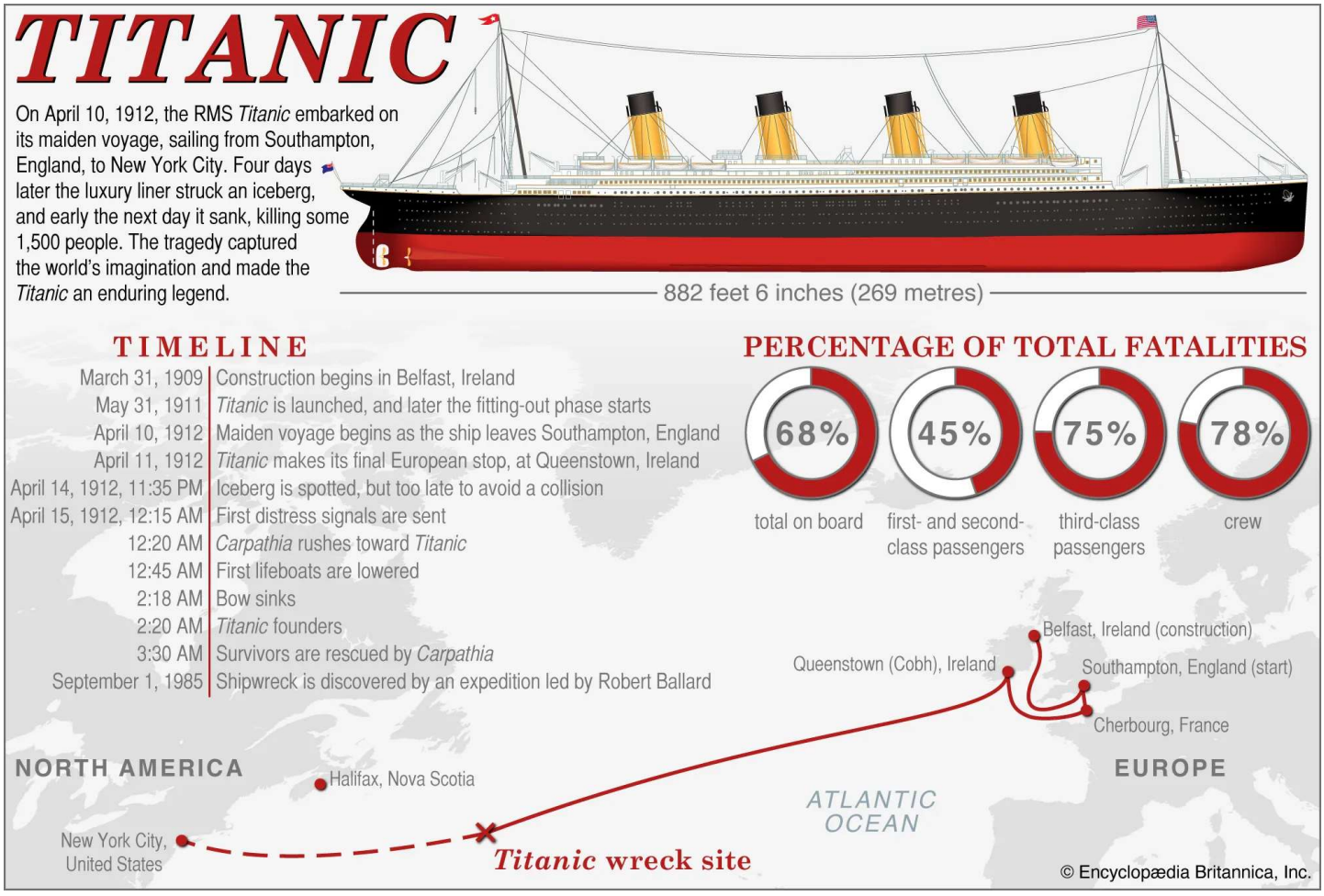


Image source: https://www.britannica.com/topic/Titanic (https://urldefense.com/v3/__https://www.britannica.com/topic/Titanic__;!!B24N9PvjPQId!fnwPWmS_yqt-7Y_YM4b1w9nn4pSBkC9qcYx_c3shHuDOziqCPhB81nCUOlsxtap2n8h1IyN7WJXETjlEkTVutcKMbvuh8uzvSQ$)

1) Data Preparation: Show the information of the dataset. E.g.# of observations, # of attributes, data types, missing values, etc.

In order to work with the data better, columns are converted into more appropriate data types. For example, the `titanic_original$Age` is converted from data type "number" to "integer". Additionally, `titanic_original$Sex` (character data type) is used to create a new column `titanic_original$Sex_M_F` which is a "Factor" data type, in order to identify male vs female data. The `titanic_original` dataset is then further cleaned up by removing columns that are not being used in this final project: `$Name`, `$Siblings.Spouses.Aboard`, `$Parents.Children.Aboard`. Also, the numerical values (Age and Fare) were placed into groups in order to better formulate conclusions.

```
titanic_original$Age<-as.integer(titanic_original$Age)
titanic_original$Sex_M_F <- with(titanic_original, ifelse(substr(titanic_original$Sex,1,1) =="f", "F", "M"))
titanic_original[,c(1:2,9)]<-lapply(titanic_original[,c(1:2,9)], factor)
titanic<-titanic_original[,c(1:2,5,8:9)]
titanic_dim<-dim(titanic)
summ_age<-summary(titanic$Age)
summ_fare<-summary(titanic$Fare)

titanic$AgeGroup <- cut(titanic$Age,
                        breaks = c(-Inf
                                   ,5 ,10 ,15,20,25,30,35,40,45,50,55,60 ,65,70,75,80,85
                                   , Inf),

                        labels = c("0-4 years"
                                   ,"5-9 years","10-14 years","15-19 years","20-24 years"
                                   ,"25-29 years","30-34 years","35-39 years","40-44 years"
                                   ,"45-49 years","50-54 years","55-59 years","60-64 years"
                                   ,"65-69 years","70-74 years","75-79 years","80-84 years"
                                   ,"85+ years"),
                        right = FALSE)

titanic$FarePrice <- cut(titanic$Fare,
                        breaks = c(-Inf
                                   ,10 ,20,30,40,50,60,70,80,90,100,200,300
                                   , Inf),

                        labels = c("0-9 £","10-19 £","20-29 £","30-39 £","40-49 £","50-59 £","60-69 £","70-79 £","80-89
£","90-99 £","100-199 £","200-299 £","300+ £"),
                        right = FALSE)

str(titanic)
```

```
## 'data.frame':    887 obs. of  7 variables:
##  $ Survived : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ Pclass   : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...
##  $ Age      : int  22 38 26 35 35 27 54 2 27 14 ...
##  $ Fare     : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Sex_M_F  : Factor w/ 2 levels "F","M": 2 1 1 1 2 2 2 1 1 ...
##  $ AgeGroup : Factor w/ 18 levels "0-4 years","5-9 years",..: 5 8 6 8 8 6 11 1 6 3 ...
##  $ FarePrice: Factor w/ 13 levels "0-9 £","10-19 £",..: 1 8 1 6 1 1 6 3 2 4 ...
```

```
summary(titanic)
```

```
##  Survived Pclass       Age             Fare         Sex_M_F        AgeGroup
##  0:545    1:216   Min.   : 0.00   Min.   :  0.000   F:314   20-24 years:158
##  1:342    2:184   1st Qu.:20.00   1st Qu.:  7.925   M:573   25-29 years:135
##           3:487   Median :28.00   Median : 14.454           30-34 years:112
##                   Mean   :29.46   Mean   : 32.305           15-19 years:110
##                   3rd Qu.:38.00   3rd Qu.: 31.137           35-39 years: 87
##                   Max.   :80.00   Max.   :512.329           40-44 years: 62
##                                                             (Other)    :223
##      FarePrice
##  0-9 £    :333
##  10-19 £  :178
##  20-29 £  :136
##  30-39 £  : 64
##  50-59 £  : 39
##  100-199 £: 33
##  (Other)  :104
```

```
glimpse(titanic)
```

```
## Rows: 887
## Columns: 7
## $ Survived  <fct> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, …
## $ Pclass    <fct> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 3, 3, …
## $ Age       <int> 22, 38, 26, 35, 35, 27, 54, 2, 27, 14, 4, 58, 20, 39, 14, 55…
## $ Fare      <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51.8625, 2…
## $ Sex_M_F   <fct> M, F, F, F, M, M, M, M, F, F, F, F, M, M, F, F, M, M, F, F, …
## $ AgeGroup  <fct> 20-24 years, 35-39 years, 25-29 years, 35-39 years, 35-39 ye…
## $ FarePrice <fct> 0-9 £, 70-79 £, 0-9 £, 50-59 £, 0-9 £, 0-9 £, 50-59 £, 20-29…
```

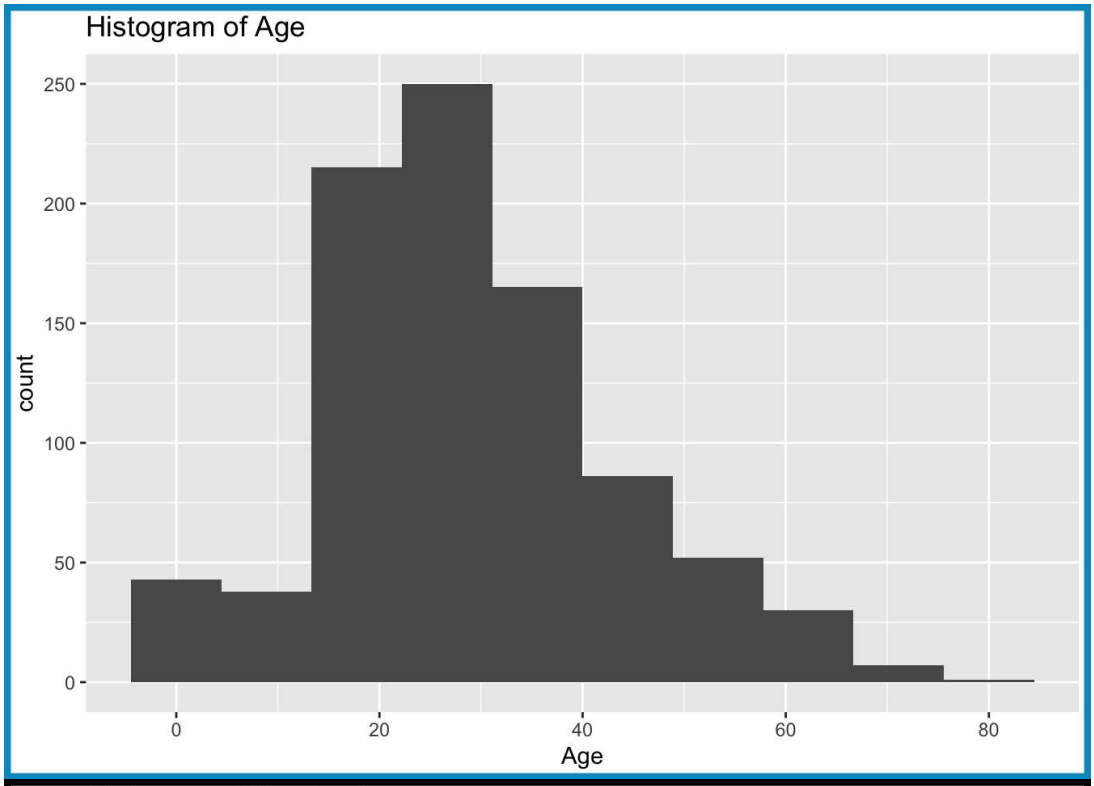
```
sum(is.na(titanic))
```

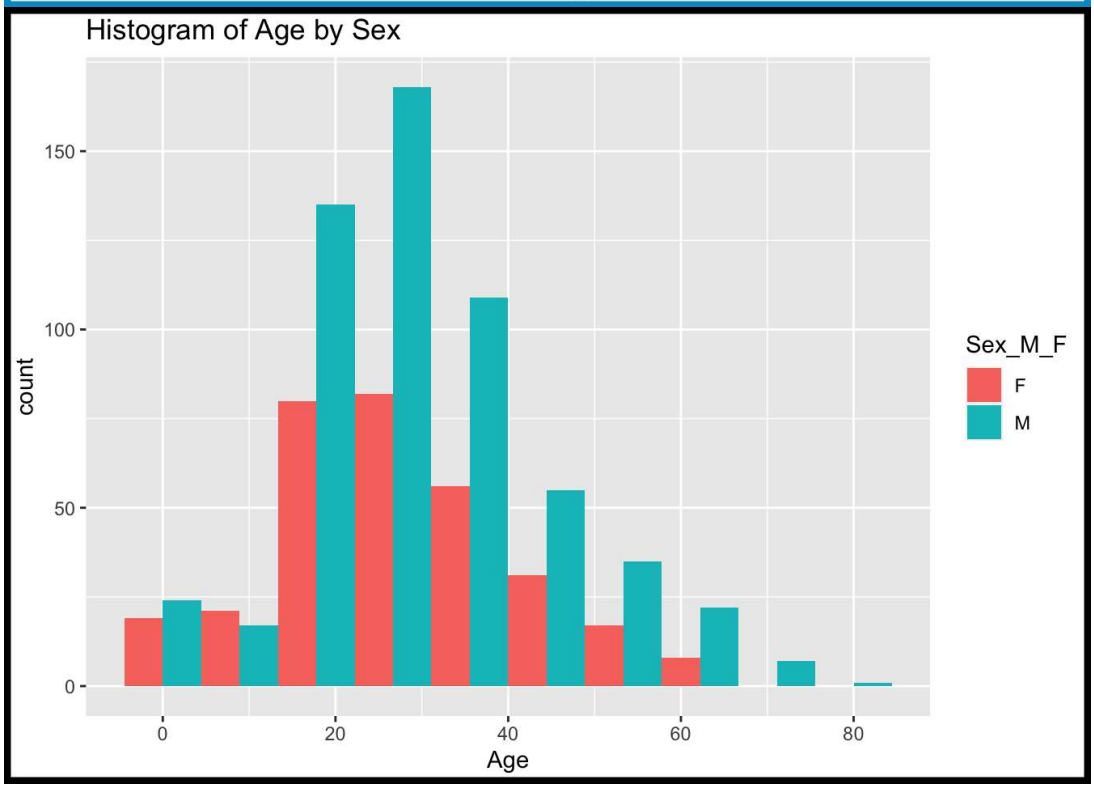
```
## [1] 0
```

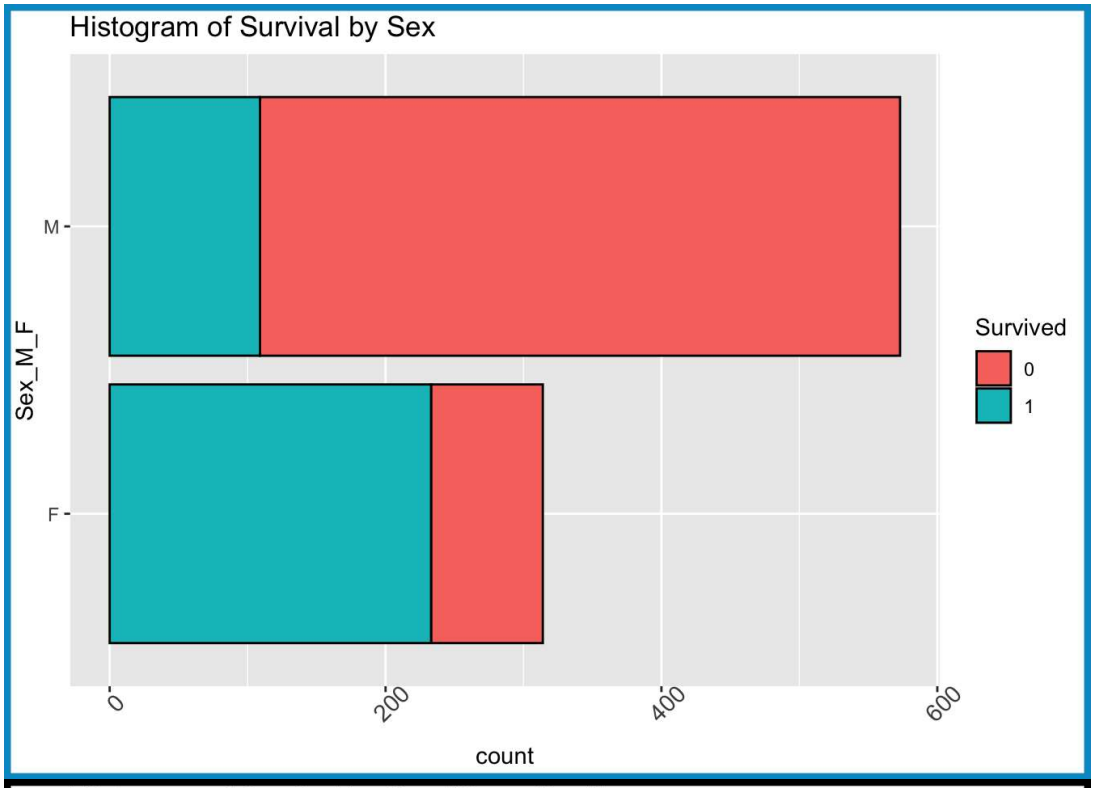
The `titanic` dataset has 887 observations of 7 attributes which are either data type: factor, integer, or number. There are no missing values.
By using the `summary` function, we are able to tell that the average age of a passenger on the Titanic was 29.4554679 years old with 80 being the oldest. The average price paid for fare was £32.3054202

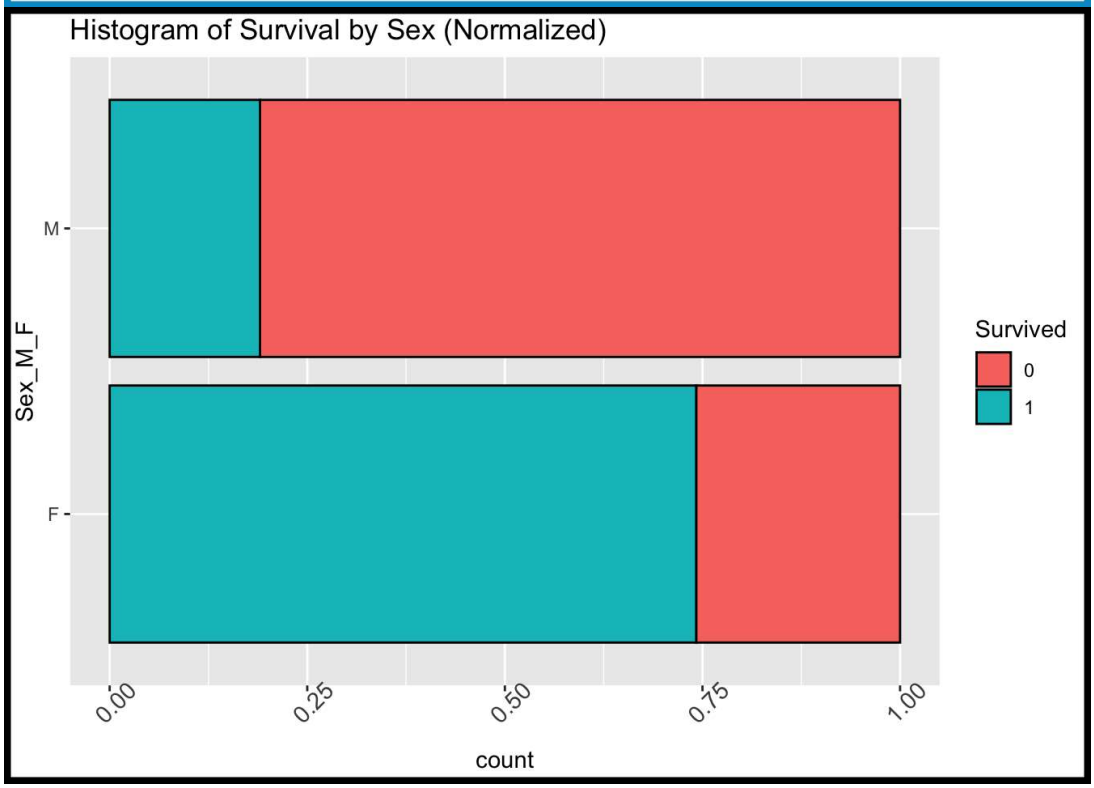
2) Data Exploration(EDA): Data Visualization, at least one histogram, one boxplot, and one overlay histogram with your conclusion

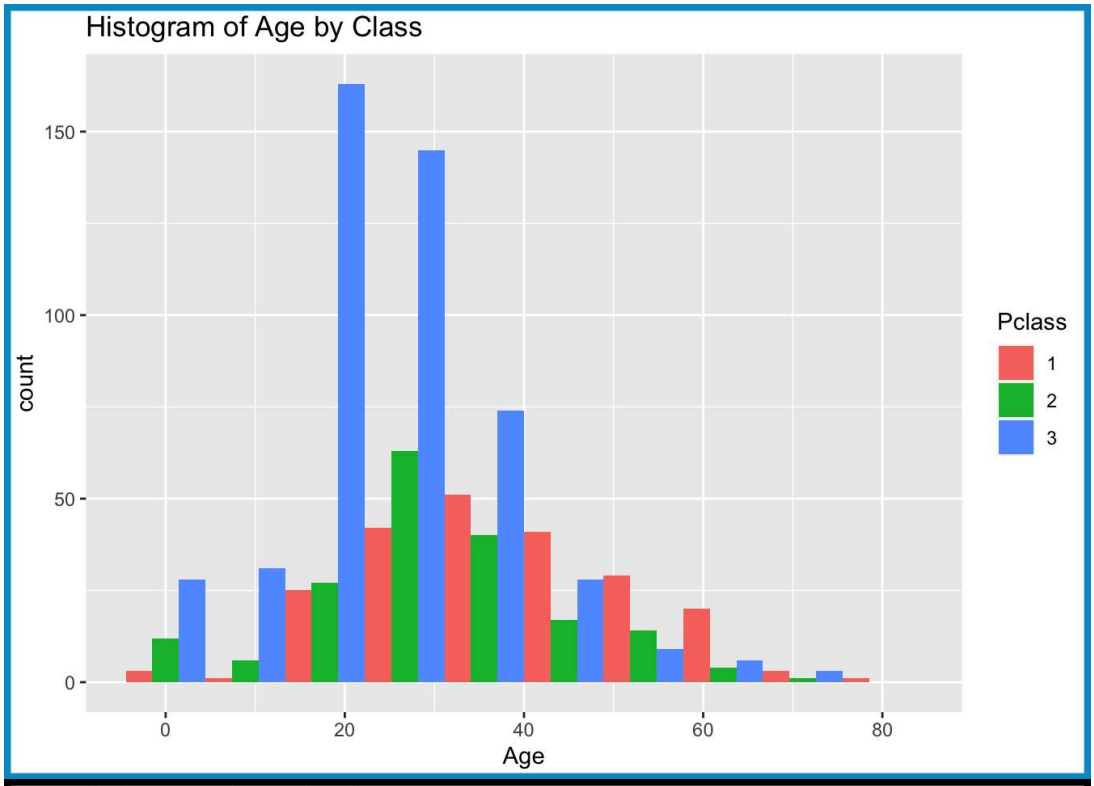
Histograms and Overlay Histograms

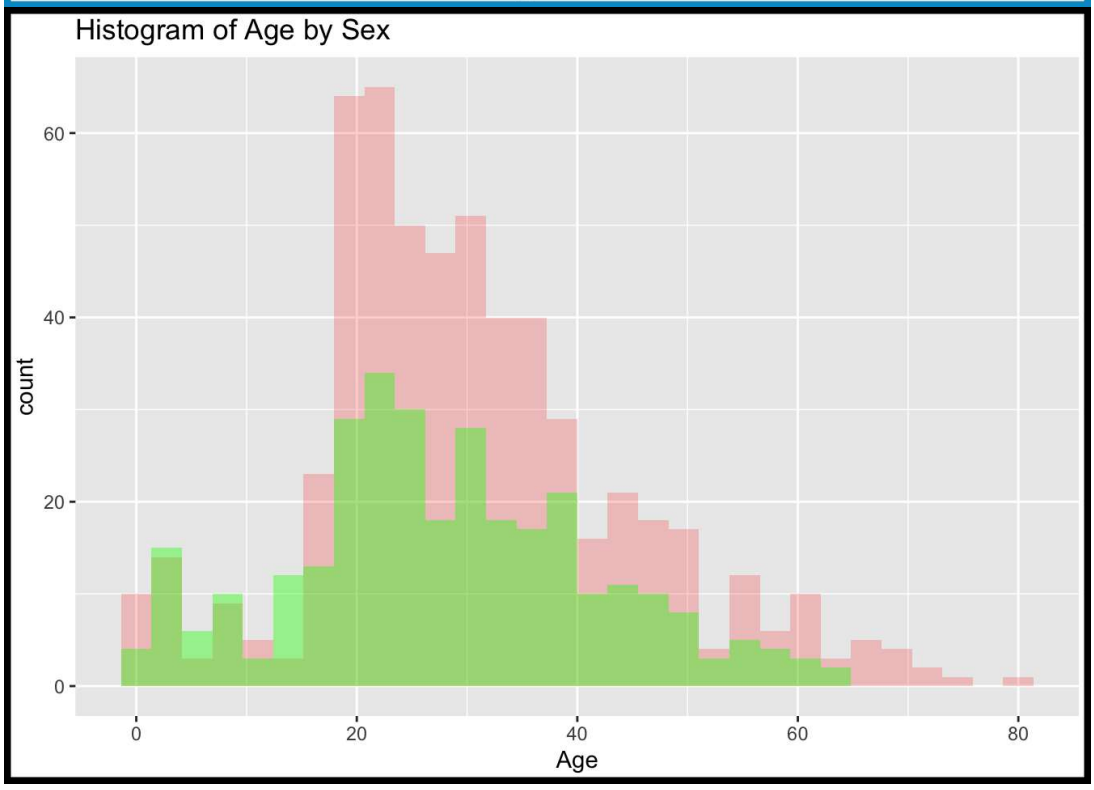

Histogram of Age
count
250
200
150
100
50
0
0
20
40
60
80
Age

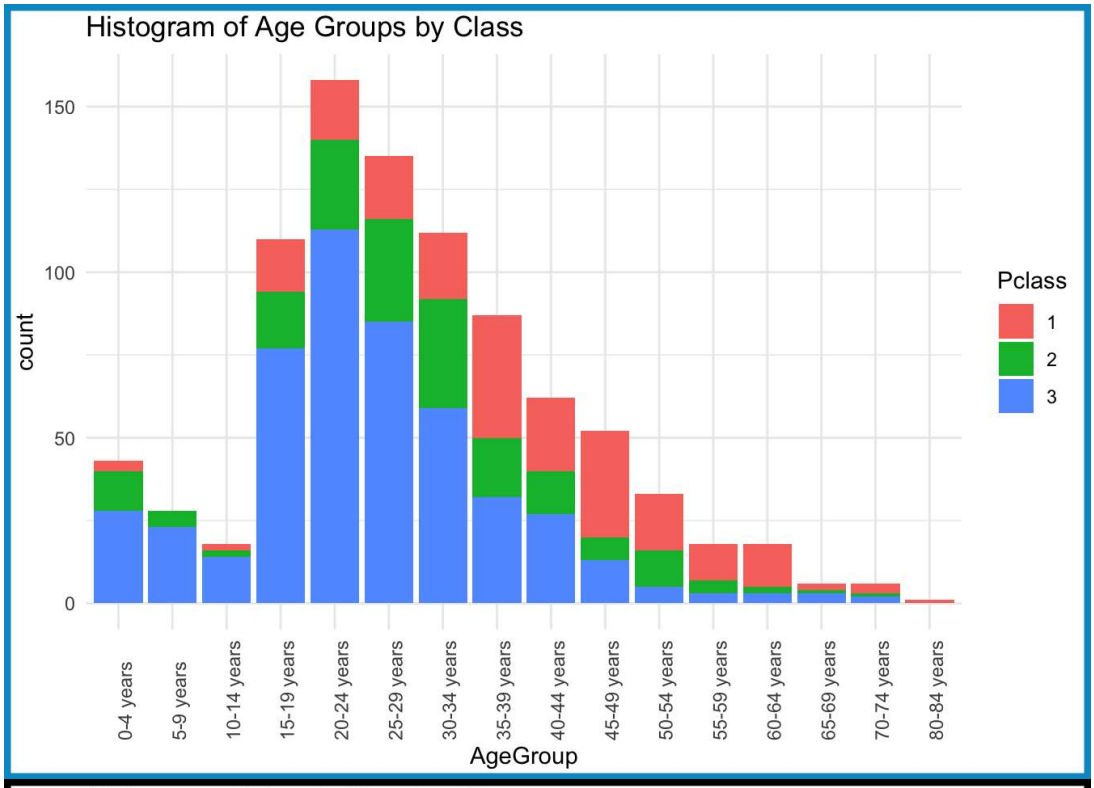

Histogram of Age by Sex
count
150
100
50
0
0
20
40
60
80
Age
Sex_M_F
F
M

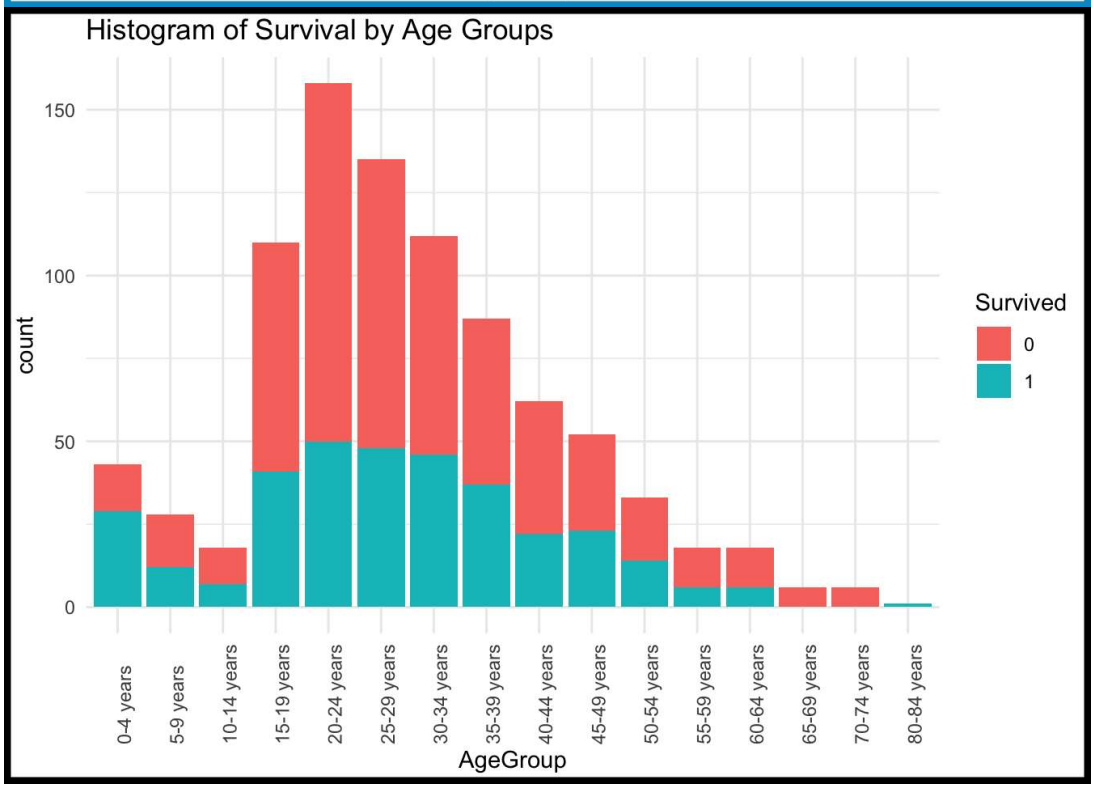

Histogram of Survival by Sex
Sex_M_F
M
F
0
200
400
600
count
Survived
0
1

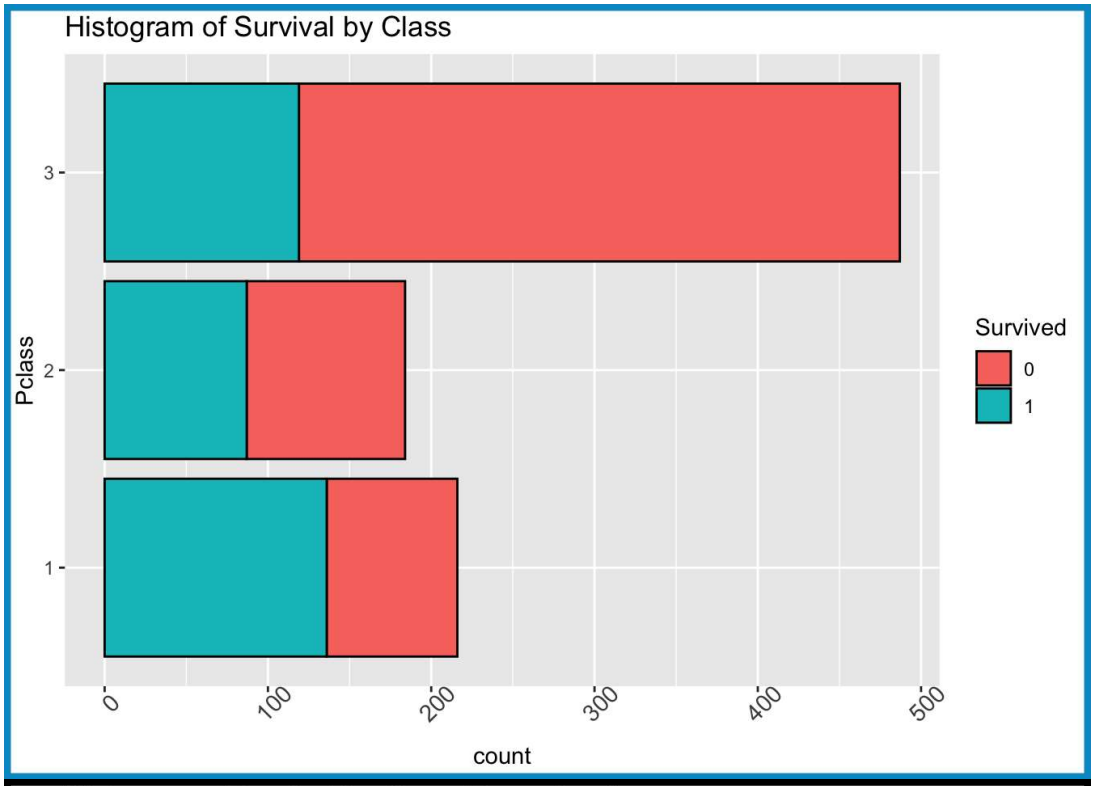

Histogram of Survival by Sex (Normalized)
Sex_M_F
M
F
0.00
0.25
0.50
0.75
1.00
count
Survived
0
1

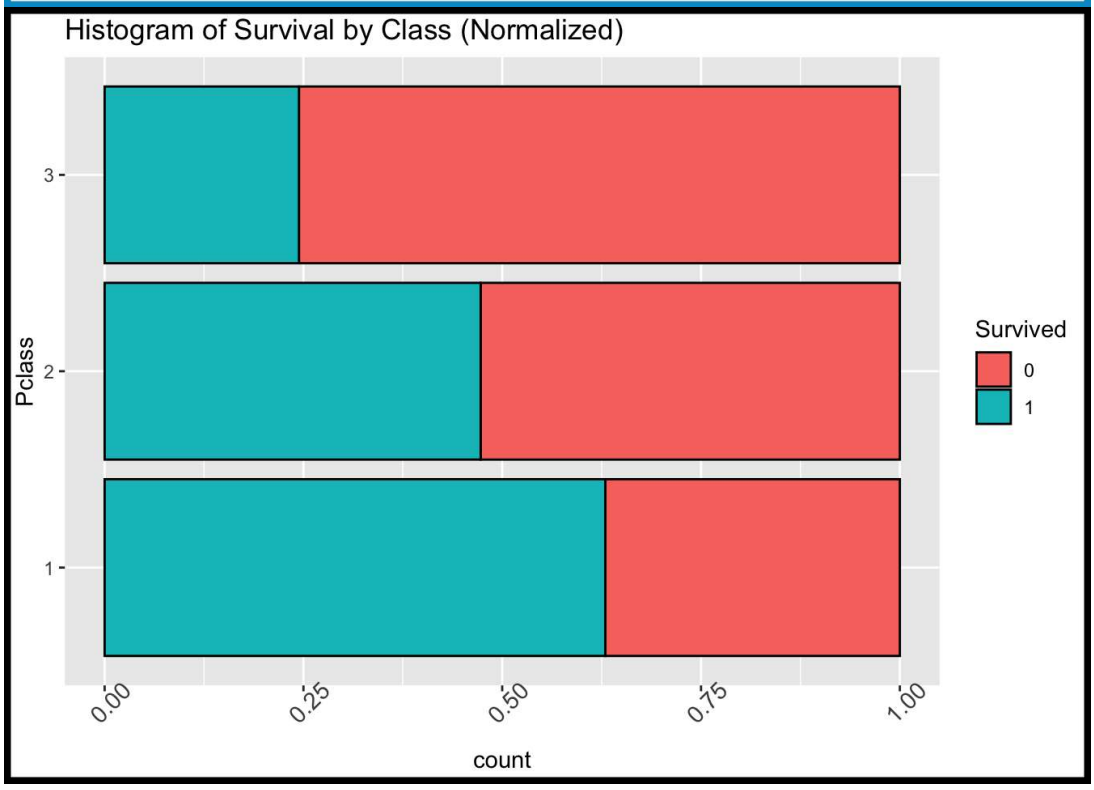

Histogram of Age by Class
count
Age
Pclass
1
2
3
0
50
100
150
0
20
40
60
80


Histogram of Age by Sex
count
Age
0
20
40
60
0
20
40
60
80

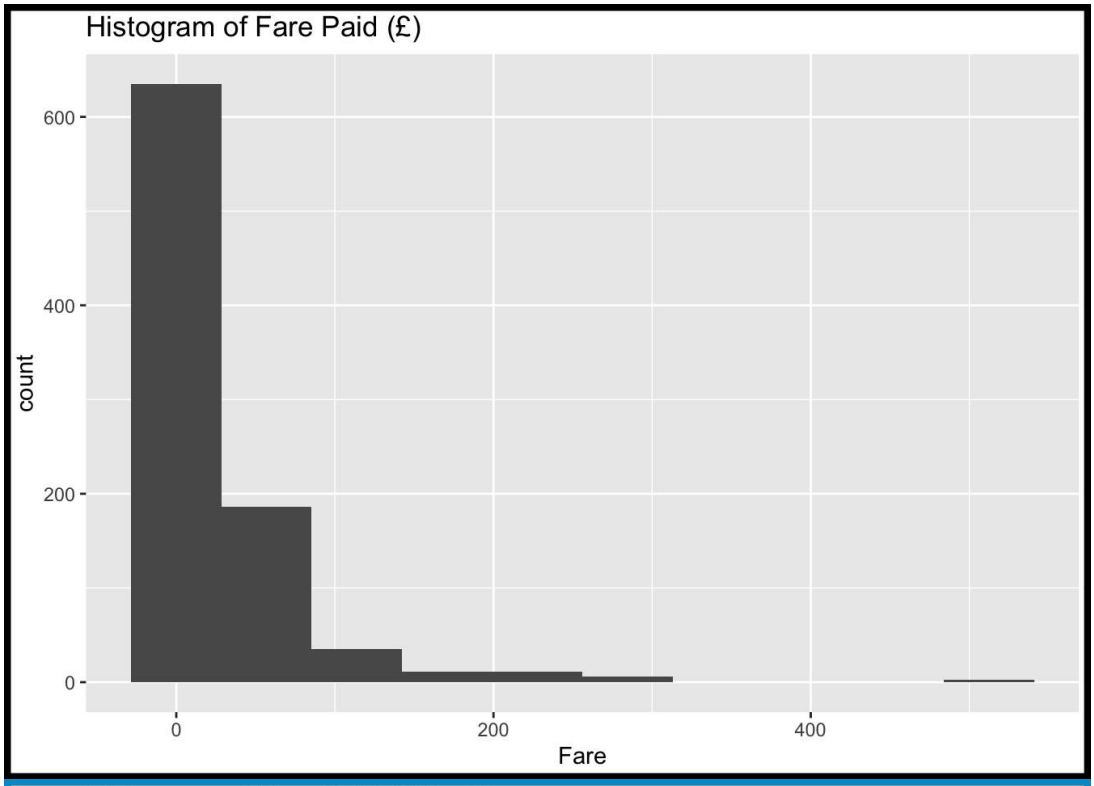

Histogram of Age Groups by Class
count
150
100
50
0
0-4 years
5-9 years
10-14 years
15-19 years
20-24 years
25-29 years
30-34 years
35-39 years
40-44 years
45-49 years
50-54 years
55-59 years
60-64 years
65-69 years
70-74 years
80-84 years
AgeGroup
Pclass
1
2
3

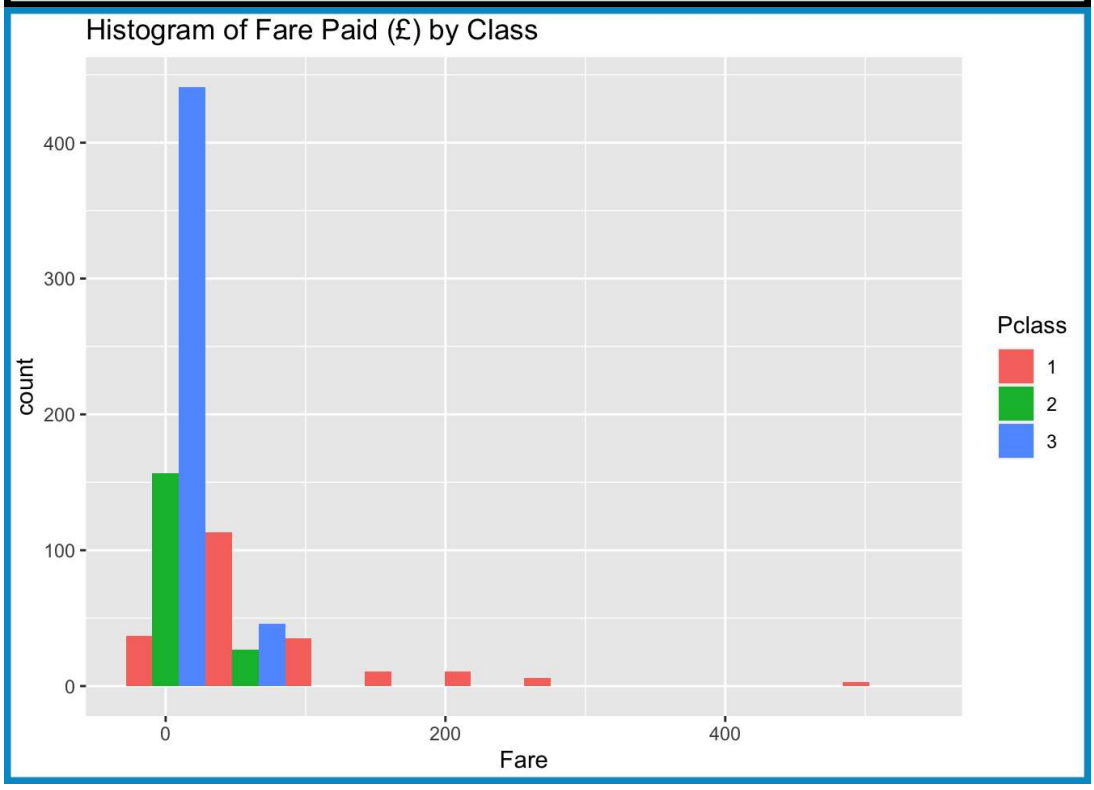

Histogram of Survival by Age Groups
count
150
100
50
0
0-4 years
5-9 years
10-14 years
15-19 years
20-24 years
25-29 years
30-34 years
35-39 years
40-44 years
45-49 years
50-54 years
55-59 years
60-64 years
65-69 years
70-74 years
80-84 years
AgeGroup
Survived
0
1

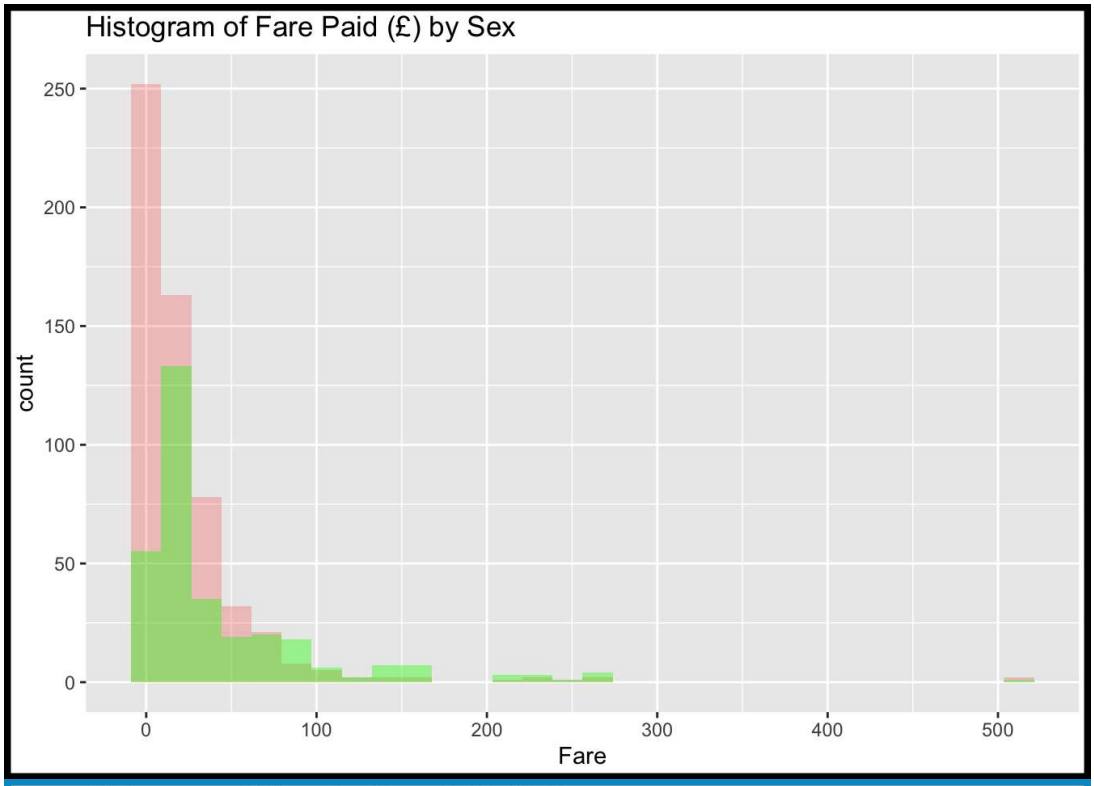


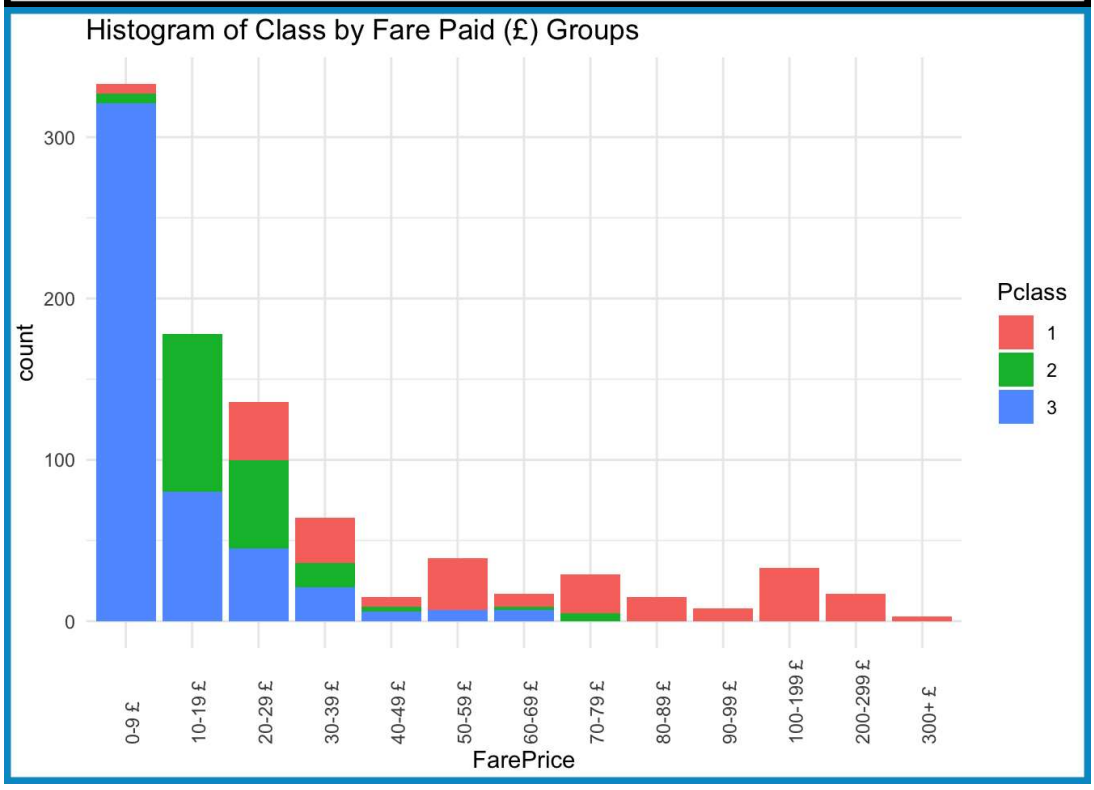


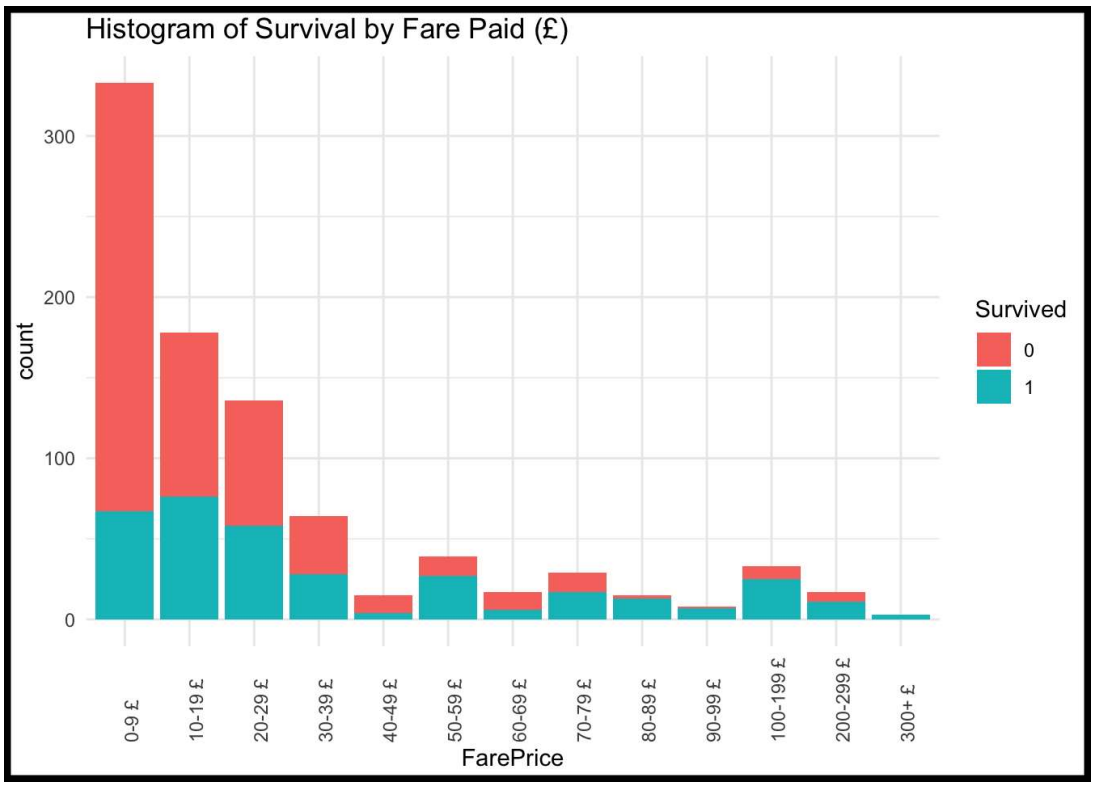
Passengers on the Titanic were mostly between the ages of 20 to 40 years of age. This remains consistent amongst the men and women. It is disheartening to see that there were many young passengers amongst all 3 classes; however, most of them were 3rd Class. 1st Class had the highest percentage of survivors.


Histogram of Fare Paid (£)
600
400
200
0
count
0
200
400
Fare


Histogram of Fare Paid (£) by Class
400
300
200
100
0
count
0
200
400
Fare
Pclass
1
2
3

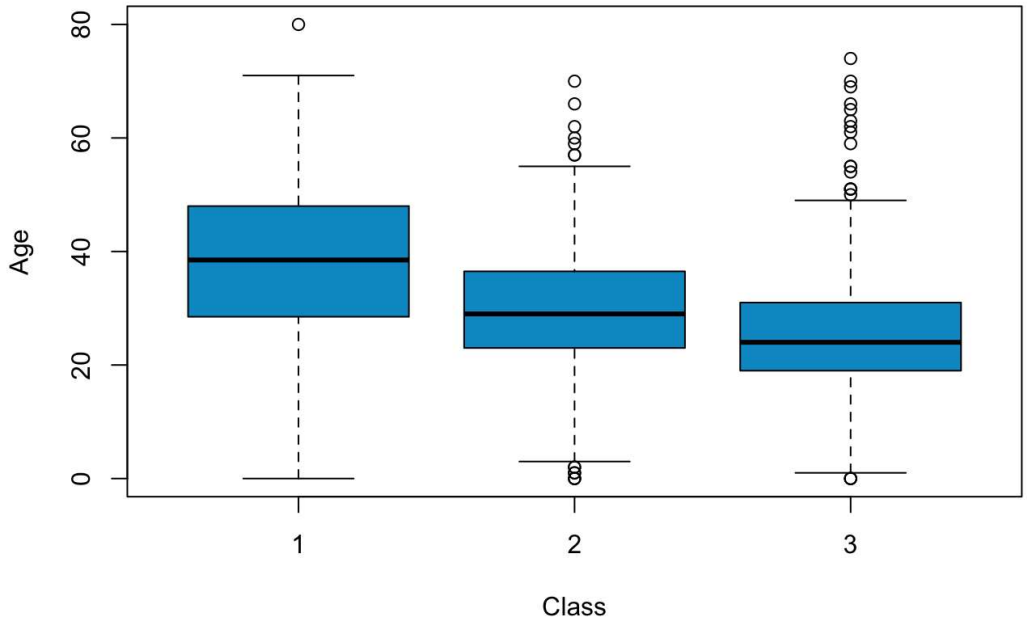

Histogram of Fare Paid (£) by Sex
count
250
200
150
100
50
0
0
100
200
300
400
500
Fare

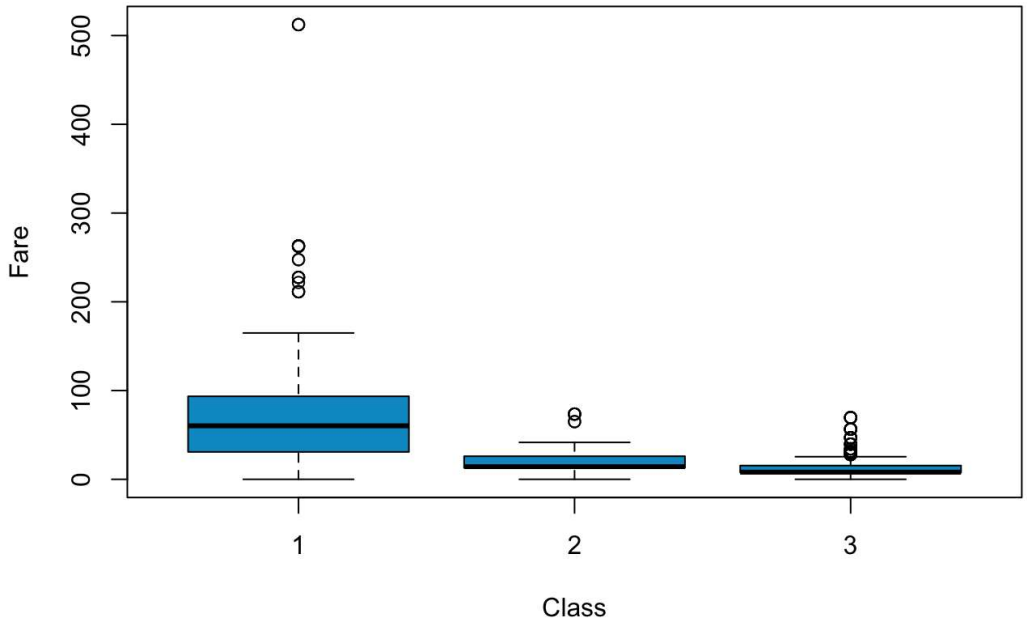

Histogram of Class by Fare Paid (£) Groups
count
300
200
100
0
0-9 £
10-19 £
20-29 £
30-39 £
40-49 £
50-59 £
60-69 £
70-79 £
80-89 £
90-99 £
100-199 £
200-299 £
300+ £
FarePrice
Pclass
1
2
3

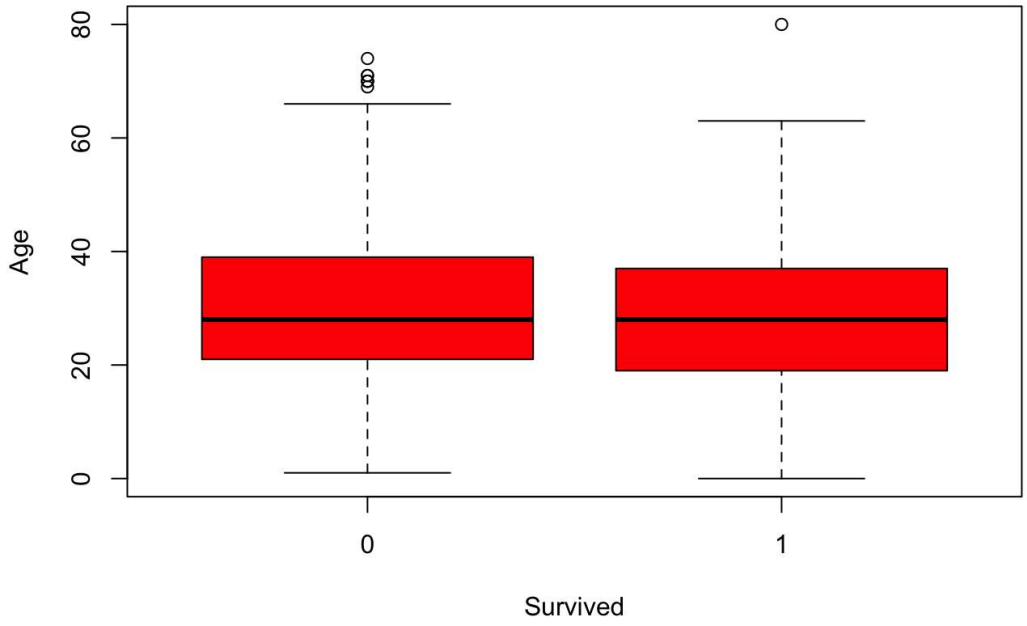


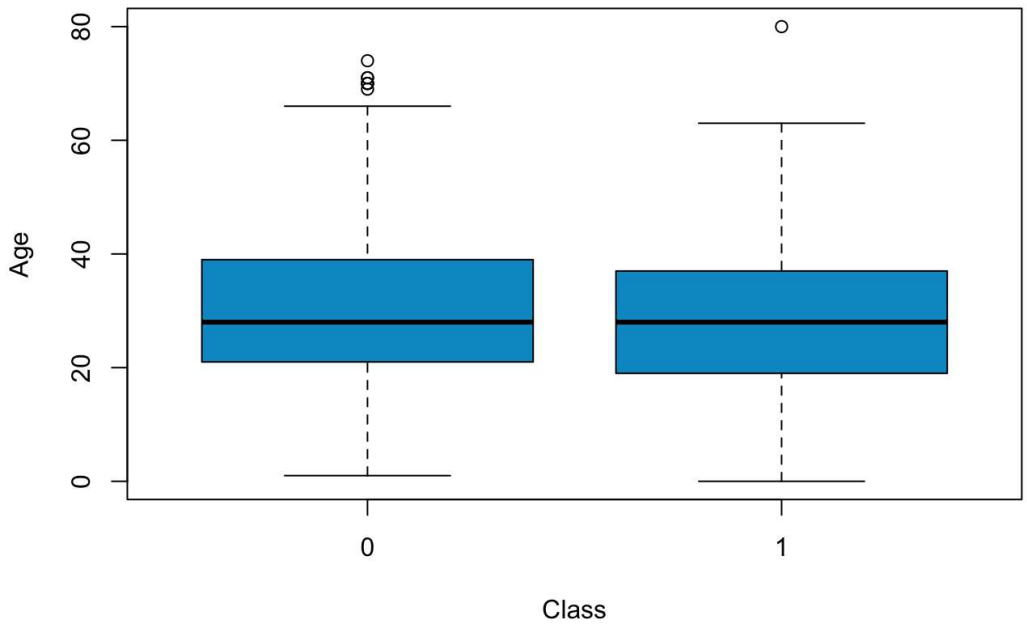
3rd Class passengers made up the majority of those traveling on the Titanic. The majority of 3rd Class passengers paid under 20£ while all 1st Class passengers paid 20£ or more. It appears that one had a better chance of survival if they paid 20£ or more, which is more likely to be 1st Class passengers.

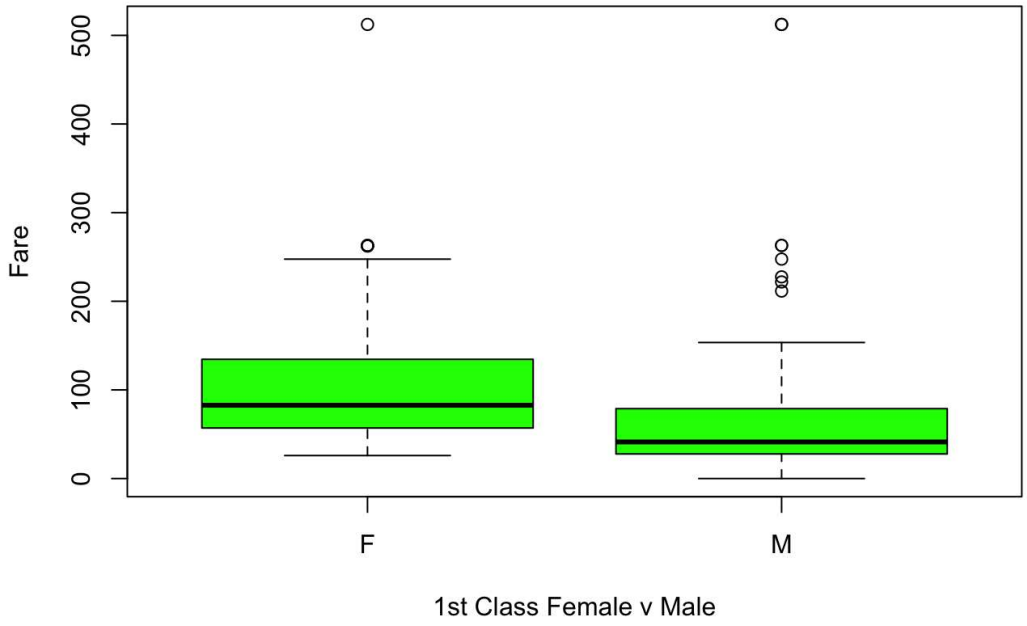
Boxplots


Age
0
20
40
60
80
1
2
3
Class

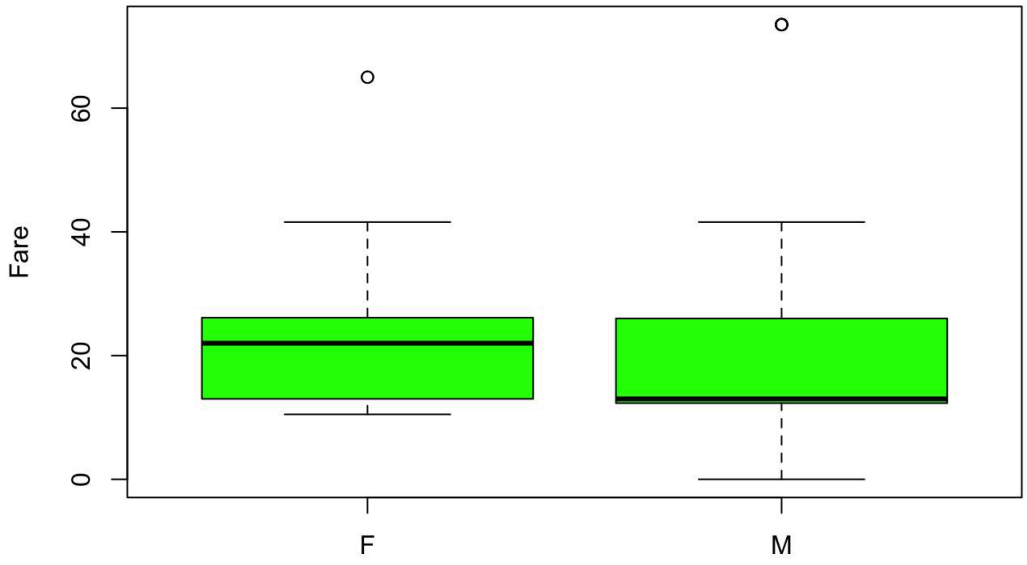

Fare
0
100
200
300
400
500
1
2
3
Class


Age
0
20
40
60
80
0
1
Survived

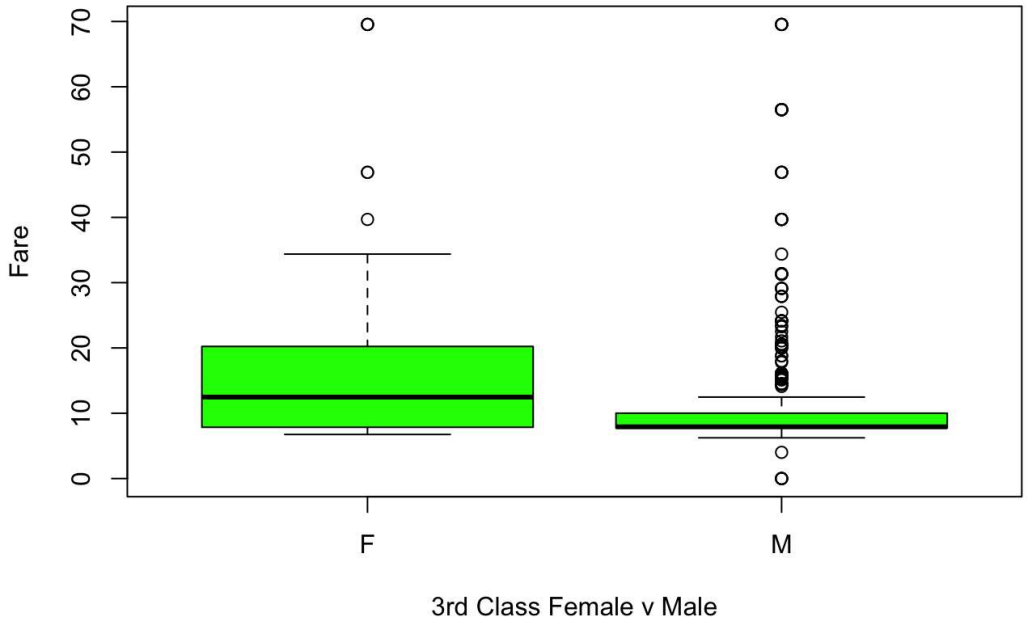

Age
0
20
40
60
80
0
1
Class



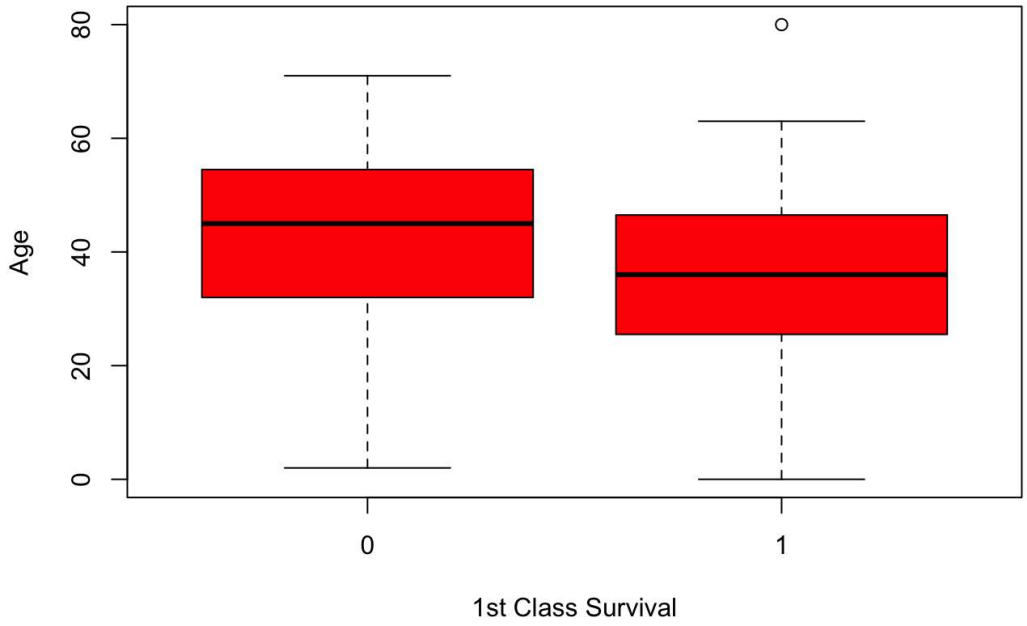
1st Class Female v Male



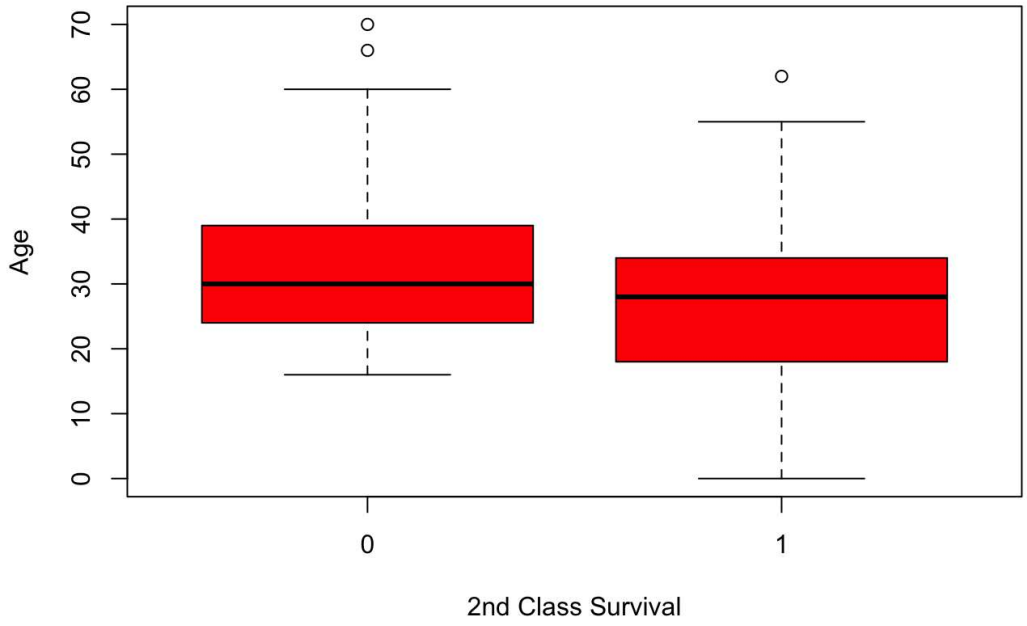
2nd Class Female v Male

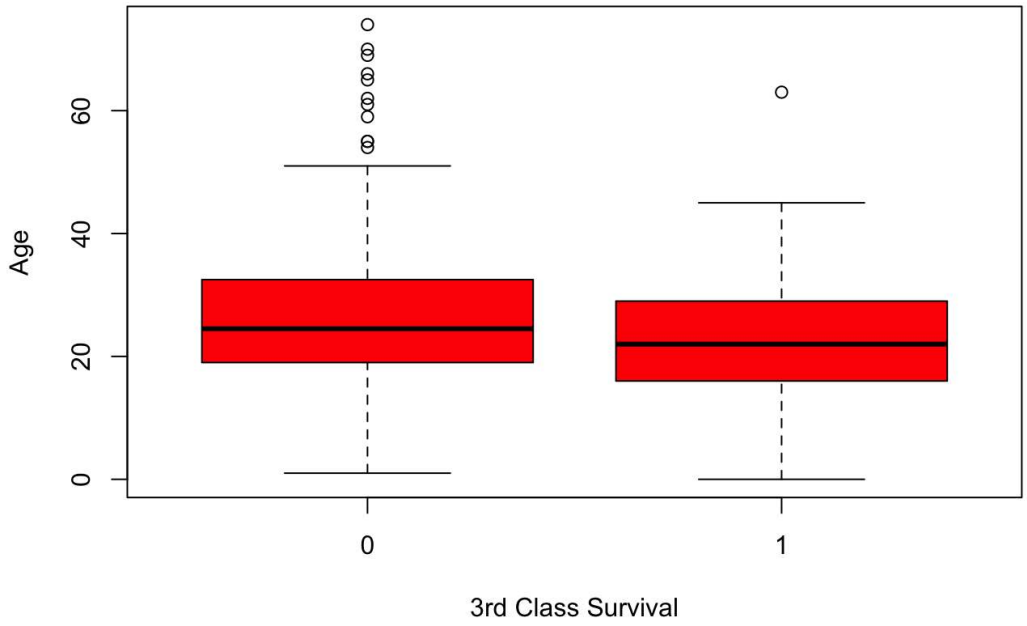


3rd Class Female v Male



1st Class Survival





3) Data Analysis:

i) Hypothesis Testing: Construct a hypothesis testing with null and alternative hypotheses. Use the appropriate test to get the conclusion.

ii) Build a linear regression model with subset selection. Please indicate all significant attributes, assess your model, and predict in a test dataset.

Split dataset into 80:20 train and test data with name `TitanicTraining` and `TitanicTest` respectively

```
i <- sample(2, nrow(titanic), replace=TRUE, prob=c(0.8, 0.2))
TitanicTraining <- titanic[i==1,]
TitanicTest <- titanic[i==2,]
str(TitanicTraining)
```

```
## 'data.frame':    729 obs. of  7 variables:
##  $ Survived : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 2 2 2 2 ...
##  $ Pclass   : Factor w/ 3 levels "1","2","3": 3 1 1 3 3 1 3 2 3 1 ...
##  $ Age      : int  22 38 35 35 27 54 27 14 4 58 ...
##  $ Fare     : num  7.25 71.28 53.1 8.05 8.46 ...
##  $ Sex_M_F  : Factor w/ 2 levels "F","M": 2 1 1 2 2 2 1 1 1 1 ...
##  $ AgeGroup : Factor w/ 18 levels "0-4 years","5-9 years",..: 5 8 8 8 6 11 6 3 1 12 ...
##  $ FarePrice: Factor w/ 13 levels "0-9 £","10-19 £",..: 1 8 6 1 1 6 2 4 2 3 ...
```

```
summary(TitanicTraining)
```

```
##  Survived Pclass       Age             Fare         Sex_M_F        AgeGroup
##  0:443    1:186   Min.   : 0.00   Min.   :  0.000   F:254   20-24 years:131
##  1:286    2:147   1st Qu.:21.00   1st Qu.:  7.925   M:475   25-29 years:109
##           3:396   Median :28.00   Median : 14.454           30-34 years: 91
##                   Mean   :29.75   Mean   : 32.768           15-19 years: 88
##                   3rd Qu.:38.00   3rd Qu.: 31.275           35-39 years: 76
##                   Max.   :80.00   Max.   :512.329           40-44 years: 50
##                                                             (Other)    :184
##      FarePrice
##  0-9 £    :274
##  10-19 £  :145
##  20-29 £  :107
##  30-39 £  : 56
##  50-59 £  : 31
##  100-199 £: 29
##  (Other)  : 87
```

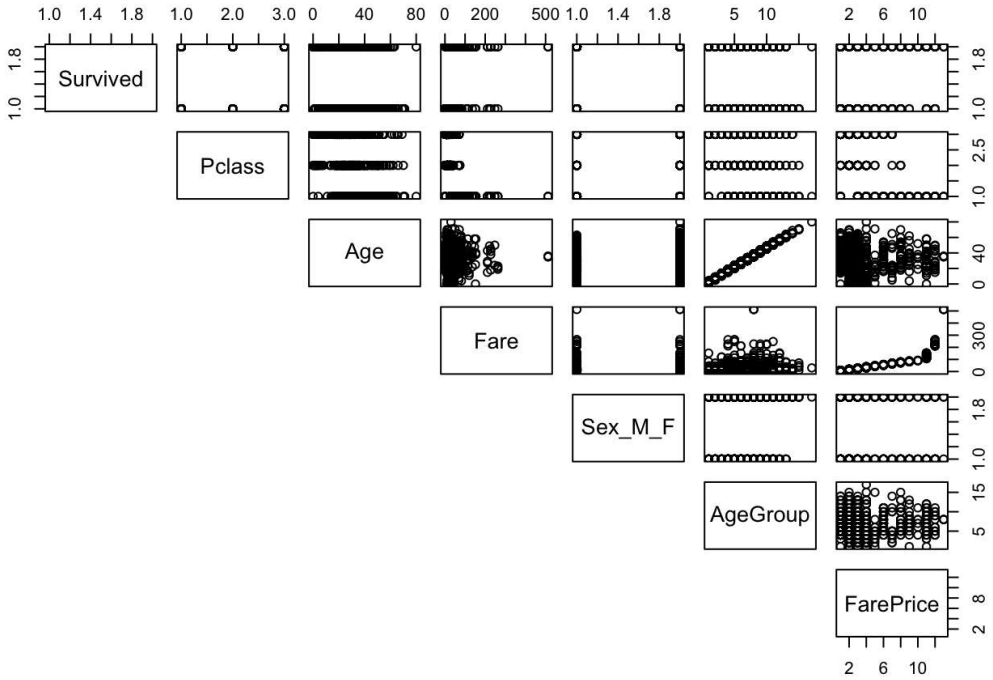
```
str(TitanicTest)
```

```
## 'data.frame':    158 obs. of  7 variables:
##  $ Survived : Factor w/ 2 levels "0","1": 2 1 1 2 1 1 1 1 1 1 ...
##  $ Pclass   : Factor w/ 3 levels "1","2","3": 3 3 1 2 3 3 3 3 2 3 ...
##  $ Age      : int  26 2 40 3 16 7 21 4 21 26 ...
##  $ Fare     : num  7.92 21.07 27.72 41.58 21.68 ...
##  $ Sex_M_F  : Factor w/ 2 levels "F","M": 1 2 2 1 2 2 2 2 2 2 ...
##  $ AgeGroup : Factor w/ 18 levels "0-4 years","5-9 years",..: 6 1 9 1 4 2 5 1 5 6 ...
##  $ FarePrice: Factor w/ 13 levels "0-9 £","10-19 £",..: 1 3 3 5 3 4 1 3 8 2 ...
```

```
summary(TitanicTest)
```

```
##  Survived Pclass      Age            Fare         Sex_M_F        AgeGroup
##  0:102    1:30   Min.   : 0.0   Min.   :  7.050   F:60    20-24 years:27
##  1: 56    2:37   1st Qu.:19.0   1st Qu.:  7.896   M:98    25-29 years:26
##           3:91   Median :26.5   Median : 14.456           15-19 years:22
##                  Mean   :28.1   Mean   : 30.170           30-34 years:21
##                  3rd Qu.:36.0   3rd Qu.: 27.900           40-44 years:12
##                  Max.   :74.0   Max.   :263.000           35-39 years:11
##                                                           (Other)    :39
##    FarePrice
##  0-9 £  :59
##  10-19 £:33
##  20-29 £:29
##  30-39 £: 8
##  50-59 £: 8
##  70-79 £: 6
##  (Other):15
```

```
summ_train_age<-summary(TitanicTraining$Age)
summ_train_fare<-summary(TitanicTraining$Fare)
train<-dim(TitanicTraining)
test<-dim(TitanicTest)
pairs(TitanicTraining, lower.panel = NULL)
```




```
cor(TitanicTraining[3:4])
```

```
##            Age      Fare
## Age  1.0000000 0.1084753
## Fare 0.1084753 1.0000000
```

> The `TitanicTraining` dataset has 729 observations of 7 variables and `TitanicTest` dataset has 158 observations of 5 attributes which are either data type: Factor, integer, or number. There are no missing values.
> By using the `summary` function, we are able to tell that the average age of a passenger on the Titanic was 29.7489712 years old with 80 being the oldest. The average price paid for fare was £32.7681516

i) Hypothesis Testing: Construct a hypothesis testing with null and alternative hypotheses.
We want to investigate the independence of `Pclass` and `Survived` in `TitanicTraining` dataset. Here

- H_0: (**null hypothesis**) The two variables are independent.
- H_1: (**alternative hypothesis**) The two variables are dependent.

Test for Independence (Categorical Data)

We wish to determine whether a passenger's chance of survival is independent of their class. Members of a random testing sample of 729 passengers on the Titanic are classified as to whether they are in 1st, 2nd, or 3rd class and whether or not they survived.

Contingency table for Class

```
##
##       1   2   3
##   0  69  76 298
##   1 117  71  98
```

Display the proportion for Class

```
##
##              1          2          3
##   0 0.09465021 0.10425240 0.40877915
##   1 0.16049383 0.09739369 0.13443073
```

Chi-squared Test for Class

```
##
##  Pearson's Chi-squared test
##
## data:  contTable
## X-squared = 83.634, df = 2, p-value < 2.2e-16
```

> Since we get a p-value less than the significance level of 0.05, we can reject the null hypothesis and conclude that the two variables are dependent. We conclude that a passenger's `Pclass` and his or her chance of survival are not independent.

Cramer's V (phi) Coefficient for Class

We can use the function `cramerV` in package `rcompanion` to calculate Cramer's V value.

```
#calculate Cramer's V
cramerV(contTable)
```

```
## Cramer V
##   0.3387
```

The range of Cramer's V value is from 0 to 1. The value we got here is very small. Even though `Pclass` and `Survived` are dependent, they don't have a very strong association.

i) Hypothesis Testing: Construct a hypothesis testing with null and alternative hypotheses.
We want to investigate the independence of `Sex_M_F` and `Survived` in `TitanicTraining` dataset. Here

- H_0: (**null hypothesis**) The two variables are independent.

- H_1: (**alternative hypothesis**) The two variables are dependent.

Test for Independence (Categorical Data)

We wish to determine whether a passenger's chance of survival is independent of their sex. Members of a random testing sample of 729 passengers on the Titanic are classified by sex and whether or not they survived.

Contingency table for Sex

```
##
##       F   M
##   0  63 380
##   1 191  95
```

Display the proportion for Sex

```
##
##              F          M
##   0 0.08641975 0.52126200
##   1 0.26200274 0.13031550
```

Chi-squared Test for Class

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contTable2
## X-squared = 209.19, df = 1, p-value < 2.2e-16
```

> Since we get a p-value less than the significance level of 0.05, we can reject the null hypothesis and conclude that the two variables are dependent. We conclude that a passenger's sex and his or her chance of survival are not independent.

Cramer's V (phi) Coefficient for Class

We can use the function `cramerV` in package `rcompanion` to calculate Cramer's V value.

```
#calculate Cramer's V
cramerV(contTable2)
```

```
## Cramer V
##   0.5386
```

The range of Cramer's V value is from 0 to 1. The value we got here is closer to 1. `Sex_M_F` and `Survived` are dependent and have a strong association.

Association between One Numerical Variable and One Categorical Variable

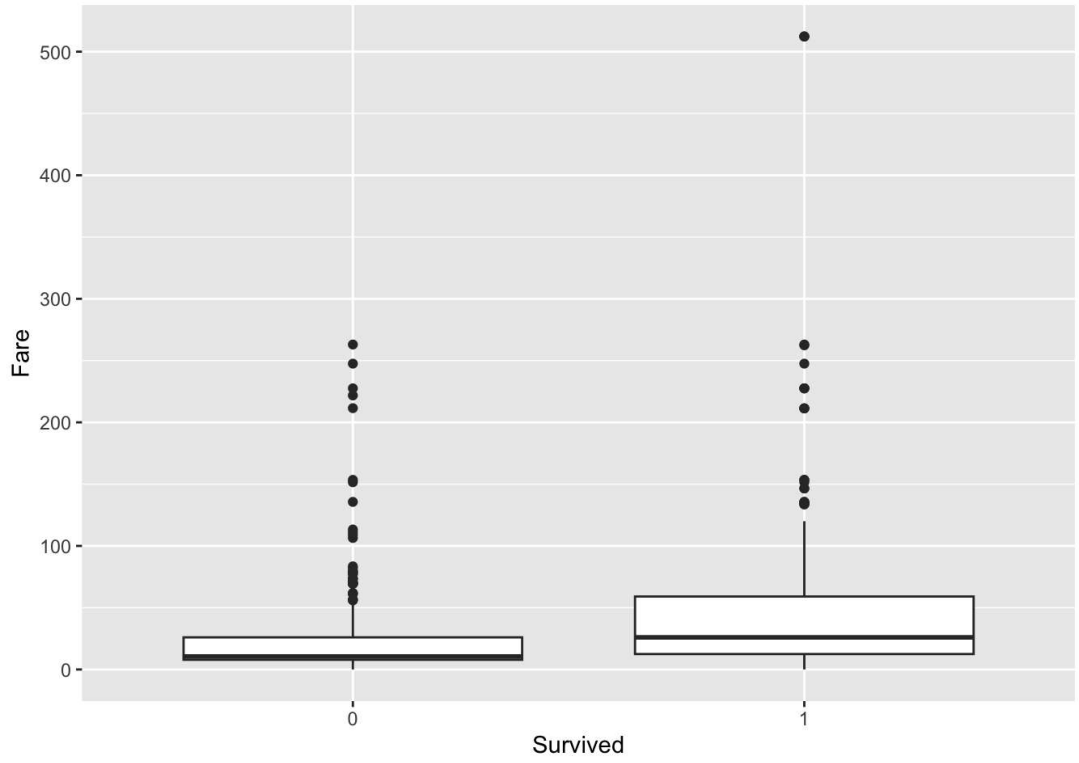
We want to investigate the independence of `Fare` and `Survived` in `TitanicTraining` dataset. Here

- H_0: (**null hypothesis**) The two variables are independent.
- H_1: (**alternative hypothesis**) The two variables are dependent.

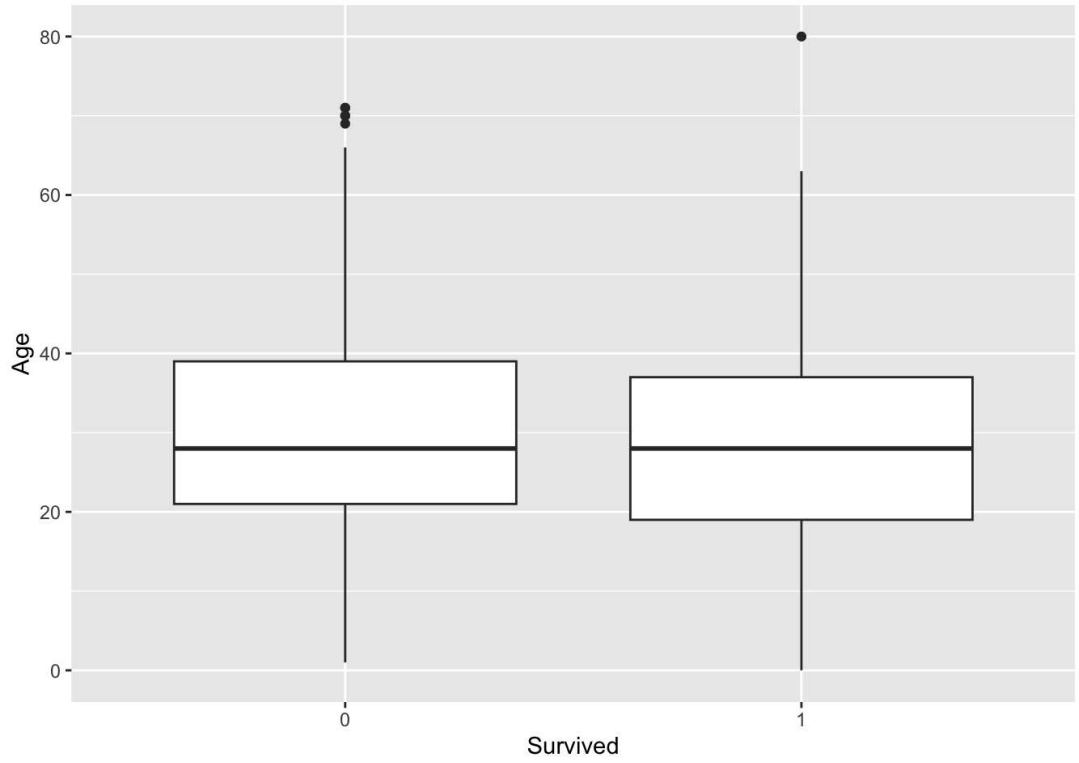
We want to investigate the independence of `Age` and `Survived` in `TitanicTraining` dataset. Here

- H_0: (**null hypothesis**) The two variables are independent.
- H_1: (**alternative hypothesis**) The two variables are dependent.

```
ggplot(TitanicTraining, aes(x = Survived , y = Fare )) + geom_boxplot()
```




```
ggplot(TitanicTraining, aes(x = Survived , y = Age )) + geom_boxplot()
```



We can use **ANOVA test** to check the association between one numerical variable and one categorical variable with `aov` function. ANOVA(AOV) is short for ANalysis Of VAriance.

```
aov1 <- aov(Fare ~ Survived, data = TitanicTraining)
summary(aov1)
```

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## Survived      1  125944  125944   51.74 1.58e-12 ***
## Residuals   727 1769563    2434
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov2 <- aov(Age ~ Survived, data = TitanicTraining)
summary(aov2)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Survived      1   1135    1135   5.763 0.0166 *
## Residuals   727 143228     197
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation:

- The Df column displays the degrees of freedom for the independent variable (the number of levels in the variable minus 1), and the degrees of freedom for the residuals (the total number of observations minus one and minus the number of levels in the independent variables).
- The Sum Sq column displays the sum of squares (a.k.a. the total variation between the group means and the overall mean).
- The Mean Sq column is the mean of the sum of squares, calculated by dividing the sum of squares by the degrees of freedom for each parameter.
- The F-value column is the test statistic from the F test. This is the mean square of each independent variable divided by the mean square of the residuals. The larger the F value, the more likely it is that the variation caused by the independent variable is real and not due to chance.
- The Pr(>F) column is the p-value of the F-statistic. This shows how likely it is that the F-value calculated from the test would have occurred if the null hypothesis of no difference among group means were true.

The p-value of the `Survived` variable is very low (1.5771895^{-12}), so it appears that the `Fare` has a real impact on the `Survived`.
The p-value of the `Survived` variable is slightly low (0.0166193), so it appears that the `Age` has a slight impact on the `Survived`.

ii) Build a linear regression model with subset selection. Please indicate all significant attributes, assess your model, and predict in a test dataset.

Simple Linear Regression Model

```
fitlm <- lm(Fare~Age, data=TitanicTraining[1:5])
summary(fitlm)
```

```
##
## Call:
## lm(formula = Fare ~ Age, data = TitanicTraining[1:5])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -40.34 -22.89 -17.51   1.46 477.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.0748     4.3966   4.793 1.99e-06 ***
## Age           0.3931     0.1336   2.942  0.00336 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.76 on 727 degrees of freedom
## Multiple R-squared:  0.01177,    Adjusted R-squared:  0.01041
## F-statistic: 8.656 on 1 and 727 DF,  p-value: 0.003363
```

```
lm_summary<-summary(fitlm)
#calculate residual sum of squares
rss1<-deviance(fitlm)
```

A simple linear regression of `Fare` by `Age` using `TitanicTest` shows that there is a weak relationship between the target (`Fare`) and the input (`Age`) because the p-value is 0.0033628. The residual standard error (RSE) is 50.7604235 with R^2 of 0.0117669. However, the F-statistic is greater than 1 (8.6563852), but not by much.

*Predict `Fare` in `TitanicTest` and calculate `MAE` and `MSE` .

```
ypred <-predict(object = fitlm, newdata = TitanicTest[1:5])
sum1<-summary(ypred)
mae1<-MAE(y_pred = ypred, y_true = TitanicTest$Fare)
mse1<-MSE(y_pred = ypred, y_true = TitanicTest$Fare)
```

Min: 21.0748424
1st Qu: 28.5430964
Median: 31.4910915
Mean: 32.1204946
3rd Qu: 35.2252185
Max: 50.1617266
After prediction, we can get MAE is 26.0524472 and MSE is 1866.2324597.

Multiple Linear Regression

```
fitlm2 <- lm(Fare~., data=TitanicTraining[1:5])
matrix_coef <- summary(fitlm2)$coefficients  # Extract coefficients in matrix
lm_summary2<-summary(fitlm2)
summary(fitlm2)
```

```
##
## Call:
## lm(formula = Fare ~ ., data = TitanicTraining[1:5])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -82.02 -11.67  -3.48   4.90 431.08
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 104.6315     7.1817  14.569  < 2e-16 ***
## Survived1     1.9188     3.9539   0.485 0.627619
## Pclass2     -67.0104     4.7079 -14.234  < 2e-16 ***
## Pclass3     -73.2631     4.3295 -16.922  < 2e-16 ***
## Age          -0.4483     0.1217  -3.685 0.000245 ***
## Sex_M_FM     -9.1611     3.7768  -2.426 0.015525 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.86 on 723 degrees of freedom
## Multiple R-squared:  0.3632, Adjusted R-squared:  0.3588
## F-statistic: 82.46 on 5 and 723 DF,  p-value: < 2.2e-16
```

```
#calculate residual sum of squares
rss2<-deviance(fitlm2)
```

A multiple linear regression of `Fare` by all features (except group attributes: AgeGroup and FairPrice) in `TitanicTraining` shows that there is a strong relationship between the target (`Fare`) and multiple inputs because the p-value is very small:
`Pclass2` 1.0405765^{-40}
`Pclass3` 2.3023181^{-54}
`Age` 2.3023181^{-54}
`Sex_M_FM` 2.4544235^{-4}
The residual standard error (RSE) is 40.8610563 with R^2 of 0.3631576. Lastly, the F-statistic is greater than 1 (82.4577341).

- Predict `Fare` in `TitanicTest` and calculate `MAE` and `MSE` .

```
ypred2 <-predict(object = fitlm2, newdata =TitanicTest[1:5])
sum2<-summary(ypred2)
mae2<-MAE(y_pred = ypred2, y_true = TitanicTest$Fare)
mse2<-MSE(y_pred = ypred2, y_true = TitanicTest$Fare)
```

Min: -10.9690474
1st Qu: 12.3440458
Median: 19.0570087
Mean: 29.1426129
3rd Qu: 28.5304478
Max: 103.7348457
After prediction, we can get MAE is 16.5275685 and MSE is 996.4735854.

Subset Selection Linear Regression Model

Forward Stepwise

```
# Create a null model
intercept_only <- lm(Fare ~ 1, data=TitanicTraining[1:5])
# Create a full model
all <- lm(Fare~., data=TitanicTraining[1:5])
# perform forward step-wise regression
forward <- stepAIC (intercept_only, direction='forward',scope = formula(all))
```

```
## Start:  AIC=5734.36
## Fare ~ 1
##
##            Df Sum of Sq     RSS    AIC
## + Pclass    2    639185 1256322 5438.5
## + Survived  1    125944 1769563 5686.2
## + Sex_M_F   1     61987 1833520 5712.1
## + Age       1     22304 1873203 5727.7
## <none>                  1895507 5734.4
##
## Step:  AIC=5438.53
## Fare ~ Pclass
##
##            Df Sum of Sq     RSS    AIC
## + Age       1     32739 1223583 5421.3
## + Sex_M_F   1     23868 1232454 5426.5
## + Survived  1     15599 1240723 5431.4
## <none>                  1256322 5438.5
##
## Step:  AIC=5421.28
## Fare ~ Pclass + Age
##
##            Df Sum of Sq     RSS    AIC
## + Sex_M_F   1   16050.2 1207533 5413.7
## + Survived  1    6619.7 1216963 5419.3
## <none>                  1223583 5421.3
##
## Step:  AIC=5413.65
## Fare ~ Pclass + Age + Sex_M_F
##
##            Df Sum of Sq     RSS    AIC
## <none>                  1207533 5413.7
## + Survived  1     393.2 1207140 5415.4
```

```
# view results of forward stepwise regression
forward$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Fare ~ 1
##
## Final Model:
## Fare ~ Pclass + Age + Sex_M_F
##
##
##        Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                              728    1895507 5734.363
## 2  + Pclass  2 639185.00       726    1256322 5438.527
## 3     + Age  1  32739.41       725    1223583 5421.277
## 4 + Sex_M_F  1  16050.24       724    1207533 5413.651
```

```
aic_all<-forward$anova
aic1<-aic_all$AIC
# view final model
lm_summary3<-summary(forward)
#calculate residual sum of squares
rss3<-deviance(forward)
```

The model resulting from a forward stepwise regression with `TitanicTraining` is:

`Fare ~ Pclass + Age + Sex_M_F`

- Predict `Fare` in `TitanicTest` and calculate `MAE` and `MSE`.

```
ypred_forward <-predict(object = forward, newdata = TitanicTest[1:5])
sum3<-summary(ypred_forward)
mae3<-MAE(y_pred = ypred_forward, y_true = TitanicTest$Fare)
mse3<-MSE(y_pred = ypred_forward, y_true = TitanicTest$Fare)
```

Min: -11.3400751
1st Qu: 12.7005057
Median: 19.4983151
Mean: 29.2280135
3rd Qu: 28.6687613
Max: 105.9041253
After prediction, we can get MAE is 16.4820697 and MSE is 992.9390946.

Backward Stepwise

```
backward <- stepAIC (all, direction='backward')
```

```
## Start:  AIC=5415.41
## Fare ~ Survived + Pclass + Age + Sex_M_F
##
##            Df Sum of Sq     RSS    AIC
## - Survived  1       393 1207533 5413.7
## <none>                  1207140 5415.4
## - Sex_M_F   1      9824 1216963 5419.3
## - Age       1     22677 1229816 5427.0
## - Pclass    2    523313 1730452 5673.9
##
## Step:  AIC=5413.65
## Fare ~ Pclass + Age + Sex_M_F
##
##           Df Sum of Sq     RSS    AIC
## <none>                 1207533 5413.7
## - Sex_M_F  1     16050 1223583 5421.3
## - Age      1     24922 1232454 5426.5
## - Pclass   2    595997 1803529 5702.1
```

```
backward$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Fare ~ Survived + Pclass + Age + Sex_M_F
##
## Final Model:
## Fare ~ Pclass + Age + Sex_M_F
##
##
##         Step Df Deviance Resid. Df Resid. Dev      AIC
## 1                              723    1207140 5415.414
## 2 - Survived  1 393.2038       724    1207533 5413.651
```

```
aic_all2<-backward$anova
aic2<-aic_all2$AIC
lm_summary4<-summary(backward)
#calculate residual sum of squares
rss4<-deviance(backward)
```

The model resulting from a backward stepwise regression with `TitanicTraining[1:5]` is:

`Fare ~ Pclass + Age + Sex_M_F`

*Predict `Fare` in `TitanicTest` and calculate `MAE` and `MSE`.

```
#Get MAE and MSE
ypred_backward <-predict(object = backward, newdata = TitanicTest[1:5])
sum4<-summary(ypred_backward)
mae4<-MAE(y_pred = ypred_backward, y_true = TitanicTest$Fare)
mse4<-MSE(y_pred = ypred_backward, y_true = TitanicTest$Fare)
```

Min: -11.3400751
1st Qu: 12.7005057
Median: 19.4983151
Mean: 29.2280135
3rd Qu: 28.6687613
Max: 105.9041253
After prediction, we can get MAE is 16.4820697 and MSE is 992.9390946.

Model Assessment

Comparison of all the linear regression models:

Model 1 (simple linear regression) is the worst, only having R^2: 0.0117669. Model 3 (forward regression) & 4 (backward regression) are identical and have the best results; mostly by having larger F-statistic values than Model 2. Model 2 has a slightly smaller `RSS`, slightly lower R^2, slightly smaller `MAE`, slightly larger `MSE` and smaller F-statistic.

Model 1: Simple linear regression of `Fare` by `Age` using `TitanicTraining[1:5]`
R^2: 0.0117669
RSE: 50.7604235
RSS: 1.8732032^{6}
MAE: 26.0524472
MSE: 1866.2324597
F-statistic: 8.6563852

Model 2: Multiple linear regression of `Fare` by all features in `TitanicTraining[1:5]`
R^2: 0.3631576
RSE: 40.8610563
RSS: 1.2071395^{6}
MAE: 16.5275685
MSE: 996.4735854
F-statistic: 82.4577341

Model 3: Forward stepwise regression with `TitanicTraining[1:5]`
`Fare ~ Pclass + Age + Sex_M_F`
R^2: 0.3629501
RSE: 40.8394773
RSS: 1.2075327^{6}
MAE: 16.4820697
MSE: 992.9390946
F-statistic: 103.1221818
AIC: 5413.6513017

Model 4: Backward stepwise regression with `TitanicTraining[1:5]`
`Fare ~ Pclass + Age + Sex_M_F`
R^2: 0.3629501
RSE: 40.8394773
RSS: 1.2075327^{6}
MAE: 16.4820697
MSE: 992.9390946
F-statistic: 103.1221818
AIC: 5413.6513017