R Notebook

Load and check Data

```
# Loading packages
library(ggplot2) # visualization
library('randomForest') # classification algorithm
```

```
## randomForest 4.6-12
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(stringr)
library(missForest)
```

```
## Warning: package 'missForest' was built under R version 3.3.3
```

```
## Loading required package: foreach
```

```
## Loading required package: itertools
```

```
## Warning: package 'itertools' was built under R version 3.3.3
```

```
## Loading required package: iterators
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:randomForest':
##
##     combine
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Import and combine Data

```
train <- read.csv('C:/Users/Aditya/Documents/R/Projects/Titanic/train.csv', stringsAsFactors = F)
test  <- read.csv('C:/Users/Aditya/Documents/R/Projects/Titanic/test.csv', stringsAsFactors = F)

full  <- bind_rows(train, test) # bind training & test data

full[which(is.na(full$Survived)),2] <- 0 #addressing NA values due to combining of train and test datas

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

Metadata information

```
# check data
str(full)
```

```
## 'data.frame':    1309 obs. of  12 variables:
##  $ PassengerId: int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Survived   : Factor w/ 2 levels "0","1": 1 2 2 2 1 1 1 1 2 2 ...
##  $ 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)"
##  $ 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  "" "C85" "" "C123" ...
##  $ Embarked   : chr  "S" "C" "S" "S" ...
```

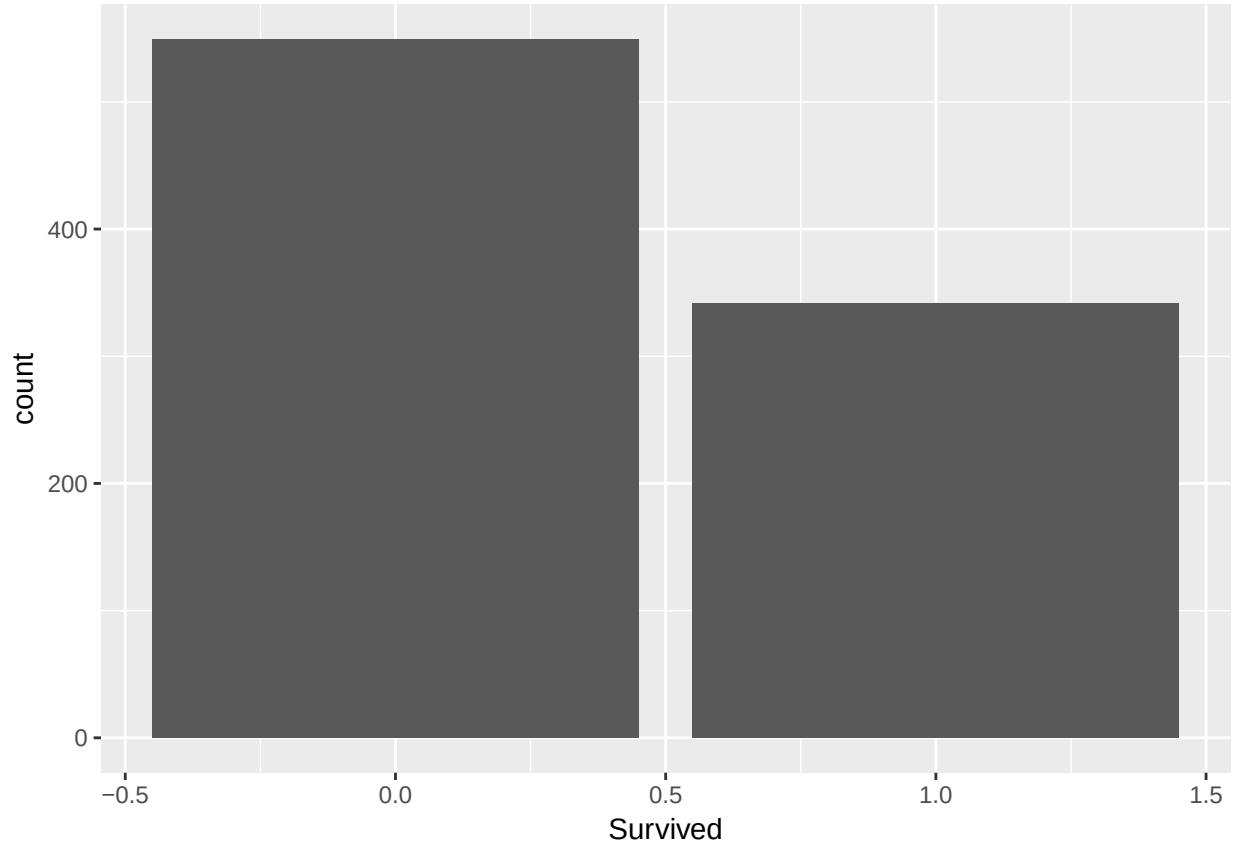
First few rows of the dataset

```
head(full)
```

```
##   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
## 1                             Braund, Mr. Owen Harris   male  22     1
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
## 3                              Heikkinen, Miss. Laina female  26     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
## 5                            Allen, Mr. William Henry   male  35     0
## 6                                    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
```

Survival Rates

```
ggplot(train , aes(x = Survived)) + geom_bar()
```




```
table(train$Survived)
```

```
##
##   0   1
## 549 342
```

549 people persished , 342 survived

Missing Values

```
sapply(full, function(full) sum(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           0           0
```

Above table shows the number of missing values in each feature. 263 observations are missing in Age and 1 in Fare.

imputing Blank values in Embarked

```
which(full$Embarked == "")
```

```
## [1]  62 830
```

```
full[c(62, 830),]
```

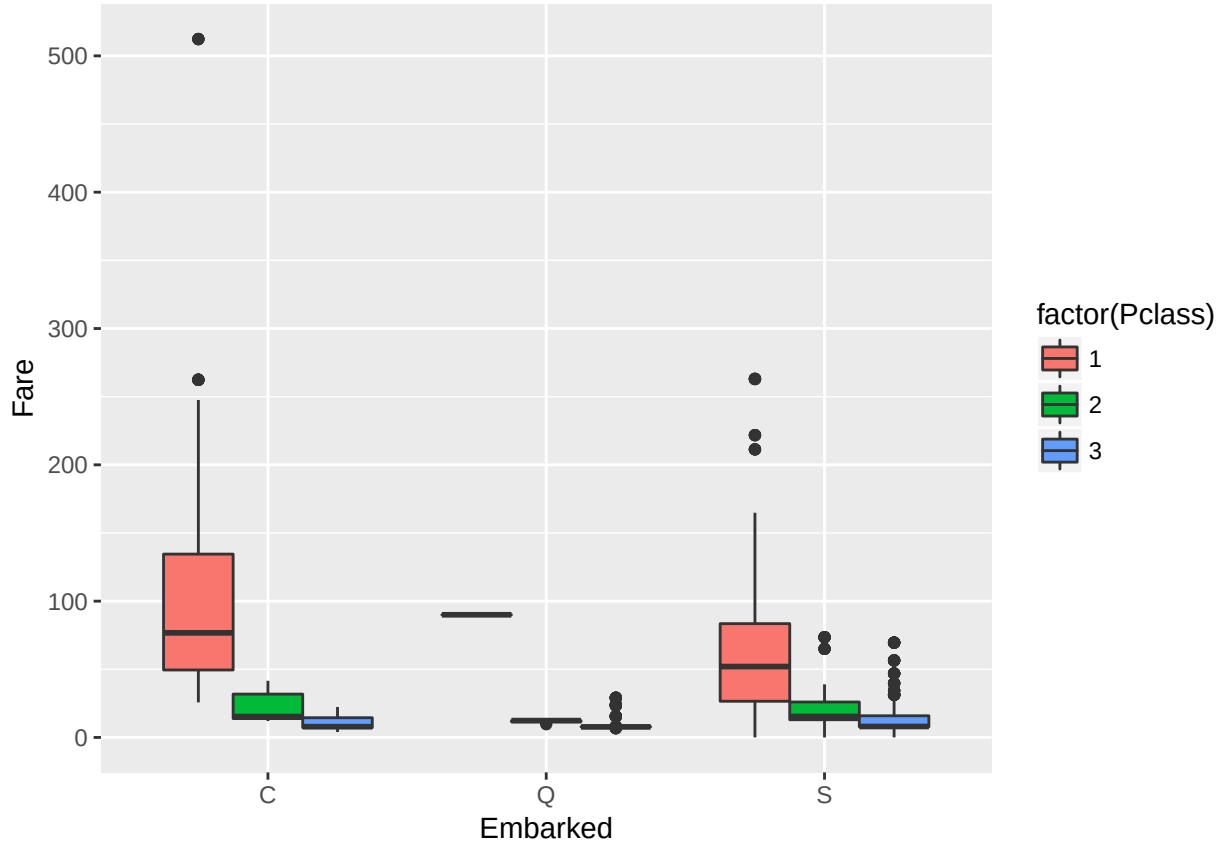
```
##     PassengerId Survived Pclass                                      Name
## 62           62        1      1                       Icard, Miss. Amelie
## 830         830        1      1 Stone, Mrs. George Nelson (Martha Evelyn)
```

```
##        Sex Age SibSp Parch Ticket Fare Cabin Embarked
## 62  female  38     0     0 113572   80   B28
## 830 female  62     0     0 113572   80   B28
```

```
embark_fare <- full %>%
  filter(PassengerId != 62 & PassengerId != 830)

ggplot(embark_fare, aes(x = Embarked, y = Fare, fill = factor(Pclass))) + geom_boxplot()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```




```
full$Embarked[c(62, 830)] <- 'C'
```

Imputing missing value in Fare

```
which(is.na(full$Fare))
```

```
## [1] 1044
```

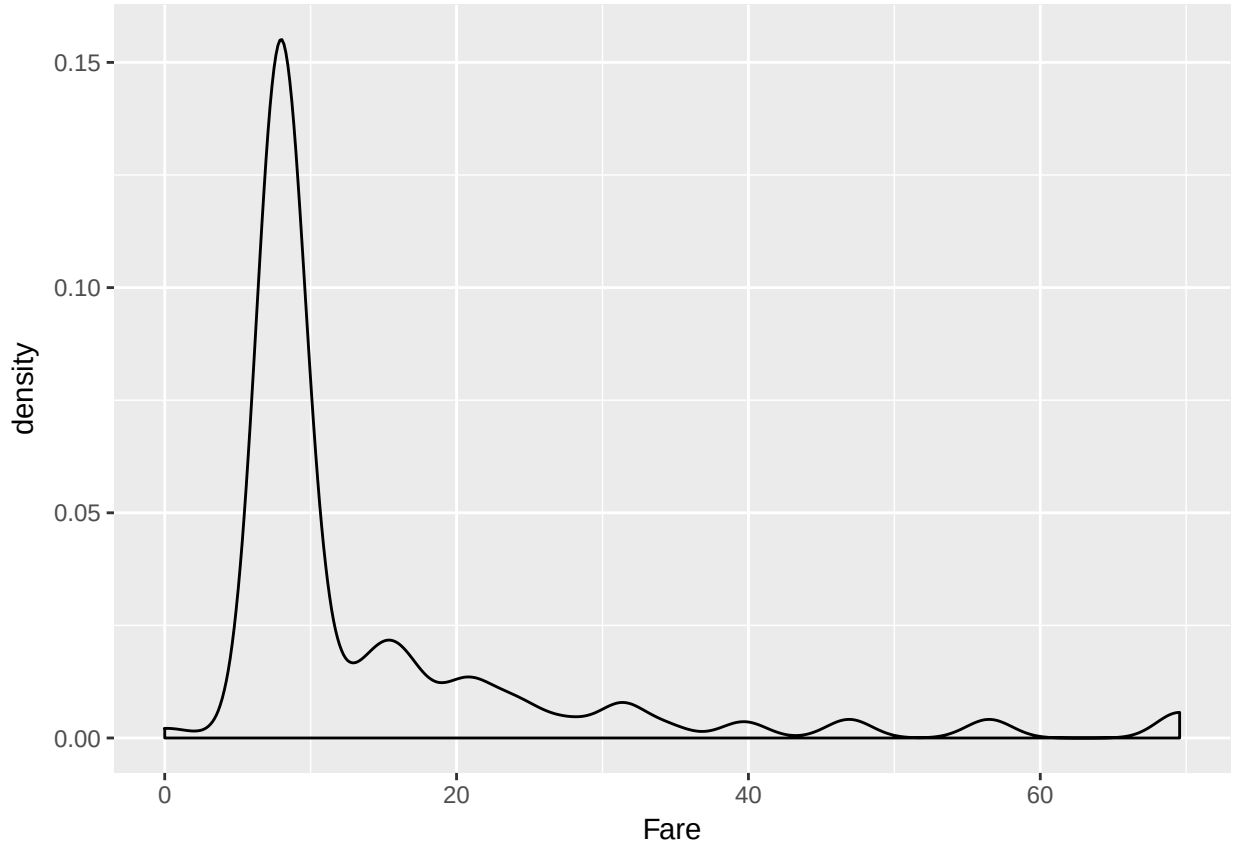
```
full[1044, ]
```

```
##      PassengerId Survived Pclass               Name  Sex  Age SibSp Parch
## 1044        1044        0      3 Storey, Mr. Thomas male 60.5     0     0
##      Ticket Fare Cabin Embarked
## 1044   3701   NA              S
```

```
thirdclass <- full[full$Pclass == 3 & full$Embarked == 'S' ,  ]

ggplot(thirdclass , aes(Fare)) + geom_density()
```

```
## Warning: Removed 1 rows containing non-finite values (stat_density).
```




```
median(thirdclass$Fare , na.rm = TRUE)
```

```
## [1] 8.05
```

```
full$Fare[1044] <- median(thirdclass$Fare , na.rm = TRUE)
```

Imputing missing values in males Ages

```
Male_Ages <- full[ which(is.na(full$Age)) & full$Sex == 'male' , ]
```

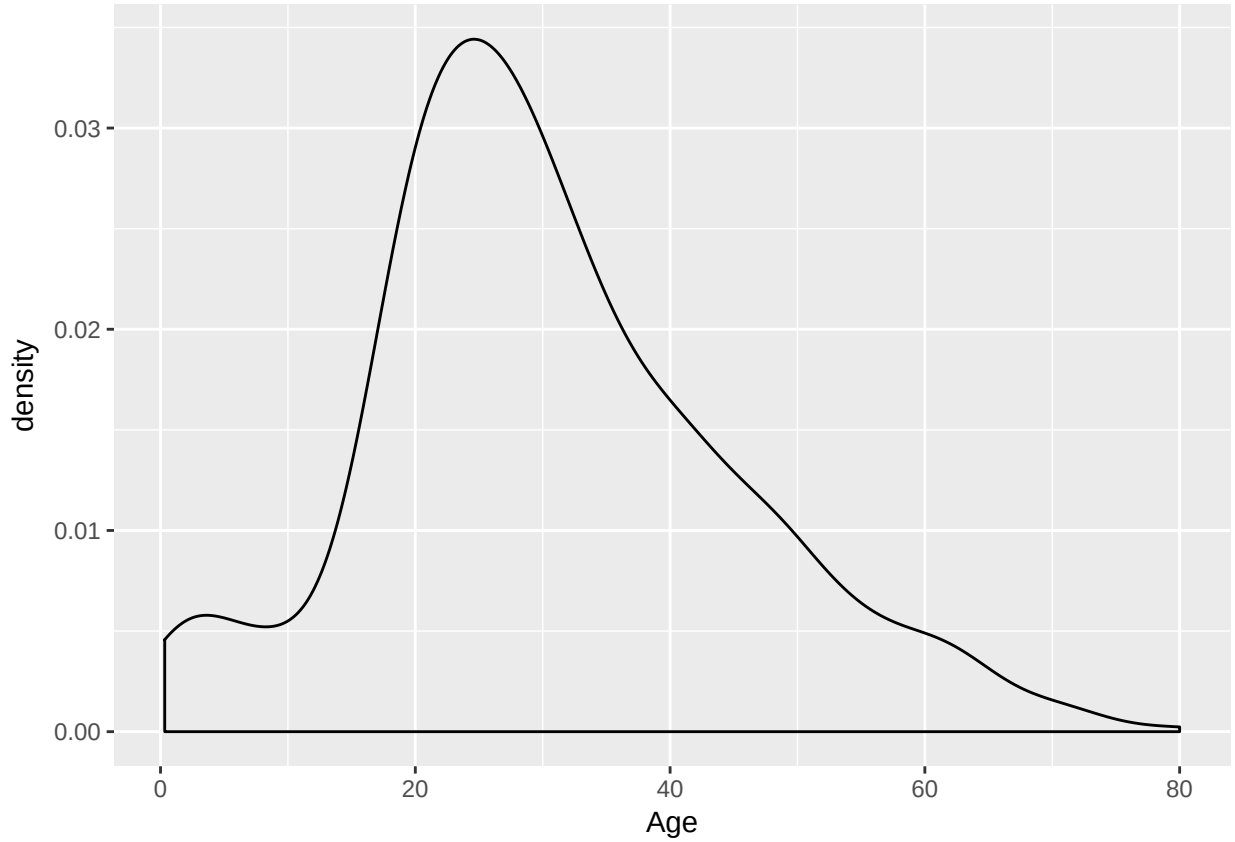
```
## Warning in which(is.na(full$Age)) & full$Sex == "male": longer object
## length is not a multiple of shorter object length
```

```
median(Male_Ages$Age , na.rm =  TRUE)
```

```
## [1] 28
```

```
ggplot(Male_Ages , aes(Age)) + geom_density()
```

```
## Warning: Removed 185 rows containing non-finite values (stat_density).
```




```
full$Age[is.na(full$Age) == TRUE & full$Sex == 'male'] <- median(Male_Ages$Age , na.rm =  TRUE)
```

Imputing missing values in females Ages

```
Female_Ages <- full[ which(is.na(full$Age)) & full$Sex == 'female' , ]
```

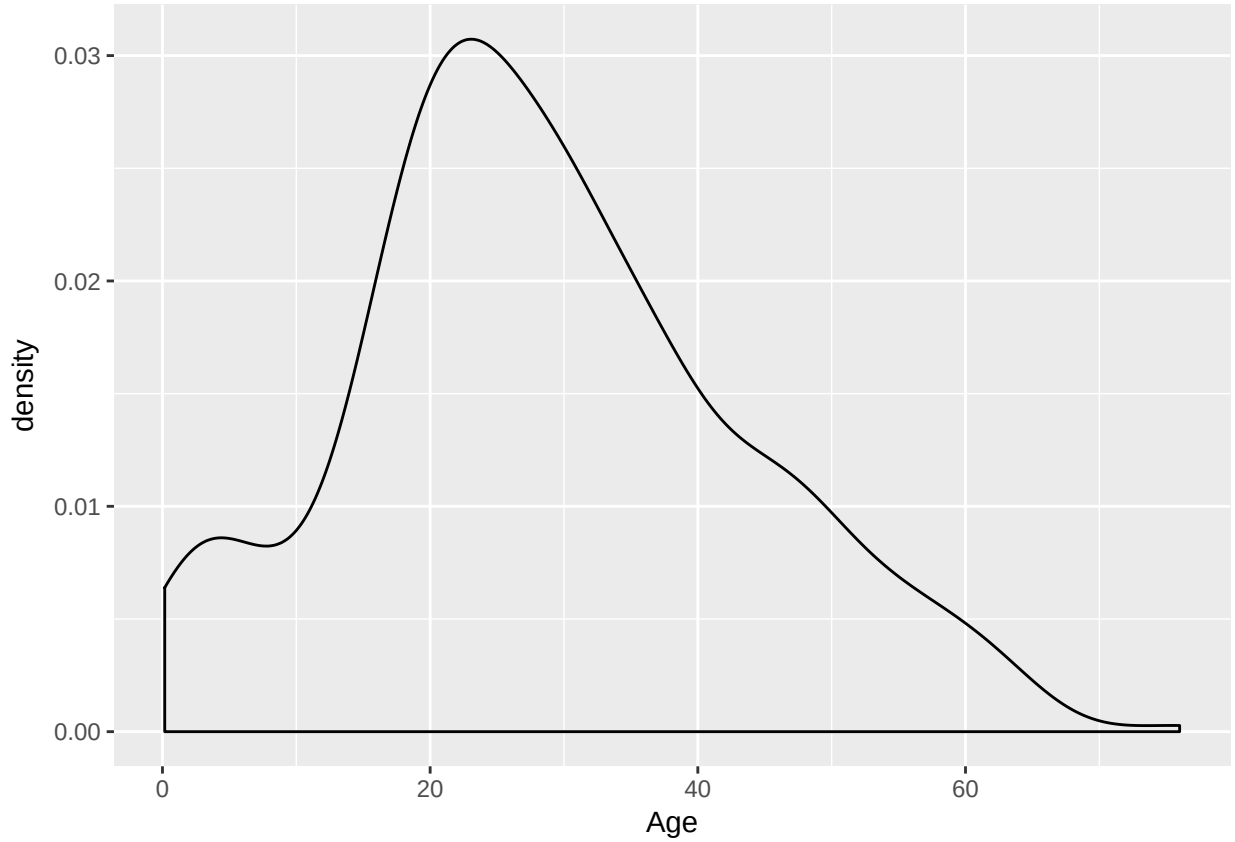
```
## Warning in which(is.na(full$Age)) & full$Sex == "female": longer object
## length is not a multiple of shorter object length
```

```
median(Female_Ages$Age , na.rm =  TRUE)
```

```
## [1] 27
```

```
ggplot(Female_Ages , aes(Age)) + geom_density()
```

```
## Warning: Removed 78 rows containing non-finite values (stat_density).
```




```
full$Age[is.na(full$Age) == TRUE & full$Sex == 'female'] <- median(Female_Ages$Age , na.rm =  TRUE)
```

Feature Engineering

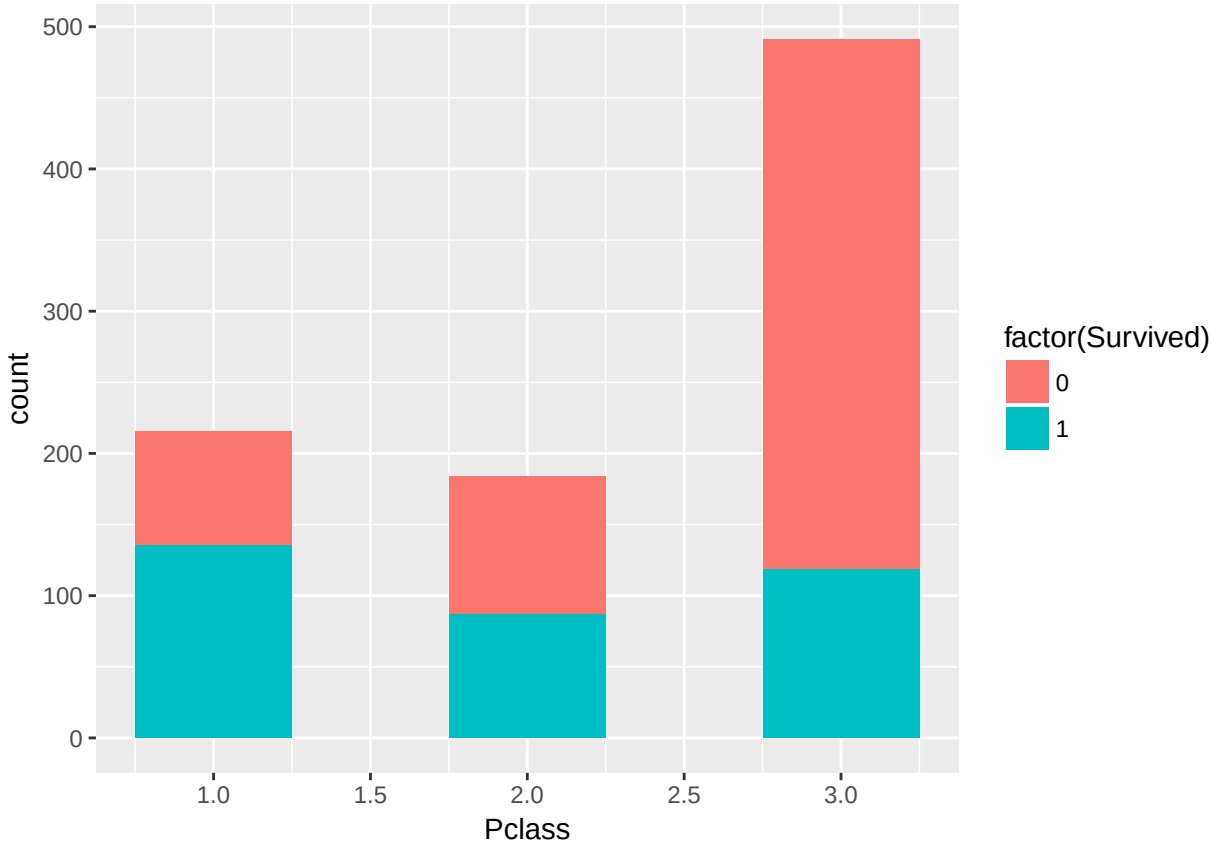
```
# Grab title from passenger names
full$Title <- gsub('(.*, )|(\\..*)', '', full$Name)

# Show title counts by sex
table(full$Sex, full$Title)
```

```
##
##          Capt Col Don Dona  Dr Jonkheer Lady Major Master Miss Mlle Mme
##   female    0   0   0    1   1        0    1     0      0  260    2   1
##   male      1   4   1    0   7        1    0     2     61    0    0   0
##
##           Mr Mrs  Ms Rev Sir the Countess
##   female   0 197   2   0   0            1
##   male   757   0   0   8   1            0
```

Passerger Class and Survival

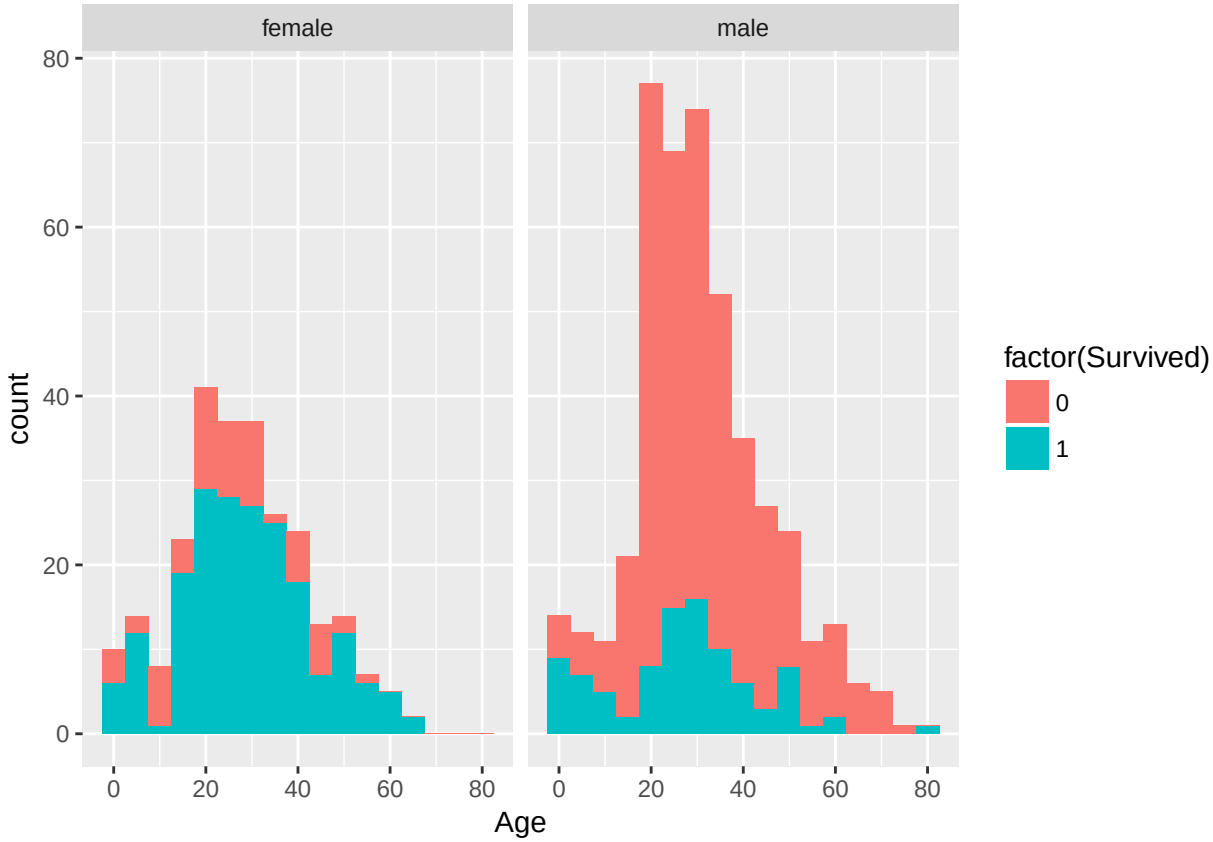
```
ggplot(train , aes(x = Pclass , fill = factor(Survived))) + geom_bar(width = 0.5)
```



This graph shows that passengers in 3rd class had a low survival rate as compared to 1st class.

```
ggplot(train, aes(Age,fill = factor(Survived))) +
    geom_histogram(binwidth = 5) + facet_grid(.~Sex)
```

```
## Warning: Removed 177 rows containing non-finite values (stat_bin).
```



This graph shows the age and gender distribution and their survival rate. We see that females had a far more chance to survive as compared to males

Split back to Train and test data

```
train <- full[1:891,]
test <- full[892:1309,]
```

Converting to factor variables

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


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

Check if there are any missing Values

```
sapply(full, function(full) sum(is.na(full)))
```

```
## PassengerId    Survived      Pclass        Name         Sex         Age
##           0           0           0           0           0           0
##       SibSp       Parch      Ticket        Fare       Cabin    Embarked
##           0           0           0           0           0           0
##       Title
##           0
```

Random Forest Model

```
rf_model <- randomForest(factor(Survived) ~ Pclass + Sex + Age + SibSp + Parch +
                                            Fare + Embarked , data = train )

## Look at variable importance:
round(importance(rf_model), 2)
```

```
##          MeanDecreaseGini
## Pclass              33.53
## Sex                101.12
## Age                 54.66
## SibSp               16.00
## Parch               12.30
## Fare                64.91
## Embarked            12.38
```

Prediction

```
# Predict using the test set
prediction <- predict(rf_model, test)

# Save the solution to a dataframe with two columns: PassengerId and Survived (prediction)
solution <- data.frame(PassengerID = test$PassengerId, Survived = prediction)

# Write the solution to file
write.csv(solution, file = 'rf_mod_Solution.csv', row.names = F)
```