CIS_544_Final_Project_Chen

Chen

December 4, 2016

Introduction

Using data provided by `www.kaggle.com` , our goal is to apply machine-learning techniques to successfully predict which passengers survived the sinking of the Titanic. Features like ticket price, age, sex, and class will be used to make the predictions.

Data Set

We were given 891 passenger samples for our training set and their associated labels of whether or not the passenger survived. For each passenger, we were given his/her passenger class, name, sex, age, number of siblings/spouses aboard, number of parents/children board, ticket number, fare, cabin embarked, and port of embarkation.

Exploratory Data Analysis (EDA)

Prepare the Data

```
##read the train and test csv file and save in Titanic.train; and Tinanic.test.
Titanic.train<-read.csv(file="C:\\Users\\Simon_000\\Downloads\\train.csv")
Titanic.test<-read.csv(file="C:\\Users\\Simon_000\\Downloads\\test.csv")
```

```
str(Titanic.train)  ## Show the structure of the train dataset.
```

```
## '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       : Factor w/ 891 levels "Abbing, Mr. Anthony",..: 109 191 358 277 16 559 520 629 417 5
81 ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ 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     : Factor w/ 681 levels "110152","110413",..: 524 597 670 50 473 276 86 396 345 133
 ...
##  $ Fare       : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Cabin      : Factor w/ 148 levels "","A10","A14",..: 1 83 1 57 1 1 131 1 1 1 ...
##  $ Embarked   : Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```
dim(Titanic.train)  ## Show the row and column of the train dataset, that means there are 891 row and
 12 column, that are 891 observations and 12 variables.
```

```
## [1] 891  12
```

```
str(Titanic.test)  ## Show the structure of the test dataset.
```

```
## '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       : Factor w/ 418 levels "Abbott, Master. Eugene Joseph",..: 210 409 273 414 182 370 85
 58 5 104 ...
##  $ Sex        : Factor w/ 2 levels "female","male": 2 1 2 2 1 2 1 2 1 2 ...
##  $ 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     : Factor w/ 363 levels "110469","110489",..: 153 222 74 148 139 262 159 85 101 270
 ...
##  $ Fare       : num  7.83 7 9.69 8.66 12.29 ...
##  $ Cabin      : Factor w/ 77 levels "","A11","A18",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Embarked   : Factor w/ 3 levels "C","Q","S": 2 3 2 3 3 3 2 3 1 3 ...
```

```
dim(Titanic.test)  ## Show the row and column of the test dataset, that means there are 418 row and 11
 column, that are 418 observations and 11 variables.
```

```
## [1] 418  11
```

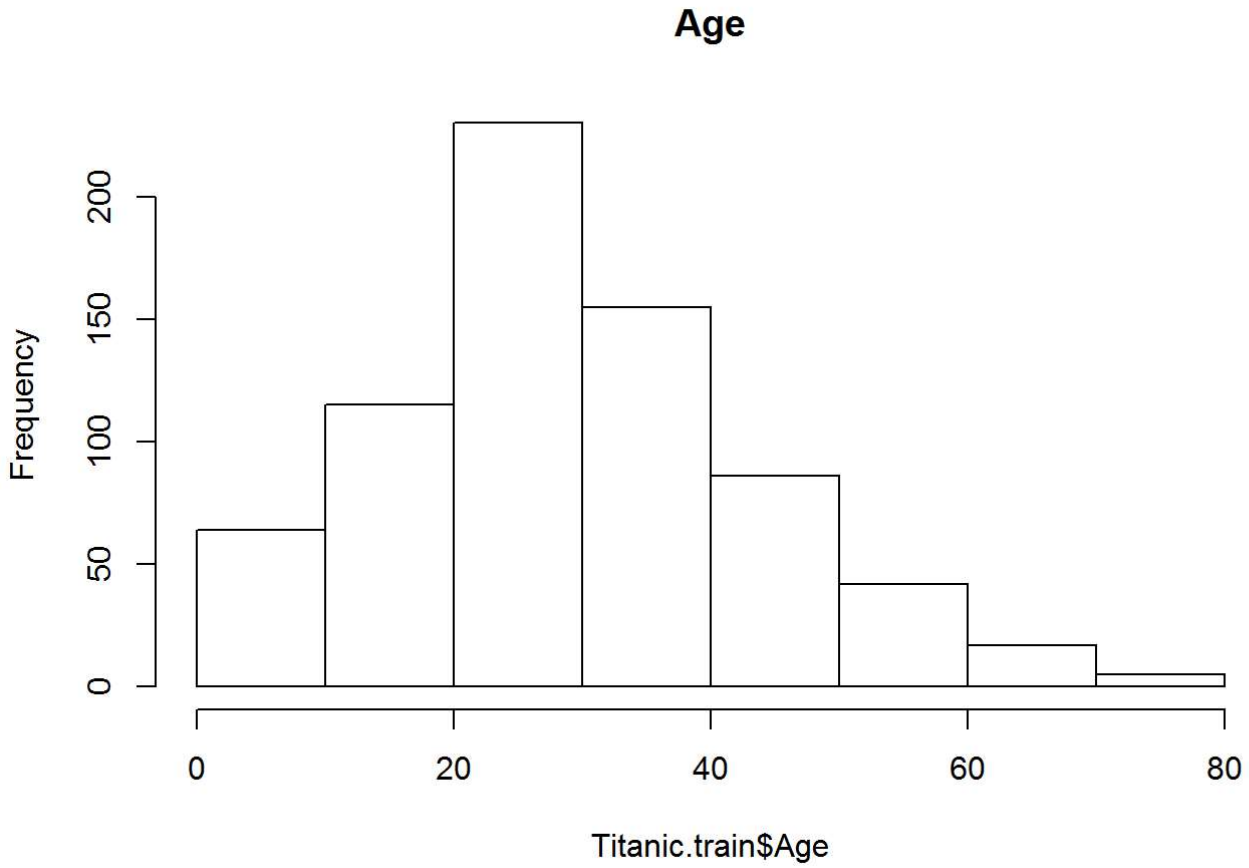
```
summary(Titanic.train)  ## Summary of the train dataset.
```

```
##   PassengerId       Survived          Pclass
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000
##  Median :446.0   Median :0.0000   Median :3.000
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##                                     Name         Sex           Age
##  Abbing, Mr. Anthony                   :  1   female:314   Min.   : 0.42
##  Abbott, Mr. Rossmore Edward           :  1   male  :577   1st Qu.:20.12
##  Abbott, Mrs. Stanton (Rosa Hunt)      :  1                Median :28.00
##  Abelson, Mr. Samuel                   :  1                Mean   :29.70
##  Abelson, Mrs. Samuel (Hannah Wizosky):  1                3rd Qu.:38.00
##  Adahl, Mr. Mauritz Nils Martin        :  1                Max.   :80.00
##  (Other)                               :885                NA's   :177
##      SibSp           Parch             Ticket         Fare
##  Min.   :0.000   Min.   :0.0000   1601     :  7   Min.   :  0.00
##  1st Qu.:0.000   1st Qu.:0.0000   347082   :  7   1st Qu.:  7.91
##  Median :0.000   Median :0.0000   CA. 2343:  7   Median : 14.45
##  Mean   :0.523   Mean   :0.3816   3101295  :  6   Mean   : 32.20
##  3rd Qu.:1.000   3rd Qu.:0.0000   347088   :  6   3rd Qu.: 31.00
##  Max.   :8.000   Max.   :6.0000   CA 2144  :  6   Max.   :512.33
##                                   (Other)  :852
##          Cabin     Embarked
##              :687    :  2
##  B96 B98     :  4   C:168
##  C23 C25 C27:  4   Q: 77
##  G6          :  4   S:644
##  C22 C26     :  3
##  D           :  3
##  (Other)     :186
```

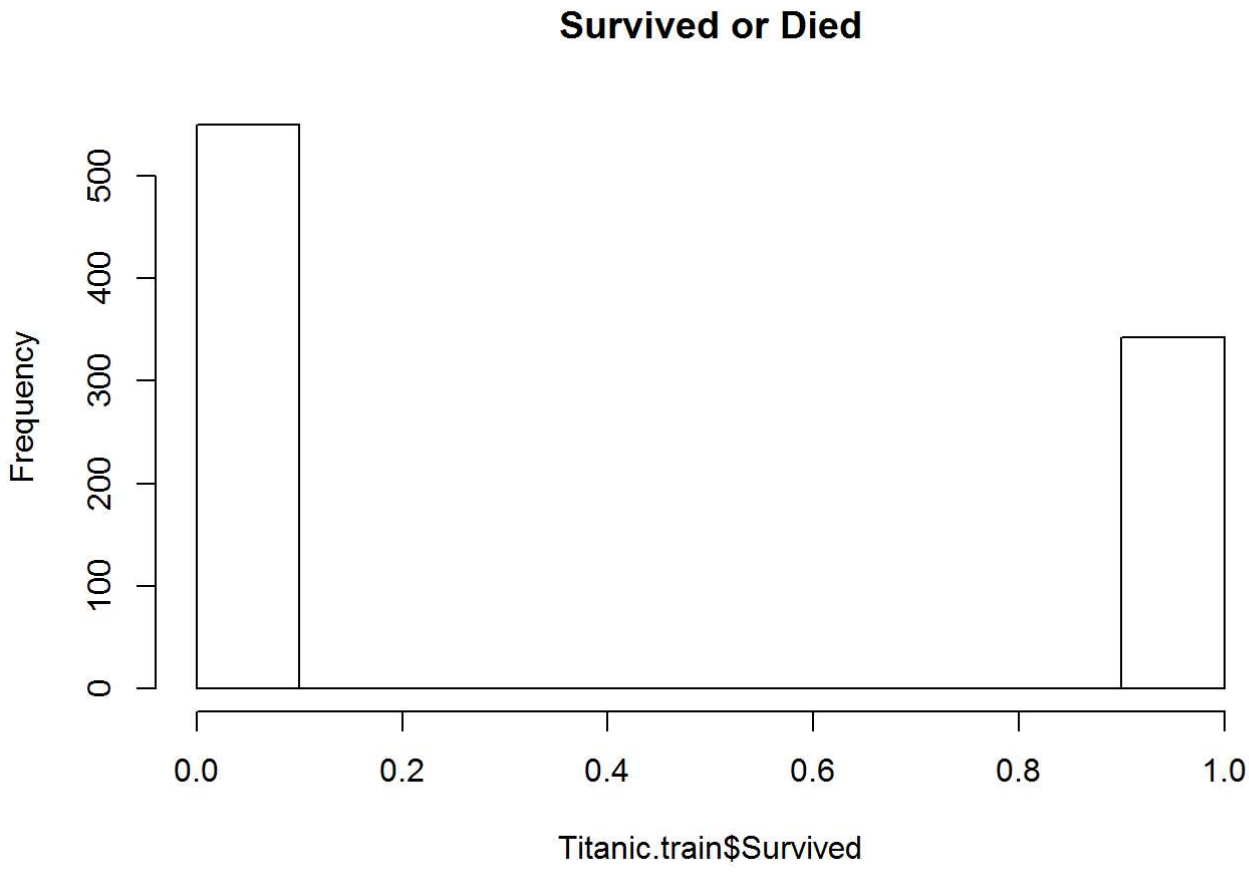
```
summary(Titanic.test)   ## Summary of the test dataset.
```

```
##   PassengerId         Pclass
##  Min.   : 892.0   Min.   :1.000
##  1st Qu.: 996.2   1st Qu.:1.000
##  Median :1100.5   Median :3.000
##  Mean   :1100.5   Mean   :2.266
##  3rd Qu.:1204.8   3rd Qu.:3.000
##  Max.   :1309.0   Max.   :3.000
##
##                                          Name         Sex
##  Abbott, Master. Eugene Joseph            :  1   female:152
##  Abelseth, Miss. Karen Marie              :  1   male  :266
##  Abelseth, Mr. Olaus Jorgensen            :  1
##  Abrahamsson, Mr. Abraham August Johannes :  1
##  Abrahim, Mrs. Joseph (Sophie Halaut Easu):  1
##  Aks, Master. Philip Frank                :  1
##  (Other)                                  :412
##       Age             SibSp            Parch             Ticket
##  Min.   : 0.17   Min.   :0.0000   Min.   :0.0000   PC 17608:  5
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000   113503  :  4
##  Median :27.00   Median :0.0000   Median :0.0000   CA. 2343:  4
##  Mean   :30.27   Mean   :0.4474   Mean   :0.3923   16966   :  3
##  3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000   220845  :  3
##  Max.   :76.00   Max.   :8.0000   Max.   :9.0000   347077  :  3
##  NA's   :86                                        (Other) :396
##       Fare                    Cabin     Embarked
##  Min.   :  0.000                 :327   C:102
##  1st Qu.:  7.896   B57 B59 B63 B66:  3   Q: 46
##  Median : 14.454   A34           :  2   S:270
##  Mean   : 35.627   B45           :  2
##  3rd Qu.: 31.500   C101          :  2
##  Max.   :512.329   C116          :  2
##  NA's   :1         (Other)       : 80
```

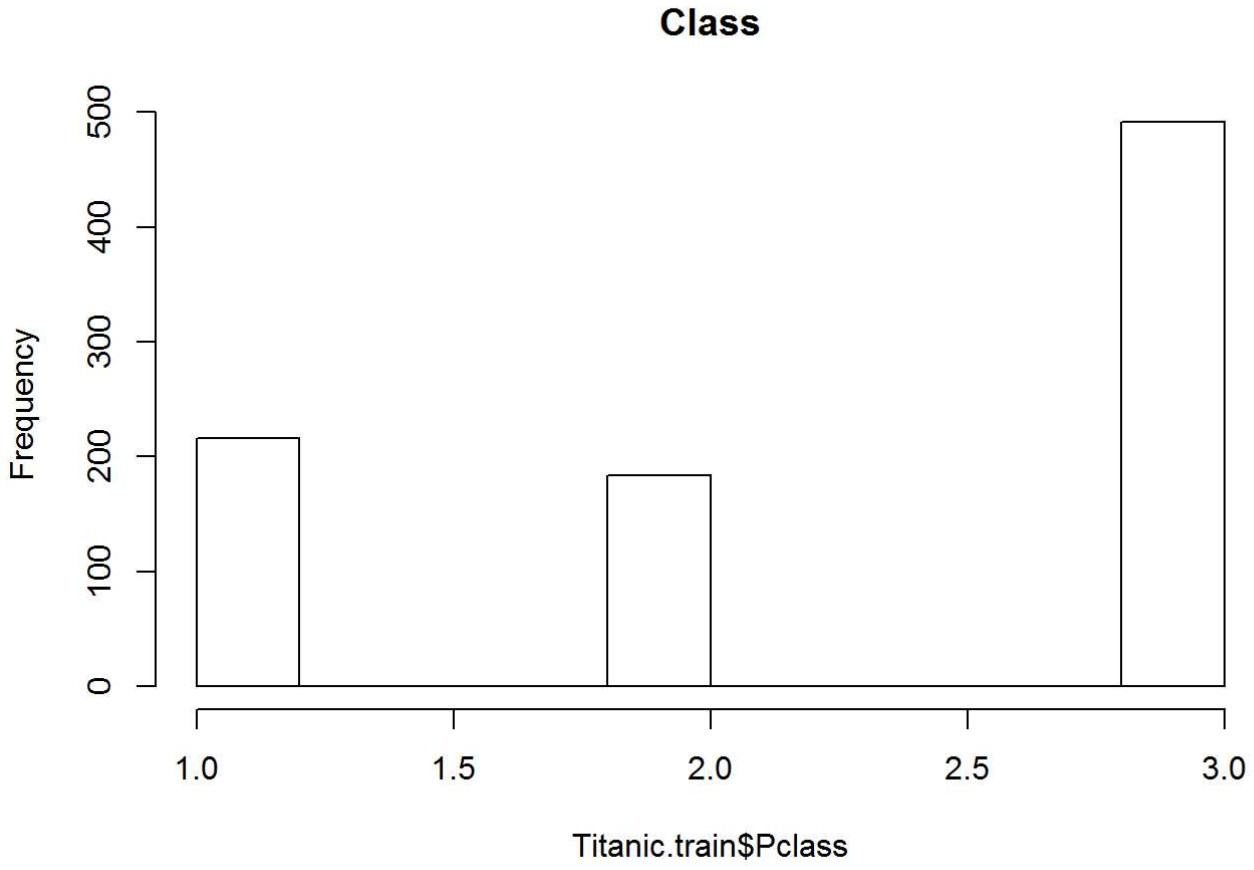
```
hist(Titanic.train$Age, main="Age")
```




```
hist(Titanic.train$Survived, main="Survived or Died")
```



```
hist(Titanic.train$Pclass, main="Class")
```




```
table(Titanic.train$Survived) ## Base on the observation if the data, we conclusion that there are 392 person died and 549 person survived.
```

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

```
require(Amelia)  ## use the Amelia library which about the missing data.
```

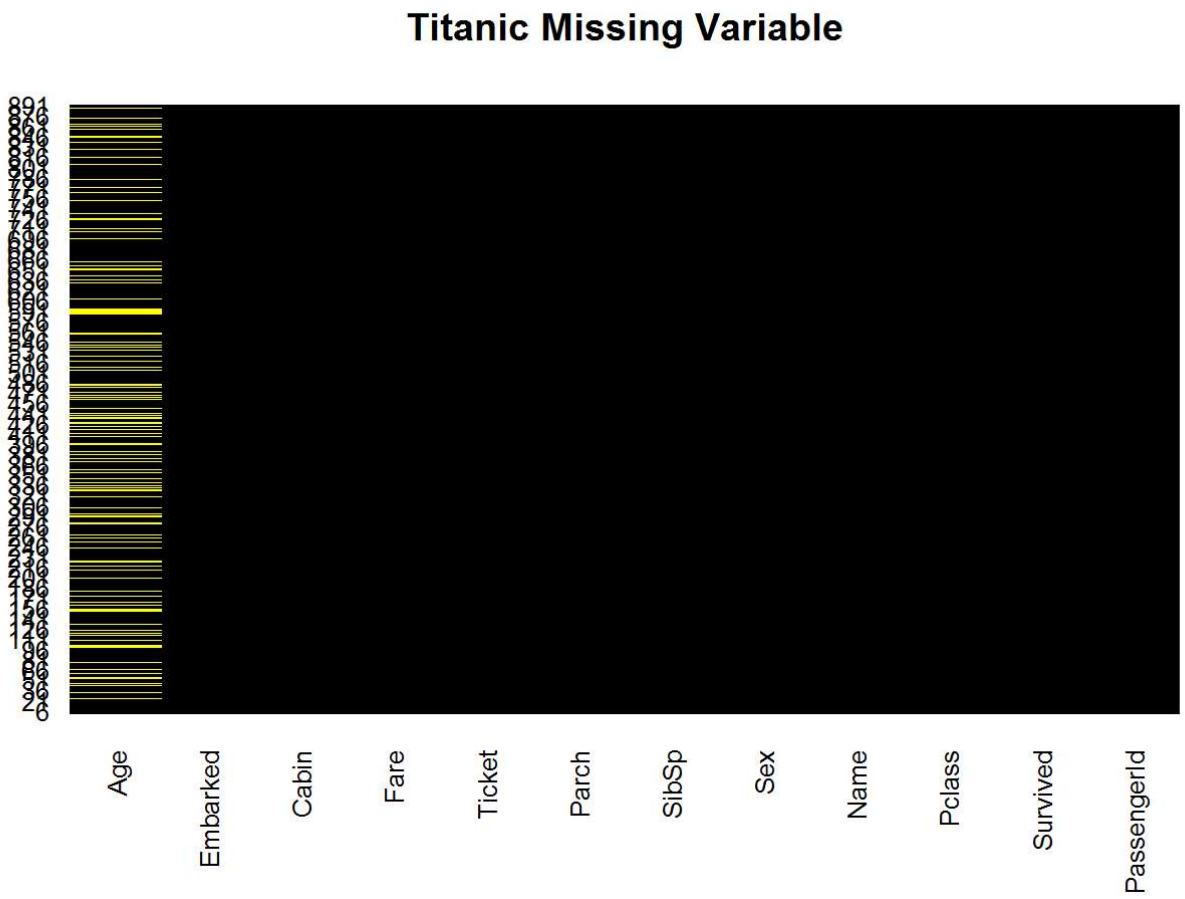
```
## Loading required package: Amelia
```

```
## Warning: package 'Amelia' was built under R version 3.3.2
```

```
## Loading required package: Rcpp
```

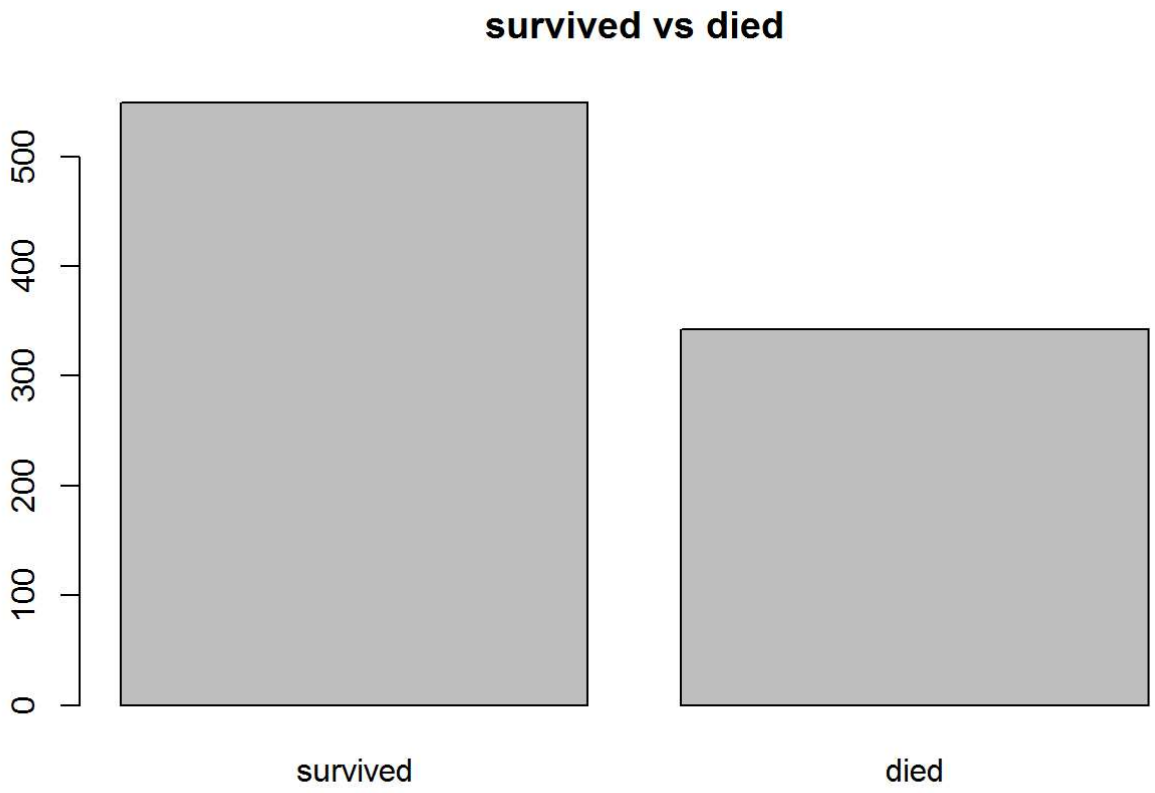
```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.4, built: 2015-12-05)
## ## Copyright (C) 2005-2016 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
missmap(Titanic.train, main="Titanic Missing Variable",col=c("yellow", "black"), legend=FALSE)   ##we can see that in the summary of the train dataset there has 177 missing data for the Age column, so we can use the function missmap().
```



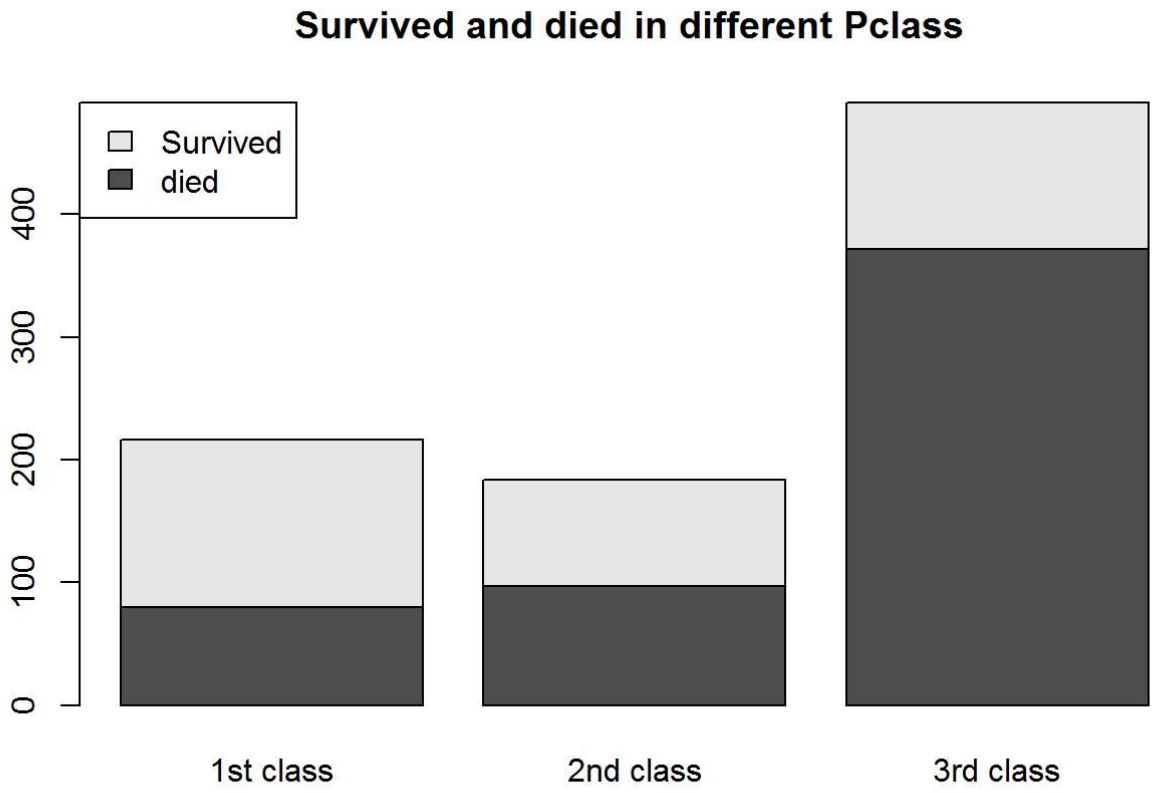
The compare between the survived and died

```
barplot(table(Titanic.train$Survived),names.arg=c("survived","died"),main="survived vs died")  ## Create a barplot for the Survived and died in Titanic.
```



compare between the survived and died in different class

```
survive.rate.class=table(Titanic.train$Survived,Titanic.train$Pclass)
barplot(survive.rate.class,names.arg=c("1st class","2nd class","3rd class"),main="Survived and died in
 different Pclass ",legend.text=c("died","Survived"),args.legend=list(x="topleft"))
```



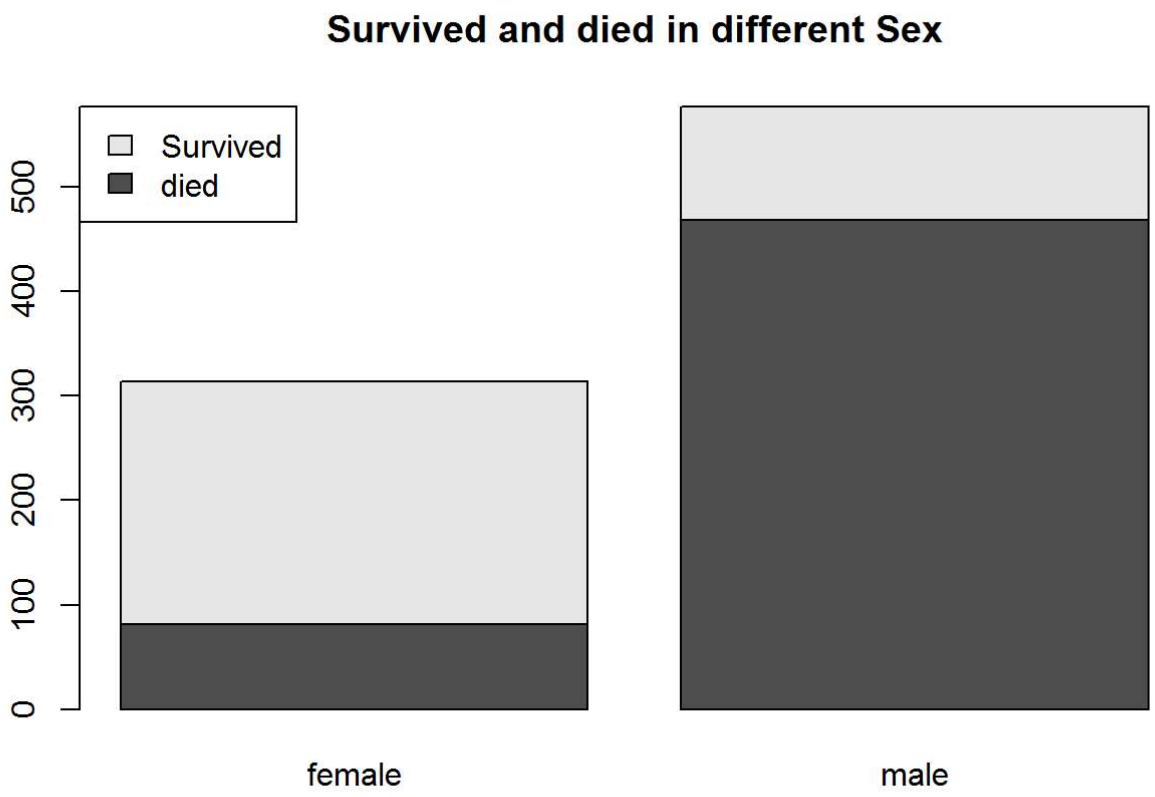

```
round((survive.rate.class[2,]/colSums(survive.rate.class))*100,2)
```

```
##     1     2     3
## 62.96 47.28 24.24
```

The Survived rate probability for the 1st 2nd class and 3rd class are 62.96%, 47.28%, and 24.24%. So upper class have more probability to alive.

compare between the survived and died in different Sex

```
survive.rate.sex=table(Titanic.train$Survived,Titanic.train$Sex)
barplot(survive.rate.sex,names.arg=c("female","male"),
        main="Survived and died in different Sex",
        legend.text=c("died","Survived"),
        args.legend=list(x="topleft"))
```



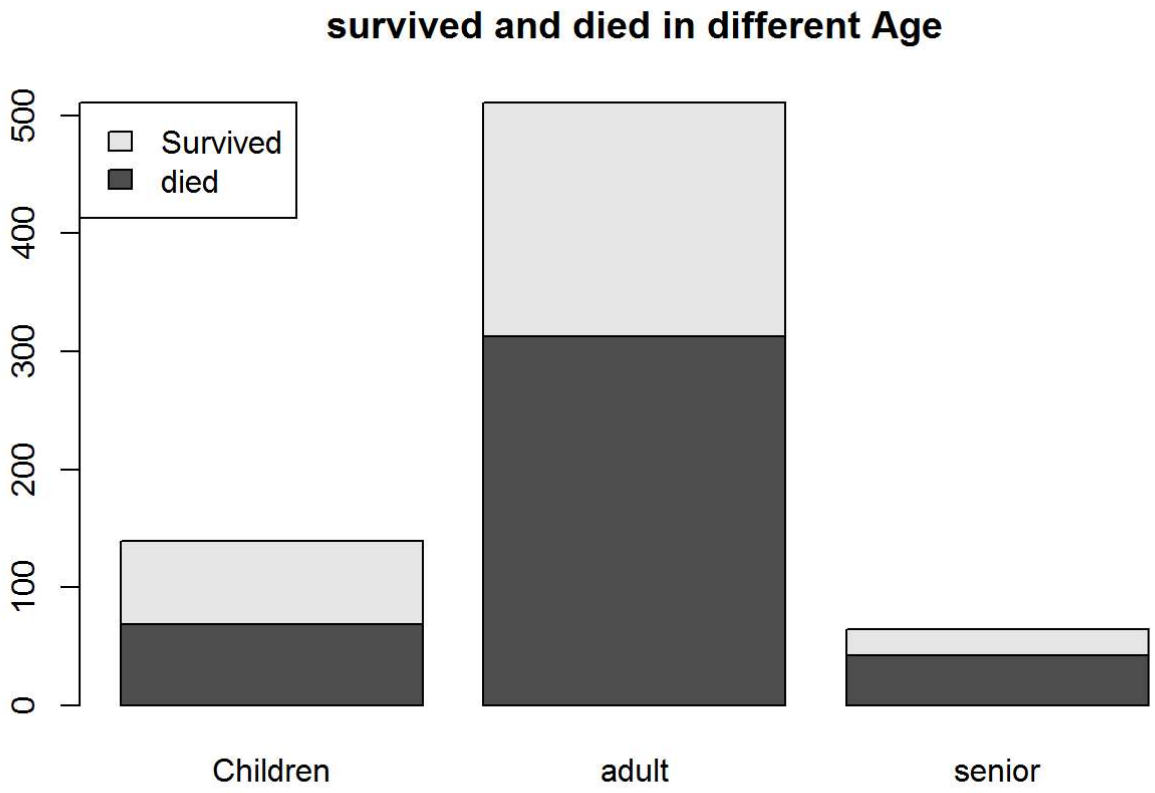

```
round((survive.rate.sex[2,]/colSums(survive.rate.sex))*100,2)
```

```
## female   male
##  74.20  18.89
```

18.89% male have probability of 18.89% to alive, and female have probability of 74.20% to alive.

compare between the survived and died in different Age

```
age.breaker=c(0,18,50,100)
age.cut= cut(Titanic.train$Age,breaks=age.breaker,labels=c("Children","adult","senior"))
Titanic.train$age.cut=age.cut
survive.rate.age=table(Titanic.train$Survived,Titanic.train$age.cut)
barplot(survive.rate.age,
        main="survived and died in different Age",
        legend.text=c("died","Survived"),
        args.legend=list(x="topleft"))
```




```
round((survive.rate.age[2,]/colSums(survive.rate.age))*100,2)
```

```
## Children    adult   senior
##    50.36    38.75    34.38
```

we use the age 0-18 for children,18-50 for adult,50-100 for senior, we conclude that children have the probability of 50.36% to alive, senior is more easy to die then adult.

Principal Components Regression

```
library(pls)
```

```
## Warning: package 'pls' was built under R version 3.3.2
```

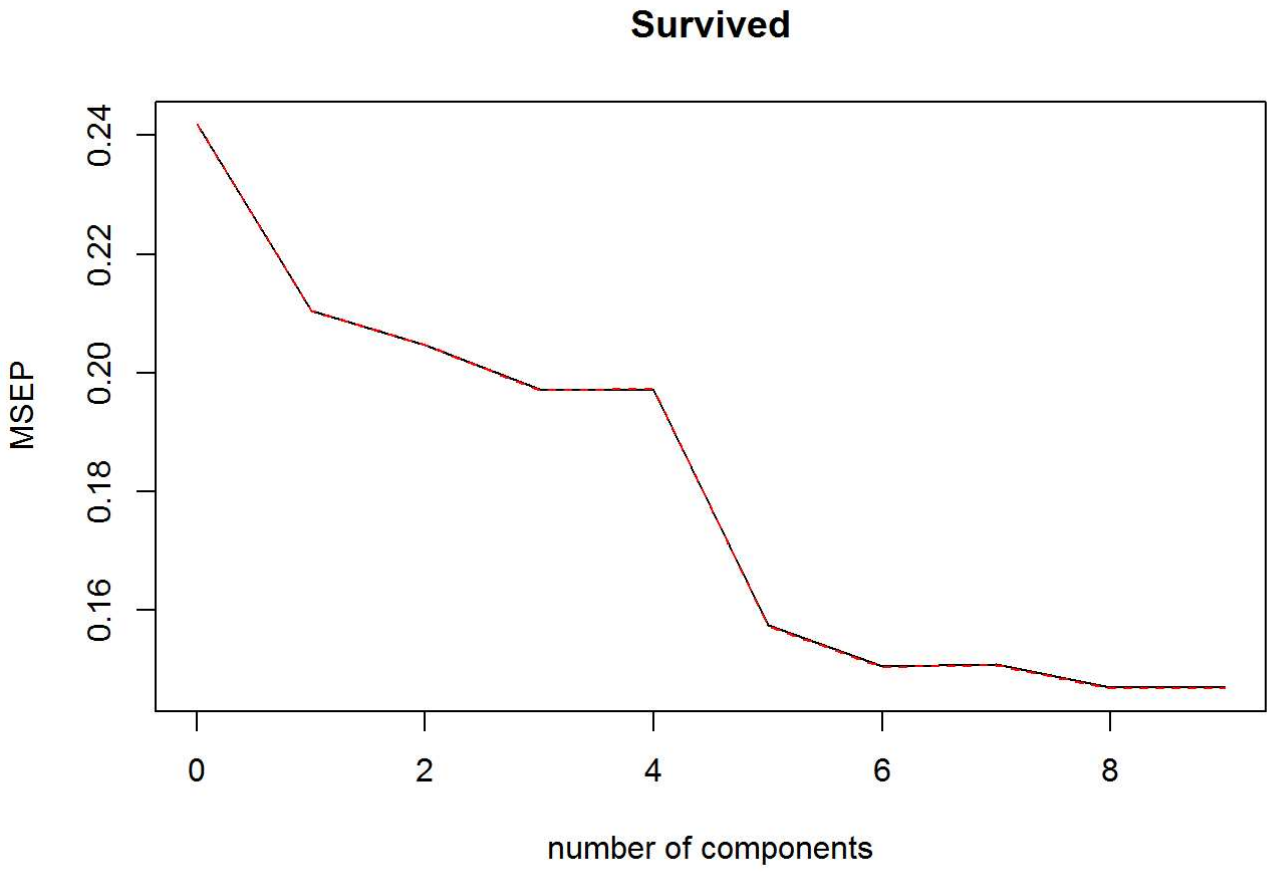
```
##
## Attaching package: 'pls'
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

```
set.seed(200)
pcr.fit = pcr(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked, data = Titanic.train, s
cale = T,      validation = "CV")
summary(pcr.fit)
```

```
## Data:    X dimension: 714 9
##  Y dimension: 714 1
## Fit method: svdpc
## Number of components considered: 9
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           0.4918   0.4587   0.4525   0.4442   0.4439   0.3970   0.3880
## adjCV        0.4918   0.4587   0.4525   0.4440   0.4441   0.3968   0.3878
##        7 comps  8 comps  9 comps
## CV      0.3885   0.3834   0.3834
## adjCV   0.3883   0.3832   0.3832
##
## TRAINING: % variance explained
##           1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X           26.18    45.67    61.95    72.99    82.60    89.89     96.0
## Survived    13.10    15.78    19.17    19.55    35.39    38.70     38.7
##           8 comps  9 comps
## X           99.92   100.00
## Survived    40.32    40.34
```

```
validationplot(pcr.fit, val.type = "MSEP")
```



Tree

```
library(rpart)  ## load the rpart library which is more powerful than tree library
fit <- rpart(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare, data=Titanic.train, method="class")
plot(fit)
text(fit)   # Create the tree plot base on Survived which related to the different attribute.
```