ml prrojfinal

pradeep

2022-12-18

```
library(ggplot2)
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.2.2
```

```
library(dplyr)
```

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

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

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

```
library(corrgram)
```

```
## Warning: package 'corrgram' was built under R version 4.2.2
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.2.2

## corrplot 0.92 loaded
```

```
library(caTools)
library(Amelia)
```

```
## Warning: package 'Amelia' was built under R version 4.2.2

## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.1, built: 2022-11-18)
## ## Copyright (C) 2005-2022 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'
```

```
## The following object is masked from 'package:corrgram':
##
##     panel.fill
```

```
data.train <- read.csv("C:/Users/prade/Downloads/titanic.csv")
head(data.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 Parch
## 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              S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250              S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500              S
## 6           330877  8.4583              Q
```

```
str(data.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)"
##  $ 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" ...
```

```
summary(data.train)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  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
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin              Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

```
sum(is.na(data.train))
```

```
## [1] 177
```

```
which(is.na(data.train), arr.ind = T)
```

```
##        row col
##   [1,]   6   6
##   [2,]  18   6
##   [3,]  20   6
##   [4,]  27   6
##   [5,]  29   6
##   [6,]  30   6
##   [7,]  32   6
##   [8,]  33   6
##   [9,]  37   6
##  [10,]  43   6
```

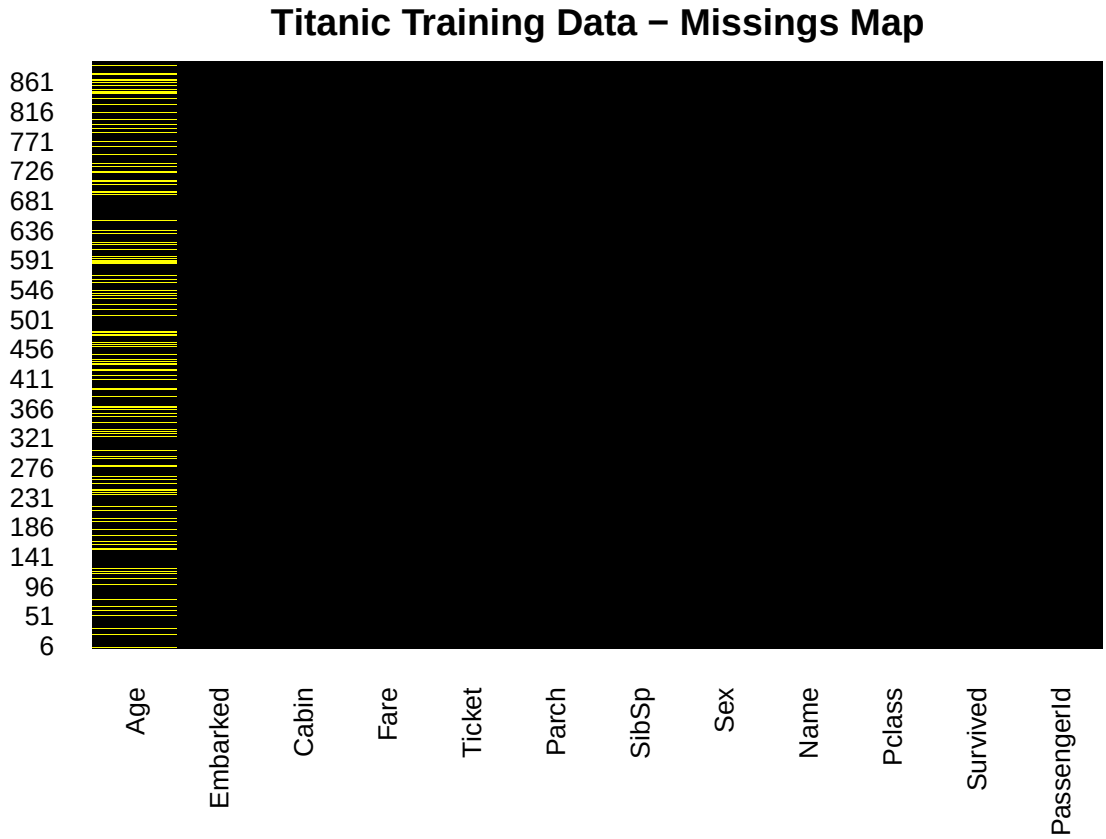
```
##  [11,]  46   6
##  [12,]  47   6
##  [13,]  48   6
##  [14,]  49   6
##  [15,]  56   6
##  [16,]  65   6
##  [17,]  66   6
##  [18,]  77   6
##  [19,]  78   6
##  [20,]  83   6
##  [21,]  88   6
##  [22,]  96   6
##  [23,] 102   6
##  [24,] 108   6
##  [25,] 110   6
##  [26,] 122   6
##  [27,] 127   6
##  [28,] 129   6
##  [29,] 141   6
##  [30,] 155   6
##  [31,] 159   6
##  [32,] 160   6
##  [33,] 167   6
##  [34,] 169   6
##  [35,] 177   6
##  [36,] 181   6
##  [37,] 182   6
##  [38,] 186   6
##  [39,] 187   6
##  [40,] 197   6
##  [41,] 199   6
##  [42,] 202   6
##  [43,] 215   6
##  [44,] 224   6
##  [45,] 230   6
##  [46,] 236   6
##  [47,] 241   6
##  [48,] 242   6
##  [49,] 251   6
##  [50,] 257   6
##  [51,] 261   6
##  [52,] 265   6
##  [53,] 271   6
##  [54,] 275   6
##  [55,] 278   6
##  [56,] 285   6
##  [57,] 296   6
##  [58,] 299   6
##  [59,] 301   6
##  [60,] 302   6
##  [61,] 304   6
##  [62,] 305   6
##  [63,] 307   6
##  [64,] 325   6
```

```
##  [65,] 331   6
##  [66,] 335   6
##  [67,] 336   6
##  [68,] 348   6
##  [69,] 352   6
##  [70,] 355   6
##  [71,] 359   6
##  [72,] 360   6
##  [73,] 365   6
##  [74,] 368   6
##  [75,] 369   6
##  [76,] 376   6
##  [77,] 385   6
##  [78,] 389   6
##  [79,] 410   6
##  [80,] 411   6
##  [81,] 412   6
##  [82,] 414   6
##  [83,] 416   6
##  [84,] 421   6
##  [85,] 426   6
##  [86,] 429   6
##  [87,] 432   6
##  [88,] 445   6
##  [89,] 452   6
##  [90,] 455   6
##  [91,] 458   6
##  [92,] 460   6
##  [93,] 465   6
##  [94,] 467   6
##  [95,] 469   6
##  [96,] 471   6
##  [97,] 476   6
##  [98,] 482   6
##  [99,] 486   6
## [100,] 491   6
## [101,] 496   6
## [102,] 498   6
## [103,] 503   6
## [104,] 508   6
## [105,] 512   6
## [106,] 518   6
## [107,] 523   6
## [108,] 525   6
## [109,] 528   6
## [110,] 532   6
## [111,] 534   6
## [112,] 539   6
## [113,] 548   6
## [114,] 553   6
## [115,] 558   6
## [116,] 561   6
## [117,] 564   6
## [118,] 565   6
```

```
## [119,] 569   6
## [120,] 574   6
## [121,] 579   6
## [122,] 585   6
## [123,] 590   6
## [124,] 594   6
## [125,] 597   6
## [126,] 599   6
## [127,] 602   6
## [128,] 603   6
## [129,] 612   6
## [130,] 613   6
## [131,] 614   6
## [132,] 630   6
## [133,] 634   6
## [134,] 640   6
## [135,] 644   6
## [136,] 649   6
## [137,] 651   6
## [138,] 654   6
## [139,] 657   6
## [140,] 668   6
## [141,] 670   6
## [142,] 675   6
## [143,] 681   6
## [144,] 693   6
## [145,] 698   6
## [146,] 710   6
## [147,] 712   6
## [148,] 719   6
## [149,] 728   6
## [150,] 733   6
## [151,] 739   6
## [152,] 740   6
## [153,] 741   6
## [154,] 761   6
## [155,] 767   6
## [156,] 769   6
## [157,] 774   6
## [158,] 777   6
## [159,] 779   6
## [160,] 784   6
## [161,] 791   6
## [162,] 793   6
## [163,] 794   6
## [164,] 816   6
## [165,] 826   6
## [166,] 827   6
## [167,] 829   6
## [168,] 833   6
## [169,] 838   6
## [170,] 840   6
## [171,] 847   6
## [172,] 850   6
```

```
## [173,] 860   6
## [174,] 864   6
## [175,] 869   6
## [176,] 879   6
## [177,] 889   6
```

```
#Data Cleaning
data.train[data.train$Survived == 1,]$Survived <- 'Y'
data.train[data.train$Survived == 0,]$Survived <- 'N'

data.train$Survived <- as.factor(data.train$Survived)
data.train$Pclass <- as.factor(data.train$Pclass)
data.train$Sex <- as.factor(data.train$Sex)
data.train$Parch <- factor(data.train$Parch)
data.train$SibSp <- factor(data.train$SibSp)
data.train$Embarked <- factor(data.train$Embarked)

missmap(data.train, main="Titanic Training Data - Missings Map", col=c("yellow", "black"), legend=FALSE)
```

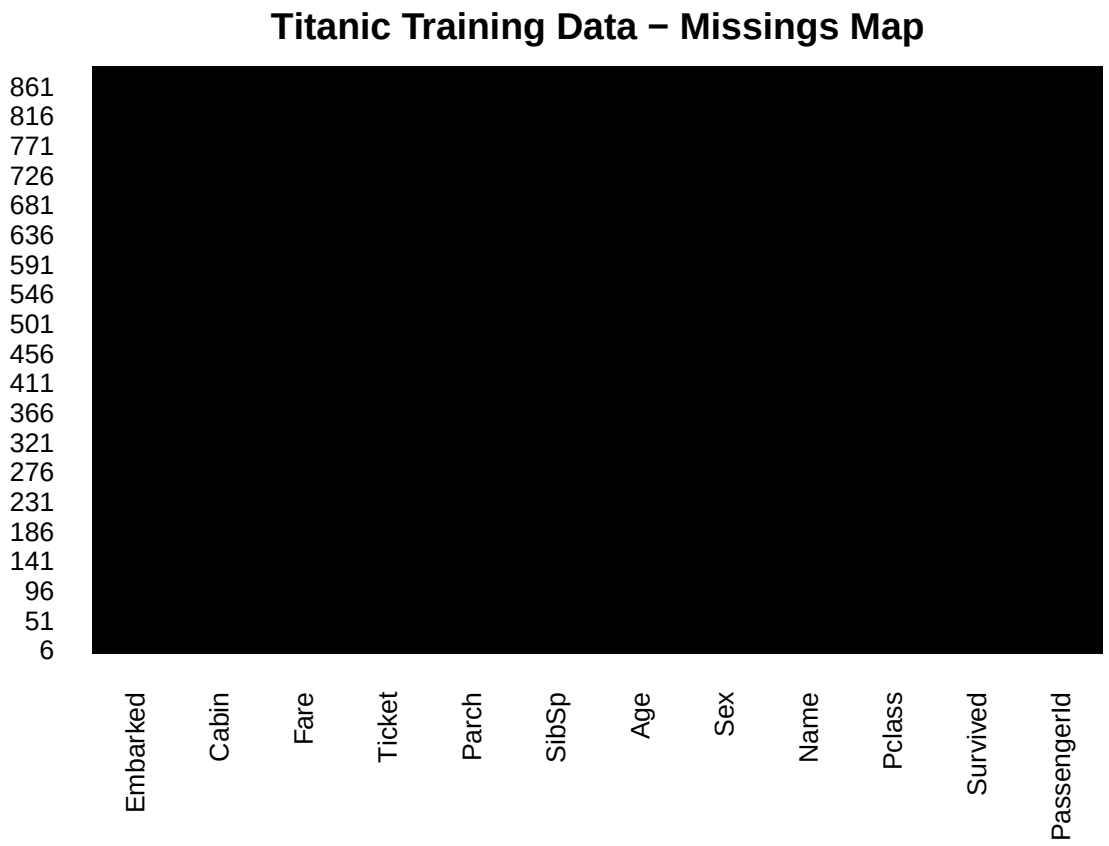



```
summary(data.train$Age)
```

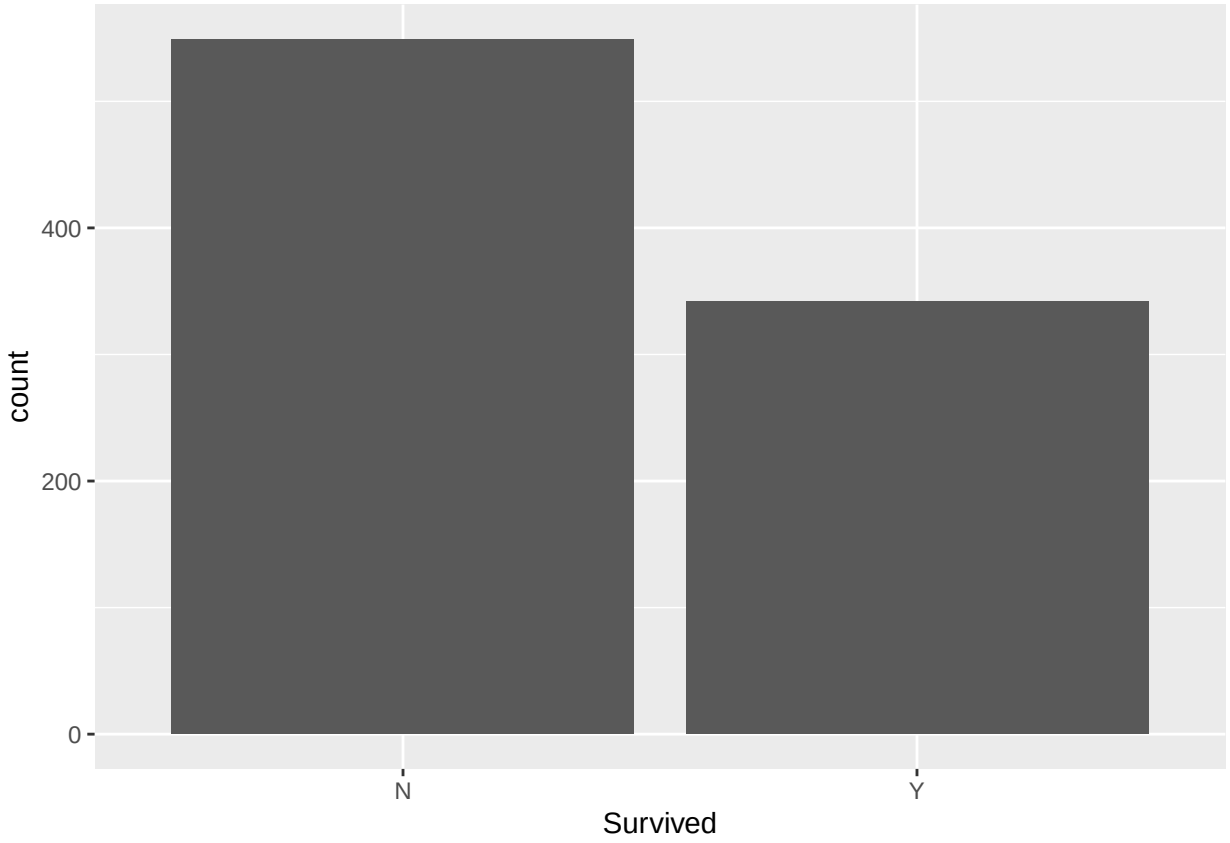
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.42   20.12   28.00   29.70   38.00   80.00     177
```

```r
#mean = 29.70, because its 177 missing values ~20% of our data, I rather input the mean instead of dele

data.train$Age[is.na(data.train$Age)] <- mean(data.train$Age, na.rm = TRUE)
```

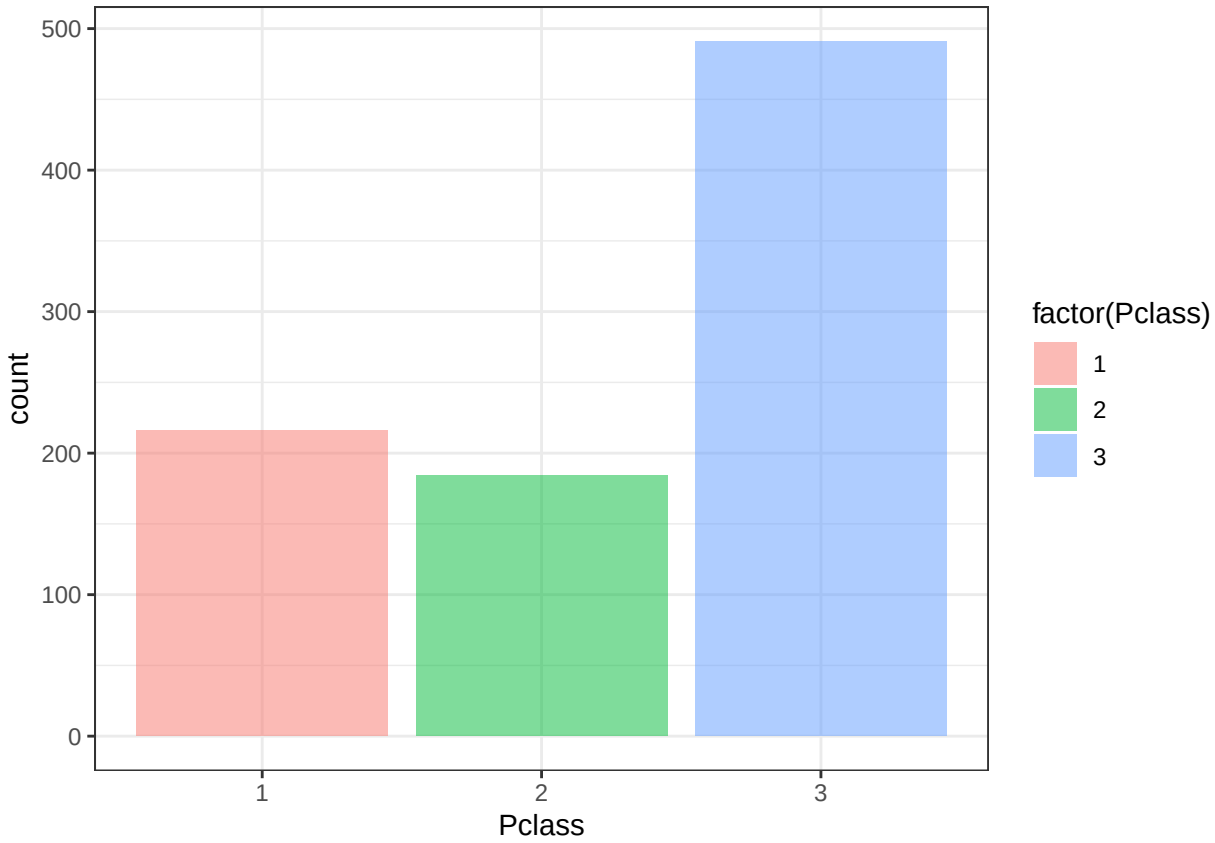
```r
#Exploratory Data Analysis
missmap(data.train, main="Titanic Training Data - Missings Map", col=c("yellow", "black"), legend=FALSE)
```



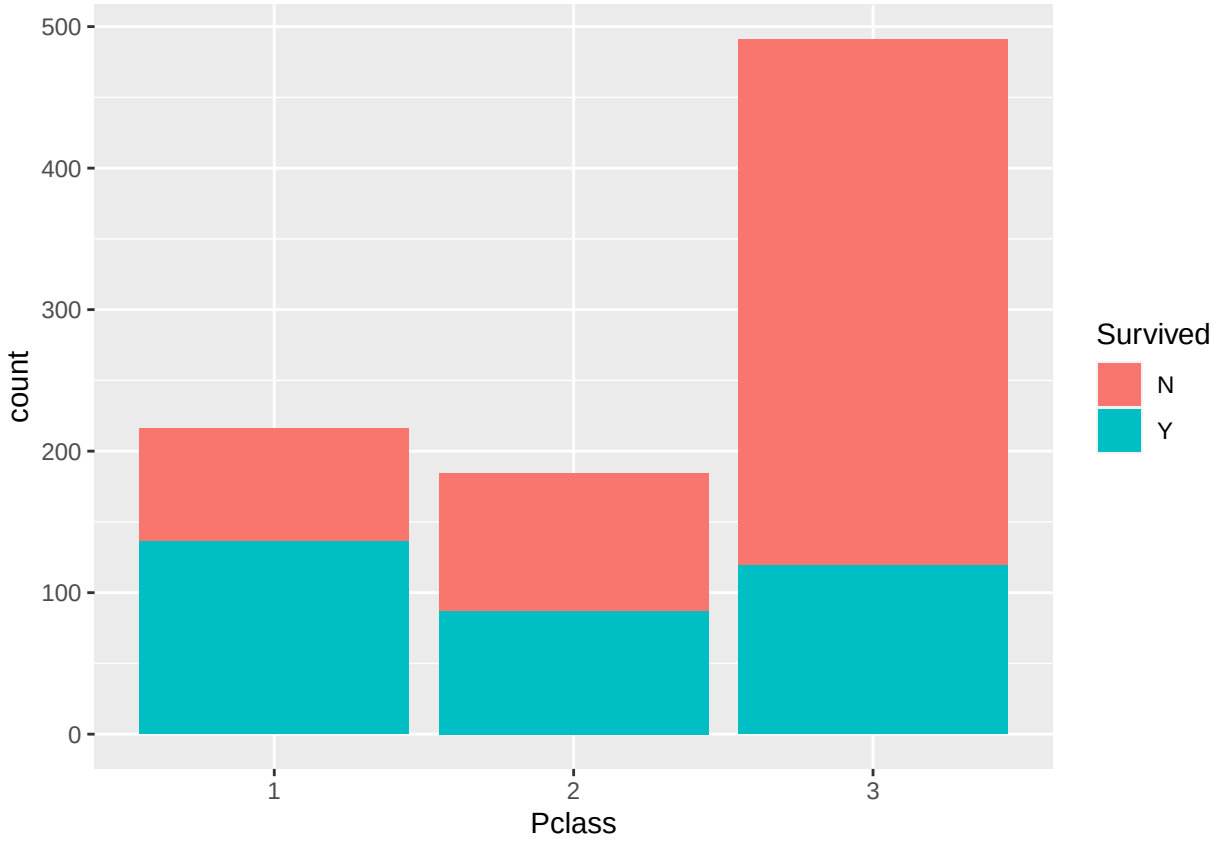

```r
ggplot(data.train, aes(Survived)) + geom_bar()
```



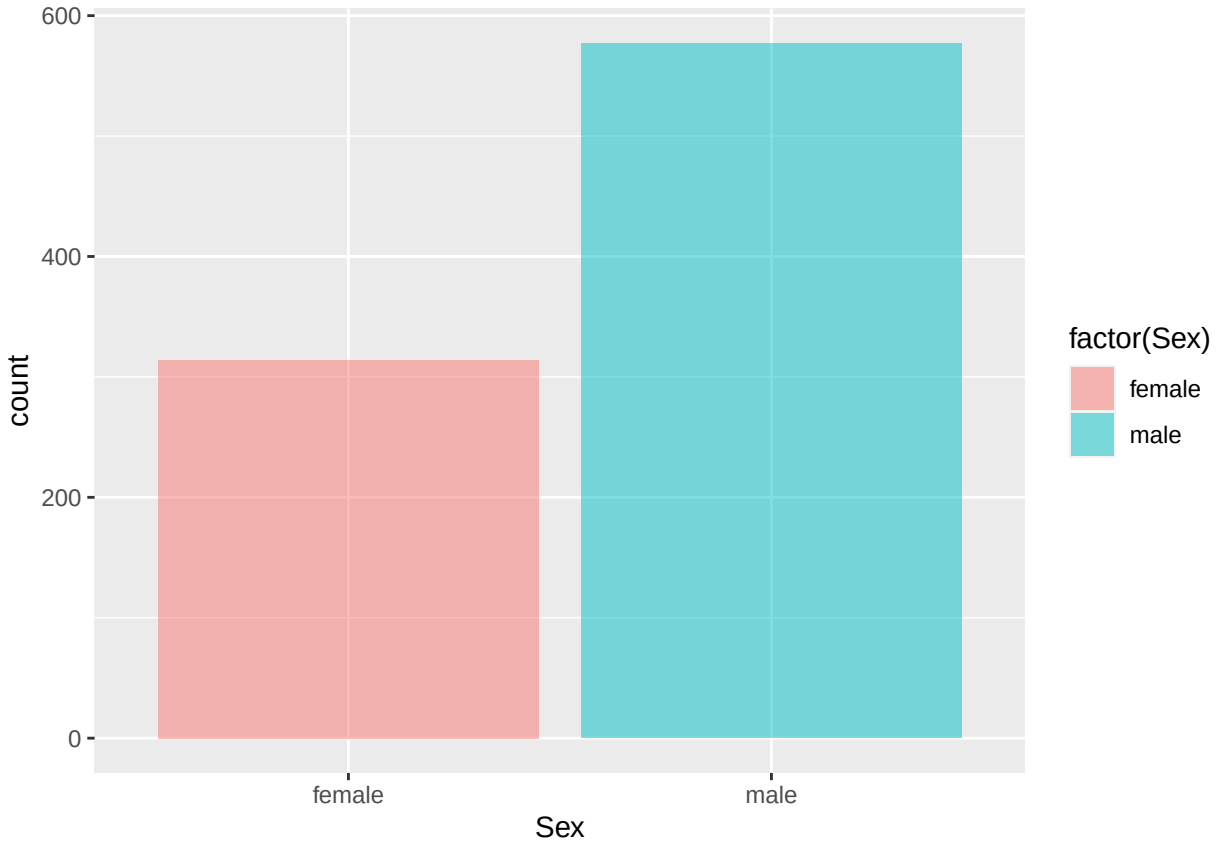

```
#549 died and 342 survived

#lets look at class
ggplot(data.train,aes(Pclass)) + geom_bar(aes(fill=factor(Pclass)),alpha=0.5) + theme_bw()
```



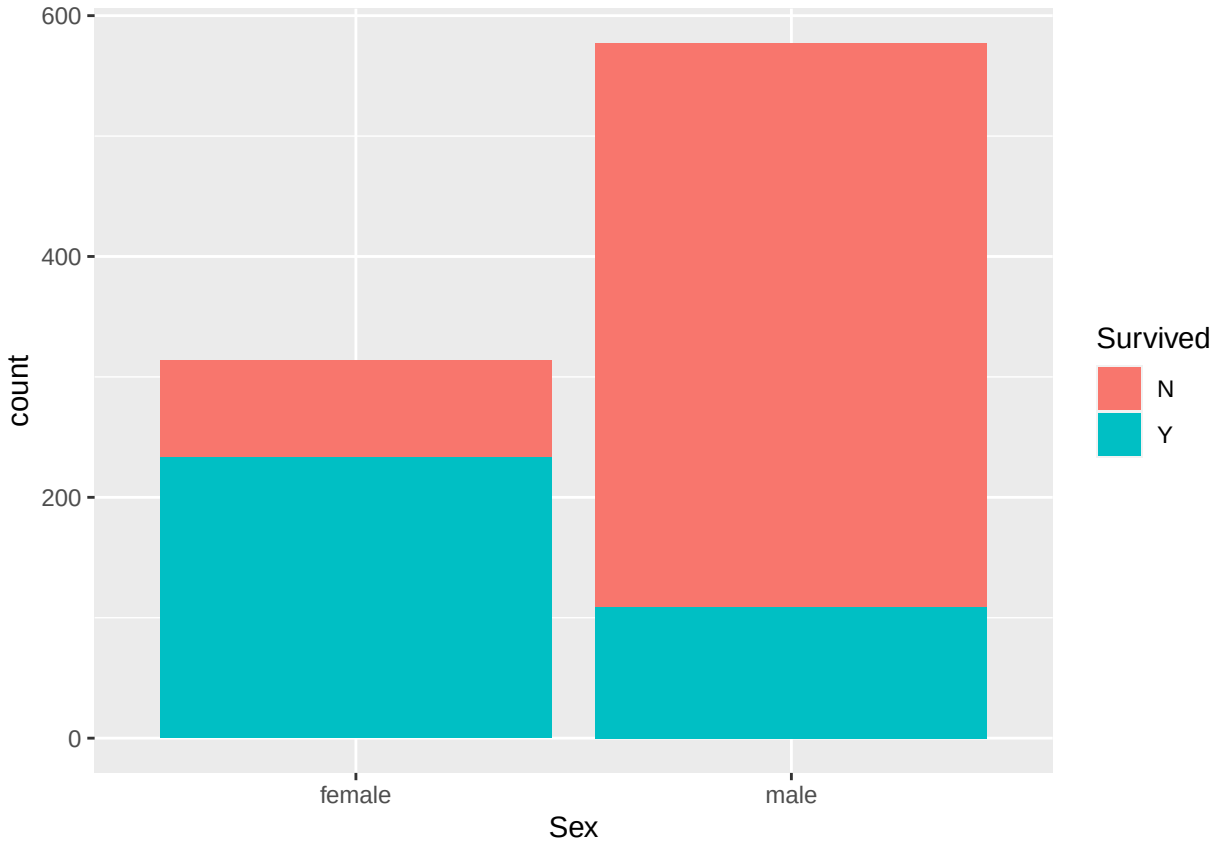

```
ggplot(data.train) + geom_bar(aes(x = Pclass,fill = Survived))
```



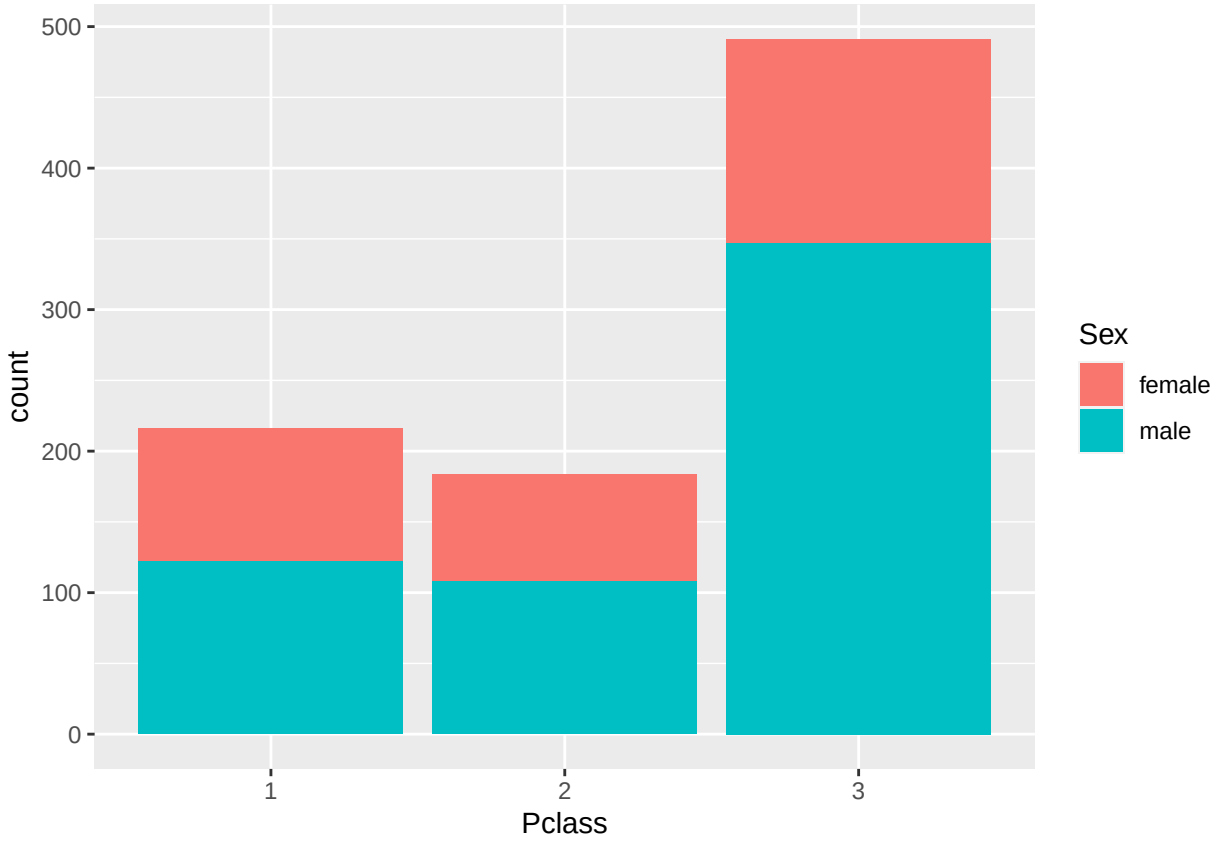

```
#more first class passengers survived compared to 3rd class that had a higher count.

#lets look at the sex
ggplot(data.train,aes(Sex)) + geom_bar(aes(fill=factor(Sex)),alpha=0.5)
```



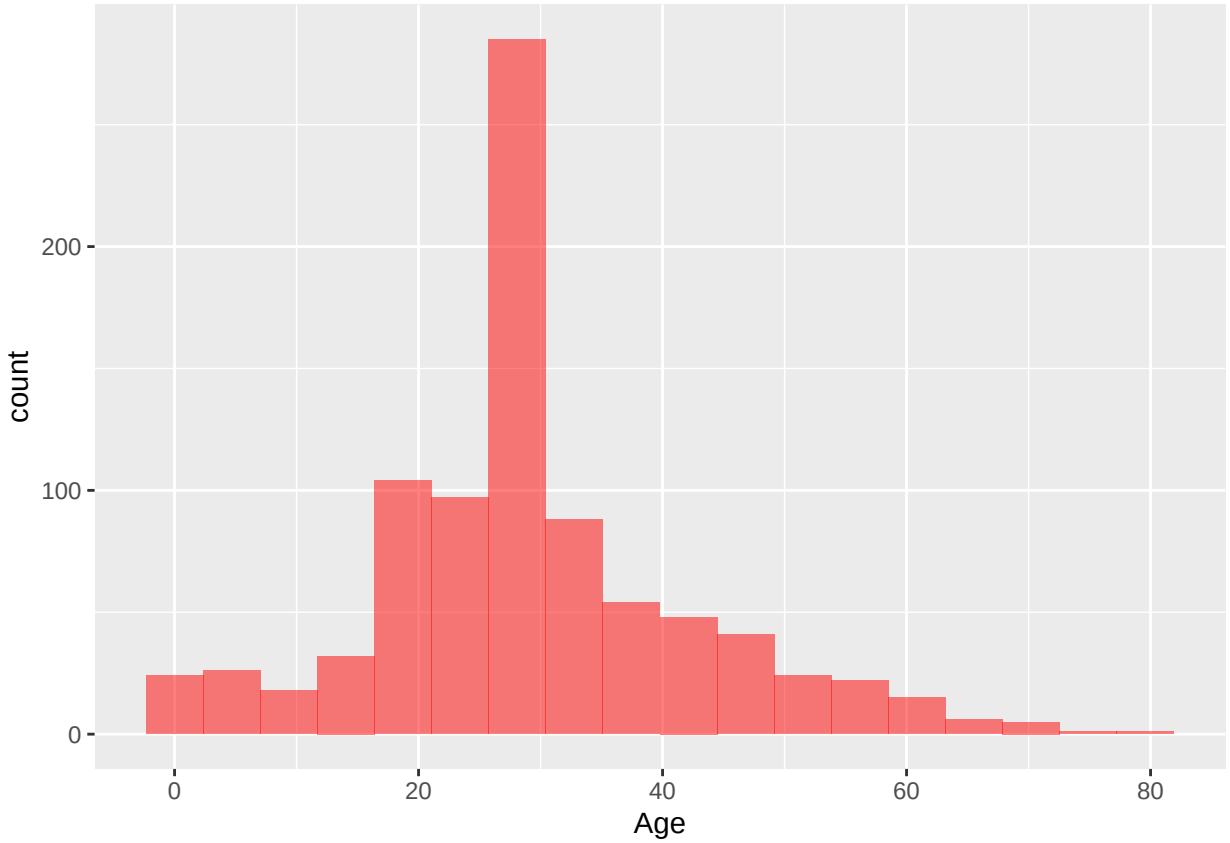

```
ggplot(data.train) + geom_bar(aes(x = Sex,fill = Survived))
```



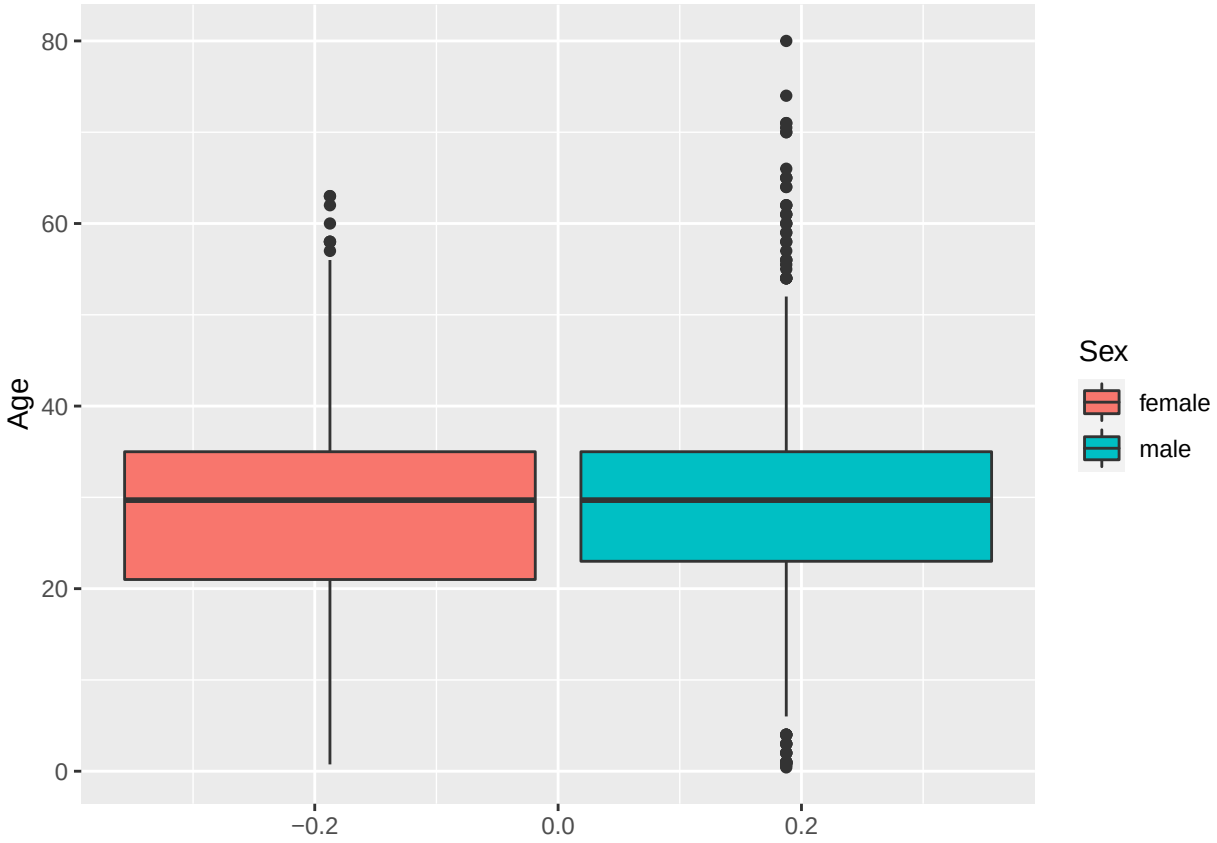

```
#although there were more males than females, females had a higher survival rate than men.

#Distribution of Male to females in first class passengers
ggplot(data.train) + geom_bar(aes(x = Pclass,fill = Sex))
```




```
#lets take a look at age
ggplot(data.train,aes(Age)) + geom_histogram(fill='red',bins=18,alpha=0.5)
```




```
ggplot(data.train,aes(y = Age)) + geom_boxplot(aes(fill=Sex))
```




```
#Feature engineering
data.train <- select(data.train,-PassengerId,-Name,-Ticket,-Cabin)
head(data.train)
```

```
##   Survived Pclass    Sex      Age SibSp Parch    Fare Embarked
## 1        N      3   male 22.00000     1     0  7.2500        S
## 2        Y      1 female 38.00000     1     0 71.2833        C
## 3        Y      3 female 26.00000     0     0  7.9250        S
## 4        Y      1 female 35.00000     1     0 53.1000        S
## 5        N      3   male 35.00000     0     0  8.0500        S
## 6        N      3   male 29.69912     0     0  8.4583        Q
```

```
str(data.train)
```

```
## 'data.frame':    891 obs. of  8 variables:
##  $ Survived: Factor w/ 2 levels "N","Y": 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 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...
##  $ Age     : num  22 38 26 35 35 ...
##  $ SibSp   : Factor w/ 7 levels "0","1","2","3",..: 2 2 1 2 1 1 1 4 1 2 ...
##  $ Parch   : Factor w/ 7 levels "0","1","2","3",..: 1 1 1 1 1 1 1 2 3 1 ...
##  $ Fare    : num  7.25 71.28 7.92 53.1 8.05 ...
##  $ Embarked: Factor w/ 4 levels "","C","Q","S": 4 2 4 4 4 3 4 4 4 2 ...
```

```r
#building the model
log.model1 <- glm(Survived ~ . , family = binomial,data = data.train)
summary(log.model1)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial, data = data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7733  -0.6184  -0.4178   0.5847   2.4637
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.827e+01  1.664e+03   0.011  0.99124
## Pclass2     -9.840e-01  3.028e-01  -3.250  0.00115 **
## Pclass3     -2.036e+00  3.020e-01  -6.741 1.57e-11 ***
## Sexmale     -2.667e+00  2.031e-01 -13.131  < 2e-16 ***
## Age         -3.663e-02  8.380e-03  -4.372 1.23e-05 ***
## SibSp1       9.961e-02  2.239e-01   0.445  0.65636
## SibSp2      -2.750e-01  5.352e-01  -0.514  0.60739
## SibSp3      -2.186e+00  7.196e-01  -3.038  0.00238 **
## SibSp4      -1.699e+00  7.611e-01  -2.233  0.02556 *
## SibSp5      -1.598e+01  9.582e+02  -0.017  0.98669
## SibSp8      -1.594e+01  7.579e+02  -0.021  0.98322
## Parch1       3.780e-01  2.888e-01   1.309  0.19067
## Parch2       7.679e-02  3.795e-01   0.202  0.83966
## Parch3       3.034e-01  1.053e+00   0.288  0.77335
## Parch4      -1.592e+01  1.056e+03  -0.015  0.98797
## Parch5      -1.267e+00  1.172e+00  -1.081  0.27965
## Parch6      -1.654e+01  2.400e+03  -0.007  0.99450
## Fare         2.253e-03  2.490e-03   0.905  0.36552
## EmbarkedC   -1.458e+01  1.664e+03  -0.009  0.99301
## EmbarkedQ   -1.448e+01  1.664e+03  -0.009  0.99306
## EmbarkedS   -1.489e+01  1.664e+03  -0.009  0.99286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  765.31  on 870  degrees of freedom
## AIC: 807.31
##
## Number of Fisher Scoring iterations: 15
```

```r
#or I could have used new.step.model <- step(model)


log.model2 <- glm(Survived ~ Pclass + Age + Sex + SibSp , family = binomial, data = data.train)
summary(log.model2)
```

```
##
```

```
## Call:
## glm(formula = Survived ~ Pclass + Age + Sex + SibSp, family = binomial,
##     data = data.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8259  -0.5998  -0.4326   0.6147   2.4463
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   3.914045   0.406967   9.618  < 2e-16 ***
## Pclass2      -1.220029   0.264067  -4.620 3.83e-06 ***
## Pclass3      -2.288869   0.245340  -9.329  < 2e-16 ***
## Age          -0.041263   0.008001  -5.157 2.51e-07 ***
## Sexmale      -2.708998   0.195246 -13.875  < 2e-16 ***
## SibSp1        0.142801   0.210125   0.680  0.49676
## SibSp2       -0.142497   0.519745  -0.274  0.78396
## SibSp3       -2.073128   0.685422  -3.025  0.00249 **
## SibSp4       -1.668710   0.744919  -2.240  0.02508 *
## SibSp5      -16.004773 956.874492  -0.017  0.98666
## SibSp8      -15.833751 753.839723  -0.021  0.98324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  779.24  on 880  degrees of freedom
## AIC: 801.24
##
## Number of Fisher Scoring iterations: 15
```

```
#from the P-values, there is a strong relationship between Class, Sex and Age.
#reject null hypothesis.
```

```
#Testing the model
#preparing the test data set
data.test <- read.csv("C:/Users/prade/Downloads/test.csv")
str(data.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
##  $ 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  "" "" "" "" ...
##  $ Embarked   : chr  "Q" "S" "Q" "S" ...
```

```
summary(data.test)
```

```
##   PassengerId         Pclass          Name               Sex
##  Min.   : 892.0   Min.   :1.000   Length:418         Length:418
##  1st Qu.: 996.2   1st Qu.:1.000   Class :character   Class :character
##  Median :1100.5   Median :3.000   Mode  :character   Mode  :character
##  Mean   :1100.5   Mean   :2.266
##  3rd Qu.:1204.8   3rd Qu.:3.000
##  Max.   :1309.0   Max.   :3.000
##
##       Age             SibSp            Parch           Ticket
##  Min.   : 0.17   Min.   :0.0000   Min.   :0.0000   Length:418
##  1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.0000   Class :character
##  Median :27.00   Median :0.0000   Median :0.0000   Mode  :character
##  Mean   :30.27   Mean   :0.4474   Mean   :0.3923
##  3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :76.00   Max.   :8.0000   Max.   :9.0000
##  NA's   :86
##       Fare            Cabin             Embarked
##  Min.   :  0.000   Length:418         Length:418
##  1st Qu.:  7.896   Class :character   Class :character
##  Median : 14.454   Mode  :character   Mode  :character
##  Mean   : 35.627
##  3rd Qu.: 31.500
##  Max.   :512.329
##  NA's   :1
```

```
#like our test data, there are 86 NA's observed in the Age Colomn which makes up ~22% of our data

data.test$Pclass <- as.factor(data.test$Pclass)
data.test$Sex <- as.factor(data.test$Sex)
data.test$Parch <- factor(data.test$Parch)
data.test$SibSp <- factor(data.test$SibSp)
data.test$Embarked <- factor(data.test$Embarked)

data.test$Age[is.na(data.test$Age)] <- mean(data.test$Age, na.rm = TRUE)
data.test <- select(data.test,-PassengerId,-Name,-Ticket,-Cabin)
str(data.test)
```

```
## 'data.frame':    418 obs. of  7 variables:
##  $ Pclass  : Factor w/ 3 levels "1","2","3": 3 3 2 3 3 3 3 2 3 3 ...
##  $ 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   : Factor w/ 7 levels "0","1","2","3",..: 1 2 1 1 2 1 1 2 1 3 ...
##  $ Parch   : Factor w/ 8 levels "0","1","2","3",..: 1 1 1 1 2 1 1 2 1 1 ...
##  $ Fare    : num  7.83 7 9.69 8.66 12.29 ...
##  $ Embarked: Factor w/ 3 levels "C","Q","S": 2 3 2 3 3 3 2 3 1 3 ...
```

```
data.test$SurvivedP <- predict(log.model2,newdata = data.test,type = 'response')
head(data.test$SurvivedP)
```

```
## [1] 0.07534114 0.45725767 0.07087715 0.09993694 0.70270020 0.15956000
```

```
head(data.test)
```

```
##   Pclass    Sex  Age SibSp Parch    Fare Embarked  SurvivedP
## 1      3   male 34.5     0     0  7.8292        Q 0.07534114
## 2      3 female 47.0     1     0  7.0000        S 0.45725767
## 3      2   male 62.0     0     0  9.6875        Q 0.07087715
## 4      3   male 27.0     0     0  8.6625        S 0.09993694
## 5      3 female 22.0     1     1 12.2875        S 0.70270020
## 6      3   male 14.0     0     0  9.2250        S 0.15956000
```

```
#since the survival values range from 1 to 0, where 1 = survived and 0 = died
#creating an if column

data.test$predictedsurvival<-ifelse(data.test$SurvivedP>0.5, 1, 0)
head(data.test)
```

```
##   Pclass    Sex  Age SibSp Parch    Fare Embarked  SurvivedP predictedsurvival
## 1      3   male 34.5     0     0  7.8292        Q 0.07534114                 0
## 2      3 female 47.0     1     0  7.0000        S 0.45725767                 0
## 3      2   male 62.0     0     0  9.6875        Q 0.07087715                 0
## 4      3   male 27.0     0     0  8.6625        S 0.09993694                 0
## 5      3 female 22.0     1     1 12.2875        S 0.70270020                 1
## 6      3   male 14.0     0     0  9.2250        S 0.15956000                 0
```