MET CS 555 Term Project

10 points

1. Assignment Description

Select a small data set from the available public data sets (you can find a list of public data sets here http://www.teymourian.de/public-data-sets-for-data-analytic-projects/).

Describe a research scenario and specify a research question based on data analytic methods that we learned in our class, for example, methods like, *one and two sample means, t-test, correlation tests, simple and multiple linear regression, ANOVA and ANCOVA, one and two-Sample Tests for Proportions and logistic regression*.

Clean up your data and reduce it to no more than 500 observations if your data set is large.

2. Research Scenario Description (no more than 200 words)

Describe your research scenario in no more than 200 words. This is a general description of the use case. Similar to our class examples, we first describe the overall scenario and then specify a specific research question based on it.

This is the legendary Titanic Kaggle competition –

The analysis is simple: create a model that predicts which passengers survived the Titanic shipwreck.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this, I'll build a predictive model that answers the question: "what sorts of people were more likely to survive?" using passenger data (ie name, age, gender, socio-economic class, etc).

3. Describe the data set (no more than 200 words)

Describe briefly the data set. Describe each column of the data set if you use the column in your analysis. Clean up your data before usage, for example, you can remove the outliers. Remove unused columns. If possible provide a link to the main data set source.

Link to dataset:
https://www.kaggle.com/c/titanic
With 891 rows and 12 columns before preprocessing.
Attribute Information

- survival : 0 = No, 1 = Yes
- pclass : the ticket class with values; 1 = 1st, 2 = 2nd, 3 = 3rd
- sex: Female or Male
- Age: Age in years
- sibsp: # of siblings / spouses aboard the Titanic
- parch: # of parents / children aboard the Titanic
- ticket: Ticket number
- fare: Passenger fare
- cabin: Cabin number
- embarked: Port of Embarkation with values; C = Cherbourg, Q = Queenstown, S = Southampton

cabin has are 687 NA values so I decided to drop it

pclass: A proxy for socio-economic status (SES)
1st = Upper
2nd = Middle
3rd = Lower

age: Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5

sibsp: The dataset defines family relations in this way...
Sibling = brother, sister, stepbrother, stepsister
Spouse = husband, wife (mistresses and fiancés were ignored)

parch: The dataset defines family relations in this way...
Parent = mother, father
Child = daughter, son, stepdaughter, stepson
Some children travelled only with a nanny, therefore parch=0 for them.

3. Research Question (no more than 100 words)

Describe briefly in one or two sentences the main research question. This is similar to the last sentence of our class examples.

We would like to answer the following questions:

- Which of the features seem to have a significant influence (significance level(α) = 0.05) on an individual who survived the sinking?
- Do these features still have significant influence, after adjusting for relevant covariants(age, gender…)?
- How good is our classification model (i.e., how accurately does our model predict that an individual is going to survive or not)?

4. Your solution R code

Copy your R code here. Start from read the data from a data file. Keep the following data read line.

This is similar to one of our R code examples.

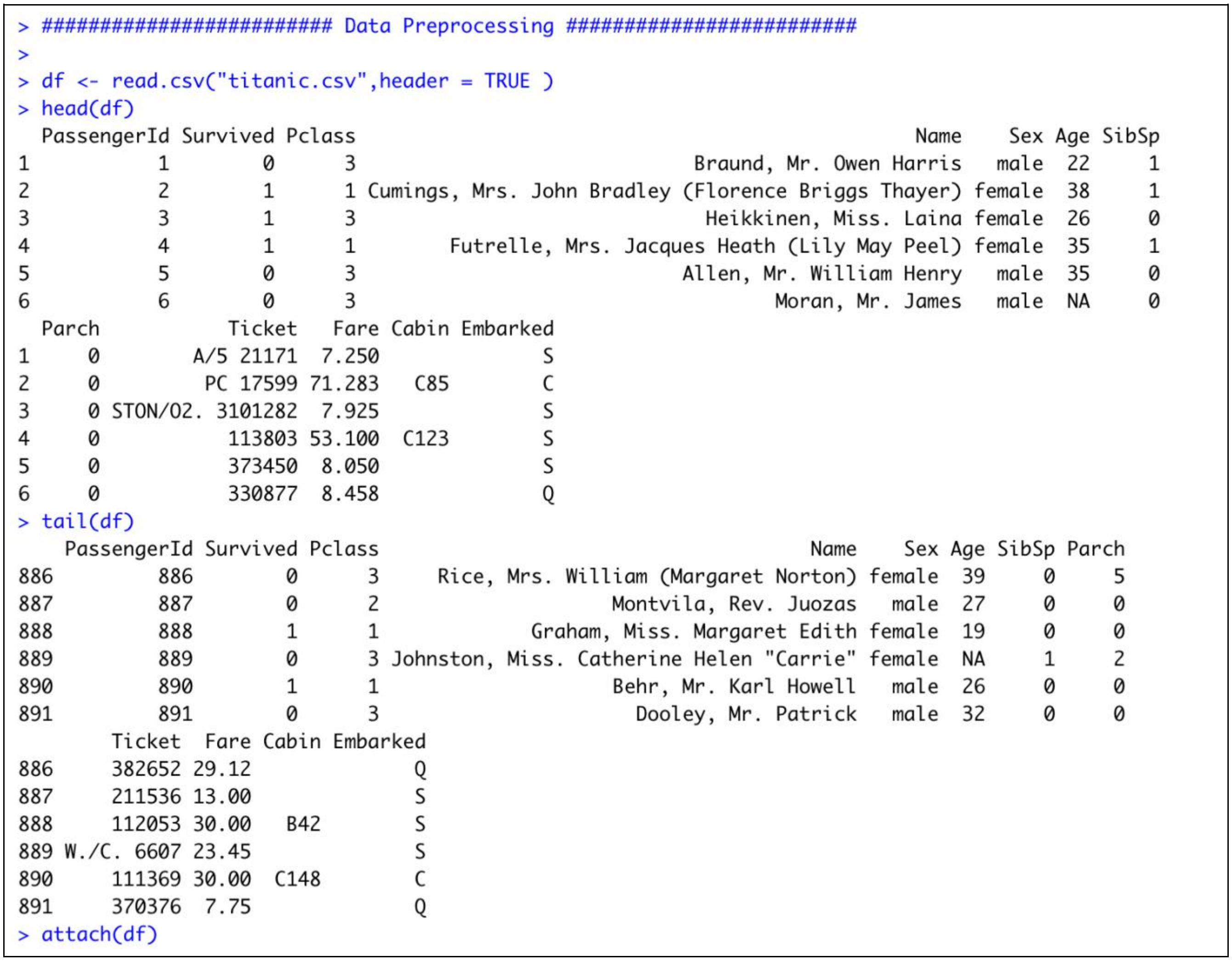
```
######################### Titanic Survival #########################

# CS555 Final Project
# @author : Mohamed Faadil
# BU ID : U87311082
# Dataset used: https://www.kaggle.com/c/titanic

options(digits = 4 , scipen = 4)
setwd("/Users/sahilkhanna/Downloads")
library(tidyverse)
#install.packages('caTools')
library(caTools)
#install.packages("ggcorrplot")
library("ggcorrplot")
# install.packages("fastDummies")
library(fastDummies)
# install.packages("lsmeans")
library("lsmeans")
#install.packages("ggfortify") ## installation required
library(ggfortify)
#install.packages("olsrr") ## installation required
library("olsrr")
library(car)
library(GGally)


######################### Data Preprocessing #########################

df <- read.csv("titanic.csv",header = TRUE )
head(df)
tail(df)
attach(df)
colnames(df_clean)
df_clean <- (df[c(2,3,5,6,7,8,10,12)])
df_clean <- na.omit(df_clean)
df_clean <- head(df_clean,500)
nrow(df_clean)

######################### Exploratory Data Analysis #########################
df_dummy <- fastDummies::dummy_cols(df_clean,
                                    select_columns = c("Sex","Embarked"))
colnames(df_dummy)
df_dummy <- df_dummy[c(1,2,4,5,6,7,9,10,12,13,14)]
corr <- (cor(df_dummy[,-9]))
ggcorrplot(corr, hc.order = F, outline.col = "white",
           lab = TRUE, title= "Correlation heatmap")

#ggpairs(df_dummy, title="Correlation Pairplot")

summary(df_clean)
summary(subset(df_clean, df_clean$Survived ==0))
summary(subset(df_clean, df_clean$Survived ==1))



ggplot(df_clean, aes(x=Survived, y=Age, fill=Sex)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=2, notch=FALSE) +
  labs(title="Box plot - Test Scores",
       x = "Survived",
       y = "Age")
#Bar Plots Survived vs Non-Survived
ggplot(df_clean, aes(x=Survived, y=Age, fill=Sex)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8,
               outlier.size=2, notch=FALSE) +
  labs(title="Box plot",
       x = "Survived",
       y = "Age")

ggplot(df_clean, aes(x=Survived, y=Age, fill=Sex)) +
  geom_bar(stat="identity") +
  labs(title="Bar plot",
       x = "Survived",
       y = "")
ggplot(df_clean, aes(x=Pclass, y=Survived, fill=Sex)) +
  geom_bar(stat="identity") +
  labs(title="Bar plot",
       x = "Passenger Class",
       y = "")

# Building logistic regression classification models
## Passenger Class
m <- glm(Survived~Pclass,
         data = df_dummy , family = binomial)
summary(m)
```

```
#Adjusting for age:
library(car)
Anova(glm(Survived~Pclass + Age,
        data = df_dummy , family = binomial))

## Sex
m <- glm(Survived~Sex_female,
        data = df_dummy , family = binomial)
summary(m)

#Adjusting for age:
Anova(glm(Survived~Sex_female+Age,
         data = df_dummy , family = binomial))

## Fare
m <- glm(Survived~Fare,
        data = df_dummy , family = binomial)
summary(m)

#Adjusting for age:
Anova(glm(Survived~Fare+Age,
         data = df_dummy , family = binomial))

## Family
m <- glm(Survived~Parch,
        data = df_dummy , family = binomial)
summary(m)

#Adjusting for age:
Anova(glm(Survived~Parch+Age,
         data = df_dummy , family = binomial))

## Siblings
m <- glm(Survived~SibSp,
        data = df_dummy , family = binomial)
summary(m)

#Adjusting for age:
Anova(glm(Survived~SibSp+Age,
         data = df_dummy , family = binomial))


#building the final logistic regression classification model

set.seed(64)
split = sample.split(df_dummy$Survived, SplitRatio = 0.8)
training_set = subset(df_dummy, split == TRUE)
test_set = subset(df_dummy, split == FALSE)
training_set[-1] = scale(training_set[-1])
test_set[-1] = scale(test_set[-1])
```

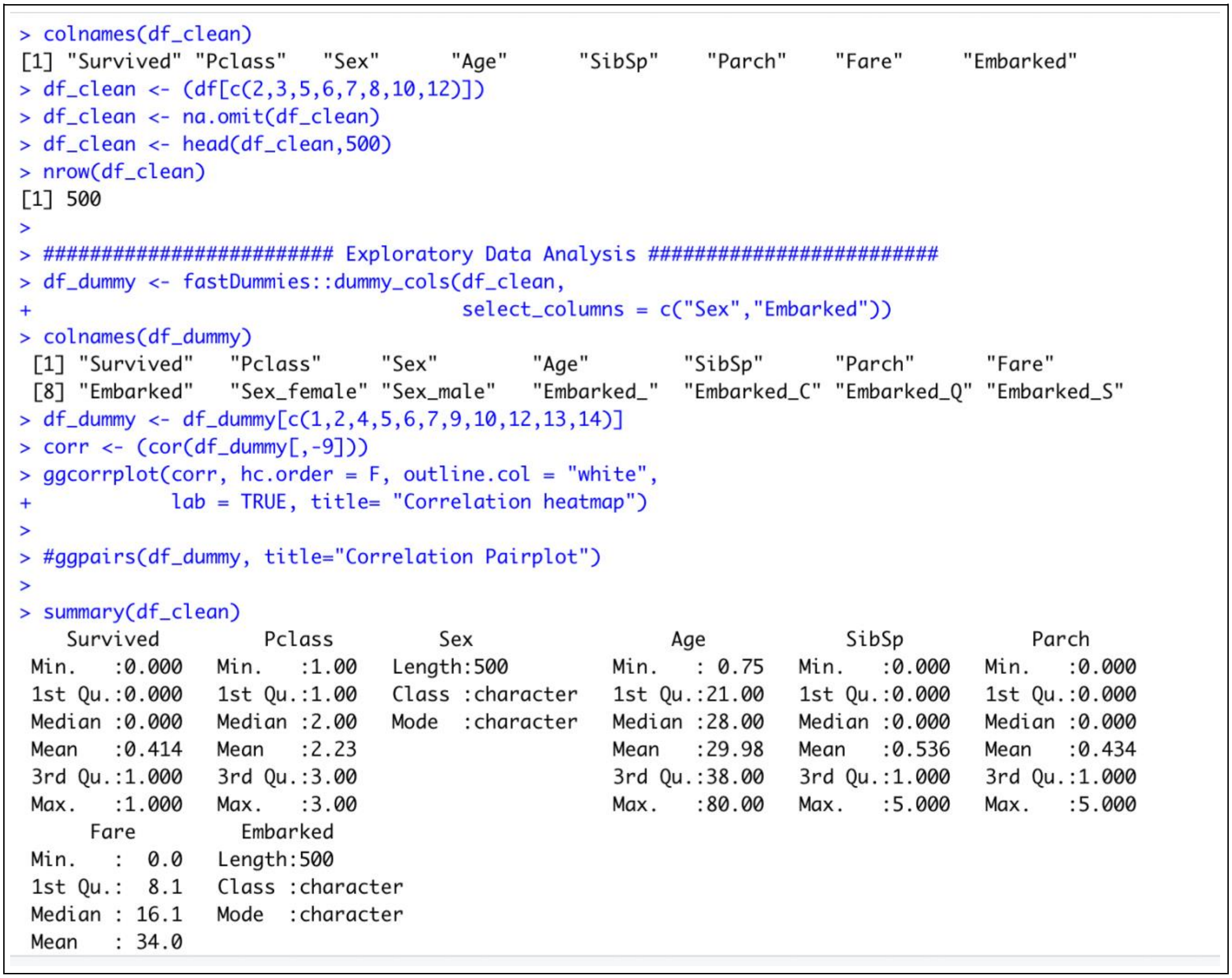
```
classifier = glm(formula = Survived ~Pclass+Sex_female+Fare+Age,
                 family = binomial,
                 data = training_set)

summary(classifier)
# Predicting the Test set results
model.probs <- predict(classifier, test_set, type = "response")
model.pred <- rep(0, length(model.probs))
model.pred[model.probs > 0.5] <- 1

# Making the Confusion Matrix
table(model.pred, test_set$Survived)

# Accuracy
1- mean(model.pred != test_set$Survived)

# ROC curve
#install.packages("pROC")
library(pROC)
g <- roc(test_set$Survived ~ model.probs)
print(g)
plot(g, main = "ROC curve")
```

5. Execute your R code, Copy and Paste results here in this Box.

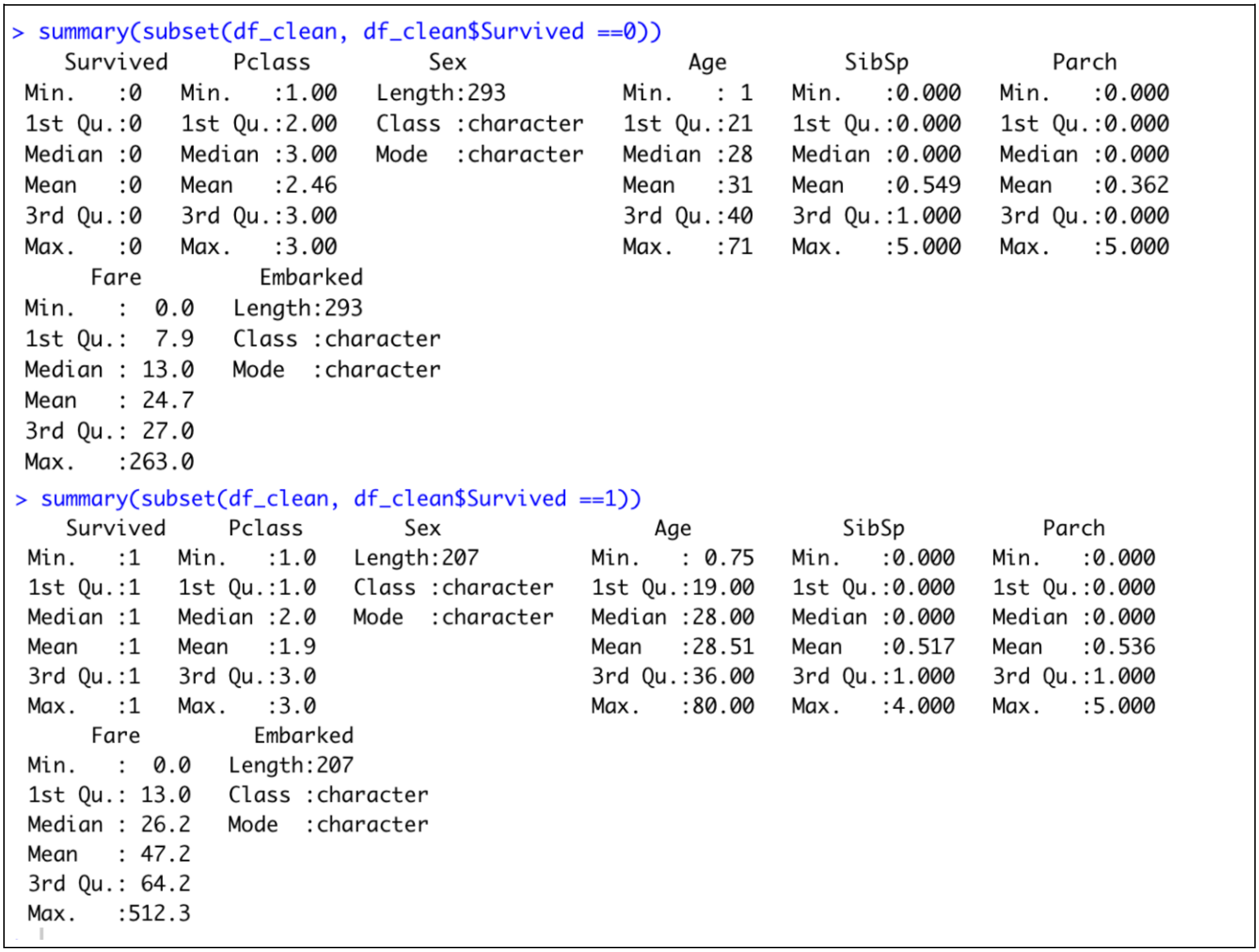
Run your code and copy the output of your code to here.



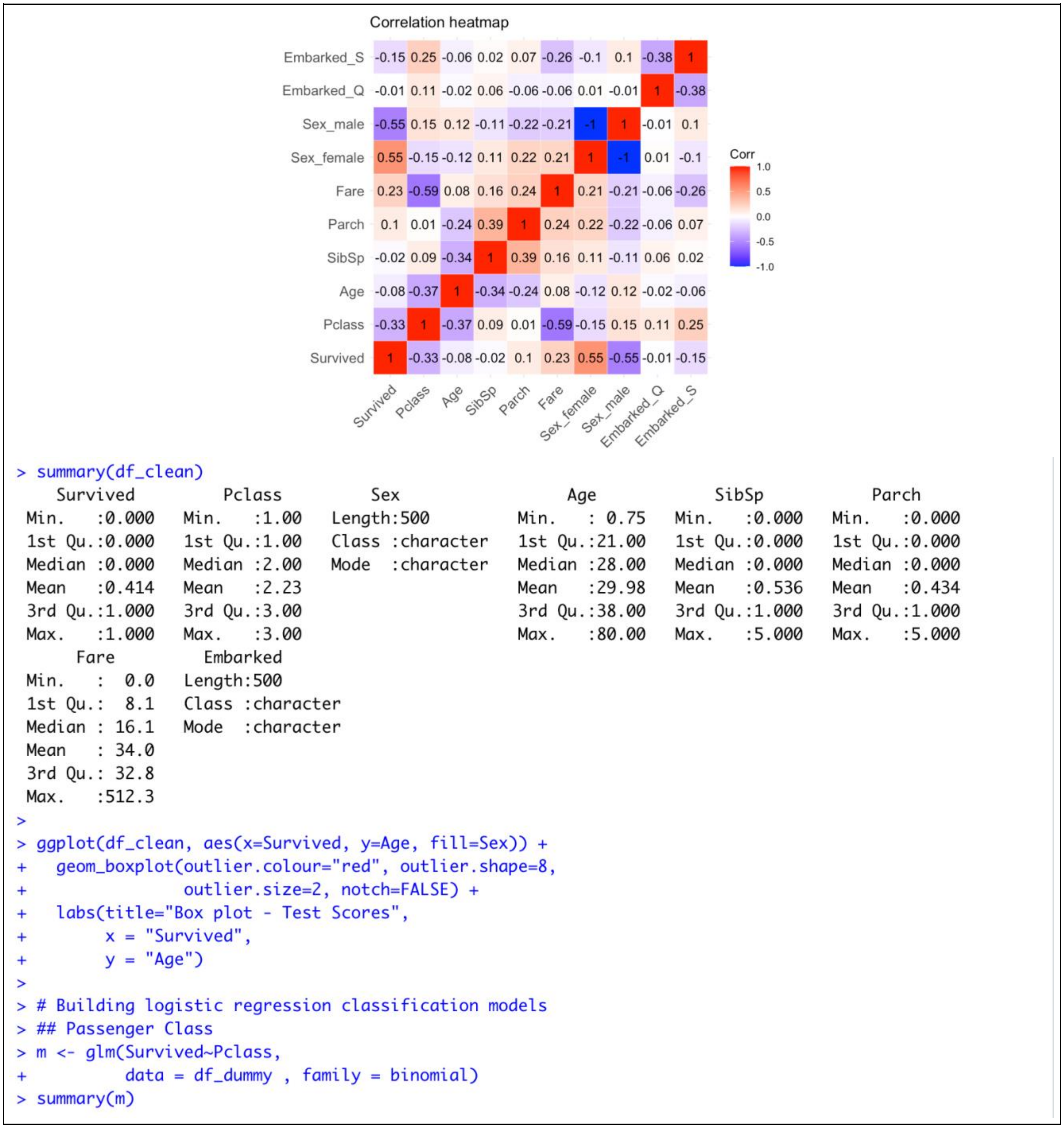
```
> ######################## Data Preprocessing ########################
>
> df <- read.csv("titanic.csv",header = TRUE )
> head(df)
  PassengerId Survived Pclass                                                Name    Sex Age SibSp
1           1        0      3                             Braund, Mr. Owen Harris   male  22     1
2           2        1      1 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38     1
3           3        1      3                              Heikkinen, Miss. Laina female  26     0
4           4        1      1        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35     1
5           5        0      3                            Allen, Mr. William Henry   male  35     0
6           6        0      3                                    Moran, Mr. James   male  NA     0
  Parch           Ticket   Fare Cabin Embarked
1     0        A/5 21171  7.250                S
2     0         PC 17599 71.283   C85          C
3     0 STON/O2. 3101282  7.925                S
4     0           113803 53.100  C123          S
5     0           373450  8.050                S
6     0           330877  8.458                Q
> tail(df)
    PassengerId Survived Pclass                                     Name    Sex Age SibSp Parch
886         886        0      3     Rice, Mrs. William (Margaret Norton) female  39     0     5
887         887        0      2                    Montvila, Rev. Juozas   male  27     0     0
888         888        1      1             Graham, Miss. Margaret Edith female  19     0     0
889         889        0      3 Johnston, Miss. Catherine Helen "Carrie" female  NA     1     2
890         890        1      1                    Behr, Mr. Karl Howell   male  26     0     0
891         891        0      3                      Dooley, Mr. Patrick   male  32     0     0
        Ticket  Fare Cabin Embarked
886     382652 29.12                Q
887     211536 13.00                S
888     112053 30.00   B42          S
889 W./C. 6607 23.45                S
890     111369 30.00  C148          C
891     370376  7.75                Q
> attach(df)
```



```
> colnames(df_clean)
[1] "Survived" "Pclass"   "Sex"      "Age"      "SibSp"    "Parch"    "Fare"     "Embarked"
> df_clean <- (df[c(2,3,5,6,7,8,10,12)])
> df_clean <- na.omit(df_clean)
> df_clean <- head(df_clean,500)
> nrow(df_clean)
[1] 500
>
> ######################## Exploratory Data Analysis ########################
> df_dummy <- fastDummies::dummy_cols(df_clean,
+                                     select_columns = c("Sex","Embarked"))
> colnames(df_dummy)
 [1] "Survived"   "Pclass"     "Sex"        "Age"        "SibSp"      "Parch"      "Fare"      
 [8] "Embarked"   "Sex_female" "Sex_male"   "Embarked_"  "Embarked_C" "Embarked_Q" "Embarked_S"
> df_dummy <- df_dummy[c(1,2,4,5,6,7,9,10,12,13,14)]
> corr <- (cor(df_dummy[,-9]))
> ggcorrplot(corr, hc.order = F, outline.col = "white",
+            lab = TRUE, title= "Correlation heatmap")
>
> #ggpairs(df_dummy, title="Correlation Pairplot")
>
> summary(df_clean)
    Survived         Pclass          Sex                 Age            SibSp           Parch      
 Min.   :0.000   Min.   :1.00   Length:500         Min.   : 0.75   Min.   :0.000   Min.   :0.000  
 1st Qu.:0.000   1st Qu.:1.00   Class :character   1st Qu.:21.00   1st Qu.:0.000   1st Qu.:0.000  
 Median :0.000   Median :2.00   Mode  :character   Median :28.00   Median :0.000   Median :0.000  
 Mean   :0.414   Mean   :2.23                      Mean   :29.98   Mean   :0.536   Mean   :0.434  
 3rd Qu.:1.000   3rd Qu.:3.00                      3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:1.000  
 Max.   :1.000   Max.   :3.00                      Max.   :80.00   Max.   :5.000   Max.   :5.000  
      Fare          Embarked        
 Min.   :  0.0   Length:500        
 1st Qu.:  8.1   Class :character  
 Median : 16.1   Mode  :character  
 Mean   : 34.0                     
```

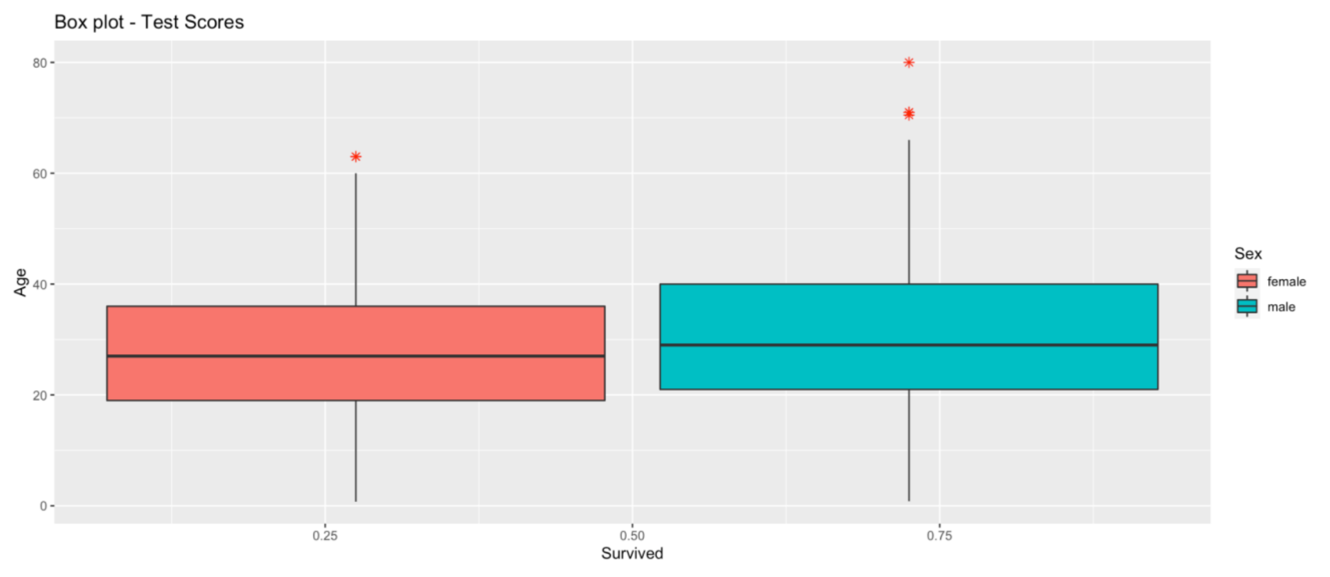


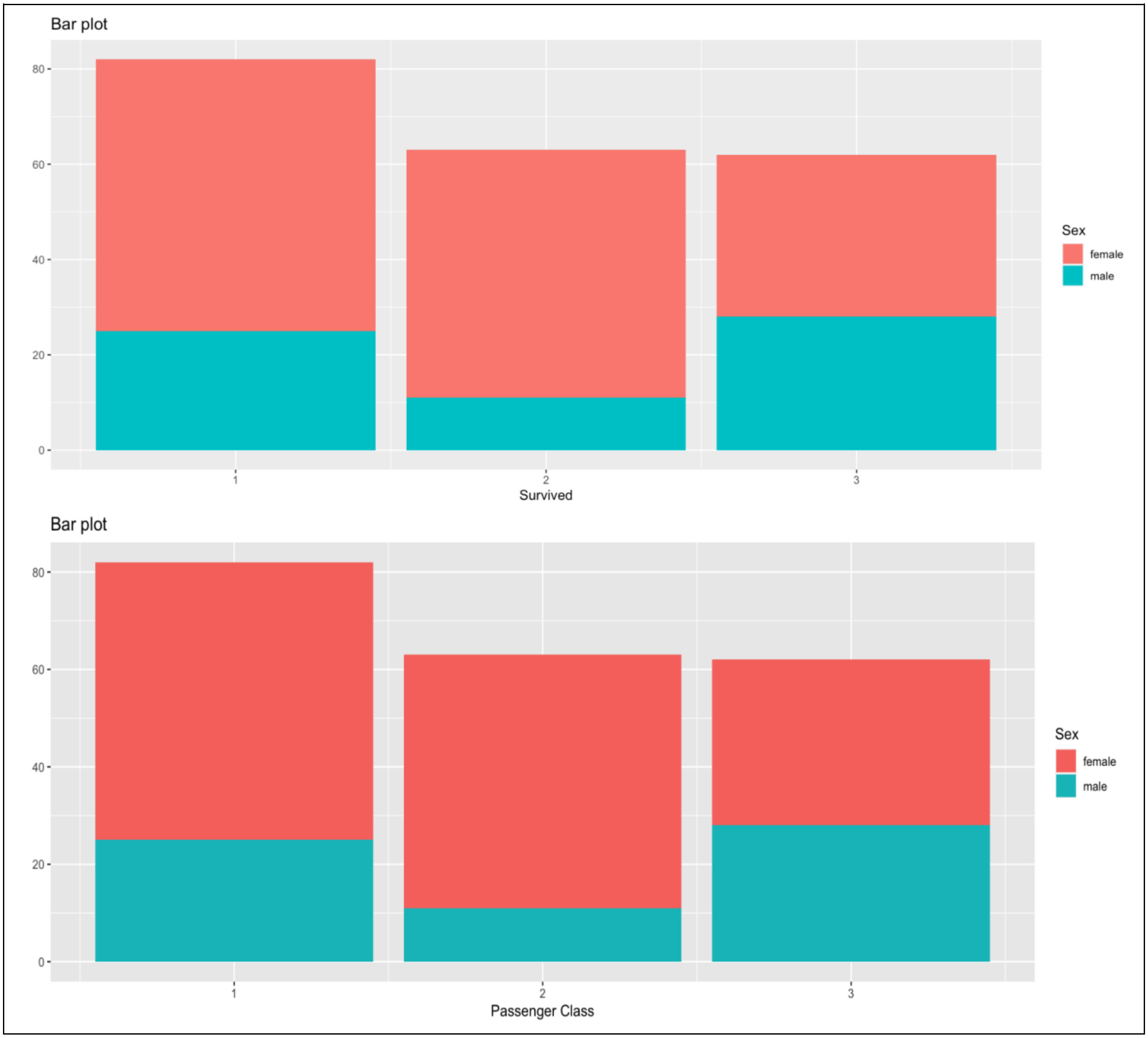
```
> summary(subset(df_clean, df_clean$Survived ==0))
    Survived      Pclass          Sex                 Age          SibSp           Parch      
 Min.   :0   Min.   :1.00   Length:293         Min.   : 1   Min.   :0.000   Min.   :0.000  
 1st Qu.:0   1st Qu.:2.00   Class :character   1st Qu.:21   1st Qu.:0.000   1st Qu.:0.000  
 Median :0   Median :3.00   Mode  :character   Median :28   Median :0.000   Median :0.000  
 Mean   :0   Mean   :2.46                      Mean   :31   Mean   :0.549   Mean   :0.362  
 3rd Qu.:0   3rd Qu.:3.00                      3rd Qu.:40   3rd Qu.:1.000   3rd Qu.:0.000  
 Max.   :0   Max.   :3.00                      Max.   :71   Max.   :5.000   Max.   :5.000  
      Fare         Embarked        
 Min.   :  0.0   Length:293        
 1st Qu.:  7.9   Class :character  
 Median : 13.0   Mode  :character  
 Mean   : 24.7                     
 3rd Qu.: 27.0                     
 Max.   :263.0                     
> summary(subset(df_clean, df_clean$Survived ==1))
    Survived     Pclass         Sex                 Age             SibSp           Parch      
 Min.   :1   Min.   :1.0   Length:207         Min.   : 0.75   Min.   :0.000   Min.   :0.000  
 1st Qu.:1   1st Qu.:1.0   Class :character   1st Qu.:19.00   1st Qu.:0.000   1st Qu.:0.000  
 Median :1   Median :2.0   Mode  :character   Median :28.00   Median :0.000   Median :0.000  
 Mean   :1   Mean   :1.9                      Mean   :28.51   Mean   :0.517   Mean   :0.536  
 3rd Qu.:1   3rd Qu.:3.0                      3rd Qu.:36.00   3rd Qu.:1.000   3rd Qu.:1.000  
 Max.   :1   Max.   :3.0                      Max.   :80.00   Max.   :4.000   Max.   :5.000  
      Fare         Embarked        
 Min.   :  0.0   Length:207        
 1st Qu.: 13.0   Class :character  
 Median : 26.2   Mode  :character  
 Mean   : 47.2                     
 3rd Qu.: 64.2                     
 Max.   :512.3                     
```


Correlation heatmap

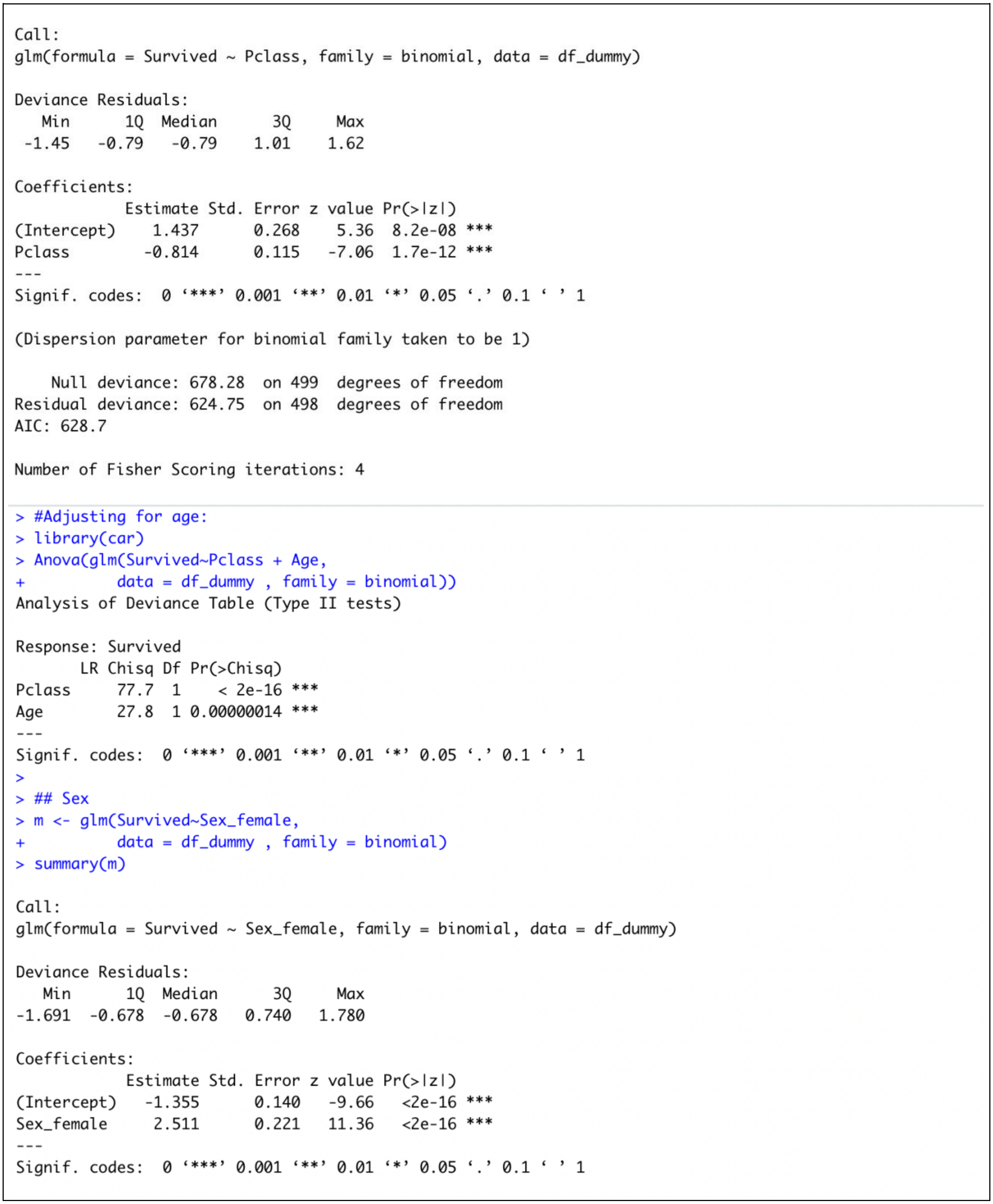
	Survived	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Embarked_Q	Embarked_S
Embarked_S	-0.15	0.25	-0.06	0.02	0.07	-0.26	-0.1	0.1	-0.38	1
Embarked_Q	-0.01	0.11	-0.02	0.06	-0.06	-0.06	0.01	-0.01	1	-0.38
Sex_male	-0.55	0.15	0.12	-0.11	-0.22	-0.21	-1	1	-0.01	0.1
Sex_female	0.55	-0.15	-0.12	0.11	0.22	0.21	1	-1	0.01	-0.1
Fare	0.23	-0.59	0.08	0.16	0.24	1	0.21	-0.21	-0.06	-0.26
Parch	0.1	0.01	-0.24	0.39	1	0.24	0.22	-0.22	-0.06	0.07
SibSp	-0.02	0.09	-0.34	1	0.39	0.16	0.11	-0.11	0.06	0.02
Age	-0.08	-0.37	1	-0.34	-0.24	0.08	-0.12	0.12	-0.02	-0.06
Pclass	-0.33	1	-0.37	0.09	0.01	-0.59	-0.15	0.15	0.11	0.25
Survived	1	-0.33	-0.08	-0.02	0.1	0.23	0.55	-0.55	-0.01	-0.15

```
> summary(df_clean)
    Survived          Pclass           Sex                  Age            SibSp            Parch      
 Min.   :0.000   Min.   :1.00   Length:500         Min.   : 0.75   Min.   :0.000   Min.   :0.000  
 1st Qu.:0.000   1st Qu.:1.00   Class :character   1st Qu.:21.00   1st Qu.:0.000   1st Qu.:0.000  
 Median :0.000   Median :2.00   Mode  :character   Median :28.00   Median :0.000   Median :0.000  
 Mean   :0.414   Mean   :2.23                      Mean   :29.98   Mean   :0.536   Mean   :0.434  
 3rd Qu.:1.000   3rd Qu.:3.00                      3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:1.000  
 Max.   :1.000   Max.   :3.00                      Max.   :80.00   Max.   :5.000   Max.   :5.000  
      Fare          Embarked        
 Min.   :  0.0   Length:500        
 1st Qu.:  8.1   Class :character  
 Median : 16.1   Mode  :character  
 Mean   : 34.0                     
 3rd Qu.: 32.8                     
 Max.   :512.3                     
> 
> ggplot(df_clean, aes(x=Survived, y=Age, fill=Sex)) +
+   geom_boxplot(outlier.colour="red", outlier.shape=8,
+                outlier.size=2, notch=FALSE) +
+   labs(title="Box plot - Test Scores",
+        x = "Survived",
+        y = "Age")
> 
> # Building logistic regression classification models
> ## Passenger Class
> m <- glm(Survived~Pclass,
+          data = df_dummy , family = binomial)
> summary(m)
```

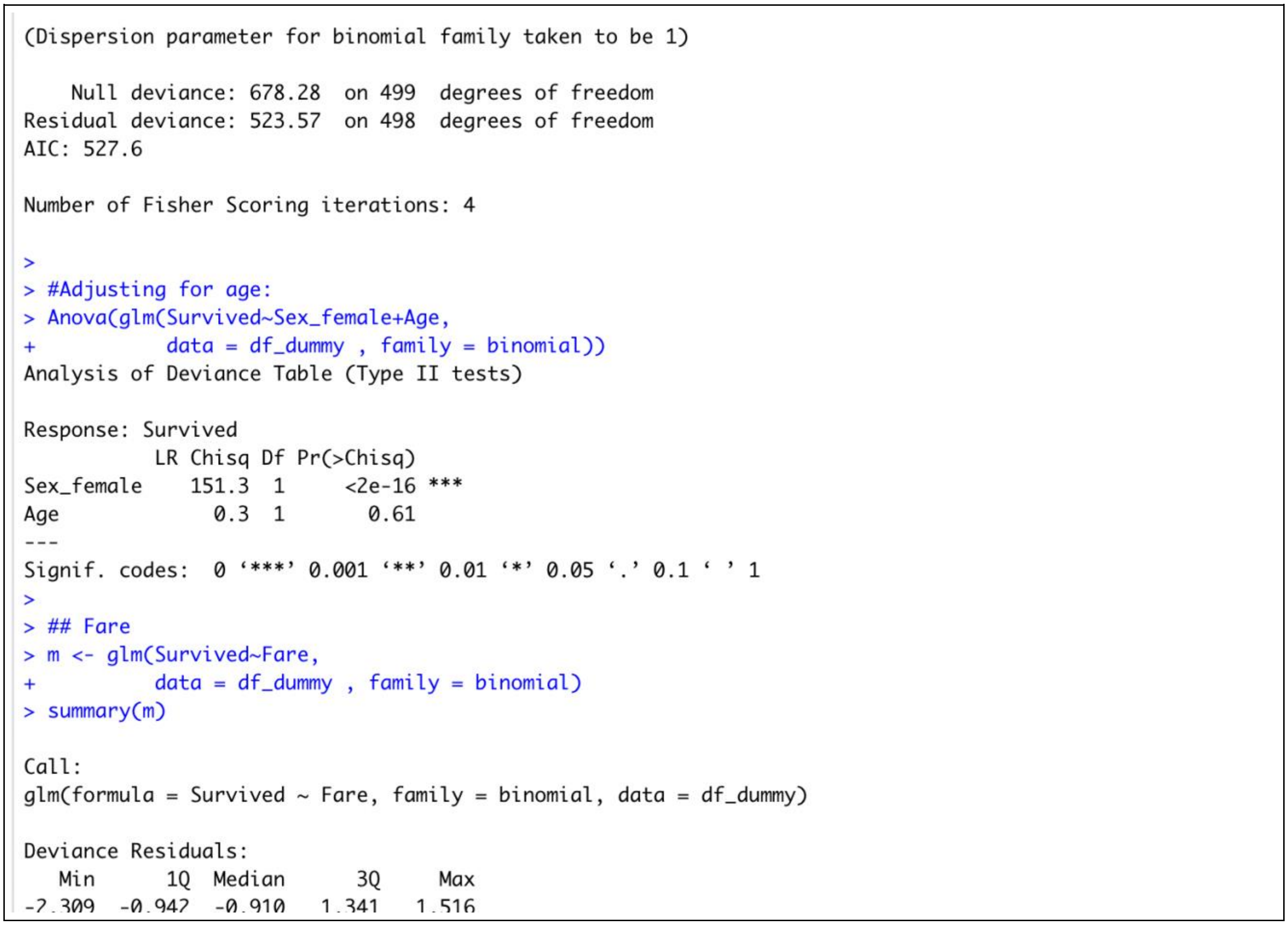

Box plot - Test Scores
Age
80
60
40
20
0
Survived
0.25
0.50
0.75
Sex
female
male


Bar plot

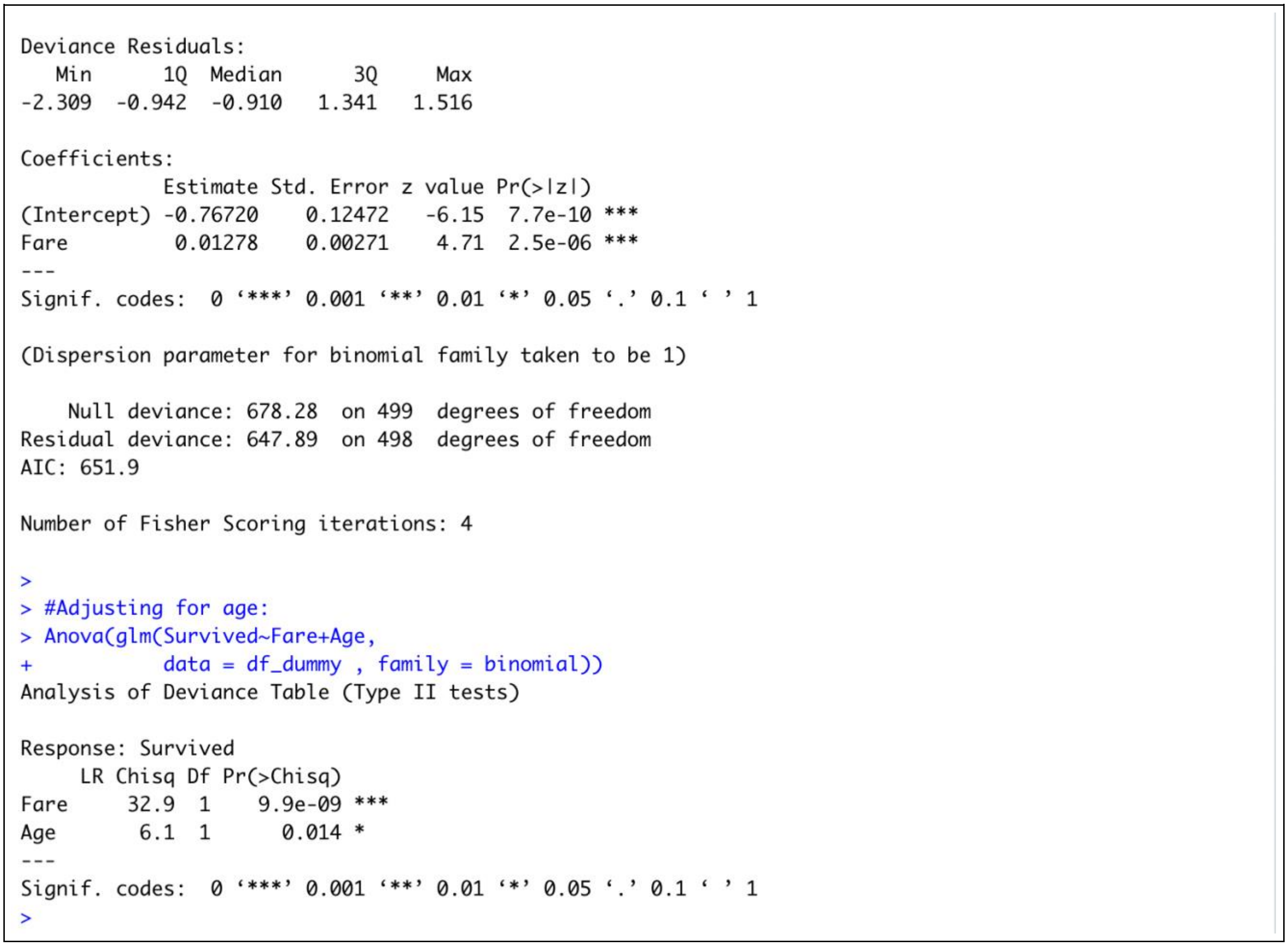
Bar plot



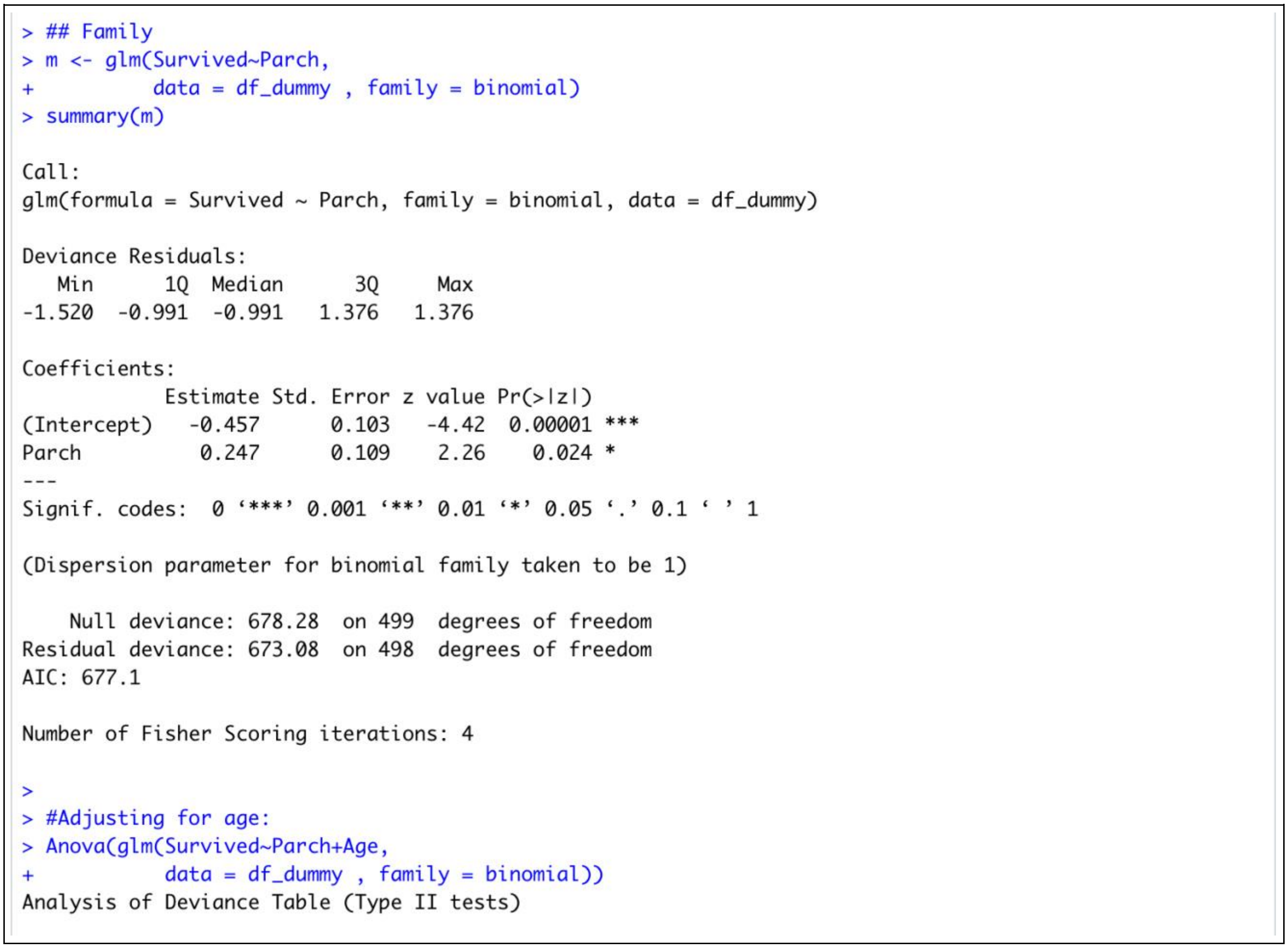
```
Call:
glm(formula = Survived ~ Pclass, family = binomial, data = df_dummy)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
 -1.45   -0.79   -0.79    1.01    1.62

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.437      0.268    5.36  8.2e-08 ***
Pclass        -0.814      0.115   -7.06  1.7e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 678.28  on 499  degrees of freedom
Residual deviance: 624.75  on 498  degrees of freedom
AIC: 628.7

Number of Fisher Scoring iterations: 4

> #Adjusting for age:
> library(car)
> Anova(glm(Survived~Pclass + Age,
+          data = df_dummy , family = binomial))
Analysis of Deviance Table (Type II tests)

Response: Survived
       LR Chisq Df Pr(>Chisq)
Pclass     77.7  1     < 2e-16 ***
Age        27.8  1 0.00000014 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> ## Sex
> m <- glm(Survived~Sex_female,
+          data = df_dummy , family = binomial)
> summary(m)

Call:
glm(formula = Survived ~ Sex_female, family = binomial, data = df_dummy)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.691  -0.678  -0.678   0.740   1.780

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.355      0.140   -9.66   <2e-16 ***
Sex_female     2.511      0.221   11.36   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



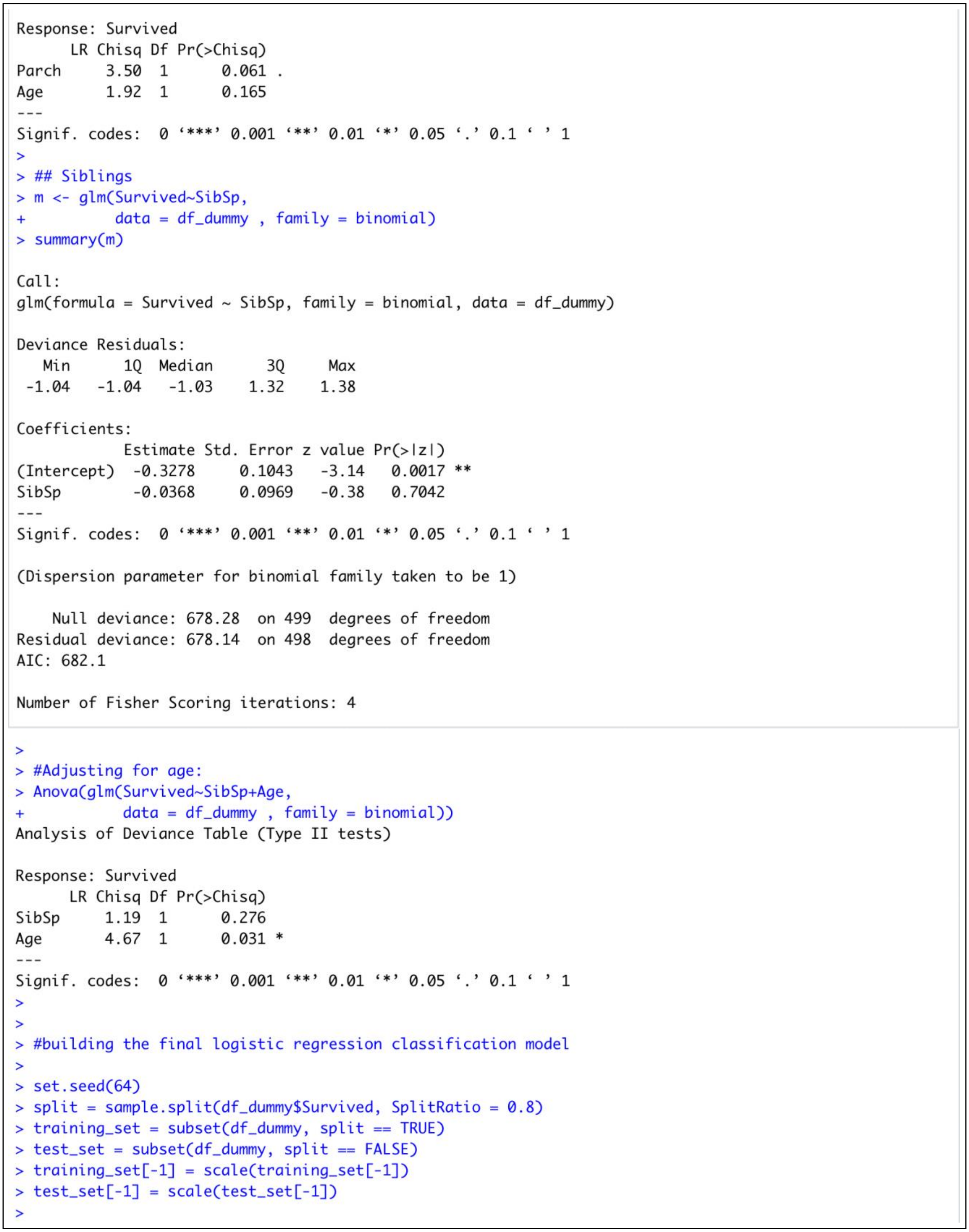
```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 678.28  on 499  degrees of freedom
Residual deviance: 523.57  on 498  degrees of freedom
AIC: 527.6

Number of Fisher Scoring iterations: 4

>
> #Adjusting for age:
> Anova(glm(Survived~Sex_female+Age,
+           data = df_dummy , family = binomial))
Analysis of Deviance Table (Type II tests)

Response: Survived
           LR Chisq Df Pr(>Chisq)
Sex_female    151.3  1     <2e-16 ***
Age             0.3  1       0.61
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> ## Fare
> m <- glm(Survived~Fare,
+          data = df_dummy , family = binomial)
> summary(m)

Call:
glm(formula = Survived ~ Fare, family = binomial, data = df_dummy)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.309  -0.942  -0.910   1.341   1.516
```



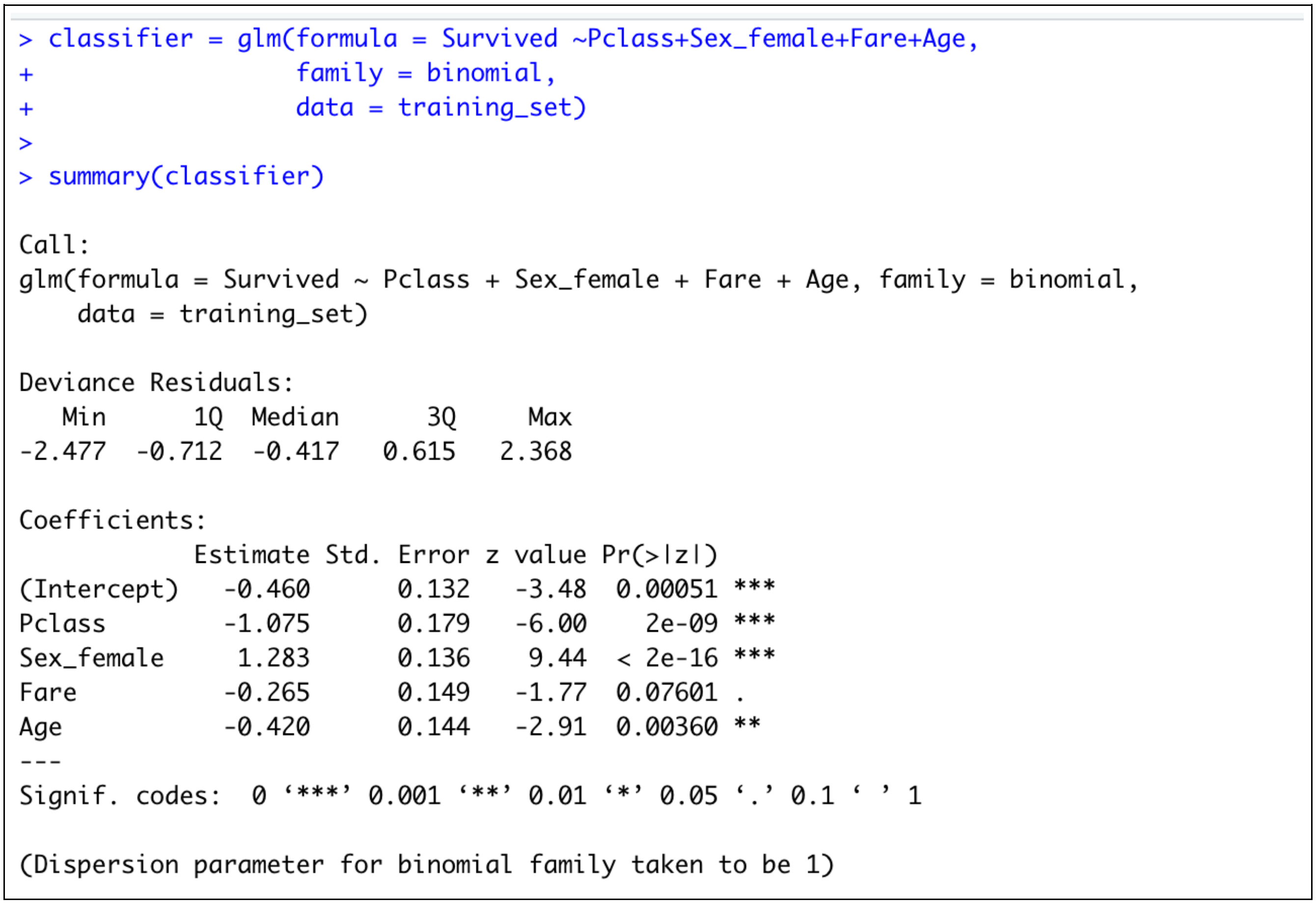
```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.309  -0.942  -0.910   1.341   1.516

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.76720    0.12472   -6.15  7.7e-10 ***
Fare         0.01278    0.00271    4.71  2.5e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 678.28  on 499  degrees of freedom
Residual deviance: 647.89  on 498  degrees of freedom
AIC: 651.9

Number of Fisher Scoring iterations: 4

>
> #Adjusting for age:
> Anova(glm(Survived~Fare+Age,
+           data = df_dummy , family = binomial))
Analysis of Deviance Table (Type II tests)

Response: Survived
     LR Chisq Df Pr(>Chisq)
Fare     32.9  1    9.9e-09 ***
Age       6.1  1      0.014 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```



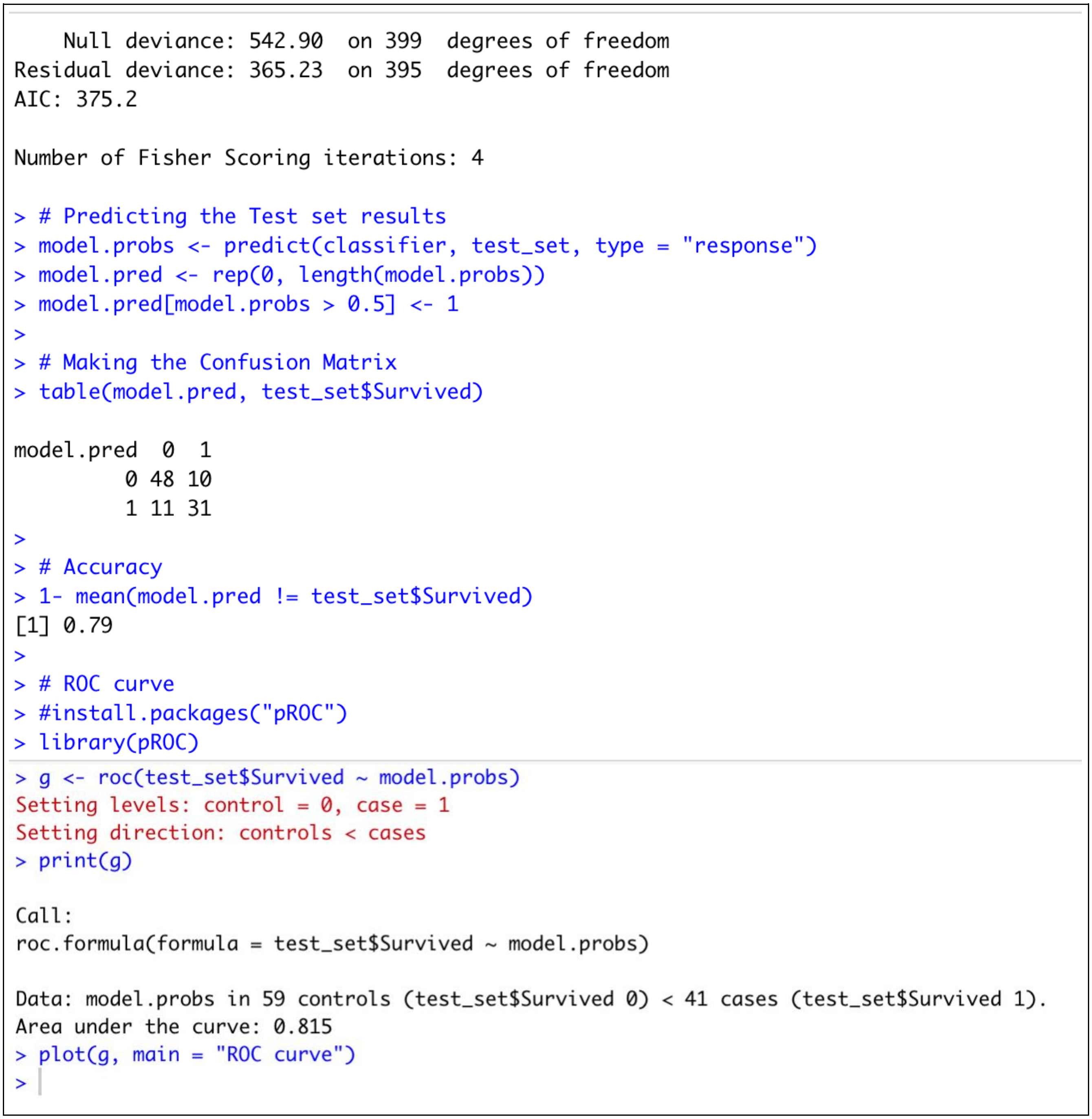
```
> ## Family
> m <- glm(Survived~Parch,
+          data = df_dummy , family = binomial)
> summary(m)

Call:
glm(formula = Survived ~ Parch, family = binomial, data = df_dummy)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.520  -0.991  -0.991   1.376   1.376

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.457      0.103   -4.42  0.00001 ***
Parch          0.247      0.109    2.26    0.024 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 678.28  on 499  degrees of freedom
Residual deviance: 673.08  on 498  degrees of freedom
AIC: 677.1

Number of Fisher Scoring iterations: 4

>
> #Adjusting for age:
> Anova(glm(Survived~Parch+Age,
+           data = df_dummy , family = binomial))
Analysis of Deviance Table (Type II tests)
```



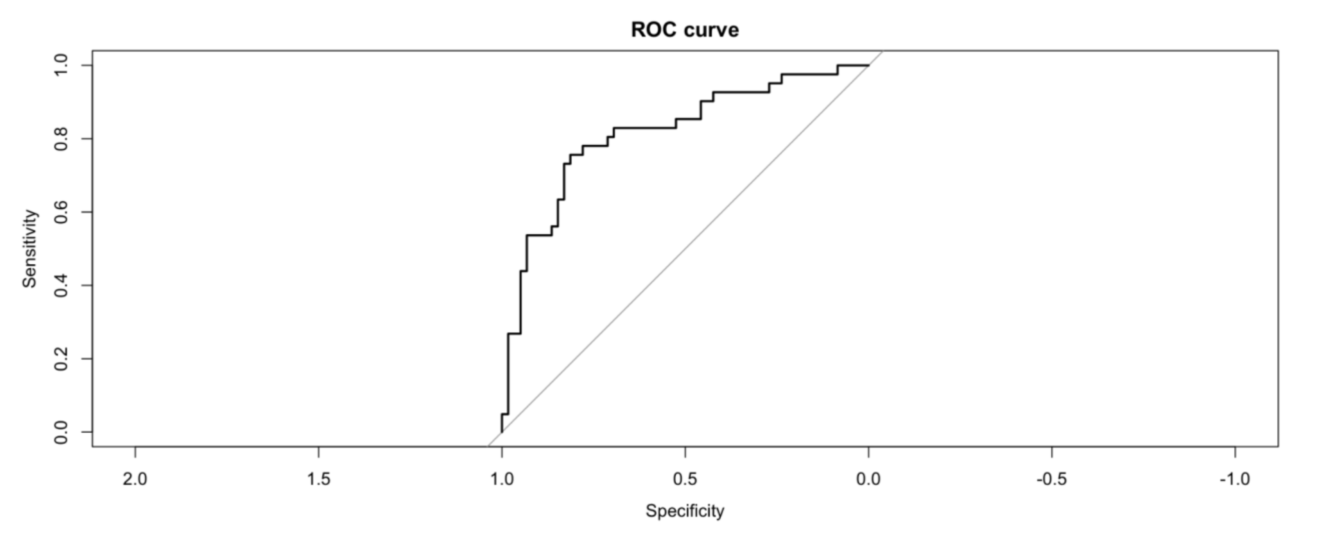
```
Response: Survived
      LR Chisq Df Pr(>Chisq)
Parch     3.50  1      0.061 .
Age       1.92  1      0.165
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> ## Siblings
> m <- glm(Survived~SibSp,
+          data = df_dummy , family = binomial)
> summary(m)

Call:
glm(formula = Survived ~ SibSp, family = binomial, data = df_dummy)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
 -1.04   -1.04   -1.03    1.32    1.38

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3278     0.1043   -3.14   0.0017 **
SibSp        -0.0368     0.0969   -0.38   0.7042
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 678.28  on 499  degrees of freedom
Residual deviance: 678.14  on 498  degrees of freedom
AIC: 682.1

Number of Fisher Scoring iterations: 4

>
> #Adjusting for age:
> Anova(glm(Survived~SibSp+Age,
+           data = df_dummy , family = binomial))
Analysis of Deviance Table (Type II tests)

Response: Survived
      LR Chisq Df Pr(>Chisq)
SibSp     1.19  1      0.276
Age       4.67  1      0.031 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
> #building the final logistic regression classification model
>
> set.seed(64)
> split = sample.split(df_dummy$Survived, SplitRatio = 0.8)
> training_set = subset(df_dummy, split == TRUE)
> test_set = subset(df_dummy, split == FALSE)
> training_set[-1] = scale(training_set[-1])
> test_set[-1] = scale(test_set[-1])
>
```



```
> classifier = glm(formula = Survived ~Pclass+Sex_female+Fare+Age,
+                  family = binomial,
+                  data = training_set)
>
> summary(classifier)

Call:
glm(formula = Survived ~ Pclass + Sex_female + Fare + Age, family = binomial,
    data = training_set)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.477   -0.712   -0.417    0.615    2.368

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.460      0.132   -3.48  0.00051 ***
Pclass        -1.075      0.179   -6.00    2e-09 ***
Sex_female     1.283      0.136    9.44  < 2e-16 ***
Fare          -0.265      0.149   -1.77  0.07601 .
Age           -0.420      0.144   -2.91  0.00360 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```



```
    Null deviance: 542.90  on 399  degrees of freedom
Residual deviance: 365.23  on 395  degrees of freedom
AIC: 375.2

Number of Fisher Scoring iterations: 4

> # Predicting the Test set results
> model.probs <- predict(classifier, test_set, type = "response")
> model.pred <- rep(0, length(model.probs))
> model.pred[model.probs > 0.5] <- 1
>
> # Making the Confusion Matrix
> table(model.pred, test_set$Survived)

model.pred  0  1
         0 48 10
         1 11 31
>
> # Accuracy
> 1- mean(model.pred != test_set$Survived)
[1] 0.79
>
> # ROC curve
> #install.packages("pROC")
> library(pROC)
> g <- roc(test_set$Survived ~ model.probs)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
> print(g)

Call:
roc.formula(formula = test_set$Survived ~ model.probs)

Data: model.probs in 59 controls (test_set$Survived 0) < 41 cases (test_set$Survived 1).
Area under the curve: 0.815
> plot(g, main = "ROC curve")
>
```



6. State Your Conclusion (no more than 100 words)

State the conclusion so that a non-statistician can understand.

We conclude that Passenger Class, gender, fare were the significant parameters in survival. Initially, the female gender parameter showed some sort of correlation with passenger survival. These parameters were: Pclass, gender, age, Fare and # of family members. ANCOVA was performed, adjusting for age on such parameters, and finally it was found that age wasn't a significant predictor if gender is the predictor. And, finally, Pclass, gender, and fare were the only significant predictors, once adjusted for age.

Based on a logistic regression model, using Pclass, gender, Fare as its features, we could predict the survival with an accuracy of 79%. Another assessment of this model, the ROC curve, showed good results as the area under the ROC curve was 0.815.

Solution Submission

1. **Fill up this word file and upload it.**
2. **Upload your data set. This is the data set after cleaning (a small CSV file)**

3. **Upload your R file as a file with the name “mini-project-solution.R”**

Grading will be done based on

1. **The originality of selected data set and data analysis approach**
2. **Data Preparation set and cleanup**
3. **General Correctness of data analysis**
4. **Quality of your R code and output results**
5. **Correct final conclusion**