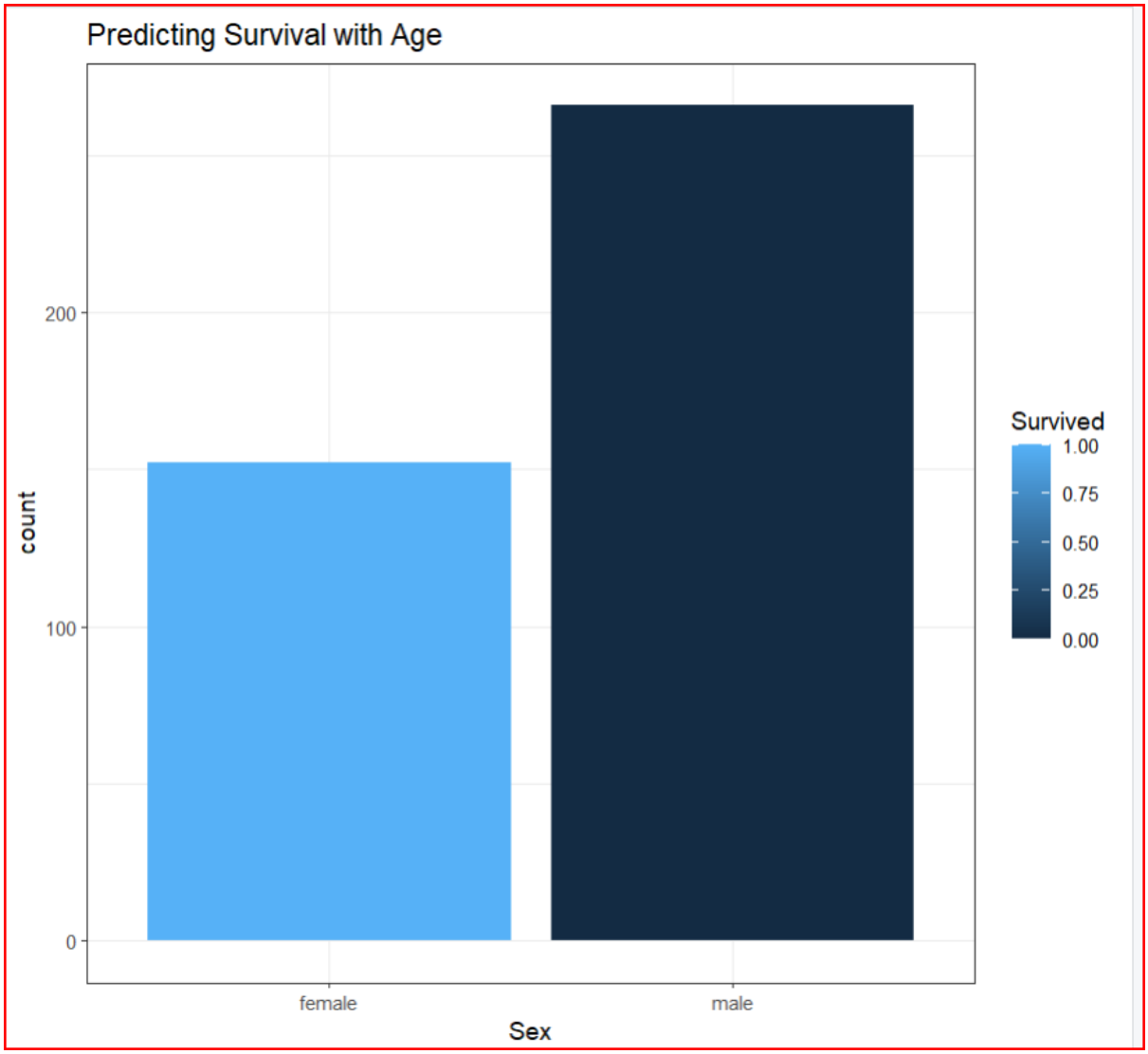
TASK 1: TITANIC SURVIVAL PREDICTION

```r
#loading necessary libraries
library(tidyverse)
library(dplyr)
library(ggplot2)
library(caTools)

#Read the Titanic dataset
tested<-read.csv("internship tasks/tested.csv")
```

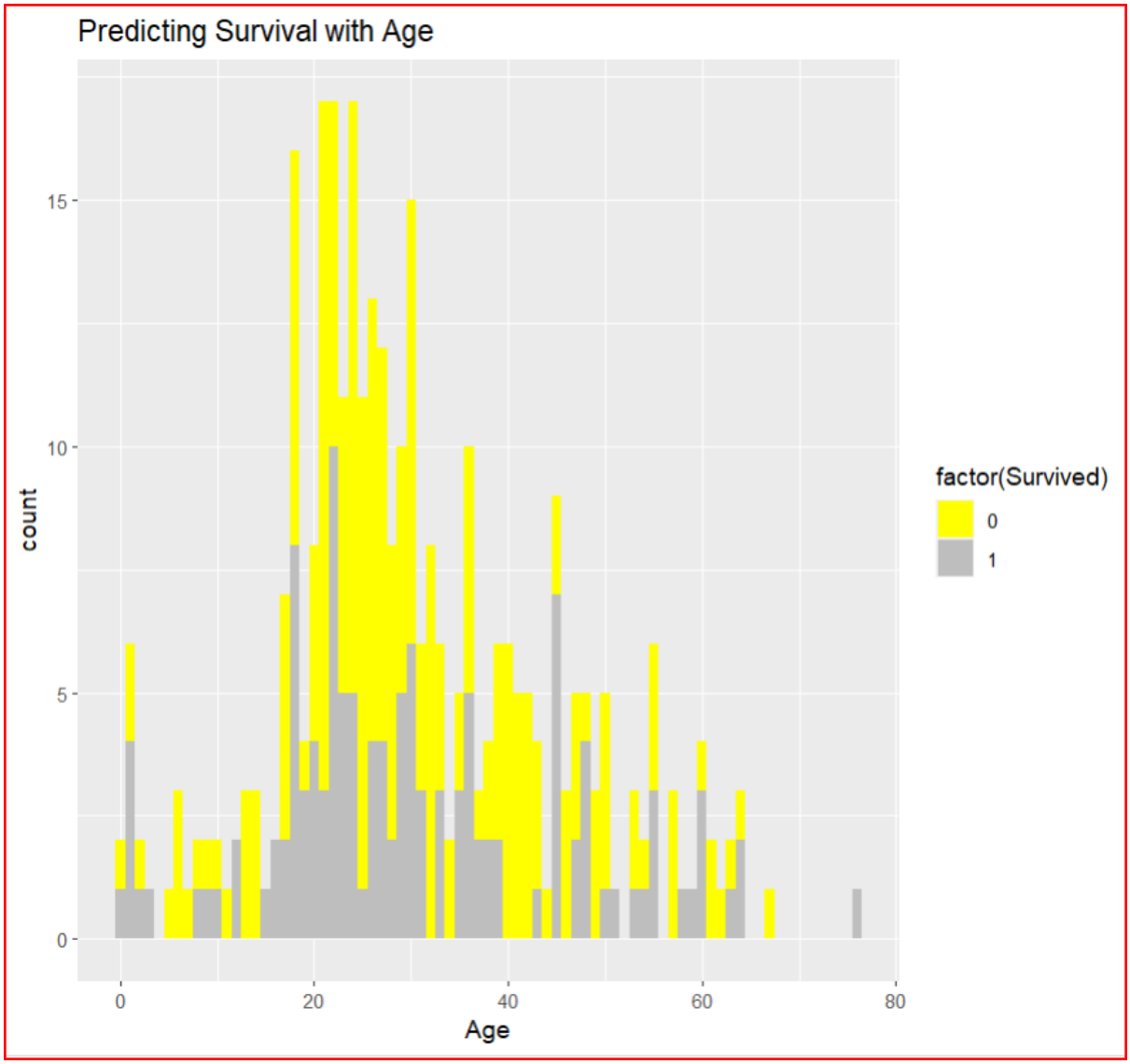
```
> #Check for any missing values
> any(is.na(tested))
[1] TRUE
>
> #Display the structure of the dataset
> str(tested)
'data.frame':   418 obs. of  12 variables:
 $ PassengerId: int  892 893 894 895 896 897 898 899 900 901 ...
 $ Survived   : int  0 1 0 0 1 0 1 0 1 0 ...
 $ Pclass     : int  3 3 2 3 3 3 3 2 3 3 ...
 $ Name       : chr  "Kelly, Mr. James" "Wilkes, Mrs. James (Ellen Needs)" "Myles, Mr. Thomas
Francis" "Wirz, Mr. Albert" ...
 $ 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" ...
> #Explore the data in the tested data frame
> dim(tested)
[1] 418  12
> head(tested)
  PassengerId Survived Pclass                                          Name    Sex  Age
1         892        0      3                              Kelly, Mr. James   male 34.5
2         893        1      3              Wilkes, Mrs. James (Ellen Needs) female 47.0
3         894        0      2                     Myles, Mr. Thomas Francis   male 62.0
4         895        0      3                              Wirz, Mr. Albert   male 27.0
5         896        1      3 Hirvonen, Mrs. Alexander (Helga E Lindqvist) female 22.0
6         897        0      3                    Svensson, Mr. Johan Cervin   male 14.0
  SibSp Parch  Ticket    Fare Cabin Embarked
1     0     0  330911  7.8292              Q
2     1     0  363272  7.0000              S
3     0     0  240276  9.6875              Q
4     0     0  315154  8.6625              S
5     1     1 3101298 12.2875              S
6     0     0    7538  9.2250              S
```

```
> tail(tested)
    PassengerId Survived Pclass                           Name    Sex  Age SibSp Parch
413        1304        1      3 Henriksson, Miss. Jenny Lovisa female 28.0     0     0
414        1305        0      3             Spector, Mr. Woolf   male   NA     0     0
415        1306        1      1   Oliva y Ocana, Dona. Fermina female 39.0     0     0
416        1307        0      3   Saether, Mr. Simon Sivertsen   male 38.5     0     0
417        1308        0      3            Ware, Mr. Frederick   male   NA     0     0
418        1309        0      3       Peter, Master. Michael J   male   NA     1     1
                Ticket     Fare Cabin Embarked
413             347086   7.7750              S
414          A.5. 3236   8.0500              S
415           PC 17758 108.9000  C105        C
416 SOTON/O.Q. 3101262   7.2500              S
417             359309   8.0500              S
418               2668  22.3583              C
> #Removing the 11th row from dataset which is Cabin having missing values
> tested<-tested[-11]
> tail(tested)
    PassengerId Survived Pclass                           Name    Sex  Age SibSp Parch
413        1304        1      3 Henriksson, Miss. Jenny Lovisa female 28.0     0     0
414        1305        0      3             Spector, Mr. Woolf   male   NA     0     0
415        1306        1      1   Oliva y Ocana, Dona. Fermina female 39.0     0     0
416        1307        0      3   Saether, Mr. Simon Sivertsen   male 38.5     0     0
417        1308        0      3            Ware, Mr. Frederick   male   NA     0     0
418        1309        0      3       Peter, Master. Michael J   male   NA     1     1
                Ticket     Fare Embarked
413             347086   7.7750        S
414          A.5. 3236   8.0500        S
415           PC 17758 108.9000        C
416 SOTON/O.Q. 3101262   7.2500        S
417             359309   8.0500        S
418               2668  22.3583        C
> #Count the missing values
> sum(is.na(tested))
[1] 87
> sum(is.na(tested$Age))
[1] 86
> colSums(is.na(tested))
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp
          0           0           0           0           0          86           0
```

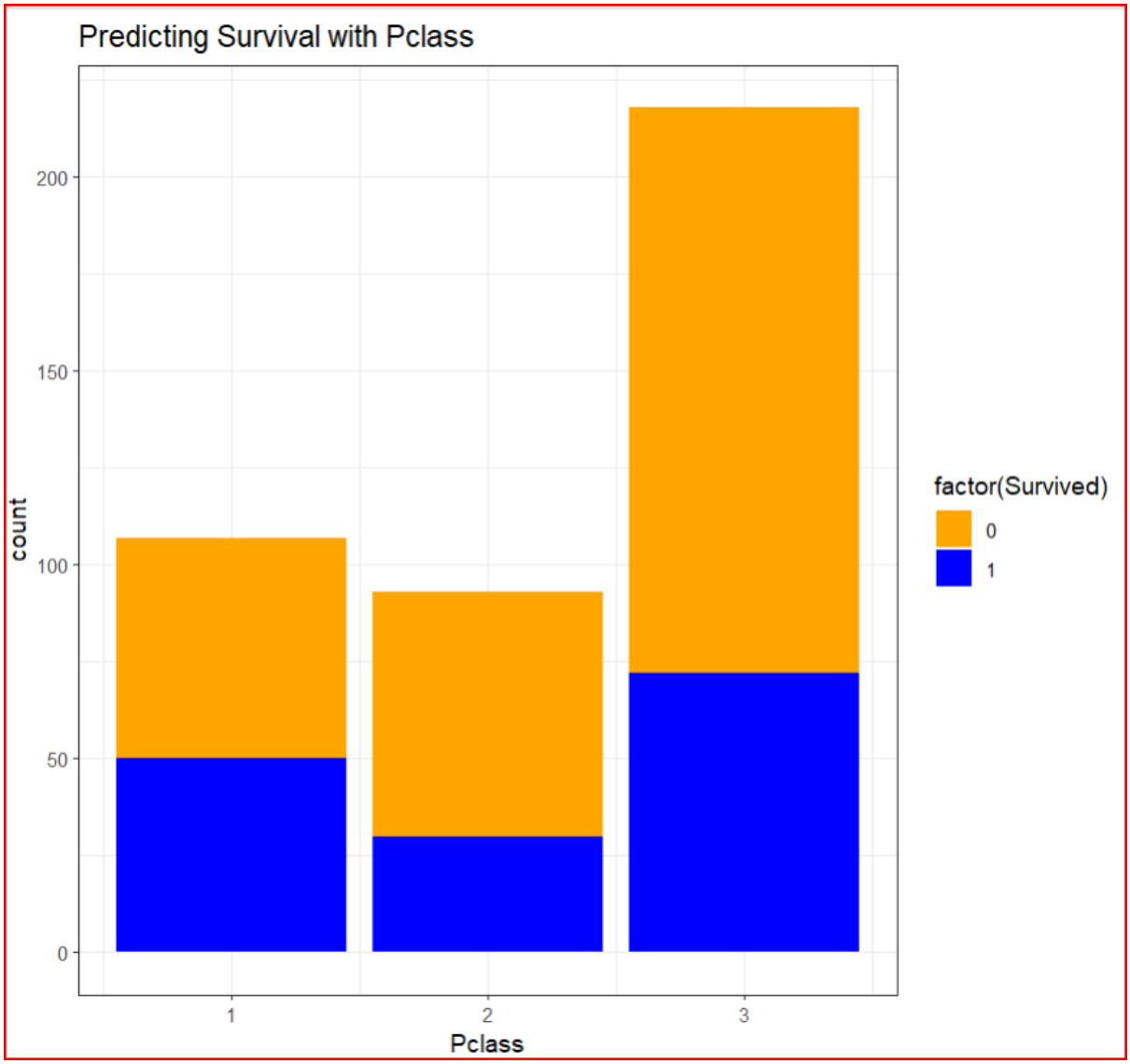
```
> colSums(is.na(tested))
PassengerId    Survived      Pclass        Name         Sex         Age       SibSp
          0           0           0           0           0          86           0
      Parch      Ticket        Fare    Embarked
          0           0           1           0
> #Remove rows with missing values
> tested.clean<-na.omit(tested)
> nrow(tested.clean)
[1] 331
> #Finding the summary of the data
> table(tested.clean$Survived)

  0   1
204 127
> table(tested.clean$Sex)

female   male
   127    204
> summary(tested.clean$Survived)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3837  1.0000  1.0000
> names(tested.clean)
 [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"         "Age"
 [7] "SibSp"       "Parch"       "Ticket"      "Fare"        "Embarked"
> var(tested.clean$Fare)
[1] 3748.936
> sd(tested.clean$Fare)
[1] 61.22856
> #Predicting Survival with Sex
> ggplot(tested, aes(x = Sex, fill = Survived)) +
+   geom_bar() +
+   #scale_fill_manual(values = c("blue", "orange"))
+   theme_bw() +
+   labs(x = "Sex", title = "Predicting Survival with Age")
>
```


Predicting Survival with Age
count
200
100
0
female
male
Sex
Survived
1.00
0.75
0.50
0.25
0.00

```
> #Predicting Survival with Age
> ggplot(tested, aes(x = Age, fill = factor(Survived))) +
+   geom_histogram(binwidth = 1) +
+   scale_fill_manual(values = c("yellow", "gray")) +
+   labs(x = "Age", title = "Predicting Survival with Age")
```

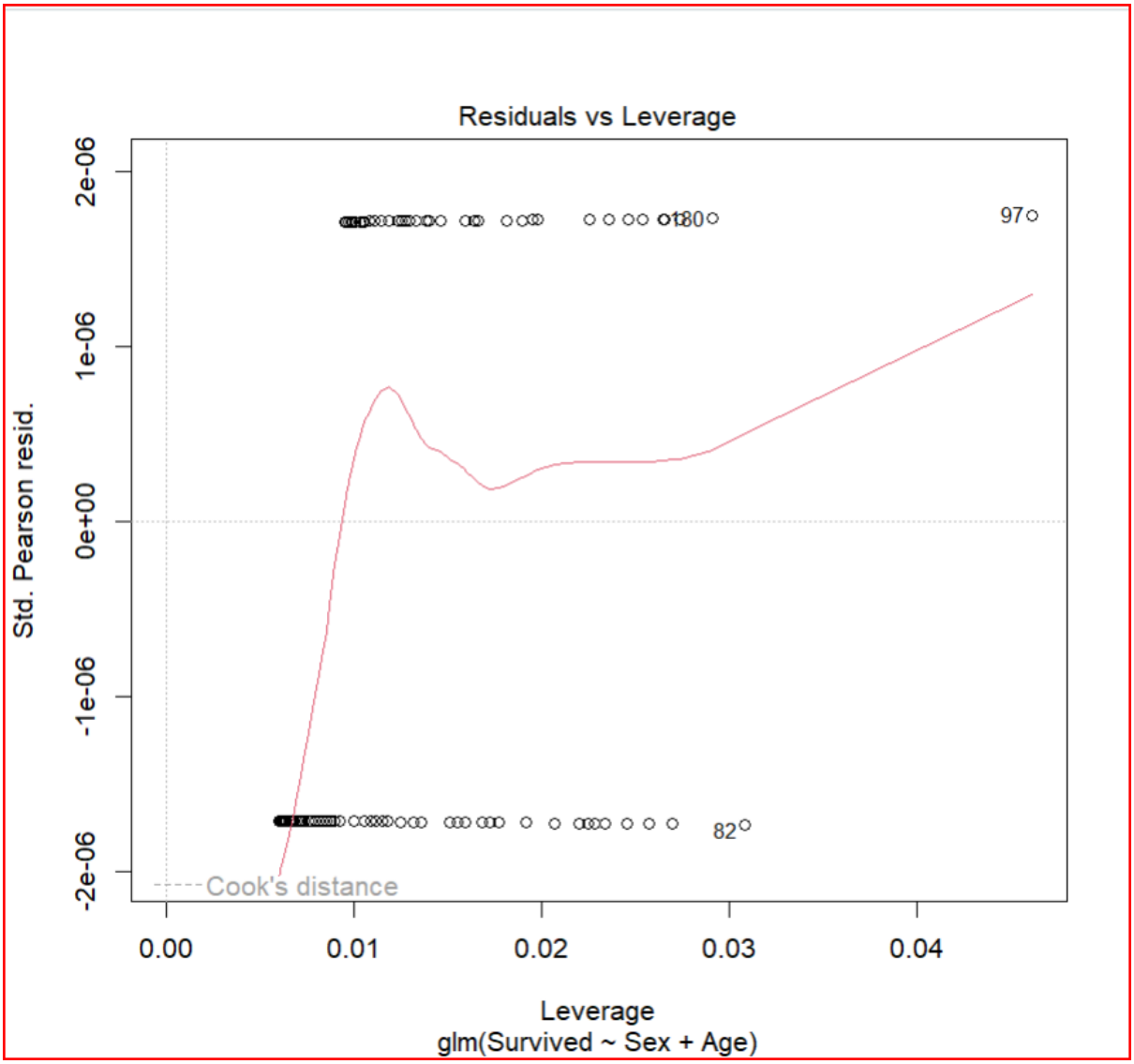


```
> #Predicting Survival with Pclass
> ggplot(tested, aes(x = Pclass, fill = factor(Survived))) +
+   geom_bar(binwidth = 1) +
+   scale_fill_manual(values = c("orange", "blue"))+
+   theme_bw() +
+   labs(x = "Pclass", title = "Predicting Survival with Pclass")
```

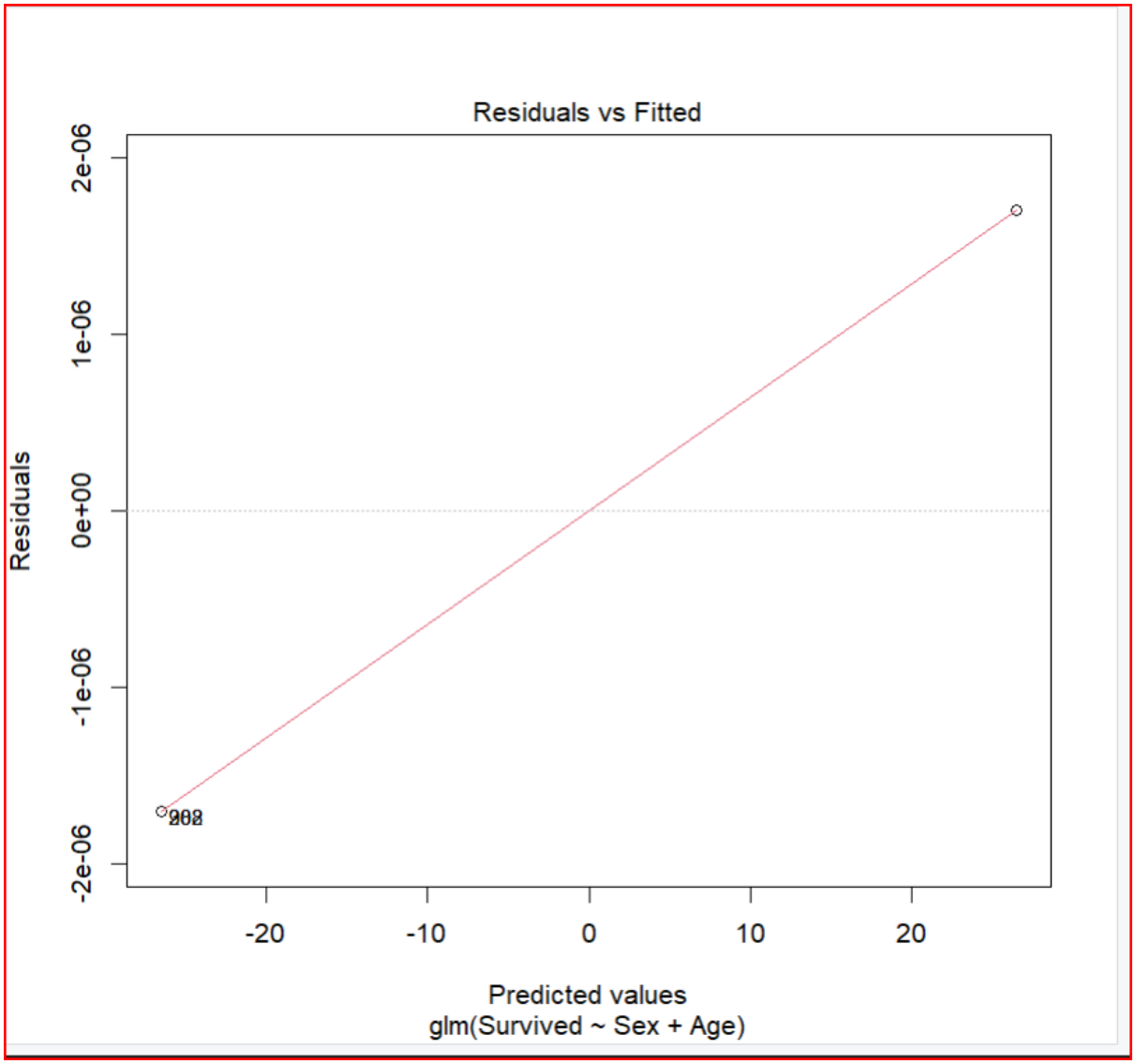


```
> #Model the data to train and test data
> set.seed(123)
> data_sample = sample.split(tested$Survived, SplitRatio=0.80)
> train_data = subset(tested,data_sample==TRUE)
> test_data = subset(tested,data_sample==FALSE)
> dim(train_data)
[1] 335  11
> dim(test_data)
[1] 83 11
> #Using Logistic Regression Model
> Logistic_Model <- glm(Survived~ Sex + Age,train_data,family = binomial())
Warning message:
glm.fit: algorithm did not converge
> summary(Logistic_Model)

Call:
glm(formula = Survived ~ Sex + Age, family = binomial(), data = train_data)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.409e-06  -2.409e-06  -2.409e-06   2.409e-06   2.409e-06

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.657e+01  5.962e+04   0.000    1.000
Sexmale     -5.313e+01  4.519e+04  -0.001    0.999
Age         -1.748e-13  1.551e+03   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3.5031e+02  on 263  degrees of freedom
Residual deviance: 1.5316e-09  on 261  degrees of freedom
  (71 observations deleted due to missingness)
AIC: 6

Number of Fisher Scoring iterations: 25
```

```
> plot(Logistic_Model)
```



```
Hit <Return> to see next plot: prediction <- predict(Logistic_Model,train_data,type = "respon
se", probability = TRUE)
```



```
summary(prediction)
```