Titanic passenger survival prediction using classification models

Bin Han

2025-07-08

Table of contents

1 Project description

This project aims to develop a predictive model to estimate the survival chances of Titanic passengers and to identify the key factors that influenced their outcomes. Using logistic regression on a cleaned version of the Titanic dataset, we examined the relationship between survival and various passenger characteristics.

A correlation analysis of numerical variables revealed that survival was positively associated with ticket fare and the number of parents or children aboard, while it was negatively

associated with passenger class and age. Passenger ID and the number of siblings/spouses aboard showed little to no correlation with survival.

After stepwise feature selection, the final logistic regression model included four key predictors: passenger class (Pclass), sex (Sex), age (Age), and number of siblings/spouses aboard (SibSp). All were statistically significant, with males, older passengers, and those in lower classes having a significantly lower chance of survival. The model demonstrated good performance, **achieving an accuracy of 80.9%**.

Additionally, we simulated predictions for 12 of the most common passenger profiles. **Results consistently showed that female passengers had a much higher survival probability, and survival likelihood decreased with increasing age.** This model provides a clear and interpretable view of the main factors affecting survival on the Titanic.

2 Titanic data analysis

Our analysis is to develop a model to determine the survival chances of Titanic passengers and identify the key factors influecing their outcomes.

2.1 Loading data and R packages

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to be
```

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
library(car)
```

```
Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

```r
data<- read.csv("titanic.csv")
```

3 Exploratory data analysis

3.1 Data overview

```r
glimpse(data)
```

```
Rows: 891
Columns: 12
$ PassengerId <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,~
$ Survived    <int> 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1~
$ Pclass      <int> 3, 1, 3, 1, 3, 3, 1, 3, 3, 2, 3, 1, 3, 3, 3, 2, 3, 2, 3, 3~
$ Name        <chr> "Braund, Mr. Owen Harris", "Cumings, Mrs. John Bradley (Fl~
$ Sex         <chr> "male", "female", "female", "female", "male", "male", "mal~
$ Age         <dbl> 22, 38, 26, 35, 35, NA, 54, 2, 27, 14, 4, 58, 20, 39, 14, ~
$ SibSp       <int> 1, 1, 0, 1, 0, 0, 0, 3, 0, 1, 1, 0, 0, 1, 0, 0, 4, 0, 1, 0~
$ Parch       <int> 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 1, 0, 0, 5, 0, 0, 1, 0, 0, 0~
$ Ticket      <chr> "A/5 21171", "PC 17599", "STON/O2. 3101282", "113803", "37~
$ Fare        <dbl> 7.2500, 71.2833, 7.9250, 53.1000, 8.0500, 8.4583, 51.8625,~
$ Cabin       <chr> "", "C85", "", "C123", "", "", "E46", "", "", "", "G6", "C~
$ Embarked    <chr> "S", "C", "S", "S", "S", "Q", "S", "S", "S", "C", "S", "S"~
```

```r
summary(data)
```

```
  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
```

Using count function to know the distribution of the character variables.

```
data|>
  count(Survived, name="count")
```

```
  Survived count
1        0   549
2        1   342
```

In our dataset, we have 342 Survived cases, compared to 549 dead cases.

```
data|>
  count(Pclass, name="count")
```

```
  Pclass count
1      1   216
2      2   184
3      3   491
```

In our data set, we have 216 passengers with class 1, 184 passengers with Class 2, 491 passengers with Class 3.

```
data|>
  count(Embarked, name="Count")
```

```
  Embarked Count
1              2
2        C   168
3        Q    77
4        S   644
```

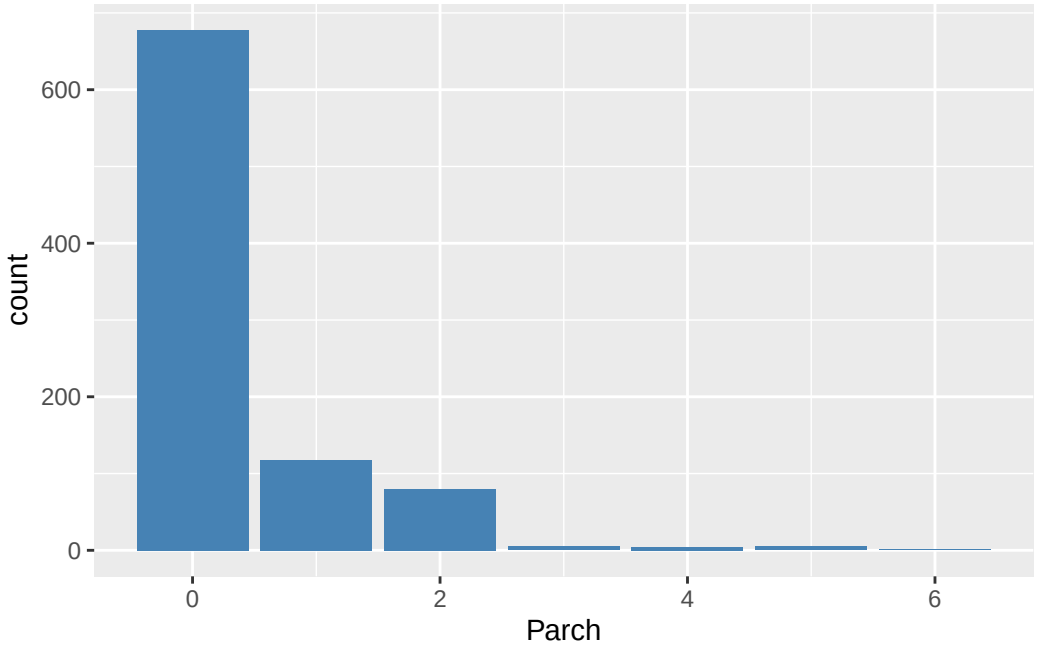
In the data set, there are 644 passengers boarded the ship at Southampton. 77 passengers board on the ship at Queenstown, 168 passengers boarded on the ship at Cherbourg.

```
data|>
  count(SibSp, name="count")
```

```
  SibSp count
1     0   608
2     1   209
3     2    28
4     3    16
5     4    18
6     5     5
7     8     7
```

From the table, we can see 608 passengers on board have no siblings and spouse aboard, 209 passengers on board have one siblings or wife aboard. However, 5 passengers on board have 5 siblings or spouse aboard. 8 passengers on board have 7 siblings or spouse aboard.

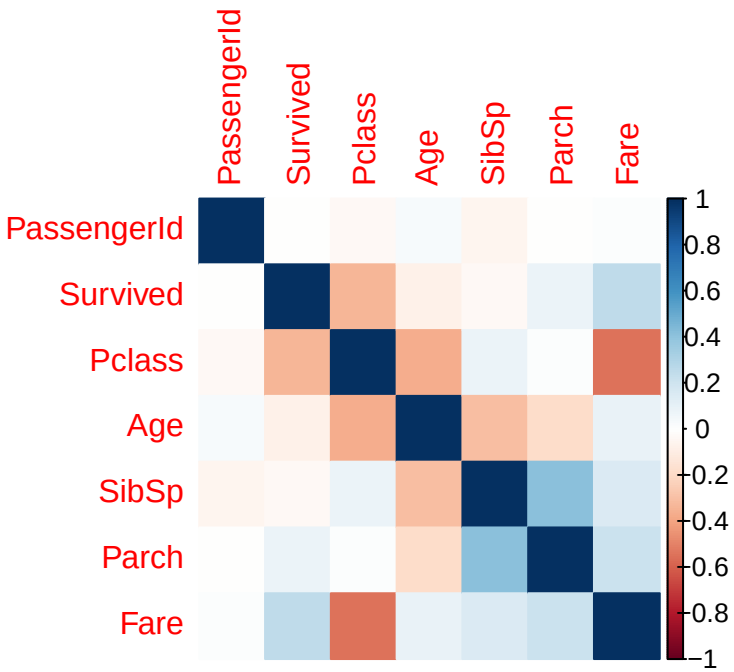
```
data%>%
  ggplot(aes(x=Parch))+
  geom_bar(fill="steelblue")
```



From the bar plot, we can see most passengers have no parents/children Aboard. Very few passengers have 3, 4 and 5 parents/Children Aboard.

3.2 Compute correlations

```
data %>%
  select((where(is.numeric)))%>%
  cor(,use="pairwise.complete.obs")%>%
  corrplot(method="color")
```



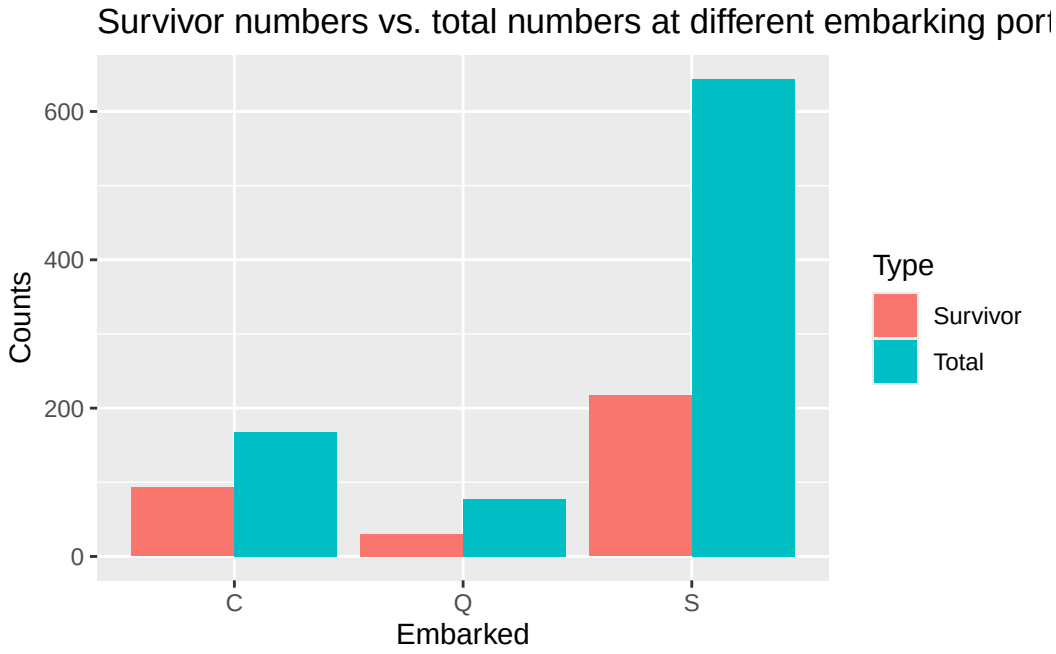
From the correlation plot of numerical variable, we can see the survived rate has propositional relation to the ticket Fare, the number of parents/children aboard. and the survived chances has inverse propositional relation to the Passenger class and the age of passenger. It seems has no clear relation with the passenger ID and the number of siblings/spouses aboard.

3.3 Boxplots and bar plots

First convert the character variable to factor variable.

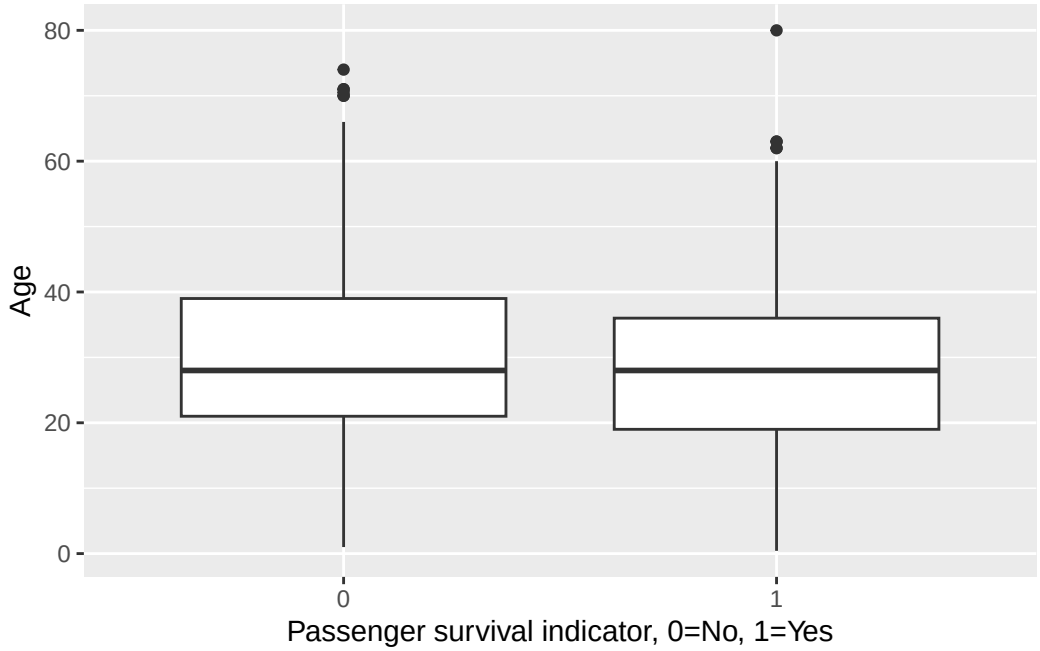
```
data%>%
  mutate(across(where(is.character), as.factor))
```

```
data%>%
  filter(Embarked !="")%>%
  group_by(Embarked) %>%
  summarise(Survivor=sum(Survived==1),Total=n(), Rate=Survivor/Total)%>%
  pivot_longer(cols=c(Survivor, Total), names_to = "Type", values_to="Counts")%>%
  ggplot(aes(x=Embarked, y=Counts, fill=Type))+
  geom_col(position="dodge")+
  labs(title= "Survivor numbers vs. total numbers at different embarking port")
```



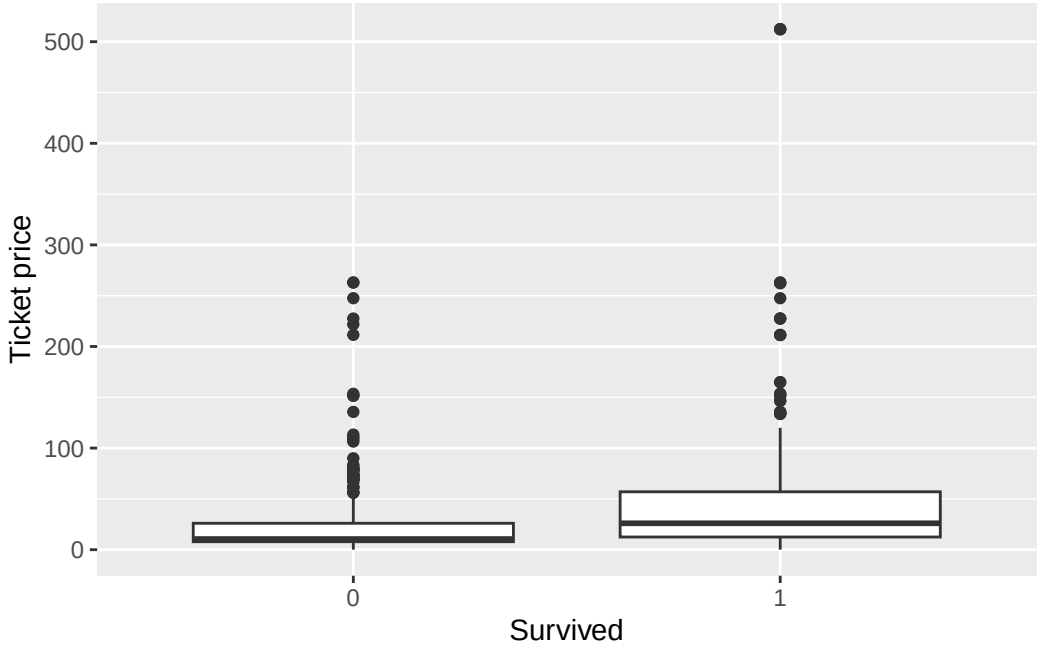
From the computation, we see the passengers from Embarked port at Cherbourg have Survivor Rate 55.4%, compared to 39.0% for Queenstown and 33.7% for Southampton.

```
ggplot(data, aes(x=as.character(Survived), y=Age))+
  geom_boxplot(na.rm=TRUE)+
  labs(x="Passenger survival indicator, 0=No, 1=Yes")
```



Fro the box plot, we see there is no big age mean difference for the survivors and dead people. Survivor ages are slightly younger than the dead group.

```
ggplot(data, aes(x=as.character(Survived), y=Fare))+
  geom_boxplot()+
  labs(x="Survived",
       y="Ticket price")
```

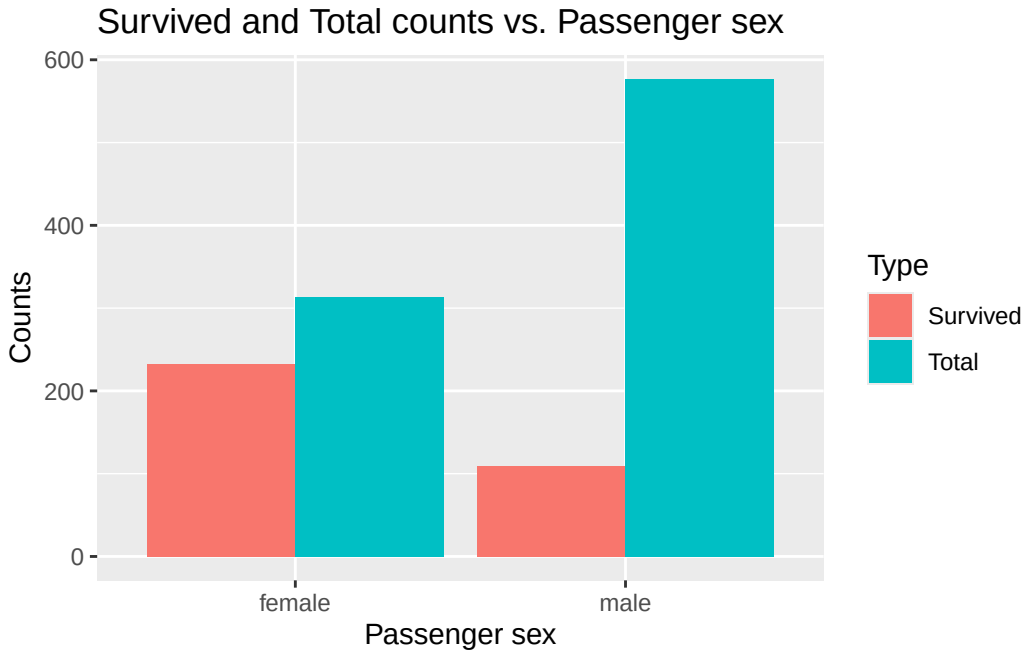


From the box plot, the Survived group has more expensive ticket than the death group.

```
data%>%
  group_by(Pclass)%>%
  summarise(Survived=sum(Survived==1), Total=n(), Rate=Survived/Total)%>%
  pivot_longer(cols=c(Survived, Total), names_to="Type", values_to="Count")%>%
  ggplot(aes(x=Pclass, y=Count, fill=Type))+
  geom_col(position="dodge")+
  labs(title="Passenger class vs. Survived and Total counts")
```

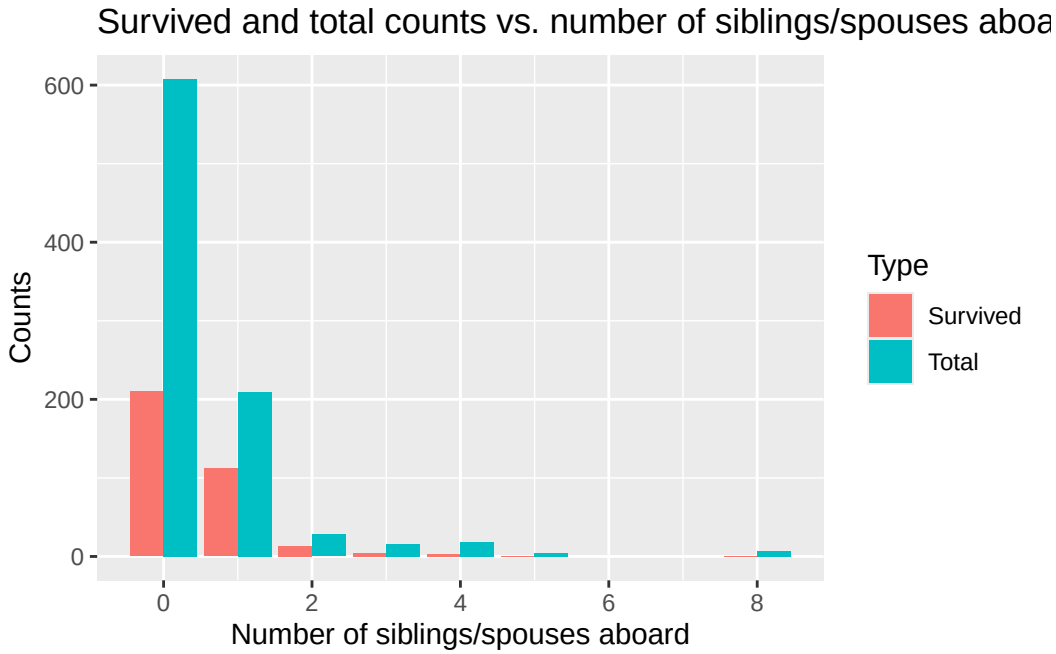
From the bar chart, the 1st passenger class has 63.0% survivor rate, compared to the 47.3% for 2nd class, 24.2% for 3rd class.

```
data%>%
  group_by(Sex) %>%
  summarise(Survived=sum(Survived==1),Total=n(), Rate=Survived/Total) %>%
  pivot_longer(cols=c(Survived, Total), names_to="Type", values_to="Counts")%>%
  ggplot(aes(x=Sex, y=Counts,fill=Type))+
  geom_col(position="dodge")+
  labs(title="Survived and Total counts vs. Passenger sex",
       x="Passenger sex")
```



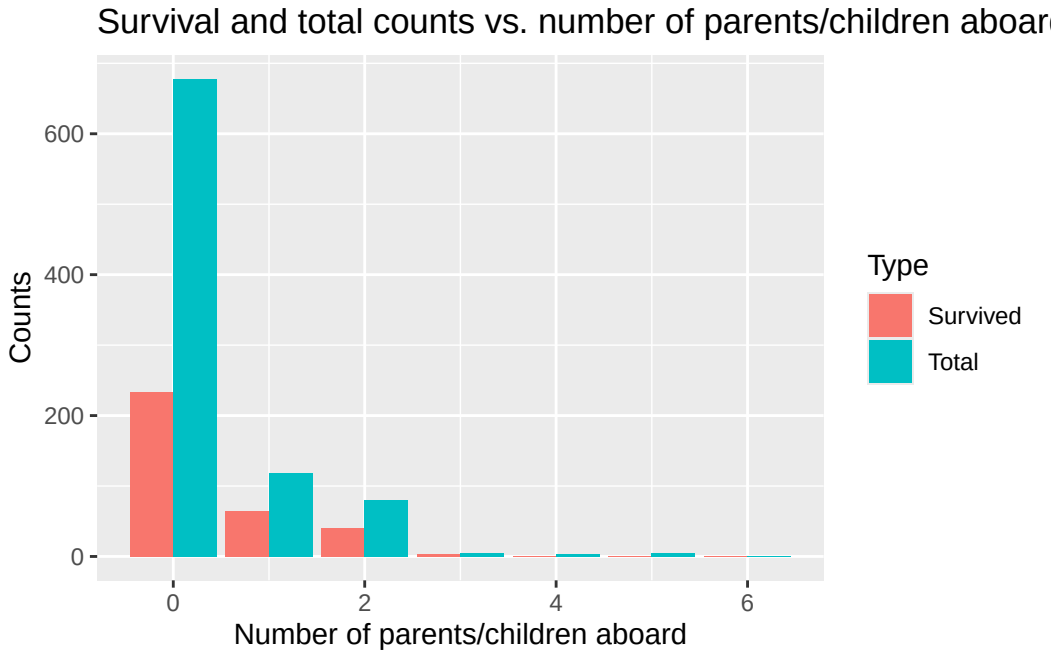
From the computation, the female survival rate is 74.2%, compared to 18.9% for males. This means women had four times the chance of being saved compared to men.

```
data%>%
  group_by(SibSp) %>%
  summarise(Survived=sum(Survived==1), Total= n(), Rate=Survived/Total)%>%
  pivot_longer(cols=c(Survived, Total), names_to="Type", values_to="Counts")%>%
  ggplot(aes(x=SibSp, y=Counts, fill=Type))+
  geom_col(position="dodge")+
  labs(title="Survived and total counts vs. number of siblings/spouses aboard",
       x="Number of siblings/spouses aboard")
```



From the computation, the highest survivor rate is 53.6% for 1 siblings/spouses aboard, the second one is 46.4% for 2 siblings/spouses aboard, the third one is 34.5% for non siblings/spouses aboard, the fourth is 25% for 3 siblings/spouse aboard, the fifth is 16.7% for 4 siblings aboard. No survivor rate for 5 and 8 siblings/spouse group.

```
data%>%
  group_by(Parch) %>%
  summarise(Survived=sum(Survived==1), Total=n(), Rate=Survived/Total)%>%
  pivot_longer(cols=c(Survived, Total), names_to="Type", values_to="Counts")%>%
  ggplot(aes(x=Parch, y=Counts, fill=Type))+
  geom_col(position="dodge")+
  labs(x="Number of parents/children aboard",
       title="Survival and total counts vs. number of parents/children aboard")
```



From the computation, the highest survived rate is 60% for 3 parents/children aboard. The second is 55.1% for 1 parents/children aboard. the third is 50% for 2 parents/children aboard. The fourth is 34.4% for non parents/children aboard. the fifth is 20% for 5 parents/children aboard. the other groups have 0% survived rate.

4 Logistic regression

4.1 Initial model

```
data_clean<-data%>%
  filter(Embarked!="")%>%
  select(-PassengerId, -Name, -Ticket, -Cabin)
model1<-glm(Survived~., data=data_clean, family=binomial)
summary(model1)
```

```
Call:
glm(formula = Survived ~ ., family = binomial, data = data_clean)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.637407   0.634550   8.884  < 2e-16 ***
Pclass      -1.199251   0.164619  -7.285 3.22e-13 ***
```

```
Sexmale     -2.638476   0.222256 -11.871  < 2e-16 ***
Age         -0.043350   0.008232  -5.266 1.39e-07 ***
SibSp       -0.363208   0.129017  -2.815  0.00487 **
Parch       -0.060270   0.123900  -0.486  0.62666
Fare         0.001432   0.002531   0.566  0.57165
EmbarkedQ   -0.823545   0.600229  -1.372  0.17005
EmbarkedS   -0.401213   0.270283  -1.484  0.13770
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 960.90  on 711  degrees of freedom
Residual deviance: 632.34  on 703  degrees of freedom
  (177 observations deleted due to missingness)
AIC: 650.34

Number of Fisher Scoring iterations: 5
```

4.2 VIF and multicollinearity

```
vif(model1)
```

```
             GVIF Df GVIF^(1/(2*Df))
Pclass   1.887735  1        1.373949
Sex      1.191874  1        1.091730
Age      1.424118  1        1.193364
SibSp    1.274012  1        1.128721
Parch    1.274248  1        1.128826
Fare     1.538521  1        1.240371
Embarked 1.139441  2        1.033173
```

4.3 Further improving our model

```
model2=step(model1, direction="both", trace=1)
```

```
Start:  AIC=650.34
Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked

           Df Deviance    AIC
- Parch     1   632.58 648.58
```

```
- Fare      1   632.68 648.68
- Embarked  2   635.30 649.30
<none>          632.34 650.34
- SibSp     1   640.91 656.91
- Age       1   662.75 678.75
- Pclass    1   686.64 702.64
- Sex       1   808.42 824.42

Step:  AIC=648.58
Survived ~ Pclass + Sex + Age + SibSp + Fare + Embarked

           Df Deviance    AIC
- Fare      1   632.82 646.82
- Embarked  2   635.54 647.54
<none>          632.58 648.58
+ Parch     1   632.34 650.34
- SibSp     1   642.73 656.73
- Age       1   662.96 676.96
- Pclass    1   689.97 703.97
- Sex       1   813.74 827.74

Step:  AIC=646.82
Survived ~ Pclass + Sex + Age + SibSp + Embarked

           Df Deviance    AIC
- Embarked  2   636.18 646.18
<none>          632.82 646.82
+ Fare      1   632.58 648.58
+ Parch     1   632.68 648.68
- SibSp     1   642.75 654.75
- Age       1   663.68 675.68
- Pclass    1   719.21 731.21
- Sex       1   816.03 828.03

Step:  AIC=646.18
Survived ~ Pclass + Sex + Age + SibSp

           Df Deviance    AIC
<none>          636.18 646.18
+ Embarked  2   632.82 646.82
+ Fare      1   635.54 647.54
+ Parch     1   636.09 648.09
- SibSp     1   646.70 654.70
- Age       1   669.11 677.11
- Pclass    1   741.12 749.12
- Sex       1   821.40 829.40
```

```r
summary(model2)
```

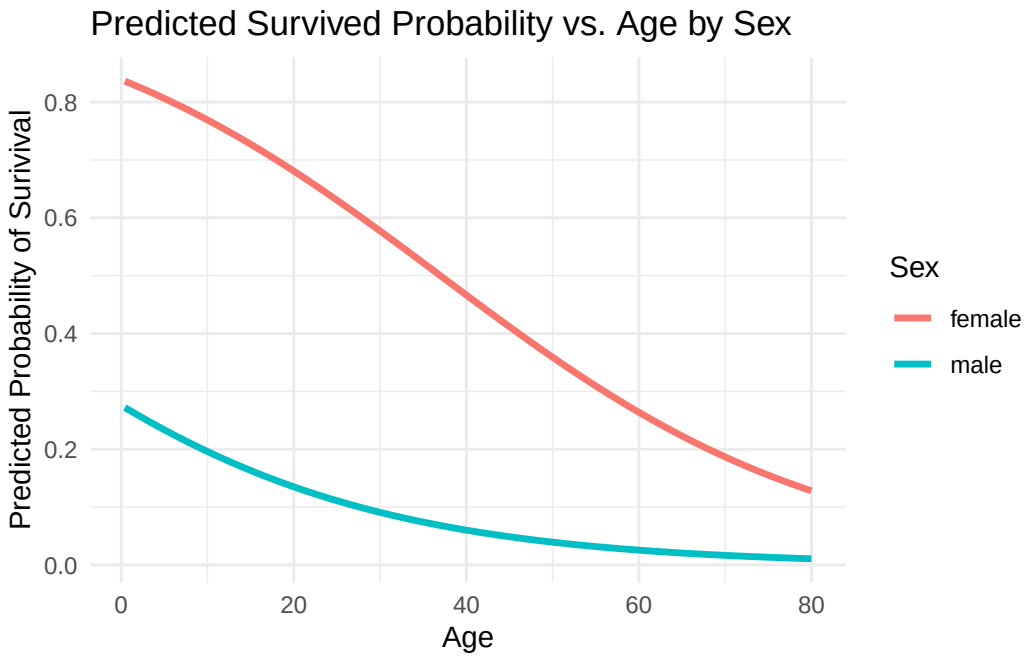
```
Call:
glm(formula = Survived ~ Pclass + Sex + Age + SibSp, family = binomial,
    data = data_clean)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.59083    0.54342  10.288  < 2e-16 ***
Pclass      -1.31392    0.14091  -9.324  < 2e-16 ***
Sexmale     -2.61477    0.21473 -12.177  < 2e-16 ***
Age         -0.04459    0.00817  -5.457 4.83e-08 ***
SibSp       -0.37465    0.12093  -3.098  0.00195 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 960.90  on 711  degrees of freedom
Residual deviance: 636.18  on 707  degrees of freedom
  (177 observations deleted due to missingness)
AIC: 646.18

Number of Fisher Scoring iterations: 5
```

5 Model validation

5.1 Prediction plot

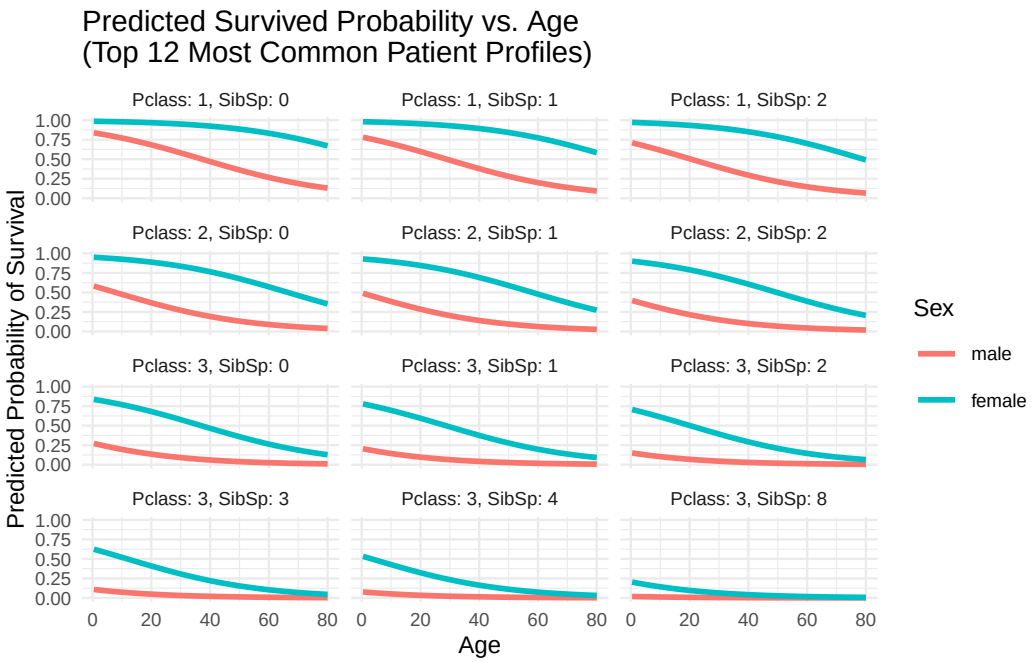
```r
# Vary sex
data1<-data%>%
  filter(!is.na(Age))
new_data <- expand.grid(
  Age = seq(min(data1$Age), max(data1$Age), length.out = 100),
  Sex = factor(c("male", "female")),
  Pclass=3,
  SibSp=0
)

# Predict
new_data$predicted_prob <- predict(model2, newdata = new_data, type = "response")
```

```r
# Plot by sex
ggplot(new_data, aes(x = Age, y = predicted_prob, color = Sex)) +
  geom_line(linewidth = 1.2) +
  labs(
    title = "Predicted Survived Probability vs. Age by Sex",
    x = "Age",
    y = "Predicted Probability of Surivival"
  ) +
  theme_minimal()
```




```r
library(purrr)
# 1. Find top 12 profiles (excluding sex)
top_profiles <- data %>%
  group_by(Pclass, SibSp) %>%
  summarise(count = n(), .groups = "drop") %>%
  arrange(desc(count)) %>%
  slice_head(n = 12)

# 2. Generate prediction data for each profile + both sexes
age_seq <- seq(min(data1$Age), max(data1$Age), length.out = 100)

# Expand to both Male and Female
new_data <- top_profiles %>%
  mutate(profile_id = row_number()) %>%
  group_split(profile_id) %>%
```

```
  map_dfr(function(profile) {
    profile_info <- profile %>% select(-count, -profile_id)

    # Add both sexes
    sexes <- c("male", "female")

    map_dfr(sexes, function(sex_val) {
      profile_replicated <- profile_info[rep(1, length(age_seq)), ]
      profile_replicated$Sex <- factor(sex_val, levels = c("male", "female"))
      profile_replicated$Age <- age_seq
      profile_replicated
    })
  })

# 3. Predict
new_data$predicted_prob <- predict(model2, newdata = new_data, type = "response")

# 4. Create facet label
new_data$profile_label <- with(new_data, paste0(
  "Pclass: ", Pclass, ", SibSp: ", SibSp
))

# 5. Plot with one curve per sex in each profile
ggplot(new_data, aes(x = Age, y = predicted_prob, color = Sex)) +
  geom_line(linewidth = 1) +
  facet_wrap(~ profile_label, ncol = 3) +
  labs(
    title = "Predicted Survived Probability vs. Age\n(Top 12 Most Common Patient Profiles)
    x = "Age",
    y = "Predicted Probability of Survival",
    color = "Sex"
  ) +
  theme_minimal(base_size = 9) +
  theme(strip.text = element_text(size = 7))
```



We plot the most 12 common profile of the passenger and use our model to predict the survival rate. It shows female always have higher survival rate. And the survival rate decrease as the age increases.

5.2 Model accuracy

```
model_data<- model.frame(model2)
actual<-model_data$Survived

Predict_probs<-predict(model2, type="response")
predict_class<-ifelse(Predict_probs>0.5, 1, 0)

Confusion_matrix<-table(Actual=actual, Predict=predict_class)
Confusion_matrix
```

```
      Predict
Actual   0   1
     0 366  58
     1  78 210
```

```
accuracy=mean(actual==predict_class)
print(paste("The model accuracy is", round(accuracy*100, 2), "%."))#%80.9%
```

```
[1] "The model accuracy is 80.9 %."
```