HARVARD EXTENSION SCHOOL

EXT CSCI E-106 Statistical Data Modeling Class Group Project Template

Author One

Anjan Chakravarti

Author Three

Author Four

Author Five

Author Six

Author Seven

Author Eight

Author Nine

23 November 2025

Abstract

This is the location for your abstract. It must consist of two paragraphs.

Classify whether a passenger on board the maiden voyage of the RMS Titanic in 1912 survived given their age, sex and class. Sample-Data-Titanic-Survival.csv to be used in the Final Project

Variable	Description
pclass	**Passanger Class, could be 1st, 2nd or 3rd**
survived	*Survival Status: 0=No, 1=Yes*
name	*Name of the Passanger*
Sex	*Sex*
sibsp	*Number of Siblings or Spouses aboard*
parch	*Number of Parents or Children aboard*
ticket	*Ticket Number*
fare	*Passenger Fare*
cabin	*Cabin number, “C85” would mean the cabin is on deck C and is numbered 85.*
embarked	*Port of Embarkation: C=Cherburg, S=Southampton, Q=Queenstown*
boat	*Lifeboat ID, if passenger survived*
body	*Body number (if passenger did not survive and body was recovered*
home.dest	*The intended home destination of the passenger*

Instructions:

0. Join a team with your fellow students with appropriate size (Up to Nine Students total) If you have not group by the end of the week of April 11 you may present the project by yourself or I will randomly assign other stranded student to your group. I will let know the final groups in April 11.
1. Load and Review the dataset named “Titanic_Survival_Data.csv”
2. Create a model development document that describes the model using this template, input the name of the authors, Harvard IDs, the name of the Group, all of your code and calculations, etc..:

II. Description of the data and quality (15 points)

Here you need to review your data, the statistical test applied to understand the predictors and the response and how are they correlated. Extensive graph analysis is recommended. Is the data continuous, or categorical, do any transformation needed? Do you need dummies?

Let's firstly important the libraries we will be using as well as importing the dataset. It looks like the half the dataset is composed of character variables and the other half are doubles. pclass and survived and are numeric but should be treated as categorical variables due to the fact that pclass has just three possible values and survived has just two. Additionally body just corresponds to essentially an ID number if a person didn't survive and should be treated as an identifier variable rather than a true numeric one.

```
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(ggplot2)
library(vcd)
```

```
## Loading required package: grid
```

```
titanic <- read_csv("/Users/Anjan/Downloads/Titanic_Survival_Data_and_Template/Titanic_Survival_Data.csv")
```

```
## Rows: 1310 Columns: 14
```

```
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr (7): name, sex, ticket, cabin, embarked, boat, home.dest
## dbl (7): pclass, survived, age, sibsp, parch, fare, body
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(titanic)
```

pclass <dbl>	survived <dbl>	name <chr>	sex <chr>	age <dbl>	sib… <dbl>
1	1	Allen, Miss. Elisabeth Walton	female	29.0000	0
1	1	Allison, Master. Hudson Trevor	male	0.9167	1
1	0	Allison, Miss. Helen Loraine	female	2.0000	1
1	0	Allison, Mr. Hudson Joshua Creighton	male	30.0000	1
1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	female	25.0000	1
1	1	Anderson, Mr. Harry	male	48.0000	0

6 rows | 1-8 of 14 columns

Numerical Summary

Let's look at summary to quickly look at the statistcal percentiles for each variable

```
summary(titanic)
```

```
##      pclass         survived         name               sex
##  Min.   :1.000   Min.   :0.000   Length:1310        Length:1310
##  1st Qu.:2.000   1st Qu.:0.000   Class :character   Class :character
##  Median :3.000   Median :0.000   Mode  :character   Mode  :character
##  Mean   :2.295   Mean   :0.382
##  3rd Qu.:3.000   3rd Qu.:1.000
##  Max.   :3.000   Max.   :1.000
##  NA's   :1       NA's   :1
##       age              sibsp            parch          ticket
##  Min.   : 0.1667   Min.   :0.0000   Min.   :0.000   Length:1310
##  1st Qu.:21.0000   1st Qu.:0.0000   1st Qu.:0.000   Class :character
##  Median :28.0000   Median :0.0000   Median :0.000   Mode  :character
##  Mean   :29.8811   Mean   :0.4989   Mean   :0.385
##  3rd Qu.:39.0000   3rd Qu.:1.0000   3rd Qu.:0.000
##  Max.   :80.0000   Max.   :8.0000   Max.   :9.000
##  NA's   :264       NA's   :1        NA's   :1
##       fare            cabin             embarked            boat
##  Min.   :  0.000   Length:1310        Length:1310        Length:1310
##  1st Qu.:  7.896   Class :character   Class :character   Class :character
##  Median : 14.454   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 33.295
##  3rd Qu.: 31.275
##  Max.   :512.329
##  NA's   :2
##       body        home.dest
##  Min.   :  1.0   Length:1310
##  1st Qu.: 72.0   Class :character
##  Median :155.0   Mode  :character
##  Mean   :160.8
##  3rd Qu.:256.0
##  Max.   :328.0
##  NA's   :1189
```

This summary is only really useful for age, sibsp, parch, and fare as the other 3 double variables aren't truly numeric data quantities. Age spans from essentially newborn to age 80 with the median and mean both being around 28-29 years. Most of the passengers were in between 21 and 39 years old. It looks like most passengers were travelling alone or with their spouse. Most didn't have parents or children with them although a few had large families with them.

Missing variables

```
colSums(is.na(titanic))
```

```
##    pclass  survived      name       sex       age     sibsp     parch    ticket
##         1         1         1         1       264         1         1         1
##      fare     cabin  embarked      boat      body home.dest
##         2      1015         3       824      1189       565
```

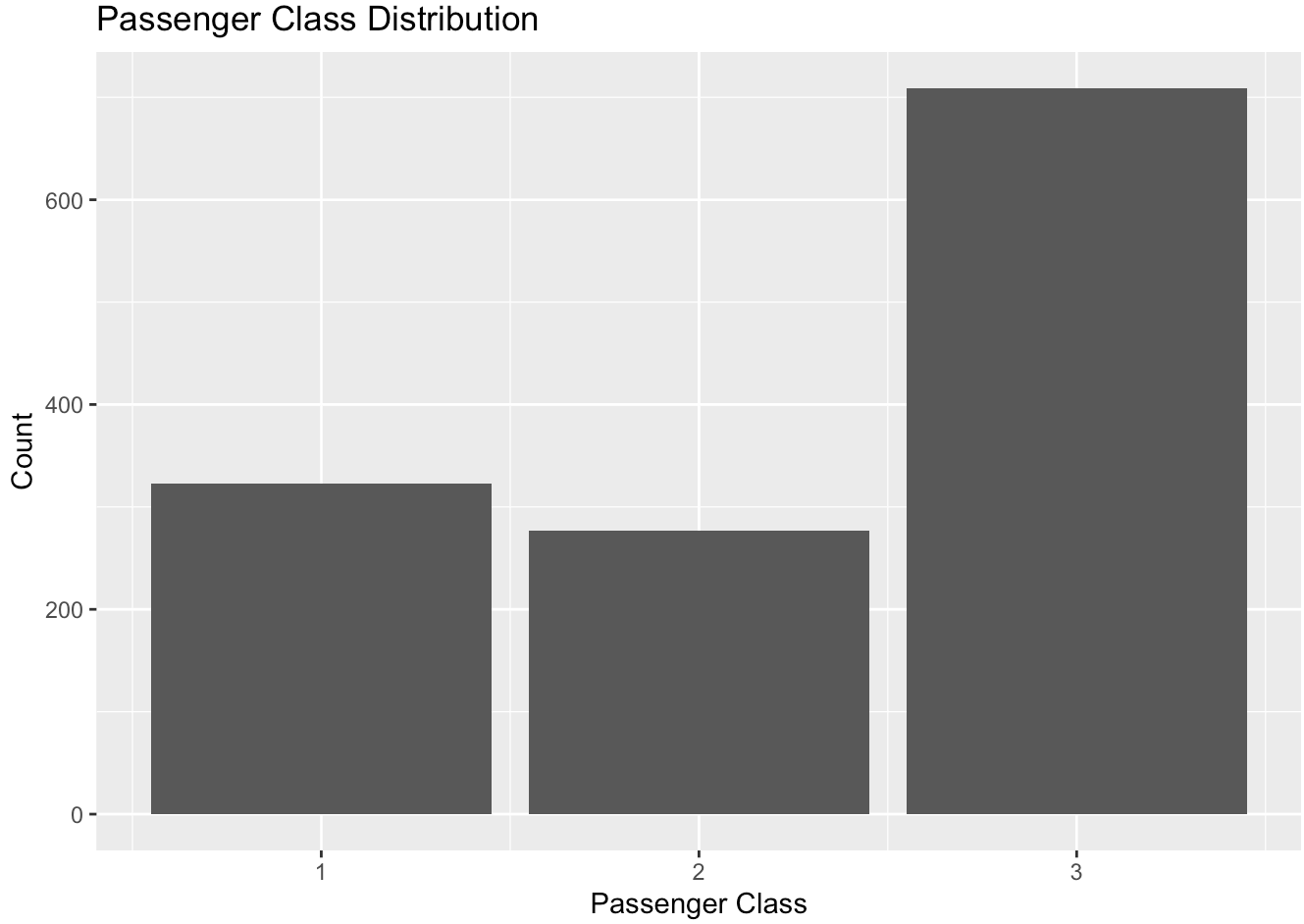
Looking at missing values, body has the majority missing with 1189/1310 being not recorded. Cabin, boat, and home.dest have a lot of missing data points. Body, Cabin, Boat, and home.dest have a large amount of missing points and shouldn't be considered for statistical analysis. The remainder of categories have a small amount of

missing values and can be considered although age should be inspected to see if it is worth including given it has 264 missing values and the rest have just 1 or 2.

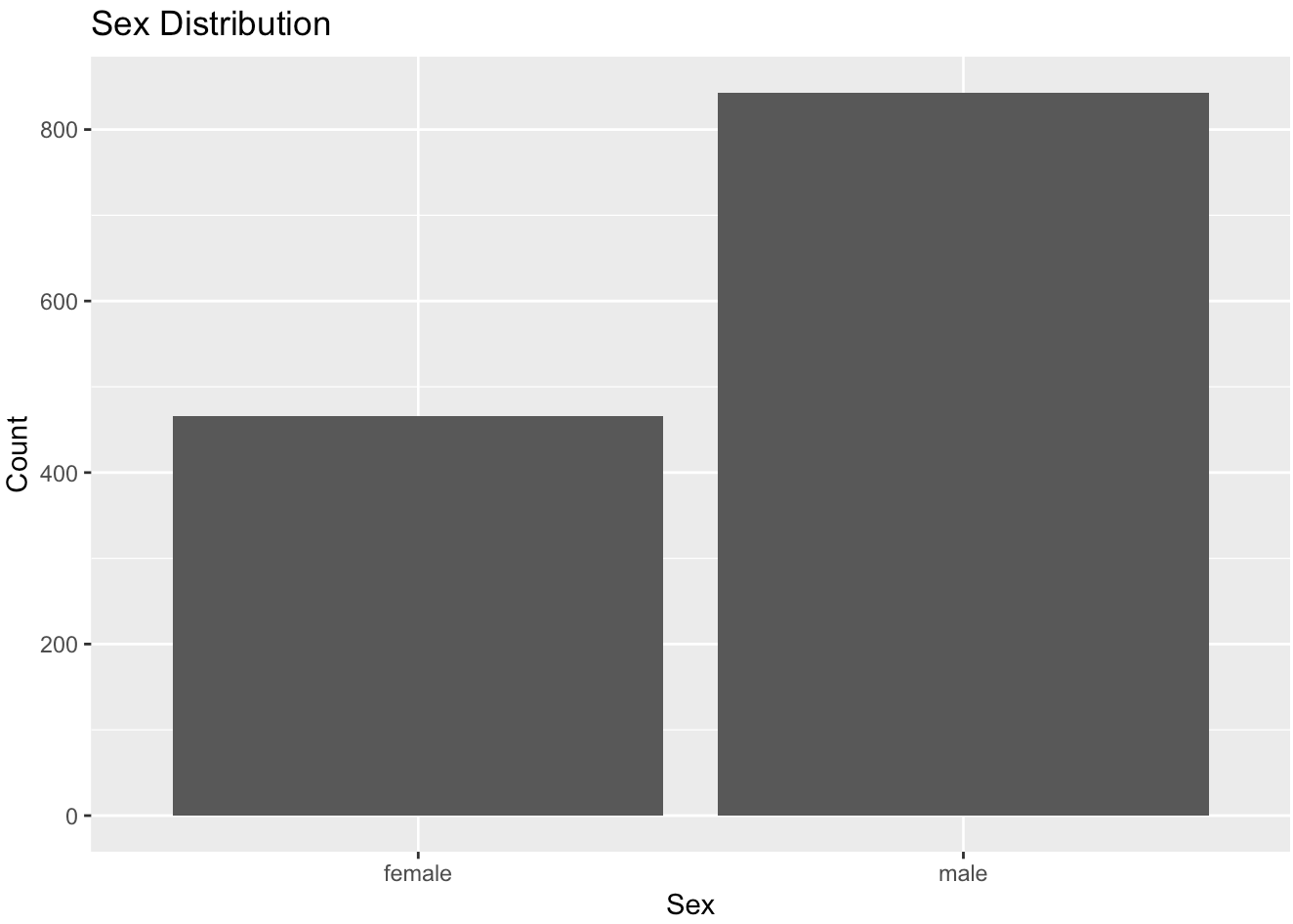
Count Distributions for Passenger Class, Sex, and Point of Embarkation (bar graphs)

Let's look at some histograms to see how the essentially categorical variables pclass, sex and the character variable point of embarkation.

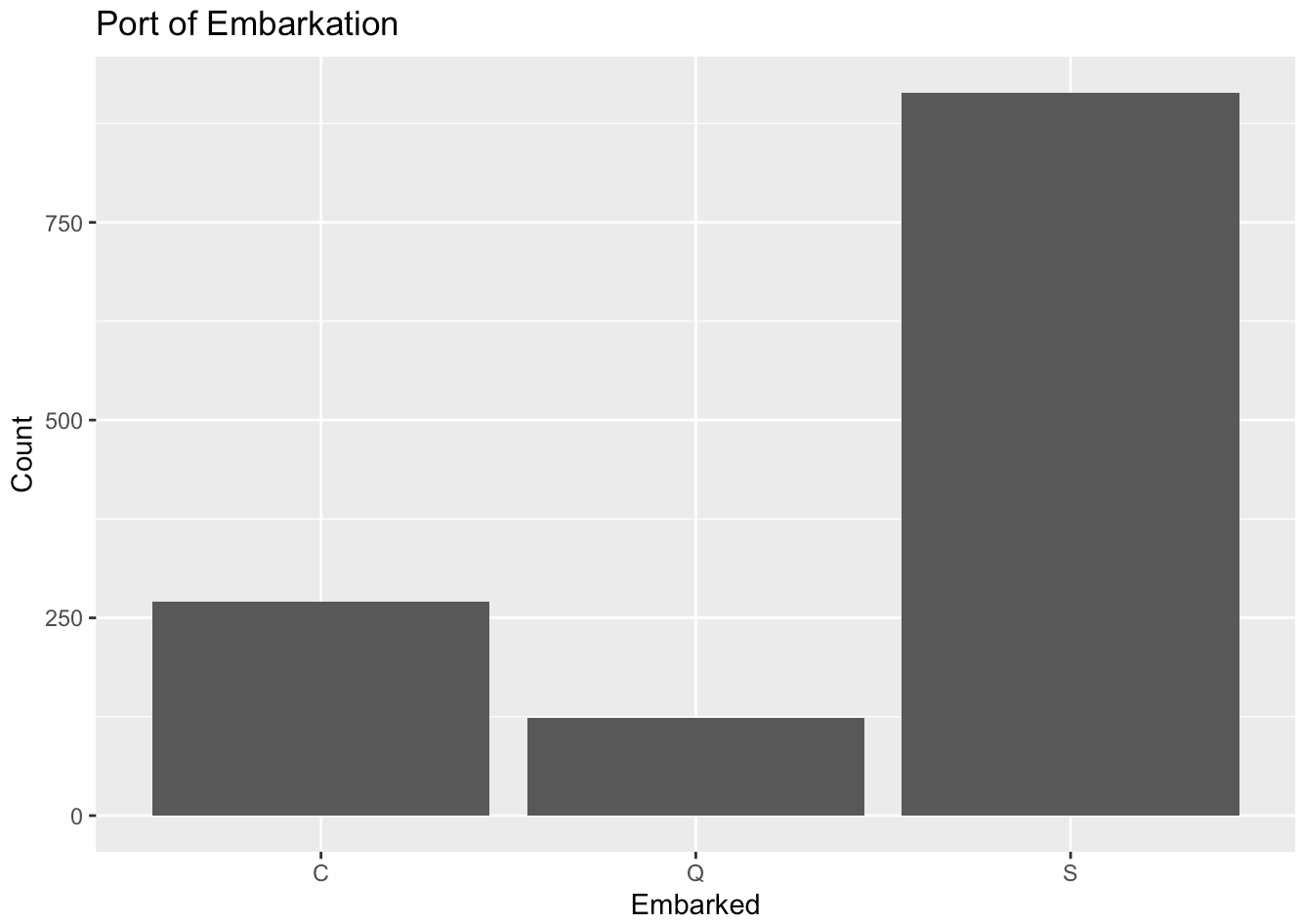
```
ggplot(titanic %>% filter(!is.na(pclass)), aes(x = pclass)) +
  geom_bar() +
  labs(title = "Passenger Class Distribution",
       x = "Passenger Class", y = "Count")
```




```
ggplot(titanic %>% filter(!is.na(sex)), aes(x = sex)) +
  geom_bar() +
  labs(title = "Sex Distribution",
       x = "Sex", y = "Count")
```




```
ggplot(titanic %>% filter(!is.na(embarked)), aes(x = embarked)) +
  geom_bar() +
  labs(title = "Port of Embarkation",
       x = "Embarked", y = "Count")
```



In the plot for pclass we can see that the majority are from class 3 with it having more than double that of class 1. Class 1 and 2 seem to have a similar number at around 300 while class seems to have around 700.

It looks like there were a lot more male travellers at around 850, while female travlelers numbered around just 475. This could prove a useful point to know when looking at fatality rates as the gender ratio is quite disparate.

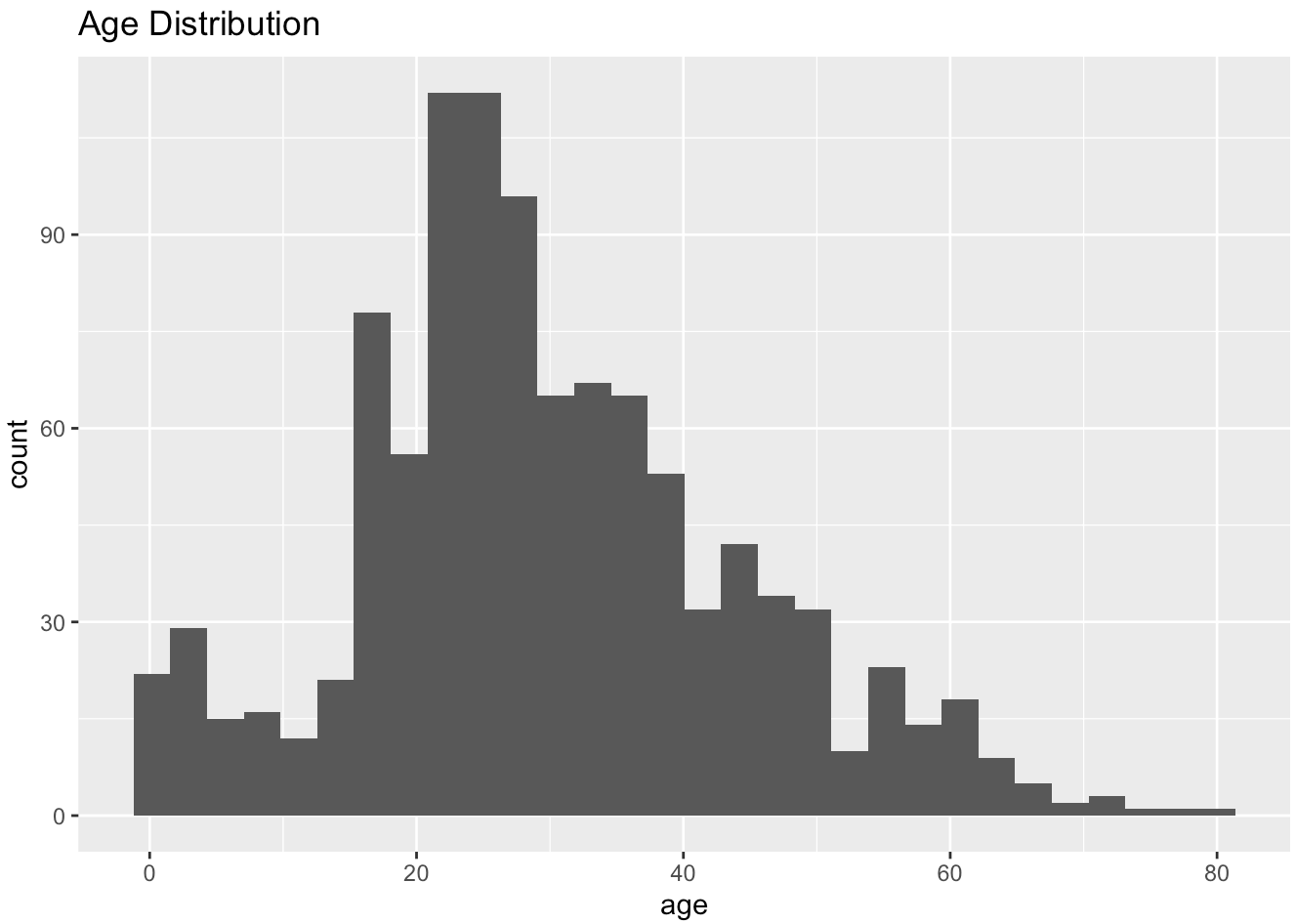
The most popular embarkation point was by far Southhampton at about 900 travellers while Cherburg has about 250 followed by Queenstown with the least at just about 125.

Histograms of Age Distributions and Fare (histograms)

Let's look at histograms of age distribution and fare.

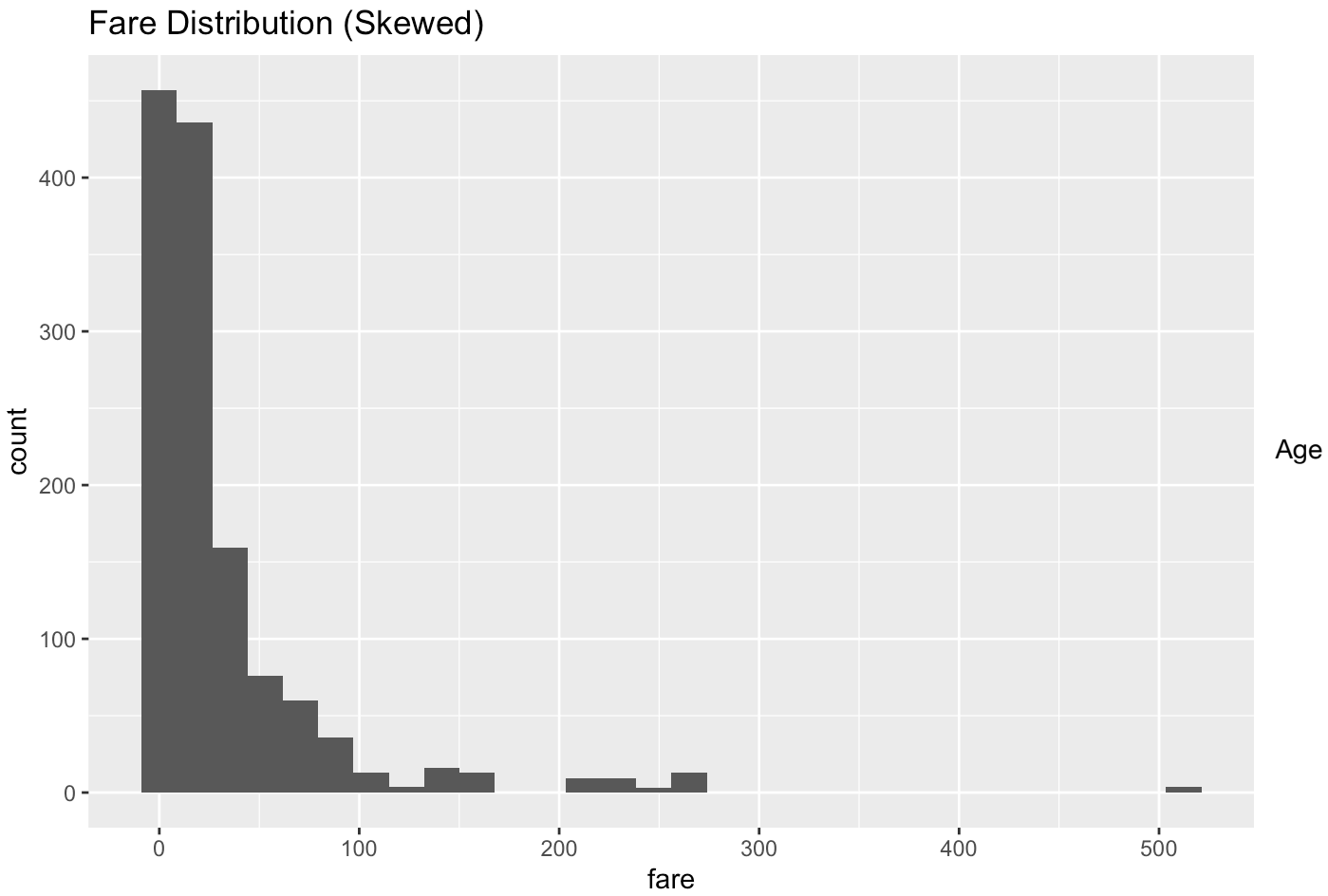
```
ggplot(titanic, aes(x = age)) +
  geom_histogram(bins = 30) +
  labs(title = "Age Distribution")
```

```
## Warning: Removed 264 rows containing non-finite outside the scale range
## (`stat_bin()`).
```




```
ggplot(titanic, aes(x = fare)) +
  geom_histogram(bins = 30) +
  labs(title = "Fare Distribution (Skewed)")
```

```
## Warning: Removed 2 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



looks skewed right with the biggest block of travllers being in their 20s a large amount of travelers also in their 30s. There weren't as many travellers with ages greater than 40 and less than 18.

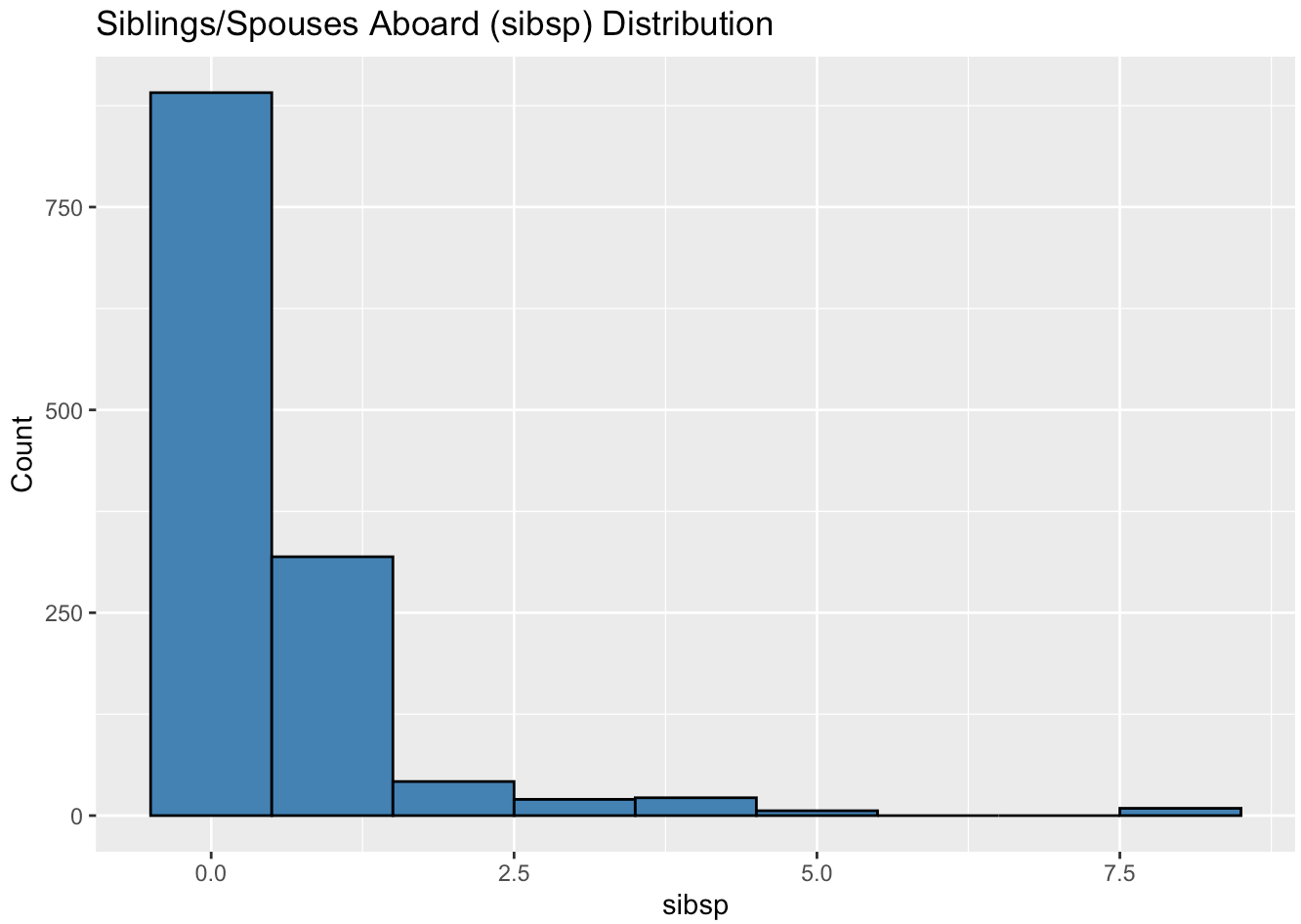
The majority of fares were less than 50 pounds with the bulk being around 10-20 pounds. There are some extreme outliers with a cluster at about 150, one at 225, and one quite far at around over 500 pounds. The majority of fares were on the lower end and this makes sense with what we know about 3rd class being the bulk of the passengers.

Histograms of Siblings/Spouses and Parents/Children aboard (histograms)

Let's look at histograms for sibsp and parch to see how roughly how many siblings/spouses and parents/children each passenger had.

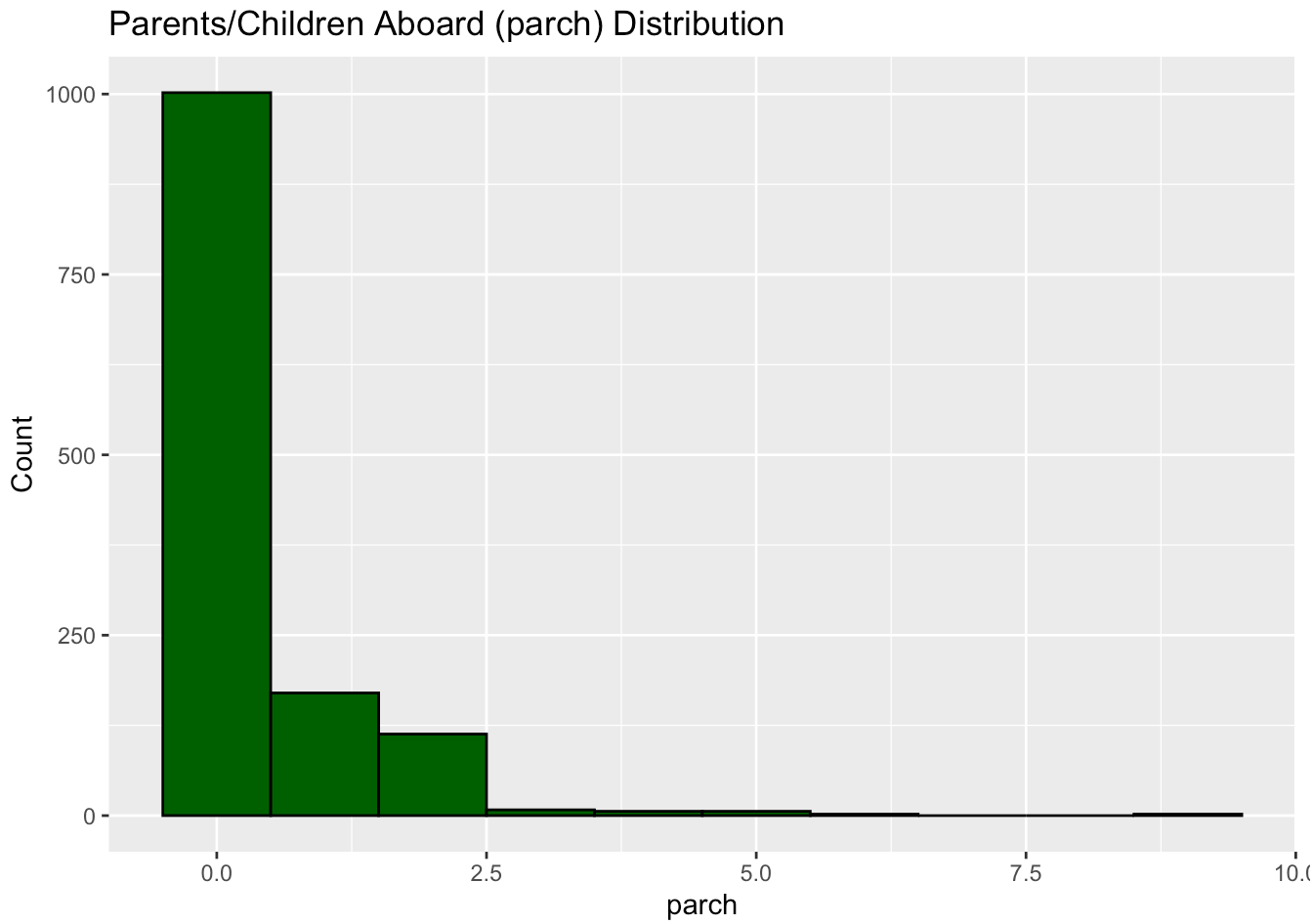
```
ggplot(titanic, aes(x = sibsp)) +
  geom_histogram(binwidth = 1, fill = "steelblue", color = "black") +
  labs(title = "Siblings/Spouses Aboard (sibsp) Distribution",
       x = "sibsp", y = "Count")
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).
```




```
ggplot(titanic, aes(x = parch)) +
  geom_histogram(binwidth = 1, fill = "darkgreen", color = "black") +
  labs(title = "Parents/Children Aboard (parch) Distribution",
       x = "parch", y = "Count")
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_bin()`).
```



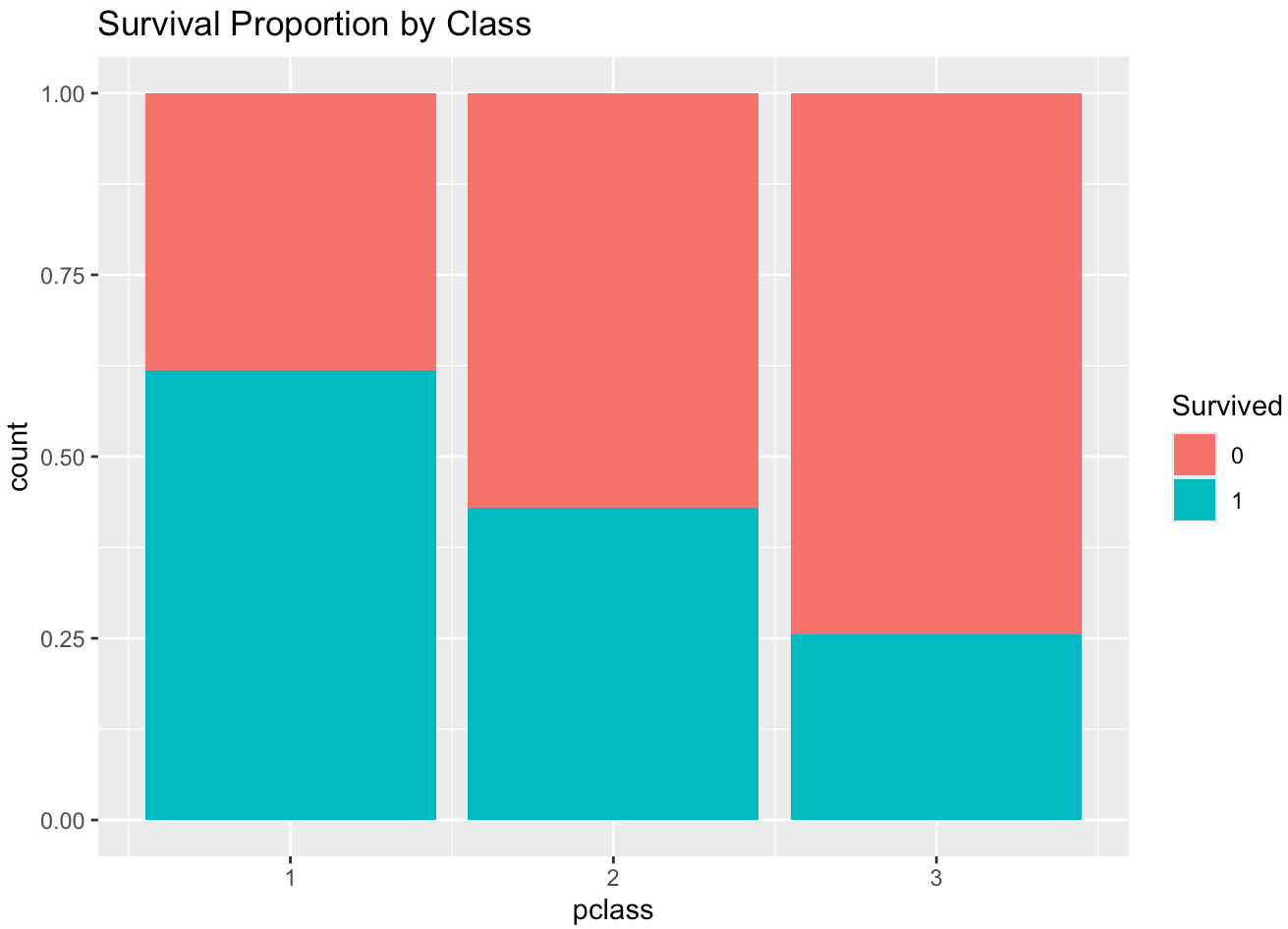
It looks like most travelers were on their own with the majority in both categories reporting zero. If they did have parents or children aboard most had less than 2. It seemed that most travelers were also alone at about 900 with about 325 having their spouse or a sibling aboard.

Survival proportion by Pclass and Sex (stacked barcharts)

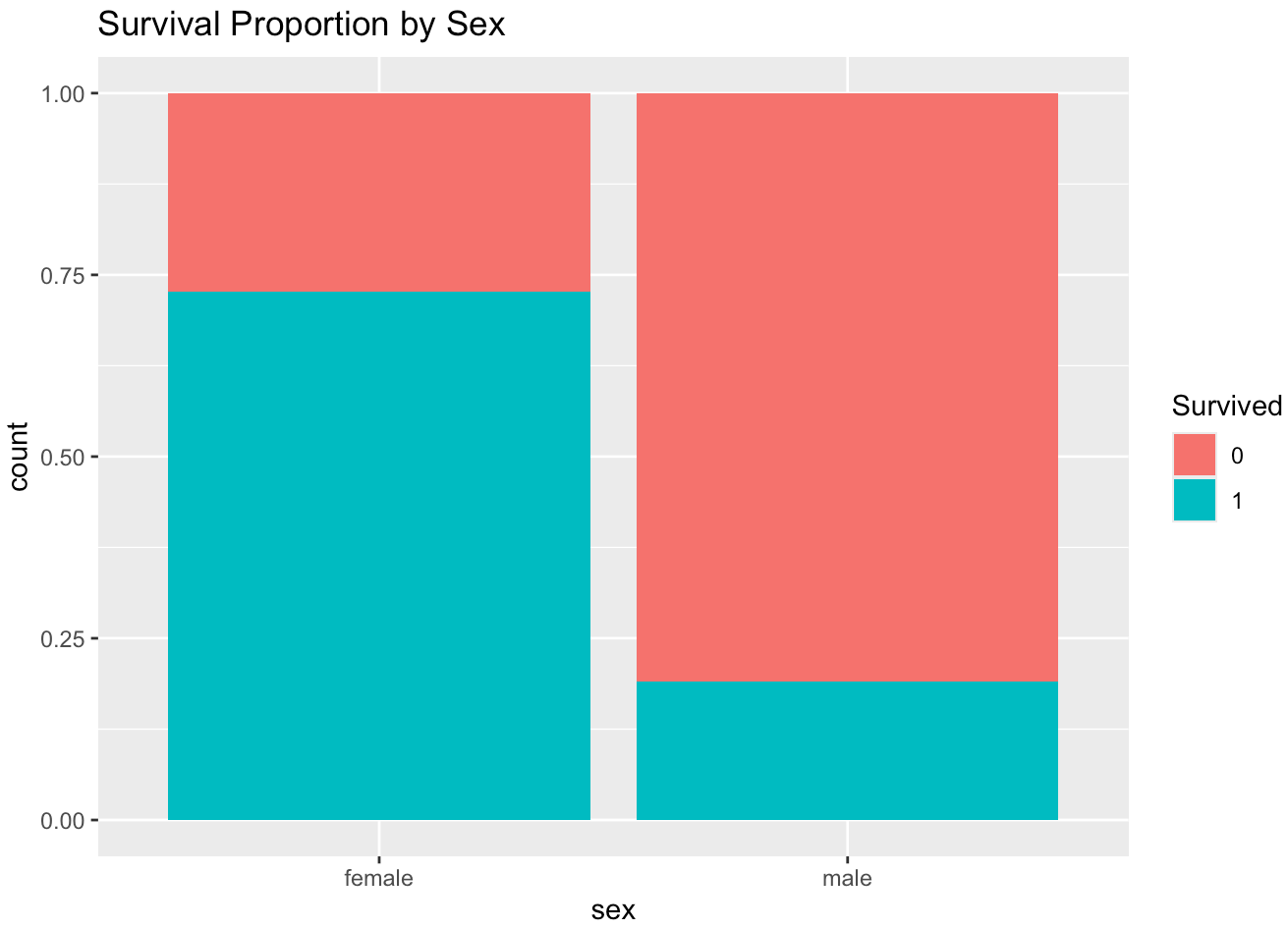
Let's look at two stacked barcharts that will illustrate key information. We know from historical records that women and children were evacuated first on lifeboats so we expect that trend to show up. We also expect to see a difference in survival by passenger class due to both the larger number of 3rd class passengers as well as potential preferential treatment to 1st class passengers.

```
ggplot(titanic, aes(x = pclass, fill = factor(survived))) +
  geom_bar(position = "fill") +
  labs(title = "Survival Proportion by Class", fill = "Survived")
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_count()`).
```




```
titanic %>%
  filter(!is.na(sex), !is.na(survived)) %>%
  ggplot(aes(x = sex, fill = factor(survived))) +
  geom_bar(position = "fill") +
  labs(title = "Survival Proportion by Sex",
       fill = "Survived")
```



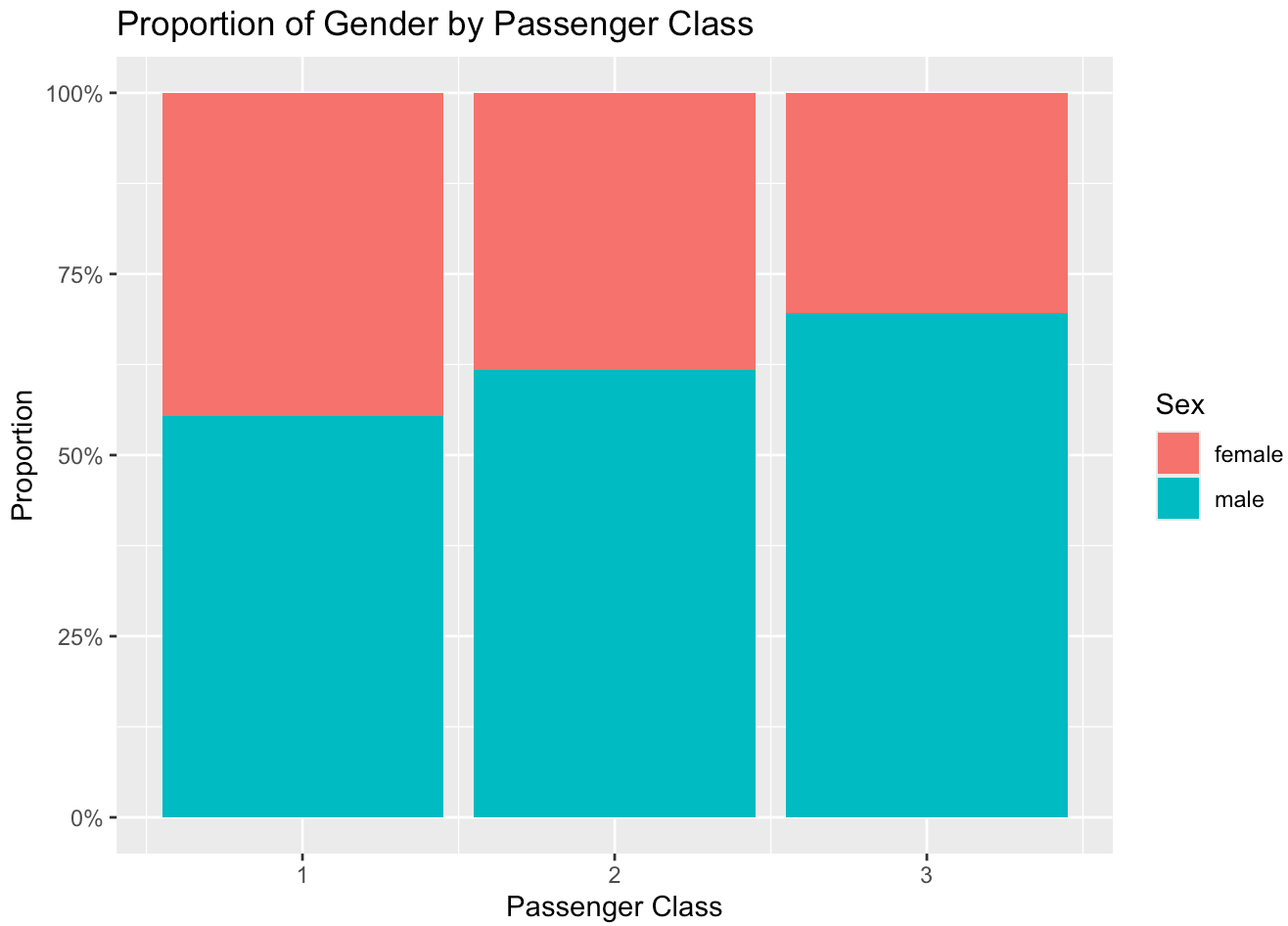
As can be seen there is a clear difference in survival rates by class with rates decrease from 1st to 2nd class and 2nd to 3rd class in a similar amount for each. 1st class looks to have a survival rate of about 62.5% with 2nd having about 42.5% and 3rd class having a survival rate of just 25%.

Additionally there is a massive gender survival disparity with females at almost a 75% survival rate versus just about 20% for males. The bulk of the passengers were also males indicating a very large death toll for males with seemingly most lifeboats being allocated to female passengers and mostly those in 1st class.

Gender Proportion by Passenger Class (stacked barcharts)

Let's see what proportion each gender makes up in each passenger class.

```
ggplot(titanic %>% filter(!is.na(sex), !is.na(pclass)),
       aes(x = pclass, fill = sex)) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Gender by Passenger Class",
       x = "Passenger Class",
       y = "Proportion",
       fill = "Sex") +
  scale_y_continuous(labels = scales::percent)
```

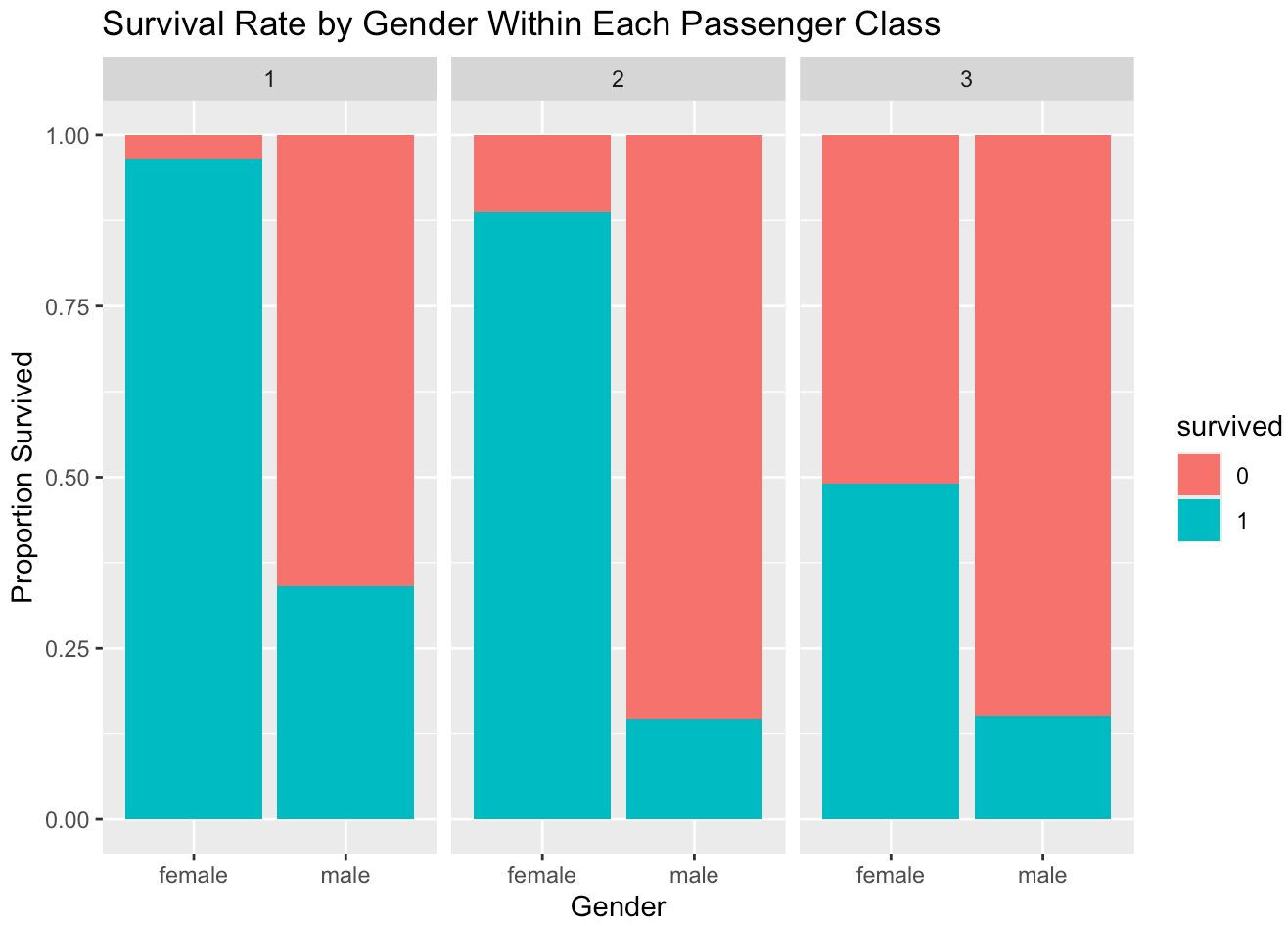


In looking at the gender split of passenger class by gender, we can see that 1st class was close to a 50-50 gender split whereas 3rd class was closer to 70% males. This means that we should expect to see a very high survival rate for women in 1st class and a very high death toll for men in 3rd class.

Survival Rate by Gender split by Passenger Class (stacked barcharts)

Let's look at survival rate split by gender and also split by class.

```
titanic %>%
  filter(!is.na(sex), !is.na(pclass), !is.na(survived)) %>%   # remove NAs
  mutate(
    survived = factor(survived),
    sex = factor(sex),
    pclass = factor(pclass)
  ) %>%
  ggplot(aes(x = sex, fill = survived)) +
    geom_bar(position = "fill") +
    facet_wrap(~ pclass) +
    labs(title = "Survival Rate by Gender Within Each Passenger Class",
         y = "Proportion Survived",
         x = "Gender")
```



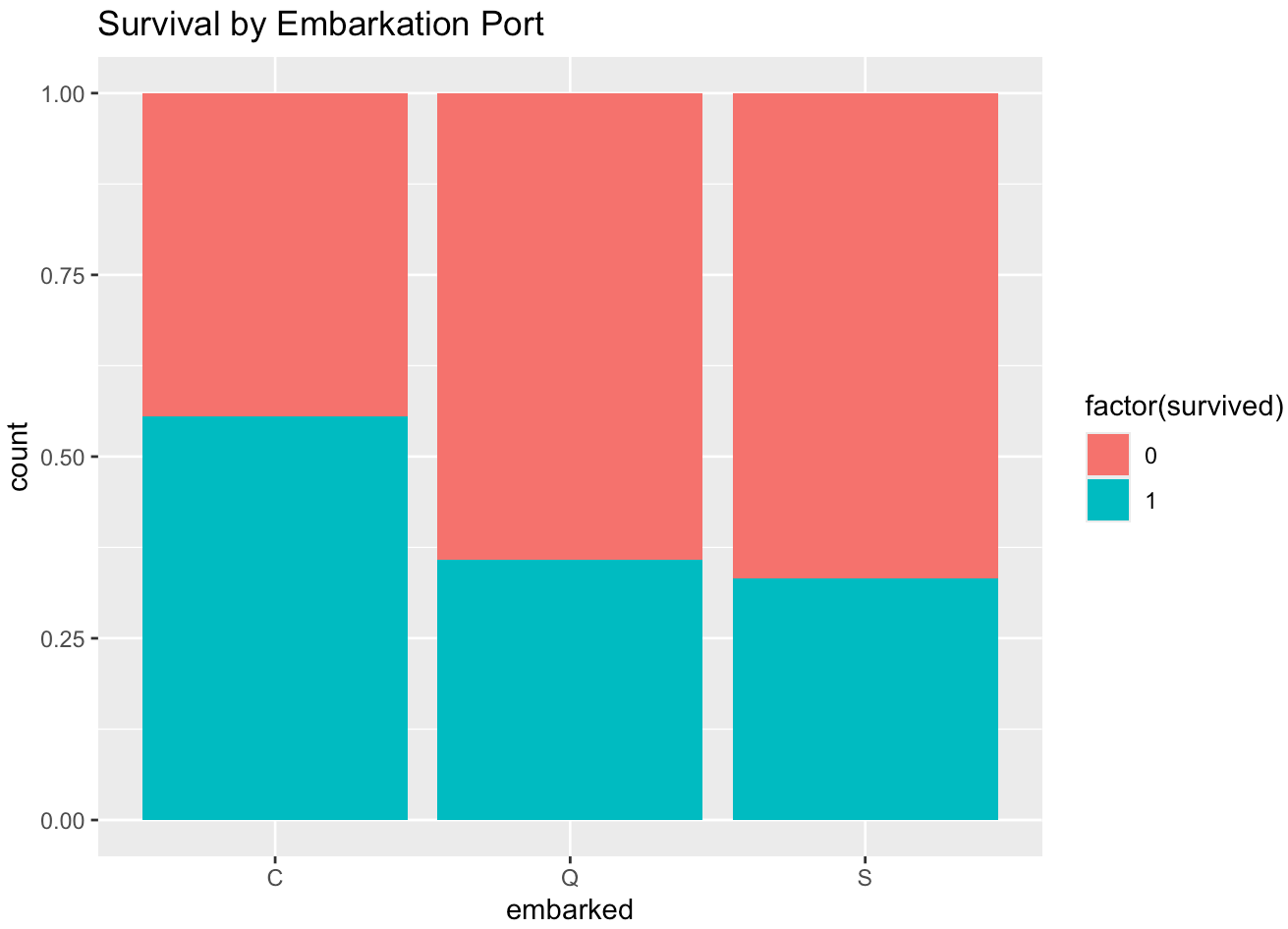
Gender is the strongest predictor of survival — females consistently survived at much higher rates. Passenger class heavily moderates this effect: 1st class women had nearly universal survival. 3rd class women had only about a 50% chance. Men in all classes had low survival, particularly in 3rd class.

This supports the historical “women and children first” policy and the structural disadvantages faced by 3rd-class passengers during evacuation.

Survival by Embarkation Port (stacked barchart)

Let’s look at survival rate by embarkation port to see if there is a relationship.

```
ggplot(titanic %>% filter(!is.na(embarked), !is.na(survived)),
aes(x=embarked, fill=factor(survived))) +
  geom_bar(position="fill") +
  labs(title="Survival by Embarkation Port")
```

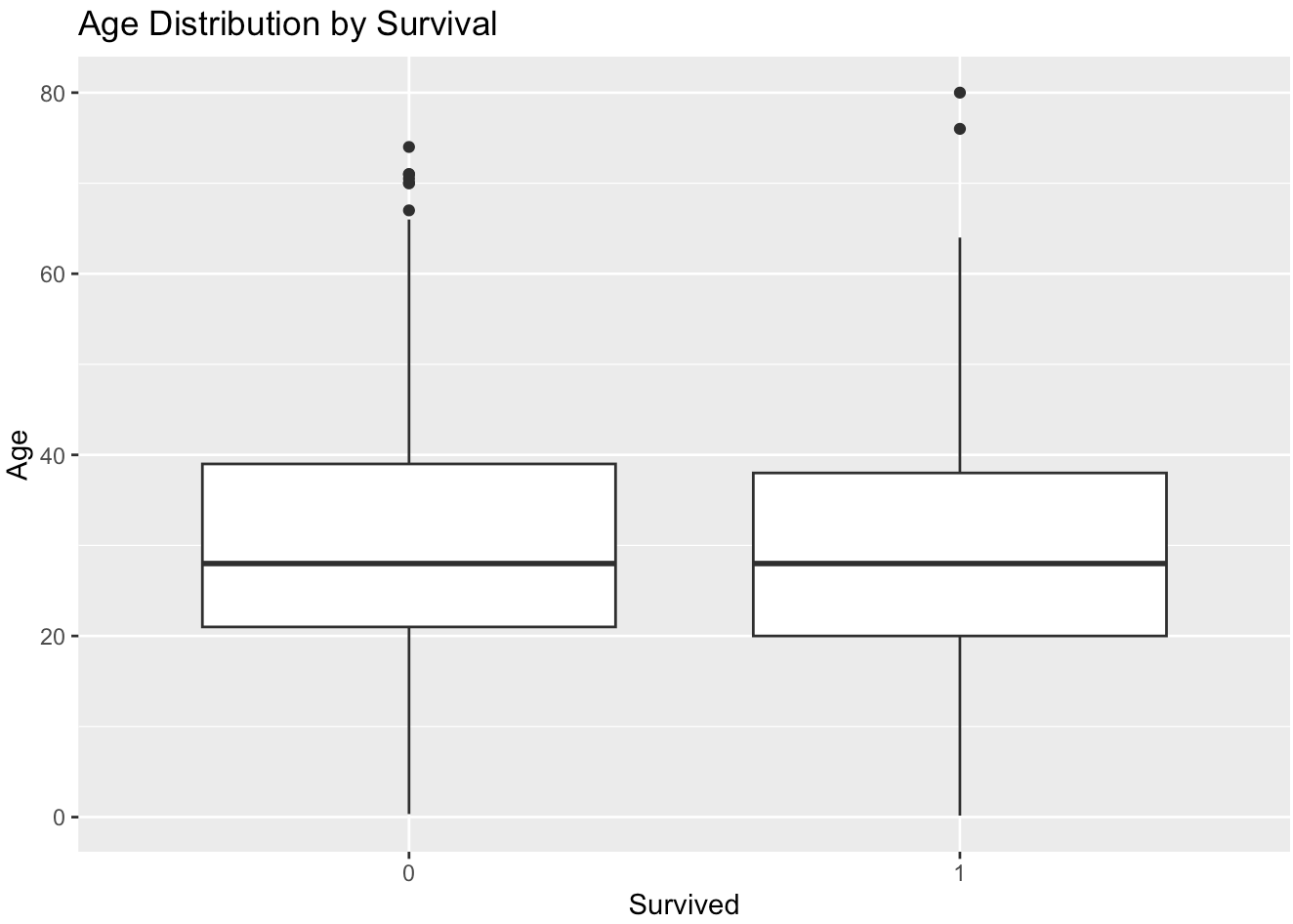


It looks like Cherburg has a very high survival rate, much higher than both Queenstown and Southampton. Cherburg passengers had a greater than 50% chance of survival while the other ports were at just around 37%. This is a noticeable result although the disparity isn't as strong as some of the other predictor variables.

Age Distributions by Survival (boxplots)

Let's see if age has any noticeable relationship.

```
ggplot(titanic %>% filter(!is.na(age), !is.na(survived)),
       aes(x = factor(survived), y = age)) +
  geom_boxplot() +
  labs(title="Age Distribution by Survival",
       x="Survived",
       y="Age")
```



It appears based on these side by side boxplots that Age doesn't seem to have much of a relationship with Survived. The 25th, 50th, and 75th percentiles have very similar values.

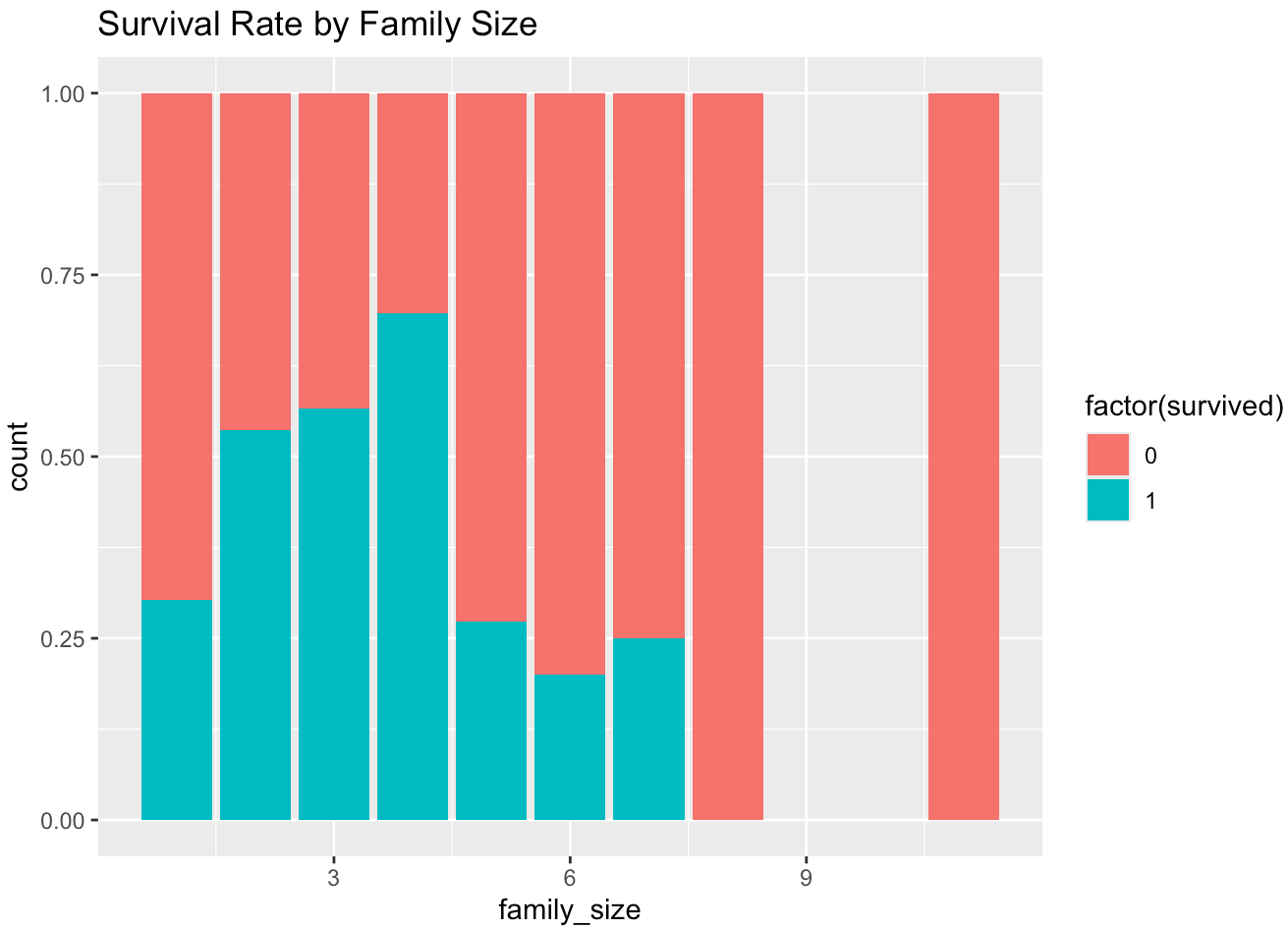
Survival Rate by Family Size (stacked barchart)

Let's create a family size column and see if survival rate has a relationship with it.

```
titanic$family_size <- titanic$sibsp + titanic$parch + 1

ggplot(titanic, aes(x=family_size, fill=factor(survived))) +
  geom_bar(position="fill") +
  labs(title="Survival Rate by Family Size")
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_count()`).
```

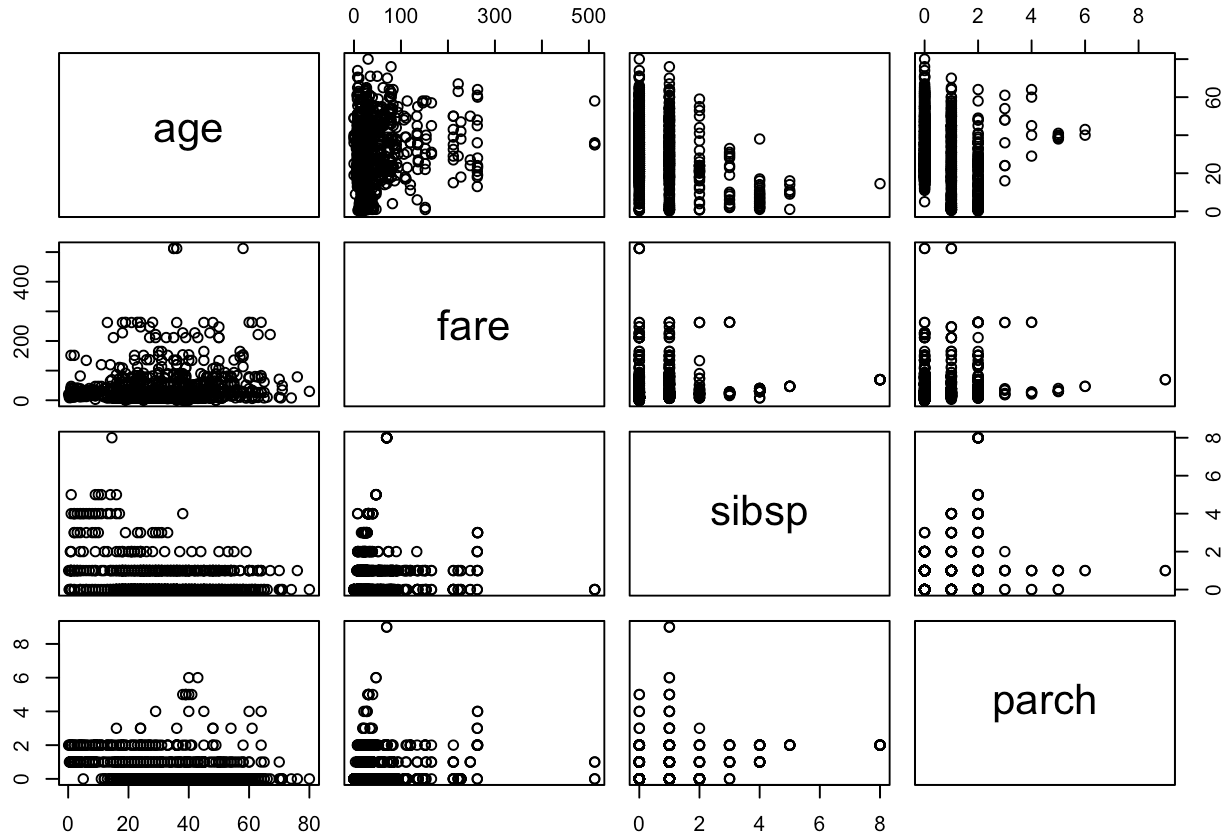


Families with 4 people seem to have the highest rate. Families of 2 and 3 also have high values with families of size greater than 4 having very low rates of survival.

Side by Side scatterplots

Let's look at scatterplots of each truly numerical variable side by side.

```
pairs(~ age + fare + sibsp + parch, data=titanic)
```

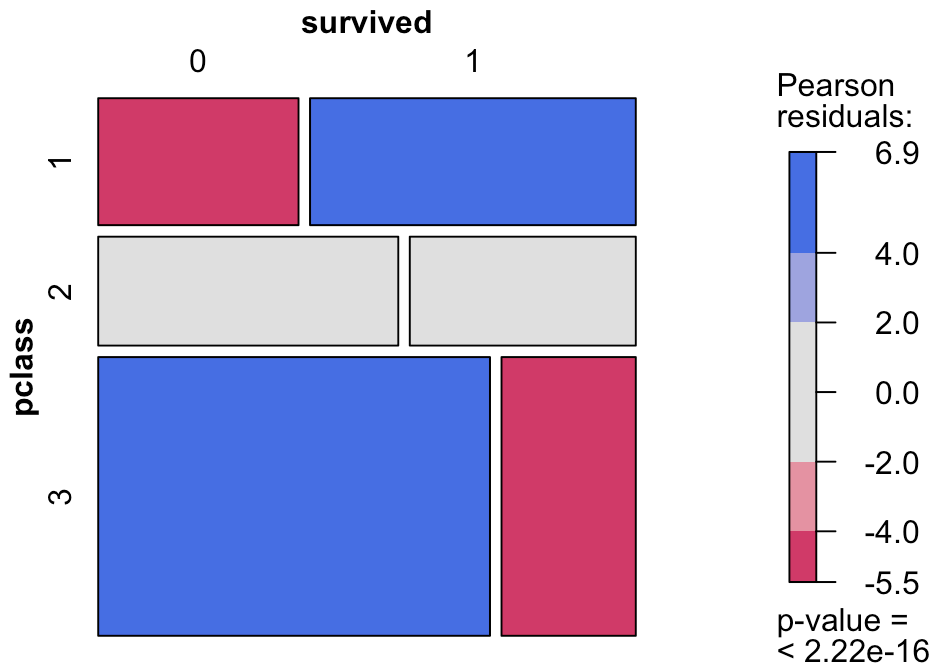


It looks like age and fare don't have a noticeable relationship with sibsp and parch seeming to have a positive correlation which makes sense as some groups of large travelers could have been travelling together.

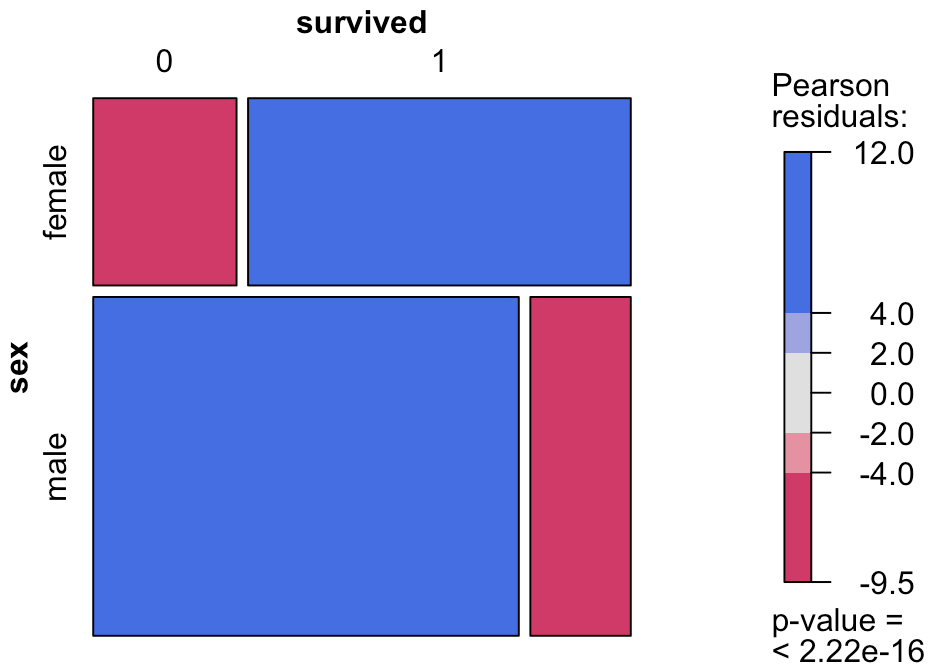
Survived by Pclass and Survived by Sex (mosaic plots)

Let's look at mosaic plots of look at the correlation betwen Pclass and Sex with Survived as these appeared to be the most significant predictor variables from the above analysis.

```
mosaic(~ pclass + survived, data = titanic, shade = TRUE)
```




```
mosaic(~ sex + survived, data = titanic, shade = TRUE)
```



The size of each rectangle corresponds to the number of passengers in that category. The color shows whether more or fewer passengers were observed than expected under a model of independence.

The first mosaic plot shows a strong association between passenger class and survival (χ^2 test $p < 2.22e\text{-}16$). First-class passengers survived at rates much higher than expected, while third-class passengers had a disproportionately high death rate. Second-class passengers were close to the expected pattern. These residual patterns indicate that socioeconomic status, represented by passenger class, was a major determinant of survival on the Titanic.

The second mosaic plot reveals a strong association between sex and survival on the Titanic (χ^2 $p < 2.22e\text{-}16$). Females survived at much higher rates than expected under independence, while males died at significantly higher rates than expected. This pattern reflects the disproportionate protection given to women during evacuation and confirms sex as one of the strongest predictors of survival in the dataset.

Convert categorical variables into factors for next steps.

```
# Convert categorical variables to factors for future analysis
titanic$pclass    <- factor(titanic$pclass)
titanic$sex       <- factor(titanic$sex)
titanic$survived  <- factor(titanic$survived)
titanic$embarked  <- factor(titanic$embarked)

str(titanic)
```

```
## spc_tbl_ [1,310 × 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ pclass     : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ survived   : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 2 1 2 1 ...
##  $ name       : chr [1:1310] "Allen, Miss. Elisabeth Walton" "Allison, Master. Hudson
Trevor" "Allison, Miss. Helen Loraine" "Allison, Mr. Hudson Joshua Creighton" ...
##  $ sex        : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
##  $ age        : num [1:1310] 29 0.917 2 30 25 ...
##  $ sibsp      : num [1:1310] 0 1 1 1 1 0 1 0 2 0 ...
##  $ parch      : num [1:1310] 0 2 2 2 2 0 0 0 0 0 ...
##  $ ticket     : chr [1:1310] "24160" "113781" "113781" "113781" ...
##  $ fare       : num [1:1310] 211 152 152 152 152 ...
##  $ cabin      : chr [1:1310] "B5" "C22 C26" "C22 C26" "C22 C26" ...
##  $ embarked   : Factor w/ 3 levels "C","Q","S": 3 3 3 3 3 3 3 3 3 1 ...
##  $ boat       : chr [1:1310] "2" "11" NA NA ...
##  $ body       : num [1:1310] NA NA NA 135 NA NA NA NA NA 22 ...
##  $ home.dest  : chr [1:1310] "St Louis, MO" "Montreal, PQ / Chesterville, ON" "Montre
al, PQ / Chesterville, ON" "Montreal, PQ / Chesterville, ON" ...
##  $ family_size: num [1:1310] 1 4 4 4 4 1 2 1 3 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   pclass = col_double(),
##   ..   survived = col_double(),
##   ..   name = col_character(),
##   ..   sex = col_character(),
##   ..   age = col_double(),
##   ..   sibsp = col_double(),
##   ..   parch = col_double(),
##   ..   ticket = col_character(),
##   ..   fare = col_double(),
##   ..   cabin = col_character(),
##   ..   embarked = col_character(),
##   ..   boat = col_character(),
##   ..   body = col_double(),
##   ..   home.dest = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```