Titanic passanger survival analysis

Analysis objective

Identify the characteristics of passengers who survived and those who did not. Determine if certain groups of passengers had higher survival rates based on the features in the dataset.

Loading data

Load data into dataframe and printing general insights.

```
titanic_train <- read.csv("Titanic_train.csv")
str(titanic_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 T
 $ 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" ...
```

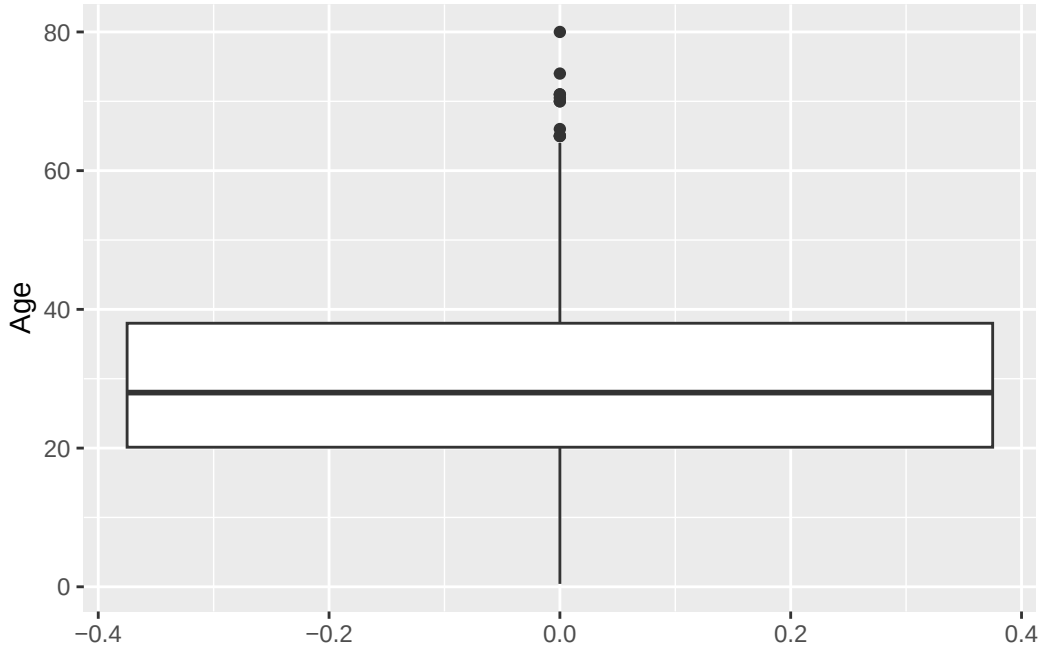
Exploratory Analysis

Performing visual exploratory analysis using the ggplot2 library.

```
library(ggplot2)
p <- ggplot(data = titanic_train)
p
```

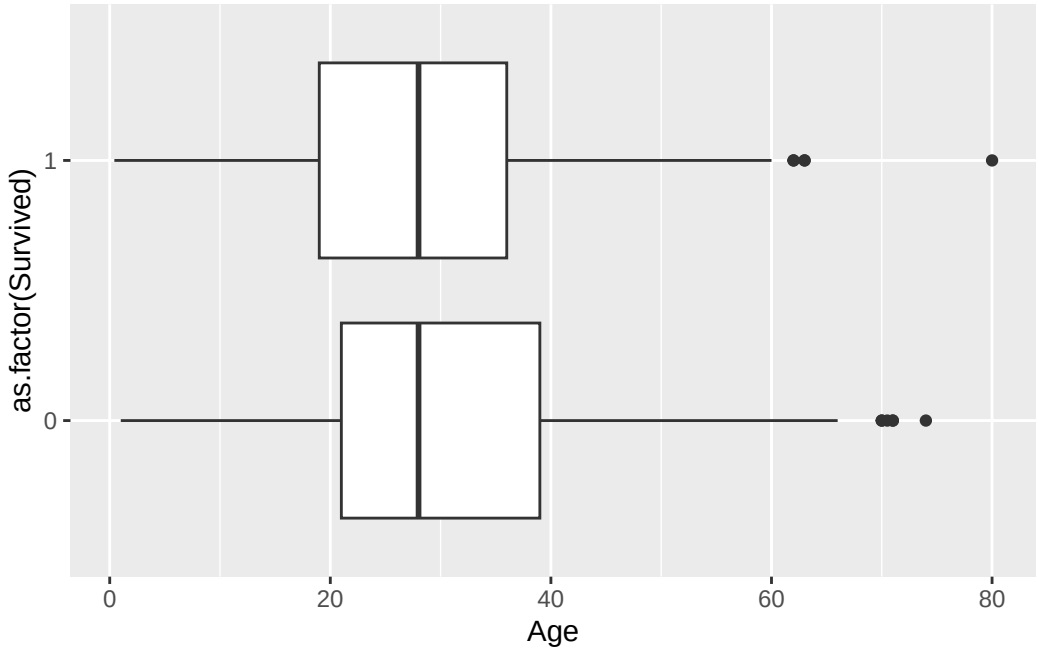
Maybe age defines groups with diferent survival chances.

```
p + geom_boxplot(aes(y = Age))
```



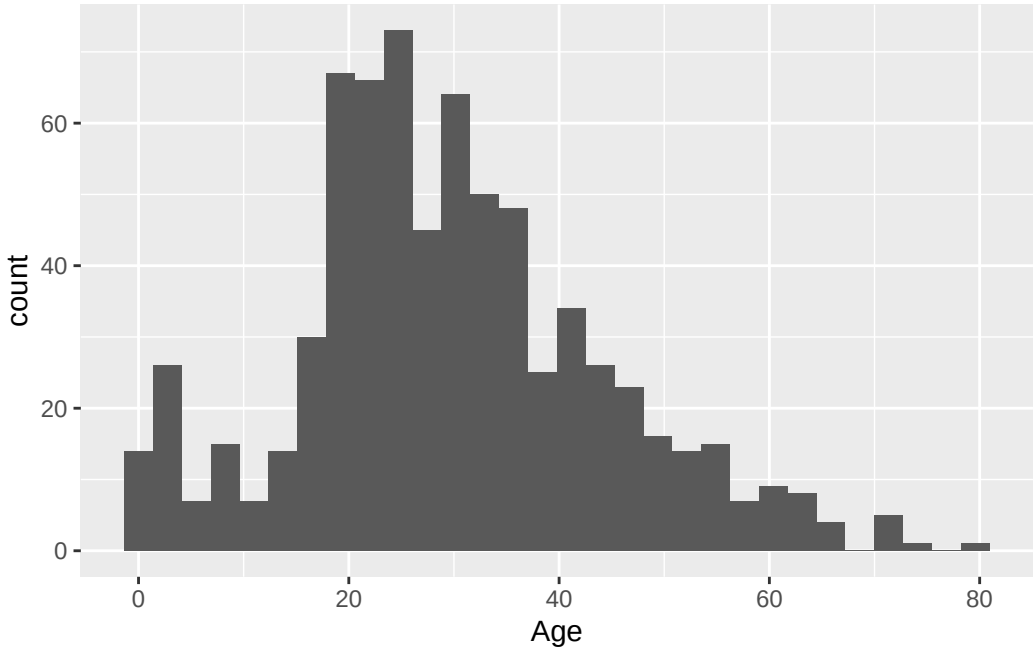
Sex can also define survival chances.

```
p + geom_boxplot(aes(x = Age, y = as.factor(Survived)))
```



Box plots did not reveal groups, analyzing distributions using histograms, starting with age, same as before.

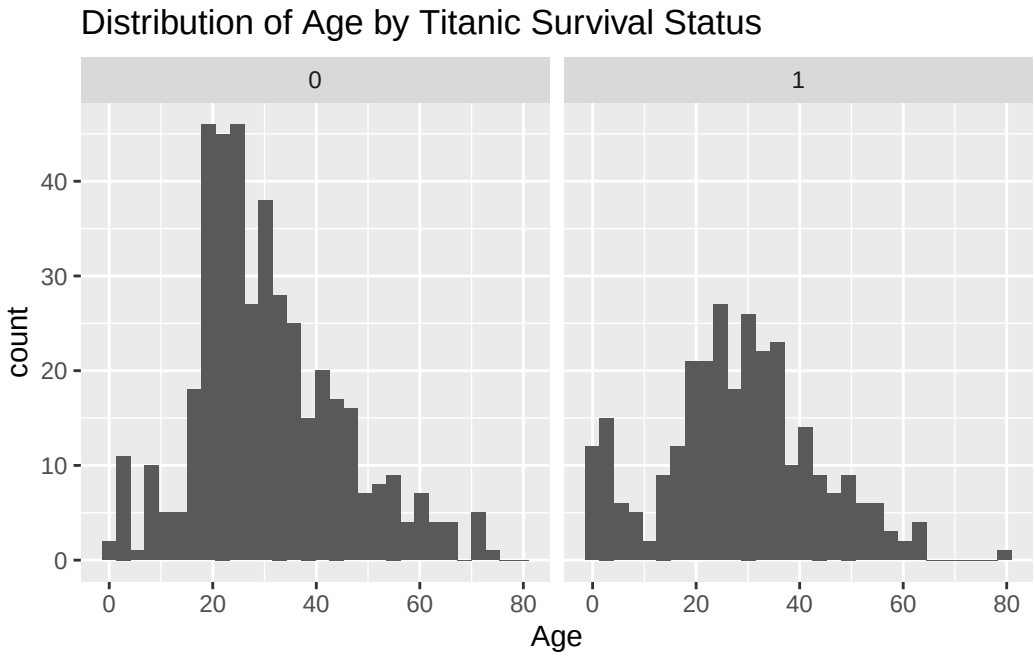
```
p + geom_histogram(aes(x = Age))
```



Groups of different ages are observed.
Are there any outliers?

Looking at the survival cpunts for the different ages.

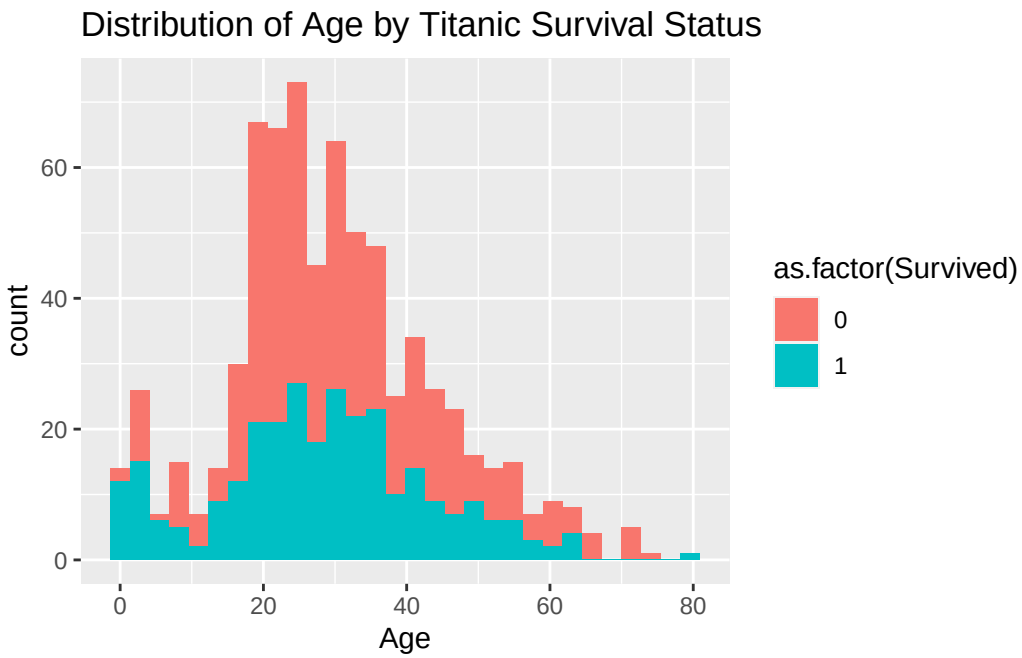
```
p + geom_histogram(aes(x = Age)) +
    facet_grid(cols = vars(Survived)) +
    ggtitle("Distribution of Age by Titanic Survival Status")
```



Side by side histograms show an age group that seems to have higher survival count. But it isn't entire easy to see.

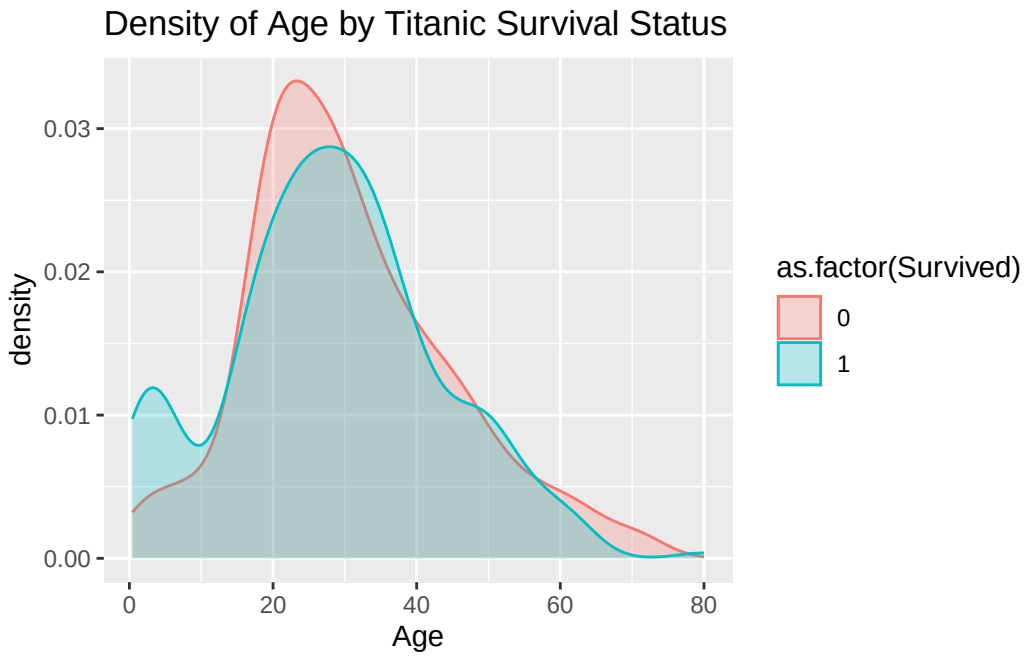
Plotting overlaping histograms to better compare survival and no survival counts.

```
p + geom_histogram(aes(x = Age, fill = as.factor(Survived))) +
    ggtitle("Distribution of Age by Titanic Survival Status")
```



Using a density plot to better see the comparison between survival and not.

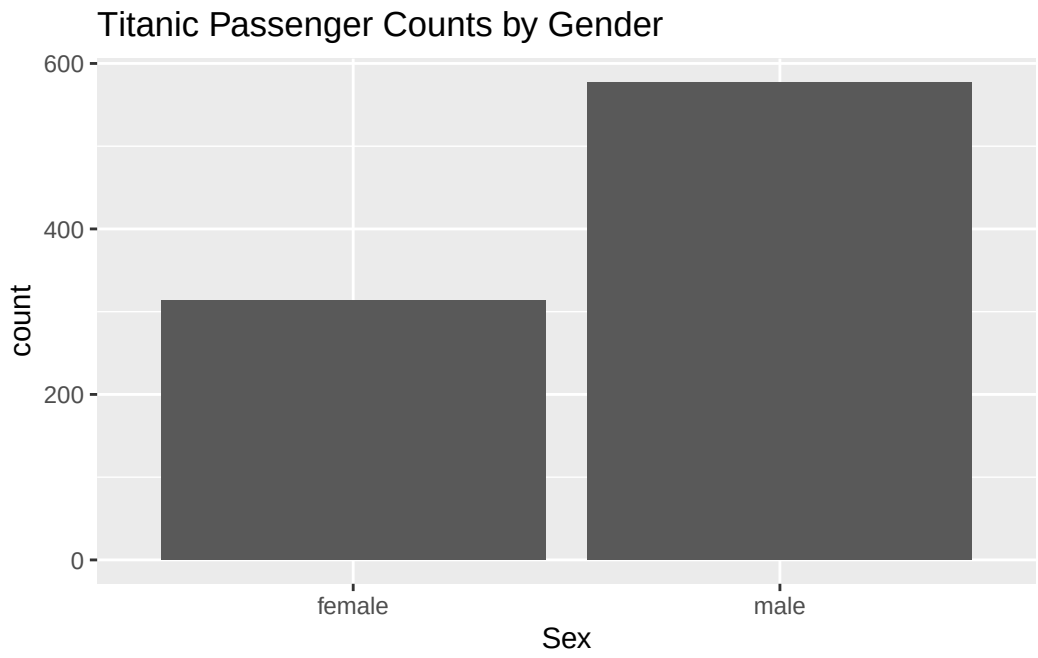
```
p + geom_density(aes(x = Age,
                     color = as.factor(Survived),
                     fill = as.factor(Survived)),
                     alpha = 0.25) +
    ggtitle("Density of Age by Titanic Survival Status")
```



Now we can see that young pasangers have a peak in survival counts. *First insight: many children survived.*

Now lets looking at sex in more detail. Using a bar chart to start exploration.

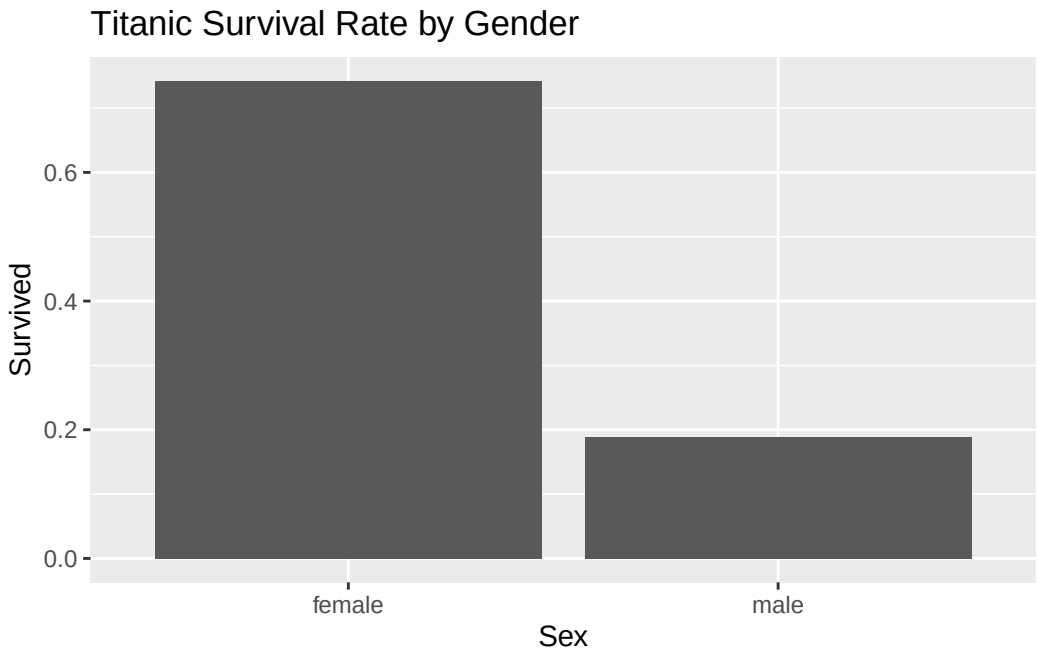
```
p + geom_bar(aes(x = Sex)) +
    ggtitle("Titanic Passenger Counts by Gender")
```



More male than female passangers observed.

To know if more females than males survived, we can't use absolute counts because there are many more male pasangers. A bar plot using relative rates will give better insights.

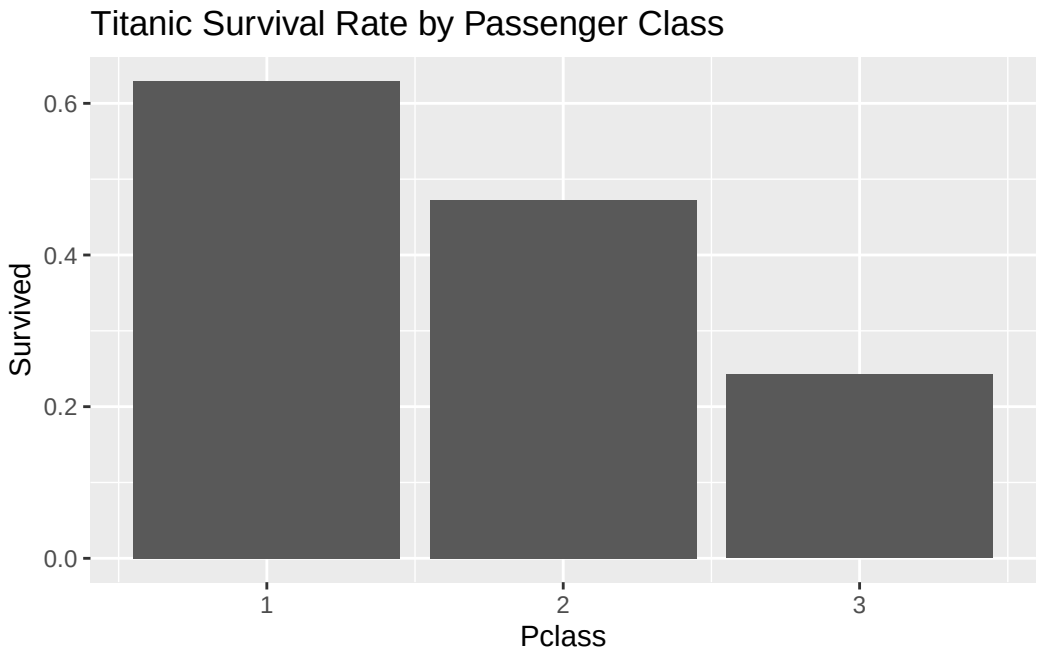
```
p + geom_bar(aes(x = Sex,
                 y = Survived),
                 stat = "summary",
                 fun = "mean") +
    ggtitle("Titanic Survival Rate by Gender")
```



The bar shows female survival rate is much greated than male.

We can do a similar analysis for pasangers of different classes.

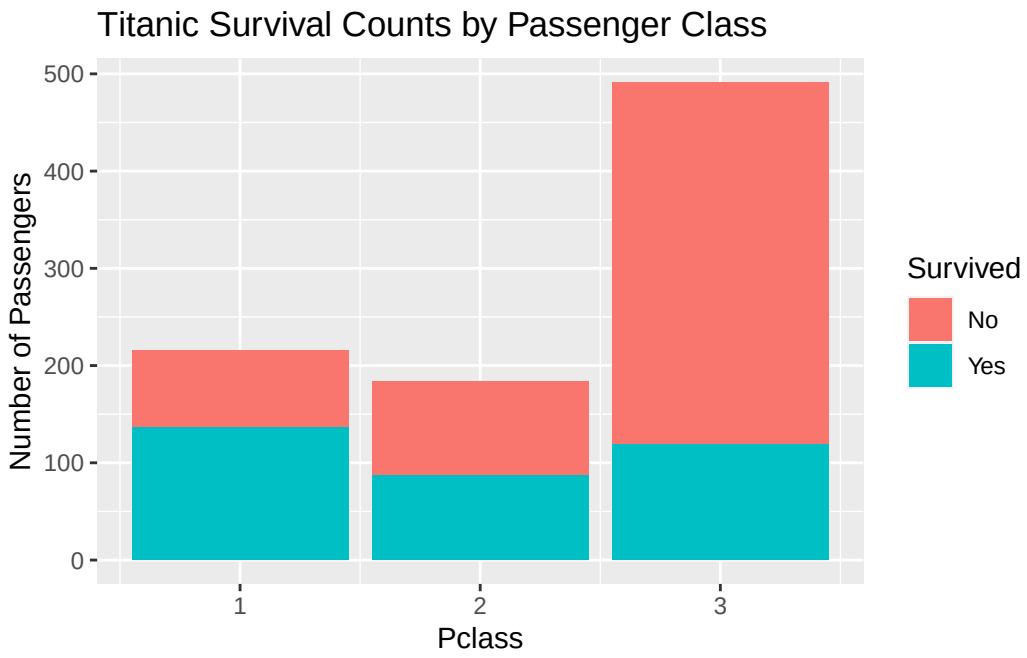
```
p + geom_bar(aes(x = Pclass,
                 y = Survived),
                 stat = "summary",
                 fun = "mean") +
    ggtitle("Titanic Survival Rate by Passenger Class")
```



First class pasangers have disproportionaly higher survival rates than other classes.

Looking a pasanger counts, to find better insights.

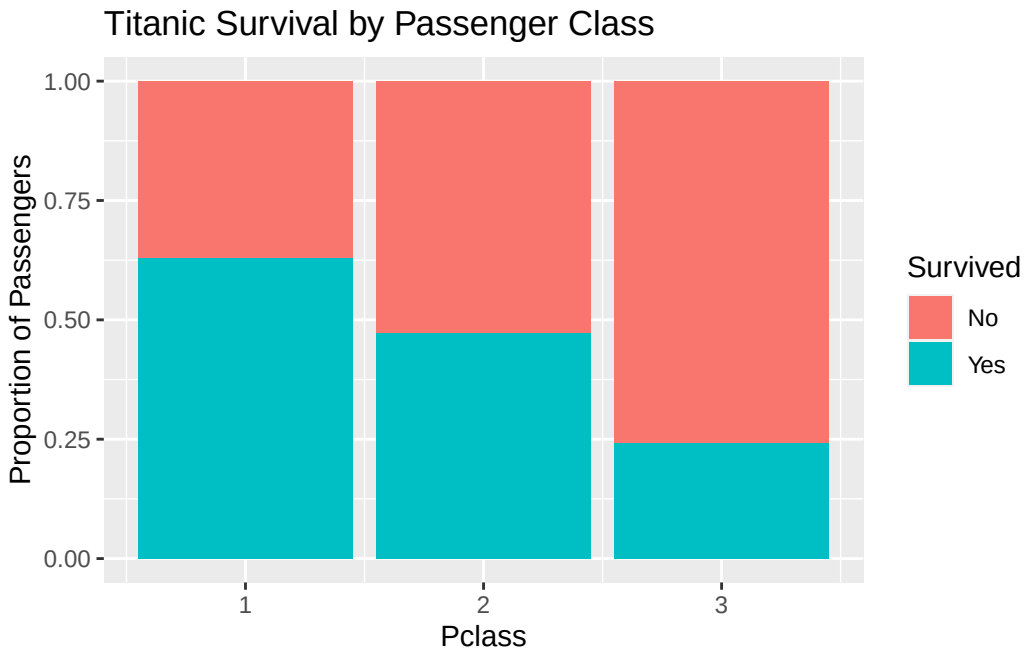
```
p + geom_bar(aes(x = Pclass,
                 fill = factor(Survived,
                 labels = c("No", "Yes")))) + labs(fill = "Survived") + ylab("Number of
```



Again survival rates are lower for 2 and 3 class pasangers, but not as easy to read as the rate bar plot.

Plotting proportions normalized to 1 so have a similar plot oas the rates bar plot.

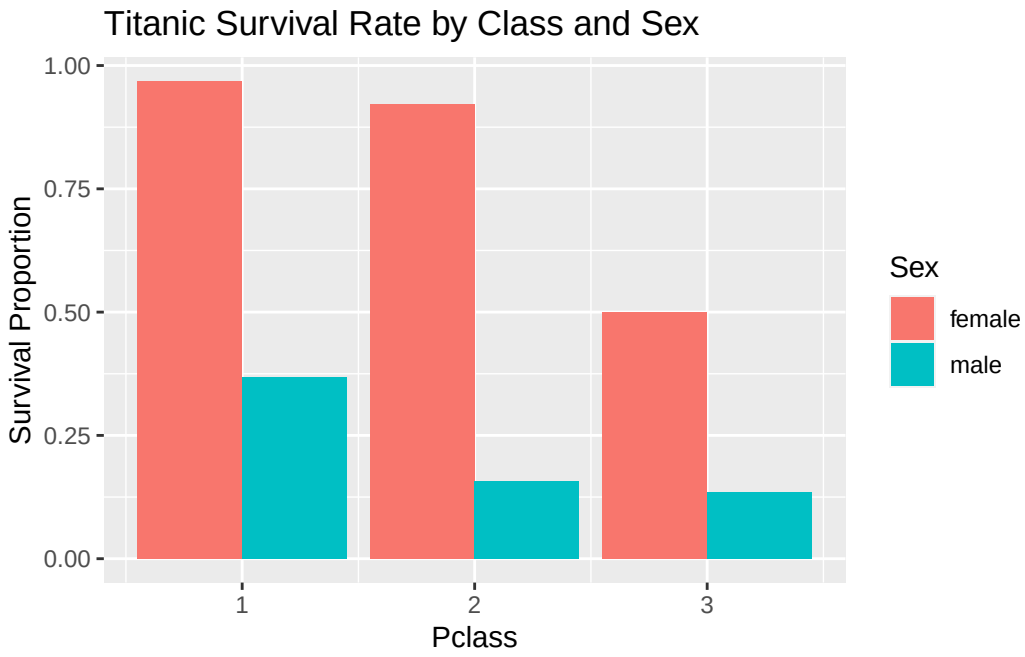
```
p + geom_bar(aes(x = Pclass,
                 fill = factor(Survived,
                               labels = c("No", "Yes"))),
                 position = "fill") +
    labs(fill = "Survived") +
    ylab("Proportion of Passengers") +
    ggtitle("Titanic Survival by Passenger Class")
```



Insight continues showing first class has higher survival. *Second Insight: first class survival is higher than 2nd and 3rd.*

Now we can combine pasanger sex and class together.

```
p + geom_bar(aes(x = Pclass,
                 y = Survived,
                 fill = Sex),
                 position = "dodge",
                 stat = "summary",
                 fun = "mean") +
    ylab("Survival Proportion") +
    ggtitle("Titanic Survival Rate by Class and Sex")
```



Third Insight: females in first and second class had higher survival.

Conclusion

- Children had high survival
- First class survival is higher than 2nd and 3rd
- Females in first and second class had higher survival