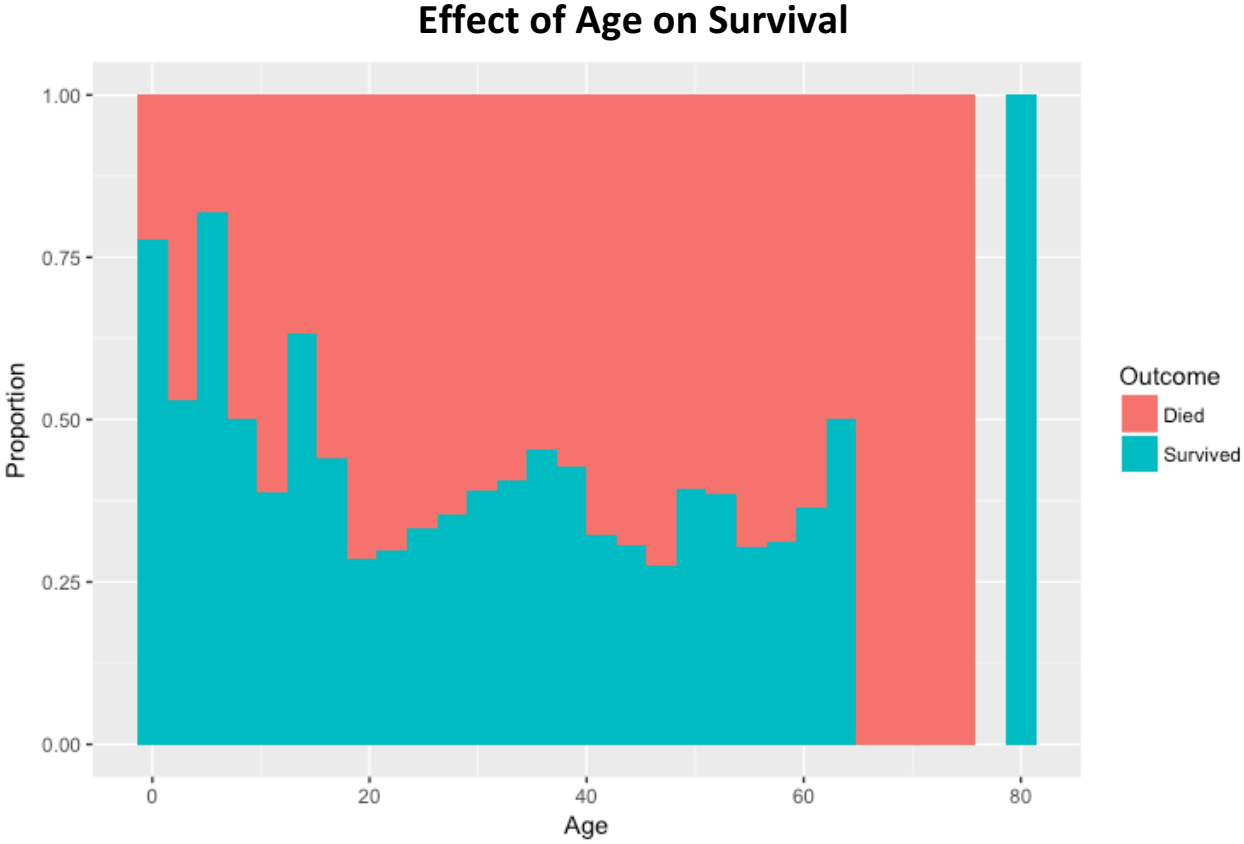
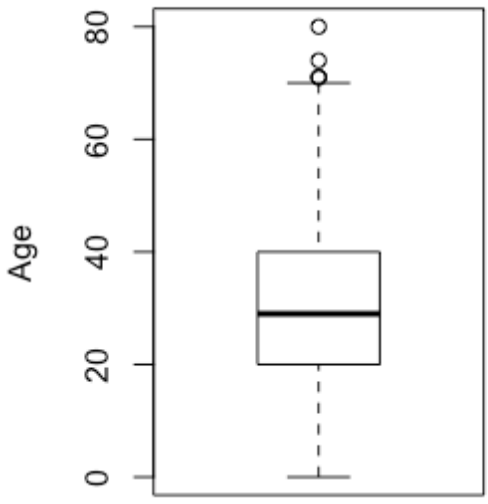
Titanic Passenger Data Analysis

(Q1) Exploratory Analysis

Our aim is to find correlation between different variables and whether a passenger survived or not.

```
library(ggplot2)
data = read.table("/Users/alferinkhenry/Desktop/255\ Extension/titanic.csv", header = TRUE, sep = ',')
Outcome = factor(data$Survived, labels = c('Died','Survived'))
ggplot(data, aes(Age, fill=Outcome)) + geom_histogram(position='fill')+ylab("Proportion")
```

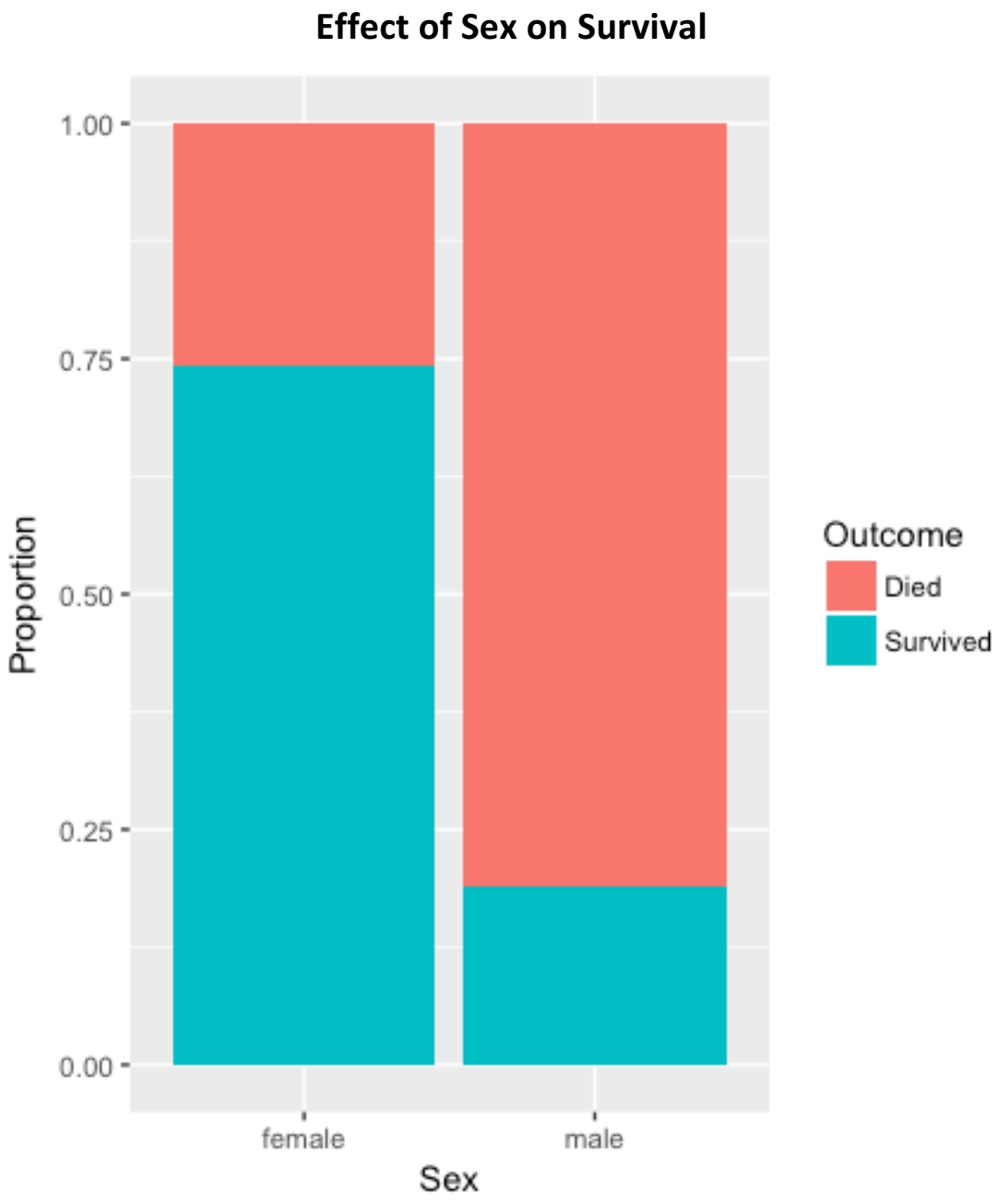




Proportionally, more young people (under 20) survived than the rest, and above the age of 60 just about all died. But most people were between 20 and 40 years old.

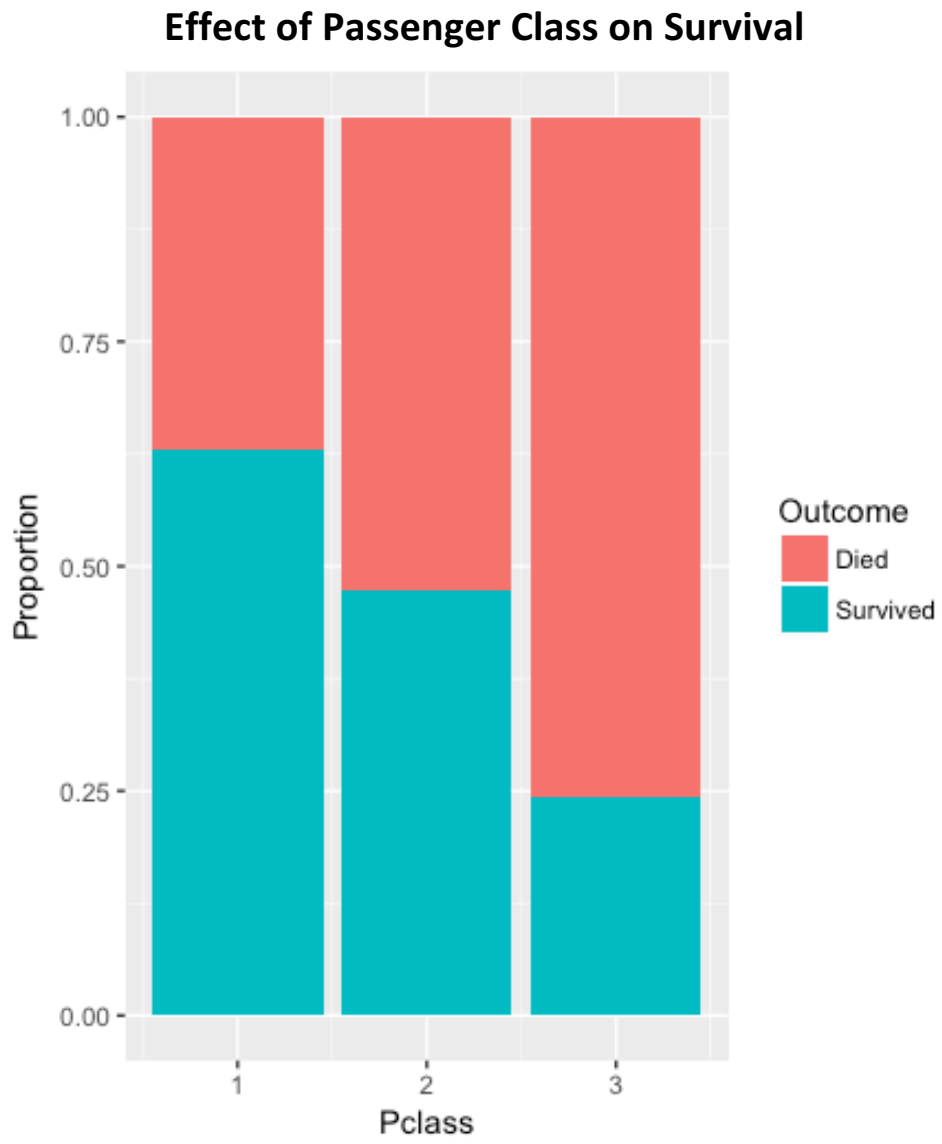
One interesting observation is that between 20 and 40, the proportion is increasing; perhaps this is because those who are younger wouldn't have been able to afford higher class tickets.

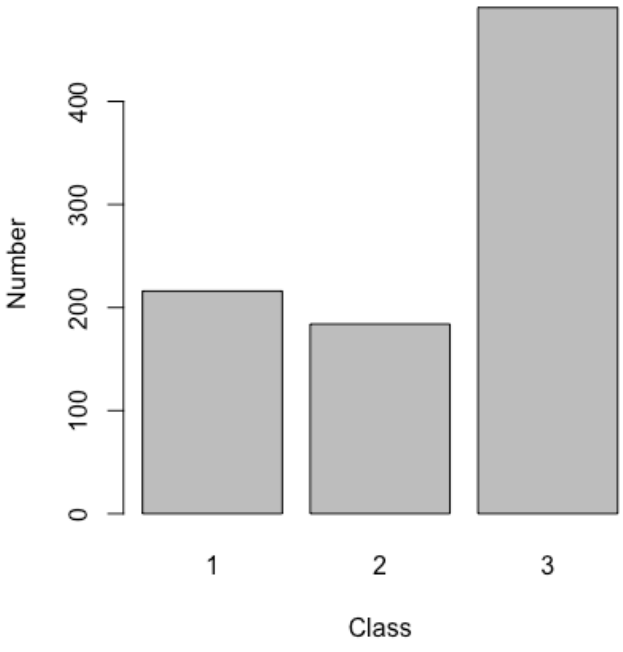
```
ggplot(data,aes(Sex,fill=Outcome)) + geom_bar(position = 'fill') +ylab('Proportion')
```



Perhaps more males died because they felt an obligation to save the females.

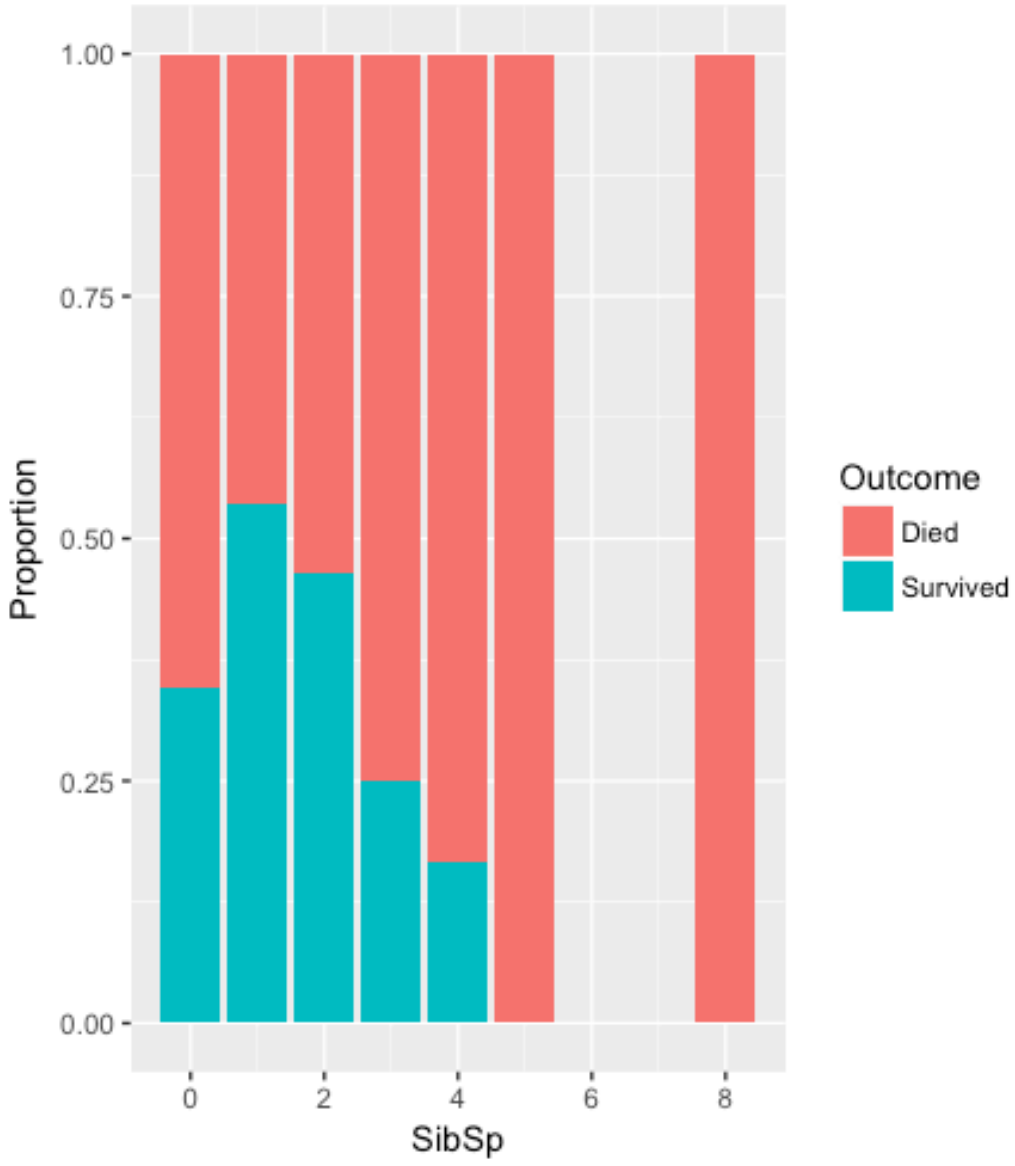
```
ggplot(data,aes(Pclass,fill=Outcome)) + geom_bar(position = 'fill') +ylab('Proportion')
```

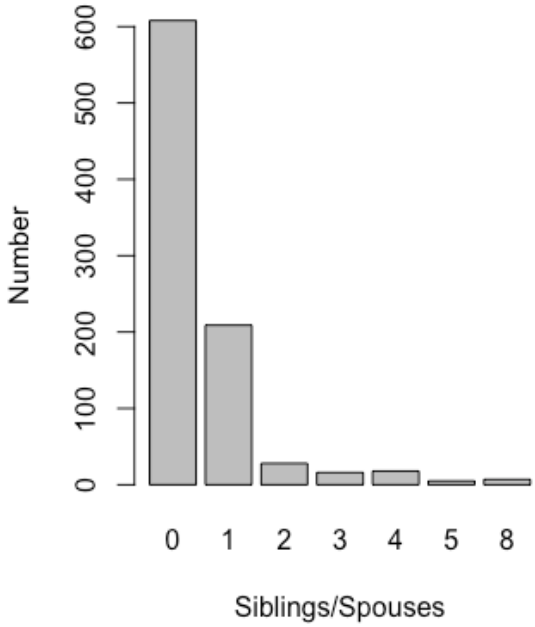




Here we see a clear relationship between passenger class and survival. A person in first class, for example, was more likely to survive than someone in third class.
The graph on the left shows that there is enough data to show this.

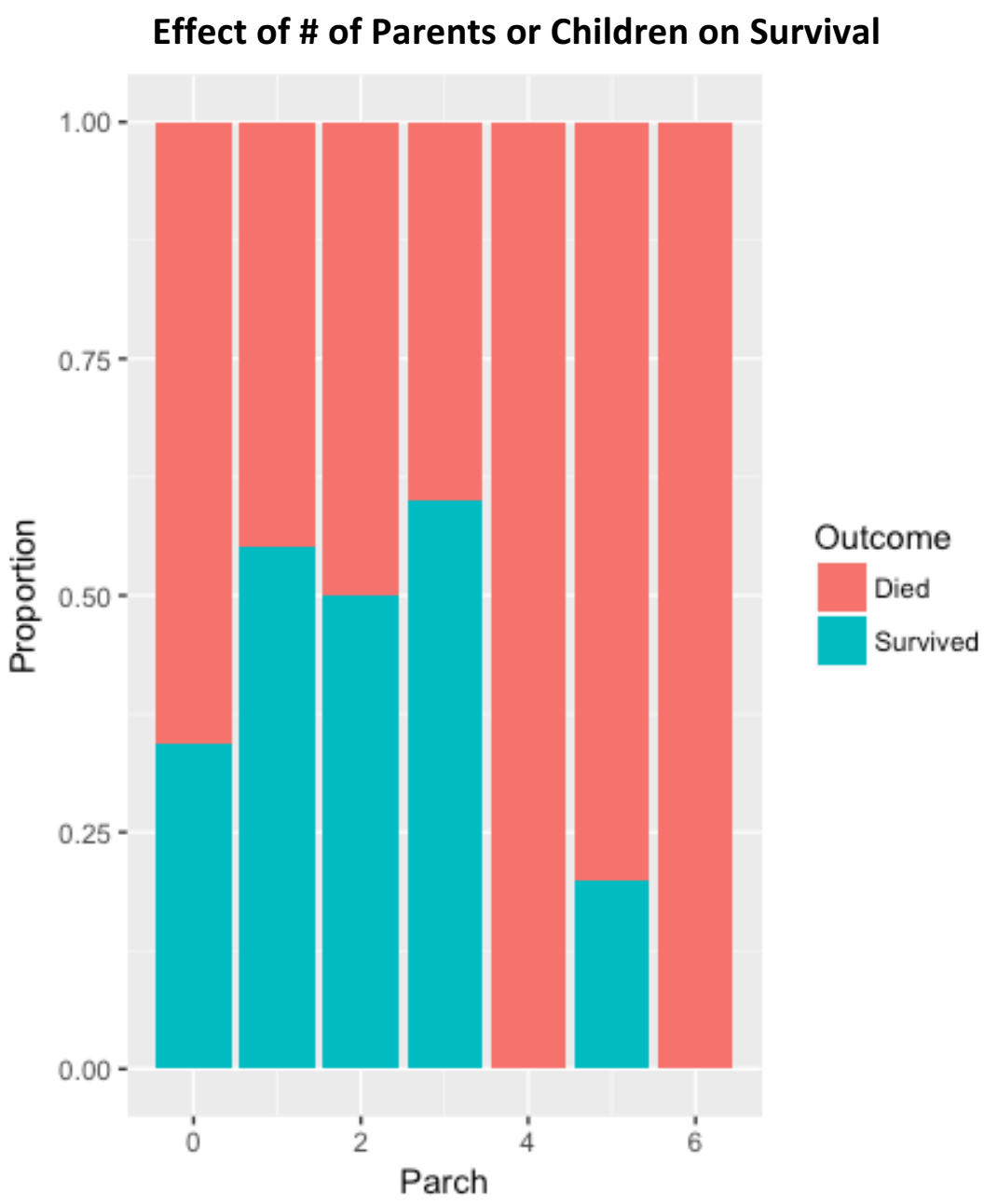
ggplot(data,aes(SibSp,fill=Outcome)) + geom_bar(position = 'fill') +ylab('Proportion')

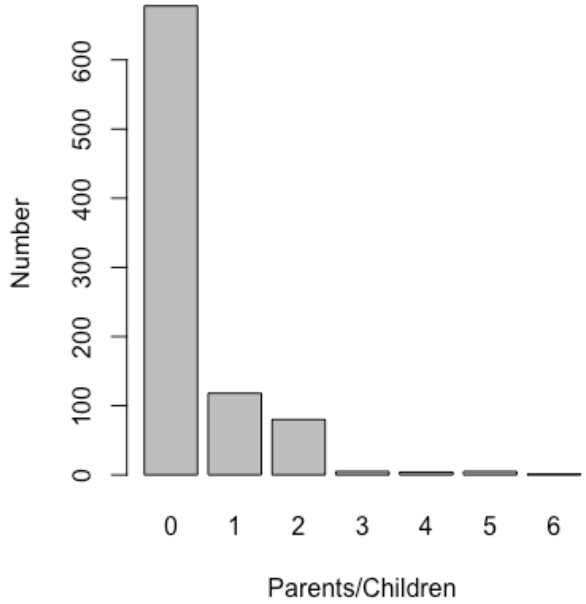




There seems to be a trend that people with one or two siblings or spouses were more likely to survive. In fact, there were a greater proportion of survivors that had one sibling/spouse compared with those who had none.

```
ggplot(data,aes(Parch,fill=Outcome)) + geom_bar(position = 'fill') +ylab('Proportion')
```





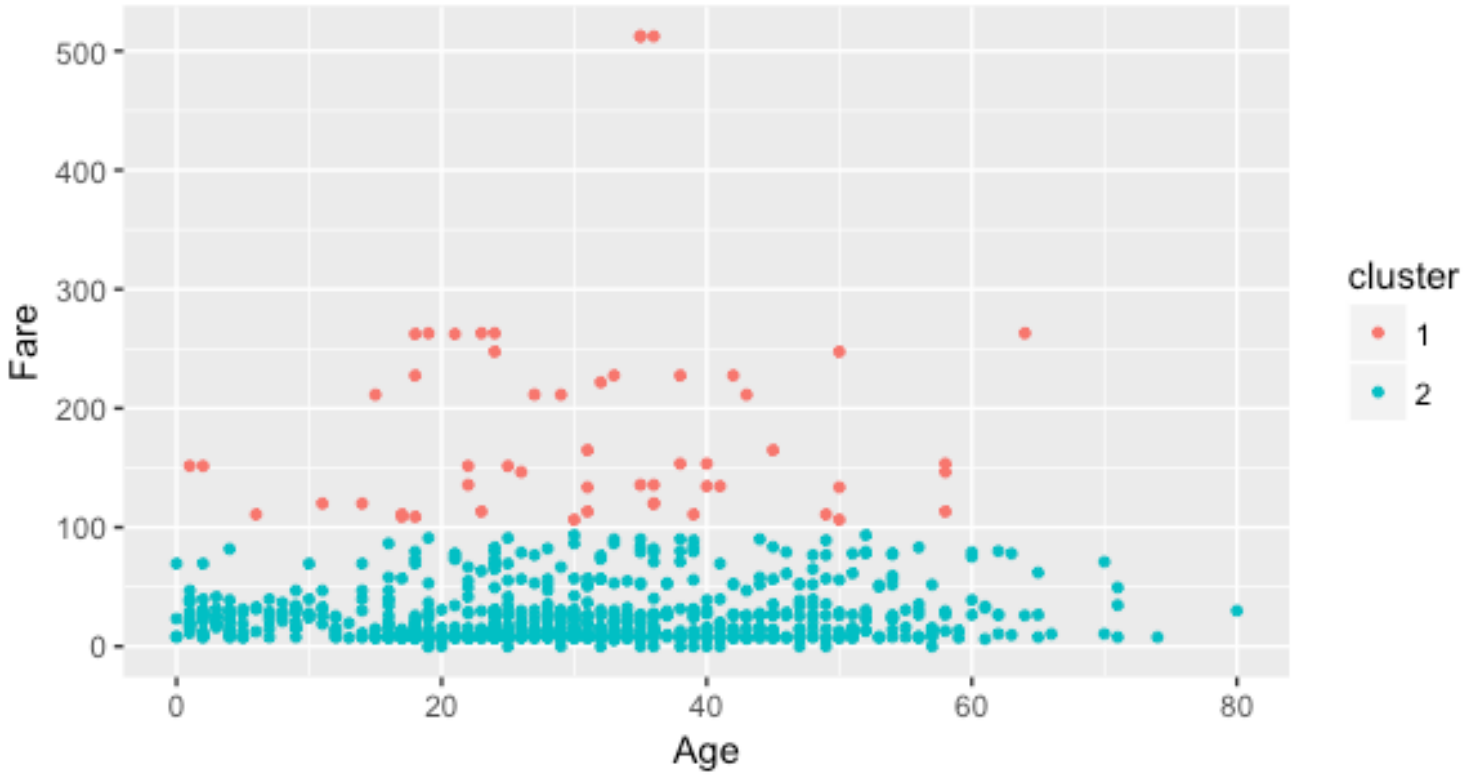
It's interesting that these graphs have a very similar shape as the graphs for the effect of Siblings/Spouses on survival.

(Q2) k-Means Clustering

```
library(ggplot2)
set.seed(99)
mydata = read.table("/Users/alferinkhenry/Desktop/255\ Extension/titanic.csv", header = TRUE, sep = ',')

clusters = kmeans(mydata[c('Age','Fare')], 2, nstart = 20)
mydata$cluster = factor(clusters$cluster)

graph = ggplot() +
  geom_point(data = mydata,
             aes(x = Age,
                 y = Fare,
                 color = cluster),
             size = 1)
graph
```



The data was split at a horizontal line going through Fare=100. Most of the data is at the bottom of the graph at the place of lower cost. If you look at the graph I made above showing how many people there were in each class you see also that most of the passengers were in the third passenger class. Perhaps we should split the data into three clusters instead of two because there are three passenger classes. Currently, cluster 2 probably refers to

(Q2c)

A person that is 30 years old and pays a fare of $20 would be grouped into cluster 2. The probability that they would survive based on the probability that anyone in cluster 2 would survive is 0.36. Code I used to get this result:

```
# take all data points that are in cluster 2.
mysubset = subset(mydata, cluster == 2, select = Survived)
mysubset = unlist(mysubset)
# we can simply sum the whole vector because we only have binary values.
sum(mysubset)/length(mysubset)
```

Additional Clusterings

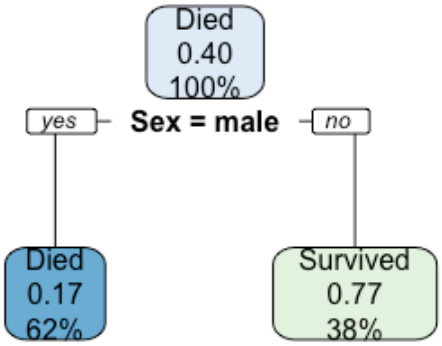
The graph on the left shows the relationship between age and how many siblings/spouses one has on board. The probability of surviving if in cluster 2 is 0.36, and 0.40 if in cluster 1. The clusters don't look very distinct; furthermore, these values don't tell us much because the overall probability of surviving on the titanic is 0.38, and they don't differ much from this value. However, the graph on the right which compares passenger class and how many siblings/spouses a passenger had on board shows a bit more distinct clustering. The probability of surviving in cluster 2 on this graph is 0.40, but in cluster 1 the probability is 0.15, which is unique. While this cluster doesn't group survivors together, it does seem to group a subset of the casualties quite well.

(Q3)

```
library(rpart)
library(rpart.plot)
set.seed(50)
mydata = read.table("/Users/alferinkhenry/Desktop/255\ Extension/titanic.csv", header = TRUE, sep = ',')
train = sample(891,250)
mydata$Survived = factor(mydata$Survived,labels=c('Died','Survived'))
tree = rpart(Survived~Sex,data=mydata,subset=train,method="class")
rpart.plot(tree)
table(Predicted=predict(tree,mydata[-train,],type="class"),Actuals=mydata[-train,"Survived"])
```

Sex as Independent Variable

Tree Using Training Data



Test Data vs Actual

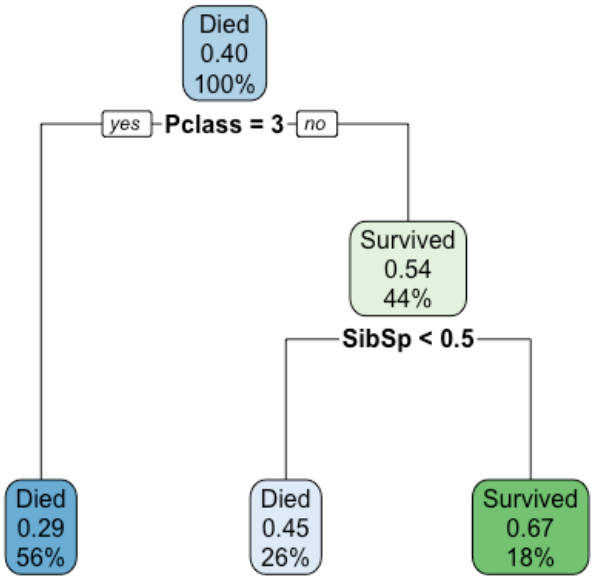
	Actuals	
Predicted	Died	Survived
Died	340	82
Survived	59	160

In the above tree, the root node shows that 40% of the people (from the training data) died; only 17% of the males survived, but 77% of the females survived. So, the tree is telling us that if someone is male, they are most likely to have died, and if female, they are most likely to have survived. Looking at the confusion matrix, if we add up the values in the 'correctly predicted' diagonal and divide by the total values, we get an indication of how accurately the tree classifies data. For the above,

$$\text{Accuracy} = \text{Correct/Total} = 0.78,$$

Which is an 'ok' result.

Pclass and SibSp as Independent Variables

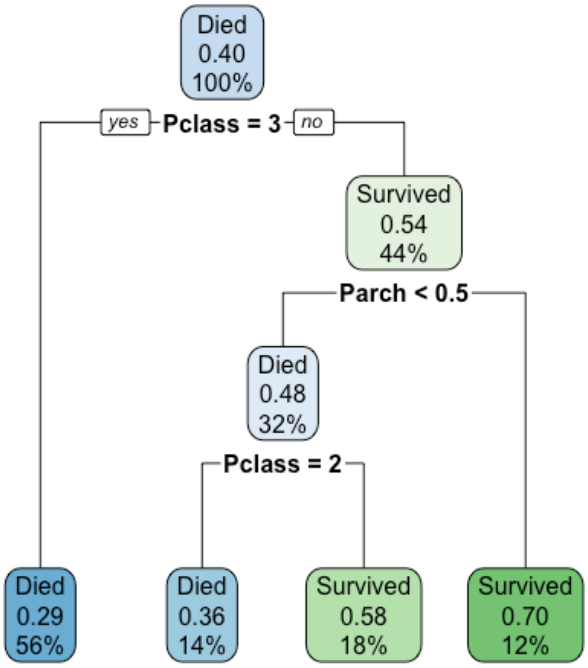


Test Data vs Actual

	Actuals	
Predicted	Died	Survived
Died	367	176
Survived	32	66

Accuracy = 433/641 = 0.68

Pclass and Parch as Independent Variables



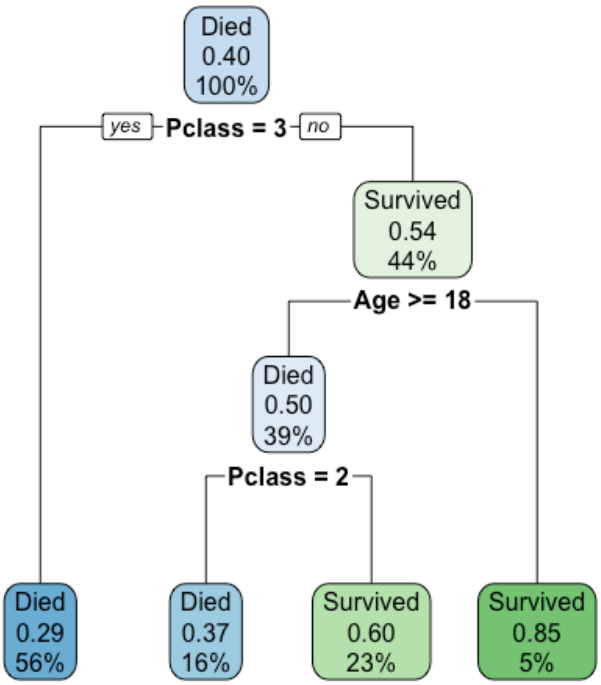
Test Data vs Actual

	Actuals	
Predicted	Died	Survived
Died	336	114
Survived	63	128

Accuracy = 0.72

Pclass, Age, and Parch as Independent Variables

(Using maxdepth = 3)



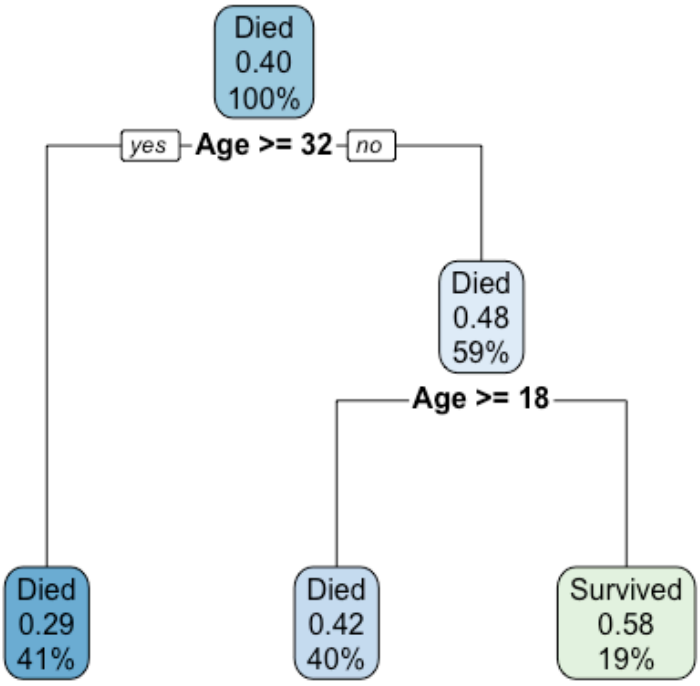
Test Data vs Actual

	Actuals	
Predicted	Died	Survived
Died	342	128
Survived	57	114

Accuracy = 0.71

Age as Independent Variable

(Using maxdepth = 2)



Test Data vs Actual

	Actuals	
Predicted	Died	Survived
Died	350	185
Survived	49	57

Accuracy = 0.63

Out of all the 5 tests above using decision trees, the most accurate tree was the one that used only Sex as the independent variable with an accuracy rating of 0.78, the next most accurate was the one that used Pclass and Parch as independent variables. So from the information I have acquired with the trees, probably the most useful one will be the one that classifies based on Sex.