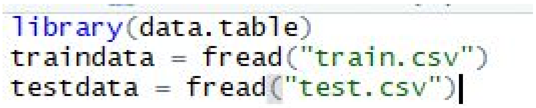
R version 3.4.3

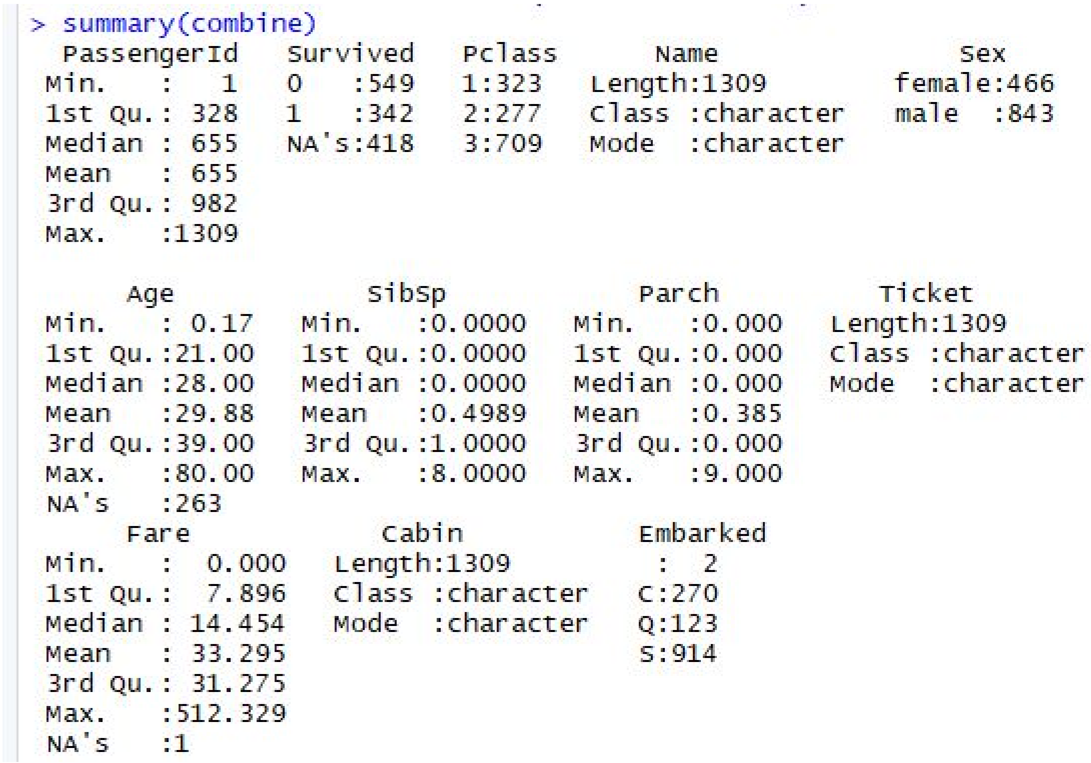
R packages used: data.table, ggplot2, dplyr, corrplot, forcats, stringr.

Analysis based on combination of Train.csv and Test.csv

To read the files, I used fread which is usually helpful in reading BIG files.

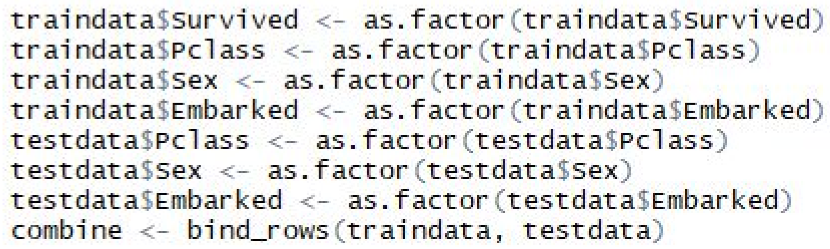


```
library(data.table)
traindata = fread("train.csv")
testdata = fread("test.csv")
```



```
> summary(combine)
  PassengerId     Survived    Pclass       Name                Sex     
 Min.   :   1    0    :549   1:323   Length:1309        female:466  
 1st Qu.: 328    1    :342   2:277   Class :character   male  :843  
 Median : 655    NA's:418    3:709   Mode  :character               
 Mean   : 655                                                       
 3rd Qu.: 982                                                       
 Max.   :1309                                                       
                                                                    
      Age            SibSp            Parch          Ticket         
 Min.   : 0.17   Min.   :0.0000   Min.   :0.000   Length:1309       
 1st Qu.:21.00   1st Qu.:0.0000   1st Qu.:0.000   Class :character  
 Median :28.00   Median :0.0000   Median :0.000   Mode  :character  
 Mean   :29.88   Mean   :0.4989   Mean   :0.385                     
 3rd Qu.:39.00   3rd Qu.:1.0000   3rd Qu.:0.000                     
 Max.   :80.00   Max.   :8.0000   Max.   :9.000                     
 NA's   :263                                                        
      Fare             Cabin            Embarked
 Min.   :  0.000   Length:1309          :  2   
 1st Qu.:  7.896   Class :character   C:270   
 Median : 14.454   Mode  :character   Q:123   
 Mean   : 33.295                      S:914   
 3rd Qu.: 31.275                              
 Max.   :512.329                              
 NA's   :1                                    
```

Before combining the files, I changed the variables to factors.

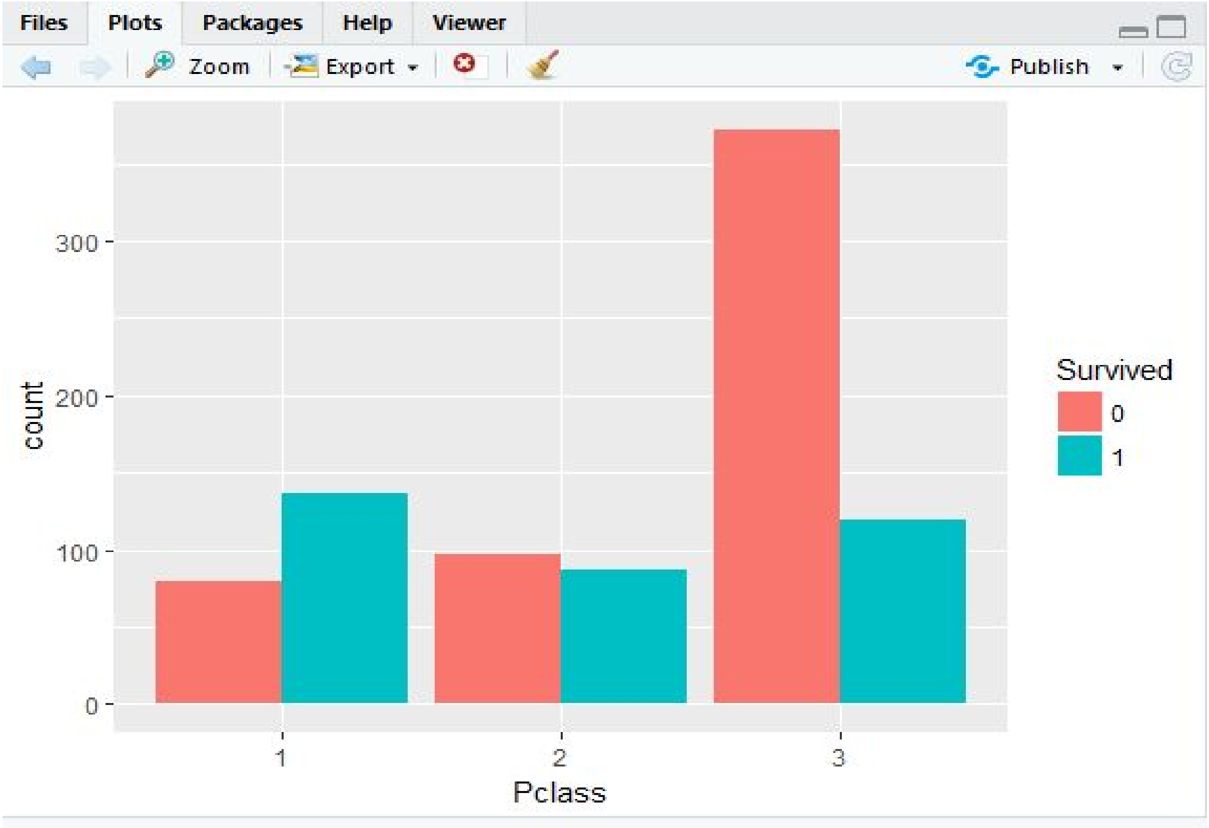


```
traindata$Survived <- as.factor(traindata$Survived)
traindata$Pclass <- as.factor(traindata$Pclass)
traindata$Sex <- as.factor(traindata$Sex)
traindata$Embarked <- as.factor(traindata$Embarked)
testdata$Pclass <- as.factor(testdata$Pclass)
testdata$Sex <- as.factor(testdata$Sex)
testdata$Embarked <- as.factor(testdata$Embarked)
combine <- bind_rows(traindata, testdata)
```

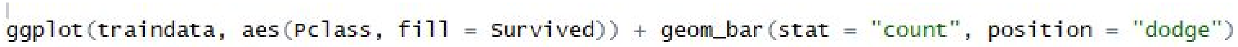
Each variable in the files is analyzed. From this summary, we can tell there are more males than females; there are missing data from the Survived, Age and Fare variables (this is because the test file excludes the Survived data). And clearly, A LOT of people died.

Analysis of Train.csv File

Number of survivors and nonsurvivors per cabin class.



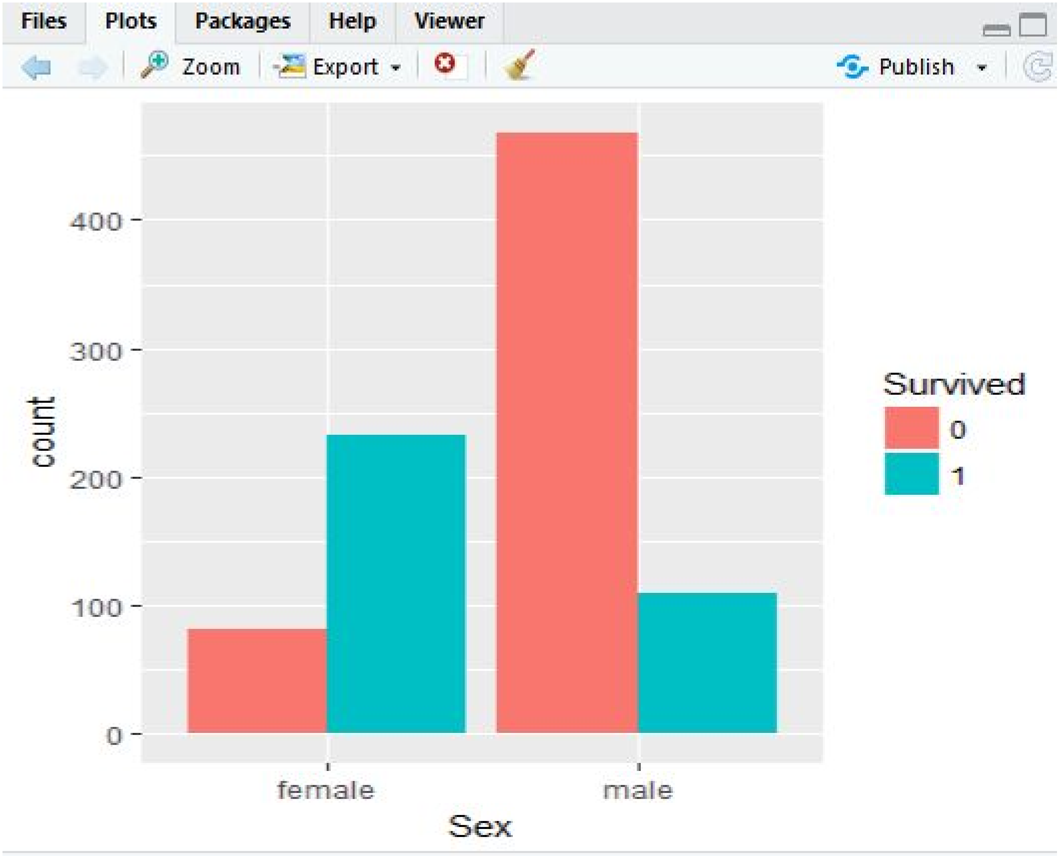
This bar graph was created using ggplot.



```
ggplot(traindata, aes(Pclass, fill = Survived)) + geom_bar(stat = "count", position = "dodge")
```

More people survived from 1st class than 2nd and 3rd. Most of the fatalities were from 3rd class. As seen in the summary, there were a lot of 3rd class passengers so it makes sense that a lot of 3rd class passengers died. And as seen in the movie, 3rd class passengers were treated like cow poop.

Number of survivors per sex

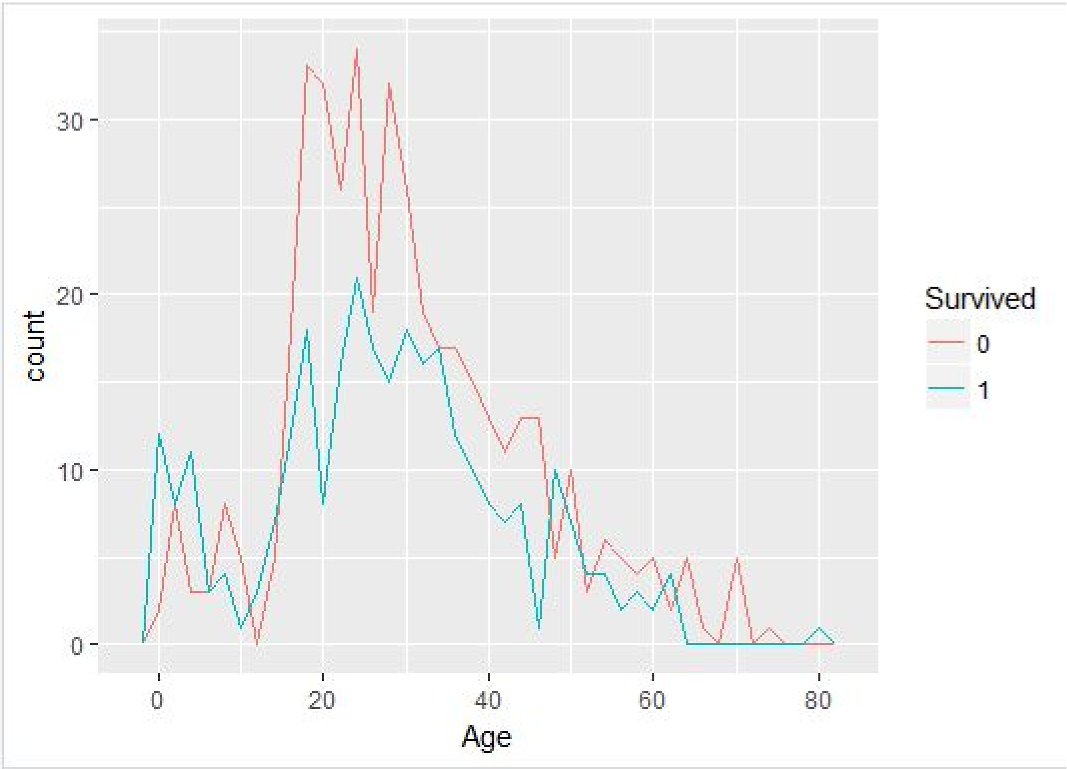


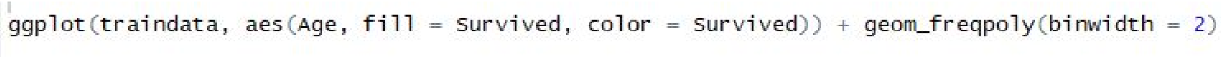
The ggplot function is also used here. It is a very useful function from the ggplot2 package. It can also be used to create different types of graphs which you will see soon.

```
ggplot(traindata, aes(Sex, fill = Survived)) + geom_bar(stat = "count", position = "dodge")
```

More males than females died in the disaster. Not only because there are more male passengers than females, but also because women and children were allowed on the lifeboats first.

Distribution of survivors based on age



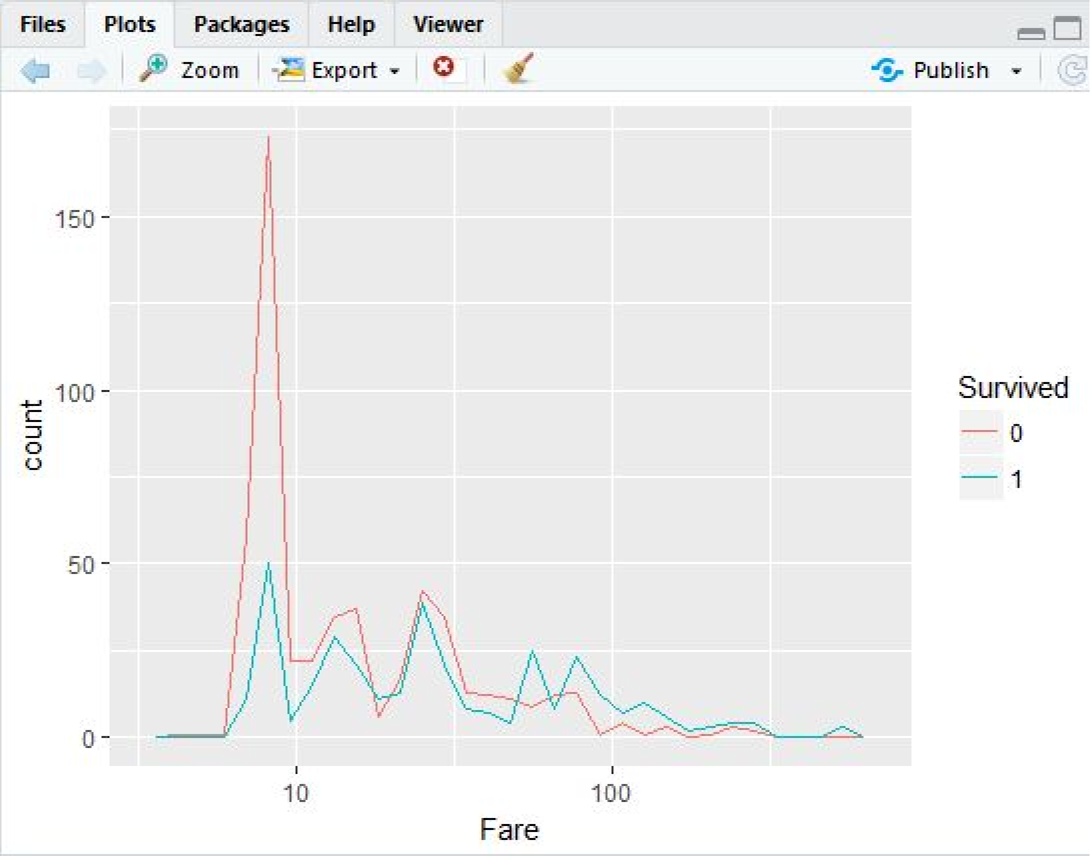


```
ggplot(traindata, aes(Age, fill = Survived, color = Survived)) + geom_freqpoly(binwidth = 2)
```

As shown above, passengers' age ranges from 20-30. This makes a lot of sense because the mean age of passengers is 29.88. Passengers in this age range have a lower survival rate then those ranging from 0-6.

P.S this graph is called a frequency polygon which is similar to a histogram and is really helpful in viewing the shape of a distribution and comparing sets of data.

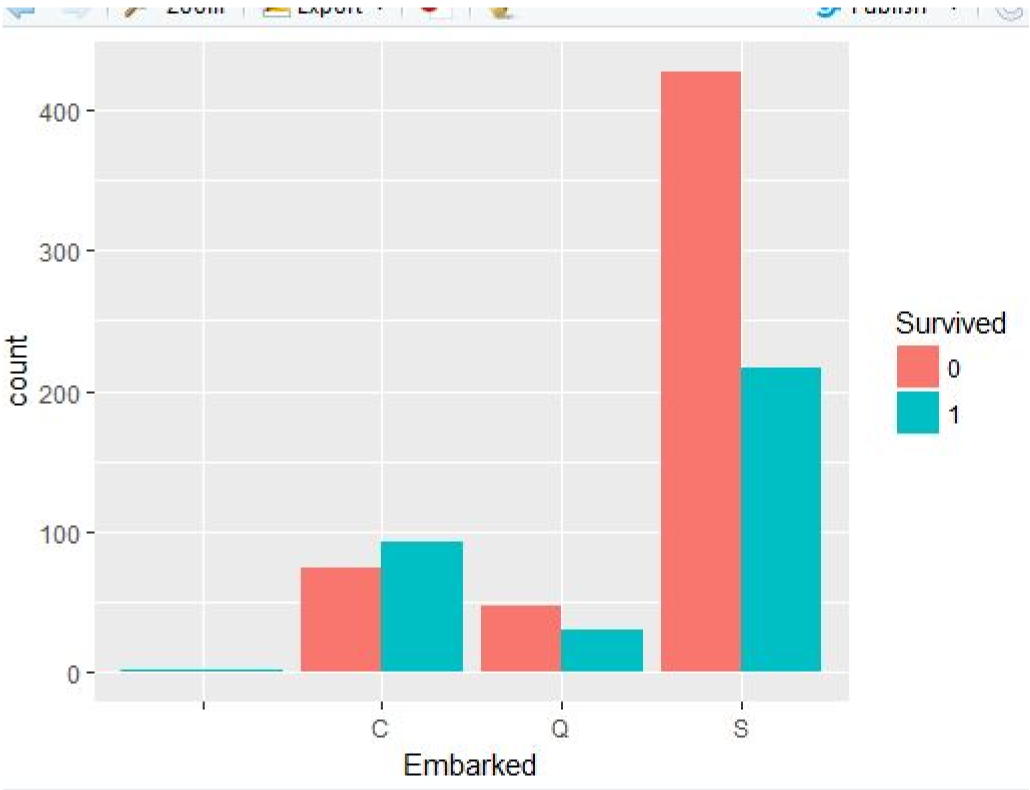
Distribution of survivors based on fare



```
ggplot(traindata, aes(Fare, fill = Survived, color = Survived)) + geom_freqpoly(binwidth = 1) + scale_x_log10()
```

It is obvious that those that paid the least are from the 3rd class. Third class had the most passengers and were not given top priority to access lifeboats.

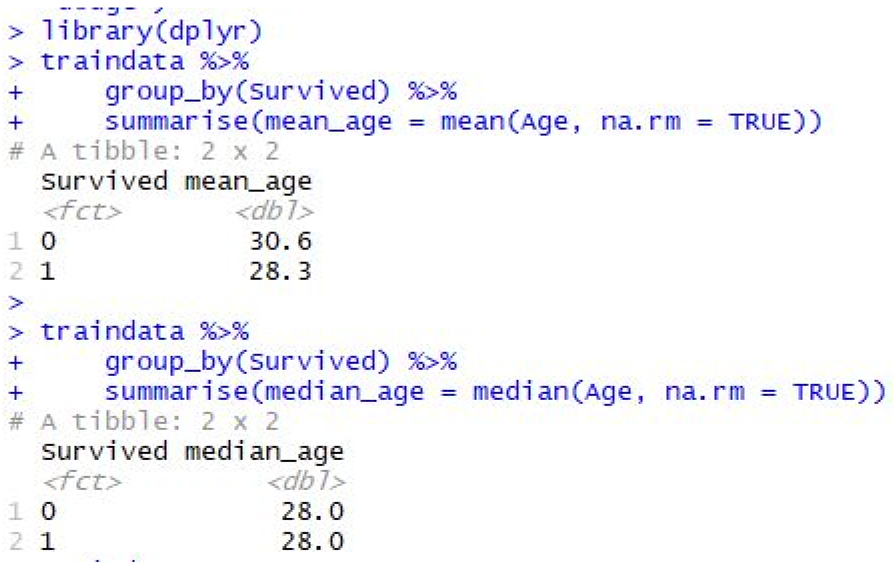
Number of survivors per port of embarkment



```
ggplot(traindata, aes(Embarked, fill = Survived)) + geom_bar(stat = "count", position = "dodge")
```

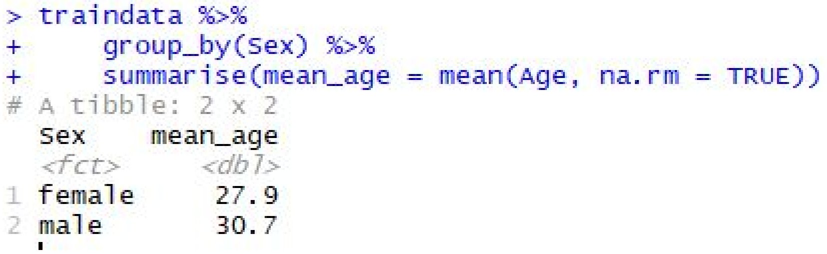
There is missing data for a few of the passengers, hence the empty value. A LOT of people embarked from Southampton. Southampton is the port of origin before stopping at Cherbourg and finally Queenstown.

Average and Median Age for Survivors & Nonsurvivors



```
> library(dplyr)
> traindata %>%
+     group_by(Survived) %>%
+     summarise(mean_age = mean(Age, na.rm = TRUE))
# A tibble: 2 x 2
  Survived mean_age
  <fct>       <dbl>
1 0            30.6
2 1            28.3
>
> traindata %>%
+     group_by(Survived) %>%
+     summarise(median_age = median(Age, na.rm = TRUE))
# A tibble: 2 x 2
  Survived median_age
  <fct>         <dbl>
1 0              28.0
2 1              28.0
```

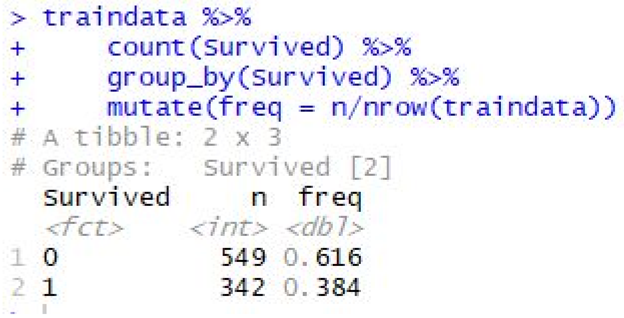
As seen in my previous analysis, there were a lot of young passengers (20- 30). The mean age for fatalities is 30.6 and that for survivors is 28.3. My guess is maybe because the men were generally older than the female and we know more men died (will analyze this next). The median age is 28 which further supports the fact that there were a lot of young passengers.



```
> traindata %>%
+     group_by(Sex) %>%
+     summarise(mean_age = mean(Age, na.rm = TRUE))
# A tibble: 2 x 2
  Sex    mean_age
  <fct>     <dbl>
1 female     27.9
2 male       30.7
```

Above is the mean age for men and women. As previously stated, male passengers were older.

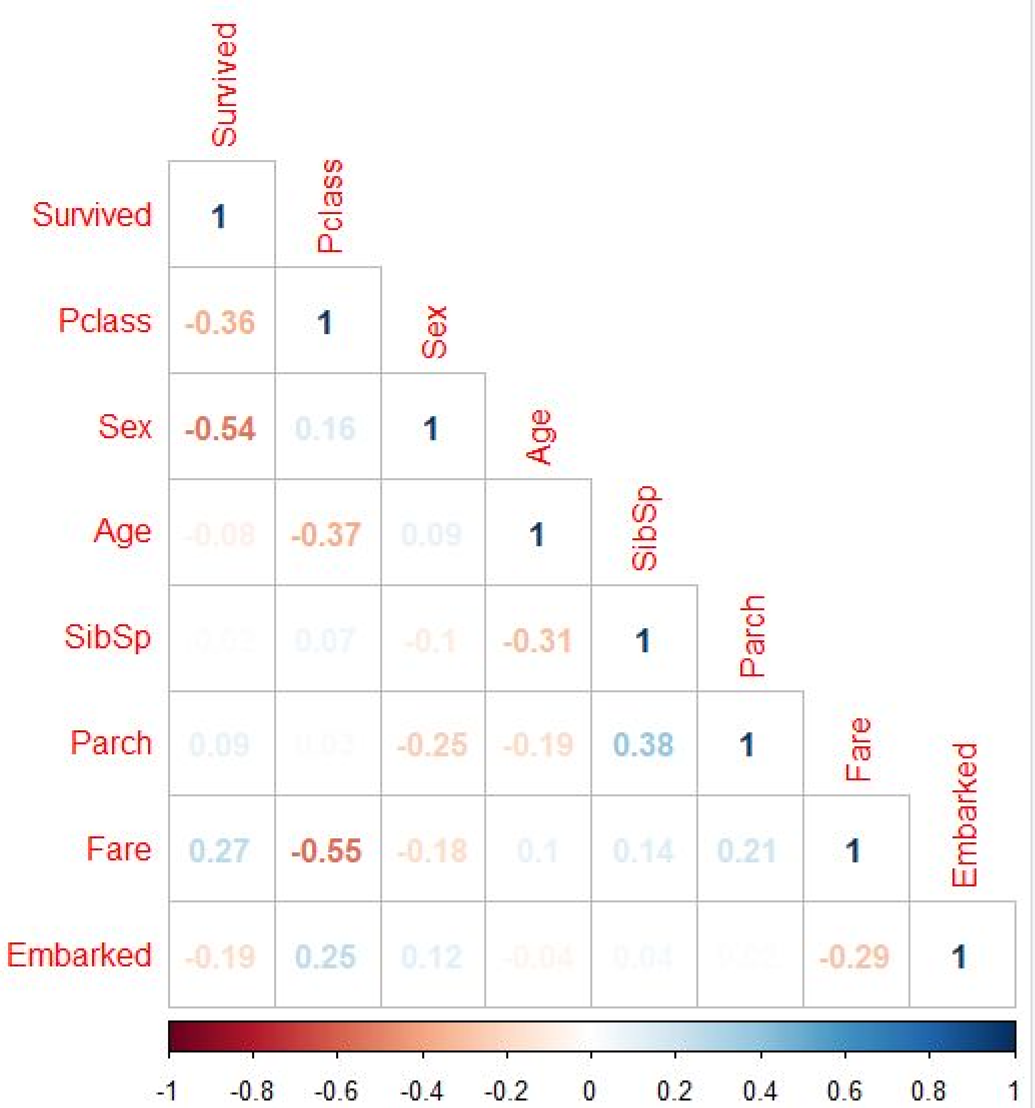
Percentage of Survivors & Nonsurvivors

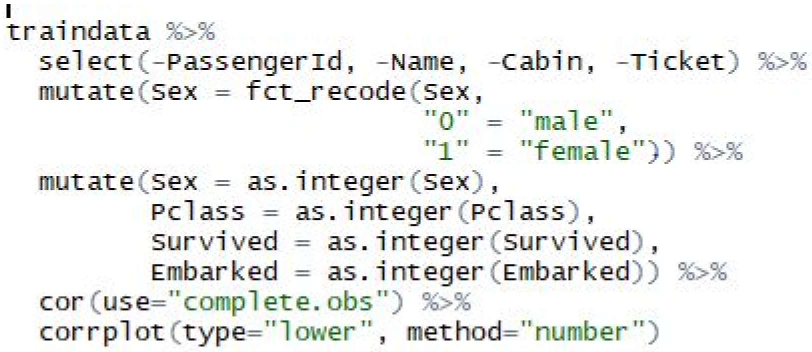


```
> traindata %>%
+     count(Survived) %>%
+     group_by(Survived) %>%
+     mutate(freq = n/nrow(traindata))
# A tibble: 2 x 3
# Groups:   Survived [2]
  Survived     n  freq
  <fct>    <int> <dbl>
1 0          549 0.616
2 1          342 0.384
```

61.6% died and 38.4% survived

Correlation between variables



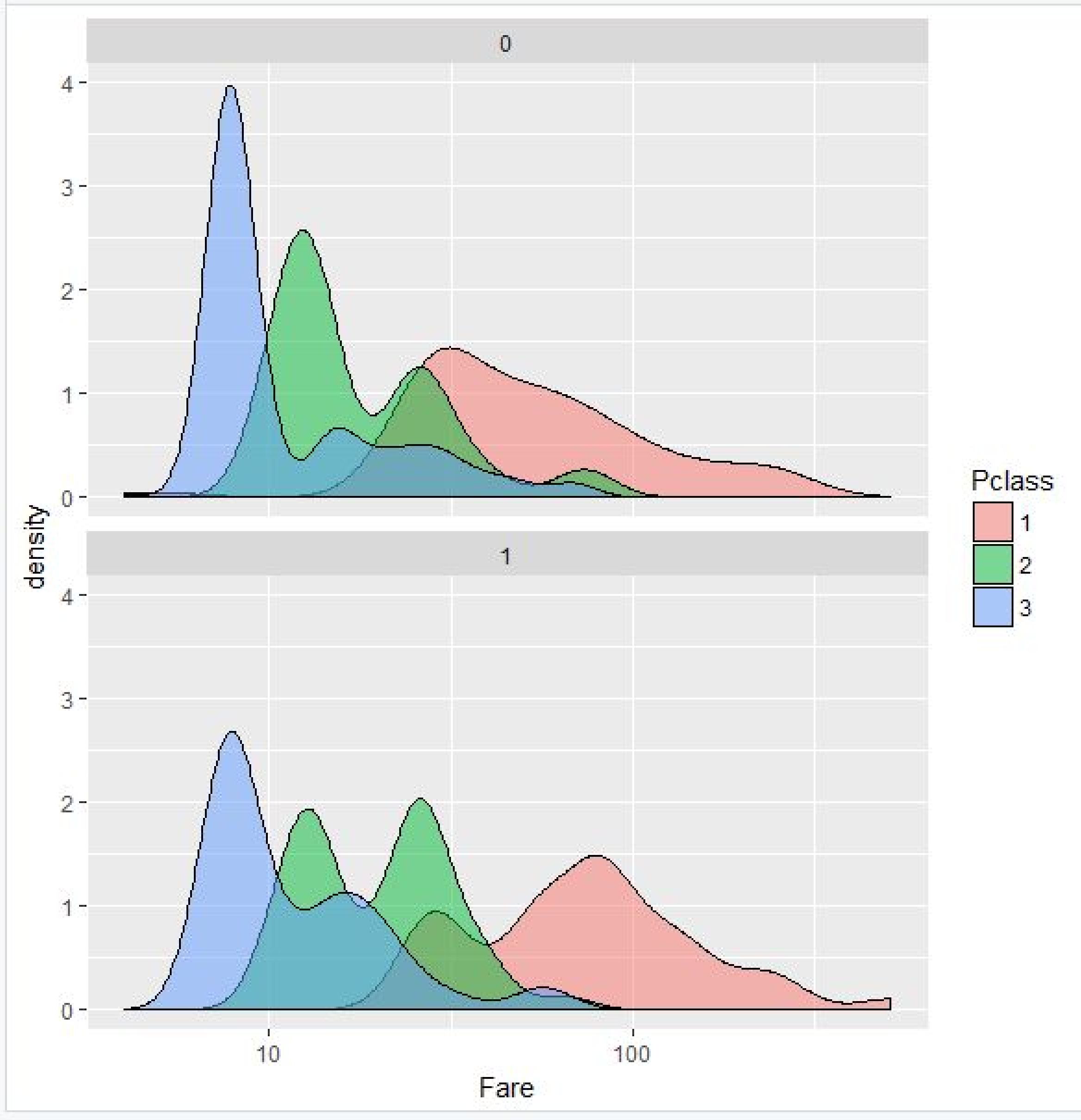


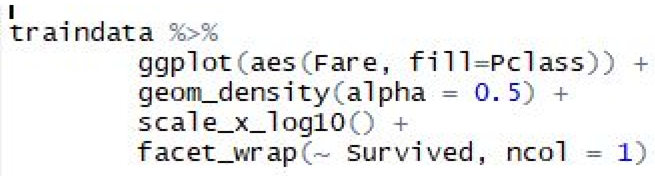
```
traindata %>%
  select(-PassengerId, -Name, -Cabin, -Ticket) %>%
  mutate(Sex = fct_recode(Sex,
                          "0" = "male",
                          "1" = "female")) %>%
  mutate(Sex = as.integer(Sex),
         Pclass = as.integer(Pclass),
         Survived = as.integer(Survived),
         Embarked = as.integer(Embarked)) %>%
  cor(use="complete.obs") %>%
  corrplot(type="lower", method="number")
```

Survived correlates the most with Sex, followed by Pclass. Pclass correlates the most with Fare, followed by Age. SibSp correlates the most by Parch (I am guessing because the more siblings, parents you have, the bigger the family).

Analyzing how the correlation between variables affects the behaviour of the Survived variable

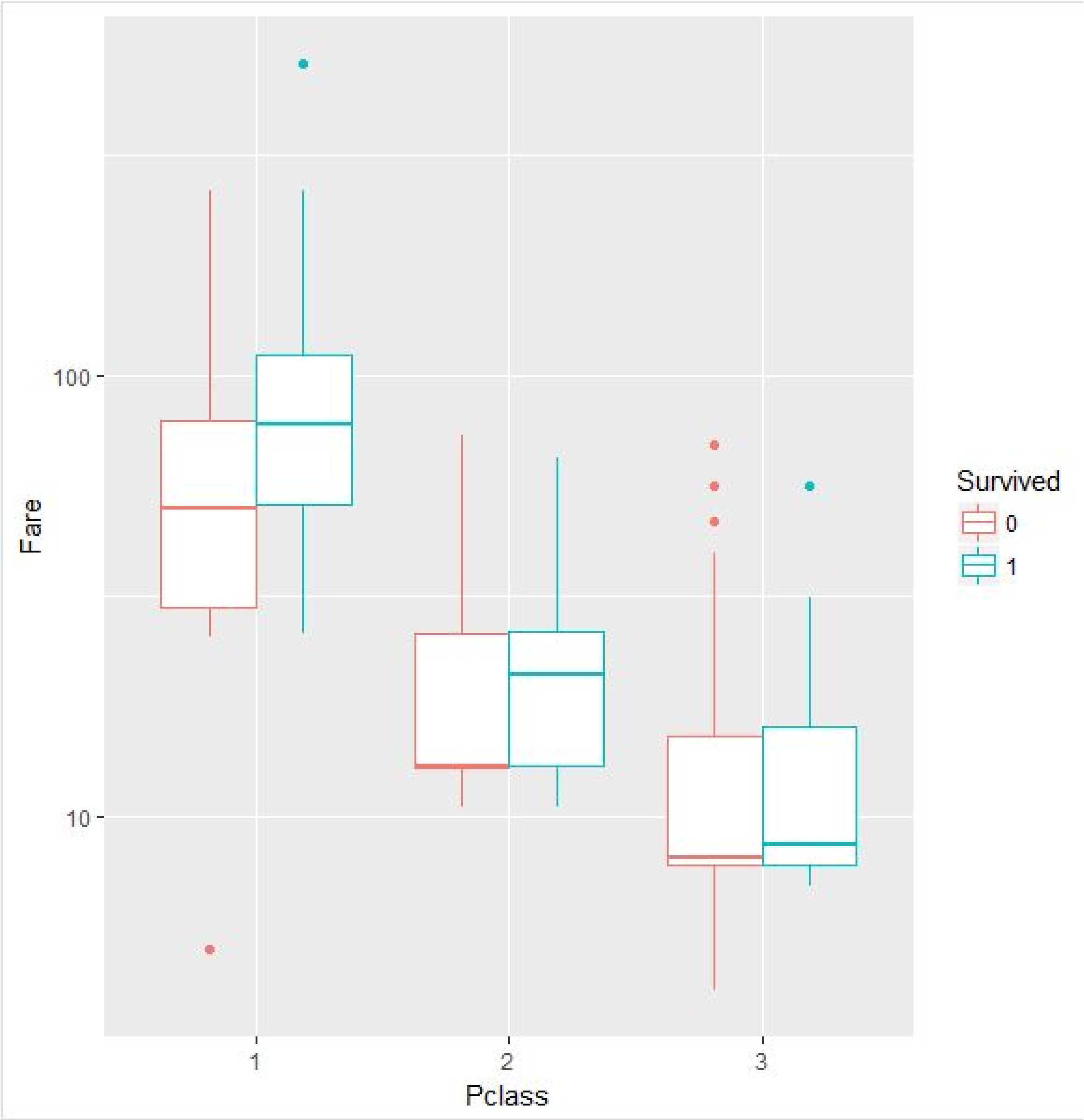
1. Pclass vs. Fare

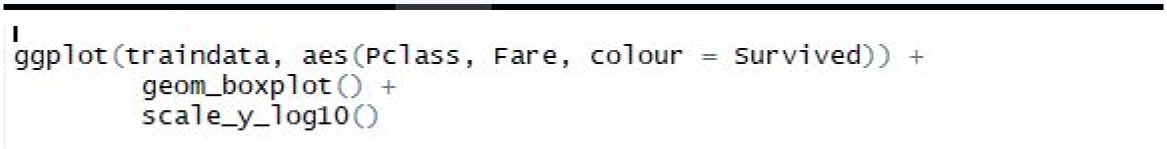




```
traindata %>%
        ggplot(aes(Fare, fill=Pclass)) +
        geom_density(alpha = 0.5) +
        scale_x_log10() +
        facet_wrap(~ Survived, ncol = 1)
```

It is obvious that passengers that paid the highest fare were in the 1st class cabins and those with the lowest were in the 3rd class cabins. In the top graph of those who did not survive, the three plots gravitate towards the left of the x-axis which again supports the notion that the richer you are, the higher your survival rate. Furthermore, there are variations in fare price within the classes with some outliers. For instance, it appears that a few passengers in 3rd class paid more than some passengers in 2nd class. I am assuming this has to do with where in the ship their cabins were situated (closer to lifeboats). The graph below further analyzes these variations.

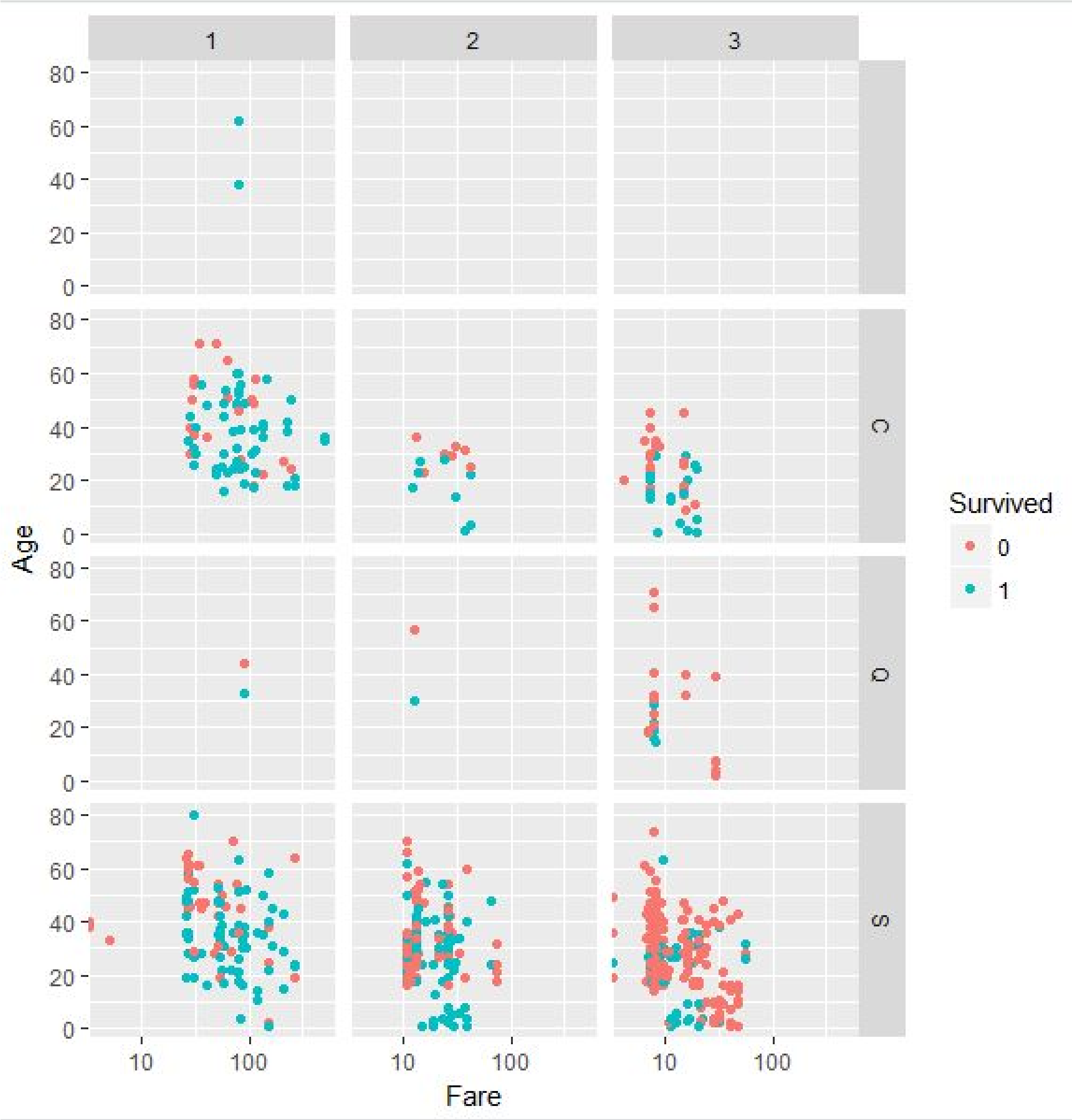


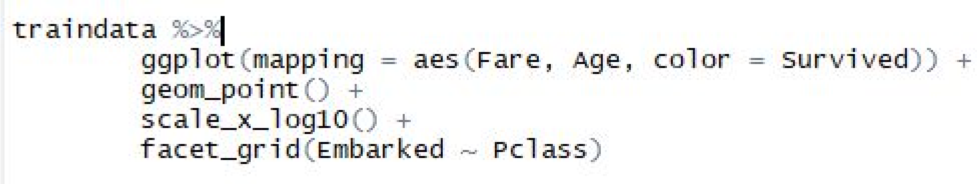


```
ggplot(traindata, aes(Pclass, Fare, colour = Survived)) +
        geom_boxplot() +
        scale_y_log10()
```

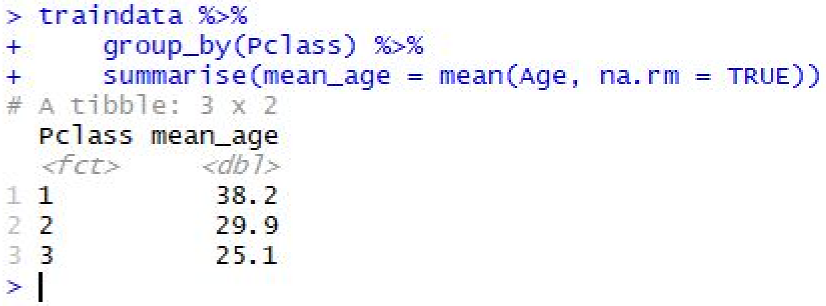
This box plot shows the minimum & maximum values, 1st & 3rd quartiles, and the median value of the fare for each class filled with the Survived variable. The dots outside the boxes are outliers. The median fare of those who survived is higher than that of those who did not survive which again suggests that within each class, there are difference fare prices depending on where the cabins are situated.

2. Fare, Age, Pclass, Embarked

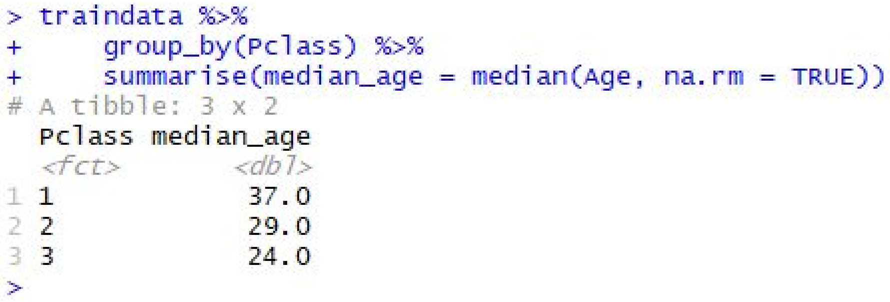



```
traindata %>%
        ggplot(mapping = aes(Fare, Age, color = Survived)) +
        geom_point() +
        scale_x_log10() +
        facet_grid(Embarked ~ Pclass)
```

There is a significant difference in the upper left corner to the lower right corner. Those with 1st class ticket were generally older than those with 2nd and 3rd class.



```
> traindata %>%
+     group_by(Pclass) %>%
+     summarise(mean_age = mean(Age, na.rm = TRUE))
# A tibble: 3 x 2
  Pclass mean_age
  <fct>     <dbl>
1 1          38.2
2 2          29.9
3 3          25.1
> 
```

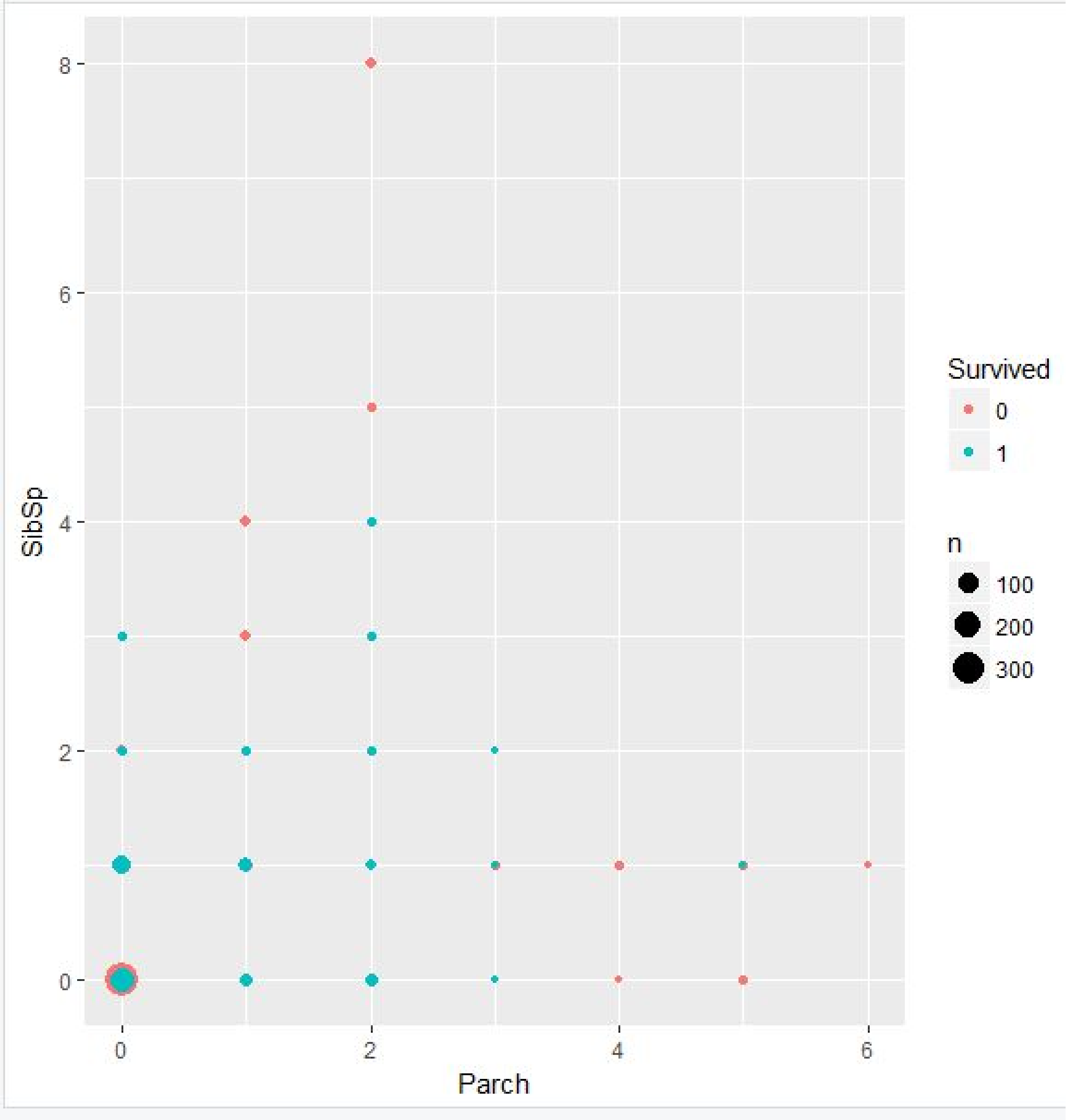


```
> traindata %>%
+     group_by(Pclass) %>%
+     summarise(median_age = median(Age, na.rm = TRUE))
# A tibble: 3 x 2
  Pclass median_age
  <fct>       <dbl>
1 1            37.0
2 2            29.0
3 3            24.0
> 
```

And as stated numerous times, A LOT of third class passengers died. The port of embarkment does not play a role in the survival rate of the passengers, but since a lot of people embarked from Southampton, it makes sense that a lot of those passengers died. It also appears that there weren’t a lot of 2nd class passengers.

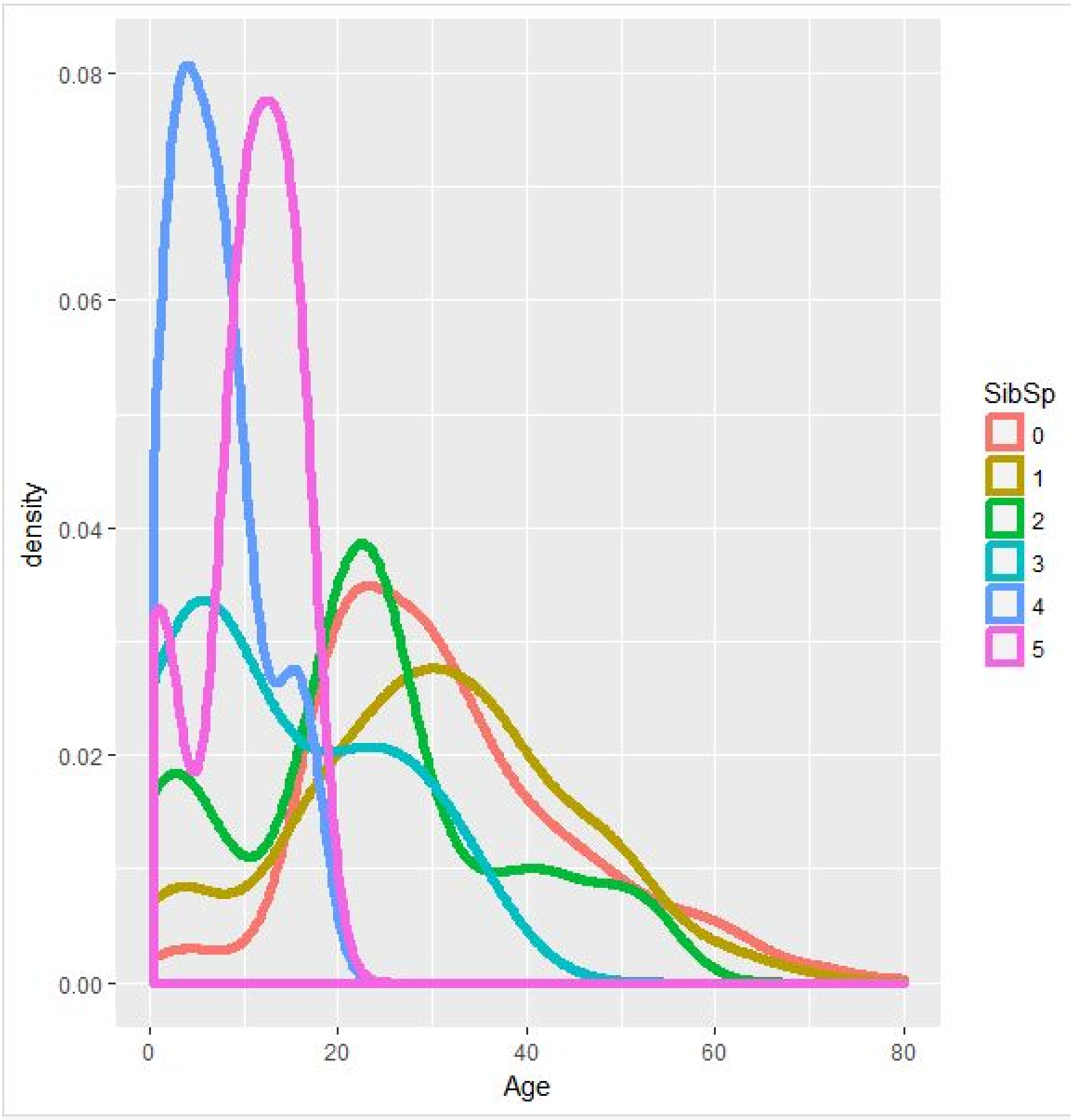
Family

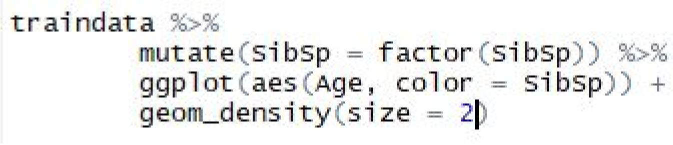
1. SibSp & Parch



Illustrated by the two large circles in the bottom left, a lot of passengers travelled alone and a lot of them died. It also appears that people with smaller families have a higher rate of survival than those with large families.

2. Age & SibSp

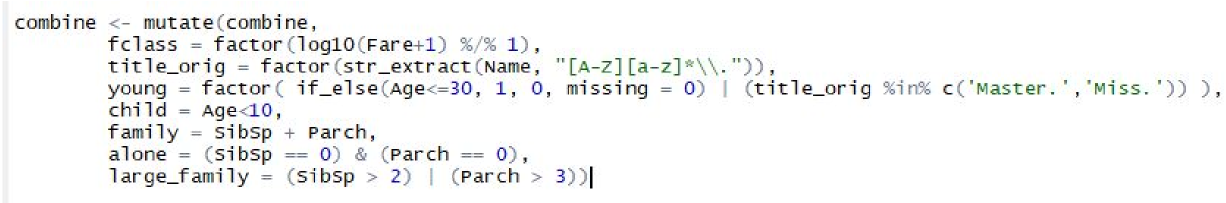




```
traindata %>%
        mutate(SibSp = factor(SibSp)) %>%
        ggplot(aes(Age, color = SibSp)) +
        geom_density(size = 2)
```

The distribution of those with more Siblings is narrow and concentrated in the lower age range (children travelling with family).

Adding new variables



```
combine <- mutate(combine,
        fclass = factor(log10(Fare+1) %/% 1),
        title_orig = factor(str_extract(Name, "[A-Z][a-z]*\\.")),
        young = factor( if_else(Age<=30, 1, 0, missing = 0) | (title_orig %in% c('Master.','Miss.')) ),
        child = Age<10,
        family = SibSp + Parch,
        alone = (SibSp == 0) & (Parch == 0),
        large_family = (SibSp > 2) | (Parch > 3))
```

Added new variables to aid in the analysis. First line makes it easier to group the fare values into low, medium, and high. Also added young, child, family, alone, and large_family.