"Titanic" by Mark Peters

I mentioned during last week's class that I waste a lot of time looking for datasets that I want to work on. There are many reasons behind this, both functional and preferential, but the main one is the overwhelming number of potential datasets to choose from. Being the kind of person who has difficulty choosing a meal in a restaurant, this is tortuous for me.

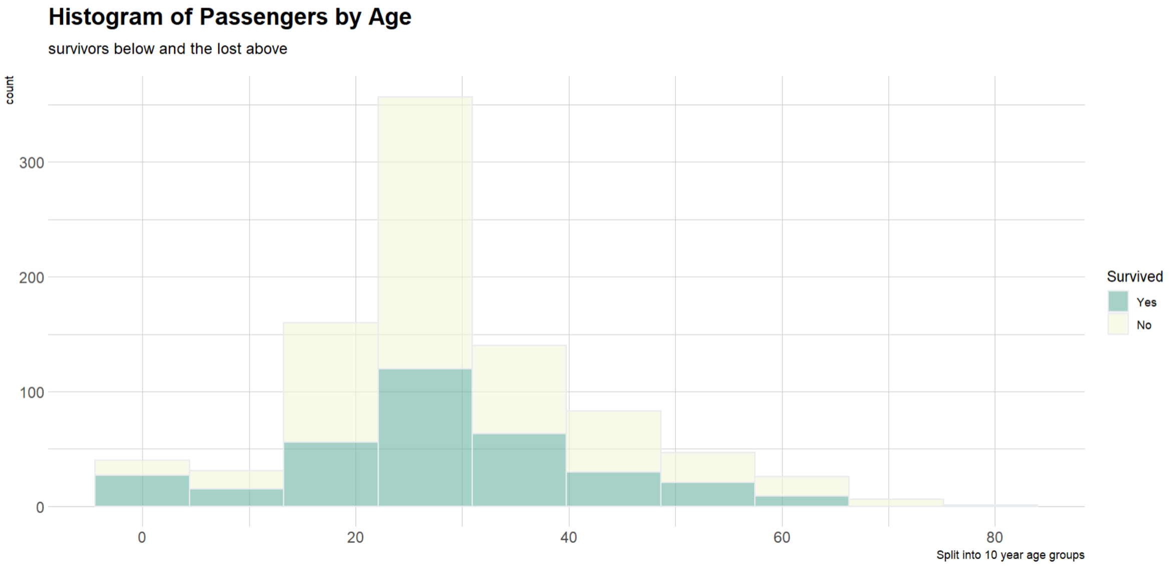
Which is why I was so happy to find this 'Titanic' dataset. I'm not sure if it contains all the features necessary to cover the range of statistical calculations that this course requires, but I'll do my best to make it work even if the fit isn't quite right.

The main reason I chose this dataset is that I have a distant relative who died on the Titanic. No word of a lie, here is a link to a 'fandom' page if you don't believe me. We both come from Ballydrehid which is a small village in Co. Tipperary, Ireland. Her name was Catherine 'Katie' Peters, and an interesting side note is that my niece has the same name as her.

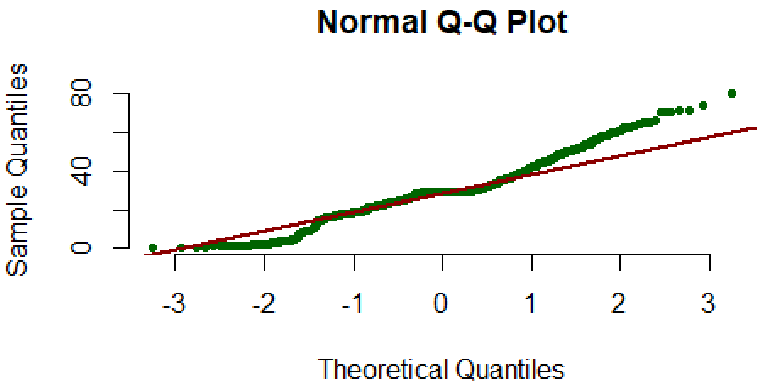
In 1912, the Titanic sank on her maiden voyage with the loss of most of its passengers and crew. This is a dataset with details of 891 out of the 1502 passengers that set out on that fateful voyage. Unfortunately, we don't have any information for the crew and many of the passengers, but it is a big enough set to ask and explore some interesting questions about who survived, why they survived, and why others didn't. And those are the questions I'd like to concentrate on for this assignment.

The headings for the collected data are Passenger ID (**numeric**), whether they survived (**categorical/logical**), ticket class (**ordinal**), Name (**categorical**), Sex (**categorical**), Age (**qualitative/nominal**), number of siblings/spouses onboard (**quantitative**), number of parents/children onboard (**quantitative**), Ticket number (**ordinal**), Fare (**quantitative**), Cabin (**ordinal**), Embarked (**nominal**).

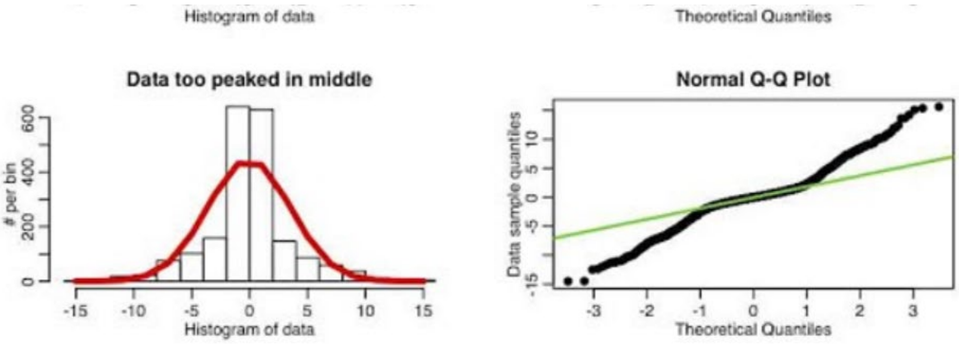
Normal Distribution Curve

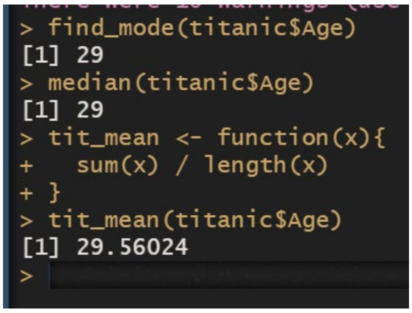


My first plot was a histogram of the range of ages of all passengers, it was interesting to note that the graph displayed (generally) a normal distribution, though it is heavily concentrated in the centre.



I also carried out a **QQ plot** and found that the points fell along or close to the line indicating normal distribution, the rise of the line to the right indicates that the data is very high in the centre.



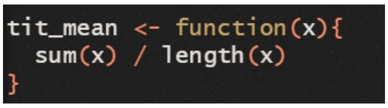


```
There were 10 warnings (use
> find_mode(titanic$Age)
[1] 29
> median(titanic$Age)
[1] 29
> tit_mean <- function(x){
+   sum(x) / length(x)
+ }
> tit_mean(titanic$Age)
[1] 29.56024
>
```

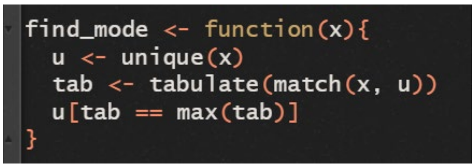
The normal curve result was further supported by exploring the '**central tendency'** with the mode, mean and median values for passenger ages. Both the mode and the median were **29**, with the mean just half a point higher at **29.56**.

To find the mode I used the function you supplied in class; the median can be found by sorting the set of numbers and finding the number at the midpoint (see caption for how I calculated in R); the mean can be found by adding all the values in the set and dividing by the number of said values (see caption for how I did it in R).

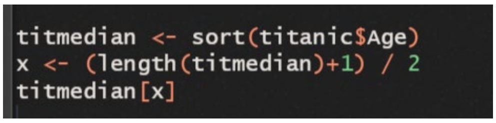
$$\bar{y} = \frac{\sum y}{n}$$



```
tit_mean <- function(x){
  sum(x) / length(x)
}
```



```
find_mode <- function(x){
  u <- unique(x)
  tab <- tabulate(match(x, u))
  u[tab == max(tab)]
}
```



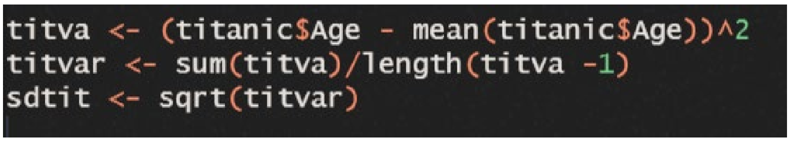
```
titmedian <- sort(titanic$Age)
x <- (length(titmedian)+1) / 2
titmedian[x]
```

Note that as my set of values were odd, I added one to the length and divided by two to get the midpoint, but if it were even I would have had to sum the two midpoint values and halve the total (see attached photo).

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th} \text{ observation}}{2}$$

The **mode, median and mean** for the average age of the **survivors** is almost the same: mode = 29, median

= 29, mean = **28.44**. Only the mean has changed, just over a year lower than the mean for the pre-sinking figure.



```
titva <- (titanic$Age - mean(titanic$Age))^2
titvar <- sum(titva)/length(titva -1)
sdtit <- sqrt(titvar)
```

The variance of ages on boarding is **169**, and the standard deviation **13**. I wrote some clumsy code to achieve (roughly) the same figures as using the built-in functions.

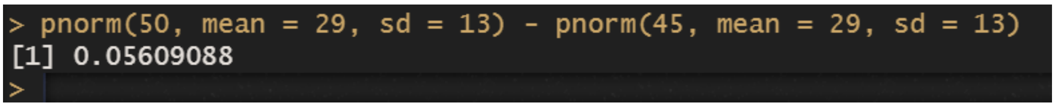
Taking what we have learned we can then calculate the **coefficient of variation** (for the original passenger complement):

$$CV\ (\%) = \left(\frac{Standard\ deviation}{Mean}\right) \times 100$$

CV = (13 / 29.56) * 100 = **43.97%.** This reflects the range of ages represented on board, from new-borns to octogenarians.

We can extrapolate some interesting predictions from these results, such as that **68%** of the passengers were between the ages of **16** and **42**, which we get from adding and subtracting one standard deviation from the mean of 29.

The variance and standard deviation of the survivor's ages are roughly the same as the full complement. Due to the deaths of many young men the range variance goes from **189** to **169**, but the sd remains the same at **13**.

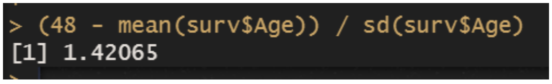


```
> pnorm(50, mean = 29, sd = 13) - pnorm(45, mean = 29, sd = 13)
[1] 0.05609088
>
```

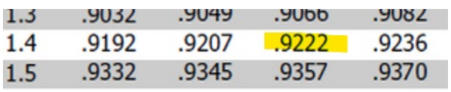
Following on from what we did in class, I ran a 'pnorm' for anyone on the ship around my own age (just for fun). I checked to see what the percentage of those onboard between the ages of 45 and 50 was; the answer was 5.5%, the same calculation for survivors was 5%, though the odds of me surviving depended much more on my social class than my age group (and gender as well of course).

Z-Score

Using the z-score calculation I checked for the percentage of survivors who would have been younger than me (assuming I survived, of course, which as we've seen would have been highly unlikely). Using the z-score equation $z = (x - \mu)/\sigma$ with the survivor/age dataset I get a whopping score of:



```
> (48 - mean(surv$Age)) / sd(surv$Age)
[1] 1.42065
```

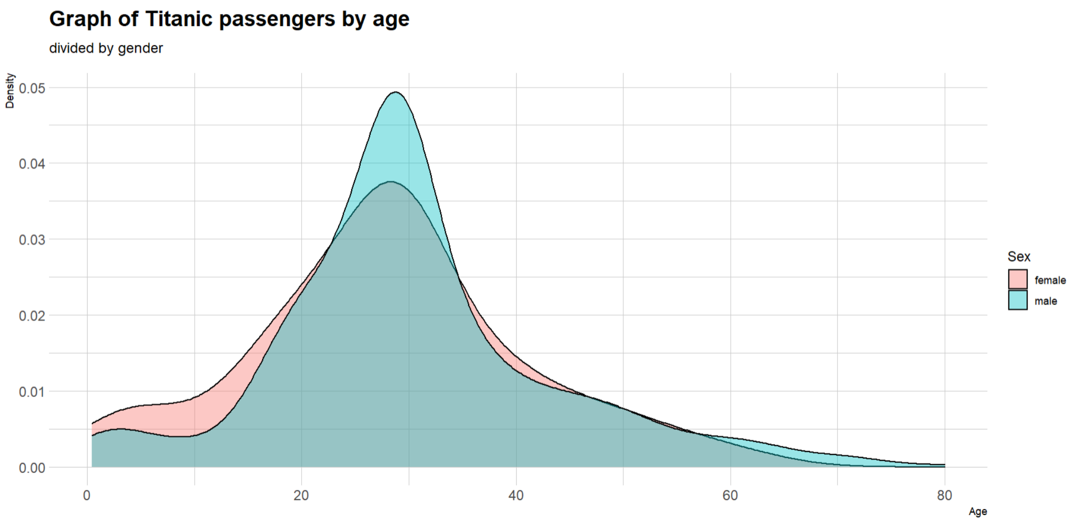


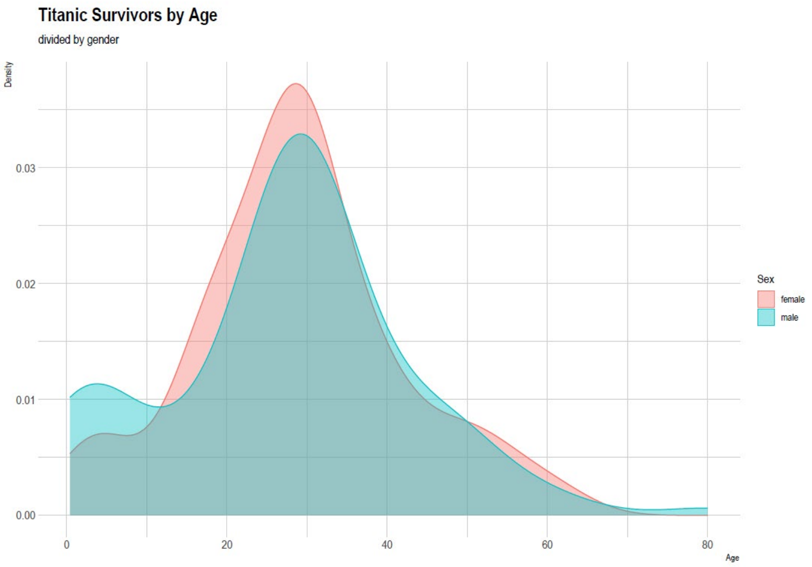
1.3	.9032	.9049	.9066	.9082
1.4	.9192	.9207	.9222	.9236
1.5	.9332	.9345	.9357	.9370

which translates on the z-score table to a depressing **92.2%.** That's the percentage of survivors who would have been younger than me, which is another way of saying "I hadn't a hope in hell of getting off that ship alive". An unsettling prediction.

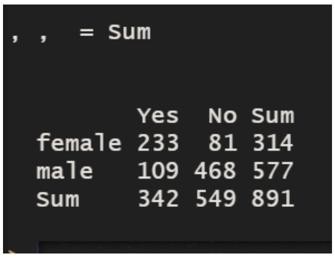
Survivor's Tale

Moving on to matters less egocentric, we find that by superimposing the graph of the spread of ages on the ship between genders, that the disposition of both graphs is similar even though the number of men around the mean is far larger. When this graph is compared to the same graph but for the survivors that central peak has disappeared and the number of women around the mean exceeds that of the men. A startling visual for how many men around the age of 29 lost their lives.





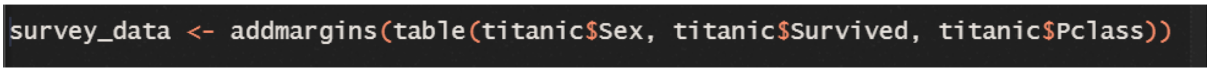
The story gets interesting when social class indicators in the form of ticket prices is introduced, it is startling to see how unequal the death toll was between first and third-class passengers. Here is a concise overview of how many survived and how many died, divided by gender.



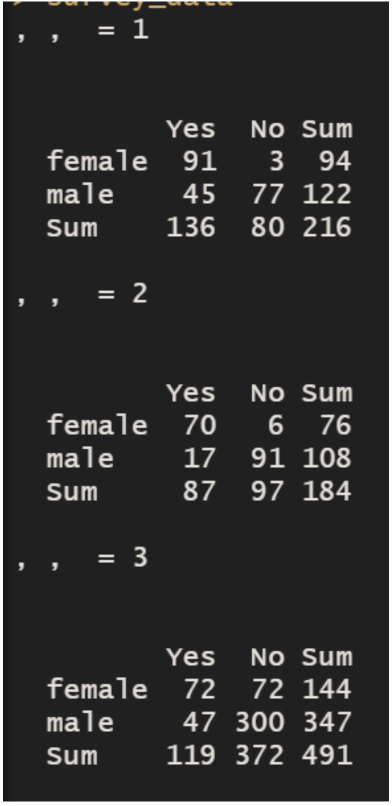
```
, ,  = Sum


        Yes  No Sum
 female 233  81 314
 male   109 468 577
 Sum    342 549 891
```

Of the registered passengers 549 of the 891 who set out drowned in the sinking, that is a percentage of **61.6**, though the ratios change greatly when we consider gender. Of the 577 males who began the journey only 109 - or **18,9%** - survived, and that includes male children. Of the women 233 survived – or **74,3%** - of the 314 total. So clearly 'women and children first' was a true policy of ocean-going vessels of the time and not just a stereotyped assumption.

Let's look at the breakdown of these figures by ticket class, taken from the 'addmargins' and 'table' functions.



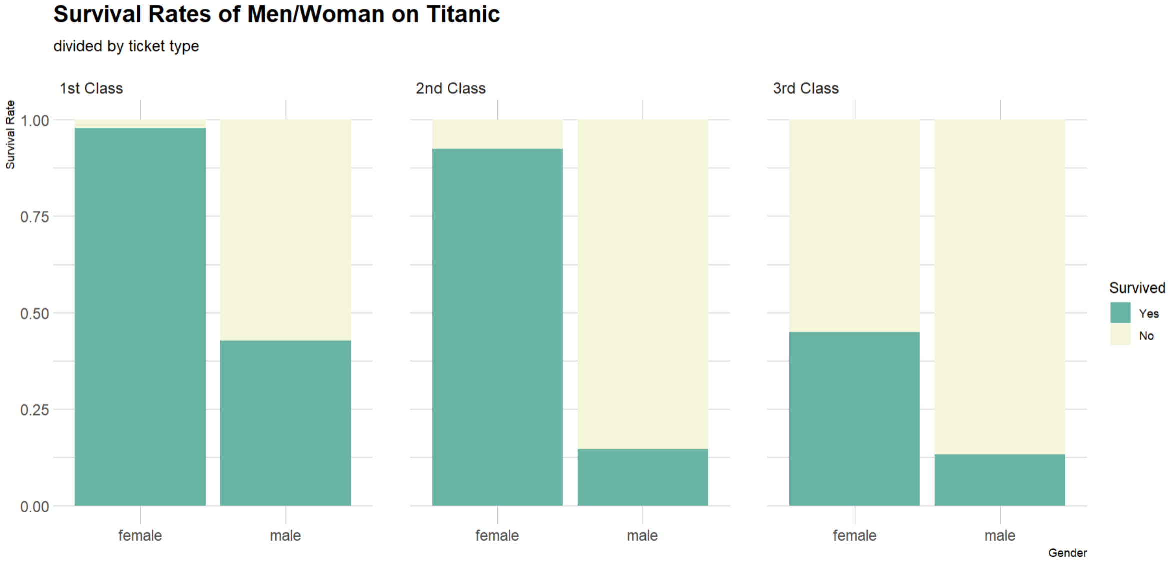
```
survey_data <- addmargins(table(titanic$Sex, titanic$Survived, titanic$Pclass))
```



```
> survey_data
, ,  = 1

        
         Yes  No Sum
  female  91   3  94
  male    45  77 122
  Sum    136  80 216

, ,  = 2

        
         Yes  No Sum
  female  70   6  76
  male    17  91 108
  Sum     87  97 184

, ,  = 3

        
         Yes  No Sum
  female  72  72 144
  male    47 300 347
  Sum    119 372 491
```

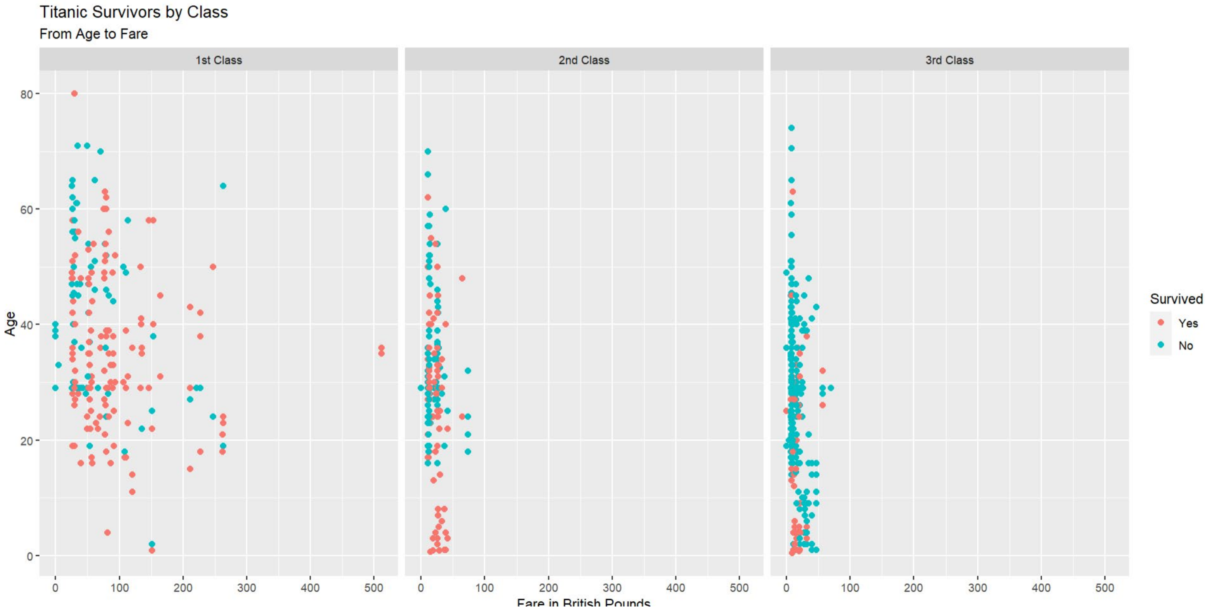
Of the first-class ticket holders only 3 women died, so Rose from the film ‘Titanic’ really had to go out of her way to get herself nearly killed. The percentage of men who survived from this group is much higher than 2nd and 3rd class – at 37%. 2nd class survival rate for men was 15,7% and 3rd class was a meagre 13,5%. The survival rate for 2nd class women was nearly as high as the first – only 6 women died. However, the rate for 3rd class female passengers is strikingly different from the first two, only 50% of women survived. Unfortunately for my relative she was in this category. Both she and her fiancé were both lost that day.

These statistics can be illustrated strikingly with the help of a bar chart. It represents plainly and starkly how much social class and financial means mattered when it came to surviving the disaster and is one of the main reasons why the story has remained relevant ever since.



Linear Regression

As interesting as this dataset was to me personally, it does not lend itself to the productive plotting of variables. Far too many of the columns are ordinal or qualitative in nature. The best I could do was to compare the ages of the passengers and the fare they paid for their tickets. As you will see there is little in the way of correlation between these two values. There is always something to learn from any comparison, even if it is only that the variables have little in common.



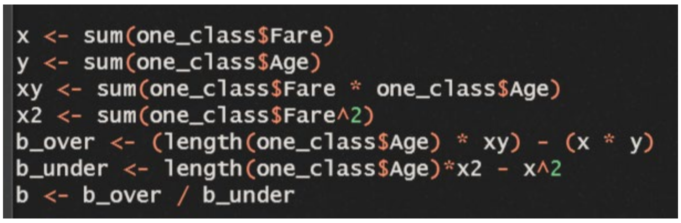
An initial glance reveals that there is little variation in either the second or third classes. First class fares are quite varied and show how one size does not fit all, and what was included in the ticket depended on how much passengers were willing to spend. With more focus on 1st class tickets, we'll see if any conclusions can be made (other than how unsurprising it is that the outliers on the far right, who paid the most for tickets, survived in the end).

R can easily and simply compute the least-squares equation, but the challenge is to see how easily it can be done without it.

$$b = \frac{n\sum xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2}$$

I broke the equation down and wrote code to calculate each part:

The x axis is 'Fare' and the y 'Age'. I created vectors for each set of numbers, including the sum of two vectors and the square of one. I created two further variables, one to represent what was above the division line and one for under. Then I calculated for 'b' and, shock of all shocks, the value was the same as the R calculated version.

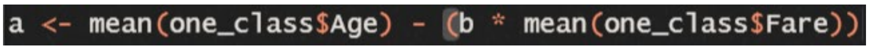


```
x <- sum(one_class$Fare)
y <- sum(one_class$Age)
xy <- sum(one_class$Fare * one_class$Age)
x2 <- sum(one_class$Fare^2)
b_over <- (length(one_class$Age) * xy) - (x * y)
b_under <- length(one_class$Age)*x2 - x^2
b <- b_over / b_under
```

b = -0.03169847

I then plugged this number and the means of the two columns from the dataset into the line equation:

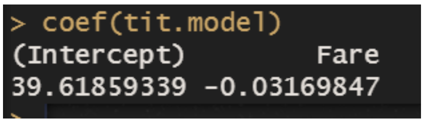
$$a = \bar{y} - b\bar{x}$$



```
a <- mean(one_class$Age) - (b * mean(one_class$Fare))
```

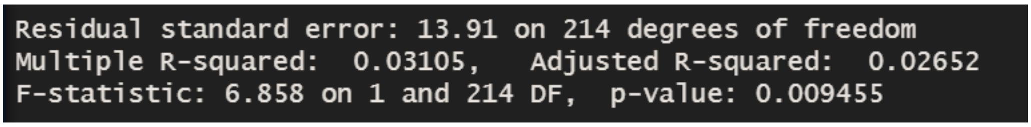
a = 39.61859339.

The **same answer** as R gave me.



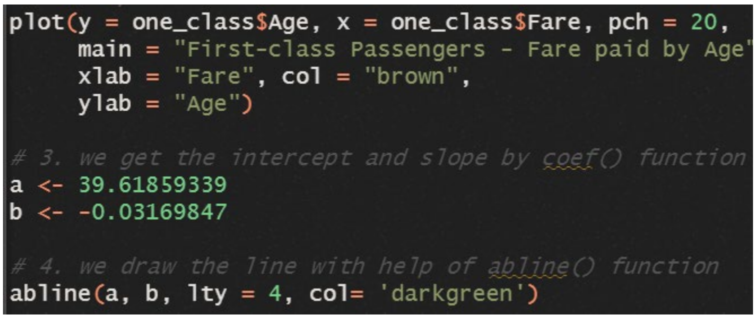
```
> coef(tit.model)
(Intercept)        Fare
39.61859339 -0.03169847
```

I was delighted to be able to work out the equation on my own and was not at all surprised to see that the coefficient of determination was as low as it was.



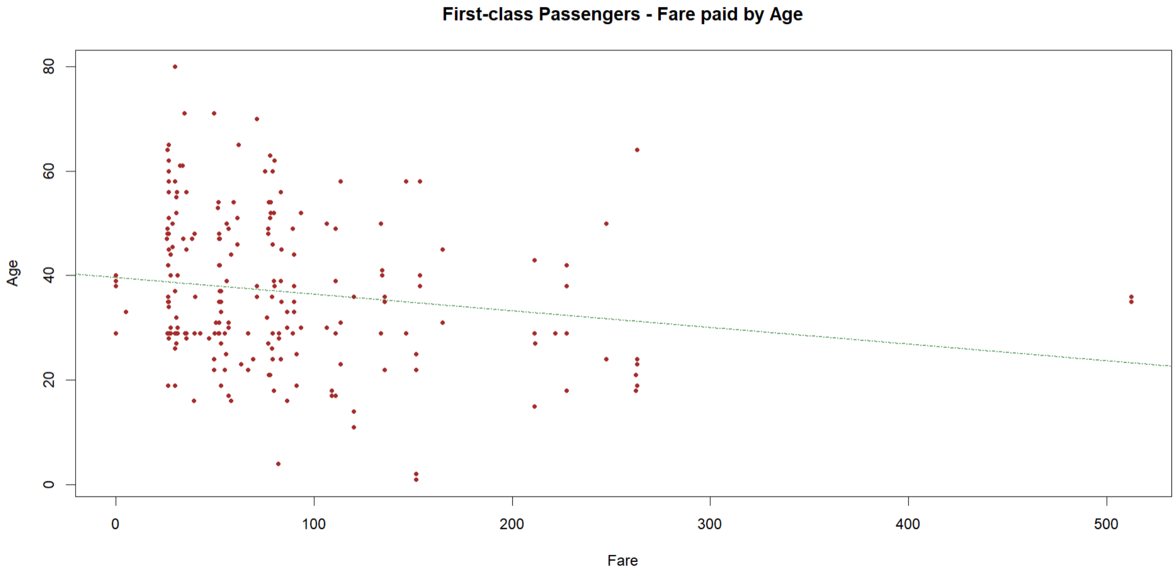
```
Residual standard error: 13.91 on 214 degrees of freedom
Multiple R-squared:  0.03105,	Adjusted R-squared:  0.02652
F-statistic: 6.858 on 1 and 214 DF,  p-value: 0.009455
```

As you can see, r^2 or ‘the coefficient of determination, is a meagre **3.1%.** As I predicted, it was a reach to hope that any kind of significant correlation could be found between these two sets.

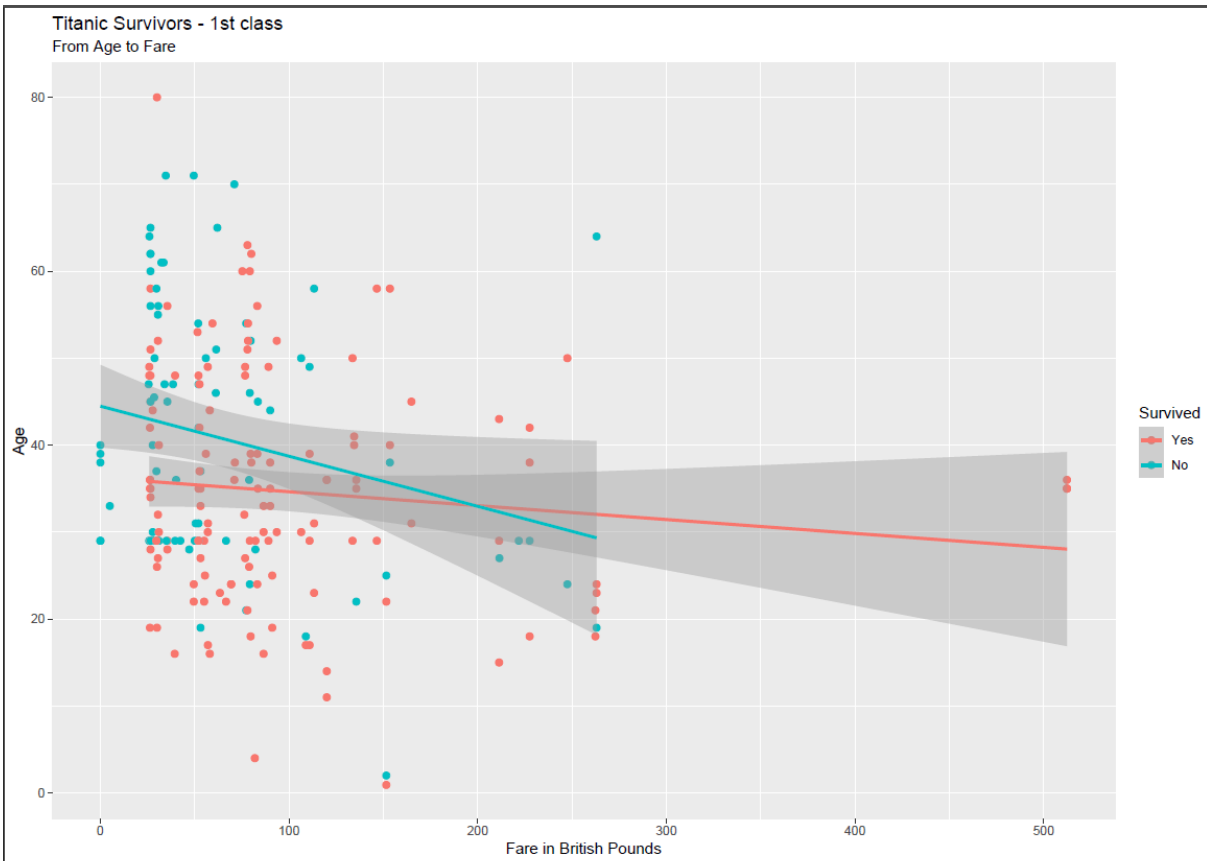


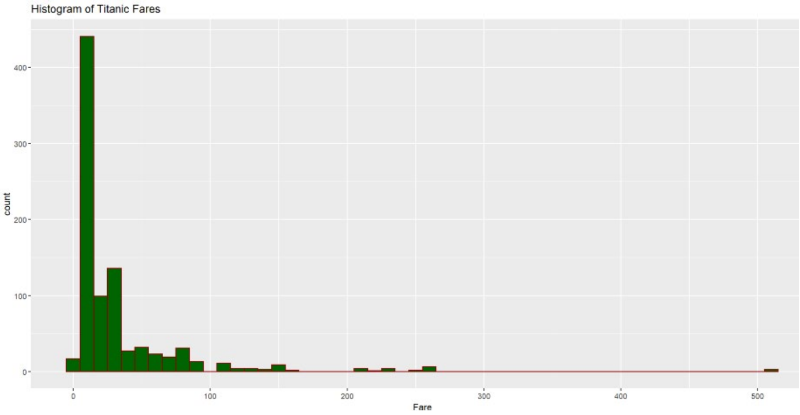
```
plot(y = one_class$Age, x = one_class$Fare, pch = 20,
     main = "First-class Passengers - Fare paid by Age"
     xlab = "Fare", col = "brown",
     ylab = "Age")

# 3. we get the intercept and slope by coef() function
a <- 39.61859339
b <- -0.03169847

# 4. we draw the line with help of abline() function
abline(a, b, lty = 4, col= 'darkgreen')
```

I plotted the graph using the model you gave us, so I could enter my own numbers and create **my own** (notice the pride) least squares line. I have printed the graph below.

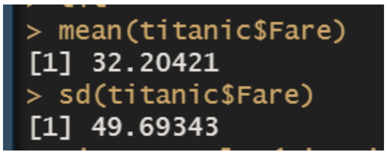


Using geom_smooth in ggplot, I created a similar graph though with the added benefit of being able to see a least squares line for the survivors and for the drowned.





The range of values for Fares is quite spread out and not normally distributed, though according to **Chebyshev's theorem** we can say with confidence that, despite the range, 75% of all values will fall within 2 standard deviations of the mean.



```
> mean(titanic$Fare)
[1] 32.20421
> sd(titanic$Fare)
[1] 49.69343
```

This means that at least **75%** of all fares lie between **0 and 82.5** pounds, which tallies with what we can see in the graph above.

Conclusion

My conclusion from this investigation with graphs and statistics is that the evidence clearly shows how important class was to surviving the sinking. This was primarily because third class ticket holders were kept below deck (which the film shows to devastating effect), and that there was little interest in releasing them onto an already chaotic top deck.

It can also clearly be seen that ‘women and children first’ was a genuine policy so long as you were in first or second class. As I’ve noted, my relative was in third and both she, her female friend, and her fiancé lost their lives. Such disasters become mythology through storytelling, but the clear resonances with the disparity of freedoms, rights and advantages between different social groups in modern times means that this story continues to resonate today, around the world and in many different cultures.