# Titanic Coursework

By Daniel McHugh

Contents

[Titanic Coursework 1](#_Toc535418152)

[Set up and Data Handling 3](#_Toc535418153)

[Data analyst 5](#_Toc535418154)

[**Histograms and QQplots** 6](#_Toc535418155)

[**Outliers and Boxplots** 8](#_Toc535418156)

[**Normality and Variance** 9](#_Toc535418157)

[**Liner Regression** 11](#_Toc535418158)

[**Correlation** 13](#_Toc535418159)

[Hypothesis One 14](#_Toc535418160)

[Hypothesis Two 15](#_Toc535418161)

[Hypothesis Three 16](#_Toc535418162)

[Hypothesis Four 17](#_Toc535418163)

[Hypothesis Five 18](#_Toc535418164)

[Hypothesis Six 19](#_Toc535418165)

[Hypothesis Seven 20](#_Toc535418166)

[Hypothesis Eight 21](#_Toc535418167)

[Hypothesis Nine 22](#_Toc535418168)

[Hypothesis Ten 23](#_Toc535418169)

[Figure 1(Shows the Distribution of the data Age, Fare and SibSp with a QQnorm plot) 6](#_Toc535418170)

[Figure 2(Shows the outliers in the data using Boxplots). 8](#_Toc535418171)

[Figure 3(Shapiro-Wilk test on dataset) 11](#_Toc535418172)

[Figure 4(Levene Test on different values in the dataset Age, Fare, Sex, SibSp) 11](#_Toc535418173)

[Figure 5(Liner Regression on some data in our dataset) 12](#_Toc535418174)

[Figure 6(Age vs. Fare on a scatter plot with Regression line and smoothing applied) 13](#_Toc535418175)

[Figure 7(Pearson Correlation on our data) 14](#_Toc535418176)

[Figure 8(Partial correlation on our data) 14](#_Toc535418177)

[Figure 9(Total amount of people that survived) 15](#_Toc535418178)

[Figure 10(Sex vs Survival) 16](#_Toc535418179)

[Figure 11(Embarked vs. Survived) 17](#_Toc535418180)

[Figure 12(Survived vs. Class) 18](#_Toc535418181)

[Figure 13 (Survive vs. SipSp) 19](#_Toc535418182)

[Figure 14(Survive vs. Age) 20](#_Toc535418183)

[Figure 15(Survived vs. Class with Sex) 21](#_Toc535418184)

[Figure 16(Fare and Class vs. Survived 22](#_Toc535418185)

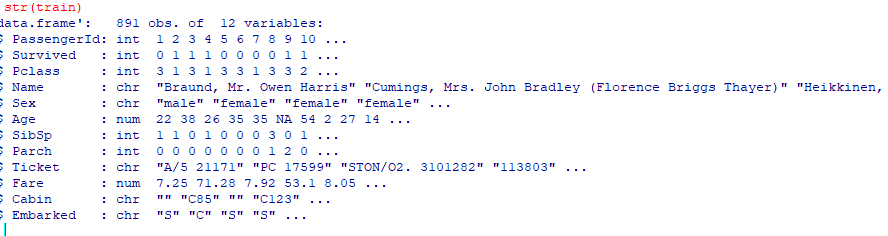
[Figure 17(Average fares on each port with class) 23](#_Toc535418186)

[Figure 18(Class and Sex with Embark) 24](#_Toc535418187)

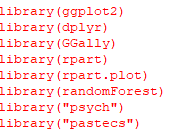
# Set up and Data Handling

Before the data can be analysed it needs to be put into R to do this we run this simple code which tells R where the file is located and where all that information should be placed. As you can see we are reading in test.csv and train from the desktop and placing all the data into test and train variables also below that you can see when you look into the variable train we get the data back.

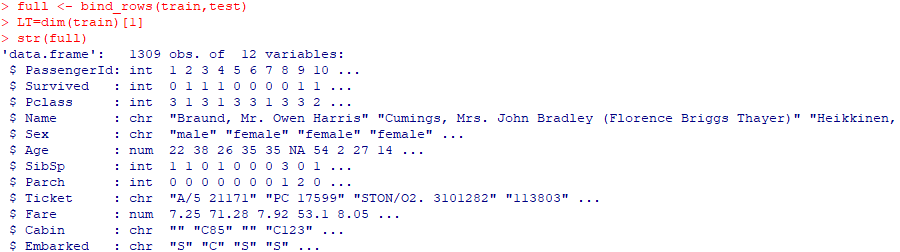




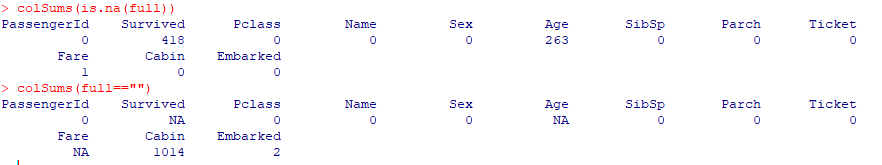
These are packages that will be used in order to analysis our data.



As we have two dataset these need to be but together to do this the function bind\_rows is called in this will bind the datasets and put them in a new variable called “full”.



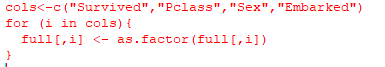
At this stage the data has to be normalized this means returning our data to a normal state so when the data is analysed we have fewer outliners. Below is the code to see what data is equal to zero if this is the case the data needs to be changed.

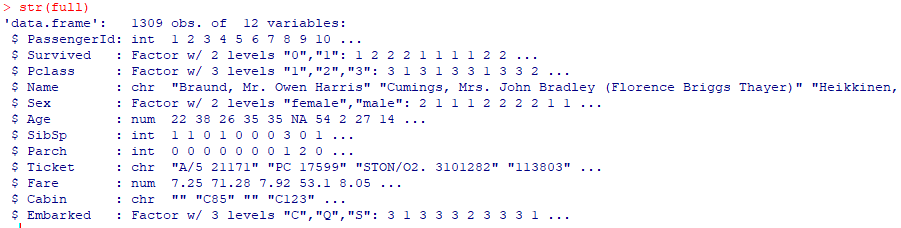


Below is the code that will normalized the data and show us how much data is in each field.



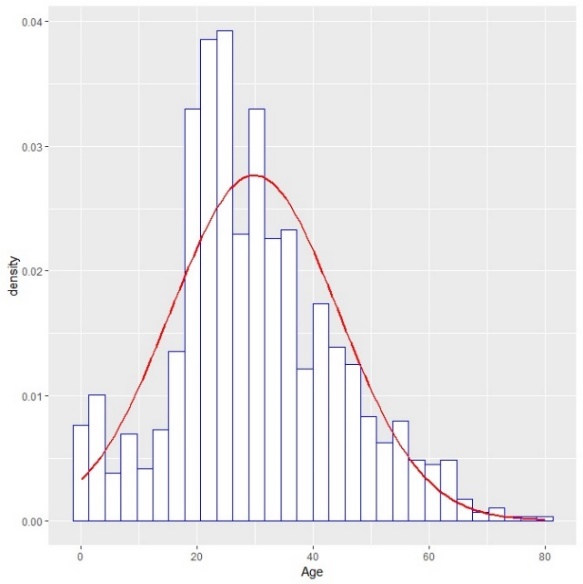
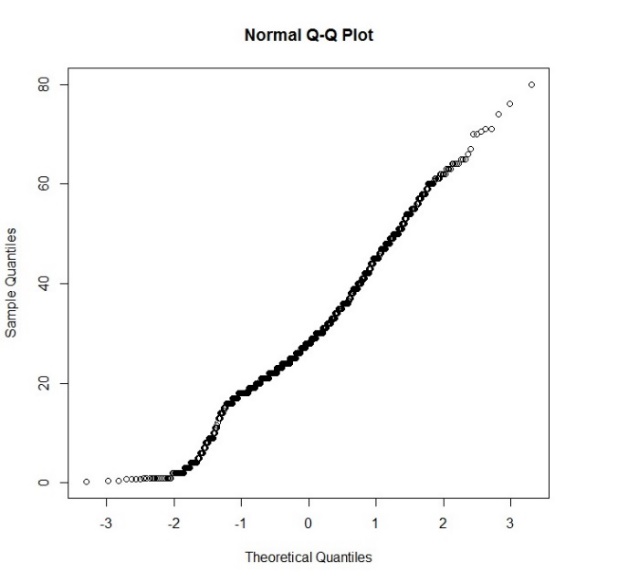
Below is the method to change some of our titles data type to factor and further below is the result of our data handing in our main data frame.

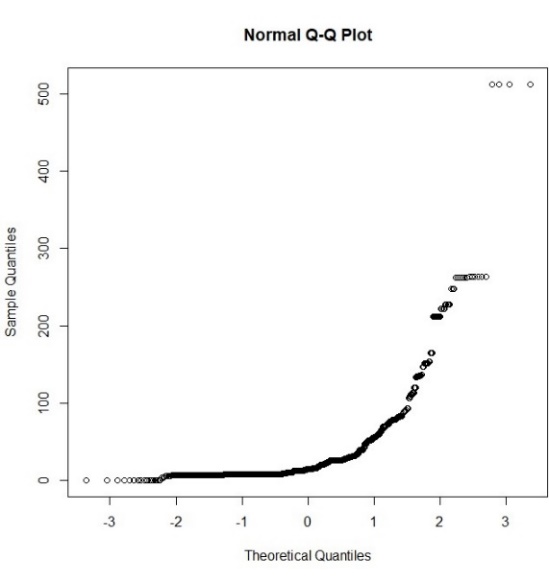
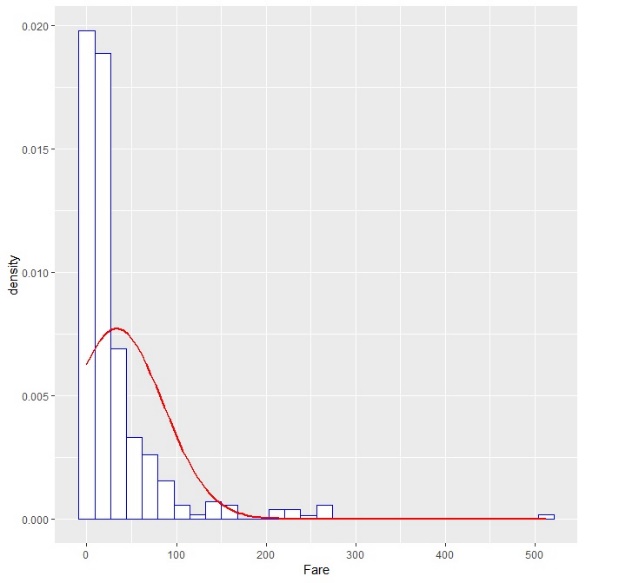




# Data analyst

In this section the data that was collected will be analysed this will allow us to have a closer look at our data hopefully this will tell us how the data is acting.



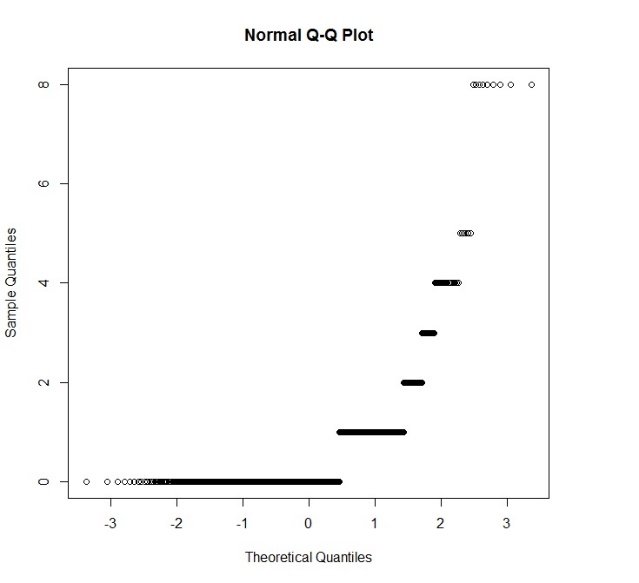
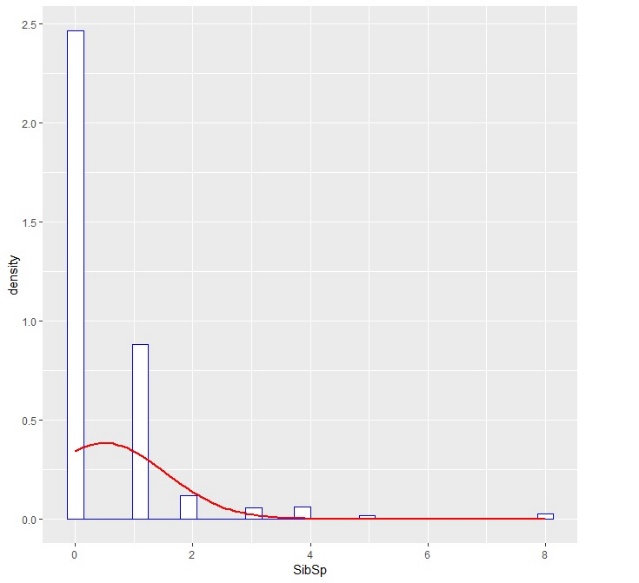


Figure 1(Shows the Distribution of the data Age, Fare and SibSp with a QQnorm plot)

## **Histograms and QQplots**

As you can see in figure 1 this is the data that has been collected you can see that three data values have been plotted these are Age, Fare and SibSp to plot these values two types of graphs have been used the first type is a Histogram (Left side) this allows us to better display the distribution of our data and the second one is a type of QQplot this one is called a QQnorm plot (Right side) which gives us a probability plot by comparing their quantiles against each other, using both of these graphs it displays if our data has a normal distributed or not.

Looking at the histogram in figure 1 first which tell us if our data is distributed normally by the skew (Red Line) the value Age has a normal distribution this is shown by the way the skew has been drawn this is known as a bell curve which tells us the data inside age has less of a chance to produce unusually values these are known as outliers which will be talked about after this part. Next is Fare and the SibSp histogram as these two have the same distribution it makes sense to talk about them together as the graph shows both of these have a positive skew this means that the right tail is longer and the distribution is more concentrated on the left side this distribution tells us that both Fare and SibSp data is right-skewed this is because of the tail being drawn out to the right also this tells us that the mean of these data is more to the right which gives it, it’s skew.

Looking at the QQplots in figure 1 this tells us about the quantiles in our data what this means is the fraction or percent of point below the given value these values are 25% for the lower quartile, 50% for the median and 75% for the upper quartile using these values lets us find any outliers that would give us unusually results this can also be used to see how data is being distributed but it is less clear than the histogram that is used.

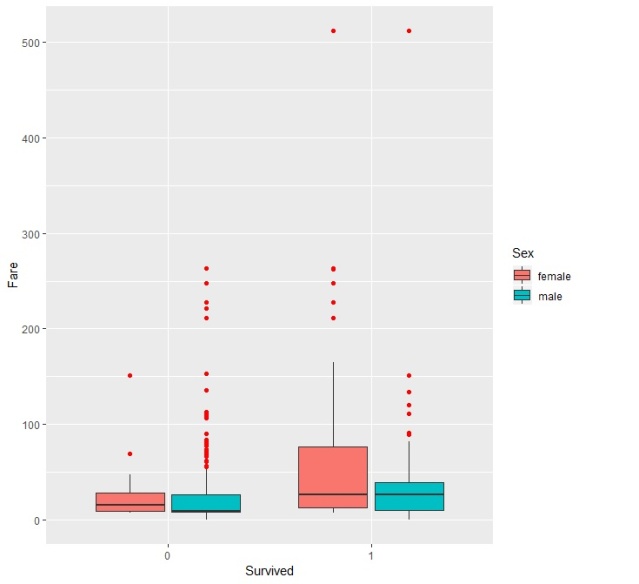
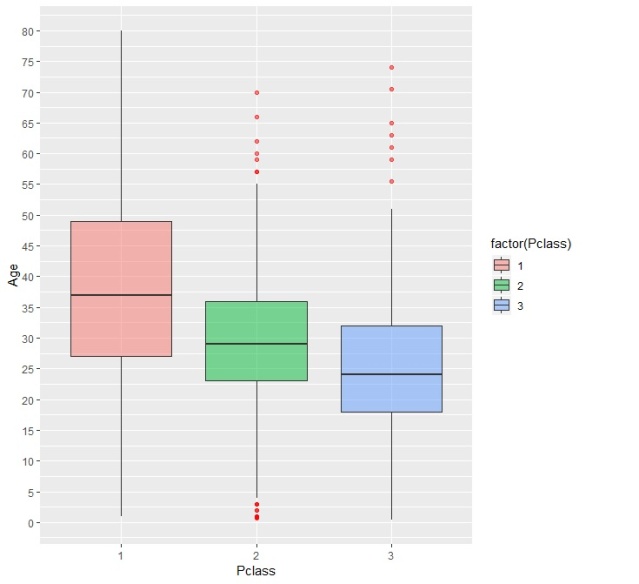
Looking at Age's QQplot as example of a good QQplot as shown already in the histogram the data inside Age is normal that being said just because the distribution of the data is normal doesn't mean that all the data is normal it just means that a higher percent of the whole data is normal so one way to see this kind of data is by using a QQplot. In Age you can see that some data is unusually as you can see the data forms a line any data that doesn't follow the line is seen as odd for example at the top you can that some of the data has broken off the line and formed a group this data could be the result of outliers in our dataset.

Now looking at the other two you can see that they are very different compared to age this is because of the amount of data for these two values either their wasn't enough to form a solid line or our data is not distributed normally, looking at Fare first some kind of line has formed meaning that the data has some kind of distribution also this has told us that some data in fare is odd this is shown by the group of data way above the other data points in fare and finally looking at SipSp this data is very strange as no line as formed but groups of data have meaning that either the data inside SipSp is very few or a lot of the data has the same value, not only that but clear odd data can be seen at the very top of the grid this could be outlier data or just odd data.

Below is the code that was used to create these plots.







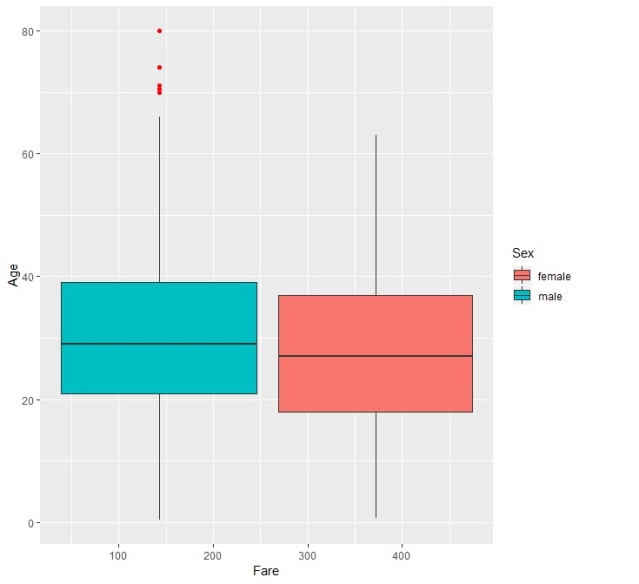
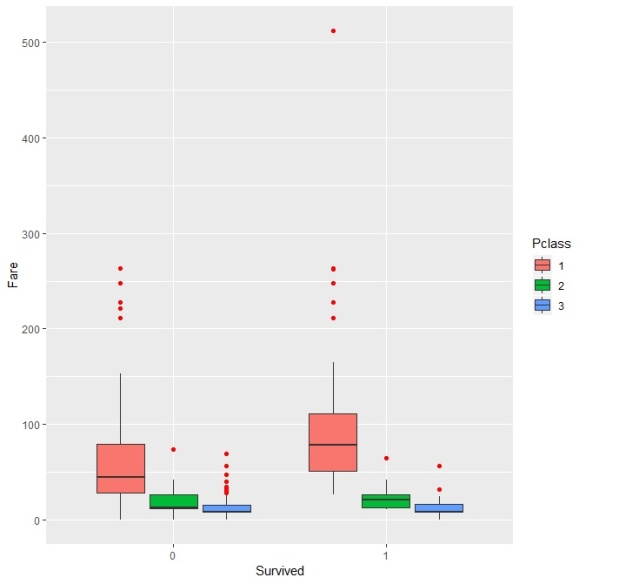
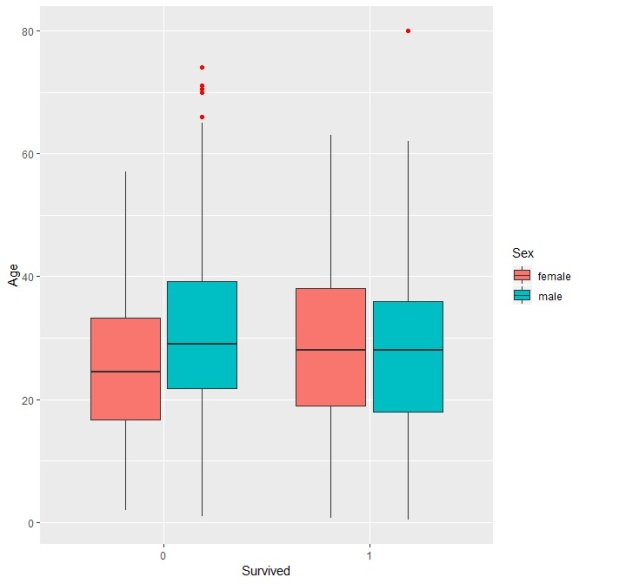


Figure 2(Shows the outliers in the data using Boxplots the values used to create these plots are Age vs. Class (Top left), Fare vs. Survived with a fill of Sex (Top right), Age vs. Survived with a fill of Sex (Middle left), Fare vs. Survived with a fill of Class (Middle right), Age vs. Fare with a fill of Sex(Bottom left).

## **Outliers and Boxplots**

Figure 2 shows us the data in a Boxplot format with small red dots informing us about the outliers in our data, the reason why boxplots have been used is because they display the data really nicely telling us a lot about our data from how the boxes have been form this will be explained later as it is important to reading data in this format, another reason why boxplots have been used is because they clearly show outlier values in our data in a format that anyone can understand.

Boxplots are very useful tools that not only show outliers they also tell us the minimum and maximum value in our data but more importantly they tell us about the quantiles of our data this is show by the box itself as the bottom line tells us the minimum value in our data to find this the box plot does this calculation "Q1 - 1.5\*IQR" next is the interquartile range (IQR) this is made up of three values Q1, median/Q2 and Q3 each will be talked about to get a better understanding of boxplots. Lowest line that form the box is called "Q1" or first quartile this number is created from the middle number between the smallest number and the median of the dataset the percent values do not change so this would be 25% as already detailed above, next is the centre line this is called the "median" or "Q2" this is the middle value of the data for example in the first boxplot Age vs. Class in the red box the median would be 37 the percent value is 50% and the last value is "Q3" or third quartile this is the middle value between the median and the highest value of the data after that comes the maximum this is the highest value in the data the calculation to find the max is "Q3 + 1.5\*IQR" also the lines that come out of the boxes are called whiskers.

Each of these boxplots represents the outliers that have been found in the dataset as before with the histograms and the QQplots the graphs will be explained to better understand what they represent. In figure 2 the first plot is Age vs. Class as you can see three boxes have formed these are 1st class (Red), 2nd class (Green) and 3rd class (Blue) with Age so each box is a representation of the different Ages within the three different class meaning that group every person that is in first class in one group and display the range of ages the reason to do this is it allows us to see the median age of the class which might be important later on, As you can see there are some outliers in this boxplot these are shown as red circles this tells us that some ages in classes two and three are unusually these could be errors within our dataset.

The next boxplots that will be looked at will be Fare vs. Survived with a fill of Class and Fare vs. Survived with a fill of Sex as these two are closely related they will be talked about together the first one that will be talked about is Fare vs. Survived with a fill of Class, in this boxplot there seems to be a lot more outliers compared to any other graph by taking a closer look most of the outliers seem to be on one point which is males that didn't survive and the amount they paid this tells us that the data for males that didn't survive and paid a higher amount is unusually which seem odd as surly the more you paid the higher up on the boat you would be, the other outliers which is uncommon are the ones at the very top of the graph but as there is only two point it's not too concerning as it is normal to get two or three points of data that is far from the group.

Fare vs. Survived with a fill of Sex in this boxplot there seem to be a concentration of outliers in 3rd class people that didn't survive this tells us that the data for people in third class that didn't survive and paid a higher amount is unusually which is different from the like example as people in third class had an overall lower chance compared to everyone else.

Below is the code that was used to create these plots.



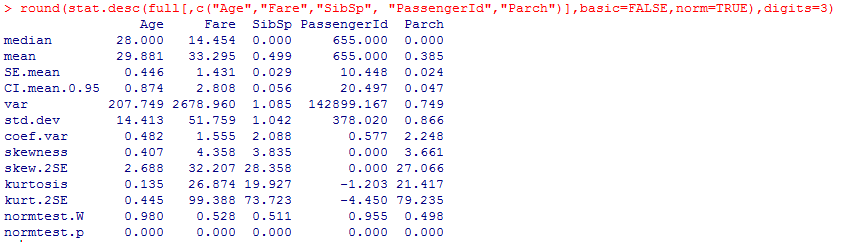
## **Normality and Variance**

In this section two main test's will be performed these the Shapiro and Levene test which test normality and variance both of these are important as they tell us important facts about our data, both of these will be explained as they are shown. But before the main test some basic statistics summary will be done the first one is using a package called psych and the second one is using a package called pastecs, some of the important data that can be taken is skew, kurtosis and mean these values tell us what our distribution is and what shape our data will take.

**psych**

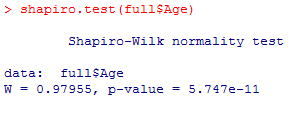
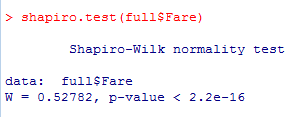
## describe(psych).JPG

**pastecs**



**Shapiro Test**This test allows us to see if our data has a normal distribution or not this test has been used for years now and is one of the best methods to check if data has normality, the way Shapiro test is done is by testing the null hypothesis so if the p-value that is created is greater than the alpha value which is normally 0.05 then the null hypothesis is not rejected. Figure 3 shows what the distribution is like on our data as you can see the first test is Age the p-value is 5.74 which is good it means that the distribution is normal which is correct as talked about before, but the other three test have the same p value this is because the p value is much smaller than 0.05 this tells us that the data is not normally distributed.

**Levene Test**Levene’s test, tests the null hypothesis that the variances in different groups are equal (i.e., the difference between the variances is zero). If Levene’s test is significant at p-value is less than 0.05 then we can conclude that the null hypothesis is incorrect and that the variances are significantly different however if Levene's test is non-significant meaning that p-values is greater than 0.05 then the variances are roughly equal and the assumption is correct. Figure 4 shows the variances in our dataset in this test both the mean and the median are tested, an example of variances is shown in the first test here Age is tested against Survived and Sex each time the p-value is above 0.05 this tells us that these values variances is equal and the null hypothesis is correct, other values like Parch and Survived have a p-value less than 0.05 telling us that these values have no variance meaning that our null hypothesis is incorrect.

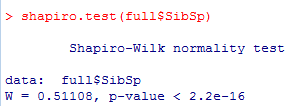
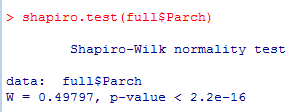
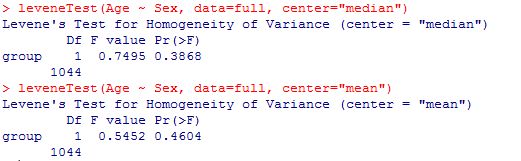
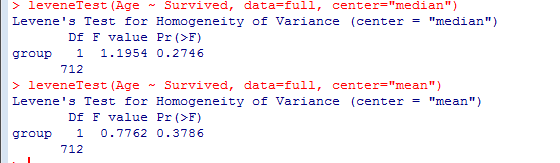
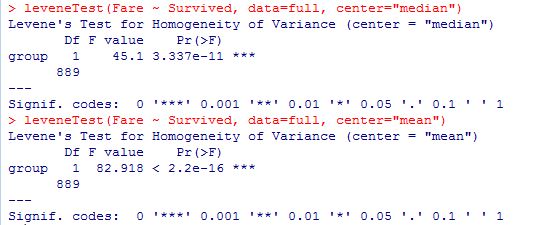
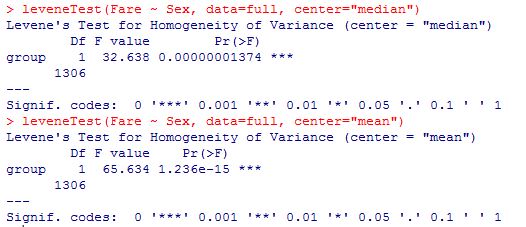
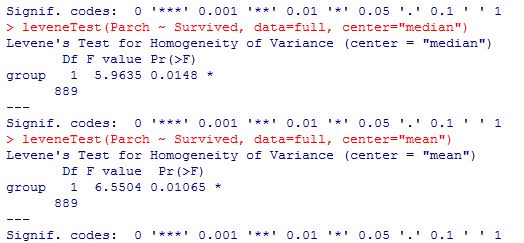
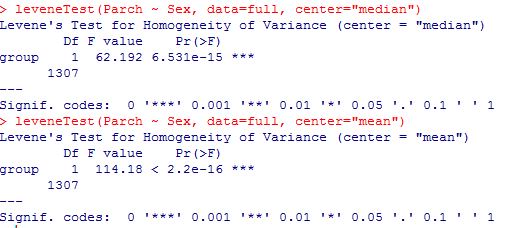
 

Figure 3(Shapiro-Wilk test on dataset)







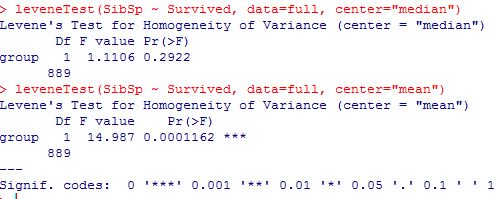
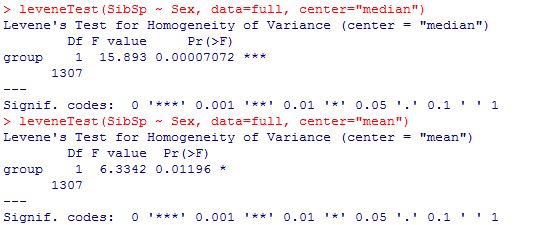


Figure 4(Levene Test on different values in the dataset Age, Fare, Sex, SibSp)

## **Liner Regression**

Linear regression is a commonly used type of predictive analysis the idea of regression is to examine if a set of predictor variables do a good job in predicting an outcome variable also which variables in particular are significant predictors of the outcome variable. To better understand regression certain terminology is used for example dependent and independent are used a lot to talk about variables in regression a dependent variable is a type of variable that measures the affect of the independent variable these are also known as Predicted variables, independent variables are manipulated or will change depending on what other variable is used with it.

Below are two different ways to represent regression using our dataset as you can see both of these have the same data values but present different information, figure 5 is a Liner regression using Rcmdr package in this figure not only can you see details about the regression but it also gives us the Residuals and the coefficients values, residuals have already been talked about above but coefficient is a new term that has not been talked about yet what this means is how the two variables relate to each other.

The summary statistics in figure 5 tells us a lot about the relation between the two variables to see if these two are statistically significant the first value that must be looked at is both p-values in the linear model in some models the p-value is written as "Pr(>|t|)" it means the same , as explained in other sections the p-value is very important in seeing how significant these variables are to each other an important factor to remember when dealing with the p-value is the significance level normally this level is "< 0.05" which has already been detailed in other sections. But that's not the only value that is important the t value is useful as well as a high t value indicates the likely that the coefficient is not equal to zero purely by chance so the higher the t value the better as this tells us that there is some relation between the two variables also this model has a more visually way of showing us how significant the variable is to the other by the amount of stars these are known as significance stars and the more stars the more significant the variable is.

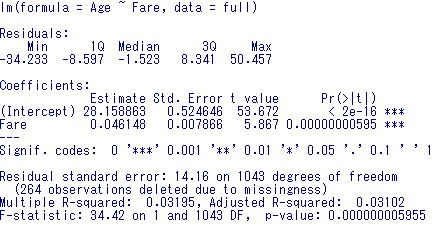
Other important values are the R-squared or Multiple R-squared and the Adjusted R-squared the R-squared value tells us the proportion of variation in the dependent variable and the Adjusted R-squared penalizes total value for the number of terms in the model, also one other value is important that is the F-statistic this values tells us the goodness of fit this means how good is this model on the data that has been given in most cases the higher the fit the better the model.

In figure 6 a scatter plot has been created using Age and Fare also the use of a regression line has been added to allow us to see the relationship of the two variables, not only that but the grey outline is used to display the residuals as well giving us a full display of what our data looks like in regression form.

The code to create the Liner regression model and the scatter plot will be below here:  
Liner Model:  


Scatter plot code:  


Figure (Liner Regression on some data in our dataset)



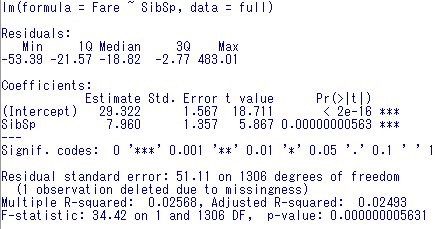
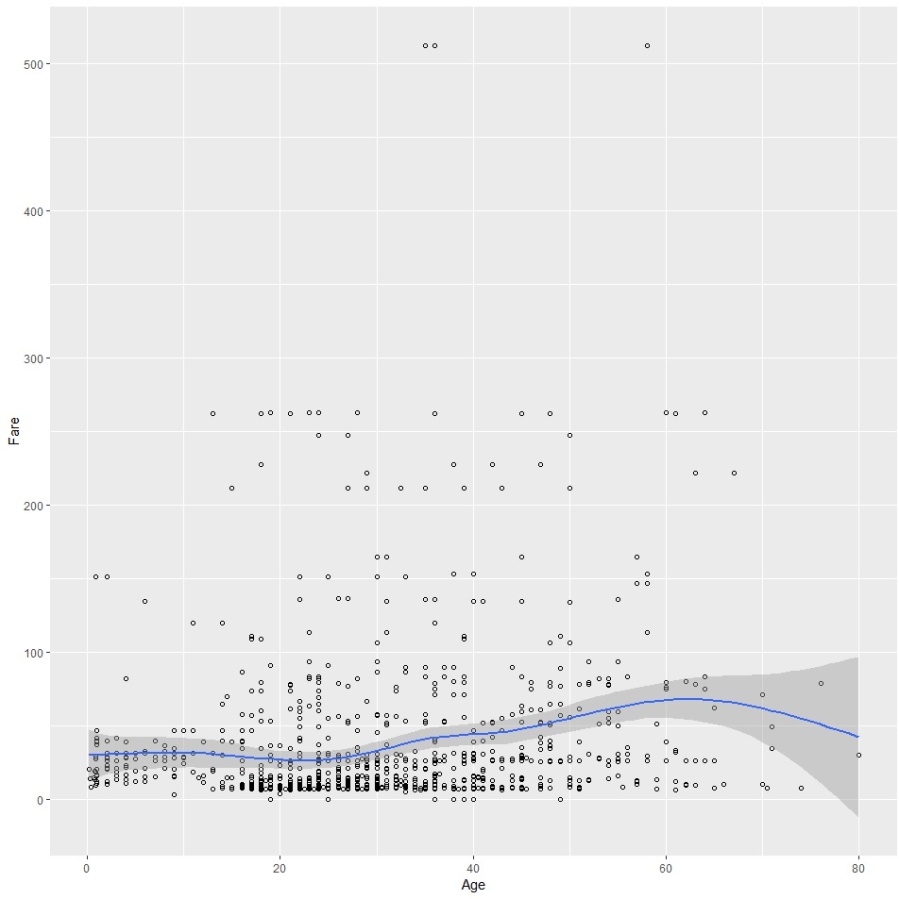


Figure (Age vs. Fare on a scatter plot with Regression line and smoothing applied)



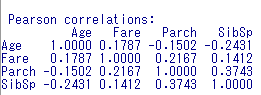
## **Correlation**

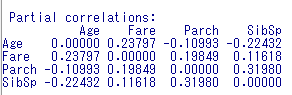
Correlation is a statistical technique that can show whether two variables relate and how strongly these pairs are related an example could be height and weight if these two variables do relate how strong is their relation to each other and does this relation cause one variable to depend on another, Correlation can tell you just how much of the variation in peoples weights is related to their heights. Although this correlation is fairly obvious our data may contain unsuspected correlations that could only be seen using the different methods in correlation, one of the main reason for doing correlation on our data is to get a better understand of how our data is related and to find data that could be related without our knowledge.

In figure 7 and 8 two different correlation methods have been used the first one is called Pearson correlation this method uses numbers between -1 and 1 this number indicates the extent to which two variables are linearly related, meaning that the numbers displayed in figure 7 tells us if there is a correlation between two variables and how strong that correlation is for example Parch and SibSp have a 0.3743 this means that there is a positive relation and the strength would be medium so in this example the data says that while Parch increases SibSp tends to increase as well.

The opposite of this would be SibSp and Age in this relation a negative number is given this tell us that these two have a negative relation with a medium strength so the data is saying that while SibSp increases age doesn't increase. In this method numbers that are positive are called positive relation and numbers that are negative are called negative relations if a number between two variables is zero it means there is no relation.

Figure 8 correlation method is Partial correlation which is a measure of the strength and direction of a linear relationship between two continuous variables whilst controlling for the effect of one or more other continuous variables this other is known as covariates, like with Pearson correlation numbers are given to show the relation between variables -1 is a perfect negative, 0 is no relation and 1 is a perfect positive relation. Using this method multiple variables overlap to try and explain certain patterns in our data from this certain variable can be seen to directly influence other variables meaning that if variable A changes then B will change as well.

Figure (Pearson Correlation on our data)

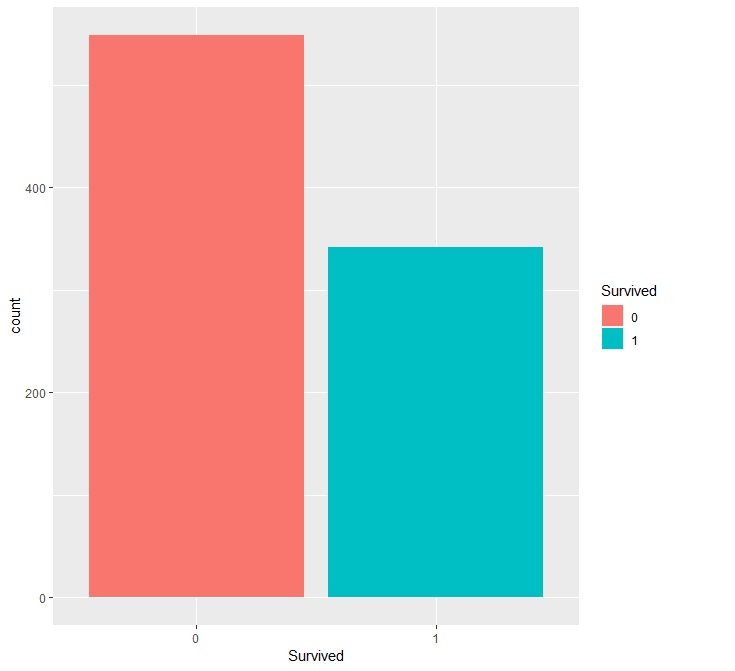
Figure (Partial correlation on our data)

To create these correlations this code was used:  
Pearson correlation:  


Partial correlation:  


Hypothesis One - Out of the total amount of people that bordered titanic more people died than survived.

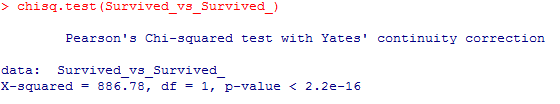
Alternative - Out of the total amount of people that bordered titanic more people survived than died.

Figure (Total amount of people that survived)



As you can see by figure 9 the total amount of people have been split into two different groups zero is if they died and one is if they survived as you can see well over half of the total amount of people did die that bordered the titanic. If we calculated these two groups into a percentage the amount of people that survived would be 38% this is because only 342 people survived while the people that died would be 62% this is because 549 people didn't survive as you can the percentage values of the two are very different telling us more people died than survived.

Below is a Chi square this test's to see if our null hypothesis is true by testing to see if the two variable are independent or not. Since our P-value is less than 0.05 we can say that these two are not independent meaning our null hypothesis is correct.



Hypothesis Two – Females on the ship have a higher chance of surviving over males.  
Alternative - Females on the ship have a lower chance of surviving over males.

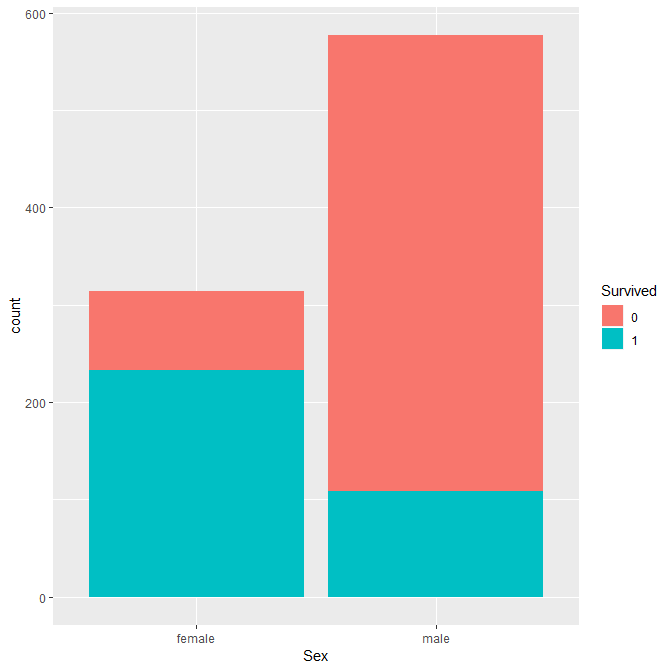


Figure 10(Sex vs. Survival)

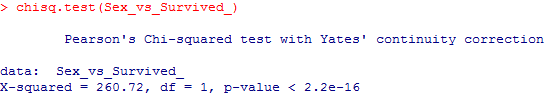


Table 1(Chance to Survive vs. Sex)

As you can see by figure 10 more females survived compared to males this could be because when the ship was sinking women and child were first off the boat meaning more woman had access to lifeboats over males another factor could be that most of the passage were males on the ship so the death ratio for males would be higher.

And to further reinforce our Hypothesis Table 1 shows the overall chances that you would survive based on your sex, as you see 74% of the people that survived were female while only 18% were male also this can be said about the people that died as well 81% were male while only 25% were female.

Below is a Chi square this test's to see if our null hypothesis is true by testing to see if the two variable are independent or not. Since our P-value is less than 0.05 we can say that these two are not independent meaning our null hypothesis is correct.



Hypothesis Three - People that Embarked at Cherbourg where more likely to survive.

Alternative - People that Embarked at Cherbourg where more likely to die.

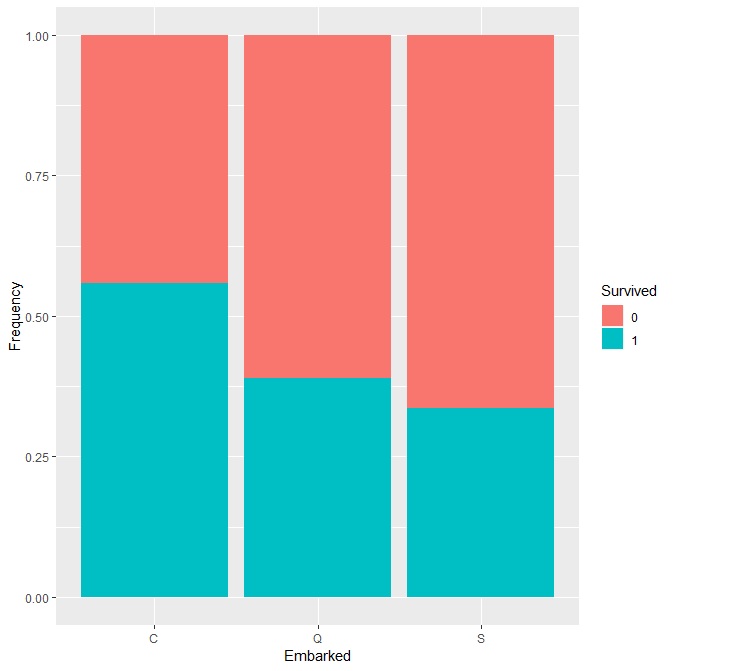


Figure 11(Embarked vs. Survived)  

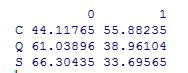
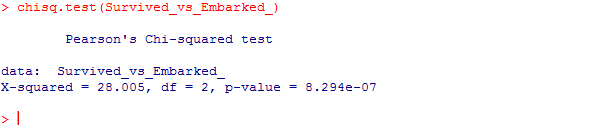



Table 2(Chance of Survive vs. Embarked)

As you can see by figure 11 people that embarked at Cherbourg survived over those that embarked at other ports this could be because the amount of people that boarded at Cherbourg was less than other ports or that most people that survived boarded at Cherbourg.

And to further reinforce our Hypothesis Table 2 shows the overall chances that you would survive based on where you embarked from as you see 55% of the people that survived boarded at "C" which would be Cherbourg, also this tells us the chances that you wouldn't survive based on where you embarked from this would be "S" which stands for Southampton.

Below is a Chi square this test's to see if our null hypothesis is true by testing to see if the two variable are independent or not. In this case our P value is above 0.05 meaning these two are independent so we choose our alternative.



Hypothesis Four – People in higher classes had a better chance to survive.

Alternative - People in higher classes had the same chance to survive as people in the lower classes.

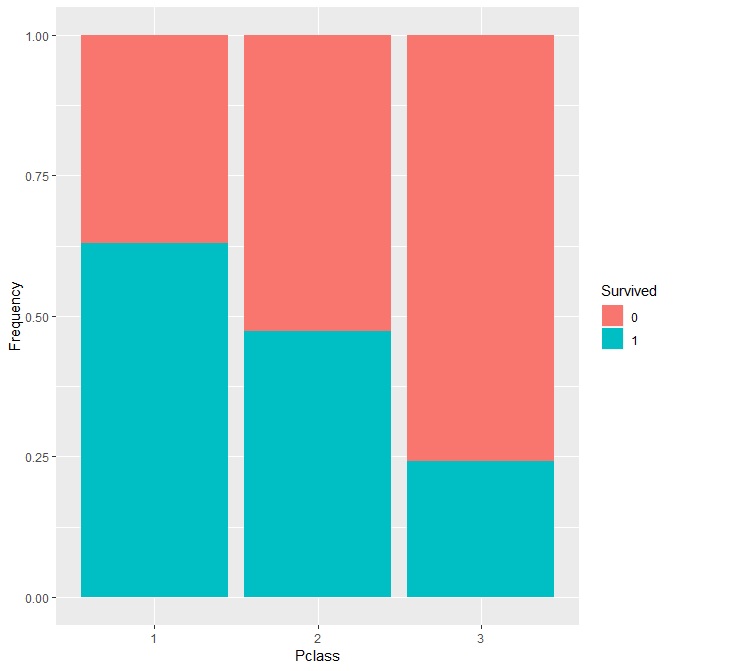


Figure 12(Survived vs. Class)

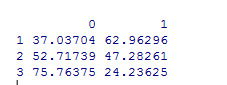
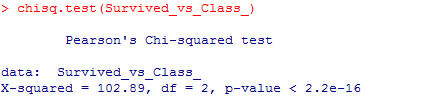


Table 3(Chance of Survive vs. Class)

As you can see by figure 12 people that were in the higher classes survived more than the other classes this could be as they were on the top of the ship they would have access to the lifeboats quicker than other people meaning more first class people got off the ship.

And to further reinforce our Hypothesis Table 3 shows the overall chances that you would survive based on your class as you see if you were in first class you had a 62% chance of surviving compared to people in third class which only had a 24% chance of surviving also from this we can see the chances of you not surviving with third class being at 75% and first class being at 37%.

Below is a Chi square this test's to see if our null hypothesis is true by testing to see if the two variable are independent or not. Since our P-value is less than 0.05 we can say that these two are not independent meaning our null hypothesis is correct.



Hypothesis Five – People that had at least one siblings and Spouse had the same chance of surviving as people that had one sibling or spouse.

Alternative - People that had at least one siblings and Spouse had a greater or less chance of surviving as people that had one sibling or spouse.

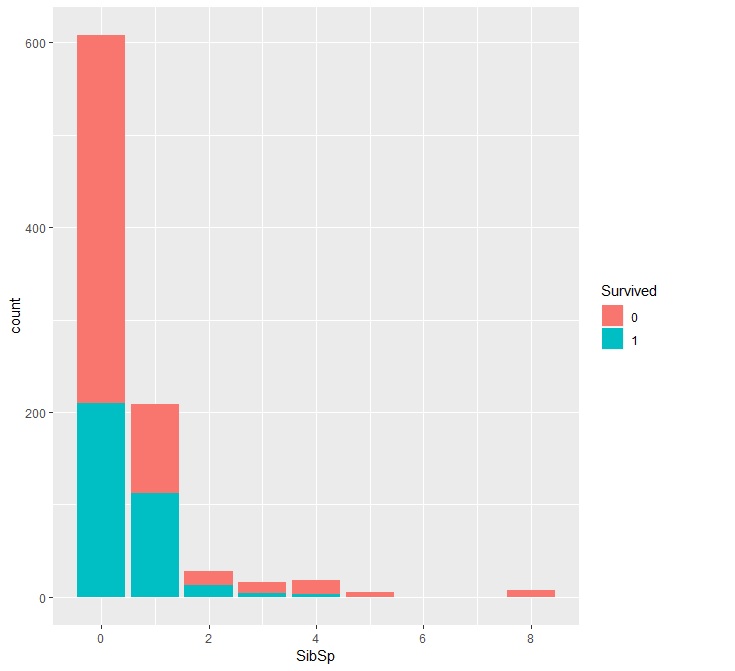


Figure 13 (Survive vs. SipSp)  

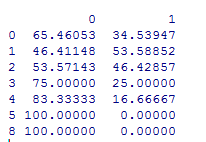



Table 4 (Chance of Survive vs. SipSp)

As you can see by figure 13 people that had fewer or no siblings and Spouses on the ship survived more compared to people that had bigger families this could be because of many reason one might be that someone lost their child and went looking for them another reason could be that someone had a brother/sister on a lower class and they went looking for them.

And to further reinforce our Hypothesis Table 4 shows the overall chances that you would survive based on the amount of siblings and Spouses you had on the ship, 88% is the chance you get if you had either zero siblings and Spouses on the ship or you has at least one sibling or spouse on the ship compare that to the other result which is also 88% this is if you had a siblings and Spouses on the ship and you have more than one sibling and Spouse on the ship.

Hypothesis Six - People in the age range of 20 to 40 died the most.

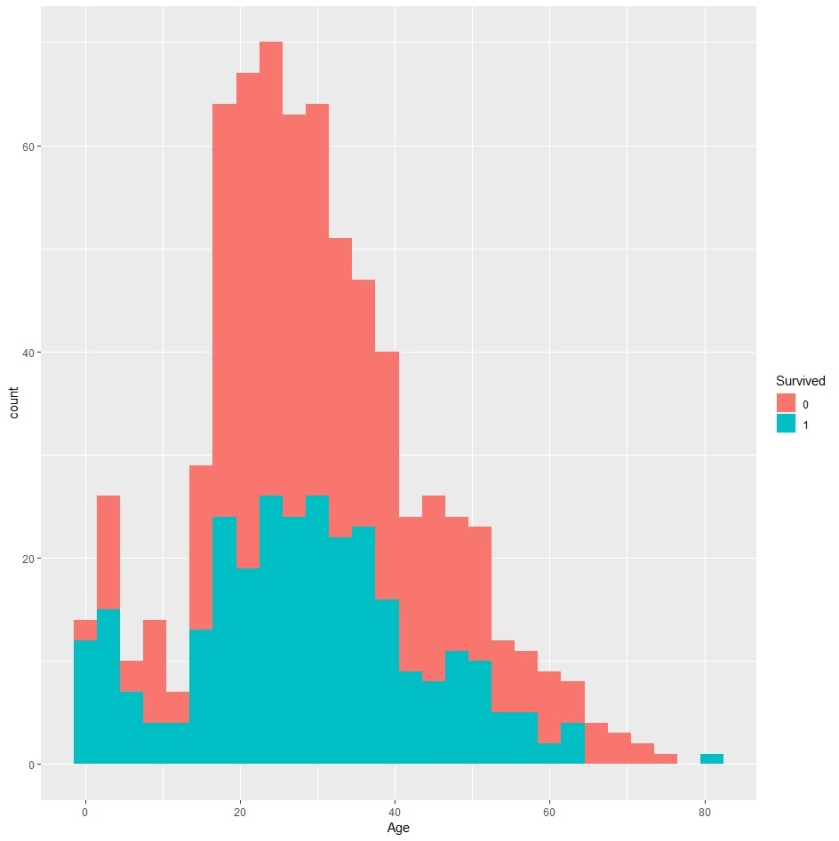
Alternative - People in the age range of 20 to 40 survived the most.

Figure 14(Survive vs. Age)



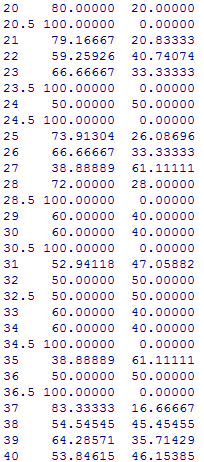


Table 5(Chances of Survive vs. Age)

As you can see by figure 14 people that were in the age range of 20 and 40 died the most this could be because most of the younger passengers were in third class which has proven to have the lowest chance of surviving out of the other classes.

And to further reinforce our Hypothesis Table 5 shows the overall chances that you would survive based on your age.

Hypothesis Seven – Females in every class survived more compared to males in every class.

Alternative - Females in every class had the same chance of surviving compared to males in every class.

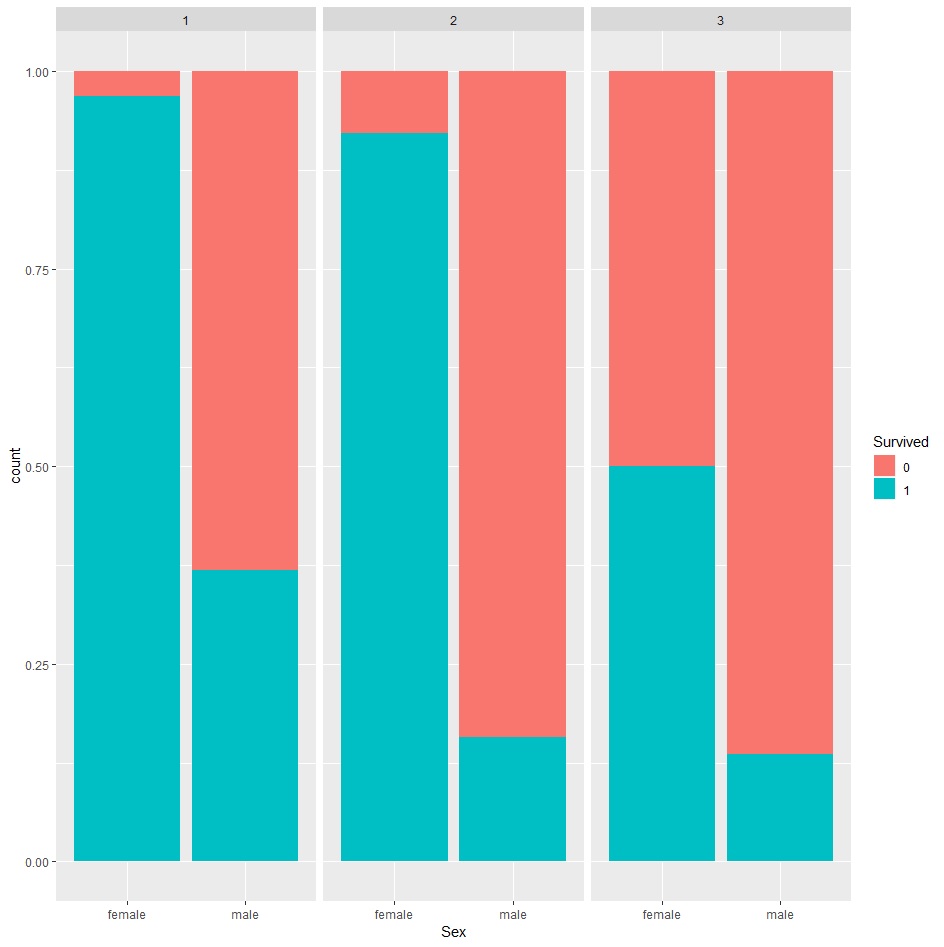
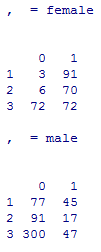


Figure 15(Survived vs. Class with Sex)



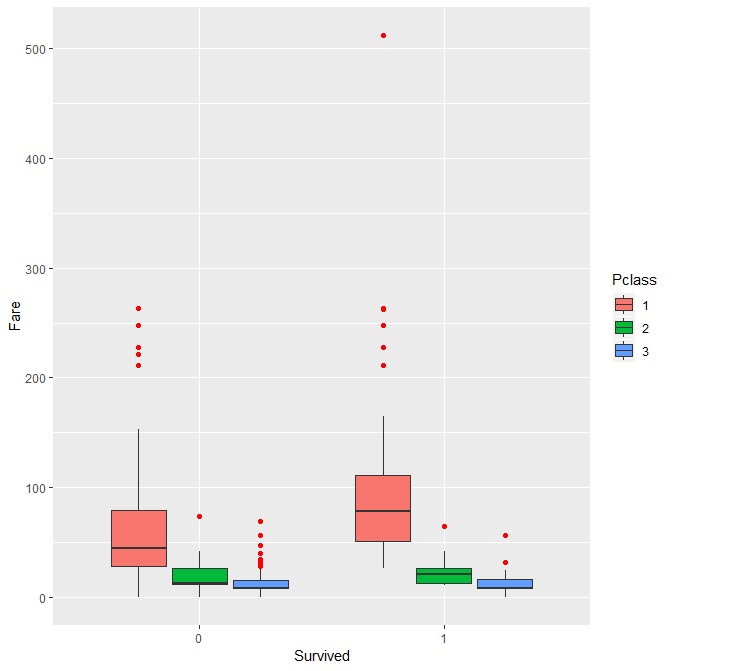
Table 6(Chances of Survived vs. Class with Sex)

As you can see by figure 15 females in each class have an overwhelming amount of survived compared to males, if we take a look at the first class were the amount of male survives is the highest it doesn't come anywhere near the amount of females this would be that females were allowed on the lifeboats first meaning more were able to get off the boat.

And to further reinforce our Hypothesis Table 6 shows the overall chances that you would survive based on your sex and class.

Hypothesis Eight - The amount of fare paid in each class directly relates to survival.

Alternative - The amount of fare paid in each class has no relation to survival.



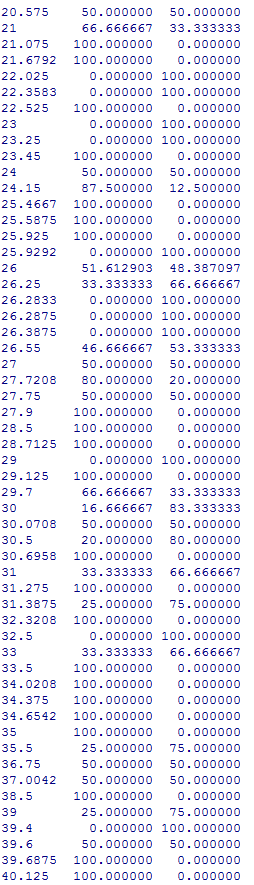
Figure 16(Fare and Class vs. Survived  

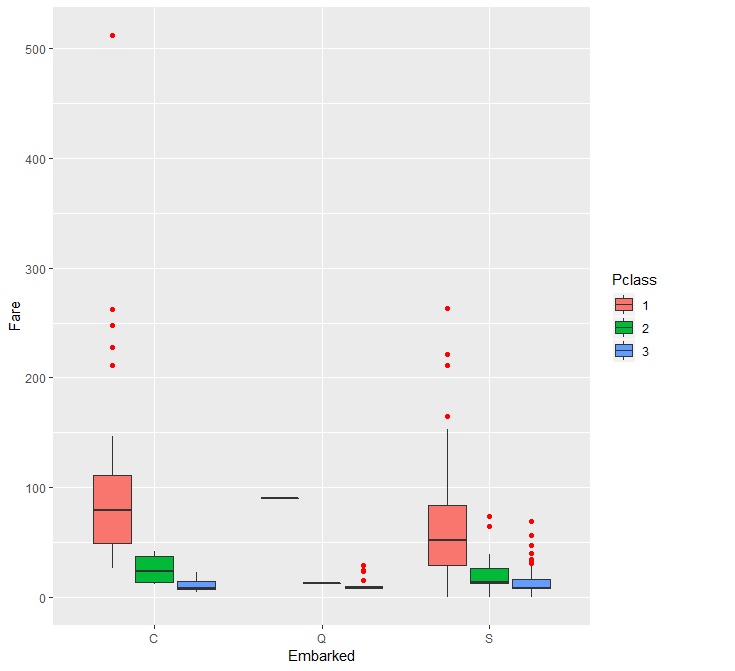

Table 7(Chances of Survive vs. Fare Sample)

As you can see by figure 16 the amount that people paid impacts their chances of survival if we look at class one you can see that out of the people that paid more their chances are a lot higher than the other first class that didn't pay as much this is shown to us by the black line in the centre of the box this tells us the medium value of their fare.

And to further reinforce our Hypothesis Table 7 shows the overall chances that you would survive based on your fare, this is a sample of the total fares starting at 20.575 and ending at 40.125.

Hypothesis Nine - People paid higher fares at Cherbourg on average in every class compared to the other places.

Alternative - People paid the same fares at Cherbourg as they would have at the other ports.

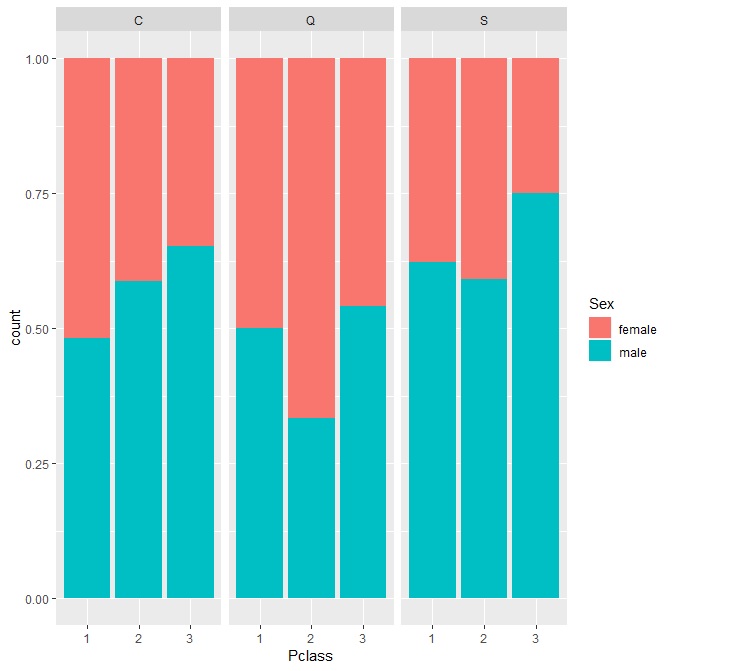
Figure (Average fares on each port with class)



As you can see by figure 17 the average fare prices in C are higher than those in Q and S in each class this is shown by the boxes and the position of the black line in the center of the box as you can see people that boarded at C in all class have a higher black line meaning that the people in C on average paid more. One of the reason for this could be because of the amount of people that boarded at those ports which looks more likely as in Q the boxes haven't formed meaning not many people boarded at Q. Another reason could be that fares were priced differently at each port also each class was priced differently at each port as well.

Hypothesis Ten - Most of the females boarded titanic in Queenstown and were most present in second class.

Alternative - Females boarded titanic evenly and were evenly spread out thought the different classes.

Figure (Class and Sex with Embark)  


As you can see by figure 18 the amount of females that boarded the ship in Queenstown compared to the other ports is the highest this tells us that most of the passages that came onto the ship in Queenstown were female. Also by looking at figure 18 it clearly shows us that the amount of females in second class was highest compared to the other two classes.

# Conclusion

In this work the titanic dataset was analysed to see what statistically data could be collected, using this data helped to create our hypothesises which allowed us to prove many different theories using the data that was collected. Our main goal was to prove that depending on which sex a person was would impact there chance of surviving the titanic event as detailed above female were more likely to survive regardless of any additional factors like class, this has been a constant fact as more hypothesis testing has been done to reinforce the main point that females were and did survive more compared to males.