SOR 1232

HYPOTHESIS TESTING AND STATISTICAL MODELLING

INVESTIGATING PASSENGERS' SURVIVAL RATE FROM THE RMS TITANIC (1912)

Compiled by: Daniel Sumler Tristan Oa Galea and Quentin Falzon

Tutor: Dr David Suda

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES (CONTINUED)

INTRODUCTION

Describing the Dataset

The dataset which we chose to analyze for this assignment contains data representing the survivors of the tragic Titanic incident. It includes several factors which may or may not have affected the number of passengers that survived.

We chose this dataset as it can help us come to an interesting conclusion of what variables affected the likelihood of survival of any individual aboard the Titanic on that fateful day. Such variables are listed on the next page.

Thus, the aim of this assignment is to determine if the given data can be manipulated through numerous tests and reveal any factors that affected a passenger's likelihood of survival.

An example of these tests was using simple box plots to compare age and survival rate. Other tests such as the Mann-Whitney test and Shapiro-Wilk test were also utilized, as can be seen later in this assignment.

The dataset used in this assignment can be found below through the following URL:

http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3.xls

The dataset was downloaded in XLS format. This was a challenge that we had to overcome, as this was to be used in the SPSS program. Therefore, we had to convert this file into .sav format which would allow it to be recognized and read by SPSS.

It should also be noted that there are multiple *Null* values across the .sav file. This is because not every piece of information could be gathered from the passengers on board due to various reasons

and complications. In order to keep the dataset as accurate as possible, these *Null* values were left untouched and not manipulated in any way.

DEFINITIONS

Factors

To explain in simple terms, a factor is the term used for a Categorical Variable. These can take numerical values from a predefined set of integers.

These factors are again split into two smaller groups, namely **Nominal** and **Ordinal** variables.

Nominal variables are ones which have been coded numerically but in which order is not important. A relevant example of a **nominal** variable would be *Sex*. In our dataset, *Sex* was changed to numerical values in order to comply with tests in SPSS. In this variable, *Female* is set as 0 while *Male* is set as 1. Other **nominal** variables can be found in the **Factors/Covariates Used** section of the assignment.

Ordinal variables are variables which have been coded numerically but in which order is important. For example, in our dataset, *Pclass* is an **ordinal** factor due to the importance of order between 1st, 2nd and 3rd class travel. This gives further importance to the fact that a sense of order is required when assigning this factor.

A **Factor Frequency** table was created in order to view how frequently factors appeared in the dataset. These can be seen in the **Descriptive Statistics & Illustrations** section of the documentation.

Covariates

To explain in simple terms, Covariates are often described as Quantitative variables.

A Quantitative variable describes something which can be counted or measured such as the numbers of persons in a queue or the Summer temperature.

Quantitative variables can either be **discrete** or **continuous**. A **discrete** covariate can be described as a value which is a whole number (does not under any circumstances contain a decimal point). To use an example from our dataset, the *Age* variable is a **discrete** covariate. This is because a person cannot be assigned an age which contains a decimal point. The same principle applies when describing the *Parch* variable.

On the other hand, a **continuous** covariate can be described as a value which contains a decimal point (although these values may be whole numbers from time to time – but that is up to chance). To use another example from our dataset, the *Fare* variable is a **continuous** covariate. A *Fare* may be a value which contains a decimal point (for example 54.99) or a value that is a whole number (for example 72.00).

Multiple **Covariate Frequency** tables were created and can be found in the **Descriptive Statistics & Illustrations** section of the documentation.

Dependent Variables

To explain in simple terms, a **dependent variable** is the opposite of an **independent variable**.

An **independent variable** is a variable which does not change as another variable changes. For example, the *Sex* of a passenger is an **independent variable** which does not change due to the result of another variable.

On the other hand, a **dependent variable** is one which changes when another variable changes. For example, if one were to calculate the total cost of an item with tax, and the **tax percentage** variable changed, the total cost of the item would also change.

After heavy consideration into our dataset, we discovered that the *Fare* variable is a dependent variable which depends on the *Pclass* variable. This was considered because, by using intuition, a higher class ticket would cost more, therefore making the *Fare* cost more.

In order to back up this claim, a **bar chart** and **table** have been created to display this information. These can be seen in the **Descriptive Statistics & Illustrations** section of the documentation, accompanied by a description of the outputs.

FACTORS / COVARIATEES USED

The variables used are as follows:

- Survived (Nominal Factor) - [Yes/No] -> Shows whether a passenger survived or not
- Sex (Nominal Factor) - [Male/Female] -> Shows the sex of the passenger
- Age (Scale Covariate) - [Discrete Range] -> Shows the age of the passenger
- Pclass (Ordinal Factor) - [1st, 2nd, 3rd] -> Shows the rank of class the passenger had
- Fare (Scale Covariate) - [Continuous Range] -> Shows how much a passenger paid to board – Dependent Variable – Depends on *Pclass*.
- Embarked (Nominal Factor) - [Cherbourg, Queenstown, Southampton] -> Shows where the passengers boarded the ship
- Sibsp (Scale Covariate) - [Discrete Range] -> Shows number of siblings/spouses onboard
- Parch (Scale Covariate) - [Discrete Range] -> Shows number of parents/children onboard

The Ordinal and Nominal Factors had to be given values in SPSS whilst the scale covariates were left as default. The values given to the factors and why they were given these values is described in the next section of this writeup.

Values given to the Factors

As previously mentioned, the different factors were given particular values related to them. Before this was done, some factors were refactored from a particular data type to another to be able to use them for certain tests. The ones which were refactored include the *sex* and *embarked* nominal factors

The sex covariate was originally a string value containing "Male" and "Female" as possible data selections. This data was then converted from string to integers to be able to be processed when performing certain tests (e.g. clustered bar graphs). This was converted using the following selections:

Transform -> Recode into different variables... -> New values were assigned to the previous string values in order to convert the data type.

The following values were encoded to these variables:

Pclass - 1 : "1st", 2 : "2nd", 3 : "3rd"

Survived - 0 : "no", 1 : "yes"

Sex : - 0 : "female", 1 : "male"

Embarked - 1 : "Cherbourg", 2 : "Queenstown", 3 : "Southampton"

AIMS AND OBJECTIVES

What we want to do with this dataset?

To answer this question clearly, with this dataset, we wish to conduct multiple tests in order to come to a valid conclusion on which factors affected the chance of survival of an individual aboard the Titanic. Factors include ones which are not seen to be obvious (for example, if one were to have a physical disability, their likelihood of survival would obviously be lower than someone who doesn't. Factors like these weren't included in this set for that obvious reason.).

Dependent Variables

From the previously stated list of Factors and Covariates, there is only one Covariate that depends on another Factor. From the below bar chart and table, it is obvious that the *Fare* covariate depends on the *Pclass* factor. The reason behind this is that a passenger with a high class ticket would have to pay a larger fare to obtain said ticket. One could also say that *Fare* and *Pclass* are directly proportional (as one goes up, the other goes up too).

Below, one can see a table which displays the mean of *Fare* with respect to which *Pclass* it falls under.

Means

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
fare * pclass	1308	99.8%	2	0.2%	1310	100.0%

Report

fare

pclass	Mean	N	Std. Deviation
1st	87.508992	323	80.4471782
2nd	21.179196	277	13.6071221
3rd	13.302889	708	11.4943585
Total	33.295479	1308	51.7586682

Table 3.1 - Means

As seen above, the first table depicts the **Case Processing Summary** which shows the amount of data entries which were used in order to calculate the mean. As can be seen, 99.8% of the entered data was used. The remaining 0.2% was left unused because a *Null* value was found in the dataset.

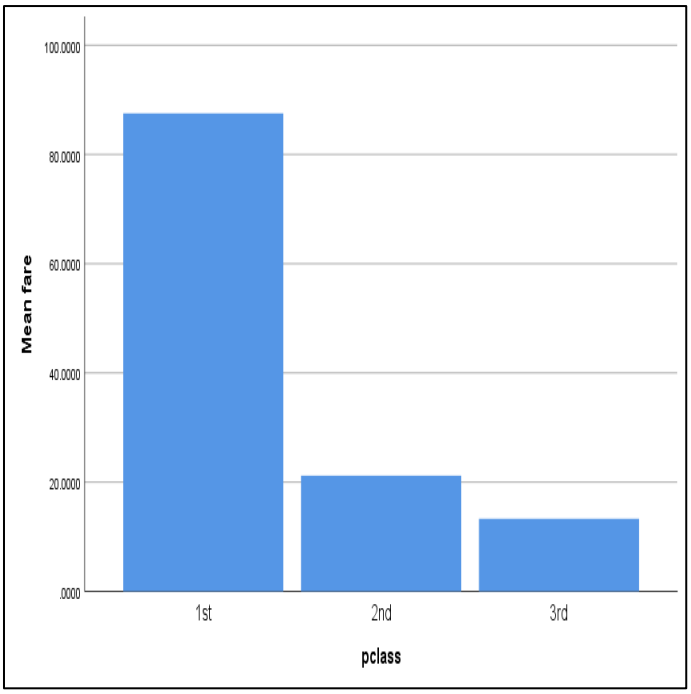

Figure 3.1 - A plot of Mean Fare against Passenger Class

The above Bar Graph depicts the mean of *Fare* on the y-axis and the category of *Pclass* in which they fall under. From these two graphs, it is clear to see that the *Fare* variable depends on the *Pclass* factor as there is significant differences between each category.

DESCRIPTIVE STATISTICS AND ILLUSTRATIONS

In this section of the documentation, we will be talking about the measures of location and dispersion of a number of variables in our data set. This topic is an essential one to cover as it is important to know how many variables we are dealing with in this situation.

First of all, factor frequency tables have been created (as was referenced in the **Factors** section previously. Below, one may see the frequency tables together with an explanation of the output.

survived

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	no	809	61.8	61.8	61.8
	yes	500	38.2	38.2	100.0
	Total	1309	99.9	100.0	
Missing	System	1	.1		
Total		1310	100.0		

Table 4.1 - Survived Descriptive Statistics

sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		1	.1	.1	.1
	female	466	35.6	35.6	35.6
	male	843	64.4	64.4	100.0
	Total	1310	100.0	100.0	

Table 4.2 - Sex Descriptive Statistics

pclass

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1st	323	24.7	24.7	24.7
	2nd	277	21.1	21.2	45.8
	3rd	709	54.1	54.2	100.0
	Total	1309	99.9	100.0	
Missing	System	1	.1		
Total		1310	100.0		

Table 4.3 - Passenger Class Descriptive Statistics

embarked

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		3	.2	.2	.2
	Cherbourg	270	20.6	20.6	20.8
	Queenstown	123	9.4	9.4	30.2
	Southampton	914	69.8	69.8	100.0
	Total	1310	100.0	100.0	

Table 4.4 - Embarked Descriptive Statistics

As we can see from the **Factor Frequency** tables above, the frequency of the *survived*, *sex*, *Pclass* and *embarked* factors from the dataset are displayed. The **Frequency** column shows the amount of times they appear in the dataset. The **Percent** column translates the frequency into an overall percentage. The **Valid Percent** column shows a percentage that does NOT include missing cases. The **Cumulative Percent** column sums up the current rows percentage with the previous row's percentage. Finally, the **Missing System** row name shows the amount of times a *Null* value was found to be in the dataset, and was not included in the analysis.

Measure of Location for Covariates

In order to calculate the measure of location for the covariates in our data set, the mean, median and mode were calculated. Covariates were chosen in this particular case because the fact that the values are numeric would make more sense in the context of calculating the mean, median and mode. The outputs are as shown below, followed by an explanation.

Statistics

		age	sibsp	parch	fare
N	Valid	1046	1309	1309	1308
	Missing	264	1	1	2
Mean		29.88	.50	.39	33.295479
Median		28.00	.00	.00	14.454200
Mode		24	0	0	8.0500

Table 4.5 - Covariate Statistics

As seen in the table above, the column names *Age*, *sibsp*, *parch* and *fare* all represent the covariate variables of the same name. The column **Valid** shows the amount of valid entries in the dataset for

each variable respectively, while the **Missing** row shows the amount of invalid entries in the dataset for each variable. The **Mean, Median** and **Mode** rows show the mean, median and mode values for each variable respectively. Below, one can see illustrations of the *age* and *fare* variables in the form of bar charts and histograms, and the *parch* and *sibsp* variables in the form of pie charts.

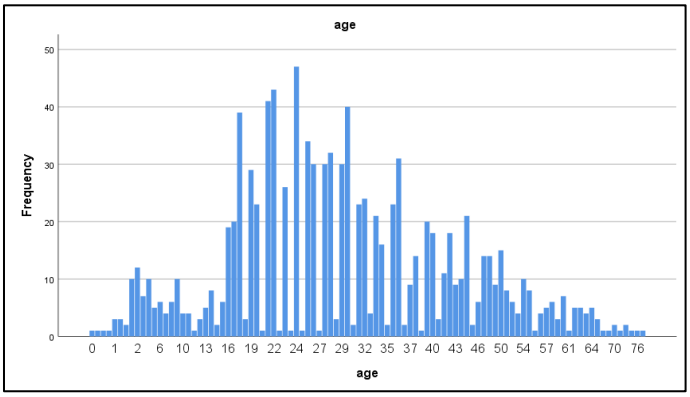

Figure 4.1 - Bar Chart showing Age Frequency

The above **bar chart** graph shows the amount of times each value of *age* was present in the dataset. A bar chart is frequently used to display discrete data (such as our *age* variable). The y-axis represents the amount of times the ages on the x-axis appear. For example, we can obviously tell that the mode of people aboard the titanic on that day were aged 24. On the other hand, multiple values of *age* appear just once in the dataset. This bar chart can also sufficiently be used in order to calculate the mean, median and mode of the data.

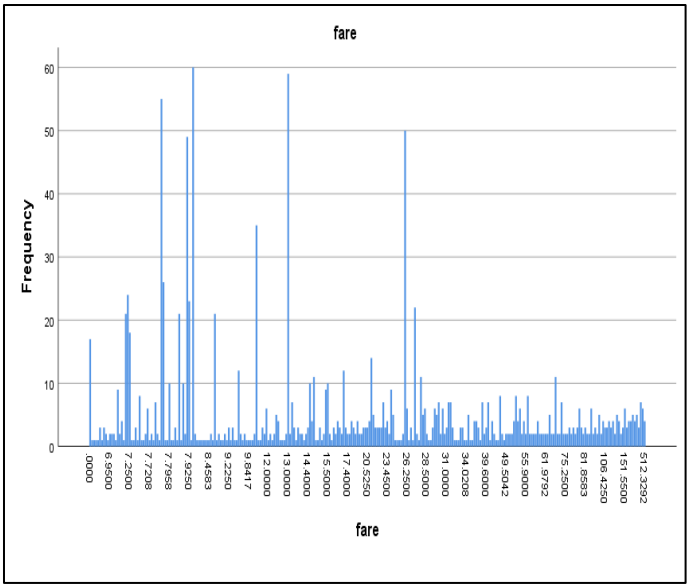

Figure 4.2 - Histogram showing Fare Values Frequency

The above diagram is a **histogram**. This type of chart is used specifically for continuous values of data. In our case, the *fare* values are continuous. In this chart, the *fare* variable values are plotted on the x-axis, while the y-axis contains the frequency of each value of *fare*. For example, by looking at the chart, we can see that the most common fare was between the 7.9 and 8.4 values. This type of chart comes in handy when representing continuous data and when calculating the mean, median and mode.

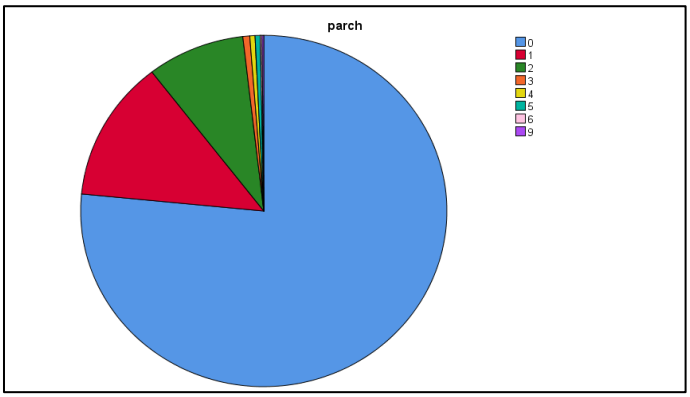

Figure 4.3 - Pie Chart showing different parch values proportionally

The above **pie chart** shows a colour-coded frequency chart for different values of the *parch* variable. Pie charts are used to display frequencies of discrete data. As we can see in the above

chart, the mode of *parch* in the dataset is 0, as it has the biggest chunk of the chart dedicated to it. The next highest frequency is 1, and so on. This chart is very good for representing discrete frequencies and calculating the mode (if it is obvious enough) but does not come in handy when calculating the mean or median. Below, one can find another example of a pie chart, this time being used with *sibsp*.

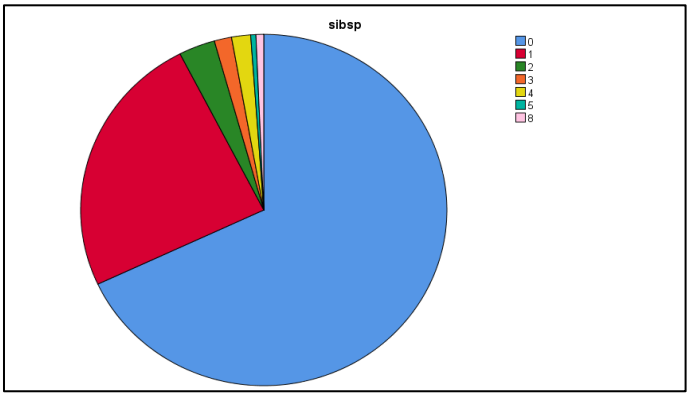

Figure 4.4 - Pie Chart showing different sibsp values proportionally

Measures of Dispersion

When talking about measures of dispersion of a dataset, we are specifically talking about calculations such as the **Range**, **Quartiles**, **Interquartiles**, **Standard Deviation** and **Variance**.

The **Range** refers to the maximum and minimum values in the variable.

The **Quartiles** are slightly more complicated. The dataset is split into four equal parts : the **lower quartile**, the **median**, the **upper quartile** and **above the upper quartile**. The quartiles are calculated by ordering the data into size order and determining the median. The data is now split into an upper and lower half based on the result of the median. The lower quartile is the middle of the lower half of the split, and the upper quartile is the middle of the upper half of the split.

The **Interquartile Ranges** are a measure of dispersion which are equal to the difference between the Upper Quartile and the Lower Quartile (both of which were discussed above). This is used in order to build boxplots (which are displayed below) and display a simple graph representation of probability.

The **Standard Deviation** is used in order to express by how much the values of a set of values differ from the mean value of the set. A low standard deviation means that the values of the set are close to the mean, while a high standard deviation means that the numbers are further away from the mean, more spread out.

The **Variance** is used in order to tell us how far a set of numbers is spread out from the mean. It describes how much a random variable value varies from its expected value.

Below is a table which describes the values of the above attributes in our dataset.

Statistics

		age	sibsp	parch	fare
N	Valid	1046	1309	1309	1308
	Missing	264	1	1	2
Std. Deviation		14.413	1.042	.866	51.7586682
Variance		207.749	1.085	.749	2678.960
Range		80	8	9	512.3292
Minimum		0	0	0	.0000
Maximum		80	8	9	512.3292
Percentiles	25	21.00	.00	.00	7.895800
	50	28.00	.00	.00	14.454200
	75	39.00	1.00	.00	31.275000

Table 4.6 - Measures of Dispersion

Below, are box plot diagrams, scatter diagrams and simple bar charts together with explanations detailing them.

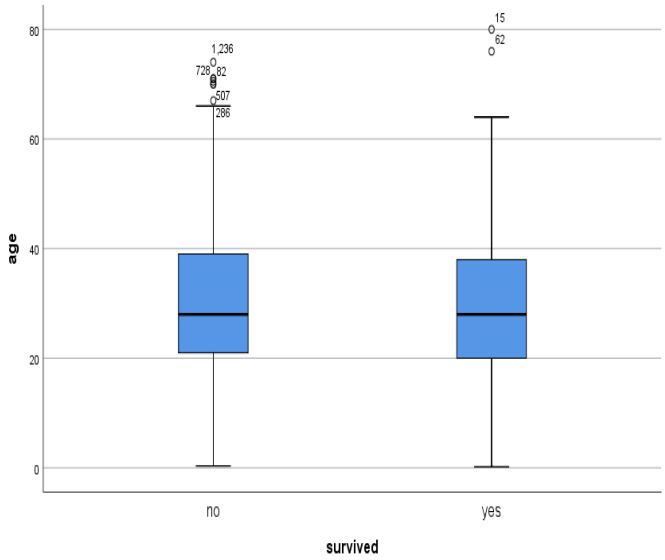

Figure 4.5 - Box Plot of Age against Survived

The diagram above is known as a box plot. In this diagram, the *survived* variable is plotted on the x-axis while the *age* variable is plotted on the y-axis, meaning there is a box plot for each x-axis state, *yes* and *no*. The lowest point of the graph is known as the **minimum**. This is the minimum value that occurred of that variable. For example, the youngest death on the titanic was a child aged just above the age of 0. The lower point of the blue box is called the **lower quartile** (this was explained above). For example, the **lower quartile** of the *yes* box is exactly 20. The line in the middle of the box represents the **median** of the data. In both of these boxes, the median lies between the 20 and 40 range of ages. The line at the top of the box represents the **upper quartile** of the data. In both cases, the **upper quartile** lies just below the *age* value 40 line. The line at the top of the chart represents the **maximum** value in the dataset. For example, in this case, at least two persons over the age of 60 were present in the dataset, with one surviving and the other perishing. Finally, the values seen above the top line are referred to as **outliers**. These are pieces of data that are an abnormal distance from other points in the graph. For example, 15 and 62 are outliers of the *yes* box.

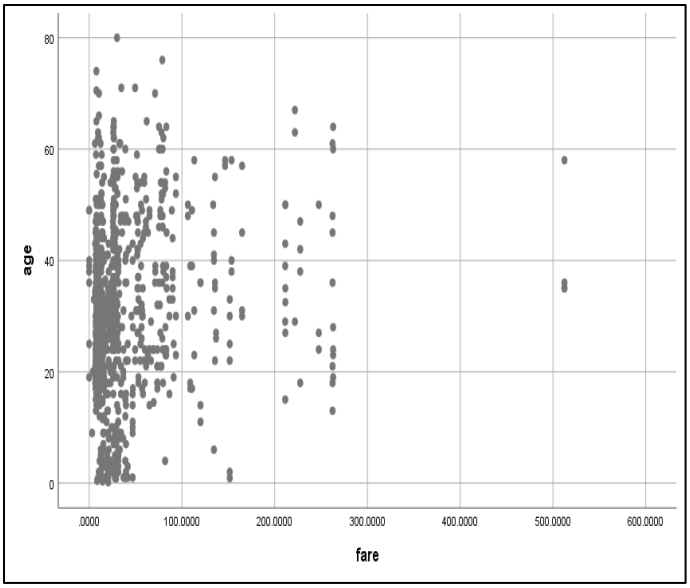

Figure 4.6 - Scatter Diagram of Age against Fare

Above, one can find a **scatter diagram.** Every dot on the diagram represents an entry of data in the dataset. Here, every *age* is plotted against every *fare*. A scatter diagram makes it easy for us to spot correlations between the two chosen variables. For example, in this case, we can see that the majority of people at all ages had a *fare* price under 100, the most being between the ages of 20 and 40. On the other end of the spectrum, we can see that the most expensive tickets (costing more than 500!) were bought by people in their late 30s and late 50s. It must be noted that a scatterplot only gives meaningful data based on the variables that are used. For example, if one were to use the *survival* variable against the *age* variable, every dot would be connected in a column-like way. In this case, there are much better diagrams to find this information, such as a bar chart.

PARAMETRIC / NON-PARAMETRIC TESTS

Shapiro-Wilk Test

The Shapiro-Wilk test is being conducted on this dataset due to it being less than 2000 elements. If the dataset contained more than 2000 elements, the Kolmogorov-Smirnov would have been used.

The Shapiro-Wilk Test is being conducted in order to find out whether the optimal test to perform would be an **Independent Sample T-Test** or a **Mann-Whitney Test**. In this case, the **Independent Sample T-test** requires a normal distribution of data, which is what the Shapiro-Wilk Test checks for. If the data is not normally distributed, the **Mann-Whitney Test** will be used.

Before examining the results for the Shapiro-Wilk Test, it is important to first state the Null Hypothesis and the Alternate Hypothesis:

H0 (Null Hypothesis) - The data is normally distributed. Therefore, Independent Sample T-Test is required to continue. In this case, H0 will be rejected if the p value is ≤ 0.05 and will be accepted if the p value is > 0.05.

H1 (Alternate Hypothesis) - The data is NOT normally distributed. Therefore, the Mann-Whitney Test is required to continue. H1 will be accepted if the p-value is ≤ 0.05.

All of the covariates in our dataset (*age*, *parch*, *sibsp*, *fare*) were put through the Shapiro-Wilk Test. The outputs that have been included are the table outputs and the graph outputs. Graph outputs were included due to the very low p-value. Tests were also conducted using *survived* as a factor, but the results of the test showed nothing meaningful, and were therefore not included. The following shows the Normality Tests for the covariates independently.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
age	.079	1045	.000	.980	1045	.000
sibsp	.364	1045	.000	.585	1045	.000
parch	.426	1045	.000	.559	1045	.000
fare	.282	1045	.000	.545	1045	.000

a. Lilliefors Significance Correction

Table 5.1 - Kolmogorov-Smirnov and Shapiro-Wilk test results

The table above shows the results of the Kolmogorov-Smirnov Test and the Shapiro-Wilk Test. As discussed previously, we shall only be observing the results of the Shapiro-Wilk Test due to the number of elements in the dataset. The important data in this table is the **Sig.** Column which represents the p value. As we can see, the p value is smaller than 0.05, meaning H0 is rejected and H1 is accepted. This therefore means that none of the covariates are normally distributed.

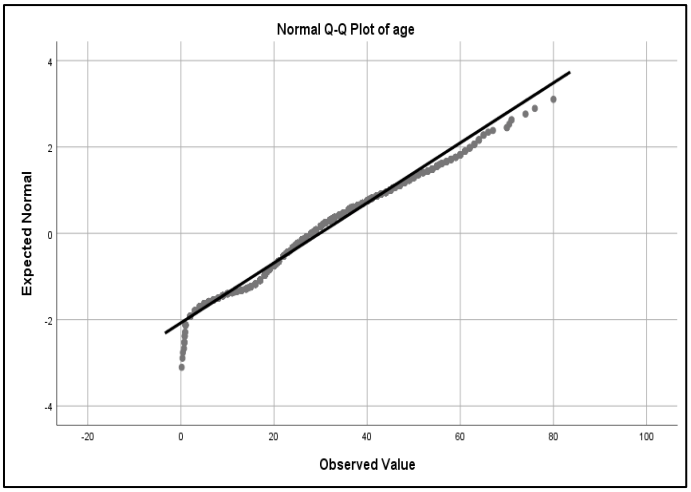

Figure 5.1 - Q-Q Plot for Age covariate

The above graph depicts a Q-Q Plot for the *age* covariate. The bold line in the middle represents what an ideal normal distribution should appear as. The black dots surrounding the bold line are the individual elements in our dataset. As we can see, the data clearly differs from the ideal normal distribution line, meaning that it is not normally distributed.

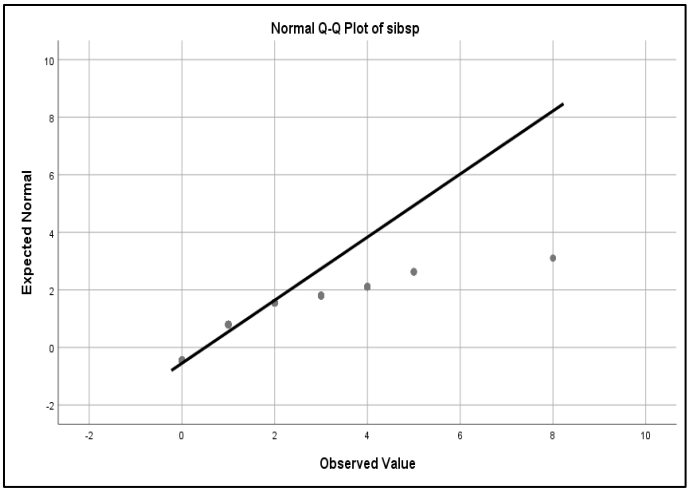

Figure 2.2 - Q-Q Plot for sibsp covariate

The graph above depicts a Q-Q Plot for the *sibsp* covariate. As we can see, there are much less values of *sibsp* in our dataset than in the first graph, but it is also evident that these values stray away from the ideal normal line. This shows that *sibsp* covariate is not normally distributed either.

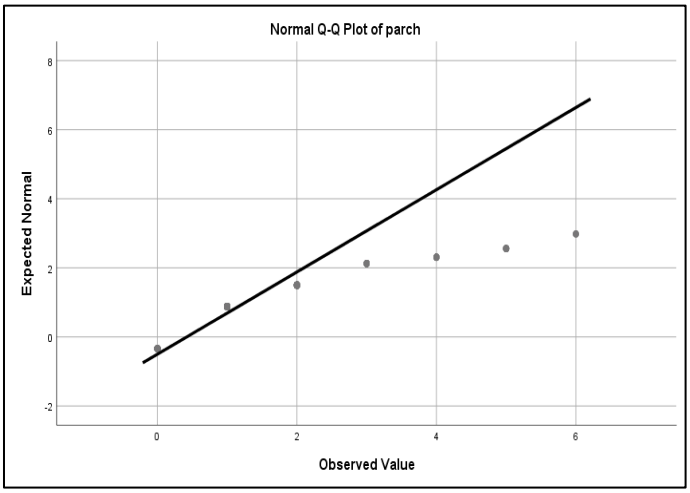

Figure 5.3 - Q-Q Plot for parch covariate

The graph above depicts another Q-Q Plot, this time in terms of the *parch* covariate. As per usual, the elements in the dataset significantly vary from the ideal normal distribution line. This therefore shows that the *parch* covariate is not normally distributed either.

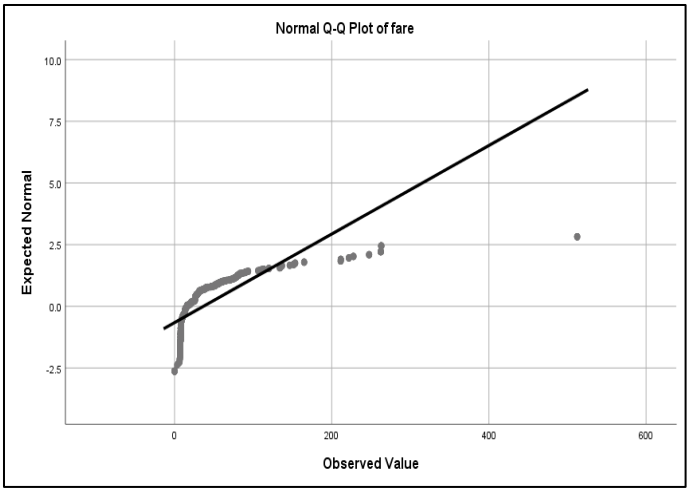

Figure 5.4 - Q-Q Plot for fare covariate

Finally, the last Q-Q Plot, which is in terms of the *fare* covariate, clearly shows that the elements differ from the ideal normal distribution line as per usual. This shows that the *fare* covariate is not normally distributed, like the other covariates.

Mann-Whitney Test

Due to our result in the Shapiro-Wilk Test, it was evident that in order to continue, we were going to have to use the Mann-Whitney Test.

The Mann-Whitney Test is used to compare differences between two independent groups. In our case, we are trying to compare whether the covariates in our dataset had any influence over the *survived* factor (for example, whether the age of a person had any influence over whether they survived or not). After these tests, we will be able to draw our own conclusions about these matters, while also supplying sufficient evidence in the form of test outputs.

First of all, the test variables which were used were the *age*, *parch*, *sibsp* and *fare* covariates. These were put up against the Grouping Variable, which in our case was the *survived* factor.

Ranks

	survived	N	Mean Rank	Sum of Ranks
age	no	619	533.95	330513.50
	yes	427	508.35	217067.50
	Total	1046		
sibsp	no	809	634.67	513444.00
	yes	500	687.90	343951.00
	Total	1309		
parch	no	809	619.33	501041.50
	yes	500	712.71	356353.50
	Total	1309		
fare	no	808	567.19	458288.50
	yes	500	795.60	397797.50
	Total	1308		

Table 5.2 - Mann-Whitney Test results

The above table is returned as a result of the Mann-Whitney Test. Every covariate variable is separated into two categories, the *survived yes* and *survived no* categories. On top of this, one can also see the most important column, the **Mean Rank** column. This column shows how the mean varies from one category to another. This output alone means nothing to us though, we must first have confirmation about whether the means are **Statistically Significant**. To explain in simple terms, **Statistically Significant** means that a certain outcome has happened not by chance. For example, if we were to say that *age* is statistically significant, that would mean that the larger **Mean Rank** value for the *survived : no* category, wouldn't be down to chance, meaning that *age* did affect whether a person would have survived or not.

In order to tell whether we are dealing with statistical significance, we will have to form some hypotheses :

H_0 (Null Hypothesis) - The median of the different groups is NOT significantly different. H_0 is accepted if $p > 0.05$.

H_1 (Alternate Hypothesis) - The median of the different groups is significantly different. H_1 is accepted if $p < 0.05$.

These hypotheses must now be checked on each covariate value in order to accept/reject them.

Test Statistics[a]

	age	sibsp	parch	fare
Mann-Whitney U	125689.500	185799.000	173396.500	131452.500
Wilcoxon W	217067.500	513444.000	501041.500	458288.500
Z	-1.347	-3.024	-5.862	-10.629
Asymp. Sig. (2-tailed)	.178	.002	.000	.000

a. Grouping Variable: survived

Table 5.3 - Further results from Mann-Whitney Test

The above table shows further results from the Mann-Whitney Test. The most important result is found in the **Asymp. Sig. (2-tailed)** row. This shows us the p-value of the respected covariates. We will now address the hypotheses for each covariate respectively :

- *age* – the p-value for the *age* covariate is > 0.05, meaning that the Null Hypothesis H_0 is accepted, while the Alternate Hypothesis H_1 is rejected. This means that the **Mean Rank** of age is not significantly different, and whether a person survived or not was down to chance.
- *sibsp* – the p-value for the *sibsp* covariate is < 0.05, meaning that the Alternate Hypothesis is accepted, while the Null Hypothesis is rejected. This means that the **Mean Rank** of siblings was not down to chance and had an influence over whether a passenger survived or not. *For example, this MAY be because siblings stuck close to eachother while the incident was happening and may have either survived or died together.*
- *parch* - the p-value for the *parch* covariate is < 0.05, meaning that the Alternate Hypothesis is accepted while the Null Hypothesis is rejected. This means that the **Mean Rank** of parents/children onboard was not down to chance and had an influence over whether a passenger survived or not. *For example, this MAY be because parents and children were prioritized when helping people escape from the sinking ship.*
- *fare* – The p-value of the *fare* covariate is < 0.05, meaning that the Alternate Hypothesis is accepted while the Null Hypothesis is rejected. This means that the **Mean Rank** of fare price was not down to chance and had an influence over whether a passenger survived or

not. *For example, this MAY be because passengers who paid a higher fare price were prioritized over people who paid lower fare prices during the rescue mission.*

To sum up, from these results, we can see that the **Mean Ranks** of the *sibsp* (people with a higher number of siblings onboard survived), *parch* (people with a higher number of parents/children onboard survived) and *fare* (more people who paid a higher fare survived when compared to those who paid a lesser fare) covariates are significantly different, meaning that they had a role to play in whether a passenger survived or not. The same cannot be said about the *age* covariate though, as its **Mean Rank** was proven to not be significantly different, meaning it had no influence over whether a passenger survived or not.

Chi-Square Test

After conducting these previous tests, our next step was to conduct the **Chi-Square Test**. This test is being conducted in order to check whether there is an association between the *survived* variable and the qualitative variables – meaning the *sex, embarked,* and *class* variables. The choice to use this test came from the fact that the independence of the quantitative variables (*age, sibsp, parch* and *fare*) has already been tested using the **Mann-Whitney test**, which is a test that is unable to measure associativity between qualitative variables.

First of all, we must define the hypotheses which we will be testing. In this case they are:

H0 (Null Hypothesis) - if p-value > 0.05, then there is no association between the two variables.

H1 (Alternate Hypothesis) - if p-value < 0.05, then there is an association between the two variables.

In each of the following tests, we will obtain a **Chi-Square Value** which shall be used in order to determine whether the Null Hypothesis or the Alternate Hypothesis will be accepted.

The **Degree of Freedom** is also shown in each test. This can also be calculated by using the formula :

(Number of rows –1) * (Number of Columns –1)

We will be looking at the 2-sided **Asymp. Sig.** Output to check if the p-value adheres to the Null or Alternate Hypothesis.

Each **Chi-Square Test** has the following output :

X^2 (degree of freedom) = Chi-Square Value, p < Asymp. Sig. (2-Sided).

In total, three **Chi-Square test** were conducted, one for each qualitative variable.

1. ***sex* against *survived*** – association

In order to perform this test, the variables are put into a crosstab table. *Sex* is placed in the Rows window and *survived* is placed in the Columns window. The table below shows the Observed Value of the crossed variables as well as the Expected Values.

sex * survived Crosstabulation

			survived		
			no	yes	Total
sex	female	Count	127	339	466
		Expected Count	288.0	178.0	466.0
	male	Count	682	161	843
		Expected Count	521.0	322.0	843.0
Total		Count	809	500	1309
		Expected Count	809.0	500.0	1309.0

Table 5.4 - Crosstab table for Chi-Square Test

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	365.887[a]	1	.000		
Continuity Correction[b]	363.618	1	.000		
Likelihood Ratio	372.921	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	365.607	1	.000		
N of Valid Cases	1309				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 178.00.

b. Computed only for a 2x2 table

Table 5.5 - Chi-Square results (sex against survived)

The table above shows the result of the **Chi-Square Test**. The **Pearson Chi-Square Value** (365.887) is the most important value which will be used in order to reject or accept our Null

Hypothesis. The p-value is in fact calculated by the answer of P(X^2 > 365.887). In this case p < 0.001.

The **Degree of Freedom** is also shown to be 1.

The 2-sided **Asymp. Sig.** Output shows the p-value to be p < 0.001. This shows that p < 0.05, meaning that the Null Hypothesis is rejected and the Alternate Hypothesis is accepted. Therefore, this shows that there IS an association between the two variables. **X^2 (2) = 365.887, p < 0.0017**

2. ***Pclass* against *survived* -** association

pclass * survived Crosstabulation

			survived		
			no	yes	Total
pclass	1st	Count	123	200	323
		Expected Count	199.6	123.4	323.0
	2nd	Count	158	119	277
		Expected Count	171.2	105.8	277.0
	3rd	Count	528	181	709
		Expected Count	438.2	270.8	709.0
Total		Count	809	500	1309
		Expected Count	809.0	500.0	1309.0

Table 5.6 - Crosstab between pclass and survived variables

The above table is portraying the crosstabulation between the *Pclass* and *survived* variables. We can again view the observed values and the expected values. These are calculated in the same way the first test was.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	127.859[a]	2	.000
Likelihood Ratio	127.765	2	.000
Linear-by-Linear Association	127.709	1	.000
N of Valid Cases	1309		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 105.81.

Table 5.7 - Chi-Square results (pclass against survived)

The above table again shows the **Pearson Chi-Square Value** together with the degree of freedom and the **Asymp. Sig.** Output. As we can see, the p-value is again $p < 0.001$, meaning that the Null Hypothesis is again rejected and the Alternate Hypothesis is accepted. Therefore, this means that there IS an association between the two variables. $\boldsymbol{X^2}$ **(2) = 127.859, p < 0.001**

3. *Embarked* against *survived* - *associated*

embarked * survived Crosstabulation

			survived		
			no	yes	Total
embarked	Cherbourg	Count	120	150	270
		Expected Count	167.1	102.9	270.0
	Queenstown	Count	79	44	123
		Expected Count	76.1	46.9	123.0
	Southampton	Count	610	304	914
		Expected Count	565.7	348.3	914.0
Total		Count	809	498	1307
		Expected Count	809.0	498.0	1307.0

Table 5.8 - Crosstab between embarked and survived variables

The above table shows the crosstabulation between the *embarked* and *survived* variables. Observed Values and Expected Values are shown as usual. These are calculated the same way the first two tests were.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	44.242[a]	2	.000
Likelihood Ratio	43.173	2	.000
Linear-by-Linear Association	40.821	1	.000
N of Valid Cases	1307		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 46.87.

Table 5.9 - Chi-Square results (embarked against survived)

The above table shows us the **Pearson Chi-Square Value** as usual, the degrees of freedom as well as the usual **Asymp. Sig.** Value. Like in the previous tests, the p-value is $p < 0.001$, which again

tells us that we must reject the Null Hypothesis and accept the Alternate Hypothesis. Therefore, the two variables are associated. $\boldsymbol{X^2}$ **(2) = 44.242, p < 0.001**

In conclusion to our **Chi-Square Tests**, it seems that all three quantitative variables are associated with the *survived* variable due to the p-values leading us to reject the Null Hypothesis and accept the Alternate Hypothesis. This gives us more insight into what variables affect the *survival* rate, and therefore brings us closer to answering our initial question.

BINARY LOGISTIC REGRESSION

A binary logistic regression model is a regression model used to model the effect of explanatory variables on a response variable, in which the possible outcomes of the latter can only comprise of two categories.

For this assignment, 2 binary logistic regressions were performed in which one is more accurate than the other as will be explained soon. Due to this similarity, both will be explained in general and reference to either the first or the second one will be made accordingly when needed.

Dummy variables were used in both regressions for the pclass ordinal and embarked nominal variables as can be seen below:

dummyclass1	dummyclass2	dummyembarked1	dummyembarked2
1.00	.00	.00	.00
1.00	.00	.00	.00
1.00	.00	.00	.00
1.00	.00	.00	.00
1.00	.00	.00	.00
1.00	.00	.00	.00
1.00	.00	.00	.00
1.00	.00	.00	.00
1.00	.00	.00	.00
1.00	.00	1.00	.00
1.00	.00	1.00	.00
1.00	.00	1.00	.00
1.00	.00	1.00	.00

Table 6.1 - Dummy Variables

Now that these were created, it was time to move on to testing of the different variables. To be able to detect possible cases of multicollinearity in the explanatory variables of the general linear model to be created, a Spearman Correlation and a Test Binary Logistic Regression were drawn up.

Starting off with the Spearman Correlation, the following variables were analyzed against each other:

- survived
- age
- sibsp
- parch
- fare
- sex
- dummyclass1
- dummyclass2
- dummyembarked1
- dummyembarked2

This gave a Spearman Correlation which can be seen below:

Correlations

			survived	age	sibsp	parch	fare	sex	dummyclass1	dummyclass2	dummyembarked1	dummyembarked2
Spearman's rho	survived	Correlation Coefficient	1.000	-.042	.084**	.162**	.294**	-.529**	.279**	.051	.183**	-.015
		Sig. (2-tailed)	.	.178	.002	.000	.000	.000	.000	.066	.000	.576
		N	1309	1046	1309	1309	1308	1309	1309	1309	1307	1307
	age	Correlation Coefficient	-.042	1.000	-.130**	-.216**	.193**	.063*	.378**	.005	.084**	-.029
		Sig. (2-tailed)	.178	.	.000	.000	.000	.042	.000	.877	.007	.350
		N	1046	1046	1046	1046	1045	1046	1046	1046	1044	1044
	sibsp	Correlation Coefficient	.084**	-.130**	1.000	.438**	.446**	-.181**	.060*	.012	.031	-.084**
		Sig. (2-tailed)	.002	.000	.	.000	.000	.000	.031	.668	.270	.002
		N	1309	1046	1309	1309	1308	1309	1309	1309	1307	1307
	parch	Correlation Coefficient	.162**	-.216**	.438**	1.000	.400**	-.245**	.016	.022	.036	-.124**
		Sig. (2-tailed)	.000	.000	.000	.	.000	.000	.563	.435	.197	.000
		N	1309	1046	1309	1309	1308	1309	1309	1309	1307	1307
	fare	Correlation Coefficient	.294**	.193**	.446**	.400**	1.000	-.242**	.653**	.093**	.216**	-.267**
		Sig. (2-tailed)	.000	.000	.000	.000	.	.000	.000	.001	.000	.000
		N	1308	1045	1308	1308	1308	1308	1308	1308	1306	1306
	sex	Correlation Coefficient	-.529**	.063*	-.181**	-.245**	-.242**	1.000	-.107**	-.029	-.068*	-.089**
		Sig. (2-tailed)	.000	.042	.000	.000	.000	.	.000	.297	.014	.001
		N	1309	1046	1309	1309	1308	1309	1309	1309	1307	1307
	dummyclass1	Correlation Coefficient	.279**	.378**	.060*	.016	.653**	-.107**	1.000	-.297**	.328**	-.166**
		Sig. (2-tailed)	.000	.000	.031	.563	.000	.000	.	.000	.000	.000
		N	1309	1046	1309	1309	1308	1309	1309	1309	1307	1307
	dummyclass2	Correlation Coefficient	.051	.005	.012	.022	.093**	-.029	-.297**	1.000	-.135**	-.122**
		Sig. (2-tailed)	.066	.877	.668	.435	.001	.297	.000	.	.000	.000
		N	1309	1046	1309	1309	1308	1309	1309	1309	1307	1307
	dummyembarked1	Correlation Coefficient	.183**	.084**	.031	.036	.216**	-.068*	.328**	-.135**	1.000	-.164**
		Sig. (2-tailed)	.000	.007	.270	.197	.000	.014	.000	.000	.	.000
		N	1307	1044	1307	1307	1306	1307	1307	1307	1307	1307
	dummyembarked2	Correlation Coefficient	-.015	-.029	-.084**	-.124**	-.267**	-.089**	-.166**	-.122**	-.164**	1.000
		Sig. (2-tailed)	.576	.350	.002	.000	.000	.001	.000	.000	.000	.
		N	1307	1044	1307	1307	1306	1307	1307	1307	1307	1307

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 6.2 - Spearman Correlation

From this correlation, the correlation coefficient of each individual crossed pair was checked. Pairs crossed against themselves giving a coefficient value of 1 were ignored whilst all the others having a value greater than 0.5 (which is considered to be quite high) were listed down. The ones standing out are listed below:

- survived crossed with sex at -0.529 collinearity
- fare crossed with dummyclass1 at 0.653 collinearity

However, the purpose of the Spearman Correlation is specifically to identify any clashes with the dummy variables created. Thus, it was clear that fare and class would not be of significance if placed in the same regression model as these gave a high collinearity value.

Moving on, the next step was to create the Test Binary Logistic Regression. This regression was created by selecting survived as the response variable together with age, sibsp, parch, fare, sex

and embarked as explanatory variables. In this regression model, the sole purpose was to simply check for any p-values greater than 0.05. The relevant output considered for this particular test can be observed below:

Likelihood Ratio Tests

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept	975.735[a]	.000	0	.
age	986.260	10.525	1	.001
sibsp	988.426	12.691	1	.000
parch	976.260	.525	1	.469
fare	1001.619	25.885	1	.000
sex	1249.172	273.437	1	.000
embarked	1004.436	28.702	2	.000

Table 6.3 - Likelihood Ratio Test

The table outputted above known as the table for the Likelihood Ratio Tests contains some information related to the exploratory variables. Other tables were also outputted when performing this regression model but these will not be mentioned as they are irrelevant for the purpose of this test. However, the only section relevant to us is the significant value column above denoted by Sig. .This column represents the significant values of all the separate exploratory variables taken into consideration. From all, parch is the only one considered to be insignificant as it's value is much higher than 0.05. Thus, as a final conclusion from this regression test, it was decided to exclude parch completely from any future binary logistic regressions.

As a general conclusion from the correlation and regression test, the following had to be applied for successful Binary Logistic Regressions:

- Apart from all the explanatory variables, only one of either class or fare could be used in the same regression
- parch cannot be used in any regression

With these outliners set and keeping in mind that survived is the response variable to be used in each case, 2 Binary Logistic Regressions were created to be able to view the different models separately as well as identify which one gives a more accurate prediction.

Starting from the first regression (which will be referred to as Regression A), this was created by setting survived as the response variable together with age, sibsp, fare, sex and embarked as explanatory variables. This regression obeys both rules set because neither parch nor class show up as part of the explanatory variables given that fare has already been chosen as one of them. Similarly, the second regression (Regression B) was formulated by allocating survived as the response variable once again together with age, sibsp, pclass, sex and embarked as the corresponding explanatory variables. Regression B is similar to A in the sense that the only difference between the 2 is that one makes use of the fare explanatory variable whilst the other uses pclass instead.

The first table showing up for the models is the Case Processing Summary which gives a brief overview of data item information. It gives count information on the different variables as well as computes a percentage value corresponding to that count. Further count information is then given at the bottom of the table relating the valid, missing, total and subpopulation values. The tables for both models A (top) and B (bottom) can be viewed below:

Case Processing Summary

		N	Marginal Percentage
survived	no	618	59.3%
	yes	425	40.7%
sex	female	386	37.0%
	male	657	63.0%
embarked	Cherbourg	212	20.3%
	Queenstown	50	4.8%
	Southampton	781	74.9%
Valid		1043	100.0%
Missing		267	
Total		1310	
Subpopulation		947[a]	

a. The dependent variable has only one value observed in 924 (97.6%) subpopulations.

Table 6.4 - Case Processing Summary for Model A

Case Processing Summary

		N	Marginal Percentage
survived	no	619	59.3%
	yes	425	40.7%
pclass	1st	282	27.0%
	2nd	261	25.0%
	3rd	501	48.0%
sex	female	386	37.0%
	male	658	63.0%
embarked	Cherbourg	212	20.3%
	Queenstown	50	4.8%
	Southampton	782	74.9%
Valid		1044	100.0%
Missing		266	
Total		1310	
Subpopulation		637[a]	

a. The dependent variable has only one value observed in 567 (89.0%) subpopulations.

Table 6.5 - Case Processing Summary for Model B

Next, we have a Model Fitting Information table for both A (top) and B (bottom). The

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1373.848	1378.798	1371.848			
Final	990.260	1024.909	976.260	395.589	6	.000

Table 6.6 - Model Fitting Information for model A

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1213.276	1218.226	1211.276			
Final	771.191	810.797	755.191	456.085	7	.000

Table 6.7 - Model Fitting Information for model B

The Goodness-of-Fit table up next assesses the Pearson discrepancy between the current model and the full model while also assessing the Deviance discrepancy between the current and full model as well. This also outputs the separate degree of freedom for each case and the corresponding significant value. The table for Regression A (top) while that of Regression B can be seen after.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	927.173	940	.611
Deviance	942.622	940	.470

Table 6.8 - Goodness of Fit for Regression A

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	651.800	629	.257
Deviance	639.810	629	.374

Table 6.9 - Goodness of Fit for Regression B

The next table in considered one of the most important tables in the binary logistic regression output as it gives the actual relationship between the explanatory variables and the response variable. The tables seen in the screenshots below correspond to Regression A (top) and B (bottom) respectively. Based on the relationship between all the explanatory variables and the response variable survived, the values below were obtained for each case. Taking into consideration just the age variable since the concept is similar for all the variables, the different values in the table will be explained. Starting out with B, this can be understood as the odds ratio when taking the exponential value of B (separate column). This means that the older a person was, the age together with exponent value of B give the chance of survival. The last 2 columns give the 95% calculated confidence interval for each case. The sig. column gives the significant value of that particular variable which as can be observed in all cases, neither exceeds 0.05 due to the regulations set at the start. The df column gives the degree of freedom for each case whilst the standard error and Wald test statistics are also provided. The middle columns essentially give information about significance test for that estimated coefficient.

Parameter Estimates

survived[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
no	Intercept	1.097	.207	28.027	1	.000			
	age	.018	.006	9.881	1	.002	1.019	1.007	1.030
	sibsp	.377	.102	13.625	1	.000	1.459	1.194	1.782
	fare	-.009	.002	20.140	1	.000	.991	.987	.995
	[sex=0]	-2.534	.168	227.663	1	.000	.079	.057	.110
	[sex=1]	0[b]	.	.	0	.	.	.	.
	[embarked=1]	-.856	.203	17.787	1	.000	.425	.285	.632
	[embarked=2]	1.121	.393	8.149	1	.004	3.068	1.421	6.625
	[embarked=3]	0[b]	.	.	0	.	.	.	.

a. The reference category is: yes.

b. This parameter is set to zero because it is redundant.

Table 6.10 - Parameter Estimates for Regression A

Parameter Estimates

survived[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
no	Intercept	1.054	.228	21.382	1	.000			
	age	.038	.007	32.705	1	.000	1.039	1.026	1.053
	sibsp	.332	.103	10.401	1	.001	1.394	1.139	1.706
	[pclass=1]	-2.071	.239	75.170	1	.000	.126	.079	.201
	[pclass=2]	-.944	.203	21.637	1	.000	.389	.261	.579
	[pclass=3]	0[b]	.	.	0	.	.	.	.
	[sex=0]	-2.633	.176	222.882	1	.000	.072	.051	.102
	[sex=1]	0[b]	.	.	0	.	.	.	.
	[embarked=1]	-.669	.213	9.881	1	.002	.512	.338	.777
	[embarked=2]	.802	.409	3.856	1	.050	2.231	1.002	4.970
	[embarked=3]	0[b]	.	.	0	.	.	.	.

a. The reference category is: yes.

b. This parameter is set to zero because it is redundant.

Table 6.11 - Parameter Estimates for Regression B

The final table to be compared for both tests is the Classification Table. This is a fairly straightforward table giving the final comparison of accuracy of the binary logistic regression. Comparing both A's (left) table and B's (right) table, we are able to identify which of the 2 tests was most accurate. In hindsight, both seem to be more or less of the same accuracy, however the regression B performed gave 1 % more accuracy to the model computed. Thus, these 2 tables can be used as a conclusion to the aim of these regressions, which was comparing both and finding the most accurate one.

Classification

Observed	Predicted		Percent Correct
	no	yes	
no	524	94	84.8%
yes	138	287	67.5%
Overall Percentage	63.5%	36.5%	77.8%

Table 6.12 - Concluding Classification Table for Regression A

Classification

Observed	Predicted no	Predicted yes	Percent Correct
no	526	93	85.0%
yes	128	297	69.9%
Overall Percentage	62.6%	37.4%	78.8%

Table 6.13 - Concluding Classification Table for Regression B

Following this, the estimated probabilities of a person surviving or not were calculated. A sample of the answers obtained can be viewed below where EST1_1 = 1- EST2_1.

EST1_1	EST2_1
.07	.93
.34	.66
.04	.96
.61	.39
.09	.91
.70	.30
.29	.71
.62	.38
.28	.72
.74	.26
.61	.39
.04	.96
.03	.97
.07	.93
.89	.11
.	.
.32	.68
.08	.92
.04	.96
.42	.58
.68	.32

Table 6.14 - Survival Probabilities

In the end, analysis of leverage values and residuals were performed together with the computation of the Cook's Distance as can be seen in the sample screenshot taken below:

PRE_1	PCP_1	ACP_1	Leverage	PearsonResid...	CooksDistance
1	.93	.93	.004	-.281	.000
1	.66	.66	.023	-.723	.002
1	.96	.04	.004	5.055	.013
0	.61	.61	.009	.792	.001
1	.91	.09	.005	3.252	.006
0	.70	.30	.008	-1.510	.002
1	.71	.71	.014	-.637	.001
0	.62	.62	.008	.787	.001
1	.72	.72	.014	-.621	.001
0	.74	.74	.016	.595	.001
0	.61	.61	.011	.798	.001
1	.96	.96	.003	-.192	.000
1	.97	.97	.002	-.183	.000
1	.93	.93	.004	-.265	.000
0	.89	.11	.010	-2.790	.010
.	.	.	.	.	.
1	.68	.32	.011	1.466	.003
1	.92	.92	.005	-.301	.000
1	.96	.96	.003	-.213	.000
1	.58	.42	.010	1.164	.002
0	.68	.32	.008	-1.444	.002
1	.82	.82	.007	-.469	.000
1	.67	.67	.011	-.709	.001

Table 6.15 - Cook's Distance, Pearson Residual and other Values

CONCLUSION

Throughout the course of this assignment, we were required to analyze our data by conducting basic descriptive tests. These helped give us a better understanding and a holistic view of the data we were dealing with. Based off of this, certain relevant parametric / non-parametric tests were chosen in order to help us reject or accept the hypothesis in each case. For example, the Mann-Whitney test revealed that the passengers' age had no substantial effect on their survival rate. However, as revealed by the Chi-Square test the passengers' class on board the ship did affect their chances of survival. This makes sense as they were most probably given different levels of priority, or been located at a safer part of the vessel. Finally, two binary logistic regression models were fitted onto the data, differing from each other by omitting one variable in the second regression. It was found from the respective classification tables that regression B was overall 1% more accurate than regression A.