Fundamentals of Information Visualisation (COMP3042) Coursework

Student Name – Eashan Yasinda Samaranayake AKD (20311843)

Chapter 1: Description of Data

The dataset used for this project is the Titanic dataset. The **goal** is to identify the features that affected the survival rate of the passengers on the Titanic ship on April 14th, 1912. Table 1 shows the different types of features and a short description.

Table 1– *Description of the Features*

Column Name (Feature)	Data Type	Description
PassengerId	Integer (Ordered), Depends	Each passenger's ID in the dataset.
Survived	Integer (Nominal), Measure, Bivariate	Mentions if the passenger survived or not.
Pclass	Integer (Ordered), Dimension, Trivariate	The passenger's cabin class.
Name	String (Nominal), Depends	The passenger's name.
Age	Integer (Q-Ratio), Dimension	The passenger's age.
SibSp	Integer (Ordered), Dimension, Hyper-variate	How many siblings/spouses the passenger had on board?
Parch	Integer (Ordered), Dimension, Hyper-variate	How many parents/children the passenger had on board.
Ticket	String (Nominal), Dimension	Ticket number.
Fare	Float (Q-Interval), Dimension	Ticket (fair) price.
Cabin	String (Nominal), Dimension	Passenger's cabin number.
Embarked	Character (Nominal), Dimension, Trivariate	Which station the passenger embarked from.
Gender	Integer (Nominal), Dimension, Bivariate	Passenger's gender.
Name Title	String (Nominal), Depends	Passenger's name title (Mr, Mrs, etc).

Chapter 2: Objective, Questions and Audience

Objective – To identify how wealth, age, gender, sibling/spouse count, parent/children count, and how the embarkation point impacted the rates of survival rates of the passengers.

Audience – Researchers/Scientists that are keen on the events from the past such as the Titanic or any person that may find answers to the following questions in Table 2 insightful.

Table 2 – *Questions to be Answered*

Initial Questions	Further Questions
How did the wealth of a passenger affect their survival?	What is the effect of Gender on Class as a Subcategory?
How did gender affect the survival rate?	What is the effect of having Siblings/Spouses on board?
How did the age of passengers play a role in their survival?	What is the effect of having Parents/Children on board?
	Where did most travellers board from?
	How much did each person pay based on fair/class?
	Which Embarkation Point has the lowest rate of survival? And why?

Chapter 3: Initial Pre-processing of Data

Note – specific cleaning is done later throughout the project in addition to these preliminary cleaning.

Initially, the missing values were handled. **Cabin**, **Age,** and **Embarked** columns had 77%, 19.9%, and 0.22% missing values respectively. The **Cabin** column was deleted since the majority of the data is missing and it's not a valuable feature to predict a passenger's survival rate. The mean of the age column (29) was used to fill in the missing values. The most frequently occurring embarkation point (Southampton or 'S') was used to replace the 2 missing values in **Embarked**.

Next, **Survived** column was converted to Factor type while the **Age** column was converted to Integer Type. The numbers in the **Pclass** column were changed to their respective classes (1 to "First Class") and the numbers in **Survived** column were changed to their meaning (1 to "Survived" and 0 to "Did not Survive").

In addition, the **Gender** column is in 1 and 0. However, it's important to identify if 1 is Male or Female before further analysis. Therefore, using the **NameTitle** column, mapping has been performed where any passenger with **NameTitle** "Mr", "Dr", "Master", "Don", "Rev", "Major", "Sir", "Col", "Jonkheer" are mapped (renamed) to "Male". The rest are mapped to "Female" in the **Gender** column.

Finally, other unnecessary columns were removed to make the dataset compact and ready for analysis. **PassengerId**, **Ticket, and NameTitle** columns were removed as they didn't contribute any important information.

Chapter 4: Fitness of Data

All the questions listed in Chapter 2 can be answered completely by this dataset used after the pre-processing mentioned in Chapter 3.

Unit of Analysis – The key behind this analysis is to understand the different features that affected the rate of survival for different passengers on the Titanic. The dependent attribute (measure) in this dataset is the **Survived** column.

Access to Data – I could obtain the dataset from Kaggle.

Data Allowance – Since the dataset is not sensitive/confidential, I could use all the columns for the analysis if needed. However, as mentioned in Chapter 3, the data was cleaned thoroughly before use.

Data Usability – The dataset did have a lot of missing values, but they were all handled. Certain outliers were rejected from some plots for more reliable visualisations. Necessary data cleaning was done when needed throughout this project.

Data Accuracy and Consent – The data contained in the dataset is accurate and verified by Lake Forest College University (US). The university has publicly published this dataset for use by any student/developer.

Chapter 5: Initial Questions

5.1 **Question:** How did the wealth of a passenger affect their survival?

A great way to refer to a passenger's wealth is the cabin class they were in as ticket prices for a first-class are much more expensive than a third-class ticket for example (this will be analysed in section 5.6 in this report).

Data Cleaning/Filtering – A subset of the original data frame was created with the filter where all passengers survived (*Survived="Survived*"). The graphs that are created from this data frame (Figure 1,2) are **interactive**. Only the Pclass column with filtering was used to create the pie charts. The passenger counts were mutated to obtain the percentages column (pie chart slices).

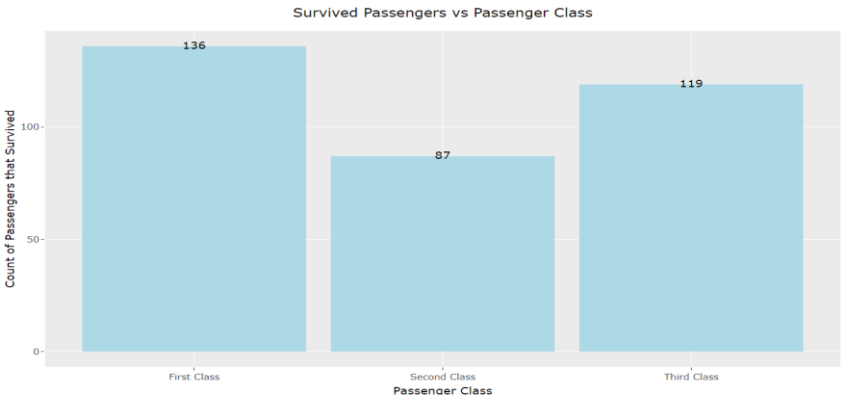


Figure 1: Bar chart (Passenger Count vs Class)

- Reason for this Visualisation - A bar chart was chosen since they can visualize the count of the people that survived with increased readability. This bar chart follows the law of **figure/ground** and the **law of similarity** (by following the same colour tone).

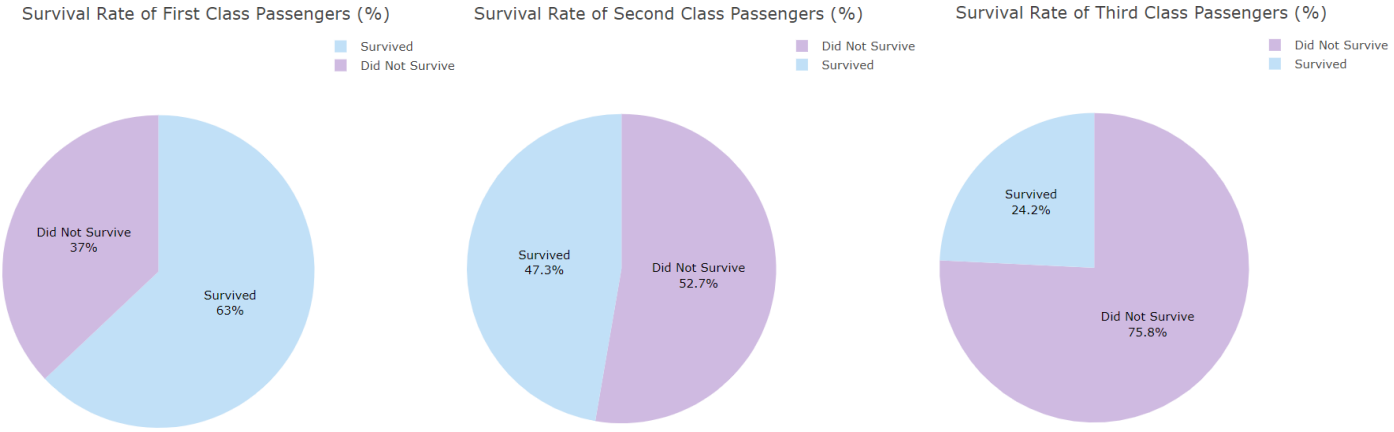


Figure 2: Pie Chart (Survival Rate vs Class)

- Reason for this Visualisation - However, it's even more accurate to know the percentage of how many passengers in each class survived as the passengers in each class are not constant. For this, pie charts will be used. The colours for both the bar chart and pie chart are kept constant since singular colours (**law of simplicity**) make it easier for the reader to understand the analysis than multiple, bright colours.

Answer – As seen in Figure 1 and Figure 2, it's clear that passengers in the Third Class had the lowest survival rate (**24%** survival). Being in First Class led to the best rate of survival (**63%** survival) while Second class led to the next best chances of survival (**47.3%**). So, it's safe that the passenger's wealth had an impact on their survival.

5.2 **Question:** How did gender affect the survival rate?

Upon understanding the effect wealth had on survival, let's understand the effect gender had on survival. To do so, the counts of the surviving passengers can be analysed.

- Data Cleaning/Filtering – A smaller data frame was created where the **Gender** column was manipulated to obtain the counts of each gender. The other features were discarded. The resulting graph shown below is **interactive**.

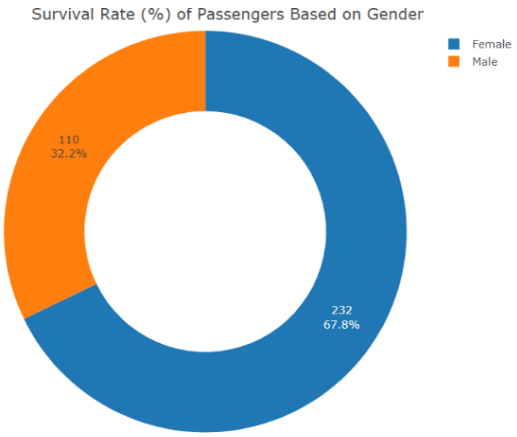


Figure 3: Donut Chart (Survival Rate/Count vs Gender)

- Reason for this Visualisation - A donut chart is used for this visualisation. This is because each slice of the donut represents a gender. It's also easier to understand due to the bitonal colours being used. It also aligns with the **law of figure/ground** as it's plotted on a white background for a complete contrast.

Answer – It's evident that from the 342 passengers that survived, **232(67.84%)** were female while only **110(32.16%)** were male. This is because of the Birkenhead drill imposed in 1892. It stated that women and children must always be saved first in a life-threatening situation.

5.3 **Question:** How did the age of passengers play a role in their survival?

To understand the link between age and survival, the dataset was split into two subsets which include children (aged under 18 years) and adults. This makes the analysis succinct.

- Data Cleaning/Filtering – Initially, 2 subset data frames were created where the **Age** column of one data frame was less than 18 and the other was greater than or equal to 18 to classify the children and adults respectively. These 2 data frames were further filtered based on the **Survived** column. Using these 4 data frames, the pie charts and density-enabled histograms were obtained. All graphs are **interactive**.

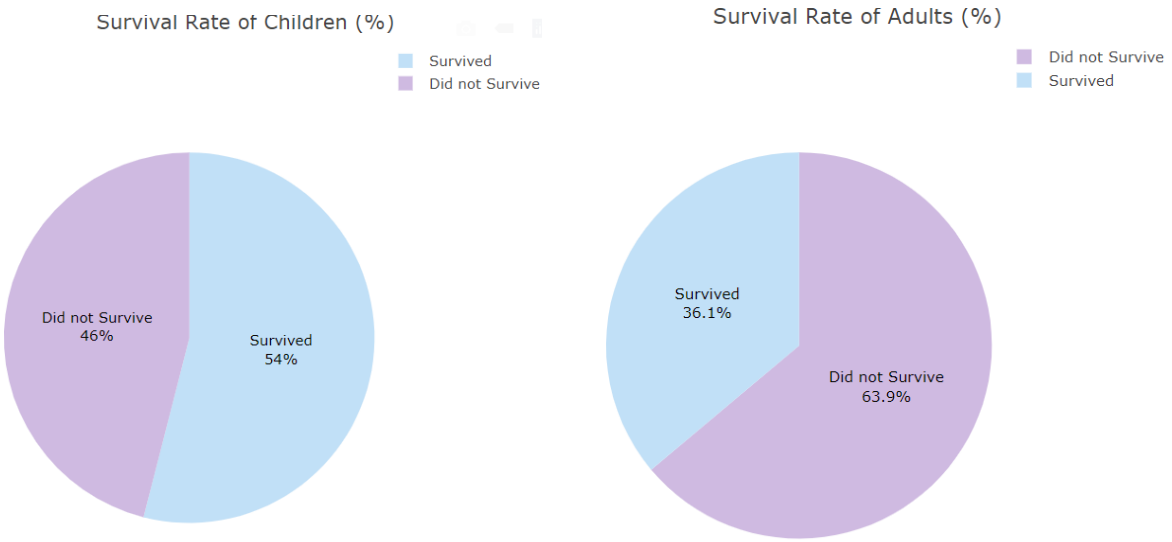


Figure 4: Pie Chart (Age Survival Rates)

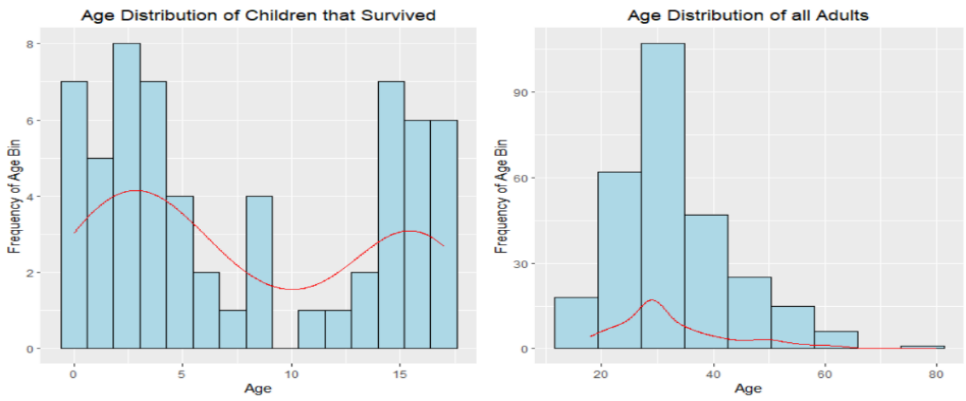


Figure 5: Density Curve Enabled Histogram (Age vs Death Frequency)

- Reason for this Visualisation - A pie chart was chosen to provide the percentages of survival of the children/adults while the density curve supported histogram gives a clearer understanding of the frequency of children/adults that survived for each age group. The histogram was used here to represent continuous, quantitative, and univariate data such as **Age**.

Answer – From Figure 4, most of the children (**53.98%**) managed to survive. This is once again due to the Birkenhead rule which prioritised the survival of women and children in the case of an emergency. Most of the children that survived were from the ages **1-5** and **15,16** (Figure 5). On the other hand, adults had a generally lower survival rate **(36.12%)** and as shown on the density line, most adults that lived were from the ages of **25-35**.

Chapter 6: Further Questions

6.1 **Question**: What was the effect of Gender with Class as a Subcategory?

In 4.1 and 4.2, the effects of wealth and gender were visualised, respectively. However, what happens if they are visualised in depth, together? 4.1 mentioned that passengers in first class had the best rate of survival while 4.2 mentioned that women had a higher chance of survival. So, for example, would a chance of a man in first class be higher than a woman in third class?

- Data Cleaning/Filtering – Only the **Pclass** and **Gender** column was filtered for this. However, obtaining the Gender column consisted of heavy data cleaning and mapping as explained previously in Table 2. The generated heatmap is **interactive**.

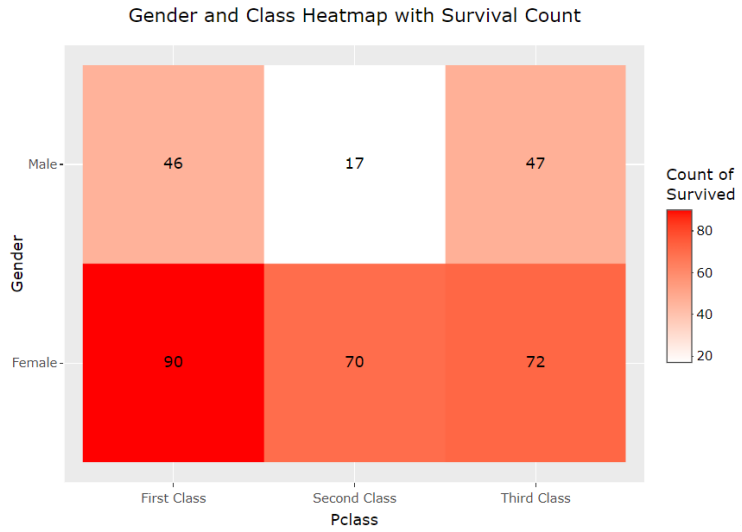


Figure 6: Heatmap (Gender vs Class Death Count)

- Reason for this Visualisation - A heatmap was chosen for this visualisation. This was because this question requires cyclical data to be displayed. It makes it easy for the reader to get a quick count of how many men in a certain class lived compared to how many women in a certain class lived.

Answer – Figure 6 states that more women in the third class (**72**) survived over the men in the first class (**46**) despite there being more men onboard as seen from the visualisations before. This proves that wealthier men didn't have any advantage over women in any class (as women had higher survival counts).

6.2 **Question:** Effect of having Siblings/Spouse Onboard

One column (**Parch**) in the dataset mentions the count of siblings or spouses a certain passenger had on board with them. It would be interesting to see if this affected the chances of survival as they might want to save their siblings/spouse in the case of an emergency.

- Data Cleaning/Filtering – Only the **Survived** measure and **SibSp** dimension were filtered for this. However, to create the facet grid, the original dataset was filtered with the different counts of the Sibling/Spouse (**SibSp**) column. All the graphs created are **interactive**.

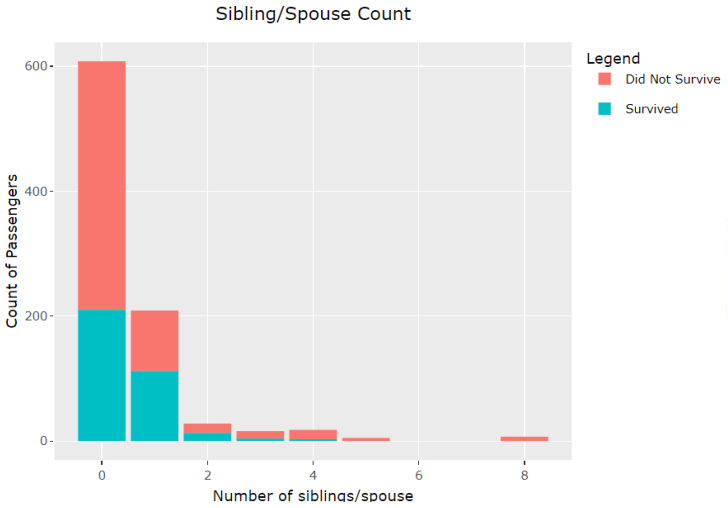


Figure 7: Stacked bar chart
(Distinct Sibling/Spouse Count and Effect on Survival)

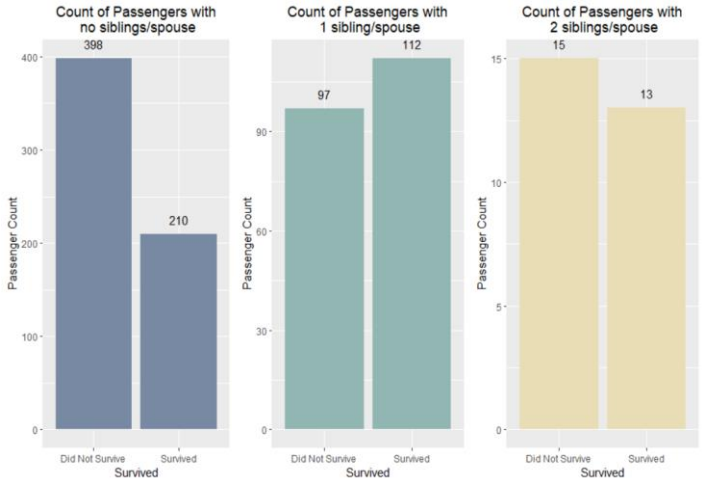


Figure 8: Facet Grid of Bar Charts
(Survival Rate of 0,1,2 Sibling/Spouse Count)

- Reason for this Visualisation - To understand the correlation between survival rate and the number of siblings/spouses, a stacked bar chart was used. Stacked bar charts allow the reader to get a glance at the different groups underneath one category. This graph follows the **figure/ground** and **simplicity** laws. This interactive graph enables the user to see the exact counts by hovering the cursor over each stack in the bar. From the graph obtained in Figure 8, a zoomed-in version supported with a facet grid enables us to analyse deeper into the death and survived count of passengers with 0 siblings/spouse, 1 sibling/spouse and 2 siblings/spouse (Figure 7). A facet grid was chosen with a bar chart to distinguish the different sibling/spouse counts and to individually focus on the counts. It also prevents the need for redundant graphs. This grid follows the **law of simplicity** and **figure/ground**.

Answer - Passengers with **5** siblings/spouses or above did not survive, this may be because they tried to save the family before them. Most passengers with no other siblings/spouse on board had a high death rate (**398**) while numbers 1-3 (**97, 15, 12**) had a decent rate of survival. Having just 1 sibling or spouse on board led to an increased survival rate of **53.5%** while any other count of siblings/spouse(s) led to a higher chance of death.

6.3 **Question:** What was the effect of having parents/children on board?

As 5.1 and 5.2 discussed the effects of having siblings/spouse(s) on board, this section will dive into the effects of having parents/children on board. Did it also impact a passenger's survival rate?

- Data Cleaning/Filtering – This analysis focuses mostly on the **Parch** column (and **Survived** column). An initial bar graph was plotted to understand the overall survival count of the parent/children numbers. It was clear that only 0, 1, and 2 had a decent count to make a comparison. So, a filter was used on the original dataset to obtain where the **Parch** column was equal to 0,1,2. The visualisations are **interactive**.

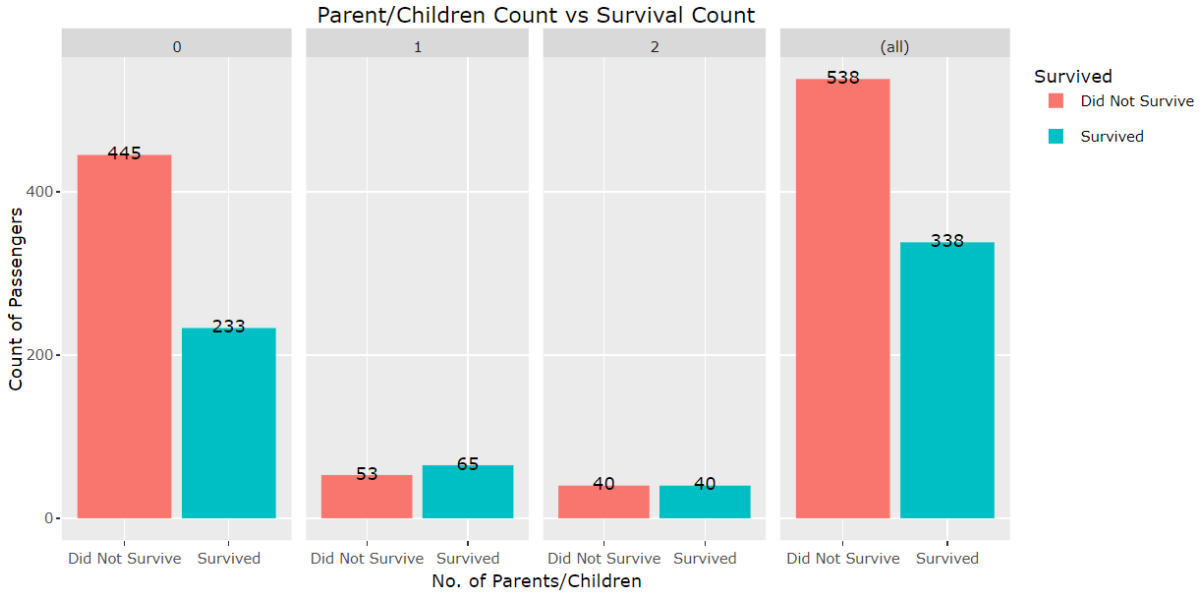


Figure 9: Facet Grid of Bar Charts (Survival Rate of 0,1,2 Parents/Children)

- Reason for this Visualisation - A facet graph allows multiple bar charts to react on 2 categorical variables which makes it easier to understand the correlations as the number of parents/children is changed.

Answer – Seems like having just 1 parent or child on board with a passenger increased their chances of survival by **55%**. This may be because they helped each other out in case of emergency.

6.4 **Question:** Where did most travellers board from?

Upon performing further analysis, it's evident that not all travellers boarded in from the same embarkation point. The titanic had three embarkation points. It would be interesting to understand where most passengers got in from.

- Data Cleaning/Filtering – A filter was used on the original dataset to obtain 3 different data frames where the **Embarked** column had distinct values (Embarked='S', Embarked='C', Embarked='Q'). These data frames were mutated and used to produce the pie charts and the bar charts. All the visualisations are **interactive**.

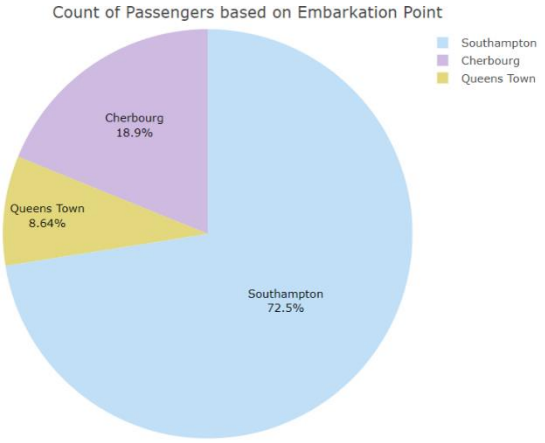


Figure 10: Pie Chart (Percentage of Passengers vs Embarkation

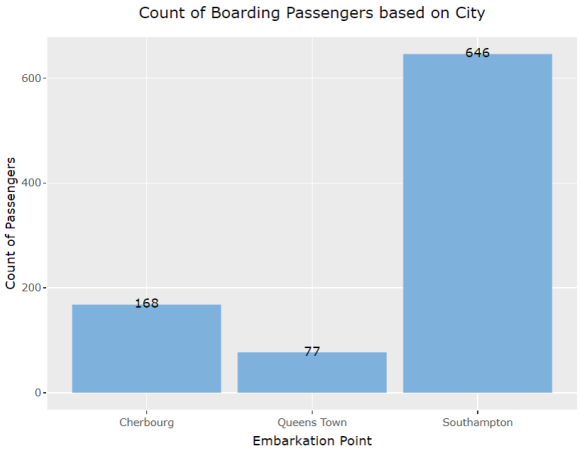


Figure 11: Bar Chart (Count of Passengers vs Embarkation Points)

- Reason for this Visualisation - The use of a pie chart makes it easy to see the distinct embarkation points and their percentages. When the pie chart is combined with a bar chart as shown on the right, we can understand the count of each embarkation point as well.

Answer – As shown in figures 10 and 11, the embarkation point with the highest number of passengers is Southampton with **646 passengers (72.5%)**.

6.5 **Question:** How much did each person pay based on fair/class?

Though there were three distinct passenger classes as seen before, the costs paid by each passenger to board into their respective classes were different. The ticket price for a certain class was not fixed. Therefore, let's see the average prices per class on the Titanic.

- Data Cleaning/Filtering – The **Fare** column was filtered out. Outlier handling was performed when plotting a boxplot on the **Fare** column. This is because some people in first class paid over 300 for the ticket. This made the graph stretch out a lot. Therefore, a y-axis limit was set to 300 which makes the other 2 classes more visible on the boxplot.

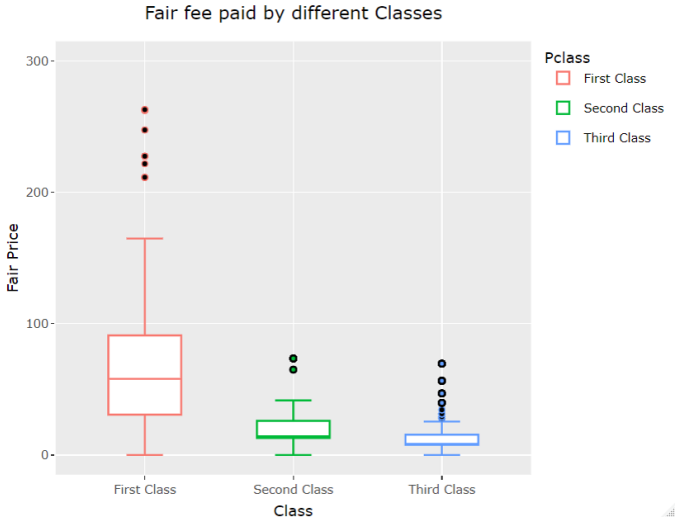


Figure 12: Boxplot (Ticket price vs Class)

- Reason for this Visualisation - A boxplot was used to depict the differences in the ticket prices for each passenger class. This is because a boxplot can show the median, interquartile range, boundaries, and outliers clearly. The plot also follows the rule of figure/ground to make the data readable. **Colour** was used as a perceptual property to emphasize the different classes.

Answer – The average (median) ticket price for first class, second class and third class were **57.98, 14.25 and 8.05** respectively.

6.6 **Question:** Which Embarkation Point has the lowest rate of survival?

Now that we've seen which embarkation point had the highest number of passengers (**Southampton**), with further analysis, it's possible to see which embarkation point led to the lowest survival rate.

- Data Cleaning/Filtering **-** This question used filtering on **Embarked** column to identify the different embarkation points as done in section 6.4. The resulting visualisation is an interactive heatmap and facet grid.

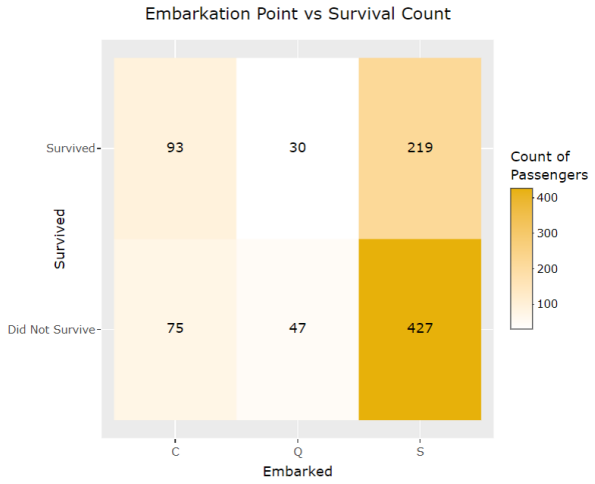


Figure 13: Heatmap (Embarkation Point vs Survival Counts)

- Reason for this Visualisation - A heatmap is chosen for this as this consists of cyclical data. A gradient that is easily identified is chosen so the reader can get a quick correlation between the embarkation point and the survival counts. **Density** (the depth of colour) was used as visual encoding.

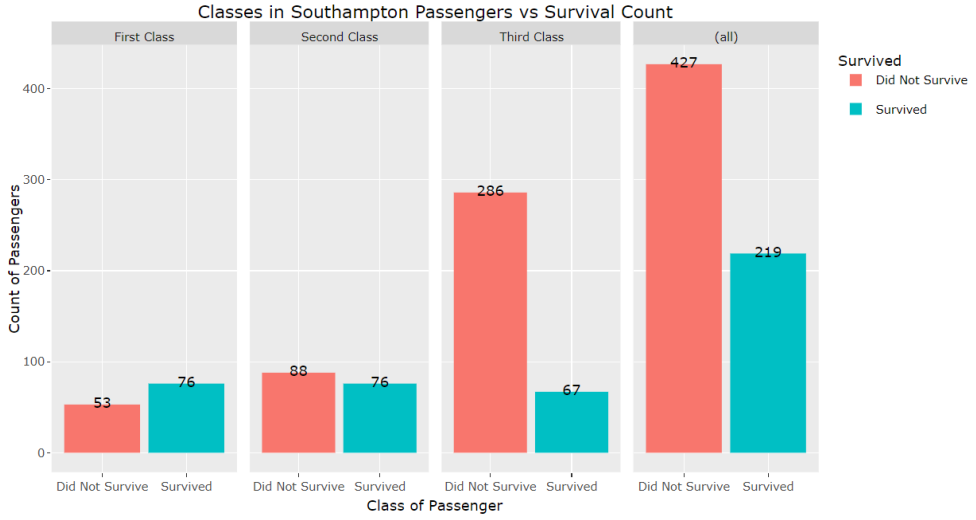


Figure 14: Facet Grid (Southampton vs Survival Counts for each Class)

Answer – The embarkation point with the most deaths was Southampton with a **66.1%** death rate. Most embarking from Cherbourg had a survival rate of **55.4%**. The reason behind Southampton having such a high death rate is that they have a high count of third-class passengers **(54.6%)**. As seen in figure 14, third-class passengers had a high death count as seen in Figure 14.

Chapter 7: Reflection on the Development Process

I started working on this coursework very early on. So, as I learnt more from the lecture and lab sessions, I learnt that there were better, more effective methods to create visualisations using R. Though I am new to R, I have done a few internships and even jobs for my university on data analysis/visualisation.

Initially, I created my basic plots using the built-in R functions. These graphs were not visually pleasing and did not have a lot of space for manipulation. I then used libraries such as **dply** and **tidyverse** to make handling data frames easier. Next, I was able to improve my plots using the libraries such as **ggplot2**, and **gridExtra**. However, I also wanted to make the graphs interactive. This was when I learnt about **plotly** from Dr Marina and Dr Hafeez in the lecture/lab sessions.

All in all, R is a great language to perform data analysis/visualisations in. Coming from a python-heavy background, I'm thankful to my lecturers for making me fluent in R as it will help me greatly in the industry.

Appendix

This section consists of the same figures from the report in larger dimensions to make it easier for the reader/examiner to read it if required.

Figure 1

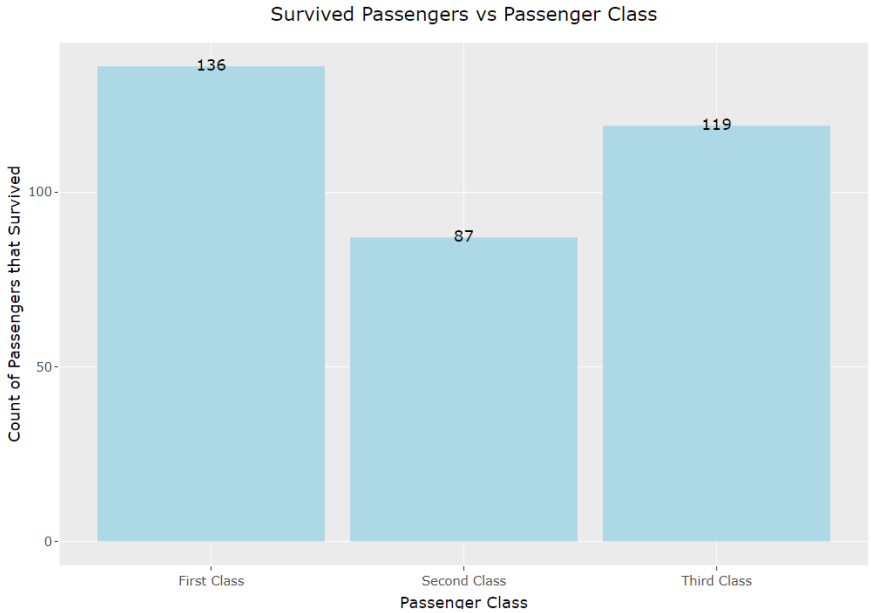
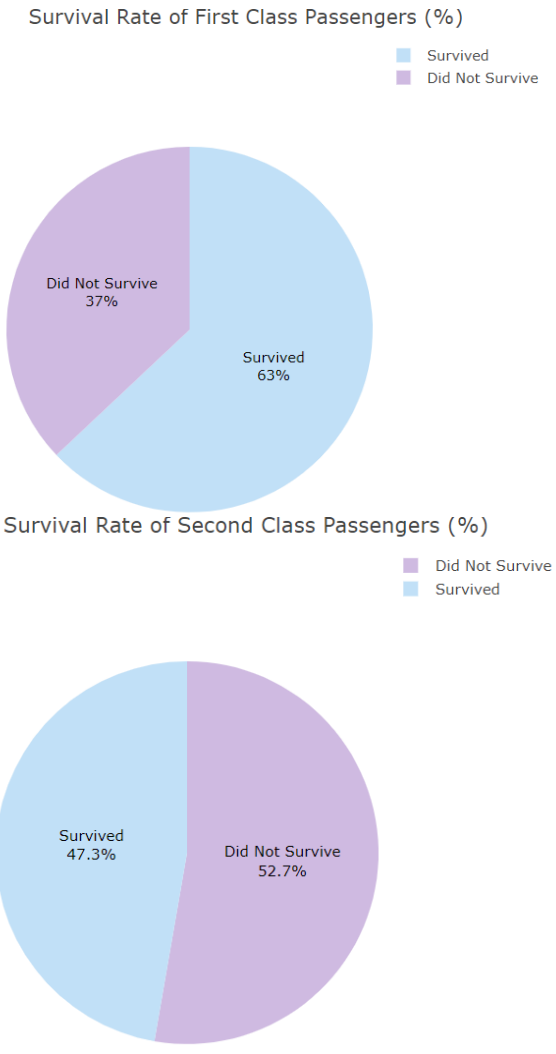


Figure 2



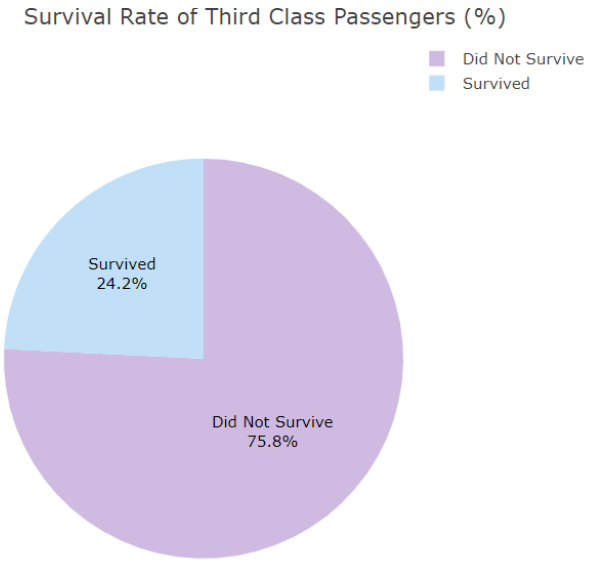


Figure 3

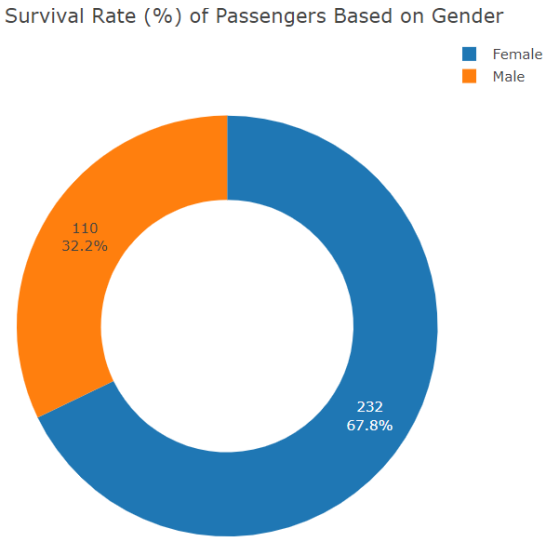


Figure 4

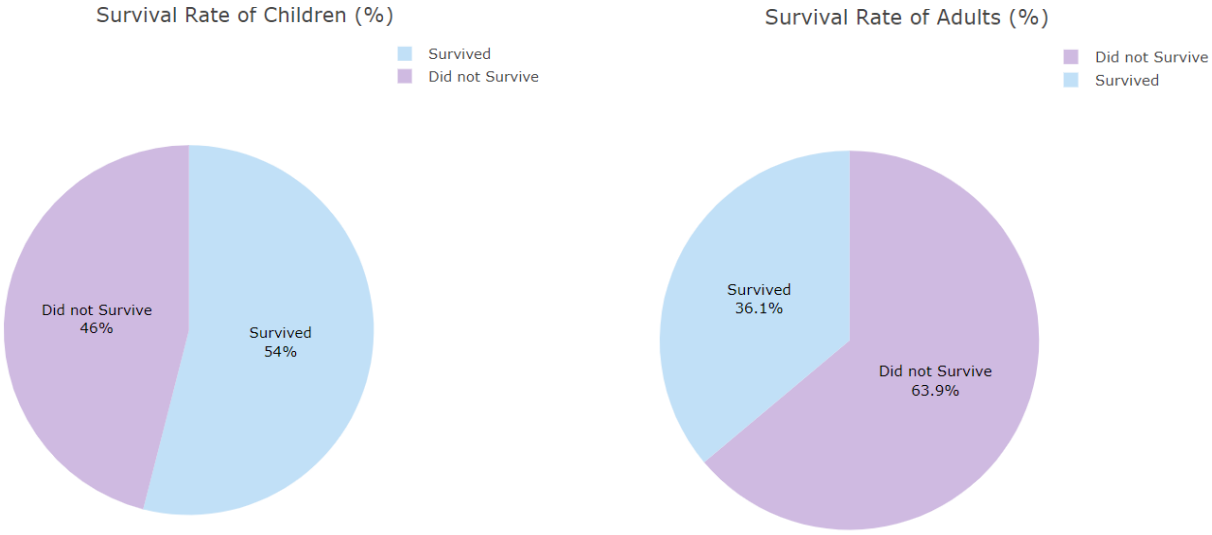


Figure 5

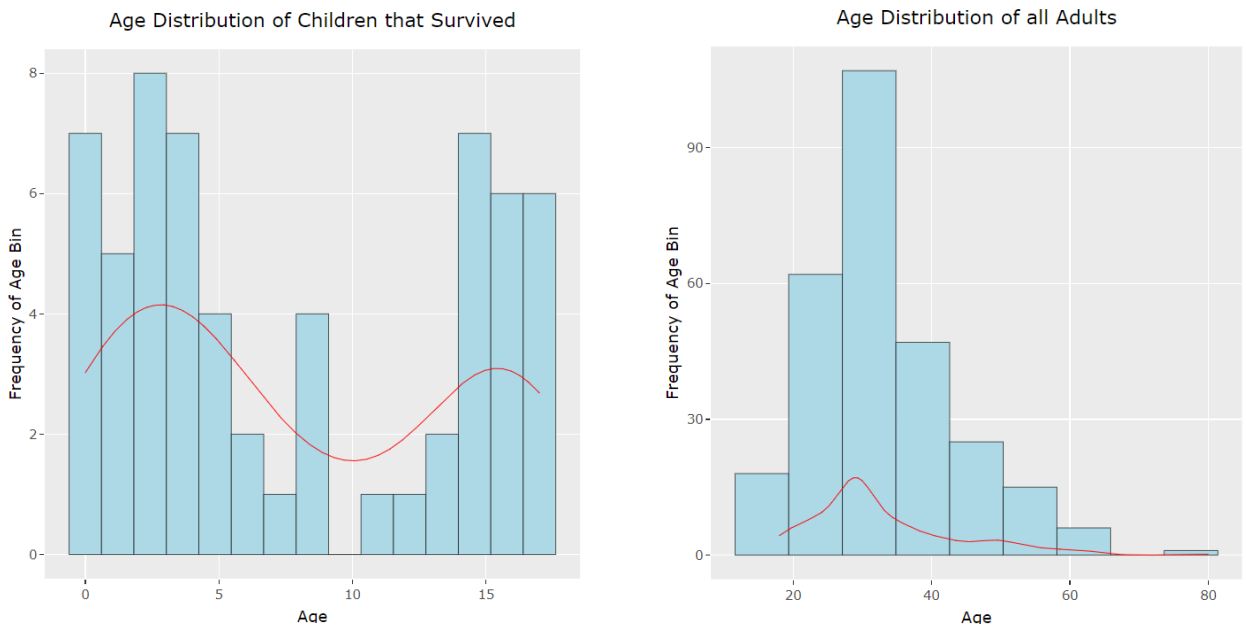


Figure 6

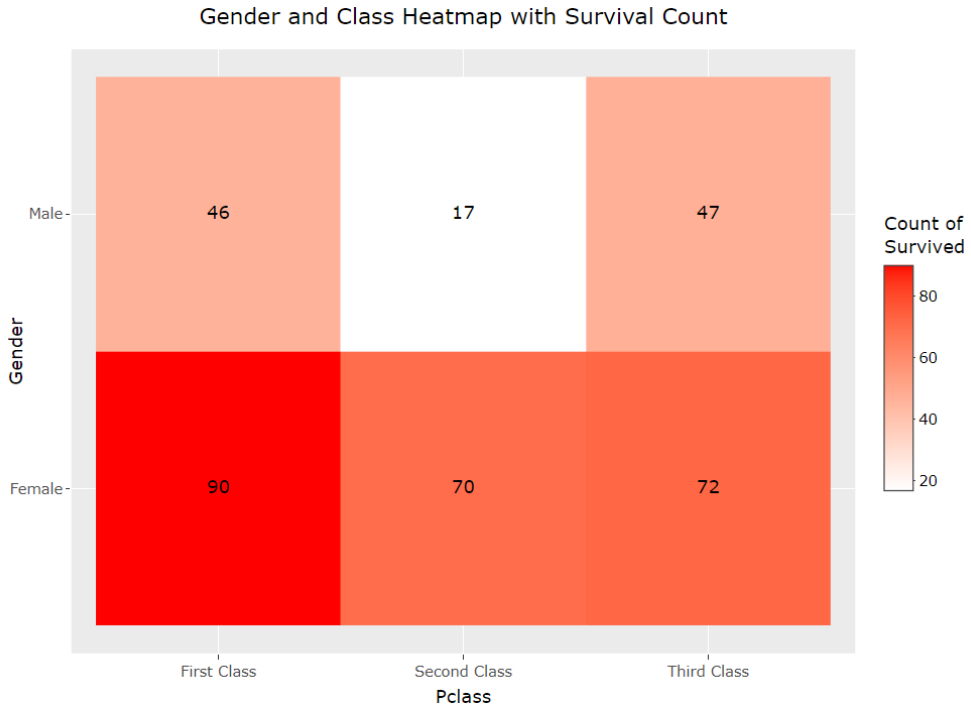


Figure 7

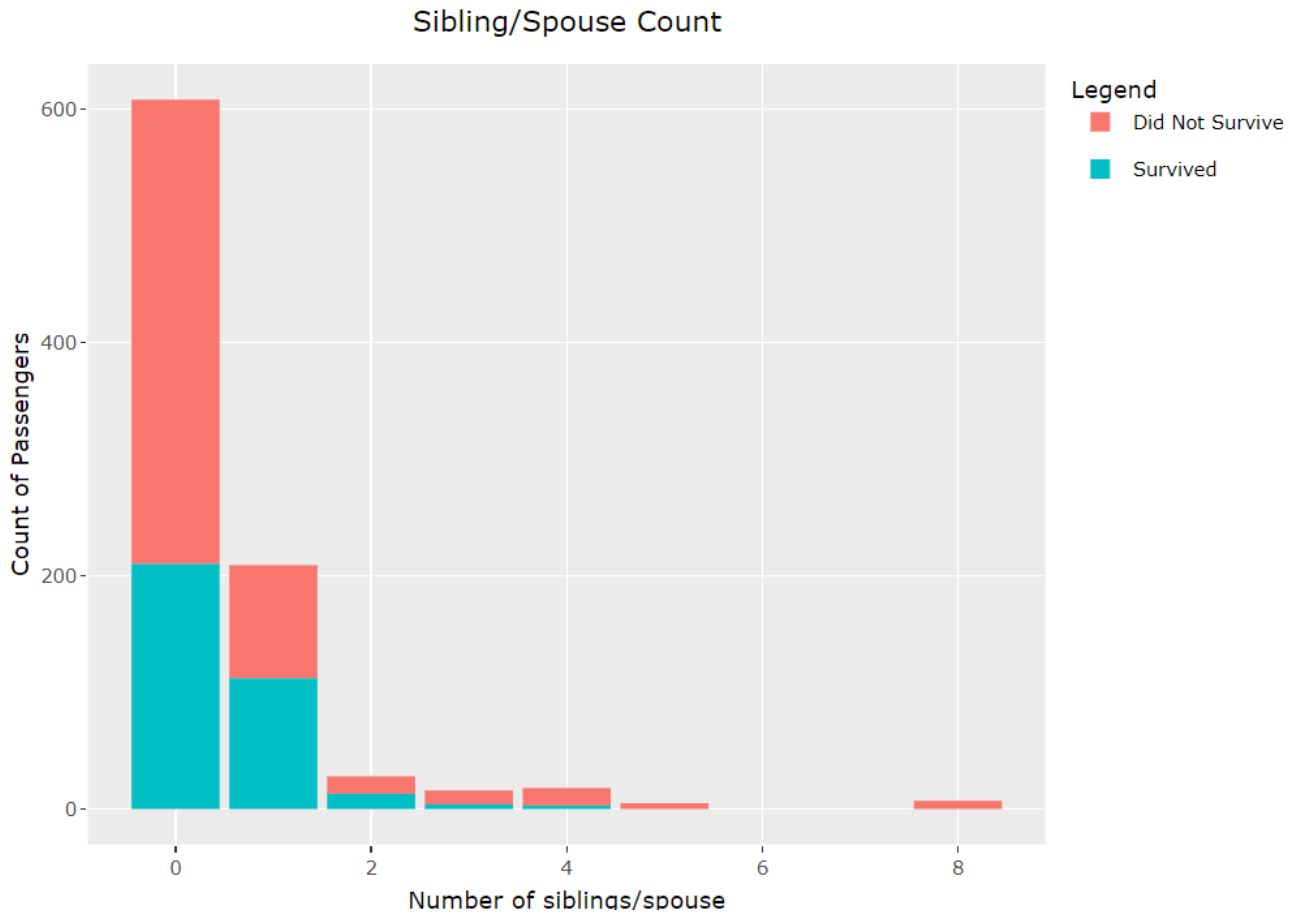


Figure 8

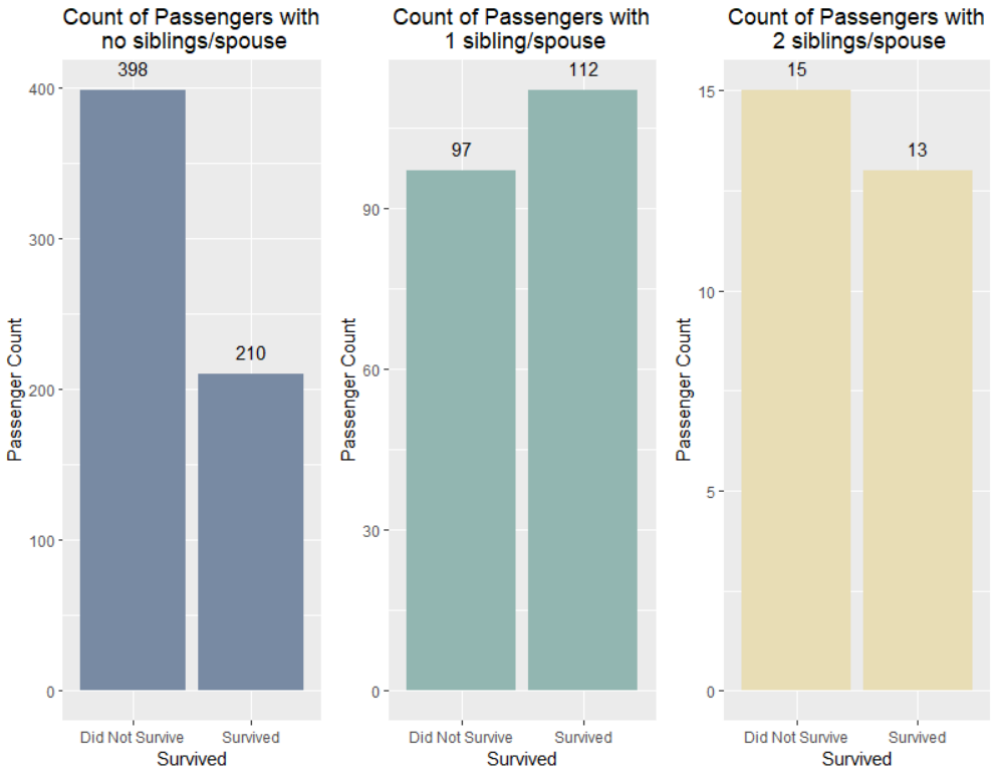


Figure 9

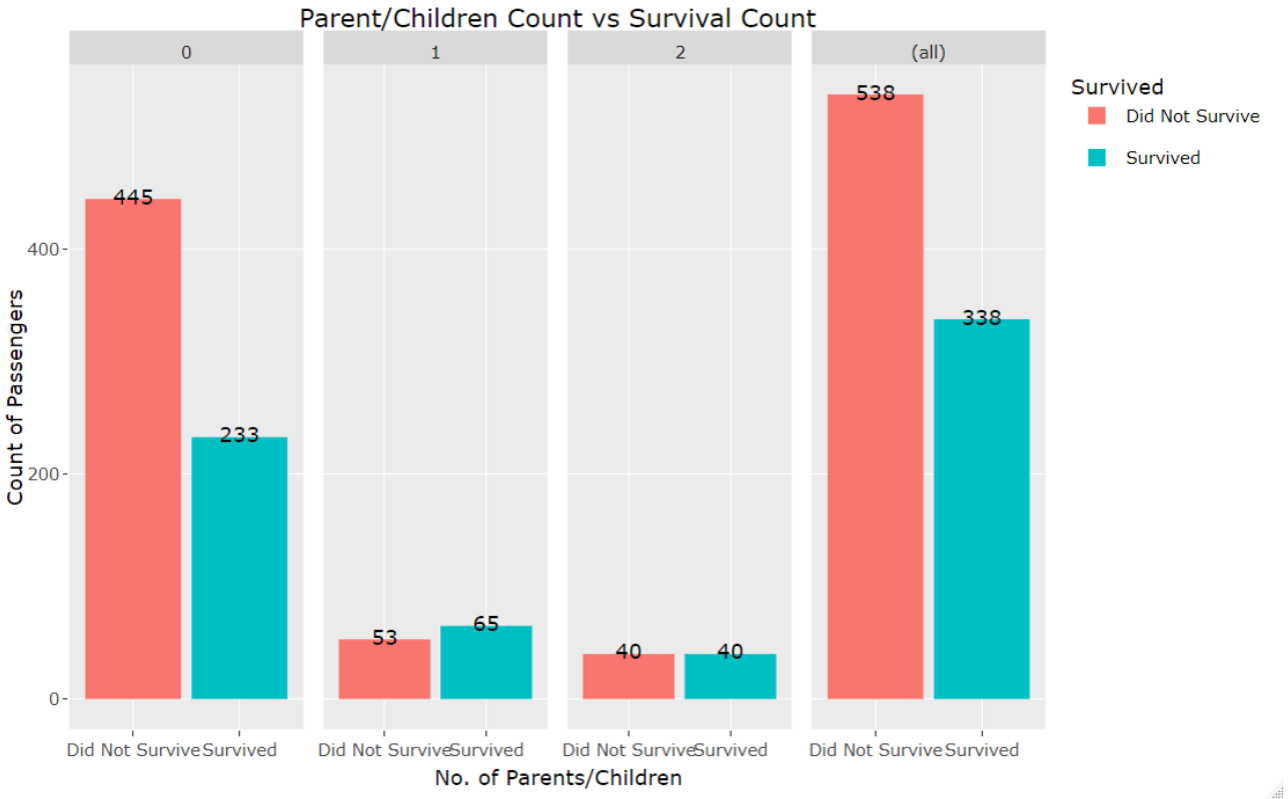


Figure 10

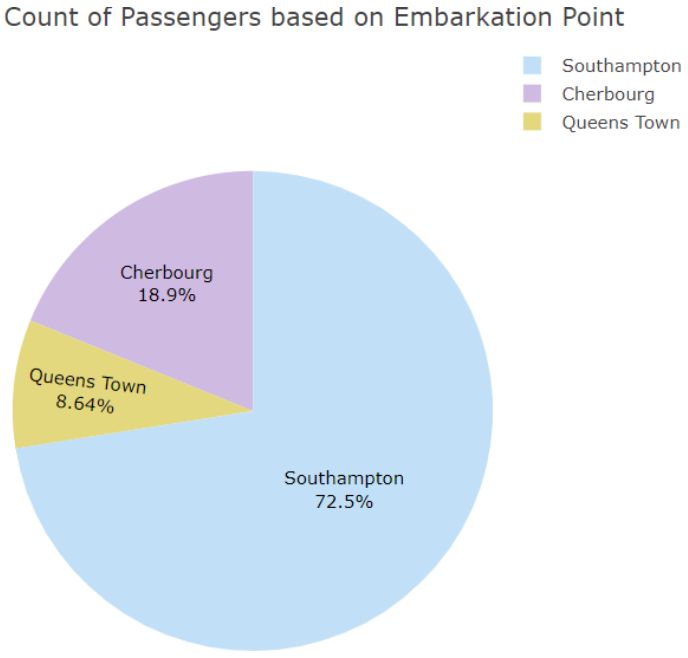


Figure 11

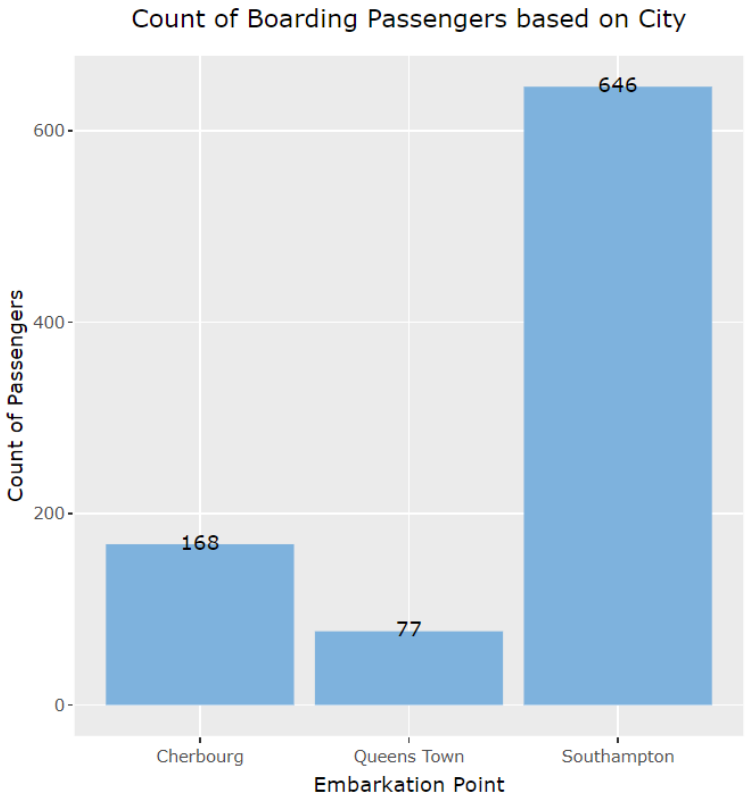


Figure 12

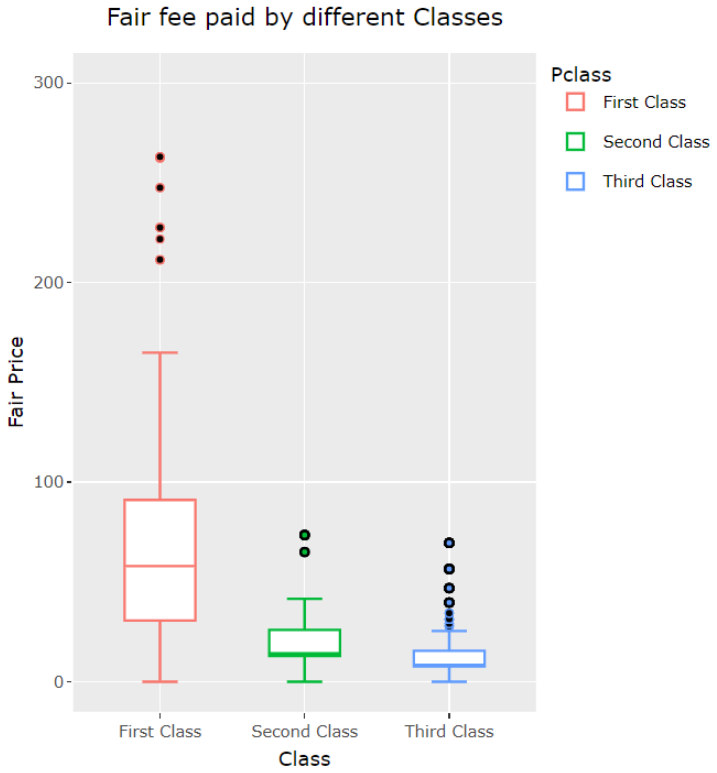


Figure 13

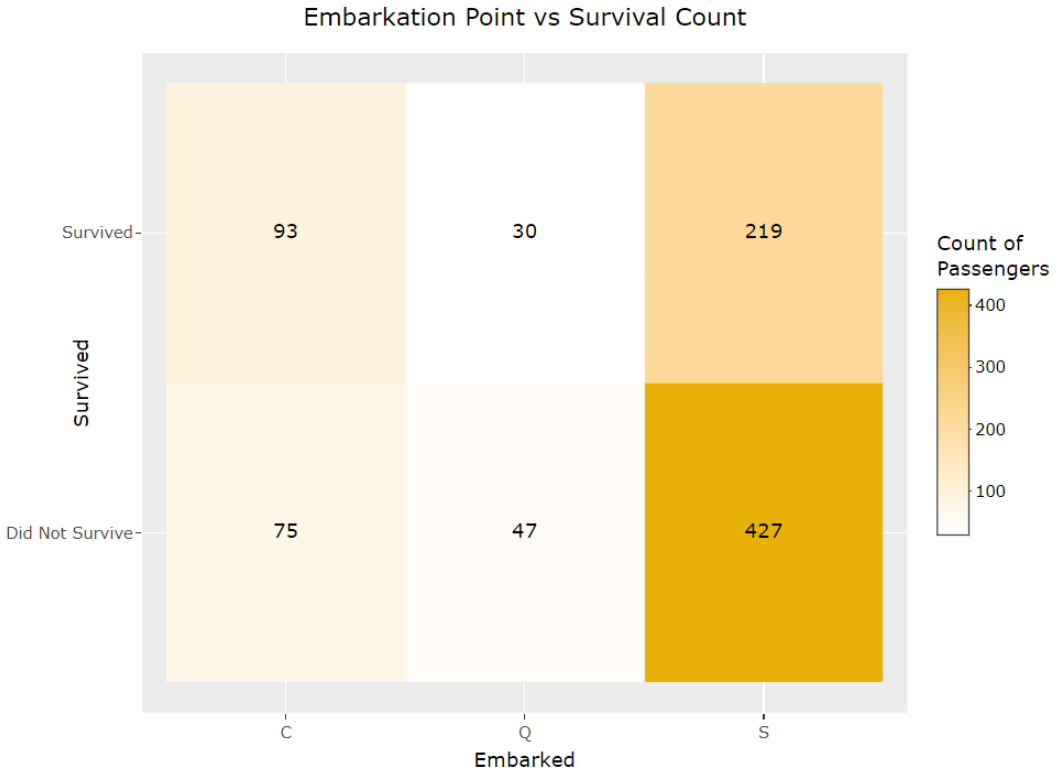


Figure 14