Data Analysis Nanodegree Project 2 : Investigate a Dataset

BY

SHAWAR NAWAZ

Abstract

The sinking of the RMS Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships.

One of the reasons that the shipwreck led to such loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others, such as women, children, and the upper-class.

In this challenge, I will analyze a sample of 892 people of the dataset to show a better understanding of what kind of people were more likely to survive the tragedy.

Introduction

The titanic dataset has tons of variables to work with and analyze survival rates. However, it is important to understand how the data is represented. After looking at the data and really getting a feel for it, I decided to post some questions that came to my mind while looking at the data. However, to actually begin analyzing the data, I had to make sure that the data was in a format that would permit analysis. For example, a dataset with lots of missing values would lead to a skewed form of the analysis. Thus, before beginning my analysis, it was imperative that I went to the process of data wrangling and cleaning that would assist me in answering the questions posted.

Once the dataset was ready for investigation, I go back to answering the questions I had thought of. Some questions came to me at the beginning of my analysis whereas other questions came up while hunting for answers to previous ones. The questions I have posted below will be highlighted as such for the sake of convenience.

Most of the solutions will shown in the form of numbers and graphs for ease of understanding. A summary of the methods and results will be posted in each section.

To investigate this dataset, it is important that the reader firstly understands the dataset we are dealing with. Thus, a summary of what all variables used in this dataset mean is given below:-

VARIABLE DESCRIPTIONS:

‘survival’ Survival

(0 = No; 1 = Yes)

‘Pclass’ Passenger Class

(1 = 1st; 2 = 2nd; 3 = 3rd)

‘name’ Name

‘Sex’ Sex

‘age’ Age

‘sibsp’ Number of Siblings/Spouses Aboard

‘parch’ Number of Parents/Children Aboard

‘ticket’ Ticket Number

‘fare’ Passenger Fare

‘cabin’ Cabin

‘embarked’ Port of Embarkation

(C = Cherbourg; Q = Queenstown; S = Southampton)

SPECIAL NOTES:

Pclass is a proxy for socio-economic status (SES)

1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower

Age is in Years; Fractional if Age less than One (1)

If the Age is Estimated, it is in the form xx.5

With respect to the family relation variables (i.e. sibsp and parch)

some relations were ignored. The following are the definitions used

for sibsp and parch.

Sibling: Brother, Sister, Stepbrother, or Stepsister of Passenger Aboard Titanic

Spouse: Husband or Wife of Passenger Aboard Titanic (Mistresses and Fiances Ignore d)

Parent: Mother or Father of Passenger Aboard Titanic

Child: Son, Daughter, Stepson, or Stepdaughter of Passenger Aboard Titanic

Other family relatives excluded from this study include cousins,

nephews/nieces, aunts/uncles, and in-laws. Some children travelled

only with a nanny, therefore parch=0 for them. As well, some

travelled with very close friends or neighbors in a village, however,

the definitions do not support such relations.

Methodology

Firstly, I had a look at a section of the dataset, which looked a little something like this.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S

The first thing I noticed is a ‘Nan’ value in the Age column. Age is a parameter I was looking to work with extensively so it was important to fix this before I moved ahead. Firstly, it was important to find out how many of the ‘Nan’ values were present in the column.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
```

The Age section of this info shows ‘714 non-null’ which means that there are about 177 ‘Nan’ values in the column. The ‘Embarked’ section also shows 2 missing ‘Nan’ values which would be an easy fix compared to that of Age.

In my first step to fix the age issue, I decided to separate values for men and women and calculate their average ages and std deviations separately. The results were as follows:-

Men - average = 30.73

Std deviation = 14.68

Number of missing ages = 124

Women – average = 27.92

Std deviation = 14.11

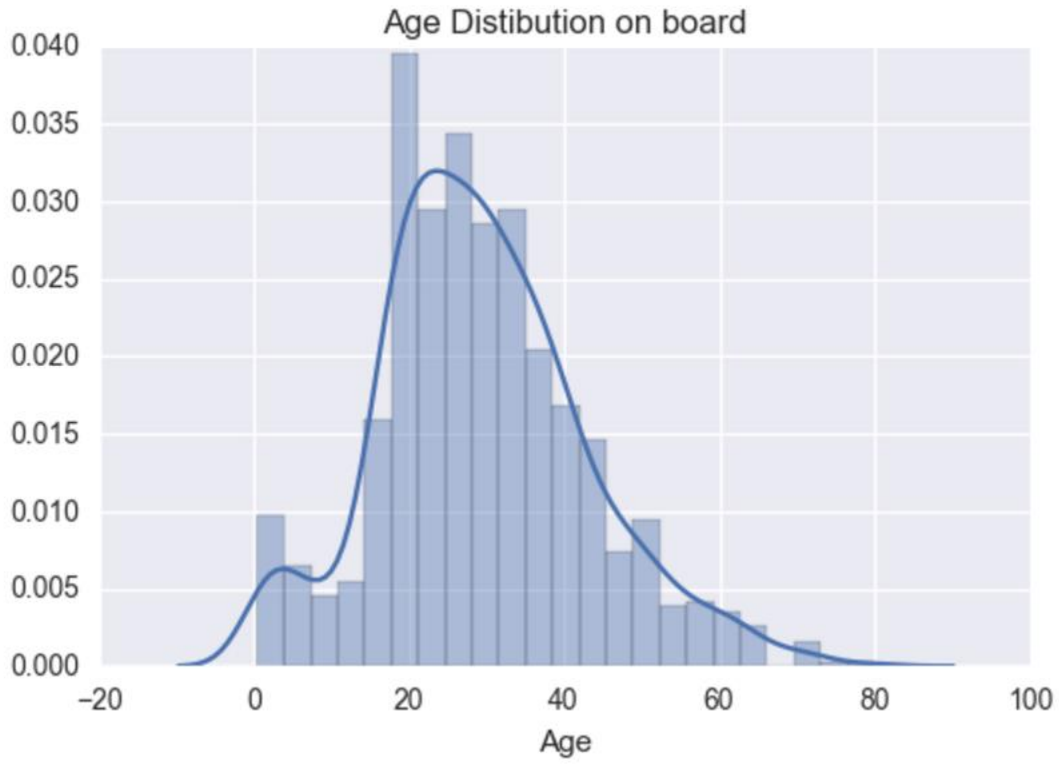
Number of missing ages = 53

The methodology I used to fill these ages was to create random array of ages for each gender. Since I didn't want to bother the average and std deviation of the final ages too much, I decided that the random array should consist of values within a range of 1 std deviation from the mean of each gender. (code in supporting files)

Now that the ages have been filled, I decided to pose my first question.

Q1) What is the mean and median of the ages of this dataset and whether it falls on a normal distribution?

To answer this question, I had to take a look at the age distribution of people on board the ship. Graphically, this was represented as :-



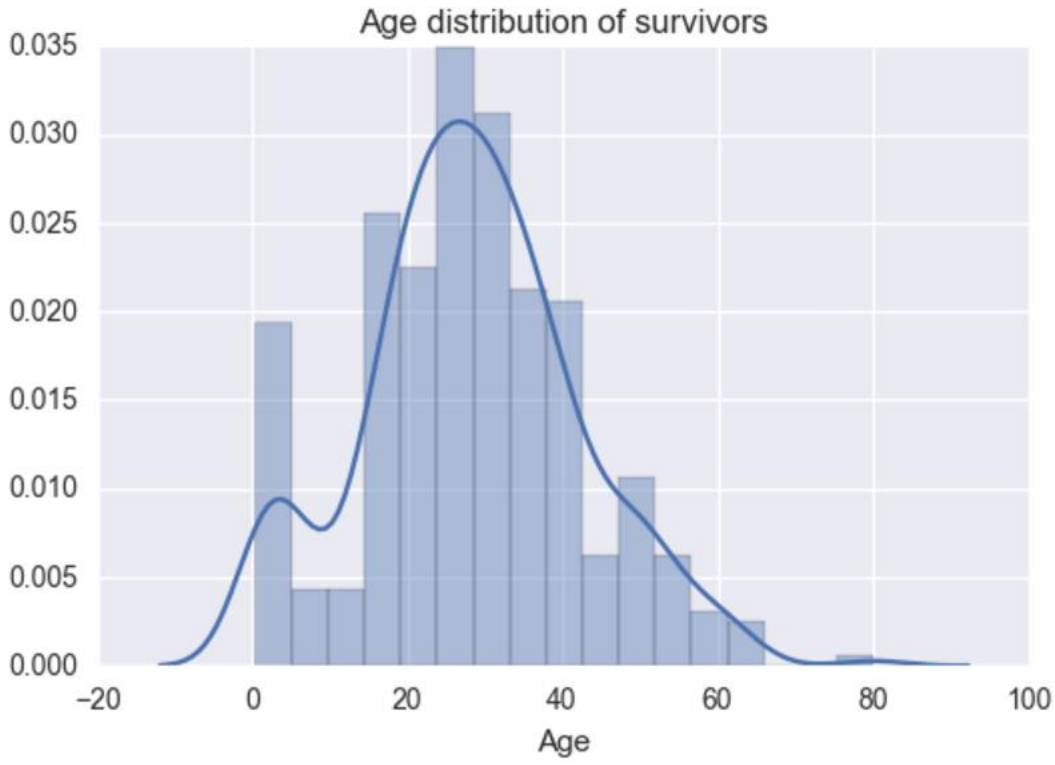
As we can see, the dataset falls within a normal distribution with a chunk of ages roughly within the 20-35 year mark. We can also confirm that it is a normal distribution by looking at the mean, median and mode of the dataset.

Mean = 29.38
Std deviation = 13.490178
Median = 28
Mode = 28.0

Since the mean, median and mode are roughly the same, it shows that the data values fall under a normal distribution.

Q2) What were the age groups that were most likely to survive?

I also wanted to check the distribution of ages of survivors. This is depicted in the graph below.

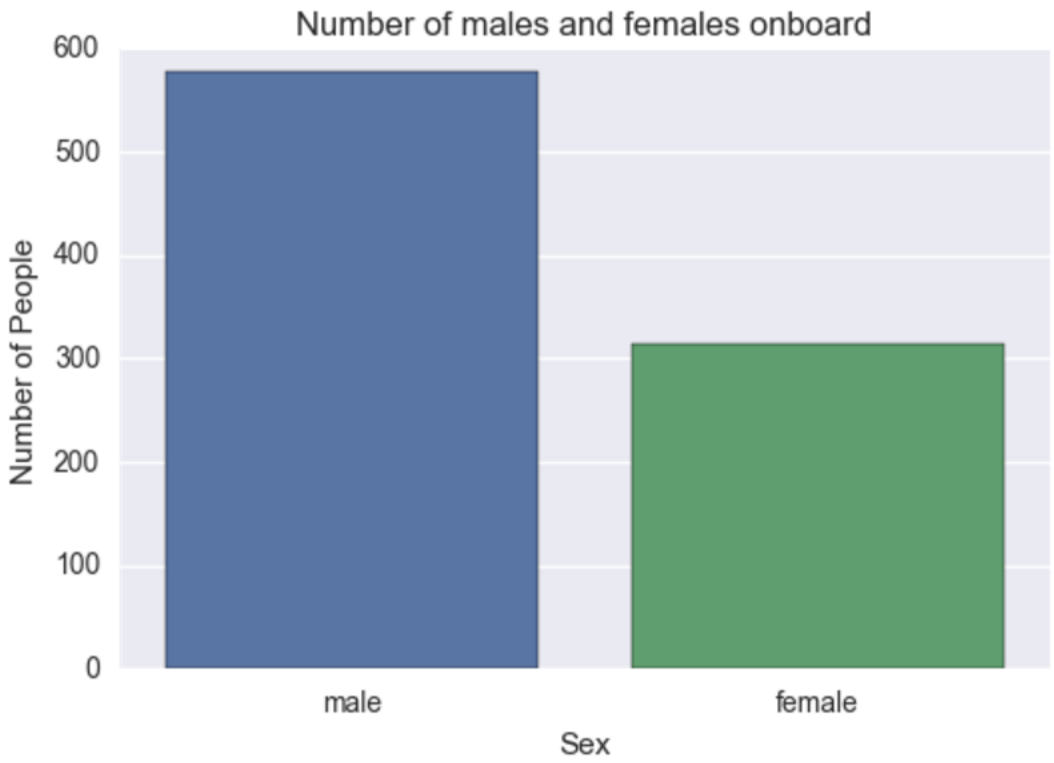


From here, it is quite obvious to see that most of the survivors were either infants or of an age group between 17 – 35.

A more detailed description of the survival proportions based on each age is given in the supporting document (graph title - 'Proportion of Survivors for each Age')

I wanted to check how many males and females were present on board during the cruise.

It turned out that the number of males were 577 (65%) and the number of females were 314 (35%).



This was good to know, but I realized that usually in catastrophic situations, it is general practice that the women and children are evacuated first. I was curious to know if that was the case here too, but realized there was no classification for children in the dataset. So I decided to create my own.

The code I used was as follows:-

```
titanic['Type'] = 0
for index, value in titanic.iterrows():
    if value['Age'] <= 10:
        titanic['Type'].loc[index] = 'Child'
    if value['Age'] > 10:
        if value['Sex'] == 'female':
            titanic['Type'].loc[index] = 'Woman'
        else:
            titanic['Type'].loc[index] = 'Man'
```

The new column that incorporated these values was named ‘Type’. Passengers under the age of 10 were classified as children. This led me to question no. 3.

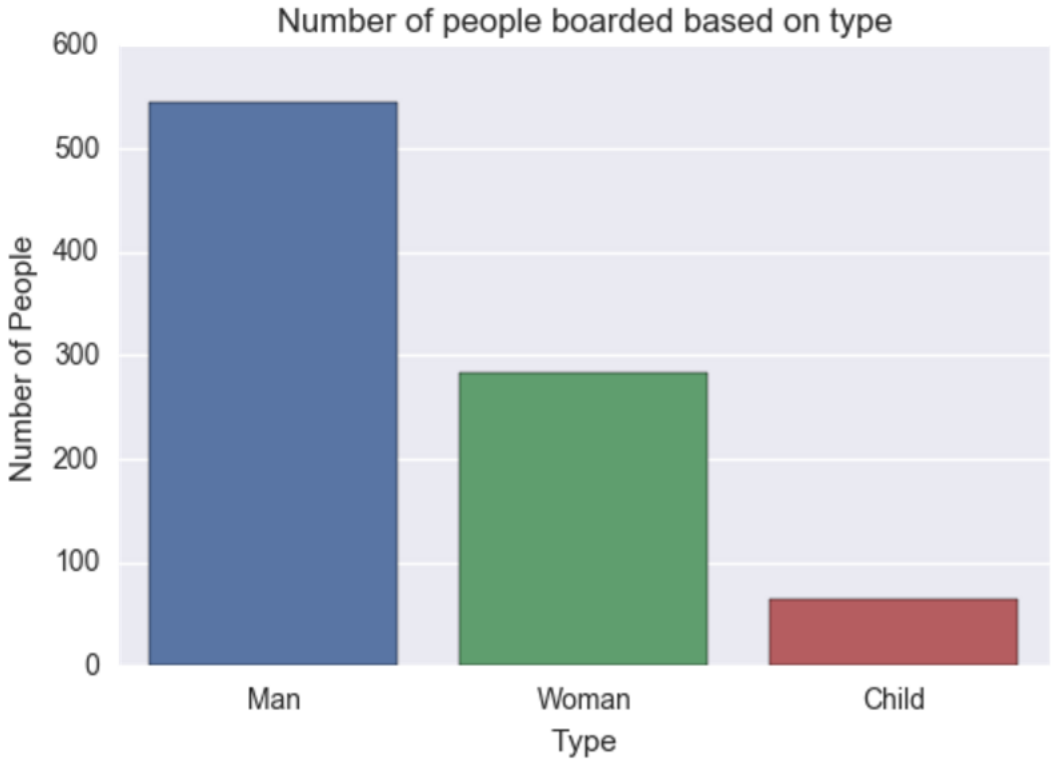
Q3) Who had a better chance of survival among men, women and children?

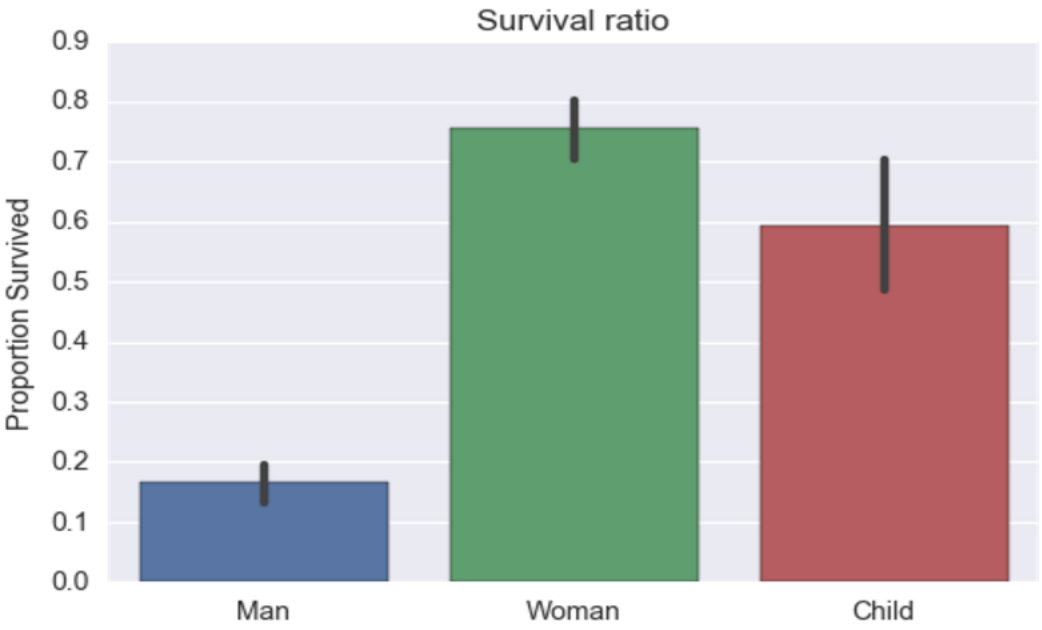
I decided to dig into this data to see what proportion of each had survived.

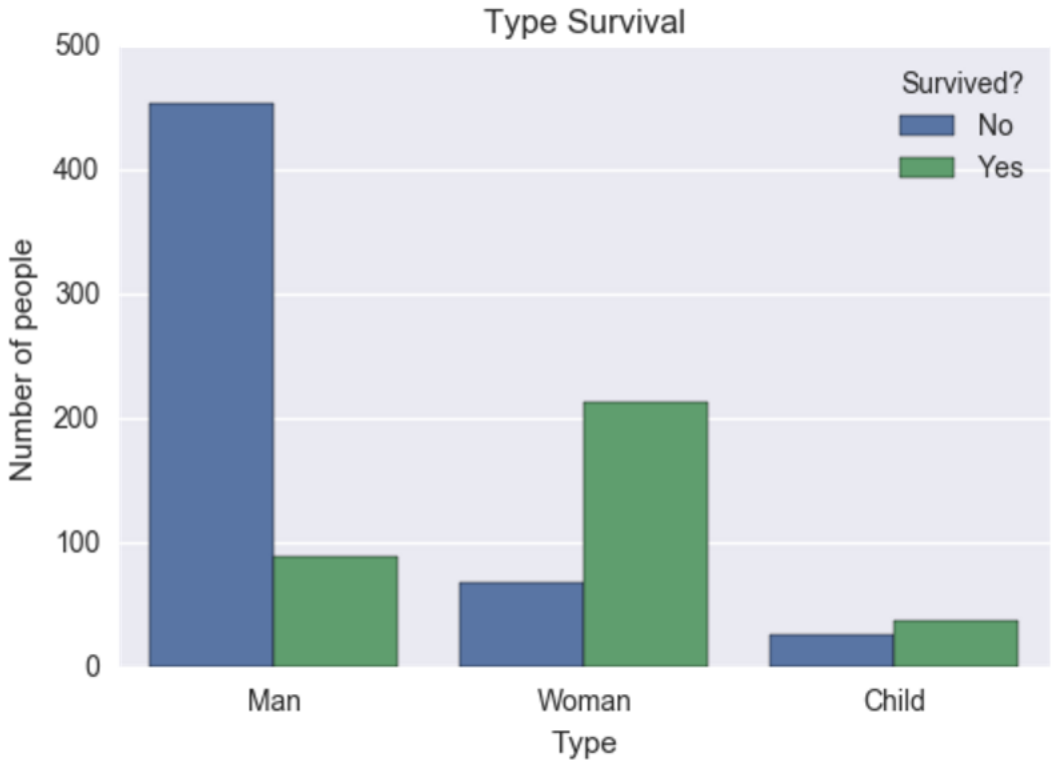
The number of men on board = 544 (61% of population)

The number of women on board = 283 (32% of population)

The number of children on board = 64 (7 % of population)

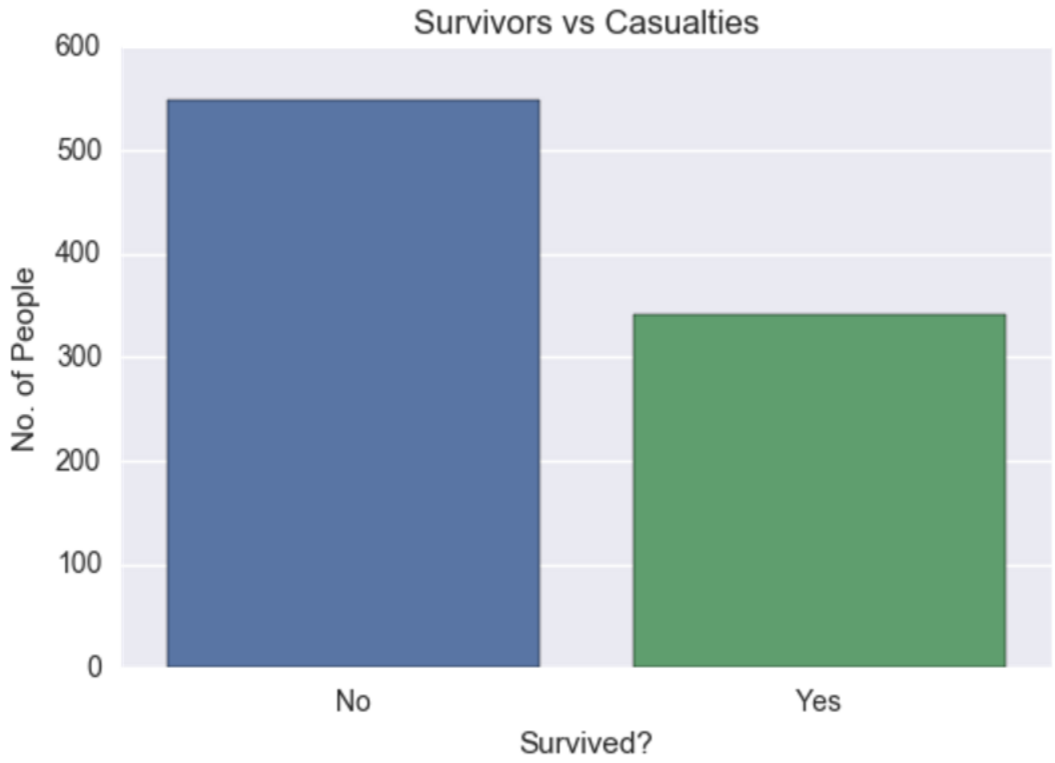






From the graphs above, it is fairly obvious that women and children were given more preference to the men in terms of rescue operations. 61% of the population that boarded the titanic were men, yet not even 20% of them survived. Women, on the other hand constituted 32% of the population, 75% of them being survivors. Children constituted 7% of the population

and nearly 60% of them were survivors. I compared these stats to that of the overall survival rates. Of the population on the titanic, 549 did not survive in total (61%) which means overall survival was only 39%.



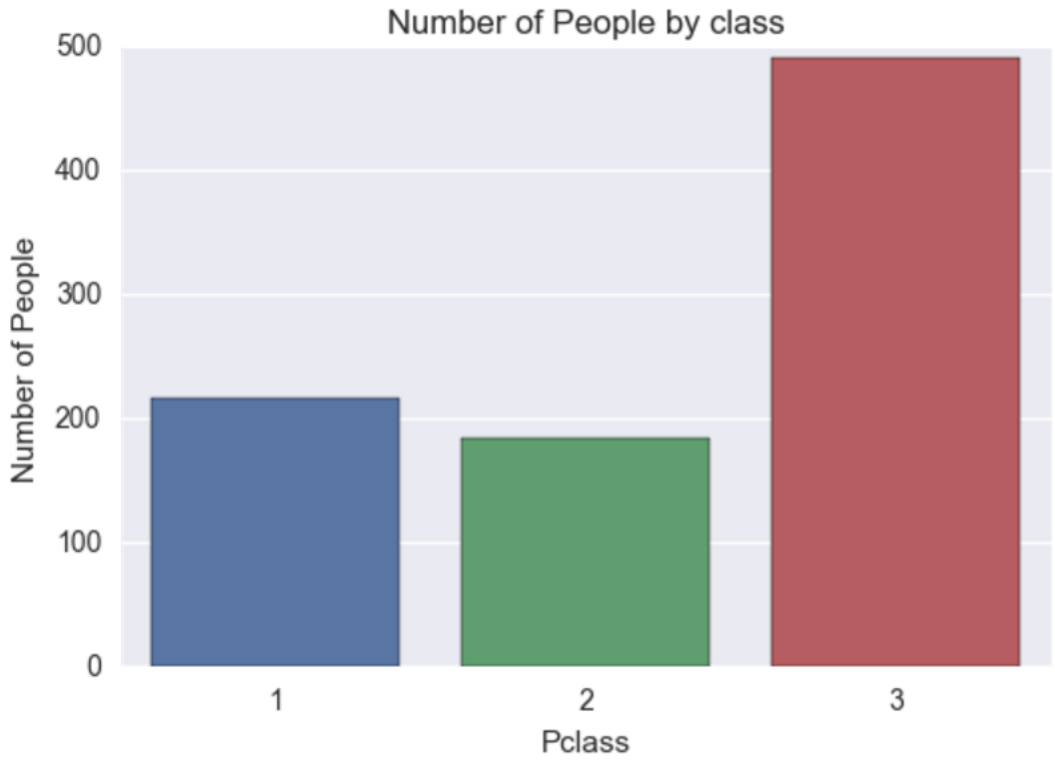
Women and children both have much higher rates of survival.

This could probably be due to standard procedures during times of emergencies, where the women and children are generally evacuated first. **The graphs certainly depict so, but this is not enough to prove causation.**

The next variable I wanted to look at were the different socio economic classes. This brought me to question 4.

4) Are survival trends different for different socio-economic classes?

Firstly, lets look at the number of people that had boarded from each class.



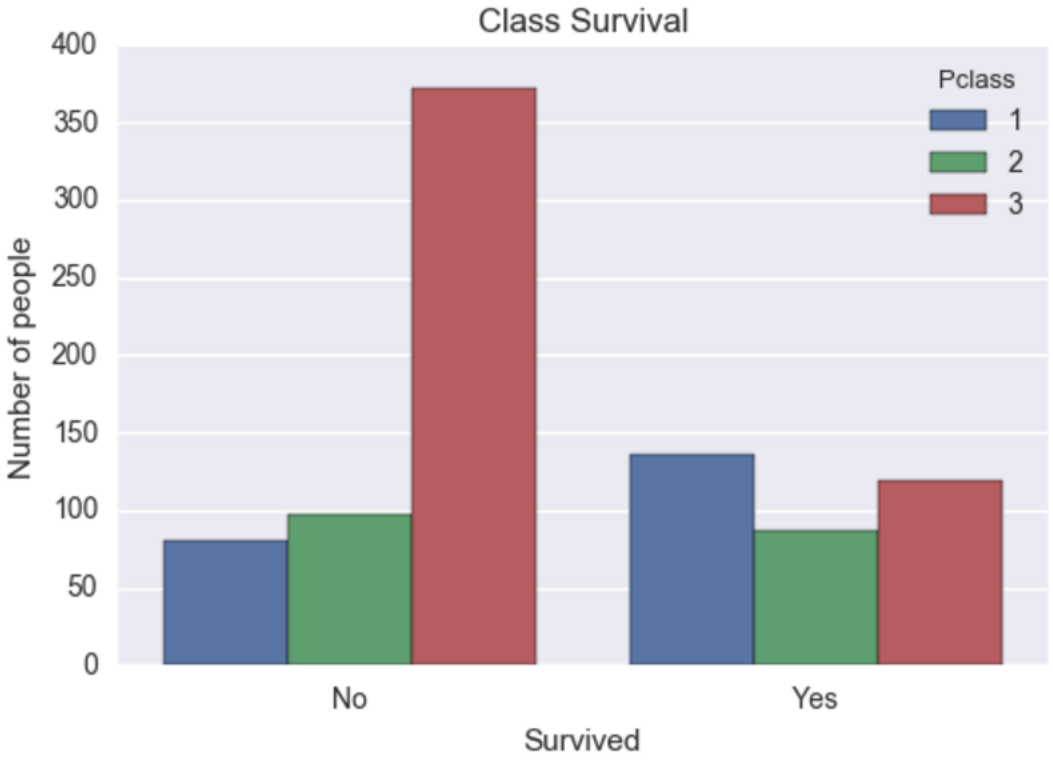
No. of people - Class 1 = 216 (24% of the population)

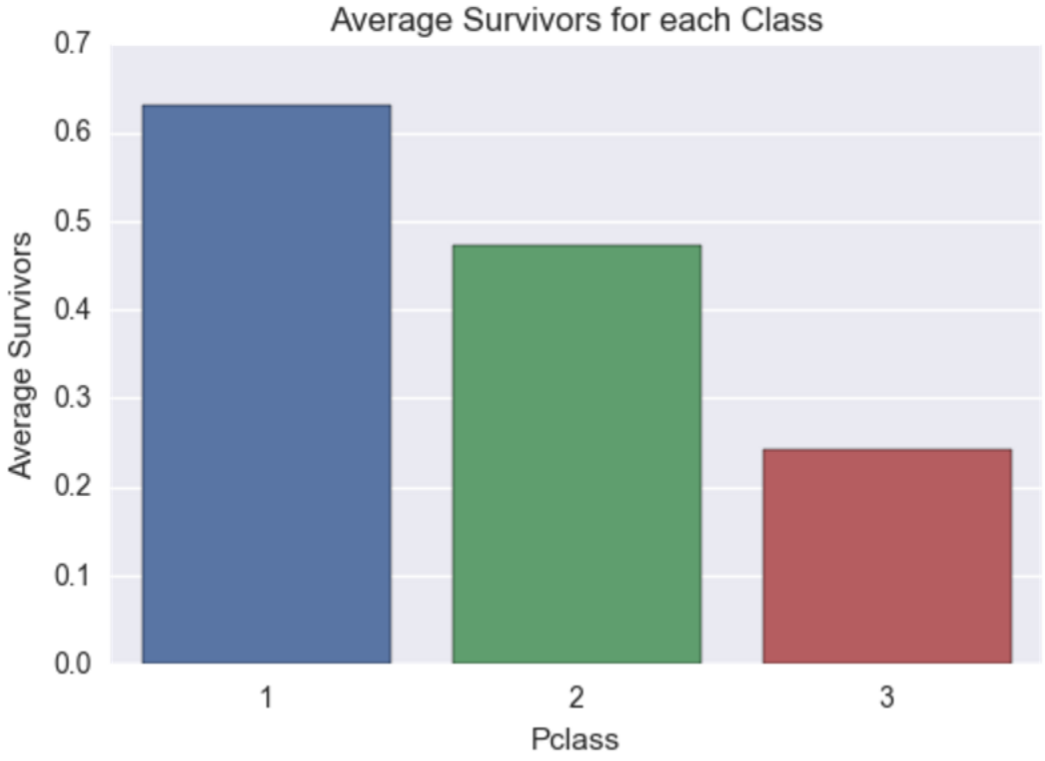
Class 2 = 184 (21%)

Class 3 = 491 (55%)

An overwhelming majority of people were of class 3.

Now let's look at the proportion of people that had survived for each class.





These graphs show that people of a higher class had higher survival proportions than lower classes. In fact, the first two graphs show that even though the number of people of class 3 that

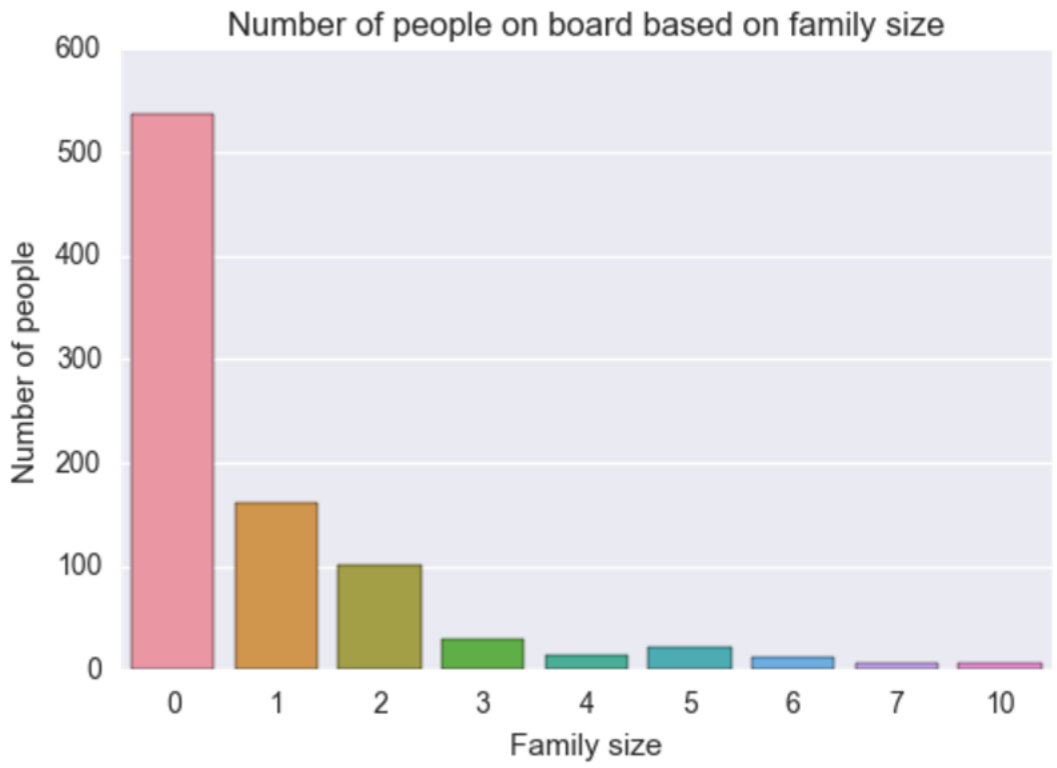
boarded were more than double that of class 1, people of class 1 had more number of survivors than that of class 3.

This may be a result of preference shown to people of higher classes or that cabins belonging to people of higher classes were in locations that allowed easier access to lifeboats and escape routes. **More tests would be required to see if there was indeed a causation factor or if the data are merely correlated.**

Another variable I was looking to investigate were the sibling and parents sections. It didn't seem convenient working with them separately so I decided to add up the two columns and create a third column called 'Family'. This sum would show how many relatives a person had on board and could be useful to our analysis for question 5.

5) Do families tend to survive the bigger they are or were people who travelled alone more likely to survive?

Firstly, to tackle this question, I made a countplot to show the number of people who had different sizes of families on board.



In numbers, Number of people in each category :-

0 – 537

1 – 161

2 – 102

3 – 29

4 – 14

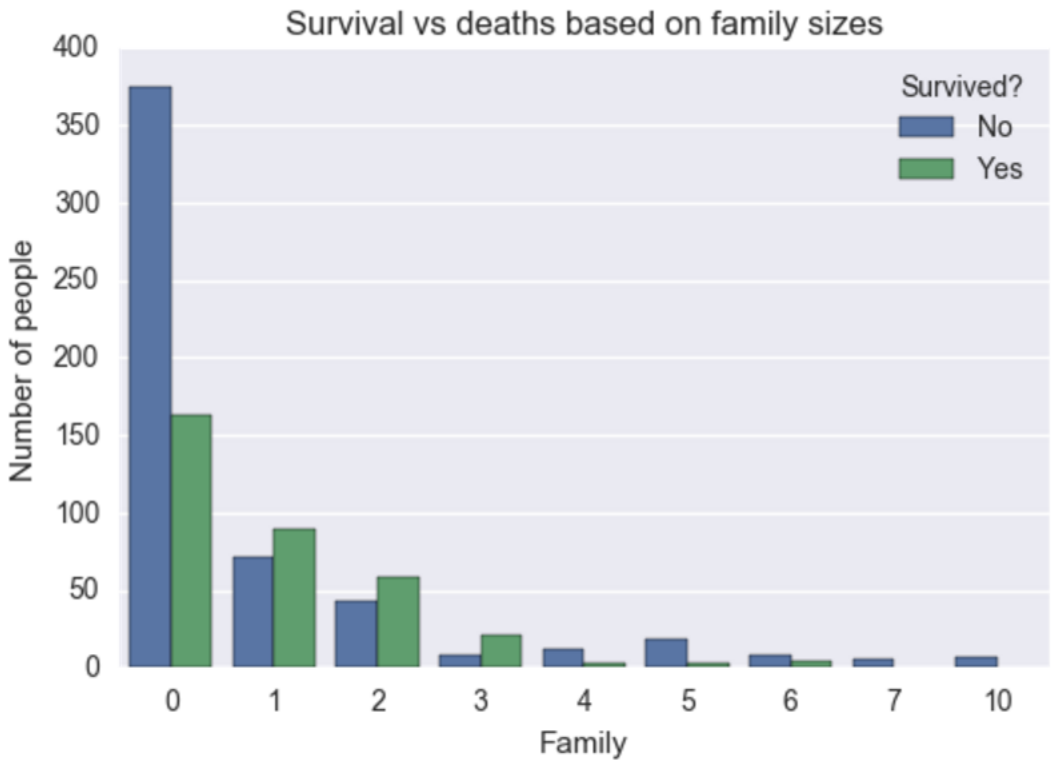
5 – 22

6 – 12

7 – 6

10 – 7

And the number of people that survived based on each category:-



This graph shows that most number of deaths came in situations where a person was travelling alone. However, it is not a proper description of the rest of the people as there are too many differing family sizes. This creates a situation where some categories have much lesser data

points than others. So, in order to fix these issues, I decided to create a new column called 'Relation' which categorizes people in three ways:-

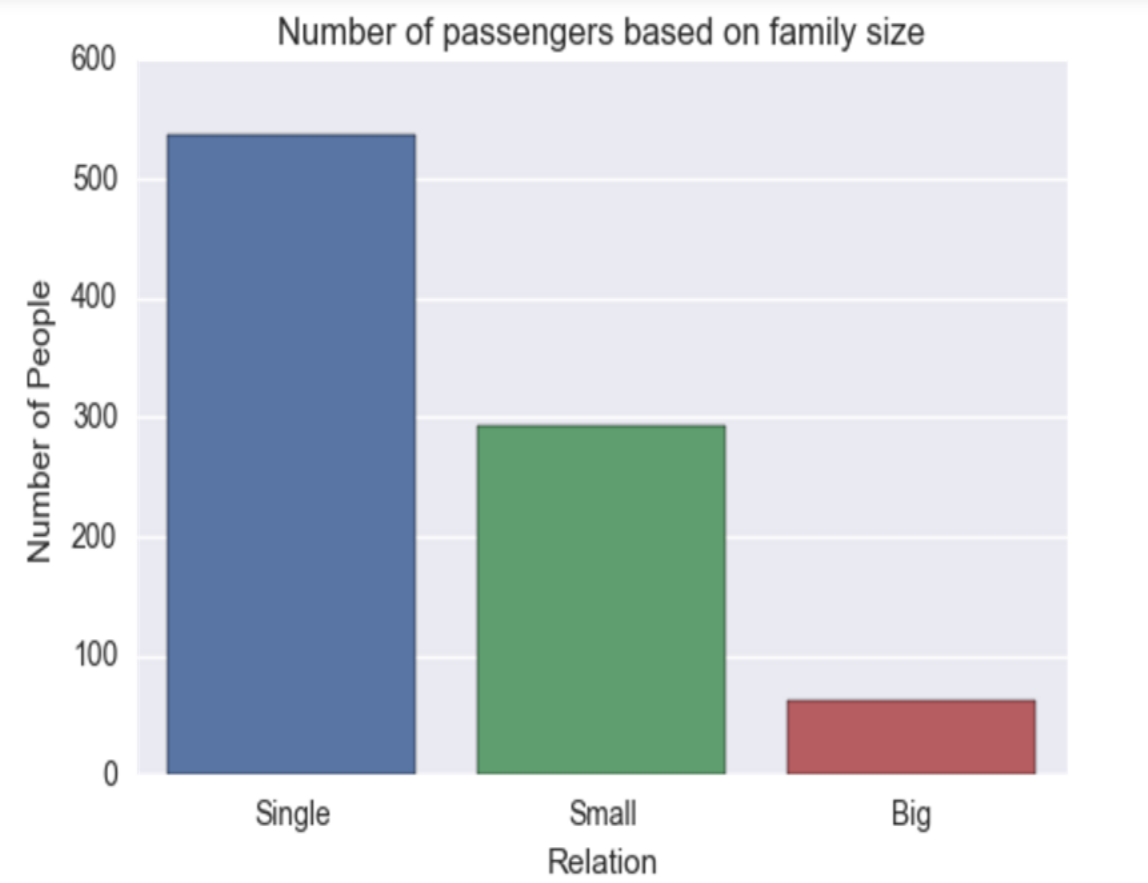
'Single' – if family members = 0

'Small' – if family members > 0 and <= 3

'Big' – if family members > 3

The code to create this column is as follows:-

```
titanic['Relation'] = 0
for index, value in titanic.iterrows():
    if value['Family'] > 3:
        titanic['Relation'].loc[index] = 'Big'
    if value['Family'] > 0 and value['Family'] <= 3:
        titanic['Relation'].loc[index] = 'Small'
    if value['Family'] == 0:
        titanic['Relation'].loc[index] = 'Single'
```



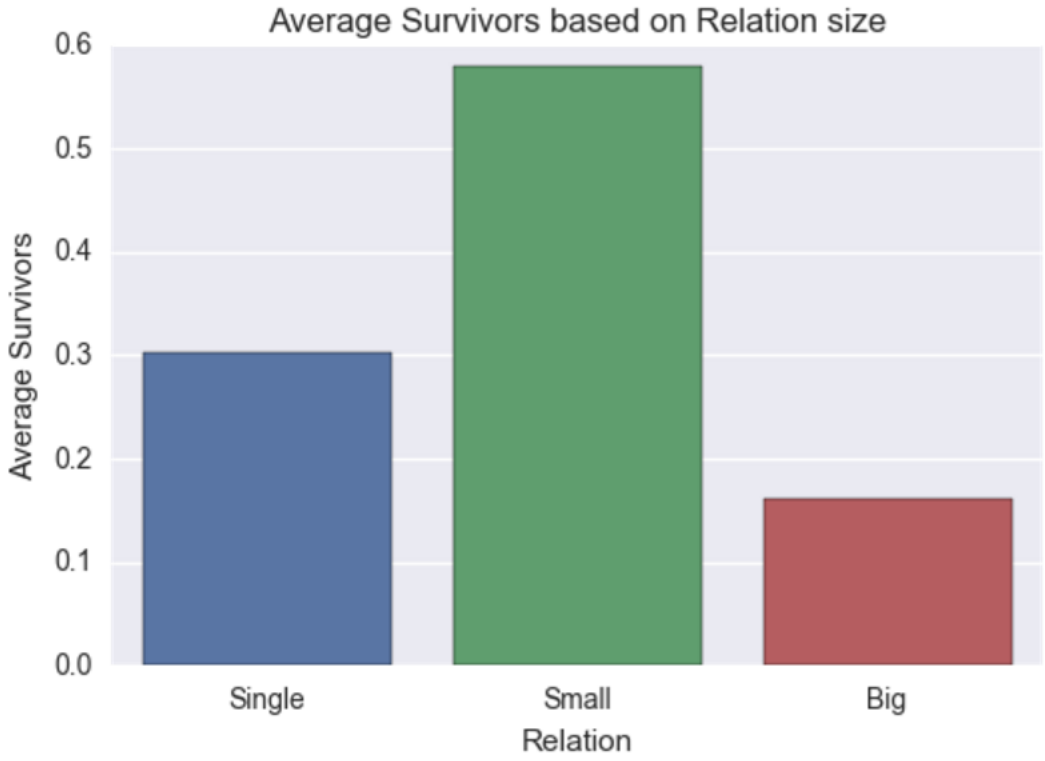
The number of people boarded based on Relation –

Single – 537 (60%)

Small – 292 (33%)

Big – 62 (7%)

And their respective survival rates –



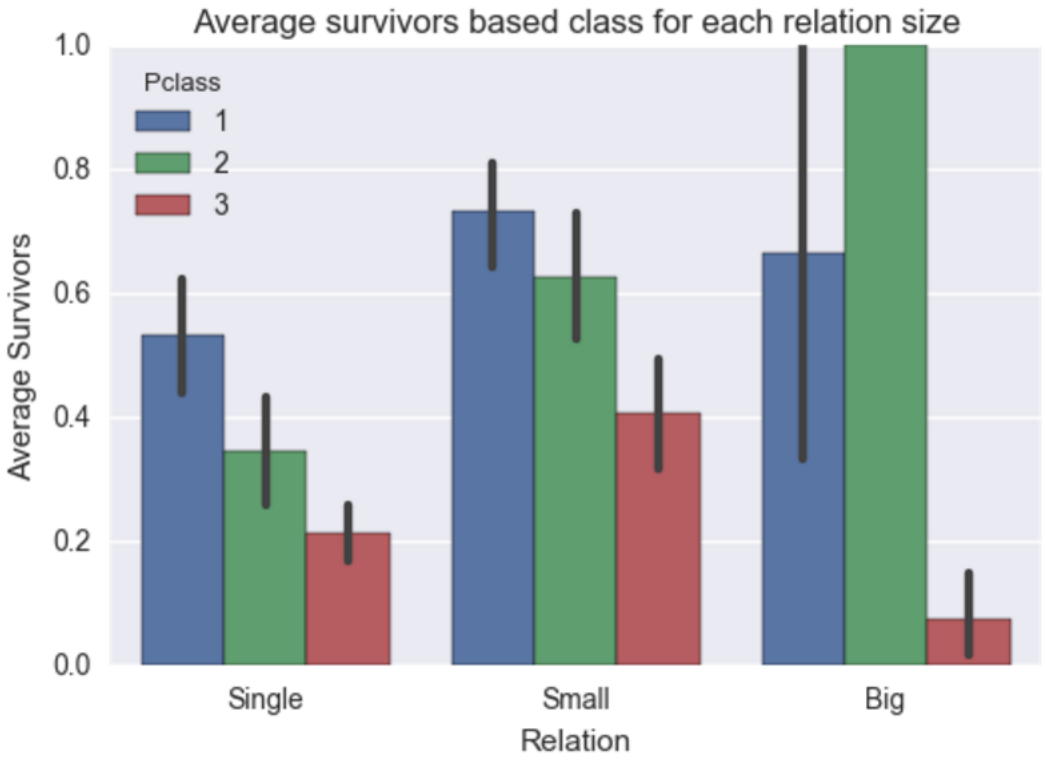
From these graphs, it's quite obvious to see that smaller families tended to survive more than big families and single people. This may be because it was probably easier to look out for and help each other in smaller families than in large families. Single people probably had nobody looking out for them. **More statistical tests would be required to prove these statements rather than just make inferences based on correlation.**

Now that we've looked at Age, Sex, Pclass and family sizes separately, we've begun to have a good feel about our data. So I thought it was time to go a little deeper into the data set, by analyzing certain variables wrt other variables. For example, we saw that women and children had a higher survival rate than men, but is that true for each class. We've answered each question earlier separately, but this would give us answers on a combination of different questions.

Lets look at class and family sizes together. The pivot table function helps us to look at this data conveniently in a tabular format.

	Survived		
Relation	**Big**	**Single**	**Small**
Pclass			
1	0.666667	0.532110	0.732673
2	1.000000	0.346154	0.628205
3	0.074074	0.212963	0.407080

Graphically, this would be represented as follows:-

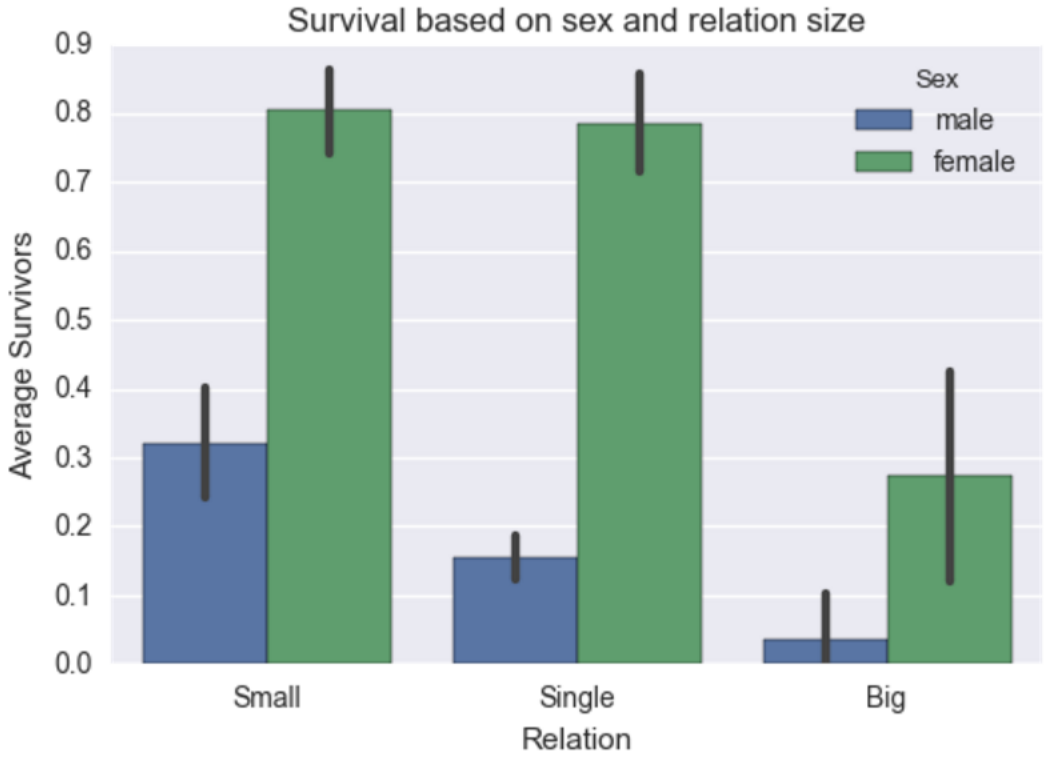


The table and the graph shows us two trends that we've seen before. One is that, in general, people of higher classes seemed to have survived in greater proportions than lower classes

(with the exception of class 2 in big families). Also, smaller families tended to survive better than bigger families and single people. These stats shed further light on question 4 and 5. **Of course these are mere correlations and not causation.**

I was curious to see why the class 2 in big families seemed to have a better survival proportion than expected. How did 100% of these people survive. A deeper look showed me that there were only 2 people in this category and they managed to survive.

Let us continue to stick with relation size for a while and analyze survivals for each sex.

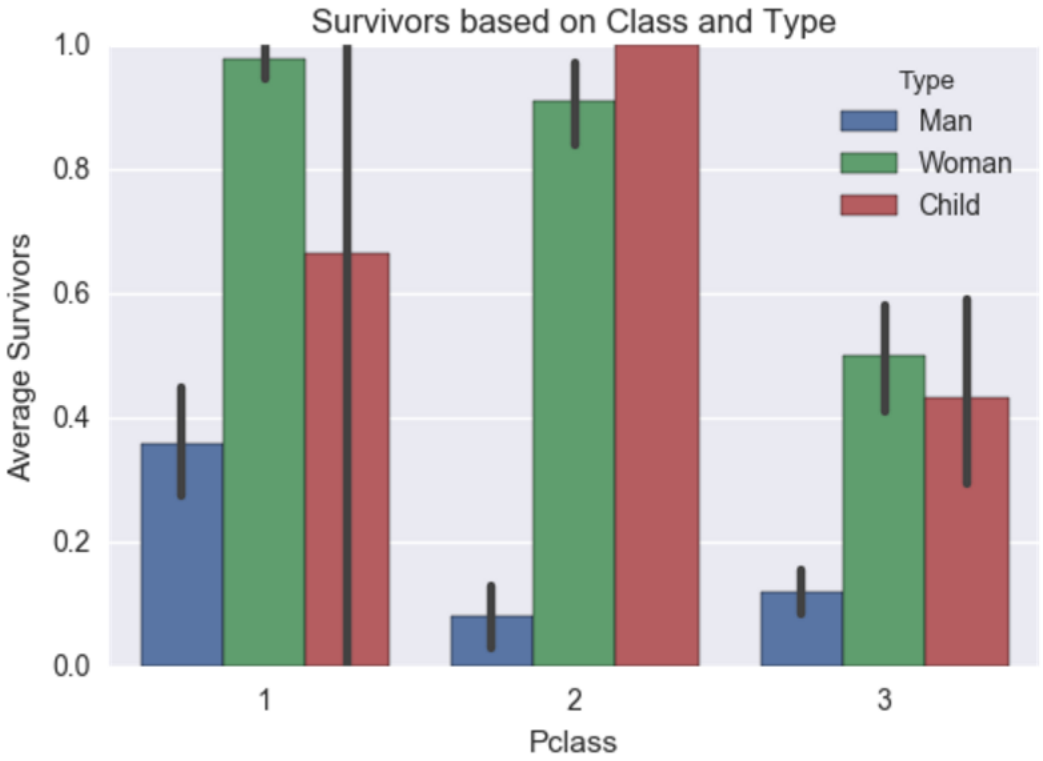


In each of these cases, we see that women have higher survival proportions than men. We see yet again that smaller families tended to survive better than people that were single or in big families. (Question 3 & 5)

***Note: Causation is not being implied, but rather correlations were being highlighted**

*Note: I ignored the cases for children here because children would end up coming under the category of single, which means they had no family members(A case where a child was travelling with a nanny). There was only 1 person in this situation and it didn't seem to be a fair assessment.

Moving on to classes. I decided to look deeper into how children and women compared to men of each class.



A point to be highlighted here is that when it comes to class 3, survival rates for even women and children drop significantly but are still higher than that of men. There were 17 children of Class 2 and all of them managed to survive. I found it strange at first that only just over 60% of children in class 1 survived. But a deeper look into it showed me that there were only 3 children in this category. 1 death dragged down the survival percentage to 66.6% (data in supporting document)

Another crazy statistic to look out for here is survival of men in class 3. The number of men of this class that had boarded were 325. Of them, 39 people survived.

Class 2 men shared a similar fate. Out of 99 men that had boarded, only 8 survived.

Contrast these to the women of class 1. 93 women of this category boarded, with 91 survivors.

These stats were analyzed to answer questions 3 & 4.

***Note – This does not imply that either variable caused the survival rates to change. It is a mere correlation.**

Pivoted tables below show this in detail.

TOTAL BOARDED

	Survived		
Type	Child	Man	Woman
Pclass			
1	3	120	93
2	17	99	68
3	44	325	122

TOTAL SURVIVED

	Survived		
Type	Child	Man	Woman
Pclass			
1	2	43	91
2	17	8	62
3	19	39	61

PROPORTION SURVIVED

	Survived		
Type	Child	Man	Woman
Pclass			
1	0.666667	0.358333	0.978495
2	1.000000	0.080808	0.911765
3	0.431818	0.120000	0.500000

Now, we look even deeper into the dataset. Let us analyze Class, Sex and relation size together.

This pivoted table should give us a good understanding.

	Survived					
Relation	Big		Single		Small	
Sex	female	male	female	male	female	male
Pclass						
1	0.750000	0.600000	0.970588	0.333333	1.000000	0.404762
2	1.000000	0.250000	0.906250	0.097222	0.909091	0.281250
3	0.242424	0.066667	0.616667	0.121212	0.529412	0.245283

Again, we can see that female survivals are higher than men in every section.

Another point to observe is that as we move from class 1 -> class 3, survival proportions drop. (Except for women in big families of class 2. Closer observation shows that there were only 2 women in this category)

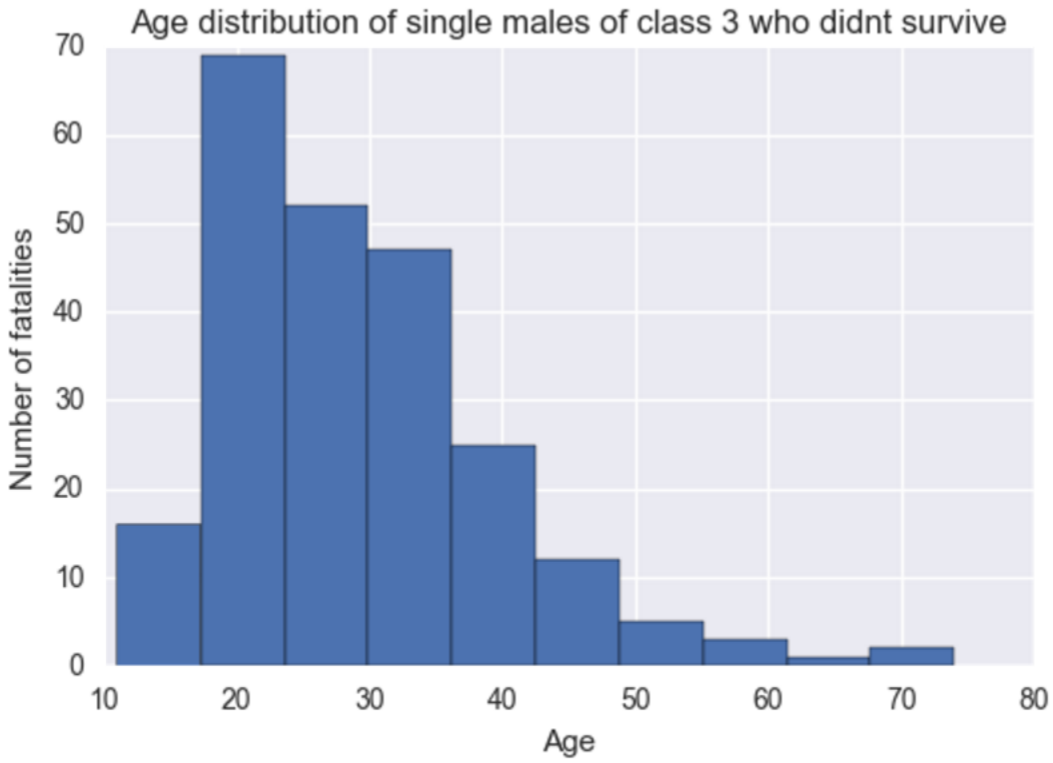
This should give us a better understanding of Questions 3,4 and 5 in tandem.

***Note – The above observations were to show correlation and I don't imply that any factor causes the survival rates to change.**

We've seen from correlations that men of class 3 tended to have lesser survival rates than anyone else. This led me to question 6.

Q6) Who tended to be the most unfortunate on the titanic.

Looking deeper into men of class 3, I decided to plot a distribution of the ages of these people who were single and had not survived.



This histogram showed us that of these men, people from age group 18 – 35 had the highest number of fatalities. **This doesn't prove that if you were a person from this category, you were unlikely to survive, but rather shows an interesting correlation to survival on the titanic.**

I also wanted to investigate the points of embarkment of different people.

This led me to Question 7

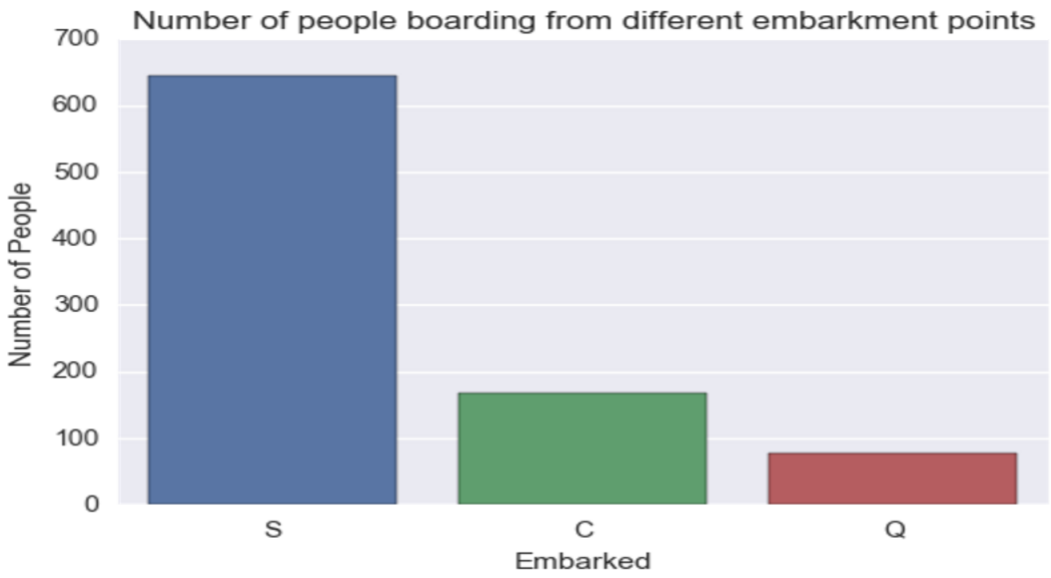
Q7) Where did most of the people embark on their journey from? What were the proportions of people of different classes from different embarking points?

Firstly, we'll need to go back to the info of the dataset.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived       891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age            714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
```

I noticed that 'Embarked' had two 'Nan' values. This isn't too many so I simply put in S (Southampton) to fill it up.

Then we have a look at how many people embarked from different points.

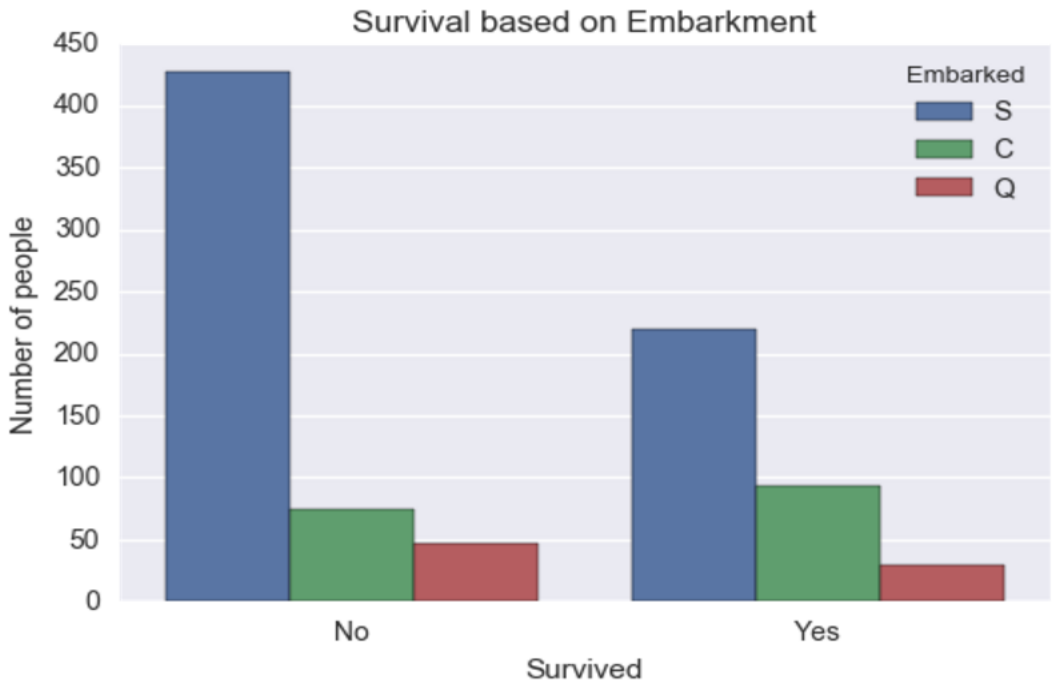


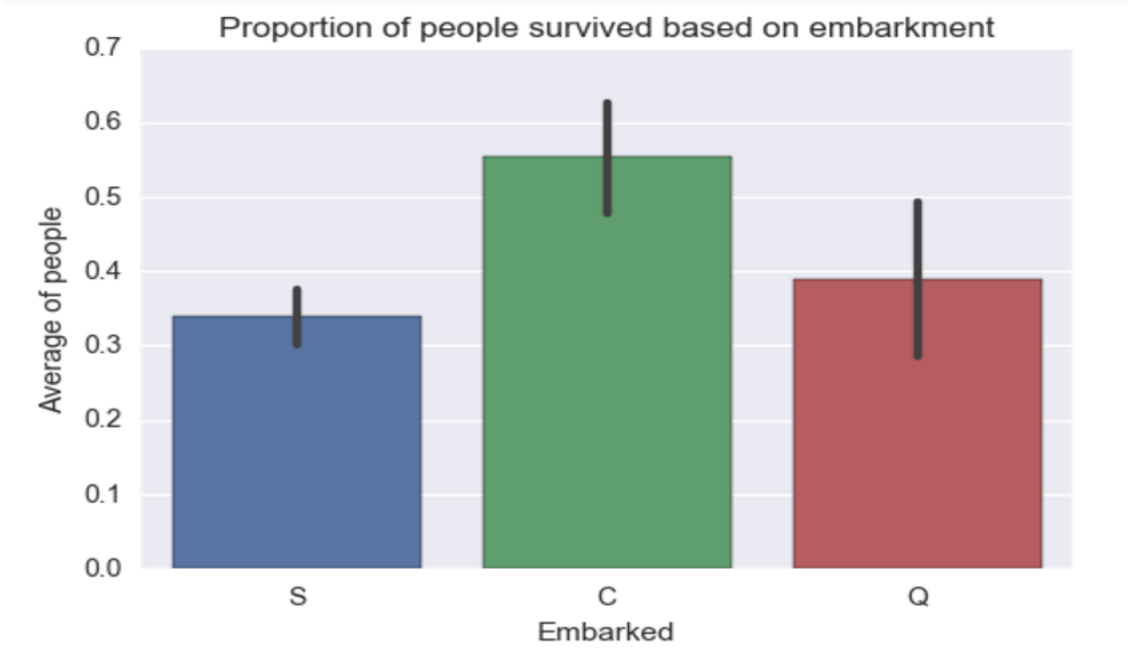
People embarked from Southampton = 644 (72%)

People embarked from Cherbourg = 168(19%)

People embarked from Queenstown = 77(9%)

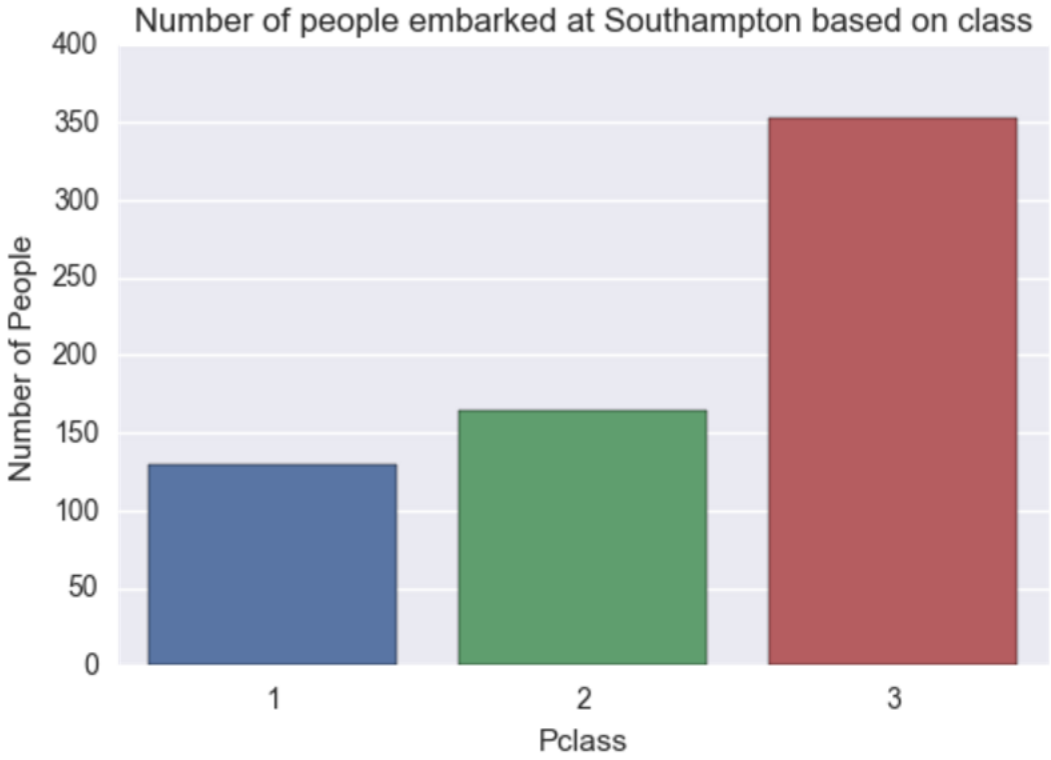
A huge majority of people boarded from Southampton. Now let's look at the survival numbers from each port.

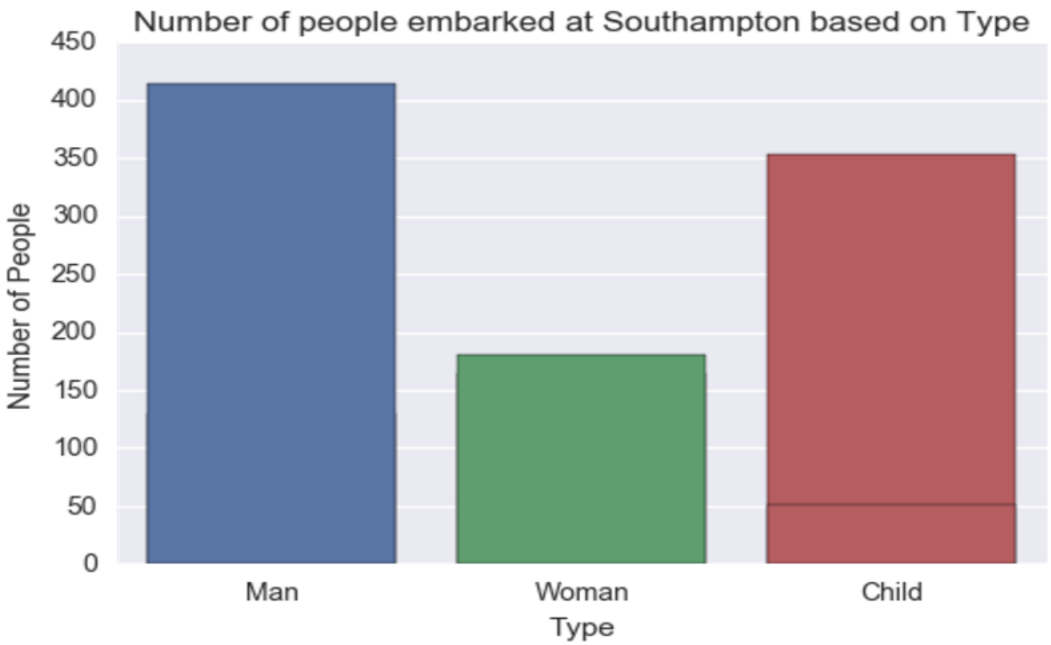




Its fairly obvious to see that most number of casualties were people who embarked at Southampton. Let's investigate further.

I wanted to check what proportion of people of each class boarded at Southampton.



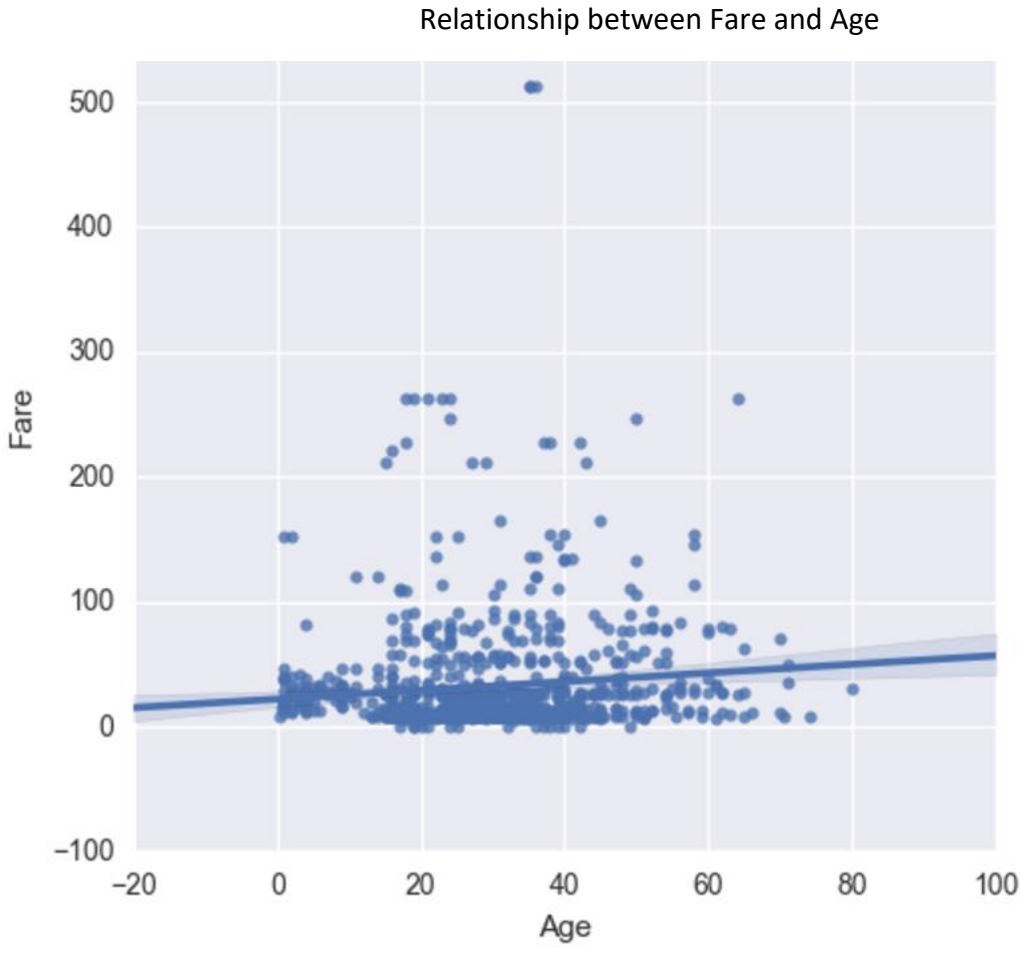


Most of the people that boarded at Southampton were people of class 3 or men. Its not really a surprise that we see high mortality rates for people who boarded from Southampton. **However, to prove causation would require more statistical tests. This could merely be a correlation.**

Moving on from fatalities, I wanted to have a look at fares people had paid. This led me to my final question.

Q8) Was there any correlation between the Ages of passengers and the fare that they paid? Did passengers tend to book fancier cabins if they were travelling with more family members?

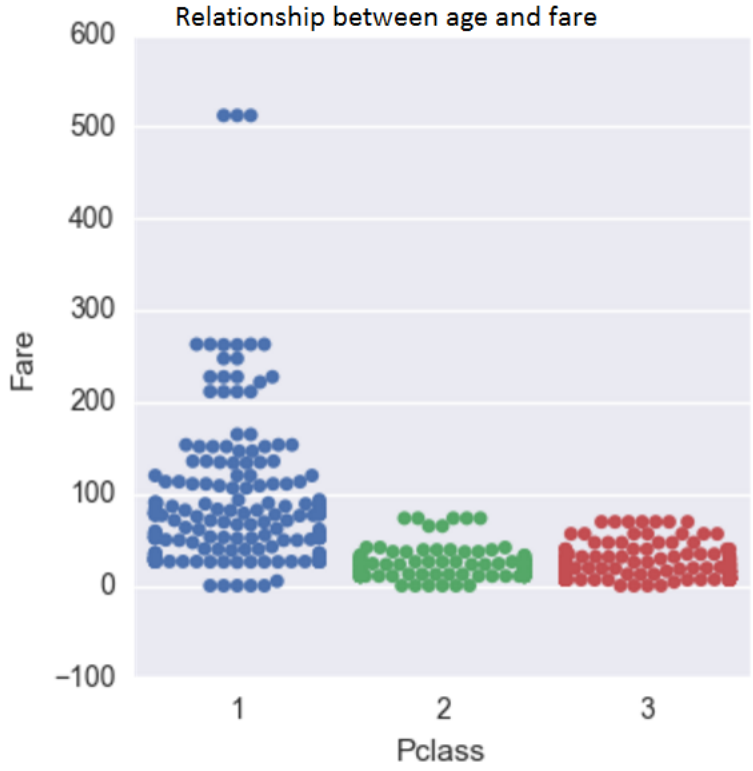
Firstly, I wanted to check if people tended to spend more the older they got.



R2 = 0.011

This plot shows us that there is almost no relationship between Age and Fares paid.

I also wanted to check how fares were distributed among classes.



I noticed that there were a lot of people of class 1 who had paid fares similar to those of class 2 and 3. I wanted to check if maybe people were paying for better cabins if they were travelling with a lot of family members.

	Fare		
Relation	**Big**	**Single**	**Small**
Pclass			
1	193.184615	63.672514	92.826686
2	34.482507	14.066106	28.026602
3	34.546032	9.272052	14.751445

The pivoted table and the graph shows that there seems to be a correlation that people who were accompanied by bigger families tended to spend more and get better cabins**. To prove the causation would require more extensive statistical tests.**

Results

1) Mean and median (After filling out missing ages) = 29.38 and 28 respectively. Mode is also = 28, Thus showing a normal distribution.

2) Women and children had a much better chance of surviving than that of men.

3) Survival trends skewed very much in favor of the upper class people (1 & 2). This was true for both sexes and for all age groups as well.

4) Ages that were most likely to survive were in the range of 17 - 35(depending on the family sizes they were supporting) and also infant children.

5) Single people had the highest number of casualties but they also had a huge number of people that had embarked on the journey. Women who were single or were accompanied by small families had very good survival rates. Men who were accompanied by small families had better survival rates when compared to men who were single or men with big families.

6) Data correlations show us that single men of class 3 within the age group of 17-35 were most likely not to survive

7) Most of the people embarked on their journey from Southampton with the highest proportion of people from Pclass = 3. Passengers that had embarked at Southampton also had the highest proportion of fatalities. People who boarded from Cherbough had the best survival rates.

8) Correlation between age and fares was very poor. There were trends that showed that people tended to pay for more expensive tickets when travelling with bigger families.

****NOTE- It is important to understand that the above results were only a description of correlation. Further statistical tests will be required to prove that these factors caused the survival rates to change.**

Conclusion

In conclusion, I will summarize the steps taken to analyze this dataset.

The first step is always to just look at the data and to familiarize yourself with it and understand the variables involved.

Secondly, it is good practice to pose a few simple questions about your dataset and to see if you have the means to answer these questions. This may also lead you to find anomalies and missing values, etc. For example, in this dataset, posing a simple question like average age of people on the titanic led me to realize that there were lots of data points in the age column that were missing.

Later, I managed to wrangle and clean the data in order to remove any obstacles from my analysis.

Then I got back to answering my questions about age, class, families, etc that led me deeper into the rabbit hole to look for answers based on many variables at once.

Graphically depicting these values to see various correlations help you understand what aboard the titanic.

The only thing missing from my analysis would be performing any sort of statistical test to prove causation between variables and survival rates.

Recommendation

As stated above, performing additional statistical analysis would help to prove causation of certain variables.

Tests like t-tests are very difficult to implement because it would involve simulating the test conditions. However, t-tests tests could be simulated in a way that calculates the amount of time taken by all people to reach certain 'safe zones' (that simulate location of life boats) based on various hypothesis statements.

Other alternate hypothesis could that it would take lesser time to evacuate people If cabins catering to people of class 3 had better access to lifeboats vs the null hypothesis stating that there would be no change in time taken.

Another alternate hypothesis could be to state that it would take lesser time for people to escape onto lifeboats if they were properly instructed on what to do in case of an emergency vs no instructions at all as opposed to the null hypothesis that there would be no change in time taken.

Many experiments can be conducted based on location of lifeboats, instructions, announcements, mode of evacuation, etc by testing one variable while keeping all other constant.

References

- https://www.kaggle.com/c/titanic/data
- http://stackoverflow.com/
- Python for Data Analysis - Wes McKinney