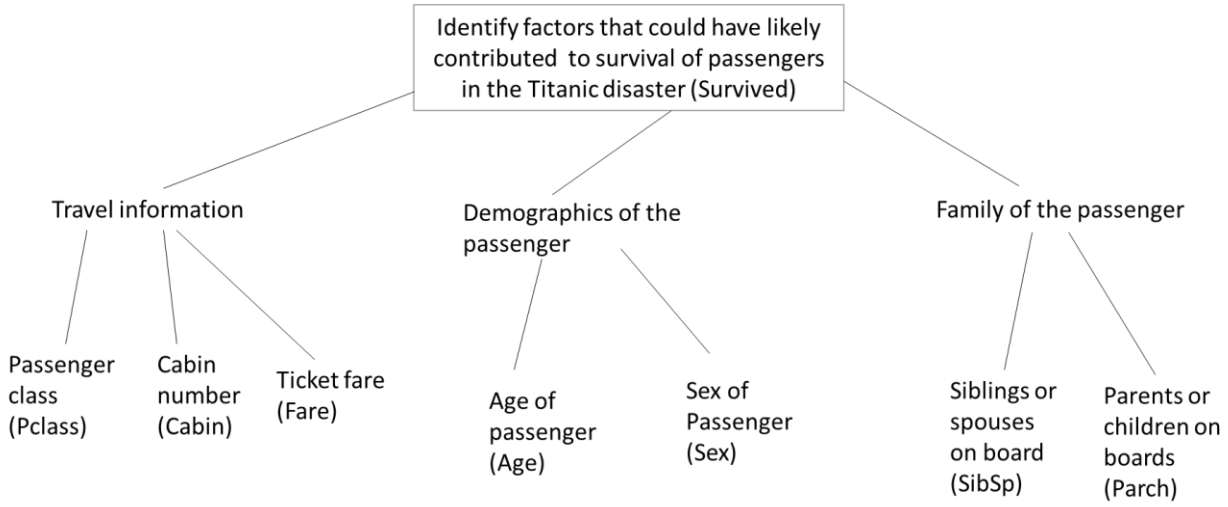
P2: Titanic Dataset

1. QUESTIONS PHASE

To come up with questions, structured pyramid analysis plan was used with dependent variable on the top of the pyramid and the independent variables below. The name of the variables is given in brackets wherever applicable. The aim of this data analysis was to answer the question: *Which factors could have likely contributed to the survival of passengers of Titanic?*



Hence, in detail the following questions are analyzed:

- *Was there a preference for passengers from 1st class when compared to 2nd and 3rd class passengers?*
- *Similarly, did fare of the ticket play a role in survival of passengers?*
- *Did demographics of the passenger have an effect on their survival? If so what was it? Age or Sex?*
- *Finally, did passengers travelling with families suffer more than individual passengers?*

In this analysis, the major factors that affected survival of passengers are researched. For the factors identified, statistical test is performed to prove their relationship with survival of passengers.

2. WRANGLING PHASE

Wrangling phase consists of data acquisition and cleaning. The data was obtained as a '.csv' file. It was imported into python as a pandas DataFrame. Cleaning of data involved checking of the imported data for accuracy, data redundancy and errors in data entry. (Nedarc.org, 2016)

The majority of the data was found to be in a good state with no duplicate records. The numerical values of the Dataframe were checked using box plot. 'Fare' column had a few outliers, but it is expected since only a few tickets are expensive while the majority are priced at a moderate range.

However, while finding the minimum and maximum of each column, 'Fare' values were found to have zeros. Hence the mean fare of the respective passenger class was given to those zero values.

3. EXPLORATION

This phase involves analyzing data to answer the questions from first phase. The following data was studied:

1. Travel Class Vs Survival
2. Effect of Fare
3. Age vs Survival
4. Sex vs Survival
5. Age and Sex combined analysis
6. Effect of Family on Survival

Cabin number from the questions phase were not studied since the data was not fully available for 'Cabin' column. Various visualizations are presented for the grouped data. The description of each graph is presented in the ipython notebook itself, right below the graphs generated.

4. CONCLUSION:

Each visualization from analysis is discussed to determine major findings.

1. Travel Class Vs Survival: We see from the graph that the number of people who survived from first and second class is about fifty percent of their group. Only one in three third class passengers seem to have survived the accident. Therefore, there is likely to be some kind of preference for 1^{st} and 2^{nd} class passengers.

2. Effect of fare: The graph mean fare of tickets for passengers who survived and didn't survive conveys that higher fare corresponds to survival and lower fare to death. This is effectively same as the previous graph.

3. Age vs Survival: This analysis consists of three graphs: two histograms and a bar chart. First histogram shows the distribution of general population. It shows a normal distribution of age of passengers on board. The second figure has two histograms plotted in it: one for passengers who survived and one for who didn't survive. From this it is evident that both histograms follow normal distribution. However, close inspection reveals that the age histogram of people who survived contains most of the values in the first bin from population histogram. Thus, there seems to be some preference for people who were young at that time. Hence further analysis is done. Finally, in the third graph (age classification vs count of passengers), we can see that survival percentage of children is the best of all age groups. This confirms the indications from histogram.

4. Sex vs Survival: The graph depicts count of males and females separated according to survival. We observe that the number of males were more than number of females. Also, that the ratio of survival is high for females than males.

5. Age and Sex combined analysis: Since we found that age and sex of passengers affected their survival, both age and sex data from previous analysis is used to find the combined effect. The resulting graph of this analysis gives a clear picture of what might have affected the decision whether passengers got a

lifeboat or not. We see that young passengers and ladies have been given preference, while men have suffered the most.

6. Effect of Family on Survival: There are two graphs in this section. First line graph depicts the percentage of survival for people with siblings or spouses on board. Second one depicts the percentage of survival for passengers with parents or children on board. The first graph suggests that passengers with many siblings or spouses had a lesser chance of survival. However, the same is not reflected in the second graph for parents and children on board. Hence, the results are inconclusive and family does not seem to have an effect on the survival chance of passengers.

From the analysis we found that passenger class, and demographics of the passenger have major impact on survival. Hence a statistical test is pursued to find if both factors really increased survival chances on that doomed day (April 14, 1912).

4.1 Z-test for Demographics of passenger (Age and Sex):

A two proportions z-test test is performed where first sample consists of passengers who are female and have age under or equal to 10. The second one is a random sample from population. Z-test assumes that the distribution of sample means follows a normal distribution according to central limit theorem and the sample is big enough (n>30) (Wikipedia, 2016). Two proportions Z-test is selected since the variables are categorical and the samples are independent. Also, the population is at least 10 times larger than the sample (Stattrek.com, 2016).

Hypothesis:

H0: There is no significant difference between the two population proportions. $P_1 <= P_2$
H1: There is a significant between the two population proportions $P_1 > P_2$
where P is the proportion of the population.

Result:

The analysis is done in ipython notebook. The results reveal that H0 is rejected and the test is significant at $p < 0.05$. Hence we can say that age and sex were factors that determined the survival of passengers on RMS Titanic.

4.2 Z-test Passenger Class:

Similarly, Z-test for passenger class is performed. The first sample consists of passengers who are only from first class. The second sample consists of random passengers from population.

Hypothesis:

H0: There is no significant difference between the two population proportions. $P_1 <= P_2$
H1: There is a significant between the two population proportions $P_1 > P_2$
where P is the proportion of the population.

Result:

The analysis is done in ipython notebook. The results reveal that H0 is rejected and the test is significant at $p < 0.05$. Hence we can say that Passenger class was a factor that determined the survival of passengers on RMS Titanic.

Thus we prove that Travel class, Age and Sex of the passenger affected their chance of survival.

5. LIMITATIONS OF THE WORK

Even though we have proved causation through statistical tests after data analysis. There are several limitations of this work and the results must be interpreted with caution.

- First, there are errors present in the dataset analyzed, which could impact the results. Best effort has been taken to minimize such errors. For e.g. Age data and particularly Cabin number data is missing for a lot of records. Also, the value of fare is found to be zero in a few records, which has been replaced with mean fare amount for the respective travel class. Hence, the analysis is done with the information available.
- Out of 1316 passengers, data for only 891 passengers is available which could introduce bias in our results. Hence, data must be as big as possible to do dependable analysis.
- Moreover, some of the charts used in the data analysis might mislead to believe certain relationships. They might not be appropriate for such an analysis. Thus the readers are advised to interpret it with caution.
- When the data set is quite big and the more analysis we do, the more probability is to find statistically significant correlations. They may just turn out to be spurious and deceive the analyst.
- The questions asked in this analysis may be wrong and may not lead to the results that are hidden in the data. Hence It is important to ask the right questions to get the right answer. The questions in this report aim at identifying factors that affected survival of passenger of Titanic. However, there may be few questions that may better answer this question that are not addressed in this report.

6. REFERENCES

DataFrame, C. (2016). *Converting a Pandas GroupBy object to DataFrame*. [online] Stackoverflow.com. Available at: http://stackoverflow.com/questions/10373660/converting-a-pandas-groupby-object-to-dataframe [Accessed 9 Apr. 2016].

Nedarc.org. (2016). *NEDARC - Purpose of Data Cleaning*. [online] Available at: http://www.nedarc.org/tutorials/analyzingData/cleanTheData/purposeDataCleaning.html [Accessed 9 Apr. 2016].

Pandas.pydata.org. (2016). *pandas: powerful Python data analysis toolkit — pandas 0.18.0 documentation*. [online] Available at: http://pandas.pydata.org/pandas-docs/version/0.18.0/ [Accessed 9 Apr. 2016].

Matplotlib.org. (2016). *pyplot — Matplotlib 1.5.1 documentation*. [online] Available at: http://matplotlib.org/api/pyplot_api.html [Accessed 9 Apr. 2016].

Faculty.csupueblo.edu. (2016). *Z Test summary*. [online] Available at: http://faculty.csupueblo.edu/paul.Chacon/156Spr05/HYPO_z_test_summary.html [Accessed 9 Apr. 2016].

Wikipedia. (2016). *Z-test*. [online] Available at: https://en.wikipedia.org/wiki/Z-test [Accessed 9 Apr. 2016].

Stattrek.com. (2016). *Hypothesis Test: Difference in Proportions*. [online] Available at: http://stattrek.com/hypothesis-test/difference-in-proportions.aspx [Accessed 10 Apr. 2016].