Assignment 2

Data Visualization

Harisaipravin Sv - 17pw13

The Great Sink of Titanic

Facts !

The RMS Titanic was the world's largest passenger ship when it entered service, measuring 269 metres (882 feet) in length, and the largest man-made moving object on Earth. The largest passenger vessel is now Harmony of the Seas, at 362.12 metres.

Dataset

The dataset is derived from kaggle. The link to the dataset is given below.

https://www.kaggle.com/c/titanic

The dataset consists of 11 columns,

- ➔ PassengerId
- ➔ Pclass
- ➔ Name
- ➔ Sex
- ➔ Age
- ➔ SibSp
- ➔ Parch
- ➔ Ticket
- ➔ Fare
- ➔ Cabin
- ➔ Embarked

Data Imputation

There are few missing data, which is imputed using data imputation techniques. Here we use fit_transform from the sklearn library in python which uses deep learning methods to predict the Nan rows.

Preprocessing

Further, the age is converted from numerical to categorical data type for easier visualization.

Domains to check

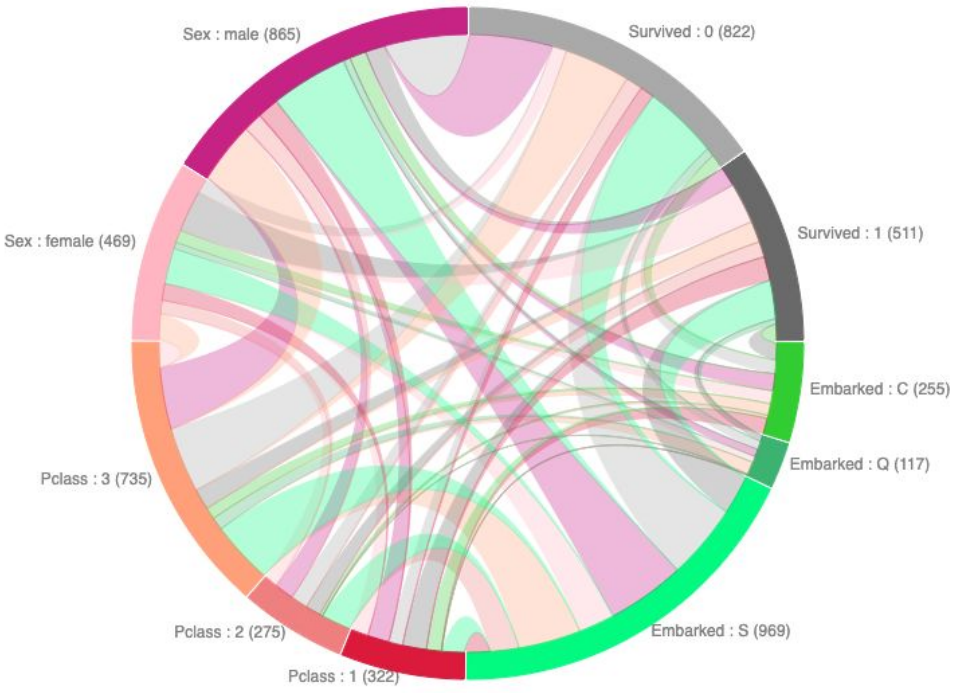
1. Who is more likely to survive - (Male or Female) & (Class) ?
2. Which Age is more likely to survive ?
3. Would Couples survive or singles survive ?
4. Single chart that interprets all information.

Data Investigation

Sex and Names are Inter related to each other. The name section consists of prefix like Mr. And Mrs. which can be used to determine sex.

The Pclass and ticket are directly related to each other.

Main Chart - Strip plot



Titanic Visualization

Data attributes considered :-

- Passenger Survival
- Class of passengers
- Sex of Passengers
- Port of Embarkation

Some key findings from this chart include,

1. Overall, **65%** of the passengers were males, while **35%** were females.
2. **74%** of the females survived, as compared to only **19%** of the males.
3. Overall, the survival rate was **38%**.
4. While **63%** of I-Class passengers survived, only **24%** from III-Class could make it.

A chord diagram represents flows or connections between several entities. The size of the arc is proportional to the importance of the flow.

The chord diagram is chosen here to give detailed view on variations of data between the class of the passengers and their sex.

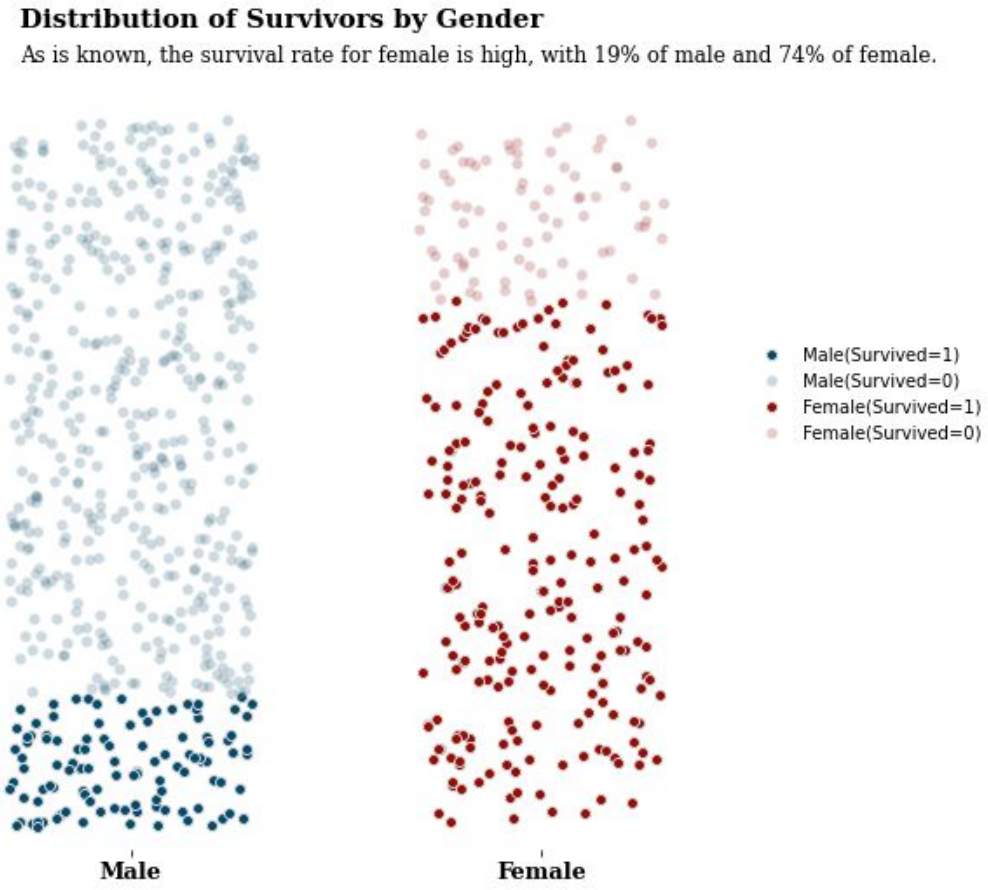
As expected, the survival rate of women was very high when compared to that of men. There was a difference of **55 %** between the survival rate of women and men.

The given bar + scatter chart below gives a detailed view of the total passengers in each sex followed by their survival rate.

A strip plot is a scatter plot where one of the variables is categorical, here we calculate the mean first and mark the difference using colours.

Distribution By Gender

female	0.742038
male	0.188908



The Dark blue specifies the Male survivors, while light blue represents male non survivors. Similarly, red represents female survivors and light red represents female non survivors.

Care is taken when assigning the colors to the strip chart. Both are vibrant colors which would attract attention with ease.

Distribution By PClass

Similar to gender distribution, this chart provides distribution by PClass.Here there are 3 classes followed by crew members.

Some useful facts that can be derived are :-

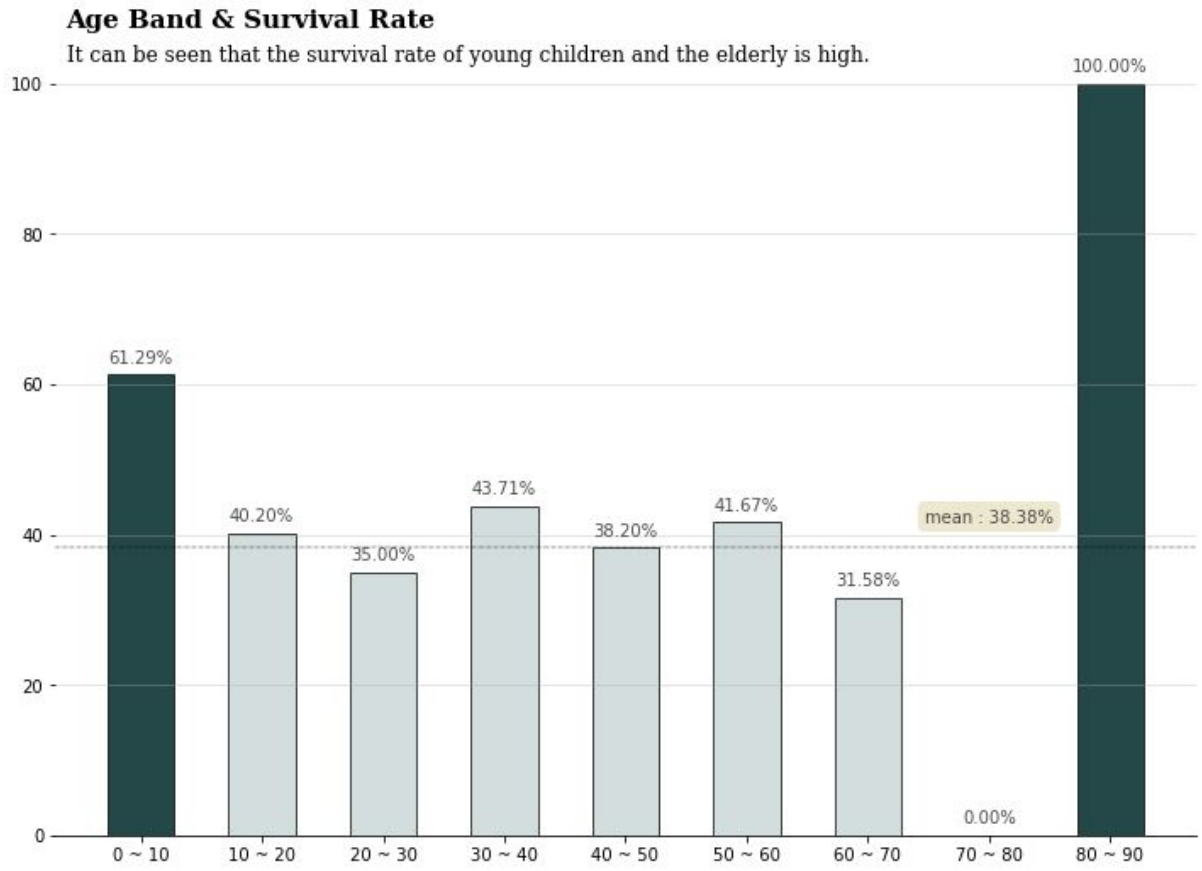
97.1 % of first class women survived. - Rose

Only 13.6 % of third class men survived. - Jack

The data and results derived from these charts are very favourable to what actually happened to Rose and Jack.

Rose lived while Jack sinked.

Survival by Age



This is a simple bar graph which is coloured based on its frequency.

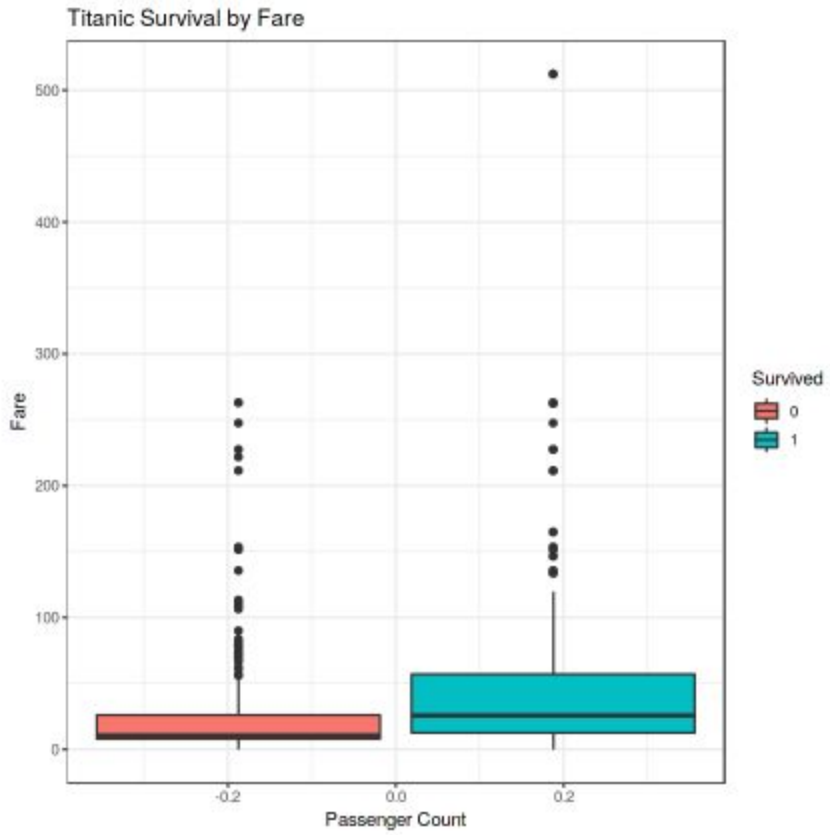
During the disaster, The priority was first given to elderly and young people as they are most vulnerable. But due to the lack of safety boats, The middle aged people were left stranded in the sinking boat.

Further, on analysing the chord chart, we could find that around **80 percent** of crew members died in the boat and all of them were middle aged.

This is very alarming as the crew members risked their life to save those of the passengers.

Survival by Fare

To explore multiple data visualization methods, we make use of box plots to analyse survival rates by fare.

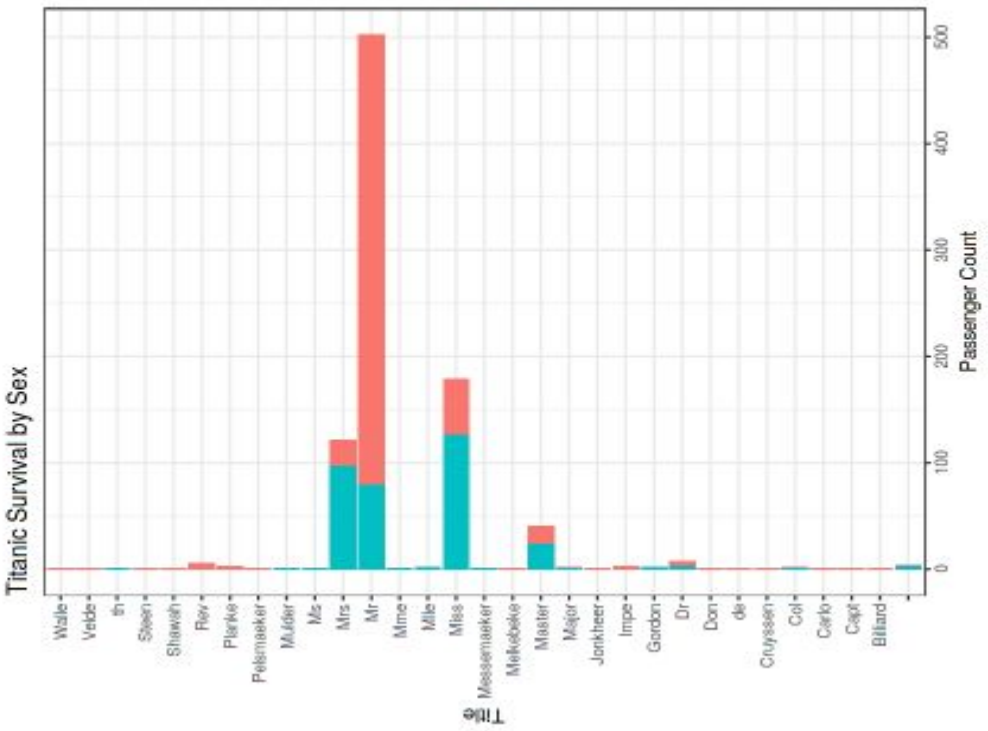


A box plot or box plot is a method for graphically depicting groups of numerical data through their quartiles. It is used to analyse the quartile ranges and outliers.

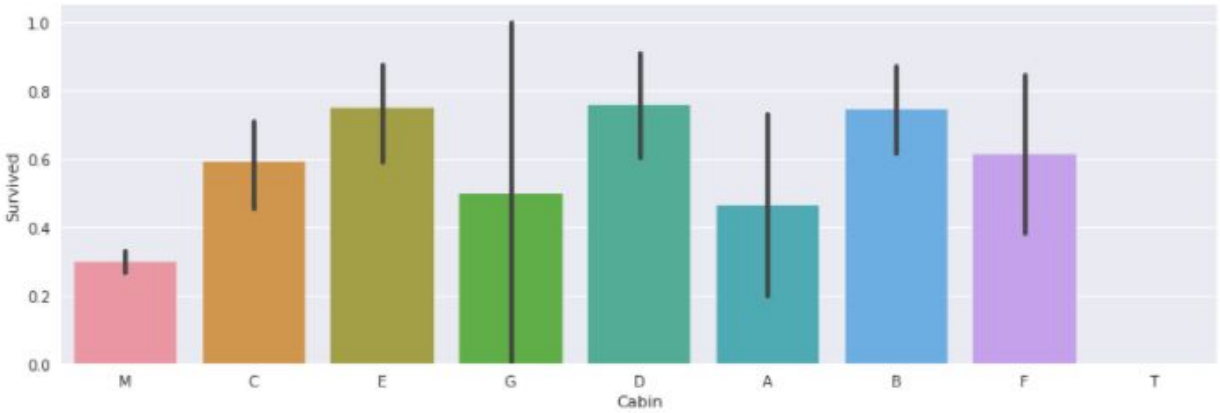
Here it is evident that the area covered by surviving passengers lies higher than the area covered by dead passengers. On close analysis, it shows that if the ticket fare is above 30, then most of them have survived.

The richest man, whose ticket was worth 510 dollars was also saved, and he was given a single boat for his family alone. Which later proved fatal for others.

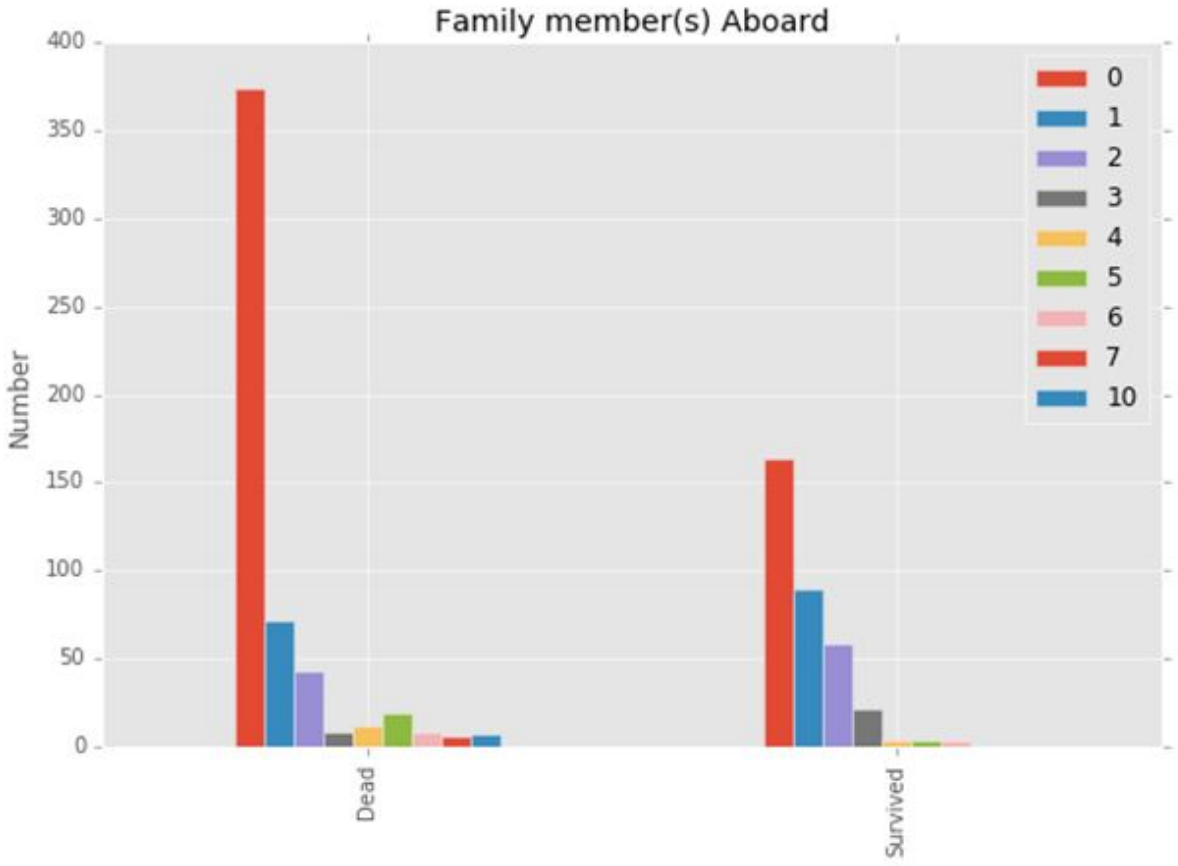
Survival by Prefix in name



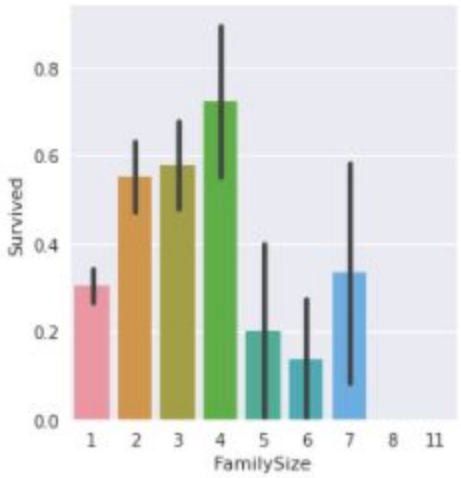
Survival by Cabin they were present



Survival by Amount of Family members Aboard



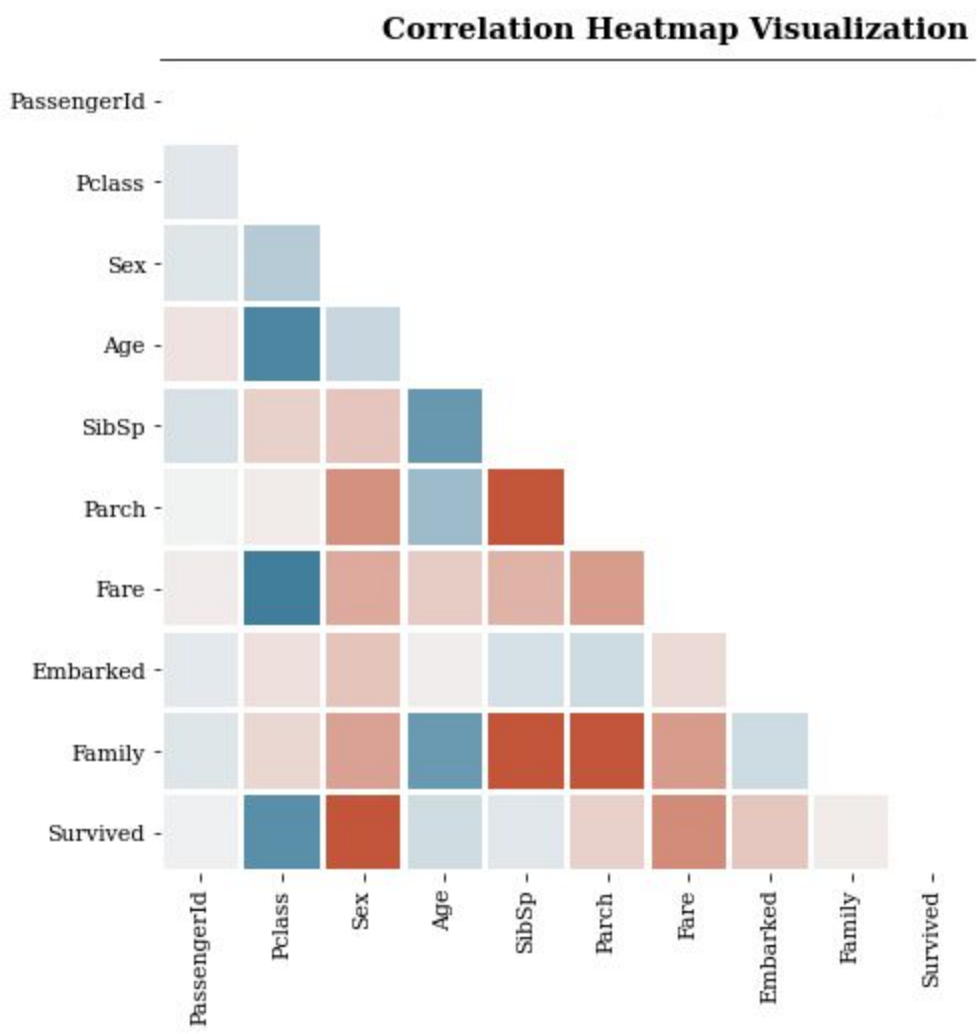
It could be seen that those with kids or those with family in the ship were more likely to survive the disaster than the single people who were on board.



This chart gives a detailed view on the size of the family on board of the people who have survived.

Thus family people are more likely to survive.

HeatMap



A heat map (or heatmap) is a data visualization technique that shows the magnitude of a phenomenon as color in two dimensions.

This heatmap summarizes all the data and inter relates them using their magnitude.

This heat map along with a dynamic strip plot provided at the start of the report gives a complete overview of the titanic dataset. It is also made interactive with hover on properties.

Timeline of Titanic

Titanic starts on April 10th and sink by 2.20 AM on April 15th

Quick Facts:-

- There were only 23 female crew members.
- There were only 20 lifeboats as compared to 65 which it was designed for.
- It was over 882 feet long.That's almost three football fields.
- The ship cost $7.5 million to build.
- A woman who survived the Titanic's sinking later survived the sinking of another ship.
- The 100-foot chunk of ice was traced back to a glacier in Greenland.
- It took on 400 tons of water per minute after it hit the iceberg.
- The ship's lookouts couldn't access the binoculars.
- -2°C – the temperature of the sea water in the area where the Titanic sank.
- 269.1 metres – the length of the Titanic (882 feet 9 inches).
- 825 tons – the amount of coal used per day.

References:-

Theory And data

1. https://titanicfacts.net/
2. https://en.wikipedia.org/
3. https://ovayozabalogun.medium.com/
4. https://www.kaggle.com/

Coding And Visualization

1. https://towardsdatascience.com/
2. https://d3js.org/
3. http://chartmaker.visualisingdata.com/
4. https://public.tableau.com/