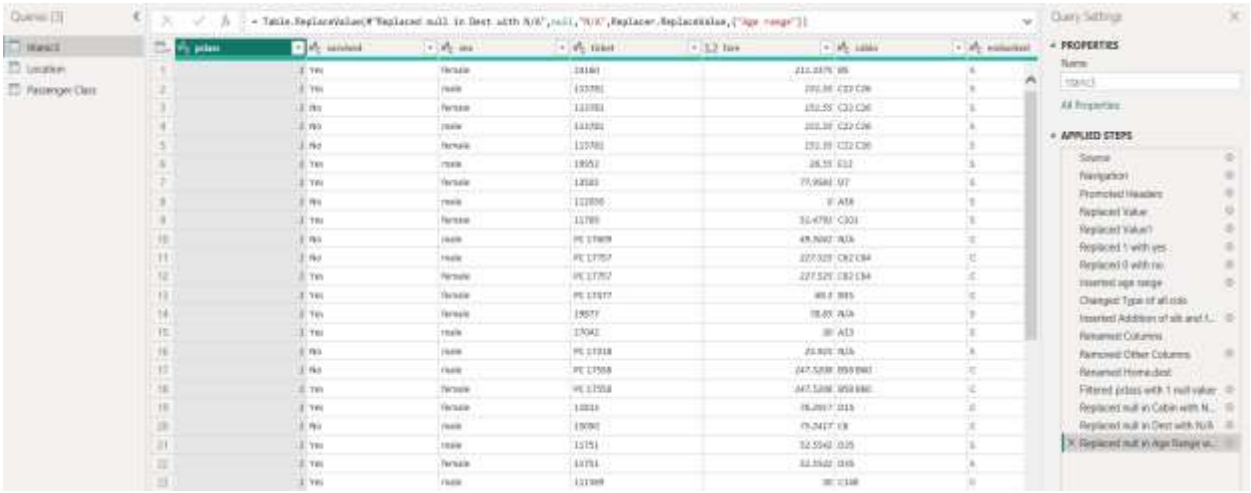
Introduction

The Titanic dataset provided for the assignment had 1310 rows and 14 columns, namely pclass, survived, name, age, sex, sibsp, parch, ticket, fare, cabin, embarked, boat, body and home.dest. The objective was to build a data model, describe the dataset and perform suitable visualization.

Cleaning the data

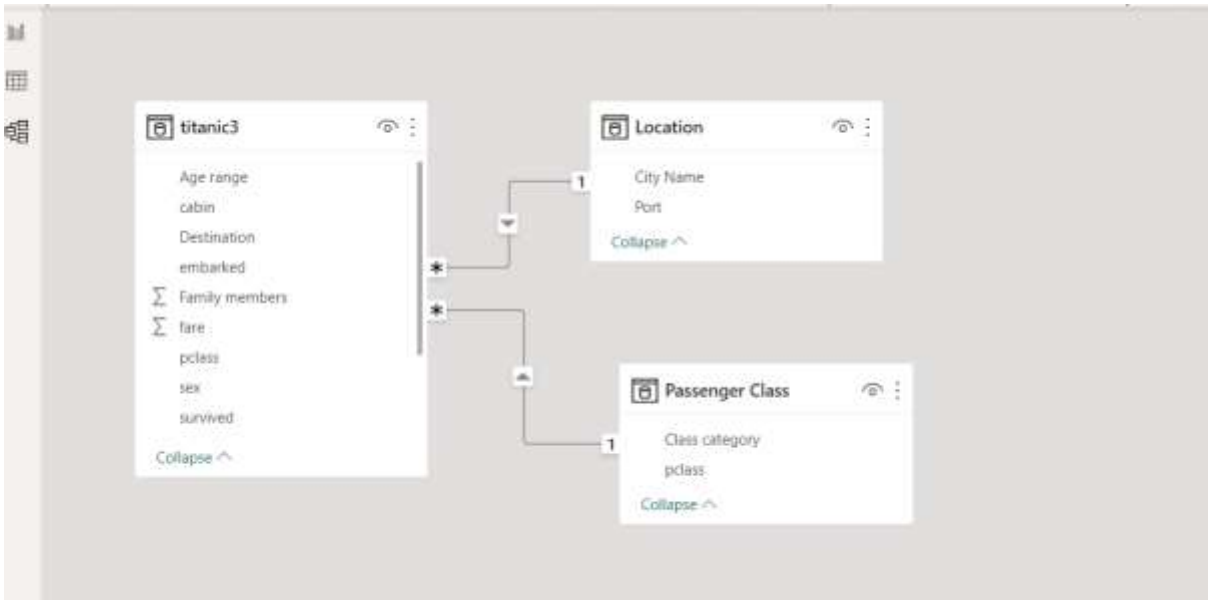
In order to conduct exploratory analysis from the data and find out which predictors could affect the survival rate I loaded the file in Power BI and the first step was to clean the data in Power Query editor.



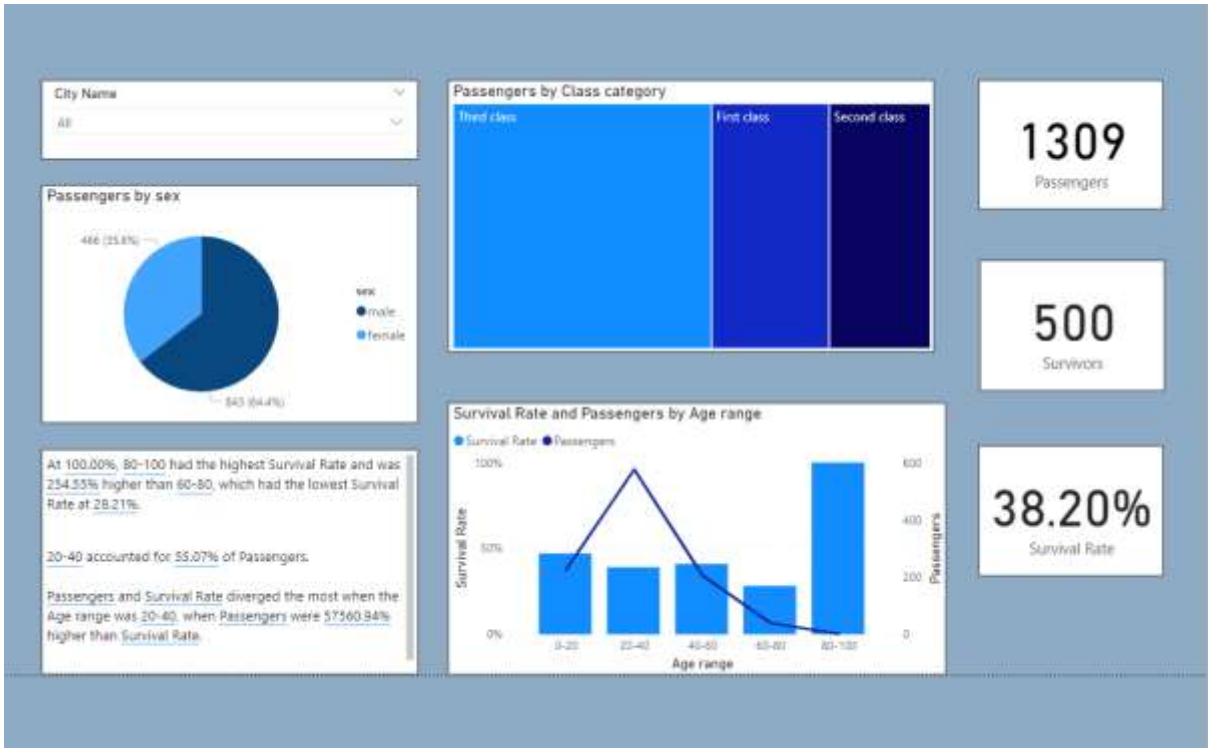
Cleaning involved changing the binary values in the Survival column to 'yes' for 1 and 'No' for 0, ensuring that the data types for each column were relevant to their values. I also added two new columns – one for Age Range constructed with reference to the ages and another was an addition of sibling/spouse and parent/children, called Family Members. Now I also noticed that there were some missing values in pclass, Cabin, Destination and Age Range columns. For pclass, there was only one null row that was null across all the columns, so I removed it. For Cabin, Destination and Age Range, I replaced the null values with N/A as I felt it did not make sense to remove the rows completely because they contained vital information in the other columns. Finally, I loaded only the columns that I felt were relevant to the analysis onto Power BI from Query window.

Building relationship model

Next step was building a data model. I created two more Excel files called Location and Passenger Class. In Location, I added two columns – one with the unique IDs of embarked ports C,S,Q and another column that assigned Cherbourg, Southampton and Queenstown to these values. In the Passenger Class file, the unique ID column was pclass with values 1,2,3 and another column Class category with First, Second and Third class. In Power BI, I loaded these two files additionally and built a one-to-many relationship with the titanic table. The relationship table is shown below.



Visualization and analysis



The data shows that there were 1309 passengers of which 500 people survived, bringing the survival rate to 38.2%. The passengers were categorized on age, gender, class and city.

- 64.4% of the passengers were male and 35.6% female. But the survival rate of females was 72.75%, a high number as compared to a mere 19% for males.
- There were more third-class passengers as compared to second or first class. But I noticed that for First class passengers, the survival rate for all age groups was fairly high, bringing the effective rate to 61.92%.
- Maximum number of passengers were between the ages of 20-40 (107, to be exact), as seen from the line stacked chart and there was only 1 old passenger aged between 80-100.

- Kids between 0-20 had an overall high survival rate of 47.11% as compared to other age groups.
- Interestingly, Cherbourg passengers had the highest survival rate amongst other ports, bringing it to 55.56%. Of the 270 passengers from Cherbourg, 150 people survived. Furthermore, kids and young people aged between 0-20 had the highest survival rate of 72.5%.

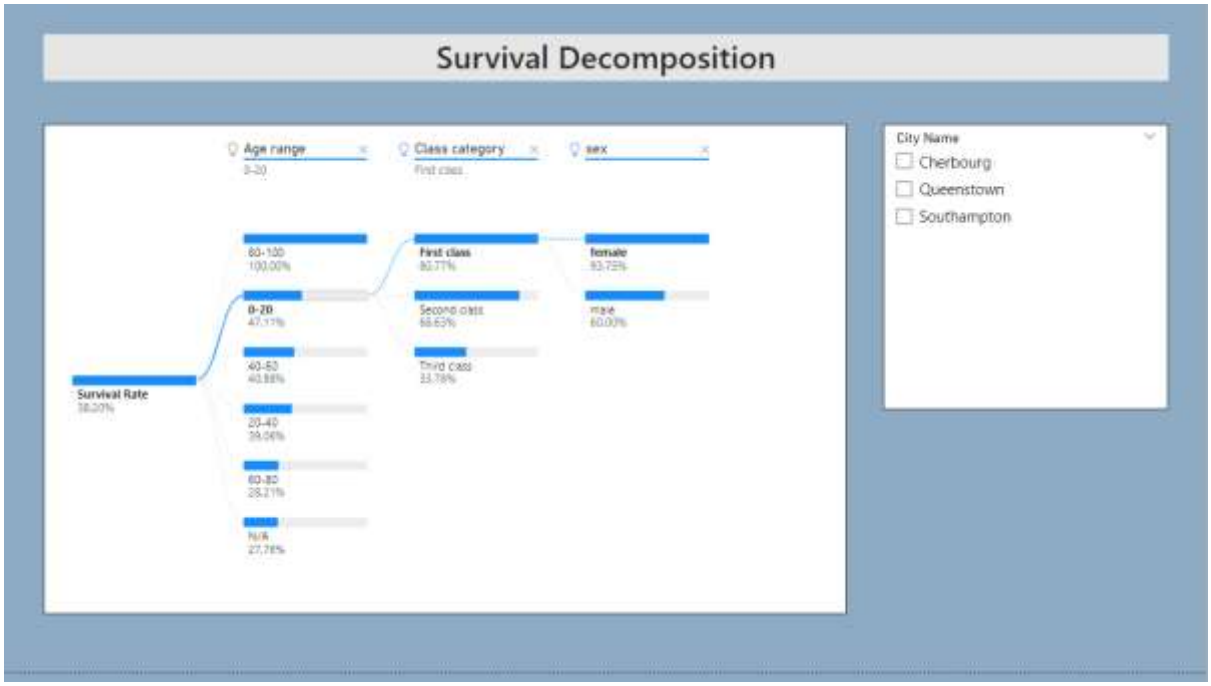
This brings us to determine factors that actually influenced the survival chances of an individual.

As seen from the diagram above, one's chances of survival were to be greatly impacted if the person was:

- a female
- belonging to first class
- boarded from Cherbourg
- between the ages of 0-20.

Adversely, being a male would indicate that chances of dying was 3.21 times more likely than females.

Finally, we calculate the exact percentages of survival rates under the influence of predictors as shown in the diagram below. We see that females belonging to first class and aged 0-20 would have had a 93.75% chances of survival.



Overall, I have tried to summarize my findings from the metadata and throw some light on factors that might have played a role in deciding the fate of the passengers aboard the Titanic.