UCD Michael Smurfit Graduate Business School

M.Sc. Business Analytics

MIS41430 – Mastering Big Data

Assignment 2

Professor: Dr. Martin Perry

Name of Student	Student ID	Email ID
Adithya Vivek	22200615	adithya.vivek@ucdconnect.ie

Table of Contents

Introduction

This Power BI visualisation report utilises the infamous RMS Titanic sinking dataset for exploratory data analysis (EDA). This analysis aims to investigate the variables that affected the Titanic passengers' survival rates. This report aims to gain insights and develop a predictive model to determine the characteristics of those more likely to escape the tragedy by examining variables such as passenger class, age, Gender, socioeconomic level, and more.

Data Cleaning, Transformation and Modelling

In the Titanic3 Excel file containing 1309 rows of data, we have changed column names ***survived*** to ***survival***, ***sex*** to ***gender***. In the ***survived*** column, we have replaced ***1*** with ***Survived*** and ***0*** with ***Died***. Another column is created- ***Age Range***. This categorises the age in a range of 20 years: ***0-20***, ***20-40***, ***40-60***, ***60-80*** and ***80-100***. The number of siblings/ spouses abroad column(***sibsp***) and the number of parents/ children abroad column(***parch***) values have been added into a new column (***Family Members***). ***sex*** column has been renamed to ***Gender***. Finally, chosen columns are- ***pclass***, ***survival***, ***Gender***, ***ticket***, ***fare***, ***embarked***, ***Age Range*** and ***Family Members***.

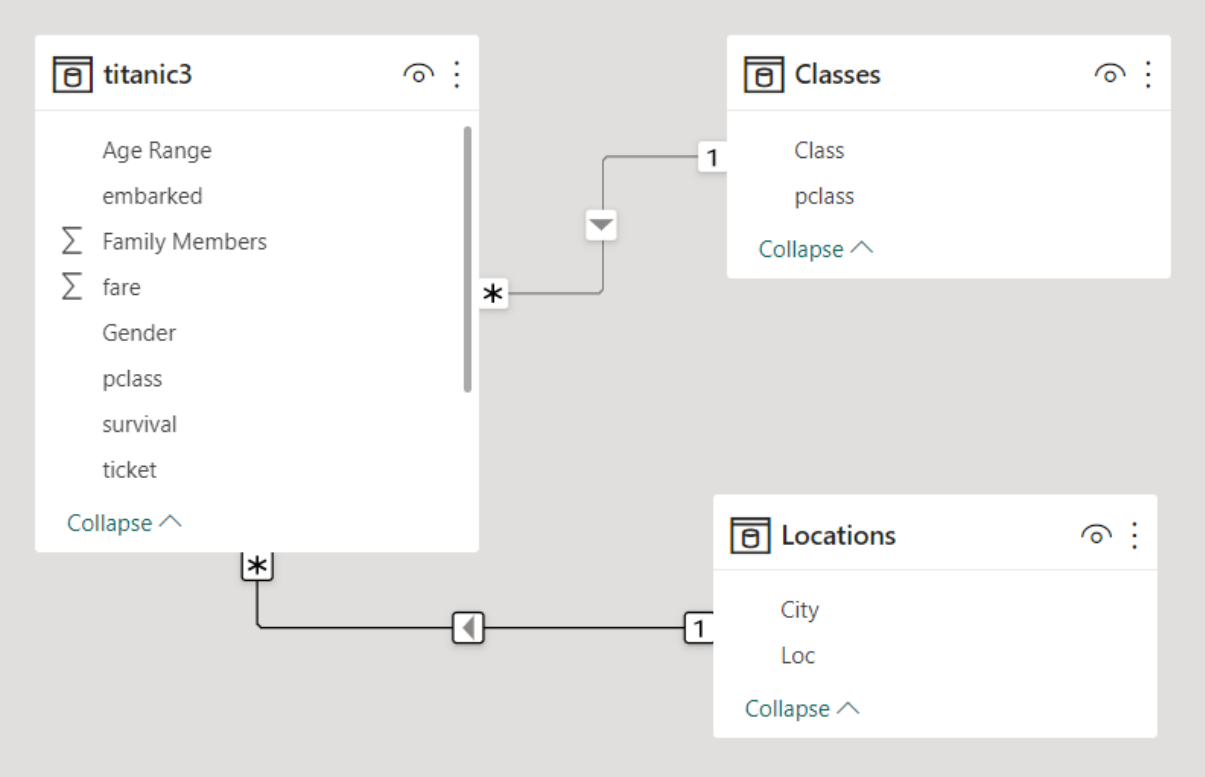
pclass	survival	Gender	ticket	fare	embarked	Age Range	Family Members
1	Survived	female	24160	211.3375	S	20-40	0
1	Survived	male	19952	26.55	S	40-60	0
1	Died	male	112050	0	S	20-40	0
1	Died	male	PC 17609	49.5042	C	60-80	0
1	Survived	female	PC 17477	69.3	C	20-40	0
1	Survived	female	19877	78.85	S	20-40	0
1	Survived	male	27042	30	S	80-100	0
1	Died	male	PC 17318	25.925	S		0
1	Survived	female	11813	76.2917	C	20-40	0
1	Died	male	13050	75.2417	C	20-40	0

Another Excel sheet has 2 sheets of data, one containing ***Classes*** where the ***pclass*** has been defined a ***Class*** and another containing the ***Locations***, where each ***Loc*** has been denoted with a ***City***. Given below are the visuals for the same-

pclass	Class
1	First
2	Second
3	Third

Loc	City
S	Southampton
C	Cherbourg
Q	Queenstown

Next, among the three tables, relationships have been established between ***pclass*** column of each ***Titanic3*** and ***Classes*** table and another relationship between the ***embarked*** column of the ***titanic3*** table and the ***Loc*** column of the ***Locations*** table. Given below is the visual for the same-



3 new measures have been created and placed in separate cards. The new measures used are ***Passengers***, ***Survivors*** and ***Survival Rate*** where the formulas for the same have been defined as follows-

- ***Passengers = COUNT(titanic3[ticket])***
- ***Survivors = CALCULATE ([Passengers] , titanic3[survival] = "Survived")***
- ***Survival Rate = DIVIDE([Survivors] , [Passengers] , 0***)

Insights and Analysis

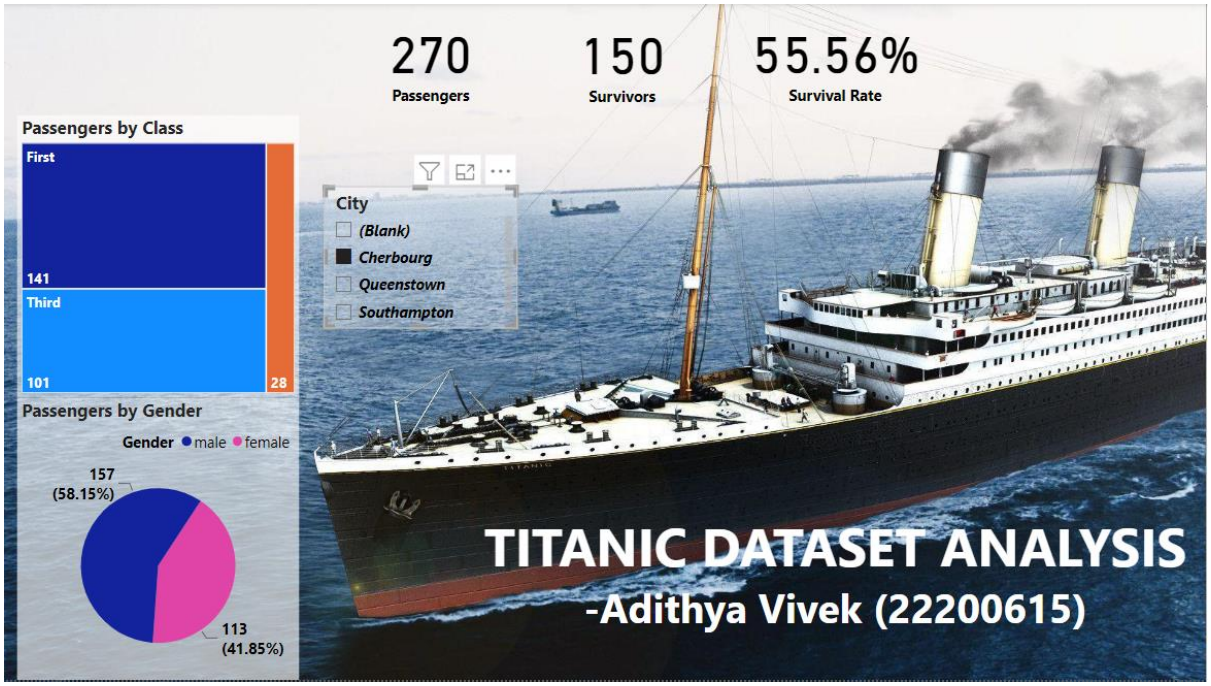
In Power BI, one dashboard has been created having 5 sheets namely- Overview, Influencers, Survival Analysis, Fare Analysis and Age Range.

Sheet 1: Overview

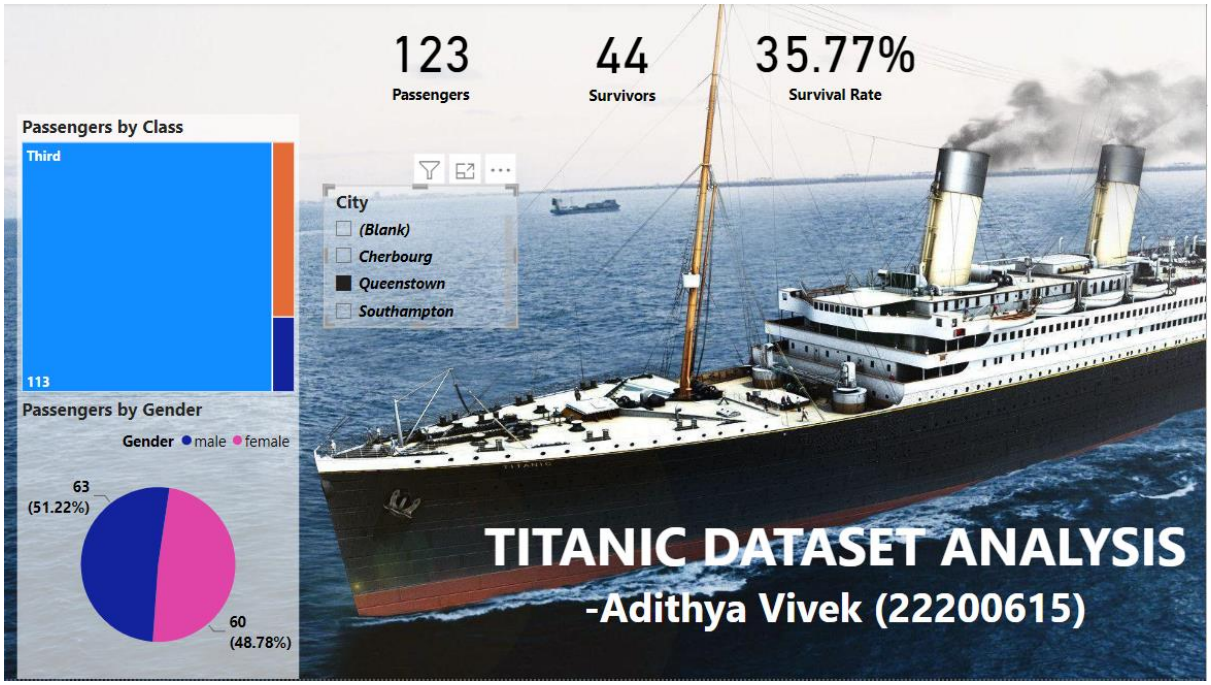
From the Titanic dataset, it could be interpreted that there were ***1309*** Passengers, out of which ***500*** survived, showing a ***38.20%*** survival rate. We have created a treemap, pie chart and a slicer in this sheet. The treemap is categorized by class. The pie chart is categorized by ***Gender***, and the field is chosen as the ***City*** in the slicer.

Key Visualizations-

- ***Third Class Passengers*** were the highest number, which stood out at ***709***.
- On Classifying the ***passengers*** by ***gender***, ***64.4%*** were ***males***, and the rest ***35.6%*** were ***females***.
- Of the 3 locations where the ***passengers*** embarked, ***passengers*** who embarked at ***Cherbourg*** had the ***highest survival rate*** of ***55.56%***. The maximum number of passengers who embarked at Cherbourg were first-class passengers, with a count of ***141***. Gender ratio of male: female is close to ***60:40***.



- ***Passengers*** who embarked from ***Queenstown*** had a survival rate of ***35.77%***, and the maximum number of passengers was third class, with ***113*** being the count. ***Gender*** ratio ***male: female*** is almost the same, ***50:50***.



- Majority of the passengers that is ***914*** of the ***1309 passengers*** embarked at ***Southampton***. ***Survival Rate*** can be seen close to ***33%***. Third class passengers is seen to be the highest with ***495*** passengers. Gender ratio of ***male: female*** is close to ***70:30***.

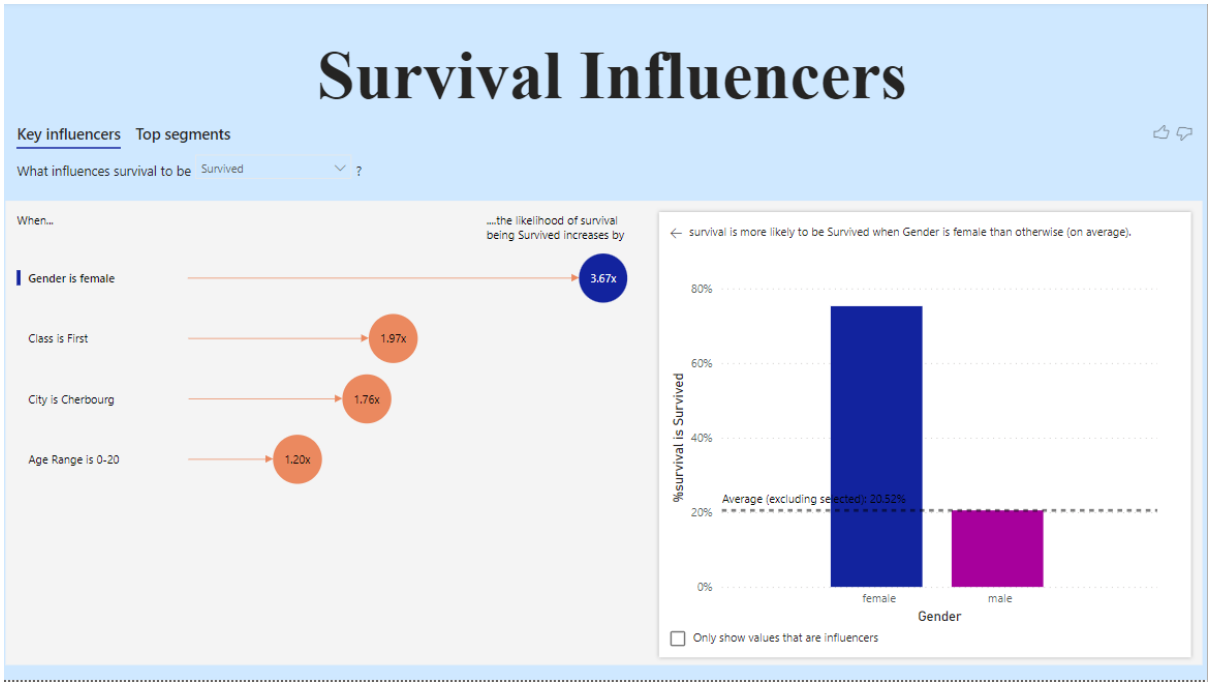
- The embarkment location is unknown for the ***two passengers***, but they were ***female*** and of ***first class***. They both ***survived***.

Sheet 2: Influencers

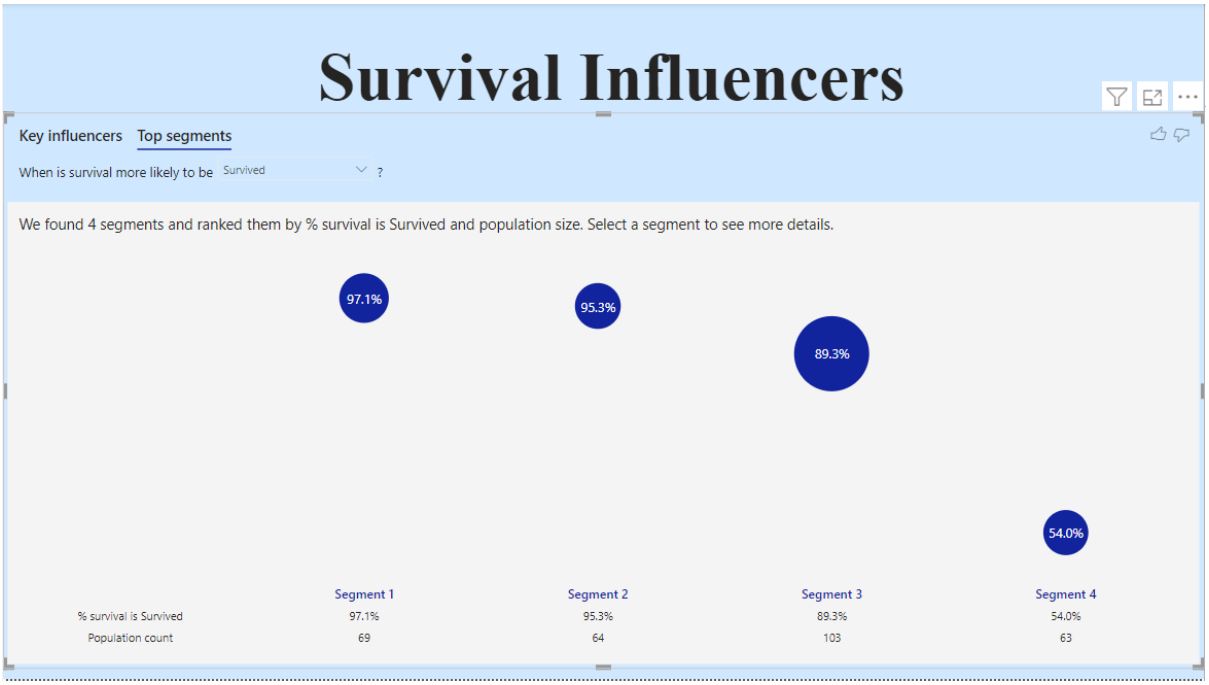
In this sheet, the ***Key influencers*** feature has been used. This feature will tell the factors on passengers either ***surviving*** or ***dying***. Through the filter option, the ***Age Range*** has been excluded for passengers whose ***age*** was not mentioned in the dataset.

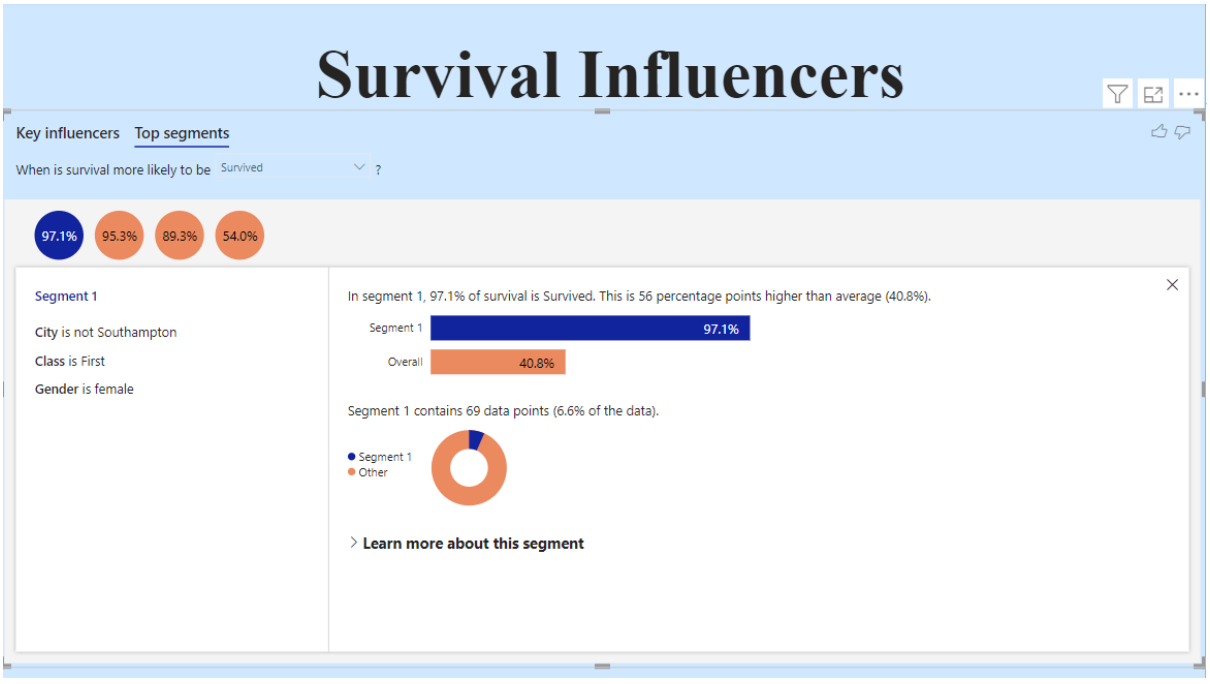
Key Visualisations-

- Factors influencing ***passengers*** surviving increase ***3.67 times*** when ***gender*** is ***female*** when class is ***first*** surviving chances increase by ***1.97 times*** when the ***City*** is ***Cherbourg***, surviving chances are ***1.76x times*** more. ***Age range*** between ***0-20*** has surviving chances increase by ***1.2x times***.

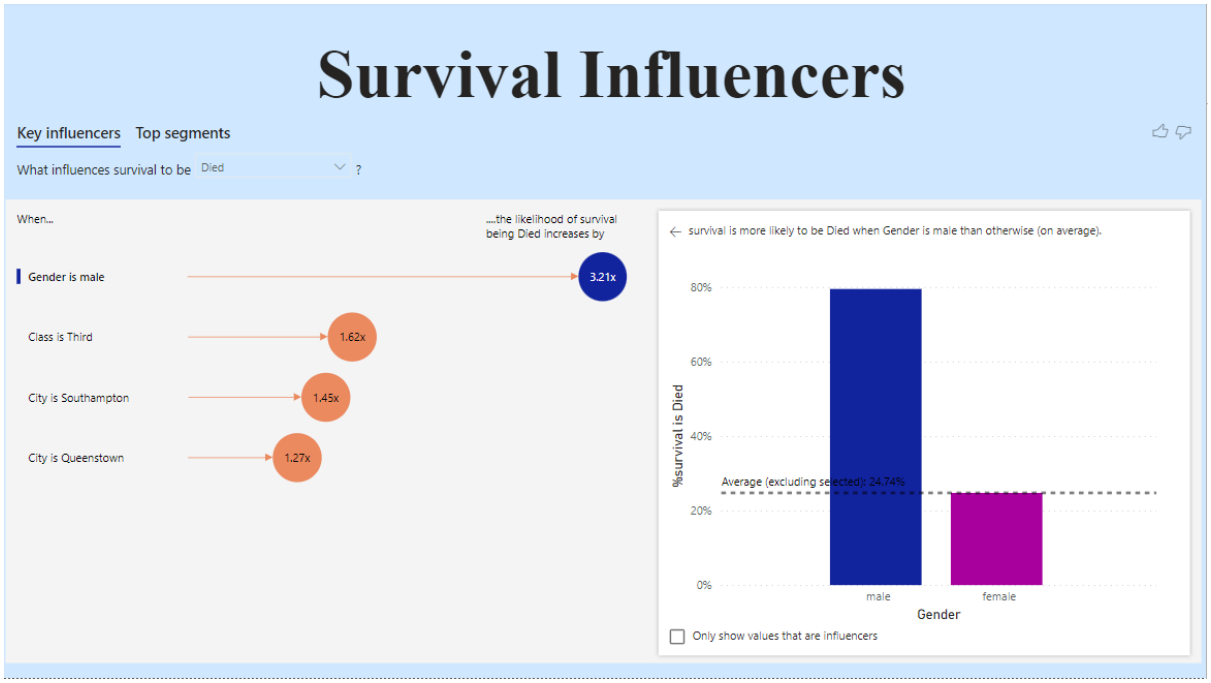


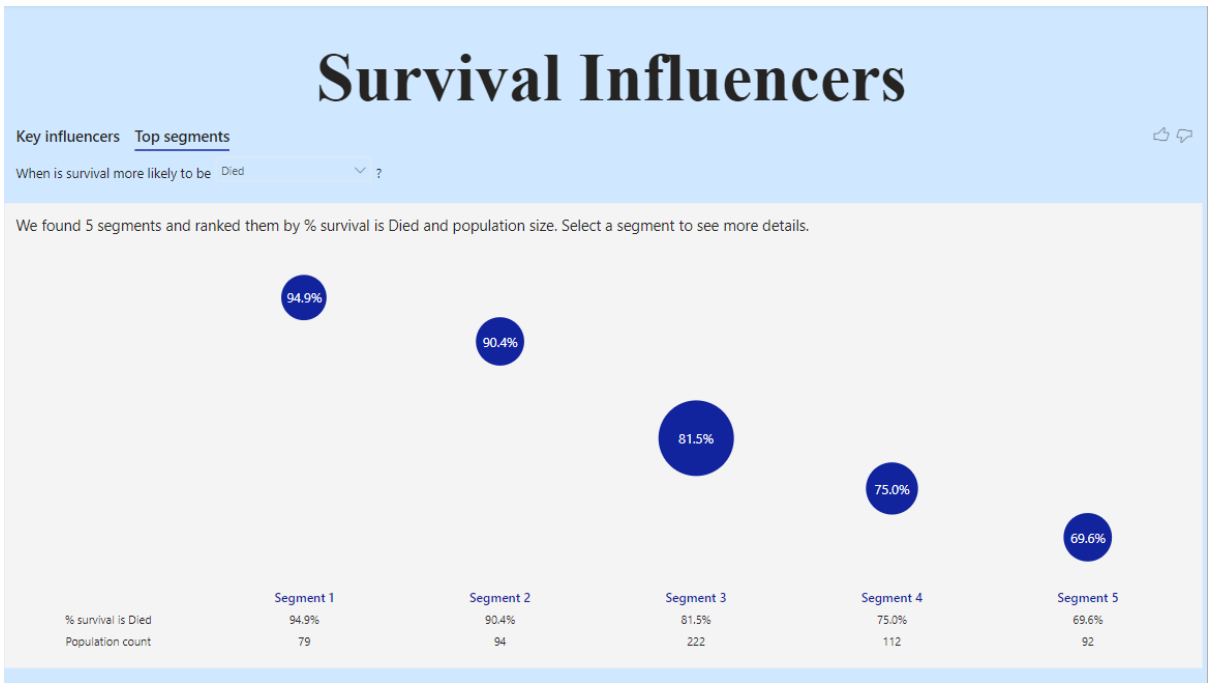
- Top Segments shows the different segments and is ranked by ***survival %*** based on the condition of ***survived*** and ***population size***.





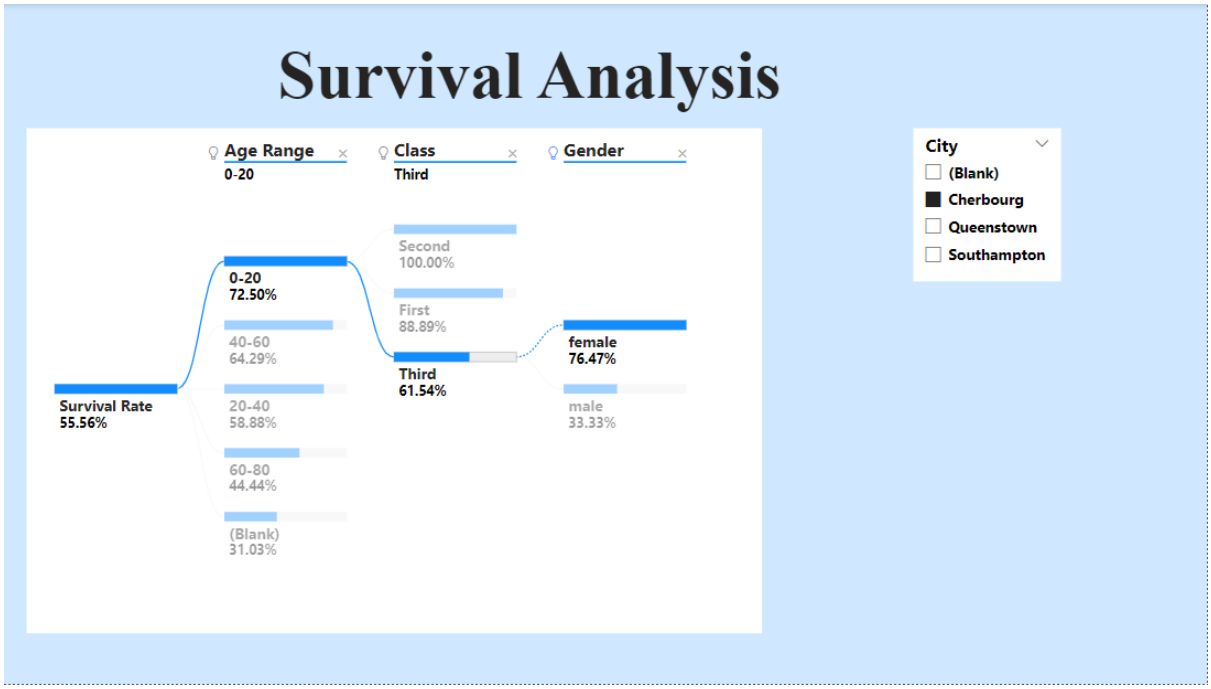
- Similarly, the factors which influence the passengers dying are shown in the screenshots as below-


Survival Influencers
Key influencers
Top segments
What influences survival to be Died ?
When...
....the likelihood of survival being Died increases by
Gender is male
3.21x
Class is Third
1.62x
City is Southampton
1.45x
City is Queenstown
1.27x
survival is more likely to be Died when Gender is male than otherwise (on average).
80%
60%
40%
20%
0%
%survival is Died
Average (excluding selected): 24.74%
male
female
Gender
Only show values that are influencers


Survival Influencers
Key influencers
Top segments
When is survival more likely to be Died ?
We found 5 segments and ranked them by % survival is Died and population size. Select a segment to see more details.
94.9%
90.4%
81.5%
75.0%
69.6%
Segment 1
Segment 2
Segment 3
Segment 4
Segment 5
% survival is Died
94.9%
90.4%
81.5%
75.0%
69.6%
Population count
79
94
222
112
92

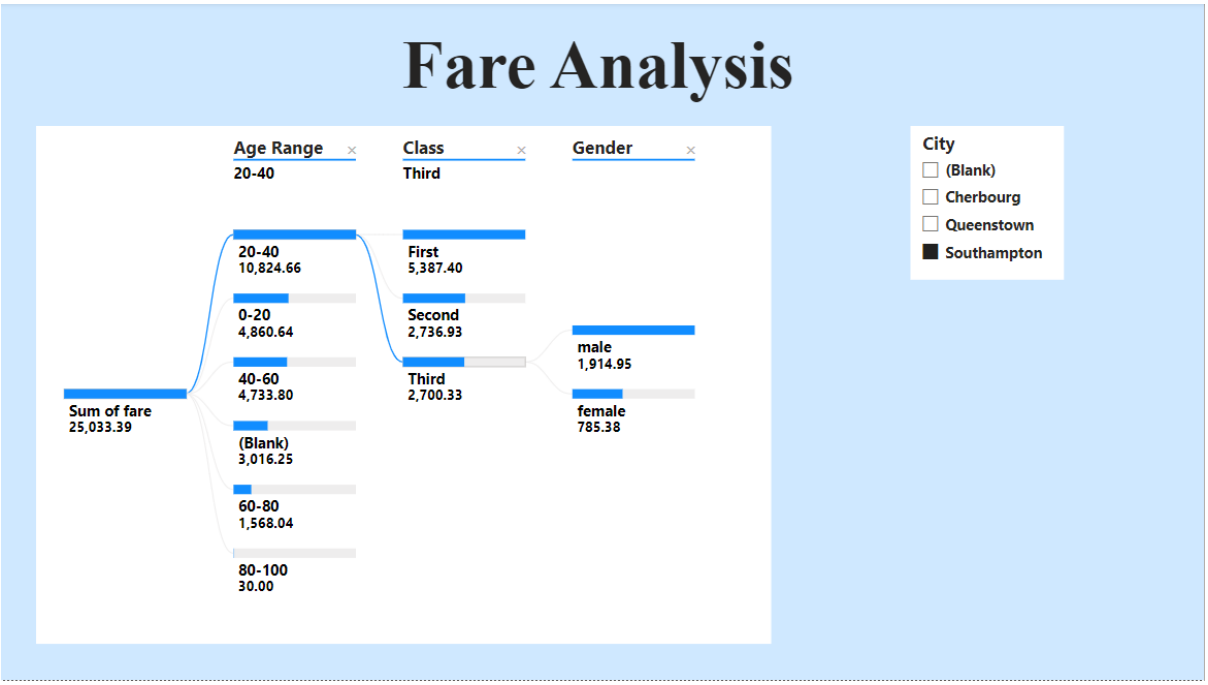
Sheet 3: Survival Analysis

A ***Decomposition tree*** for analysing ***Survival Rate*** and a slicer for ***City*** have been used in this sheet. Based on the ***Survival Rate***, it is classified in terms of ***Age Range***, ***Class*** and ***Gender***. One can browse and click on each node to get insights into the ***Survival Rate*** depending on what ***Age Range***, ***Class*** and ***Gender*** is chosen. The ***Survival rate*** can further be detailed in terms of ***City*** when a particular ***City*** is chosen in the slicer. Given Below is the screenshot of one scenario-



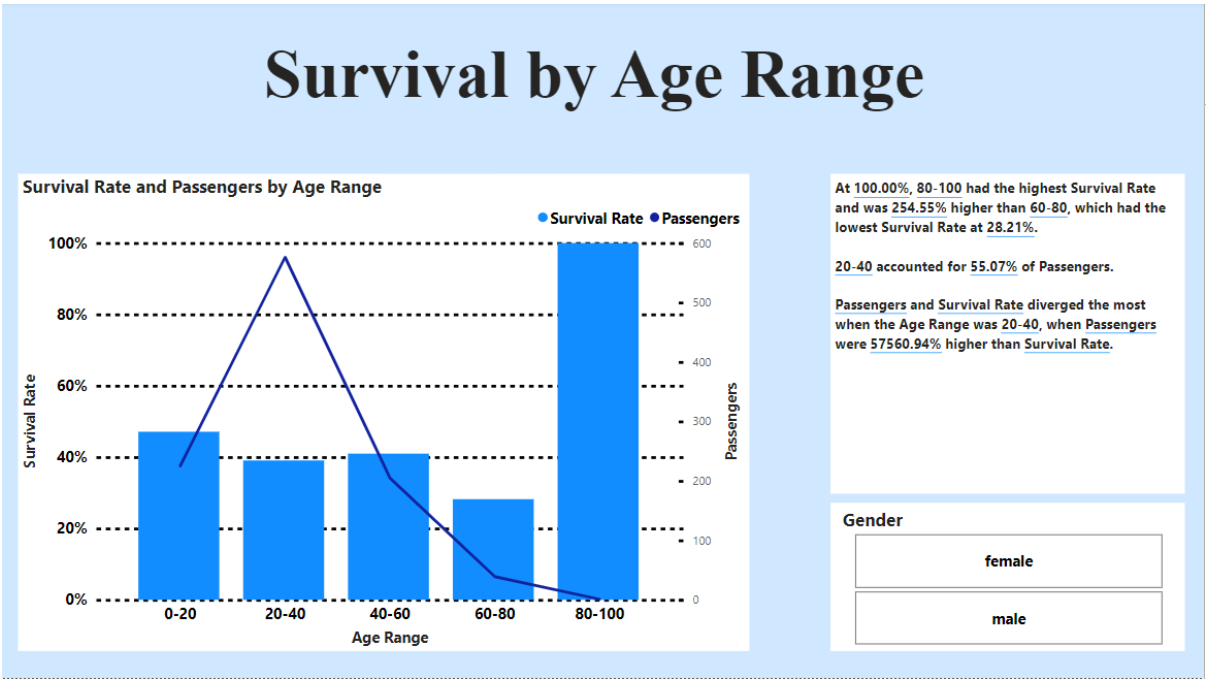
Sheet 4: Fare Analysis

In the Fare Analysis Sheet, a ***decomposition tree*** shows the ***Sum of Fare*** collected from all the ***passengers***. A slicer has been created, which allows the selection of the ***City***. There is a breakdown of the ***Sum of Fare*** in terms of ***Age Range***, ***Class*** and ***Gender***. One can infer the distribution of the ***Sum of Fare*** by the ***Age Range*** of ***passengers***, ***class***, and ***Gender***. Each node can be clicked upon, and data can be inferred according to the node selected in each ***Age Range***, ***Class*** and ***Gender***. Given Below is the screenshot of one such scenario-



Sheet 5: Age Range

In the ***Age Range*** sheet, a line and clustered column chart have been used to simultaneously infer the comparison of ***Survival Rate*** and ***Passengers*** by ***Age Range***. Through Filter, ***Age Range*** which are ***null***, have been unchecked and not used. A slicer has been used for the ***Gender*** if we want to segregate based on either male or female. Please find the screenshot given below for the sheet-



Conclusion

Finally, the Power BI visualizations used in this report have given insightful information about the Titanic dataset. We have determined the main variables that affected survival rates using exploratory data analysis. The decomposition tree demonstrated the vital influence of passenger class and gender, while the line and clustered column chart illustrated patterns in survival based on age groups. These results lay the groundwork for more investigation and predictive modelling, which helps us comprehend the elements that affected the Titanic survivor composition.