University College Dublin

Michael Smurfit Graduate Business School

MSc. Business Analytics

MIS41430: Mastering Big Data with Citizen Development

Title: **Assignment 2**

Date of Submission: 10th July 2023

Student name – **Pooja Padmacharan Dash**

Student number - **22200862**

Declaration: This document is submitted in the form of a report to UCD Smurfit Business School in part fulfilment of the requirements for the degree of MSc Business Analytics.

Introduction

The Titanic dataset provides information about the passengers aboard the RMS Titanic, which sank in 1912. It includes details such as passenger class, name, age, gender, family relationships, ticket information, fare price, cabin number, port of embarkation, survival status, and more. The dataset is used to analyze factors influencing passenger survival and gain insights into demographics and socio-economic dynamics of the time.

Dataset Description

The Titanic dataset consists of information about the passengers aboard the RMS Titanic, including various attributes that can be used to analyze factors affecting survival. It contains a total of 1309 rows (observations) and 14 columns (variables).

- **pclass:** Passenger class (integer: 1, 2, or 3). This column represents the socio-economic status of the passengers and can be used to study the correlation between class and survival rates.
- **survived**: Survival status (text: "yes" or "no"). This column represents the target variable for survival analysis and is essential for studying the factors affecting survival rates.
- **name**: Name of the passenger (text). While names may not directly impact survival, they can be used for identification purposes or for further analysis, such as extracting titles (Mr., Mrs., etc.).
- **sex:** Gender of the passenger (text: "male" or "female"). Gender is often considered a significant factor in survival analysis, as women were given priority during the evacuation.
- **age:** Age of the passenger (numeric). Age can play a role in survival, as there might have been priority given to certain age groups during rescue operations.
- **sibsp**: Number of siblings/spouses aboard (integer). This column indicates the presence of family members on board, which could impact survival probabilities and evacuation decisions.
- **parch**: Number of parents/children aboard (integer). Similar to SibSp, the presence of parents or children could influence survival chances.
- **ticket:** Ticket number (text). Ticket numbers can be used for further analysis or to identify patterns related to survival.
- **fare:** Fare price (numeric). Fare could be correlated with passenger class and might reflect socio-economic status, which could impact survival.
- **cabin**: Cabin number (text). Cabin information may be useful for studying the location of passengers on the ship and its proximity to lifeboats. However, this column has a significant number of missing values.
- **embarked:** Port of embarkation (text: "C" for Cherbourg, "Q" for Queenstown, "S" for Southampton). The embarkation port may have some correlation with socio-economic status and potentially affect survival rates.
- **body:** Dead body number (numeric). This column indicates the identification number of deceased passengers. While relevant for understanding casualties, it is not applicable to survival analysis.
- **boat:** Lifeboat number (integer). This number can be used to identify the survived passenger on the ship.
- **homeDest:** Home/destination of the passenger (text). Home or destination information may not directly impact survival but could provide insights into the demographics of the passengers.

Distribution of Missing Data Points and Their Potential Impact

The Titanic dataset contains missing data points, primarily in the "Age," "Cabin," and "Embarked" columns. The "Age" column has missing values, which could affect age-related analysis and predictions. From the age column 20% of data are unknown.

The "Cabin" column has a substantial number of missing values, which limits its usability for studying the location-based factors of survival. The total of 77% of data from the cabin column are unknown/Blank

The "Embarked" column has 2 missing values, which can be handled during data pre-processing.

Data Exploration and Visualization

After importing the CSV file into Power BI and accessing the Power Query Editor for data cleaning, I performed the following cleanup steps:

Added a conditional column named "**Survived_Flag**" to replace the values 1 and 0 in the "Survived" column with "Yes" and "No" respectively. This makes the column more understandable and user-friendly.

Added a conditional column named "**Sibling_Spouse**" as well as Parent _child to replace the values 0 as "No" and with value as "Yes" . This makes the column more understandable and user-friendly.

Added another conditional column named "**Passenger_Class"** to identify the passenger class as "First Class," "Second Class," or "Third Class," replacing the values 1, 2, and 3 in the "Pclass" column. This improves clarity and ease of interpretation.

Added a conditional column named "**Embark_Port"** to replace the values "S," "C," and "Q" in the "Embarked" column with their corresponding names "Southampton," "Cherbourg," and "Queenstown" respectively. This ensures clarity and avoids confusion.

Added a conditional column **Age_category** to categorize different ages into specific age groups such as 0-18 (age <= 18), 19-30 (age <= 30), 31-50 (age <= 50), and "Adult" (age > 50). I used a nested IF formula to handle missing values and categorized them as "Unknown." This grouping simplifies analysis and provides meaningful insights. It is worth noting that I rounded down the age values since the year is not completed.

Checked and corrected any wrongly labelled data types for each column to ensure accuracy and consistency. Once the data cleaning was complete, I closed the Power Query Editor and applied the changes in Power BI.

Moving to the Power BI interface, I performed additional tasks:

1. Created two measures: "Total Died" and "Total Survived" using the CALCULATE formula. These measures allow for distinct representation of the number of passengers who died and survived on the ship, enabling clear visualization of the survival outcomes.
2. Created a measure to count the total number of passengers present on the ship using the COUNTROWS formula. This measure provides an accurate count of passengers, which is useful for various calculations and comparisons.

By implementing these steps, the data in Power BI is now properly categorized, cleaned, and ready for analysis and visualization.

Visualisation Insights

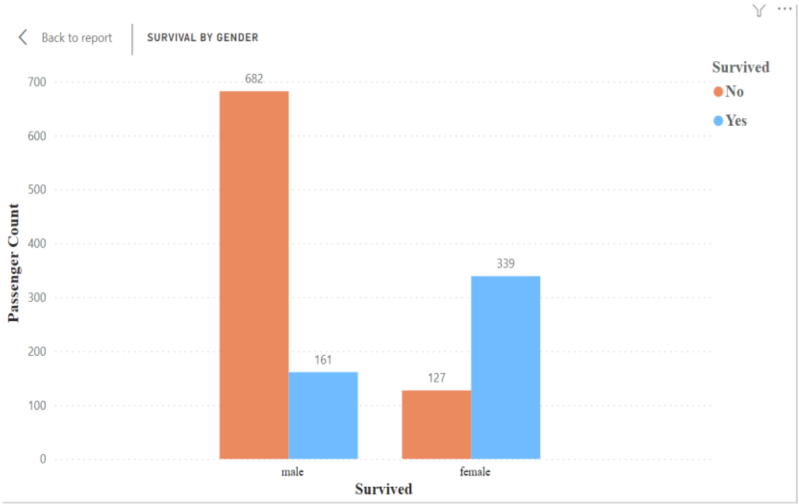
The Titanic dataset comprises records of 1309 passengers, with 64% being male and 36% female. The goal of the analysis is to identify which variables had the greatest impact on survival rates. This analysis utilizes the available characteristics in the dataset to determine how well each characteristic correlates with survival.

Out of the total 1309 passengers, only 500 (38%) survived while 809 (62%) perished. Below, I provide observations on the survival rate based on various columns in the dataset.

Survival by Gender

The survival rate of females was higher than that of males. This indicates that females were prioritized for rescue.

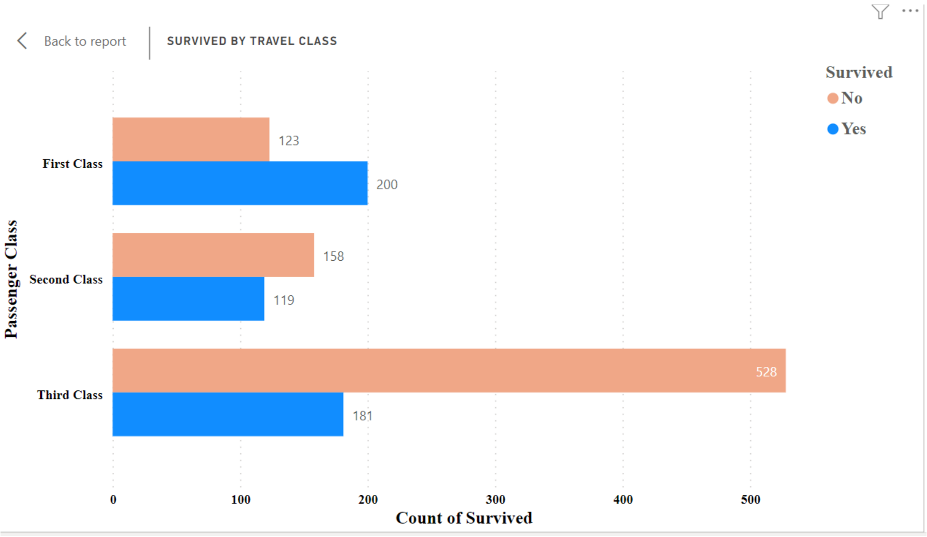
Figure 1 Passenger Count vs Survived



Survival by Travel Class

Passengers in first class had the highest survival rate at 200 (40%), followed by third class at 181 (36%) and second class at 119 (24%). This reflects the fact that passengers who paid higher fares had a higher chance of survival.

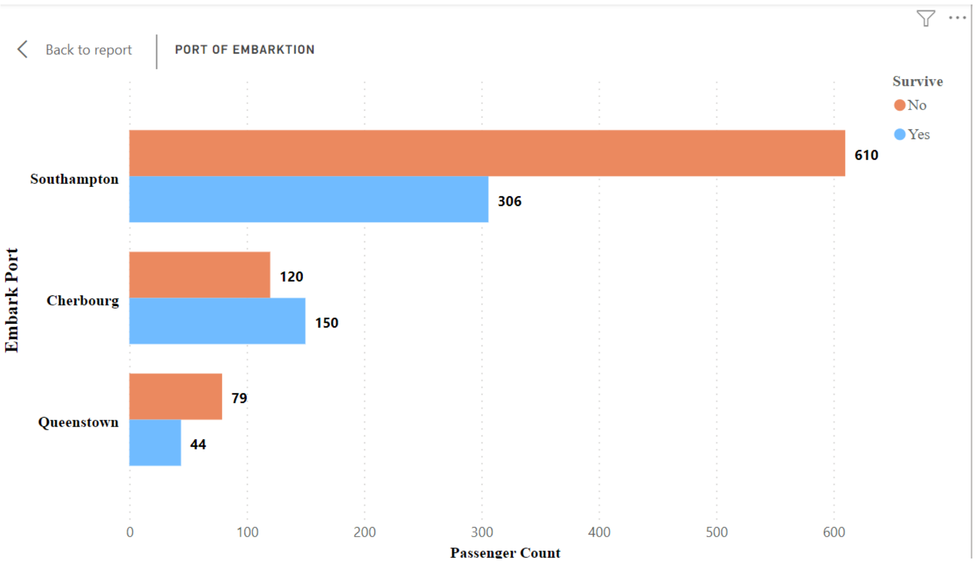
Figure 2 Passenger Class vs Count of Survived



Survival by Port of Embarkation

Approximately 70% of passengers embarked from Southampton, 20% from Cherbourg, and 10% from Queenstown. Passengers who boarded from Cherbourg had the highest survival rate (55%), followed by Queenstown (35%) and Southampton (33%).

Figure 3 Embark Port vs Passenger Count

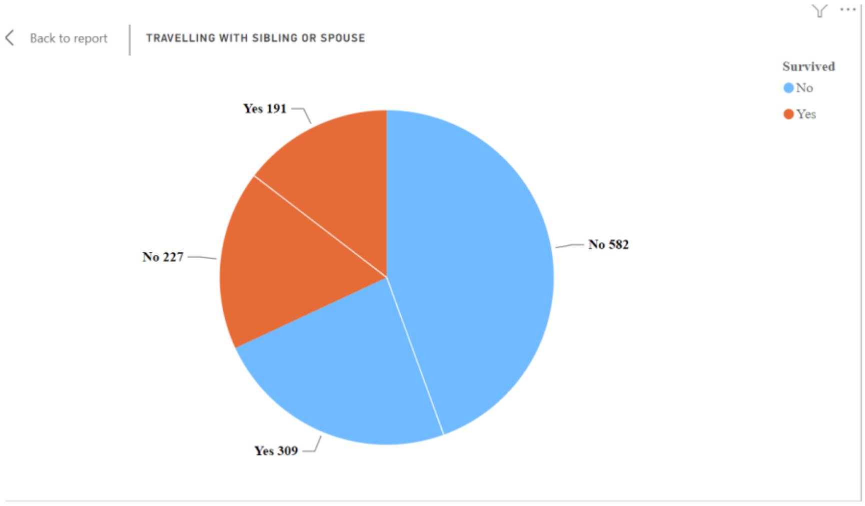


Note: There were 2 missing values in the dataset for the "Embarked" column. I imputed the missing values with the most common port of embarkation (Southampton) based on the assumption that the missing values were likely from the most common category.

Travelling with Sibling or Spouse:

The below pie chart shows that the passengers traveling with siblings, or a spouse had a higher survival rate than those traveling alone.

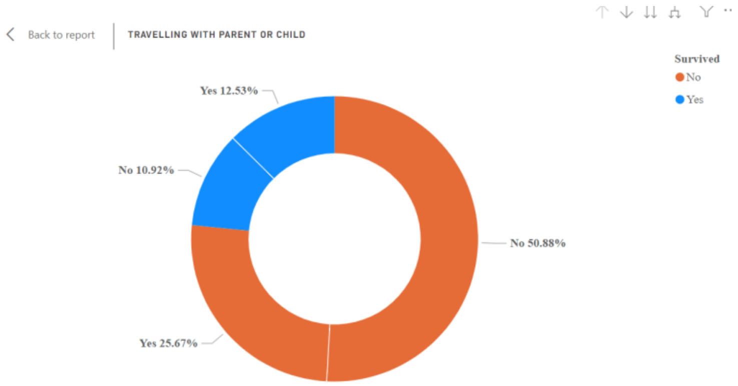
Figure 4 Survival rate of passengers travelling with siblings or spouse



Travelling with Parent or Child

The donut chart shows that larger proportion of passengers who did not survive travelled alone compared to those who were accompanied by parents or children.

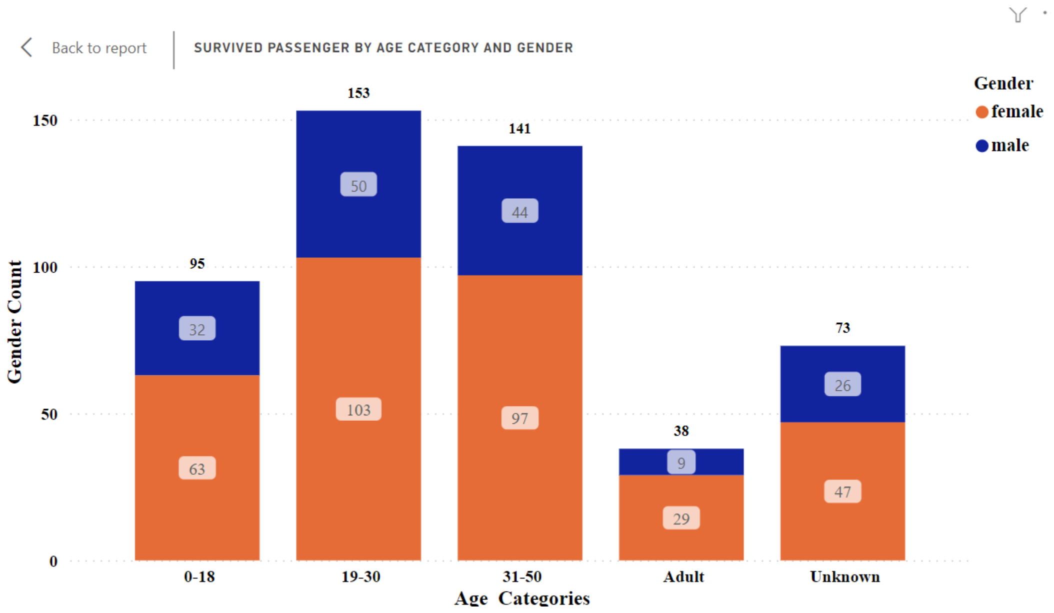
Figure 5 Survival rate of passengers travelling alone and accompanied



Survived Passengers by Age Category and Gender

The visual represent the survival rates by age category relation to identify the age range that was most affected. The age categories were defined as follows: 0-18 (age <= 18), 19-30 (age <= 30), 31-50 (age <= 50), and Adult (age > 50). Approximately 20% of the data in the age column were missing. I created an "unknown" category to represent the missing values and analysed the survival rate within this category.

Figure 6 Survival rate by Age category and Gender



Among the known age groups, the highest number of survivors belonged to the 19-30 age group, with 153 passengers, of which 103 were female and 50 were male. The second-highest number of survivors were in the 31-50 age group, with 141 passengers, of which 97 were female and 44 were male. Within the "unknown" age category, a total of 73 passengers survived, with 47 being female and 26 being male.

Summary:

During the analysis of the Titanic disaster dataset, several characteristics were observed that correlated with the survival rate of the passengers.

Gender: Female passengers were prioritized and had a higher survival rate.

Family: Traveling with family members increased the chances of survival.

Age: The 19-30 age group had the highest survival rate.

Class: First-class passengers had a higher survival rate compared to other classes.

Embarkation: Passengers boarding from Southampton had the lowest survival rate.These factors played a significant role in determining the passengers' survival outcomes during the Titanic disaster.