A logo of a university

Description automatically generated

MIS41430 Mastering Big Data

|  |  |
| --- | --- |
| **Student Name and Number** | Leena Mary Varghese (23200412) |
| **Assessment Title** | Mastering Big Data Assignment 2 |
| **Module Code** | MIS41430 |
| **Module Title** | Mastering Big Data |
| **Module Coordinator** | Prof. Martin Perry |
| **Date Submitted** | 12-07-2024 |

**TITANIC DATASET ANALYSIS REPORT**

**ABSRACT**

The Titanic disaster in April 1912 is a significant historical event that has been thoroughly examined by researchers and data analysts. This report provides a comprehensive analysis of the Titanic dataset, focusing on demographic details, ticket information, and survival status. The key findings highlight how factors such as passenger class, gender, and cabin position can affect the chances of survival. The analysis includes detailed visualizations highlighting the differences in survival rates based on factors such as socio-economic status and family dynamics.

**INTRODUCTION**

The sinking of the Titanic is one of the most tragic and well-known disasters in history. This report explores the Titanic dataset, examining the relationships between passenger characteristics and survival outcomes. Using Power BI, the study reveals patterns influenced by class, age, gender, and embarkation port. The primary goal is to gain a deeper understanding of the factors that determined passenger survival during the disaster.

**DESCRIPTION OF DATASET**

The Titanic dataset is a comprehensive collection of data points related to the passengers aboard Titanic.

Number of Rows: The dataset contains 1,309 rows, representing passengers on the Titanic.

Number of Columns: The dataset has 14 columns, providing different types of information about each passenger.

|  |  |  |
| --- | --- | --- |
| **TABLE NAME** | **DESCRIPTION** | **DATA TYPE** |
| Survived | Survival status (0 = No, 1 = Yes). Indicates whether the passenger survived the disaster. | Integer |
| |  | | --- | | Pclass | | |  | | --- | | Passenger class (1 = 1st class, 2 = 2nd class, 3 = 3rd class). Represents the socio-economic status of the passenger. | | |  | | --- | | Integer | |
| name | Full Name of the passengers | String |
| sex | Gender of the passenger (male or female) | String |
| age | Age of the passenger in years. Some entries may contain fractional values for infants. | Float |
| sibsp | Number of siblings/spouses the passenger had aboard | Integer |
| parch | Number of parents/children the passenger had aboard | Integer |
| ticket | Ticket number of the passenger | String |
| fare | Fare or price paid by the passenger for the ticket | Float |
| cabin | Cabin number of the passenger | String |
| embarked | Represents the port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton) | String |
| boat | The lifeboat number in which the passenger was rescued. (if they survived) | String |
| body | The body number, if the passenger's body was recovered. | Float |
| home.dest | Home/destination of the passenger | String |
|  |  |  |

**DATA CLEANING**

To ensure the dataset was ready for analysis, several data cleaning steps were undertaken. The following procedures were performed:

**Handling Missing Values**:

* **Age**: Missing values in the Age column were filled with the median age of the passengers. This approach was chosen to maintain the dataset's overall age distribution.
* **Embarked**: Missing values in the Embarked column were filled with the mode, which is 'S' (Southampton), as it was the most common embarkation point.

**Creating New Features**:

* **Family Size**: Created a new feature by summing the SibSp (number of siblings/spouses) and Parch (number of parents/children) columns, providing a holistic view of the family size for each passenger
* **Embarked:** Converted the codes to their respective full names for better clarity and readability in the analysis.  
   C was converted to Cherbourg

Q was converted to Queenstown  
 S was converted to Southampton

* **Passengers**: The values in the Survived column were modified to make the data more interpretable. The original Survived column contained binary values (0 = No, 1 = Yes) indicating whether a passenger survived the Titanic disaster. These values were changed into strings, 1 to ‘survived’ and 0 to ‘died’ more descriptive labels.
* **Home and Destination**: Information about passengers' home and destination, was split into two separate columns: Home and Destination.

**DATA MODELLING**

In Power BI, I structured the data model with a main fact table and two supporting dimension tables:  
  
**Main Table:**

Titanic main Dataset: Contains key metrics and attributes such as survived, pclass, fare, age, boat, body, cabin, destination, embarked, family size, first name, and other passenger details.

**Dimension Tables:**

Coordinates: Includes embarkation ports and their respective latitude and longitude.  
Classes: Contains class details, including descriptions of each class and the corresponding pclass values.

**Relationships:**

Titanic Main dataset is linked to Coordinates via the embarked column.  
Titanic Main datset is linked to Classes via the pclass column.

**DATA ANALYSIS**

**1. Overview**

A screenshot of a map

Description automatically generated

The overview offers a comprehensive summary of the key metrics of the Titanic dataset.

* The total number of Titanic passengers is displayed, showing a higher number of male passengers compared to females.
* Additionally, the survival metrics are highlighted. The overall survival rate was 38.2%, indicating that a majority of the passengers did not survive the disaster. The breakdown of survivors by gender is also illustrated using gauges.It highlights a significantly higher survival rate among female passengers (about 68%) compared to male passengers (about 32%).
* The map visualizes the geographic distribution of passengers based on their port of embarkation. The map provides insight into the number of passengers who boarded the Titanic at each port: Cherbourg, Queenstown, and Southampton.

**2. Survival Analysis**

A screenshot of a computer screen

Description automatically generated

This page offers a deeper understanding of the factors influencing passenger survival rates.

* The pie charts show the proportion of passengers who survived versus those who did not. Also, the survival rates among male and female passengers. Female passengers had a much higher survival rate (72.75%) compared to male passengers (19.1%). This reflects the "women and children first" protocol that was followed during the evacuation.
* The bar chart shows the survival status categorized by cabin. Passengers in certain cabins, particularly those located in first class, had higher survival rates. For example, cabins such as C23, C25, and C27 had higher survival counts. This shows that passengers in higher-class cabins had better access to lifeboats or received priority during the evacuation. In contrast, third-class passengers, located at the bottom of the ship, faced lower survival rates as the water flooded from the bottom up.
* The line chart illustrates the survival rate by age. Younger passengers, particularly children aged 0-10, had higher survival rates. The survival rate decreases with age, particularly for passengers aged 20-40, suggesting that the evacuation likely prioritized younger passengers.

**3. Passenger Class Analysis**

A screenshot of a computer

Description automatically generated

Passenger class analysis provides a detailed examination of passengers' survival rates based on their class and gender. This analysis highlights the significant impact of socio-economic status on survival during the Titanic disaster.  
Interactive filters and buttons allow users to view data segmented by passenger class (1st, 2nd, 3rd) and survival status (died, survived). This setup allows for an in-depth analysis of how class impacted survival rates across different categories.

* A donut chart shows that the majority of male (58.48%) and female (46.35%) passengers were in third class, and they did not survive, showing the preferential treatment received by wealthier passengers during the evacuation.
* The analysis reveals that female passengers, particularly those in first class, had a significantly higher chance of survival. In contrast, males, especially those in the third class, had the lowest survival rates. This highlights the "women and children first" evacuation protocol, which prioritized women and those in higher socio-economic classes.
* The data highlights the significant socio-economic differences on the Titanic. First-class passengers had greater access to lifeboats and evacuation resources, which led to higher survival rates.

**4. Key Influencers Analysis**

A screenshot of a computer

Description automatically generated

This analysis provides a detailed understanding of the variables that had the greatest impact on the chances of survival.

* **Gender**: Gender was the most influential factor in determining survival. Female passengers were 3.81 times more likely to survive compared to male passengers.
* **Passenger Class**: Passenger class was another crucial factor. Passengers in the first class were 2.04 times more likely to survive compared to those in other classes.
* **Family Size**: Passengers with a family size between 1 and 4 were 1.93 times more likely to survive compared to those with larger or no families.
* **Embarkation Port**: Passengers who embarked from Cherbourg were 1.65 times more likely to survive compared to those who embarked from other ports.

**5. Analysis of Survival Rates by Location (Embarkation Port and Hometown)**

A screenshot of a computer

Description automatically generated

**6. Analysis of Passenger Class and Family Size**

A screenshot of a computer

Description automatically generated

1. A Sankey diagram that shows the flow of passengers from different classes (1st, 2nd, 3rd) to their assigned lifeboats. Passengers in the first class had better access to lifeboats; second-class passengers had fewer but still noticeable pathways; and third-class passengers had the most limited access, reflecting their lower survival rates.
2. A bar chart displaying the number of passengers who survived and died, categorized by family size.

**CONCLUSION**

This analysis of the Titanic dataset reveals significant differences in survival rates based on gender, socio-economic status, family size, and embarkation port. Female and first-class passengers had higher survival rates, highlighting the preferential treatment during the evacuation. Larger family sizes and third-class status were associated with lower survival rates, emphasizing the impact of socio-economic factors on survival outcomes. These findings highlight the importance of fair safety measures and efficient evacuation protocols in enhancing survival rates in maritime disasters.