Project 1 - Exploring Titanic Database

**PART 1**

Understanding the Business Context

**What are these data for?**

This titanic dataset contains 12 data fields which are PassengerId, Survived, Pclass, Name, Sex, Age, Sibsp, Parch, Ticket, Fare, Cabin and Embarked. These data fields will be used in the analysis to determine which factors do affect the survival rates of the passenger onboard.

**Why do we need this database?**

We need this database to perform the analysis in determining the factors that affected the survival rates of passengers in the sinking of the titanic tragedy. By using this database, we could evaluate the relationship and or how related each variable to the survival rates of passengers.

**Where are these data collected?**

This titanic dataset is available on Kaggle.

**PART 2**

Understanding the Technical Context

**How are these data collected?**

These data were downloaded from Kaggle. Kaggle is an online community of data scientists and machine learning practitioners where it allows users to find and publish data sets, explore and build models using the database.

**Where are the sources of these data?**

The data might be extracted from various sources such as titanic passenger record, Wikipedia and other internet sources.

**Is the data complete? Would there be missing pieces of data?**

This titanic dataset was divided by two group which is training dataset and test dataset. In this project, the training dataset will be analysed. In the training set, Kaggle provide the outcome or also known as the “ground truth” for each passenger. If there are missing values in the dataset, it will be identified in the next step.

**PART 3**

Understanding the Tables & Fields

**How many tables do we have?**

The dataset only has one table.

**What are the tables? and what are these tables representing?**

There is passengers table which contained all the details for each passenger such as passenger ID, name, age and sex.

**What are the fields in the tables? What is the meaning of each of the field?**

Table 1: The definition of each data field in the table

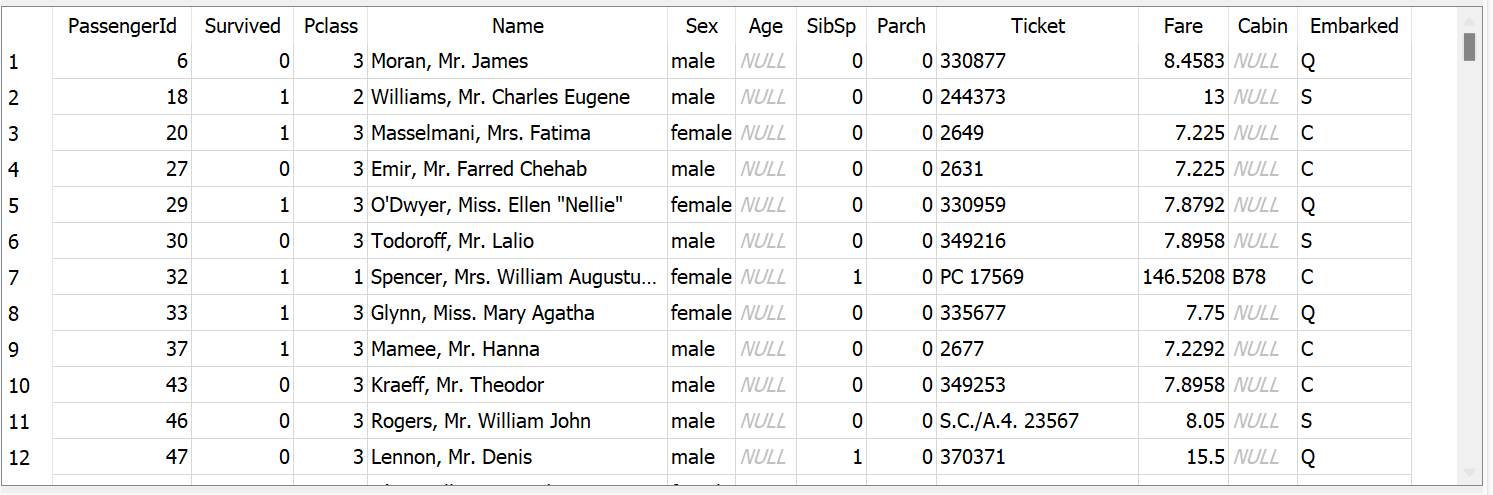
|  |  |
| --- | --- |
| Data field | Definition |
| PassengerId | Number ID of passengers |
| Survived | Survival status of passenger |
| Pclass | Ticket class |
| Name | Name of passenger |
| Sex | Sex |
| Age | Age in years |
| Sibsp | Siblings / spouses aboard the Titanic |
| Parch | Parents / children aboard the Titanic |
| Ticket | Ticket number |
| Fare | Passenger fare |
| Cabin | Cabin number |
| Embarked | Port of Embarkation |

**Are there any null data in the dataset? If there is any, can you determine which field that has null?**

Yes, there is null data in the dataset which are the Age, Cabin and Embarked. The query used to determine the null data are as follows below.

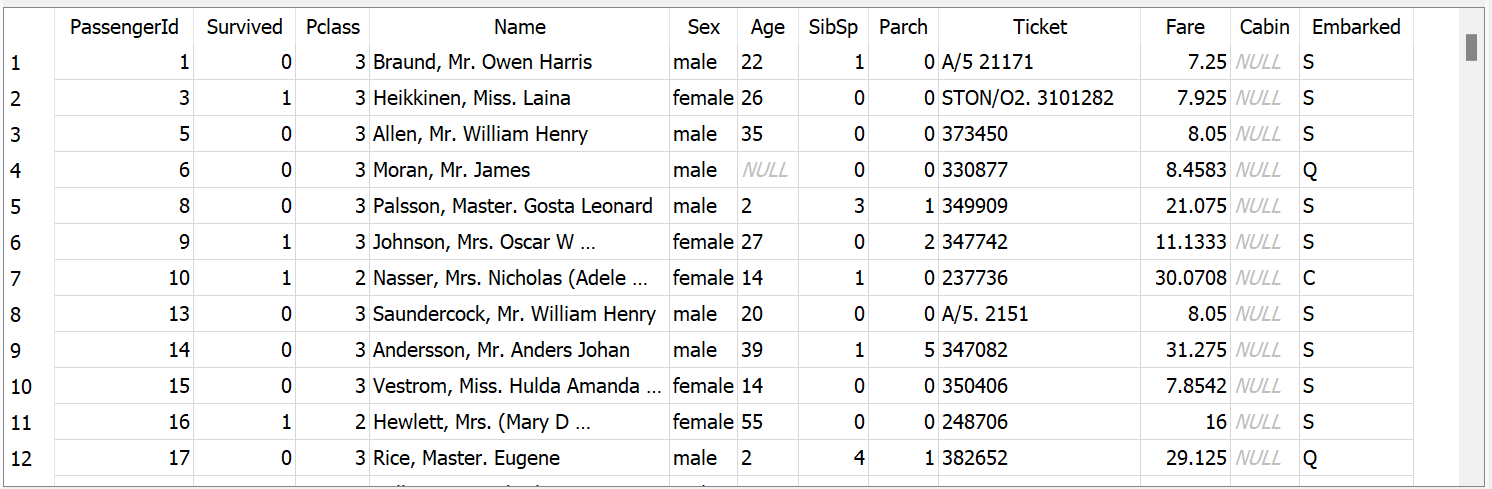
1. Query: SELECT \* FROM passengers WHERE Age is null

Result:



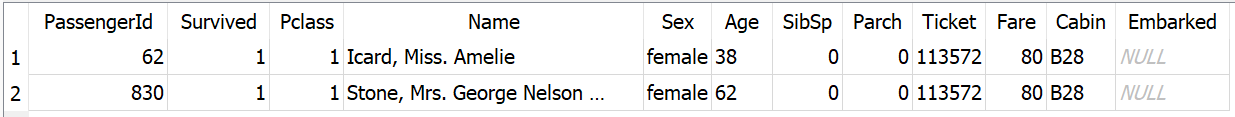
2. Query: Query: SELECT \* FROM passengers WHERE Cabin is null

Result:



3. Query: SELECT \* FROM passengers WHERE Embarked is null

Result:

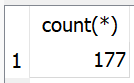


**Are there any missing data in the dataset? If there is any, can you determine the amount of missing in each field?**

Yes, there is missing data in the dataset which are the Age, Cabin and Embarked. The query used to determine the amount of missing data are as follows below.

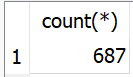
1. Query: SELECT COUNT(\*) FROM passengers WHERE Age is null or Age = 0

Result:



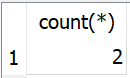
1. Query: SELECT COUNT(\*) FROM passengers WHERE Cabin is null or Cabin = 0

Result:



1. Query: SELECT COUNT(\*) FROM passengers WHERE Embarked is null or Embarked = 0

Result:

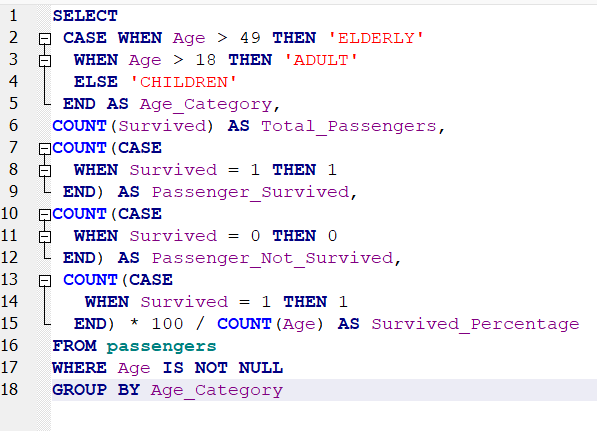


**PART 4**

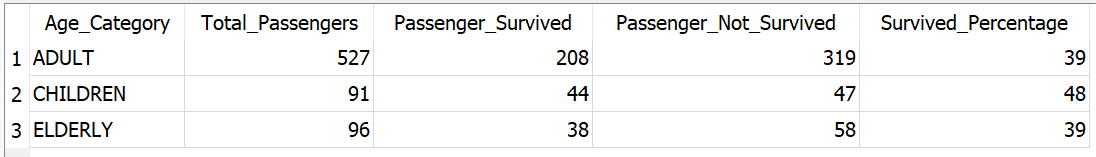
Free Exploration

**Are children and elderlies have a higher survival rate in this accident?**

Query:



Result:

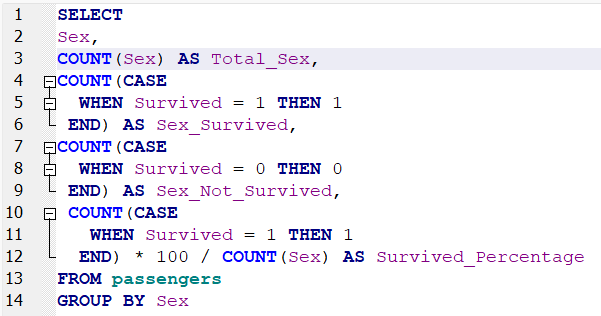


Conclusion:

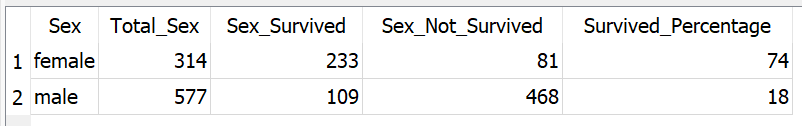
The result obtained shows that the survival percentage for adult is 39% followed by children 48% and lastly for elderly is 39%. For the survival percentage of adult and elderly, both have the same amount which is 39%. The highest survival percentage belongs to children which the passenger’s age are under 18 years old. Based on this result, we can conclude that children have the highest survival rate compared to other age group passenger which are the adult and elderly.

**Are females more likely to survive in this incident?**

Query:



Result:



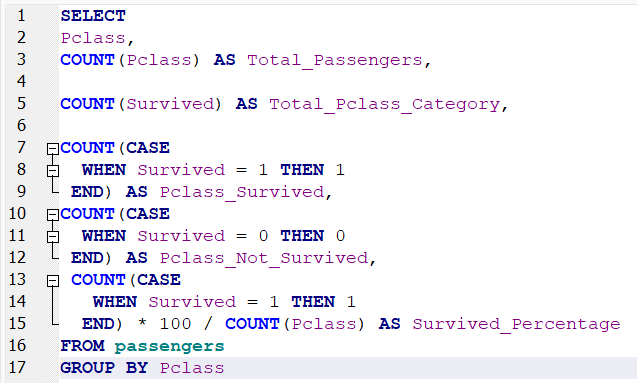
Conclusion:

The result obtained shows that the survival percentage for female is 74% and for male is 18%. Based on this result, we can conclude that female passenger has a higher survival rate compared male passenger.

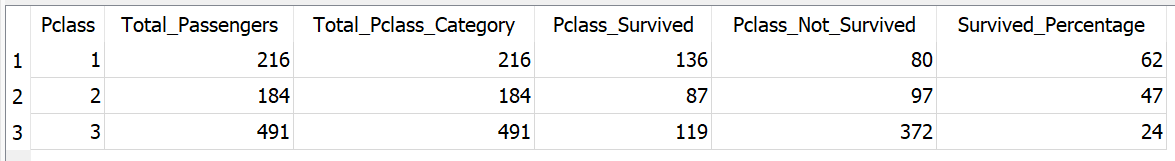
**Are rich people have a higher survival rate because they can get onboard to the**

**rescue boat sooner (like what is shown in the movie)?**

Query:



Result:

****

Conclusion:

The result obtained shows that the survival percentage for Ticket Class 1 is 62%, followed by Ticket Class 2 47% and lastly for Ticket Class 3 is 24%. The highest survival percentage belongs to passengers that bought Ticket Class 1 which the survival rate is 62%. Based on this result, we can conclude that passenger that bought Ticket Class 1 has the higher survival rate compared to other ticket class.