1. **brief introduction to the background where the data set comes from and a general justification for data science for the intended domain (there are quite a lot of effort already). It is realized that many of you may not have sufficient knowledge about the historic event. This means some very basic understanding of the history needs to be achieved through some research.**

The Titanic was a part of one of history's most famous catastrophes. The RMS Titanic was a British passenger liner that sunk in the North Atlantic Ocean after colliding with an iceberg on its way from Southampton to New York City. Hundreds of people died because of the disaster. Some people were rescued after surviving until aid arrived, while many more died hopelessly waiting for relief.

Our database contains information on those who were alive at the time of the accident, including practically every passenger but not the crew members (1, 317 passengers but in our data set we missed some data). And this dataset is well-known within the data science community because of a challenge involving whether or not the passengers survived, which the community attempted to address in various ways using various methodologies (approaches). This also aids in the comprehension of data science fundamentals.

1. **A detail description on data and all variables (may include derived ones). You may use some simple descriptive statistics to help you on this part of the work.**

The very basic information’s of the dataset variables are provided below table (Numeric, categorical, and string)

|  |  |  |
| --- | --- | --- |
| Type of feature | Feature | Feature Values |
|  |  |  |
| Numeric variables | No of Parents or Children on Board | Number of parents or child on board on the day |
|  | Age | Age in year |
|  | No of Siblings or Spouses on Board | Number sibling or spouse on board on the day |
|  | Fare | Passenger fare |
|  |  |  |
| String | Name | Name of the passenger |
|  | Cabin | Cabin number |
|  | Ticket | Ticket number |
|  |  |  |
| Categorical Variable | Passenger class | Ticket class (First, Second and Third) |
|  | Sex | Sex (string: ‘male’, ‘female’) |
|  | Port of Embarkation | Port of Embarkation (Southampton,  Cherbourg, Queenstown) |
|  |  |  |

First, we look at how the dataset look like (First five rows of the data).

A picture containing table

Description automatically generated

Graphical user interface

Description automatically generated with low confidenceExplain the variables via statistics using correlation matrix of the variable (For all the numerical variables)

The Age have negative correlation with Number of Siblings or Spouse on Board and Number of Parents or Children on Board.

The below figure shows the data types in details of the dataset.

Graphical user interface, text, application

Description automatically generated

1. **A summary on data quality assessment**.

Graphical user interface, text, application, email

Description automatically generated We may take a look at the null values in the variables to see if the data is of good quality. The variable and Boolean value combination below displays whether or not the variable has null values.

Graphical user interface

Description automatically generated with low confidenceBy checking the output, the Passenger fare, Cabin, Port of Embarkation and lifeboat are having null values, also we provided the number of null values each variable have.

1. **A comprehensive reporting on data visualization with justification on the suitability of the visual representations used together with any patterns shown.**

Below chart showing that the correlation between the Age and passenger who survived in the incident. We can see there some conclude observation.

* Babies (under the age of four) had a high survival rate.
* The oldest passenger (Age = 80) made it out alive.
* A large number of people aged 15 to 25 died.
* The majority of travelers are between the ages of 15 and 35.

Chart, histogram

Description automatically generated

Then we can visualize the data for the columns survived and passenger class to identify the correlation pattern.

* The majority of passengers were in Passenger class= Three, however the majority did not survive.
* The majority of infant travelers in Passenger class= Two and Passenger class= Three survived. The majority of passengers in Passenger class=1 One survived.

Chart, histogram

Description automatically generated

From this visualization we can conclude that the higher fare payers survived more and from Queenstown are who paid high and survived in according the output of our data.

Chart, bar chart

Description automatically generated

1. **Summarize the understandings obtained from the data set. Comments on any further possible data science operations that may exploit the data set.**

Overall, while visualizing the data from the dataset, we looked at the data kinds, difficulties that the data set had, and discovered some patterns. To describe our data, we may classify it as numerical (continuous variables such as fee, discrete variables such as age), categorical (port of embarkation, gender/sexuality, passenger class), and string (name). Also, because the ticket number contains both letters and digits, it might have a mixed value.

And, in terms of quality problems, we focused on the missing value problem, displaying all variables with values and the amount of data they included. Also, we can see that the dataset's age is displayed as a float value, but it should be integer values, therefore this missing datatype should be taken into account when exploring the dataset.

We discovered some correlations between age and surviving passengers, sex and survived passengers and passenger class, and survived passenger with sex category in the visualization section. The third passenger class was higher, and the survivor passenger was at the bottom of the class. The Queenstown top paid and most surviving in the dataset in the Passenger port of embarkation and fare payment.