**Statistics for Data Science(UE19CS203)**

**Project Report**

**PROJECT TITLE**

# Predicting the Survival of Titanic Passengers

**Team details**

|  |  |  |
| --- | --- | --- |
| NAME | SRN | SECTION |
| Manoj Kumar | PES2UG19CS222 | D |
| Minal HR | PES2UG19CS234 | D |
| Nikhil M Adyapak | PES2UG19CS257 | D |
| M Satvika | PES2UG19CS207 | D |

**Abstract:**

The RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning hours of 15 April 1912, after it collided with an iceberg during its maiden voyage from Southampton to New York City. There were an estimated 2,224 passengers and crew aboard the ship, and more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history.

With the use of data science and dataset consisting of 891 rows in the train set and 418 rows in the test set, the research attempts to determine the correlation between factors such as age, sex, passenger class, fare etc. to the chance of survival of the passengers. These factors may or may not have impacted the survival rates of the passengers.

We will be mainly focusing on the feature Pclass of the passengers in this dataset

**Introduction**

Titanic disaster is one of the most famous shipwrecks in the world history. Titanic was a British cruise liner that sank in the North Atlantic Ocean a few hours after colliding with an iceberg. While there are facts available to support the cause of the shipwreck, there are various speculations regarding the survival rate of passengers in the Titanic disaster. Over the years, data of survived as well as deceased passengers has been collected.

The prime objective of this project is to analyse Titanic disasterto determine a correlation between the survival of passengers and characteristics of the passengers.

**Dataset:**

The data we’ve selected is the passenger list who survived after the titanic disaster. The data set contains the details of total 891 passengers. We found the source for our dataset on kaggle. Each passenger’s information is given in detail such as name of the passenger, sex, age, his or her passenger class, number of siblings or spouse on board, number of parents or children aboard, cabin, ticket number, fare of the ticket and embarkation.

Few columns:

Survived- If the passenger lived or died

Pclass- passenger division

SibSp- no.of siblings/ spouse of a person onboard

Parch- no. of parents/children each passenger was touring with

Fare- Ticket fare

Embarked-Implies where the passenger mounted from

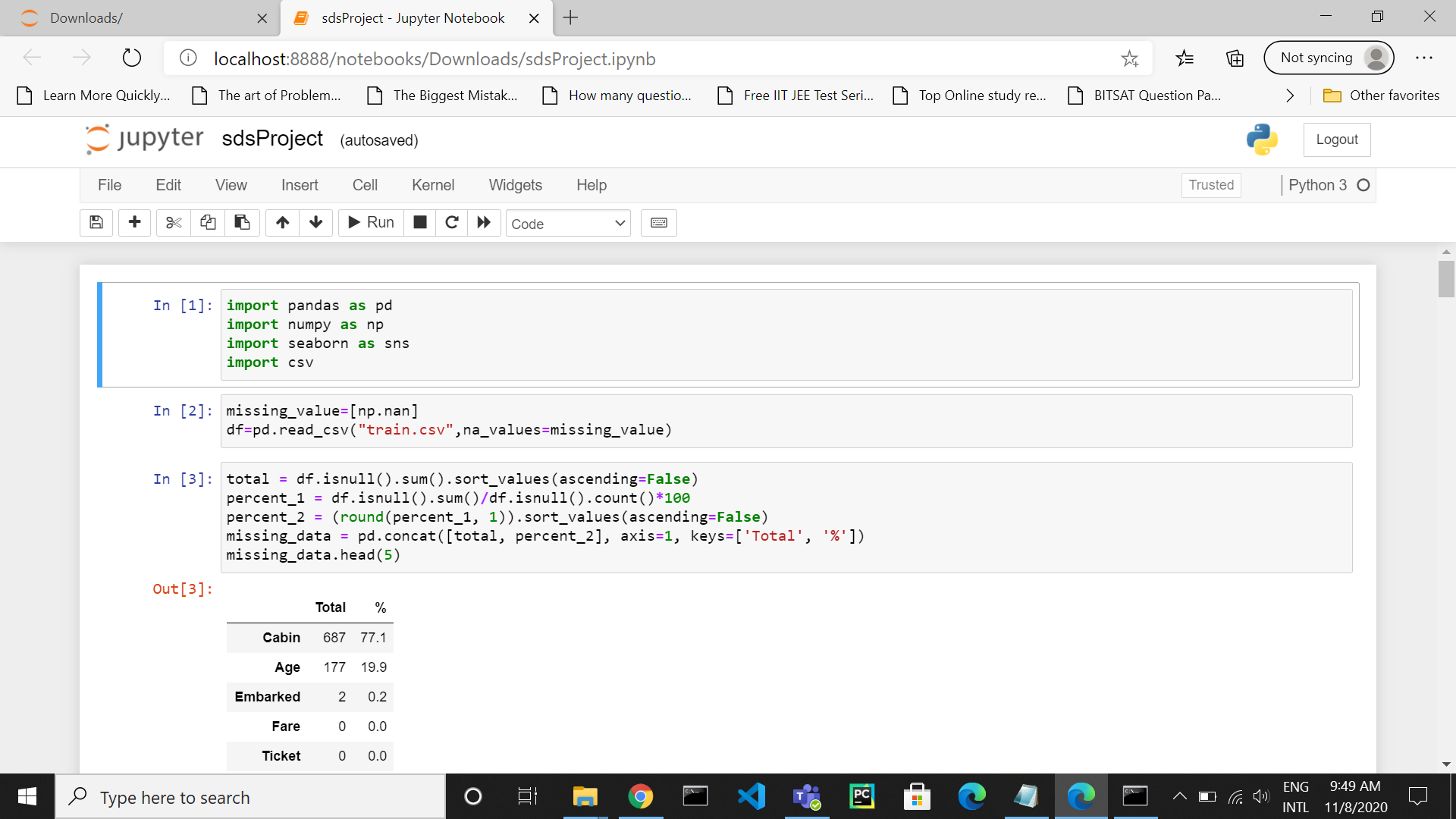
S-Southampton

C-Cherbourg

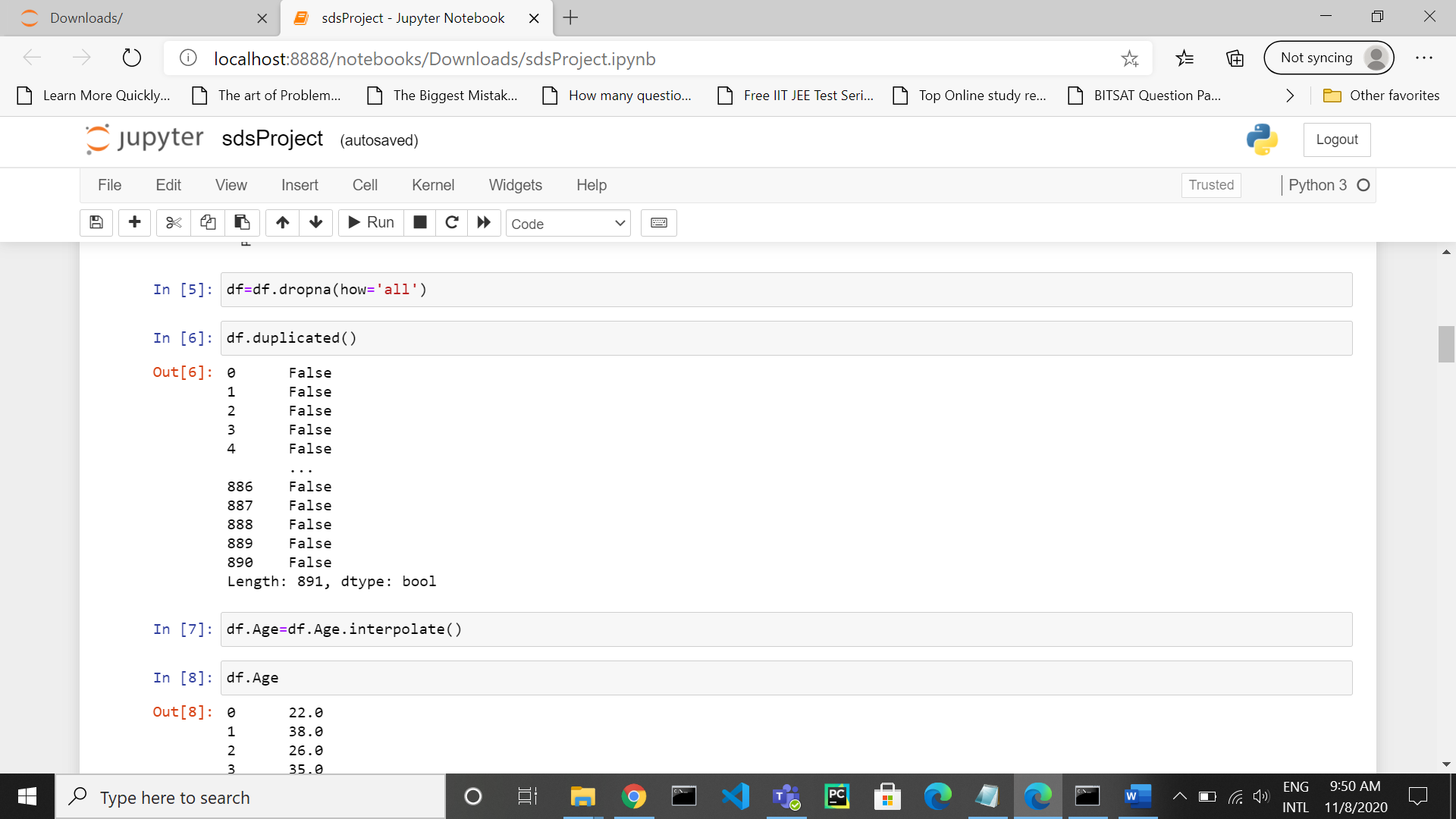
Q-Queenstown

**Preprocessing or Data Cleaning :**

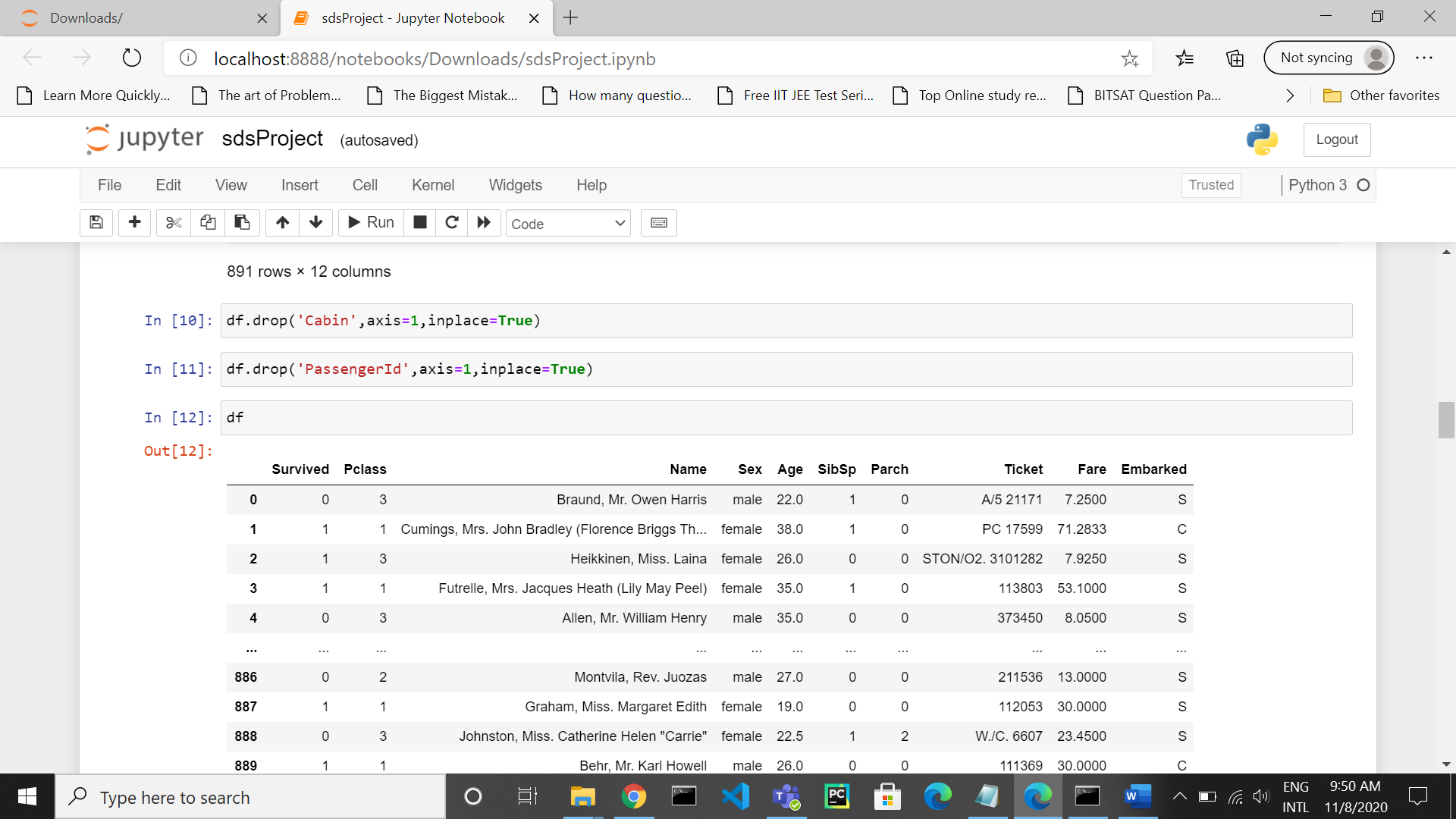
In our dataset there are two columns, age and cabin, having missing values. Column cabin has 77% missing values and age has almost 20% missing values.



* Initially we cleaned the dataset by removing all rows having null or na as values by using the dropna function( df=df.dropna(how='all') , where df is the dataset )We cleaned the dataset by entirely removing the cabin since 77% values were missing.

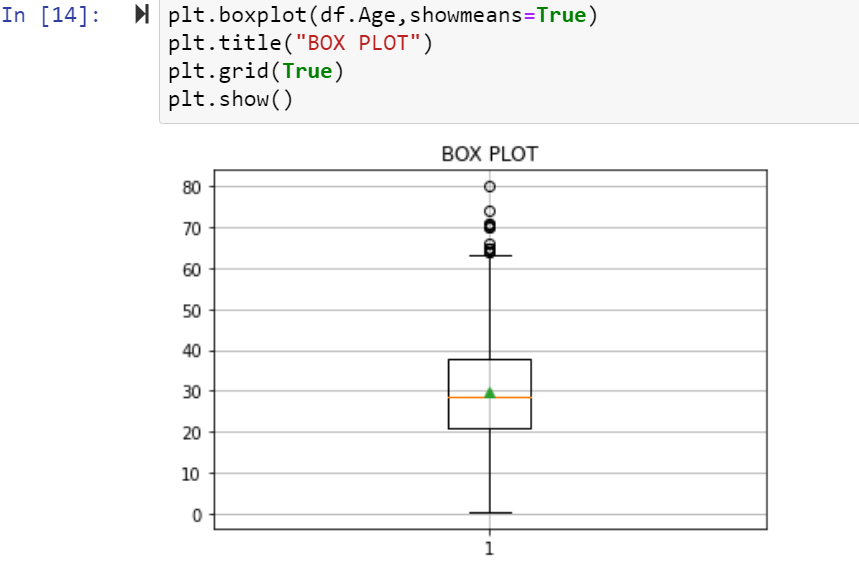


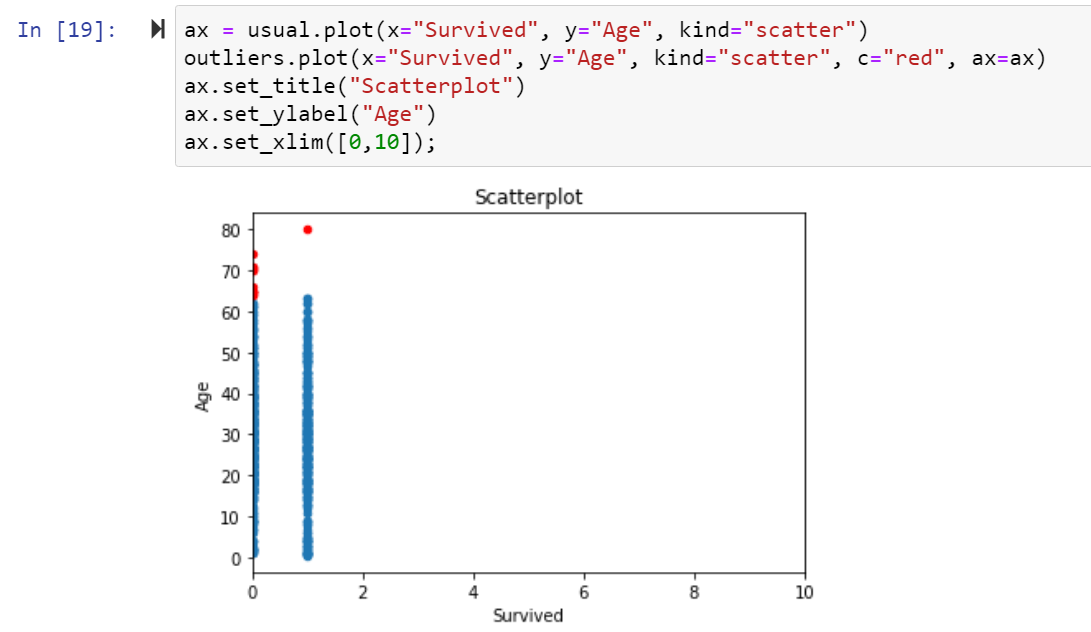
* We removed passengerId column since we don’t use it here in our project. Our main focus is on passenger class. We inserted the missing values for age by the technique of interpolation where the null values of age is added by the average values of the adjacent top and bottom age values. ( function used to do that was df.Age=df.Age.interpolate() , where df is the dataset used )



**Exploratory Data Analysis :**

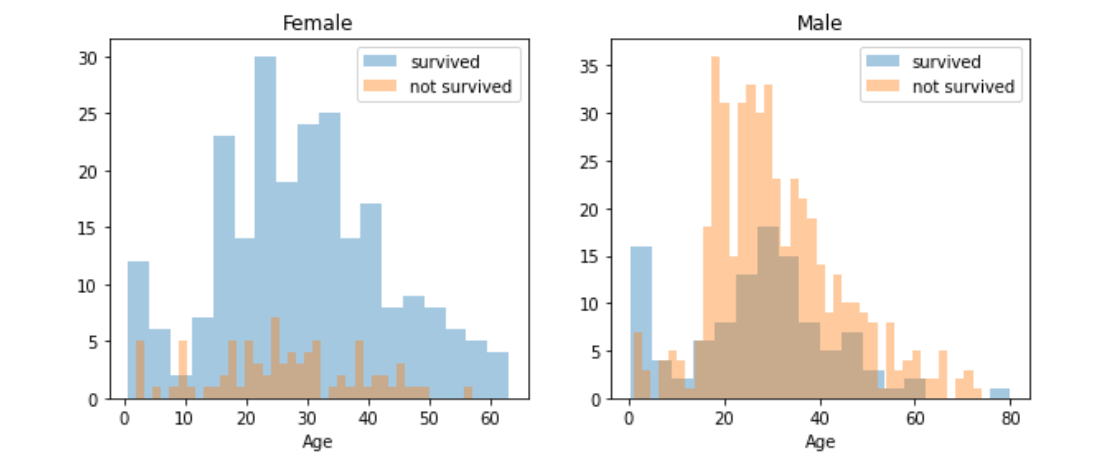
We constructed a box-plot for age and found a few outliers which we discarded.

keeping 1 as survived and 0 otherwise, we also constructed a scatterplot graph age vs survived.



There are 3 ports S,C and Q that represents the boarding points for the passengers.Another graph is visualised based on the port

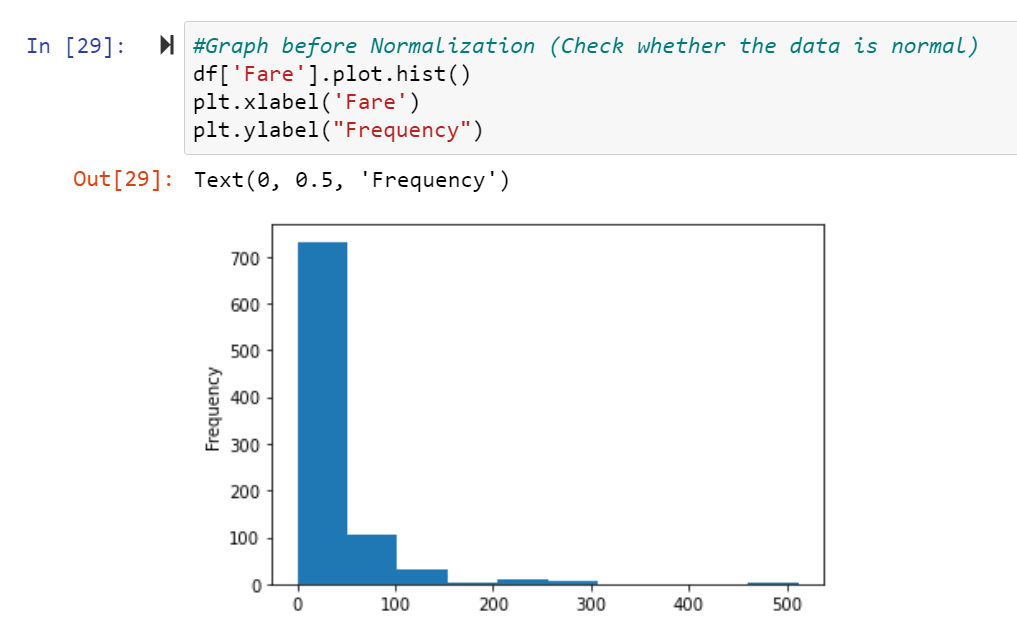
where the people had got on the ship and whether they were male or female and deduced who had a higher chance of surviving.



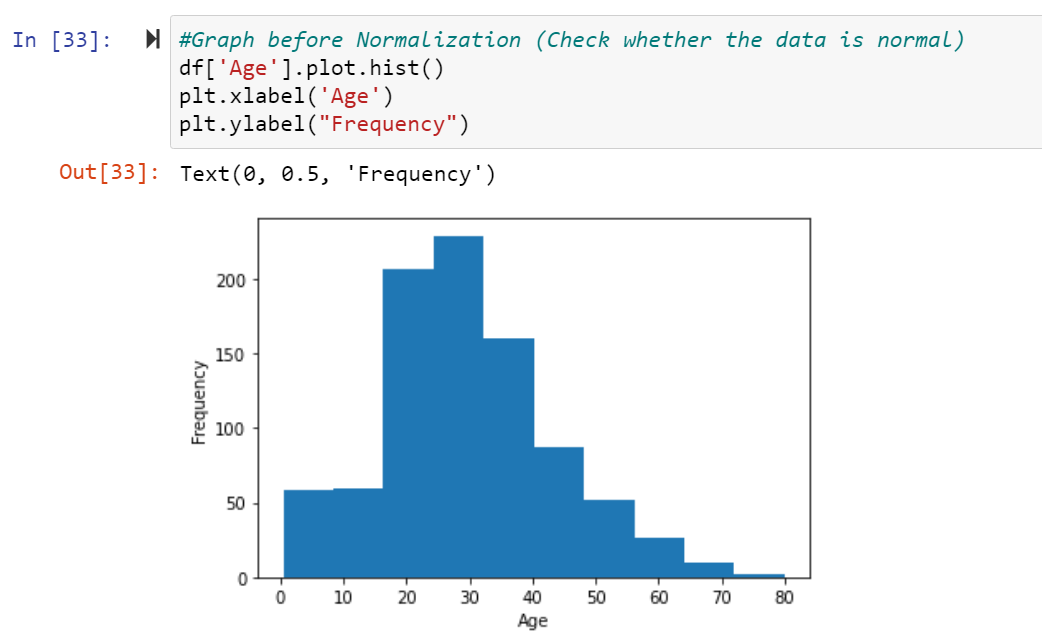
Normalization is the process of organizing a database to reduce

redundancy and improve data integrity.

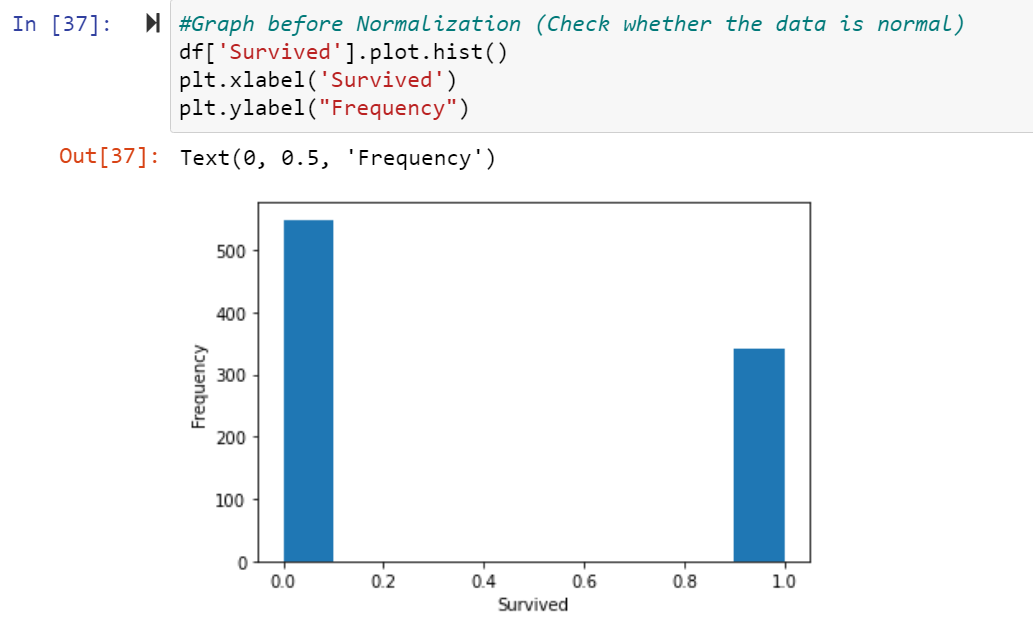
FREQUENCY VS FARE



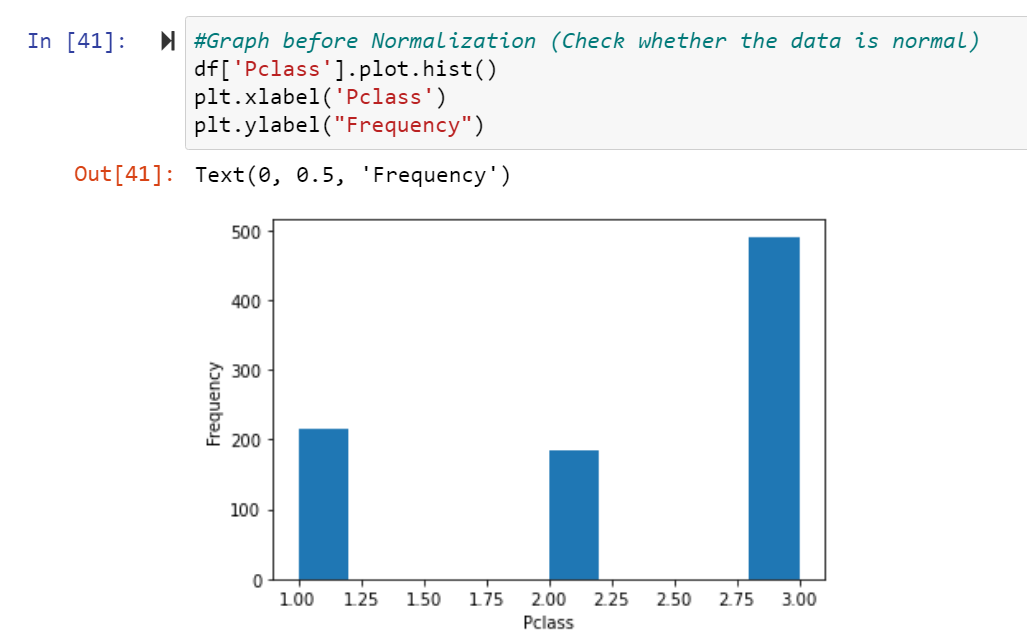
FREQUENCY VS AGE



FREQUENCY VS SURVIVED



**Our main focus is the PClass . the graph before normalization is represented below :**



WHY IS NORMALIZATION NEEDED?

The goal of data **normalization** is to reduce and even eliminate data

redundancy, an important consideration for application developers

because it is incredibly difficult to stores objects in a relational database

that maintains the same information in several places.

How does it affect dataset?

Normalization reduces redundancy. Redundancy is the unnecessary

repetition of data. It can cause problems with storage, reterival and

updation of data. Redundancy can lead to:



**Inconsistencies:-**errors are more likely to occur when facts are

repeated.

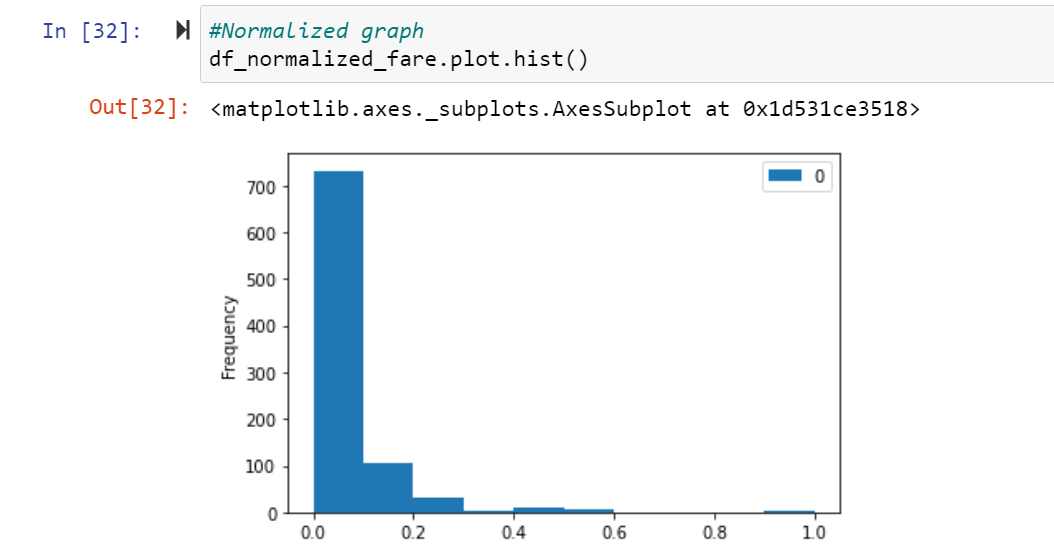


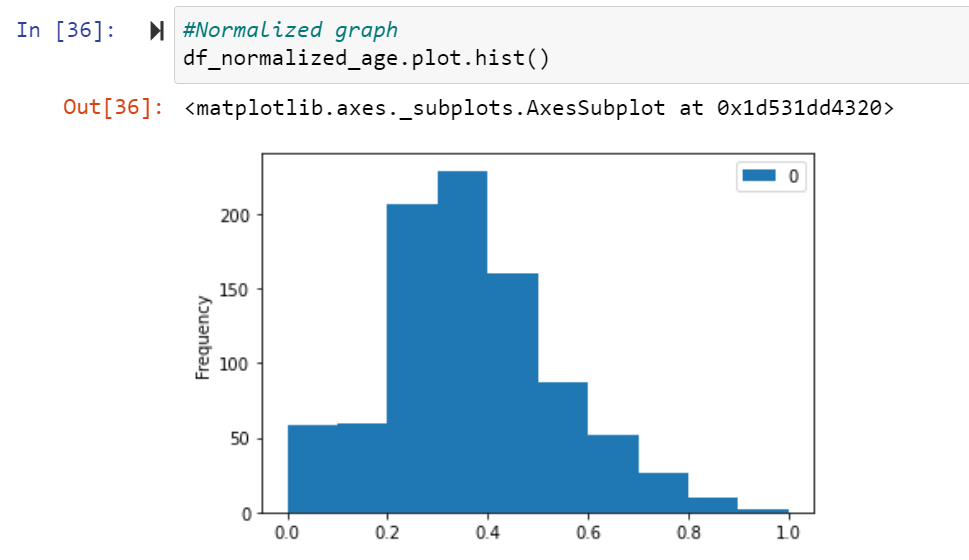
**Update anomalies:-**inserting, modifying and deleting data may

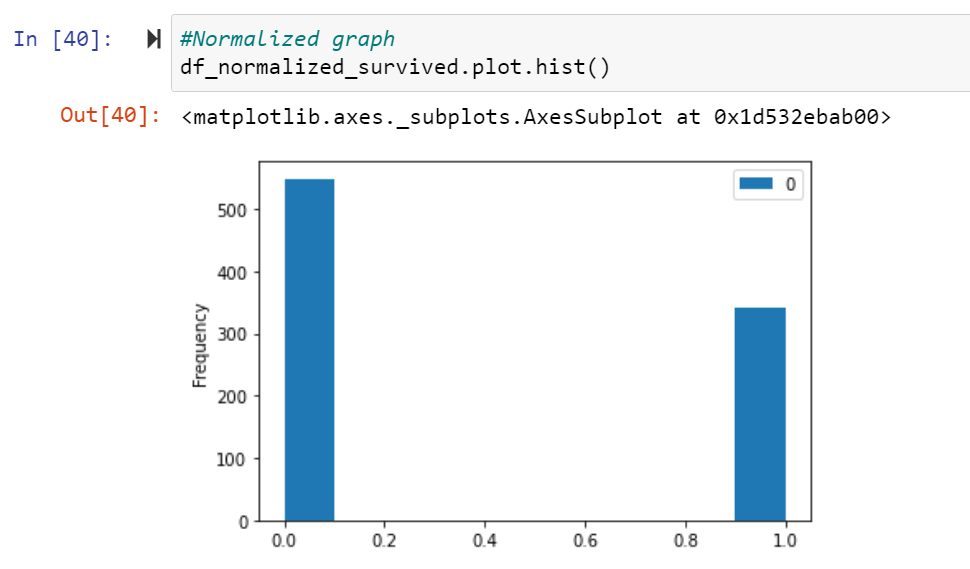
cause inconsistencies. Inconsistency occurs when we perform

updation or deletion of data in one relation, while forgetting to make

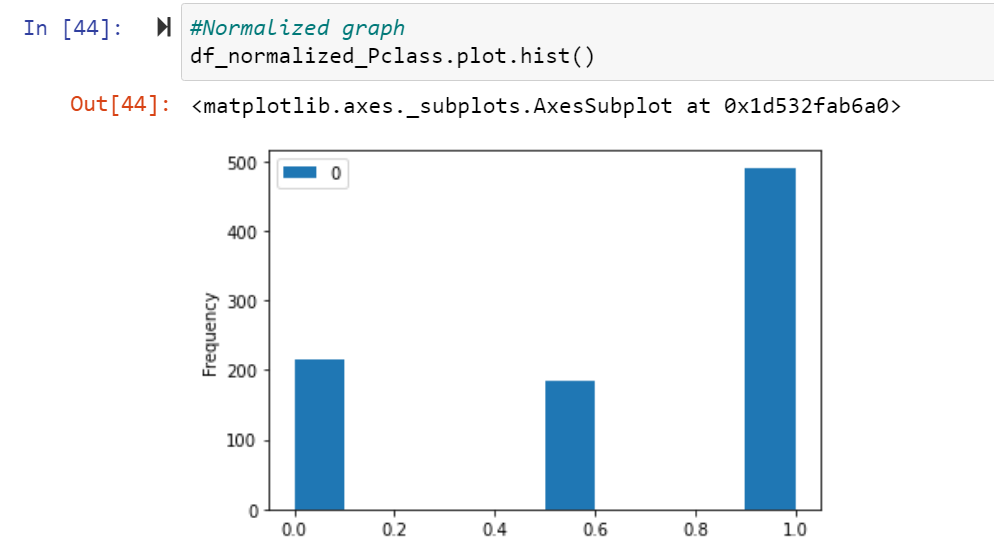
corresponding changes in other relations





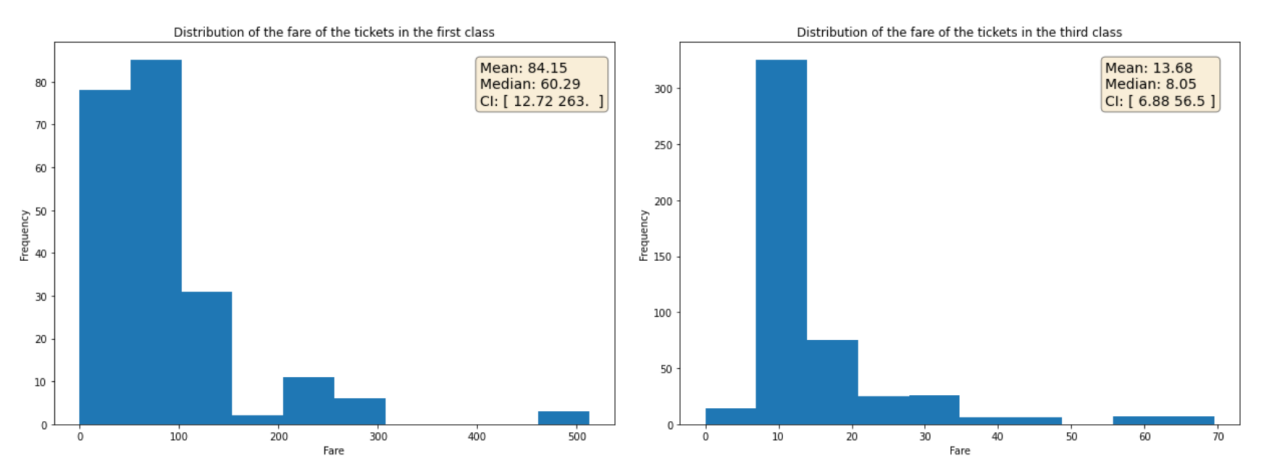


**Normalized graph of Frequency vs P-class**

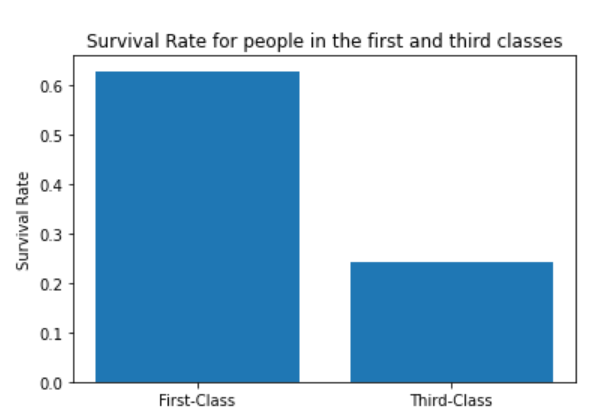


**Hypothesis Testing :**

hypothesis testing has to be conducted in order to assess the significance of that conclusion



 we can see that the class feature is a representation of the socio-economic status of the people in the ship and we can assume that those in the first-class group are the rich.



Looking at the plot, it’s very obvious that first-class people had a higher rate of survival. That will be our assumption for now . we need to prove it and conclude .

* Null Hypothesis: The socio-economic class of the people didn’t have an effect on the survival rate.
* Alternative Hypothesis: The socio-economic class of the people affected their survival rate.

According to our code and results, The P-value obtained is definitely much smaller than the significance level of 0.05 that we set. That P-value means that if we assume that the null hypothesis is true, then the probability of observing that effect by random is small according to the values obtained. Hence rejecting the Null hypothesis.

The conclusion is that the provided sample proves a significant correlation between the socioeconomic class and the survival rate.But we still can’t establish causation between these two features.We can just make a generalized induction that richer people had a better chance of survival at the ship.

**Results and Discussion**

The purpose of Titanic dataset is to use the existing features of passengers onboard Titanic as predictors to predict their survival outcome, for 0 being dead and 1 being survived from the tragic ship crash. Even from the logistic regression model, we can easily see that the Titanic survival outcome is highly depended on several predictors, such as sex, age and passenger class. In particular, female are more likely to have survived than male while keeping other predictors conditions constant, older people are less likely to have survived while keeping other predictors conditions constant; and lastly, people from a lower class are less likely to have survived keeping other predictors conditions constant. The conclusion is that the provided sample proves a significant correlation between the socioeconomic class and the survival rate.But we still can’t establish causation between these two features.We can just make a generalized induction that richer people had a better chance of survival at the ship.