Exploratory Data Analysis and Hypothesis Testing on Titanic Dataset

APOORVA KULKARNI

ABOUT TITANIC DATASET

The data consists of demographic and traveling information for 891 of the Titanic passengers, and the goal is to predict the survival of these passengers.

The Titanic dataset is also the subject of the introductory competition on Kaggle.com

INDEX

- Data Overview
- Data Analysing
- Data Cleaning
- Converting Categorical Features
- Numeric Data
- Numeric Data Analysis
- Hypothesis Testing

Data Overview

The Titanic data has 891 rows and 12 columns. Top 5 rows are given below.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Missing values in Dataset

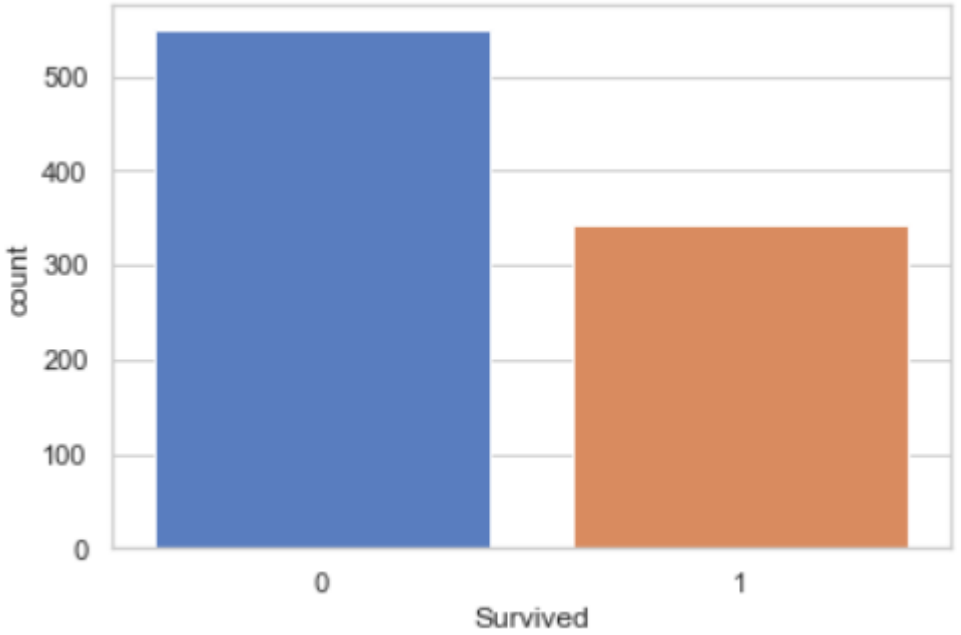
Approximately 20% of Age data is missing and 77% of cabin data is missing.

We can also visualize missing data using seaborn

	False	True
PassengerId	891.0	NaN
Survived	891.0	NaN
Pclass	891.0	NaN
Name	891.0	NaN
Sex	891.0	NaN
Age	714.0	177.0
SibSp	891.0	NaN
Parch	891.0	NaN
Ticket	891.0	NaN
Fare	891.0	NaN
Cabin	204.0	687.0
Embarked	889.0	2.0

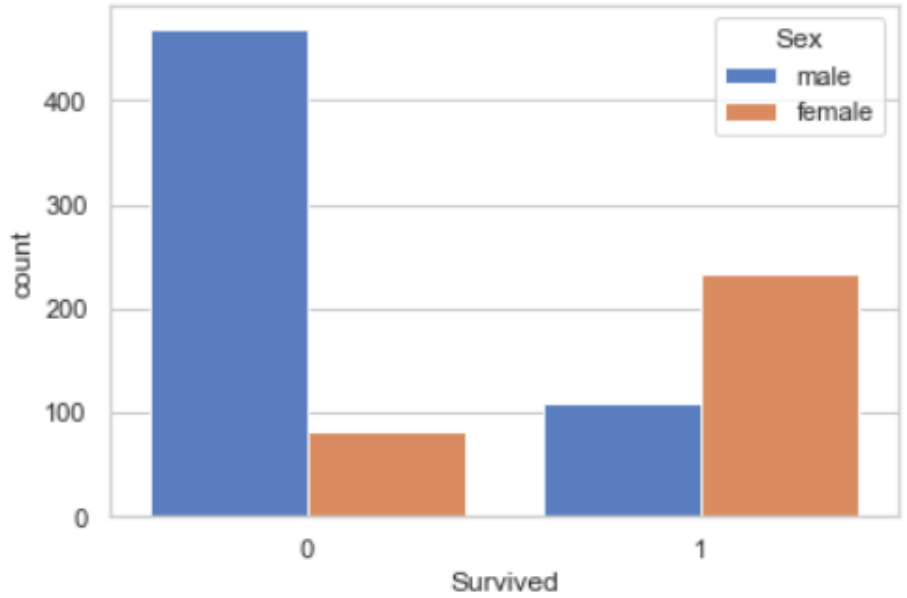
Data Analysing

Most of the people died in the Titanic Tragedy , only around 300 people survived.

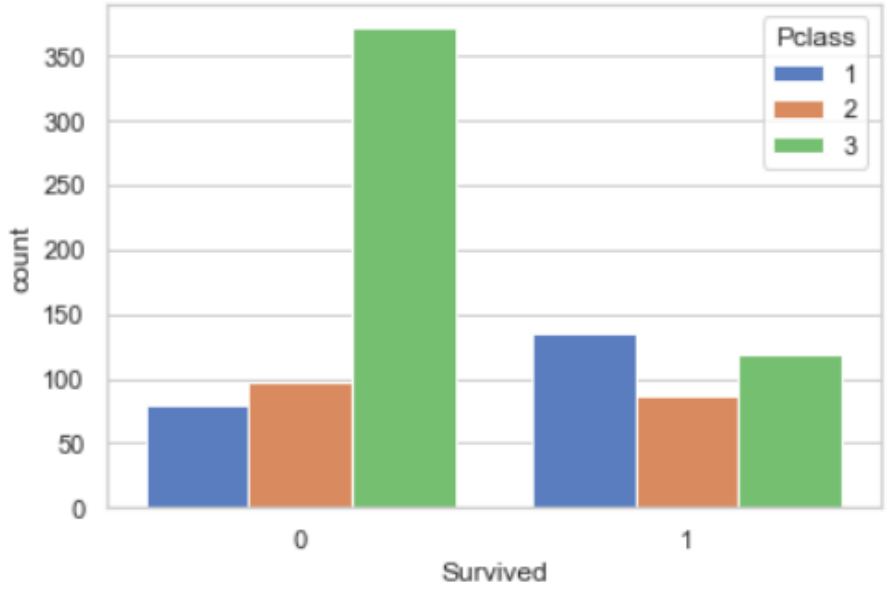


Data Analysing

SURVIVAL BASED ON GENDER



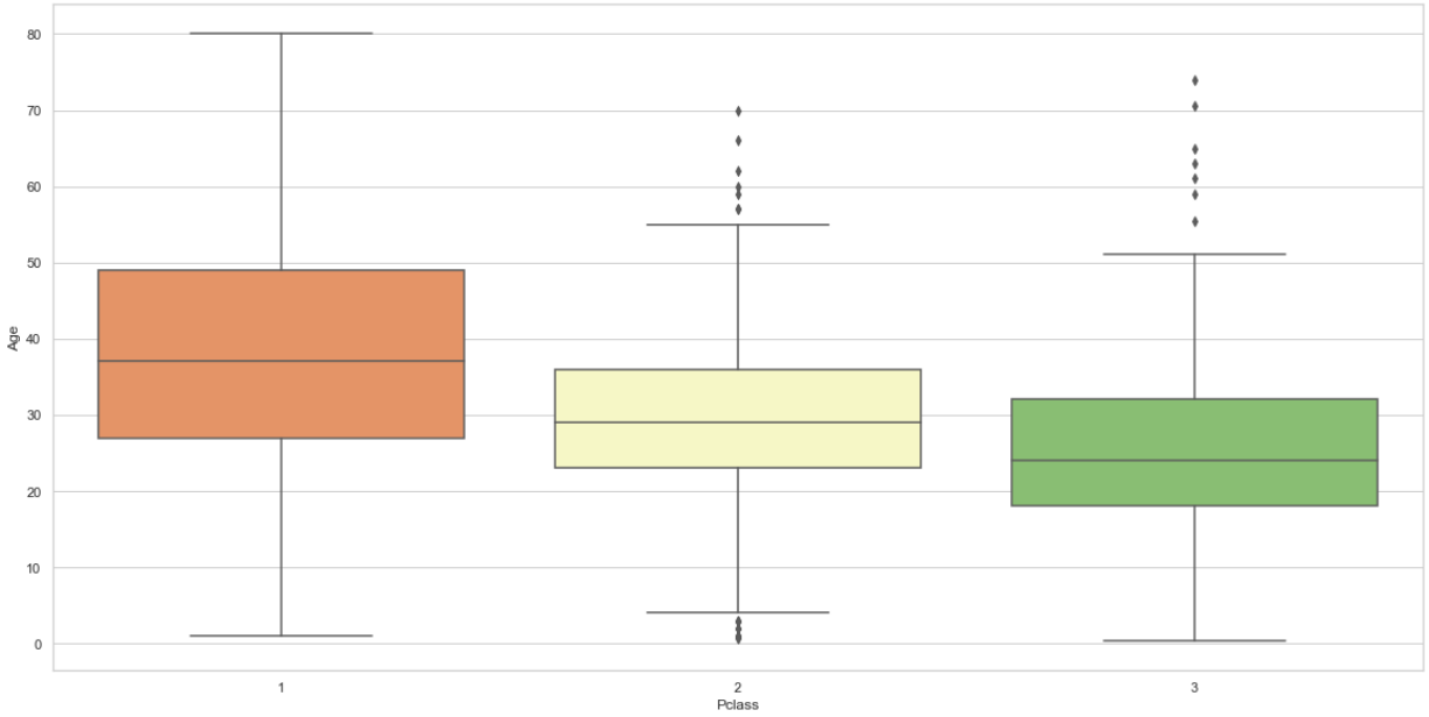
SURVIVAL BASED ON SOCIO-ECONOMIC CLASS



There were more males than females aboard the ship, roughly double the amount. More males died in Titanic Tragedy than females. The majority of the people traveling, had tickets to the 3rd class.

Data Cleaning

We have to fill the missing data in Age Feature. Traditionally we fill calculating mean of all passengers ages (imputation) or We can check average age by passenger class. Here a function written to fill the missing value



Boxplot (Age vs Pclass)

```
def filling_age(col):
    Age=col[0]
    Pclass=col[1]

    if pd.isnull(Age):

        if Pclass==1:
            return 37

        elif Pclass ==2:
            return 29

        else:
            return 24

    else:
        return Age
```

Median is taken from boxplot and filled in function

Converting Categorical Features

- Features such ‘Sex’ , ‘Embarked’ is encoded using Label Encoder
- Remove Name, Ticket and Cabin feature as they are not useful for prediction.

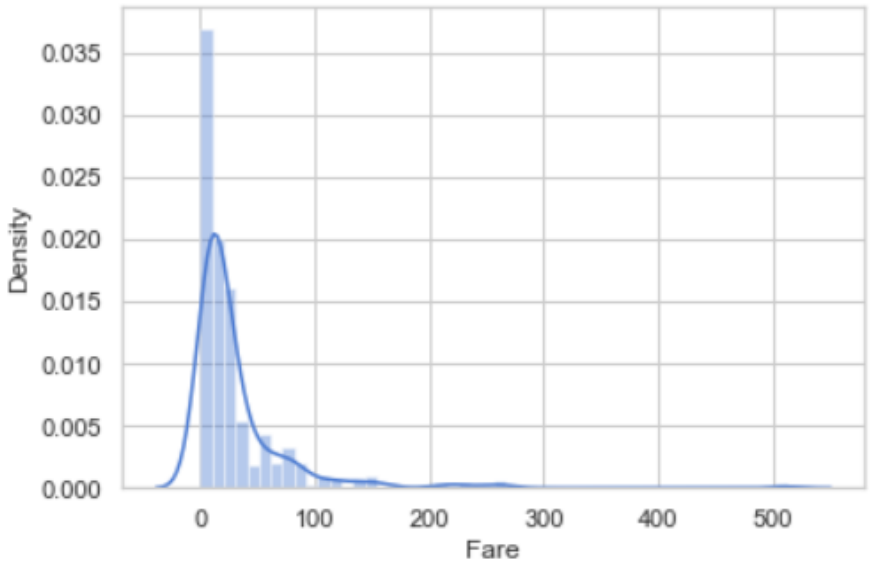
Numeric Data

- Remove the features ‘PassengerId’, 'SibSp','Parch’.
- Now we have 891 rows and 6 columns
- Plot the Pairplot to check skewness and outliers.

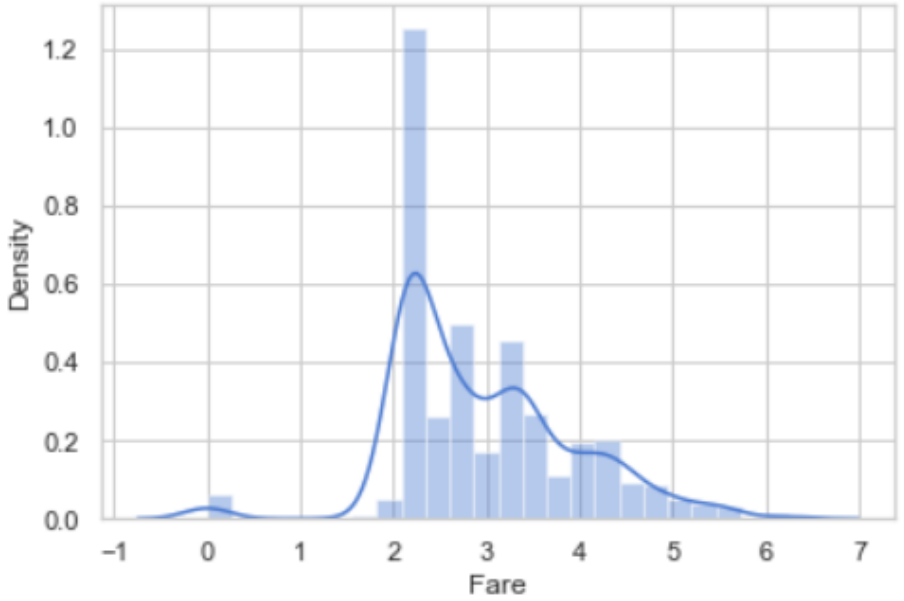
Numeric Data

Fare Feature is positively skewed. Hence Log Transformation is applied.

BEFORE LOG TRANSFORMATION



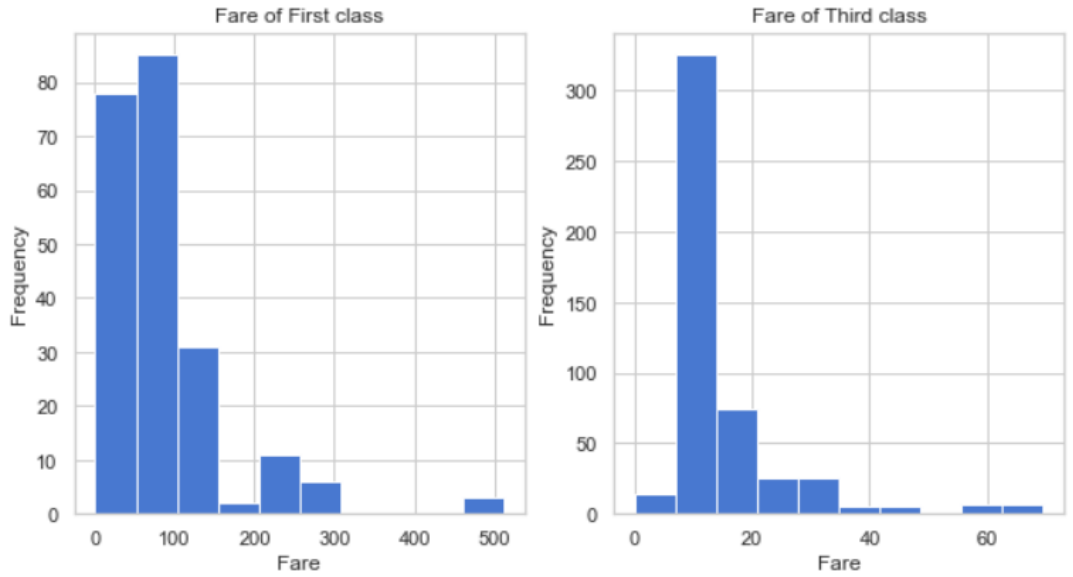
AFTER LOG TRANSFORMATION



Numeric Data Analysis

The First class ticket has mean of 84.15 USD which is much lower than mean of third class ticket.

Hence, we can assume that people travelling with first class ticket were rich.



Numeric Data Analysis

We can see in the above plot that First Class passengers had Higher Survival Rate than third Class passengers.

Hence, Hypothesis test is conducted to check whether the survival rate was affected by socio-economic class of passengers.

Hypothesis Testing

Hypothesis Test is conducted to check whether the socio-economic status of the passenger affected their survival rate

Null Hypothesis(H0): The socio-economic class of the people didn’t have an effect on the survival rate.

Alternative Hypothesis(Ha): The socio-economic class of the people affected their survival rate.

Hypothesis Testing

Z test is conducted.

Here a sample of 100 mean is taken for each population.

Significance level is 0.01.

Z score is calculated and 28.16 is obtained

P value(Two tailed test) is 1.6027274360886918e-174 is obtained

Conclusion

P value is obtained is much lower to our significance level so we can comfortably reject our Null Hypothesis. The provided sample proves that there is correlation between socio-economic status and survival rate. Hence there was more chance for rich passenger to survive in the titanic tragedy.