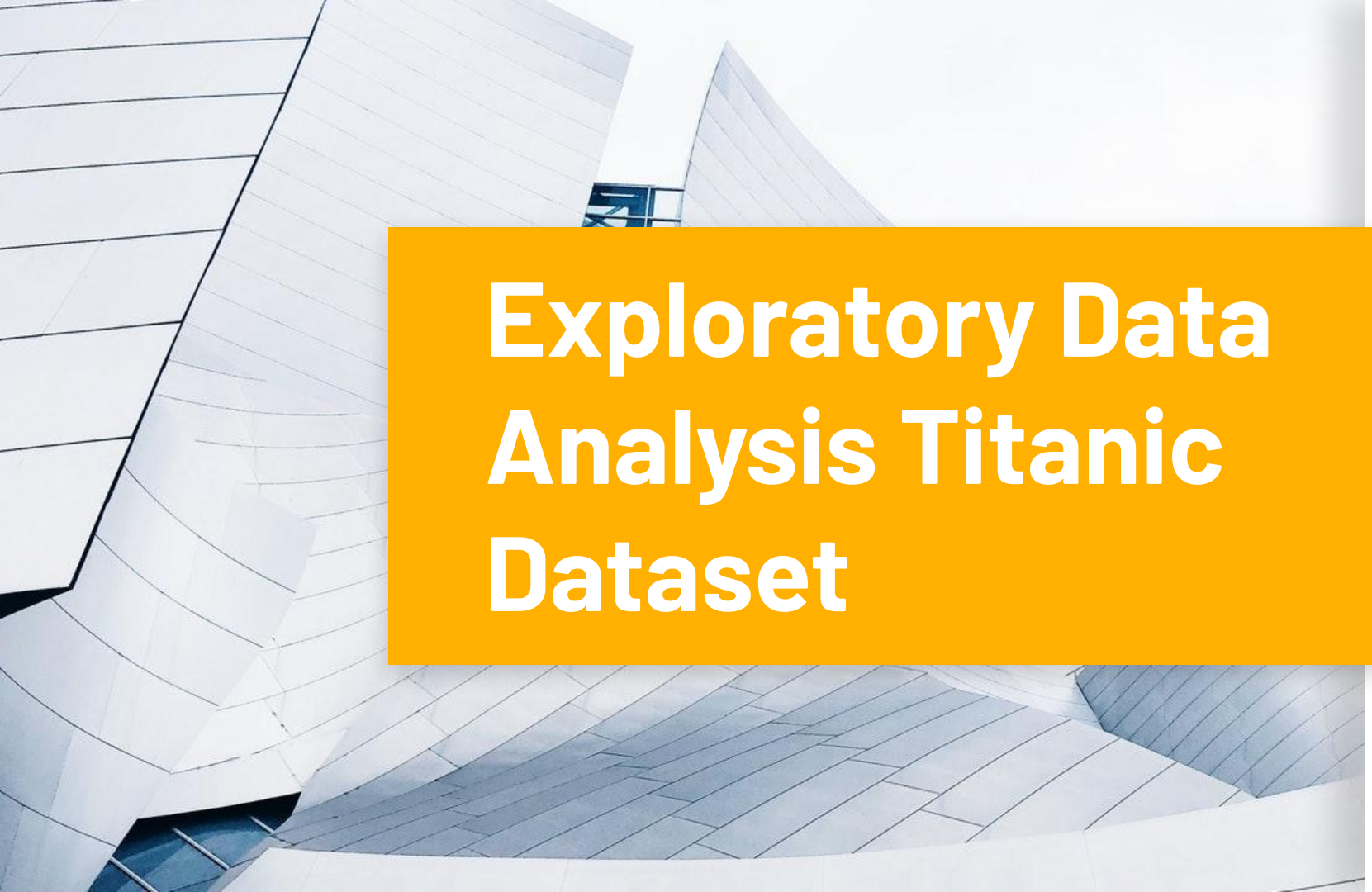


Exploratory Data Analysis Titanic Dataset

Initial plan for data exploration

1. Data Overview
2. Delete Columns Not Useful Discriminate Target
3. Missing Values Identification And Treatment
4. Outliers Detection And Treatment
5. Feature Engineering
6. Visualizations on insights
7. Hypothesis Testing

1. Data Overview

1. The dataset has 12 rows and 891 columns.
2. There are 7 numerical columns and 5 columns with text data.
3. There are some columns with missing values
4. Dataset has columns with categorical data like Pclass,Survived,Sex,Embarked

Brief description of the data set and a summary of its attributes

PassengerId : int64 : Ids given to pax while boarding cruise

Survived : int64 : Survival status of pax,(0 = No; 1 = Yes)

Pclass : int64 : Class of pax in cruise. (1 = 1st; 2 = 2nd; 3 = 3rd)

Name : object : Name of Pax

Sex : object : Gender of pax

Age : float64 : Age of pax

SibSp : int64: Number of Siblings/Spouses Aboard

Parch : int64 : Number of Parents/Children Aboard

Ticket : object : Ticket Number of pax

Fare : float64 : Passenger Fare

Cabin : object : Cabin

Embarked : object : Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Brief description of the data set and a summary of its attributes

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

```
train_df.describe()
```

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	14.526497	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Delete Columns Not Useful Discriminate Target

1. There are columns like
    i. 1. PassengerId
    ii. 2. Cabin
    iii. 3. Ticket
    iv. 4. Name
2. Which are not useful to predict the target variable.
3. Hence have deleted those column

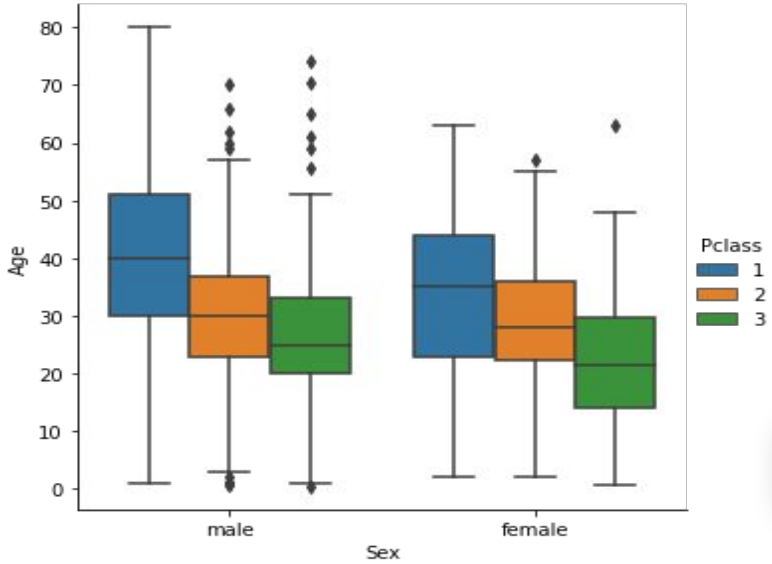
Missing Values Identification And Treatment

1. Column Age and Embarked have null values.
2. Embarked column has only 0.22% values are null hence imputed them with mode.
3. Age column have 19.87% null values which is significant number hence decided to input those values rather than deleting them.

Survived	0.00
Pclass	0.00
Name	0.00
Sex	0.00
Age	19.87
SibSp	0.00
Parch	0.00
Fare	0.00
Embarked	0.22

Missing Values Identification And Treatment

1. **Null values of Age column has been filled by finding its correlation with Pclass and taking means of those values.**

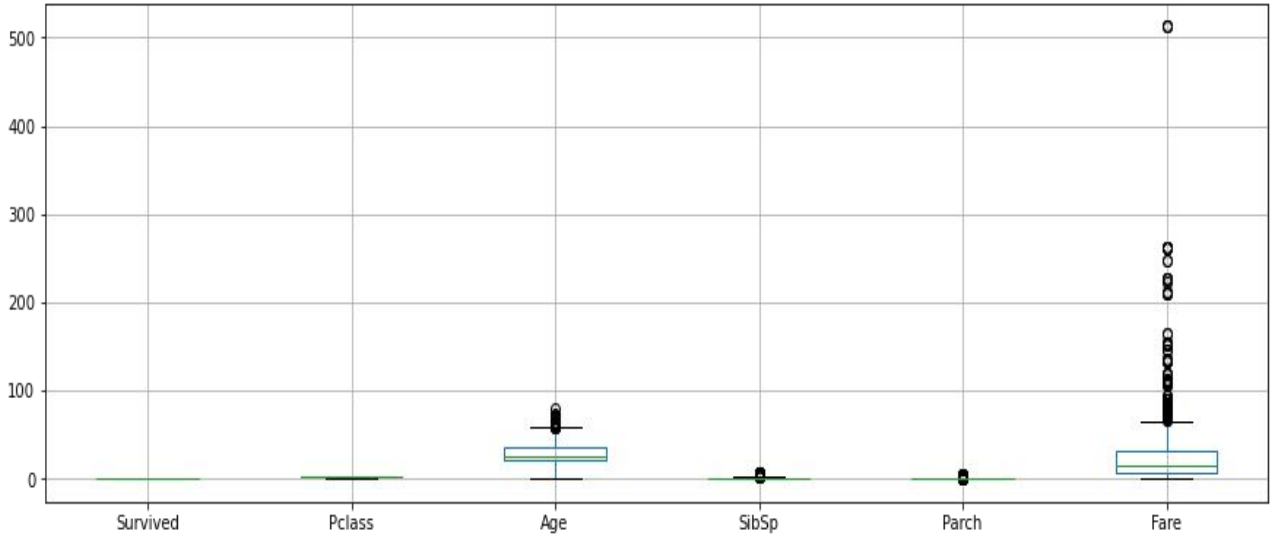


Outliers Detection And Treatment

1. There are outlier detected in below columns.
    a. Age
    b. SibSp
    c. Parch
    d. Fare
2. Since outliers in SibSp and Parch are valid values which would not affect predictions significantly.

Outliers Detection And Treatment

1. **Outlier under column Age and Fare has been removed using Z-score methods.**

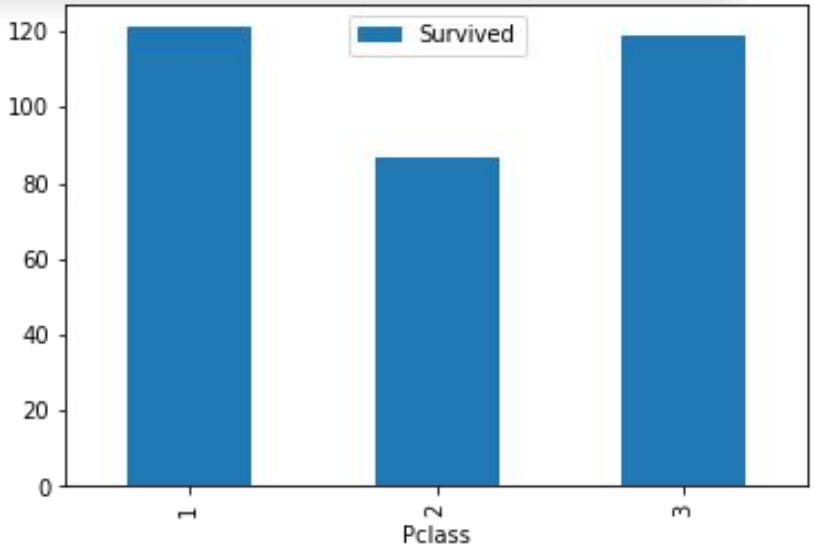


Feature Engineering

1. Checked the distribution of numerical columns.
2. Found that numerical columns are following the approximate normal distribution
3. Checked the categorical columns and identified the requirement of encoding.
4. For Sex column nominal encoding applied.
    a. Female = F
    b. Male = M

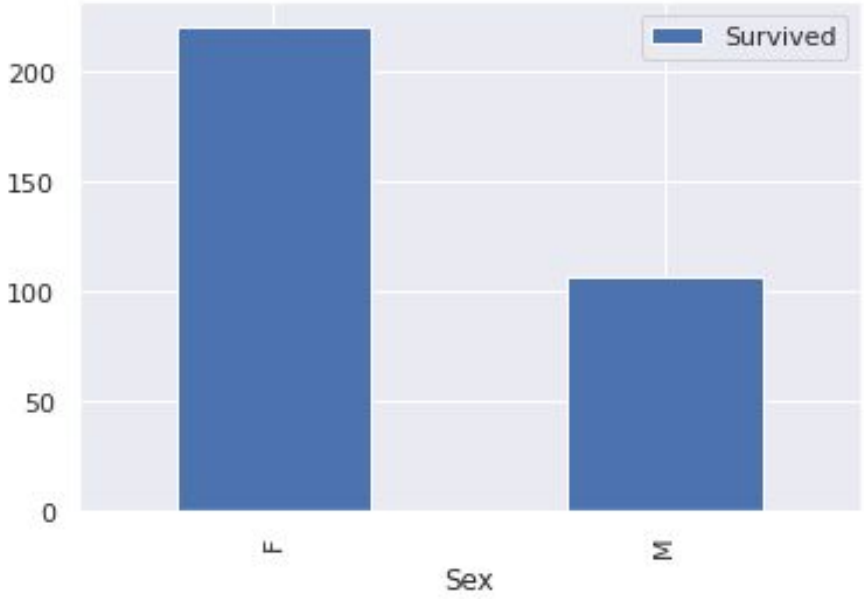
Visualizations on insights

1. Survival status of second
2. class is less as compared
3. to first and second class



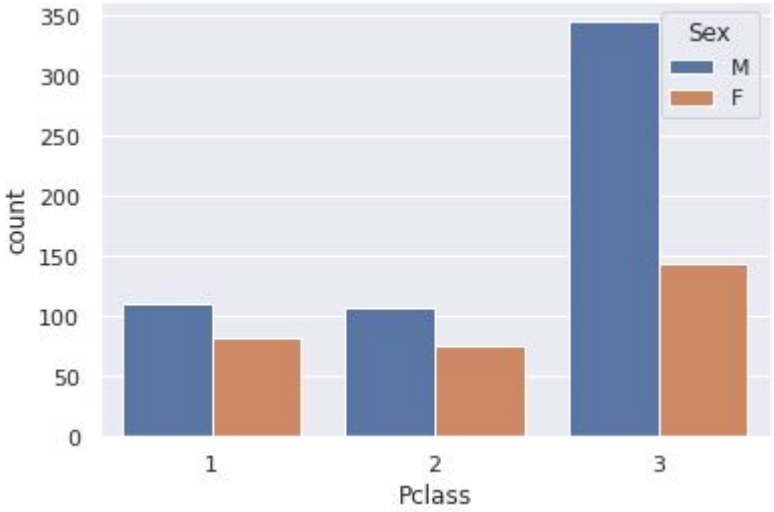
Visualizations on insights

1. It can be seen that
2. more female survived
3. than men



Visualizations on insights

1. Male often choose to
2. travel in third class
3. where female are less
4. likely to choose third
5. class



Hypothesis Testing : Pclass Against Survival Rate

1. Main Purpose : Check if Plcass has affecting the survival rate
2. Null Hypothesis: The socio-economic class of the people didn't have an effect on the survival rate.
3. Alternative Hypothesis: The socio-economic class of the people affected their survival rate.
4. P_value : 5.645603538613823e-147
5. Null Hypothesis Rejected
6. It indicating that socio-economic class of the people may have an effect on survival rate.
7.

Hypothesis Testing : Sex Against Survival Rate

1. Main Purpose : Check if gender has affecting the survival rate
2. Null Hypothesis: Gender of the people didn't have an effect on the survival rate.
3. Alternative Hypothesis: Gender of the people affected their survival rate.
4. P_value : Is closed to zero
5. Null Hypothesis Rejected
6. It indicating that gender of the people may have an effect on survival rate.
7.

Hypothesis Testing : Sex against Pclass

1. Main Purpose : Check if gender has affecting to choose the Pclass
2. Null Hypothesis: Gender of the people didn't have an effect on the Pclass.
3. Alternative Hypothesis: Gender of the people affected their Pclass.
4. P_value : 2.1701131381042185e-19
5. Null Hypothesis Rejected
6. It indicating that gender of the people does affect while choosing the Pclass
7.

Conclusion

- As shown in analysis, logistic regression can be a good machine learning algorithm to predict the survival status of passenger.

- Provided datasets proves a significant correlation between the socioeconomic class and the survival rate.

- Provided datasets proves a significant correlation between the sex and the survival rate.