IBM PEER GRADED ASSIGNMENT

Data Set Chosen - Titanic Prediction Set

Source - https://www.kaggle.com/c/titanic/data

1) Description

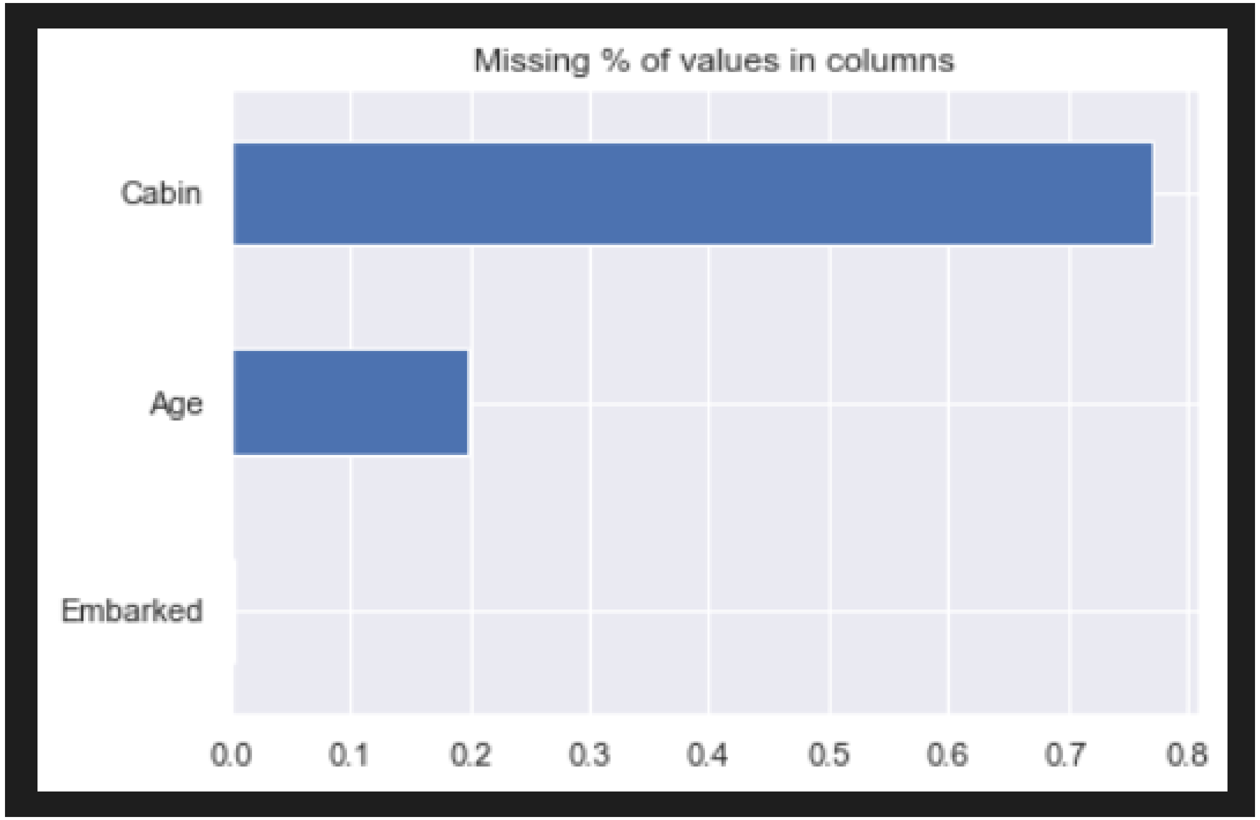
- The sinking of the Titanic is one of the most infamous shipwrecks in history.

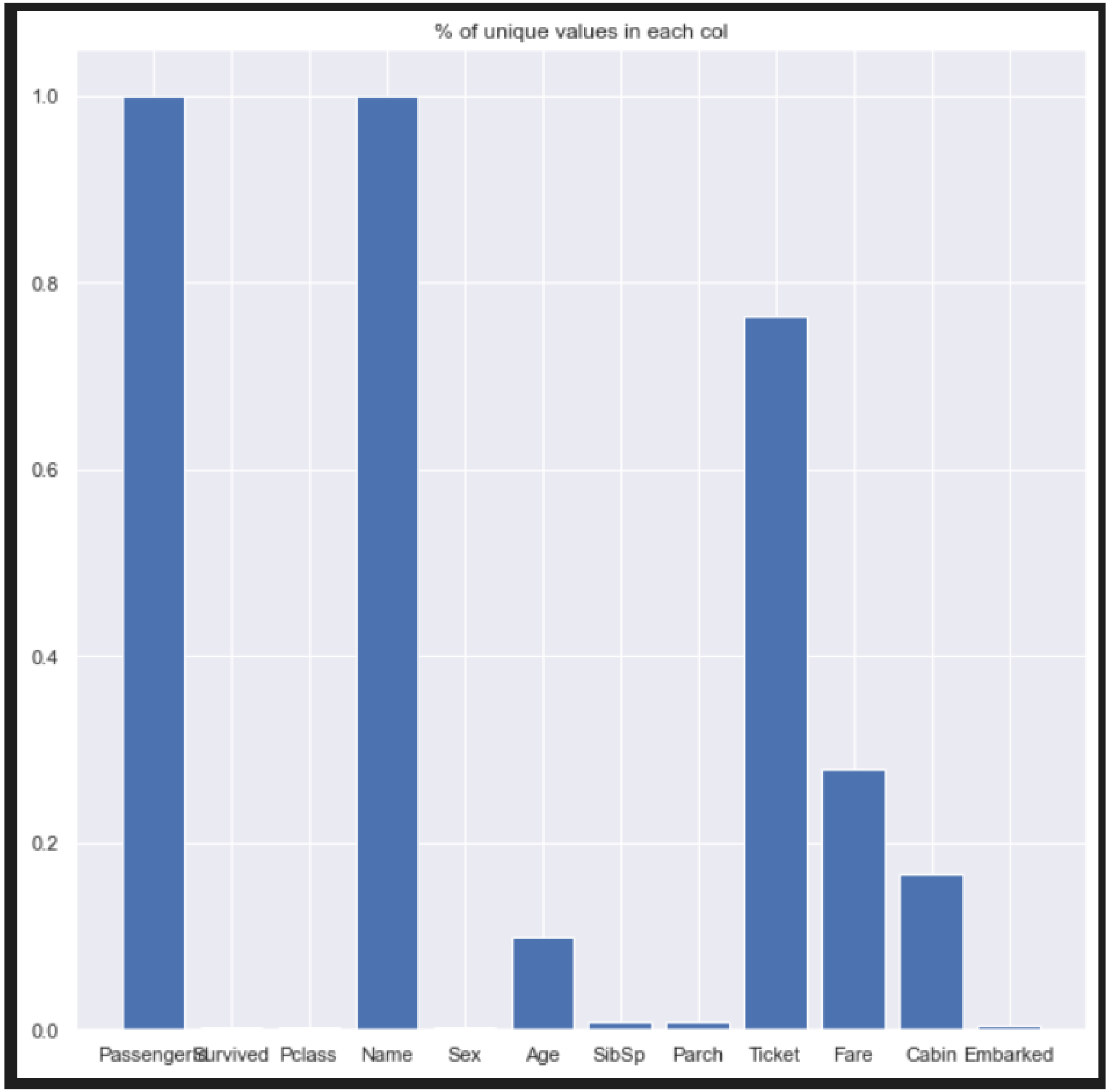
- On April 15, 1912, during her maiden voyage, the widely considered “unsinkable” RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren’t enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

- While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.This Datsets contains data about all the passengers abord Titanic and wethere they survived or not

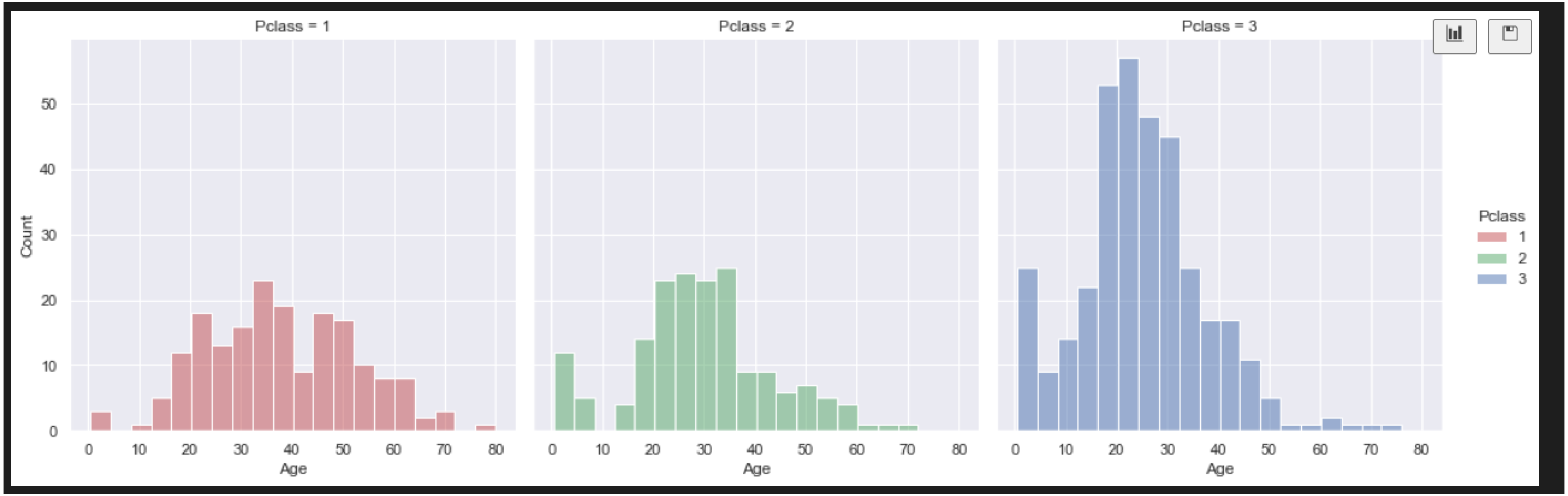
2)Data Exploration

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   PassengerId  891 non-null    int64
 1   Survived     891 non-null    int64
 2   Pclass       891 non-null    int64
 3   Name         891 non-null    object
 4   Sex          891 non-null    object
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64
 7   Parch        891 non-null    int64
 8   Ticket       891 non-null    object
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object
 11  Embarked     889 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

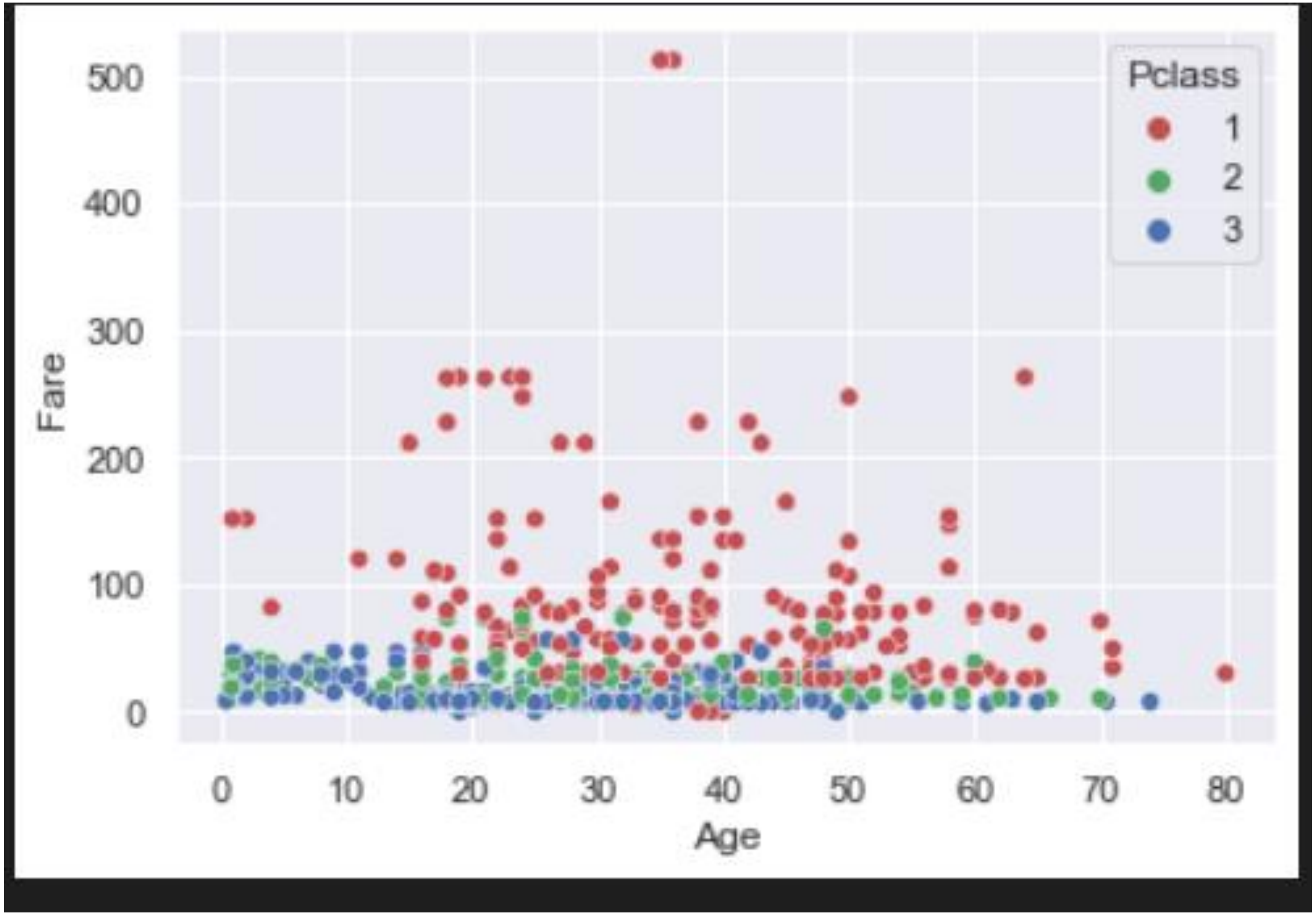



% of unique values in each col
1.0
0.8
0.6
0.4
0.2
0.0
PassengerId
Survived
Pclass
Name
Sex
Age
SibSp
Parch
Ticket
Fare
Cabin
Embarked

Exploring on Passenger Class and Age



Exploring on Age and Fare



3) Data Cleaning and Feature Engineering

- Drop Columns with many Unique Values (Name , Passenger Id , Ticket)
- Drop Columns with many missing values (Cabin)
- Fill Missing Data in Age with median grouped by Passenger Class and in Embarked with mode

After Data cleaning

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Survived  891 non-null    int64
 1   Pclass    891 non-null    int64
 2   Sex       891 non-null    object
 3   Age       891 non-null    float64
 4   SibSp     891 non-null    int64
 5   Parch     891 non-null    int64
 6   Ticket    891 non-null    object
 7   Fare      891 non-null    float64
 8   Embarked  891 non-null    object
dtypes: float64(2), int64(4), object(3)
memory usage: 62.8+ KB
```

Feature Engineering Using Ordinal Encoder

Before Encoding

```
one_hot_encode_cols = data.dtypes[data.dtypes == np.object]  # filtering by string categoricals
```

	0	1	2	3	4
Sex	male	female	female	female	male
Embarked	S	C	S	S	S

```python
# Ordinal Encoding
from sklearn.preprocessing import OrdinalEncoder
enc = OrdinalEncoder()
data[["Sex","Embarked"]] = enc.fit_transform(data[["Sex","Embarked"]])
data.info()
```

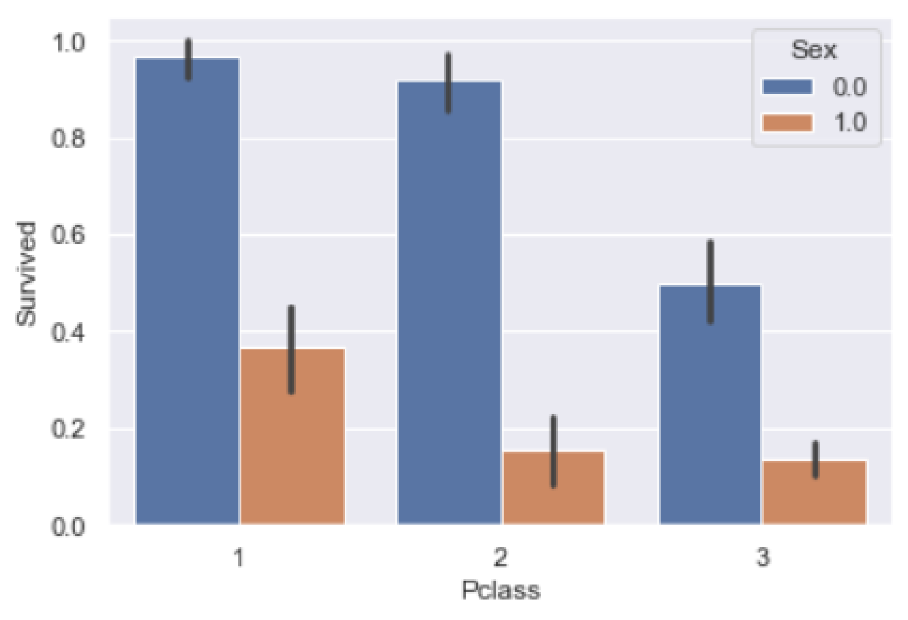
After Encoding

```
data.head()
```

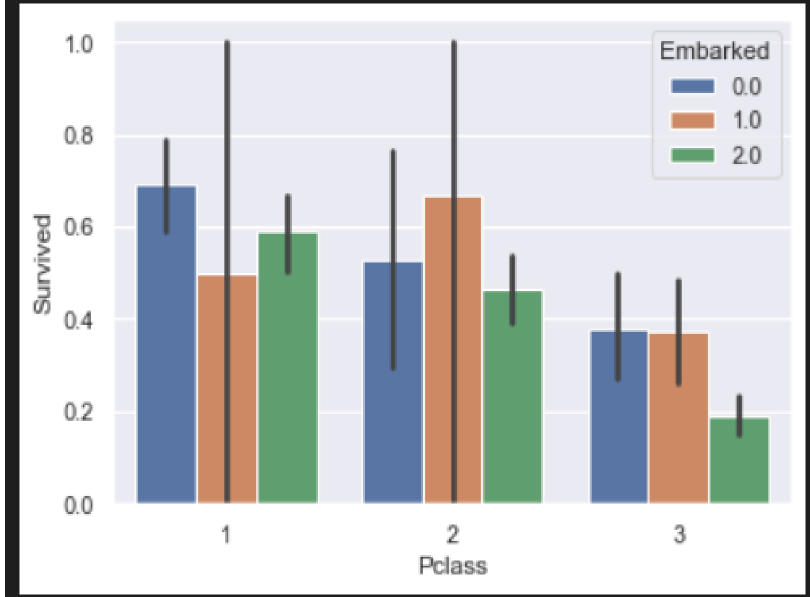
	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	1.0	22.0	1	0	7.2500	2.0
1	1	1	0.0	38.0	1	0	71.2833	0.0
2	1	3	0.0	26.0	0	0	7.9250	2.0
3	1	1	0.0	35.0	1	0	53.1000	2.0
4	0	3	1.0	35.0	0	0	8.0500	2.0

Before Encoding Sex column had values male and female but now it has 0 and 1
Embarked column had values as C,S and Q. After Encoding it has values 0,1 and 2

4) Exploratory Data Analysis

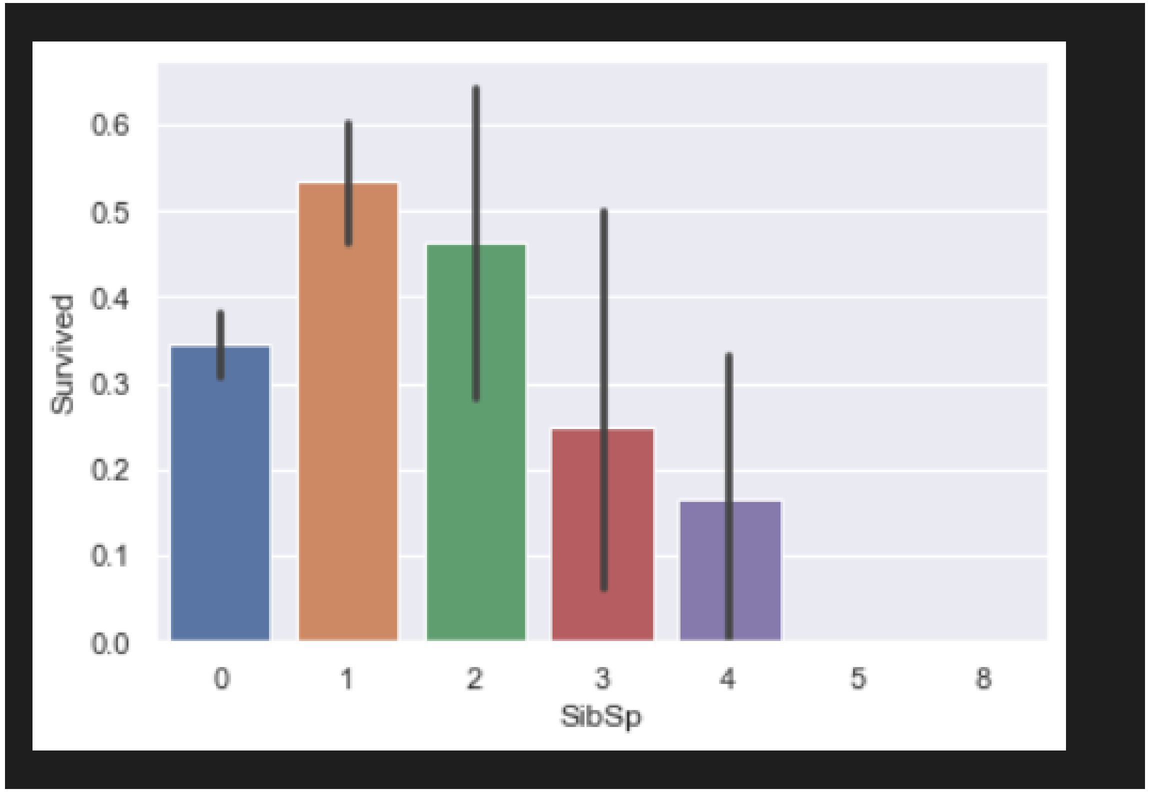


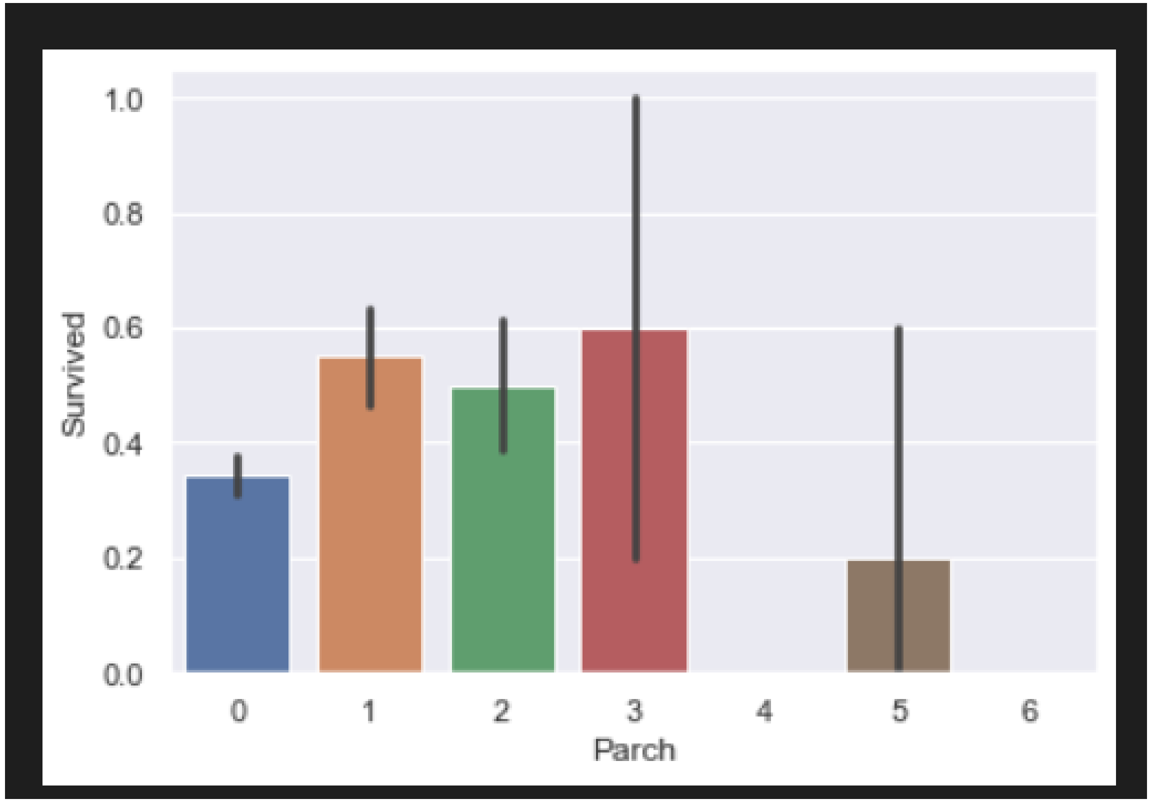
It can be seen ,Female generally have a high percentage survival rate then men and also Pclass plays an infuence on survival Rate



Place Embarked doenst play much influence however it can seen people embarked from cherbourg in third class have a very low rate of survival , this maybe beacuse they might be in the first decks which flooded or someother reason

-> further Data needed for further investigation





We dont see any noticable influence from Sibling Spouse and Parent children on Survival Rate

However one thing observed is more number of siblings spouse or parent children leads to less survival rate but this might also be due to a small proportion of people wit many parent children or spouse siblings

5) Hypothesis Testing

Null Hypothesis :

A person who didnt survive is from 3rd class and a men

Alternate Hypothesis 1 :

A person who didnt survive is from any class and a men

Alternate Hypothesis 2:

A person who didnt survive is from 3rd class and any gender

6) Testing Null Hypothesis

% of men in 3rd class who survived : 0.08080808080808081

From Null Hypothesis we can see that men from third class survivng rate is 8 %

7) Further analysis Recommended

Many of age data was missing and thus I took the median age data grouped by Passenger Class

The cabin column was deleted due to lack of 80% data, however cabin is an important information as the people in lower cabins which flooded first had low rate of survival. This column should be analyzed

Fare prices can be more analyzed on Pclass and Sex

8) Quality of Dataset

The Quality of this dataset is medicore as most of the cabin column was null which is an important information. Also age column had a lot of missing values which led to assuming median values which makes the dataset biased towards the middle aged people. Even after such backlashes the data set still provides enough data for us to conclude that Sex and Passenger Class have high influence on survival rate.

It can also be noticed that there isn't much difference in Ticket Fare values in second and third class.

Further Data on Cabin information , Life boat ratios ,age information can help us conclude better hypothesis and more accurate machine learning Models