Titanic Classification By Shantanu Garain

Importing dependencies

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

Loding data into our Notebook

In [2]:
```
# load csv file in our pandas dataframe
df = pd.read_csv('train.csv')
```

Looking for top 5 Rows of our dataset

In [3]:
```
df.head()
```

Out[3]:

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

Looking for bottom 5 Rows of our dataset

In [4]:
```
df.tail()
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Emb
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN	
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42	
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	

Size of our data

In [5]:
```
df.size
```

Out[5]:
```
10692
```

Shape of our data

In [6]:
```
# number of rows and Columns
df.shape
```

Out[6]:
```
(891, 12)
```

Some important information about the columns of our data

In [7]:
```
df.info()
```

```
<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
```

Some statistical information about the data:

In [8]:
```
df.describe()
```

Out[8]:

	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

DATA PREPROCESSING AND DATA CLEANING:

Checking Null values in each column:

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

Check missing values and remove them

In [10]:
```
# We see most missing values are in Cabin column so we remove them
df = df.drop(columns='Cabin', axis=1)
```

replacing the missing values in "Age" column with mean values

In [11]:
```
df['Age'].fillna(df['Age'].mean(), inplace=True)
```

Finding the mode value of "Embarked" column

In [12]:
```
df['Embarked'].mode()
```

Out[12]:
```
0    S
Name: Embarked, dtype: object
```

Replacing the missing values in "Embarked" column with mode value

In [13]:
```
df['Embarked'].fillna(df['Embarked'].mode()[0], inplace=True)
```

In [14]:
```
df.isnull().sum()
```

Out[14]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

Exploratory Data Analysis

Finding the total number of people survived and not survived

In [15]:
```
df['Survived'].value_counts()
```

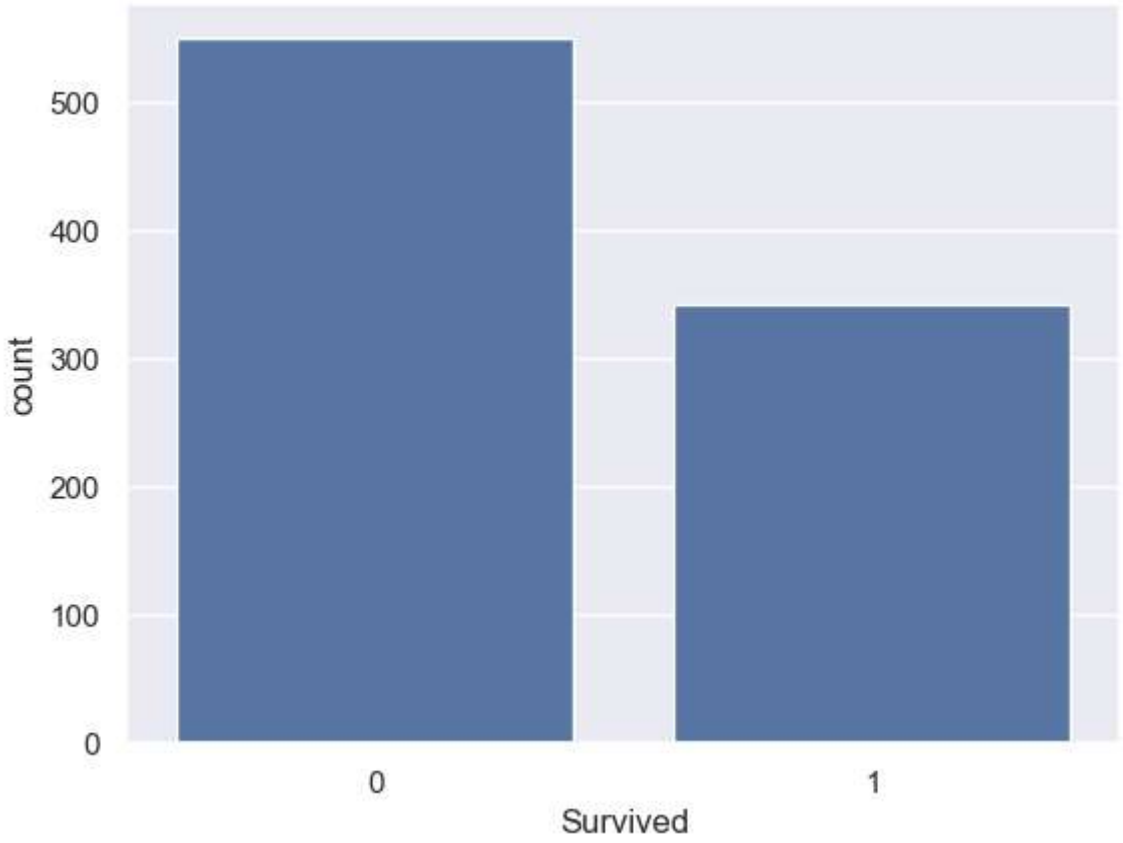
Out[15]:
```
Survived
0    549
1    342
Name: count, dtype: int64
```

Visualizing our data

In [16]:
```
sns.set()
```

Count plot of didn't survived and survived

In [17]:
```
# people didn't survived represent with '0'
# and survived represent as 1
sns.countplot(x='Survived', data=df)
plt.show()
```



Count plot for "Sex" column

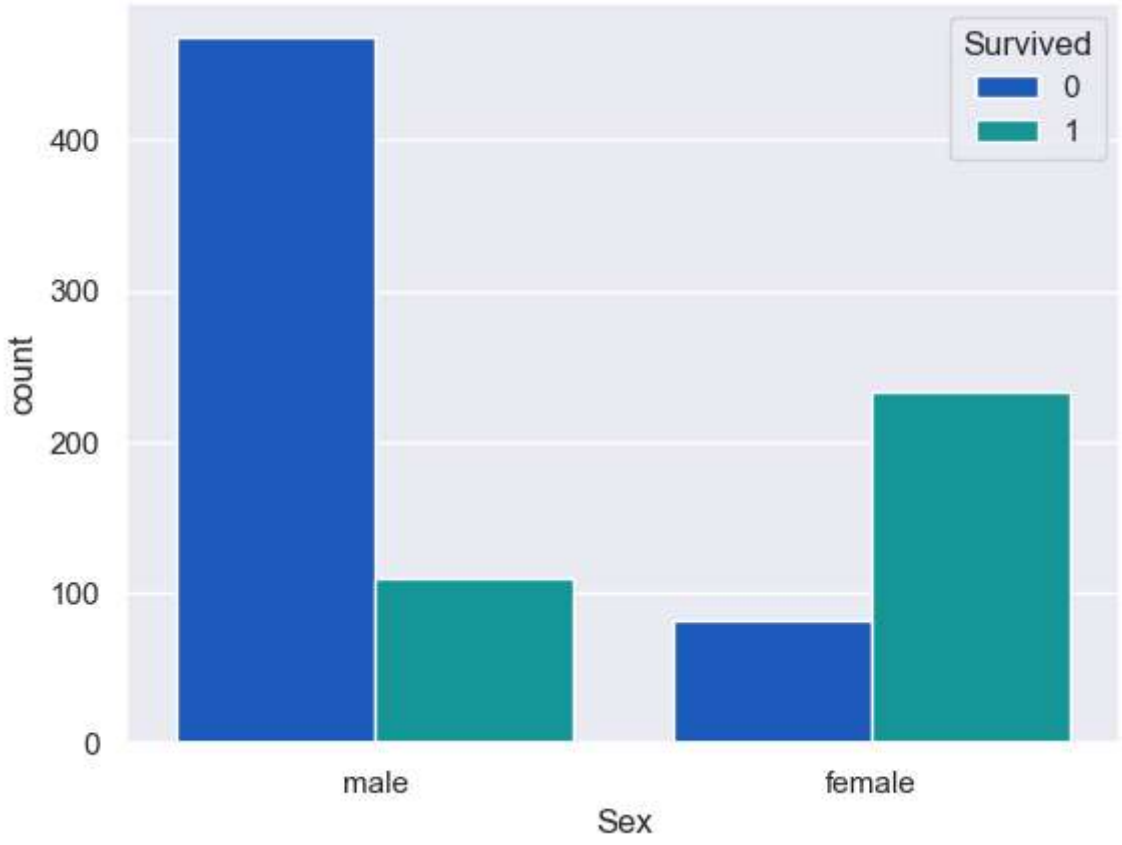
Finding the Male and Female ratio

In [18]:
```
df['Sex'].value_counts()
```

Out[18]:
```
Sex
male      577
female    314
Name: count, dtype: int64
```

In [19]:
```
sns.countplot(x='Sex', hue='Survived', data=df, palette='winter')
```

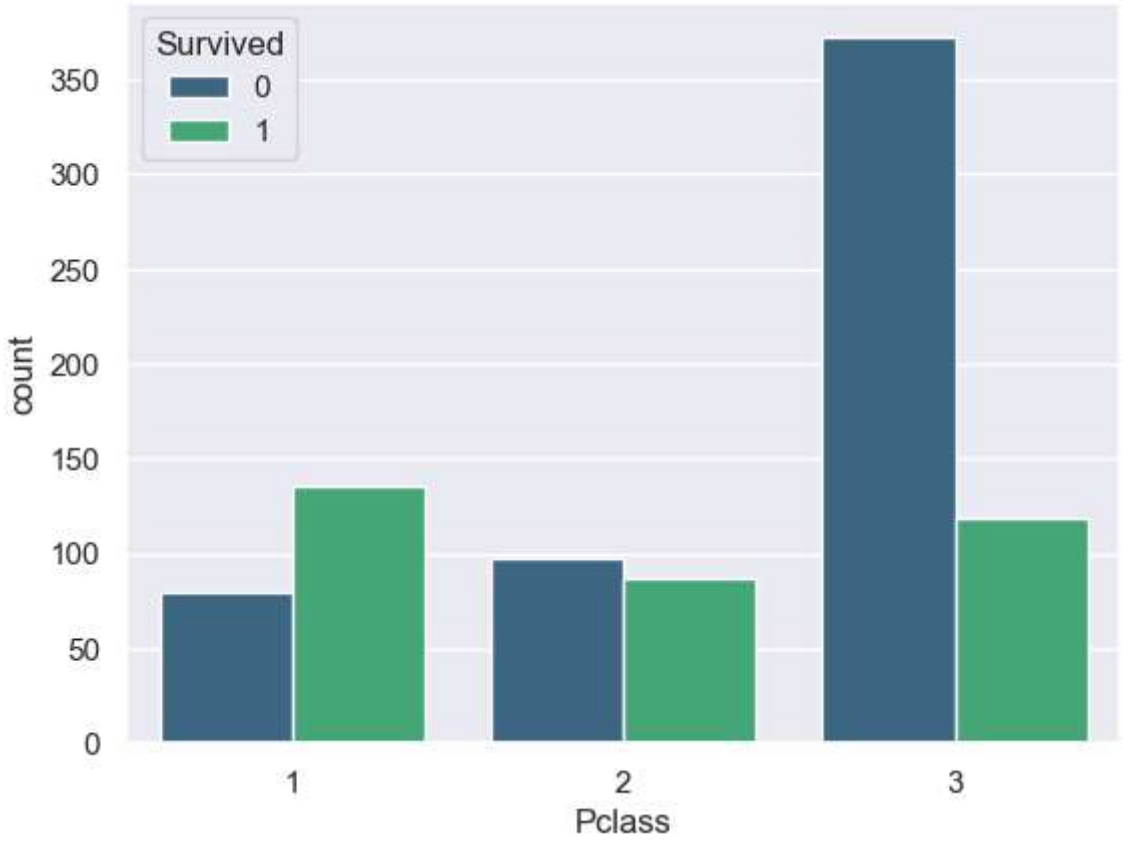
Out[19]:
```
<Axes: xlabel='Sex', ylabel='count'>
```



Count plot for Survived column

In [20]:
```
sns.countplot(x='Pclass', hue='Survived', data=df, palette='viridis')
```

Out[20]:
```
<Axes: xlabel='Pclass', ylabel='count'>
```

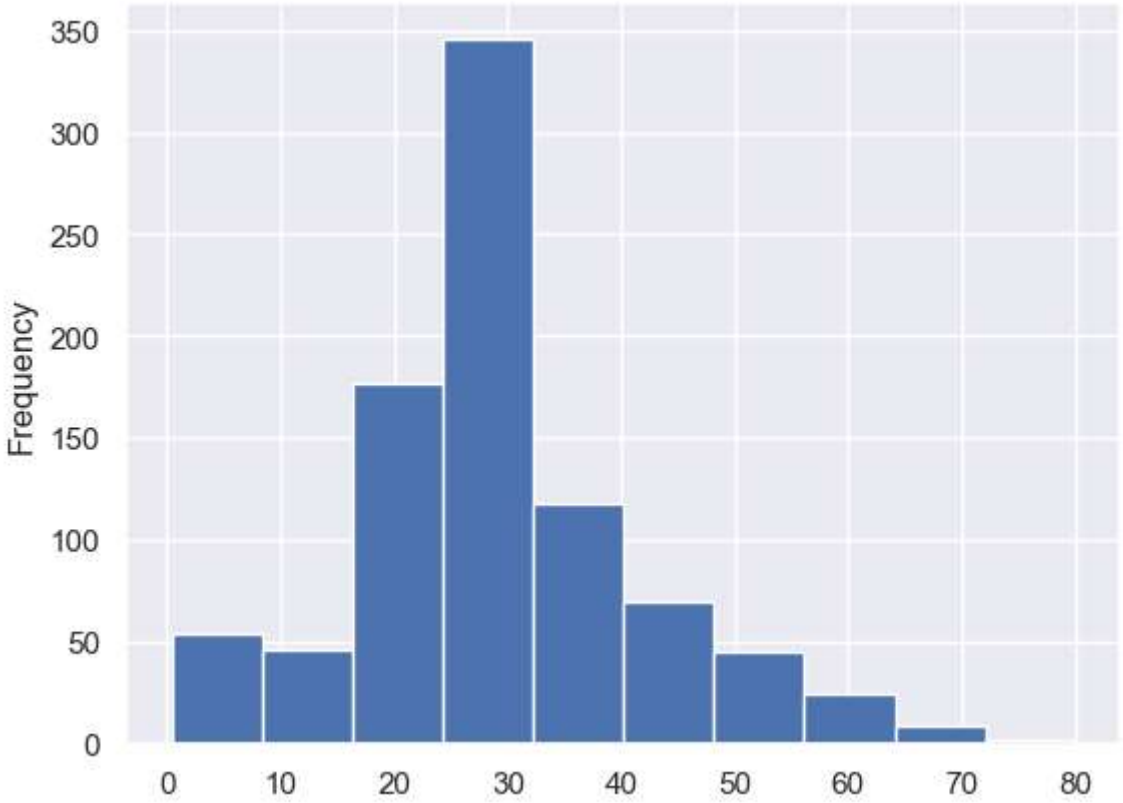


- We see that the passengers who did not survived belong to the 3rd class
- and the 1st class passengers are more likely to survive

Ploting the 'Age' group

In [21]:
```
df['Age'].plot.hist()
```

Out[21]:
```
<Axes: ylabel='Frequency'>
```

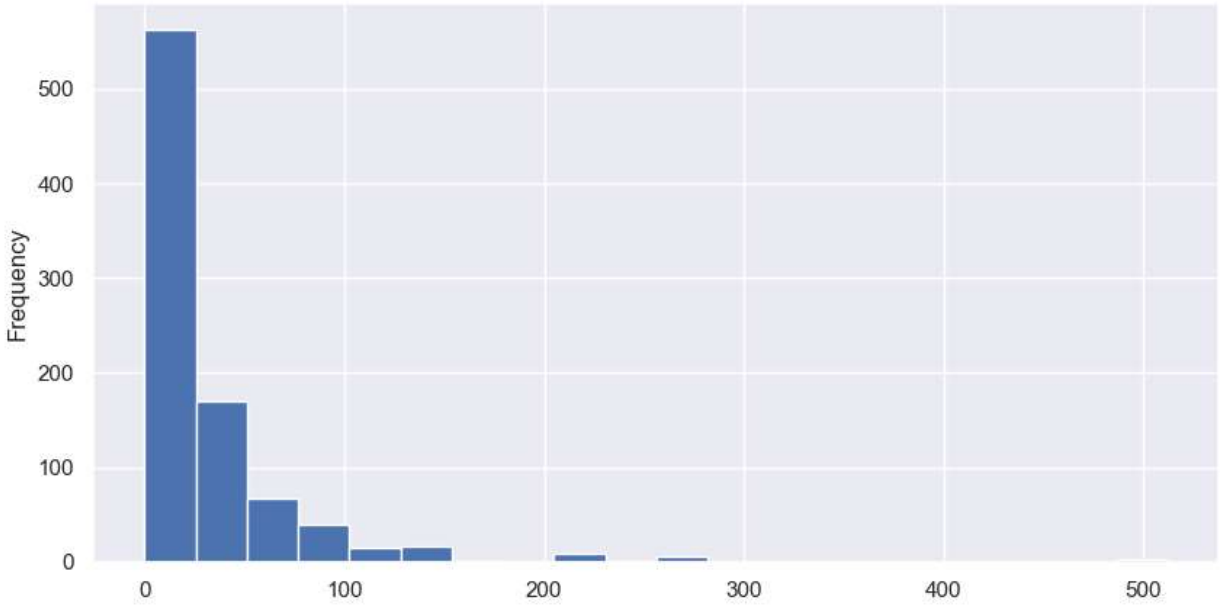


- We notice that higher age group travelling are among the young age between 20-40

Ploting the 'Fare' plot

In [22]:
```
df['Fare'].plot.hist(bins=20, figsize=(10,5))
```

Out[22]:
```
<Axes: ylabel='Frequency'>
```

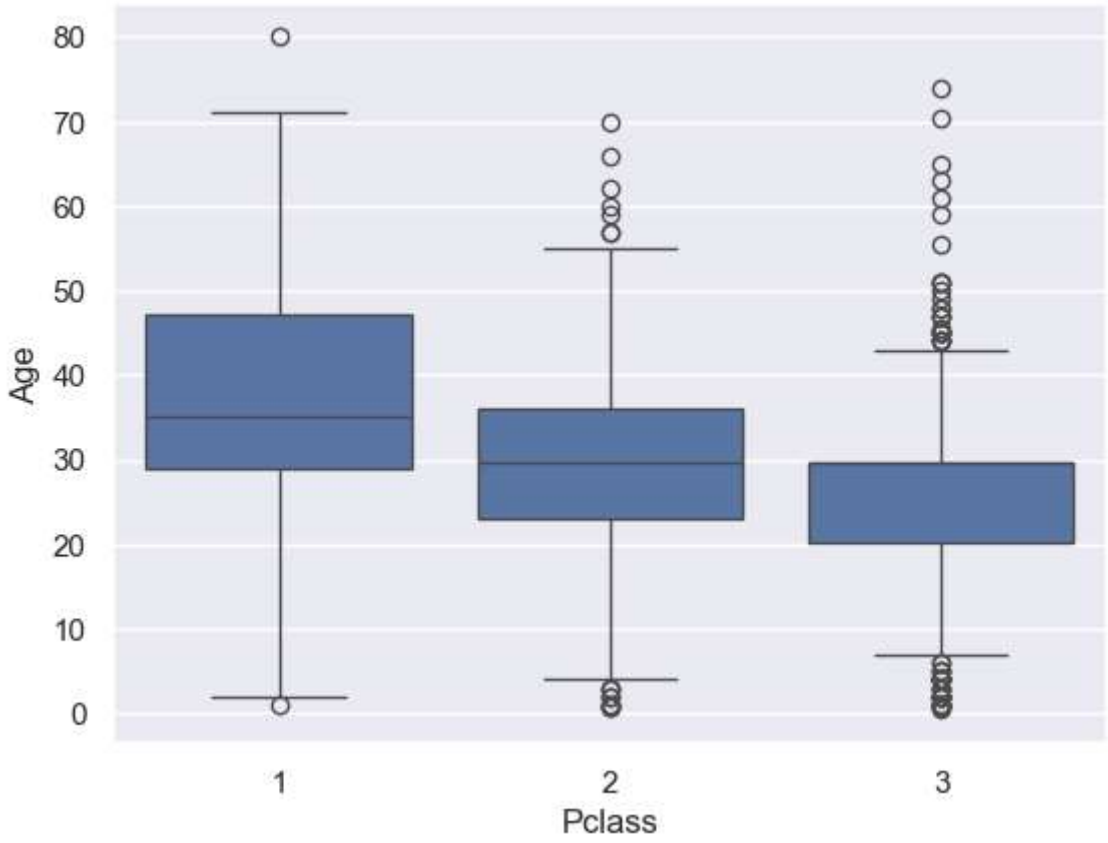


- We see that most of the tickets bought are under fare 100

- also very few are on the higher side of fare i.e 220-500 range

In [23]:
```
sns.boxplot(x='Pclass', y='Age', data=df)
```

Out[23]:
```
<Axes: xlabel='Pclass', ylabel='Age'>
```



- We can see that older age group are travelling more in class 1 and 2 compared to class 3

In [24]:
```
df['Sex'].value_counts()
```

Out[24]:
```
Sex
male      577
female    314
Name: count, dtype: int64
```

In [25]:
```
df['Embarked'].value_counts()
```

Out[25]:
```
Embarked
S    646
C    168
Q     77
Name: count, dtype: int64
```

In [26]:
```
# Converting categorical Columns to numerical

df.replace({'Sex':{'male':0,'female':1}, 'Embarked':{'S':0,'C':1,'Q':2}}, inplace=True
```

In [27]:
```
df.head()
```

Out[27]:

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

Separating features & Target

In [28]:
```
X = df.drop(columns = ['PassengerId','Name','Ticket','Survived'], axis=1)
Y = df['Survived']
```

In [29]:
```
print(X)
```

```
     Pclass  Sex        Age  SibSp  Parch     Fare  Embarked
0         3    0  22.000000      1      0   7.2500         0
1         1    1  38.000000      1      0  71.2833         1
2         3    1  26.000000      0      0   7.9250         0
3         1    1  35.000000      1      0  53.1000         0
4         3    0  35.000000      0      0   8.0500         0
..      ...  ...        ...    ...    ...      ...       ...
886       2    0  27.000000      0      0  13.0000         0
887       1    1  19.000000      0      0  30.0000         0
888       3    1  29.699118      1      2  23.4500         0
889       1    0  26.000000      0      0  30.0000         1
890       3    0  32.000000      0      0   7.7500         2

[891 rows x 7 columns]
```

In [30]:
```
print(Y)
```

```
0      0
1      1
2      1
3      1
4      0
      ..
886    0
887    1
888    0
889    1
890    0
Name: Survived, Length: 891, dtype: int64
```

Spliting the data into Training data & Test Data

In [31]:
```
X_train, X_test, Y_train, Y_test = train_test_split(X,Y, test_size=0.2, random_state=2
```

In [32]:
```
print(X.shape, X_train.shape, X_test.shape)
```

```
(891, 7) (712, 7) (179, 7)
```

Model Training

Logistic Regression

In [33]:
```
model = LogisticRegression()
```

In [34]:
```
# Train our model with Training data
model.fit(X_train, Y_train)
```

```
C:\Users\SHANTANU GARAIN\.conda\envs\CampusX\Lib\site-packages\sklearn\linear_model\_
logistic.py:460: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

Out[34]:
```
LogisticRegression()
```

Model Evaluation

Accuracy Score of Training data

In [35]:
```
X_train_prediction = model.predict(X_train)
```

In [36]:
```
print(X_train_prediction)
```

```
[0 1 0 0 0 0 0 1 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 0 1 0 0 1 0 1 1 0 0 1 0 1
 0 0 0 0 0 0 1 1 0 0 1 0 1 0 1 0 0 0 0 0 0 1 0 1 0 0 1 1 0 0 1 1 0 1 0 0 1
 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 0 1 0 0 0 1 1 1 0 1 0 0 0 0 0 1 0 0 0
 1 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 1 1 1 1 1 0 0 1 1 1 0 0 1 0 0
 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 1 0 1 1 1
 0 0 0 1 0 0 0 1 0 0 1 0 0 0 1 1 0 1 0 0 0 0 0 1 1 0 1 1 1 1 0 0 0 0 0 0 0
 0 1 0 0 1 1 1 0 0 1 0 1 1 1 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 1 0 0 0
 0 0 0 0 0 0 1 0 1 0 0 1 0 0 1 0 1 0 1 1 0 0 0 0 1 0 1 0 0 1 0 0 0 1 0 0 0
 0 1 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 1 0 0 0 1 0 1 0 0 0 0 0 0 1 1 0 1 1
 0 1 1 1 0 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 1 0 0 0 0 1 1 0 0 0 1 0 1 1 1 0 0
 0 0 1 0 0 0 1 1 0 0 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 1 1 1 0 1 1 0 0 0
 0 1 0 1 0 0 1 1 0 0 0 0 1 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 0 0 0 0 1 1 0 0
 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0 1 1 0 1 0 0 1 0 0 0 1 1 0 1 0
 0 0 0 0 1 0 0 1 0 1 1 0 0 1 0 0 1 0 0 0 1 0 1 1 0 0 1 1 0 1 0 1 1 1 0 1 0
 0 1 0 0 1 0 0 1 0 0 0 0 1 1 0 0 1 0 1 0 0 0 0 0 0 1 1 1 0 0 1 1 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0
 0 0 1 0 0 0 0 0 1 0 1 0 1 0 0 0 1 0 1 1 1 0 0 0 1 0 1 0 0 0 1 1 1 0 0 1 1
 0 0 0 1 0 1 0 0 0 0 0 1 1 0 1 1 1 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0
 1 0 0 1 0 1 0 0 0 1 1 1 1 1 0 0 1 1 0 1 1 1 1 0 0 0 1 1 0 0 1 0 0 0 0 0 0
 0 0 0 1 1 0 0 1 0]
```

In [37]:
```
training_data_accuracy = accuracy_score(Y_train, X_train_prediction)
print("Accuracy score of training data : ", training_data_accuracy)
```

```
Accuracy score of training data :  0.8075842696629213
```

Accuracy Score of our Test data:

In [38]:
```
X_test_prediction = model.predict(X_test)
```

In [39]:
```
print(X_test_prediction)
```

```
[0 0 1 0 0 0 0 0 0 0 1 1 0 0 1 0 0 1 0 1 1 0 1 0 1 1 0 0 0 0 0 0 0 1 1
 0 0 0 0 0 1 0 0 1 1 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 1 0 0 0 1 0 1 0
 1 0 0 0 1 0 1 0 0 0 1 1 0 0 1 0 0 0 0 0 0 1 0 1 0 0 1 0 1 1 0 1 1 0 0 0 0
 0 0 0 1 1 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 1 1 1 1 0 1 0 0
 0 1 0 0 0 0 1 0 0 1 1 0 1 0 0 0 1 1 0 0 1 0 0 1 1 1 0 0 0 0 0]
```

In [40]:
```
test_data_accuracy = accuracy_score(Y_test, X_test_prediction)
print("Accuracy score of test data : ", test_data_accuracy)
```

```
Accuracy score of test data :  0.7821229050279329
```

We see the training data and test data are predict well and it's not much difference