TITANIC SURVIVAL DATA PREDICTION

Importing All The Necessary Libraries

In [2]:
```
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  #splits data into training and testing model
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

Data Collection, Loading and Pre Processing

In [3]:
```
#Loading the CSV file into a Pandas Dataframe
data=pd.read_csv('C:/Users/NIdhi Aggarwal/Downloads/archive/Titanic-Dataset.csv')
```

In [4]:
```
#The number of rows and columns in the data
data.shape
```

Out[4]:
```
(891, 12)
```

In [5]:
```
#Knowing the columns with null or not null values
data.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
```

In [6]:
```
#Displaying the first 5 rows of the dataset.
data.head()
```

Out[6]:

	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

In [7]:
```
#Finding out the number of NULL values in each column
data.isnull().sum()
```

Out[7]:
```
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
```

Handling The Missing Values In The Dataset

In [8]:
```
#Dropping Cabin Column
data=data.drop(['Cabin'],axis=1)
```

In [9]:
```
#Fill null values in Age column with mean value
data['Age'].fillna(data['Age'].mean(),inplace=True)
```

In [10]:
```
print(data['Embarked'].mode())
```

```
0    S
Name: Embarked, dtype: object
```

In [11]:
```
#Fill null values of Embarked column with the mode value
data['Embarked'].fillna(data['Embarked'].mode()[0], inplace=True)
```

In [12]:
```
data.isnull().sum()
```

Out[12]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

In [13]:
```
data.describe()
```

Out[13]:

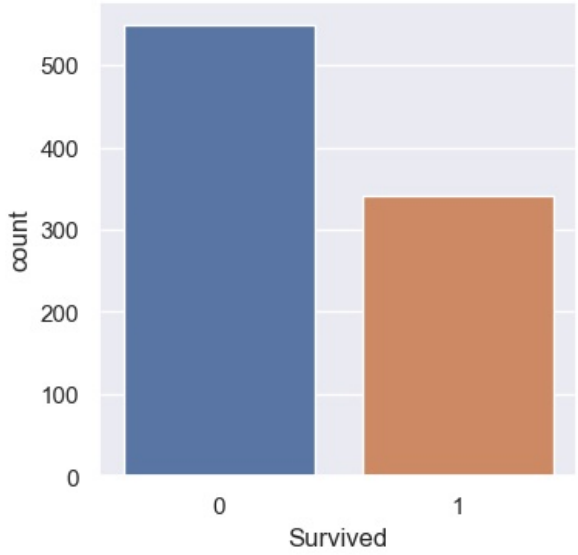
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	446.000000	0.383838	2.308642	29.699118	0.523008	0.381594	32.204208
std	257.353842	0.486592	0.836071	13.002015	1.102743	0.806057	49.693429
min	1.000000	0.000000	1.000000	0.420000	0.000000	0.000000	0.000000
25%	223.500000	0.000000	2.000000	22.000000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	29.699118	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	35.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

Data Visualization

In [14]:
```
#Number of people who survived or not survived
data['Survived'].value_counts()
```

Out[14]:
```
Survived
0    549
1    342
Name: count, dtype: int64
```

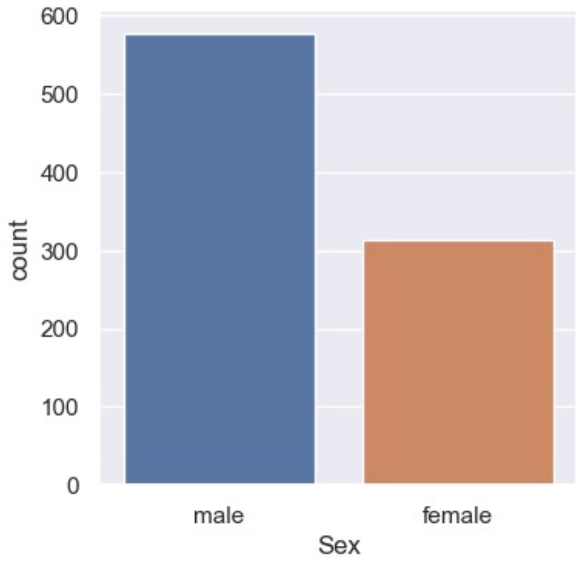
In [15]:
```
#Count plot for the number of people who survived or not
sns.set()
plt.figure(figsize=(4,4))
sns.countplot(x='Survived',data=data)
plt.show()
```



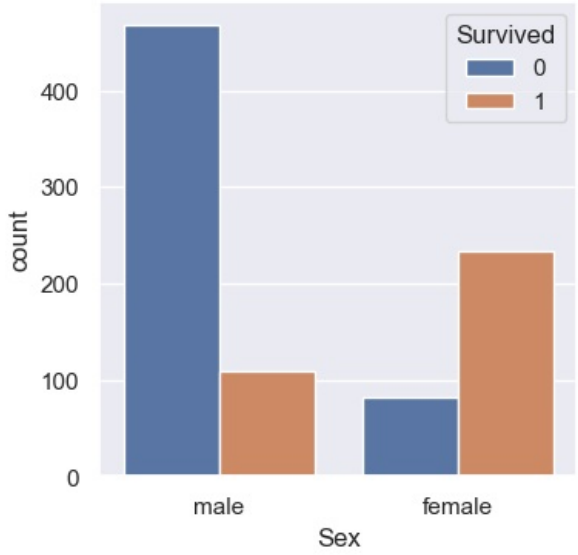
In [16]:
```
#Finding the number of males amnd females onboard
data['Sex'].value_counts()
```

Out[16]:
```
Sex
male      577
female    314
Name: count, dtype: int64
```

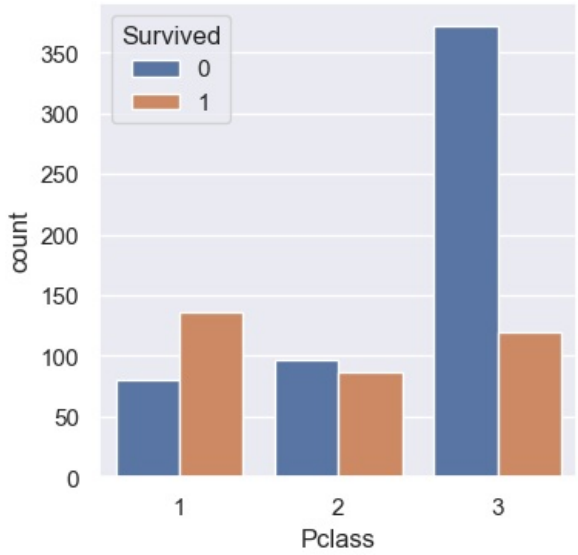
In [17]:
```
#Count plot for the males or females
plt.figure(figsize=(4,4))
sns.countplot(x='Sex',data=data)
plt.show()
```



In [18]:
```
#Count plot for Survived v/s Sex
sns.set()
plt.figure(figsize=(4,4))
sns.countplot(x='Sex',hue='Survived',data=data)
plt.show()
```



In [19]:
```
#Count plot foe Suvived v/s Class of ticket
plt.figure(figsize=(4,4))
sns.countplot(x='Pclass', hue='Survived',data=data)
plt.show()
```



In [20]:
```
data['Sex'].value_counts()
```

Out[20]:
```
Sex
male      577
female    314
Name: count, dtype: int64
```

In [21]:
```
data['Embarked'].value_counts()
```

Out[21]:
```
Embarked
S    646
C    168
Q     77
Name: count, dtype: int64
```

In [22]:
```
data.replace({'Sex':{'male':0,'female':1},'Embarked':{'S':0,'C':1,'Q':2}},inplace=True)
```

In [23]:
```
data.head()
```

Out[23]:

	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

In [24]:
```
A=data.drop(columns=['PassengerId','Ticket','Name','Survived'],axis=1)
B=data['Survived']
```

In [25]:
```
print(A)
```

```
     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 [26]:
```
print(B)
```

```
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
```

Splitting Data Into Training And Testing Data

In [27]:
```
a_train,a_test,b_train,b_test=train_test_split(A, B, test_size=0.2, random_state=2)
```

In [28]:
```
print(A.shape, a_train.shape, a_test.shape)
```

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

Training The Model

Using Logistic Regression And Evaluation Of Data

In [29]:
```
model=LogisticRegression()
```

In [30]:
```
model.fit(a_train,b_train)
```

```
C:\Users\NIdhi Aggarwal\anaconda3\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[30]:
```
▾ LogisticRegression
LogisticRegression()
```

In [31]:
```
a_train_predict=model.predict(a_train)
```

In [32]:
```
print(a_train_predict)
```

```
[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 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 0
 1 0 0 1 0 1 0 0 0 1 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 [33]:
```
train_accuracy_score=accuracy_score(b_train,a_train_predict)
```

In [34]:
```
print("The Accuracy Score for the Training Data is: ",train_accuracy_score)
```

```
The Accuracy Score for the Training Data is:  0.8075842696629213
```

In [35]:
```
a_test_predict=model.predict(a_test)
```

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

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

In [37]:
```
test_accuracy=accuracy_score(b_test,a_test_predict)
```

In [38]:
```
print("The Accuracy Score for the Testing Data is: ",test_accuracy)
```

```
The Accuracy Score for the Testing Data is:  0.7821229050279329
```

In [ ]:
```

```