About Titanic Dataset

The Titanic dataset is a famous dataset that contains information about passengers aboard the Titanic ship, which sank in 1912 after colliding with an iceberg. The dataset is often used in data science and machine learning education and competitions as a starting point for exploring data analysis and predictive modeling techniques.

The Titanic dataset contains information about **1309** passengers, including their age, gender, ticket class, cabin, port of embarkation, and whether they survived or not. The goal of many analyses and models built on the Titanic dataset is to predict whether a given passenger would have survived the disaster.

The variables in the Titanic dataset are as follows: **PassengerId: Unique identifier for each passenger Survived: Whether the passenger survived (0 = No, 1 = Yes) Pclass: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd) Name: Passenger name Sex: Passenger gender Age: Passenger age SibSp: Number of siblings/spouses aboard the Titanic Parch: Number of parents/children aboard the Titanic Ticket: Ticket number Fare: Passenger fare Cabin: Cabin number Embarked: Port of embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)** As mentioned earlier, the main objective of many analyses and models built on the Titanic dataset is to predict whether a given passenger would have survived the disaster, based on their demographic and travel information. This is a **binary classification problem**, where the target variable is **Survived** and the predictors are the other variables in the dataset.

Importing Libraries

```
import pandas as pd
```

Data Loading

```
data=pd.read_csv('/content/titanic.csv')
```

```
data.head(5)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.28
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.92
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.10
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05

```
data.tail(5)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fa
1304	1305	0	3	Spector, Mr. Woolf	male	NaN	0	0	A.5. 3236	8.05
1305	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.0	0	0	PC 17758	108.90
1306	1307	0	3	Saether, Mr. Simon Sivertsen	male	38.5	0	0	SOTON/O.Q. 3101262	7.25
1307	1308	0	3	Ware, Mr. Frederick	male	NaN	0	0	359309	8.05
1308	1309	0	3	Peter, Master. Michael J	male	NaN	1	1	2668	22.35

Data Dimention:- No. of Rows and Columns

```
data.shape
```

```
(1309, 12)
```

Double-click (or enter) to edit

Double-click (or enter) to edit

```
print("Number of Rows",data.shape[0])
print("Number of Columns",data.shape[1])
```

```
Number of Rows 1309
Number of Columns 12
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1309 entries, 0 to 1308
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  1309 non-null   int64  
 1   Survived     1309 non-null   int64  
 2   Pclass       1309 non-null   int64  
 3   Name         1309 non-null   object 
 4   Sex          1309 non-null   object 
 5   Age          1046 non-null   float64
 6   SibSp        1309 non-null   int64  
 7   Parch        1309 non-null   int64  
 8   Ticket       1309 non-null   object 
 9   Fare         1308 non-null   float64
 10  Cabin        295 non-null    object 
 11  Embarked     1307 non-null   object 
dtypes: float64(2), int64(5), object(5)
memory usage: 122.8+ KB
```

Get Overall Statistics About The Dataframe

```
data.describe(include='all')
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Tick
count	1309.000000	1309.000000	1309.000000	1309	1309	1046.000000	1309.000000	1309.000000	13
unique	NaN	NaN	NaN	1307	2	NaN	NaN	NaN	9
top	NaN	NaN	NaN	Connolly, Miss. Kate	male	NaN	NaN	NaN	C 23
freq	NaN	NaN	NaN	2	843	NaN	NaN	NaN	
mean	655.000000	0.377387	2.294882	NaN	NaN	29.881138	0.498854	0.385027	Na
std	378.020061	0.484918	0.837836	NaN	NaN	14.413493	1.041658	0.865560	Na
min	1.000000	0.000000	1.000000	NaN	NaN	0.170000	0.000000	0.000000	Na
25%	328.000000	0.000000	2.000000	NaN	NaN	21.000000	0.000000	0.000000	Na
50%	655.000000	0.000000	3.000000	NaN	NaN	28.000000	0.000000	0.000000	Na
75%	982.000000	1.000000	3.000000	NaN	NaN	39.000000	1.000000	0.000000	Na
max	1309.000000	1.000000	3.000000	NaN	NaN	80.000000	8.000000	9.000000	Na

Data Preprocessing & Data Cleaning

Data Filtering

```
data.columns
```

```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

```
data[['Name','Age']]
```

	Name	Age
0	Braund, Mr. Owen Harris	22.0
1	Cumings, Mrs. John Bradley (Florence Briggs Th...	38.0
2	Heikkinen, Miss. Laina	26.0
3	Futrelle, Mrs. Jacques Heath (Lily May Peel)	35.0
4	Allen, Mr. William Henry	35.0

```
sum(data['Sex']=='male')
```

```
843
```

```
data[data['Sex']=='male'].head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.250(
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.050(
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.458:
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.862:
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.075(

```
sum(data['Survived']==1)
```

```
494
```

Check Missing (Null) Values In The *Dataset*

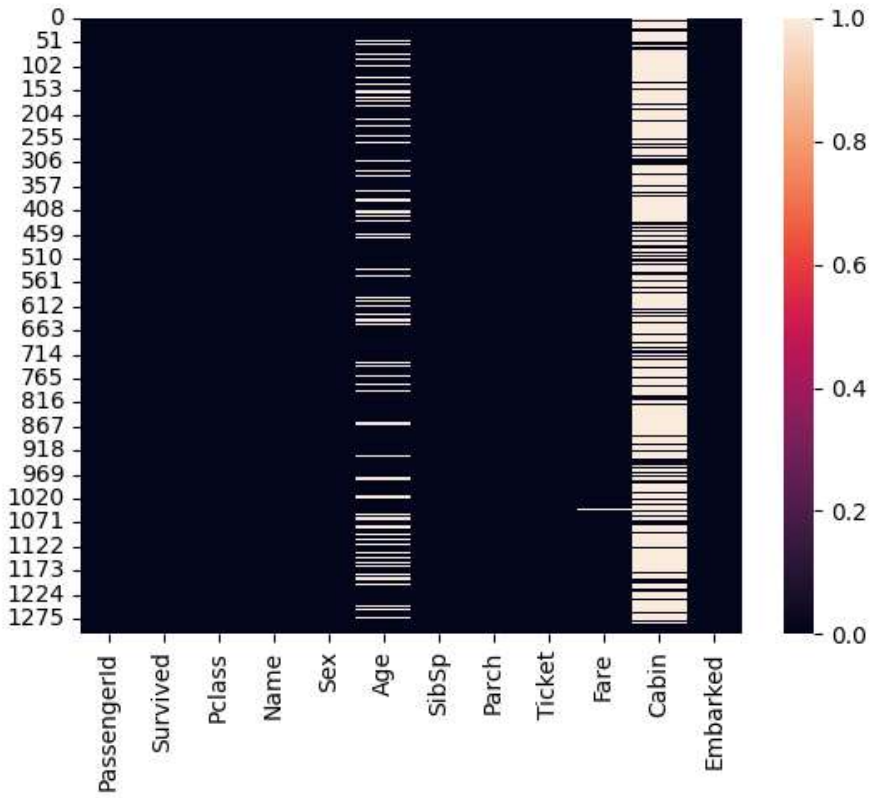
```
data.isnull().sum()
```

```
PassengerId       0
Survived          0
Pclass            0
Name              0
Sex               0
Age             263
SibSp             0
Parch             0
Ticket            0
Fare              1
Cabin          1014
Embarked          2
dtype: int64
```

```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
sns.heatmap(data.isnull())
```

```
<Axes: >
```




```
per_missing = data.isnull().sum() * 100 / len(data)
```

Drop the Column

```
data.drop('Cabin', axis=1,inplace=True)
```

```
data.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            263
SibSp            0
Parch            0
Ticket           0
Fare             1
Embarked         2
dtype: int64
```

Handle Missing Values

```
data['Embarked'].mode()
```

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

```
data['Embarked'].fillna('S',inplace=True)
```

```
data.isnull().sum()
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            263
SibSp            0
Parch            0
Ticket           0
Fare             1
Embarked         0
dtype: int64
```

```
data['Age']
```

```
0       22.0
1       38.0
2       26.0
3       35.0
4       35.0
        ...
1304     NaN
1305    39.0
1306    38.5
1307     NaN
1308     NaN
Name: Age, Length: 1309, dtype: float64
```

```
data['Age'].fillna(data['Age'].mean(), inplace = True)
```

```
data.isnull().sum()
```

```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           1
Embarked       0
dtype: int64
```

```
data.head()
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticke
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 2117
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 1759
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2 310128
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	11380
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	37345

```
data['Sex'].unique()
```

```
array(['male', 'female'], dtype=object)
```

```
data['Gender']=data['Sex'].map({'male':1, 'female':0})
```

```
data.head(5)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Far
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.250
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.283
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.100
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.050

Data Encoding

```
x=data['Sex'].map({'male':1, 'female':0})
```

```
data['Embarked'].unique()
```

```
array(['S', 'C', 'Q'], dtype=object)
```

```
pd.get_dummies(data,columns=['Embarked'])
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ge
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.2500	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.2833	
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.9250	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.1000	
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.0500	
...	...	...	...	...	...	...	...	...	...	...	...
1304	1305	0	3	Spector, Mr. Woolf	male	29.881138	0	0	A.5. 3236	8.0500	
1305	1306	1	1	Oliva y Ocana, Dona. Fermina	female	39.000000	0	0	PC 17758	108.9000	
1306	1307	0	3	Saether, Mr. Simon Sivertsen	male	38.500000	0	0	SOTON/O.Q. 3101262	7.2500	
1307	1308	0	3	Ware, Mr. Frederick	male	29.881138	0	0	359309	8.0500	
1308	1309	0	3	Peter, Master. Michael J	male	29.881138	1	1	2668	22.3583	

1309 rows × 14 columns

```
data1=pd.get_dummies(data,columns=['Embarked'],drop_first=True)
```

```
data1.head(1)
```

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Gender	Emb
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25	1	

Visual Analysis

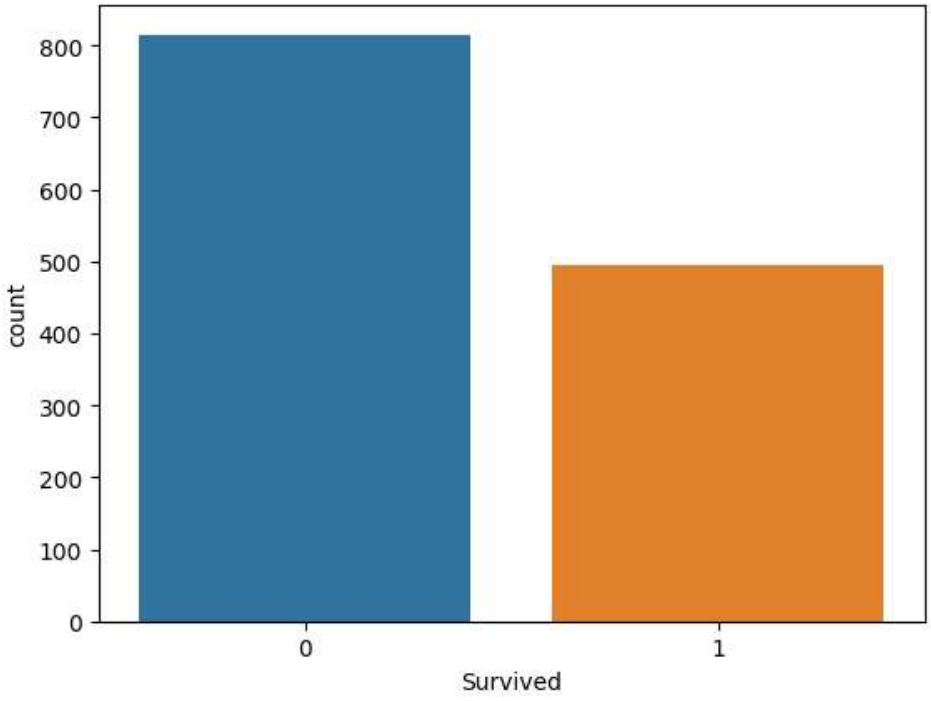
How Many People Survived And How Many Died?

```
data['Survived'].value_counts()
```

```
0    815
1    494
Name: Survived, dtype: int64
```

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.countplot(x='Survived',data=data)
```

```
<Axes: xlabel='Survived', ylabel='count'>
```



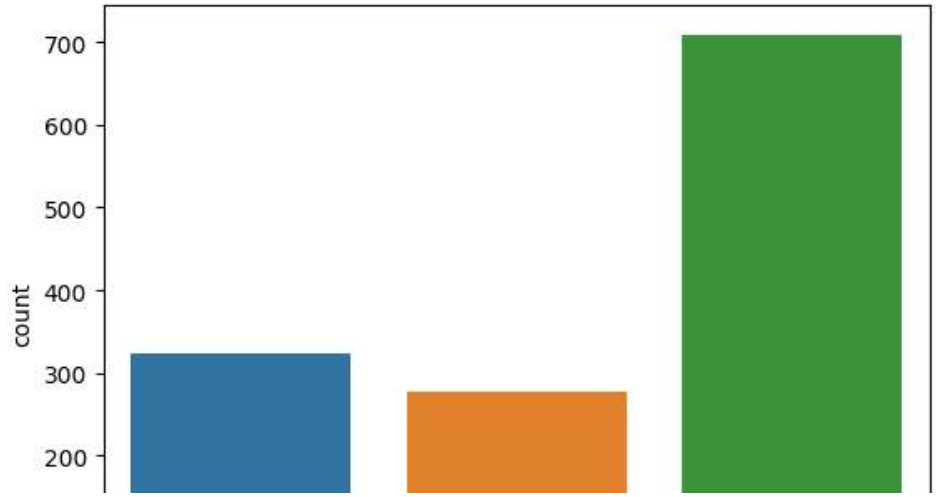
How Many Passengers Were In First Class, Second Class, and Third Class?

```
data['Pclass'].value_counts()
```

```
3    709
1    323
2    277
Name: Pclass, dtype: int64
```

```
sns.countplot(x='Pclass', data=data)
```

```
<Axes: xlabel='Pclass', ylabel='count'>
```



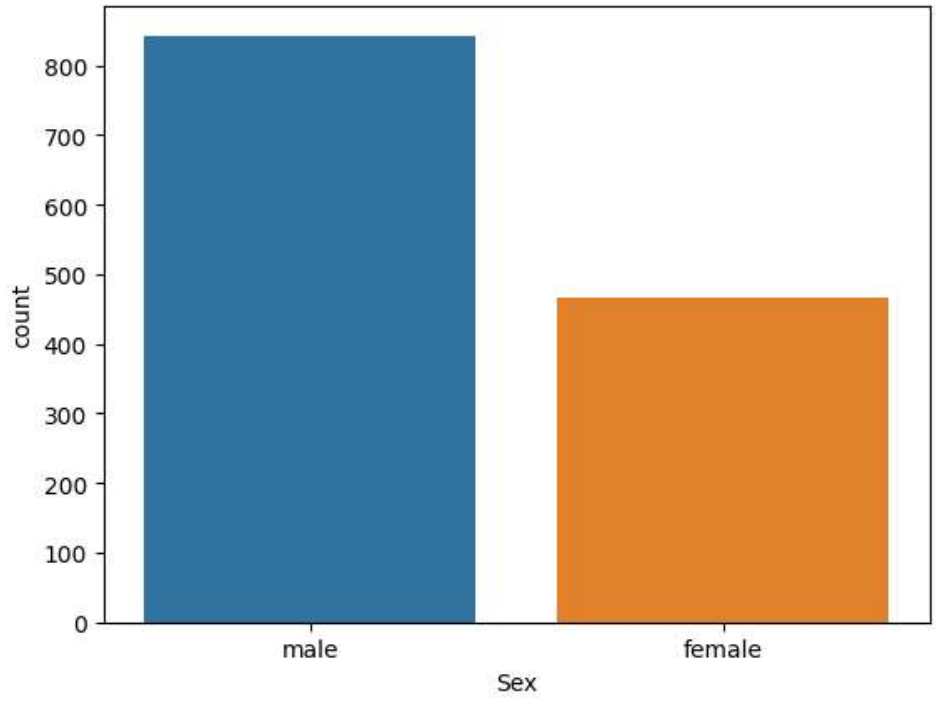
Number of Male And Female Passengers

```
data['Sex'].value_counts()
```

```
male      843
female    466
Name: Sex, dtype: int64
```

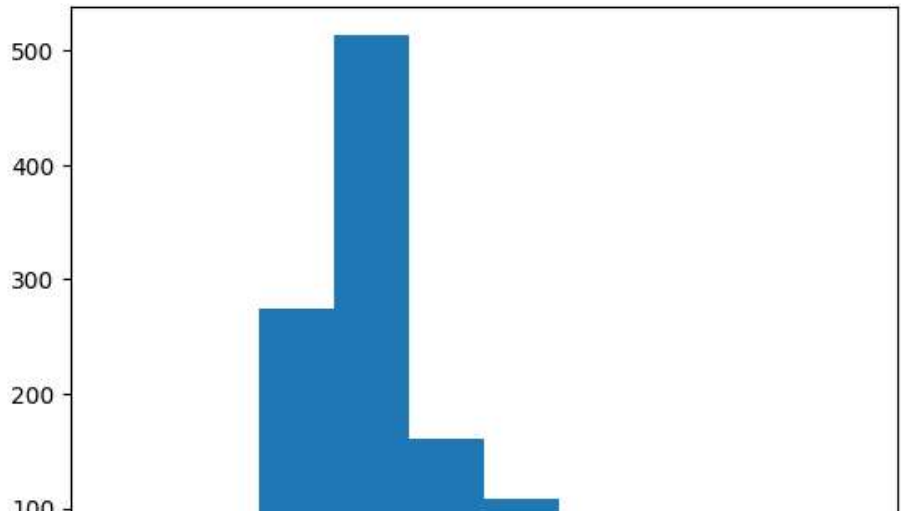
```
sns.countplot(x ='Sex', data = data)
```

```
<Axes: xlabel='Sex', ylabel='count'>
```

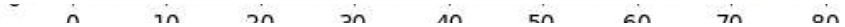



```
plt.hist(data['Age'])
```

```
(array([ 72.,  62., 274., 513., 161., 108.,  65.,  41.,  10.,   3.]),
 array([ 0.17 ,  8.153, 16.136, 24.119, 32.102, 40.085, 48.068, 56.051,
        64.034, 72.017, 80.   ]),
 <BarContainer object of 10 artists>)
```



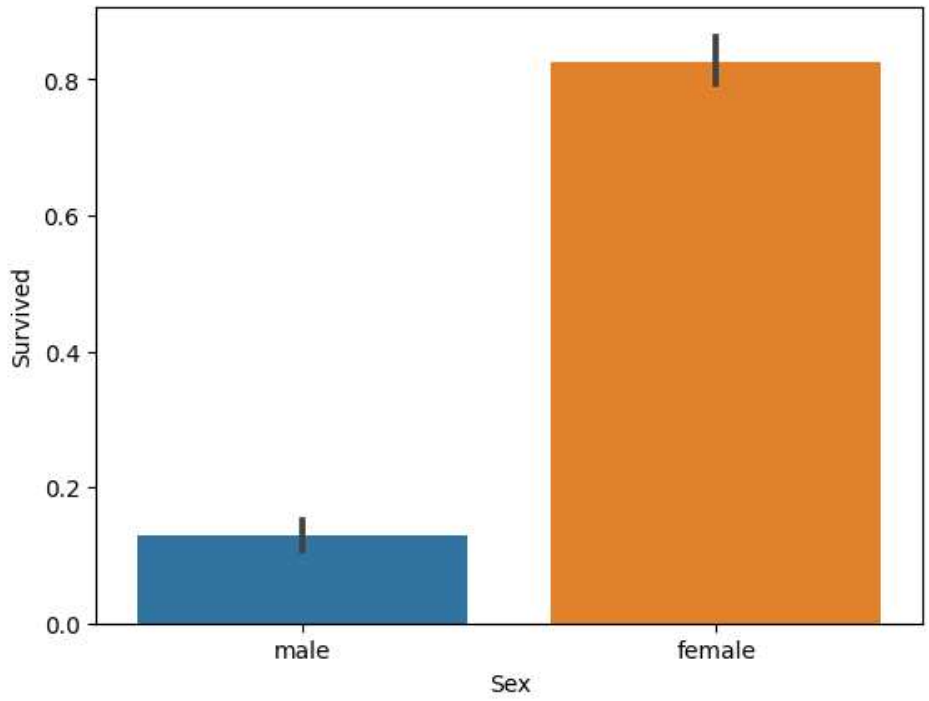
12. Bivariate Analysis



How Has Better Chance of Survival Male or Female?

```
sns.barplot(x='Sex',y='Survived',data=data)
```

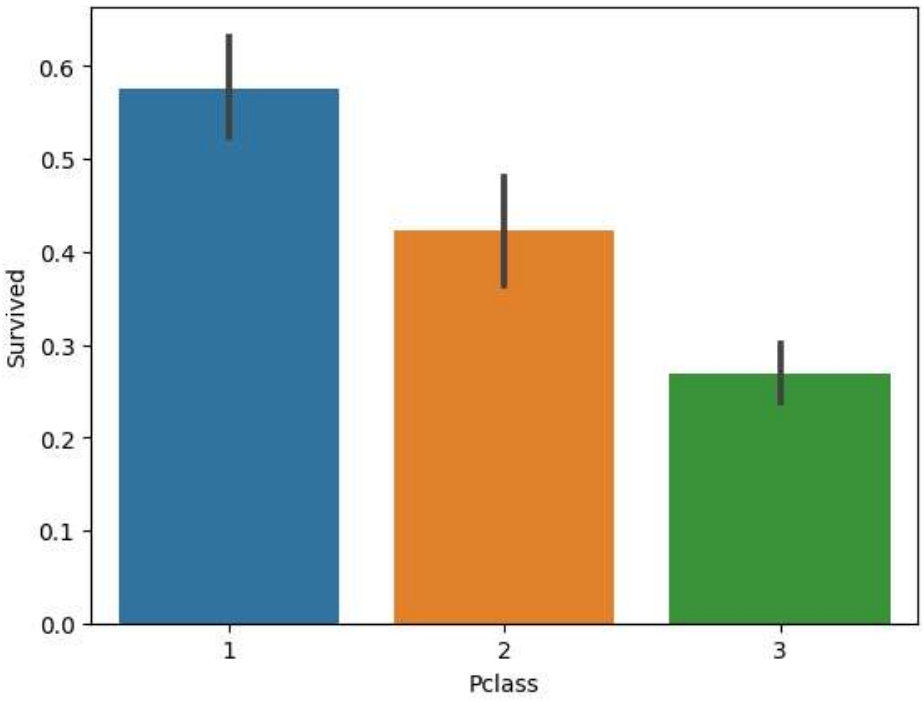
```
<Axes: xlabel='Sex', ylabel='Survived'>
```



Which Passenger Class Has Better Chance of Surviva(First, Second, Or Third Class)?

```
sns.barplot(x="Pclass", y="Survived",data=data)
```

```
<Axes: xlabel='Pclass', ylabel='Survived'>
```



✓ 1s completed at 10:41 PM