In [1]:
```python
import pandas as pd
```

In [2]:
```python
data = pd.read_csv('train.csv')
```

In [3]:
```python
data.head()
```

Out[3]:

	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 [4]:
```python
data.tail(3)
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN	Q

In [5]:
```python
data.shape
```

Out[5]:
```
(891, 12)
```

In [6]:
```python
print('Number of Rows: ', data.shape[0])
print('Number of columns: ', data.shape[1])
```

```
Number of Rows:  891
Number of columns:  12
```

In [7]:
```python
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 [8]:
```python
data.describe(include='all')
```

Out[8]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
count	891.000000	891.000000	891.000000	891	891	714.000000	891.000000	891.000000	891	891.000000	204	889
unique	NaN	NaN	NaN	891	2	NaN	NaN	NaN	681	NaN	147	3
top	NaN	NaN	NaN	Perreault, Miss. Anne	male	NaN	NaN	NaN	CA. 2343	NaN	G6	S
freq	NaN	NaN	NaN	1	577	NaN	NaN	NaN	7	NaN	4	644
mean	446.000000	0.383838	2.308642	NaN	NaN	29.699118	0.523008	0.381594	NaN	32.204208	NaN	NaN
std	257.353842	0.486592	0.836071	NaN	NaN	14.526497	1.102743	0.806057	NaN	49.693429	NaN	NaN
min	1.000000	0.000000	1.000000	NaN	NaN	0.420000	0.000000	0.000000	NaN	0.000000	NaN	NaN
25%	223.500000	0.000000	2.000000	NaN	NaN	20.125000	0.000000	0.000000	NaN	7.910400	NaN	NaN
50%	446.000000	0.000000	3.000000	NaN	NaN	28.000000	0.000000	0.000000	NaN	14.454200	NaN	NaN
75%	668.500000	1.000000	3.000000	NaN	NaN	38.000000	1.000000	0.000000	NaN	31.000000	NaN	NaN
max	891.000000	1.000000	3.000000	NaN	NaN	80.000000	8.000000	6.000000	NaN	512.329200	NaN	NaN

Data Filtering

In [9]:
```python
data.columns
```

Out[9]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [10]:
```python
data[['Name', 'Age']]
```

Out[10]:

	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
...	...	...
886	Montvila, Rev. Juozas	27.0
887	Graham, Miss. Margaret Edith	19.0
888	Johnston, Miss. Catherine Helen "Carrie"	NaN
889	Behr, Mr. Karl Howell	26.0
890	Dooley, Mr. Patrick	32.0

891 rows × 2 columns

In [11]:
```python
sum(data['Sex'] == 'male')
```

Out[11]:
```
577
```

In [12]:
```python
data[data['Sex'] == 'male'].head()
```

Out[12]:

	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
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	NaN	Q
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	NaN	S

In [13]:
```python
data.columns
```

Out[13]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [14]:
```python
sum(data['Survived'] ==1)
```

Out[14]:
```
342
```

In [15]:
```python
data[data['Survived'] ==1]
```

Out[15]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
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
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	NaN	S
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	NaN	C
...	...	...	...	...	...	...	...	...	...	...	...	...
875	876	1	3	Najib, Miss. Adele Kiamie "Jane"	female	15.0	0	0	2667	7.2250	NaN	C
879	880	1	1	Potter, Mrs. Thomas Jr (Lily Alexenia Wilson)	female	56.0	0	1	11767	83.1583	C50	C
880	881	1	2	Shelley, Mrs. William (Imanita Parrish Hall)	female	25.0	0	1	230433	26.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C

342 rows × 12 columns

Check Null Values In The Dataset

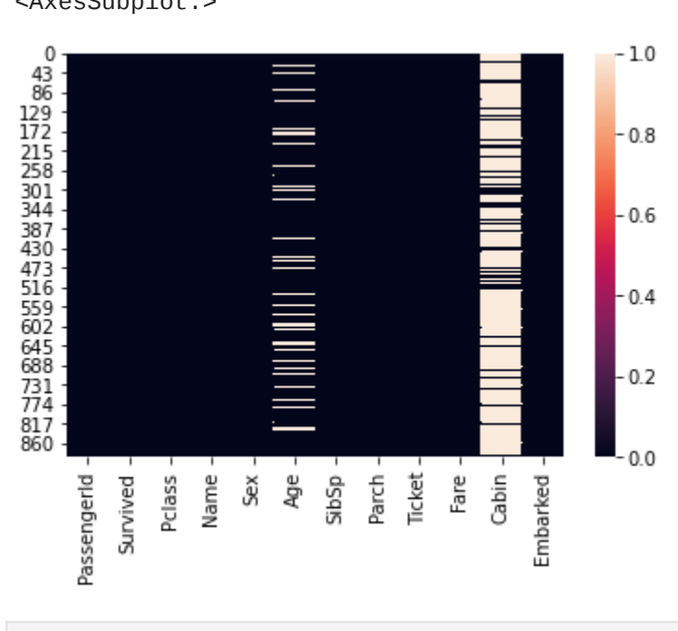
In [16]:
```python
data.isnull().sum()
```

Out[16]:
```
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
```

In [17]:
```python
import seaborn as sns
import matplotlib.pyplot as plt
```

In [18]:
```python
sns.heatmap(data.isnull())
```

Out[18]:
```
<AxesSubplot:>
```



In [19]:
```python
per_missing = data.isnull().sum()*100 / len(data)
per_missing
```

Out[19]:
```
PassengerId     0.000000
Survived        0.000000
Pclass          0.000000
Name            0.000000
Sex             0.000000
Age            19.865320
SibSp           0.000000
Parch           0.000000
Ticket          0.000000
Fare            0.000000
Cabin          77.104377
Embarked        0.224467
dtype: float64
```

Drop tha column

In [20]:
```python
data.drop('Cabin', axis=1, inplace=True)
```

In [21]:
```python
data.isnull().sum()
```

Out[21]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Embarked         2
dtype: int64
```

Handle missing values

In [22]:
```python
data.columns
```

Out[22]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Embarked'],
      dtype='object')
```

In [23]:
```python
data['Embarked'].mode()
```

Out[23]:
```
0    S
dtype: object
```

In [24]:
```python
data['Embarked'].fillna('S', inplace=True)
```

In [25]:
```python
data.isnull().sum()
```

Out[25]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Embarked         0
dtype: int64
```

In [26]:
```python
data['Age']
```

Out[26]:
```
0      22.0
1      38.0
2      26.0
3      35.0
4      35.0
       ...
886    27.0
887    19.0
888     NaN
889    26.0
890    32.0
Name: Age, Length: 891, dtype: float64
```

In [27]:
```python
data['Age'].fillna(data['Age'].mean(), inplace=True)
```

In [28]:
```python
data.isnull().sum()
```

Out[28]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Embarked       0
dtype: int64
```

Categorical Data Encoding

In [29]:
```python
data.head()
```

Out[29]:

	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

In [30]:
```python
data['Sex'].unique()
```

Out[30]:
```
array(['male', 'female'], dtype=object)
```

In [31]:
```python
data['Sex'].map({'male': 1, 'female': 0})
```

Out[31]:
```
0      1
1      0
2      0
3      0
4      1
      ..
886    1
887    0
888    0
889    1
890    1
Name: Sex, Length: 891, dtype: int64
```

In [32]:
```python
data['Gender'] = data['Sex'].map({'male': 1, 'female': 0})
```

In [33]:
```python
data.head(1)
```

Out[33]:

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

In [34]:
```python
x=data['Sex'].map({'male': 1, 'female': 0})
```

In [35]:
```python
data.insert(5, 'Gender_new', x)
```

In [36]:
```python
data.head(1)
```

Out[36]:

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

In [38]:
```python
data['Embarked'].unique()
```

Out[38]:
```
array(['S', 'C', 'Q'], dtype=object)
```

In [40]:
```python
pd.get_dummies(data, columns=['Embarked'])
```

Out[40]:

	PassengerId	Survived	Pclass	Name	Sex	Gender_new	Age	SibSp	Parch	Ticket	Fare	Gender	Embarked_C	Embarked_Q	Embarked_S
0	1	0	3	Braund, Mr. Owen Harris	male	1	22.000000	1	0	A/5 21171	7.2500	1	0	0	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	0	38.000000	1	0	PC 17599	71.2833	0	1	0	0
2	3	1	3	Heikkinen, Miss. Laina	female	0	26.000000	0	0	STON/O2. 3101282	7.9250	0	0	0	1
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	0	35.000000	1	0	113803	53.1000	0	0	0	1
4	5	0	3	Allen, Mr. William Henry	male	1	35.000000	0	0	373450	8.0500	1	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	1	27.000000	0	0	211536	13.0000	1	0	0	1
887	888	1	1	Graham, Miss. Margaret Edith	female	0	19.000000	0	0	112053	30.0000	0	0	0	1
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	0	29.699118	1	2	W./C. 6607	23.4500	0	0	0	1
889	890	1	1	Behr, Mr. Karl Howell	male	1	26.000000	0	0	111369	30.0000	1	1	0	0
890	891	0	3	Dooley, Mr. Patrick	male	1	32.000000	0	0	370376	7.7500	1	0	1	0

891 rows × 15 columns

In [42]:
```python
data1 = pd.get_dummies(data, columns=['Embarked'], drop_first= True)
```

In [44]:
```python
data1.head(1)
```

Out[44]:

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

What is Univariate Analysis?

In [45]:
```python
data.columns
```

Out[45]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Gender_new', 'Age',
       'SibSp', 'Parch', 'Ticket', 'Fare', 'Embarked', 'Gender'],
      dtype='object')
```

In [ ]:

How Many People Survived And How Many Died?

In [48]:
```python
data['Survived'].value_counts()
```

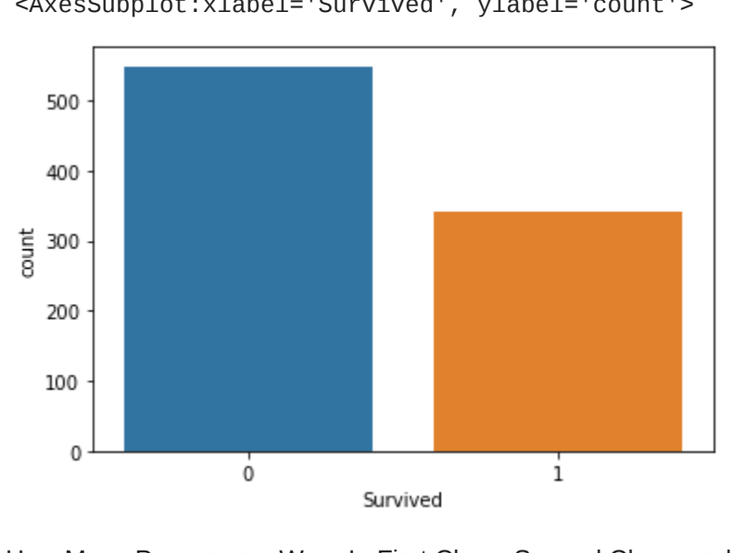
Out[48]:
```
0    549
1    342
Name: Survived, dtype: int64
```

In [50]:
```python
import seaborn as sns
import matplotlib.pyplot as plt
```

In [51]:
```python
sns.countplot(data['Survived'])
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other ar
guments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[51]:
```
<AxesSubplot:xlabel='Survived', ylabel='count'>
```



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

In [52]:
```python
data.columns
```

Out[52]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Gender_new', 'Age',
       'SibSp', 'Parch', 'Ticket', 'Fare', 'Embarked', 'Gender'],
      dtype='object')
```

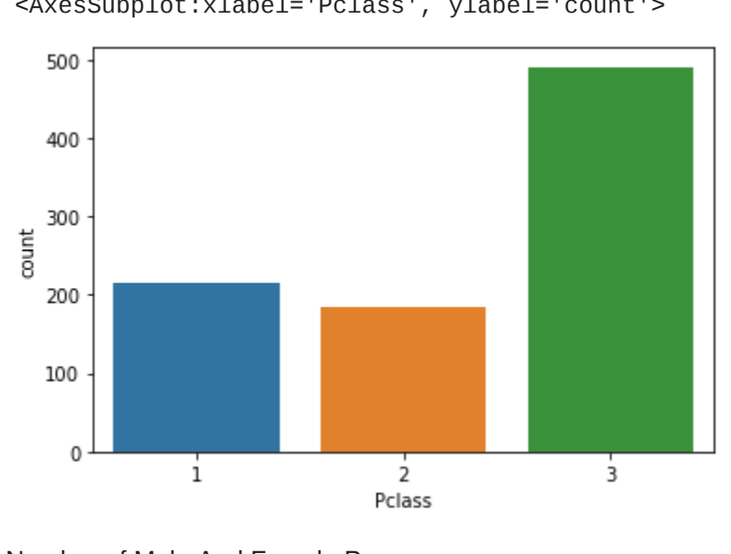
In [54]:
```python
data['Pclass'].value_counts()
```

Out[54]:
```
3    491
1    216
2    184
Name: Pclass, dtype: int64
```

In [55]:
```python
sns.countplot(data['Pclass'])
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other ar
guments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[55]:
```
<AxesSubplot:xlabel='Pclass', ylabel='count'>
```



Number of Male And Female Passengers

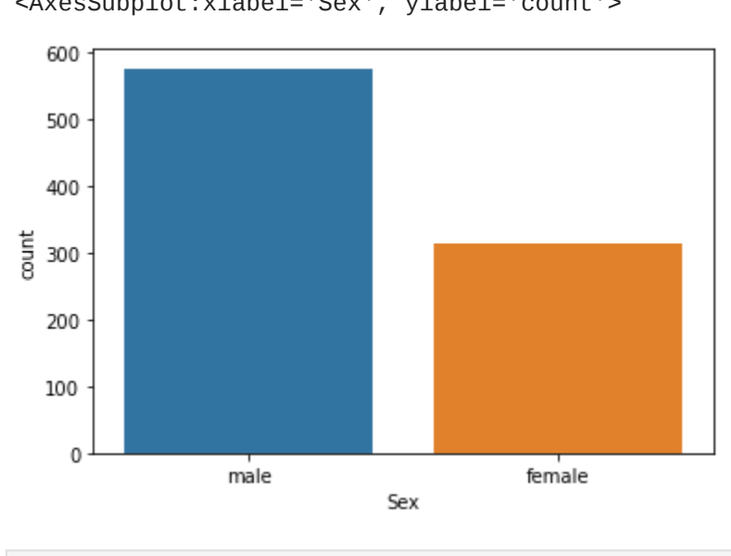
In [57]:
```python
data['Sex'].value_counts()
```

Out[57]:
```
male      577
female    314
Name: Sex, dtype: int64
```

In [58]:
```python
sns.countplot(data['Sex'])
```

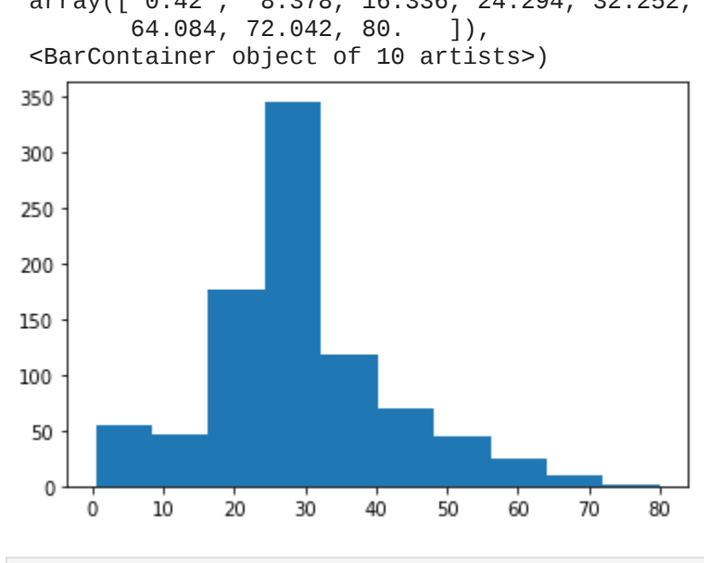
```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other ar
guments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[58]:
```
<AxesSubplot:xlabel='Sex', ylabel='count'>
```



In [59]:
```python
plt.hist(data['Age'])
```

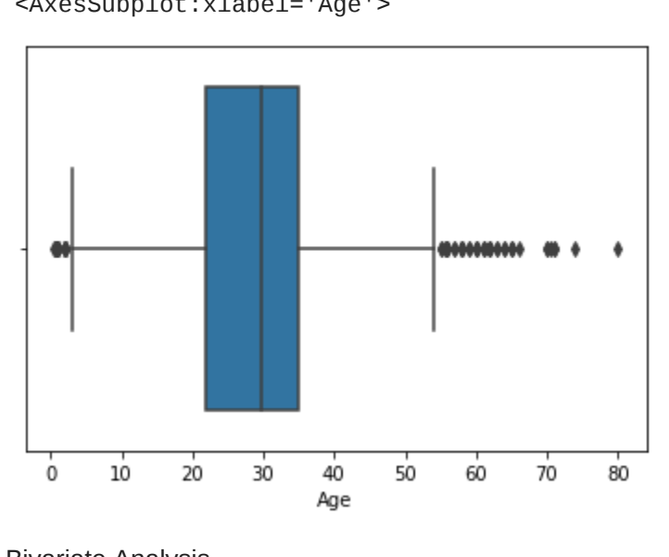
Out[59]:
```
(array([ 54.,  46., 177., 346., 118.,  70.,  45.,  24.,   9.,   2.]),
 array([ 0.42 ,  8.378, 16.336, 24.294, 32.252, 40.21 , 48.168, 56.126,
        64.084, 72.042, 80.   ]),
 <BarContainer object of 10 artists>)
```



In [61]:
```python
sns.boxplot(data['Age'], orient='v')
```

```
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other ar
guments without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
C:\ProgramData\Anaconda3\lib\site-packages\seaborn\_core.py:1319: UserWarning: Vertical orientation ignored with only `x` specified.
  warnings.warn(single_var_warning.format("Vertical", "x"))
```

Out[61]:
```
<AxesSubplot:xlabel='Age'>
```



Bivariate Analysis

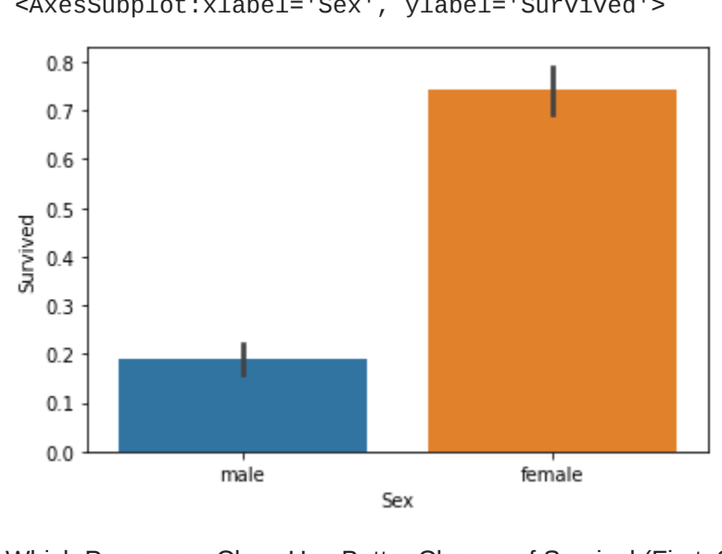
In [62]:
```python
data.columns
```

Out[62]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Gender_new', 'Age',
       'SibSp', 'Parch', 'Ticket', 'Fare', 'Embarked', 'Gender'],
      dtype='object')
```

How Has Better Chance of Survival Male or Female?

In [64]:
```python
sns.barplot(x='Sex', y='Survived', data=data)
```

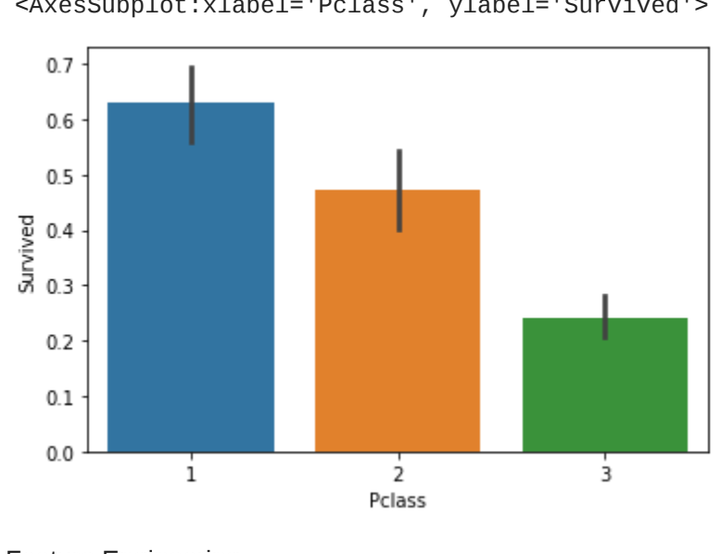
Out[64]:
```
<AxesSubplot:xlabel='Sex', ylabel='Survived'>
```



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

In [65]:
```python
sns.barplot(x='Pclass', y='Survived', data=data)
```

Out[65]:
```
<AxesSubplot:xlabel='Pclass', ylabel='Survived'>
```



Feature Engineering

In [66]:
```python
data.columns
```

Out[66]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Gender_new', 'Age',
       'SibSp', 'Parch', 'Ticket', 'Fare', 'Embarked', 'Gender'],
      dtype='object')
```

In [67]:
```python
data['Family_size']=data['SibSp'] + data['Parch']
```

In [68]:
```python
data.head(1)
```

Out[68]:

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

In [69]:
```python
data.columns
```

Out[69]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Gender_new', 'Age',
       'SibSp', 'Parch', 'Ticket', 'Fare', 'Embarked', 'Gender',
       'Family_size'],
      dtype='object')
```

In [70]:
```python
data['Fare_Per_Person']=data['Fare'] / (data['Family_size'] + 1)
```

In [71]:
```python
data.head(1)
```

Out[71]:

	PassengerId	Survived	Pclass	Name	Sex	Gender_new	Age	SibSp	Parch	Ticket	Fare	Embarked	Gender	Family_size	Fare_Per_Person
0	1	0	3	Braund, Mr. Owen Harris	male	1	22.0	1	0	A/5 21171	7.25	S	1	1	3.625

In [ ]: