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

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

In [3]:
```
data
```

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
...	...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	NaN	S
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000	B42	S
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500	NaN	S
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000	C148	C
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	NaN	Q

891 rows × 12 columns

In [4]:
```
data.head()
```

Out[4]:

	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 [5]:
```
data.tail(3)
```

Out[5]:

	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 [6]:
```
data.shape
```

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

In [7]:
```
print('Number of Columns', data.shape[0])
print('Number of Rows', data.shape[1])
```

```
Number of Columns 891
Number of Rows 12
```

In [8]:
```
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 [9]:
```
data.describe(include='all')
```

Out[9]:

	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	Kent, Mr. Edward Austin	male	NaN	NaN	NaN	1601	NaN	C23 C25 C27	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 [10]:
```
data.columns
```

Out[10]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [11]:
```
data[['Name', 'Age']]
```

Out[11]:

	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 [12]:
```
sum(data['Sex'] == 'male')
```

Out[12]:
```
577
```

In [13]:
```
data[data['Sex'] == 'male'].head()
```

Out[13]:

	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 [14]:
```
data.columns
```

Out[14]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Cabin', 'Embarked'],
      dtype='object')
```

In [15]:
```
sum(data['Survived'] == 1)
```

Out[15]:
```
342
```

In [16]:
```
data[data['Survived'] == 1]
```

Out[16]:

	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

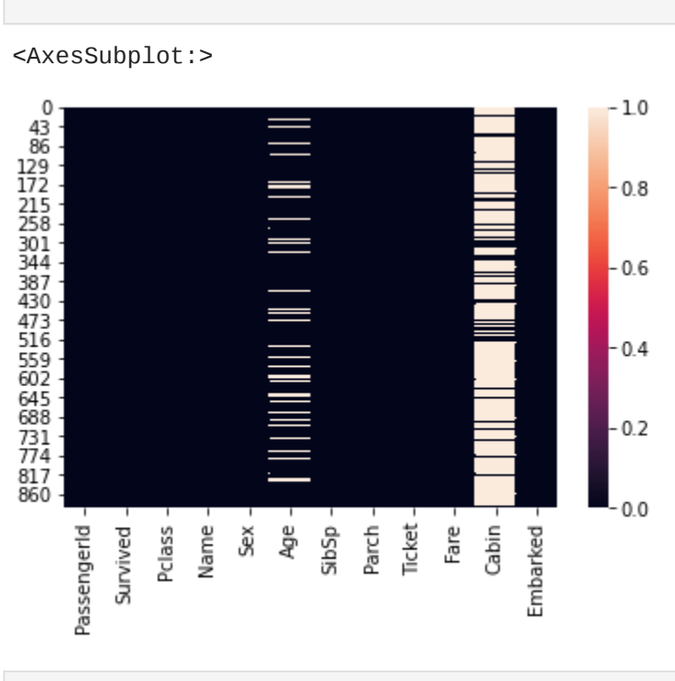
In [17]:
```
data.isnull().sum()
```

Out[17]:
```
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 [20]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [21]:
```
sns.heatmap(data.isnull())
```

Out[21]:
```
<AxesSubplot:>
```



In [23]:
```
per_missing = data.isnull().sum() * 100 / len(data)
per_missing
```

Out[23]:
```
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 the Column

In [26]:
```
data.drop('Cabin', axis=1, inplace=True)
```

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

Out[28]:
```
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 [29]:
```
data.columns
```

Out[29]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Age', 'SibSp',
       'Parch', 'Ticket', 'Fare', 'Embarked'],
      dtype='object')
```

In [32]:
```
data['Embarked'].mode()
```

Out[32]:
```
0    S
dtype: object
```

In [33]:
```
data['Embarked'].fillna('S', inplace=True)
```

In [34]:
```
data.isnull().sum()
```

Out[34]:
```
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 [37]:
```
data['Age'].fillna(data['Age'].mean(), inplace=True)
```

In [36]:
```
data['Age'].mean()
```

Out[36]:
```
29.69911764705882
```

In [38]:
```
data.isnull().sum()
```

Out[38]:
```
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 [39]:
```
data.head()
```

Out[39]:

	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 [41]:
```
data['Sex'].unique()
```

Out[41]:
```
array(['male', 'female'], dtype=object)
```

In [49]:
```
data['Gender'] = data['Sex'].map({'male': 1, 'female': 0})
```

In [52]:
```
data.head()
```

Out[52]:

	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	7.2500	S	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C	0
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S	0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S	0
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S	1

In [53]:
```
x= data['Sex'].map({'male': 1, 'female': 0})
```

In [54]:
```
data.insert(5, 'Gender_New', x)
```

In [55]:
```
data.head()
```

Out[55]:

	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	7.2500	S	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	0	38.0	1	0	PC 17599	71.2833	C	0
2	3	1	3	Heikkinen, Miss. Laina	female	0	26.0	0	0	STON/O2. 3101282	7.9250	S	0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	0	35.0	1	0	113803	53.1000	S	0
4	5	0	3	Allen, Mr. William Henry	male	1	35.0	0	0	373450	8.0500	S	1

In [56]:
```
data['Embarked'].unique()
```

Out[56]:
```
array(['S', 'C', 'Q'], dtype=object)
```

In [57]:
```
pd.get_dummies(data, columns=['Embarked'])
```

Out[57]:

	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 [59]:
```
data1 = pd.get_dummies(data, columns=['Embarked'], drop_first=True)
```

In [61]:
```
data1.head(1)
```

Out[61]:

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

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

How Many People Survived And How Many Died?

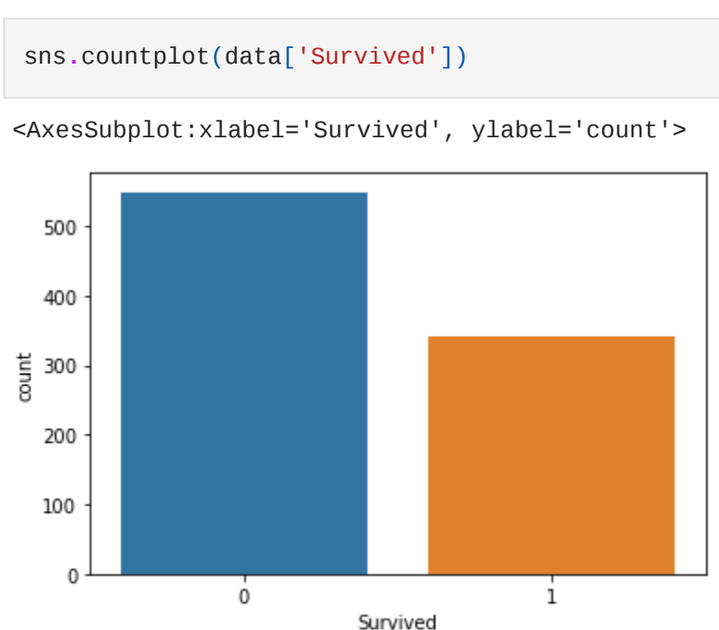
In [65]:
```
data['Survived'].value_counts()
```

Out[65]:
```
0    549
1    342
Name: Survived, dtype: int64
```

In [69]:
```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [72]:
```
sns.countplot(data['Survived'])
```

Out[72]:
```
<AxesSubplot:xlabel='Survived', ylabel='count'>
```



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

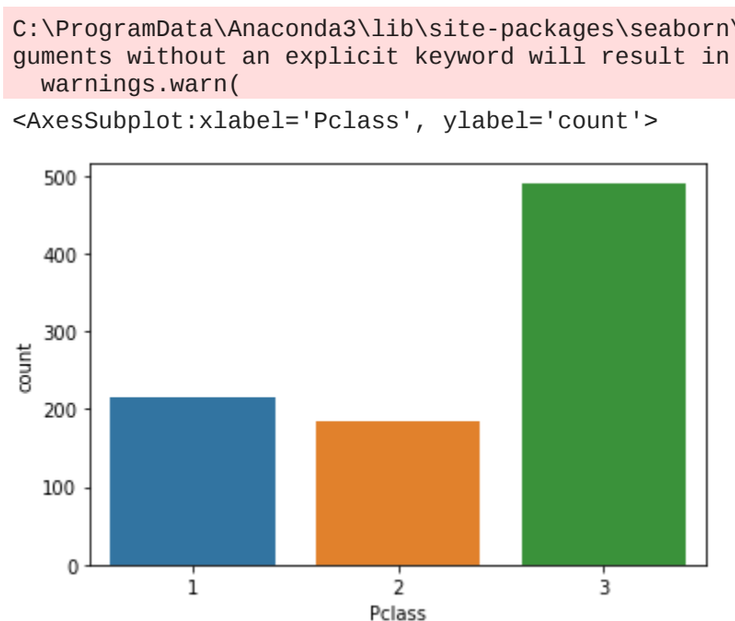
In [73]:
```
data['Pclass'].value_counts()
```

Out[73]:
```
3    491
1    216
2    184
Name: Pclass, dtype: int64
```

In [74]:
```
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[74]:
```
<AxesSubplot:xlabel='Pclass', ylabel='count'>
```



Number of Male And Female Passengers

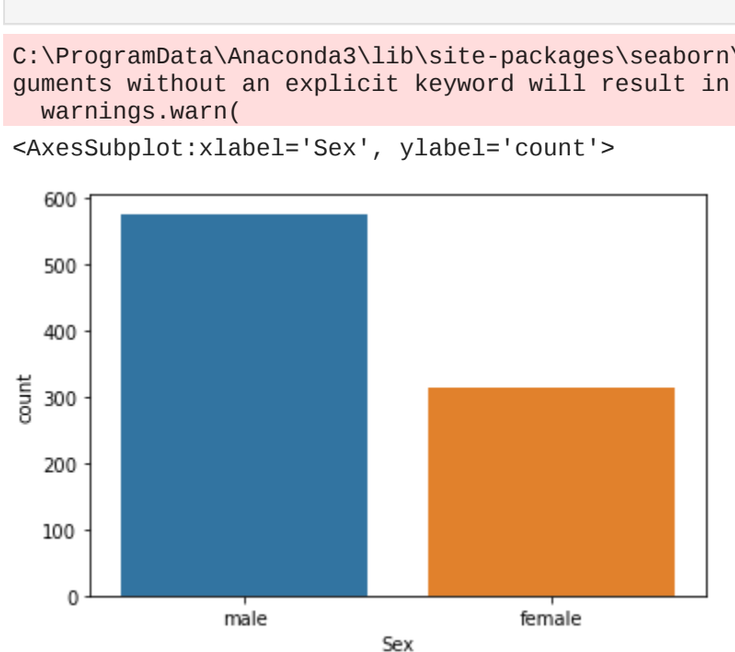
In [76]:
```
data['Sex'].value_counts()
```

Out[76]:
```
male      577
female    314
Name: Sex, dtype: int64
```

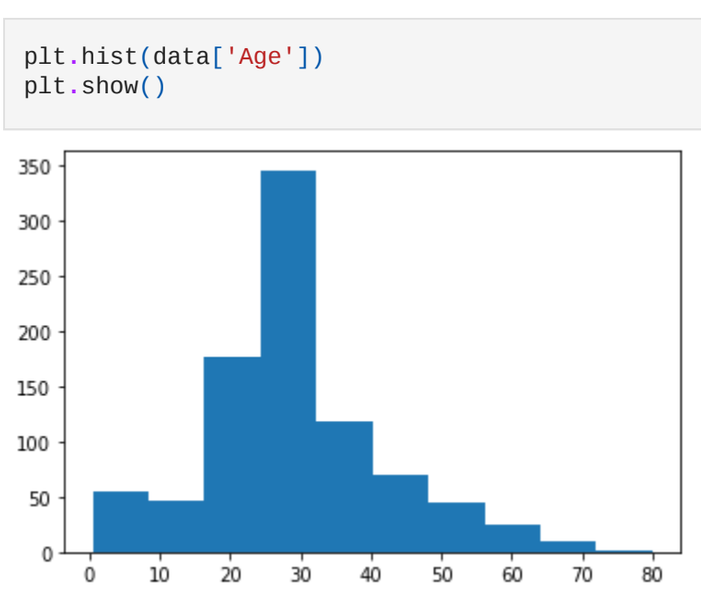
In [80]:
```
sns.countplot(data['Sex'])
```

```
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[80]:
```
<AxesSubplot:xlabel='Sex', ylabel='count'>
```



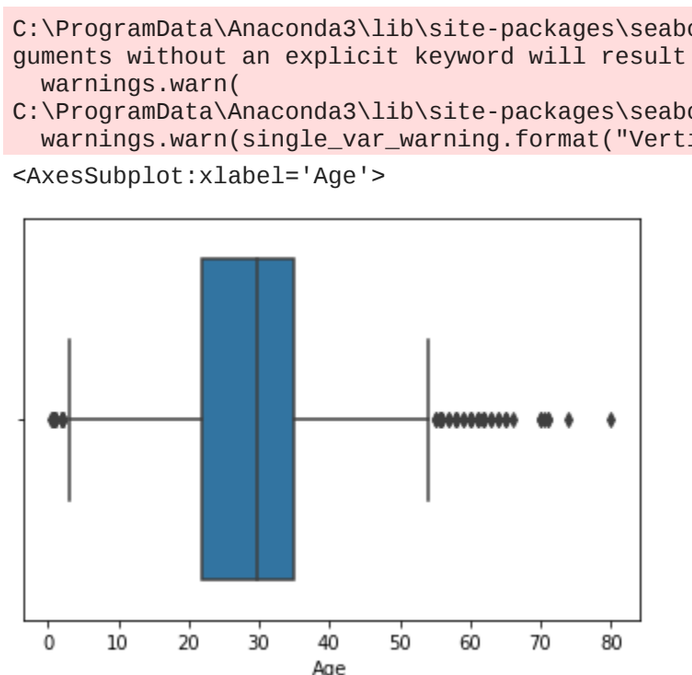
In [83]:
```
plt.hist(data['Age'])
plt.show()
```



In [87]:
```
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[87]:
```
<AxesSubplot:xlabel='Age'>
```



Bivariate Analysis

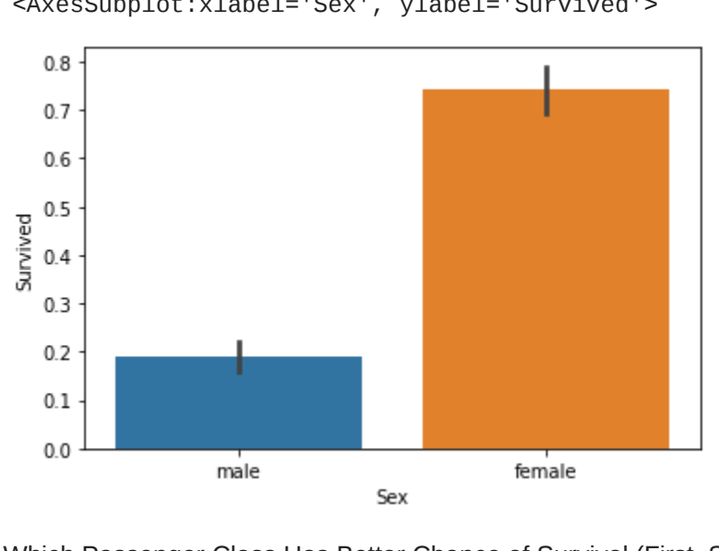
How Has Better Chance of Survival Male or Female?

In [88]:
```
data.columns
```

Out[88]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Gender_New', 'Age',
       'SibSp', 'Parch', 'Ticket', 'Fare', 'Embarked', 'Gender'],
      dtype='object')
```

In [92]:
```
sns.barplot(x='Sex', y='Survived', data=data)
```

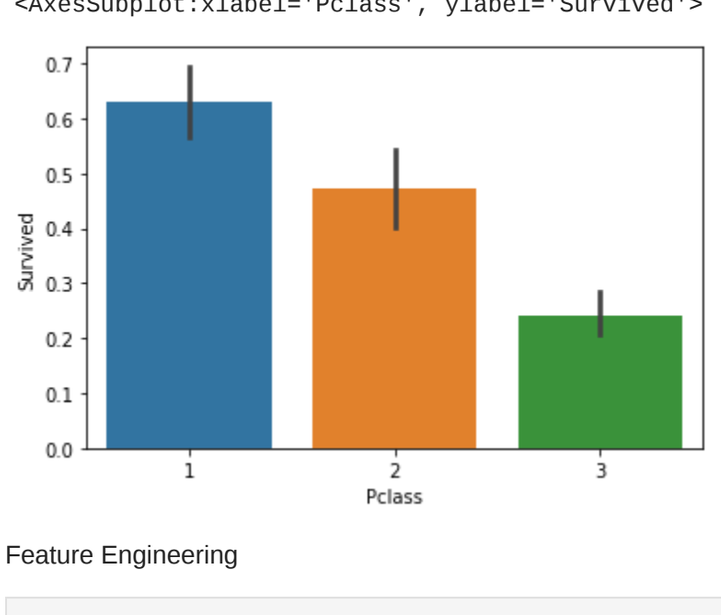
Out[92]:
```
<AxesSubplot:xlabel='Sex', ylabel='Survived'>
```



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

In [94]:
```
sns.barplot(x='Pclass', y='Survived', data=data)
```

Out[94]:
```
<AxesSubplot:xlabel='Pclass', ylabel='Survived'>
```



Feature Engineering

In [95]:
```
data.columns
```

Out[95]:
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Gender_New', 'Age',
       'SibSp', 'Parch', 'Ticket', 'Fare', 'Embarked', 'Gender'],
      dtype='object')
```

In [98]:
```
data['Family_Size'] = data['SibSp'] + data['Parch']
```

In [99]:
```
data.head()
```

Out[99]:

	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	7.2500	S	1	1
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	0	38.0	1	0	PC 17599	71.2833	C	0	1
2	3	1	3	Heikkinen, Miss. Laina	female	0	26.0	0	0	STON/O2. 3101282	7.9250	S	0	0
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	0	35.0	1	0	113803	53.1000	S	0	1
4	5	0	3	Allen, Mr. William Henry	male	1	35.0	0	0	373450	8.0500	S	1	0

In [100…
```
data.columns
```

Out[100…
```
Index(['PassengerId', 'Survived', 'Pclass', 'Name', 'Sex', 'Gender_New', 'Age',
       'SibSp', 'Parch', 'Ticket', 'Fare', 'Embarked', 'Gender',
       'Family_Size'],
      dtype='object')
```

In [102…
```
data['Fare_Per_Person'] = data['Fare'] / (data['Family_Size'] + 1)
```

In [103…
```
data.head()
```

Out[103…

	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	7.2500	S	1	1	3.62500
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	0	38.0	1	0	PC 17599	71.2833	C	0	1	35.64165
2	3	1	3	Heikkinen, Miss. Laina	female	0	26.0	0	0	STON/O2. 3101282	7.9250	S	0	0	7.92500
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	0	35.0	1	0	113803	53.1000	S	0	1	26.55000
4	5	0	3	Allen, Mr. William Henry	male	1	35.0	0	0	373450	8.0500	S	1	0	8.05000

In [ ]: