In [1]:
```
#Objective is to find the factors made people more likely to survive th
e sinking of the titanic
```

In [2]:
```
#dataset analysis based
#feature description
#passebgerid
#Pclass[passenger class] :  first class passenger ;second class passeng
er and 3rd class passenger.
#Survival[categorical variable] : '0'-people who didnot survive; '1'-pe
ople who survived
#name-Name
#Sex-Gender
#Age-Age
#Sibsp- No of siglings/spouse
#Parch-No of parents/children
#embarked - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Sou
thampton)
```

Steps involved in model building

In [ ]:
```
Step1:collecting data
Step2:Analysing data - Go through the various features and analize the
data.
Step3:Data Wrangling - cleaning data i.e to remove the unnecessary data
 and treating the missing value.
Step4:Splitting data - Split data into train and test dataset.
Step5:Accuracy check- to check how much accurate your values are.
```

In [3]:
```
#Importing the libraries

import pandas as pd
import numpy as np
```

```
import seaborn as sns
import math
import matplotlib.pyplot as plt
%matplotlib inline
```

In [4]:
```
#panda library for data analysis purpose
#numpy[numerical python] for the scientific computation
#Seaborn for statistical plotting
#matplotlib for data visualization
#math function to calculate basic mathematic function
```

In [5]:
```
titanic = pd.read_csv('C:/Users/ELCOT/Desktop/titanic.train.csv')
```

In [6]:
```
titanic
```

Out[6]:

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

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.0000
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.4500
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.0000
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500

891 rows × 12 columns

In [7]:
```
#in total there are 891 observations(rows) and 12 variables(features)
titanic.shape
```

Out[7]: (891, 12)

In [8]:
```
#to fetch only first 5 rows of the dataset
titanic.head()
```

Out[8]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	N

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
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	N
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	N

In [9]:
```
#to glance only the variables
titanic.columns
```

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

In [10]:
```
#To know each variables datatype
titanic.dtypes
```

Out[10]:
```
PassengerId      int64
Survived         int64
Pclass           int64
Name            object
Sex             object
Age            float64
SibSp            int64
```

```
Parch            int64
Ticket          object
Fare           float64
Cabin           object
Embarked        object
dtype: object
```

In [11]:
```
titanic.head(7,)
```

Out[11]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	N
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	N
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	N
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	N
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E

In [12]:
```
titanic.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 [13]:
```
#to fetch only paricular column
Pclass_titanic=titanic['Pclass']
```

In [14]:
```
Pclass_titanic
```

Out[14]:
```
0      3
1      1
2      3
3      1
4      3
      ..
886    2
887    1
888    3
889    1
890    3
Name: Pclass, Length: 891, dtype: int64
```

In [15]:
```
len(titanic)
```

Out[15]: 891

In [16]:
```
print("# of passengers in the data:" +str(len(titanic.index)))
```

```
# of passengers in the data:891
```

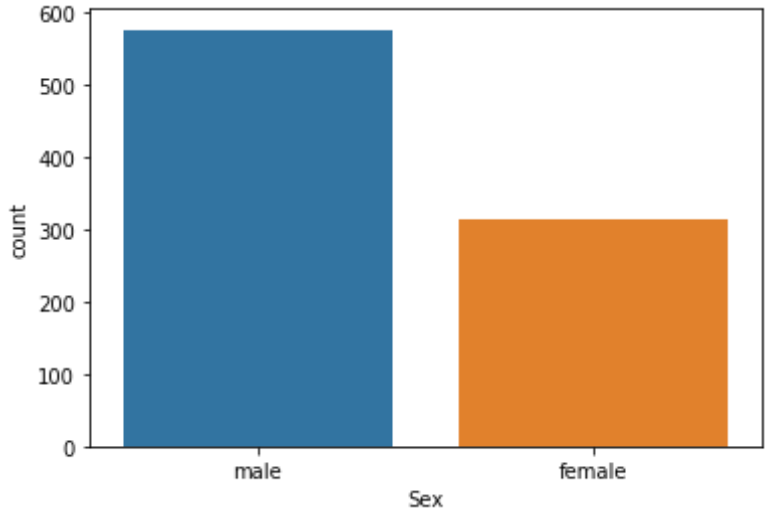
In [17]:
```
len(titanic.index)
```

Out[17]: 891

Analysing the variable clearly plotting the graph

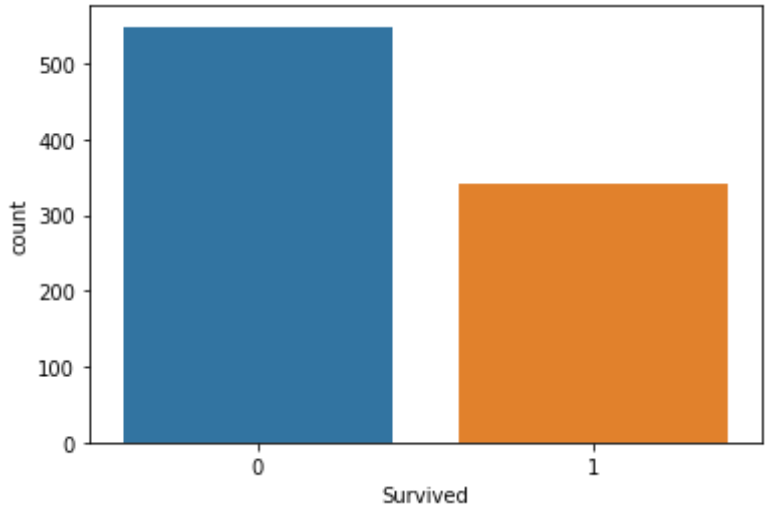
In [18]:
```
import seaborn as sns
sns.countplot(x='Sex',data=titanic)
```

Out[18]: <matplotlib.axes._subplots.AxesSubplot at 0x2828ea46788>



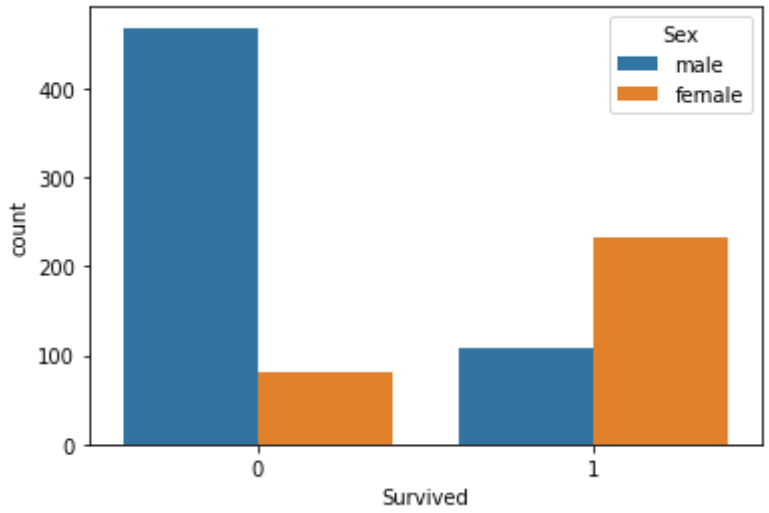
In [19]:
```
sns.countplot(x='Survived',data=titanic)
```

Out[19]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2828eb2adc8>
```



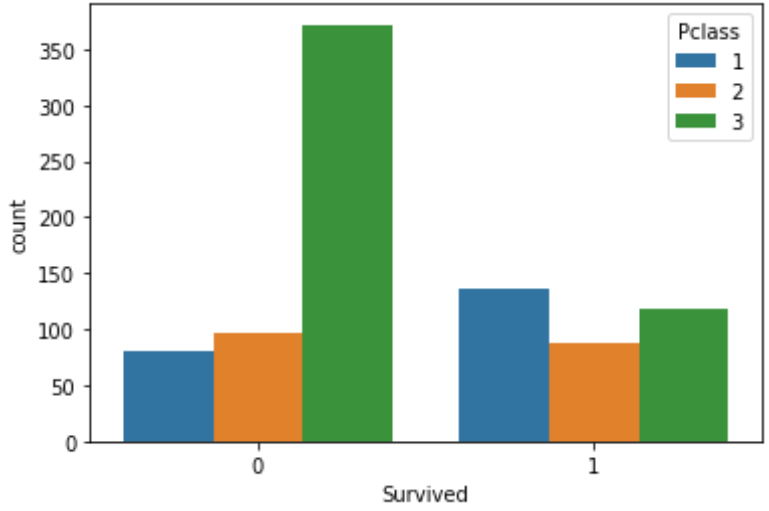
In [20]:
```
sns.countplot(x='Survived',hue='Sex',data=titanic)
```

Out[20]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2828eb93bc8>
```



In [21]:
```
sns.countplot(x='Survived',hue='Pclass',data=titanic)
```

Out[21]: <matplotlib.axes._subplots.AxesSubplot at 0x2828ec14e08>



In [22]:
```
Age_titanic=titanic['Age']
```

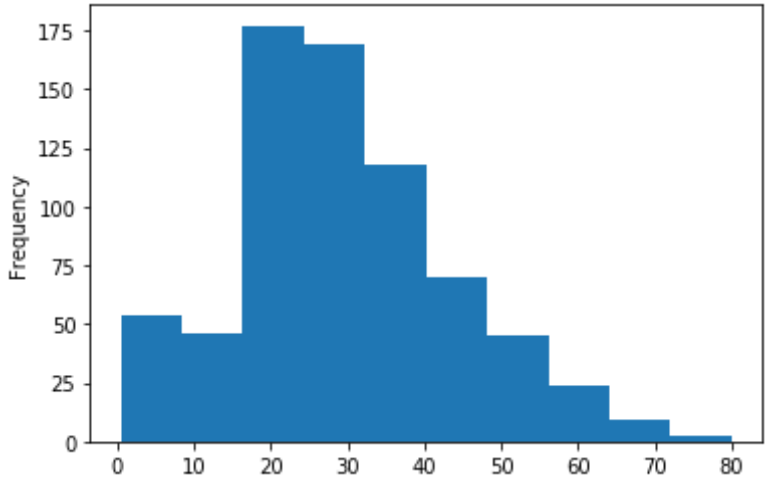
In [23]:
```
import pandas as pd
```

In [24]:
```
Age_titanic
```

Out[24]:
```
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
```

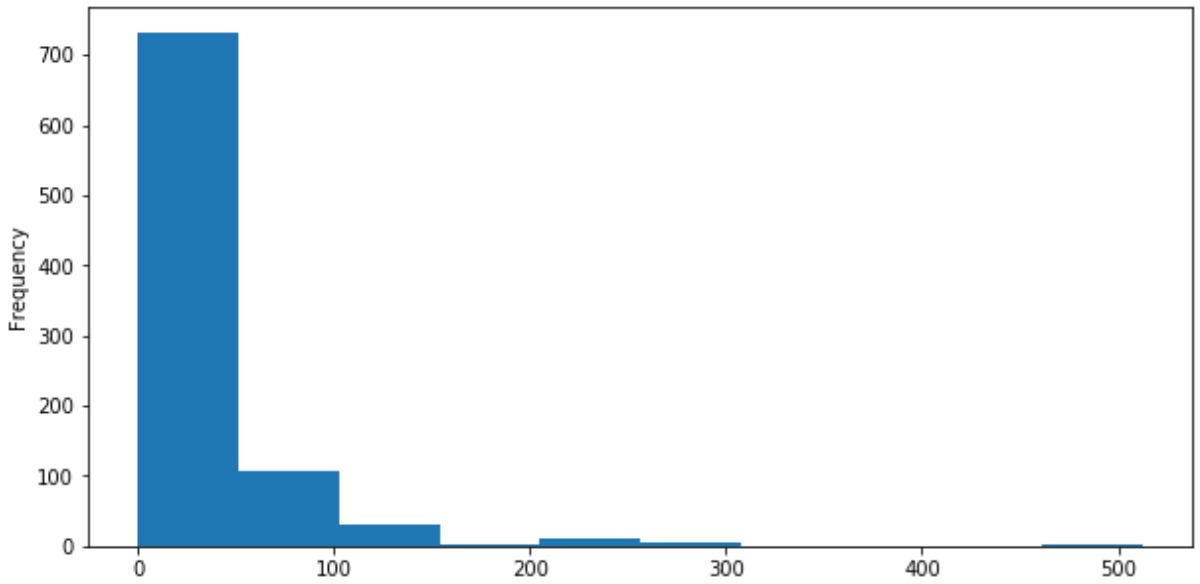
In [25]:
```
Age_titanic.plot.hist()
```

Out[25]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2828ecc7cc8>
```



In [26]:
```
titanic['Fare'].plot.hist(bins=10,figsize=(10,5))
```

Out[26]: <matplotlib.axes._subplots.AxesSubplot at 0x2828ed61a48>



In [27]:
```
titanic['Fare']
```

Out[27]:
```
0       7.2500
1      71.2833
2       7.9250
3      53.1000
4       8.0500
        ...
886    13.0000
887    30.0000
888    23.4500
889    30.0000
890     7.7500
Name: Fare, Length: 891, dtype: float64
```

In [28]:
```
titanic.head()
```

Out[28]:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	N
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	N
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	N

In [29]:
```
titanic.drop("Cabin", axis=1, inplace=True)
```

In [30]:
```
titanic.head()
```

Out[30]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	En
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	

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

Data Wrangling/Cleaning

In [31]:
```
titanic.isnull()
```

Out[31]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
0	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked
886	False	False	False	False	False	False	False	False	False	False	False
887	False	False	False	False	False	False	False	False	False	False	False
888	False	False	False	False	False	True	False	False	False	False	False
889	False	False	False	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False	False	False	False

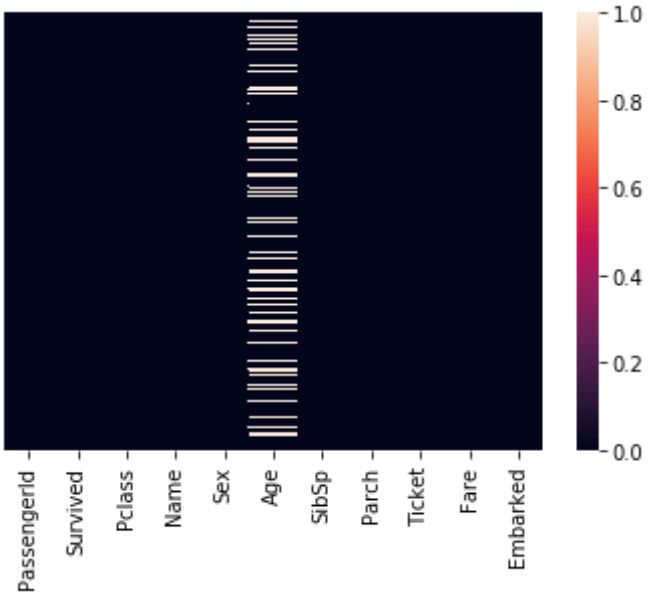
891 rows × 11 columns

In [32]:
```
titanic.isnull().sum()
```

Out[32]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Embarked         2
dtype: int64
```

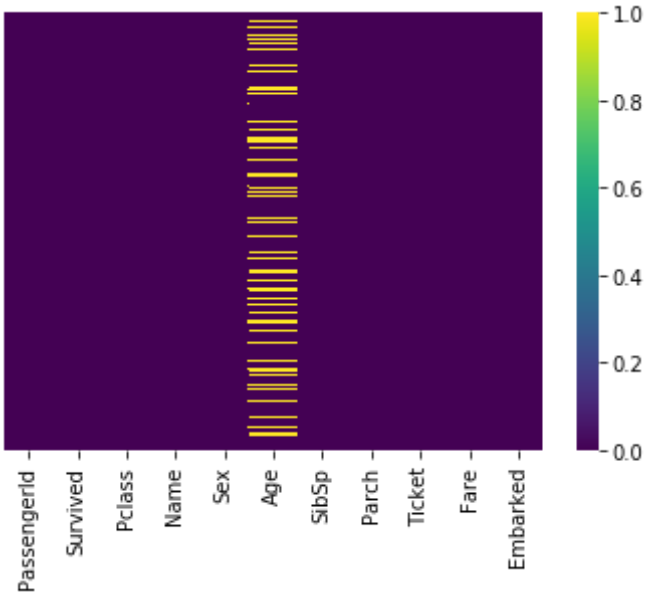
In [33]:
```
sns.heatmap(titanic.isnull(), yticklabels=False)
```

Out[33]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2828ee3b188>
```



In [34]:
```
sns.heatmap(titanic.isnull(), yticklabels=False, cmap='viridis')
```

Out[34]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2828eefdf48>
```



In [35]:
```
titanic.isnull().sum()
```

Out[35]:
```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Embarked         2
dtype: int64
```

In [36]:
```
titanic.dropna(inplace=True)
```

In [37]:
```
titanic.isnull().sum()
```

Out[37]:
```
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 [38]:
```
titanic.head()
```

Out[38]:

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

In [39]:
```
Sex=pd.get_dummies(titanic['Sex'],drop_first=True)
Sex.head()
```

Out[39]:

	male
0	1
1	0
2	0
3	0
4	1

In [40]:
```
Embarked=pd.get_dummies(titanic['Embarked'],drop_first=True)
Embarked.head()
```

Out[40]:

	Q	S
0	0	1
1	0	0
2	0	1
3	0	1
4	0	1

In [41]:
```
Pclass=pd.get_dummies(titanic['Pclass'],drop_first=True)
Pclass.head()
```

Out[41]:

	2	3
0	0	1
1	0	0
2	0	1
3	0	0

	2	3
4	0	1

In [42]:
```
titanic=pd.concat([titanic,Sex,Embarked,Pclass],axis=1)
```

In [43]:
```
titanic.head()
```

Out[43]:

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

In [44]:
```
titanic.drop(['Pclass','Sex','Embarked','PassengerId','Name'],axis=1,inplace=True)
```

In [45]:
```
titanic.head()
```

Out[45]:

	Survived	Age	SibSp	Parch	Ticket	Fare	male	Q	S	2	3
0	0	22.0	1	0	A/5 21171	7.2500	1	0	1	0	1
1	1	38.0	1	0	PC 17599	71.2833	0	0	0	0	0
2	1	26.0	0	0	STON/O2. 3101282	7.9250	0	0	1	0	1
3	1	35.0	1	0	113803	53.1000	0	0	1	0	0
4	0	35.0	0	0	373450	8.0500	1	0	1	0	1

In [46]:
```
titanic.drop(['Ticket'],axis=1,inplace=True)
```

Train Data

In [47]:
```
#let us build the model on the train data and predict the output on the
 test data
y=titanic['Survived']
X=titanic.drop(['Survived'],axis=True)
```

In [ ]:
```
import sklearn
```

In [49]:
```
from sklearn.model_selection import train_test_split
```

In [53]:
```
X_train, X_test, y_train, y_test = train_test_split(  X, y, test_size=
0.3, random_state=1)
```

In [51]:
```
from sklearn.linear_model import LogisticRegression
```

In [52]:
```
logmodel=LogisticRegression()
```

In [55]:
```
logmodel.fit(X_train,y_train)
```

```
C:\Users\ELCOT\Anaconda3\ANA\lib\site-packages\sklearn\linear_model\_lo
gistic.py:940: 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
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG)
```

Out[55]:
```
LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=
True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='auto', n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs', tol=0.0001, verbo
se=0,
                   warm_start=False)
```

In [56]:
```
predictioon=logmodel.predict(X_test)
```

In [57]:
```
from sklearn.metrics import classification_report
```

In [58]:
```
classification_report(y_test,predictioon)
```

Out[58]:
```
'              precision    recall  f1-score   support\n\n           0
      0.80      0.81      0.81       126\n           1       0.72
0.72      0.72        88\n\n    accuracy                           0.77
       214\n   macro avg       0.76      0.76      0.76       214\nweig
hted avg       0.77      0.77      0.77       214\n'
```

In [59]:
```
from sklearn.metrics import confusion_matrix
```

In [60]:
```
confusion_matrix(y_test,predictioon)
```

Out[60]:
```
array([[102,  24],
       [ 25,  63]], dtype=int64)
```

Accuracy Check

In [68]:
```
from sklearn.metrics import accuracy_score
```

In [69]:
```
accuracy_score(y_test,predictioon)
```

Out[69]:
```
0.7710280373831776
```

In [ ]:
```
-
```