NAME: M.SIDDARTHA REG.NO: 21BCE9247 ASSIGNMENT-3

1.IMPORT THE NECESSARY LIBRARIES

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

2.IMPORT THE DATASET

In [2]:
```
df=pd.read_csv("Titanic-Dataset.csv")
```

In [3]:
```
df
```

Out[3]:

	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
...	...	...	...	...	...	...	...	...	...	...
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 [4]:
```
df.shape
```

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

In [5]:
```
df.describe()
```

Out[5]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
count	891.000000	891.000000	891.000000	714.000000	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	14.526497	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	20.125000	0.000000	0.000000	7.910400
50%	446.000000	0.000000	3.000000	28.000000	0.000000	0.000000	14.454200
75%	668.500000	1.000000	3.000000	38.000000	1.000000	0.000000	31.000000
max	891.000000	1.000000	3.000000	80.000000	8.000000	6.000000	512.329200

In [6]:
```
df.head()
```

Out[6]:

	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 [7]:
```
df.tail()
```

Out[7]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.00	NaN
887	888	1	1	Graham, Miss. Margaret Edith	female	19.0	0	0	112053	30.00	B42
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	NaN	1	2	W./C. 6607	23.45	NaN
889	890	1	1	Behr, Mr. Karl Howell	male	26.0	0	0	111369	30.00	C148
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.75	NaN

In [8]:
```
df.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
```

3.CHECKING FOR NULL VALUES

In [9]:
```
df.isnull().any()
```

Out[9]:
```
PassengerId    False
Survived       False
Pclass         False
Name           False
Sex            False
Age             True
SibSp          False
Parch          False
Ticket         False
Fare           False
Cabin           True
Embarked        True
dtype: bool
```

In [10]:
```
df.isnull().sum()
```

Out[10]:
```
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 [11]:
```
mean1 = df["Age"].mean()
```

In [12]:
```
mean1
```

Out[12]:
```
29.69911764705882
```

In [13]:
```
df["Age"] = df["Age"].fillna(mean1)
```

In [14]:
```
df["Age"]
```

Out[14]:
```
0      22.000000
1      38.000000
2      26.000000
3      35.000000
4      35.000000
         ...
886    27.000000
887    19.000000
888    29.699118
889    26.000000
890    32.000000
Name: Age, Length: 891, dtype: float64
```

In [15]:
```
model = df["Cabin"].mode()
```

In [16]:
```
model
```

Out[16]:
```
0        B96 B98
1    C23 C25 C27
2             G6
dtype: object
```

In [17]:
```
df["Cabin"] = df["Cabin"].fillna(model[2])
```

In [18]:
```
df["Cabin"]
```

Out[18]:
```
0        G6
1       C85
2        G6
3      C123
4        G6
       ...
886      G6
887     B42
888      G6
889    C148
890      G6
Name: Cabin, Length: 891, dtype: object
```

In [19]:
```
mode2 = df["Embarked"].mode()
```

In [20]:
```
mode2
```

Out[20]:
```
0    S
dtype: object
```

In [22]:
```
df["Embarked"] = df["Embarked"].fillna(mode2[0])
```

In [24]:
```
df["Embarked"]
```

Out[24]:
```
0      S
1      C
2      S
3      S
4      S
      ..
886    S
887    S
888    S
889    C
890    Q
Name: Embarked, Length: 891, dtype: object
```

In [25]:
```
df.isnull().any()
```

Out[25]:
```
PassengerId    False
Survived       False
Pclass         False
Name           False
Sex            False
Age            False
SibSp          False
Parch          False
Ticket         False
Fare           False
Cabin          False
Embarked       False
dtype: bool
```

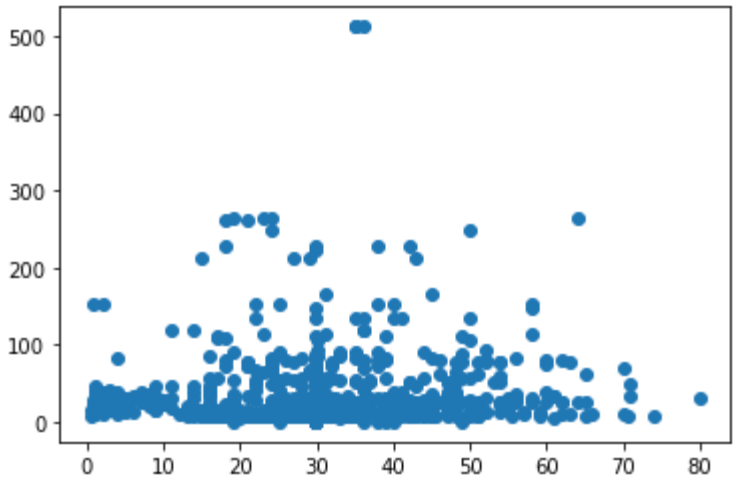
In [26]:
```
df.isnull().sum()
```

Out[26]:
```
PassengerId    0
Survived       0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
```

4.DATA VISUALISATION

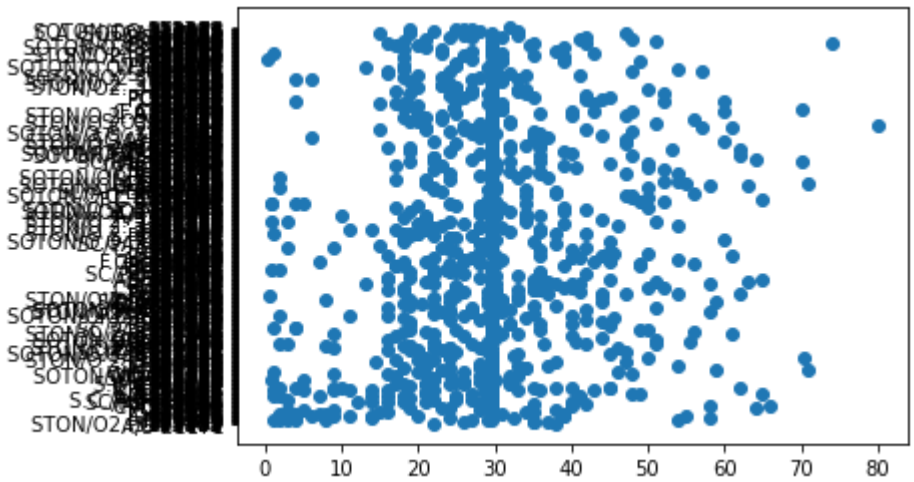
In [27]:
```
plt.scatter(x=df["Age"],y=df["Fare"])
```

Out[27]:
```
<matplotlib.collections.PathCollection at 0x1f1f3f0d340>
```



In [28]:
```
plt.scatter(x=df["Age"],y=df["Ticket"])
```

Out[28]:
```
<matplotlib.collections.PathCollection at 0x1f1f46afa30>
```



In [29]:
```
cor = df.corr()
```

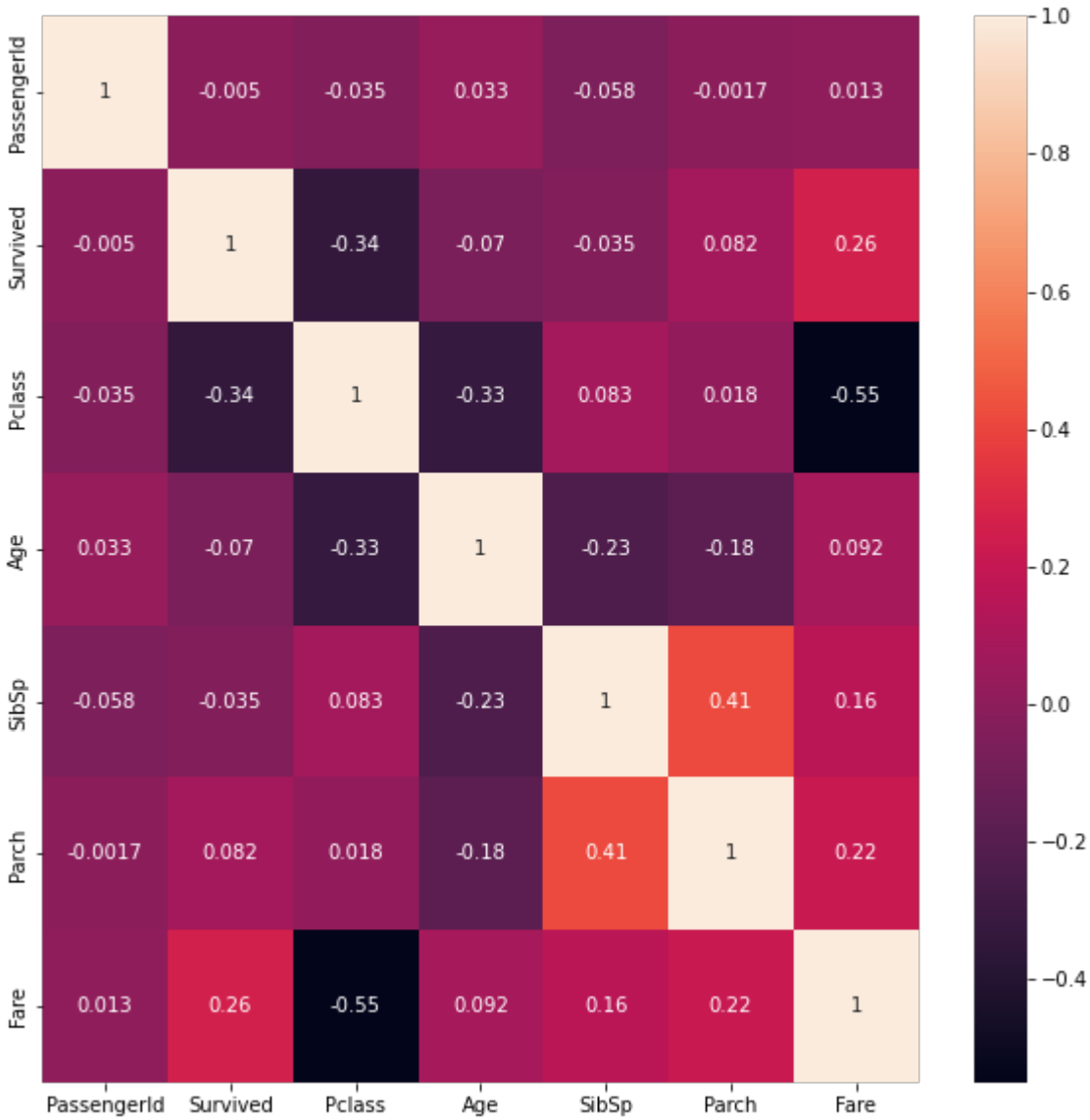
In [30]:
```
cor
```

Out[30]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005007	-0.035144	0.033207	-0.057527	-0.001652	0.012658
Survived	-0.005007	1.000000	-0.338481	-0.069809	-0.035322	0.081629	0.257307
Pclass	-0.035144	-0.338481	1.000000	-0.331339	0.083081	0.018443	-0.549500
Age	0.033207	-0.069809	-0.331339	1.000000	-0.232625	-0.179191	0.091566
SibSp	-0.057527	-0.035322	0.083081	-0.232625	1.000000	0.414838	0.159651
Parch	-0.001652	0.081629	0.018443	-0.179191	0.414838	1.000000	0.216225
Fare	0.012658	0.257307	-0.549500	0.091566	0.159651	0.216225	1.000000

In [31]:
```
plt.figure(figsize=(10,10))
sns.heatmap(cor,annot=True)
```

Out[31]: <AxesSubplot:>

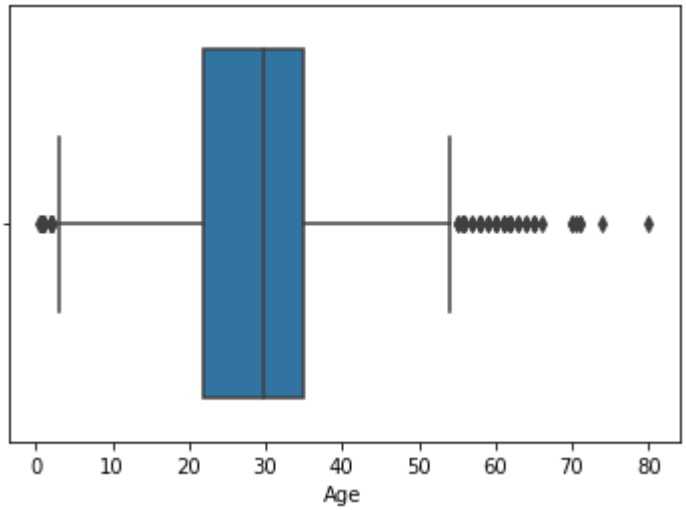


5.OUTLIER DETECTION

In [32]:
```
sns.boxplot(df["Age"])
```

```
C:\Users\SIDDU\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureW
arning: Pass the following variable as a keyword arg: x. From version 0.12, t
he only valid positional argument will be `data`, and passing other arguments
without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[32]: <AxesSubplot:xlabel='Age'>



In [34]:
```
A_q1 = df.Age.quantile(0.25)
A_q3 = df.Age.quantile(0.75)
```

In [35]:
```
IQR = A_q3-A_q1
```

In [36]:
```
IQR
```

Out[36]: 13.0

In [37]:
```
upper_limit = A_q3+1.5*IQR
```

In [38]:
```
upper_limit
```

Out[38]: 54.5

In [39]:
```
med = df.Age.median()
```

In [40]:
```
med
```

Out[40]: 29.69911764705882

In [41]:
```
df["Age"] = np.where(df["Age"]>upper_limit,med,df["Age"])
```

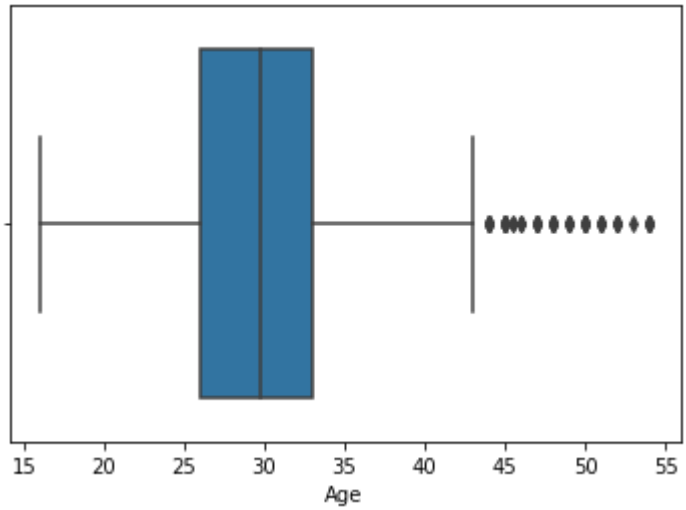
In [42]:
```
lower_limit =  A_q3-1.5*IQR
```

In [43]:
```
df["Age"] = np.where(df["Age"]<lower_limit,med,df["Age"])
```

In [51]:
```
sns.boxplot(df.Age)
```

```
C:\Users\SIDDU\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureW
arning: Pass the following variable as a keyword arg: x. From version 0.12, t
he only valid positional argument will be `data`, and passing other arguments
without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[51]: <AxesSubplot:xlabel='Age'>



In [44]:
```
from scipy import stats
```

In [45]:
```
Age_zscore = stats.zscore(df.Age)
```

In [46]:
```
Age_zscore
```

Out[46]:
```
0     -1.035650
1      0.948887
2     -0.539516
3      0.576786
4      0.576786
         ...
886   -0.415482
887   -1.407751
888   -0.080701
889   -0.539516
890    0.204686
Name: Age, Length: 891, dtype: float64
```

In [47]:
```
df_z = df[np.abs(Age_zscore)<=3]
```

In [48]:
```
df_z
```

Out[48]:

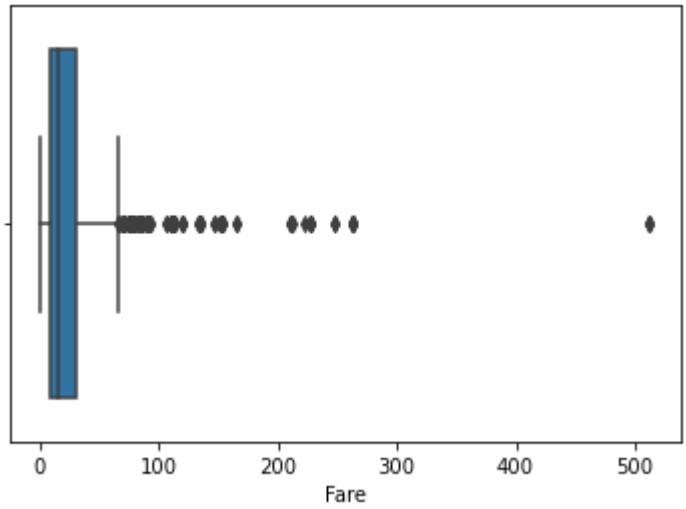
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	F
0	1	0	3	Braund, Mr. Owen Harris	male	22.000000	1	0	A/5 21171	7.2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.000000	1	0	PC 17599	71.2
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.9
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.1
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.0
...	...	...	...	...	...	...	...	...	...	
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.0
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.0
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.4
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.0
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.7

891 rows × 12 columns

In [52]:
```
sns.boxplot(df.Fare)
```

```
C:\Users\SIDDU\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureW
arning: Pass the following variable as a keyword arg: x. From version 0.12, t
he only valid positional argument will be `data`, and passing other arguments
without an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[52]:
```
<AxesSubplot:xlabel='Fare'>
```



6.SPLITTING DEPENDENT AND INDEPENDENT VARIABLES

In [53]:
```
df.drop(['Name'],axis=1,inplace=True)
```

In [54]:
```
df
```

Out[54]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
0	1	0	3	male	22.000000	1	0	A/5 21171	7.2500	G6
1	2	1	1	female	38.000000	1	0	PC 17599	71.2833	C85
2	3	1	3	female	26.000000	0	0	STON/O2. 3101282	7.9250	G6
3	4	1	1	female	35.000000	1	0	113803	53.1000	C123
4	5	0	3	male	35.000000	0	0	373450	8.0500	G6
...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	male	27.000000	0	0	211536	13.0000	G6
887	888	1	1	female	19.000000	0	0	112053	30.0000	B42
888	889	0	3	female	29.699118	1	2	W./C. 6607	23.4500	G6
889	890	1	1	male	26.000000	0	0	111369	30.0000	C148
890	891	0	3	male	32.000000	0	0	370376	7.7500	G6

891 rows × 11 columns

In [55]:
```
df.drop(['Ticket'],axis=1,inplace=True)
```

In [56]:
```
df
```

Out[56]:

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
0	1	0	3	male	22.000000	1	0	7.2500	G6	S
1	2	1	1	female	38.000000	1	0	71.2833	C85	C
2	3	1	3	female	26.000000	0	0	7.9250	G6	S
3	4	1	1	female	35.000000	1	0	53.1000	C123	S
4	5	0	3	male	35.000000	0	0	8.0500	G6	S
...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	male	27.000000	0	0	13.0000	G6	S
887	888	1	1	female	19.000000	0	0	30.0000	B42	S
888	889	0	3	female	29.699118	1	2	23.4500	G6	S
889	890	1	1	male	26.000000	0	0	30.0000	C148	C
890	891	0	3	male	32.000000	0	0	7.7500	G6	Q

891 rows × 10 columns

In [58]:
```
df.drop(['PassengerId'],axis=1,inplace=True)
```

In [59]:
```
df
```

Out[59]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked
0	0	3	male	22.000000	1	0	7.2500	G6	S
1	1	1	female	38.000000	1	0	71.2833	C85	C
2	1	3	female	26.000000	0	0	7.9250	G6	S
3	1	1	female	35.000000	1	0	53.1000	C123	S
4	0	3	male	35.000000	0	0	8.0500	G6	S
...	...	...	...	...	...	...	...	...	...
886	0	2	male	27.000000	0	0	13.0000	G6	S
887	1	1	female	19.000000	0	0	30.0000	B42	S
888	0	3	female	29.699118	1	2	23.4500	G6	S
889	1	1	male	26.000000	0	0	30.0000	C148	C
890	0	3	male	32.000000	0	0	7.7500	G6	Q

891 rows × 9 columns

In [60]:
```
df.drop(['Cabin'],axis=1,inplace=True)
```

In [61]:
```
df
```

Out[61]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	0	3	male	22.000000	1	0	7.2500	S
1	1	1	female	38.000000	1	0	71.2833	C
2	1	3	female	26.000000	0	0	7.9250	S
3	1	1	female	35.000000	1	0	53.1000	S
4	0	3	male	35.000000	0	0	8.0500	S
...	...	...	...	...	...	...	...	...
886	0	2	male	27.000000	0	0	13.0000	S
887	1	1	female	19.000000	0	0	30.0000	S
888	0	3	female	29.699118	1	2	23.4500	S
889	1	1	male	26.000000	0	0	30.0000	C
890	0	3	male	32.000000	0	0	7.7500	Q

891 rows × 8 columns

In [62]:
```
y = df["Survived"]
```

In [63]:
```
y
```

Out[63]:
```
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
```

In [64]:
```
x = df.drop("Survived",axis = 1)
```

In [65]:
```
x
```

Out[65]:

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	male	22.000000	1	0	7.2500	S
1	1	female	38.000000	1	0	71.2833	C
2	3	female	26.000000	0	0	7.9250	S
3	1	female	35.000000	1	0	53.1000	S
4	3	male	35.000000	0	0	8.0500	S
...	...	...	...	...	...	...	...
886	2	male	27.000000	0	0	13.0000	S
887	1	female	19.000000	0	0	30.0000	S
888	3	female	29.699118	1	2	23.4500	S
889	1	male	26.000000	0	0	30.0000	C
890	3	male	32.000000	0	0	7.7500	Q

891 rows × 7 columns

7.Encoding

In [66]:
```
from sklearn.preprocessing import LabelEncoder
```

In [67]:
```
lr = LabelEncoder()
```

In [68]:
```
lr
```

Out[68]:
```
LabelEncoder()
```

In [70]:
```
x["Sex"] = lr.fit_transform(x["Sex"])
```

In [71]:
```
x["Sex"]
```

Out[71]:
```
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: int32
```

In [72]:
```
x.head()
```

Out[72]:

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	1	22.0	1	0	7.2500	S
1	1	0	38.0	1	0	71.2833	C
2	3	0	26.0	0	0	7.9250	S
3	1	0	35.0	1	0	53.1000	S
4	3	1	35.0	0	0	8.0500	S

In [73]:
```
x["Embarked"] = lr.fit_transform(x["Embarked"])
```

In [74]:
```
x["Embarked"]
```

Out[74]:
```
0      2
1      0
2      2
3      2
4      2
      ..
886    2
887    2
888    2
889    0
890    1
Name: Embarked, Length: 891, dtype: int32
```

In [75]:
```
x.head()
```

Out[75]:

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
0	3	1	22.0	1	0	7.2500	2
1	1	0	38.0	1	0	71.2833	0
2	3	0	26.0	0	0	7.9250	2
3	1	0	35.0	1	0	53.1000	2
4	3	1	35.0	0	0	8.0500	2

In [76]:
```
x["Sex"].nunique()
```

Out[76]:
```
2
```

In [77]:
```
x["Embarked"].nunique()
```

Out[77]:
```
3
```

In [78]:
```
y.head()
```

Out[78]:
```
0    0
1    1
2    1
3    1
4    0
Name: Survived, dtype: int64
```

8.SPLITTING THE TRAIN AND TEST DATA

In [79]:
```
from sklearn.model_selection import train_test_split
```

In [80]:
```
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2,random_stat
```

In [82]:
```
x_train
```

Out[82]:

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
140	3	0	29.699118	0	2	15.2458	0
439	2	1	31.000000	0	0	10.5000	2
817	2	1	31.000000	1	1	37.0042	0
378	3	1	20.000000	0	0	4.0125	0
491	3	1	21.000000	0	0	7.2500	2
...	...	...	...	...	...	...	...
835	1	0	39.000000	1	1	83.1583	0
192	3	0	19.000000	1	0	7.8542	2
629	3	1	29.699118	0	0	7.7333	1
559	3	0	36.000000	1	0	17.4000	2
684	2	1	29.699118	1	1	39.0000	2

712 rows × 7 columns

In [83]:
```
y_train
```

Out[83]:
```
140    0
439    0
817    0
378    0
491    0
      ..
835    1
192    1
629    0
559    1
684    0
Name: Survived, Length: 712, dtype: int64
```

In [84]:
```
x_test
```

Out[84]:

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
495	3	1	29.699118	0	0	14.4583	0
648	3	1	29.699118	0	0	7.5500	2
278	3	1	29.699118	4	1	29.1250	1
31	1	0	29.699118	1	0	146.5208	0
255	3	0	29.000000	0	2	15.2458	0
...	...	...	...	...	...	...	...
780	3	0	29.699118	0	0	7.2292	0
837	3	1	29.699118	0	0	8.0500	2
215	1	0	31.000000	1	0	113.2750	0
833	3	1	23.000000	0	0	7.8542	2
372	3	1	19.000000	0	0	8.0500	2

179 rows × 7 columns

In [85]:
```
y_test
```

Out[85]:
```
495    0
648    0
278    0
31     1
255    1
      ..
780    1
837    0
215    1
833    0
372    0
Name: Survived, Length: 179, dtype: int64
```

In [86]:
```
x_train.shape
```

Out[86]:
```
(712, 7)
```

In [87]:
```
x_test.shape
```

Out[87]:
```
(179, 7)
```

In [88]:
```
y_train.shape
```

Out[88]:
```
(712,)
```

In [89]:
```
y_test.shape
```

Out[89]:
```
(179,)
```

9.FEATURE SCALING

In [90]:
```
from sklearn.preprocessing import StandardScaler
```

In [91]:
```
sc = StandardScaler()
```

In [92]:
```
sc
```

Out[92]:
```
StandardScaler()
```

In [93]:
```
x_train = sc.fit_transform(x_train)
```

In [94]:
```
x_train
```

Out[94]:
```
array([[ 0.81925059, -1.37207547, -0.08095892, ...,  1.95926403,
        -0.33167904, -1.98156574],
       [-0.38096838,  0.72882288,  0.08297642, ..., -0.47741019,
        -0.42640542,  0.5790056 ],
       [-0.38096838,  0.72882288,  0.08297642, ...,  0.74092692,
         0.10261958, -1.98156574],
       ...,
       [ 0.81925059,  0.72882288, -0.08095892, ..., -0.47741019,
        -0.48162887, -0.70128007],
       [ 0.81925059, -1.37207547,  0.71306931, ..., -0.47741019,
        -0.28868112,  0.5790056 ],
       [-0.38096838,  0.72882288, -0.08095892, ...,  0.74092692,
         0.14245584,  0.5790056 ]])
```

In [95]:
```
x_test = sc.fit_transform(x_test)
```

In [96]:
```
x_test
```

Out[96]:
```
array([[ 0.86022947,  0.77344314, -0.07992613, ..., -0.46006628,
        -0.39903373, -1.80134224],
       [ 0.86022947,  0.77344314, -0.07992613, ..., -0.46006628,
        -0.54333564,  0.61394061],
       [ 0.86022947,  0.77344314, -0.07992613, ...,  0.88996427,
        -0.09267286, -0.59370081],
       ...,
       [-1.50871015, -1.29291987,  0.07224618, ..., -0.46006628,
         1.66506862, -1.80134224],
       [ 0.86022947,  0.77344314, -0.86356362, ..., -0.46006628,
        -0.53698145,  0.61394061],
       [ 0.86022947,  0.77344314, -1.33146852, ..., -0.46006628,
        -0.53289154,  0.61394061]])
```