Assignment-3

Perform Data preprocessing on Titanic dataset

Name- Rishima Chowdhury

Reg no-21BCE1097

Email- rishima.chowdhury2021@vitstudent.ac.in

Campus- VIT Chennai

Phone Number- 6290559155

Data Preprocessing

1. import the necessary libraries
2. import the dataset
3. Handling null values
4. Data visualization
5. outlier detection
6. Seperate Dependent and Independent variables.
7. Encoding
8. Splitting into training and testing set
9. Perform feature scaling

1. Import the necessary libraries.

In [2]:
```
import numpy as np
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
```

2. import the dataset

In [3]:
```
dataset=pd.read_csv("Titanic-Dataset.csv")
dataset.head()
```

Out[3]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN

In [4]:
```
dataset.tail()
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Em
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	

PassengerId= Unique ID of the passengers travelling

Survived= whether survived or not

```
0=no 1=yes
```

Pclass= Tickect booked class

```
1= first class, 2= second class 3=third class
```

Name= Name of the passenger

Sex= gender of the passenger

Age= Ages in years

SibSp= Number of siblings/Spouse aboard the Titanic

Parch= Number of parents/children aboard

Ticket= Ticket Number

Fare= Passenger Fare

Cabin= Cabin Number

Embarked= implies where the traveler mounted from. There are three possible values for Embark — Southampton, Cherbourg, and Queenstown.

In [5]:
```
dataset.shape
```

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

In [6]:
```
dataset.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 [7]:
```
dataset.describe()
```

Out[7]:

	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

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
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 [8]:
```
dataset.corr()
```

```
<ipython-input-8-c187c74d1e71>:1: FutureWarning: The default value of numeric_only i
n DataFrame.corr is deprecated. In a future version, it will default to False. Selec
t only valid columns or specify the value of numeric_only to silence this warning.
  dataset.corr()
```

Out[8]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005007	-0.035144	0.036847	-0.057527	-0.001652	0.012658
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307
Pclass	-0.035144	-0.338481	1.000000	-0.369226	0.083081	0.018443	-0.549500
Age	0.036847	-0.077221	-0.369226	1.000000	-0.308247	-0.189119	0.096067
SibSp	-0.057527	-0.035322	0.083081	-0.308247	1.000000	0.414838	0.159651
Parch	-0.001652	0.081629	0.018443	-0.189119	0.414838	1.000000	0.216225
Fare	0.012658	0.257307	-0.549500	0.096067	0.159651	0.216225	1.000000

3. Checking for Null Values

In [9]:
```
dataset.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]:
```
dataset.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]:
```
dataset["Age"].fillna(dataset["Age"].mean(),inplace=True)
```

In [12]:
```
dataset["Cabin"].fillna(dataset["Cabin"].mode()[0],inplace=True)
```

In [13]:
```
dataset["Embarked"].fillna(dataset["Embarked"].mode()[0],inplace=True)
```

In [14]:
```
dataset.isnull().any()
```

Out[14]:
```
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 [15]:
```
dataset.tail()
```

Out[15]:

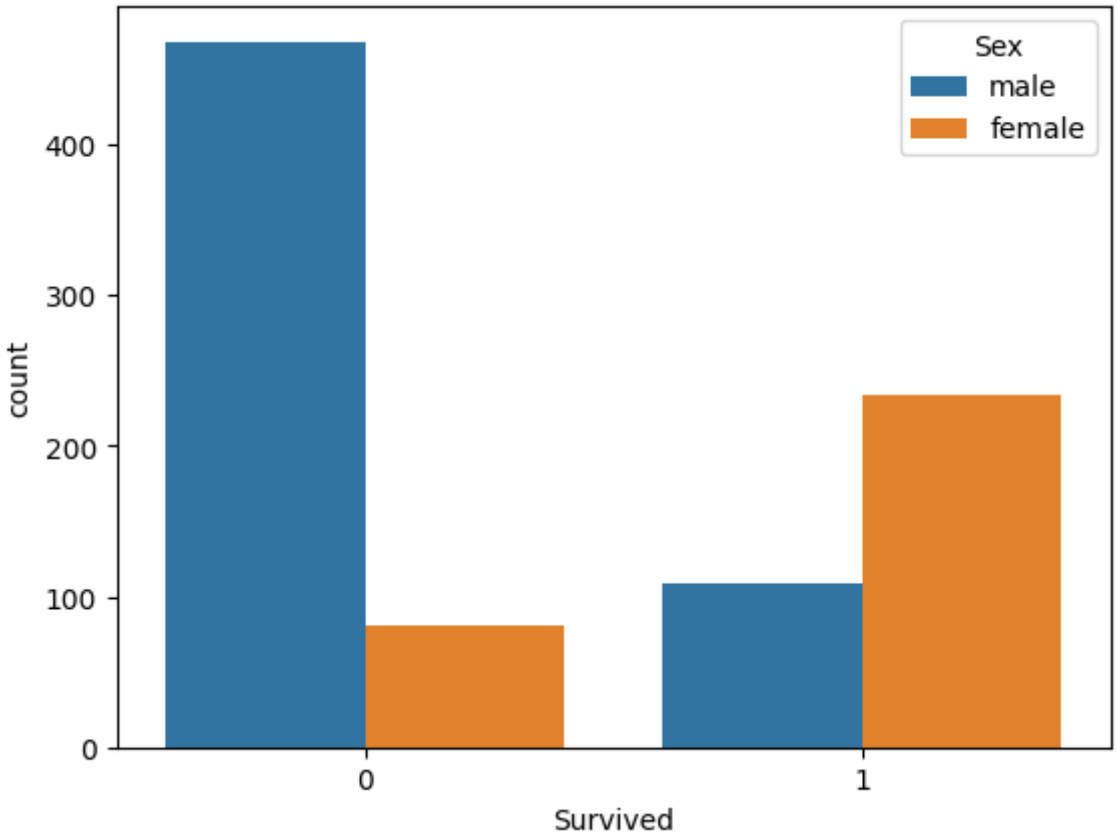
	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.00	B96 B98
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.00	B42
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.699118	1	2	W./C. 6607	23.45	B96 B98
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.00	C148

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.75	B96 B98

4. Data Visualization

In [16]:
```
sns.countplot(x='Survived',data=dataset,hue='Sex')
```

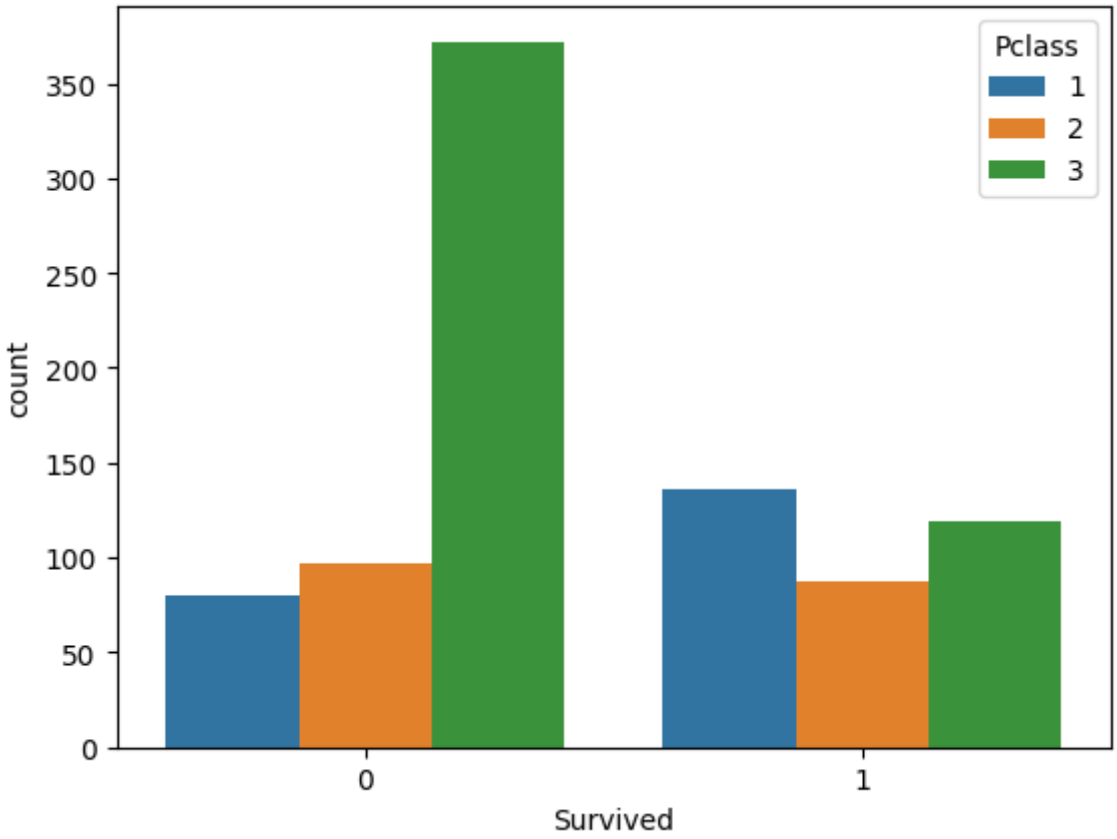
Out[16]:
```
<Axes: xlabel='Survived', ylabel='count'>
```



With the help of the coutplot we can see that large number of male did not survived (more than 400).

In [17]:
```
sns.countplot(x='Survived',data=dataset,hue = 'Pclass')
```

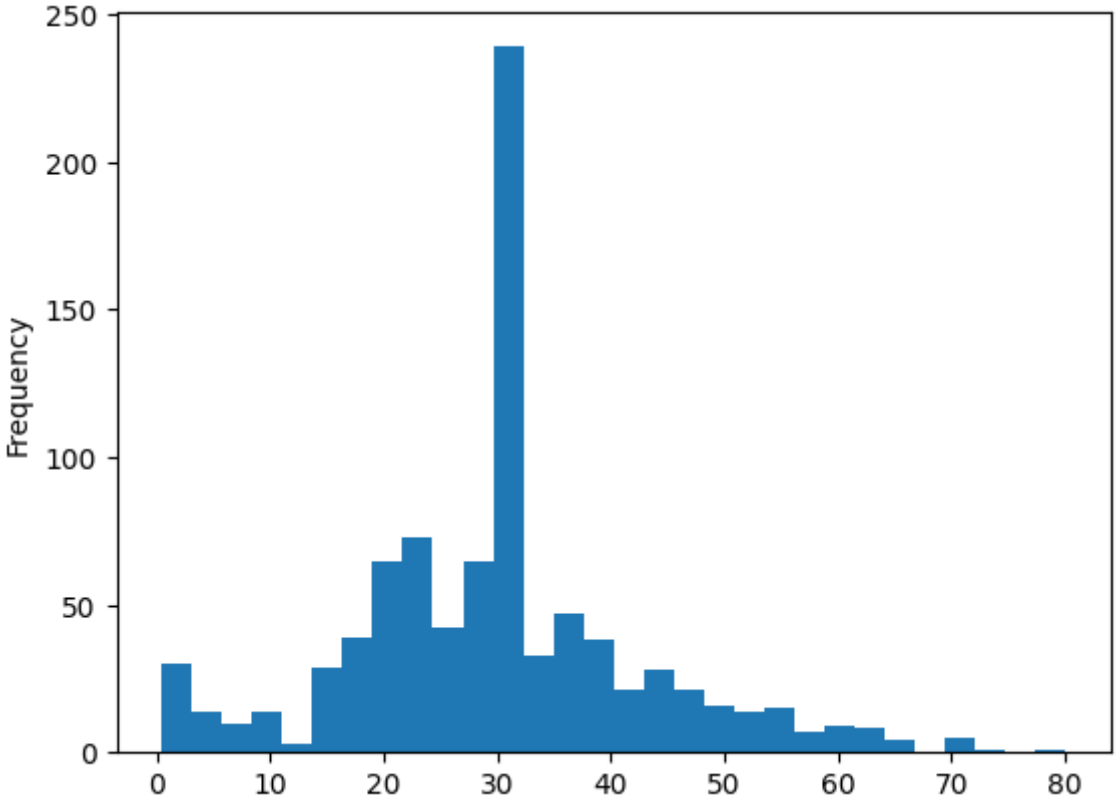
Out[17]:
```
<Axes: xlabel='Survived', ylabel='count'>
```



With help of this graph we can see that the large number of the 3rd class or say lower class people did not survived.

In [18]:
```
dataset['Age'].plot.hist(bins=30)
```

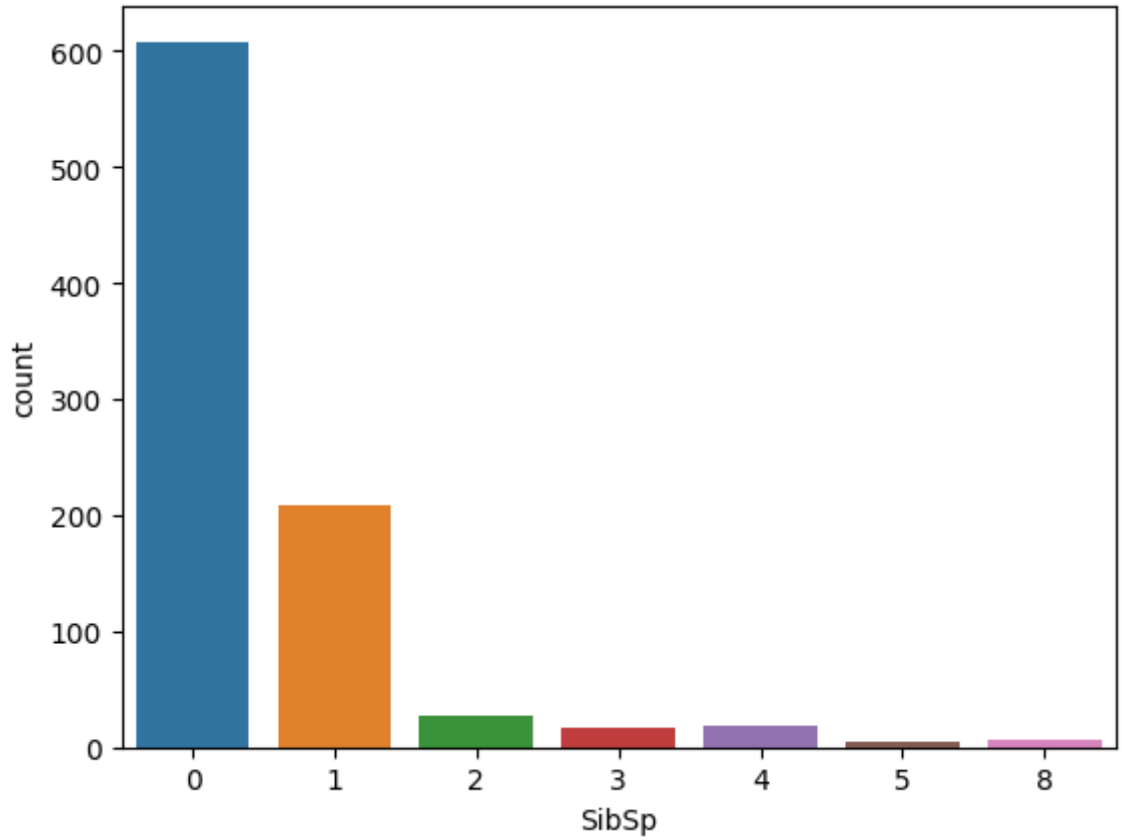
Out[18]:
```
<Axes: ylabel='Frequency'>
```



This histogram helps us to understand that many of the passengers that were present in the titanic were of age range 28-30 years.

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

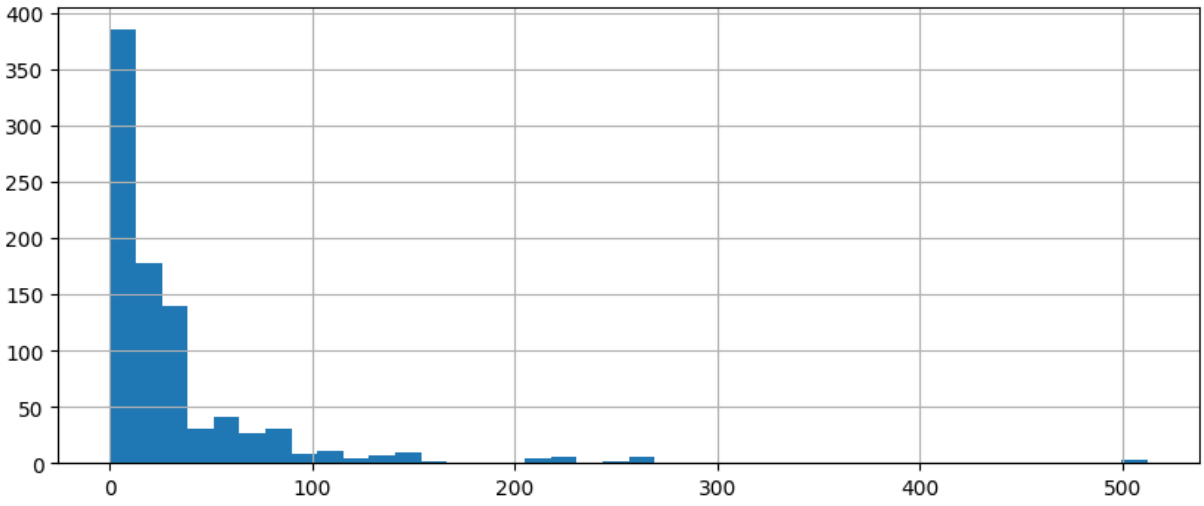
Out[19]:
```
<Axes: xlabel='SibSp', ylabel='count'>
```



Here we can see that majority of the passengers were single and second highest were we can say pessengers with their 1 sibling or spouse etc

In [20]:
```
dataset['Fare'].hist(bins=40,figsize=(10,4))
```

Out[20]:
```
<Axes: >
```



With this histogram this we can find that cheaper tickets were sold, most passengers were lower class.

5. Outlier Detector

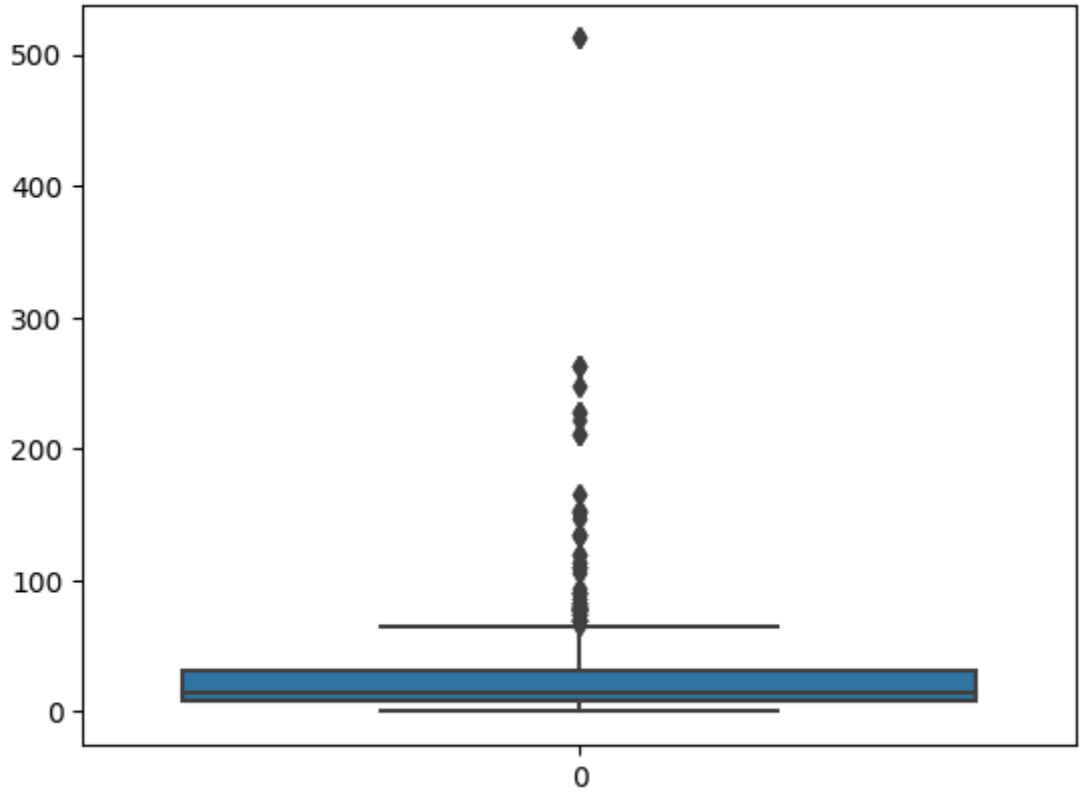
In [21]:
```
dataset.head()
```

Out[21]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	B96 B98
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	B96 B98
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	B96 B98

In [22]:
```
sns.boxplot(dataset["Fare"])
```

Out[22]:
```
<Axes: >
```



In [23]:
```
q1=dataset.Fare.quantile(0.25)
q3=dataset.Fare.quantile(0.75)
print(q1)
print(q3)
IQR=q3-q1
IQR
upper_limit = q3+1.5*IQR
upper_limit
lower_limit = q1-1.5*IQR
lower_limit
dataset.median()
```

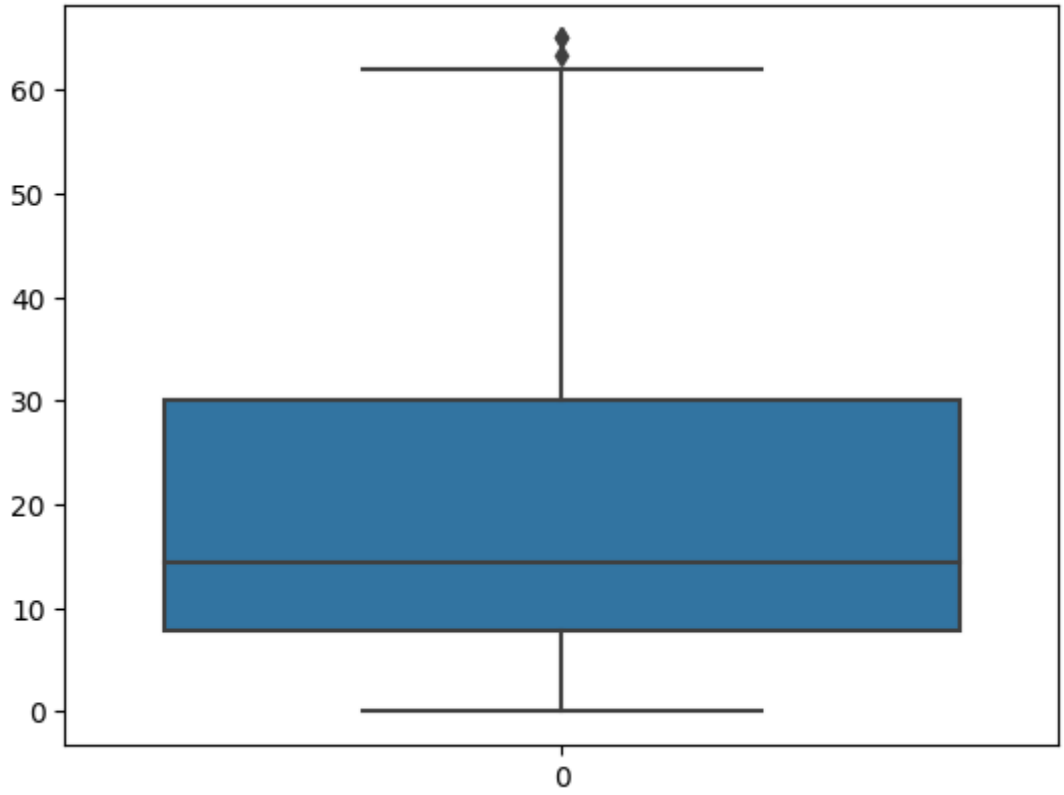
```
7.9104
31.0
<ipython-input-23-d2ca88a1885e>:11: FutureWarning: The default value of numeric_only in DataFrame.median is deprecated. In a future version, it will default to False. In addition, specifying 'numeric_only=None' is deprecated. Select only valid columns or specify the value of numeric_only to silence this warning.
  dataset.median()
```

Out[23]:
```
PassengerId    446.000000
Survived         0.000000
Pclass           3.000000
Age             29.699118
SibSp            0.000000
Parch            0.000000
Fare            14.454200
dtype: float64
```

In [24]:
```
dataset['Fare']= np.where(dataset['Fare']>upper_limit,30,dataset['Fare'])
```

In [25]:
```
sns.boxplot(dataset["Fare"])
```

Out[25]:
```
<Axes: >
```



6. Seperate Dependent and Independent variables.

In [26]:
```
dataset.head()
```

Out[26]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	E
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.250	B96 B98	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	30.000	C85	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925	B96 B98	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.100	C123	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.050	B96 B98	

In [27]:
```
y=dataset.iloc[:,1:2]
```

In [28]:
```
x=dataset.drop(labels='Survived',axis=1)
```

In [29]:
```
x.head()
```

Out[29]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.250	B96 B98	S
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	30.000	C85	C
2	3	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925	B96 B98	S
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.100	C123	S
4	5	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.050	B96 B98	S

In [30]:
```
y
```

Out[30]:

	Survived
0	0
1	1
2	1
3	1
4	0
...	...
886	0
887	1
888	0
889	1
890	0

891 rows × 1 columns

6. Encoding

In [31]:
```
dataset.head()
```

Out[31]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	E
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.250	B96 B98	
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	30.000	C85	
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925	B96 B98	
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.100	C123	
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.050	B96 B98	

In [32]:
```
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
```

In [33]:
```
x["Sex"]=le.fit_transform(x["Sex"])
x["Sex"]
```

Out[33]:
```
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 [34]:
```
x["Sex"].value_counts()
```

Out[34]:
```
1    577
0    314
Name: Sex, dtype: int64
```

In [35]:
```
x["Sex"].nunique()
```

Out[35]:
```
2
```

In [36]:
```
x.Embarked.value_counts()
```

Out[36]:
```
S    646
C    168
Q     77
Name: Embarked, dtype: int64
```

In [37]:
```
x.shape
```

Out[37]:
```
(891, 11)
```

In [38]:
```
embarked=pd.get_dummies(x["Embarked"],drop_first=True)
embarked
```

Out[38]:

	Q	S
0	0	1
1	0	0
2	0	1
3	0	1
4	0	1
...	...	...
886	0	1
887	0	1
888	0	1
889	0	0
890	1	0

891 rows × 2 columns

In [39]:
```
x=pd.concat([x,embarked],axis=1)
```

In [40]:
```
x.head()
```

Out[40]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Q
0	1	3	Braund, Mr. Owen Harris	1	22.0	1	0	A/5 21171	7.250	B96 B98	S	0
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	PC 17599	30.000	C85	C	0

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Q
2	3	3	Heikkinen, Miss. Laina	0	26.0	0	0	STON/O2. 3101282	7.925	B96 B98	S	0
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	113803	53.100	C123	S	0
4	5	3	Allen, Mr. William Henry	1	35.0	0	0	373450	8.050	B96 B98	S	0

In [41]:
```
x.drop(["Embarked"],axis=1,inplace=True)
```

In [42]:
```
x.head()
```

Out[42]:

	PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Q	S
0	1	3	Braund, Mr. Owen Harris	1	22.0	1	0	A/5 21171	7.250	B96 B98	0	1
1	2	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	0	38.0	1	0	PC 17599	30.000	C85	0	0
2	3	3	Heikkinen, Miss. Laina	0	26.0	0	0	STON/O2. 3101282	7.925	B96 B98	0	1
3	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	0	35.0	1	0	113803	53.100	C123	0	1
4	5	3	Allen, Mr. William Henry	1	35.0	0	0	373450	8.050	B96 B98	0	1

In [43]:
```
x.drop(["Name"],axis=1,inplace=True)
```

In [44]:
```
x.head()
```

Out[44]:

	PassengerId	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Q	S
0	1	3	1	22.0	1	0	A/5 21171	7.250	B96 B98	0	1
1	2	1	0	38.0	1	0	PC 17599	30.000	C85	0	0
2	3	3	0	26.0	0	0	STON/O2. 3101282	7.925	B96 B98	0	1
3	4	1	0	35.0	1	0	113803	53.100	C123	0	1

	PassengerId	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Q	S
4	5	3	1	35.0	0	0	373450	8.050	B96 B98	0	1

In [45]:
```
x.drop(["Ticket"],axis=1,inplace=True)
```

In [46]:
```
x.drop(["Cabin"],axis=1,inplace=True)
```

In [47]:
```
x.head()
```

Out[47]:

	PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare	Q	S
0	1	3	1	22.0	1	0	7.250	0	1
1	2	1	0	38.0	1	0	30.000	0	0
2	3	3	0	26.0	0	0	7.925	0	1
3	4	1	0	35.0	1	0	53.100	0	1
4	5	3	1	35.0	0	0	8.050	0	1

I have dropped the Name, Ticket and Cabin columns because they are all object type and can't be converted from categorical to numerical type.

8. Splitting into training and testing set

We need to split a dataset into train and test sets to evaluate how well our machine learning model performs. The train set is used to fit the model, and the statistics of the train set are known. The second set is called the test data set, this set is solely used for predictions.

In [48]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=0)
```

In [49]:
```
x_train.shape,x_test.shape
```

Out[49]:
```
((623, 9), (268, 9))
```

In [50]:
```
y_train.shape,y_test.shape
```

Out[50]:
```
((623, 1), (268, 1))
```

9. Perform feature scaling

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.

In [51]:
```
from sklearn.preprocessing import StandardScaler
sc=StandardScaler()
```

In [52]:
```
x_train=sc.fit_transform(x_train)
x_test=sc.fit_transform(x_test)
```

In [53]:
```
x_train
```

Out[53]:
```
array([[ 1.59014094, -1.5325562 ,  0.72592065, ...,  0.49567463,
        -0.31426968,  0.59774449],
       [-1.52952238, -1.5325562 , -1.37756104, ...,  0.75298828,
        -0.31426968, -1.67295561],
       [-0.23515275,  0.84844757,  0.72592065, ...,  2.01345222,
        -0.31426968,  0.59774449],
       ...,
       [ 0.70655928,  0.84844757,  0.72592065, ..., -0.90774382,
         3.18198052, -1.67295561],
       [ 0.43528421,  0.84844757, -1.37756104, ..., -0.1867659 ,
        -0.31426968,  0.59774449],
       [ 0.91970398, -0.34205431,  0.72592065, ...,  1.42424126,
        -0.31426968,  0.59774449]])
```

In [54]:
```
x_test
```

Out[54]:
```
array([[ 0.21119888,  0.77963055,  0.76537495, ..., -0.29235767,
        -0.29158231, -1.51942159],
       [ 0.8106727 ,  0.77963055,  0.76537495, ..., -0.82457025,
        -0.29158231,  0.65814518],
       [-0.63903523,  0.77963055,  0.76537495, ...,  0.83755883,
         3.42956335, -1.51942159],
       ...,
       [ 0.70096507,  0.77963055,  0.76537495, ..., -0.29267353,
        -0.29158231, -1.51942159],
       [ 1.35137458,  0.77963055, -1.30654916, ..., -0.82874579,
        -0.29158231,  0.65814518],
       [-1.47751496, -1.64991582,  0.76537495, ...,  0.72937985,
        -0.29158231, -1.51942159]])
```

In [ ]:
```

```