```
NAME: CHINNAM SRI SAI SUMANTH REDDY
REG.NO: 21BCE9129
```

Data Preprocessing

1. Import the Libraries

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

2.Importing the dataset

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

In [3]:
```
dataset
```

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

Out[6]:

	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

3.Checking for Null values

In [7]:
```
dataset.isnull().any()
```

Out[7]:
```
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 [8]:
```
dataset.isnull().sum()
```

Out[8]:
```
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 [9]:
```
dataset['Age'].fillna(dataset['Age'].mean(),inplace=True)
```

In [10]:
```
dataset['Embarked'].fillna(dataset['Embarked'].mode()[0],inplace=True)
```

In [11]:
```
dataset['Has_Cabin'] = np.where(dataset['Cabin'].isnull(),'No','Yes')
```

In [12]:
```
dataset.drop('Cabin',axis=1,inplace=True)
```

4.Data Visualization

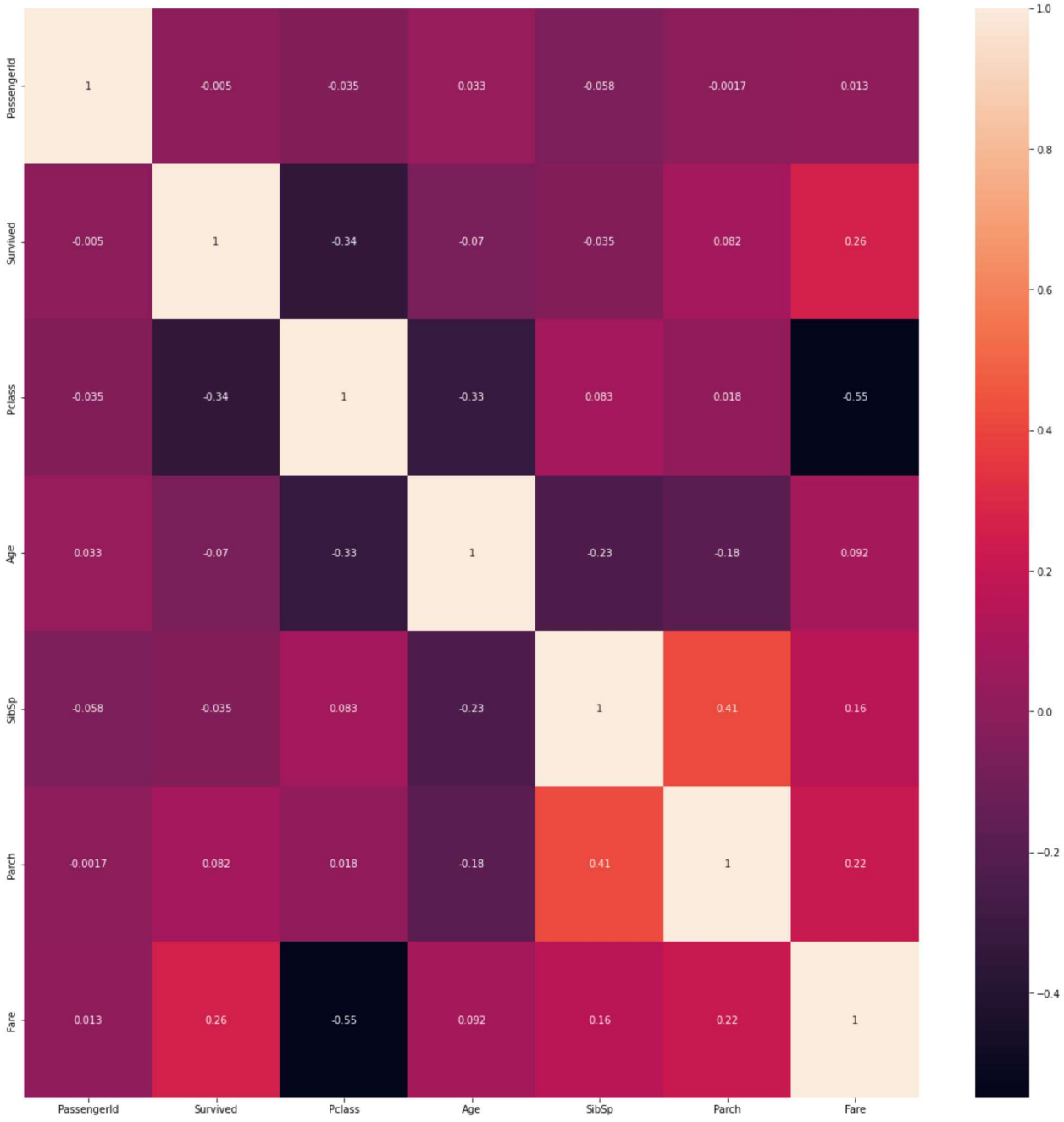
In [13]:
```
cor = dataset.corr()
cor
```

Out[13]:

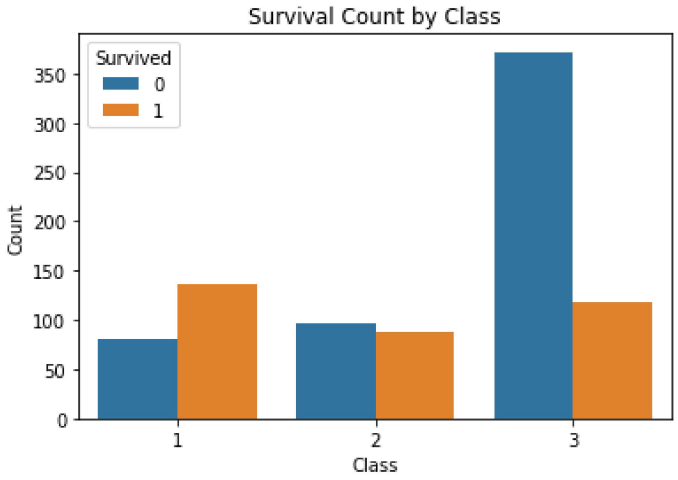
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005007	-0.035144	0.033207	-0.057527	-0.001652	0.012658
Survived	-0.005007	1.000000	-0.338481	-0.069809	-0.035322	0.081629	0.257307
Pclass	-0.035144	-0.338481	1.000000	-0.331339	0.083081	0.018443	-0.549500
Age	0.033207	-0.069809	-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 [14]:
```
plt.subplots(figsize=(20,20))
sns.heatmap(cor,annot=True)
```

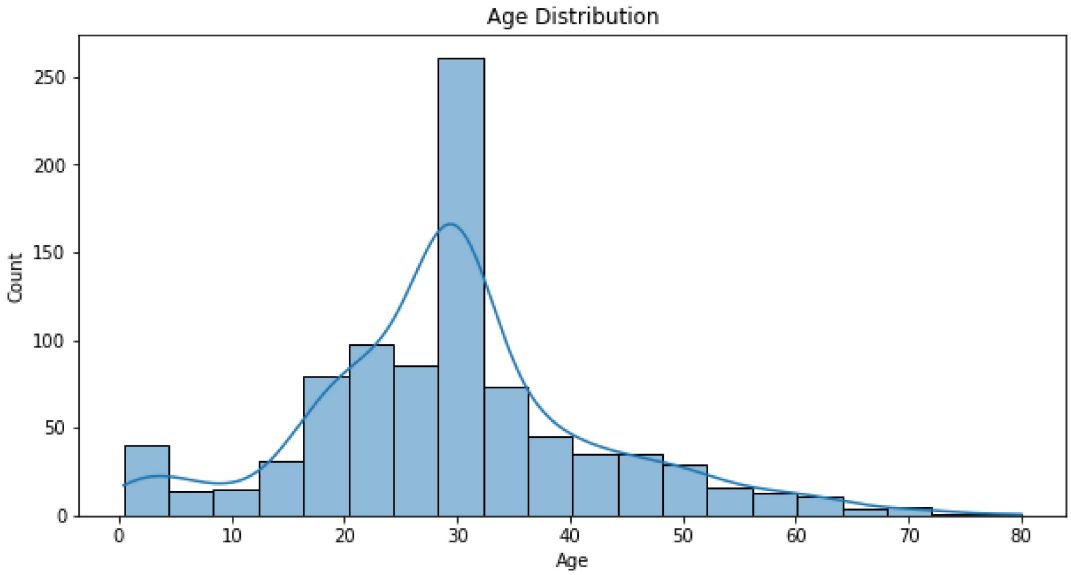
Out[14]: <AxesSubplot:>



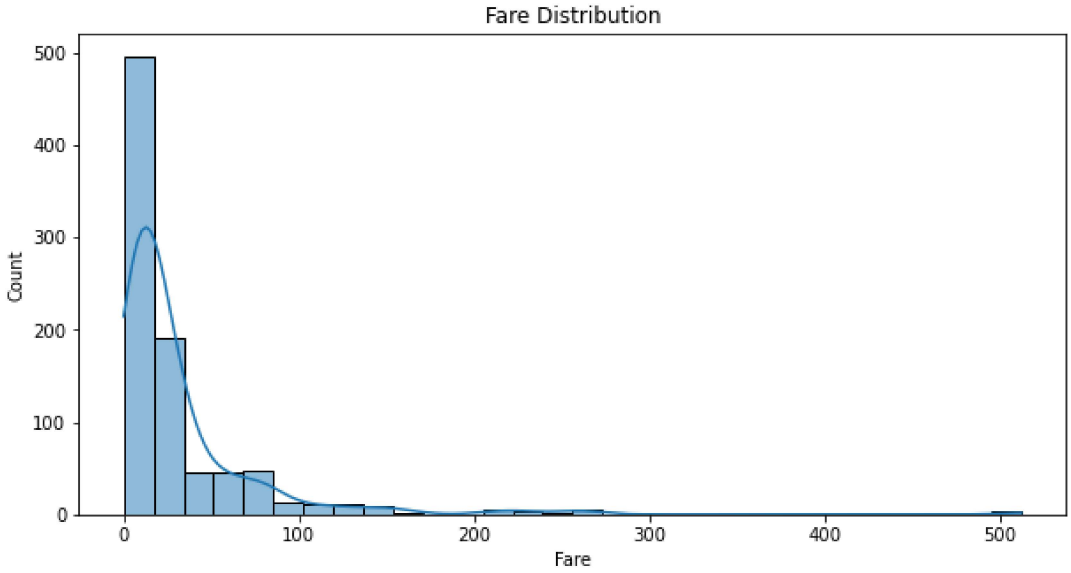
In [15]:
```
#survival count by class
sns.countplot(data=dataset,x='Pclass',hue='Survived')
plt.title('Survival Count by Class')
plt.xlabel('Class')
plt.ylabel('Count')
plt.show()
```



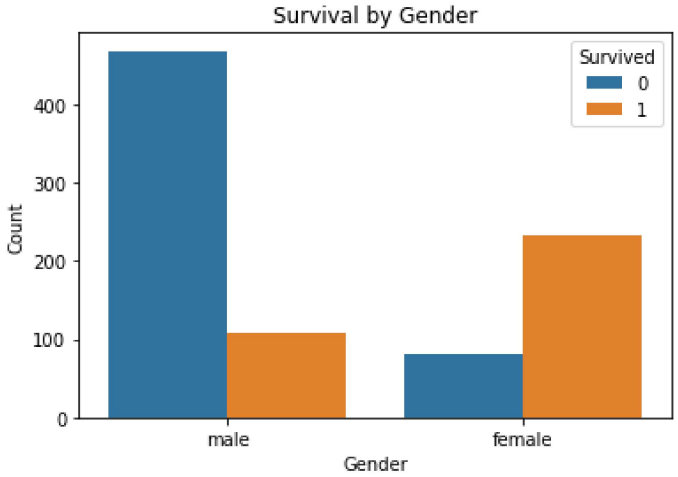
In [16]:
```
#Age distribution
plt.figure(figsize=(10,5))
sns.histplot(data=dataset,x='Age',bins=20,kde=True)
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Count')
plt.show()
```



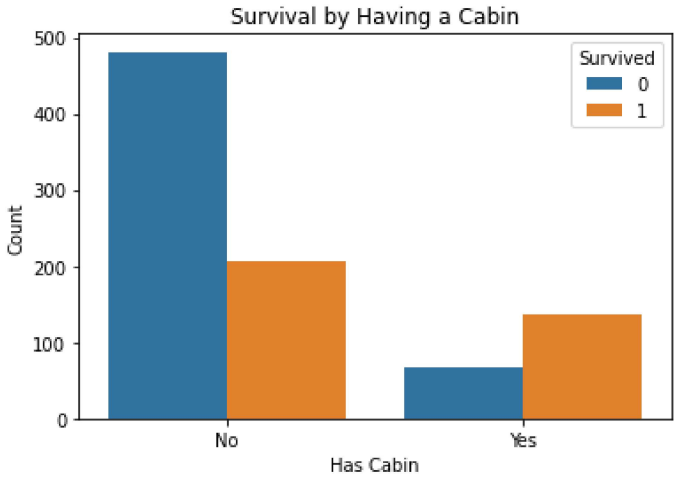
In [17]:
```
#Fare Distribution
plt.figure(figsize=(10,5))
sns.histplot(data = dataset,x='Fare',bins=30,kde=True)
plt.title('Fare Distribution')
plt.xlabel('Fare')
plt.ylabel('Count')
plt.show()
```



In [18]:
```
sns.countplot(data=dataset,x='Sex',hue='Survived')
plt.title('Survival by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```



In [19]:
```
sns.countplot(data=dataset, x='Has_Cabin', hue='Survived')
plt.title('Survival by Having a Cabin')
plt.xlabel('Has Cabin')
plt.ylabel('Count')
plt.show()
```

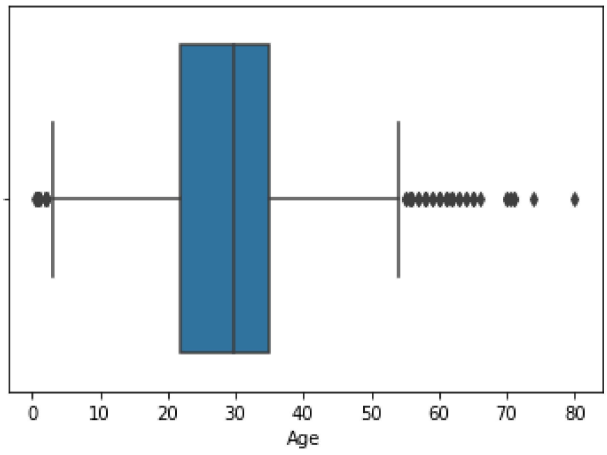


Outlier Detection

In [20]:
```
sns.boxplot(dataset['Age'])
```

```
C:\Users\suman\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass th
e following variable as a keyword arg: x. From version 0.12, the only valid positional argum
ent will be `data`, and passing other arguments without an explicit keyword will result in a
n error or misinterpretation.
  warnings.warn(
```

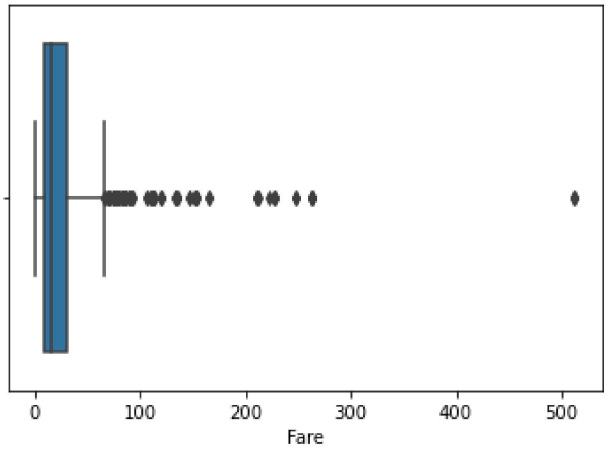
Out[20]: <AxesSubplot:xlabel='Age'>



In [21]:
```
sns.boxplot(dataset['Fare'])
```

```
C:\Users\suman\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass th
e following variable as a keyword arg: x. From version 0.12, the only valid positional argum
ent will be `data`, and passing other arguments without an explicit keyword will result in a
n error or misinterpretation.
  warnings.warn(
```

Out[21]: <AxesSubplot:xlabel='Fare'>



In [22]:
```
Q1_age = dataset['Age'].quantile(0.25)
print(Q1_age)
Q3_age = dataset['Age'].quantile(0.75)
print(Q3_age)
IQR_age = Q3_age - Q1_age
print(IQR_age)
```

```
22.0
35.0
13.0
```

In [23]:
```
lower_bound_age = Q1_age - 1.5 * IQR_age
print(lower_bound_age)
upper_bound_age = Q3_age + 1.5 * IQR_age
print(upper_bound_age)
```

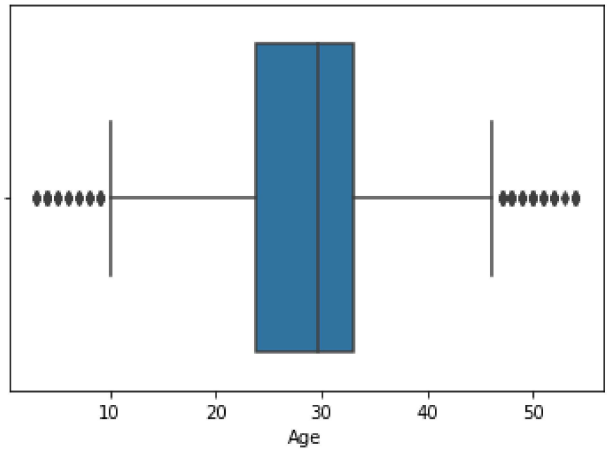
```
2.5
54.5
```

In [24]:
```
dataset['Age']=np.where(dataset['Age']>upper_bound_age,dataset['Age'].median(),dataset['Age']]
dataset['Age']=np.where(dataset['Age']<lower_bound_age,dataset['Age'].median(),dataset['Age']]
```

In [25]:
```
sns.boxplot(dataset['Age'])
```

```
C:\Users\suman\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass th
e following variable as a keyword arg: x. From version 0.12, the only valid positional argum
ent will be `data`, and passing other arguments without an explicit keyword will result in a
n error or misinterpretation.
  warnings.warn(
```

Out[25]: <AxesSubplot:xlabel='Age'>



In [26]:
```
Q1_fare = dataset['Fare'].quantile(0.25)
Q3_fare = dataset['Fare'].quantile(0.75)
IQR_fare = Q3_fare - Q1_fare
```

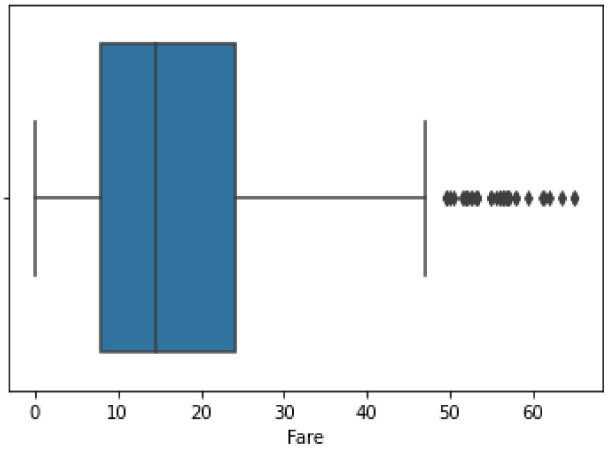
In [27]:
```
lower_bound_fare = Q1_fare - 1.5 * IQR_fare
upper_bound_fare = Q3_fare + 1.5 * IQR_fare
```

In [28]:
```
dataset['Fare']=np.where(dataset['Fare']>upper_bound_fare,dataset['Fare'].median(),dataset['Fa
dataset['Fare']=np.where(dataset['Fare']<lower_bound_fare,dataset['Fare'].median(),dataset['Fa
```

In [29]:
```
sns.boxplot(dataset['Fare'])
```

```
C:\Users\suman\anaconda3\lib\site-packages\seaborn\_decorators.py:36: FutureWarning: Pass th
e following variable as a keyword arg: x. From version 0.12, the only valid positional argum
ent will be `data`, and passing other arguments without an explicit keyword will result in a
n error or misinterpretation.
  warnings.warn(
```

Out[29]: <AxesSubplot:xlabel='Fare'>



Splitting Dependent and Independent variables

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

Out[30]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Embarked	Has_Cabin
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	S	No
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	14.4542	C	Yes
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	S	No
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	S	Yes
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	S	No

In [36]:
```
print(dataset.Name.nunique())
print(dataset.PassengerId.nunique())
print(dataset.Ticket.nunique())
```

```
891
891
681
```

In [40]:
```
dataset.drop(columns=['PassengerId', 'Name', 'Ticket'],inplace=True)
```

In [41]:
```
dataset.head()
```

Out[41]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Has_Cabin
0	0	3	male	22.0	1	0	7.2500	S	No
1	1	1	female	38.0	1	0	14.4542	C	Yes
2	1	3	female	26.0	0	0	7.9250	S	No
3	1	1	female	35.0	1	0	53.1000	S	Yes
4	0	3	male	35.0	0	0	8.0500	S	No

In [42]:
```
x=dataset.drop('Survived',axis=1)
y=dataset['Survived']
print(x)
print(y)
```

```
     Pclass     Sex        Age  SibSp  Parch     Fare Embarked Has_Cabin
0         3    male  22.000000      1      0   7.2500        S        No
1         1  female  38.000000      1      0  14.4542        C       Yes
2         3  female  26.000000      0      0   7.9250        S        No
3         1  female  35.000000      1      0  53.1000        S       Yes
4         3    male  35.000000      0      0   8.0500        S        No
..      ...     ...        ...    ...    ...      ...      ...       ...
886       2    male  27.000000      0      0  13.0000        S        No
887       1  female  19.000000      0      0  30.0000        S       Yes
888       3  female  29.699118      1      2  23.4500        S        No
889       1    male  26.000000      0      0  30.0000        C       Yes
890       3    male  32.000000      0      0   7.7500        Q        No

[891 rows x 8 columns]
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
```

Perform Encoding

In [43]:
```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
```

In [56]:
```
x['Sex'] = le.fit_transform(x['Sex'])
x['Embarked'] = le.fit_transform(x['Embarked'])
x['Has_Cabin'] = le.fit_transform(x['Has_Cabin'])
```

In [58]:
```
x.head(2)
```

Out[58]:

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Has_Cabin
0	3	1	22.0	1	0	7.2500	2	0
1	1	0	38.0	1	0	14.4542	0	1

Feature Scaling

In [49]:
```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
```

In [59]:
```
x[['Age','Fare']] = sc.fit_transform(x[['Age','Fare']])
```

In [60]:
```
x.head()
```

Out[60]:

	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked	Has_Cabin
0	3	1	-0.708584	1	0	-0.797554	2	0
1	1	0	0.924948	1	0	-0.230556	0	1
2	3	0	-0.300201	0	0	-0.744429	2	0
3	1	0	0.618661	1	0	2.811012	2	1
4	3	1	0.618661	0	0	-0.734591	2	0

Splitting Data into Train and Test

In [61]:
```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=0.2,random_state=0)
```

In [62]:
```
print(x_train.shape)
print(x_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(712, 8)
(179, 8)
(712,)
(179,)
```

In [63]:
```
print(x_train)
```

```
     Pclass  Sex       Age  SibSp  Parch      Fare  Embarked  Has_Cabin
140       3    0  0.077463      0      2 -0.168255         0          0
439       2    1  0.210278      0      0 -0.541767         2          0
817       2    1  0.210278      1      1  1.544212         0          0
378       3    1 -0.912775      0      0 -1.052357         0          0
491       3    1 -0.810680      0      0 -0.797554         2          0
..      ...  ...       ...    ...    ...       ...       ...        ...
835       1    0  1.027043      1      1 -0.230556         0          1
192       3    0 -1.014871      1      0 -0.750001         2          0
629       3    1  0.077463      0      0 -0.759516         1          0
559       3    0  0.720756      1      0  0.001289         2          0
684       2    1  0.077463      1      1  1.701289         2          0

[712 rows x 8 columns]
```

In [64]:
```
print(x_test)
```

```
     Pclass  Sex       Age  SibSp  Parch      Fare  Embarked  Has_Cabin
495       3    1  0.077463      0      0 -0.230234         0          0
648       3    1  0.077463      0      0 -0.773943         2          0
278       3    1 -2.240020      4      1  0.924090         1          0
31        1    0  0.077463      1      0 -0.230556         0          1
255       3    0  0.006086      0      2 -0.168255         0          0
..      ...  ...       ...    ...    ...       ...       ...        ...
780       3    0 -1.627446      0      0 -0.799191         0          0
837       3    1  0.077463      0      0 -0.734591         2          0
215       1    0  0.210278      1      0 -0.230556         0          1
833       3    1 -0.606488      0      0 -0.750001         2          0
372       3    1 -1.014871      0      0 -0.734591         2          0

[179 rows x 8 columns]
```

In [65]:
```
print(y_train)
```

```
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 [66]:
```
print(y_test)
```

```
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 [ ]:
```

```