TASK-1

Cleaning the Titanic Dataset by removing missing values and Outliers

Importing the Python libraries

In [1]:

```python
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Importing the dataset

In [2]:

```python
df = pd.read_csv('train.csv')                          #reading the file.
df
```

Out[2]:

	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	female	NaN	1	2	W./C. 6607	23.4500	NaN	S

				Helen "Carrie"								
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 [3]: df.head(10)                                  #displaying top 10 rows
```

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
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
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

In [4]:
```
df.tail(10)                                        #displaying last 10 rows
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarkec
881	882	0	3	Markun, Mr. Johann	male	33.0	0	0	349257	7.8958	NaN	S
882	883	0	3	Dahlberg, Miss. Gerda Ulrika	female	22.0	0	0	7552	10.5167	NaN	S
883	884	0	2	Banfield, Mr. Frederick James	male	28.0	0	0	C.A./SOTON 34068	10.5000	NaN	S
884	885	0	3	Sutehall, Mr. Henry Jr	male	25.0	0	0	SOTON/OQ 392076	7.0500	NaN	S
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250	NaN	C
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	C

Getting the datatypes of all columns

In [5]:
```
df.dtypes
```

Out[5]:
```
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
```

Statistical details of Dataframe

In [9]:

```
df.describe()
```

Out[9]:

	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

Data Cleaning

Counting the No of missing values in each column

In [5]:

```
df.isnull().sum()
```

Out[5]:

```
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
```

Calculating the percentage of missing values in Dataframe

In [6]:

```
missing_values=(df.isnull().sum()/len(df))*100
print(missing_values)
```

```
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
```

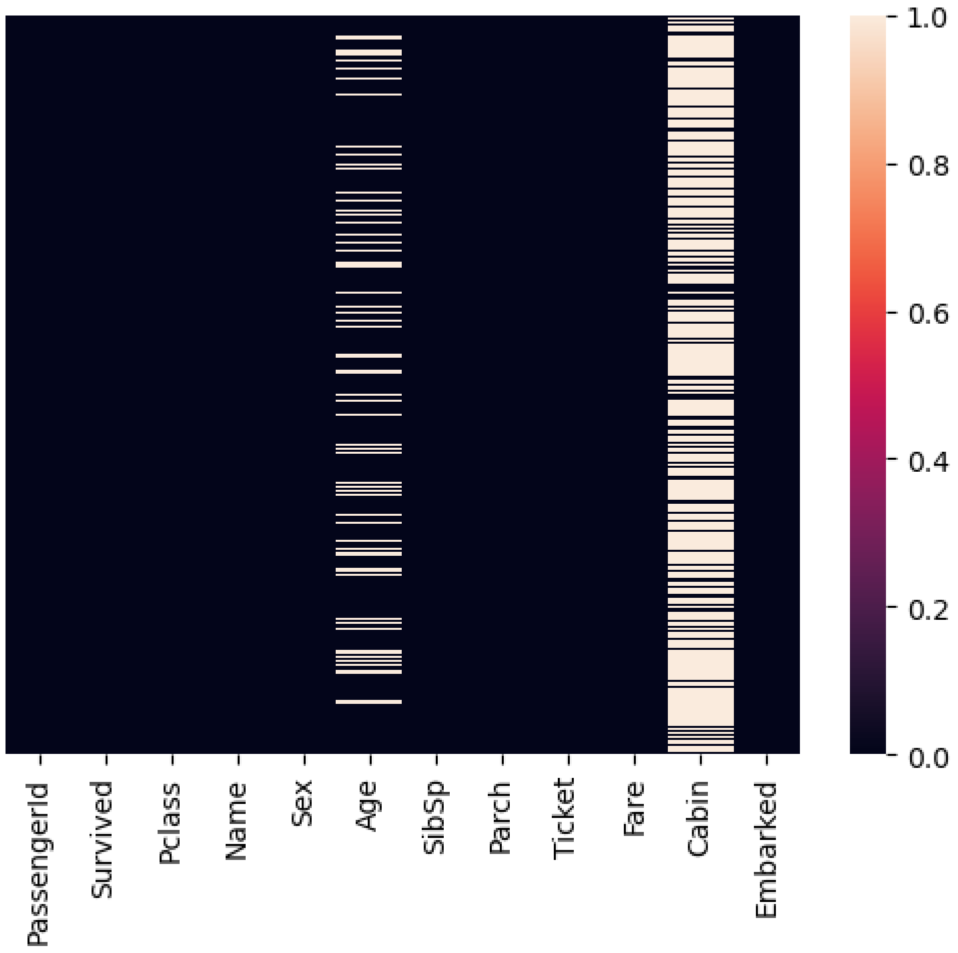
Visualizing the missing data

In [4]:

```
sns.heatmap(df.isnull(), yticklabels=False)
```

Out[4]:

```
<Axes: >
```



In [5]:

```
df.dropna(how="all")                                          #dropping when whole ro
```

Out[5]:

	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,	female	35.0	1	0	113803	53.1000	C123	S

				Mrs. Jacques Heath (Lily May Peel)								
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

Dropping the null values in the column 'Embarked'

In [9]:

```
df.dropna(subset=['Embarked'], inplace=True)
```

In [14]:

```
df.isnull().sum()
```

Out[14]:

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         0
dtype: int64
```

Counting the No of unique values in the Cabin column of the Dataframe

In [15]:

```
df["Cabin"].value_counts()
```

Out[15]:

```
B96 B98        4
G6             4
C23 C25 C27    4
E101           3
C22 C26        3
              ..
E34            1
C7             1
C54            1
E36            1
C148           1
Name: Cabin, Length: 146, dtype: int64
```

Calculating the mode of Cabin column

In [16]:

```
df["Cabin"].mode()
```

Out[16]:

```
0        B96 B98
1    C23 C25 C27
2             G6
Name: Cabin, dtype: object
```

Replacing the missing values with mode values in Cabin column

In [7]:

```
df['Cabin'].fillna(df['Cabin'].mode()[2],inplace=True)   # we replace with any of three 0
```

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            0
Embarked         0
dtype: int64
```

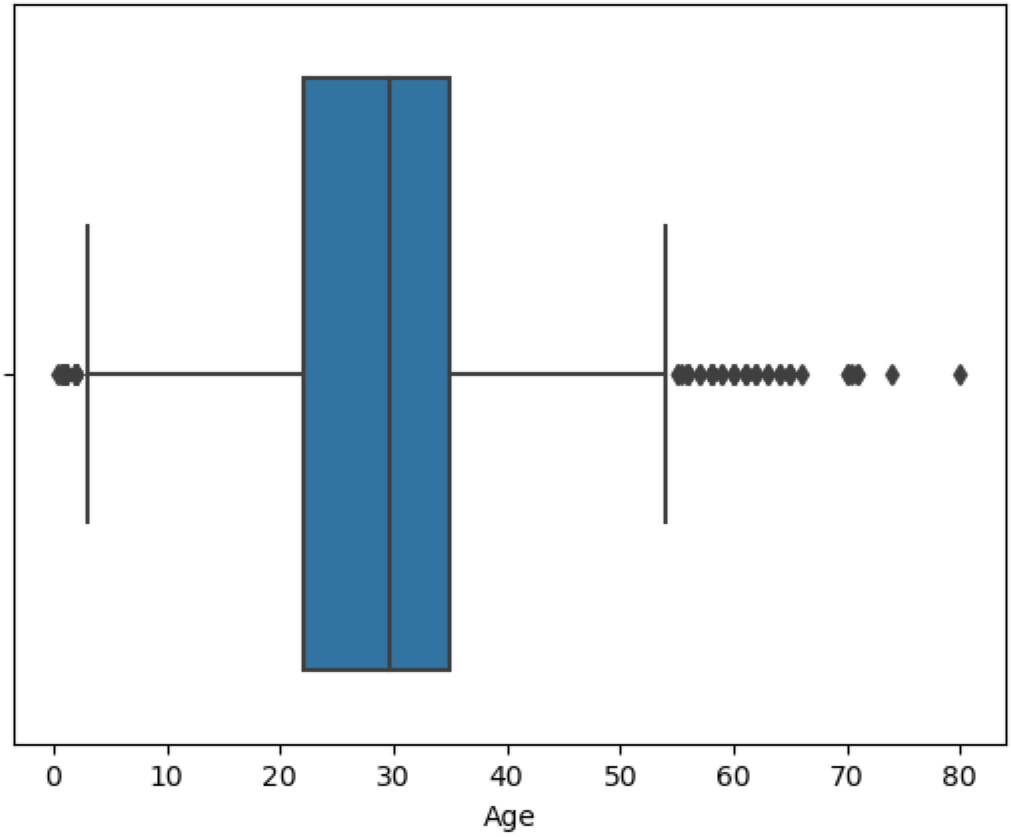
Visualization of Outliers in Age Column

In [70]:

```
sns.boxplot(x=df["Age"],showfliers=True)
```

Out[70]:

```
<Axes: xlabel='Age'>
```



Removing the outliers in Age Column

In [20]:

```
Q1=df.Age.quantile(0.25)
Q3=df.Age.quantile(0.75)

Q1,Q3
```

Out[20]:

```
(20.0, 38.0)
```

In [22]:

```
IQR=Q3-Q1
IQR
```

Out[22]:

```
18.0
```

In [25]:

```
lower_limit=Q1-1.5*IQR
upper_limit=Q3+1.5*IQR
lower_limit,upper_limit
```

Out[25]:

```
(-7.0, 65.0)
```

In [27]:

```
df[(df.Age<lower_limit)|(df.Age>upper_limit)]
```

Out[27]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
33	34	0	2	Wheadon, Mr. Edward H	male	66.0	0	0	C.A. 24579	10.5000	G6	S
96	97	0	1	Goldschmidt, Mr. George B	male	71.0	0	0	PC 17754	34.6542	A5	C
116	117	0	3	Connors, Mr. Patrick	male	70.5	0	0	370369	7.7500	G6	Q

493	494	0	1	Artagaveytia, Mr. Ramon	male	71.0	0	0	PC 17609	49.5042	G6	C
630	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male	80.0	0	0	27042	30.0000	A23	S
672	673	0	2	Mitchell, Mr. Henry Michael	male	70.0	0	0	C.A. 24580	10.5000	G6	S
745	746	0	1	Crosby, Capt. Edward Gifford	male	70.0	1	1	WE/P 5735	71.0000	B22	S
851	852	0	3	Svensson, Mr. Johan	male	74.0	0	0	347060	7.7750	G6	S

In [37]:

```
no_outlier=df[(df.Age>=lower_limit)&(df.Age<=upper_limit)]
no_outlier
```

Out[37]:

	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	G6	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	G6	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	G6	S
...	...	...	...	...	...	...	...	...	...	...	...	...
885	886	0	3	Rice, Mrs. William (Margaret Norton)	female	39.0	0	5	382652	29.1250	G6	Q
886	887	0	2	Montvila, Rev. Juozas	male	27.0	0	0	211536	13.0000	G6	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.	male	26.0	0	0	111369	30.0000	C148	C

				Karl Howell								
890	891	0	3	Dooley, Mr. Patrick	male	32.0	0	0	370376	7.7500	G6	Q

704 rows × 12 columns

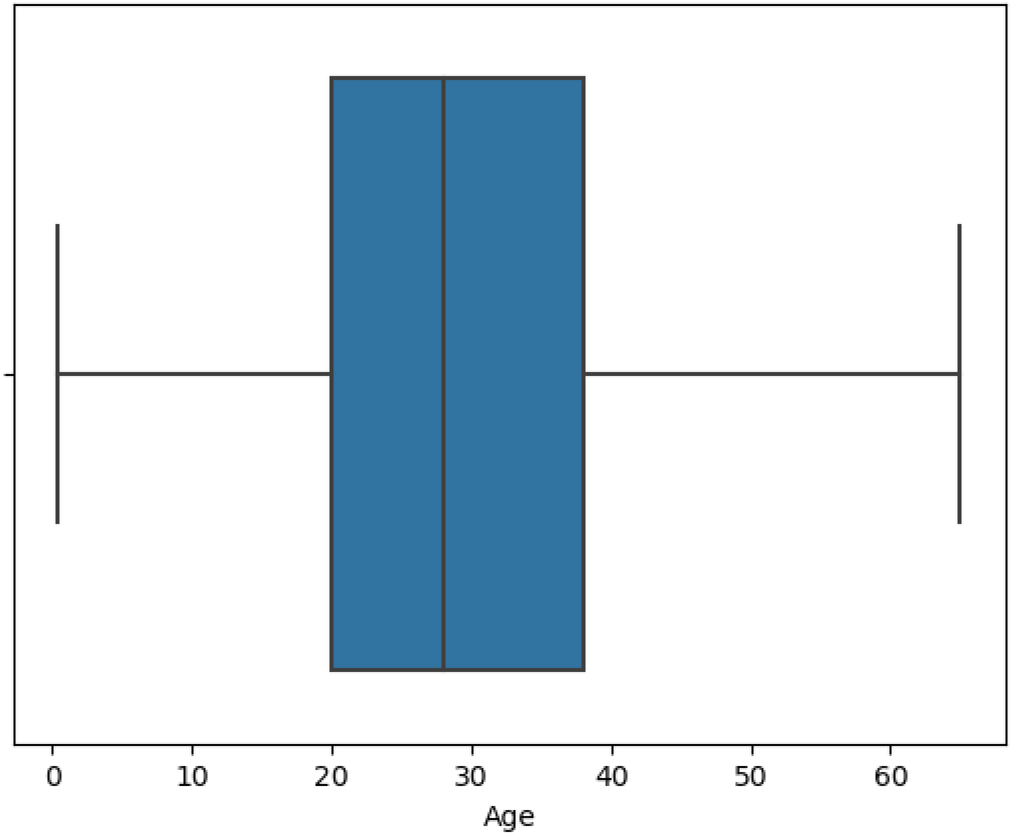
Visualization of Age Column without Outliers

In [79]:

```
sns.boxplot(x=no_outlier["Age"],showfliers=True)
```

Out[79]:

```
<Axes: xlabel='Age'>
```



Calculating the mean of Age column

In [40]:

```
df["Age"].mean()
```

Out[40]:

```
29.64209269662921
```

Replacing the null values of Age by mean

In [45]:

```
df['Age'].fillna(df['Age'].mean(),inplace=True)
```

In [46]:

```
df.isnull().sum()
```

Out[46]:

```
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
```

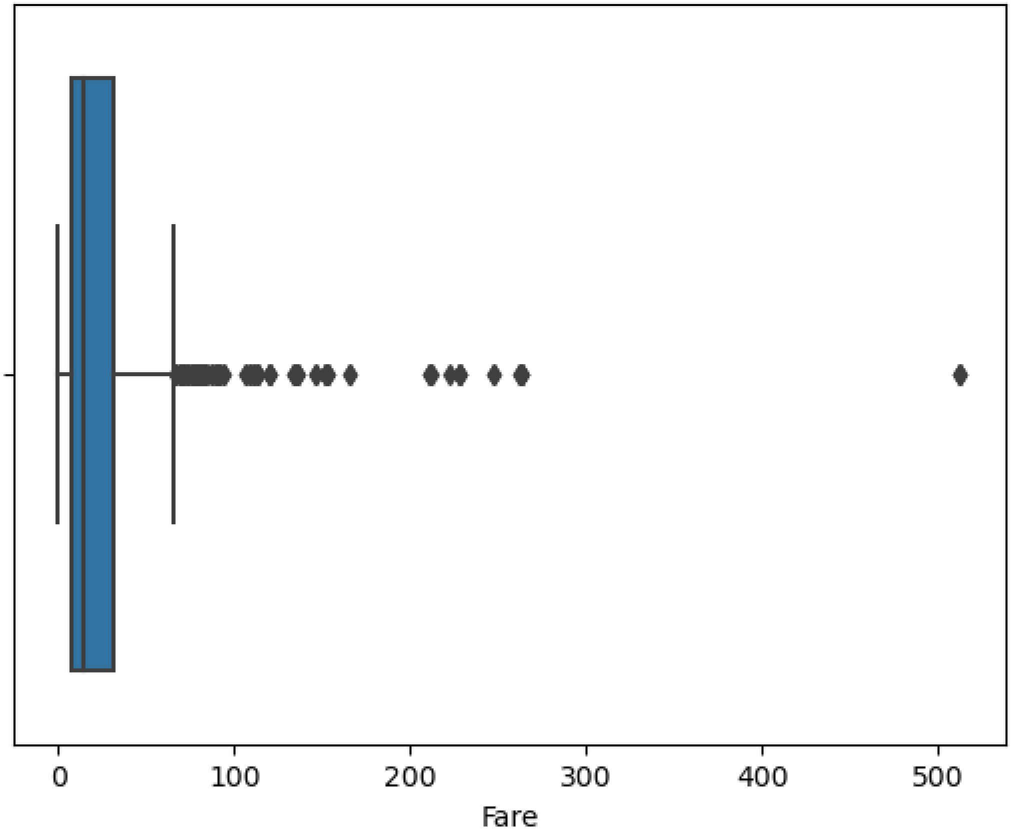
Visualization of Fare Column with outliers

In [71]:

```
sns.boxplot(x=df["Fare"],showfliers=True)
```

Out[71]:

```
<Axes: xlabel='Fare'>
```



Removing the outliers in Fare column

In [76]:

```
Q1=df.Fare.quantile(0.25)
Q3=df.Fare.quantile(0.75)

IQR=Q3-Q1
lower_limit=Q1-1.5*IQR
upper_limit=Q3+1.5*IQR
no_outlier_fare=df[(df.Fare>=lower_limit)&(df.Fare<=upper_limit)]
no_outlier_fare
```

Out[76]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embai
0	1	0	3	Braund, Mr. Owen	male	22.000000	1	0	A/5 21171	7.2500	G6	

				Harris							
2	3	1	3	Heikkinen, Miss. Laina	female	26.000000	0	0	STON/O2. 3101282	7.9250	G6
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.000000	1	0	113803	53.1000	C123
4	5	0	3	Allen, Mr. William Henry	male	35.000000	0	0	373450	8.0500	G6
5	6	0	3	Moran, Mr. James	male	29.642093	0	0	330877	8.4583	G6
...	...	...	...	...	...	...	...	...	...	...	...
886	887	0	2	Montvila, Rev. Juozas	male	27.000000	0	0	211536	13.0000	G6
887	888	1	1	Graham, Miss. Margaret Edith	female	19.000000	0	0	112053	30.0000	B42
888	889	0	3	Johnston, Miss. Catherine Helen "Carrie"	female	29.642093	1	2	W./C. 6607	23.4500	G6
889	890	1	1	Behr, Mr. Karl Howell	male	26.000000	0	0	111369	30.0000	C148
890	891	0	3	Dooley, Mr. Patrick	male	32.000000	0	0	370376	7.7500	G6

775 rows × 12 columns

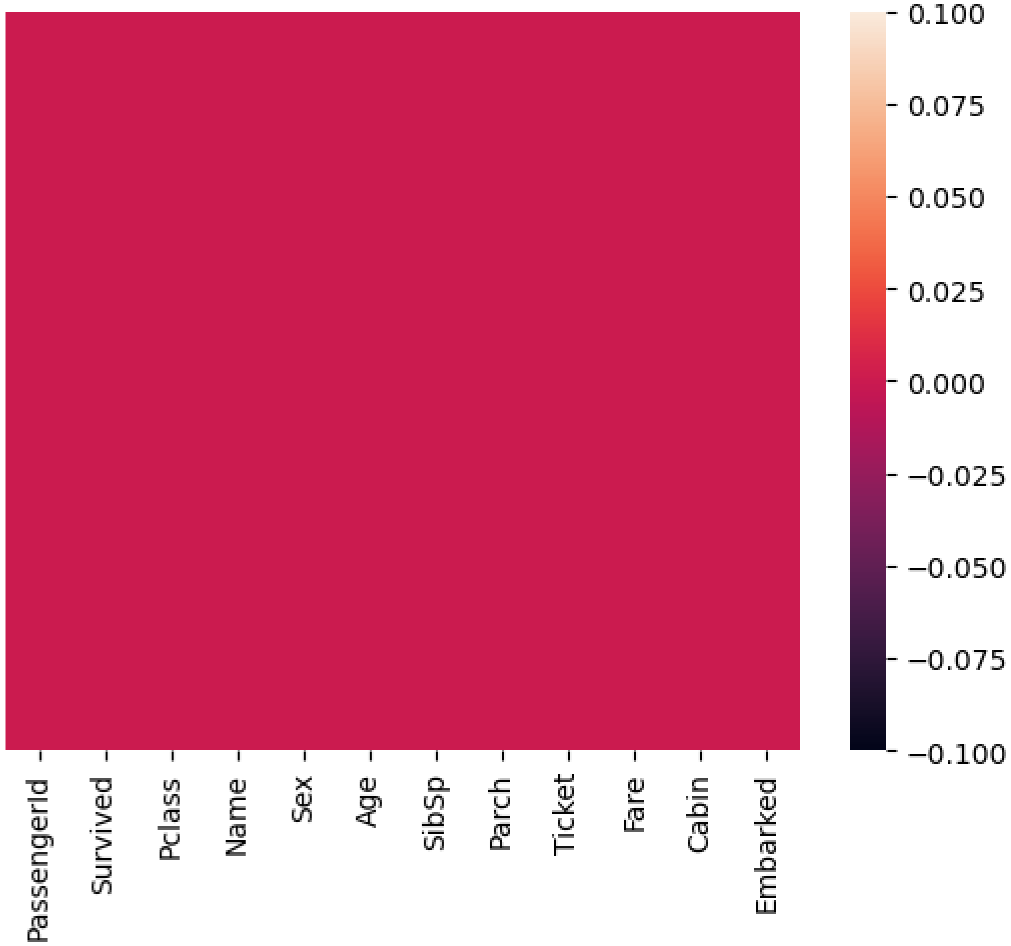
Data is cleaned and outliers has been removed

In [85]:

```
sns.heatmap(df.isnull(), yticklabels=False)
```

Out[85]:

```
<Axes: >
```



Importing the second file

```
In [48]: df1=pd.read_csv("gender_submission.csv")
```

Checking the null values

```
In [49]: df1.isnull().sum()
```

```
Out[49]: PassengerId    0
         Survived       0
         dtype: int64
```

file does not contain null values

Importing the third file

```
In [50]: df2=pd.read_csv("test.csv")
```

Checking the null values

```
In [51]: df2.isnull().sum()
```

Out[51]:
```
PassengerId      0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch            0
Ticket           0
Fare             1
Cabin          327
Embarked         0
dtype: int64
```

Statistical details of Dataframe

In [52]:
```
df2.describe()
```

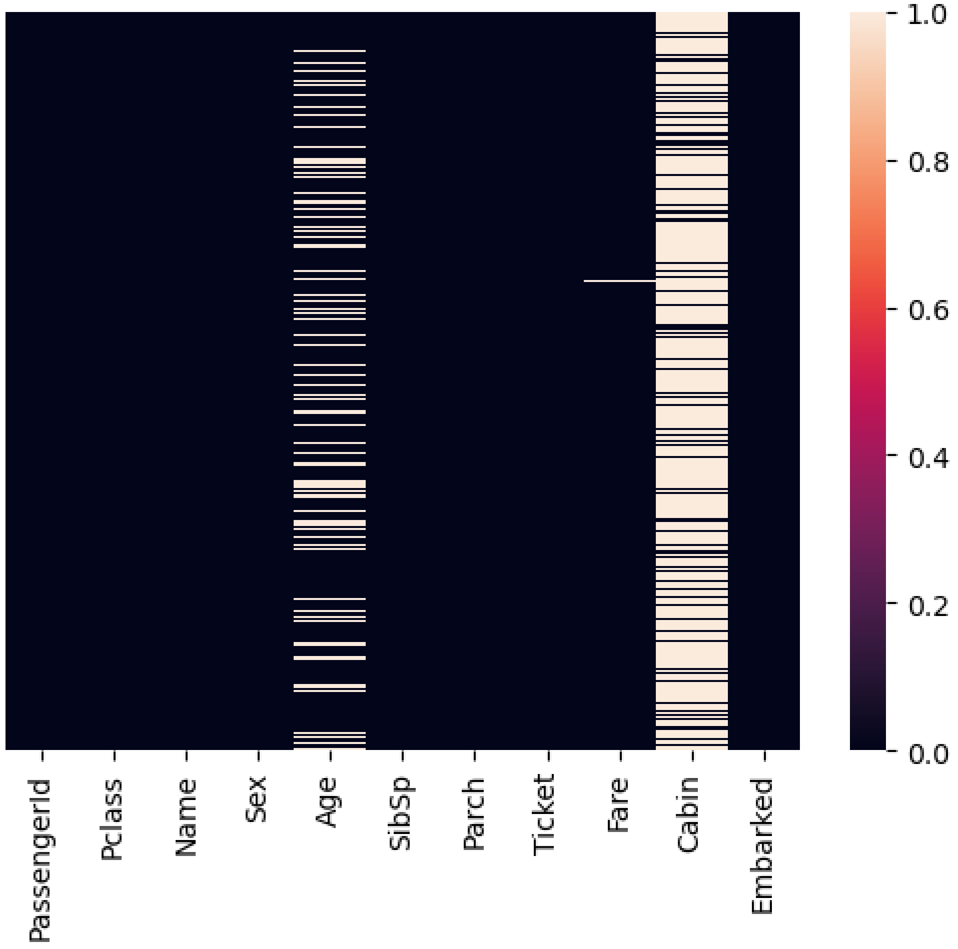
Out[52]:

	PassengerId	Pclass	Age	SibSp	Parch	Fare
count	418.000000	418.000000	332.000000	418.000000	418.000000	417.000000
mean	1100.500000	2.265550	30.272590	0.447368	0.392344	35.627188
std	120.810458	0.841838	14.181209	0.896760	0.981429	55.907576
min	892.000000	1.000000	0.170000	0.000000	0.000000	0.000000
25%	996.250000	1.000000	21.000000	0.000000	0.000000	7.895800
50%	1100.500000	3.000000	27.000000	0.000000	0.000000	14.454200
75%	1204.750000	3.000000	39.000000	1.000000	0.000000	31.500000
max	1309.000000	3.000000	76.000000	8.000000	9.000000	512.329200

Visualizing the missing data

In [54]:
```
sns.heatmap(df2.isnull(), yticklabels=False)
```

Out[54]:
```
<Axes: >
```



calculating the mean of Age column

In [55]:

```
df2["Age"].mean()
```

Out[55]:

```
30.272590361445783
```

Replacing the null values of Age by mean

In [60]:

```
df2['Age'].fillna(df['Age'].mean(),inplace=True)
```

Dropping the null values of fare column

In [ ]:

```
df2.dropna(subset=['Fare'], inplace=True)
```

In [61]:

```
df2.isnull().sum()
```

Out[61]:

```
PassengerId      0
Pclass           0
Name             0
Sex              0
Age              0
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          326
```

```
Embarked          0
dtype: int64
```

Calculating the mode of Cabin column

In [62]:

```
df2["Cabin"].mode()
```

Out[62]:

```
0    B57 B59 B63 B66
Name: Cabin, dtype: object
```

Replacing the missing values with mode values in Cabin column

In [68]:

```
df2['Cabin'].fillna(df2['Cabin'].mode()[0],inplace=True)
```

Null Values are removed

In [69]:

```
df2.isnull().sum()
```

Out[69]:

```
PassengerId    0
Pclass         0
Name           0
Sex            0
Age            0
SibSp          0
Parch          0
Ticket         0
Fare           0
Cabin          0
Embarked       0
dtype: int64
```