Task :- 1

In [1]:

```
# basic python package
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import sklearn
```

In [2]:

```
# importing the required files
# IMPORTING DATASETS

train_df = pd.read_csv("train.csv")
test_df = pd.read_csv("test.csv")
```

In [3]:

```
#top 5 rows of dataset
train_df.head()
```

Out[3]:

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

In [4]:

```
train_df.head(10)
```

Out[4]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarke
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	unknown	
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	unknown	
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	unknown	
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	unknown	
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	unknown	
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	unknown	
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	unknown	

In [5]:

```
#top bottom rows
train_df.tail()
```

Loading [MathJax]/extensions/Safe.js

Out[5]:

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

In [6]:

```
train_df.shape
```

Out[6]:

```
(891, 12)
```

In [7]:

```
train_df.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
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]:

```
#to know the columns of the dataset
train_df.columns
```

Out[8]:

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

In [9]:

```
test_df.columns
```

Out[9]:

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

In [10]:

```
train_df.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]: train_df.size
```

```
Out[11]: 10692
```

```
In [12]: train_df.count()
```

```
Out[12]: PassengerId    891
         Survived       891
         Pclass         891
         Name           891
         Sex            891
         Age            714
         SibSp          891
         Parch          891
         Ticket         891
         Fare           891
         Cabin          891
         Embarked       889
         dtype: int64
```

```
In [13]: print(train_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        891 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
None
```

```
In [14]: train_df["Age"].value_counts()
```

Out[14]:

```
24.00    30
22.00    27
18.00    26
19.00    25
28.00    25
         ..
36.50     1
55.50     1
0.92      1
23.50     1
74.00     1
Name: Age, Length: 88, dtype: int64
```

In [15]:

```
train_df["Sex"].value_counts()
```

Out[15]:

```
male      577
female    314
Name: Sex, dtype: int64
```

In [16]:

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

Out[16]:

```
unknown        687
C23 C25 C27      4
G6               4
B96 B98          4
C22 C26          3
              ...
E34              1
C7               1
C54              1
E36              1
C148             1
Name: Cabin, Length: 148, dtype: int64
```

In [17]:

```
train_df["Cabin"]
```

Out[17]:

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

In [18]:

```
train_df["Fare"].value_counts()
```

Out[18]:

```
8.0500     43
13.0000    42
7.8958     38
7.7500     34
26.0000    31
           ..
35.0000     1
28.5000     1
6.2375      1
14.0000     1
10.5167     1
Name: Fare, Length: 248, dtype: int64
```

In [19]:

```
train_df["Fare"]
```

Loading [MathJax]/extensions/Safe.js

```
Out[19]: 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 [20]: train_df["Survived"]
```

```
Out[20]: 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 [21]: train_df["Survived"].value_counts()
```

```
Out[21]: 0    549
         1    342
         Name: Survived, dtype: int64
```

```
In [22]: #show null values
         train_df.isnull()
```

Out[22]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...
886	False	False	False	False	False	False	False	False	False	False	False	False
887	False	False	False	False	False	False	False	False	False	False	False	False
888	False	False	False	False	False	True	False	False	False	False	False	False
889	False	False	False	False	False	False	False	False	False	False	False	False
890	False	False	False	False	False	False	False	False	False	False	False	False

891 rows × 12 columns

```
In [23]: #how many null values
         print(train_df.isnull().sum())
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin            0
Embarked         2
dtype: int64
```

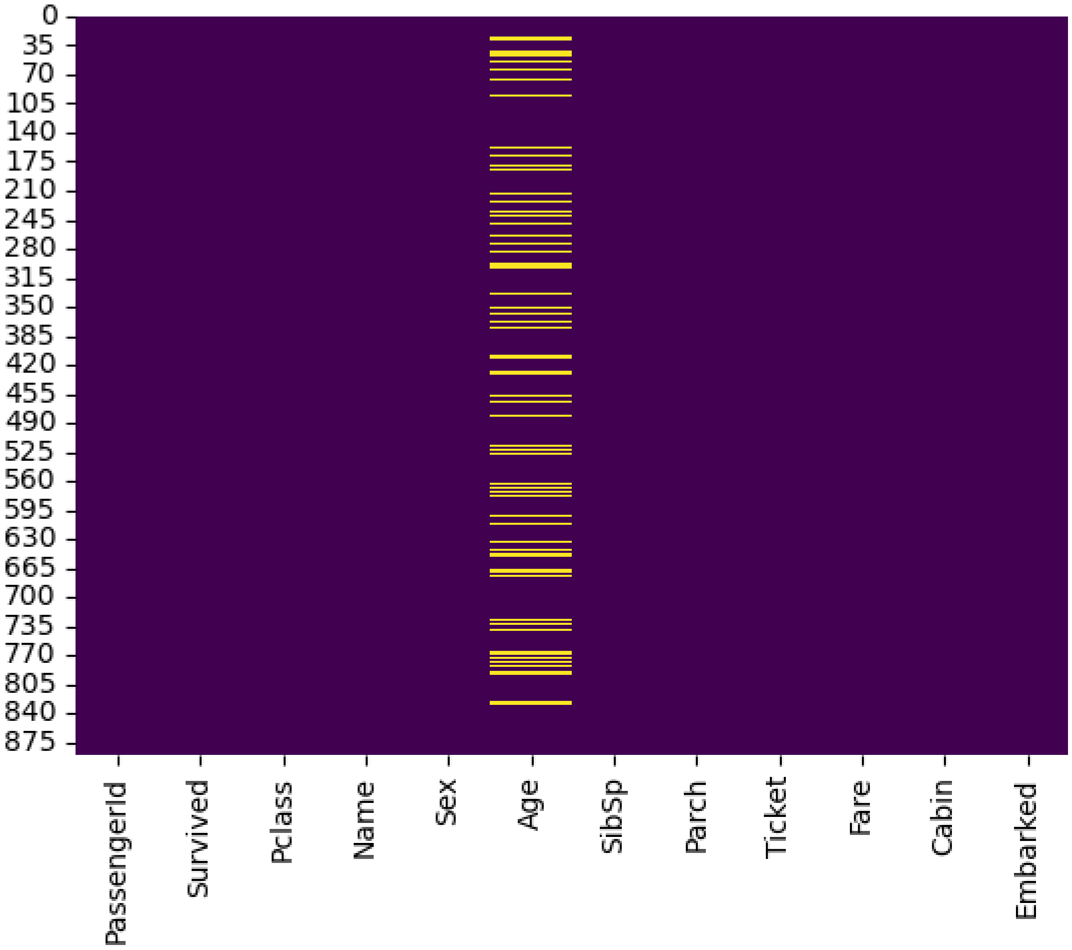
In [24]:

```
#percentage of missing values
missing_percentage = (train_df.isnull().sum()/len(train_df)*100)
print(missing_percentage)
```

```
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           0.000000
Embarked        0.224467
dtype: float64
```

In [25]:

```
sns.heatmap(train_df.isnull(), cmap = 'viridis', cbar = False)
plt.show()
```



In [26]:

```
#to remove null values use fillna
train_df.Cabin = train_df.Cabin.fillna("unknown")
print(train_df.isnull().sum())
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin            0
Embarked         2
dtype: int64
```

In [27]:

```
train_df.head(10)
```

Out[27]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarke
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	unknown	
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	unknown	
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	unknown	
5	6	0	3	Moran, Mr. James	male	NaN	0	0	330877	8.4583	unknown	
6	7	0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	
7	8	0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.0750	unknown	
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333	unknown	
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708	unknown	

Handling the missing values¶

so here for the numerical values fill it with mean and for categorical value fill it with "Unknown"

In [28]:

```
categorical_columns = train_df.select_dtypes(include = ["int64"]).columns
train_df[categorical_columns] = train_df[categorical_columns].fillna("Unknown")
```

In [29]:

```
# check if any missing values is left after handling
print(train_df.isnull().sum())
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin            0
Embarked         2
dtype: int64
```

In [30]:

```
#so here you can seenumber of elements in each columns are equal now that mean there is
train_df.count()
```

Out[30]:

```
PassengerId    891
Survived       891
Pclass         891
Name           891
Sex            891
Age            714
SibSp          891
Parch          891
Ticket         891
Fare           891
Cabin          891
Embarked       889
dtype: int64
```

In [31]:

```
train_df.count().T
```

Out[31]:

```
PassengerId    891
Survived       891
Pclass         891
Name           891
Sex            891
Age            714
SibSp          891
Parch          891
Ticket         891
Fare           891
Cabin          891
Embarked       889
dtype: int64
```

In [32]:

```
train_df.to_csv('train.csv', index = False)
```

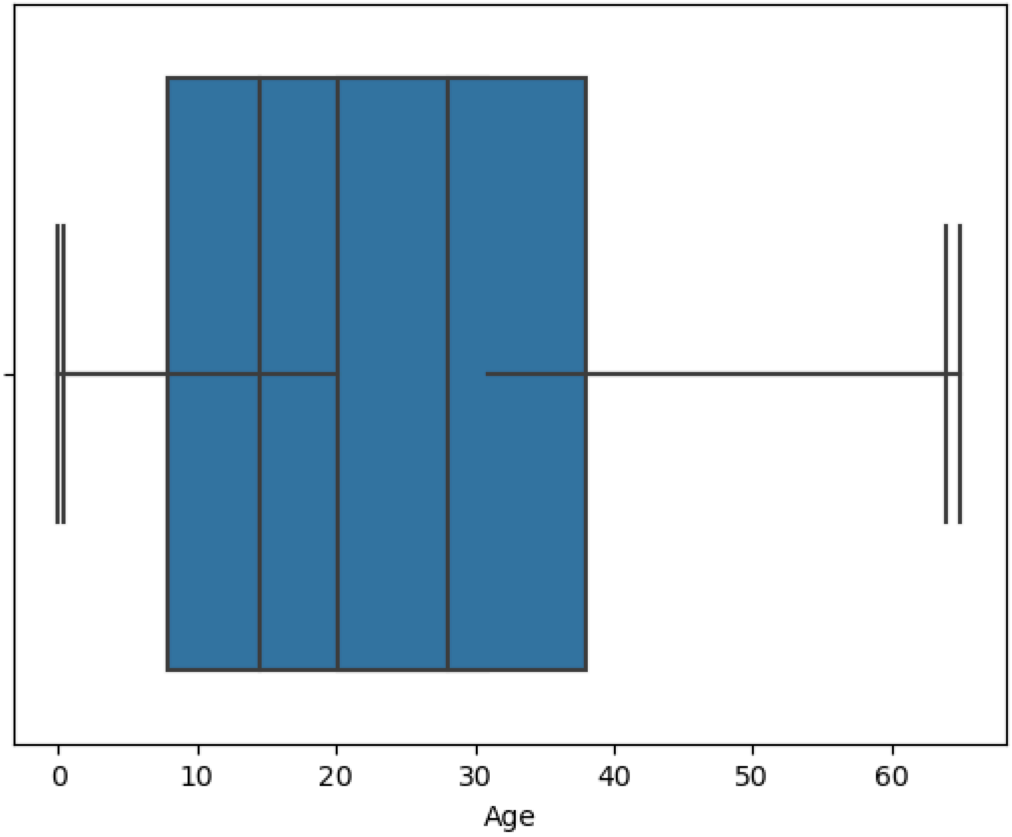
Visualization of Outliers in Dataset

In [33]:

```
sns.boxplot(x = train_df.Fare, showfliers = False)
sns.boxplot(x=train_df.Age, showfliers=False)
```

Out[33]:

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



```
In [ ]:
```

```
In [ ]:
```

Loading [MathJax]/extensions/Safe.js