```
import numpy as np
data=[1,2,2,2,3,1,1,15,2,2,2,3,1,1,2]
mean =np.mean(data)
std=np.std(data)
print('mean of the dataset is',mean)
print('std.deviation is',std)
threshold=3
outlier=[]
for i in data:
    z =(i-mean)/std
    if z > threshold:
       outlier.append(i)
print('outlier in dataset is',outlier)
```

```
mean of the dataset is 2.6666666666666665
std.deviation is 3.3598941782277745
outlier in dataset is [15]
```

Interquratile range to detect outliers in data Q1 represents the 25 th percentiile of data. Q2 represents the 50th percentile of the data. Q3 repersents the 75th percentile of the data. If the data set has 2n/2n+1 data points ,then Q1=median of the dataset Q2=median of the n smallest data points Q3=median of n highest data points IQR=Q3-Q1

Double-click (or enter) to edit

```
import numpy as np
import pandas as pd
import seaborn as sns
data=[6,2,3,4,5,1,50]
sort_data=np.sort(data)
sort_data
```

```
array([ 1,  2,  3,  4,  5,  6, 50])
```

```
Q1=np.percentile(data,25,interpolation='midpoint')
Q2=np.percentile(data,50,interpolation='midpoint')
Q3=np.percentile(data,75,interpolation='midpoint')

print('Q1 25 percentile of the given data is,',Q1)
print('Q1 50 percentile of the given data is,',Q2)
print('Q1 75 percentile of the given data is,',Q3)

IQR=Q3-Q1
print('Interquartile range is',IQR)
```

```
Q1 25 percentile of the given data is, 2.5
Q1 50 percentile of the given data is, 4.0
Q1 75 percentile of the given data is, 5.5
Interquartile range is 3.0
```

```
low_lim=Q1-1.5*IQR
up_lim=Q3+1.5*IQR
print('low_limit',low_lim)
print('up_limit is',up_lim)
```

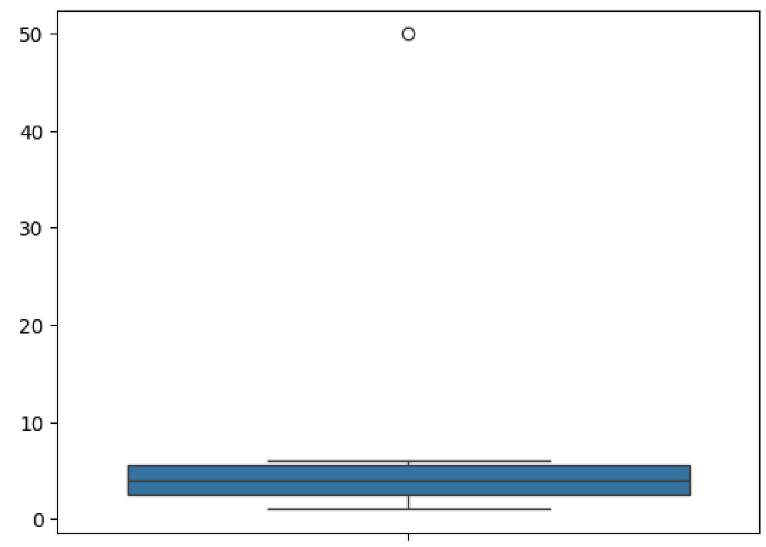
```
low_limit -2.0
up_limit is 10.0
```

```
outlier=[]
for x in data:
  if((x> up_lim)or(x<low_lim)):
    outlier.append(x)
print('outlier in the dataset is',outlier)
```

```
outlier in the dataset is [50]
```

```
sns.boxplot(data)
```

```
<Axes: >
```



```
def load_data():
    df_all=pd.read_csv('/content/train.csv')
    return df_all.loc[:300,['Survived','Pclass','Sex','Cabin','Embarked']]
df=load_data()
```

```
df.Cabin.duplicated()
```

```
0      False
1      False
2      False
3       True
4       True
       ...
296     True
297     True
298     True
299    False
300    False
Name: Cabin, Length: 301, dtype: bool
```

```
df.duplicated()
```

```
0      False
1      False
2      False
3      False
4      False
       ...
296     True
297     True
298     True
299    False
300    False
Length: 301, dtype: bool
```

```
df.duplicated(subset=['Survived','Pclass','Sex'])
```

```
0      False
1      False
2      False
3      False
4      False
       ...
296     True
297     True
298     True
299     True
300     True
Length: 301, dtype: bool
```

```
df.Cabin.duplicated().sum()
```

```
230
```

```
df.duplicated().sum()
```

```
199
```

```
df.loc[df.duplicated(keep='first'),:]
```

	Survived	Pclass	Sex	Cabin	Embarked
5	1	2	female	NaN	S
6	0	3	male	NaN	S
7	0	2	male	NaN	S
11	0	3	male	NaN	S
12	0	3	male	NaN	S
...	...	...	...	...	...
294	0	2	female	NaN	S
295	1	3	female	NaN	C
296	1	3	female	NaN	S
297	0	3	male	NaN	S
298	1	3	male	NaN	S

199 rows × 5 columns

```
df.loc[df.duplicated(keep='last'),:]
```

	Survived	Pclass	Sex	Cabin	Embarked
2	1	2	female	NaN	S
3	0	2	male	NaN	S
4	0	3	male	NaN	S
5	1	2	female	NaN	S
6	0	3	male	NaN	S
...	...	...	...	...	...
285	1	3	male	NaN	S
287	0	3	male	NaN	S
288	0	3	male	NaN	S
289	0	3	male	NaN	S
291	0	3	male	NaN	S

199 rows × 5 columns

```
df.loc[df.duplicated(keep=False),:]
```

```
         Survived  Pclass     Sex  Cabin  Embarked
```

```
df.drop_duplicates()
```

	Survived	Pclass	Sex	Cabin	Embarked
0	0	1	male	C30	S
1	1	1	female	D33	C
2	1	2	female	NaN	S
3	0	2	male	NaN	S
4	0	3	male	NaN	S
...	...	...	...	...	...
271	1	1	male	C93	S
278	0	1	male	C111	C
286	1	1	male	C148	C
299	1	1	female	D21	S
300	1	2	male	F2	S

102 rows × 5 columns

Double-click (or enter) to edit

Double-click (or enter) to edit

Double-click (or enter) to edit

Double-click (or enter) to edit

```
Start coding or generate with AI.
```

```
Start coding or generate with AI.
```

```
Start coding or generate with AI.
```

```
Start coding or generate with AI.
```

```
Start coding or generate with AI.
```

```
Start coding or generate with AI.
```