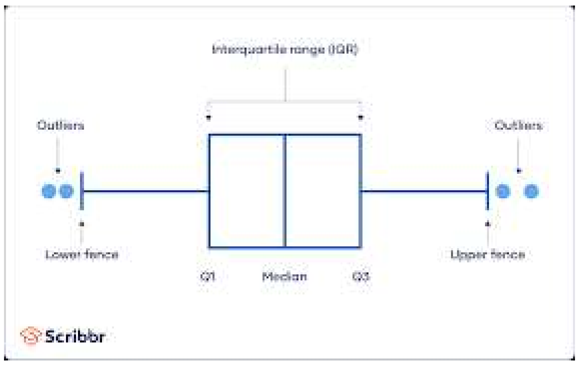
Outlier deduction



Outlier deduction involves recognizing and handling data points significantly different from the main distribution to enhance data analysis accuracy and maintain statistical integrity.

In [1]:
```
## Import necessary libraries
import pandas as pd
import numpy as np
import seaborn as sns

import warnings
warnings.filterwarnings("ignore")
```

In [2]:
```
#dataset
dataset = [10,30,50,60,70,80,320,120,40,50,60,130,140,150,300,360]
```

IQR

The Interquartile Range (IQR) technique involves calculating the range between the first quartile (Q1) and the third quartile (Q3) of a dataset. Data points outside the range (Q1 - 1.5 * IQR) to (Q3 + 1.5 * IQR) are considered outliers and can be removed or investigated further.

1. Sort the data
2. Calculate Q1(25%) and Q3(75%)
3. IQR(Q3-Q1)
4. Find the lower fence(q1-1.5*(iqr)
5. Find the upper fence(q1+1.5*(iqr)

In [3]:
```
# 1. Sort the data
dataset = sorted(dataset)
```

In [4]:
```
dataset
```

Out[4]:
```
[10, 30, 40, 50, 50, 60, 60, 70, 80, 120, 130, 140, 150, 300, 320, 360]
```

In [5]:
```
## 2. Calculate Q1(25%) and Q3(75%)
q1,q3 = np.nanpercentile(dataset,[25,75])
print(q1,q3)
```

```
50.0 142.5
```

In [6]:
```
## 3. IQR(Q3-Q1)
IQR = q3-q1
print(IQR)
```

```
92.5
```

In [7]:

```
## 4. Find the lower fence(q1-1.5*(iqr)
## 5. Find the upper fence(q1+1.5*(iqr)

lower_fence = q1-(1.5*IQR)
upper_fence = q3+(1.5*IQR)
print(lower_fence)
print(upper_fence)
```

```
-88.75
281.25
```

In [8]:

```
# find the outliers
df = pd.DataFrame(dataset,columns=["D"])
outlier = df.loc[df["D"]>=281]
print(outlier.values)
```

```
[[300]
 [320]
 [360]]
```

box plot:

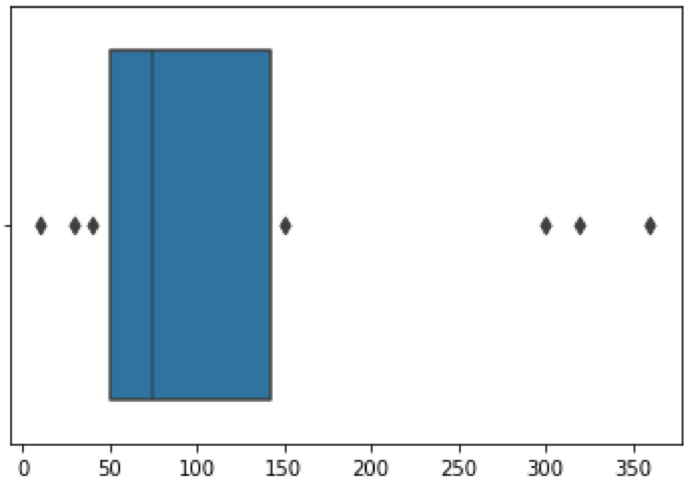
A box plot (box-and-whisker plot) visually represents the distribution of a dataset and can help identify outliers, as well as calculate Q1 and Q3. The "box" covers the interquartile range (IQR), while "whiskers" extend to 1.5 times IQR. Points beyond the whiskers are potential outliers.

In [9]:

```
sns.boxenplot(dataset)
```

Out[9]:

```
<AxesSubplot:>
```



Observation:

1. In the given box plot, the lower quartile (Q1(25%)) is 50, the median is 70, and the upper quartile (Q3(75%)) is 150. Any data point greater than 300 is considered an outlier.