STATISTICS

(SESSION-8)

OUTLIERS :

- Outliers are the observations having very Huge value or very Small value.
- Mean will affect by Outliers.
- For Example :

Assume that Indian Income : 1L , 2L , 3L , 4L

- Mean = $\frac{1+2+3+4}{4}$ = 2.5

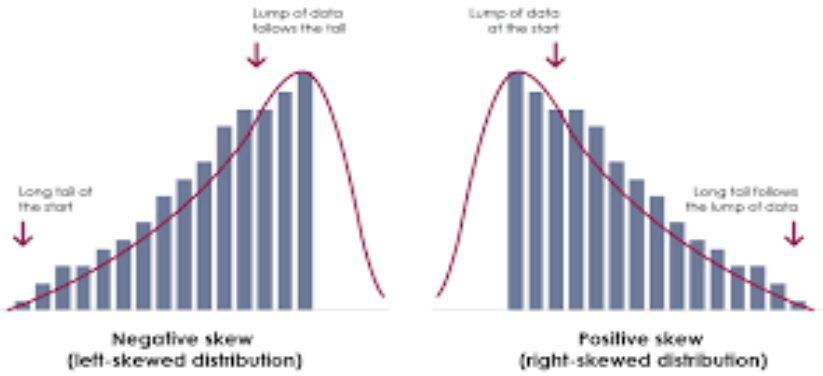
Suppose we added an unusual Observation 100Crs.

- Mean = $\frac{1+2+3+4+100\text{cr}}{5}$ = 20cr

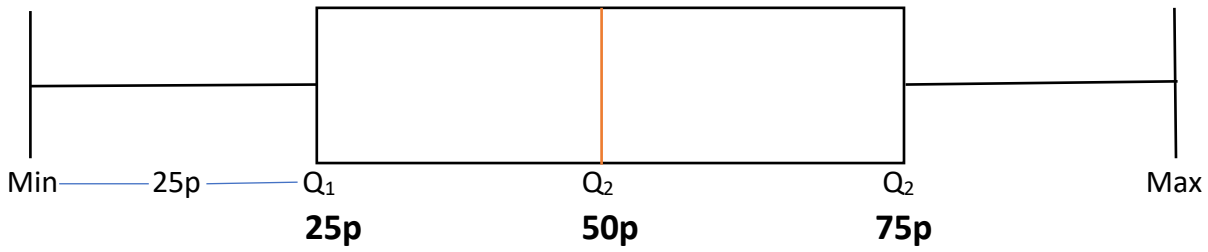
Here , Outlier is 100crs.

Because of Outliers data will skew

- Positive outliers means huge value , data will skew to the positive(Right) side.
- Negative outliers means very small value , data will skew to the Nehgative (Left) side.



BOX PLOT :



In the above diagram

Q_1 = 25p Value

Q_2 = 50p Value

Q_3 = 75p Value

Outliers will exist after Q_3 point and below Q_1 point

Upper bound = Q_3 + ?

Lower bound = Q_2 - ?

IQR : (Inter Quartile Range**)**

In order to find the outliers we need to travel from Q_3 to above Q_1 to below

The travel distance based on Middle 50% of data

That middle 50% of data is called as IQR : Inter Quartile Range

IQR = $Q_3 – Q_1$ (Middle data)

Upper bound = Q_3 + IQR

Lower bound = Q_1 – IQR

The Upper bound and Lower bound cutoff varies based on , How many times of IQR we are using

Upper bound = Q_3 + k * IQR

Lower bound = Q_1 – k * IQR

Generally we will use k = 1.5 and k = 3

When k = 1.5 : (Mild Outliers)

Upper bound = Q_3 + 1.5 * IQR

Lower bound = Q_1 – 1.5 * IQR

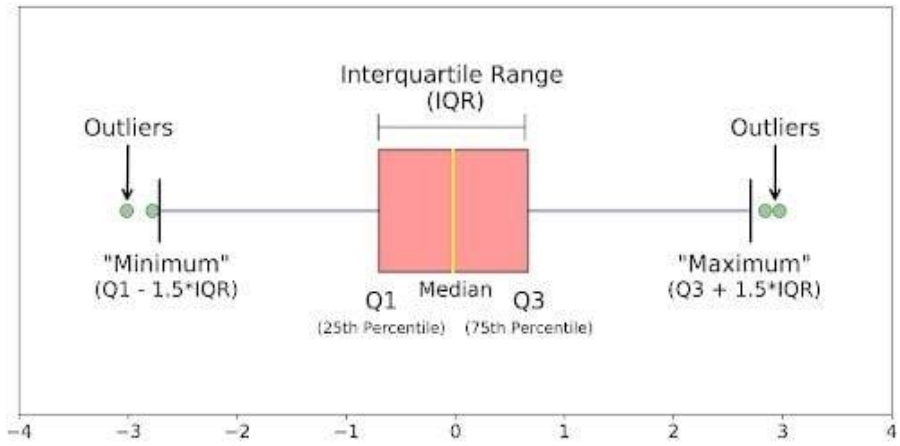
When k = 3 : (Huge Outliers)

Upper bound = Q_3 + 3 * IQR

Lower bound = Q_1 – 3 * IQR

In Python we use by default k = 1.5 only

Middle line represents median = 50p of Data



Example : Let Q_1 =10K Q_2 =1Lakh Q_3 =5Lakhs

Outliers : Q_3 + 1.5 * IQR

= Q_3 + 1.5 * (Q_3 – Q_1)

= 5 + 1.5 * (5L – 10K)

= 5 + 1.5 * (4.9L) = 12.35

If a person is earning 12.35 per month or more he is considered as a outlier

Outliers : Q_3 - 1.5 * IQR

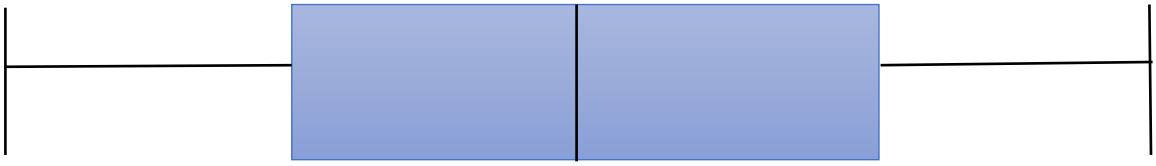
= Q_3 - 1.5 * (Q_3 – Q_1)

= 5 - 1.5 * (5L – 10K)

= 5 - 1.5 * (4.9L) = - 2.35

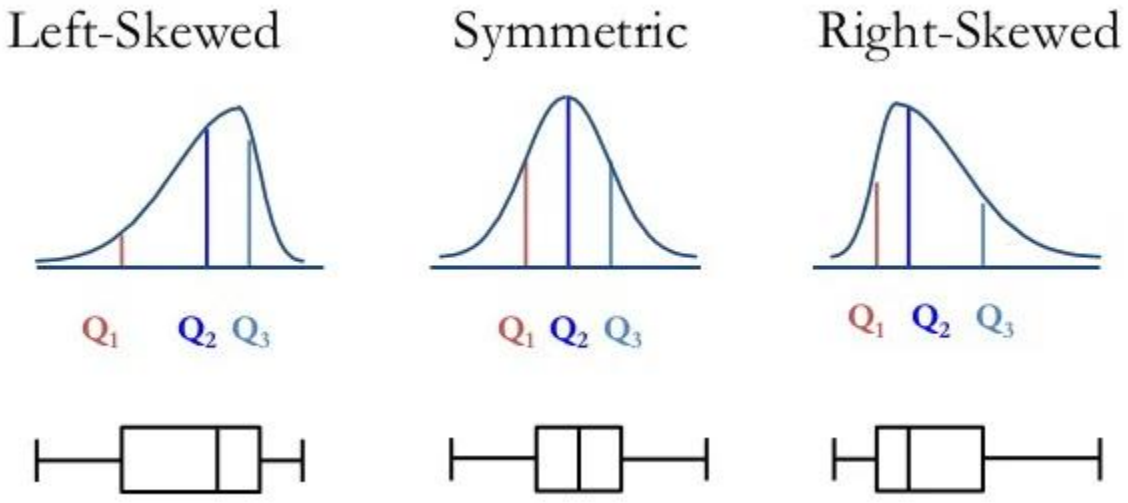
If a person is earning - 2.35 per month or more he is considered as a outli

Normal Distribution



Positive Skew

Negative Skew



DEALING OUTLIER :

Drop the outlier :

- If any outliers are present , we can drop the outlier If the outlier has 2% of data.
- Suppose a data has 100 observations in that 2 observations consider as outlier.
- So Outlier perecentage is 2%.
- If you drop 2% of data then we have 98% of data available.
- It is enough to train the ML Model.
    - Drop the Outlier is generally not recommended.
    - If we drop the outlier means we are droping the information.

Fill with Median Value :

- We know that , Outlier doesnot affect the Median.
- So , that it is good practice we can fill Outliers with median value.

Fill with Q_3 and Q_1 (Cap the Values) :

Caping method .

- More than Q_3 Outliers replace with Q_3 value.
    - Positive side Outliers can replace with Upper bound value
- Less than Q_1 Outliers replace with Q_1 value.
    - Negative side Outliers can replace with Lower bound value