**Statistics**

**Percentile –**

* **Percentile = (No. of values below x / n) \* 100** ……………………….. Formula

Data = [2,2,3,4,5,5,6,7,8,7,9,8,9]

(6/12) \*100 = 50%

Length of the data is 14. Hence 6 is 50th percentile of data means 50% of data is less then element 6.

* Q1. What is the 25% of data?

Value = [(Percentile / 100)\*(n+1)] = (1/4)\*(14) = 3.5 … Hence 25% of data is between 3 and 4.

**Removing Outlier From Dataset–**

**Steps:**

1. Sort the given data.
2. Find out Q1 and Q3 of the data
3. Calculate the IQR, Upper fence and lower fence.
4. Check if any element of dataset is falling outside of the range. If yes, then remove it as it is outlier.

Data which fall outside of the lower fence and upper fence is going to be outlier.

**Lower Fence = Q1 – 1.5(IQR)**

**Upper Fence = Q3 + 1.5(IQR)**

Where: Q1 = 25% of the data

Q3 = 75% of the data

IQR = inter quartile range (Q3-Q1)

Let’s take one example:

Age = [1,2,2,2,3,3,4,5,5,6,6,6,6,7,8,8,9]

Q1 = 3, Q3 = 6

Hence IQR = Q3-Q1 = 6-3 = 3

Now calculate for upper and lower fence:

Lower fence = 3 – 1.5(3) = -1.5

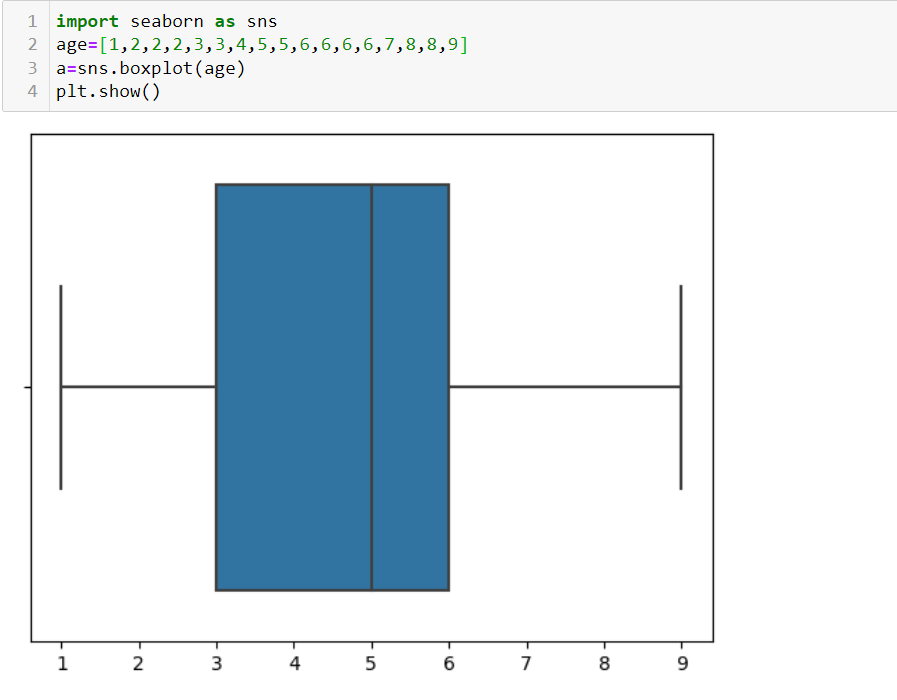
Upper fence = 7 + 1.5(3) = 11.5

Data fall outside of this range (-1.5 to 11.5) is nothing but outlier. In age data all the value are fall inside the range hence no data element is outlier here.

**5 Number Summary** –

1. Minimum element – 1
2. Q1 – 3
3. Median – 5
4. Q3 – 6
5. Maximum element – 9

Note- We can plot this on box plot as below



Hence it is proved that it is not Gaussian distribution as median is not in the center.

**Co-variance –**

Co-variance defines relation between two independent features or X and Y.

**Cov(x,y) = (x-mean(x)) \* (y-mean(y)) / N-1**

Let’s take one example to calculate Co-variance –

X = 2,3,4,5 y = 5,6,7,8

Cov(x,y) = 1/3[(0.5 \* 0.5) + (1.5 \* 1.5)] = 1.67

Whenever you have +ve value of co-variance then x and y are +ve correlated and if value is –ve then it is inversely correlated.

**Disadvantage of Co-variance:**

By using co-variance user can quantify the relation but not able to find how much directly correlated.

For instance, value of co-variance is not limited it can be anything such as +2.5, 1000, -250. There is no any specific range of co-variance, hence user cannot find exact correlation between two features even if he knows the co-variance value. To overcome this we use **Pearson correlation.**

**Pearson correlation Co-efficient:**

In Pearson Correlation co-efficient technique entire co-variance values is limited between **-1 to +1.**

**Pearson Correlation co-efficient = Cov(x,y) /S.D.(x) \* S.D.(y)**

This correlation technique useful to feature selection process.

**Spearman Correlation—**

Spearman Correlation is useful for non-linear data.

Spearman Correlation -- Cov [( Rg(x) , Rg(y)) / S.D.(Rg(x)) \* S.D.(Rg(y)]

Here in dataset user assign rank to the feature in descending order and then user can calculate co-variance.

X = 1,2,5,3,6,7 Y= 6,8,3,6,9,1

Rg(x) = 6,5,3,4,2,1 Rg(y) = 3,2,5,4,1,6