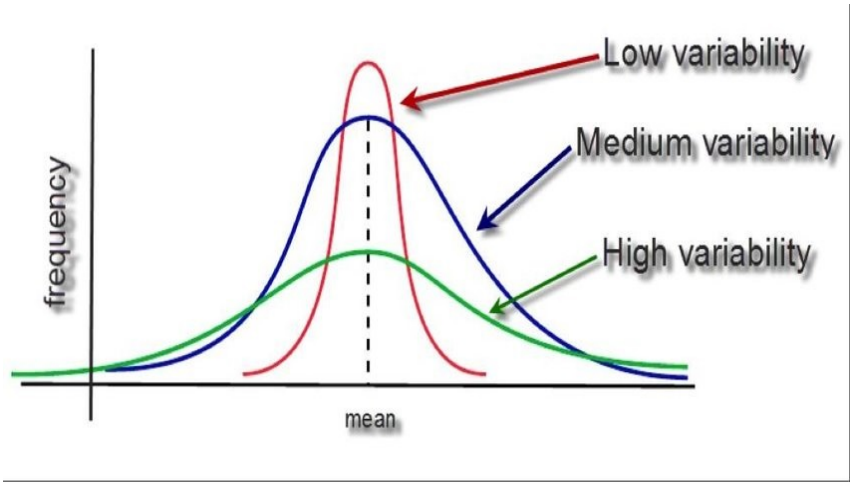
Measures of Dispersion (Variability)

Data variability also known as spread or dispersion, refers to how spread out a set of data is. Variability gives users a way to describe how much data sets vary and allows users to use statistics to compare their data to other sets of data.

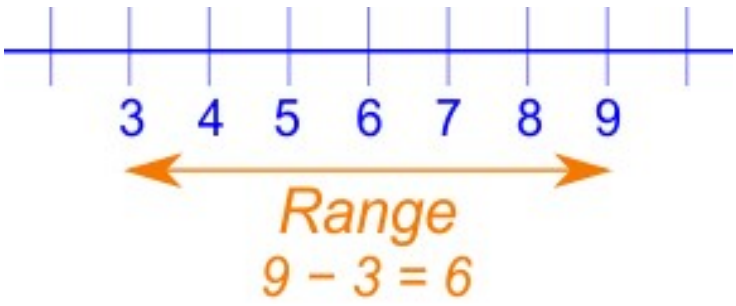
The four main ways to describe variability in a data set are:

- Range
- Variance
- Standard deviation.
- Interquartile range



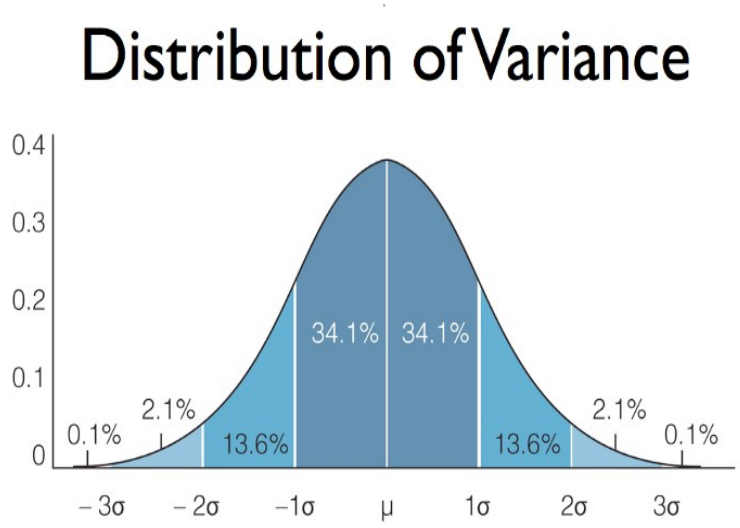
1. Range:

- In statistics, the range is the spread of your data from the lowest to the highest value in the distribution.
- It is a commonly used measure of **variability**.
- The range is calculated by subtracting the lowest value from the highest value.
- Large range means high variability, a small range means low variability in a distribution.
- For example, if the given data set is {2, 5, 8, 10, 3}, then the range will be 10 – 2 = 8.



2. Variance:

- Variance is a measure of how data points differ from the **mean**.
- A small variance indicates that the data points tend to be very close to the mean.
- A high variance indicates that the data points are very spread out from the mean.
- Therefore, it is called a **measure of spread of data from mean**.
- Variance is the average of the squared distances from each point to the mean.



- Variance is the sum of squares of differences between all numbers and mean (μ). The mathematical formula for variance is as follows

$$\text{Variance} = \sigma^2 = \frac{\Sigma(x_i - \mu)^2}{n}$$

where: *xi = Each Sample*
μ = Mean
n= total number of samples

Let's say we have values: 5, 7, 9, and 3.

1. Calculate the mean

$$\overline{x} = \frac{\sum x}{n} \Rightarrow \overline{x} = \frac{x1+x2+x3.....xn}{n}$$

$$\textbf{Mean} = \frac{5+7+9+3}{4}$$

$$= 6$$

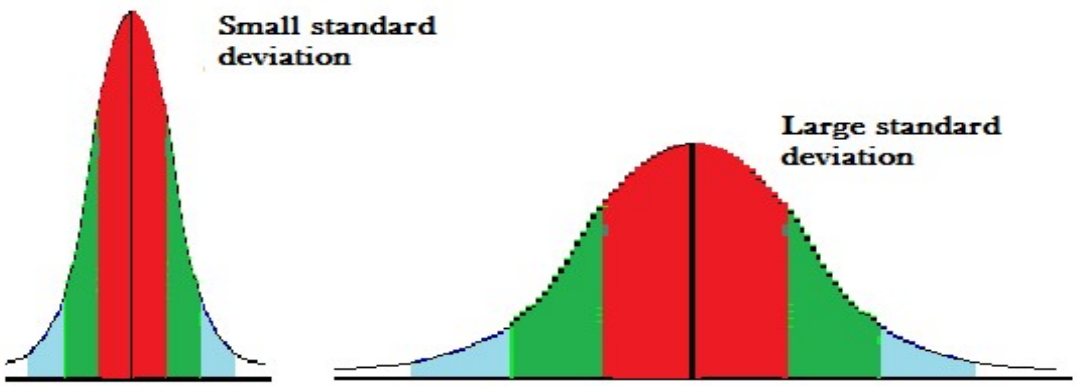
2. Subtract mean from all observation to find the distance of all observation from mean.

$$\textbf{Variance} = \frac{(5-6)^2 + (7-6)^2 + (9-6)^2 + (3-6)^2}{4}$$

$$= \frac{1+1+9+9}{4} \Rightarrow 5$$

3. Standard deviation

- Standard deviation is a squared root of the variance to get original values.
- Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out.



Population SD:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Sample SD:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

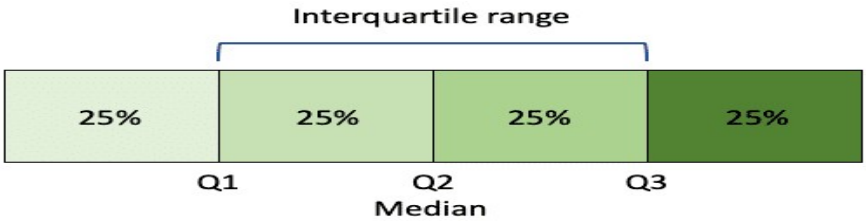
Finding measures of Variability on simple Data

```
# Importing the NumPy module
import numpy as np
# Taking a list of elements
list = [2, 4, 4, 4, 5, 5, 7,25]
X= list
# Calculating variance using var()
x1=np.var(list)
print("varaince of list is :",x1)
range=np.max(X)- np.min(X)
print("range is:",range)
SD =np.std(X)
print("Standard devaition is", SD)
```

```
varaince of list is : 48.0
range is: 23
Standard devaition is 6.928203230275509
```

4. Interquartile Range:

- Interquartile Range, (IQR) is a property that is used to measure variability.
- The IQR divides a dataset into quartiles. These quartiles store data after the data gets sorted in ascending order and split into 4 equal parts. The first, second, and third quartiles are called Q1, Q2, and Q3. These quartiles are the values that separate the 4 equal parts.



The following is the percentile distribution of the data amongst Q1, Q2, and Q3-

- 25^{th} percentile of the data is represented in Q1
- 50^{th} percentile of the data is represented in Q2
- 75^{th} percentile of data is represented in Q3

The measurement of IQR gives an insight into the width of distribution as most of the points of the dataset are contained in this range.

- In a dataset that contains even or odd elements of data points, then-
- Q2 is the median
- Q1 is the median of *x* smallest points of data i.e., lower half of data
- Q3 is the median of *x* highest points of data i.e., Upper half of data

Problem) Find the median, lower quartile, upper quartile, interquartile range and range of the following numbers.
12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25, 65

Solution:

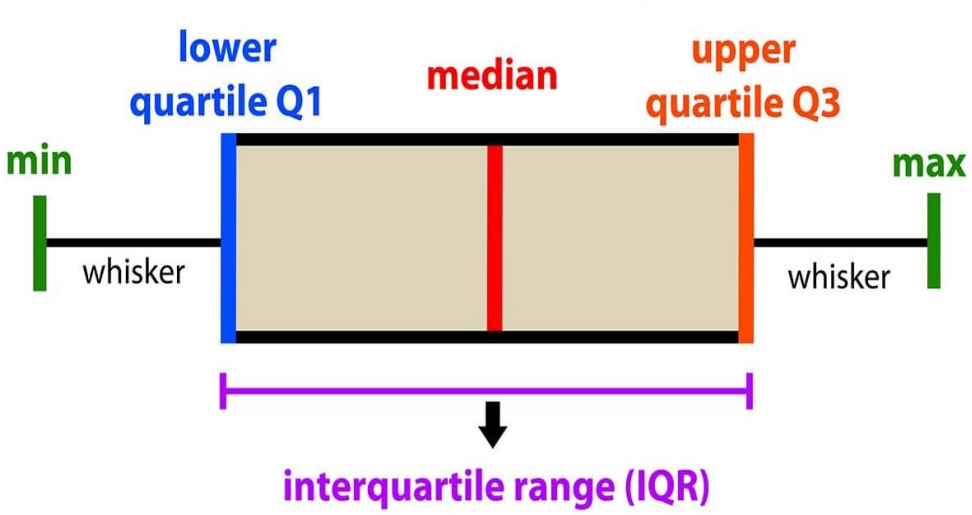
- Find the median, lower quartile, upper quartile, interquartile range and range of the following numbers.
  12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25, 65
- **Solution:**
  First, arrange the data in ascending order:

5, 7, 12, 14, 15, 22, 25, 30, 36, 42, 53, 65
↑ lower quartile or first quartile
↑ median or second quartile
↑ upper quartile or third quartile

- **Lower quartile or first quartile(Q1) =** $\frac{12+14}{2} = 13$
- **Median or second quartile(Q2) =** $\frac{22+25}{2} = 23.5$
- **Upper quartile or third quartile(Q3) =** $\frac{36+42}{2} = 39$
- **Interquartile range** = Upper quartile − lower quartile
  = 39 − 13 = 26
- **Range** = largest value − smallest value
  = 65 − 5 = 60

BOX PLOT:

- The box and whisker plot, sometimes simply called the box plot, is a type of graph that help visualize the five-number summary.
- Box plots are a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum").
- It doesn't show the distribution in as much detail as histogram does, but it's especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set.
- A box plot is ideal for comparing distributions because the center, spread and overall range are immediately apparent.

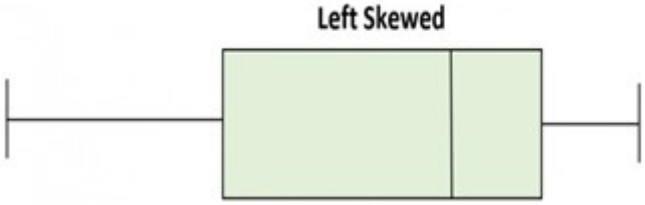


In a box and whisker plot:

- The left and right sides of the box are the lower and upper quartiles. The box covers the interquartile interval, where 50% of the data is found.
- The vertical line that split the box in two is the median. Sometimes, the mean is also indicated by a dot or a cross on the box plot.
- The whiskers are the two lines outside the box, that go from the minimum to the lower quartile (the start of the box) and then from the upper quartile (the end of the box) to the maximum.
- Box plot can be displayed in both vertical & horizontal pattern.
- The box plot shape will show if a data set is normally distributed or skewed.
- When the median is in the middle of the box, and the whiskers are about the same on both sides of the box, then the distribution is symmetric.

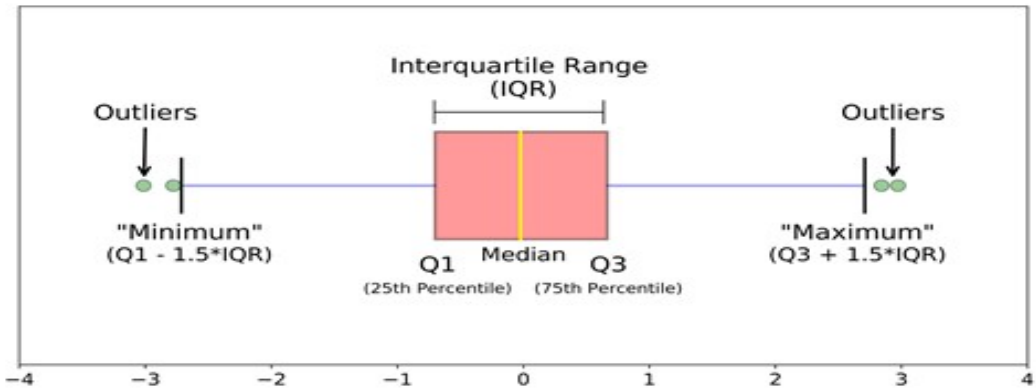
- When the median is closer to the bottom of the box, and if the whisker is shorter on the lower end of the box, and longer to the upper side then the distribution is positively skewed (skewed right).

- When the median is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is negatively skewed (skewed left).



Outliers:

- Outliers are those data points that are significantly different from the rest of the dataset.
- They are often abnormal observations that skew the data distribution, and arise due to incorrect data entry or false observations.
- The outlier formula — also known as the 1.5 IQR rule — is a Thumb rule used for identifying outliers.
- The outlier formula designates outliers based on an upper and lower boundary .i.e.,
- Anything above Q3 + 1.5 x IQR is an outlier
- Anything below Q1 - 1.5 x IQR is an outlier



Finding Outliers Using Boxplot:

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
speed = [60,66,87,88,90,66,103,67,94,78,77,85,66,66,66,65,200,500,800,]
y=speed
plt.figure(figsize=(10, 4))
plt1 = sns.boxplot(speed)
plt.show()
```

Outlier Detection Methods:

1. Standard Deviation Method
2. Z-Score Method
3. IQR Method

Understanding outliers detection with simple Data

```
data1 = [11,12,24,11,14,12,15,11,12,13,12,13,14,102,12,11,13,14,107,15,11,12,14,108,11,14,16,60,140]
len(data1)
```

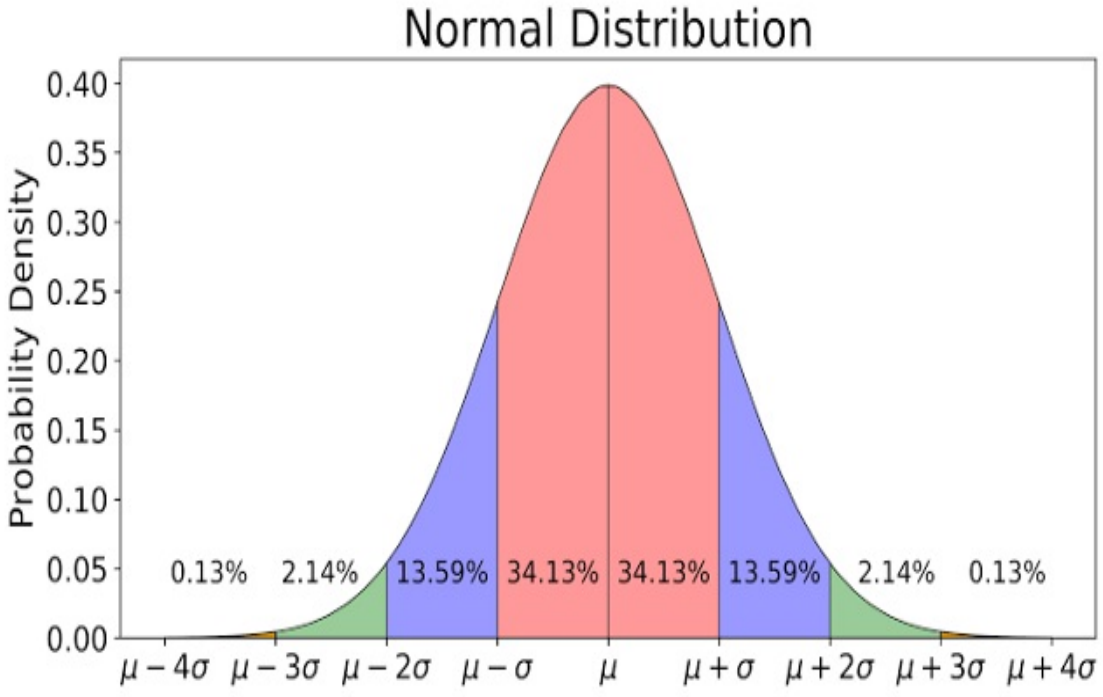
29

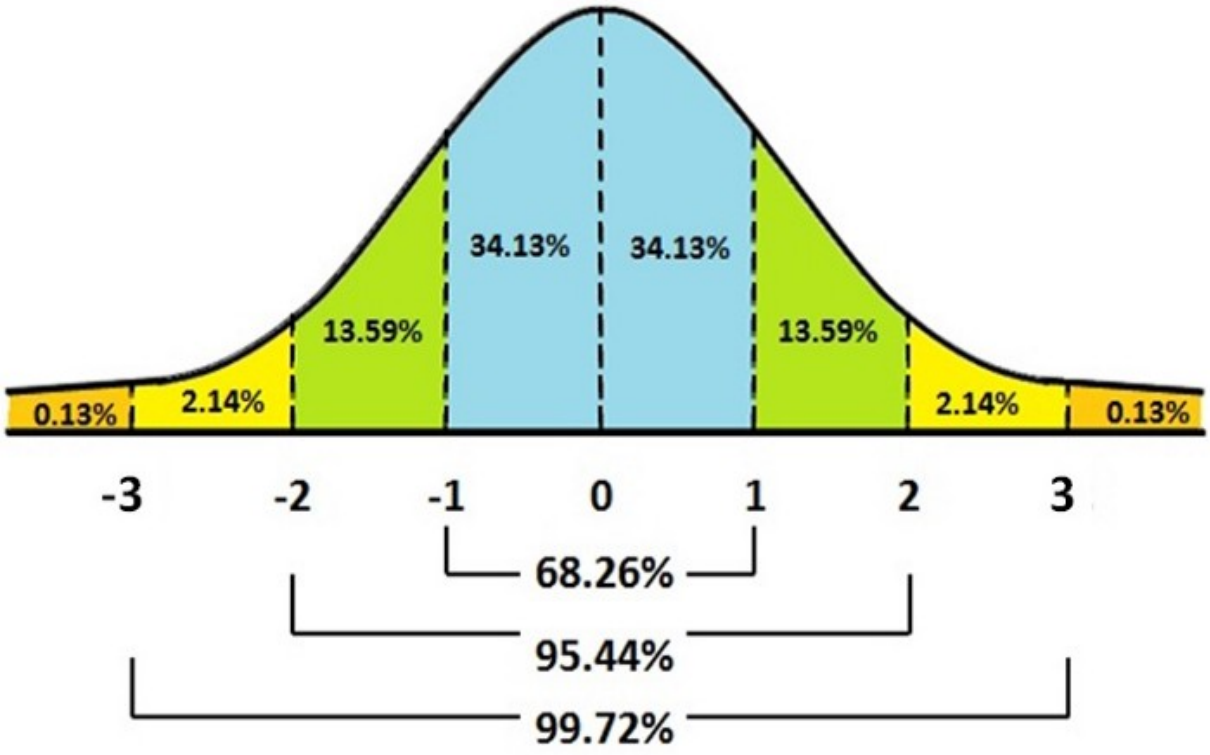
+ Code + Markdown

```
X1=np.mean(data1)
print("The mean value is",X1)
SD =np.std(data1)
print("Standard devaition is", SD)
```

```
The mean value is 28.75862068965517
Standard devaition is 35.76325910128688
```

1. Standard Deviation Method





Outlier detection using Normal Standard Deviation method

```
import numpy as np
lower_limit = np.mean(data) - 1* np.std(data1)
upper_limit = np.mean(data) + 1* np.std(data1)
print(lower_limit, upper_limit)
```

```
-12.281777619805396 59.24474058276836
```

```
outliers = []
def detect_outliers(data1):
  for i in data1:
    if (i<lower_limit or i>upper_limit):
      outliers.append(i)
  return outliers

outliers = detect_outliers(data1)
outliers
```

```
import numpy as np
lower_limit = np.mean(data1) - 2* np.std(data1)
upper_limit = np.mean(data1) + 2* np.std(data1)
print(lower_limit, upper_limit)
```

```
-42.76789751291858   100.28513889222893
```

```
outliers = []
def detect_outliers(data1):
  for i in data1:
    if (i<lower_limit or i>upper_limit):
      outliers.append(i)
  return outliers

outliers = detect_outliers(data1)
outliers
```

```
[102, 107, 108, 140]
```

```
import numpy as np
lower_limit = np.mean(data1) - 3* np.std(data1)
upper_limit = np.mean(data1) + 3* np.std(data1)
print(lower_limit, upper_limit)
```

```
-78.53115661420546      136.0483979935158
```

```
outliers = []
def detect_outliers(data):
  for i in data:
    if (i<lower_limit or i>upper_limit):
      outliers.append(i)
  return outliers

outliers = detect_outliers(data1)
outliers
```

```
[140]
```

2. Z Score Method:

- Z score is also called standard score.
- Z score tells how many standard deviations away a data point is from the mean. i.e.,

Z score detects outliers based threshold value

$$Z = \frac{x - \mu}{\sigma}$$

```
outliers = []
def detect_outliers(data1):
  threshold = 2.0
  mean = np.mean(data1)
  std = np.std(data1)

  for x in data1:
    z_score = np.abs((x - mean)/std)
    if np.ceil(z_score) > threshold:
      outliers.append(x)
  return outliers
```

+ Code + Markdown

```
outliers_pts = detect_outliers(data1)
outliers_pts
```

```
[102, 107, 108, 140]
```

3. IQR (Inter Quartile Range)

IQR (Inter Quartile Range) Inter Quartile Range approach to finding the outliers is the most commonly used and most trusted approach used in the research field.
IQR = Quartile3 – Quartile1

To define the outlier base value is defined above and below datasets normal range namely Upper and Lower bounds, define the upper and the lower bound (1.5*IQR value is considered) :

upper = Q3 +1.5*IQR lower = Q1 – 1.5*IQR

In the above formula as according to statistics, the 0.5 scale-up of IQR (new_IQR = IQR + 0.5*IQR) is taken, to consider all the data between 2.7 standard deviations in the Gaussian Distribution.

Example1:

Consider following dataset
2, 19, 22, 27, 29, 30, 32, 35, 52, 59

Median(Q2)=(29+30)/2 = 29.5

Median of Lower half of data(Q1) 2, 9, 2, 27, 29

Q1=22

Median of Upper half of data(Q3)
30, 32, 35, 52, 59

Q3=35

Inter-quartile range, IQR.

Inter-quartile range is the difference between Q3 and Q1.

IQR = Q3 - Q1 = 35 - 22 = 13

Find the upper boundary

Upper boundary = Q3 + 1.5 IQR = 35 + (1.5)(13) = 54.5

Find the lower boundary

Lower boundary = Q1 - 1.5 IQR = 22 - (1.5)(13) = 2.5
Identify the outliers

The outliers are any data points that lie above the upper boundary or below the lower boundary. In this case, the outliers are 2 and 59.

Example2: The runs scored by a cricket team in a league of 12 matches – 100,120,110,150,110,140,130,170,120,220,140,110. Draw box plot

STEP 1) To draw a box plot for the given data first we need to arrange the data in ascending order Ascending Order -

100,110,110,110,120,120,130,140,140,150,170,220

STEP 2) Find the minimum, first quartile, median, third quartile and IQR

A) Median(Q2) 100,110,110,110,120,120,130,140,140,150,170,220

Median (Q2) = (120+130)/2 = 125 ; Since there were even values

B) Find the First Quartile we take the first six values and find their median. 100,110,110,110,120,120

Q1 = (110+110)/2 = 110

C) For the Third Quartile, we take the next six and find their median. 130,140,140,150,170,220 Q3 = (140+150)/2 = 145

Note: If the total number of values is odd then we exclude the Median while calculating Q1 and Q3. Here since there were two central values we included them.

D) Calculate the Inter Quartile Range. IQR = Q3-Q1 = 145-110 = 35

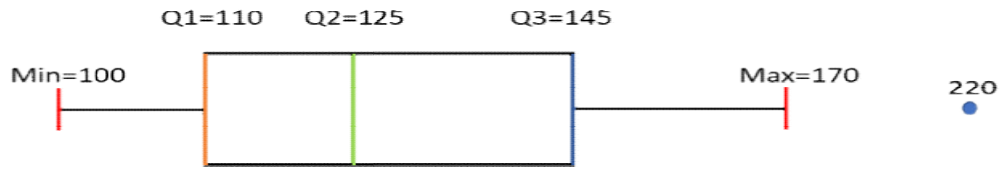
STEP 3: Calculate the Upper and Lower Limits to find the minimum and maximum values and also the outliers if any.

Lower Limit = Q1-1.5*IQR = 110-1.5*35 = 57.5 Upper Limit = Q3+1.5*IQR = 145+1.5*35 = 197.5

So the minimum and maximum between the range [57.5,197.5] for our given data are –
Minimum = 100
Maximum = 170

The outliers which are outside this range are –
Outliers = 220

we have all the information, so we can draw the box plot which is as below-



Example3:

Find Q1, Q2, and Q3 for the following data set. Identify any outliers, and draw a box-and- whisker plot.

{5,40,42,46,48,49,50,50,52,53,55,56,58,75,102}

STEP 1: Data set has 15 values, arranged in increasing order. 5,40,42,46,48,49,50, 50, 52, 53, 55, 56, 58, 75, 102

STEP 2: Find the minimum, first quartile, median, third quartile and IQR

A) Median(Q2)

5,40,42,46,48,49,50, 50, 52, 53,55, 56, 58, 75, 102

Median (Q2) = 50;

B) Find the First Quartile we take the first six values and find their median. 5,40,42,46,48,49,50

Q1 = 46

C) For the Third Quartile, we take the next six and find their median. 52,53,55,56,58,75,102

Q3 = 56

D) Calculate the Inter Quartile Range. IQR = Q3-Q1 = 56-46 = 10

STEP 3: Calculate the Upper and Lower Limits to find the minimum and maximum values and also the outliers if any.

Lower Limit = Q1-1.5*IQR = 46−15=31

Upper Limit = Q3+1.5*IQR = 56+15=71

So the minimum and maximum between the range [31,71] for our given data are –

Minimum = 31

Maximum = 71

The outliers which are outside this range are –

Outliers = 5, 75, 102

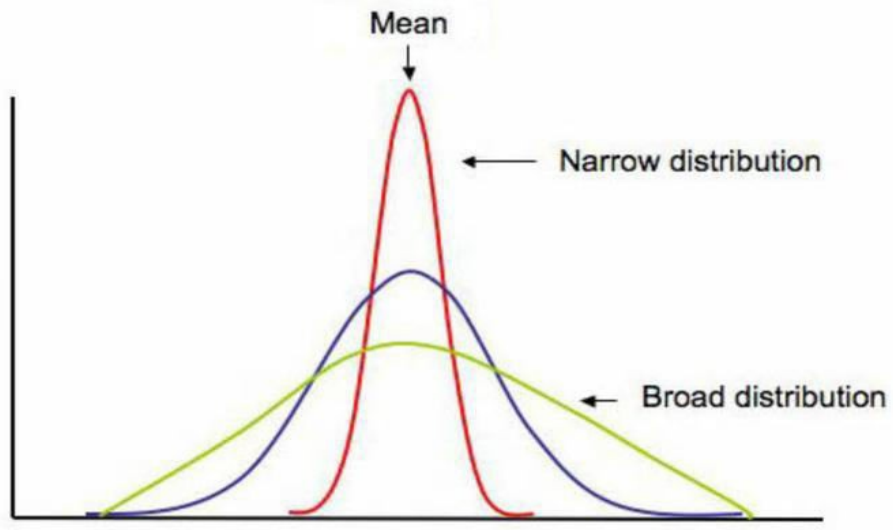
40 and 58 are shown as the ends of the whiskers, with the outliers plotted separately.

Mean absolute deviation

The mean absolute deviation of a dataset is the average distance between each data point and the mean. It gives us an idea about the variability in a dataset.

The measure of spread represents the amount of dispersion in a data-set. i.e how spread-out are the values of data-set around the central value (example- mean/mode/median). It tells how far away the data points tend to fall from the central value.

- The lower value of the measure of spread reflects that the data points are close to the central value. In this case, the values in a data-set are more consistent.
- Further, the distance of the data points from the central-value, the greater is the spread. whereas here, the values are not much consistent.



Using the above diagram, we can infer that the narrow distribution represents a lower spread, and the broad distribution represents a higher spread.

Formula for mean absolute deviation

$$\text{MAD} = \frac{\sum |x_i - \bar{x}|}{n}$$

- Here || gives the absolute value that means all negative deviation (distance) made positive.

Example 1) Find the mean absolute deviation of the following data set: 10,15,15,17,18,21

Solution: Mean (X_i) = (10+15+15+17+18+21)/6 =16

Data points	Distance from mean($\|X_i$-X|)
10	|10-16|=6
15	|15-16|=1
15	|15-16|=1
17	|17-16|=1
18	|18-16|=2
21	|21-16|=5

average of all the absolute values=(6+1+1+1+2+5)/6=16/6=2.67

Example 2) Find the mean absolute deviation of the following data set: 26, 46, 56, 45, 19, 22, 24.

Solution:

Given set of data is:

26, 46, 56, 45, 19, 22, 24

Mean = (26 + 46 + 56 + 45 + 19 + 22 + 24)/7 = 238/7 = 34

Data points	**Distance from mean(|Xi-X|**
26	|26-34|=8
46	|46-34|=12
56	|56-34|=22
45	|45-34|=11
19	|19-34|=15
22	|22-34|=12
24	|24-34|=10

Average of all the absolute values:

(8+12+22+11+15+12+10)/7=90/7=12.857

Therefore, the mean absolute deviation of the given data set is 12.857.

Covariance:

Covariance signifies the direction of the linear relationship between the two variables. By direction we mean if the variables are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).

The value of covariance between 2 variables is achieved by taking the summation of the product of the differences from the means of the variables as follows:

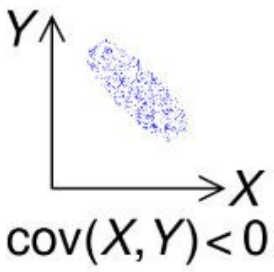
Where,

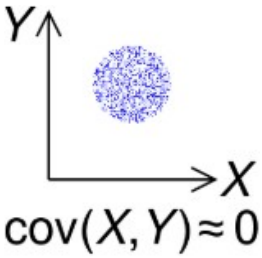
$$\mathrm{COV}_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

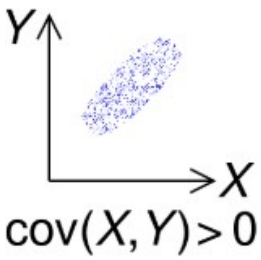
x = the independent variable
y = the dependent variable
n = number of data points in the sample

$\bar{x}$ = the mean of the independent variable x
$\bar{y}$ = the mean of the dependent variable y

The upper and lower limits for the covariance depend on the variances of the variables involved. These variances, in turn, can vary with the scaling of the variables. Even a change in the units of measurement can change the covariance. Thus, covariance is only useful to find the direction of the relationship between two variables and not the magnitude. Below are the plots which help us understand how the covariance between two variables would look in different directions.







Example:

X	Y
10	40
12	48
14	56
8	32

Step 1: Calculate Mean of X and Y

Mean of X : (10+12+14+8) / 4 = 11

Mean of Y : (40+48+56+32) / 4 = 44

Step 2: Substitute the values in the formula

X	Y	$x_i - \bar{x}$	$y_i - \bar{y}$
10	40	10 – 11 = -1	40 – 44 = – 4
12	48	12 – 11 = 1	48 – 44 = 4
14	56	14 – 11 = 3	56 – 44 = 12
8	32	8 – 11 = -3	32 – 44 = -12

Substitute the above values in the formula

Cov(x,y) = (-1) (-4) +(1)(4)+(3)(12)+(-3)(-12) / 4-1

Cov(x,y) = 80/3 = 26.67

Hence, Co-variance for the above data is 26.67.

```
import numpy as np
height = [5.1,5.2,5.3,5.4,5.5]
weight=[60.3,59.2,63.6,88.4,68.7]
print("covariance",np.cov(height,weight))
```

```
covariance [[2.50000e-02 1.15000e+00]
 [1.15000e+00 1.43183e+02]]
```

```
print("covariance",np.cov(height,weight)[0,1])
```

```
covariance 1.1500000000000026
```

Correlation:

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two numerically measured continuous variables. It not only shows the kind of relation (in terms of direction) but also how strong the relationship is. Thus, we can say the correlation values have standardized notions, whereas the covariance values are not standardized and cannot be used to compare how strong or weak the relationship is because the magnitude has no direct significance. It can assume values from -1 to +1.

To determine whether the covariance of the two variables is large or small, we need to assess it relative to the standard deviations of the two variables.

To do so we have to normalize the covariance by dividing it with the product of the standard deviations of the two variables, thus providing a correlation between the two variables.

The main result of a correlation is called the correlation coefficient.

$$Correlation = \frac{Cov\,(x, y)}{\sigma x * \sigma y}$$

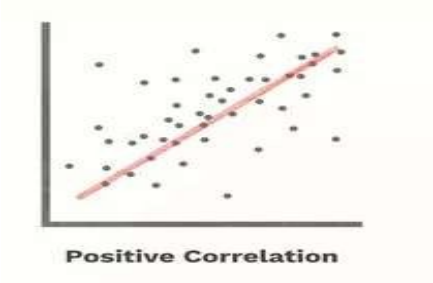
where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

Types of correlation

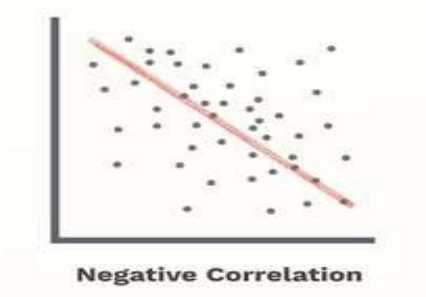
Positive correlation

If with increase in random variable A, random variable B increases too, or vice versa.



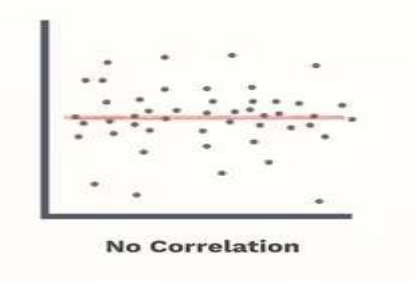
Negative correlation

If increase in random variable A leads to a decrease in B, or vice versa.



No correlation –

When both the variables are completely unrelated and change in one leads to no changein other.



Example

X	Y
10	40
12	48
14	56
8	32

Step 1: Calculate Mean of X and Y
Mean of X : 10+12+14+8 / 4 = 11
Mean of Y = 40+48+56+32/4 = 44

Step 2: Substitute the values in the formula

xi –$\bar{x}$	yi – $\bar{y}$
10 – 11 = -1	40 – 44 = – 4
12 – 11 = 1	48 – 44 = 4
14 – 11 = 3	56 – 44 = 12
8 – 11 = -3	32 – 44 = -12

Step 3: Now substitute the obtained answer in Correlation formula

Substitute the above values in the formula

Cov(x,y) = (-1) (-4) +(1)(4)+(3)(12)+(-3)(-12) / 4-1

Cov(x,y) = 80/3 = 26.67

Hence, Co-variance for the above data is 26.67.

$$Correlation = \frac{Cov\ (x, y)}{\sigma x * \sigma y}$$

- Before substitution we have to find standarddeviation of x and y
- Let's take the data for X as mentioned in the tablethat is 10,12,14,8
- To find standard deviation

$$s = \sqrt{\frac{\Sigma\ (X - \bar{x})^2}{n - 1}}$$

Step 1: Find the mean of x that is $\bar{x}$
10+14+12+8 /4 = 11

Step 2: Find each number deviation: Subtract each scorewith mean to get mean deviation
10 – 11 = -1
12 – 11 = 1
14 – 11 = 3
8 – 11 = -3

Step 3: Square the mean deviation obtained

Mean Deviation Value	Squared Mean Deviation Value
-1	1
1	1
3	9
-3	9

Step 4: Sum the squares
1+1+9+9 = 20

Step5: Find the variance
Divide the sum of squares with n-1 that is 4-1 = 3
20 /3 = 6.6

Step 6: Find the square root
Sqrt of 6.6 = 2.581

Therefore, Standard Deviation of x = 2.581
Find for Y using same method

The Standard Deviation of y = 10.29
Correlation = 26.67 /(2.581 * 10.29)
Correlation = 1.0041

```
height = [5.1,5.2,5.3,5.4,5.5]
weight=[60.3,59.2,63.6,88.4,68.7]
print("correlation",np.corrcoef(height,weight))
```

```
correlation [[1.         0.60782997]
 [0.60782997 1.        ]]
```

```
print("correlation",np.corrcoef(height,weight)[0,1])
```

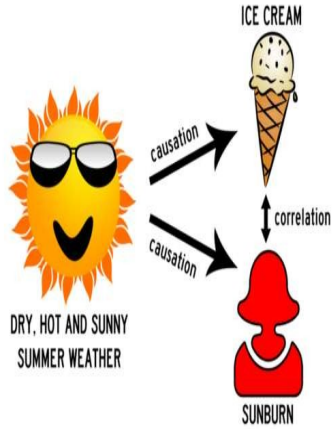
```
correlation 0.6078299663324911
```

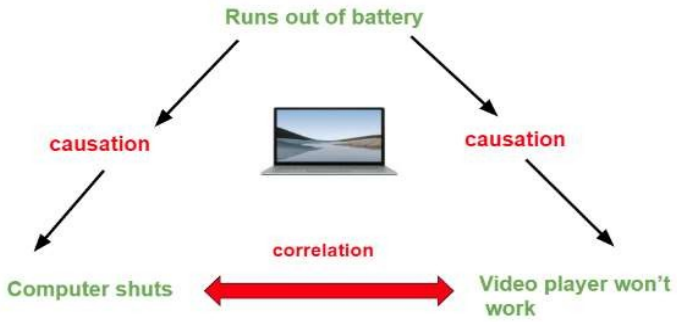
Causation

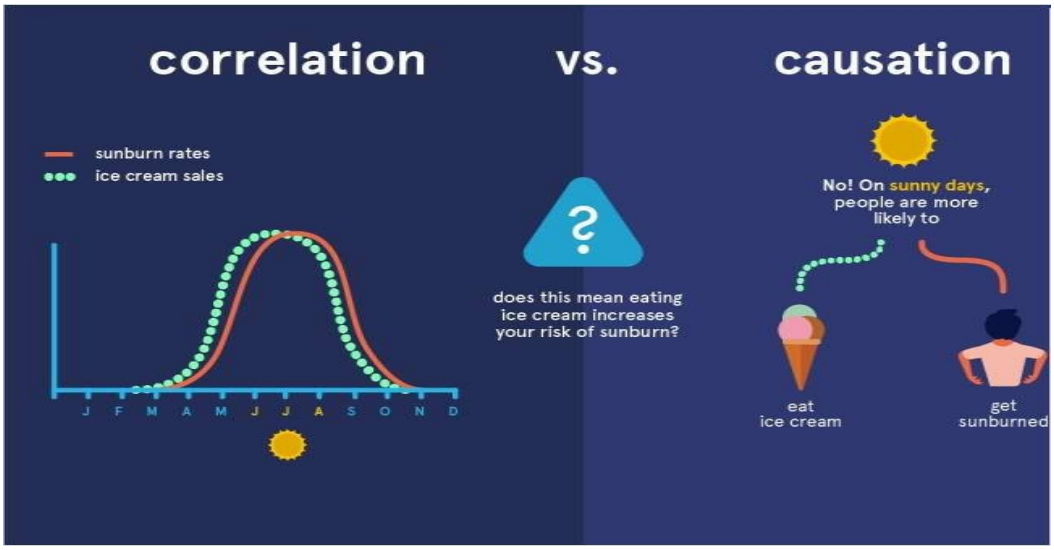
Causation means that changes in one variable bringsabout changes in the other; there is a cause-and- effect relationship between variables. The two variables are correlated with each other andthere is also a causal link between them.

Correlation and Causation can exist at the same time also, so definitely correlation doesn't imply causation.Below example is to show this difference more clearly

No battery in computer causes computer to shut and also causes video player to stop ,shows causality of battery over laptop and video player. The moment computer shuts, video player also shuts shows both are correlated. More specifically positively correlated.






correlation
vs.
causation
sunburn rates
ice cream sales
J F M A M J J A S O N D
does this mean eating ice cream increases your risk of sunburn?
No! On sunny days, people are more likely to
eat ice cream
get sunburned