Probability and Statistics for Data Science Part-1

towardsdatascience.com/probability-and-statistics-for-data-science-part-1-3eed6051c40d

Badreesh Shetty

November 22, 2018

Probability and Statistics form the basis of Data Science. The probability theory is very much helpful for making the prediction. Estimates and predictions form an important part of Data science. With the help of statistical methods, we make estimates for the further analysis. Thus, statistical methods are largely dependent on the theory of probability. And all of probability and statistics is dependent on Data.

Data

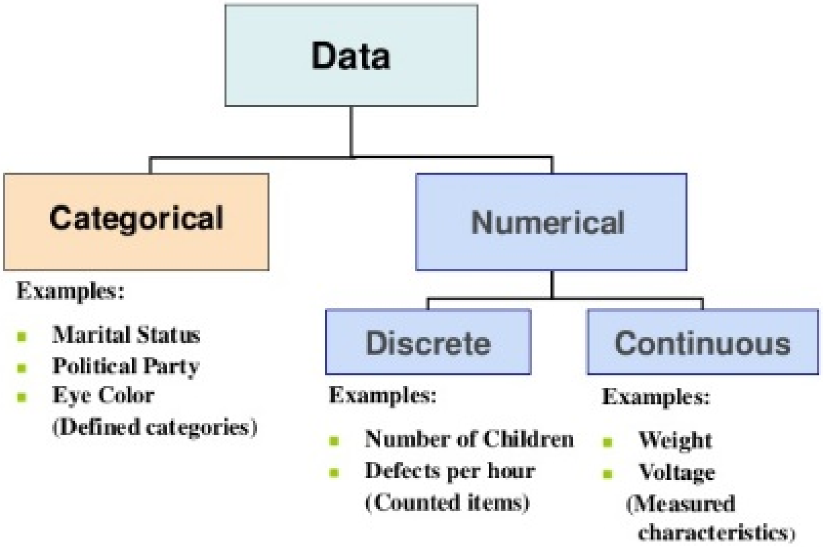
Data is the collected information(observations) we have about something or facts and statistics collected together for reference or analysis.

> Data—a collection of facts (numbers, words, measurements, observations, etc) that has been translated into a form that computers can process

Why does Data Matter?

- Helps in understanding more about the data by identifying relationships that may exist between 2 variables.
- Helps in predicting the future or forecast based on the previous trend of data.
- Helps in determining patterns that may exist between data.
- Helps in detecting fraud by uncovering anomalies in the data.

Data matters a lot nowadays as we can infer important information from it. Now let’s delve into how data is categorized. Data can be of 2 types categorical and numerical data. For Example in a bank, we have regions, occupation class, gender which follow categorical data as the data is within a fixed certain value and balance, credit score, age, tenure months follow numerical continuous distribution as data can follow an unlimited range of values.

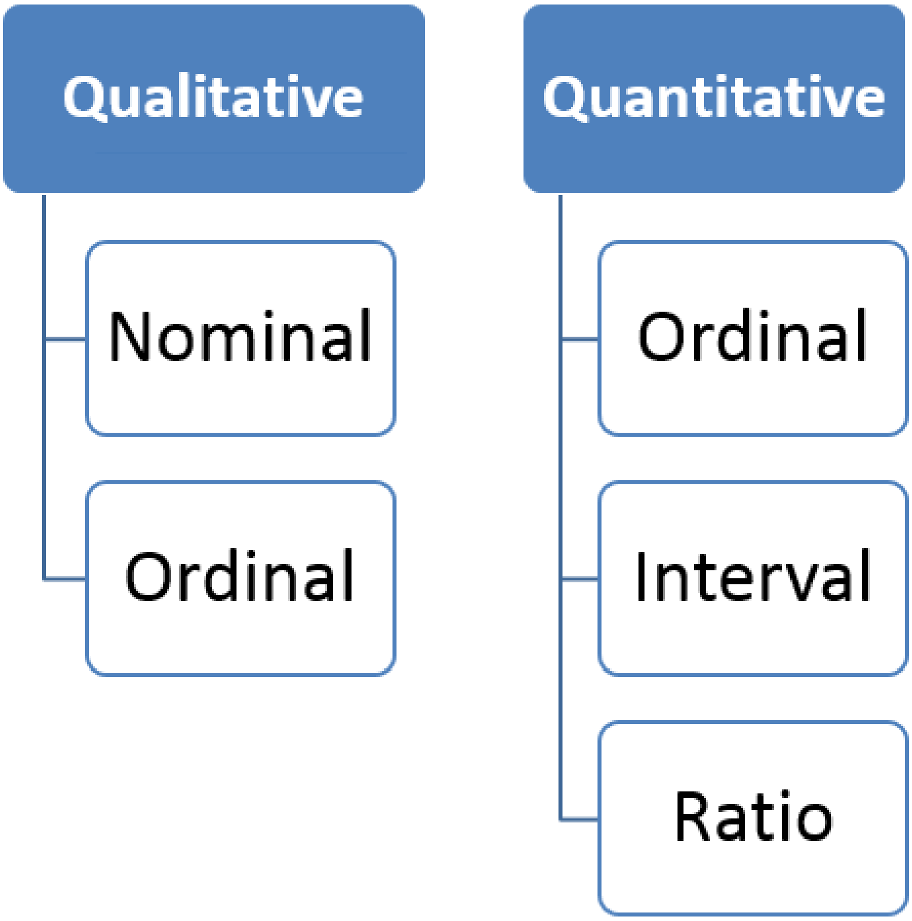


Note: Categorical Data can be visualized by Bar Plot, Pie Chart, Pareto Chart. Numerical Data can be visualized by Histogram, Line Plot, Scatter Plot

Descriptive Statistics

A descriptive statistic is a summary statistic that quantitatively describes or summarizes features of a collection of information. It helps us in knowing our data better. It is used to describe the characteristics of data.

Measurement level of Data



The qualitative and quantitative data is very much similar to the above categorical and numerical data.

Nominal: Data at this level is categorized using names, labels or qualities. eg: Brand Name, ZipCode, Gender.

Ordinal: Data at this level can be arranged in order or ranked and can be compared. eg: Grades, Star Reviews, Position in Race, Date

Interval: Data at this level can be ordered as it is in a range of values and meaningful differences between the data points can be calculated. eg: Temperature in Celsius, Year of Birth

Ratio: Data at this level is similar to interval level with added property of an inherent zero. Mathematical calculations can be performed on these data points. eg: Height, Age, Weight

Check this out:DATA | Mahrita Harahap

Population or Sample Data

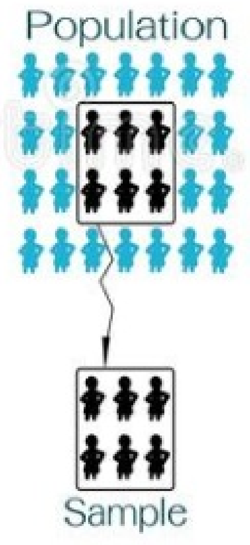
Before performing any analysis of data, we should determine if the data we’re dealing with is population or sample.

Population: Collection of all items (N) and it includes each and every unit of our study. It is hard to define and the measure of characteristic such as mean, mode is called parameter.

Sample: Subset of the population (n) and it includes only a handful units of the population. It is selected at random and the measure of the characteristic is called as statistics.

For Example, For example, say you want to know the mean income of the subscribers to a movie subscription service(parameter). We draw a random sample of 1000 subscribers and determine that their mean income(x̄) is $34,500 (statistic). We conclude that the population mean income (μ) is likely to be close to $34,500 as well.

Now before looking at distributions of data. Let’s take a look at measures of data.



Measures of Central Tendency

The measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location. They are also classed as summary statistics.

Mean: The mean is equal to the sum of all the values in the data set divided by the number of values in the data set i.e the calculated average. **It susceptible to outliers** when unusual values are added it gets skewed i.e deviates from the typical central value.

Median: The median is the middle value for a dataset that has been arranged in order of magnitude. Median is a better alternative to mean as it is less affected by outliers and skewness of the data. The median value is much closer than the typical central value.

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

If the total number of values is odd then

If the total number of values is even then

$$\text{Median} = (\frac{n+1}{n})^{th}\ term$$

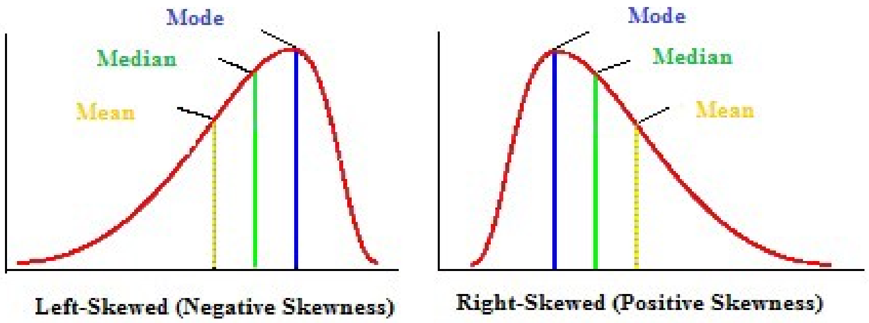
$$\text{Median} = (\frac{(\frac{n}{2})^{th} term + (\frac{n}{2}+1)^{th} term}{2})^{th}\ \ term$$

Mode: The mode is the most commonly occurring value in the dataset. The mode can, therefore sometimes consider the mode as being the most popular option.

For Example, In a dataset containing {13,35,54,54,55,56,57,67,85,89,96} values. Mean is 60.09. Median is 56. Mode is 54.

Measures of Asymmetry

Skewness: Skewness is the asymmetry in a statistical distribution, in which the curve appears distorted or skewed towards to the left or to the right. Skewness indicates whether the data is concentrated on one side.

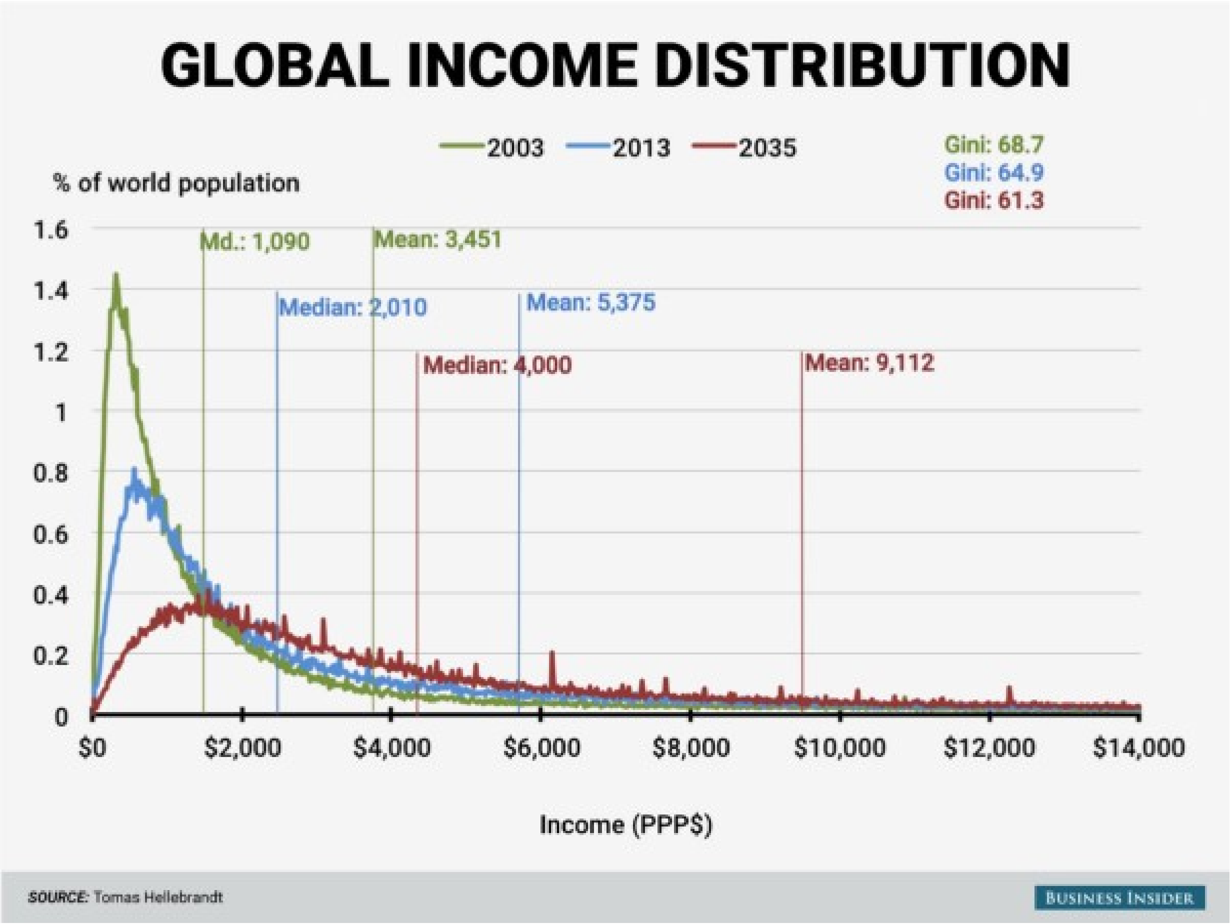

Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

Source: Pearson Mode Skewness: Definition and Formulas

Positive Skewness: Positive Skewness is when the mean>median>mode. The outliers are skewed to the right i.e the tail is skewed to the right.

Negative Skewness: Negative Skewness is when the mean<median<mode. The outliers are skewed to the left i.e the tail is skewed to the left.

Skewness is important as it tells us about where the data is distributed.



SOURCE: Tomas Hellebrandt

For eg: Global Income Distribution in 2003 is highly right-skewed.We can see the mean $3,451 in 2003(green) is greater than the median $1,090. It suggests that the global income is not evenly distributed. Most individuals incomes are less than $2,000 and less number of people with income above $14,000, so the skewness. But it seems in 2035 according to the forecast income inequality will decrease over time.

Measures of Variability(Dispersion)

The measure of central tendency gives a single value that represents the whole value; however, the central tendency cannot describe the observation fully. The measure of dispersion helps us to study the variability of the items i.e the spread of data.

Remember: Population Data has N data points and Sample Data has (n-1) data points. (n-1) is called Bessel's Correction and it is used to reduce bias.

Range: The difference between the largest and the smallest value of a data, is termed as the range of the distribution. Range does not consider all the values of a series, i.e. it takes only the extreme items and middle items are not considered significant. eg: For {13,33,45,67,70} the range is 57 i.e(70–13).

Variance: Variance measures how far is the sum of squared distances from each point to the mean i.e the dispersion around the mean.

Variance is the average of all squared deviations.

Note: *The units of values and variance is not equal so we use another variability measure.*

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n} \text{ for populations}$$

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \text{ for samples}$$

Standard Deviation: AsVariance suffers from unit difference so standard deviation is used. The square root of the variance is the standard deviation. It tells about the concentration of the data around the mean of the data set.

Population standard deviation: σ

= square root of the population variance

$$\sigma = \sqrt{\sigma^2}$$

Sample standard deviation: s

= square root of the sample variance, so that

$$s = \sqrt{s^2}$$

For eg: {3,5,6,9,10} are the values in a dataset.

$$\text{Mean} = \frac{3+5+6+9+10}{5} = 6.6$$

$$\text{Variance} = \frac{(3-6.6)^2 + (5-6.6)^2 + (6-6.6)^2 + (9-6.6)^2 + (10-6.6)^2}{5}$$

$$= \frac{12.96 + 2.56 + 0.36 + 5.76 + 11.56}{5} = \frac{33.2}{5} = 6.64$$

$$\text{Standard Deviation} = \sqrt{\text{Variance}} = \sqrt{6.64} = 2.576$$

Coefficient of Variation(CV): It is also called as the relative standard deviation. It is the ratio of standard deviation to the mean of the dataset.

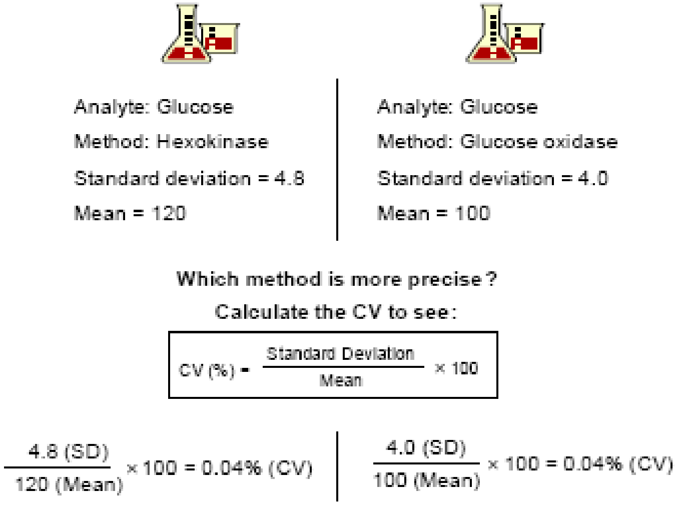
Standard deviation is the variability of a single dataset. Whereas the coefficient of variance can be used for comparing 2 datasets.

CV for a population:

$$CV = \frac{\sigma}{\mu} * 100\%$$

CV for a sample:

$$CV = \frac{s}{\bar{x}} * 100\%$$



Analyte: Glucose
Method: Hexokinase
Standard deviation = 4.8
Mean = 120

Analyte: Glucose
Method: Glucose oxidase
Standard deviation = 4.0
Mean = 100

Which method is more precise?
Calculate the CV to see:

$$\boxed{CV\ (\%) = \frac{\text{Standard Deviation}}{\text{Mean}} \times 100}$$

$$\frac{4.8\ (SD)}{120\ (Mean)} \times 100 = 0.04\%\ (CV)$$

$$\frac{4.0\ (SD)}{100\ (Mean)} \times 100 = 0.04\%\ (CV)$$

From the above example, we can see that the CV is the same. Both methods are precise. So it is perfect for comparisons.

Measures of Quartiles

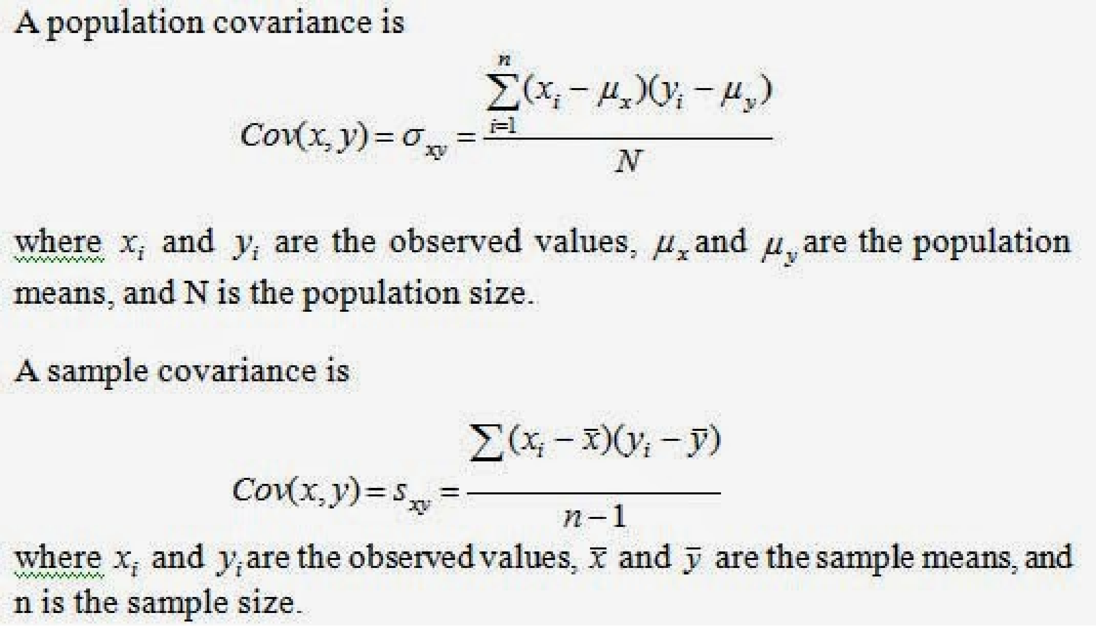
Quartiles are better at understanding as every data point considered.

Check my previous post — In the Boxplot Section, I have elaborated on Quartiles.

Measures of Relationship

Measures of relationship are used to find the comparison between 2 variables.

Covariance: Covariance is a measure of the relationship between the variability of 2 variables i.e It measures the degree of change in the variables, when one variable changes, will there be the same/a similar change in the other variable.



A population covariance is

$$Cov(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

where x_i and y_i are the observed values, μ_x and μ_y are the population means, and N is the population size.

A sample covariance is

$$Cov(x,y) = s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

where x_i and y_i are the observed values, $\bar{x}$ and $\bar{y}$ are the sample means, and n is the sample size.

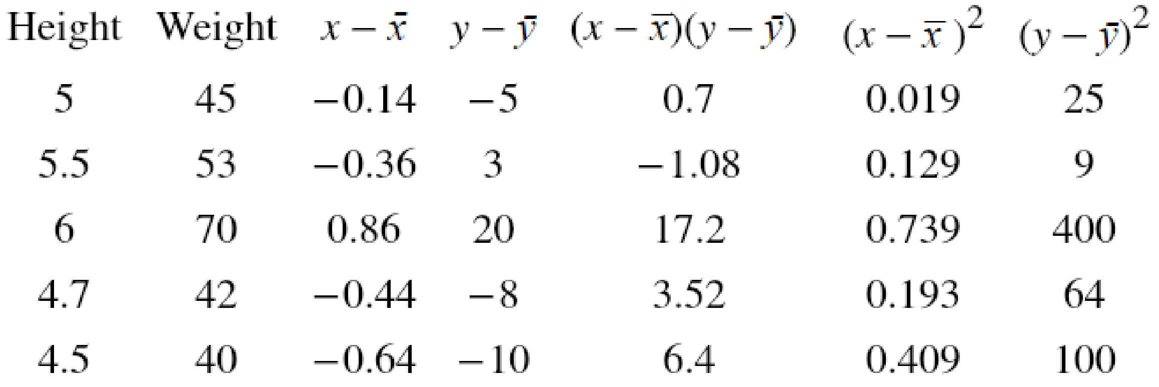
Covariance does not give effective information about the relation between 2 variables as it is not normalized.

Correlation: Correlation gives a better understanding of covariance. It is normalized covariance. Correlation tells us how correlated the variables are to each other. It is also called as Pearson Correlation Coefficient.

$$Correlation = \rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

The value of correlation ranges from -1 to 1. -1 indicates negative correlation i.e with an increase in 1 variable independent there is a decrease in the other dependent variable.1 indicates positive correlation i.e with an increase in 1 variable independent there is an increase in the other dependent variable.0 indicates that the variables are independent of each other.

For Example,



Height	Weight	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
5	45	-0.14	-5	0.7	0.019	25
5.5	53	-0.36	3	-1.08	0.129	9
6	70	0.86	20	17.2	0.739	400
4.7	42	-0.44	-8	3.52	0.193	64
4.5	40	-0.64	-10	6.4	0.409	100

$$\text{Sum(Height)} = 25.7 \ \text{Mean(Height)} = 5.14$$

$$\text{Sum(Weight)} = 250 \ \text{Mean(Weight)} = 50$$

$$\sum(x-\bar{x})(y-\bar{y}) = 26.74$$

$$\sum(x-\bar{x})^2 = 1.489$$

$$\sum(y-\bar{y})^2 = 598$$

$$\text{Correlation} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{26.74}{\sqrt{1.489}\sqrt{598}} = \frac{26.54}{1.220 * 24.454} = 0.889$$

Correlation 0.889 tells us Height and Weight has a positive correlation. It is obvious that as the height of a person increases weight too increases.

Note: Correlation does not imply causation, Spurious Correlation for some strange correlations.

Conclusion

In this article, we learnt about descriptive statistics which helps us to know about our data better by understanding crucial characteristics in a dataset.