

Module 5 - Business Data Analytics

Understand business analytics and develop business intelligence.

(17 hours)

In this section, we will discuss:

- Introduction to business analytics and concepts of business analytics.
- Trends in business analytics.
- Introduction to Big Data Analytics

Introduction to business analytics and Concepts of business analytics

What is Business Analytics?

- Business analytics (BA) is the iterative, methodical exploration of an organization's data, with an emphasis on statistical analysis.
- Business analytics is used by companies that are committed to making data-driven decisions.

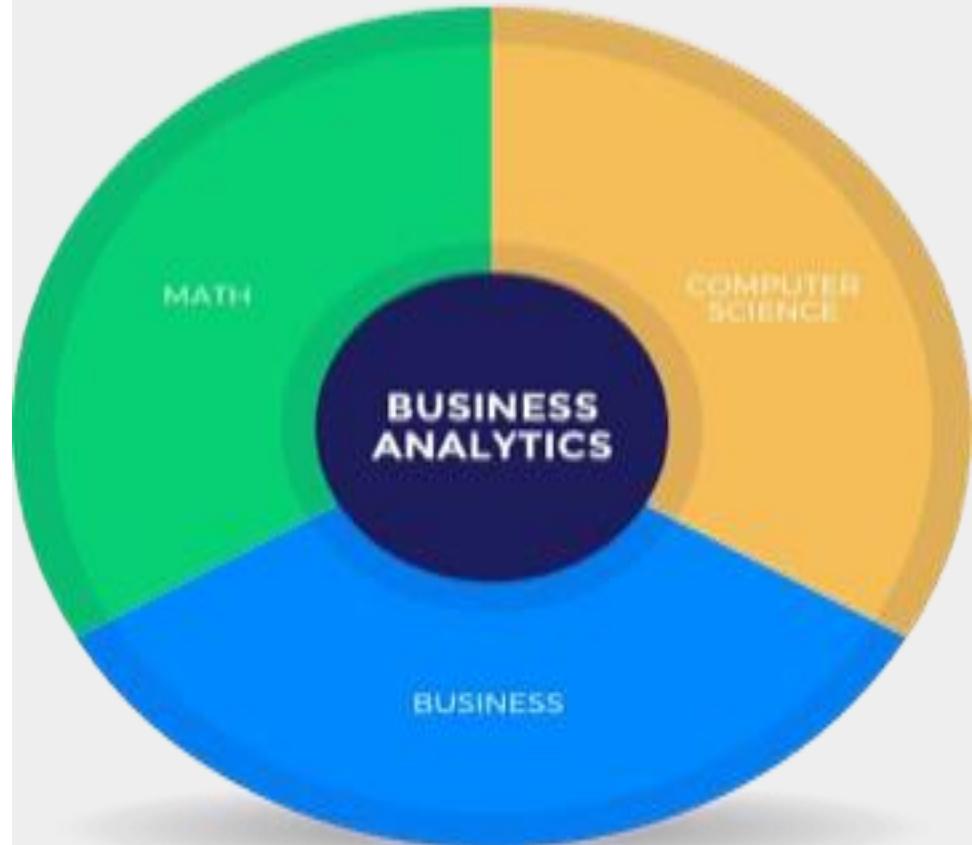


Image Source: <https://www.businessanalytics.com/>

Introduction to business analytics and Concepts of business analytics

What is Business Analytics?(Contd)

- Business Analytics is "the study of data through statistical and operations analysis, the formation of predictive models, application of optimization techniques, and the communication of these results to customers, business partners, and college executives".



Image Source: <https://www.proschoolonline.com/certification-business-analytics-course/what-is-b>

Introduction to business analytics and Concepts of business analytics

What is Business Analytics?(Contd)

- It adopts quantitative methods and evidence is required for data to build certain models for businesses and make profitable decisions. Thus, Business Analytics majorly depends on and uses Big Data(large volume of data) .



Introduction to business analytics and Concepts of business analytics

Understanding Business Analytics

- Business Analytics is the procedure through which information is dissected after studying past performances and issues, to devise a successful plan for the future.
- Big Data or large amounts of data is used to derive solutions.

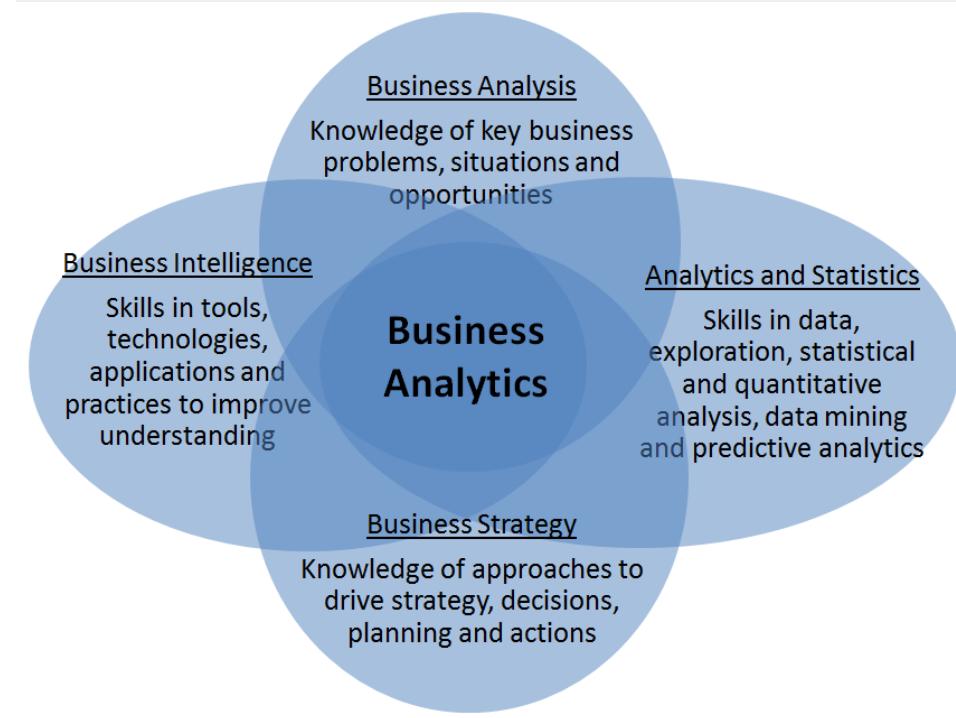


Image Source: <https://www.indiaeducation.net/managementstreams/business-analytics.html>

Introduction to business analytics and Concepts of business analytics

Understanding Business Analytics

- This method of going about a business or this outlook towards building and sustaining a business is vital to the economy and industries that thrive in the economy.



Introduction to business analytics and Concepts of business analytics

Components of Business Analytics

- Define Objective
- Data Aggregation
- Data Cleaning
- Analytical Methodology
- Evaluation and Validation
- Reporting and Data Visualisation

Components of Business Analytics

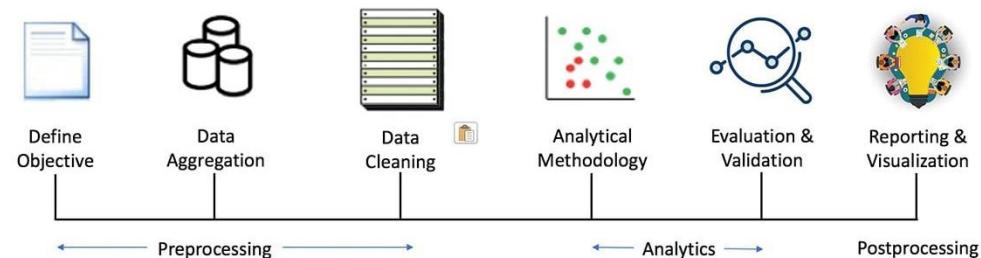


Image Source: <https://www.analytixlabs.co.in/blog/what-is-business-analytics>

Introduction to business analytics and Concepts of business analytics

Types of Business Analytics Methods

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics

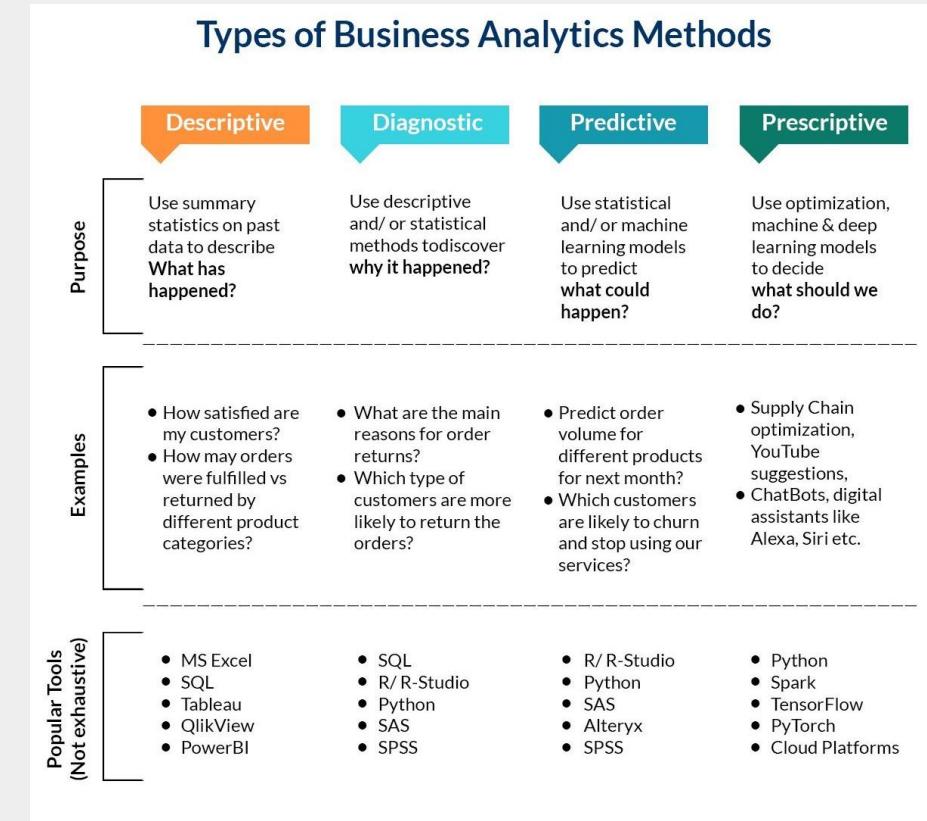


Image Source: <https://www.analytixlabs.co.in/blog/what-is-business-analytics/>

Introduction to business analytics and Concepts of business analytics

Uses and Benefits of Business Analytics

- To carry out data mining and exploring new data to find new patterns and relationships.
- To carry out statistical and quantitative analysis to provide explanations for certain occurrences.

Benefits of Business Analytics



Introduction to business analytics and Concepts of business analytics

Uses and Benefits of Business Analytics

- Test previous decisions are taken with the help of A/B testing and multivariate testing.
- Deploy predictive modeling to predict future outcomes

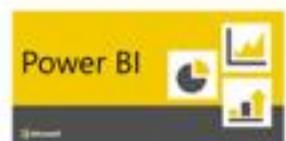


Image Source: <https://www.datapine.com/blog/benefits-of-business-intelligence-and-business-an>

Introduction to business analytics and Concepts of business analytics

Business Analytics Tools

- SQL
- Tableau/ QlikView/ Power BI
- Birt
- Python •R
- MS Excel
- Sisense
- Clear Analytics
- Pentaho BI
- MicroStrategy



Introduction to business analytics and Concepts of business analytics

Applications of Business Analytics

- Marketing
- Finance
- Human Resources
- Manufacturing



Image Source: <https://www.proschoolonline.com/certification-business-analytics-course/what-is-b>

Trends in Business Analytics

Business Analytics Trends For 2021

- Data Quality Management
- Data Discovery/Visualization
- Artificial Intelligence
- Predictive and Prescriptive Analytics
- Tools
- Collaborative Business Intelligence
- Data-driven Culture



Image Source: <https://www.datapine.com/blog/benefits-of-business-intelligence-and-business-an>

Trends in Business Analytics

Business Analytics Trends For 2021

- Augmented Analytics
- Mobile BI
- Data Automation
- Embedded Analytics
- Natural language processing

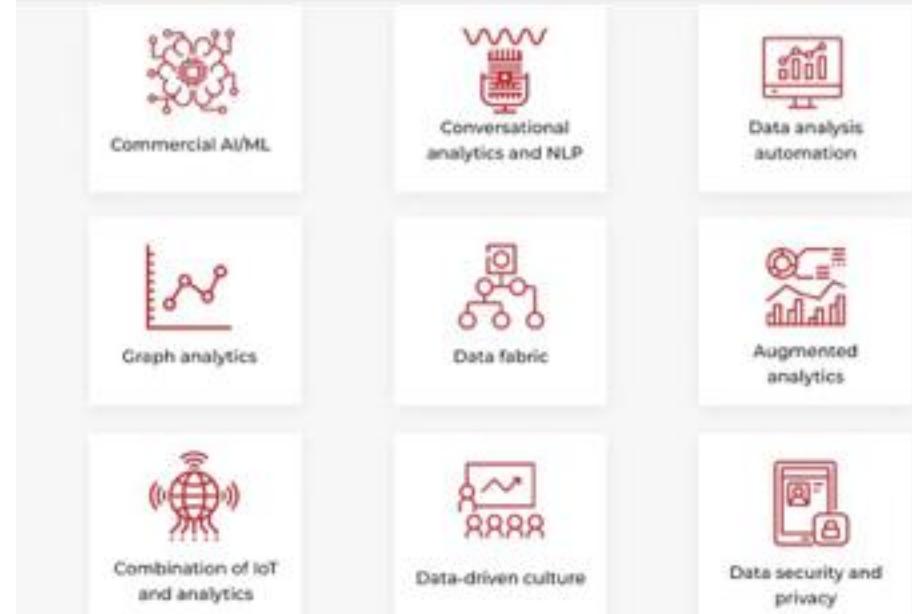


Image Source: <https://codeit.us/blog/top-data-and-analytics-trends>

Descriptive analytics

What is Descriptive Analytics?

- Descriptive analytics is a statistical method that is used to search and summarize historical data in order to identify patterns or meaning.
- Descriptive analytics are based on standard aggregate functions in databases

Descriptive Analytics

Business Intelligence and Data mining

"The simplest class of analytics, one that allows you to condense big data into smaller, more useful nuggets of information." Dr. Michael Wu



90% of organizations use descriptive analytics.



Analyses the data coming in real-time and historical data for insights on how to approach the future.



Most of the social analytics are descriptive analytics.

Descriptive analytics

What is Descriptive Analytics?
(Contd..)

- For example, in an online learning course with a discussion board, descriptive analytics could determine how many students participated in the discussion, or how many times a particular student posted in the discussion forum.



Image Source: <https://www.valamis.com/hub/descriptive-analytics>

Descriptive analytics

How does descriptive analytics work?

- Data aggregation and data mining are two techniques used in descriptive analytics to discover historical data.
- Data is first gathered and sorted by data aggregation in order to make the datasets more manageable by analysts.



Image Source: <https://www.dataversity.net/fundamentals-descriptive-analytics/>

Descriptive analytics

How does descriptive analytics work?
(Contd..)

- Data mining describes the next step of the analysis and involves a search of the data to identify patterns and meaning.
- Identified patterns are analyzed to discover the specific ways that learners interacted with the learning content and within the learning environment.



Image Source: <https://www.sisense.com/glossary/descriptive-analytics/>

Descriptive analytics

Examples of descriptive analytics

- Tracking course enrolment's, course compliance rates,
- Recording which learning resources are accessed and how often
- Summarizing the number of times a learner posts in a discussion board
- Tracking assignment and assessment grades

Data consolidation
Our data scientists bring together large volumes of collected data relevant to the objective.



Image Source: <https://www.vertical-leap.uk/blog/data-science-for-marketers-part-2-descriptive-v->

Descriptive analytics

Examples of descriptive analytics (Contd..)

- Comparing pre-test and post-test assessments
- Analyzing course completion rates by learner or by course
- Collating course survey results
- Identifying length of time that learners took to complete a course

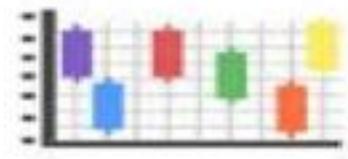
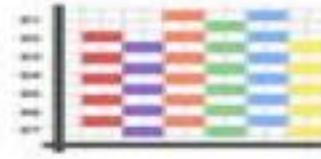
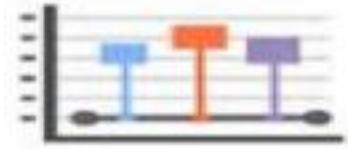


Image Source: <https://www.vectorstock.com/royalty-free-vector/data-analytics-icons-flat-pack-vec>

Descriptive analytics

Advantages of descriptive analytics

- Quickly and easily report on the Return on Investment (ROI) by showing how performance achieved business or target goals.
- Identify gaps and performance issues early - before they become problems.



Image Source: <https://forums.bsdinsight.com/threads/descriptive-predictive-and-prescriptive-anal>

Descriptive analytics

Advantages of descriptive analytics
(Contd..)

- Identify specific learners who require additional support, regardless of how many students or employees there are
- Identify successful learners in order to offer positive feedback or additional resources.
- Analyze the value and impact of course design and learning resources.

TOO BUSY



MUCH BETTER



Image Source: <https://econsultancy.com/analytics-approaches-every-marketer-should-know-1-de>

Introduction to Big Data Analytics

What is Data?

- The quantities, characters, or symbols on which operations are performed by a computer, which may be stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media.



Image Source: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcSCqNFP8VjcmqJX2EyEd-2mOaHcwSqTiXQVjCP1ISvmclxoMYvCms5tQ_9imGeKaTmuBaA&usqp=CAU

Introduction to Big Data Analytics

What is Big Data?

- **Big Data** is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size



Image Source:

https://www.guru99.com/images/Big_Data/061114_0759_WhatIsBigDa1.jpg

Introduction to Big Data Analytics

Example of Big Data

- The **New York Stock Exchange** is an example of Big Data that generates about **one terabyte** of new trade data per day.



Image Source:

https://www.guru99.com/images/Big_Data/061114_0759_WhatIsBigDa2.jpg

Introduction to Big Data Analytics

Example of Big Data Social Media

- The statistic shows that **500+terabytes** of new data get ingested into the databases of social media site **Facebook**, every day. This data is mainly generated in terms of photo and video uploads, message exchanges, putting comments etc.



Image Source:

https://www.guru99.com/images/Big_Data/061114_0759_WhatIsBigDa3.jpg

Introduction to Big Data Analytics

Example of Big Data

- A single **Jet engine** can generate **10+terabytes** of data in **30 minutes** of flight time. With many thousand flights per day, generation of data reaches up to many **Petabytes**.



Image Source:

https://www.guru99.com/images/Big_Data/061114_0759_WhatIsBigDa4.jpg

Introduction to Big Data Analytics

Types Of Big Data

Following are the types of Big Data:

- Structured
 - Unstructured
 - Semi-structured



Image Source:

https://www.guru99.com/images/Big_Data/061114_0759_WhatIsBigData.jpg

Introduction to Big Data Analytics

Structured Big Data

Any data that can be stored, accessed and processed in the form of fixed format is termed as a ‘structured’ data.



Image Source:

https://www.guru99.com/images/Big_Data/061114_0759_WhatIsBigData.jpg

Introduction to Big Data Analytics

Unstructured Big Data

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it.



Image Source:

https://www.guru99.com/images/Big_Data/061114_0759_WhatIsBigDa4.jpg

Introduction to Big Data Analytics

Semi-structured Big Data

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational DBMS.

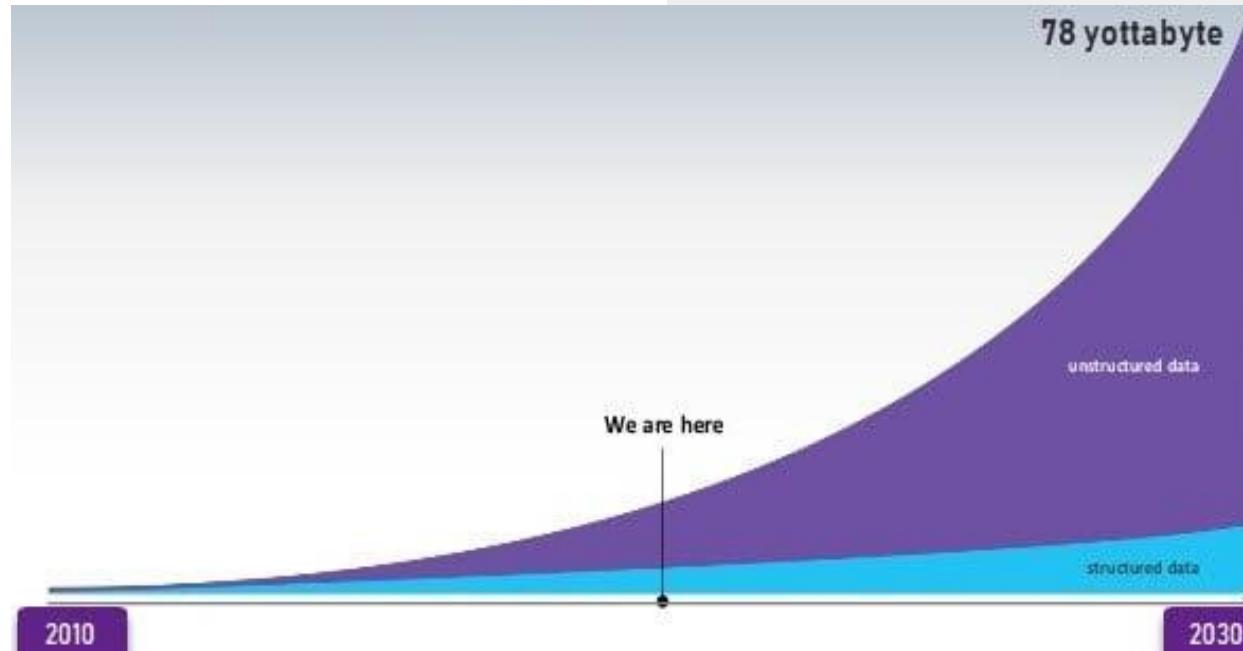


Image Source:

https://www.guru99.com/images/Big_Data/061114_0759_WhatIsBigData.jpg

Introduction to Big Data Analytics

Data Growth over the years



Introduction to Big Data Analytics

Characteristics Of Big Data

Big data can be described by the following characteristics:

- Volume
- Variety
- Velocity
- Variability



Image Source: <https://qph.cf2.quoracdn.net/main-qimg-b093fe2a8f3d7ed42897bd85c33a4075-lq>

Statistics

(16 hours)

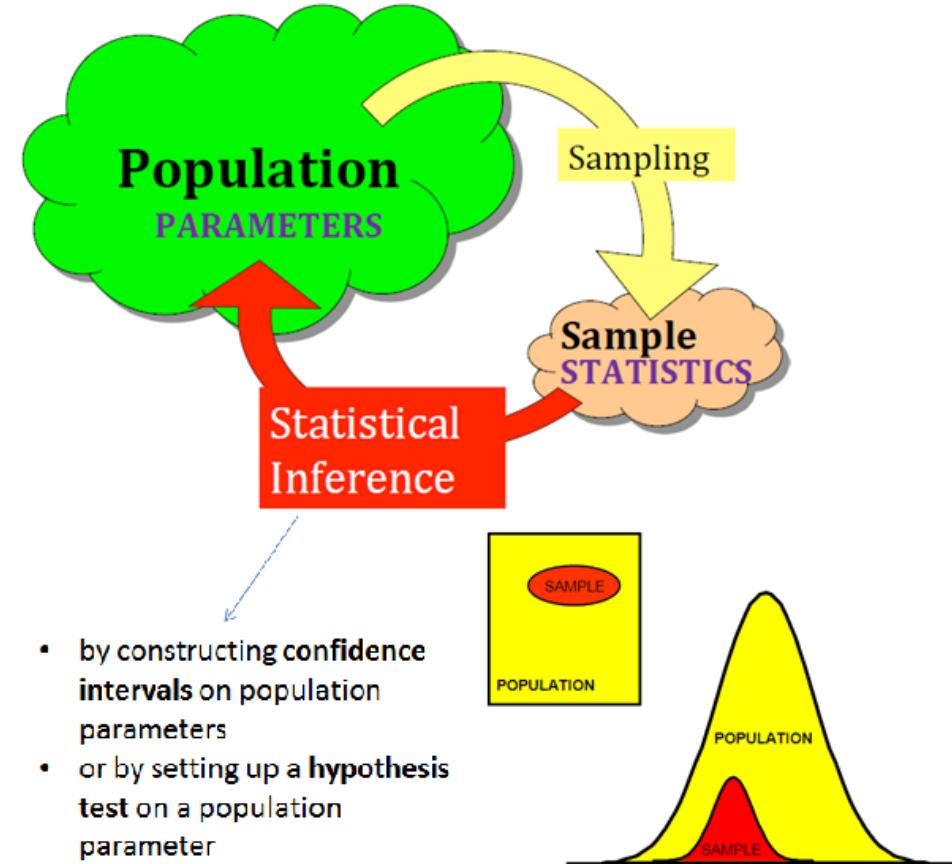
In this section, we will discuss:

- Views in Laravel with complete conditional and looping construct.
- Controllers and its usage.
- Complete database connectivity with DB
- Complete Database connectivity with Eloquent Model and its working.

Inferential Statistics

Introduction to Inferential statistics

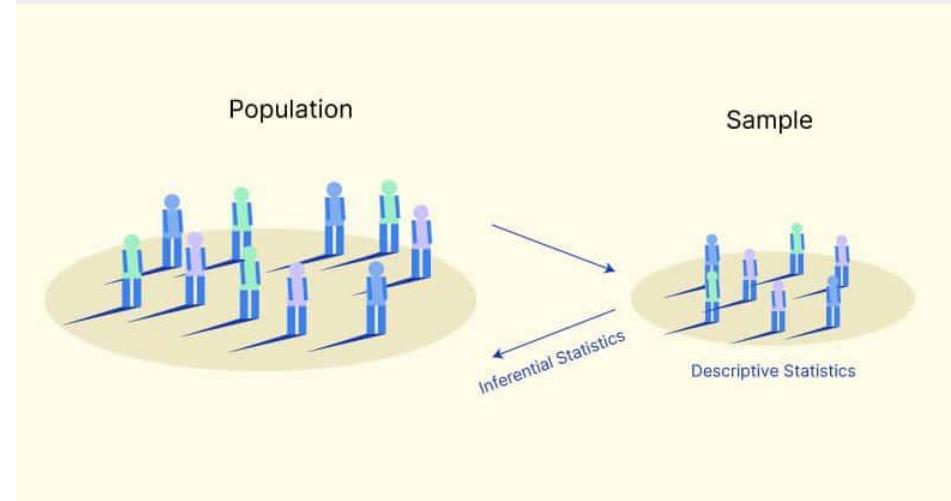
- Inferential statistics is a scientific discipline that uses mathematical tools to make forecasts and projections by analyzing the given data.
- This is of use to people employed in such fields as engineering, economics, biology, the social sciences, business, agriculture and communications.



Inferential Statistics

Advantages of Inferential statistics

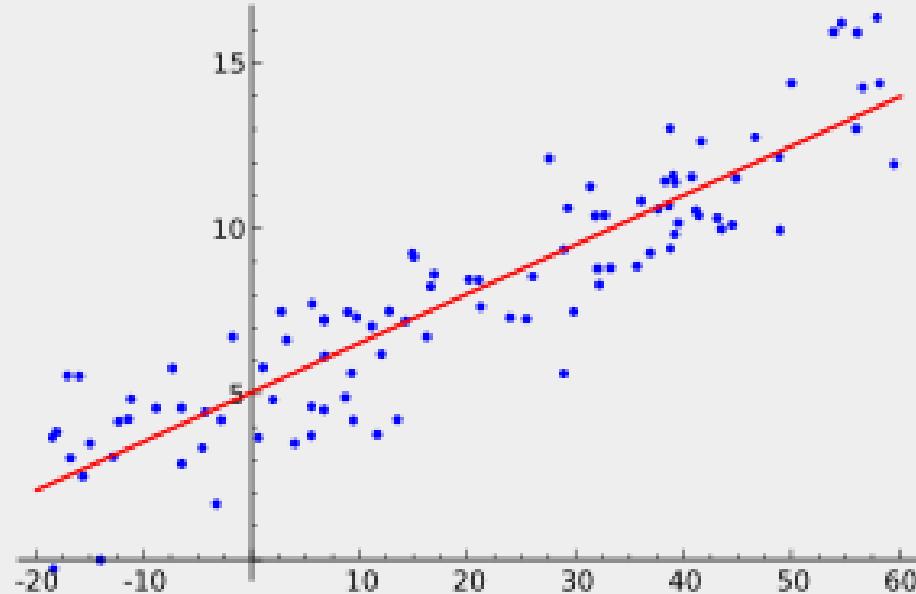
- A precise tool for estimating population.
- Highly structured analytical methods.



Inferential Statistics

Inferential Statistics Examples

- **Regression Analysis** is one of the most popular analysis tools.
- Regression analysis is used to predict the relationship between independent variables and the dependent variable.



Inferential Statistics

Inferential Statistics Examples

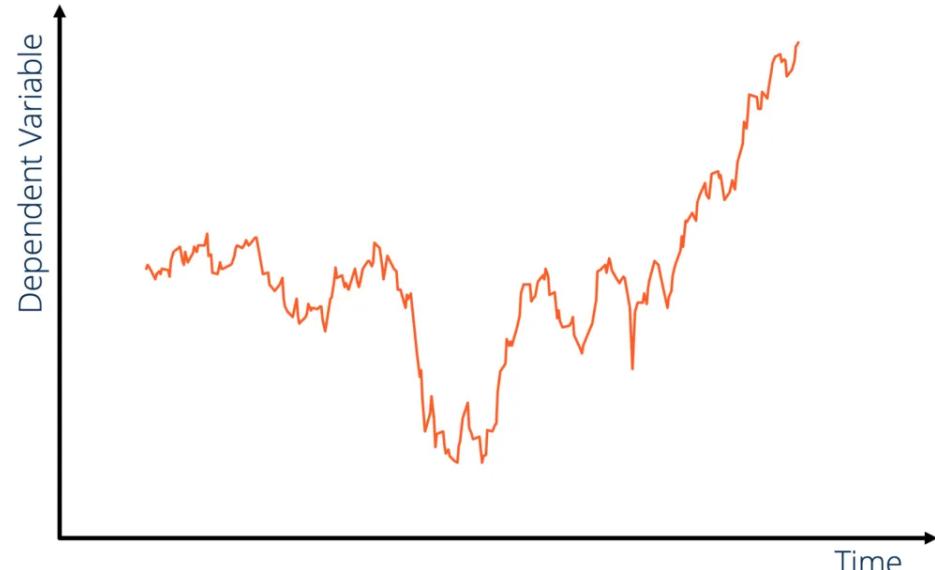
- **Hypothesis testing** is a statistical test where we want to know the truth of an assumption or opinion that is common in society.
- **Confidence interval** or confidence level is a statistical test used to estimate the population by using samples.

Inferential Statistics

Inferential Statistics Examples

- **Time series analysis** is one type of statistical analysis that tries to predict an event in the future based on pre-existing data.
- With this method, we can estimate how predictions a value or event that appears in the future.

Time-Series Analysis



Descriptive Statistics

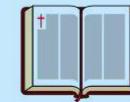
Introduction to descriptive statistics

- It is used to describe the basic features of data in a study.
- Descriptive statistics deals with the processing of data without attempting to draw any inferences from it.
- The data are presented in the form of tables and graphs.

Descriptive Statistics are procedures to organize, summarize, and present data in an informative way.

EXAMPLE 1:

The average test score for the students in a class, to give a descriptive sense of the typical scores.



EXAMPLE 2:

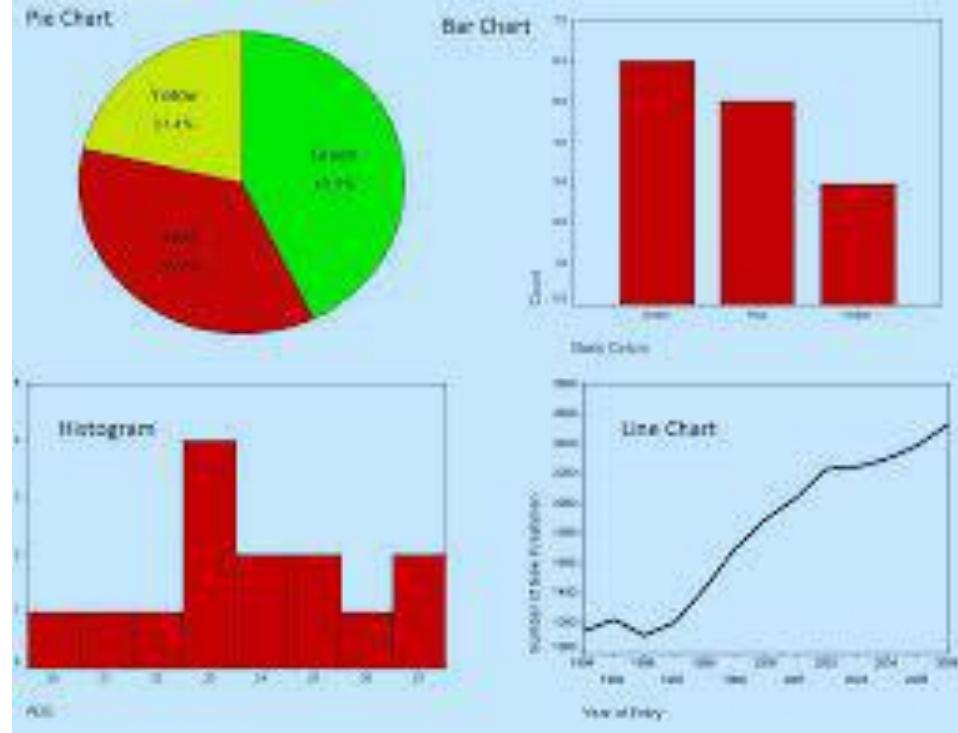
According to Consumer Reports, there were 2.5 problems per one copying machines reported during 2009.



Descriptive Statistics

Introduction to descriptive statistics

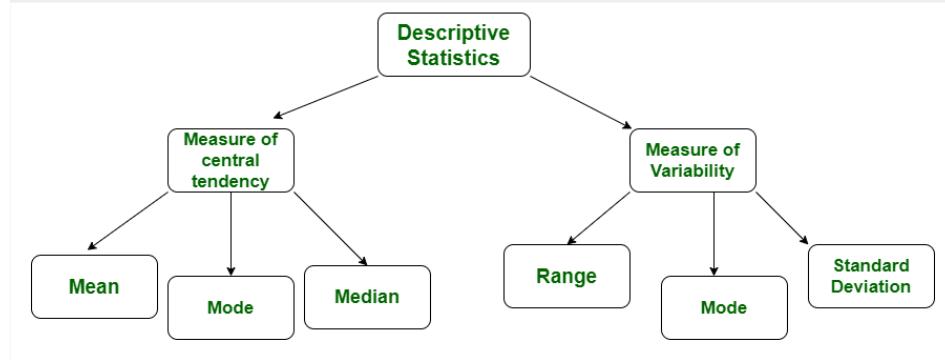
- The characteristics of the data are described in simple terms.
- Events that are dealt with include everyday happenings such as accidents, prices of goods, business, incomes, epidemics, sports data, population data.



Descriptive Statistics

Types of descriptive statistics

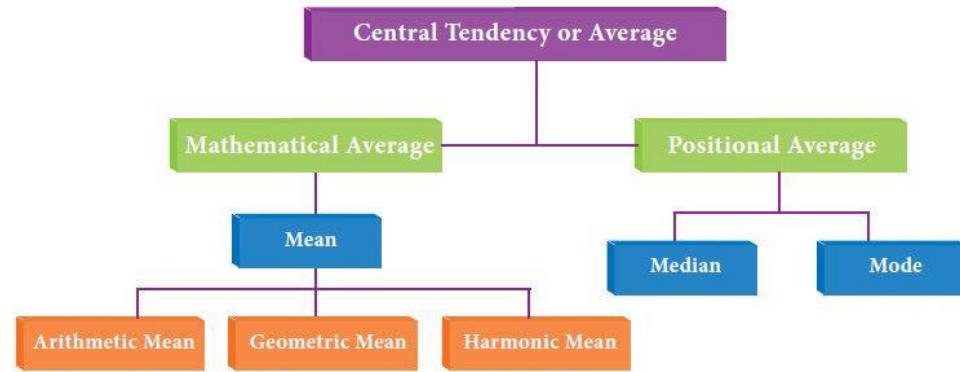
- All descriptive statistics are either measures of central tendency or measures of variability, also known as measures of dispersion.



Descriptive Statistics

Measure of Central Tendency

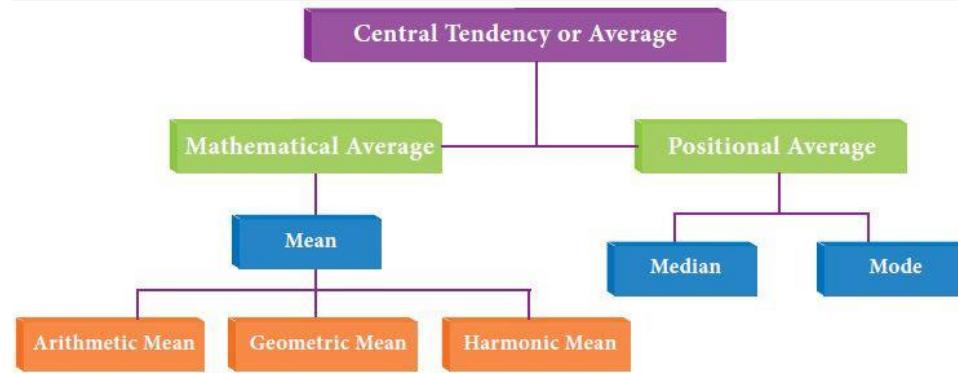
- Measures of central tendency focus on the average or middle values of data sets.
- These measures indicate where most values in a distribution fall and are also referred to as the central location of a distribution.



Descriptive Statistics

Measure of Central Tendency continued...

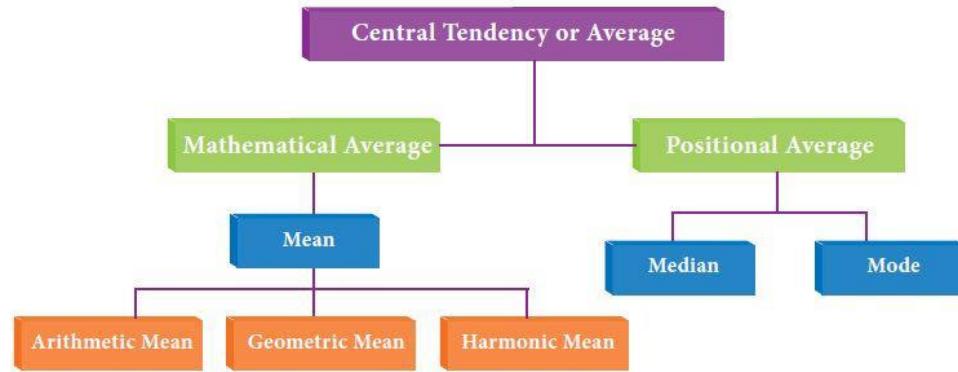
- We can think of it as the tendency of data to cluster around a middle value.
- In statistics the three most common measures of central tendency are the mean, median and mode.



Descriptive Statistics

Measure of Central Tendency continued...

- Each of these measures calculates the location of the central point using a different method.
- Choosing the best measure of central tendency depends on the type of data we have.



Measure of Central Tendency

Mean

- The mean is the arithmetic average, and it is probably the measure of central tendency that you are most familiar.
- Calculating the mean is very simple.

Mean Formula

$$\text{Mean} = \frac{\text{Sum of All Data Points}}{\text{Number of Data Points}}$$

$$\text{Mean} = \text{Assumed Mean} + \frac{\text{Sum of All Deviations}}{\text{Number of Data Points}}$$

Measure of Central Tendency

Mean continued...

- We just add up all of the values and divide by the number of observations in your dataset.
- $x_1 + x_2 + x_3 + \dots + x_n$

n

Mean

$$\text{Mean} = \frac{\text{Sum of Data Points}}{\text{Number of Data Points}}$$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Data Set: 6,4,10,3,7

$$\bar{x} = \frac{6 + 4 + 10 + 3 + 7}{5} = \frac{30}{5} = 6$$

Measure of Central Tendency

Mean continued...

- The calculation of the mean incorporates all values in the data.
- If you change any value, the mean changes.
- However, the mean doesn't always locate the center of the data accurately.

Mean

$$\text{Mean} = \frac{\text{Sum of Data Points}}{\text{Number of Data Points}}$$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

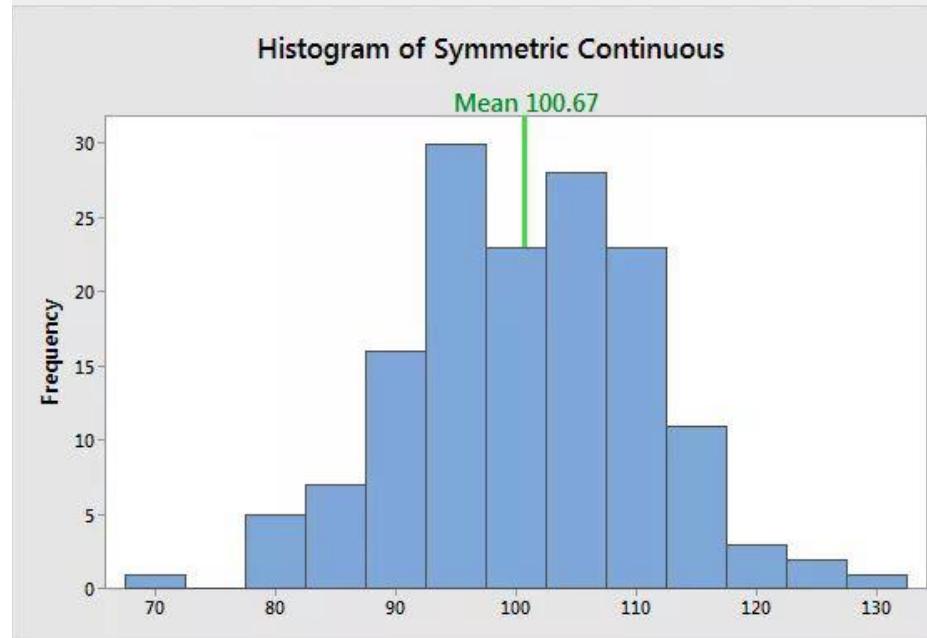
Data Set: 6,4,10,3,7

$$\bar{x} = \frac{6 + 4 + 10 + 3 + 7}{5} = \frac{30}{5} = 6$$

Measure of Central Tendency

Mean continued...

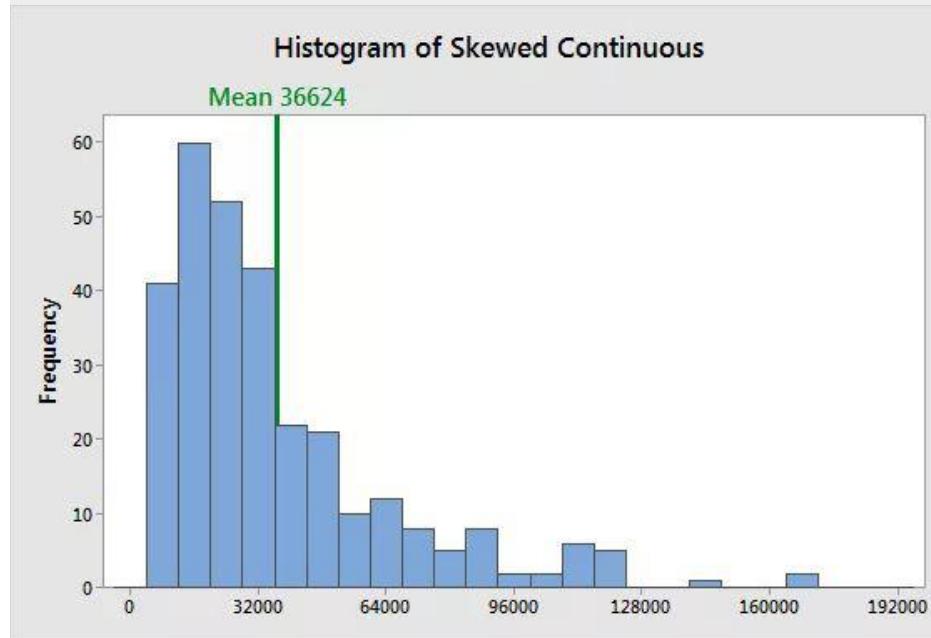
- In a symmetric distribution, the mean locates the center accurately.



Measure of Central Tendency

Mean continued...

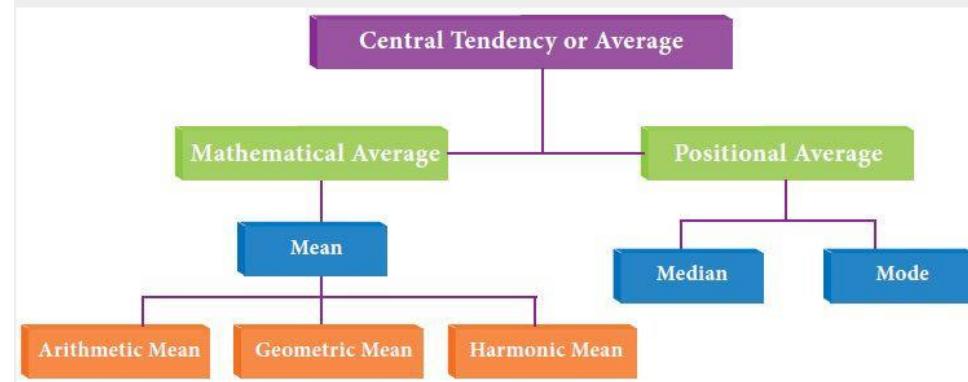
- However, in a skewed distribution, the mean can miss the mark.
- This problem occurs because outliers have a substantial impact on the mean.
- Extreme values in an extended tail pull the mean away from the center.
- As the distribution becomes more skewed, the mean is drawn further away from the center.



Measure of Central Tendency

Median

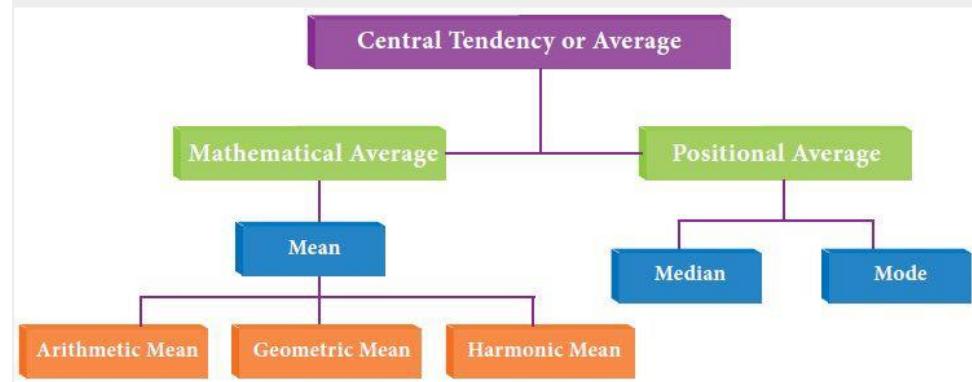
- The median is the middle value.
- It is the value that splits the dataset in half.
- To find the median, order your data from smallest to largest, and then find the data point that has an equal amount of values above it and below it.



Measure of Central Tendency

Median continued...

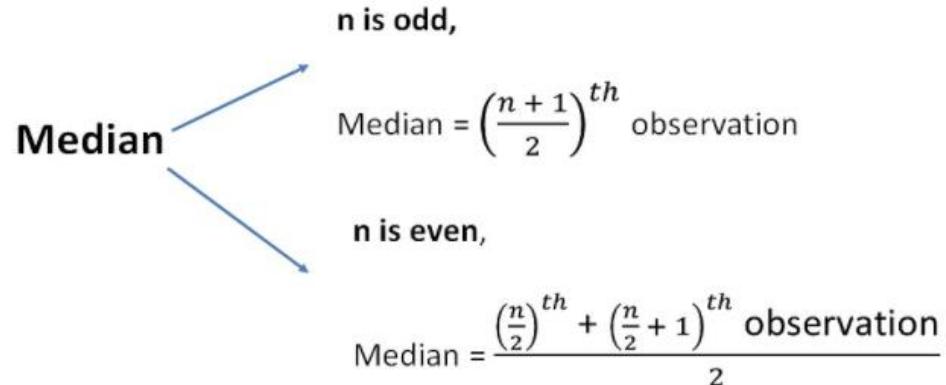
- The method for locating the median varies slightly depending on whether your dataset has an even or odd number of values.



Measure of Central Tendency

Median continued...

- In the dataset with the odd number of observations, notice how the number 12 has six values above it and six below it.
- Therefore, 12 is the median of this dataset.



Measure of Central Tendency

Median continued...

- When there is an even number of values, you count in to the two innermost values and then take the average.
- The average of 27 and 29 is 28.
- Consequently, 28 is the median of this dataset.

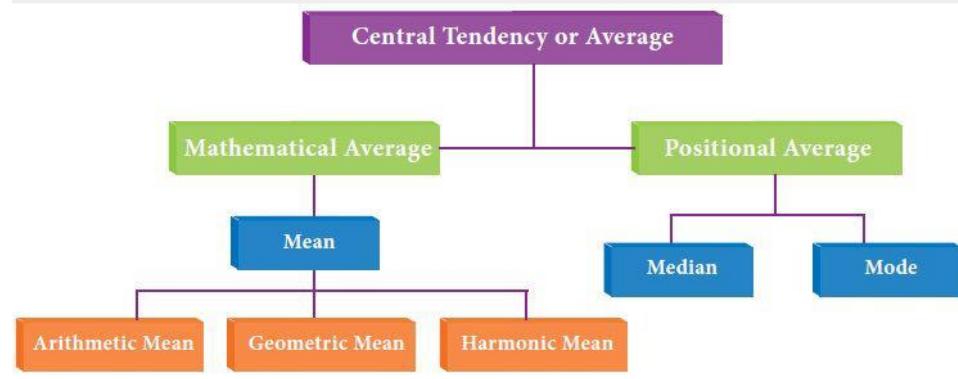
Median Odd
23
21
18
16
15
13
12
10
9
7
6
5
2

Median Even
40
38
35
33
32
30
29
27
26
24
23
22
19
17

Measure of Central Tendency

Mode

- The mode is the value that occurs the most frequently in your data set.
- On a bar chart, the mode is the highest bar.
- If the data have multiple values that are tied for occurring the most frequently, you have a multimodal distribution.
- If no value repeats, the data do not have a mode.



Measure of Central Tendency

Mode continued....

- In the dataset, the value 5 occurs most frequently, which makes it the mode.
- These data might represent a 5-point Likert scale.

$$\text{Mode} = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times h$$

Measure of Central Tendency

Mode continued....

- Typically, you use the mode with categorical, ordinal, and discrete data.
- In fact, the mode is the only measure of central tendency that you can use with categorical data—such as the most preferred flavor of ice cream.
- However, with categorical data, there isn't a central value because you can't order the groups.

Mode
5
5
5
4
4
3
2
2
1

Measure of Central Tendency

Mode continued....

- With ordinal and discrete data, the mode can be a value that is not in the center.
- Again, the mode represents the most common value.

Mode
5
5
5
4
4
3
2
2
1

Arithmetic mean

Definition

- Arithmetic Mean is the most common and easily understood measure of central tendency.
- We can define mean as the value obtained by dividing the sum of measurements with the number of measurements contained in the data set and is denoted by the symbol \bar{x}

Arithmetic mean

Arithmetic Mean for three types of series

- Individual Data Series
- Discrete Data Series
- Continuous Data Series

Arithmetic mean

Individual Data Series

$$\bar{x} = \sum_{i=1}^n X_i$$

- When data is given on individual basis.
- Following is an example of individual series:

Items:

5 10 20 30 40 50 60 70

Arithmetic mean

Individual Data Series continued...

$$\bar{x} = \frac{\sum x}{N}$$

- Alternatively, we can write same formula as follows:

$$\bar{x} = \frac{\sum x}{N}$$

Arithmetic mean

Individual Data Series continued...

Where –

- $X_1, X_2, X_3, \dots, X_n$ = individual observation of variable.
- $\sum x$ = sum of all observations of the variable
- N = Number of observations

$$\bar{x} = \frac{\sum x}{N}$$

Arithmetic mean

Individual Data Series continued...

Example:

Problem Statement:

- Calculate Arithmetic Mean for the following individual data:

Items:

14 36 45 70 105

$$\bar{x} = \frac{\sum x}{N}$$

Arithmetic mean

Individual Data Series continued...

Solution:

- Based on the above mentioned formula, Arithmetic Mean \bar{x} will be:
- The Arithmetic Mean of the given numbers is 54.

$$\begin{aligned}\bar{x} &= \frac{14+36+45+70+105}{5} \\ &= \frac{270}{5} \\ &= 54\end{aligned}$$

Arithmetic mean

Discrete Data Series

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \dots + f_nx_n}{N}$$

- When data is given along with their frequencies. Following is an example of discrete series:

Items : 5 10 20 30 40 50 60 70

Frequency: 2 5 1 3 12 0 5 7

Arithmetic mean

Discrete Data Series continued...

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \dots + f_nx_n}{N}$$

- For discrete series, the Arithmetic Mean can be calculated using the following formula.

Formula

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \dots + f_nx_n}{N}$$

Arithmetic mean

Discrete Data Series continued...

$$\bar{x} = \frac{\sum fx}{\sum f}$$

- Alternatively, we can write same formula as follows:

Formula

$$\bar{x} = \frac{\sum fx}{\sum f}$$

Arithmetic mean

Continuous Data Series

$$\bar{x} = \frac{f_1m_1 + f_2m_2 + f_3m_3 + \dots + f_nm_n}{N}$$

- When data is given based on ranges along with their frequencies. Following is an example of continuous series:

Items: 0-5 5-10 10-20 20-30 30-40

Frequency: 2 5 1 3 12

Arithmetic mean

Continuous Data Series continued...

Where –

- N = Number of observations.
- $f_1, f_2, f_3, \dots, f_n$ = Different values of frequency f.
- $m_1, m_2, m_3, \dots, m_n$ = Different values of mid points for ranges.

$$\bar{x} = \frac{f_1m_1 + f_2m_2 + f_3m_3 + \dots + f_nm_n}{N}$$

Arithmetic mean

Geometric mean

- Geometric mean of n numbers is defined as the nth root of the product of n numbers.

Formula :

$$GM = \sqrt[n]{x_1 \times x_2 \times x_3 \dots x_n}$$

$$GM = \sqrt[n]{x_1 \times x_2 \times x_3 \dots x_n}$$

Arithmetic mean

Geometric mean continued...

Where –

- n = Total numbers.
- x_i = numbers.

Where –

- n = Total numbers.
- x_i = numbers.

Harmonic mean

What is Harmonic Mean ?

- Harmonic mean is a type of average that is calculated by dividing the number of values in a data series by the sum of the reciprocals ($1/x_i$) of each value in the data series.
- A harmonic mean is one of the three Pythagorean means (the other two are arithmetic mean and geometric mean).
- The harmonic mean always shows the lowest value among the Pythagorean means.

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}}$$

Harmonic mean

Formula of Harmonic Mean

- The general formula for calculating a harmonic mean is:
- Harmonic mean = $n / (\sum 1/x_i)$
- Where: n – the number of the values in a dataset
- x_i – the point in a dataset
- The weighted harmonic mean can be calculated using the following formula:
- Weighted Harmonic Mean = $(\sum w_i) /(\sum w_i/x_i)$ Where:
- w_i – the weight of the data point
- x_i – the point in a dataset.

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}}$$

Harmonic mean

Formula of Harmonic Mean

- The general formula for calculating a harmonic mean is:
- Harmonic mean = $n / (\sum 1/x_i)$
- Where: n – the number of the values in a dataset
- x_i – the point in a dataset
- The weighted harmonic mean can be calculated using the following formula:
- Weighted Harmonic Mean = $(\sum w_i) /(\sum w_i/x_i)$ Where:
- w_i – the weight of the data point
- x_i – the point in a dataset.

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}}$$

Harmonic mean

Example of Harmonic Mean

- Firstly, we need to find the P/E ratios of each company. Remember that the P/E ratio is essentially the market capitalization divided by the earnings.
- P/E (Company A) = (\$1 billion) / (\$20 million) = 50
- P/E (Company B) = (\$20 billion) / (\$5 billion) = 4

Example: Find the harmonic mean of the following data {8, 9, 6, 11, 10, 5} ?

Solution:

Given data: {8, 9, 6, 11, 10, 5}

$$\text{So Harmonic mean} = \frac{6}{\frac{1}{8} + \frac{1}{9} + \frac{1}{6} + \frac{1}{11} + \frac{1}{10} + \frac{1}{5}}$$

$$H = \frac{6}{0.7336} = 7.560$$

$$\text{Harmonic mean}(H) = 7.560$$

Harmonic mean

Example of Harmonic Mean

- We must use the weighted harmonic mean to calculate the P/E ratio of the index. Using the formula for the weighted harmonic mean, the P/E ratio of the index can be found in the following way:
$$\text{P/E (Index)} = (0.4+0.6) / (0.4/50 + 0.6/4) = 6.33$$
- Note that if we calculate the P/E ratio of the index using the weighted arithmetic mean, it would be significantly overstated:
$$\text{P/E (Index)} = 0.4 \times 50 + 0.6 \times 4 = 22.4$$

Example: Find the harmonic mean of the following data {8, 9, 6, 11, 10, 5} ?

Solution:

Given data: {8, 9, 6, 11, 10, 5}

$$\text{So Harmonic mean} = \frac{6}{\frac{1}{8} + \frac{1}{9} + \frac{1}{6} + \frac{1}{11} + \frac{1}{10} + \frac{1}{5}}$$

$$H = \frac{6}{0.7336} = 7.560$$

$$\text{Harmonic mean(H)} = 7.560$$

Median in Raw and Grouped Data

Median in Raw Data

- The median of raw data is the number which divides the observations when arranged in an order (ascending or descending) in two equal parts.

©math-only-math.com

Median of Raw Data

©math-only-math.com

Arrange the raw data in ascending or descending order.

If n = Number of variates in the data, then

Median = $\frac{n+1}{2}$ th variate, when n is odd.

Median = $\frac{1}{2} \left\{ \frac{n}{2} \text{th variate} + \left(\frac{n}{2} + 1 \right) \text{th variate} \right\}$,
when n is even.

Median in Raw and Grouped Data

Method of finding median

- Take the following steps to find the median of raw data.
- Step I: Arrange the raw data in ascending or descending order.
- Step II: Observe the number of variates in the data. Let the number of variates in the data be n . Then find the median as following.
 - (i) If n is odd then [Math Processing Error] th variate is the median

©math-only-math.com

Median of Raw Data

©math-only-math.com

Arrange the raw data in ascending or descending order.

If n = Number of variates in the data, then

Median = $\frac{n+1}{2}^{th}$ variate, when n is odd.

Median = $\frac{1}{2} \left\{ \frac{n}{2}^{th} \text{ variate} + \left(\frac{n}{2} + 1 \right)^{th} \text{ variate} \right\}$,
when n is even.

Median in Raw and Grouped Data

Method of finding median

- (ii) If n is even then the mean of $\left(\frac{n}{2}\right)$ th and $\left(\frac{n}{2} + 1\right)$ th variates is the median, i.e.,
- median = $\left(\frac{n}{2}\right)$.

©math-only-math.com

Median of Raw Data

©math-only-math.com

Arrange the raw data in ascending or descending order.

If n = Number of variates in the data, then

Median = $\frac{n+1}{2}$ th variate, when n is odd.

Median = $\frac{1}{2} \left\{ \frac{n}{2} \text{th variate} + \left(\frac{n}{2} + 1 \right) \text{th variate} \right\}$,
when n is even.

©math-only-math.com

©math-only-math.com

©math-only-math.com

©math-only-math.com

Median in Raw and Grouped Data

Solved Examples on Median of Raw Data

- Find the median of the ungrouped data.
15, 18, 10, 6, 14

- Solution:

Arranging variates in ascending order,
we get 6, 10, 14, 15, 18.

The number of variates = 5, which is odd.

- Therefore, median = [Math Processing Error]th variate = 3rd variate

@math-only-math.com

Median of Raw Data

@math-only-math.com

Arrange the raw data in ascending or descending order.

If n = Number of variates in the data, then

Median = $\frac{n+1}{2}$ th variate, when n is odd.

Median = $\frac{1}{2} \left\{ \frac{n}{2} \text{th variate} + \left(\frac{n}{2} + 1 \right) \text{th variate} \right\}$,

@math-only-math.com

@math-only-math.com

when n is even.

Median in Raw and Grouped Data

Finding Median for Grouped Data

- Median is the value which occupies the middle position when all the observations are arranged in an ascending or descending order. It is a positional average.
- (i) Construct the cumulative frequency distribution.
- (ii) Find $(N/2)$ th term
- (iii) The class that contains the cumulative frequency $N/2$ is called the median class.

Median in Raw and Grouped Data

Finding Median for Grouped Data

- (iv) Find the median by using the formula:
- Where l = Lower limit of the median class,
- f = Frequency of the median class
- c = Width of the median class,
- N = The total frequency (Σf)
- m = cumulative frequency of the class preceding the median class

$$\text{Median} = l + \frac{\left(\frac{N}{2} - m \right)}{f} \times c$$

Median in Raw and Grouped Data

Solved Examples on Median of Grouped Data

- A researcher studying the behavior of mice has recorded the time (in seconds) taken by each mouse to locate its food by considering 13 different mice as 31,33, 63, 33, 28, 29, 33, 27, 27, 34, 35,28, 32. Find the median time that mice spent in searching its food.
- 31, 33, 63, 33, 28, 29, 33, 27, 27, 34,35, 28, 32
- Ascending order of given data is 27, 27, 28, 28, 29, 31, 32, 33, 33, 33,34, 35, 63
- Middle value is 7th observation

Mode in Raw and Grouped Data

Finding the Mode in Raw Data

- To find the mode, or modal value, it is best to put the numbers in order. Then count how many of each number. A number that appears most often is the mode.
- 3, 7, 5, 13, 20, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29
- In order these numbers are:

Mode in Raw and Grouped Data

Finding the Mode in Raw Data

- 3, 5, 7, 12, 13, 14, 20, 23, 23, 23, 23, 29, 39, 40, 56
- This makes it easy to see which numbers appear most often.
- This makes it easy to see which numbers appear most often.
- In this case the mode is 23.

Mode in Raw and Grouped Data

Finding the Mode in Grouped Data

- In some cases (such as when all values appear the same number of times) the mode is not useful. But we can group the values to see if one group has more than the others.
- Example: {4, 7, 11, 16, 20, 22, 25, 26, 33}
- Each value occurs once, so let us try to group them.

Mode in Raw and Grouped Data

Finding the Mode in Grouped Data

- We can try groups of 10:
- 0-9: 2 values (4 and 7)
- 10-19: 2 values (11 and 16)
- 20-29: 4 values (20, 22, 25 and 26)
- 30-39: 1 value (33)
- In groups of 10, the "20s" appear most often, so we could choose 25 (the middle of the 20s group) as the mode.

Standard Deviation

Standard Deviation Formulas

- The Standard Deviation is a measure of how spread out numbers are.
- You might like to read this simpler page on Standard Deviation first.
- But here we explain the formulas.
- The symbol for Standard Deviation is σ (the Greek letter sigma).
- This is the formula for Standard Deviation:

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

Standard Deviation

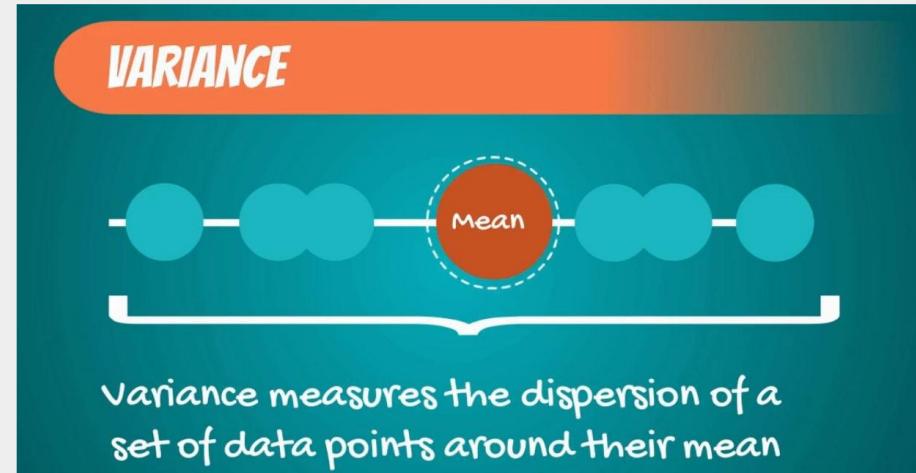
Steps for Standard Deviation

- Say we have a bunch of numbers like 9, 2, 5, 4, 12, 7, 8, 11.
- To calculate the standard deviation of those numbers:
- 1. Work out the Mean (the simple average of the numbers)
- 2. Then for each number: subtract then Mean and square the result
- 3. Then work out the mean of those squared differences.
- 4. Take the square root of that and we are done!

Variance

What is Variance?

- Variance is the expected value of the squared deviation of a random variable from its mean.
- In short, it is the measurement of the distance of a set of random numbers from their collective average value.
- Variance is used in statistics as a way of better understanding a data set's distribution.



Variance

How does Variance work?

- Variance is calculated by finding the square of the standard deviation of a variable, and the covariance of the variable with itself.
- In the formula above, u represents the mean of the data points, x is the value of an individual data point, and N is the total number of data points.

$$\sigma^2 = \frac{\sum(\chi - \mu)^2}{N}$$

Variance

How to Calculate Variance? (Continued)

- Steps to Calculate Variance:
 1. List elements of data set. The following are ages of students pursuing a Master's degree:
 2. Data set 1: 28,25,26,27,31,32,24
- Calculate the mean.
- $(28 + 25 + 26 + 27 + 31 + 32 + 24) / 7 = 27.57$

Mean Formula

$$\text{Mean} = \frac{\text{Sum of All Data Points}}{\text{Number of Data Points}}$$

$$\text{Mean} = \text{Assumed Mean} + \frac{\text{Sum of All Deviations}}{\text{Number of Data Points}}$$

Variance

How to Calculate Variance? (Continued)

- Find the deviation from the mean for each data point.

Given Data Point(x)	Difference between given data point(Age) and Mean ($X - \bar{X}$)
28	$(28 - 27.57) = 0.43$
25	$(25 - 27.57) = -2.57$
26	$(26 - 27.57) = -1.57$
27	$(27 - 27.57) = -0.57$
31	$(31 - 27.57) = 3.43$
32	$(32 - 27.57) = 4.43$
24	$(24 - 27.57) = -3.53$

Variance

How to Calculate Variance? (Continued)

- Square it

$$(X - \bar{X})^2$$

$$(0.43)^2 = 0.1849$$

$$(-2.57)^2 = 6.6049$$

$$(-1.57)^2 = 2.4649$$

$$(-0.57)^2 = .3249$$

$$(3.43)^2 = 11.7649$$

$$(4.43)^2 = 19.6249$$

$$(-3.53)^2 = 12.4609$$

Variance

How to Calculate Variance? (Continued)

- The average of all squared differences is the variance. To find it, add all squared variances and divide the sum by a number of elements in data set (n).
- To find the standard deviation in ages of students pursuing Master's, we calculate the square root of the variance

$$\Rightarrow (0.1849 + 6.6049 + 2.4649 + .3249 + 11.76 + 19.6249 + 12.4609) / 7$$

$$\Rightarrow 53.4303 / 7 = 7.6329$$

$$\Rightarrow \text{Variance} = 7.6329$$

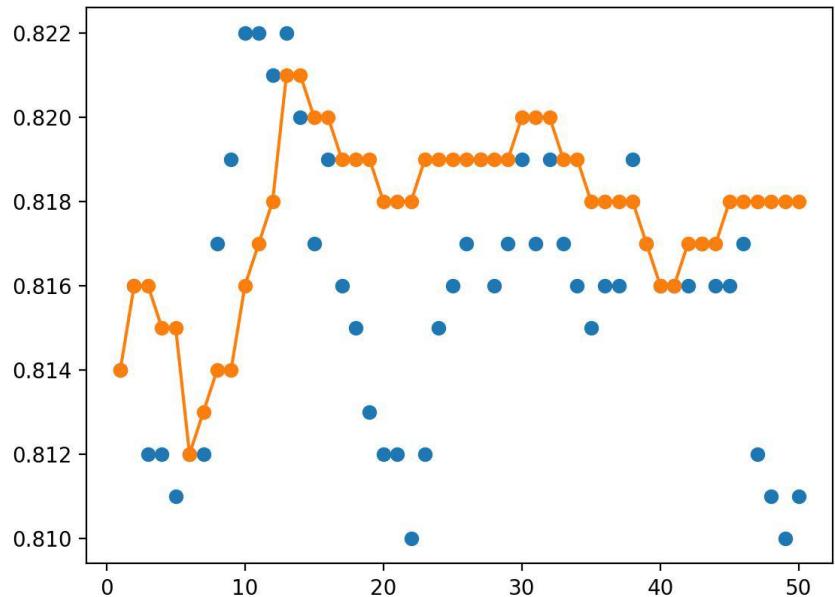
$\Rightarrow \text{Standard Deviation} = \sqrt{\text{Variance}}$

$$\sqrt{7.6329} = 2.7627$$

Variance

Applications of Variance

- Variance plays a major role in interpreting data in statistics.
- The most common application of variance is in polls.
- For opinion polls, the data gathering agencies cannot invest in collecting data from the entire population.
- They set criteria for sampling the population based on ethnicity, income group, regions, education level, salary and religion, so that the population is completely represented by the samples.



Properties of Variance and standard deviation

Properties of Variance

- Variance is a numerical value that describes the variability of observations from its arithmetic mean.
- Variance is nothing but an average of squared deviations.
- Variance is denoted by sigma-squared (σ^2)
- Variance is expressed in square units which are usually larger than the values in the given dataset.
- Variance measures how far individuals.

For unclassified data:

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

For grouped frequency distribution:

$$\sigma^2 = \frac{\sum f_i(x_i - \mu)^2}{N}$$

Properties of Variance and standard deviation

Properties of Variance

- In statistics variance is defined as the measure of variability that represents how far members of a group are spread out.
- It finds out the average degree to which each observation varies from the mean.
- When the variance of a data set is small, it shows the closeness of the data points to the mean whereas a greater value of variance represents that the observations are very dispersed around

For unclassified data:

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

For grouped frequency distribution:

$$\sigma^2 = \frac{\sum f_i(x_i - \mu)^2}{N}$$

Median in Raw and Grouped Data

Finding Median for Grouped Data

- Median is the value which occupies the middle position when all the observations are arranged in an ascending or descending order. It is a positional average.
- (i) Construct the cumulative frequency distribution.
- (ii) Find $(N/2)$ th term
- (iii) The class that contains the cumulative frequency $N/2$ is called the median class.

Properties of Variance and standard deviation

Properties of Standard Deviation

- Standard deviation is a measure that quantifies the amount of dispersion of the observations in a dataset.
- The low standard deviation is an indicator of the closeness of the scores to the arithmetic mean and a high standard deviation represents.
- The scores are dispersed over a higher range of values.

For unclassified data:

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

For grouped frequency distribution:

$$\sigma = \sqrt{\frac{\sum f_i(x_i - \mu)^2}{N}}$$

Properties of Variance and standard deviation

Properties of Standard Deviation

- Standard deviation is a measure of the dispersion of observations within a data set relative to their mean.
- The standard deviation is the root mean square deviation.
- standard deviation is labelled as sigma (σ).
- standard deviation which is expressed in the same units as the values in the set of data.
- Standard Deviation measures how much

For unclassified data:

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

For grouped frequency distribution:

$$\sigma = \sqrt{\frac{\sum f_i(x_i - \mu)^2}{N}}$$

Properties of Variance and standard deviation

Example : To find Standard Deviation and Variance

- Marks scored by a student in five subjects are 60, 75, 46, 58 and 80 respectively.
- You have to find out the standard deviation and variance.
- First of all, you have to find out the mean,

$$\text{Mean} = \frac{60+75+46+58+80}{5} = 63.8$$

So the average (mean) marks are 63.8

Properties of Variance and standard deviation

Example : To find Standard Deviation and Variance

- Now calculate the variance
- Where, X = Observations
- A = Arithmetic Mean
- Both variance and standard deviation are always positive.
- If all the observations in a data set are identical, then the standard deviation and variance will be zero.

X	A	(X-A)	(X-A) ²
60	63.8	-3.8	14.44
75	63.8	11.2	125.44
46	63.8	-17.8	316.84
58	63.8	5.8	33.64
80	63.8	16.2	262.44

$$\text{Variance} = \frac{14.44 + 125.44 + 316.84 + 33.64 + 262.44}{5} = 150.56$$

So the variance is 150.56

And Standard deviation is –

$$\sigma = \sqrt{150.56} = 12.27$$

Properties of Variance and standard deviation

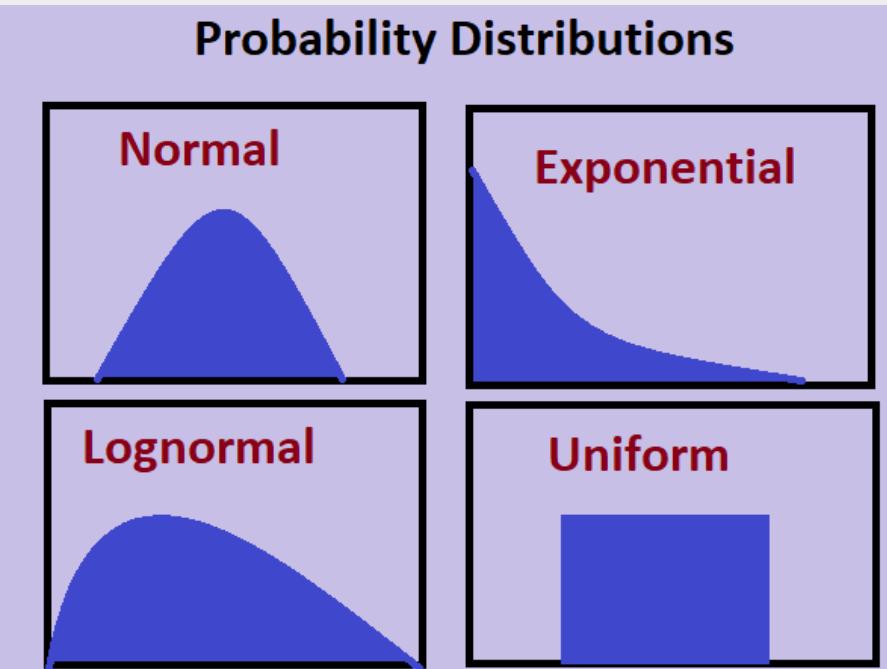
Basis of comparison	Variance	Standard Deviation
Basic Definition	Variance can be defined as the numerical value which is nothing but the variability of all the observations from its arithmetic average or the arithmetic mean.	The standard deviation can be defined as the measure of the dispersion of the numerical values in a given set of data from their average or the mean.
Symbol / Label (in general)	Variance can be labeled as Sigma ² (σ^2)	The standard deviation can be labeled as Sigma (σ)
Usefulness	Variance can help in determining the size of the data spread.	If one wants to measure the absolute measure of the variability of dispersion, then the standard deviation is the right choice.
What does it indicate?	How far are the individuals or the observations in a group that are spread out?	How many observations or the individuals of a data set differ from its average or the mean.
Units expressed in	Variance is always expressed in squared units.	The unit of standard deviation is the same as the observations.
Mostly used?	While deciding upon the asset allocation before investing any of the funds.	Standard deviation can be used to measure the stock market or the stock's variability either on a daily, weekly, or monthly basis.
Calculation Methodology	Variance can be calculated by taking the average or the mean of the squared deviation of each of the observations in the data set from its arithmetic mean or average.	One just needs to take a square root of the variance.



Types of Distributions

Data Distribution

- Data distribution is a function that determines the values of a variable and quantifies relative frequency, it transforms raw data into graphical methods to give valuable information.



Types of Distributions

Uniform Distribution

- Uniform distribution can either be discrete or continuous where each event is equally likely to occur. It has a constant probability constructing a rectangular distribution.
- A variable X is said to be uniformly distributed if the density function is:

$$f(x) = \frac{1}{b-a}$$

for $-\infty < a \leq x \leq b < \infty$

The graph of a uniform distribution curve looks like



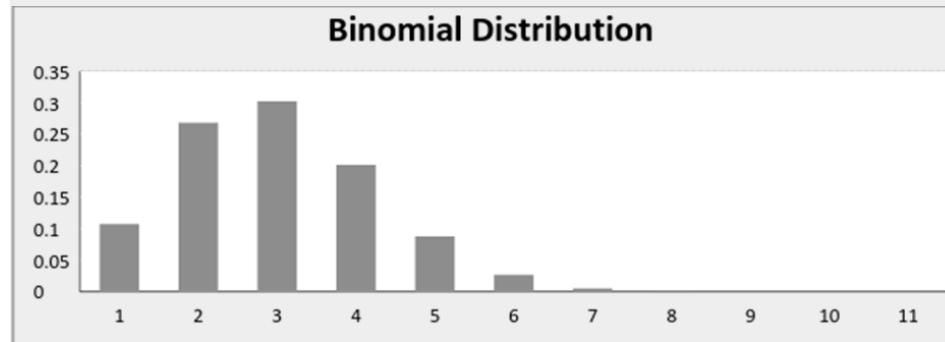
Types of Distributions

Binomial Distribution

- A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose and where the probability of success and failure is same for all the trials is called a Binomial Distribution.
- The mathematical representation of binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

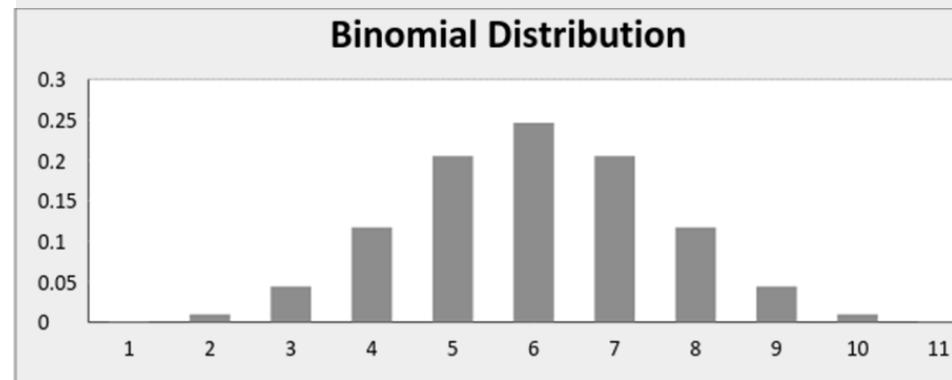
A binomial distribution graph where the probability of success does not equal the probability of failure looks like



Types of Distributions

Binomial Distribution continued..

- When probability of success = probability of failure, in such a situation the graph of binomial distribution looks like



Types of Distributions

Normal Distribution

- Being a continuous distribution, the normal distribution is most commonly used in data science.
- A very common process of our day to day life belongs to this distribution-income distribution, average employees report, average weight of a population, etc.

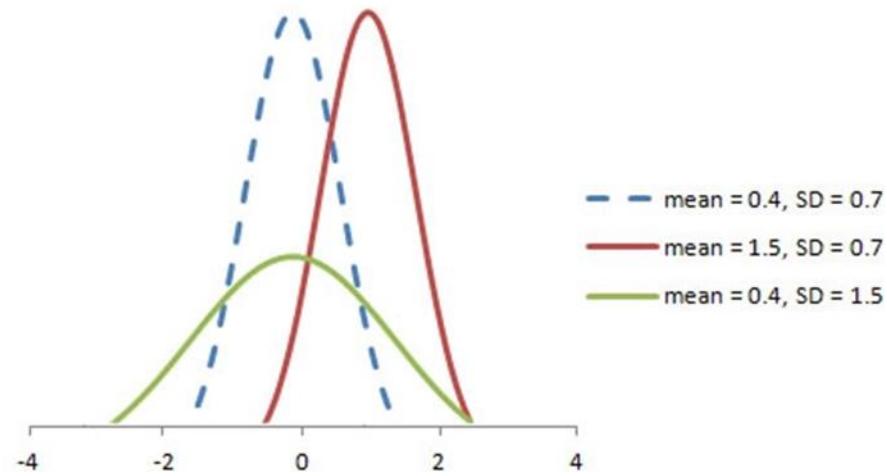
Types of Distributions

Normal Distribution

- The PDF of a random variable X following a normal distribution is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad \text{for } -\infty < x < \infty.$$

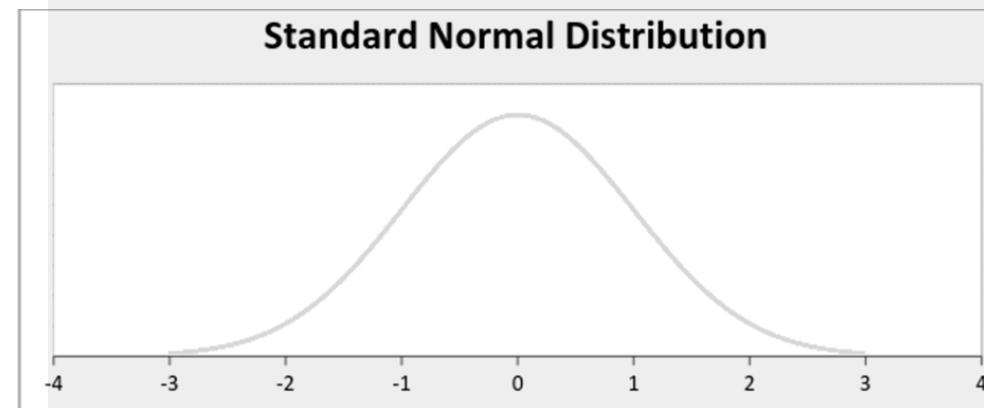
The graph of a random variable $X \sim N(\mu, \sigma)$ is shown below.



Types of Distributions

Normal Distribution

- A standard normal distribution is defined as the distribution with mean 0 and standard deviation

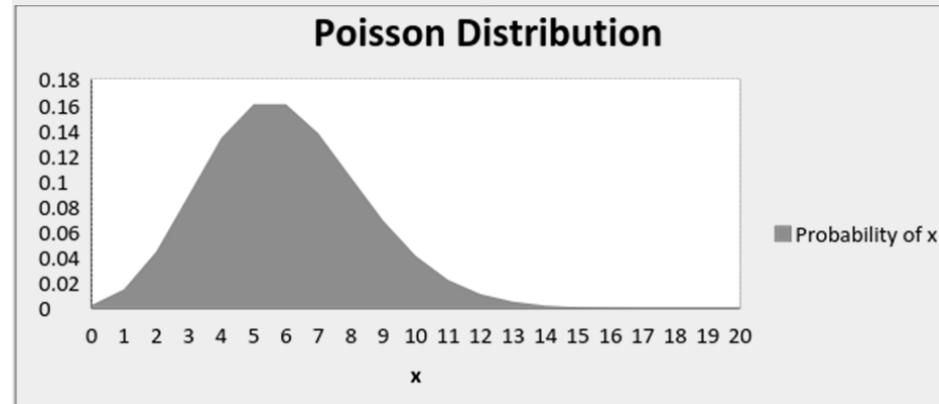


Types of Distributions

Poisson Distribution

- A distribution is called Poisson distribution when the following assumptions are valid:
- Any successful event should not influence the outcome of another successful event.
- The probability of success over a short interval must equal the probability of success over a longer interval.
- The probability of success in an interval approaches zero as the interval becomes smaller.

The graph of a Poisson distribution is shown below:



Types of Distributions

Poisson Distribution

- Some notations used in Poisson distribution are:
- λ is the rate at which an event occurs,
- t is the length of a time interval,
- And X is the number of events in that time interval.
- Here, X is called a Poisson Random Variable and the probability distribution of X is called Poisson distribution.
- Let μ denote the mean number of events in an interval of length t . Then, $\mu = \lambda*t$.
- The PMF of X following a Poisson distribution is given by:

$$P(X = x) = e^{-\mu} \frac{\mu^x}{x!}$$

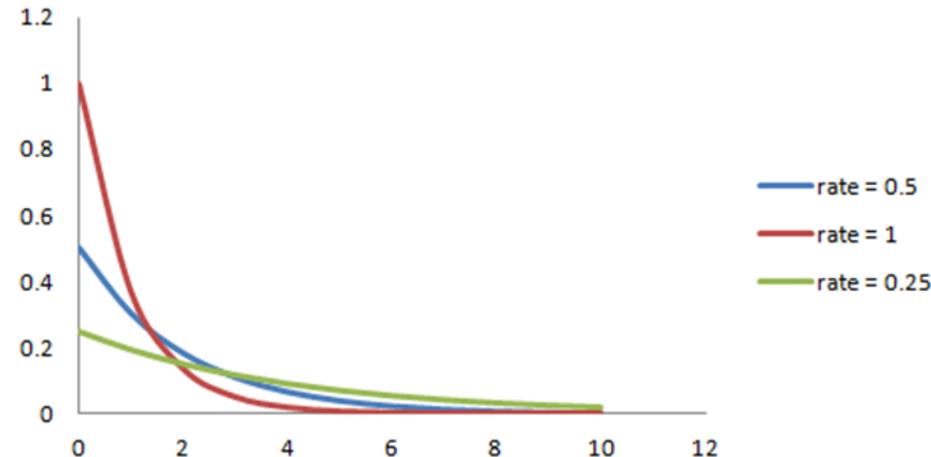
for $x = 0, 1, 2, \dots$

Types of Distributions

Exponential Distribution

- Like the Poisson distribution, exponential distribution has the time element; it gives the probability of a time duration before an event takes place.
- A random variable X is said to have an exponential distribution with PDF:
- $f(x) = \{ \lambda e^{-\lambda x}, x \geq 0 \}$

Exponential Distribution



Sampling

What is sampling?

- Sampling is a technique of selecting individual members or a subset of the population to make statistical inferences from them and estimate characteristics of the whole population.

Sampling

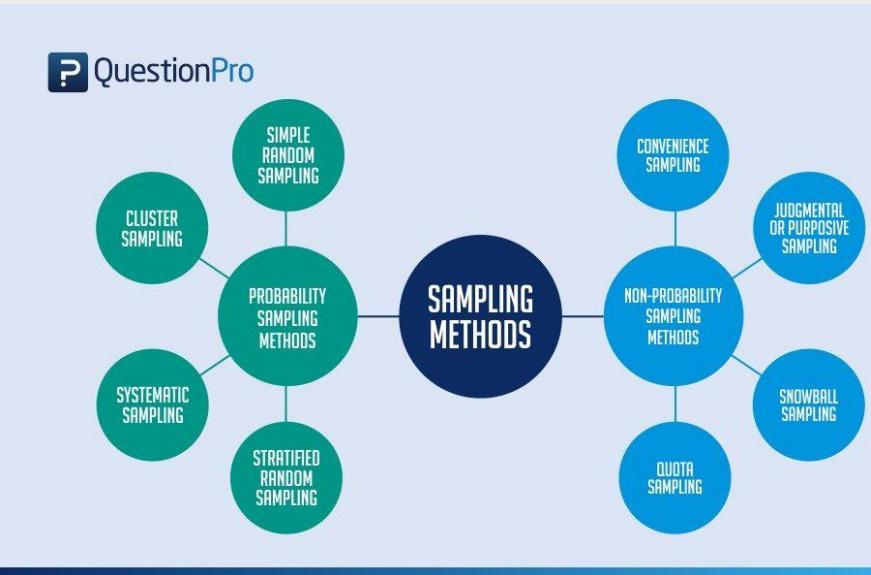
Sampling Example

- If a drug manufacturer would like to research the adverse side effects of a drug on the country's population, it is almost impossible to conduct a research study that involves everyone.
- In this case, the researcher decides a sample of people from each demographic and then researches them, giving him/her indicative feedback on the drug's behavior.

Sampling

Types of sampling: sampling methods

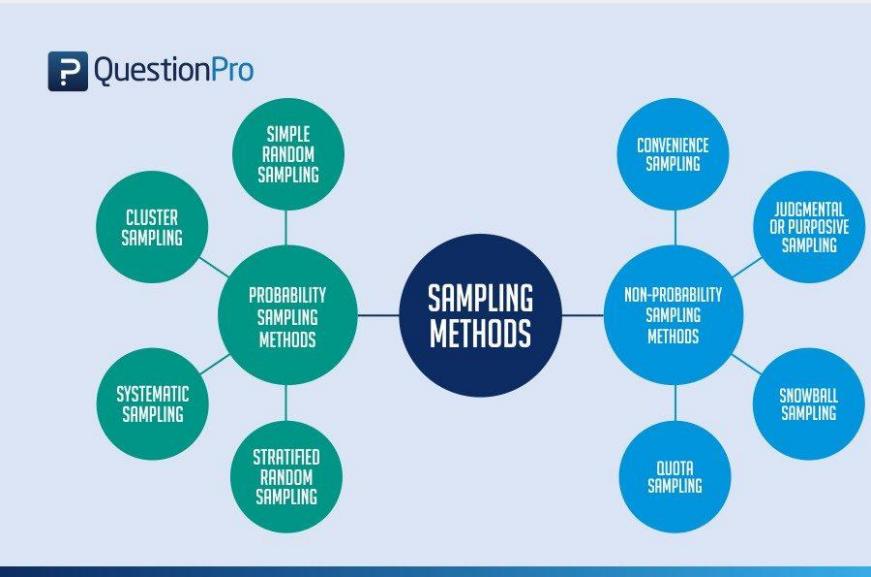
- **Probability sampling:** Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly.
- All the members have an equal opportunity to be a part of the sample with this selection parameter.



Sampling

Types of sampling: sampling methods

- **Non-probability sampling:** In non-probability sampling, the researcher chooses members for research at random.
- This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.



Sampling

Types of probability sampling

- Probability sampling is a sampling technique in which researchers choose samples from a larger population using a method based on the theory of probability.
- This sampling method considers every member of the population and forms samples based on a fixed process.

Types of probability sampling

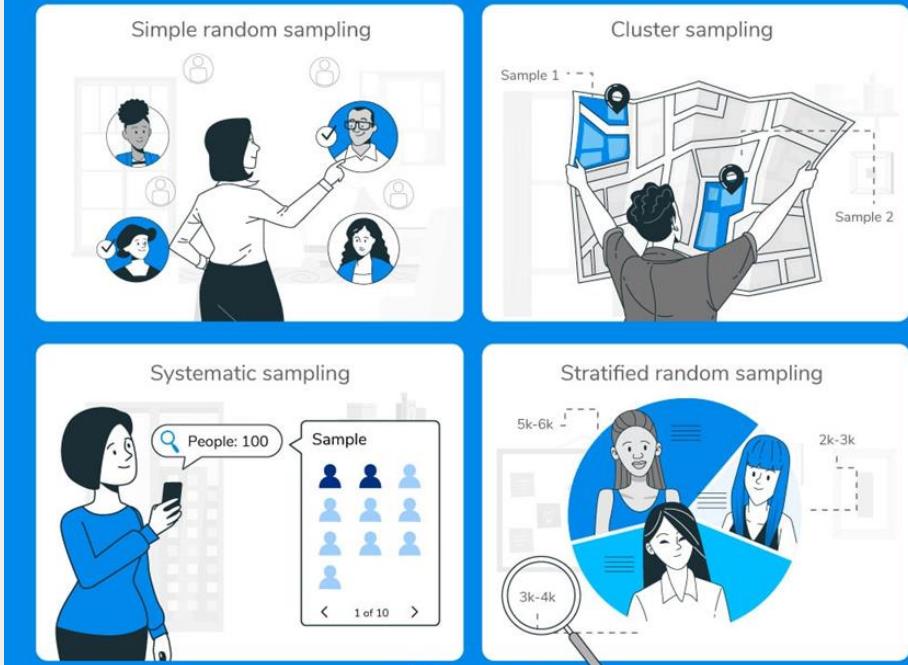


Sampling

Types of probability sampling

- **Simple random sampling:** It is a reliable method of obtaining information where every single member of a population is chosen randomly, merely by chance.
- **Cluster sampling:** It sampling is a method where the researchers divide the entire population into sections or clusters that represent a population.

Types of probability sampling



Sampling

Types of probability sampling

- **Systematic sampling:** This method to choose the sample members of a population at regular intervals. It requires the selection of a starting point for the sample and sample size that can be repeated at regular intervals.
- **Stratified random sampling:** It is a method in which the researcher divides the population into smaller groups that don't overlap but represent the entire population.

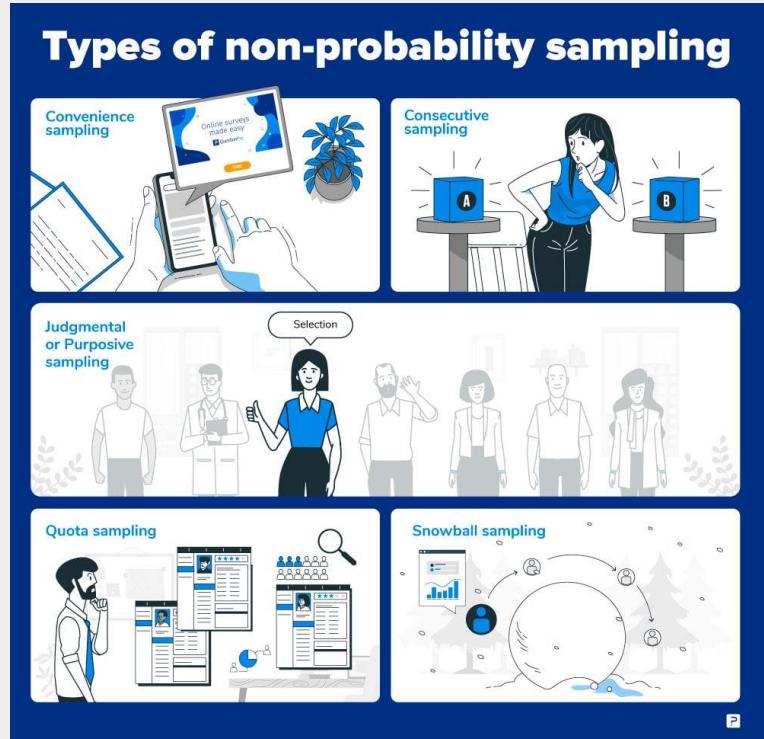
Types of probability sampling



Sampling

Types of non-probability sampling

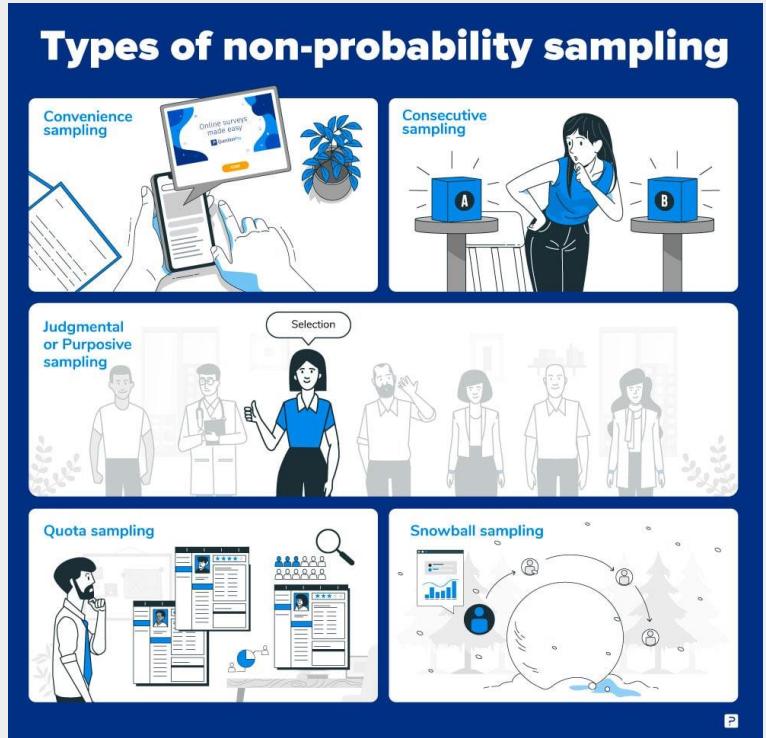
- **Convenience sampling:** This method is dependent on the ease of access to subjects such as surveying customers at a mall or passers-by on a busy street.
- **Judgmental or purposive sampling:** Judgmental or purposive samples are formed by the discretion of the researcher. Researchers purely consider the purpose of the study, along with the understanding of the target audience.



Sampling

Types of non-probability sampling

- **Snowball sampling:** It is a sampling method that researchers apply when the subjects are difficult to trace.
- **Quota sampling:** In Quota sampling, the selection of members in this sampling technique happens based on a pre-set standard. .



Business Analytics

(17 hours)

In this section, we will discuss:

- Probability Theories
- Bayes' Theorem
- Maximum Likelihood
- Hypothesis Testing
- Central limit theorem
- Chi-square test

Probability Theories

What Is Probability Theory?

- Probability theory is a branch of mathematics focusing on the analysis of random phenomena. It is an important skill for data scientists using data affected by chance.

Probability Theory



The probability of getting number "3" with one throw?
 $\frac{1}{6}$



The probability of getting number "3" with double throw?

$$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$


Probability Theories

Practical Uses for Probability Theory

- Data scientists to model situations
- Business world
- Business world
- Clinical trials

Probability Theory



The probability of getting number "3" with one throw?

$$\frac{1}{6}$$

The probability of getting number "3" with double throw?

$$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$


Image Source:
<https://i.pinimg.com/originals/c7/44/33/c74433effae9a094ff059a6d016d547b.jpg>

Probability Theories

Types of Probability?

- Classical
- Relative Frequency
- Subjective Probability

Types of Probability

Classical. (Also referred to as **Theoretical**). The number of outcomes in the sample space is known, and each outcome is equally likely to occur.

Empirical. (Also referred to as **Statistical** or **Relative Frequency**). The frequency of outcomes is measured by experimenting.

Subjective. You estimate the probability by making an “educated guess”, or by using your intuition.

Probability Theories

Probability Theory Examples

Probability theory is a tool employed by researchers, businesses, investment analysts and countless others for risk management and scenario analysis.

- Epidemiology
- Insurance
- Small Business
- Meteorology

Probability Theory



The probability of getting number "3" with one throw?

$$\frac{1}{6}$$

The probability of getting number "3" with double throw?

$$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

Image Source:
<https://i.pinimg.com/originals/c7/44/33/c74433effae9a094ff059a6d016d547b.jpg>

Probability Theories

Advantages and Disadvantages of Probability Theory

- Classical
- Relative Frequency
- Subjective

Types of Probability

Classical. (Also referred to as **Theoretical**). The number of outcomes in the sample space is known, and each outcome is equally likely to occur.

Empirical. (Also referred to as **Statistical** or **Relative Frequency**). The frequency of outcomes is measured by experimenting.

Subjective. You estimate the probability by making an “educated guess”, or by using your intuition.

Probability Theories

What is the probability formula?

$$P(A) = \frac{\text{Number of favorable outcomes to A}}{\text{Total number of possible outcomes}}$$

- $P(A)$ = favorable outcomes/total outcomes

Image Source:

https://d138zd1kt9iqe.cloudfront.net/media/seo_landing_files/probability-formula-1-1634878086.png

Probability Theories

How Data Scientists Use Probability Theory

- Probability allows data scientists to assess the certainty of outcomes of a particular study or experiment.
- Today's data scientists need to have an understanding of the foundational concepts of probability theory including key concepts involving probability distribution, statistical significance, hypothesis testing and regression.

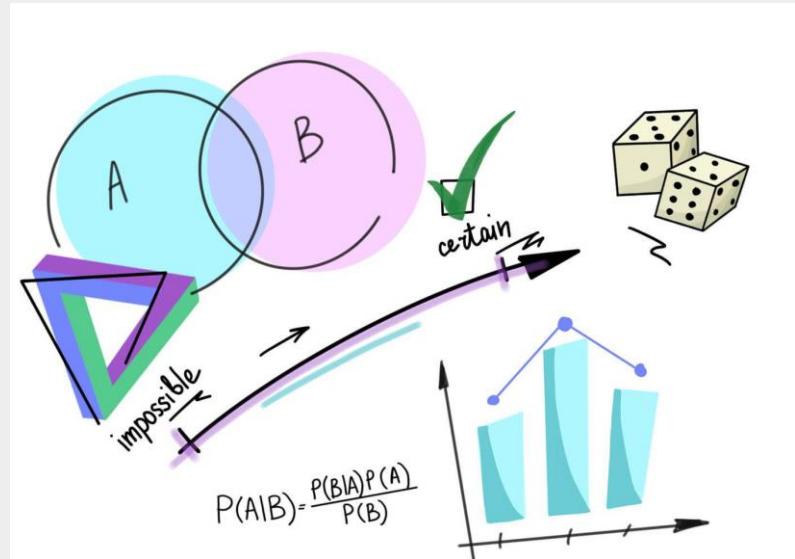


Image Source: https://luminousmen.com/media/data-science-probability_2.jpg

Bayes' Theorem

What is the Bayes' Theorem?

- The Bayes' theorem is a mathematical formula used to determine the conditional probability of events. Essentially, the Bayes' theorem describes the probability of an event based on prior knowledge of the conditions that might be relevant to the event.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Image Source: : <https://cdn.corporatefinanceinstitute.com/assets/bayes-theorem.png>

Bayes' Theorem

Formula for Bayes' Theorem

Where:

- $P(A|B)$ – the probability of event A occurring, given event B has occurred
- $P(B|A)$ – the probability of event B occurring, given event A has occurred
- $P(A)$ – the probability of event A
- $P(B)$ – the probability of event B

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Theorem

Formula for Bayes' Theorem

A special case of the Bayes' theorem is when event A is a binary variable. In such a case, the theorem is expressed in the following way

Where:

- $P(B|A^-)$ – the probability of event B occurring given that event A^- has occurred
- $P(B|A^+)$ – the probability of event B occurring given that event A^+ has occurred

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A^-)P(A^-) + P(B|A^+)P(A^+)}$$

Bayes' Theorem

Example of Bayes' Theorem

$$P(A|B) = \frac{0.60 \times 0.04}{0.60 \times 0.04 + 0.35 \times (1 - 0.04)} = 0.067 \text{ or } 6.67\%$$

- Using the Bayes' theorem, we can find the required probability:

Maximum likelihood

What is Maximum likelihood

$$L(\theta) = \prod_{i=1}^n f(x_i|\theta)$$

- Maximum likelihood is a widely used technique for estimation with applications in many areas including time series modeling, panel data, discrete data, and even machine learning.

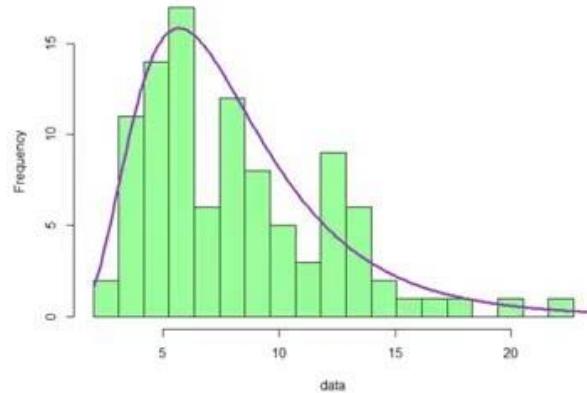
Image Source: https://4.bp.blogspot.com/-FewCGHC23oU/WliHLS_APPI/AAAAAAAABv8/gvldpT8aG4VEPVu6DnqhIXoSABg0F_YgCLcBGAs/s1600/Maximum-Likelihood-Estimation-in-Machine-Learning.jpg

Maximum likelihood

What is Maximum Likelihood Estimation?

- Maximum likelihood estimation is a statistical method for estimating the parameters of a model. In maximum likelihood estimation, the parameters are chosen to maximize the likelihood that the assumed model results in the observed data.

Maximum Likelihood Estimation



EART125: Statistics and Data Analysis in the Geosciences UC Santa Cruz

Image Source: https://i.ytimg.com/vi/2vh98ful3_M/hqdefault.jpg

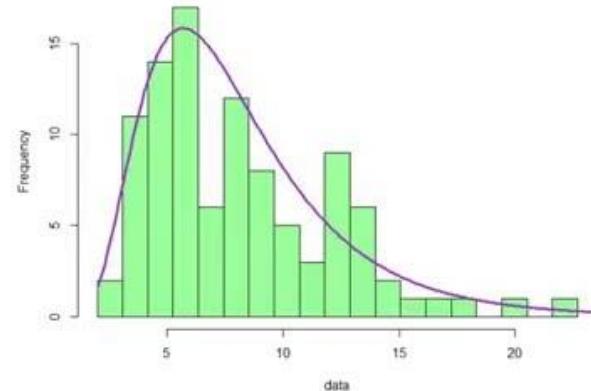
Maximum likelihood

What is Maximum Likelihood Estimation?

In order to implement maximum likelihood estimation we must:

- Assume a model, also known as a data generating process, for our data.
- Be able to derive the likelihood function for our data, given our assumed model (we will discuss this more later).

Maximum Likelihood Estimation



EART125: Statistics and Data Analysis in the Geosciences UC Santa Cruz

Image Source: https://i.ytimg.com/vi/2vh98ful3_M/hqdefault.jpg

Maximum likelihood

Advantages of Maximum Likelihood Estimation

- If the model is correctly assumed, the maximum likelihood estimator is the most efficient estimator.
- It provides a consistent but flexible approach which makes it suitable for a wide variety of applications, including cases where assumptions of other models are violated.
- It results in unbiased estimates in larger samples.

Maximum Likelihood Estimation (2)

- Advantages:
 1. they become unbiased minimum variance estimators as the sample size increases
 2. they have approximate normal distributions and approximate sample variances that can be calculated and used to generate confidence bounds
 3. likelihood functions can be used to test hypotheses about models and parameters
- Disadvantages:
 1. With small numbers of failures (less than 5, and sometimes less than 10 is small), MLE's can be heavily biased and the large sample optimality properties do not apply
 2. Calculating MLE's often requires specialized software for solving complex non-linear equations.

Maximum likelihood

Disadvantages of Maximum Likelihood Estimation

- It relies on the assumption of a model and the derivation of the likelihood function which is not always easy.
- Like other optimization problems, maximum likelihood estimation can be sensitive to the choice of starting values.

Maximum Likelihood Estimation (2)

- Advantages:
 1. they become unbiased minimum variance estimators as the sample size increases
 2. they have approximate normal distributions and approximate sample variances that can be calculated and used to generate confidence bounds
 3. likelihood functions can be used to test hypotheses about models and parameters
- Disadvantages:
 1. With small numbers of failures (less than 5, and sometimes less than 10 is small), MLE's can be heavily biased and the large sample optimality properties do not apply
 2. Calculating MLE's often requires specialized software for solving complex non-linear equations.

Maximum likelihood

Disadvantages of Maximum Likelihood Estimation

- Depending on the complexity of the likelihood function, the numerical estimation can be computationally expensive.
- Estimates can be biased in small samples.

Maximum Likelihood Estimation (2)

- Advantages:
 1. they become unbiased minimum variance estimators as the sample size increases
 2. they have approximate normal distributions and approximate sample variances that can be calculated and used to generate confidence bounds
 3. likelihood functions can be used to test hypotheses about models and parameters
- Disadvantages:
 1. With small numbers of failures (less than 5, and sometimes less than 10 is small), MLE's can be heavily biased and the large sample optimality properties do not apply
 2. Calculating MLE's often requires specialized software for solving complex non-linear equations.

Maximum likelihood

What is the Likelihood Function?

- Maximum likelihood estimation hinges on the derivation of the likelihood function. For this reason, it is important to have a good understanding of what the likelihood function is and where it comes from.
- Let's start with the very simple case where we have one series with 10 independent observations: 5, 0, 1, 1, 0, 3, 2, 3, 4, 1.

The Likelihood Function

- ◆ How good is a particular θ ? It depends on how likely it is to generate the observed data

$$L(\theta : D) = P(D | \theta) = \prod_m P(x[m] | \theta)$$

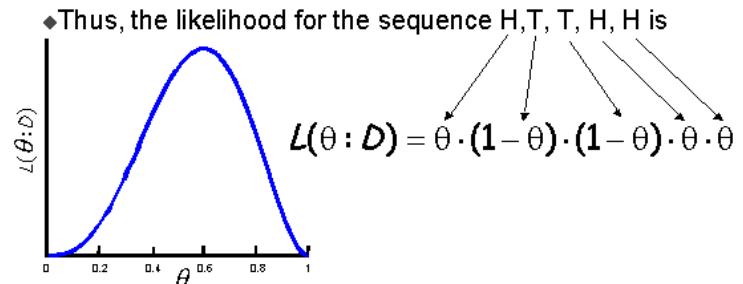


Image Source: <http://ai.stanford.edu/~moises/tutorial/img032.GIF>

What is the Likelihood Function?

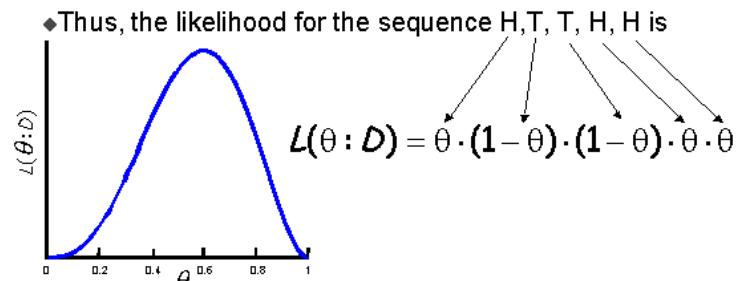
The Probability Density

- The first step in maximum likelihood estimation is to assume a probability distribution for the data. A probability density function measures the probability of observing the data given a set of underlying model parameters.

The Likelihood Function

- ◆ How good is a particular θ ?
It depends on how likely it is to generate the observed data

$$L(\theta : D) = P(D | \theta) = \prod_m P(x[m] | \theta)$$



What is the Likelihood Function?

The Probability Density

- The Poisson probability density function for an individual observation, is given by

$$f(y_i|\theta) = \frac{e^{-\theta}\theta^{y_i}}{y_i!}$$

- Because the observations in our sample are independent, the probability density of our observed sample can be found by taking the product of the probability of the individual observations:

$$f(y_1, y_2, \dots, y_{10}|\theta) = \prod_{i=1}^{10} \frac{e^{-\theta}\theta^{y_i}}{y_i!} = \frac{e^{-10\theta}\theta^{\sum_{i=1}^{10} y_i}}{\prod_{i=1}^{10} y_i!}$$

The Likelihood Function

- How good is a particular θ ? It depends on how likely it is to generate the observed data

$$L(\theta : D) = P(D | \theta) = \prod_m P(x[m] | \theta)$$

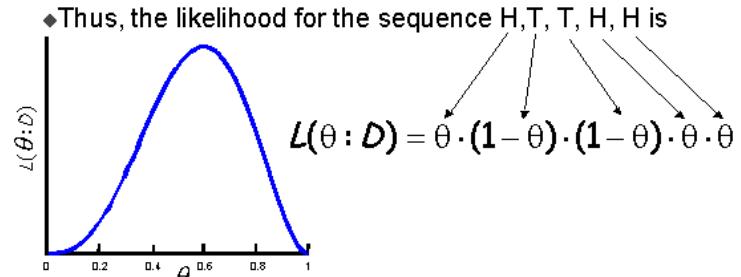


Image Source: <http://ai.stanford.edu/~moises/tutorial/img032.GIF>

What is the Likelihood Function?

The Likelihood Function

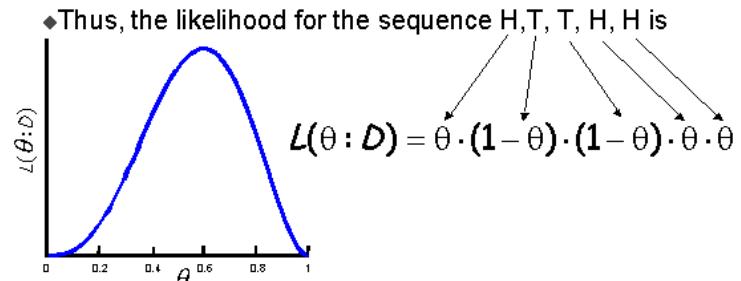
The differences between the likelihood function and the probability density function are nuanced but important.

- A probability density function expresses the probability of observing our data given the underlying distribution parameters. It assumes that the parameters are known.

The Likelihood Function

- ◆ How good is a particular θ ?
It depends on how likely it is to generate the observed data

$$L(\theta : D) = P(D | \theta) = \prod_m P(x[m] | \theta)$$



What is the Likelihood Function?

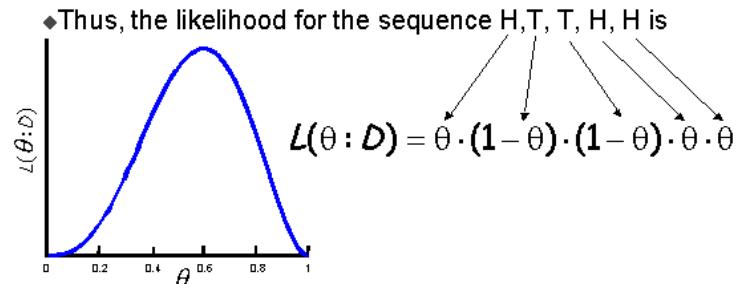
The Likelihood Function

- The likelihood function expresses the likelihood of parameter values occurring given the observed data. It assumes that the parameters are unknown.

The Likelihood Function

- ◆ How good is a particular θ ?
It depends on how likely it is to generate the observed data

$$L(\theta : D) = P(D | \theta) = \prod_m P(x[m] | \theta)$$



What is the Likelihood Function?

The Likelihood Function

- Mathematically the likelihood function looks similar to the probability density:

$$L(\theta | y_1, y_2, \dots, y_{10}) = f(y_1, y_2, \dots, y_{10} | \theta)$$

- For our Poisson example, we can fairly easily derive the likelihood function

$$L(\theta | y_1, y_2, \dots, y_{10}) = \frac{e^{-10\theta} \theta^{\sum_{i=1}^{10} y_i}}{\prod_{i=1}^{10} y_i!} = \frac{e^{-10\theta} \theta^{20}}{207,360}$$

The Likelihood Function

- How good is a particular θ ?
It depends on how likely it is to generate the observed data

$$L(\theta : D) = P(D | \theta) = \prod_m P(x[m] | \theta)$$

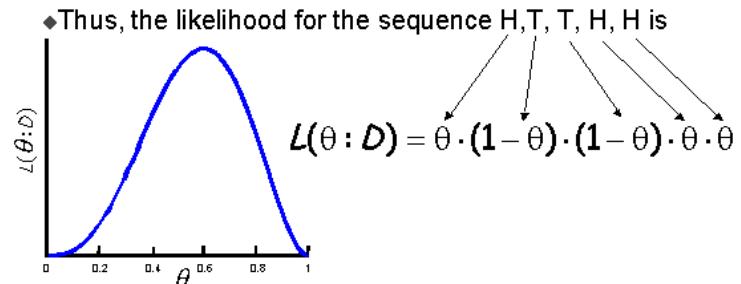


Image Source: <http://ai.stanford.edu/~moises/tutorial/img032.GIF>

What is the Likelihood Function?

The Log-Likelihood Function

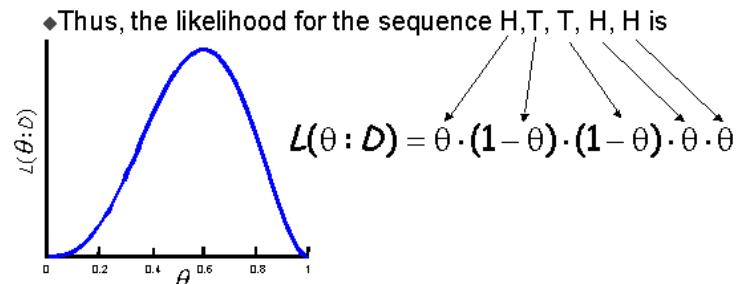
- In practice, the joint distribution function can be difficult to work with and the of the likelihood function is used instead. In the case of our Poisson dataset the log-likelihood function is:

$$\ln(L(\theta|y)) = -n\theta + \ln \sum_{i=1}^n y_i - \ln \theta \sum_{i=1}^n y_i! = -10\theta + 20 \ln(\theta) - \ln(207,360)$$

The Likelihood Function

- How good is a particular θ ?
It depends on how likely it is to generate the observed data

$$L(\theta : D) = P(D | \theta) = \prod_m P(x[m] | \theta)$$



What is the Likelihood Function?

The Maximum Likelihood Estimator

- A graph of the likelihood and log-likelihood for our dataset shows that the maximum likelihood occurs when $\theta= 2$. This means that our maximum likelihood estimator, $\hat{\theta}_{MLE} = 2$.

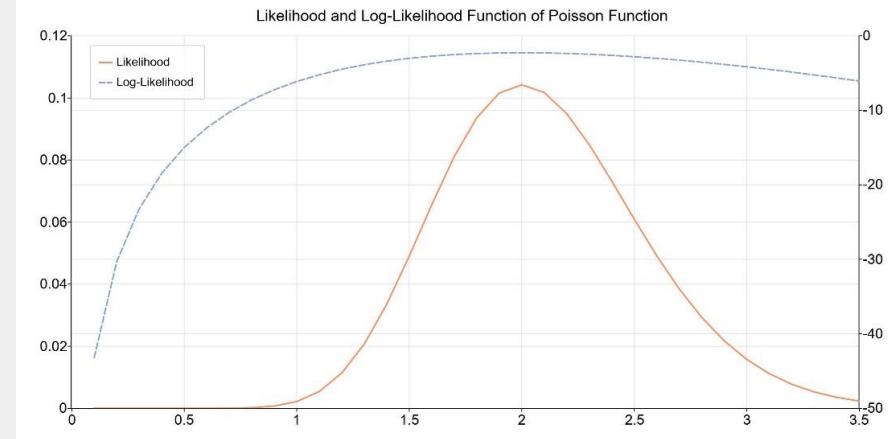


Image Source: <https://www.aptech.com/wp-content/uploads/2020/09/poisson-likelihood-function.jpeg>

What is the Likelihood Function?

Maximum Likelihood Estimation (MLE) vs Bayesian Estimation.

Maximum Likelihood Estimation (MLE) vs Bayesian Estimation.

	MLE	Bayesian Estimation
Predictions	We make predictions utilizing the latent variables in the density function to compute a probability.	We make predictions using the posterior distribution and the parameters which are considered as the random variables.
Situations to working with	Data with minimal values and the knowledge of prior is low. We can use MLE.	Data with sparse value and knowledge about the reliability of priors is high. We can use Bayesian estimation.
Complexity	MLE is less complex because we require to compute only the likelihood function	Bayesian estimation is more complex because the computation requires the likelihood function, evidence, and prior.

Hypothesis Testing

What Is Hypothesis Testing?

- Hypothesis testing is an act in statistics whereby an analyst tests an assumption regarding a population parameter. The methodology employed by the analyst depends on the nature of the data used and the reason for the analysis.
- https://miro.medium.com/max/862/1*VXXdieFiYCgR6v7nUaq01g.jpeg

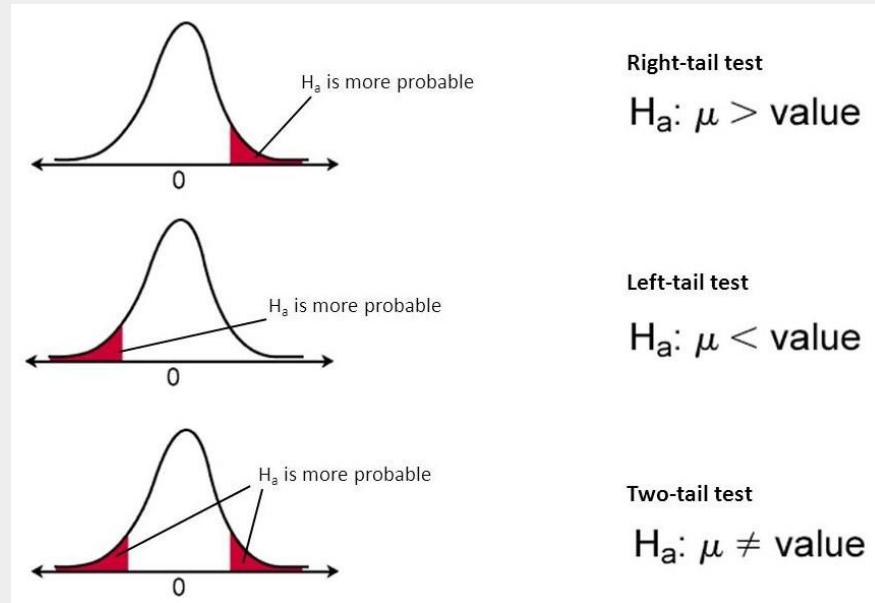


Image Source: https://miro.medium.com/max/862/1*VXXdieFiYCgR6v7nUaq01g.jpeg

Hypothesis Testing

What Is Hypothesis Testing?

- Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. Such data may come from a larger population, or from a data-generating process. The word "population" will be used for both of these cases in the following descriptions.

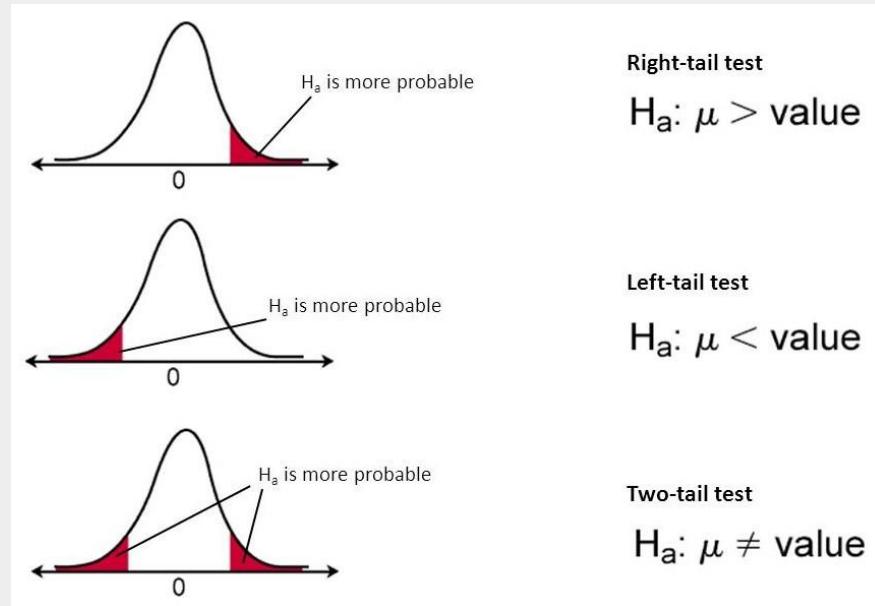
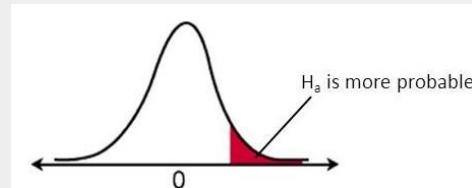


Image Source: https://miro.medium.com/max/862/1*VXXdieFiYCgR6v7nUaq01g.jpeg

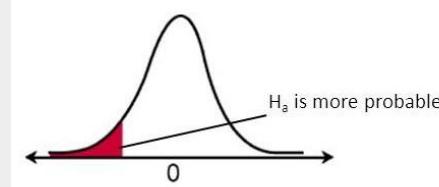
Hypothesis Testing

What Is Hypothesis Testing?

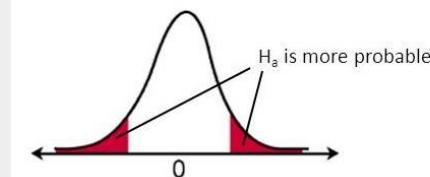
- Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data.
- The test provides evidence concerning the plausibility of the hypothesis, given the data.
- Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.



Right-tail test
 $H_a: \mu > \text{value}$



Left-tail test
 $H_a: \mu < \text{value}$



Two-tail test
 $H_a: \mu \neq \text{value}$

Image Source: https://miro.medium.com/max/862/1*VXxdieFiYCgR6v7nUaq01g.jpeg

Hypothesis Testing

How Hypothesis Testing Works

- In hypothesis testing, an analyst tests a statistical sample, with the goal of providing evidence on the plausibility of the null hypothesis.
- Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.

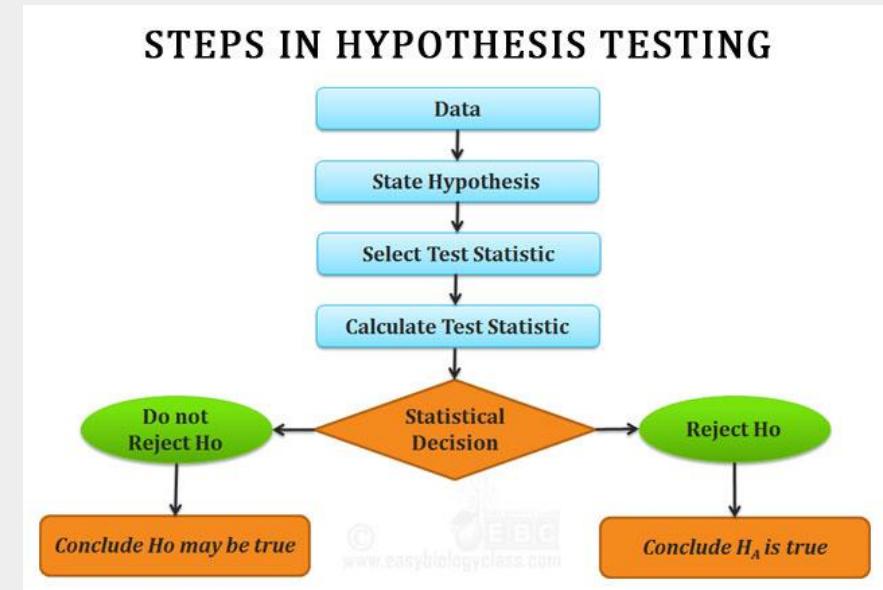


Image Source: https://miro.medium.com/max/642/0*EsLPwer1RYCSczmw

Hypothesis Testing

Types of Hypothesis Testing

- There are two types of Hypothesis Testing
 - i. Null hypothesis
 - ii. Alternative hypothesis

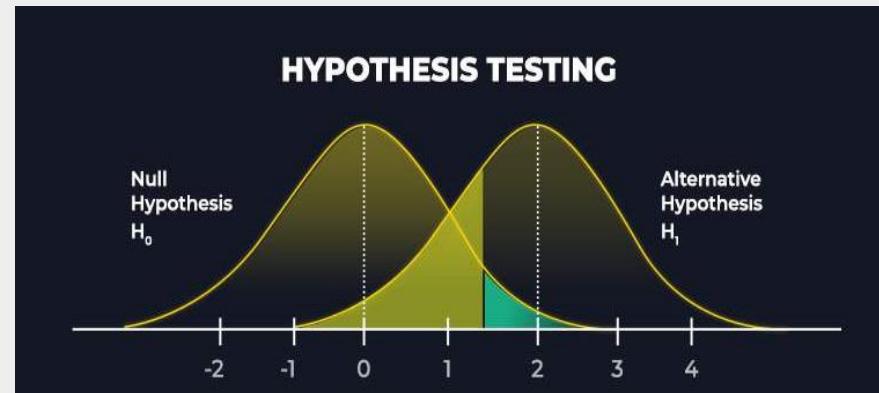


Image Source:

https://www.analyticssteps.com/backend/media/thumbnail/6735922/4237247_1626434645_H0THESIS%20TESTINGArtboard%201.jpg

Hypothesis Testing

Types of Hypothesis Testing

- **Null Hypothesis:** It is denoted by H_0 . A null hypothesis is the one in which sample observations result purely from chance. This means that the observations are not influenced by some non-random cause.

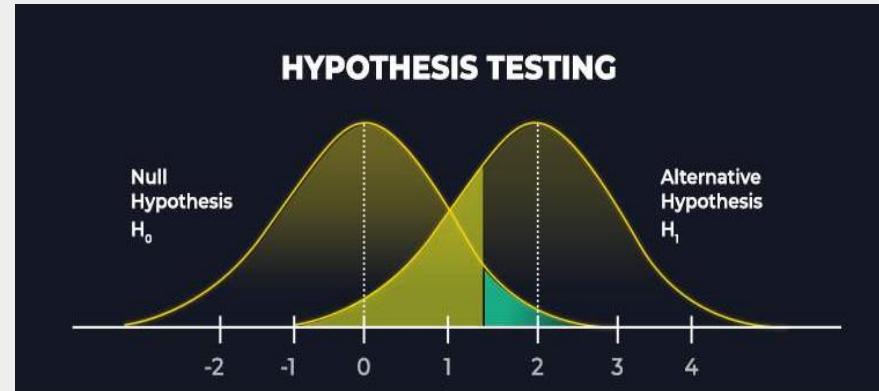


Image Source:

https://www.analyticssteps.com/backend/media/thumbnail/6735922/4237247_1626434645_H0THESIS%20TESTINGArtboard%201.jpg

Hypothesis Testing

Types of Hypothesis Testing

- **Alternative Hypothesis:** It is denoted by H_a or H_1 . An alternative hypothesis is the one in which sample observations are influenced by some non-random cause. A hypothesis test concludes whether to reject the null hypothesis and accept the alternative hypothesis or to fail to reject the null hypothesis. The decision is based on the value of X and R .

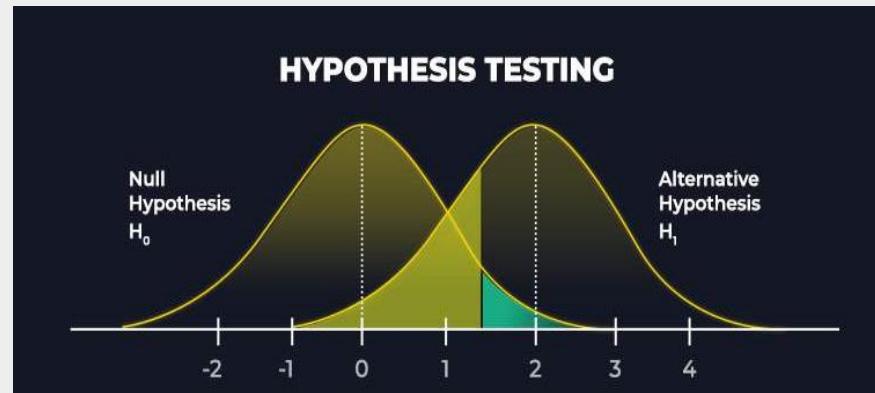


Image Source:

https://www.analyticssteps.com/backend/media/thumbnail/6735922/4237247_1626434645_HYPOTHESIS%20TESTINGArtboard%201.jpg

Hypothesis Testing

Steps in Hypothesis Testing

- Stating the Hypotheses
- Making Statistical Assumptions
- Formulating an Analysis Plan
- Investigating Sample Data
- Interpreting Results

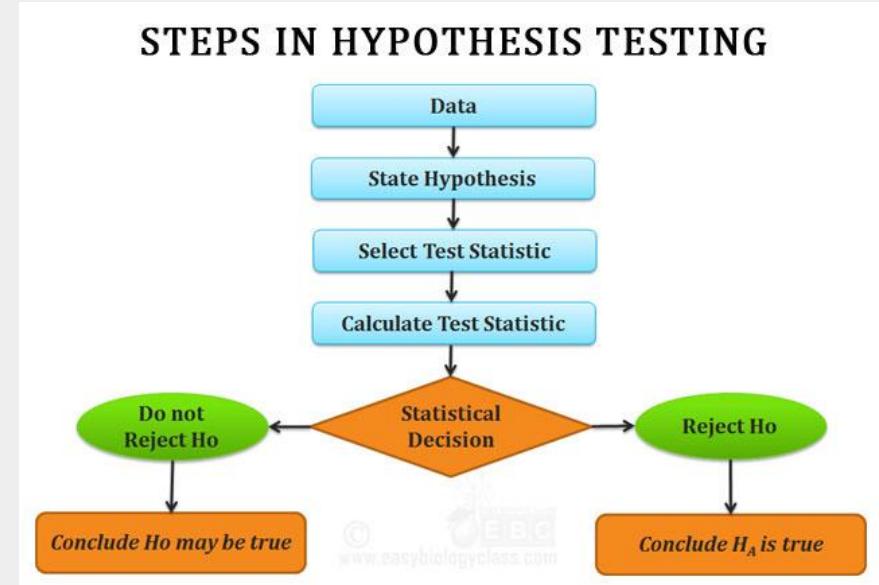


Image Source: https://miro.medium.com/max/642/0*EsLPwer1RYCSczmw

Hypothesis Testing

Accepting or Rejecting Null Hypothesis

- This is an extension of the last step - interpreting results in the process of hypothesis testing. A null hypothesis is accepted or rejected basis P value and the region of acceptance.
- P value – it is a function of the observed sample results.

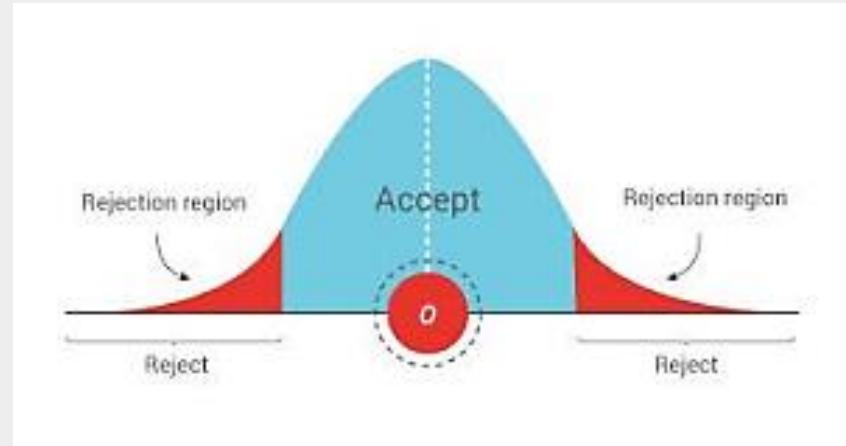


Image Source: <https://i0.wp.com/www.iedunote.com/img/25143/hypothesis-testing.png?resize=299%2C158&quality=100&ssl=1>

Central limit theorem

What is the Central Limit Theorem (CLT)?

- The Central Limit Theorem (CLT) is a statistical concept that states that the sample mean distribution of a random variable will assume a near-normal or normal distribution if the sample size is large enough.

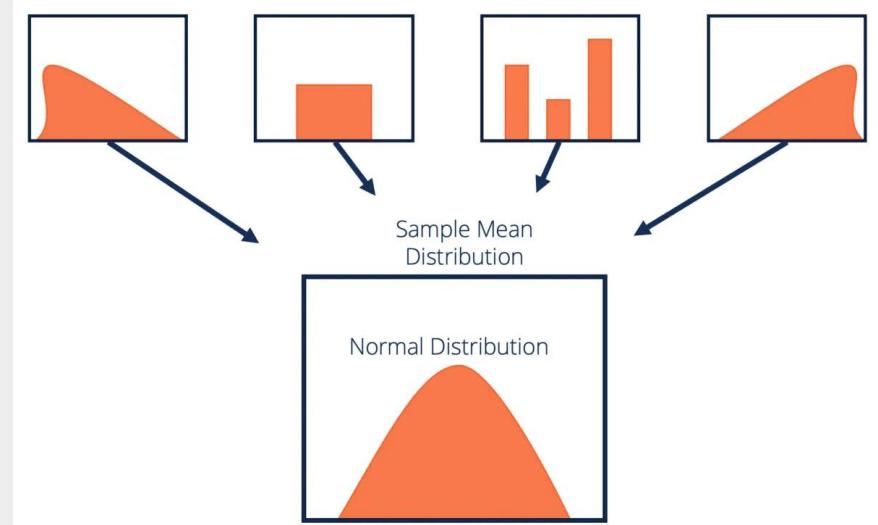


Image Source: <https://cdn.corporatefinanceinstitute.com/assets/Central-Limit-Theorem-CLT-Diagram-1200x734.png>

Central limit theorem

How Does the Central Limit Theorem Work?

- The central limit theorem forms the basis of the probability distribution. It makes it easy to understand how population estimates behave when subjected to repeated sampling. When plotted on a graph, the theorem shows the shape of the distribution formed by means of repeated population samples.

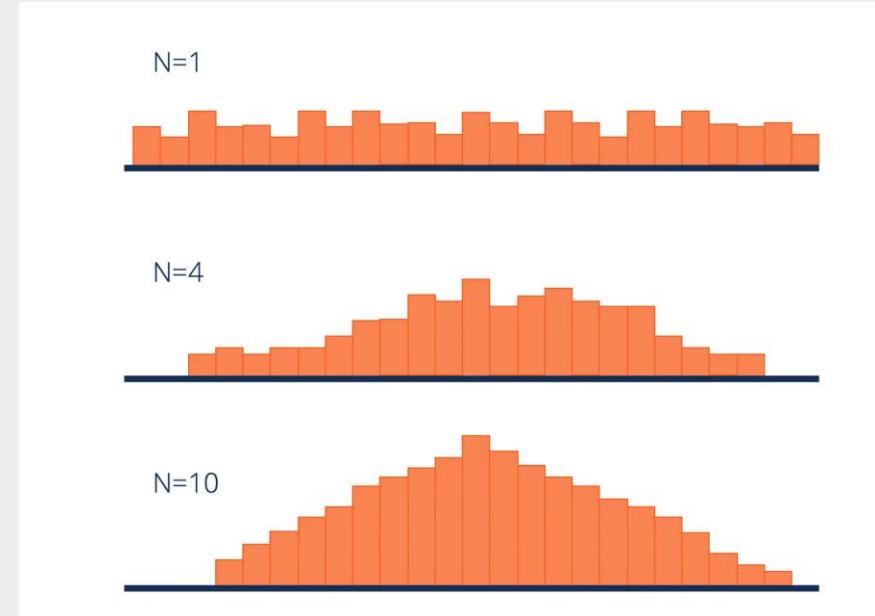


Image Source: <https://cdn.corporatefinanceinstitute.com/assets/Central-Limit-Theorem-CLT-How-it-works-and-how-it-arises.png>

Central limit theorem

How Does the Central Limit Theorem Work?

- From the figure above, we can deduce that despite the fact that the original shape of the distribution was uniform, it tends towards a normal distribution as the value of n (sample size) increases.

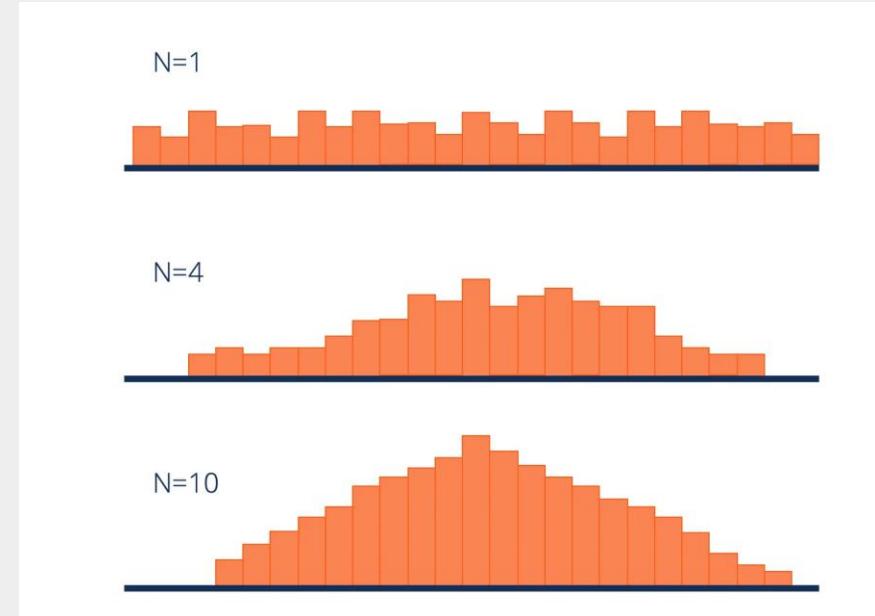


Image Source: <https://cdn.corporatefinanceinstitute.com/assets/Central-Limit-Theorem-CLT-How-it-works-and-how-it-arises.png>

Central limit theorem

How Does the Central Limit Theorem Work?

- Apart from showing the shape that the sample means will take, the central limit theorem also gives an overview of the mean and variance of the distribution. The sample mean of the distribution is the actual population mean from which the samples were taken.

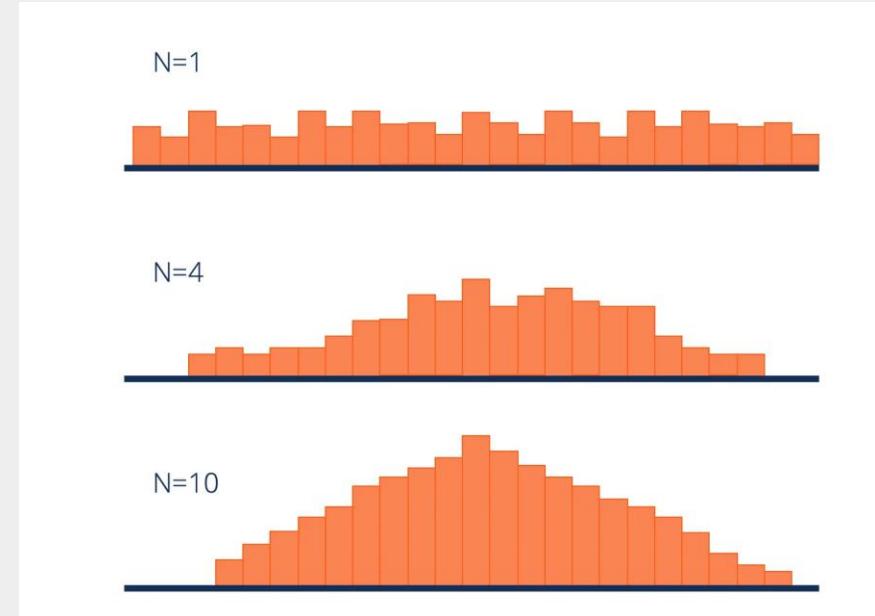


Image Source: <https://cdn.corporatefinanceinstitute.com/assets/Central-Limit-Theorem-CLT-How-it-works-and-how-it-arises.png>

Central limit theorem

How Does the Central Limit Theorem Work?

- The variance of the sample distribution, on the other hand, is the variance of the population divided by n . Therefore, the larger the sample size of the distribution, the smaller the variance of the sample mean.

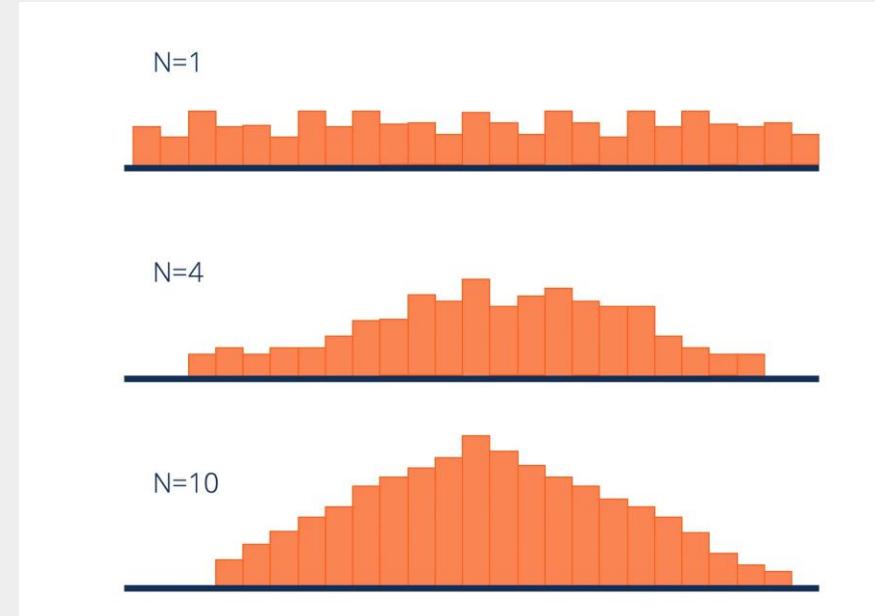


Image Source: <https://cdn.corporatefinanceinstitute.com/assets/Central-Limit-Theorem-CLT-How-it-works-and-how-it-arises.png>

Chi-square test

What Is a Chi-Square Statistic?

- A chi-square (χ^2) statistic is a test that measures how a model compares to actual observed data. The data used in calculating a chi-square statistic must be random, raw, mutually exclusive, drawn from independent variables, and drawn from a large enough sample.

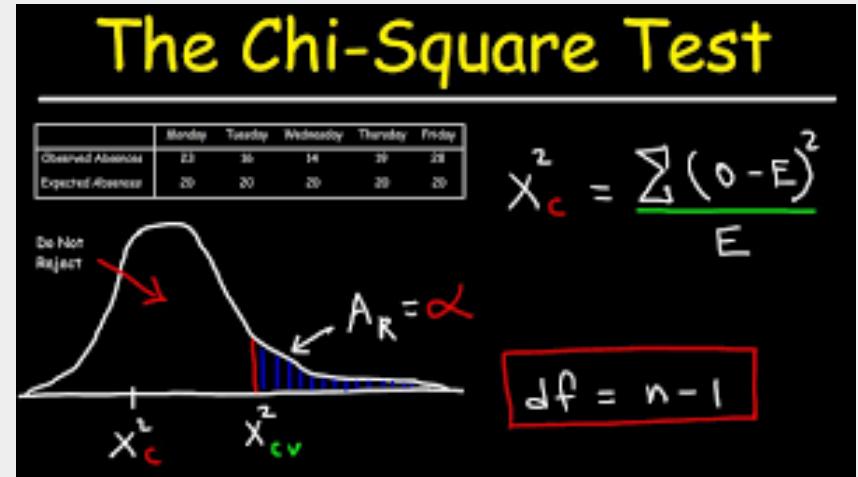


Image Source: <https://i.ytimg.com/vi/HKDqjYSLt68/maxresdefault.jpg>

Chi-square test

What Is a Chi-Square Statistic?

- Chi-square tests are often used in hypothesis testing.
- The chi-square statistic compares the size of any discrepancies between the expected results and the actual results, given the size of the sample and the number of variables in the relationship.

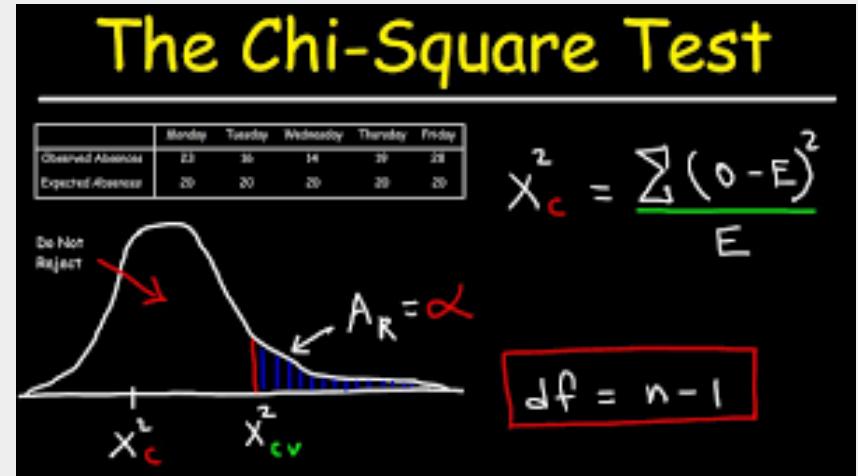


Image Source: <https://i.ytimg.com/vi/HKDqjYSLt68/maxresdefault.jpg>

Chi-square test

The Formula for Chi-Square Is

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

O = Degrees of freedom

O = Observed value(s)

E = Expected value(s)

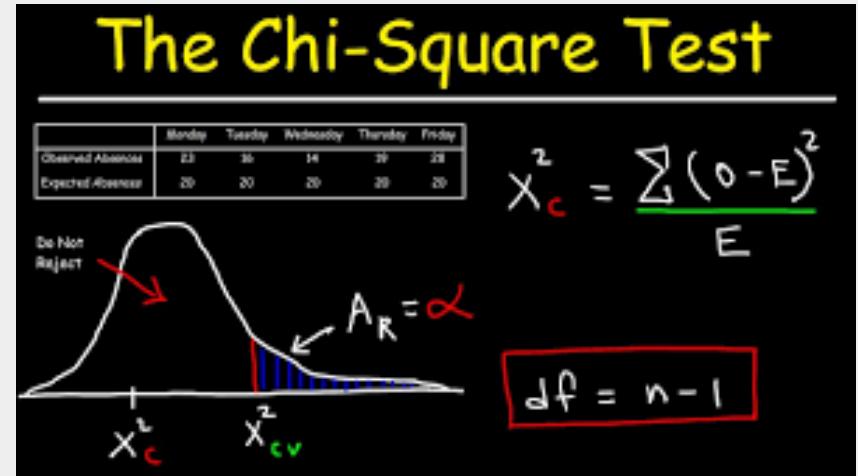


Image Source: <https://i.ytimg.com/vi/HKDqjYSLt68/maxresdefault.jpg>

Chi-square test

What Does a Chi-Square Statistic Tell You?

There are two main kinds of chi-square tests:

- The test of independence, which asks a question of relationship, such as, "Is there a relationship between student sex and course choice?";

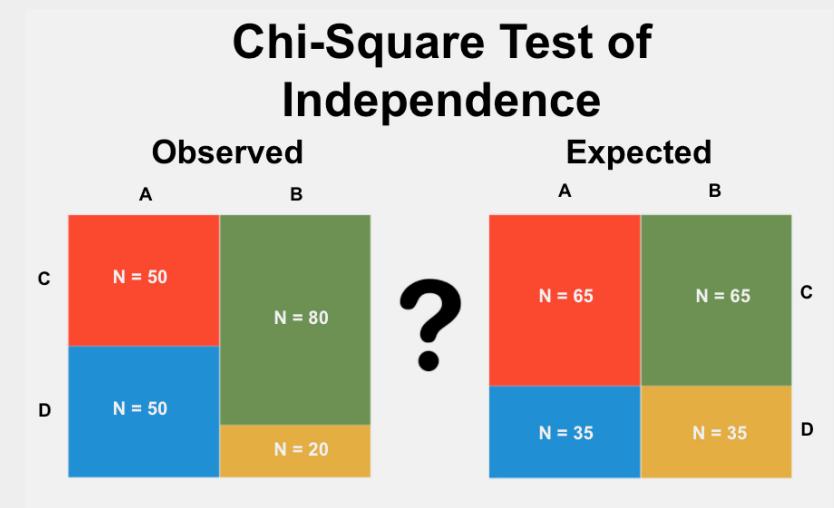


Image Source: <https://www.statstest.com/wp-content/uploads/2020/10/Chi-Square-Test-of-Independence-1.jpg>

Chi-square test

What Does a Chi-Square Statistic Tell You?

- The goodness-of-fit test, which asks something like "How well does the coin in my hand match a theoretically fair coin?"

Chi-square analysis is applied to categorical variables and is especially useful when those variables are nominal (where order doesn't matter, like marital status or gender).

Chi-Square Goodness of Fit Test

Observed

Group 1 N = 70
Group 2 N = 20
Group 3 N = 40
Group 4 N = 70



Expected

Group 1 N = 20
Group 2 N = 40
Group 3 N = 60
Group 4 N = 80

Image Source: <https://www.statstest.com/wp-content/uploads/2020/09/Chi-Square-Goodness-Fit-Test-1024x622.jpg>

Chi-square test

What is a chi-square test used for?

- Chi-square is a statistical test used to examine the differences between categorical variables from a random sample in order to judge goodness of fit between expected and observed results.

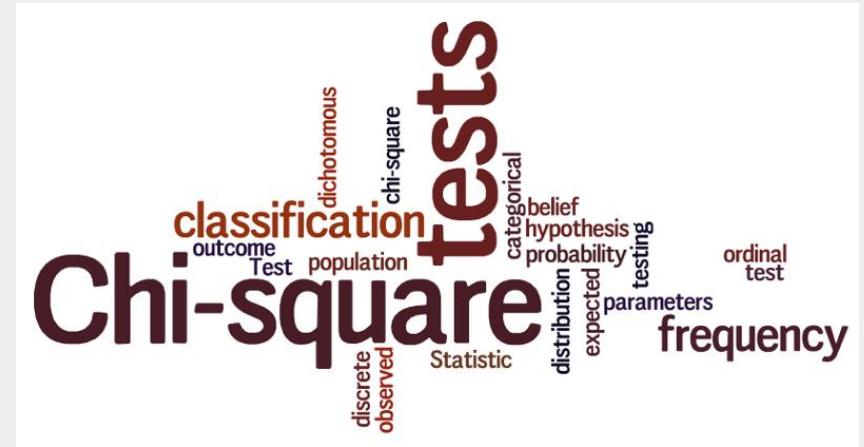


Image Source: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcR7Hd2uJDxTmKlfIWnu_U6Hzq4X8J5DRfRewA&usg=CAU

Chi-square test

Who uses chi-square analysis?

- Researchers
- Nominal or Ordinal variable



Image Source: <https://statswork.com/blog/wp-content/uploads/2019/10/chi-square-infographics.png>

Data Analytics using Python

(30 Hours)

In this section, we will discuss:

- Data mining, wrangling, data manipulation techniques
- Data cleaning and pre-processing techniques
- Data analytics project lifecycle
- Numerical Computing using NumPy Library
- Multidimensional data handling using Pandas Library
- Data Visualization using Matplotlib
- Advanced data visualization using seaborn
- Pandas profiling for report generation
- Need for data visualization

Data mining, wrangling,data manipulation techniques

Data Mining

- Data Mining is defined as extracting information from huge sets of data.
- Data mining is a process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems.



Data mining, wrangling,data manipulation techniques

Why is data mining important?

- Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

Computerization of businesses produce huge amount of data

- ❑ How to make best use of data?
- ❑ Knowledge discovered from data can be used for competitive advantage.

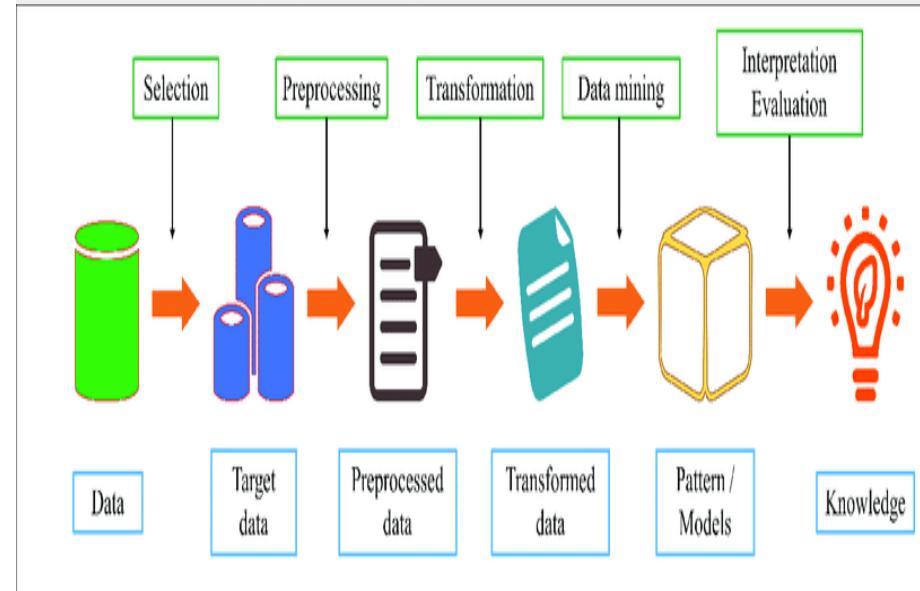
Online e-businesses are generate even larger data sets

- ❑ Online retailers (e.g., amazon.com) are largely driving by data mining.
- ❑ Web search engines are information retrieval (text mining) and data mining companies

Data mining, wrangling,data manipulation techniques

Data mining Steps

- Understand Business
- Understand the Data
- Prepare the Data
- Model the Data
- Evaluate the Data
- Deploy the Solution



Data mining, wrangling,data manipulation techniques

Benefits of Data Mining

- It helps companies gather reliable information.
- It's an efficient, cost-effective solution compared to other data applications.
- It helps businesses make profitable production and operational adjustments.
- Data mining uses both new and legacy systems.
- It helps businesses make informed decisions.

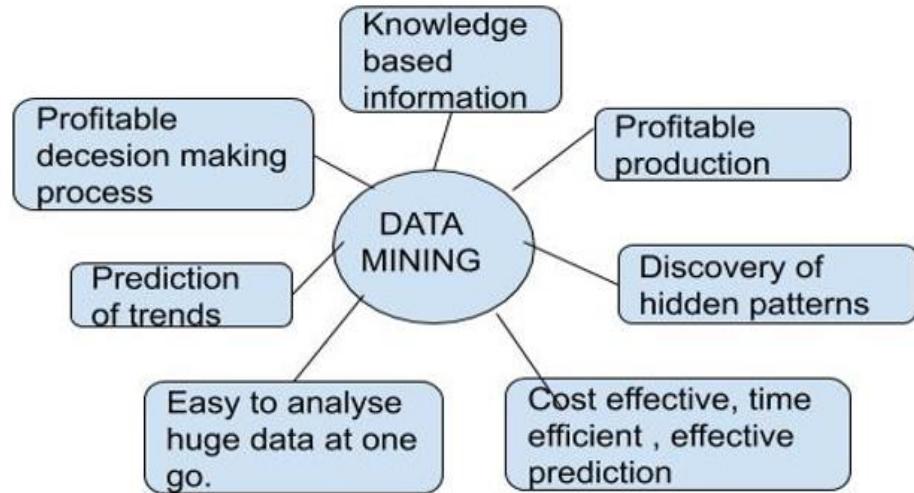


Image Source: <https://www.includehelp.com/basics/data-mining-introduction-benefits-disadvantages-and-applications.aspx>

Data mining, wrangling,data manipulation techniques

Data Mining Tools

- Artificial Intelligence
- Association Rule Learning
- Clustering
- Classification
- Data Analytics
- Data Cleansing and Preparation
- Data Warehousing
- Machine Learning
- Regression



Data mining, wrangling,data manipulation techniques

Data Mining Applications

- Retail
- Financial services
- Insurance
- Manufacturing
- Entertainment
- Healthcare



Data mining, wrangling,data manipulation techniques

Data Wrangling

- Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis.
- Data Wrangling is also known as Data Munging.



Data mining, wrangling,data manipulation techniques

What is the Purpose of Data Wrangling?

- The primary purpose of data wrangling can be described as getting data in coherent shape

DATA WRANGLING

Process of transforming and mapping data from one raw data form into another form with the intent of making it more appropriate and valuable for various tasks

Data mining, wrangling,data manipulation techniques

Data Wrangling Steps

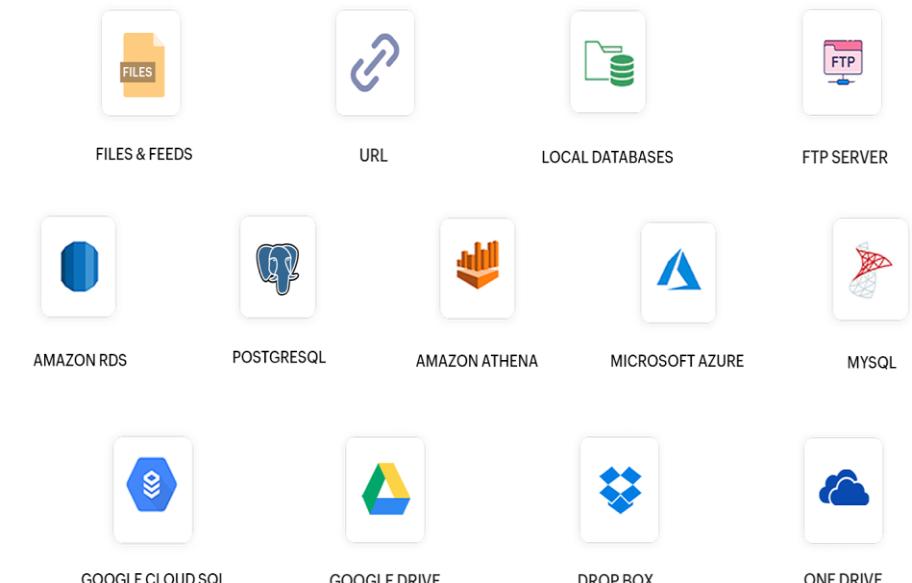
- Data Discovery
- Data Structuring
- Data Cleaning
- Data Enriching
- Data Validating
- Data Publishing



Data mining, wrangling,data manipulation techniques

Data Wrangling Tools

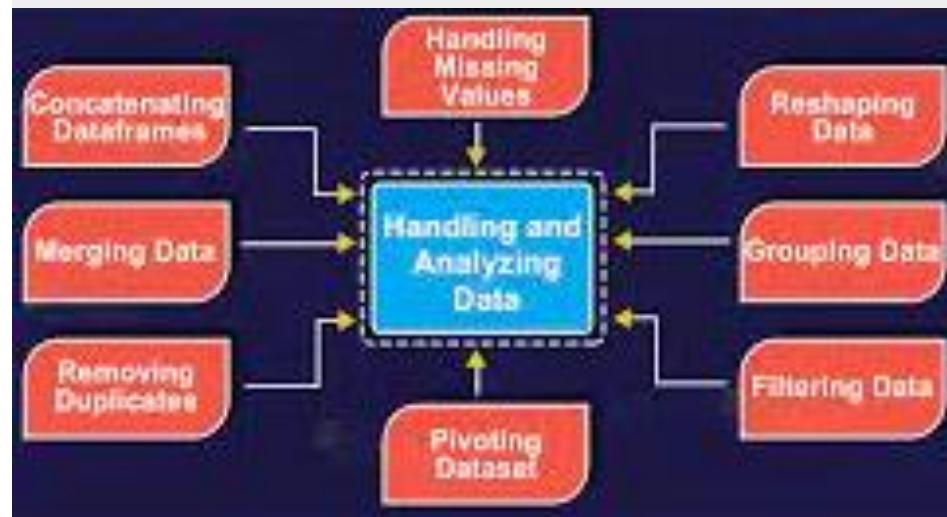
- Excel Power Query / Spreadsheets
- OpenRefine
- Google DataPrep
- Tabula
- DataWrangler
- CSVKit



Data mining, wrangling,data manipulation techniques

Data Wrangling in Python

- Numpy
- Pandas
- Matplotlib
- Plotly
- Theano



Data mining, wrangling,data manipulation techniques

Data Manipulation

- Pandas is an open-source python library that is used for data manipulation and analysis.
- It provides many functions and methods to speed up the data analysis process.



Data mining, wrangling,data manipulation techniques

What are Data Manipulations?

- Data manipulation with python is defined as a process in the python programming language that enables users in data organization in order to make reading or interpreting the insights from the data more structured and comprises of having better design.



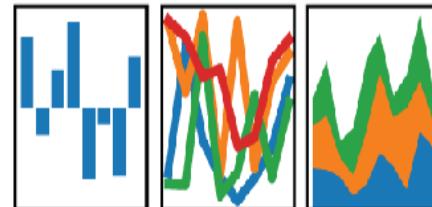
Data mining, wrangling,data manipulation techniques

Pandas

- Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks.
- It is built on top of another package named Numpy, which provides support for multi-dimensional arrays.

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



	BandName	WavelengthMax	WavelengthMin
0	CoastalAerosol	450	430
1	Blue	510	450
2	Green	590	530
3	Red	670	640
4	NearInfrared	880	850
5	ShortWaveInfrared_1	1650	1570
6	ShortWaveInfrared_2	2290	2110
7	Cirrus	1380	1360

Data mining, wrangling,data manipulation techniques

9 Effective Pandas Techniques in Python for Data Manipulation

- Pivot Table
- Boolean Indexing
- Apply Function
- Crosstab
- Merge DataFrames
- Sorting DataFrames
- Plotting
- Cut function
- Impute Missing Values

Dataset

Name	Gender	Age
John	Male	45
Sammy	Female	6
Stephan	Male	4
Joe	Female	36
Emily	Female	12
Tom	Male	43

Pivot Table

Gender	%Gender	Age Group	Count
Male	50%	>18 years	2
		<18 years	1
Female	50%	>18 years	1
		<18 years	2



Data cleaning and pre-processing techniques

Data Cleaning

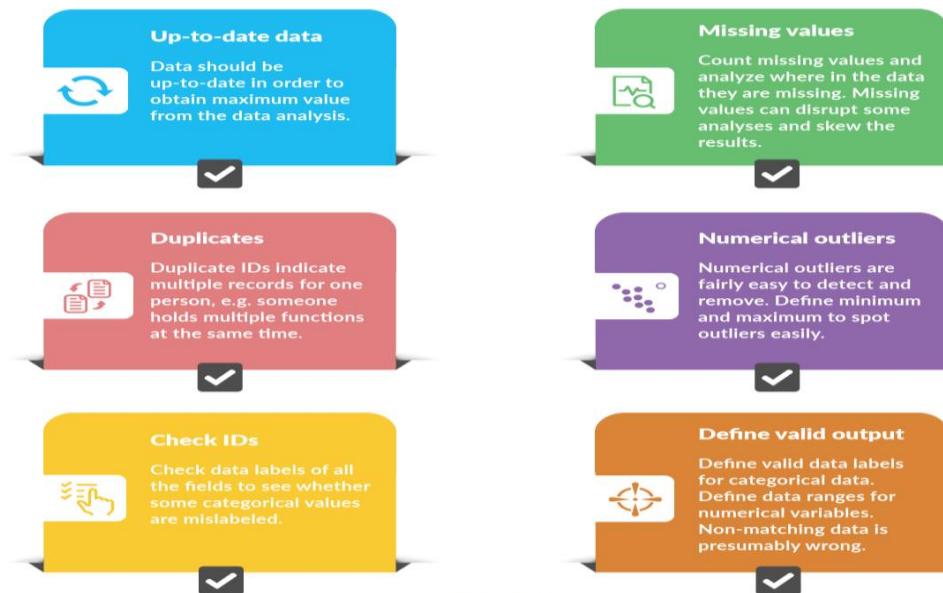
- Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.



Data cleaning and pre-processing techniques

Why data cleaning is essential?

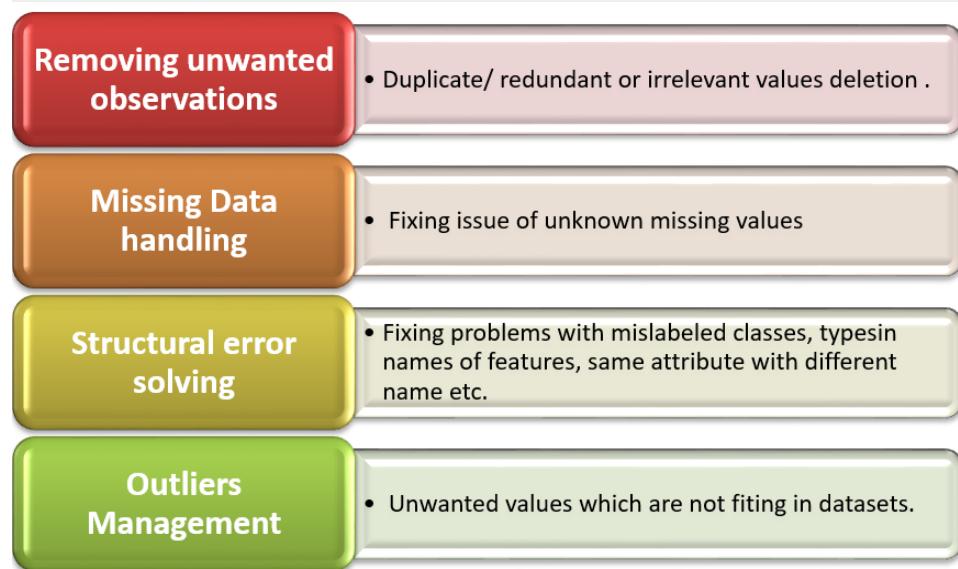
- Error-Free Data
- Data Quality
- Accurate and Efficient
- Complete Data
- Maintains Data Consistency



Data cleaning and pre-processing techniques

8 effective data cleaning techniques

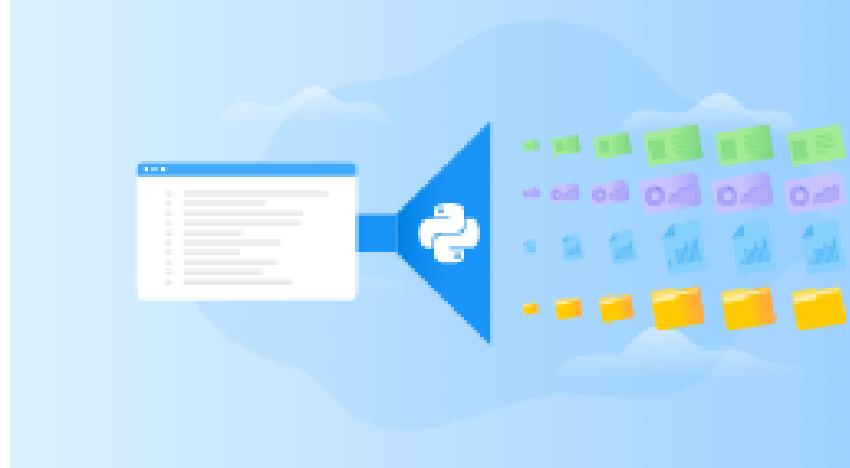
- Remove duplicates
- Remove irrelevant data
- Standardize capitalization
- Convert data type
- Clear formatting
- Fix errors
- Language translation
- Handle missing values



Data cleaning and pre-processing techniques

6 Steps to Manipulate and Cleanse Data with Python

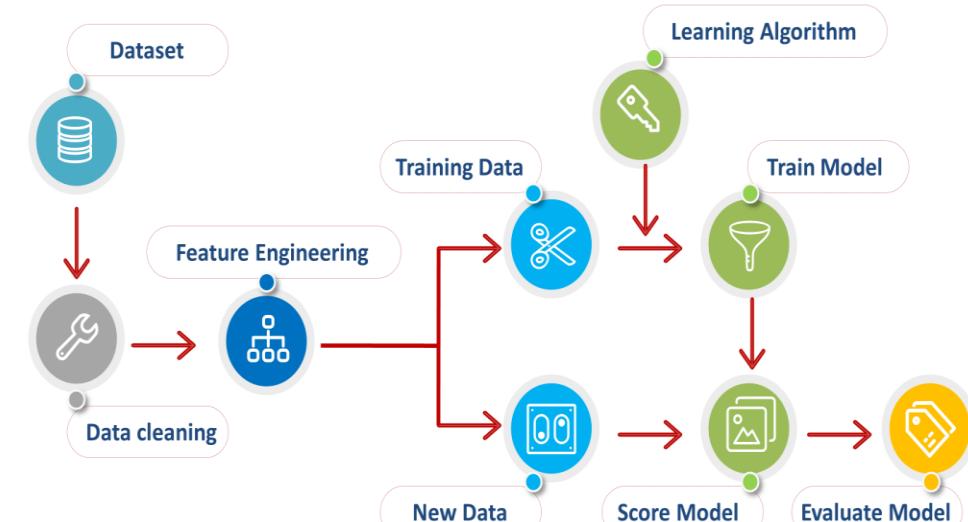
- Imputing Missing Values
- Outlier and Anomaly Detection
- X-Variable Cleaning Methods
- Y-Variable Cleaning Methods
- Merging DataFrames
- Parsing Dates



Data cleaning and pre-processing techniques

Data Pre-processing

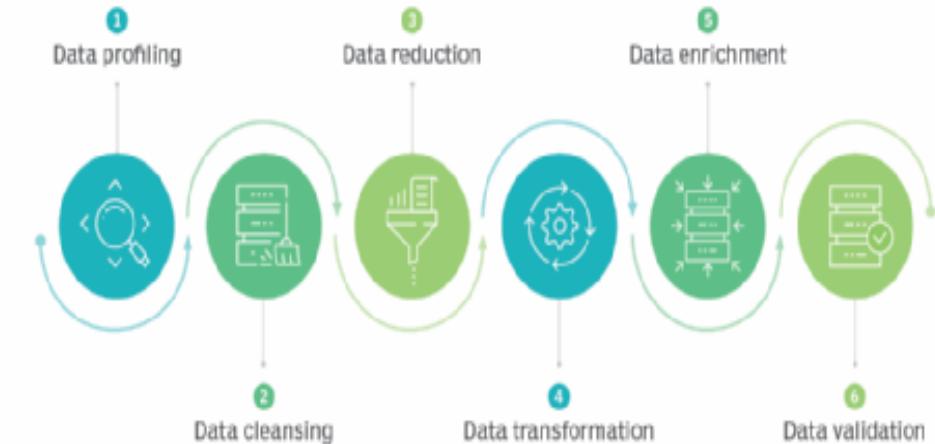
- Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format.



Data cleaning and pre-processing techniques

Data Pre-processing Steps

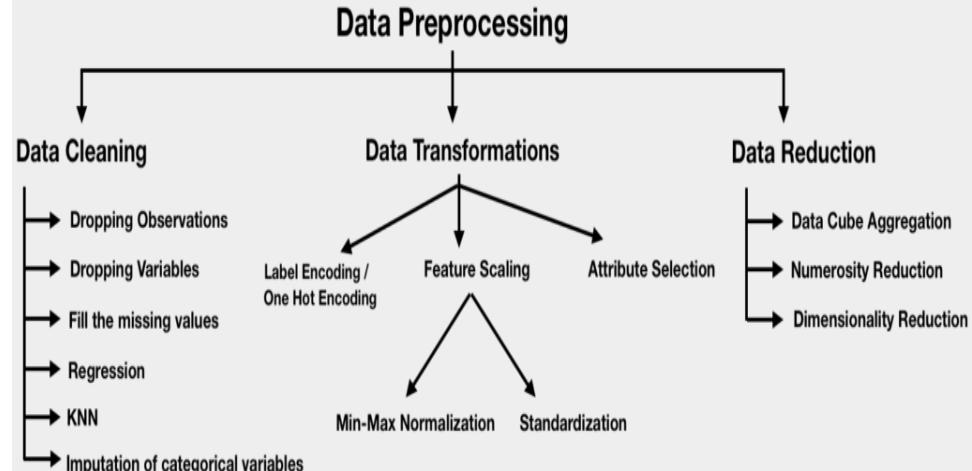
- Data quality assessment
- Data cleaning
- Data transformation
- Data reduction



Data cleaning and pre-processing techniques

Steps involved in data pre-processing

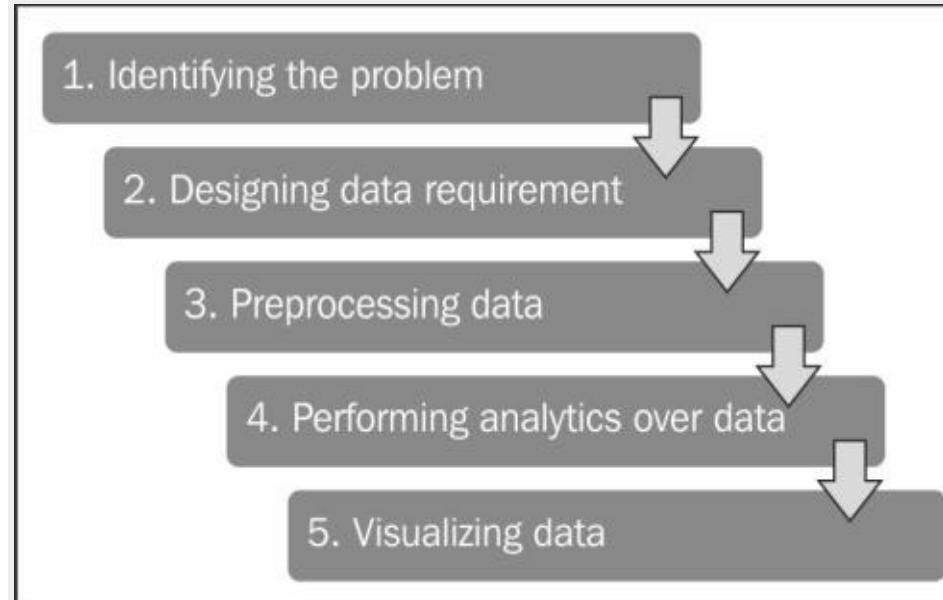
- Importing the required Libraries
- Importing the data set
- Handling the Missing Data.
- Encoding Categorical Data.
- Splitting the data set into test set and training set.
- Feature Scaling.



Data analytics project lifecycle

Project lifecycle

- Domain/Business Understanding
- Data collection/Data Exploration
- Data Cleaning
- Feature Engineering
- Modelling



Data analytics project lifecycle

Project lifecycle

- Evaluation
- Optimization and Tuning
- Deploy the Model to production
- Monitor the performance during Production



Numerical Computing using NumPy Library

What is Numerical Computing?

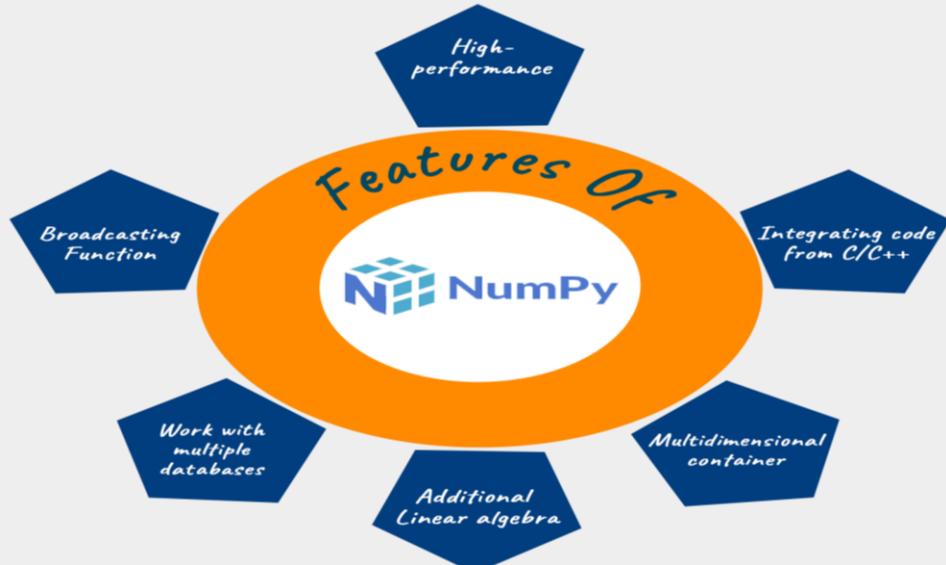
- Numerical computing is an approach for solving complex mathematical problems using only simple arithmetic operations.
- Numerical Python has a fixed-size, homogeneous (fixed-type), multi-dimensional array type and lots of functions for various array operations.

- Numerical computing is an approach for solving complex mathematical problems using only simple arithmetic operations
- The approach involves, in most of the cases, formulation of mathematical models of physical situations that can be solved with arithmetic operations
- It requires development, analysis and use of algorithm
- Algorithm is a systematic procedure that solves a problem or a number of problems
- Its efficiency may be measured by the number of steps in the algorithm, the computer time, and the amount of memory (of the computing instrument) that is required

Numerical Computing using NumPy Library

What is NumPy in Python?

- NumPy is an open-source library available in Python, which helps in mathematical, scientific, engineering, and data science programming.
- It is a very useful library to perform mathematical and statistical operations in Python.



Numerical Computing using NumPy Library

Why use NumPy?

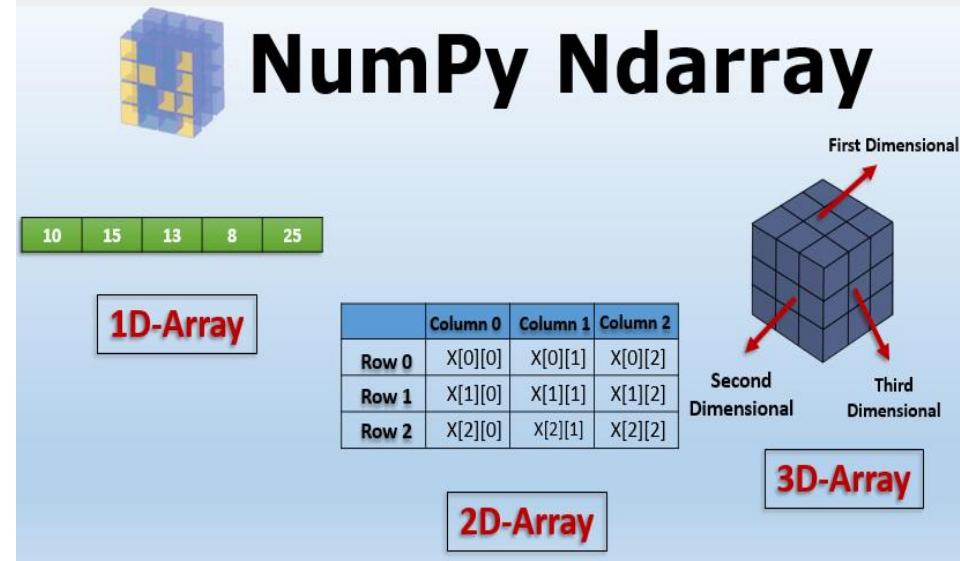
- NumPy is memory efficient, meaning it can handle the vast amount of data more accessible than any other library.



Numerical Computing using NumPy Library

Creating Arrays

- The array object in NumPy is called ndarray.
- We can create a NumPy ndarray object by using the array() function.



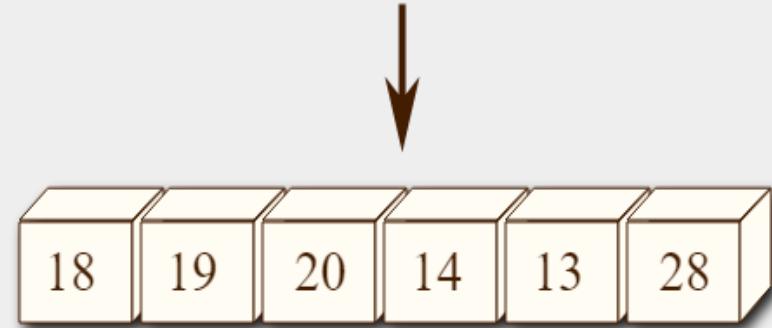
Numerical Computing using NumPy Library

Random Numbers using NumPy

NumPy offers the random module to work with random numbers.

- `rand()`
- `randint()`
- `choice()`

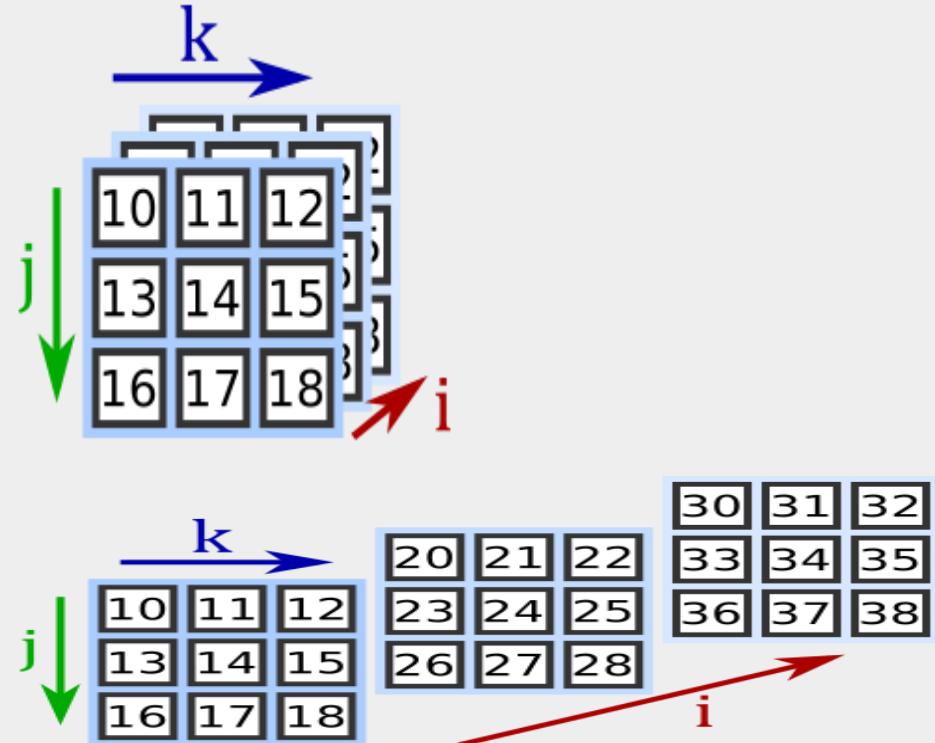
`np.random.randint(low=10, high=30, size=6)`



Numerical Computing using NumPy Library

Indexing and Slicing in Python

- Array indexing is the same as accessing an array element.
- Slicing in python means taking elements from one given index to another given index.



Numerical Computing using NumPy Library

Statistical Functions in Python

- **median:** This will return the median along the specified axis.
- **average:** This will return the weighted average along the specified axis.
- **mean:** This will return the arithmetic mean along the specified axis.
- **std:** This will return the standard deviation along the specified axis.
- **var:** This will return the variance along the specified axis.



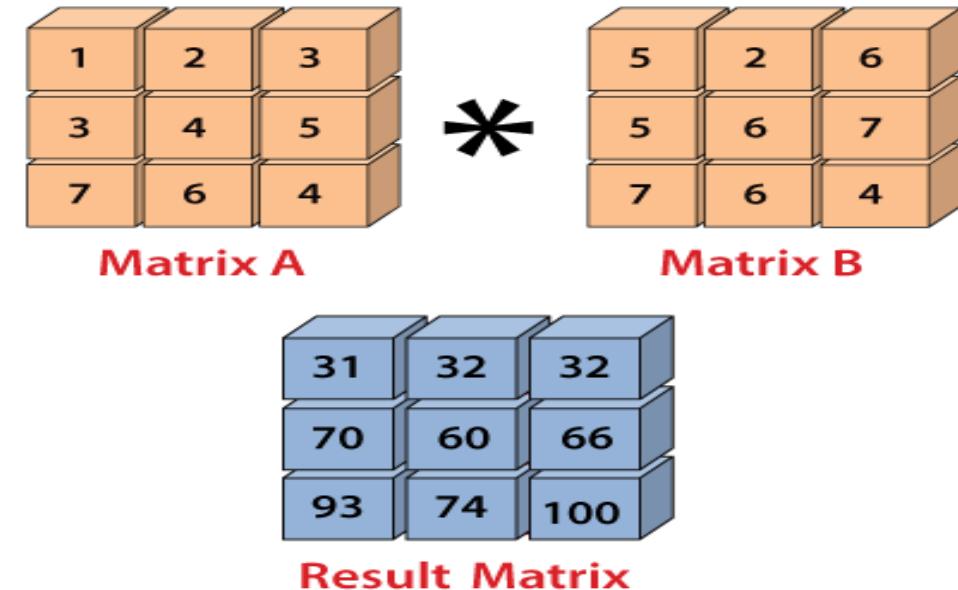
Image Source:

https://d3mxt5v3yxgcsr.cloudfront.net/courses/3391/course_3391_image.jpg

Numerical Computing using NumPy Library

Matrix Multiplication in Python

- The Numpy matmul() function is used to return the matrix product of 2 arrays.



Multidimensional data handling using Pandas Library

What is PANDAS

- PANDAS (PANel Data) is a high level data manipulation tool used for analysis data.
- It is very easy to import and export data using the Pandas library which has a very rich set of functions.



Multidimensional data handling using Pandas Library

What is PANDAS

- Pandas have three important data structures, namely- Series, DataFrame, and Panel to make the process of analyzing data organized, effective and efficient.



Multidimensional data handling using Pandas Library

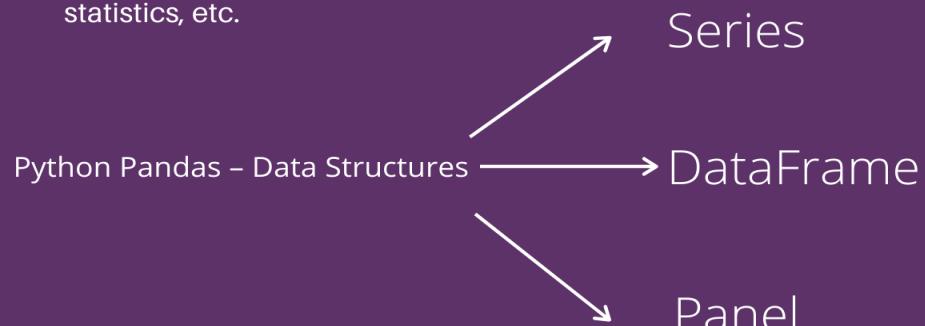
Data Structure in Pandas

Pandas deals with 3 data structure

- Series
- Data Frame
- Panel

Python Pandas Module

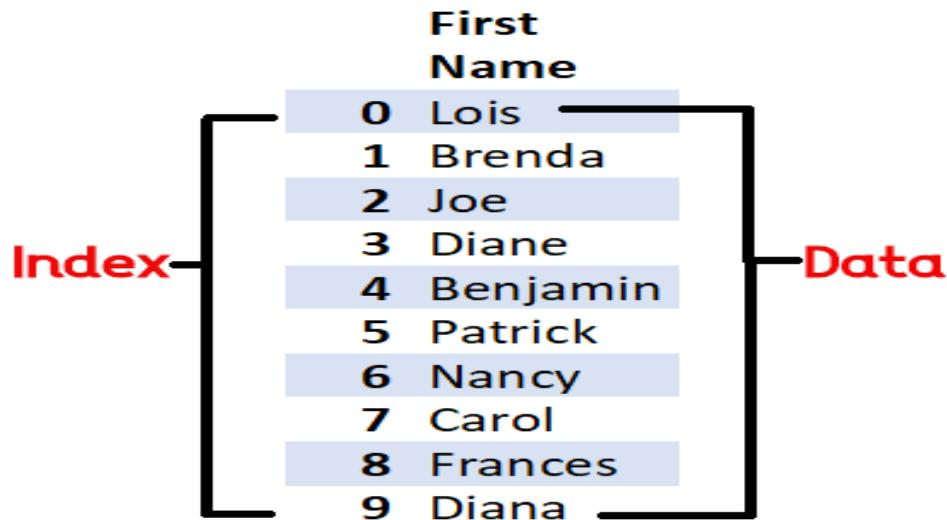
Pandas is basically an open-source Python module. It has a wide scope of use in the field of computing, data analysis, statistics, etc.



Multidimensional data handling using Pandas Library

Series

- Series is a one-dimensional array like structure with homogeneous data, which can be used to handle and manipulate data.

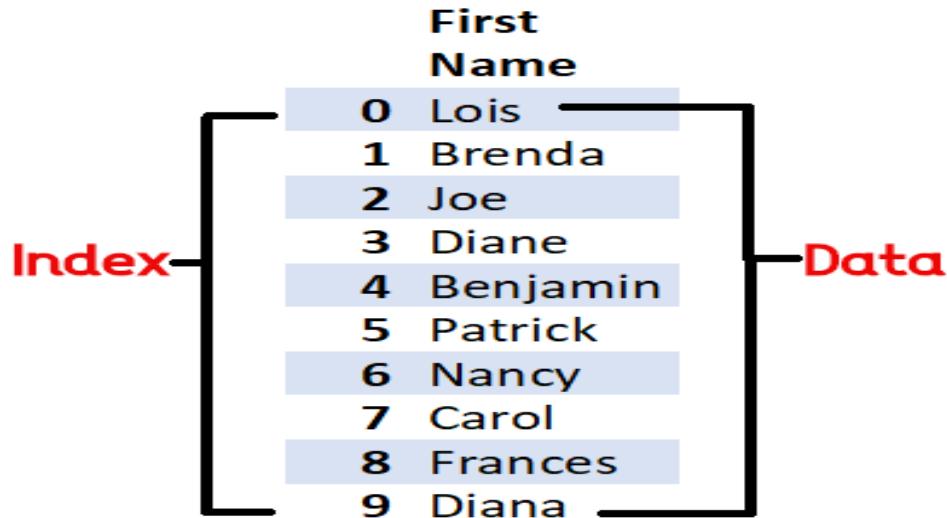


Multidimensional data handling using Pandas Library

Series

It has two parts

- Data part (An array of actual data)
- Associated index with data (associated array of indexes or data labels)

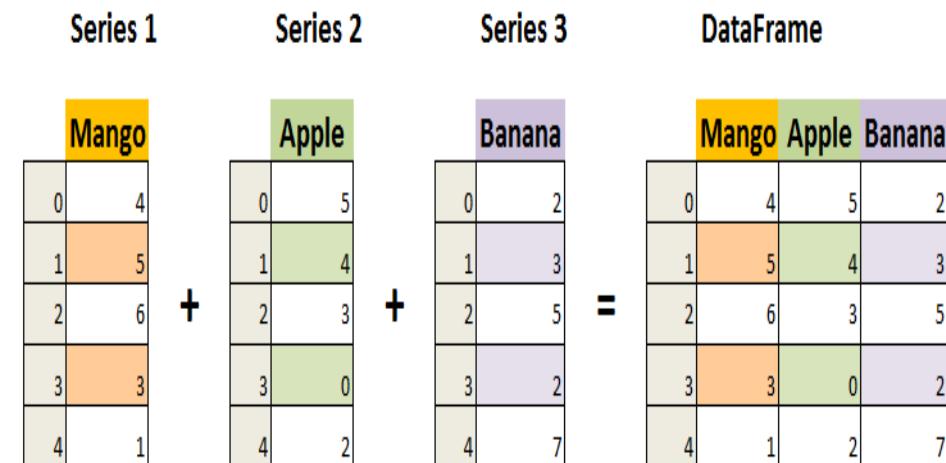


Multidimensional data handling using Pandas Library

Creation of Series

There are different ways in which a series can be created in Pandas.

- Creation of Series from Scalar Values
- Creation of Series from NumPy Arrays
- Creation of Series from Dictionary

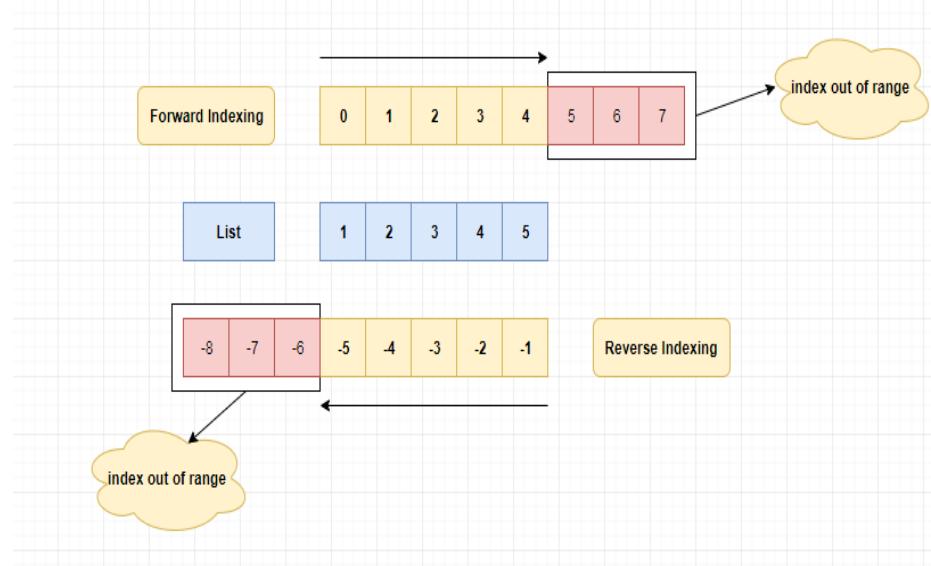


Multidimensional data handling using Pandas Library

Accessing Elements of a Series

There are two common ways for accessing the elements of a series: Indexing and Slicing.

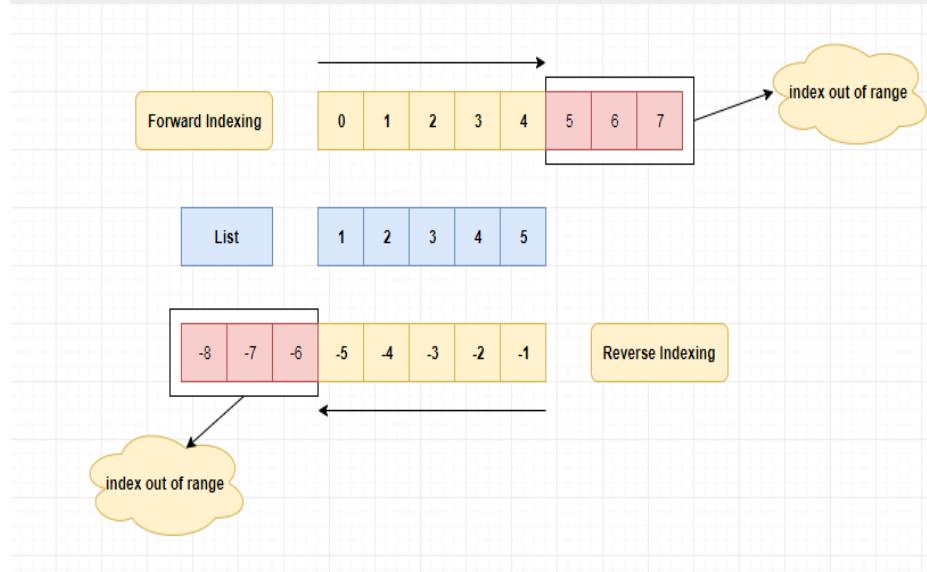
- Indexing
- Slicing



Accessing Elements of a Series

Indexing

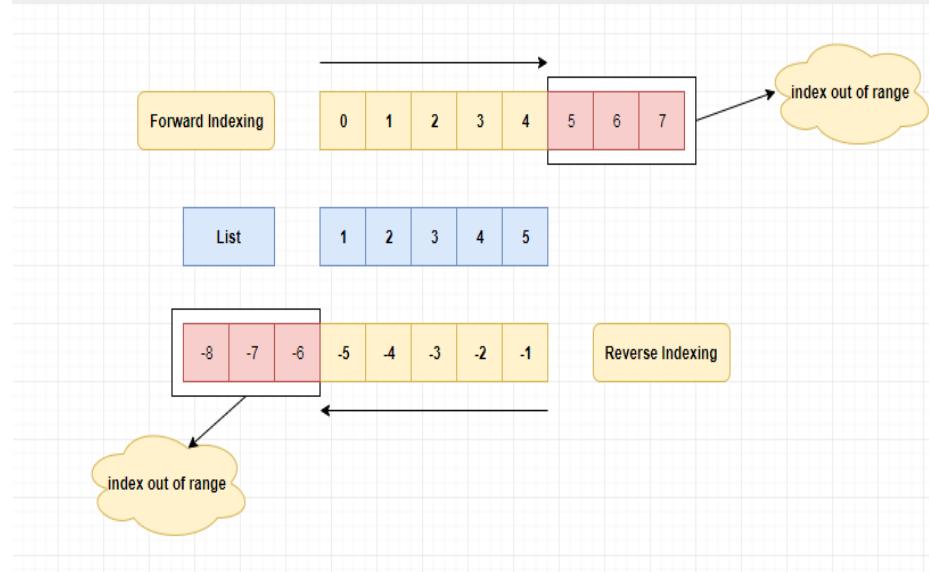
- Indexing in Series is similar to that for NumPy arrays, and is used to access elements in a series.
- Indexes are of two types: positional index and labelled index.
- Positional index takes an integer value that corresponds to its position in the series starting from 0, whereas labelled index takes any user-defined label as index.



Accessing Elements of a Series

Slicing

- Sometimes, we may need to extract a part of a series.
- This can be done through slicing. This is similar to slicing used with NumPy arrays.
- We can define which part of the series is to be sliced by specifying the start and end parameters [start :end] with the series name.

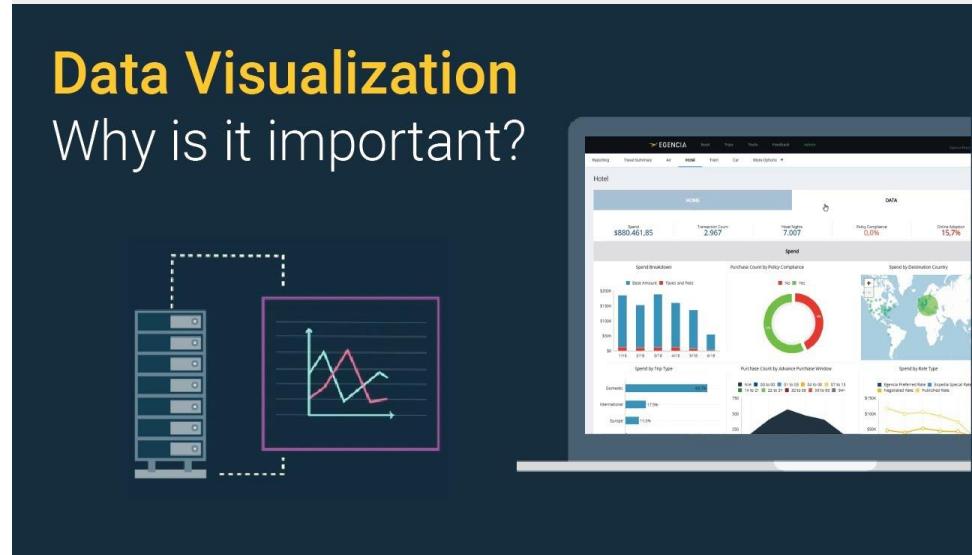


Data Visualization using Matplotlib

Why are visualizations important?

- Visualizations are the easiest way to analyze and absorb information.
- Visuals help to easily understand the complex problem.
- They help in identifying patterns, relationships, and outliers in data.

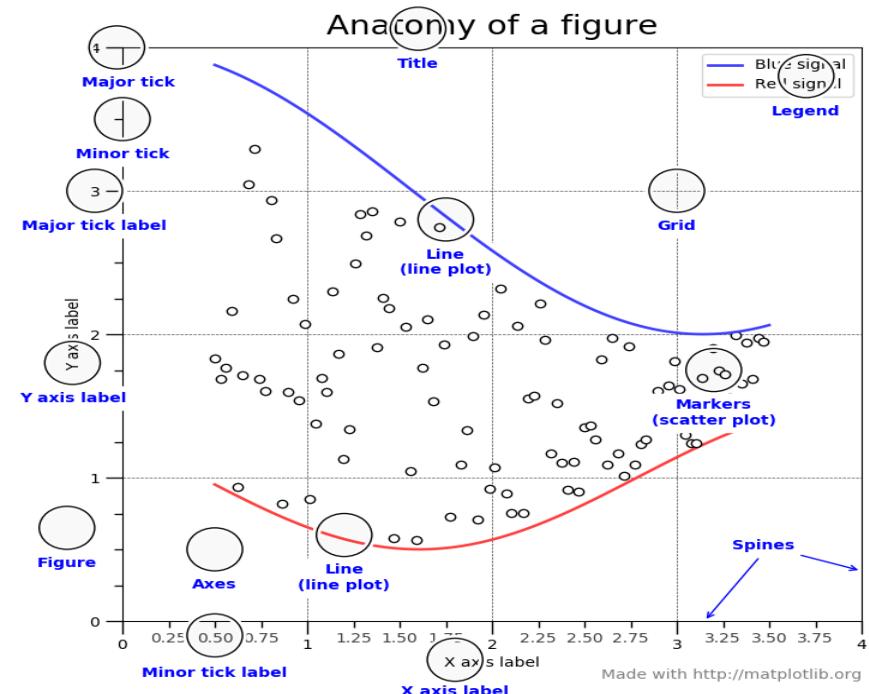
Data Visualization
Why is it important?



Data Visualization using Matplotlib

Matplotlib

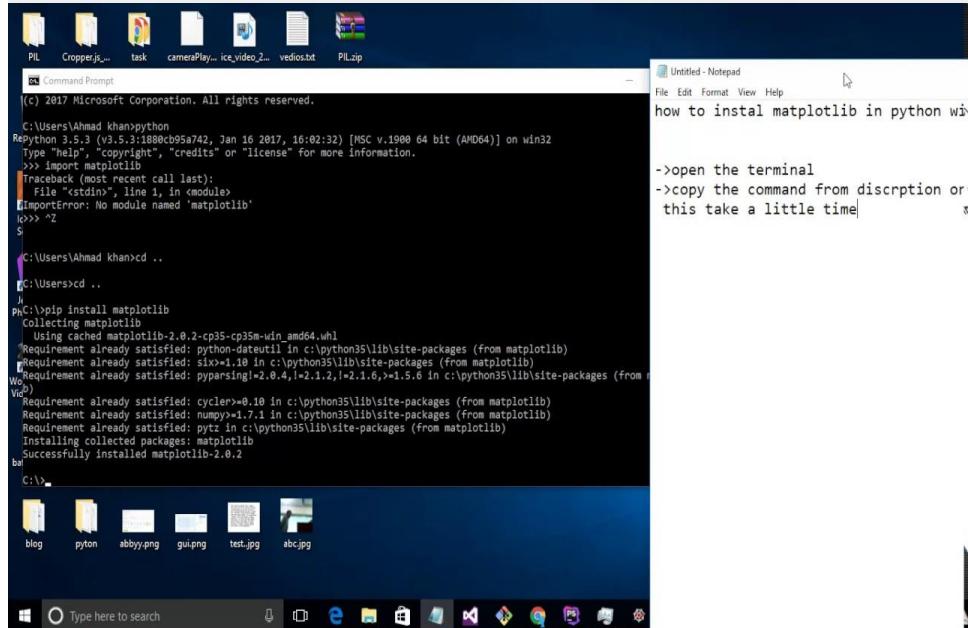
- Matplotlib is a 2-D plotting library that helps in visualizing figures.
- Matplotlib emulates Matlab like graphs and visualizations.
- Matlab is not free, is difficult to scale and as a programming language is tedious.



Data Visualization using Matplotlib

Installing Matplotlib

- Type !pip install matplotlib in the Jupyter Notebook or if it doesn't work in cmd type conda install -c conda-forge matplotlib .
- This should work in most cases.



The screenshot shows a Windows desktop environment. At the top, there's a taskbar with various icons. Below the taskbar, a Command Prompt window is open, showing the following text:

```
(c) 2017 Microsoft Corporation. All rights reserved.  
Python 3.5.3 (v3.5.3:1388ccb95a742, Jan 16 2017, 16:02:32) [MSC v.1900 64 bit (AMD64)] on win32  
Type "help", "copyright", "credits" or "license" for more information.  
>>> import matplotlib  
Traceback (most recent call last):  
  File "<stdin>", line 1, in <module>  
ImportError: No module named 'matplotlib'  
>>> ^Z  
S  
  
C:\Users\Ahmad Khan>d ..  
C:\Users>d ..  
C:\>pip install matplotlib  
Collecting matplotlib>=2.0.2-cp35-cp35m-win_amd64.whl  
  Using cached matplotlib-2.0.2-cp35-cp35m-win_amd64.whl  
Requirement already satisfied: python-dateutil in c:\python35\lib\site-packages (from matplotlib)  
Requirement already satisfied: six>=1.10 in c:\python35\lib\site-packages (from matplotlib)  
Requirement already satisfied: pytz in c:\python35\lib\site-packages (from matplotlib)  
Requirement already satisfied: pyparsing>=2.0.4,!=2.1.2,!=2.1.6,>=1.5.6 in c:\python35\lib\site-packages (from matplotlib)  
Requirement already satisfied: cycler>=0.10 in c:\python35\lib\site-packages (from matplotlib)  
Requirement already satisfied: numpy>=1.7.1 in c:\python35\lib\site-packages (from matplotlib)  
Requirement already satisfied: pytz in c:\python35\lib\site-packages (from matplotlib)  
Installing collected packages: matplotlib  
Successfully installed matplotlib-2.0.2  
ba  
C:\>
```

To the right of the Command Prompt, a Notepad window titled "Untitled - Notepad" contains the following text:

```
->open the terminal  
->copy the command from discription or  
this take a little time|
```

Data Visualization using Matplotlib

Things to follow

- Plotting of Matplotlib is quite easy.
Generally, while plotting they follow the same steps in each and every plot.
- Matplotlib has a module called pyplot which aids in plotting figure.
- The Jupyter notebook is used for running the plots.

The screenshot shows a Windows desktop environment. At the top, there's a taskbar with several icons. Below it is a Start button. The main area features two windows: a 'Command Prompt' window and a 'Untitled - Notepad' window. The Command Prompt window shows Python code being run to install matplotlib, with output indicating the package is successfully installed. The Notepad window contains instructions: '->open the terminal', '->copy the command from discription or this take a little time'. The desktop background is dark blue.

```
(c) 2017 Microsoft Corporation. All rights reserved.
C:\Users\Ahmad khan>python
Python 3.5.3 (v3.5.3:188ccb95a742, Jan 16 2017, 16:02:32) [MSC v.1900 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> import matplotlib
Traceback (most recent call last):
  File "<stdin>", line 1, in <module>
ImportError: No module named 'matplotlib'
>>> ^Z
S

C:\Users\Ahmad khan>cd ..

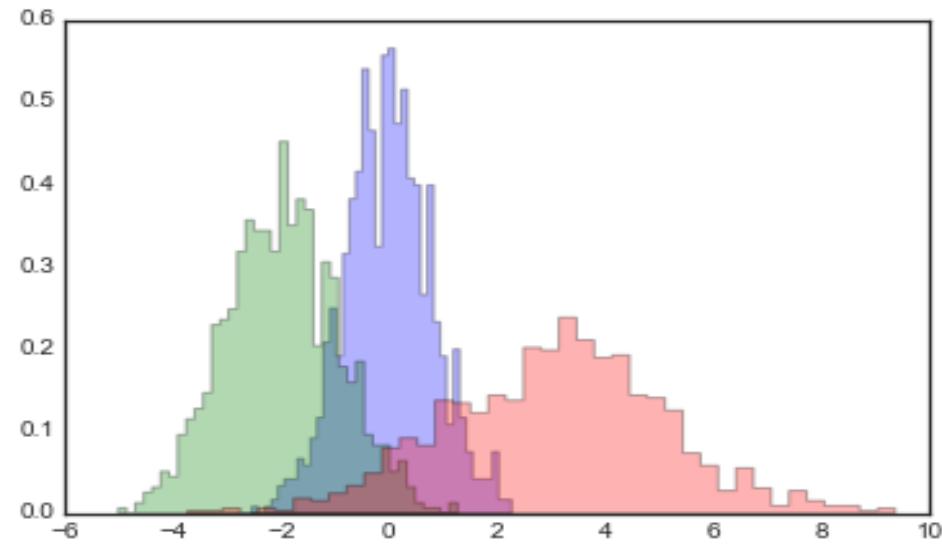
C:\Users>cd ..
j
PhC:\>pip install matplotlib
Collecting matplotlib>=2.0.2-cp35-cp35m-win_amd64.whl
  Using cached matplotlib-2.0.2-cp35-cp35m-win_amd64.whl
Requirement already satisfied: python-dateutil in c:\python35\lib\site-packages (from matplotlib)
Requirement already satisfied: six>=1.10 in c:\python35\lib\site-packages (from matplotlib)
Requirement already satisfied: pytz in c:\python35\lib\site-packages (from matplotlib)
Requirement already satisfied: pyparsing>=2.0.4,>=2.1.2,>=2.1.6,>=1.5.6 in c:\python35\lib\site-packages (from Wic)
Requirement already satisfied: cython>=0.10 in c:\python35\lib\site-packages (from matplotlib)
Requirement already satisfied: numpy>=1.7.1 in c:\python35\lib\site-packages (from matplotlib)
Requirement already satisfied: pytz in c:\python35\lib\site-packages (from matplotlib)
Installing collected packages: matplotlib
Successfully installed matplotlib-2.0.2
ba
C:\>.
```

->open the terminal
->copy the command from discription or
this take a little time|

Data Visualization using Matplotlib

Histogram

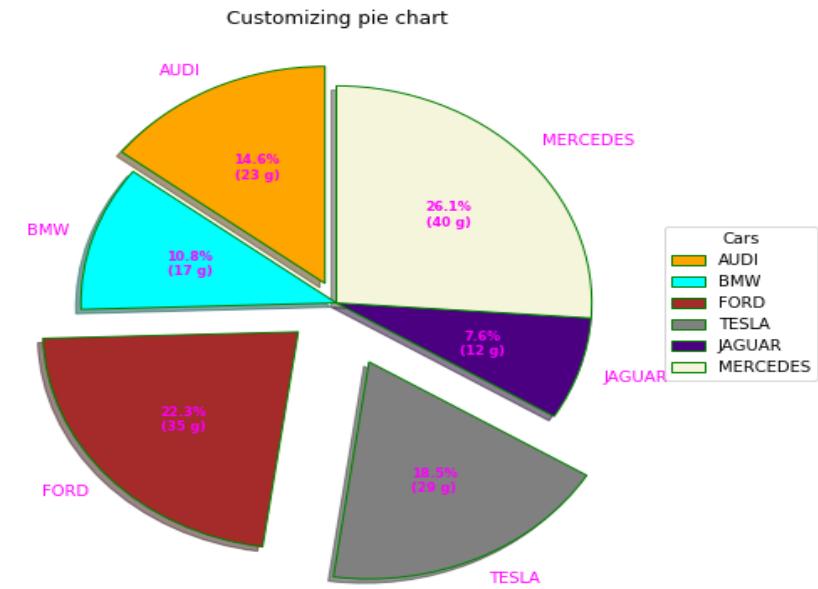
- A histogram takes in a series of data and divides the data into a number of bins.
- It then plots the frequency data points in each bin (i.e. the interval of points).
- It is useful in understanding the count of data ranges.



Data Visualization using Matplotlib

Histogram: Pie chart

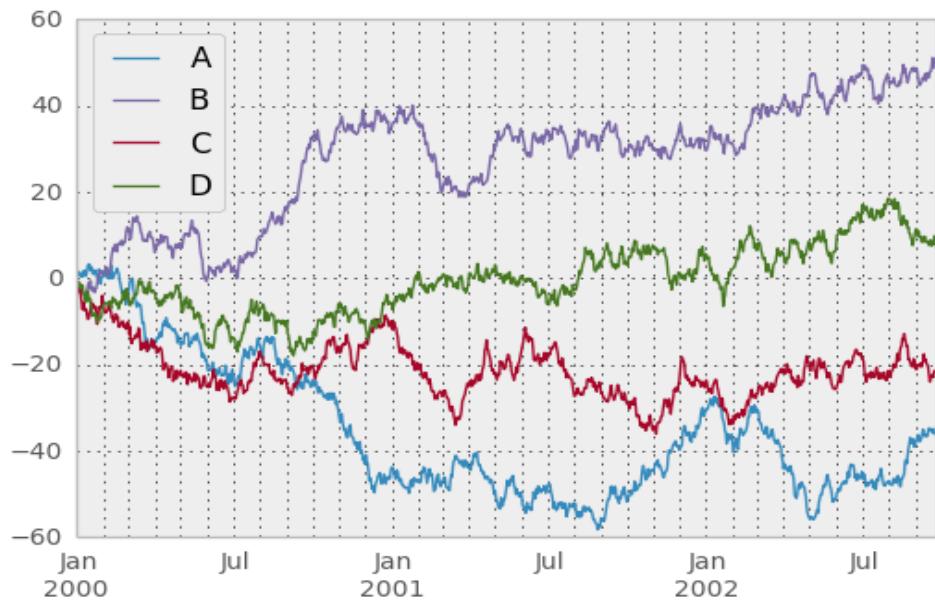
- It is a circular plot which is divided into slices to illustrate numerical proportion. The slice of a pie chart is to show the proportion of parts out of a whole.
- When to use: Pie chart should be used seldom used as It is difficult to compare sections of the chart. Bar plot is used instead as comparing sections is easy.



Data Visualization using Matplotlib

Time Series by line plot

- Time series is a line plot and it is basically connecting data points with a straight line.
- It is useful in understanding the trend over time.
- It can explain the correlation between points by the trend.
- An upward trend means positive correlation and downward trend means a negative correlation.



Data Visualization using Matplotlib

Time Series by line plot

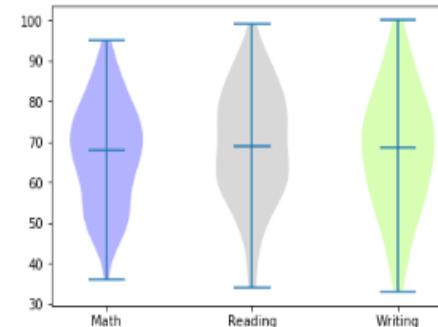
- Violin plot is a better chart than boxplot as it gives a much broader understanding of the distribution.
- It resembles a violin and dense areas point the more distribution of data otherwise hidden by box plots

Violin Plots

Similar to boxplots, except they can show the density of the data points around a particular value with their widths

In [7]:

```
1 vp = plt.violinplot(exam_scores_array,  
2                      showmedians=True)  
3  
4 plt.xticks([1, 2, 3], ['Math', 'Reading', 'Writing'])  
5  
6 for i in range(len(vp['bodies'])):  
7     vp['bodies'][i].set(facecolor=colors[i])  
8  
9 plt.show()
```



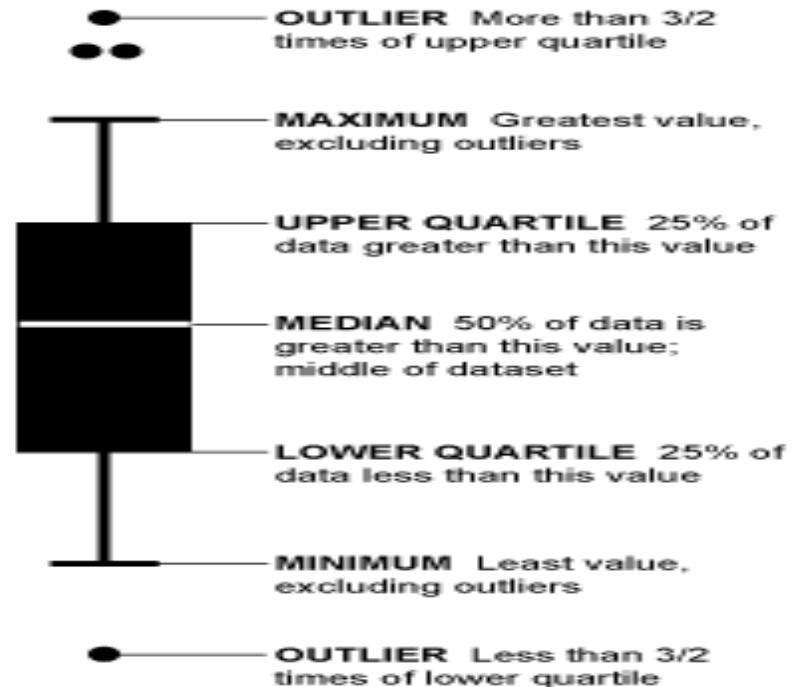
Image

Source:https://miro.medium.com/max/1204/1*J9OnuX8f5BjlB3XZiyHkVA.png

Boxplot and Violinplot

Boxplot

- Boxplot gives a nice summary of the data.
- It helps in understanding our distribution better.



Data Visualization using Matplotlib

TwinAxis

- TwinAxis helps in visualizing plotting 2 plots w.r.t to the y-axis and same x-axis.

In [2]: 1 austin_weather.head()

Out[2]:

	Date	TempHighF	TempAvgF	TempLowF	DewPointHighF	DewPointAvgF	DewPointLowF	HumidityHighPercent	HumidityAvgPercent	HumidityLowPercent
0	2013-12-21	74	60	45	67	49	43	93	75	57
1	2013-12-22	56	48	39	43	36	28	93	68	43
2	2013-12-23	58	45	32	31	27	23	78	52	27
3	2013-12-24	61	46	31	36	28	21	89	56	22
4	2013-12-25	58	50	41	44	40	36	88	71	58

5 rows × 21 columns

Extracting Date,Avg Temperature and Avg Wind Speed columns

In [3]: 1 austin_weather = austin_weather[['Date', 'TempAvgF', 'WindAvgMPH']].head(30)
2 austin_weather

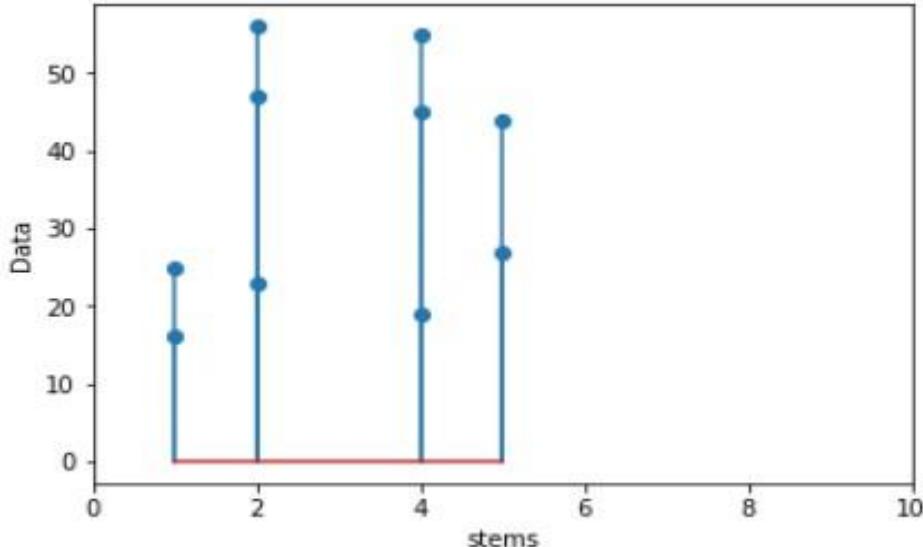
Out[3]:

	Date	TempAvgF	WindAvgMPH
0	2013-12-21	60	4
1	2013-12-22	48	6

Data Visualization using Matplotlib

Stack Plot and Stem Plot

- Stack plot visualizes data in stacks and shows the distribution of data over time.
- Stemplot even takes negative values, so the difference is taken of data and is plotted over time.



Data Visualization using Matplotlib

Bar Plot

- Bar Plot shows the distribution of data over several groups.
- It is commonly confused with a histogram which only takes numerical data for plotting.
- It helps in comparing multiple numeric values.

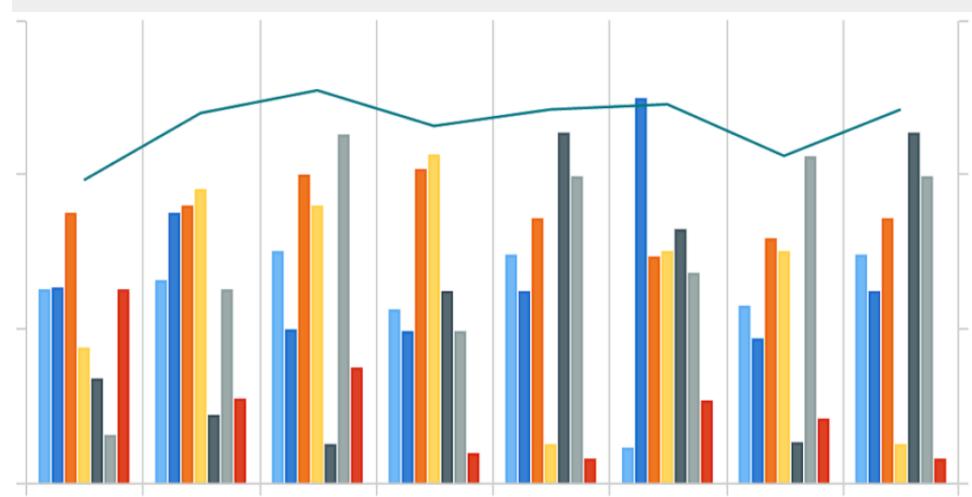
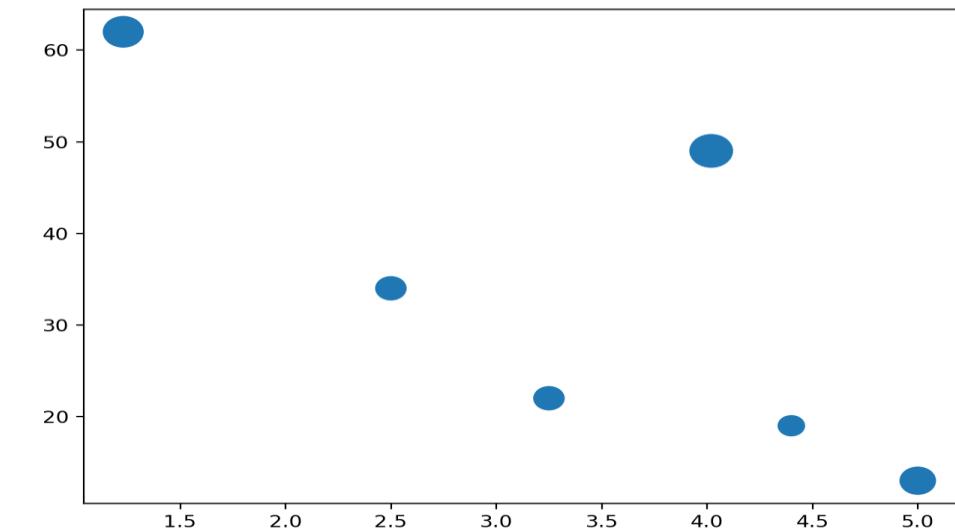


Image Source:<https://shanelynnwebsite-mid9n9g1q9y8tt.netdna-ssl.com/wp-content/uploads/2020/04/data-visualisation-bar-charts-in-python-pandas-1-1024x441.png>

Data Visualization using Matplotlib

Scatter Plot

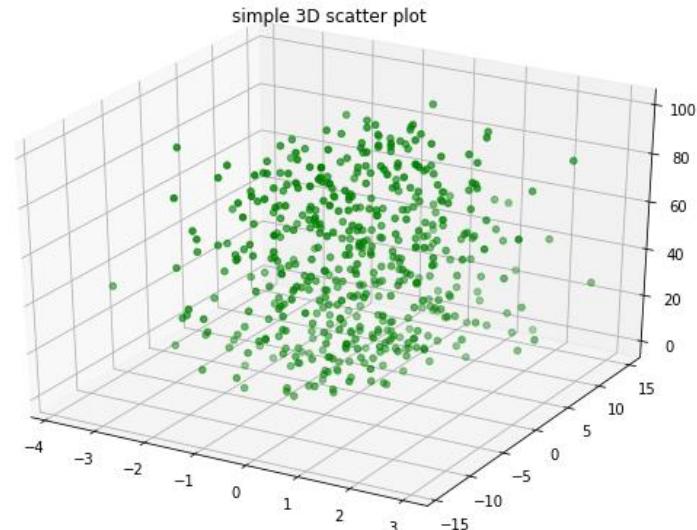
- Scatter plot helps in visualizing 2 numeric variables.
- It helps in identifying the relationship of the data with each variable i.e correlation or trend patterns.
- It also helps in detecting outliers in the plot.



Data Visualization using Matplotlib

3D Scatterplot

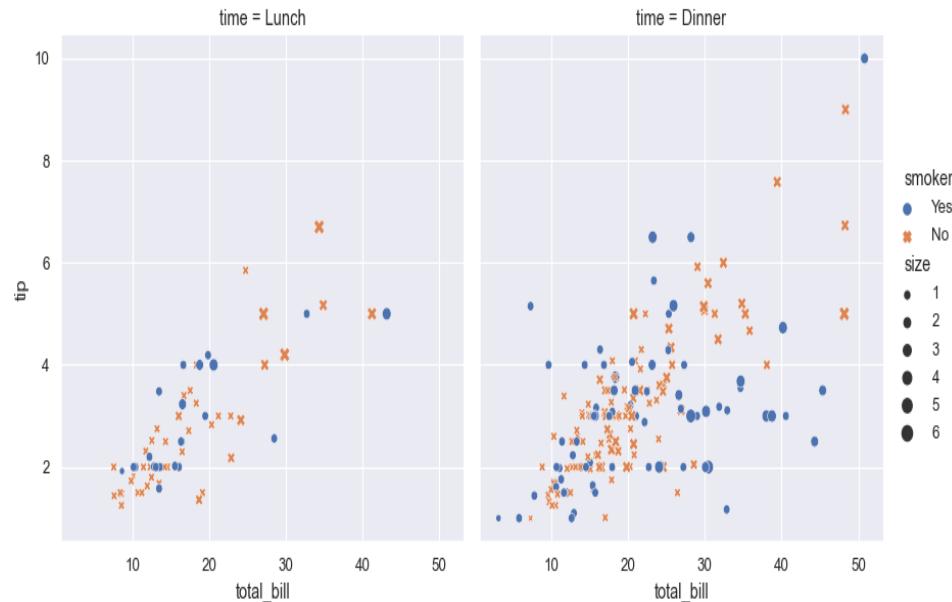
- 3D Scatterplot helps in visualizing 3 numerical variables in a three-dimensional plot.



Advanced data visualization using seaborn

What is Seaborn?

- Matplotlib is the king of Python data visualization libraries and makes it a breeze to explore tabular data visually.
- Seaborn is another Python data visualization library built on top of Matplotlib that introduces some features that weren't previously available, and, in this tutorial, we'll use Seaborn.



Advanced data visualization using seaborn

Installing the libraries and loading the data

We will start by installing the libraries and importing our data. Running the below command will install the Pandas, Matplotlib, and Seaborn libraries for data visualization:

- pip install pandas matplotlib seaborn

```
!pip install jovian --upgrade --quiet
```

```
# installing the upgraded version of the libraries
!pip install pandas matplotlib seaborn --upgrade --quiet
```

```
# importing the libraries
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
%matplotlib inline
```

%matplotlib inline is a special command to ensure that our plots are shown and embedded within the Jupyter notebook itself. Without this command, sometimes plots may

Advanced data visualization using seaborn

Installing the libraries and loading the data

Now, let's import the libraries under their standard aliases:

- import matplotlib.pyplot as plt
- import pandas as pd
- import seaborn as sns

```
!pip install jovian --upgrade --quiet
```

```
# installing the upgraded version of the libraries
!pip install pandas matplotlib seaborn --upgrade --quiet
```

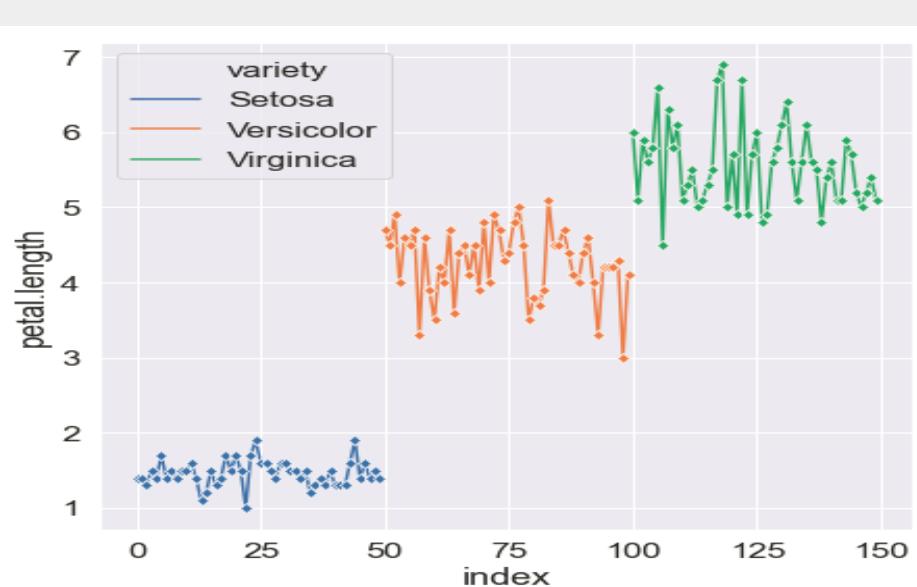
```
# importing the libraries
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
%matplotlib inline
```

%matplotlib inline is a special command to ensure that our plots are shown and embedded within the Jupyter notebook itself. Without this command, sometimes plots may

Advanced data visualization using seaborn

Performing univariate analysis with Seaborn

- The goal of EDA is simple — get to know your dataset at the deepest level possible.
- Becoming intimate with the data and learning its relationships between its variables is an absolute must.

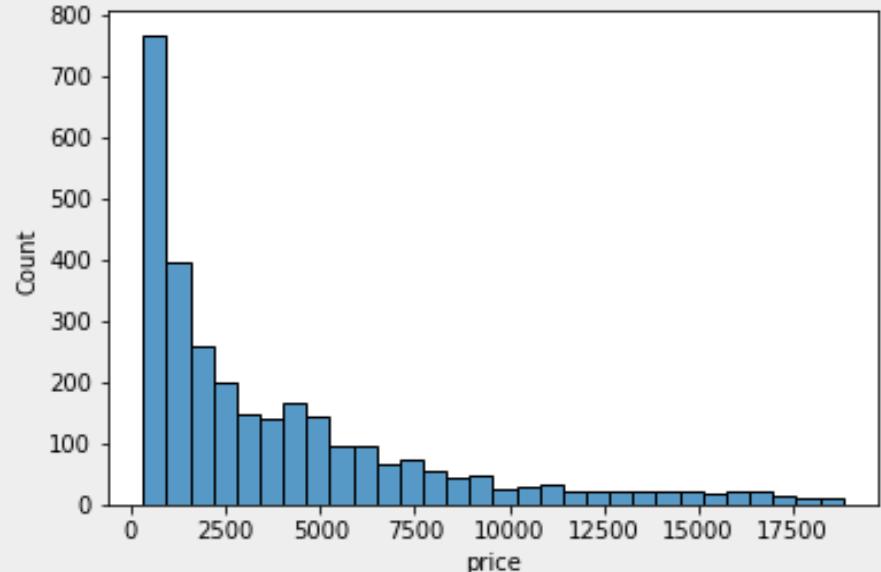


Advanced data visualization using seaborn

Creating histograms in Seaborn

Now, we create our first plot, which is a histogram:

- `sns.histplot(x=sample["price"])`

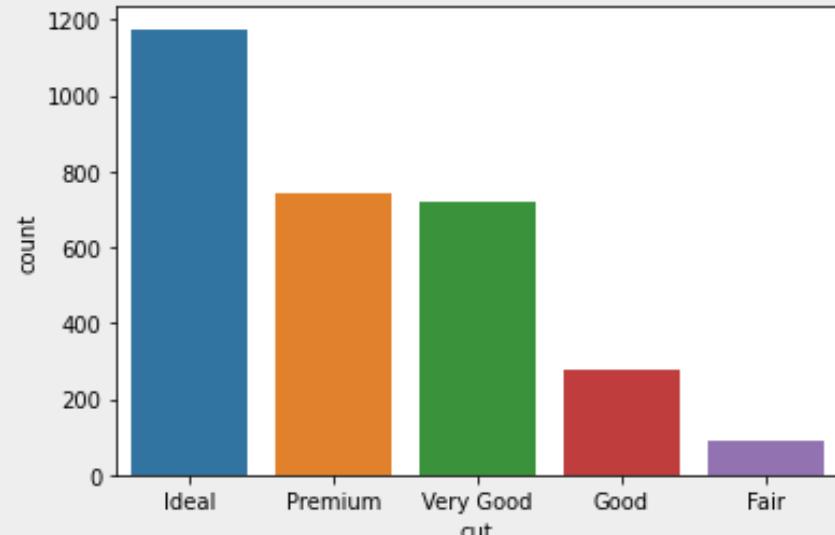


Advanced data visualization using seaborn

Creating count plots in Seaborn

The most common plot for categorical features is a countplot. Passing the name of a categorical feature in our dataset to Seaborn's countplot draws a bar chart, with each bar height representing the number of diamonds in each category.

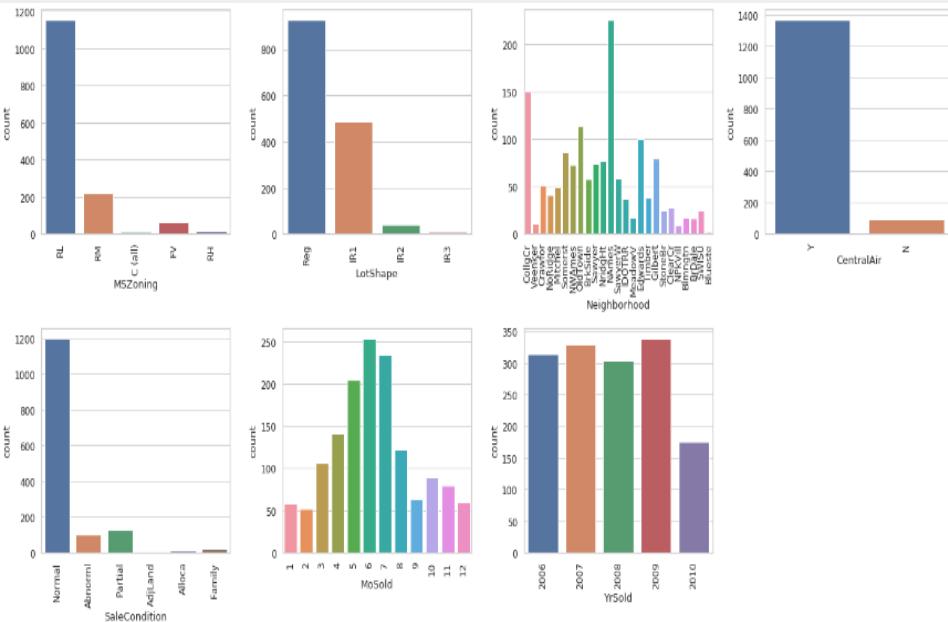
- `sns.countplot(sample["cut"])`



Advanced data visualization using seaborn

Performing bivariate analysis with Seaborn

- Now, let's look at the relationships between two variables at a time.
- Let's start with the connection between diamond carats and price.



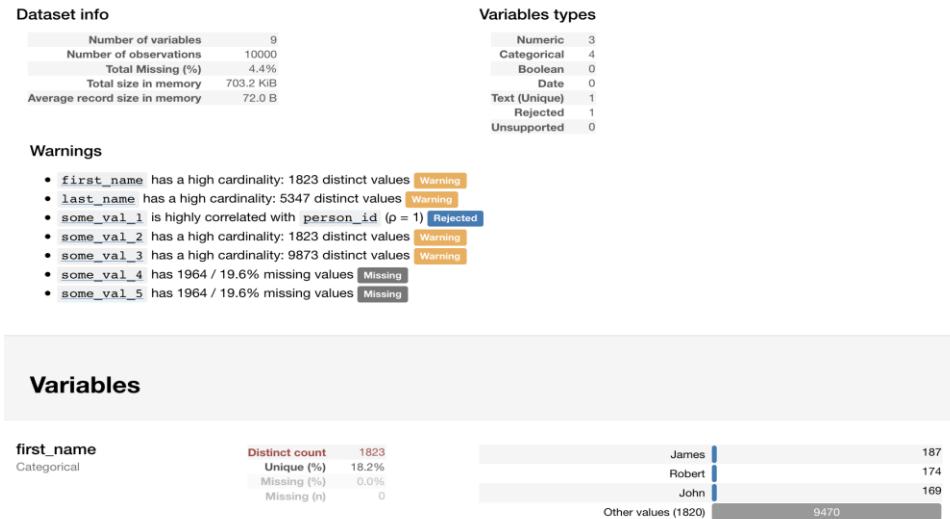
Image

Source: https://miro.medium.com/max/1400/1*cIOI1p56qlHaLAbb6MxrKg.png

Pandas profiling for report generation

What is Pandas Profiling

- Pandas Profiling is an open-source python library, which allows you to do your EDA very quickly.
- By the way, it also generates an interactive HTML report, which you can show to anyone.

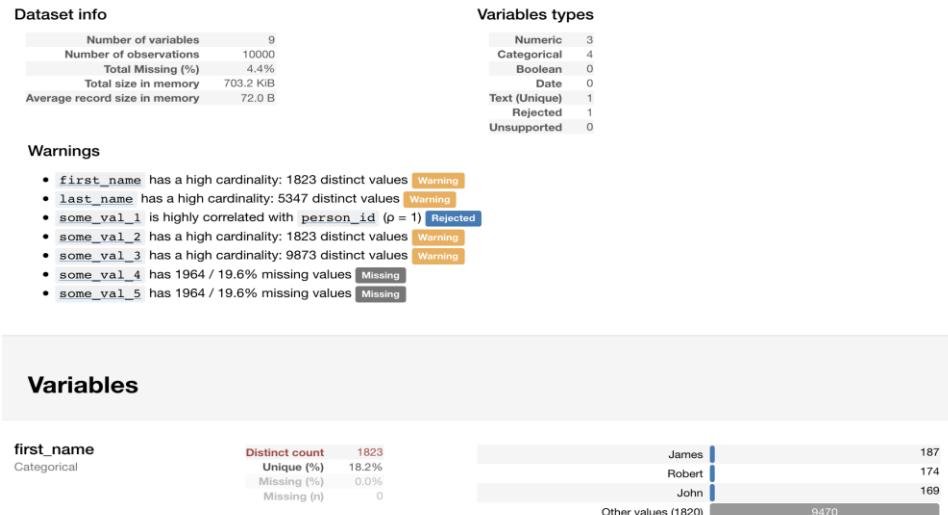


Pandas profiling for report generation

What is Pandas Profiling

These are some of the things you get in your report:

- Type inference
- Essentials
- Quantile statistics
- Descriptive statistics
- Most frequent values
- Histogram



Pandas profiling for report generation

How to install Pandas Profiling

First of all, you need to install the package.

#installing Pandas Profiling

- !pip install <https://github.com/pandas-profiling/pandas-profiling/archive/master.zip> -q



Pandas profiling for report generation

How to install Pandas Profiling

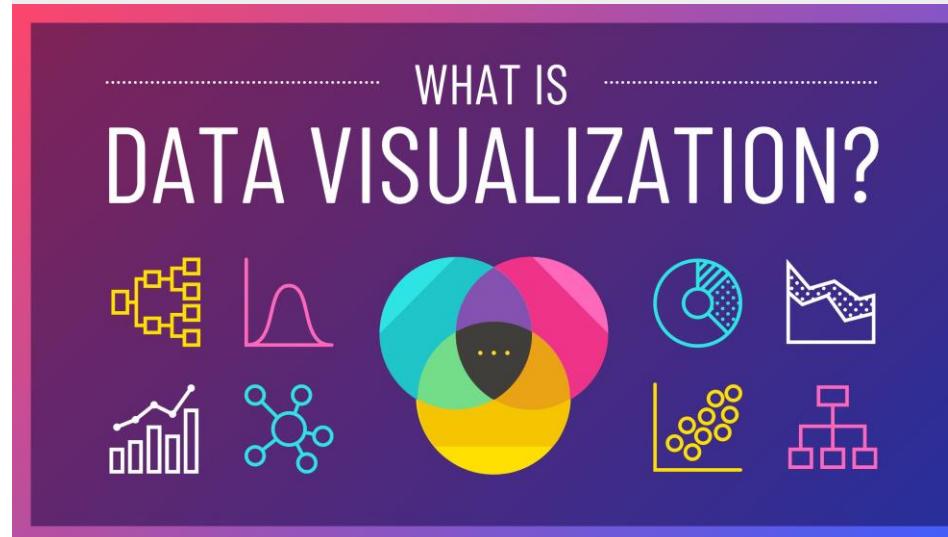
- Then import both pandas and panda_profiling.
- We will be using the Titanic dataset to complete our analysis so import that as well.
- After you import it, you should always take a look at your dataset and then merely link report to it:



Need for data visualization

What is Data Visualization?

- Data visualization gives us a clear idea of what the information means by giving it visual context through maps or graphs.
- This makes the data more natural for the human mind to comprehend and therefore makes it easier to identify trends, patterns, and outliers within large data sets.

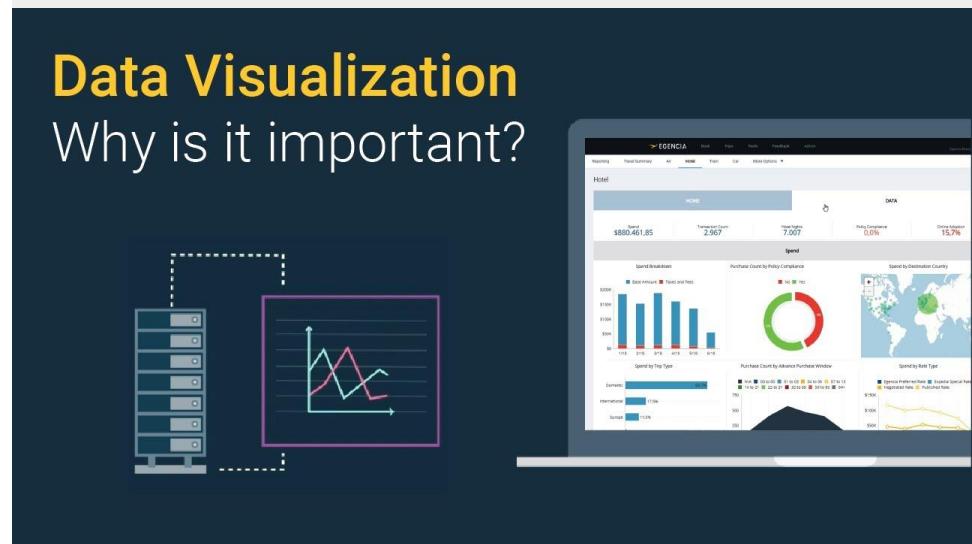


Need for data visualization

Why is Data Visualization Important?

- Data visualization uses visual data to communicate information in a manner that is universal, fast, and effective.
- This practice can help companies identify which areas need to be improved, which factors affect customer satisfaction and dissatisfaction, and what to do with specific products.

Data Visualization Why is it important?



Need for data visualization

What Are The Benefits of Data Visualization?

- Data visualization positively affects an organization's decision-making process with interactive visual representations of data.
- Businesses can now recognize patterns more quickly because they can interpret data in graphical or pictorial forms.

Benefits of Data Visualization



Need for data visualization

Which Data Visualization Techniques are Used?

- There are many different methods of putting together information in a way that the data can be visualized.
- Depending on the data being modeled, and what its intended purpose is, a variety of different graphs and tables may be utilized to create an easy to interpret dashboard.

Data
visualisation
techniques and
tools

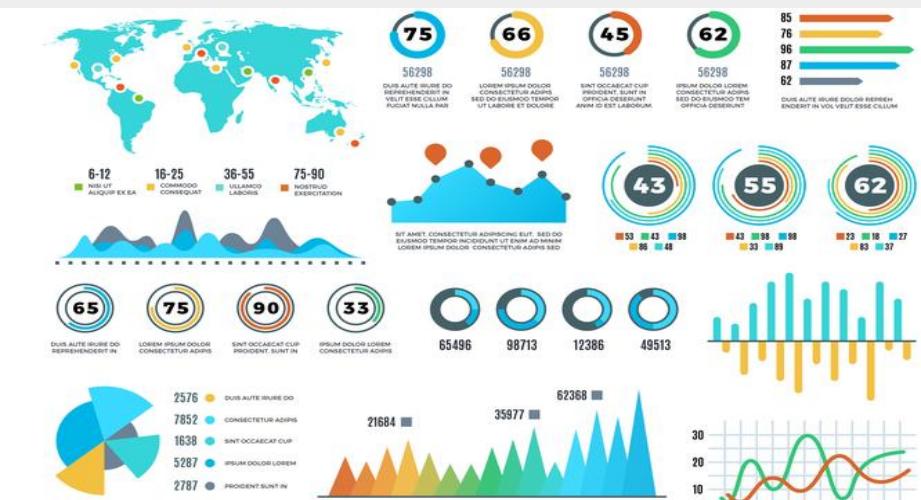
whatagraph



Need for data visualization

Who Uses Data Visualization?

- Data visualization is used across all industries to increase sales with existing customers and target new markets and demographics for potential customers.



Fundamentals of Predictive Analytics using Machine Learning techniques (40 hr)

In this section, we will discuss:

- Machine learning and its types & applications
- Supervised machine learning techniques
- Classification vs. regression
- Understanding Regression and types
- Linear regression using OLS
- Multi-Variate Linear Regression
- Correlation concepts
- Metrics- Loss function, MSE, RMSE, MAE, R2 Score

Machine learning and its types & Applications

Machine learning

AI Approach that uses a system that is capable of learning from experience without having to be programmed.

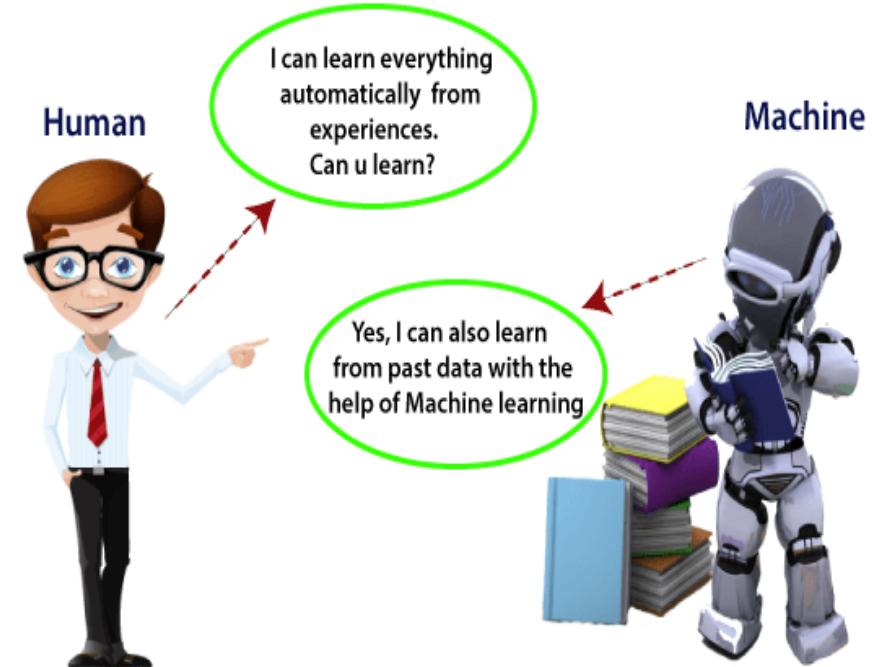


Image Source:

<https://static.javatpoint.com/tutorial/machine-learning/images/introduction-to-machine-learning.png>

Machine learning and its types & Applications

How does Machine Learning Work?

Learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it.

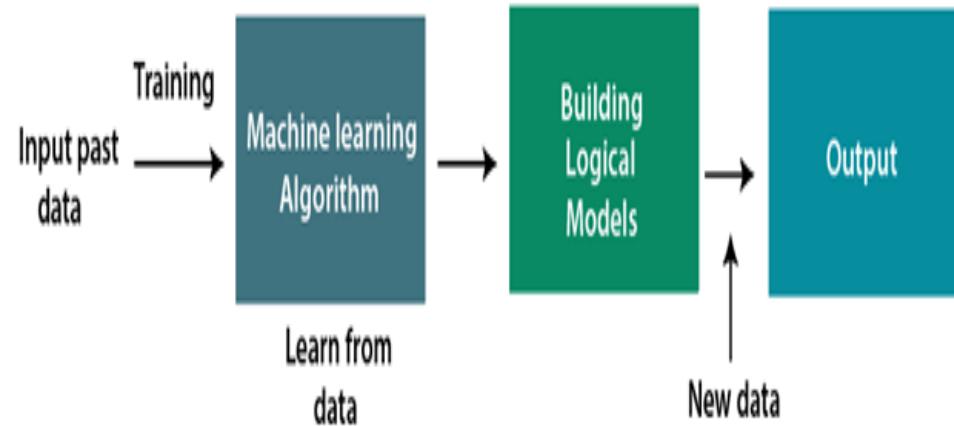


Image Source:

<https://static.javatpoint.com/tutorial/machine-learning/images/introduction-to-machine-learning2.png>

Machine learning and its types & Applications

Applications

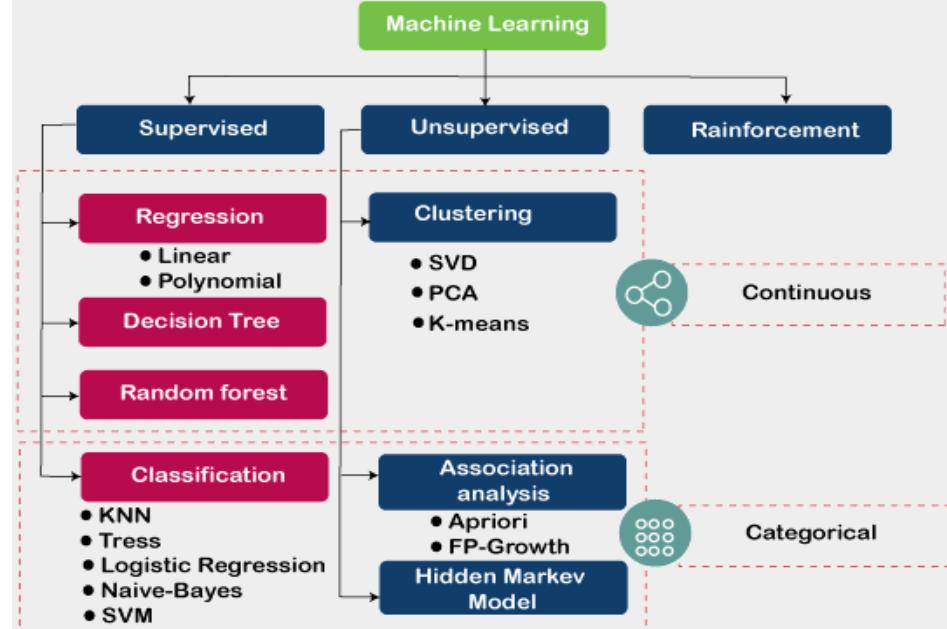
- Image Recognition
- Speech Recognition
- Traffic prediction
- Product recommendations
- Self-driving cars
- Email Spam and Malware Filtering
- Virtual Personal Assistant
- Online Fraud Detection
- Stock Market trading
- Medical Diagnosis
- Automatic Language Translation



Machine learning and its types & Applications

Types

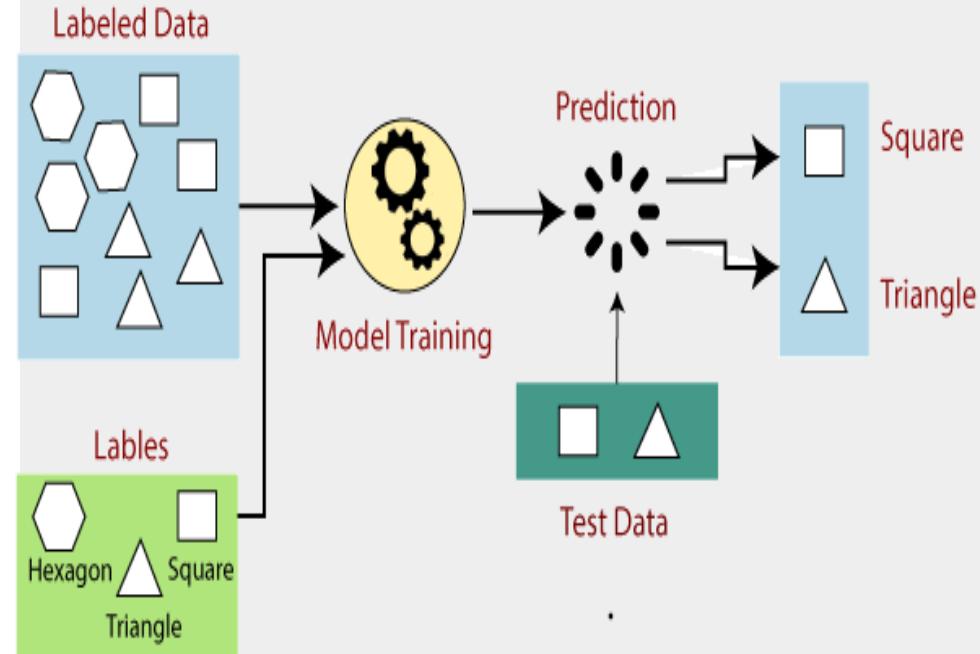
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning



Machine learning and its types & Applications

Supervised Learning

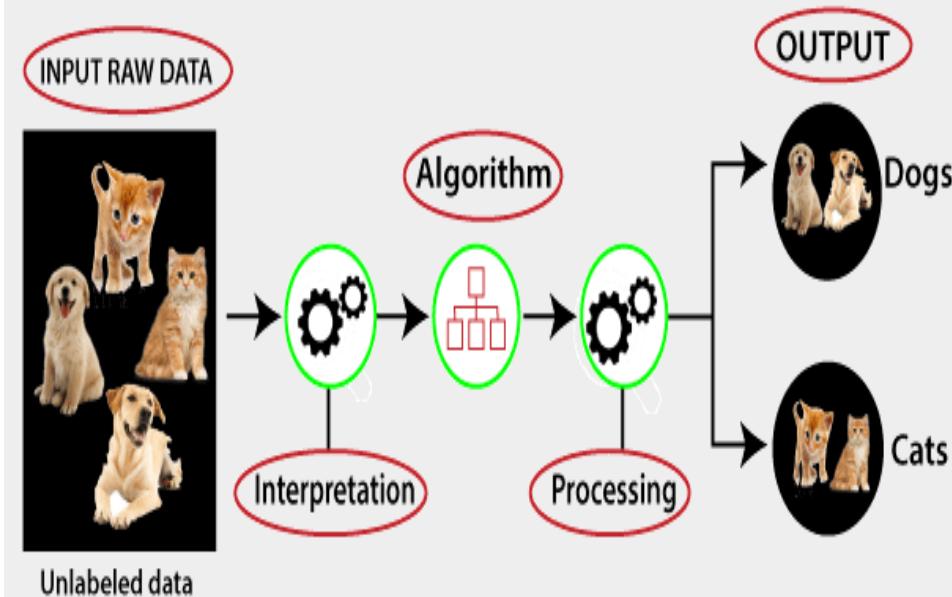
- It is a process of learning algorithm from the training dataset.
- Using input and output variable, an algorithm is used to learn the mapping function from the input to the output.



Machine learning and its types & Applications

Unsupervised Learning

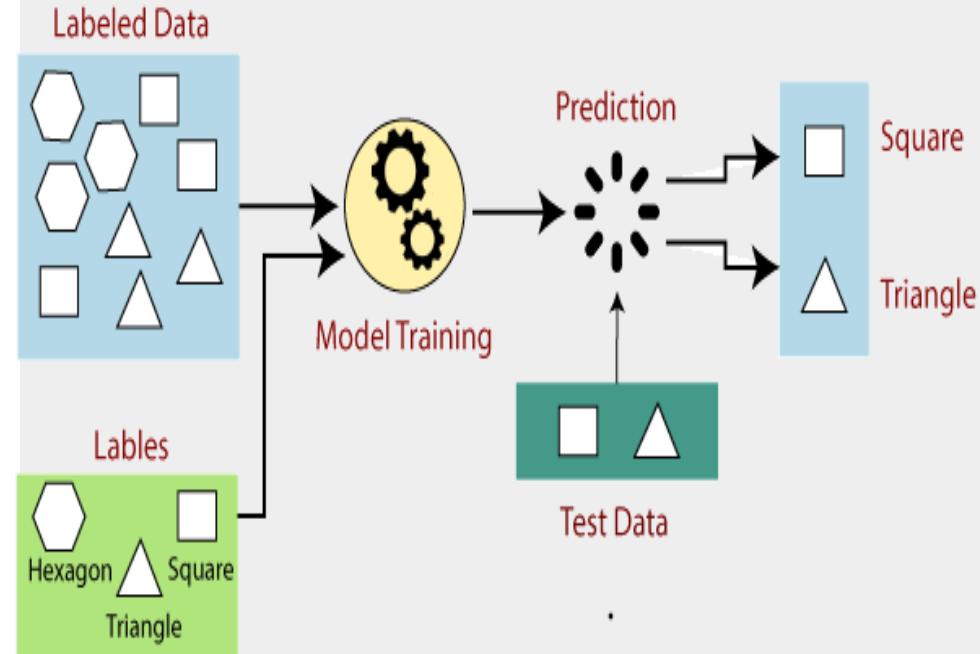
- Modeling the underlying or hidden structure or distribution in the data in order to learn more about the data.
- Only have input data and no corresponding output variables.



Machine learning and its types & Applications

Supervised Learning

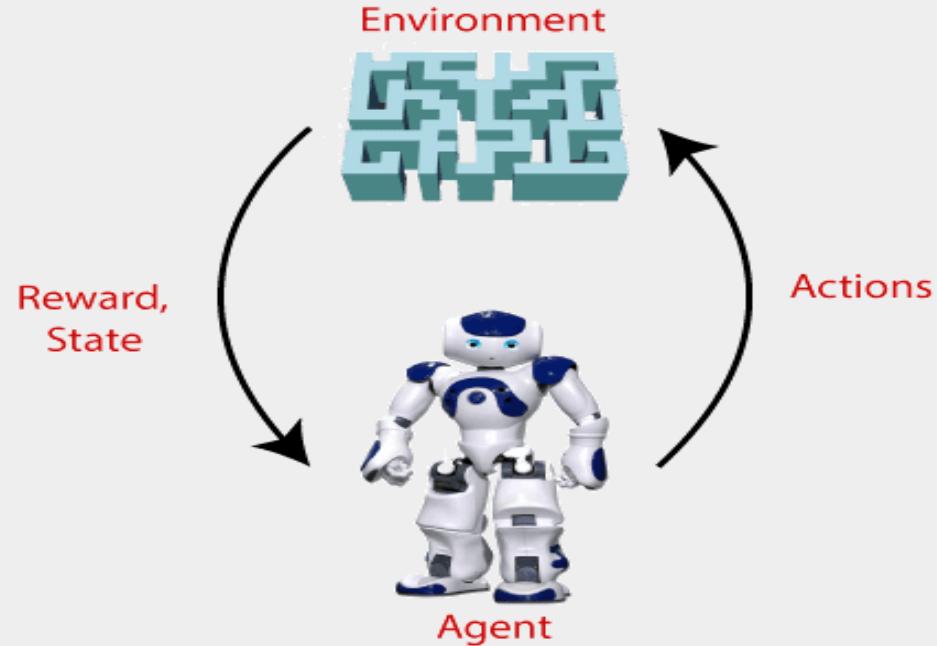
- It is a process of learning algorithm from the training dataset.
- Using input and output variable, an algorithm is used to learn the mapping function from the input to the output.



Machine learning and its types & Applications

Reinforcement Learning

- Model keeps on increasing its performance using a reward feedback to learn the behavior or pattern.
- Markov decision process, Bellman's equation, Q-learning, SARSA (state-Action-Reward-State-Action), Deep Q-network



Supervised machine learning techniques

Classification

- Classification is a process of finding a function which helps in dividing the dataset into classes based on different parameters.
- Output is having discrete value.

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Supervised machine learning techniques

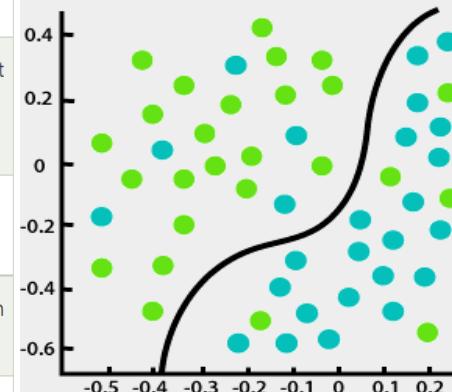
Regression

- Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.
- Output is having continuous value.

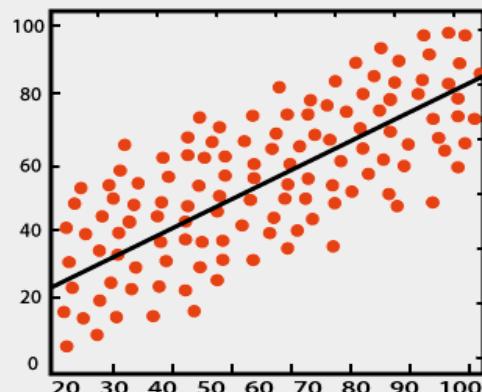
Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Classification vs. Regression

Regression Algorithm	Classification Algorithm
In Regression, the output variable must be of continuous nature or real value.	In Classification, the output variable must be a discrete value.
The task of the regression algorithm is to map the input value (x) with the continuous output variable(y).	The task of the classification algorithm is to map the input value(x) with the discrete output variable(y).
Regression Algorithms are used with continuous data.	Classification Algorithms are used with discrete data.
In Regression, we try to find the best fit line, which can predict the output more accurately.	In Classification, we try to find the decision boundary, which can divide the dataset into different classes.
Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.	Classification Algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.
The regression Algorithm can be further divided into Linear and Non-linear Regression.	The Classification algorithms can be divided into Binary Classifier and Multi-class Classifier.



Classification

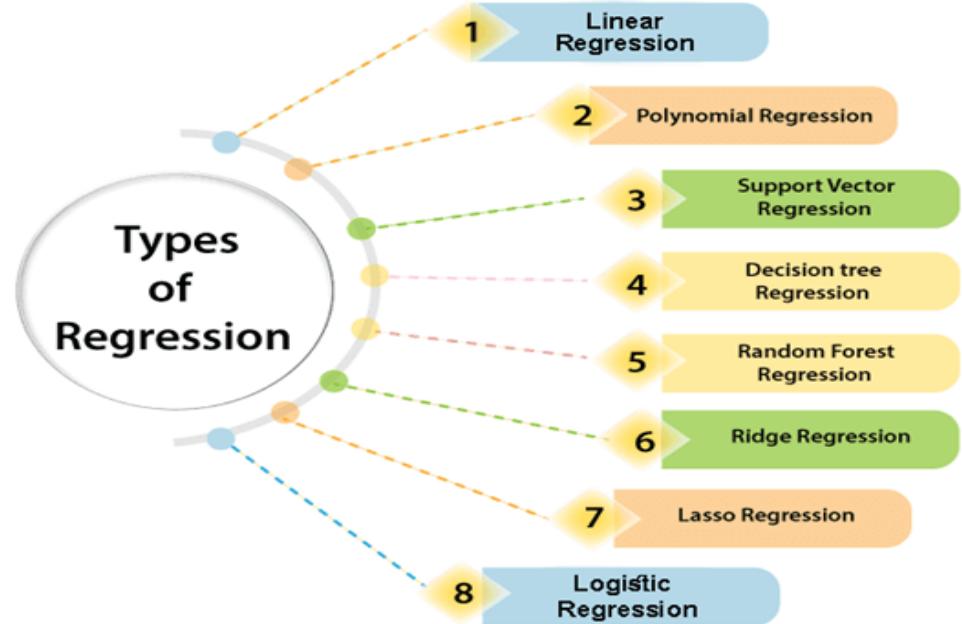


Regression

Understanding Regression and Types

Types

- Linear Regression
- Logistic Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression
- Ridge Regression
- Lasso Regression:



Linear regression using OLS

Ordinary Least Squares

- Ordinary least squares, or linear least squares, estimates the parameters in a regression model by minimizing the sum of the squared residuals.
- This method draws a line through the data points that minimizes the sum of the squared differences between the observed values and the corresponding fitted values.

$$m = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$
$$b = \bar{y} - m * \bar{x}$$

x = independent variables

\bar{x} = average of independent variables

y = dependent variables

\bar{y} = average of dependent variables

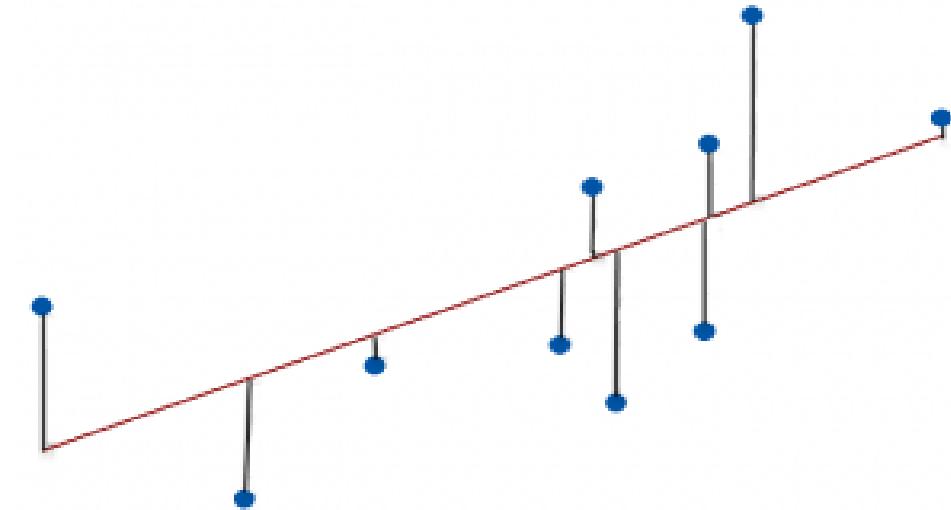
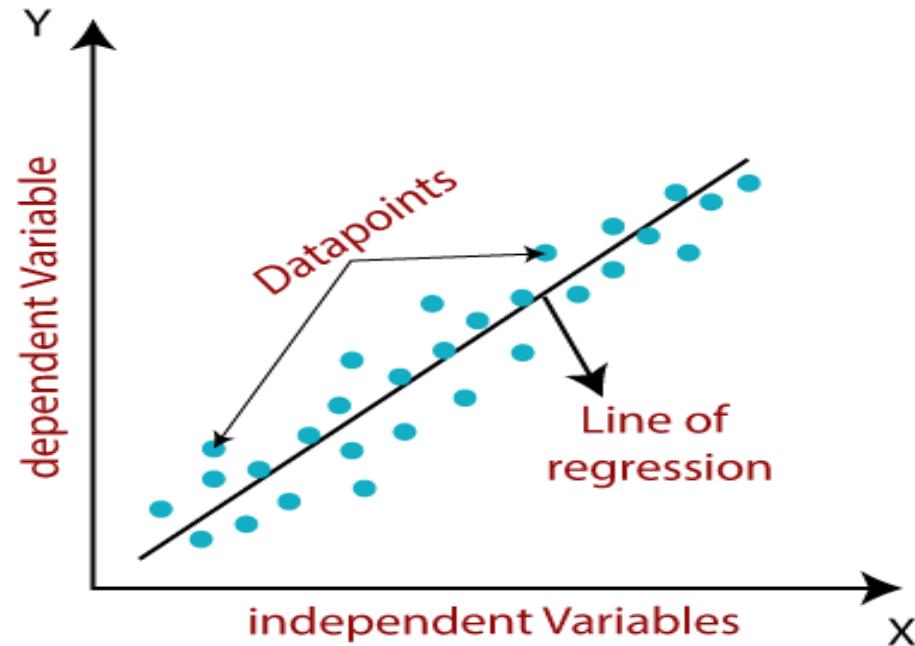


Image Source: <https://statisticsbyjim.com/glossary/ordinary-least-squares/#:~:text=Ordinary%20least%20squares%2C%20or%20linear.and%20the%20corresponding%20fitted%20values>

Understanding Regression and Types

Linear Regression

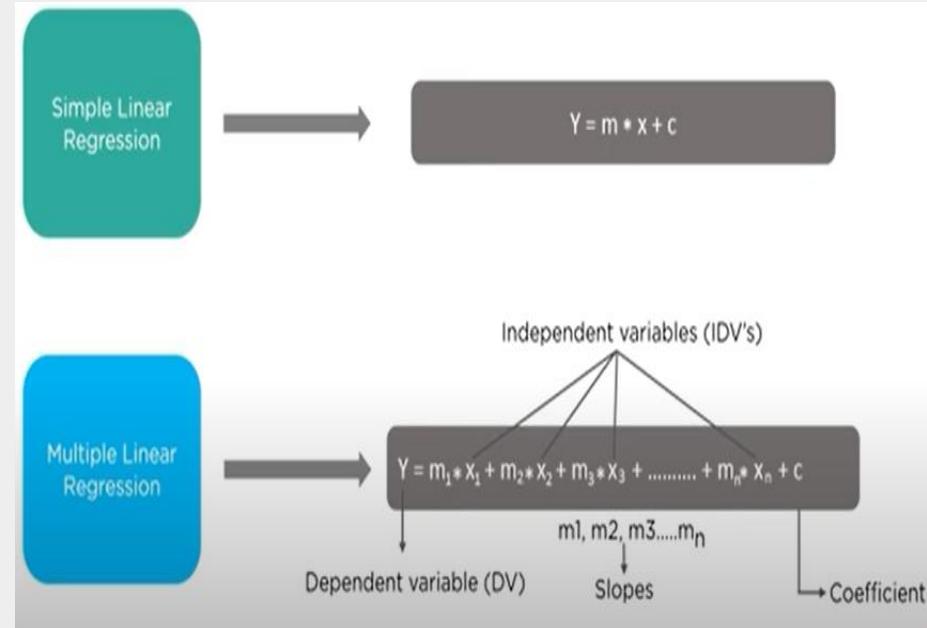
- Relationship between a dependent (y) and one or more independent (x) variables.



Understanding Regression and Types

Linear and Multiple Regression

- Multiple regression is a broader class of regressions that encompasses linear and nonlinear regressions with multiple explanatory variables.
- Whereas linear regress only has one independent variable impacting the slope of the relationship, multiple regression incorporates multiple independent variables.

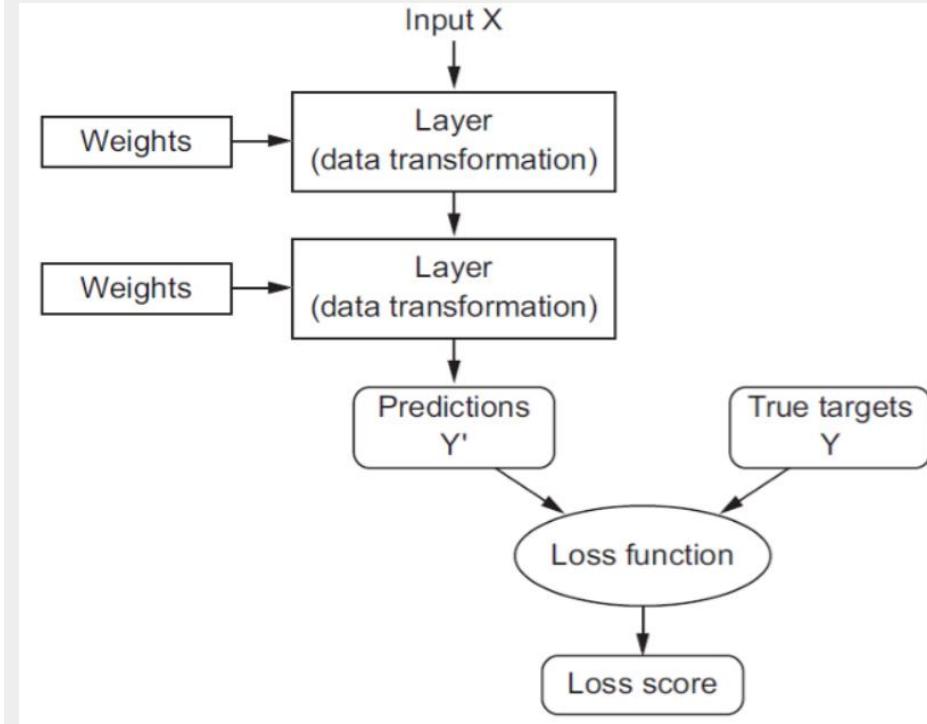


Metrics

Loss Function

A function that calculates loss for 1 data point is called the loss function.

$$\text{Squared Error} = (y_i - \hat{y}_i)^2$$



Metrics

Mean Squared Error (MSE) / Mean Squared Deviation (MSD)

- It basically calculates the difference between the estimated and the actual value, squares these results and then computes their average.
- MSE can only assume non-negative values .
- $\hat{y}_i \rightarrow$ Predicted, $y_i \rightarrow$ Actual

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

Metrics

Root Mean Squared Error (RMSE) / Root Mean Squared Deviation (RMSD)

- RMSE calculates the average of the squared errors across all samples but, in addition, takes the square root of the result.

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

Metrics

Mean Absolute Error (MAE)

- It simply calculates the absolute value of the errors and then takes the average of these values.

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n}$$

Metrics

R Squared (R^2) / Coefficient of Determination

- R Squared (R^2) represents the proportion of the variance for the dependent variable y that's explained by the independent variables X .
- R^2 explains to what extent the variance of one variable explains the variance of the second variable
- If the R^2 of a model is 0.75, then approximately 75% of the observed variation can be explained by the model's features..

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

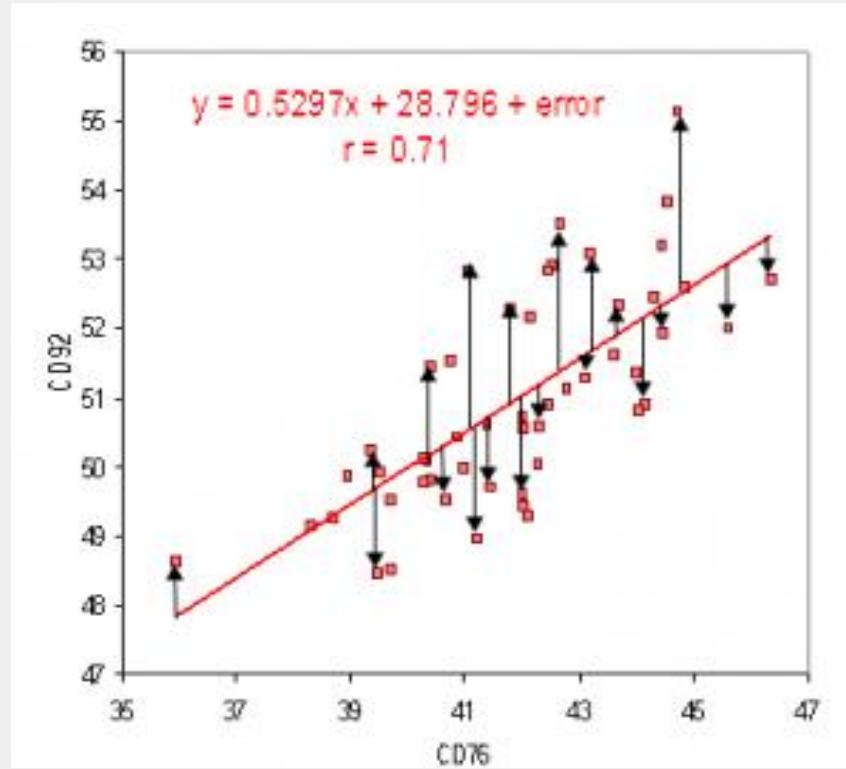
In this section, we will discuss:

- Residuals in Regression
- Polynomial features
- Classification techniques
- Types of distance metrics
- KNN Classification
- Gradient Decent

Residuals in Regression

What is it?

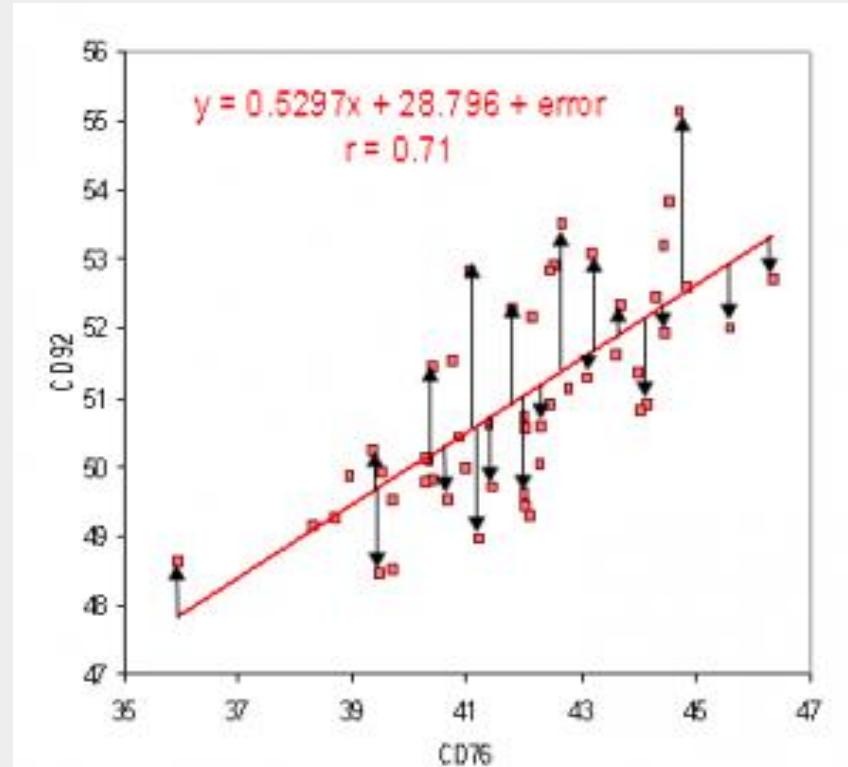
- A residual is the vertical distance between a data point and the regression line.
- Each data point has one residual.



Residuals in Regression

Types of residual

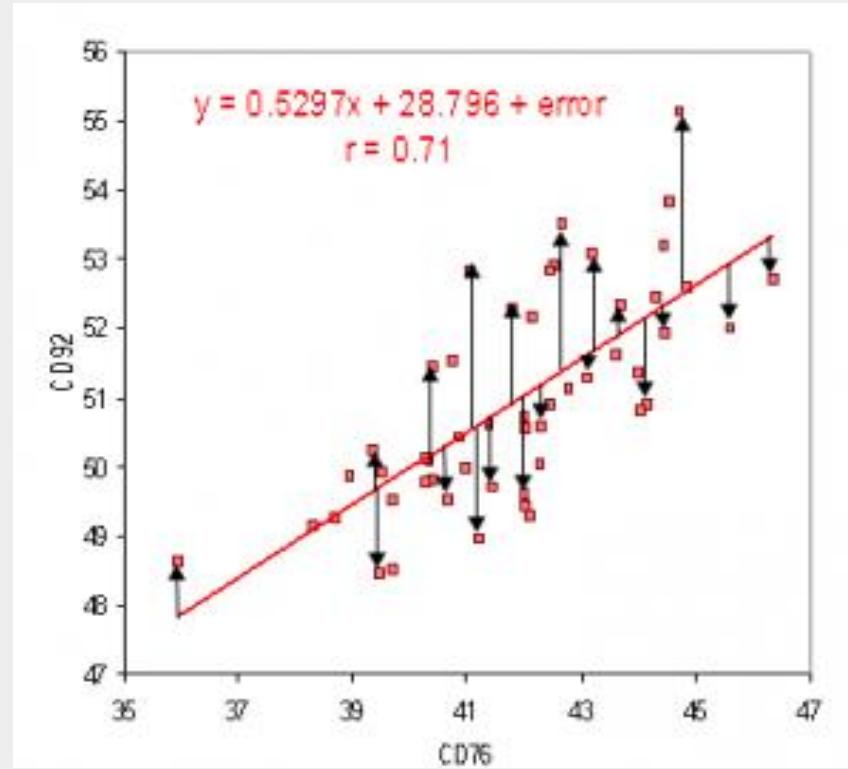
- Positive if they are above the regression line,
- Negative if they are below the regression line,
- Zero if the regression line actually passes through the point



Residuals in Regression

More about residual

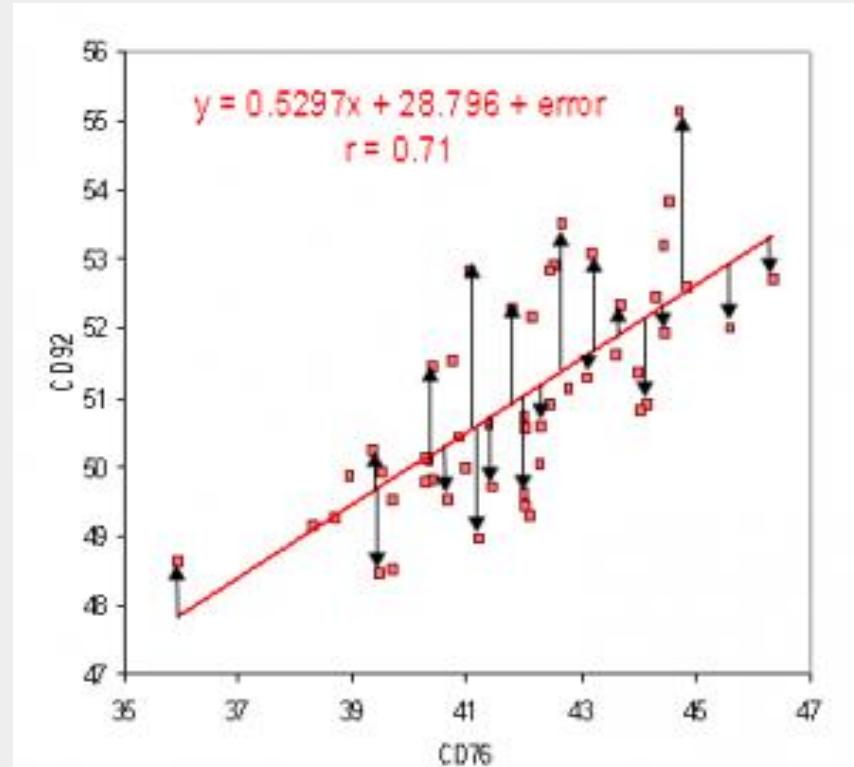
- As residuals are the difference between any data point and the regression line, they are sometimes called “errors.”
- In other words, the residual is the error that isn’t explained by the regression line.



Residuals in Regression

More about residual

- The residual(e) can also be expressed with an equation.
- Residual =
Observed value – predicted value
$$e = y - \hat{y}$$



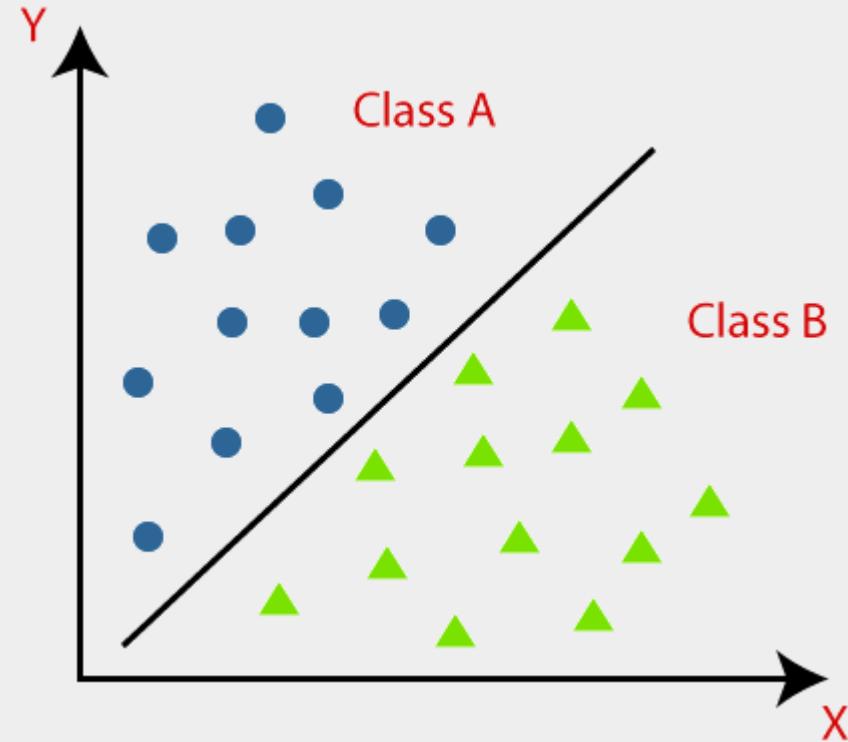
Polynomial features

- Polynomial features are those features created by raising existing features to an exponent.
- Polynomial features are a type of feature engineering, e.g. the creation of new input features based on the existing features.
- The “degree” of the polynomial is used to control the number of features added, e.g. a degree of 3 will add two new variables for each input variable.

Classification techniques

What Is Classification?

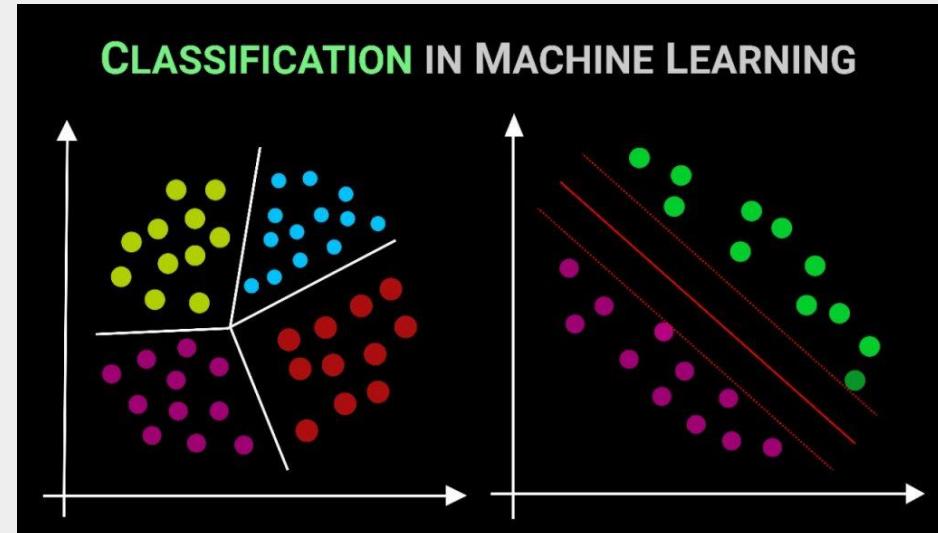
- Classification is the process of recognizing, understanding, and grouping ideas and objects into preset categories or “sub-populations.”
- Classification algorithms in machine learning use input training data to predict the likelihood that subsequent data will fall into one of the predetermined categories.



Classification techniques

Popular Classification Algorithms

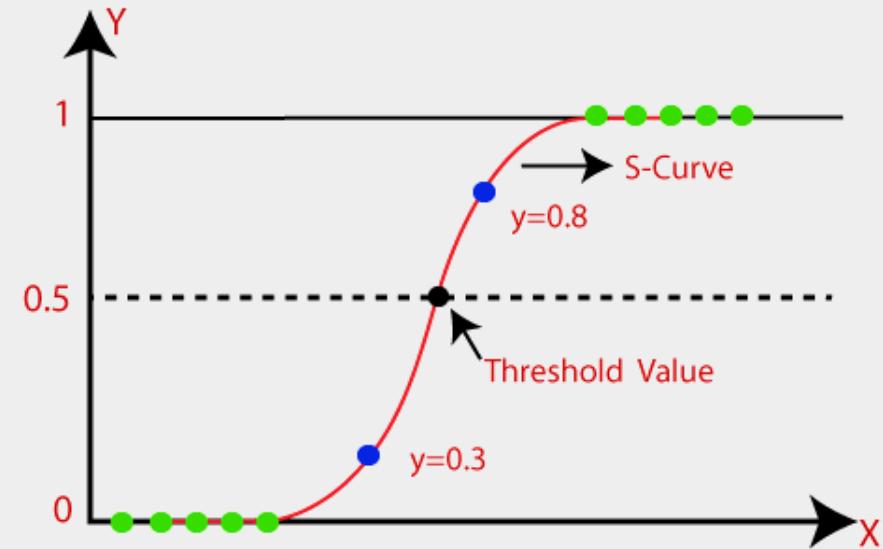
- Logistic Regression
- Naive Bayes
- K-Nearest Neighbors
- Decision Tree
- Support Vector Machines



Classification techniques

Logistic Regression

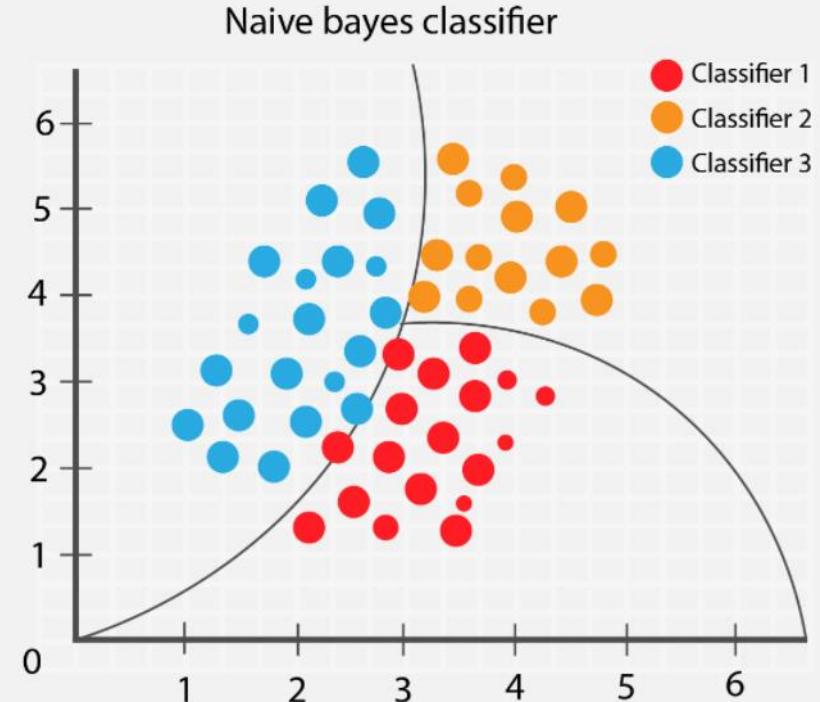
- Logistic Regression is a calculation used to predict a binary outcome: either something happens, or does not.
- This can be exhibited as Yes/No, Pass/Fail, Alive/Dead, etc.



Classification techniques

Naive Bayes

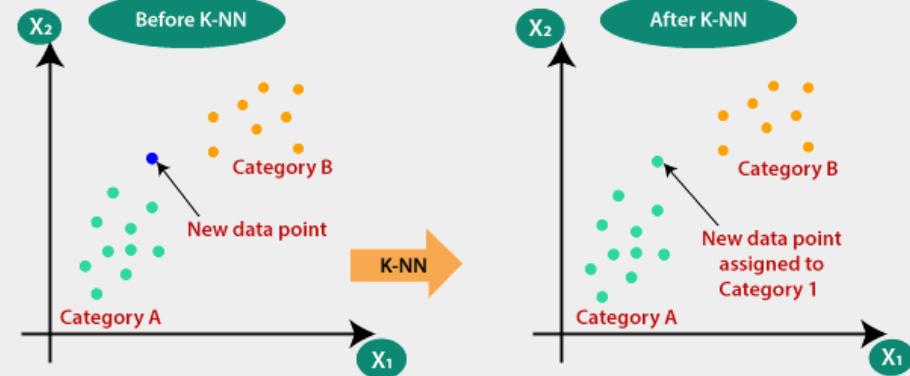
- Naive Bayes calculates the possibility of whether a data point belongs within a certain category or does not.
- In text analysis, it can be used to categorize words or phrases as belonging to a preset “tag” (classification) or not.



Classification techniques

K-nearest Neighbors

- K-nearest neighbors (k-NN) is a pattern recognition algorithm that uses training datasets to find the k closest relatives.
- When k-NN is used in classification, you calculate to place data within the category of its nearest neighbour.



Classification techniques

Decision Tree

- A decision tree is a supervised learning algorithm that is able to order classes on a precise level.
- It works like a flow chart, separating data points into two similar categories at a time from the “tree trunk” to “branches,” to “leaves,” where the categories become more finitely similar.
- This creates categories within categories, allowing for organic classification with limited human supervision.

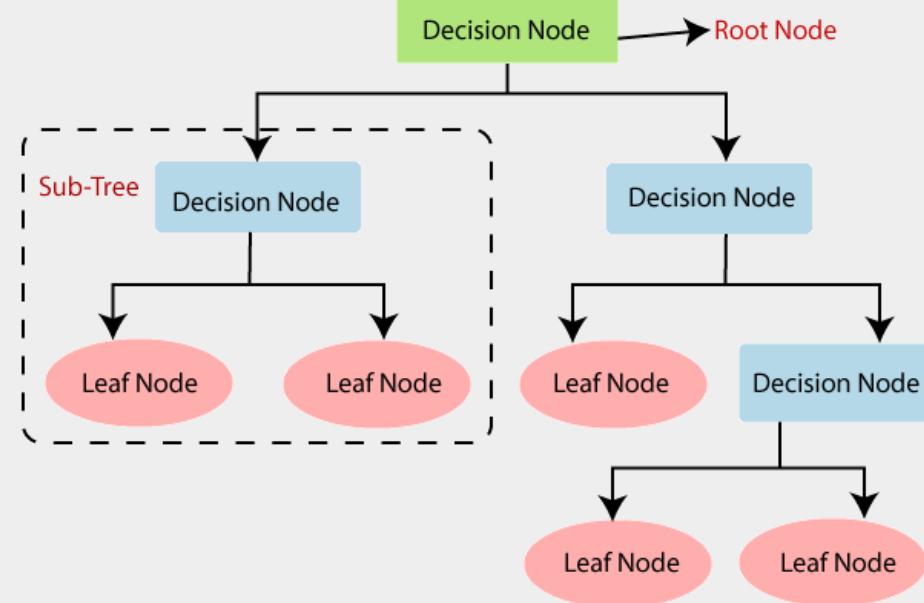
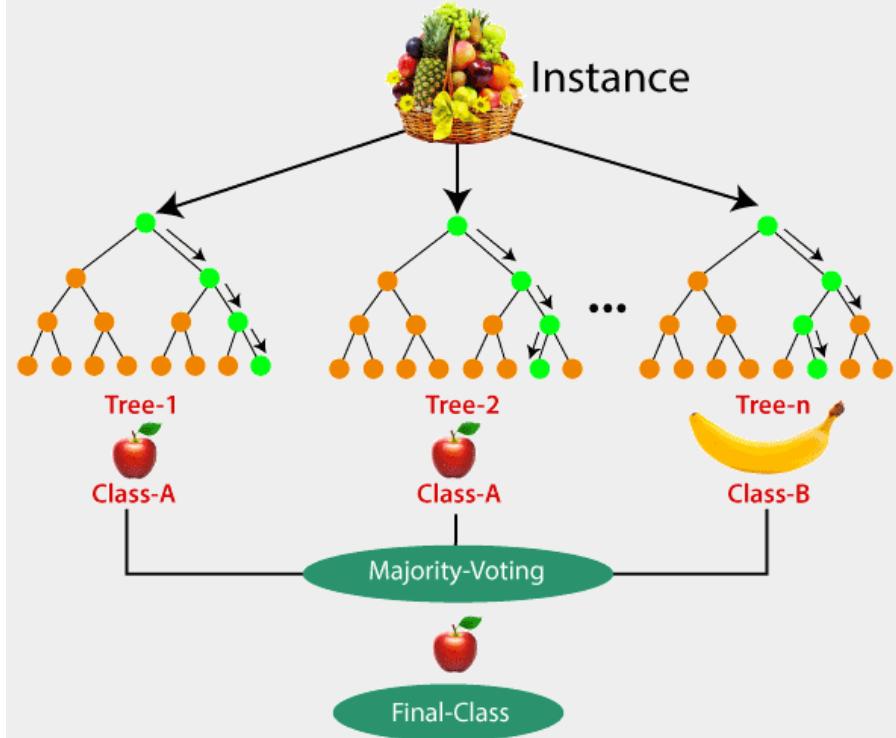


Image Source: <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>

Classification techniques

Random Forest

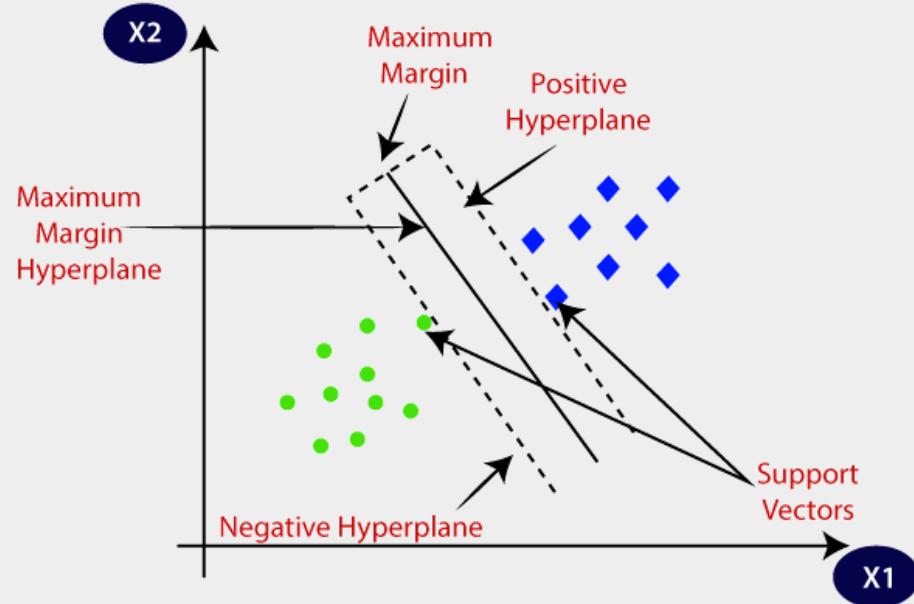
- The random forest algorithm is an expansion of decision tree
- You first construct a multitude of decision trees with training data, then fit your new data within one of the trees as a “random forest.”



Classification techniques

Support Vector Machines

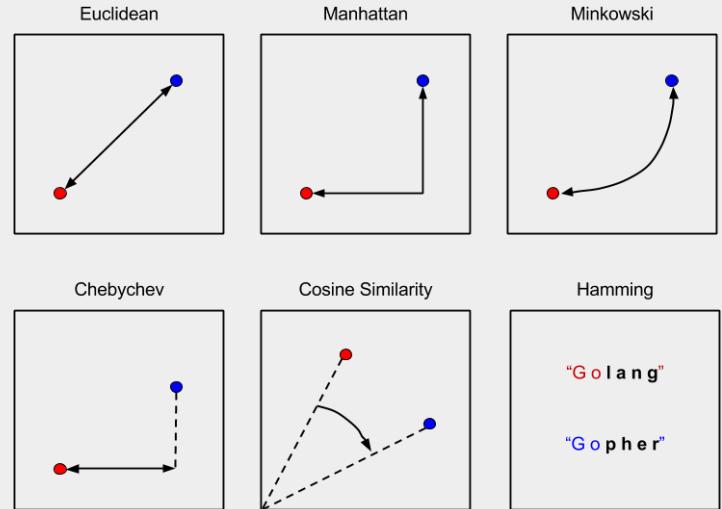
- A support vector machine (SVM) uses algorithms to train and classify data within degrees of polarity
- The SVM then assigns a hyperplane that best separates the tags.
- In two dimensions this is simply a line.



Distance metrics

What is it?

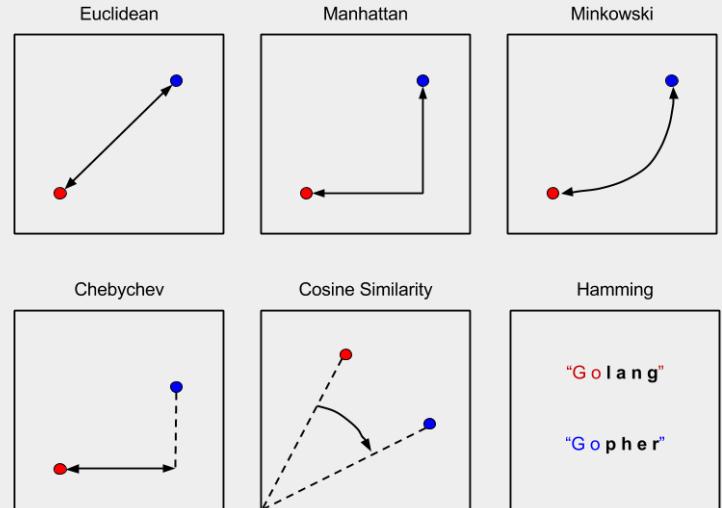
- A distance measure is an objective score that summarizes the relative difference between two objects in a problem domain.
- Different distance measures must be chosen and used depending on the types of the data.



Distance metrics

Types

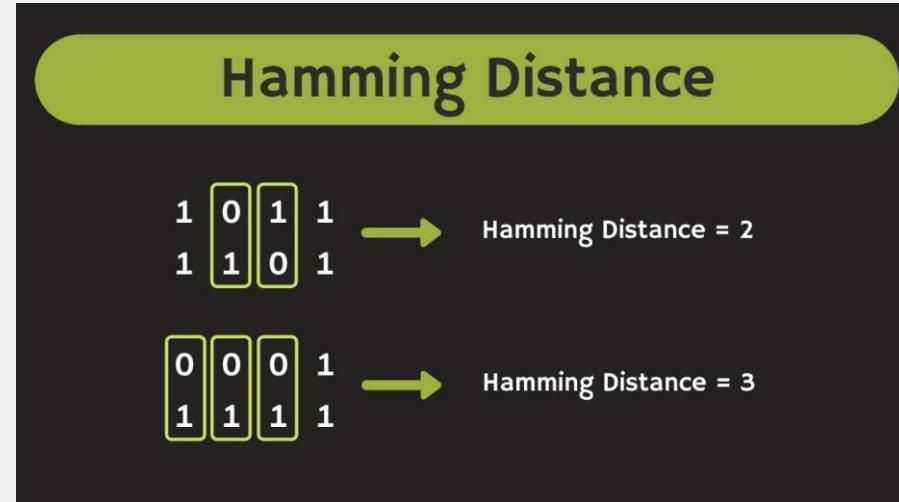
- Following are the 4 most commonly used distance measures in machine learning:
- Hamming Distance
- Euclidean Distance
- Manhattan Distance
- Minkowski Distance



Distance metrics

Hamming Distance

- Hamming distance calculates the distance between two binary vectors, also referred to as binary strings or bitstrings.

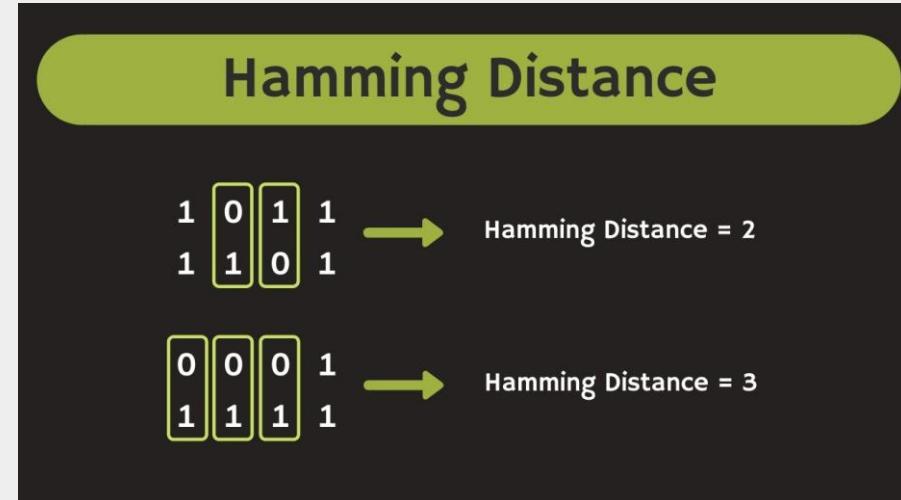


Distance metrics

Hamming Distance

- Example :
- Suppose there are two strings 1101 1001 and 1001 1101.

$11011001 \oplus 10011101 = 01000100$. Since, this contains two 1s, the Hamming distance, $d(11011001, 10011101) = 2$.



Distance metrics

Euclidean Distance

- Euclidean distance calculates the distance between two real-valued vectors.
- Euclidean distance is calculated as the square root of the sum of the squared differences between the two vectors.
- $\text{EuclideanDistance} = \sqrt{\sum_{i=1}^N (v1[i] - v2[i])^2}$

Euclidean Distance

$$\text{Euclidean}(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$



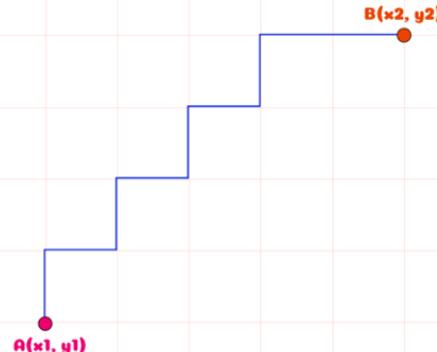
Distance metrics

Manhattan Distance

- The Manhattan distance, also called the Taxicab distance or the City Block distance, calculates the distance between two real-valued vectors.
- The taxicab name for the measure refers to the intuition for what the measure calculates: the shortest path that a taxicab would take between city blocks (coordinates on the grid).

Manhattan Distance

$$\text{Manhattan}(A, B) = |x_1 - x_2| + |y_1 - y_2|$$



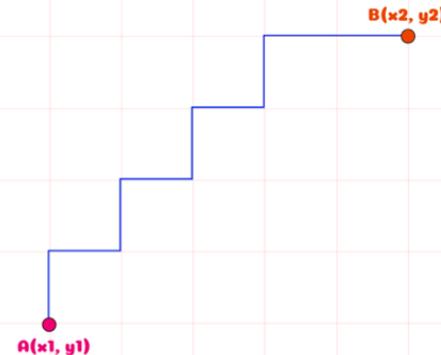
Distance metrics

Manhattan Distance

- Manhattan distance is calculated as the sum of the absolute differences between the two vectors.
- The Manhattan Distance between two points (X_1, Y_1) and (X_2, Y_2) is given by $|X_1 - X_2| + |Y_1 - Y_2|$.

Manhattan Distance

$$\text{Manhattan}(A, B) = |x_1 - x_2| + |y_1 - y_2|$$



Distance metrics

Minkowski Distance

- Minkowski distance calculates the distance between two real-valued vectors.
- It is a generalization of the Euclidean and Manhattan distance measures and adds a parameter, called the “order” or “ p ”, that allows different distance measures to be calculated.

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$\text{dist} = \left(\sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k th attributes (components) or data objects p and q .

Distance metrics

Minkowski Distance

- For example:

Given two vectors, vect1 as (4, 2, 6, 8) and vect2 as (5, 1, 7, 9). Their Minkowski distance for $p = 2$ is given by, $(|4 - 5|^2 + |2 - 1|^2 + |6 - 7|^2 + |8 - 9|^2)^{1/2}$ which is equal to 2.

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$\text{dist} = \left(\sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k th attributes (components) or data objects p and q .

KNN Classification

Concept

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

KNN Classifier



Image source: <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>

KNN Classification

Concept

- K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN Classifier



Image source: <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>

KNN Classification

Why do we need a K-NN Algorithm?

- With the help of K-NN, we can easily identify the category or class of a particular dataset.
- Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x_1 , so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm.

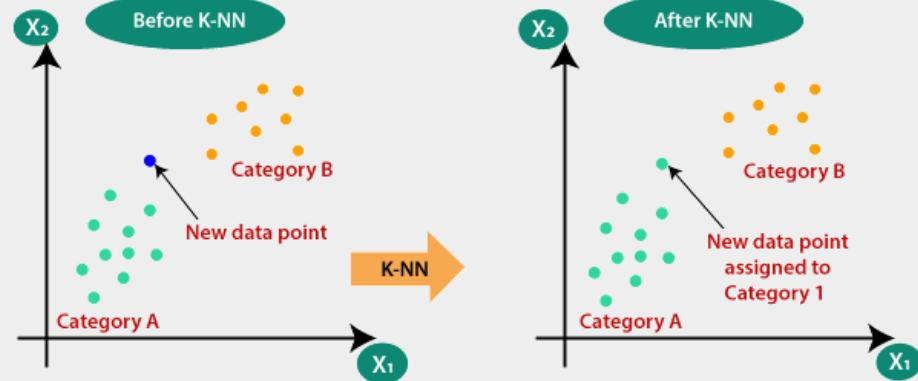
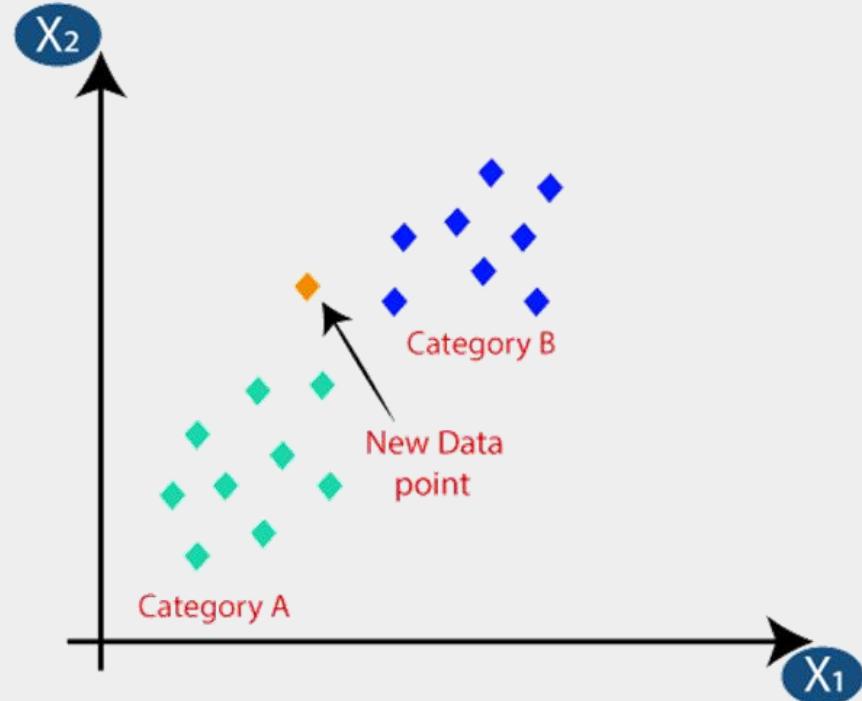


Image Source: <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>

KNN Classification

How does K-NN work?

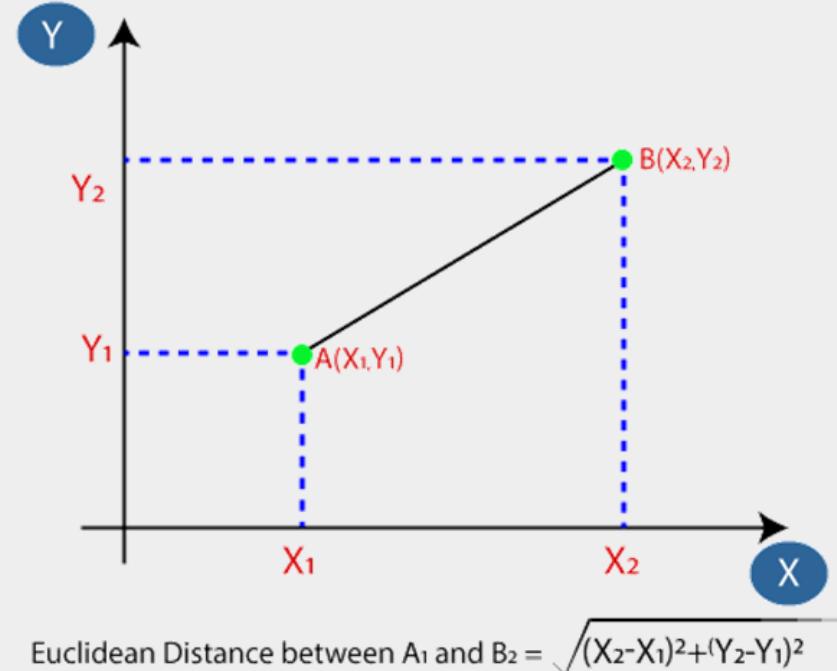
- Step-1: Select the number K of the neighbors
- Step-2: Calculate the Euclidean distance of K number of neighbors



KNN Classification

How does K-NN work?

- Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
- Step-4: Among these k neighbors, count the number of the data points in each category.



KNN Classification

How does K-NN work?

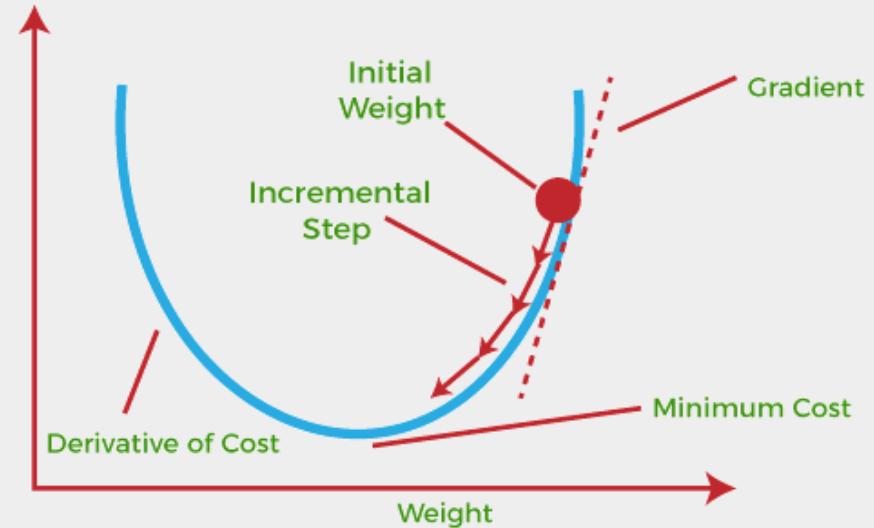
- Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
- Step-6: Our model is ready.



Gradient Descent

Concept

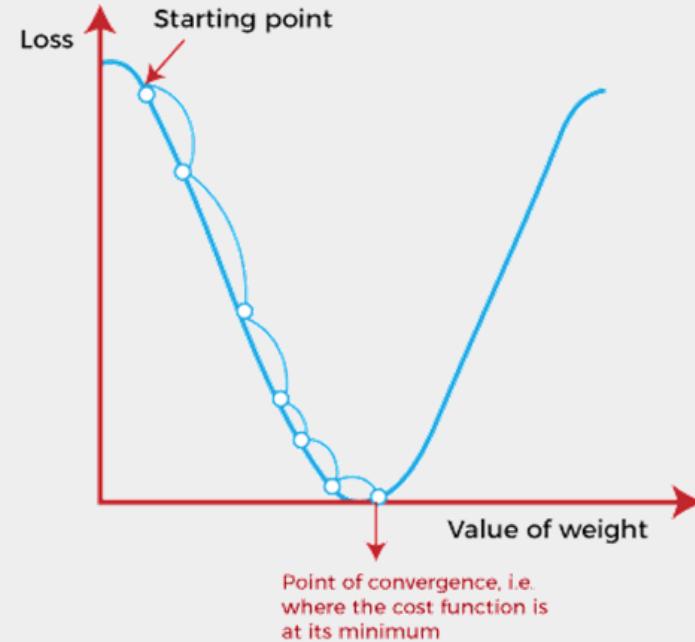
- Gradient descent (GD) is an iterative first-order optimisation algorithm used to find a local minimum/maximum of a given function.
- This method is commonly used in machine learning (ML) and deep learning(DL) to minimise a cost/loss function (e.g. in a linear regression).



Gradient Descent

How does Gradient Descent work?

- The equation for simple linear regression is given as:
- $Y = mX + c$
- Where ' m ' represents the slope of the line, and ' c ' represents the intercepts on the y-axis.



Gradient Descent

What is Cost-function?

- The cost function is defined as the measurement of difference or error between actual values and expected values at the current position and present in the form of a single real number.
- It helps to increase and improve machine learning efficiency by providing feedback to this model so that it can minimize error and find the local or global minimum.

COST FUNCTION IN MACHINE LEARNING

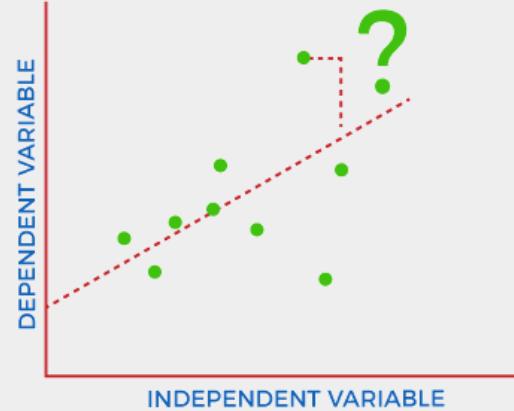


Image Source:: <https://www.javatpoint.com/cost-function-in-machine-learning>

Gradient Descent

The Gradient Descent Algorithm

- Gradient Descent method's steps are:
 1. Choose a starting point (initialisation)
 2. Calculate gradient at this point
 3. Make a scaled step in the opposite direction to the gradient (objective: minimise)
 4. Repeat points 2 and 3 until one of the criteria is met:
 5. Maximum number of iterations reached
 6. Step size is smaller than the tolerance.

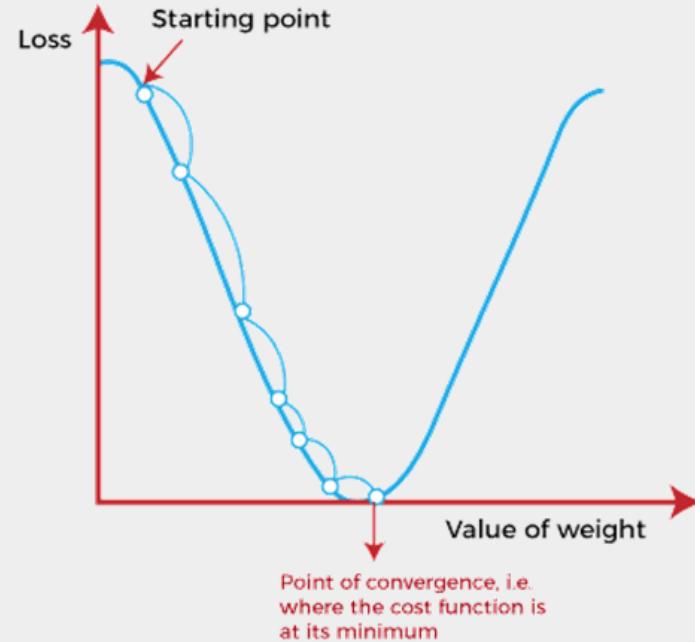


Image Source:: <https://www.javatpoint.com/gradient-descent-in-machine-learning>

In this section, we will discuss:

- Logistic Regression
- Evaluation – Confusion Matrix, Precision, Recall, F1 Score, Accuracy
- Python Library – Sci-kit Learn

Logistic Regression

What is Logistic Regression?

- Logistic regression is a **statistical model that in its basic form uses a logistic function to model a binary dependent variable**, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

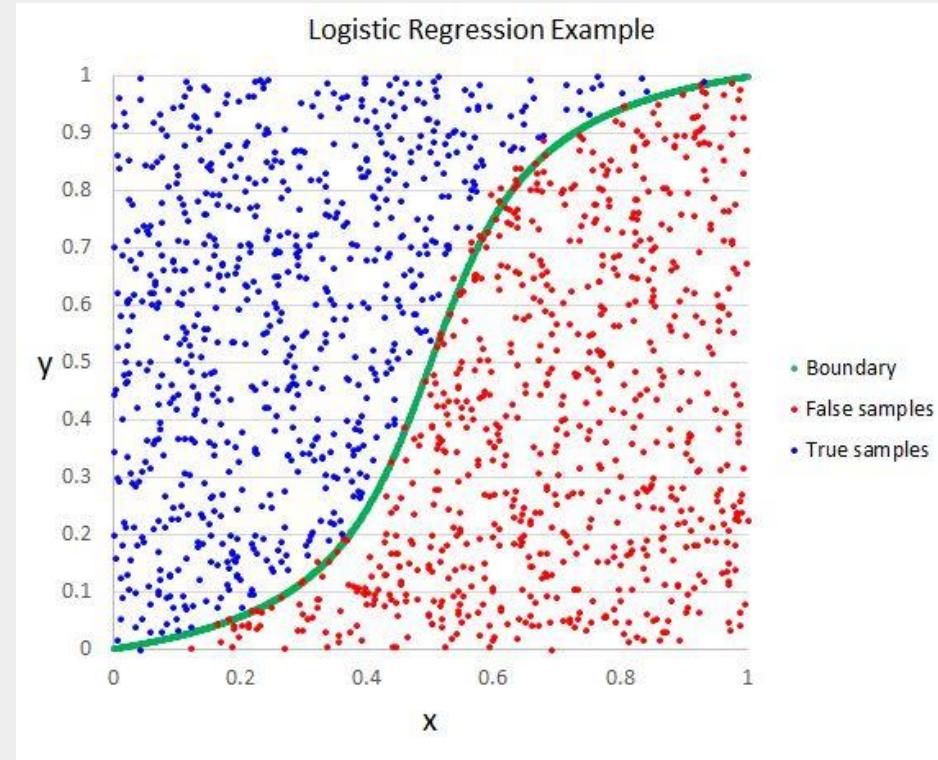
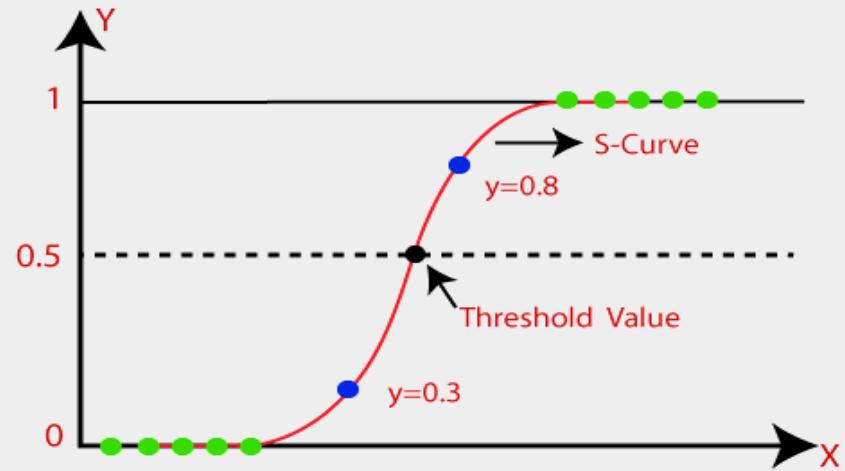


Image Source: <https://helloacm.com/wp-content/uploads/2016/03/logistic-regression-example.jpg>

Logistic Regression

Logistic Regression

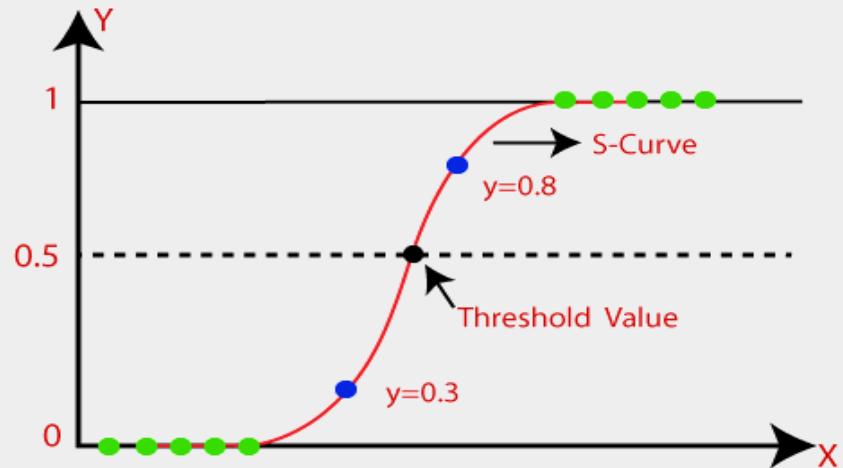
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification. The below image is showing the logistic function:



Logistic Regression

Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1.
- The S-form curve is called the Sigmoid function or the logistic function.



Logistic Regression

Logistic Regression Equation

- We know the equation of the straight line can be written as:
- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by $(1-y)$:
- But we need range between $-\infty$ to $+\infty$, then take logarithm of the equation it will become:



$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$



$$\frac{y}{1-y}; \text{ 0 for } y=0, \text{ and infinity for } y=1$$



$$\log \left[\frac{y}{1-y} \right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$

Logistic Regression

Types Logistic Regression Equation

- Binomial
- Multinomial
- Ordinal



Logistic Regression

Steps in Logistic Regression

- Data Pre-processing step
- Fitting Logistic Regression to the Training set
- Predicting the test result
- Test accuracy of the result (Creation of Confusion matrix)
- Visualizing the test set result.

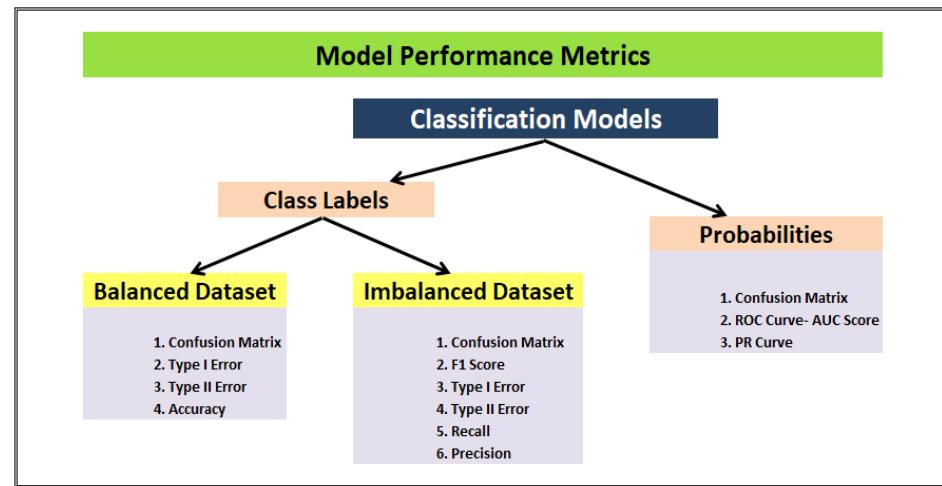
Note:- Check out theory Manual for Example



Model Performance Metrics

Logistic Regression Evaluation metrics

- Confusion matrix
- Precision
- Recall
- F1 Score
- Accuracy



Model Performance Metrics

Confusion matrix

- A confusion matrix is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known.

		Actual Value	
		Yes (1)	No (0)
Predicted Value	Yes (1)	TP	FP
	No (0)	FN	TN

TP= True Positive
FP= False Positive
FN= False Negative
TN= True Negative

Model Performance Metrics

Confusion matrix

- TP (True-positives):
- TN (True-negatives):
- FP (False-positives):
- FN (False-negatives):

		Actual Value	
		Yes (1)	No (0)
Predicted Value	Yes (1)	500 (TP)	100 (FP)
	No (0)	200 (FN)	200 (TN)

Model Performance Metrics

Accuracy

- Accuracy is the proximity of measurement results to the true value. It tell us how accurate our classification model is able to predict the class labels given in the problem statement.
- $\text{Accuracy} = (\text{TP} + \text{TN}) / \text{Total customers}$

		Actual Value	
		Yes (1)	No (0)
Predicted Value	Yes (1)	500 (TP)	100 (FP)
	No (0)	200 (FN)	200 (TN)

$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN})$
 $= (500 + 200) / (500 + 100 + 200 + 200)$
70%

Model Performance Metrics

Recall

- **Recall/ Sensitivity/ TPR (True Positive Rate)** attempts to answer the following question:
- What proportion of actual positives was identified correctly?
- Recall is generally used in use cases where the truth-detection is of utmost importance

		Actual Value	
		Yes (1)	No (0)
Predicted Value	Yes (1)	700 (TP)	80 (FP)
	No (0)	200 (FN)	20 (TN)

Recall/ True Positive Rate/ Sensitivity = $\frac{TP}{TP+FN}$

$=\frac{700}{700+200}$

78%

Model Performance Metrics

Precision

- **Precision** attempts to answer the following question:
- **What proportion of positive identifications was actually correct?**
- Precision is generally used in cases where it's of utmost importance not to have a high number of False positives

		Actual Value	
		Yes (1)	No (0)
Predicted Value	Yes (1)	600 (TP)	200 (FP)
	No (0)	100 (FN)	100 (TN)

Precision = $\frac{TP}{TP+FP}$
 $=\frac{600}{(600+200)}$
75%

Model Performance Metrics

F1 score

- **F1 score** (also **F-score** or **F-measure**) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score

The traditional F-measure or balanced F-score (**F₁ score**) is the harmonic mean of precision and recall:

$$F_1 = \left(\frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \right) = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

Python Library: Sci-Kit Learn

Sci-kit learn

- Open-source ML library for Python. Built on NumPy, SciPy, and Matplotlib.
- [Scikit-learn](#) is a library in Python that provides many unsupervised and supervised learning algorithms.



Python Library: Sci-Kit Learn

Functionality of Sci-kit learn

- Regression
- Classification
- Clustering
- Model Selection
- Pre-processing



Python Library: Sci-Kit Learn

Install Sci-kit learn

- Using pip

```
pip install -U scikit-learn
```

- Using conda

```
conda install scikit-learn
```



Python Library: Sci-Kit Learn

Features Sci-kit learn

- Supervised Learning algorithms
- Unsupervised Learning algorithms
- Clustering
- Cross Validation
- Feature extraction
- Feature selection
- Open Source



Python Library: Sci-Kit Learn

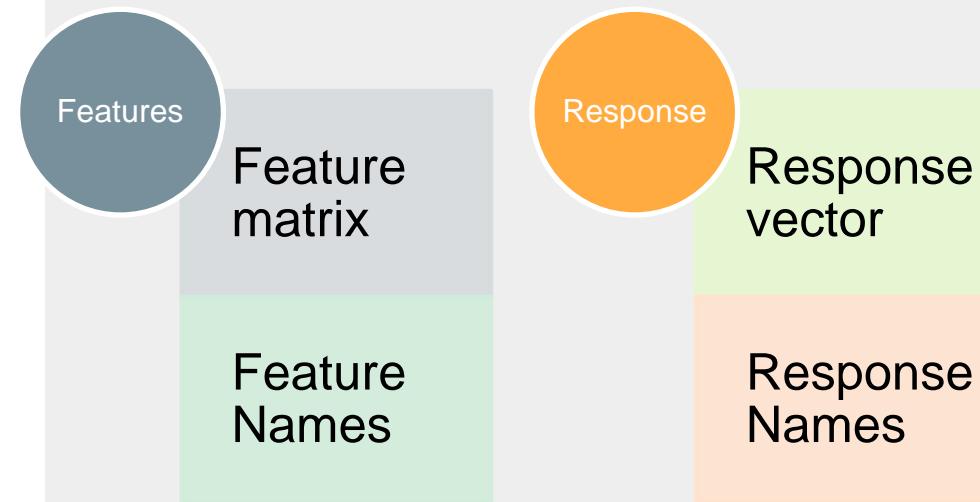
Sci-kit learn – Dataset Loading

Feature

- Features Matrix
- Features Names

Response

- Response Vector
- Response Names



Python Library: Sci-Kit Learn

Sci-kit learn – Dataset Loading

Feature

- Features Matrix
- Features Names

Response

- Response Vector
- Response Names

```
from sklearn.datasets import load_iris
```

```
iris = load_iris()
```

```
X = iris.data
```

```
y = iris.target
```

```
feature_names = iris.feature_names
```

```
target_names = iris.target_names
```

```
print("Feature names:", feature_names)
```

```
print("Target names:", target_names)
```

```
print("\nFirst 10 rows of X:\n", X[:10])
```

Python Library: Sci-Kit Learn

Sci-kit learn – Splitting dataset

```
from sklearn.model_selection import  
train_test_split
```

```
from sklearn.datasets import load_iris  
iris = load_iris()  
  
X = iris.data  
y = iris.target  
  
from sklearn.model_selection import train_test_split  
  
X_train, X_test, y_train, y_test = train_test_split(  
    X, y, test_size = 0.3, random_state = 1  
)  
  
print(X_train.shape)  
print(X_test.shape)  
  
print(y_train.shape)  
print(y_test.shape)
```

Python Library: Sci-Kit Learn

Sci-kit learn – Linear Regression

- This supervised ML model is used when the output variable is continuous and it follows linear relation with dependent variables. It can be used to forecast sales in the coming months by analyzing the sales data for previous months.

```
from sklearn.linear_model import LinearRegression  
  
from sklearn.metrics import mean_squared_error, r2_score  
  
regression_model = LinearRegression()  
  
regression_model.fit(x_train, y_train)  
  
y_predicted = regression_model.predict(x_test)  
  
rmse = mean_squared_error(y_test, y_predicted)
```