DATA SCIENCE

MCA 3RD SEMESTER
SIKKIM UNIVERSITY

-PRATIKSHYA SHARMA

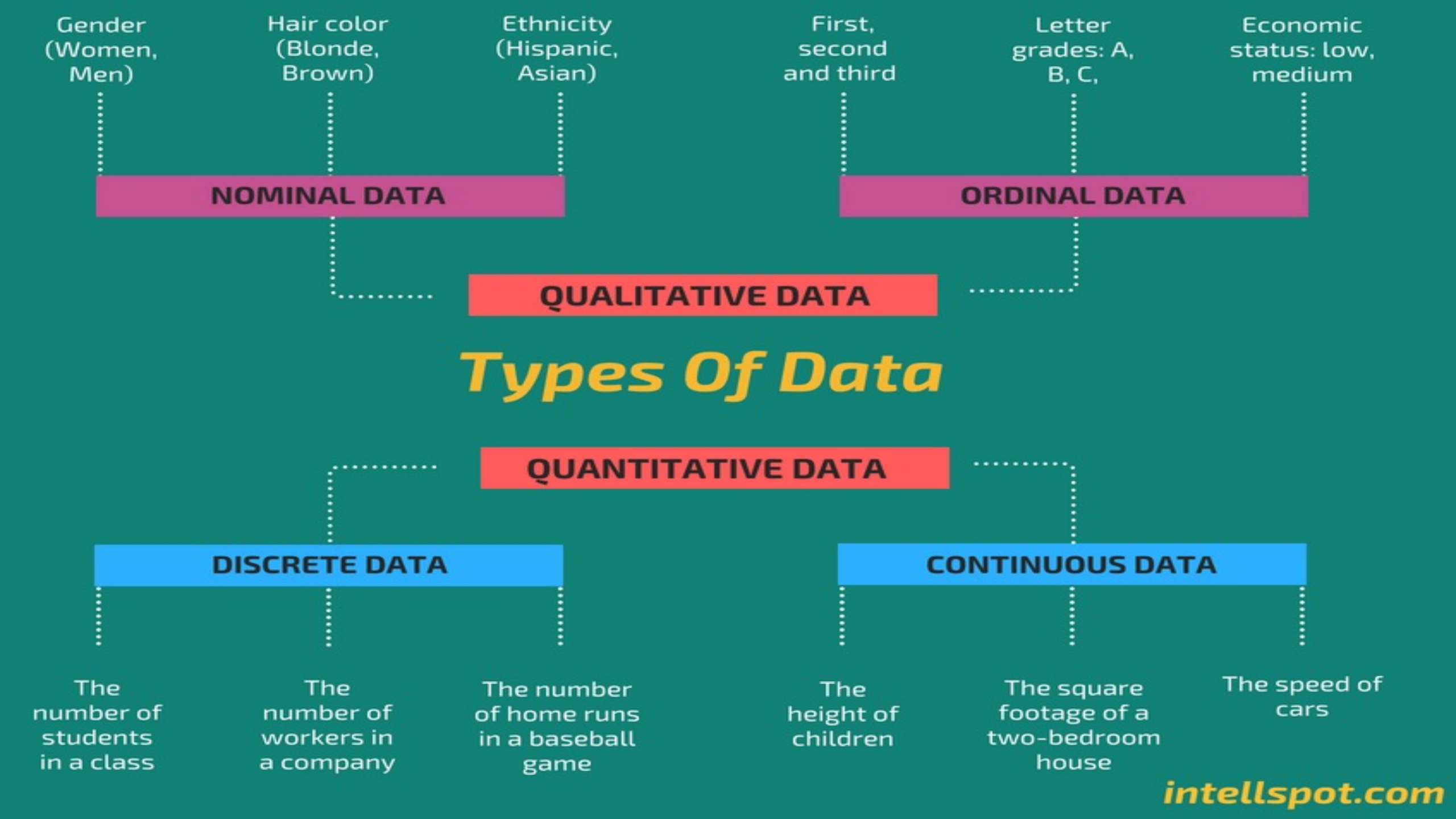
Contents

Introduction to Data Types

- Quantitative vs Qualitative data types
- Nominal , Ordinal, Nominal vs Ordinal
- Discrete, Continuous, Discrete vs Continuous
- Ratio

Introduction to Data Preparation

- Normalization
- Discretization
- Missing Value Estimation
- Sampling
- Feature Selection


Gender (Women, Men)
Hair color (Blonde, Brown)
Ethnicity (Hispanic, Asian)
First, second and third
Letter grades: A, B, C,
Economic status: low, medium
NOMINAL DATA
ORDINAL DATA
QUALITATIVE DATA
Types Of Data
QUANTITATIVE DATA
DISCRETE DATA
CONTINUOUS DATA
The number of students in a class
The number of workers in a company
The number of home runs in a baseball game
The height of children
The square footage of a two-bedroom house
The speed of cars
intellspot.com

Quantitative Data

- Quantitative data seems to be the easiest to explain. It answers key questions such as “how many, “how much” and “how often”.
- Quantitative data can be expressed as a number or can be quantified. Simply put, it can be measured by numerical variables.
- Quantitative data are easily amenable to statistical manipulation and can be represented by a wide variety of statistical types of graphs and charts such as line, bar graph, scatter plot, and etc.

- **Examples of quantitative data:**
- Scores on tests and exams e.g., 85, 67, 90 and etc.
- The weight of a person or a subject.
- Your shoe size.
- The temperature in a room.
- There are 2 general types of quantitative data: discrete data and continuous data.

Qualitative Data

- Qualitative data can't be expressed as a number and can't be measured.
- Qualitative data consist of words, pictures, and symbols, not numbers.
- Qualitative data is also called categorical data because the information can be sorted by category, not by number.
- Qualitative data can answer questions such as "how this has happened" or and "why this has happened".

- **Examples of qualitative data:**
- Colors e.g. the color of the sea
- Your favorite holiday destination such as Hawaii, New Zealand and etc.
- Names as John, Patricia,…..
- Ethnicity such as American Indian, Asian, etc.
- More you can see on our post qualitative vs quantitative data.
- There are 2 general types of qualitative data: nominal data and ordinal data.

QUANTITATIVE VS QUALITATIVE DATA

QUANTITATIVE DATA

Quantitative data can be expressed as a number or can be quantified. Simply put, quantitative data can be measured by numerical variables.



EXAMPLES

- Scores on tests and exams e.g. 85, 67, 90 and etc.
- The weight of a person or a subject.
- Your shoe size.
- The temperature in a room.

QUALITATIVE DATA

Qualitative data can't be expressed as a number and can't be measured. Qualitative data consist of words, pictures, and symbols, not numbers.

EXAMPLES

- Colors e.g. the color of the sea
- Your favorite holiday destination such as Hawaii, New Zealand.
- Names as John, Patricia,.....
- Ethnicity such as American Indian, Asian, etc.


Nominal Data

- Nominal data is used just for labeling variables, without any type of quantitative value. The name ‘nominal’ comes from the Latin word “nomen” which means ‘name’.
- The nominal data just name a thing without applying it to order. Actually, the nominal data could just be called “labels.”

- **Examples of Nominal Data:**
- Gender (Women, Men)
- Hair color (Blonde, Brown, Brunette, Red, etc.)
- Marital status (Married, Single, Widowed)
- Ethnicity (Hispanic, Asian)
- As you see from the examples there is no intrinsic ordering to the variables.
- Eye color is a nominal variable having a few categories (Blue, Green, Brown) and there is no way to order these categories from highest to lowest.

Nominal Data

What is your gender?

- M – Male
- F – Female

What is your hair color?

- 1 – Brown
- 2 – Black
- 3 – Blonde
- 4 – Gray
- 5 – Other

Where do you live?

- A – North of the equator
- B – South of the equator
- C – Neither: In the international space station

Ordinal Data

- Ordinal data shows where a number is in order. This is the crucial difference from nominal types of data.
- Ordinal data is data which is placed into some kind of order by their position on a scale. Ordinal data may indicate superiority.
- However, **you cannot do arithmetic with ordinal numbers** because they only show sequence.
- Ordinal variables are considered as “in between” qualitative and quantitative variables.
- In other words, the ordinal data is qualitative data for which the values are ordered.
- In comparison with nominal data, the second one is qualitative data for which the values cannot be placed in an ordered.
- We can also assign numbers to ordinal data to show their relative position. But we cannot do math with those numbers. For example: “first, second, third…etc.”

- **Examples of Ordinal Data:**
- The first, second and third person in a competition.
- Letter grades: A, B, C, and etc.
- When a company asks a customer to rate the sales experience on a scale of 1-10.
- Economic status: low, medium and high.

Ordinal Data

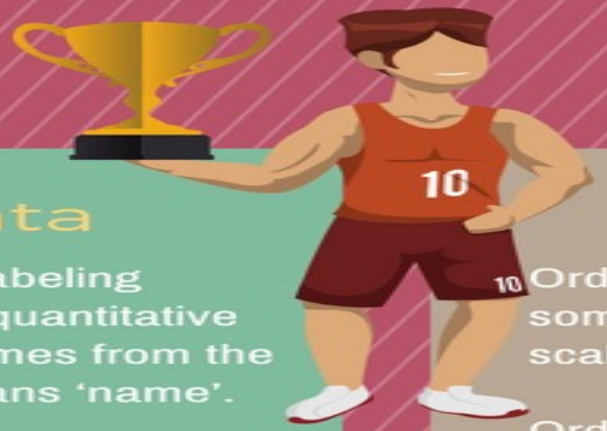
How do you feel today?

- (•) 1 – Very Unhappy
- ( ) 2 – Unhappy
- ( ) 3 – OK
- ( ) 4 – Happy
- ( ) 5 – Very Happy

How satisfied are you with our service?

- (•) 1 – Very Unsatisfied
- ( ) 2 – Somewhat Unsatisfied
- ( ) 3 – Neutral
- ( ) 4 – Somewhat Satisfied
- ( ) 5 – Very Satisfied

Nominal vs Ordinal Data



Nominal Data

Nominal data is used just for labeling variables, without any type of quantitative value. The name 'Nominal' comes from the Latin word "nomen" which means 'name'.

The nominal data just name a thing without applying it to an order. Actually, the nominal data could just be called "labels."

Ordinal data

Ordinal data is data which is placed into some kind of order by their position on a scale.

Ordinal data may indicate superiority. However, you cannot do arithmetic with ordinal numbers because they only show sequence.

Examples

- Gender (Women, Men)
- Hair color (Blonde, Brown, Brunette, Red, etc.)
- Marital status (Married, Single, Widowed)
- Ethnicity (Hispanic, Asian)

Examples

- The first, second and third person in a competition.
- Letter grades: A, B, C, and etc.
- When a company asks a customer to rate the sales experience on a scale of 1-10.
- Economic status: low, medium and high.

Nominal vs. Ordinal Data

Nominal Data	Ordinal Data
Nominal data can’t be quantified, neither they have any intrinsic ordering	Ordinal data gives some kind of sequential order by their position on the scale
Nominal data is qualitative data or categorical data	Ordinal data is said to be “in-between” qualitative data and quantitative data
They don’t provide any quantitative value, neither we can perform any arithmetical operation	They provide sequence and can assign numbers to ordinal data but cannot perform the arithmetical operation
Nominal data cannot be used to compare with one another	Ordinal data can help to compare one item with another by ranking or ordering
Examples: Eye colour, housing style, gender, hair colour, religion, marital status, ethnicity, etc	**Examples:** Economic status, customer satisfaction, education level, letter grades, etc

Discrete Data

- Discrete and continuous data are the two key types of quantitative data.
- In statistics, marketing research, and data science, many decisions depend on whether the basic data is discrete or continuous.

- Discrete data is a count that involves only integers. The discrete values cannot be subdivided into parts.
- For example, the number of children in a class is discrete data. You can count whole individuals. You can't count 1.5 kids.
- To put in other words, discrete data can take only certain values. The data variables cannot be divided into smaller parts.
- It has a limited number of possible values e.g., days of the month.

- **Examples of discrete data:**
- The number of students in a class.
- The number of workers in a company.
- The number of home runs in a baseball game.
- The number of test questions you answered correctly

Continuous Data

- Continuous data is information that could be meaningfully divided into finer levels. It can be measured on a scale or continuum and can have almost any numeric value.
- For example, you can measure your height at very precise scales — meters, centimeters, millimeters and etc.
- You can record continuous data at so many different measurements – width, temperature, time, and etc. This is where the key difference from discrete types of data lies.
- The continuous variables can take any value between two numbers. For example, between 50 and 72 inches, there are literally millions of possible heights: 52.04762 inches, 69.948376 inches and etc.
- A good great rule for defining if a data is continuous or discrete is that if the point of measurement can be reduced in half and still make sense, the data is continuous.

- **Examples of continuous data:**
- The amount of time required to complete a project.
- The height of children.
- The square footage of a two-bedroom house.
- The speed of cars.

Discrete vs Continuous Data

DISCRETE

Discrete data is a count that involves only integers. The discrete values cannot be subdivided into parts. For example, the number of children in a class is discrete data. You can't count 1.5 kids.

EXAMPLES

- The number of students in a class.
- The number of workers in a company.
- The number of home runs in a baseball game.
- The number of test questions you answered correctly

PICS

CONTINUOUS

Continuous data could be meaningfully divided into finer levels. It can be measured on a scale or continuum and can have any numeric value. For example, you can measure your height at very precise scales — meters, centimeters, millimeters, etc.

EXAMPLES

- The amount of time required to complete a project.
- The height of children.
- The square footage of a two-bedroom house.
- The speed of cars.

PICS

Discrete vs. Continuous Data

Discrete Data	Continuous Data
Discrete data are countable and finite; they are whole numbers or integers	Continuous data are measurable; they are in the form of fraction or decimal
Discrete data are represented mainly by bar graphs	Continuous data are represented in the form of a histogram
The values cannot be divided into subdivisions into smaller pieces	The values can be divided into subdivisions into smaller pieces
Discrete data have spaces between the values	Continuous data are in the form of a continuous sequence
Examples: Total students in a class, number of days in a week, size of a shoe, etc	**Example:** Temperature of room, the weight of a person, length of an object, etc

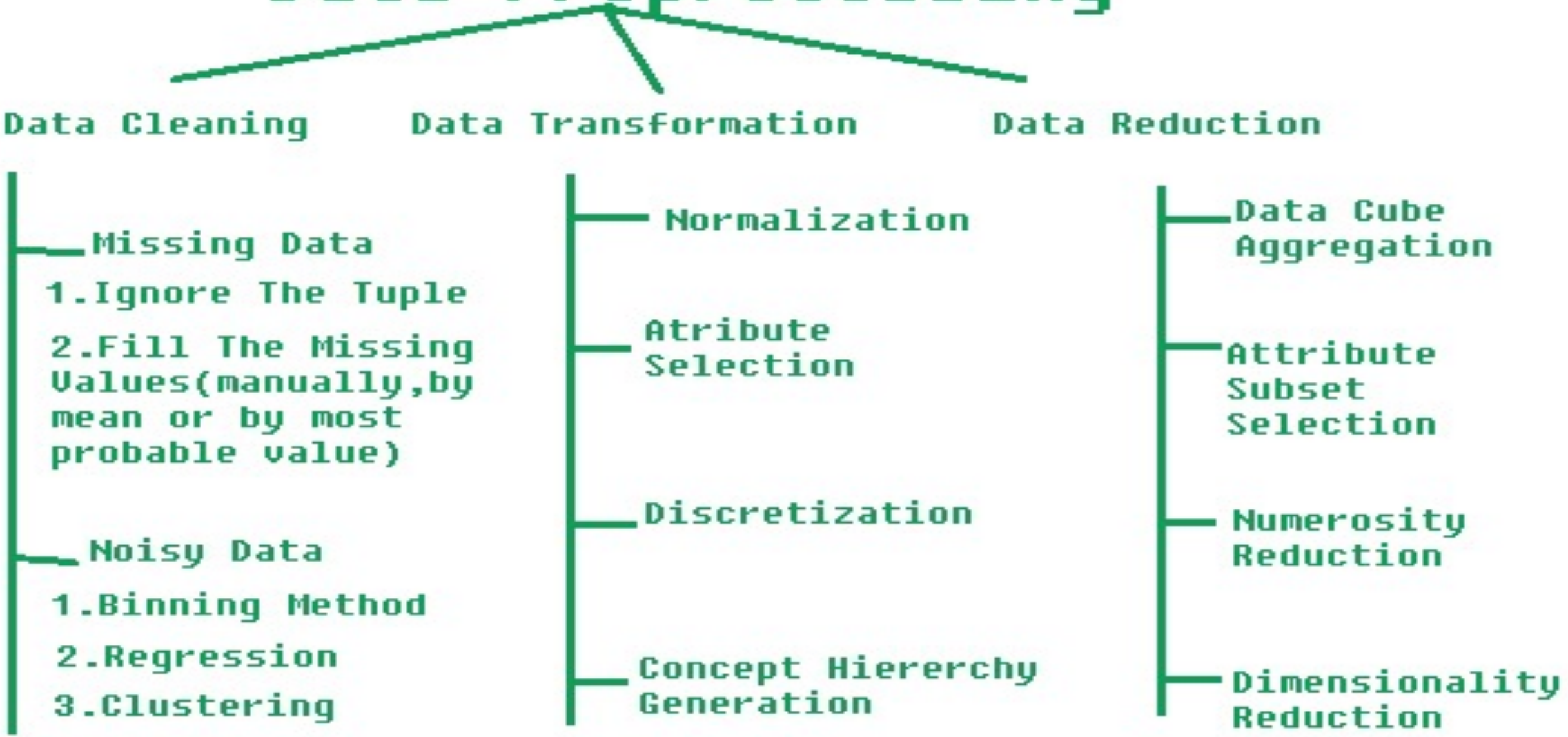
Ratio

- Ratio scale data is quantitative in nature due to which all quantitative analysis techniques can be used to calculate ratio data.

- **Ratio Scale Examples**
- The following questions fall under the Ratio Scale category:
- What is your daughter's current height?
  - Less than 5 feet.
  - 5 feet 1 inch – 5 feet 5 inches
  - 5 feet 6 inches- 6 feet
  - More than 6 feet
- What is your weight in kilograms?
  - Less than 50 kilograms
  - 51- 70 kilograms
  - 71- 90 kilograms
  - 91-110 kilograms
  - More than 110 kilograms

DATA PREPARATION

Data Preprocessing



DATA PREPARATION

- NORMALIZATION
- DISCRETIZATION
- MISSING VALUE ESTIMATION
- SAMPLING
- FEATURE SELECTION

DATA PREPARATION

NORMALIZATION

Normalization

- Data normalization is mainly needed to minimize or exclude duplicate data.
- **Normalization in data mining** is a beneficial procedure as it allows achieving certain advantages as mentioned below:

➢ It is a lot easier to apply data mining algorithms on a set of normalized data.

➢ The results of data mining algorithms applied to a set of normalized data are more accurate and effective.

➢ Once the data is normalized, the extraction of data from databases becomes a lot faster.

➢ More specific data analyzing methods can be applied to normalized data.

Popular Techniques for Data Normalization

- **Min Max Normalization**
- **Decimal Scaling Normalization**
- **Z-Score Normalization**

Min-Max Normalization

- What is easier to understand – the difference between 200 and 1000000 or the difference between 0.2 and 1?
- When the difference between the minimum and maximum values is less, the data becomes more readable.
- The min-max normalization functions by converting a range of data into a scale that ranges from 0 to 1.

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

Example.

- Suppose a company wants to decide on a promotion based on the years of work experience of its employees.
- So, it needs to analyze a database that looks like this:

Employee Name	**Years of Experience**
ABC	8
XYZ	20
PQR	10
MNO	15

Employee Name	Years of Experience
ABC	8
XYZ	20
PQR	10
MNO	15

- The minimum value is 8
- The maximum value is 20
- As this formula scales the data between 0 and 1,
- The new min is 0
- The new max is 1
- Here, V stands for the respective value of the attribute, i.e., 8, 10, 15, 20
- After applying the min-max normalization formula, the following are the **V'** values for the attributes:
- For 8 years of experience: **v'= 0**
- For 10 years of experience: **v' = 0.16**
- For 15 years of experience: **v' = 0.58**
- For 20 years of experience: **v' = 1**

So, the min-max normalization can reduce big numbers to much smaller values. This makes it extremely easy to read the difference between the ranging numbers.

Decimal Scaling Normalization

- Decimal scaling is another technique for **normalization in data mining**.
- It functions by converting a number to a decimal point.
- Normalization by decimal scaling follows the method of standard deviation.
- In decimal scaling normalization, the decimal point of values of the attributes is moved.
- The movement of the decimal points in decimal scaling normalization is dependent upon the maximum values amongst all values of the attribute.

Formula: $$v' = \frac{v}{10^j}.$$

Here:

- V’ is the new value after applying the decimal scaling
- V is the respective value of the attribute
- Now, integer J defines the movement of decimal points. So, how to define it? It is equal to the number of digits present in the maximum value in the data table.

- Suppose a company wants to compare the salaries of the new joiners. Here are the data values:

Employee Name	Salary
ABC	10,000
XYZ	25,000
PQR	8,000
MNO	15,000

Now, look for the maximum value in the data.
In this case, it is 25,000.
Now count the number of digits in this value.
In this case, it is ‘5’.
So here ‘j’ is equal to 5, i.e 100,000.
This means the V (value of the attribute) needs to be divided by 100,000 here.

Name	Salary	Salary after Decimal Scaling
ABC	10,000	0.1
XYZ	25, 000	0.25
PQR	8, 000	0.08
MNO	15,000	0.15

After applying the zero decimal scaling formula, here are the new values:

Thus, decimal scaling can tone down big numbers into easy to understand smaller decimal values.

Also, data attributed to different units becomes easy to read and understand once it is converted into smaller decimal values.

Z-Score Normalization

- Z-Score value is to understand how far the data point is from the mean.
- Technically, it measures the standard deviations below or above the mean.
- It ranges from -3 standard deviation up to +3 standard deviation. Z-score **normalization in data mining** is useful for those kinds of data analysis wherein there is a need to compare a value with respect to a mean(average) value, such as results from tests or surveys.
- Thus, Z-score normalization is also popularly known as Standardization.
- The following formula is used in the case of z-score normalization on every single value of the dataset.

Formula : New value = (x – μ) / σ

- Here:
- **x:** Original value
- **μ:** Mean of data
- **σ:** Standard deviation of data

Data
3
5
5
8
9
12
12
13
15
16
17
19
22
24
25
134

Suppose we have the following dataset:

- Therefore, we can find that the mean of this dataset is 21.2 also the standard deviation is 29.8.
- If we have to perform z score normalization on the first value of the dataset,
- Then according to the formula it will be,
- New value = $(x - \mu) / \sigma$
- New value = (3-21.2)/ 29.8
- $\therefore$ New value = -0.61
- By performing z score normalization on each of the value of the dataset, we will get the following chart.

Data	Z score normalized value
3	-0.61
5	-0.54
5	-0.54
8	-0.44
9	-0.41
12	-0.31
12	-0.31
13	-0.28
15	-0.21
16	-0.17
17	-0.14
19	-0.07
22	0.03
24	0.09
25	0.13
134	3.79

- The mean of this normalized dataset is 0 and the standard deviation is 1.
- For example, a person's weight is 150 pounds.
- Now, if there is a need to compare that value with the average weight of a population listed in a vast table of data, Z-score normalization is needed to study such values, especially if someone's weight is recorded in kilograms.

Difference between Min Max normalization and Z Score Normalization:

Min Max normalization	Z Score Normalization
•For scaling the minimum and maximum values of the feature are used. •Applicable when the features are of different sizes •The values are scaled between the range of [0,1] or [-1, 1] •Gets easily affected by outliers •A transformer named MaxMinScaler is available in Scikit-Learn •This method transforms an n-dimensional data into an n-dimensional unit hypercube •Best if the distribution is unknown •Also known as Scaling Normalization	•For scaling mean deviation and the standard deviation is used. •Useful when want to maintain a zero mean and unit standard deviation. •No fixed range is present •Not that much affected by outliers. •Transformer named StandardScaler is available in Scikit-Learn to perform the task. •This method translates data to the mean vector of the original data and then either squeezes or expands it. •Best when Normal or Gaussian distribution •Also known as Standardization

PROBLEMS:

1. Normalize the following group of data using min-max normalization– 1000,2000,3000,9000
2. Normalise using decimal scale

F63

	A	B	C	D	E
1					
2	https://T4Tutorials.com				
3	Id	Dependa	Sal	Eucledean	Id
4	E101	3	50000	0	E101
5	E105	5	50000	49999.37304	E110
6	E110	3	45000	5000	E113
7	E113	3	57000	7000	E114
8	E111	6	43000	7000.000643	E112
9	E114	3	42000	8000	E107
10	E109	5	40000	10000.0002	E108
11	E112	4	39000	11000.00005	E102
12	E108	4	38000	12000.00004	E104
13	E107	3	35000	15000	E105
14	E102	4	65000	15000.00003	E103
15	E104	4	35000	15000.00003	E109
16	E103	3	70000	20000	E106
17	E106	1	30000	20000.0001	E111

Perform Z-score normalization

Question

marks
8
10
15
20

Answer

marks	marks after z-score normalization
8	-1.14
10	-0.7
15	0.3
20	1.4

DATA PREPARATION

DISCRETIZATION

Discretization

- This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
- It is a process of transforming continuous data into set of small intervals or into discrete buckets by grouping it.
- It can significantly improve its efficiency by replacing a constant quality attribute with its discrete values.

Data Discretization Example

- Suppose we have an attribute of age with the following values.
- **Table:** Before discretization

Age	10,11,13,14,17,19,30, 31, 32, 38, 40, 42,70 , 72, 73, 75

- **Table:** Before discretization

Attribute	**Age**	**Age**	**Age**
	10,11,13,14,17,19	30, 31, 32, 38, 40, 42	70 , 72, 73, 75
After Discretization	Young	Mature	Old

What are some famous techniques of data discretization?

- Histogram
- Binning
- Clustering analysis
- Decision tree analysis

Histogram

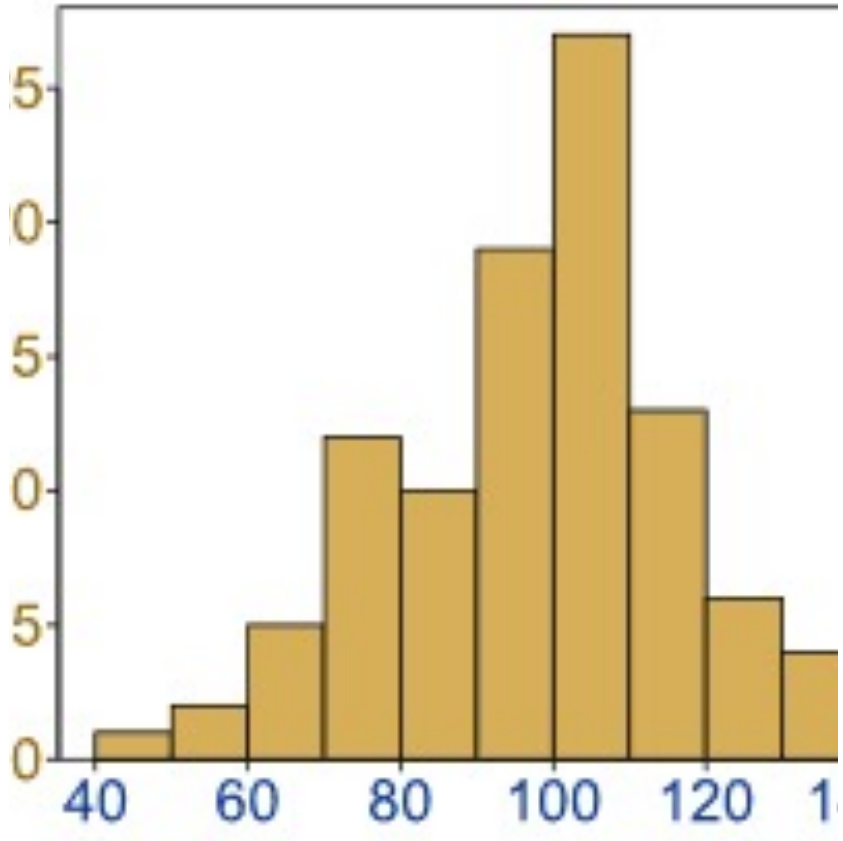
- It is a graphical representation of the distribution of a dataset.
- A histogram is a plot that lets you show the underlying frequency distribution or the **probability distribution** of a single **continuous numerical variable.**
- Histograms are two-dimensional plots with two axes; the vertical axis is a frequency axis whilst the horizontal axis is divided into a range of numeric values (intervals or **bins**) or time intervals.
- The frequency of each bin is shown by the area of vertical rectangular bars.
- Each bar covers a range of continuous numeric values of the variable under study.
- The vertical axis shows frequency values derived from counts for each bin.
- The midpoint value is the one that gives the name to the interval.
- When a numerical value corresponds exactly to one of the boundaries of the interval, it will be assigned to the left or right interval according to the default setting of the visualization tool.

Creating a Histogram

Steps you must go through to create a Histogram.

Step 1 – Minimum Data Points

- To accurately analyze a data set, *it's commonly recommended that you have at least 50 data points*. Without an adequate amount of data, you cannot make reasonable conclusions about your data.
- Basically, you may miss the pattern in the variation.
- On the flip side of this requirement, one of the strengths of the Histogram is that it allows you to easily analyze large data sets.



Creating a Histogram

Step 2 – Number of Bins

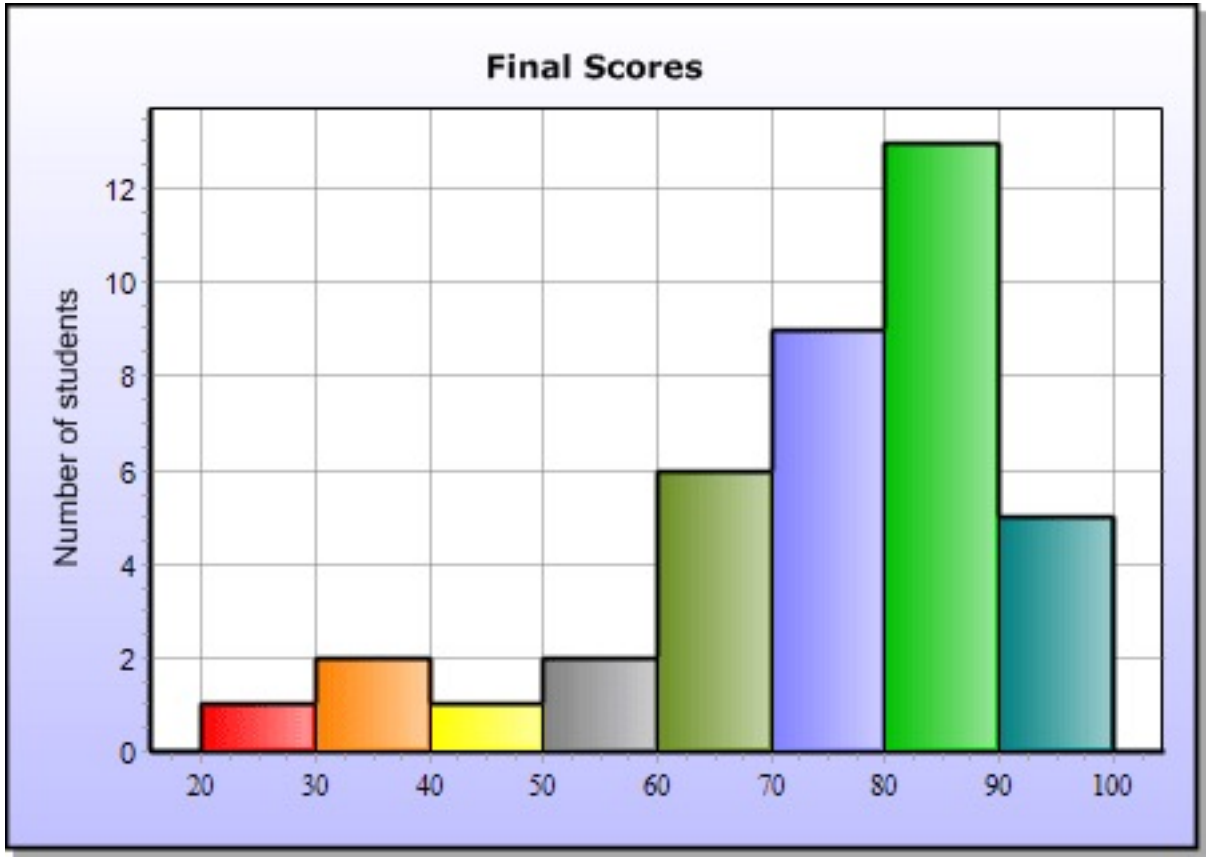
- Now that you've collected an adequate amount of data, it's time to calculate the number of Bars, sometimes called Bins or Ranges, for your data set. *The number of Bars for your Histogram will depend on the number of data points you collected.*
- Selecting the correct number of Bins is important as it can drastically affect the appearance of your data, which might lead you to the wrong conclusion.

Creating a Histogram

- ***Step 3 – Determine Bin Width***
- Once you've determined the number of Bins for your Histogram, it's time to calculate the Width or Range of each individual Bin.
- To do that *you take the entire Range of the data (Max data point minus Min data point) and divide by the total number of Bins*.

Creating a Histogram

- So for example, let's say you're creating a Histogram of Student's Test Scores on an exam and the maximum score was 100 and the minimum score was 20; then your Range is 80 i.e., (100 – 20).
- Then you can divide your data Range (80), by the total number of Bins, lets say 8 in this instance. So the Width of each Bin is 80/ 8 = 10.
- Similar to selecting the right number of total Bins, it's important that you keep all the Bin widths the same or this will skew the distribution of the data.



Create a Histogram for following data

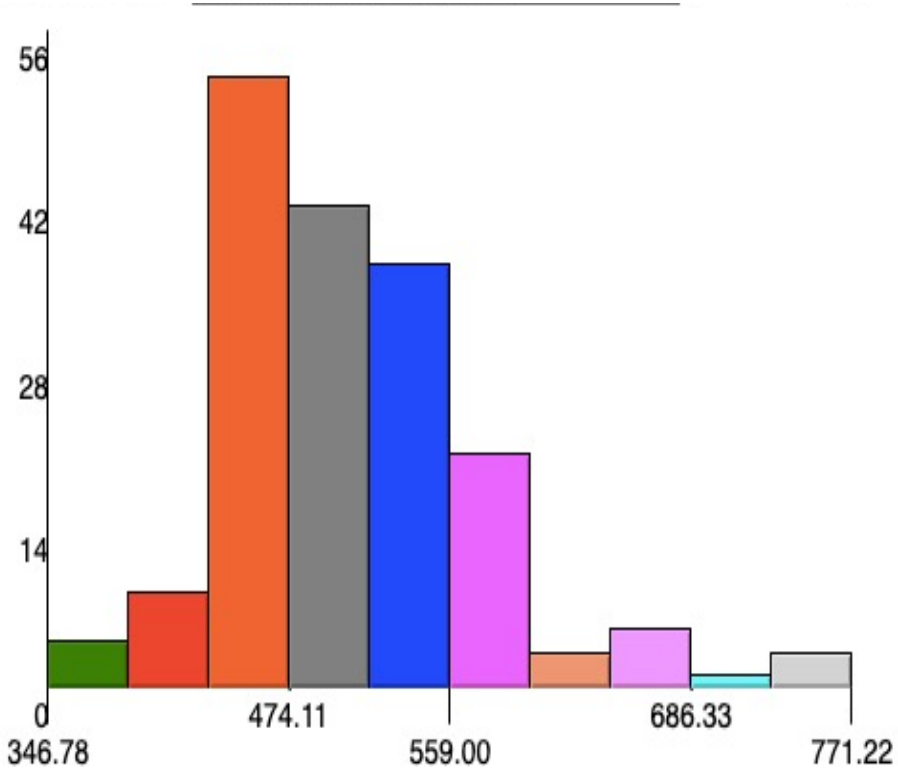
490	485	482	585	548	644	505	650	515
499	521	449	466	589	489	450	512	450
459	495	464	549	549	506	477	445	537
575	526	501	736	485	660	557	446	490
575	474	670	654	480	444	446	474	446
513	500	740	585	545	551	553	449	410
382	441	590	574	451	583	370	529	526
525	750	700	621	448	457	533	538	560
510	495	590	542	487	440	496	450	560
542	476	450	616	480	470	513	570	540
368	456	452	547	540	486	403	499	502
564	440	468	554	470	413	496	375	642
509	547	472	514	529	470	543	515	590
530	479	447	592	445	408	533	445	480
	501	520	531	460	440	471	571	557
	476	506	550	457	596	404	442	468
	457	570	507	560	442	439	492	524
	444	474	441	495	544	459	456	445
	444	532	551	480	528	393	428	479
	467	472	450	430	559	470	536	

Frequency Range

Tabular Data

Frequency Range	Hits
> 346.778 <= 389.222	4
> 389.222 <= 431.666	8
> 431.666 <= 474.110	52
> 474.110 <= 516.554	41
> 516.554 <= 558.998	36
> 558.998 <= 601.442	20
> 601.442 <= 643.886	3
> 643.886 <= 686.330	5
> 686.330 <= 728.774	1
> 728.774 <= 771.218	3

Clear Close



Interval Size = 42.444 Update Interval

Minimum value of Y-axis = 0

Maximum value of Y-axis = 56

Update Y-axis

Min Frequency = 1

Max Frequency = 52

N = 173

Mean = 507.173

Std. Dev. = 68.283

Minimum value of X-axis = 346.778 Set X Min

Binning

- It is a pre-processing technique used to reduce the effects of minor observation errors.
- The original data values are divided into small intervals known as bins, and then they are replaced by a general value calculated for that bin.
- This has a soothing effect on the input data and may also reduce the chances of over fitting in the case of small datasets.
- Binning of continuous variables introduces non-linearity and tends to improve the performance of the model.
- It can also be used to identify missing values or outliers.

How do you Bin Data?

There are two methods of dividing data into bins and binning data:

1. Equal Frequency Binning: Bins have an equal frequency.

For example, equal frequency:

Input: [5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]

Output:

[5, 10, 11, 13]

[15, 35, 50, 55]

[72, 92, 204, 215]

2. Equal Width Binning: Bins have equal width with a range of each bin are defined as:

[min + w], [min + 2w] …. [min + nw] where w = (max - min) / (no of bins).

For example, equal Width:

Input: [5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215]

Output:

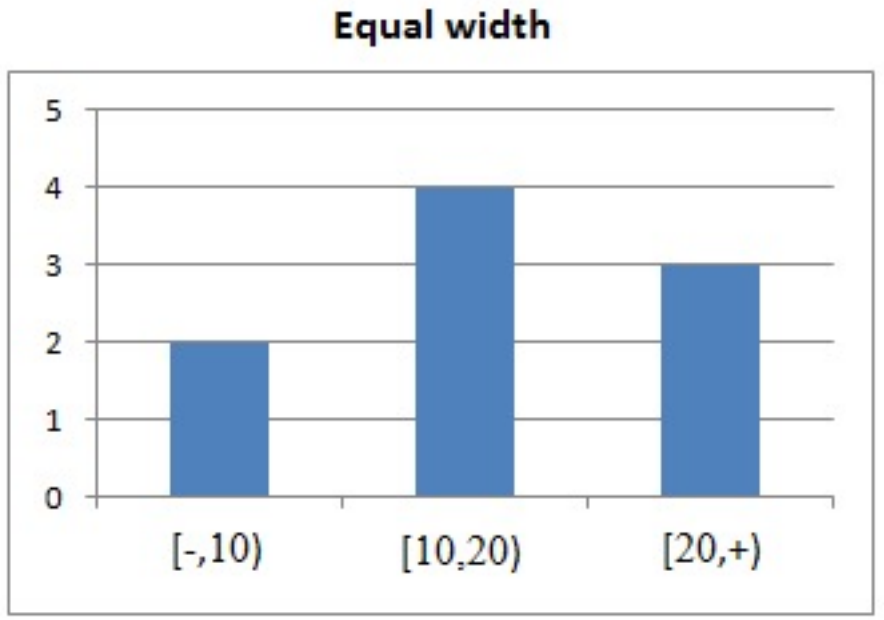
[5, 10, 11, 13, 15, 35, 50, 55, 72]

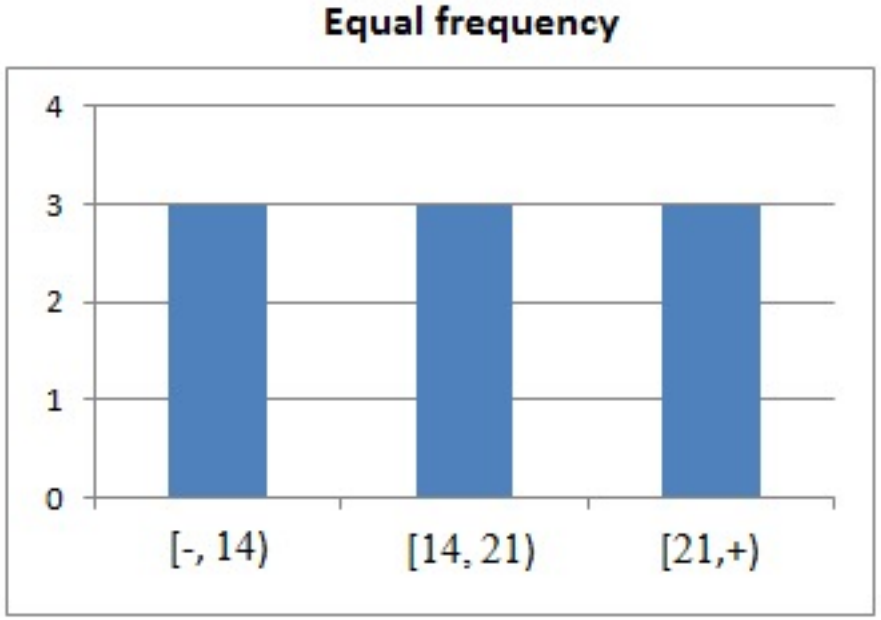
[92]

[204, 215]

The algorithm divides the data into k groups which each group contains approximately same number of values. For the both methods, the best way of determining k is by looking at the histogram and try different intervals or groups.

- **Data :** 0, 4, 12, 16, 16, 18, 24, 26, 28
- **Equal width**
  - Bin 1: 0, 4 [-,10)
  - Bin 2: 12, 16, 16, 18 [10,20)
  - Bin 3: 24, 26, 28 [20,+)
- **Equal frequency**
  - Bin 1: 0, 4, 12 [-, 14)
  - Bin 2: 16, 16, 18 [14, 21)
  - Bin 3: 24, 26, 28 [21,+)





Cluster Analysis

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

Cluster:

The given data is divided into different groups by combining similar objects into a group. This group is nothing but a cluster. A cluster is nothing but a collection of similar data which is grouped together.

For example, consider a dataset of vehicles given in which it contains information about different vehicles like cars, buses, bicycles, etc. As it is unsupervised learning there are no class labels like Cars, Bikes, etc for all the vehicles, all the data is combined and is not in a structured manner.

Now our task is to convert the unlabelled data to labelled data and it can be done using clusters.

K-Mean (A centroid based Technique)

- The K means algorithm takes the input parameter K from the user and partitions the dataset containing N objects into K clusters so that resulting similarity among the data objects inside the group (intracluster) is high but the similarity of data objects with the data objects from outside the cluster is low (intercluster).
- The similarity of the cluster is determined with respect to the mean value of the cluster.
- At the start randomly k objects from the dataset are chosen in which each of the objects represents a cluster mean(centre).
- For the rest of the data objects, they are assigned to the nearest cluster based on their distance from the cluster mean.
- The new mean of each of the cluster is then calculated with the added data objects.

Algorithm: K mean:

Input: K: The number of clusters in which the dataset has to be divided D: A dataset containing N number of objects

Output: A dataset of K clusters

Method:

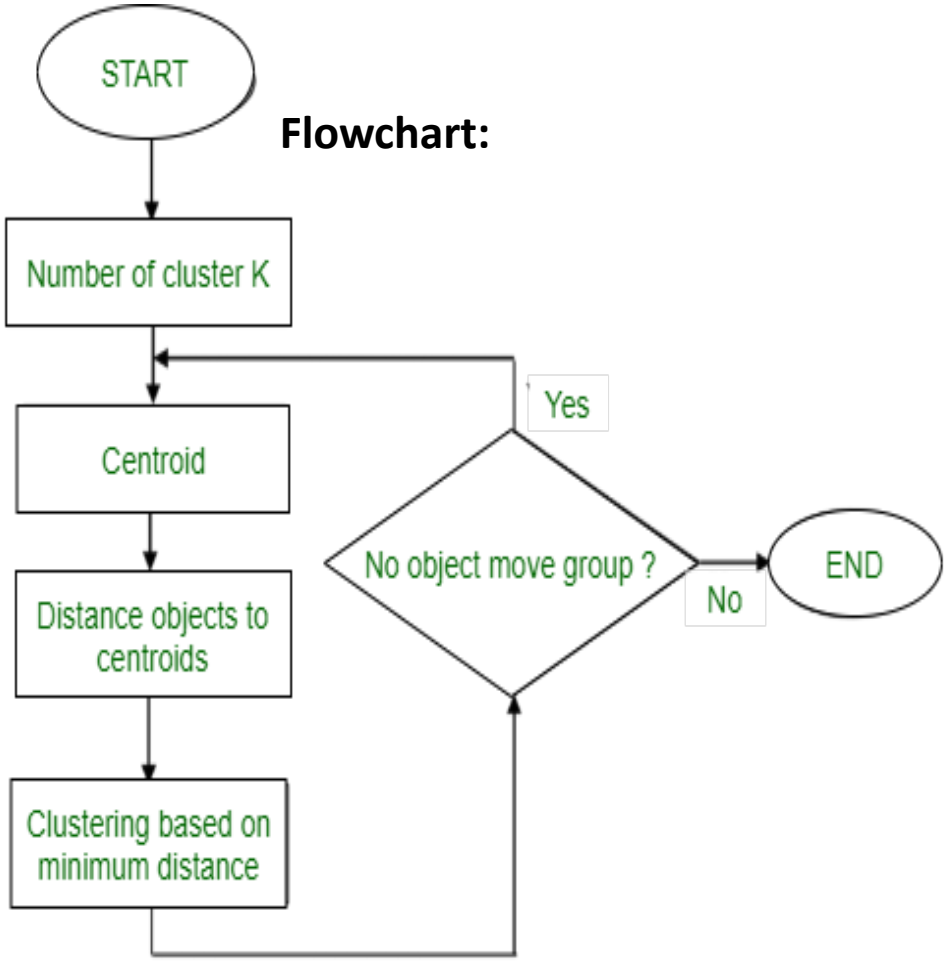
Randomly assign K objects from the dataset(D) as cluster centres(C)

(Re) Assign each object to which object is most similar based upon mean values.

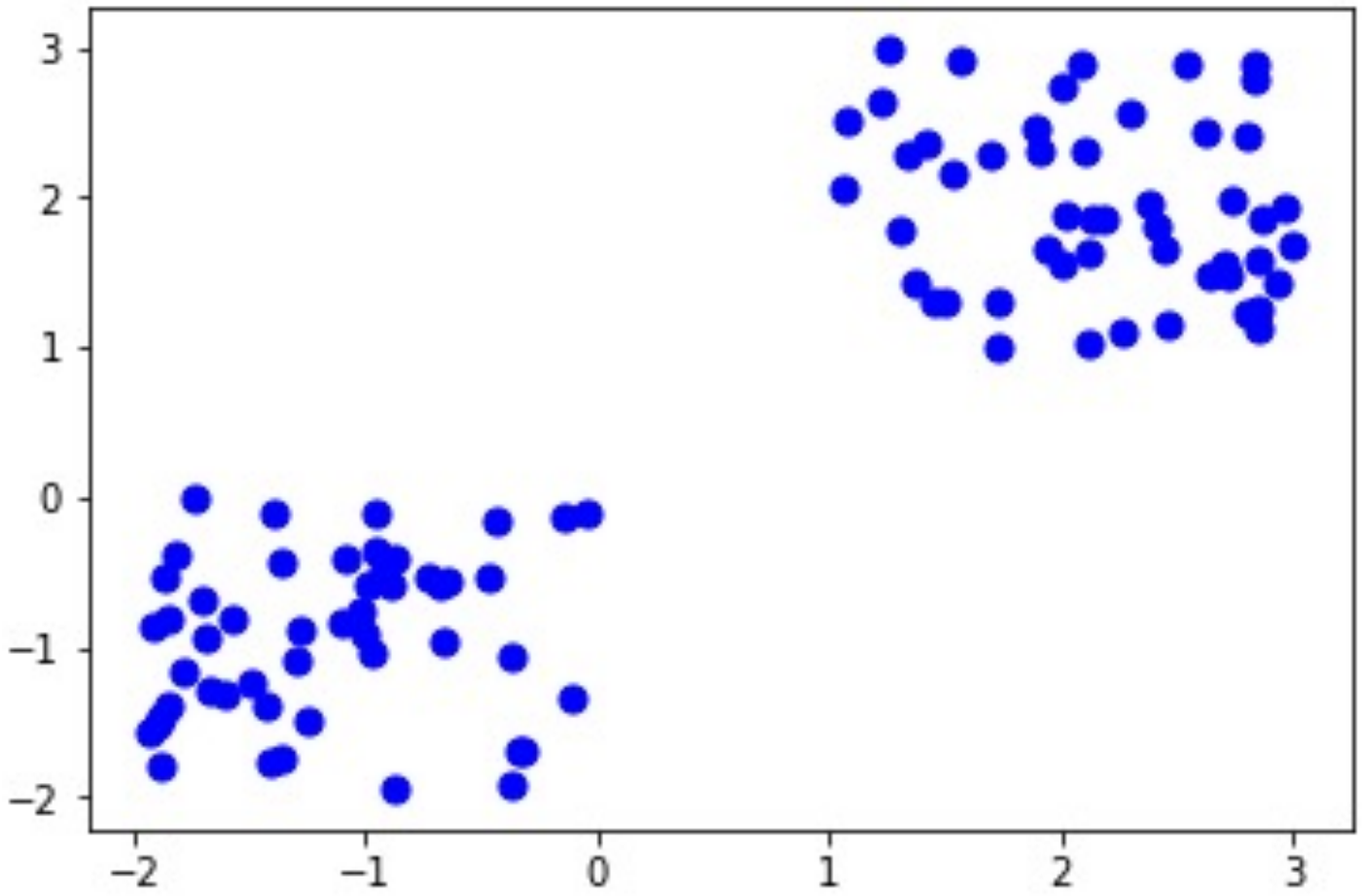
Update Cluster means, i.e., Recalculate the mean of each cluster with the updated values.

Repeat Step 4 until no change occurs.

Flowchart:



- **Figure –** K-mean Clustering



Example:

Suppose we want to group the visitors to a website using just their age as follows:

Data: 16, 16, 17, 20, 20, 21, 21, 22, 23, 29, 36, 41, 42, 43, 44, 45, 61, 62, 66

Initial Cluster:

K=2 Centroid(C1) = 16 [16] Centroid(C2) = 22 [22]

Note: These two points are chosen randomly from the dataset.

Iteration-1:

C1 = 16.33 [16, 16, 17] C2 = 37.25 [20, 20, 21, 21, 22, 23, 29, 36, 41, 42, 43, 44, 45, 61, 62, 66]

Iteration-2:

C1 = 19.55 [16, 16, 17, 20, 20, 21, 21, 22, 23]
C2 = 46.90 [29, 36, 41, 42, 43, 44, 45, 61, 62, 66]

Iteration-3:

C1 = 20.50 [16, 16, 17, 20, 20, 21, 21, 22, 23, 29]
C2 = 48.89 [36, 41, 42, 43, 44, 45, 61, 62, 66]

Iteration-4:

C1 = 20.50 [16, 16, 17, 20, 20, 21, 21, 22, 23, 29]
C2 = 48.89 [36, 41, 42, 43, 44, 45, 61, 62, 66]

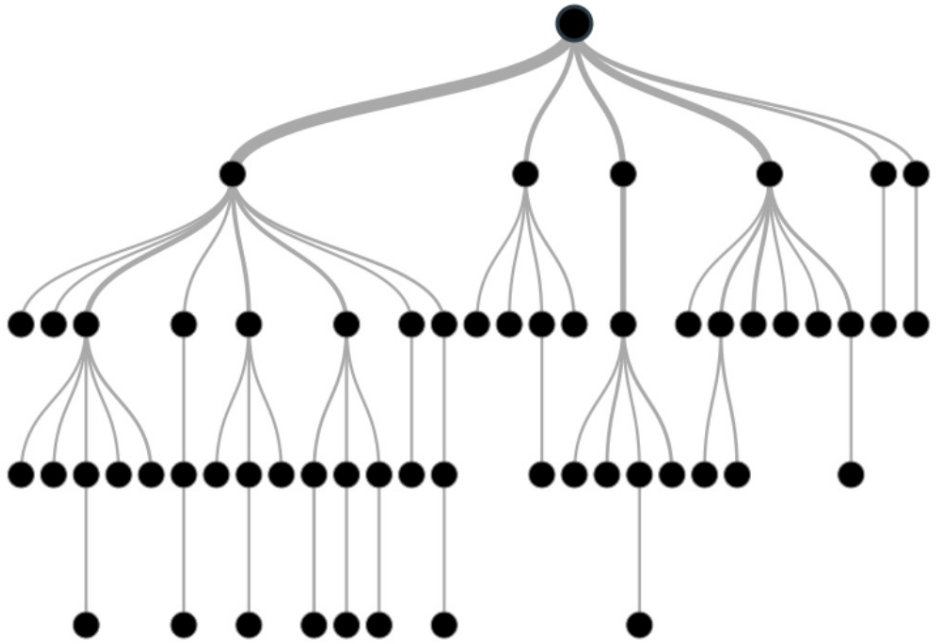
No change

Between Iteration 3 and 4, so we stop.

Therefore, we get the clusters (16-29) and (36-66) as 2 clusters we get using K Mean Algorithm.

Decision Tree Analysis

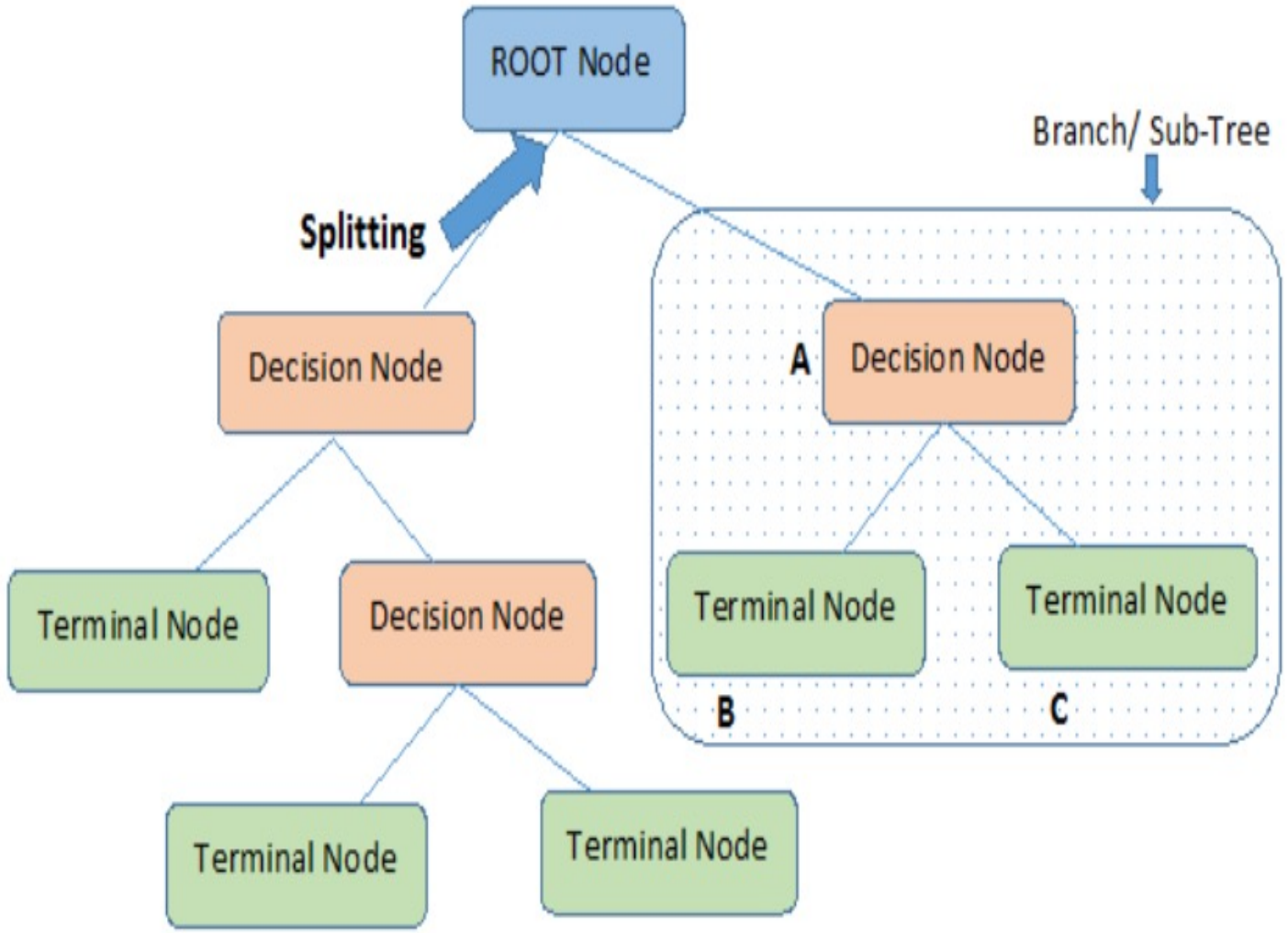
In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions.



- Decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.

Before learning more about decision trees let's get familiar with some of the terminologies.

- ***Root Nodes*** – It is the node present at the beginning of a decision tree from this node the population starts dividing according to various features.
- ***Decision Nodes*** – the nodes we get after splitting the root nodes are called Decision Node
- ***Leaf Nodes*** – the nodes where further splitting is not possible are called leaf nodes or terminal nodes
- ***Sub-tree*** – just like a small portion of a graph is called sub-graph similarly a sub-section of this decision tree is called sub-tree.
- ***Pruning*** – is nothing but cutting down some nodes to stop overfitting.



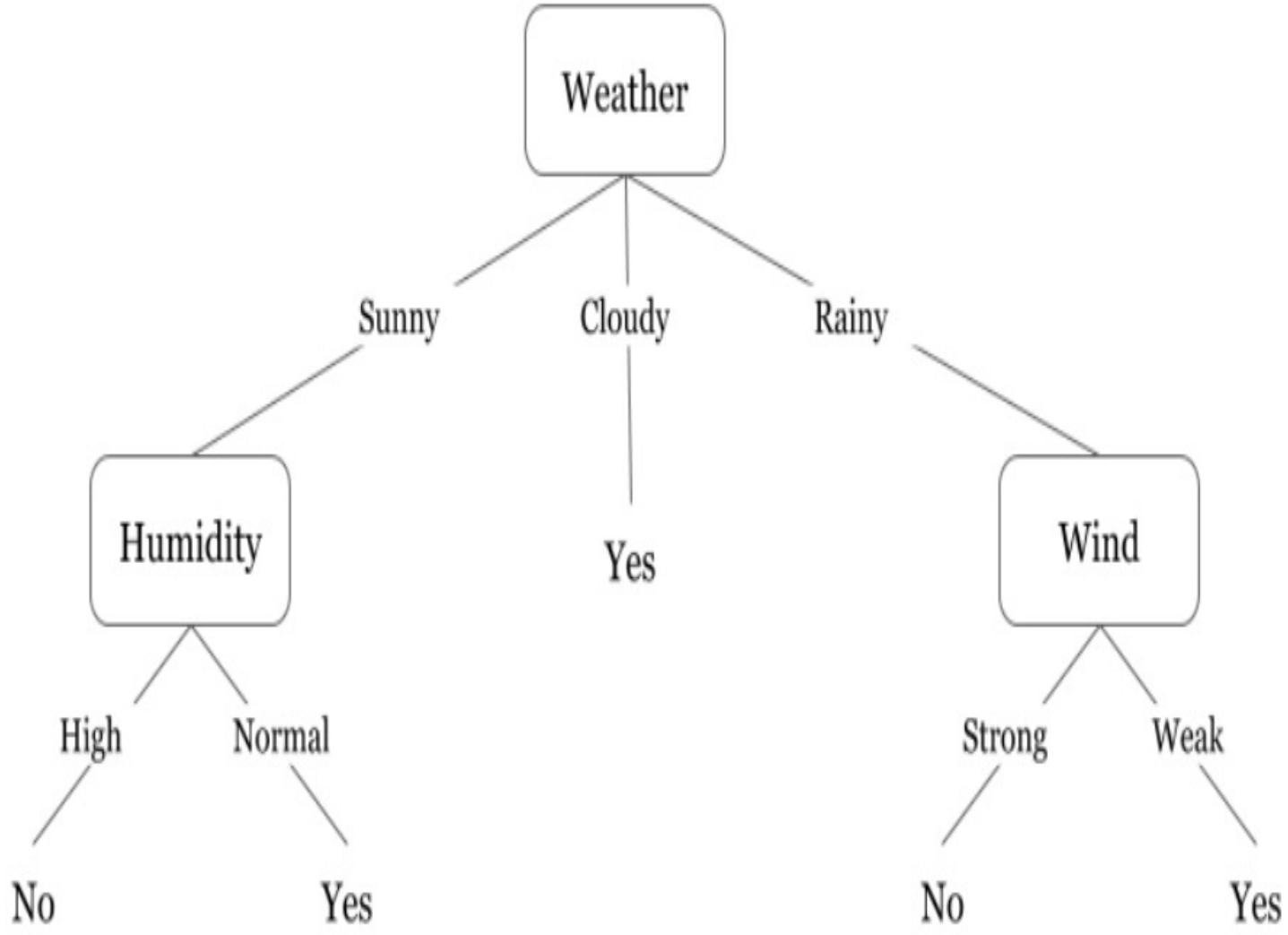
Example - decision tree

Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No

Decision trees are upside down which means the root is at the top and then this root is split into various several nodes.

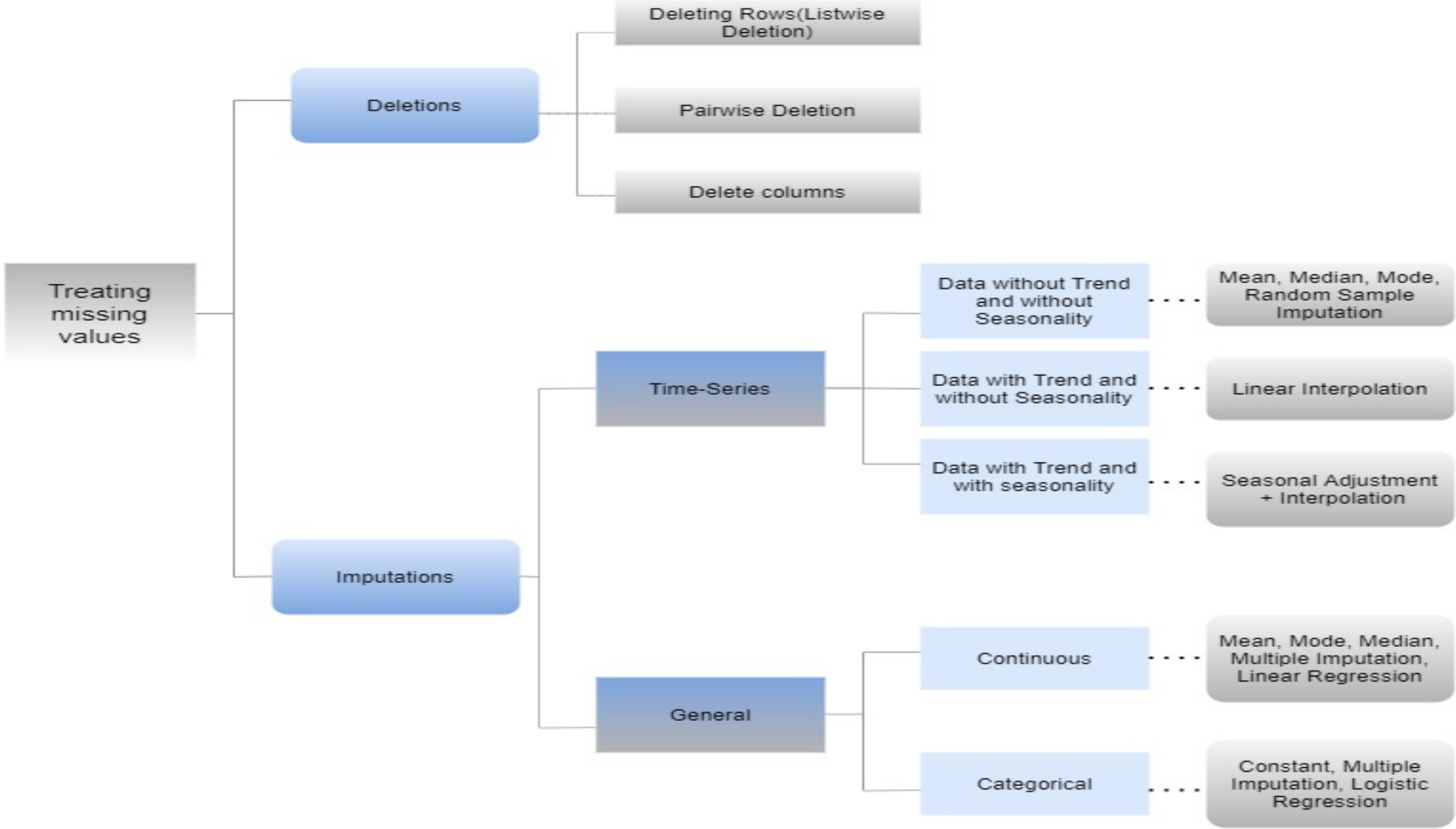
Decision trees are nothing but a bunch of if-else statements in layman terms.

It checks if the condition is true and if it is then it goes to the next node attached to that decision.



DATA PREPARATION

MISSING VALUE ESTIMATION


Treating missing values
Deletions
Deleting Rows(Listwise Deletion)
Pairwise Deletion
Delete columns
Imputations
Time-Series
Data without Trend and without Seasonality
Mean, Median, Mode, Random Sample Imputation
Data with Trend and without Seasonality
Linear Interpolation
Data with Trend and with seasonality
Seasonal Adjustment + Interpolation
General
Continuous
Mean, Mode, Median, Multiple Imputation, Linear Regression
Categorical
Constant, Multiple Imputation, Logistic Regression

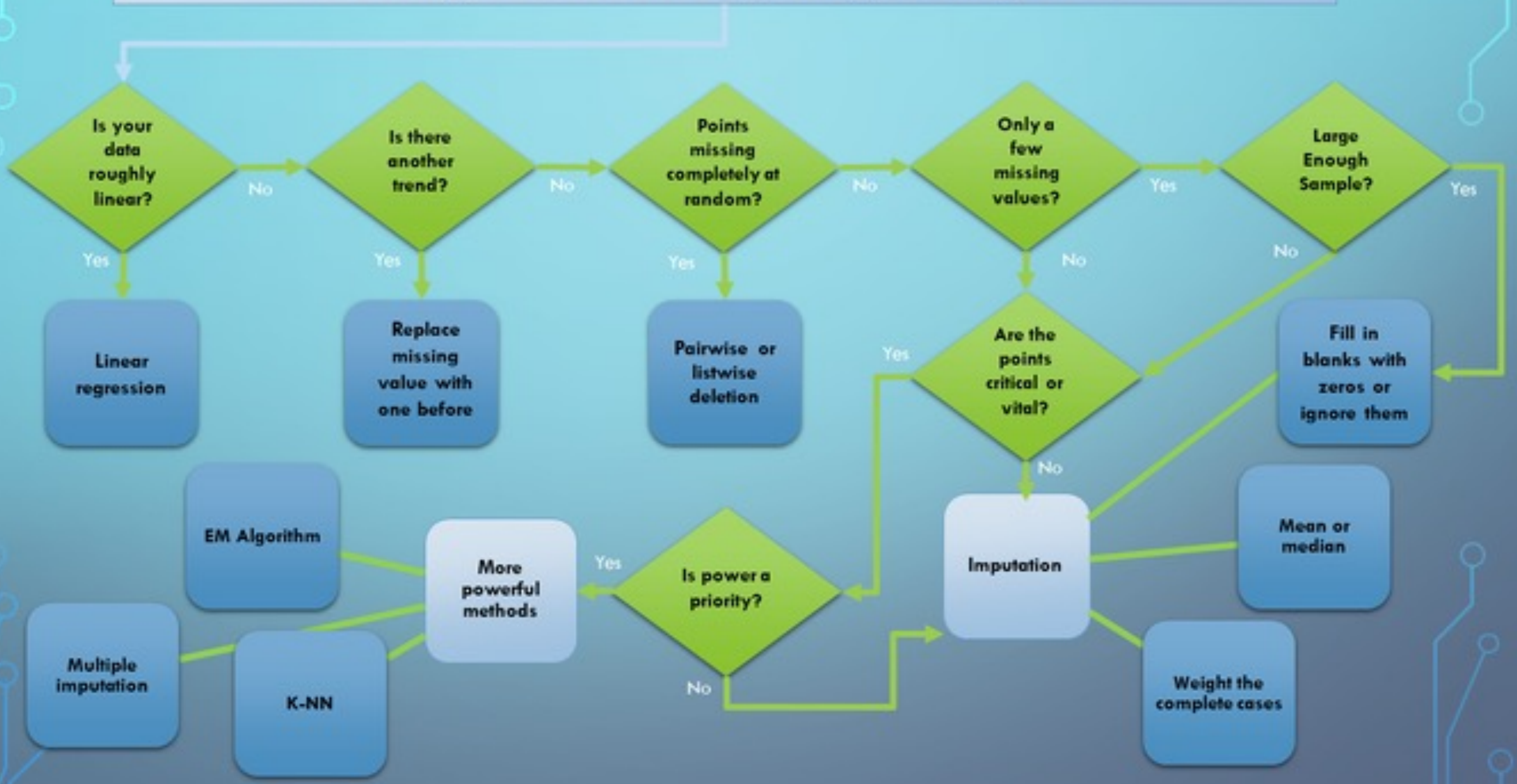
MISSING VALUE ESTIMATION

- Imputation is that the method of substituting missing data with substituted values. Two types of Imputations are majorly categorized
- General
- Time-Series
- **General Data**
- General data is mainly imputed by mean, mode, median, Linear Regression, Logistic Regression, Multiple Imputations, and constants.
- Further **General** data is divided into two types **Continuous** and **Categorical**.
- Here we are attending to take one dataset and that we shall apply some imputation techniques.

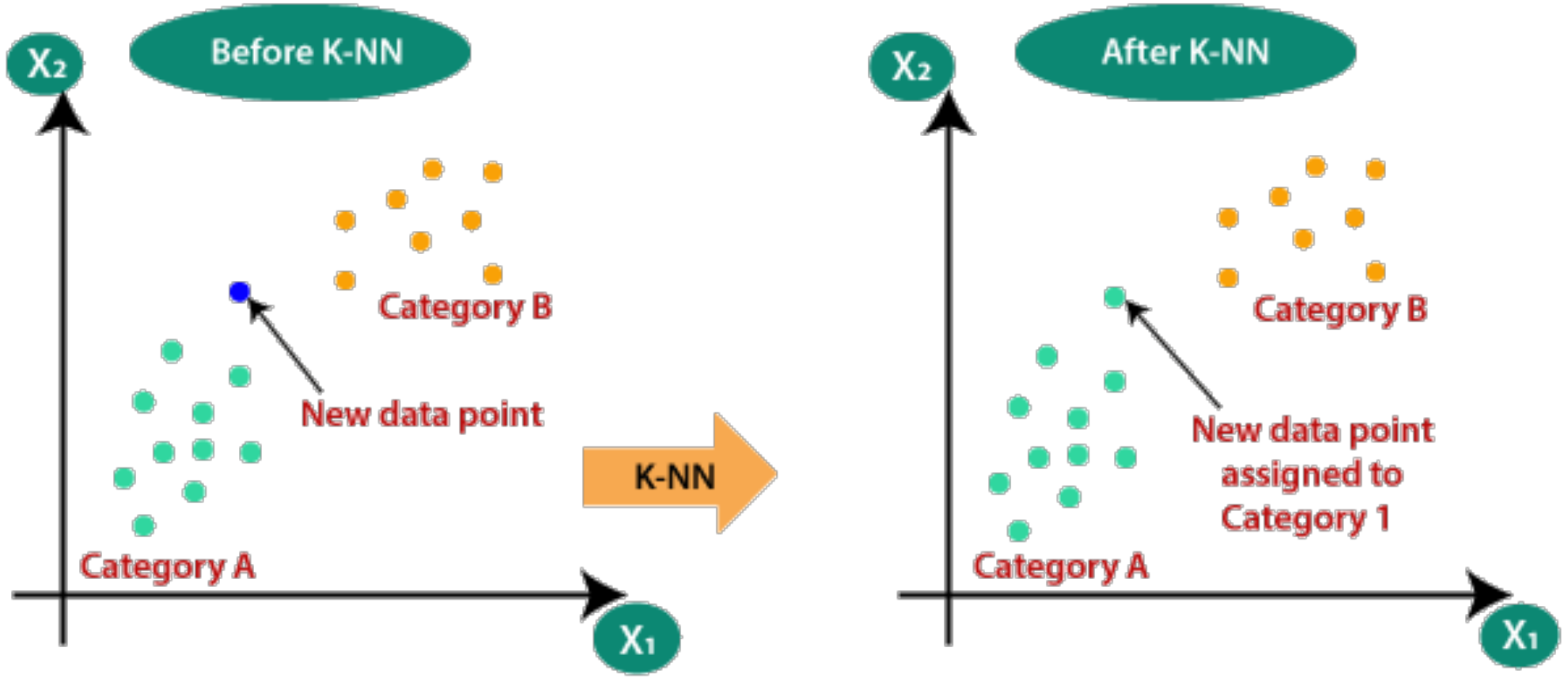
Example

0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	nan	6
3	4	Female	23	16	77
4	5	Female	nan	17	40
5	6	Female	22	17	76
6	7	Female	35	18	6
7	8	nan	23	18	94
8	9	Male	64	19	3
9	10	Female	30	19	72

Choosing a Method for Replacing Missing Values



KNN

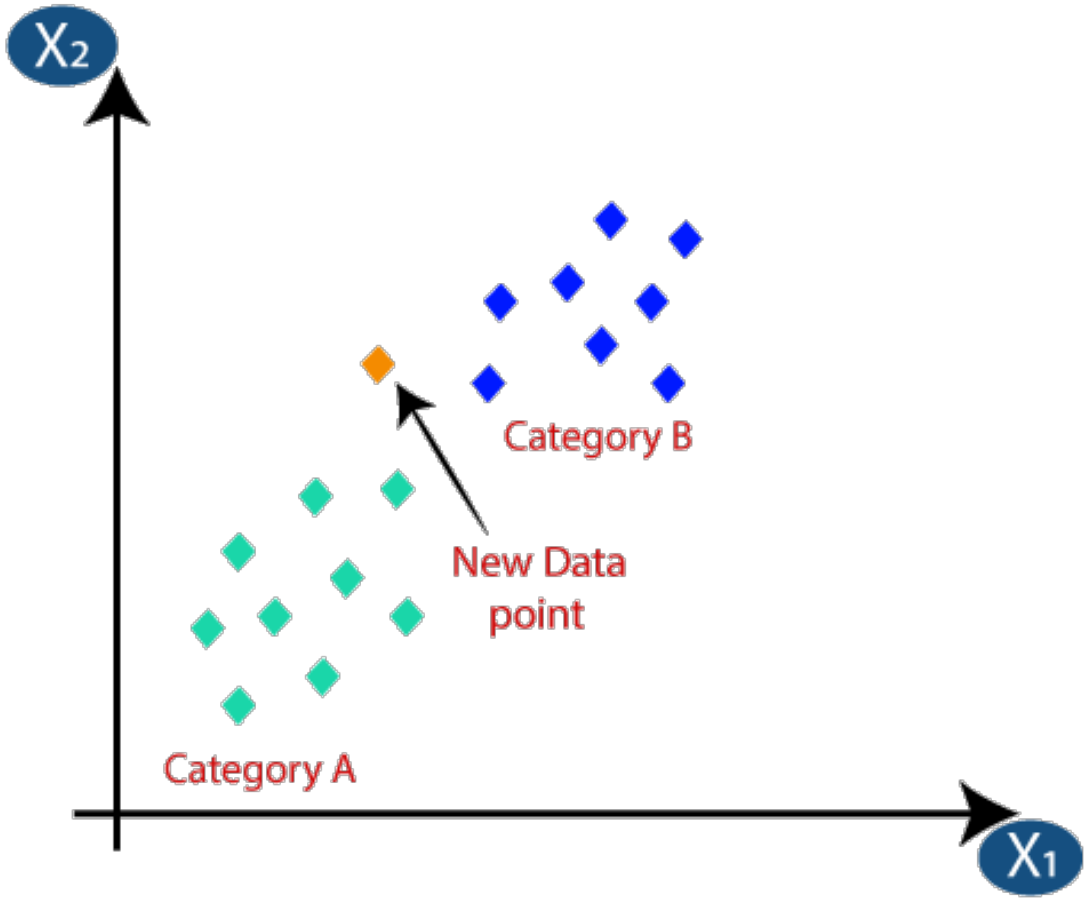

Before K-NN
X2
Category B
New data point
Category A
X1
K-NN
After K-NN
X2
Category B
New data point
assigned to
Category 1
Category A
X1

How does K-NN work?

The K-NN working can be explained on the basis of the below algorithm:

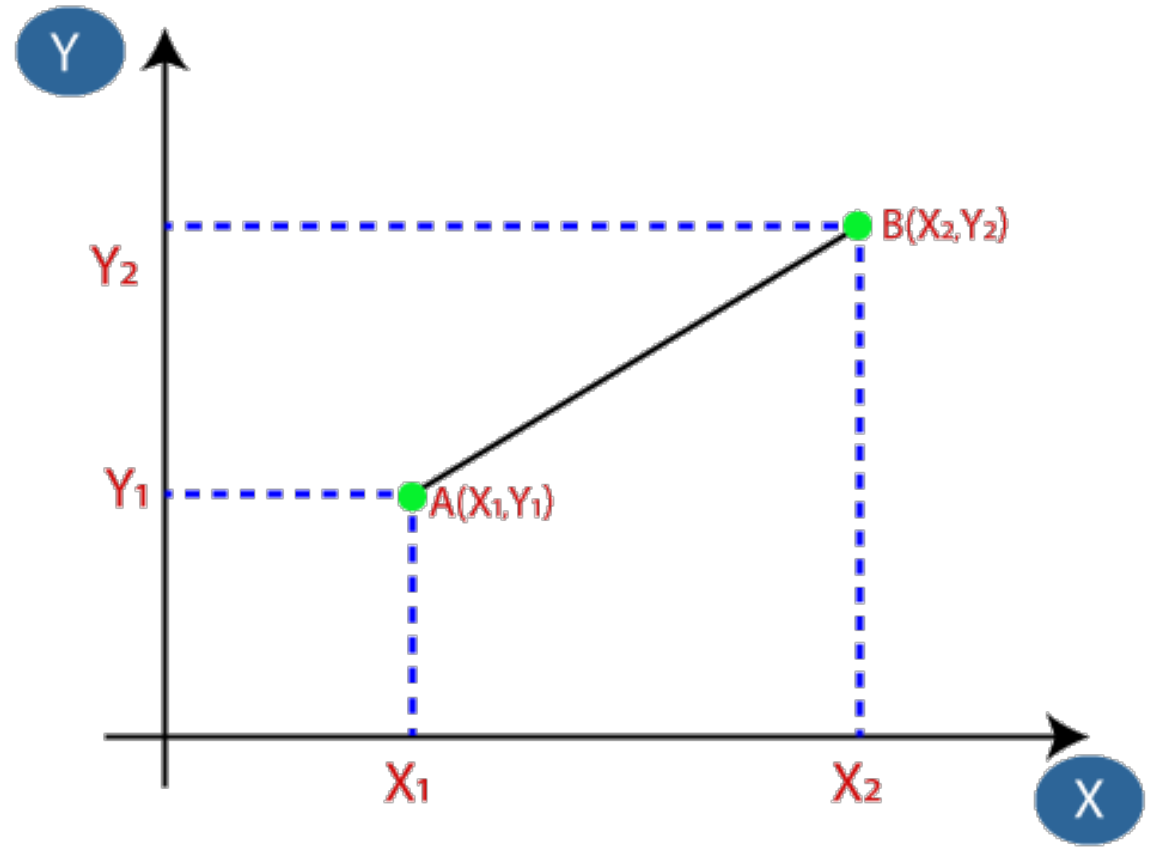
- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **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.

Suppose we have a new data point and we need to put it in the required category. Consider the below image:



•Firstly, we will choose the number of neighbors, so we will choose the k=5.

•Next, we will calculate the **Euclidean distance** between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



Euclidean Distance between A_1 and $B_2 = \sqrt{(X_2-X_1)^2+(Y_2-Y_1)^2}$

By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the image:

•As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

Distance functions

Euclidean

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan

$$\sum_{i=1}^{k}\left|x_i - y_i\right|$$

Minkowski

$$\left(\sum_{i=1}^{k}\left(\left|x_i - y_i\right|\right)^q\right)^{1/q}$$

How to select the value of K in the K-NN Algorithm?

- Below are some points to remember while selecting the value of K in the K-NN algorithm:
- There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
- A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
- Large values for K are good, but it may find some difficulties.

Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

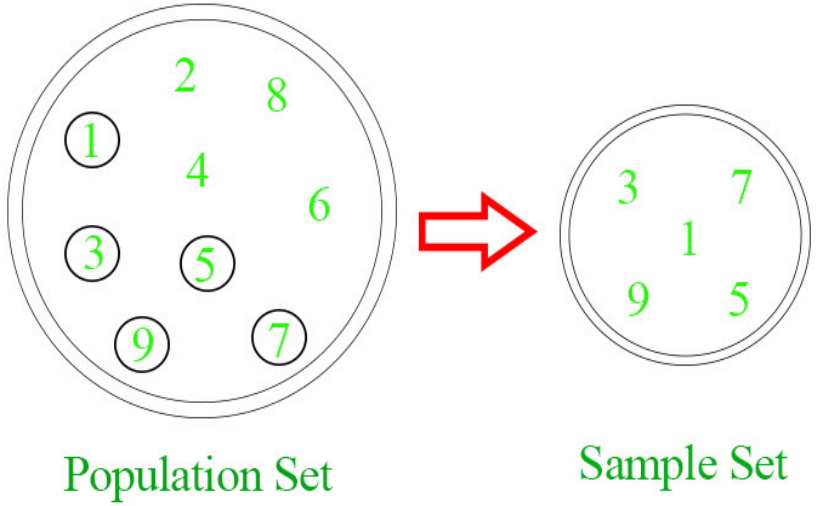
- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all the training samples.

Sampling

- In the world of Statistics, the very first thing to be done before any estimation is to create a Sample set from the entire Population Set.
- The Population set can be seen as the entire tree from where data is collected whereas the Sample Set can be seen as the branch in which the actual study of observations and estimation is done.
- Population tree is a very large set and making the study of observations on it can be very exhausting, both time and money-wise alike.
- Thus, to cut down on the amount of time and as well as resources, a Sample Set is created from the Population set.

Process of Sampling

1. Identifying the Population set.
2. Determination of the size of our sample set.
3. Providing a medium for the basis of selection of samples from the Population medium.
4. Picking out samples from the medium using one of many Sampling techniques like Simple Random, Systematic or Stratified Sampling.
5. Checking whether the formed sample set, contains elements actually matches the different attributes of population set, without large variations in between.
6. Checking for errors or inaccurate estimations in the formed sample set, that may or may not have occurred
7. The set which we get after performing the above steps actually contributes to the Sample Set.



- Population is the whole set of variables, elements, entities which are considered for a statistical study. It is also known as the universal set from where actual inferences are drawn. Population set consists of all the attributes of individuals or elements under consideration, but doing estimations on a Population is very exhausting resources as well as time-wise alike.

Example: Consider the mean weight of all men on Earth. This here, is considered a hypothetical population because it includes all men that have ever lived on earth which includes people who will exist in the future and also people who have lived earlier before us. But there comes an anomaly, while doing such measurement which is not all men in the population tray are observable (consider men, who will exist in the future and also men, who have lived before and doesn't exist right now). Also, performing statistics on the population sample (if hypothetically possible) would require a great deal of time as well as resources, which will be exhaustive and inefficient as well. Thus what is perform instead is to take a subset from the available population and perform statistics on them and interpolate inferences about the entire population. Taking out a subset, makes the task easier as the time required to scrutinize the subset is lesser than the time required to scrutinize the whole set of Population. Statistics is performed on the sample set to draw conclusions about the entire population tray. Calculations are considered to be a conclusion of the population set because it doesn't measure with the actual data of the population set and is not free from errors. This is obvious as sample set is used as a medium frame, having fewer members and thus some information is lost. (which results in errors).

Methods and Types of sampling:

1. Simple Random Sampling
2. Systematic Sampling
3. Stratified Sampling

These are the most widely used Sampling Processes with each having their both advantages as well as disadvantages. Let us look at each of these sampling methods in details:

1. **Simple Random Sampling**: Simple Random Sampling is the most elementary form of sampling. In this method, all the elements in populations are first divided into random sets of equal sizes. Random sets have no defining property among themselves, i.e one set cannot be identified from another set based on some specific identifiers. Thus, every element has an equal property of being selected.

P(of getting selected) = $< strong > \frac{1}{2} < /strong >$

- The basic methods for employing SRS are:
- Choose the Population Set
- Identify the basis of Sampling
- Use of random number/session generators to pick an element from each set.
- Less exhaustive with respect to time as it is the most elementary form of sampling
- Very useful for population set with very less number of elements
- SRS can be employed anywhere, anytime even without the use of special random generators
- Not efficient for large population sets
- Causes the most number of errors out of the three mentioned methods of sampling
- There are chances of bias and then SRS won’t be able to provide a correct result
- Does not provide a specific identifier to separate statistically similar samples

2. Systematic Sampling:

Systematic Sampling is also known as a type of probability sampling. It is much more accurate than SRS and also the standard error formation percentage is very low but not error-free. In this method, first, the population tray elements are arranged based on a specific order or scheme properly known as being sorted. It can be of any order, which totally depends upon the person performing the statistics. The elements are first arranged either ascendingly, descending, lexicographically or any other known methods deemed fit by the tester. Although the start point needs to be random every time. After being arranged, then the sample elements are picked based on a pre-defined interval set or function. **Example**: In a random set of numbers with elements ranging from 1 to 100. The elements are first sorted either in ascending or descending order. Then let's say every 4th element is picked to be a part of the sampling frame. This kind of sampling is known as Systematic Sampling.

- **P(of getting selected) = [depends upon the ordered population tray after it has been sorted]**

- The basic methods of employing Systematic Random Sampling are :-Choosing the Population Set wisely
- Checking whether Systematic Sampling will be the efficient method or not.
- If Yes, then Application of an sorting method to get an ordered pair of population elements.
- Choosing a periodicity to crawl out elements.
- Accuracy is higher than SRS.
- Standard probability of error is lesser .
- No problem for bias to creep in during creation of sample frame.
- Not much efficient when comes to the time wise
- Periodicity in population tray elements can lead to absurd results.
- Systematic sampling can either provide the most accurate result or an impossible one.

3. Stratified Sampling:

Stratified Sampling is the most complex type of Sampling Method out of all the three methods mentioned above. It is a hybrid method concerning both simple random sampling as well as systematic sampling. It is one of the most advanced types of sampling method available, providing near accurate result to the tester. In this method, the population tray is divided into sub-segments also known as stratum(singular). Each stratum can have their own unique property. After being divided into different sub-stratum, SRS or Systematic Sampling can be used to create and pick out samples for performing statistics.

The elementary methods for Stratified Sampling are :

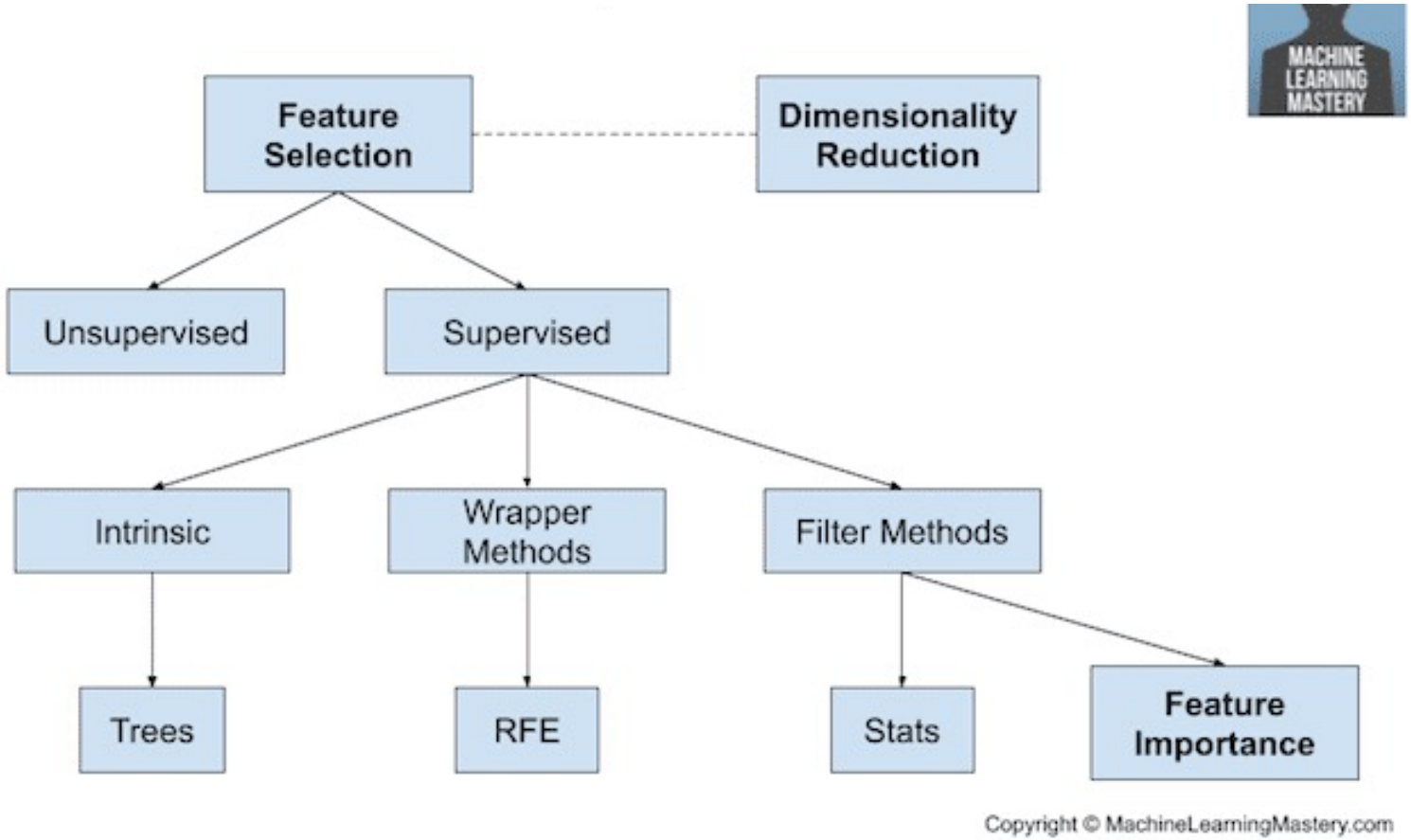
1. Choosing the population tray wisely.
2. Checking for periodicity or any other features, so that they can be divided into different strata
3. Dividing the population tray into sub-sets and sub-groups on the basis of selective property.
4. Using SRS or Systematic Sampling of each individual strata to form the sample frame.
5. We can even apply different sampling methods to different sub-sets.
6. Provide results with high accuracy measurements.
7. Different results can be desired just by changing the Sampling method.
8. This method also compares different strata when samples are being drawn.
9. Inefficient and Expensive when comes to resources as well as money.
10. This method will fail only in rare cases where homogeneity in elements is present.

- These three are the widely used methods of Sampling which are being done nowadays. Each of them has their own advantages as well as disadvantages. So, the sampling method must be chosen wisely, because a wrong choice can lead to erroneous answers.

Feature Selection

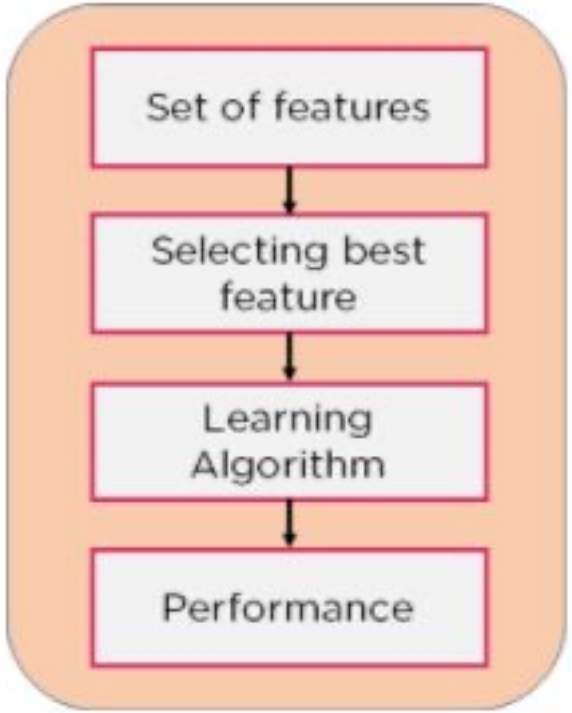
- We can summarize feature selection as follows.
- **Feature Selection**: Select a subset of input features from the dataset.
  - **Unsupervised**: Do not use the target variable for selecting the feature importance of Input variable (e.g. remove redundant variables).
    - Correlation
  - **Supervised**: Use the target variable (e.g. remove irrelevant I/P features).
    - **Wrapper Method**: Search for well-performing subsets of features.
      - Recursive Feature Elimination (RFE)
    - **Filter Method**: Select subsets of features based on their relationship with the target.
      - Statistical Methods
      - Feature Importance Methods
    - **Intrinsic**: Algorithms that perform automatic feature selection during training.
      - Decision Trees
- **Dimensionality Reduction**: Project input data into a lower-dimensional feature space.

Feature Selection



Feature Selection

- In machine learning and statistics, feature selection, also known as **variable selection**, **attribute selection** or **variable subset selection**, is the process of reducing the number of input variables when developing a predictive model.
- Feature selection techniques are used for several reasons:
- It reduces model complexity by dropping some irrelevant features.
- Helps ML algorithm to train a model faster.
- Reduction of dimensionality helps in avoid overfitting.

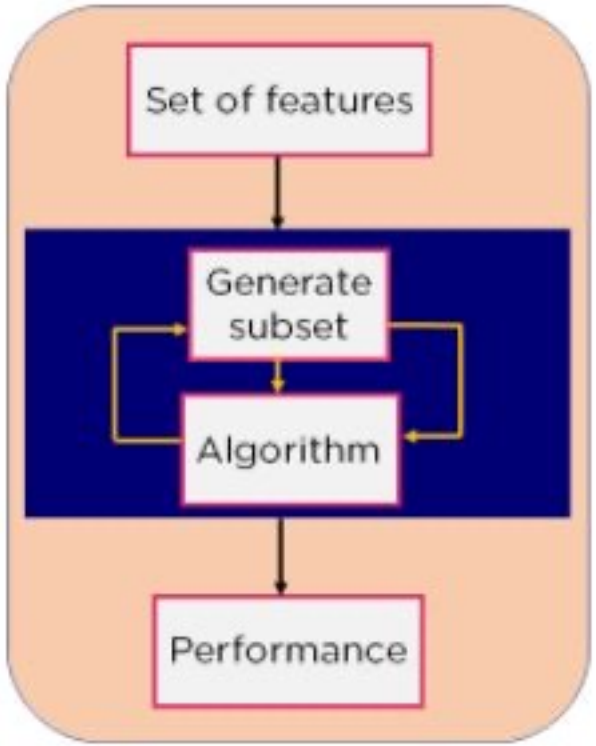


Methodologies used for feature selection.

1. Filter Method:

- Filter feature selection methods use statistical techniques to evaluate the relationship **between each input variable and the target variable**, and these scores are used as the basis to choose (filter) those input variables that will be used in the model.
- The statistical measures used in filter-based feature selection are generally calculated one input variable at a time with the target variable. As such, they are referred to as **univariate statistical measures**. This may mean that any interaction between input variables is not considered in the filtering process.
- **Note:-** In this case, the existence of correlated predictors makes it possible to select important, but redundant, predictors. The obvious consequences of this issue are that too many predictors are chosen and, as a result, collinearity problems arise.

- **2. Wrapper Method:**
- Wrapper feature selection methods create many models with different subsets of input features and select those features that result in the best performing model according to a performance metric.
- These methods are unconcerned with the variable types, although they can be computationally expensive.
- Recursive Feature Elimination (RFE) is a good example of a wrapper feature selection method.



1. **Forward Selection**: Forward selection is an iterative method in which we start with having no feature in the model. In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.

2. **Backward Elimination**: In backward elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.

3. **Recursive Feature elimination**: It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.

3. Embedded/Intrinsic Method:

- Embedded methods combine the qualities of filter and wrapper methods.