## *Disclaimer*

- This presentation is purely for academic purpose and does not carry any commercial value.
- All images and photos used in this presentation are property of respective image holder(s) and due credit is provided to them. Images are used only for indicative purpose and does not carry any other meaning.
- All information and data in this slide are collected from open domain.

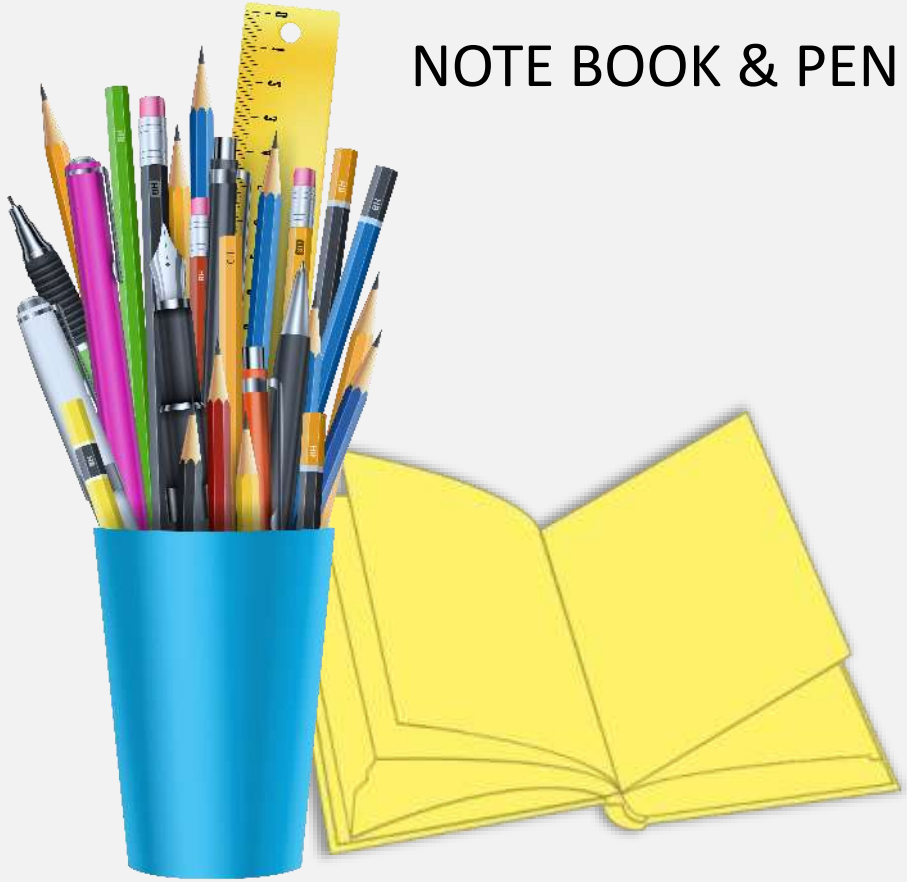
**MANISH GODSE, Ph.D.(IIT Bombay)**


# Request & Instructions

# PLEASE OPEN

CALCULATOR

NOTE BOOK & PEN



LAPTOP OR DESKTOP,
IF YOU HAVE.

# PLEASE FOLLOW THIS

SILENCE

MUTE MIC

RAISE HAND

NO CHAT

SILENT MODE

# DATA PREPARATION

# BOOKS & REFERENCES


- https://scikit-learn.org/stable/modules/preprocessing.html
- https://machinelearningmastery.com/model-based-outlier-detection-and-removal-in-python/

# Table of Contents





# DATA VECTOR-MATRIX AND TABLE

1

# VECTOR & MATRIX

**VECTOR**

- Vector is an ordered set of ‘n’ equal objects in order of ‘n’.
- Vector is a matrix having only one column or one row. In other words, vector is 1×n matrix.
- Vectors used in data analysis are scalar vectors.
- Example –

Row Vector, $a = \begin{bmatrix} 11 & 22 & 33 \end{bmatrix}$ Column vector, $a = \begin{bmatrix} 11 \\ 22 \\ 33 \end{bmatrix}$

# VECTOR & MATRIX

**MATRIX**

- A matrix is a rectangular array of numbers arranged in rows and columns.
- Example – Matrix of 2×3

$$\begin{bmatrix} 11 & 12 \\ 21 & 22 \\ 31 & 32 \end{bmatrix}$$

# VECTOR-MATRIX TO TABLE

- Data can be arranged in a **STRUCTURED** or **UNSTRUCTURED** form.
- The structured data is arranged in a **TABLE** or **MATRIX** form having **ROWS** and **COLUMNS**.

| Released Date | Movie | India Hindi Net | India Gross | Overseas | Worldwide | Budget | Profit | Verdict | % Profit on Investment |
|---|---|---|---|---|---|---|---|---|---|
| 02-Oct-19 | War | 303.34 | 375.00 | 96.00 | 471.00 | 150.00 | 321.00 | SuperHit | 214 |
| 30-Aug-19 | Saaho | 145.67 | 359.00 | 80.00 | 439.00 | 350.00 | 89.00 | Average | 25 |
| 21-Jun-19 | Kabir Singh | 278.80 | 330.00 | 47.00 | 377.00 | 55.00 | 322.00 | Blockbuster | 585 |
| 11-Jan-19 | URI The Surgical Strike | 244.14 | 293.75 | 48.00 | 341.75 | 70.00 | 271.75 | Blockbuster | 388 |
| 05-Jun-19 | Bharat | 212.03 | 251.00 | 70.00 | 321.00 | 165.00 | 156.00 | Hit | 95 |
| 27-Dec-19 | Good Newwz | 205.09 | 242.00 | 74.00 | 316.00 | 60.00 | 256.00 | Blockbuster | 427 |
| 25-Oct-19 | Housefull 4 | 210.30 | 247.00 | 49.00 | 296.00 | 175.00 | 121.00 | Hit | 69 |
| 15-Aug-19 | Mission Mangal | 203.08 | 236.00 | 54.00 | 290.00 | 40.00 | 250.00 | Blockbuster | 625 |

# DEPENDENT & INDEPENDENT VARIABLE

2

# INDEPENDENT AND TARGET VARIABLES

**DEPENDENT VARIABLE**

It is a variable whose value will change depending on the value of other variables. It is expected to change as the changes occur in independent variable.

**INDEPENDENT VARIABLE**

It is a variable that will not change by the other variable(s). A value of independent variable depends on the data gathered during real life scenarios or experiments.

**EXAMPLE**

Blood pressure will change because of salt intake, stress level and age. Hence in this example blood pressure is a dependent variable, while salt intake, stress level and age are independent variables.

***Variable is also named as FEATURE.***

# WHY DATA PREPARATION ?

3

# WHY DATA PREPARATION?

- The **raw data** may have quality issues and measurement errors. Hence data must be processed to make it **suitable for analytics**.
- The objectives of data processing are
  - To **improve** data **quality**
  - To **modify** the **data** so that it can fits to machine learning algorithm or technique.
- For example, a continuous attribute like length can be transformed into an attribute with discrete categories, such as short, medium, or long, in order to apply a particular technique. As another example, the

# DATA QUALITY

4

# DATA QUALITY

Data is often **not perfect** or as required to solve the problem. Though machine learning techniques can **tolerate** some level of **imperfection** in the data, however good data quality improves the **quality** of the **model output**. Hence data quality issues has to be addressed before using to construct the model.

Data quality problems occurs due to

- Human error in recording the data, like typo error.
- Limitations of data recording machines
- Incorrect or incomplete or old business processes and data collection mechanism

# DATA QUALITY

Quality problem includes

- Presence of **noise** and **outliers** in the data
- **Missing**, **inconsistent**, or **duplicate** data
- **Value** recorded may be **different** from true value.
- Data may be **biased** or, in some other way, **unrepresentative** of the population that the data is supposed to describe.

Data quality problems can be managed by below options.

- **Data Cleaning** - the detection and correction of data quality problems
- **Algorithms** - The use of algorithms that can tolerate poor data quality.

# DATA MEASUREMENT ERROR

5

# ERROR TYPE

- **MEASUREMENT ERROR**

  This error occurs due to problem in the measurement process.

- **DATA COLLECTION ERROR**

  This error is in the data object or attribute value like missing value, duplicate value, wrong value, etc.

# PRECISION, BIAS, AND ACCURACY

- Errors are measured by Precision, Bias, and Accuracy.
- **PRECISION**

  The precision is the closeness of repeated measurements to one another, when measured with the same quantity. Precision is measured by the standard deviation.

- **BIAS**

  A systematic variation of measurements from the quantity being measured. Bias is measured by taking the difference between the mean of the set of values and the known value of the quantity being measured.

- **ACCURACY**

  The accuracy is the closeness of measurements to the true value of the quantity being measured. Accuracy depends on precision and bias.

# NOISE & ARTIFACT

- **NOISE** is the random component of a measurement error in the data.

  It may be the distortion of a value or the addition of spurious objects. Hence noise is meaningless data . A system cannot understand and interpret noise correctly.
- **NOISY DATA or CORRUPT DATA** is the problem in machine learning algorithms as algorithms can think of noise to be a pattern and can start generalizing from it. Hence model will be faulty and output will not be correct.
- **ARTIFACT** is the deterministic distortions of the data.

# OUTLIER

Outliers are –

- Data objects having **characteristics different** from most of the other data.
- Values of an attribute are **unusual** with respect to the typical values for that attribute.

Outliers can be **legitimate data** objects or values. Outliers and noise are not the same.

In fraud and network intrusion detection, outliers are the very important as they are unusual objects or events from among a large number of normal ones. Hence in these problems outlier represent fraud and network intrusion. These problems falls under **anomaly detection**.

# MISSING VALUE

- It may be possible that data may be missing one or more attribute values.
- Data may be missing because
  - It is not **shared**. e.g., someone shared to decline to the age or weight.
  - It may be missing because of **human error**. e.g., someone forgot to fill the data in the form.
  - Data may be **skipped** if not mandatory. e.g., Someone has kept the age field blank as it was not not mandatory. Hence data is missing.

# DUPLICATE DATA

- A data set may have data objects that are **duplicates**.
- Sometime data duplicate may not be the same. For example, two persons having the same name but may be living at two different locations.
- The term **deduplication** is often used to refer to the process of dealing with these issues.

# INCONSISTENT DATA

- Data may have inconsistent values.
- Inconsistencies may occur due to human or system error. E.g., Height shown as negative, weight shown below zero value are the examples of inconsistent data.

# DATA PREPARATION TASKS

6

# DATA PREPARATION TASKS

Data preparation is the process of **CLEANING** and **TRANSFORMING** raw data so that it is useful in analytics. Data has to be read and analyzed so that it can be prepared for analytics.

Data preparation tasks are as below.

- **DATA CLEANING** - Identify and correct errors in the data.
- **FEATURE SELECTION** - Identify input variables which are most relevant to the task.
- **DATA TRANSFORMATION** - Change the scale or distribution of variables.
- **FEATURE ENGINEERING** - Derive new variables from available data.
- **DIMENSIONALITY REDUCTION** - Creating compact projections of the data.

# DATA CLEANING

7

# DATA CLEANING

- Data cleaning process to remove quality problems or errors in the data.
- When combining data from multiple sources, there are chances of data duplication, mislabeled data, missing data. This data need to corrected as it may lead to **incorrect** or **unreliable outcome from the model**.
- The data cleaning process varies from dataset to dataset.

Data cleaning tasks are as below.

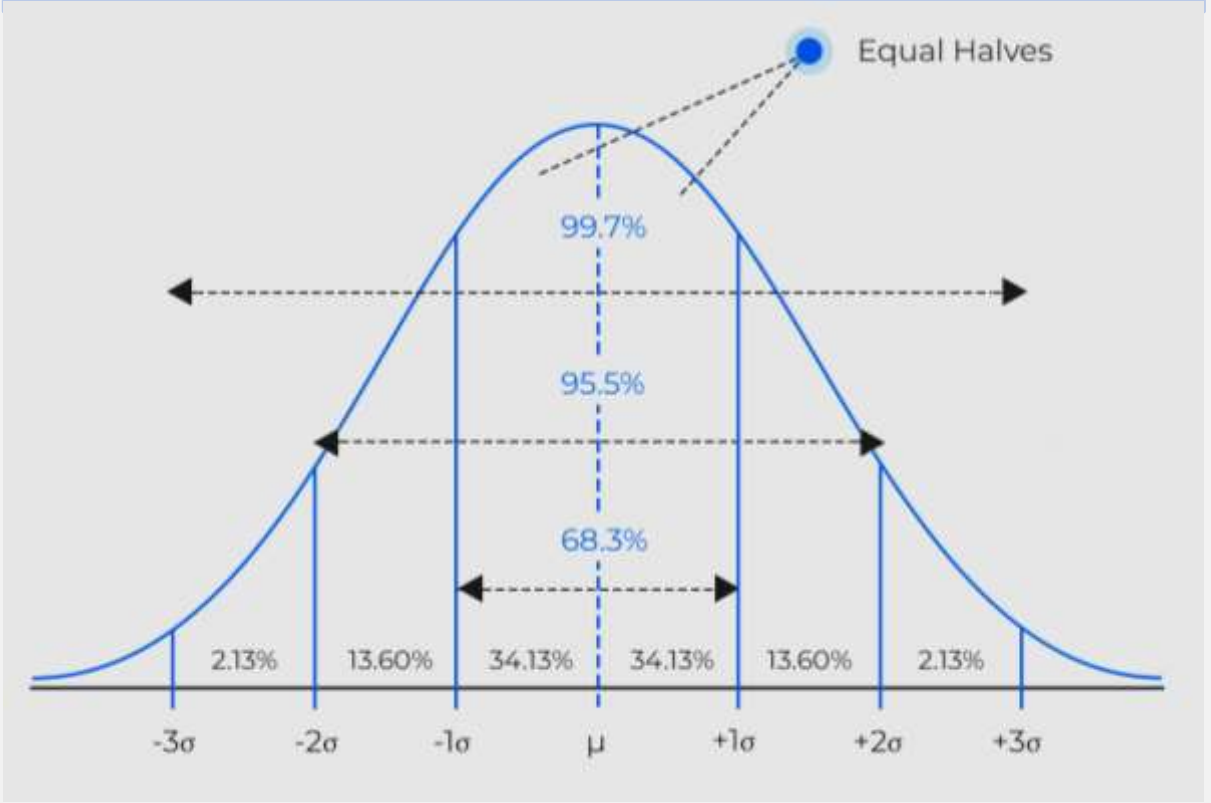
- Remove **DUPLICATE** or Irrelevant **Observations**
- Handle **MISSING VALUES**
- Manage **OUTLIERS**

# NORMAL DISTRIBUTION

In many practical applications, data sets exhibit a **symmetric mound-shaped** or **bell-shaped distribution**. For data having a bell-shaped distribution:

- Approximately 68% of the data values will be within one standard deviation of the mean.
- Approximately 95% of the data values will be within two standard deviations of the mean.
- Almost all of the data values will be within three standard deviations of the mean.

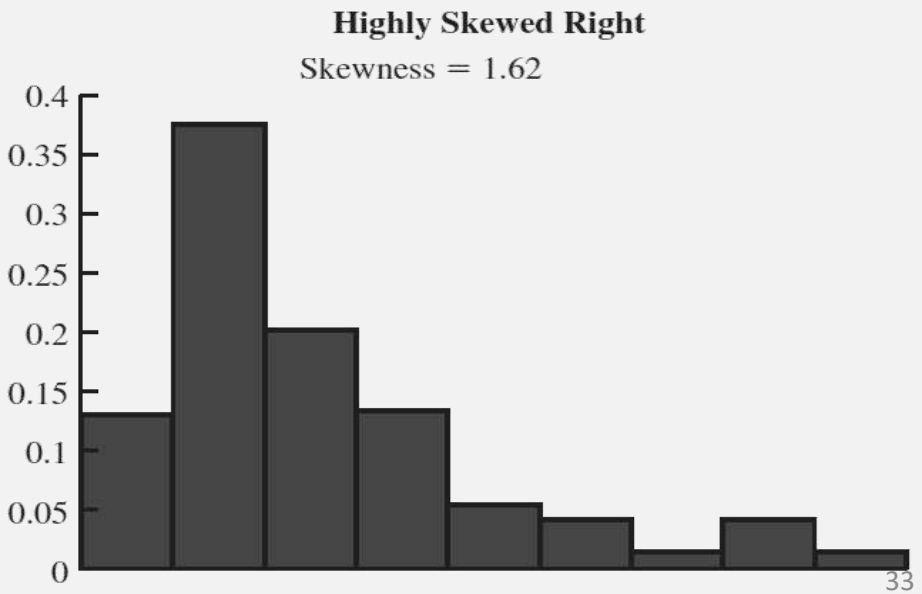
**A Symmetric Mound-shaped or Bell-shaped Distribution**



# SKEWNESS

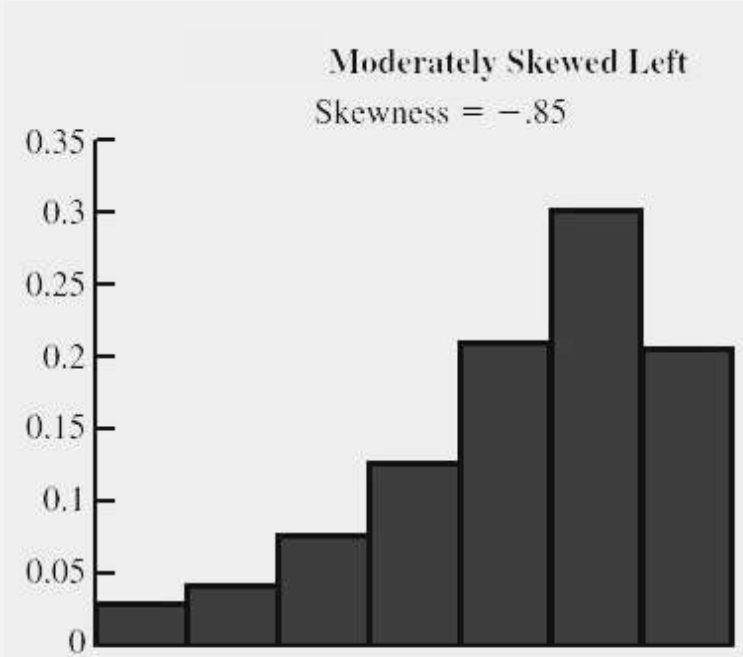
- For a **symmetric** distribution, mean and median are equal.
- When the data are **positively** skewed, the mean will usually be greater than the median.
- When the data are **negatively** skewed, the mean will usually be less than the median.
- The **median** provides the preferred measure of location when the data are highly skewed.

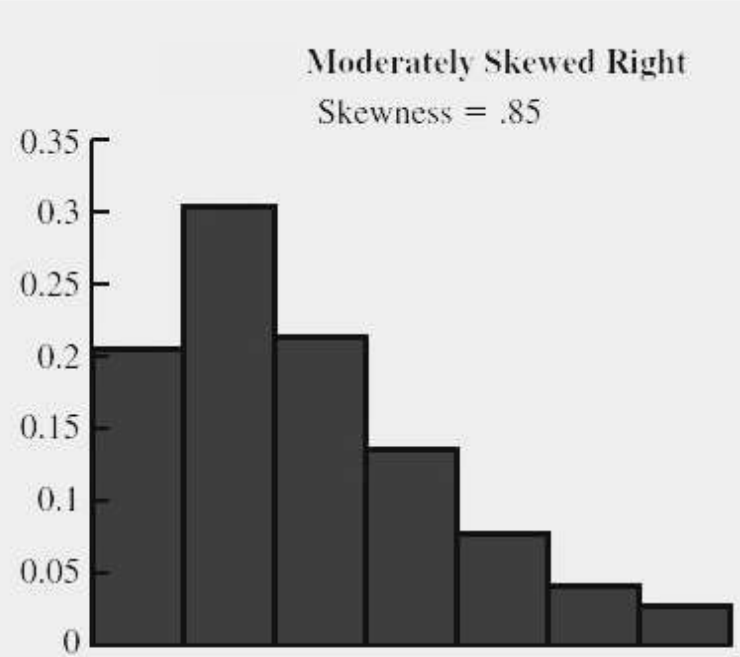
An important numerical measure of the shape of a distribution is called **skewness.**

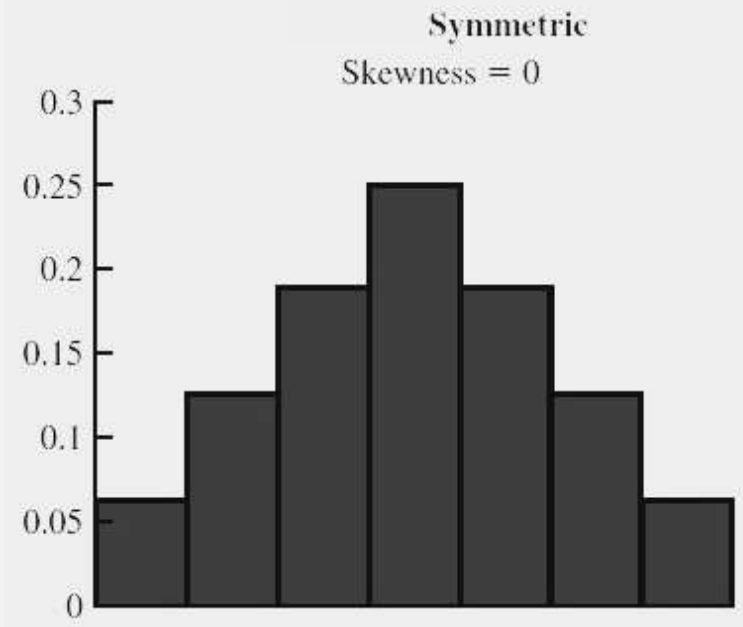


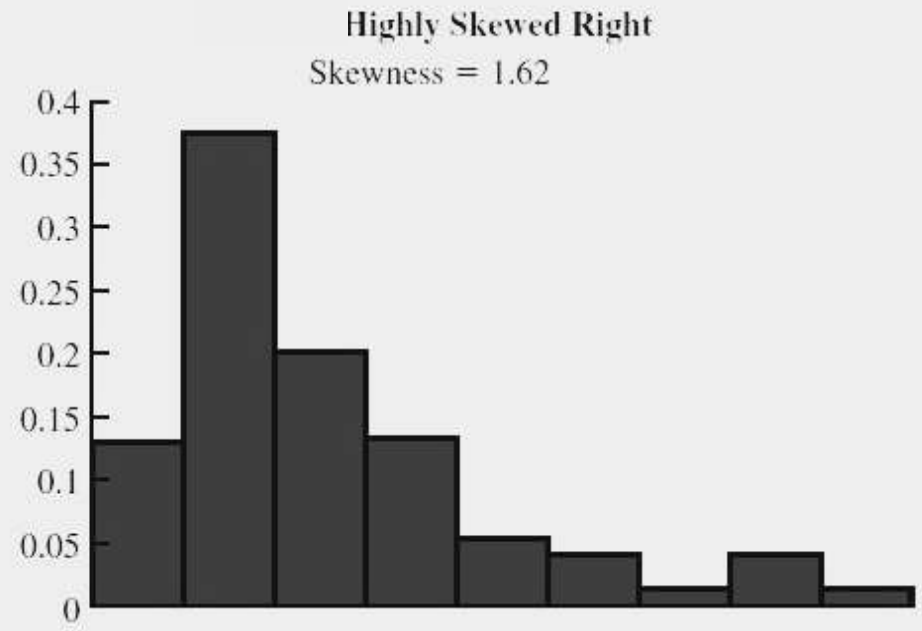

Histogram in Panel D are customer purchases at a women's apparel store. The mean purchase amount is $77.60 and the median purchase amount is $59.70. The relatively few large purchase amounts tend to increase the mean, while the median remains unaffected by the large purchase amounts.

# SKEWNESS EXAMPLES



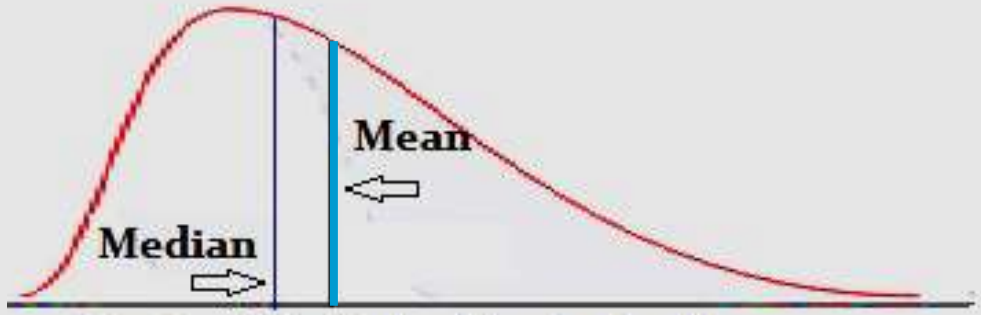



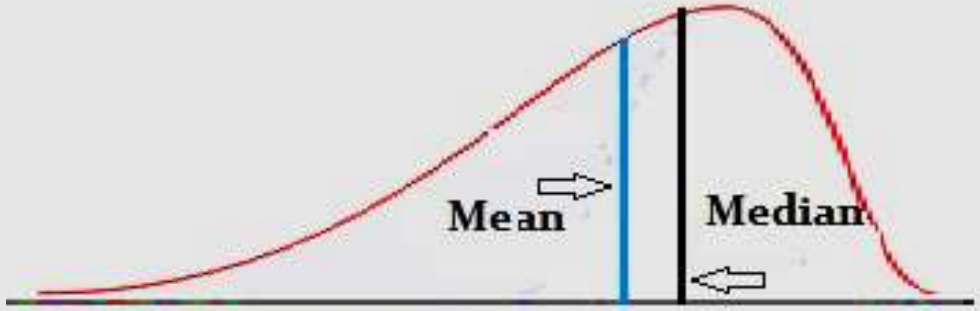
# SKEWNESS

- For data skewed to the right, the skewness is **positive**.
  **MEAN > MEDIAN**
- For data skewed to the left, the skewness is **negative**.
  **MEAN < MEDIAN**
- If the data are symmetric, the skewness is **zero**.
  **MEAN = MEDIAN**

$$\text{Skewness} = \frac{n}{(n-1)(n-2)} \sum \left( \frac{x_i - \bar{x}}{s} \right)^3$$




Right Skewed Distribution : Mean is on the right side of Median.



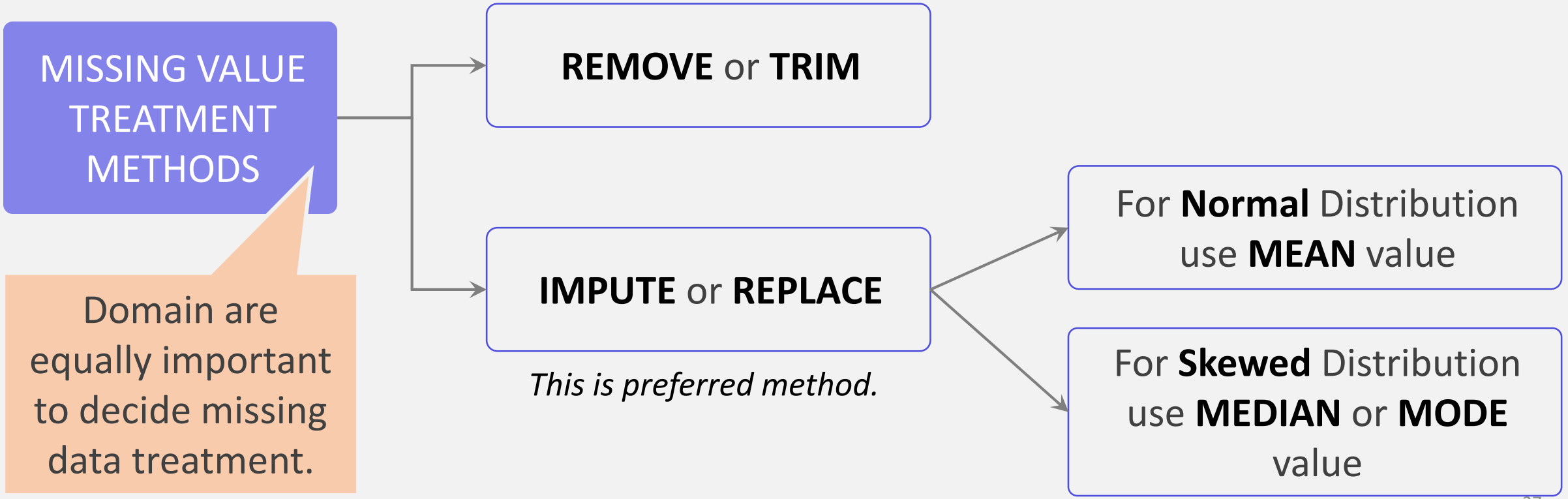

Left Skewed Distribution : Mean is on the left side of Median.

# MEAN OR MEDIAN, WHICH IS BETTER?

- The mean is influenced by **extremely small** and **large data values**.
- With the extremely high starting salary included, the median provides a better measure of central location than the mean.
- We can generalize to say that whenever a data set contains **extreme values**, the **median** is often the **preferred measure** of central location.

# MISSING DATA TREATMENT

Missing data can be trimmed or imputed. The central tendency measures such as mean, median or mode are used for imputation. However, they can't be used randomly. Data distribution and domain knowledge is important to decide use of mean, median or mode for the imputation of numeric features.



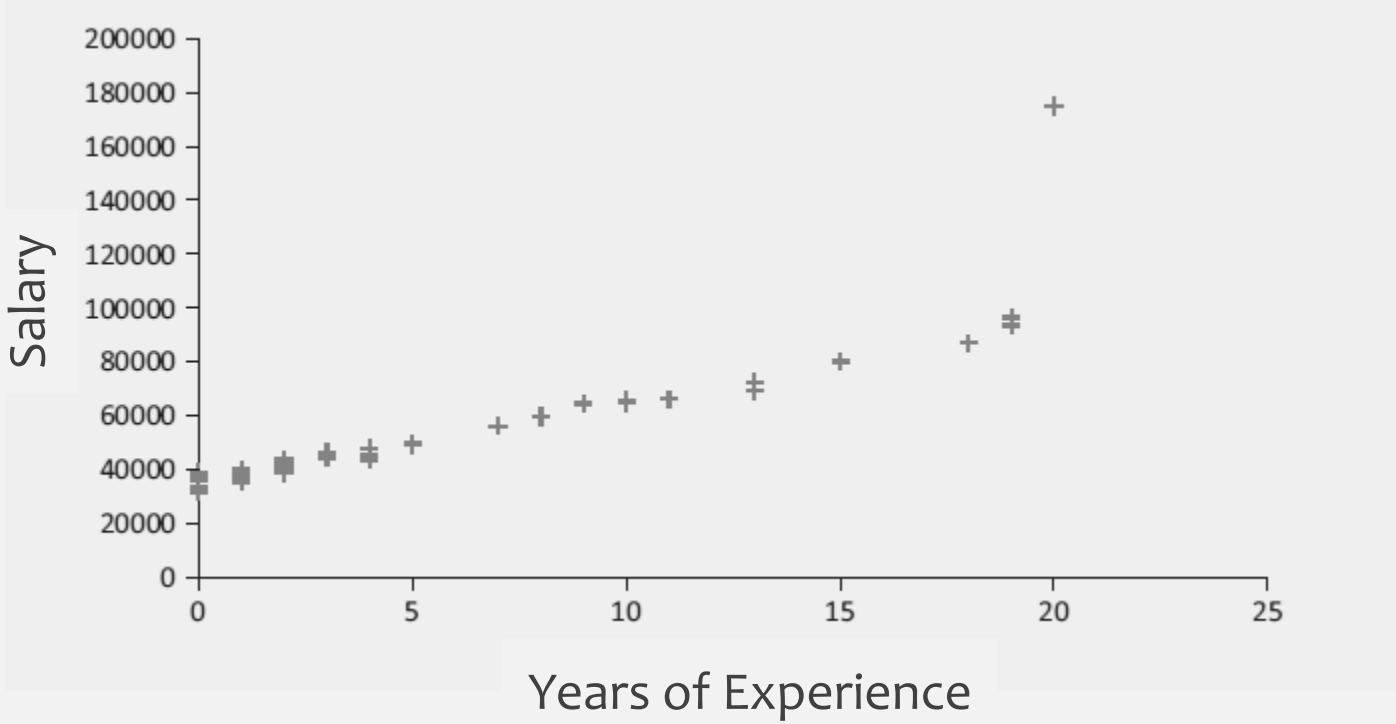
# OUTLIER

Sometimes a data set will have one or more observations with **unusually large** or **unusually small** values. These extreme values are called **outliers**.

- An outlier may be a data value that has been **incorrectly recorded**. If so, it can be **corrected** before further analysis.
- An outlier may also be from an observation that was **incorrectly included** in the data set; if so, it can be **removed**.
- Finally, an outlier may be an unusual data value that has been **recorded correctly** and belongs in the data set. In such cases it should **remain**.

Standardized values (*z*-scores) can be used to identify outliers.

# EXAMPLE - OUTLIER




**OUTLIER & ACTION**

1. If an outlier is clearly not a member of the population of interest, then it is best to delete it from the analysis.
2. If it isn't clear whether outliers are members of the relevant population, then it is better impute or replace it.

CEO's salary is not determined in the same way as the salaries for typical employees. Hence it is outlier. It can be deleted as it is not representing population.

# OUTLIER AND Z-SCORE

- It is a good idea to check for outliers before making decisions based on data analysis.
- Outliers should not necessarily be deleted, but their accuracy and appropriateness should be verified.
- The empirical rule allows us to conclude that for data with a bell-shaped distribution, almost all the data values will be within **three standard deviations of the mean**. Hence, in using *z*-scores to identify **outliers**, we recommend treating any data value with a *z*-score **less than 3 or greater than 3 as an outlier**. Such data values can then be reviewed for accuracy and to determine whether they belong in the data set.

# FEATURE SELECTION

8

# FEATURE SELECTION

- This process helps to **identify important input** variables that impact or are relevant to output of modelling.
- This process support to **select a subset** of input features from the dataset.
- This process helps in **reducing the number of input variables**. Hence improving the performance of the model and reducing computational cost.
- Feature selection methods are as below.
  - **UNSUPERVISED**: Correlation method
  - **SUPERVISED**: Use the target variable (e.g. remove irrelevant variables).
    - **Wrapper**: Search subset using Recursive Feature Elimination (RFE)
    - **Filter**: Select subsets of features based on their relationship with the target using Statistical Methods or Feature Importance Methods
    - **Intrinsic**: Automatic feature selection using Decision Trees

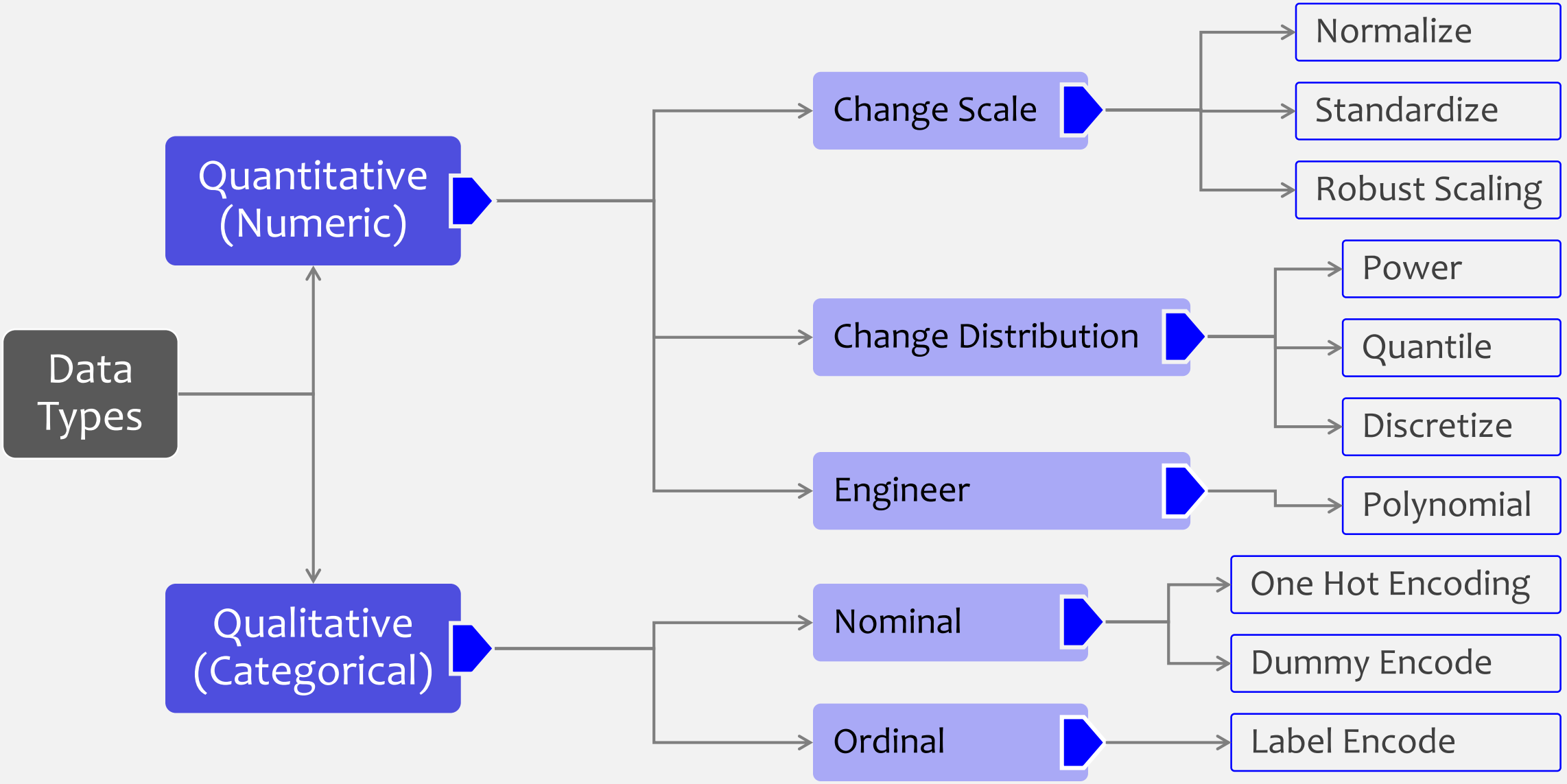
# DATA TRANSFORMATION

9

# DATA TRANSFORMATION

- Data transformation is used to change the data type or distribution of data variables.
- The technique used in transformation depends on data type. Below is the summary of data types.
- **NUMERIC** Data Type: Data presented as numbers.
  - **INTEGER**: Number without fractional part. E.g. 1 , 2 , 3 , 4 , 5
  - **REAL**: Number with fractional part. E.g. 0.1 , 2.1 , 3.0 , 0.44 , 5.56
- **Categorical** Data Type: Label values.
  - **ORDINAL**: Data with ordered labels. E.g. $1^{st}$, $2^{nd}$, $3^{rd}$, $4^{th}$ , $5^{th}$
  - **NOMINAL**: Labels with no rank ordering.
  - **BOOLEAN**: Values True and False.

# DATA TRANSFORMATION TECHNIQUES




*Encode is to convert a categorical variable to number.*

# CHANGE SCALE OR DATA SCALING

- **RESCALING** of a data vector is to add or subtract a constant and then multiply or divide by a constant. It changes the units of measurement of the data. E.g. Convert a temperature from Celsius to Fahrenheit.
- **NORMALIZATION** is the rescaling of a data vector by the minimum and range of the vector. It converts a data elements between 0 and 1. Normalization requires accurate estimate of minimum and maximum values. Outliers may impact normalization.
- **STANDARDIZATION** of a data vector means subtracting a measure of location and dividing by a measure of scale. It converts a random variable with mean 0 and standard deviation 1. For example, if the vector contains random values with a Gaussian distribution, then subtract the mean and divide by the standard deviation to standardize the vector.

# WHEN & WHY STANDARDIZATION?

- Standardization is used, when data has input values with differing scales.
- Standardization assumes that observations fit a Gaussian distribution (bell curve). If data standardized for not well bell curve, then result may not be reliable.
- Standardization of input or target variables has to be done cautiously as it discards important information, However it also makes the training process better.
- It is common to standardize each input to the same range or the same standard deviation.
- If some inputs are more important than others, it may help to scale the inputs such that the more important ones have larger variances and/or ranges.

# FEATURE ENGINEERING

10

# FEATURE ENGINEERING

- Feature engineering is a process of **CREATING NEW VARIABLE** form the existing data.
- In other words, it derives new variables from existing variables.
- Being new variable is created, knowledge of data and domain is must. New features also depends on the quality of data and problem to be solved using the data.
- Why feature engineering?
  1. Identify better input variables, hence better outcome of model.
  2. Better features reduces complexity, enhances flexibility and reduces computational cost of modeling.

# DIMENSIONALITY REDUCTION

11

# DIMENSIONALITY REDUCTION

- **DIMENSIONALITY REDUCTION** is the process of reducing the number of input variables required in machine learning model.
- Dimensionality reduction creates compact projections of the data.
- If input variables are more, then the dataset representation may very sparse. It may be possible that this spared sample dataset may not be representing population correctly. This is referred to as the **CURSE OF DIMENSIONALITY**.
- Dimensionality reduction methods –
  1. **Feature Selection**
  2. **Matrix Factorization Method ( or Linear Algebra Method)**
  3. **Manifold Learning Method (or Projection Method)**
  4. **Autoencoders**

# FEATURE SELECTION METHODS

These methods use scoring or statistical methods to select features. It also helps in deleting less important features. Thus helps in reducing dimensionality.

Feature selection techniques has two methods - Wrapper Methods and Filter Methods.

- Wrapper Methods
    1. Recursive Feature Elimination (RFE)
- Filter Methods
    1. Pearson's correlation
    2. Chi-Squared test

# MATRIX FACTORIZATION METHODS

This uses linear algebra based matrix factorization to reduce dimensionality. Matrix factorization reduces a dataset matrix into its constituent parts.

Methods in matrix factorization are -

1. Principal Components Analysis (PCA)
2. eigendecomposition
3. Singular Value Decomposition

# MANIFOLD LEARNING METHODS

Manifold learning method (or projection method) is used to create a low-dimensional projection of high-dimensional data.

Low-dimensional representation preserves the salient structure or relationships in the data.

Manifold learning techniques are -

1. Kohonen Self-Organizing Map (SOM)
2. Sammons Mapping
3. Multidimensional Scaling (MDS)
4. t-distributed Stochastic Neighbor Embedding (t-SNE)

# AUTOENCODERS

Deep learning (neural networks) methods are used to perform dimensionality reduction.

Manifold learning techniques are -

1. LSTM Autoencoders

# QUESTION AND ANSWERS