NARSIMHA REDDY ENGINEERING COLLEGE

DEPARTMENT- CSE

INTRODUCTION TO DATA SCIENCE

P. LAKSHMI PRASANNA

ASST.PROFESSOR

UNIT I

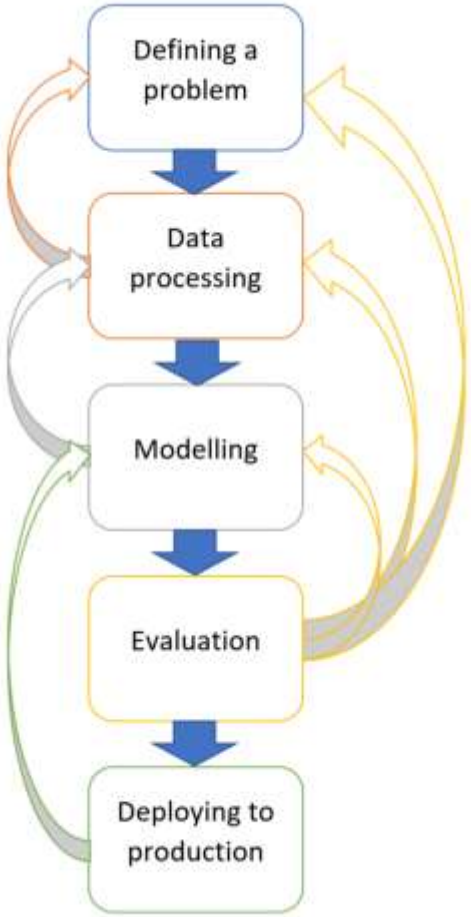
Data science:

- Data science is the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions. Data science uses complex machine learning algorithms to build predictive models.

Roles in Data Science

- Data Analyst
- Data Engineers
- Database Administrator
- Machine Learning Engineer
- Data Scientist
- Data Architect
- Statistician
- Business Analyst
- Data and Analytics Manager

Stages in a data science project



Applications of data science in various fields

- In Search Engines
- In Transport
- In Finance
- In E-Commerce
- In Health Care
- Image Recognition
- Targeting Recommendation
- Medicine and Drug Development

Data security issues

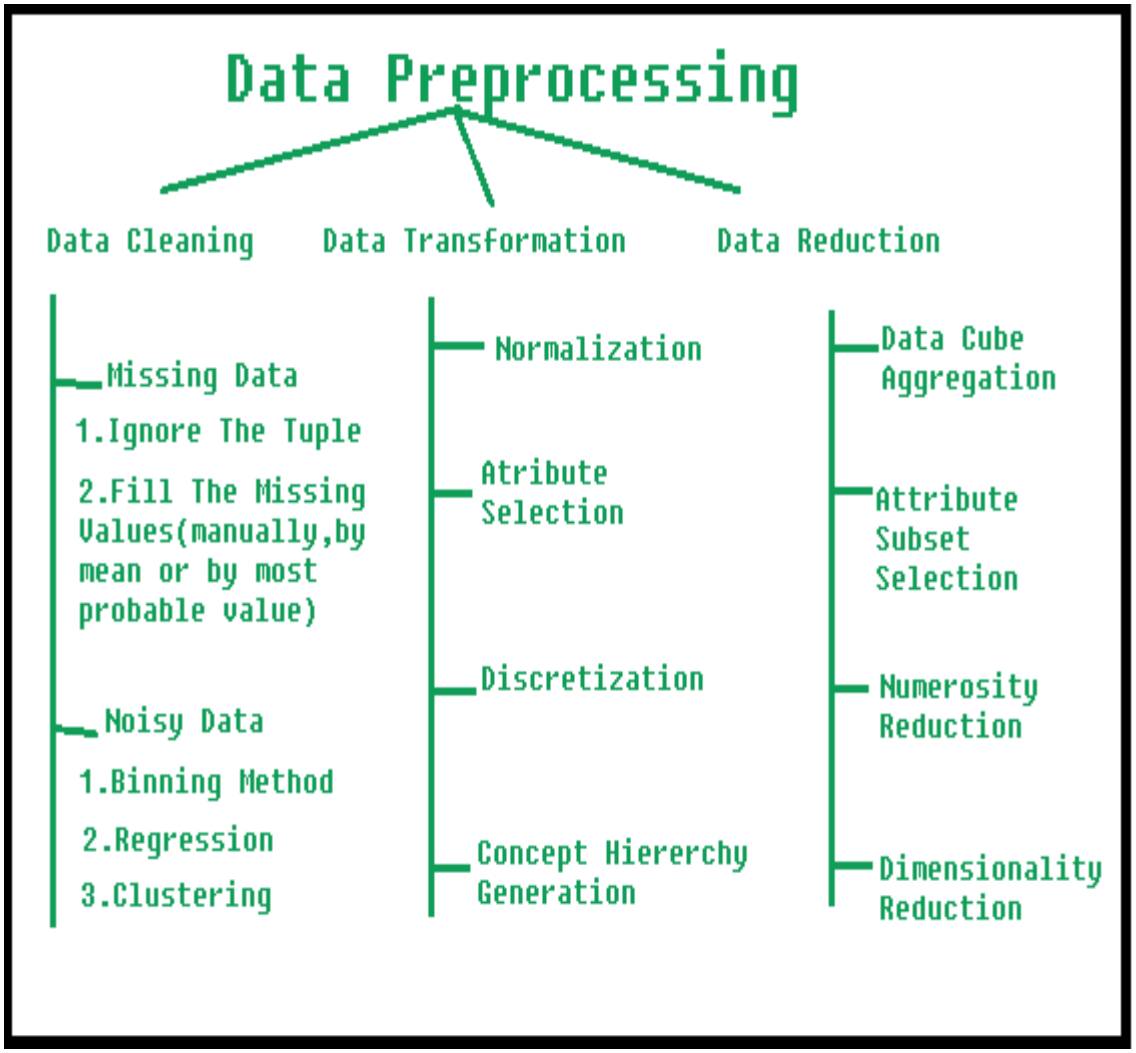
- Data security is the process of protecting corporate data and preventing data loss through unauthorized access. This includes protecting your data from attacks that can encrypt or destroy data, such as ransomware, as well as attacks that can modify or corrupt your data. Data security also ensures data is available to anyone in the organization who has access to it.
- **Access control**—ensuring that anyone who tries to access the data is authenticated to confirm their identity, and authorized to access only the data they are allowed to access.
- **Data protection**—ensuring that even if unauthorized parties manage to access the data, they cannot view it or cause damage to it. Data protection methods ensure encryption, which prevents anyone from viewing data if they do not have a private encryption key, and data loss prevention mechanisms which prevent users from transferring sensitive data outside the organization.

UNIT –II
DATA COLLECTION AND PREPROCESSING

- **DATA COLLECTION:**
- Data collection is the process of collecting, measuring and analyzing different types of information using a set of standard validated techniques. The main objective of data collection is to gather information-rich and reliable data, and analyze them to make critical business decisions.
- There are two main methods of data collection in research based on the information that is required, namely:
- Primary Data Collection
- Secondary Data Collection

Data preprocessing

- **key steps in data preprocessing**
- Data profiling
- Data cleansing
- Data reduction
- Data transformation
- Data enrichment
- Data validation


Data Preprocessing
Data Cleaning
Data Transformation
Data Reduction
Missing Data
1.Ignore The Tuple
2.Fill The Missing Values(manually,by mean or by most probable value)
Noisy Data
1.Binning Method
2.Regression
3.Clustering
Normalization
Atribute Selection
Discretization
Concept Hiererchy Generation
Data Cube Aggregation
Attribute Subset Selection
Numerosity Reduction
Dimensionality Reduction

Data cleaning

- Data cleaning is fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.
- **Steps of Data Cleaning**
- 1. Remove duplicate or irrelevant observations
- 2. Fix structural errors
- 3.Filter unwanted outliers
- 4. Handle missing data
- 5. Validate and QA

Methods of Data Cleaning

- **Ignore the tuples**
- **Fill the missing value**
- **Binning method**
- **Regression**
- **Clustering**

Data Integration

- **Data Integration** is a data preprocessing technique that combines data from multiple heterogeneous data sources into a coherent data store and provides a unified view of the data. These sources may include multiple data cubes, databases, or flat files.
-
- The data integration approaches are formally defined as triple <G, S, M> where,
  G stand for the global schema,
  S stands for the heterogeneous source of schema,
  M stands for mapping between the queries of source and global schema.

- There are mainly 2 major approaches for data integration – one is the “tight coupling approach” and another is the “loose coupling approach”.
- **Tight Coupling:**
- Here, a data warehouse is treated as an information retrieval component.
- In this coupling, data is combined from different sources into a single physical location through the process of ETL – Extraction, Transformation, and Loading.
- **Loose Coupling:**
- Here, an interface is provided that takes the query from the user, transforms it in a way the source database can understand, and then sends the query directly to the source databases to obtain the result.
- And the data only remains in the actual source databases.

Data Transformation

- Data transformation is a technique used to **convert** the raw data into a suitable format that efficiently eases data mining and retrieves strategic information. Data transformation includes data cleaning techniques and a data reduction technique to convert the data into the appropriate form. Data transformation changes the format, structure, or values of the data and converts them into clean, usable data
- Data Transformation Techniques
- Data smoothing
- Attribute construction
- Data aggregation
- Data normalization
- Data discretization
- Data generalization

Data Reduction

- ***Data reduction*** techniques ensure the integrity of data while reducing the data. Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume.
- Techniques of Data Reduction
- Dimensionality reduction
- Numerosity reduction
- Data cube aggregation
- Data compression
- Discritization operation

Data Discretization

- Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss.
- There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization.

- Some Famous techniques of data discretization
- **Histogram analysis**
- Histogram refers to a plot used to represent the underlying frequency distribution of a continuous data set. Histogram assists the data inspection for data distribution. For example, Outliers, skewness representation, normal distribution representation, etc.
- **Binning**
- Binning refers to a data smoothing technique that helps to group a huge number of continuous values into smaller values. For data discretization and the development of idea hierarchy, this technique can also be used.
- **Cluster Analysis**
- Cluster analysis is a form of data discretization. A clustering algorithm is executed by dividing the values of x numbers into clusters to isolate a computational feature of x.

UNIT 3
DESCRIPTIVE STATISTICS

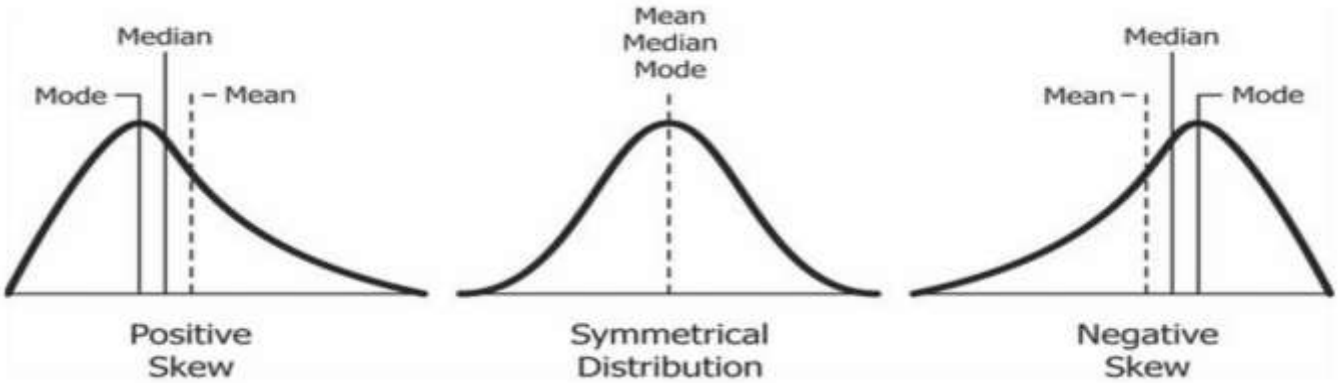
- **Descriptive statistics:**
- Descriptive statistics describe, show, and summarize the basic features of a dataset found in a given study, presented in a summary that describes the data sample and its measurements. It helps analysts to understand the data better.
- Descriptive statistics represent the available data sample and does not include theories, inferences, probabilities, or conclusions. That’s a job for inferential statistics

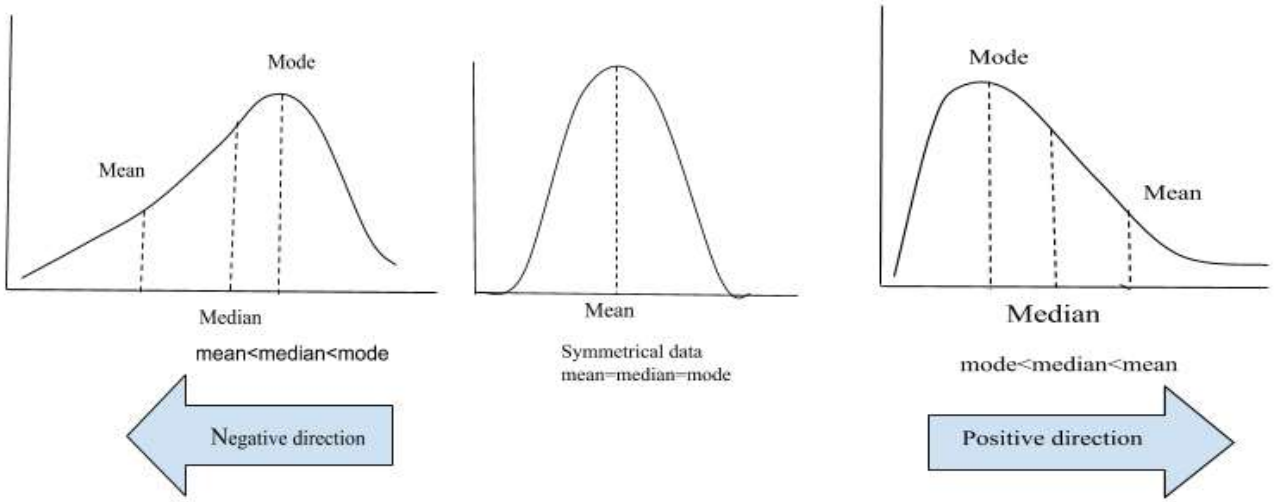
- **Types of Descriptive Statistics**
- Descriptive statistics break down into several types, characteristics, or measures. Some authors say that there are two types. Others say three or even four. In the spirit of working with averages, we will go with three types.
- Distribution, which deals with each value's frequency
- Central tendency, which covers the averages of the values
- Variability (or dispersion), which shows how spread out the values are

- **Distribution (also called Frequency Distribution)**
- Datasets consist of a distribution of scores or values. Statisticians use graphs and tables to summarize the frequency of every possible value of a variable, rendered in percentages or numbers. For instance, if you held a poll to determine people's favorite Beatle, you'd set up one column with all possible variables (John, Paul, George, and Ringo), and another with the number of votes.
- Statisticians depict frequency distributions as either a graph or as a table.
- **Measures of Central Tendency**
- Measures of central tendency estimate a dataset's average or center, finding the result using three methods: mean, mode, and median.
- **Mean**. The mean is also known as "M" and is the most common method for finding averages. You get the mean by adding all the response values together, dividing the sum by the number of responses, or "N." For instance, say someone is trying to figure out how many hours a day they sleep in a week. So, the data set would be the hour entries (e.g., 6,8,7,10,8,4,9), and the sum of those values is 52. There are seven responses, so N=7. You divide the value sum of 52 by N, or 7, to find M, which in this instance is 7.3.
- **Mode.** The mode is just the most frequent response value. Datasets may have any number of modes, including "zero." You can find the mode by arranging your dataset's order from the lowest to highest value and then looking for the most common response. So, in using our sleep study from the last part: 4,6,7,8,8,9,10. As you can see, the mode is eight.
- **Median.** Finally, we have the median, defined as the value in the precise center of the dataset. Arrange the values in ascending order (like we did for the mode) and look for the number in the set's middle. In this case, the median is eight.

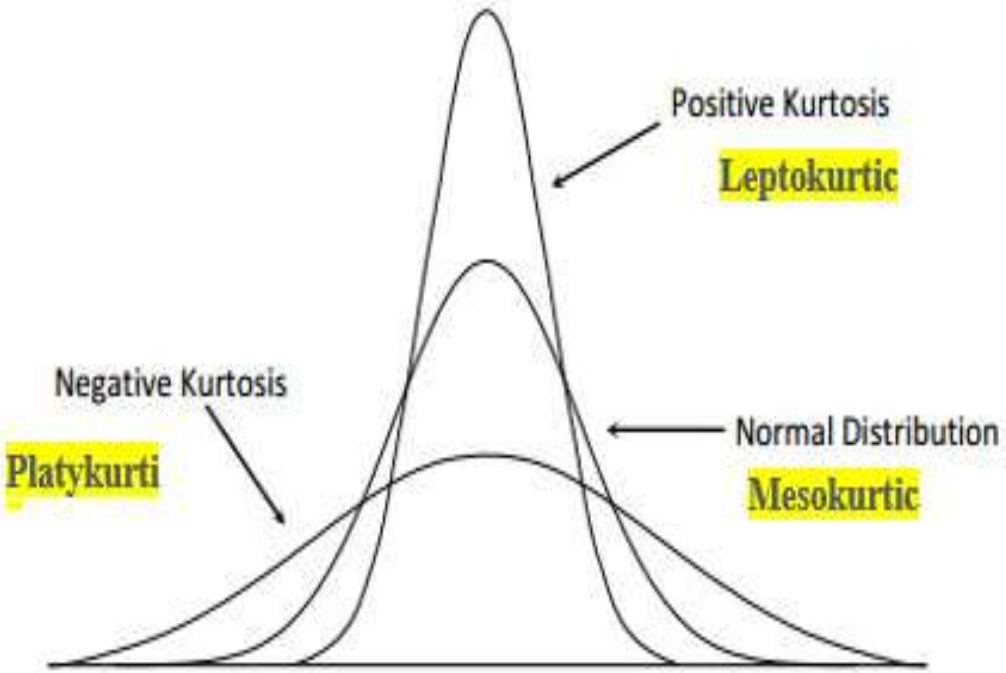
- **Variability (also called Dispersion)**
- The measure of variability gives the statistician an idea of how spread out the responses are. The spread has three aspects — range, standard deviation, and variance.
- **Range.** Use range to determine how far apart the most extreme values are. Start by subtracting the dataset's lowest value from its highest value. Once again, we turn to our sleep study: 4,6,7,8,8,9,10. We subtract four (the lowest) from ten (the highest) and get six. There's your range.
- **Standard Deviation**. This aspect takes a little more work. The standard deviation (s) is your dataset's average amount of variability, showing you how far each score lies from the mean. The larger your standard deviation, the greater your dataset's variable. Follow these six steps:
- List the scores and their means.
- Find the deviation by subtracting the mean from each score.
- Square each deviation.
- Total up all the squared deviations.
- Divide the sum of the squared deviations by N-1.
- Find the result's square root.

- **Skewness:**
- Skewness is the measure of the asymmetry of an ideally symmetric probability distribution and is given by the third standardized moment. If that sounds way too complex, don’t worry! Let me break it down for you.
  the normal distribution is the probability distribution without any skewness.
- there are two types of skewness:
- Positive Skewness
- Negative Skewness


Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew

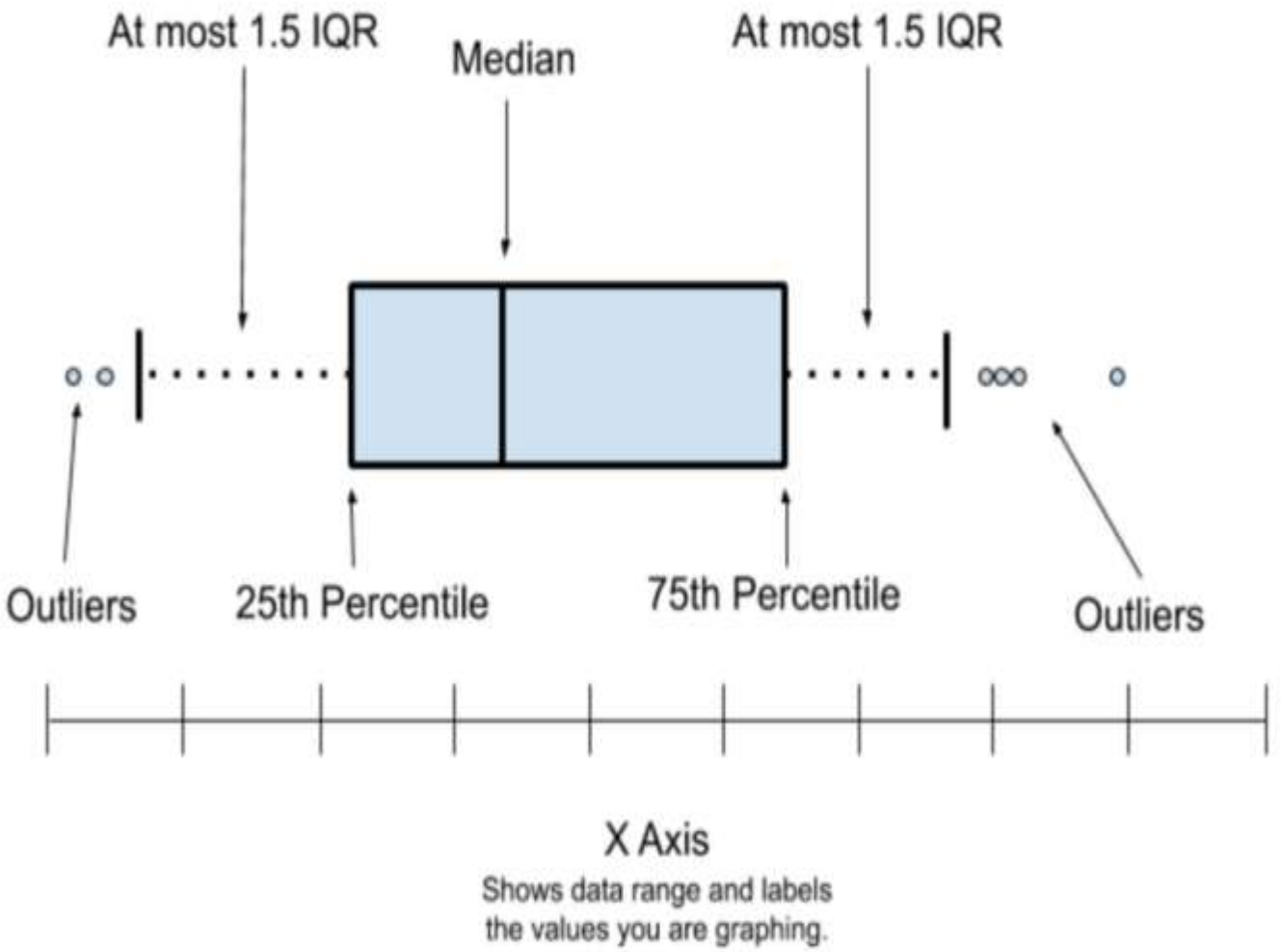

Mode
Mean
Median
mean<median<mode
Negative direction
Mean
Symmetrical data
mean=median=mode
Mode
Mean
Median
mode<median<mean
Positive direction

- **Kurtosis:**
- Kurtosis measures the "heaviness of the tails" of a distribution (in compared to a normal distribution). Kurtosis is positive if the tails are "heavier" then for a normal distribution, and negative if the tails are "lighter" than for a normal distribution. The normal distribution has kurtosis of zero.
- The kurtosis of a distribution or sample is equal to the 4th central moment divided by the 4th power of the standard deviation, minus 3.
- To calculate the kurtosis of a sample:
- i) subtract the mean from each value to get a set of deviations from the mean;
- ii) divide each deviation by the standard deviation of all the deviations;
- iii) average the 4th power of the deviations and subtract 3 from the result.


Positive Kurtosis
Leptokurtic
Negative Kurtosis
Platykurti
Normal Distribution
Mesokurtic

- **Types of excess kurtosis**
- Leptokurtic or heavy-tailed distribution (kurtosis more than normal distribution).
- Mesokurtic (kurtosis same as the normal distribution).
- Platykurtic or short-tailed distribution (kurtosis less than normal distribution).
- **Leptokurtic (kurtosis > 3)**
- **platykurtic (kurtosis < 3)**
- **Mesokurtic (kurtosis = 3)**

- **Box plot:**
- A box plot also known as Five Number Summary, summarizes data using the median, upper quartile, lower quartile, and the minimum and maximum values. It allows you to see important characteristics of the data at a glance(visually). This also help us to visualize outliers in the data set.
- **Box plot or Five Number Summary has below five information.**
- 1.Median
- 2. Lower Quartile(25th Percentile)
- 3.Upper Quartile(75th Percentile)
- 4. Minimum Value
- 5.Maximum Value


At most 1.5 IQR
Median
At most 1.5 IQR
Outliers
25th Percentile
75th Percentile
Outliers
X Axis
Shows data range and labels
the values you are graphing.

Pivot table

- A pivot table is a **summary tool** that wraps up or summarizes information sourced from bigger tables. These bigger tables could be a database, an Excel spreadsheet, or any data that is or could be converted in a table-like form. The data summarized in a pivot table might include sums, averages, or other statistics which the pivot table groups together in a meaningful way.
- **Examples of pivot table:**
- Pivoting in Python with Pandas
- Pivoting in Google Sheets
- Grouping in PostgreSQL
- Aggregation Functions
- Missing Data

Heat map

- Heatmaps visualize the data in a 2-dimensional format in the form of colored maps. The color maps use hue, saturation, or luminance to achieve color variation to display various details. This color variation gives visual cues to the readers about the magnitude of numeric values.
- **Uses of HeatMap**
- Business Analytics
- Website
- Exploratory Data Analysis
  Molecular Biology
- Geovisualization
- Marketing and Sales

- **Types of HeatMaps**
- Typically, there are two types of Heatmaps:

1. Grid Heatmap: The magnitudes of values shown through colors are laid out into a matrix of rows and columns, mostly by a density-based function. Below are the types of Grid Heatmaps

- ***Clustered Heatmap:*** The goal of Clustered Heatmap is to build associations between both the data points and their features. This type of heatmap implements clustering as part of the process of grouping similar features.
- The order of the rows in Clustered Heatmap is determined by performing hierarchical cluster analysis of the rows. Clustering positions similar rows together on the map. Similarly, the order of the columns is determined.
- **Correlogram:** A correlogram replaces each of the variables on the two axes with numeric variables in the dataset. Each square depicts the relationship between the two intersecting variables, which helps to build descriptive or predictive statistical models.

2. Spatial Heatmap: Each square in a Heatmap is assigned a color representation according to the nearby cells' value. The location of color is according to the magnitude of the value in that particular space. These Heatmaps are data-driven "paint by numbers" canvas overlaid on top of an image.

Correlation statistics

- **Correlation**
- Correlation measures the relationship between two variables.
- We mentioned that a function has a purpose to predict a value, by converting input (x) to output (f(x)). We can say also say that a function uses the relationship between two variables for prediction.
- **Correlation Coefficient**
- The correlation coefficient measures the relationship between two variables.
- The correlation coefficient can never be less than -1 or higher than 1.
- 1 = there is a perfect linear relationship between the variables (like Average_Pulse against Calorie_Burnage)
- 0 = there is no linear relationship between the variables
- -1 = there is a perfect negative linear relationship between the variables (e.g. Less hours worked, leads to higher calorie burnage during a training session)

ANOVA

- **ANOVA** stands for analysis of variance and, as the name suggests, it helps us understand and compare variances among groups. Before going in detail about ANOVA, let's remember a few terms in statistics:
- **Mean**: The average of all values.
- **Variance**: A measure of the variation among values. It is calculated by adding up squared differences of each value and the mean and then dividing the sum by the number of samples.

- **Standard deviation**: The square root of variance.
- In order to understand the motivation behind ANOVA, or some other statistical tests, we should learn two simple terms: population and sample.
- **Population** is all elements in a group.
- **Sample** is a subset of a population.

$$Variance = \frac{\sum(x_i - \text{mean})^2}{N}$$

X_i = value i

N = number of values

Mean = average of all values

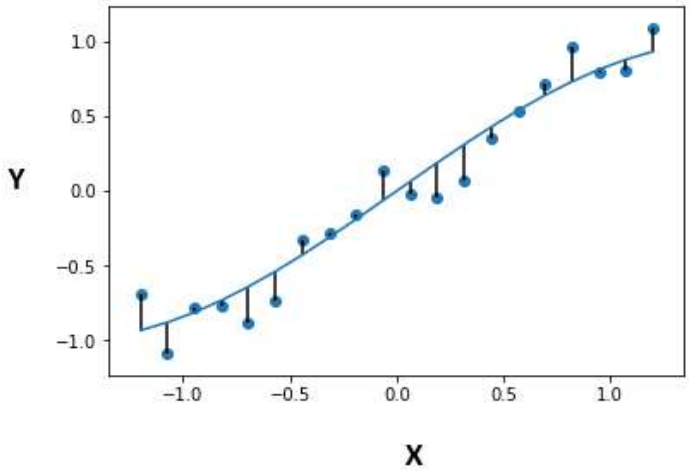
UNIT – 4
MODEL DEVELOPMENT

- **Regression Analysis**
- **Regression analysis** is a predictive modelling technique that assesses the relationship between dependent (i.e., the goal/target variable) and independent factors. Forecasting, time series modelling, determining the relationship between variables, and predicting continuous values can all be done using regression analysis

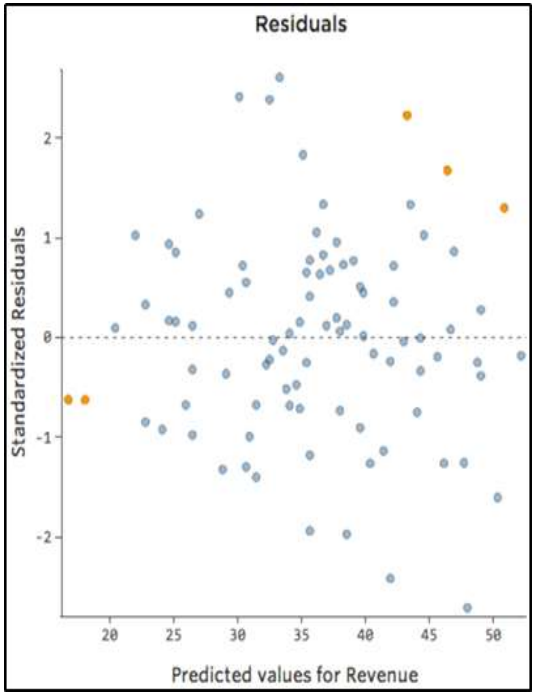
- **Simple Linear Regression:** The association between two variables is established using a straight line in Simple Linear Regression. It tries to create a line that is as near to the data as possible by determining the slope and intercept, which define the line and reduce regression errors. There is a single x and y variable
- Equation: Y = mX+c
- **Multiple Linear Regression:** Multiple linear regressions are based on the presumption that both the dependent and independent variables, or Predictor and Target variables, have a linear relationship. There are two types of multilinear regressions: linear and nonlinear. It has one or more x variables and one or more y variables, or one dependent variable and two or more independent variables
- Equation: Y = m1X1+m2X2+m3X3+..c
- Where,
- Y = Dependent Variable
  m = Slope
  X = Independent Variable
  c = Intercept

Model evaluation using visualization

- **Residual Plot:**
- **Residuals**
- A residual is a measure of how far away a point is vertically from the regression line. Simply, it is the error between a predicted value and the observed actual value. $Residual\ (\epsilon) = y - \hat{y}$



- **Residual Plots**
- A typical residual plot has the residual values on the Y-axis and the independent variable on the x-axis. The below fig
- is a good example of how a typical residual plot looks like.



Distribution plots

- **Polynomial Regression**
- In polynomial regression, the relationship between the independent variable x and the dependent variable y is described as an nth degree polynomial in x. Polynomial regression, abbreviated E(y |x), describes the fitting of a nonlinear relationship between the value of x and the conditional mean of y. It usually corresponded to the least-squares method.
-
- **Types of Polynomial Regression**
- A quadratic equation is a general term for a second-degree polynomial equation. This degree, on the other hand, can go up to nth values. Polynomial regression can so be categorized as follows:
- 1. Linear – if degree as 1
- 2. Quadratic – if degree as 2
- 3. Cubic – if degree as 3 and goes on, on the basis of degree.

Data science pipeline

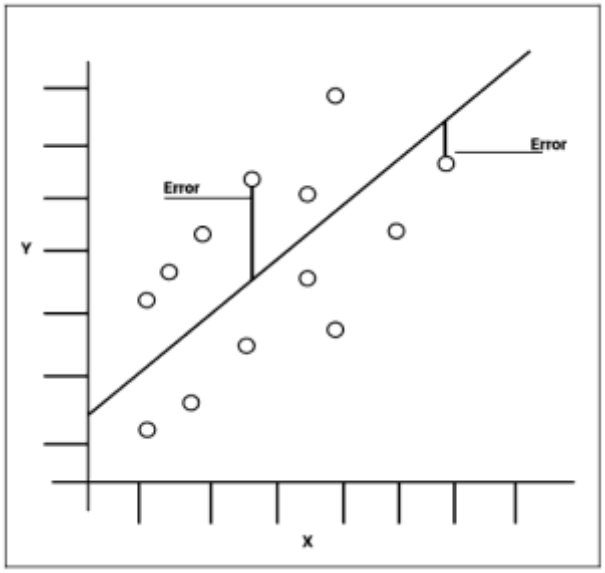
- A Data Science Pipeline is a collection of processes that transform raw data into actionable business answers. **Data Science Pipelines** automate the flow of data from source to destination, providing you with insights to help you make business decisions.
- **Key Features of Data Science Pipelines**
- Here is a list of key features of the Data Science Pipeline:
- Continuous and Scalable Data Processing
- Cloud-based Elasticity and Agility.
- Data Processing Resources that are Self-Contained and Isolated.
- Access to a Large Amount of Data and the ability to self-serve.
- Disaster Recovery and High Availability
- Allow users to Delve into Insights at a Finer Level.
- Removes Data silos and Bottlenecks that cause Delays and Waste of Resources.

- **Working of Data Science Pipeline:**
- It is critical to have specific questions you want data to answer before moving raw data through the pipeline. This allows users to focus on the right data in order to uncover the right insights.
- The Data Science Pipeline is divided into several stages, which are as follows:
- Obtaining Information
- Data Cleansing
- Data Exploration and Modeling
- Data Interpretation
- Data Revision

MEASURES FOR IN – SAMPLE EVALUATION

- A way to numerically determine how good the model fits the dataset.
- Two important measures to determine the fit of a model:
- Mean squared error(MSE)
- R squared ($R^{\wedge 2}$)
- **Mean squared error(MSE)**
-
- The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss.

- Lesser the MSE => Smaller is the error => Better the estimator.
- The Mean Squared Error is calculated as:
- MSE = (1/n) * Σ(actual – forecast)2
- where:
- Σ – a symbol that means “sum”
- n – sample size
- actual – the actual data value
- forecast – the predicted data value

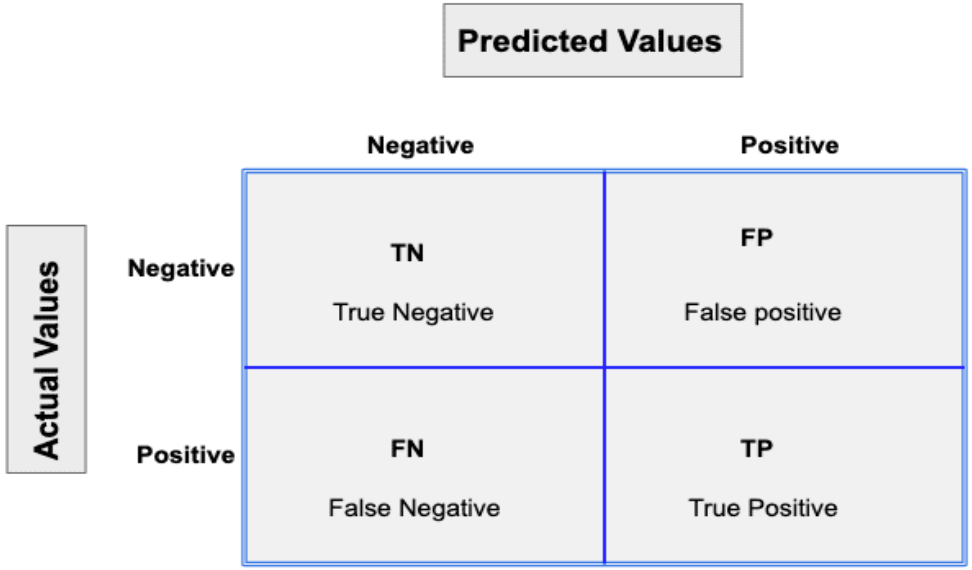


- **R squared ($R^{\wedge 2}$)**
- R-squared is a metric of correlation. Correlation is measured by **“r”** and it tells us how strongly two variables can be related. A correlation closer to +1 means a strong relationship in the positive direction, while -1 means a stronger relationship in the opposite direction. A value closer to 0 means that there is not much of a relationship between the variables.

UNIT V
MODEL EVALUATION

- **EVALUATION METRICS**
- Model Evaluation Metrics define the evaluation metrics for evaluating the performance of a machine learning model, which is an integral component of any data science project. It aims to estimate the generalization accuracy of a model on the future (unseen/out-of-sample) data.
- **Confusion Matrix**
- A confusion matrix is a matrix representation of the prediction results of any binary testing that is often used to **describe the performance of the classification model** (or "classifier") on a set of test data for which the true values are known.

- Each prediction can be one of the four outcomes, based on how it matches up to the actual value:
- True Positive (TP): Predicted True and True in reality.
- True Negative (TN): Predicted False and False in reality.
- False Positive (FP): Predicted True and False in reality.
- False Negative (FN): Predicted False and True in reality.

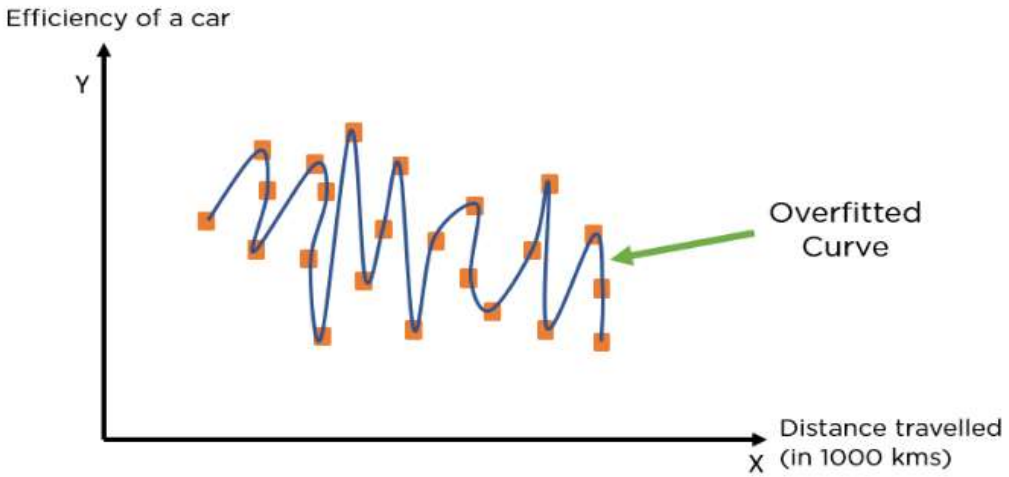


- **CROSS VALIDATION**

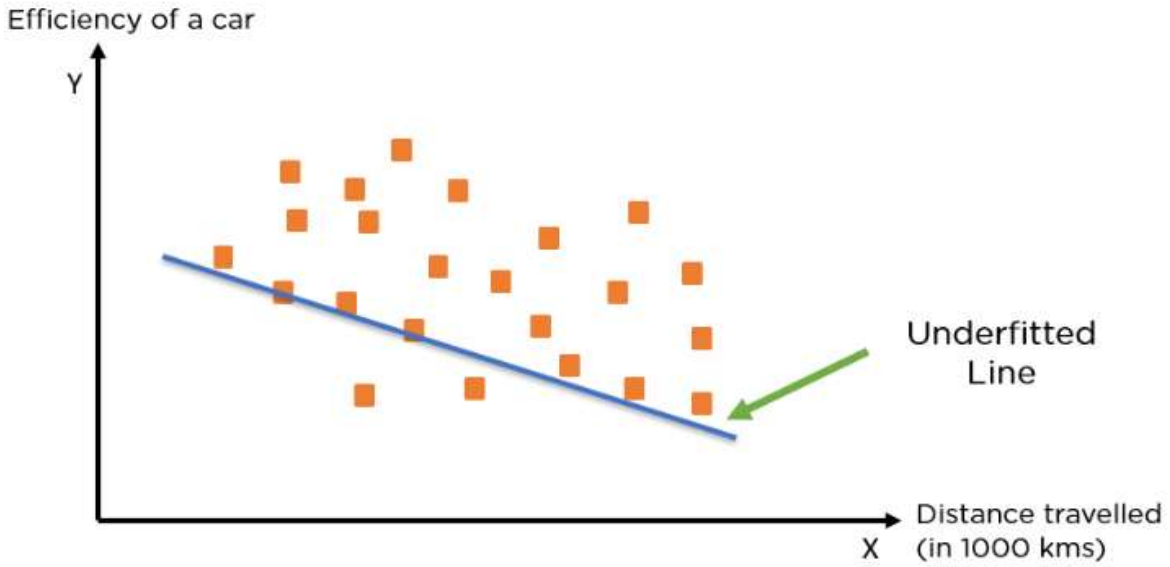
- Cross validation is a technique for assessing how the statistical analysis generalises to an independent data set.It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data. Using cross-validation, there are high chances that we can detect over-fitting with ease.

Overfitting and Underfitting

- Overfitting:
- When a model performs very well for training data but has poor performance with test data (new data), it is known as overfitting. In this case, the machine learning model learns the details and noise in the training data such that it negatively affects the performance of the model on test data. Overfitting can happen due to low bias and high variance.



- Underfitting:
- When a model has not learned the patterns in the training data well and is unable to generalize well on the new data, it is known as underfitting. An underfit model has poor performance on the training data and will result in unreliable predictions. Underfitting occurs due to high bias and low variance.



- **Ridge Regression**
- Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.
- The cost function for ridge regression:
- ***Min(||Y – X(theta)||^2 + λ||theta||^2)***
- Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function.

GRID SEARCH

- Grid-search is used to find the optimal hyperparameters of a model which results in the most ‘accurate’ predictions.
- Grid search refers to a technique used to identify the optimal hyperparameters for a model. Unlike parameters, finding hyperparameters in training data is unattainable. As such, to find the right hyperparameters, we create a model for each combination of hyperparameters.

- **Cross validation**
- We have mentioned that cross-validation is used to evaluate the performance of the models. Cross-validation measures how a model generalizes itself to an independent dataset. We use cross-validation to get a good estimate of how well a predictive model performs.
- With this method, we have a pair of datasets: an independent dataset and a training dataset. We can partition a single dataset to yield the two sets. These partitions are of the same size and are referred to as folds. A model in consideration is trained on all folds, bar one.

K-fold cross-validation with K as 5

Grid search implementation

- Load dataset.
- Import GridSearchCV, svm and SVR
- Set estimator parameters.
- Specify hyperparameters and range of values.
- Evaluation.
- Fitting the data.