



**CHRIST**  
(DEEMED TO BE UNIVERSITY)  
BANGALORE · INDIA

# Advanced Data Analytics

## Unit-1 Introduction

### MISSION

CHRIST is a nurturing ground for an individual's holistic development to make effective contribution to the society in a dynamic environment

### VISION

Excellence and Service

### CORE VALUES

Faith in God | Moral Uprightness  
Love of Fellow Beings  
Social Responsibility | Pursuit of Excellence

## Overview of data analytics

- **Traditional Data Analysis** focuses on summarizing and interpreting existing data using basic statistical methods and tools, making it suitable for straightforward business insights and reporting.
- **Advanced Data Analysis** involves sophisticated techniques, including machine learning and big data technologies, to derive deeper and more actionable insights, predictive models, and strategic recommendations from complex datasets.

# Traditional Data Analysis Vs Advanced Data Analysis

	Traditional Data Analysis	Advanced Data Analysis
<b>Techniques Used</b>	Descriptive Statistics, Basic Inferential Statistics, Data Visualization	Ad. Stat Methods, AI&ML, Data Mining, Predictive and Prescriptive Analytics
<b>Tools</b>	Spreadsheets, Basic Statistical Software	Advanced Programming Languages, Big Data Technologies, Advanced Statistical Software, Data Visualization Tools
<b>Data Size and Complexity</b>	smaller datasets, Focus on structured data.	large-scale, complex((text, images, etc.), process and analyze massive amounts of data.
<b>Objective</b>	summary and straightforward interpretation, generating reports, dashboards, and basic insights	uncover deeper insights, trends, and patterns, predictive models for decision-making and strategy, get leveraging data-driven insights
<b>Required Skill Level</b>	basic statistical and data handling	knowledge of advanced statistics, mathematics, and programming. data scientists, machine learning engineers, and advanced analysts with specialized training.

## Example of Traditional Data Analysis

- **Scenario:** A retail company wants to understand its sales performance over the last year.
- **Approach:**
  - **Data Collection,** Descriptive Statistics, Data Visualization, Basic Inferential Statistics
- **Outcome:**
  - Summary of sales performance, highlighting months with peak sales and identifying low-performing months
  - Basic insights into which product categories are performing well and which need attention
  - Simple correlations between sales and advertising spend.

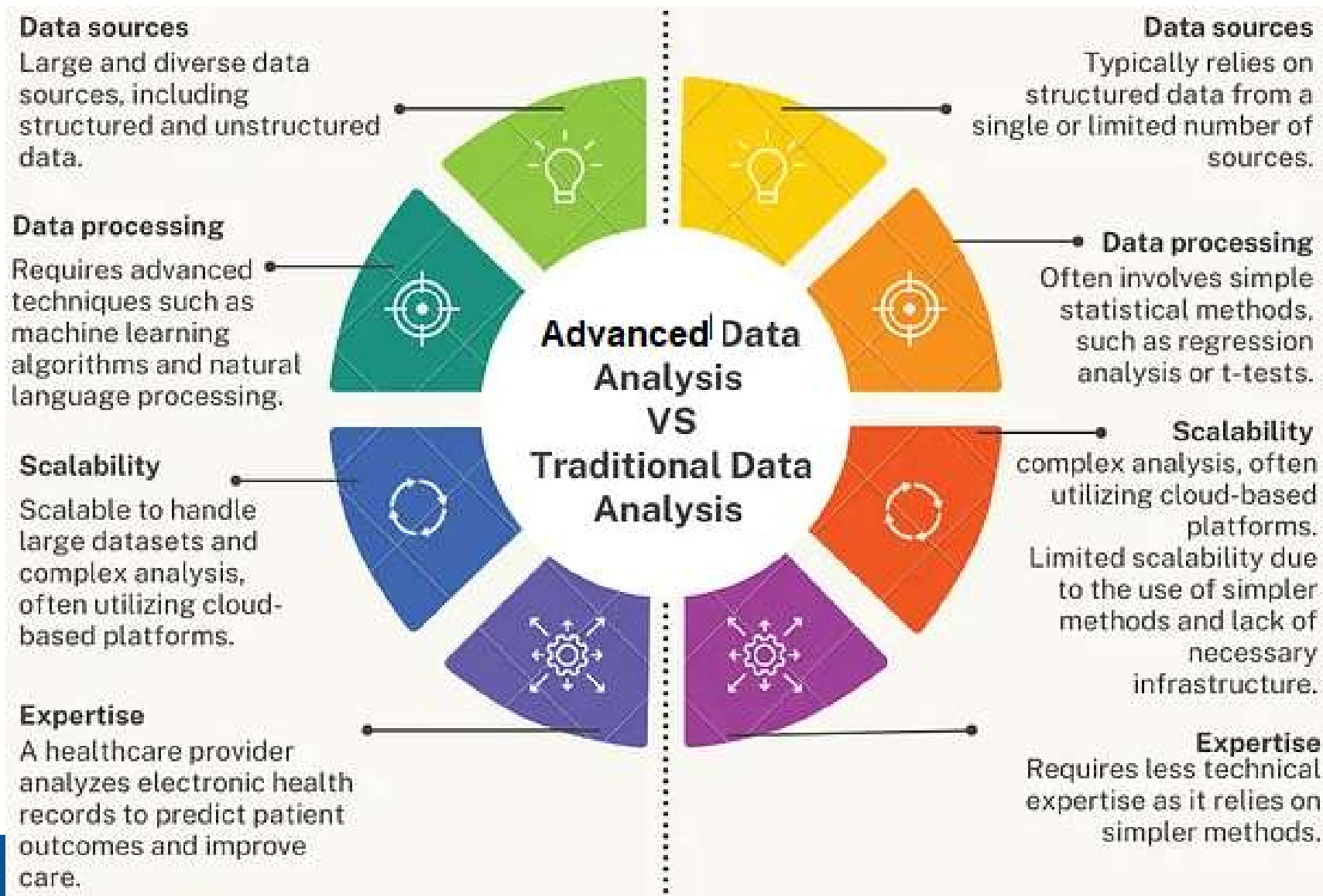
## Example of Advanced Data Analysis

- **Scenario:** The same retail company wants to predict future sales and optimize inventory management.
- **Approach**
  - Data Collection, Data Pre-processing, Advanced Techniques, Big Data Technologies, Data Visualization, Predictive and Prescriptive Analytics
- **Outcome:**
  - Accurate sales forecasts helping in better inventory management and strategic planning.
  - Identification of customer segments and personalized marketing strategies to boost sales.
  - Insights from customer reviews leading to product improvements and better customer satisfaction.
  - Optimized operations and reduced costs through data-driven decision-making.

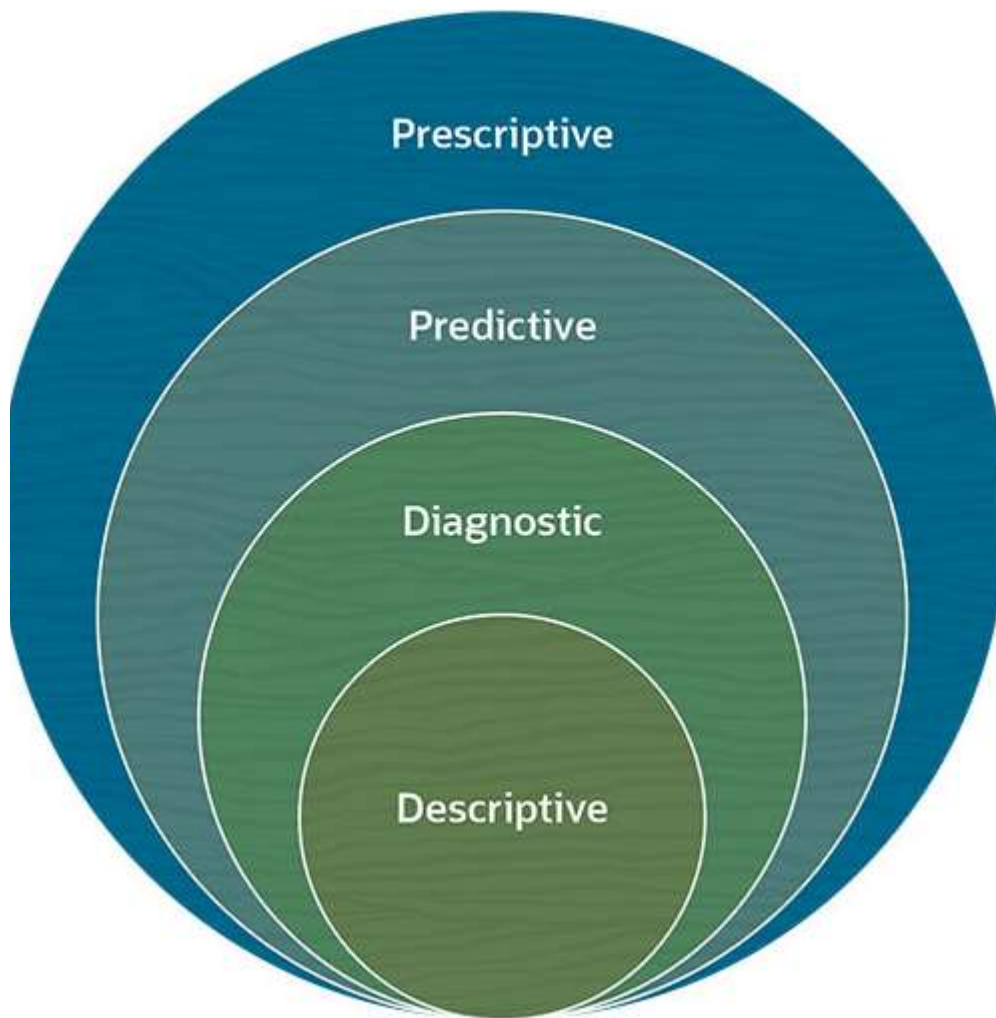
## Example of Advanced Data Analysis

Type of Multimedia Needs Assessment	How Advanced Data Analysis Can Be Used	Real-World Examples
Audio Analysis	Acoustic analysis can be used to identify specific sounds or speech patterns	A speech recognition tool might use acoustic analysis to improve accuracy and identify specific speakers
Video Analysis	Facial recognition can be used to analyze facial expressions and emotions	A video production company might use facial recognition to analyze audience reactions to different actors or characters
Image Analysis	Color analysis can be used to understand the emotional impact of different color schemes	A social media platform might use color analysis to understand which types of images are most engaging to users

# Need for Advanced Data Analytics

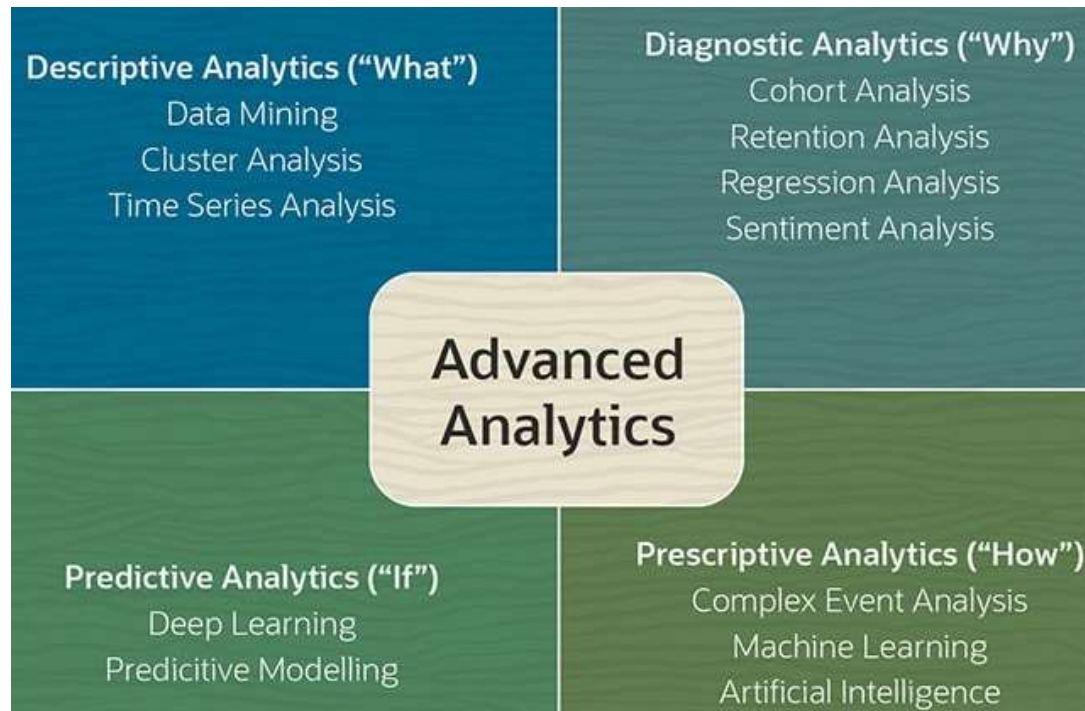


# Advanced data analytics techniques Example



- **Descriptive analytics(What):**
  - A retail company analyzing its sales data to understand its sales performance over the last year.
- **Diagnostic analytics(Why):**
  - A retail company investigating the reasons behind a sudden drop in sale in February.
- **Predictive analytics(If):**
  - A retail company forecasting sales for the next quarter to optimize inventory management .
- **Prescriptive analytics(How):**
  - A retail company optimizing its inventory levels based on sales forecasts. to ensure it has the right amount of inventory to meet predicted sales while minimizing costs.





- **Descriptive analytics(What):**
  - focuses on the aggregation of data
  - Data and text mining, cluster analysis and summary statistics
- **Diagnostic analytics(Why):**
  - answers why something happened
  - regression analysis, sensitivity analysis and PCA
- **Predictive analytics(If):**
  - estimate what could happen if certain conditions
  - predictive modeling and deep-learning techniques.
- **Prescriptive analytics(How):**
  - focuses on how to achieve a particular outcome.
  - simulation analysis, AI, ML, NN

# Introduction to advanced data analytics techniques

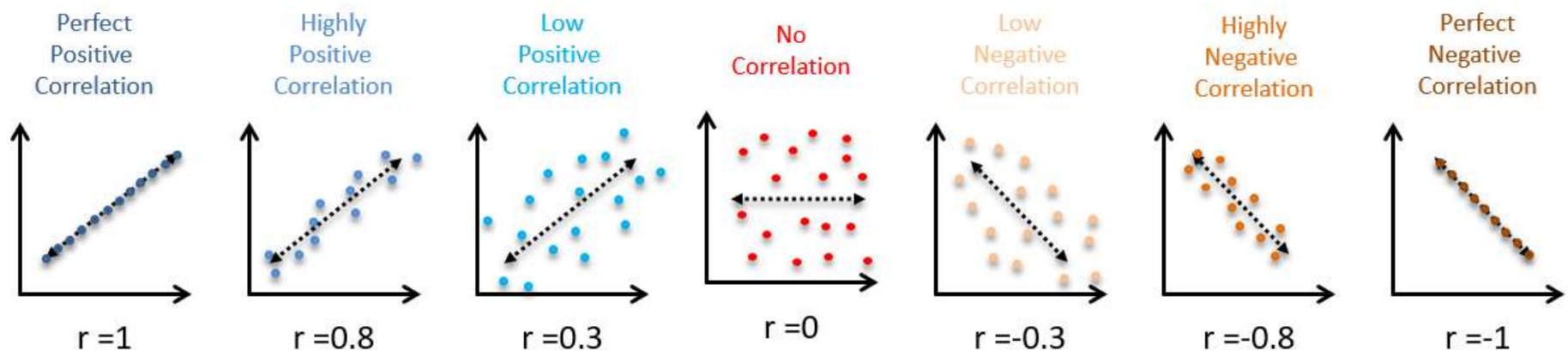
Advanced Data Analysis Technique	Description	Example
<b>Qualitative Data Analysis</b>	Examining non-numerical data such as text, images, and audio to gain insights into multimedia needs and preferences	Conducting interviews with focus groups to understand the attitudes and beliefs of a target audience towards different types of multimedia content
<b>Quantitative Data Analysis</b>	Analyzing numerical data to make statistical inferences about multimedia needs and preferences	Conducting a survey to gather data on the frequency of media consumption and preferences among a target audience
<b>Content Analysis</b>	Systematically examining media content to identify themes and patterns that can inform multimedia design	Analyzing social media posts to understand the types of content that are most popular among a target audience
<b>Social Network Analysis</b>	Analyzing the relationships between individuals and groups in a social network to identify key influencers and communication patterns	Mapping out the social connections between members of a target audience to identify individuals who are likely to share multimedia content with their networks

# Regression & Co-Relation

# Correlation

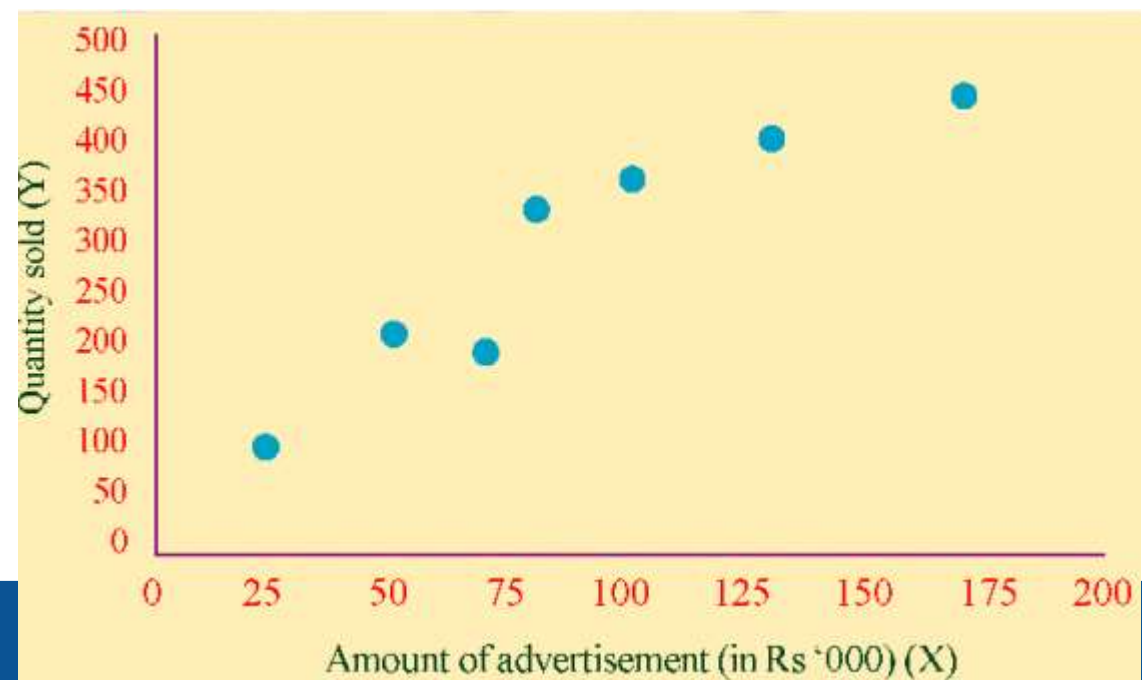
- Degree of relationship between two variables in a bivariate distribution
- Example
  - Price of product and Demand
  - Height and Weight

## Scatter Plots & Correlation Examples



**Case (i):** Consider a certain brand of television. The amount utilised for advertisements and quantity sold in different years are given below.

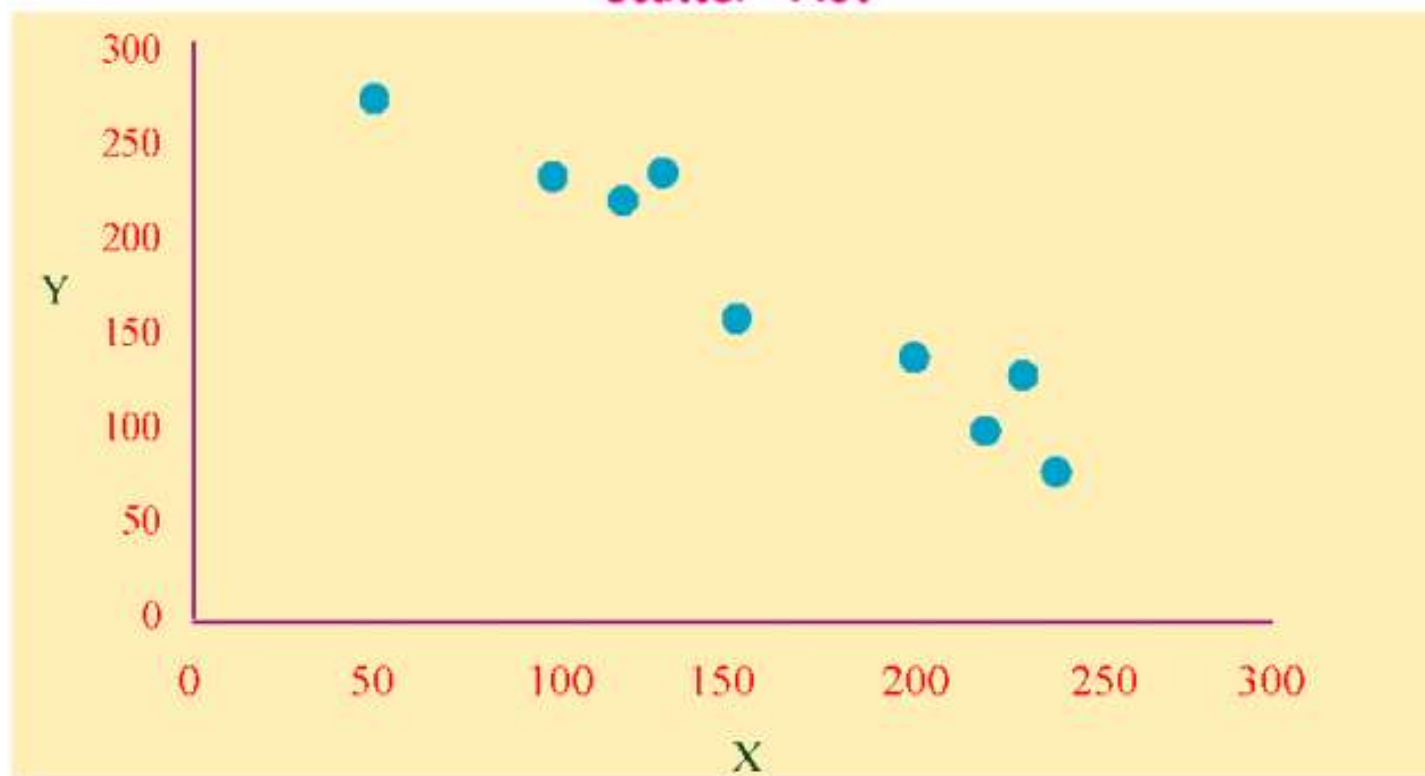
Amount of advertisement (in Rs '000) (X)	25	50	70	80	100	130	170
Quantity sold (Y)	100	220	200	340	370	410	450



**Case (ii):** Consider the case of CFL lamps. The price and quantity sold in different months are given below.

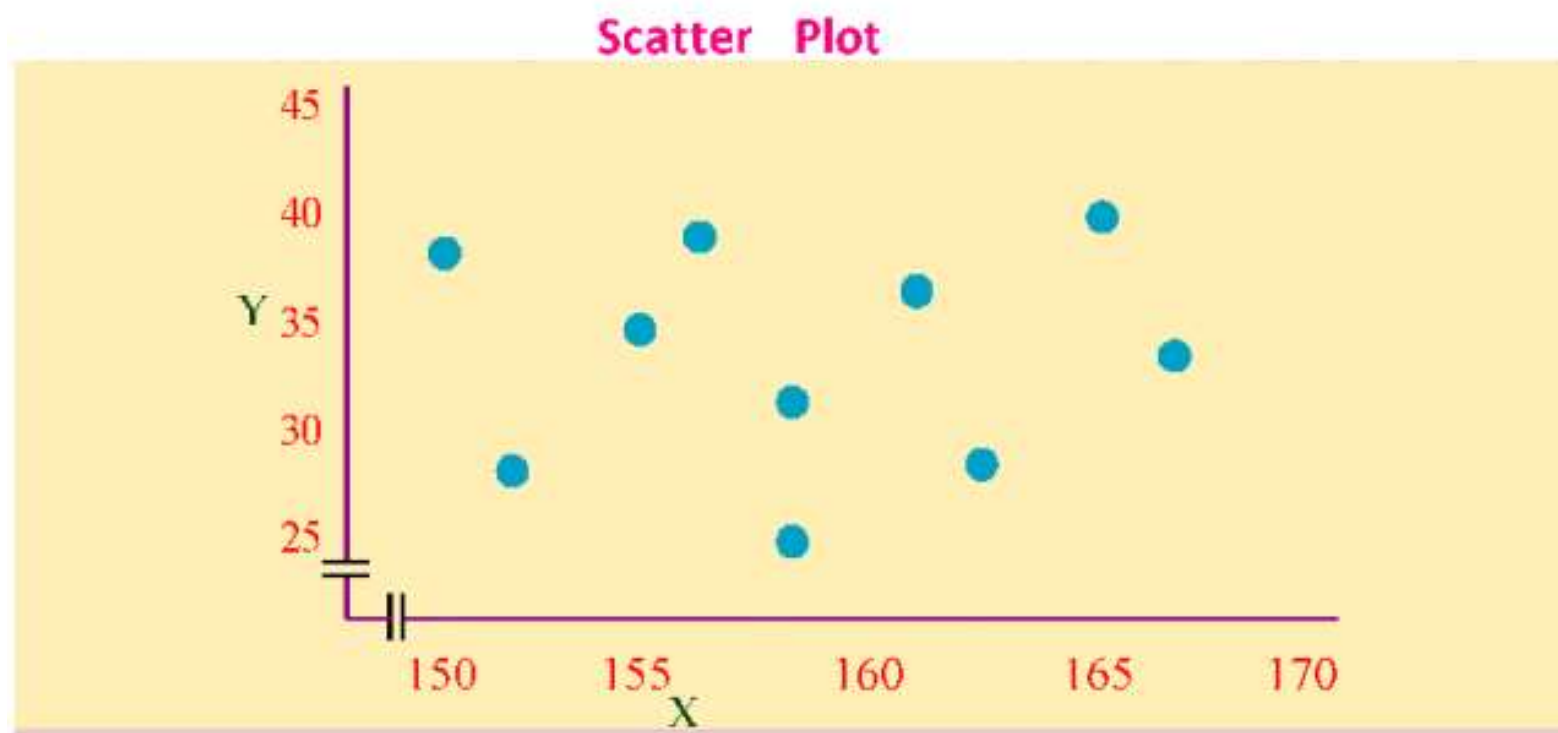
Price (X in Rs.)	50	100	120	130	150	200	220	230	240
Quantity sold (Y)	275	234	220	235	160	140	100	130	80

Scatter Plot



**Case (iii):** The height in cms. and marks in English out of 50 of 10 students are given as follows.

Height in cms. (X)	150	165	155	156	158	163	158	162	152	167
Marks in English out of 50 (Y)	38	40	35	39	25	27	32	37	28	34





# Correlation Vs Regression

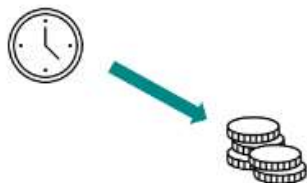
Basis	Correlation	Regression
Meaning	A statistical measure that defines co-relationship or association of two variables.	Describes how an independent variable is associated with the dependent variable.
Dependent and Independent variables	No difference	Both variables are different.
Usage	To describe a linear relationship between two variables.	To fit the best line and estimate one variable based on another variable.
Objective	To find a value expressing the relationship between variables.	To estimate values of a random variable based on the values of a fixed variable.



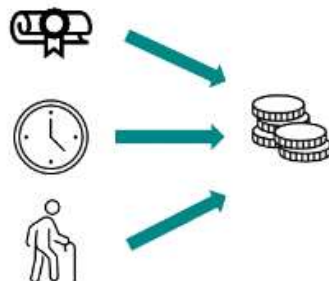
# Linear Regression

- Regression Analysis is a mathematical measure of the nature of relation between 2 or more variables.
- A regression analysis makes it possible to infer or predict another variable based on one or more variables
- The variable to be inferred is called the **dependent variable** (criterion). The variables used for prediction are called **independent variables** (predictors)
- Examples of a regression
  - Simple linear regression :Does the **weekly working time** have an influence on the **hourly wage** of employees?
  - Multiple lineare regression: Do the **weekly working time** and the **age of employees** have an influence on their **hourly wage**?
  - Logistic regression : Do the **weekly working time** and the **age** of employees have an influence on the probability that they are **at risk of burnout**?

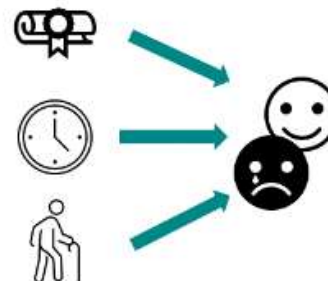
Simple Linear Regression



Multiple Linear Regression

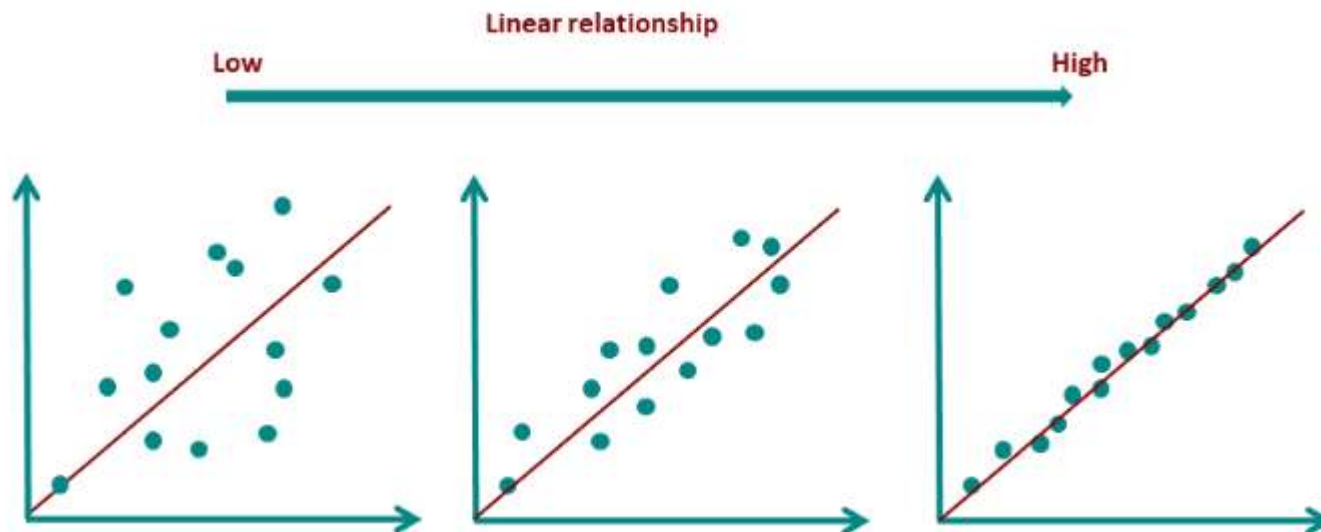


Logistic Regression



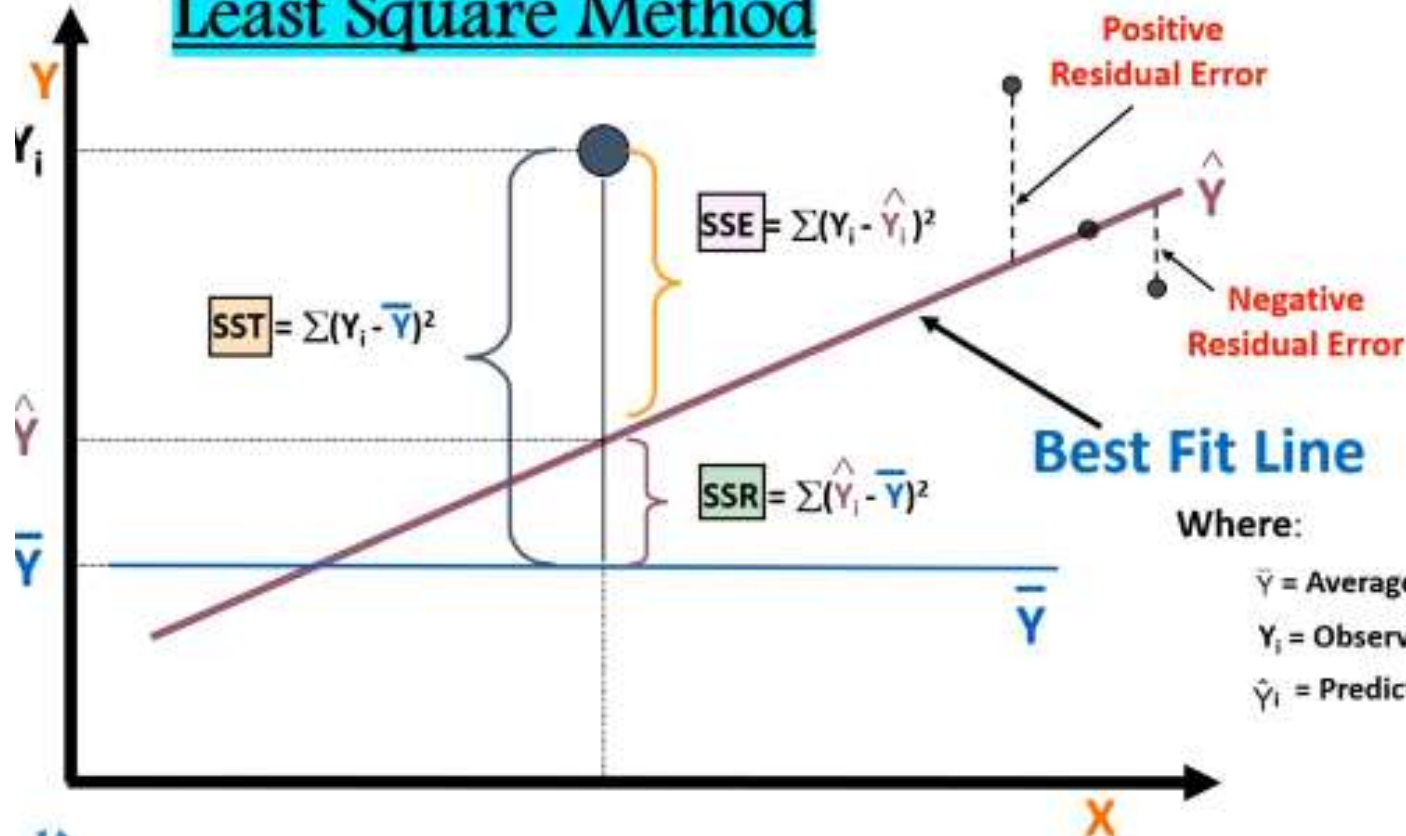
# Simple Linear Regression

- The goal of a **simple linear regression** is to predict the value of a dependent variable based on an independent variable.
- Best Fit line - describes the linear relationship between the dependent and independent variable



# Simple Linear Regression

## Least Square Method



$$SST = \sum (Y_i - \bar{Y})^2$$

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2$$

$$SSE = \sum (Y_i - \hat{Y}_i)^2$$

Where:

$\bar{Y}$  = Average value of the dependent variable

$Y_i$  = Observed values of the dependent variable

$\hat{Y}_i$  = Predicted value of Y for the given  $X_i$  value

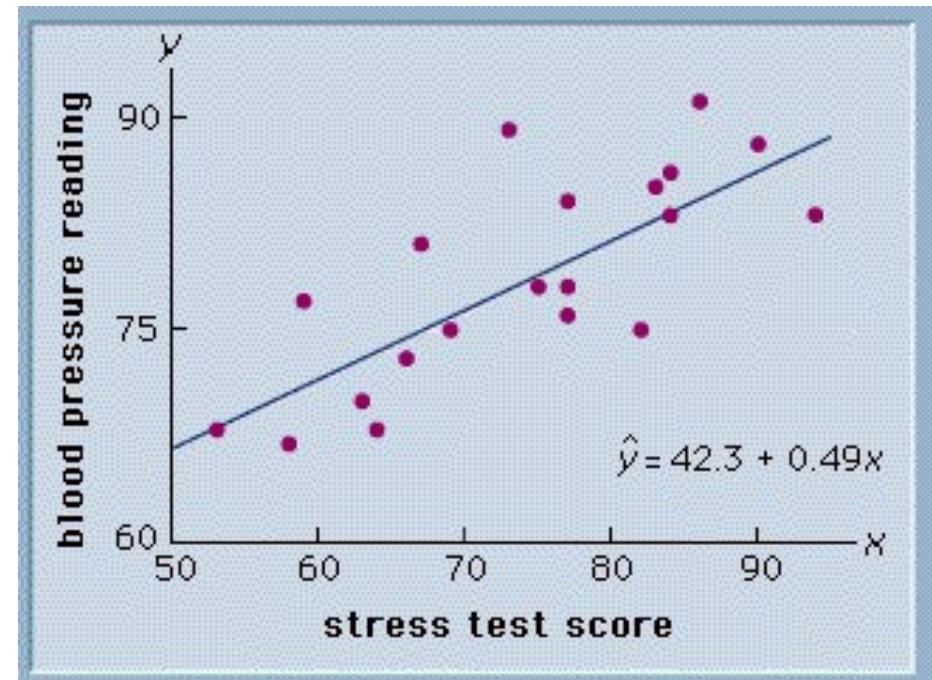
- Why do we need SST, SSR, and SSE? We can use them to calculate the R-squared, conduct F-tests in regression analysis, and combine them with other **goodness-of-fit** measures to evaluate regression models.

# Simple Linear Regression

- $a$  : point of intersection with the y-axis
- $b$  : gradient of the straight line

$$\hat{y} = b \cdot x + a$$

$\hat{y}$ : Estimated dependent variable  
 $b$ : Slope  
 $x$ : Independent variable  
 $a$ : y intercept



The regression equation of  $Y$  on  $X$ ,

$$\hat{Y} - \bar{y} = b_{YX} (x - \bar{x})$$

$$b_{YX} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}$$

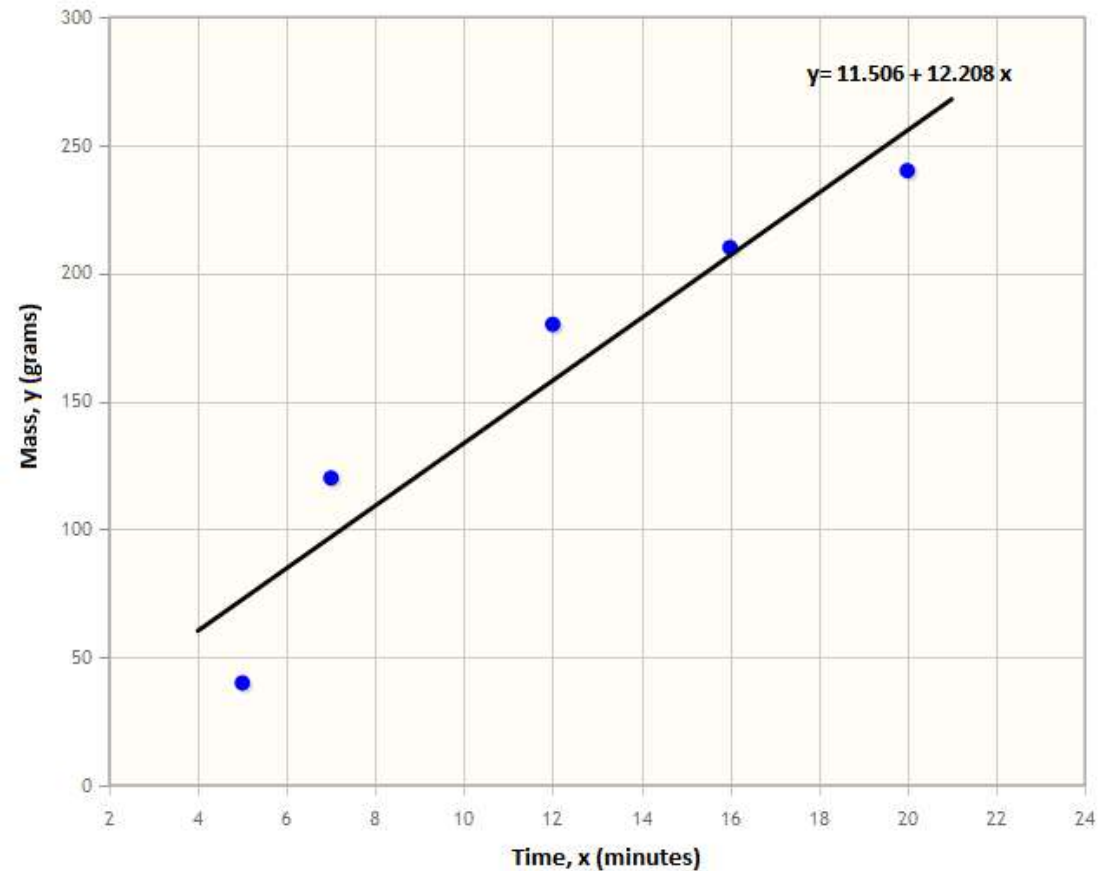
Regression equation of  $X$  on  $Y$ ,

$$\hat{X} - \bar{x} = b_{XY} (y - \bar{y})$$

$$b_{XY} = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n y_i^2 - \left( \sum_{i=1}^n y_i \right)^2}$$

- Consider the example below where the mass,  $y$  (grams), of a chemical is related to the time,  $x$  (seconds), for which the chemical reaction has been taking place according to the table

Time, $x$ (seconds)	5	7	12	16	20
Mass, $y$ (grams)	40	120	180	210	240



# Multiple Linear Regression

- **Definition:** Multiple Linear Regression is a statistical technique used to model the relationship between one dependent variable and two or more independent variables. The goal is to predict the value of the dependent variable based on the values of the independent variables.
- Predicting house prices based on features like size, number of bedrooms, and age of the house.

Mathematical Representation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

where:

- $y$  is the dependent variable.
- $x_1, x_2, \dots, x_n$  are the independent variable
- $\beta_0$  is the intercept.
- $\beta_1, \beta_2, \dots, \beta_n$  are the coefficients.
- $\epsilon$  is the error term.



## Example

Suppose we have a dataset containing information about houses, and we want to predict the house price based on the size (square footage), number of bedrooms, and age of the house.

Dataset:

Size (sqft)	Bedrooms	Age (years)	Price (\$)
2100	3	20	500000
1600	2	15	400000
2400	4	25	600000
1400	2	10	350000
3000	5	30	700000

Model:

$$\text{Price} = \beta_0 + \beta_1 \times \text{Size} + \beta_2 \times \text{Bedrooms} + \beta_3 \times \text{Age} + \epsilon$$

By fitting this model, we estimate the coefficients  $(\beta_0, \beta_1, \beta_2, \beta_3)$  and use them to predict house prices based on size, number of bedrooms, and age.

# Multivariate Linear Regression

- **Definition:** Multivariate Linear Regression is a statistical technique used to model the relationship between two or more dependent variables and two or more independent variables. It aims to predict multiple dependent variables simultaneously based on the values of the independent variables.
- Predicting both house prices and rental prices based on features like size, number of bedrooms, and location.

Mathematical Representation:

$$\begin{cases} y_1 = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \dots + \beta_{1n}x_n + \epsilon_1 \\ y_2 = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \dots + \beta_{2n}x_n + \epsilon_2 \\ \vdots \\ y_m = \beta_{m0} + \beta_{m1}x_1 + \beta_{m2}x_2 + \dots + \beta_{mn}x_n + \epsilon_m \end{cases}$$

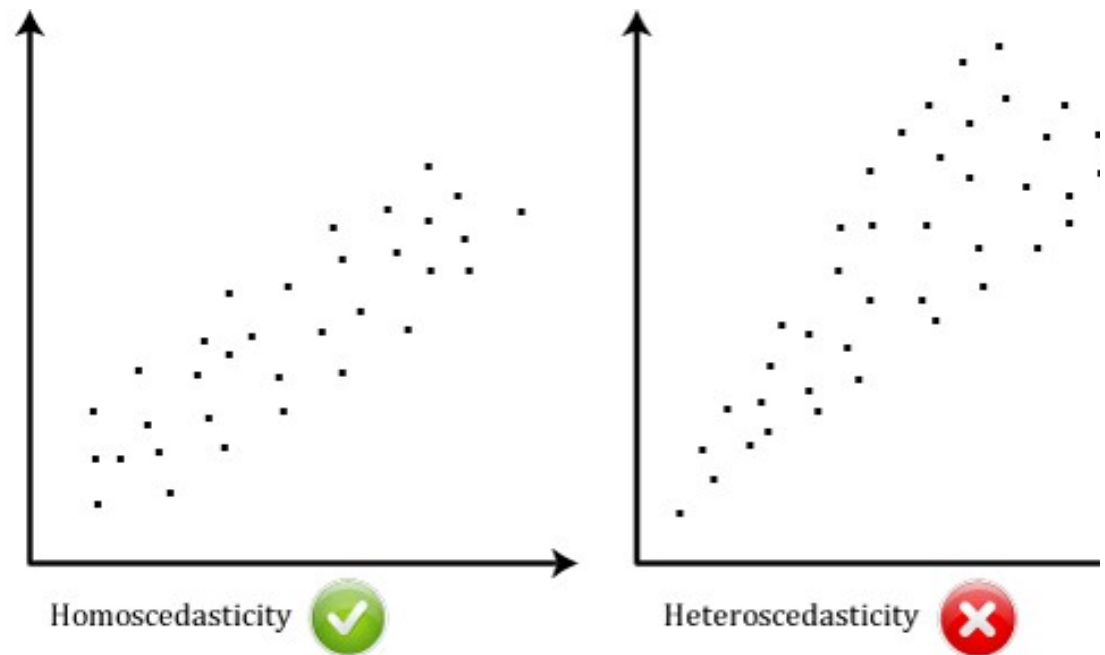
where:

- $y_1, y_2, \dots, y_m$  are the dependent variables.
- $x_1, x_2, \dots, x_n$  are the independent variables.
- $\beta_{10}, \beta_{20}, \dots, \beta_{m0}$  are the intercepts for each dependent variable.
- $\beta_{1i}, \beta_{2i}, \dots, \beta_{mi}$  are the coefficients for each dependent variable with respect to each independent variable.
- $\epsilon_1, \epsilon_2, \dots, \epsilon_m$  are the error terms for each equation.



Multivariate linear regression relies on several key assumptions:

- **Linearity:** The relationship between the dependent variable and each independent variable is linear.
- **Independence:** Observations are independent of each other.
- **Homoscedasticity:** The variance of residuals (errors) is constant across all levels of the independent variables.
- **Normality:** The residuals (errors) are normally distributed.
- **No Multicollinearity:** Independent variables are not highly correlated with each other



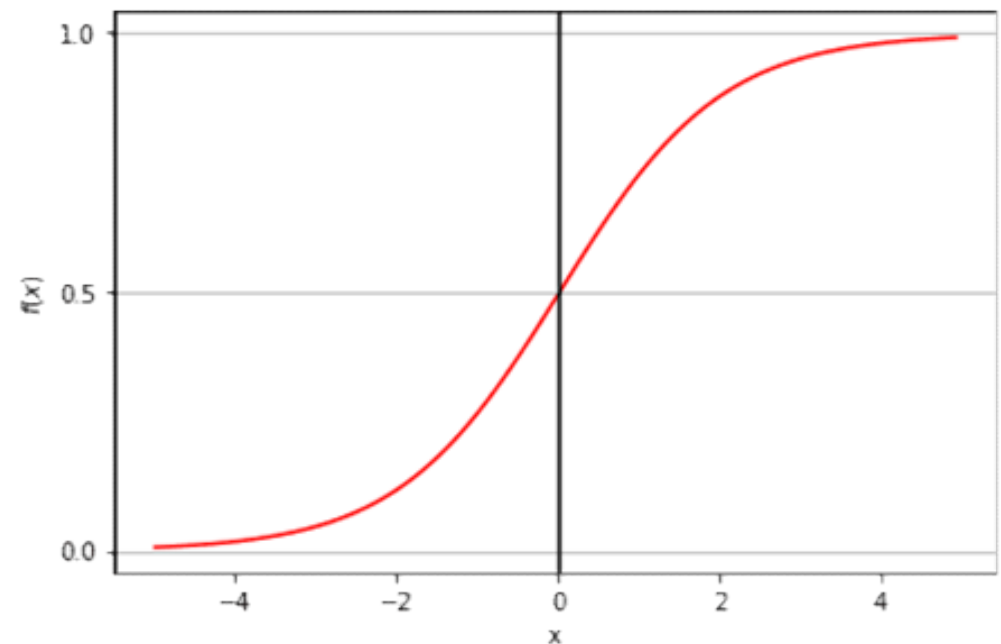
# Logistic Regression

- Logistic regression is a statistical method for predicting binary classes. target variable is categorical in nature.
- The outcome or target variable is BINARY(0/1) in nature-two possible classes.
- Example- cancer detection problem(0/1).
- It computes the probability of an event occurrence.
- It uses a log of odds as the dependent variable. Predicts the probability of occurrence of a binary event utilizing a logit function.
- **Properties of Logistic Regression:**
  - The dependent variable in logistic regression follows Bernoulli Distribution.
  - Estimation is done through maximum likelihood.
  - No R Square, Model fitness is calculated through Concordance, KS-Statistics.

# Sigmoid Function

- map it into a value between 0 and 1
- sigmoid function value  $\leq 0.5$  then output = 1 or YES
- sigmoid function value  $> 0.5$  then output = 0 or NO

$$f(x) = \frac{1}{1 + e^{-x}}$$



## Types of Logistic Regression

- **Binary Logistic Regression:** The target variable has only two possible outcomes such as Spam or Not Spam, Cancer or No Cancer.
- **Multinomial Logistic Regression:** The target variable has three or more nominal categories such as predicting the type of Wine.
- **Ordinal Logistic Regression:** the target variable has three or more ordinal categories such as restaurant or product rating from 1 to 5.
- Reference
  - <https://www.datacamp.com/tutorial/understanding-logistic-regression-python>

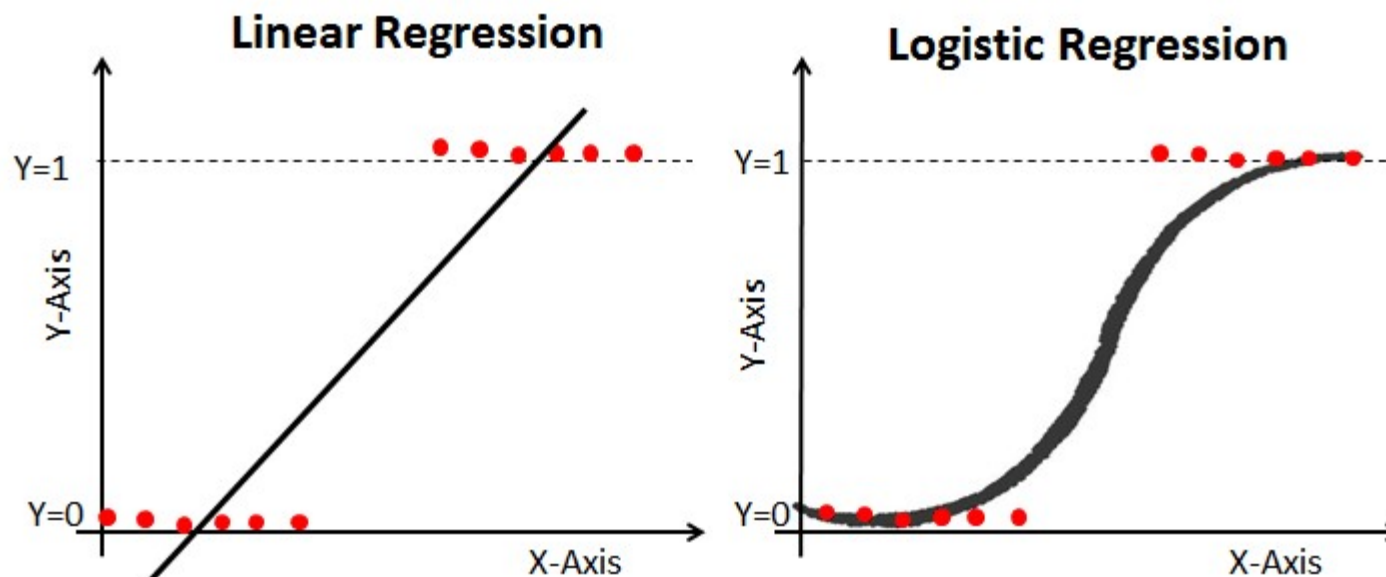
# Linear Regression Vs. Logistic Regression

- **Linear regression**

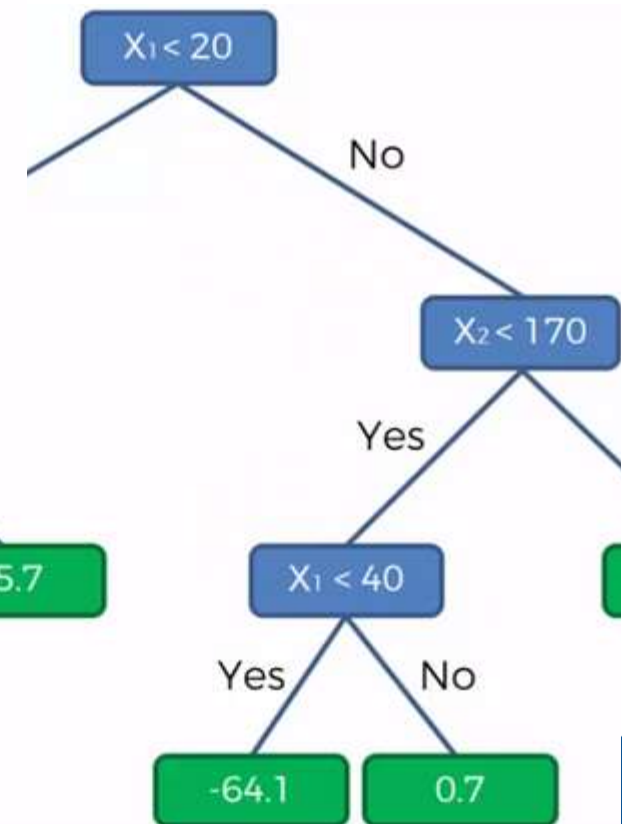
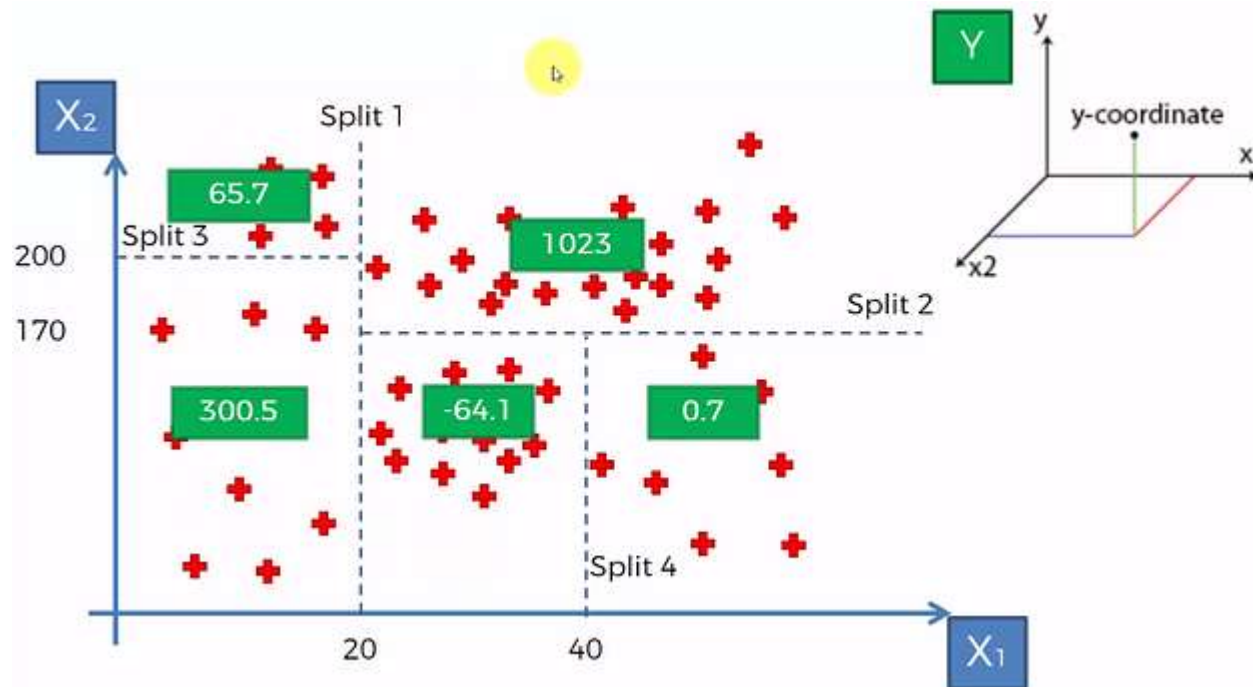
- gives you a continuous output
- Example- house price and stock price prediction.
- Estimated using Ordinary Least Squares (OLS)

- **Logistic regression**

- gives you a discrete output(0/1)
- predicting whether a patient has cancer or not, the customer will churn or Not
- estimated using Maximum Likelihood Estimation (MLE) approach.



# Decision Tree Regression



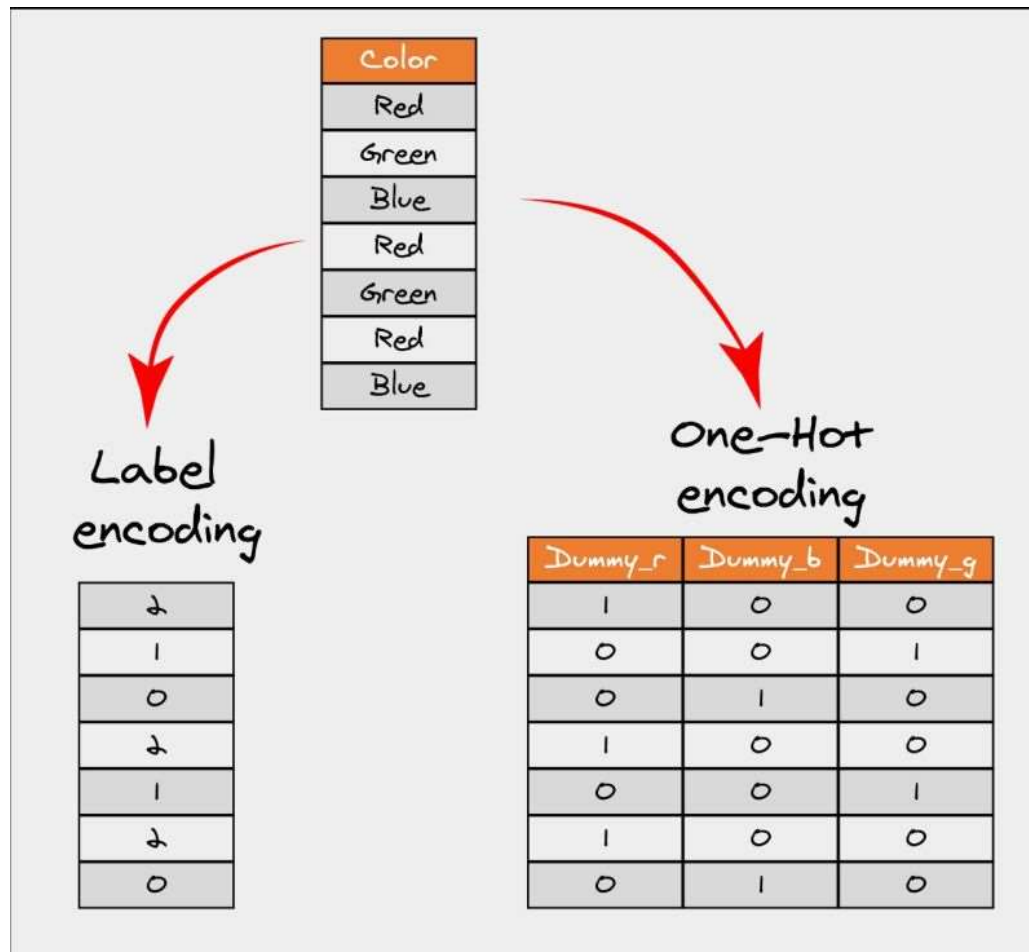
- Reference

- <https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/>

## Class Assignment

- <https://realpython.com/linear-regression-in-python/>
- Identify any dataset
- Implement Linear, Multiple and Multivariate Regression using Python
- Calculate error measures and compare

# Preprocessing- Label Encoding Vs One-Hot Encoding

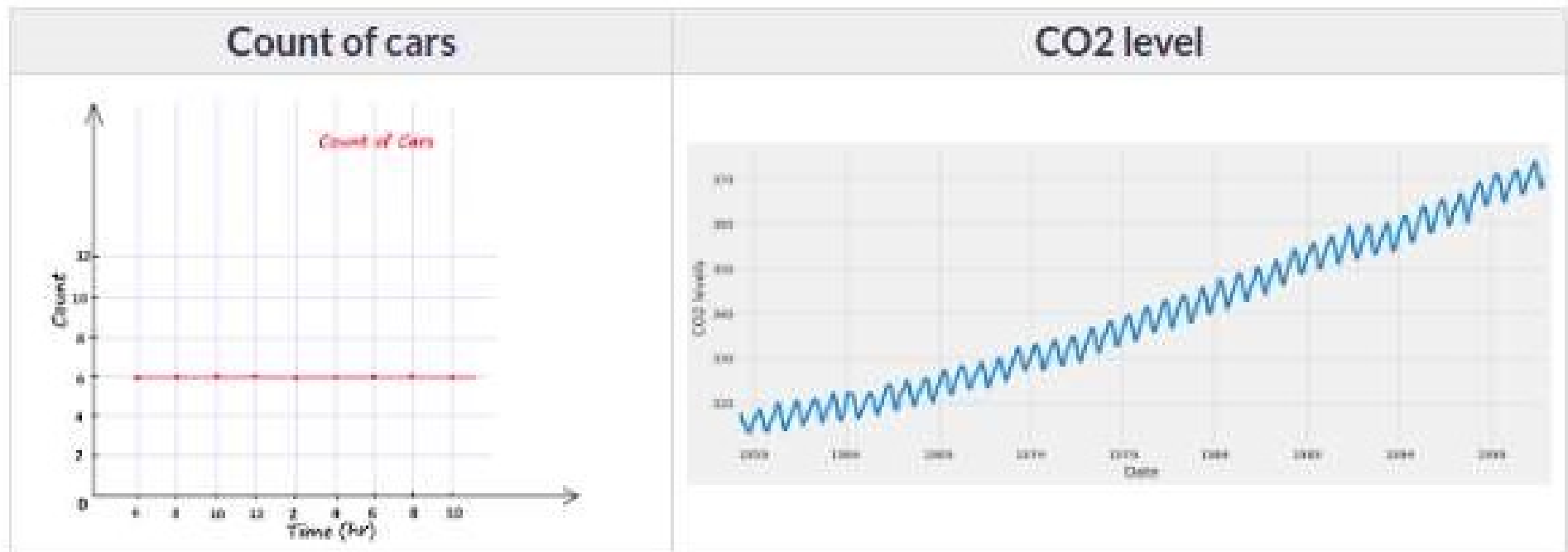


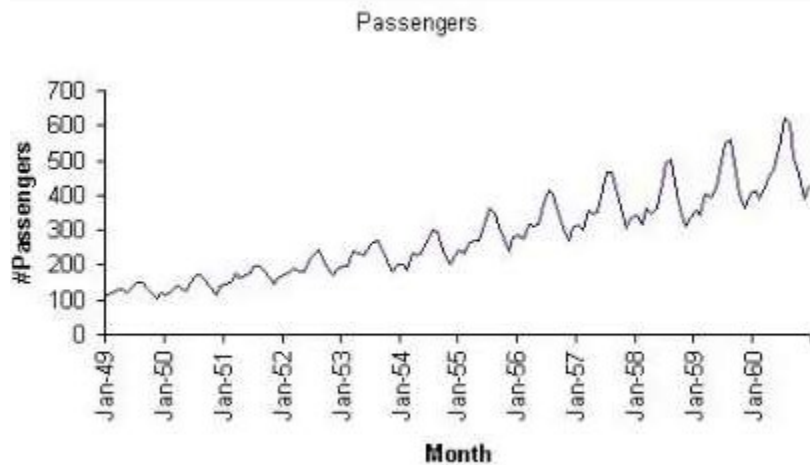
- <https://medium.com/data-folks-indonesia/powering-up-your-pandas-part-ii-label-encoding-and-one-hot-encoding-dac0fce045da>



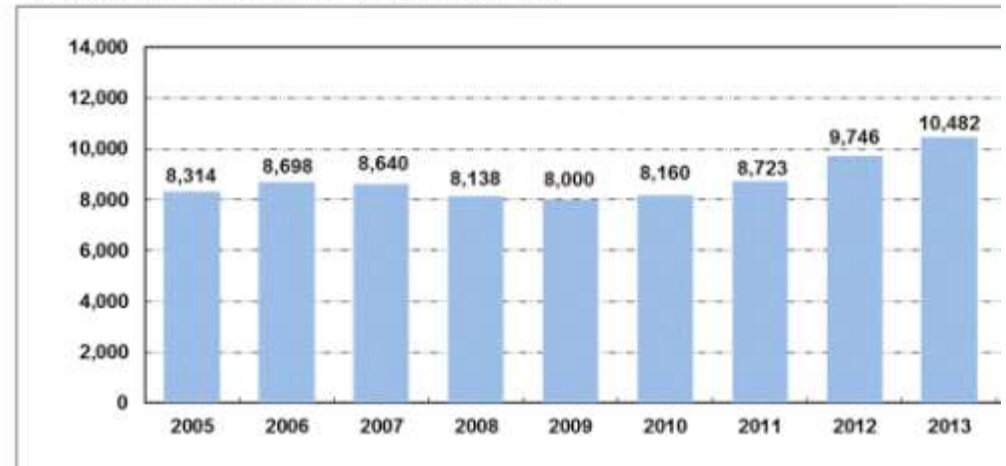
# Statistical Methods for Data Analysis: Time series analysis & Forecasting

- Let's Understand, What is Time Series Data?



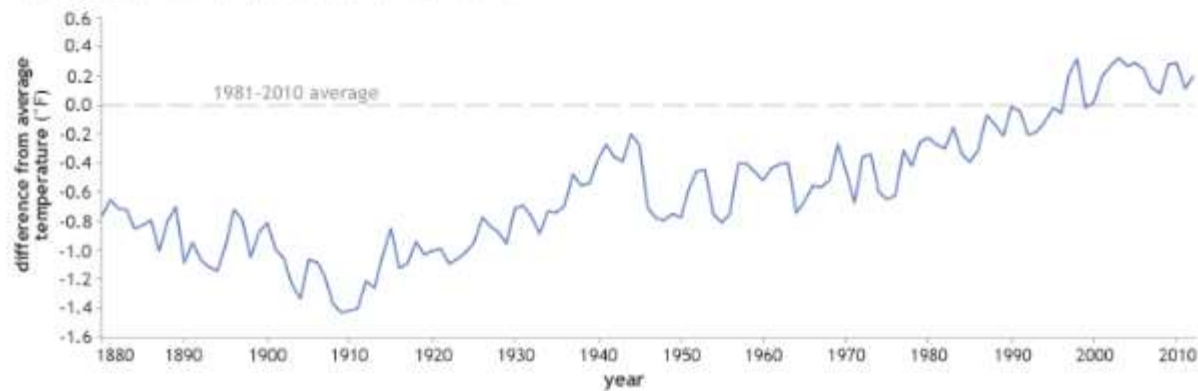


Visitors to Universal Studios Japan (Unit: million)



Source: Universal Studios Japan

Yearly sea surface temperature anomaly 1880-2012



# Time Series Data

- Refers to a sequence of data points collected or recorded at specific time intervals.
- These intervals can be continuous or discrete and are usually uniform (e.g., Second, Minutes, hourly, daily, Weekly, monthly, yearly).
- The primary characteristic of time series data is that each data point is associated with a specific point in time, which makes the temporal order of observations crucial
- Time-series forecasting uses historical and current data **to predict future values over a period** or at a specific point in the future.
- **Examples of Time Series Data**
  - **Finance:** Daily stock prices, quarterly earnings reports, interest rates.
  - **Economics:** Monthly unemployment rates, GDP growth rates, inflation rates.
  - **Meteorology:** Daily temperatures, yearly rainfall totals, hurricane frequency.
  - **Health:** Daily patient counts, weekly infection rates, annual birth rates.
  - **Retail:** Daily sales figures, monthly revenue, yearly customer foot traffic

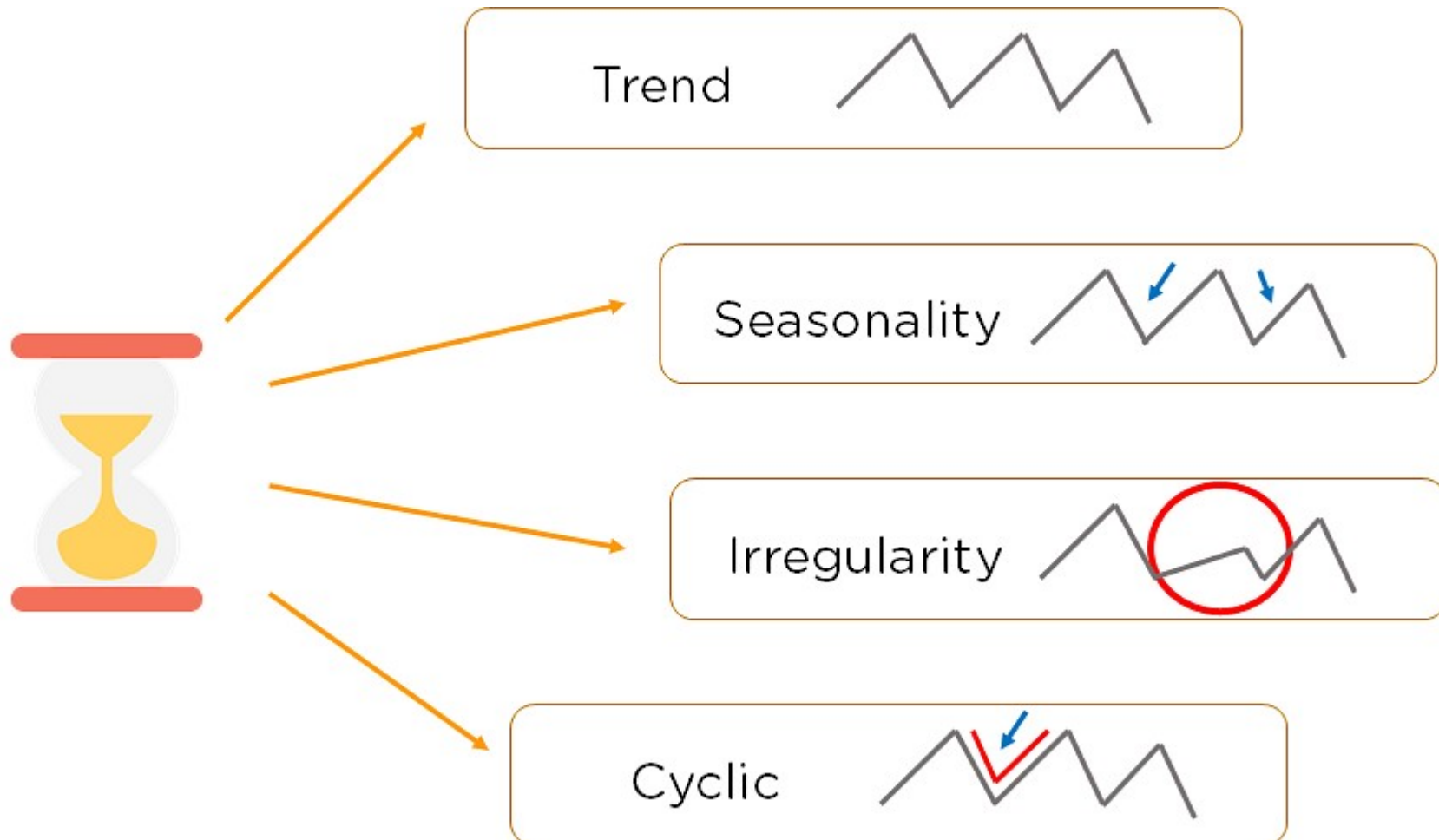
## Objectives of Time Series Analysis








- To understand how time series works and what factors affect a certain variable(s) at different points in time.
- Time series analysis will provide the consequences and insights of the given dataset's features that change over time.
- Supporting to derive the predicting the future values of the time series variable.
- Assumptions: There is only one assumption in TSA, which is “stationary,” which means that the origin of time does not affect the properties of the process under the statistical factor.
- Significance of Time Series
  - Analyzing the historical dataset and its patterns
  - Understanding and matching the current situation with patterns derived from the previous stage.
  - Understanding the factor or factors influencing certain variable(s) in different periods

## Time-based analyses and results

- **Forecasting:** Predicting any value for the future.
- **Segmentation:** Grouping similar items together.
- **Classification:** Classifying a set of items into given classes.
- **Descriptive analysis:** Analysis of a given dataset to find out what is there in it.
- **Intervention analysis:** Effect of changing a given variable on the outcome.

# Components of Time Series Analysis



	Trend	Seasonality	Cyclical	Irregularity
Time	Fixed Time Interval	Fixed Time Interval	Not Fixed Time Interval	Not Fixed Time Interval
Duration	Long and Short Term	Short Term	Long and Short Term	Regular/Irregular
Visualization				
Nature - I	Gradual	Swings between Up or Down	Repeating Up and Down	Errored or High Fluctuation
Nature – II	Upward/Down Trend	Pattern repeatable	No fixed period	Short and Not repeatable
Prediction Capability	Predictable	Predictable	Challenging	Challenging
Market Model				Highly random/Unforeseen Events – along with white noise.

# Time Series Analysis

Time series analysis involves methods to analyze time series data to extract meaningful statistics and other characteristics of the data. The main goals are:

- **Descriptive Analysis:** Summarizing the historical data to understand what has happened over a certain period.
- **Trend Analysis:** Identifying the long-term movement in the data.
- **Seasonality Analysis:** Detecting and understanding regular patterns or cycles.
- **Forecasting:** Predicting future data points based on past patterns.


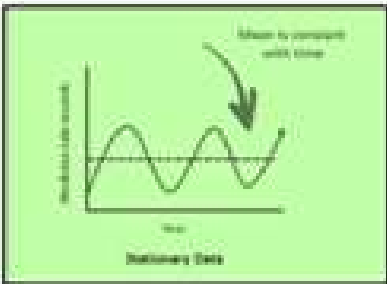
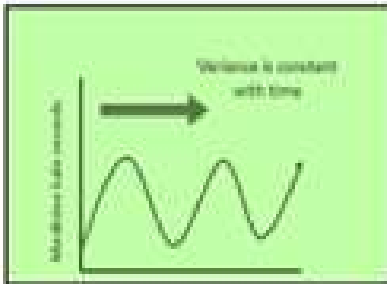
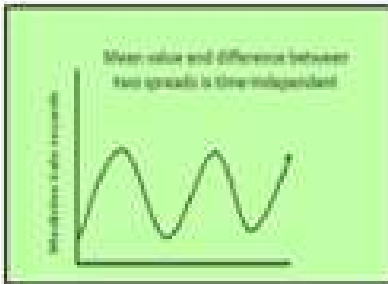

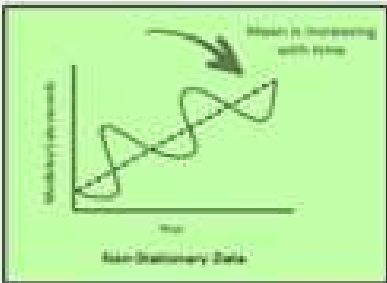
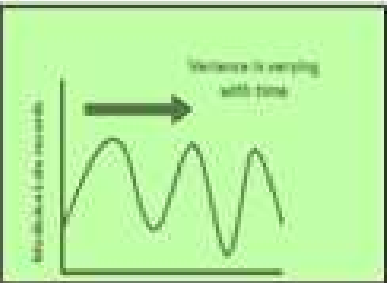
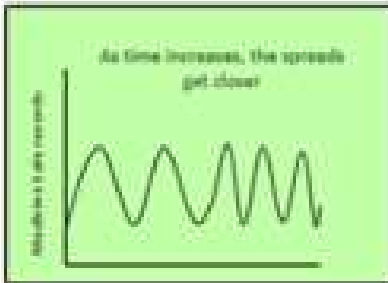


## Limitations of Time Series Analysis

- Similar to other models, the missing values are not supported by TSA
- The data points must be linear in their relationship.
- Data transformations are mandatory, so they are a little expensive.
- Models mostly work on Uni-variate data.

# Data Types of Time Series

- **Stationary:** Mean and Variance should be constant with respect to time, Covariance measures the relationship between two variables
- **Non-Stationary:** mean-variance or covariance is changing with respect to time

	MEAN	Variance	Covariance
<b>Stationary</b> 			
<b>Non-Stationary</b> 			

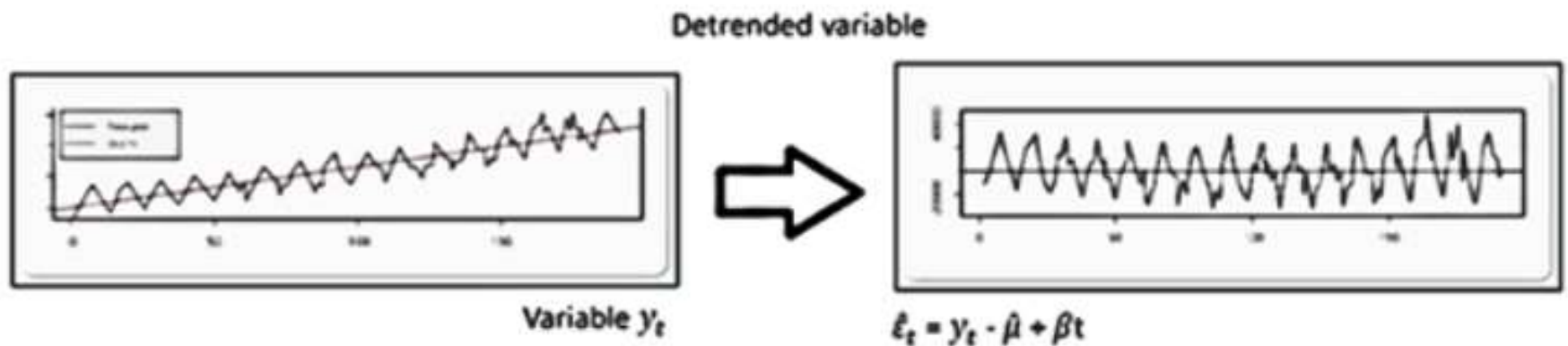
## Methods to Check Stationarity

- There are two **Statistical Tests** available to test if the dataset is stationary:
  - Augmented Dickey-Fuller (ADF) Test or Unit Root Test
  - Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test
- Converting Non-Stationary Into Stationary
  - Detrending
  - Differencing
  - Transformation-Power Transform, Square Root, and Log Transfer. The most commonly use one is Log Transfer.

- Null Hypothesis (H0): Series is non-stationary
- Alternate Hypothesis (HA): Series is stationary
  - p-value  $> 0.05$  Fail to reject (H0)
  - p-value  $\leq 0.05$  Accept (H1)

# Detrending

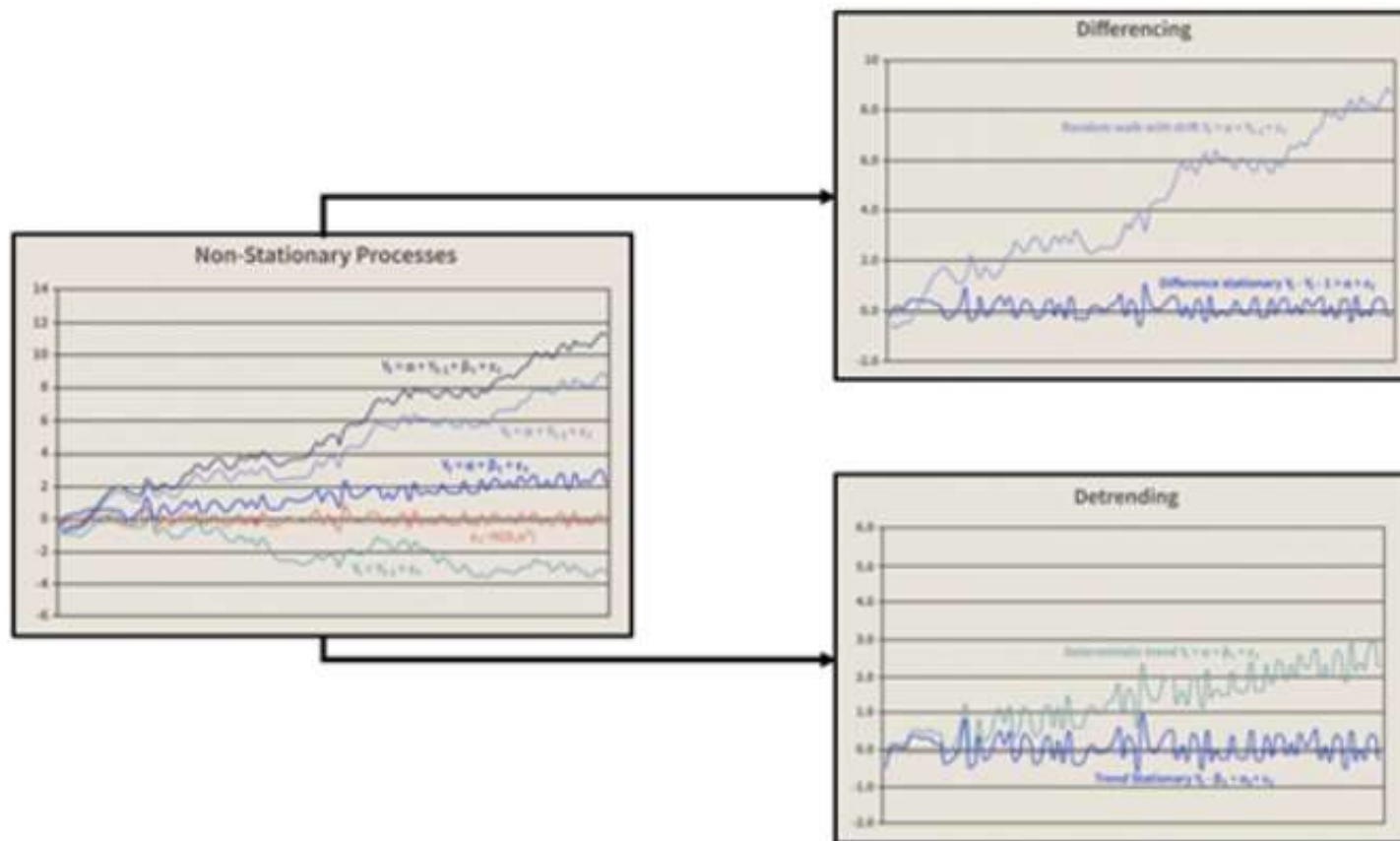
- Removing the trend effects from the given dataset and showing only the differences in values from the trend.
- It always allows cyclical patterns to be identified.



Designed by Author (Shanthababu)

# Differencing

- Simple transformation of the series into a new time series, which we use to remove the series dependence on time and stabilize the mean of the time series
- Trend and seasonality are reduced during this transformation.
- $Y_t = Y_t - Y_{t-1}$        $Y_t = \text{Value with time}$



# Common Techniques in Time Series Analysis

- Moving Averages: Smoothing the data by averaging data points within a sliding window.
- Exponential Smoothing: Applying exponentially decreasing weights to past observations.
- ARIMA (AutoRegressive Integrated Moving Average): A combination of autoregressive and moving average models to forecast future points.
- Seasonal Decomposition: Breaking down the series into trend, seasonal, and irregular components.
- LSTM (Long Short-Term Memory Networks): Using neural networks designed for sequence prediction to capture long-term dependencies

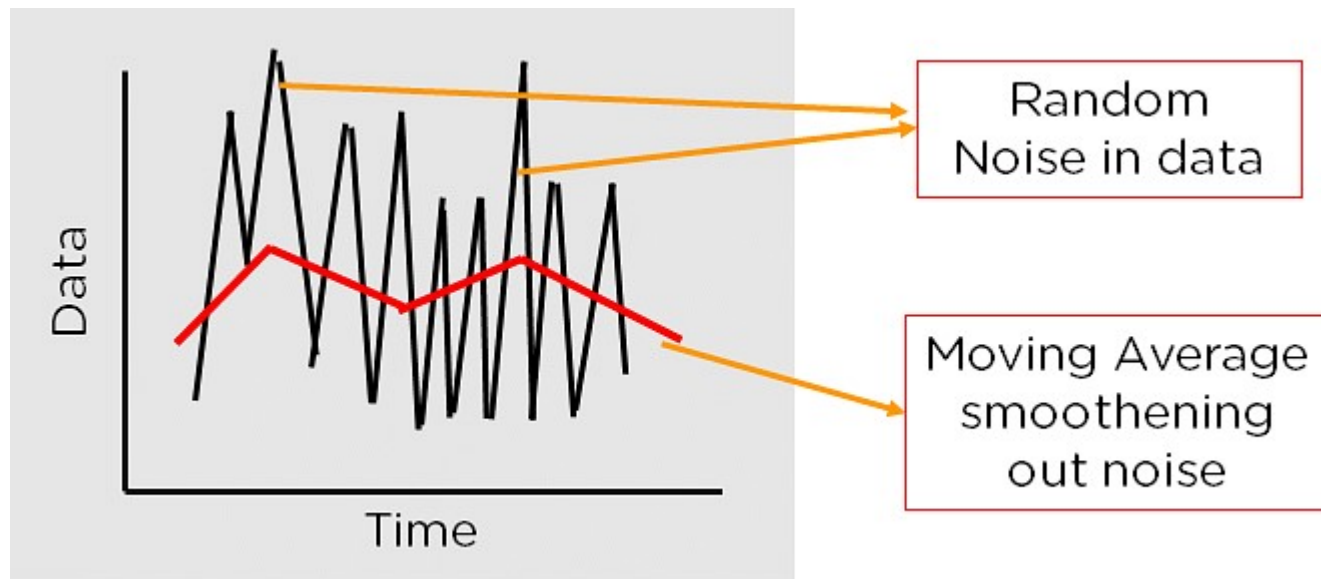
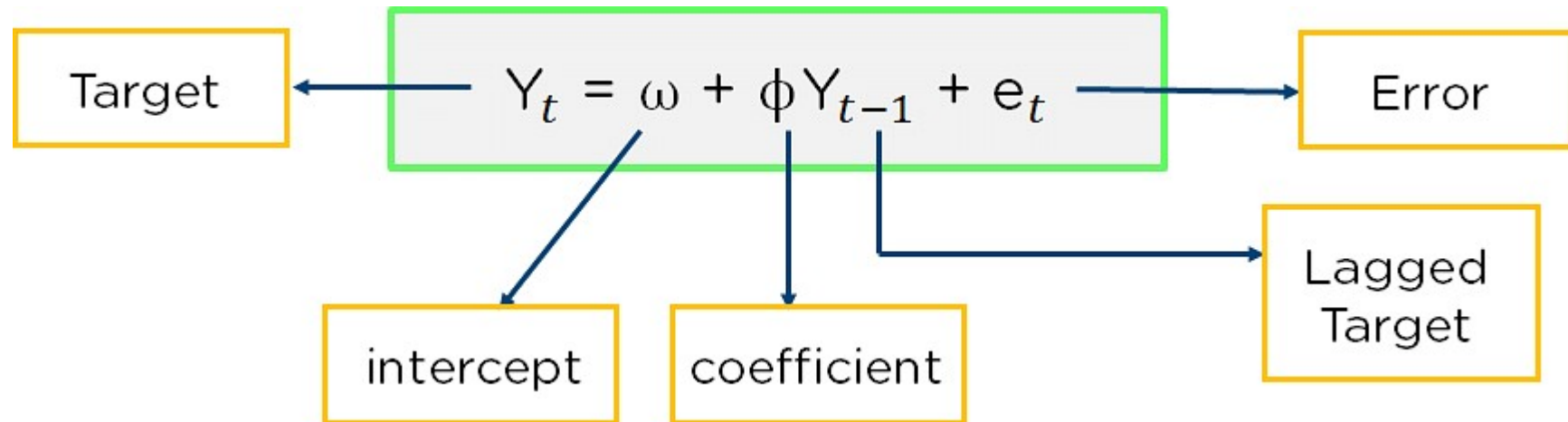
# Auto-Regressive Model

The equation for the AR model (Let's compare  $Y=mX+c$ )

$$Y_t = C + b_1 Y_{t-1} + b_2 Y_{t-2} + \dots + b_p Y_{t-p} + E_{rt}$$

## Key Parameters

- $p$ =past values
- $Y_t$ =Function of different past values
- $E_{rt}$ =errors in time
- $C$ =intercept





## The Moving Average (MA) (

$$SMA_t = \frac{x_t + x_{t-1} + x_{t-2} + \dots + x_{M-(t-1)}}{M}$$

- The value of MA is calculated by taking average data of the time-series within k periods.
  - Simple Moving Average (SMA),
  - Cumulative Moving Average (CMA)
  - Exponential Moving Average (EMA)

	A	N	O	P	Q
1	Any	Avg Temp SMA			
2	1780	14.075			
3	1781	14.71667			
4	1782	13.63333	= (N2+N3+N4)/3		
5	1783	14.4	14.25		
6	1784	13.61667	13.88333		
7	1785	14.15833	14.05833		
8	1786	14.19167			
9	1787	14.025			
10	1788	14.275			
11	1789	13.91667			
12	1790	14.45833			
13	1791	14.44167			
14	1792	14.64167			
15	1793	14.29167			
16	1794	14.66667			
17	1795	14.25833			

$$CMA_t = \frac{x_1 + x_2 + x_3 + \dots + x_t}{t}$$

	A	N	O	P	Q
1	Any	Avg Temp	SMA	CMV	
2	1780	14.075		14.075	
3	1781	14.71667		14.39583	
4	1782	13.63333	14.14167	14.14167	
5	1783	14.4	14.25	=(N2+N3+N4+N5)/4	
6	1784	13.61667	13.88333	14.08833	
7	1785	14.15833	14.05833	14.1	
8	1786	14.19167			
9	1787	14.025			
10	1788	14.275			
11	1789	13.91667			
12	1790	14.45833			
13	1791	14.44167			
14	1792	14.64167			
15	1793	14.29167			
16	1794	14.66667			
17	1795	14.25833			
18	1796	13.75			
19	1797	14.04167			
20	1798	15.075			
21	1799	14.5			
22	1800	14.18333			
23	1801	14			

temperature\_TSA

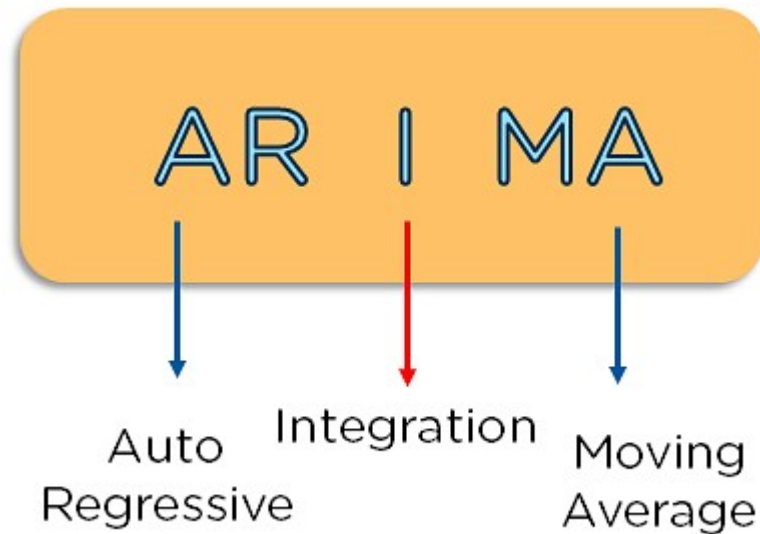
$$EMA_t = \begin{cases} x_0 & t = 0 \\ \alpha x_t + (1 - \alpha) EMA_{t-1} & t > 0 \end{cases}$$

	A	N	O	P	Q	R	S
1	Any	Avg Temp	SMA	CMV	EMA		
2	1780	14.075		14.075	14.075		
3	1781	14.71667		14.39583	14.13917		
4	1782	13.63333	14.14167	14.14167	14.60833		
5	1783	14.4	14.25	14.20625	13.71		
6	1784	13.61667	13.88333	14.08833	14.32167		
7	1785	14.15833	14.05833	14.1	=0.1*N7+0.9*N6		
8	1786	14.19167					
9	1787	14.025					
10	1788	14.275					
11	1789	13.91667					
12	1790	14.45833					
13	1791	14.44167					
14	1792	14.64167					
15	1793	14.29167					
16	1794	14.66667					
17	1795	14.25833					
18	1796	13.75					
19	1797	14.04167					
20	1798	15.075					
21	1799	14.5					
22	1800	14.18333					
23	1801	14					

temperature\_TSA

Model	Use When	Key Identification Tools	Characteristics
AR	Strong autocorrelation in stationary data	PACF	Uses past values
MA	Short-term dependencies in stationary data	ACF	Uses past errors
ARIMA	Non-stationary data with complex dependencies	ACF and PACF	Combines AR, differencing, and MA

## ARIMA Model



- $p \Rightarrow$  lag order  $\Rightarrow$  No of lag observations.
- $d \Rightarrow$  degree of differencing  $\Rightarrow$  No of times that the raw observations are differenced.
- $q \Rightarrow$  order of moving average  $\Rightarrow$  the size of the moving average window

## Implementation steps for ARIMA

- Plot a time series format
- Find the difference for making constant on mean by removing the trend
- Make the variable constant by applying the log transformation
- Note down the different log transformations for making constant on both mean and variance
- Plot ACF and PACF and identify the potential autoregressive and moving average models
- Discover the best fit for the ARIMA model
- Forecast or predict the value using the best fit for the ARIMA model
- Plott ACF and PACF for residuals of the ARIMA model, and ensure no information is left.

## Implementation of time series analysis based on web data.

- <https://towardsdatascience.com/time-series-analysis-and-forecasting-of-web-service-metrics-1e15d7fb72c2>
- [https://data.gov.in/search?title=TEMP\\_ANNUAL\\_SEASONAL\\_MEAN](https://data.gov.in/search?title=TEMP_ANNUAL_SEASONAL_MEAN)

# Complexities of modern datasets

- SELF STUDY

# Recent Technologies and Frameworks for Data Analytics

- SELF STUDY



# Ethical considerations related to data analytics

- Why

- Protection of Individual Privacy
- Building and Maintaining Trust
- Ensuring Fairness and Preventing Bias
- Compliance with Laws and Regulations
- Accountability and Responsibility
- Mitigating Harm
- Promoting Positive Social Outcomes
- Transparency and Openness
- Economic and Reputational Benefits
- Moral and Ethical Duty



# Ethical considerations related to data analytics:

## 1. Privacy

- **Issue:** Data analytics often involves the collection and analysis of personal data. Ensuring the privacy of individuals is a major concern.
- **Considerations:**
  - Collect only necessary data and avoid excessive data collection.
  - Use anonymization and pseudonymization techniques to protect personal information.
  - Implement strong data security measures to prevent unauthorized access.

## 2. Consent

- **Issue:** Individuals must be aware that their data is being collected and analyzed, and they should consent to this usage.
- **Considerations:**
  - Obtain informed consent from individuals before collecting their data.
  - Clearly communicate how the data will be used, stored, and shared.
  - Provide individuals with options to opt-out or withdraw their consent.

## Ethical considerations related to data analytics:

### 3. Transparency

- **Issue:** Transparency about data collection and analysis processes is crucial to maintain trust.
- **Considerations:**
  - Clearly disclose data collection practices and purposes to data subjects.
  - Be transparent about the algorithms and methodologies used in data analysis.
  - Provide accessible explanations of how decisions based on data analytics are made.

### 4. Bias and Fairness

- **Issue:** Data analytics can perpetuate and even amplify existing biases in the data, leading to unfair outcomes.
- **Considerations:**
  - Ensure that the data used for analysis is representative and unbiased.
  - Regularly audit and test algorithms for bias and take corrective actions if biases are found.
  - Consider the impact of data-driven decisions on different groups and strive for fairness.

## Ethical considerations related to data analytics:

### 5. Accountability

- **Issue:** There must be accountability for the outcomes of data analytics and decisions made based on data.
- **Considerations:**
  - Establish clear lines of responsibility for data governance and analytics processes.
  - Implement mechanisms for auditing and oversight of data analytics activities.
  - Be prepared to explain and justify data-driven decisions and actions.

### 6. Data Security

- **Issue:** Protecting data from breaches and unauthorized access is critical.
- **Considerations:**
  - Implement robust cybersecurity measures to protect data.
  - Regularly update security protocols and conduct security audits.
  - Train employees on data security best practices and potential threats.

## Ethical considerations related to data analytics:

### 7. Purpose Limitation

- **Issue:** Data should be used only for the specific purposes for which it was collected.
- **Considerations:**
  - Define clear and legitimate purposes for data collection and use.
  - Avoid using data for purposes beyond the original intent without obtaining additional consent.
  - Regularly review data usage practices to ensure compliance with purpose limitations.

### 8. Impact on Society

- **Issue:** The use of data analytics can have broader societal impacts that need to be considered.
- **Considerations:**
  - Assess the potential societal impacts of data analytics projects.
  - Engage with stakeholders, including the public, to understand concerns and perspectives.
  - Strive to use data analytics to promote positive social outcomes and mitigate negative impacts.

# Ethical considerations related to data analytics:

## 9. Regulatory Compliance

- **Issue:** Compliance with laws and regulations governing data protection and privacy is essential.
- **Considerations:**
  - Stay informed about relevant data protection laws (e.g., GDPR, CCPA) and ensure compliance.
  - Implement data governance frameworks that align with regulatory requirements.
  - Regularly review and update data policies and practices to maintain compliance.

# Role of data analytics in Text, Social Media, and Image

## Text Analytics

- **Text analytics/** text mining / NLP, involves analyzing unstructured text data to extract meaningful information and insights.
- **Key Applications:**
  - **Sentiment Analysis** - sentiments in customer reviews, social media posts, or feedback
  - **Topic Modelling** - Identify topics or themes within a large corpus of text,
  - **Text Classification** - Categorize text into predefined categories
  - **Named Entity Recognition (NER)** -Identify and classify named entities in text
  - **Keyword Extraction** -Extract significant keywords or phrases from text
- **Tools and Techniques:**
  - **NLP Libraries:** NLTK, spaCy, Gensim
  - **Machine Learning Algorithms:** Naive Bayes, SVM, Random Forests
  - **Deep Learning Models:** RNNs, LSTMs, Transformers (e.g., BERT, GPT)

# Image Analytics

- **Image analytics** involves extracting meaningful information from images using computer vision and machine learning techniques
- **Key Applications**
  - Image Classification
  - Object Detection
  - Image Segmentation
  - Image Enhancement
  - **Content-Based Image Retrieval (CBIR)**
- **Tools and Techniques**
  - **Computer Vision Libraries:** OpenCV, PIL (Python Imaging Library)
  - **Deep Learning Frameworks:** TensorFlow, Keras, PyTorch
  - **Pre-trained Models:** VGG, ResNet, Inception for image classification; YOLO, SSD for object detection



# Social Media Analytics

- **Social media analytics** involves collecting and analyzing data from social media platforms to understand user behavior, track brand reputation, and gauge public opinion
- **Key Applications:**
  - Brand Monitoring
  - Influencer Analysis
  - Campaign Analysis
  - Trend Analysis
  - Customer Insights
- **Tools and Techniques:**
  - **Social Media Platforms:** APIs from Twitter, Facebook, Instagram, LinkedIn
  - **Analytics Tools:** Hootsuite, Sprout Social, Brandwatch
  - **Sentiment Analysis:** Determine the sentiment of social media posts and comments
  - **Network Analysis:** Analyze social networks to find connections and influential nodes

# Brand Monitoring

- **Brand Monitoring:** Track mentions of a brand, product, or service to understand public perception and respond to customer feedback.
- **Benefits**
  - Increases brand awareness through listening
  - Effective monitoring enhances customer relationships
  - Gain a competitive advantage by tracking trends
  - Avoid and prevent social media crises before they take off
  - Empower the rest of your organization with timely data



# Brand Monitoring

- Explore any one FREE social media monitoring tool
- <https://www.youtube.com/watch?v=Dxn0ZtyXeGM>

# Influencer Analysis

- Data-driven process used to identify and evaluate individuals who have the potential to influence others' opinions, behaviors, and purchasing decisions within a particular niche or industry.
- Benefits of Influencer Analysis
  - Targeted Marketing -reach highly specific audiences
  - Increased Engagement
  - Enhanced Credibility -when they provide genuine endorsements.
  - Cost-Effective
  - Measurable Results -Influencer campaigns can be tracked and analyzed for effectiveness
- Challenges in Influencer Analysis
  - Fake Followers - fake accounts
  - Maintaining Authenticity – genuine or Not
  - Finding the Right Fit
  - Managing Relationships -requires careful management.
- <https://www.modash.io/blog/influencer-analysis-tools>

# Sentiment Analysis



## Self Check

- **How is Advanced Data Analytics different from traditional data analysis?**
- **What are the critical challenges in implementing Advanced Data Analytics?**
- **What skills are required to become an Advanced Data Analyst?**
- **What are the ethical considerations in Advanced Data Analytics?**
- **Can Advanced Data Analytics be applied to small businesses?**
- **How can organizations ensure data security in Advanced Data Analytics?**

## MCA-Lab Exercises

1. Implementation of Correlation and Regression
2. Implementation of time series analysis

## Topic to Discuss

1. Why can't Robots Click the "I'm Not a Robot" Box on Websites