B. Tech CSE

Course : Data Science

Course Code : 20UCS701D

Syllabus

- **Unit 1: Introduction:** State of the practice in Analytics- BI Vs Data Science, Current Analytical Architecture, Data Analytic Life Cycle: Overview, phase 1- Discovery, Phase 2- Data preparation, Phase 3- Model Planning, Phase 4- Model Building, Phase 5- Communicate Results, Phase 6 Operationalize, Descriptive Statistics, Probability Distributions
- **Unit 2: Inferential Statistics:** Inferential Statistics through hypothesis tests, Permutation & Randomization Test Statistical Methods for Evaluation: Hypothesis Testing, Difference of Means, Wilcoxon Rank-Sum Test, Type I and Type II Errors, Power and Sample Size
- **Unit 3: Regression & ANOVA:** Regression, ANOVA Interactive (Analysis of Variance), teaching using Regression- linear, logistics, reasons to choose and cautions, additional regression models

Syllabus

- **Unit 4: Machine Learning Introduction and Concepts :** Differentiating algorithmic and model based frameworks, Regression: Ordinary Least Squares, Ridge Regression, Lasso Regression, K Nearest Neighbours, Regression & Classification
- **Unit 5: Supervised Learning with Regression and Classification techniques :** Variance Dichotomy, Model Validation Approaches, Logistic Regression, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Regression and Classification Trees, Support Vector Machines, Ensemble Methods: Random Forest, Neural Networks, Deep learning
- **Unit 6: Unsupervised Learning and data modeling :** Clustering, Associative Rule Mining, LogicalModelling: Converting a conceptual model to logical model, Integrity constraints, Normalization.

Books for reference

- 1.Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: springer, 2009.

- 2.David Dietrich, Barry Hiller, “Data Science & Big Data Analytics”, EMC education services, Wiley publications, 2012

Unit1:
Introduction to Data Science

Contents

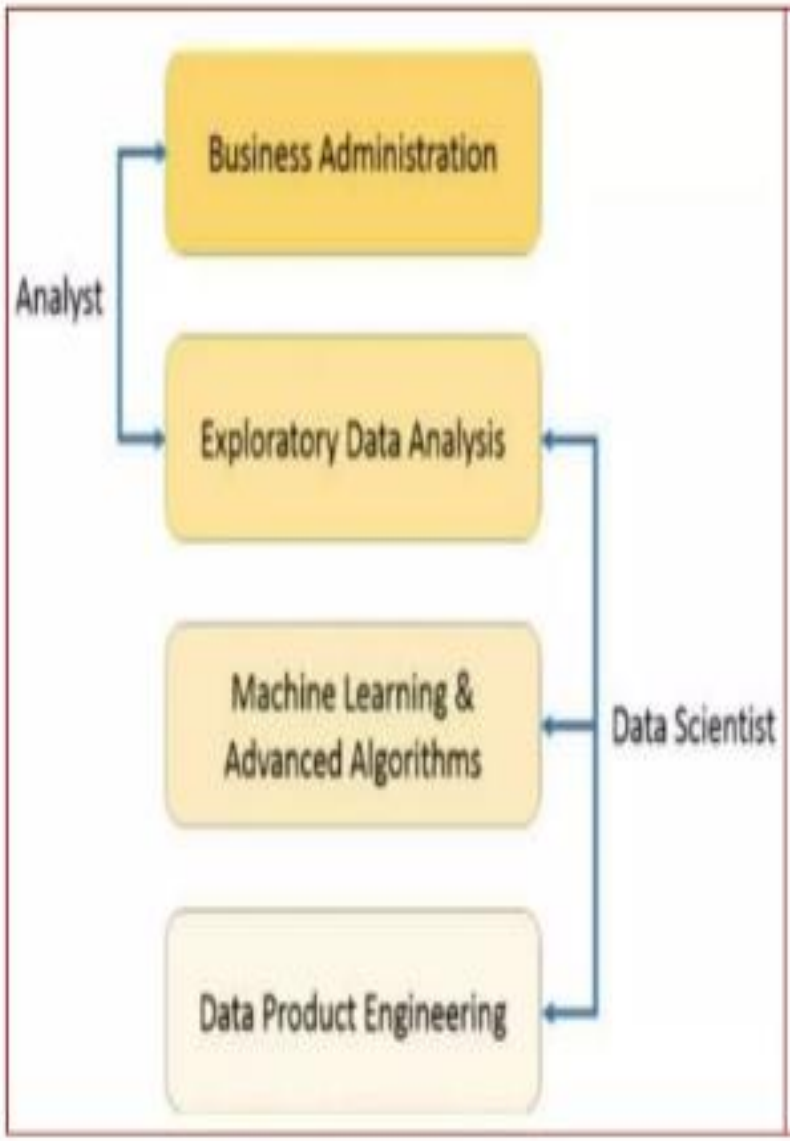
- What is Data Science?
- Why Data Science?
- Data Science Applications
- State of the Practice in Analytics
- BI Vs Data Science
- Current Analytic Architecture

What is Data Science?

- **Data science** is an interdisciplinary field that uses statistics, scientific methods, processes, algorithms, and systems **to extract knowledge and insights** from structured and unstructured data and apply knowledge and actionable insights from data across a broad range of application domains.
- Data Science is related to Big Data, Machine Learning & Data Mining.


Visualization
Scientific Methods
Statistical Modeling
Real-world Applications
Data Science
Statistical Computing
Data Consulting
Data Technology
Data Research

What is Data Science?



Why Data Science?

- Traditionally, the data that we had was mostly structured and small in size which could be analyzed by using simple BI tools.
- Business Intelligence Tools:
  - Microsoft Power BI
  - Tableau
  - Qlik
  - SAS BI

Why Data Science?

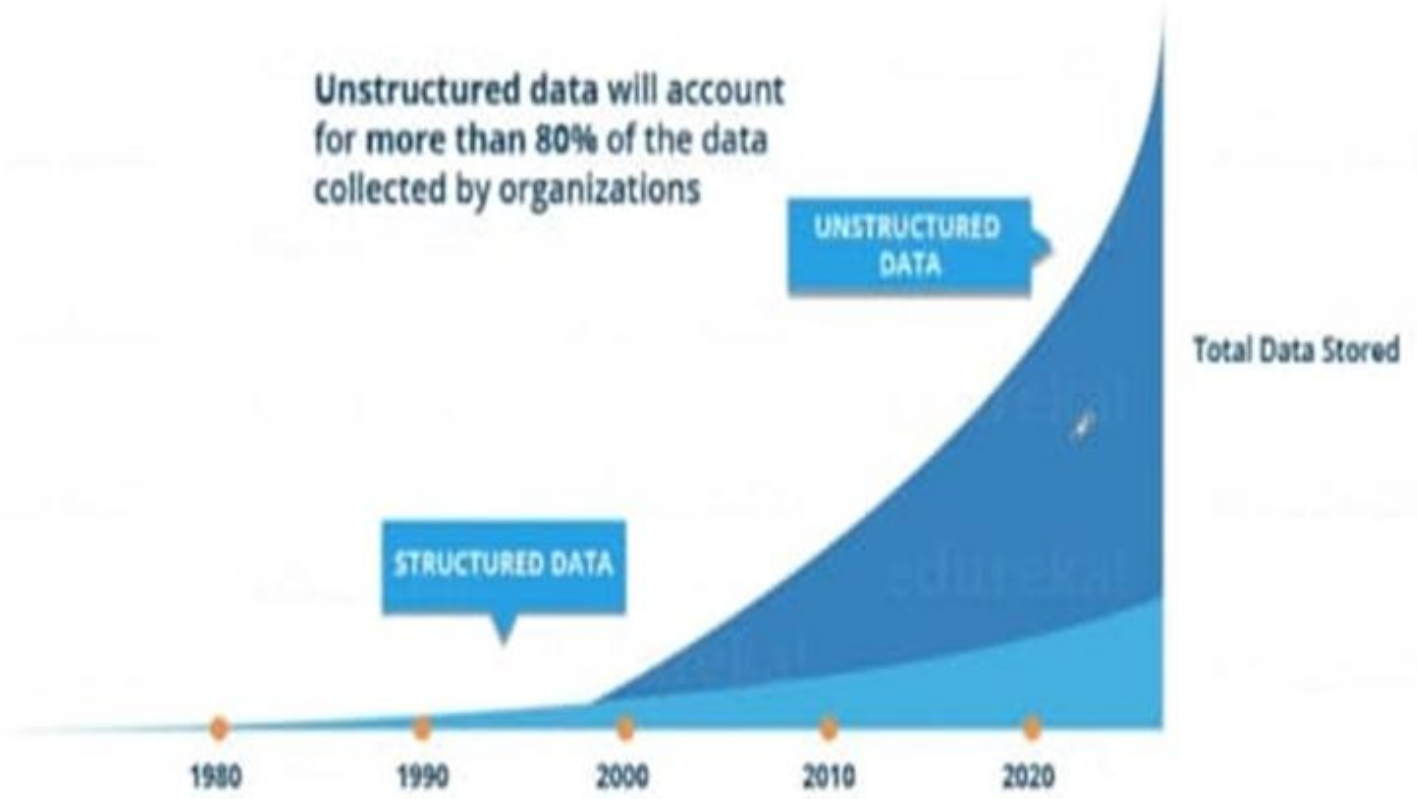
- Today most of the data is unstructured or semi-structured & Lots of data is being generated, collected and warehoused
  - e-commerce
  - Financial transactions, bank/credit transactions
  - Healthcare
  - Online trading and purchasing
  - Social Media & Web
  - Manufacturing & Automotive
  - Internet & Easily available high-end computers

Why Data Science?

- Simple BI tools are not capable of processing this huge volume & variety of data.

- That's why we need more complex and advanced analytical tools and algorithms for processing, analyzing and drawing meaningful insights out of it.

Why Data Science?



Why Data Science is Important?

- Every business has data but its business value depends on how much they know about the data they have.
- It can help businesses to increase the business value of its available data, which in turn can help them take a competitive advantage against their competitors.
- It can help us to know our customers better,
- It can help us to optimize our processes
- It can help us to take better decisions.
- Because of data science, data has become a strategic asset.

Data Science Applications

What Is Statistics?

1. **Collecting Data**
   e.g., Survey
2. **Presenting Data**
   e.g., Charts & Tables
3. **Characterizing Data**
   e.g., Average

Data Analysis

Why?

Decision-Making

What Is Statistics?

- **Statistics** is the science of data.
- It involves
  - collecting,
  - classifying,
  - summarizing,
  - organizing,
  - analyzing, and
  - interpreting numerical information.

Statistics: Two Processes

Describing sets of data

and

Drawing conclusions (making estimates, decisions, predictions, etc. about sets of data based on sampling)

State of the Practice in Analytics

- Current business problems provide many opportunities for organizations to become more analytical and data driven.

Business Driver	Examples
Optimize business operations	Sales, pricing, profitability, efficiency
Identify business risk	Customer churn, fraud, default
Predict new business opportunities	Upsell, cross-sell, best new customer prospects
Comply with laws or regulatory requirements	Anti-Money Laundering, Fair Lending, Base l II-III, Sarbanes-Oxley (SOX)

State of the Practice in Analytics summary

- Organizations have been trying –
  - To reduce customer churn
  - Increase sales and
  - Increase cross-sell
  - To find a new opportunity

Business Intelligence (BI)

- **Business Intelligence is the process for analyzing data and presenting actionable information that will help in decision-making**.

- It is a set of methodologies, processes, theories that transform raw data into useful information to help companies make better decisions.

- BI technologies include reporting, online analytical processing, data mining, text mining, predictive and prescriptive analytics.

Business Intelligence (BI)

- BI can be used by enterprises to support a wide range of **business decisions- ranging from operational to strategic**.

Eg.

- Operational – basic operating decisions include product positioning or pricing.

- Strategic – priorities, goals, and directions at the broadest level.

BI and Data Science

Features	Business Intelligence (BI)	Data Science
Data Sources	Structured (Usually SQL, often Data Warehouse)	Structured, Semi-structured & Unstructured (Logs, Cloud data, SQL, NoSQL, text)
Approach	Statistics and Visualization	Statistics, Machine Learning, Graph analysis, NLP
Focus	Past and Present	Present and Future
Tools	Microsoft BI, Tableau, QlikView, SAS BI	R, Python, BigML, Weka, RapidMiner

S. No.	Factor	Data Science	Business Intelligence
1.	Concept	It is a field that uses mathematics, statistics and various other tools to discover the hidden patterns in the data.	It is basically a set of technologies, applications and processes that are used by the enterprises for business data analysis.
2.	Focus	It focuses on the future.	It focuses on the past and present.
3.	Data	It deals with both structured as well as unstructured data.	It mainly deals only with structured data.
4.	Flexibility	Data science is much more flexible as data sources can be added as per requirement.	It is less flexible as in case of business intelligence data sources need to be pre-planned.
5.	Method	It makes use of the scientific method.	It makes use of the analytic method.
6.	Complexity	It has a higher complexity in comparison to business intelligence.	It is much simpler when compared to data science.
7.	Expertise	It’s expertise is data scientist.	It’s expertise is the business user.
8.	Questions	It deals with the questions of what will happen and what if.	It deals with the question of what happened.

S. No.	Factor	Data Science	Business Intelligence
9.	Storage	The data to be used is disseminated in real-time clusters.	Data warehouse is utilized to hold data.
10.	Integration of data	The ELT (Extract-Load-Transform) process is generally used for the integration of data for data science applications.	The ETL (Extract-Transform-Load) process is generally used for the integration of data for business intelligence applications.
11.	Tools	It's tools are SAS, BigML, MATLAB, Excel, etc.	It's tools are InsightSquared Sales Analytics, Klipfolio, ThoughtSpot, Cyfe, TIBCO Spotfire, etc.
12.	Usage	Companies can harness their potential by anticipating the future scenario using data science in order to reduce risk and increase income.	Business Intelligence helps in performing root cause analysis on a failure or to understand the current status.
13.	Business Value	Greater business value is achieved with data science in comparison to business intelligence as it anticipates future events.	Business Intelligence has lesser business value as the extraction process of business value carries out statically by plotting charts and KPIs (Key Performance Indicator).
14.	Handling data sets	The technologies such as Hadoop are available and others are evolving for handling understandingItsItsarge data sets.	The sufficient tools and technologies are not available for handling large data sets.

Data Scientist, Data Analyst, Data Engineer

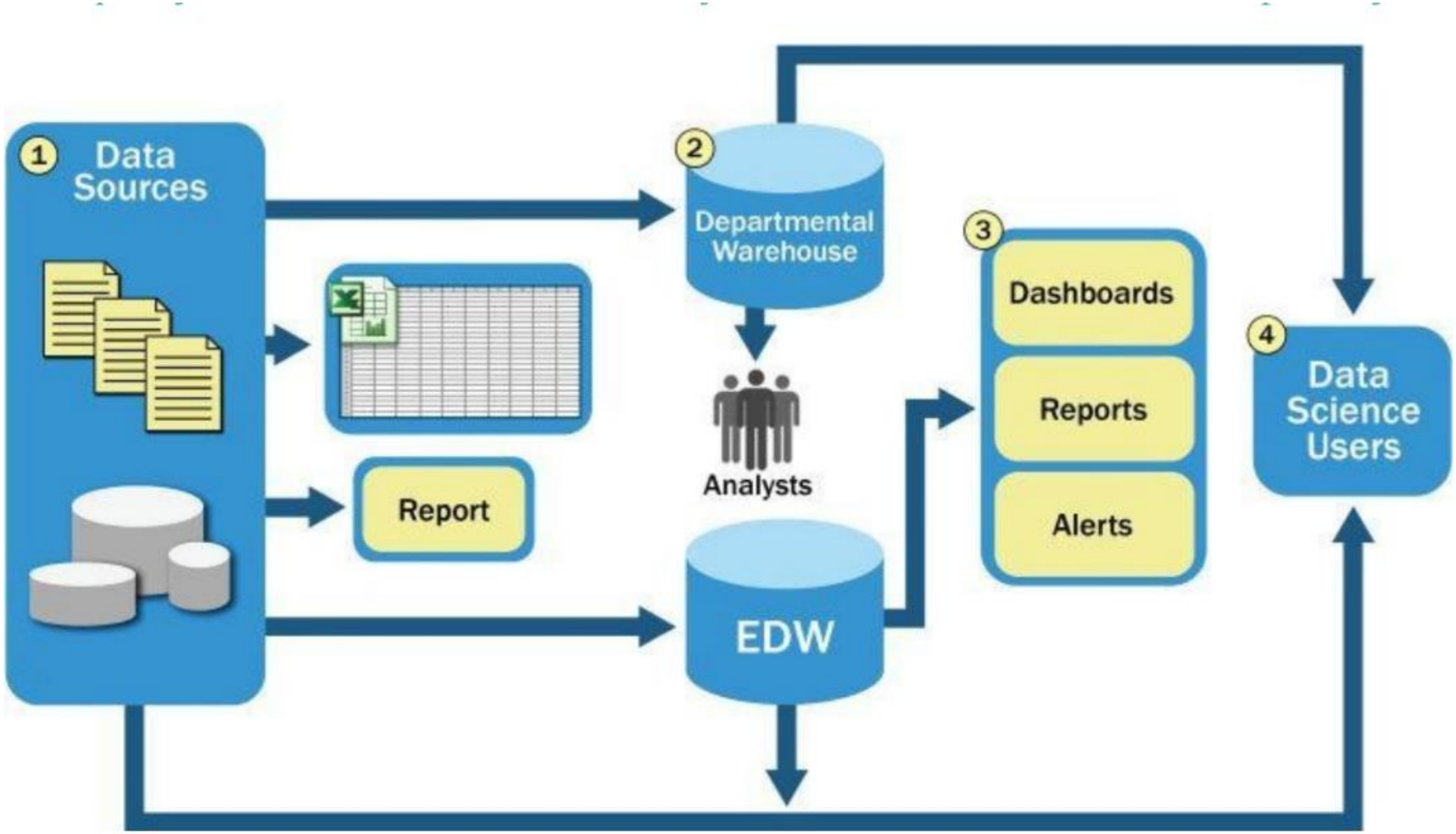
1. Data Scientist:- Good statistics, Data Scientist implies the ability to work with large volumes of data generated by organizational processes. Data driven, decision making.

2. Data Analyst :- collect, process and perform statistical analysis of data. Their goals- how data can be used to answer questions and solve problems.

3. Data Engineer :- A specialist in data wrangling. Data engineers take the messy data…and build the infrastructure for real, tangible analysis. Perform ETL, enrich and clean data.

Current Analytical Architecture

- Identify various components of Analytical Architecture



Current Analytical Architecture

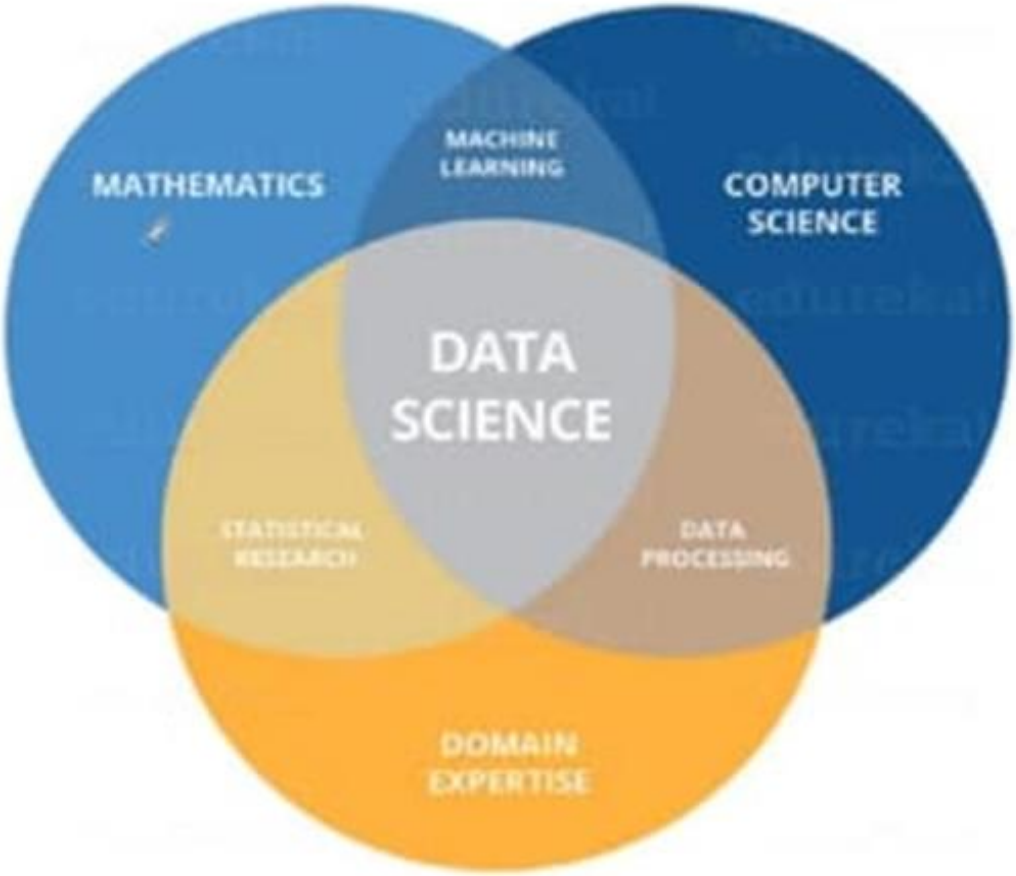
1. Data Sources – Data is generated here from different sources.

2. Data Warehouse – Systematic way of storing data.

3. EDW – High-quality data for visualizing.

4. Data Scientist – Uses all the data from data sources, data warehouse, and EDW for decision-making and predictions.

Skills required for Data Scientist



Data Analytic Life Cycle/ Data Science Process

Phase 1. Discovery – Frame or define the business problem

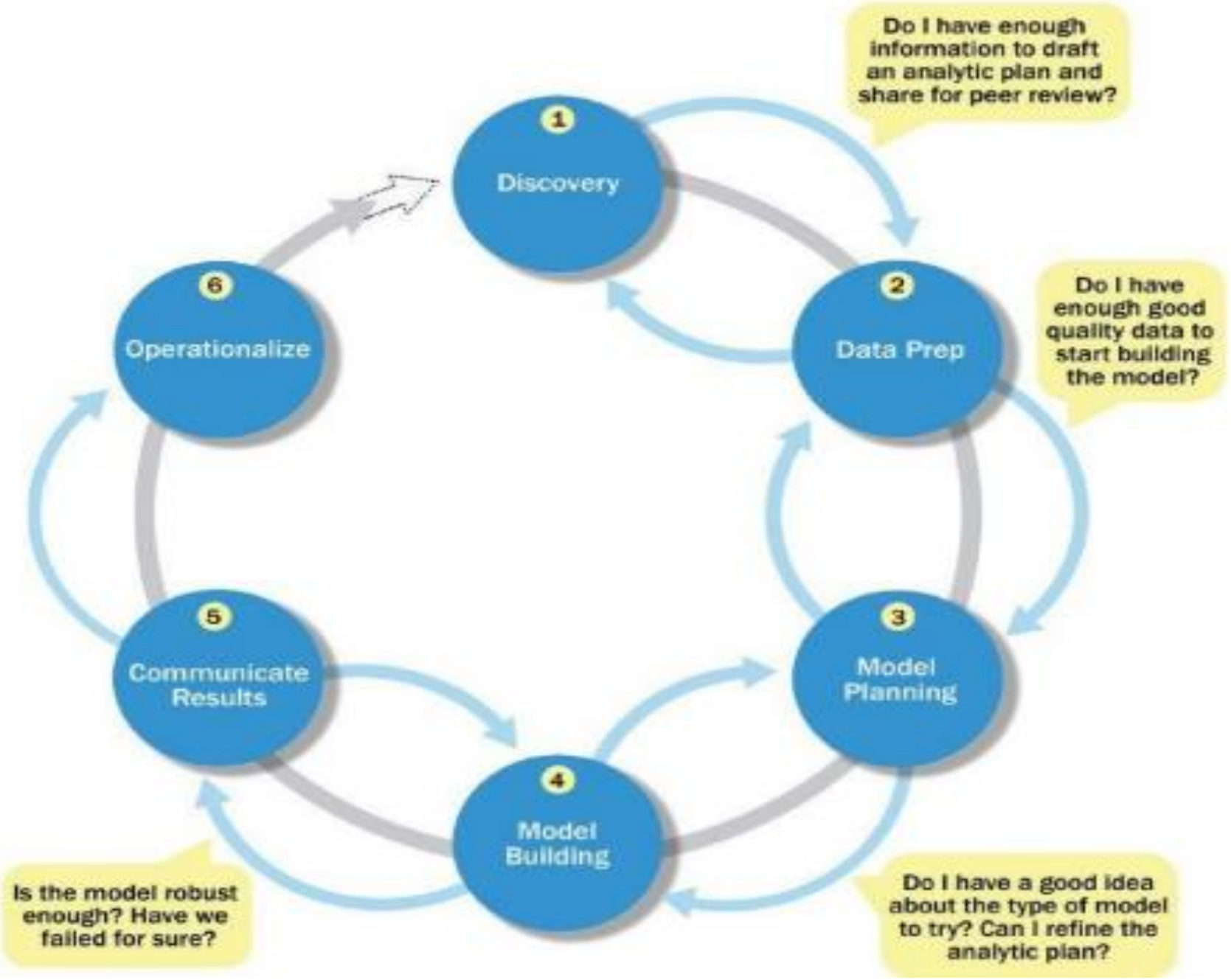
Phase 2. Data preparation – Collect the raw data needed for your problem

Phase 3. Model planning

Phase 4. Model Building

Phase 5. Communicate results

Phase 6. Operationalize


Do I have enough information to draft an analytic plan and share for peer review?
1
Discovery
2
Data Prep
Do I have enough good quality data to start building the model?
3
Model Planning
Do I have a good idea about the type of model to try? Can I refine the analytic plan?
4
Model Building
Is the model robust enough? Have we failed for sure?
5
Communicate Results
6
Operationalize

Phase 1. Discovery

Phase 1. Discovery

- Before you begin the project, it is important to understand the various specifications, requirements, priorities and required budget. You must possess the ability to ask the right questions. Here, you assess if you have the required resources present in terms of people, technology, time and data to support the project.

Phase 1. Discovery

- In this phase, the data science team must learn and investigate the problem, develop context and understanding, and learn about the data sources needed and available for the project.

- In addition, the team formulates initial hypotheses that can later be tested with data.

Phase 1. Discovery

1.1 Learning the Business Domain : Understanding the domain area of the problem is essential.

1.2 Resources: The team needs to assess the resources available to support the project. e.g. Resources include technology, tools, systems, data, and people.

1.3 Framing the Problem : Set Objectives, Framing is the process of stating the analytics problem to be solved.

Phase 1. Discovery

1.4 Identifying Key Stakeholders : to identify the key stakeholders and their interests in the project.

1.5 Interviewing the Analytics Sponsor : The team must use its knowledge and expertise to identify the true underlying problem and appropriate solution.

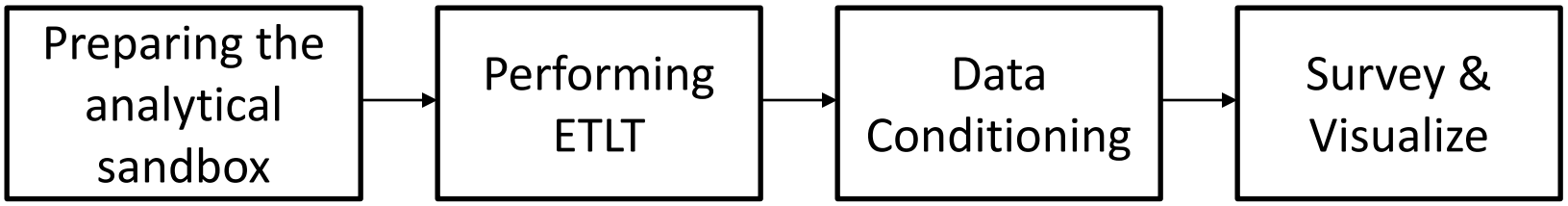
1.6 Developing Initial Hypotheses

1.7 Identifying Potential Data Sources

Phase 2. Data preparation

Phase 2. Data preparation

- In this phase, you require an analytical sandbox in which you can perform analytics for the entire duration of the project.
- You need to explore, preprocess and condition data prior to modeling.
- Further, you will perform ETLT (Extract, Transform, Load and Transform) to get data into the sandbox.



Phase 2. Data preparation

2.1 Preparing the Analytic Sandbox

2.2 Learning About the Data

2.3 Performing ETLT

2.4 Data Conditioning

2.5 Survey and Visualize

2.6 Common Tools for the Data Preparation Phase

Phase 2. Data preparation

2.3 Performing ETLT – Extract, Transform, Load, Transform

Extract – the raw, unprepared data from source application & database

Transform – applies to one data source at a time. Fast and simple because they transform each source independently, it relates to data formats, data cleansing and masking/removing sensitive data .

Load – the prepared data into the data warehouse

Transform – It relates to integrating multiple data sources and other transformers that apply to data from multiple sources at the same time.

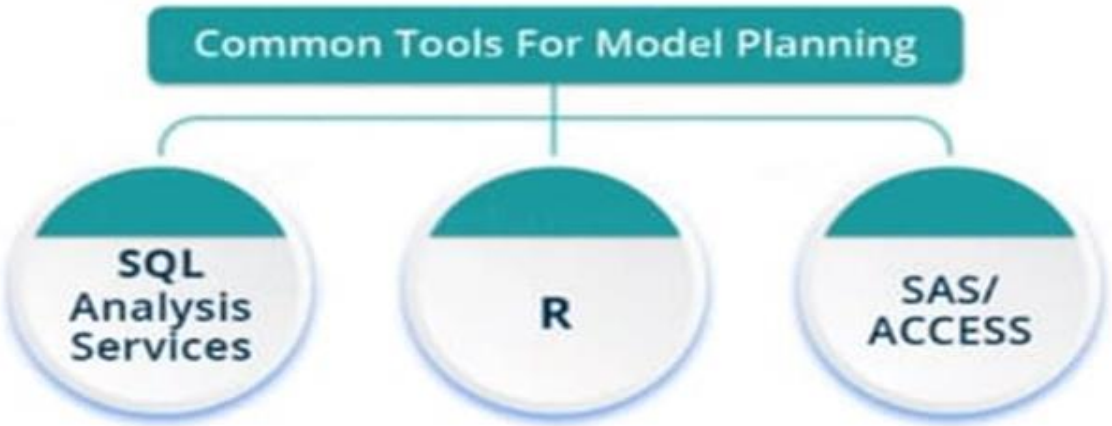
Phase 3. Model planning

Phase 3. Model planning

- Here, you will determine the methods and techniques to draw the relationships between variables.
- These relationships will set the base for the algorithms which you will implement in the next phase.
- You will apply Exploratory Data Analytics(EDA) using various statistical formulas and visualization tools.
  - EDA (Exploratory Data Analysis) :
    - Perform Statistical & visual analysis
    - Discover & handle errors/outliers
    - Shortlist predictive modelling technique

Phase 3. Model planning

- Data Exploration and Variable Selection
- Model Selection
- Common Tools for the Model Planning Phase



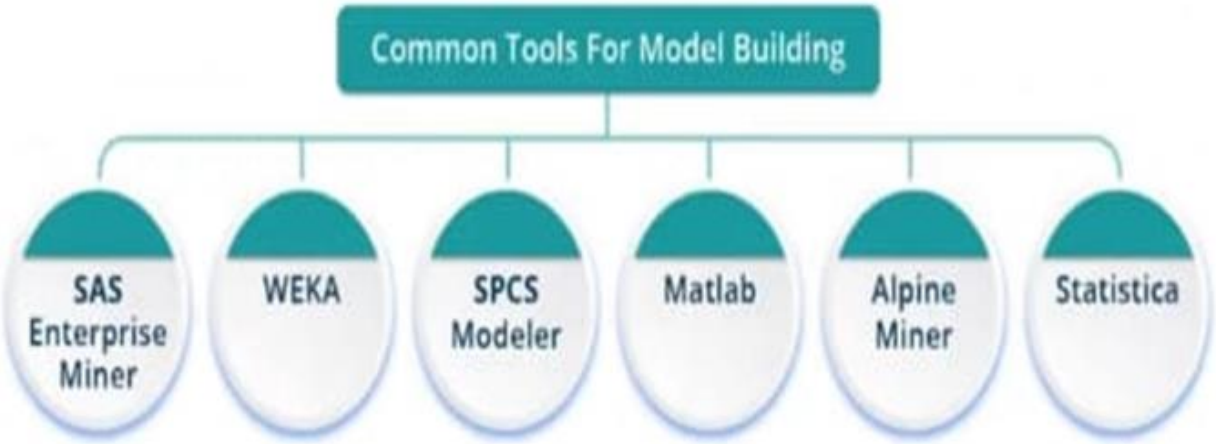
Phase 4. Model Building

Phase 4. Model Building

- In this phase you will develop datasets for training and testing purposes.
- Here, you need to consider whether your existing tools will suffice for running the models or they will need a more robust environment(like fast and parallel processing)
- You will analyze various learning techniques like classification, association & clustering to build the model.

Phase 4. Model Building

- Common Tools for the Model Building Phase



Phase 5. Communicate results

Phase 5. Communicate results

- Now it is important to evaluate if you have been able to achieve the goal that you had planned in the first phase.
- So, in the last phase, you identify all the key findings, communicate to the stakeholders, and determine if the results of the project are a success or a failure based on the criteria developed in Phase 1.

Phase 6. Operationalize

Phase 6. Operationalize

- In this phase, you deliver final reports, briefings, code and technical documents.
- In addition, sometimes a pilot project is also implemented in a real-time production environment.
- This will provide you with a clear picture of the performance and other related constraints on a small scale before full deployment.

Summary

- Phase-1 : Discovery
  - Clearly defined business problem
  - Set Success Criteria
  - Define clear data science objectives.

- Phase-2 : Data preparation
  - Understand data points and constraints
  - Formulate data analytics strategy
  - Perform required transformation

Summary

- Phase-3 : Model planning
  - Break Business problems to data science problems
  - Identify Mschine Learning Model

  - EDA (Exploratory Data Analysis) :
    - Perform Statistical & visual analysis
    - Discover & handle errors/outliers
    - Shortlist predictive modelling technique

Summary

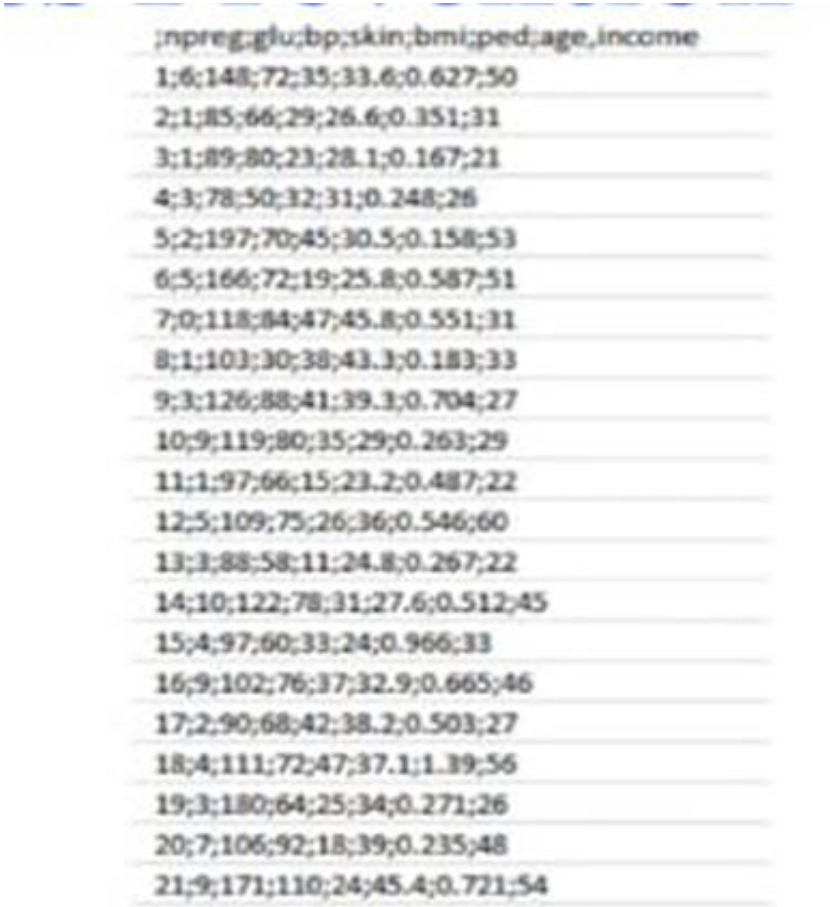
- Phase-4 : Model Building
  - Experiment with multiple models
  - Choose the most optimal model
  - Create a feedback loop
- Phase-5 : Communicate results
  - Break Business problems to data science problems
  - Identify Mschine Learning Model
- Phase-6 : Operationalize
  - Evaluation

Case Study : Diabetes Prevention

- What if we could predict the occurrence of diabetes and take appropriate measures beforehand to prevent it?

Case Study : Diabetes Prevention

- First, we will collect the data based on the medical history of the patient as discussed in phase-1



;npreg;glu;bp;skin;bmi;ped;age,income
1;6;148;72;35;33.6;0.627;50
2;1;85;66;29;26.6;0.351;31
3;1;89;80;23;28.1;0.167;21
4;3;78;50;32;31;0.248;26
5;2;197;70;45;30.5;0.158;53
6;5;166;72;19;25.8;0.587;51
7;0;118;84;47;45.8;0.551;31
8;1;103;30;38;43.3;0.183;33
9;3;126;88;41;39.3;0.704;27
10;9;119;80;35;29;0.263;29
11;1;97;66;15;23.2;0.487;22
12;5;109;75;26;36;0.546;60
13;3;88;58;11;24.8;0.267;22
14;10;122;78;31;27.6;0.512;45
15;4;97;60;33;24;0.966;33
16;9;102;76;37;32.9;0.665;46
17;2;90;68;42;38.2;0.503;27
18;4;111;72;47;37.1;1.39;56
19;3;180;64;25;34;0.271;26
20;7;106;92;18;39;0.235;48
21;9;171;110;24;45.4;0.721;54

Case Study : Diabetes Prevention

- **Attributes:**
- npreg – Number of times pregnant
- glucose – Plasma glucose concentration
- bp – Blood pressure
- skin – Triceps skinfold thickness
- bmi – Body mass index
- ped – Diabetes pedigree function
- age – Age
- income – Income

Case Study : Diabetes Prevention

- Step 2: Now, once we have the data, we need to clean and prepare the data for data analysis.
- This data has lot of inconsistencies like missing values, blank columns, abrupt values and incorrect data format which need to be cleaned.
- Here, we have organized the data into a single table under different attributes- making it look more structured.

Case Study : Diabetes Prevention

	npreg	glu	bp	skin	bmi	ped	age	income
1	6	148	72	35	33.6	0.627	50	
2	1	85	66	29	26.6	0.351	31	
3	1	89	6600	23	28.1	0.167	21	
4	3	78	50	32	31	0.248	26	
5	2	197	70	45	30.5	0.158	53	
6	5	166	72	19	25.8	0.587	51	
7	0	118	84	47	45.8	0.551	31	
8	one	103	30	38	43.3	0.183	33	
9	3	126	88	41	39.3	0.704	27	
10	9	119	80	35	29	0.263	29	
11	1	97	66	15	23.2	0.487	22	
12	5	109	75	26	36	0.546	60	
13	3	88	58	11	24.8	0.267	22	
14	10	122	78	31	27.6	0.512	45	
15	4		60	33	24	0.966	33	
16	9	102	76	37	32.9	0.665	46	
17	2	90	68	42	38.2	0.503	27	
18	4	111	72	47	37.1	1.39	56	
19	3	180	64	25	34	0.271	26	
20	7	106	92	18		0.235	48	
21	9	171	110	24	45.4	0.721	54	

Case Study : Diabetes Prevention

- This data has a lot of inconsistencies.
- In the column npreg, “one” is written in words, whereas it should be in numeric form like 1.
- In column BP one of the value is 6600 which is impossible (at least for humans) as bp cannot go upto such a huge value
- As you can see the income column is blank and also makes no sense in predicting diabetes therefore it is redundant to have it here and should be removed from the table.
- So we will clean and preprocess this data by removing the outliers, filling up the null values and normalizing the data type. If you remember this is our second phase which is data pre-processing.
- Finally, we get the clean data that can be used for the analysis.

Case Study : Diabetes Prevention

	npreg	glu	bp	skin	bmi	ped	age
1	6	148	72	35	33.6	0.627	50
2	1	85	66	29	26.6	0.351	31
3	1	89	80	23	28.1	0.167	21
4	3	78	50	32	31	0.248	26
5	2	197	70	45	30.5	0.158	53
6	5	166	72	19	25.8	0.587	51
7	0	118	84	47	45.8	0.551	31
8	1	103	30	38	43.3	0.183	33
9	3	126	88	41	39.3	0.704	27
10	9	119	80	35	29	0.263	29
11	1	97	66	15	23.2	0.487	22
12	5	109	75	26	36	0.546	60
13	3	88	58	11	24.8	0.267	22
14	10	122	78	31	27.6	0.512	45
15	4	97	60	33	24	0.966	33
16	9	102	76	37	32.9	0.665	46
17	2	90	68	42	38.2	0.503	27
18	4	111	72	47	37.1	1.39	56
19	3	180	64	25	34	0.271	26
20	7	106	92	18	39	0.235	48
21	9	171	110	24	45.4	0.721	54

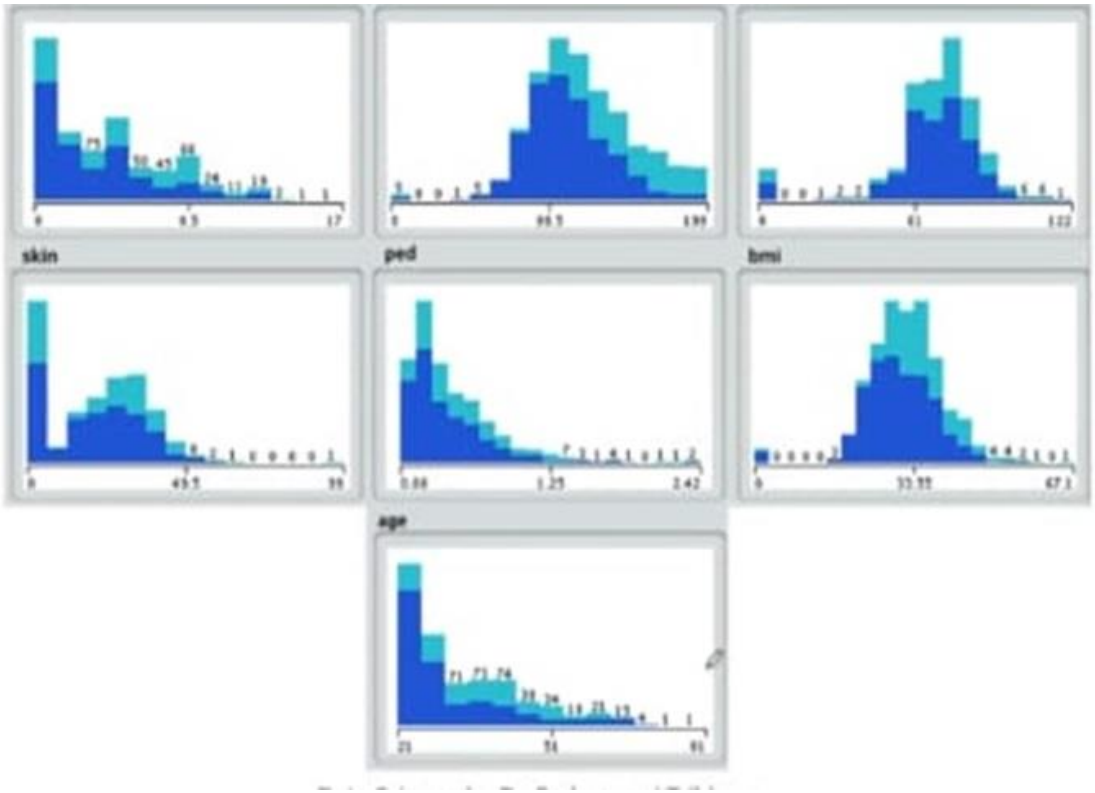
Case Study : Diabetes Prevention

- Step3: Now let's do some analysis as discussed earlier in phase3.
- First, we will load the data into the analytical sandbox and apply various statistical functions on it.

e.g. R has functions like describe which gives us the number of missing values and unique values. We can also use the summary function which gives us statistical information like mean, median, range, min or max values.

Then, we use visualization techniques like histograms, line graphs, box plots to get a fair idea for the distribution of data.

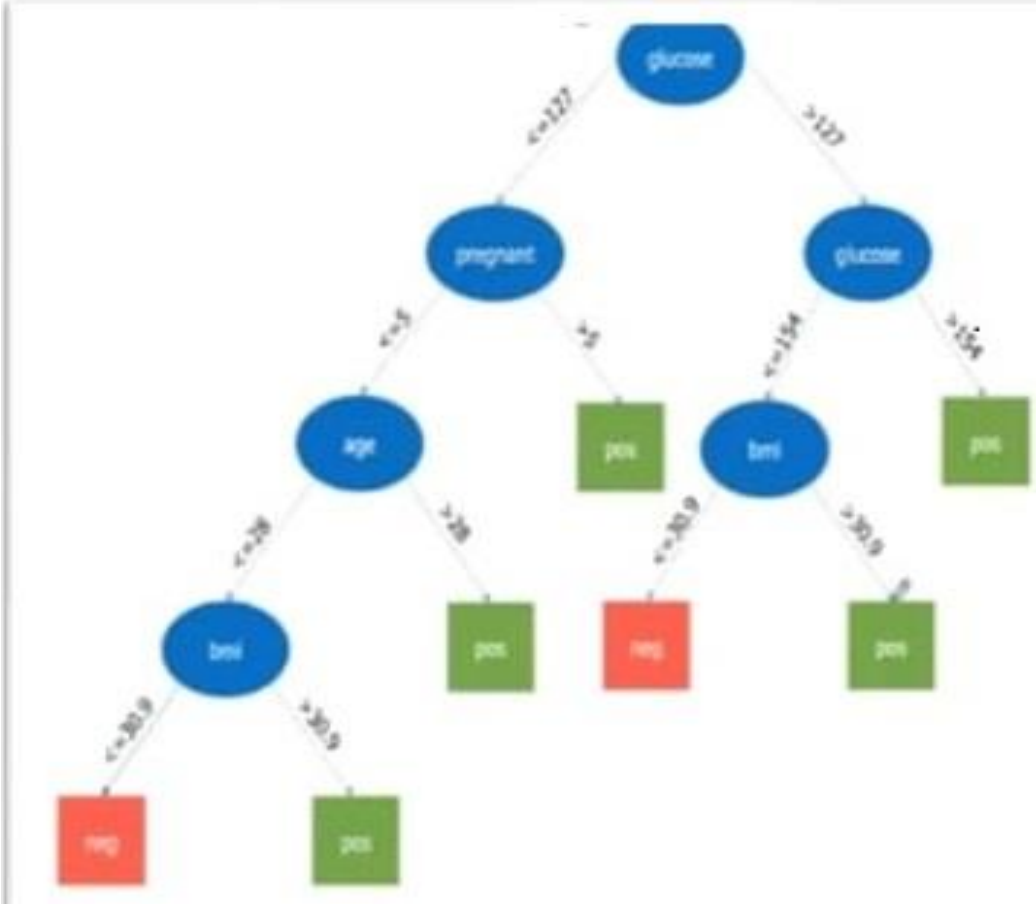
Case Study : Diabetes Prevention



Case Study : Diabetes Prevention

- Now based on insights derived from the previous step, the best fit for this kind of problem is the decision tree. Let's see how?
- Since, we have particularly used decision tree because it takes all attributes into consideration in one go, like the ones which have a linear relationship as well as those which have a nonlinear relationship. In our case, we have a linear relationship between npreg and age, whereas the nonlinear relationship between npreg and ped.
- Decision tree models are also robust as we can use the different combination of attributes to make various trees and then finally implement the one with maximum efficiency.

Case Study : Diabetes Prevention

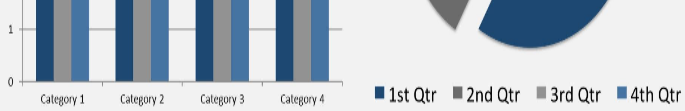


Here, the most important parameter is the level of glucose, so it is our root node. Now, the current node and its value determine the next important parameter to be taken. It goes on until we get the result in terms of *pos* or *neg*. Pos means the tendency of having diabetes is positive and neg means the tendency of having diabetes is negative.

Case Study : Diabetes Prevention

- Step 5 in this phase we will run a small pilot project to check if our results are appropriate. We will also look for performance constraints if any. If the results are not accurate then we need to replan and rebuild the model

- Step 6: once we have executed the project successfully, we will share the output for full deployment



FREQUENCY DISTRIBUTION TABLE	
Category	Frequency
Black	12
Brown	5
Blond	3
Red	7

Measures of Central Tendency

MEAN

The mean is equal to the sum of all the values in the data set divided by the number of values in the data set.

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

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

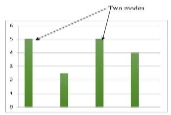
MEDIAN

It is the mid score of a dataset that has been arranged in order of magnitude. In order to calculate the median, suppose we have the following dataset.

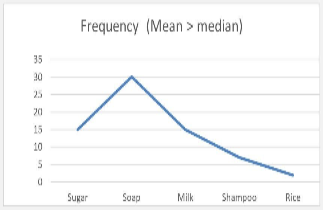
10 15 15 15 20 20 20 20 30 30

MODE

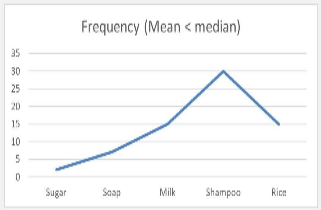
It is a value that most often score in our dataset. A dataset can have no mode, one mode or multiple modes.



Measure of Asymmetry



Right or Positive Skewness



Left or Negative Skewness

Variance and Standard Deviation

Introduction to Descriptive

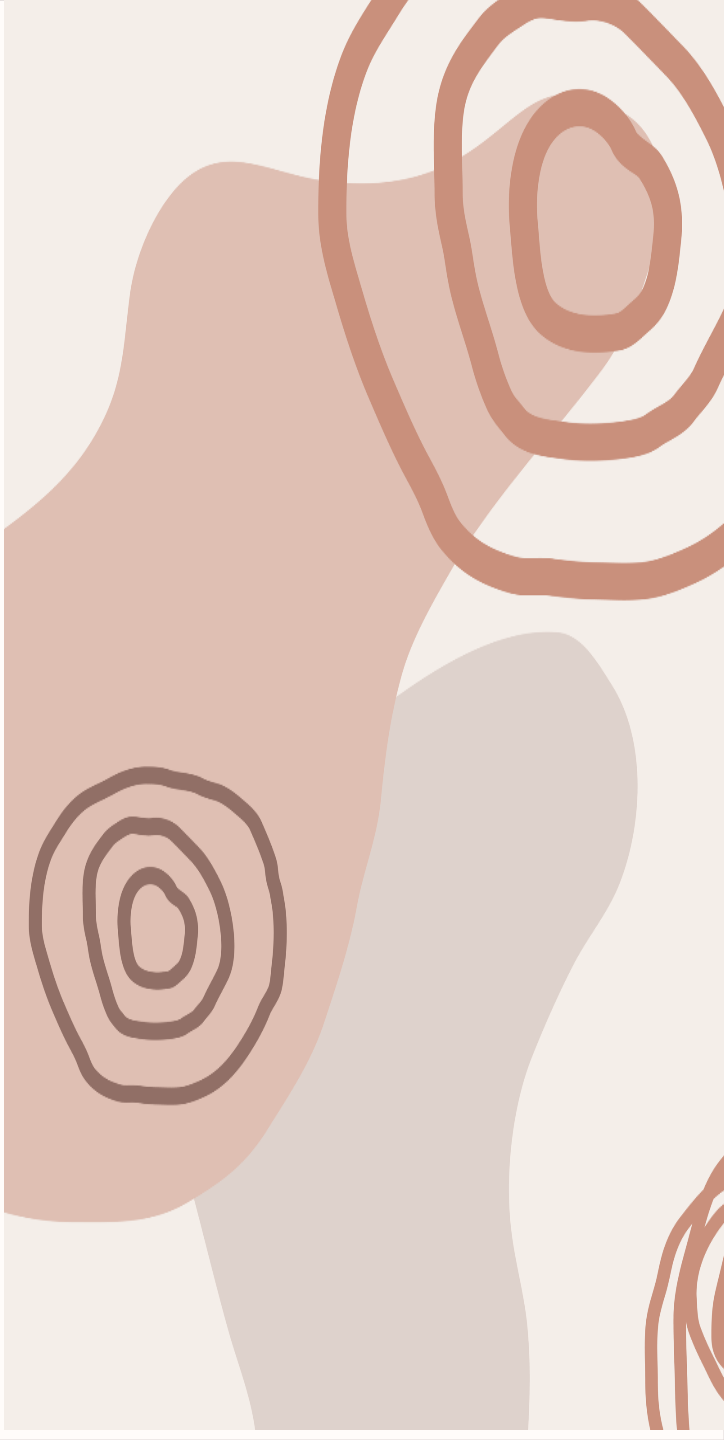
Welcome to this presentation on Descriptive Statistics. The purpose of this presentation is to provide you with a comprehensive understanding of this branch of statistics and how it is relevant to computer engineering

What is Descriptive Statistics?

•Descriptive statistics is the practice of summarizing and presenting data in a meaningful way.

•It involves utilizing different statistical measures and graphical representations to make sense of complex data sets.

• It is extremely useful in computer engineering as it allows us to analyze data, make decisions, and evaluate performance.



Why Descriptive Statistics

1 **Measure Performan**

Utilize measures of central tendency variability, such as the mean and standard deviation, analyze computer system

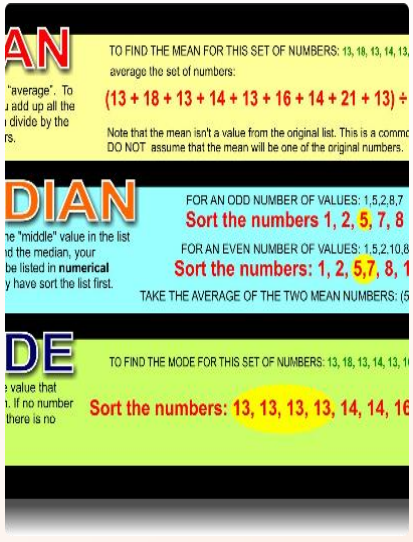
2 **Data Visualization**

Use histograms, scatter plots, and other types of graphs to display large amounts of data in a compact & understandable manner.

3 **Data Collection**

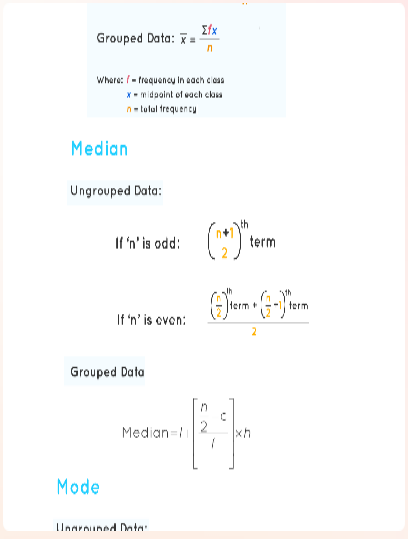
Understand how to choose the right sample and how to collect data that is representative a larger population.

Measures of Central Tendency



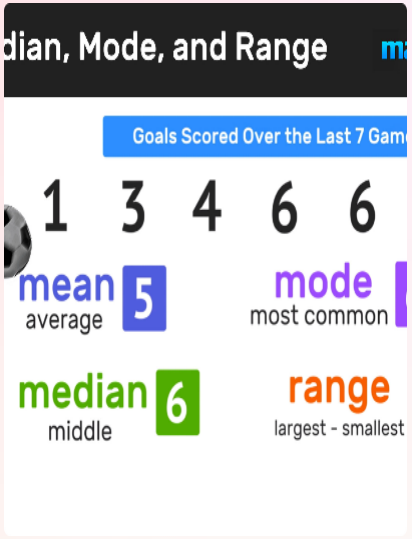
Mean

The most commonly used used measure of central tendency, calculated by adding all the values and dividing by the number of of observations.



Median

The middle value in a dataset when all values are sorted from smallest to largest. Useful for skewed data where the mean is not representative.



Mode

The value that occurs most frequently in a dataset. Useful for categorical data.

Measures of Variability

Range

The difference between the largest and smallest values in a dataset. Useful for providing idea of how spread out the data is.

Variance

The average of the squared differences from the mean. Useful for understanding the amount variation in a dataset.

Standard Deviation

The square root of the variance. Useful for understanding the distribution of data.

Why Descriptive Statistics

•Descriptive statistics is an essential tool in computer engineering as it helps us to understand the data generated in various projects at a deeper level.

•By analyzing the data statistically, we can identify trends, patterns, and relationships that may be otherwise difficult to notice.

•This helps us to make more informed research decisions which result in better solutions.

Why Use Descriptive Statistics in Computer Engineering?

- Descriptive statistics is an essential tool in computer engineering as it helps us to understand the data generated in various projects at a deeper level.
- By analyzing the data statistically, we can identify trends, patterns, and relationships that may be otherwise difficult to notice.
- This helps us to make more informed research decisions which result in better solutions.

Types of Statistical Data

1. Numerical (Discrete & Continuous)
2. Categorical
3. Ordinal

- Discrete data represents items that can be counted. It is also called digital or binary data.
- Continuous data types are called Analog Data in information technology. Continuous data represent measurements; their possible values cannot be counted and can only be described using intervals on the real number line.

Types of Statistical Data

1. Numerical (Discrete & Continuous)
2. Categorical
3. Ordinal

Categorical data is qualitative and classifies variables based on groups or categories, for example, gender. Numerical data is quantitative and classifies variables based on numbers, such as age or temperature. Ordinal data mixes numerical and categorical data.

Categorical or Qualitative data examples

Gender, occupation, product type

Numerical data examples

Age, height, income temperature

Ordinal data examples

Rating a restaurant
On scale of 0 to 4

Categorical and Numerical Data

Data can be categorized as either categorical or numerical. Categorical data is qualitative and classifies variables based on groups or categories, for example, gender. Numerical data is quantitative and classifies variables based on numbers, such as age or temperature.

Categorical data examples

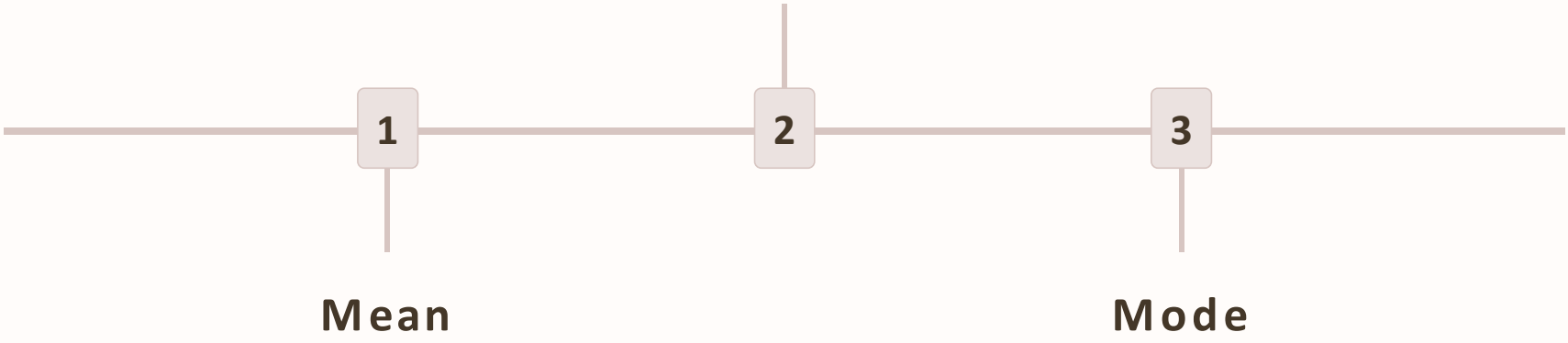
Gender, occupation, product type

Numerical data examples

Age, height, temperature, income

Measures of Central Tendency

Measures of central tendency are used to identify the "center" or average value of a of a dataset. There are three common measures of central tendency- mean, median, median, and mode.



Mean

The mean (or average) is the most popular and well known measure of central tendency.
It can be used in discrete and continuous data, most often with continuous data.
The mean is equal to the sum of all the values in the data set divided by number of values in the data set.
The formula for calculating the mean:

For a sample of data: $\bar{x} = (\sum x) / n$

For a population of data: $\mu = (\sum x) / N$

Where: $\bar{x}$ is the sample mean,
μ is the population mean,
$\sum x$ represents the sum of all individual data points in the dataset,
n is the number of data points in the sample,
N is the number of data points in the population.

Mean : Disadvantages

1. Sensitive to outliers: The mean is highly sensitive to extreme values, known as outliers, in the dataset. Even a single outlier can significantly affect the mean, pulling it away from the central tendency and potentially misrepresenting the average value of the majority of the data.

2. Not appropriate for skewed distributions: In datasets with skewed distributions (where the data is not symmetrical), the mean may not accurately represent the typical value.

3. Not suitable for ordinal data & Categorical data: Attempting to calculate the mean for categories like colors, genders, or regions would not provide meaningful results.

4. Not robust to errors or missing data: An error or missing value can distort the mean, data imputation techniques may be necessary to handle missing data before calculating the mean.

Median

- The median is the middle score for a set of data that has been arranged in of magnitude.
- The median is less affected by outliers and skewed data.
- For an Odd number of scores, take middle scores.
- For the Even number of scores, take the middle two scores and average the result.

e.g. Let's consider a dataset of exam scores: {85, 90, 78, 92, 88, 87}
Step 1: Sort the dataset: {78, 85, 87, 88, 90, 92}
Step 2: Since the number of data points (n) is 6 (even), we take the two middle values, which are 87 and 88.
Step 3: Calculate the median: Median = (87 + 88) / 2 = 87.5

Mean vs Median

- The median may be a better indicator if a set of scores has an Outlier.
- An outlier is an extreme value that differs greatly from other values. values.

- When the sample size is large and does not include outliers, the mean score usually provides a better measure of central tendency. tendency.

Mode

- Mode is the most frequent score in our data set.
- On a histogram it represents the highest bar in a bar chart or histogram.

Type of Variable	**Best Measure of Central Tendency**
Nominal	Mode
Ordinal	Median
Interval/Ratio(not skewed)	Mean
Internal/Ratio(skewed)	Median

Measures of Dispersion

Measures of dispersion characterize how spread-out the data is. There are three common measures of range, variance, and standard deviation.

1

Range

The difference between the maximum and minimum values in a dataset.

2

Variance

An average of the squared differences from the mean.

3

Standard Deviation

The square root of the variance.

Variance

In a population, variance is the average squared deviation from the population mean.

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
$\sigma^2 =$ population variance	$s^2 =$ sample variance
$x_i =$ value of i^{th} element	$x_i =$ value of i^{th} element
$\mu =$ population mean	$\bar{x} =$ sample mean
$N =$ population size	$n =$ sample size

Standard Deviation

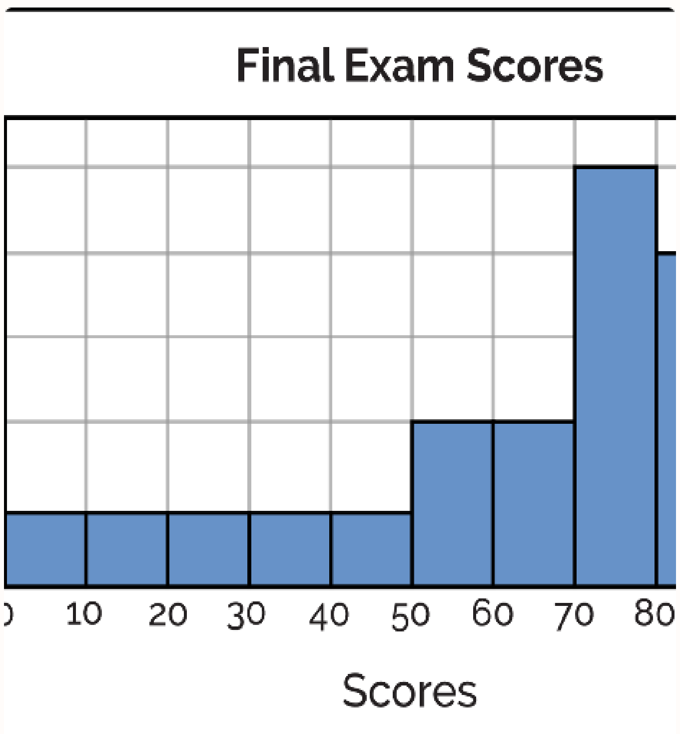
The standard deviation is the square root of the variance.
e.g.

Population	Sample
$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}$
X - The Value in the data distribution μ - The population Mean N - Total Number of Observations	X - The Value in the data distribution $\bar{x}$ - The Sample Mean n - Total Number of Observations

5◇λ◇ ë¹α↯Ώ~◇λ¹NZÇ≡◦ÞZ¹ά↯≡α

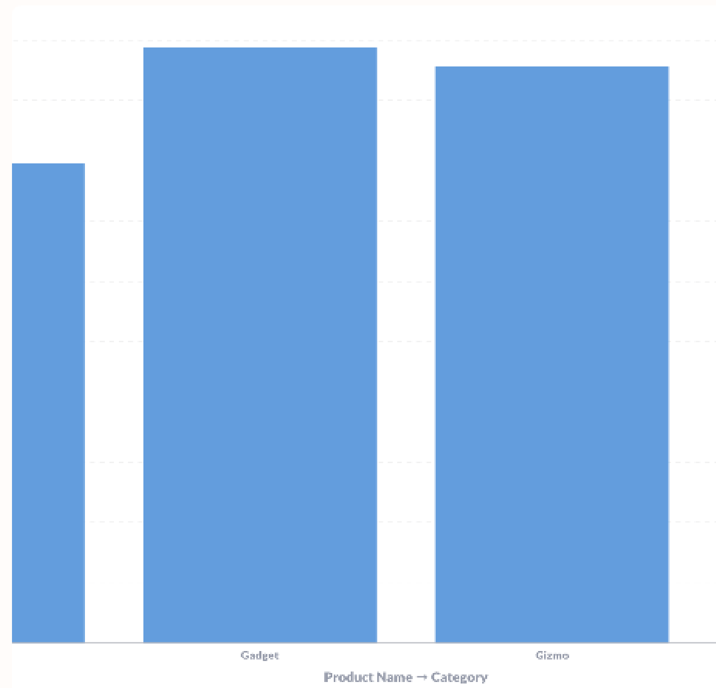
Data visualization techniques help to communicate information visually, making it easier to understand the data. There are four common visualization techniques- histograms, bar charts, scatter plots, and box plots.

Histograms



Graphical representations of the distribution of numerical data.

Bar Charts

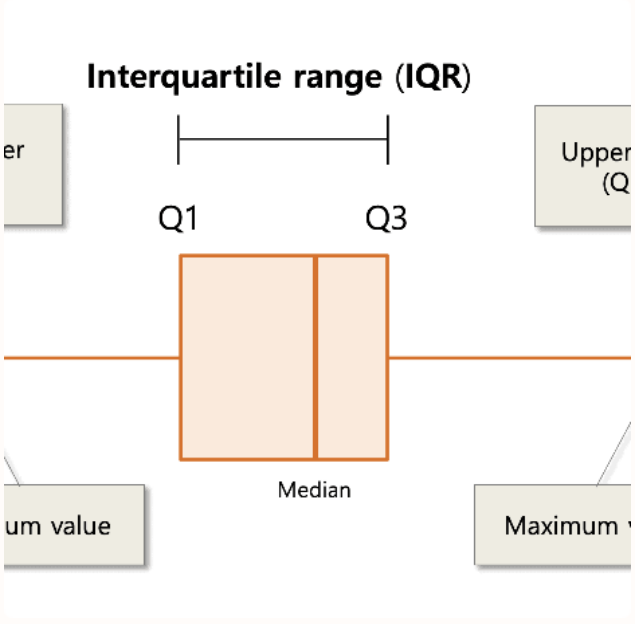


Visualizations for comparing categorical data.

Data Visualization Techniques

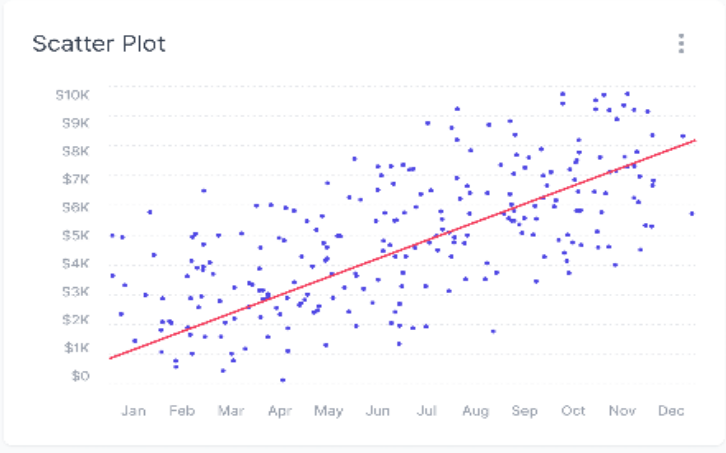
Data visualization techniques help to communicate information visually, making it easier to understand the data. There are four common visualization techniques- histograms, bar charts, scatter plots, and box plots.

Box Plots

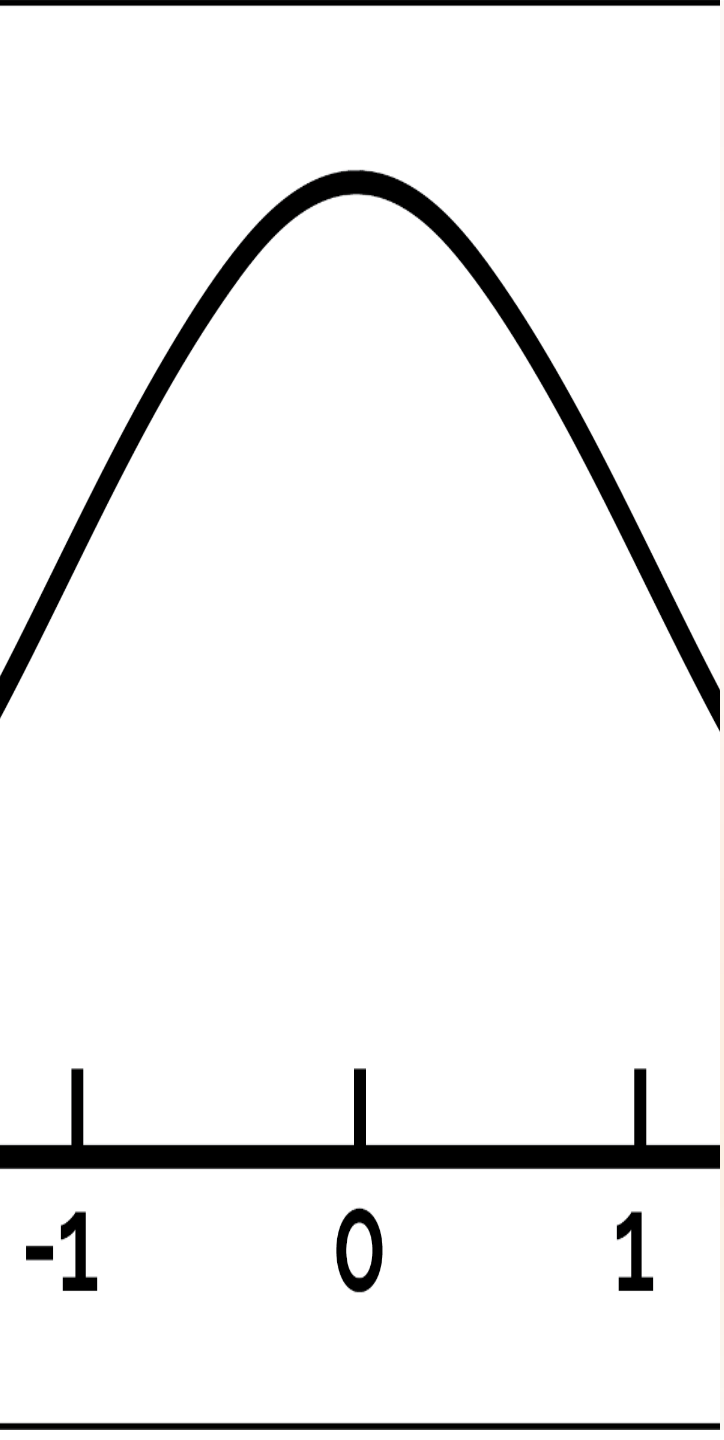


Visualizations for displaying the distribution and outliers in numerical data.

Scatter Plots



Graphical representations for analyzing relationships between two numerical variables.



Normal Distribution

Many natural phenomena and large-scale systems follow a normal distribution, also known as a bell In computer engineering, this can be seen in things like overall system response times, network throughput, disk access speeds. Understanding normal distribution allows engineers predict and optimize these systems for better performance.

The Mean: Real-Time Application

1

5=÷-1Z9\INZ

The arithmetic mean is the sum of all data points divided by the number of data points.

2

Real-Time Example: Finance

The mean can be used to calculate the average daily stock a given period, providing insights into the stock's overall performance.

The Median: Real-Time Application

Real-Time Example: Healthcare

The median can be used to determine the determine the median patient age, providing a more representative measure measure of the population's age distribution.

1

2

3

Definition

Ç≈Î ˙ ÎÁ¬ć—¬7≈Î ˙ —ÁÁ?Î]—?Ŕ —ć—
Á—7—Î7 ^≈Îć—7-]—?Ŕ- —)Î —))—ć ÎÁ—ć
—-€Ĉć Á—ć ‹)ÁÎ-€Ĉć Á—ć ‹)ÁÎ)ت

Real-Time Example: Education

The median can be applied to determine a student's better, especially when the contains outliers.

The Mode: Real-Time Application

5≒ 1Z9\NZ

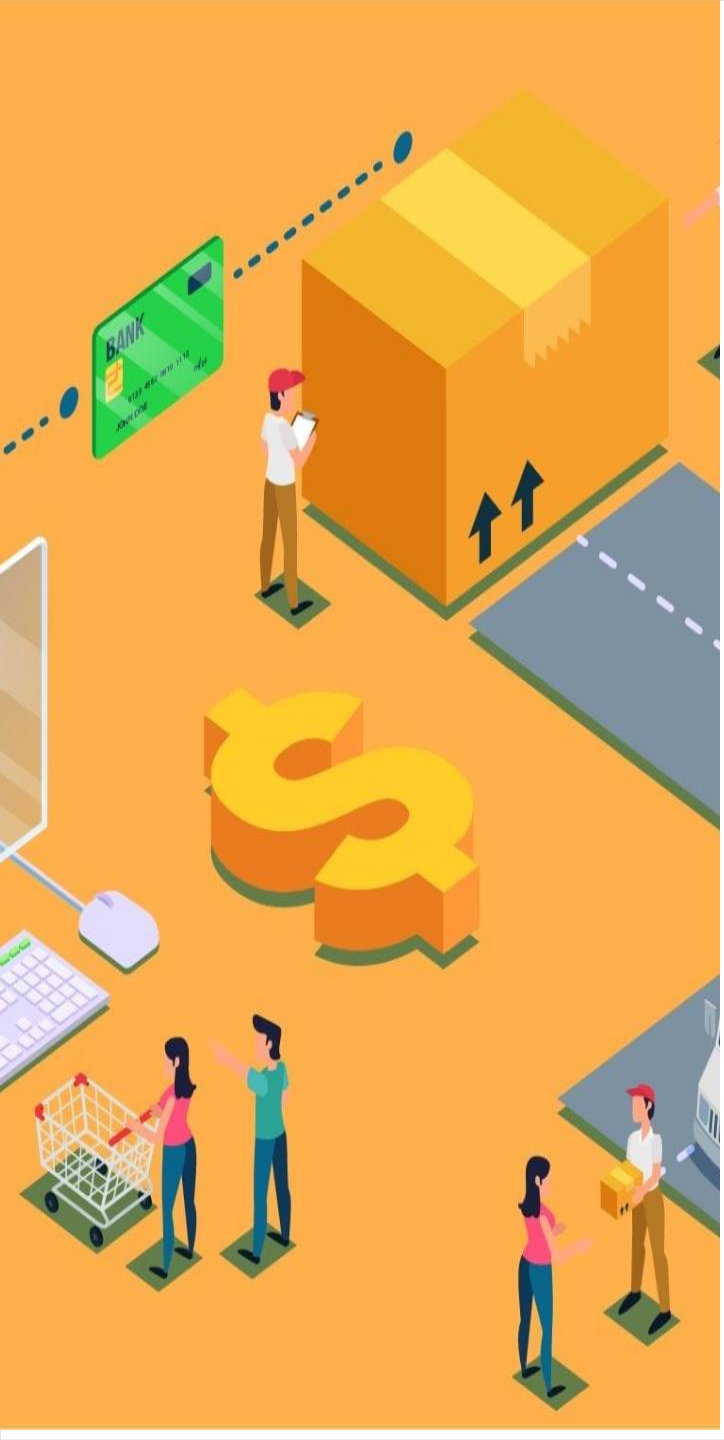
The mode is the value that appears most frequently in a dataset. dataset.

Real-Time Example: Marketing Analysis

The mode can be identify the most product category, businesses focus their marketing efforts on high-demand items.

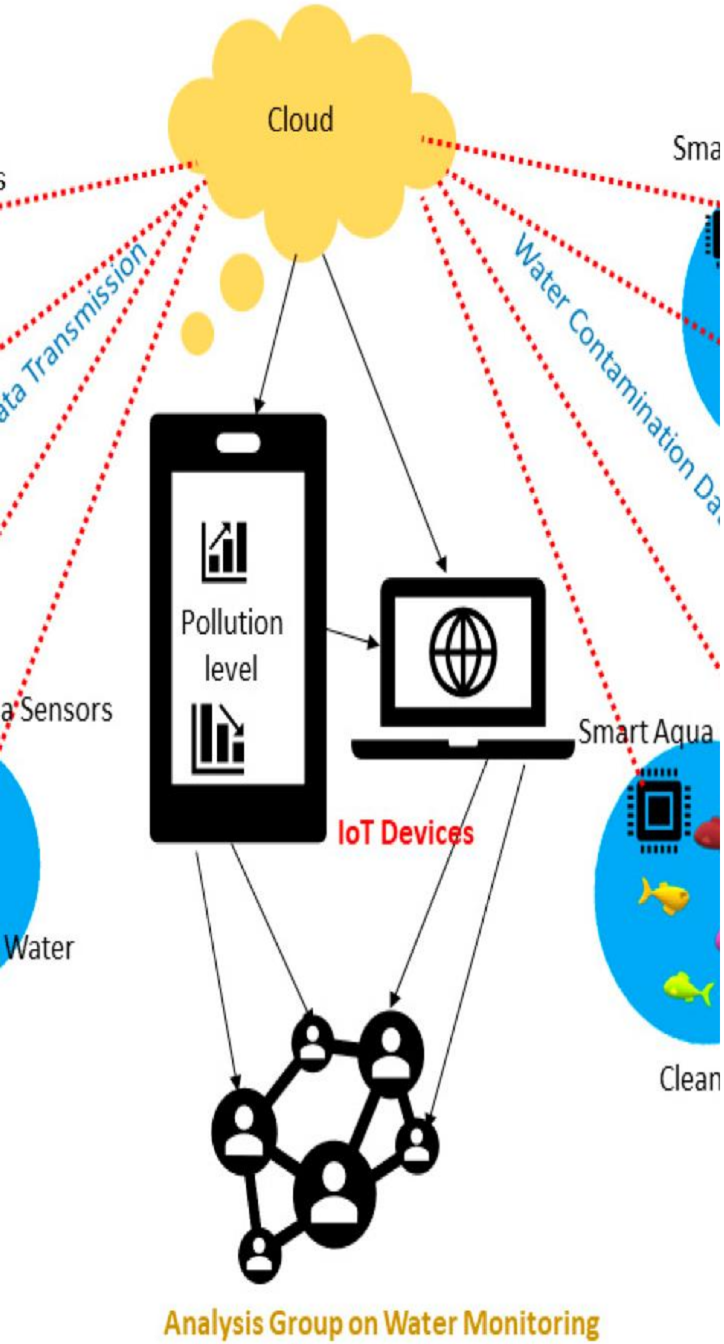
Real-Time Example: Public Public Health

The mode can be identify the most prevalent diseases in population or region.



Real-Time Use Case: Customer Behavior Analysis Analysis

Measures of central tendency play a vital role in in understanding customer behavior in e-commerce. commerce. By analyzing the mean, median, and and mode of various metrics such as purchase amount, browsing time, or customer ratings, businesses can gain insights into customer preferences, identify key trends, and optimize their their marketing strategies accordingly.



w≐ΏϘξΕ≐ Üα≐/◇α≑ণ{≐ZαŃ 5◇λ◇ ! ZΏ]αα

In environmental monitoring, sensors gather data on data on various parameters such as temperature, temperature, humidity, and air quality. Measures of Measures of central tendency, such as the mean, mean, median, and mode, can be employed to to detect anomalies, understand typical values, and values, and make informed decisions based on the on the sensor readings.

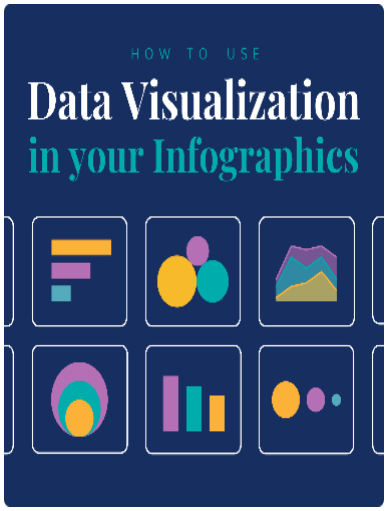
Real-Time Use Case: Performance Monitoring

Measures of central tendency are utilized to monitor system performance in software software engineering. By calculating the mean response time, median latency, or mode of mode of resource utilization, engineers can identify performance bottlenecks, optimize optimize system parameters, and ensure efficient operation of computer systems and and networks.

How to Choose the Right Measure of Central Tendenc

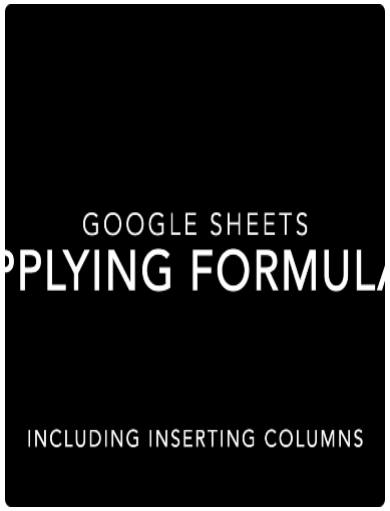
Define Your Research Objective

Identify what you to learn from the



Visualize the Data

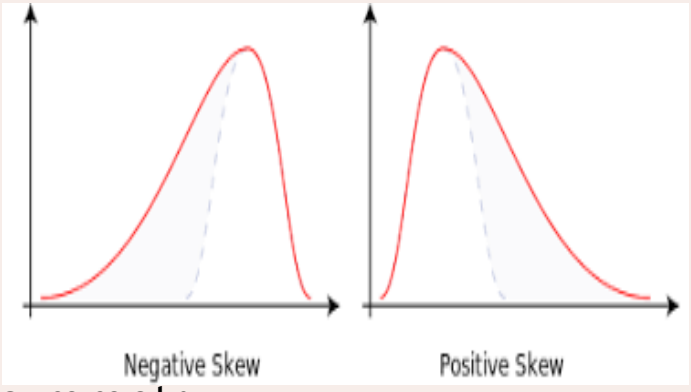
Check if the data follows a normal distribution or has outliers. Use histograms, box plots, or scatter plots to help visualize the data



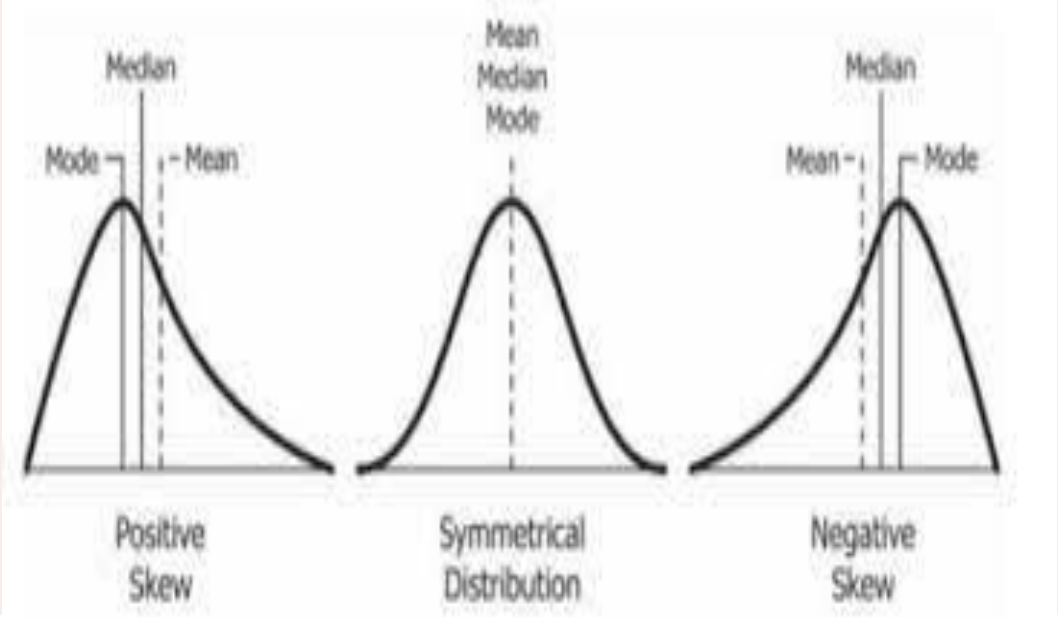
Apply the Formula

Select the measure of central tendency tendency that best matches your research research objective and data distribution. distribution. Calculate the mean, median, median, or mode according to their respective formulas.

Skewness



- Skewness is a measure of symmetry or lack of symmetry.
- A distribution or data set, is symmetric if it looks the same to the left and the center point.
- Skewness tells us about the direction of variation of the data set.



Importance of Skewness

1

Positive Skewness Skewed)

[illegible] analyse.

2

Negative Skewness (Left (Left Skewed)

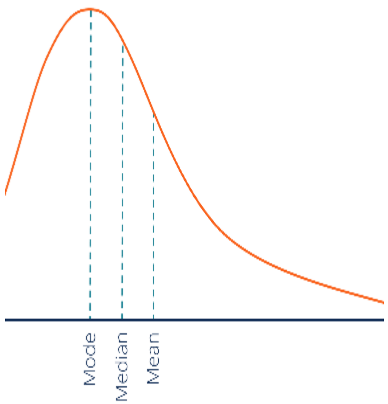
In a negatively distribution, most of the data values are clustered on the side of the Negative skewness affect the mean, making it less than median. Be sure to this into account interpreting results.

3

Symmetric (Zero Skewness)

In a symmetric distribution, the evenly distributed around the central point with the tail extending equally both sides. The median, and mode approximately equal a symmetric distribution.

Positive Skewness: Concentration on the Right



Dέ◇ÏÞ¹◦ʻΩw=Ïέ=α=ZΛ◇λΊΝΖ

In a positively skewed distribution, the tail of the data points extends more towards the right side. An example could be the income distribution, where a few individuals have significantly higher earnings.

Real-World Example

In a real-world scenario, a business may have a few high-end customers who generate a significant portion of the company's revenue, leading to a positively skewed sales distribution.

Application in Business

Positive skewness helps a business identify its highest value customers and focus its resources on serving them better, improving customer retention and long-term profitability.

Negative Skewness: Concentration on the Left the Left

1 Definition

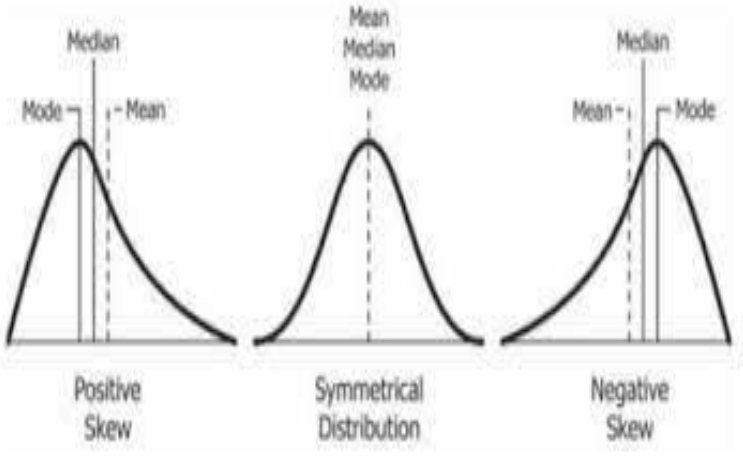
A negatively skewed distribution has a tail that extends more towards the left side. An example could be the age distribution of retirement, where most people retire at a similar age, but a few individuals retire at younger ages.

2 Impact on Data Models

Understanding negative skewness is essential for building accurate data science models models and predictions. For For example, skewed data may require normalization or transformation n to correct the the distribution. distribution.

Practical Applications

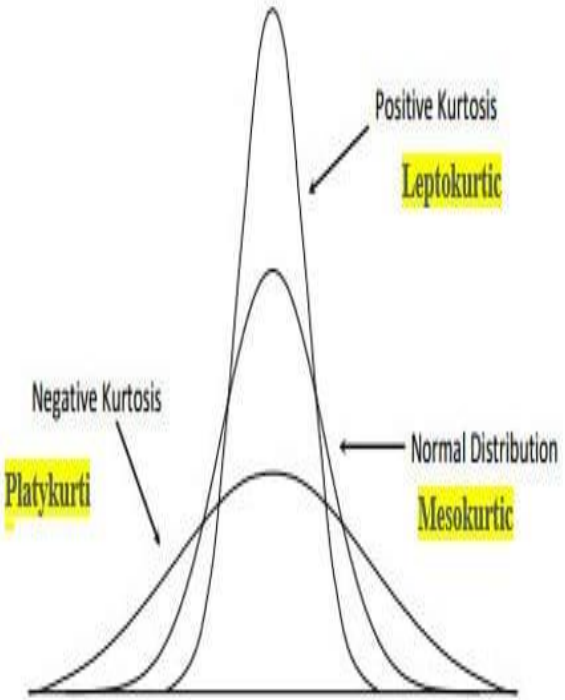
Companies can use negative skewness analysis to identify areas of employee skill deficiency and overall through training developme programs.



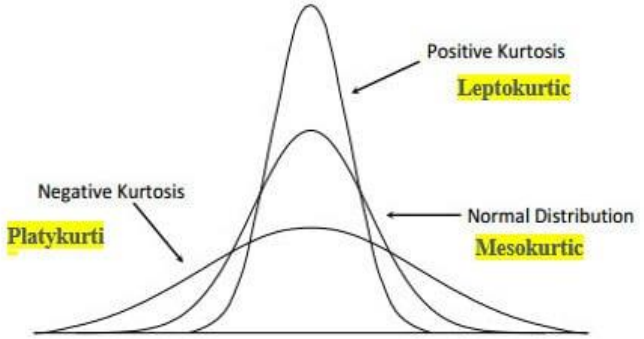
Limitations : Negative skew can make it challenging to determine the optimal threshold for decision making. For example, in loan defaults, a narrow range of data near the left tail could lead to errors in predictions.

Kurtosis

- Kurtosis is a parameter that describes the shape of a random variable's probability distribution whereas skewness measures the lack of symmetry of the frequency curve.
- Kurtosis refers to the degree of presence of outliers (extreme values) in the distribution.
- Various frequency curves can be divided into 3 categories depending on the shape of their peak.



The Relevance of Kurtosis in Data Analysis



Leptokurtic (Positive Kurtosis)

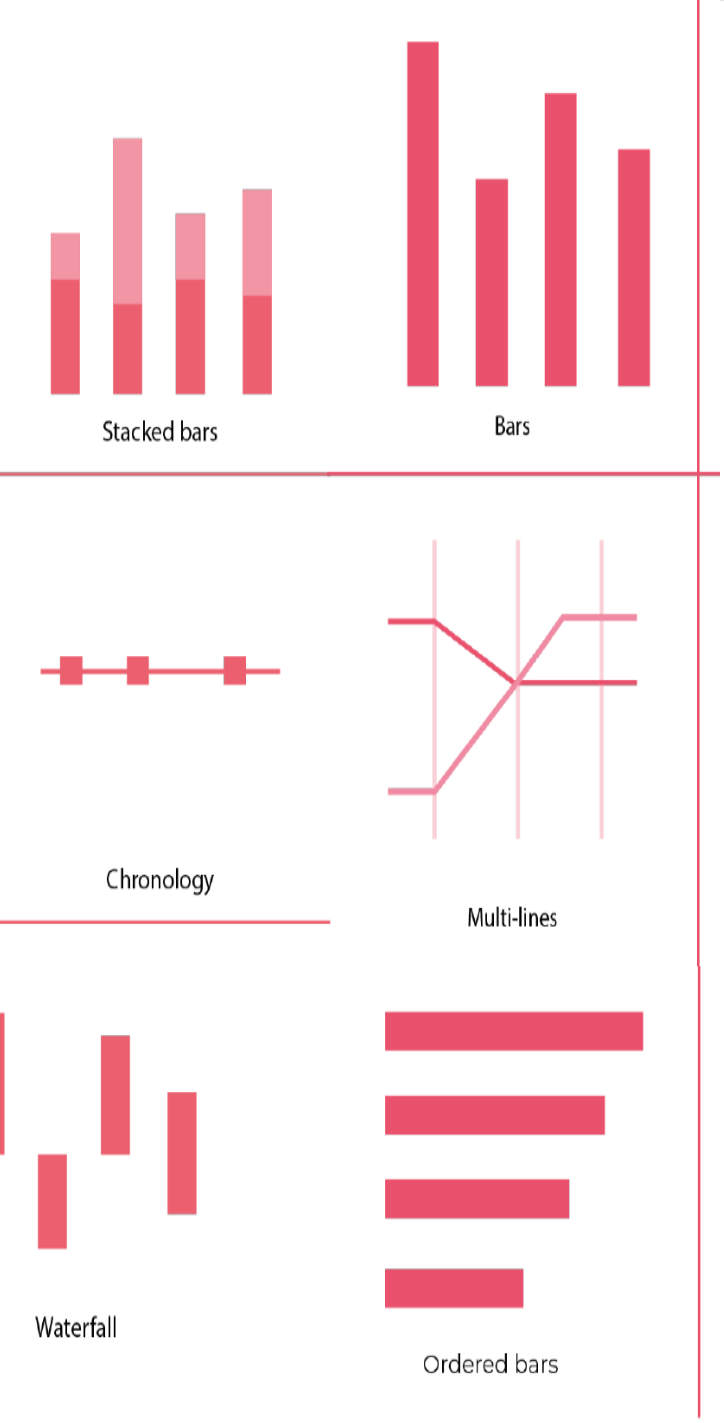
The peak of the distribution is higher and sharper than the normal distribution.
This indicates that the dataset has heavy tails, and extreme values (outliers) are more likely to occur.

Mesokurtic (Normal Kurtosis)

In a mesokurtic distribution, the peak of the distribution is similar to the normal distribution, and the tails have a similar shape.

Platykurtic (Negative Kurtosis)

The peak of the distribution is flatter than the normal distribution. This suggests that the dataset has lighter tails, and extreme values are less likely to occur.



Summary and Conclusion of Measures of central tendency

- That help us better understand data behavior and and make data-driven decisions.
- Real-time applications of measures of central tendency include finance, healthcare, marketing marketing analysis, environmental monitoring, e-monitoring, e-commerce, and performance monitoring.
- When choosing the right measure of central tendency, consider your research objective, visualize visualize the data, and apply the formula that matches your data distribution.

Probability Distribution

Probability distributions

- Probability distributions describe how the values of a random variable are distributed.

- Understanding these distributions is essential for making predictions, and performing statistical analyses.

Random Variables

- A random variable is a variable that can take on different values, each with an associated probability, depending on the outcome of a random event.
- There are two main types of random variables:
- **Discrete Random Variables**
- **Continuous Random Variables**

Discrete Random Variables

- A discrete random variable can take on a finite or countably infinite number of values.
- Each value has an associated probability.
- **Example:**
- The number of heads obtained when flipping a coin three times is a discrete random variable. The possible values are 0, 1, 2, and 3.

Continuous Random Variables

- A continuous random variable can take on any value within a given range. The probabilities are described using a probability density function (PDF).
- **Example:**
- The height of adult men in a population is a continuous random variable. It can take on any value within a certain range, say from 150 cm to 200 cm.

Types of Probability Distributions

- **Discrete Probability Distributions**
- **Continuous Probability Distributions**

1. Discrete Probability Distributions

- A discrete probability distribution describes the probability of occurrence of each value of a discrete random variable.
- Eg: Binomial Distribution, Poisson Distribution

Binomial Distribution

- The binomial distribution models the number of successes in a fixed number of independent Bernoulli trials with the same probability of success.

Formula:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where:

- n is the number of trials,
- k is the number of successes,
- p is the probability of success in each trial,
- $\binom{n}{k}$ is the binomial coefficient.

Binomial Distribution

- Suppose we flip a fair coin 10 times. The probability of getting heads (success) in each flip is 0.5. We want to find the probability of getting exactly 4 heads.

$$P(X = 4) = \binom{10}{4}(0.5)^4(0.5)^6 = \frac{10!}{4!6!}(0.5)^{10} \approx 0.205$$

Q. suppose there are 12 multiple choice questions in an English class quiz. Each question has 5 possible answers, and only 1 of them is correct. Find the probability of having 4 or less correct answer if a student attempts to answer every question at random?

- Ans:
- The number of trials (n) is 12 (since there are 12 questions).
- The probability of success (p) is 1/5 or 0.2 (since each question has 5 possible answers and only 1 correct answer).
- We need to find the probability of having k correct answers for k=0,1,2,3, and 4.

$$P(X \leq 4) = \sum_{k=0}^{4} \binom{12}{k} (0.2)^k (0.8)^{12-k}$$

We can calculate this step by step for each k:

1. For $k = 0$:
   $$P(X = 0) = \binom{12}{0} (0.2)^0 (0.8)^{12} = 1 \cdot 1 \cdot (0.8)^{12} \approx 0.0687$$
2. For $k = 1$:
   $$P(X = 1) = \binom{12}{1} (0.2)^1 (0.8)^{11} = 12 \cdot 0.2 \cdot (0.8)^{11} \approx 0.2061$$
3. For $k = 2$:
   $$P(X = 2) = \binom{12}{2} (0.2)^2 (0.8)^{10} = \frac{12 \cdot 11}{2 \cdot 1} \cdot 0.04 \cdot (0.8)^{10} \approx 0.2835$$
4. For $k = 3$:
   $$P(X = 3) = \binom{12}{3} (0.2)^3 (0.8)^9 = \frac{12 \cdot 11 \cdot 10}{3 \cdot 2 \cdot 1} \cdot 0.008 \cdot (0.8)^9 \approx 0.2362$$
5. For $k = 4$:
   $$P(X = 4) = \binom{12}{4} (0.2)^4 (0.8)^8 = \frac{12 \cdot 11 \cdot 10 \cdot 9}{4 \cdot 3 \cdot 2 \cdot 1} \cdot 0.0016 \cdot (0.8)^8 \approx 0.1304$$

Now, summing these probabilities:

$$P(X \leq 4) = 0.0687 + 0.2061 + 0.2835 + 0.2362 + 0.1304 \approx 0.9249$$

So, the probability of having 4 or fewer correct answers is approximately 0.9249, or 92.49%.

Solution in R

- >dbion(0, size=12, prob=0.2) +
  dbion(1, size=12, prob=0.2) +
  dbion(2, size=12, prob=0.2) +
  dbion(3, size=12, prob=0.2) +
  dbion(4, size=12, prob=0.2) ➔0.9274

Or (Cummulative probability)

>pnorm(4, size=12, prob=0.2) ➔0.9274 or 92.7%

Poisson Distribution

- The Poisson distribution models the number of events occurring in a fixed interval of time or space when these events happen with a known constant mean rate and independently of the time since the last event.

Formula:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where:

- λ is the average number of events in the interval,
- k is the number of events,
- e is Euler's number (approximately 2.71828).

- Suppose a call center receives an average of 5 calls per hour. We want to find the probability that exactly 3 calls will be received in the next hour.

$$P(X = 3) = \frac{5^3 e^{-5}}{3!} = \frac{125e^{-5}}{6} \approx 0.140$$

- Q. If there are 12 cars crossing a bridge per minute on average, find the probability of having seventeen or more cars crossing the bridge in a particular minute.
- Ans: ppois(16, lambda=12) ➔ 0.89871 (upper tail)
- >ppois(16, lambda=12,lower=FALSE) = 0.1012
- The probability of having seventeen or more cars crossing the bridge in a particular minute is approximately 0.101 or 10.1%.

Binomial Vs Poisson Distribution

- **Binomial Distribution**:
- Fixed number of trials (n).
- Each trial has two possible outcomes.
- Constant probability of success (p).
- Use when you are counting the number of successes in a fixed number of trials.
- **Poisson Distribution**:
- Events occur independently.
- Constant average rate (λ-lambda).
- No fixed number of trials, but rather events occurring over a continuous interval (time, area, etc.).
- Use when you are counting the number of events in a fixed interval.

Continuous Probability Distributions

- A continuous probability distribution describes the probabilities of the possible values of a continuous random variable.
- **Example: Normal Distribution**
- The normal distribution, also known as the Gaussian distribution, is characterized by its bell-shaped curve. It is defined by two parameters: the mean (μ) and the standard deviation (σ - sigma).