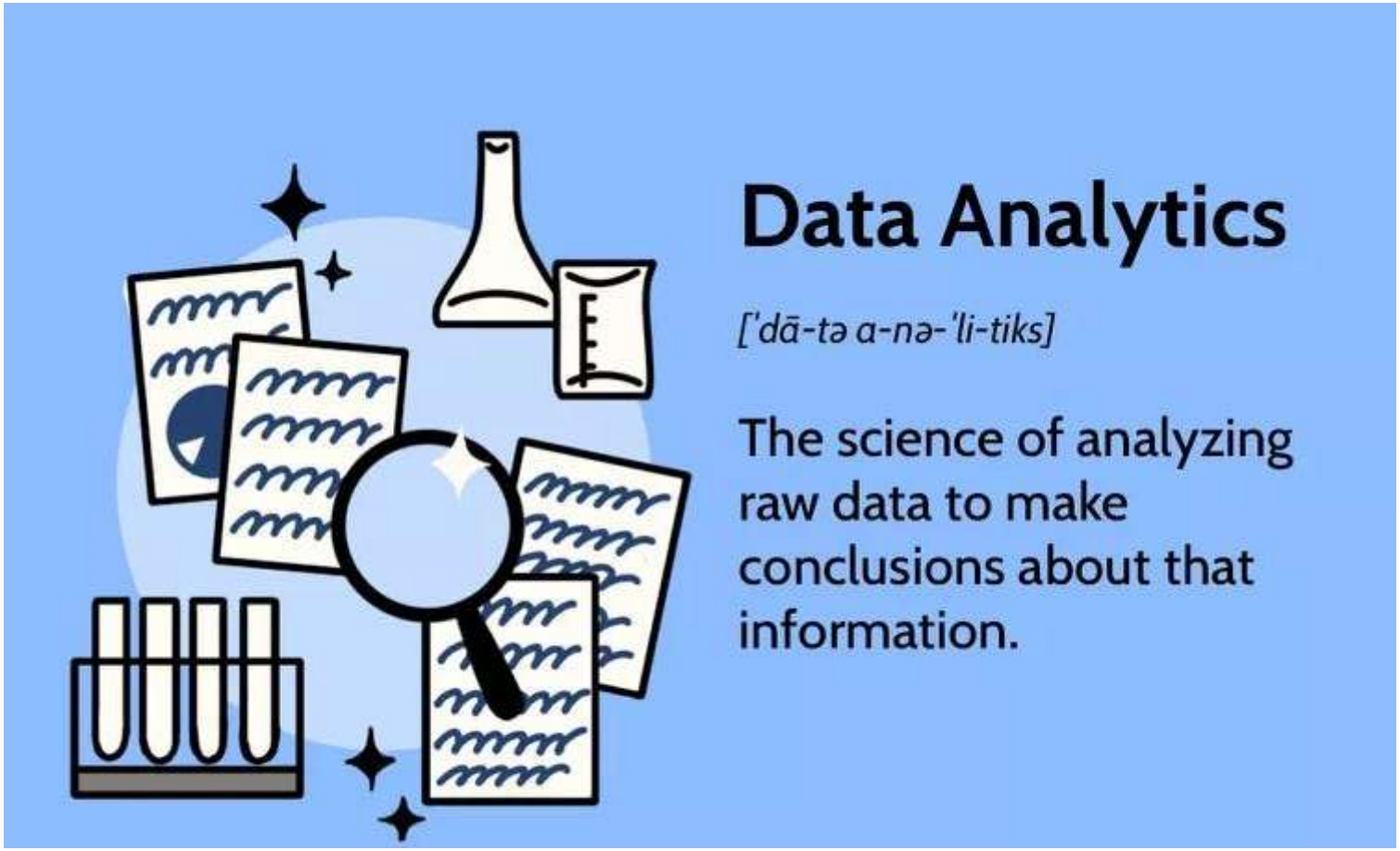
2.Data Analytics

Agenda

- Data Analytics Overview,
- Importance of Data Analytics,
- Types of Data Analytics,
- Descriptive Analytics,
- Diagnostic Analytics,
- Predictive Analytics,
- Prescriptive Analytics,
- Benefits of Data Analytics,
- and Data Visualization Techniques.

- Self-Learning Topics: Case Studies of Data analytics.

Introduction to Data Analysis.



What Is Data Analytics?

- Data analytics **is the science of analyzing** raw data to make conclusions about information.
- Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.
- Data analytics is a broad term that encompasses many diverse types of data analysis.
- Data analytics techniques can reveal trends and **metrics** that would otherwise be lost in the mass of information.

- This information can then be used to optimize processes to increase the overall efficiency of a business or system.

Data analytics: Important

- Data analytics is important because it helps businesses optimize their performances. Implementing it into the business model means companies can help reduce costs by identifying more efficient ways of doing business and by storing large amounts of data.

- *A company can also use data analytics to make better business decisions and help analyze customer trends and satisfaction, which can lead to new and better products and services.*

Why Is Data Analytics Important?

- Implementing data analytics into the business model means companies can help reduce costs by identifying more efficient ways of doing business. A company can also use data analytics to make better business decisions.

Data Analysis Steps

Determine the data requirements or how the data is grouped.

Data analytics is the process of collecting it. This can be done through a variety of sources such as computers, online sources, cameras, environmental sources, or through personnel.

The data must be organized after it's collected so it can be analyzed.

The data is then cleaned up before analysis.

Wholeness of Data Analytics

- Business is the act of doing something productive to serve someone's needs, and thus earn a living and make the world a better place.
- Business activities are recorded on paper or using electronic media, and then these records become data.
- There is more data from customers' responses and on the industry as a whole.
- All this data can be analyzed and mined using special tools and techniques to generate patterns and intelligence, which reflect how the business is functioning.

Wholeness of Data Analytics

- These patterns can then be fed back into the business as new ideas so that it can evolve to become more effective and efficient in serving the needs of all stakeholders.

- And the cycle continues on leading to an exponential growth in data, intelligence, and business effectiveness (Figure 1.1).

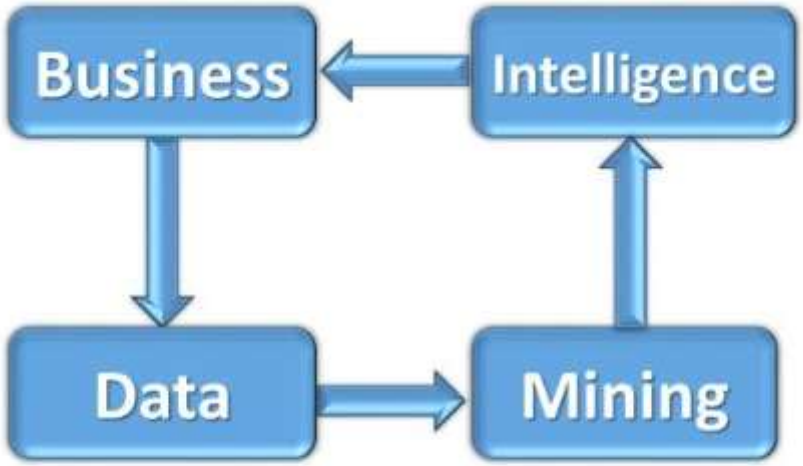


Figure 1.1: Business Intelligence and Data Mining (BIDM) Cycle

Types of Data Analytics

- **Descriptive analytics:** This describes what has happened over a given period of time. Have the number of views gone up? Are sales stronger this month than last?
- **Diagnostic analytics:** This focuses more on why something happened. It involves more diverse data inputs and a bit of hypothesizing. Did the weather affect beer sales? Did that latest marketing campaign impact sales?
- **Predictive analytics:** This moves to what is likely going to happen in the near term. What happened to sales the last time we had a hot summer? How many weather models predict a hot summer this year?
- **Prescriptive analytics:** This suggests a course of action. We should add an evening shift to the brewery and rent an additional tank to increase output if the likelihood of a hot summer is measured as an average of these five weather models and the average is above 58%,

Data Analytics Techniques

- **Regression analysis** entails analyzing the relationship between dependent variables to determine how a change in one may affect the change in another.
- **Factor analysis** entails taking a large data set and shrinking it into a smaller data set. The goal of this maneuver is to attempt to discover hidden trends that would otherwise have been more difficult to see.
- **Cohort analysis** is the process of breaking a data set into groups of similar data, often into a customer demographic. This allows data analysts and other users of data analytics to further dive into the numbers relating to a specific subset of data.
- **Monte Carlo simulations** model the probability of different outcomes happening. They're often used for risk mitigation and loss prevention. These simulations incorporate multiple values and variables and often have greater forecasting capabilities than other data analytics approaches.
- **Time series analysis** tracks data over time and solidifies the relationship between the value of a data point and the occurrence of the data point. This data analysis technique is usually used to spot cyclical trends or to project financial forecasts.

Data Analytics Tools

- Data analytics has rapidly evolved in technological capabilities in addition to a broad range of mathematical and statistical approaches to crunching numbers. Data analysts have a broad range of software tools to help acquire data, store information, process data, and report findings.

- Data analytics has always had loose ties to spreadsheets and Microsoft Excel. Data analysts also often interact with raw programming languages to transform and manipulate databases.

- SAS is an analytics platform that can assist with data mining. Apache Spark is an open-source platform useful for processing large sets of data.

Data Analytics Tools

- There are many tools used in data analytics, depending on the type of data, the analysis method, and the desired outcome. Some of the most popular tools are:
- **Microsoft Excel**: A spreadsheet software that can be used for data wrangling and visualization[1].
- **Python**: A programming language that can be used for data analysis, machine learning, web development, and more. It has thousands of free libraries that provide various functionalities[1].
- **R**: A programming language that is specialized for statistical analysis and data visualization. It has a large community of users and developers who create packages for different purposes[1].
- **Jupyter Notebook**: An interactive web-based platform that allows users to write and run code, display graphs, and document their work[1].

Data Analytics Tools

- **Apache Spark**: A distributed computing framework that can process large-scale data in parallel. It can be used for data engineering, machine learning, streaming analytics, and more[1].
- **SAS**: A software suite that offers solutions for data management, analytics, business intelligence, and more. It is widely used in industries such as healthcare, banking, and government[1].
- **Microsoft Power BI**: A business analytics service that can connect to various data sources, create interactive dashboards, and share insights[1].
- **Tableau**: A data visualization software that can create interactive charts, maps, graphs, and more. It can also connect to various data sources and perform basic analysis[2].
- **IBM Watson Analytics**: A cloud-based service that uses artificial intelligence to automate data preparation, analysis, and visualization. It can also generate natural language summaries and recommendations[2].
- These are just some examples of the tools used in data analytics. There are many more tools available for different purposes and preferences.

The Role of Data Analytics

- Data analytics can enhance operations, efficiency, and performance in numerous industries by shining a spotlight on patterns. Implementing these techniques can give companies and businesses a competitive edge. The process includes four basic steps of analysis.

What's the difference between data analytics and data science?

- The terms "data science" and "data analytics" tend to be used interchangeably.
- However, they are two different fields and denote two distinct career paths.
- It's important to recognize that data science and data analytics work together, and both make extremely valuable contributions to business.

What's the difference between data analytics and data science?

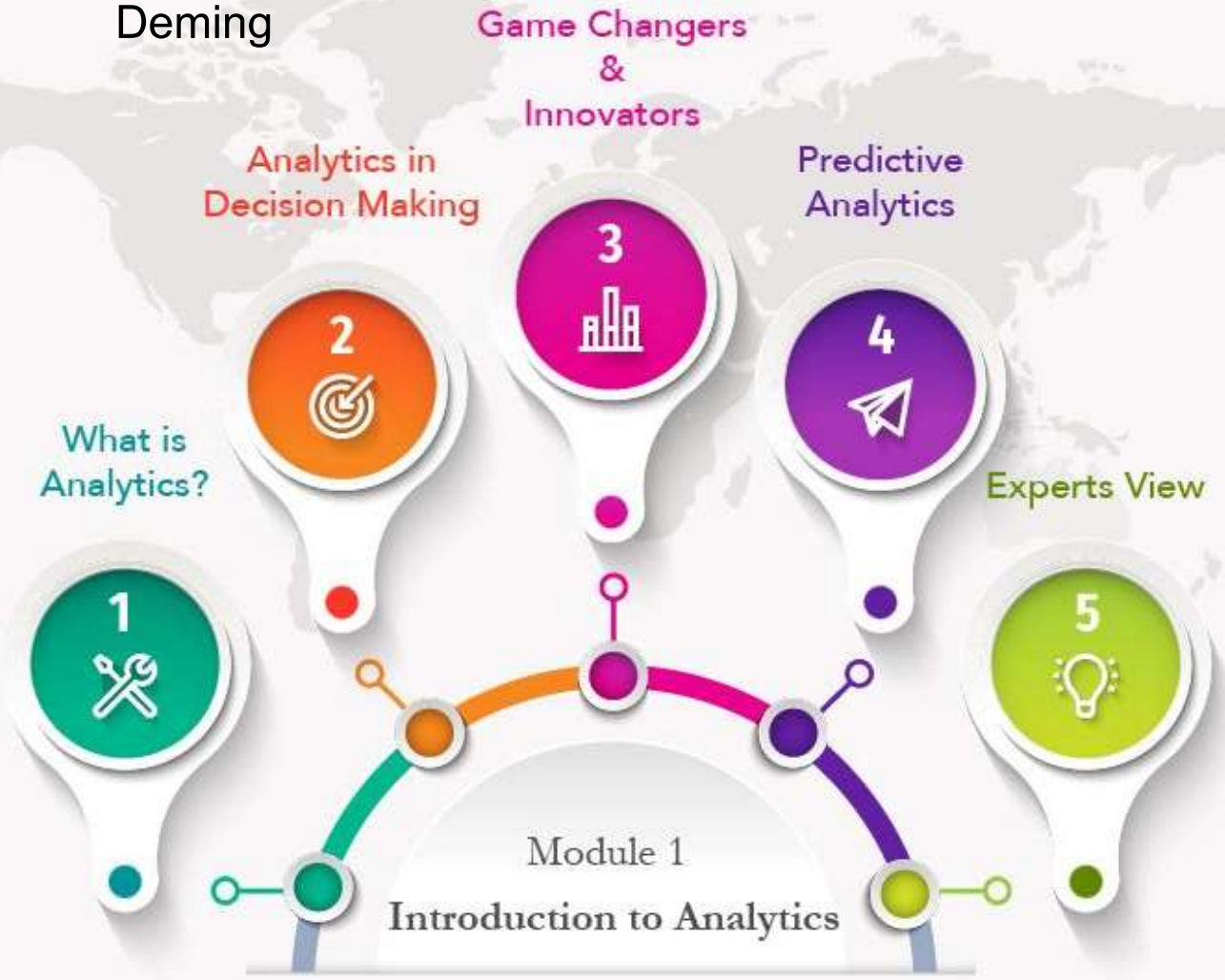
- A data analyst is usually part of the Business Intelligence team, and their work often has a direct impact on the decision-making occurring within the team.

- A data analyst's key responsibilities are:
- Interpreting data and identifying patterns using statistical techniques.
- Developing databases and data collection systems to optimize statistical efficiency.
- Working with management to understand business priorities.
- Filtering and cleaning data to ensure efficiency in data collection.

- Data scientists take a more science-based approach to data handling. The work of a data scientist incorporates mathematical knowhow, computer skills, and business acumen.
- A data scientist's key responsibilities are:
- Locating valuable sources of data and developing processes to gather such data.
- Building data analysis models to address business problems.
- Creating algorithms and predictive models to test data.
- Presenting findings and information using data visualisation techniques.
- Suggesting solutions and strategies for overcoming business problems.

"In God we trust; all others must bring Data"

- Edwards Deming



- Analytics is a body of knowledge consisting of statistical, mathematical, and operations research techniques; artificial intelligence techniques such as machine learning and deep learning algorithms; data collection and storage; data management processes such as data extraction, transformation and loading (ETL); and computing and big data technologies such as Hadoop, Spark, and Hive that create value by developing actionable items from data. The primary macro-level objectives of analytics are problem solving and decision making.

- "Analytics help organizations to create value by solving problems effectively and assisting in decision making"

Types of data

- Univariate/Multivariate
  1. Univariate: When we use one variable to describe a person, place, or thing.
  2. Multivariate: When we use two or more variables to measure a person, place or thing. Variables may or may not be dependent on each other.
- Cross-sectional data/Time-ordered data (business, social sciences)
  1. Cross-Sectional: Measurements taken at one time period
  2. Time-Ordered: Measurements taken over time in chronological sequence.

The type of data will dictate (in part) the appropriate data-analysis method.

Measurement Scales

- **Nominal or Categorical Scale**
  - Classification of people, places, or things into categories (e.g. age ranges, colors, etc.).
  - Classifications must be mutually exclusive (every element should belong to one category with no ambiguity).
  - Weakest of the four scales. No category is greater than or less (better or worse) than the others. They are just different.
- **Ordinal or Ranking Scale**
  - Classification of people, places, or things into a ranking such that the data is arranged into a meaningful order (e.g. poor, fair, good, excellent).
  - Qualitative classification only

Measurement Scales (business, social sciences)

1. Interval Scale

   • Data classified by ranking.

   • Quantitative classification (time, temperature, etc).

   • Zero point of scale is arbitrary (differences are meaningful).

2. Ratio Scale

   • Data classified as the ratio of two numbers.

   • Quantitative classification (height, weight, distance, etc).

   • Zero point of scale is real (data can be added, subtracted, multiplied, and divided).

Univariate Analysis/Descriptive Statistics

- Descriptive Statistics

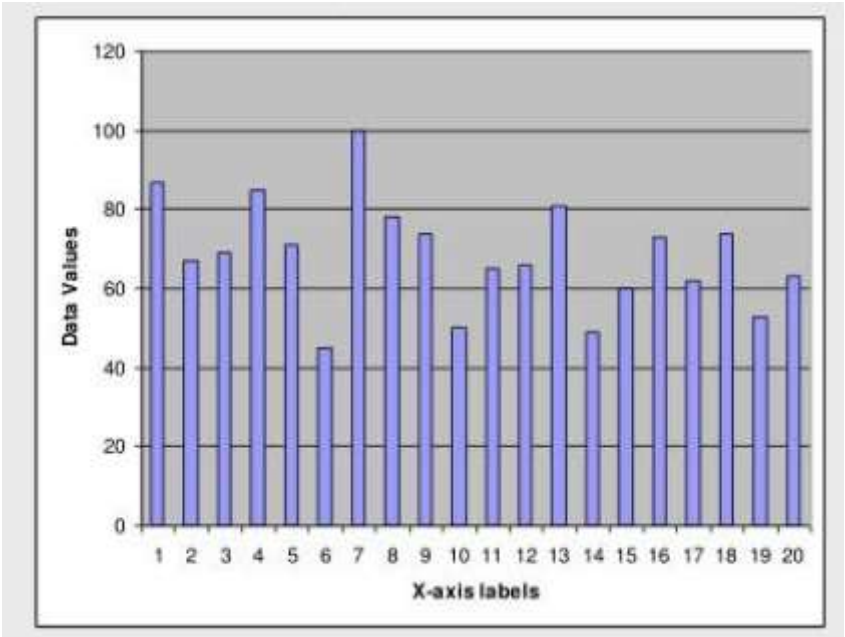
1. The Range
2. Min/Max
3. Average
4. Median
5. Mode
6. Variance
7. Standard Deviation
8. Histograms and Normal Distributions

Distributions

- Descriptive statistics are easier to interpret when graphically illustrated.
- However, charting each data element can lead to very busy and confusing charts that do not help interpret the data.
- Grouping the data elements into categories and charting the frequency within these categories yields a graphical illustration of how the data is distributed throughout its range. Univariate Analysis/Histograms

Univariate Analysis/Histograms

- With just a few columns this chart is difficult to interpret. It tells you very little about the data set. Even finding the Min and Max can be difficult.



- The data can be presented such that more statistical parameters can be estimated from the chart (average, standard deviation).

Univariate Analysis/Histograms

- Frequency Table

• The first step is to decide on the categories and group the data appropriately.

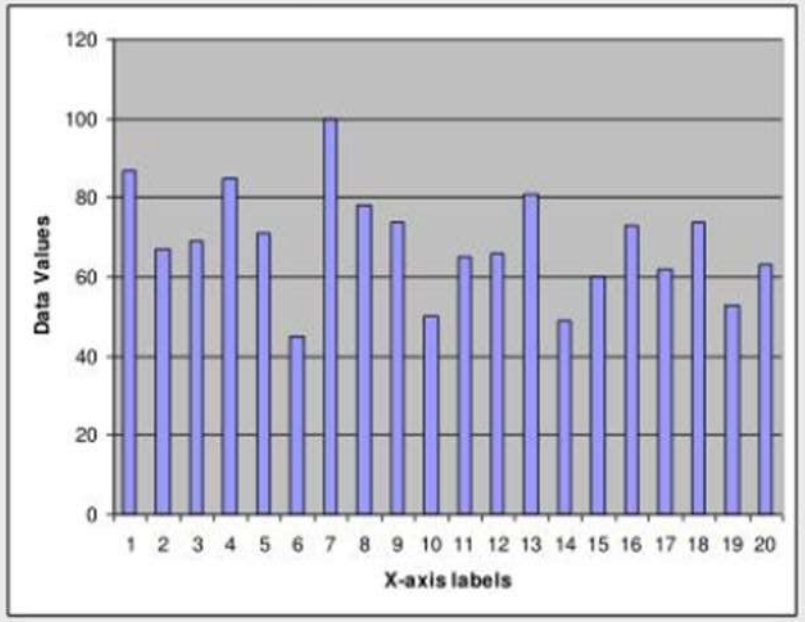
Eg: (45, 49, 50, 53, 60, 62, 63, 65, 66, 67, 69, 71, 73, 74, 74, 78, 81, 85, 87, 100)

Category Labels	Frequency
0-50	3
51-60	2
61-70	6
71-80	5
81-90	3
>90	1

Histogram

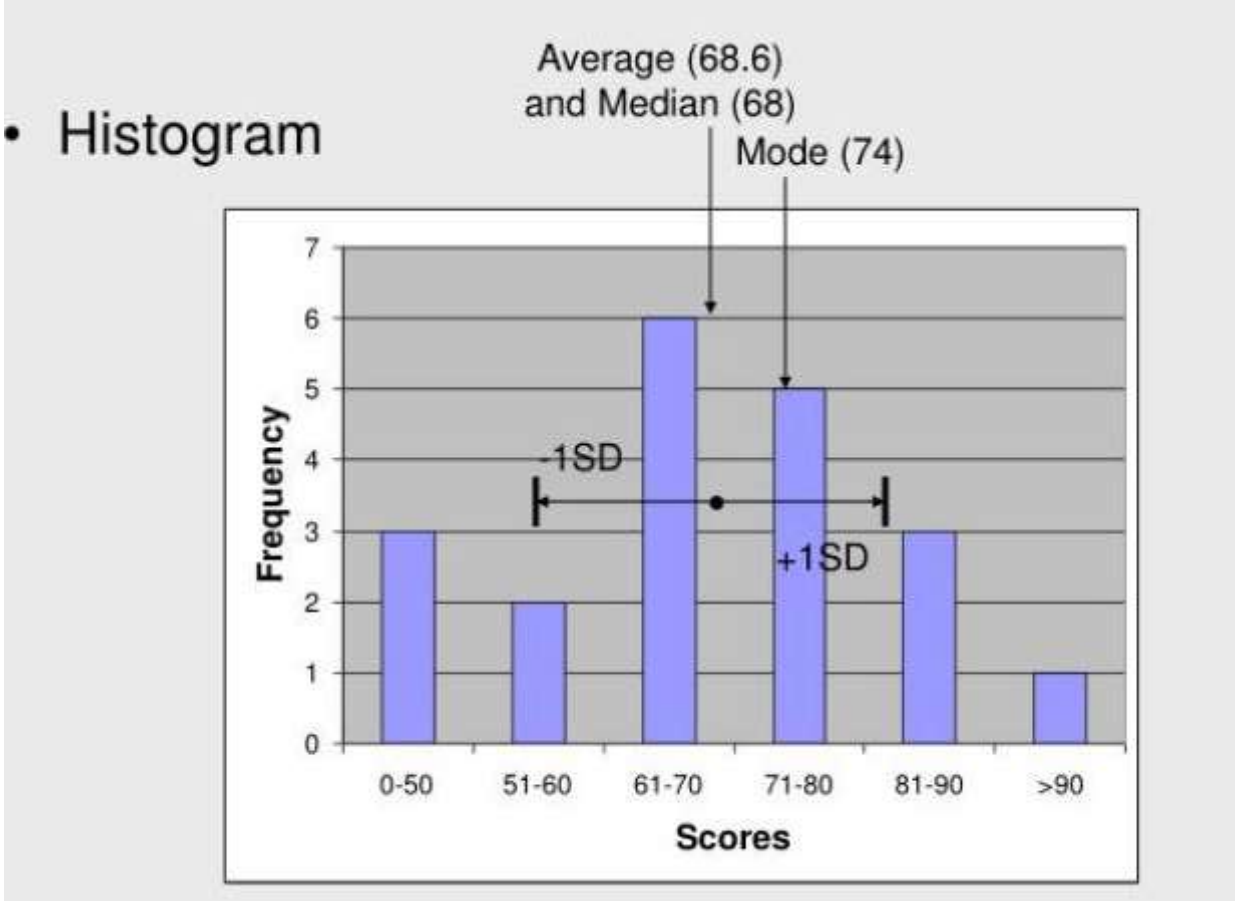
- A histogram is simply a column chart of the frequency table.

Category Labels	Frequency
0-50	3
51-60	2
61-70	6
71-80	5
81-90	3
>90	1



Univariate Analysis/Normal Distributions

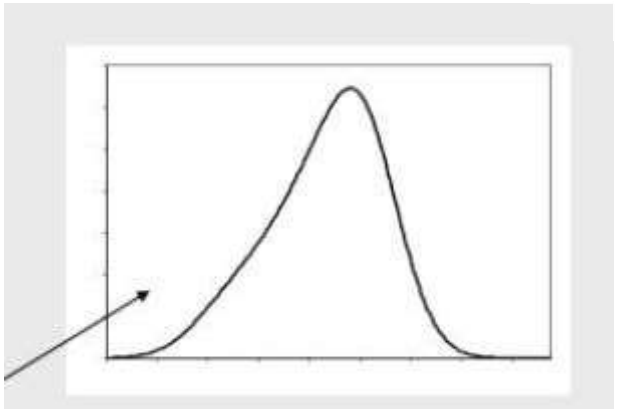
- Histogram Univariate Analysis/Histograms Average (68.6) and Median (68) Mode (74) -1SD +1SD

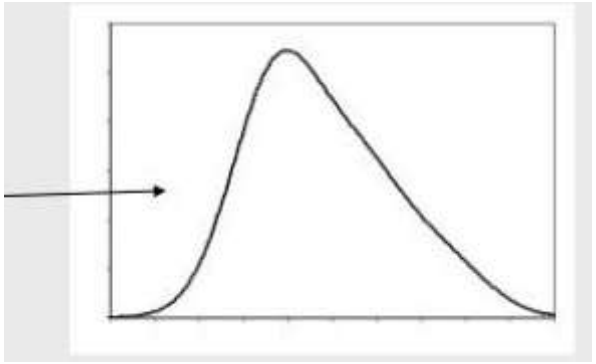


Univariate Analysis/Skewed Distributions

- When data are skewed, the mean and SD can be misleading
- Skewness sk= 3(mean-median)/SD
  - If sk>|1| then distribution is non-symmetrical

• Positively Skewed • Mean>Median • Sk is positive

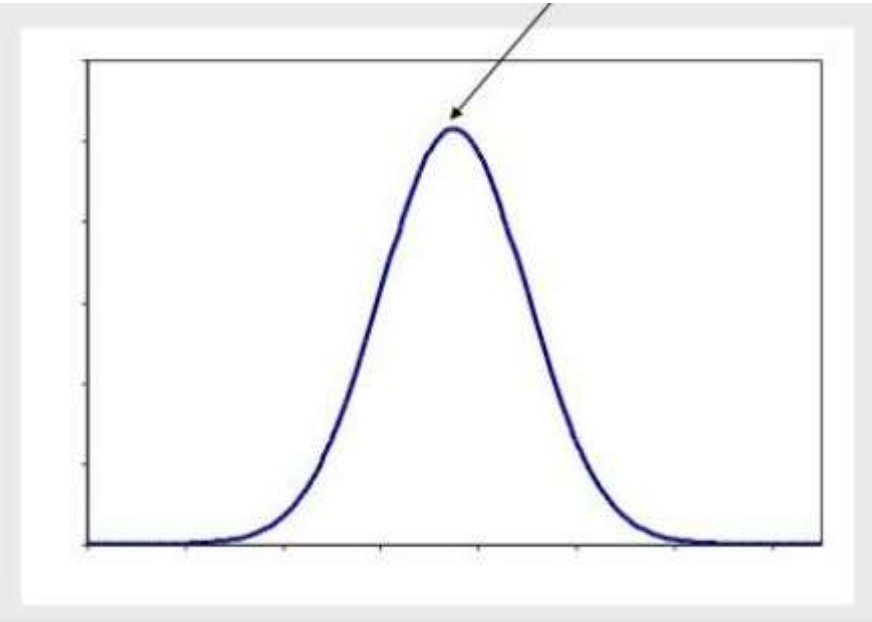




• Negatively skewed • Mean<Median• Sk is negative

Univariate Analysis/Normal Distributions

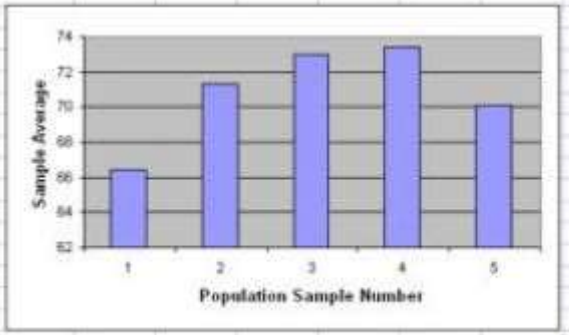
- Distributions that can be described mathematically as Gaussian are also called Normal
- **The Bell curve** : • Symmetrical • Mean ≈ Median
  - Mean, Median, Mode



Central Limit Theorem

- Regardless of the shape of a distribution, the distribution of the sample mean based on samples of size N approaches a normal curve as N increases.
- N must be less than the entire sample N=10

	Population	S1	S2	S3	S4	S5	
2	63	73	74	74	74	49	
3	49	50	62	60	62	60	
4	62	71	69	73	74	62	
5	60	63	67	69	73	63	
6	53	62	66	67	71	73	
7	66	74	78	85	78	74	
8	66	81	81	87	87	74	
9	50	60	63	66	65	78	
10	45	45	53	49	50	81	
11	67	85	100	100	100	87	
12	69	66.4	71.3	73	73.4	70.1	Average
13	71						
14	73						
15	74						
16	74						
17	78						
18	81						
19	85						
20	87						
21	100						



The Range

- Difference between minimum and maximum values in a data set
- Larger range usually (but not always) indicates a large spread or deviation in the values of the data set.

Eg: (73, 66, 69, 67, 49, 60, 81, 71, 78, 62, 53, 87, 74, 65, 74, 50, 85, 45, 63, 100)

The Average (Mean)

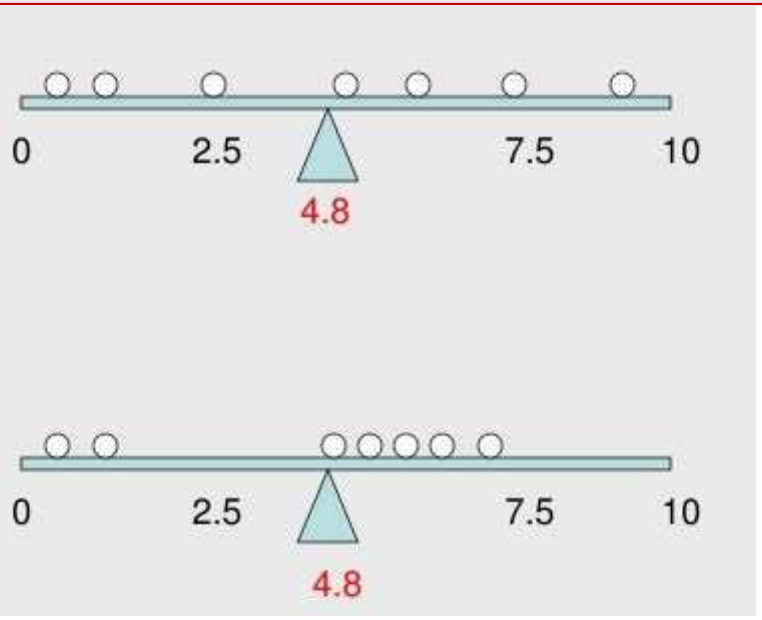
Sum of all values divided by the number of values in the data set. • One measure of central location in the data set.

$$\text{Average} = \frac{1}{N}\sum_{i=1}^{N} m_i$$

Average=(73+66+69+67+49+60+81+71+78+62+53+87+74+65+74+ 50+85+45+63+100)/20 = 68.6

Excel function: AVERAGE()

The data may or may not be symmetrical around its average value



The Median

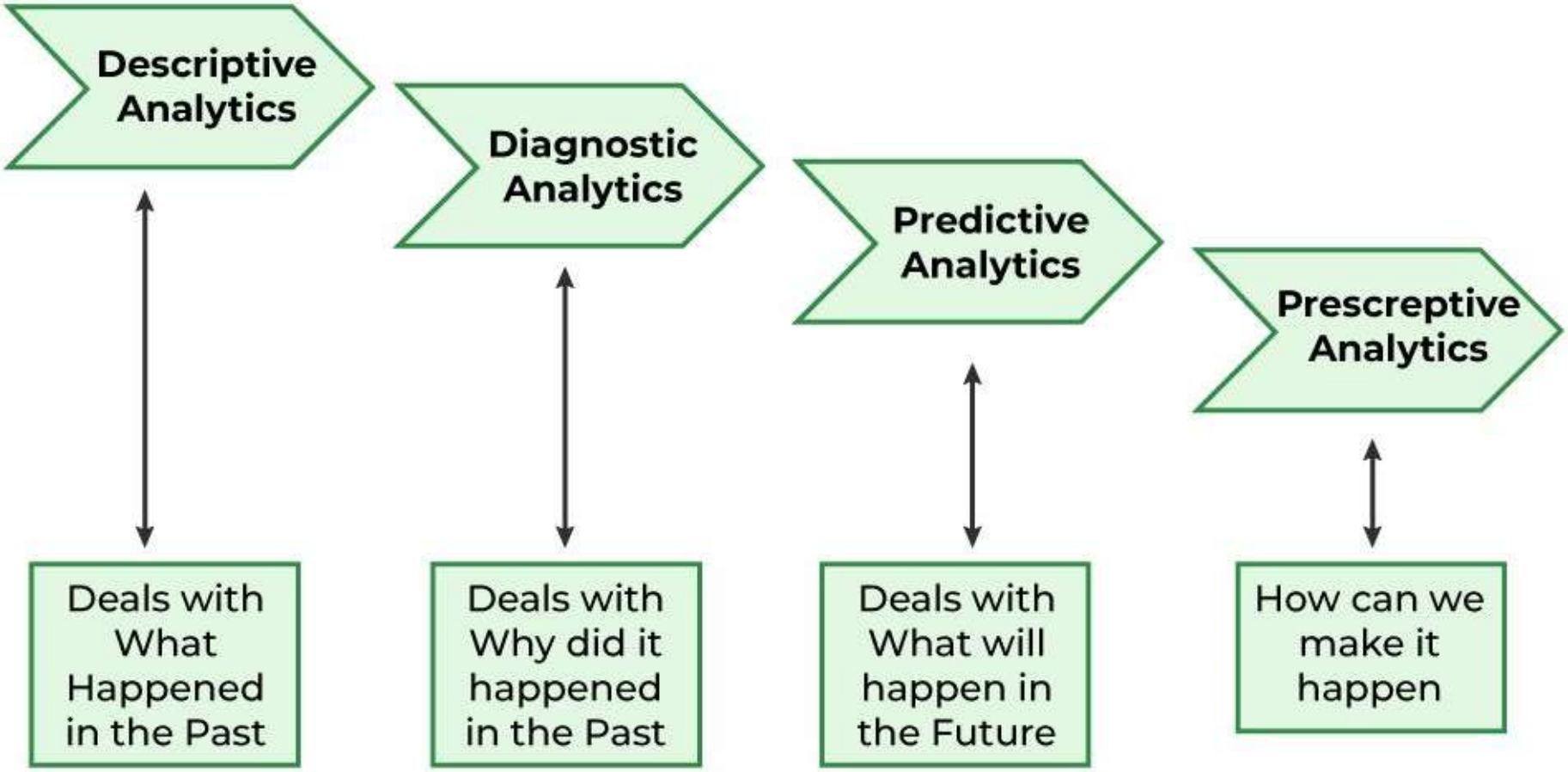
- The middle value in a sorted data set. Half the values are greater and half are less than the median.
- Another measure of central location in the data set. (45, 49, 50, 53, 60, 62, 63, 65, 66, 67, 69, 71, 73, 74, 74, 78, 81, 85, 87, 100) Median: 68 (1, 2, 4, 7, 8, 9, 9)
- Excel function: MEDIAN() Univariate Analysis/Descriptive Statistics

WHAT IS DATA ANALYTICS?

- Data analytics (DA) is the process of examining data sets to discover trends and draw conclusions about the information contained within them. Data analytics is increasingly being performed with the assistance of specialized systems and software.
- THERE ARE FOUR TYPES OF DATA ANALYTICS:

1. Descriptive Analysis
2. Diagnostic Analysis
3. Predictive Analysis
4. Prescriptive Analysis

Types of Data Analytics,



Descriptive Analysis

- The simplest way of understanding the functioning of **descriptive analytics** is the answer to **the question of what**.

- Consider the following example for greater clarity.
- What happened in each product category can be deduced from the data available on income and manufacturing.
- Divide your data into the following categories:
  - The earnings per product category
  - The monthly revenue or income
  - Total number of products manufactured in each category per month

Descriptive Analysis

Ecommerce businesses that use a customer's browsing and purchasing history to make product recommendations.

Financial organizations that need help determining whether a customer is likely to pay their credit card bill on time.

Marketers who analyze data to determine the likelihood that new customers will respond favorably to a given campaign or product offering.

DIAGNOSTIC ANALYSIS

- Regardless of the type of analysis, past events will always be necessary.
- This method, which is a step above descriptive analysis, requires you to compare it to other data.
- The goal is to investigate the why question.
  - Assume you need to investigate why a certain profit was made or not made.
  - The analysis will focus on sales and gross profit in this case.

DIAGNOSTIC ANALYSIS

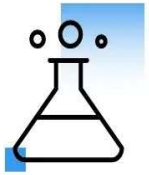
- Examples of diagnostic analytics include:
- Why did year-over-year sales go up?
- Why did a certain product perform above expectations?
- Why did we lose customers in Q3?

PREDICTIVE ANALYSIS

- Regardless of the type of analysis, past events will always be necessary.
- Predictive analytics is an advanced form of data analytics that attempts to answer the question, **"What might happen next?"**
- Predictive analytics is the process of using data to forecast future outcomes. The process uses data analysis, machine learning, artificial intelligence, and statistical models to find patterns that might predict future behavior. Organizations can use historic and current data to forecast trends and behaviors seconds, days, or years into the future with a great deal of precision.

PREDICTIVE ANALYSIS

Calculating client risk in the insurance industry to determine what plans and rates an account should be offered.



Discovering what features to include in a new product to ensure its success in the market, possibly by analyzing data like customer surveys and market research to identify what features are most desirable for customers and prospects.

Identifying tactics to optimize patient care in healthcare, like assessing the risk for developing specific health problems in the future and targeting treatment decisions to reduce those risks.

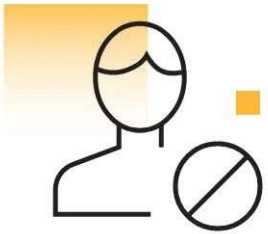
PRESCRIPTIVE ANALYSIS

- Finally, you will be able to recommend a course of action using prescriptive analytics.
- Remember that the goal of this stage is to eliminate future problems.
- While the technical processes will immerse themselves here, you must notice a promising trend.

PRESCRIPTIVE ANALYSIS

Why did year-over-year sales go up?

Why did a certain product perform above expectations?



Why did we lose customers in Q3?

Summary

- Data analytics is quite an important process of breaking down the functioning of your organization.
- The primary reason behind that being the help it offers to businesses. Meaning, optimizing your performances accordingly becomes swift and efficient.

- By implementing the required data analytics into your business model would reduce incurred costs or failure. Moreover, it helps you identify multiple efficient ways of going about with your daily business!

Summary

	Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics
Purpose	To answer what has happened	To answer why something has happened	To consider what might happen	To suggest a course of action
Hierarchy status	For basic foundational purposes	For basic foundational purposes	Higher in sophistication: building a structure	Highest in sophistication: implementing a structure

Summary

	Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics
Factors considered	Past or historical events: data that is internal to the organization or company	Past or historical events: data that is internal to the organization or company	– Past or historical events: data that is internal to the organization or companyUsed further for future possibilities	– Considering historical events as well as outside trends: data internal to the organization and external tendencies or changes. – Further puts the possible forecasted outcomes in comparison with predicted consequences

Summary

	Descriptive Analytics	Diagnostic Analytics	Predictive Analytics	Prescriptive Analytics
Benefits	– Quick and easy access to returns on investment – Better for spotting problematic gaps or phases in a report	– For determining causality if any – Concise and comprehensive understanding of these causalities	– For smarter detection purposes Helps prioritise agenda	– For streamlining the decision-making of your organization- To overcome problems quicker by preparing for the different possibilities