Data Analytics

Assignment 1

Unit-01

1.Compare the terms data analysis and data analytics and justify how each is applied in real-word scenarios.

Ans:- 1. Data Analytics : Data analytics is an important field that involves the process of collecting, processing, and interpreting data to uncover insights and help in making decisions. Data analytics is the practice of examining raw data to identify trends, draw conclusions, and extract meaningful information. This involves various techniques and tools to process and transform data into valuable insights that can be used for decision-making.

2. Data Analysis : It is the technique of observing, transforming, cleaning, and modeling raw facts and figures with the purpose of developing beneficial information and acquiring profitable conclusions.

S.No.	Data Analytics	Data analysis
1.	It is described as a traditional form or generic form of analytics.	It is described as a particularized form of analytics.
2.	It includes several stages like the collection of data and then the inspection of business data is done.	To process data, firstly raw data is defined in a meaningful manner, then data cleaning and conversion are done to get meaningful information from raw data.
3.	It supports decision making by analyzing enterprise data.	It analyzes the data by focusing on insights into business data.
4.	It uses various tools to process data such as Tableau, Python, Excel, etc.	It uses different tools to analyze data such as Rapid Miner, Open Refine, Node XL, KNIME, etc.
5.	Descriptive analysis cannot be performed on this.	A Descriptive analysis can be performed on this.
6.	One can find anonymous relations with the help of this.	One cannot find anonymous relations with the help of this.
7.	It does not deal with inferential analysis.	It supports inferential analysis.

2.List out types of data analytics and briefly explain each with neat diagram.

Ans:- Types of Data Analytics

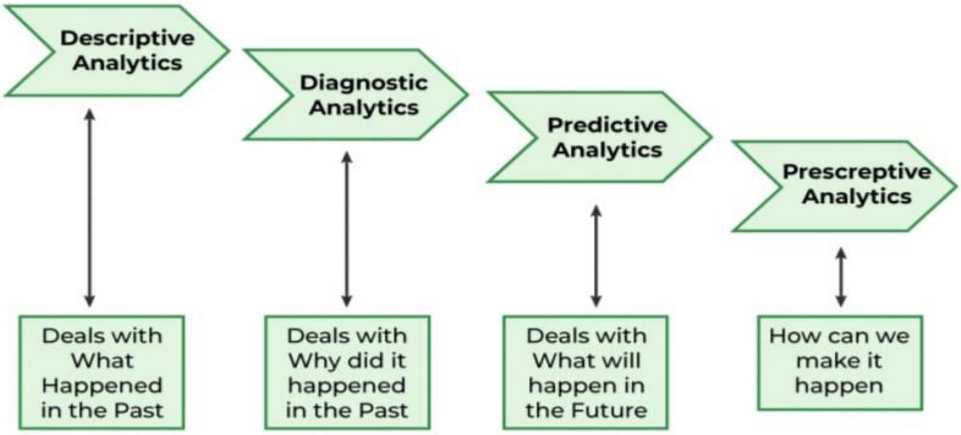
There are four major types of data analytics:

1. Predictive (forecasting)

2. Descriptive (business intelligence and data mining)

3. Prescriptive (optimization and simulation)

4. Diagnostic analytics


Descriptive Analytics
Diagnostic Analytics
Predictive Analytics
Prescreptive Analytics
Deals with What Happened in the Past
Deals with Why did it happened in the Past
Deals with What will happen in the Future
How can we make it happen

1. Descriptive analytics: What happened?

Descriptive analytics uses historical data from a single internal source to describe what happened. For example: How many people viewed the website? Which products had the most defects? This type of analytics employs simple mathematical and statistical tools, such as spreadsheets, instead of complex calculations to create visualizations , like bar charts or line graphics to describe a data set. Used by most businesses, descriptive analytics forms the crux of everyday reporting, especially through dashboards.

2. Diagnostic analytics: Why did it happen?

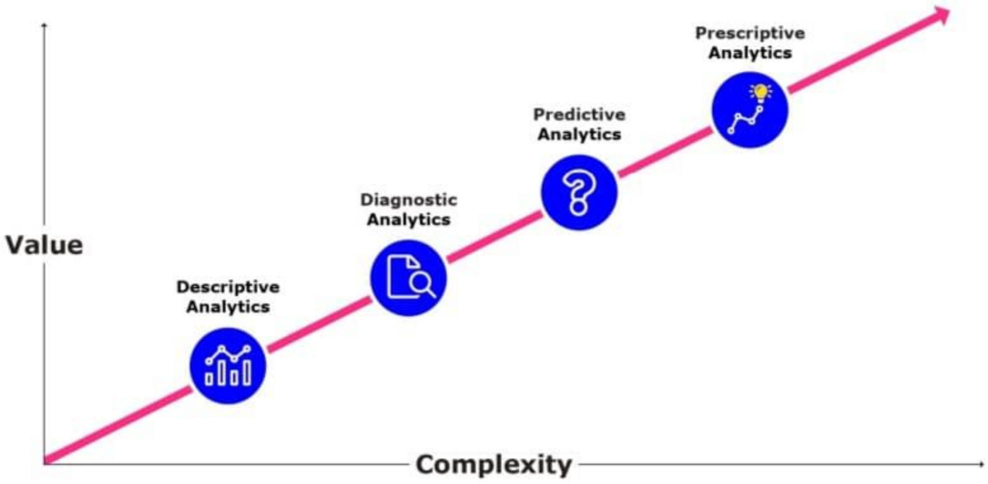
Diagnostic analytics is a form that dives deep into historical data to identify anomalies, find patterns, identify correlations, and determine causal relationships. Though diagnostic analytics can be performed manually, the rise of big data has pushed analysts to employ machine-learning techniques for the analysis. Unlike humans, computers can process vast amounts of data, recognise patterns, detect anomalies, and expose 'unusual' events. They can apply analytical techniques from a portfolio of algorithms to identify drivers of change and determine causation.

3. Predictive analytics: What might happen next?

As an organisation increases its analytical maturity and embarks on predictive analytics, it shifts its focus from understanding historical events to creating insights about a current or future state. Predictive analytics lies at the intersection of classical statistical analysis and modern artificial intelligence (AI) techniques. By employing predictive analytics, organisations identify the likelihood of possible outcomes, which can guide them on the best course of action. Predictive analytics is used in many sectors, such as the aerospace industry to predict the effect of maintenance operations on fuel use, and the manufacturing industry to predict future requirements and optimise warehouse stocking accordingly. Simple predictive models can be created using tools such as spreadsheets or Tableau.

4. Prescriptive analytics: What do I need to do?

Prescriptive analytics is the most complex type of analytics. It combines internal data, external sources, and machine-learning techniques to provide the most effective recommendations for business decisions. In prescriptive analytics, a decision-making process is applied to descriptive and predictive models. This leads to finding a combination of existing conditions and possible decisions that are likely to have the most effect in the future. This process is both complex and resource-intensive but, when done well, can provide immense value to an organisation.



3.Discuss the need of data analytics.

Ans:- The need for data analytics has become increasingly apparent in today's data-driven world due to several factors:

1. **Decision Making:** Businesses and organizations need to make informed decisions to remain competitive and adapt to changing market conditions. Data analytics provides insights and evidence-based recommendations that enable better decision-making across various domains, including marketing, operations, finance, and human resources.
2. **Identifying Trends and Patterns:** Data analytics helps identify trends, patterns, and correlations within large datasets that might not be immediately apparent to human analysts. By analyzing historical data, organizations can uncover valuable insights into customer behavior, market trends, and operational inefficiencies, allowing them to anticipate future trends and plan accordingly.
3. **Improving Efficiency and Performance:** Analyzing data can reveal inefficiencies and bottlenecks in business processes, allowing organizations to streamline operations, reduce costs, and improve overall efficiency. By optimizing resource allocation and workflow processes, companies can enhance productivity and performance.
4. **Enhancing Customer Experience:** Understanding customer behavior and preferences is crucial for delivering personalized and targeted experiences. Data analytics enables organizations to segment customers based on demographics, purchasing history, and interactions, allowing for targeted marketing campaigns, product recommendations, and customer service improvements.
5. **Risk Management:** Data analytics plays a vital role in risk management by identifying potential risks and threats to businesses, such as fraud, cybersecurity breaches, and market fluctuations. By analyzing historical data and using predictive modeling techniques, organizations can anticipate and mitigate risks before they escalate into significant problems.
6. **Innovation and Product Development:** Data analytics fuels innovation by providing insights into market demands, consumer preferences, and emerging trends. By analyzing data from various sources, organizations can identify unmet needs and develop new products or services that better meet customer requirements.
7. **Compliance and Regulation:** In heavily regulated industries such as finance, healthcare, and utilities, compliance with laws and regulations is paramount. Data analytics helps organizations ensure compliance by monitoring and analyzing vast amounts of data to identify any irregularities or non-compliance issues.

8. **Competitive Advantage:** In today's highly competitive business environment, gaining a competitive edge is essential for long-term success. Data analytics can provide organizations with unique insights and actionable intelligence that can differentiate them from competitors, whether through innovative products, superior customer service, or operational excellence.

Overall, the need for data analytics stems from the growing volume, variety, and velocity of data generated in today's digital age. By harnessing the power of data analytics, organizations can unlock valuable insights, drive innovation, and stay ahead in an increasingly competitive marketplace.

4.What is meant by sample? Explain sampling with replacement and without replacement with proper example.

Ans:- In statistics and research methodology, a sample refers to a subset of a population that is selected for analysis. Samples are used because it is often impractical or impossible to study an entire population, so researchers collect data from a representative subset to make inferences about the population as a whole.

Sampling with replacement and sampling without replacement are two different methods for selecting elements from a population.

1. **Sampling with Replacement:**
   - In sampling with replacement, each element selected from the population is returned to the population before the next selection is made. This means that the same element can be selected more than once in the sample.
   - Example: Let's say you have a bag with five balls numbered 1 through 5. If you're sampling with replacement and you draw a ball numbered 3, you record its number and put it back in the bag before drawing again. So, there's a chance you might draw the same ball (number 3) again in the next draw.
   - Suppose you need to select three balls with replacement. The possible outcomes for each draw could be: 1, 2, 3, 4, or 5. Therefore, each draw is independent of the others, and the same ball might be selected more than once.
2. **Sampling without Replacement:**
   - In sampling without replacement, each element selected from the population is not returned to the population before the next selection is made. Once an element is selected, it is removed from the population, and subsequent selections are made from the reduced population.

- Example: Using the same bag of five numbered balls, if you're sampling without replacement and you draw a ball numbered 3, you record its number and do not put it back in the bag before the next draw. So, there's no chance of drawing the same ball (number 3) again in the next draw.
- If you need to select three balls without replacement, the possible outcomes for each draw would depend on the balls remaining in the bag after each draw. For example, if you draw the ball numbered 3 first, for the second draw, you would only have four balls left in the bag.

Sampling with replacement allows for the possibility of selecting the same element more than once, while sampling without replacement ensures that each element is selected only once in the sample. The choice between these methods depends on the specific research question and the goals of the study.

5. Classify sampling types. Briefly explain Probability Sampling Techniques with proper diagram.

Ans:- Types of Sampling :-1)Probability Sampling 2)Non-Probability Sampling

1)Probability sampling - Random selection techniques are used to select the sample.

2) Non-probability sampling - Non-random selection techniques based on certain criteria are used to select the sample.

Types of Sampling Techniques in Data Analytics-

Probability Sampling Techniques:-Probability Sampling Techniques are one of the important types of sampling techniques. Probability sampling allows every member of the population a chance to get selected. It is mainly used in quantitative research when you want to produce results representative of the whole population.

1. Simple Random Sampling

In simple random sampling, the researcher selects the participants randomly. There are a number of data analytics tools like random number generators and random number tables used that are based entirely on chance.

Example: The researcher assigns every member in a company database a number from 1 to 1000 (depending on the size of your company) and then use a random number generator to select 100 members.

2. Systematic Sampling

In systematic sampling, every population is given a number as well like in simple random sampling. However, instead of randomly generating numbers, the samples are chosen at regular intervals.

Example: The researcher assigns every member in the company database a number. Instead of randomly generating numbers, a random starting point (say 5) is selected. From that number onwards, the researcher selects every, say, 10th person on the list (5, 15, 25, and so on) until the sample is obtained.

3. Stratified Sampling

In stratified sampling, the population is subdivided into subgroups, called strata, based on some characteristics (age, gender, income, etc.). After forming a subgroup, we can then use random or systematic sampling to select a sample for each subgroup. This method allows you to draw more precise conclusions because it ensures that every subgroup is properly represented.

Example: If a company has 500 male employees and 100 female employees, the researcher wants to ensure that the sample reflects the gender as well. So the population is divided into two subgroups based on gender.

4. Cluster Sampling

In cluster sampling, the population is divided into subgroups, but each subgroup has similar characteristics to the whole sample. Instead of selecting a sample from each subgroup, we randomly select an entire subgroup. This method is helpful when dealing with large and diverse populations.

Example: A company has over a hundred offices in ten cities across the world which has roughly the same number of employees in similar job roles. The researcher randomly selects 2 to 3 offices and uses them as the sample.

6. Classify sampling types. Briefly explain Non Probability Sampling Techniques with proper diagram.

Ans:- Non-Probability Sampling Techniques:-Non-Probability Sampling Techniques is one of the important types of Sampling techniques. In non-probability sampling, not every individual has a chance of being included in the sample. This sampling method is easier and cheaper but also has high risks of sampling bias. It is often used in exploratory and qualitative research with the aim to develop an initial understanding of the population.

1. Convenience Sampling

In this sampling method, the researcher simply selects the individuals which are most easily accessible to them. This is an easy way to gather data, but there is no way to tell if the sample is representative of the entire population. The only criteria involved is that people are available and willing to participate.

Example: The researcher stands outside a company and asks the employees coming in to answer questions or complete a survey.

2. Voluntary Response Sampling

Voluntary response sampling is similar to convenience sampling, in the sense that the only criterion is people are willing to participate. However, instead of the researcher choosing the participants, the participants volunteer themselves.

Example: The researcher sends out a survey to every employee in a company and gives them the option to take part in it.

3.Purposive Sampling

In purposive sampling, the researcher uses their expertise and judgment to select a sample that they think is the best fit. It is often used when the population is very small and the researcher only wants to gain knowledge about a specific phenomenon rather than make statistical inferences.

Example: The researcher wants to know about the experiences of disabled employees at a company. So the sample is purposefully selected from this population.

4. Snowball Sampling

In snowball sampling, the research participants recruit other participants for the study. It is used when participants required for the research are hard to find. It is called snowball sampling because like a snowball, it picks up more participants along the way and gets larger and larger.

Example: The researcher wants to know about the experiences of homeless people in a city. Since there is no detailed list of homeless people, a probability sample is not possible. The only way to get the sample is to get in touch with one homeless person who will then put you in touch with other homeless people in a particular area.

7. What is sampling distribution and explain its types.

Ans:- A sampling distribution refers to a probability distribution of a statistic that comes from choosing random samples of a given population. Also known as a finite-sample distribution, it represents the distribution of frequencies on how spread apart various outcomes will be for a specific population.The sampling distribution depends on multiple factors – the statistic, sample size, sampling process, and the overall population. It is used to help calculate statistics such as means, ranges, variances, and standard deviations for the given sample.

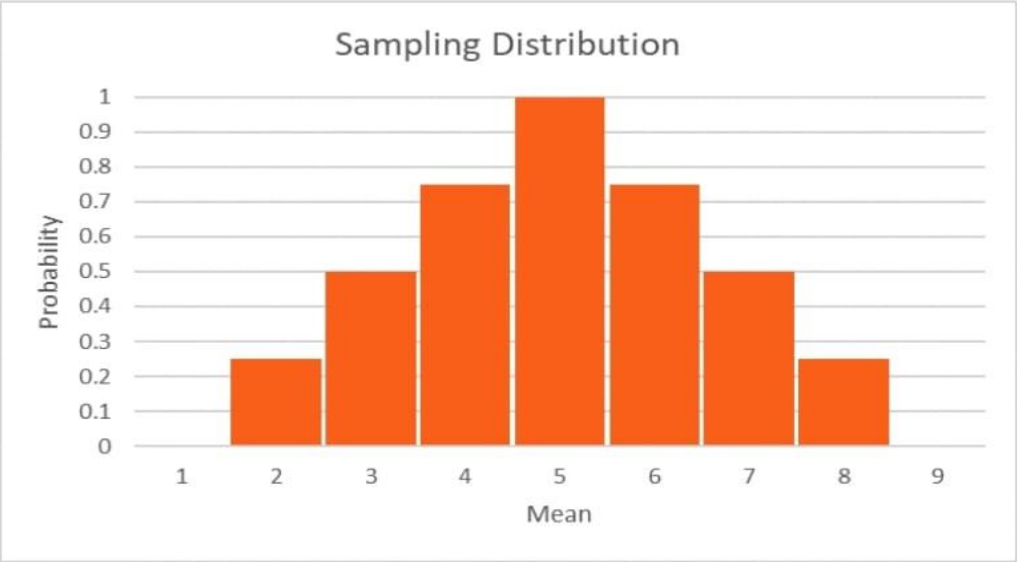
How Does it Work?

1. Select a random sample of a specific size from a given population.

2. Calculate a statistic for the sample, such as the mean, median, or standard deviation.

3. Develop a frequency distribution of each sample statistic that you calculated from thestep above.

4. Plot the frequency distribution of each sample statistic that you developed from the step above. The resulting graph will be the sampling distribution.



Types of Sampling Distribution

1. Sampling distribution of mean

As shown from the example above, you can calculate the mean of every sample group chosen from the population and plot out all the data points. The graph will show a normal distribution, and the center will be the mean of the sampling distribution, which is the mean of the entire population.

2. Sampling distribution of proportion

It gives you information about proportions in a population. You would select samples from the population and get the sample proportion. The mean of all the sample proportions that you calculate from each sample group would become the proportion of the entire population.

3. T-distribution

T-distribution is used when the sample size is very small or not much is known about the population. It is used to estimate the mean of the population, confidence intervals, statistical differences, and linear regression.

Practical Example

Suppose you want to find the average height of children at the age of 10 from each continent. You take random samples of 100 children from each continent, and you compute the mean for each sample group.

For example, in South America, you randomly select data about the heights of 10-year-old children, and you calculate the mean for 100 of the children. You also randomly select data from North America and calculate the mean height for one hundred 10-year-old children. As you continue to find the average heights for each sample group of children from each continent, you can calculate the mean of the sampling distribution by finding the mean of all the average heights of each sample group. Not only can it be computed for the mean, but it can also be calculated for other statistics such as standard deviation and variance.

8. With proper example explain sampling distribution of mean and justify it with central limit theorem.

Ans:- The sampling distribution of the mean is a probability distribution that describes the possible values that the mean of a sample can take from a population. It provides information about the variability of sample means when multiple samples are drawn from the same population.

Example:

Suppose we have a population of students and their exam scores. The population has a mean score of 80 and a standard deviation of 10. We are interested in the sampling distribution of the mean exam score for samples of size n.

1. **Sampling Distribution of the Mean:**
    - We take multiple samples, each of size n, from the population and calculate the mean of each sample.
    - The sampling distribution of the mean shows all the possible sample means and their associated probabilities.
    - As the sample size increases, the sampling distribution of the mean tends to become more normally distributed, regardless of the shape of the population distribution. This is known as the Central Limit Theorem.
2. **Central Limit Theorem (CLT):**
    - The Central Limit Theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the shape of the population distribution.
    - According to the CLT, if we take a large enough number of samples from any population (even if the population distribution is not normal), the distribution of the sample means will be approximately normally distributed.
    - The mean of the sampling distribution of the mean is equal to the population mean, and the standard deviation of the sampling distribution of the mean (also known as the standard error) is equal to the population standard deviation divided by the square root of the sample size: $n\sigma$, where σ is the population standard deviation.

Let's say we take multiple samples of size 30 from our population of students. According to the Central Limit Theorem, as we increase the number of samples, the distribution of sample means will become approximately normal, centered around the population mean (80 in this case), and with a standard deviation of $1030 \approx 1.83\,3010 \approx 1.83$.

This means that even if the original population distribution of exam scores was not normal, the distribution of sample means will tend to be normal as long as the sample size is sufficiently large.

The sampling distribution of the mean provides insights into the behavior of sample means drawn from a population, and the Central Limit Theorem justifies why the sampling distribution tends to be approximately normal, regardless of the shape of the population distribution, as the sample size increases.

9. A population consists of four member 3,7,11,15.Consider all possible sample size two which can be drawn with replacement from population. Find the population mean , population standard deviation, the mean of sampling distribution of mean & standard deviation of sampling distribution of mean.

Ans:- With Replacement

Population Mean=9

Population standard deviation=root 20

Mean of sampling distribution=9

Standard deviation of sampling distribution of mean=10

Standard Deviation=Standard Error

10. A population consists of four member 3,7,11,15.Consider all possible sample size two which can be drawn with replacement from population. Find the population mean ,population standard deviation, the mean of sampling distribution of mean &amp; standard deviation of sampling distribution of mean.

Ans:- Without Replacement

Population Mean=9

Population standard deviation=root20

Mean of sampling distribution=9

Standard deviation of sampling distribution of mean=2.5820s

Standard Deviation=Standard Error

11. Briefly Explain the term Hypothesis and illustrate its types with proper example.

Ans:- **Hypothesis:** In Statistics, a hypothesis is defined as a formal statement, which gives the explanation about the relationship between the two or more variables of the specified population. It helps the researcher to translate the given problem to a clear explanation for the outcome of the study. It clearly explains and predicts the expected outcome.

Types of Hypothesis:

1.Simple Hypothesis

It shows a relationship between one dependent variable and a single independent variable. **For example –** If you eat more vegetables, you will lose weight faster. Here, eating more vegetables is an independent variable, while losing weight is the dependent variable.

2.Complex Hypothesis

It shows the relationship between two or more dependent variables and two or more independent variables.

For example-Eating more vegetables and fruits leads to weight loss, glowing skin, and reduces the risk of many diseases such as heart disease.

3.Directional Hypothesis

It shows how a researcher is intellectual and committed to a particular outcome. The relationship between the variables can also predict its nature.

For example- children aged four years eating proper food over a five-year period are having higher IQ levels than children not having a proper meal. This shows the effect and direction of the effect.

4.Non-directional Hypothesis

It is used when there is no theory involved. It is a statement that a relationship exists between two variables, without predicting the exact nature (direction) of the relationship.

5.Null Hypothesis

It provides a statement which is contrary to the hypothesis. It's a negative statement, and there is no relationship between independent and dependent variables. The symbol is denoted by "HO".

6.Associative and Causal Hypothesis

Associative hypothesis occurs when there is a change in one variable resulting in a change in the other variable. Whereas, the causal hypothesis proposes a cause and effect interaction between two or more variables.