INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal, Hyderabad - 500 043

COMPUTER SCIENCE AND ENGINEERING(AI &ML)

DEFINITION AND TERMINOLGY

Department	**COMPUTER SCIENCE AND E ENGINEERING(AI & ML)**				
Course Title	**STATISTICAL FOUNDATIONS OF DATA SCIENCE**				
Course Code	ACAC07				
Program	B.Tech				
Semester	V	CSE(AI & ML)			
Course Type	Electove				
Regulation	IARE-UG20				
Course Structure	Theory			Practical	
	Lecture	Tutorials	Credits	Laboratory	Credits
	3	0	3	-	-
Course Coordinator	Ms. K. Anjali, Assistant Professor				

COURSE OBJECTIVES:

The students will try to learn:

I	The fundamental knowledge on basics of data science.
II	The basic principles of data acquisition, exploring and modeling data efficiently.
III	The foundations of probability and statistics for data science.
IV	The current scope, potential applications of data science.

COURSE OUTCOMES:

After successful completion of the course, students should be able to:

CO 1	**Recall** the categories and levels of data using steps involved in data science.	Remember
CO 2	**Demonstrate** the data pre-processing terms for improving the quality of dataset using processes such as feature generation and feature selection	Understand
CO 3	**Solve** mathematical problems using various arithmetic and more challenging forms of math.	Apply
CO 4	**Apply** probability theorems and approaches for calculating the number of outcomes of the events.	Apply
CO 5	**Illustrate** the obtaining and sampling data in statistics to quantify and visualize our data.	Understand
CO 6	**Summarize** the concepts of communication by using the visualization and presenting strategies.	Understand

DEFINITION AND TERMINOLOGY:

S.No	DEFINITION	CO's
MODULE I		
FLAVORS OF DATA		
1	**What do you mean by data?**	CO 1
	Data is defined as facts or figures, or information that's stored in or used by a computer. An example of data is information collected for a research paper. An example of data is an email.	
2	**Define data science?**	CO 1
	Data science is the field of study that combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data.	
3	**Explain organized data?**	CO 1, CO 6
	This refers to data that is sorted into a row/column structure, where every row represents a single observation and the columns represent the characteristics of that observation.	
4	**Discuss Unorganized data?**	CO 1
	This is the type of data that is in the free form, usually text or raw audio/signals that must be parsed further to become organized.	
5	**Explain data model?**	CO 1
	A data model refers to an organized and formal relationship between elements of data, usually meant to simulate a real-world phenomenon	
6	**Write short notes on Machine learning?**	CO 1
	Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.Machine learning allows the user to feed a computer algorithm an immense amount of data and have the computer analyze and make data-driven recommendations and decisions based on only the input data.	

7	**Discuss Probabilistic model?**	CO 3
	Probabilistic modeling is a statistical technique used to take into account the impact of random events or actions in predicting the potential occurrence of future outcomes.For example, if you live in a cold climate you know that traffic tends to be more difficult when snow falls and covers the roads. We could go a step further and hypothesize that there will be a strong correlation between snowy weather and increased traffic incidents.	
8	**What is meant by Statistical model?**	CO 3
	A Statistical Model is the use of statistics to build a representation of the data and then conduct analysis to infer any relationships between variables or discover insights. Machine Learning is the use of mathematical and or statistical models to obtain a general understanding of the data to make predictions	
9	**Define the Exploratory data analysis (EDA) .**	CO 1
	Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to.EDA build a robust understanding of the data, issues associated with either the info or process. it's a scientific approach to get the story of the data.	
10	**Explain Quantitative data?**	CO 1
	Quantitative data are measures of values or counts and are expressed as numbers. Quantitative data are data about numeric variables (e.g. how many; how much; or how often).What is an example of quantitative data in science? Quantitative data could include its length, weight, number of toes on each paw, how high it can jump, how many ounces of food it eats each data, and its body temperature	
11	**Write short notes on Qualitative data?**	CO 1
	Qualitative data describes qualities or characteristics. It is collected using questionnaires, interviews, or observation, and frequently appears in narrative form. For example, it could be notes taken during a focus group on the quality of the food at Cafe Mac, or responses from an open-ended questionnaire.	

12	**What is meant by median?** At the ordinal level, the median is usually an appropriate way of defining the center of the data. The mean, however, would be impossible because division is not allowed at this level. We can also use the mode like we could at the nominal level.	CO 1
13	**Give the measures of variation?** Measures of variation in statistics are ways to describe the distribution or dispersion of your data. In other words, it shows how far apart data points are from each other. Statisticians use measures of variation to summarize their data	CO 1, CO 3
14	**Explain the term data mining?** Data mining is used to explore increasingly large databases and to improve market segmentation. By analysing the relationships between parameters such as customer age, gender, tastes, etc., it is possible to guess their behaviour in order to direct personalised loyalty campaigns.	CO 1
15	**Discuss the ratio level in data science?** The ratio level contains all of the features of the other 3 levels. At the ratio level, values can be categorized, ordered, have equal intervals and take on a true zero. While nominal and ordinal variables are categorical variables, interval and ratio variables are quantitative variables.	CO 1
16	**Annonate the term a Standard deviation?** Standard deviation is a number that describes how spread out the observations are. A mathematical function will have difficulties in predicting precise values, if the observations are ”spread”. Standard deviation is a measure of uncertainty.	CO 1
17	**Elaborate Discrete data?** When values in a data set are countable and can only take certain values, it is called discrete data. For example, number of students in a class, number of players required in a team, etc. We can easily count the variables in a discrete data.	CO 1
18	**Explore the continuous data in short?** Continuous data is data that falls in a constant sequence.Continuous data is data that can take any value. Height, weight, temperature and length are all examples of continuous data. Some continuous data will change over time; the weight of a baby in its first year or the temperature in a room throughout the day. Examples of continuous data: The amount of time required to complete a project. The height of children. The amount of time it takes to sell shoes. The amount of rain, in inches, that falls in a storm.	CO 1

19	**Illustrate nominal level data?** Nominal data is data that can be labelled or classified into mutually exclusive categories within a variable. These categories cannot be ordered in a meaningful way. For example, for the nominal variable of preferred mode of transportation, you may have the categories of car, bus, train, tram or bicycle.	CO 1
20	**Annotate nominal variable in research?** A nominal variable is a type of variable that is used to name, label or categorize particular attributes that are being measured. It takes qualitative values representing different categories, and there is no intrinsic ordering of these categories..	CO 1

MODULE II		
DATA PRE-PROCESSING AND FEATURE SELECTION		
1	**Illustrate the data cleaning?** Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled.	CO 2
2	**Why data cleaning is important in data science?** he main aims of data cleaning is to keep as much of a dataset intact as possible. This helps improve the reliability of your insights. Data cleaning is not only important for data analysis. It's also important for general business housekeeping (or 'data governance').	CO 2
3	**Illustrate data integration?** Data integration is a common industry term referring to the requirement to combine data from multiple separate business systems into a single unified view, often called a single view of the truth. This unified view is typically stored in a central data repository known as a data warehouse.	CO 2
4	**Explain data reduction and why is it important?** Data reduction is the process of reducing the amount of capacity required to store data. Data reduction can increase storage efficiency and reduce costs. Storage vendors will often describe storage capacity in terms of raw capacity and effective capacity, which refers to data after the reduction.	CO 2
5	**Give the steps of data transformation?** he Data Transformation Process Explained in Four Steps Step 1: Data interpretation. Step 2: Pre-translation data quality check. Step 3: Data translation. Step 4: Post-translation data quality check.	CO 2
6	**Discuss data transformation?** Data transformation is the process of converting data from one format, such as a database file, XML document or Excel spreadsheet, into another. Transformations typically involve converting a raw data source into a cleansed, validated and ready-to-use format.	CO 2

7	**Enlist amd explain the benefits of data transformation?** Here are some of the biggest benefits of data transformation: 1.Better Organization: Transformed data is easier for both humans and computers to use. 2.Improved Data Quality: Data transformation can help your organization eliminate quality issues such as missing values and other inconsistencies. 3.Perform Faster Queries: You can quickly and easily retrieve transformed data thanks to it being stored and standardized in a source location. 4.Better Data Management: Businesses are constantly generating data from more and more sources. Data transformation refines your metadata, so it's easier to organize and understand. 5.More Use Out of Data: While businesses may be collecting data constantly, a lot of that data sits around unanalyzed. Transformation makes it easier to get the most out of your data by standardizing it and making it more usable.	CO 3
8	**Write a short notes on Data Discretization?** Data discretization is defined as a process of converting continuous data attribute values into a finite set of intervals and associating with each interval some specific data value.	CO 3
9	**Explain feature generation ?** Feature Generation (also known as feature construction, feature extraction or feature engineering) is the process of transforming features into new features that better relate to the target. This can involve mapping a feature into a new feature using a function like log, or creating a new feature from one or multiple features using multiplication or addition.	CO 4
10	**Explore the term feature selection.** Feature selection is the process of isolating the most consistent, non-redundant, and relevant features to use in model construction. Methodically reducing the size of datasets is important as the size and variety of datasets continue to grow.	CO 2
11	**How are filter method and wrapper methods of feature selection are different?** Wrapper methods measure the "usefulness" of features based on the classifier performance. In contrast, the filter methods pick up the intrinsic properties of the features (i.e., the "relevance" of the features) measured via univariate statistics instead of cross-validation performance.	CO 2
12	**Write short notes on filter based feature selection?** The user requirement(s) document (URD) or user requirement(s) specification (URS) is a document usually used in software engineering that specifies what the user expects the software to be able to do.	CO 2

13	**Which method can be used for feature selection?**	CO 1
	There are two main types of feature selection techniques: supervised and unsupervised, and supervised methods may be divided into wrapper, filter and intrinsic.	
14	**Explore the term wrapper method in python?**	CO 2
	wrappers are the functionality available in Python to wrap a function with another function to extend its behavior. Now, the reason to use wrappers in our code lies in the fact that we can modify a wrapped function without actually changing it. They are also known as decorators.	
15	**Describe the term deciosion tree algorithm?**	CO 2
	A decision tree is a non-parametric supervised learning algorithm, which is utilized for both classification and regression tasks. It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.	
16	**Why are deciosion trees useful?**	CO 2
	Decision trees help you to evaluate your options. Decision Trees are excellent tools for helping you to choose between several courses of action. They provide a highly effective structure within which you can lay out options and investigate the possible outcomes of choosing those options.	
17	**Give the pros and cons of deciosion trees?**	CO 2
	The Advantages of decision trees 1.Good for interpreting data in a highly visual way. 2.Good for handling a combination of numerical and non-numerical data. 3.Requires minimal preparation or data cleaning before use. Disadvantages of decision trees 1.Overfitting (where a model interprets meaning from irrelevant data) can become a problem if a decision tree's design is too complex. 2.They are not well-suited to continuous variables (i.e. variables which can have more than one value, or a spectrum of values).	
18	**Explore the random forest.**	CO 5
	A random forest is a supervised algorithm that uses an ensemble learning method consisting of a multitude of decision trees, the output of which is the consensus of the best answer to the problem. Random Forest can be used for classification or regression.	
19	**Is random forest better than decision tree?**	CO 2
	With that said, random forests are a strong modeling technique and much more robust than a single decision tree. They aggregate many decision trees to limit overfitting as well as error due to bias and therefore yield useful results.	

20	**What is random forest best for?**	CO 3
	Random forests is great with high dimensional data since we are working with subsets of data. It is faster to train than decision trees because we are working only on a subset of features in this model, so we can easily work with hundreds of features.	

MODULE III		
BASIC MATHEMATICS AND PROBABILITY FOR DATA SCIENCE		
1	**Describe the term vector?**	CO 3
	A vector is a quantity or phenomenon that has two independent properties: magnitude and direction. The term also denotes the mathematical or geometrical representation of such a quantity. Examples of vectors in nature are velocity, momentum, force, electromagnetic fields, and weight.	
2	**What are vectors are used for?**	CO 3
	Vectors can be used to represent physical quantities. Most commonly in physics, vectors are used to represent displacement, velocity, and acceleration. Vectors are a combination of magnitude and direction and are drawn as arrows.	
3	**Explain the term matrix?**	CO 3
	A matrix is a 2-dimensional representation of arrays of numbers. Matrices (plural) have two main characteristics that we need to be aware of. The dimension of the matrix, denoted as n x m (n by m), tells us that the matrix has n rows and m columns. Matrices are generally denoted using a capital, bold-faced letter, such as X.	
4	**Explain the term set.**	CO 5
	A set is a well-defined collection of objects. Each object in a set is called an element of the set. Two sets are equal if they have exactly the same elements in them. A set that contains no elements is called a null set or an empty set.	
5	**How do you define a set?**	CO 5
	In Maths, sets are a collection of well-defined objects or elements. A set is represented by a capital letter symbol and the number of elements in the finite set is represented as the cardinal number of a set in a curly bracket ....For example: 1,2,3,4 is a set of numbers.	
6	**Illustrate the term Linear algebra in statistics?**	CO 5
	Linear algebra is about linear combinations. That is, using arithmetic on columns of numbers called vectors and arrays of numbers called matrices, to create new columns and arrays of numbers. Linear algebra is the study of lines and planes, vector spaces and mappings that are required for linear transforms.	
7	**Why do you need linear algebra for statistics?**	CO 3
	You must learn linear algebra in order to be able to learn statistics. Especially multivariate statistics. Statistics and data analysis are another pillar field of mathematics to support machine learning.	

8	**Discuss the probability in statistics?**	CO 3
	The probability is the measure of the likelihood of an event to happen. It measures the certainty of the event. The formula for probability is given by; P(E) = Number of Favourable Outcomes/Number of total outcomes. It is a mathematical tool used to study randomness. It deals with the chance (the likelihood) of an event occurring. For example, if you toss a fair coin four times, the outcomes may not be two heads and two tails.	
9	**How do you find probability in statistics?**	CO 3
	Divide the number of events by the number of possible outcomes. After determining the probability event and its corresponding outcomes, divide the total number of ways the event can occur by the total number of possible outcomes. For instance, rolling a die once and landing on a three can be considered one event.	
10	**Give the five rules of probability?**	CO 4
	-Probability Rule One (For any event A, $0 \leq P(A) \leq 1$). -Probability Rule Two (The sum of the probabilities of all possible outcomes is 1). -Probability Rule Three (The Complement Rule). -Probabilities Involving Multiple Events. -Probability Rule Four (Addition Rule for Disjoint Events).	
11	**Give the difference between Bayesian and frequentist approach for machine learning?**	CO 3
	Both frequentist and Bayesian are statistical approaches to learning from data. But there is a broad distinction between the frequentist and Bayesian. The frequentist learning is only depended on the given data, while the Bayesian learning is performed by the prior belief as well as the given data.	
12	**Explain frequentist model?**	CO 3
	n a frequentist model, probability is the limit of the relative frequency of an event after many trials. In other words, this method calculates the probability that the experiment would have the same outcomes if you were to replicate the same conditions again.	
13	**Why is baysian statistics better than frequentist?**	CO 4
	the frequentist approach assigns probabilities to data, not to hypotheses, whereas the Bayesian approach assigns probabilities to hypotheses. Furthermore, Bayesian models incorporate prior knowledge into the analysis, updating hypotheses probabilities as more data become available.	
14	**Annotate collectively exhaustive event?**	

CO 3	Collectively exhaustive is defined as ensuring all of your survey responses cover the realm of all possible answers a respondent can select. It means the collection of responses are exhaustive of all answer responses.	
15	**How do you know if events are collectively exhaustive?** In probability, a set of events is collectively exhaustive if they cover all of the probability space: i.e., the probability of any one of them happening is 100	CO 3
16	**Are collectively exhaustive are independent?** Two or more events are collectively exhaustive if they cover entire sample space. Two or more events are independent if occurance or failure of one does not affect occurance or failure of other.	CO 3
17	**Explain conditional probability?** Conditional probability is defined as the likelihood of an event or outcome occurring, based on the occurrence of a previous event or outcome. Conditional probability is calculated by multiplying the probability of the preceding event by the updated probability of the succeeding, or conditional, event.	CO 4
18	**Give the difference between probability and conditional probabiliity?** The rule of thumb is that when provided a probability for an event occurring under some condition, you are being presented a conditional probability. Here, ”when a student is absent” is a condition, under which the probability for the event ”student being sick” is being measured.	CO 4
19	**Give the properties of conditional probability?** Property 1: Let E and F be events of a sample space S of an experiment, then we have P(S|F) = P(F|F) = 1. Property 2: f A and B are any two events of a sample space S and F is an event of S such that P(F) 0, then P((A B)|F) = P(A|F) + P(B|F) – P((A B)|F).	CO 4
20	**Explain the term random variable?** A random variable is a numerical description of the outcome of a statistical experiment. A random variable that may assume only a finite number or an infinite sequence of values is said to be discrete; one that may assume any value in some interval on the real number line is said to be continuous.	CO 4

MODULE IV		
STATISTICS FOR DATA SCIENCE		
1	**Is data science basically statistics?** Statistics is a mathematically-based field which seeks to collect and interpret quantitative data. In contrast, data science is a multidisciplinary field which uses scientific methods, processes, and systems to extract knowledge from data in a range of forms.	CO 5
2	**How do you take sampling data?** The most straightforward way to sample data is with simple random sampling. Essentially, the subset is built of observations that were chosen from a larger set purely by chance; Each observation has the same chance of being selected from the larger set. Simple random sampling is extremely simple and easy to implement.	CO 5
3	**Explain the data sampling?** The most straightforward way to sample data is with simple random sampling. Essentially, the subset is built of observations that were chosen from a larger set purely by chance; Each observation has the same chance of being selected from the larger set. Simple random sampling is extremely simple and easy to implement.	CO 5
4	**What do you mean by empirical rule?** The empirical rule, also referred to as the three-sigma rule or 68-95-99.7 rule, is a statistical rule which states that for a normal distribution, almost all observed data will fall within three standard deviations (denoted by) of the mean or average (denoted by µ).	CO 5
5	**Give the purpose of empirical rule?** In most cases, the empirical rule is of primary use to help determine outcomes when not all the data is available. It allows statisticians – or those studying the data – to gain insight into where the data will fall, once all is available. The empirical rule also helps to test how normal a data set is.	CO 4
6	**Illustrate the term point estimates?** point estimation, in statistics, the process of finding an approximate value of some parameter—such as the mean (average)—of a population from random samples of the population.	CO 5
7	**Annonate the functions of point estimates?** point estimation, in statistics, the process of finding an approximate value of some parameter—such as the mean (average)—of a population from random samples of the population.	CO 5

8	**Discuss confidence level.**	CO 4
	A confidence interval is the mean of your estimate plus and minus the variation in that estimate. This is the range of values you expect your estimate to fall between if you redo your test, within a certain level of confidence. Confidence, in statistics, is another way to describe probability.	
9	**How are confidence intervals used ?**	CO 2
	Statisticians use confidence intervals to measure uncertainty in a sample variable. For example, a researcher selects different samples randomly from the same population and computes a confidence interval for each sample to see how it may represent the true value of the population variable. The resulting datasets are all different where some intervals include the true population parameter and others do not.	
10	**Write short notes on hypothesis tests?**	CO 5
	Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution. This assumption is called the null hypothesis and is denoted by H0.	
11	**Give the importance of hypothesis testing?**	CO 5
	Hypothesis testing is used to assess the plausibility of a hypothesis by using sample data. The test provides evidence concerning the plausibility of the hypothesis, given the data. Statistical analysts test a hypothesis by measuring and examining a random sample of the population being analyzed.	
12	**How do you perform hypothesis tests?**	CO 5
	Five Steps in Hypothesis Testing: Specify the Null Hypothesis. Specify the Alternative Hypothesis. Set the Significance Level (a) Calculate the Test Statistic and Corresponding P-Value. Drawing a Conclusion.	
13	**What test is used for hypothesis testing?**	CO 5
	A t-test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population. A t-test looks at the t-statistic, the t-distribution values, and the degrees of freedom to determine the statistical significance.	

14	**How do you know which statistical test to use?** Three criteria are decisive for the selection of the statistical test, which are as follows: the number of variables, types of data/level of measurement (continuous, binary, categorical) and. the type of study design (paired or unpaired).	CO 5
15	**What does measurements mean in statistics?** Statistical measures are a descriptive analysis technique used to summarise the characteristics of a data set. This data set can represent the whole population or a sample of it. Statistical measures can be classified as measures of central tendency and measures of spread.	CO 5
16	**Annotate the term arithmetic mean of data set?** The arithmetic mean of a dataset is found by adding up all of the values and then dividing it by the number of data values.	CO 5
17	**Explore sampling distribution in statistics?** A sampling distribution is a probability distribution of a statistic obtained from a larger number of samples drawn from a specific population. The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.	CO 5
18	**Which sampling distribution should be used and why?** We might use either distribution when standard deviation is unknown and the sample size is very large. We use the t-distribution when the sample size is small, unless the underlying distribution is not normal. The t distribution should not be used with small samples from populations that are not approximately normal.	CO 5
19	**Give the characteristics of sampling distribution?** In general, a sampling distribution will be normal if either of two characteristics is true: the population from which the samples are drawn is normally distributed or. the sample size is equal to or greater than 30.	CO 5
20	**How do sampling distributions relate to hypothesis testing?** When you perform a hypothesis test of a single population mean using a normal distribution (often called a z-test), you take a simple random sample from the population. The population you are testing is normally distributed or your sample size is sufficiently large.	CO 5

MODULE V		
COMMUNICATING DATA		
1	**Write short notes on effective visualization?**	CO 6
	Data visualizations should have a clear purpose and audience. Choose the right type of viz or chart for your data. Use text and labels to clarify, not clutter. Use color to highlight important information or to differentiate or compare. Avoid misleading visualizations.	
2	**What makes a visualization ineffective?**	CO 6
	Avoid using colors with negligible contrast. Avoid using too many colors. Avoid using conventional colors to convey opposite meanings. Pay heed to the needs of people who might be colorblind.	
3	**Do you think that data visualization is accurate or misleading Why?**	CO 6
	Misleading data visualizations might be intentional, if the creator has an agenda to promote. Or they might be the result of errors, the creator not understanding the data or the data visualization process, or allowing engaging or even beautiful visual design to get in the way of clear communication.	
4	**Give the elements of effective data visualization? ?**	CO 6
	Successful data visualization will be achieved when the the four elements are present: information, story, goal and visual inform.	
5	**Explain the Simpson's paradox ?**	CO 6
	Simpson's Paradox is a statistical phenomenon where an association between two variables in a population emerges, disappears or reverses when the population is divided into subpopulations.	
6	**Why is Simpson a paradox?**	CO 6
	Simpson's paradox, also called Yule-Simpson effect, in statistics, an effect that occurs when the marginal association between two categorical variables is qualitatively different from the partial association between the same two variables after controlling for one or more other variables.	
7	**Discuss Simpson's paradox and how does it pertain to confounding?**	CO 6
	Simpson's paradox is an extreme condition of confounding in which an apparent association between two variables is reversed when the data are analyzed within each stratum of a confounding variable.	
8	**Illustrate the term Verbal Communication and examples?**	CO 6
	Verbal communication is the use of sounds and words to express yourself, especially in contrast to using gestures or mannerisms (non-verbal communication). An example of verbal communication is saying "No" when someone asks you to do something you don't want to do.	

9	**Describe the importance of verbal communication??**	CO 6
	Verbal communication helps us think.The ability most often used to distinguish humans from other animals is our ability to reason and communicate. With language, we are able to reflect on the past, consider the present, and ponder the future. We develop our memories using language.	
10	**Explain verbal communication skills?**	CO 6
	Verbal communication skills refer to the way you relay a message through words. This can include the way you speak and the way you write. The primary goal of verbal communication is to use language to convey information clearly and concisely.	
11	**Give strategies of presentation?**	CO 6
	5 Strategies to Make a Good Presentation Great Get the structure right. When doing a presentation, preparation is essential. Know your stuff. Next, make sure you know your stuff. Make eye contact. Use more imagery than text. Be the best version of yourself.	
12	**Why does Simpson's Paradox happen?**	CO 6
	Simpson's Paradox happens because disaggregation of the data (e.g., splitting it into subgroups) can cause certain subgroups to have an imbalanced representation compared to other subgroups. This might be due to the relationship between the variables, or simply due to the way that the data has been partitioned into subgroups.	
13	**Write short notes on Correlation ?**	CO 6
	Correlation is a quantitative metric between -1 and 1 that measures how two variables move with each other. If two variables have a correlation close to -1, it means that as one variable increases, the other decreases, and if two variables have a correlation close to +1, it means that those variables move together in the same direction—as one increases, so does the other, and vice versa.	
14	**Explore the term Causation ?**	CO 6
	Causation is the idea that one variable affects another. For example, we can look at two variables: the average hours of TV watched in a day and a 0-100 scale of work performance (0 being very poor performance and 100 being excellent performance). One might expect that these two factors are negatively correlated, which means that as the number of hours of TV watched increases in a 24 hour day, your overall work performance goes down.	

15	**Explain the term histogram?** The histogram is a popular graphing tool. It is used to summarize discrete or continuous data that are measured on an interval scale. It is often used to illustrate the major features of the distribution of the data in a convenient form.	CO 6
16	**Give the benefits of using a histogram?** The main advantages of a histogram are its simplicity and versatility. It can be used in many different situations to offer an insightful look at frequency distribution. For example, it can be used in sales and marketing to develop the most effective pricing plans and marketing campaigns.	CO 6
17	**Detailing the strengths and weaknesses of a histogram??** What are the strengths and weaknesses of a histogram? The strength of a histogram is that it provides an easy-to-read picture of the location and variation in a data set. There are, however, two weaknesses of histograms that you should bear in mind: The first is that histograms can be manipulated to show different pictures.	CO 6
18	**When should you use a histogram instead of a bar graph?** IHistograms are used to show distributions of variables while bar charts are used to compare variables. Histograms plot quantitative data with ranges of the data grouped into bins or intervals while bar charts plot categorical data.	CO 6
19	**Give the difference between graph and histogram?** A bar graph is the graphical representation of categorical data using rectangular bars where the length of each bar is proportional to the value they represent. A histogram is the graphical representation of data where data is grouped into continuous number ranges and each range corresponds to a vertical bar.	CO 6
20	**Annotate the data visualization and why is it used?** Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets.	CO 6

Course Coordinator:
Ms. K. Anjali, Assistant Professor

HOD, CSE(AI & ML)