USCS606: Data Science SEM VI

UNIT I

1. What is data? Explain types of data. Explain different sources of data?

ANS.

> Data – a collection of facts (numbers, words, measurements, observations, etc) that has been translated into a form that computers can process.

Human's vs Machines

> Human-readable (also known as unstructured data) refers to information that only humans can interpret and study, such as an image or the meaning of a block of text.

> If it requires a person to interpret it, that information is human-readable.

> Machine-readable (or structured data) refers to information that computer programs can process.

> A program is a set of instructions for manipulating data and when we take data and apply a set of programs, we get software.

> In order for a program to perform instructions on data, that data must have some kind of uniform structure.

> Types of Data:

I) Personal data

>Personal data is anything that is specific to you. It covers your demographics, your location, your email address and other identifying factors.

II) Transactional data

>Transactional data is anything that requires an action to collect. You might click on an ad, make a purchase, visit a certain web page, etc.

III) Web data

> Web data is a collective term which refers to any type of data you might pull from the internet, whether to study for research purposes or otherwise.

IV) Sensor data:

> Sensor data is produced by objects and is often referred to as the Internet of Things.

> There can be two types of data:

- Primary Data
- Qualitative Data
- Quantitative Data
- Secondary Data
- Internal Data
- External Data

2. Explain data, information and knowledge triangle.

ANS:

>Data is defined as the collection of facts and details like text, figures, observations, symbols or simply description of things, event or entity gathered with a view to drawing inferences.

>It is the raw fact, which should be processed to gain information.

>It is the unprocessed data that contains numbers, statements and characters before it is refined by the researcher

>The term data is derived from Latin term 'datum' which refers to 'something given'.

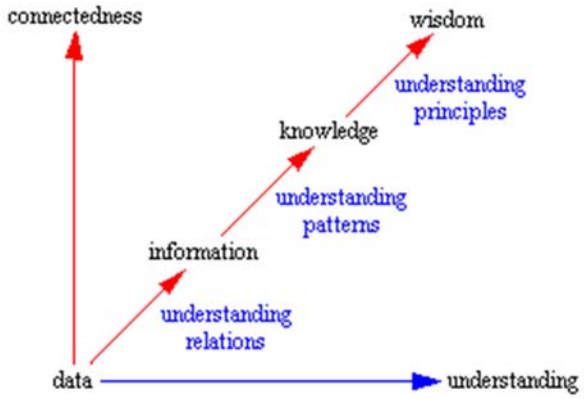
>The concept of data is connected with scientific research, which is collected by various organisations, government departments, institutions and non-government agencies for a variety of reasons.

>Information is described as that form of data which is processed, organized, and specific and structured, which is presented in the given setting.

>It assigns meaning and improves the reliability of the data, thus ensuring understandability and reduces uncertainty.

>When the data is transformed into information, it is free from unnecessary details or immaterial things, which has some value to the researcher.

>The term information discovered from the Latin word ‘informare’, which refers to ‘give form to’.
>Raw data is not at all meaningful and useful as information.
>It is refined and cleaned through purposeful intelligence to become information.
>Therefore data is manipulated through tabulation, analysis and similar other operations which enhance the explanation and interpretation.



I) Data
> Data is raw. It simply exists and has no significance beyond its existence (in and of itself).
>It can exist in any form, usable or not.
>It does not have meaning of itself.
>In computer parlance, a spreadsheet generally starts out by holding data.

II) Information
> Information is data that has been given meaning by way of relational connection.
>This "meaning" can be useful, but does not have to be.
>In computer parlance, a relational database makes information from the data stored within it.

III) Knowledge
> Knowledge is the appropriate collection of information, such that it's intent is to be useful.
>Knowledge is a deterministic process.
>When someone "memorizes" information (as less-aspiring test-bound students often do), then they have amassed knowledge.
>This knowledge has useful meaning to them, but it does not provide for, in and of itself, an integration such as would infer further knowledge

IV) Understanding
> Understanding is an interpolative and probabilistic process.
>It is cognitive and analytical.
>It is the process by which I can take knowledge and synthesize new knowledge from the previously held knowledge.
>The difference between understanding and knowledge is the difference between "learning" and "memorizing".
>People who have understanding can undertake useful actions because they can synthesize new knowledge, or in some cases, at least new information, from what is previously known (and understood).
>That is, understanding can build upon currently held information, knowledge and understanding itself.

V) Wisdom
> Wisdom is an extrapolative and non-deterministic, non-probabilistic process.

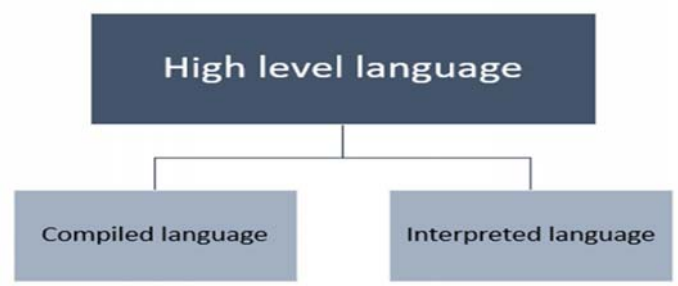
>It calls upon all the previous levels of consciousness, and specifically upon special types of human programming (moral, ethical codes, etc.).
>It beckons to give us understanding about which there has previously been no understanding, and in doing so, goes far beyond understanding itself.
>It is the essence of philosophical probing.
>Unlike the previous four levels, it asks questions to which there is no (easily-achievable) Answer, and in some cases, to which there can be no humanly-known Answer period.
>Wisdom is therefore, the process by which we also discern, or judge, between right and wrong, good and bad

3. Explain the difference between Data and Information.
ANS :

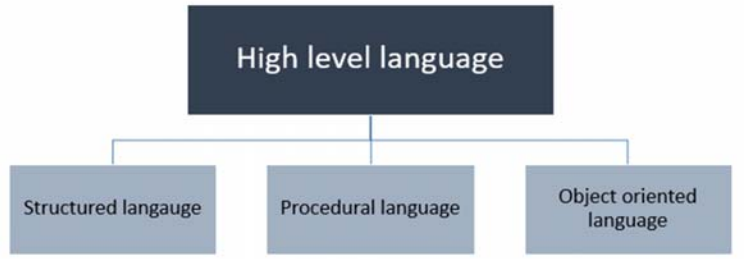
BASIS FOR COMPARISON	DATA	INFORMATION
Meaning	Data means raw facts gathered about someone or something, which is bare and random.	Facts, concerning a particular event or subject, which are refined by processing is called information.
What is it?	It is just text and numbers.	It is refined data.
Based on	Records and Observations	Analysis
Form	Unorganized	Organized
Useful	May or may not be useful.	Always
Specific	No	Yes
Dependency	Does not depend on information.	Without data, information cannot be processed.

4. Write a short note High Level Languages.
ANS:
>A high-level language (HLL) is a programming language such as C,
>ForTran or Pascal that enables a programmer to write programs that are more or less independent of a particular type of computer.
> Such languages are considered high-level because they are closer to human languages and further from machine languages.
>In contrast, assembly languages are considered low-level because they are very close to machine languages.
>High level language provides higher level of abstraction from machine language.
>They do not interact directly with the hardware.
>Rather, they focus more on the complex arithmetic operations, optimal program efficiency and easiness in coding.
>High level programs require compilers/interpreters to translate source code to machine language.

>We can compile the source code written in high level language to multiple machine languages.
>Thus, they are machine independent language.
>High level languages are grouped in two categories based on execution model – compiled or interpreted languages.



>We can also classify high level language several other categories based on programming paradigm



>Advantages of High level language
I) High level languages are programmer friendly. They are easy to write, debug and maintain.
II) It provide higher level of abstraction from machine languages.
III) It is machine independent language.
IV) Easy to learn.
V) Less error prone, easy to find and debug errors.
VI) High level programming results in better programming productivity.

>Disadvantages of High level language
I) It takes additional translation times to translate the source to machine code.
II) High level programs are comparatively slower than low level programs.
III) Compared to low level programs, they are generally less memory efficient.
IV) Cannot communicate directly with the hardware.

5. Explain the components of IDE.
ANS:
>An integrated development environment (IDE)
>An integrated development environment (IDE) is a software suite that consolidates basic tools required to write and test software.
>Developers use numerous tools throughout software code creation, building and testing. Development tools often include text editors, code libraries, compilers and test platforms.
>Without an IDE, a developer must select, deploy, integrate and manage all of these tools separately.
>An IDE brings many of those development-related tools together as a single framework, application or service.

>The integrated toolset is designed to simplify software development and can identify and minimize coding mistakes.

>Common features of integrated development environments
I) An IDE typically contains a code editor, a compiler or interpreter, and a debugger, accessed through a single GUI(GUI).
II) The user writes and edit source code in the code editor.
III) The compiler trAnslates the source code into a readable language that is executable for a computer and the debugger tests the software to solve any issues or bugs.

>Benefits of using IDEs
I) An IDE can improve the productivity of software developers thanks to fast setup and standardization across tools.
II) Without an IDE, developers spend time deciding what tools to use for various tasks, configuring the tools and learning how to use them.
III) Many or even all of the necessary dev-test tools are included in one integrated development environment.
IV) IDEs are also designed with all their tools under one user interface.
V) An IDE can standardize the development process by organizing the necessary features for software development in the UI.

>Types of IDEs and available tools
I) Developers must match the IDE they use with the type of application they want to produce. For example, if a developer wants to create an application on iOS, then they need an IDE that supports Apple's swiftprogramming language. Types of IDEs range from web-based and cloud-based to mobile, language-specific or multi-language.
II) Web-based IDEs suit web-based application development in HTML, Javascript or similar programming languages. Microsoft's Visual Studio Code is an example of a web-based IDE with features such as a code editor, syntax highlighting, code completion and debugging.
More popular IDE tools include Net Beans, Eclipse.

6. What is EDA? Explain tools of EDA.
ANS :
> In statistics Exploratory Data Analysis (EDA) is an approach analyzingdata sets to summarize their main characteristics, often with visual methods
>Exploratory Data Analysis (EDA) is the first step in your data analysis process.
>Here, you make sense of the data you have and then figure out what questions you want to ask and how to frame them, as well as how best to manipulate your available data sources to get the Answers you need.
>What is EDA?" and to tackle specific tasks such as:
I) Spotting mistakes and missing data;
II) Mapping out the underlying structure of the data.
III) Identifying the most important variables.
IV) Listing anomalies and outliers.
V) Testing a hypotheses / checking assumptions related to a specific model.
VI) Establishing a parsimonious model (one that can be used to explain the data with minimal predictor variables).

VII) Estimating parameters and figuring out the associated confidence intervals or margins of error.
>Tools and Techniques:
I) To conduct exploratory data analysis are S-Plus and R.
II) The latter is a powerful, versatile, open-source programming language that can be integrated with many BI platforms.
>Typical graphical techniques used in EDA are:
I) Box plot
II) Histogram
III) Multi-vari chart
(See Answer below)

7. What is EDA and data visualization? Explain methods to visualize data.
ANS :
>In statistics Exploratory Data Analysis (EDA) is an approach analyzing data to summarize their main characteristics, often with visual methods
>Exploratory Data Analysis (EDA) is the first step in your data analysis process.
>Here, you make sense of the data you have and then figure out what questions you want to ask and how to frame them, as well as how best to manipulate your available data sources to get the Answers you need.
>Data visualization is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data.
>To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools.
>Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message.
>Effective visualization helps users analyze and reason about data and evidence.
>It makes complex data more accessible, understandable and usable.
>Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task.
>There are many types of visualizations. Some of the most famous are: **line plot**, **scatter plot**, **histogram**, **box plot**, **bar chart**, and **pie chart**
I) Line Plot
>A type of plot which displays information as a series of data pointscalled “markers” connected by straight lines.
>In this type of plot, we need the measurement points to be ordered (typically by their x-axis values).
>This type of plot is often used to visualize a trend in data over intervals of time - a time series.
II) Scatter plot
> This type of plot shows all individual data points.
>Here, they aren’t connected with lines.
>Each data point has the value of the x-axis value and the value from the y-axis values.
>This type of plot can be used to display trends or correlations.
III) Histogram
> An accurate representation of the distribution of numeric data.
>To create a histogram, first, we divide the entire range of values into a series of intervals, and second, we count how many values fall into each interval**.**
>The intervals are also called bins**.**
>The bins are consecutive and non-overlapping intervals of a variable.

>They must be adjacent and are often of equal size.

IV) Box plot

> It is also called the box-and-whisker plot: a way to show the distribution of values based on the five-number summary: minimum, first quartile, median, third quartile, and maximum.

>The minimum and the maximum are just the min and max values from our data.

>The median is the value that separates the higher half of a data from the lower half.

>It's calculated by the following steps: order your values, and find the middle one.

>In a case when our count of values is even, we actually have 2 middle numbers, so the median here is calculated by summing these 2 numbers and divide the sum by 2. For example, if we have the numbers 1, 2, 5, 6, 8, 9, your median will be (5 + 6) / 2 = 5,5.

>The first quartile is the median of the data values to the left of the median in our ordered values. For example, if we have the numbers 1, 3, 4, 7, 8, 8, 9, the first quartile is the median from the 1, 3, 4 values, so it's 3.

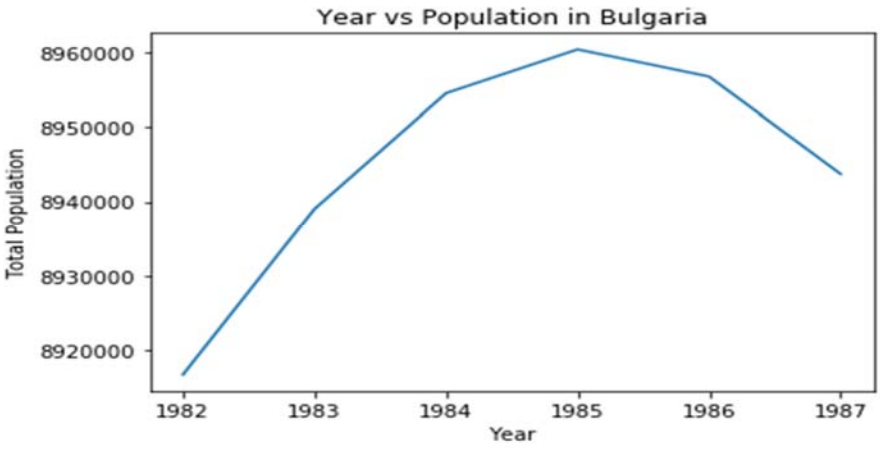
>The third quartile is the median of the data values to the right of the median in our ordered values. For example, if we use these numbers 1, 3, 4, 7, 8, 8, 9 again, the third quartile is the median from the 8, 8, 9values, so it's 8.

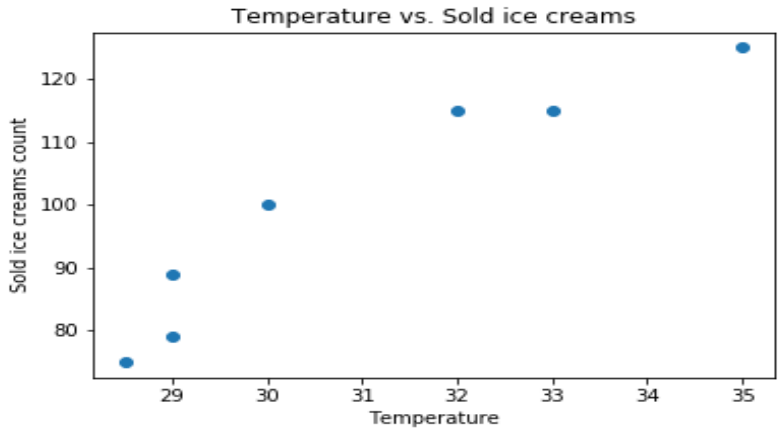
>I also want to mention one more statistic here. That is the IQR (Interquartile Range). The IQR approximates the amount of spread in the middle 50% of the data. The formula is the third quartile - the first quartile.

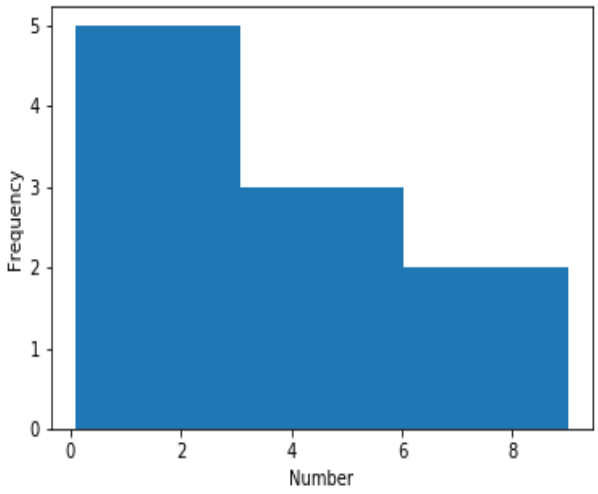
>This type of plot can also show outliers. An outlier is a data value that lies outside the overall pattern.

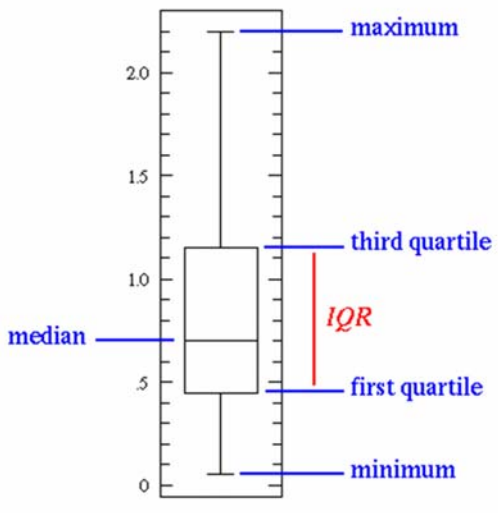
>They are visualized as circles. When we have outliers, the minimum and the maximum are visualized as the min and the max values from the values which aren't outliers.

>There are many ways to identify what is an outlier. A commonly used rule says that a value is an outlier if it's less than the first quartile - 1.5 * IQR or high than the third quartile + 1.5 * IQR.









8. What is primary data? Explain primary data collection methods.
ANS:

>Primary data is data originated for the first time by the researcher through direct efforts and experience, specifically for the purpose of addressing his research problem.
>It is also known as the first hand or raw data.
>Primary data collection is quite expensive, as the research is conducted by the organisation or agency itself, which requires resources like investment and manpower.
>The data collection is under direct control and supervision of the investigator.
>The data can be collected through various methods like surveys, observations, physical testing, mailed questionnaires, questionnaire filled and sent by enumerators, personal interviews, telephonic interviews, focus groups, case studies, etc.

Primary Data Collection Methods

>Primary data collection methods can be divided into two groups:
• Quantitative and • Qualitative.

Quantitative data collection methods:

> They are based in mathematical calculations in various formats.
>Methods of quantitative data collection and analysis include questionnaires with closed-ended questions, methods of correlation and regression, mean, mode and median and others.
>Quantitative methods are cheaper to apply and they can be applied within shorter duration of time compared to qualitative methods. Moreover, due to a high level of standardization of quantitative methods, it is easy to make comparisons of findings.

Qualitative data collection methods:

> On the contrary, do not involve numbers or mathematical calculations.
>Qualitative research is closely associated with words, sounds, feeling, emotions, colours and other elements that are non-quantifiable.
>Qualitative studies aim to ensure greater level of depth of understanding and qualitative data collection methods include interviews, questionnaires with open-ended questions, focus groups, observation, game or role-playing, case studies etc.

Data Collection Methods

>Data collection is a process of collecting information from all the relevant sources to find Answers to the research problem, test the hypothesis and evaluate the outcomes.
>Data collection methods can be divided into two categories:
•Secondary methods of data collection and •Primary methods of data collection.

9. What is secondary data? Explain secondary data collection methods.

ANS:

>Secondary data is a type of data that has already been published in books, newspapers, magazines, journals, online portals etc.
>There is an abundance of data available in these sources about your research area in business studies, almost regardless of the nature of the research area.
>Therefore, application of appropriate set of criteria to select secondary data to be used in the study plays an important role in terms of increasing the levels of research validity and reliability

Data Collection Methods:

>Data collection is a process of collecting information from all the relevant sources to find Answers to the research problem, test the hypothesis and evaluate the outcomes.

>Data collection methods can be divided into two categories: secondary methods of data collection and primary methods of data collection.

Secondary Data Collection Methods

>Secondary data is a type of data that has already been published in books, newspapers, magazines, journals, online portals etc.

>There is an abundance of data available in these sources about your research area in business studies, almost regardless of the nature of the research area.

>Therefore, application of appropriate set of criteria to select secondary data to be used in the study plays an important role in terms of increasing the levels of research validity and reliability

10. Distinguish between primary and secondary data.

ANS:

BASIS FOR COMPARISON	PRIMARY DATA	SECONDARY DATA
Meaning	Primary data refers to the first hand data gathered by the researcher himself.	Secondary data means data collected by someone else earlier.
Data	Real time data	Past data
Process	Very involved	Quick and easy
Source	Surveys, observations, experiments, questionnaire, personal interview, etc.	Government publications, websites, books, journal articles, internal records etc.
Cost effectiveness	Expensive	Economical
Collection time	Long	Short
Specific	Always specific to the researcher's needs.	May or may not be specific to the researcher's need.
Available in	Crude form	Refined form
Accuracy and Reliability	More	Relatively less

11. Write a short note on Data Analysis.

ANS:

>The process of evaluating data using analytical and logical reasoning to examine each component of the data provided.

>This form of analysis is just one of the many steps that must be completed when conducting a research experiment.

>Data from various sources is gathered, reviewed, and then analyzed to form some sort of finding or conclusion.

>There are a variety of specific data analysis method, some of which include data mining, text analytics, business intelligence, and data visualizations.

>Data extraction is where data is analyzed and crawled through to retrieve relevant information from data sources (like a database) in a specific pattern.

>Further data processing is done, which involves adding metadata and other data integration; another process in the data workflow.

>The majority of data extraction comes from unstructured data sources and different data formats.

>This unstructured data can be in any form, such as tables, indexes, and analytics.

12. How to handle missing data in a dataset?

ANS.

>To prevent missing data and dropout, but missing values need to be dealt with.

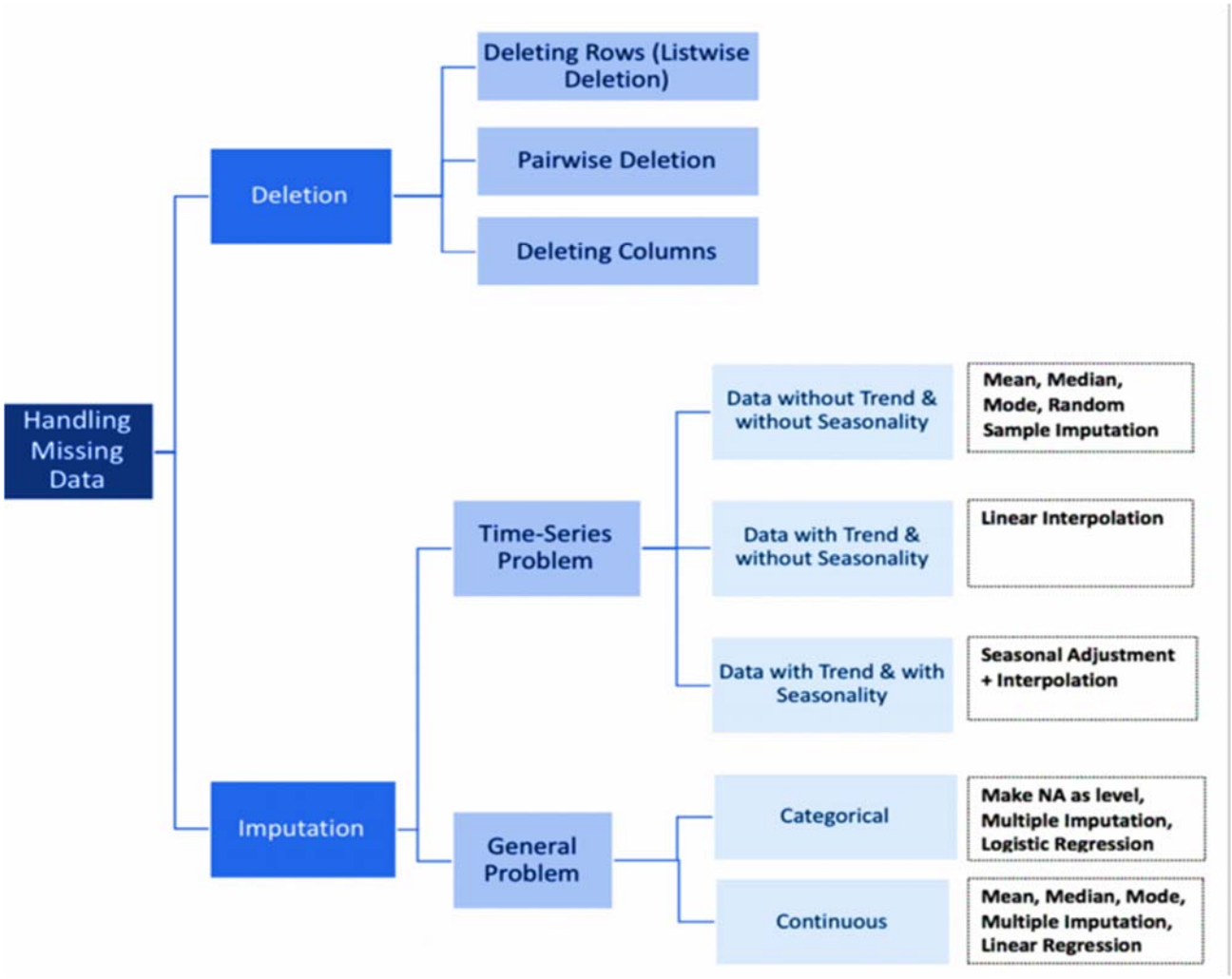
>First, determine the pattern of your missing data.

>There are three types of missing data:

I) Missing Completely at Random: There is no pattern in the missing data on any variables. This is the best you can hope for.

II) Missing at Random: There is a pattern in the missing data but not on your primary dependent variables such as likelihood to recommend or SUS Scores.

III) Missing Not at Random: There is a pattern in the missing data that affect your primary dependent variables. For example, lower-income participants are less likely to respond and thus affect your conclusions about income and likelihood to recommend. Missing not at random isyour worst-case scenario.



13. What is data normalization? Illustrate any one type of data normalization technique with an example.
ANS:

>Data normalization means transforming all variables in the data to a specific range.
>This step is very important when dealing with parameters of different units and scales.
>For example, some data mining techniques use the Euclidean distance.
>Therefore, all parameters should have the same scale for a fair comparison between them.
>Two methods are usually well known for rescaling data.
>*Normalization*, which scales all numeric variables in the range [0,1].
>One possible formula is given below:
Eg. Consider marks: 8, 10, 15, and 20
Min : 8 and Max :20

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

For Marks as 8:

$$MinMax = \frac{(V - Min\ marks)}{Max\ omarks - Min\ marks}(newMax - newMin)^{\square} + newMin$$

$$MinMax = \frac{(8 - 8)}{20 - 8} * (1 - 0)^{\square} + 0$$

$$MinMax = \frac{(0)}{12} * 1$$

$$MinMax = 0$$

For Marks as 10:

$$MinMax = \frac{(10 - 8)}{20 - 8} * (1 - 0)^{\square} + 0$$

$$MinMax = \frac{(2)}{12} * 1$$

$$MinMax = 0.16$$

For Marks as 15:

$$MinMax = \frac{(15 - 8)}{20 - 8} * (1 - 0)^{\square} + 0$$

$$MinMax = \frac{(7)}{12} * 1$$

$$MinMax = \mathbf{0.58}$$

For Marks as 20:

$$MinMax = \frac{(20 - 8)}{20 - 8} * (1 - 0)^{\square} + 0$$

$$MinMax = \frac{(12)}{12} * 1$$

$$MinMax = 1$$

Figure: min max normalization scaling

marks	marks after Min-Max normalization
8	0
10	0.16
15	0.58
20	1

>Other ways of Normalization are

I) Decimal Scaling –

>In this technique, the computation is generally scaled in terms of decimals.

>It means that the result is generally scaled by multiplying or dividing it with pow(10,k).

II) Standard Deviation method –
>In this method, the d.d is normalized by using the formula: [x - mean(x)] * sd(x)

III) By eliminating outliers –
>Outliers are a common sighting while dealing with data.
>Their presence create quite a lot of hassles in the computations.
>So, eliminating them is a very clever idea.
>So, detect your outliers from the box-plots and refine your data by eliminating them.

14. Write a short note on the following smoothing techniques with the help of an example:-
a. Smoothing by bin meAns
b. Smoothing by bin boundaries
ANS:
>Equal-width (distance) partitioning:

- It divides the range into N intervals of equal size: uniform grid
- If A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B-A)/N.
- The most straightforward
- But outliers may dominate presentation
- Skewed data is not handled well.

>Equal-depth (frequency) partitioning:

- It divides the range into N intervals, each containing approximately same number of samples
- Good data scaling
- Managing categorical attributes can be tricky.

Eg.
Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
* Partition into equal-frequency (equi-depth) bins:
- Bin 1: 4, 8, 9, 15 - Bin 2: 21, 21, 24, 25 - Bin 3: 26, 28, 29, 34

* Smoothing y b y bin means:
- Bin 1: 9, 9, 9, 9 - Bin 2: 23, 23, 23, 23 - Bi 3 29 29 29 29 Bin 3: 29, 29, 29, 29

* Smoothing by bin boundaries:
- Bin 1: 4, 4, 4, 15 - Bin 2: 21, 21, 25, 25 - Bin 3: 26, 26, 26, 34

15. Write a short note on data cleaning and data extraction Or Explain the concept of data cleaning. Why is data cleaning needed?

ANS:

>Data cleansing or data cleaning is the process of identifying and removing (or correcting) inaccurate records from a dataset, table, or database and refers to recognising unfinished, unreliable, inaccurate or non-relevant parts of the data and then restoring, remodelling, or removing the dirty or crude data.

>After cleaning, a dataset should be uniform with other related datasets in the operation.

>The discrepancies identified or eliminated may have been basically caused by user entry mistakes, by corruption in storage or transmission, or by various data dictionary descriptions of similar items in various stores.

>With the rise of big data, data cleaning has become more important than ever before.

>Every industry – banking, healthcare, retail, hospitality, education – is now navigating in a large ocean of data and as the data pool is getting bigger, the variables of things going wrong too are getting larger.

>Each fault becomes difficult to find when you can't just look at the whole dataset in a spreadsheet on your computer.

>In fact, this could be true for a variety of reasons.

>The majority of data extraction comes from unstructured data sources and different data formats.

>This unstructured data can be in any form, such as tables, indexes, and analytics.

>Data extraction is where data is analyzed and crawled through to retrieve relevant information from data sources (like a database) in a specific pattern.

>Further data processing is done, which involves adding metadata and other data integration; another process in the data workflow.

16. Write a short note on Data Analysis and Data Modeling.

ANS:

>The process of evaluating data using analytical and logical reasoning to examine each component of the data provided.

>This form of analysis is just one of the many steps that must be completed when conducting a research experiment.

>Data from various sources is gathered, reviewed, and then analyzed to form some sort of finding or conclusion.

>There are a variety of specific data analysis method, some of which include data mining, text analytics, business intelligence, and data visualizations.

>Data modelling is the process of creating a data model for the data to be stored in a Database.

>This data model is a conceptual representation of

- Data objects
- The associations between different data objects
- The rules.

>Data modelling helps in the visual representation of data and enforces business rules, regulatory compliances, and government policies on the data.

>Data Models ensure consistency in naming conventions, default values, semantics, and security while ensuring quality of the data.

>Data model emphasizes on what data is needed and how it should be organized instead of what operations need to be performed on the data.

>Data Model is like architect's building plan which helps to build a conceptual model and set the relationship between data items.

>The two types of Data Models techniques are

1. Entity Relationship (E-R) Model

2. UML (Unified Modelling Language)

17. What is Qualitative data? Explain its types.

ANS:

>Qualitative data is defined as the data that approximates and characterizes.
>Qualitative data can be observed and recorded.
>This data type is non-numerical in nature.
>This type of data is collected through methods of observations, one-to-one interview, conducting focus groups and similar methods.
>Qualitative data in statistics is also known as categorical data.
>Data that can be arranged categorically based on the attributes and properties of a thing or a phenomenon.
>Qualitative data collection is exploratory in nature, it involves in-depth analysis and research.
>Qualitative data collection methods are mainly focused on gaining insights, reasoning, and motivations hence they go deeper in terms of research.
>Since the qualitative data cannot be measured, this leads to the preference for methods or data collection tools that are structured to a limited extent.
>Being Categorical, they can be classified on the basis of their collection methods which are as follows:

1. One-to-One Interviews: One of the most commonly used data collection instrument for qualitative research, mainly because of its personal approach. The interviewer or the researcher collects data directly from the interviewee on a one-to-one basis. The interview may be informal and unstructured – conversational. The questions asked are mostly open-ended questions, spontaneous, with the interviewer letting the flow of the interview dictate the next questions to be asked.

2. Focus groups: This is done in a group discussion setting. The group is limited to 6-10 people and a moderator is assigned to moderate the ongoing discussion. Depending on the data which is sorted, the members of a group may have something in common. For example, a researcher conducting a study on track runners will choose athletes who are track runners or were track runners and have sufficient knowledge of the subject matter.

3. Record keeping: This method makes use of the already existing reliable documents and similar sources of information as the data source. This data can be used in a new research. This is similar to going to a library. There one can go over books and other reference material to collect relevant data that can likely be used in the research.

4. Process of observation: In this qualitative data collection method, the researcher immerses himself/ herself in the setting where his respondents are, and keeps a keen eye on the participants and takes down notes. This is known as the process of observation. Besides taking notes, other documentation methods, such as video and audio recording, photography and similar methods can be used.

5. Longitudinal studies: This data collection method is performed on the same data source repeatedly over an extended period of time. It is an observational research method that goes on for a few years and in some cases can go on for even decades. The goal of this data collection method is to find correlations through an empirical study of subjects with common traits.

6. Case studies: In this method, data is gathered by in-depth analysis of case studies. The versatility of this method is demonstrated in how this method can be used to analyze both simple and complex subjects. The strength of this method is how judiciously it uses a combination of one or more qualitative data collection methods to draw inferences.

18. What is Quantitative data? Explain its types.

ANS:

>Quantitative data is defined as the value of data in the form of counts or numbers where each data-set has a unique numerical value associated with it.
>This data is any quantifiable information that can be used for mathematical calculations and statistical analysis, such that real-life decisions can be made based on these mathematical derivations.
>Quantitative data is used to Answer questions such as "How many?", "How often?", "How much?".
>This data can be verified and can also be conveniently evaluated using mathematical techniques.
>For example, there are quantities corresponding to various parameters, for instance, "How much did that laptop cost?" is a question which will collect quantitative data.
>There are values associated with most measuring parameters such as pounds or kilograms for weight, dollars for cost etc.
>Quantitative data makes measuring various parameters controllable due to the ease of mathematical derivations they come with.
>Quantitative data is usually collected for statistical analysis using surveys, polls or questionnaires sent across to a specific section of a population.
>The retrieved results can be established across a population

TYPES

>The most common types of quantitative data are as below:

I) Counter: Count equated with entities. For example, the number of people who download a particular application from the App Store.

II) Measurement of physical objects: Calculating measurement of any physical thing. For example, the HR executive carefully measures the size of each cubicle assigned to the newly joined employees.

III) Sensory calculation: Mechanism to naturally "sense" the measured parameters to create a constant source of information. For example, a digital camera converts electromagnetic information to a string of numerical data.

IV) Projection of data: Future projection of data can be done using algorithms and other mathematical analysis tools. For example, a marketer will predict an increase in the sales after launching a new product with thorough analysis.

V) Quantification of qualitative entities: Identify numbers to qualitative information. For example, asking respondents of an online survey to share the likelihood of recommendation on a scale of 0-10.

19. Describe the various types of data collection methods.

ANS:

>There are 2 types of data. Discussed below are the types of data.

I) Primary Data –

> It refers to the data that the investigator collects for the very first time.
>This type of data has not been collected either by this or any other investigator before.
>A primary data will provide the investigator with the most reliable first-hand information about the respondents.
>The investigator would have a clear idea about the terminologies uses, the statistical units employed, the research methodology and the size of the sample.
>Primary data may either be internal or external to the organization.

II) Secondary Data –

> It refers to the data that the investigator collects from another source.
>Past investigators or agents collect data required for their study.
>The investigator is the first researcher or statistician to collect this data.
>Moreover, the investigator does not have a clear idea about the intricacies of the data.

>There may be ambiguity in terms of the sample size and sample technique.
>There may also be unreliability with respect to the accuracy of the data.

20. Describe the types of observational methods used in data collection.
ANS:
>There are different types of observation method of data collection in research.
> The important one's are listed below:
1. Casual and Scientific Observation
>An observation may be either casual or scientific.
>Casual observation occurs without any previous preparation.
>It is a matter of chance that the right thing is observed at the right time and in the right place.
>Scientific observation, on the other hand, is carried out with due preparations and is done with the help of right tools of measurement experienced enumerators and under able guidance.
>Scientific observations yield thorough and accurate data.

2. Simple and Systematic Observation
>An observation may be either Simple or Systematic.
> Simple Observation is found in almost all research studies, at least in the initial stages of exploration.
>Its practice is not very standardized.
>It befits the heuristic nature of exploratory research.
>Participant studies are also usually classified as simple observation because participant roles do not permit systematic observation.
>Systematic observation, on the other hand, employs standardized procedures, training of observers, schedules for recording and other devices to control the observer sometimes even the subject.
>Clearly some systematization is valuable in research observation, but the situation often limits what can be done.
>A systematic observation is a scientific observation too.

3. Subjective and Objective Observations
>An observation may be either Subjective or Objective.
>In every act of observation there are two components namely, the object (or what is observed) and the subject (or the observer). It may be that some times one may have to observe one's own immediate experience. That is called Subjective Observation or Self-Observation or introspection.
> Prejudices and biases are generally parts of subjective observation. Many data of psychological interest are gathered by the method of subjective observation. To avoid such prejudices, the observer takes stock of him and discovers what prejudices and biases will prevent impartial study and disinterested points of view. Persistent self-observation and criticism by others may ultimately overcome prejudice and biases. Such introspection may have another social value i.e., it sensitizes the observer to the problems of others and creates sympathetic insight which facilitates, at least to some degree, the understanding of people's behavior in similar circumstances and similar cultural contexts.
> The net result is impartial subjective observation. When the observer is an entity apart from the thing observed, the observation of this type is objective.

4. Factual and Inferential Observation

>Observation may be either factual or inferential.
>In factual observation things or phenomena observed with naked eyes are reported.
>In inferential observation behavior or psychological aspects are observed.

5. Direct and Indirect Observation
>Observation may be either Direct or Indirect.
>In the case of direct Observation the observer is physically present and personally monitors what takes place.
>This approach is very flexible of events and behavior as they occur.
>He is also free to shift places, change the focus of the observation, on concentrate unexpected events if they should occur.
>In indirect observation recording is done by mechanical, photographic or electronic means.
>For example a special motion picture camera which takes one frame every second is mounted in a department of a large store to study customer and employee movement.

6. Behavioral and Non-behavioral Observations
>Observation may be either behavioral or non-behavioral.
>The concept of observation involves not only watching but also listening and reading.
>Thus, observation includes the full range of monitoring behavioral and non-behavioral activities and conditions.
>Non-verbal analysis, linguistic analysis, extra-linguistic analysis and spatial analysis are the four major categories of behavioral observational study of persons.
>Record analysis, physical condition analysis and physical process analysis are the three major categories of non-behavioral study of persons.

21. Explain the process of Web crawling.
ANS:
(refer to ans no 8 in UNIT II)

22. Discuss the importance of scatter plot in data analysis. How it can be viewed in R?
ANS:
>Scatterplots show many points plotted in the Cartesian plane.
>Each point represents the values of two variables.
>One variable is chosen in the horizontal axis and another in the vertical axis.
>The simple scatterplot is created using the plot() function.
>The basic syntax for creating scatterplot in R is –

- plot(x, y, main, xlab, ylab, xlim, ylim, axes)

>Following is the description of the parameters used −

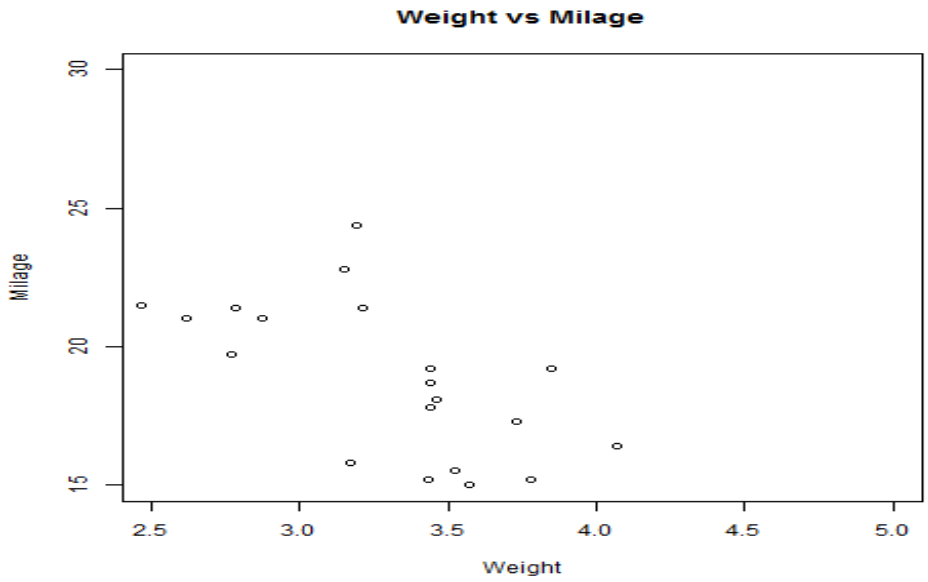
- x is the data set whose values are the horizontal coordinates.
- y is the data set whose values are the vertical coordinates.
- main is the tile of the graph.
- xlab is the label in the horizontal axis.
- ylab is the label in the vertical axis.
- xlim is the limits of the values of x used for plotting.
- ylim is the limits of the values of y used for plotting.

- axes indicates whether both axes should be drawn on the plot.

Ex. # Get the input values.
input<- mtcars[,c('wt','mpg')]

# Give the chart file a name.
png(file = "scatterplot.png")

Plot the chart for cars with weight between 2.5 to 5 and mileage between 15 and 30.
plot(x = input$wt,y = input$mpg,
xlab = "Weight",
ylab = "Milage",
xlim = c(2.5,5),
ylim = c(15,30),
main = "Weight vs Milage"
)
Save the file



23. Write short notes on the following data visualization techniques: Explain wrt to R
a. Line chart
b. Dendrograms
ANS:
>A line chart is a graph that connects a series of points by drawing line segments between them.
>These points are ordered in one of their coordinate (usually the x-coordinate) value.
>Line charts are usually used in identifying the trends in data.
>The plot() function in R is used to create the line graph.
Syntax

>The basic syntax to create a line chart in R is −

- plot(v,type,col,xlab,ylab)

>Following is the description of the parameters used −

- v is a vector containing the numeric values.
- type takes the value "p" to draw only the points, "l" to draw only the lines and "o" to draw both points and lines.
- xlab is the label for x axis.
- ylab is the label for y axis.
- main is the Title of the chart.
- col is used to give colors to both the points and lines.

Example

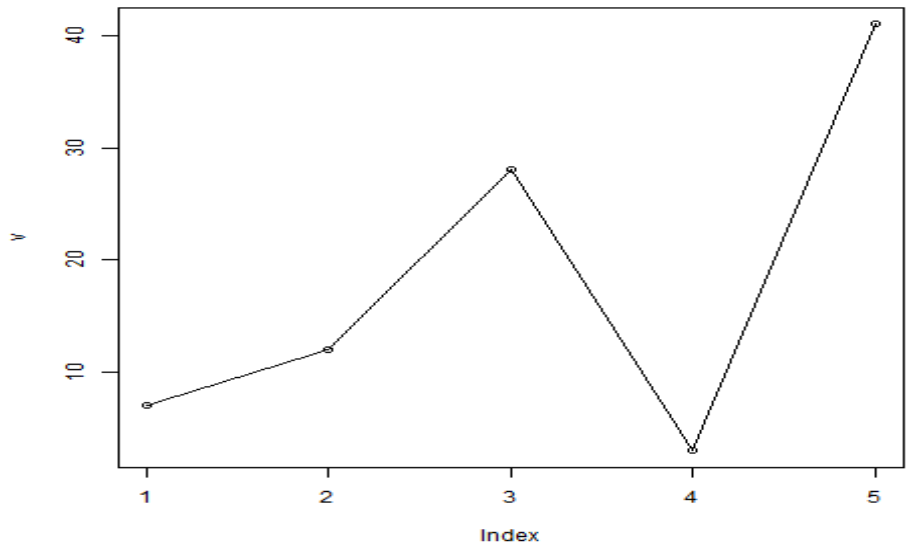
A simple line chart is created using the input vector and the type parameter as "O". The below script will create and save a line chart in the current R working directory.

```
# Create the data for the chart.
v <- c(7,12,28,3,41)

# Give the chart file a name.
png(file = "line_chart.jpg")

# Plot the bar chart.
plot(v,type = "o")
```

When we execute the above code, it produces the following result



b) Dendograms

>A dendrogram is a type of tree diagram showing hierarchical clustering — relationships between similar sets of data.

>They are frequently used in biology to show clustering between genes or samples, but they can represent any type of grouped data

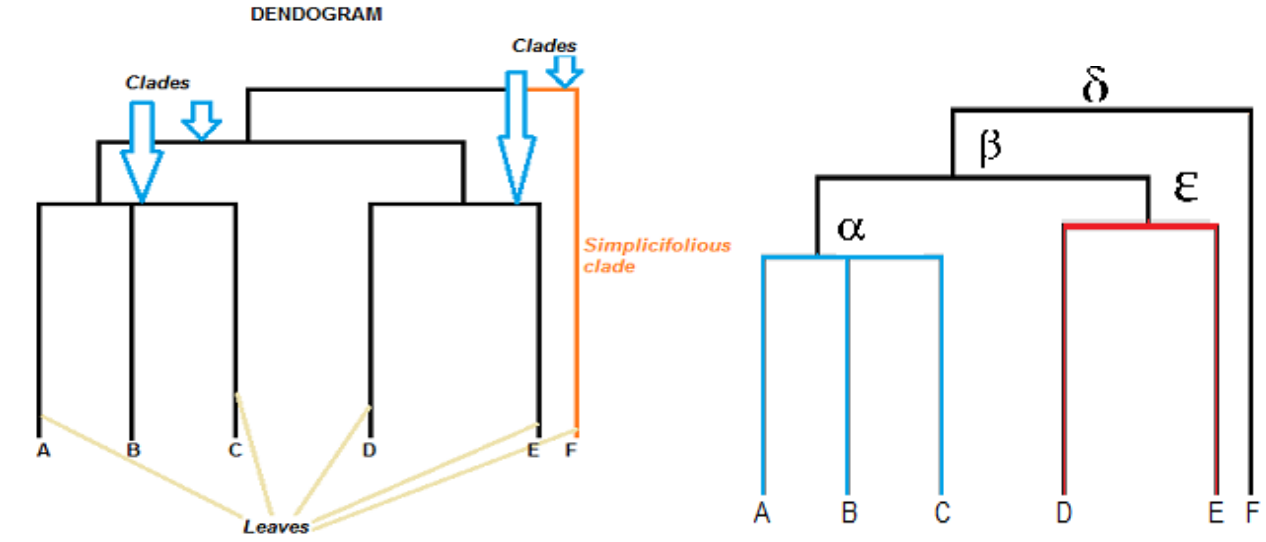
>A dendogram can be a column graph (as in the image below) or a row graph.
>Some dendograms are circular or have a fluid-shape, but software will usually produce a row or column graph.
>No matter what the shape, the basic graph comrpises of the same parts:

- The *clade* is the branch. Usually labeled with Greek letters from left to right (e.g. α β, δ…).
- Each clade has one or more *leaves*. The leaves in the image below are:
    - Single (simplicifolius): F
    - Double (bifolius): D E
    - Triple (trifolious): A B C

>To read a dendogram:
The clades are arranged according to how similar (or dissimilar) they are.
Clades that are close to the same height are similar to each other; clades with different heights are dissimilar — **the greater the difference in height, the more dissimilarity** (you can measure similarity in many different ways; One of the most popular measures is Pearson's Correlation Coefficient).

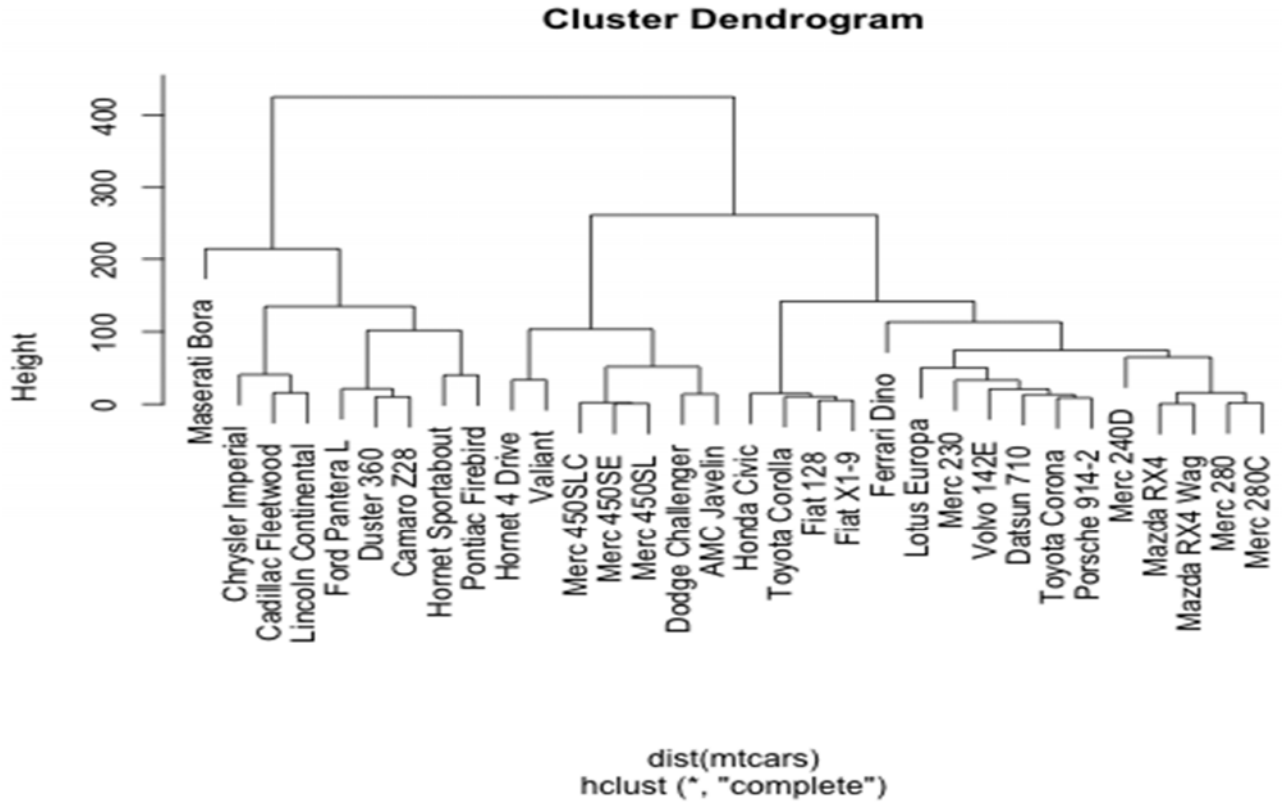
- Leaves A, B, and C are more similar to each other than they are to leaves D, E, or F.
- Leaves D and E are more similar to each other than they are to leaves A, B, C, or F.
- Leaf F is substantially different from all of the other leaves.

Note that on the graph below, the same clave, β joins leaves A, B, C, D,and E. That means that the two groups (A,B,C& D,E) are more similar to each other than they are to F.



.
For the most basic type of dendrogram, we will use the mtcars dataset and we'll calculate a hierarchical clustering with the function hclust (with the default options).

```
# prepare hierarchical cluster
hc = hclust(dist(mtcars))
# very simple dendrogram
plot(hc)
```



24. What is a Box plot? Describe the process to identify an outlier with Box plot. Explain wrt to R.
ANS.
>Boxplots are a measure of how well distributed is the data in a data set. It divides the data set into three quartiles. This graph represents the minimum, maximum, median, first quartile and third quartile in the data set. It is also useful in comparing the distribution of data across data sets by drawing boxplots for each of them.
>Boxplots are created in R by using the boxplot() function.

The basic syntax to create a boxplot in R is −
boxplot(x, data, notch, varwidth, names, main)
Following is the description of the parameters used −

- x is a vector or a formula.
- data is the data frame.
- notch is a logical value. Set as TRUE to draw a notch.
- varwidth is a logical value. Set as true to draw width of the box proportionate to the sample size.
- names are the group labels which will be printed under each boxplot.
- main is used to give a title to the graph.

>We use the data set "mtcars" available in the R environment to create a basic boxplot. Let's look at the columns "mpg" and "cyl" in mtcars.

```
input <- mtcars[,c('mpg','cyl')]
print(head(input))
```

When we execute above code, it produces following result –

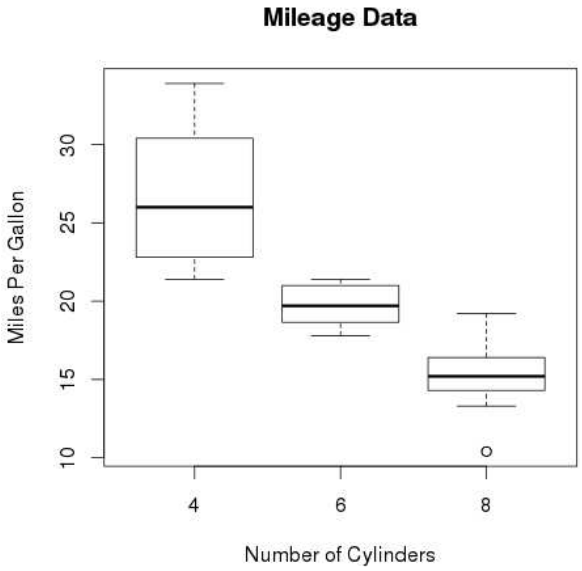
```
                   mpg  cyl\
Mazda RX4         21.0   6
Mazda RX4 Wag     21.0   6
Datsun 710        22.8   4
Hornet 4 Drive    21.4   6
Hornet Sportabout 18.7   8
Valiant           18.1   6
```

The below script will create a boxplot graph for the relation between mpg (miles per gallon) and cyl (number of cylinders).

Live Demo

```
# Give the chart file a name.
png(file = "boxplot.png")

# Plot the chart.
boxplot(mpg ~ cyl, data = mtcars, xlab = "Number of Cylinders",
   ylab = "Miles Per Gallon", main = "Mileage Data")
```



25. Distinguish between structured and unstructured data.

ANS:

	Structured Data	Unstructured Data
Characteristics	• Pre-defined data models • Usually text only • Easy to search	• No pre-defined data model • May be text, images, sound, video or other formats • Difficult to search
Resides in	• Relational databases • Data warehouses	• Applications • NoSQL databases • Data warehouses • Data lakes
Generated by	Humans or machines	Humans or machines
Typical applications	• Airline reservation systems • Inventory control • CRM systems • ERP systems	• Word processing • Presentation software • Email clients • Tools for viewing or editing media
Examples	• Dates • Phone numbers • Social security numbers • Credit card numbers • Customer names • Addresses • Product names and numbers • Transaction information	• Text files • Reports • Email messages • Audio files • Video files • Images • Surveillance imagery

26. Explain data management.

ANS:

>Data management is the practice of organizing and maintaining data processes to meet ongoing information lifecycle needs.

>Emphasis on data management began with the electronics era of data processing, but data management methods have roots in accounting, statistics, logistical planning and other disciplines that predate the emergence of corporate computing in the mid-20th century.

>It is an administrative process that includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data for its users.

>Organizations and enterprises are making use of Big Data more than ever before to inform business decisions and gain deep insights into customer behavior, trends, and opportunities for creating extraordinary customer experiences.

>The best way to manage data, and eventually get the insights needed to make data-driven decisions, is to begin with a business question and acquire the data that is needed to answer that question.

>A few data management best practices organizations and enterprises should strive to achieve include:

- Simplify access to traditional and emerging data
- Scrub data to infuse quality into existing business processes
- Shape data using flexible manipulation techniques

>Managing your data is the first step toward handling the large volume of data, both structured and unstructured, that floods businesses daily.

>It is only through data management best practices that organizations are able to harness the power of their data and gain the insights they need to make the data useful.

>In fact, data management via leading data management platforms enables organizations and enterprises to use data analytics in beneficial ways, such as:

- Personalizing the customer experience
- Adding value to customer interactions
- Identifying the root causes of marketing failures and business issues in real- time
- Reaping the revenues associated with data-driven marketing
- Improving customer engagement
- Increasing customer loyalty

UNIT II

1. What is Data Curation?

ANS:

>Data curation is the management of data throughout its lifecycle.i.e from its initial phase of creation and initial storage to the time when it is archived for posterity or becomes obsolete and is deleted.

>The main purpose of data curation is to ensure that data is reliably retrievable for future research purposes or reuse.

>It provides a graphical, high-level overview of the stages required for successful curation and preservation of data from initial conceptualization or receipt through the iterative curation cycle.

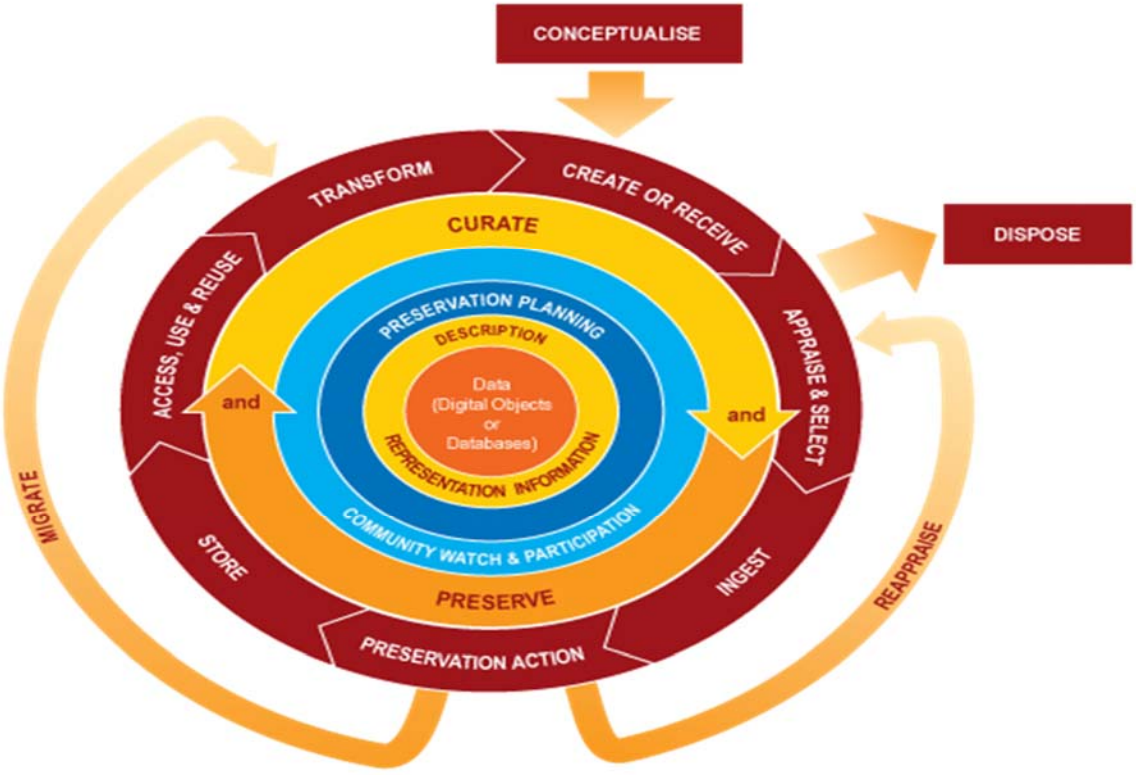
>In the modern era of big data the curation of data has become more prominent, particularly for software processing high volume and complex data systems.

>In broad terms, curation meAns a range of activities and processes done to create, manage, maintain, and validate a component.

>Specifically, data curation is the attempt to determine what information is worth saving and for how long.

2. Explain with the help of a diagram the Data Curation Life Cycle.

ANS:



>Data curation consists of 3 types of action

I) Full lifecycle actions:

- **Description and Representation Information**

  > Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term.

>Collect and assign representation information required to understand and render both the digital material and the associated metadata.

- **Preservation Planning**
  >Plan for preservation throughout the curation lifecycle of digital material.
  >This would include plAns for management and administration of all curation lifecycle actions.

- **Community Watch and Participation**
  > Maintain a watch on appropriate community activities, and participate in the development of shared standards, tools and suitable software.
-
- **Curate and Preserve**
  >Be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle.

II) Sequential actions

- **Conceptualize**
  > Conceive and plan the creation of data, including capture method and storage options.

- **Create or Receive**
  >Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation.
  >Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centers, and if required assign appropriate metadata.

- **Appraise and Select**
  > Evaluate data and select for long-term curation and preservation. Adhere to documented guidance, policies or legal requirements.

- **Ingest**
  > TrAnsfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.

- **Preservation Action**
  > Undertake actions to ensure long-term preservation and retention of the authoritative nature of data.
  > Preservation actions should ensure that data remains authentic, reliable and usable while maintaining its integrity.
  >Actions include data cleaning, validation, assigning preservation metadata, assigning representation information and ensuring acceptable data structures or file formats.

- **Store**
  > Store the data in a secure manner adhering to relevant standards.

- **Access, Use and Reuse**
  > Ensure that data is accessible to both designated users and reusers, on a day-to-day basis.
  > This may be in the form of publicly available published information.
  >Robust access controls and authentication procedures may be applicable

- **TrAnsform**
  > Create new data from the original, for example:

a) by migration into a different format, or
b) by creating a subset, by selection or query, to create newly derived results, perhaps for publication

III) Occasional actions

- **Dispose**
  > Dispose of data, which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements.
  >Typically data may be trAnsferred to another archive, repository, data center or other custodian. In some instances data is destroyed. The data's nature may, for legal reasons, necessitate secure destruction.

- **Reappraise**
  > Return data which fails validation procedures for further appraisal and re-selection.

- **Migrate**
  >Migrate data to a different format. This may be done to accord with the storage environment or to ensure the data's immunity from hardware or software obsolescence.

3. Write about the 5 V's of data.

ANS:

•Volume
→ Large amounts of data generated every second (emails, twitter messages, videos, sensor data...)
• Velocity
→ The speed of data moving in and out data management systems (videos going viral...)
→ "on-the-fly"
• Variety
→ Different data formats in terms of structured or unstructured (80%) data
• Value
→ Insights we can reveal within the data
• Veracity
→ Trustworthiness of the data

3. Write a short note on structured data and unstructured data. Write their sources

ANS.

Unstructured data

>Unstructured data is a data that is which is not organized in a pre-defined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database.
>So for Unstructured data, there are alternative platforms for storing and managing, it is increasingly prevalent in IT systems and is used by organizations in a variety of business intelligence and analytics applications.

>Example: Word, PDF, Text, Media logs.
>Characteristics of Unstructured Data

- It is not based on Schema
- It is not suitable for relational database
- 90% of unstructured data is growing today
- It includes digital media files, Word doc. ,pdf files,
- It is stored in NoSQL database

Structured data

>Structured data includes mainly text, these data are easily processed.
>These data are easily entered, stored and analyzed. Structured data are stored in the form of rows and columns which is easily managed with the a language called "structured query language"
>Relational model is a data model that supports structured data and manage it in the form of row and table and process the content of the table easily.
>XML also supports structured data.
>Most of the content of the web pages are in the XML forms. These content are included in structured data, companies like Google uses structured data to find on the web to understand the content of the page.
>This way most of the Google search is done with the help of structured data.
>Since starting of the revolution of database, network, hierarchical, relational, object relational data model deals with structured data.
>Structured data is a data whose elements are addressable for effective analysis.
>It has been organized into a formatted repository that is typically a database.
>It concerns all data which can be stored in database SQL in table with rows and columns.
>They have relational key and can easily mapped into pre-designed fields.
>Today, those data are most processed in development and simplest way to manage information. Example: Relational data.
>Characteristics of structured data:

- Structured data has various data type: date, name, number, characters, address
- These data are arranged in a defined way
- Structured data are handle through SQL
- Structured data are dependent on schema, it is a schema based
- These data can easily interact with computer

>Sources of structured data:

- Computer- or machine-generated: Machine-generated data generally refers to data that is created by a machine without human intervention.
- Human-generated: This is data that humAns, in interaction with computers, supply

5. Explain Semi-structured data in detail.
ANS:
Semi- Structured Data

>Semi-structured data is information that does not reside in a rational database but that have some organizational properties that make it easier to analyze.
>Can sometimes, be stored in the relational database
>Semi-structured data includes e-mails, XML and JSON. Semi structured data is not fit for relational database where it is expressed with the help of edges, labels and tree structures.

>These are represented with the help of trees and graphs and they have attributes, labels. These are schema-less data.
>Data models which are graph based can store semi-structured data. MongoDB is a NOSQL model that support JSON (semi-structured data).
>Data consist of tags and which are self-describing are generally semi-structured data. They are different from structured and unstructured data.
>Data object Model, Objects Exchange Model, Data Guide are famous data model that express semi structured data.
>Concepts for semi-structured data model

- document instance,
- document schema,
- elements attributes,
- elements relationship sets.

>Characteristics of Semi-structured Data:

- It is not based on Schema
- It is represented through label and edges
- It is generated from various web pages
- It has multiple attributes

>Example:

- Semi-structured data falls in the middle between structured and unstructured data. It contains certain aspects that are structured, and others that are not.
- For example, X-rays and other large images consist largely of unstructured data – in this case, a great many pixels.
- It is impossible to search and query these X-rays in the same way that a large relational database can be searched, queried and analyzed.
- After all, all you are searching against are pixels within an image. Fortunately, there is a way around this.
- Although the files themselves may consist of no more than pixels, words or objects, most files include a small section known as metadata.
- This opens the door to being able to analyze unstructured data.

6. Distinguish between Structured, Semi-Structured and Unstructured Data.
ANS:

PROPERTIES	STRUCTURED DATA	SEMI-STRUCTURED DATA	UNSTRUCTURED DATA
Technology	It is based on Relational database table	It is based on XML/RDF	It is based on character and binary data
Transaction management	Matured transaction and various concurrency technique	Transaction is adapted from DBMS not matured	No transaction management and no concurrency
Version management	Versioning over tuples,row,tables	Versioning over tuples or graph is possible	Versioned as whole
Flexibility	It is sehema dependent and less flexible	It is more flexible than structuded data but less than flexible than unstructured data	it very flexible and there is abbsence of schema
Scalability	It is very difficult to scale DB schema	It's scaling is simpler than sstructured data	It is very scalable
Robustness	Very robust	New technology, not very spread	–
Query performance	Structured query allow complex joining	Queries over anonymous nodes are possible	Only textual query are possible

7. What is data Authentication and Authorization?

ANS:

>While authentication verifies the user's identity, authorization verifies that the user in question has the correct permissions and rights to access the requested resource.

Authentication

>Authentication is used by a system when the system needs to know exactly who is accessing their information or site.

>Used by a client when the user needs to know that the server is system it claims to be.

>The user or computer has to prove its identity to the system or user.

>Usually, authentication by a system entails the use of a user name and password. Other ways can be through cards, retina scAns, voice recognition, and fingerprints.

>Authentication by a client usually involves the system giving a certificate to the system in which a trusted third party which states that the system belongs to the entity (such as a bank) that the system expects it to.

>Does not determine what tasks the individual can do or what files the individual can see.

>Identifies and verifies who the person or system is.

Authorization

>Authorization is a process by which a System determines if the user has permission to use a resource or access a file.

>Authorization is usually coupled with authentication so that the system has some concept of who the user is that is requesting access.

>The type of authentication required for authorization may vary; passwords may be required in some cases but not in others.

>In some cases, there is no authorization; any user may be use a resource or access a file simply by asking for it.

8. What is Web Crawling?

ANS:

>Web crawling makes use of web crawlers.

>A web crawler (also known as a web spider or web robot) is a program or automated script which browses the World Wide Web in a methodical, automated manner.

>This process is called Web crawling or spidering.

>Many legitimate sites, in particular search engines, use spidering as a meAns of providing up-to-date data.

>Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.

>Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code.

>Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for spam).

>Web scraping, web harvesting, or web data extraction is data scraping used for extracting data from websites.

>It is a technique for converting the data present in unstructured format (HTML tags) over the web to the structured format which can easily be accessed and used.

>A web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when you view the page).

>Therefore, web crawling is a main component of web scraping, to fetch pages for later processing.

>Once fetched, then extraction can take place. The content of a page may be parsed, searched, reformatted, its data copied into a spreadsheet, and so on.

>There are several ways of scraping data from the web. Some of the popular ways are:

I) Human Copy-Paste: This is a slow and efficient way of scraping data from the web. This involves humans themselves analyzing and copying the data to local storage.

II) Text pattern matching: Another simple yet powerful approach to extract information from the web is by using regular expression matching facilities of programming languages.

III) API Interface: Many websites like Facebook, Twitter, LinkedIn, etc. provides public and/ or private APIs which can be called using standard code for retrieving the data in the prescribed format.

IV) DOM Parsing: By using the web browsers, programs can retrieve the dynamic content generated by client-side scripts. It is also possible to parse web pages into a DOM tree, based on which programs can retrieve parts of these pages.

>Web scraping can be used to get current prices for the current market scenario, and e-commerce more generally. We will use web scraping to get the data from an e-commerce site.

>In R, rvest is one popular package to do web crawling.

>Let's implement it and see how it works. We will scrape the Amazon website for the price comparison of a product called 64 gb pendrive of sandisk
>For example - Step 1 -

```
library(xml2)
library(rvest)
library(stringr)
url<-https://www.amazon.in/SanDisk-Cruzer-Blade-Flash-Drive/dp/B00BX5FOCK/ref=sr_1_1?s=electronics&ie=UTF8&qid=1552494518&sr=1-1&keywords=64+gb+pen+drive
webpage<- read_html(url)
```

9. Write a note on GitHub.
ANS:
>GitHub is a web-based version-control and collaboration platform for software developers.
>GitHub, which is delivered through a software-as-a-service (SaaS) business model, was started in 2008 and was founded on Git, an open source code management system created by Linus Torvalds to make software builds faster.
>GitHub is used to store the source code for a project and track the complete history of all changes to that code.
>It allows developers to collaborate on a project more effectively by providing tools for managing possibly conflicting changes from multiple developers.
>GitHub allows developers to change, adapt and improve software from its public repositories for free, but it charges for private repositories, offering various paid plAns.
>Each public or private repository contains all of a project's files, as well as each file's revision history.
>Repositories can have multiple collaborators and can be either public or private.
>Three important terms used by developers in GitHub are fork, pull request and merge.
>A fork, also known as a branch, is simply a repository that has been copied from one member's account to another member's account.
>Forks and branches allow a developer to make modifications without affecting the original code.
>If the developer would like to share the modifications, she can send a pull request to the owner of the original repository.
>If, after reviewing the modifications, the original owner would like to pull the modifications into the repository, she can accept the modifications and merge them with the original repository.
>Commits are, by default, all retained and interleaved onto the master project, or can be combined into a simpler merge via commit squashing.

10. What is MongoDB? State its features.
ANS:
>MongoDB (www.mongodb.com) is the project name for the "hu(mongo)us database" system.
>It is maintained by a company called 10gen as open source and is freely available under the GNU AGPL v3.0 license.
>Commercial licenses with full support are available from 10gen (www.10gen.com).
>MongoDB is growing in popularity and may be a good choice for the data store supporting your big data implementation.

>MongoDB is composed of databases containing “collections.” A collection is composed of “documents,” and each document is composed of fields.
>Just as in relational databases, you can index a collection. Doing so increases the performance of data lookup.
>Unlike other databases, however, MongoDB returns something called a “cursor,” which serves as a pointer to the data.
>This is a very useful capability because it offers the option of counting or classifying the data without extracting it.
>Natively, MongoDB supports BSON, the binary implementation of JSON documents.
>These are some important features of MongoDB:

1. Support ad hoc queries
   - In MongoDB, you can search by field, range query and it also supports regular expression searches.
2. Indexing
   - You can index any field in a document.
3. Replication
   - MongoDB supports Master Slave replication.
   - A master can perform Reads and Writes and a Slave copies data from the master and can only be used for reads or back up (not writes)
4. Duplication of data
   - MongoDB can run over multiple servers.
   - The data is duplicated to keep the system up and also keep its running condition in case of hardware failure.

5. Load balancing
   - It has an automatic load balancing configuration because of data placed in shards.
6. Supports map reduce
   - MapReduce to support analytics and aggregation of different collections/documents.
7. Uses JavaScript instead of Procedures.
8. Provides high performance.
9. Stores files of any size easily without complicating your stack.
10. Easy to administer in the case of failures.
11. A sharding service:
    - A sharding service that distributes a single database across a cluster of servers in a single or in multiple data centers.
    - The service is driven by a shard key.
    - The shard key is used to distribute documents intelligently across multiple instances.
12. A Grid-based file system (GridFS)
    - Enables the storage of large objects by dividing them among multiple documents.

11. What is MongoDB? State its advantages over RDBMS.
ANS:
>Refer the above question for the Answer of “what is MongoDB”
>MongoDB advantages over DBMS:

- In recent days, MongoDB is a new and popularly used database. It is a document based, non-relational database provider.

- Although it is 100 times faster than the traditional database but it is early to say that it will broadly replace the traditional RDBMS. But it may be very useful in term to gain performance and scalability.
- A Relational database has a typical schema design that shows number of tables and the relationship between these tables, while in MongoDB there is no concept of relationship.

>Performance analysis of MongoDB and RDBMS

- In relational database (RDBMS) tables are using as storing elements, while in MongoDB collection is used.
- In the RDBMS, we have multiple schema and in each schema we create tables to store data while, MongoDB is a document oriented database in which data is written in BSON format which is a JSON like format.
- MongoDB is almost 100 times faster than traditional database systems.

12. How to create, use, show and delete databases in MongoDB? Give example.

ANS:

>There is no create database command in MongoDB.

>Actually, MongoDB do not provide any command to create database.

>It may be look like a weird concept, if you are from traditional SQL background where you need to create a database, table and insert values in the table manually.

>Here, in MongoDB you don't need to create a database manually because MongoDB will create it automatically when you save the value into the defined collection at first time.

>You also don't need to mention what you want to create; it will be automatically created at the time you save the value into the defined collection.

>If there is no existing database, the following command is used to create a new database. If the database already exists, it will return the existing database

->use DATABASE_NAME

Example:

>use javatpointdb

Swithched to db javatpointdb

To check the database list, use the command show dbs:

Example:

```
>show dbs
```

javatpointdb 0.078GB

The dropDatabase command is used to drop a database. It also deletes the associated data files. It operates on the current database.

```
db.dropDatabase()
```

This syntax will delete the selected database. In the case you have not selected any database, it will delete default "test" database.

13. What is a collection in MongoDB? Give two different examples of creating collections

ANS:

>MongoDB stores data in the form of documents, which are stored in collections, while collections are stored in the database.

>A collection can store a number of documents that might not be the same in structure.

>Since MongoDB is a Schema-free DB, a collection can store documents with varying structures.

>Compared to other RDBMS (Relational Database Management Systems) such as MySQL in which the structure or organization of data is defined by a schema, MongoDB allows storing documents of different structures in the same collection.
>This meAns users don't have to define columns and its datatype.
>Two MongoDB documents might belong to the same collection, but can store different data structures.
>In MongoDB, db.createCollection(name, options) is used to create collection also MongoDB creates collection automatically when you insert some documents.

Syntax:
db.createCollection(name, options)
Here,
Name: is a string type, specifies the name of the collection to be created.
Options: is a document type, specifies the memory size and indexing of the collection. It is an optional parameter

Example
>use test
Switched to db test
>db.createCollection("SSSIT")
{"ok": 1 }

>MongoDB creates collections automatically when you insert some documents.
>For example: Insert a document named "seomount" into a collection named SSSIT. The operation will create the collection if the collection does not currently exist.
>db.SSSIT.insert({"name":"seomount"})
>show collections
SSSIT

14. How can you see data stored in MongoDB? Explain any two methods with example.
ANS:
>One can see data stored in MongoDB using 2 methods:
1.Show collections command
Example:
>db.SSSIT.insert({"name" : "seomount"})
>show collections
SSSIT
2. find command
Syntax:
db.collection_name.find(collection name)

15. What is MongoDB? Write commands for the following:
ANS :Refer to Ans 14 of unit II.
>Explain find function of MongoDB.

- db.collection_name.find((collection name)

>How can you update information in MongoDB?

>In MongoDB, update() method is used to update or modify the existing documents of a collection.
Syntax:

- db.COLLECTION_NAME.update(SELECTIOIN_CRITERIA, UPDATED_DATA)

Example:
Consider an example which has a collection name javatpoint. Insert the following documents in collection:

```
db.javatpoint.insert(
  {
    course: "java",
    details: {
      duration: "6 months",
      Trainer: "Sonoo jaiswal"
    },
    Batch: [ { size: "Small", qty: 15 }, { size: "Medium", qty: 25 } ],
    category: "Programming language"
  }
)
```

After successful insertion, check the documents by following query:

```
>db.javatpoint.find(javatpoint)
```

Output:

```
{ "_id" : ObjectId("56482d3e27e53d2dbc93cef8"), "course" : "java", "details" :
{ "duration" : "6 months", "Trainer" : "Sonoojaiswal" }, "Batch" :
[ {"size" : "Small", "qty" : 15 }, { "size" : "Medium", "qty" : 25 } ],
 "category" : "Programming language" }
```

Update the existing course "java" into "android":

```
>db.javatpoint.update({'course':'java'},{$set:{'course':'android'}})
```

Check the updated document in the collection:

```
>db.javatpoint.find(javatpoint)
```

Output:

```
{ "_id" : ObjectId("56482d3e27e53d2dbc93cef8"), "course" : "android", "details" :
{ "duration" : "6 months", "Trainer" : "Sonoojaiswal" }, "Batch" :
[ {"size" : "Small", "qty" : 15 }, { "size" : "Medium", "qty" : 25 } ],
 "category" : "Programming language"}
```

Give examples of how to delete records in MongoDB.

MongoDB Delete documents
In MongoDB, the db.colloction.remove() method is used to delete documents from a collection. The remove() method works on two parameters.
1. Deletion criteria: With the use of its syntax you can remove the documents from the collection.
2. JustOne: It removes only one document when set to true or 1.

Syntax:
db.collection_name.remove (DELETION_CRITERIA)

Remove all documents
If you want to remove all documents from a collection, pass an empty query document {} to the remove() method. The remove() method does not remove the indexes.
Let's take an example to demonstrate the remove() method. In this example, we remove all documents from the "javatpoint" collection.

db.javatpoint.remove({})

State the use of limit and skip methods.
1) limit() method in the Mongo database
In the Mongo universe, the limit() method specifies the number of documents returned in response to a particular Mongo query i.e. developers can append the limit() method to the db.collection_name.find() query. The cursor.limit() method has the following prototype form:
Mongo database 'limit()' Syntax

```
1    > db.collection_name.find(<query string>).limit(<number>)
```

Where:
The query_string is an optional input argument that retrieves the documents from a collection on the basis of a specified choice criteria
The number is an input integer argument that specifies the maximum number of documents to be returned

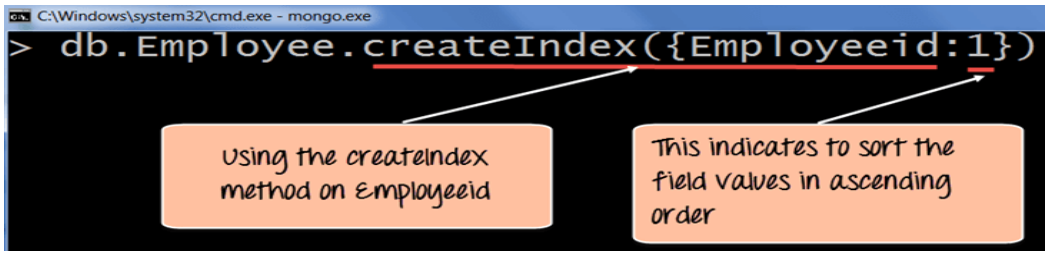
2) skip() method in the Mongo database

In the Mongo universe, the skip() method skips the given number of the documents within the cursor object. In other words, developers can control from where the Mongo database begins returning the results and the cursor.skip() method has the following prototype form:
Mongo database 'skip()'.
Syntax

```
> db.collection_name.find(<query_string>).limit(<number>).skip(<offset>)
```

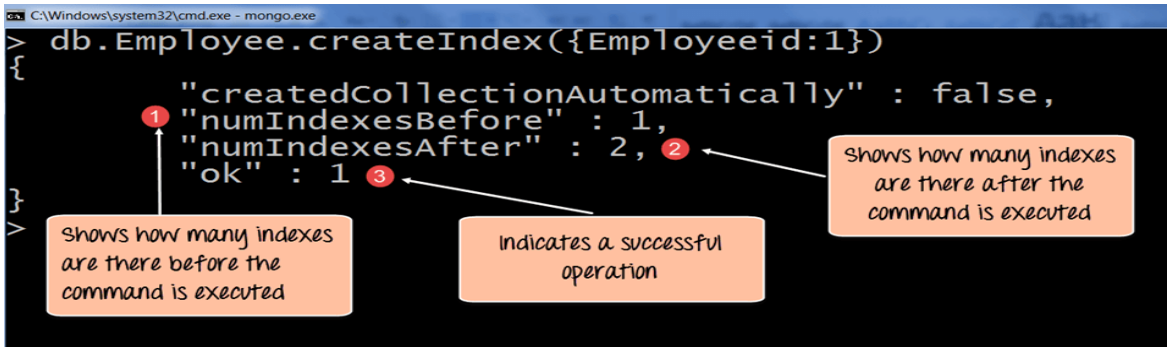
Where:
The query_string is an optional input argument that retrieves the documents from a collection on the basis of a specified choice criteria
The number is an input integer argument that specifies the maximum number of documents to be returned
The offset is an input integer argument that specifies the number of documents to be skipped in a result set

16. How to create indexes in MongoDB? Give example.
ANS:
>Indexes are very important in any database, and with MongoDB it's no different.
>With the use of Indexes, performing queries in MongoDB becomes more efficient.
>Creating an Index in MongoDB is done by using the "createIndex" method.
>The following example shows how add index to collection.
>Let's assume that we have our same Employee collection which has the Field names of "Employeeid" and "EmployeeName".



db.Employee.createIndex({Employeeid:1])

Output:



17. What is NoSQL? What are its features?

ANS:

>NoSQL database stands for "Not Only SQL" or "Not SQL."

>NoSQL is a non-relational DMS that does not require a fixed schema, avoids joins, and is easy to scale.

>NoSQL database is used for distributed data stores with humongous data storage needs.

>NoSQL is used for big data and real-time web apps. For example in companies like Twitter, Facebook, Google

>A NoSQL database system encompasses a wide range of database technologies that can store structured, semi-structured, unstructured and polymorphic data.

>The data structures used by NoSQL databases are different from those used by default in relational databases which makes some operations faster in NoSQL.

>Features of NoSQL

I) Non-relational –

NoSQL databases never follow the relational model and never provides tables with flat fixed-column records.

II) Schema-free –

NoSQL databases are either schema-free or have relaxed schemas, Do not require any sort of definition of the schema of the data, Offers heterogeneous structures of data in the same domain.

III) Simple API –

Offers easy to use interfaces for storage and querying data provided, Web-enabled databases running as internet-facing services.

IV) Distributed –

Multiple NoSQL databases can be executed in a distributed fashion.

V) Offers auto –

> Scaling and fail-over capabilities, Shared Nothing Architecture. This enables less coordination and higher distribution.

18. State advantages of NoSQL over DBMS.
ANS:
>NoSQL databases differ from older, relational technology in four main areas:
I) Data models:
>A NoSQL database lets you build an application without having to define the schema first unlike relational databases which make you define your schema before you can add any data to the system.
>No predefined schema makes NoSQL databases much easier to update as your data and requirements change.
II) Data structure:
>Relational databases were built in an era where data was fairly structured and clearly defined by their relationships.
>NoSQL databases are designed to handle unstructured data (e.g., texts, social media posts, video, and email) which makes up much of the data that exists today.
III) Scaling:
>It's much cheaper to scale a NoSQL database than a relational database because you can add capacity by scaling out over cheap, commodity servers.
>Relational databases, on the other hand, require a single server to host your entire database.
>To scale, you need to buy a bigger, more expensive server.
IV) Development model:
>NoSQL databases are open source whereas relational databases typically are closed source with licensing fees baked into the use of their software.
>With NoSQL, you can get started on a project without any heavy investments in software fees upfront.

21. What is NoSQL? Briefly explain its types
ANS:
>NoSQL database stands for "Not Only SQL" or "Not SQL."
>NoSQL is a non-relational DMS, that does not require a fixed schema, avoids joins, and is easy to scale. NoSQL database is used for distributed data stores with humongous data storage needs.
>NoSQL is used for Big data and real-time web apps. For example in companies like Twitter, Facebook, Google
>NoSQL database system encompasses a wide range of database technologies that can store structured, semi-structured, unstructured and polymorphic data.
>The data structures used by NoSQL databases are different from those used by default in relational databases which makes some operations faster in NoSQL.
>There are 4 basic types of NoSQL databases:
I) Key-Value Store –
>It has a Big Hash Table of keys & values {Example- Riak, Amazon S3 (Dynamo)}
II) Document-based Store –
>It stores documents made up of tagged elements. {Example- CouchDB}
III) Column-based Store –
>Each storage block contains data from only one column, {Example- HBase, Cassandra}

IV) Graph-based–

>A network database that uses edges and nodes to represent and store data. {Example- Neo4J}

22. Explain any one type of NoSQL technology.
ANS:
Key Value Store NoSQL Database
>The schema-less format of a key value database like Riak is just about what you need for your storage needs.
>The key can be synthetic or auto-generated while the value can be String, JSON, BLOB (basic large object) etc.
>The key value type basically, uses a hash table in which there exists a unique key and a pointer to a particular item of data.
>A bucket is a logical group of keys – but they don't physically group the data. There can be identical keys in different buckets.
>Performance is enhanced to a great degree because of the cache mechanisms that accompany the mappings.
>To read a value you need to know both the key and the bucket because the real key is a hash (Bucket+ Key).
>There is no complexity around the Key Value Store database model as it can be implemented in a breeze.
>Not an ideal method if you are only looking to just update part of a value or query the database.
>When we try and reflect back on the CAP theorem, it becomes quite clear that key value stores are great around the Availability and Partition aspects but definitely lack in Consistency.
>Example: Consider the data subset represented in the following table. Here the key is the name of the 3Pillar country name, while the value is a list of addresses of 3PiIllar centers in that country.

Key	Value
"India"	{"B-25, Sector-58, Noida, India – 201301"
"Romania"	{"IMPS Moara Business Center, Buftea No. 1, Cluj-Napoca, 400606",City Business Center, Coriolan Brediceanu No. 10, Building B, Timisoara, 300011"}
"US"	{"3975 Fair Ridge Drive. Suite 200 South, Fairfax, VA 22033"}

>The key can be synthetic or auto-generated while the value can be String, JSON, BLOB (basic large object) etc.
>This key/value type database allow clients to read and write values using a key as follows:
I) Get(key), returns the value associated with the provided key.
II) Put(key, value), associates the value with the key.
III) Multi-get(key1, key2, .., keyN), returns the list of values associated with the list of keys.
IV) Delete(key), removes the entry for the key from the data store.

23. Write a note on data trAnsformation.
ANS:

>In computing, extract, trAnsform, load (ETL) is a process in database usage to prepare data for analysis, especially in data warehousing
>Extract, TrAnsform, Load
>Downloading data programmatically from the Web
>Processing XML and JSON formats
>Scraping and parsing data from raw HTML sources
>Interacting with APIs
>In data transformation process data are transformed from one format to another format that is more appropriate for data mining.
>Some Data Transformation Strategies:-
1 Smoothing
Smoothing is a process of removing noise from the data.
2 Aggregation
Aggregation is a process where summary or aggregation operations are applied to the data.
3 Generalization
In generalization low-level data are replaced with high-level data by using concept hierarchies climbing.

4 Normalization
Normalization scaled attribute data so as to fall within a small specified range, such as 0.0 to 1.0.
5 Attribute Construction
In Attribute construction, new attributes are constructed from the given set of attributes.

24. Explain how you can read JSON file in R with the help of an example.
ANS:
>JSON file stores data as text in human-readable format. Json stands for JavaScript Object Notation. R can read JSON files using the rjson package.
I) Install rjson Package
>In the R console, you can issue the following command to install the rjson package.
>install.packages("rjson")
II) Input Data
>Create a JSON file by copying the below data into a text editor like notepad. Save the file with a .json extension and choosing the file type as all files(*.*).

```
{
"ID":["1","2","3","4","5","6","7","8"],
"Name":["Rick","Dan","Michelle","Ryan","Gary","Nina","Simon","Guru"],
"Salary":["623.3","515.2","611","729","843.25","578","632.8","722.5"],

"StartDate":["1/1/2012","9/23/2013","11/15/2014","5/11/2014","3/27/2015","5/21/2013",
"7/30/2013","6/17/2014"],
"Dept":["IT","Operations","IT","HR","Finance","IT","Operations","Finance"]
}
```

III) Read the JSON File
>The JSON file is read by R using the function from JSON(). It is stored as a list in R.

```
# Load the package required to read JSON files.
```

```
library("rjson")
# Give the input file name to the function.
result<-fromJSON(file ="input.json")
# Print the result.
print(result)
```

25. Explain how you can read XML file in R with the help of an example.

ANS:

>XML is a file format which shares both the file format and the data on the World Wide Web, intranets, and elsewhere using standard ASCII text.

>It stands for Extensible Markup Language (XML).

>Similar to HTML it contains markup tags.

>But unlike HTML where the markup tag describes structure of the page, in xml the markup tags describe the meaning of the data contained into the file.

>You can read a xml file in R using the "XML" package.

>This package can be installed using following command.

```
install.packages("XML")
```

I) Input Data

>Create aXMl file by copying the below data into a text editor like notepad. Save the file with a .xml extension and choosing the file type as all files(*.*).

II) Reading XML File

>The xml file is read by R using the function xmlParse(). It is stored as a list in R.

```
# Load the package required to read XML files.
library("XML")
# Also load the other required package.
library("methods")
# Give the input file name to the function.
result<-xmlParse(file ="input.xml")
# Print the result.
print(result)
```

26. Write a note on XPATH.

ANS:

>XPath is a major element in the XSLT standard.

> It can be used to navigate through elements and attributes in an XML document.

> It stands for XML Path Language

> It uses "path like" syntax to identify and navigate nodes in an XML document

> It contains over 200 built-in functions

> It is a major element in the XSLT standard

> It is a W3C recommendation

>XPath specification specifies seven types of nodes which can be the output of execution of the XPath expression.

I) Root

II) Element

III) Text

IV) Attribute
V) Comment
VI) Processing Instruction
VII) Namespace

27. Explain the following XPATH expressions.

- **a.** /
- **b.** //
- **c.** .
- **d.** ..
- **e.** @

ANS:

S.No.	Expression & Description
1	**node-name** Select all nodes with the given name "nodename"
2	**/** Selection starts from the root node
3	**//** Selection starts from the current node that match the selection
4	**.** Selects the current node
5	**..** Selects the parent of the current node
6	**@** Selects attributes

28. What is web scraping?

ANS:

>Web scraping is a process of automating the extraction of data in an efficient and fast way.

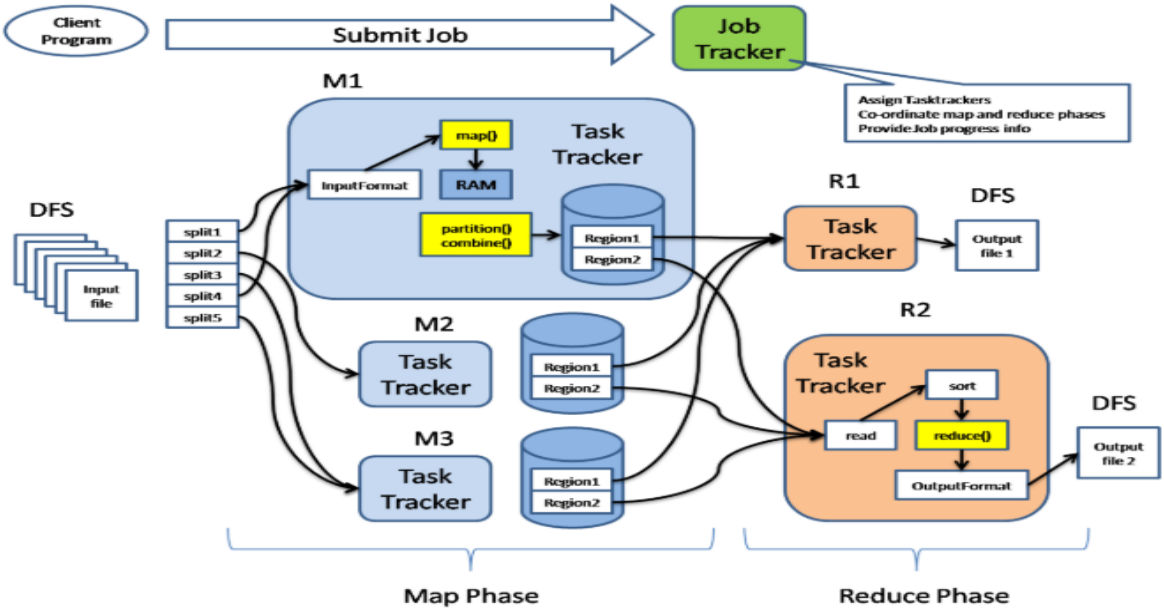
>With the help of web scraping, you can extract data from any website, no matter how large is the data, on your computer.
>Moreover, websites may have data that you cannot copy and paste.
>Web scraping can help you extract any kind of data that you want.
>When you extract web data with the help of a web scraping too, you would be able to save the data in a format such as CSV. You would then be able to retrieve, analyze and use the data the way you want.
> Some web scraping applications are:
I) Competitor Price Monitoring
II) Monitoring MAP Compliance
III) Fetching Images and Product Descriptions
IV) Monitoring Consumer Sentiment

29. Explain various ways to do web scraping.
ANS:
> Different ways to do web scraping are:
I) Human Copy-Paste
>This is a slow and efficient way of scraping data from the web. This involves humAns themselves analyzing and copying the data to local storage.
II) Text pattern matching
>Another simple yet powerful approach to extract information from the web is by using regular expression matching facilities of programming languages.
III) API Interface
>Many websites like Facebook, Twitter, LinkedIn, etc. provides public and/ or private APIs which can be called using standard code for retrieving the data in the prescribed format.
IV) DOM Parsing
>By using the web browsers, programs can retrieve the dynamic content generated by client-side scripts. It is also possible to parse web pages into a DOM tree, based on which programs can retrieve parts of these pages.

30. Explain version control in detail.
ANS:
1. Back up (almost) everything created by a human being as soon as it is created. This includes scripts and programs of all kinds, software packages that your project depends on, and documentation.
2. Keep changes small. Each change should not be so large as to make the change tracking irrelevant. For example, a single change such as Revise script file that adds or changes several hundred lines is likely too large, as it will not allow changes to different components of an analysis to be investigated separately. Similarly, changes should not be broken up into pieces that are too small. As a rule of thumb, a good size for a single change is a group of edits that you could imagine wanting to undo in one step at some point in the future.
3. Share changes frequently. Everyone working on the project should share and incorporate changes from others on a regular basis. Do not allow individual investigator's versions of the project repository to drift apart, as the effort required to merge differences goes up faster than the size of the difference. This is particularly important for the manual versioning procedure describe below, which does not provide any assistance for merging simultaneous, possibly conflicting, changes.

4. Create, maintain, and use a checklist for saving and sharing changes to the project. The list should include writing log messages that clearly explain any changes, the size and content of individual changes, style guidelines for code, updating to-do lists, and bAns on committing half-done work or broken code.
5. Store each project in a folder that is mirrored off the researcher's working machine using a system such as Dropbox or a remote repository such as GitHub. Synchronize that folder at least daily. It may take a few minutes, but that time is repaid the moment a laptop is stolen or its hard drive fails.

31. Write a note on Map Reduce architecture.
ANS:
>MapReduce is mainly used for parallel processing of large sets of data stored in Hadoop cluster.
>Initially, it is a hypothesis specially designed by Google to provide parallelism, data distribution and fault-tolerance. MR processes data in the form of key-value pairs.
>A key-value (KV) pair is a mapping element between two linked data items - key and its value.



>Map reduce architecture consists of mainly two processing stages.
>First one is the map stage and the second one is reduce stage.
>The actual MR process happens in task tracker.
>In between map and reduce stages, Intermediate process will take place.
>Intermediate process will do operations like shuffle and sorting of the mapper output data.
>The Intermediate data is going to get stored in local file system.
I) Mapper Phase
>In Mapper Phase the input data is going to split into 2 components, Key and Value.
>The key is writable and comparable in the processing stage.
>Value is writable only during the processing stage.
>Suppose, client submits input data to Hadoop system, the Job tracker assigns tasks to task tracker.
>The input data that is going to get split into several input splits.
II) Intermediate Process
>The mapper output data undergoes shuffle and sorting in intermediate process.

>The intermediate data is going to get stored in local file system without having replications in Hadoop nodes.
>This intermediate data is the data that is generated after some computations based on certain logics.
>Hadoop uses a Round-Robin algorithm to write the intermediate data to local disk.
>There are many other sorting factors to reach the conditions to write the data to local disks.

III) Reducer Phase

>Shuffled and sorted data is going to pass as input to the reducer.
>In this phase, all incoming data is going to combine and same actual key value pairs is going to write into hdfs system.
>Record writer writes data from reducer to hdfs.
>The reducer is not so mandatory for searching and mapping purpose.

32. Write a note on HBase.

ANS:

>One of the most popular columnar databases is HBase (http://hbase. apache.org).
>It, too, is a project in the Apache Software Foundation distributed under the Apache Software License v2.0.
>HBase uses the Hadoop file system and MapReduce engine for its core data storage needs.
>The design of HBase is modeled on Google's BigTable (an efficient form of storing nonrelational data).
>Therefore, implementations of HBase are highly scalable, sparse, distributed, persistent multidimensional sorted maps.
>The map is indexed by a row key, column key, and a timestamp; each value in the map is an uninterpreted array of bytes.
>When your big data implementation requires random, real-time read/write data access, HBase is a very good solution.
>It is often used to store results for later analytical processing.
>Important characteristics of HBase include the following:

I) Consistency:

>Although not an "ACID" implementation, HBase offers strongly consistent reads and writes and is not based on an eventually consistent model.
>This meAns you can use it for high-speed requirements as long as you do not need the "extra features" offered by RDBMS like full trAnsaction support or typed columns.

II) Sharding:

>Because the data is distributed by the supporting file system, HBase offers trAnsparent, automatic splitting and redistribution of its content.

III) High availability:

>Through the implementation of region servers, HBase supports LAN and WAN failover and recovery.
>At the core, there is a master server responsible for monitoring the region servers and all metadata for the cluster.

IV) Client API:

>HBase offers programmatic access through a Java API.

V) Support for IT operations:

>Implementers can expose performance and other metrics through a set of built-in web pages.

33. Compare HBase& RDBMS in brief.

ANS:

HBase	RDBMS
Column-oriented	Row oriented (mostly)
Flexible schema, add columns on the fly	Fixed schema.
Good with sparse tables,	Not optimized for sparse tables.
Joins using MR –not optimized	Optimized for joins.
Tight integration with MR	Not really...
Horizontal scalability –just add hardware	Hard to shard and scale
Good for semi-structured data as well as Un-structured data	Good for structured data

33. Describe in detail cloud services.
ANS:
Refer question 37.

34. Explain in detail homogeneous distributed database and heterogeneous distributed database.
ANS:
>In a homogeneous distributed database, all the sites use identical DBMS and operating systems.
>Its properties are −
• The sites use very similar software.
• The sites use identical DBMS or DBMS from the same vendor.
• Each site is aware of all other sites and cooperates with other sites to process user requests.
• The database is accessed through a single interface as if it is a single database.
>There are two types of homogeneous distributed database −
• Autonomous − each database is independent that functions on its own. They are integrated by a controlling application and use message passing to share data updates.
• Non-autonomous − Data is distributed across the homogeneous nodes and a central or master DBMS co-ordinates data updates across the sites.

Heterogeneous Database
>In a heterogeneous distributed database, different sites have different operating systems, DBMS products and data models.
>Its properties are −
• Different sites use dissimilar schemas and software.
• The system may be composed of a variety of DBMSs like relational, network, hierarchical or object oriented.
• Query processing is complex due to dissimilar schemas.
• TrAnsaction processing is complex due to dissimilar software.
• A site may not be aware of other sites and so there is limited co-operation in processing user requests.
>There are two types of heterogeneous distributed database −

• Federated − The heterogeneous database systems are independent in nature and integrated together so that they function as a single database system.
• Un-federated − The database systems employ a central coordinating module through which the databases are accessed.

35. Write a short note on Hadoop Architecture. State its advantages.
ANS:
>Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models.
>The Hadoop framework application works in an environment that provides distributed storage and computation across clusters of computers.
> It is designed to scale up from single server to thousands of machines, each offering local computation and storage.
> It has three core components, plus ZooKeeper if you want to enable high availability:
I) Hadoop Distributed File System (HDFS)
> The Hadoop Distributed File System (HDFS) is the underlying file system of a Hadoop cluster.
> It provides scalable, fault-tolerant, rack-aware data storage designed to be deployed on commodity hardware.
II) MapReduce
III) Yet Another Resource Negotiator (YARN)
IV) ZooKeeper

> Advantages of Hadoop are
I) Hadoop framework allows the user to quickly write and test distributed systems. It is efficient, and it automatic distributes the data and work across the machines and in turn, utilizes the underlying parallelism of the CPU cores.
II) Hadoop does not rely on hardware to provide fault-tolerance and high availability (FTHA), rather Hadoop library itself has been designed to detect and handle failures at the application layer.
III) Servers can be added or removed from the cluster dynamically and Hadoop continues to operate without interruption.
IV) Another big advantage of Hadoop is that apart from being open source, it is compatible on all the platforms since it is Java based.

36. Write a short note on AWS.
ANS:
> Many businesses use cloud based services; as a result various companies have started building and providing such services.
>Amazon began the trend, with Amazon Web Services (AWS).
>While AWS began in 2006 as a side business, it now makes $14.5 billion in revenue each year.
>Other leaders in this area include:
I) Google—Google Cloud Platform (GCP)
II) Microsoft—Azure Cloud Services
III) IBM—IBM Cloud
>Cloud services are useful to businesses of all sizes—small companies benefit from the low cost, as compared to buying servers.

>Larger companies gain reliability and productivity, with less cost, since the services run on optimum energy and maintenance.
>These services are also powerful tools that you can use to ease your work.
>Setting up a Hadoop cluster to work with Spark manually could take days if it's your first time, but AWS sets that up for you in minutes.
>We are going to focus on AWS here because it comes with more products relevant to data scientists.
>In general, we can say familiarity with AWS helps data scientists to:
I) Prepare the infrastructure they need for their work (e.g. Hadoop clusters) with ease.
II) Easily set up necessary tools (e.g. Spark).
III) Decrease expenses significantly—such as by paying for huge Hadoop clusters only when needed.
IV) Spend less time on maintenance, as there's no need for tasks like manually backing up data.
V) Develop products and features that are ready to launch without needing help from engineers (or, at least, needing very little help).

UNIT III

1. Explain the general idea of model selection techniques in Machine learning. / List and explain features of Model selection.

ANS:

>Model selection is the process of choosing between different machine learning approaches - e.g. SVM, logistic regression, etc - or choosing between different hyper parameters or sets of features for the same machine learning approach - e.g. deciding between the polynomial degrees/complexities for linear regression.

>The choice of the actual machine learning algorithm (e.g. SVM or logistic regression) is less important than you'd think - there may be a "best" algorithm for a particular problem, but often its performance is not much better than other well-performing approaches for that problem.

>There may be certain qualities you look for in a model:

- Interpretable - can we see or understand why the model is making the decisions it makes?
- Simple - easy to explain and understand
- Accurate
- Fast (to train and test)
- Scalable (it can be applied to a large dataset)

2. Explain the concept of regularization.

ANS:

>Regularizations are techniques used to reduce the error by fitting a function appropriately on the given training set and avoid over fitting.

>Consider the training dataset comprising ofindependent variables X=(x1,x2….xn) and the corresponding target variables t=(t1,t2,…tn).

>X are random variables lying uniformly between [0,1].

>The target dataset 't' is obtained by substituting the value of X into the function sin(2πx) + adding some Gaussian noise into it.

> This is a form of regression that constrains/ regularizes or shrinks the coefficient estimates towards zero.

>This technique discourages learning a more complex or flexible model, so as to avoid the risk of over fitting.

>A simple relation for linear regression looks like this.

$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$

Here Y represents the learned relation and β represents the coefficient estimates for different variables or predictors(X).

>The fitting procedure involves a loss function, known as residual sum of squares or RSS.

>The coefficients are chosen, such that they minimize this loss function.

>Regularization is a technique used for tuning the function by adding an additional penalty term in the error function.

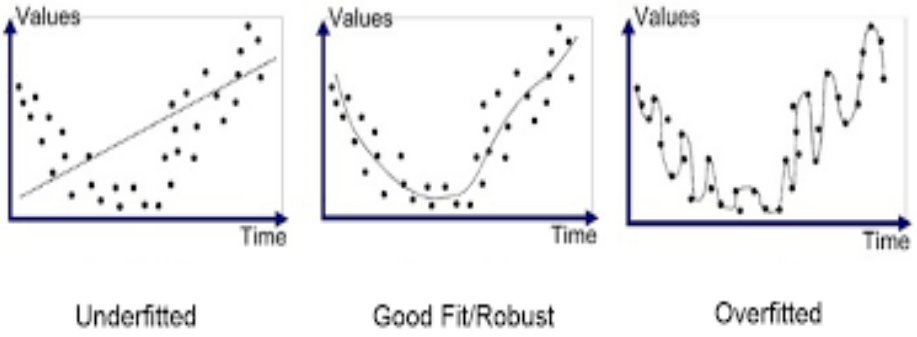
>The additional term controls the excessively fluctuating function such that the coefficients don't take extreme values.

>This technique of keeping a check or reducing the value of error coefficients are called shrinkage methods or weight decay in case of neural networks.

>Overfitting can also be controlled by increasing the size of training dataset.

>One of the major aspects of training your machine learning model is avoiding overfitting.

>The model will have a low accuracy if it is overfitting.
>This happens because your model is trying too hard to capture the noise in your training dataset.
>By noise we mean the data points that don't really represent the true properties of your data, but random chance.
>Learning such data points, makes your model more flexible, at the risk of overfitting.
>The concept of balancing bias and variance, is helpful in understanding the phenomenon of overfitting.



3. What is bias? What is variance? Write a note on bias / variance trade off?
ANS:
>In supervised machine learning an algorithm learns a model from training data.
>The goal of any supervised machine learning algorithm is to best estimate the mapping function (f) for the output variable (Y) given the input data (X).
>The mapping function is often called the target function because it is the function that a given supervised machine learning algorithm aims to approximate.
>The prediction error for any machine learning algorithm can be broken down into three parts:

- Bias Error
- Variance Error
- Irreducible Error

>The irreducible error cannot be reduced regardless of what algorithm is used.
>It is the error introduced from the chosen framing of the problem and may be caused by factors like unknown variables that influence the mapping of the input variables to the output variable.
>Bias are the simplifying assumptions made by a model to make the target function easier to learn.
>Generally, parametric algorithms have a high bias making them fast to learn and easier to understand but generally less flexible.

>In turn, they have lower predictive performance on complex problems that fail to meet the simplifying assumptions of the algorithms bias.

- Low Bias: Suggests less assumptions about the form of the target function.
- High-Bias: Suggests more assumptions about the form of the target function.

>Examples of low-bias machine learning algorithms include: Decision Trees, k-Nearest Neighbours and Support Vector Machines.
>Examples of high-bias machine learning algorithms include: Linear Regression, Linear Discriminant Analysis and Logistic Regression.
>Variance is the amount that the estimate of the target function will change if different training data was used.
>The target function is estimated from the training data by a machine learning algorithm, so we should expect the algorithm to have some variance.
>Ideally, it should not change too much from one training dataset to the next, meaning that the algorithm is good at picking out the hidden underlying mapping between the inputs and the output variables.
>Machine learning algorithms that have a high variance are strongly influenced by the specifics of the training data.
>This meAns that the specifics of the training have influences the number and types of parameters used to characterize the mapping function.

- Low Variance: Suggests small changes to the estimate of the target function with changes to the training dataset.
- High Variance: Suggests large changes to the estimate of the target function with changes to the training dataset.

>Generally, nonparametric machine learning algorithms that have a lot of flexibility have a high variance.
>For example, decision trees have a high variance that is even higher if the trees are not pruned before use.
>Examples of low-variance machine learning algorithms include: Linear Regression, Linear Discriminant Analysis and Logistic Regression.
>Examples of high-variance machine learning algorithms include: Decision Trees, k-Nearest Neighbours and Support Vector Machines.

Trade Off in Variance & Bias

>The goal of any supervised machine learning algorithm is to achieve low bias and low variance. In turn the algorithm should achieve good prediction performance.
>Parametric or linear machine learning algorithms often have a high bias but a low variance.
>Non-parametric or non-linear machine learning algorithms often have a low bias but a high variance.
>The parameterization of machine learning algorithms is often a battle to balance out bias and variance.
>Below are two examples of configuring the bias-variance trade-off for specific algorithms:

- The k-nearest neighbours' algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbours that contribute t the prediction and in turn increases the bias of the model.
- The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.
- There is no escaping the relationship between bias and variance in machine learning.
- Increasing the bias will decrease the variance.
- Increasing the variance will decrease the bias.

- There is a trade-off at play between these two concerns and the algorithms you choose and the way you choose to configure them are finding different balances in this trade-off for your problem. In reality, we cannot calculate the real bias and variance error terms because we do not know the actual underlying target function. Nevertheless, as a framework, bias and variance provide the tools to understand the behaviour of machine learning algorithms in the pursuit of predictive performance.

4. What are AIC, BIC? State their mathematical formula.

ANS:

>Akaike information criterion (AIC) (Akaike, 1974) is a fined technique based on in-sample fit to estimate the likelihood of a model to predict/estimate the future values.

>A good model is the one that has minimum AIC among all the other models.

>Bayesian information criterion (BIC) (Stone, 1979) is another criteria for model selection that measures the trade-off betweenmodel fit and complexity of the model.

>Ideally, A lower AIC or BIC value indicates a better fit.

AIC

>AIC is a statistical model selector for estimating the skill of the given dataset.

>It is a model selector and gives number of models it estimates the quality of each model in comparison with other models and provides with the optimal model.

>If any model is estimated on a particular set of data, AIC score will provide an estimation of that model performance for the new data set.

>AIC is the in-sample error of the estimated model.

>To select the model using AIC, the model which provides smallest AIC value with respect to other is chosen.

>To avoid the risk of overfitting AIC provides penalty by the term 2*d to reduce complexity.

BIC

>Partially based on AIC

>Also known as Schwarz criterion, used for selection of an appropriate model out of many available finite set of models.

>Based upon the likelihood function

>Used to solve the problem of overfitting faced by models it does so by introducing the penalty term for the numbers of parameters or features in the model.

>Used in time series data and linear regression.

>AIC, BIC mathematical formula. The following equations are used to estimate the AIC and BIC of a model:

AIC=-2*lnL+2*k

BIC=-2*lnL+2*lnN*k

where L is the value of the likelihood,

N is the number of recorded measurements,

k is the number of estimated parameters

5. Write a note on cross validation.

ANS:

>Cross-validation, aka rotation estimation/ out-of-sample testing is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.

>Statistical technique to evaluate predictive models by recursively making partitions of the original sample into training sets to train the model and a test set to revaluate it.
>Estimates the skill of machine learning models.
>One round of cross validation involves the partitioning/portioning a sample of data into various subsetsand performs the analysis on one subset also known as training set and validating the analysis on other subsets which is known as the validation of the testing set.

6. What do you mean by - LASSO regression, Ridge Regression?
ANS:
LASSO regression
>LASSO (Least Absolute Shrinkage and Selection Operator or LASSO) is a regression analysis method that performs both variable selection and regularizationin order to enhance the prediction accuracy and interpretability of the statistical model it produces.
>The only difference is instead of taking the square of the coefficients, magnitudes are taken into account.
>This type of regularization (L1) can lead to zero coefficients i.e. some of the features are completely neglected for the evaluation of output.
>So Lasso regression not only helps in reducing over-fitting but it can help us in feature selection.
>Just like Ridge regression the regularization parameter (lambda) can be controlled

Ridge Regression
>Ridge regression is an extension for linear regression. It's basically a regularized linear regression model.
>The λ parameter is a scalar that should be learned as well, using a method called cross validation
>It is a technique used by data that suffers multi co linearity (independent variables are highly correlated.
>In multi co linearity, even though the least squares estimates are unbiased, their variances are large which in turn result in the deviation of the observed value far from the true value
>By adding a degree of bias to the regression estimates, ridge regression is able to reduce standard errors.

7. Write down the difference between Lasso and Ridge regression.
ANS: Refer Ans No-6

8. Explain feature extraction.
ANS:
>TrAnsformation of input data into a set of features.
>Features are distinctive properties of input patterns that help in differentiating between the categories of input patterns. Learn more in: Non-Manual Control Devices: Direct Brain-Computer Interaction
>The process to represent raw image in a reduced form to facilitate decision making such as pattern detection, classification or recognition. Learn more in: Feature Extraction Techniques: Fundamental Concepts and Survey
>A technique that reduces the amount of input data by distilling its representative descriptive attributes. Learn more in: Using Supervised Machine Learning to Explore Energy Consumption Data in Private Sector Housing
>It is a process of deriving new features from the original features in order to reduce the cost of feature measurement, increase classifier efficiency, and allow higher classification accuracy. Learn more in: Big Data Mining Based on Computational Intelligence and Fuzzy Clustering
>TrAnsforming the input data into the set of features is called feature extraction.

>If the features extracted are carefully chosen, it is expected that the features set will perform the desired task using the reduced representation instead of the full size input. Learn more in: Computational Intelligence for Pathological Issues in Precision Agriculture
>Feature extraction refers to the extraction of linguistic items from the documents to provide a representative sample of their content.
>Distinctive vocabulary items found in a document are assigned to the different categories by measuring the importance of those items to the document content. Learn more in: Text Mining
>Finding of representative features of a determined problem from samples with different characteristics. Learn more in: Artificial Intelligence in Computer-Aided Diagnosis
>The process by which a new set of discriminative features is obtained from those available.
>Classification is performed using the new set of features. Learn more in: Component Analysis in Artificial Vision
>This is a process which is used to obtain certain characteristics which are intrinsic and discriminate of a thing. Learn more in: State of the Art in Writer's Off-Line Identification

9. What is Supervised Learning / Unsupervised Learning?
ANS:
Supervised Learning –
>Supervised learning is where you have both input variables (x) and output variables(Y) and can use an algorithm to derive the mapping function from the input to the output.
>Further classified into Classification and Regression.
>Supervised learning is typically done in the context of classification, when we want to map input to output labels, or regression, when we want to map input to a continuous output.
>Common algorithms in supervised learning include logistic regression, naive bayes, support vector machines, artificial neural networks, and random forests.
>In both regression and classification, the goal is to find specific relationships or structure in the input data that allow us to effectively produce correct output data.
>When conducting supervised learning, the main considerations are model complexity, and the bias-variance tradeoff.
>Noisy, or incorrect, data labels will clearly reduce the effectiveness of your model.
Unsupervised Learning –
>Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.
>The most common unsupervised learning method is cluster analysis, which is used for exploratory data analysis to find hidden patterns or grouping in data.
>This method combines highly correlated variables together to form a smaller number of an artificial set of variables which is called “principal components” that account for most variance in the data.
>Principal component analysis is a method of extracting important variables (in form of components) from a large set of variables available in a data set.
>It extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible.

10. Differentiate between Supervised Learning and Unsupervised Learning?
ANS: Refer Ans No- 9

11. What is Regression? Explain its types.

ANS:

>Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable (s)** (predictor).

>This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

>For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

>Regression analysis is an important tool for modelling and analyzing data.

>Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized.

>I'll explain this in more details in coming sections.

>Types of regression are:

1. Linear Regression

>It is one of the most widely known modeling technique.

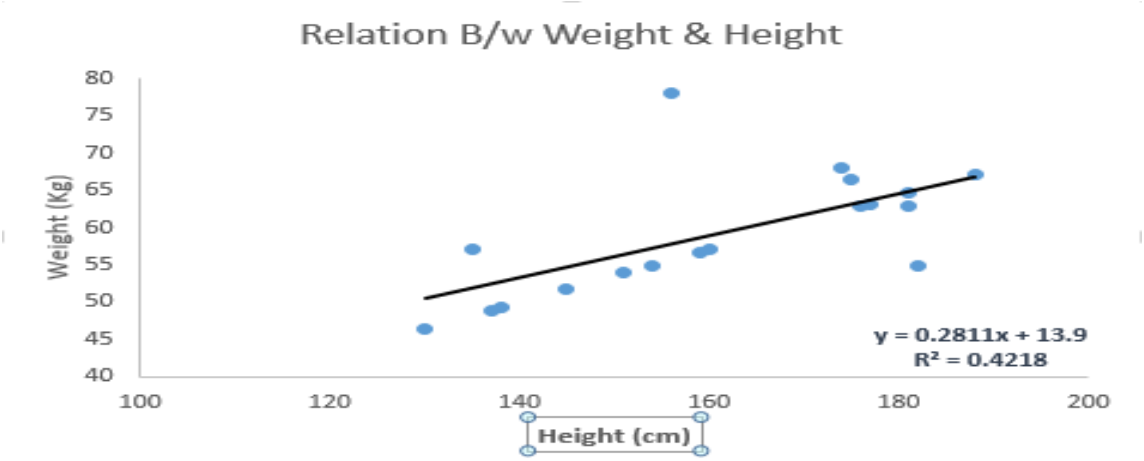
>Linear regression is usually among the first few topics which people pick while learning predictive modeling.

>In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

>Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).

>It is represented by an equation **Y=a+b*X + e**, where a is intercept, b is slope of the line and e is error term.

>This equation can be used to predict the value of target variable based on given predictor variable(s).



>Simple Linear Regression: It is characterized by one independent variable. Consider the price of the house based only one field that is the size of the plot then that would be a simple linear regression.

>Multiple Linear Regression: It is characterized by multiple independent variables. The price of the house if depends on more that one like the size of the plot area, the economy then it is considered as multiple linear regression which is in most real-world scenarios.

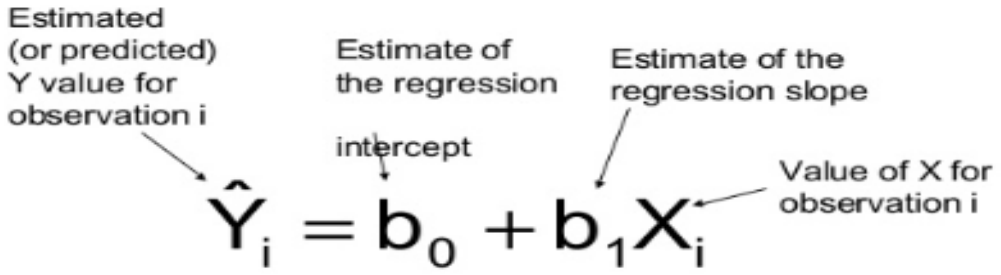
2. Logistic Regression

>It's a classification algorithm that is used where the response variable is categorical.

>The idea of Logistic Regression is to find a relationship between features and probability of particular outcome.

>Predicting the values where only two outcomes /possible values like 1 or 0, True or False, Yes or No
>By the virtue of a logistic curve which takes value 0 or 1.
>E.g. When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.
>Binomial Logistic Regression, where the response variable has two values 0 and 1 or pass and fail or true and false. Y={0,1}
>Multinomial Logistic Regression deals with situations where the response variable can have three or more possible values.Y={0,1,2,3}

12. Explain Simple Regression and Multiple Regression with its equation.
ANS:
Simple Regression –
>It is characterized by one independent variable.
>Consider the price of the house based only one field that is the size of the plot then that would be a simple linear regression.



Multiple Regression –
>Multiple regressions are an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables.
>The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable).
>The variables we are using to predict the value of the dependent variable are called the independent variables (or sometimes, the predictor, explanatory or regress or variables)
>The multiple linear regression equation is as follows:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p ,$$

Where
$\hat{Y}$ is the predicted or expected value of the dependent variable,
X_1 through X_p are p distinct independent or predictor variables,
b_0 is the value of Y when all of the independent variables (X_1 through X_p) are equal to zero, and
b_1 through b_p are the estimated regression coefficients.
>Each regression coefficient represents the change in Y relative to a one unit change in the respective independent variable.
>In the multiple regression situation, b_1, for example, is the change in Y relative to a one unit change in X_1, holding all other independent variables constant (i.e., when the remaining independent variables are held at the same value or are fixed).

>Again, statistical tests can be performed to assess whether each regression coefficient is significantly different from zero.

13. Explain the general model of regression. Give the idea wrt R /What are regression trees? Give an idea wrt R.

ANS:

Regression tree - when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).

Refer to Q. 20 Unit III.

15. What is Linear Model? Explain with the help of example.

ANS:

>Linear models describe a continuous response variable as a function of one or more predictor variables.

>They can help you understand and predict the behavior of complex systems or analyze experimental, financial, and biological data.

>The generalized linear model (GLZ) is a way to make predictions from sets of data.

>It takes the idea of a general linear model (for example, a linear regression equation) a step further.

>A general linear model (GLM) is the type of model you probably came across in elementary statistics.

>Ordinary least squares regression is one example of a GLM.

>They are also found in ANOVA and T Tests.

>The *generalized* linear model on the other hand, is much more complex, drawing from an array of different distributions to find the "best fit" model.

>The model uses, among other techniques, Bayesian hypothesis testing to predict outcomes.

16. Explain Least Square Method used for finding the best fit line in linear models

ANS:

Linear regression is a linear approach to modelling the relationship between a dependent variable and one or more independent variables. In the case of one independent variable it is called simple linear regression. For more than one independent variable, the process is called multiple linear regression.

Let **X** be the independent variable and **Y** be the dependent variable. We will define a linear relationship between these two variables as follows:

$$Y = mX + c$$

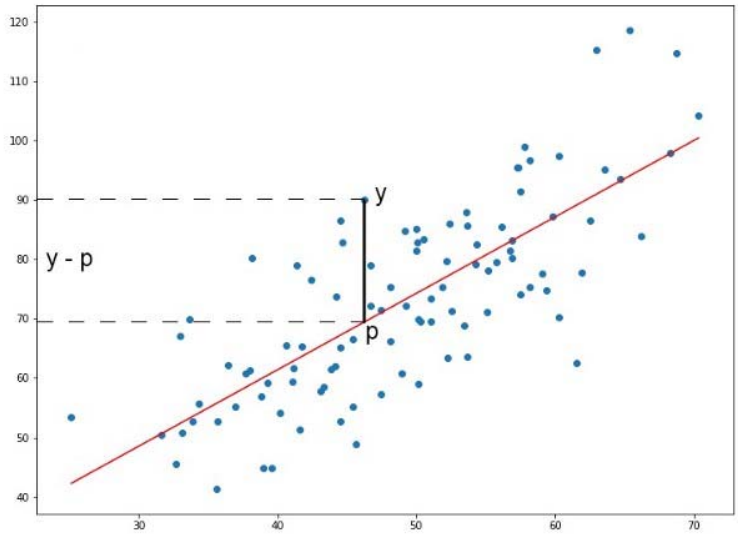
This is the equation for a line. **m** is the slope of the line and **c** is the y intercept. We will use this equation to train our model with a given dataset and predict the value of **Y** for any given value of **X**.

Error has to be reduced,

So to minimize the error we need a way to calculate the error in the first place. A **loss function** in machine learning is simply a measure of how different the predicted value is from the actual value.

Today we will be using the **Quadratic Loss Function** to calculate the loss or error in our model. It can be defined as:

$$L(x) = \sum_{i=1}^{n} (y_i - p_i)^2$$



We are squaring it because, for the points below the regression line **y—p** will be negative and we don't want negative values in our total error.Least Squares method

Now that we have determined the loss function, the only thing left to do is minimize it. This is done by finding the partial derivative of **L**, equating it to 0 and then finding an expression for **m** and **c**. After we do the math, we are left with these equations:

$$m = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$c = \bar{y} - m\bar{x}$$

Here $\bar{x}$ is the mean of all the values in the input **X** and $\bar{y}$ is the mean of all the values in the desired output **Y**. This is the Least Squares method.

17. What is Time Series Analysis? Explain its components.

ANS:

>Time series analysis is a statistical technique that deals with time series data, or trend analysis.

>A time series is a collection of observations of well-defined data items obtained through repeated measurements over time.

>For example, measuring the value of retail sales each month of the year would comprise a time series.

>This is because sales revenue is well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series.
>Time series data meAns that data is in a series of particular time periods or intervals.
> Components of time series analysis:

I) Trend

>Behavior of a particular feature at a particular amount of time
>The ABS trend is defined as the 'long term' movement in a time series without calendar related and irregular effects, and is a reflection of the underlying level.
>It is the result of influences such as population growth, price inflation and general economic changes.

II) Seasonality

>Pattern which repeats at a constant frequency.
>The seasonal component consists of effects that are reasonably stable with respect to timing, direction and magnitude.
>It arises from systematic, calendar related influences such as:

- Natural Conditions
  >Weather fluctuations that are representative of the season (uncharacteristic weather patterns such as snow in summer would be considered irregular influences)

- Business and Administrative procedures
  >Start and end of the school term

- Social and Cultural behavior
- Christmas

III) Cycles

>Seasonality patterns but not repeating at regular frequency. (Task completion time)
>Cycle has often been described as a non-fixed pattern usually of at least 2 years in duration. The length of the cycle is described as the period.

18. Explain different concepts in Time Series Analysis.

ANS:

> Different concepts in Time Series Analysis are:

I) Dependence

>Dependence refers to the association of two observations with the same variable, at prior time points.

II) Stationarity

>Shows the mean value of the series that remains constant over a time period; if past effects accumulate and the values increase toward infinity, then stationarity is not met.

III) Differencing

>Used to make the series stationary, to De-trend, and to control the auto-correlations; however, some time series analyses do not require differencing and over-differenced series can produce inaccurate estimates.

IV) Specification

>May involve the testing of the linear or non-linear relationships of dependent variables by using models such as ARIMA, ARCH, GARCH, VAR, Co-integration, etc.

V) Exponential smoothing in time series analysis

>This method predicts the one next period value based on the past and current value.

>It involves averaging of data such that the nonsystematic components of each individual case or observation cancel out each other.
>The exponential smoothing method is used to predict the short term predication.
>Alpha, Gamma, Phi, and Delta are the parameters that estimate the effect of the time series data.
>Alpha is used when seasonality is not present in data.
>Gamma is used when a series has a trend in data.
>Delta is used when seasonality cycles are present in data.
>A model is applied according to the pattern of the data.
Some more concepts are as follows
Multiple Regression Analysis: Used when two or more independent factors are involved-widely used for intermediate term forecasting. Used to assess which factors to include and which to exclude. Can be used to develop alternate models with different factors.
Nonlinear Regression: Does not assume a linear relationship between variables-frequently used when time is the independent variable.
Trend Analysis: Uses linear and nonlinear regression with time as the explanatory variable-used where pattern over time.
Decomposition Analysis: Used to identify several patterns that appear simultaneously in a time series-time consuming each time it is used-also used to deseasonalize a series
Moving Average Analysis: Simple Moving Averages-forecasts future values based on a weighted average of past values-easy to update.
Weighted Moving Averages: Very powerful and economical. They are widely used where repeated forecasts required-uses methods like sum-of-the-digits and trend adjustment methods.
Adaptive Filtering: A type of moving average which includes a method of learning from past errors-can respond to changes in the relative importance of trend, seasonal, and random factors.

19. Explain Forecasting. List the steps in forecasting.
ANS:
>Estimating the likelihood of an event taking place in the future, based on available data.
>Statistical forecasting concentrates on using the past to predict the future by identifying trends, patterns and business drives within the data to develop a forecast.
>Statistical forecast uses mathematical formulas to identify the patterns and trends while testing the results for mathematical reasonableness and confidence.
>Steps in forecasting are:
I) Developing the Basis: Define the goal or perspective
II) Get the required data: via different data collection methods
III) Explore and visualize the series: EDA
IV) Preprocess the data: Data Cleansing, Data extraction,
V) Partition the series: Data uniformity increases accuracy.
VI) Apply suitable forecasting method
VII) Evaluate and compare the different series & their analytics
VIII) Implement the final forecasting system

20. What is Decision Tree? Explain with the help of example.
ANS:

>Decision Tree learning is one of the predictive modeling techniques.
>Decision trees used in data mining are of two main types:
I) Classification tree –When the predicted outcome is the class to whichthe data belongs.
II) Regression tree –When the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).
>Decision Tree breaks down a data set into smaller subsets and presents association between target variable (dependent) and independent variables as a tree structure.
>A final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches and leaf node represents a classification or decision.
> The term Classification and Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by Breiman et al.
>**Example** –
"PassengerId: Type should be integers
Survived:Survived or Not
Pclass:Class of Travel
Name:Name of Passenger
Sex:Gender
Age:Age of Passengers
SibSp:Number of Sibling/Spouse aboard
Parch:Number of Parent/Child aboard
Ticket
Fare
Cabin
Embarked:The port in which a passenger has embarked. C - Cherbourg, S - Southampton, Q = Queenstown"

```
titanic<-read.csv(file.choose(),header = T,sep=",")
summary(titanic)
names(titanic)
install.packages("partykit")
install.packages("CHAID",repos = "http://R-Forge.R-project.org",type="source")
library(CHAID)
library(partykit)
titanic$Survived<-as.factor(titanic$Survived)
summary(titanic$Survived)
names(titanic)
tree<-chaid(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,data=titanic)
class(titanic$Survived)

library(rpart)
fit<- rpart(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,data=titanic,method="class")
plot(fit)
text(fit)
install.packages('rattle')
install.packages('rpart.plot')
install.packages('RColorBrewer')
```

```
library(rattle)
library(rpart.plot)
library(RColorBrewer)
fancyRpartPlot(fit)
Prediction <- predict(fit, titanic, type = "class")
Prediction

hitters<-read.csv(file.choose(),sep=",",header = T)
summary(hitters)
reg.tree<- rpart(Salary ~ Years + Hits, data = hitters)
rpart.plot(reg.tree, type = 4)
reg.tree$variable.importance
install.packages("MASS")
library(MASS)
set.seed(1984)
library(rpart)
train<- sample(1:nrow(hitters), nrow(hitters)/2)
tree_baseball<- rpart(Salary ~ Hits + HmRun + Runs + RBI + Walks + Years + Errors, subset = train, data = hitters)
library(rpart.plot)
rpart.plot(tree_baseball)
tree_baseball$variable.importance
```

21. What is Information Gain and Entropy? Explain with its formula.

ANS:

>The core algorithm for building decision trees – ID3 uses Entropy and Information Gain to construct the tree.

>**Entropy –** ID3 algorithm uses entropy to calculate the homogeneity of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.

>Entropy of a category is calculated as:

$P1*\log_2(P1) – P2*\log_2(P2)$

where

P1 is the proportion of class 1 and P2 is the proportion of class 2

>Entropy at the variable level can be derived by adding weighted averages of all classes. Weights are the proportion of respondents in each class to total respondents.

>Let us consider survey data from three cities depicting shopper's preferred brand

City	Brand A	Brand B	% of votes for Brand A	% of votes for Brand B	Number of Voters
Delhi	90	310	22.5%	77.5%	400
Chennai	10	90	10%	90%	100
Mumbai	100	100	50%	50%	200

>Entropy for each city is calculated as

- Delhi: $– 0.225* \log_2 (0.225) – 0.775* \log_2 (0.775) = 0.76919$

- Chennai: $-0.1 * \log_2(0.1) - 0.9 * \log_2(0.9) = 0.46900$
- Mumbai: $-0.5 * \log_2(0.5) - 0.5 * \log_2(0.5) = 1$

>Entropy at the variable level is

- $0.57*0.76919 + 0.14* 0.46900 + 0.29*1 = 0.79225$

>**Information Gain –**It is based on the decrease in entropy after a dataset is split on an attribute.

>Constructing a decision tree is all about finding attribute that returns the highest information gain.

>Information gain = Entropy of sample (dependent variable) - Average Entropy of any of the independent variable.

>Information gain can be interpreted as ability of reducing theUncertainty (Entropy) and hence increase predictability.

>Entropy for complete sample (Based on proportion of respondents in category to the total number of respondents) – $(0.286)*\log(0.286)-(0.714)*\log(0.714) = 0.86312$

>Information gain

- $0.86312 - 0.79225 = 0.070868$

22. Write a short note on Classification Tree.

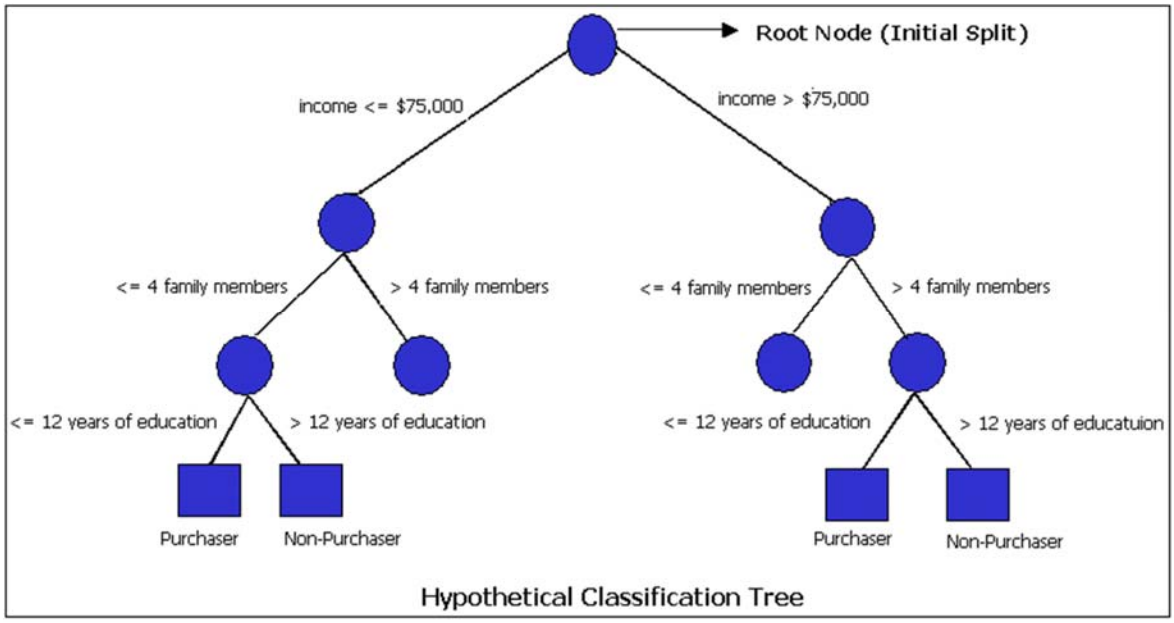
ANS:

>Classification tree methods (i.e., decision tree methods) are recommended when the data mining task contains classifications or predictions of outcomes, and the goal is to generate rules that can be easily explained and translated into SQL or a natural query language.

>A Classification tree labels, records, and assigns variables to discrete classes. A Classification tree can also provide a measure of confidence that the classification is correct.

>A Classification tree is built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches.

>Initially, a Training Set is created where the classification label (i.e., purchaser or non-purchaser) is known (pre-classified) for each record. Next, the algorithm systematically assigns each record to one of two subsets on the some basis (i.e., income > $75,000 or income <= $75,000). The object is to attain an homogeneous set of labels (i.e., purchaser or non-purchaser) in each partition. This partitioning (splitting) is then applied to each of the new partitions. The process continues until no more useful splits can be found. The heart of the algorithm is the rule that determines the initial split rule (displayed in the following figure).



23. What is Logistic Regression? Explain with its equation.

ANS:

>It's a classification algorithm that is used where the response variable is categorical.

>The idea of Logistic Regression is to find a relationship between features and probability of particular outcome.

>Predicting the values where only two outcomes /possible values like 1 or 0, True or False, Yes or No

>By the virtue of a logistic curve which takes value 0 or 1.

>E.g. When we have to predict if a student passes or fails in an exam when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail.

>Binomial Logistic Regression, where the response variable has two values 0 and 1 or pass and fail or true and false. Y={0,1}

>Multinomial Logistic Regression deals with situations where the response variable can have three or more possible values.Y={0,1,2,3}

24. Write down the difference between logistic and linear regression.

ANS:

>The simplest form of regression analysis is the linear regression, where the relation between the variables is a linear relationship. In statistical terms, it brings out the relationship between the explanatory variable and the response variable. For example, using regression we can establish the relation between the commodity price and the consumption based on data collected from a random sample. Regression analysis will produce a regression function of the data set, which is a mathematical model that best fits the data available. This can easily be represented by a scatter plot. Graphically regression is equivalent to finding the best fitting curve for the given data set. The function of the curve is the regression function. Using the mathematical model the usage of a commodity can be predicted for a given price.

Logistic regression is comparable to multivariate regression, and it creates a model to explain the impact of multiple predictors on a response variable. However, in logistic regression, the end result variable should be

categorical (usually divided; i.e., a pair of attainable outcomes like death or survival, though special techniques enable more categorised information to be modelled). A continuous outcome variable may be transformed into a categorical variable, to be used for logistical regression; however, collapsing continuous variables in this manner is mostly discouraged because it reduces the accuracy.

ANS:

1. Variable Type : Linear regression requires the dependent variable to be continuous i.e. numeric values (no categories or groups).
While Binary logistic regression requires the dependent variable to be binary - two categories only (0/1). Multinominal or ordinary logistic regression can have dependent variable with more than two categories.
2. Algorithm : Linear regression is based on least square estimation which says regression coefficients should be chosen in such a way that it minimizes the sum of the squared distances of each observed response to its fitted value.

While logistic regression is based on Maximum Likelihood Estimation which says coefficients should be chosen in such a way that it maximizes the Probability of Y given X (likelihood). With ML, the computer uses different "iterations" in which it tries different solutions until it gets the maximum likelihood estimates.

3. Equation :
Multiple Regression Equation :

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots\ldots\ldots + b_kX_k$$

Linear Regression Equation

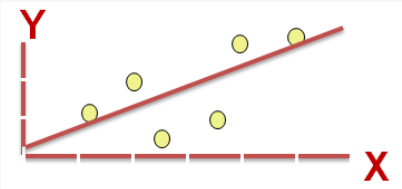
Y is target or dependent variable, b0 is intercept. x1,x2,x3...xk are predictors or independent variables. b1,b2,b3....bk is coefficients of respective predictors.
Logistic Regression Equation :
P(y=1) = e(b0 + b1x1 + b2x2 +-----bkxk) / (1+e(b0+b1x1+ b2x2+------bkxk))
Which further simplifies to :
P(y=1) = 1 / (1+ exp -(b0+b1x1+ b2x2+------bkxk))

$$p = \frac{1}{1 + e^{-(b_0 + b_1X_1 + b_2X_2 + --- + b_kX_k)}}$$

Logistic Regression Equation

The above function is called logistic or sigmoid function.

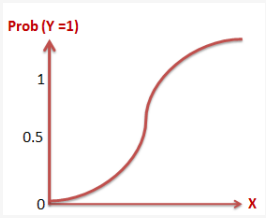
4. Curve :
Linear Regression : Straight line



Straight Line : Linear Regression

Linear regression aims at finding the best-fitting straight line which is also called a regression line. In the above figure, the red diagonal line is the best-fitting straight line and consists of the predicted score on Y for each possible value of X. The distance between the points to the regression line represent the errors.

Logistic Regression : S Curve



Logistic S-shaped curve

Changing the coefficient leads to change in both the direction and the steepness of the logistic function. It means positive slopes result in an S-shaped curve and negative slopes result in a Z-shaped curve.

5. Linear Relationship : Linear regression needs a linear relationship between the dependent and independent variables. While logistic regression does not need a linear relationship between the dependent and independent variables.

6. Normality of Residual : Linear regression requires error term should be normally distributed. While logistic regression does not require error term should be normally distributed.

7. Homoscedasticity : Linear regression assumes that residuals are approximately equal for all predicted dependent variable values. While Logistic regression does not need residuals to be equal for each level of the predicted dependent variable values.

8. Sample Size : Linear regression requires 5 cases per independent variable in the analysis.While logistic regression needs at least 10 events per independent variable.

9. Purpose : Linear regression is used to estimate the dependent variable incase of a change in independent variables. For example, relationship between number of hours studied and your grades. Whereas logistic regression is used to calculate the probability of an event. For example, an event can be whether customer will attrite or not in next 6 months.

10. Interpretation : Betas or Coefficients of linear regression is interpreted like below -

Keeping all other independent variables constant, how much the dependent variable is expected to increase/decrease with an unit increase in the independent variable.
In logistic regression, we interpret odd ratios -

The effect of a one unit of change in X in the predicted odds ratio with the other variables in the model held constant.

11. Distribution:

Linear regression assumes normal or gaussian distribution of dependent variable. Whereas, Logistic regression assumes binomial distribution of dependent variable. Note : Gaussian is the same as the normal distribution. See the implementation in R below –

R Code :

Create sample data by running the following script

```
set.seed(123)
y = ifelse(runif(100) < 0.5, 1,0)
x = sample(1:100,100)
y1 = sample(100:1000,100, replace=T)
```

Linear Regression

```
glm(y1 ~ x, family = gaussian(link = "identity"))
Coefficients:
(Intercept)          x
  600.6152     -0.8631

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:     7339000
Residual Deviance: 7277000  AIC: 1409
```

Logistic Regression

```
glm(y ~ x, family = binomial(link = "logit"))
Coefficients:
(Intercept)                    x
                -0.024018                0.005279

Degrees of Freedom: 99 Total (i.e. Null);  98 Residual
Null Deviance:                    137.2
Residual Deviance: 136.6  AIC: 140.6
```

12. Link Function: Linear regression uses Identity link function of gaussian family. Whereas, logistic regression uses Logit function of Binomial family.

13. Computational Time: Linear regression is very fast as compared to logistic regression as logistic regression is an iterative process of maximum likelihood.

25. Write a short note on separating hyper planes. /Write a short note on SVM.

ANS:

>Hyperplane: It is basically a generalization of plane.

> In one dimension, an hyperplane is called a point.

> In two dimensions, it is a line.

> In three dimensions, it is a plane.

> In more dimensions you can call it anhyperplane.

>SVM or Support Vector Machine is a linear model for classification and regression problems.

>It can solve linear and non-linear problems and work well for many practical problems.

>The algorithm creates a line or a hyperplane which separates the data into classes.

>Support vector machines so called as SVM is a supervised learning algorithmbased on the idea of finding a hyperplane that best separates the features into different domains which can be used for classification and regression problems as support vector classification (SVC) and support vector regression (SVR).

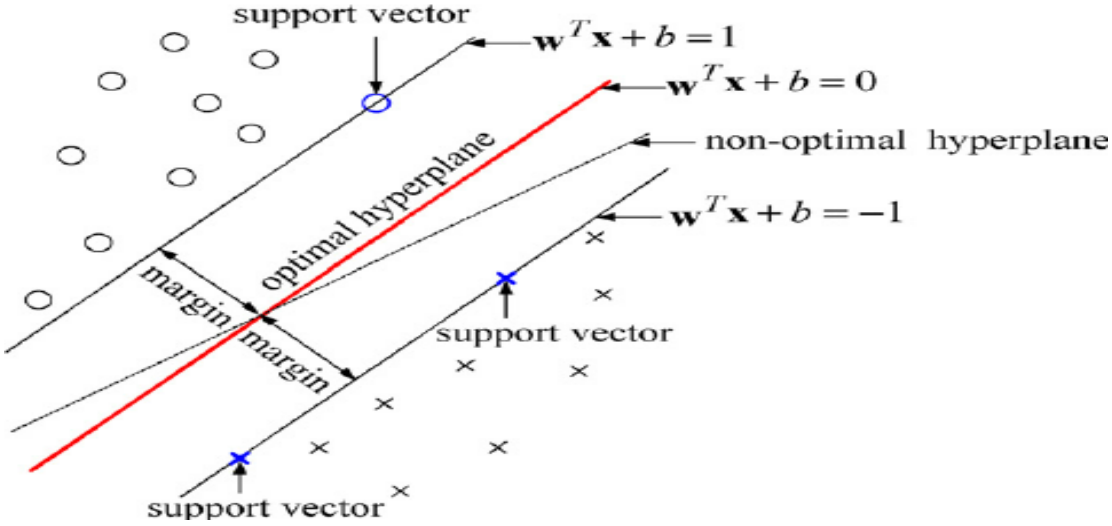
>It is used for smaller dataset as it takes too long to process.

>It is a generalization of a simple classifier called maximum margin classifier.

>SVM is composed of the idea of coming up with an Optimal hyperplane which will clearly classify the different classes(in this case they are binary classes).

>Idea is to choose the hyperplane which is at maximize margin from both the classes of training data.

>Maximizing the distance between the nearest points of each class (minimum of functional margins of all the examples) and the hyperplane would result in an optimal separating hyperplane.



Wi = vectors(W0,W1,W2,W3……Wm)
b = biased term (W0)
X = variables.

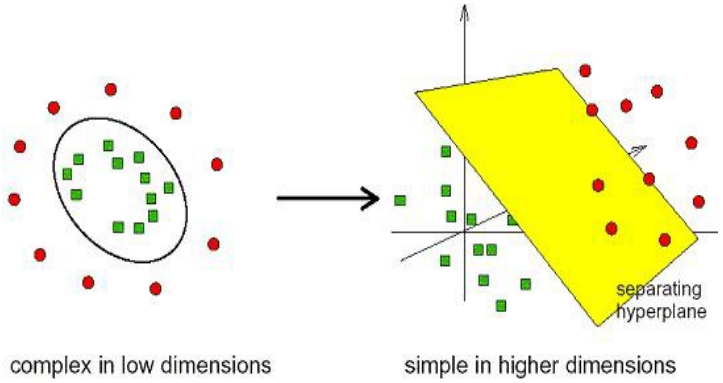
>Support Vectors: Support Vectors are simply the co-ordinates of data points which are nearest to the optimal separating hyperplane.

How to find this optimal separating Hyperplane?

> More the farther SV points, from the hyperplane, more is the probability of correctly classifying the points in their respective region or classes.

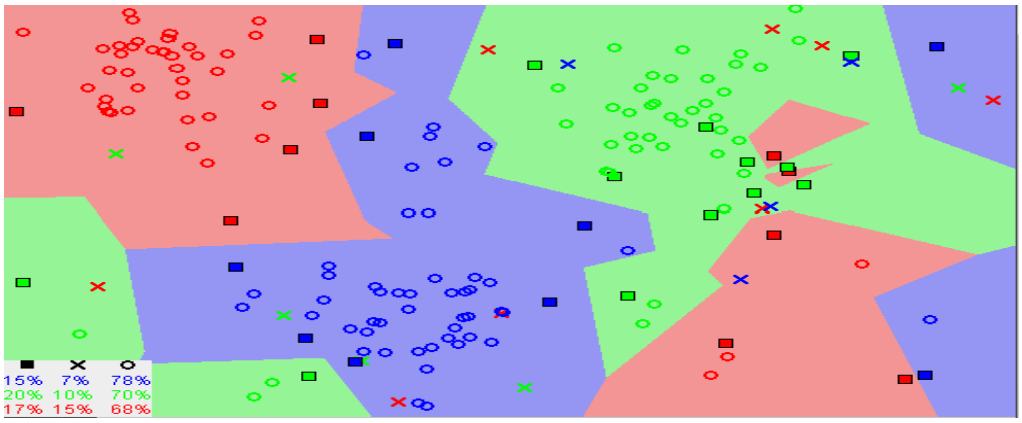
>SV points are very critical in determining the hyperplane because if the position of the vectors changes the hyperplane's position is altered.

>Technically this hyperplane can also be called as margin maximizing hyperplane.
Hard Margin
if the points are linearly separable then only our hyperplane is able to distinguish between them and if any outlier is introduced then it is not able to separate them. So these type of SVM is called as hard margin SVM, real data is messy and cannot be separated perfectly with a hyperplane.So, the constraint of maximizing the margin of the line that separates the classes must be relaxed. This is called the soft margin classifier. This allows some points in the training data to violate the separating line.
A tuning parameter is introduced called C which defines the amount of violation of the margin allowed.In the case of non-linearly separable data points SVM uses the kernel trick.
The idea stems from the fact that if the data cannot be partitioned by a linear boundary in its current dimension,then projecting the data into a higher dimensional space may make it linearly separable.



Pros
It is really effective in higher dimension.
Best algorithm if you data are separable. That two classes are not mixed.
Only support vectors affect the optimally spaced hyperplane. So, it is less affected by outliers.
Cons
On large dataset it takes too much time. Mainly because of kernel function calculations and finding optimal hyperplane in higher dimensions.
Can not perform well in case of overlapping classes.
Can only give you 0–1 classification. Probably estimates computation are really expensive.

26. What is K-NN? Explain with the help of an example.
ANS:
>The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.



>In K meAns algorithm,for each test data point, at the K nearest training data points andtake the most frequently occurring classes and assign that class to the test data.

>Therefore, K represents the number of training data points lying in proximity to the test data point which we are going to use to find the class.

>Predictions are made for a new instance (x) by searching through the entire training set for the K most similar instances (the neighbors) and summarizing the output variable for those K instances

>KNN has no model other than storing the entire dataset, so there is no learning required.

>Steps for K-NN

- Load the training and test data
- Choose the value of K
- For each point in test data:
  - find the Euclidean distance to all training data points
  - store the Euclidean distances in a list and sort it
  - choose the first k points
  - assign a class to the test point based on the majority of classes present in the chosen points
- End

>The KNN Algorithm

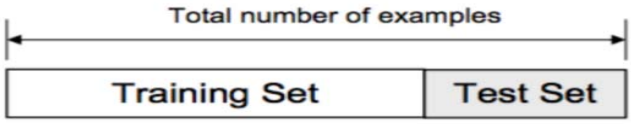
- Load the data
- Initialize K to your chosen number of neighbors
- For each example in the data
  - Calculate the distance between the query example and the current example from the data.
  - Add the distance and the index of the example to an ordered collection
- Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
- Pick the first K entries from the sorted collection
- Get the labels of the selected K entries
- If regression, return the mean of the K labels
- If classification, return the mode of the K labels

27. Explain Model, Train Data and Test Data.

ANS:

>The data we use is usually split into training data and test data.

>The training set contains a known output and the model learns on this data in order to be generalized to other data later on.
>We have the test dataset (or subset) in order to test our model's prediction on this subset.



>"Training data" and "testing data" refer to subsets of the data you wish to analyze.
>If a supervised machine learning algorithm is being used to do something to your data (ex. to classify data points into clusters), the algorithm needs to be "trained".
>Some examples of supervised machine learning algorithms are Support Vector Machines (SVM) and Linear Regression.
>They can be used to classify or cluster data that has many dimensions, allowing us to clump data points that are similar together.
>These algorithms need to be trained with a subset of the data (the "training set") being analyzed before they are used on the "test set".
>Essentially, the training provides an algorithm an opportunity to infer a general solution for some new data it gets presented, much in the same way we as humans train so we can handle new situations in the future.

28. What is Dimension Reduction? Explain its methods.
ANS:
>Dimensionality reduction is simply, the process of reducing the dimension of your feature set.
>Your feature set could be a dataset with a hundred columns (i.e features) or it could be an array of points that make up a large sphere in the three-dimensional space.
>Dimensionality reduction is bringing the number of columns down to say, twenty or converting the sphere to a circle in the two-dimensional space.
> Advantages of Dimensionality reduction:

- Less misleading data meAns model accuracy improves.
- Less dimensions mean less computing. Less data meAns that algorithms train faster.
- Less data meAns less storage space required.
- Less dimensions allow usage of algorithms unfit for a large number of dimensions
- Removes redundant features and noise.

>The most common and well known dimensionality reduction methods are the ones that apply linear trAnsformations, like
I) PCA (Principal Component Analysis):
>Popularly used for dimensionality reduction in continuous data, PCA rotates and projects data along the direction of increasing variance.
>The features with the maximum variance are the principal components.
II) Factor Analysis:
> A technique that is used to reduce a large number of variables into fewer numbers of factors.
>The values of observed data are expressed as functions of a number of possible causes in order to find which are the most important.
>The observations are assumed to be caused by a linear trAnsformation of lower dimensional latent factors and added Gaussian noise.

III) LDA (Linear Discriminant Analysis):
> Projects data in a way that the class separability is maximised.
>Examples from same class are put closely together by the projection.
>Examples from different classes are placed far apart by the projection
Why is Dimensionality Reduction required?
Here are some of the benefits of applying dimensionality reduction to a dataset:
Space required to store the data is reduced as the number of dimensions comes down
Less dimensions lead to less computation/training time
Some algorithms do not perform well when we have a large dimensions. So reducing these dimensions needs to happen for the algorithm to be useful
It takes care of multicollinearity by removing redundant features. For example, you have two variables – 'time spent on treadmill in minutes' and 'calories burnt'. These variables are highly correlated as the more time you spend running on a treadmill, the more calories you will burn. Hence, there is no point in storing both as just one of them does what you require.
It is very difficult to visualize data in higher dimensions so reducing our space to 2D or 3D may allow us to plot and observe patterns more clearly.

Done by the following:

- Missing Value Ratio
- Low Variance Filter
- High Correlation filter
- Backward Feature Elimination
- Forward Feature Selection
- Factor Analysis

29. Explain Filter method, Wrapping Method and Embedded method of Data selection.
ANS:

- Top reasons to use feature selection are:
- It enables the machine learning algorithm to train faster.
- It reduces the complexity of a model and makes it easier to interpret.
- It improves the accuracy of a model if the right subset is chosen.
- It reduces overfitting.

I) Filter Method:
>Filter methods are generally used as a preprocessing step.
>The selection of features is independent of any machine learning algorithms.
>Instead, features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable.
>The correlation is a subjective term here. For basic guidance, you can refer to the following table for defining correlation co-efficients.

II) Wrapper Method:

>In wrapper methods, we try to use a subset of features and train a model using them.

>Based on the inferences that we draw from the previous model, we decide to add or remove features from your subset.

>The problem is essentially reduced to a search problem. These methods are usually computationally very expensive.

>Some common examples of wrapper methods are forward feature selection, backward feature elimination, recursive feature elimination, etc.

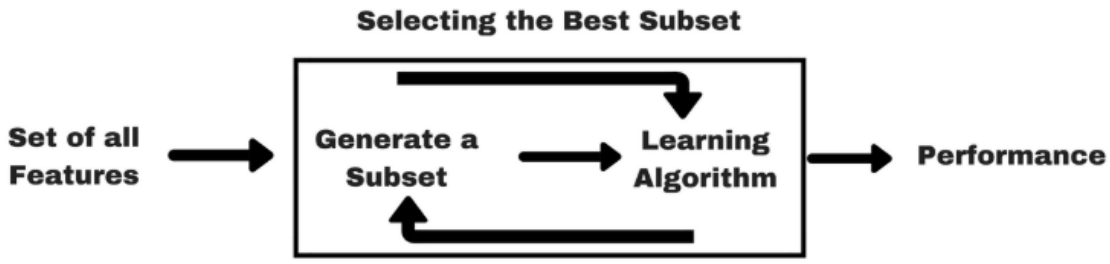
III) Forward Selection:

>Forward selection is an iterative method in which we start with having no feature in the model.

>In each iteration, we keep adding the feature which best improves our model till an addition of a new variable does not improve the performance of the model.

Backward Elimination: In backward elimination, we start with all the features and removes the least significant feature at each iteration which improves the performance of the model. We repeat this until no improvement is observed on removal of features.

Recursive Feature elimination: It is a greedy optimization algorithm which aims to find the best performing feature subset. It repeatedly creates models and keeps aside the best or the worst performing feature at each iteration. It constructs the next model with the left features until all the features are exhausted. It then ranks the features based on the order of their elimination.



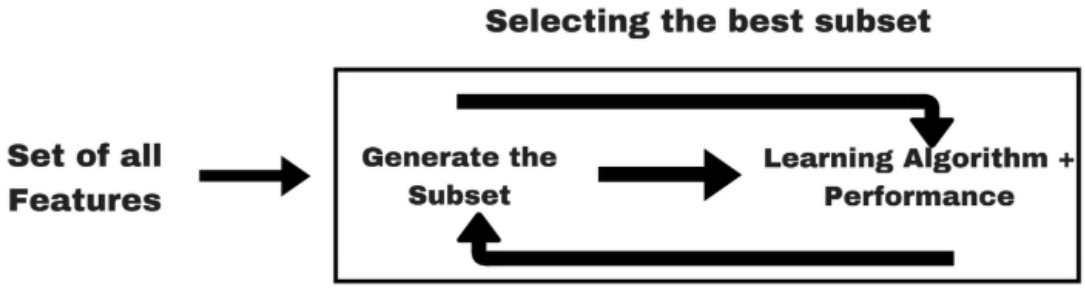
Embedded methods combine the qualities' of filter and wrapper methods. It's implemented by algorithms that have their own built-in feature selection methods.

Some of the most popular examples of these methods are LASSO and RIDGE regression which have inbuilt penalization functions to reduce overfitting.

Lasso regression performs L1 regularization which adds penalty equivalent to absolute value of the magnitude of coefficients.

Ridge regression performs L2 regularization which adds penalty equivalent to square of the magnitude of coefficients.

Other examples of embedded methods are Regularized trees, Memetic algorithm, Random multinomial logit.



30. Write a short note on Feature Selection and Data Extraction

ANS:

>Feature Selection is one of the core concepts in machine learning which hugely impacts the performance of your model.

>The data features that you use to train your machine learning models have a huge influence on the performance you can achieve.

>Irrelevant or partially relevant features can negatively impact model performance.

>Feature selection and Data cleaning should be the first and most important step of your model designing.

>Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

>Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features.

How to select features and what are Benefits of performing feature selection before modeling your data?

· Reduces Overfitting: Less redundant data meAns less opportunity to make decisions based on noise.
· Improves Accuracy: Less misleading data meAns modeling accuracy improves.
· Reduces Training Time: fewer data points reduce algorithm complexity and algorithms train faster.

Feature Selection Methods:
I will share 3 Feature selection techniques that are easy to use and also gives good results.
1. Univariate Selection
2. Feature Importance
3.Correlation Matrix with Heatmap

31. Explain Smoothing and its methods.

ANS:

>Smoothing is a very powerful technique used all across data analysis.

>Other names given to this technique are *curve fitting* and *low pass filtering*.

>It is designed to detect trends in the presence of noisy data in cases in which the shape of the trend is unknown.

>The *smoothing* name comes from the fact that to accomplish this feat, we assume that the trend is *smooth*, as in a smooth surface.

>In contrast, the noise, or deviation from the trend, is unpredictably wobbly
Refer to answer below for methods.

32. Write a short note on Binning method of Data Smoothing.(included in Unit I)

33. Write a short note on Smoothing and Aggregation.

ANS:

Smoothing is a very powerful technique used all across data analysis. It is designed to estimate f(x)f(x) when the shape is unknown, but assumed to be *smooth*. The general idea is to group data points that are expected to have similar expectations and compute the average, or fit a simple parametric model.

When data collected over time displays random variation, smoothing techniques can be used to reduce or cancel the effect of these variations. When properly applied, these techniques smooth out the random variation in the time series data to reveal underlying trends.

XLMiner features four different smoothing techniques: Exponential, Moving Average, Double Exponential, and Holt-Winters. Exponential and Moving Average are relatively simple smoothing techniques and should not be performed on data sets involving seasonality. Double Exponential and Holt-Winters are more advanced techniques that can be used on data sets involving seasonality.

Exponential Smoothing

Exponential Smoothing is one of the more popular smoothing techniques due to its flexibility, ease in calculation, and good performance. Exponential Smoothing uses a simple average calculation to assign exponentially decreasing weights starting with the most recent observations. New observations are given relatively more weight in the average calculation than older observations. The Exponential Smoothing tool uses the following formulas.

$S_0 = x_0$

$S_t = \alpha x_{t-1} + (1-\alpha)st\text{-}1,\ t > 0$

where

original observations are denoted by $\{x_t\}$ starting at $t = 0$

α is the smoothing factor which lies between 0 and 1

Exponential Smoothing should only be used when the data set contains no seasonality. The forecast is a constant value that is the smoothed value of the last observation.

Moving Average Smoothing

In Moving Average Smoothing, each observation is assigned an equal weight, and each observation is forecasted by using the average of the previous observation(s). Using the time series $X_1, X_2, X_3, ..., X_t$, this smoothing technique predicts X_{t+k} as follows :

$S_t = \text{Average}\ (x_{t-k+1}, x_{t-k+2}, ..., x_t),\ t = k, k+1, k+2, ...N$

where, k is the smoothing parameter.

XLMiner allows a parameter value between 2 and t-1 where t is the number of observations in the data set. Note that when choosing this parameter, a large parameter value will oversmooth the data, while a small parameter value will undersmooth the data. The past three observations will predict the future observations. As with Exponential Smoothing, this technique should not be applied when seasonality is present in the data set.

Double Exponential Smoothing

Double Exponential Smoothing can be defined as the recursive application of an exponential filter twice in a time series. Double Exponential Smoothing should not be used when the data includes seasonality. This technique introduces a second equation that includes a trend parameter; thus, this technique should be used when a trend is inherent in the data set, but not used when seasonality is present. Double Exponential Smoothing is defined by the following formulas.

$S_t = A_t + B_t$, t = 1,2,3,..., N

Where, $A_t = ax_t + (1- a) S_{t-1}$ $0< a <= 1$

$B_t = b (A_t - A_{t-1}) + (1 - b) B_{t-1}$ $0< b <= 1$

The forecast equation is: $X_{t+k} = A_t + K B_t$, K = 1, 2, 3, ...

where, a denotes the Alpha parameter, and b denotes the trend parameters. These two parameters can be entered manually. XLMiner includes an optimize feature that will choose the best values for alpha and trend parameters based on the Forecasting Mean Squared Error. If the trend parameter is 0, then this technique is equivalent to the Exponential Smoothing technique. (However, results may not be identical due to different initialization methods for these two techniques.)

Holt-Winters Smoothing

Holt Winters Smoothing introduces a third parameter (g) to account for seasonality (or periodicity) in a data set. The resulting set of equations is called the Holt-Winters method, after the names of the inventors. The Holt-Winters method can be used on data sets involving trend and seasonality (a, b , g). Values for all three parameters can range between 0 and 1.

The following three models associated with this method.

Multiplicative: $X_t = (A_t + B_t)* S_t + e_t$ where A_t and B_t are previously calculated initial estimates. S_t is the average seasonal factor for the t^{th} season.

$A_t = ax_t/S_{t-p} + (1-a)(A_{t-1} + B_{t-1})$

$B_t = b(A_t + A_{t-1}) + (1 - b)B_{t-1}$

$S_t = gx_t/A_t + (1 - g)S_{t-p}$

Additive: $X_t = (A_t + B_t) + S_t + e_t$

No Trend: b = 0, so, $X_t = A * S_t + e_t$

Aggregation

Data aggregation is a type of data and information mining process where data is searched, gathered and presented in a report-based, summarized format to achieve specific business objectives or processes and/or conduct human analysis.

Data aggregation may be performed manually or through specialized software.

Data aggregation is a component of business intelligence (BI) solutions. Data aggregation personnel or software search databases find relevant search query data and present data findings in a summarized format that is meaningful and useful for the end user or application.

Data aggregation generally works on big data or data marts that do not provide much information value as a whole.

Data aggregation's key applications are the gathering, utilization and presentation of data that is available and present on the global Internet.

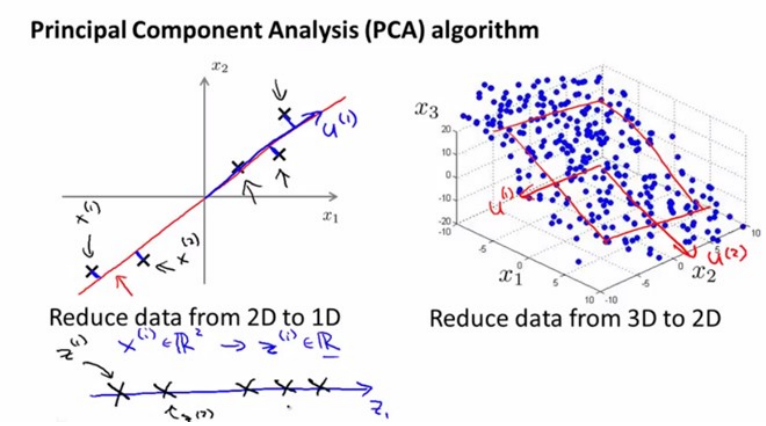
34. What is Principal Component Analysis (PCA)? Explain with the help of an example.

ANS:

The main idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset, up to the maximum extent. The same is done by transforming the variables to a new set of variables, which are known as the principal components (or simply, the PCs) and are orthogonal, ordered such that the retention of variation present in the original variables decreases as we move down in the order. So, in this way, the 1st principal component retains maximum variation that was present in the original components. The principal components are the eigenvectors of a covariance matrix, and hence they are orthogonal.

Importantly, the dataset on which PCA technique is to be used must be scaled. The results are also sensitive to the relative scaling. As a layman, it is a method of summarizing data. Imagine some wine bottles on a dining table. Each wine is described by its attributes like colour, strength, age, etc. But redundancy will arise because many of them will measure related properties. So what PCA will do in this case is summarize each wine in the stock with less characteristics.

Intuitively, Principal Component Analysis can supply the user with a lower-dimensional picture, a projection or "shadow" of this object when viewed from its most informative viewpoint.



Dimensionality: It is the number of random variables in a dataset or simply the number of features, or rather more simply, the number of columns present in your dataset.
Correlation: It shows how strongly two variable are related to each other. The value of the same ranges for -1 to +1. Positive indicates that when one variable increases, the other increases as well, while negative indicates the other decreases on increasing the former. And the modulus value of indicates the strength of relation.
Orthogonal: Uncorrelated to each other, i.e., correlation between any pair of variables is 0.
Eigenvectors: Eigenvectors and Eigenvalues are in itself a big domain, let's restrict ourselves to the knowledge of the same which we would require here. So, consider a non-zero vector v. It is an eigenvector of a square matrix A, if Av is a scalar multiple of v. Or simply:
$Av = \lambda v$
Here, v is the eigenvector and λ is the eigenvalue associated with it.

Covariance Matrix: This matrix consists of the covariances between the pairs of variables. The (i,j)th element is the covariance between i-th and j-th variable.
Technically, a principal component can be defined as a linear combination of optimally-weighted observed variables. The output of PCA are these principal components, the number of which is less than or equal to the number of original variables. Less, in case when we wish to discard or reduce the dimensions in our dataset. The PCs possess some useful properties which are listed below:

The PCs are essentially the linear combinations of the original variables, the weights vector in this combination is actually the eigenvector found which in turn satisfies the principle of least squares.
The PCs are orthogonal..
The variation present in the PCs decrease as we move from the 1st PC to the last one, hence the importance.
The least important PCs are also sometimes useful in regression, outlier detection, etc.

Implementing PCA on a 2-D Dataset
Step 1: Normalize the data

First step is to normalize the data that we have so that PCA works properly. This is done by subtracting the respective means from the numbers in the respective column. So if we have two dimensions X and Y, all X become $\boldsymbol{x}$- and all Y become $\boldsymbol{y}$-. This produces a dataset whose mean is zero.

Step 2: Calculate the covariance matrix

Since the dataset we took is 2-dimensional, this will result in a 2x2 Covariance matrix.

Please note that Var[X1] = Cov[X1,X1] and Var[X2] = Cov[X2,X2].

Step 3: Calculate the eigenvalues and eigenvectors

Next step is to calculate the eigenvalues and eigenvectors for the covariance matrix. The same is possible because it is a square matrix. λ is an eigenvalue for a matrix A if it is a solution of the characteristic equation:

$$\det(\lambda I - A) = 0$$

Where, I is the identity matrix of the same dimension as A which is a required condition for the matrix subtraction as well in this case and 'det' is the determinant of the matrix. For each eigenvalue λ, a corresponding eigen-vector v, can be found by solving:

$$(\lambda I - A)v = 0$$

Step 4: Choosing components and forming a feature vector:

We order the eigenvalues from largest to smallest so that it gives us the components in order or significance. Here comes the dimensionality reduction part. If we have a dataset with n variables, then we have the corresponding n eigenvalues and eigenvectors. It turns out that the eigenvector corresponding to the highest eigenvalue is the principal component of the dataset and it is our call as to how many eigenvalues we choose to proceed our analysis with. To reduce the dimensions, we choose the first p eigenvalues and ignore the rest. We do lose out some information in the process, but if the eigenvalues are small, we do not lose much.

Next we form a feature vector which is a matrix of vectors, in our case, the eigenvectors. In fact, only those eigenvectors which we want to proceed with. Since we just have 2 dimensions in the running example, we can either choose the one corresponding to the greater eigenvalue or simply take both.

Feature Vector = (eig1, eig2)

Step 5: Forming Principal Components:

This is the final step where we actually form the principal components using all the math we did till here. For the same, we take the trAnspose of the feature vector and left-multiply it with the trAnspose of scaled version of original dataset.

NewData = FeatureVectorT x ScaledDataT
Here,
NewData is the Matrix consisting of the principal components,
FeatureVector is the matrix we formed using the eigenvectors we chose to keep, and
ScaledData is the scaled version of original dataset
('T' in the superscript denotes trAnspose of a matrix which is formed by interchanging the rows to columns and vice versa. In particular, a 2x3 matrix has a trAnspose of size 3x2)

If we go back to the theory of eigenvalues and eigenvectors, we see that, essentially, eigenvectors provide us with information about the patterns in the data. In particular, in the running example of 2-D set, if we plot the eigenvectors on the scatterplot of data, we find that the principal eigenvector (corresponding to the largest eigenvalue) actually fits well with the data. The other one, being perpendicular to it, does not carry much information and hence, we are at not much loss when deprecating it, hence reducing the dimension.

All the eigenvectors of a matrix are perpendicular to each other. So, in PCA, what we do is represent or transform the original dataset using these orthogonal (perpendicular) eigenvectors instead of representing on normal x and y axes. We have now classified our data points as a combination of contributions from both x and y. The difference lies when we actually disregard one or many eigenvectors, hence, reducing the dimension of the dataset

35. Explain PCA. List the steps in PCA.
ANS:
REFER ANSWER 35 for PCA.

36. Explain K-means clustering with the help of an example.
ANS:
>K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.
>Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labelled, outcomes.
>A cluster refers to a collection of data points aggregated together because of certain similarities.
>You'll define a target number k, which refers to the number of centroids you need in the dataset.
>A centroid is the imaginary or real location representing the centre of the cluster.
>Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.
>In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.
>The 'means' in the K-means refers to averaging of the data; that is, finding the centroid.
How the K-means algorithm works
>To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids
>It halts creating and optimizing clusters when either:

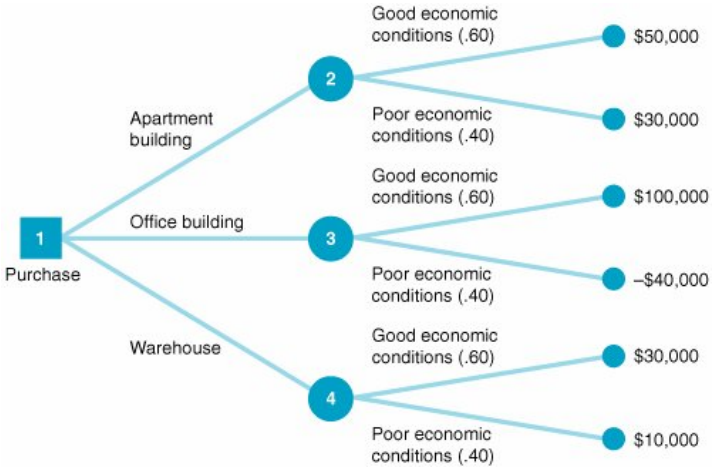
>The centroids have stabilized—there is no change in their values because the clustering has been successful.
>The defined number of iterations has been achieved.

37. What is Hierarchical Clustering?
ANS:
>Clustering, in one sentence, is the extraction of natural groupings of similar data objects.
Hierarchical Clustering
>As mentioned before, hierarchical clustering relies using these clustering techniques to find a hierarchy of clusters, where this hierarchy resembles a tree structure, called a dendrogram.
>Hierarchical clustering is the hierarchical decomposition of the data based on group similarities
Finding hierarchical clusters
>There are two top-level methods for finding these hierarchical clusters:

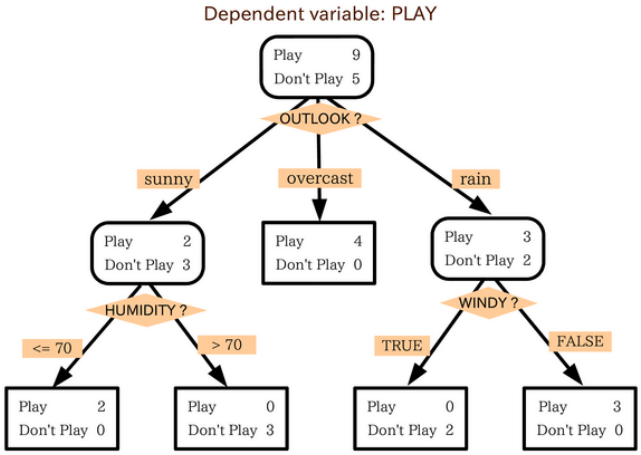
- Agglomerative clustering uses a bottom-up approach, wherein each data point starts in its own cluster. These clusters are then joined greedily, by taking the two most similar clusters together and merging them.
- Divisive clustering uses a top-down approach, wherein all data points start in the same cluster. You can then use a parametric clustering algorithm like K-MeAns to divide the cluster into two clusters. For each cluster, you further divide it down to two clusters until you hit the desired number of clusters.

>Both of these approaches rely on constructing a similarity matrix between all of the data points, which is usually calculated by cosine or Jaccard distance.

38. Explain ensemble methods.
ANS:
>Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model.
A Decision Tree determines the predictive value based on series of questions and conditions. For instance, this simple Decision Tree determining on whether an individual should play outside or not. The tree takes several weather factors into account, and given each factor either makes a decision or asks another question. In this example, every time it is overcast, we will play outside. However, if it is raining, we must ask if it is windy or not? If windy, we will not play. But given no wind, tie those shoelaces tight because were going outside to play.



Decision Trees can also solve quantitative problems as well with the same format. In the Tree to the left, we want to know wether or not to invest in a commercial real estate property. Is it an office building? A Warehouse? An Apartment building? Good economic conditions? Poor Economic Conditions? How much will an investment return? These questions are Answered and solved using this decision tree.

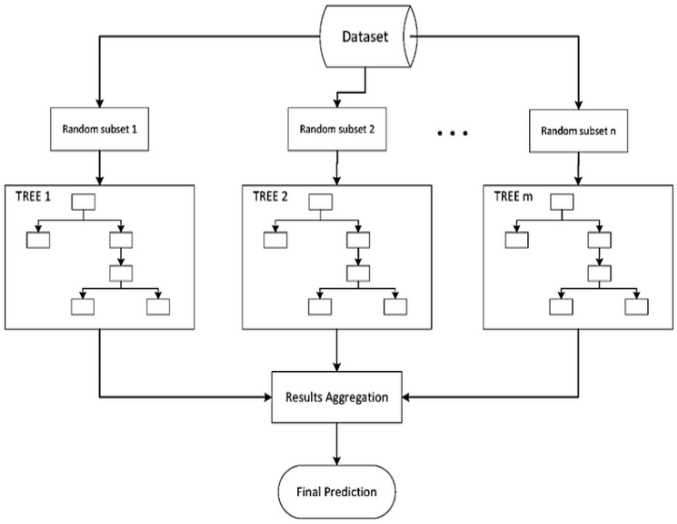


When making Decision Trees, there are several factors we must take into consideration: On what features do we make our decisions on? What is the threshold for classifying each question into a yes or no Answer? In the first Decision Tree, what if we wanted to ask ourselves if we had friends to play with or not. If we have friends, we will play every time. If not, we might continue to ask ourselves questions about the weather. By adding an additional question, we hope to greater define the Yes and No classes.

This is where Ensemble Methods come in handy! Rather than just relying on one Decision Tree and hoping we made the right decision at each split, Ensemble Methods allow us to take a sample of Decision Trees into account, calculate which features to use or questions to ask at each split, and make a final predictor based on the aggregated results of the sampled Decision Trees.

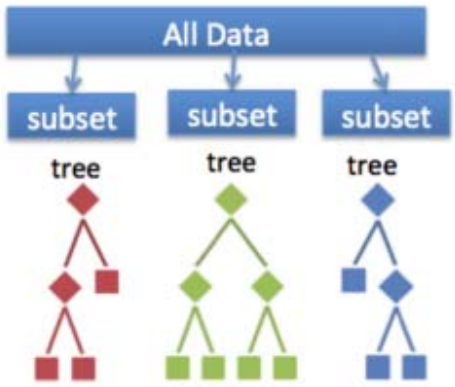
Types of Ensemble Methods

BAGGing, or Bootstrap AGGregating. BAGGing gets its name because it combines Bootstrapping and Aggregation to form one ensemble model. Given a sample of data, multiple bootstrapped subsamples are pulled. A Decision Tree is formed on each of the bootstrapped subsamples. After each subsample Decision Tree has been formed, an algorithm is used to aggregate over the Decision Trees to form the most efficient predictor. The image below will help explain:



Given a Dataset, bootstrapped subsamples are pulled. A Decision Tree is formed on each bootstrapped sample. The results of each tree are aggregated to yield the strongest, most accurate predictor.

2. Random Forest Models. Random Forest Models can be thought of as BAGGing, with a slight tweak. When deciding where to split and how to make decisions, BAGGed Decision Trees have the full disposal of features to choose from. Therefore, although the bootstrapped samples may be slightly different, the data is largely going to break off at the same features throughout each model. In contrary, Random Forest models decide where to split based on a random selection of features. Rather than splitting at similar features at each node throughout, Random Forest models implement a level of differentiation because each tree will split based on different features. This level of differentiation provides a greater ensemble to aggregate over, ergo producing a more accurate predictor. Refer to the image for a better understanding.



A random forest takes a random subset of features from the data, and creates n random trees from each subset. Trees are aggregated together at end.

Similar to BAGGing, bootstrapped subsamples are pulled from a larger dataset. A decision tree is formed on each subsample. HOWEVER, the decision tree is split on different features (in this diagram the features are represented by shapes).

39. What is Overfitting and underfitting? Explain with the help of Diagram.

ANS:

Overfitting in Machine Learning

>Overfitting refers to a model that models the training data too well.

Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This meAns that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function. As such, many nonparametric machine learning algorithms also include parameters or techniques to limit and constrain how much detail the model learns.

For example, decision trees are a nonparametric machine learning algorithm that is very flexible and is subject to overfitting training data. This problem can be addressed by pruning a tree after it has learned in order to remove some of the detail it has picked up.

Underfitting in Machine Learning

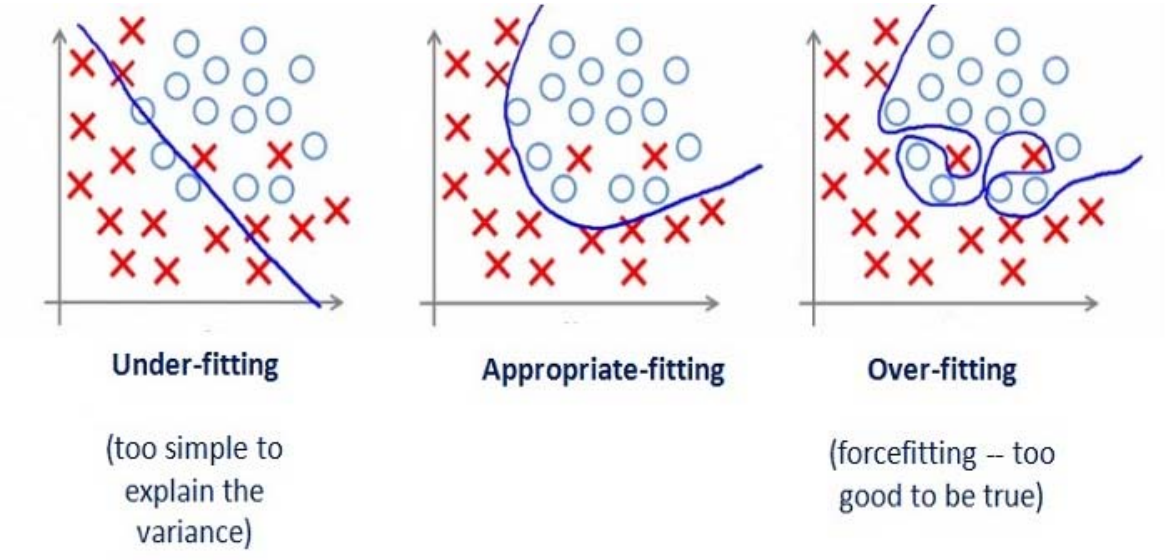
>Underfitting refers to a model that can neither model the training data nor generalize to new data.

>An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data.

>Underfitting is often not discussed as it is easy to detect given a good performance metric.

>The remedy is to move on and try alternate machine learning algorithms.

>Nevertheless, it does provide a good contrast to the problem of overfitting.



40. What is Regularization? Why do we need to use Regularization? Explain with its equation.

ANS:

Regularization

>This is a form of regression that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.

>A simple relation for linear regression looks like this. Here Y represents the learned relation and β represents the coefficient estimates for different variables or predictors(X).

Y ≈ β0 + β1X1 + β2X2 + …+ βpXp

The fitting procedure involves a loss function, known as residual sum of squares or RSS. The coefficients are chosen, such that they minimize this loss function.

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

Now, this will adjust the coefficients based on your training data. If there is noise in the training data, then the estimated coefficients won't generalize well to the future data. This is where regularization comes in and shrinks or regularizes these learned estimates towards zero.

Ridge Regression

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

In ridge regression, the RSS is modified by adding the shrinkage quantity. Now, the coefficients are estimated by minimizing this function. Here, λ is the tuning parameter that decides how much we want to penalize the flexibility of our model. The increase in flexibility of a model is represented by increase in its coefficients, and if we want to minimize the above function, then these coefficients need to be small. This is how the Ridge regression technique prevents coefficients from rising too high. Also, notice that we shrink the estimated association of each variable with the response, except the intercept β0, This intercept is a measure of the mean value of the response when xi1 = xi2 = …= xip = 0.

When λ = 0, the penalty term has no e□ect, and the estimates produced by ridge regression will be equal to least squares. However, as λ→∞, the impact of the shrinkage penalty grows, and the ridge regression coe□cient estimates will approach zero. As can be seen, selecting a good value of λ is critical. Cross validation comes in handy for this purpose. The coefficient estimates produced by this method are also known as the L2 norm.

The coefficients that are produced by the standard least squares method are scale equivariant, i.e. if we multiply each input by c then the corresponding coefficients are scaled by a factor of 1/c. Therefore, regardless of how the predictor is scaled, the multiplication of predictor and coefficient(Xjβj) remains the same. However, this is not the case with ridge regression, and therefore, we need to standardize the predictors or bring the predictors to the same scale before performing ridge regression. The formula used to do this is given below.

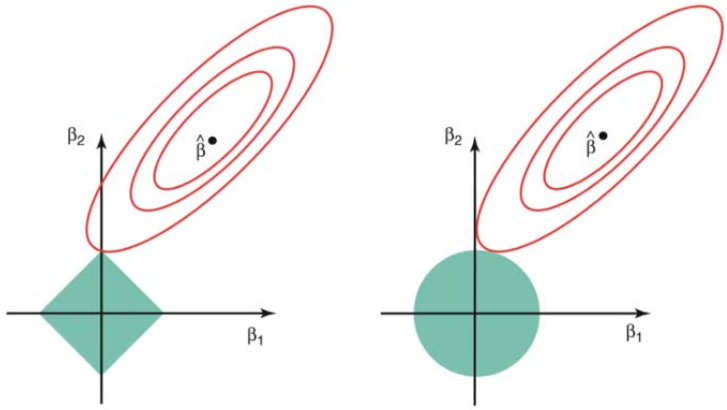
$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}},$$

Lasso

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Lasso is another variation, in which the above function is minimized. Its clear that this variation differs from ridge regression only in penalizing the high coefficients. It uses |βj|(modulus)instead of squares of β, as its penalty. In statistics, this is known as the L1 norm.

Lets take a look at above methods with a different perspective. The ridge regression can be thought of as solving an equation, where summation of squares of coefficients is less than or equal to s. And the Lasso can be thought of as an equation where summation of modulus of coefficients is less than or equal to s. Here, s is a constant that exists for each value of shrinkage factor λ. These equations are also referred to as constraint functions.
Consider there are 2 parameters in a given problem. Then according to above formulation, the ridge regression is expressed by $\beta 1^2 + \beta 2^2 \leq s$. This implies that ridge regression coefficients have the smallest RSS (loss function) for all points that lie within the circle given by $\beta 1^2 + \beta 2^2 \leq s$.
Similarly, for lasso, the equation becomes,$|\beta 1|+|\beta 2| \leq s$. This implies that lasso coefficients have the smallest RSS (loss function) for all points that lie within the diamond given by $|\beta 1|+|\beta 2| \leq s$.
The image below describes these equations.



The above image shows the constraint functions (green areas), for lasso (left) and ridge regression (right), along with contours for RSS (red ellipse). Points on the ellipse share the value of RSS. For a very large value of s, the green regions will contain the centre of the ellipse, making coefficient estimates of both regression techniques, equal to the least squares estimates. But, this is not the case in the above image. In this case, the lasso and ridge regression coefficient estimates are given by the first point at which an ellipse contacts the constraint region. Since ridge regression has a circular constraint with no sharp points, this intersection will not generally occur on an axis, and so the ridge regression coe□cient estimates will be exclusively non-zero. However, the lasso constraint has corners at each of the axes, and so the ellipse will often intersect the constraint region at an axis. When this occurs, one of the coe□cients will equal zero. In higher dimensions (where parameters are much more than 2), many of the coe□cient estimates may equal zero simultaneously.
This sheds light on the obvious disadvantage of ridge regression, which is model interpretability. It will shrink the coefficients for least important predictors, very close to zero. But it will never make them exactly zero. In other words, the final model will include all predictors. However, in the case of the lasso, the L1 penalty has the e□ect of forcing some of the coe□cient estimates to be exactly equal to zero when the tuning parameter λ is su□ciently large. Therefore, the lasso method also performs variable selection and is said to yield sparse models.
What does Regularization achieve?
A standard least squares model tends to have some variance in it, i.e. this model won't generalize well for a data set different than its training data. Regularization, significantly reduces the variance of the model, without substantial increase in its bias. So the tuning parameter λ, used in the regularization techniques described above, controls the impact on bias and variance. As the value of λ rises, it reduces the value of coefficients and thus reducing the variance. Till a point, this increase in λ is beneficial as it is only reducing the variance (hence avoiding overfitting), without losing any important properties in the data. But after certain value, the model starts

loosing important properties, giving rise to bias in the model and thus underfitting. Therefore, the value of λ should be carefully selected.
This is all the basic you will need, to get started with Regularization. It is a useful technique that can help in improving the accuracy of your regression models. A popular library for implementing these algorithms is Scikit-Learn. It has a wonderful api that can get your model up an running with just a few lines of code in python.

41. What is L1 Regularization? Explain with its equation. / What is L2 Regularization? Explain with its equation.
ANS:
1. L1 Regularization
2. L2 Regularization
A regression model that uses L1 regularization technique is called Lasso Regression and model which uses L2 is called Ridge Regression.
The key difference between these two is the penalty term.
Ridge regression adds "squared magnitude" of coefficient as penalty term to the loss function. Here the highlight

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

ed part represents L2 regularization element.

Cost function
Here, if lambda is zero then you can imagine we get back OLS. However, if lambda is very large then it will add too much weight and it will lead to under-fitting. Having said that it's important how lambda is chosen. This technique works very well to avoid over-fitting issue.
Lasso Regression (Least Absolute Shrinkage and Selection Operator) adds "absolute value of magnitude" of coefficient as penalty term to the loss function.

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Cost function
Again, if lambda is zero then we will get back OLS whereas very large value will make coefficients zero hence it will under-fit.
The key difference between these techniques is that Lasso shrinks the less important feature's coefficient to zero thus, removing some feature altogether. So, this works well for feature selection in case we have a huge number of features.
Traditional methods like cross-validation, stepwise regression to handle overfitting and perform feature selection work well with a small set of features but these techniques are a great alternative when we are dealing with a large set of features.
As An Error Function
L1-norm loss function is also known as least absolute deviations (LAD), least absolute errors (LAE). It is basically minimizing the sum of the absolute differences (S) between the target value (Y_i) and the estimated values ($f(x_i)$):

$$S = \sum_{i=1}^{n} |y_i - f(x_i)|.$$

L2-norm loss function is also known as least squares error (LSE). It is basically minimizing the sum of the square of the differences (S) between the target value (Y_i) and the estimated values ($f(x_i)$):

$$S = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

The differences of L1-norm and L2-norm as a loss function can be promptly summarized as follows:

L2 loss function	L1 loss function
Not very robust	Robust
Stable solution	Unstable solution
Always one solution	Possibly multiple solutions

42. Explain Penalized Regression.

Ans

>Penalized regression methods are examples of modern approaches to model selection. Because they produce more stable results for correlated data or data where the number of predictors is much larger than the sample size, they are often preferred to traditional selection methods.

>Unlike subset selection methods, penalized regression methods do not explicitly select the variables; instead they minimize the RSS by using a penalty on the size of the regression coefficients. This penalty causes the regression coefficients to shrink toward zero. This is why penalized regression methods are also known as shrinkage or regularization methods.

>If the shrinkage is large enough, some regression coefficients are set to zero exactly. Thus, penalized regression methods perform variable selection and coefficient estimation simultaneously.

>A penalized regression method yields a sequence of models, each associated with specific values for one or more tuning parameters. Thus you need to specify at least one tuning method to choose the optimum model (that is, the model that has the minimum estimated prediction error).

>Popular tuning methods for penalized regression include fit criteria (such as AIC, SBC, and the Cp statistic), average square error on the validation data, and cross validation.

>Understanding the bias-variance tradeoff is crucial in understanding penalized regression. The bias-variance tradeoff can be best explained by the mean square error (MSE) of a model, which is basically its expected prediction error.

>Penalized regression methods introduce bias in coefficient estimation by continuously shrinking the regression coefficients. However, this shrinkage provides a decrease in variance. This is called the bias-variance trade-off.

>Penalized regression methods keep all the predictor variables in the model but constrain (regularize) the regression coefficients by shrinking them toward zero. If the amount of shrinkage is large enough, these methods can also perform variable selection by shrinking some coefficients to zero.
