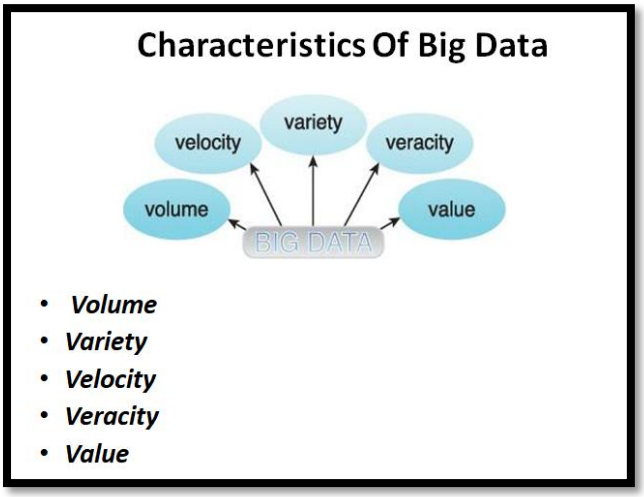
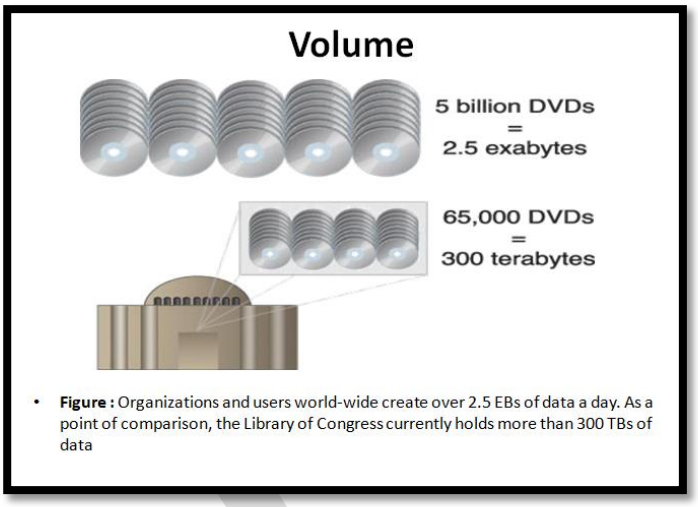
1. Briefly discuss the characteristics of Big Data?





- **Figure :** Organizations and users world-wide create over 2.5 EBs of data a day. As a point of comparison, the Library of Congress currently holds more than 300 TBs of data

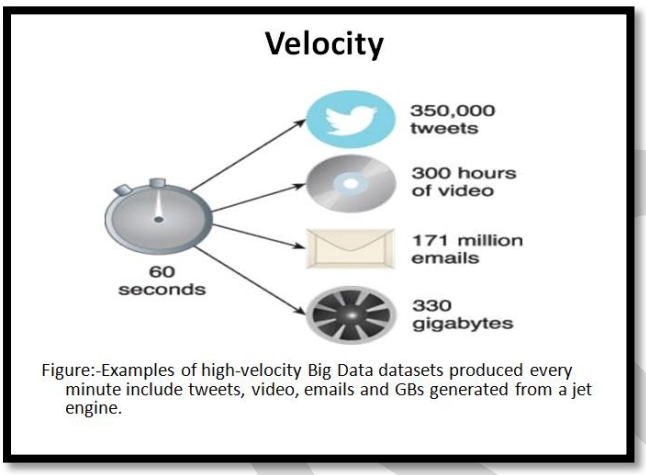
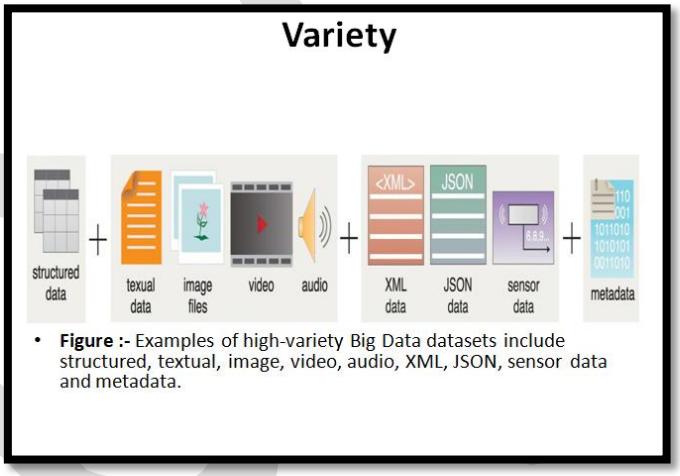


Figure:-Examples of high-velocity Big Data datasets produced every minute include tweets, video, emails and GBs generated from a jet engine.

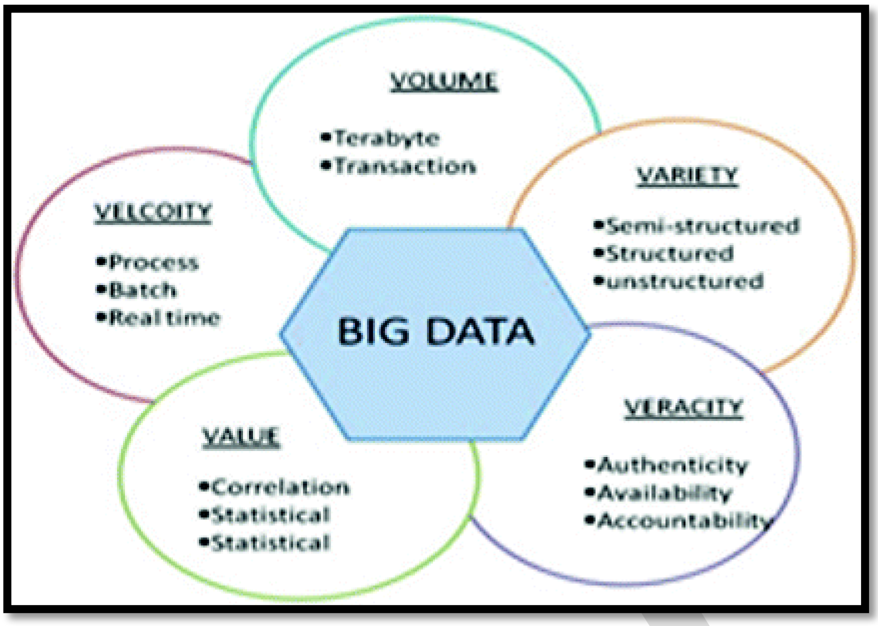


- **Figure :-** Examples of high-variety Big Data datasets include structured, textual, image, video, audio, XML, JSON, sensor data and metadata.

Veracity & VALUE

Veracity refers to the quality or fidelity of data. Data with a high signal-to-noise ratio has more veracity than data with a lower ratio.

Value is defined as the usefulness of data for an enterprise. The value characteristic is intuitively related to the veracity characteristic in that the higher the data fidelity, the more value it holds for the business.



~~~~~~

2. What are the different types of data formats give an example for each type
~~~~~~

Data formats:

Data can mean many different things, and there are many ways to classify it. Two of the more common are:

- Primary and Secondary. Primary data is data that you collect or generate. Secondary data is created by other researchers, and could be their primary data, or the data resulting from their research.
- Qualitative and Quantitative. Qualitative refers to text, images, video, sound recordings, observations, etc. Quantitative refers to numerical data.

There are typically five main categories that it can be sorted into for management purposes. The category that you choose will then have an effect upon the choices that you make throughout the rest of your data management plan.

Observational

- Captured in real-time
- Cannot be reproduced or recaptured. Sometimes called "unique data".
- Examples include sensor readings, telemetry, survey results, images, and human observation.

Experimental

- Data from lab equipment and under controlled conditions
- Often reproducible, but can be expensive to do so
- Examples include gene sequences, chromatograms, magnetic field readings, and spectroscopy.

Simulation

- Data generated from test models studying actual or theoretical systems
- Models and metadata where the input more important than the output data
- Examples include climate models, economic models, and systems engineering.

Derived or compiled

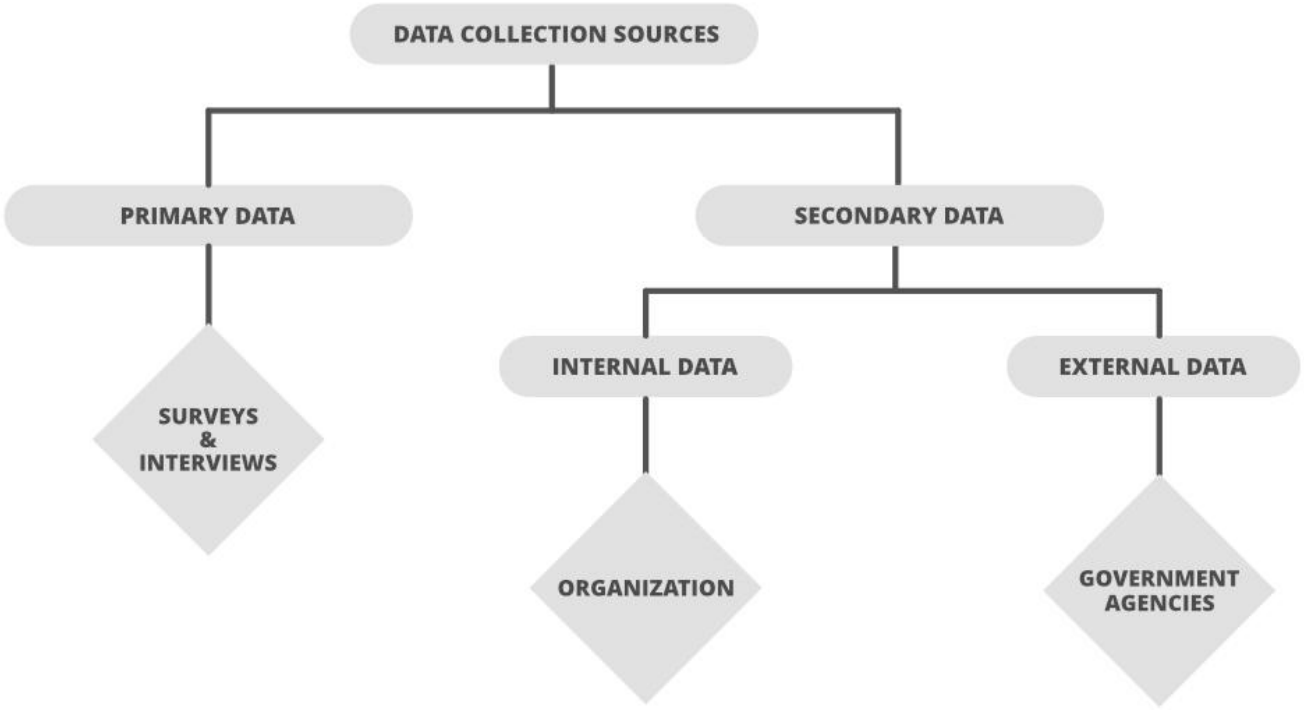
- The results of data analysis, or aggregated from multiple sources
- Reproducible (but very expensive)
- Examples include text and data mining, compiled database, and 3D models

Reference or canonical

- Fixed or organic collection datasets, usually peer-reviewed, and often published and curate
- Examples include gene sequence databanks, census data, chemical structures.

Data can come in many forms. Some common ones are text, numeric, multimedia, models, audio, code, software, discipline specific (i.e., FITS in astronomy, CIF in chemistry), video, and instrument.

3. Describe how data is gathered from different data sources



1.Primary data:

The data which is Raw, original, and extracted directly from the official sources is known as primary data. This type of data is collected directly by performing techniques such as questionnaires, interviews, and surveys. The data collected must be according to the demand and requirements of the target audience on which analysis is performed otherwise it would be a burden in the data processing.

Few methods of collecting primary data:

1. Interview method:

The data collected during this process is through interviewing the target audience by a person called interviewer and the person who answers the interview is known as the interviewee. Some basic business or product related questions are asked and noted down in the form of notes, audio, or video and this data is stored for processing. These can be both structured and unstructured like personal interviews or formal interviews through telephone, face to face, email, etc.

2. Survey method:

The survey method is the process of research where a list of relevant questions are asked and answers are noted down in the form of text, audio, or video. The survey method can be obtained in both online and offline mode like through website forms and email. Then that survey answers are stored for analyzing data. Examples are online surveys or surveys through social media polls.

3. Observation method:

The observation method is a method of data collection in which the researcher keenly observes the behavior and practices of the target audience using some data collecting tool and stores the observed data in the form of text, audio, video, or any raw formats. In this method, the data is collected directly by posting a few questions on the participants. For example, observing a group of customers and their behavior towards the products. The data obtained will be sent for processing.

4. Experimental method:

The experimental method is the process of collecting data through performing experiments, research, and investigation. The most frequently used experiment methods are CRD, RBD, LSD, FD.

CRD- Completely Randomized design is a simple experimental design used in data analytics which is based on randomization and replication. It is mostly used for comparing the experiments.
RBD- Randomized Block Design is an experimental design in which the experiment is divided into small units called blocks. Random experiments are performed on each of the blocks and results are drawn using a technique known as analysis of variance (ANOVA). RBD was originated from the agriculture sector.
LSD – Latin Square Design is an experimental design that is similar to CRD and RBD blocks but contains rows and columns. It is an arrangement of NxN squares with an equal amount of rows and columns which contain letters that occurs only once in a row. Hence the differences can be easily found with fewer errors in the experiment. Sudoku puzzle is an example of a Latin square design.
FD- Factorial design is an experimental design where each experiment has two factors each with possible values and on performing trail other combinational factors are derived.

2. Secondary data:

Secondary data is the data which has already been collected and reused again for some valid purpose. This type of data is previously recorded from primary data and it has two types of sources named internal source and external source.

Internal source:

These types of data can easily be found within the organization such as market record, a sales record, transactions, customer data, accounting resources, etc. The cost and time consumption is less in obtaining internal sources.

External source:

The data which can't be found at internal organizations and can be gained through external third party resources is external source data. The cost and time consumption is more because this contains a huge amount of data. Examples of external sources are Government publications, news publications, Registrar General of India, planning commission, international labor bureau, syndicate services, and other non-governmental publications.

Other sources:

Sensors data: With the advancement of IoT devices, the sensors of these devices collect data which can be used for sensor data analytics to track the performance and usage of products.
Satellites data: Satellites collect a lot of images and data in terabytes on daily basis through surveillance cameras which can be used to collect useful information.
Web traffic: Due to fast and cheap internet facilities many formats of data which is uploaded by users on different platforms can be predicted and collected with their permission for data analysis. The search engines also provide their data through keywords and queries searched mostly.

Reference: https://www.geeksforgeeks.org/different-sources-of-data-for-data-analysis/#:~:text=The%20data%20which%20is%20Raw,questionnaires%2C%20interviews%2C%20and%20surveys

~~~~~~

4. Discuss in detail different parsing techniques

~~~~~~

5. Discuss in detail the process of Data preparation

1. Gather data

The data preparation process begins with finding the right data. This can come from an existing data catalog or can be added ad-hoc.

2. Discover and assess data

After collecting the data, it is important to discover each dataset. This step is about getting to know the data and understanding what has to be done before the data becomes useful in a particular context.

Discovery is a big task, but Talend's data preparation platform offers visualization tools which help users profile and browse their data.

3. Cleanse and validate data

Cleaning up the data is traditionally the most time consuming part of the data preparation process, but it's crucial for removing faulty data and filling in gaps. Important tasks here include:

- Removing extraneous data and outliers.
- Filling in missing values.
- Conforming data to a standardized pattern.
- Masking private or sensitive data entries.

Once data has been cleansed, it must be validated by testing for errors in the data preparation process up to this point. Often times, an error in the system will become apparent during this step and will need to be resolved before moving forward.

4. Transform and enrich data

Transforming data is the process of updating the format or value entries in order to reach a well-defined outcome, or to make the data more easily understood by a wider audience. Enriching data refers to adding and connecting data with other related information to provide deeper insights.

5. Store data

Once prepared, the data can be stored or channeled into a third party application—such as a business intelligence tool—clearing the way for processing and analysis to take place.

Reference: https://www.talend.com/resources/what-is-data-preparation/

~~~~~~

6. Distinguish between the categories of Data Analytics?

Text Analysis- Used to discover a pattern in large data sets using databases or data mining tools. Used to transform raw data into business information.

Statistical Analysis- Includes collection, Analysis, interpretation, presentation, and modeling of data. 2 types-

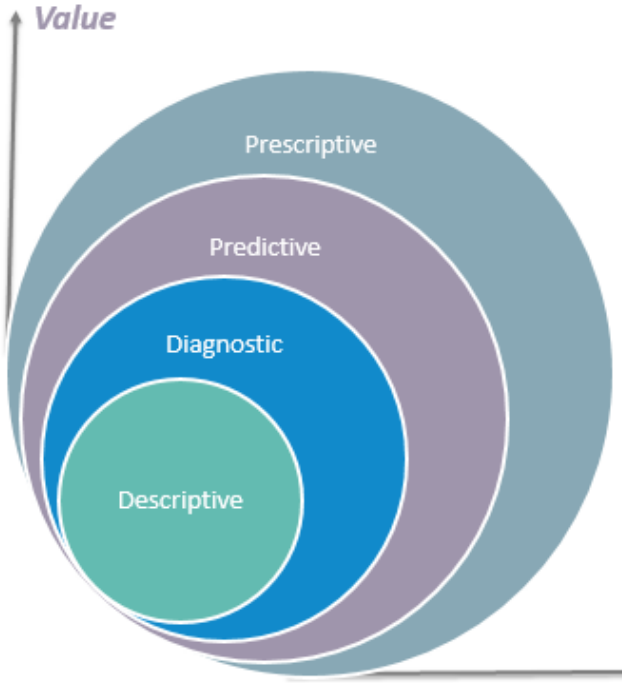
- Descriptive Analysis- Analyses complete data or a sample of summarized numerical data. It shows mean and deviation for continuous data whereas percentage and frequency for categorical data.
- Inferential Analysis - analyses sample from complete data. In this type of Analysis, you can find different conclusions from the same data by selecting different samples. Diagnostic Analysis

Diagnostic Analysis - Diagnostic Analysis shows "Why did it happen?" by finding the cause from the insight found in Statistical Analysis. This Analysis is useful to identify behavior patterns of data.

Predictive Analysis - Predictive Analysis shows "what is likely to happen" by using previous data. Analysis makes predictions about future outcomes based on current or past data.

Prescriptive Analysis - Prescriptive Analysis combines the insight from all previous Analysis to determine which action to take in a current problem or decision. Based on current situations and problems, they analyze the data and make decisions.
~~~~~~

4 types of Data Analytics



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

Diagnostic vs. Descriptive vs. Predictive vs. Prescriptive Analytics

The four main types of advanced analytics have some similarities, but are mainly defined by their differences. Here is a summary of how they operate:

Diagnostic	Descriptive	Predictive	Prescriptive
Uses historical data	Uses historical data	Uses historical data	Uses historical data
Identifies data anomalies	Reconfigures data into easy-to-read formats	Fills in gaps in available data	Estimates outcomes based on variables
Highlights data trends	Describes the state of your business operations	Creates data models	Offers suggestions about outcomes
Investigates under-lying issues	Learns from the past	Forecasts potential future outcomes	Uses algorithms, AI and machine leanring
Answers "Why" Questions	Answer "What" Questions	Answers "What Might Happen?"	Answers "If, Then" Questions

SelectHub

Reference : https://www.scnsoft.com/blog/4-types-of-data-analytics

~~~~~~

7. Write a short notes on different categories of data with examples

8. Write short notes on types of data and types of variables.

BOTH ~ SAME ANSWER

TYPES OF DATA

Quantitative – Quantitative data is defined as the value of data in the form of counts or numbers where each data-set has an unique numerical value associated with it. Ex:-Numbers, tests, counting, measuring

Types of Quantitative Data:

- Counter
- Measurement of physical objects
- Sensory calculation
- Projection of data
- Quantification of qualitative entities

Quantitative Data: Collection Methods:

- Surveys : Longitudinal Studies, Cross-sectional Studies. Fundamental Levels of Measurement, Use of Different Question Types,
- Survey Distribution and Survey Data Collection :- Email, Buy respondents, Embed survey in a website, Social distribution, QR code, SMS survey, QuestionPro app, API integration
- One-on-one Interviews:- Face-to-Face Interviews, Online/Telephonic Interviews, Computer Assisted Personal Interview

Steps to conduct Quantitative Data Analysis

- Relate measurement scales with variables
- Connect descriptive statistics with data
- Decide a measurement scale
- Select appropriate tables to represent data and analyze collected data

Advantages of Quantitative Data

- Conduct in-depth research
- Minimum bias
- Accurate results

Disadvantages of Quantitative Data

- Restricted information
- Depends on question types

Qualitative – Qualitative data is defined as the data that approximates and characterizes. Ex:-Words, images, observations, conversations, photographs. 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 interviews, 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. **Importance**- Qualitative data is important in determining the particular frequency of traits or characteristics. It allows the statistician or the researchers to form parameters through which larger data sets can be observed. Qualitative data provides the means by which observers can quantify the world around them.

Qualitative Data Collection Methods-

- One-to-One Interviews
- Focus groups
- Record keeping
- Process of observation
- Longitudinal studies
- Case studies
~~~~~~

Deductive Approach: The deductive approach involves analyzing qualitative data based on a structure that is predetermined by the researcher. A researcher can use the questions as a guide for analyzing the data.

Inductive Approach: The inductive approach, on the contrary, is not based on a predetermined structure or set ground rules/framework. It is more time consuming and a thorough approach to qualitative data analysis.

Steps to Qualitative Data Analysis

Step 1: Arrange your Data

Step 2: Organize all your Data

Step 3: Set a Code to the Data Collected

Step 4: Validate your Data

1.Accuracy of your research design or methods.

2.Reliability, which is the extent to which the methods produce accurate data consistently.

Step 5: Concluding the Analysis Process

Advantages of Qualitative Data:

- It helps in-depth analysis
- Understand what customers think
- Rich data

Disadvantages of Qualitative Data:

- Time-consuming
- Not easy to generalize
- Dependent on the researcher's skills

Ex-

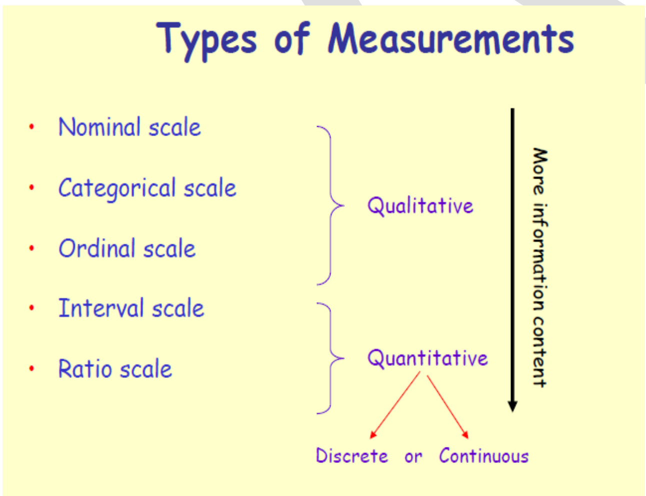
The cake is orange, blue, and black in colour (qualitative).

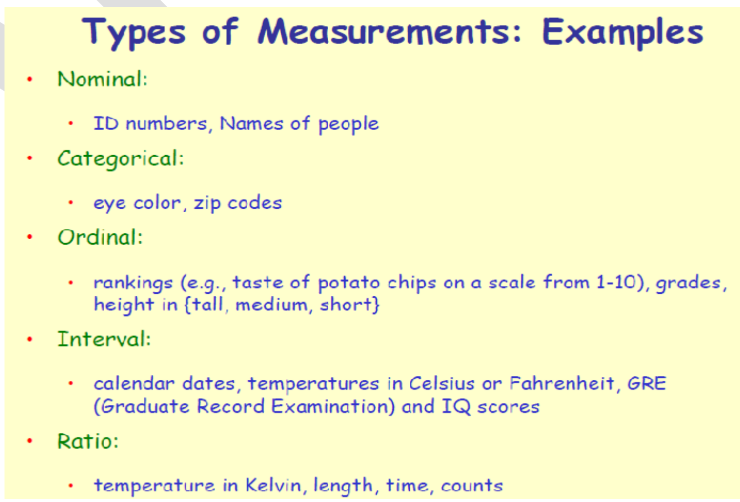
Females have brown, black, blonde, and red hair (qualitative).

Quantitative Data:- There are four cakes and three muffins kept in the basket (quantitative).

One glass of fizzy drink has 97.5 calories (quantitative).

Type of Variables





~~~~~~

9. Illustrate how Data distribution is used in Data understanding

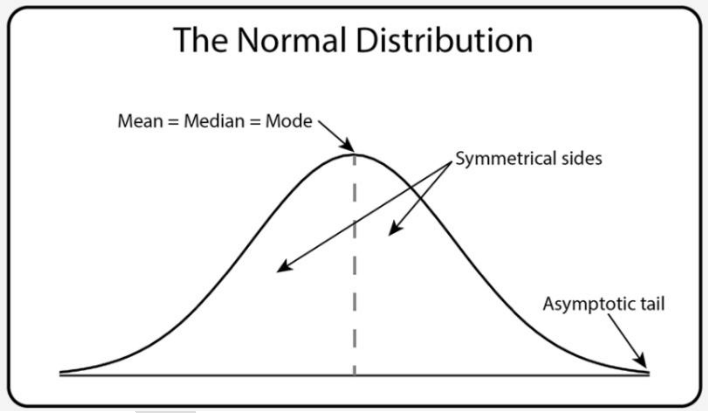
Distribution Statistics:

- One of the first steps in understanding data is to describe it using **distribution statistics.**
- These often include the sample size, mean (average), and various quantiles or percentiles.
- Looking at numbers, so we will look to visualize our data to get a better sense of the distributions.

Relative Histogram:
~~~~~~

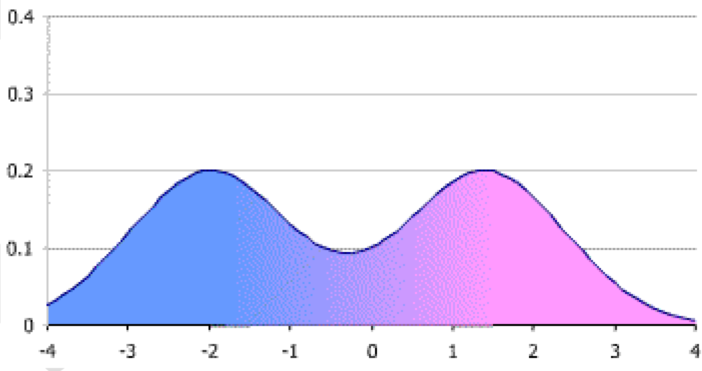
- Histograms can be thought of as bar graphs where the width of the bar describes a range of values in your data and the height of the bar describes how many observations fall into that range.
- Histograms generally have either the raw count (absolute) or the percentage (relative) of observations plotted on the vertical axis. While each will produce the same shape of bars, the percentage is often preferred for inferential purposes.

Normal Distribution:



- Often referred to by its "bell-shaped" curve.
- It offers a number of desirable statistical properties, including symmetry.
- Symmetric distributions may be split in half vertically, resulting in the left and right halves mirroring each other. Consequently, this results in the mean and the median being equal.
- In the example below, we see the distribution of our gym customers by their age. Because this data is (approximately) normally distributed, we can again tell that the average customer age is around 55. About 50% of customers are younger than 55, while about 50% are 55 or older.

Bi-modal Distribution:



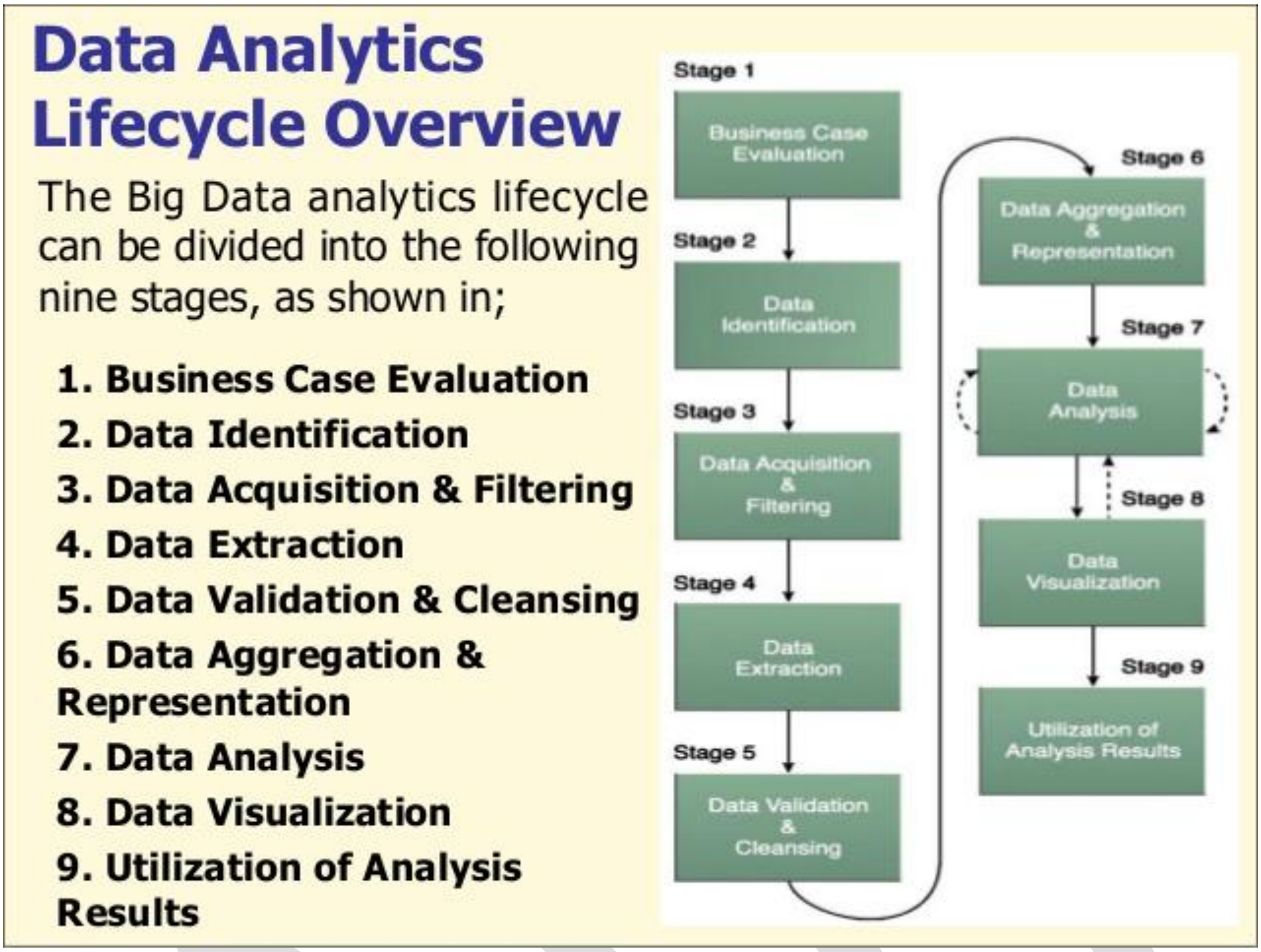
- Another descriptive measure is the mode, which is the value (or range) with the highest count of observations.
- In a histogram, the mode is easily seen as the tallest bar or "peak".
- Occasionally, we may come across data that has two distinct peaks. This is referred to as a bi-modal distribution.
- For example, the distribution of heights in a sample of adults might have two peaks, one for women and one for men.

Skewed Distribution:

- A more commonly encountered distribution in *business* applications is a **skewed distribution**, in which the data tends to start with a peak and then gradually taper off to a long tail—or vice versa.
- When the data tapers off to the right, it's referred to as a right-skewed distribution.
- Likewise, a left-skewed distribution begins with a tail and gradually increases to its peak.
- Right-skewed data is often seen in business when looking at customer distributions by sales. When this occurs, it suggests that the customers are generally lower spenders and that higher value customers are somewhat scarce.
- One consequence of right-skewed data is that the mean is larger than the median, and many times this difference can be drastic.

~~~~~~

10. With a neat sketch explain the big data life cycle
~~~~~~



~~~~~~

11. Explain the concept of Central tendency in data analysis

CENTRAL TENDENCY

Central tendency is defined as "the statistical measure that identifies a single value as representative of an entire distribution."[2] It aims to provide an accurate description of the entire data. It is the single value that is most typical/representative of the collected data. The term "number crunching" is used to illustrate this aspect of data description. The mean, median and mode are the three commonly used measures of central tendency.

Measures of Central Tendency

These are statistics that attempt to describe typical scores that reflect how the data is similar. The average is a commonly used term; in statistics this includes 3 different expressions: the mean, median and mode.

The appropriateness of which measure to use depends on the data type (see below), and its use:

Data type	Average		
	Mode	Median	Mean
Nominal	✓		
Ordinal	✓	✓	
~~~~~~

Interval/Ratio	✓	✓	✓

Reference: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3127352/

~~~~~~

12. Define data segmentation and analyse the key benefits of the data segmentation

Data Segmentation is the process of taking the data you hold and dividing it up and grouping similar data together based on the chosen parameters so that you can use it more efficiently within marketing and operations. Data segmentation is how you divide and organize your data into defined groups, so you can sort through it and view it more easily. Examples of Data Segmentation could be:

- Gender
- Customers vs. Prospects
- Industry

The key benefits of Data Segmentation are:

- You will be able create messaging that is tailored and sophisticated to suit your target market – appealing to their needs better.
- It allows you to easier conduct an analysis of your data stored in your database, helping to identify potential opportunities and challenges based within it.
- Enables you to mass-personalise your marketing communications, reducing costs.

Reference: https://www.experian.co.uk/business/glossary/data-segmentation/#:~:text=Data%20Segmentation%20is%20the%20process,efficiently%20within%20marketing%20and%20operations

~~~~~~

13. What are the different data tables used in data cleaning process?

~~~~~~

14. Mention the implementation of graphs in data cleaning process?
~~~~~~