Data Science Notes

Unit 1

*EDA

Q1] Compute mean, median , mode for 15,10,18,20,28,32.

→ Mean: Add up all the numbers in the set and divide by the total count of numbers:

15+10+18+20+28+32 / 6 = 20.5

Median: Arrange the numbers in the set in ascending or descending order and find the middle number. If there are two middle numbers, take their average: Arranging the numbers in ascending order:

{10, 15, 18, 20, 28, 32}

The median of the set is (18 + 20) / 2 = 19

Mode: The mode is the number that appears most frequently in the set: The given set has no repeating numbers, so there is no mode.

Q2] Computer mean, variance and standard deviation for 1,3,4,6,5.

→ Mean: (1 + 3 + 4 + 6 + 5) / 5 = 19 / 5 = 3.8

Variance: Find the difference between each number in the set and the mean, square the differences, add up the squares, and divide by the total count of numbers minus one:

Variance = [(1-3.8)^2 + (3-3.8)^2 + (4-3.8)^2 + (6-3.8)^2 + (5-3.8)^2] / (5-1) = (7.84+ 0.64 + 0.04 + 4.84 + 1.44) / 4 = 3.7

Standard deviation: Take the square root of the variance: Standard deviation = sqrt(3.7) = 1.923

Q3] Distinguish between primary and secondary data.

→

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

Q4] Describe various types of data collection methods.

→ (Primary and secondary)There are many different types of data collection methods used by researchers to gather information for their studies. Here are some common examples:

1. Surveys: Surveys are a popular method of data collection that involves asking respondents to answer questions about a particular topic or issue. Surveys can be conducted through various means such as face-to-face interviews, telephone interviews, online surveys, or paper questionnaires.
2. Interviews: Interviews involve one-on-one conversations between the researcher and the respondent. Interviews can be conducted in person or over the phone, and can be structured or unstructured.
3. Observations: Observations involve watching and recording behaviour in a natural or controlled setting. Observations can be participant or non-participant.
4. Experiments: Experiments can be conducted in a laboratory or in a natural setting, and can be controlled or uncontrolled.
5. Case Studies: Case studies involve in-depth investigation of a particular individual, group, or situation.
6. Focus Groups: Focus groups involve a group of people discussing a particular topic or issue in a moderated setting. Focus groups can be conducted in person or online.
7. Document Analysis: Document analysis involves analysing existing documents such as government reports, company records, or media articles to gather information on a particular topic.
8. Online Tracking: Online tracking involves collecting data on users' behaviour online. This data can be collected through cookies, tracking pixels, or other online tracking technologies.

Q5] Describe the types of observational methods used in data collection.

→ Observational methods are a type of data collection method that involve watching and recording behaviour in a natural or controlled setting.

1. Naturalistic Observation:
    - This method involves observing behaviour in a natural setting without interfering with the environment or the behaviour being observed.
    - This method is often used in field research.

1. Participant Observation:
    - This method involves the researcher actively participating in the behaviour being observed.
    - The researcher may disguise themselves as a member of the group being observed to gain a better understanding of the behaviour being studied.
    - This method is often used in ethnographic research.

2. Controlled Observation:
    - This method involves observing behaviour in a controlled setting, such as a laboratory or a simulated environment.
    - The researcher can manipulate the environment to test a hypothesis.

3. Structured Observation:
    - This method involves the researcher observing behaviour according to a predetermined set of criteria.
    - The researcher may use a checklist or a rating scale to record the frequency or intensity of specific behaviours.
    - This method is used in studies of child development or animal behaviour.

4. Unstructured Observation:
    - This method involves the researcher observing behaviour without a predetermined set of criteria.
    - The researcher may record detailed notes or create a narrative of the behaviour being observed.
    - This method is often used in exploratory research.

Q6] Explain the process of web crawling.

→ Web crawling, also known as web scraping, is the process of automatically collecting data from websites using software programs called web crawlers or spiders. Here's an overview of the web crawling process:

1. Identification of the Target Website: The first step in web crawling is to identify the target website from which data needs to be collected. This can be a single website or a list of websites.
2. Crawling: The web crawler starts by sending a request to the website's server for the web page or pages to be crawled. The server responds by sending back the HTML code for the requested page.
3. Parsing: The web crawler then parses the HTML code to extract the data of interest. This involves identifying the relevant HTML tags, such as or <p>, that contain the data to be collected.
4. Data Extraction: Once the relevant HTML tags have been identified, the web crawler extracts the data and stores it in a structured format, such as a database or a spreadsheet. The data can include text, images, links, or other types of content.
5. Follow-up Crawling: If the target website contains links to other pages or websites, the web crawler can follow these links and crawl the linked pages as well. This process can continue recursively until all relevant pages have been crawled.
6. Data Cleaning: After the data has been collected, it may require cleaning or pre-processing to remove any duplicates, irrelevant information, or formatting errors.
7. Storage: The final step in web crawling is to store the collected and cleaned data in a suitable format for analysis or further processing.

Q7] Why is data cleaning required?

→ Data cleaning is the process of detecting and correcting errors, inconsistencies, and discrepancies in data to improve its quality and accuracy. It is an essential step in the data analysis process and is required for several reasons, including:

1. Improving Accuracy: Data cleaning helps to ensure that the data is accurate and error-free. By identifying and correcting errors and inconsistencies in the data, we can have confidence in the results of our analysis.
2. Removing Duplicate Data: Data cleaning can help to identify and remove duplicate data, which can occur due to data entry errors, system glitches, or other reasons. Removing duplicate data can help to improve the efficiency of data analysis and reduce the risk of errors.
3. Handling Missing Data: Data cleaning can also help to handle missing data by imputing values for missing data points. This can help to prevent bias in the analysis and ensure that the results are representative of the entire dataset.
4. Standardizing Data: Data cleaning can help to standardize data by correcting inconsistencies in the way that data is recorded or stored. For example, data cleaning can help to standardize date formats, currency symbols, or units of measurement, making it easier to analyze the data.
5. Enhancing Data Quality: Data cleaning can help to enhance data quality by ensuring that the data is complete, consistent, and accurate. This can improve the usefulness of the data for decision-making and other applications.

Q8] How to handle missing data in a dataset?

→ Handling missing data is an important step in data analysis as it can have a significant impact on the accuracy and reliability of the results. Here are some common methods for handling missing data in a dataset:

1. Delete the Missing Data
2. Mean/Mode/Median Imputation
3. Regression Imputation
4. Multiple Imputation
5. Leave-One-Out Imputation

The choice of method for handling missing data depends on the type of data, the amount of missing data, and the research question being investigated.

It is important to carefully consider the strengths and limitations of each method and to evaluate the impact of missing data on the results of the analysis.

Q9] What is data normalization? Illustrate any one type of data normalization technique with an example.

→ Data normalization is the process of organizing and transforming data in a database to improve its efficiency and reduce redundancy.

It involves structuring the data in a standardized way so that it can be easily searched, analysed, and updated.

The goal of data normalization is to eliminate data anomalies, such as data redundancy, inconsistency, and dependency, that can lead to data quality issues and increase the risk of errors in data analysis.

There are different levels of data normalization:

1. First Normal Form (1NF
2. Second Normal Form (2NF
3. Third Normal Form (3NF)
4. Fourth Normal Form (4NF)
5. Fifth Normal Form (5NF)

For example: Suppose we have a database table called "Order Details" with the following columns:

- Order ID

- Product Name
- Product Category
- Quantity
- Price

To normalize this data to 2NF, we can split the "Order Details" table into two separate tables:

* "Orders" table with columns:

  - Order ID (primary key)
  - Quantity
  - Price

* "Products" table with columns:

  - Product Name (primary key)
  - Product Category

Q10] Smoothing by bin means & Smoothing by boundaries .

→ Smoothing by bin means:

1. Smoothing by bin means is a method used to reduce noise or variability in a dataset by grouping the data into bins and replacing the original values with the mean value of each bin.
2. Smoothing by bin also have some limitations. For example, it may not work well if the data is highly skewed or has a non-uniform distribution.
3. The steps for smoothing by bin means are as follows:

- Divide the range of the data into equally spaced intervals or bins.
- Calculate the mean value of the data within each bin.
- Replace the original values with the mean value of the bin to which they belong.

→ Smoothing by boundaries:

1. Smoothing by boundaries is a data smoothing technique that involves replacing extreme or outlier values in a dataset with the nearest non-outlier values.
2. Smoothing by boundaries also have some limitations. For example, it may not work if the dataset is highly skewed.
3. The steps for smoothing by boundaries are as follows:

- Identify the extreme values or outliers in the dataset.
- Determine the direction and distance to the nearest non-outlier values on both sides of each extreme value.
- Replace each extreme value with the nearest non-outlier value in the corresponding direction.

Q11] What is heat map? Explain its importance to visualise the existing patterns in the datasets

→ A heat map is a graphical representation of data that uses colours to display the relative magnitude or density of values across a two-dimensional grid or map.

Its importance to visualise the existing patterns in the datasets are :

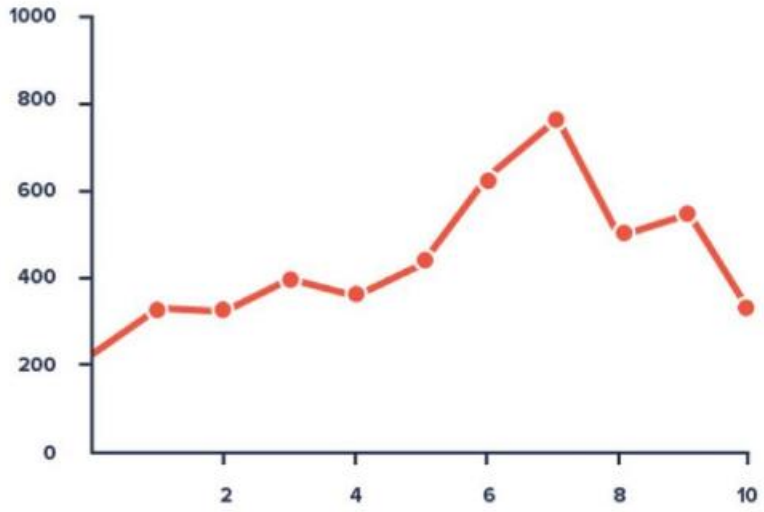
1. Heatmaps allow us to easily identify areas of high and low values in a dataset.
2. Heatmaps can highlight subtle differences in data that might be missed in other types of visualizations.
3. Heatmaps can be used to identify correlations between different variables within a dataset.

4. Heatmaps are often used in scientific research, finance, marketing, and other fields etc.
5. Heatmaps can be interactive, allowing users to explore the data in more detail.

Q12] Short note on Line Chart and Dendrograms.

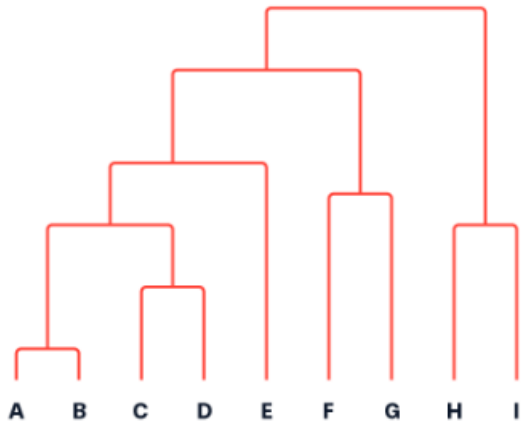
→ Line Charts:

1. A line chart is a graph that connects a series of points by drawing line segments between them.
2. Line charts are usually used in identifying the trends in data.
3. The plot () function in R is used to create the line graph
4. The basic syntax to create line chart in R is – Plot(v,type,col,xlab,ylab)
5. In a line chart, the horizontal axis represents the independent variable, such as time or category, while the vertical axis represents the dependent variable, such as a numerical value or percentage.
6. They are often used in business, economics, science, and other fields to track changes in stock prices, sales figures, weather patterns, or other types of data.



→ Dendrograms:

1. A dendrogram is a type of tree diagram used to represent hierarchical relationships between data points or objects.
2. Dendrograms are commonly used in data analysis, clustering, and classification.
3. It is also used in biology to show clustering between genes or samples.
4. It can be a column graph or a row graph.



Q13] What is Boxplot? Describe the process to identify the outlier with box plot.

→ A boxplot, also known as a box and whisker plot, is a type of data visualization that provides a graphical representation of the distribution of a dataset.

It shows the median, quartiles, range, and any outliers of a dataset.

Boxplots are created in R by using the bloxplot() function.

The process to identify the outlier with a boxplot is as follows:

1. Draw a vertical line to represent the range of the data, from the minimum to the maximum value.
2. Draw a box around the range of the middle 50% of the data, with the median (50th percentile) marked as a line inside the box.
3. Draw vertical lines, known as whiskers, extending from the box to represent the range of the data beyond the middle 50%.

4. Identify any points that fall outside the whiskers as potential outliers.

Q14] Distinguish between structured and unstructured data.

→

	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

Q15] Discuss some applications of unstructured data.

→

1. Unstructured data does not have pre- defined data model.
2. It includes texts, images, sound , videos and other formats.
3. It is difficult to search.
4. Some applications of unstructured data are :

- Natural Language Processing (NLP)
- Image and Video Analysis
- Voice and Speech Analysis
- Fraud Detection
- Health Care
- Marketing and Advertising

- Cybersecurity

Q16] What is data? State and explain different types of data.

→ Data is a collection of facts (numbers , words , measurements etc) that has been translated into a form that computer can process.

Data can take many forms, including numbers, text, images, audio, and video.

The different types of data are as follows :

- Personal data:

It is anything that is specific to you.

It covers your demographics, you location, you email address, and other identifying factors.

- Transactional data:

It is anything that requires an action to collect.

You might click on an ad, make a purchase, visit a certain web page, etc.

- Web data:

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

- Sensor data:

It is produced by objects and is often referred to as the IOT.

There can be 2 types of data :

- Primary and Secondary data
- Qualitative and Quantitative data
- Internal and External data

Q17] Write a short note on EDA.

→

1. EDA stands for exploratory data analysis
2. It involves analysing and visualizing data to gain insights
3. EDA helps to identify patterns, trends, and relationships in the data
4. It is an important step in the data analysis process
5. EDA helps to uncover important information about the data
6. Common EDA techniques include histograms, scatterplots, box plots, and correlation matrices
7. EDA helps to identify outliers or anomalies in the data that may require further investigation
8. It can also help to identify potential biases or confounding factors that may impact the accuracy or validity of the analysis
9. EDA helps to ensure that the data is properly understood and analysed in a way that produces accurate and meaningful insights.

Q18] Explain any two types of data visualization in R along with example.

→

1. There are many types of data visualization in R.
2. Some of them are: Line plot, Scatter plot, Histogram, Boxplot, Bar charts, Heatmaps, etc.
3. Two of them are explained as follows:

Box plot: A boxplot, also known as a box and whisker plot, is a type of data visualization that provides a graphical representation of the distribution of a dataset.

It shows the median, quartiles, range, and any outliers of a dataset.

Boxplots are created in R by using the bloxplot() function.

Heatmaps: A heat map is a graphical representation of data that uses colours to display the relative magnitude or density of values across a two-dimensional grid or map.

It can be used to identify correlations between different variables within a dataset.

They are often used in scientific research, finance, marketing, and other fields etc.

It can be interactive, allowing users to explore the data in more detail.

Q19] State and explain different types of data sources.

→ Same answer as that of Q16.

Q20] State various tasks done under data management.

→ Data management involves a wide range of tasks related to the acquisition, storage, processing, and distribution of data. Some of the key tasks involved in data management include:

1. Data collection:

   This involves the process of gathering data from various sources, such as surveys, administrative records, or sensor networks.

2. Data entry and coding:

   Once data has been collected, it needs to be entered into a database or other storage system. This can involve coding or categorizing data to make it easier to analyse.

3. Data cleaning:

Data cleaning involves identifying and correcting errors or inconsistencies in the data. This may include removing duplicate records, correcting misspelled or mislabelled data, or dealing with missing data.

4. Data storage:

   Data must be stored in a secure and reliable manner. This may involve using a database management system, cloud storage, or other storage solutions.

5. Data security and privacy:

   Data management also involves ensuring the security and privacy of data. This includes implementing appropriate security measures to prevent unauthorized access or breaches, as well as complying with relevant laws and regulations related to data privacy.

Q21] Explain data collection.

→

Data collection is the process of gathering information or data from various sources.

The purpose of data collection is to obtain accurate and reliable data that can be used for analysis, research, or decision-making.

Proper data collection is essential for accurate data analysis and decision-making, and can help to identify trends, patterns, and insights that can inform business strategies and improve organizational performance.

1. Defining the research question: The first step in data collection is to clearly define the research question or objective.

2. Selecting a data collection method: There are many different methods for collecting data, including surveys, interviews, observations, and administrative records.
3. Developing a data collection plan: A data collection plan should outline the procedures for collecting data, including the target population, sample size, and sampling method.
4. Pretesting the data collection instruments: Before collecting data, it is important to pretest the data collection instruments, such as surveys or questionnaires, to ensure they are valid and reliable.
5. Collecting the data: The actual data collection process involves gathering data from the selected sources. This may involve conducting surveys or interviews, observing behaviour, or extracting data from administrative records.
6. Verifying the data: Once the data has been collected, it's important to verify its accuracy and completeness.
7. Storing the data: The collected data should be stored in a secure and organized manner to ensure it is easily accessible and can be retrieved when needed.
8. Analysing the data: After the data has been collected and verified, it can be analysed to answer the research question or objective.
9. Communicating the results: Finally, the results of the data analysis should be communicated to the relevant stakeholders.

Q22] What id data cleaning? why it is done? how it is done?

→ Same answer as that of Q7

Q23] Write a note on data analysis.

→

1. Data analysis is the process of transforming raw data into useful insights and conclusions.
2. The goal of data analysis is to extract meaningful information from the data to inform decisions and solve problems.
3. Data analysis can involve a variety of techniques, including descriptive statistics, inferential statistics, data visualization, machine learning, text mining, and data mining.
4. Descriptive statistics involve summarizing and describing the main features of the data, such as the mean, median, mode, and standard deviation.
5. Data analysis is used in various fields, including business, science, healthcare, and social sciences, to improve operations, make better decisions, and develop new products or services.
6. Some common techniques used in data analysis include:

- Descriptive statistics
- Inferential statistics
- Data visualization
- Machine learning
- Text mining
- Data mining

Q24] Explain data modelling in data science.

→

1. Data modelling is the process of creating a conceptual representation of the data and the relationships between the data elements.
2. It helps to organize, analyse, and interpret complex data and build a logical framework that defines the structure, properties, and constraints of the data.
3. There are different types of data models, such as conceptual, logical, and physical models, that describe different aspects of the data.
4. Data modelling is a critical step in data science as it helps to ensure the accuracy, consistency, and completeness of the data and supports effective data analysis and decision-making.
5. In data science, data modelling involves the following steps:

- Understanding the business problem
- Identifying entities and attributes
- Defining relationships between entities
- Creating a conceptual model
- Creating a logical model
- Creating a physical model