UNIT 1

Short answer questions

1. What is big data? What are the characteristics of big data?

A. Big data is a collection of large, complex, and diverse data sets that are difficult to manage. It can be structured, semi-structured, or unstructured.

The characteristics of big data are:

i. **Volume**-How much data is there?
ii. **Variety**-How diverse are different types of data?
iii. **Velocity-**At what speed is new data generated?
iv. **Veracity**- How accurate is the data?

2. What is streaming data? Give an example.

A. The data flows into the system when an event happens instead of being loaded into a data stored in a batch is known as streaming data.

Examples: "What's happening" on Twitter, Stock price, Live Sporting event like cricket/ football.

3. List out the different forms of data.

A. Data is a collection of facts, numbers, words, some observations which can be used to learn something. The data can be raw and unprocessed which can be interpreted to become information.

Data can be in different forms like:

- Text
- Numbers
- Audio
- Video
- Images
- graphs

4. Define research goal for placement data.

A. A project starts by understanding the what, the why, and the how of your project. Continue asking questions and devising examples until you grasp the exact business expectations.

Research goal for placement data:

What?

- Analyze campus placement data to identify trends.

How?

- Use data cleaning, exploratory analysis, predictive modeling, and visualization tools to extract insights and create dashboards.

Why?

- To improve placements

5. What are errors? What are the types of errors?

A. Error refers to inaccuracies or inconsistencies in the data that can occur during the collection, processing, or storage stages.

The different types of errors are:

- **Interpretation error:** It occurs when data, information, or a situation is misunderstood.
  Examples: age> 300, percentage>100
- **Inconsistency error:** It occurs when data or information contains conflicting, contradictory, or mismatched values.
  Examples:"Male" in one place and **"M"** in other place, **"F"** and **"f"** are different.

6. Write the problems and solutions faced if we don't correct errors as early as possible?

A. The problems faced if we don't correct errors as early as possible are:

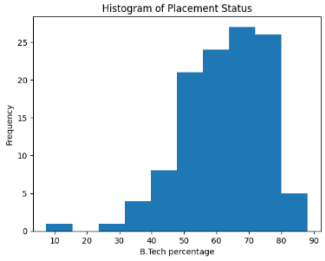
i. Costly Mistakes Due to Data Anomalies.
ii. Data Errors Can Reveal Business Process Issues.
iii. Errors may occur due to equipment or software issues.

7. Define data base and data mart with example.

A. **Data base** is used to store large amount of data in one place which helps organizations to quickly access, manage, modify, update, organize and retrieve their data. Databases are normally controlled using a database management system.
Example: college placement data.
Data mart is a subset of data warehouse. It contains small and selected part of data which is available in a large data warehouse.
Example: Placement data of **Vishnu society** is **Data Warehouse**.
Placement data of **Vishnu Institute of Technology** is **Data Base.**
Placement data of **AI&DS** department is **DataMart**.

8. What are outliers? What is the danger with outliers?

A. An outlier is an observation that seems to be distant from other observations.



The danger with outliers are:

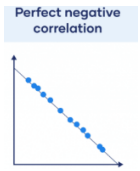
i. Outliers lead to incorrect interpretations or predictions.
ii. Important information might also miss.

9. What is correlation? Explain its types.

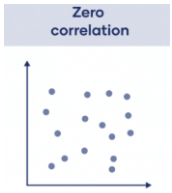
A. Correlation is a statistical measure that describes the strength and direction of the relationship between two variables.
The different types of correlations are:

i. **Positive correlation:** Both variables move in the same direction. When one variable increases, the other variable also increases.
Example: communication skills and coding skills

ii. **Negative correlation:** variables move in opposite directions. When one variable increases, the other variable decreases.
Example: B.tech result in CGPA and SSC result in percentage



iii. **Zero correlation:** no relationship between the variables. There is no clear relationship between the two variables.
Example: roll number and marks



10. Define data warehouse and data lake with examples.

A. A **data warehouse** is a repository of data from multiple data sources. Data can be in structured, unstructured or in semi-structured format.
Examples: placement data of colleges under Vishnu society.
A **data lake** can store both structured, semi-structured, and unstructured data in its raw form or unprocessed form.