I. MULTIPLE CHOICE

Choose the best answer.

1. Which of the following is a key difference between ETL and ELT processes? A. ETL stores unstructured data B. ELT requires data transformation before loading C. ETL transforms data before loading D. ELT uses fewer resources
2. What does data normalization in databases primarily achieve? A. Increases redundancy B. Improves speed of data entry C. Removes data anomalies and redundancy D. Adds unnecessary complexity
3. In stratified sampling, the population is: A. Divided randomly B. Sampled without replacement C. Grouped by shared characteristics D. Sampled only from one region
4. Which data type allows infinite measurable values within a range? A. Nominal B. Discrete C. Ordinal D. Continuous
5. Which of the following is a disadvantage of unstructured data? A. Hard to store B. Poor for sentiment analysis C. Requires special tools like NLP for analysis D. Cannot be stored in any format
6. What tool is most suitable for creating dynamic dashboards in data analytics? A. PowerPoint B. MySQL C. Power BI D. Notepad
7. Which statistical measure is least affected by outliers? A. Mean B. Mode C. Standard deviation D. Median
8. Which file type is structured and typically used in tabular datasets? A. .mp4 B. .csv C. .jpg D. .pdf
9. The primary goal of exploratory data analysis (EDA) is to: A. Build predictive models B. Clean data C. Summarize main characteristics using visuals D. Deploy models
10. Which of the following is a categorical variable? A. GPA B. Weight C. Payment method D. Temperature
11. What is the typical role of data wrangling in analytics? A. Visualizing results B. Standardizing and cleaning raw data C. Calculating statistical

models D. Exporting final data

12. Which technique is best for reducing dimensionality? A. Linear regression B. Decision tree C. Principal Component Analysis D. Naive Bayes
13. A discrete variable: A. Cannot be counted B. Can be a decimal C. Is always binary D. Takes only whole number values
14. Which sampling method is best when the population is widely spread geographically? A. Stratified B. Systematic C. Random D. Cluster
15. What does a data pipeline include? A. Only extraction B. Extraction, transformation, and loading C. Visualization tools D. Regression analysis
16. Which statistical test compares means of two groups? A. Chi-square B. ANOVA C. T-test D. Logistic regression
17. Which chart is best for showing data distribution? A. Pie chart B. Bar chart C. Histogram D. Line chart
18. The variance of a dataset indicates: A. The average value B. The middle value C. The spread of data D. The most frequent value
19. Power Pivot in Excel is best used for: A. Writing code B. Building macros C. Managing data models D. Creating documents
20. Which technique best applies to group customers based on purchasing behavior? A. Regression analysis B. Clustering C. Classification D. Dimensionality reduction

II. IDENTIFICATION

Write the correct term for each description.

1. Type of sampling that selects entire groups randomly: _____
2. A chart that shows frequency distribution: _____
3. The central number in an ordered dataset: _____
4. Process of preparing raw data for analysis: _____
5. Tool used to visualize relationships in large datasets: _____

6. A free public data repository often used by data scientists: _____
7. Step in ETL where data is moved to a target database: _____
8. Variable type that represents ordered categories: _____
9. Python library used for data manipulation: _____
10. A variable with no meaningful order: _____
11. Excel function to compute average: _____
12. SQL command to combine tables: _____
13. The extent to which data values differ from the mean: _____
14. Common platform for uploading machine learning datasets: _____
15. The most frequently occurring value in a dataset: _____
16. Type of analysis used before predictive modeling: _____
17. Data that does not follow any predefined format: _____
18. Programming language widely used in analytics and ML: _____
19. Process of visual data summary: _____
20. Sampling technique where every element has equal chance: _____
21. Term for the numeric outcome of a survey or experiment: _____
22. A software used to build dashboards and KPIs: _____
23. A continuous variable that can be precisely measured: _____
24. A logical group or category in categorical data: _____
25. Common function in SQL to retrieve data: _____
26. A small part of population used for analysis: _____
27. File format used for structured text-based data tables: _____
28. A calculated value showing how data is spread out: _____
29. An ordinal variable commonly found in surveys: _____

III. FILL IN THE BLANKS

1. _____ is the process of converting raw data into a usable format.
2. The _____ is used when analyzing the spread of data around the mean.
3. In data visualization, a _____ chart is used for showing proportions.
4. _____ sampling method is used to ensure each subgroup is represented.
5. _____ variable can take on infinite values within a range.
6. The _____ is the value that appears most frequently.
7. Data collected through online surveys is usually considered _____ data.
8. _____ is the first step in the ETL process.
9. Power BI is developed by _____.
10. _____ is the measure of how data values vary from the mean.

IV. STATISTICS

Scenario 1:

Dataset: Test scores: 85, 90, 75, 95, 100

1. Compute the mean: _____
2. Identify the median: _____
3. Determine the mode: _____
4. Calculate the variance (population): _____
5. Calculate the standard deviation (population): _____

Scenario 2:

Branch Sales (in thousands): 120, 130, 115, 125, 140, 135, 150, 160, 120, 145

6. What is the mean sales value? _____

7. What is the standard deviation (population)? _____

8. If the company wants to focus on above-average branches, how many meet this criterion? _____

9. Compute the variance: _____

10. Identify the mode: _____

Scenario 3:

Exam scores: 65, 70, 75, 80, 85, 90, 95, 100

11. After removing the highest and lowest values, compute the new mean: _____

12. Compute the variance: _____

V. SITUATIONAL ANALYSIS

Read each scenario and answer the objective questions.

Scenario 1: ETL in Retail Analytics

1. What data process model are you performing? _____
2. What phase involves removing duplicated records and standardizing date formats? _____
3. Name one tool you can use to load clean data into a database: _____

Scenario 2: Data Governance in Healthcare

4. What policy ensures proper control over access to sensitive data? _____

5. What term is used for information such as insurance number and diagnosis? _____

6. What global regulation applies to protecting patient data in the EU? _____

Scenario 3: Sampling and Survey Design

7. What sampling method is used in this case? _____

8. What type of variable is “Age Group”? _____

9. What kind of analytics is used when analyzing customer preference data? _____

Scenario 4: Data Cleaning and Consistency

10. What type of data issue does the “Province” column have? _____

11. What process is needed to make all province names uniform? _____

12. What do you call the step of making all dates follow the same format? _____

Scenario 5: Data Visualization and Interpretation

13. What chart is most appropriate to show the relationship between study hours and grades? _____

14. What method is used to test the strength of this relationship? _____

15. What technique allows prediction of final grades based on hours studied? _____

VI. TRUE OR FALSE with CORRECTION

Write TRUE if the statement is correct. If FALSE, write FALSE and correct it.

1. ETL stands for Extract, Transfer, and Load.
2. Unstructured data is best stored in relational databases.
3. Data governance involves only data backup and recovery.
4. A variable that can only take whole number values is called a discrete variable.
5. Data transformation occurs after loading in the ELT process.
6. A nominal variable has a meaningful order.
7. The mean is affected by extreme values in a dataset.
8. Power BI is used for data visualization and report creation.
9. Removing duplicate records is part of the data cleaning process.
10. Data lineage refers to tracking where data came from and how it was processed.

Let me know if you'd like this formatted into a Word doc or printable test sheet!