Data Science

1. Data Science is primarily concerned with:
   A. Extracting insights from data
   B. Web design
   C. Software development
   D. Network security
   Answer: A
2. Which role focuses on building predictive models?
   A. Business Analyst
   B. Database Administrator
   C. Data Analyst
   D. Data Scientist
   Answer: D
3. An application of Data Science in healthcare includes:
   A. Hardware optimization
   B. Disease prediction
   C. Website hosting
   D. Network routing
   Answer: B
4. The first step in the Data Science workflow is:
   A. Model building
   B. Data visualization
   C. Problem definition
   D. Data collection
   Answer: C
5. Data preparation mainly involves:
   A. Reporting results
   B. Model evaluation
   C. Deploying models
   D. Cleaning and transforming data
   Answer: D
6. Which tool is commonly used for interactive Python notebooks?
   A. Jupyter Notebook
   B. VS Code
   C. Spyder
   D. PyCharm
   Answer: A
7. SQL is mainly used for:
   A. Model deployment
   B. Data visualization
   C. Database querying
   D. Statistical testing
   Answer: C
8. A Pandas Series is best described as:
   A. Two-dimensional data
   B. Visualization object
   C. One-dimensional labeled array
   D. Database table
   Answer: C

9. A Pandas DataFrame consists of:
    A. Single values
    B. Multiple Series
    C. Only rows
    D. Only columns
    Answer: B
10. Filtering a DataFrame is used to:
    A. Merge datasets
    B. Change column names
    C. Select specific rows
    D. Plot graphs
    Answer: C
11. Sorting data in Pandas is done to:
    A. Encode categories
    B. Normalize data
    C. Clean missing values
    D. Rearrange rows based on values
    Answer: D
12. Grouping data is useful for:
    A. Removing duplicates
    B. Feature encoding
    C. Aggregation operations
    D. Visualization only
    Answer: C
13. Handling inconsistent data improves:
    A. Dataset size
    B. Visualization style
    C. Data quality
    D. Model complexity
    Answer: C
14. Matplotlib is primarily used for:
    A. Machine learning
    B. Static data visualization
    C. Database management
    D. Web scraping
    Answer: B
15. Seaborn is built on top of:
    A. NumPy
    B. Plotly
    C. Pandas
    D. Matplotlib
    Answer: D
16. Descriptive statistics are used to:
    A. Summarize data
    B. Train models
    C. Predict future outcomes
    D. Test hypotheses
    Answer: A

17. Which measure of central tendency is affected most by outliers?
    A. Mode
    B. Mean
    C. Median
    D. Range
    Answer: B
18. Standard deviation measures:
    A. Central value
    B. Data frequency
    C. Data spread
    D. Data direction
    Answer: C
19. Covariance indicates:
    A. Central tendency
    B. Data range
    C. Strength of relationship
    D. Direction of relationship
    Answer: D
20. Correlation values range between:
    A. −100 and +100
    B. 0 and 1
    C. −1 and +1
    D. −10 and +10
    Answer: C
21. Inferential statistics are used to:
    A. Visualize patterns
    B. Describe samples
    C. Clean data
    D. Draw population conclusions
    Answer: D
22. Simple random sampling ensures:
    A. Ordered samples
    B. Equal selection probability
    C. Stratified groups
    D. Clustered selection
    Answer: B
23. Data wrangling mainly focuses on:
    A. Data preparation
    B. Model evaluation
    C. Data storytelling
    D. Data deployment
    Answer: A
24. Removing outliers helps to:
    A. Increase variance
    B. Improve model reliability
    C. Increase noise
    D. Reduce dataset size
    Answer: B

25. Normalization scales data typically between:
    A. Mean and variance
    B. –1 and 1
    C. 0 and 1
    D. 1 and 100
    Answer: C
26. Feature engineering aims to:
    A. Improve model performance
    B. Reduce rows
    C. Visualize trends
    D. Remove labels
    Answer: A
27. Encoding categorical variables is required because:
    A. Categories remove noise
    B. Labels are optional
    C. Text improves accuracy
    D. Algorithms need numeric input
    Answer: D
28. Combining datasets is commonly done using:
    A. Normalization
    B. Filtering
    C. Merging or joining
    D. Encoding
    Answer: C
29. Time-series plots are best for:
    A. Spatial data
    B. Category comparison
    C. Trend analysis over time
    D. Frequency distribution
    Answer: C
30. Folium is used for:
    A. Statistical testing
    B. Line plots
    C. Bar charts
    D. Geospatial visualization
    Answer: D
31. Plotly dashboards are known for being:
    A. Offline only
    B. Interactive
    C. Static
    D. Text-based
    Answer: B
32. Pivot tables in Excel help to:
    A. Normalize data
    B. Encode data
    C. Clean missing values
    D. Summarize large datasets
    Answer: D

33. Conditional formatting is used to:
    A. Sort columns
    B. Remove duplicates
    C. Highlight patterns visually
    D. Apply formulas
    Answer: C
34. Data storytelling focuses on:
    A. Model training
    B. Raw data display
    C. Communicating insights
    D. Complex equations
    Answer: C
35. Power BI is mainly used for:
    A. Business intelligence and reporting
    B. Data collection
    C. Database design
    D. Machine learning
    Answer: A
36. Data modeling in Power BI involves:
    A. Writing Python code
    B. Scheduling refresh
    C. Creating relationships
    D. Cleaning data
    Answer: C
37. Scalars, vectors, and matrices belong to:
    A. Data mining
    B. Statistics
    C. Visualization
    D. Linear algebra
    Answer: D
38. Vectors are mainly used to represent:
    A. Text data
    B. Single values
    C. Ordered numerical data
    D. Categorical labels
    Answer: C
39. CRISP-DM is a framework for:
    A. Database management
    B. Data mining process
    C. Machine learning algorithms
    D. Data visualization
    Answer: B
40. The final stage of CRISP-DM is:
    A. Data understanding
    B. Modeling
    C. Evaluation
    D. Deployment
    Answer: D