1. What is the purpose of Exploratory Data Analysis (EDA)?

A) Cleaning the data

B) Developing predictive models

C) Visualizing data and creating hypotheses

D) Data preprocessing

\*E) Understanding data content and creating hypotheses

2. You are working on a project where you need to analyze customer preferences for movie genres (e.g., Action, Romance). What type of measurement scale is most suitable for this categorical data?

A) Nominal scale

B) Interval scale

C) Ordinal scale

D) Ratio scale

\*E) Ordinal scale

3. During the "Making sense of data" section, which type of data is described as having no quantitative value and is often used for labeling variables?

A) Continuous data

B) Ordinal data

C) Categorical data

D) Ratio data

\*E) Nominal data

4. If you are dealing with a dataset of **temperatures** in different cities, what type of data are you likely analyzing?

A) Discrete data

B) Categorical data

C) Ordinal data

D) Continuous data

\*E) Continuous data

5. You are conducting EDA on a dataset and find a variable labeled "Rank." Is this variable most likely to be considered as numerical or categorical data?

A) Numerical data

B) Categorical data

C) It could be either numerical or categorical.

D) It depends on the range of values.

\*E) Numerical data if it represents a continuous scale; otherwise, categorical data.

6. Consider a dataset that includes information about types of drugs, such as Stimulants, Opioids, and Cannabis. What kind of categorical data classification do these types fall under?

A) Binary categorical

B) Polytomous variables

C) Ordinal categorical

D) Continuous categorical

\*E) Nominal categorical

7. Imagine you have a dataset with a variable labeled "Movie Genre" that includes values like "Action," "Drama," and "Comedy." What is the primary reason for categorizing this variable as nominal data?

A) The order of movie genres matters.

B) The values are continuous.

C) The values represent a range.

D) The values are quantitative.

\*E) The values are labels without numerical importance.

8. Why is it important to understand different data distributions in data analysis?

A) It helps choose the best machine learning algorithm.

B) It ensures data security and privacy.

\*C) It aids in selecting appropriate descriptive statistics.

D) It determines the type of data visualization libraries to use.

E) It simplifies data collection processes.

9. What is the primary use of the mean and standard deviation in data analysis?

A) Identifying outliers

\*B) Summarizing normally distributed data

C) Handling categorical data

D) Interpreting summary statistics

E) Selecting visualization techniques

10. In a positively skewed data distribution, how does the mean compare to the median?

A) They are always equal.

B) The mean is always less than the median.

\*C) The mean is typically greater than the median.

D) The mean and median have no relationship.

E) The mean is typically equal to the mode.

11. What is the primary difference between the mean and median in a symmetric distribution?

A) The mean is always greater than the median.

B) The median is always greater than the mean.

\*C) They are similar in a symmetric distribution.

D) The mean and median have no relationship in a symmetric distribution.

E) The mean is typically equal to the mode.

12. When should quartiles be used as descriptive statistics in data analysis?

A) In normally distributed data

B) In symmetric data distributions

C) In bimodal data distributions

D) In negatively skewed data

\*E) In skewed data distributions

13. In a bimodal distribution, what does it mean when two modes are identified?

A) The distribution is symmetric.

B) The distribution is negatively skewed.

C) The mean is equal to the median.

\*D) There are two distinct peaks or clusters in the data.

E) The data has no central tendency.

14. What is the primary purpose of descriptive statistics in data science?

A) Developing machine learning models

B) Collecting and organizing data

C) Selecting appropriate visualization techniques

\*D) Formulating simple summaries of data

15. Which type of data visualization is suitable for showing the central tendency and spread of data, including outliers?

A) Histograms

B) Scatter plots

C) Line charts

D) Pie charts

\*E) Box plots

16. What do measures of central tendency help identify in a dataset?

A) Outliers

B) Data distributions

C) Measures of variability

D) Data transformation techniques

\*E) The central or representative value

17. How is the range calculated for a dataset?

A) By squaring the differences between data points and the mean

B) By finding the middle value in a dataset

C) By dividing the number of data points by the sum of values

D) By taking the square root of the variance

\*E) By subtracting the minimum value from the maximum value

18. What does the standard deviation provide a measure of in a dataset?

A) The central tendency

B) The variance

C) The median

D) The skewness

\*E) The average distance between data points and the mean

19. When is it most appropriate to use box plots for data visualization?

A) In positively skewed data distributions

B) In symmetric data distributions

C) In bimodal data distributions

D) In normally distributed data

\*E) To show central tendency, spread, and outliers in data.

20. Why is kurtosis important in statistics?

A) It measures the thickness of data tails.

B) It assesses the symmetry of data distribution.

C) It calculates the percentiles of a dataset.

D) It determines the standard deviation of a dataset.

\*E) It helps understand how data tails compare to those of a normal distribution.

21. What does high kurtosis indicate about a data distribution?

A) Thin tails and fewer extreme values

B) A symmetric distribution

C) A high mean value

D) Few outliers

\*E) Heavy tails and the presence of extreme values

22. If the correlation coefficient (r) is close to -1, what does it indicate about the relationship between two variables?

A) No relationship

B) A strong positive relationship

C) A strong negative relationship

D) A weak positive relationship

\*E) A strong inverse relationship

23. In correlation analysis, what does a positive correlation mean?

A) There is no relationship between variables.

B) As one variable increases, the other decreases.

\*C) As one variable increases, the other also increases.

D) The variables are unrelated.

E) The correlation coefficient cannot be positive.

24. If the correlation coefficient is 0, what can be concluded about the relationship between two variables?

\*A) There is no linear relationship between the variables.

B) The variables are strongly positively correlated.

C) The variables are strongly negatively correlated.

D) The variables are perfectly correlated.

E) The variables have a weak positive relationship.

25. Scenario: A user receives an email with a link to update their account credentials, but the email has several grammatical errors and an unusual sender address. What type of cybersecurity threat is this likely to be?

A) Malware infection

B) Insider threat

C) DDoS attack

\*D) Phishing attempt

E) Ransomware attack

26. Scenario: A company's firewall logs show a sudden surge in incoming network traffic from a foreign IP address, targeting a specific internal server. What type of cybersecurity event is occurring?

A) Insider threat

B) Phishing attack

\*C) DDoS attack

D) Ransomware attack

E) SQL injection attack