Quiz Questions: Exploratory Data Analysis

Question 1: EDA Purpose

Question: What is the PRIMARY purpose of Exploratory Data Analysis?

A) Test specific hypotheses
B) Build predictive models
C) Discover patterns and generate hypotheses
D) Clean the data
E) Create final visualizations for presentation

Correct Answer: C

Explanation: EDA is fundamentally about discovery and hypothesis generation, not confirmation. While data cleaning and visualization are part of EDA, the main goal is to explore data openly to find patterns, anomalies, and relationships that generate questions for further investigation. Hypothesis testing and model building come after EDA.

Question 2: Skewness Interpretation

Question: A distribution has mean=100, median=85, mode=80. What type of skew does it have?

A) No skew (symmetric)
B) Left skewed (negative)
C) Right skewed (positive)
D) Bimodal
E) Cannot determine from this information

Correct Answer: C

Explanation: When Mean > Median > Mode, the distribution is right-skewed (positively skewed). The mean is pulled toward the long right tail by high values. In this case: 100 > 85 > 80 indicates a right-skewed distribution, common in income, house prices, and other naturally bounded-below datasets.

Question 3: Correlation Coefficient

Question: A correlation coefficient r = -0.85 between study hours and sleep hours means:

A) Studying causes less sleep
B) Strong positive relationship

C) Weak negative relationship
D) Strong inverse relationship (more study → less sleep)
E) The variables are independent

Correct Answer: D

Explanation: $r = -0.85$ indicates a strong negative (inverse) relationship. The magnitude (0.85) shows strength (strong = $|r| > 0.7$), and the negative sign shows inverse direction. However, correlation does NOT prove causation (ruling out A). It simply means when one variable increases, the other tends to decrease.

Question 4: Box Plot Interpretation

Question: In a box plot, the box represents:

A) The full data range
B) The middle 50% of data (IQR)
C) All values within 1 standard deviation
D) The 95% confidence interval
E) Outliers only

Correct Answer: B

Explanation: The box in a box plot spans from Q1 (25th percentile) to Q3 (75th percentile), representing the Interquartile Range (IQR) - the middle 50% of data. The line inside the box is the median (Q2). Whiskers extend to show range, and points beyond whiskers are outliers.

Question 5: Histogram Bins

Question: When creating a histogram, too few bins can result in:

A) More accurate representation
B) Loss of detail in the distribution
C) Too many outliers
D) Biased mean calculation
E) Faster computation

Correct Answer: B

Explanation: Too few bins (e.g., only 3 bins) over-smooth the distribution, hiding important features like multiple peaks or specific patterns. Too many bins create noise. The right number depends on data size and distribution - common rules include Sturges' rule or using $\sqrt{n}$ bins.

Question 6: Missing Data

Question: You find 40% of income values are missing, and missingness is higher for high earners. This is:

A) MCAR (Missing Completely at Random)
B) MAR (Missing at Random)
C) MNAR (Missing Not at Random)
D) Random deletion
E) Acceptable and can be ignored

Correct Answer: C

Explanation: This is MNAR (Missing Not at Random) because the probability of missingness depends on the unobserved value itself (high earners more likely not to report income). This is problematic because simple imputation methods will bias results. MCAR would require missingness to be completely independent of all variables.

Question 7: Summary Statistics

Question: For a dataset with extreme outliers, which measure is MOST reliable for central tendency?

A) Mean
B) Median
C) Mode
D) Midrange
E) All are equally reliable

Correct Answer: B

Explanation: The median is robust to outliers because it only depends on the middle value(s), not all values. Mean is sensitive to extremes (one huge value drastically changes the mean). For highly skewed data or data with outliers, median better represents the "typical" value. Example: median income vs mean income in US data.

Question 8: Scatter Plot Pattern

Question: A scatter plot shows points in a clear curve (not line). What does this suggest?

A) No relationship exists
B) Linear relationship
C) Nonlinear relationship

D) Data error
E) Perfect correlation

Correct Answer: C

Explanation: A curved pattern indicates a nonlinear relationship. Variables are related, but not in a straight-line fashion. This is important because Pearson correlation (r) only measures linear relationships and might show weak correlation even when a strong nonlinear relationship exists. Consider log transformation or polynomial regression.

Question 9: Standard Deviation

Question: Dataset A has σ=2, Dataset B has σ=20. What can you conclude?

A) Dataset A has higher mean
B) Dataset B has more spread/variability
C) Dataset A is more accurate
D) Dataset B is normally distributed
E) Cannot compare without seeing the data

Correct Answer: B

Explanation: Standard deviation (σ) measures spread/variability. σ=20 indicates values are more spread out from the mean than σ=2. This doesn't tell us about means, accuracy, or distribution shape. Note: σ units must match data units for meaningful comparison.

Question 10: Categorical Analysis

Question: To visualize the relationship between two categorical variables, the BEST chart type is:

A) Scatter plot
B) Histogram
C) Stacked bar chart or heatmap
D) Line chart
E) Box plot

Correct Answer: C

Explanation: For two categorical variables, stacked bar charts show proportions clearly, or heatmaps/crosstabs display the contingency table. Scatter plots are for two numeric variables. Histograms for one numeric variable. Box plots for one categorical, one numeric. Line charts for time series.

Question 11: Outlier Detection

Question: Using IQR method, if Q1=50 and Q3=100, values above what threshold are outliers?

A) 100
B) 125
C) 150
D) 175
E) 200

Correct Answer: D

Explanation: IQR = Q3 - Q1 = 100 - 50 = 50 Upper outlier boundary = Q3 + 1.5 × IQR = 100 + 1.5 × 50 = 100 + 75 = 175 Values above 175 are considered outliers. Lower boundary would be Q1 - 1.5 × IQR = 50 - 75 = -25.

Question 12: Bimodal Distribution

Question: A bimodal distribution suggests:

A) Data entry errors
B) Normal distribution
C) Two distinct subgroups in the data
D) High variance
E) Outliers present

Correct Answer: C

Explanation: Bimodal (two peaks) distributions typically indicate two distinct subpopulations mixed together. Example: heights of adults (male peak ~5'10", female peak ~5'4"). This suggests you should analyze groups separately or create a categorical variable to distinguish them. Not necessarily an error - often meaningful!

Question 13: Correlation Matrix

Question: In a correlation heatmap, what does a correlation of 1.0 between two DIFFERENT variables indicate?

A) Normal situation
B) Variables are identical or one is perfect function of other
C) Strong relationship

D) The variables are independent
E) Data error

Correct Answer: B

Explanation: Correlation r=1.0 between different variables means perfect linear relationship - one variable is an exact linear function of the other (y = a + bx with no noise). This is unusual and suggests: duplicate columns, one derived from other (e.g., Celsius and Fahrenheit), or data entry error. Should investigate!

Question 14: Sample Size

Question: You have 10 million records but limited time. What's the BEST EDA approach?

A) Analyze all data no matter how long it takes
B) Random sample of reasonable size (e.g., 10,000) for initial EDA
C) Only look at first 100 rows
D) Skip EDA and go straight to modeling
E) Analyze only categorical variables

Correct Answer: B

Explanation: For very large datasets, random sampling enables quick EDA without losing representativeness. A well-chosen sample (10,000-100,000 rows) captures patterns while being manageable. After initial EDA on sample, you can: (1) verify findings on full data, (2) focus detailed analysis on interesting subsets, (3) use parallel processing for full dataset if needed.

Question 15: EDA Report

Question: An EDA report should include all EXCEPT:

A) Data quality issues found
B) Distribution visualizations
C) Final business recommendations
D) Correlation analysis
E) Unexpected patterns discovered

Correct Answer: C

Explanation: EDA is exploratory - it generates questions and hypotheses but doesn't make final recommendations. Final recommendations require confirmatory analysis, hypothesis testing, and consideration of business context beyond data patterns. EDA reports should document findings, patterns, and suggested next steps, but not jump to conclusions. Keep exploration separate from decision-making!