Quiz 1: Exploratory Data Analysis

Name: _____ Grade: ___/5

1. What is the key difference between a histogram and a density plot when visualizing the distribution of a quantitative variable?

   a) A histogram displays frequencies as bars, while a density plot uses a continuous curve to represent the distribution.
   b) A histogram is suitable for small datasets, while a density plot is better for large datasets.
   c) A histogram can only be used for quantitative variables, while a density plot can be used for both quantitative and qualitative variables.
   d) A histogram shows the exact frequencies of each value, while a density plot provides an estimated probability density.

2. Which of the following is **NOT** a recommended step in exploratory data analysis (EDA)?

   a) Identify and address extreme values in the data.
   b) Calculate and interpret summary statistics like mean, median, and standard deviation.
   c) Begin by analyzing all variables in the dataset simultaneously.
   d) Visualize the distribution of quantitative variables using histograms or density plots.

3. What is the primary distinction between experimental data and observational data concerning the source of variation in the conditioning variable (x)?

   a) Experimental data involve random sampling, while observational data rely on convenience sampling.
   b) Experimental data are collected in a controlled setting, while observational data are gathered from real-world scenarios.
   c) In experimental data, researchers manipulate the conditioning variable, while in observational data, variation in x occurs naturally or due to factors outside the researcher's control.
   d) Experimental data are typically longitudinal, while observational data are cross-sectional.

4. What is a latent variable, and how is it typically incorporated into data analysis?

   a) A latent variable is a directly observable and easily measurable variable that is central to the research question.
   b) A latent variable is an abstract concept that cannot be directly measured and is represented in the analysis using proxy variables.
   c) A latent variable is an extraneous factor that is not of primary interest but needs to be controlled for in the analysis.
   d) A latent variable is a variable that mediates the relationship between the independent and dependent variables.

5. When is it appropriate to consider dropping extreme values from a dataset during exploratory data analysis?

   a) Always, as extreme values always indicate errors in data collection.
   b) When the extreme values are confirmed errors or when they are not relevant to the research question and would unduly influence the analysis.
   c) When the dataset is particularly large, as extreme values have less impact on large samples.
   d) Never, as extreme values are crucial for understanding the true spread and nature of the data.