**1. Data Visualization**

* **Multi-box plots (Question 1)**: Primary advantage is to compare the spread, central tendency, and skewness of data across multiple groups, and to easily spot outliers within each dataset.
* **Scatter plots (Question 7)**: Best for visualizing relationships and potential correlations between two continuous variables, or assessing if one variable predicts another.
* **Histograms (Question 18, 35)**: Consist of bars representing intervals (bins) of data, with bar height reflecting the frequency of data points within each interval. Ideal for analyzing the distribution of a single variable.
* **Bar charts (Question 13)**: Best for comparing discrete categories of data, highlighting differences and trends.
* **Pie charts (Question 14)**: Useful for showing proportions (2-5 categories), but not for evenly distributed data, identifying trends over time, or adding detailed visualizations.
* **Box plots (Question 15, 33)**: Excellent for identifying outliers (e.g., points beyond whiskers, Q1 – 1.5×IQR or Q3 + 1.5×IQR) and visualizing spread, quartiles, and median. Easiest for spotting outliers compared to line charts, scatter plots, or histograms.

**2. Measures of Central Tendency**

* **Mean, Median, Mode (Question 25, 39)**: Describe the *location* (central tendency) of data.
  + **Mean**: Average, highly affected by outliers (Question 17).
  + **Median**: Middle value, less affected by outliers, not suited for nominal variables (Question 2).
  + **Mode**: Most frequent value, unaffected by outliers, not suited for nominal variables.
* **Outliers’ Impact (Question 17)**: An outlier significantly higher than other values most affects the mean, not the median or mode.

**3. Measures of Spread/Dispersion**

* **Range, Variance, Interquartile Range (IQR) (Question 20)**: Describe variability or dispersion in data, crucial for comprehensive statistical analysis.
* **Standard Deviation (Question 24, 27)**: Measures spread from the mean; high variance indicates data points are widely spread out from the mean.
* **Normal Distribution (Question 27)**: Approximately 68% of data falls within 1 SD, 95% within 2 SD, and 99.7% within 3 SD of the mean.

**4. Outliers and Binning**

* **Identifying Outliers (Question 19)**: Use IQR method: Outliers fall below Q1 – 1.5×IQR or above Q3 + 1.5×IQR (Q1 = first quartile, Q3 = third quartile).
* **Binning Outliers (Question 16)**: Involves creating categories or bins to group data points, smoothing out extreme values and reducing their impact on analysis (does not remove or scale them directly).

**5. Distribution Shapes**

* **Skewness (Question 11)**:
  + *Right-skewed*: Tail on the right, mean > median.
  + *Left-skewed*: Tail on the left, mean < median.
* **Kurtosis (Question 22, 32)**: Measures height and tail thickness of a distribution:
  + *Leptokurtic*: Sharp peak, more outliers (heavy tails).
  + *Platykurtic*: Flat peak, fewer outliers (light tails).
  + *Mesokurtic*: Normal distribution shape (moderate peak, balanced tails).
* **Normal Distribution (Question 8)**: Symmetric, single peak (unimodal), mean = median = mode. Exception: Tails taper off unequally (not true for normal distribution).
* **Uniform Distribution (Question 28)**: Values equally likely across a range, resulting in a flat, rectangular shape.

**6. Missing Data**

* **Types of Missing Data (Questions 23, 26, 36)**:
  + *Missing Completely At Random (MCAR)*: Missingness has no relationship to any data; probability of missingness is not influenced by other variables.
  + *Missing At Random (MAR)*: Missingness related to observed data (e.g., missing income data due to city of residence, not income itself).
  + *Not Missing At Random (NMAR)*: Missingness related to the missing value itself (e.g., high earners not reporting income due to sensitivity).
* **Handling Missing Data (Questions 5, 31, 34)**:
  + *Multiple Imputation*: Predicts missing values multiple times using different models, combining predictions for a final estimate.
  + *Simple Imputation*: Replaces missing values with mean, median, or mode.
  + *Ethical Concern*: Don’t remove observations with missing data without knowing why. Imputation is inappropriate if >20% of data points are missing in a variable.

**7. Descriptive Statistics**

* **Definition (Question 4)**: EDA = Exploratory Data Analysis, used to summarize and explore data patterns.
* **Importance (Question 21, 37)**: Reveals inconsistencies (outliers, missing values), identifies patterns/trends for hypothesis formulation, and informs data quality.
* **Ethical Communication (Question 9)**: Avoid reporting only selected statistics to support a narrative; include context, limitations, and transparency about data collection and potential bias.

**8. Data Transformation and Standardization**

* **Logarithmic Transformation (Question 30)**: Reduces outliers’ impact, compresses data range, stabilizes variance, useful for data spanning orders of magnitude. Exception: Does not shift data by adding a constant to address skewness.
* **Z-Score (Question 40)**: Standardizes data by calculating how many standard deviations a data point is from the mean.

**9. General Data Analysis**

* **Primary Reason for Visualizing Data (Question 10)**: Simplify complex data, reveal patterns, trends, and insights difficult to identify through numbers alone.
* **Data Quality (Question 21)**: Descriptive statistics help identify inconsistencies and anomalies (e.g., outliers, missing values) indicating data quality issues.
* **Structurally Missing Data (Question 38)**: Acceptable when intentionally left blank (e.g., pet ownership responses from non-pet owners).