Data Visualization and Summary Statistics: Principles and Methods

I. Data Summarization Methods

Two types: numerical (statistics) and graphical (visuals). Graphical summaries are easier for interpretation and communication.

II. Visualizing Qualitative Data

• Pie Chart – shows category proportions but hard for direct comparison.
• Dot Plot – easier category comparison but less intuitive for proportions.

III. Visualizing Quantitative Data

• Bar Graph – displays counts clearly across categories.
• Histogram – shows density and proportion by area, ideal for analyzing distribution.

IV. Box and Whisker Plot

Displays minimum, Q1, median, Q3, and maximum. Compact for comparing groups but less detailed than histograms.

V. Scatter Plot

Plots two variables to reveal relationships or trends. Effective for identifying correlations and non-linear patterns.

VI. Principle of Small Multiples

Displays related graphs in the same format and scale for pattern recognition and contextual comparison (e.g., temperature trends).

VII. Importance of Context

Reference bands or lines help interpret data relative to norms. Context prevents misleading conclusions.

VIII. Common Pitfalls

Flashy 3D designs distort perception. Simple 2D graphics (dot plots, bar graphs) communicate comparisons more accurately.

IX. Measures of Central Tendency

• **Mean (Average)** – Sum of all values divided by count. Best for symmetric data without outliers.
• **Median** – The midpoint where half of the values lie above and half below. Preferred for skewed data or when outliers exist.
• **Guideline:** Use mean for symmetric data, median for skewed data (e.g., income, house prices).

X. Measures of Spread and Variability

• **Percentiles & Quartiles** – Divide ordered data into 100 or 4 parts, showing data position and variation.
• **Five-Number Summary** – Minimum, Q1, Median, Q3, Maximum; forms the basis of a Box Plot.
• **Interquartile Range (IQR)** = Q3 – Q1; measures the central 50% spread of the data.
• **Standard Deviation** – Quantifies the average distance from the mean, best for symmetric data but sensitive to outliers.
• **Rule of Thumb:** Median & IQR for skewed data; Mean & SD for symmetric distributions.

Sources: Mean and Median; Percentiles, the Five-Number Summary, and Standard Deviation; Summarize All Information; Box and Whisker Plot and Scatter Plot; Pie Chart, Bar Graph, and Histograms; Providing Context Is Key; Pitfalls When Visualizing Information.