Data Visualization Exam - Solutions

Questions

1. **Primary purpose of a box plot in EDA:** A box plot (or box-and-whisker plot) is used to visually display the distribution of a dataset, showing key statistics like median, quartiles, and potential outliers. It helps in understanding the data's central tendency, spread, and skewness, and in identifying outliers.

2. **Most suitable chart for two quantitative variables:** A **scatter plot** is most suitable for showing the relationship between two quantitative variables (e.g., temperature vs. daily sales), as it plots individual data points to reveal correlations, trends, or clusters.

3. **Fundamental design requirement for bar chart value axis:** The value axis (usually the y-axis) **must start at zero** to avoid misrepresenting the proportional differences between categories.

4. **Phase to address duplicate records:** Duplicate records should be addressed during the **data cleaning/preprocessing phase**, before any analysis or modeling.

5. **Role of a 'Fact' table in dimensional data modeling:** A Fact table stores quantitative **measures (metrics)** of business processes (e.g., sales amount) and contains foreign keys linking to dimension tables, enabling analysis across dimensions.

6. **OLAP operation to view data for a single year:** The operation is **SLICE**—it selects a subset of the cube by fixing one dimension (e.g., Year = 2024).

Exercice: SoufConnect 5G Optimization

Part1: Descriptive Statistics & Boxplot

Dataset: 22, 25, 25, 28, 30, 32, 33, 35, 35, 36, 38, 40, 42, 45, 48, 50, 55, 62, 85, 110

1. Descriptive Statistics

- **Mean:** 46.8 ms
- **Median:** (36+38)/2 = 37 ms
- **Mode:** 25 ms and 35 ms
- **Population Variance:** 445.96
- **Standard Deviation:** 21.12 ms

2. Minimum, Maximum, Range, Quartiles

- **Minimum:** 22 ms
- **Maximum:** 110 ms
- **Range:** $110 - 22 = 88$ ms
- **First Quartile (Q1):** Median of first 10 values $= (30 + 32)/2 = 31$ ms
- **Third Quartile (Q3):** Median of last 10 values $= (48 + 50)/2 = 49$ ms

3. Interquartile Range (IQR) & Outliers

- **IQR:** $Q3 - Q1 = 49 - 31 = 18$ ms
- **Lower fence:** $Q1 - 1.5 \times IQR = 31 - 27 = 4$ ms
- **Upper fence:** $Q3 + 1.5 \times IQR = 49 + 27 = 76$ ms
- **Outliers:** Values below 4 ms (none) or above 76 ms $\rightarrow$ **85 ms and 110 ms** are outliers.

4. Boxplot & Shape

- **Boxplot components:**
  - Lower whisker: 22 ms (smallest non-outlier)
  - Q1: 31 ms, Median: 37 ms, Q3: 49 ms
  - Upper whisker: 62 ms (largest non-outlier)
  - Outliers: 85 ms and 110 ms plotted as individual points.
- **Shape:** The distribution is **right-skewed** (positive skew) because:
  - Median (37) is closer to Q1 (31) than to Q3 (49).
  - Right whisker is longer, and outliers are on the high end.

Part2: Histograms & Visualization Principles

1. Histograms

Histogram A (Bin width = 10 ms):

Bin (ms)	Frequency
20–29	4
30–39	7
40–49	4
50–59	2
60–69	1
70–79	0
80–89	1
90–99	0
100–109	0
110–119	1

Histogram B (Bin width = 50 ms):

Bin (ms)	Frequency
0–49	15
50–99	4
100–149	1

2. Misleading Histogram

- **Histogram B** is misleading for identifying network instability because its wide bin width (50 ms) obscures important details like the distribution shape, gaps, and outliers.

3. Violation of Visualization Principles

- Histogram B violates the principle of **appropriate bin selection**: bins that are too wide can hide variability, clusters, and outliers, leading to oversimplification.

4. Potential Misleading Impact

- Decision-makers might perceive network performance as acceptable (most data in 0–49 ms bin) and miss the high-latency issues (outliers at 85 ms and 110 ms). This could lead to underestimating network instability, delaying infrastructure improvements, and affecting customer satisfaction.

Part3: Data Warehousing & OLAP

1. Star Schema

2. OLAP Operations

1. **Total DataVolumeMB per Region in January 2026:**
    - **SLICE** on Time (Month="January", Year=2026)
    - **ROLL-UP** on Subscriber dimension (from individual to Region level)

2. **Total CostAmount for "Apple" devices with 5G during "Night" shift in 2025:**
    - **DICE** (multiple conditions: Brand="Apple", 5GEnabled="Yes", Shift="Night", Year=2025)
3. **Moving from yearly to monthly to daily view of PlanType usage:**
    - **DRILL-DOWN** on Time dimension (Year → Month → Day)
4. **Average DurationSeconds for Year 2025:**
    - **SLICE** on Time (Year=2025) and aggregate (average) over the slice.