Theoretical Questions - Data Science 2nd Assignment

Mahla Entezari

Theoretical Questions

Question 1:

Explain the difference between correlation and causation with an example from a real-world dataset.

Answer: Correlation means two things happen together, but one doesn't cause the other. Causation means one thing causes another to happen.

Example: In summer, ice cream sales and drowning incidents both go up. They are correlated. But ice cream doesn't cause drowning. Hot weather is the real cause of both.

Question 2:

(a) **What are the major types of issues found in raw data, and how do they affect analysis?**

Answer: Common problems in raw data:

- Missing values – some data is not there.
- Noisy data – random errors or strange values.
- Inconsistent data – different formats or spelling.
- Duplicate data – repeated entries.
- Outliers – very high or low values that don't fit.

These problems can lead to wrong results.

(b) **List the four major tasks of data preprocessing.**

Answer:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction

(c) **What are some methods for handling missing values in a dataset?**

Answer:

- Remove rows with missing data
- Fill with average, median, or most common value
- Predict missing values using models
- Use algorithms that handle missing data

Question 3:

Describe the 'binning' method for managing noisy data. Give an example.

Answer: Binning is grouping values into bins (or ranges) and smoothing the values inside each bin.

Example: For exam scores: 51, 52, 53, 98. You can group into two bins: 50–60 and 90–100. Replace values in each bin with their average (e.g., 52 for the first bin).

Question 4:

(a) **Discuss the importance of data quality in EDA.**

Answer: If data has errors or is not clean, your results won't be correct. Outliers and inconsistencies can give false trends or wrong patterns.

(b) **Scenario where bad data leads to wrong results:**

Answer: If some prices are wrongly entered as 0, it may look like a business is losing money.

(c) **How EDA helps fix these problems:**

Answer: EDA uses:

- Graphs to spot outliers
- Summary stats to check data ranges
- Value counts to see duplicates or format errors

Question 5:

What is normalization, and why is it important? Name three methods.
Answer: Normalization makes all values fall in a similar range, like 0 to 1. It helps compare data fairly.
Methods:

- Min-Max Scaling
- Z-score Standardization
- Decimal Scaling

Question 6:

What is the goal of data reduction, and what techniques are commonly used?
Answer: Data reduction means keeping only useful data to make analysis faster and simpler.
Techniques:

- Removing unnecessary columns
- Sampling a small part of data
- PCA (Principal Component Analysis)

Question 7:

(a) **Why is data visualization powerful for storytelling?**

Answer: It helps people quickly understand trends and insights. It makes data easy to follow.

(b) **Compare simple and storytelling visualizations.**

Answer: A simple chart shows numbers. A storytelling chart shows a clear message with titles, highlights, and colors. For example, a plain line chart vs. a line chart showing how a new policy changed sales, with notes and colors.

Question 8:

(a) **What factors decide the best chart type?**

Answer: It depends on:

- Type of data (numbers, categories)
- Goal (compare, show trend, show part of whole)
- Audience (technical or not)

(b) **How do distribution charts help in EDA?**

Answer: They show how values are spread. They answer questions like: Are most values low or high? Are there outliers?

(c) **How does a heatmap help?**

Answer: A heatmap of a correlation matrix shows which variables are related. It helps find patterns in many variables.

Question 9:

Compare bar charts, line charts, and pie charts.

Answer:

- **Bar chart:** Best for comparing categories.
- **Line chart:** Best for showing trends over time.
- **Pie chart:** Shows parts of a whole, but not good for many categories.