Data Types and Data Collection

Interview Questions

(Practice Project)

PW SKILLS

Easy

1. **Question: What is the difference between continuous and discrete data?**

Answer: Continuous data can take any value within a specified range and is often measured (e.g., height, temperature). Discrete data consists of countable values and is often represented by integers (e.g., number of students in a class).

2. **Question: What are the four levels of measurement?**

Answer: The four levels of measurement are nominal, ordinal, interval, and ratio.

3. **Question: What is qualitative data?**

Answer: Qualitative data describes attributes or qualities and is often represented in categories or labels, such as colors or types of cars.

4. **Question: What is structured data?**

Answer: Structured data is highly organized and formatted in a predefined structure, such as tables with rows and columns, making it easily searchable (e.g., databases, spreadsheets).

5. **Question: Define nominal level of measurement.**

Answer: Nominal data categorizes items into distinct groups without any inherent order, such as gender or eye color.

6. **Question: What is a primary data source?**

Answer: Primary data sources refer to information collected directly from original sources for a specific analysis, such as through surveys, interviews, or experiments.

7. **Question: What is time series data?**

Answer: Time series data is a type of longitudinal data where data points are recorded at consistent time intervals, such as daily stock prices.

Intermediate

1. **Question: How does interval data differ from ratio data?**

Answer: Interval data has meaningful intervals between values but lacks a true zero point (e.g., temperature in Celsius). Ratio data includes a true zero point and has meaningful intervals (e.g., weight, height).

2. **Question: What are the advantages of using primary data?**

Answer: Primary data is accurate, relevant, and up-to-date for the specific analysis since it is collected firsthand and tailored to the research objectives.

3. **Question: Explain the concept of semi-structured data.**

Answer: Semi-structured data does not follow a strict schema but has some organizational properties, making it easier to analyze than unstructured data (e.g., JSON files, XML files).

4. **Question: What are the key characteristics of cross-sectional data?**

Answer: Cross-sectional data is collected at a single point in time, providing a snapshot of variables for comparison across different entities.

5. **Question: Why might imbalanced data be problematic in data science?**

Answer: Imbalanced data can lead to biased models, especially in classification tasks, where the model might perform well on the majority class but poorly on the minority class.

6. **Question: What are some common methods for collecting primary data?**

Answer: Common methods include surveys, interviews, observations, experiments, and focus groups.

7. **Question: What are the disadvantages of using secondary data?**

Answer: Secondary data may not be directly relevant, might lack accuracy and reliability, and could be outdated.

Hard

1. **Question: How would you handle an imbalanced dataset in a classification problem?**

Answer: Techniques include resampling methods (e.g., oversampling the minority class, undersampling the majority class), using algorithms that are robust to imbalance (e.g., SMOTE, cost-sensitive learning), and evaluating models with metrics beyond accuracy, like precision, recall, or F1 score.

2. **Question: Describe a situation where longitudinal data analysis is preferable to cross-sectional analysis.**

Answer: Longitudinal data analysis is preferable when studying changes over time, such as tracking the health outcomes of patients over several years to understand the long-term effects of a treatment.

3. **Question: What are the challenges in analyzing unstructured data, and how can they be addressed?**

Answer: Challenges include lack of predefined structure, difficulty in searching and organizing, and complexity in processing. These can be addressed using natural language processing (NLP) techniques, machine learning algorithms, and tools like text mining and image recognition.

4. **Question: How do you determine the appropriate level of measurement for your data?**

Answer: The level of measurement is determined by the nature of the data and the type of analysis required. For example, categories without order are nominal, ordered categories are ordinal, intervals without a true zero are intervals, and data with a true zero point is ratio.

5. **Question: Can secondary data be used for predictive modeling? If so, what precautions should be taken?**
Answer: Yes, secondary data can be used for predictive modeling. Precautions include verifying the data's relevance, accuracy, and timeliness, understanding the data collection methods, and potentially cleaning or adjusting the data to fit the specific analysis.

6. **Question: What are some examples of data that could be both structured and semi-structured, and how would you approach analyzing it?**

Answer: Emails can be both structured (e.g., metadata like date, sender, recipient) and semi-structured (e.g., body text). Analyzing it involves separating structured parts for standard analysis and applying text mining techniques to the semi-structured content.

7. **Question: How would you justify the use of a ratio scale in measuring economic data, such as income or expenditure?**

Answer: The ratio scale is justified because it allows for meaningful comparison between values, supports all arithmetic operations, and includes a true zero point, making it suitable for measuring economic data like income or expenditure where zero represents the absence of income or spending.