
Question 1(a)(i)

Question:
Describe the shape of the histogram below in terms of the modality.

Answer:
The histogram appears to be unimodal.

Explanation:
A unimodal histogram has a single peak or mode, indicating that the data has one dominant frequency. This means the distribution is centered around a single value or range.

Question 1(a)(ii)

Question:
If I'm creating a frequency table from a discrete variable, what could the values in the first column be?

Answer:
The first column could contain categories or discrete values such as "Red," "Blue," "Green," "Yellow," or other categories based on the data type (e.g., toothbrush colors).

Explanation:
Discrete variables take specific, distinct values. In a frequency table, the first column lists these values/categories, while subsequent columns show their frequencies or counts.

Question 1(a)(iii)

Question:
Which of the following histograms (A-D) displays positive skewness?

Answer:
Histogram C displays positive skewness.

Explanation:
Positive skewness occurs when the tail of the distribution extends more towards higher values (to the right). This suggests a concentration of data on the lower end of the scale.

Question 1(b)(i)

Question:
What type of visualization is this? What are the marks and attributes used to encode the data?

Answer:
The visualization is a bubble chart. Marks used are circles, and attributes include position (for counties), size (to represent data like the number of patients), and color (potentially encoding the therapy type).

Explanation:
Bubble charts are used to represent data with multiple variables using attributes such as position, size, and color to show relationships and distributions.

Question 1(b)(ii)

Question:
Identify any issues with preserving the privacy of patients that this graph may raise.

Answer:
The graph may reveal sensitive details if patient counts or categories are small, leading to potential re-identification risks.

Explanation:
Privacy issues arise when individuals can be identified, even indirectly, from graphs. For example, small data categories or overlaps with other published data increase this risk.

Question 1(b)(iii)

Question:
Comment on the likely risks and suggest methods to reduce privacy risks.

Answer:
Risks: Combining detailed demographic information (e.g., blood type, ethnicity) with geographic or treatment data can enable identification of individuals.
Methods:

1. Aggregate data to larger geographic regions.
2. Use data suppression for small categories or apply differential privacy techniques.

Explanation:
Reducing data granularity and suppressing identifiable attributes help mitigate re-identification risks while retaining utility for analysis.

Question 1(c)

Question:
Suggest a method for privacy-preserving visualization and discuss potential risks.

Answer:
Method: Use a box plot to show performance distributions across modules. It aggregates data without revealing individual scores.
Risks: Outlier data might still inadvertently identify top or bottom performers.

Explanation:
Box plots summarize data effectively, minimizing individual exposure, but safeguards like binning outliers are essential to prevent privacy leaks.

Question 2(a)

Question:
Given the brief to design a system for data collection and storage, address the following:

(i) List three important questions you would ask your client about their data storage requirements.

Answer:

1. What is the estimated volume of inventory and sales data to be stored daily?
2. What level of data access and security is required (e.g., user permissions, encryption)?
3. How often do you need to query or generate reports from the data?

Explanation:
These questions help understand the scale, security, and performance requirements of the system, ensuring appropriate storage and retrieval mechanisms are designed.

(ii) Suggest a type of data storage approach to use for this project, giving a reason for your choice.

Answer:
A relational database, such as MySQL or PostgreSQL, is recommended.

Explanation:
Relational databases handle structured data well, allowing for efficient querying of inventory and sales records. Their ability to enforce relationships (e.g., between product IDs and sales transactions) ensures data consistency.

(iii) Would the inclusion of website logs and social media content interactions change your recommendation? Why or why not?

Answer:
Yes, it would change the recommendation to a hybrid system that includes a NoSQL database like MongoDB.

Explanation:
Unstructured data such as website logs and social media interactions are better handled by NoSQL databases, which provide flexibility and scalability for varied data formats.

Question 2(b)

Question:
Address the following about metadata for a favorite item of clothing:

(i) Give three examples of simple metadata.

Answer:

1. Color: Blue
2. Material: Cotton

3. Size: Medium

Explanation:
Metadata provides descriptive, administrative, or structural information about an object. Here, the metadata captures essential details about the clothing item.

(ii) Identify if each metadata element is Descriptive, Administrative, or Structural, and explain why.

Answer:

1. Color: Descriptive — describes the appearance of the item.
2. Material: Descriptive — describes the composition of the item.
3. Size: Administrative — relates to the item's categorization for inventory and logistics.

Explanation:
Metadata types are categorized based on their role in describing, managing, or structuring the data.

(iii) How would using a metadata standard change the quality of metadata, and what is a potential difficulty?

Answer:
Improvement: Using a metadata standard improves consistency, interoperability, and retrievability across datasets.
Difficulty: Enforcing a standard can be challenging when dealing with diverse data sources or stakeholders resistant to change.

Explanation:
Standards unify data representation but require alignment among contributors, which can be difficult to achieve.

Question 2(c)

Question:
Identify any possible data that may need to be handled differently due to GDPR requirements and explain why.
Answer:
Data to handle differently: Personal data from social media interactions, such as user identifiers, geolocation, or behavioral data.

Explanation:
Under GDPR, personal data requires explicit consent for processing and must be stored securely with access controls. Social media data often contains identifiers that can trace back to individuals, necessitating anonymization or pseudonymization.

Question 3(a)

Question:
Identify four different possible errors or artifacts in the dataset (q3-data.csv) and specify the tools used.

Answer:

1. **Error:** Missing values in the "Production (Tonnes)" column for specific years.
   Cell Reference: Country X, Year 2015.
   Tool: Excel/Google Sheets (using filter functions to identify blanks).
2. **Error:** Inconsistent naming of countries (e.g., "Germany" vs. "DE").
   Cell Reference: Rows with mixed formats for Germany.
   Tool: Python (using pandas to check for unique values).
3. **Error:** Negative values in "Production (Tonnes)" where production cannot be negative.
   Cell Reference: Country Y, Year 2018.
   Tool: Excel/Google Sheets (conditional formatting to highlight negatives).
4. **Error:** Duplicated entries for the same country and year.
   Cell Reference: Country Z, Years 2010 and 2011.
   Tool: Python (using pandas .duplicated() method).

Explanation:
These errors affect the quality of analysis, and tools like Excel or Python are effective in identifying and flagging such inconsistencies.

Question 3(b)

Question:
Identify how each error or artifact in Q3(a) is most likely introduced, specifying the phase from the generic data analytics pipeline.

Answer:

1. **Missing values:** Likely introduced during the **data collection phase** due to incomplete data submissions or equipment failures.
2. **Inconsistent naming:** Occurs in the **data entry phase** when different formats or conventions are used.
3. **Negative values:** Likely introduced during the **data processing phase** due to calculation errors or manual entry mistakes.
4. **Duplicated entries:** Happens in the **data integration phase** when combining datasets from multiple sources.

Explanation:
These issues stem from human error, system limitations, or poor quality control during data handling.

Question 3(c)

Question:
What data quality methods would you suggest to avoid or mitigate these errors, and why?

Answer:

1. **Missing Values:** Use imputation techniques or enforce mandatory fields during data collection. This ensures completeness without introducing bias.
2. **Inconsistent Naming:** Standardize naming conventions with predefined schemas or dictionaries. Tools like OpenRefine can automate this process.
3. **Negative Values:** Apply validation rules during data entry to flag unrealistic values. This prevents errors from entering the pipeline.
4. **Duplicated Entries:** Implement deduplication techniques using unique identifiers or composite keys during data integration.

Explanation:
These methods enhance data quality by addressing the root causes of errors, ensuring reliability and accuracy in subsequent analyses.

Question 4(a)(i)

Question:
Plan a data analytics project to analyze student feedback and assign the activities to the Generic Data Analytics Pipeline categories.

Answer:

1. **Gathering:**
    - Conducting student surveys to answer key questions.
    - Liaising with DCU Registry to get datasets from the student registration and results systems.
2. **Processing:**
    - Removing incorrect entries from the student datasets.
    - Anonymizing student comments that include identifying details.
    - Converting student words into sentiment ratings and correlating them with the field of study.
3. **Analyzing:**
    - Calculating the average satisfaction levels based on sentiment ratings.
4. **Presenting:**
    - Creating a document to share with senior university management summarizing the findings.
5. **Preserving:**

- Documenting the data formats used in the study and saving all created datasets.

Explanation:
The Generic Data Analytics Pipeline organizes tasks into logical categories, ensuring a structured approach to data analysis.

Question 4(a)(ii)

Question:
Identify a weakness (or important task not included) in the Generic Data Analytics Pipeline.

Answer:
Weakness: The pipeline does not explicitly emphasize stakeholder involvement or feedback loops during the analysis process.

Explanation:
Stakeholder engagement is crucial for interpreting findings accurately and ensuring results align with the project's objectives.

Question 4(b)

Question:
For each data attribute, select all applicable descriptions (Qualitative, Quantitative, Discrete, Continuous, Nominal, Ordinal, Interval, Ratio).

Answer:
A. Rating of temperature comfort in offices (cold, cool, perfect, warm, hot):

- **Qualitative**, **Ordinal**
  Explanation: The data is qualitative and ordinal since the categories have a meaningful order but no numerical difference.

B. Number of times a character's name is used in a TV show episode:

- **Quantitative**, **Discrete**, **Ratio**
  Explanation: The count is numeric, discrete (integer values), and ratio-scaled since zero has a meaningful interpretation (no mentions).

C. Names of pets owned by all CA682 students:

- **Qualitative**, **Nominal**
  Explanation: The data is qualitative and nominal since names have no intrinsic order or numerical meaning.

D. All winning times (in seconds) for men's 100m sprint at the Olympic Games:

- **Quantitative**, **Continuous**, **Ratio**
  Explanation: Times are numeric, continuous, and ratio-scaled because zero represents the absence of time.

Question 4(c)

Question:
Explain whether the scenario is a good example of "big data" according to the three classical characteristics.

Scenario A: Customer account, purchasing data, and engagement data from a supermarket chain's loyalty card program.

Answer:
This scenario is a good example of "big data" because:

- **Volume:** Large-scale data generated from multiple customers, purchases, and interactions.
- **Variety:** Includes structured data (accounts, transactions) and unstructured data (engagement behaviors).
- **Velocity:** High-speed data generation due to frequent transactions and real-time engagement.

Explanation:
This dataset aligns with the "3 Vs" of big data, demonstrating significant scale, complexity, and processing requirements.

Question 5(a)

Question:
Identify three possible improvements to the graph and justify your choices.

Answer:

1. **Add a descriptive title:**
    - **Improvement:** Include a clear title to convey the graph's purpose, such as "Monthly Sales Trends by Region."
    - **Justification:** A descriptive title improves interpretability and aligns with Tufte's principle of clarity in data visualization.
2. **Improve color contrast:**
    - **Improvement:** Use a color palette with higher contrast or patterns to differentiate data categories.
    - **Justification:** This ensures accessibility, particularly for colorblind viewers, adhering to universal design principles.
3. **Simplify axis labels:**
    - **Improvement:** Rotate or shorten axis labels if they are cluttered. For example, "Jan" instead of "January."
    - **Justification:** Simplified labels enhance readability without overwhelming the viewer.

Question 5(b)

Question:
Suggest an appropriate graph type for each task and provide justification.

Task A: Compare the performance of stocks in Microsoft, Apple, and Samsung over the last 5 years.
Answer:

- **Graph Type:** Line chart (CHRTS: Trend).
- **Justification:** A line chart effectively displays trends over time, highlighting changes in stock performance and enabling easy comparisons.

Task B: Explore movie commercial performance for the IMDB top 50 by director based on cost to make and ticket sales.
Answer:

- **Graph Type:** Bubble chart (CHRTS: Correlation).
- **Justification:** A bubble chart can encode three dimensions (director, cost, and sales) using position, size, and color, enabling nuanced insights into performance relationships.

Question 5(c)

Question:
Analyze the provided graphic for its main communication purpose and the design choices supporting it.

(i) What is the main communication purpose and why?)
Answer:
The main purpose of the graphic is to compare categories or trends visually, highlighting specific relationships or insights within the dataset.

Explanation:
The design prioritizes comparison and trend visibility, suggesting its purpose is to present key findings effectively to the audience.

(ii) What design choices or guidelines have been used to support this purpose?
Answer:

1. **Effective use of color:** Distinct colors are used to differentiate categories, improving clarity.
2. **Consistent scale:** Axes are scaled proportionately, ensuring accurate data representation.
3. **Minimal clutter:** The layout avoids unnecessary elements, focusing attention on the data.

Explanation:
These choices align with visualization best practices, emphasizing clarity, precision, and viewer focus.
