Question 1(a):
(i) List three important questions you would ask your client.
(ii) Describe the data and/or file formats that you are likely to use in collecting the data.
(iii) Suggest a type of database system to use for this project, giving a reason for your choice.

Answer:
(i) List three important questions you would ask your client:

1. What are the specific objectives for understanding water usage patterns and customer sentiment?
2. How frequently will the data be updated or accessed?
3. What is the expected volume and variety of data sources to be integrated into the system?

(ii) Data formats:

- Historical water usage: CSV, Excel, or database table formats.
- Water processing plant data: JSON, XML for logs and API outputs.
- Maps of water pipelines: Geospatial formats like Shapefiles or GeoJSON.
- Population information: CSV or Excel from the Central Statistics Office.
- Survey responses: JSON or CSV for structured storage.

(iii) Recommended database:

- Type: Relational Database Management System (RDBMS) like PostgreSQL.
- Reason: It can handle structured data with relationships (e.g., linking households to water usage data) and supports geospatial extensions (PostGIS) for maps and spatial queries.

Explanation:
The questions aim to align system design with the client's needs. Suggested data formats ensure compatibility with typical analysis tools. PostgreSQL supports advanced analytics and geospatial queries, making it suitable for this project.

Question 1(b):
Categorize the following data attributes as Qualitative or Quantitative; Discrete or Continuous (if appropriate) and Nominal, Ordinal, Interval, or Ratio.

(i) Type of pet (e.g., cat, dog, bird, fish)
(ii) Number of pets currently owned
(iii) Weight of the pets
(iv) Happiness of pet owners (self-rated from 1 to 5)

Answer:
(i) Type of pet:

- **Qualitative**

- **Nominal**

(ii) Number of pets currently owned:

- **Quantitative**
- **Discrete**
- **Ratio**

(iii) Weight of the pets:

- **Quantitative**
- **Continuous**
- **Ratio**

(iv) Happiness of pet owners (self-rated from 1 to 5):

- **Qualitative**
- **Ordinal**

Explanation:

- Type of pet is qualitative and nominal as it represents categories without inherent order.
- Number of pets is quantitative, discrete (integer count), and ratio since it has a true zero.
- Weight is quantitative, continuous (measured), and ratio (a true zero exists).
- Happiness is qualitative and ordinal because the ratings imply a ranking but do not specify precise intervals between ranks.

Question 1(c):
Which of the following descriptions of data ([A], [B], or [C]) are most likely to be classified as "big data"?

[A] The "Titanic" dataset showing passenger details from the final voyage of the ship.
[B] Records from Spotify of the tracks listened to by each user (est. 232M users).
[C] Sales records from the DCU merchandise store.

Answer:
[B] Records from Spotify of the tracks listened to by each user.

Explanation:
This dataset qualifies as big data due to its massive volume (232M users generating numerous records), variety (different types of data like track metadata and user interactions), and velocity (data generated continuously in real-time). The "Titanic" dataset is small and static, and DCU's sales records are likely limited in scale.

Question 1(d):
Describe the process of scraping data from a website. Give two rules to remember when using this as a data source.

Answer:

- **Process:**
    1. Identify the target website and inspect the page structure (HTML, CSS).
    2. Use tools like Python's BeautifulSoup or Scrapy to extract the desired data.
    3. Clean and format the data for analysis.
    4. Store the data in a structured format (e.g., CSV, database).
- **Rules:**
    1. Adhere to the website's terms of service and robots.txt file.
    2. Avoid overloading the website's servers by limiting the scraping frequency.

Explanation:
The process ensures efficient data extraction while the rules maintain ethical and legal compliance, protecting against misuse and potential server damage.

Question 2(a):
Using the UK data archive Data Management Lifecycle, explain how you could go about this task and give examples of data analytics tasks, methods, and tools at each stage.

Answer:

1. **Creating Data:** Collect data from water usage logs, surveys, and geospatial sources. Use tools like online forms or APIs.
2. **Processing Data:** Clean and transform data using Python (Pandas) or ETL tools (e.g., Talend).
3. **Analyzing Data:** Apply machine learning algorithms (e.g., clustering for water usage patterns) using tools like Python (scikit-learn) or R.
4. **Preserving Data:** Store processed data in secure, redundant storage like AWS S3 or on-premises databases.
5. **Giving Access to Data:** Share insights through dashboards (Tableau, Power BI) or APIs.
6. **Re-Using Data:** Enable others to access datasets and metadata for further analysis, using standards like Dublin Core.

Explanation:
Each stage involves systematic management to ensure data usability and insights while maintaining data integrity and access.

Question 2(b):
List and explain four potential issues with using metadata created by human users.

Answer:

1. **Inconsistency:** Users may use different terms for the same concept (e.g., “DOB” vs. “Date of Birth”), leading to difficulties in searching.
2. **Errors:** Typographical errors in metadata can mislead or exclude relevant data from searches.
3. **Subjectivity:** Descriptions might reflect personal interpretations, causing variations in metadata quality.
4. **Incomplete Metadata:** Users may skip providing crucial metadata, reducing data discoverability and utility.

Explanation:
Human-generated metadata is prone to errors, inconsistencies, and subjectivity, making automated data integration and searches more challenging.

Question 2(c):
Identify and explain two potential problems with making data open or using open data.

Answer:

1. **Privacy Concerns:** Open datasets might inadvertently include sensitive information, leading to privacy violations.
2. **Data Quality Issues:** Open data may be outdated, incomplete, or lack proper validation, compromising its reliability for analysis.

Explanation:
Opening data must balance transparency with ethical considerations. Ensuring data quality and privacy safeguards is essential to avoid misuse and misinterpretation.

Question 2(d):
How does HDFS prevent or limit data corruption errors?

Answer:
HDFS replicates data blocks across multiple DataNodes. If a block becomes corrupted on one node, the system retrieves it from another node with a valid copy. The NameNode monitors replication to ensure data redundancy.

Explanation:
Replication provides fault tolerance, and continuous monitoring ensures corrupted blocks are replaced, minimizing data loss or corruption risks.

Question 3(a):
(i) Give simple example metadata describing your pen.
(ii) Categorize as Descriptive, Administrative, or Structural metadata.
(iii) How could a standard be used for this metadata?
(iv) Identify one problem with enforcing a standard.

Answer:
(i) Metadata:

- Color: Blue
- Ink Type: Gel
- Manufacturer: Pilot
- Weight: 15g

(ii) Categorize as Descriptive, Administrative or Structural metadata:

- Descriptive: Color, Ink Type.
- Administrative: Manufacturer.
- Structural: Weight.

(iii) A standard like Dublin Core can define consistent attributes for similar objects, improving metadata integration and reuse.

(iv) Problem: Standards may not cover specific requirements or adapt well to unique use cases, leading to limited adoption.

Explanation:
Standards improve metadata usability but require flexibility to address diverse user needs effectively.

Question 3(b):
Two examples of data glitches and their effects on decision-making.

Answer:

1. **Duplicate Entries:** May lead to overestimation of water usage, causing unnecessary investment in infrastructure.
2. **Missing Values:** Gaps in population data can result in inaccurate forecasts for water demand, misguiding resource allocation.

Explanation:
Data glitches reduce data quality, introducing bias or errors that affect analysis outcomes and decisions.

Question 3(c):
Three possible data errors in the sample expenses table and cleaning methods.

Answer:

1. **Inconsistent Currency Format:** Use a script to standardize formats to €, e.g., 5 to €5.
2. **Missing Data (Simon’s June Expense):** Apply interpolation or impute using averages.

3. **Miscalculated Totals:** Recalculate totals based on cleaned data.

Explanation:
Cleaning ensures uniformity, completeness, and accuracy, enabling reliable analysis.

Question 3(d):
Example of sensitive data under GDPR and actions to take after theft.

Answer:

- **Sensitive Data:** Names and addresses of DCU students.
- **Actions:**
    1. Notify the Data Protection Officer (DPO) and assess breach impact.
    2. Report the breach to the Data Protection Commission within 72 hours.
    3. Inform affected individuals if their data is at risk.
    4. Secure other systems to prevent further breaches.

Explanation:
Compliance with GDPR ensures proper response, minimizing harm and legal consequences.

Question 4(a):

(i) **What visual communication goals are evident?**

- Present data insights clearly for decision-making.
- Highlight key trends or patterns effectively.

(ii) **Identify two design principles and explain how the graphic applies them.**

1. **Clarity:** The graphic uses clear labels and axis titles to convey meaning without ambiguity.
2. **Emphasis:** Important data points are highlighted using larger or bold elements to direct attention.

(iii) **The figure has been converted to greyscale. What colors would you recommend to highlight important points and why?**

- Use red for warnings or critical data and blue for positive trends or information. These colors stand out distinctly and are widely recognized.

(iv) **Identify two attributes that the graphic uses to encode data.**

1. Position along an axis (e.g., bar height or line position).
2. Size (e.g., larger elements for greater values).

Explanation:
The visual design principles enhance comprehension and make key insights accessible. Thoughtful use of color adds clarity and focus.

Question 4(b):
Identify the Gestalt principles of visualization present in images [A], [B], and [C].

Answer:

- [A] **Proximity:** Objects close to each other are perceived as groups.
- [B] **Similarity:** Elements with the same shape or color are seen as related.
- [C] **Continuity:** The viewer perceives connected points as a smooth path or line.

Explanation:
Gestalt principles describe how visual elements are organized for intuitive interpretation, improving data visualization clarity.

Question 4(c):
In visualizations, what should designers be careful of when using 2D shapes like circles to represent quantities?

Answer:
Designers must scale the **area** of the circle (not diameter) proportionally to the quantity. Misrepresenting sizes can distort data perception.

Explanation:
The human eye interprets area rather than linear dimensions; incorrect scaling may exaggerate or understate differences.

Question 4(d):
Which is of greater importance in a visualization - luminance (brightness/contrast) or color (hue)?

Answer:
Luminance is more important because brightness and contrast differences are universally perceptible, while color perception varies among individuals (e.g., colorblind users).

Explanation:
Luminance ensures accessibility and clarity across a broader audience, enhancing visualization effectiveness.

Question 5(a):

(i) **Identify two specific problems with the graph design.**

1. Overlapping data points make it difficult to distinguish values.
2. Poor axis labeling fails to clearly describe the data represented.

(ii) **Sketch an alternative graph and describe it:**

- Use a **bar chart** to show the importance of CA682 compared to other modules.
- Highlight CA682 with a distinct color (e.g., blue) to emphasize its significance.

- Label x-axis (Modules) and y-axis (Importance Rating).

(iii) **Label components on the sketch:**

- X-axis: Modules.
- Y-axis: Importance Rating.
- Title: "Importance of Modules in the Data Analytics Course."
- Marks: Bars represent importance scores.

Explanation:
A bar chart with clear labels and distinct colors improves readability and focus.

Question 5(b):
Suggest an appropriate graph type for the following tasks:

[A] **Scatter plot** to show the relationship between temperature and water consumption. Justification: Captures correlations effectively.
[B] **Line chart** to depict improvement in sales over time. Justification: Ideal for showing trends across years.
[C] **Pie chart** for the most popular method of travel. Justification: Illustrates proportions in categorical data.
[D] **Box plot** for grade distribution. Justification: Summarizes spread, median, and outliers clearly.

Explanation:
The chosen graph types align with the data structure and task goals, enhancing interpretability.

Question 5(c):
Describe the four stages of understanding when viewing a graphic or chart and explain why 3D effects in 2D graphs hinder understanding.

Answer:

- **Stages of Understanding:**
  1. Perceive: Identify visual elements.
  2. Interpret: Assign meaning to elements.
  3. Analyze: Compare and extract relationships.
  4. Decide: Form conclusions based on insights.
- **Problem with 3D Effects:**
  3D effects distort perspective, making it hard to compare values accurately. Misleading visual elements increase cognitive load, hindering interpretation.

Explanation:
Effective design simplifies perception and analysis. Avoiding unnecessary 3D effects ensures clarity and accurate comparisons.