Question 1(a)

Question: (i) List three important questions to ask the client.
(ii) Describe the data and/or file formats likely to be used for collecting the data.
(iii) Suggest a type of database system to use, giving a reason for your choice.

Answer: (i)

1. What specific metrics (e.g., duration, spending, demographics) are most valuable?
2. How often is the data updated, and what is the desired time frame for analysis?
3. Are there privacy or security concerns with the data collected?

(ii) Likely data/file formats:

- **CSV/Excel**: For structured numerical data like cash register receipts.
- **JSON/GeoJSON**: For unstructured data or location data.
- **Video/image formats (e.g., MP4, JPEG)**: For camera footage to analyse customer counts.

(iii)

- **Suggested Database System**: NoSQL database like MongoDB.
- **Reason**: The data involves multiple formats and requires flexibility in structure, which NoSQL handles better than relational databases.

Explanation:
Questions address the scope, timeline, and compliance aspects. The data formats and database choice match the multi-source and semi-structured nature of the project.

Question 1(b)

Question:
(i) Why is it useful to categorize data?
(ii) Give possible categories for given data examples.

Answer:
(i) Categorization helps organize data into meaningful groups, making it easier to analyze and interpret trends. It reduces complexity and enhances decision-making.

(ii) Categories:
a. **Gender**: Male, Female, Other.
b. **Qualification**: Leaving Cert, Bachelors, Grad Cert, Masters, PhD.
c. **Shoe Size**: Numerical ranges (e.g., <6, 6-8, >8).
d. **House Prices**: Ranges (e.g., <€200k, €200k-€500k, >€500k).
e. **Global Birth Rate**: Low (<10), Medium (10-20), High (>20).

Explanation:
Categorization simplifies analysis, e.g., filtering data based on groups like "PhD holders" or "high birth rate regions."

Question 1(c)

Question:
Describe the activities at each stage of a generic data analytics pipeline and name a tool useful for each stage.

Answer:
1. Gathering: Collecting raw data from sources like surveys, APIs, or IoT devices.

- Tool: **Python** (using libraries like requests or pandas).

2. Processing: Cleaning and transforming the data into a usable format.

- Tool: **OpenRefine** for deduplication and formatting.

3. Analysing: Applying statistical or machine learning techniques to extract insights.

- Tool: **R** or **Python** with libraries like scikit-learn or pandas.

4. Presenting: Visualizing the results through charts or dashboards.

- Tool: **Tableau** or **Power BI**.

5. Preserving: Storing the data and results for future use.

- Tool: **PostgreSQL** or **Amazon S3**.

Explanation:
Each stage addresses a critical aspect of handling data, ensuring actionable insights and reproducibility. Tools enhance efficiency and accuracy at each stage.

Question 2(a)

Question:
(i) Provide simple metadata for describing buildings.
(ii) Why is a standard useful for this metadata?
(iii) Identify a problem with enforcing a standard.

Answer:
(i) Example metadata:

1. Name: McNulty Building.
2. Location: Dublin City University, Dublin, Ireland.
3. Capacity: 500 students.
4. Year Built: 1995.

(ii) **Utility of Standards:**

- Ensures consistency and compatibility across systems for data exchange.
- Facilitates integration with existing frameworks and enhances searchability.

(iii) **Problem:**

- Standards may be overly rigid, making them difficult to adapt to unique or evolving requirements.

Explanation:
Metadata helps in cataloguing resources effectively. However, enforcing standards can be challenging due to variability in data use cases.

Question 2(b)

Question:
Explain the role of the map-reduce algorithm in the Hadoop ecosystem and its utility for large data sets.

Answer:
Role of Map-Reduce:
Map-Reduce breaks down a large dataset into smaller, manageable chunks. The "Map" step processes data in parallel across distributed nodes, while the "Reduce" step aggregates the results.

Utility:
It efficiently processes massive datasets by leveraging distributed computing, making it ideal for tasks like log analysis or indexing large-scale web data.

Explanation:
Map-Reduce's parallel processing capability and fault tolerance allow for scalable big data analysis.

Question 2(c)

Question:
(i) What is a REST API and its use for data collection?
(ii) Identify components in a given URL.
(iii) What is JSON and how is it used by REST APIs?

Answer:
(i) **REST API:** A web service interface following REST principles, enabling data retrieval or submission over HTTP. It's used to query or send structured data between systems.

(ii) URL Components:

- **Base URL:** http://www.example.com/rest.
- **Resource:** CUSTOMER.
- **Parameters:** sortBy=age and country=US.

(iii) **JSON (JavaScript Object Notation):** A lightweight, readable data-interchange format used by REST APIs to encode and decode structured data, such as records or hierarchical objects.

Explanation:
REST APIs simplify data interaction. JSON is preferred for its simplicity and compatibility with various programming languages.

Question 2(d)

Question:
Why is varied and unstructured data difficult to store in traditional relational databases?

Answer:
(i) **Varied Data:** Contains different types like text, images, and numerical data, requiring flexible schema management.
(ii) **Unstructured Data:** Lacks predefined schema, making it hard to fit into fixed relational database tables.

Explanation:
Relational databases require rigid schema definitions, unsuitable for dynamic or unstructured data like social media posts.

Question 3(a)

Question:
For each data source, identify a possible cause and consequence of poor-quality data.

Answer:
A. Census Data:

- Cause: Outdated information.
- Consequence: Misleading demographic insights.

B. Survey Data:

- Cause: Non-representative sample.
- Consequence: Biased results affecting policy decisions.

C. Traffic Data:

- Cause: Faulty sensors.
- Consequence: Underestimating road infrastructure needs.

D. Social Media Data:

- Cause: Noise/spam.
- Consequence: Misinterpretation of public opinion.

E. Predictive Model:

- Cause: Incorrect assumptions.
- Consequence: Flawed future projections.

Explanation:
Data quality issues propagate errors throughout the pipeline, compromising decision-making.

Question 3(b)

Question:
Pick one data source from Question 3(a). Provide an approach for cleaning the data and a method to enforce better quality.

Answer:
Data Source: Survey Data
Approach to Cleaning: Remove incomplete responses and standardize answers (e.g., "Yes/No" format).
Enforce Better Quality: Use digital surveys with validation (e.g., mandatory fields, dropdown menus).

Explanation:
Cleaning ensures consistent and usable data. Enforcing quality at collection reduces the need for post-processing.

Question 3(c)

Question:
Identify three data quality issues addressable using OpenRefine, and state the pipeline stage for its use.

Answer:
Issues Addressed:

1. Removing duplicates.
2. Standardizing formats (e.g., date formats).
3. Resolving typos or inconsistencies.

Pipeline Stage: Processing

Explanation:
OpenRefine is a powerful tool for pre-analysis data cleaning, ensuring clean, consistent inputs.

Question 3(d)

Question:
Can you let a colleague use patient test data for cancer research? Why or why not?

Answer:
No, unless the data is anonymized and proper ethical and legal protocols (e.g., GDPR compliance) are followed.

Explanation:
Sharing sensitive data without safeguards violates privacy laws and ethical standards.

Question 4(a)

Question:
Suggest an appropriate graph for each visualization task and justify your choice.

Answer:
A. **Scatter Plot**: Captures the relationship between vitamin C intake and cold duration.
B. **Stacked Area Chart**: Shows trends in government expenditure by category over time.
C. **Heat Map**: Visualizes disease spread geographically.
D. **Box Plot**: Compares grade distributions across modules.

Explanation:
Graphs are chosen based on their suitability for highlighting patterns, distributions, or relationships in the data.

Question 4(b)

Question:
Identify three problems with Figure 1, suggest improvements, and specify a chart type.

Answer:
Problems:

1. Cluttered labels.
2. Poor color contrast.
3. Overlapping data points.

Improvements:

- Simplify labels.
- Use a contrasting color scheme.
- Space out data points or aggregate similar values.

Chart Type: Bar Chart

Explanation:
Simplifying design and choosing appropriate visualizations enhance readability.

Question 4(c)

Question:
Identify marks and attributes used in Figure 2.

Answer:
Marks: Points (data locations) and lines (trends).
Attributes: Position, color (indicating categories), and size (value magnitude).

Explanation:
Marks represent data points, while attributes enhance interpretability.

Question 4(d)

Question:
Provide two design considerations for a data-driven Visa advertisement.

Answer:

1. Use a bold, clean layout to emphasize key data points.
2. Incorporate company branding with consistent fonts and color schemes.

Explanation:
Visual design helps focus audience attention and reinforces brand identity.

Question 5(a)

Question:
Sketch a graph using the provided data, indicating key components.

Answer:

- **Title:** "Programming Languages of Conference Attendees."
- **X-Axis:** Language names.
- **Y-Axis:** Percentage of attendees.
- **Legend:** Colors for each language.
- **Tick Marks:** Indicate percentages at 10% intervals.

Explanation:
Clear labeling ensures easy interpretation of the graph.

Question 5(b)

Question:
Explain how the Gestalt theory of proximity/similarity can be used in Q5(a).

Answer:
Group related elements (e.g., similar colors for languages with similar use cases) to highlight patterns.
Pre-attentive Processing: Yes, it allows immediate perception without detailed analysis.

Explanation:
Gestalt principles enhance visual appeal and cognitive clarity.

Question 5(c)

Question:
Describe the searchlight model of visual attention and its use in graphic design.

Answer:
Searchlight Model: Attention is directed like a focused beam to specific elements of interest.
Application: Use contrasting colors or motion to draw viewers' attention.

Explanation:
Strategic design ensures critical information is noticed first.

Question 5(d)

Question:
Name and sketch examples for categorical and relational data visualizations.

Answer:
(i) **Categorical:** Bar chart for survey results.
(ii) **Relational:** Scatter plot for correlation between variables.

Explanation:
Graph types are chosen to suit data nature and analysis goals.