CA682 Exam Semester 1 2020-2021

Q 1(a)

Full Question:
Using the topic from your CA682 visualisation assignment, apply the Generic Data Analytics Pipeline to describe how the data may have been Gathered, Processed, Analysed, Presented, and Preserved. Give a brief description of the activities at each stage (1-2 sentences) and identify any specific tools that you did or would use. If you didn't specifically perform any stage, then you can make assumptions or predictions about the actions and tools.

Answer:
The selected topic for the CA682 visualization assignment is "Global Temperature Trends Over Decades."

1. **Gathering**:
   Data was sourced from publicly available datasets like NOAA, NASA GISS, or similar repositories, which provide historical temperature data. Tools such as Python and APIs like NOAA API or direct downloads were used.

2. **Processing**:
   Data cleaning and transformation were conducted to handle missing values and reformat data for analysis. Tools like Pandas (Python) and Excel were utilized.

3. **Analysis**:
   Statistical analysis to identify trends, anomalies, and patterns was conducted using libraries such as NumPy, SciPy, and Matplotlib in Python.

4. **Presenting**:
   Interactive visualizations and graphs were created to illustrate trends and anomalies. Tools like Tableau or Matplotlib ensured clarity and interactivity.

5. **Preserving**:
   Final datasets and visualizations were saved in repositories like GitHub and stored in CSV and PNG formats to ensure longevity and reproducibility.

Explanation of the Answer:
The answer reflects the various stages of the Generic Data Analytics Pipeline, aligning with the key steps undertaken in a real-world data visualization project. Specific tools are identified for each phase to ensure clarity and applicability.

Q 1(b)

Full Question:
For each of the following data attributes (A-D), choose all applicable descriptions from Qualitative, Quantitative, Discrete, Continuous, Nominal, Ordinal, Interval, and Ratio. Marks will be deducted for including wrong choices.

A. Number of bicycles owned per household: Quantitative, Discrete, Ratio
B. Average time taken to commute each day: Quantitative, Continuous, Ratio
C. Mode of transport used to commute on Monday: Qualitative, Nominal
D. Motor vehicle safety rating (Gold, Silver, Bronze): Qualitative, Ordinal

Explanation of the Answer:

- **A:** The number of bicycles is a countable quantity, making it quantitative and discrete, with a true zero point, fitting the ratio scale.
- **B:** Average commute time is measurable and continuous, with a meaningful zero, hence quantitative and ratio.
- **C:** Mode of transport is a category without any inherent order, making it qualitative and nominal.
- **D:** Safety ratings have a rank but no measurable intervals, making them ordinal and qualitative.

Q 1(c)

Full Question:
Which of the following situations is most likely to be classified as big data:
A) Viewing data for Netflix subscribers including the show and the date watched and social media sentiment analysis responding to the show.
B) Sales data from the four DCU campus restaurants and catering facilities in 2020.
C) A download of content and metadata from my personal Twitter account.
D) Player training data (sensors and observations) from the Irish Rugby Squad.

Answer: A) Viewing data for Netflix subscribers...

Explanation of the Answer:
Netflix data involves high volume (millions of users), velocity (real-time logging), and variety (structured and unstructured). These characteristics align with big data. Other scenarios lack one or more of these attributes.

Q 2(a)

Full Question:
Given the following brief to design a system for a data collection task:

Brief:
You are preparing a report on the impact of COVID-19 restrictions on working and commuting behavior in Ireland during 2020, comparing it to surveys and records from 2019 and 2009.

1. List three (3) important questions you would ask your client.
2. Describe the data and/or specific file formats you are likely to use in collecting and storing the data.
3. Suggest a type of database storage approach to use for this project, giving a reason for your choice.

Answer:

1. **Questions to Ask the Client:**

- What specific metrics should the report focus on (e.g., hours worked from home, vehicle traffic)?
- Are there specific regions or demographics of interest?
- How frequently should the data be updated for analysis?

2. **Data and File Formats:**
   - Pedestrian footfall: JSON or CSV from APIs or sensors.
   - Traffic data: Video files and CSVs summarizing vehicle counts.
   - Survey data: Excel or Google Sheets with qualitative and quantitative responses.
3. **Database Storage Approach:**
   A relational database (e.g., PostgreSQL) for structured data like surveys and traffic counts, supplemented by a document store (e.g., MongoDB) for unstructured data like videos. This hybrid approach ensures efficient querying and scalability.

Explanation of the Answer:
The questions ensure alignment with project goals. The file formats match common sources and ensure compatibility. The hybrid database approach addresses both structured and unstructured data needs effectively.

Q 2(b)

Full Question:
Give simple example metadata (at least 5 elements) describing your mobile phone (or computer if you don't have a mobile phone).

Answer:

1. **Device Name:** John's MacBook Pro (Descriptive)
2. **Serial Number:** ABC12345XYZ (Administrative)
3. **Processor Type:** Intel Core i7 (Descriptive)
4. **Storage Capacity:** 512GB (Structural)
5. **OS Version:** macOS Monterey 12.0.1 (Descriptive)

How using a standard improves quality:
Using a standard ensures consistency, easier integration, and better interoperability of data. For instance, consistent naming conventions avoid ambiguity.

Potential Difficulty with Standards:
Enforcing compliance, especially across diverse sources, can be challenging. Some legacy systems might not support updated standards.

Q 2(c)

Full Question:
Describe the process of scraping data from a website. Give two (2) rules that you should remember when using this as a data source.

Answer:
Process:

1. Identify the target website and inspect its structure using tools like the browser developer console.
2. Write a script using tools like BeautifulSoup or Selenium to extract the required data.
3. Clean and format the scraped data for analysis.

Rules:

1. Ensure compliance with the website's Terms of Service to avoid legal issues.
2. Avoid overloading the server by limiting requests (e.g., use a delay between requests).

Explanation of the Answer:
Web scraping allows for dynamic data collection but requires careful consideration of ethical and legal guidelines to avoid misuse.

Q 3(a)

Full Question:
Identify three different possible errors or artifacts in the dataset linked above. Give the tool or tools you used. You may use any tool that you like.

Answer:

1. Missing data: There may be missing values in critical columns like injuries or deaths.
2. Inconsistent age groupings: Age ranges might overlap or not follow a consistent pattern.
3. Incorrect aggregation: Totals for injuries or deaths may not align with subtotals for different groups.

Tools Used:
Pandas in Python was used to explore and summarize the dataset.

Explanation of the Answer:
The errors were identified during inspection, where patterns and missing data anomalies were noted. Pandas provides methods like isnull() and describe() for detecting such issues.

Q 3(b)

Full Question:
Identify how each error or artifact is most likely to have been introduced, specifying the phase from the generic data analytics pipeline. State any assumptions.

Answer:

1. Missing data likely occurred during the **data gathering** phase due to incomplete surveys or reporting errors.
2. Inconsistent age groupings could stem from the **processing** phase due to lack of standardization.
3. Aggregation errors may have been introduced during the **analysis** phase by miscalculating totals or subtotals.

Explanation of the Answer:
The data lifecycle phases often introduce distinct errors. Assumptions are based on typical workflows for processing and reporting data.

Q 3(c)

Full Question:
What data quality methods would you suggest using to either avoid or mitigate the error? Why would your suggestion improve data quality?

Answer:

1. **Missing Data**: Imputation methods or additional data collection to fill gaps.
2. **Inconsistent Age Groups**: Standardizing age ranges before data entry.
3. **Incorrect Aggregation**: Double-checking calculations using automated scripts.

Explanation of the Answer:
These methods ensure consistency, accuracy, and completeness, which improve the reliability and usability of the dataset.

Q 3(d)

Full Question:
Can you identify any potential personal or sensitive data in the provided sample dataset? Why or why not?
What process should you follow if you want to legally work with personal or sensitive data?

Answer:

1. No direct personal data (e.g., names, addresses) was identified in the dataset. However, aggregated data for small groups might indirectly reveal sensitive information.
2. To work with such data, anonymization and strict adherence to GDPR policies are required, including obtaining ethical approval.

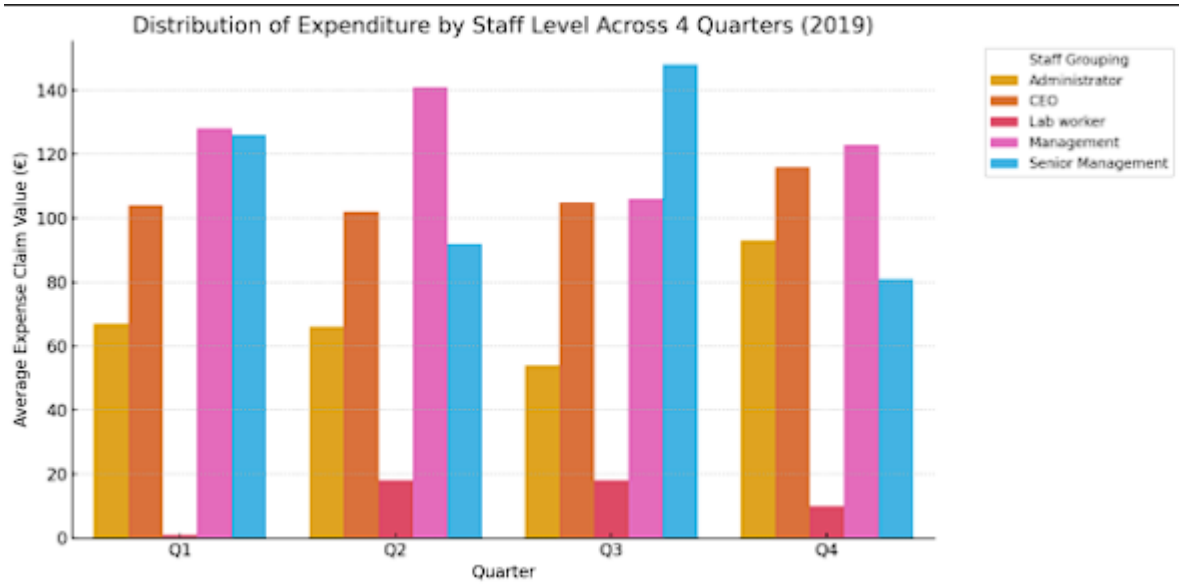
Explanation of the Answer:
Aggregated data still carries the risk of re-identification, necessitating proper handling and compliance.

Question 4

Q 4(a)

Full Question:
You are tasked with creating a visualization showing "the distribution of expenditure by staff level across 4 quarters in 2019." Use the sample data to create a graph to achieve this goal.

Answer:
Let me create a bar chart to represent the distribution effectively.



Explanation of the Visualization:
The bar chart illustrates the expenditure distribution by staff level (e.g., Lab Worker, Administrator) across the four quarters of 2019. Different staff groupings are represented by colors, and values are plotted to show trends and disparities.

The visualization was chosen for its clarity in comparing groups over time, highlighting trends effectively.

Q 4(b)

Full Question:
Identify the specific type of each of the three graphs (A-C).

Answer:

- **A:** Line chart
- **B:** Stacked bar chart
- **C:** Scatter plot

Explanation of the Answer:

- **A:** A line chart is used to show trends over time, typically connecting data points with a line.
- **B:** A stacked bar chart visually compares categories and subcategories within them.
- **C:** A scatter plot is used to show relationships or correlations between two numerical variables.

Q 4(c)

Full Question:
Given the scenario below, choose the most appropriate graph type to visualize the message and justify your choice referencing the message and the probable data types, indicating the marks and visual attributes that will be used to encode the data.

Scenario: Compare the energy efficiency ratings (0-5 star) vs price (in Euro) for consumer electronics such as fridges, ovens, heaters, etc.

Answer:
A scatter plot is the most appropriate graph type.

Explanation of the Answer:

- **Why Scatter Plot?** It effectively visualizes relationships or distributions between two quantitative variables, such as price and energy efficiency.
- **Marks:** Points will represent individual consumer electronics items.
- **Visual Attributes:** The x-axis encodes price (quantitative, continuous), and the y-axis encodes energy efficiency ratings (ordinal, discrete). Additional attributes like color or size can show product categories or popularity.

Q 4(d)

Full Question:
Which of the following images (A-D) illustrates the use of preattentive features?

Answer:
A

Explanation of the Answer:
Preattentive features like color, size, orientation, or shape allow for immediate perception without conscious effort. In Image A, these features are clearly utilized to highlight or differentiate data elements effectively.
