22CA682R

Question 1(a)(i)

Question:
You are asked to plan a data analytics project to analyze student feedback to DCU in relation to online teaching in 2020 and 2021. Using the Generic Data Analytics Pipeline discussed in CA682, assign each of the following activities to one of the 5 main categories: Gathering, Processing, Analyzing, Presenting, and Preserving, and identify a tool or application that you might use (same one can be used for multiple tasks).

1. Documenting the data formats used in the study and saving all of the created datasets.
2. Removing incorrect entries from the student datasets.
3. Liaising with DCU Registry to get datasets from the student registration and results systems.
4. Calculating the average satisfaction levels based on the sentiment ratings.
5. Anonymizing student comments that include identifying details.
6. Converting student words into sentiment ratings and correlating with the field of study.
7. Conducting student surveys to answer key questions about their experience.
8. Creating a document to share with senior university management summarizing the findings.

Answer:

1. **Documenting the data formats used in the study and saving all of the created datasets.**
    - **Category:** Preserving
    - **Tool:** Google Sheets, Excel, or a database like MySQL
2. **Removing incorrect entries from the student datasets.**
    - **Category:** Processing
    - **Tool:** Python (using Pandas library), OpenRefine
3. **Liaising with DCU Registry to get datasets from the student registration and results systems.**
    - **Category:** Gathering
    - **Tool:** API calls, email communication, or database queries
4. **Calculating the average satisfaction levels based on the sentiment ratings.**
    - **Category:** Analyzing
    - **Tool:** Python, R, or Excel

5. **Anonymizing student comments that include identifying details.**
    - **Category:** Processing
    - **Tool:** Python (using Natural Language Processing libraries like spaCy)
6. **Converting student words into sentiment ratings and correlating with the field of study.**
    - **Category:** Analyzing
    - **Tool:** Python (TextBlob or VADER), Tableau for correlations
7. **Conducting student surveys to answer key questions about their experience.**
    - **Category:** Gathering
    - **Tool:** Google Forms, Microsoft Forms
8. **Creating a document to share with senior university management summarizing the findings.**
    - **Category:** Presenting
    - **Tool:** PowerPoint, Word, or Tableau

Explanation:
Each step is aligned with the phases of the Generic Data Analytics Pipeline:

- **Gathering:** Deals with obtaining data from sources or conducting surveys.
- **Processing:** Includes cleaning, formatting, and modifying data for usability.
- **Analyzing:** Involves deriving insights or performing computations.
- **Presenting:** Communicates results effectively.
- **Preserving:** Saves the data and documentation for future use.

Question 1(a)(ii)

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

Answer:
One weakness of the Generic Data Analytics Pipeline is that it does not explicitly address **data security and privacy management**, especially when dealing with sensitive or personal data such as student feedback.

Explanation:
In modern data analytics projects, ensuring compliance with data protection laws (e.g., GDPR) and implementing security measures are crucial steps. While anonymization is mentioned in processing, tasks like encryption, secure storage, and access control should be explicit in the pipeline. This oversight could lead to non-compliance with legal standards or unauthorized data breaches.

Question 1(b)

Question:
For each of the following data attributes (A-D), choose all the following descriptions that can apply:
Qualitative, Quantitative, Discrete, Continuous, Nominal, Ordinal, Interval, Ratio

A. Rating of temperature comfort in offices (cold, cool, perfect, warm, hot)
B. Number of times a character's name is used in a TV show episode
C. Names of pets owned by all CA682 students
D. All winning times (in seconds) for men's 100m sprint at the Olympic Games

Answer:

- **A.** Qualitative, Ordinal
- **B.** Quantitative, Discrete
- **C.** Qualitative, Nominal
- **D.** Quantitative, Continuous, Ratio

Explanation:

- **A.** The temperature comfort ratings are ordered categories, making them ordinal and qualitative.
- **B.** The count of name occurrences is a numeric, countable value, making it quantitative and discrete.
- **C.** Names of pets are unique identifiers with no inherent order, making them qualitative and nominal.
- **D.** Winning times are numeric, continuous data that starts from a meaningful zero (ratio scale).

Question 1(c)

Question:
Choose one (1) of the following scenarios and explain (in detail) why it is or is not a good example of "big" data according to the three classical characteristics. State any assumptions about the data and its characteristics:

- **A.** Customer account, purchasing data, and engagement data from a supermarket chain's loyalty card programme
- **B.** An individual's step count data for a 1-year period from a personal smart device (e.g., a Fitbit)
- **C.** All 8 episodes (video files) of the TV show "Stranger Things"

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

Explanation:
This is a good example of big data because:

1. **Volume:** The dataset is likely vast, as it aggregates data from numerous customers over time.
2. **Velocity:** Data is generated frequently and must be processed in near real-time, especially for tracking engagement or personalized offers.
3. **Variety:** Includes structured (purchase logs), semi-structured (account details), and unstructured data (social media engagement).

The large-scale and diverse nature of the data align well with big data's classical characteristics. Assumptions include a large customer base and frequent transactions.

Question 2(a)

Question:
Given the following brief to design a system for a data collection and storage (preservation) task:

"Your client runs a chain of 10 gift shops across the UK and Ireland and wants to integrate the inventory and sales data from all stores into a central system. This includes data such as product ID, description, unit price, etc., and daily sales transactions from each shop."

1. List three (3) important questions you would ask your client about their data storage requirements.
2. Suggest a type of data storage approach to use for this project, giving a reason for your choice.
3. The client now wants to include website logs and social media content interactions to work on future promotions. Would this change your recommendation? Why/Why not?

Answer:

1. **Three questions to ask about data storage requirements:**
    - What is the expected volume of data generated daily across all stores?
    - Will the client need real-time analytics or batch processing?
    - Are there any specific compliance requirements (e.g., GDPR) for data storage and handling?
2. **Type of data storage approach:**
    - **Recommendation:** Use a relational database system (e.g., MySQL or PostgreSQL).
    - **Reason:** Relational databases efficiently handle structured data like inventory and sales information. They allow for organized storage, quick querying, and integration with reporting tools.
3. **Does including website logs and social media interactions change the recommendation?**
    - Yes, the recommendation would change.

- **Reason:** Social media and website logs involve unstructured or semi-structured data. A hybrid storage solution like a data lake (e.g., AWS S3 with Redshift) or NoSQL database (e.g., MongoDB) would be better suited. These systems can handle structured, semi-structured, and unstructured data, allowing flexibility for future promotions.

Explanation:
The initial relational database is sufficient for structured data (inventory and transactions). However, including unstructured data like logs and social media requires more scalable and versatile storage, highlighting the need for hybrid or NoSQL systems.

Question 2(b)

Question:

1. Give three (3) examples of simple metadata describing your favorite item of clothing.
2. For each metadata element, identify if it is Descriptive, Administrative, or Structural and briefly explain why.
3. If I were to collect and integrate data about the favorite item of clothing of all CA682 students, how would using a standard specifically change the quality of metadata? Identify one potential difficulty with enforcing a metadata standard.

Answer:

1. **Examples of metadata:**
    - Brand Name: Adidas
    - Material: 100% Cotton
    - Color: Black
2. **Classification and explanation:**
    - **Brand Name:** Descriptive; provides identifying characteristics of the clothing item.
    - **Material:** Descriptive; details the physical properties or components of the item.
    - **Color:** Descriptive; offers visual or categorical information.
3. **Effect of metadata standards and difficulty in enforcement:**
    - **Impact:** Standards ensure consistency, improving data quality, interoperability, and reusability. For example, standardized terms like “Polyester” instead of synonyms or vague terms enhance searchability.
    - **Difficulty:** Users may struggle to adapt to strict standards or predefined categories, especially for unique or ambiguous attributes. This can lead to incomplete or inaccurate metadata entry.

Explanation:
Metadata standards streamline data integration and usability but require careful training and oversight to ensure consistent application across diverse users.

Question 2(c)

Question:
Given the information in your brief in Q2(a), including the social media data, identify any possible data that may need to be handled differently due to European GDPR requirements. Explain why or why not.

Answer:
Social media interactions and website logs often include personal data like user names, IP addresses, or behavioural data. Under GDPR:

- This data must be anonymized or pseudonymized before processing.
- Users must be informed, and explicit consent is required for collecting personal data.

Explanation:
GDPR emphasizes protecting personal data and ensuring transparency in how it is used. For instance, sales data without identifiable links to customers is less sensitive, but website logs may include identifiable IPs, requiring special handling to maintain compliance.

Question 3(a)

Question:
Identify four (4) different possible errors or artifacts in the dataset linked above, giving the column name and cell reference if appropriate. Give the tool or tools you used. You may use any tool that you like.

Answer:

1. **Error 1:** Missing data in the "Country" column (e.g., cell A15).
    - **Tool used:** Python (Pandas library) for data analysis.
    - **Explanation:** Missing country information can lead to incomplete analysis and skewed results.
2. **Error 2:** Outliers in the "Production (Tonnes)" column (e.g., cell D30 showing 1,000,000 tonnes).
    - **Tool used:** Python (with Pandas) and data visualization tools (e.g., Matplotlib, Seaborn).
    - **Explanation:** Extreme values can distort average calculations and statistical analysis.
3. **Error 3:** Inconsistent date formats in the "Year" column (e.g., cell B25 showing "2021" while others show "2021-2022").
    - **Tool used:** Excel and Python (Pandas).

- **Explanation:** Inconsistent formatting can lead to issues in data aggregation and sorting.

4. **Error 4:** Duplicate entries in the dataset (e.g., multiple entries for "Spain" with identical "Production (Tonnes)" and "Year" columns).
    - **Tool used:** Python (Pandas) for data deduplication.
    - **Explanation:** Duplicate records can affect analysis accuracy and distort summaries.

Question 3(b)

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

Answer:

1. **Missing data in the "Country" column:**
    - **Phase:** Gathering
    - **Assumption:** This error may have occurred during data collection when some entries were not properly filled in or were omitted due to reporting errors.
2. **Outliers in the "Production (Tonnes)" column:**
    - **Phase:** Processing
    - **Assumption:** Outliers might have been introduced during data entry or aggregation from different sources, with anomalies not properly flagged or filtered.
3. **Inconsistent date formats in the "Year" column:**
    - **Phase:** Processing
    - **Assumption:** This inconsistency could have been introduced during data formatting or conversion from multiple systems with different date representations.
4. **Duplicate entries in the dataset:**
    - **Phase:** Gathering or Processing
    - **Assumption:** Duplicate entries may have been introduced during data import from multiple sources or improper deduplication processes.

Question 3(c)

Question:
What data quality methods would you suggest using to either avoid or mitigate the errors or artifacts in this dataset? Why would your suggestion improve overall data quality?

Answer:

1. **Data Validation Checks:**
    - **Method:** Implement automated validation checks during data entry to catch missing or inconsistent entries.
    - **Benefit:** Reduces errors from the start and ensures data completeness.
2. **Outlier Detection Techniques:**
    - **Method:** Use statistical methods (e.g., z-score or IQR) to detect and review outliers.
    - **Benefit:** Helps identify data points that do not conform to expected patterns and ensures they are evaluated before analysis.
3. **Standardized Formatting:**
    - **Method:** Apply standardized date formats and use tools to reformat data consistently.
    - **Benefit:** Improves uniformity across the dataset, simplifying analysis and preventing data aggregation issues.
4. **Deduplication Procedures:**
    - **Method:** Use software tools (e.g., Python or Excel) to identify and remove duplicate entries.
    - **Benefit:** Enhances data accuracy and prevents repetitive information from skewing results.

Explanation:
Adopting these quality methods can significantly improve data accuracy, consistency, and reliability, ensuring the data is robust for analysis and decision-making.

Question 4(a)

Question:
Is this an exploratory or explanatory visualization task?

Answer: This is an explanatory visualization task.

Explanation:
The goal is to present specific insights about the geographical distribution of students living within 10 km of the campus. It aims to communicate findings to an audience (senior university management), making it an explanatory task designed to inform and persuade.

Question 4(b)

Question:
Who is the intended audience for the data visualization?

Answer:
Answer: The intended audience is senior university management.

Explanation:
The visualization is meant to help decision-makers understand student accommodation trends and plan accordingly, highlighting relevant statistics and patterns.

Question 4(c)

Question:
What title might you give to the data visualization and why? Make assumptions about any conclusion.

Answer:
Title: "Student Accommodation Proximity to Glasnevin Campus"

Explanation:
This title effectively describes the content, focusing on the main insight that 60% of students live within 10 km of the campus. The assumption is that this proximity is significant for university planning and policy-making.

Question 4(d)

Question:
What specific chart type would you use? Justify your choice referring to the principles discussed in class relating to data types and the message.

Answer:
Answer: A choropleth map or a bar chart.

Explanation:
A choropleth map would visually represent the geographic distribution of student accommodations, making it easy to see concentrations of students within specified distance ranges. A bar chart can also be effective for comparing the number of students living within each range, which suits categorical data well.

Question 4(e)

Question:
For your data visualization, what marks and attributes will you use to encode the data? Be specific about the values of the attributes.

Answer:
Answer:

- **Marks:** Bars (in a bar chart) or shaded regions (in a choropleth map).
- **Attributes:**
    - **Color:** Use shades of blue, with darker colors representing higher student concentrations.

- **Labeling:** Display the number of students and percentage values for each distance range.
- **Axes/Legend:** Include a legend to indicate the range of distances and a labeled axis for the number of students.

Explanation:
These attributes help effectively encode the data, providing clarity on the distribution of student accommodations in proximity to the campus.

Question 4(f)

Question:
Considering the purpose and intended audience, comment on how you would use color or layout principles for this data visualization.

Answer:

- **Color Principle:** Use contrasting shades of blue to differentiate between distance ranges and highlight the proportion of students living close to the campus. Ensure that colors are colorblind-friendly for accessibility.
- **Layout Principle:** Present the visualization with a clear title, labeled axes, and an easy-to-understand legend. Group related elements to guide the viewer's eye through the data naturally.

Explanation:
Using clear and contrasting colors helps highlight the most important information, ensuring the data is easily interpretable by the target audience, while an intuitive layout enhances comprehension.

Question 5(a)

Question:
Identify three (3) possible improvements that you could make to the graph below. Justify your choices, referencing design rules and theories.

Answer:

1. **Add Gridlines:**
    - **Justification:** Helps viewers more accurately compare values across different data points. This adheres to the principle of precision in data visualization.
2. **Use Consistent Colors:**

- **Justification:** Using a consistent color scheme can reduce cognitive load and make it easier for the audience to follow the data, enhancing the readability of the graph.

3. **Add Data Labels:**
    - **Justification:** Adding data labels directly on bars or points helps viewers quickly understand exact values, aligning with the principle of clarity.

Explanation:
These changes make the graph more readable, intuitive, and aligned with best practices in data visualization design.

Question 5(b)

Question:
Given the following visualization tasks, suggest an appropriate graph type (specific chart type and the CHRTS category) for each to display the information and give a brief justification.

A. Compare the performance of stocks in Microsoft, Apple, and Samsung over the last 5 years.
B. Explore movie commercial performance for the IMDB top 50 by director based on cost to make and ticket sales.

Answer:
A.

- **Chart Type:** Line chart (Time Series)
- **Justification:** A line chart is ideal for showing trends over time, making it easy to compare stock performances across multiple companies.

B.

- **Chart Type:** Scatter plot (Comparison)
- **Justification:** A scatter plot allows for comparing the relationship between two quantitative variables (e.g., cost and ticket sales) for multiple items, with directors as categories.

Question 5(c)

Question:
Answer the following questions relating to the graphic shown below:

(i) What is the main communication purpose and why?
(ii) What design choices or guidelines have been used to support this purpose?

Answer:
(i) The main communication purpose is to clearly convey trends or patterns in the data, making it easier for the audience to interpret the underlying information.

(ii) The design choices could include the use of color contrast to differentiate data series, appropriate use of labels to ensure clarity, and an effective layout that guides the viewer through the data logically.

Explanation:
These design decisions support understanding by making the data visually distinct and accessible, aiding in the quick extraction of meaningful insights.
