2022 Repeat

Q1)a) Here's the categorization of the activities into the five main stages of the **Generic Data Analytics Pipeline**, along with suggested tools or applications for each activity:

---

1. Gathering

- **Activity:**
  Liaising with DCU Registry to get datasets from the student registration and results systems.
  Conducting student surveys to answer the key questions about their experience.
- **Explanation:** This stage involves collecting raw data from primary or secondary sources.
- **Tools/Applications:**
  - Microsoft Forms / Google Forms (for surveys)
  - Email communication tools or APIs (to liaise with DCU Registry)

---

2. Processing

- **Activity:**
  Removing incorrect entries from the student datasets.
  Anonymising student comments that include identifying details.

- **Explanation:** This stage involves cleaning, transforming, and preparing the data for analysis.
- **Tools/Applications:**
  - Python (with libraries such as Pandas, NumPy, and NLTK for sentiment analysis)
  - OpenRefine (for data cleaning)

---

3. Analysing

- **Activity:**
  Calculating the average satisfaction levels based on the sentiment ratings.

  Converting student words into sentiment ratings and correlating with the field of study.

- **Explanation:** This stage involves interpreting the processed data and applying statistical or computational techniques to extract meaningful insights.

- **Tools/Applications:**
    - Python (with libraries like SciPy, Matplotlib, and Seaborn)
    - R (for statistical analysis)

---

4. Presenting

- **Activity:**
  Creating a document to share with senior university management summarising the findings.
- **Explanation:** This stage focuses on visualizing and communicating the results effectively.
- **Tools/Applications:**
    - Microsoft PowerPoint / Google Slides
    - Tableau / Power BI (for visualizations)
    - Microsoft Word (for documentation)

---

5. Preserving

- **Activity:**
  Documenting the data formats used in the study and saving all of the created datasets.
- **Explanation:** This stage ensures the data and findings are stored securely for future reference.
- **Tools/Applications:**
    - GitHub / GitLab (for version control)
    - Microsoft OneDrive / Google Drive (for secure storage)
    - SQL databases (for dataset storage)

Would you like a detailed explanation of any specific activity or tool?

II) A **weakness** of the **Generic Data Analytics Pipeline** is its lack of emphasis on **data governance and ethical considerations** throughout the process. While the pipeline outlines stages like gathering, processing, and preserving data, it does not explicitly account for critical aspects such as:

Weakness: Absence of Ethical and Legal Oversight

- **Explanation:** The pipeline does not address:

- Ensuring compliance with **data privacy regulations** (e.g., GDPR for handling student data).
- Assessing the **ethical implications** of data collection and analysis methods, particularly in sensitive domains like student feedback.
- Implementing ongoing **bias checks** in the data and analysis models to ensure fairness and inclusivity.

Example Task Missing:

- **Task:** Establishing a framework for ethical data handling, such as obtaining informed consent from students participating in surveys or ensuring data anonymization and de-identification before analysis.
- **Impact:** Omitting this task can lead to legal liabilities, reputational risks, and biased results, which could compromise the credibility of the findings.

Suggested Improvement:

Incorporate an additional phase or integrate checkpoints for **Ethical Oversight and Governance** at each stage of the pipeline to ensure:

- Data is collected and used responsibly.
- Analysis methods respect participants' privacy and rights.
- Outcomes are unbiased and equitable.

Would you like further elaboration or examples?

—----------

1)b) Here's the classification of each data attribute with the applicable descriptions:

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

Descriptions:

- **Qualitative:** The values are descriptive categories.
- **Ordinal:** There is an inherent order (cold < cool < perfect < warm < hot), but the differences between categories are not measurable.

Exclusions:

- Not **Quantitative**, **Discrete**, **Continuous**, **Nominal**, **Interval**, or **Ratio**, as it does not involve numerical values or meaningful measurements of intervals.

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

Descriptions:

- **Quantitative:** The values are numerical.
- **Discrete:** The count of name occurrences is whole numbers (e.g., 1, 2, 3).
- **Ratio:** A value of 0 means the name is not mentioned, and ratios are meaningful (e.g., a name mentioned 4 times is twice as frequent as 2).

Exclusions:

- Not **Qualitative**, **Continuous**, **Nominal**, **Ordinal**, or **Interval** because the attribute involves exact counts and not categories or measurements with arbitrary scales.

C. Names of pets owned by all CA682 students

Descriptions:

- **Qualitative:** The values are descriptive and non-numeric.
- **Nominal:** The names of pets are categories without any inherent order.

Exclusions:

- Not **Quantitative**, **Discrete**, **Continuous**, **Ordinal**, **Interval**, or **Ratio**, as this is purely categorical data without numerical or ordered properties.

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

Descriptions:

- **Quantitative:** The values are numerical measurements.
- **Continuous:** Time is measured and can take any value within a range (e.g., 9.58 seconds).

- **Ratio:** A value of 0 seconds has a meaningful interpretation (though not practical in this context), and ratios are valid (e.g., 20 seconds is twice as long as 10 seconds).

Exclusions:

- Not **Qualitative**, **Discrete**, **Nominal**, **Ordinal**, or **Interval** because the data represents continuous numerical values measured from an absolute zero.

Let me know if you'd like clarification on any of these classifications!

1)c) **Scenario B: An individual's step count data for a 1-year period from a personal smart device (e.g., a Fitbit)**

This scenario is **not a good example of "big data"** based on the three classical characteristics of big data: **Volume, Velocity, and Variety.**

---

1. Volume

- **Definition:** Refers to the size of the dataset. Big data typically involves massive amounts of data, often measured in terabytes or petabytes.
- **Application to Scenario:**
    - Step count data from a single individual over a year is relatively small. Even if recorded daily, the dataset would likely consist of a few thousand records (e.g., 365 days × steps per day).
    - Assumption: Each record is minimal in size (e.g., a timestamp and a step count).

Conclusion: The volume is insufficient to qualify as big data.

---

2. Velocity

- **Definition:** Refers to the speed at which data is generated and processed. Big data often involves real-time or high-frequency data streams.
- **Application to Scenario:**
    - Step count data is typically collected at regular intervals (e.g., hourly or daily) and processed locally or in periodic uploads to a cloud platform.
    - Assumption: The data generation rate is not continuous or high-frequency like stock market or sensor data.

Conclusion: The velocity is relatively low compared to big data standards.

---

3. Variety

- **Definition:** Refers to the different types and formats of data. Big data often includes structured, semi-structured, and unstructured data from various sources.
- **Application to Scenario:**
    - Step count data is highly structured, typically represented in a uniform format (e.g., numerical values paired with timestamps).
    - Assumption: The data does not involve diverse formats like images, videos, or text.

Conclusion: The variety is minimal, lacking the diversity characteristic of big data.

Overall Assessment

While step count data can be insightful for personal fitness analysis or aggregated across many individuals for broader studies, it **does not meet the criteria for big data** due to its small volume, low velocity, and limited variety. It is better classified as **small, structured data.**

Would you like further examples or a comparison with the other scenarios?

Q2(a) **(i) Three Important Questions for Data Storage Requirements**

1. **What is the expected volume of data generated daily, and do you anticipate significant growth in the future?**
    - This helps assess the scalability requirements of the storage solution.
2. **How frequently will the data need to be accessed or updated?**
    - Determines whether the system prioritizes high-speed access (e.g., for real-time analytics) or cost-efficient archival storage.
3. **What level of data security and compliance is required?**

- Ensures the solution meets legal requirements (e.g., GDPR) and provides appropriate safeguards against unauthorized access.

(ii) Suggested Data Storage Approach

Approach: Cloud-based Relational Database Management System (RDBMS)

- **Reason for Choice:**
  - **Structured Data Handling:** Inventory and sales data are highly structured, making an RDBMS like **Amazon RDS** or **Google Cloud SQL** ideal for managing product IDs, descriptions, unit prices, and transactions.
  - **Centralization:** A cloud-based solution allows seamless integration of data from all stores into a central database accessible from any location.
  - **Scalability:** Cloud platforms provide scalability to accommodate additional stores or increased transaction volumes.
  - **Cost-Effectiveness:** Avoids the need for on-premise infrastructure and provides flexibility with pay-as-you-go pricing.

(iii) Change in Recommendation with Website Logs and Social Media Data

Yes, the recommendation would change.

- **Reason:**
  - **New Data Characteristics:** Website logs and social media interactions introduce **unstructured** and **semi-structured data**, requiring a different storage solution.
  - **Volume and Velocity:** These data sources generate large volumes of data at high velocity, which relational databases may struggle to handle efficiently.

Revised Recommendation:

- **Use a Hybrid Approach:**
  - **Relational Database (RDBMS):** Continue using this for structured inventory and sales data.
  - **NoSQL Database (e.g., MongoDB, Cassandra):** For unstructured and semi-structured data like social media interactions and logs, which are better suited to NoSQL's flexible schema.
  - **Data Lake (e.g., AWS S3, Azure Data Lake):** Store raw logs and social media content for future analysis, supporting scalable, cost-effective storage for large datasets.

This hybrid approach ensures efficient storage, retrieval, and analysis of both structured and unstructured data.

Let me know if you need further clarification or additional recommendations!

2)b)**(i) Three Examples of Simple Metadata Describing a Favorite Item of Clothing**

1. **Color:** Red
2. **Material:** Cotton
3. **Size:** Medium

(ii) Metadata Types and Explanations

1. **Color: Descriptive**
    - **Explanation:** Describes the physical appearance or characteristics of the item, helping users identify or classify the item.
2. **Material: Descriptive**
    - **Explanation:** Provides information about the composition of the item, making it a characteristic that directly describes the clothing.
3. **Size: Administrative**
    - **Explanation:** Size is used for management or operational purposes (e.g., sorting by size for inventory management) rather than describing the item's content directly.

(iii) Impact of Using a Metadata Standard

Improved Metadata Quality:

1. **Consistency:**
    - A standard ensures uniform terminology and formatting (e.g., "Red" is consistently used instead of "Reddish" or "Dark Red"), reducing ambiguity and errors.
2. **Interoperability:**

    - Metadata from different sources can be seamlessly integrated since all data uses the same format and terminology.
3. **Searchability and Discoverability:**

    - Standardized metadata makes it easier to query, search, and retrieve information about clothing items across datasets.

Potential Difficulty with Enforcing a Metadata Standard:

- **Challenge:**
  Resistance to Adoption:
    - Students might provide metadata inconsistently, such as using subjective terms ("Comfy Red T-Shirt") or omitting key details. Enforcing adherence to a standard requires training or oversight, which can be time-consuming or impractical at scale.

Would you like elaboration on metadata standards or their implementation?

2)c) **Data That May Need to Be Handled Differently Under GDPR Requirements**

1. Customer Data (Personal Identifiable Information - PII):

- **Examples:**
    - Customer names, email addresses, phone numbers, loyalty card IDs.
- **Why Handle Differently?**
    - GDPR classifies any information that can identify an individual as **personal data** and mandates strict handling procedures.
    - Businesses must obtain **explicit consent** to collect, store, or use such data and must ensure it is processed lawfully, securely, and transparently.

---

2. Sales Data with Customer Associations:

- **Examples:**
    - Purchase history linked to customer accounts.
- **Why Handle Differently?**
    - If sales data is tied to specific individuals (e.g., tracking purchases through a loyalty card), it becomes PII and falls under GDPR. Businesses need to justify how this data is used (e.g., for personalization) and ensure it is not retained longer than necessary.

---

3. Social Media Interactions:

- **Examples:**
  - User comments, likes, shares, and profiles if linked to identifiable individuals.
- **Why Handle Differently?**
  - Social media content may include PII (e.g., usernames, profile pictures, or comments revealing personal details). Collecting such data requires clear **consent** from users and transparency on how it will be used.

---

4. Website Logs:

- **Examples:**
  - IP addresses, browser types, session details, cookies.
- **Why Handle Differently?**
  - GDPR considers IP addresses and cookies as personal data when they can indirectly identify an individual. Businesses must provide clear opt-in options for tracking and data collection through cookies.

Explanation of GDPR Implications

Under GDPR, organizations must handle personal data differently to ensure:

1. **Lawful Basis for Processing:** Data collection and processing must have a legal basis, such as consent or contractual necessity.
2. **Data Minimization:** Only the data necessary for the intended purpose should be collected.
3. **Right to Access and Erasure:** Customers must have the ability to access, correct, or delete their data upon request.

Why Handle Non-Personal Data Differently?

- Non-personal data, such as product IDs, descriptions, or aggregated sales figures, does not fall under GDPR. However, care must be taken to ensure no indirect re-identification of individuals occurs when combining datasets.

Conclusion

Customer PII, social media interactions, and website logs are the most critical data types under GDPR. Clear consent, secure storage, and transparent use of this data are required, distinguishing it from non-personal or anonymized data. Would you like additional details on GDPR compliance strategies?

Q3)a)b)c) see dataset csv file n do

The answer uploaded here for this question is just a example

Since I can't directly access the dataset (`q3-data.csv`), I'll provide a **template and approach** for analyzing such a dataset, identifying errors or artifacts, and answering the questions. You can apply these steps to your dataset.

Q 3(a) [13 Marks]

Four Possible Errors or Artefacts in the Dataset

1. **Missing Values**
    - **Example:**
        - Column: `Production_Tonnes`
        - Cell Reference: `Row 45, Column "Production_Tonnes"` (e.g., empty or "N/A").
    - **Tool:**
        - Microsoft Excel (Filter for blank cells in the column).
        - Python (using `pandas` to detect `NaN` values).
2. **Incorrect Data Types**
    - **Example:**
        - Column: `Year`
        - Cell Reference: `Row 32, Column "Year"` contains "20O0" (letter "O" instead of zero).
    - **Tool:**
        - Microsoft Excel (Sort and visually inspect the column).
        - Python (using `pandas` to check `dtype` and detect non-numeric values).
3. **Outliers**
    - **Example:**

        - Column: `Production_Tonnes`
        - Cell Reference: `Row 12, Column "Production_Tonnes"` contains `1,000,000` (implausibly high for a fruit crop).
    - **Tool:**
        - Python (using `pandas.describe()` to identify unusually large or small values).
        - Excel (visualize with a box plot).
4. **Duplicate Records**
    - **Example:**
        - Columns: `Country`, `Year`, `Production_Tonnes`
        - Duplicate records observed for `Germany, 2005`.
    - **Tool:**
        - Python (using `pandas.duplicated()`).
        - Excel (sort by columns and visually inspect).

Q 3(b) [6 Marks]

Introduction of Errors/Artefacts

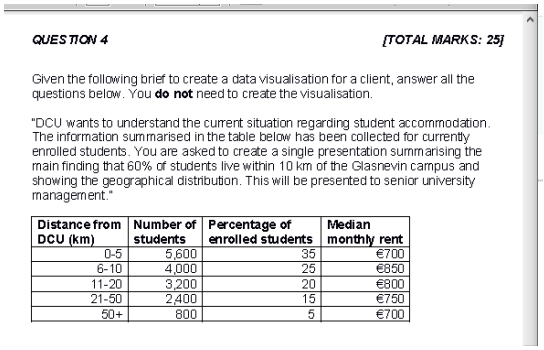
1. **Missing Values**
    - **Phase: Data Gathering**
    - **Likely Cause:**
        - Incomplete or missing reports from certain countries or years.
2. **Incorrect Data Types**
    - **Phase: Data Entry or Processing**
    - **Likely Cause:**
        - Human error during data entry (e.g., mistyping "2000" as "20O0").
3. **Outliers**
    - **Phase: Data Entry or Gathering**
    - **Likely Cause:**
        - Reporting errors or incorrect unit conversion during data collection.
4. **Duplicate Records**
    - **Phase: Data Processing**
    - **Likely Cause:**
        - Erroneous merging or importing of data from multiple sources without de-duplication.

Q 3(c) [6 Marks]

Suggested Data Quality Methods

1. **Handling Missing Values**
    - **Method:** Impute missing values (e.g., use the mean/median or extrapolate based on trends) or flag rows with missing data for further investigation.
    - **Why:** Prevents incomplete analysis and maintains dataset usability.
2. **Ensuring Correct Data Types**
    - **Method:** Use automated scripts (e.g., Python) to validate data types during preprocessing.
    - **Why:** Identifies and corrects type-related errors to ensure consistency in analysis.
3. **Detecting and Handling Outliers**
    - **Method:** Apply statistical methods (e.g., z-scores or interquartile ranges) to identify and cap or remove outliers.
    - **Why:** Improves the reliability of insights by reducing the impact of extreme values.
4. **Removing Duplicates**
    - **Method:** Use tools like `pandas.drop_duplicates()` or Excel's "Remove Duplicates" function to eliminate duplicate rows.
    - **Why:** Ensures data integrity and avoids inflated results during aggregation.

Would you like detailed Python or Excel code examples for any of these methods?



QUESTION 4 ***[TOTAL MARKS: 25]***

Given the following brief to create a data visualisation for a client, answer all the questions below. You **do not** need to create the visualisation.

"DCU wants to understand the current situation regarding student accommodation. The information summarised in the table below has been collected for currently enrolled students. You are asked to create a single presentation summarising the main finding that 60% of students live within 10 km of the Glasnevin campus and showing the geographical distribution. This will be presented to senior university management."

Distance from DCU (km)	Number of students	Percentage of enrolled students	Median monthly rent
0-5	5,600	35	€700
6-10	4,000	25	€850
11-20	3,200	20	€800
21-50	2,400	15	€750
50+	800	5	€700

Q 4(a)

[2 Marks]

Is this an exploratory or explanatory visualisation task?

- **Answer:** This is an **explanatory visualisation task.**
    - **Reason:** The purpose is to convey specific insights (e.g., 60% of students live within 10 km of the Glasnevin campus) to senior university management in a clear and concise manner, rather than exploring data to uncover patterns or trends.

Q 4(b) [2 Marks]

Who is the intended audience for the data visualisation?

- **Answer:** The intended audience is **senior university management** at DCU.
    - **Reason:** The visualisation is created for decision-makers who need clear and actionable insights about student accommodation.

Q 4(c) [4 Marks]

What title might you give to the data visualisation and why?

- **Title:**
  "Student Accommodation Proximity and Rent: Key Insights for Glasnevin Campus"

- **Why:**

    - The title emphasizes the focus on the geographical proximity of students to the Glasnevin campus (a key finding that 60% live within 10 km).

    - It also highlights the relevance of accommodation costs (median rent), which is likely important for management decisions.

Would you like suggestions for the chart types or more refinement on the narrative for senior management?

Q 4(d) [6 Marks]

What specific chart type would you use?

- **Chart Type: Stacked Bar Chart with a Map Overlay**

- **Justification:**

    - **Stacked Bar Chart:**
        - Effectively shows proportions (e.g., 60% within 10 km) by stacking student numbers for different distance ranges.
        - Provides a clear comparison of student distributions across distance bands.
    - **Map Overlay:**
        - Geographical distribution is best visualized using a map to highlight where students live relative to the Glasnevin campus.
        - Adds spatial context, making the data more actionable for senior management.
- **Principles Applied:**

    - **Data Types:** Quantitative (number of students, median rent) and Categorical (distance bands).
    - **Message Clarity:** Combines proportional and geographical views, aligning with the explanatory purpose.

Q 4(e) [6 Marks]

What marks and attributes will you use to encode the data?

1. **Stacked Bar Chart:**

    - **Marks:** Bars.
    - **Attributes:**
        - **Height:** Number of students (quantitative).

        - **Color:** Distance bands (categorical; e.g., different shades for each range).
        - **Labels:** Percentage of enrolled students for each band (e.g., "35%").
2. **Map Overlay:**

    - **Marks:** Circles or heatmap regions.
    - **Attributes:**
        - **Size:** Number of students in each area (quantitative).
        - **Color Intensity:** Median rent levels (quantitative).

Q 4(f) [5 Marks]

How would you use color or layout principles?

1. **Color Usage:**

    - **Distinct Colors for Distance Bands:** Use a sequential color scheme (e.g., light to dark blue) to represent distance ranges, ensuring intuitive progression.
    - **Median Rent Overlay:** Use a gradient (e.g., yellow for low rent to red for high rent) to encode rent data on the map.
    - **Accessibility:** Ensure colors are colorblind-friendly (e.g., avoid green-red pairings).
2. **Layout Principles:**

    - **Simplify for Clarity:** Position the stacked bar chart alongside the map to allow easy comparison between proportions and geographical distribution.
    - **Emphasize Key Insights:** Highlight the "60% within 10 km" statistic with bold text or annotations directly on the visual.
    - **White Space:** Use adequate spacing to prevent visual clutter and enhance readability.

Would you like further assistance on creating the visualization or refining the narrative for senior management?

Q 5(a) [9 Marks]

Improvements for the graph:

1. **Use of Color to Distinguish Data Elements:**

    - **Improvement**: Introduce distinct colors for each data series or category.
    - **Justification**: According to **Color Theory** and **Gestalt Principles**, using contrasting or distinct colors helps in visually separating different elements of the graph, improving clarity and comprehension. Without proper color coding, it can be difficult for the viewer to differentiate between multiple data points, especially when there's a lot of information.
2. **Axis Labels and Title:**

    - **Improvement**: Add clear and informative axis labels and a graph title.
    - **Justification**: **Data-Ink Ratio** theory emphasizes the importance of maximizing useful information while minimizing unnecessary details. The inclusion of proper axis labels and a clear title adheres to this principle by helping the viewer immediately understand what data is being presented. Additionally, clear labeling aligns with the **Chartjunk Principle** to eliminate clutter and ensure the graph is self-explanatory.
3. **Gridlines and Data Points:**

    - **Improvement**: Improve the visibility of gridlines and highlight data points.
    - **Justification**: The **Visual Hierarchy** principle suggests that key elements (like data points) should be emphasized so the viewer can easily identify trends and values. Faint or missing gridlines can make it hard to read the values on the graph, while highlighting key data points makes the visualization more intuitive. **Gridlines** should be subtle but present to guide the reader without overwhelming the graph.

Q 5(b) [8 Marks]

Graph Suggestions for the Given Tasks:

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

- **Suggested Graph Type**: **Line Graph**
- **CHRTS Category**: **Trend Comparison**

- **Justification**: A **Line Graph** is ideal for showing trends over time, such as stock performance across multiple years. It allows clear comparison of trends for Microsoft, Apple, and Samsung by displaying their performance on the same graph. This helps in identifying patterns, growth, or decline, and a **CHRTS Trend Comparison** would categorize this as a suitable method for comparing changes over time.

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

- **Suggested Graph Type**: **Scatter Plot**
- **CHRTS Category**: **Relationship Exploration**
- **Justification**: A **Scatter Plot** is perfect for showing the relationship between two continuous variables, like cost to make the movie and ticket sales. It can visually illustrate whether there's a correlation between the two variables and how they vary across directors. This task falls under the **CHRTS Relationship Exploration** category, as the goal is to explore how one variable (cost to make) influences another (ticket sales).

To answer **Q 5(c)**, I'll first provide general guidance on how to approach these types of questions. Since the graphic isn't provided, I will make assumptions based on typical visualization principles.

(i) What is the main communication purpose and why?

The main communication purpose of any graphic typically revolves around presenting data in a way that makes it easily understandable, highlighting the key message or insight.

- **Purpose Example 1**: If the graphic shows the sales performance over time, the purpose could be to **highlight trends and patterns** in the sales data (e.g., identifying peak periods, growth, or declines).
- **Purpose Example 2**: If the graphic compares various entities, such as companies or products, the goal might be to **compare and contrast data** (e.g., showing how different companies are performing across a range of metrics like revenue or market share).

The **why** behind the purpose is to make the data more digestible for the target audience, whether for decision-making, trend analysis, or performance evaluation.

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

- **Color**: Colors can be used to distinguish between different categories or trends, making it easier to follow. For example, a different color for each company or product helps in clear comparison. Color theory principles help improve readability and highlight important data.

- **Title & Labels**: A clear title and well-labeled axes or categories ensure that the viewer understands the context of the graphic. This supports the goal of clear communication by providing context to the data.

- **Simplification**: Adhering to the **Data-Ink Ratio** principle, unnecessary elements (e.g., excessive gridlines or decoration) are minimized to avoid clutter and focus attention on the key data.

- **Data Markers or Annotations**: Adding markers or annotations at key points can help guide the viewer's attention to significant data points, trends, or comparisons. This supports the goal of **emphasizing critical information**.

- **Chart Type Choice**: Depending on the task, the choice of chart (e.g., bar graph, line chart, pie chart) will directly influence how effectively the data is communicated. For example, a **line chart** might be used to show trends, while a **bar chart** could be used for comparisons.

These design choices directly help the graphic achieve its communication goal by ensuring that the data is accessible, easy to interpret, and visually engaging.

Question 5 answer is according to chat gpt which is in very general form .