

**AUGUST/RESIT EXAMINATIONS 2021/2022**

**MODULE:** CA682 - Data Management and Visualisation

**PROGRAMME(S):**

MCM M.Sc. in Computing

CAPT PhD-track

MEPT PhD-track

MEPM MEng

ECSAO Study Abroad (Engineering & Computing) **YEAR OF STUDY:** 1,2,3,4,O,X

|  |  |  |
| --- | --- | --- |
| **EXAMINER(S):** |  |  |
| Suzanne Little | (Internal) | (Ext:6052) |
| Prof. Giorgio Giacinto | (External) | External |
| Dr. Katarina Domijan | (External) | External |
| Prof. Joseph Cannataci | (External) | External |

**TIME ALLOWED:** 3 Hours

**INSTRUCTIONS: Answer Q1 then answer 3 questions from Q2-Q5. All**

**questions carry equal marks.** Type or paste your answers into the

|  |
| --- |
| indicated boxes |

.

Upload a single document (word or pdf) to loop.

**By submitting this exam, you declare (1) that all of the work is your own;**

**(2) that you did not seek whole or partial solutions for any part of your submission from others; and (3) that you did not and will not discuss,**

**exchange, share, or publish complete or partial solutions for this exam or any part of it during the exam.**

Please note that where a candidate answers more than the required number of questions, the examiner will mark all questions attempted and then select the highest scoring ones.

***There are no additional requirements for this paper.***

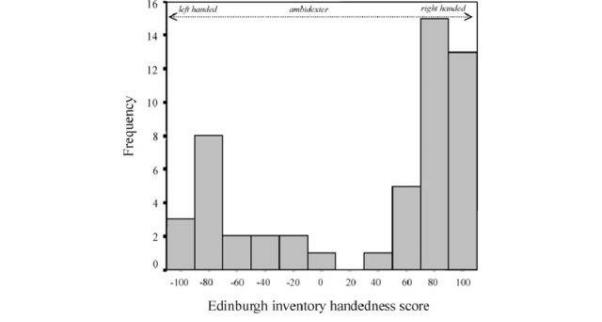
CA682D - Data Management and Visualisation

August/Resit Examinations 2021/2022 Page 1 of 12

***QUESTION 1 [TOTAL MARKS: 25]***

**Q 1(a) [7 Marks]**

(i) Describe the shape of the histogram below in terms of the modality. [2 marks]



|  |
| --- |
| Answer:  **Modality** in a histogram refers to the number of peaks in the distribution.  ●  **Unimodal:** A histogram with one clear peak.  ●  **Bimodal:** A histogram with two clear peaks.  ●  **Multimodal:** A histogram with more than two clear peaks.  To determine the modality, you would examine the histogram and identify the number of distinct peaks.  Describing a Bimodal Histogram  Let's look at an example of how to describe a bimodal histogram. Imagine a histogram representing the height of students in a university. The histogram has two distinct peaks:  ●  **Peak 1:** Centered around 160 cm, representing a large group of female students.  ●  **Peak 2:** Centered around 175 cm, representing a large group of male students.  **Description:**  The histogram exhibits **bimodality**, indicating the presence of two distinct groups within the data. This suggests a potential **categorical variable** influencing the height distribution, likely gender in this case. The first peak, centered around 160 cm, is likely representative of female students, while the second peak, centered around 175 cm, likely represents male students.  **Key Points to Include in a Description:**  ●  **Modality:** State that the histogram is bimodal.  ●  **Peak Locations:** Identify the approximate center points of the peaks.  ●  **Potential Explanations:** Offer possible reasons for the two peaks, referencing potential categorical variables influencing the distribution.  Remember, the sources focus on general data visualization concepts and do not provide specific examples of bimodal histograms. Therefore, the height example above is not derived from the sources and is for illustrative purposes only. |

(ii) If I’m creating a frequency table from a *discrete* variable (e.g., the count of current primary toothbrush colours for CA682 students) what could the

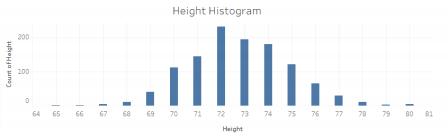
values in the first column be? [2 marks]

|  |  |
| --- | --- |
| **?** | **Frequency** |
| ? | 100 |
| ? | 52 |
| ? | 23 |
| ? | 40 |

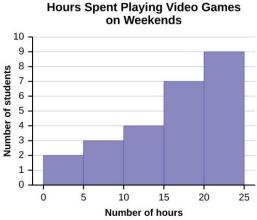
|  |
| --- |
| Answer:The first column of a frequency table typically lists the **distinct values of the variable**. In this case, the variable is "current primary toothbrush colours for CA682 students."  Possible values for the first column could be:  ●  Red  ●  Blue  ●  Green  ●  Yellow  ●  White  ●  Multicolored  ●  Other  These values represent the different categories of toothbrush colours reported by students. You might need to adjust these categories based on the actual responses collected. For instance,3 highlights the importance of anticipating all possible answers in survey design, using the example of a toothbrush colour survey with limited options. |

(iii) Which of the following histograms (A-D) displays *positive skewness*?

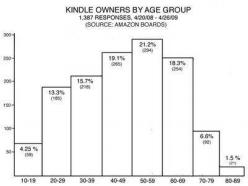
[3 marks] A:



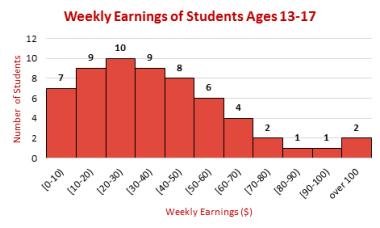
B:



C:



D:

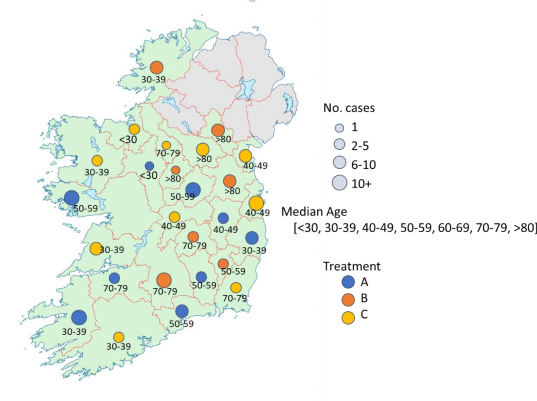


|  |
| --- |
| Answer:Identifying Positive Skewness Of the given histograms, **histogram D** displays positive skewness.  **Positive Skewness** is evident in a histogram when the **tail of the distribution extends further to the right** and the **peak of the distribution lies to the left**. [outside source] In simpler terms, the majority of the data points are clustered towards the lower end of the scale, with a few higher values stretching the tail towards the right.  A **histogram** is a graphical representation of the **frequency distribution of a dataset**. It uses bars to represent the number of data points (frequency) that fall within specific intervals (bins). [outside source]  The other histograms presented do not exhibit positive skewness. Histograms A and C appear to be relatively symmetrical, while histogram B leans towards negative skewness with its tail extending to the left. |

CA682D – Data Management and Visualisation

August/Resit Examinations 2021/2022 Page 3 of 12

**Q 1(b) [10 Marks]**

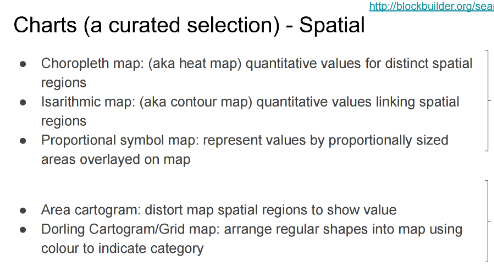


The figure above was created to show the distribution of patients with a particular

long term medical condition that is being treated with three different therapies (A, B &

C) according the county where the therapy is administered. The median age of patients in each county is also included.

(i) What type of visualisation is this (specific name)? What are the marks and attributes used to encode the data? [3 marks]



(i) Visualisation Type, Marks, and Attributes

The visualisation in question is a **proportional symbol map**12. This type of map uses **area marks** overlaid on a geographical map to represent values12. The larger the area of the mark, the higher the value it represents. In this specific case, the area of the circles likely represents the number of patients with the medical condition in each county.

The attributes used to encode the data in the visualisation are:

●

**Position (X & Y):** This attribute is used to encode the geographical location of each county on the map34.

●

**Area:** This attribute is used to encode the number of patients with the condition in each county34.

●

**Colour Hue:** This attribute may be used to encode the type of therapy (A, B or C) being administered in each county456. The question describes the three therapies but the colours are not specified.

●

**Median Age:** The question also notes that the median age of patients in each county is included. This could be encoded by varying the colour saturation or lightness of the area marks5. Again, the exact encoding method is not described in the question.

(ii) Identify any issues with preserving the privacy of patients that this graph may raise? [3 marks]

(ii) Privacy Issues

The graph raises several issues regarding patient privacy:

●

**Small Population Counties:** In counties with small populations, the visualisation could potentially reveal the identities of individuals if only a small number of patients are receiving a particular therapy.

●

**Unique Combinations:** If the number of patients with the condition in a county is small, and there's a unique combination of therapy and median age, it might be possible to deduce the identity of an individual.

●

**External Data:** Even if the graph alone doesn't reveal individual identities, someone with access to external data sources (e.g., local news reports about specific cases) might be able to use the information in the graph to identify patients.

(iii) The graph is released in a report that also includes further aggregated

information such as the date of diagnosis, blood type, gender and ethnicity of the patient by county. Comment on the likely risks associated with this report and suggest two methods that may be used to reduce the privacy

risks. [4 marks]

|  |
| --- |
| Answer:(iii) Risks of Aggregated Information and Mitigation Methods  Releasing the graph alongside aggregated data on diagnosis date, blood type, gender, and ethnicity significantly increases the risk of **patient re-identification**.  **Reasoning:** By combining these variables, an attacker could potentially narrow down the possible individuals in a county who match the characteristics presented in the report. This is particularly true for smaller counties where the number of patients is already low.  **Methods to Reduce Privacy Risks:**  1.  **Data Suppression:** Suppress data for counties with a population below a certain threshold. This prevents the release of highly specific information that could easily lead to individual identification.  2.  **Data Generalisation:** Instead of providing precise median ages, group ages into broader ranges (e.g., 30-39, 40-49). This makes it harder to pinpoint specific individuals based on age.  **Note:** It's important to be aware that even with these methods, there's always a risk of re-identification when dealing with personal health information. It's crucial to consider the potential consequences of data release and implement appropriate safeguards to protect patient privacy.  copy\_all  thumb\_up  thumb\_down  keep\_pinSave to note |

CA682D – Data Management and Visualisation

August/Resit Examinations 2021/2022 Page 4 of 12

**Q 1(c) [8 Marks]**

You are asked to create a visualisation to show 120 first year students so they can understand their comparative performance across all modules (sample data table headings shown below). Suggest a method for communicating the information that

would be privacy preserving and discuss any potential risks that should be considered.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Student ID** | **Student Name** | **CA100** | **CA105** | **CA121** | **CA167** |
| 123456789 | Alex Smith | 64% | 72% | 56% | 85% |

|  |
| --- |
| Answer:Visualizing Student Performance While Preserving Privacy To visualize the comparative performance of 120 first-year students across all modules while preserving their privacy, a **histogram with aggregated data** can be used.  Here's how it could work:   * **Data Aggregation:** Instead of displaying individual student scores, group the marks into ranges (e.g., 60-69%, 70-79%, 80-89%). This creates aggregated data that represents the distribution of student performance without revealing individual scores. * **Separate Histograms per Module:** Create a separate histogram for each module (CA100, CA105, CA121, CA167). The x-axis of each histogram would represent the mark ranges, and the y-axis would represent the number of students falling within each range. * **Anonymous Representation:** Do not include any personally identifiable information like Student ID or Student Name in the visualization. Each student's performance is represented anonymously within the aggregated data.  Potential Risks and Mitigation Strategies Even with aggregated data, there are potential privacy risks to consider:  **Small Groups within Ranges:** If a particular mark range has very few students, it might be possible to infer the identity of a student based on their known performance in other modules.   * + **Mitigation:** Suppress data for ranges with fewer than a predetermined number of students. For example, if a range has fewer than 5 students, do not display a bar for that range.   **External Knowledge:** Someone with external knowledge about specific students, like their strengths in certain subjects, could potentially use the histograms to deduce individual performance.   * + **Mitigation:** Carefully consider the level of detail provided in the histograms and the potential for external knowledge to compromise privacy. It might be necessary to further generalize the data or limit the number of modules visualized together.  Example using the provided data: For module CA100, the histogram could have a bar for the 60-69% range representing one student (Alex Smith). However, if there were fewer than 5 students in this range, the data would be suppressed, protecting privacy. Additional Considerations  * **Contextual Information:** The level of risk associated with the visualization depends on the context. Sharing this data within a secure online learning platform accessed only by authorized individuals might pose a lower risk than displaying it publicly. * **Communication:** Clearly communicate to students how the data is aggregated and anonymized to ensure they understand that individual performance is not being revealed.  Relevant Information from the Sources: The sources provided focus on data visualization techniques and principles but do not directly address the specific scenario of visualizing student performance while preserving privacy. The concepts of data aggregation, anonymity, and the consideration of privacy risks are based on general data privacy principles. Note: The information about data privacy principles and mitigation strategies is based on general knowledge about data privacy and might need further verification from specific data privacy regulations or guidelines relevant to your institution. |

***[End of Question 1]***

CA682D – Data Management and Visualisation

August/Resit Examinations 2021/2022 Page 5 of 12

***QUESTION 2 [TOTAL MARKS: 25]***

**Q 2(a) [10 Marks]**

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 to 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.

When designing a data collection and storage system for the client's gift shop chain, it's important to understand their specific requirements. Here are three crucial questions to ask:

1.

**"How much data do I have now? What rate will I get new data?"**1 This question is vital to determine the scale and growth rate of the data, informing decisions about storage capacity, scalability, and potential future needs.

2.

**"Does the data need to be processed before loading?"**1 Understanding any pre-processing requirements is crucial to design an efficient data pipeline. This includes tasks like data cleaning, transformation, or validation before storage.

3.

**"How many queries will be run? Will they be concurrent? How many users?"**1 This question addresses the anticipated usage patterns of the system. Knowing the volume and concurrency of queries helps determine the database technology and infrastructure required to ensure optimal performance and responsiveness.

1. Suggest a type of data storage approach to use for this project, giving a reason for your choice.

A **relational database** is recommended for this project.

**Reasons:**

●

**Structured Data:** The data described (product ID, description, unit price, sales transactions) is highly structured and fits well into a relational model with tables, rows, and columns.23 Relational databases excel in managing structured data, ensuring data integrity and facilitating efficient querying using SQL.3

●

**Querying and Reporting:** The client's need to integrate inventory and sales data suggests a requirement for querying and reporting on the data. Relational databases offer powerful querying capabilities, allowing for complex analysis and the generation of insightful reports.

(iii) 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?

|  |
| --- |
| 1. (ii) (iii)   **(iii) Including Website Logs and Social Media Data**  Including website logs and social media content interactions would likely necessitate a change in the data storage approach.  **Reasons:**  ●  **Unstructured Data:** Website logs and social media content are largely unstructured data.2 While relational databases are suitable for structured data, they may not be the optimal choice for managing and querying unstructured data like text, images, and video.  ●  **NoSQL Databases:** A **NoSQL database**, which is designed to handle unstructured data, might be a more appropriate choice for storing website logs and social media content.35 Different types of NoSQL databases, such as document, column, or graph databases, could be considered based on the specific needs of the project.3  ●  **Polyglot Persistence:** A combination of a relational database for the structured inventory and sales data and a NoSQL database for the unstructured website and social media data could provide a comprehensive solution. This approach, known as **polyglot persistence**, leverages the strengths of different database technologies for managing diverse data types.56  **Note:** Specific recommendations on NoSQL database types and their integration with the relational database would require further analysis of the nature and volume of the website and social media data. |

**Q 2(b) [9 Marks]**

1. Give three (3) examples of simple metadata describing your favourite item of clothing.

(ii) For each metadata element, identify if it is Descriptive, Administrative or Structural and briefly explain why.

(iii) If I was to collect and integrate data about the favourite item of clothing of all CA682 students then, in your own words, how would using a standard specifically change the quality of metadata data? Identify one potential

difficulty with enforcing a metadata standard.

|  |
| --- |
| Simple Metadata and its impact on Data Quality **(i) Simple Metadata Examples**  Here are three examples of simple metadata describing a favourite item of clothing:   * **Size:** This indicates the garment's physical dimensions, crucial for fit and understanding. For example, "Size: Medium" provides a general understanding of the garment's dimensions. * **Date of Purchase:** Recording when the item was acquired provides historical context and potential insights into purchasing patterns. An example would be "Date of Purchase: 2023-04-15." * **Condition:** Describing the current state of the garment is useful for managing its lifecycle. An example is "Condition: Excellent," suggesting minimal wear and tear.   **(ii) Metadata Element Classification**   * **Size:** **Descriptive** metadata. "Size" is an inherent property of the clothing and helps in identifying and understanding the garment. * **Date of Purchase:** **Administrative** metadata. This element focuses on the management aspect of the item rather than its inherent characteristics. It relates to when the object was acquired, which aligns with the definition of administrative metadata as pertaining to "the object’s creation and management". * **Condition:** **Descriptive** metadata. "Condition" describes the current state of the garment which is an intrinsic property of the clothing item.   **(iii) Impact of Standards on Metadata Quality**  Using a **metadata standard** when collecting data about favourite clothing items of CA682 students would significantly enhance the quality of the metadata in several ways:   * **Consistency:** A standard ensures everyone uses the same terms and formats for recording metadata, like "Size" or "Date of Purchase". This consistency makes the data much easier to aggregate, analyze, and compare across different records. * **Interoperability:** Standards promote interoperability, meaning the data can be easily shared and used across different systems or platforms. This is particularly important when integrating data from various sources. * **Findability:** Consistent terminology and structure based on a standard make it easier to search and retrieve specific information from the dataset.   **Potential Difficulty with Enforcing a Standard**  One challenge with enforcing a metadata standard is the potential for **subjectivity** in interpretation. Even with a defined standard, individuals might apply terms differently based on their understanding or personal judgment. For example, the "Condition" of a garment might be assessed differently by various people. This subjectivity can introduce inconsistencies in the metadata, hindering the effectiveness of the standard.  This response draws on information from the sources and integrates insights from our previous conversation about metadata and clothing. However, the specific example of "Condition" as metadata and the potential difficulty with subjective interpretation are not directly supported by the provided sources. You may want to verify these points independently. |

CA682D – Data Management and Visualisation

August/Resit Examinations 2021/2022 Page 6 of 12

**Q 2(c) [6 Marks]**

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: GDPR Considerations for Gift Shop Data **Based on the provided brief and the information from the sources, several types of data in this gift shop system might need special handling due to European GDPR requirements:**  **Sales Transactions:** Daily sales transactions likely contain information about individual customers, such as:   * + **Customer Names and Addresses:** These are **directly identifying** personal data and fall under GDPR.   + **Payment Card Details:** Credit card numbers, security codes, and transaction details are considered **sensitive personal data**, requiring strict security measures and explicit consent for processing.   + **Purchase History:** This data, while not always directly identifying, can be combined with other information to indirectly identify individuals, especially when it involves unique products or specific purchase patterns.   **Website Logs and Social Media Data:** The brief mentions the client's desire to integrate website logs and social media interactions. This data can contain various personal data points, potentially requiring different handling due to GDPR:   * + **IP Addresses and Online Identifiers:** These can indirectly identify individuals, especially when combined with other data.   + **User Activity and Browsing History:** This information can reveal personal interests, preferences, and potentially sensitive information.   + **Social Media Posts and Interactions:** These could contain personal opinions, beliefs, or even health information, falling under sensitive data.   **GDPR Principles and Considerations:**   * **Lawful Basis for Processing:** The client must have a legitimate reason (lawful basis) for collecting and processing this data, such as fulfilling orders, marketing with consent, or complying with legal obligations. * **Data Minimization:** The client should collect only the necessary data for the specified purpose. For example, if collecting customer addresses is only for shipping, they should not collect additional personal details unless there's a legitimate reason and consent. * **Consent and Transparency:** The client must obtain clear and unambiguous consent from individuals for collecting and processing their data. This includes being transparent about what data is collected, how it's used, and for what purpose. * **Data Security:** The client must implement appropriate technical and organizational security measures to protect personal data from unauthorized access, disclosure, alteration, or destruction. * **Data Retention:** The client should retain personal data only for as long as necessary for the specified purpose. Once the purpose is fulfilled, the data should be securely deleted.   **Specific Actions for GDPR Compliance:**   * **Data Mapping and Risk Assessment:** The client should identify all personal data collected, assess the risks associated with processing this data, and document their data processing activities. * **Privacy Policy and Consent Mechanisms:** The client needs a clear and accessible privacy policy explaining their data collection practices and obtain explicit consent for processing sensitive data. * **Data Security Measures:** The client should implement encryption, access controls, and other security measures to safeguard personal data. * **Data Subject Rights:** The client must respect individuals' rights under GDPR, such as the right to access their data, rectify inaccuracies, and request data deletion.   **Important Note:** The information provided in this response is based on a general understanding of GDPR principles as presented in the sources. It is not legal advice and should not be considered as such. For specific legal guidance, consult with a data protection professional. |

***[End of Question 2]***

CA682D – Data Management and Visualisation

August/Resit Examinations 2021/2022 Page 7 of 12

***QUESTION 3 [TOTAL MARKS: 25]***

Q3 requires the dataset (q3-data.csv) provided in loop – [[…]]

Please download the data linked above and use it to answer the questions below.

The dataset contains information about fruit crop production (in Tonnes per year) by European country for all the years 2000 to 2020 inclusive.

**Q 3(a) [13 Marks]**

Identify **four (4) different** possible errors or artefacts 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: Tool: |

**Q 3(b) [6 Marks]**

Identify how each error or artefact in Q3(a) is **most likely** to have been introduced,

specifying the phase from the generic data analytics pipeline. State any assumptions.

|  |
| --- |
| Answer: Analysing Potential Errors and Artefacts in a Data Analytics Pipeline To answer your query, the content of “Q3(a)” is required. The sources provided do not include this information. To effectively identify how errors or artefacts might have been introduced into the dataset mentioned in "Q3(a)", more information is needed about the errors and the dataset itself.  However, the sources offer a detailed overview of the generic data analytics pipeline and common errors associated with each phase. This information can be used to make assumptions about the potential sources of errors, once the specifics of "Q3(a)" are provided.  **Generic Data Analytics Pipeline and Potential Errors**  The sources outline the following phases in a data analytics pipeline:  **Gathering:** This phase involves collecting data from various sources, which can lead to the following errors:   * + **Manual entry errors:** Typos, inconsistencies in formatting, and missing values due to human error.   + **Poor survey or interface design:** Leading questions, limited options, and inadequate data validation can introduce inaccurate or incomplete data.   + **Lack of standards:** Inconsistent formats, uncontrolled vocabularies, and differing units of measurement can create integration challenges.   + **Data retrieval issues:** Misunderstandings about source data, the need for derived data, or plain mistakes in data extraction can result in errors.   **Processing:** This phase focuses on cleaning, aligning, and integrating data, where errors can occur due to:   * + **Inappropriate pre-processing:** Incorrect aggregation, converting nulls to default values, and data loss during transmission can compromise data quality.   + **Data integration challenges:** Combining data from multiple sources can introduce errors due to heterogeneous data, differing definitions, time synchronization issues, and legacy data formats.   **Analysing:** The analysis phase involves applying statistical methods, machine learning, and exploration techniques. Potential errors in this phase include:   * + **Scale and performance issues:** Large datasets can lead to computational limitations, affecting analysis accuracy.   + **Incorrect statistical assumptions:** Misinterpreting confidence bounds or relying on inappropriate models can lead to misleading results.   + **Insufficient domain expertise:** Lack of understanding of the data context can result in misinterpretations and incorrect conclusions.   + **Casual empiricism:** Using arbitrary numbers to support preconceptions can bias the analysis.   **Presenting:** This phase involves creating visualisations and communicating insights, where errors can occur due to:   * + **Misleading visualisations:** Inappropriate chart types, misleading scales, or poor design choices can distort the representation of data.   + **Lack of context and explanation:** Visualisations without proper context or explanations can lead to misinterpretations.   **Preserving:** This phase focuses on storing, managing, and re-using data. Errors in this phase can include:   * + **Format conversion errors:** Converting between data formats can lead to rounding errors or data loss.   + **Lack of metadata:** Missing or inadequate metadata can make it difficult to understand and interpret data in the future.   + **Technical issues:** Data loss due to transmission errors, disk failures, or data corruption can compromise data integrity.   + **No documentation for missing values:** Failing to document the format or reason for missing values can create confusion during analysis.   **Assumptions and Additional Considerations:**   * **Data Source:** The type and origin of the data in "Q3(a)" are unknown. Different data sources have varying levels of data quality and potential error sources. * **Tools and Methods Used:** The specific tools and methods used in each phase of the data analytics pipeline can influence the types of errors introduced. * **Human Error:** Human error is a significant factor in data quality issues. It can occur at any stage of the pipeline.   **Please provide the specifics of "Q3(a)" to receive a more tailored analysis of the potential error sources.** For example, you could provide:   * **The type of data in "Q3(a)."** * **A description of the identified errors or artefacts.** * **Any context about the data analytics pipeline used to process this data.**   With this information, a more comprehensive and accurate assessment of how each error was likely introduced can be provided. |

**Q 3(c) [6 Marks]**

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

|  |
| --- |
| Answer: Suggestions for Improving Data Quality in the Dataset To answer your question accurately, the specific errors or artefacts present in the dataset are needed. However, based on the previous conversation and the information in the sources, some general data quality methods that could be employed to avoid or mitigate errors are:  **1. Data Gathering Phase:**   * **Implementing Process Mandates:** This involves establishing clear guidelines and standards for data entry, collection, and formatting. For instance, enforcing schema or rule restrictions to ensure data consistency and using controlled vocabularies for specific fields (e.g., using "Dublin City University" consistently instead of variations like "DCU" or "Dublin City Uni").   + **Why it improves data quality:** Process mandates reduce human error by minimizing inconsistencies, typos, and formatting issues, leading to cleaner data from the outset. * **Building in Integrity Checks and Entry Constraints:** This includes using data validation techniques to ensure data accuracy during the data entry process. For example, setting constraints on numerical fields (e.g., age should be a positive integer), using drop-down menus for categorical data, and implementing checks for duplicate entries.   + **Why it improves data quality:** Integrity checks prevent inaccurate or invalid data from entering the system, improving the accuracy and reliability of the dataset. * **Process Management:** This focuses on establishing a culture of data quality by rewarding accurate data entry, encouraging data sharing, and establishing clear data stewardship roles.   + **Why it improves data quality:** This approach promotes awareness and responsibility for data quality throughout the organization, leading to better data management practices.   **2. Data Processing Phase:**   * **Building Reliable Transmission Protocols:** Using reliable methods for data transfer, such as using a relay server, can help prevent data loss during transmission.   + **Why it improves data quality:** Ensures that data is transferred completely and accurately, reducing the risk of missing or corrupted data. * **Verification:** Implementing checksums, verification parsers, and checks to ensure that uploaded data fits expected patterns can help identify and correct errors during data delivery.   + **Why it improves data quality:** These methods provide an extra layer of validation, ensuring that the processed data adheres to the predefined quality standards and expected structures. * **Data Browsing and Exploration:** Thoroughly exploring and reviewing the data before and after integration helps identify hidden problems and ensures the integration process yielded the expected results.   + **Why it improves data quality:** Allows for the detection of inconsistencies, anomalies, and unexpected results arising from the integration of different datasets.   **3. Data Storage Phase:**   * **Documenting and Publishing Data Format Standards:** Clearly defining and documenting the format standards for each data field, including whether it is a string, float, or integer, and how rounding or approximation should be handled, can prevent conversion errors.   + **Why it improves data quality:** This practice ensures consistency in data interpretation and reduces the risk of errors arising from format conversions or misinterpretations. * **Recording Comprehensive Metadata:** Maintaining thorough metadata that explains the meaning and context of each data field is crucial. For instance, including information like the units of measurement, data source, and date of collection.   + **Why it improves data quality:** Metadata provides crucial context for understanding and interpreting data, making it easier to identify potential errors and ensure the data's long-term usability.   **4. Data Analysis and Retrieval Phase:**   * **Data Exploration:** This involves using various techniques to determine appropriate models, identify data bugs, and develop domain expertise. For instance, visualizing data distributions, checking for outliers, and understanding relationships between variables.   + **Why it improves data quality:** Provides a deeper understanding of the data, allowing for the identification of potential data quality issues and informing decisions about appropriate cleaning and analysis techniques. * **Continuous Analysis:** Regularly analysing data for stability and consistency helps detect changes or anomalies that might indicate data quality problems.   + **Why it improves data quality:** Early detection of issues allows for timely interventions and prevents errors from propagating through subsequent analysis stages.   **5. General Data Quality Methods:**   * **Implementing Data Quality Constraints:** Establishing and enforcing both static (based on schema) and dynamic (based on workflow) constraints can help prevent and detect data quality issues. For example, static constraints might include defining data types, ranges, and mandatory fields, while dynamic constraints could enforce business rules or data validation checks.   + **Why it improves data quality:** Constraints act as automated checks, ensuring that data adheres to predefined rules and preventing inconsistencies or invalid data. * **Using Data Quality Metrics:** Defining and measuring data quality metrics, such as accuracy, completeness, consistency, and timeliness, helps quantify data quality and track improvements over time.   + **Why it improves data quality:** Metrics provide a tangible measure of data quality, enabling the identification of areas for improvement and demonstrating the impact of data quality initiatives.   **6. Tools for Data Cleaning:**   * **Spreadsheets:** For smaller datasets, spreadsheets can be effective for basic data cleaning tasks, such as identifying missing values, correcting typos, and removing duplicates. * **Purpose-Built Tools:** Tools like RapidMiner, Tableau Prep, and OpenRefine offer more advanced features for data cleaning, transformation, and standardization. * **General Purpose Languages:** Languages like Python and R provide flexibility and customization for complex data cleaning tasks, allowing for the creation of custom scripts and algorithms.   **Important Considerations:**   * **Data Size and Complexity:** The choice of data quality methods and tools will depend on the size and complexity of the dataset. * **Frequency of Data Cleaning:** Regular data cleaning procedures should be established, especially for dynamic datasets. * **Documentation:** All data cleaning steps and decisions should be documented to ensure transparency and reproducibility.   **Disclaimer:** While the sources provide valuable insights into data quality management, they do not offer specific advice on how to address particular errors in a given dataset. Therefore, the provided suggestions are general recommendations based on common data quality practices and principles. |

***[End of Question 3]***

CA682D – Data Management and Visualisation

August/Resit Examinations 2021/2022 Page 8 of 12

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

**Q 4(a) [10 Marks]**

(i) You are asked to plan a data analytics project to *analyse 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,

Analysing, 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. Anonymising student comments that include identifying details.

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

7. Conducting student surveys to answer the key questions about their experience.

8. Creating a document to share with senior university management summarising the findings.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Answer:  Gathering Processing Analysing Presenting Preserving Analysing Student Feedback on Online Teaching Here's a breakdown of the activities involved in a data analytics project for analysing student feedback on online teaching at DCU, categorized according to the Generic Data Analytics Pipeline:   |  |  |  | | --- | --- | --- | | **Activity** | **Category** | **Tool/Application** | | 1. Documenting data formats and saving datasets. | **Preserving** | Python (Jupyter Notebooks) | | 2. Removing incorrect entries from datasets. | **Processing** | OpenRefine | | 3. Liaising with DCU Registry to obtain datasets. | **Gathering** | N/A (This is a human interaction task) | | 4. Calculating average satisfaction levels from sentiment ratings. | **Analysing** | Python (using libraries like Pandas, NumPy) | | 5. Anonymising student comments containing identifying details. | **Processing** | Python (using libraries for text processing) | | 6. Converting student words into sentiment ratings and correlating with field of study. | **Analysing** | Python (using sentiment analysis libraries) | | 7. Conducting student surveys. | **Gathering** | Google Forms, SurveyMonkey | | 8. Creating a document summarizing findings for senior university management. | **Presenting** | Microsoft Word, PowerPoint |   **Explanation:**   * **Gathering:** This stage involves collecting data from various sources. In this case, liaising with the DCU Registry to obtain existing datasets and conducting student surveys are the primary gathering activities. * **Processing:** This stage focuses on cleaning and preparing the data for analysis. Removing incorrect entries and anonymising student comments to ensure data privacy fall under processing. * **Analysing:** This stage involves extracting meaningful insights from the data. Calculating average satisfaction levels from sentiment ratings and correlating sentiment with fields of study are analytical tasks. * **Presenting:** This stage focuses on communicating the findings effectively. Creating a summary document for university management is a presentation activity. * **Preserving:** This stage involves storing and managing the data for future use and reproducibility. Documenting data formats and saving all created datasets are crucial for preservation.   **Tool/Application Rationale:**   * **Python (Jupyter Notebooks):** A versatile programming language well-suited for data analysis and visualization. It offers numerous libraries for data manipulation, statistical analysis, and sentiment analysis, making it suitable for Processing, Analysing, and Preserving. * **OpenRefine:** A powerful tool specifically designed for data cleaning and transformation. It's particularly useful for tasks like removing incorrect entries and standardizing data formats. * **Google Forms/SurveyMonkey:** Online survey platforms ideal for creating and distributing student surveys, making data collection efficient and organized. * **Microsoft Word/PowerPoint:** Standard tools for creating professional documents and presentations for communicating findings to senior management.   **Important Note:** The tool choices are based on the sources and common practices in data analytics. Specific tools might vary depending on the project's specific requirements and the team's familiarity with different software. |

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

|  |
| --- |
| Answer: Weaknesses of the Generic Data Analytics Pipeline While the Generic Data Analytics Pipeline provides a structured framework for data analysis projects, it has some limitations. One important task that is often overlooked in this pipeline is **data ethics and privacy**. The sources emphasize the importance of data quality and cleaning but don't explicitly address the ethical considerations and potential risks associated with data handling.   * **Ethical considerations** encompass responsible data collection, storage, use, and sharing practices. These considerations become particularly crucial when dealing with sensitive personal data, such as student feedback, which may contain identifiable information. * **Data privacy** focuses on protecting individuals' personal information and ensuring that data is used and shared in accordance with relevant laws and regulations, such as the European GDPR.   **Addressing Data Ethics and Privacy in the Pipeline:**  To address this weakness, the Generic Data Analytics Pipeline should incorporate explicit steps for:   * **Data Ethics Assessment:** Evaluating the ethical implications of the project, considering potential biases, fairness, transparency, and accountability in data handling. * **Privacy Compliance:** Ensuring compliance with relevant data protection regulations, including obtaining informed consent for data collection, implementing appropriate data security measures, and anonymising or pseudonymising data where necessary. * **Data Governance:** Establishing clear policies and procedures for data management, access control, and data sharing to ensure responsible and ethical data practices.   **Integration of Ethics and Privacy:**  Integrating these steps into the existing pipeline stages can strengthen the ethical and privacy aspects of data analytics projects. For instance:   * **Gathering:** Ensuring ethical data collection practices and obtaining informed consent from participants. * **Processing:** Anonymising or pseudonymising sensitive data to protect individual privacy. * **Analysing:** Considering potential biases in data analysis and interpreting results responsibly. * **Presenting:** Communicating findings in a transparent and ethical manner, avoiding misleading visualizations or interpretations. * **Preserving:** Implementing secure data storage and access control measures to protect data confidentiality and integrity.   By explicitly addressing data ethics and privacy, the Generic Data Analytics Pipeline can be enhanced to ensure responsible and trustworthy data-driven insights. |

CA682D – Data Management and Visualisation

August/Resit Examinations 2021/2022 Page 9 of 12

**Q 4(b) [8 Marks]**

For each of the following data attributes (A-D), choose all of the following

descriptions that can apply. Marks will be deducted for including wrong choices.

*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 Data Attributes and their Descriptions Here's a breakdown of the data attributes and their corresponding descriptions, based on the information from the sources and common data analysis concepts:  **A. Rating of temperature comfort in offices (cold, cool, perfect, warm, hot)**   * **Qualitative:** This attribute describes a quality or characteristic, rather than a numerical measurement. * **Ordinal:** The categories have a natural order (from cold to hot), indicating a ranking or progression.   **B. Number of times a character's name is used in a TV show episode**   * **Quantitative:** This attribute represents a numerical measurement, the count of character name occurrences. * **Discrete:** The values can only take on whole numbers (you can't have half a mention of a name). * **Ratio:** There's a true zero point (a character can be mentioned zero times), and ratios are meaningful (e.g., a character mentioned 10 times is mentioned twice as often as a character mentioned 5 times).   **C. Names of pets owned by all CA682 students**   * **Qualitative:** This attribute describes a characteristic (the type of pet), not a numerical value. * **Nominal:** The categories have no inherent order or ranking.   **D. All winning times (in seconds) for men's 100m sprint at the Olympic Games**   * **Quantitative:** This attribute represents a numerical measurement of time in seconds. * **Continuous:** The values can theoretically take on any value within a given range (limited by the precision of timing devices). * **Ratio:** There's a true zero point (0 seconds), and ratios are meaningful (e.g., a time of 10 seconds is twice as long as a time of 5 seconds).   **Key Concepts:**  **Qualitative vs. Quantitative:** Qualitative data deals with non-numerical descriptions or categories, while quantitative data involves numerical measurements.  **Discrete vs. Continuous:** Discrete data can only take on specific values (usually whole numbers), while continuous data can theoretically take on any value within a given range.  **Nominal, Ordinal, Interval, Ratio:** These are the four levels of data measurement, increasing in complexity and the types of mathematical operations allowed.   * + **Nominal:** Categories with no inherent order or ranking.   + **Ordinal:** Categories with a natural order but no defined intervals.   + **Interval:** Ordered data with equal intervals between values but no true zero point.   + **Ratio:** The most informative level, with a true zero point and meaningful ratios. |

**Q 4(c) [7 Marks]** Choose **one (1)** of the following scenarios and explain (in your own words and 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: Analysis of Supermarket Loyalty Card Data as "Big" Data **Scenario A**, customer account, purchasing data, and engagement data from a supermarket chain’s loyalty card program, can be considered a good example of "big" data, based on the three classical characteristics: volume, velocity, and variety.  **Assumptions:**   * The supermarket chain is large and has a significant number of customers participating in the loyalty program. * Data is collected on a regular basis (e.g., with every purchase). * The data collected includes various attributes, such as customer demographics, purchase history, product preferences, engagement with promotional offers, website browsing history, and social media interactions.   **Analysis Based on the Three V's:**   * **Volume:** The volume of data generated by a large supermarket chain’s loyalty program is likely to be substantial. Each customer generates data with every transaction, interaction, and point earned. This data accumulates rapidly, especially considering the number of customers, products, and stores involved. * **Velocity:** The data is generated at a high rate as customers shop frequently and interact with the loyalty program through various touchpoints. Real-time data streams from point-of-sale systems, online platforms, and mobile apps contribute to the rapid data influx. * **Variety:** The loyalty program captures diverse data types. This includes structured data from customer accounts and transactions, semi-structured data from website logs and social media interactions, and potentially unstructured data like customer reviews or images.   **Example Supporting Variety:**  For instance, a single customer might generate data points from:   * **Structured Data:** Customer ID, purchase amount, date and time of purchase, items purchased, loyalty points earned, payment method. * **Semi-structured Data:** Website browsing history (pages visited, products viewed, searches performed), interactions with promotional emails or app notifications. * **Unstructured Data:** Social media posts mentioning the supermarket or its products, customer feedback surveys, product reviews.   **Challenges and Considerations:**  While this scenario exhibits characteristics of "big" data, challenges may arise:   * **Data Storage and Management:** Handling the massive and diverse data requires robust data storage solutions, potentially utilizing technologies like NoSQL databases. * **Data Cleaning and Processing:** Ensuring data quality and consistency across various sources is crucial for accurate analysis. Data cleaning tools like OpenRefine might be necessary. * **Data Security and Privacy:** Protecting sensitive customer information in compliance with regulations like GDPR is essential.   **Conclusion:**  Given these factors, the loyalty card data scenario is a good example of "big" data due to its potential volume, velocity, and variety. However, addressing the inherent challenges through appropriate data management, cleaning, and security measures is crucial for effectively leveraging this data for business insights and decision-making. Is an Individual's Step Count Data "Big Data?" **Scenario B**, an individual's step count data for a one-year period from a personal smart device, is likely **not** a good example of "big" data, despite the potential for a large amount of data points, when considering the classical characteristics of volume, velocity, and variety.  **Assumptions:**   * The smart device records step count data continuously throughout the day. * Data is stored locally on the device and potentially synced to a cloud service.   **Analysis Based on the Three V's:**   * **Volume:** While a year's worth of step count data might seem like a lot, it is unlikely to reach the volume typically associated with "big" data. A single data point (step count) is likely recorded at frequent but regular intervals (e.g., every minute). This results in a large number of data points over a year but may not be massive enough to require specialized big data technologies for storage or processing. Traditional data storage and analysis tools are usually sufficient for this volume of data. * **Velocity:** The rate at which step count data is generated is relatively consistent and predictable. The smart device records data at a set frequency, resulting in a steady data stream rather than the high-velocity bursts or fluctuations seen in scenarios like social media trends or financial markets. * **Variety:** This scenario exhibits low variety. The data primarily consists of step counts, possibly with timestamps. While additional data points like heart rate or sleep patterns might be collected, the overall data structure remains relatively simple and homogeneous. This limited variety doesn't pose the same challenges as handling diverse data formats and sources that characterize "big" data scenarios.   **Conclusion:**  An individual's step count data, while potentially voluminous over a year, lacks the high velocity and significant variety typically associated with "big" data. Therefore, it's not a strong example of "big" data based on the classical three V's definition. Conventional data analysis methods and tools are generally sufficient for managing and analyzing this type of personal health data. Is "Stranger Things" a Good Example of Big Data? **Scenario C**, encompassing all 8 episodes (video files) of the TV show "Stranger Things," presents a somewhat ambiguous case when assessed against the criteria of "big" data. Let's analyze this using the three classic characteristics: volume, velocity, and variety.  **Assumptions:**   * We are considering high-definition video files, likely in a format like 1080p or 4K resolution. * The data includes only the raw video files, without additional metadata or associated information like subtitles, audio tracks, or production files.   **Analysis Based on the Three V's:**  **Volume:** Video files, especially in high definition, are inherently large. A single episode of a TV show can easily range from several gigabytes to tens of gigabytes in size. 8 episodes combined would likely result in a dataset measuring in the tens to hundreds of gigabytes. This volume is considerable but might not reach the petabyte or exabyte scale often associated with "big" data in fields like social media analytics or scientific research.   * + However, the definition of "big" data is constantly evolving. What was considered "big" a decade ago might be considered commonplace today. In the context of personal media consumption or even smaller production studios, this dataset's size could be considered "big."   **Velocity:** The velocity in this scenario is low. The data is static; it was created at a particular time and remains unchanged. There's no continuous data stream or rapid generation of new data points. The 8 video files are a fixed dataset, unlike scenarios with real-time data feeds or frequent updates.  **Variety:** The variety in this scenario is also low. We are dealing solely with video files, likely in a standardized format. There's no diversity in data types or sources as seen in datasets that encompass text, images, sensor data, social media interactions, etc.  **Conclusion:**  Classifying "Stranger Things" video files as "big" data is not straightforward. The volume can be significant, but the lack of high velocity and variety makes it a weak example based on the classical three V's definition.  However, it's crucial to consider the context. In situations where resources for storage, processing, and analysis are limited, this dataset could pose challenges akin to those encountered with "big" data.  Ultimately, the determination depends on the specific analytical tasks and available resources. For tasks like simple playback or basic editing, the dataset might not require specialized "big data" tools. However, for more complex analysis involving computer vision techniques, scene recognition, or extraction of detailed information from the videos, the volume might necessitate the use of distributed computing frameworks or other "big data" approaches. |

***[End of Question 4]***

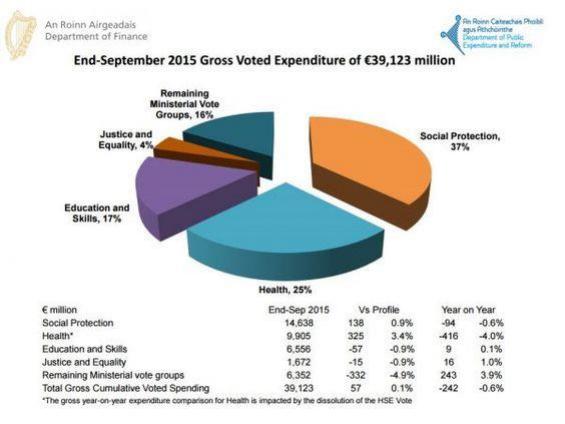
CA682D – Data Management and Visualisation

August/Resit Examinations 2021/2022 Page 10 of 12

***QUESTION 5 [TOTAL MARKS: 25]***

**Q 5(a) [9 Marks]**

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



|  |
| --- |
| Answer: |

**Q 5(b) [8 Marks]**

Given the following visualisation 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.

|  |
| --- |
| A:  B: |

CA682D – Data Management and Visualisation

August/Resit Examinations 2021/2022 Page 11 of 12

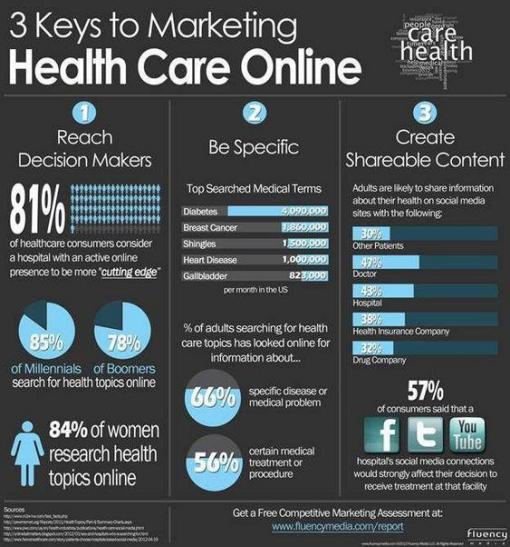
**Q 5(c) [8 Marks]**

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?

|  |
| --- |
| (i) (ii) |



***[End of Question 5]***

***[END OF EXAM]***

CA682D – Data Management and Visualisation

August/Resit Examinations 2021/2022 Page 12 of 12