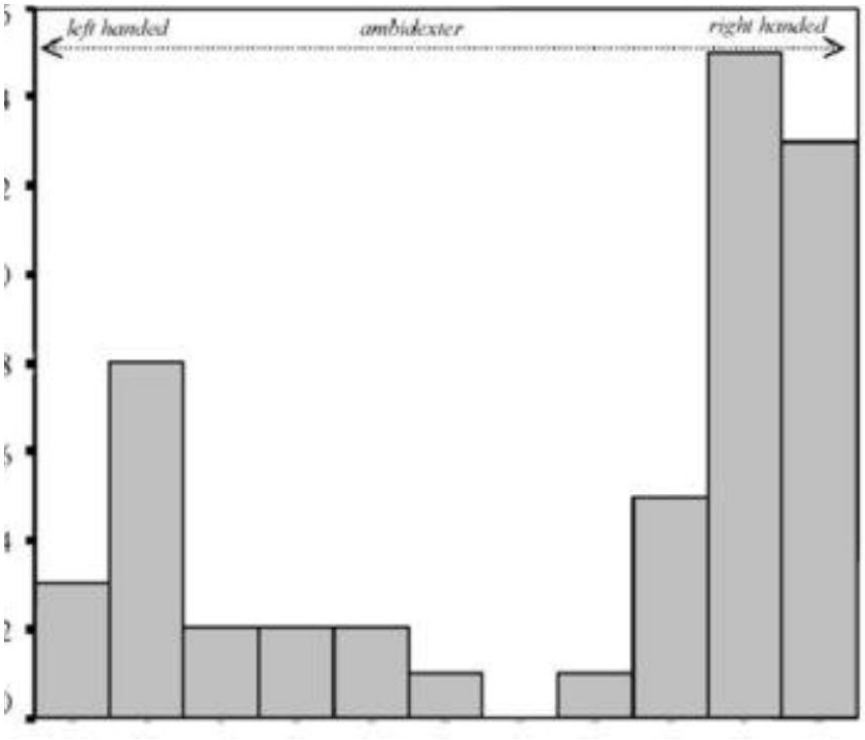


QUESTION 1 A

Q 1(a) [7 Marks]

(i) Describe the shape of the histogram below in terms of the modality.

[2 marks]

(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

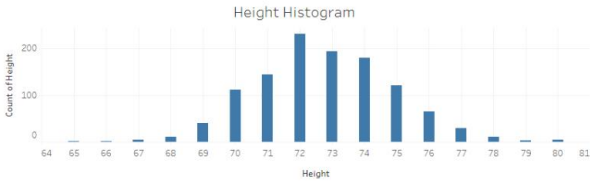
? 52

? 23

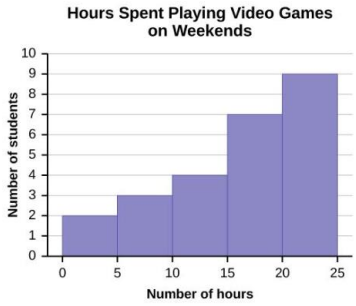
? 40

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

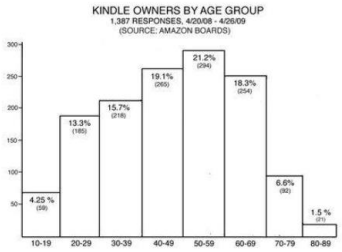
A:



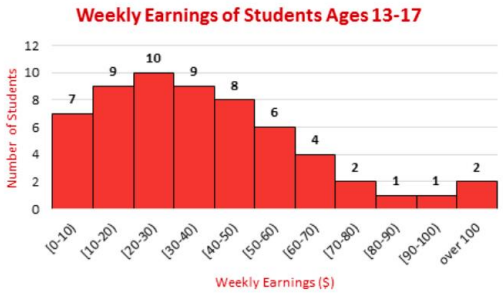
B:



C:



D:



[3 marks]

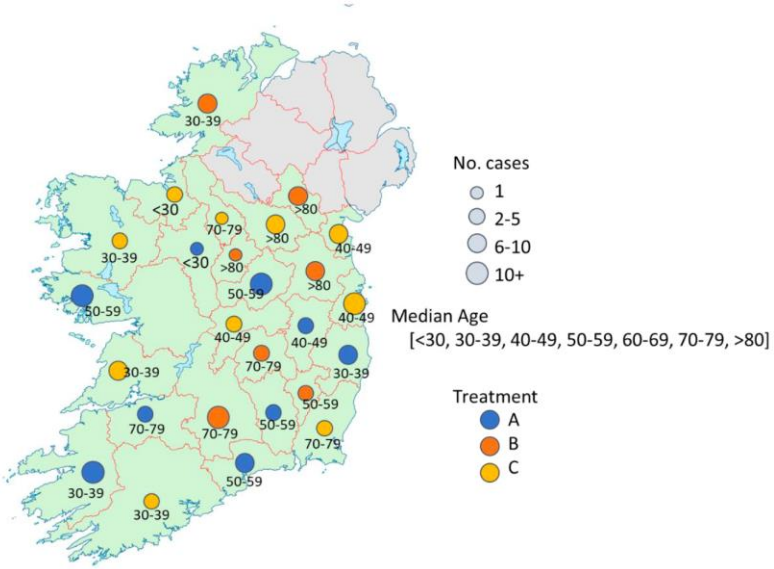
A:

B:

C:

D:

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]

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

(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]

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%

ANSWERS:

Question 1

Q1(a) [7 Marks]

1. **Describe the shape of the histogram in terms of modality (2 Marks):**
    - The modality of a histogram refers to the number of peaks it has:
        - **Unimodal**: One peak.
        - **Bimodal**: Two peaks.
        - **Multimodal**: More than two peaks.
    - Describe the histogram based on its visual shape (e.g., "unimodal with one peak centered around X").
2. **Values in the first column of a frequency table for discrete variables (2 Marks):**
    - Example values could represent categories or discrete states, such as toothbrush colors: "Red," "Blue," "Green," "Yellow."
3. **Identify histogram with positive skewness (3 Marks):**
    - A histogram has **positive skewness** if most data is concentrated on the left, with a long tail extending to the right.
    - Select the histogram (A-D) with this characteristic.

Q1(b) [10 Marks]

1. **What type of visualization is shown? (3 Marks):**
    - Identify the type, e.g., a **stacked bar chart, bubble chart, or scatter plot.**
    - **Marks and attributes** used:
        - Marks: Bars, points, or shapes.
        - Attributes: Size, position, and color encode data.
2. **Privacy issues raised by the graph (3 Marks):**
    - Risks include identifying patients in small counties due to unique combinations of age, therapy type, and condition.
    - Example: Small sample sizes could allow re-identification when combined with other datasets.
3. **Risks and methods to reduce privacy concerns (4 Marks):**
    - Risks:
        - Further aggregation (e.g., blood type and ethnicity) increases re-identification risks.
    - Methods:
        - **Anonymization**: Remove or mask identifying features.
        - **Data minimization**: Share only aggregated data (e.g., totals instead of detailed breakdowns).

Q1(c) [8 Marks]

Privacy-preserving method to show student performance:

1. **Visualization method**:
    - A **box plot** or **heatmap** showing overall performance distributions by module without revealing individual scores.
    - Alternatively, assign anonymized IDs to students.
2. **Potential risks**:
    - Indirect identification: Patterns (e.g., top scorer) could be used to infer identities.

- Data breaches: Secure the visualization to prevent unauthorized access.

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."

(i) List three (3) important questions you would ask your client about their data storage requirements.

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

(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?

Q 2(b) [9 Marks]

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

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

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.

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.

Question 2

Q2(a) [10 Marks]

1. **Three important questions to ask the client about their data storage requirements (3 Marks):**
    - **Data Volume**: What is the expected size of inventory and sales data over time?
    - **Access Needs**: Who needs access to the data, and how frequently will it be updated?
    - **Data Retention Policy**: How long does the client want to store historical data for analysis?
2. **Type of data storage approach and reason for the choice (4 Marks):**
    - **Recommended Approach**: Cloud-based relational database system (e.g., MySQL on AWS RDS or PostgreSQL).
    - **Reason**: It supports structured data with relationships (e.g., product IDs linked to transactions) and offers scalability for future needs.
3. **Effect of including website logs and social media content (3 Marks):**
    - **Would the recommendation change?**: Yes.

- **Reason**: Website logs and social media interactions are unstructured data and require different handling.
    - Use a hybrid approach combining a relational database for inventory/sales data and a NoSQL database (e.g., MongoDB or Elasticsearch) for unstructured data.

Q2(b) [9 Marks]

1. **Three examples of metadata for your favorite clothing item (3 Marks):**
    - Material: Cotton.
    - Size: Medium.
    - Color: Black.
2. **Metadata type and explanation (3 Marks):**
    - **Material**: Descriptive metadata (describes the product's characteristics).
    - **Size**: Structural metadata (relates to fit and design).
    - **Color**: Descriptive metadata (visual characteristic of the item).
3. **Effect of using a standard and difficulty in enforcement (3 Marks):**
    - **Effect**: A metadata standard ensures consistency and interoperability, improving the quality and usability of metadata (e.g., same format for size: "M" vs. "Medium").
    - **Difficulty**: Standards may be inconsistently applied by users, leading to errors or deviations.

Q2(c) [6 Marks]

Possible data requiring different handling due to GDPR requirements:

1. **Example of sensitive data**:
    - Customer interactions from social media may include personal data like names, photos, or locations.
    - Website logs might contain IP addresses, which are classified as personal data under GDPR.
2. **Explanation**:

- GDPR requires special handling of personal data, such as anonymization or consent for processing.
- Sensitive data must be stored securely, and access should be limited to authorized personnel.

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.

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.

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?

ANSWERS:

Q3(a) [13 Marks]

1. **Four possible errors or artifacts in the dataset (linked q3-data.csv):**
    - **Missing Data**: Certain cells may have no values, e.g., production data for a country in a specific year is blank.
    - **Inconsistent Units**: Some rows might use "Tonnes" while others use "Kilograms," leading to inconsistencies.
    - **Outliers**: Extremely high or low values (e.g., production of 1,000,000 Tonnes for a small country) could indicate data entry errors.
    - **Duplicate Entries**: The same country and year may appear multiple times with slight variations.
2. **Tools used**:
    - Use **Excel** or **Python (pandas)** to identify missing data, duplicates, or outliers.
    - **Visualization tools** like Tableau or Power BI can help spot anomalies.

Q3(b) [6 Marks]

1. **Introduction of errors and their pipeline phases**:
    - **Missing Data**: Likely introduced during the **data collection** phase due to incomplete submissions or communication errors.
    - **Inconsistent Units**: Occurs during the **data processing** phase if standardization rules were not applied.
    - **Outliers**: May stem from errors during **data entry** in the collection phase.
    - **Duplicate Entries**: Can occur during the **data integration** phase when combining multiple datasets.
2. **Assumptions**:
    - Errors result from manual processes or lack of validation during data collection and processing.

Q3(c) [6 Marks]

1. **Data quality methods to avoid or mitigate errors**:
    - **Missing Data**: Use imputation techniques to fill gaps (e.g., mean or median substitution for numerical data).
    - **Inconsistent Units**: Standardize all units during data preprocessing using a common scale (e.g., converting Kilograms to Tonnes).
    - **Outliers**: Apply statistical tests to identify and review extreme values; remove or adjust if necessary.
    - **Duplicate Entries**: Use de-duplication algorithms in tools like Python (pandas .drop_duplicates()function).
2. **Why these methods improve quality**:
    - Ensures consistency, reduces noise, and enhances the reliability of insights derived from the dataset.

QUESTION 4 [TOTAL MARKS: 25]

Q 4(a) [10 Marks]

(i) 1. 2. 3. 4. 5. 6. 7. 8. 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).

Documenting the data formats used in the study and saving all of the created

datasets.

Removing incorrect entries from the student datasets.

Liaising with DCU Registry to get datasets from the student registration and

results systems.

Calculating the average satisfaction levels based on the sentiment ratings.

Anonymising student comments that include identifying details.

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

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

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

Answer:

Gathering

Processing

Analysing

Presenting

Preserving

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

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

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"

ANSWERS:

Question 4

Q4(a) [10 Marks]

Mapping activities to categories in the Generic Data Analytics Pipeline:

1. **Gathering**:
    - Conducting student surveys to answer the key questions about their experience.
    - Liaising with DCU Registry to get datasets from the student registration and results systems.
    - **Tool**: Google Forms, SQL query tools.
2. **Processing**:
    - Removing incorrect entries from the student datasets.
    - Anonymizing student comments that include identifying details.
    - **Tool**: Python (pandas), Excel, or dedicated anonymization tools.
3. **Analysing**:

- Calculating the average satisfaction levels based on the sentiment ratings.
- Converting student words into sentiment ratings and correlating with field of study.
- **Tool**: Python (Natural Language Toolkit), R.

4. **Presenting**:
   - Creating a document to share with senior university management summarizing the findings.
   - **Tool**: PowerPoint, Tableau, or Word.
5. **Preserving**:
   - Documenting the data formats used in the study and saving all the created datasets.
   - **Tool**: Cloud storage solutions like Google Drive or AWS S3.

Weakness of the Generic Data Analytics Pipeline:

- It doesn't explicitly address **data security and privacy** throughout the pipeline, which is critical for sensitive datasets.
- **Suggestion**: Incorporate a security-focused phase to ensure data protection at every step.

Q4(b) [8 Marks]

Descriptions for data attributes (A-D):

1. **Rating of temperature comfort (cold, cool, perfect, warm, hot)**:
   - **Qualitative** (describes a subjective category).
   - **Ordinal** (ordered categories without measurable intervals).
2. **Number of times a character's name is used in a TV show episode**:
   - **Quantitative** (numerical count).
   - **Discrete** (countable values).
   - **Ratio** (zero means no occurrences; ratios are meaningful).
3. **Names of pets owned by CA682 students**:

    - **Qualitative** (categorical information).
    - **Nominal** (categories without order).
4. **Winning times for men's 100m sprint at the Olympics (in seconds)**:
    - **Quantitative** (numerical measurement).
    - **Continuous** (can take any fractional value).
    - **Ratio** (zero represents an absence of time; ratios are meaningful).

Q4(c) [7 Marks]

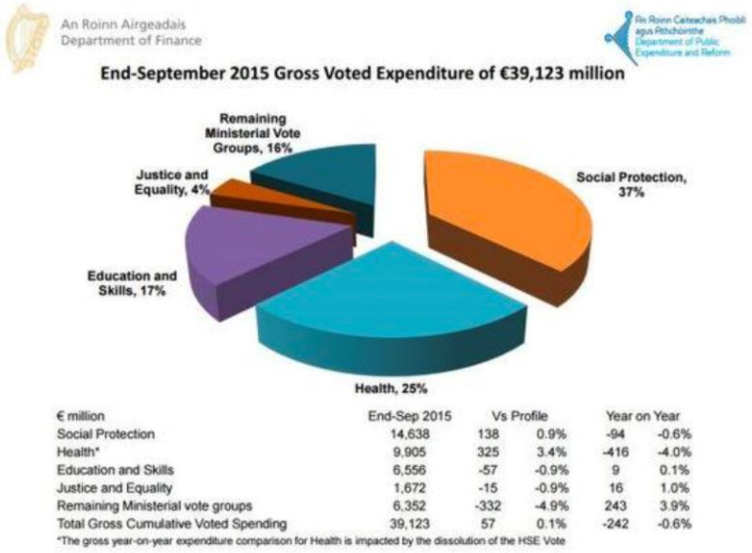
Scenario Analysis for Big Data:
Option A: Customer account, purchasing data, and engagement data from a supermarket loyalty program.

- **Volume**: Large data volume due to multiple customers and transactions.
- **Velocity**: Real-time or near-real-time data generation (e.g., purchases, interactions).
- **Variety**: Structured data (purchases) and unstructured data (engagement interactions).
- **Conclusion**: This qualifies as big data due to its high volume, velocity, and variety.

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.



€ million	End-Sep 2015	Vs Profile		Year on Year	
Social Protection	14,638	138	0.9%	-94	-0.6%
Health*	9,905	325	3.4%	-416	-4.0%
Education and Skills	6,556	-57	-0.9%	9	0.1%
Justice and Equality	1,672	-15	-0.9%	16	1.0%
Remaining Ministerial vote groups	6,352	-332	-4.9%	243	3.9%
Total Gross Cumulative Voted Spending	39,123	57	0.1%	-242	-0.6%

*The gross year-on-year expenditure comparison for Health is impacted by the dissolution of the HSE Vote

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:

5(c) [8 Marks]

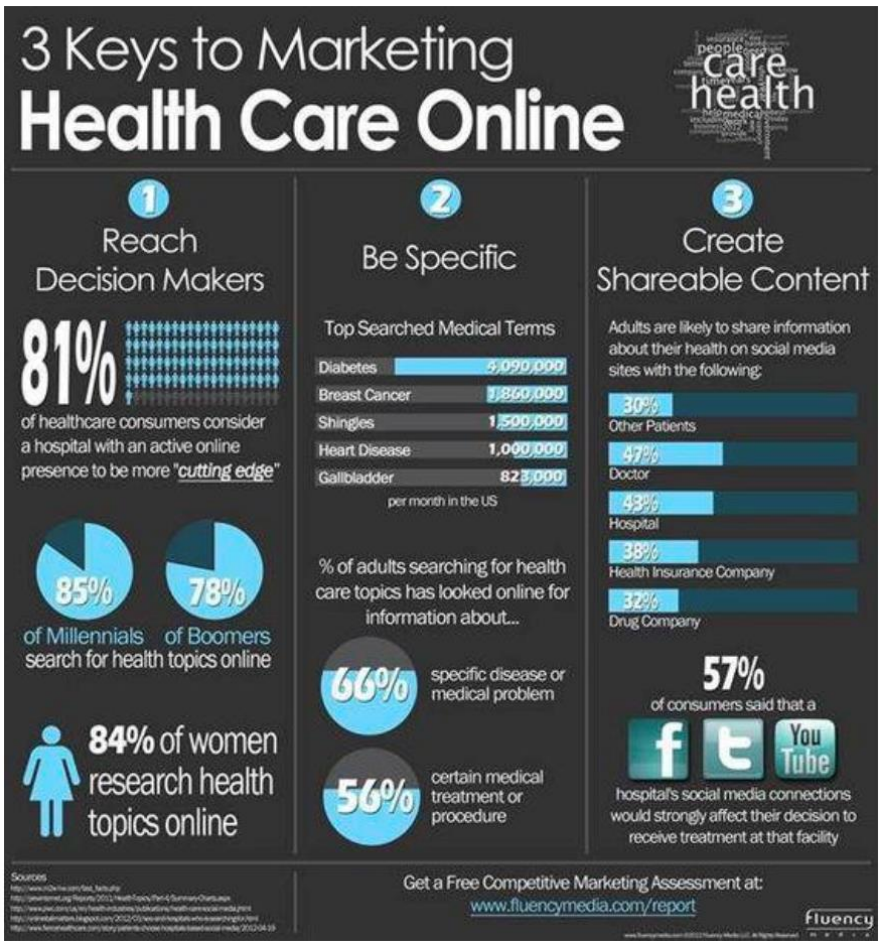
Answer the following questions relating to the graphic shown below:

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

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

(i)

(ii)



ANSWERS:

Q5(a) [9 Marks]

Identify three possible improvements to the graph below. Justify your choices, referencing design rules and theories.

Without the actual graph, general improvements could be:

1. **Increase Label Clarity**: Ensure all axes are labeled clearly and concisely. For example, if the x-axis and y-axis lack titles or have unclear labels, viewers might misinterpret the data. Proper labeling adheres to Tufte's principle of eliminating ambiguity.

2. **Improve Color Choices**: Use a colorblind-friendly palette to ensure accessibility for all viewers. Avoid colors that may confuse such as red/green combinations, aligning with universal design principles.
3. **Minimize Data Overload**: If the graph contains excessive data points or clutter, consider simplifying by using aggregation or reducing non-essential elements. Following the Gestalt principle of simplicity can make the graph more digestible.

Q5(b) [8 Marks]

Suggest an appropriate graph type and CHRTS category for the tasks.

- **Task A**: Compare the performance of stocks in Microsoft, Apple, and Samsung over the last 5 years.
  Graph Type: Line chart
  CHRTS Category: Trends
  Justification: A line chart effectively shows changes over time for multiple entities, making it easy to compare stock performance trends.
- **Task B**: Explore movie commercial performance for the IMDB top 50 by director based on cost to make and ticket sales.
  Graph Type: Scatter plot
  CHRTS Category: Relationships
  Justification: A scatter plot would visualize the relationship between production cost and ticket sales for each movie, allowing identification of outliers or patterns.

Q5(c) [8 Marks]

Answer questions related to the graphic shown.

(i) **What is the main communication purpose and why?**
The main purpose is likely to compare, trend, or relate data points depending on the type of chart. For example, if it's a scatter plot, the purpose might be to highlight relationships between two quantitative variables.

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

- Use of clear legends or direct labeling to ensure the chart is self-explanatory.
- Adequate spacing and alignment of data points to avoid overlap, supporting data legibility.

- Use of appropriate scaling to avoid distortion, maintaining truthful representation of data values.