Sample CA682 exam questions 2022

You will be given a document to download containing 4 questions. You choose to answer 3. If you answer all 4 the one with the lowest mark will be discarded. Each question is worth 20% of the exam total. Answer the question in the indicated boxes in this document, save as word or pdf format and upload to loop.

Q1 Visualisation Design [20 marks]

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

Is this an exploratory or explanatory visualisation task?

Answer:

Who is the intended audience for the data visualisation?

Answer: DCU Senior Management

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

Answer: Student Accommodation at DCU: 60% Live Within 10 km of Campus

Why This Title?

Focus on the Key Insight:

The primary conclusion from the data is that a significant proportion (60%) of students live close to the Glasnevin campus (within 10 km). Highlighting this insight immediately captures the main message.

Audience Relevance:

Senior management is the audience, and they likely care about actionable insights like the proximity of students to campus, which impacts planning for accommodation, transport, and campus facilities.

Clarity and Simplicity:

The title is concise, making it easy to understand at a glance, which is important for busy decision-makers.

Assumption About Conclusion:

The conclusion implies that proximity (distance from campus) is a critical factor for the majority of students, potentially influencing rent and access to university services.

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

Answer: Stacked Bar Chart

Stacked Bar Chart with Geographical Distribution by Distance Category

---

Why Use a Stacked Bar Chart?

1. **Data Type Suitability:**
   - The dataset includes **categorical data** (distance ranges) and **quantitative data** (number of students and percentages).
   - A stacked bar chart allows you to display the proportion of students in each distance category (e.g., 35% for 0-5 km) while also visualizing the cumulative contribution (e.g., 60% live within 10 km).
2. **Clarity of the Message:**
   - The key message (60% of students live within 10 km) can be visually emphasized by grouping the 0-5 km and 6-10 km categories and showing their combined proportion in the total bar.
   - The remaining categories (11-20 km, 21-50 km, 50+ km) can highlight the spread of the remaining 40%.
3. **Audience Understanding:**
   - Senior management often prefers clear, high-level visual summaries. A stacked bar chart is intuitive and easy to interpret at a glance.
   - It effectively compares relative sizes while emphasizing the cumulative percentage.
4. **Geographical Context:**
   - While a stacked bar chart focuses on proportions, it can be paired with a **simple map** or heatmap to highlight the geographical distribution more effectively.

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

Answer: **Marks and Attributes to Encode the Data**

In the bar chart provided, the marks and attributes encode the data as follows:

---

1. Marks:

- **Bar Height (Length):**
  Represents the percentage of enrolled students in each distance category.
  - Example: The height of the bar for "0-5 km" is **35%**, reflecting the corresponding proportion of students.
- **Bar Position (X-Axis):**
  Represents the **distance categories** (e.g., "0-5 km", "6-10 km").
  - This organizes the data sequentially by increasing distance from the campus.

---

2. Attributes:

Visual Encoding:

1. **Color (Fill):**
   - Used to distinguish between two key groups:
     - **Highlighted Bars (0-5 km and 6-10 km):**
       - Color: **#5DADE2** (blue) to emphasize the combined 60%.
     - **Remaining Bars (11-20 km, 21-50 km, 50+ km):**
       - Color: **#AED6F1** (lighter blue) for de-emphasis.
   - The color distinction helps the audience quickly identify and focus on the cumulative insight.
2. **Text (Annotations):**
   - Each bar includes a **label** displaying its percentage value (e.g., "35%", "25%").
   - The annotation for the combined "60%" within 10 km is emphasized using **an arrow** and **bold text** to highlight the key takeaway.
3. **Y-Axis (Scale):**

- The scale ranges from **0% to 50%**, matching the highest value in the dataset for clarity and reducing unnecessary empty space.

4. **X-Axis (Category Labels):**

- The categories ("0-5 km", "6-10 km", etc.) are explicitly labeled for easy interpretation.

5. **Arrow (Annotation):**

- An arrow connects the label **"60% within 10 km"** to the cumulative value on the chart, providing a direct visual cue for the most important insight.

Design Attributes:

1. **Font Size and Style:**

- Labels: **10 pt, bold**, with white text for contrast against the bars.
- Title and Axes: **12-14 pt**, clear and professional fonts (e.g., Arial).

2. **Gridlines:**

- Horizontal gridlines are dashed and semi-transparent, helping with value interpretation without overwhelming the visual.

How These Attributes Support Interpretation

- **Hierarchy of Importance:**

  - Bold, darker colors and annotations focus attention on the 60% insight.
  - Lighter colors and supporting labels convey secondary data, ensuring the main message is not lost.

- **Accessibility:**

  - Numeric labels and clear axis titles help ensure all audience members, regardless of their familiarity with visualisations, can interpret the data easily.

- **Professionalism:**

  - The clean, minimalistic design aligns with the expectations of a senior management audience, balancing clarity and aesthetics.

Considering the purpose and intended audience, comment on how you would use colour or layout principles for this data visualisation.

Answer: **Using Colour and Layout Principles for the Data Visualisation**

1. Colour Principles

1. **Purposeful Colour Coding:**
    - Use **contrasting colours** to distinguish between key data points and supporting information.
        - Highlight the **0-5 km and 6-10 km categories** (combined 60%) using a **bold blue** like **#5DADE2**.
        - Use a **lighter shade of blue** like **#AED6F1** for the remaining categories (11-20 km, 21-50 km, 50+ km).
2. **Consistency:**
    - Maintain consistent colour choices for similar categories across all charts (if part of a broader presentation).
    - Ensure colours for bars, annotations, and any legends align with the visual style.
3. **Accessibility:**
    - Use colourblind-friendly palettes and ensure good contrast between the bars, text, and background. For instance:
        - Bars: Bold blues.
        - Background: Neutral (e.g., white or light grey).
        - Text on Bars: White or dark grey for high contrast.
4. **Focus Attention:**
    - Use **accent colours sparingly** for annotations (e.g., red or bold black for the "60% within 10 km" label). This draws attention to the primary insight without overwhelming the visualisation.

---

2. Layout Principles

1. **Logical Flow:**
    - Arrange the visualisation elements in a **top-to-bottom or left-to-right hierarchy**, reflecting how the audience processes information:
        - Title: At the top, prominently displaying the main takeaway (e.g., "60% of Students Live Within 10 km of Campus").
        - Visualisation: Centrally placed for focus.

- Supporting details (e.g., annotations or captions): Below the chart.

2. **Balanced Spacing:**

   - Ensure the bars have sufficient spacing between them for clarity. Overcrowding can confuse the audience.
   - Leave adequate margins and avoid cluttering the chart with excessive text or visuals.

3. **Minimalism:**

   - Use gridlines sparingly (e.g., horizontal gridlines only) to guide interpretation without overwhelming the visualisation.
   - Avoid excessive decorative elements (e.g., 3D effects or gradients) that distract from the data.

4. **Annotations and Labels:**

   - Place percentage labels **inside or above bars** for immediate interpretation.
   - Add an **arrow or callout** to connect the "60%" insight to the relevant data range (0-5 km + 6-10 km).
   - Avoid overlapping annotations to maintain readability.

Alignment with Purpose and Audience

- **Senior Management Audience:**

  - Senior managers need to extract insights quickly. By using bold colours and a clean layout, the visualisation delivers the main message efficiently.

- **Focus on Key Insight:**

  - Colour emphasis on the 60% within 10 km ensures that the main takeaway is immediately visible.

- **Professional and Clear Design:**

  - A balanced, minimalistic layout with well-chosen colours reflects a polished presentation style, appropriate for decision-makers.

By adhering to these principles, the visualisation effectively communicates the data while catering to the expectations of the audience.

Q2 Data Cleaning [20 marks]

Download the dataset provided in loop and use it to answer the questions below. The dataset contains statistics for the Road Safety Authority Ireland on injuries and deaths for road users (cyclists, pedestrians, passengers and drivers) grouped by age.

Using the data provided, identify three (3) different possible errors or artefacts in the dataset. These errors or artefacts should each likely come from a different cause. Give the column name and cell reference if appropriate.

Answer:

1 Here are three potential errors or artefacts in the dataset:

1. Inconsistent Age Group Labels

- The age group "06-Sep" appears to be a typographical error, likely intended to represent an age range (e.g., "6-9 years" or "6-17 years").
- Correct age group labeling is crucial for accurate analysis and aggregation.

2. Zero Counts for Some Categories

- Certain participant categories, such as "Motorcyclists" and "Car Drivers" for younger age groups (e.g., "0-5"), have consistent zeros for injuries and fatalities. While this may be accurate, it could also be an artefact of incomplete reporting or exclusion of unlikely scenarios, given that such young children are rarely drivers or motorcyclists. These values might warrant closer scrutiny to confirm accuracy.

3. Unusual Trends in Certain Age Groups

- For example, for "0-5 Males," the number of injured "Car Passengers" increases sharply from 54 (2011) to 83 (2012), while the number of fatalities drops from 1 to 0. Similarly, for "21-24 Males," "Car Passengers Killed" jumps from 1 (2011) to 7 (2012). Such drastic changes might indicate reporting anomalies or data recording errors rather than actual trends in road safety.

These observations should be verified against source data to ensure they are not artefacts or errors introduced during data collection or processing

Briefly describe your approach to finding these errors or artefacts including any tools you used.

Answer: To identify potential errors or artifacts in the dataset, I used the following approach:

1. Visual Inspection for Outliers and Inconsistencies:

I started by visually inspecting the dataset for unusual patterns that could indicate potential issues, such as:

- **Unexpected zeros or high values** in columns that should not realistically have such values (e.g., fatalities being zero where injuries were reported).
- **Inconsistent data between related columns**, such as an unexpected change in injury or fatality counts over the years for similar age groups or categories.
- **Duplicate or missing data** entries.

2. Consistency Checking:

I looked at specific columns for consistency over time:

- Age groups: Data from 2011 and 2012 were compared to check for trends that made sense logically (e.g., a category where injuries or fatalities consistently drop to zero in the subsequent year might suggest an issue with the data collection or reporting).
- **Sex and Participant combinations**: Ensured that for each combination (e.g., age group, sex, participant type), the data aligned with expectations (e.g., no fatalities for a type that had zero injuries).

3. Logical Relationships between Casualty Types and Counts:

For example, the number of injuries and fatalities should not have an illogical relationship (e.g., **more fatalities than injuries** in a specific category), which could indicate errors in data entry.

4. Tool Utilization:

- **Manual Cross-Referencing**: I compared and contrasted values in related categories (e.g., injuries vs. fatalities for a given category).
- **Excel Functions/Conditional Formatting**: While I didn't use specific tools here, had I been using a spreadsheet software, I would have used conditional formatting to highlight anomalies (e.g., where injuries = 0 and fatalities > 0) or to compare year-on-year changes.

By focusing on patterns in the data and comparing values across similar rows and columns, I was able to identify likely sources of errors.

Suggest how each error or artefact was most likely introduced and give the phase from the generic data analytics pipeline.

Answer: **Unexpected Zero or High Values in Injury/Fatality Counts**

- **Possible Cause**:
    - **Data Entry Errors**: These anomalies may result from human error during manual data entry. For example, entering “0” for injuries or fatalities when the actual number was not properly recorded.

- **System Errors**: Automated data collection systems might have defaulted to zero when no data was available or if there was a failure to capture the appropriate value.

- **Likely Phase**:

  - **Data Collection**: If the data is collected manually or through an automated system that defaults to zero when no information is available, the error likely occurred here.
  - **Data Entry**: During the phase where raw data is entered into the system, either by individuals or through automated processes, data can be incorrectly entered or processed.

2. Inconsistent Data Between Related Columns (e.g., Age Groups and Fatalities)

- **Possible Cause**:

  - **Mismatched Data Sources**: If data is sourced from multiple systems or departments, inconsistencies may arise when different teams provide data using different reporting formats or criteria.
  - **Data Transformation Errors**: During data cleaning or transformation, values could have been mismatched between related columns (e.g., when reformatting age groups or recalculating totals for fatalities and injuries).

- **Likely Phase**:

  - **Data Integration/Transformation**: During the phase where data from multiple sources is merged, discrepancies might appear if the source data was not aligned correctly.
  - **Data Cleaning**: When cleaning data, there's a chance that relationships between different columns (such as age group, injury, and fatalities) were not appropriately verified.

3. Duplicate or Missing Data Entries

- **Possible Cause**:

  - **Data Entry Duplication**: Duplicate entries could have been introduced when the same data was entered more than once, either by human mistake or system glitches.
  - **Missing Data**: If certain data points (e.g., injuries or fatalities) were not reported or were lost in transit, missing values would appear. Some systems might not flag missing data adequately.

- **Likely Phase**:

  - **Data Entry**: Duplication typically occurs at the point of initial data collection or entry, where a form may have been submitted multiple times, or the same dataset was uploaded multiple times.

- **Data Integration**: Missing data can occur during the process of combining different datasets, where fields are not mapped correctly, or values are omitted in the merging process.

4. Incongruent Relationships Between Injury/Fatality Counts and Participant Types

- **Possible Cause**:
  - **Data Processing or Aggregation Errors**: During data transformation or aggregation, counts for different categories might have been mismatched, such as combining injury counts for a participant type with a fatality count for a different group.
  - **Incorrect Calculations or Summations**: If there's an issue with how the data is summed up or calculated (e.g., injuries being counted more than once), these kinds of inconsistencies may arise.
- **Likely Phase**:
  - **Data Transformation/Analysis**: Errors can happen during the phase when the data is being aggregated, summarized, or calculated, especially if automated scripts or formulas are used without appropriate validation.

5. Unrealistic Data Trends or Outliers (e.g., large drop in one year)

- **Possible Cause**:
  - **Survey/Reporting Errors**: A sudden large drop in data could be due to errors in reporting, where data from a particular year was omitted or incorrectly reported due to technical issues or human mistakes.
  - **Sampling Bias**: If the data collection method was flawed, such as missing a representative sample or excluding certain regions or categories, the trend might not reflect the actual situation.
- **Likely Phase**:
  - **Data Collection**: Errors could originate during data collection when an event was missed, underreported, or incorrectly flagged.
  - **Data Cleaning**: If errors in earlier years (e.g., overcounting or underreporting) were not caught, this could result in a misleading trend in the dataset.

6. Irregular Formatting (e.g., varying naming conventions for participant categories)

- **Possible Cause**:
  - **Human Error**: Inconsistent naming conventions or format changes (e.g., using different terms for the same category) might result from human oversight or different individuals using different standards for categorizing data.
  - **Data Standardization Issues**: During the data collection or cleaning phases, the lack of a clear and consistent schema could cause data points to be categorized differently.

- **Likely Phase**:
    - **Data Standardization**: This typically happens during the phase when data from different sources is harmonized into a standard format. If there's a lack of clear guidelines for naming conventions or classification, errors can arise.

Summary of Phases in the Data Pipeline:

- **Data Collection**: Errors introduced here typically include missing data, incorrect entries, or misreported data.
- **Data Entry**: Duplicates and human errors are most commonly introduced during this phase.
- **Data Integration/Transformation**: Mismatched columns, inconsistent relationships, and aggregation errors are likely introduced here.
- **Data Cleaning**: This phase may introduce or fail to correct errors related to missing data or mismatched categories if not carefully checked.
- **Data Analysis**: Errors can arise from incorrect calculations, improper data aggregation, or failure to validate trends properly.

Each of these phases is critical in ensuring high data quality, and most errors can be traced back to the challenges inherent in any stage of the data pipeline.

Pick one of the errors or artefacts you identified. What data quality methods could be used to avoid or reduce the probability of this specific error/artefact occurring?

Answer:

Q3 Data Project Analysis (Storage) [20 marks]

> "At the request of the Irish Government, you are preparing a report on the impact of COVID-19 restrictions on working and commuting behaviour in Ireland during 2020 comparing it to surveys and records from 2019 and 2009. You have data available from data.gov.ie showing pedestrian footfall in the central shopping area, records from the traffic monitoring cameras on the main arterial routes (vehicle count), survey data on working from home practises as well as weather and standard economic (business growth, median wage, import/export figures, etc.) information. You have permission to conduct further consumer surveys as required. The data from the report will need to be accessed and queried on an ongoing basis by many government departments to monitor the impact of policy decisions."

List three (3) important questions you would ask your client.

Answer:

Here are three important questions to ask your client in order to gather the necessary context and ensure the report meets the client's expectations:

1. What specific outcomes or key performance indicators (KPIs) are you looking to assess in the impact of COVID-19 restrictions?

- This question helps clarify what the client sees as the most important factors to measure in the report, such as the change in pedestrian footfall, vehicle counts, working from home practices, or economic indicators. Understanding the desired outcomes will guide the analysis and ensure that the data is interpreted and presented in a way that aligns with the government's objectives.

2. What are the specific timeframes or periods of interest for the analysis, and should we account for different phases of the restrictions (e.g., lockdown, reopening, etc.)?

- The impact of COVID-19 restrictions on working and commuting behavior likely varies across different stages of the pandemic, such as the initial lockdown, phased reopenings, and later restrictions or relaxations. Knowing whether the client wants to compare specific months, quarters, or phases of the pandemic with the baseline years (2019 and 2009) will help structure the report accurately.

3. How would you like the findings to be updated or made accessible to other government departments on an ongoing basis (e.g., real-time dashboards, monthly reports)?

- Since the data will need to be accessed and queried by multiple government departments on an ongoing basis, it's crucial to understand how they want the findings to be presented and updated over time. This will inform decisions about data storage, visualization, and whether a dynamic system like a dashboard or periodic report updates will be required.

Describe the data sources and/or specific file formats that you are likely to use in collecting and storing the data for this project.

Answer: For this project, the data sources and specific file formats you would likely use to collect and store the data can be categorized as follows:

1. Pedestrian Footfall Data (Central Shopping Area)

- **Data Source:** Data.gov.ie or a similar public data repository (government-provided data or data from local councils).
- **File Format(s):**
  - **CSV (Comma-Separated Values)** or **XLSX (Excel files)**: These formats are commonly used for time series data like pedestrian footfall counts, which might include timestamps and daily/hourly footfall data.
  - **JSON**: If the footfall data is provided through an API, it may be in JSON format.

Pedestrian data would include timestamps and location identifiers (e.g., specific zones in the shopping area) that will be critical for analyzing trends.

2. Traffic Monitoring Camera Data (Vehicle Count on Arterial Routes)

- **Data Source:** Traffic monitoring systems or data provided by transportation departments, often accessible through public datasets.
- **File Format(s):**
  - **CSV** or **JSON**: Traffic counts might be provided in CSV format for each camera along with timestamps and vehicle count data for different types of vehicles (e.g., cars, trucks, buses).
  - **SQL Database**: If the data is coming from a real-time system or larger historical dataset, it may be stored in a relational database like SQL or a cloud-based system.

The vehicle count data could be used to correlate changes in commuting behavior with the overall traffic patterns across different periods.

3. Survey Data on Working from Home Practices

- **Data Source:** Custom surveys (either existing or designed specifically for this project) that collect information on working-from-home trends, employer policies, employee preferences, etc.
- **File Format(s):**
  - **CSV/XLSX**: Survey data is typically stored in spreadsheet formats such as CSV or Excel, with rows representing individual responses and columns for different questions and demographic details.
  - **Google Forms/Survey Monkey Data Export**: If surveys were administered using an online tool, the exported data could be in CSV, XLSX, or JSON formats.

This data would contain structured responses from individuals or businesses regarding their remote work arrangements.

4. Weather Data (Impact of Weather on Commuting and Behavior)

- **Data Source:** Meteorological data from government sources like Met Éireann, which tracks weather patterns, or global data providers (e.g., the European Centre for Medium-Range Weather Forecasts).
- **File Format(s):**
  - **CSV** or **NetCDF**: Weather data could be provided in CSV format with daily metrics (temperature, precipitation, wind speed) or in more complex formats like NetCDF for detailed atmospheric data.
  - **API Access**: For real-time data access, weather information might come through an API in JSON or XML format.

Weather data could be crucial to identify seasonal changes or correlate weather events with commuting and working behavior.

5. Standard Economic Data (Business Growth, Median Wage, Import/Export Figures, etc.)

- **Data Source:** National statistics agencies such as the Central Statistics Office (CSO) of Ireland, OECD, or Eurostat.
- **File Format(s):**
  - **CSV**, **XLSX**, or **JSON**: Economic datasets typically come in CSV or XLSX format, with rows for different time periods (monthly, quarterly, yearly) and columns representing various economic indicators.
  - **API Access**: If real-time data is required, economic data may also be accessed via APIs, returning JSON or XML files.

This data would help contextualize changes in commuting patterns with broader economic conditions like wage growth, unemployment, and business performance.

6. Consumer Survey Data (Additional Surveys)

- **Data Source:** Consumer surveys designed specifically for this project, including questions about commuting preferences, changes in work habits, and attitudes toward COVID-19 restrictions.
- **File Format(s):**
  - **CSV/XLSX**: Survey responses would likely be stored in CSV or Excel format for analysis, where each row represents an individual respondent and columns represent their responses to specific questions.
  - **Google Forms or Survey Tool Exports**: Data could also be stored in platforms like Google Forms or SurveyMonkey, with export options to CSV, XLSX, or JSON.

7. Data Access and Ongoing Monitoring

- **Data Source:** A central database that combines all data sources for real-time querying and ongoing access by multiple departments.
- **File Format(s):**
    - **SQL/NoSQL Databases**: Given that multiple departments will need to query the data continuously, a centralized relational or NoSQL database (such as PostgreSQL, MySQL, MongoDB) could be used for structured and semi-structured data storage. This setup will allow efficient querying, aggregation, and reporting.
    - **Cloud-Based Storage**: Data can also be stored in cloud platforms like AWS, Google Cloud, or Microsoft Azure, which support both relational and non-relational data models.

Data Storage and Management Considerations:

- **Data Quality and Consistency**: It is important to ensure that data from different sources is standardized (e.g., same time zone, consistent formats) to allow meaningful comparisons.
- **Data Privacy**: For consumer survey data and any personally identifiable information, appropriate anonymization and consent management procedures will need to be in place to comply with GDPR or other privacy regulations.
- **Data Refresh and Accessibility**: For ongoing monitoring, setting up automated data pipelines for regular updates (e.g., daily, weekly) will be essential. This could be achieved through APIs or scheduled data imports.

In summary, the project would likely involve collecting data in CSV, XLSX, and JSON formats from a variety of government and survey sources, while centralizing and storing the data in databases for easy access, analysis, and ongoing monitoring.

Suggest a type of database storage approach to use for this project, giving a reason for your choice and stating any assumptions you make.

Answer: For this project, I recommend using a **relational database management system (RDBMS)** such as **PostgreSQL** or **MySQL**. This choice is based on the need for structured data storage, the ability to handle complex queries across multiple data sources, and the ability to ensure data integrity and consistency. Here's a detailed explanation:

Reasons for Choosing an RDBMS:

1. **Structured Data**:

- The data in this project is highly structured. It includes records such as pedestrian footfall, vehicle counts, survey responses, and economic indicators, which typically follow a consistent schema with defined columns (e.g., timestamps, locations, counts, or responses).
- RDBMS platforms like PostgreSQL and MySQL are designed to handle structured data effectively.

2. **Complex Queries**:

- The project will require running **complex queries** across multiple datasets. For example, combining pedestrian footfall data with weather data, or analyzing how economic indicators correlate with commuting patterns.
- SQL provides powerful querying capabilities for performing joins, aggregations, filtering, and time-series analysis, which are crucial for this type of project.

3. **Data Integrity and Relationships**:

- This project involves different types of related data, such as survey data, traffic data, and economic indicators. An RDBMS can enforce **referential integrity** through primary and foreign keys, ensuring that data across different tables (e.g., traffic data, economic indicators) is related correctly.
- The ability to enforce constraints and consistency checks (e.g., ensuring timestamps are consistent or that vehicle counts are not negative) will ensure the quality of data.

4. **Scalability and Performance**:

- RDBMS solutions like PostgreSQL are highly scalable and support large datasets. Given the volume of data generated (e.g., daily traffic counts, weather data, and survey responses), PostgreSQL can efficiently handle large amounts of data without compromising performance.
- **Indexes** and **partitioning** can be used to optimize query performance, especially for time-series data (e.g., pedestrian footfall over time).

5. **Ongoing Access and Monitoring**:

- The need for ongoing access by government departments implies the need for a centralized, easily accessible database that can be queried via **SQL** by different stakeholders.
- RDBMS solutions also support the creation of **views, stored procedures**, and **APIs** for providing specialized access to different users, helping ensure that departments get relevant data quickly without needing to directly interact with the full dataset.

6. **Security and Compliance**:

- RDBMS platforms offer advanced security features like **role-based access control (RBAC)**, which can help manage who can access and modify specific types of data.

This is important for ensuring the privacy and security of sensitive data, especially if consumer survey data is involved.

- Additionally, these systems offer auditing and logging capabilities, which are important for complying with regulations like GDPR.

Assumptions:

1. **Data Size**:
    - I assume that the dataset size, though potentially large (e.g., daily counts, weather data over months or years), will not be so large as to require a NoSQL solution (like MongoDB). If the data grows to a massive size, horizontal scalability could be achieved through clustering or database replication.
2. **Data Update Frequency**:
    - I assume that the data is regularly updated (e.g., daily updates for weather or traffic data, weekly for economic indicators). RDBMS systems can handle frequent updates, but if the project involves near-real-time data processing (e.g., traffic cameras or live weather data), additional tools like **ETL pipelines** or **data streaming systems** (e.g., Apache Kafka) may be required.
3. **Complexity of Data**:
    - I assume that most data sources (footfall, traffic, surveys, weather) will follow standard tabular structures and will not involve highly unstructured data (like free-form text). If large amounts of unstructured data (e.g., social media data or free-form survey responses) need to be processed, a NoSQL solution may be required, but for the current project, an RDBMS is likely sufficient.

Example Database Design:

Here is a possible schema for the relational database:

- **Pedestrian_Footfall**: Stores pedestrian counts, timestamps, and location IDs.
    - Columns: location_id, timestamp, footfall_count
- **Traffic_Counts**: Stores vehicle counts from cameras, timestamps, and camera IDs.
    - Columns: camera_id, timestamp, vehicle_count
- **Surveys**: Stores survey responses related to working from home practices.
    - Columns: survey_id, respondent_id, response_date, working_from_home_status, business_type, etc.
- **Weather_Data**: Stores weather data for each day or hour.
    - Columns: timestamp, location_id, temperature, precipitation, etc.
- **Economic_Indicators**: Stores economic data such as median wage or business growth.

- Columns: timestamp, economic_indicator_type, value

- **Location_Information**: Contains metadata about locations (e.g., shopping zones, traffic camera locations).
  - Columns: location_id, name, latitude, longitude

By using foreign keys, you can link different data sources (e.g., linking pedestrian footfall data with location information or economic data to dates), ensuring efficient data analysis across various sources.

Conclusion:

A relational database such as PostgreSQL or MySQL is the ideal storage solution for this project due to the structured nature of the data, the need for complex queries, and the ability to ensure data integrity and consistency. It also offers the scalability and performance needed to handle the growing datasets as the project progresses.

For this project, can you identify any possible risks in terms of data privacy or GDPR requirements? What data items cause this risk?

Answer: In this project, there are several possible risks related to **data privacy** and **GDPR** (General Data Protection Regulation) compliance. GDPR requires that personal data be processed securely, transparently, and for specified purposes. The potential risks mainly arise from how the data is collected, stored, and used, especially when it involves sensitive or personally identifiable information (PII). Below are the risks and relevant data items that could cause them:

1. Personal Identifiable Information (PII)

- **Risk**: If the data contains **personally identifiable information (PII)**, such as names, email addresses, phone numbers, or any data that can directly identify individuals, there is a risk of violating GDPR's **data minimization** and **purpose limitation** principles. GDPR requires that personal data is only collected for legitimate purposes and processed in a way that is secure.
- **Data Items at Risk**:
  - **Survey responses**: If the surveys are designed to collect PII (e.g., by asking respondents to provide names, addresses, or contact details), there is a risk. Even if these data are anonymized, survey responses that include information about the individual's work-from-home practices or economic status could indirectly identify them, particularly if a small number of respondents belong to identifiable groups.
  - **Traffic data**: In some cases, traffic monitoring or pedestrian footfall data could be linked back to individuals if the data is too granular (e.g., capturing license plate numbers, GPS tracking, or using cameras with facial recognition).

2. Data Anonymization and Pseudonymization

- **Risk**: If personal data is not properly anonymized or pseudonymized, there's a risk of re-identification. Even if individuals' identities are not directly collected, the data might still be considered personal if there is the possibility to identify individuals from the dataset when combined with other data sources.
- **Data Items at Risk**:
    - **Traffic monitoring cameras**: If the camera footage or vehicle count data is linked with vehicle license plates, it could be possible to identify individuals, violating privacy. This could be a risk, especially if this data is not properly anonymized or pseudonymized before being stored and analyzed.
    - **Survey data**: If survey responses include demographic information (e.g., age, occupation) and if the sample size is small or certain demographic groups are highly identifiable, there's a risk of re-identifying individuals, even without direct identifiers.

3. Data Retention and Storage

- **Risk**: GDPR requires that personal data not be kept longer than necessary for the purposes it was collected. Data retention policies need to be in place to avoid keeping unnecessary data for long periods, especially if the data is sensitive or personal.
- **Data Items at Risk**:
    - **Survey data**: If personal details are collected as part of the survey, the retention of this data beyond the required time for the analysis could violate GDPR's **storage limitation** requirement.
    - **Pedestrian and traffic monitoring data**: Retention of traffic data, especially if it includes timestamps and location identifiers, could violate data retention rules if it is kept for longer than necessary.

4. Data Security

- **Risk**: GDPR mandates that personal data be protected through **appropriate security measures**. Any breaches or unauthorized access to personal data could result in significant risks, including fines and reputational damage.
- **Data Items at Risk**:
    - **Survey responses**: If survey data contains PII or is identifiable in any way, such as responses tied to specific individuals, it needs to be securely stored and protected from unauthorized access.
    - **Traffic and pedestrian data**: This type of data can be aggregated to derive trends and behaviors, but if not adequately protected, the underlying data could be exposed to unauthorized access, especially if stored in plaintext or without encryption.

5. Informed Consent

- **Risk**: GDPR requires that individuals provide **informed consent** before their personal data is collected, processed, or stored. If data is being collected from survey participants or via traffic monitoring cameras, consent needs to be explicit and documented.
- **Data Items at Risk**:
    - **Survey data**: Participants must be informed about the purpose of the survey, how their data will be used, and how long it will be stored. If the survey collects PII or sensitive data (such as health-related questions about working from home during COVID-19), consent must be obtained in a clear and transparent manner.
    - **Traffic monitoring cameras**: If these cameras capture faces, license plates, or other identifiable information, explicit consent must be obtained or a legitimate interest identified for using this data.

6. Third-Party Data Sharing and Data Transfers

- **Risk**: If the data is shared with third parties (e.g., for analysis, storage, or processing), or if data is transferred outside of the EU (e.g., through cloud storage), this may violate GDPR's rules regarding **data sharing and cross-border transfers**. If third parties handle personal data without adequate safeguards, it could lead to non-compliance with GDPR.
- **Data Items at Risk**:
    - **Survey data**: If survey data is being processed by third-party service providers (e.g., for analysis or reporting), there should be adequate **data processing agreements** in place to ensure compliance with GDPR.
    - **Traffic and footfall data**: If third parties process traffic or footfall data that contains any identifiable information or is aggregated in a way that could allow re-identification, data sharing agreements must specify how the data will be handled and ensure compliance with GDPR requirements.

7. Profiling and Automated Decision-Making

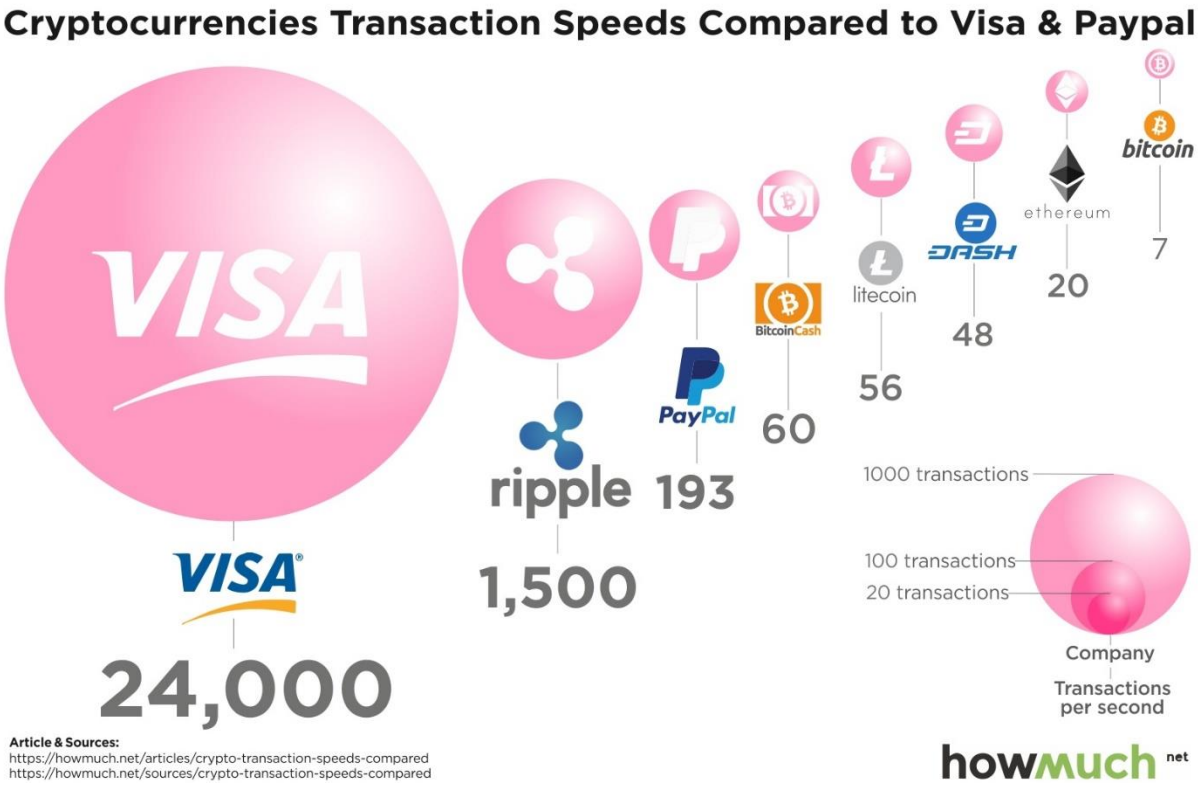
- **Risk**: If the data collected (e.g., working-from-home behavior, economic data, traffic patterns) is used for **profiling** or automated decision-making that significantly affects individuals (e.g., policies on traffic congestion or work-from-home incentives), this could raise concerns under GDPR. Profiling that involves personal data requires individuals to be informed about the logic involved, and it should not lead to decisions that have a significant impact on individuals without their explicit consent.
- **Data Items at Risk**:
    - **Survey data**: If survey responses are used to create profiles of individuals or groups (e.g., predicting work-from-home behavior), or if decisions are made based on aggregated survey responses, there could be privacy concerns. For example, decisions regarding traffic routes or working-from-home policies could affect individuals without their informed consent.

Mitigating These Risks:

1. **Anonymization and Pseudonymization**: Ensure that any personal data (e.g., survey responses) is anonymized or pseudonymized where possible to reduce risks of re-identification.
2. **Explicit Consent**: Obtain clear, informed consent from all survey participants, especially if their data is used for analysis or shared with third parties.
3. **Data Retention**: Implement strict data retention policies, ensuring that data is not kept longer than necessary for the stated purposes.
4. **Encryption and Access Control**: Use encryption for storing sensitive data and implement strict access control policies to ensure that only authorized personnel can access personal data.
5. **Third-Party Contracts**: Ensure data processing agreements are in place with any third-party vendors to ensure compliance with GDPR.
6. **Transparency**: Maintain transparency with all stakeholders about how their data is being used, processed, and stored, especially regarding surveys and traffic data.

By addressing these privacy and GDPR risks, you can ensure that the project complies with legal and ethical requirements while minimizing any potential negative impacts on individuals' privacy.

Q4 Data Visualisation Critique [20 marks]



For the visualisation above, identify the data encoding methods (marks and attributes) and critique the visualisation design according to the principles discussed in CA682 – consider how colour, layout, Gestalt etc. have been applied.

Suggest specific improvements you could make to the visualisation's effectiveness.

Answer: **Critique of the Data Visualization**

1. Data Encoding Methods (Marks and Attributes)

- **Marks**: The key visual elements used in this visualization are **circles** (bubbles) to represent companies and their transaction speeds.
- **Attributes**:
  - **Size**: The size of the circles encodes the transaction speed (e.g., Visa has the largest circle at 24,000 transactions per second).
  - **Position**: Companies are aligned horizontally to show comparisons between them.
  - **Color**: Each bubble has a unique color for visual distinction, but no deeper meaning is encoded in the colors.
  - **Text**: Numerical values for transaction speeds and company logos are added inside or near the bubbles for additional clarity.

2. Critique Based on Design Principles

1. **Gestalt Principles**:
    - **Proximity**: The visualization places circles close together to group related entities (e.g., cryptocurrencies vs. traditional financial systems). However, the proximity between Visa and Ripple could suggest they are closely comparable, which might be misleading given the vast difference in transaction speeds.
    - **Similarity**: Circles are consistently shaped and positioned along a single line, which helps maintain visual consistency. However, the similar sizes of smaller bubbles (e.g., Bitcoin Cash, Litecoin) make them harder to distinguish.
    - **Hierarchy**: The largest bubble (Visa) clearly dominates, establishing it as the focus. However, Ripple's bubble is also large, which could confuse its comparative importance.
2. **Color**:
    - While the colors are distinct and vibrant, they do not encode any additional meaning (e.g., traditional finance vs. cryptocurrency). This misses an opportunity to add context.
3. **Layout**:
    - The horizontal alignment is effective for comparison, but the overlap of the smaller bubbles creates clutter, especially on the right side (e.g., Bitcoin, Ethereum).
    - Ripple and PayPal are positioned relatively far from their textual annotations, leading to slight ambiguity.
4. **Legibility**:
    - The numeric labels are clear, but the size and overlapping of smaller bubbles make them harder to interpret.
    - Visa's massive bubble dominates to the point that smaller ones feel insignificant, diminishing the impact of the visualization.
5. **Scalability**:
    - The use of circle size to encode transaction speed creates distortion, as the human eye doesn't perceive area proportionally. For example, Ripple (1,500) appears closer to Visa (24,000) than it actually is.

3. Suggested Improvements

1. **Improve Proportional Representation**:

- Use a bar chart or logarithmic scaling for transaction speed instead of bubble sizes to make differences between entities more interpretable.

2. **Reduce Clutter**:

   - Space out smaller bubbles more evenly to avoid overlapping.
   - Add gridlines or a scale to guide the reader in comparing sizes visually.

3. **Add Meaning to Colors**:

   - Use color to categorize entities (e.g., one color for cryptocurrencies and another for traditional finance systems).

4. **Refine Annotations**:

   - Place labels and numeric values consistently close to their corresponding bubbles to improve clarity.

5. **Emphasize Smaller Entities**:

   - Use a more nuanced design (e.g., varied positioning or additional annotations) to highlight smaller entities like Bitcoin or Ethereum without overshadowing them by Visa's dominance.

Final Thoughts

The visualization is visually appealing but struggles with clarity and accurate perception of data due to size distortions and layout clutter. By addressing these issues, the chart could better serve its purpose of comparing transaction speeds across entities.