# 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: explanatory visualisation task

Who is the intended audience for the data visualisation?

Answer: senior university management

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

Answer: Answer: "Geographical Distribution of DCU Students and Accommodation Trends"

Reasoning: The title should clearly communicate the focus of the visualisation, which is the distribution of students by distance from the university campus and key accommodation trends (e.g., rent prices). The assumption here is that students living closer to the campus have a higher proportion of lower rent costs, and the title reflects both the geographical aspect and the rent trend.

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: Answer: A choropleth map(色阶图) or symbol map would be the best choice for this visualisation.

Justification:

Geographical Distribution: Since the task is about showing the geographical distribution of students by distance from the campus, a map visualisation allows the audience to quickly grasp where the majority of students live.

Percentage & Rent Information: On the map, you can use different colours (or size of symbols) to represent key information, such as the percentage of students within each distance range and average rents.

Clear Messaging: This type of visualisation effectively communicates spatial relationships and distribution patterns, which are essential to answering the client’s query.

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: I would use geospatial marks (such as points or polygons) to represent the distances (0-5 km, 6-10 km, etc.) from DCU. The marks could be placed on the map to show the areas where students are located.

Attributes:

Colour: The colour scale could represent the percentage of students in each distance range, for instance, darker shades for larger percentages (e.g., darker green for the 0-5 km range which has 35% of students).

Size: The size of the markers could reflect the median monthly rent, with larger symbols indicating higher rents (e.g., a larger circle for the 6-10 km range with a €850 median rent).

Shape: Different shapes could be used to represent different ranges of students (e.g., circles for 0-5 km, squares for 6-10 km, etc.), but this would need to be simple and not overly cluttered.

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

Answer: Colour:

Purpose: Colour should help to distinguish different distance bands (e.g., 0-5 km, 6-10 km, etc.). It can also highlight key trends, such as the higher concentration of students within a 10 km radius of the campus.

Principles: Use contrasting colours (极为不同的颜色)to make the visualisation easy to interpret. For instance, use a gradient scale(梯度分级) for the percentage of students (e.g., from light to dark green) and a different colour scale for rent (e.g., a gradient from light blue for lower rents to dark blue for higher rents).

Layout:

Purpose: A clear, uncluttered map(整齐的地图) layout will ensure that the information is accessible and easy to read by senior management.

Principles: Organise the map so that it highlights the most important data (60% of students within 10 km) prominently. It should avoid unnecessary elements that could distract the viewer from the key message.

Legibility: Make sure all text (e.g., distance bands, rent, percentages) is large enough to be legible, with a clear legend explaining the colours and sizes used in the map. Keep a balance between aesthetics and clarity.

Focus on Key Insights: Highlight the area where 60% of students live (within 10 km), perhaps by using a border or slightly larger symbols for this area. This helps senior management focus on the core message.

### 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

2

3

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

Answer:

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

Answer:

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:

What specific insights or policy decisions do you intend to draw from this analysis?

(Understanding the goals helps tailor data collection and analysis to meet their objectives.)

How frequently will the data need to be updated and accessed by government departments?

(This informs the database design and infrastructure for ongoing access and scalability.)

Are there any existing data integration standards or platforms used by government departments that we should align with?

(Ensures compatibility with current systems, simplifying deployment and access.)

Data Granularity:

What level of detail do you require for the data? For example, do you need hourly, daily, or weekly pedestrian footfall data?

Historical Data Availability:

Do you have comprehensive historical data on working and commuting behavior for 2019 and 2009? How complete and consistent is this data?

Future Data Requirements:

Beyond monitoring the impact of policy decisions, do you have any additional specific use cases? For instance, will the data be used to predict future trends or identify potential issues?

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: Data Sources and File Formats:

data.gov.ie:

Likely provides data in various formats such as CSV, JSON, or XML.

Traffic Monitoring Camera Records:

These may be video files or aggregated data stored in proprietary formats, databases, or binary files.

Survey Data on Working Practices:

Typically stored in spreadsheets (e.g., XLS or XLSX) or CSV files.

Weather and Economic Data:

Obtainable from various sources, often available via APIs, CSV downloads, or databases.

Consumer Surveys:

Data can be collected using online survey tools or paper forms and stored in spreadsheets or databases.

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: Database Storage Approach:

Recommended Database: Relational Database (e.g., PostgreSQL or MySQL)

Reason:

Government departments need continuous access and querying capabilities to monitor policy impacts.

Relational databases support efficient data retrieval and complex queries through SQL.

They ensure data integrity and consistency, which is crucial for storing sensitive information.

Assumptions:

Most data is structured and can be easily mapped into tables within a relational database.

The data volume is manageable within standard government infrastructure.

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

Answer: Data Privacy and GDPR Risks:

Potential Risks:

Consumer Survey Data:

May contain Personally Identifiable Information (PII) such as names, addresses, or contact details.

Traffic Monitoring Camera Records:

Video recordings could include facial images or vehicle license plates, which can be used to identify individuals.

Mitigation Measures:

Anonymization or Pseudonymization: Remove or replace PII in datasets.

Data Encryption: Protect data at rest and in transit using encryption technologies.

Access Controls: Implement strict access control measures to limit data access.

Data Retention Policy: Retain data only as long as necessary and securely delete it when no longer needed.

Before implementing the project, consult a Data Protection Officer (DPO) to ensure full compliance with GDPR requirements.

### Q4 Data Visualisation Critique [20 marks]

Chart

Description automatically generated with medium confidence

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: