

**AUGUST/RESIT EXAMINATIONS 2018/2019**

**MODULE:** CA682A - Data Management and Visualisation

**PROGRAMME(S):**

MCM M.Sc. in Computing **YEAR OF STUDY:** 2

|  |  |  |
| --- | --- | --- |
| **EXAMINER(S):** |  |  |
| Suzanne Little | (Internal) | (Ext:6052) |
| Dr. Brian MacNamee | (External) | External |
| Dr. Dan Emanoil Grigoras | (External) | External |

**TIME ALLOWED:** 3 Hours

**INSTRUCTIONS:** Answer **four (4)** questions from the 5 available.

**PLEASE DO NOT TURN OVER THIS PAGE UNTIL YOU ARE INSTRUCTED TO DO SO.**

The use of programmable or text storing calculators is expressly forbidden.

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.***

CA682A - Data Management and Visualisation

August/Resit Examinations 2018/2019 Page 1 of 8

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

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

Given the following brief to design a system for a data collection task, list three (3) important questions you would ask your client and suggest a type of database

system to use, giving a reason for your choice.

“We are collecting data to use for a marketing campaign by DCU to increase public transport use when travelling to campus. Data sources include public surveys, records from Transport Ireland and information from DCU estates.”

## Questions for the Client and Database Choice

Given the brief, here are three important questions to ask the client:

* **What specific information do you need to collect from public surveys, Transport Ireland records, and DCU estates to inform the marketing campaign?** This question aims to clarify the data requirements, which will inform the database design. Understanding the specific data points needed, such as demographics, travel patterns, and preferences, will help determine the appropriate data types and structures for the database.
* **What is the expected volume of data from each source, and what is the anticipated rate of data growth?** This is crucial for choosing a database system that can handle the volume and velocity of the data. If the volume is large and the growth rate is rapid, a database system designed for handling big data, like a NoSQL database, might be more suitable.
* **What kind of analysis do you plan to perform on the collected data?** Understanding the analysis needs will help determine the database system's capabilities. For example, if complex queries and calculations are required, a relational database management system (RDBMS) with strong analytical capabilities would be appropriate.

**Suggested Database System**: A relational database system (RDBMS) is recommended for this project. Here's why:

* **Structured Data**: The sources of data (public surveys, Transport Ireland records, and DCU estates information) are likely to be structured. RDBMS excels at handling structured data, organizing it into tables with defined relationships.
* **Data Integrity**: RDBMS provides mechanisms for enforcing data integrity, ensuring data accuracy and consistency. This is crucial for making reliable decisions based on the collected data. For example, constraints can be set to ensure that data entries are within specific ranges or adhere to predefined formats.
* **Querying and Analysis**: RDBMS allows for complex querying and analysis using SQL (Structured Query Language). This is essential for extracting meaningful insights from the data and making informed decisions about the marketing campaign.

However, depending on the answers to the questions posed to the client, particularly regarding data volume and growth, a NoSQL database might be a better choice if the data becomes too large or unstructured for an RDBMS to handle efficiently.

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

(i) Why is it useful to categorise data?

(ii) You have data from the last 5 track and field meets run by Athletics Ireland. Identify some different category descriptions for the following pieces of data:

A. Athlete profile

B. List of races

C. Gold, Silver and Bronze winners

D. Times of the winners from the sprint races

### The Benefits of Data Categorisation

(i) **Categorising data is useful for several reasons:**

* **Organisation and Structure:** Categorisation helps bring order to potentially chaotic datasets, making it easier to manage, store, and retrieve information. By grouping similar data points, we can structure it in a meaningful way that facilitates efficient analysis and interpretation.
* **Meaningful Analysis:** Different categories often lend themselves to different types of analysis. Categorising data allows us to apply appropriate statistical methods and data visualisation techniques based on the nature of the information. For instance, categorical data might be best represented with bar charts, while numerical data could be visualised using histograms or scatterplots.
* **Pattern Identification:** By categorising data, we can more readily identify patterns and trends that might otherwise be obscured within a large, unorganised dataset. For example, grouping athletes by age or event type can reveal performance patterns related to those categories.
* **Effective Communication:** Well-categorised data supports clear and concise communication of findings. By presenting information in a structured and logical manner, we enhance the audience's comprehension and facilitate informed decision-making.

(ii) **Athletics Ireland Data Categories**

Here are some different category descriptions for data from Athletics Ireland track and field meets, keeping in mind the importance of using area rather than diameter to represent quantities, as discussed in our conversation history:

**A. Athlete Profile**

* **Demographics:** Age, Gender, Location
* **Performance History:** Personal best times, Event specialisation, Previous competition results
* **Club Affiliation:** Club name, Coach
* **Physical Attributes:** Height, Weight (can be represented by area of a circle in a visualisation)

**B. List of Races**

* **Event Type:** Sprints, Middle Distance, Long Distance, Field Events (e.g., Javelin, High Jump)
* **Distance/Discipline:** 100m, 200m, 800m, etc.
* **Age Group:** Junior, Senior, Masters
* **Gender:** Men's, Women's

**C. Gold, Silver and Bronze Winners**

* **Athlete Name and ID**
* **Event:** Specific race or field event
* **Medal:** Gold, Silver, Bronze
* **Winning Time/Distance/Height:** Depending on the event

**D. Times of the Winners from the Sprint Races**

* **Event:** 100m, 200m, 400m
* **Athlete Name and ID**
* **Winning Time:** Measured in seconds
* **Date and Location** of the meet

These categories provide a structured way to organise and analyse the Athletics Ireland data, facilitating insights into athlete performance, event trends, and potential areas for improvement. Remember that these are just examples, and you might choose different categories or subcategories depending on your specific analytical goals.

**Q 1(c) [6 Marks]**

Give two (2) advantages & two (2) disadvantages of non-relational databases and give an example of a when a non-relational database would be useful.

Here are two advantages and two disadvantages of non-relational databases, as well as an example of when a non-relational database would be useful:

### Non-Relational Database Advantages and Disadvantages

* **Advantages**:
  + **Large volumes of data**: NoSQL databases are designed to handle large volumes of structured, semi-structured, and unstructured data. This makes them well-suited for applications that generate massive amounts of data, such as social media platforms, e-commerce websites, and sensor networks.
  + **Efficient, scale-out architecture**: They use a scale-out architecture, which means they can be easily scaled horizontally by adding more servers to the system. This is in contrast to traditional relational databases, which often require expensive, monolithic hardware upgrades to scale.
* **Disadvantages**:
  + **Lack of standardization**: There are many different types of NoSQL databases, each with its own strengths and weaknesses. This lack of standardization can make it difficult to choose the right database for a particular application, and can also make it difficult to migrate data from one NoSQL database to another.
  + **Limited querying capabilities**: They typically do not support the same rich querying capabilities as relational databases, such as SQL. This can make it more difficult to perform complex data analysis.

### When to Use a Non-Relational Database

A non-relational database would be useful for:

* **A social media platform** that needs to store large amounts of user data, such as posts, comments, and likes. For example, a social media platform that needs to store large amounts of user data, such as posts, comments, and likes. The data is often semi-structured or unstructured and changes frequently, making it a good fit for a document-oriented NoSQL database.

The sources provide a general overview of different data storage approaches but do not offer specific real-world use cases. The social media platform example is not from the sources and is a common use case for non-relational databases. You may want to verify this information independently.

**Q 1(d) [5 Marks]**

Given a generic data analytics pipeline – Gathering, Processing, Analysing,

Presenting, Preserving – describe in 1 or 2 sentences the activities that can occur at each stage.

Here is a description of the activities that can occur at each stage of a generic data analytics pipeline:

* **Gathering**: This stage involves **collecting data from various sources**, such as files, databases, surveys and sensors. Tools like spreadsheets or Python can be used for this purpose.
* **Processing**: In this stage, the **collected data is cleaned**, transformed and prepared for analysis. Tools like OpenRefine or Python scripts can be used for data cleaning.
* **Analysing**: This stage focuses on **extracting insights from the processed data using statistical methods or machine learning algorithms**. This often involves using specific tools like Python libraries or statistical software.
* **Presenting**: The insights derived from the analysis are then **presented using data visualisations, reports, or dashboards** to effectively communicate findings. Tools like Tableau, spreadsheets, or Python visualisation libraries can be used.
* **Preserving**: This final stage involves **storing and managing the data for future use and ensuring data integrity**. This can involve using databases, data warehouses, or cloud storage solutions.

It's important to note that this is a generic pipeline, and the specific activities and tools used at each stage will vary depending on the project and the data being analyzed.

***[End of Question1]***

CA682A – Data Management and Visualisation

August/Resit Examinations 2018/2019 Page 2 of 8

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

**Q 2(a) [8 Marks]**

**(i)** What is metadata?

(ii) Name three (3) things that metadata is used for. (iii) Give examples of different types of metadata.

### Understanding Metadata

**(i) What is metadata?** Metadata is information that describes and defines the content of data, making it easier to find and utilize.

**(ii) Name three (3) things that metadata is used for.** Metadata serves several purposes, including:

* + **Finding, Locating, Identifying, Selecting, Obtaining, and Navigating data.**
  + **Using data effectively.**
  + **Managing rights and data.**

**(iii) Give examples of different types of metadata.** The sources highlight three primary types of metadata:

* + **Descriptive metadata**: This type of metadata focuses on what the information object is about. It encompasses inherently intrinsic properties. An example is a book's title, author, or subject keywords.
  + **Administrative metadata**: This type of metadata deals with the who, what, why, and where of an object's creation and management. It primarily concerns extrinsic properties. Examples include information about file creation date, file format, or access rights.
  + **Structural metadata**: This type of metadata provides information about the structure, format, and composition of the data being described. It can be either intrinsic or extrinsic. An example is the arrangement of chapters in a book or the schema of a database.

**Q 2(b) [7 Marks]**

Give simple example metadata describing the module CA682. Why would a standard be useful for this type of metadata and identify one (1) problem with enforcing a

standard.

### Simple Metadata for CA682 and the Use of Standards

Here are some simple examples of metadata that could describe the CA682 module:

* **Title:** CA682 - Data Management and Visualization
* **Description:** This module covers the fundamentals of data management and visualization, including data collection, cleaning, analysis, and presentation. Students will learn how to work with large datasets, apply data quality measures, and create effective data visualizations.
* **Keywords:** Data Management, Data Visualization, Data Analysis, Big Data, Databases, Python, Tableau, OpenRefine.
* **Module Coordinator:** Dr. Suzanne Little

**Why Use a Standard?**

A metadata standard would be useful for describing CA682 (and other modules) because it would ensure consistency and interoperability. If all modules were described using the same standard, it would be easier to:

* **Search for modules** based on specific criteria (e.g., keywords, module coordinator, level).
* **Compare modules** from different institutions or departments.
* **Integrate module information** into different systems (e.g., student management systems, online course catalogues).

A good example is the **Dublin Core** standard. It proposes a simple set of metadata elements that are generic enough to be used across various disciplines and applications.

**Problems with Enforcing a Standard:**

One problem with enforcing a metadata standard is that it can be difficult to get people to adhere to it. Some common reasons for this include:

* **Lack of awareness:** People may not be aware of the standard or its benefits.
* **Perceived complexity:** The standard may be perceived as too complex or time-consuming to implement.
* **Lack of incentives:** There may be no incentives for people to use the standard.

In addition to these factors, the sources also highlight potential semantic overlap between some elements of the Dublin Core standard and the possibility of requiring judgment from the person assigning the metadata, which can introduce inconsistencies.

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

Big data is characterised by its Volume, Velocity, Variety and recently Veracity. Veracity refers to how reliable the data is.

(i) Why does big data have a particular issue with data uncertainty?

(ii) Give one (1) method for checking the veracity of a big data source.

### Data Uncertainty in Big Data

**(i) Why does big data have a particular issue with data uncertainty?** While the sources do not explicitly address the relationship between big data and data uncertainty, they offer insights into the complexities of data quality, which can be extrapolated to understand the challenges of veracity in big data. Big data is often characterised by its volume, velocity, and variety. The massive scale of big data often involves collecting data from diverse sources, which increases the likelihood of inconsistencies, errors, and biases.

The sources emphasize that data generated by people can be messy, leading to issues like manual entry errors, typos, and inconsistencies. When dealing with big data, the sheer volume amplifies these potential issues. Additionally, the high velocity of data influx in big data scenarios can make it difficult to implement real-time quality checks, potentially allowing unreliable data to enter the system.

**(ii) Give one (1) method for checking the veracity of a big data source.** One method for checking the veracity of a big data source, as suggested by the sources, is **Verification**. This involves implementing checks and using a verification parser to ensure that the uploaded data conforms to an expected pattern. For instance, if you are collecting data on customer orders, you could use a verification parser to check that the order IDs are in the correct format and that the order dates are within a reasonable range.

This response draws on information from the sources and previous turns in the conversation to address the query. The sources do not directly discuss data uncertainty in the context of big data; therefore, the provided response is an inference based on related concepts in the sources. You may want to independently verify this information.

**Q 2(d) [4 Marks]**

Describe the process of scraping data from a website. Give two (2) rules that you should remember when using this as data source.

***[End of Question2]***

CA682A – Data Management and Visualisation

August/Resit Examinations 2018/2019 Page 3 of 8

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

**Q 3(a) [8 Marks]**

You are collecting data to for a marketing campaign to increase public transport use to travel to DCU. You have the following data sources:

A. Survey of current transport patterns of staff and students.

B. Access logs from Transport Ireland app filtered by GPS location for the Glasnevin Campus.

C. Map of transport options for the DCU campuses.

D. Medical research data showing improved heart health from regular walking based on user’swearing fitbit sensors.

For **each** of these sources, identify one (1) possible cause and consequence of poor quality data.

Here is an analysis of the possible causes and consequences of poor quality data for each of your data sources:

### A. Survey of current transport patterns of staff and students

* **Possible cause of poor quality data:** **Poor survey or interface design** could lead to inaccurate or incomplete data. For example, if the survey only provides limited options for modes of transport, respondents might be forced to select an option that doesn't accurately reflect their usual travel habits. Similarly, an interface design that is confusing or difficult to navigate could lead to respondents skipping questions or providing incorrect information.
* **Possible consequence of poor quality data:** Inaccurate data from the survey could lead to **misinformed decisions about the marketing campaign**. For instance, if the survey data suggests that a large proportion of staff and students already use public transport, the campaign might focus on encouraging more frequent use rather than targeting those who currently don't use public transport at all. This could result in a less effective campaign overall.

### B. Access logs from Transport Ireland app filtered by GPS location for the Glasnevin Campus

* **Possible cause of poor quality data:** **Errors during data gathering**, such as GPS location inaccuracies or data transmission problems, could lead to unreliable data. For example, if a user's GPS location is slightly off, they might be recorded as being at the Glasnevin Campus when they are actually somewhere else. Similarly, network overload could lead to data packets being lost during transmission, resulting in incomplete or missing data.
* **Possible consequence of poor quality data:** If the access logs contain inaccurate or incomplete data, it could lead to **misleading insights about public transport usage patterns**. For example, if a significant number of users are incorrectly recorded as being at the Glasnevin Campus, the data might overestimate the number of people using public transport to travel to DCU. This could lead to an underestimation of the need for improved public transport services.

### C. Map of transport options for the DCU campuses.

* **Possible cause of poor quality data:** **Lack of standards for format or controlled vocabulary for fields** could lead to inconsistencies and make it difficult to integrate the map data with other data sources. For example, if different terms are used to refer to the same transport option (e.g., "bus stop" vs. "bus station"), it could be challenging to combine data from the map with data from other sources, such as the Transport Ireland app access logs.
* **Possible consequence of poor quality data:** Inconsistent or poorly formatted data could hinder the **effectiveness of the marketing campaign**. For example, if the map data cannot be reliably integrated with data on student and staff travel patterns, it might be difficult to identify areas where public transport services are lacking or to target specific groups with relevant information. This could limit the impact of the campaign.

### D. Medical research data showing improved heart health from regular walking based on users wearing Fitbit sensors.

* **Possible cause of poor quality data:** While the provided data sources do not directly mention the potential causes of poor quality data for medical research data, the example of the **Mars orbiter** highlights the importance of accurate data and the potential consequences of errors. A possible cause of poor quality data in this scenario could be **errors in data entry** or **data manipulation**, such as typos when recording Fitbit data or incorrect unit conversions during analysis.
* **Possible consequence of poor quality data:** If the medical research data contains errors, it could lead to **inaccurate conclusions about the benefits of walking**. For example, if errors in data entry or analysis inflate the apparent improvement in heart health, the marketing campaign might overstate the health benefits of walking to DCU. This could damage the credibility of the campaign and undermine its effectiveness.

Remember, **high-quality data is essential for making informed decisions**. By carefully considering the potential causes of poor data quality and implementing appropriate data cleaning and verification methods, you can increase the reliability of your data and improve the effectiveness of your marketing campaign.

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

Pick one of the data sources listed in Q3(a).

1. Give an example of an approach to cleaning data that you could use. (ii) Give an example of how you could enforce better data quality.

## Cleaning and Enforcing Quality in Transport App Data

Choosing **Data Source B: Access logs from Transport Ireland app filtered by GPS location for the Glasnevin Campus.**

**(i) Examples of approaches to cleaning data:**

* **Removing Duplicate Records:** The access logs might contain duplicate entries for the same user and journey. This could happen due to technical glitches or inconsistencies in data recording. Identifying and removing these duplicates ensures that each journey is counted only once. \*\*\*\*
* **Handling Missing Values:** Some access logs might have missing data points, such as the user ID or the GPS location. There are several ways to handle missing values, including:
  + Deleting records with missing values.
  + Replacing missing values with estimated values (e.g., using an average or median).
  + Using algorithms that can handle missing data. \*\*\*\*
* **Addressing GPS Inaccuracies:** GPS data can be inaccurate due to various factors like signal interference or device limitations. To improve data quality, techniques like:
  + **Filtering Outliers:** Extreme GPS values that are clearly out of range can be removed.
  + **Smoothing Techniques:** Apply algorithms to smooth out GPS trajectories and reduce noise.
  + **Snapping to Road Network:** Match GPS points to the nearest road on a map to correct for inaccuracies.

While the sources do not directly discuss these specific techniques for handling GPS inaccuracies, they emphasize the importance of data cleaning and the use of various tools and methods to improve data quality. You may want to independently verify the suitability of these methods for your specific dataset.

**(ii) Example of enforcing better data quality:**

* **Implement Data Validation Checks:** To prevent errors during data gathering, implement validation checks within the Transport Ireland app. For example: \*\*\*\*
  + Require users to enable location services for accurate GPS data collection.
  + Implement checks to ensure that GPS coordinates fall within a reasonable range for the Glasnevin Campus.
  + Prompt users to verify their location if the app detects any inconsistencies.

These preemptive measures can help in reducing errors at the data gathering stage and improving the overall veracity of the access logs. \*\*\*\*

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

(i) What are **constraints** with respect to Data Quality? Ensure you define and distinguish between static and dynamic constraints.

(ii) What are two (2) potential problems with enforcing constraints to improve data quality?

## Constraints in Data Quality

**(i) What are constraints with respect to Data Quality? Ensure you define and distinguish between static and dynamic constraints.**

Constraints in data quality are rules or limitations set to ensure the accuracy, consistency, and reliability of data. They define acceptable values, formats, and relationships within a dataset. The sources differentiate between two main types of constraints:

* **Static Constraints:** Based on the schema or structure of the data. These constraints focus on the inherent properties of individual data fields and are typically enforced at the database level. Examples include:
  + **Nulls not allowed:** Specifying whether a field can be left empty.
  + **Field domains:** Defining the valid range of values for a specific field (e.g., age must be a positive integer).
  + **Foreign key constraints:** Ensuring relationships between tables are maintained (e.g., a customer ID in an orders table must correspond to an existing customer in the customers table).
* **Dynamic Constraints:** Based on the workflow or business rules associated with the data. They involve more complex conditions and calculations and often depend on factors beyond the individual data fields. Examples include:
  + **Orders above $200 are processed by Biller 2:** This constraint dictates a specific action based on the order value, influencing the workflow of order processing.

The sources highlight that many data quality problems can be addressed using static constraints implemented through the database schema, while dynamic constraints address problems arising from workflow and business logic.

**(ii) What are two (2) potential problems with enforcing constraints to improve data quality?**

While constraints are essential for maintaining data quality, their enforcement can also present challenges:

1. **Overly strict constraints can lead to data loss:** If constraints are too rigid, they might reject valid data points that deviate slightly from the defined rules. For example, a constraint requiring all addresses to have a postal code might reject valid addresses in regions where postal codes are not used. This can result in the loss of valuable data and introduce bias into the dataset.
2. **Enforcing constraints can be complex and resource-intensive:** Implementing and maintaining a large number of constraints, especially dynamic constraints that involve intricate business rules, can require significant computational resources and technical expertise. This complexity can make data management more challenging and costly, particularly for large and complex datasets.

While the sources do not directly address these specific problems, they do mention that data quality is a complex issue and achieving perfect data is unlikely. Additionally, the sources highlight the importance of balancing the benefits of constraints with the potential downsides. The 80-20 rule suggests that a few constraints can address most data quality issues, but capturing all edge cases might require thousands of constraints, leading to increased complexity.

This information from the sources provides insights into the potential problems with enforcing constraints, emphasizing the need for a balanced and thoughtful approach to data quality management.

**Q 3(d) [5 Marks]**

Open datasets are made freely available for all people to access. Identify and

explain two (2) potential problems that may arise in making data open or using open data.

***[End of Question3]***

CA682A – Data Management and Visualisation

August/Resit Examinations 2018/2019 Page 4 of 8

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

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

Given the following visualisation tasks, suggest an appropriate graph type (specific chart type not just the category) for each to display the information and give a brief justification.

A. Summary of voter choices in the Irish Election 2016

B. Annual income for Computer Science students grouped by university

C. Population trend for Ireland over the last decade

D. Average rental prices for each Irish County in 2016

Here are some appropriate graph types for each of the visualisation tasks, with justifications:

### A. Summary of voter choices in the Irish Election 2016

* **Stacked Bar Chart**: This chart type would be useful for showing the proportion of votes each party received. Each bar would represent a constituency, and segments within each bar would represent the different parties, with the size of each segment reflecting the proportion of votes received. This allows for quick comparisons between constituencies and an understanding of the overall distribution of votes. While the sources don't specifically mention the 2016 Irish Election, they do provide various examples of chart types and their suitability for different data types. The stacked bar chart is commonly used for showing parts of a whole, making it appropriate for summarizing election results.

### B. Annual income for Computer Science students grouped by university

* **Box-and-Whisker Plot:** This type of chart effectively displays the distribution of income data, showing the median, quartiles, and potential outliers for each university. This allows for a nuanced comparison of income ranges and central tendencies across different universities. This is a common chart type for statistical analysis.

### C. Population trend for Ireland over the last decade

* **Line Chart:** This is a straightforward way to visualize trends over time. The x-axis would represent the years, and the y-axis would represent the population. The line would clearly show the increase or decrease in population over the decade. Line charts are frequently used to showcase changes in quantitative values over time.

### D. Average rental prices for each Irish County in 20

* **Choropleth Map:** This map type uses color shading to represent data values for different geographic regions. Each county would be shaded according to its average rental price, allowing for an immediate visual comparison of rental costs across the country. Choropleth maps are specifically designed for displaying data associated with distinct spatial regions.

**Q 4(b) [7 Marks]**

In the appendix, Figure 1 shows a graph. Identify three (3) problems with the design and suggest a better method for showing the information, giving a specific chart type that could be used. You do not need to view the appendix in colour.

## Issues with the Pie Chart and a Suggested Alternative

The provided image (Figure 1 – a 3d pie-chart) from the appendix of source presents "Sales 2006-2012" as a 3D pie chart. Here are three problems with this design choice:

1. **Distortion of Data:** 3D pie charts inherently distort the perception of data by altering the relative size of segments. The segments closer to the viewer appear larger than those further away, making accurate comparisons challenging. This violates the principle of **trustworthiness**, a key aspect of good data visualisation outlined by Andy Kirk in source.
2. **Difficult to Compare Segment Sizes:** In a pie chart, the human eye struggles to accurately compare the areas of different segments, especially when they are close in size. This difficulty is compounded in a 3D pie chart by the added visual complexity and distortion. While humans are good at judging linear distances, using area for encoding data can make comparisons challenging.
3. **Unnecessary Chart Junk:** The 3D effect adds no value to the data being presented and only serves as a distraction. The drop shadow and slanted perspective add visual clutter, hindering the clarity of the chart. Source emphasizes the importance of removing unnecessary elements from graphs to enhance the data-to-pixel ratio.

**A Better Alternative**

To present the "Sales 2006-2012" data more effectively, a **grouped bar chart** would be a suitable alternative. This type of chart:

* Allows for **clear and accurate comparison** of sales figures across different years by directly comparing the heights of the bars.
* Reduces the risk of **data distortion** as the bars are presented in a 2D plane.
* Facilitates the **perception of trends** over time by visually connecting the bars for each year.

This choice aligns with source's advice to prioritize attention over distraction and the advice given in source to switch to a bar chart for representing data differences clearly.

**Q 4(c) [6 Marks]**

Explain what D3.js is and give 2 examples of things it doesn't do. Explain the difference between rules and selectors (CSS) and how they are used in D3.js.

## Explaining D3.js, its limitations, and the use of CSS rules and selectors

* D3.js (Data-Driven Documents) is a JavaScript library for creating interactive and dynamic data visualisations within web browsers.
* It works by transforming data into standard web formats like HTML, SVG, and CSS.
* This means it leverages the browser's existing tools but provides an easier interface for using those tools for data visualisation.

### What D3.js doesn't do:

* **Support older browsers:** D3.js may not work optimally in older browsers due to limitations in their support for modern web standards.
* **Generate prepared visualisations for you:** Unlike tools like Excel and Tableau, which offer pre-built chart options, D3.js requires you to code the visualisations.
  + This also means that you don't typically use D3.js for data processing or analysis tasks.

### Rules and Selectors in CSS and their application in D3.js:

The sources do not explicitly cover the distinction between CSS rules and selectors or their specific use in D3.js. However, based on general web development knowledge, this is an explanation:

**CSS Rules:** Define how HTML elements should be styled. They consist of two parts:

* + **Selector:** Identifies the HTML element to be styled.
  + **Declaration:** Specifies the styling properties and their values.

**Example:**

p { /\* Selector: targets all <p> elements \*/

color: blue; /\* Declaration: sets text color to blue \*/

font-size: 16px; /\* Declaration: sets font size to 16 pixels \*/

}

**Selectors:** Allow you to target specific HTML elements based on their tag name, class, ID, attributes, or their relationship to other elements. Some common selector types:

* + **Element Selector:** p (targets all paragraph elements)
  + **Class Selector:** .highlight (targets elements with the class "highlight")
  + **ID Selector:** #header (targets the element with the ID "header")
  + **Attribute Selector:** [type="text"] (targets input elements with the type attribute set to "text")

#### How this applies to D3.js:

* D3.js uses selectors to manipulate HTML and SVG elements, applying data-driven attributes and styles to create visualisations.
* For instance, you can use D3.js to select all bars in a bar chart and set their height based on data values.
* You can also use CSS rules to define the appearance of elements in a D3.js visualisation, such as the colour, size, and position of shapes or text.

Please note that this information regarding the use of CSS rules and selectors in D3.js is based on general web development knowledge and is not explicitly stated in the sources. You may want to verify this information independently.

**Q 4(d) [4 Marks]**

How can design rules help to make better data visualisations? Give an example.

## How Design Rules Improve Data Visualisations

Design rules, often referred to as principles or guidelines, are crucial for enhancing the effectiveness of data visualisations. They provide a framework for making informed design choices that promote clarity, accuracy, and engagement. The sources emphasize the importance of understanding and applying these rules to create visualisations that effectively communicate insights from data.

**Here's how design rules contribute to better data visualisations:**

* **Trustworthiness:** Design rules ensure that the visualisation accurately represents the underlying data without distortion or manipulation. Principles like avoiding inappropriate color palettes or unnecessary chart junk contribute to creating a trustworthy visualisation that viewers can rely on.
* **Accessibility:** Design rules promote inclusivity by making visualisations understandable and usable for a wider audience, including individuals with visual impairments. Using clear labels, appropriate font sizes, and color combinations that consider color blindness are examples of design choices that enhance accessibility.
* **Elegance:** A well-designed visualisation is not only informative but also visually appealing and engaging. Applying design principles related to layout, typography, and color harmony can elevate the aesthetics of a visualisation, making it more enjoyable and memorable for the viewer.
* **Cognitive Load Reduction:** Design rules aim to minimize the mental effort required to understand the information presented in a visualisation. By using familiar chart types, minimizing clutter, and employing effective visual hierarchies, designers can reduce cognitive load and facilitate comprehension.
* **Purposeful Storytelling:** Design rules guide the creation of visualisations that effectively convey a specific message or story. By carefully selecting chart types, color schemes, and visual elements, designers can craft a narrative that resonates with the audience and leads to a better understanding of the data.

**An Example: Enhancing a Scatter Plot**

Consider a scatter plot showing the relationship between two variables, such as income and life expectancy. Applying design rules can significantly improve its effectiveness:

* **Using Clear Labels and Titles:** Providing descriptive labels for the axes and a concise title helps viewers quickly grasp the context and purpose of the visualisation.
* **Choosing Appropriate Color and Shape:** Using distinct colors and shapes for different data points, if applicable, can aid in visual differentiation and highlight patterns. Ensuring color choices consider potential color vision deficiencies enhances accessibility.
* **Minimizing Clutter:** Removing grid lines, unnecessary annotations, and excessive data points can reduce visual noise and allow the key trends to stand out.
* **Adding Trend Lines:** If appropriate, adding a trend line can clearly illustrate the overall relationship between the variables.

By adhering to these design rules, the scatter plot becomes more trustworthy, accessible, elegant, and easier to understand. It effectively tells a story about the relationship between income and life expectancy, enabling viewers to extract meaningful insights.

The sources consistently emphasize that data visualisation is a communication process. By understanding and applying design rules, we can create visualisations that effectively bridge the gap between data and understanding, empowering viewers to make informed decisions and explore complex relationships.

***[End of Question4]***

CA682A – Data Management and Visualisation

August/Resit Examinations 2018/2019 Page 5 of 8

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

**Q 5(a) [7 Marks]**

In the appendix, Figure 2 shows a graphic.

What visual communication goals are evident?

Identify two (2) design principles and explain how the graphic applies them to fulfil the communication goals.

**Q 5(b) [6 Marks]**

Using Figure 2, explain and give examples of two (2) gestalt principles and how they are used.

**Q 5(c) [6 Marks]**

In visualisation it is important to understand pre-attentive processing. Explain what a pre-attentive feature is and describe an experiment to determine if a feature is pre- attentive or not.

**Q 5(d) [6 Marks]**

Correctly match the following depth cues for human vision.

|  |  |
| --- | --- |
| Occlusion | A. moving the head slightly to create differences in the sensed images |
| Convergence | B. difference in direction of our eyes when looking at closer objects |
| Accommodation | C. images sensed by our two eyes are slightly different and this difference is used to determine depth |
| Aerial Haze | D. blocking of more distant objects by closer objects |
| Binocular Disparity | E. objects on the far horizon look hazy due to particles in the air |
| Motion Parallax | F. muscle tension from re-focussing the eye |

Which cue is (mostly) used to create the 3D effect in movies?

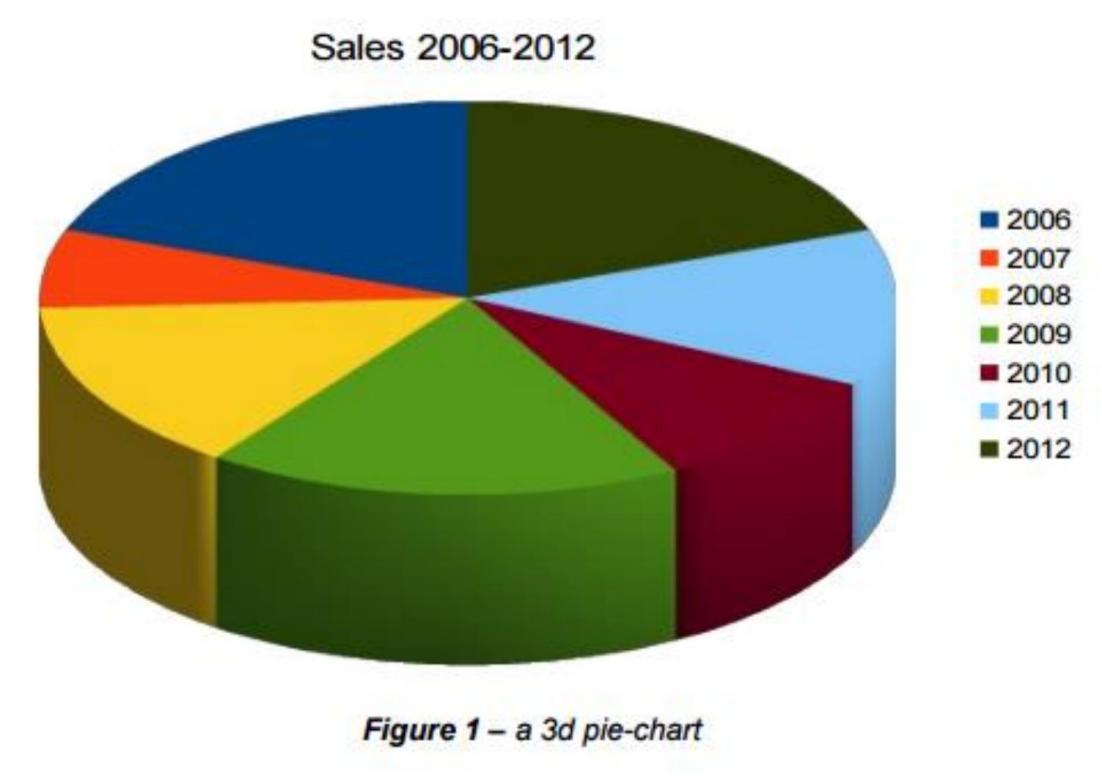
***[End of Question5]***

CA682A – Data Management and Visualisation

August/Resit Examinations 2018/2019 Page 6 of 8

***APPENDICES***

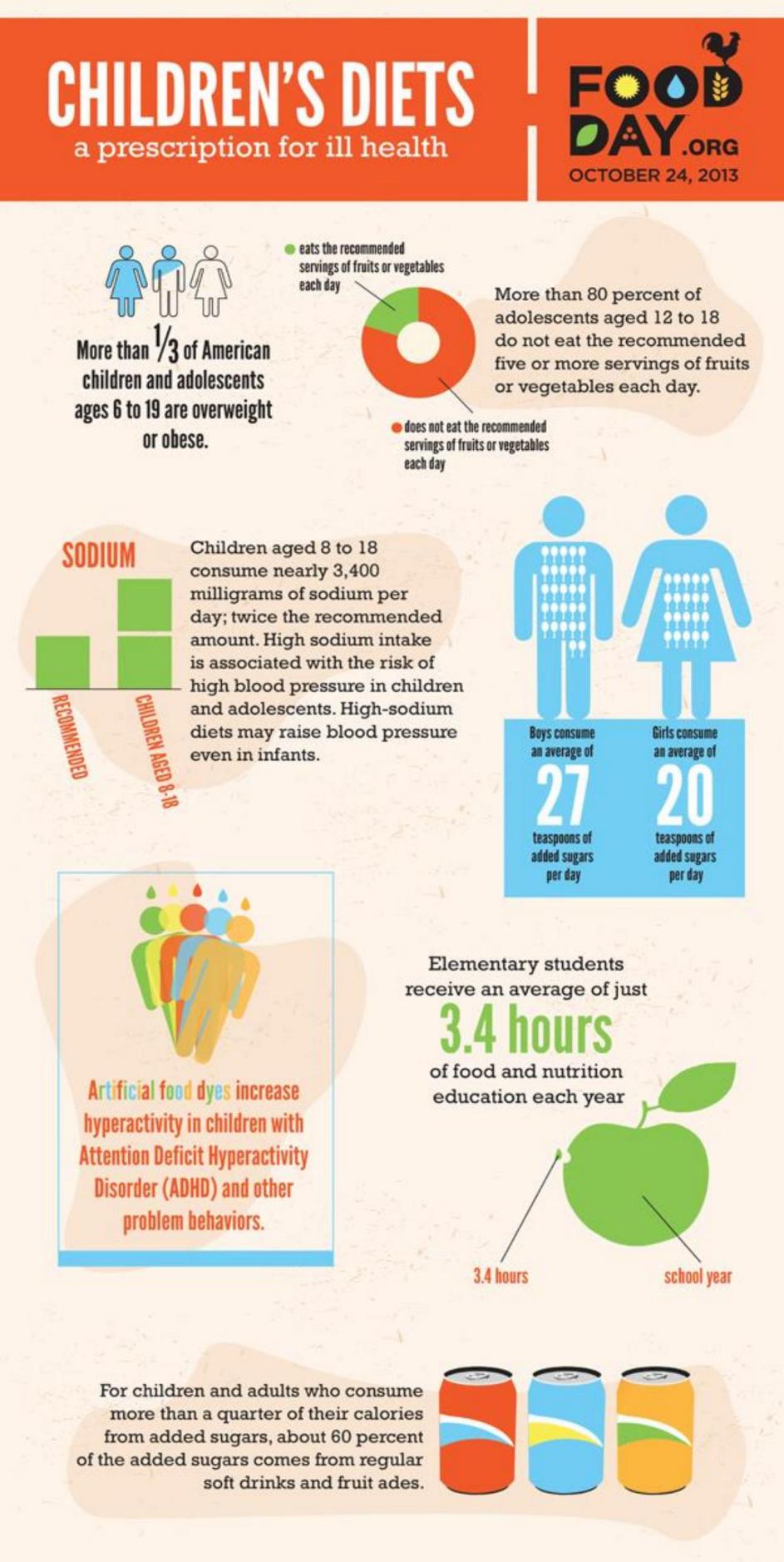
***Figure 1***



CA682A – Data Management and Visualisation

August/Resit Examinations 2018/2019 Page 7 of 8

***Figure 2***



***[END OF APPENDICES] [END OF EXAM]***

CA682A – Data Management and Visualisation

August/Resit Examinations 2018/2019 Page 8 of 8