## 

**SEMESTER 1** **EXAMINATIONS** **2020/2021**

**MODULE:** CA682, CA682A - Data Management and Visualisation

### PROGRAMME(S):

|  |  |
| --- | --- |
| MCM | M.Sc. in Computing |
| CAPT | PhD-track |
| EEPT | PhD-track |
| LGPD | PhD |
| EEPM | Meng |
| CAPD | PhD |
|  |  |

**YEAR OF STUDY:** 1,2,3

**EXAMINER(S):**

|  |  |  |
| --- | --- | --- |
| Suzanne Little | (Internal) | (Ext:6052) |
| Dr. Dan Emanoil Grigoras | (External) | External |
| Dr. Katarina Domijan | (External) | External |
| Prof. Joseph Cannataci | (External) | External |

**TIME ALLOWED:** 3 Hours

**INSTRUCTIONS: Answer 4 questions of the 5 provided. All questions carry equal marks.** Sketches or graphs should be pasted into the document. Type or paste your answers into the indicated boxes.

**PLEASE UPLOAD A SINGLE DOCUMENT WITH YOUR ANSWERS.**

**By submitting this exam, you declare (1) that all of the work is your own; (2) that you did not seek whole or partial solutions for any part of your submission from others; and (3) that you did not and will not discuss, exchange, share, or publish complete or partial solutions for this exam or any part of it during the exam.**

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

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

## QUESTION 1 [TOTAL MARKS: 25]

### Q 1(a) [12 Marks]

Using the topic from your CA682 visualisation assignment, apply the Generic Data Analytics Pipeline to describe how the data may have been Gathered, Processed, Analysed, Presented and Preserved. Give a brief description of the activities at each stage (1-2 sentences) and identify any specific tools that you did or would use. If you didn’t specifically perform any stage then you can make assumptions or predictions about the actions and tools.

If you didn’t complete a visualisation assignment then write about a scenario based on analysing *student feedback to DCU in relation to online teaching in 2020*.

Gathering

Processing

Analysis

Presenting

Preserving

The FIFA dataset you chose from Kaggle is a great example of a "big" dataset, as described in the sources. It has volume (6GB of data and millions of rows), variety (information about players, coaches, and teams), and while not explicitly stated, likely some degree of velocity if you were accessing real-time updates. However, you point out that finding a free dataset with all these features is challenging.

Data Cleaning

The sources highlight the importance of data quality, and data cleaning is a crucial step in ensuring that. Data scientists often spend a significant amount of time cleaning and organizing data before analyzing it. The sources list a variety of potential problems that can occur during data gathering, data delivery, and data integration, many of which are caused by human error or inconsistencies.

The cleaning steps you describe for the FIFA dataset, encoding, standardization, grouping and selection, and combining data, all address potential data quality issues.●

Encoding: The need to convert the dataset from ASCII to UTF-8 underscores the fact that data often comes in various formats, and proper encoding is necessary for accurate representation. The sources offer examples of different file formats, including CSV, TSV, JSON, and XML, and emphasize that some formats may be proprietary, adding another layer of complexity to data processing.●

Standardization: The steps you outline, such as adding a gender field, removing empty rows, cleaning extra symbols, and standardizing dates, are common data cleaning practices. This process ensures consistency and makes the data easier to analyze. The sources emphasize the importance of data quality metrics to assess the effectiveness of cleaning efforts.●

Grouping and Selection: Grouping players by country, removing duplicates, and selecting the player with the highest totalStats are all forms of data transformation, aimed at preparing the data for meaningful analysis. This process allows for easier comparison and analysis of player performance within and across different countries.●

Combining Data: The use of Spark to merge the six processed datasets leverages a powerful tool designed for handling large datasets. The sources identify Spark as a valuable tool for data processing and analysis, particularly in a "big data" context. Combining datasets allows for a more holistic view of the data and facilitates more comprehensive analysis.

●

tool

Gathering: Kaggle for obtaining the dataset.●

Processing: We used IntelliJ IDEA, command prompt (CMD)and Spark for data cleaning and optimization, handling tasks like removing duplicates, standardizing formats, and processing large datasets efficiently.

For file management, Xftp facilitated secure file transfers between our Windows system and the Linux server, while Xshell enabled remote terminal access for managing Spark jobs and HDFS.

Spreadsheet tools (e.g., Excel) were used for quick data validation and spot-checking before bulk processing.

Analysing: Spark for data exploration and analysis.●

Presenting: PyCharm and Python visualization libraries(Matplotlibor)

Preserving:MYSQL HDFS

●

Gathering: The data was captured from Kaggle, an open data platform. You chose a football game dataset from Kaggle that contains information about male and female players, teams, and coaches for the years 2018 to 2023. The data was originally stored in six CSV files.●

Processing: The dataset required cleaning and transformation. We cleaned the data by converting it from ASCII to UTF-8, standardizing dates, and removing duplicates. We also added a gender field and calculated the total stats for each player. This process involved using tools like PowerShell to check the encoding and Spark to process and combine the data. The data was then uploaded to HDFS for efficient storage and processing.●

JOIN operations in Spark can be computationally expensive, especially for large datasets, leading to performance issues like stack overflow errors.

To avoid using the JOIN operation in Spark, which can be resource-intensive for large datasets, we can convert the six processed arrays into maps and then merge them into a single, larger map using a key-value mapping approach.

Analysing: we analysed the data using Spark to determine the best player from each country based on their total score. This analysis would have involved exploring the data and identifying trends in player performance across different countries.

Players often care about the strongest players. However, we also need to consider their potential, which shows their future ability.

We aim to help gamers choose players based on three factors:

(1)Favorite country: Choose the strongest player from that country.

(2)High potential: Select players with the best future potential.

(3)Appearance: Pick players they find good-looking.

Presenting: At the very start,the results of our analysis were presented in a simple table displayed on a webpage using Thymeleaf as the front-end. Although this effectively displays the dataset in a table format, it is not the final result. Later, my partner created an interactive world map using Python. When we click on a country on the left-side map, the corresponding player’s information is automatically shown on the right side. This enhancement improves interactivity, making it more engaging and informative compared to a static table. Additionally, we used radar charts to represent various player statistics, providing a more intuitive and visual way to understand their performance data.

Preserving: Since we had already processed the data from six CSV files into a smaller, structured dataset with no missing fields and complete information, we stored this dataset in a MySQL database for easier access and retrieval.

Although the results were already stored on HDFS, we used MySQL for storage to support local visualization development on our Windows 10 system.

### Q 1(b) [8 Marks]

For each of the following data attributes (A-D), choose all of the following descriptions that can apply. Marks will be deducted for including wrong choices.

Qualitative, Quantitative, Discrete, Continuous, Nominal, Ordinal, Interval, Ratio

A Number of bicycles owned per household:

Quantitative Discrete Ratio

B Average time taken to commute each day:

Quantitative Continuous Ratio

C Mode of transport used to commute on Monday:

Qualitative Nominal

D Motor vehicle safety rating (Gold, Silver, Bronze):

Qualitative Ordinal

### Q 1(c) [5 Marks]

Which of the following situations is **most likely** to be classified as big data:

A Viewing data for Netflix subscribers including the show and the date watched and social media sentiment analysis responding to the show.

B Sales data from the four DCU campus restaurants and catering facilities in 2020.

C A download of content and metadata from my personal twitter account.

D Player training data (sensors and observations) from the Irish Rugby Squad.

Answer: Answer: A

Explanation:

The scenario most likely to be classified as big data is A: Viewing data for Netflix subscribers, including the show, date watched, and social media sentiment analysis responding to the show.

In your own words, referring to the classical characteristics for big data, explain why your choice is most likely to produce big data.

Answer: Big data is typically characterized by:●

Volume: A vast quantity of data. Netflix, with its enormous global subscriber base, generates massive amounts of viewing data daily. The inclusion of social media sentiment analysis adds another layer of data, potentially encompassing millions of tweets, posts, and comments.●

Velocity: The speed at which data is generated and processed. Netflix viewing data is continuously generated in real time as users stream content. Social media responses also flow in at a rapid pace.●

Variety: The diversity of data sources and formats. Netflix viewing data includes structured data (show title, date, duration) and unstructured data from social media (text, emojis, images, etc.). Sentiment analysis requires processing diverse formats and understanding natural language.

The other scenarios are less likely to constitute big data:●

B: Sales data from DCU campus restaurants: While this involves transactional data, it's limited to four restaurants within a specific timeframe. The volume and velocity are unlikely to be on the scale of Netflix's global data.●

C: Personal Twitter data: An individual's Twitter data is relatively small compared to the vast data generated by platforms like Netflix.●

D: Irish Rugby Squad training data: Data from sensors and observations for one sports squad might be complex but is likely confined in volume compared to the streaming habits and social media interactions of millions of Netfli

## [End of Question 1]

## 

## QUESTION 2 [TOTAL MARKS: 25]

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

Given the following brief to design a system for a data collection task:

“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.”

1. List three (3) important questions you would ask your client.
2. Describe the data and/or specific file formats that you are likely to use in collecting and storing the data.
3. Suggest a type of database storage approach to use for this project, giving a reason for your choice.

Answer:

(i)

Clear Objectives and Deliverables: What are the specific objectives of the report? What key insights or policy-relevant information does the government aim to extract from this analysis? Defining clear objectives from the outset guides the entire data analysis process and ensures the final report meets the client's needs.●

Target Audience and Reporting Needs: Who is the intended audience for the report, and what are their data literacy levels and information needs? Understanding the target audience helps determine the appropriate level of detail, presentation style, and communication strategy. Will the report be primarily for high-level policymakers, or will it also be used by analysts and researchers within government departments?●

Key Policy Areas and Restrictions: Which specific COVID-19 restrictions are of primary interest for the analysis? For example, is the focus on workplace closures, travel limitations, or public gathering bans? Focusing on the specific restrictions helps identify the most relevant data sources and analysis techniques.●

Geographic Scope and Granularity: What level of geographic detail is required for the analysis (national, regional, county, city)? What time granularity is needed for the data (daily, weekly, monthly)? The desired geographic scope and granularity impact data collection, storage, and analysis methods.●

Data Accessibility and Quality: What is the accessibility of the data from data.gov.ie and other sources? What are the potential data quality issues with the existing datasets? Assessing data accessibility and quality early on helps anticipate challenges and plan for data cleaning and integration steps.●

Commuting Behaviour Definition: How does the client define "commuting behaviour?" Does this include all forms of transportation, or is there a specific focus on public transport or private car usage? A clear definition of the term helps select appropriate data sources and analysis methods.●

Comparison with Previous Data: How should the 2020 data be compared to the 2019 and 2009 data, given potential differences in data collection methods and available metrics? Establishing a consistent methodology for comparison ensures the findings are valid and meaningful.●

Ethical and Privacy Considerations: What are the ethical and privacy implications of using the available data sources, particularly personal data collected from surveys or GPS tracking? How will data be anonymized or aggregated to protect individual privacy while still providing meaningful insights?

(ii)

Existing Data Sources:○

Pedestrian Footfall: likely in CSV or Excel format, with timestamps and counts for specific locations within the central shopping area.○

Traffic Monitoring Data: likely in CSV or database format, with timestamps and vehicle counts for various points on the main arterial routes.○

Working from Home Survey Data: potentially in a variety of formats including SPSS, CSV, or Excel, depending on how the surveys were conducted and stored.○

Weather Data: typically available in CSV or API formats from meteorological services, containing daily or hourly readings for various parameters.○

Economic Data: often published in Excel, CSV, or PDF formats by government statistical agencies, with data aggregated at national or regional levels.●

New Consumer Surveys:○

Online surveys using platforms like SurveyMonkey or Google Forms would yield data in CSV or Excel formats.○

Focus groups or interviews would produce transcripts or notes, which could be stored as text files or in qualitative data analysis software.

(iii)

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

This approach is recommended for the following reasons:●

Structured Data: The majority of the data involved (footfall, traffic counts, economic indicators) is highly structured, fitting well within a relational database model. This allows for efficient querying and aggregation based on shared attributes (e.g., time, location).●

Data Integrity and Consistency: Relational databases enforce data constraints and relationships, ensuring data quality and consistency across the dataset.●

Accessibility for Government Departments: Many government departments likely already use relational databases, facilitating data sharing and integration across various systems.●

Support for Ongoing Monitoring: Relational databases offer tools and functionalities for querying, aggregating, and visualising data, supporting the need for ongoing monitoring and policy impact assessments.

### Q 2(b) [10 Marks]

1. Give simple example metadata (at least 5 elements) describing your mobile phone (or your computer if you don’t have a mobile phone).
2. For each metadata element, identify if it is Descriptive, Administrative or Structural.
3. If I was to collect and integrate data about the electronic devices used by all CA682 students then, in your own words, how would using a standard change the quality of metadata data? Identify one potential difficulty with enforcing a metadata standard.

Answer:

(i)

Device Name: iPhone 13 (Descriptive)●

Operating System: iOS 16.2 (Descriptive)●

Storage Capacity: 256GB (Descriptive)●

Date Purchased: 2022-10-01 (Administrative)●

File Structure: Hierarchical file system with folders and subfolders (Structural)

(ii)

（●

Types of Metadata●

Descriptive: This type of metadata provides information about the content or characteristics of the object itself. In this case, the device name, operating system, and storage capacity directly describe the features of the mobile phone.●

Administrative: This type of metadata deals with the management and creation of the object. The date purchased provides information about when and potentially where the phone was acquired, aiding in tracking its lifecycle.●

Structural: This type of metadata describes how the information or data within the object is organised. The hierarchical file structure of the mobile phone determines how files and folders are arranged and accessed.

Descriptive Metadata: Descriptive metadata provides information about the content and characteristics of the data object. It aims to help users understand what the data is about.●

Examples of descriptive metadata include:○

Device name of a mobile phone○

Operating system of a mobile phone○

Storage capacity of a mobile phone○

Title of a book○

Author of a book○

Year of publication of a book○

Genre of a book○

Summary of a book○

Description of a book○

Location of a book●

Administrative Metadata: Administrative metadata relates to the management and creation of the data object, covering its extrinsic properties. It can include details about:○

When the data was created○

Who created the data○

How the data has been managed○

File format of the data●

Examples of administrative metadata include:○

Date purchased of a mobile phone○

Date and time an image was taken●

Structural Metadata: Structural metadata describes how the information or data within the object is organized, covering its structure, format, and composition. It can address:○

How the data is arranged○

How different parts of the data relate to each other○

File structure of a mobile phone●

Examples of structural metadata include:○

Hierarchical file structure of a mobile phone○

Schema of a database）

(iii)

Impact of a Standard on Metadata Quality

Using a standard for collecting data about electronic devices used by CA682 students would significantly improve the quality of the metadata in several ways.

Enhanced Consistency and Comparability: A standard would ensure that everyone uses the same terms and definitions for describing their devices, leading to a more consistent and comparable dataset. For example, instead of having variations like "phone model," "device name," or "brand," a standard might specify a consistent term like "device\_model."

Improved Data Integration and Analysis: Consistent metadata makes it easier to integrate data from different students and perform meaningful analysis. For instance, you could easily compare the distribution of operating systems or storage capacities across the student population without needing to manually clean and standardize the data.

Facilitated Data Discovery and Retrieval: Standardized metadata allows for more effective searching and retrieval of specific devices or device characteristics. For example, if everyone uses the same terms for describing screen size, a researcher could easily find all students with devices exceeding a particular screen size.

Potential Difficulty with Enforcing a Metadata Standard

One major challenge with enforcing a metadata standard is ensuring compliance and consistent application across all students. Students might interpret or apply the standard differently, leading to inconsistencies or errors in the metadata. Some students might be more diligent in following the standard than others, resulting in variations in the completeness and accuracy of the metadata.

To mitigate this difficulty, it's crucial to:●

Provide clear and comprehensive documentation of the metadata standard.●

Offer training and support to students on how to correctly apply the standard.●

Implement validation mechanisms to check for adherence to the standard and identify potential errors.

### Q 2(c) [5 Marks]

### In your own words, describe the process of scraping data from a website. Give two (2) rules that you should remember when using this as data source.

Answer:

Web scraping is the process of automatically extracting data from websites. It involves using software tools to retrieve the underlying HTML code of a webpage and then parsing this code to identify and extract the desired information. This information can then be stored in a structured format for further analysis or use. The sources explain that this process typically involves three key steps:

1.

Parse HTML: The first step is to retrieve the HTML code of the target website. This code contains the structure and content of the webpage, including text, images, tables, and other elements. Tools known as "crawlers" or "spiders" are often used to fetch this HTML data.

2.

Match Patterns: Once the HTML code is obtained, the next step is to identify the specific data elements to be extracted. This often involves using regular expressions or other pattern-matching techniques to locate the relevant information within the HTML structure. The crawler needs to be able to recognize and distinguish between different elements like headings, paragraphs, lists, and tables.

3.

Identify Links (Repeat): Many websites organize information across multiple pages. To capture a complete dataset, the crawler must identify and follow links to other relevant pages. This can be done recursively, with the crawler extracting data from each linked page until all relevant content is gathered.

Rule 1

Rule 2

●

Data Quality: The sources emphasize the importance of data quality throughout the data lifecycle, particularly during the "Gathering" phase. Website structures can be inconsistent, and data might be presented in various formats, making it prone to errors or inconsistencies. Scraping tools may not always accurately extract data, leading to missing values, incorrect data types, or formatting problems. It's essential to thoroughly clean and validate the scraped data before using it for any analysis or decision-making. This might involve checking for missing values, correcting data types, standardizing formats, and removing duplicates.●

Respect Website Terms and Conditions: Websites often have terms of service or robots.txt files that specify how their data can be accessed and used. Violating these terms could have legal consequences. It's crucial to always review a website's terms of service and robots.txt file before scraping to ensure compliance and respect for the website owner's policies. Excessive scraping can also burden a website's server, potentially disrupting its normal operation. Be mindful of the frequency and volume of your scraping requests to avoid causing unintended problems.

## [End of Question 2]

## QUESTION 3 [TOTAL MARKS: 25]

### Q3 requires the dataset (Q3-D) provided online - <https://loopexam.dcu.ie/mod/resource/view.php?id=8654>

### Please download and use it to answer the question. 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.

### Q 3(a) [7 Marks]

Identify 3 different possible errors or artefacts in the dataset linked above. Give the tool or tools you used. You may use any tool that you like.

Answer:

1

2

3

Tool

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Error/Artefact | Likely Phase Introduced | Tool(s) | Data Quality Method(s) | Example | Data Quality Principle Affected |
| **Inconsistent Date Formats** | Gathering, Processing | Spreadsheets, OpenRefine, Python with Pandas | Standardization, Data validation rules | A dataset containing dates in various formats like "23/12/2023", "12-23-2023", "20231223". This can lead to errors in sorting, filtering, and calculations.1 | Consistency |
| **Missing Data** | Gathering, Integration | Spreadsheets, OpenRefine, Python with Pandas, SQL | Imputation, Rule-based data entry, Default values | A customer database with missing values in the "Address" field. This makes it difficult to segment customers for targeted marketing.2 | Completeness |
| **Language Inconsistency** | Gathering, Processing | OpenRefine, Python with libraries for language processing | Language-specific cleaning and transformation | A social media dataset containing posts in different languages, making sentiment analysis challenging.1 | Accuracy, Interpretability |
| **Sampling Bias (e.g., Gender Bias)** | Gathering | Statistical software (R, Python), Visualization tools (Tableau) | Stratified sampling, Quota sampling | A survey on political opinions that over-represents a particular demographic, leading to skewed results.3 | Credibility |
| **Unclear Units/Column Meanings** | **preserving** Data Storage | OpenRefine, Data cataloguing tools | Metadata documentation, Data dictionaries | A spreadsheet with a column named "Sales" but no indication of whether it's in Euros, Dollars, or units sold.4 | Interpretability |
| **Outliers or Typos** | Gathering, Processing | Spreadsheets, OpenRefine, Python with Pandas, Statistical software | Data validation rules, Double-entry, Automated data capture, Outlier detection techniques (e.g., box plots, z-scores) | A dataset with an unusually high temperature reading (e.g., 100°C in Ireland) that might indicate a sensor malfunction or data entry error.5 | Accuracy |
| **Data Redundancy** | **Processing** Data Integration | Database management systems (SQL, NoSQL), De-duplication tools | De-duplication, Entity resolution | A customer database with multiple entries for the same customer, potentially due to merging data from different sources. This can lead to inflated customer counts and inaccurate analysis.67 | Uniqueness |
| **Incorrect Data Types** | Processing | Spreadsheets, OpenRefine, Python with Pandas, R | Data type conversion, Schema validation | A column containing numerical data stored as text, preventing calculations and statistical analysis.4 | Accuracy |
| **Inconsistent Capitalization** | Gathering, Processing | OpenRefine, Python with Pandas | Standardization, String functions (e.g., to\_lowercase(), to\_uppercase()) | A dataset with inconsistent capitalization in a "City" column (e.g., "Dublin", "DUBLIN", "dublin"). This can affect sorting and grouping operations.1 | Consistency |
| **Time Gaps in Time Series Data** | Gathering, Processing | Time series analysis tools (e.g., Python with Pandas, R) | Interpolation techniques, Data imputation methods for time series data | A dataset tracking stock prices with missing data for a specific trading day, leading to inaccurate trend analysis.8 | Completeness |

# Install packages if not already installed

install.packages(c("readr", "dplyr", "stringr", "ggplot2"))

# Load libraries

library(readr)

library(dplyr)

library(stringr)

library(ggplot2)

# Load the CSV file

file\_path <- "C:/Users/zhou/Desktop/document/rsa\_2011-2012-V.csv"

data <- read\_csv(file\_path)

# Display the first few rows and structure of the data

print(head(data))

str(data)

# Identify rows with non-English characters in the Date column

non\_english\_rows <- data[grepl("[^ -~]", data$Age), ]

print(non\_english\_rows)

# Check the data type of Age column

print(class(data$Age))

# Check data types for each column

print(sapply(data, class))

# Get a summary of the data

summary(data)

# Check for missing values in each column

missing\_data <- colSums(is.na(data))

print(missing\_data)

# Find duplicate rows

duplicates <- data[duplicated(data), ]

print(duplicates)

#局部重复

duplicates\_by\_columns <- data[duplicated(data[c("Age", "Sex", "Participant","CasualtyType")]), ]

print(duplicates\_by\_columns)

# Ensure the numeric columns (`2011`, `2012`) are of numeric type

data$`2011` <- as.numeric(data$`2011`)

data$`2012` <- as.numeric(data$`2012`)

# Boxplot for columns `2011` and `2012` to check for outliers

boxplot(data$`2011`, main = "Boxplot for 2011 Data", ylab = "Values")

boxplot(data$`2012`, main = "Boxplot for 2012 Data", ylab = "Values")

z\_scores\_2011 <- scale(data$`2011`, center = TRUE, scale = TRUE)

outliers\_2011 <- data[which(abs(z\_scores\_2011) > 3), ]

print(outliers\_2011)

# Check unique values in categorical columns

print(unique(data$Age))

print(unique(data$Sex))

print(unique(data$Participant))

print(unique(data$CasualtyType))

# Calculate IQR for 2011 column

Q1 <- quantile(data$`2011`, 0.25)

Q3 <- quantile(data$`2011`, 0.75)

IQR\_value <- Q3 - Q1

# Detect outliers

lower\_bound <- Q1 - 1.5 \* IQR\_value

upper\_bound <- Q3 + 1.5 \* IQR\_value

outliers\_IQR <- data[data$`2011` < lower\_bound | data$`2011` > upper\_bound, ]

print(outliers\_IQR)

# Standardize inconsistent text formats (example: convert 'Sex' column to uppercase)

data$Sex <- str\_to\_upper(data$Sex)

# Filter data to check specific Age group or Participant inconsistencies

age\_check <- data %>% filter(Age == "0 - 5")

print(age\_check)

# Save the cleaned dataset to a new file (optional)

write\_csv(data, "C:/Users/zhou/Desktop/document/rsa\_2011-2012-V\_cleaned.csv")

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

Identify how each error or artefact is **most likely** to have been introduced, specifying the phase from the generic data analytics pipeline. State any assumptions.

Answer:Data Gathering Phase●

Inconsistent data entry: These errors typically occur during the data gathering phase. Manual entry, as seen in examples like "DUC or DCU", and differing formats for the same information, such as "DCU or Dublin City University?", highlight the human factor in introducing inconsistencies.●

Typographical errors: Similar to inconsistent data entry, typos are commonly introduced in the data gathering phase, especially when data is entered manually.●

Poor survey or interface design: Issues stemming from poorly designed surveys or interfaces often manifest during the data gathering phase. Leading questions or limited response options can result in skewed or incomplete data.

Data Storage Phase●

Incorrect data types: While this error can occur during data gathering, it can also be introduced during the data storage phase. Problems arise when converting between formats, such as deciding between string or float, or interpreting years like 1918 or 2018. Limitations in database fields, like storing rounded values, can further exacerbate this.●

Missing values: The data storage phase can also contribute to missing values. Technical issues like network dropouts or disk failures can lead to data loss. Additionally, a lack of documentation regarding the format of missing values (empty cells, NaN, -999) can create ambiguity in later stages.

Data Integration Phase●

Duplicate records: Duplicate records frequently arise during data integration. This occurs when merging datasets from different sources, as in the example of combining company records.●

Date Formatting and Data Type Misinterpretation: The sources often mention challenges with date formats. For example, Excel can interpret data differently based on system settings. It's possible the system processing the data mistook "6-9" for a date, attempting to fit it into a standard format like DD/MM/YYYY. The hyphen could have been misinterpreted as a separator.

Time synchronization errors: These errors, leading to missing values, are common in the data integration phase. Combining data with different time periods or incompatible time windows can result in gaps or misaligned data.

Data Analysis Phase●

Outliers: While outliers might be genuine, they could also signal errors introduced during previous phases. The data analysis phase is where these outliers are often detected.

General Considerations

It's important to note that these are just potential points of error introduction. Specific causes can vary depending on the dataset and processes. Additionally, some errors, like incorrect data types or missing values, can be introduced at multiple stages.

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

What data quality methods would you suggest using to either avoid or mitigate the error? Why would your suggestion improve data quality?

Answer:Data Gathering Phase

Data Validation Rules:

What it addresses: Inconsistent data entry and typographical errors.

How it improves data quality: Establishing validation rules, such as enforcing consistent formatting and limiting acceptable input values, helps prevent errors like "DUC or DCU" or typos. For example, using dropdown menus or pre-defined lists ensures consistency.

Suggested method: Implement front-end input validation with dropdowns or type checks and employ back-end checks for any discrepancies.

Clear Survey/Interface Design:

What it addresses: Poor survey or interface design.

How it improves data quality: Ensuring that surveys or data entry forms are well-designed (e.g., avoiding leading questions or offering clear choices) improves the quality of responses and reduces ambiguity.

Suggested method: Use concise and unambiguous language, and include guidance or examples for users to follow.

Data Storage Phase

Data Type Enforcement:

What it addresses: Incorrect data types.

How it improves data quality: Enforcing correct data types (e.g., integer, string, date) at the database schema level ensures that the data is stored in the appropriate format, reducing the risk of errors like storing numerical values as text.

Suggested method: Implement strong data type constraints in the database schema and regularly review data type conversions.

Automated Data Backups and Monitoring:

What it addresses: Missing values.

How it improves data quality: Regular automated backups and real-time data monitoring help mitigate the risk of data loss due to network failures or disk issues, ensuring that data remains intact and consistent.

Suggested method: Schedule regular backups and implement automated monitoring to detect potential issues.

Handling Missing Values:

What it addresses: Missing values and ambiguities in their representation.

How it improves data quality: Establishing a consistent approach for representing missing data (e.g., using NULL, NaN, or a designated placeholder) reduces confusion later in the analysis phase.

Suggested method: Define clear rules for handling and representing missing values at the time of data storage.

Data Integration Phase

De-duplication:

What it addresses: Duplicate records.

How it improves data quality: Applying de-duplication techniques, such as unique identifiers or data matching algorithms, ensures that duplicate records are removed when integrating datasets from different sources.

Suggested method: Use data matching algorithms or fuzzy matching to identify and remove duplicates, and implement automatic de-duplication processes during data integration.

Date Standardization:

What it addresses: Date formatting and data type misinterpretation.

How it improves data quality: Standardizing date formats across datasets ensures that data is interpreted correctly, reducing issues like misinterpreting "6-9" as a date.

Suggested method: Convert all dates to a single, consistent format (e.g., YYYY-MM-DD) and perform pre-processing checks on incoming data for formatting issues.

Time Synchronization:

What it addresses: Time synchronization errors.

How it improves data quality: Aligning time periods or time zones during data integration ensures that data points are correctly synchronized, preventing gaps or misaligned data.

Suggested method: Implement time zone normalization and ensure that timestamps are consistently synchronized across data sources.

Data Analysis Phase

Outlier Detection:

What it addresses: Outliers.

How it improves data quality: Using statistical methods (e.g., Z-scores or interquartile range) to identify and review outliers helps ensure that true outliers are distinguished from data entry or processing errors.

Suggested method: Use automated outlier detection techniques during data cleaning and analysis, followed by manual review where necessary.

### Q 3(d) [6 Marks]

1. Can you identify any potential personal or sensitive data in the provided sample dataset? Why or why not?
2. What process should you follow if you want to legally work with personal or sensitive data?

Answer:

(i)

The provided dataset excerpts do not contain any readily identifiable personal or sensitive data.

personal or sensitive data typically include the following:

Personal Identification Information (PII):

Examples: Full names, identification numbers (e.g., Social Security Number, passport numbers), contact details (e.g., phone numbers, email addresses), and physical addresses.

Reason: This data can identify an individual or could be used in combination with other information to identify them.

Sensitive Personal Data (as per GDPR or other privacy laws):

Examples: Racial or ethnic origin, political opinions, religious beliefs, trade union membership, genetic or biometric data, health-related information, and sexual orientation.

Reason: This type of data requires more stringent protection due to the risk of misuse.

Financial Data:

Examples: Bank account numbers, credit card numbers, transaction records.

Reason: This is considered sensitive because misuse can lead to financial fraud or identity theft.

Employee or Student Data (if relevant):

Examples: Job titles, performance reviews, student grades, or any data that can be linked to an individual’s employment or academic record.

Reason: This could be used for profiling, decision-making, or discrimination.

Health Data:

Examples: Medical records, health conditions, treatments, medications.

Reason: Sensitive health data is protected under laws like HIPAA in the U.S. or GDPR in the EU due to privacy concerns.

(ii)Processes for Legally Working with Personal or Sensitive Data:1.

Purpose Limitation and Data Minimization: Data should be collected for specified, explicit, and legitimate purposes, and only the necessary data should be collected.2.

Consent: Obtain unambiguous consent from individuals ("data subjects") before processing their personal data. Explicit consent is required for processing sensitive data.3.

Data Security: Implement appropriate technical and organizational measures to protect personal data from unauthorized access, use, disclosure, alteration, or destruction. This includes encryption, access controls, and data breach response plans.4.

Transparency: Inform individuals about how their data will be used and their rights under GDPR, including rights to access, rectification, erasure, and restriction of processing.5.

Data Protection Officer (DPO): Organizations handling significant amounts of personal data may be required to appoint a DPO to oversee data protection compliance.6.

Accountability: Organizations are accountable for demonstrating compliance with GDPR principles.7.

Pseudonymization and Anonymization: Techniques like pseudonymization (replacing identifiers with pseudonyms) and anonymization (irreversibly removing identifying information) can be used to reduce privacy risks, but they must be implemented effectively to be truly protective.

## [End of Question 3]

## QUESTION 4 [TOTAL MARKS: 25]

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

You are tasked with creating a visualisation showing “the distribution of expenditure by staff level across 4 quarters in 2019”. Use the sample data to create a graph to achieve this goal. You can use any tool to create the graph or you can sketch (pencil/paper) or mock up (e.g., powerpoint). Annotate or describe any specific design ideas. Paste your graph, sketch or mock-up diagram in the answer section below. Marks will be awarded for appropriate choice of graph, effective visualisation design and clarity of the message.

Data: Q4a-D - <https://loopexam.dcu.ie/mod/resource/view.php?id=8651>

Answer:Title: Distribution of Expenditure by Staff Level Across 4 Quarters in 2019

X-Axis: Quarters (Q1 2019, Q2 2019, Q3 2019, Q4 2019)

Y-Axis: Expenditure (in suitable units, e.g., thousands of euros)

Legend: Different colors will represent different staff levels (e.g., Entry Level, Mid-Level, Senior Level).

Design Ideas●

Stacking: Each bar will represent a quarter, and the segments within each bar will represent the expenditure for each staff level in that quarter. The height of each segment will correspond to the expenditure amount.●

Color Choice: Use a color palette that is visually appealing and easy to distinguish between staff levels. Consider using a gradient scale where darker colors represent higher staff levels.●

Data Labels: Display the expenditure values for each staff level within each bar segment for clarity.●

Axis Labels: Ensure the axis labels are clear and include the units of measurement.●

Title: Provide a concise and informative title that accurately reflects the chart's content.

Clarity of the Message

This visualization will clearly show:●

The overall expenditure trend across the four quarters.●

The proportion of expenditure attributed to each staff level in each quarter.●

Any significant shifts in expenditure patterns between staff levels over time.

This approach aligns with the principles of effective data visualization discussed in the sources, such as using suitable graph types, good design, and clear labeling to convey information effectively. However, the specific design choices, such as the exact color palette and data labeling format, would be refined based on the specific data values and the intended audience.

Please note that while the sources provide a theoretical framework for data visualization, they do not offer specific instructions or examples for creating this particular graph. You may want to consult data visualization tools or resources for specific implementation details.

### Q 4(b) [6 Marks]

Identify the specific type of each of the three graphs (A-C).

**A**

Chart, scatter chart

Description automatically generated

Answer:Graph A: This is a scatterplot, which belongs to the relational category of charts. Scatterplots show the relationship between quantitative values for two categories and are often used to show correlations. They can include a line of best fit or a trend line. Scatterplots utilize position to encode data values.

**B**

Chart

Description automatically generated

Answer:This is a stacked bar chart, which falls under the hierarchical chart category. Stacked bar charts are a variation of the classic bar graph, which is used for comparing categories and distributions of quantitative values. They present a breakdown of values within each bar to illustrate the part-to-whole relationship within a category while allowing for comparison between categories.

**C**

Chart, line chart

Description automatically generated

Answer:Slope Graph

A slope graph belongs to the temporal category of charts, emphasizing the change in quantitative values over time or across a specific sequence of intervals. Its primary function is to visualize the "before and after" changes in quantities for a set of variables. By connecting corresponding data points with lines, a slope graph highlights the direction and magnitude of change for each variable. The steeper the slope of the line, the greater the change. Slope graphs excel at showcasing comparisons of change across different categories and are particularly effective when the focus is on highlighting the differences in trends rather than precise values.

Pie Chart

A pie chart is a classic visualization that falls under the hierarchical category of charts. Despite its widespread recognition, sources caution against its overuse. Pie charts are designed to represent part-to-whole relationships, illustrating how individual quantities contribute to a total sum. They achieve this by dividing a circle into slices, with each slice's angle corresponding to the proportion of the total it represents. However, sources point out that pie charts are often misused and can be difficult to interpret accurately, particularly when dealing with many slices or subtle differences in proportions. While visually appealing, pie charts are best suited for datasets with a limited number of categories (ideally no more than 5) and clear differences in proportions.

Line Chart

A line chart is another classic visualization that falls under the temporal category of charts. Line charts are commonly used to visualize trends and changes in quantitative values over time, making them ideal for datasets where the x-axis represents a continuous time scale. By connecting data points with a line, line charts effectively display the overall pattern of change over time, allowing viewers to quickly identify trends, fluctuations, and significant points of change. Sources emphasize that line charts are particularly powerful for showing trends in time-series data and that care should be taken to ensure that the x-axis attribute represents a connected sequence.

Maps

Maps are a unique visualization type that falls under the spatial category of charts. They are used to represent spatial patterns and relationships by encoding data onto a geographic map. Sources explain that maps can use various techniques to encode data, including overlays and distortions. Overlays, like choropleth maps, use colors or patterns to represent quantitative values for different geographic regions. Distortions, on the other hand, manipulate the size or shape of geographic areas to reflect a specific variable, as seen in area cartograms. Sources emphasize the importance of choosing appropriate map projections to minimize distortion and accurately represent the data.

In conclusion, each chart type has its unique strengths and limitations. Selecting the most appropriate chart type depends on the nature of the data, the message you want to convey, and the intended audience. The CHRTS categorization system can guide this selection process, ensuring that the chosen visualization effectively communicates the insights within the data.

A bar chart, also known as a column chart when the bars are vertical, is a classic visualization technique used for comparing categories and distributions of quantitative values. It falls under the categorical chart category. Bar charts present data using rectangular bars, where the length or height of each bar corresponds to the value it represents. They are highly versatile and can be used to compare values across different categories or to show the distribution of values within a single category. Key characteristics of bar charts include:●

Comparison: Bar charts excel at enabling quick visual comparisons of values across different categories.●

Clear Representation: The use of bars makes it easy to perceive and compare the magnitudes of the values being represented.●

Versatility: Bar charts can be presented horizontally or vertically and can be further enhanced through variations like stacked or clustered arrangements.●

Widely understood: Bar charts are generally well-understood by a broad audience, making them effective communication tools.

Histogram

A histogram, often mistaken for a bar chart, serves a distinct purpose and falls under the distribution category within the categorical charts. Unlike bar charts that display categorical data on one axis and quantitative data on the other, histograms are specifically designed to visualize the frequency and distribution of continuous quantitative data. They achieve this by dividing the data range into a set of intervals or "bins" and representing the frequency of values falling within each bin using bars. The height of each bar corresponds to the number of data points within that bin. Key features of histograms include:●

Distribution Visualization: They effectively showcase the shape, center, and spread of a continuous data distribution.●

Binning: The use of bins enables the representation of the frequency distribution of continuous data.●

Focus on Frequency: Histograms emphasize the frequency of data points within specific intervals rather than individual data point values.●

Not for Categorical Data: Unlike bar charts, histograms are not suitable for representing categorical data.

Bubble Chart

A bubble chart is a type of relational chart, extending the capabilities of a scatterplot by incorporating a third quantitative value. It utilizes position and area to encode data values. Key aspects of bubble charts are:●

Three-Variable Representation: Bubble charts enable the visualization of the relationship between three quantitative variables.●

Position Encoding: The x and y coordinates of each bubble represent values for two variables.●

Area Encoding: The size of each bubble corresponds to the value of the third variable.●

Correlations and Clusters: Bubble charts can be used to identify correlations between variables and to visualize clusters of data points.

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

Given the scenario below choose the most appropriate graph type to visualise the message and justify your choice referencing the message and the probable data types and indicating the marks and visual attributes that will be used to encode the data.

“Compare the energy efficiency ratings (0-5 star) vs price (in Euro) for consumer electronics such as fridges, ovens, heaters, etc.”

Answer:

Data Types: The scenario involves two quantitative variables: energy efficiency rating (ordinal) and price (ratio). Scatterplots excel at illustrating relationships between two quantitative variables.●

Message: The goal is to "compare" the two variables, suggesting a need to visualize potential correlations or trends. Scatterplots effectively reveal such relationships.●

Marks and Visual Attributes:○

Marks: A scatter plot employs point marks, each representing a single consumer electronic item.○

Attributes:■

Position will be used to encode the data values. The x-axis will represent the energy efficiency rating (0-5 stars), and the y-axis will represent the price in Euro.■

Color (Hue) could be used to categorize the data points further, such as differentiating between types of consumer electronics (fridges, ovens, heaters, etc.).

A well-designed scatterplot will allow viewers to quickly discern any relationship between energy efficiency ratings and prices. For example, the plot might reveal if higher-rated appliances generally come with higher price tags. It would also effectively show outliers—appliances that deviate significantly from the general trend.

### Q 4(d) [3 Marks]

Which of the following images (A-D) illustrates the use of preattentive features?

Answer:The question asks which of the images (A-D) illustrates the use of preattentive features. Preattentive features are visual properties that are detected very rapidly by the human visual system (< 200-250ms). These features are immediately perceived and can sometimes mislead the viewer. When designing visualizations, preattentive features can be used to guide attention. Some examples of preattentive features are:●

Color●

Simple Shape = Orientation, Size●

Motion●

Depth

Let's look at each image (A-D) to determine which one utilizes preattentive features:●

Image A: This image shows multiple red squares on a grey background. It does not utilize any preattentive features as all of the shapes are the same color, size, and orientation.●

Image B: This image shows multiple red squares on a grey background with one red circle. The use of a different shape (circle vs square) is a preattentive feature.●

Image C: This image shows multiple grey circles with a blue border on a white background. It does not utilize any preattentive features as all of the shapes are the same color, size, and orientation.●

Image D: This image shows multiple grey circles and grey squares on a blue background. Some of the circles are black, and some of the squares are a lighter shade of grey. The use of different colors (black vs grey vs light grey) and different shapes (circle vs square) are preattentive features.

Therefore, the images that illustrate the use of preattentive features are B and D.

**A Icon

Description automatically generated B Icon, calendar

Description automatically generated**

**C Icon

Description automatically generated D Icon

Description automatically generated**

## [End of Question 4]

## QUESTION 5 [TOTAL MARKS: 25]

### Q 5(a) [6 Marks]

In the image below identify the **main** communication purpose and explain your reasoning, referring to specific design guidelines that support your choice.

Answer:

The main communication purpose of the image is persuasion. Persuasion is a graphic communication goal where information is used to elicit a specific response from the viewer. For example, advertisements use persuasive techniques to encourage viewers to purchase a product. This image aims to persuade the viewer that "Big Data" is a growing field with ample job opportunities and encourages the viewer to learn more by visiting the website.

Several design guidelines evident in the image support the identification of persuasion as the primary communication purpose:●

Appeal to Factual and Emotional Responses: The image combines factual information like job growth statistics and skills in demand with emotionally charged language like "revolution" and "competitive advantage". This approach appeals to both the viewer's rational side and their aspirations for career success.●

Emphasis on Key Information: The image utilizes size and bold typography to draw attention to specific elements, such as the number of data jobs available and the projected salary range. This technique helps ensure that the key persuasive messages are readily absorbed by the viewer.●

Call to Action: The image includes a website URL, directing viewers to a specific source for further information. This clear call to action is a hallmark of persuasive communication, encouraging viewers to take the next step towards engagement.

The combination of these design elements strongly suggests that the image's primary purpose is to persuade viewers about the importance and opportunities presented by "Big Data".

Main Communication Purpose: The main purpose of the infographic is persuasion, specifically to convince viewers of the growing importance of big data and data analytics as a career path.

Design Guidelines Supporting Persuasion:●

Appealing to Factual and Emotional Responses: The infographic combines factual data points, such as growth statistics and job market trends, with emotional appeals by using quotes that highlight the excitement and potential of a career in data analytics. This dual approach aims to resonate with both the rational and aspirational sides of the audience.●

Eye-Catching Visuals: The infographic employs illustrations, themes, and colour grouping to attract the viewer's attention and reinforce the message. The large, bold numbers and the visual representation of increasing data storage with growing hard drives serve to quickly convey the scale and significance of the topic.●

Targeted Audience Research: Although the specific research conducted is not described in the infographic, the design and language suggest a focus on students or young professionals considering career choices. The emphasis on job market growth, salary potential, and in-demand skills aligns with the interests and concerns of this target audience.

Example:

Information

The goal of an informative response is to provide clear, concise, and accurate facts without attempting to influence opinions or beliefs.

Example Prompt: What is the role of metadata in data management?

Example Response: Metadata is data that describes other data. It provides information about the content, context, and structure of data, making it easier to find, understand, and manage. For example, metadata for a photograph could include the date it was taken, the camera model, the location, and keywords describing the subject. Metadata plays a crucial role in various data management tasks, including:●

Discovery and Retrieval: Metadata allows users to search for specific data using relevant keywords or attributes.●

Data Understanding: Metadata helps users interpret the meaning and significance of data.●

Data Quality Management: Metadata can be used to track data provenance, assess data quality, and ensure consistency.●

Data Integration: Metadata facilitates combining data from different sources by providing a common understanding of the data's meaning and structure.

Persuasion

A persuasive response aims to influence the audience's opinion or actions by presenting arguments and evidence to support a specific viewpoint.

Example Prompt: Should we invest in a new data visualization tool?

Example Response: Investing in a new data visualization tool is crucial for our company's success. The current tools lack the advanced features and capabilities necessary to effectively communicate insights from our growing datasets. A new tool would empower us to:●

Uncover Hidden Patterns: Advanced visualization techniques can reveal hidden patterns and relationships in data that would be difficult to spot using traditional methods.●

Improve Decision-Making: Clear and insightful visualizations can facilitate better-informed and faster decision-making.●

Enhance Communication: Compelling visuals can effectively communicate complex data insights to stakeholders, fostering collaboration and buy-in.

Moreover, the cost of inaction is significant. By failing to invest in modern visualization tools, we risk:●

Missed Opportunities: We may miss crucial insights that could drive revenue growth or operational efficiency.●

Competitive Disadvantage: Our competitors who utilize advanced visualization tools will have an edge in extracting value from data.

Investing in a new data visualization tool is a strategic decision that will pay dividends in the long run.

Education

An educational response aims to teach the audience about a particular topic or skill, providing explanations, examples, and guidance to promote understanding and knowledge acquisition.

Example Prompt: Explain the different types of data and how they are used in visualization.

Example Response: Data can be classified into various types, each with its own characteristics and implications for visualization:●

Quantitative Data: Quantitative data represents measurable quantities. It can be further divided into discrete data (countable values, such as the number of students in a class) and continuous data (values within a range, such as height or weight). Quantitative data is often visualized using charts like bar graphs, line charts, and scatter plots.●

Qualitative Data: Qualitative data describes qualities or characteristics. It is often categorical, such as gender, eye color, or country of origin. Qualitative data can be visualized using charts like pie charts, bar graphs, and word clouds.

Understanding the different types of data is essential for selecting appropriate visualization techniques. For example, using a pie chart to represent continuous data would be misleading. Similarly, a bar graph may not be suitable for visualizing large amounts of categorical data. The choice of visualization should align with the data type and the message you want to convey.

Entertainment

An entertaining response seeks to engage the audience in a lighthearted or enjoyable manner, using humor, storytelling, or other engaging techniques to provide amusement or enjoyment.

Example Prompt: Can you share a funny anecdote about a data visualization mishap?

Example Response: I once worked on a project where we were analyzing customer feedback data. We created a beautiful visualization showcasing the most frequent keywords used in positive reviews. We were proud of our work and presented it to the stakeholders. However, to our embarrassment, we discovered that the most frequent keyword was "not"! It turned out that our data cleaning process had inadvertently removed the "not" from negative phrases like "not satisfied" and "not recommended." Our "positive" visualization was actually highlighting the most common complaints! We had a good laugh about it and learned a valuable lesson about the importance of careful data cleaning and interpretation.

Please note that this anecdote is not from the sources provided.

A picture containing text

Description automatically generated

### Q 5(b) [4 Marks]

For the image in Q5(a) identify the most probable Sender, Receiver, Message & Medium. Justify your answer, including any assumptions.

Answer:

Sender

Receiver

Message

Medium

Justification

Components of Graphic Communication:●

Sender: The most probable sender is the University of Maryland University College (UMUC), as indicated by their logo and call to action at the bottom of the infographic.●

Receiver: The intended receiver is likely students or young professionals who are considering future career paths, particularly those interested in data analytics.●

Message: The core message is that big data is revolutionizing industries, creating significant job opportunities and demanding specific skills, making it an attractive and lucrative career choice.●

Medium: The medium is a digital infographic, likely intended for online distribution through websites, social media, and other digital platforms.

Justification and Assumptions:

The presence of the UMUC logo and website address strongly suggests that they are the sender. The design and content, particularly the emphasis on career prospects, suggest a focus on prospective students or individuals early in their careers as the target audience. The infographic's format and the inclusion of digital contact information point to online dissemination as the intended medium.

Sender: The most probable sender is JESS3, a creative agency specializing in data visualisation and infographics. This assumption is based on the "Brought to you by JESS3" caption at the top of the image and the JESS3 logo prominently displayed at the bottom right corner.●

Receiver: The intended audience, or receiver, is likely the general public or individuals interested in technology and social media trends. The image's focus on popular social media platforms and the use of visually appealing graphics suggests it aims to communicate information to a broad audience rather than a specialized group.●

Message: The primary message conveyed is the vast reach and interconnectedness of social media in May 2011. The image highlights the large user base of various platforms, with "Cell Phones" positioned centrally as a connecting hub. This arrangement emphasizes the role of mobile technology in facilitating social media's widespread adoption.●

Medium: The medium is a static infographic, specifically a network diagram. The use of bubbles to represent different platforms and the connecting lines visualise the relationships and scale of the geosocial universe. The static nature of the image allows for detailed examination and analysis of the information presented.

Additional Assumptions and Justification:●

Persuasive Element: The image might also have a persuasive element, aiming to emphasise the significance of social media as a communication and marketing channel. By visually demonstrating the extensive reach of social media, JESS3 could be subtly promoting their services in data visualisation and communication to potential clients.●

Time Context: It's crucial to note that the data in the image is specific to May 2011. The social media landscape has significantly evolved since then, with new platforms emerging and existing ones experiencing substantial growth or decline. Therefore, the image serves as a historical snapshot of the geosocial universe at a particular point in time.

### 

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

Identify 2 possible problems with the graph below and suggest an alternative or improvement. Justify your choice, referencing design rules and theories.

Answer:

The graph presented in exhibits several design flaws that hinder its effectiveness in conveying information. Here are two key problems and suggestions for improvement:●

Problem 1: Excessive Line Density and Color Similarity: The graph displays a high density of lines representing different data series, which makes it challenging to distinguish individual trends. The color choices, being very similar in hue and saturation, further exacerbate this issue. According to the perceptual ranking in, color hue and saturation are relatively weak encoding attributes for quantitative data.○

Improvement: To address this, consider using a stream graph instead. Stream graphs are well-suited for displaying changes in quantitative values within different categories over time. They use flowing, organic shapes to represent the data series, making it easier to perceive overall trends and patterns, even with multiple categories. By carefully selecting a color palette with greater contrast, individual data series can be easily distinguished.○

The sources offer examples of good and bad visualizations, and this graph could be used as a teaching example to demonstrate the problems with poorly chosen encoding methods for different types of data, and how to choose a chart type to suit the data.●

Problem 2: Lack of Clear Labeling and Context: The graph lacks clear labels for the y-axis and the individual data series. Without proper labels, it is difficult for the viewer to understand what the values on the y-axis represent and what each line corresponds to. The sources repeatedly emphasize the importance of clear labeling and context in effective data visualization.○

Improvement:■

Provide a clear and descriptive title that summarizes the key takeaway from the visualization. For instance, instead of "Match 2 1st half - half-time - 2nd half," a title like "Decibel Levels During Match 2" would be more informative.■

Label the y-axis properly, including units of measurement. A label such as "Decibel Level (dB)" would make the data more meaningful.■

Consider adding a legend or directly labeling the lines to identify each data series. This will allow viewers to easily track the trends of specific data points.

The sources also highlight that effective data visualization involves choosing the most appropriate chart type for the data and question being addressed. In this case, a stream graph offers a more visually appealing and easily interpretable representation of the changes in decibel levels over time, particularly with multiple data series.

Please note that appears to be a superior version of this visualization from a different source, likely based on feedback given on. It includes a better title, and properly labeled axes.

Choosing the Wrong Chart Type

One of the most fundamental decisions in data visualisation is selecting the appropriate chart type. This choice should be driven by the nature of the data and the story you want to tell. Using the wrong chart type can make it difficult to see patterns, compare values, or understand the relationship between variables.

For example, if you want to show the distribution of a continuous variable, like age, a histogram would be a suitable choice. Using a pie chart for this purpose would be inappropriate, as pie charts are best suited for showing parts of a whole. The sources emphasize that selecting the right chart type is critical for creating a good visualisation.

Misleading Scaling and Axes

The scale and axes of a chart provide the context for interpreting the data. If these elements are manipulated or chosen poorly, it can distort the representation of the data and lead to misleading conclusions.

One common mistake is starting the y-axis at a value other than zero. This can exaggerate the differences between data points and create a false sense of dramatic change. For instance, if a chart showing company profits starts the y-axis at €50,000 instead of €0, a small increase in profit might appear much larger than it actually is.

Another potential issue is using a logarithmic scale when a linear scale is more appropriate. Logarithmic scales are useful for displaying data that spans several orders of magnitude, but they can be confusing for viewers who are not familiar with them. Using a logarithmic scale can also make small differences appear more significant than they are.

Clutter and Overcomplexity

A cluttered chart is like a noisy room – it’s hard to focus on the important information. Including too many data points, visual elements, or unnecessary details can overwhelm the viewer and obscure the message you are trying to convey. Excessive gridlines, data labels that overlap, and cluttered legends all contribute to a sense of visual chaos.

The sources advocate for a "data-to-ink ratio" approach, which means striving for clarity and simplicity by removing any pixels that are not directly related to the data. This principle encourages removing unnecessary elements and using visual encoding channels efficiently to present a clean and impactful visualisation.

3D Effects and Chart Junk

3D effects and decorative elements, often referred to as "chart junk," might seem visually appealing, but they rarely add anything of value to a data visualization. In fact, they can often hinder interpretation by distorting the perception of the data or distracting from the key insights.

A 3D pie chart, for example, might look impressive, but it makes it harder to compare the sizes of the slices accurately compared to a 2D version. Decorative elements like background images or unnecessary borders can also distract from the data and make the chart less effective. The sources emphasize the importance of clarity and elegance over unnecessary embellishments.

●

Misusing Visual Encoding Channels: Visual encoding channels such as colour, size, shape, and position should be used thoughtfully. Using the same channel to represent multiple data attributes can create confusion. For example, using colour to represent both a quantitative variable (like sales volume) and a categorical variable (like product type) in the same chart can obscure the relationship between the variables.●

Neglecting Data Relationships and Context: Effective visualisations not only present data points but also reveal relationships and provide context. A line chart showing monthly sales, for example, becomes more insightful when annotations are added to highlight key events (like promotions) or external factors that might influence sales.●

Ignoring Human Perception: Chart designers should consider how human vision works to avoid creating misleading or confusing visuals. People tend to underestimate differences in area, so using circles to represent quantitative values might lead to inaccurate perceptions.●

Overlooking Accessibility: Visualisations should be accessible to everyone, including those with disabilities. This involves using colour palettes that are colour-blind friendly and providing alternative encodings (like textures) for people who can't perceive colour differences.

Chart

Description automatically generated

### 

### Q 5(d) [7 Marks]

For the image in Q5(c), identify the visual attributes used to encode data and discuss how the image uses and directs attention, referring to preattentive features, gestalt theory, colour and other concepts as necessary. Marks will be given for correctly identifying all visual attributes and for noting two applications of attention concepts.

Answer:

Attributes:

Attention:

Visual Attributes and Attention Concepts●

Visual Attributes: The graph encodes data using the following visual attributes:○

Position: The horizontal axis represents time, while the vertical axis represents the measured value (likely decibel level, given the likely source).○

Colour (Hue): Different colours are used to represent each data series (Mic 5, Mic6, etc.).○

Line: Lines are used as the primary mark to connect data points over time.●

Attention Concepts: The graph uses and directs attention through the following concepts:○

Preattentive Features: The use of colour as a preattentive feature allows viewers to quickly distinguish the different data series, despite the similarity issue mentioned earlier.○

Gestalt Principles: The principle of similarity is unintentionally employed, as the lines with similar colours tend to group visually, making it harder to distinguish individual trends. This similarity hinders the desired separation and analysis of each data series.

Example:○

Points: Used as marks in a scatter plot to represent individual data points.○

Lines: Can be used to connect data points in a line chart, show trends, or indicate connections between entities.○

Area: The size of shapes (circles, squares) can encode quantitative values.○

Colour (Hue): Different colours can distinguish categories or groups.○

Colour (Saturation/Lightness): Variations in colour intensity can represent gradients or different levels within a category.○

Position: The location of marks on a chart (along x and y axes) encodes data values.

Discussing Attention:●

Explain how the image guides the viewer's eye and highlights specific information. This is where you link the visual attributes to attention concepts. Sources [35, 36, 39, 40, 41, 42, 43, 44, and 45] will help you understand these concepts.●

Two Key Applications of Attention Concepts:○

Preattentive Features: These are visual properties that the human visual system processes very rapidly and effortlessly. Examples include colour, size, orientation, and motion. Mention how the image leverages these features to make certain elements stand out immediately.○

Gestalt Theory: Gestalt principles describe how humans perceive visual elements as organized patterns or groups. Principles like proximity, similarity, closure, and continuity can be applied to data visualisation. Discuss how the image uses these principles to organise information and make it easier to understand.

Example Application:

Let's say the image is a bar chart showing sales figures for different product categories:●

Visual Attributes:○

Position: The height of each bar represents the sales amount.○

Colour (Hue): Different colours distinguish each product category.●

Attention:○

Preattentive Feature (Colour): The use of distinct colours for each bar allows viewers to rapidly identify and compare sales across categories.○

Gestalt Principle (Proximity): Bars belonging to the same time period might be placed closer together, visually grouping related data and aiding in understanding temporal trends.

These principles are particularly relevant in data visualisation, where the goal is to effectively communicate complex information through visual means. By understanding and applying Gestalt principles, you can create visualisations that are more intuitive, engaging, and easy to understand.

Here are some of the key Gestalt principles and how they can be applied in data visualisation:●

Proximity: Elements that are close together are perceived as belonging to the same group. This principle can be used to group related data points, create visual hierarchy, and guide the viewer's eye through the visualisation. For example, placing bars representing sales for the same month close together in a bar chart visually reinforces the temporal connection between those data points.●

Similarity: Items that share visual characteristics (such as shape, colour, or size) are perceived as being related. Using similar colours for data points belonging to the same category helps viewers quickly grasp the categorical groupings. Conversely, using similar colours for distinct data series can lead to confusion, as seen in the previous example where the yellow and orange lines were difficult to differentiate.●

Closure: Our brains tend to complete incomplete shapes and patterns, filling in the gaps to perceive a whole. This principle can be used to create visual interest and guide the viewer's eye to the most important information. For example, a line chart with a few missing data points can still be understood because our brains naturally connect the visible points to infer the overall trend.●

Continuity: We tend to perceive elements arranged on a line or curve as belonging together, even when they are interrupted. This principle can be applied to show trends and connections in data. A line chart effectively uses continuity to represent the evolution of a variable over time, allowing the viewer to follow the flow of the data.●

Figure/Ground: This principle describes our ability to distinguish between an object (figure) and its background (ground). In data visualisation, it's important to ensure that the data stands out from the background elements so that it is easily perceived. Using a clear contrast between data points and the background colour can prevent the data from getting lost visually.●

Symmetry: Symmetrical elements are perceived as more stable and visually appealing. Applying symmetry in chart layouts and design can create a sense of balance and harmony, making the visualisation more pleasant to look at.

By applying Gestalt principles, you can:●

Improve readability and clarity: Make your visualisations more intuitive and easy to understand.●

Highlight important information: Draw attention to key data points and relationships.●

Create visual hierarchy: Guide the viewer's eye through the information in a logical and meaningful way.●

Enhance engagement: Make your visualisations more visually appealing and memorable.

## [End of Question 5] [END OF EXAM]

**Distinguishing Between Bar Charts and Histograms**

While bar charts and histograms might appear similar at first glance, they serve distinct purposes and represent different types of data. Understanding these differences is crucial for selecting the appropriate visualization and interpreting the information correctly.

●

**Bar Chart:** A bar chart is used to **compare categorical data**, with the length of each bar representing the value or frequency of that category. Source1 introduces bar charts as a fundamental type of chart, highlighting their common use for comparing quantitative values across different categories. Bar charts are versatile and well-suited for a wide range of situations, effectively displaying comparisons between categories. Source2 emphasizes that bar charts are particularly useful when comparing quantitative values derived from distinct categories.

●

**Histogram:** A histogram, on the other hand, is specifically designed to **visualize the distribution of continuous data.** It groups continuous data into intervals or bins along the x-axis and uses bars to represent the frequency or count of data points within each bin. Source3 clarifies that a histogram focuses on displaying the frequency and distribution of data, emphasizing that it should not be confused with a bar chart. The y-axis in a histogram represents the frequency count, while the x-axis represents the bins or groups into which the data instances are divided.

**Key Differences**

|  |  |  |
| --- | --- | --- |
| Feature | Bar Chart | Histogram |
| **Data Type** | Categorical data | Continuous data |
| **X-axis** | Distinct categories | Intervals or bins of continuous data |
| **Y-axis** | Value or frequency of each category | Frequency or count of data points in each bin |
| **Purpose** | Comparing values across categories | Visualizing the distribution of data |
| **Spacing** | Bars typically have spaces between them | Bars are usually adjacent, without gaps |

**Example (Not from sources):**

●

**Bar Chart:** Comparing the sales of different car models (e.g., Toyota Corolla, Honda Civic, Ford Focus). The x-axis would list the car models (categories), and the y-axis would represent the number of cars sold.

●

**Histogram:** Visualizing the distribution of student heights in a class. The x-axis would represent height ranges (e.g., 150-155cm, 155-160cm, etc.), and the y-axis would show the number of students falling within each height range.

**Beware of Misinterpretation:**

Source3 highlights the importance of distinguishing between bar charts and histograms to avoid misinterpreting the data. Misinterpreting a histogram as a bar chart can lead to incorrect conclusions about the data's distribution and relationships.

**In Summary:**

●

Bar charts are for comparing categorical data, while histograms visualize the distribution of continuous data.

●

Understanding the differences between these chart types is essential for selecting the appropriate visualization and interpreting the information accurately.