## 

**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

### 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, Interval (because 0 time would make no sense)

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: A

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

Answer:

Volume: The amount of data generated by Netflix, including subscriber viewing history and the continuous flow of social media sentiment data, is substantial. The sheer volume of data being processed and analyzed is a key characteristic of big data.

Velocity: The data is generated and updated in real-time. Subscriber viewing data and social media sentiments change dynamically, contributing to a high velocity of data generation.

Variety: The situation involves diverse types of data, including viewing history and sentiment analysis. This variety in data types adds complexity and is another hallmark of big data.

Complexity: Analyzing and deriving meaningful insights from viewer data, coupled with sentiment analysis, introduces complexity. The need to process and understand this complex information aligns with the challenges posed by big data.

## [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)

**GDPR Compliance:**

How is the project ensuring compliance with GDPR regulations, particularly when dealing with personal data collected through surveys? Are there mechanisms in place to anonymize or pseudonymize the data to protect individual privacy?

**Data Retention and Deletion Policies:**

What are the data retention policies in place, and how is the project handling the deletion of data when it is no longer needed? GDPR emphasizes the importance of not retaining personal data for longer than necessary.

**Informed Consent and Transparency:**

How is the project ensuring that participants in surveys are adequately informed about how their data will be used? Are there transparent mechanisms in place to inform individuals about data processing activities and obtain their consent?

**Data Characteristics and Representations:**

Can you provide insights into the unique characteristics of the datasets involved (e.g., pedestrian footfall, traffic data, survey responses), and how these data types will be represented or visualized to facilitate meaningful analysis by government departments?

**Data Anonymization and Consent:**

How is the project addressing GDPR compliance by anonymizing or pseudonymizing personal data collected through surveys, and what mechanisms are in place to ensure informed consent from participants?

(ii)

Pedestrian Footfall Data:

Type: Historical footfall data from central shopping areas.

Structure: Likely numerical data, capturing the volume of pedestrians over time.

Storage: CSV or Excel for historical records, SQL database for efficient retrieval.

Traffic Monitoring Camera Records (Real-time Data):

Type: Real-time vehicle count data from main arterial routes.

Structure: Continuous stream of data capturing real-time traffic conditions.

Storage: JSON format for streaming, stored in a NoSQL or time-series database.

Survey Data on Working from Home Practices:

Type: Responses from surveys on working from home practices.

Structure: Mix of categorical and numerical data, potentially containing personal information.

Storage: CSV or Excel, with anonymization/pseudonymization in a structured database.

Weather and Economic Indicators:

Type: Various economic indicators, including business growth, median wage, and import/export figures, along with weather data.

Structure: Numerical and categorical data, capturing economic and weather conditions.

Storage: CSV or Excel for accessibility, structured database for efficient management.

Permission for Further Consumer Surveys:

Type: Consumer survey data as required by ongoing needs.

Structure: Will vary based on survey questions and responses.

Storage: Consistent with existing infrastructure, adapting to the specific needs of new surveys.

(iii)

**MySQL with Elasticsearch**

**Hybrid Database Approach**

**Reason**: This project involves a mix of structured data (such as economic indicators), semi-structured data (survey responses), and real-time streaming data (traffic records). A hybrid database model allows flexibility, enabling the use of relational databases for structured data, NoSQL databases for semi-structured and real-time data, and other specialized databases as needed. This approach provides an efficient and scalable solution tailored to the varied data types and requirements of the project, ensuring optimal performance and data management.

(i) Important questions to ask the client:

What specific aspects of working and commuting behavior are of most interest to the Irish Government? This will help focus the collection and analysis of data.

What time granularity is required for the analysis? Daily, weekly, or monthly data may be needed depending on the desired insights.

What level of detail is required for the analysis? Are aggregated or individual-level data sufficient?

(ii) Data and file formats:

Pedestrian footfall data: Likely to be in CSV format, with columns for timestamp, location, footfall count, weather data, and other relevant factors.

Traffic monitoring data: Likely to be in CSV or JSON format, with columns for timestamp, location, vehicle type, counts, and other relevant parameters.

Survey data: May be in SPSS, SAS, Excel, or other proprietary formats. Data cleaning and normalization may be required.

Weather data: Likely to be in CSV, JSON, or XML format, with columns for timestamp, location, temperature, precipitation, humidity, and other relevant weather parameters.

Standard economic data: May be in PDF, Excel, or CSV format. Data cleaning and conversion may be required.

(iii) Database storage approach:

Recommended: A relational database management system (RDBMS) like MySQL or PostgreSQL.

Reasons:

Structured Data: The data from various sources is structured and well-defined, making it suitable for a relational database.

Scalability: The system needs to handle large volumes of data and multiple concurrent users, which RDBMSs are well-equipped for.

Querying Flexibility: Government departments will need to access and query data for various purposes, and RDBMSs provide powerful query capabilities.

Data Integrity: Maintaining data integrity and consistency is crucial for the reliability of the analysis, and RDBMSs enforce data integrity constraints.

Standardization: RDBMSs are widely used and standardized, ensuring compatibility with existing tools and technologies.

### 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)

Unique Device Identifier (UDI): iPhone 39 Pro Max

Manufacturer: Apple Inc.

Model Number: AL2342

Operating System Version: iOS 57.2

IMEI (International Mobile Equipment Identity): 1234567899876

(ii)

UDI: Descriptive

Manufacturer: Administrative

Model Number: Descriptive

Operating System Version: Descriptive

IMEI (International Mobile Equipment Identity): Structural

(iii)

Using a standard for metadata would improve the quality of the data in several ways:

**Consistency**: A standard ensures that all devices are described using the same terms and definitions, which makes it easier to compare data from different sources.

**Accuracy**: A standard can help to ensure that the metadata is accurate and complete, as it provides clear guidelines for what data should be collected and how it should be formatted.

**Interoperability**: A standard can make it easier to share metadata between different systems and applications. This is important for research and analysis, as it allows researchers to combine data from multiple sources.

Potential Difficulty Enforcing a Metadata Standard:

Getting everyone to comply with a metadata standard can be challenging. Different organizations may have different needs and priorities, and they may not be willing to adopt a new standard if it does not meet their specific requirements. Additionally, enforcing a standard can be costly and time-consuming, as it may require training and updating systems.

In the case of CA682 students, it might be difficult to ensure that all students' devices are accurately described using the standardized metadata elements. Students may not be familiar with the standard, or they may not have the necessary tools or resources to collect the required data. Additionally, students may be reluctant to share sensitive information about their devices, such as their IMEI numbers.

### 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:

Description of the Process:

Scraping data from a website involves the automated extraction of information from web pages. This process typically includes sending HTTP requests to the target website, retrieving the HTML content, and then parsing and extracting relevant data. In a professional context, this might involve using specialized libraries or tools, such as BeautifulSoup or Scrapy in Python, to efficiently navigate and extract structured data from the HTML structure.

Rule 1

**Respect Website Policies:**

It is crucial to review and adhere to the terms of service, robots.txt file, and any other relevant policies outlined by the website. Some websites explicitly prohibit scraping in their terms of use, while others may have restrictions on the frequency of requests. Adhering to these policies is essential to maintain ethical practices and avoid legal consequences.

Rule 2

**Ethical Use and Attribution:**

When scraping data from a website, it is imperative to ensure ethical use of the extracted information. Give proper attribution to the source, and be mindful of the potential impact on the website's server load. Avoid overloading the server with excessive requests, as this could disrupt the normal functioning of the website and violate ethical guidelines.

## [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:

**Outliers in Age Groups**:

Issue: Unusual values or outliers in the age groups, such as extremely high or low ages.

Tool: You can use data visualization tools like Python's Matplotlib or Seaborn to create box plots or histograms for each age group and identify any data points that fall outside the expected range.

**Missing Values:**

Issue: Incomplete or missing data for specific road user categories or age groups.

Tool: Pandas, a data manipulation library in Python, can be used to check for missing values. Functions like isnull() and info() help identify the presence of missing data in the dataset.

**Inconsistent or Incorrect Categorization:**

Issue: Misclassification of road user types or age groups, leading to inaccuracies in statistics.

Tool: SQL queries or Pandas can be used to examine unique values in categorical columns. You can identify inconsistent or incorrectly labeled categories by checking distinct values in the relevant columns.

import pandas as pd

# Assuming 'dataset' is the DataFrame containing the road safety data

# Check for missing values

missing\_values = dataset.isnull().sum()

# Identify unique values in the 'road\_user' and 'age\_group' columns

unique\_road\_users = dataset['road\_user'].unique()

unique\_age\_groups = dataset['age\_group'].unique()

# Display results

print("Missing Values:")

print(missing\_values)

print("\nUnique Road Users:")

print(unique\_road\_users)

print("\nUnique Age Groups:")

print(unique\_age\_groups)

--------

# for outliers

import seaborn as sns

import matplotlib.pyplot as plt

sns.boxplot(x=your\_data)

plt.show()  
  
# or

from scipy.stats import zscore

z\_scores = zscore(your\_data)

outliers = (z\_scores > 3) | (z\_scores < -3)

### 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:

To address how each error or artifact may have been introduced in the context of the generic data analytics pipeline, we need to consider potential scenarios for each phase. Below are examples for each error and the likely phase of the pipeline where they could have been introduced, along with associated assumptions:

**Outliers in Age Groups:**

Introduction Phase: Gathering

Likely Scenario: Manual data entry errors during the collection of age-related statistics, where some values were entered inaccurately.

Assumption: Assumes that the data collection involved manual input or conversion of data, and there might have been human errors during this process.

**Missing Values:**

Introduction Phase: Gathering or Processing

Likely Scenario: Incomplete data submission during the data collection phase or issues in data extraction during processing, leading to missing values.

Assumption: Assumes that the data collection process may have had incomplete submissions, or there were challenges in processing data.

**Inconsistent or Incorrect Categorization:**

Introduction Phase: Gathering or Processing

Likely Scenario: Errors in classifying road users or age groups during data collection or issues in data transformation and preprocessing.

Assumption: Assumes that there might have been challenges in standardizing or categorizing data during the gathering or processing phases.

For each error, the assumptions are based on common challenges and potential pitfalls in the data analytics pipeline

### 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:

To avoid or mitigate the identified errors, we can implement specific data quality methods at different stages of the data analytics pipeline. Here are suggestions for each error and the associated methods:

**Outliers in Age Groups:**

Data Quality Method: Outlier Detection and Validation Checks

Explanation: Implement outlier detection algorithms during the gathering phase to identify and validate age values that deviate significantly from the expected range. This can involve setting realistic bounds for age values and rejecting entries falling outside these bounds. Additionally, manual validation checks can be incorporated during data entry to catch potential errors.

**Missing Values:**

Data Quality Method: Imputation Techniques

Explanation: Apply imputation methods during the processing phase to fill in missing values. Techniques such as mean imputation, regression imputation, or machine learning-based imputation can be employed based on the nature of the missing data. Imputation helps to maintain the integrity of the dataset by providing estimates for missing values while minimizing potential biases.

**Inconsistent or Incorrect Categorization:**

Data Quality Method: Standardization and Validation Checks

Explanation: Standardize the categorization of road users and age groups during the gathering or processing phase by implementing clear guidelines and validation checks. This involves using standardized coding schemes and conducting thorough checks to ensure consistent and accurate categorization. Implementing data validation rules helps maintain data consistency and integrity throughout the analytics pipeline.

By incorporating these data quality methods, we aim to enhance the overall quality and reliability of the dataset:

Accuracy: Outlier detection, imputation, and validation checks contribute to accurate and precise data by identifying and addressing errors in age values, missing entries, and inconsistent categorization.

Completeness: Imputation techniques ensure that missing values are filled in, enhancing the completeness of the dataset.

Consistency: Standardization and validation checks help maintain consistent categorization of road users and age groups, ensuring that data adheres to predefined rules and guidelines.

### 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)

(ii)

nderstand Applicable Regulations: Familiarize with relevant data protection laws.

Data Classification and Purpose Specification: Classify and specify the purpose for data collection and processing.

Obtain Consent: Obtain explicit and informed consent if required.

Implement Security Measures: Ensure robust security measures to protect data.

Data Minimization: Collect only necessary data for the intended purpose.

Data Subject Rights: Respect individuals' rights regarding their data.

Data Breach Response: Establish procedures for detecting and responding to data breaches.

Documentation and Accountability: Maintain documentation and establish accountability.

Regular Audits and Assessments: Conduct regular audits for ongoing compliance.

## [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:

A stacked bar chart here seems suitable

Distinct Colors: Use clear and distinct colors for each staff level.

Concise Legend: Include a concise legend for color reference.

Label Expenditure: Label each section with expenditure amounts.

Total Expenditure: Show total expenditure for each quarter.

Quarterly Trends: Add a line chart for quarterly expenditure trends.

Annotations: Use annotations for outliers or key points.

Clear Axis Labels: Clearly label the x-axis (quarters) and y-axis (expenditure).

Whitespace: Utilize whitespace for clarity and aesthetics.

-------

---------------------------------------------------

Data Distribution:

Use a histogram or a box plot to show the distribution of a continuous variable.

Use a bar chart for the distribution of a categorical variable.

Relationship between Variables:

Use a scatter plot to show the relationship between two continuous variables.

Use a line chart when illustrating trends or patterns over time.

Comparison:

Use a bar chart for comparing values of different categories.

Use a line chart for comparing trends over time.

Use a stacked bar chart or a grouped bar chart for comparing subcategories within categories.

Composition:

Use a pie chart for showing the composition of a whole in terms of parts.

Use a stacked bar chart for comparing parts of a whole across different categories.

Change Over Time:

Use a line chart for showing trends or changes over time.

Use a time series plot for visualizing time-dependent data.

Correlation or Regression:

Use a scatter plot for visualizing the correlation between two continuous variables.

Use a regression line or a bubble chart for illustrating regression relationships.

Magnitude:

Use a bar chart for showing the magnitude of different categories.

Use a bubble chart for displaying the magnitude of three variables (x-axis, y-axis, and bubble size).

Distribution of Frequencies:

Use a histogram for displaying the distribution of a continuous variable.

Use a frequency polygon to connect the midpoints of histogram bars.

Part-to-Whole Relationships:

Use a stacked bar chart or a pie chart for illustrating part-to-whole relationships.

Geographical Data:

Use a map for visualizing geographical data.

### 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: Scatter plot

**B**

Chart

Description automatically generated

Answer: Stacked bar chart

**C**

Chart, line chart

Description automatically generated

Answer: Slope Chart

### 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:

**Scatter Plot!**

**Justification:**

A scatter plot is the most suitable graph type for comparing the energy efficiency ratings (0-5 stars) against the price (in Euro) for consumer electronics like fridges, ovens, heaters, etc.

**Reference to the Message and Probable Data Types:**

**Energy Efficiency Ratings (0-5 stars):**

Data Type: Ordinal (categorical with a clear order)

Marks and Visual Attributes: Each point on the scatter plot represents a product, and the color or shape of the point can be used to encode the energy efficiency rating. For example, a color gradient from red (less efficient) to green (more efficient) or different shapes for each rating level.

**Price (in Euro):**

Data Type: Continuous (numeric)

Marks and Visual Attributes: The position on the y-axis represents the price, and the distance along the axis can be used to encode the price. Larger dots or different shades can be used to represent products in the same price range.

**Additional Points for Encodings:**

**Size of Points:**

The size of each point can be used to encode a third variable if applicable (e.g., product popularity or sales volume). Larger points could represent more popular products.

**Tooltip Information:**

Hover-over tooltips can be implemented to provide additional information about each data point, such as the specific model or brand of the consumer electronic product.

-------------------------------

Most appropriate graph type: Scatter plot

Justification:

Data types: The data is quantitative for both of the two variables: energy efficiency (0-5 star rating) and price (in euro).

Purpose: The graph is used to compare the two variables, so the marks should be placed on the axes. The energy efficiency rating should be represented on the x-axis and the price on the y-axis.

Visual attributes: The marks should be circles for each data point. The size of the circles can be used to encode additional information, such as the brand or model of the appliance. The color of the circles can also be used to encode additional information, such as the type of appliance (fridge, oven, heater, etc.).

Specific marks and visual attributes:

The x-axis will represent energy efficiency rating (0-5 star rating).

The y-axis will represent price (in euro).

The marks (circles) will represent each individual appliance.

The size of the marks will be proportional to the price of the appliance.

The color of the marks will be based on the type of appliance (red for fridges, blue for ovens, green for heaters, etc.).

The scatter plot is an appropriate graph type for this scenario because it allows us to easily compare the two variables and to see any patterns or trends in the data. The use of circles for the marks allows us to easily see the relative price of different appliances, and the use of different colors for the marks allows us to see the distribution of different types of appliances.

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

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

Answer: B

**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:

**Advertisement!**

The image is designed to persuade viewers to consider a career as a data analyst. It demonstrates the growing importance of skilled data analysts in the workplace and emphasizes the career opportunities that graduates of UMUC’s data analytics program can anticipate. Through the use of compelling language, relevant visuals, and a clear call to action, the poster successfully conveys its message and encourages viewers to take the next step in exploring the program.

**Key design elements that support the communication purpose are:**

“The Big Data Revolution Is Here” headline: The headline immediately captures attention with its bold and eye-catching font, and conveys the central message of the poster: that data analytics is a rapidly growing and important field.

“Become a Data Analyst” tagline: The tagline reinforces the call to action and emphasizes the career opportunities available to data analytics graduates. It is placed prominently below the main headline, ensuring viewers are aware of the desired outcome.

Upbeat and modern color scheme: The use of bright colors, such as blue, green, and orange, creates a visually appealing and engaging experience. The color palette aligns with the modern and forward-thinking nature of data analytics, further emphasizing the career opportunities associated with the field.

Use of relevant visuals: The poster features charts and graphics that are relevant to data analytics, such as bar charts and graphs, demonstrating the types of data that data analysts analyze. These visual elements enhance the credibility of the poster and reinforce the skills and knowledge required for this profession.

Prominent placement of the UMUC logo: The UMUC logo is strategically positioned at the top of the poster, making it the first thing viewers notice. This reinforces the connection between the poster and the UMUC data analytics program, establishing a sense of brand recognition and trust.

Clear and concise call to action: The call to action, “Visit umuc.edu/data to learn more,” is simple and direct, encouraging viewers to take the next step in exploring the program. It is placed at the bottom of the poster in a contrasting color, ensuring it stands out and captures attention.

Balanced and organized layout: The poster maintains a balanced and organized layout, with ample white space and consistent spacing between elements. This makes the poster easy to read and navigate, guiding the viewer’s focus on the key message.

Visual hierarchy: The poster utilizes visual hierarchy to ensure that the most important elements are emphasized. The headline, tagline, and call to action are larger and bolder than the body text, drawing attention to the desired message.

However, the advertisement for the UMUC's data analytics program is very small and in the bottom right corner of the poster. This is a common design tactic used in advertising. The goal is to get the viewer's attention with a large and eye-catching headline and visuals, and then to slowly introduce the advertising message. This helps to ensure that the viewer sees the advertising message, but it also helps to avoid overwhelming the viewer with too much information at once.

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

The most probable sender of the poster is the University of Maryland University College (UMUC), represented by the prominent placement of the UMUC logo at the top. The poster likely aims to promote the UMUC data analytics program, aligning with the call to action urging viewers to "Visit umuc.edu/data to learn more.

Receiver

The target audience for the poster is individuals seeking career opportunities in data analytics. The poster highlights the growing importance of data analytics and emphasizes the career prospects available to graduates of the UMUC program. The use of modern language and graphics suggests an appeal to a younger demographic, potentially students or professionals in the early stages of their careers.

Message

The central message of the poster is to encourage viewers to consider a career in data analytics, emphasizing the demand for skilled professionals in this field. The poster highlights the benefits of the UMUC data analytics program, including flexible online learning and promising career opportunities.

Medium

The medium for this communication is a printed poster, indicating a physical distribution method. The poster is likely displayed in areas frequented by potential students, such as college campuses, career fairs, and libraries. The poster's design is eye-catching and engaging, suitable for capturing attention in a public setting.

**Assumptions:**

We can assume that the sender, UMUC, has conducted market research to identify the desired audience for the poster, targeting individuals with an interest in data analytics and the potential to pursue higher education in this field. The poster's design aligns with the target audience, using modern language and visuals that appeal to a younger demographic. The placement of the poster in strategic locations suggests that UMUC has a clear understanding of where potential students and professionals might encounter the message.

### 

### 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:

Data viz checklist: <https://stephanieevergreen.com/wp-content/uploads/2016/10/DataVizChecklist_May2016.pdf>

1. “Misleading scaling” in the y-axis.

2. Display is free from decoration

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:

The image uses the following visual attributes to encode data:

Size: The size of the text "47,500 +" visually encodes the number of nurses being recruited.

Color: The color red is used to highlight the most important information in the image, such as the number of nurses being recruited and the headline "RECRUITING MORE NURSES."

Position: The headline "RECRUITING MORE NURSES" is positioned at the top of the image, where it is likely to be noticed first by viewers.

Spacing: The text in the image is spaced out evenly, making it easy to read and scan.

Attributes:

Attention:

**Size preattentive feature**- makes it easy for viewers to quickly identify the overall trend of the data, which is an increase in the number of nurses recruited over time.

Even though this is data manipulation

*The image uses the following attention concepts:*

Preattentive processing: The image uses preattentive features such as size, color, and position to grab viewers' attention. This is important because it ensures that viewers will notice the most important information in the image, even if they are only scanning it quickly.

Selective attention: The image uses selective attention to direct viewers' attention to the most important information. For example, the large red text "47,500 +" is likely to be the first thing that viewers notice in the image.

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