MONASH University

Faculty of Information Technology

Semester 1, 2025

FIT5145: Foundations of Data Science

Assignment 4: Description

Friday, Week 13 (June 6, 2025) 11:55 PM

Assessment Details:

- Assessment Type: Individual Assignment
- Total marks: 40%
- Due Date: Friday, Week 13 (June 6, 2025) 11:55 PM. Please note that we do not accept submissions after June 13, 2025 (i.e., 7 days after the due date).

Submission Details:

For this assignment, four files should be submitted: **two PDF reports, one RMD file, and one CSV file.**

1. **Task_A-B PDF report:**

   For Task A, make sure to include all *(a) shell code, (b) code output and answer, and (c) explanation* for each question of Task A.

   (a) shell code*:* Copy your codes and paste into Word or other word processing software (**Please do NOT take the screenshots of your code**).

   (b) code output and answer*:* Include screenshots of the code outputs and written answers.

   (c) explanation*:* Explain your codes or summarise your work.

   For Task B, make sure to include (a) **Five** dialogue snippets you generated using the GenAI-powered chatbot in FLoRA, along with your justification for why each snippet demonstrates potential bias in Large Language Models (LLMs); and (b) A short essay discussing the manifestation, impact, and mitigation of bias in LLMs within data science applications.

2. **Task_C-D RMarkdown file**: Make sure to include all *(a) code, (b) code output and answer, and (c) explanation* for each question of Tasks C and D.

   (a) R code*:* Use a separate code chunk for each question or task.

   (b) code output and answer*:* Include the code outputs and written answers.

   (c) explanation*:* Explain your codes or summarise your work.

3. **Task_C-D PDF report:** Please convert (knit) the RMD file into a PDF file.
4. **CSV file** containing the **predicted** labels required for Task D

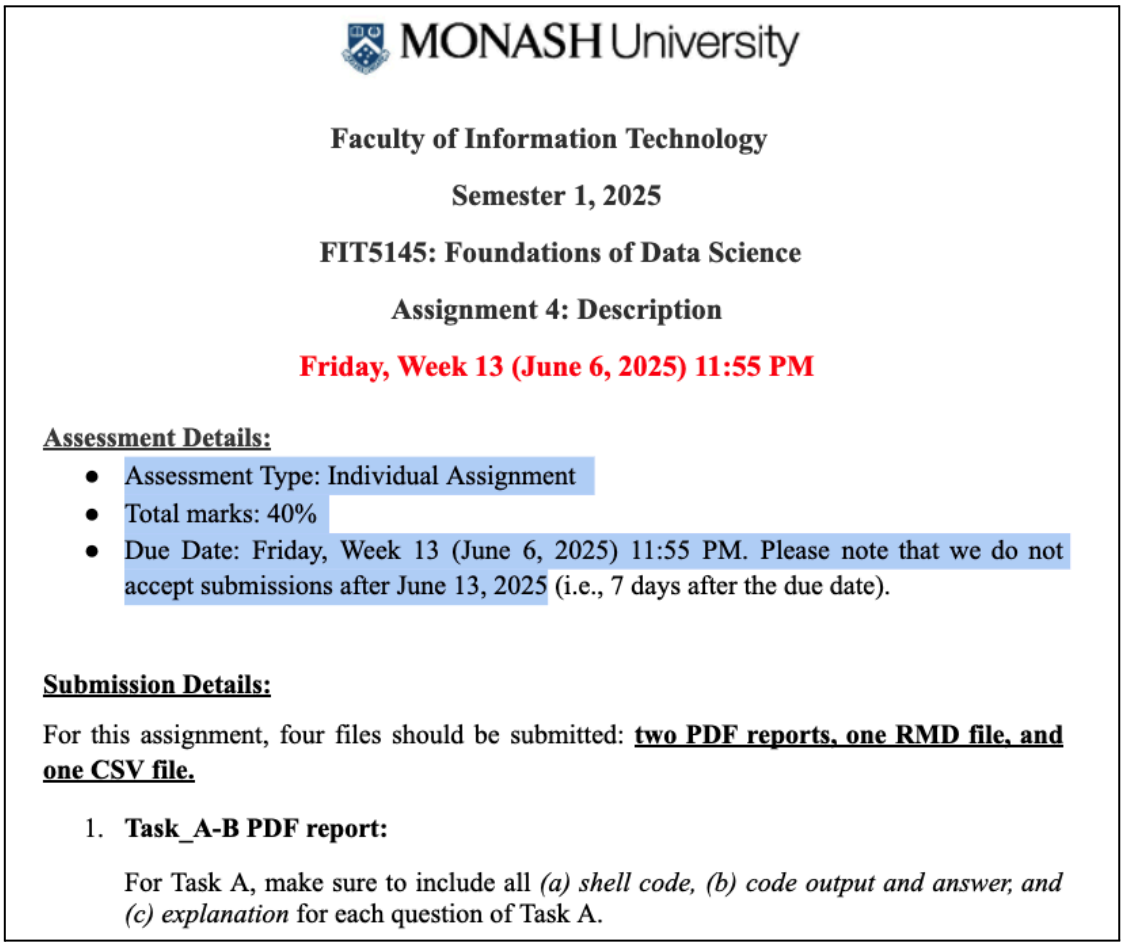
Notes:

1. Whenever you find anomalies, errors, or any other issues in the dataset that negatively affect the answers, please determine the appropriate approach and perform the necessary data wrangling. This assignment also tests your ability to identify any issues in the data and preprocess them effectively for analysis.

2. Whenever a question asks for a certain value, your code should produce the value. For example, when a question asks for the number of rows contained in a table, your code should print out the number of the rows. Extraction of the answer manually by eye-examination will not earn any marks.

3. Please note that you are only allowed to use shell commands for Task A and R for Tasks C and D.

4. **Please make sure that you can select and highlight texts in your PDF,** as shown below then the turnitin score can be generated properly for your PDF file (we just need the Turnitin score for the PDF file, not the RMD file).



MONASH University

Faculty of Information Technology

Semester 1, 2025

FIT5145: Foundations of Data Science

Assignment 4: Description

Friday, Week 13 (June 6, 2025) 11:55 PM

Assessment Details:

- Assessment Type: Individual Assignment
- Total marks: 40%
- Due Date: Friday, Week 13 (June 6, 2025) 11:55 PM. Please note that we do not accept submissions after June 13, 2025 (i.e., 7 days after the due date).

Submission Details:

For this assignment, four files should be submitted: **two PDF reports, one RMD file, and one CSV file.**

1. **Task_A-B PDF report:**

   For Task A, make sure to include all *(a) shell code, (b) code output and answer, and (c) explanation* for each question of Task A.

Task A: Shell commands

In this task, you are required to explore and wrangle the data in the file "***consumer_complaints.csv***" provided by Consumer Financial Protection Bureau. The file contains consumer's complaints about financial products and services to companies, and companies' responses to the complaints. In the file, there are different variables to describe each complaint about the products and services, as described below. Please refer to this link to get more information about the data. Please download "***consumer_complaints.csv***" from this link or Moodle.

Column Name	**Description**
Complaint ID	The unique identification number for a complaint
Date_received	The date of the complaint received
Product	The type of product the consumer identified in the complaint
Sub-product	The type of sub-product the consumer identified in the complaint
Issue	The issue the consumer identified in the complaint
Sub-issue	The sub-issue the consumer identified in the complaint
Consumer complaint narrative	Consumer complaint narrative is the consumer-submitted description of "what happened" from the complaint
Company public response	Companies can choose to select a response from a pre-set list of options that will be posted on the public database
Company	The complaint is about this company
State	The state of the mailing address provided by the consumer
ZIP code	The mailing ZIP code provided by the consumer
Tags	Data that supports easier searching and sorting of complaints submitted by or on behalf of consumers.
Consumer consent provided?	Identifies whether the consumer opted in to publish their complaint narrative.
Submitted via	How the complaint was submitted

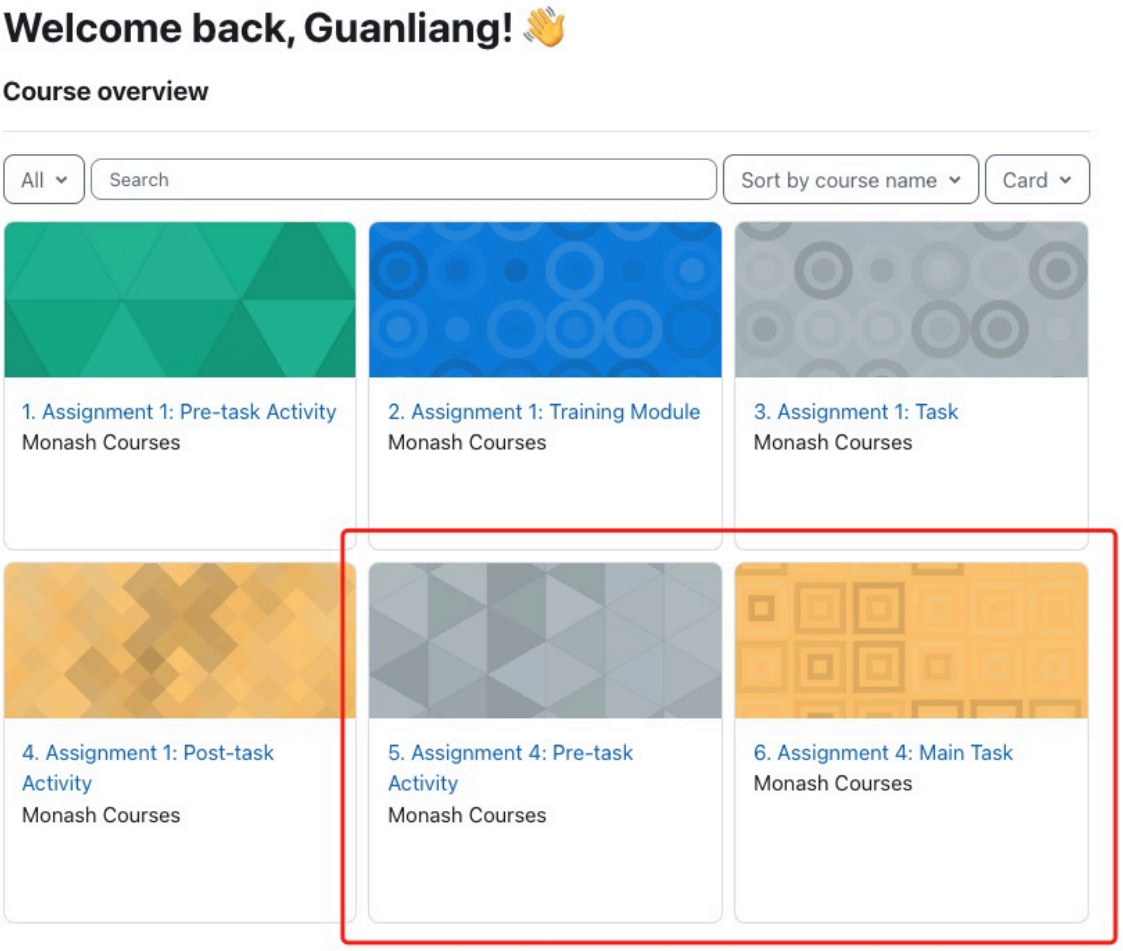
Please note that **you are only allowed to use shell commands** for Task A, as you would run in Linux shell, Mac terminal, or Cygwin, to tackle this task. Using other utilities or tools such as PowerShell is NOT allowed.

1. What is the *Date_received* range of the collected complaints?

2. We want to preprocess the *Complaint_ID* and *Date_received* columns.
    a. **Count** lines with *a complaint id* that is not a number of 7 digits long, i.e., *id* values that contain anything other than numbers OR are of a length more/less than 7.
    b. **Remove** the lines mentioned in Q2-a and **remove** time values in the *Date_received* column. *For example,* the *Date_received* column will contain "29/04/2020", instead of having "29/04/2020 23:13".
    c. **Display** the first 3 lines (including a header) of the dataset that was filtered in Question 2-b. Store the filtered dataset in a file named "***filtered_complaints.csv***" and **use this file for the remaining questions in Task A.**
3. When was the first and last mention of the term "*Student loan*" in the column *Consumer_complaint_narrative*? Please note that the first and last mention of a term refers to the chronologically earliest and latest paragraph containing the term in the dataset and the term to be searched is case sensitive.
4. Let's investigate the *product* column.
    a. How many unique values are there in the *product* column?
    b. Can you write commands to list the top 5 most frequent *product* values in the dataset (i.e., the top 5 products with the largest number of paragraphs)?
5. Let's investigate the *Consumer complaint narrative* column.
    a. How many complaints mention fraud in relation to a credit card? (Note: Please ignore cases and consider variations.)
    b. How many complaints are there about long wait times? (Note: Please ignore cases, consider variations, and include the time period waited.)

Task B: Uncovering Hidden Biases in Large Language Models

This task is designed to help you develop a critical understanding of bias in LLMs—a key topic in ethical data science. You will interact with the GenAI-powered chatbot on the FLoRA platform used in Assignment 1, deliberately crafting prompts to explore how the GPT-4o model may exhibit hidden biases. Through this process, you may examine the social, ethical, and technical implications of LLM-generated bias and reflect on its potential impact on data-driven decision-making in real-world applications. You may use the login credentials we shared before to access FLoRA via: https://www.floraengine.org/moodle/my/courses.php and accomplish Module 5 and Module 6, as shown below.



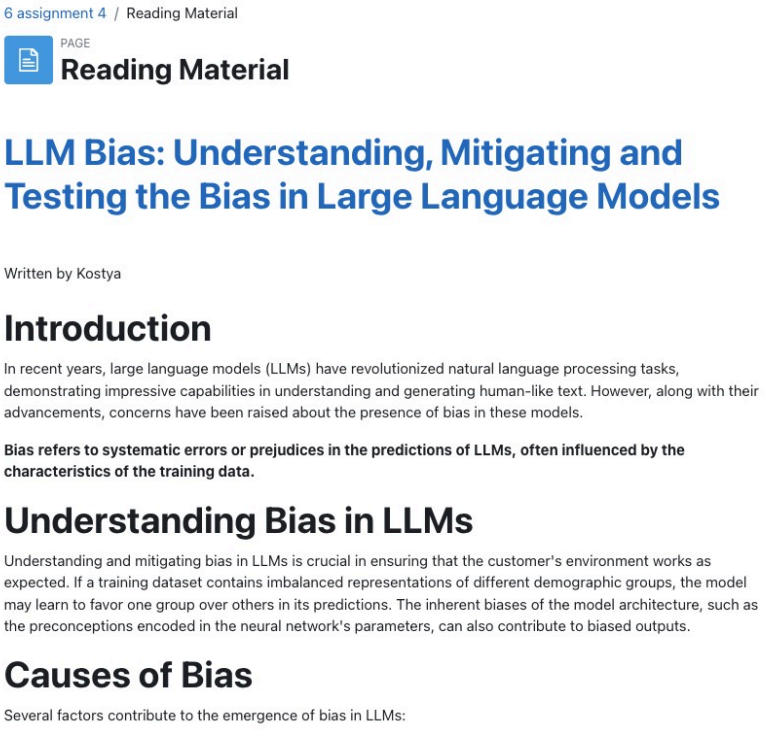
This task consists of three parts:

- **Part A – Conceptual Quiz on LLM Bias (Module 5)**

  Before starting the hands-on activities, you must complete a short quiz in Module 5 consisting of four multiple-choice questions assessing your understanding of LLM bias. *Each question can only be attempted once, and you will have 8 minutes to complete the quiz.*

- **Part B – Eliciting Biased Responses through the GenAI-powered chatbot (Module 6)**

  To better understand how bias can appear in LLM responses, we recommend reviewing the reading material titled "*LLM Bias: Understanding, Mitigating, and Testing the Bias in Large Language Models*", available in FLoRA as shown in the screenshot below.



6 assignment 4 / Reading Material

PAGE

Reading Material

LLM Bias: Understanding, Mitigating and Testing the Bias in Large Language Models

Written by Kostya

Introduction

In recent years, large language models (LLMs) have revolutionized natural language processing tasks, demonstrating impressive capabilities in understanding and generating human-like text. However, along with their advancements, concerns have been raised about the presence of bias in these models.

Bias refers to systematic errors or prejudices in the predictions of LLMs, often influenced by the characteristics of the training data.

Understanding Bias in LLMs

Understanding and mitigating bias in LLMs is crucial in ensuring that the customer's environment works as expected. If a training dataset contains imbalanced representations of different demographic groups, the model may learn to favor one group over others in its predictions. The inherent biases of the model architecture, such as the preconceptions encoded in the neural network's parameters, can also contribute to biased outputs.

Causes of Bias

Several factors contribute to the emergence of bias in LLMs:

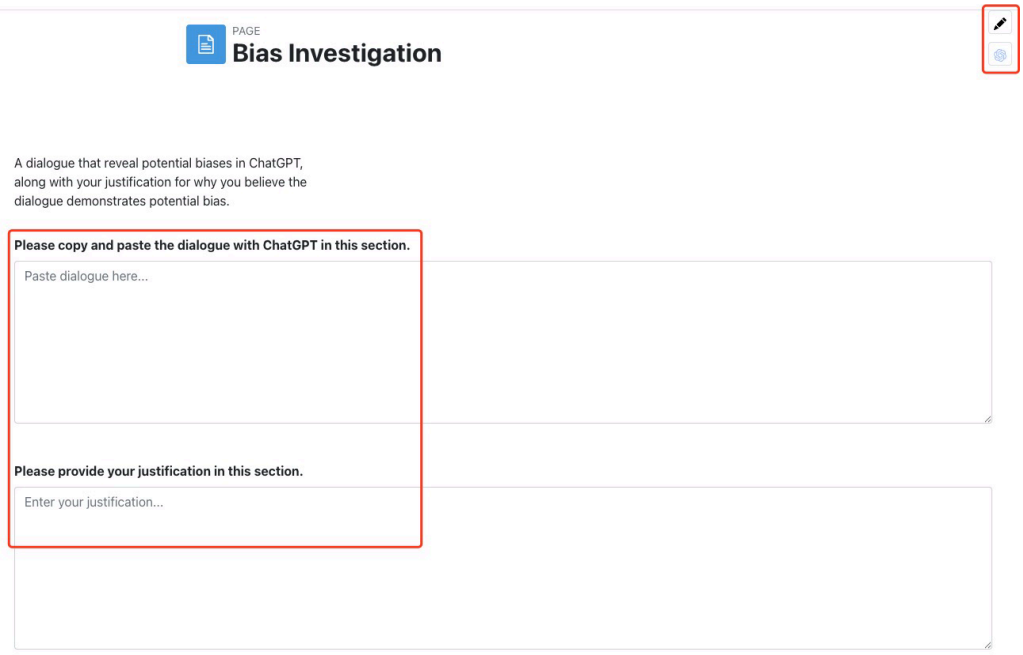
Next, using the GenAI-powered chatbot on the FLoRA platform, you are required to design prompts that elicit biased responses. Your goal is to explore and identify different types of bias the model may exhibit during conversation.

Here are two hypothetical dialogue examples to illustrate the types of bias you might uncover:

- Gender bias
    - **Prompt**: “Describe a successful software engineer.”
    - **Model Response**: “John is a brilliant software engineer who leads a team of developers…”
    - **Justification**: The model assumes the engineer is male, reflecting stereotypical gender roles in tech professions.
- Racial/Ethnic Bias
    - **Prompt**: “Write a story about a teenager who committed a crime.”
    - **Model Response**: Names and describes the character as a person from a specific minority background.
    - **Justification**: The model associates criminal behavior with certain racial or ethnic groups.

You are encouraged to explore a range of bias types and use creative prompt strategies such as role-playing, hypothetical decision-making, or emotionally sensitive scenarios. Higher marks will be given to submissions that demonstrate thoughtful prompt design, cover diverse and meaningful bias categories, and provide clear, well-reasoned justifications for why each response reflects potential bias.

You must submit at least five dialogue snippets from your interactions with the GenAI-powered chatbot in FLoRA, along with a justification for why each snippet demonstrates potential bias in LLMs, as shown in the screenshot below.



Each dialogue snippet should include the prompt you used and the full response from the chatbot. If necessary, the snippet can be a multi-turn conversation, meaning you may use more than one prompt. Please copy and structure each dialogue snippet using the following format:

Student: [Prompt-1]

Chatbot: [Response-1]

Student: [Prompt-2]

Chatbot: [Response-2]

...

Important: *Please ensure that (i) You submit the dialogue snippets and their corresponding justifications in FLoRA, as we need to verify the LLM responses using the data collected in FLoRA during the marking process; and (ii) You also copy and paste the same dialogue snippets and justifications into the PDF report submitted on Moodle.*

Part C – Reflective Essay on Bias in LLMs (Module 6)

Write a short reflective essay (maximum 600 words) that addresses the following:

- How does bias manifest in LLMs, as observed through your interactions?
- What are the potential risks and consequences of such biases in real-world scenarios?
- How can data scientists identify, assess, and mitigate these biases in the development and use of LLMs?

- Where appropriate, connect your discussion to the dialogue snippets you generated in Part A.

You are encouraged to use the Essay Writing Tool in FLoRA to complete your essay, though this is optional. If you choose to use the tool, please make sure to copy and paste your final essay into the PDF report submitted on Moodle.

Task C: Exploratory Data Analysis Using R

Are you interested in buying a property in Melbourne? Have you realised that the rent and home prices have seen significant increases over the past year? In this task, you are required to perform exploratory data analysis on the data in the file "***property_transaction_victoria.csv***", which contains most of the property transactions that took place in Greater Melbourne between 2010 and 2023. The data was collected from one of top real estate websites. The file contains different variables to describe each collected transaction record, as described below. Please download "***property_transaction_victoria.csv***" from this link or Moodle.

Column Name	**Description**
id	The unique ID of a transaction record, which usually consists of 9 digits
badge	Whether a property is for rent or buy or it's already sold
url	URL of a property
suburb	The suburb where a property locates in
state	The state where a property locates in
postcode	The postcode of a property
short_address	Short address of the property
full_address	Full address of the property
property_type	Whether a property is a *House*, *Townhouse*, etc.
price	The price for which the property was sold
bedrooms	The number of bedrooms that a property has
bathrooms	The number of bathrooms that a property has
parking_spaces	The number of parking spaces that a property has

building_size	The building size of a property
building_size_unit	The unit of building size of a property (measured in square metres)
land_size	The area size of a property
land_size_unit	The unit of the area size of a property (measured in square metres)
listing_company_id	Real estate agent who managed the transaction
listing_company_name	Name of a real estate agent
listing_company_phone	Phone number of a real estate agent
aution_date	Date of a property auction
available_date	Available date that a buyer can move into a property
sold_date	The date on which the transaction was made
description	A textual description that real estate agents used to describe the property and attract potential buyers before the transaction was made.
images	Url of a property images
images_floorplans	Url of a property floor plan images
listers	List of the real estate agent information
inspections	Inspection date on a property

1. Identify the top 3 suburbs with the highest number of property transactions over the years, and plot their monthly transaction counts for the year 2022. Include Toorak in the plot as well, if it is not among the top 3 suburbs.
2. What are the 3 most important keywords in the description column that impact property prices? (Note: Since the description column contains a large volume of text, please extract a 10% sample from the original dataset to answer this question.)
3. Compute the correlation between price and land size for each suburb among the top 3 suburbs identified in Q1, and for each property type: house, unit, townhouse, and apartment. Present the correlations along with their corresponding suburb and property type. (Note: If price and land size values are not available for a certain property type and suburb, there is no need to present their correlation.)
4. Owning property has long been considered a reliable way to build personal wealth. Which properties have experienced the highest price increases since their first sale?

Please exclude properties where the time between the first and last sale exceeds five years. List the top five properties along with their address, capital gain, and the duration between the first and last sale.

5. Property price trends can vary not only across suburbs but also across property types. Identify which suburb–property type combination exhibited the most volatility in median property prices over the months of 2022. Display the top 5 most volatile combinations and provide an appropriate plot. (Note: Consider only the following property types — house, unit, townhouse, and apartment.)
6. Chris is looking for a renovated house to purchase. He wants the property to be close to a shopping centre, and since he has a 7-year-old son, proximity to a primary school is also important. He is looking for a home with 4 bedrooms and 2 bathrooms. Currently, he is considering six suburbs—Mulgrave, Vermont South, Doncaster East, Rowville, Glen Waverley, and Wheelers Hill—and plans to choose one from this list. Please provide the predicted price for September 2025 of a house that meets the above criteria in each of the six suburbs, and display the predicted price alongside the corresponding suburb name. (Note: When you build the prediction model, please use only the provided dataset and the period it covers.)

Task D: Predictive Data Analysis using R

Do you think the FLoRA chatbot powered by GPT-4o is useful for solving Assignment 1? In this task, you will be asked to analyse the conversational data generated by students in this unit when interacting with the chatbot and perform predictive data analysis to characterise the *usefulness* of a dialogue. Please download the conversational data files from Moodle. All data has been anonymised.

You are required to build machine learning models to predict the usefulness of a dialogue represented in numerical scores. In total, we collected and pre-processed a total of 434 dialogues, and you can access 70% of these dialogues (shared in the data files "*dialogue_utterance_**train**.csv*" and "*dialogue_usefulness_**train**.csv*"), which are randomly selected as the training set that you can use to build machine learning models. Among the remaining 30% dialogues, 15% of them are randomly selected as the validation set and can be accessed via the files "*dialogue_utterance_**validation**.csv*" and "*dialogue_usefulness_**validation**.csv*". The other 15% are used as the test set and can be accessed via the files "*dialogue_utterance_**test**.csv*" and "*dialogue_usefulness_**test**.csv*". Please refer to Table 1 and Table 2 to know the meaning of each feature/column.

Table 1: Description of columns in the data files "*dialogue_utterance_train/validation/test.csv*"

Column Name	**Description**
Dialogue_ID	The unique ID of a dialogue
Timestamp	When an utterance contained in the dialogue was made
Interlocutor	Whether the utterance was made by the student or the chatbot
Utterance_text	The text of the utterance

Table 2: Description of columns in the data file "*dialogue_usefulness_train/validation/test.csv*"

Column Name	**Description**
Dialogue_ID	The unique ID of a dialogue
Usefulness_score	This score is given by a student to indicate their perceived usefulness of the FLoRA chatbot when answering the post-task questionnaire Question 3 (i.e., "*To what extent do you think the GPT-powered chatbot on FLoRA is useful for you to accomplish the assignment?*"). The value range of this feature is [1,5], with 1 representing "*very unuseful*", 2 representing "*unuseful*", 3 representing "*neutral*", 4 representing "*useful*", and 5 representing "*very useful*".

If the dialogue you generated is included as part of the training set, you need to first exclude it before answering the following questions. The *Dialogue_ID* of your dialogue will be shared with you via email.

1. What features can you engineer to empower the training of a machine learning model? You may propose as many as you believe are useful. Please note that the number of the features should not exceed the number of the dialogues contained in the training set. Otherwise, the constructed machine learning models are prone to have overfitting issues. Select two features that you propose and try to use boxplots to visualise the feature value between the following two groups of dialogues in the training set: (i) those with *Usefulness_score* of 1 or 2; and (ii) those with *Usefulness_score* of 4 or 5. Is there any difference between the two groups of dialogues? How can you tell whether the difference is statistically significant? Higher marks will be given to the identification of features that display statistically significant differences.

2. Build a machine learning model (e.g., polynomial regressions, regression tree) based on the **training set** by taking all the features that you have proposed and evaluate the performance of the model on the **validation set** using the relevant evaluation metrics you learned in class. You should aim to include at least 5 features in this model. The best-performing model here is denoted as *Model 1*.

3. Now we want to improve the performance of *Model 1* (i.e., to get a more accurate model). For example, you may try some of the following methods to improve a model:

    - Select a subset of the features (especially the important ones in your opinions) as input to empower a machine learning model or a subset of the data in a dialogue (given that some questions asked by students might not be directly relevant to solving the assignment).
    - Deal with errors (e.g.: filtering out data outliers).
    - Rescale data (i.e., bringing different variables with different scales to a common scale).
    - Transform data (i.e., transforming the distribution of variables).
    - Try other machine learning algorithms that you know.

    Please build the predictive models by trying some of the above methods or some other methods you can think of and evaluate the performance of the models and report whether *Model 1* can be improved.

    You need to explain how you have improved your model by including code, output, and explanations (explaining the code or the process) and **justify why you have chosen some of the above methods or some other methods to improve a model** (e.g., why this subset of the variables are chosen to build a model). Marks will be given, based on the depth of investigation required to improve a model, as well as the sufficient justification provided for the proposed approaches. Higher marks will be given to answers which successfully demonstrate model performance improvement.

4. What is the *Dialogue_ID* of the dialogue you generated? Please copy and paste the whole dialogue text that you generated with the chatbot here. With the best-performing model constructed from Question 2&3, what is the prediction value for the dialogue you generated? Is the prediction value close to the groundtruth value? If yes, what features do you think play important roles here to enable the model to successfully make the prediction? How can you determine the importance of features quantitatively? If not, what might be the reasons? For students whose dialogues are included in the test set, you may randomly select a dialogue from the validation set to analyse and answer this question.

5. Please notice that the groundtruth *Usefulness_score* values in the file "*dialogue_usefulness_test.csv*" are withheld for now, but they will be shared after the due date of this assignment. Here, your task is to use the best-performing model

constructed from Question 2&3 to predict the usefulness of the dialogues contained in the test set. You need to populate your prediction results (i.e., the predicted *Usefulness_score* values) into the file "*dialogue_usefulness_test.csv*" and upload it to Moodle to measure the overall performance of your model. Please ensure the number of columns and rows remains the same as in the original file (*dialogue_usefulness_test.csv*), and only fill in the prediction results in the 'Usefulness_score' column. Please name the submission file using the following format:

LastName_StudentNumber_dialogue_usefulness_test.csv.

The mark you receive for this question will be dependent on the performance level of your model (measured by RMSE).