MONASH University

Faculty of Information Technology

Semester 2, 2024

FIT5145: Foundations of Data Science

Assignments 1 & 3: Business and Data Case Study

1. **Assignment 1 (Proposal)**: Write a proposal document to introduce a data science project that you are studying. The due date of Assignment 1 is: **Friday, 23 August 2024, 11:55 PM (Week 5).**

2. **Assignment 3 (Report+Presentation):** Write a report on your case study of the data science project and prepare a 4-minute presentation on your project. The due date of Assignment 3 (the final project report and presentation slides) is: **Friday, 11 October 2024, 11:55 PM (Week 11)** and the presentation will be held **in the Week 12 applied class.**

- Both Assignment 1 and 3 are **individual** assignments.
- **Please do NOT zip your submission files.** Zip file submission will have a penalty of 20% of the total mark of the assignment.

Focus of the study

This case study needs to analyse ("study") a data science project relevant to any of the following business scenarios ("the case"): **agriculture, education, finance, gaming industry, healthcare, social media, and sports**. There are a couple of ways you can choose to do the case study. For instance, you may choose to study *how an existing data science project has been implemented in these business scenarios*. However, please notice that the project chosen is **NOT** limited to those already established or completed. *You are highly recommended to propose an entirely new and novel project of your own, e.g., how can the Australian government tackle bushfire or flood by making use of Twitter data?* That being said, you can either study an existing data science project or propose a completely new data science project. Talk to your tutors about any proposed project you are interested in.

Assignment 1: Proposal (15%)

Weight: 15% of the unit mark

Submission format: one PDF file

Size: up to 700 words.

What you need to do:

- Choose a data science project.
- Write the initial two sections, *Project Description* and *Business Model* (*References* as well to support your project) of the report, as detailed in the specification of *Assignment 3: Report + Feedback + Presentation* below. This assignment is worth 15% of your unit mark.

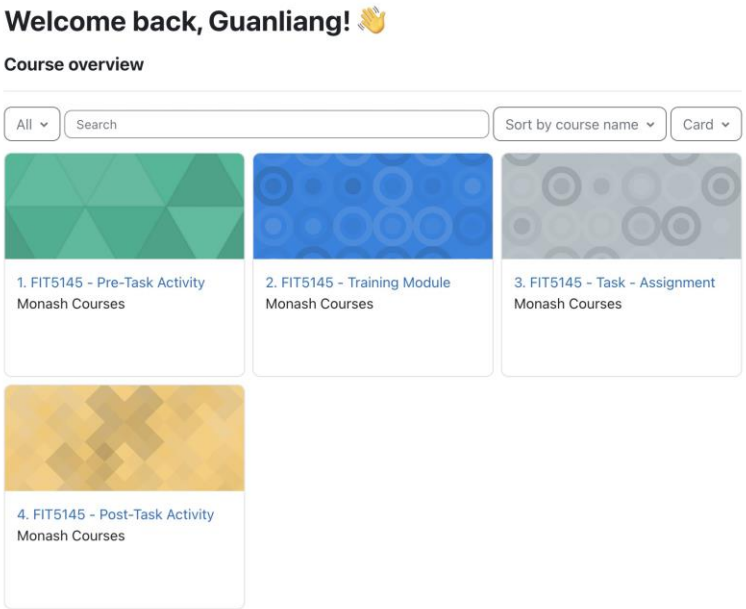
We have developed a system named FLoRA to support you to accomplish Assignment 1, which you may access via: https://www.floraengine.org/moodle/my/courses.php. You are expected to work with the

chatbot embedded in FLoRA, powered by cutting-edge Generative Artificial Intelligence (GAI) technology GPT-4o, to accomplish the assignment. You can discuss the assignment requirements with the GAI-powered chatbot, seek suggestions for potential project topics in a specific domain, and gather information relevant to a specific project topic by conversing with the chatbot. Even more, you may send the proposal draft to the chatbot and seek feedback for further improvement.

Please notice:

- We will send you the login credentials via emails for you to access the FLoRA platform.
- You are only expected to use FLoRA to accomplish Assignment 1, though you may use it for Assignment 3 as well, it is not mandatory.
- The conversational data you generate when interacting with the GAI-powered chatbot will be used for textual analysis in **Assignment 4** in this unit. That is, the conversational data will be shared with the whole class. Therefore, please notice the following:
    - Please **only** discuss your data science project with the GAI-powered chatbot **in English** and do not ask any questions that are irrelevant to your project;
    - Please do not disclose any personal or sensitive information when interacting with the GAI-powered chatbot.
    - There are **four modules** in FLoRA and you are required to complete all of them. ***Any incomplete activities in these modules will result in that your conversational data with the chatbot will not be included in the dataset used for Assignment 4 and you will not be able to answer some of the questions in Assignment 4.***

After logging to the platform, you will see there are four modules required to accomplish Assignment 1, as shown below:



Module 1: Pre-task activity

- Please provide information about yourself as well as your prior knowledge and experience in data science and GAI.

Module 2: Training Module

- We provide a set of tutorial documents to help you familiarise with FLoRA, including:
    - *The system interface*;

    - *The annotation tool*, which you can use to make annotation to the reading materials provide to help you get some inspirations about potential project ideas before discussing with the GAI-empowered chatbot;
    - *The essay writing tool*, which you can use to draft the assignment;
    - *The GAI-empowered chatbot*, which you can consult for help when solving the assignment. The Chatbot uses the GPT-4o model.
- Please notice that all these tools have been enabled in Module 2 (available on the top right corner) and you may familiarise yourself with these tools first before moving to Module 3 to start working on the assignment.

Module 3: Task - Assignment

- This is the main module in which you are expected to accomplish Assignment 1. Before discussing with the GAI-powered chatbot, please first have a look at the "inspiring" materials that may give you some initial ideas of what data science can achieve in the domains where data science is playing an increasingly important role:
    - Data Science in Agriculture
    - Data Science in Education
    - Data Science in Finance
    - Data Science in Gaming Industry
    - Data Science in Healthcare
    - Data Science in Social Media
    - Data Science in Sports

  You may use the provided annotation tool to annotate the useful information in these materials.
- After selecting the domain that you would like to work on, use the GAI-powered chatbot to get necessary help for accomplishing the assignment (e.g., seeking relevant information about a specific topic in the selected domain).
- After you finish the draft, please (i) click the "Save Essay" button to send your submission to FLoRA; and (ii) copy your project text and paste it into a word processing tool (e.g., Microsoft Word), format it if necessary, and then save it as a PDF file and submit it on Moodle as well.
- ***Please notice that the conversational data you generate with the GAI-powered chatbot will be used and shared for Assignment 4 and thus do not disclose any personal or sensitive information to the chatbot. Your project proposals or any other personal information will NOT be used or shared for Assignment 4.***
- As we will use the conversational data for Assignment 4, ideally you have one "meaningful" discussion session with the GAI-powered chatbot (instead of having multiple at different times) in Module 3 to get the help you need to accomplish Assignment 1. Prior to this, you may familiarise yourself with the chatbot (and other tools as well) in Module 2.

Module 4: Post-Task Activity

- Please share your experience in using FloRA as well as the GAI-powered chatbot to tackle Assignment 1. Your responses to these survey questions will be **mandatory** for including your conversational data to prepare an anonymised conversational dataset for Assignment 4.

For any technical issues in using FLoRA, please contact guanliang.chen@monash.edu and xinyu.li1@monash.edu

Assignment 3: Report (15%) + Feedback (5%) + Presentation (10%)

1. Assignment 3: Report

Weight: 15% of the unit mark

Submission format: one PDF file and one RMD file (for demonstration in the Characterising and Analysing Data section)

Size: up to 2500 words

This report is your analysis of how data science can be used to help solve a particular problem. In your report you need to identify the size and scope of both the problem and the data science project, as well as the requirements of enabling the project.

Please answer the following question in the FIRST page of your Assignment 3 submission:

- Have you selected a topic for Assignment 3 that is different from the one that you used for Assignment 1 (i.e., have you rewrote the first two sections of the report)?

Your report should have at least (but not limited to) the following sections:

- **Project Description**: provide a description about the data science project that you study/propose, what the objective of the project is, and what data science roles (e.g., data scientist, data engineer, system architect) are involved in this project and what are their responsibilities.
- **Business Model**: provide analysis about the business/application area the project sits in, what kind of benefits or values the project can create for the specific business area and who can benefit from, and what the challenges of the project are.
- **Characterising and Analysing Data**:
  - Discuss potential sources to collect the data, provide analysis about the characteristics of the data (e.g., the 4 V's), provide analysis on the required platforms, software, and tools for data processing and storage, according to the specific data characteristics.
  - Specify/propose the data analysis and the statistical methods (e.g., decision tree and regression tree) used in the project, provide analysis on why you choose those methods and discuss the high-level output (Note: Specifying and proposing the data analysis and statistical methods is different from the demonstration below and must be described separately).
  - Demonstration: identify a usable dataset for the proposed project and perform some basic analysis on the identified dataset to demonstrate the feasibility of the project, using R (e.g., detailing the information/features contained in the dataset, analyse the basic characteristics of the dataset, etc.), and report the analysis process and result in the demonstration section of a final report.
    Note: Students will get more marks if they use realistic data. If the realistic data is not available, please create a mockup/example dataset to clearly explain the proposition, modelling approach, and visualisations that can be derived from it. (please include a link to download the data in the final report, and upload the R markdown file created for data analysis on Moodle).
- **Standard for Data Science Process, Data Governance and Management:** describe any standard used in your data science process, and describe appropriate practices for data governance and management in the project, e.g., issues related to the accessibility, security, and confidentiality of the data as well as potential ethical concerns with the use of the data.

The sections would present aspects of Weeks 1-10 of the unit for your chosen case study.

The **maximum word limit** for the report (Assignment 3) is **2500** words. It may include some/all of your

Assignment 1, modified if needed (counted in the 2500 word total). References at the end of the report (i.e., URLs and academic publications) are not included in the word count. Note that staying within the word limit demonstrates your ability to write concisely.

2. Assignment 3: Feedback from Assignment 1

Weight: 5% of the unit mark

You need to attend the Week 7 applied class to seek feedback from your tutors regarding your Assignment 1.

Please include the following in the SECOND page of your Assignment 3 submission:

- What is the feedback given by your tutor for Assignment 1?
- Please briefly describe how you act upon the provided feedback to prepare Assignment 3 (no more than 150 words)

3. Assignment 3: Presentation (Slides + Verbal) + Peer-review Evaluation

Weight: 10% of the unit mark

Submission format: one PDF file (Slides)

Size: a maximum of 10 slides (Slides)

You need to submit your presentation slides along with your final report. The 4 minute presentation is given in Week 12 during your assigned applied class and after your presentation, the tutor will ask at least one question to the presenter (1 minute). You will also be required to review and provide feedback on presentations of other students (peer-review) during the applied class in Week 12, using the Google Form provided.

How you will be assessed

Assignment 1 proposal: The 15% awarded for your proposal is broken down into the following categories:

- clear description of the goals of the project;
- appropriateness of topic;
- clear description of the business benefits and challenges ;
- novelty/creativity (*this one is of extreme importance!*);
- overall clarity of the initial report.

Assignment 3 report: See the marking rubric to understand how we will grade your report. You will be assessed on your ability to:

- provide a description about the data science project that you study/propose, what the objective of the project is, and what data science roles (e.g., data scientist, data engineer, system architect) are involved in this project and what their responsibilities are;
- provide analysis about the business/application area the project sits in, what kind of benefits or values the project can create for the specific business area and who can benefit from, and what the challenges of the project are;
- discuss potential sources to collect the data, provide analysis about the characteristics of the data (e.g., the 4 V's), provide analysis on the required platforms, software, and tools for data processing and storage, according to the specific data characteristics;

- specify/propose the data analysis and statistical methods used in the project, provide analysis on why you choose those methods and discuss the high-level output (Note: Specifying and proposing the data analysis and statistical methods is different from the demonstration below and must be described separately);
- identify a usable dataset for the proposed project and perform some basic analysis on the identified dataset to demonstrate the feasibility of the project, using R (e.g., detailing the information/features contained in the dataset, analyse the basic characteristics of the dataset, etc.), and report the analysis process and result in the demonstration section of a final report;
  Note: Students will get more marks if they use the realistic data. If the realistic data is not available, please create a mockup/example dataset to clearly explain the proposition, modelling approach, and visualisations that can be derived from it. (please include a link to download the data in the final report, and upload the R markdown file created for data analysis on Moodle).
- describe any standard used in your data science process, and describe appropriate practices for data governance and management in the project, e.g., issues related to the accessibility, security, and confidentiality of the data as well as potential ethical concerns with the use of the data.
- think critically and creatively, providing justification and analysis;
- provide a good quality of report in terms of structure, expression, grammar and spelling.

For both assignments, make sure that any resources you use are acknowledged in your report. You may need to review the FIT citation style to make yourself familiar with appropriate citing and referencing for this assessment. Also, review the demystifying citing and referencing guide for help.

Please also make sure that the Turnitin scores will be generated properly for your submissions. If a submission receives a high Turnitin score (e.g., more than 15%), the student will likely need to provide further explanation on the project idea and a penalty might be imposed on the submission in case no proper justification is provided.

Assignment 3 Presentation (Slides + Verbal Presentation + Peer-review Evaluation): The 10% awarded is broken down into the following categories:

- Presentation (Slides + Verbal) (6%)
- Peer-review evaluation (4%)

What you need to do

Before you begin, make sure you:

- You are highly recommended to review the "inspiring" materials provided in FLoRA to select a topic that you would like to work on. Also, you are highly recommended to propose your own interesting and novel topic and please feel free to discuss it with your tutors to ensure the topic is suitable.
- Download the **marking rubric** (available on Moodle) as guidance on how you will be assessed.

Choose a data science project as a case, and then:

- **Do preliminary research** about your case and the relevant technologies by conversing with the GAI-powered chatbot
- **Write and submit your proposal** with cited references (Assignment 1)
- **Research and prepare your final report** with cited references.
- **Submit your report and do a presentation** (Assignment 3).

You are free to modify the initial proposal sections submitted for Assignment 1 (especially in response to feedback from your marker), or even change topics, when you are working on Assignment 3.

How to Submit

Once you have completed your work, take the following steps to submit your work. Penalties may be applied to your marks if the following instructions are not followed.

1. For Assignment 1, please finish the project proposal first in FLoRA, copy & paste it into a word processing tool (e.g., Microsoft Word) for the purposes of structuring/formatting or including visuals if necessary, then save the project proposal in the PDF format and submit it on Moodle.
2. Please ensure you **name the file containing your proposal/report/slides** correctly using the following format:
   LastName_StudentNumber_AssignmentNumber(_report or Rmd or slides).pdf
   e.g., Guanliang_12345678_Assignment1.pdf or Guanliang_12345678_Assignment3_report.pdf or Guanliang_12345678_Assignment3_Rmd.Rmd or Guanliang_12345678_Assignment3_slides.pdf
3. Upload your assignment file in the corresponding assignment link provided on Moodle.

Those unable to attend week 12 applied class for presentation:

Those who cannot do the presentation in their original applied class and would like to attend another applied class to do the presentation, please first contact the tutors of the class that you would like to attend to check whether the tutors have additional capacity to accommodate you and if yes, then you can join the class for presentation. We cannot guarantee that you will be admitted to another applied class for presentation if you cannot attend your original class due to the limited capacity the teaching team has.

Those who cannot do the presentation in any applied classes due to valid mitigating circumstances, can record and submit a video 4 minutes duration, through Youtube along with slides. **Please notice that, if you choose to do a video presentation, that means you will lose the 3% verbal presentation as well as the 4% peer-review evaluation in Assignment 3, i.e., only a maximum of 3% will be given for Presentation (Slides + Verbal Presentation + Peer-review Evaluation) if you fail to give a verbal presentation.** We recommend you produce a video by doing a 4 minute screen capture of your slides with voice over entered concurrently via microphone. You must upload your video through Youtube (make it unlisted, not private), and provide a LINK to your Youtube video in your submission (on the first page of the presentation slides) on Moodle. However, DO NOT include your video on Moodle as part of the submission. Please check the details of this, and confirm with a lecturer and your tutor a week in advance of presentation week.

Further advice on the assignment:

Here is some further advice from the teaching team regarding the assignment:

1. Make sure to carefully read the assignment specification above.
2. The project should be data-centred -- ideally combining multiple sources of studies to develop your own project that can solve a real-world problem.
3. The project should contain a clear statement of the problem being tackled. What is the objective/purpose of the project? Have a look at the structure of the example case studies, each one starts with a clear definition of the problem.
4. Make sure that the benefit of the project is clear. What is it? Will the project have a financial benefit, or result in a social good?
5. The report needs to be "telling a story", and to be convincing somebody to "invest in your project" so that it can be built.
6. Try not to make the project too broad. It should be an achievable data science project.

7. For both Assignment 1 and Assignment 3, you are highly recommended to include visuals to explain your idea more clearly and support your claims and in lesser words. For example, depending on your project it may (or may not) make sense to include:
    - an influence diagram showing what data is available and how it relates to the decisions and objectives of the project,
    - graphs showing some data analysis (if applicable).

    For Assignment 1, you may add the visuals after you copy and paste the project proposal into a word processing tool (e.g., Microsoft Word).
8. Read up as much as you can on the particular topic you've chosen in order to be able to describe the data (and software) requirements of the project.
9. Make it clear where the data would come from for the project:
    - Is the data proprietary? How would it be collected?
    - If the data is public, you should do some exploratory data analysis on it.
10. What preprocessing would be needed? How would the data need to be preprocessed before it can be used? What software might be needed? Can the preprocessing be distributed?
11. Finally, make sure you've seen the set of possible section headings suggested above and structure your report accordingly.