Assignment Autumn 2025

COMP7024 Programming for Data Science

Due Friday 30 May 2025 (Week 13)

1 Project Description

In this assignment there are 4 parts. For each part you should:

- Explain your rationales behind choices made answering the tasks.
- Write the appropriate R code.
- Include comments within the code to explain the algorithm.
- Test the code to ensure its correctness.
- Format and structure the code to maximise its readability.

A report must be submitted containing a cover page, the solutions to each of the four parts, and your code, as a **PDF**, to the vUWS submission site. The cover page must contain your name, student number, subject code and name, and the declaration below.

Submissions in other formats or without cover pages will have marks deducted.

Submission is due by Friday of week 13. Late submissions will receive a 10% reduction in marks for each day late.

2 Marking Criteria

This assignment is worth 40% of the subject assessment tasks. There four problems to investigate and 10 marks available for each of the four problems. In addition, there is 10 marks for using of GIT in the assignment. Therefore, the **total marks** for assignment is **50**. The marking criteria for each question is given in Table 2.

Criteria	Q1	Q2	Q3	Q4
Rationales of algorithm choices (1 mark) Code Correctness (4 marks) Comments explaining code (2 marks) Code Testing (1 mark) Code Style and Readability (2 marks)				
Total (10 marks)				

Table 1: Marking criteria for each part of this project.

For GIT, the following marking criteria will be used:

Criteria	GIT
Basic GIT command (2 marks) An appropriate number of Commit and Commit messages (3 marks) Advanced GIT (3 marks) Extra GIT (2 marks)	
Total (10 marks)	

Table 2: Marking criteria for GIT part of this project.

When writing the solutions to each of the four parts, make sure to consult the marking criteria and check that you have covered them. The project will be marked using this criteria.

For each task, there are a maximum of 1 bonus marks available for answers above and beyond the subject content. For example, extra analysis. And, your codes are expected to be kept on GIT.

3 Declaration

Before submitting the assignment, include the following declaration in a clearly visible and readable place on the cover page of your project report.

***

By including this statement, we the authors of this work, verify that:

- We hold a copy of this assignment that we can produce if the original is lost or damaged.
- We hereby certify that no part of this assignment/product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.
- No part of this assignment/product has been written/produced for us by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.
- We are aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism (**which may retain a copy on its database for future plagiarism checking**).
- We hereby certify that we have read and understand what the School of Computing, Engineering and Mathematics defines as minor and substantial breaches of misconduct as outlined in the learning guide for this unit.

***

Note: An examiner or lecturer/tutor has the right not to mark this project report if the above declaration has not been added to the cover of the report.

4 Project Tasks

You are working at the RyneThomas consulting firm as a data scientist and analyst. You are tasked to analyse the user behavior and engagement from online community. The data set is contained in three different sets:

User: User information:

- `user_id` — the unique identifier of the user.
- `name` – First name of the user.
- `review_count` – Number of reviews user submitted.
- `average_stars` – User's average rating across all reviews.
- `member_since` – The date the user join the platform.

Businesses: Business information.

- `business_id` – the unique identifier of the business.
- `name` – the name of the business.
- `city` – the city of the business.

- `state` – the state of the business.
- `business_avg_stars` – Average star rating for the businesses.
- `review_count` – Total number of reviews received.
- `categories` – Business categories or tags.
- `business_group` – Randomly assigned Group A or Group B for comparison.

Reviews: Reviews of the businesses by the users.

- `review_id` – the unique identifier of the review.
- `user_id` – The user_id of the user posted the review.
- `business_id` – The ID of the business being reviewed.
- `stars` – Star rating given by the user.
- `date` – When the review was done.
- `text` – Full text content of the review.

Your tasks are:

1. Write the code to analyse the review behaviour across user groups. The users should be grouped into 3 group: Veteran, Intermediate and New (based on their member_since date) before 2017, between 2017-2022, and after 2022 respectively. Calculate the numbers of users, their average review stars and average number of reviews per user. Tabulate the data using kable or kableextra. Visualise the Average Review Stars by User Age Group. You are required to make sure you handle the NA value in your analysis. Explain your findings.

2. Write the code to analyse the average reviews star by State. Calculate the average review star, the number of reviews and the number of unique users. Visualise the Unique users by State. You are required to make sure you take care of the NA value in your analysis. You are required to carry out a seperate analysis for both datasets (BusinessPGA and BusinessPGB). Elaborate on the finding. Is there any difference between the 2 datasets?

3. Write the code to analyse the top users and their behaviours. First, identify the top 10 users by the review count. For those top 10 users, calculate their average review stars. Tabulate the summary of the data (kable/kableextra). You are required to make sure you handle the NA value in your analysis. Visualise their rating distrubtion using ggplot2 - boxplot. Discuss your findings.

4. Write the code to analyse if there is a major difference between the review behavior of users who joined before and after 2020. For these 2 groups of users, compare their star rating behaviour and the length of the reviews (number of charaters in the review text). You are required to make sure you handle the NA value in your analysis. Visualise the average review length by the two groups. Discuss your findings.

Write a PDF report containing your code and all required analysis and results. The report is being marked using the marking criteria, so make sure that each piece of analysis covers all of the criteria. Please also ensure the use of GIT as your repo.