Project Autumn 2024

COMP1013 Analytics Programming

Due Friday 31 May 2024 (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 1.

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.

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 2 bonus marks available for answers above and beyond the subject content. For example, extra analysis.

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 Ryne-Hanson Health Research Foundation Centre as a data scientist and analyst. You are tasked to analyse the health data based on demography, cost, gender and age groups. The data set is contained in three different sets:

patients: Patient demographic data:

- `id` – the unique identifier of the patient.
- `BirthDate` – The date the patient was born.
- `DeathDate` – The date the patient died. An empty field indicates the patient is still alive.
- `Marital` – The patient's marital status. M - Married and S - Single.
- `Race` – Description of the patient's primary race.
- `Gender` – Gender. M is male, and F is female.
- `City` – The city of the patient's current address.
- `State` – The state of the patient's current address.
- `County` – The county of the patient's current address.
- `Zip` – The postal code for the patient.

Encounters: Patient encounter data.

- `id` – the unique identifier of the encounter
- `Start` – The date and time the encounter started.
- `stop` – The date and time the encounter concluded.
- `Patient` – The patient ID for the patient who went to health services
- `EncounterClass` – The type of encounter, such as, ambulatory, emergency, inpatient, wellness, or urgent care.
- `Code` – The Encounter code using the Health Standard coding of SNOMED-CT (More info at https://www.snomed.org/).
- `Description` – Description of the type of encounter.
- `Base_Encounter_Cost` – The base cost of the encounter, not including any line item costs related to medications, immunisations, procedures and or other services.
- `Total_Claim_Cost` – The total cost of the encounter, including all line items.
- `Payer_Coverage` – The amount of cost the Payer covers.

- ReasonCode – The Diagnosis code from SNOMED-CT, if this encounter targeted a specific condition.
- ReasonDescription – Description of the reason code.

conditions: Patient conditions or diagnoses.

- Start – The date the condition was diagnosed.
- Stop – The date the condition resolved, if applicable.
- Patient – Patient ID for the diagnosed patient.
- Encounter – Encounter ID to map the encounter details for this patient.
- Code – Diagnosis code from SNOMED-CT.
- Description – The description of the diagnosis/condition.

Your tasks are:

1. Write the code to analyse the distribution of COVID patients (confirmed or suspected) across counties. Write the code to investigate the distribution of the patients across age groups (e.g., 0-18, 19-35, 36-50, 51+). Visualise both the findings using the histogram. Explain your findings.

2. Filter those patients in the dataset that have contracted COVID-19 or Suspected COVID-19; ; what are the top 10 most common conditions (symptoms) related to the patients? Do the conditions differ between genders? Provide a table to rank the top 10 conditions for male and female patients separately. Elaborate on the findings.

3. Write the code to analyse the factors that might influence the hospitalisation rate (ambulatory, emergency, inpatient, urgent care) for the COVID patient (confirmed or suspected) in the dataset. Any factors in the dataset, such as age, gender, zip code, marital status, race and county, can be considered. Pick 2 of the factors and explain if there is a trend that explains the variation.

4. Write the code to investigate the characteristics of patients (confirmed or suspected) who recover from COVID-19 compared to those who don't. Consider factors such as demographics (age, gender, zip code), symptoms, and timeline of diagnosis and recovery. Analyse how these factors impact the recovery outcome.

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.