Project Trimester 2 2025

COMP1013 Analytics Programming

Due Friday 14 November 2025 (Week 11)

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 maximize 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. Apart from submitting it to the vUWS submission system, you must also submit your work on GitHub. 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 11. 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	Q5
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 (3 marks)	
An appropriate number of Commit and Commit messages (3 marks)	
Advanced GIT (4 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 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 Vietnam Vehicle Maintenance Centre as a data scientist and analyst. You are tasked with analyzing vehicle data based on engine parameters, car specifications, and various diagnostic reports. The dataset is divided into three different sets:

Engines: Engine data:

- **EngineModel** - the unique identifier (ID) of the engine.
- **EngineTypes** - the type of the engine (dohc, dohcv, l, ohc, ohcf, ohcv, rotor).
- **NumCylinders** - number of cylinders (eight, five, four, six, three, twelve, two).
- **EngineSize** - engine capacity (continuous from 60 to 320).
- **FuelSystem** - fuel supply to the engine (1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi).
- **Horsepower** - power output of an engine (continuous from 48 to 288).
- **FuelType** - the kind of fuel a vehicle uses to power its engine (diesel, gas).
- **Aspiration** - how an engine draws in air for the combustion process (std, turbo).

Automobile: Car data.

- **PlateNumber** - the unique identifier (ID) of the car
- **Manufactures** (Toyota, Honda, Audi, Bmw, Chevrolet,…).
- **BodyStyles** - shape, structure, and design of a vehicle (hardtop, wagon, sedan, hatchback, convertible).
- **DriveWheels** - the wheels that receive power from the engine to move the vehicle (4wd, fwd, rwd).
- **EngineLocation** - position of the engine within a vehicle (front, rear).
- **Wheel-base** - distance between the front and rear axles (continuous from 86.6 120.9).
- **Length** - length of a vehicle (continuous from 141.1 to 208.1).
- **Width** - width of a vehicle (continuous from 60.3 to 72.3).
- **Height** - height of a vehicle (continuous from 47.8 to 59.8).
- **CurbWeight** - total weight of a vehicle without passengers or cargo (continuous from 1488 to 4066).

- **EngineModel** - ID of an engine that a vehicle uses
- **CityMpg** - miles Per Gallon in the City (continuous from 10 to 50).
- **HighwayMpg** - miles Per Gallon on the Highway (continuous from 15 to 55)

Maintenance: Vehicle conditions or diagnoses.

- **ID** - the unique identity of the encounter
- **PlateNumber** - the unique identifier (ID) of the car.
- **Date** - vehicle entry date.
- **Troubles** - diagnosis description
- **ErrorCodes** - 0 if there is no error, 1 if the engine fails, and -1 if any other vehicle component fails.
- **Methods** - maintenance methods (Urgent care, Adjustment, Replacement, NA (no error))
- **Price** - maintenance fee.

Your tasks are:

1. Write the code to inspect the data structure and present the data: The missing values in the dataset were written as "?", replace any "?" with NA; Write code to check: after replacing, how many rows were affected in total? Does this change alter the data distribution?; Convert categorical variables **BodyStyles, FuelTypes, ErrorCodes** to factors; Replace the missing values in column **Horsepower** with the median horsepower; Select the appropriate chart type and display: horsepower distribution.
2. Write the code to analyse the distribution of the horsepower across the engine types. Write the code to investigate the distribution of the horsepower across the groups of the engine sizes (e.g., 60-90, 91-190, 191-299, 300+). Visualise both the findings using the histogram. Explain your findings.
3. Do diesel cars have higher average **CityMpg** than gasoline cars? Provide statistical evidence. How does **DriveWheels** affect fuel efficiency (**CityMpg** and **HighwayMpg**)? Elaborate on the findings. Filter out those engines in the dataset that have trouble or are suspected of having trouble; what are the top 5 most common troubles related to the engines? Do the troubles differ between engine types?
4. Write the code to show which error type (**ErrorCodes**) occurs most frequently; Write the code to analyze the factors that might influence the maintenance methods (Urgent care, Adjustment, Replacement) for the trouble vehicles (confirmed or suspected) in the dataset. Any factors in the dataset, such as **BodyStyles**, **FuelTypes**, and **ErrorCodes**, can be considered. Pick 2 of the factors and explain if there is a trend that explains the variation.
5. Add and commit files R Markdown and exported PDF in vUWS submission site, which includes the complete R source code, visualizations, statistical analyses, and relevant data files, to your Git repository with a meaningful commit message. Then push the changes to your remote repository and provide the repository link for assessment. Make sure no private or sensitive data is committed.

The report will be evaluated based on the marking criteria, so ensure each part of your analysis meets all required standards.