MONASH University

Faculty of Information Technology

Semester 1, 2025

FIT5145: Foundations of Data Science

Assignment 2: Description

Due Date: Friday, Week 8 (May 2, 2025), 11:55 PM

The aim of this assignment is to investigate and visualise data using **R**. It will test your ability to:

1. Read data files and extract related data from those files;
2. Clean and process data into the required formats;
3. Perform exploratory data analysis and visualisation;
4. Use basic tools for managing and processing data; and
5. Communicate your findings in your report.

Assessment Details:

- Assessment Type: Individual Assignment
- Total marks: 15%
- Due Date: Friday, Week 8 (May 2, 2025), 11:55 PM. Please notice that we do not accept submissions after May 9, 2025 (i.e., 7 days after the due date).

Submission Details:

You will need to submit two separate files (**PDF report and RMD file**).

1. A **report in PDF** containing your ***(a) code, (b) answer, and (c) explanation*** used to answer each question.

   (a) code: Please convert (knit) the RMD file, including your codes, directly into a PDF file (Note: Please knit RMD to HTML and print the HTML as a PDF).

   (b) answer: Please make sure to include **the code outputs and written answers** for each question.

   (c) explanation: Please explain how you answered each question (i.e.: explaining your codes or summarising your work for each question).

   Marks will be assigned to reports based on their correctness and clarity. For example, higher marks will be awarded to reports that include graphs with appropriately labelled axes and sufficient comments for the code.

2. The **RMarkdown** file: Please submit the RMarkdown file that contains your R codes. Your file should contain all the codes, and proper comments. If your work uses new libraries that have not been introduced in classes, please include instructions on how to get these libraries installed.

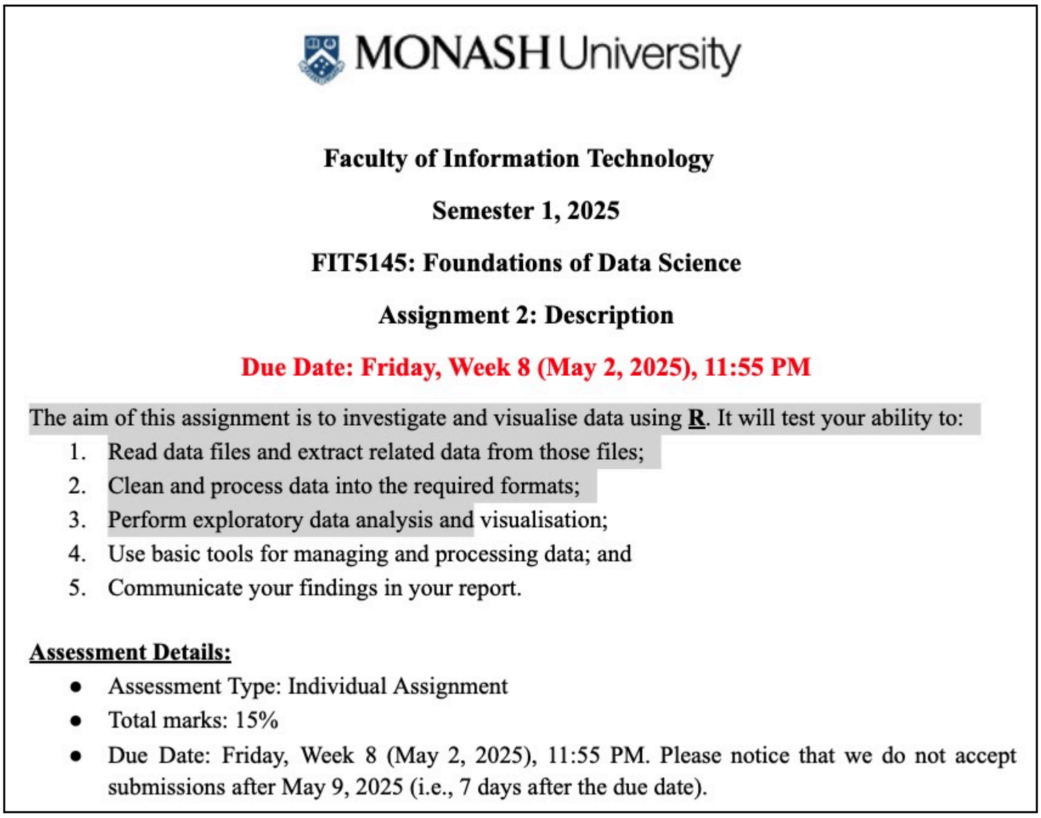
Notes:

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 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).



Faculty of Information Technology

Semester 1, 2025

FIT5145: Foundations of Data Science

Assignment 2: Description

Due Date: Friday, Week 8 (May 2, 2025), 11:55 PM

The aim of this assignment is to investigate and visualise data using **R**. It will test your ability to:

1. Read data files and extract related data from those files;
2. Clean and process data into the required formats;
3. Perform exploratory data analysis and visualisation;
4. Use basic tools for managing and processing data; and
5. Communicate your findings in your report.

Assessment Details:

- Assessment Type: Individual Assignment
- Total marks: 15%
- Due Date: Friday, Week 8 (May 2, 2025), 11:55 PM. Please notice that we do not accept submissions after May 9, 2025 (i.e., 7 days after the due date).

Assignment Task:

The National Waste Report 2022 report dataset is provided by Dept. of Climate Change, Energy, the Environment and Water in Australia. The dataset contains data on Australia's waste generation, recovery and fate for all waste streams and various material categories and has data for the period from 2005 to 2021.

There are two data files, *Wastes.csv* and *Year_State_ID.csv* with their details provided below. A more description of the dataset is available here. **These files have been modified for the task. Please download them from the Assessments page on Moodle.**

- ***Wastes.csv***

Column	**Description**
Case_ID	ID of each waste process case.
Year_State_ID	ID of Year and State.
Category	A broad classification of waste material.
Type	A more detailed classification of waste material. For example, the category 'Metals' may be split into : 'Aluminium', 'Non-ferrous metals (ex. aluminium)', and 'Iron and steel'.
Stream	Describes the source of waste, comprising three options: municipal solid waste (MSW) from households and council operations; commercial and industrial (C&I) waste; and construction and demolition (C&D) waste.
Fate	The ultimate destination of the waste, comprising five options: disposal; recycling; energy recovery; long-term storage; and waste reuse.
Tonnes	The quantity of waste.
Core_Non-core	Core waste: Waste generally managed by the waste and resource recovery sector; comprises solid non-hazardous waste, hazardous waste (including liquids) and biosolids. Non-Core waste: It includes material from mining, minerals processing, agriculture and fishing.
Description	● Describes each waste process case and provides a score and feedback for each case. ● Environmental Impact Score: Measures how much a waste process affects the environment, either positively or negatively, using a 10-point scale. - 10: Very positive impact - 1: Very negative impact

- ***Year_State_ID.csv***

Column	**Description**
ID	ID of Year and State
Year	Financial year. Data is presented for each year between 2006-2007 and 2020-2021
State	State or territory in which the waste was generated.
Economic_Growth	The increase in the production of total goods and services (e.g.: 3.07: 3.07 % increase, -1.31: -1.31% increase)

Let’s take a look at the file ***Wastes.csv***.

1. How many unique *“Category”* values are there in the data file (6 marks)?

2. How many negative feedback comments are in the 'Description' column with an environmental impact score of 2 or 3 (4 marks)?

3. For each *Category* value, write code to calculate the fractions of waste tonnes of different waste sources, and then draw a chart to visualise the fraction numbers specific to the *Category* value (3 marks).

4. Please add the 'Year' and 'State' values from *Year_State_ID.csv* to *Wastes.csv*, compute the total waste tonnes for each year and state, and store the result in a dataset named 'temp'. Then, use a single R function/command to display the statistical information (i.e., Min, Max, and Mean) of the total waste tonnes for each state in 'temp' (Note: You may use multiple functions/commands to prepare the pre-processed data table, but when you compute and display the statistical information, you need to use a single R function/command.) (6 marks).

5. Write code to draw a chart showing a yearly trend of total waste tonnes of food organics for each state. To draw the chart, please convert the Year-Year formats of all *Year* values into Year formats (3 marks).

6. Write code to display the most recycled waste *Type* and the most disposed waste *Type* with the corresponding year. Write code to display the most increased waste *Type* over years (4 marks).

7. Please investigate the factors influencing environmental impact scores. (Note: You may use variables from both *Wastes.csv* and *Year_State_ID.csv* for your analysis.) Marks will be awarded based on the depth of your investigation (including analyses and discussions) and the strength of your reasoning (how insights and conclusions are drawn) (5 marks).

8. Write code to only **keep** data records where the *Category value* is Hazardous wastes, the *Type* value is Tyres (T140) and tonnes value is more than 0, then write code to add a new column named "*Tonnes_range*" and fill it with one of the following values based on the corresponding "Tonnes" value:

   ○ [0, 10000)

   ○ [10000, 20000)

   ○ [20000, 40000)

   ○ [40000, 80000]

   Then, for each state, display a chart to show the number of cases of different *score_range* values. What do you observe? (4 marks)

9. Using both original datasets and external sources, please investigate the factors influencing the yearly trend of total C&D waste tonnes. Marks will be awarded based on the depth of your investigation (including analyses and discussions) and the strength of your reasoning (how insights and conclusions are drawn). (Note: Please provide the source links for any external data used) (7 marks).