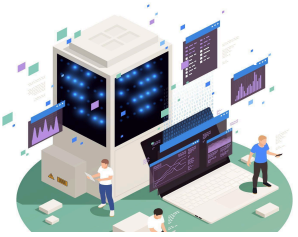
MONASH University
MALAYSIA



Assignment Sheet

Unit Name	Introduction to Data Science
Unit Code	FIT 1043
Unit Teacher Name	Ts. Dr. Sicily Ting
Assignment Name	Assignment 3 (20%)
Aim of this assignment	Exploratory Analysis of Big Data- R and Unix Shell

Learning Outcomes

This assignment assesses the following learning outcomes:

Learning Outcome Number	Learning Outcome Description
2	Demonstrate the size and scope of data storage and data processing, locate suitable resources, and classify the basic technologies in use.

Weighting

This assignment is worth [**20%**] of your overall grade for this unit.

Requirements

This assignment has the following requirements:

Assignment Type	**Individual Task (20%)**
Response Format / Hand-in Requirements	Submit a **PDF file[1]** 1. **PDF file** : a. Answers to the questions. In order to justify your answers to all the questions, make sure to i. Include **screenshots/images of the graphs or outputs** you generate (You will need to use screen-capture functionality to create appropriate images.) ii. Please be informed that you need to explain what each part of the command does for all your answers. For instance, if the code you use is 'unzip tutorial_data.zip', you need to explain that

	the code is used to uncompress the zip file. iii. **Copy and paste of your Unix code from Bash Shell and the R code (<u>Do not screenshots of your code</u>**). iv. Kindly **Do Not** copy the questions, else you might have high Turnitin similarity due to all submissions referring to the same set of questions (5% penalty) [1] You can use Word or other word processing software to format your submission then save it/convert it as a PDF file.
Response Specifications	**1 PDF file.** Zip, rar or any other similar file compression format **is not acceptable** and **will have a penalty of 10%.**
Due Date	**Week 12 : 11.55pm (MYT), Friday 24th May 2024**
Supporting Material	**Dataset needs to be downloaded from Moodle source.**
Notes:	The submission must be done via the Moodle site's submission link.

Objectives

Assignment 1 & 2 walked you through what you have learnt in Lectures 1 to 7 and also the "middle pipeline" or Collection, Wrangling, Analyse and Present of our Standard Value Chain. It provides you an introduction to the Data Science lifecycle. This assignment relates to the latter part of this unit, in the use of the BASH Shell and the R programming language to work on larger datasets. It will test your ability to:

- Read a reasonably large dataset,
- Process the dataset using BASH Shell Scripts,
    - Use online resources or the "man" pages or "-- help" to assist in the commands
- Conduct aggregation of the dataset content,
- Read data from a file in R, and
- Generate appropriate visualisations in R and output to files

Note that unlike the previous Assignments, you will notice that there is less explanation or detailed information pertaining to the data and the process. In other words, there will be less guidance and you are expected to be able to understand the requirements and provide suitable answers to the tasks.

Data

The data provided is a pre-processed data that is derived from corona_tweets_58.zip from https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset . The data is a twitter dataset for 15th May 2020, filtered by keywords that are related to the current COVID-19 situation. Instead of providing the whole twitter data for that day alone, which amounts to more than 6GB of JSON formatted data (***V**olume*), the dataset provided to you for this assignment has been processed and contains only a portion of the data. There are more than 2 million tweets and it will take more than 24 hours to download all the tweets from twitter (using a standard developer account). Note that to download 1-day data (that is only a small portion of twitter data for that day), it takes more than 1 day to do so (***V**elocity*). The original data comprises more than 6GB of JSON data which contains a variety of information such as data, text, identities, numbers and so on.

The data has been pre-processed into CSV (it's separated using tabs in order to minimize issues with commas (',') as the twitter text may contain commas as well as other fields). It has been converted from JSON and only contains a subset of the data. The file corona_tweets.csv.gz can be downloaded from Moodle and this assignment will be based on this file.

MONASH University
MALAYSIA

Assignment Tasks:

This assignment is worth **20% of this Unit's assessment.**

This assignment is to illustrate working with large data sets (in this case, just more than a million lines of data, not really large but enough for learning) and to also experience the use of shell scripts to process and aggregate data. In the whole exercise, you **must NOT uncompress** the data and store it. Once the data is aggregated and nicely formatted, you are then to read the data in R where you are to conduct further analysis. In this assignment, you only need to read the data in R and provide some visualisations.

Note, for this assignment you are required to write **shell commands** to answer all questions unless the instructions specify using **R code**.

A1: Inspecting the data

1) Copy the downloaded file to your UNIX (Linux) terminal. State the size (in Bytes or MegaBytes) of the **corona_tweets.csv.gz** file and provide the shell command that you used to determine the size.

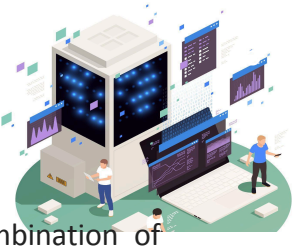
2) The first line of the CSV file contains headers that are "tab" separated. What are the header names and provide the command you used to obtain it.
Note that the command provided has to be in one line.

3) How many lines are there in the dataset? Again, provide the single line code on how you obtained it.

A2: Investigating the information from Data

1) How many unique twitter users are there in the dataset? Provide the single line code that uses the "awk" and "uniq" command. You are also required to read the "man" pages / "help" of the "uniq" command to figure out if it is sufficient to answer the question.
Provide an explanation on the code you provided.

2) For each of the sub-question below, provide the single line code (one each) and briefly explain your code.



a) How many tweets mentioned the word “vaccine” in any combination of uppercase or lowercase letters?

b) How many of those are not spelt exactly “vaccine”, or“Vaccine” but in other combination of uppercase and lowercase (e.g. VaCcine, vacciNE), and

c) Output the lines of **A2.2 (b)** into a file called Result.txt (not the number of lines but the specific lines).

A3: Data aggregation

For this part, let’s assume that you would like to know how many followers each of these twitter users have.

*Do note that twitter users may tweet more than once a day and hence they can appear more than once in this dataset.

1) Let’s group the twitter user (ID) by the number of followers that they have into the following ranges. Provide the code for them. One line of code for each of them below (10 lines).
   a) Less than or equal to 1500
   b) 1501 to 2500
   c) 2501 to 3500
   d) 3501 to 4500
   e) 4501 to 5500
   f) 5501 to 6500
   g) 6501 to 7500
   h) 7501 to 8500
   i) 8501 to 9500
   j) More than 9500

2) Create a CSV file manually with the output from **Part A3.1** . The CSV file should contain two columns, the first column is the range (e.g. “1501 to 2500” or “1.5-2.5k” or other meaningful names) and the second column is the number of twitter users.

3) **[R Code]** Use the output of the above **(A3.2)** and read it using R.

4) **[R code]** Plot the suitable visualisation (Histogram/Bar Chart/ any visualisation that you think is suitable) using the data and output it into a PNG file (.png). For submission, you just show the code, and paste the PNG image in your PDF report.

(Not part of the assignment but worth mentioning. For those who are eagle eyed, you will

find that the total number you obtain from Part A3 (1) (a-k) is different from the unique twitter users obtained from Part A2 (1), can you think of a reason? Do you trust your data? (***V**eracity*))

A4: Small Challenge

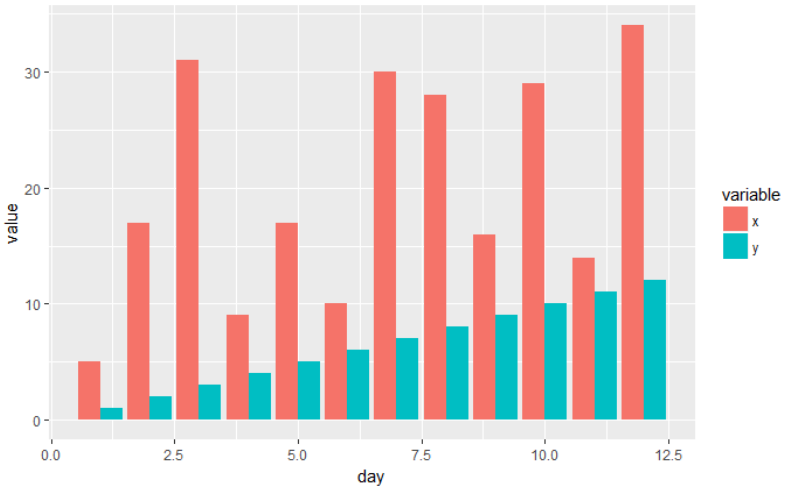
Let's assume that we want to compare against tweets that are not retweets. We can assume that this is indicated at the beginning of the text by the "`RT @`" (Note that if we are using the raw JSON data, there is a field that indicates it).

1) Provide a single line code that filters out the tweets that contain "RT @" and output the results into a compressed gz file (Hint: you need to use the opposite of gunzip).

2) Do the same process as **Part A3.1 to A3.2** with this new file.

3) **[R code]** Copy the output of the above **(A4.2)** and read it using R.

4) **[R code]** Plot a side by side bar chart to visualise the data created from **A3.2** and from **A4.2** .
(An example of a side by side grayscale bar chart is shown below - the example below is not the answer to this question). The bars should be coloured (any colour/pattern that you like).



5) Explain your findings based on the plotted chart in **(A4.4)**

Clarifications

Do use the Ed Forum (https://edstem.org/au/courses/15857/discussion/) so that other students can participate and contribute. For postings on the forum, do use it as though you are asking others (instead of your lecturer or tutors only) for their opinions or interpretation. Just note that you are not to post answers/solutions directly.

MONASH University
MALAYSIA

Congratulations!

You have completed all FIT1043 assignments and you will have only the final exams left. I do hope that you have been well introduced to the world of Data Science, which still requires significant effort and there is lots more to learn. Hopefully those skills will contribute to your lifelong learning!