FIT1043 Assignment 3 Specifications (20 %)

Version 1.0

Release on: Monday 9th May 2022 – 11.55 pm

Due date: Wednesday 25th May 2022 – 11.55 pm

Table of Content:

1. Objectives and Learning outcomes
2. Data
3. Tasks
    3.1. Format
    3.2. Tasks to do
    3.3. Sanity checks
4. Submission
5. Marking Rubric

1. Objectives and learning outcomes

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.

This assignment will test your ability to:

- Navigate the BASH Shell
- Process large file using BASH Shell
    - Use online resources or the "man" pages or the "--help" to assist in the commands Output a processed file to CSV format using BASH Shell
- Read a processed file in R
    - Conduct visualisation using R

Contribute to the following learning outcomes:

LO 4. Classify participants in a data science project: such as statistician, archivist, analyst, and systems architect;

LO 5. Classify the kinds of data analysis and statistical methods available for a data science project;

LO 6. Locate suitable resources, software and tools for a data science project;

2. Data

Format: compressed file (FB_dataset.gz) provided via Moodle site

Description: It is a compressed file that contains Facebook posts from 15 of the top mainstream media sources (e.g., BBC, CNN, Fox News, etc.) from 2012 to 2016.

Note: You will need to use either a Windows Subsystem for Linux (WSL) in Windows OS, Linux machine, a Mac terminal or Cygwin on a Windows machine for this purpose.

3. Tasks

3.1. Format

This assignment is worth 60 marks, which makes up for **20% of this Unit's assessment**. There are two parts of tasks that you need to complete for this assignment. Students that complete only Tasks A1 – A9 and B1 can only get a maximum of D (Distinction). Students that attempt tasks A10 and B2 (Challenge Questions) will be showing critical analysis skills and a deeper understanding of the task at hand and can achieve the highest grade (HD). You need to use the BASH shell and R to complete the tasks.

3.2. Tasks

Part A: Investigating Facebook Data using shell commands

Section A (45 marks)	Questions
A1 (2 mark)	What is the original file size? Decompress the file. How big is it?
A2 (2 marks)	What delimiter is used to separate the columns in the file and how many rows are there?
A3 (3 marks)	Print out the names of the columns, and how many columns are there?
A4 (4 marks)	How many unique pages are there? (You will need to understand how to interpret "unique pages" on your own)
A5 (3 marks)	What is the date range for the Facebook posts in this file? (Assume that the data is ordered by date in chronological order)
A6 (6 marks)	When was the first mention (give the date) in the file regarding "Malaysia Airlines", what was the message, and which media source first mentioned it?
A7 (4 marks)	How many times is "Cat" mentioned in the message column of the file? How did you find this? (Do not ignore the case, i.e., lower/upper case)
A8 (6 marks)	What about "Dog" (mentioned in the message column of the file)? Which animal is more popular on Facebook, Cat or Dog? (Do not ignore the case, and you are to define what is meant by popular)
A9 (8 marks)	Select the posts where "Cat" (Ignore the case) is mentioned in the post content (message field) and number of likes for those posts are greater or equal than 1,000. And generate a new file with post_id and sorted like_count and name it "cat.txt". (You need to add a screenshot of the column header, with first 5 rows and the last 5 rows, in your report).

[Challenge] A10 (7 marks)	**Challenge:** Find the total number of love_count and angry_count for "Cat" and "Dog" separately. Which animal has more positive feeling among people?

Note: Justify your answer. (Do not need to ask for clarification for this, you are to justify your interpretation of the question and your approach).

(You may need to search online to find how to sum a column of numbers using awk, and you may need to consider both love and angry count when justifying your answer)

Part B: Investigating Facebook Data using shell commands

Section B (15 marks)	**Questions**
B1 (10 marks)	We want to consider how the amount of discussion regarding Dog varies over the time period covered by the data file. i.) To answer this question, you will need to extract the timestamps for all posts referring to "Dog" (ignore case) using the BASH Shell. You will then need to read them into R and generate a histogram. [Hint: To read the data into R, first generate a file containing only the timestamp column as text. Then read the file into R as a CSV file.] R will not recognise the strings as timestamps automatically, so you'll need to convert them from text values using the strptime() function. Instructions on how to use the function is available here. You will need to write a format string, starting with "%a %b" to tell the function how to parse the particular date/time format in your file. What format string do you need to use? ii.) Once you have converted the timestamps, use the hist() function to plot the data in R.
[Challenge] *B2 (5 marks)*	**Challenge:** In this question, we want to look at a specific content type that influences engagement on Facebook. To make this task easier, we will specifically look at the number of comments posted against each of the post type (event, link, photo, status and video) for "abc-news". i.) Draw a boxplot to show the distribution of comments made against each type of post (event, link, photo, status and video) created by "abc-news". What can you infer from this plot? Which is the most engaging post type?

ii.)	You may have noticed that the presence of outliers affects the readability and interpretation of the data in the box plot. Redraw the boxplot by filtering out values (comments_count) greater than 1,000. *[Hint: You may need the library mentioned in Week 8's lab]*
iii.)	iii. Which type of post (event, link, photo, status or video) has on average been most effective for "abc-news". In other words, which post_type has the highest median comment_count.

3.3 Sanity checks

- After you are done with the tasks, do sanity checks.
    - Even though you don't need to submit the code script, you will need to double check the your file consisting of the correct code (copied from your Bash Shell/R), answer for the questions that you had attempted and images of your outputs.
- Make sure that your submission contains everything we've asked for.

4. Submission

For this assignment, you are to hand-in your work via Moodle, **only 1 well formatted PDF (**generated from your word processor, e.g. from Microsoft Word) **file is needed**.

Details for the submission:

1) Hand in a PDF file containing your answers to all the questions and numbered correspondingly.

2) Your report should include the following cases:

a) The **screenshots/images** of the **outputs/graphs** you generate in order to justify your answers to all the questions. Ensure that they are legible, such as making sure that the image resolution is sufficient.

b) Copies of all the bash command lines and R scripts you use. If your answer is wrong, you may still get half marks if your command line or script is close to correct. **Ensure that the grader can copy and paste your code from the PDF document.**

Note: Do not screenshot the code, copy and paste your code to your file.

3) Please be informed that you need to explain what each part of 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.

4) Please don't include the questions into the assignment, this will cause a high Turnitin percentage, and we will impose a **10% penalty if you include the questions** in the PDF. Note that, unlike the previous assignments, where the Turnitin of 50% is acceptable due to the display of the data and also the limited commands; for this assignment, the expectation will be much lower.

5. Marking Rubric

This assignment is **worth 60 marks**, which makes up for **20% of this Unit's assessment**. There are two parts of tasks that you need to complete for this assignment. Students that complete only Tasks A1 – A9 and B1 can only get a maximum of D (Distinction). Students that attempt tasks A10 and B2 (Challenge Questions) will be showing critical analysis skills and a deeper understanding of the task at hand and can achieve the highest grade (HD). You need to use the BASH shell and R to complete the tasks.

General points:

- **Late submission:** Late submissions will have a penalty of 10% per day, including weekends and public holidays for up to 7 days. Assessment items handed in after 7 days will not be considered.
- **Zip (compressed) file submission will be penalised** : Zip file (or any compressed file) submission will have a penalty of 10%.
- **Drafts (not submitted):** There have been many of you who left your submission in Draft mode. Please make sure to submit your assignments that are in draft mode. *Note: For this assignment, we reserve the right not to accept the assignments that are not yet submitted.*
- **Screen shots of codes are NOT acceptable** (the grader will need to be able to copy and paste your code) and there will be no marks awarded for the code portion.
- **Acknowledgement of sources:** Plagiarism or unauthorised collaboration will result in an automatic fail grade
- **Extension (Special Consideration):** If you can't complete an assessment (due to exceptional circumstances beyond your control), you may be eligible for special consideration.

- Marks: Refer to the marks beside every task. And the marking range as below:

Part A (45 marks)	
Coding Part (60 % of Part A)	Low marks: with errors High marks: Error free
Answers or justify answers (if applicable) 40% of Part A	Low marks: Poor/no justification (if applicable) High marks: Correct answer, strong justification (if applicable)

Part B (15 marks)	
Coding and Visualisation (80% of Part B)	Low marks: Some errors or misleading visuals High marks: Error free or clear visuals
Answers or justify answers (if applicable) 20% of Part B	Low marks: Poor/no justification (if applicable) High marks: Correct answer, strong justification (if applicable)

Have Fun!

Clarifications

Do use the Moodle Ed Forum so that other students can participate and contribute. For postings on the forum, do post 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 directly.

Congratulations!

You have completed your in-semester assessments for FIT1043. I hope that you have enjoyed the course assignments, starting from a very guided assignment 1, to something with a little bit of flexibility for you to try out new stuff and compete in assignment 2, and finally assignment 3.