# FIT1043 Assignment 3
# Semester 2, 2020

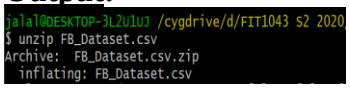
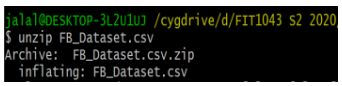
Due: 6th November 2020, 11:55pm

## Hand in Requirements:

1) Please hand in a **PDF** file containing your answers to all the questions and numbered correspondingly.
2) Your report should include the following cases:
   - The screenshots/images of the outputs/graphs you generate to justify your answers to all the questions.
   - Copies of all the bash command lines and Python/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. You must **copy-paste** your codes into your report and screenshots of the code will not be accepted and will result in **auto failure***.
3) Please be informed that you need to explain what each part of command does for all your answers in Task A. 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. For task B, you need to provide a more general explanation about how your codes work.
4) Please do not include the questions into the assignment and just include the question numbers such as 1, 2, etc. (It has a 5% penalty) to generate a more reliable Turnitin score.
5) You will be penalized by 5% of the assignment mark (5% out of 20 marks) if you submit after the due date for every day that you are late. If you could not submit your assignment before the due date, please make sure to submit your files at most 7 days after the assignment due date, we do not mark assignments which will be submitted after 13th of November11:55 pm.

***Suppose the question asked for uncompressing the FB_Dataset.csv. Example of marking is provided in Table 1.

**Table 1: Marking example**

| **Example** | **Code:** unzip FB_Dataset.csv<br><br>**Output:**<br>  <br><br><br>**Explanation:** the code is used to uncompress the zip file | **Code:**<br>  <br><br><br>**Output:**<br>  <br><br><br>**Explanation:** the code is used to uncompress the zip file | **Code:** unzip FB_Dataset.csv<br><br>**Output**:<br>  <br> |
|---|---|---|---|
| **Marks** | Full mark. | Will be rejected as the code is a screenshot. | 30% of the mark will be deducted as there is no explanation. |

## Data:

The dataset contains Facebook posts from 15 of the top mainstream media sources (e.g., ABC, BBC, etc.) from 2012 to 2016. The dataset for this assignment is shared in the Google drive and can be downloaded using the following link.

https://drive.google.com/file/d/1T2LMLTjhxGlr9V32B2FaNmtokJPRXJxY/view?usp=sharing

## Assignment Tasks:

There are two tasks that you need to complete for this assignment. Students that complete **only** Tasks **A** and **B1** can only get a **maximum of Distinction**. Students that attempt task **B2** will be showing critical analysis skills and a deeper understanding of the task at hand and can achieve the **highest grade**.

## Task A: Investigating Facebook Data using shell commands

Download the file FB_Dataset.csv.zip from the link above. Use a Unix shell to manipulate the file and answer the following questions.

1. Decompress the file. How big is it?

2. What delimiter is used to separate the columns in the file? The 2nd column is the unique identifier for a Facebook post. Print out the names of other columns in the output?

3. How many unique pages are there?

4. What is the date range for Facebook posts in this file? (Assume that the data is in order)

5. How many times has the term “Donald Trump” (ignore the case) appeared in the content of post names? When was the first mention of “Donald Trump” (ignore the case) in the post names and what was the post name? Considering different columns which you have for this post, excluding the message and post name, can you say whether people’s reactions were positive/ negative to this post? How many reactions can you see against this post?

6. Select the post id and number of likes of posts in which the term “Trump” (Ignore the case) is mentioned in the post content and the number of likes is less than 100. Then, sort the data based on the like_count (descending sort) and save it in a file named as “trump.txt”. (You need to add a screenshot of the output, including the first 5 rows and the column headers in your report).

# Task B

## Task B1: Analysis of "the-wall-street-journal" posts using shell commands and R

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 types (event, link, photo, status, and video) for "the-wall-street-journal". Extract the required information which is posted by "the-wall-street-journal" using shell commands and save the result as a CSV file named as "the-wall-street-journal.csv".

7. "the-wall-street-journal" has asked you to focus on the analysis of the post_types for which the number of comments is less than 4000. You need to read the "the-wall-street-journal.csv" file generated as mentioned above into R, filter the data based on the requirements of "the-wall-street-journal", and draw boxplots to show the distribution of comments made against each type of post (event, link, photo, status and video). You need to present one plot which contains different boxplots for different post types. What can you infer from this plot? Can you detect which one is the most engaging post type? Make sure that your plot has proper labels and a title.

8. You may have noticed that the presence of outliers affects the readability and interpretation of the data in the box plots. Redraw the boxplot by filtering out values (comments_count) less than 1000.

9. Which type of post (event, link, photo, status, or video) has on average been most effective for "the-wall-street-journal.csv"? In other words, which post_type has the highest median comment count?

## Task B2: Analysis of "abc-news" posts based on your preferences

In this task, you can use R, Python or shell commands or any combination of Python, R and Shell to answer the questions.

The 'abc-news' asked you to help them to analyse the reactions of Facebook users to the posts which they published about "Donald Trump" (ignore the case) by doing the following tasks.

10. Create a bar chart which shows the total number of reactions to the posts published by "abc-news", in which the posted message contains the term "Donald Trump"(ignore the case) for each day of the week ('Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday'). Make sure the bar chart is sorted based on the weekdays as shown in the screenshot below. Understanding what should be considered as a reaction is a part of the answer to this question which you can figure out by checking different columns of your dataset. You need to mention and justify the criterion which you choose to define a reaction to a post. (Please pay attention that the plot does not show the real values and it is created with fake data just to show you how the output should be).

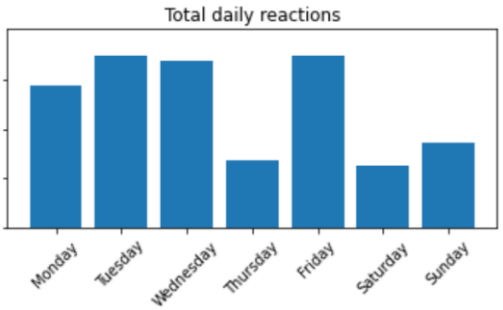



**Figure 1:Sample output for question 10**

11. Considering the created bar chart of question 10, name two days in which users have shown the most reactions to the posts. Is there any difference between the number of reactions during the weekdays and at the weekends?

12. We need to take a closer look at the total reactions in the two days which users have shown the most reactions. Create two bar charts to show the hourly total reactions for each of two days. What time did the most reactions happen on each day? Is there any similarity between the number of hourly reactions in these two days? (Please pay attention that the sample plot given below does not show the real values and it is created with fake data just to show you how the output should be presented for this question.)

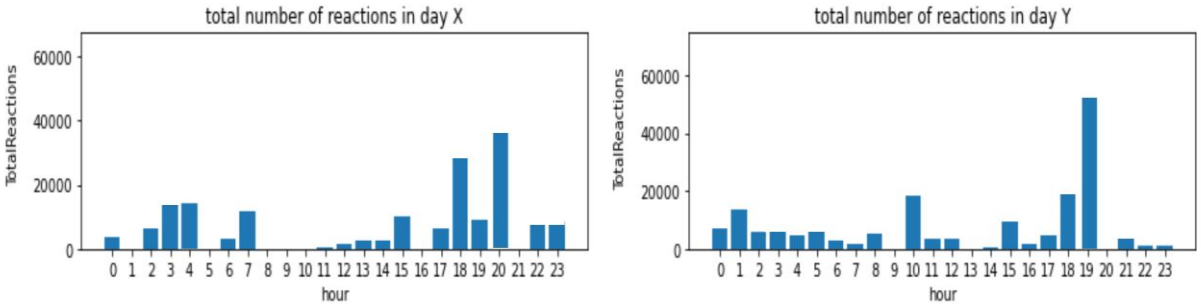



**Figure 2: Sample output for question 12**

13. Considering your exploration about the reactions in different days/times for the term "Donald Trump" in posts by "abc-news", answer the following questions.
    a) What was the day and time which had the maximum number of reactions?
    b) Do you think it is a good idea to recommend publishing a general post about Trump in the days which you found in question 11 and the peak hours which you found in question 12? What is your suggestion based on the analysis which you did in this task? Justify your answer.

**Good Luck!**