Assignment Sheet

Unit Name	Introduction to Data Science
Unit Code	FIT 1043
Unit Teacher Name	Ts. Dr. Sicily Ting
Assignment Name	Assignment 3 (15%)
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 [**15%**] of your overall grade for this unit.

Requirements

This assignment has the following requirements:

Assignment Type	**Individual Task (15%)**
Response Format / Hand-in Requirements	Submit a single **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/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	**A single PDF file**. Zip, rar or any other similar file compression format **is not acceptable** and **will have a penalty of 10%.**
Due Date	**11.55pm (MYT), Friday 21st October 2022**
Supporting Material	**Two data sets** for this assignment are in the Google shared drive: https://drive.google.com/drive/folders/1Ala0KnoHlgeXFpxVwr7OJaOzSR7xxHSN?usp=sharing Both are large, so your best bet is to download them while in the lab/studio and do the assignment there. You will need to use either a Linux machine for this or a Mac terminal or Cygwin on a Windows machine.
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:

- Navigate the BASH Shell
- Process large file using BASH Shell
    - Use online resources or the "man" pages 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

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

We will explore two datasets in this assignment :

1. **dataset_TIST2015 dataset:** There's a readme file which consists of basic info of this dataset, you will work on this dataset in Task A.

2. **Twitter_Data_1 dataset:** you will work on this dataset in Task B.

Assignment Tasks:

This assignment is worth 50 marks, which makes up for **15% of this Unit's assessment.** There are two tasks that you need to complete for this assignment. Students that complete **only** Tasks **A1-A7, B1-B4** can only get a **maximum of Distinction**. Students that attempt tasks **A8 and B5** will be showing critical analysis skills and a deeper understanding of the task at hand and can achieve the **highest grade**. You need to use the Unix shell and R to complete the tasks.

[30 marks] Task A: Investigating User global-scale check-in data collected from Foursquare Data in the Shell

Download the file dataset_TIST2015.tar, which contains user check-in data from Foursquare (https://foursquare.com/).

1) **[4 marks] Decompress the tar file and have a look at the files it contains. How many files are there in the tar file? How big is each file?**

2) **[2 marks] What delimiter is used to separate the columns in the file (dataset_TIST2015_Checkins_v2.txt) and how many columns are there?**

3) **[4 marks] The first column in the dataset_TIST2015_Checkins_v2.txt file is user_id. What are the other columns? Print out the names of the columns.**

4) **[4 marks] How many Checkins are there in the file? and how many users are there in the file?**

5) **[2 marks] What is the first and last dates in dataset_TIST2015_Checkins_v2.txt file** (Assume that the data is ordered by date in chronological order)

6) **[4 marks] How many unique venue IDs are there in the dataset_TIST2015_POIs.txt file?**

7) **[4 marks] How many unique Venue categories are included in dataset_TIST2015_POIs.txt file for France** *(Hint: FR is country code for France)*

8) **Background:** How would you select venues from Europe? Consider the structure of the data presented in the readme file. Check-ins are indexed by a Venue ID, and these are described separately in a separate file, the POI file. You can select European venues from the POI file in (at least) two ways: select items in a latitude longitude bounding box, or select items by country code. Don't be too fussed by the exact locations (include or exclude Turkey, Ukraine, etc., that is OK either way).

   [6 marks] Create an awk script to create a European subset of the POI file, and name the subset file "POIeu.txt". Investigate your European subset.

   A. Submit the created POIeu.txt along with your PDF file.
   B. What country has the most venues and what the least, with how many?
   C. Which country has the most Seafood restaurants?
   D. What is the most common (as in, how many venues) class of restaurant in Europe?

[20 marks] Task B: Investigating the Twitter Data in the Shell and Graphing in R

In this task you are working with **Twitter_Data_1.gz data file**. Please decompress the file and answer the following questions.

1) **[2 marks] How many times does the term 'Donald Trump' appear in tweets?** (Note: If the term appears two times in a tweet, we count as two)

2) **[5 marks]** ***Background:*** We want to consider how the amount of discussion regarding Donald Trump varies over the time period covered by the data file.

   To answer this question,

   - You will need to **extract the timestamps** for all tweets referring to Donald Trump.
   - 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.]* 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 are available here: (https://stat.ethz.ch/R-manual/R-devel/library/base/html/strptime.html). [Note: the histogram should be plotted in the next question (Q3).]

   - ***Question:*** 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?

3) **[6 marks] (R code) Once you've converted the timestamps, use the hist() function to plot the data.** *[Hint: you may need to set the number of bins sufficiently high to see the variation over time well.]*

4) **[3 marks] [R code] The plot has a bit of an unusual shape. Can you see a pattern before Feb 15 and what happens after that?**

5) **[4 marks] Plot a second histogram, but this time showing the distribution over number of tweets per author in the file.**
   [Hint: You'll need to count up the number of Tweets by each unique author in the Twitter file giving a file with two columns "user" and "twitter count" in the ***bash Shell*** *. Then load them into* ***R****. This is a large file so you can also just isolate the counts, sort and count them to get a summary statistics file with columns "twitter count" and "number of users".]*

Good Luck!