Name: Lim Yu Jin
Student ID: 32637888

# Data Science Assignment 3

## Part A:

1.

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ gunzip FIT1043_Dataset.gz
```

Code = Gunzip FIT1043_Dataset.gz

Result = Unzip the file (FIT1043_Dataset)

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ ls -lh FIT1043_Dataset
-rwx------+ 1 yujin yujin 193M Oct  4 16:42 FIT1043_Dataset
```

Code = ls -lh FIT1043_Dataset

Result = Show the file (FIT1043_Dataset) is in 193MB

2.

```
0,1467810672,Mon Apr 06 22:19:49 PDT 2009,NO_QUERY,scotthamilton,is upset that h
e can't update his Facebook by texting it... and might cry as a result  School t
oday also. Blah!
~
```

Code = head -1 FIT1043_Dataset | less

= /,

Result = delimeter is comma (,)

3.

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/Monash/FIT 1043/Assignme
nt 3
$ wc -l FIT1043_Dataset
1471793 FIT1043_Dataset
```

Code = wc -l FIT1043_Dataset

Result = show number of rows (1471793 rows)

Name: Lim Yu Jin
Student ID: 32637888

4.

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ head -n1 FIT1043_Dataset | grep -o "," | wc -l
5
```

Code = head -n1 FIT1043_Dataset | grep -o “,” | wc -l

Result = Show the number of delimeters in a row , which the result shown is 5. So, there are 6 columns

5.

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ awk -F ',' '{print $5}' FIT1043_Dataset | sort | uniq | wc -l
626684
```

Code = awk -F ‘,’ ‘{print $5}’ FIT1043_Dataset | sort | uniq | wc -l

Result =Sort the unique users and count the numbers of them

There are 626684 unique users

6.

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
$ head -1 FIT1043_Dataset
0,1467810672,Mon Apr 06 22:19:49 PDT 2009,NO_QUERY,scotthamilton,is upset that h

yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
$ tail -1 FIT1043_Dataset
4,2193602129,Tue Jun 16 08:40:50 PDT 2009,NO_QUERY,RyanTrevMorris,happy #charity
```

Code = head -1 FIT1043_Dataset
= tail -1 FIT1043_Dataset
Result = list out the first and last line of the data
= Date range: 6 April 2009 ~ 16 June 2009

Name: Lim Yu Jin
Student ID: 32637888

7.

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ grep Ian FIT1043_Dataset | head -1
0,1468287671,Tue Apr 07 00:45:44 PDT 2009,NO_QUERY,IanB022,Started getting mails
hots aimed at pensioners - it's all downhill now
```

Code = grep Ian FIT1043_Dataset | head -1

Result = get user Ian information which is the first appear in the dataset

User: IanB022

Date & Time: 07/04/2009, 00:45:44

Message: Started getting mailshots aimed at pensioners – it's all at downhill now

8.

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ cut -f 6 FIT1043_Dataset | grep -i "Australia" | wc -l
grep: (standard input): binary file matches
1758
```

Code = cut -f 6 FIT1043_Dataset | grep -I "Australia" | wc -l

Result = Get the number of the word "Australia"(ignoring case) in the 6$^{th}$ column by counting the number of rows

= Which is 1758 tweets

9.

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ cut -f 6 FIT1043_Dataset | grep -c -i "Australia"
1764
```

Code = cut -f 6 FIT043_Dataset | grep -c -i "Australia"

Result = Get the number of the word "Australia"(ignoring case) in the 6$^{th}$ column

= 1764

Name: Lim Yu Jin
Student ID: 32637888

10.

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ cut -f 6 FIT1043_Dataset | egrep -i -w -c "Australia"
1289
```

Code = cut -f 6 FIT1043_Dataset | grep -i -w -c “Australia”

Result = get the exact word count of “Australia” (ignoring case) in the 6th column

= 1289

11.

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ cut -f 6 FIT1043_Dataset | egrep -w -c "India"
383

yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ cut -f 6 FIT1043_Dataset | egrep -w -c "Australia"
876
```

Code = cut -f 6 FIT1043_Dataset | egrep -w -c “India”

= cut -f 6 FIT1043_Dataset | egrep -w -c “Australia”

Result = get the exact word count of “India” in the 6th column (383)

= get the exact word count of “Australia” in the 6th column (876)

Conclusion = Popular is the times the current word appears. The word ‘Australia’ is more popular than the word ‘India’ as ‘India’ appears 383 times but ‘Australia’ appears 876 times. So, Australia is more popular.

12.

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ cut -f 6 FIT1043_Dataset | egrep -w -i "India" |cut -f 5 | sort | uniq | wc -l
grep: (standard input): binary file matches
499
```

Code = cut -f 6 FIT1043_Dataset | egrep -w -i “India” | cut -f 5 | sort | uniq | wc -l

Result = count the number of unique users which use the exact word “India”

= 499 unique users

Name: Lim Yu Jin
Student ID: 32637888

13.

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ cut -f 6 FIT1043_Dataset | grep -i "India" | cut -f 1 | grep -c "4"
grep: (standard input): binary file matches
1499

yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ cut -f 6 FIT1043_Dataset | grep -i "India" | cut -f 1 | grep -c "0"
grep: (standard input): binary file matches
1669

yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ cut -f 6 FIT1043_Dataset | grep -i "India" | cut -f 1 | grep -c "2"
grep: (standard input): binary file matches
1669
```

Code = cut -f 6 FIT1043_Dataset | grep -i “India” | cut -f 1 | grep -c “4”

= cut -f 6 FIT1043_Dataset | grep -i “India” | cut -f 1 | grep -c “0”

= cut -f 6 FIT1043_Dataset | grep -i “India” | cut -f 1 | grep -c “2”

Result = count the number of different polar (0,2,4) which represents negative, neutral, positive which have the word “India” in their tweets

India：

Negative： 1669

Neutral： 1669

Positive： 1499

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ cut -f 6 FIT1043_Dataset | grep -i "Australia" | cut -f 1 | grep -c "2"
grep: (standard input): binary file matches
1758

yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ cut -f 6 FIT1043_Dataset | grep -i "Australia" | cut -f 1 | grep -c "0"
grep: (standard input): binary file matches
1758

yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ cut -f 6 FIT1043_Dataset | grep -i "Australia" | cut -f 1 | grep -c "4"
grep: (standard input): binary file matches
1589
```

Code = cut -f 6 FIT1043_Dataset | grep -i “Australia” | cut -f 1 | grep -c “4”

= cut -f 6 FIT1043_Dataset | grep -i “Australia” | cut -f 1 | grep -c “0”

Name: Lim Yu Jin
Student ID: 32637888

= cut -f 6 FIT1043_Dataset | grep -i “Australia” | cut -f 1 | grep -c “2”

Result = count the number of different polar (0,2,4) which represents negative, neutral, positive which have the word “Australia” in their tweets

Australia：

Negative： 1758

Neutral： 1758

Positive： 1589

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ printf '%s\n' Negative,1758 Neutral,1758 Positive,1589 | paste -sd ',' >> sent
iment-australia.csv

yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ printf '%s\n' Negative,1669 Neutral,1669 Positive,1499 | paste -sd ',' >> sent
iment-india.csv
```

Code = printf ’%\n’ Negative,1758 Neutral,1758 Positive,1589 | paste -sd ‘,’ >> sentiment-australia.csv

= printf ’%\n’ Negative,1669 Neutral,1669 Positive,1499 | paste -sd ‘,’ >> sentiment-india.csv

Result = save the data of Australia: Negative,1758 Neutral,1758 Positive,1589 into sentiment-australia.csv file

save the data of India: Negative,1669 Neutral,1669 Positive,1499 into sentiment-india.csv file

Name: Lim Yu Jin
Student ID: 32637888

# Part B:

1.

(i)

```
yujin@LAPTOP-B27UBBGC /cygdrive/c/users/yujin/downloads/monash/FIT 1043/Assignme
nt 3
$ awk -F ',' '$6 ~ "Australia" || $6 ~"australia" {print $3}' FIT1043_Dataset >>
timestamps.csv
```

Code = awk -F ‘,’ ‘$6 ~ “Australia” || $6 ~“australia” {print $3}’ FIT1043_Dataset >> timestamps.csv

Result = get the time column which is column index 3 which its column 6 have the word “Australia”(ignoring index) and save into a csv file name timestamps.csv

(ii)

```
> DF = read.csv("timestamps.csv")
```

Code = DF = read.csv(“timestamps.csv”)

Result = read the file timestamps.csv which I created in Part B (A)I and named it as DF.

```
> DF1 = strptime(DF[,1],format = "%a %b %e %H:%M:%S PDT %Y")
> DF1
   [1] "2009-04-06 23:55:14 +08" "2009-04-07 01:16:43 +08"
   [3] "2009-04-07 02:35:16 +08" "2009-04-07 03:06:17 +08"
   [5] "2009-04-07 03:21:47 +08" "2009-04-07 03:58:25 +08"
   [7] "2009-04-07 05:40:06 +08" "2009-04-07 06:04:24 +08"
   [9] "2009-04-07 06:50:24 +08" "2009-04-07 07:14:51 +08"
  [11] "2009-04-07 07:19:08 +08" "2009-04-07 07:53:09 +08"
  [13] "2009-04-18 07:20:37 +08" "2009-04-18 07:54:42 +08"
  [15] "2009-04-18 08:51:32 +08" "2009-04-18 21:40:46 +08"
  [17] "2009-04-18 22:55:29 +08" "2009-04-18 23:00:47 +08"
  [19] "2009-04-18 23:35:13 +08" "2009-04-19 00:12:39 +08"
  [21] "2009-04-19 01:12:27 +08" "2009-04-19 01:51:25 +08"
  [23] "2009-04-19 06:02:23 +08" "2009-04-19 06:28:58 +08"
  [25] "2009-04-19 22:56:38 +08" "2009-04-19 23:17:04 +08"
  [27] "2009-04-20 01:13:03 +08" "2009-04-20 02:04:42 +08"
  [29] "2009-04-20 03:52:22 +08" "2009-04-20 04:23:42 +08"
  [31] "2009-04-20 04:28:44 +08" "2009-04-20 04:37:56 +08"
  [33] "2009-04-20 23:25:19 +08" "2009-04-20 23:29:30 +08"
```

Code = DF1 = strptime(DF[,1],format = “%a %b %e %H:%M:%S PDT %Y”)

Result = Format the data frame of time zone

Name: Lim Yu Jin
Student ID: 32637888

(iii)

```
getwd()
setwd("C:\\Users\\yujin\\Downloads\\MONASH\\FIT 1043\\Assignment 3")
DF = read.csv("timestamps.csv")
install.packages("tidyr")
library("tidyr")
DF1 ← separate(data = DF, col = 2, into = c("Date", "Time"), sep = " ")
install.packages("data.table")
library(data.table)
DF2 ← setDT(DF1[c(2)])[,list(Count=.N),names(DF1[c(2)])]
DF2$Date
DF3←DF2[order(as.Date(DF2$Date, format="%Y-%m-%d")),]
head(DF3)
ggplot(data = DF3, aes(x = Date, y = Count, group = 1))+
  geom_line()+
  theme(axis.text.x = element_text(angle = 90))
```

(Codes that used in PartB (A)iii in R)

Line1 & 2: change my working directory

Line 3: DF = read.csv(“timestamps.csv”)

read the file timestamps.csv which I created in Part B (A)I and named it as DF.

Line 4: install.packages(“tidyr”)

install library named “tidyr”

Line 5: library(“tidyr”)

call the library “tidyr”

Line 6: DF1 <- separate(data = DF, col = 2, into c(“Date”, “Time”), sep = “ “)

separate the date and time list in DF1 into date column and time column

Line 7: install.packages(“data.table”)

install library named “data.table”

Line 8: library(“data.table”)

call the library “data.table”

Line 9: DF2 <- setDT(DF1[c(2)])[,list(count=.N),names(DF1[c(2)])]

count the number of duplicated dates and present as a table

Name: Lim Yu Jin
Student ID: 32637888

Line 10: DF2$Date

show the date column

```
> getwd()
[1] "C:/Users/yujin/Downloads/MONASH/FIT 1043/Assignment 3"
> setwd("C:\\Users\\yujin\\Downloads\\MONASH\\FIT 1043\\Assignment 3")
> DF = read.csv("timestamps.csv")
> library("tidyr")
> DF1 ← separate(data = DF, col = 2, into = c("Date", "Time"), sep = " ")
> library(data.table)
> DF2 ← setDT(DF1[c(2)])[,list(Count=.N),names(DF1[c(2)])]
> DF2$Date
 [1] "2009-04-06" "2009-04-07" "2009-04-18" "2009-04-19" "2009-04-20" "2009-04-21" "2009-05-01"
 [8] "2009-05-02" "2009-05-03" "2009-05-04" "2009-05-09" "2009-05-10" "2009-05-11" "2009-05-13"
[15] "2009-05-14" "2009-05-16" "2009-05-17" "2009-05-18" "2009-05-21" "2009-05-22" "2009-05-26"
[22] "2009-05-27" "2009-05-28" "2009-05-29" "2009-05-30" "2009-05-31" "2009-06-01" "2009-06-02"
[29] "2009-06-03" "2009-06-04" "2009-06-05" "2009-06-06" "2009-06-07" "2009-06-14" "2009-06-15"
[36] "2009-06-16" "2009-06-17" "2009-06-18" "2009-06-19" "2009-06-20" "2009-06-21" "2009-06-22"
[43] "2009-06-23" "2009-06-24" "2009-06-25" "2009-04-17"
```

Line 11: DF3 <-DF2[order(as.Date(DF2$Date, format=%Y-%m-%d)),]

reformat the DF2 Date column into the data frame show below

Line 12: head(DF3)

show the first 6 lines of the data frame

```
> DF3←DF2[order(as.Date(DF2$Date, format="%Y-%m-%d")),]
> head(DF3)
         Date Count
1: 2009-04-06     2
2: 2009-04-07    23
3: 2009-04-17     2
4: 2009-04-18    21
5: 2009-04-19    25
6: 2009-04-20    26
> ggplot(data = DF3, aes(x = Date, y = Count, group = 1))+
+   geom_line()+
+   theme(axis.text.x = element_text(angle = 90))
>
```

Name: Lim Yu Jin
Student ID: 32637888

Line 13: ggplot(data = DF3, aes(x = Date, y = Count, group = 1))+ geom_line()+ theme(axis.text.x = element_text(angle = 90))

plot a line graph base on the DF3 and rotate the label in x-axis in 90 degrees.