Task A

1) Decompress the file. How big is it?

```
zzl@zzl:~$ cd ~/FIT5145
zzl@zzl:~/FIT5145$ unzip FB_Dataset.csv.zip
Archive:  FB_Dataset.csv.zip
  inflating: FB_Dataset.csv
```

Code: unzip FB_Dataset.csv.zip

This code can unzip the zip file and store the file in the same directory with the original name.

```
zzl@zzl:~/FIT5145$ ls -l -h FB_Dataset.csv
-rw-r--r-- 1 zzl zzl 344M Sep 11 17:21 FB_Dataset.csv
```

Code: ls -l -h FB_Dataset.csv

'-l' means the output is the long format. '-h' means human readable

2) What delimiter is used to separate the columns in the file and how many columns are there?

```
zzl@zzl:~/FIT5145$ head -1 FB_Dataset.csv
page_name,post_id,page_id,post_name,message,description,caption,post_type,status_type,likes_count,comments_count,shares_count,love_count,wow_count,haha_count,sad_co
unt,thankful_count,angry_count,post_link,picture,posted_at
```

Code: head -1 FB_Dataset.csv

This code shows the first line of the file. According to the output we can find the comma is the delimiter

```
zzl@zzl:~/FIT5145$ head -1 FB_Dataset.csv | awk -F',' '{print NF}'
21
```

There are 21 columns.

Code: head -1 FB_Dataset.csv | awk -F',' '{print NF}'

The output is the number of the columns. we use -F',' to set the delimiter and NF is a predefined variable, and the value of NF is the number of fields in the current record.

3) The 2nd column is the unique identifier for a Facebook post. What are the other columns?

Page_name: company name. page_id: is a unique ID of pages. Post_name: the name of the post. Message: the content of the post. Description: the theme of the post. Caption: the website of the post. Post_type: the type of the post. Status_type: the type of the status. Like_count: the number of the like. Comments_count: the number of the comments. Shares_count: the number of the shares. Love_count: the number of the love count. Wow_count: the number of the wow. Haha_count: the number of the haha. Sad_count: the

number of the sad. Angry_count: the number of the angry. Post_link: the link of the post. Picture: the picture in the post. Post_at: the post date.

4) How many Facebook post are there in the file?

```
zzl@zzl:~$ awk -F',' 'NR>1 {print $2}' FB_Dataset.csv|wc -l
533926
```

Answer is 533926

Code: awk -F',' 'NR>1 {print $2}' FB_Dataset.csv | wc -l

we select the second columns of the dataset expect the header, and then count the number of the lines.

5) what is the data range for Facebook posts? (Assume that the data is in order)

```
zzl@zzl:~/FIT5145$ awk -F',' '{print $21}' FB_Dataset.csv | head -n 2 | tail -1
1/1/12 0:30
zzl@zzl:~/FIT5145$ awk -F',' '{print $21}' FB_Dataset.csv | tail -1
7/11/16 23:45
```

Answer is 1/1/12 to 7/11/16

Code: awk -F',' '{print $21}' FB_Dataset.csv | head -n 2 | tail -1

awk -F',' '{print $21}' FB_Dataset.csv | tail -1

We chose the 21th column and just show the second row, because the first row is head. The second command means showing the last row of the 21th column.

6) How many unique pages are there?

```
zzl@zzl:~$ awk -F',' 'NR>1 {print $3}' FB_Dataset.csv|sort|uniq|wc -l
15
```

Answer is 15

Code: awk -F',' 'NR>1 {print $3}' FB_Dataset.csv | sort | uniq | wc -l

The code 'sort' means sort the input. The code 'uniq' means that remove duplicate information. At first, we chose the third column, after that sort them and remove duplicate information. Finally count the number of lines.

7) How many unique posts are there? [Hint: one pages can have multiple posts]

```
zzl@zzl:~$ awk -F',' 'NR>1 {print$2}' FB_Dataset.csv|awk -F'_' '{print$2}'|sort|uniq|wc -l
512808
```

Answer is 512808.

Code: awk -F',' 'NR>1 {print$2}' FB_Dataset.csv|awk -F'_' '{print$2}'|sort|uniq|wc -l

At first, the command (awk -F',' 'NR>1{print$2}' FB_Dataset.csv) select the second column expect the head. Because the second column including the page_id, we use command (awk -F'_' '{print$2}'|sort|uniq|wc -l) to select the post_id, and then sort and remove the duplicate information. Finally count the number of lines

8) when was the first mention in the file regarding "Italian Dishes" and what was the post?

```
zzl@zzl:~/FIT5145$ awk -F',' '/Italian Dishes/{print $21}' FB_Dataset.csv
11/6/15 14:01
```

Code: awk -F',' '/Italian Dishes/{print $21}' FB_Dataset.csv

The date of the first mention is 11/6/15 14:01. The code selects the row which includes 'Italian Dishes' and then print the 21th columns.

```
zzl@zzl:~$ grep -r -n 'Italian Dishes' FB_Dataset.csv
308739:the-huffington-post,18468761129_10153133124136130,18468761129,5 Brilliant Italian Dishes You Haven't Tried Before,Move over  fettuccine alfredo.,No fettuccin
e alfredo or penne alla vodka here. Each of these recipes -- for Italian 'sushi ' roast chicken with a serious kick and more -- offers a fresh approach to the belov
ed cuisine.,huff.to,link,shared_story,397,35,277,0,0,0,0,0,0,http://huff.to/1f4MMOo,https://external.xx.fbcdn.net/safe_image.php?d=AQAJulr3tN0ACu5d&w=130&h=130&url=
http%3A%2F%2Fi.huffpost.com%2Fgen%2F3029028%2Fimages%2Fo-AUTHENTIC-ITALIAN-RECIPES-facebook.jpg&cfs=1,11/6/15 14:01
```

Code: grep -r -n 'Italian Dishes' FB_Dataset.csv

The post show in the picture. The code '-r' means recursive, '-n' means print the number of the line.

9) How many times is "Barack Obama" mentioned in the file? How did you find this? (Do not ignore the case)

```
zzl@zzl:~/FIT5145$ grep -o 'Barack Obama' FB_Dataset.csv | wc -l
6831
```

The answer is 6831

Code: grep -o 'Barack Obama' FB_Dataset.csv | wc -l

The code '-o' means that print only the matched parts of a matching line, with each such part on a separate output line. At first, we select all the Barack Obama in the dataset with each such part on a separate output line, and then count the line number.

10) what about "Donald Trump"? Who is more popular on Facebook, Obama or Trump? (Do not ignore the case)

```
zzl@zzl:~/FIT5145$ grep -o 'Donald Trump' FB_Dataset.csv | wc -l
15024
```

The answer is 15024

Code: grep -o 'Donald Trump' FB_Dataset.csv | wc -l

15024 is bigger than 6831, so Trump is more popular.

11) Select the posts where "Trump" (Ignore the case) is mentioned in the post content and number of likes for those posts are greater than 100. And generate a new file with post_id and sorted like_count and name it "trump.txt". (In the output, you need to show the headers as well) [Hint: Find Trump in message column, i.e., 5th column]. Then copy and paste the first 5 lines of trump.txt in your answer.

```
zzl@zzl:~/FIT5145$ awk -F',' '$10 > 100 && $5 ~ /[Tt]rump/ || NR == 1 {print $2,$10}' FB_Dataset.csv | sort -nk 2 > trump.txt
zzl@zzl:~/FIT5145$ head -5 trump.txt
post_id likes_count
10606591490_10153445206101491 101
 131459315949_10153961477340950 101
6250307292_10154235149992293 101
8304333127_10154089866028128 101
```

Code:

awk -F',' '$10>100 && $5~/[Tt]rump/ || NR == 1 {print $2,$10}' FB_Dataset.csv | sort -nk 2 >trump.txt

head -5 trump.txt

At first, we select the specific information use the command (awk -F',' '$10>100 && $5 ~ /[Tt]rump/ || NR==1 {print $2, $10}' FB_Dataset.csv). The command (NR == 1 {print$2, $10}) means that the output including the first line and just print the second and 10th columns. The command (sort -nk 2) means that sort the data numerically base on like_count. The command (> trump.txt) means that put the information into a file which named trump.txt

12) Find the total number of love_count and angry_count for "Donald Trump" and "Barack Obama" separately. Who has more positive feeling among people? Justify your answer. [Hint 1: you will need to search online to find how to sum a column of numbers using awk. Hint 2: You will need to consider both love and angry count when justifying your answer.]

```
zzl@zzl:~$ awk -F',' '/Donald Trump/{sum+=$13} END {print sum}' FB_Dataset.csv
1565929
zzl@zzl:~$ awk -F',' '/Donald Trump/{sum+=$18} END {print sum}' FB_Dataset.csv
2198153
zzl@zzl:~$ awk -F',' '/Barack Obama/{sum+=$13} END {print sum}' FB_Dataset.csv
836659
zzl@zzl:~$ awk -F',' '/Barack Obama/{sum+=$18} END {print sum}' FB_Dataset.csv
582064
```

Code: awk -F',' '/Donald Trump/{sum+=$13}END{print sum}' FB_Dataset.csv

awk -F',' '/Donald Trump/{sum+=$18}END{print sum}' FB_Dataset.csv

awk -F',' '/Donald Trump/{sum+=$13}END{print sum}' FB_Dataset.csv

awk -F',' '/Donald Trump/{sum+=$18}END{print sum}' FB_Dataset.csv

The love_count of Donald Trump is 1565929. The angry_count of Donald Trump is 2198153.

The love_count of Barack Obama is 836659. The angry_count of Barack Obama is 582064.

```
zzl@zzl:~$ awk -F',' '/Barack Obama/{sum_love+=$13; sum_angry+=$18; rate = sum_love/(sum_love+sum_angry)}END{print rate}' FB_Dataset.csv
0.589727
zzl@zzl:~$ awk -F',' '/Donald Trump/{sum_love+=$13; sum_angry+=$18; rate = sum_love/(sum_love+sum_angry)}END{print rate}' FB_Dataset.csv
0.416019
```

Code:

awk-F',' '/Barack Obama/{sum_love+=$13; sum_angry+=$18; rate=sum_love/(sum_love+sum_angry)}END{print rate}' FB_Dataset.csv

awk-F',' '/Donald Trump/{sum_love+=$13; sum_angry+=$18; rate=sum_love/(sum_love+sum_angry)}END{print rate}' FB_Dataset.csv

The rate of Barack Obama (0.5897272) is bigger than the Donald Trump (0.416019)

Task B

1) How many times does the term 'Trump' appear in the post content? (use shell to answer to this question)

```
zzl@zzl:~/FIT5145$ grep -o 'Trump' FB_Dataset.csv | wc -l
52673
```

Code: grep -o 'Trump' FB_Dataset.csv | wc -l

There are 52673 times the term 'Trump' appear in the post content.

2) 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 posts referring to Trump using 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.] 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:

https://www.rdocumentation.org/packages/base/versions/3.6.1/topics/strptime 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?

```
zzl@zzl:~/FIT5145$ awk -F ',' '/Trump/ || NR == 1 {print $21}' FB_Dataset.csv > Trump_date.txt
zzl@zzl:~/FIT5145$
```

Code:

awk -F',' '/Trump/ || NR == 1 {print $21}' FB_Dataset.csv > Trump_date.txt

We use command 'awk' to select the values of post_at which includes Trump, and then store them into a file named Trump_date.txt

```
> trump_df <- read.csv('Trump_date.txt', header = TRUE)
> head(trump_df)
       posted_at
1  29/1/12 19:48
2  30/1/12 21:07
3   2/2/12 15:53
4    3/4/12 0:49
5   5/10/12 2:00
6 24/10/12 17:11
>
```

Code: trump_df <- read.csv('Trump_date.txt', header = TRUE)

head(trump_df)

Using R language to read the file named Trump_date.txt into a data frame format, and then display the first 6 lines.

```
> trump_df$date <- trump_df$posted_at
```

Code: trump_df$date <- trump_df$posted_at

Using R language to set a new column named date and store the same data with posted_at column.

```
> trump_df$date <- strptime(trump_df$date, "%d/%m/%y %H:%M")
```

Code: trump_df$date <- strptime(trump_df$date, "%d/%m/%y %H:%M")

Converting date format data to string format

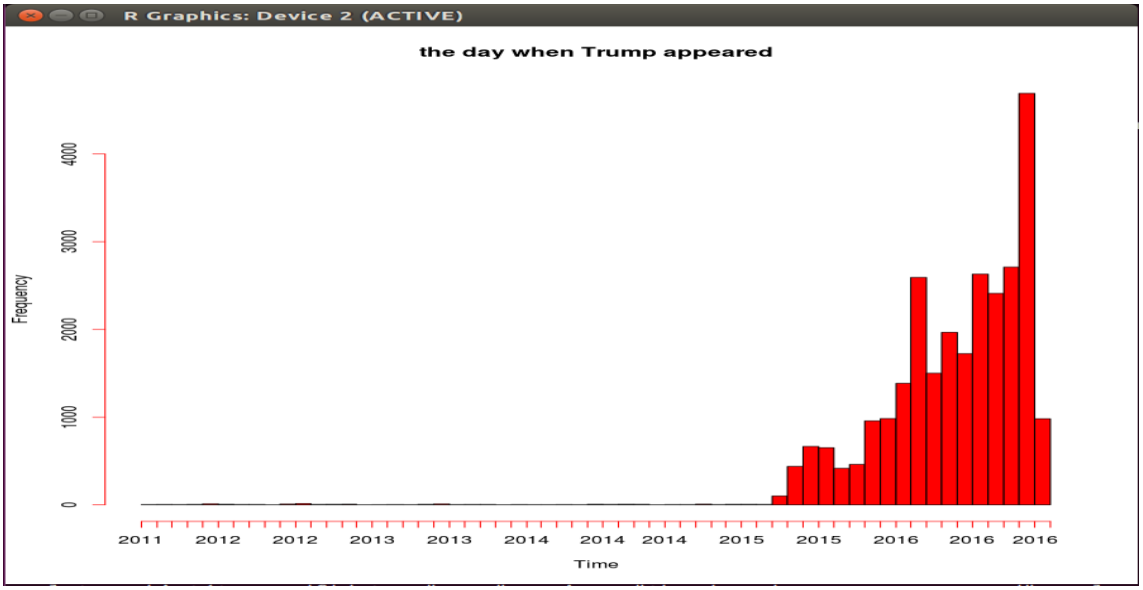
1. Once you have converted the timestamps, use the hist() function to plot the data in R.

```
> plot <- hist(trump_df$date, "months", main = "the day when Trump appeared", col = "red", xlab="Time", freq=TRUE)
>
```

Code:

plot <- hist(trump_df$date, “months”, main = “the day when Trump appeared”, col = “red”, xlab=”Time”, freq=TRUE)

Draw the line of date as a histogram. The figure named ‘the day when Trump appeared’, and the break is month.



2. The plot has a bit of an unusual shape. Describe the pattern you see.

According to the figure, very large fluctuations occurred at the end of 2016

3) In this question, we want to investigate the Facebook posts of a few top media sources. To answer this question, you will need to extract the facebook posts made on the pages of "abc-news", "cnn" and "fox-news" from your original Facebook dataset.

1. Use the unix shell to first generate a file containing all the records belonging to "abc-news", "cnn" and "fox-news" only. Then read the resulting file in R.

```
zzl@zzl:~/FIT5145$ awk -F',' '$1~"abc-news"; $1 ~ "cnn"; $1 ~ "fox-news" || NR==1{print $0}' FB_Dataset.csv > extract.txt
```

Code:

awk -F',' '$1 ~ "abc-news"; $1 ~ "cnn"; $1 ~ "fox-news" || NR==1{print $0}' FB_Dataset.csv > extract.txt

The command selects the specific row which including 'abc-news', 'cnn' and 'fox-news', and then put these rows with the header into a new file named extract.txt

2. Background: We now want to see if any relationship exists between the number of times a post is shared on Facebook and the number of likes it generates. Task: Use appropriate R code to generate a plot showing the relationship between the number of shares and the number of likes in your dataset. Do you see any relationship?

```
> name_df <- read.csv('extract.txt', header = TRUE)
```

Code:

name_df <- read.csv('extract.txt', header = TRUE)

Using the R language to read the file named extract.txt to data frame with the header.

```
> name_df<-name_df[name_df$shares_count !=0,]
> name_df<-name_df[name_df$likes_count !=0,]
> likes_count<-name_df$likes_count
> shares_count<-name_df$shares_count
```

Code:

name_df <- name_df[name_df$shares_count !=0,]

name_df <- name_df[name_df$likes_count !=0,]

likes_count<-name_df$likes_count

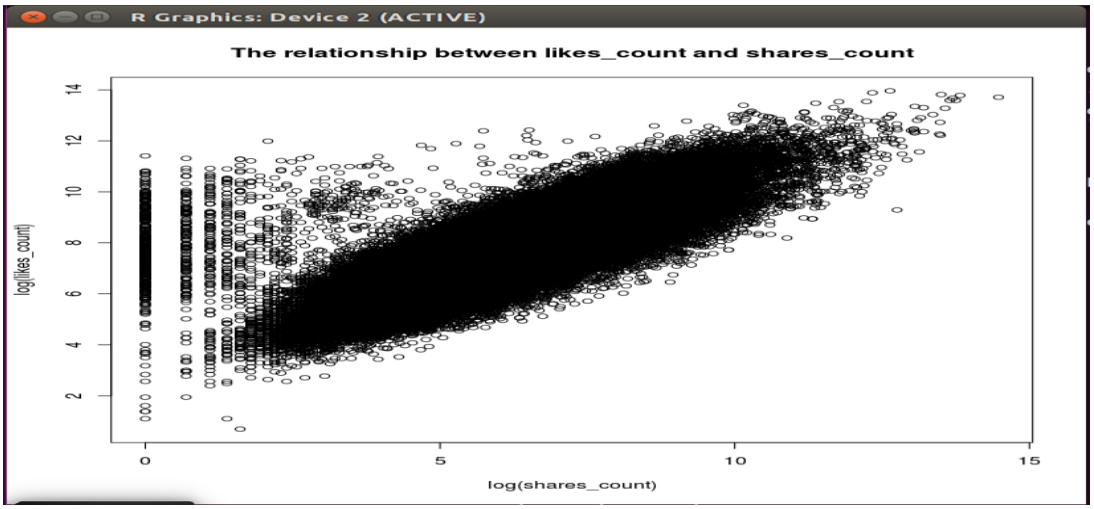
shares_count <- name_df$shares_count

Selecting the row which shares_count and likes_count is not 0.

```
> plot(log(likes_count)~log(shares_count), main="The relationship between likes_count and shares_count")
>
```

code: plot(log(likes_count)~log(shares_count), main="The relationship between likes_count and shares_count")

Using the R language to plot the data. X-axis is log(shares_count), y-axis is log(likes_count), and the figure named 'The relationship between likes_count and shares_count'.



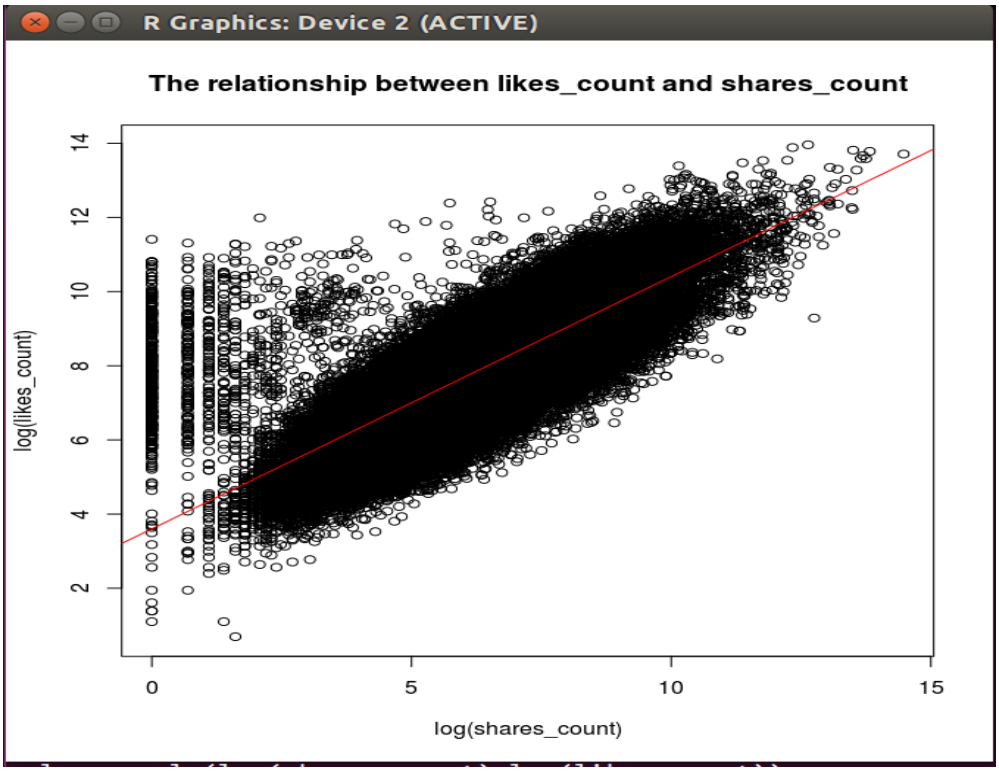
3. Fit a linear regression model using R to the above data (i.e., shares_count and likes_count) and plot the linear fit. Does it look like a good fit to you?

```
> lm.reg<-lm(log(likes_count)~log(shares_count))
> abline(lm.reg, col='red')
```

Code: lm.reg<-lm(log(likes_count)~log(shares_count))

abline(lm.reg, col='red')

Set a linear regression named lm.reg and plot this regression in the figure with red color.



4. Use the linear fit to predict the number of likes a post will generate if it is shared 0 times, 1000 times, 10000 times and 100000 times on Facebook.

```
> predict(lm.reg,newdata=data.frame(shares_count=0),interval='confidence')
   fit lwr upr
1 -Inf NaN NaN
```

Code: predict(lm.reg,newdata=data.frame(shares_count=0),interval='confidence')

We can not predict the number of likes when it shared 0 times.

```
> predict(lm.reg,newdata=data.frame(shares_count=1000),interval='confidence')
       fit      lwr      upr
1 8.301746 8.29593 8.307562
> exp(8.301746)
[1] 4030.904
```

Code: predict(lm.reg,newdata=data.frame(shares_count=1000),interval='confidence')

When it is shared 1000 times, the number of likes is 4030.904

```
> predict(lm.reg,newdata=data.frame(shares_count=10000),interval='confidence')
       fit      lwr      upr
1 9.868402 9.857677 9.879126
> exp(9.868402)
[1] 19310.46
```

Code: predict(lm.reg,newdata=data.frame(shares_count=10000),interval='confidence')

When it is shared 10000 times, the number of likes is 19310.46

```
> predict(lm.reg,newdata=data.frame(shares_count=100000),interval='confidence')
       fit      lwr      upr
1 11.43506 11.41779 11.45232
> exp(11.43506)
[1] 92508.89
```

Code: predict(lm.reg,newdata=data.frame(shares_count=100000),interval='confidence')

When it is shared 10000 times, the number of likes is 92508.89.