Task A: Investigating Facebook Data using shell commands

1) The file is 344MB
The command:

```
$ unzip FB_Dataset.csv.zip
$ ls -lh FB_Dataset.csv
```

2) The delimiter is comma (,) and there are 21 columns in total.
The command:

```
$ head -1 FB_Dataset.csv
$ awk -F ',' 'NR==1{print NF}' FB_Dataset.csv
```

3) The other column apart from second column are as the following:

1. page_name
2. page_id
3. post_name
4. message
5. description
6. caption
7. post_type
8. status_type
9. likes_count
10. comments_count
11. shares_count
12. love_count
13. wow_count
14. haha_count
15. sad_count
16. thankful_count
17. angry_count
18. post_link
19. picture
20. posted_at

The command:

```
$ awk -F ',' 'NR==1{$2="" ;print}' FB_Dataset.csv
```

4) There are 533926 posts in the file.
The command:

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

5) The date range is (1/1/12 0:30, 7/11/16 23:45) assuming that the data is already in order.
The command:

```
$ awk -F ',' '{if(NR>1) print $21}' FB_Dataset.csv | head -n 1
$ awk -F ',' '{if(NR>1) print $21}' FB_Dataset.csv | tail -n 1
```

6) There are 15 unique pages in total.
The command:

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

7) There are 533925 unique posts in total.
The command:

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

8) The first mention of “Italian Dishes” is in the post “5 Brilliant Italian Dishes You Haven't Tried Before”
The command:

```
$ awk -F ',' '/Italian Dishes/ || head {print $4}' FB_Dataset.csv
```

9) Barack Obama has been mentioned 6831 times in the file

The command:

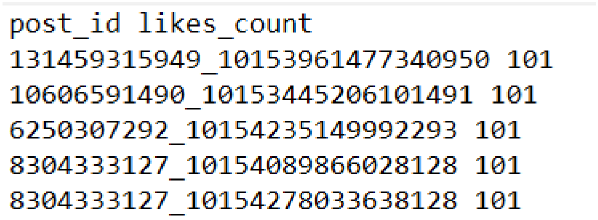
```
$ grep -o 'Barack Obama' FB_Dataset.csv | wc -l
```

10) Donald Trump has been mentioned 15024 times in the file. In comparison, Trump is more popular than Obama.

The command:

```
$ grep -o 'Donald Trump' FB_Dataset.csv | wc -l
```

11) The first 5 lines of tump.txt are



```
post_id likes_count
131459315949_10153961477340950 101
10606591490_10153445206101491 101
6250307292_10154235149992293 101
8304333127_10154089866028128 101
8304333127_10154278033638128 101
```

The command:

```
$ awk -F ',' '$5~"[Tt]rump" && $10>100 || NR==1 {print $2,$10}'
FB_Dataset.csv | sort -nk2 >trump.txt
```

12) People tend to have more positive feeling toward Obama since he has more like than angry.

	Like	Angry
Trump	1565929	2198153
Obama	836659	582064

The command:

```
$ awk -F ',' '/Donald Trump/ {sum_like+=$13; sum_angry+=$18} END{printf
"Trum FB_Dataset.csv
```

```
$ awk -F ',' '/Barack Obama/ {sum_like+=$13; sum_angry+=$18} END{printf
"Obam FB_Dataset.csv
```

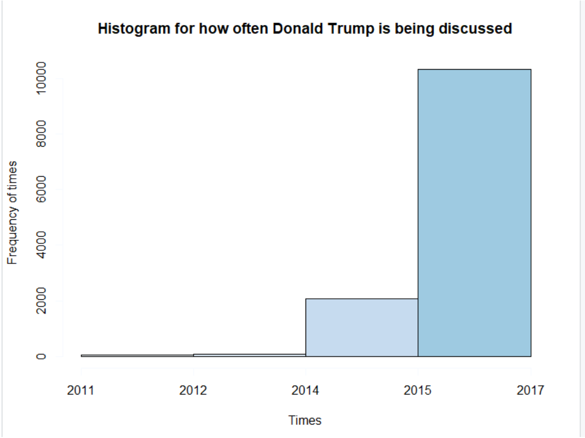
Task B: Graphing the Data in R

1) The term ‘Trump’ appears in the post content 52673 times
The command:

```
$ grep -o 'Trump' FB_Dataset.csv | wc -l
```

2)

2.1



2.2 Trump has just been discussed more from 2014 onward and he is much more famous on 2015 since the election of USA had been held around that time. Previously, few people acknowledge his presence on social media particularly, on Facebook

The command:

(In shell)

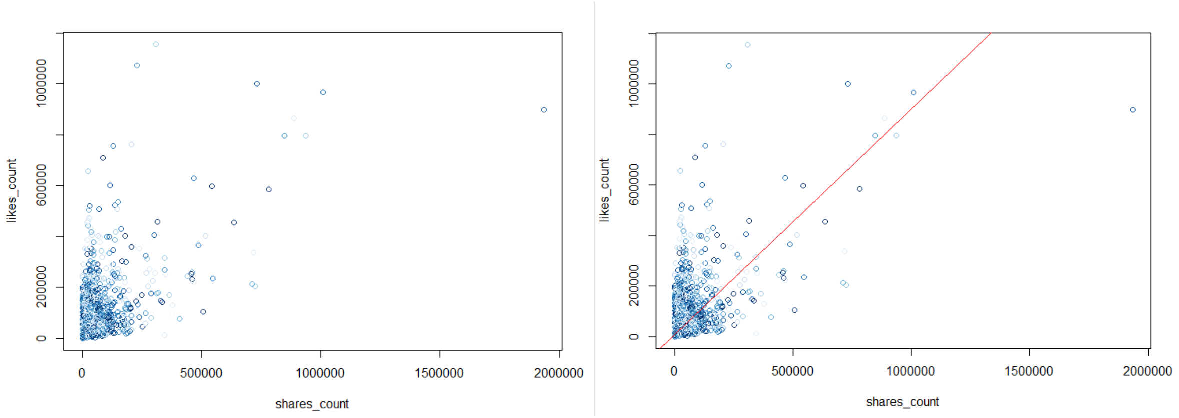
```
$ unzip -p FB_Dataset.csv.zip | awk -F ',' '/Donald Trump/ {print $21}'> 
time.txt
```

(In R)

```
df<-read.csv('time.txt',header=FALSE)
df2<-strptime(df$V1,"%d/%m/%y %H:%M")
hist(df2,breaks=5,
     freq=TRUE,
     main="Histogram for how often Donald Trump is being discussed",
     xlab = "Times",
     ylab = "Frequency of times",
     col = blues9)
```

3)

3.1-3.3 From the figure below, you can see that the more people like the post, the more they share. Therefore, we can say that there is some trend between like count and share count. Let’s look at the correlation by using cor() in R and fit linear regression.



Correlation is 0.6260144

The linear fit doesn’t seem to be a good fit since the data don’t fall around the line. Some are spreading.

3.4 The prediction of number of like(Y) on number of share (X):

X	0	1000	10000	100000
Y	5408.544	6304.821	14371.319	95036.292

(In shell)

```
$ unzip -p FB_Dataset.csv.zip | awk -F ',' '$1=="abc-news" || $1=="cnn" || $1=="fox-news" || NR==1 {print}' > page.txt
```

(In R)

```
ef<-read.csv('page.txt', header=TRUE)
ef2<- subset( ef, select=c('likes_count', 'shares_count'))
plot(likes_count~shares_count,data = ef2,col=blues9)
fit<-lm(likes_count~shares_count,data = ef2)
abline(fit,col='red')
cor(ef2)
predict(fit,data.frame(shares_count=c(0, 1000, 10000, 100000)))
```