DISHI JAIN
30759307
FIT5145 ASSIGNMENT 3

TASK - A

1. unzip FB_Dataset.csv.zip
   ls -lh FB_Dataset.csv
   Output - The file is 344 MB in size
2. head -1 FB_Dataset.csv
   Output - The delimiter used is comma
   awk -F',' 'NR==1 {print NF}' FB_Dataset.csv
   Output - There are 21 columns in the file
3. cut -d ',' -f 1,3- FB_Dataset.csv | head -1
4. cat FB_Dataset.csv | wc -l
   Ans- There are 533926 Facebook posts in the file
5. awk -F',' 'NR==2 {print $21} END{print $21}' FB_Dataset.csv
   Ans - The date range is -
   1/1/12 0:30
   7/11/16 23:45
6. awk -F ',' '{print $3}' FB_Dataset.csv | uniq | grep -v "page_id" | wc -l
   Ans - There are 15 unique pages
7. awk -F ',' 'NR>1 {print $2}' FB_Dataset.csv | tr '_' ' ' | awk -F ' ' '{print $2}' | uniq | wc -l
   Ans - There are 533925 unique posts
8. grep -m 1 "Italian Dishes" FB_Dataset.csv | awk -F ',' '{print $4,$21}'
   Ans - 5 Brilliant Italian Dishes You Haven't Tried Before 11/6/15 14:01
9. grep "Barack Obama" FB_Dataset.csv | wc -l
   So “Barack Obama” (not ignoring case) occurs 6457 times.
   I found the occurrence of “Barack Obama” in the entire file using the grep command. It returns the lines containing the keyword. Hence, I did a word count of lines.
10. grep "Donald Trump" FB_Dataset.csv | wc -l
    So Donald Trump appears 12450 times. (not ignoring case).
    Ans - Hence Donald Trump is more popular on Facebook.
11. awk -F',' 'tolower($5) ~ /trump/ && $10>100 || NR==1 {print $2,$10}' FB_Dataset.csv | sort -n -k2,2 >trump.txt
    head -6 trump.txt
    Ans -
    post_id likes_count
    10606591490_10153445206101491 101
    131459315949_10153961477340950 101
    6250307292_10154235149992293 101
    8304333127_10154089866028128 101
    8304333127_10154278033638128 101

DISHI JAIN
30759307
FIT5145 ASSIGNMENT 3

12. awk -F',' '/Donald Trump/ {sum1+=$13;sum2+=$18} END{print sum1,sum2}' FB_Dataset.csv
Love count is- 1565929
Angry count is- 2198153

awk -F',' '/Barack Obama/ {sum1+=$13;sum2+=$18} END{print sum1,sum2}' FB_Dataset.csv
Love count is- 836659
Angry count is- 582064

To calculate the person who has a more positive feeling among people, we can add the love count to the negative of angry count.
For Donald Trump -
1565929 - 2198153 = -632224

For Barack Obama -
836659 - 582064 = 254595

254595 > -632224.
Hence as the final count is more for Barack Obama, we can say that Barack Obama is more liked as compared to Donald Trump. The Love count and Angry count is hence very useful to determine the likeliness of any post.
Hence, Barack Obama has a more positive feeling amongst the people.

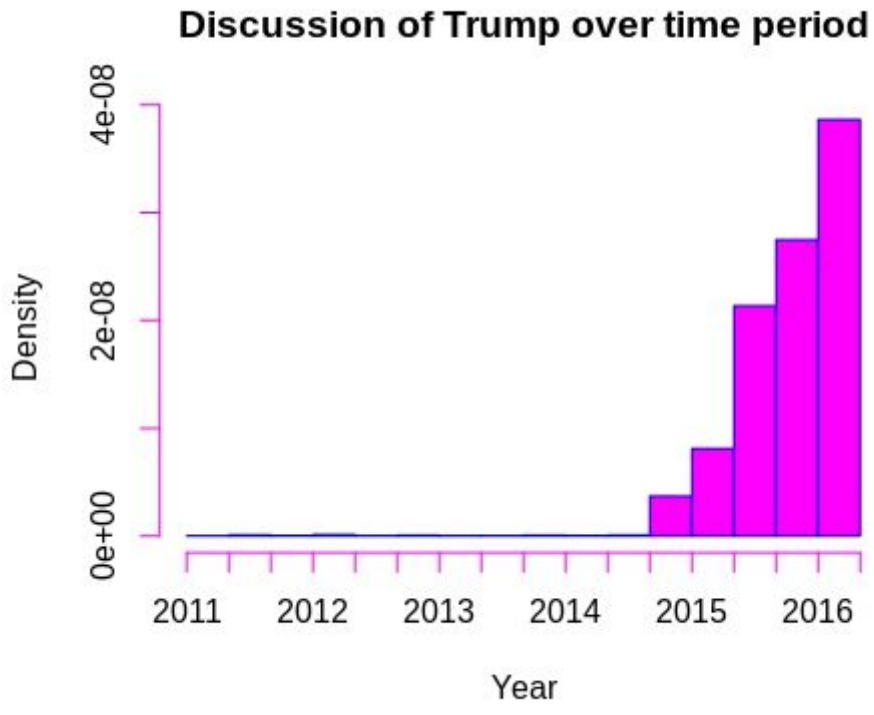
DISHI JAIN
30759307
FIT5145 ASSIGNMENT 3

TASK - B

1. cat FB_Dataset.csv | awk -F ‘,’ '{sum+=gsub(/Trump/, " ")} END{ print sum}'
   52673
   Hence “Trump” appears as it is in the post content 52673 times.

   (Reference taken from http://bigdatums.net/2016/10/18/count-number-of-occurrences-of-characters-with-awk/)

2. 1.) awk -F ',' '/Trump/ || NR==1 {print $21}' FB_Dataset.csv > datetime.txt
   In R Studio -
   output <- read.csv('datetime.txt',header=TRUE)
   time <-strptime(output[["posted_at"]], "%d/%m/%y")
   hist(time, breaks=20, main="Discussion of Trump over time period", xlab="Year" , col=’magenta’, border="blue")



DISHI JAIN
30759307
FIT5145 ASSIGNMENT 3

2.) The plot has an unusual shape. The density of discussions related to “Donald Trump” is seen to increase during the middle of the years 2014 and 2015. The pattern is hence seen to increase steadily after the middle of the year 2014. As Donald Trump came into elections during this time period, hence the discussion increased during that time. The pattern in this histogram is hence almost negligible during the beginning and increases during the end.

3. 1.)

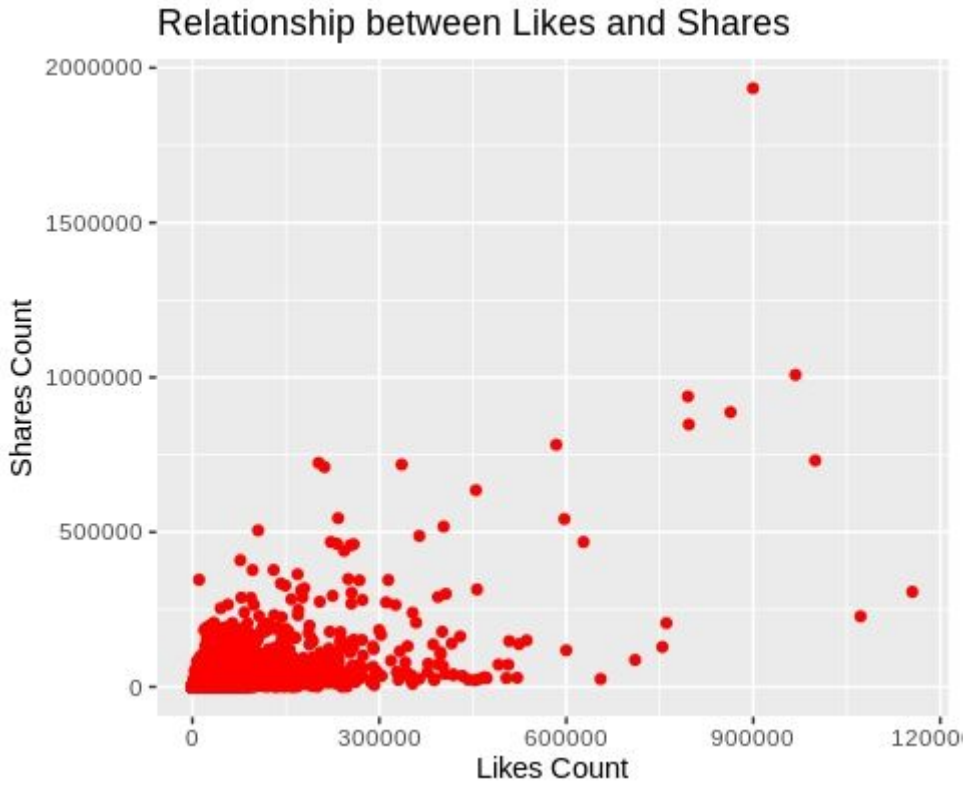
```
awk -F ',' '$1=="abc-news" || $1=="cnn" || $1=="fox-news" || NR==1 {print}'
FB_Dataset.csv > new_file.csv
```

In R Studio-

```
mydata <- read.csv('new_file.csv', header=TRUE,sep=',')
```

2.)

```
install.packages("ggplot2")
library(ggplot2)
ggplot(data=mydata, mapping=aes(x=likes_count,y=shares_count)) +
geom_point(col=”red”) + labs(x="Likes Count" , y = "Shares Count" ,title =
"Relationship between Likes and Shares")
```

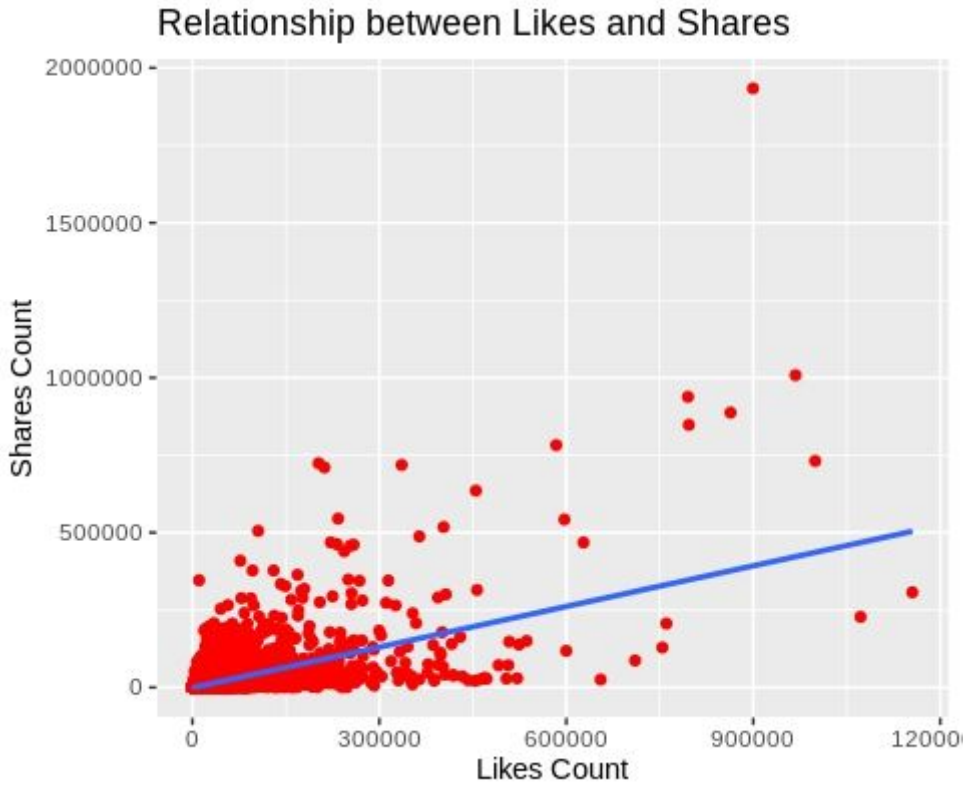


DISHI JAIN
30759307
FIT5145 ASSIGNMENT 3

No there is no relationship between Likes Count and Shares Count. There is no dependency seen between the two values. The scatter points are closely packed at one place only. A very few scatter points are spread along the graph. This increases the chances of error in prediction.

3.)

```
library(ggplot2)
ggplot(data=mydata, mapping=aes(x=likes_count,y=shares_count)) +
geom_point(col=”red”) + geom_smooth(method=lm) + labs(x="Likes Count" , y =
"Shares Count" , title = "Relationship between Likes and Shares")
```



According to me, it is a bad fit. The scatter points are closely packed for the most part. This makes the predictions difficult. We cannot predict easily the shares according to the likes count and vice versa. There are many outliers present in the graph also.

DISHI JAIN
30759307
FIT5145 ASSIGNMENT 3

4.) install.packages("tigerstats")
require(tigerstats)
new <- lmGC(likes_count ~ shares_count, data = mydata)
predict(new , x=0)
Ans - Predict likes_count is about 5409,
give or take 16410 or so for chance variation.

predict(new , x=1000)
Ans - Predict likes_count is about 6305,
give or take 16410 or so for chance variation.

predict(new , x=10000)
Ans - Predict likes_count is about 14370,
give or take 16410 or so for chance variation.

predict(new , x=100000)
Ans - Predict likes_count is about 95040,
give or take 16420 or so for chance variation.