FIT5045 Introduction to Data Science

Assignment 3 submission

Due 30 September 2019, 11.55 p.m.

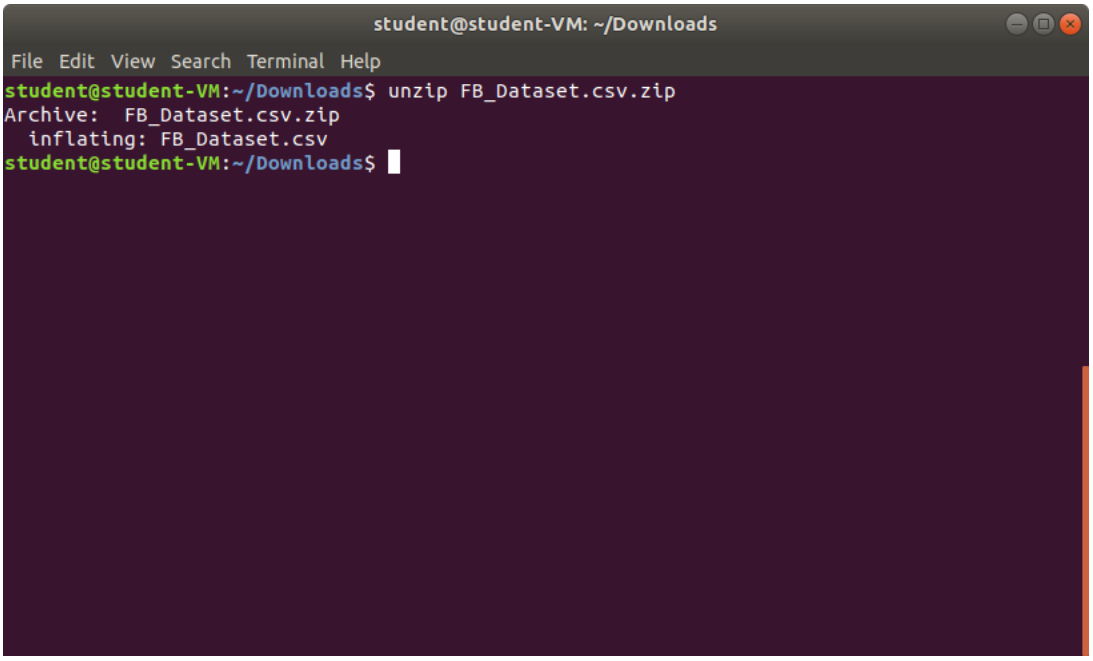
Name (as per enrolment)	Farhad Ullah Rezwan
Student ID	30270111
Allocated Tutorial Class / Day / Time	Activity 30/Thursday/18:00 to 20:00
Name of Tutor	Dilini Rajapaksha Hewa Ranasinghage
Date of Submission	27 September 2019.

Task A: Investigating Facebook Data using shell commands

1. Decompressing FB_Dataset.csv.zip

Shell command:
unzip FB_Dataset.csv.zip

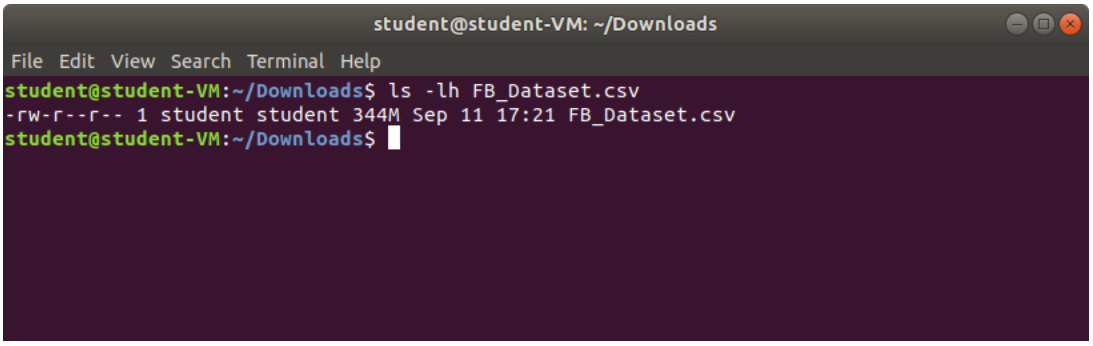
Result:



Size of csv file:

Shell command:
ls -lh FB_Dataset.csv

Result:
-rw-r--r-- 1 student student 344M Sep 11 17:21 FB_Dataset.csv



So, total file size is 344 Megabyte(MB)

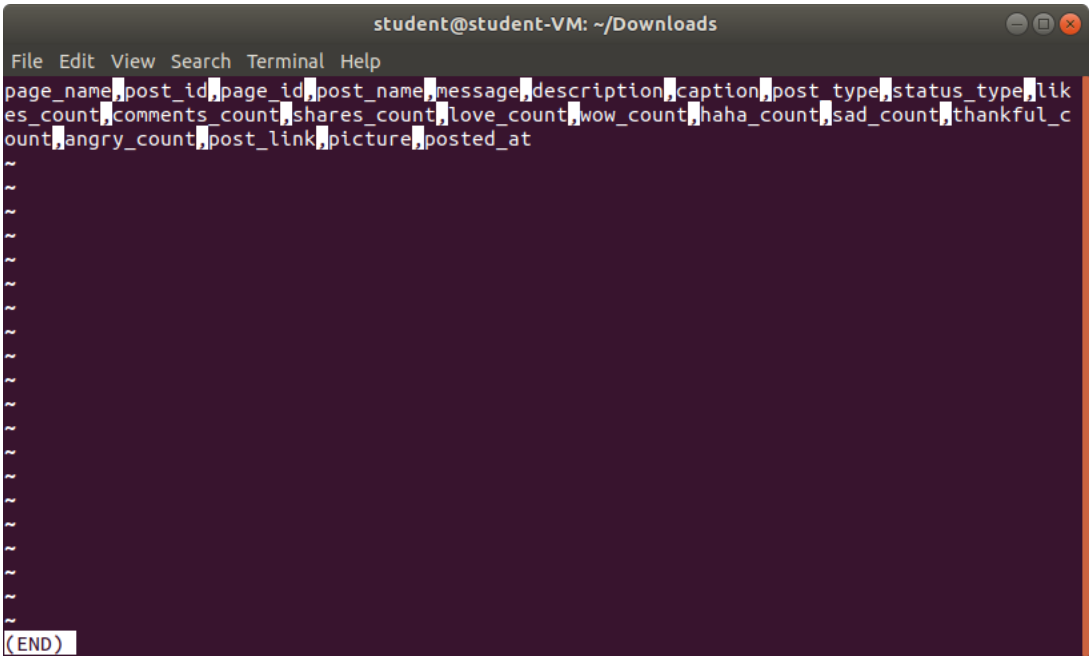
2. Delimiter used to separate the columns:

Shell command:

head -1 FB_Dataset.csv | less

Result:

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_count,thankful_count,angry_count,post_link,picture,posted_at



So, result shows that the delimiter used to separate the columns in the file is “,”

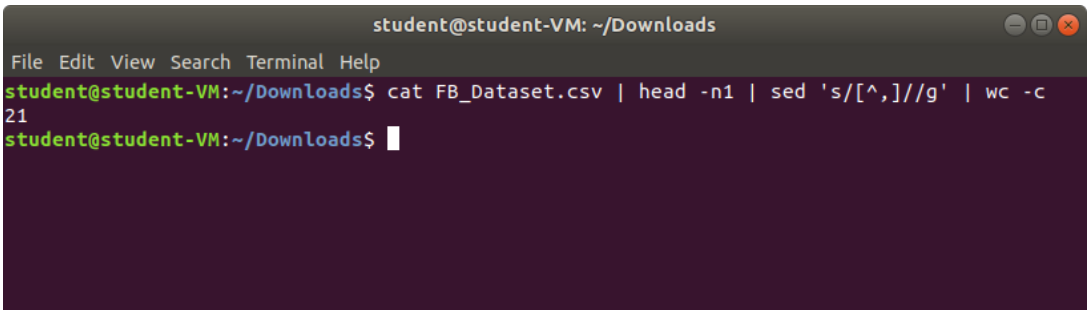
Number of columns :

Shell command:

cat FB_Dataset.csv | head -n1 | sed 's/[^,]//g' | wc -c

Result:

21



So, result shows that there are 21 columns.

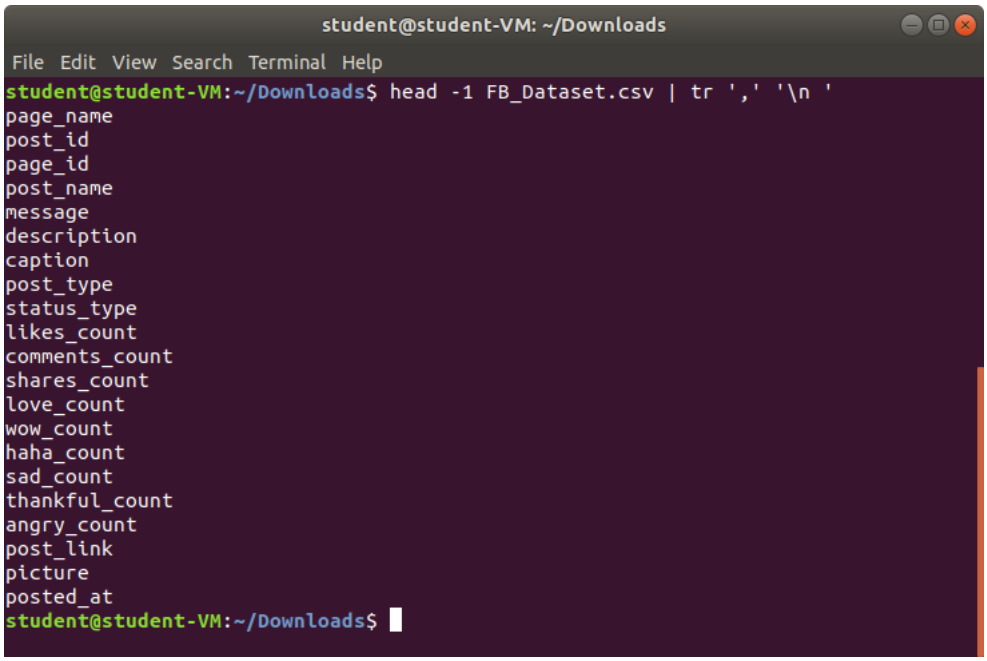
3. Unique identifier = 2nd column, Other columns:

Shell command:

```
head -1 FB_Dataset.csv | tr ',' '\n '
```

Result:

```
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_count
thankful_count
angry_count
post_link
picture
posted_at
```



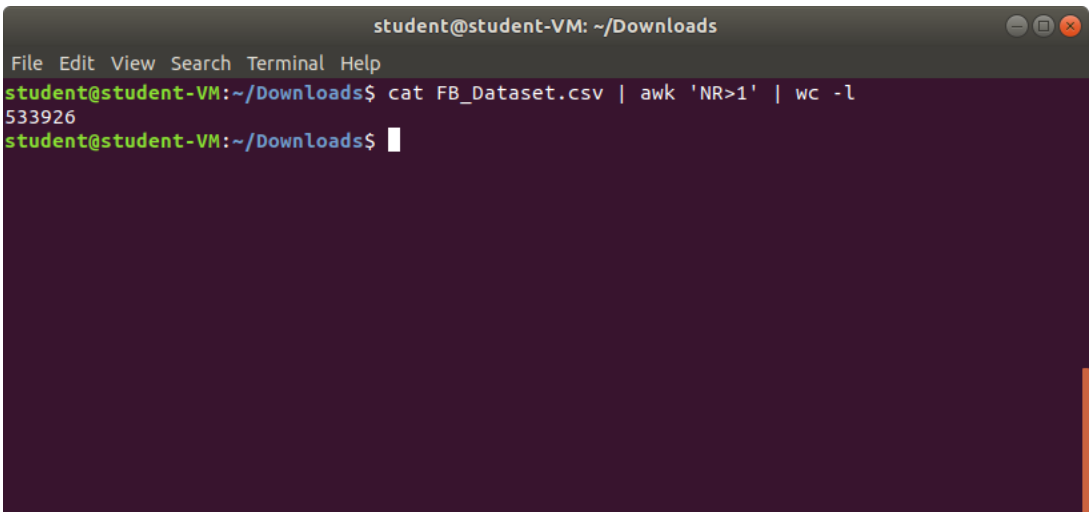
4. Number of Facebook posts in the file

Shell command:

```
cat FB_Dataset.csv | awk 'NR>1' | wc -l
```

Result:

533926



So, there are 533926 number of Facebook Posts are there in the file.

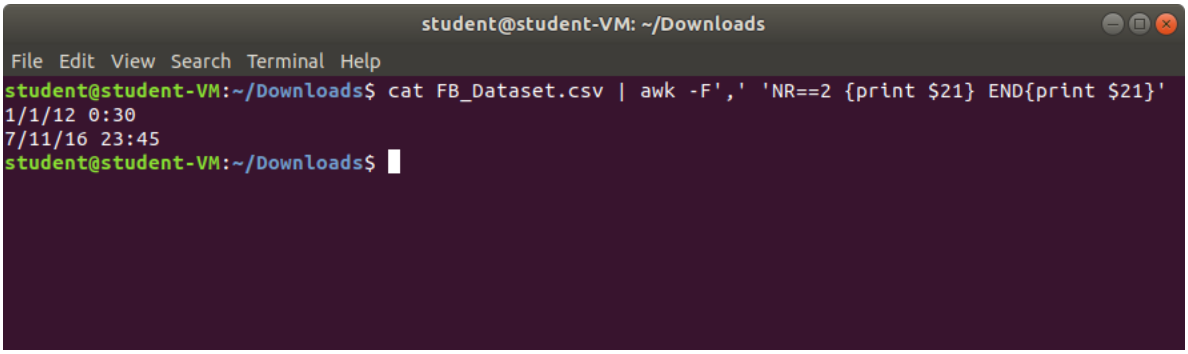
5. Date range for Facebook Posts in this file?

Shell command:

```
cat FB_Dataset.csv | awk -F',' 'NR==2 {print $21} END{print $21}'
```

Result:

1/1/12 0:30
7/11/16 23:45



So the date range is from “1/1/12 0:30” to “7/11/16 23:45”.

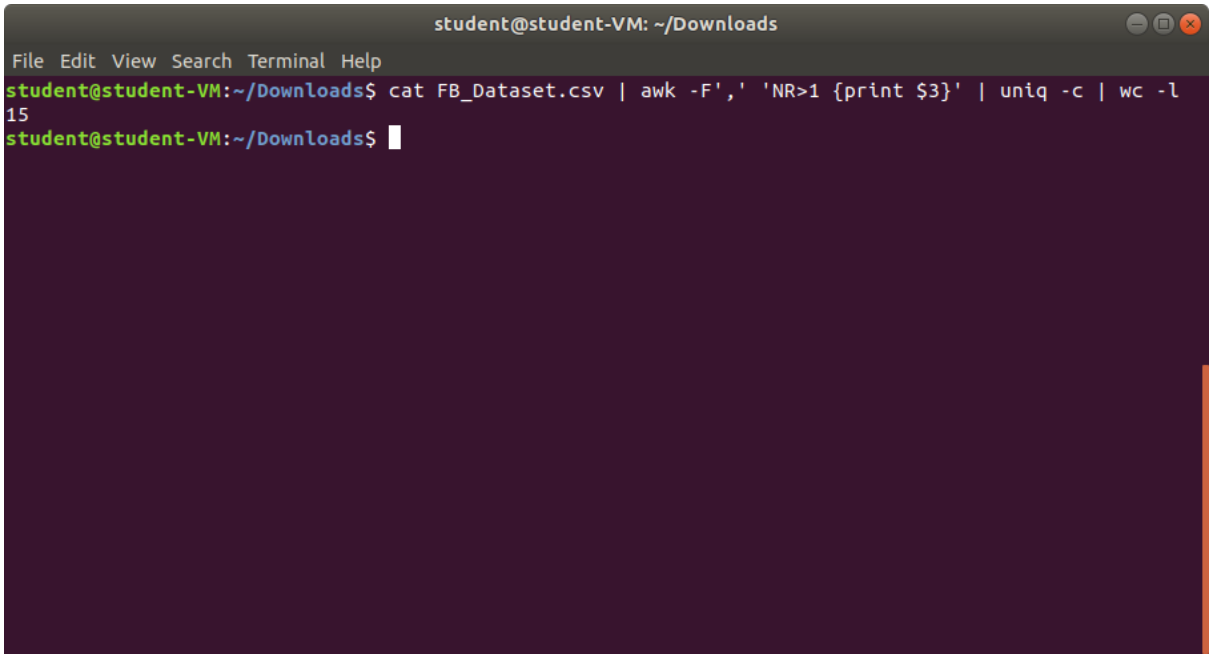
6. Number of unique pages in the file:

Shell command:

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

Result:

15



So, there are 15 unique pages in the csv file.

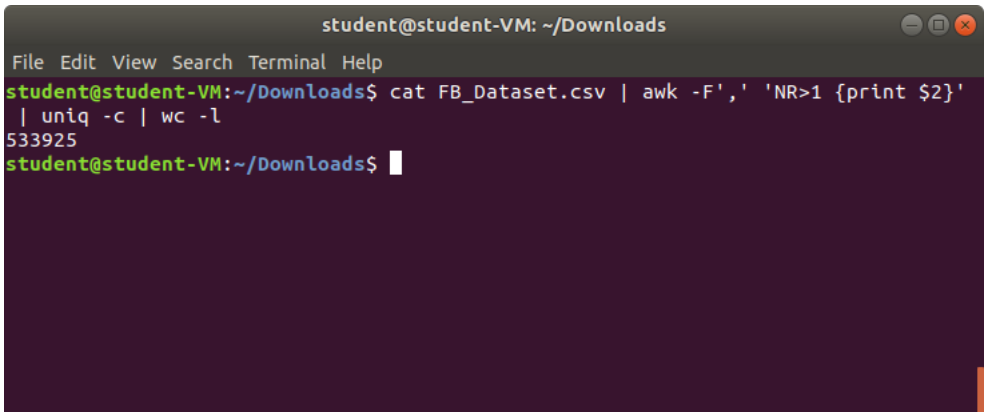
7. Number of unique posts in the file:

Shell command:

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

Result:

533925



So, there are 533925 unique posts are there.

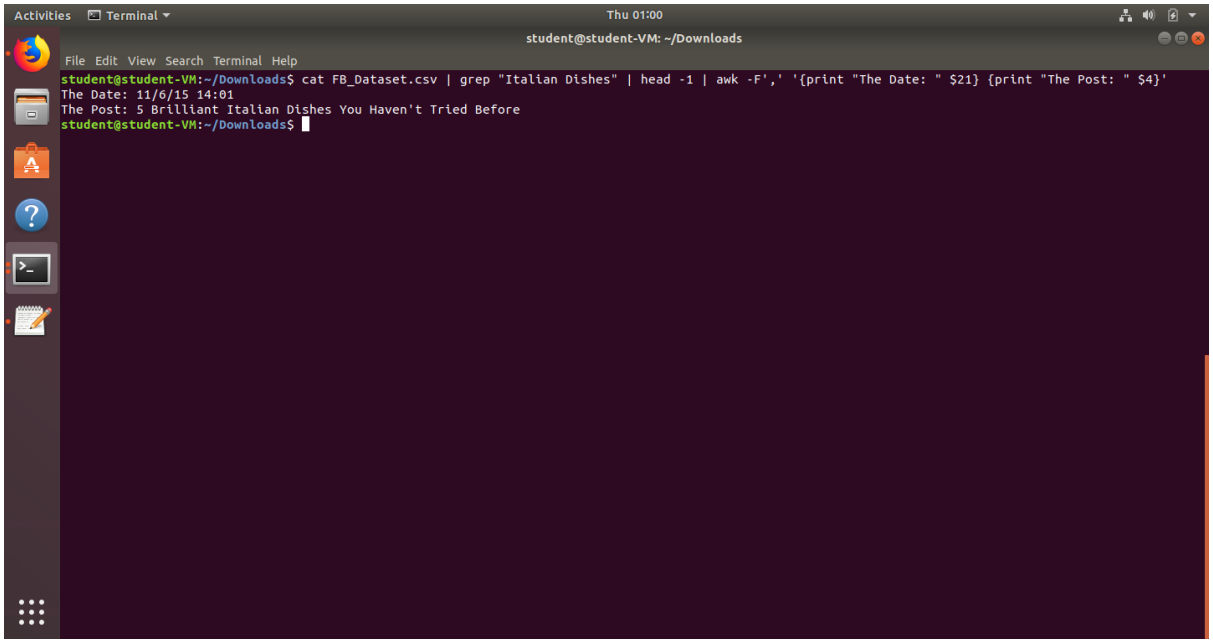
8. The first mention in the file regarding “Italian Dishes” and the post:

Shell command:

```
cat FB_Dataset.csv | grep "Italian Dishes" | head -1 | awk -F',' '{print "The Date: " $21} {print "The Post: " $4}'
```

Result:

```
The Date: 11/6/15 14:01
The Post: 5 Brilliant Italian Dishes You Haven't Tried Before
```



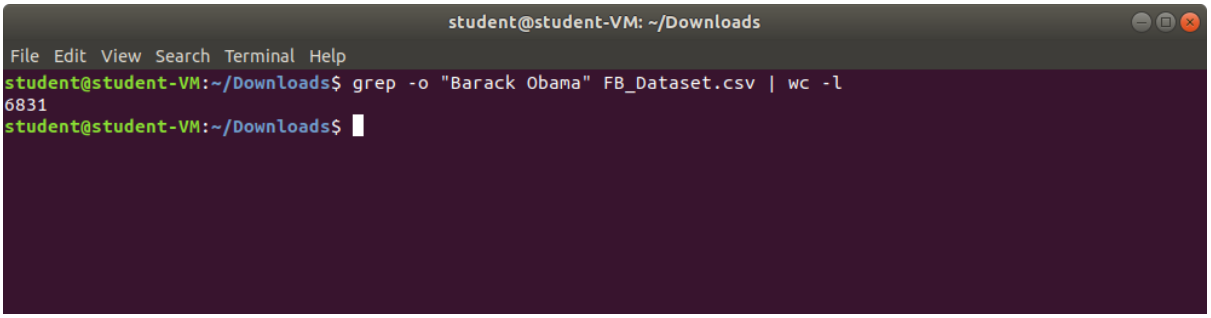
9. Number of times “Barack Obama” mentioned in the file?

Shell command:

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

Result:

```
6831
```



Barack Obama appeared 6831 times in the file.

How to find this:

Using “grep -o” command first the matching part of the post line is shown as separate line like following:

```
student@student-VM: ~/Downloads
File  Edit  View  Search  Terminal  Help
student@student-VM:~/Downloads$ grep -o "Barack Obama" FB_Dataset.csv
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
Barack Obama
```

Used pipe to use that output of the “grep -o” command and after that counted the number of lines using “wc -l” command.

10. What about “Donald Trump”, Who is more popular on Facebook, Obama or Trump?

Shell command:

grep -o "Donald Trump" FB_Dataset.csv | wc -l

Result:

15024

```
student@student-VM: ~/Downloads
File  Edit  View  Search  Terminal  Help
student@student-VM:~/Downloads$ grep -o "Donald Trump" FB_Dataset.csv | wc -l
15024
student@student-VM:~/Downloads$
```

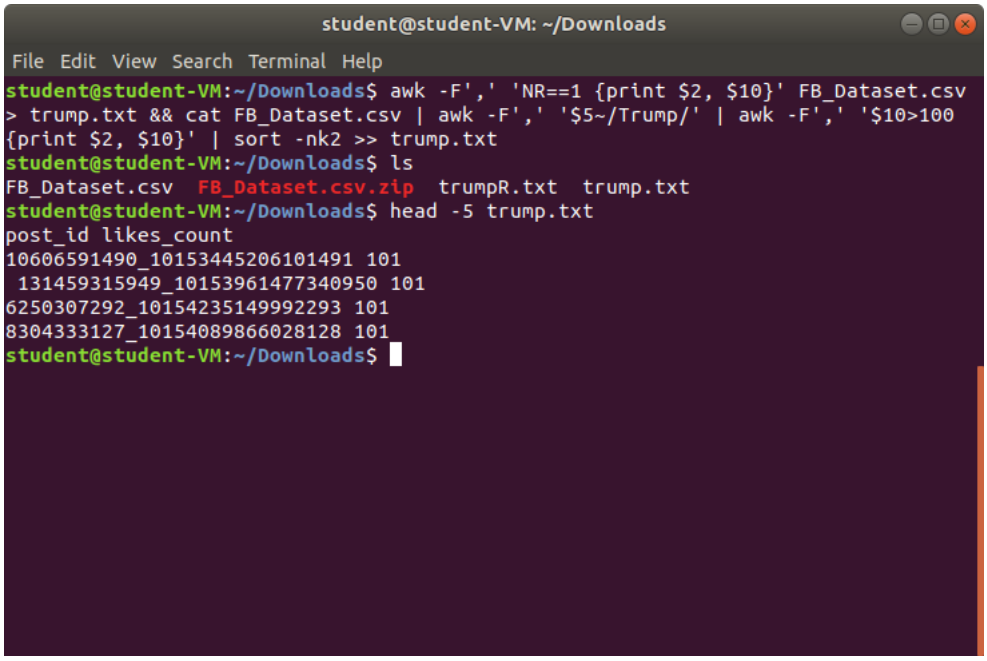
So, Donald Trump Appeared 15024 times in the file.

So, Donald Trump is more popular with 15024 appearance in the file then Barack Obama, who has 6831 appearance.

11. Select the post where “Trump” is mentioned in the post which has more then 100 number of likes, Generate a new file with post id(naming it “trump.txt”).

Shell command and Results:

```
student@student-VM:~/Downloads$ awk -F',' 'NR==1 {print $2, $10}'
FB_Dataset.csv> trump.txt && cat FB_Dataset.csv | awk -F',' '$5~/Trump/' | awk -F',' '$10>100 {print $2, $10}' | sort -nk2 >> trump.txt
student@student-VM:~/Downloads$ ls
FB_Dataset.csv  FB_Dataset.csv.zip  trumpR.txt  trump.txt
student@student-VM:~/Downloads$ head -5 trump.txt
post_id likes_count
10606591490_10153445206101491 101
 131459315949_10153961477340950 101
6250307292_10154235149992293 101
8304333127_10154089866028128 101
```



First Five line of the text:

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

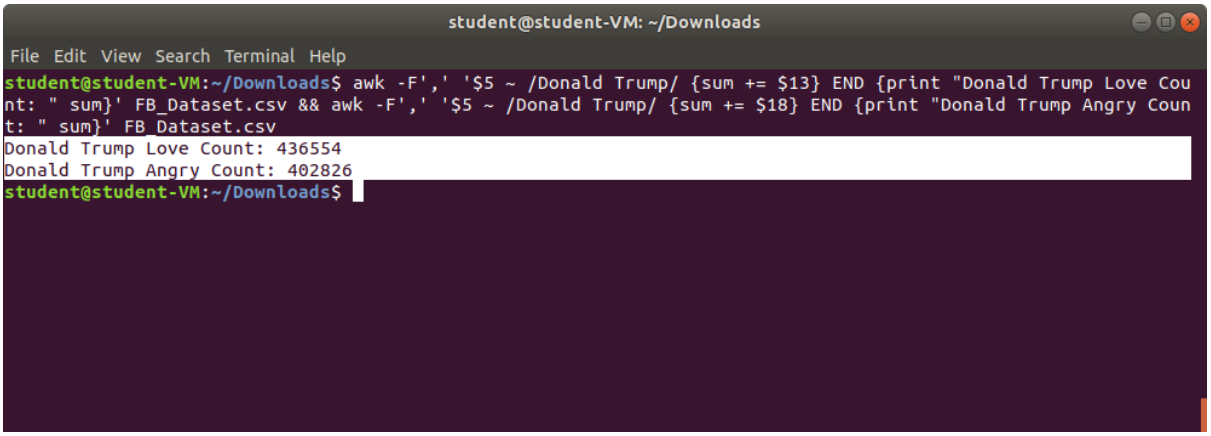
12. Find total number of love_count, and angry_count for “Donald Trump”

Shell Command:

```
awk -F',' '$5 ~ /Donald Trump/ {sum += $13} END {print "Donald Trump Love Count: " sum}' FB_Dataset.csv && awk -F',' '$5 ~ /Donald Trump/ {sum += $18} END {print "Donald Trump Angry Count: " sum}' FB_Dataset.csv
```

Result:

Donald Trump Love Count: 436554
Donald Trump Angry Count: 402826



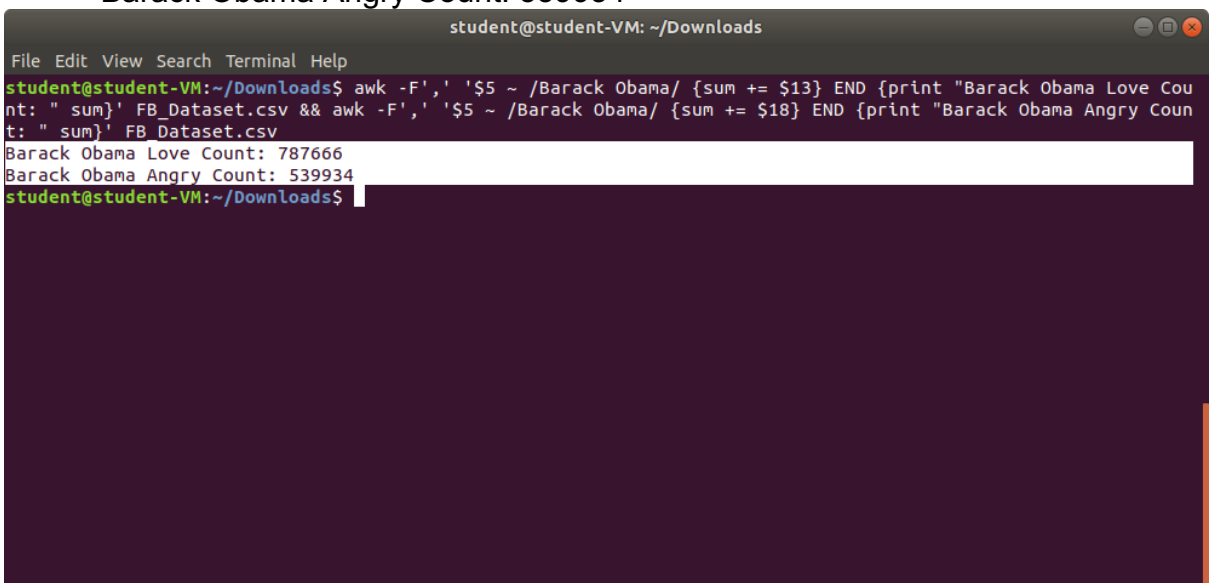
Find total number of love_count, and angry_count for "Barack Obama"

Shell Command:

awk -F',' '$5 ~ /Barack Obama/ {sum += $13} END {print "Barack Obama Love Count: " sum}' FB_Dataset.csv && awk -F',' '$5 ~ /Barack Obama/ {sum += $18} END {print "Barack Obama Angry Count: " sum}' FB_Dataset.csv

Result:

Barack Obama Love Count: 787666
Barack Obama Angry Count: 539934



Justification:

The above result shows that Barack Obama has 787666 love count whereas, Donald Trump has 436554 love count, which shows that Barack Obama has more positive feeling among the Facebook users.

Task B: Graphing the Data in R

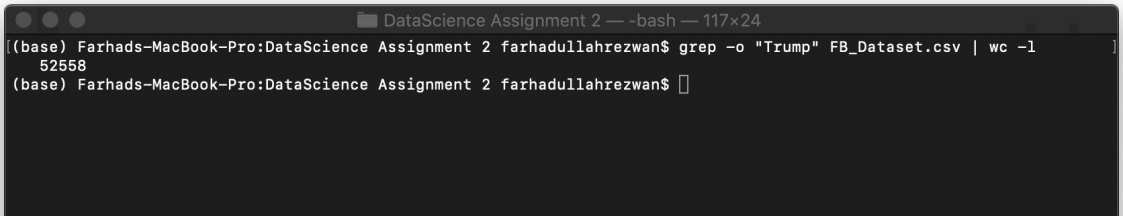
1. Number of times 'Trump' appear in the post content:

Shell command:

```
grep -o "Trump" FB_Dataset.csv | wc -l
```

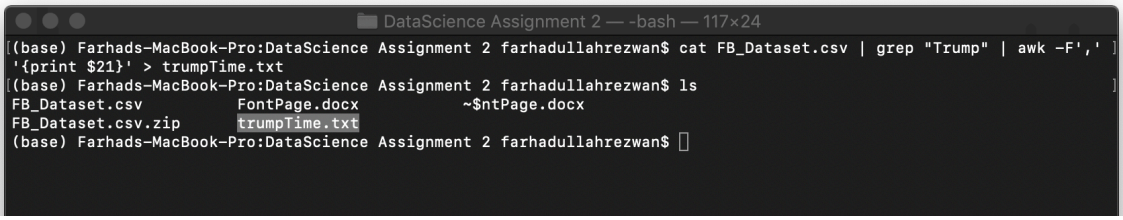
Result:

52558

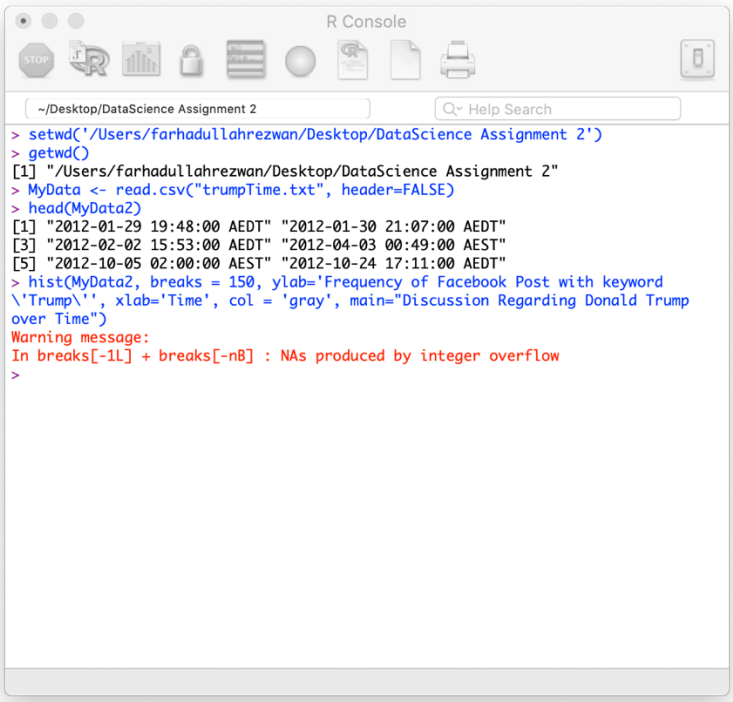


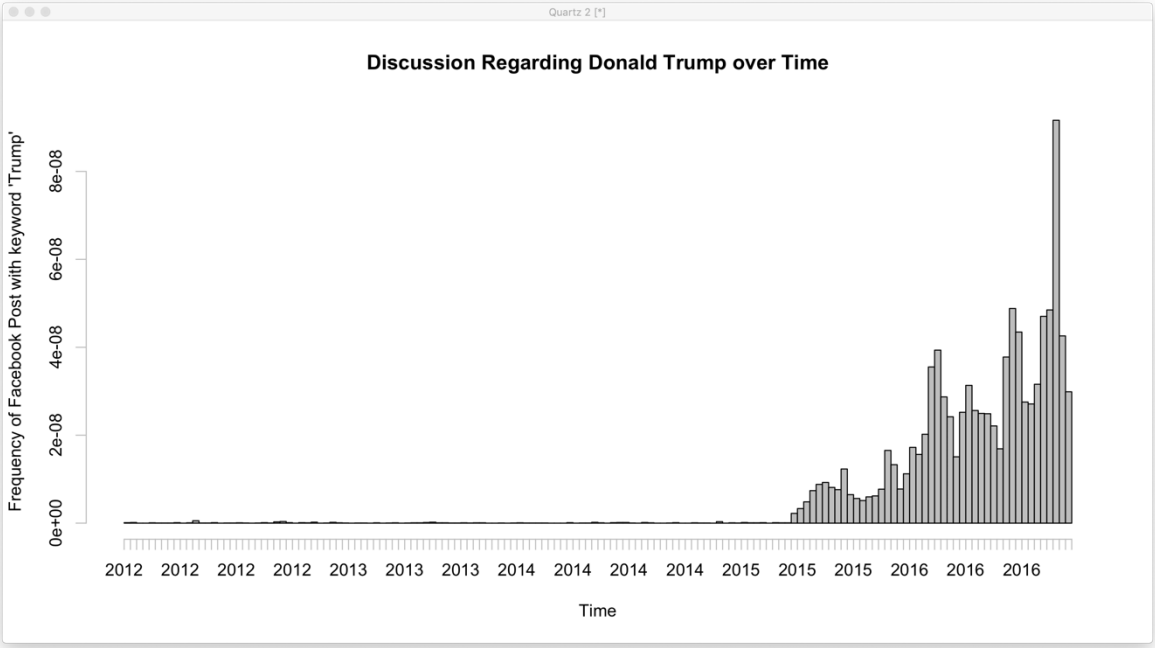
2. Converting the timestamps

Preparing text file in shell:



2.1 hist() function to plot the data in R





2.1. Pattern Description:

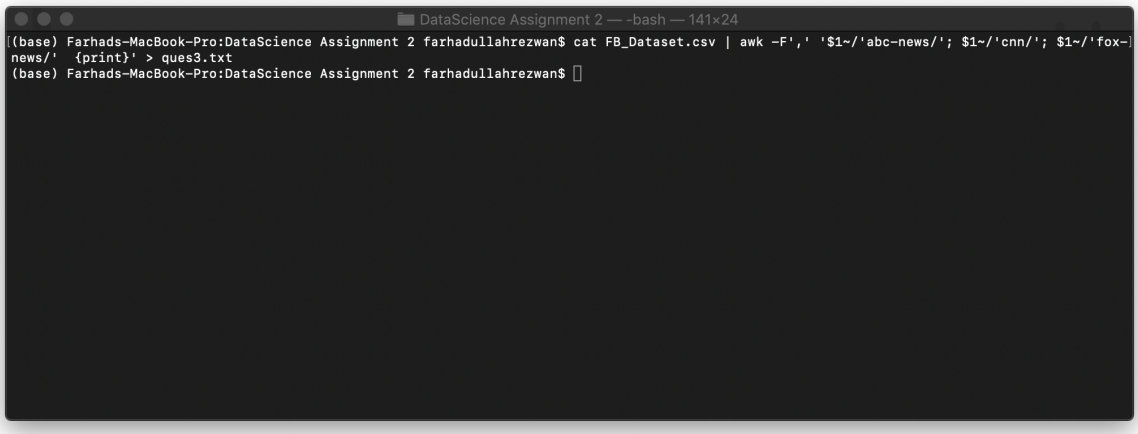
The graph has unusual shape, because the data starts from 2012, where Donald Trump was not that popular in Facebook. The pattern also shows that during the year in 2016 Donald Trump was more popular in Facebook, Due to the precedency election of December 19, 2016 in the USA.

3. Investigating Facebook posts of top media pages.

3.1 Generating file that contains the posts of top media sources:

Shell Command:

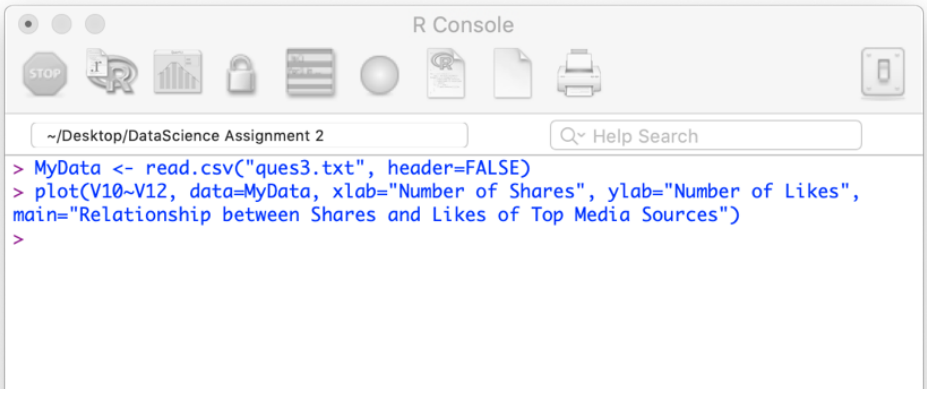
```
cat FB_Dataset.csv | awk -F',' '$1~/'abc-news/';
$1~/'cnn/'; $1~/'fox-news/'  {print}' > ques3.txt
```



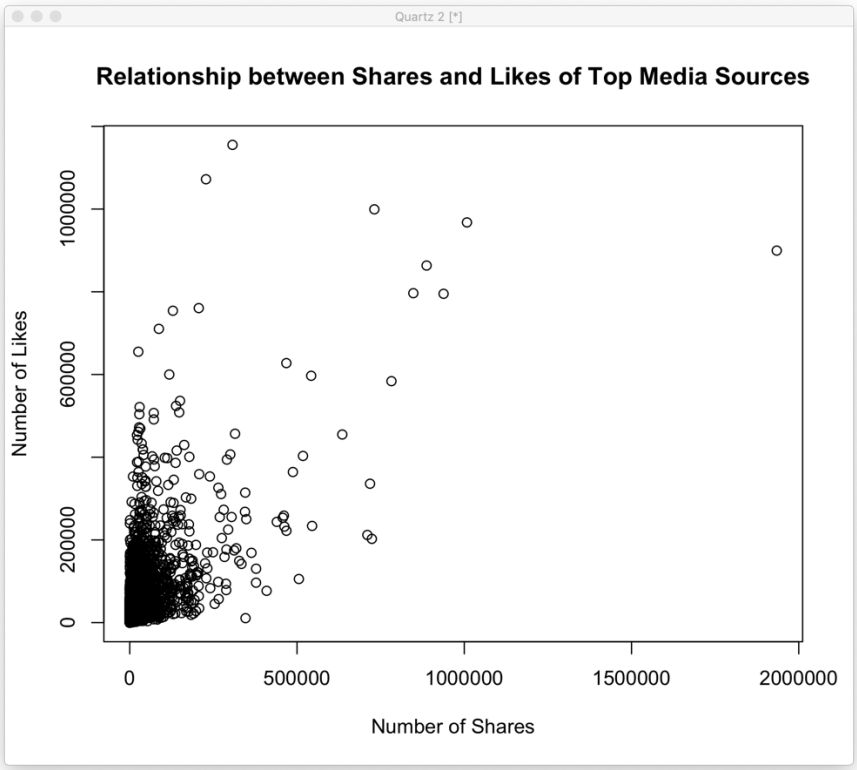
3.2 Plotting the relationship between number of times a post is shared on Facebook and the likes of it.

R Code:

```
> MyData <- read.csv("ques3.txt", header=FALSE)
> plot(V10~V12, data=MyData, xlab="Number of Shares", ylab="Number of Likes",
main="Relationship between Shares and Likes of Top Media Sources")
```



Result:

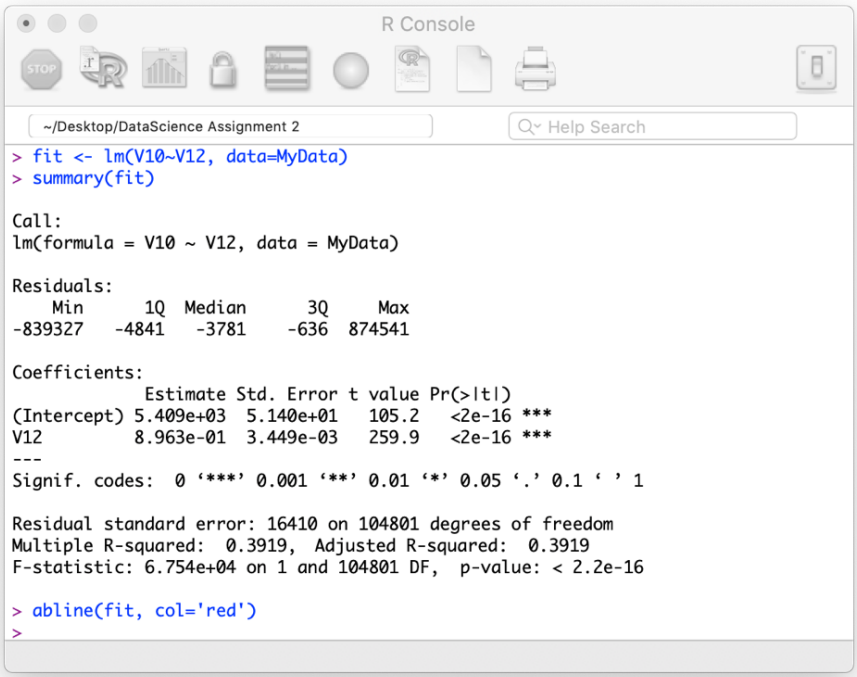


The above graph depicts the relationship between number of likes and number of share of Facebook posts of top media sources like ABC-News, CNN and Fox-News. The above graph shows that when the number of share increases of a particular posts the number of likes also increases.

3.3 Fitting Linear Regression Model:

R code and result:

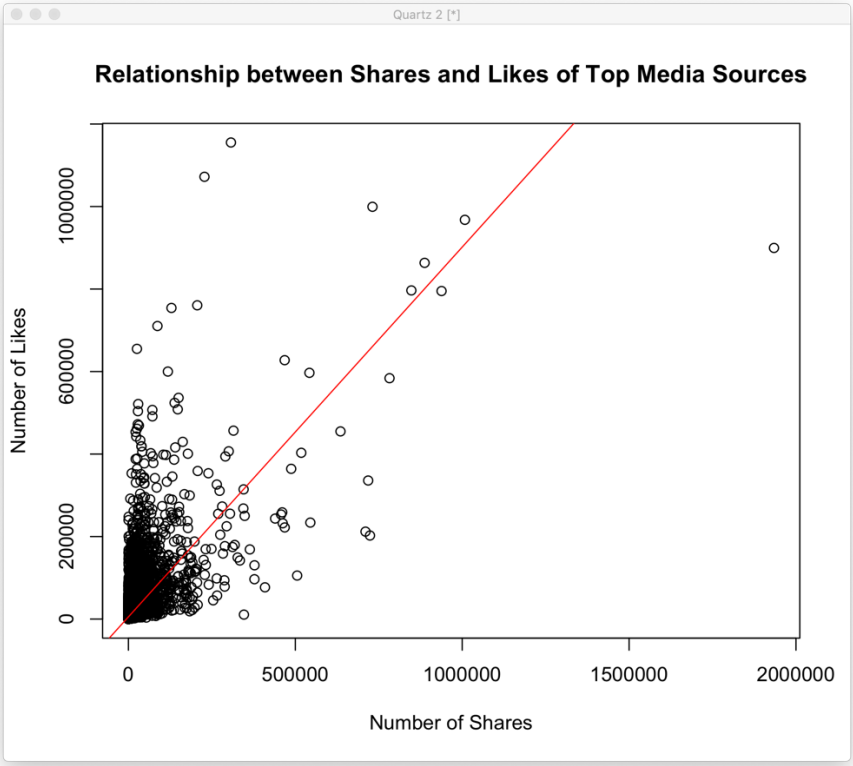
```
> fit <- lm(V10~V12, data=MyData)
> summary(fit)

Call:
lm(formula = V10 ~ V12, data = MyData)

Residuals:
   Min      1Q  Median      3Q     Max 
-839327   -4841   -3781    -636  874541 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.409e+03  5.140e+01   105.2   <2e-16 ***
V12         8.963e-01  3.449e-03   259.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16410 on 104801 degrees of freedom
Multiple R-squared:  0.3919,	Adjusted R-squared:  0.3919 
F-statistic: 6.754e+04 on 1 and 104801 DF,  p-value: < 2.2e-16
```



Result:

R code:

```
> abline(fit, col='red')
```

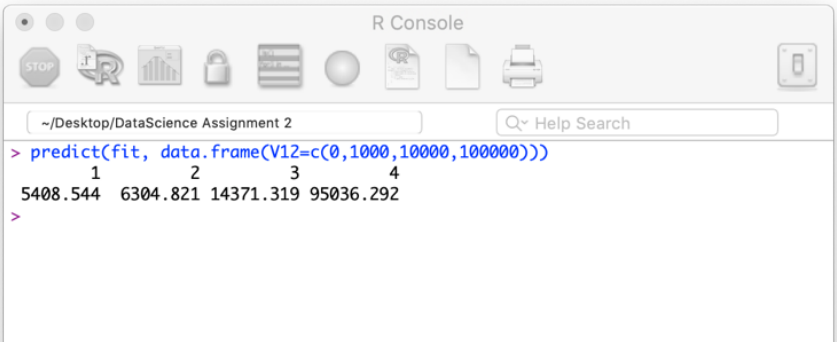


The linear regression of the relationship between Number of shares and number of likes data shows there is a positive relationship, which means when number of share increases the number of likes also increases. However, this regression model is not perfect predector of this data as the error rate is high.

4. Liner fit to predict:

R Code:

```
> predict(fit, data.frame(V12=c(0,1000,10000,100000)))
```



So the prediction using the Linear Fit shows that, number of likes(rounded) will be 5409, 6305, 14371 and 95036 when the posts are shared 0 times, 1000 times, 10000 times, and 100000 times respectively.