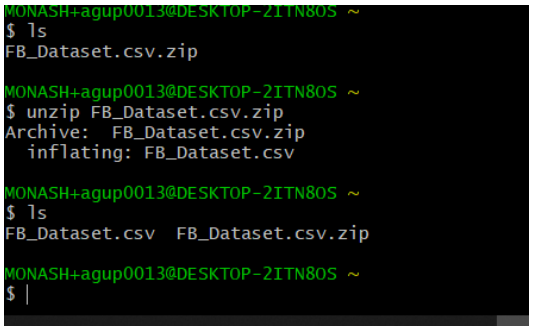
ASSIGNMENT-3

On Cygwin and R

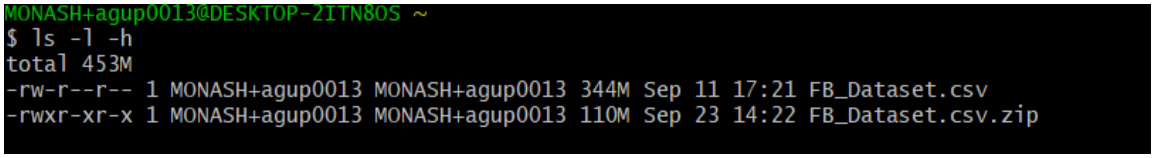
AKANSHA GUPTA
29460069

1. **Decompress the file. How big is it?**

Code: cp Downloads/FB_Dataset.csv.zip .
Code: unzip FB_Dataset.csv.zip



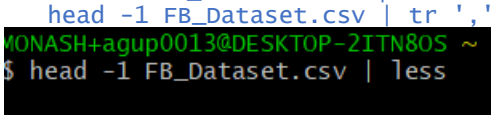
Code: ls -l -h



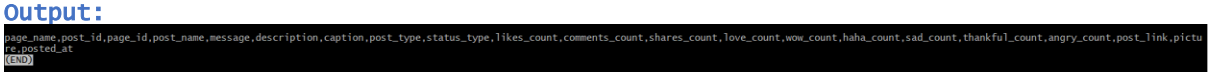
From the above output we could see that FB_Dataset.csv is 344 MB.

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

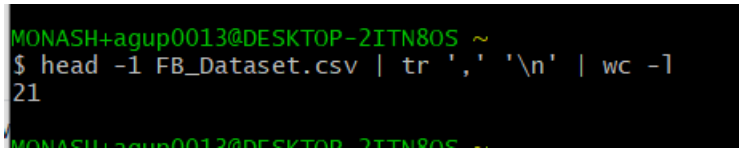
Code: head -1 FB_Dataset.csv | less
head -1 FB_Dataset.csv | tr ',' '\n' | wc -l



Output:



, is the delimeter.



The number of columns in Dataset as per the first row 21.

But from observing the dataset we could see different row have different set of columns. So, the below also helps us to identify that

Code: awk -F ',' '{print NF}' FB_Dataset.csv | sort | uniq -c

21 columns for 533907 rows
22 columns for 2 rows
23 columns for 4 rows

14 columns for 41 rows

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

Code: awk NR==1 FB_Dataset.csv

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ awk NR==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,sa
d_count,thankful_count,angry_count,post_link,picture,posted_at
```

The above is the list of columns in the dataset. Which is the required data for analysing the FB dataset.

4. **How many Facebook posts are there in the file?**

Code: cat FB_Dataset.csv | cut -d',' -f2 | tail -n +2 |wc -l

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ cat FB_Dataset.csv | cut -d',' -f2 | tail -n +2 | wc -l
533926
```

Although we can observe a single null value in the dataset. If we remove it too the count will reduce by 1.

5. **What is the date range for Facebook posts in this file? (Assume that the data is in order)**

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

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ awk -F ',' '{print $21}' FB_Dataset.csv | head -2
posted_at
1/1/12 0:30

MONASH+agup0013@DESKTOP-2ITN8OS ~
$ awk -F ',' '{print $21}' FB_Dataset.csv | tail -1
7/11/16 23:45
```

According to the given dataset and the above query, we could identify that the first post was posted on 1/1/12 at 00:30, while the last post was on 7/11/16 23:45

6. **How many unique pages are there?**

Code:cut -d',' -f3 FB_Dataset.csv | tail -n +2 | uniq| wc -l

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ cut -d',' -f3 FB_Dataset.csv | tail -n +2 | uniq| wc -l
15
```

There are 15 unique pages in the dataset. (The above result is after eliminating the header by using command tail -n +2)

7. **How many unique posts are there?**

Code: awk -F ',' '{print $2}' FB_Dataset.csv | uniq -c| tail -n +2 |wc -l

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ awk -F ',' '{print $2}' FB_Dataset.csv | uniq -c| tail -n +2 |wc -l
533925
```

We could see the 533925 unique facebook post in the whole dataset.

8. When was the first mention in the file regarding "Italian Dishes" and what was the post?

Code:`grep 'Italian Dishes' FB_Dataset.csv | cut -d ',' -f2,4,21`

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ grep 'Italian Dishes' FB_Dataset.csv |cut -d ',' -f2,4,21
18468761129_10153133124136130,5 Brilliant Italian Dishes You Haven't Tried Before,11/6/15 14:01
```

We could see that the Italian Dish was mentioned in the single comment and the post was "Brilliant Dishes You Haven't Tried Before" at 11/06/15 14:01

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

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

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ grep -o 'Barack Obama' FB_Dataset.csv|wc -l
6831
```

I found it by running the grep command is used to search a string of character in specified document and count it by using wc -l command after the pipeline. I have received the output of 6831. (Note: I have ignored the case by not using -i in the syntax)

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

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

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ grep -o 'Donald Trump' FB_Dataset.csv|wc -l
15024
```

As per the above output we could see that, Trump is the most popular post in the given dataset as he had been mentioned 15024 times which is almost more than double than the times Obama is been mentioned in the posts over FB i.e. 6831.

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.**

Code: `cut -f2,5,10 -d',' FB_Dataset.csv | grep -i 'Trump' | awk -F ',' '$3>100' |cut -d ',' -f1,3 | sort -t ',' -n -r -k2> trump.txt`

The below code will add the post_id and like_count header in the Trump.txt document.

Code: `sed -i '1ipost_id,like_count' trump.txt`

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ cut -f2,5,10 -d',' FB_Dataset.csv | grep -i 'Trump' | awk -F ',' '$3>100' |cut -d ',' -f1,3 | sort -t ',' -n -r -k2> trump.txt

MONASH+agup0013@DESKTOP-2ITN8OS ~
$ sed -i '1ipost_id,like_count' trump.txt
```

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ head -5 trump.txt
post_id,like_count
_222287356672l6_1015396016795221722,368179
5550296508_10154298504746509,248012
18468761129_10153524839811130,229187
15704546335_10154108339971336,222119
```

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.**

Code: `grep -i "Donald Trump" FB_Dataset.csv|cut -d ',' -f5,13,18| awk -F ',' '{donald_love += $2 ; donald_hate += $3} END { print donald_love, donald_hate}'`

Code: `grep -i "Barack Obama" FB_Dataset.csv|cut -d ',' -f5,13,18| awk -F ',' '{obama_love += $2 ; obama_hate += $3} END { print obama_love, obama_hate}'`

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ grep -i "Barack Obama" FB_Dataset.csv|cut -d ',' -f5,13,18| awk -F ',' '{obama_love += $2 ; obama_hate += $3} END { print obama_love, obama_hate}'
835889 581989

MONASH+agup0013@DESKTOP-2ITN8OS ~
$ grep -i "Donald Trump" FB_Dataset.csv|cut -d ',' -f5,13,18| awk -F ',' '{donald_love += $2 ; donald_hate += $3} END { print donald_love, donald_hate}'
1563384 2189425
```

From the above output, we could analyse that the love_count and the angry_count for Donald Trump is the highest as compared to Barak Obama. The range of difference between the love count is 727,495 while the difference of angry count is 1,607,436 which is higher than the love count range.
We can also analyse that Trump received more anger than Obama as per the data analysis, but there is also a huge range of difference between the love he received.
If we take a difference between love and angry count of Trump we could see that difference comes in negative but we can observe a reverse scenario for Obama, this clearly states that citizen love more to Obama than Trump.

Task B

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

Code: `cut -f5 -d',' FB_Dataset.csv | grep -i 'Trump'|wc -l`

Count with ignore case

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ cut -f5 -d',' FB_Dataset.csv | grep -i 'Trump'|wc -l
20507
```

Count with no ignore case

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ cut -f5 -d',' FB_Dataset.csv | grep -o 'Trump'|wc -l
22338
```

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

To convert the timestamp into a file

Code: `cut -d ',' -f5,21 FB_Dataset.csv | grep -i "Trump" | awk -F ',' '{print $2}' > Timestamp.txt`

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ cut -d ',' -f5,21 FB_Dataset.csv | grep -i "Trump" | awk -F ',' '{print $2}' > Timestamp.txt

MONASH+agup0013@DESKTOP-2ITN8OS ~
```

Code:
```
install.packages("readtext")
data_csv <- read.csv("Timestamp.txt", header = FALSE, sep = " ")
head(data_csv)
```

V1	V2
29/1/12	19:48
2/2/12	15:53
24/10/12	17:11
11/8/13	16:00
12/7/14	17:00
31/7/14	8:08

Code:
```
install.packages("lubridate")
formatted_datetime <- strptime(donald_timestamp, "%d/%m/%y %H:%M")
head(formatted_datetime)
```

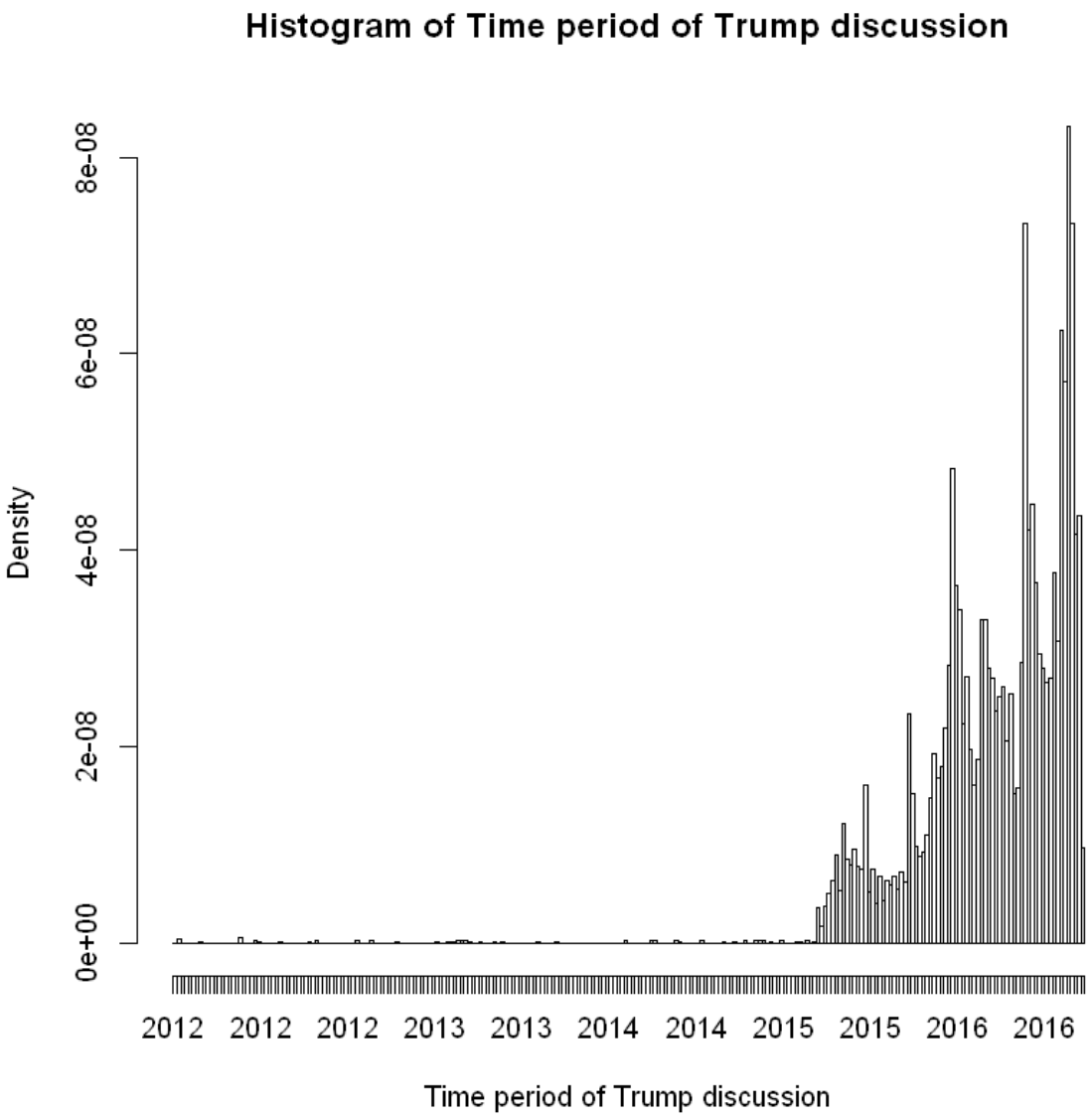
Output:

```
In [5]: install.packages("lubridate")
        formatted_datetime <- strptime(donald_timestamp, "%d/%m/%y %H:%M")
        head(formatted_datetime)

        package 'lubridate' successfully unpacked and MD5 sums checked

        The downloaded binary packages are in
                C:\Users\agup0013\AppData\Local\Temp\RtmpWYjmIs\downloaded_packages

        [1] "2012-01-29 19:48:00 AEDT" "2012-02-02 15:53:00 AEDT"
        [3] "2012-10-24 17:11:00 AEDT" "2013-08-11 16:00:00 AEST"
        [5] "2014-07-12 17:00:00 AEST" "2014-07-31 08:08:00 AEST"
```

Code: For plotting a histogram

```
hist(formatted_datetime, xlab = 'Time period of Trump discussion', frequency = TRUE, breaks = "weeks")
```

Output:



This histogram shows us the major time period where Trump was the most popular topic of discussion . As per the studies, during the period of 2016 trump was most popular as the won the most state primaries, caucuses and delegates. As well as he created 4000,000 more manufacturing jobs than before.

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

The above histogram refers a skewed distribution. We could observe that the above histogram is containing a peak which was during 2016. All the frequencies are lying on the right-hand side of the histogram during the time phase of 2015 to 2016. We can easily identify the gradual increase of Trump being a hot topic over the social media as per his accomplishments despite of all the chaos.

3.1 **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.**

Code:
```
awk -F ',' '($1=="abc-news" || "cnn" || "fox-news")' FB_Dataset.csv > FB_Report.txt
```

```
MONASH+agup0013@DESKTOP-2ITN8OS ~
$ awk -F ',' '($1=="abc-news" || "cnn" || "fox-news")' FB_Dataset.csv > FB_Report.txt

MONASH+agup0013@DESKTOP-2ITN8OS ~
```

The above shell command gives us all the post mentioned in “abc-news” **or** “cnn” **or** “fox-news” in the FB_dataset.csv and store it in a different file named FB_Report.txt

3.2 **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?**

Code:

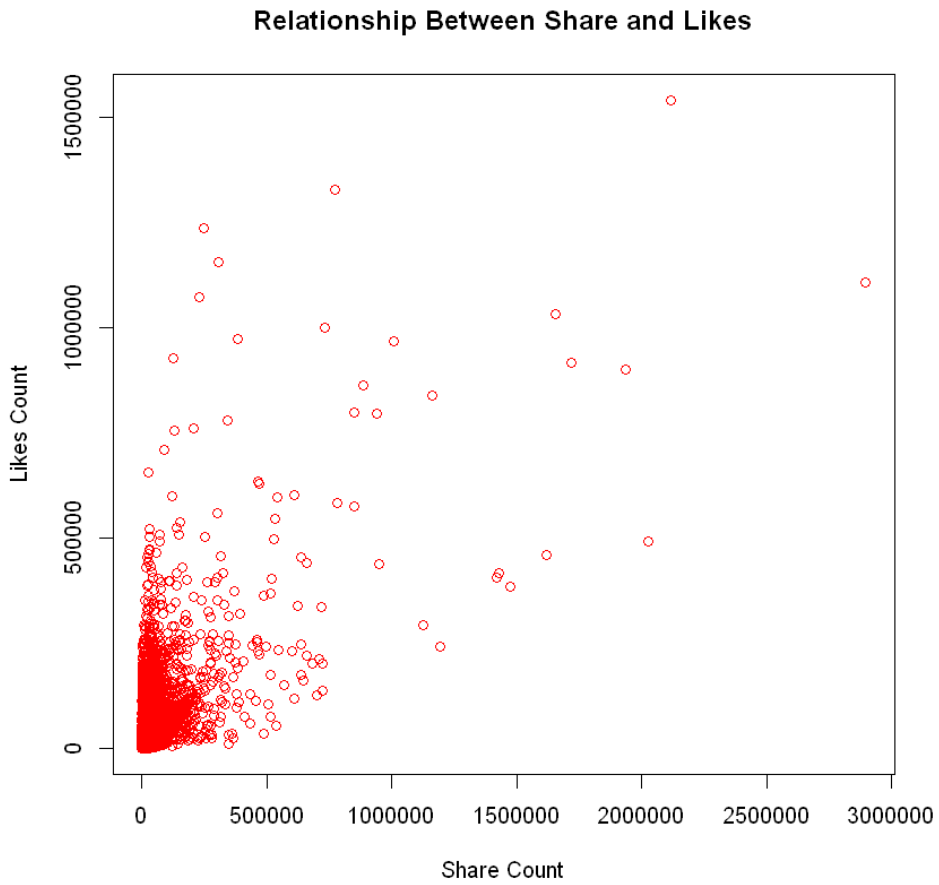
```
data_fb <- read.csv("FB_Report.txt", header = TRUE, sep = ",")
head(data_fb)
data_fb$shares_count <- factor(data_fb$shares_count)
data_fb$likes_count <- factor(data_fb$likes_count)
```

In [9]:
```
data_fb <- read.csv("FB_Report.txt", header = TRUE, sep = ",")
head(data_fb)
```

page_name	post_id	page_id	post_name	message	
abc-news	86680728811_272953252761568	86680728811	Chief Justice Roberts Responds to Judicial Ethics Critics	Roberts took the unusual step of devoting the majority of his annual report to the issue of judicial ethics.	PAUL
abc-news	86680728811_273859942672742	86680728811	With Reservations .. Obama Signs Act to Allow Detention of Citizens	Do you agree with the new law?	
abc-news	86680728811_10150499874478812	86680728811	Wishes For 2012 to Fall on Times Square	Some pretty cool confetti will rain down on New York City celebrators.	The v
abc-news	86680728811_244555465618151	86680728811	Mitt Romney Vows to Veto Dream Act if President	NULL	
abc-news	86680728811_252342804833247	86680728811	NY Pharmacy Shootout Leaves Suspect .. ATF Agent Dead	The pharmacy was held up by a man seeking prescription medication.	A shc
abc-news	86680728811_200661383359612	86680728811	The World Rings in 2012	NULL	Arour

In [10]:
```
data_fb$shares_count <- factor(data_fb$shares_count)
data_fb$likes_count <- factor(data_fb$likes_count)
```

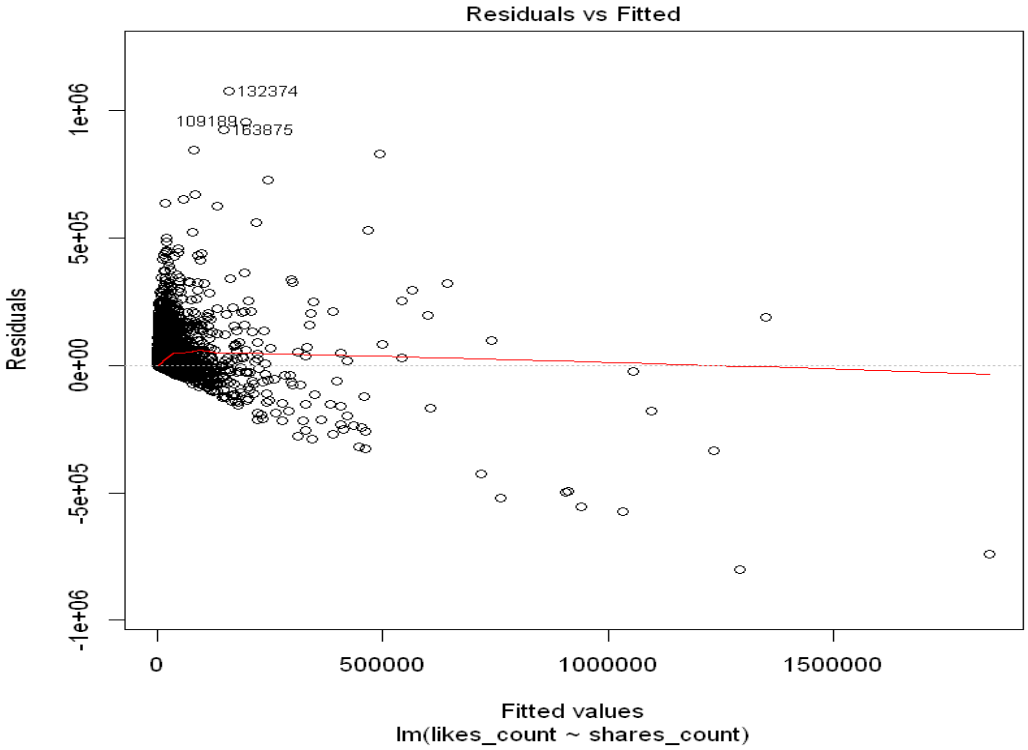
```
plot(data_fb$shares_count,data_fb$likes_count, col="red",main="Relationship Between Share and Likes",
     xlab="Share Count",
     ylab="Likes Count")
```



From this we can identify that maximum number of likes and shares lies in the range between 500000. We can also observe that as the number of likes increases in a given span of time, the number of share count also increases. Hence, likes_count and share_count is linear to each other.

3.3 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?

Code: linearMod <- lm(likes_count ~ shares_count, data=data_fb)
plot(linearMod)

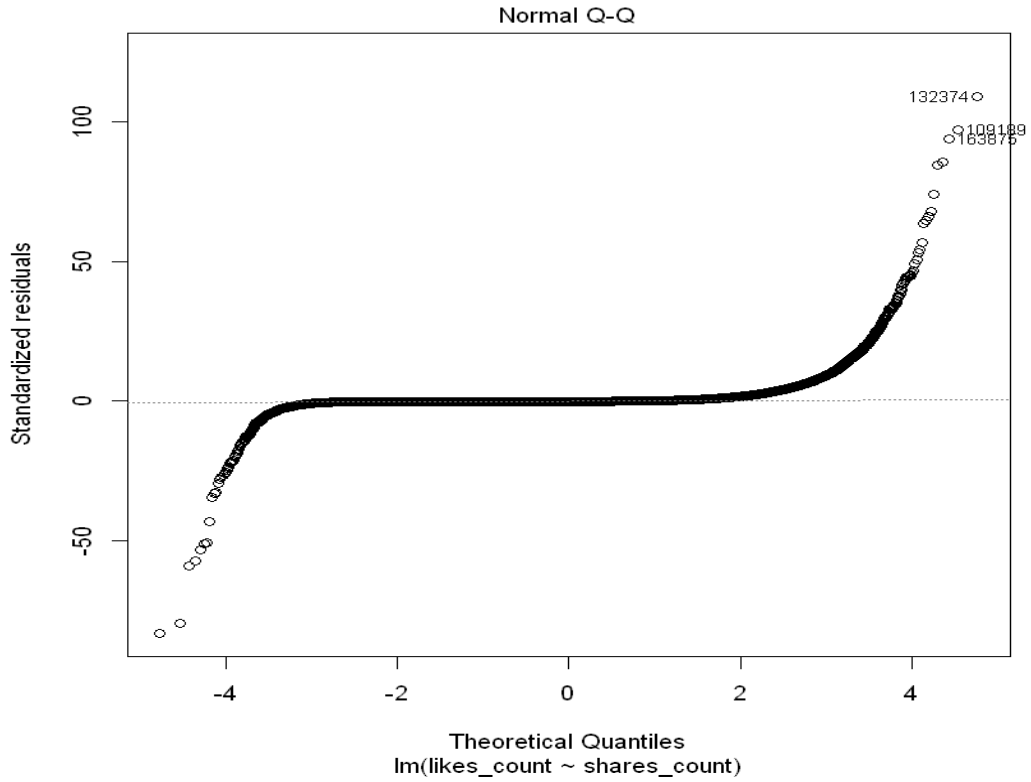


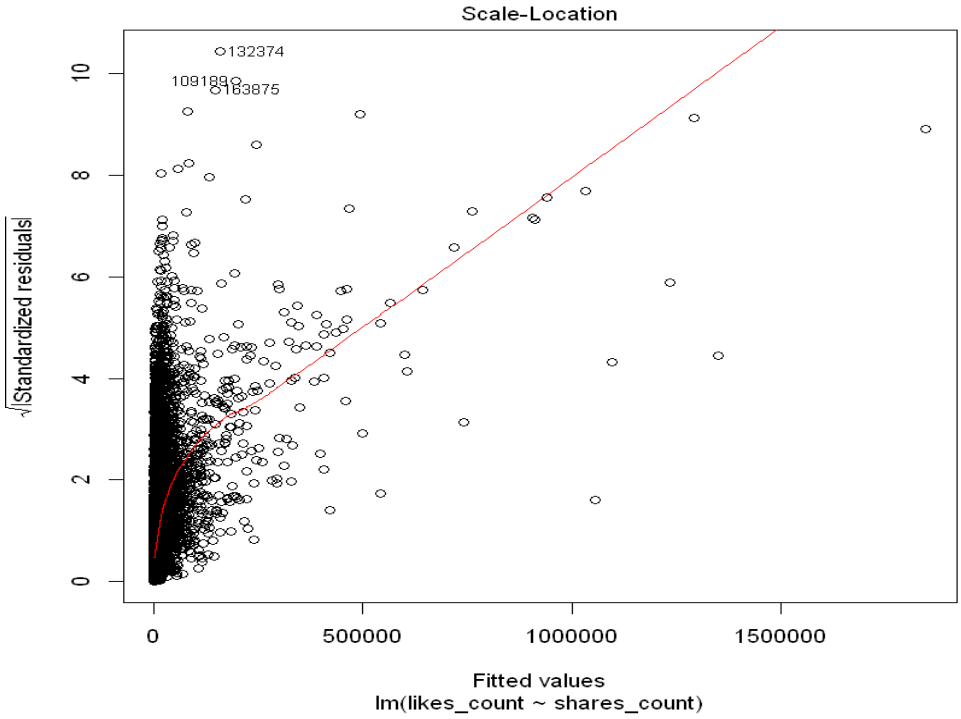
The above graph is the scatter plot of Residual on the y-axis and Fitted values over the x-axis. This plot basically detects outliers, error variance, non-linearity.

We can observe residual has a non-linear pattern which basically indicates that the model is not "well-fitted" and has a non-linear relationship between predictor and dependent.
While the plot suggests the horizontal line in the graph have no residual pattern around it, which further proves that the model is not linear as well as the variance of error terms equal

Standardized Residual and Normal Q-Q

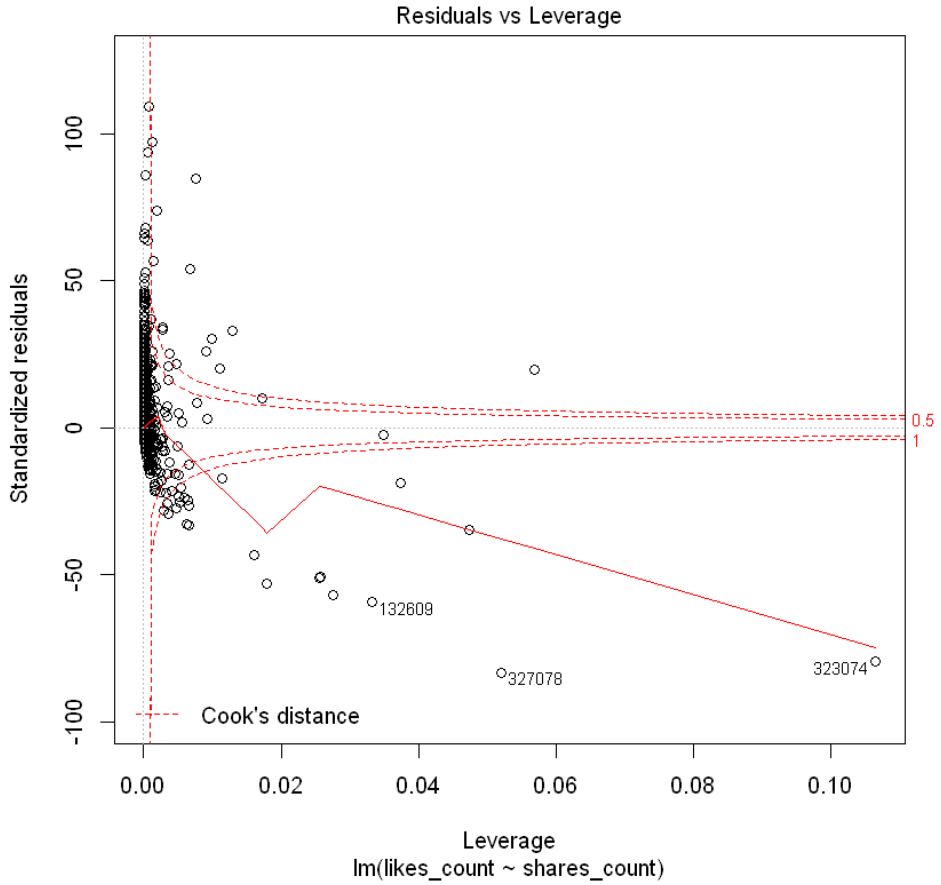
Here we could see that the point falls along the centre of the line while takes the curve at the extremities. To see whether the data matches the assumptions of normalities which is analysed by comparing the quartiles of our data against quartiles of normal distribution.
But, our dataset doesn't resembles the normality on the points like 163875, 109189 or even at 132374.





Standardized Residual and Scale-Location

From the above graph we could see that, our line slopes around at 1 and slopes till 3 but after that we could identify an exponential increase in the line. But the range within 1 till 3.5 is up because the residual of those predictor value is more spread out. Hence, we can observe that the graph is somewhat heteroscedastic.



Standardized Residual and Residual vs Leverage

The plots Residuals vs. Leverage help you define your model's important influential points. Outliers can be important, but not necessarily, although at some points in your model could be very influential within a normal range.
The points we are looking for are values in the top right or bottom right corners outside the distance line of the red dashed Cook. These are points in the model that would be influential and removing them would probably significantly change the outcomes of regression.
So from the above we could observe that the influential point will be 323074 in this scenario.

print(linearMod)

```
Call:
lm(formula = likes_count ~ shares_count, data = data_fb)

Coefficients:
 (Intercept)  shares_count
   2526.3212        0.6367
```

summary(linearMod)

```
Call:
lm(formula = likes_count ~ shares_count, data = data_fb)

Residuals:
    Min      1Q  Median      3Q     Max
-800116   -2305   -1907    -570 1075484

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.526e+03  1.356e+01   186.4   <2e-16 ***
shares_count 6.367e-01  1.112e-03   572.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9865 on 533938 degrees of freedom
Multiple R-squared:  0.3804,    Adjusted R-squared:  0.3804
F-statistic: 3.278e+05 on 1 and 533938 DF,  p-value: < 2.2e-16
```

3.4 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

Code:

For 0 times

```
set.seed(100)
 predict(linearMod, data.frame(shares_count=0))
```

For 1000 times

```
set.seed(100)
predict(linearMod, data.frame(shares_count=1000))
```

For 10000 times

```
set.seed(100)
predict(linearMod, data.frame(shares_count=10000))
```

For 100000 times

```
set.seed(100)
predict(linearMod, data.frame(shares_count=100000))
```

Task B.3.4

In [20]:
```
set.seed(100)
predict(linearMod, data.frame(shares_count=0))
```

1: 2526.32123698894

In [21]:
```
set.seed(100)
predict(linearMod, data.frame(shares_count=1000))
```

1: 3162.98360832903

In [22]:
```
set.seed(100)
predict(linearMod, data.frame(shares_count=10000))
```

1: 8892.94495038992

In [23]:
```
set.seed(100)
predict(linearMod, data.frame(shares_count=100000))
```

1: 66192.5583709988