FIT1043 Assignment 3

Yo Kogure
32134541

Introduction:

This is my attempt at Assignment 3 in 1043, by Yo Kogure, 32134541. I am using Mac OS, as well as UNIX shell(Terminal) and R application to analyze the data.
Please note that those highlighted in yellow lines are my codes that I have actually used. Comments are indicated with *//* in this pdf file and those highlighted in green are my outputs from the shell. Some texts are emphasized in **bold**.

The data I am dealing with, are a compressed file that contains Facebook posts from 15 of the top mainstream media sources (e.g., BBC, CNN, Fox News, etc.) from 2012 to 2016.

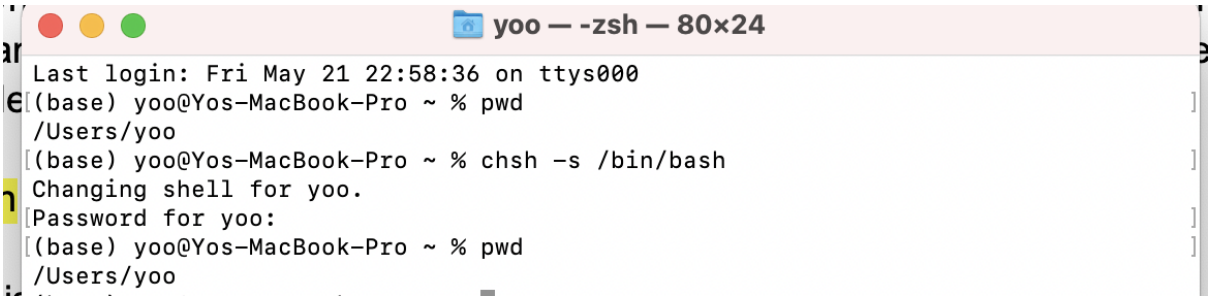
Here are the guidelines for this report. For citations, there are in-text citations as well as APA references at the end of this paper. I have also referred to Tutorial Questions, but the reference to them is not included every time.

1. **Introduction**
2. **Part A**
3. **Part B**
4. **Conclusion**
5. **References(APA format, 6th edition)**

Part A:

First, I have downloaded *FB_Dataset.gz.* My first task is to open the terminal and type in the following commands to change the default shell to BASH, as my Mac is set to zsh.

```
chsh -s /bin/bash
```



Now that BASH is available, I will create the new folder inside /Users/yoo for working with things in this assignment, and I use the following commands,

```
mkdir Assignment3  // creates file

cd Assignment3  // change directly to Assignment3
mv /Users/yoo/Downloads/FB_Dataset.gz .
```

Cd changes directly to the new folder Assignment3.
This mv command will move the data file to the current working directory, indicated with ".".

Q1)
pwd(print working directory) is used to check where I am right now.
Ls is used to list files that are in the folder.

```
(base) yoo@Yos-MacBook-Pro Assignment3 % pwd
/Users/yoo/Assignment3
(base) yoo@Yos-MacBook-Pro Assignment3 % mv /Users/yoo/Downloads/FB_Dataset.gz .

(base) yoo@Yos-MacBook-Pro Assignment3 % ls
FB_Dataset.gz
(base) yoo@Yos-MacBook-Pro Assignment3 %
```

pwd
mv /Users/yoo/Downloads/FB_Dataset.gz .
ls

Now that environment is set, I will use the following commands to see the file size.

ls -lh FB_Dataset.gz

-lh option of ls lets us see the file size.

```
(base) yoo@Yos-MacBook-Pro Assignment3 % ls -lh FB_Dataset.gz
-rw-r--r--@ 1 yoo  staff   110M May 21 23:09 FB_Dataset.gz
(base) yoo@Yos-MacBook-Pro Assignment3 %
```

110M is the original file size.

```
(base) yoo@Yos-MacBook-Pro Assignment3 % ls -lh FB_Dataset.gz
-rw-r--r--@ 1 yoo  staff   110M May 21 23:09 FB_Dataset.gz
(base) yoo@Yos-MacBook-Pro Assignment3 % gunzip FB_Dataset.gz
(base) yoo@Yos-MacBook-Pro Assignment3 % ls -lh FB_Dataset.txt
ls: FB_Dataset.txt: No such file or directory
(base) yoo@Yos-MacBook-Pro Assignment3 % ls -lh FB_Dataset
-rw-r--r--  1 yoo  staff   343M May 21 23:09 FB_Dataset
(base) yoo@Yos-MacBook-Pro Assignment3 %
```

(*I was misthinking that it was .txt format, while it is not)*

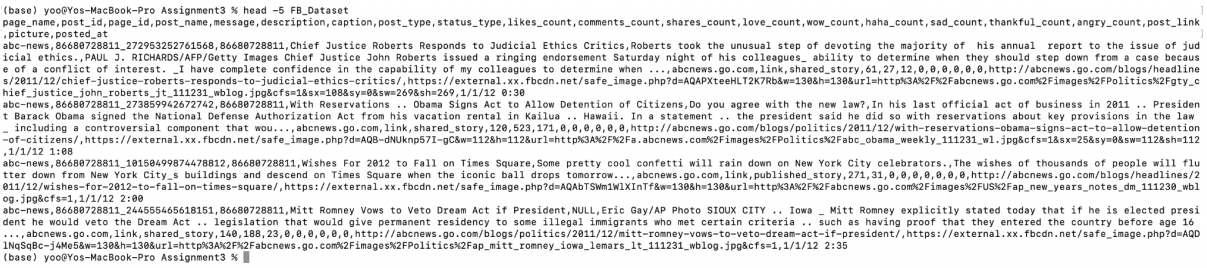
After
gunzip FB_Dataset.gz // extracts the file
and,
ls -lh FB_Dataset

I now see the file size after extracting **343M**(Megabytes).
The compression rate as percentage is 343M / 110M * 100 = 312 % (3s.f.)

Q2)
After typing

head -5 FB_Dataset

Which obtains for me first few lines,

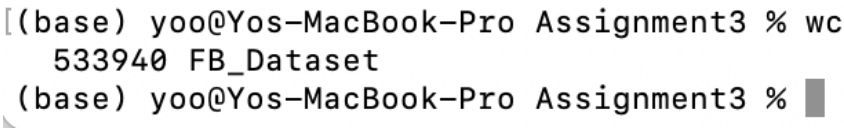


```
(base) yoo@Yos-MacBook-Pro Assignment3 % head -5 FB_Dataset
page_name,post_id,page_id,post_name,message,description,caption,post_type,status_type,likes_count,comments_count,shares_count,love_count,wow_count,haha_count,sad_count,thankful_count,angry_count,post_link
,picture,posted_at
abc-news,86680728811_2729532527615 68,86680728811,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 jud
icial ethics.,PAUL J. RICHARDS/AFP/Getty Images Chief Justice John Roberts issued a ringing endorsement Saturday night of his colleagues_ ability to determine when they should step down from a case becaus
e of a conflict of interest. _I have complete confidence in the capability of my colleagues to determine when ...,abcnews.go.com,link,shared_story,61,27,12,0,0,0,0,0,0,http://abcnews.go.com/blogs/headline
s/2011/12/chief-justice-roberts-responds-to-judicial-ethics-critics/,https://external.xx.fbcdn.net/safe_image.php?d=AQAPXteeHLT2K7Rb&w=130&h=130&url=http%3A%2F%2Fabcnews.go.com%2Fimages%2FPolitics%2Fgty_c
hief_justice_john_roberts_jt_111231_wblog.jpg&cfs=1&sx=108&sy=0&sw=269&sh=269,1/1/12 0:30
abc-news,86680728811_273859942672742,86680728811,With Reservations .. Obama Signs Act to Allow Detention of Citizens,Do you agree with the new law?,In his last official act of business in 2011 .. Presiden
t Barack Obama signed the National Defense Authorization Act from his vacation rental in Kailua .. Hawaii. In a statement .. the president said he did so with reservations about key provisions in the law
_ including a controversial component that wou...,abcnews.go.com,link,shared_story,120,523,171,0,0,0,0,0,0,http://abcnews.go.com/blogs/politics/2011/12/with-reservations-obama-signs-act-to-allow-detention
-of-citizens/,https://external.xx.fbcdn.net/safe_image.php?d=AQB-dNUknp57I-gC&w=112&h=112&url=http%3A%2F%2Fa.abcnews.com%2Fimages%2FPolitics%2Fabc_obama_weekly_111231_wl.jpg&cfs=1&sx=25&sy=0&sw=112&sh=112
,1/1/12 1:08
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 wishes of thousands of people will flu
tter down from New York City_s buildings and descend on Times Square when the iconic ball drops tomorrow...,abcnews.go.com,link,published_story,271,31,0,0,0,0,0,0,0,http://abcnews.go.com/blogs/headlines/2
011/12/wishes-for-2012-to-fall-on-times-square/,https://external.xx.fbcdn.net/safe_image.php?d=AQAbTSWm1WlXInTf&w=130&h=130&url=http%3A%2F%2Fabcnews.go.com%2Fimages%2FUS%2Fap_new_years_notes_dm_111230_wbl
og.jpg&cfs=1,1/1/12 2:00
abc-news,86680728811_2445554656181 51,86680728811,Mitt Romney Vows to Veto Dream Act if President,NULL,Eric Gay/AP Photo SIOUX CITY .. Iowa _ Mitt Romney explicitly stated today that if he is elected presi
dent he would veto the Dream Act .. legislation that would give permanent residency to some illegal immigrants who met certain criteria .. such as having proof that they entered the country before age 16
...,abcnews.go.com,link,shared_story,140,188,23,0,0,0,0,0,0,http://abcnews.go.com/blogs/politics/2011/12/mitt-romney-vows-to-veto-dream-act-if-president/,https://external.xx.fbcdn.net/safe_image.php?d=AQD
lNqSqBc-j4Me5&w=130&h=130&url=http%3A%2F%2Fabcnews.go.com%2Fimages%2FPolitics%2Fap_mitt_romney_iowa_lemars_lt_111231_wblog.jpg&cfs=1,1/1/12 2:35
(base) yoo@Yos-MacBook-Pro Assignment3 %
```

I notice that they are separated by commas. Hence, **the delimiter is comma**.

wc -l FB_Dataset

Gives me access to the number of rows in the file. Using -w instead of -l ends up counting words. Wc stands for word count.



```
(base) yoo@Yos-MacBook-Pro Assignment3 % wc -l FB_Dataset
   533940 FB_Dataset
(base) yoo@Yos-MacBook-Pro Assignment3 %
```

533940 Rows.

Q3)

From looking at the first 5 rows obtained from Q2, I can understand that the first row represents the column types(header).

```
(base) yoo@Yos-MacBook-Pro Assignment3 % head -1 FB_Dataset
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
(base) yoo@Yos-MacBook-Pro Assignment3 %
```

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

These are the texts obtained from the header. Since columns are separated by commas, after I count, there are 21 columns.

Q4)

I interpret a unique page as each row, the actual row containing 21 aspects for that single post. I noticed from the previous question, when I printed the first 5 rows with -5, it printed more than 10 lines in the BASH, but 5 lines if we consider them separated by **new lines**. These rows are what I want to find and each represents a unique post. Q3 helped me justify that each line represents each post after reading and analyzing the 21 columns, such as post_id and post_name.

This repeats Q2 in a way, but to make sure, I perform the row count again with another method, which again prints 533940.

cat FB_Dataset | wc -l // do the row count again in FB_Dataset using wc

// Cat stands for concatenate

uniq FB_Dataset Output // this uniq helps me remove duplicate rows, and output it to newly made Output file

cat Output | wc -l // In the Output file, count the number of rows again

```
(base) yoo@Yos-MacBook-Pro Assignment3 % cat FB_Dataset | wc -l
  533940
(base) yoo@Yos-MacBook-Pro Assignment3 % uniq FB_Dataset Output
(base) yoo@Yos-MacBook-Pro Assignment3 % cat Output | wc -l
  533940
(base) yoo@Yos-MacBook-Pro Assignment3 %
```

Uniq only works for adjacent lines, but I don't have to use sort because I know they are sorted by date of the post in Q5. If there are duplicates I assume they are adjacent already.

Both of these printed 533940, and I can conclude that there are no duplicate data or headers somewhere in the file, and that **533940** is the number of unique posts.

Q5)

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

The last column, 21th column indicates posted_at.
Since the first row is the header and the second column represents the very first data, I use the command to extract 2nd row 21th column.

Note that from Q4 onwards I am using “|”, this is called piping and I am using it to connect the output of a certain command to another command.

PLEASE NOTE:
As my friend Shariq Nadeem has suggested me to use this command for Mac to avoid illegal byte sequences, I have used it here.
export LC_CTYPE=C

```
cut: stdin: Illegal byte sequence
(base) yoo@Yos-MacBook-Pro Assignment3 % export LC_CTYPE=C
(base) yoo@Yos-MacBook-Pro Assignment3 %
```

And my solution part:
cut -d"," -f21 FB_Dataset | head -4

```
Yos-MacBook-Pro:Assignment3 yoo$ cut -d"," -f21  FB_Dataset | head -4
posted_at
1/1/12 0:30
1/1/12 1:08
1/1/12 2:00
Yos-MacBook-Pro:Assignment3 yoo$
```

After many tries, the above code has worked for me.
I used cut and “|” to pipe the filtered column into a head which prints the first 4 rows.

-d”,” sets the delimiter to comma, as the default one for cut is Tab. -f21 represents the 21st column, *posted_at.*

1/1/12 0:30 is the first date, and the last date can be found by piping the above command into tail -3 (it can just be -1, but for double-checking I like to inspect multiple lines). Somehow this resulted in Illegal byte sequence, I used export LC_CTYPE=C as stated.

```
Yos-MacBook-Pro:Assignment3 yoo$ cut -d"," -f21  FB_Dataset | tail -3
cut: FB_Dataset: Illegal byte sequence
14/8/12 14:54
14/8/12 15:57
14/8/12 16:55
Yos-MacBook-Pro:Assignment3 yoo$
```

Above is the failed attempt.

Below is the correct attempt.

cut -d"," -f21 FB_Dataset | tail -3

```
Yos-MacBook-Pro:Assignment3 yoo$ cut -d"," -f21  FB_Dataset | tail -3
7/11/16 23:00
7/11/16 23:30
7/11/16 23:45
Yos-MacBook-Pro:Assignment3 yoo$
```

7/11/16 23:45 is the last date in chronological order and **1/1/12 0:30** is the first date.
Date range is
1st January 2012 0:30 to **7th November 2016 23:45.**

Q6)
grep 'Malaysia Airlines' FB_Dataset | head -3

Is the code I have used. Grep stands for global regular expression print, and it searches for a certain string “Malaysia Airlines” in the dataset and pipes it to head -3, i.e., shows the first 3 rows. Again, it can just be head -1 but to reduce errors and misunderstandings I am using this way.

```
Yos-MacBook-Pro:Assignment3 yoo$ grep 'Malaysia Airlines' FB_Dataset | head -3
abc-news,86680728811_10152267754078812,86680728811,NULL,DEVELOPING: Malaysia Airlines spokesperson: Flight carrying 239 people from Kuala Lumpur to Beijing has gone missing  contact lost: http://abcn.ws/NHHeLT,NULL,NULL,status,mobile_status_update,1583,435,1526,0,0,0,0,0,0,NULL,NULL,8/3/14 0:47
abc-news,86680728811_10152267759088812,86680728811,NULL,UPDATE: Malaysia Airlines spokesperson: Flight carrying 239 people has gone missing en route to China: http://abcn.ws/NHHeLT,NULL,NULL,status,mobile_status_update,1384,377,1393,0,0,0,0,0,0,NULL,NULL,8/3/14 0:50
abc-news,86680728811_10152262883340729,86680728811,Flight Goes Missing En Route to China,UPDATE: Malaysia Airlines says passengers on the missing airliner are from 13 different nationalities http://abcn.ws/NHHeLT,A Malaysian Airlines flight with 227 passengers on board has gone missing a spokeswoman has confirmed to ABC News.,abcnews.go.com,link,shared_story,2410,398,502,0,0,0,0,0,0,http://abcn.ws/NHHeLT,https://external.xx.fbcdn.net/safe_image.php?d=AQDm-ZbZjzs-ReXS&w=130&h=130&url=https%3A%2F%2Fscontent-a-iad.xx.fbcdn.net%2Fhphotos-prn2%2Ft31.0-8%2Fq74%2Fs720x720%2F1599401_10152267849518812_1355361430_o.jpg&cfs=1,8/3/14 2:04
Yos-MacBook-Pro:Assignment3 yoo$
```

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

I will inspect the first line that’s showing, remember that:
Column 4 shows *post_name*, the post name.
Column 5 shows *message*, the message.

Column 19 gives *post_link,* if there are any(there are some NULL).
Column 21 shows *posted_at*, date of the post.

And because I am looking at news articles by official platforms, I am assuming that they won't disobey the english grammar and use "malaysia airlines" by not using capitalization. No need to consider lower cases for this one.
The first line mentioning "Malaysia Airlines" is:

abc-news,86680728811_10152267754078812,86680728811,NULL,DEVELOPING: Malaysia Airlines spokesperson: Flight carrying 239 people from Kuala Lumpur to Beijing has gone missing contact lost: http://abcn.ws/NHHeLT,NULL,NULL,status,mobile_status_update,1583,435,1526,0,0,0,0,0,0,NULL,NULL,8/3/14 0:47

This post was first mentioned by abc-news. Although post_name and post_link are not given, the message includes the article link. So to summarize:

The first mention is **8/3/14 0:47(8th March 2014 0:47).**
DEVELOPING: Malaysia Airlines spokesperson: Flight carrying 239 people from Kuala Lumpur to Beijing has gone missing contact lost: http://abcn.ws/NHHeLT
Was the message.
And this was first mentioned by **ABC News**.

Q7)

I have referred to *hek2mgl's* approach, retrieved from:
https://stackoverflow.com/questions/31038073/print-a-row-only-if-the-string-exist-in-a-specific-column
.
awk -F, '$5 ~ /Donald Trump/' FB_Dataset | wc -l // uses awk to filter out a string with a message column that contains(doesn't have to be exactly equal), and pipes it to the word counter.
"," after -F indicates that my file is comma separated.
Outputs 3298

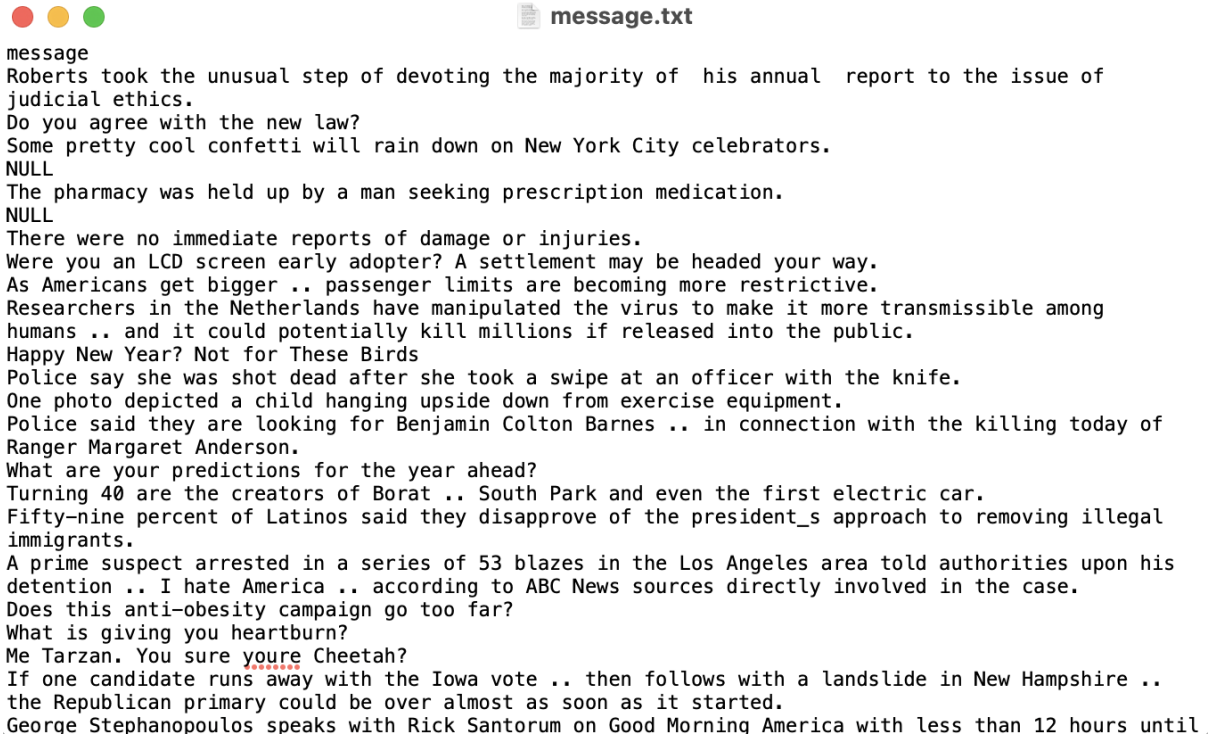
```
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /Donald Trump/' FB_Dataset | wc -l
     3298
```

I find this above approach more fitting as we are using this data in comparing popularity in Q9. Basically this method counts the number of lines that contain "Donald Trump", and there is another method that counts the number of *occurrences,* which include duplicate "Donald Trump" within the same line. Although having a person's name multiple times does indicate how enthusiastic the post is, because the news media can abuse this statistics by including the name more than 2 times, in order to ensure the fairness of my analysis I am taking this approach to count the *lines*, not *occurrences*. This is my interpretation of this question.

For your knowledge, here is my other approach that includes duplicates. This method is brainstormed from these 2 websites:

https://stackoverflow.com/questions/6741967/how-can-i-count-the-occurrences-of-a-string-within-a-file
https://unix.stackexchange.com/questions/398413/counting-occurrences-of-word-in-text-file
cut -d"," -f5 FB_Dataset > message.txt

// this extracts all data column 5 into new text file message.txt, to be filtered later



Next, I type in
grep -o "Donald Trump" message.txt | wc -l
Which extracts line containing “Donald Trump” in message.txt file and pipes it to word count(counts it). I am using the -o option(only matching) to find the number of occurrences.

```
Yos-MacBook-Pro:Assignment3 yoo$ cut -d"," -f5 FB_Dataset >  message.txt
Yos-MacBook-Pro:Assignment3 yoo$ grep -o "Donald Trump" message.txt | wc -l
     3321
```

3321
Is the number of posts including the duplicate “Donald Trump” within a single line.
Notice it is slightly higher than 3298.

But for the reason given above, I consider the first approach more appropriate so I will not be using it.

awk -F, '$5 ~ /donald trump/' FB_Dataset | wc -l
awk -F, '$5 ~ /donald Trump/' FB_Dataset | wc -l
awk -F, '$5 ~ /Donald trump/' FB_Dataset | wc -l Returned 3
awk -F, '$5 ~ /DONALD TRUMP/' FB_Dataset | wc -l Returned 6

```
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /Donald Trump/' FB_Dataset | wc -l
    3298
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /donald trump/' FB_Dataset | wc -l
       0
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /donald Trump/' FB_Dataset | wc -l
       0
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /Donald trump/' FB_Dataset | wc -l
       3
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /DONALD TRUMP/' FB_Dataset | wc -l
       6
```

I have tried combinations of other lower-upper cases, and there are a total of 3298 + 3 + 6 = 3307 posts with messages relating to Donald Trump.

As expected from mainstream media, most of them used correct capital letter form.

For the word “Donald”, although limited to media sources, it could point to something else.
awk -F, '$5 ~ /McDonald/' FB_Dataset | wc -l
For instance, I have found 625 cases for lines containing the string “McDonald”.

```
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /McDonald/' FB_Dataset | wc -l
     625
```

Therefore for this question, I am only searching by full name. I don't feel the necessity yet to deal with “Trump”, yet.

So my conclusion is, “Donald Trump” is mentioned in the message column of the file **3307** times.

Q8)
I am using a similar approach for Q7, referenced from the same source https://stackoverflow.com/questions/31038073/print-a-row-only-if-the-string-exist-in-a-specific-column .

awk -F, '$5 ~ /Barack Obama/' FB_Dataset | wc -l // , indicates the delimiter is common, and I search if column 5 contains the string “Barack Obama”, put to the word counter.
awk -F, '$5 ~ /barack obama/' FB_Dataset | wc -l
awk -F, '$5 ~ /Barack obama/' FB_Dataset | wc -l
awk -F, '$5 ~ /barack Obama/' FB_Dataset | wc -l
awk -F, '$5 ~ /BARACK OBAMA/' FB_Dataset | wc -l

```
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /Barack Obama/' FB_Dataset | wc -l
    3629
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /barack obama/' FB_Dataset | wc -l
       0
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /Barack obama/' FB_Dataset | wc -l
       0
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /barack Obama/' FB_Dataset | wc -l
       0
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /BARACK OBAMA/' FB_Dataset | wc -l
       1
```

I have searched through lines that contain a string “Barack Obama”, “barack obama”, “Barack obama”, “barack Obama” and “BARACK OBAMA”. There are a total of 3629 + 1 = 3630 Outputs.

I would like to ***define*** the term “popular” as, a person with more mentioned posts on message column on Facebook during a period of time from top media sources, taking into account their *full name*.

From my definition,“Barack Obama” with 3630 posts beats “Donald Trump” with 3307 posts, and I conclude that Obama is more popular.

WRONG APPROACH:

Initially with this task, I accidentally used column 4(post name) instead of column 5(message). I only realized it at Q9), so redid all the questions in Q7 and 8.

Below is the document I first had, where I faced a weird case of having way less “Barack Obama” compared to “Donald Trump” in the *post name.*
I would like to leave it below as it was an interesting opportunity for me to think about weird outliers and what bias is. I have decided to leave it out here as I found this text connects to my solution in Q10. I would appreciate it if you could just consider this as a fun text and a reference to Q10, thank you.

awk -F, '$4 ~ /Barack Obama/' FB_Dataset | wc -l
awk -F, '$4 ~ /barack obama/' FB_Dataset | wc -l
awk -F, '$4 ~ /Barack obama/' FB_Dataset | wc -l
awk -F, '$4 ~ /barack Obama/' FB_Dataset | wc -l

```
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$4 ~ /Barack Obama/' FB_Dataset | wc -l
     195
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$4 ~ /barack obama/' FB_Dataset | wc -l
       0
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$4 ~ /Barack obama/' FB_Dataset | wc -l
       0
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$4 ~ /barack Obama/' FB_Dataset | wc -l
       0
```

I have searched through lines that contain a string “Barack Obama”, “barack obama”, “Barack obama”, “barack obama”. I found 195 posts only and 0s for the lower-cases.

However,

```
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$4 ~ /Barack/' FB_Dataset | wc -l
     226
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$4 ~ /Obama/' FB_Dataset | wc -l
   12818
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$4 ~ /Trump/' FB_Dataset | wc -l
   19296
```

awk -F, '$4 ~ /Obama/' FB_Dataset | wc -l
Conducting a search on “Obama” returned 12818 posts. This is far more than 195, but understandable from common sense as his last name Obama is way more familiar with the public. Furthermore, there is nothing other than himself “Obama” points to.

Now it is not easy to draw conclusions.

awk -F, '$4 ~ /Trump/' FB_Dataset | wc -l

Trying a search on Trump resulted in 19296 cases.

	Donald Trump	Barack Obama
Full name	7692	195
Trump, Obama	19296	12818

It is not fair to conclude Obama is more popular with 12818 posts, while I am using “Donald Trump” - full name only.

I must be flexible here to draw out unbiased conclusions.

awk -F, '$4 ~ /Trump/' FB_Dataset | head -10

abc-news,86680728811_3228583744419787,86680728811,Trump Threatens Third Party Run,What do you think of Trump possibly throwing his hat into the ring as a Third party candidate?,Bobby Bank/WireImage/Getty Images Donald Trump is keeping the door open to a possible presidential bid .. stoking rumors again today that he could toss his hat into the ring and run as an independent if he does not believe the event ual Republican nominee can beat President Obama. _It_s not something I...,abcnews.go.com,link,shared_story,349,760,88,0,0,0,0,0,0,http://abcnews.go.com/blogs/politics/2012/01/sunday-squabbles-trump-stokes-presidential-bid-rumors-rnc-chair-says-obama-jumped-ship/,https://external.xx.fbcdn.net/safe_image.php?d=AQBh6uOvdTdfJS7r&w=130&h=130&url=http%3A%2F%2Fabcnews.go.com%2Fimages%2FPolitics%2Fgty_donald_trump_jef_111202_wblog.jpg&cfs=1&sx=91&sy=0&sw=269&sh=269,29/1/12 19:48

abc-news,86680728811_3254667008825405,86680728811,Donald Trump Staff Reaching Out to Financers .. Campaign Managers to Explore Third Party Bid,The Donald is back at it.,Donald Trump_s top political adviser told ABC News he_s speaking to _high-level political operatives_ to explore a third-party bid for the real estate mogul .. in a sign that Trump may not quite be done with the 2012 presidential race.,abcnews.go.com,link,shared_story,174,223,68,0,0,0,0,0,0,http://abcnews.go.com/blogs/politics/2012/01/donald-trump-staff-reaching-out-to-financers-campaign-managers-to-explore-third-party-bid/,https://external.xx.fbcdn.net/safe_image.php?d=AQBh6uOvdTdfJS7r&w=130&h=130&url=http%3A%2F%2Fabcnews.go.com%2Fimages%2FPolitics%2Fgty_donald_trump_jef_111202_wblog.jpg&cfs=1&sx=91&sy=0&sw=269&sh=269,30/1/12 21:07

abc-news,86680728811_1475565553611835,86680728811,Donald Trump Set To Endorse Mitt Romney In Las Vegas,Does Trumps endorsement matter to you?,Donald Trump will endorse Mitt Romney in Las Vegas on Thursday .. sources familiar with the decision tell ABC News. ,abcnews.go.com,link,shared_story,124,559,18,0,0,0,0,0,0,http://abcnews.go.com/blogs/politics/2012/02/donald-trump-set-to-endorse-mitt-romney-in-las-vegas/,https://external.xx.fbcdn.net/safe_image.php?d=AQCplIWc8D1LqJN4&w=130&h=130&url=http%3A%2F%2Fa.abcnews.com%2Fimages%2FWNT%2Fabc_wn_romney_120201_wl.jpg&cfs=1,2/2/12 15:53

abc-news,86680728811_2757005091780995,86680728811,Trump Wont Bar Transsexual From Miss Universe Canada,Should she be barred or allowed to compete?,Jenna Talackova .. the Canadian beauty queen .. won_t have to fight Donald Trump for a chance to compete for the Miss Universe Canada title after all. ,abcnews.go.com,link,shared_story,436,460,104,0,0,0,0,0,0,http://abcnews.go.com/blogs/entertainment/2012/04/trump-won-bar-transsexual-beauty-queen-from-shot-at-miss-universe-canada-title/,https://external.xx.fbcdn.net/safe_image.php?d=AQBJ6pTJMEyXb8Vi&w=130&h=130&url=http%3A%2F%2Fa.abcnews.go.com%2Fimages%2FUS%2Fgty_jenna_talackova_tk_120326_wblog.jpg&cfs=1&sx=6&sy=0&sw=269&sh=269,3/4/12 0:49

abc-news,86680728811_527836793897957,86680728811,Greed Trumps Envy in New Investment Study,Here_s the gist: We are all happy to lose money .. just as long as other people lose more.,_It is not enough that I succeed .._ the late Gore Vidal once noted. _Others must fail._ Vidal .. who also famously admitted that _a little something_ in him dies whenever a friend succeeds .. would have had a field day with a new study .. _Investment Choices with Envy and Alt_,abcnews.go.com,link,shared_story,96,12,16,0,0,0,0,0,0,http://abcnews.go.com/blogs/business/2012/10/greed-trumps-envy-in-new-investment-study/,https://external.xx.fbcdn.net/safe_image.php?d=AQAfMm85MfCkNgQr&w=130&h=130&url=http%3A%2F%2Fabcnews.go.com%2Fimages%2FBusiness%2Fgty_money_greed_envy_lpl_121004_wblog.jpg&cfs=1,5/10/12 2:00

abc-news,86680728811_1636566737386670,86680728811,Petition Urges Macys to Dump Donald Trump,Has the Donald become too controversial?,Cher tweets outrage .. is among petitions nearly 500 ..000 signatures ,abcnews.go.com,link,shared_story,2023,791,226,0,0,0,0,0,0,http://abcnews.go.com/Business/petition-urges-macys-dump-donald-trump/story?id=17710529,https://external.xx.fbcdn.net/safe_image.php?d=AQDK2fftoeVbaBgB&w=130&h=130&url=http%3A%2F%2Fa.abcnews.go.com%2Fimages%2FUS%2Fap_donald_trump_jef_121113_main.jpg&cfs=1&sx=82&sy=0&sw=310&sh=310,14/11/12 3:49

abc-news,86680728811_10151773394288812,86680728811,Donald Trump: Sen. Ted Cruz 'Perhaps Not' Eligible for White House 'If' Born in Canada,_If he was born in Canada perhaps not._ Trump told ABC_s Jonathan Karl.,ABC During an interview for _This Week_ real estate mogul Donald Trump questioned whether Texas Sen. Ted Cruz would be eligible for the White House given his Canadian birth. _If he was born in Canada perhaps not._ ,abcnews.go.com,link,shared_story,485,530,80,0,0,0,0,0,0,http://abcnews.go.com/blogs/politics/2013/08/donald-trump-sen-ted-cruz-perhaps-not-eligible-for-white-house-if-born-in-canada/,https://external.xx.fbcdn.net/safe_image.php?d=AQCAzl7uoDxkuX0X&w=130&h=130&url=http%3A%2F%2Fa.abcnews.com%2Fimages%2FPolitics%2FABC_donald_trump_this_week_jt_130811_16x9_608.jpg&cfs=1&sx=196&sy=0&sw=342&sh=342,11/8/13 16:00

abc-news,86680728811_10152623555133812,86680728811,Homeowner Who Triumphed Over Donald Trump Is Moving On,Vera Coking became a folk hero for resisting decades-long efforts by big-name developers like Donald Trump to displace her Atlantic City boardinghouse. http://abcn.ws/1knSRHs,She once called Donald Trump 'a maggot a cockroach and a crumb.' This week he remembered her as 'an impossible person.' The woman who became a folk hero for resisting decades-long efforts by big-name developers like Trump to displace her Atlantic City boardinghouse is now 91 and...,abcnews.go.com,link,shared_story,1149,46,94,0,0,0,0,0,0,http://abcn.ws/1knSRHs,https://external.xx.fbcdn.net/safe_image.php?d=AQAvWMimS_2dcyFK&w=130&h=130&url=http%3A%2F%2Fa.abcnews.com%2Fimages%2FUS%2FAP_TRUMP_140731_DG_16x9_992.jpg&cfs=1,31/7/14 8:08

abc-news,86680728811_10152623771458812,86680728811,Homeowner Who Triumphed Over Donald Trump Is Moving On,The 91-year-old woman once called Donald Trump 'a maggot a cockroach and a crumb.' He called her 'an impossible person.',She once called Donald Trump 'a maggot a cockroach and a crumb.' This week he remembered her as 'an impossible person.' The woman who became a folk hero for resisting decades-long efforts by big-name developers like Trump to displace her Atlantic City boardinghouse is now 91 and...,abcnews.go.com,link,shared_story,1348,76,87,0,0,0,0,0,0,http://abcn.ws/1odAvUl,https://external.xx.fbcdn.net/safe_image.php?d=AQAvWMimS_2dcyFK&w=130&h=130&url=http%3A%2F%2Fa.abcnews.com%2Fimages%2FUS%2FAP_TRUMP_140731_DG_16x9_992.jpg&cfs=1,31/7/14 10:48

abc-news,86680728811_10152638605648812,86680728811,Trump Sues to Get Name off Casinos,Donald Trump has a message for the two Atlantic City casinos that still bear his name: You're fired. http://abcn.ws/1v8asYb,The real estate mogul and reality TV star who presided over a casino empire in the glory days of Atlantic City filed a lawsuit on Tuesday demanding that his name be stripped from the remaining two.,abcnews.go.com,link,shared_story,678,150,64,0,0,0,0,0,0,http://abcn.ws/1v8asYb,https://external.xx.fbcdn.net/safe_image.php?d=AQAlmNGwlrQJcpep&w=130&h=130&url=http%3A%2F%2Fa.abcnews.com%2Fimages%2FPolitics%2FAP_donald_trump_jt130811_16x9_992.jpg&cfs=1,6/8/14 9:24

awk -F, '$4 ~ /Trump/' FB_Dataset | tail -10

usa-today,13652355666_10154016663165667,13652355666,Clinton tells black voters Trump would _destroy_ Obama legacy,The schedule Clinton is keeping in the final days of her campaign reflects Democrats' anxiety over whether African Americans particularly younger voters will turn out in the same numbers they did for Obama in 2008 and 2012.,Hillary Clinton is closing her campaign with a warning to African Americans that Donald Trump would _destroy_ President Obama_s legacy as her campaign seeks to motivate black voters to get to the polls.,usatoday.com,link,shared_story,10937,673,109,151,19,105,9,0,37,http://www.usatoday.com/story/news/politics/onpolitics/2016/11/06/clinton-tells-black-voters-trump-would-destroy-obama-legacy/93393316/,https://external.xx.fbcdn.net/safe_image.php?d=AQAQscm6J00DFa3t&w=130&h=130&url=http%3A%2F%2Fwww.gannett-cdn.com%2F-mm-%2F5f8a50e13a1d2975c346f32f645c6be472aff11c%2Fc%3D0-170-3000-1865%26r%3Dx633%26c%3D1200x630%2Flocal%2F-%2Fmedia%2F2016%2F11%2F06%2FUSATODAY%2FUSATODAY%2F636140369787851289-CLINTON-AND-BLACKS.JPG&cfs=1&sx=477&sy=0&sw=630&sh=630,7/11/16 1:26

usa-today,13652355666_10154017788855667,13652355666,Following scare at Trump rally life is now different for protester,'I_m very cognizant that there is going to be a portion of the U.S. population that is going to doubt my true intentions no matter what I say.',Life was fairly normal for Trump protester Austyn Crites until Saturday night.,usatoday.com,link,shared_story,13071,433,255,79,166,47,250,0,35,http://www.usatoday.com/story/news/politics/elections/2016/11/06/trump-rally-protester/93407106/,https://external.xx.fbcdn.net/safe_image.php?d=AQAd40-iss0WMnjv&w=130&h=130&url=http%3A%2F%2Fwww.gannett-cdn.com%2F-mm-%2F953e8dc17bbc9f62499e855dcc25a5a63519fedf%2Fc%3D0-184-5177-3109%26r%3Dx1683%26c%3D3200x1680%2Flocal%2F-%2Fmedia%2F2016%2F11%2F06%2FReno%2FRGJ%2F636140541995465929-IMG-7308.JPG&cfs=1&sx=276&sy=0&sw=1680&sh=1680,7/11/16 3:20

usa-today,13652355666_10154018168780667,13652355666,I know Trump. Don't vote for him: Ex-Harrah's executive,My opinion of Donald Trump from the 1980s has not changed. The negative publicity about Trump during this campaign _ his conduct toward women his business failures and his explosive temperament _ matches my dealings with him.'More in USA TODAY Opinion:,I wrote of his bluster threats intemperance and falsehoods in 1985. Nothing has changed.,usatoday.com,link,shared_story,16768,629,1119,235,105,200,15,0,66,http://www.usatoday.com/story/opinion/nation-now/2016/11/06/dont-vote-trump-former-harrahs-executive/93353464/,https://external.xx.fbcdn.net/safe_image.php?d=AQAXl9-Z7YYdZ-x7&w=130&h=130&url=http%3A%2F%2Fwww.gannett-cdn.com%2F-mm-%2F2ce46401a2b0343ad76ddfaf989acee881cf6fac%2Fc%3D0-251-513-541%26r%3Dx633%26c%3D1200x630%2Flocal%2F-%2Fmedia%2F2016%2F11%2F05%2FUSATODAY%2FUSATODAY%2F636139610203696538-trump-plaza.JPG&cfs=1,7/11/16 5:02

usa-today,13652355666_10154018830490667,13652355666,Robert De Niro knocks 'insane' Donald Trump at Hollywood Film Awards,So let me lay it out right here. We have the opportunity to prevent comedy from turning into a tragedy ' Robert De Niro said.,De Niro took to the podium to hit at Trump before the election and the crowd loved it.,usatoday.com,link,shared_story,12752,576,348,323,40,146,9,0,41,http://www.usatoday.com/story/life/movies/2016/11/07/robert-de-niro-knocks-insane-donald-trump-hollywood-film-awards/93413986/,https://external.xx.fbcdn.net/safe_image.php?d=AQAqFS7bmzn7rs5f&w=130&h=130&url=http%3A%2F%2Fwww.gannett-cdn.com%2F-mm-%2F26afbcfb592e6fa5b918b9049352d37ccf10d872%2Fc%3D0-35-2479-1435%26r%3Dx633%26c%3D1200x630%2Flocal%2F-%2Fmedia%2F2016%2F11%2F07%2FUSATODAY%2FUSATODAY%2F636140788281764380-AP-2016-HOLLYWOOD-FILM-AWARDS---SHOW-86536346.JPG&cfs=1&sx=362&sy=0&sw=630&sh=630,7/11/16 11:35

usa-today,13652355666_10154018999690667,13652355666,What would Clinton or Trump actually do?: Robert Robb,This is a character-driven race and there is much about the character of each candidate for voter s to weigh.,The endgame is all about character flaws but real-world challenges await the winner.,usatoday.com,link,shared_story,9252,368,14,77,21,80,3,0,6,http://usat.ly/2eEWf1r,https://external.xx.fbcdn.net/safe_image.php?d=AQBkgCvByurt_Fsr&w=130&h=130&url=http%3A%2F%2Fwww.gannett-cdn.com%2F-mm-%2Fdbc3069436d6ae946a9aa4cd1dcd3c0018ba7b06e%2Fc%3D0-126-977-678%26r%3Dx633%26c%3D1200x630%2Flocal%2F-%2Fmedia%2F2016%2F11%2F06%2FUSATODAY%2FUSATODAY%2F636140448186412232-hillary-donald.JPG&cfs=1&sx=203&sy=0&sw=630&sh=630,7/11/16 13:50

usa-today,13652355666_10154019447425667,13652355666,Lady Gaga calls Melania Trump's anti-bullying stance 'hypocrisy','Your husband is (one) of the most notorious bullies we have ever witnessed._,NULL,usatoday.com,link,shared_story,20570,2407,1493,1350,72,733,24,0,161,http://www.usatoday.com/story/life/entertainthis/2016/11/07/lady-gaga-calls-melania-trumps-anti-bullying-stance-hypocrisy/93413914/,https://external.xx.fbcdn.net/safe_image.php?d=AQA3sXsmI73rwlVx&w=130&h=130&url=http%3A%2F%2Fwww.gannett-cdn.com%2F-mm-%2F160ee1d252cf6a4b8ead238df6381b6b1440f5%2Fc%3D0-0-983-555%26r%3Dx633%26c%3D1200x630%2Flocal%2F-%2Fmedia%2F2016%2F11%2F07%2FUSATODAY%2FUSATODAY%2F636141401910986151l-gaga-melania.png&cfs=1&sx=300&sy=0&sw=630&sh=630,7/11/16 15:50

usa-today,13652355666_10154019884925667,13652355666,Trump conducts election eve campaign marathon,We are going to have one of the great victories of all time _ of all time ' Trump said in Virginia. 'This is going to be Brexit times 50.',Donald Trump says that enthusiasm for him and anger with the political establishment will deliver a victory that stuns his critics.,usatoday.com,link,shared_story,7930,524,33,114,25,191,11,0,49,http://www.usatoday.com/story/news/politics/elections/2016/11/07/donald-trump-campaign-final-day/93412700/,https://external.xx.fbcdn.net/safe_image.php?d=AQAdVvTW3cBk1dhK&w=130&h=130&url=http%3A%2F%2Fwww.gannett-cdn.com%2F-mm-%2Fba8543785aeb7925c1a20cbad67b2e03be19f13d%2Fc%3D0-2-4456-2519%26r%3Dx1683%26c%3D3200x1680%2Flocal%2F-%2Fmedia%2F2016%2F11%2F07%2FUSATODAY%2FUSATODAY%2F636141011594967586-AP-CAMPAIGN-2016-TRUMP-86541028.JPG&cfs=1&sx=680&sy=0&sw=1680&sh=1680,7/11/16 18:30

usa-today,13652355666_10154019934620667,13652355666,John Oliver is super sorry he asked Trump to run for president,Come on John.,Remember when people thought it might be funny if Donald Trump ran for president? John Oliver does. ,usatoday.com,link,shared_story,7314,315,34,61,23,114,4,0,0,15,http://www.usatoday.com/story/life/entertainthis/2016/11/07/john-oliver-is-super-sorry-he-asked-trump-to-run-for-president/93420392/,https://external.xx.fbcdn.net/safe_image.php?d=AQAP4AlOmvc78aM6&w=130&h=130&url=http%3A%2F%2Fwww.gannett-cdn.com%2F-mm-%2F09b97a2d7e2db39df5823c7c0318480cb71c74dc%2Fc%3D0-133-4577-2719%26r%3Dx1683%26c%3D3200x1680%2Flocal%2F-%2Fmedia%2F2016%2F03%2F21%2FUSATODAY%2FUSATODAY%2F636941649585358187-AP-TV-John-Oliver.jpg&cfs=1&sx=804&sy=0&sw=1680&sh=1680&1,7/11/16 19:30

usa-today,13652355666_10154019826825667,13652355666,Trump aide: We did not take Trump's Twitter away,No Donald Trump_s Twitter did not get taken away.,His campaign pushes back on a report that aides had stopped the nominee from tweeting.,usatoday.com,link,shared_story,6470,301,30,55,13,119,4,0,5,http://www.usatoday.com/story/news/politics/onpolitics/2016/11/07/trump-aide-we-did-not-take-trumps-twitter-away/93425352/,https://external.xx.fbcdn.net/safe_image.php?d=AQBJVHKla6sRmzYv&w=130&h=130&url=http%3A%2F%2Fwww.gannett-cdn.com%2F-mm-%2Fffc8ea4b567bacd90272521be77b2cb712d5aa8b%2Fc%3D0-207-4500-2749%26r%3Dx1683%26c%3D3200x1680%2Flocal%2F-%2Fmedia%2F2016%2F11%2F07%2FUSATODAY%2FUSATODAY%2F636141184455760711-AFP-AFP-HV1H3-86545364.JPG&cfs=1&sx=324&sy=0&sw=1680&sh=1680,7/11/16 19:50

usa-today,13652355666_1738781909779663,13652355666,De Niro calls Trump 'totally insane',Just ahead of the election Robert De Niro went off on Donald Trump on national television at the Hollywood Film Awards. http://usat.ly/2fuLMF6,NULL,NULL,video,added_video,7120,568,350,225,15,75,10,0,144,https://www.facebook.com/usatoday/videos/1738781909779663/,https://scontent.xx.fbcdn.net/v/t15.0-10/s130x130/14926259_1738779633113224_2157279791739201536_n.jpg?oh=6fad14c5b4f32e921029aecd5bcf0da8&oe=58D4969C,7/11/16 23:00


```
Yos-MacBook-Pro:Assignment3 yoo$
```

Sorry that they are hard to see - I have obtained head 10 and tails 10 lines of the search by “Trump”, and manually checked that they are actually Donald Trump related. From my understanding of general knowledge and common sense, the media are very unlikely to create posts regarding someone else or the noun including “Trump” (even if there is an article on, say, “trumpet”, it would usually be lower-cased anyway, so nouns aren't a problem).
So in my Q10, I am using “Trump” and “Obama” to keep the fairness of my approach.
~~While checking 20 lines is not statistically sufficient, I would like to ***define*** the term “popular” as, a person with more mentioned posts on Facebook during a period of time from top media sources, with taking into account their **“last name only”**.~~
~~It is impossible to draw a perfect approach ignoring outliers, however I find this the most unbiased considering the circumstances given.~~

~~From my definition, I conclude that Donald Trump is more popular with **19296** relevant posts, against Barack Obama with **12818.**~~

Q9)
I am expecting to use multiple pipings as I want to complete this with one single command without intermediate files.

I have referred to:
https://askubuntu.com/questions/987512/how-to-extract-only-values-greater-than-a-threshold-from-a-file
Column 2: post id
Column 5: message
Column 10: likes count

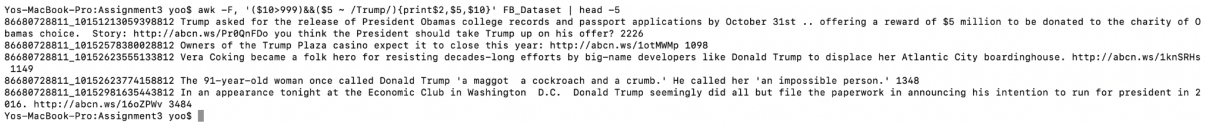
awk -F, '{if($10>999)print$2,$5,$10}' FB_Dataset | head -5
// uses awk with -F’,’, comma taken as delimiter unlike tutorial or many other online sources. It uses the if statement to see if the like count is greater than 999(greater or equal to 1000), and then prints only 3 columns that we are interested in. For this case I output into the head of line 5, and looks like this code actually works.

```
Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '{if($10>999)print$2,$5,$10}' FB_Dataset | head -5
post_id message likes_count
86680728811_193512637411477 The mother went through with it after her 911 operator gave her the okay to shoot. 1702
86680728811_275162259212376 Do you think a new Barbie should lose her locks? Or should that be left to kids with scissors? 1370
86680728811_301872166525725 _It_s unbelievable ..._ she told ABC News today. _It might as well be the lottery._ 1029
86680728811_10150532426693812 Do you approve of the tough lesson? 1494
```

This correctly shows id, message and likes count greater than or equal to 1000.

Now I have to filter for messages containing “Trump”.
awk -F, '$5 ~ /Trump/' FB_Dataset | head -10
// This was my previous code, so I will try to combine.

awk -F, '($10>999)&&($5 ~ /Trump/){print$2,$5,$10}' FB_Dataset | head -5
// I have done it! Without referring to anything, here is the code that worked after many try & errors.



I can see that among 3 columns, the last column is always greater than 999 and Trump is included in the second column.

In fact, the specification sheet says that I don't need to record the message column. So I will only adjust it to print 2 columns only:

awk -F, '($10>999)&&($5 ~ /Trump/){print$2,$10}' FB_Dataset | head -5
// $5 removed

```
Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '($10>999)&&($5 ~ /Trump/){print$2,$10}' FB
_Dataset | head -5
86680728811_10151213059398812 2226
86680728811_10152578380028812 1098
86680728811_10152623555133812 1149
86680728811_10152623774158812 1348
86680728811_10152981635443812 3484
```

```
Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '($10>999)&&($5 ~ /Trump/){print$2,$10}' FB
_Dataset | wc -l
   11809
Yos-MacBook-Pro:Assignment3 yoo$
```

awk -F, '($10>999)&&($5 ~ /Trump/){print$2,$10}' FB_Dataset | wc -1
// For your knowledge, there are 11809 rows filling this condition.

My gratitude to:
https://stackoverflow.com/questions/6438896/sorting-data-based-on-second-column-of-a-file

I have used my deducted code and connected it to pipe into sorting command.

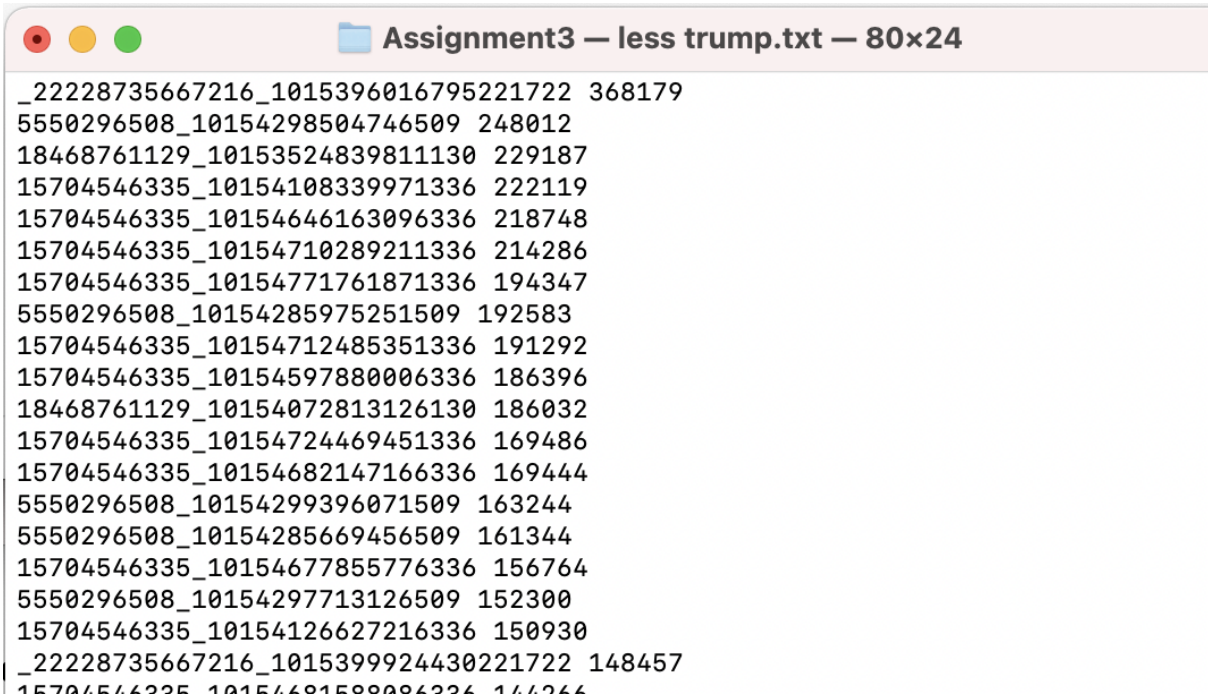
awk -F, '($10>999)&&($5 ~ /Trump/){print$2,$10}' FB_Dataset | sort -k2 -n -r | head -10

// Sort part implies sorting, -kn specifying to sort based on the second column(like count), to sort numerically by -n, and to sort in descending order and -r indicating descending order.
Here is my output:

```
Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '($10>999)&&($5 ~ /Trump/){print$2,$10}' FB
_Dataset | sort -k2 -n -r | head -10
_22228735667216_1015396016795221722 368179
5550296508_10154298504746509 248012
18468761129_10153524839811130 229187
15704546335_10154108339971336 222119
15704546335_10154646163096336 218748
15704546335_10154710289211336 214286
15704546335_10154771761871336 194347
5550296508_10154285975251509 192583
15704546335_10154712485351336 191292
15704546335_10154597880006336 186396
```

I do not have to use descending order(-r, reverse), but I prefer this because ones with more impressions come up first.

So now it is time to output to a new text file.
awk -F, '($10>999)&&($5 ~ /Trump/){print$2,$10}' FB_Dataset | sort -k2 -n -r > trump.txt
// output to a new text file. And then,
less trump.txt // have led me to the below. successful.



I must not forget to add post_id and like_count as header at the beginning of the file.

~~sed -i '1i post_id like_count' test.txt~~
~~sed -i '1i post_id like_count' trump.txt~~
I tried to use sed to insert a header at the beginning of the file, but somehow it resulted in an error. I tried many hours on this but failed.
Mainly this weird error that does not read “t”

```
(base) yoo@Yos-MacBook-Pro Assignment3 % sed  -i '1i post_id like_count' trump.t
xt
sed: 1: "trump.txt": undefined label 'rump.txt'
```

So I decided to delete the current trump.txt, create a new empty one, add space separated headers and append the data.
Let’s go :)

rm trump.txt
// Removes text
ls // lists folder to check make sure it is deleted

```
(base) yoo@Yos-MacBook-Pro Assignment3 % rm trump.txt
(base) yoo@Yos-MacBook-Pro Assignment3 % ls
FB_Dataset      Output          message.txt
(base) yoo@Yos-MacBook-Pro Assignment3 %
```

```
(base) yoo@Yos-MacBook-Pro Assignment3 % echo "post_id like_count" >> trump.txt
(base) yoo@Yos-MacBook-Pro Assignment3 %
```

echo “post_id like_count” >> trump.txt

awk -F, '($10>999)&&($5 ~ /Trump/){print$2,$10}' FB_Dataset | sort -k2 -n -r >> trump.txt
// Now I append my lines. I use double >> to show that we are appending, not updating.

```
(base) yoo@Yos-MacBook-Pro Assignment3 % echo "post_id like_count" >> trump.txt
(base) yoo@Yos-MacBook-Pro Assignment3 % awk -F, '($10>999)&&($5 ~ /Trump/){print$2,$10}' FB_Dataset | sort -k2 -n -r >> trump.txt
(base) yoo@Yos-MacBook-Pro Assignment3 % head -5 trump.txt
post_id like_count
_22228735667216_1015396016795221722 368179
5550296508_10154298504746509 248012
18468761129_10153524839811130 229187
15704546335_10154108339971336 222119
(base) yoo@Yos-MacBook-Pro Assignment3 %
```

head -5 trump.txt // prints out header for first 5 rows
And Yes! I finally made it.

tail -5 trump.txt // prints the tail, last 5 rows for my trump.txt
wc -l trump.txt // shows me that this contains 11810 lines.

```
(base) yoo@Yos-MacBook-Pro Assignment3 % tail -5 trump.txt
5550296508_10154955006956509 1001
5550296508_10154541840616509 1001
18468761129_10153650393881130 1001
86680728811_10154802229533812 1000
5550296508_10154572513421509 1000
(base) yoo@Yos-MacBook-Pro Assignment3 % wc -l trump.txt
   11810 trump.txt
(base) yoo@Yos-MacBook-Pro Assignment3 %
```

This concludes my Q9!

Q10)
love_count is at column 13 and angry_count is at column 18.

There are many ways to interpret this question, but I would like to find data where the person’s name is included in the post_name, not the message. Although message results in larger occurrences of the word than post_name, and this results in larger numbers for love_count and angry_count, I believe that people will press the reaction just based on the article title(post_name). Furthermore, the post_name is where the summary of the main context of the article is included - one can take a glance at it and judge what Trump or Obama has done, and it will usually include his name as well. I feel that I could observe people’s reactions more if we consider the post_id, because people on facebook might just read the summarized title and not click on the actual message.

In my Wrong Approach in Q8, I have talked about whether to use full name or last name only when conducting the search. For the reason given there, using “Barack Obama” in the post_name search resulted in too few numbers compared to “Obama”, because post_name

is usually summarized by the professional journalists of mainstream media so that full name is not observed very often. Plus, “Obama” is more familiar to public than his full name so to keep fairness and unbiasedness,
Although the question asks me to find counts for “Barack Obama” and “Donald Trump”, I am conducting a search on “Obama” and “Trump”, for post_name.
Lower-cases are not considered because they are grammatically wrong.
This approach does not contradict the question and I believe that this gives more reliable results.

Reference:
https://unix.stackexchange.com/questions/242946/using-awk-to-sum-the-values-of-a-column-based-on-the-values-of-another-column

I have referred to the above page for how to sum columns in BASH.

awk -F ',' '$4 ~ /Trump/ {sum += $13} END {print sum}' FB_Dataset
// This is the code I have used to sum up love_count($13) into a variable sum where post_name($4) contains the word “Trump”, where data is separated by delimiter ‘,’ and then prints out sum at the end.

```
Yos-MacBook-Pro:Assignment3 yoo$ awk -F ',' '$4 ~ /Trump/ {sum += $13} END {print sum}' FB_Dataset
3044289
```

3044289 is the love_count for Trump.

awk -F ',' '$4 ~ /Obama/ {sum += $13} END {print sum}' FB_Dataset

```
Yos-MacBook-Pro:Assignment3 yoo$ awk -F ',' '$4 ~ /Obama/ {sum += $13} END {print sum}' FB_Dataset
2015674
```

2015674 is the love_count for Obama.

awk -F ',' '$4 ~ /Trump/ {sum += $18} END {print sum}' FB_Dataset
// Similar command, but notice it is to find a sum for column 18(angry_count).

```
Yos-MacBook-Pro:Assignment3 yoo$ awk -F ',' '$4 ~ /Trump/ {sum += $18} END {print sum}' FB_Dataset
4499926
```

4499926 is the angry_count for Trump.

awk -F ',' '$4 ~ /Obama/ {sum += $18} END {print sum}' FB_Dataset

```
Yos-MacBook-Pro:Assignment3 yoo$ awk -F ',' '$4 ~ /Obama/ {sum += $18} END {print sum}' FB_Dataset
802881
```

802881 is the angry_count for Obama.

	Trump	Obama
love_count	3044289	2015674
angry_count	4499926	802881

Now I have data to draw out conclusions.
By taking a first look, Trump seems to have more love_count than Obama, but he has significantly more angry_count than Obama. We can't just conclude that Trump has more positive feelings among people.

I would like to justify my answer by saying that I will consider the ratio of love_count to the total number of people who reacted with both love_count and angry_count.
So my love ratio = love_count / (love_count + angry_count),
And whoever has the higher ratio can say that he has more positive feelings among people, because this also takes into account the number of people who have negative feelings towards them.

Love ratio for Trump:
3044289 / (3044289 + 4499926) = 0.4045 (3.s.f.)

Love ratio for Obama:
2015674 / (2015674 + 802881) = 0.7151 (3.s.f.)

Since Obama resulted in a higher love ratio, I would like to conclude that from my method and justification, **Obama has more positive feelings among people**.

Part B:

1)
I will first use the BASH shell to extract timestamps(*posted_at*, column 21) for columns for messages. The does not specify whether to consider post_name or message, but I will use *message* as "Barack Obama" has more occurrences in there compared to post_name(as I have shown in Q8).

```
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /Barack Obama/' FB_Dataset | wc -l
     3629
[Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$4 ~ /Barack Obama/' FB_Dataset | wc -l
      195
```

awk -F, '$5 ~ /Barack Obama/' FB_Dataset | wc -l
// (considering message) results in more numbers than
awk -F, '$4 ~ /Barack Obama/' FB_Dataset | wc -l
// (considering post_name) so I am using the former command.

In order to print a certain column into a csv file, I have referred to
https://stackoverflow.com/questions/19602181/how-to-extract-one-column-of-a-csv-file
So the command that I used is:

awk -F, '$5 ~ /Barack Obama/' FB_Dataset | awk -F "\"*,\"*" '{print $21}' > Obama.csv
// The left part is similar to the previous one, and for
awk -F "\"*,\"*" '{print $21}' > Obama.csv,
I am printing column 21(posted_at) from a comma separated data set and use > to assign into a new file called Obama.csv. Piping has made this possible.

```
Yos-MacBook-Pro:Assignment3 yoo$ awk -F, '$5 ~ /Barack Obama/' FB_Dataset | awk -F "\"*,\"*" '{print $2
1}' > Obama.csv
Yos-MacBook-Pro:Assignment3 yoo$ head -5 Obama.csv
24/1/12 19:26
10/4/12 11:43
19/4/12 12:43
29/4/12 14:23
6/9/12 3:33
```

This is the result, and I have successfully created Obama.csv.
Checked via: head -5 Obama.csv

24/1/12 19:26
10/4/12 11:43
19/4/12 12:43
29/4/12 14:23
6/9/12 3:33

Now out of these data, I need to decide on the format. From Q6 we know the dates of the Malaysia Airlines accident, which is 8th March 2014, and this was stored in the data as 8/3/14 0:47.
I can deduct that the format of the date is:
Day/Month/Year Hour: Minute

Referring to the
https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/strptime as given by the questions sheet,
"%d/%m/%y %H:%Mz" // will be my format for timestamps for strptime().

```
setwd("/Users/yoo/Assignment3")
```
// It sets the current working directory of Rstudio to Assignment3, my workplace for this task
```
obama <- read.csv("Obama.csv")
```
// this reads in csv file Obama and assign it to a variable called obama
```
x <- strptime(obama[,1], format = "%d/%m/%y %H:%M")
```
// this is strptime that converts obama of the given date/time format, and then assigns them all to a new variable named x. The reason obama[,1] is that this function takes in a column only, so I have specified to take in the first column(timestamp).

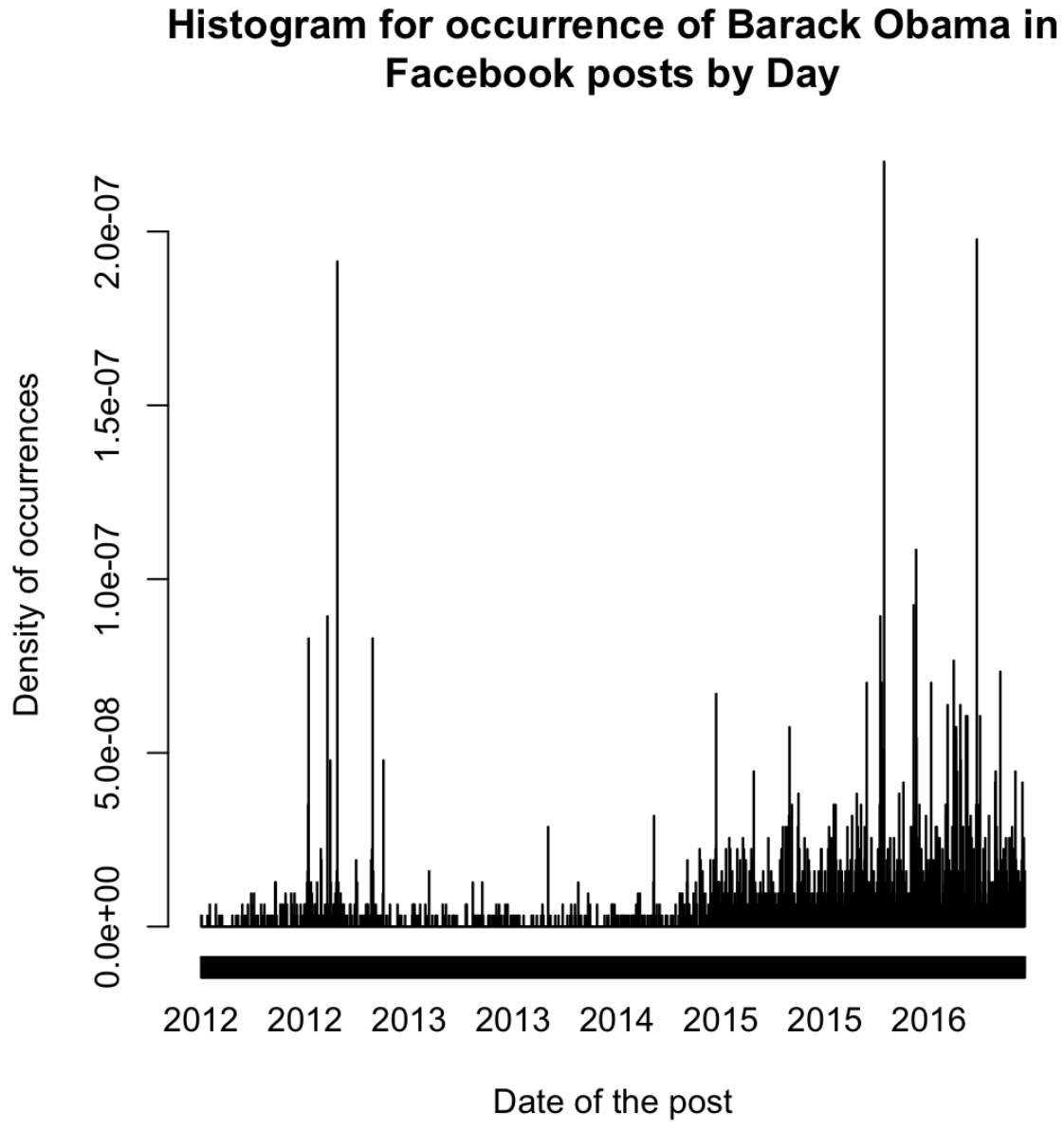
i)
I use hist() to plot the data in R.
```
hist(x, breaks = "days", main="Histogram for occurrence of Barack Obama in
Facebook posts", xlab = "Date of the post", ylab = "Density of occurrences")
```
// this is my command for plotting. Main specify the title of my histogram, and xlab shows my x label while ylab shows my y labels. Note that since this is a histogram, the area represents the occurrence of "Barack Obama" in posts in the period of time, and the y-axis does not directly indicate the value but represents density for that reason.
Here is my full script:

Question1.R*

Source on Save | Run | Source

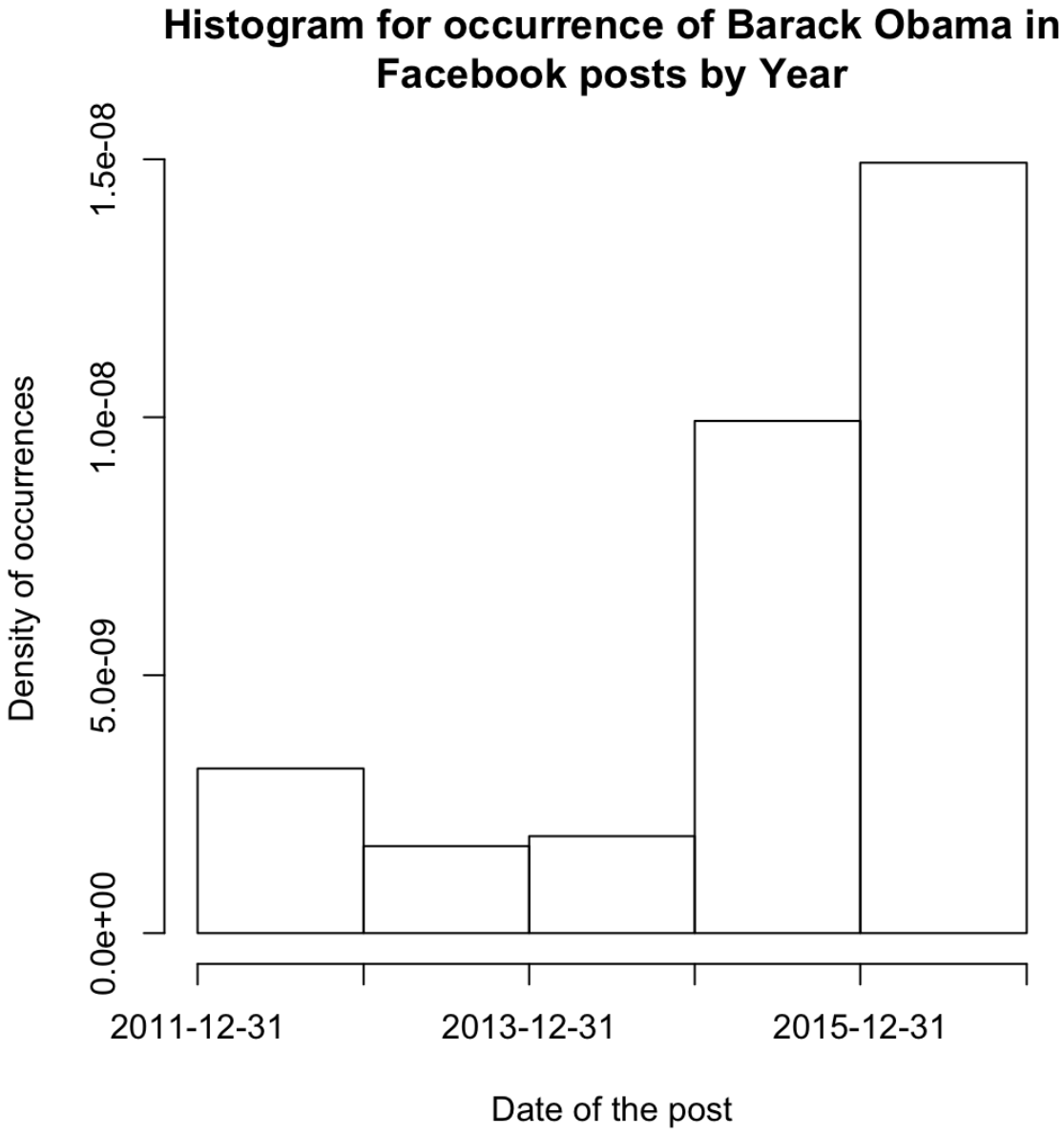
```
1  setwd("/Users/yoo/Assignment3")
2  obama <- read.csv("Obama.csv")
3  x <- strptime(obama[,1], format = "%d/%m/%y %H:%M")
4  hist(x, breaks = "days", main="Histogram for occurrence of Barack Obama in
5  Facebook posts by Day", xlab = "Date of the post", ylab = "Density of occurrences")
6
```

setwd("/Users/yoo/Assignment3")
obama <- read.csv("Obama.csv")
x <- strptime(obama[,1], format = "%d/%m/%y %H:%M")
hist(x, breaks = "days", main="Histogram for occurrence of Barack Obama in
Facebook posts by Day", xlab = "Date of the post", ylab = "Density of occurrences")



Here I specified my breaks to be “days”, so the x-axis is split by each day. However I could set them to “weeks” or “years” to visualize the data in the long-term.
This is what happens when I use the following script:

```
setwd("/Users/yoo/Assignment3")
obama <- read.csv("Obama.csv")
x <- strptime(obama[,1], format = "%d/%m/%y %H:%M")
hist(x, breaks = "years", main="Histogram for occurrence of Barack Obama in Facebook posts by Year", xlab = "Date of the post", ylab = "Density of occurrences")
```



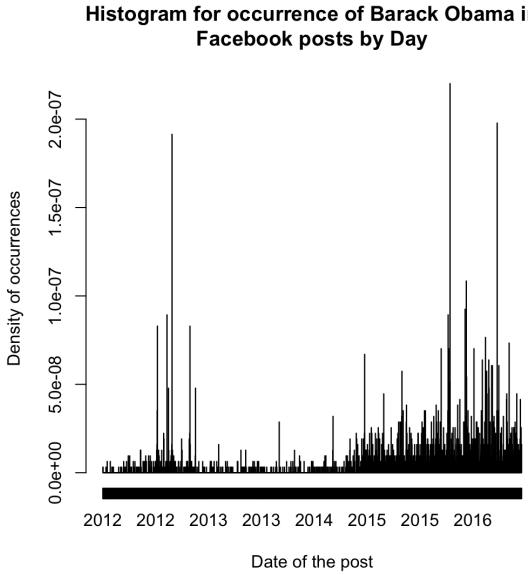
ii)

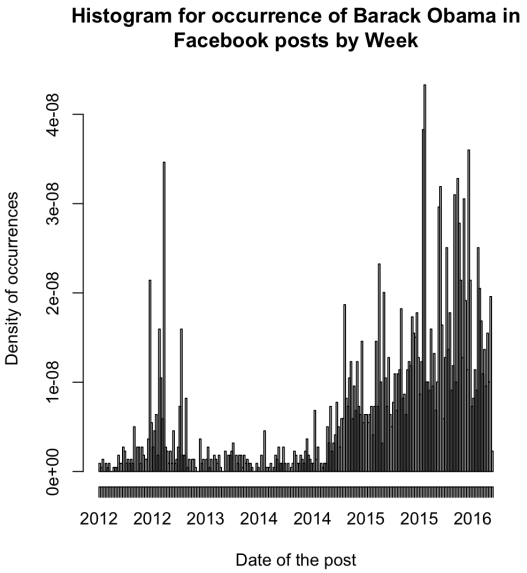
I would like to analyze the graphs obtained.
For better visualization, I have added histograms with breaks set on weeks and months as well.

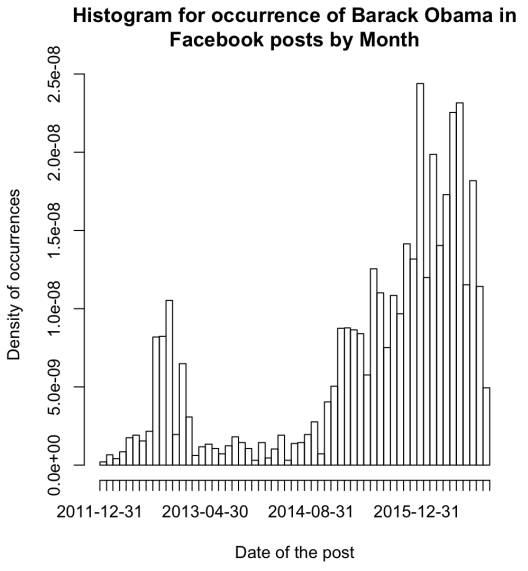
```
setwd("/Users/yoo/Assignment3")
```

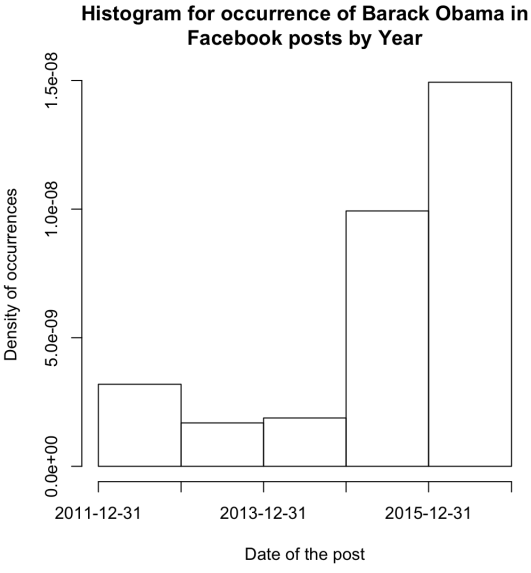
```
obama <- read.csv("Obama.csv")
x <- strptime(obama[,1], format = "%d/%m/%y %H:%M")
hist(x, breaks = "weeks", main="Histogram for occurrence of Barack Obama in
Facebook posts by Week", xlab = "Date of the post", ylab = "Density of occurrences")
// for weekly histogram

setwd("/Users/yoo/Assignment3")
obama <- read.csv("Obama.csv")
x <- strptime(obama[,1], format = "%d/%m/%y %H:%M")
hist(x, breaks = "months", main="Histogram for occurrence of Barack Obama in
Facebook posts by Month", xlab = "Date of the post", ylab = "Density of occurrences")
// for monthly histogram
```









What I could understand from them is that it has a bit of an unusual shape. From the yearly histogram and weekly histogram(clearer label of years), it can be seen that 2015 has a high number of posts relating to Obama and 2016 even tops that recording the highest number of occurrences. This can be related to multiple events that took place during that period.
I have paraphrased some events from the article below:
https://www.rollingstone.com/politics/politics-lists/2015-the-year-obama-stopped-giving-any-f-ks-40439/october-1-2015-gets-serious-about-gun-violence-178295/

From around mid-2015 to the beginning of 2016, he has started to make some actions during his presidency. On July 16 he became the first sitting president to visit the federal prison, commuting the sentence of 46 criminals that he thought were over-punished.

July 26 is the date he visited Kenya, making fun of conspiracy theorists doubting Obama's actual place of birth.
October 2015, the internet has gotten controversial after a tragic mass shooting took place in Oregon. Obama has been claiming to make gun control laws more strict from before, but this event has pushed him further into disappointment and anger to take an action to reduce innocent casualties like this.

One of the events that took place in the early part of next year, 2016 are when he visited Cuba as the first president to visit since 1928.
Late-2016 has high occurrences in the graph as well, since this is a period of time where the presidency will be passed on to either Hillary Clinton or Donald Trump by next year.

To add some data from
https://biography.yourdictionary.com/answers/timelines/barack-obama-timeline.html ,
During 2015, June 25 is the date he won in supreme court on his Affordable Care Act, protecting the majority's healthcare for citizens. On the next day June 26, Obama has claimed another victory on the supreme court making same-sex marriage legal in United States.

These major events during his presidency does support why it has such a high frequency in Medias during that time period.
Although things look low overall during 2012-2014, if you look at the histogram by Day, it does have a huge number of frequencies for a few specific days. We can assume that they are corresponding to when Obama has claimed his second term as president beating Mitt Romney on November 6, and when Obama vowed to fight against gun crimes in December after a disastrous event in Newton.

I believe that despite its weird shape, the histogram does represent the frequency of "Barack Obama" in the Facebook posts well, and they do correspond to respective large events that take place on the similar period of time.

2)

i) Recall from previous tasks:

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

I will filter abc-news for column 1(page_name) which shows the source of media for that article, and will be focusing on column 8(post_type) to see for event, link, photo, status and video. As I will be looking at column 11(comments_count) to compare engagement, I would like to extract this column as well.
I can directly create separate csv file for event, link photo, etc, but I can just filter them on Rstudio.

```
awk -F, '$1 ~ /abc-news/' FB_Dataset | awk -F "\"*,\"*" '{print $8,$11}' > comments.csv
```

// I have reused the framework of previous commands as I am sure they work. I search for rows containing abc-news, and then print columns containing post_type and comments_count, to output into a csv file comments.csv.
I then used
`less comments.csv` // to see in the bash to make sure it worked correctly

Now I am going to take this file into Rstudio.
I am using read.table because they are *space* separated, not *comma* separated.

```
setwd("/Users/yoo/Assignment3")
data <- read.table("comments.csv", header = F)  // set header to false
```

names(data) <- c("post_type", "comments_count") // I set names of the column
head(data) // show first few rows

```
> names(data) <- c("post_type", "comments_count")
> head(data)
  post_type comments_count
1      link             27
2      link            523
3      link             31
4      link            188
5      link             51
6      link             52
>
```

It correctly imports and shows data in the console
I learnt usage of subset() here: https://www.statmethods.net/management/subset.html

setwd("/Users/yoo/Assignment3") // set working directory
data <- read.table("comments.csv", header = F) // read table into variable data
names(data) <- c("post_type", "comments_count") // set column names

event <- subset(data, post_type=="event") // set subset of event, extract rows only where post_type == “event”
link <- subset(data, post_type=="link")
photo <- subset(data, post_type=="photo")
status <- subset(data, post_type=="status")
video <- subset(data, post_type=="video")
head(event) // I use head to make sure headers are set right, and data makes sense
head(link)
head(photo)
head(status)
head(video)

```
> video <- subset(data, post_type=="video")
> head(event)
      post_type comments_count
22798     event            149
> head(link)
  post_type comments_count
1      link             27
2      link            523
3      link             31
4      link            188
5      link             51
6      link             52
> head(photo)
     post_type comments_count
115      photo             71
273      photo             30
277      photo            305
301      photo            270
316      photo             92
724      photo            372
> head(status)
     post_type comments_count
256     status            229
844     status             79
1735    status            227
1878    status             22
2104    status             19
2481    status            121
> head(video)
    post_type comments_count
33      video            134
50      video             67
84      video            385
86      video            364
145     video            198
164     video             86
>
```

Take note that:
***event* only has 1 row.**

Now I will plot a **boxplot for the number of comments against different types of posts, made by abc-news on Facebook.**

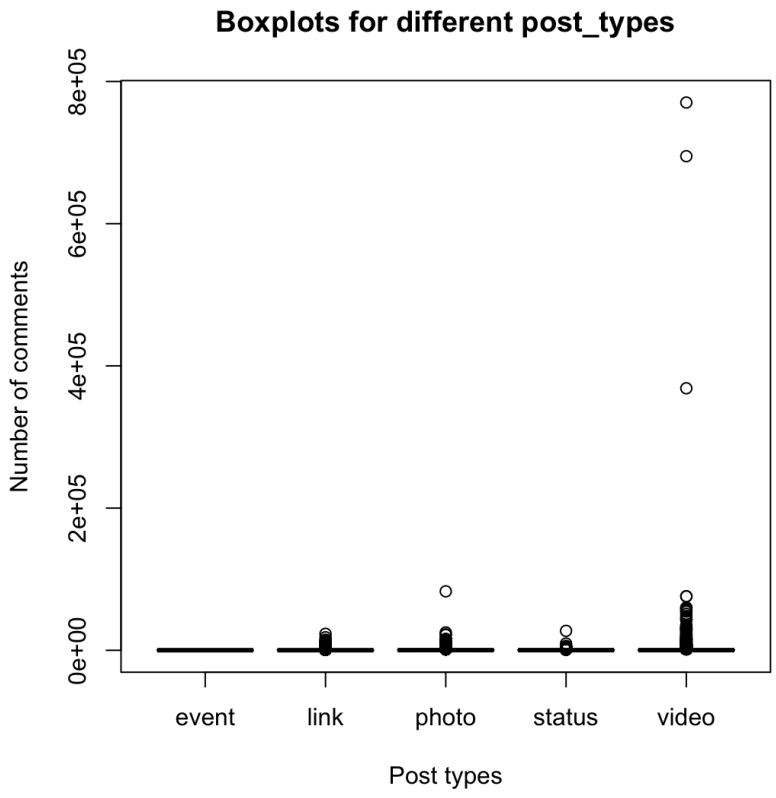
setwd("/Users/yoo/Assignment3")
data <- read.table("comments.csv", header = F)
names(data) <- c("post_type", "comments_count")

event <- subset(data, post_type=="event")

```
link <- subset(data, post_type=="link")
photo <- subset(data, post_type=="photo")
status <- subset(data, post_type=="status")
video <- subset(data, post_type=="video")

head(video[,2])
boxplot(event[,2], link[,2], photo[,2], status[,2], video[,2],
        main = "Boxplots for different post_types",
        names = c("event", "link", "photo", "status", "video"),
        xlab = "Post types", ylab = "Number of comments")
```

// boxplot() is used. Main to set the title of the plot, names to set the names for 5 different post types, xlab for x label, and ylab for y labels.
The reason I am using [,2] is to show that I am only interested in the second column(comments_count) for each data. Because dataframe in R is denoted by [row, column], [, 2] means **take all rows but only 2nd column.**



I can see from this box plot that there are few observable outliers which are too large that they are affecting the ylabel of the plot.

Although the median looks almost the same for all 5, it is because the y-axis is scaled that way that they look similar. In order to analyze further, we have to remove outliers. A post type of video records the top 3 highest number of comments, ranging from 300000-800000 comments. This is probably because videos are more likely to go viral, going through language barriers. Similar reason for that one outlier for photo at around 100000 comments.

We can't really see medians or interquartile ranges, but we know that there are many outstanding datasets in photo, link, status and video due to the number of white outliers forming seemingly black dots. On the other hand, we know from head(event) above, it only contains one element. Although data for *event* might not be enough to draw trustworthy analysis, this plot does make sense as there are no outliers for *event*.

ii)

I will modify the previous script by filtering out outliers that records comments_count greater than 1000.

I have referred to: https://dplyr.tidyverse.org/reference/filter.html

To produce a code that filters out outliers. I could just use filter(), but I prefer subset as I am more familiar with it.

```
data <- subset(data, comments_count <= 1000)
```

This is the working code that reassigns data with filtered data frames by using subset()

.

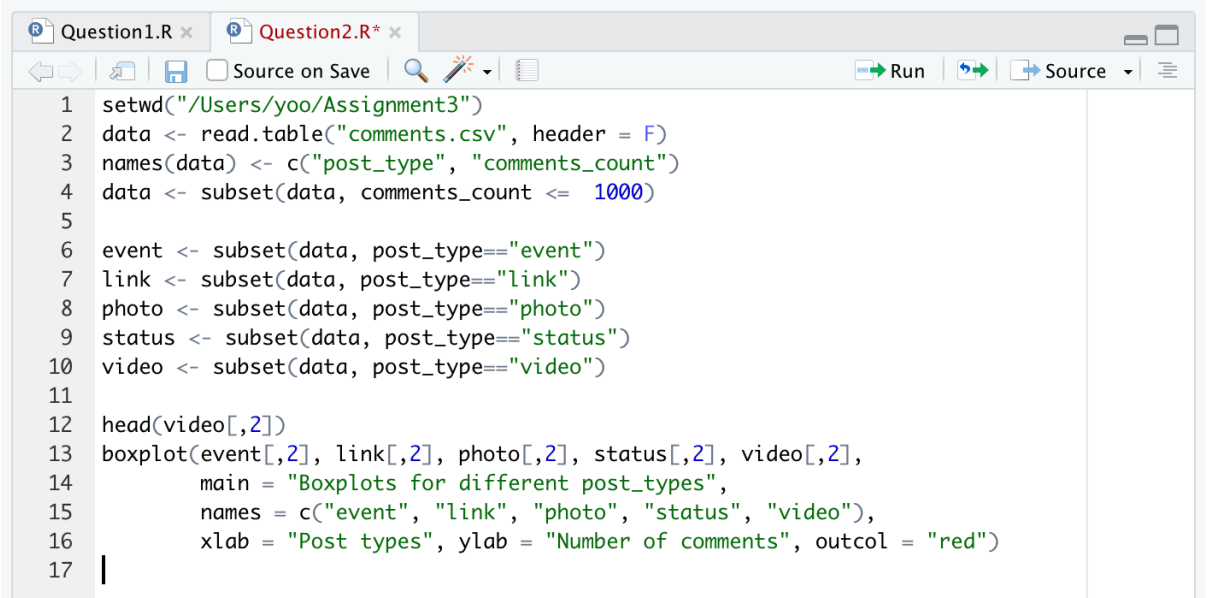
Putting them together:

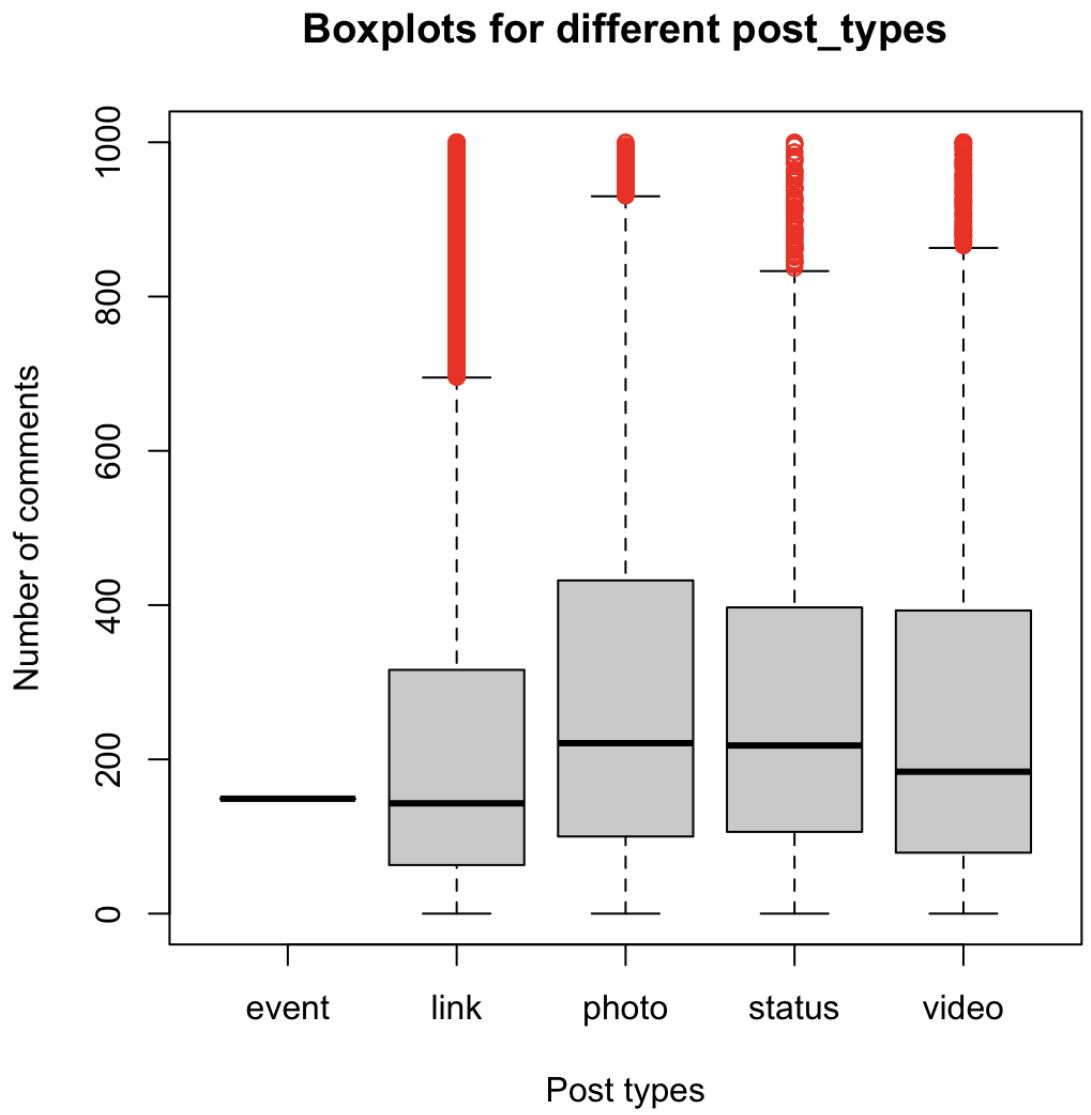
```
setwd("/Users/yoo/Assignment3")
data <- read.table("comments.csv", header = F)
names(data) <- c("post_type", "comments_count")
data <- subset(data, comments_count <=  1000)

event <- subset(data, post_type=="event")
link <- subset(data, post_type=="link")
photo <- subset(data, post_type=="photo")
status <- subset(data, post_type=="status")
video <- subset(data, post_type=="video")

head(video[,2])
boxplot(event[,2], link[,2], photo[,2], status[,2], video[,2],
        main = "Boxplots for different post_types",
        names = c("event", "link", "photo", "status", "video"),
        xlab = "Post types", ylab = "Number of comments", outcol = “red”)
```

I have added outcol = “red”, to change the color of the outlier from black to red for better visualization.





iii)

The black bold in the middle of the rectangle represents the median, while tips of that rectangle represent interquartile range. Tips of the dotted lines outside the box show maximum and minimum values.

The reason the *event* does not have a box is because there was only one data that had a post_type of event.

Photo and status has higher median among all, but photo has slightly more.

Therefore, for ABC News, posting photos has on average been most effective in attracting engagements on Facebook.

Conclusion:

This concludes the end of my assignment 3. During my journey in FIT1043, I was introduced to many aspects of data science with a wide range of softwares and techniques.
The factor that surprised me the most was that there exists gigantic data such that even Python pandas are not able to process. While the terminal does look very intimidating at first, if I played around it it was very fun to work with. Even with RStudio, there are many functions developed by many talented programmers in the R community. I know I could use filter(), but somehow I just prefer subset(), which is a fun part of wrangling with data - many different ways to do tons of things.

References:

hek2mgl. (2017, May 23). Print a row only if the string exist in a specific column. Stack Overflow. Retrieved from https://stackoverflow.com/a/31038104.

Dmitry. (2013, January 24). How can I count the occurrences of a string within a file?. Stack Overflow. Retrieved from https://stackoverflow.com/a/14510665.

Jeff Schaller. (2017, October 16). Counting occurrences of word in text file. Stack Exchange. Retrieved from https://unix.stackexchange.com/a/398414.

dessert. (2017, December 18). How to extract only values greater than a threshold from a file?. Ask Ubuntu. Retrieved from https://askubuntu.com/a/987515.

Matt Ryall. (2011, June 22). Sorting data based on second column of a file. Stack Overflow. Retrieved from https://stackoverflow.com/a/6438940.

Wildcard. (2015, November 14). Using awk to sum the values of a column, based on the values of another column. Stack Exchange. Retrieved from https://unix.stackexchange.com/a/242949.

synthesizerpatel. (2013, October 26). How to extract one column of a csv file. Stack Overflow. Retrieved from https://stackoverflow.com/a/19602188.

strptime: Date-time Conversion Functions to and from Character. (n.d.). Retrieved from https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/strptime.

Tessa Stuart. (2015, December 22). 2015: The Year Obama Stopped Giving Any F–ks. Rolling Stone. Retrieved from https://www.rollingstone.com/politics/politics-lists/2015-the-year-obama-stopped-giving-any-f-ks-40439/.

Barack Obama Timeline. (n.d.). Barack Obama Timeline. Retrieved from https://biography.yourdictionary.com/answers/timelines/barack-obama-timeline.html.

Robert I. Kabacoff, Ph.D. (2017). Subsetting Data. Retrieved from https://www.statmethods.net/management/subset.html.

Subset rows using column values. (n.d.). Retrieved from https://dplyr.tidyverse.org/reference/filter.html.

Thank you for grading!

32134541 Yo Kogure

END OF ASSIGNMENT 3