FIT1043 Assignment 3

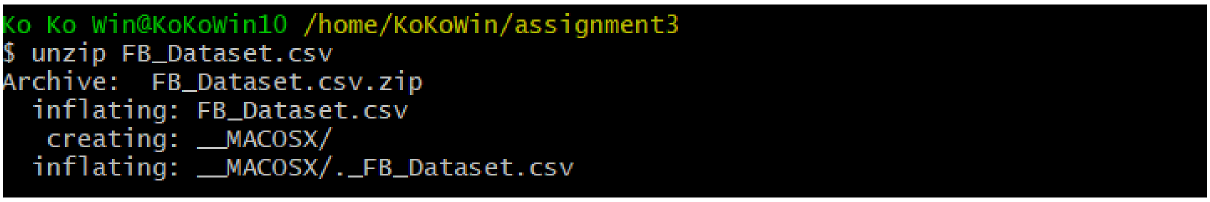
Ko Ko Win (31842305)

Remark:

All the codes are written in this font "`this is the font`". Follow by the screenshot of the code and the explanation below the screenshot.

Question 1:

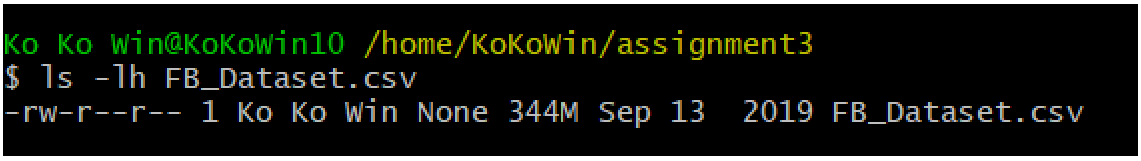
```
unzip FB_Dataset.csv
```



```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ unzip FB_Dataset.csv
Archive:  FB_Dataset.csv.zip
  inflating: FB_Dataset.csv
   creating: __MACOSX/
  inflating: __MACOSX/._FB_Dataset.csv
```

The code is used to uncompressed the zip file.

```
ls -lh FB_Dataset.csv
```

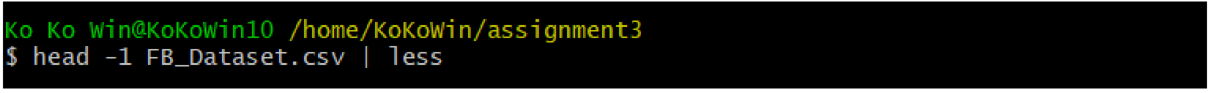


```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ ls -lh FB_Dataset.csv
-rw-r--r-- 1 Ko Ko Win None 344M Sep 13  2019 FB_Dataset.csv
```

The size of the dataset is ***344M*** which stands for 344 Megabyte. **ls** is responsible for listing the files contains in current directory.

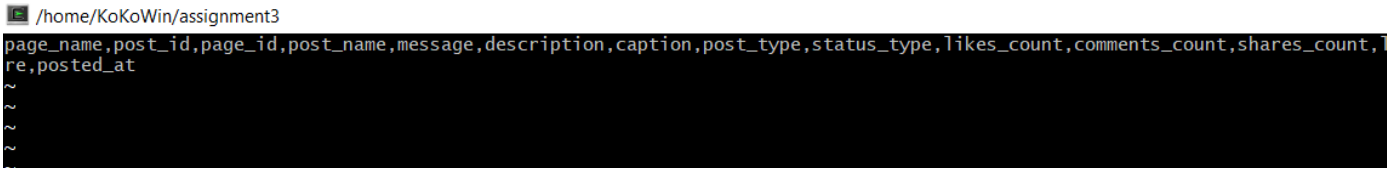
Question 2:

```
head -1 FB_Dataset.csv | less
```



```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ head -1 FB_Dataset.csv | less
```

The code ***head -1*** will print out the first column of our dataset which will be our name of column to check the delimiter that is used to separate the column. By calling less it will display the content from the file. **|** is called a pipe and it acts like a water pipe which allows the water to flow but in this case if we are using two or more commands we can use the pipe.

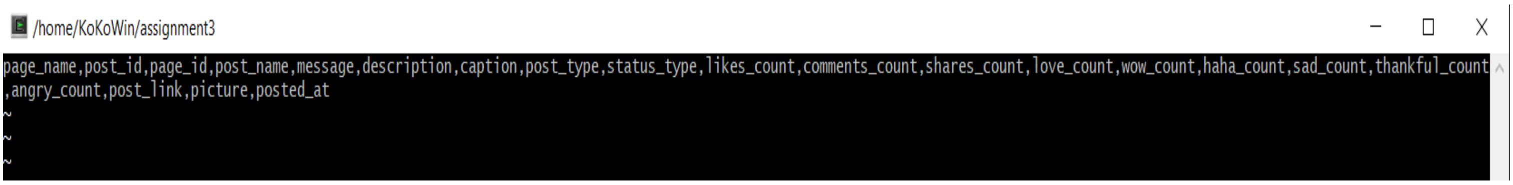


/home/KoKoWin/assignment3

```
page_name,post_id,page_id,post_name,message,description,caption,post_type,status_type,likes_count,comments_count,shares_count,l
re,posted_at
~
~
~
~
```

As we can see **‘,’** is used to separate the columns.

```
head -1 FB_Dataset.csv | less
```



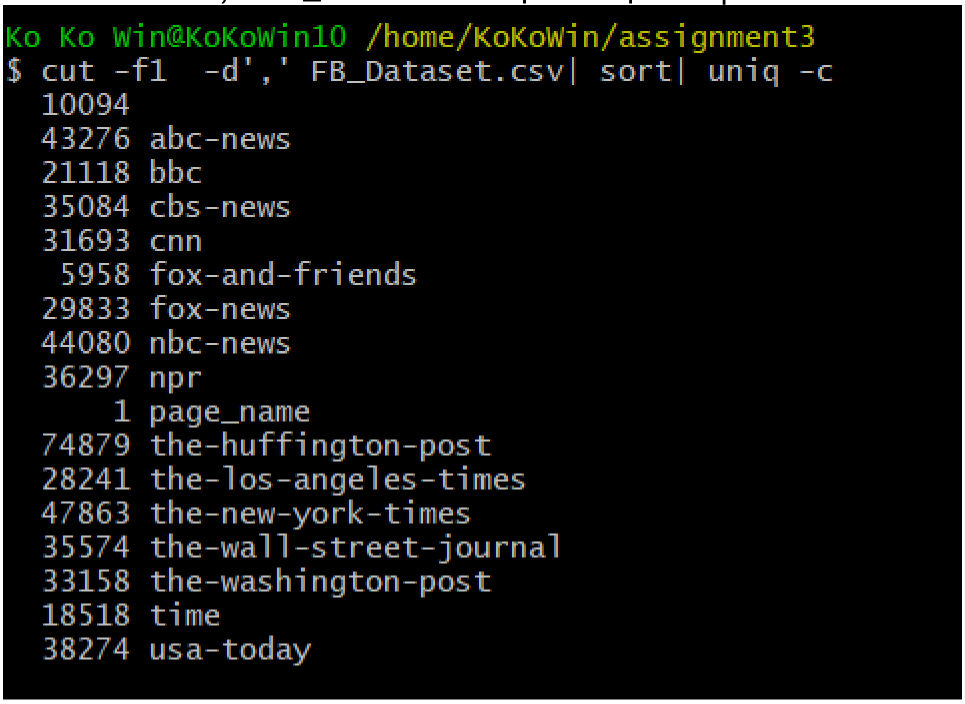
/home/KoKoWin/assignment3

```
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
~
~
~
```

The same code can be used to check the name of other columns .

Question 3:

```
cut -f1  -d',' FB_Dataset.csv| sort| uniq -c
```

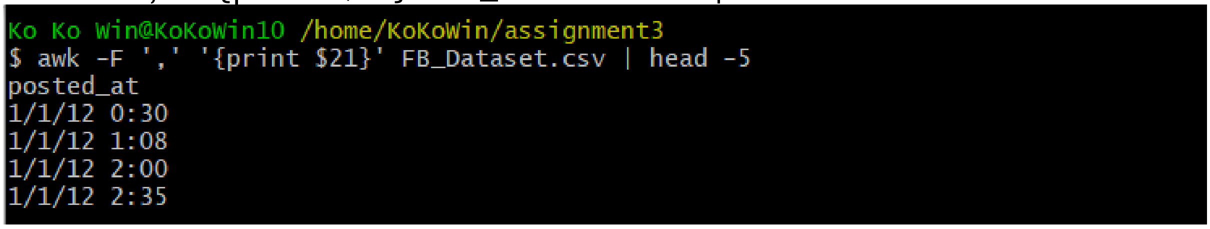


```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ cut -f1  -d',' FB_Dataset.csv| sort| uniq -c
  10094
  43276 abc-news
  21118 bbc
  35084 cbs-news
  31693 cnn
   5958 fox-and-friends
  29833 fox-news
  44080 nbc-news
  36297 npr
      1 page_name
  74879 the-huffington-post
  28241 the-los-angeles-times
  47863 the-new-york-times
  35574 the-wall-street-journal
  33158 the-washington-post
  18518 time
  38274 usa-today
```

cut is a command that allows you to cut the part of the line or field from the file. In this case we will cut the first column which is page_name. **-d’,’** is used to specify the delimiter that is used to separate the file

Question 4:

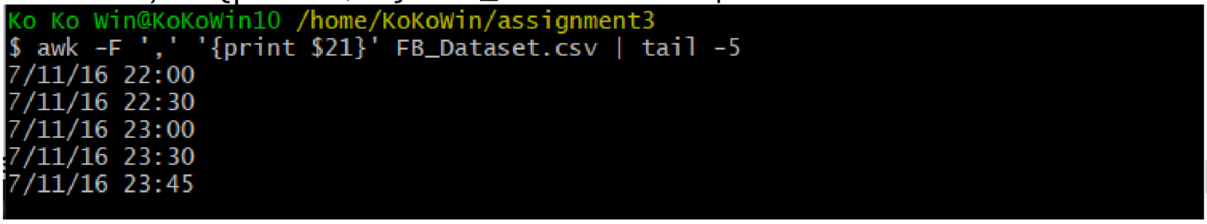
```
awk -F ',' '{print $21}' FB_Dataset.csv | head -5
```



```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ awk -F ',' '{print $21}' FB_Dataset.csv | head -5
posted_at
1/1/12 0:30
1/1/12 1:08
1/1/12 2:00
1/1/12 2:35
```

awk is a function that can process a file one line at the time and after that ***-F ‘,’*** is used to identify the delimiter and in the code the function ***head -5*** will print out the first five lines of column number 21 which is a date that is posted.

```
awk -F ',' '{print $21}' FB_Dataset.csv | tail -5
```

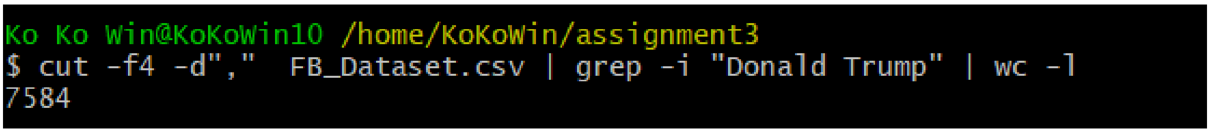


```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ awk -F ',' '{print $21}' FB_Dataset.csv | tail -5
7/11/16 22:00
7/11/16 22:30
7/11/16 23:00
7/11/16 23:30
7/11/16 23:45
```

In the above code by calling ***tail -5*** last five lines of the column number 21 will be printed which is a date that is posted. From above two lines of code we can conclude that date range for Facebook posts are 1/1/12 to 7/11/16.

Question 5:

```
cut -f4 -d","  FB_Dataset.csv | grep -i "Donald Trump" | wc -l
```

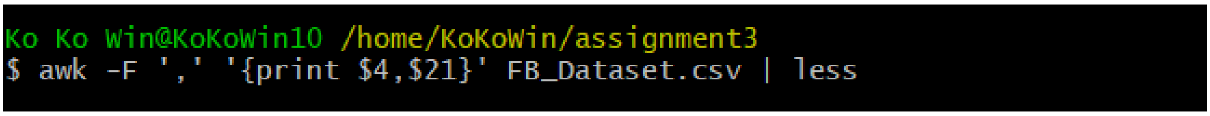


```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ cut -f4 -d","  FB_Dataset.csv | grep -i "Donald Trump" | wc -l
7584
```

cut -f4 -d”,” FB_Dataset.csv will select the 4th column of our dataset which is post_name and we specify the delimiter is the comma. ***grep -i*** is used for extracting the word “Donald

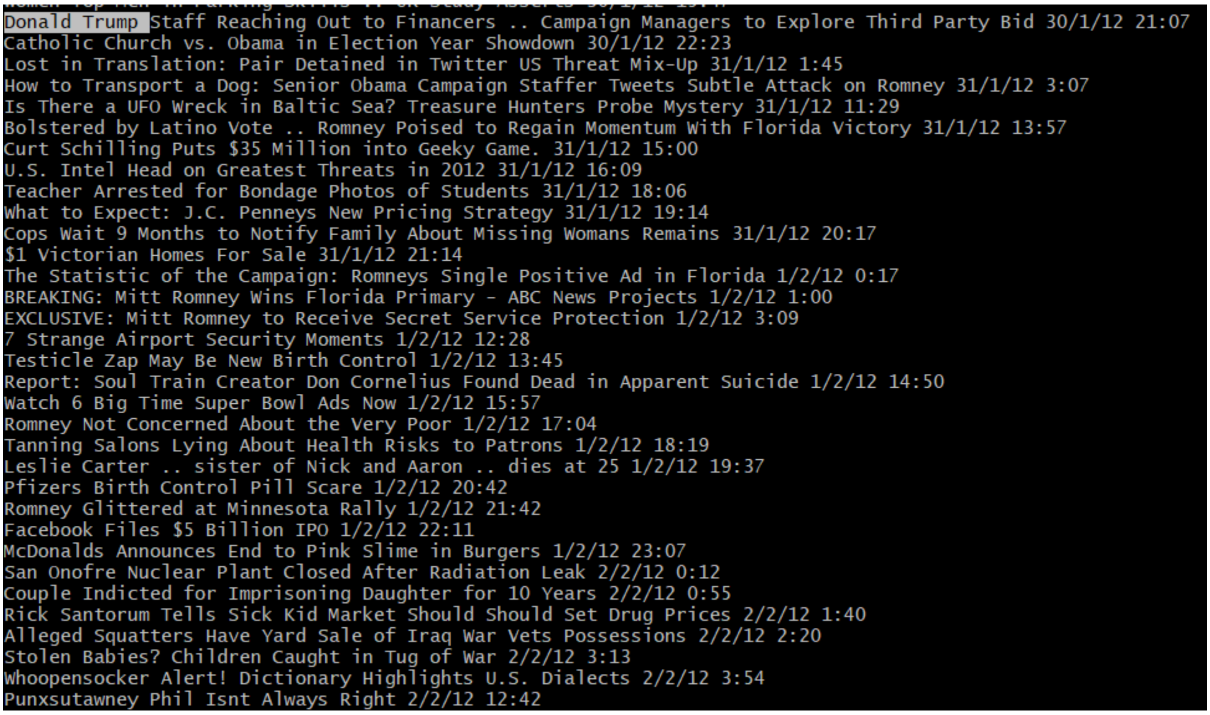
Trump” by ignoring the case. ***wc -l*** function will tell the ***wc*** which is stands for word count to count the number of lines. The word “Donald Trump” occurs 7584 times in the content of post names.

```
awk -F ',' '{print $4,$21}' FB_Dataset.csv | less
```



```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ awk -F ',' '{print $4,$21}' FB_Dataset.csv | less
```

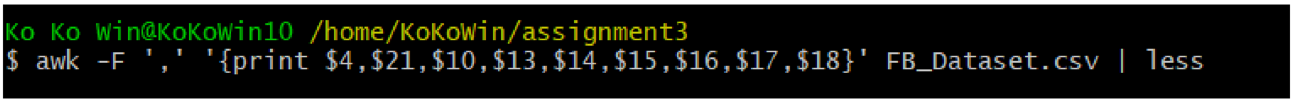
awk -F ',' '{print $4,$21}' FB_Dataset.csv will print out the 4th column and 21st column which is our post names and date posted by specifying the delimiter which columns are separated and by calling ***less*** it will load the CSV file.



```
Donald Trump Staff Reaching Out to Financers .. Campaign Managers to Explore Third Party Bid 30/1/12 21:07
Catholic Church vs. Obama in Election Year Showdown 30/1/12 22:23
Lost in Translation: Pair Detained in Twitter US Threat Mix-Up 31/1/12 1:45
How to Transport a Dog: Senior Obama Campaign Staffer Tweets Subtle Attack on Romney 31/1/12 3:07
Is There a UFO Wreck in Baltic Sea? Treasure Hunters Probe Mystery 31/1/12 11:29
Bolstered by Latino Vote .. Romney Poised to Regain Momentum With Florida Victory 31/1/12 13:57
Curt Schilling Puts $35 Million into Geeky Game. 31/1/12 15:00
U.S. Intel Head on Greatest Threats in 2012 31/1/12 16:09
Teacher Arrested for Bondage Photos of Students 31/1/12 18:06
What to Expect: J.C. Penneys New Pricing Strategy 31/1/12 19:14
Cops Wait 9 Months to Notify Family About Missing Womans Remains 31/1/12 20:17
$1 Victorian Homes For Sale 31/1/12 21:14
The Statistic of the Campaign: Romneys Single Positive Ad in Florida 1/2/12 0:17
BREAKING: Mitt Romney Wins Florida Primary - ABC News Projects 1/2/12 1:00
EXCLUSIVE: Mitt Romney to Receive Secret Service Protection 1/2/12 3:09
7 Strange Airport Security Moments 1/2/12 12:28
Testicle Zap May Be New Birth Control 1/2/12 13:45
Report: Soul Train Creator Don Cornelius Found Dead in Apparent Suicide 1/2/12 14:50
Watch 6 Big Time Super Bowl Ads Now 1/2/12 15:57
Romney Not Concerned About the Very Poor 1/2/12 17:04
Tanning Salons Lying About Health Risks to Patrons 1/2/12 18:19
Leslie Carter .. sister of Nick and Aaron .. dies at 25 1/2/12 19:37
Pfizers Birth Control Pill Scare 1/2/12 20:42
Romney Glittered at Minnesota Rally 1/2/12 21:42
Facebook Files $5 Billion IPO 1/2/12 22:11
McDonalds Announces End to Pink Slime in Burgers 1/2/12 23:07
San Onofre Nuclear Plant Closed After Radiation Leak 2/2/12 0:12
Couple Indicted for Imprisoning Daughter for 10 Years 2/2/12 0:55
Rick Santorum Tells Sick Kid Market Should Should Set Drug Prices 2/2/12 1:40
Alleged Squatters Have Yard Sale of Iraq War Vets Possessions 2/2/12 2:20
Stolen Babies? Children Caught in Tug of War 2/2/12 3:13
Whoopensocker Alert! Dictionary Highlights U.S. Dialects 2/2/12 3:54
Punxsutawney Phil Isnt Always Right 2/2/12 12:42
```

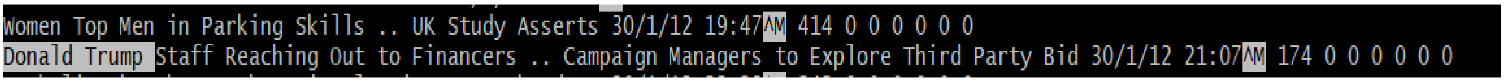
That is the first mention of the word “Donald Trump” on 30/1/12 and the name of the post is “Donald Trump Staff Reaching Out to Financers .. Campaign Managers to Explore Third Party Bid”.

```
awk -F ',' '{print $4,$21,$10,$13,$14,$15,$16,$17,$18}'
FB_Dataset.csv | less
```



```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ awk -F ',' '{print $4,$21,$10,$13,$14,$15,$16,$17,$18}' FB_Dataset.csv | less
```

To the previous code I have added `$10,$13,$14,$15,$16,$17,$18` which are columns for likes_count, love_count, wow_count, haha_count, sad_count, thankful_count, angry_count respectively.

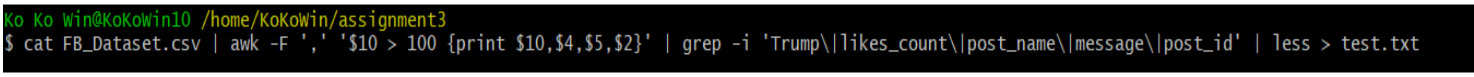


```
Women Top Men in Parking Skills .. UK Study Asserts 30/1/12 19:47^M 414 0 0 0 0 0 0
Donald Trump Staff Reaching Out to Financers .. Campaign Managers to Explore Third Party Bid 30/1/12 21:07^M 174 0 0 0 0 0 0
```

As we can see from the output above for that particular post there are 174 likes_count, 0 love_count, 0 wow_count, 0 haha_count, 0 sad_count, 0 thankful_count, 0 angry_count. From my opinion I will take likes,love,haha,thankful and wow as a positive reaction to the post. Overall, people's reaction to this post was positive as the amount of likes dominates other reacts. Moreover, I take angry_count as a reaction against the post so we can say that there is 0 reaction against this post.

Question 6:

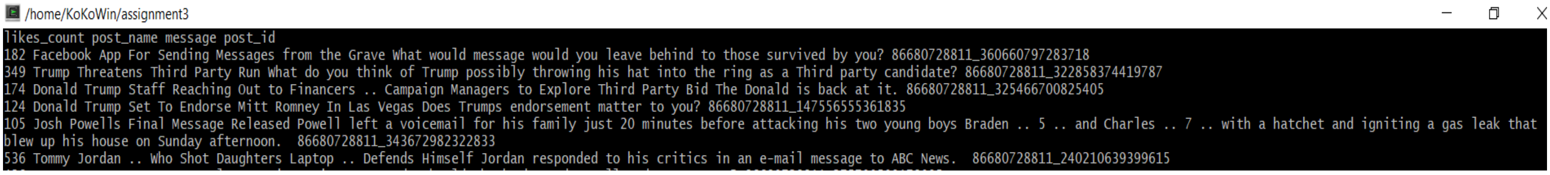
```
cat FB_Dataset.csv | awk -F ',' '$10 > 100 {print $10,$4,$5,$2}' | grep -i
'Trump\|likes_count\|post_name\|message\|post_id' | less > test.txt
```



```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ cat FB_Dataset.csv | awk -F ',' '$10 > 100 {print $10,$4,$5,$2}' | grep -i 'Trump\|likes_count\|post_name\|message\|post_id' | less > test.txt
```

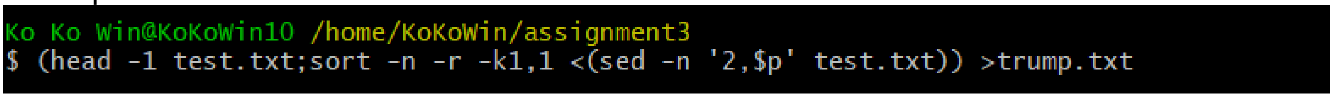
cat function is used to concatenate files together and allow us to view the content of the files. ***awk*** function is used when dealing with the columns and fields so inside that i ***used -F*** to specify the delimiter which is ***","*** furthermore ***$10 > 100*** will only select the post with more than 100 likes. After that I printed out ***$10, $4,$5, $2*** which are columns for likes_count ,post_name,message and post_id respectively. ***grep -i*** function is used to only select certain string from our dataset and ***'\|'*** is used so we can select multiple strings and select the word Trump in post_name and message column as well as the column headings. ***>*** function can allow us to save it to text file name 'test.txt' .

Output:



```
likes_count post_name message post_id
182 Facebook App For Sending Messages from the Grave What would message would you leave behind to those survived by you? 86680728811_360660797283718
349 Trump Threatens Third Party Run What do you think of Trump possibly throwing his hat into the ring as a Third party candidate? 86680728811_322858374419787
174 Donald Trump Staff Reaching Out to Financers .. Campaign Managers to Explore Third Party Bid The Donald is back at it. 86680728811_325466700825405
124 Donald Trump Set To Endorse Mitt Romney In Las Vegas Does Trumps endorsement matter to you? 86680728811_147556555361835
105 Josh Powells Final Message Released Powell left a voicemail for his family just 20 minutes before attacking his two young boys Braden .. 5 .. and Charles .. 7 .. with a hatchet and igniting a gas leak that
blew up his house on Sunday afternoon.  86680728811_343672982322833
536 Tommy Jordan .. Who Shot Daughters Laptop .. Defends Himself Jordan responded to his critics in an e-mail message to ABC News.  86680728811_240210639399615
```

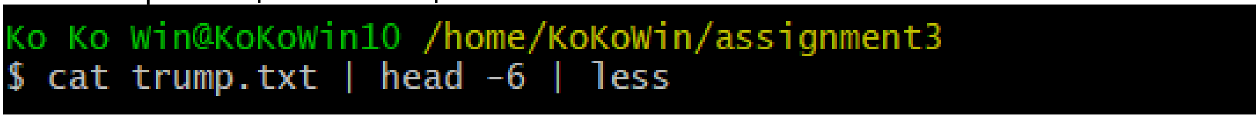
```
(head -1 test.txt;sort -n -r -k1,1 <(sed -n '2,$p' test.txt))
>trump.txt
```



```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ (head -1 test.txt;sort -n -r -k1,1 <(sed -n '2,$p' test.txt)) >trump.txt
```

The code above is used for sorting our lkes_count column in a descending order. ***head -1*** test.txt is our column names in our text files which are likes_count, post_name, post_id. ***sort -n -r -k1,1*** function will sort the numerical value in a reverse order so we will get it in a descending order for number of likes and -***k1,1*** function is to tell the program to sort the first column only which is our number of likes. Furthermore, ***sed -n '2,$p' test.txt*** function is used to tell the program to sort starting from second rows until the last line because we want to leave the column name as it is. Lastly **>** function is used to save it in a text file name ***trump.txt*** which is our final txt file.

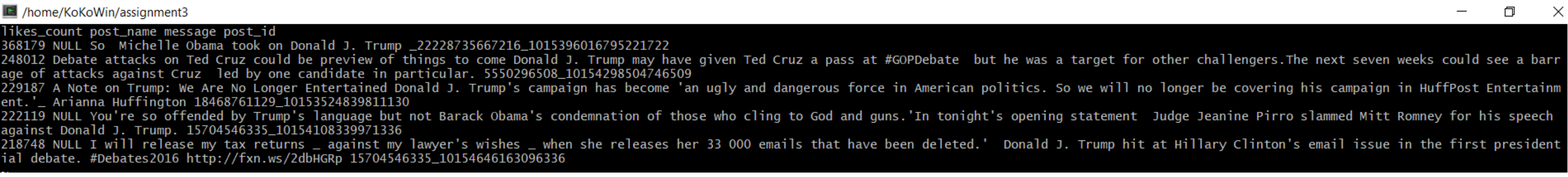
```
cat trump.txt | head -6 | less
```



```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ cat trump.txt | head -6 | less
```

The code above is used to first 6 lines of our text file including the column names.

Output:

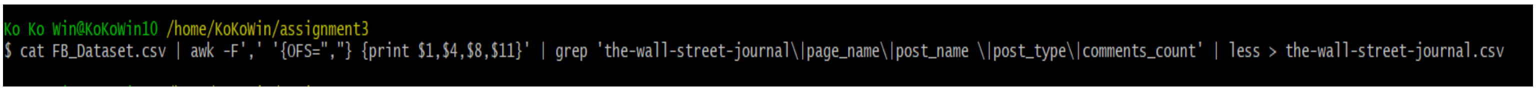


/home/KoKoWin/assignment3

```
likes_count post_name message post_id
368179 NULL So  Michelle Obama took on Donald J. Trump _22228735667216_1015396016795221722
248012 Debate attacks on Ted Cruz could be preview of things to come Donald J. Trump may have given Ted Cruz a pass at #GOPDebate  but he was a target for other challengers.The next seven weeks could see a barr
age of attacks against Cruz  led by one candidate in particular. 5550296508_10154298504746509
229187 A Note on Trump: We Are No Longer Entertained Donald J. Trump's campaign has become 'an ugly and dangerous force in American politics. So we will no longer be covering his campaign in HuffPost Entertainm
ent.'_ Arianna Huffington 18468761129_10153524839811130
222119 NULL You're so offended by Trump's language but not Barack Obama's condemnation of those who cling to God and guns.'In tonight's opening statement  Judge Jeanine Pirro slammed Mitt Romney for his speech
against Donald J. Trump. 15704546335_10154108339971336
218748 NULL I will release my tax returns _ against my lawyer's wishes _ when she releases her 33 000 emails that have been deleted.'  Donald J. Trump hit at Hillary Clinton's email issue in the first president
ial debate. #Debates2016 http://fxn.ws/2dbHGRp 15704546335_10154646163096336
```

Question 7:

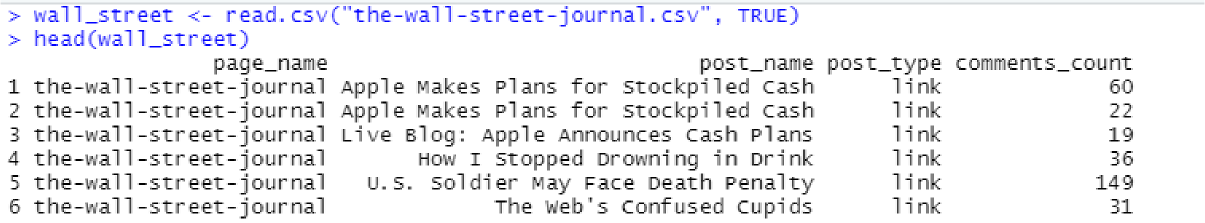
```
cat FB_Dataset.csv | awk -F',' '{OFS=","} {print $1,$4,$8,$11}' |
grep 'the-wall-street-journal\|page_name\|post_name
\|post_type\|comments_count' | less > the-wall-street-journal.csv
```



```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ cat FB_Dataset.csv | awk -F',' '{OFS=","} {print $1,$4,$8,$11}' | grep 'the-wall-street-journal\|page_name\|post_name \|post_type\|comments_count' | less > the-wall-street-journal.csv
```

OFS="," is a function is to concatenates two parameter with a space so I will have the homogeneity with the delimiters . I select columns 1,4,8 and 11 which are page_name,post_name ,post_type,comments_count. Furthermore, I select the post by the wall street journal. Lastly, the **>** function will save it to CSV file name the-wall-street-journal.csv

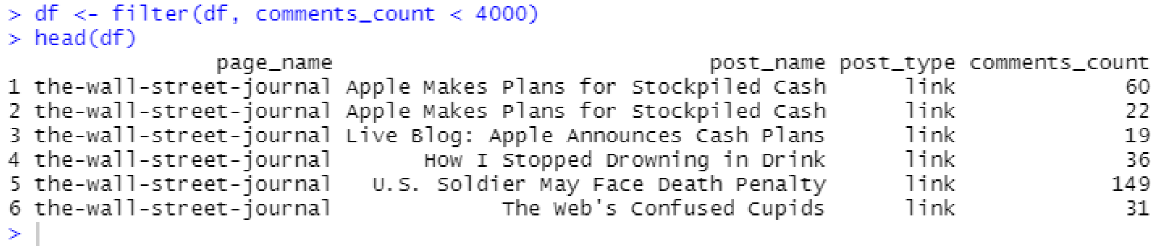
```
wall_street <- read.csv("the-wall-street-journal.csv", TRUE)
head(wall_street)
```



```
> wall_street <- read.csv("the-wall-street-journal.csv", TRUE)
> head(wall_street)
                page_name                             post_name post_type comments_count
1 the-wall-street-journal Apple Makes Plans for Stockpiled Cash      link             60
2 the-wall-street-journal Apple Makes Plans for Stockpiled Cash      link             22
3 the-wall-street-journal Live Blog: Apple Announces Cash Plans      link             19
4 the-wall-street-journal       How I Stopped Drowning in Drink      link             36
5 the-wall-street-journal   U.S. Soldier May Face Death Penalty      link            149
6 the-wall-street-journal             The Web's Confused Cupids      link             31
```

I store all the data in a name called “wall_street” and read the csv file. The parameter TRUE is passed to let the system know that first row is a header. head(wall_street) will print out first 5 rows.

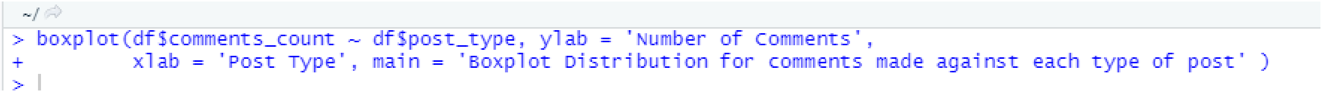
```
library(dplyr)
df <- filter(df, comments_count < 4000)
head(df)
```



```
> df <- filter(df, comments_count < 4000)
> head(df)
                page_name                             post_name post_type comments_count
1 the-wall-street-journal Apple Makes Plans for Stockpiled Cash      link             60
2 the-wall-street-journal Apple Makes Plans for Stockpiled Cash      link             22
3 the-wall-street-journal Live Blog: Apple Announces Cash Plans      link             19
4 the-wall-street-journal       How I Stopped Drowning in Drink      link             36
5 the-wall-street-journal   U.S. Soldier May Face Death Penalty      link            149
6 the-wall-street-journal             The Web's Confused Cupids      link             31
> 
```

First, I imported the library called dplyr which is used for Data Manipulation when working with dataframe. After that, I filter comments less than 4000 and store the data in a name called ‘df’ .

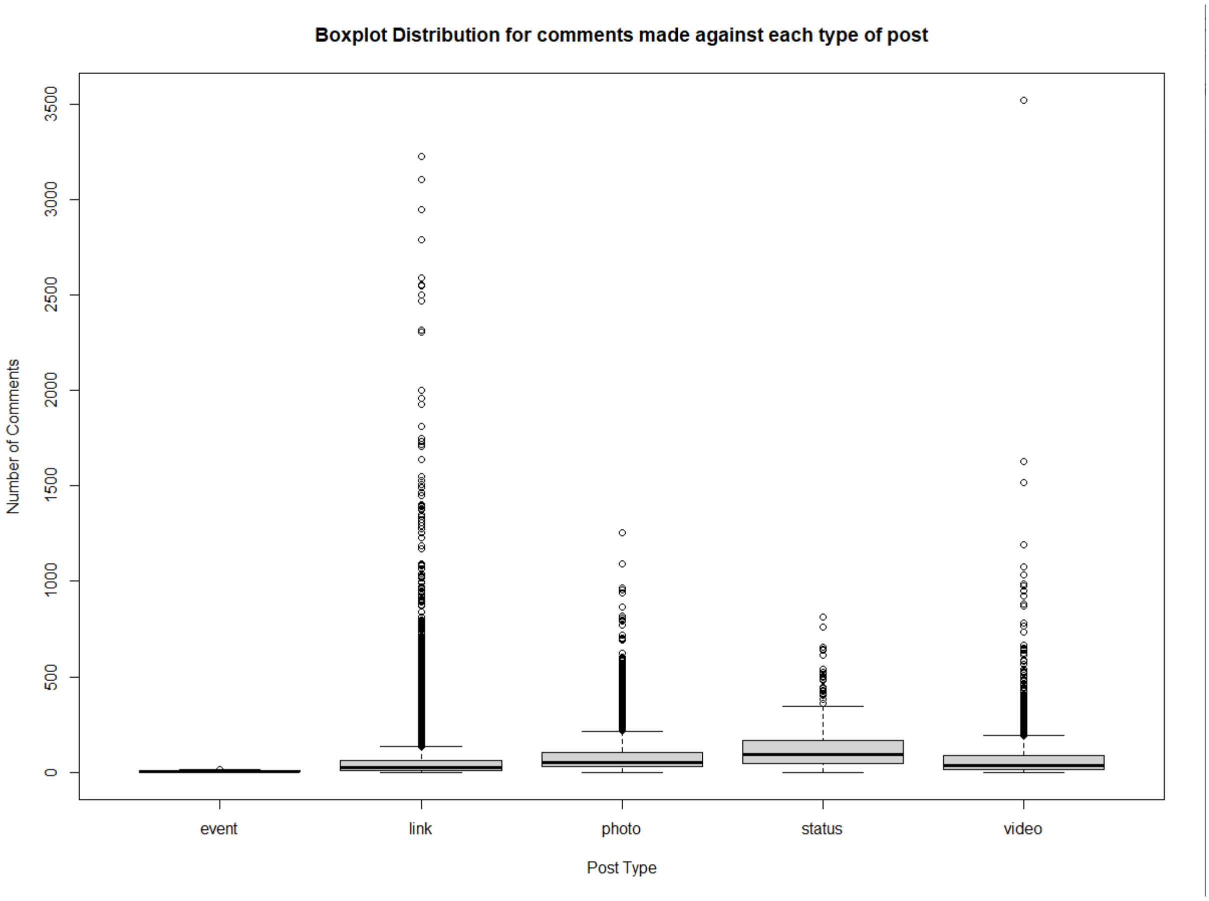
```
boxplot(df$comments_count ~ df$post_type, ylab = 'Number of
Comments',xlab = 'Post Type', main = 'Boxplot Distribution for
comments made against each type of post' )
```



```
~/
> boxplot(df$comments_count ~ df$post_type, ylab = 'Number of Comments',
+         xlab = 'Post Type', main = 'Boxplot Distribution for comments made against each type of post' )
> 
```

The above code is for plotting boxplot for number of comments made against each type of post.

Output:

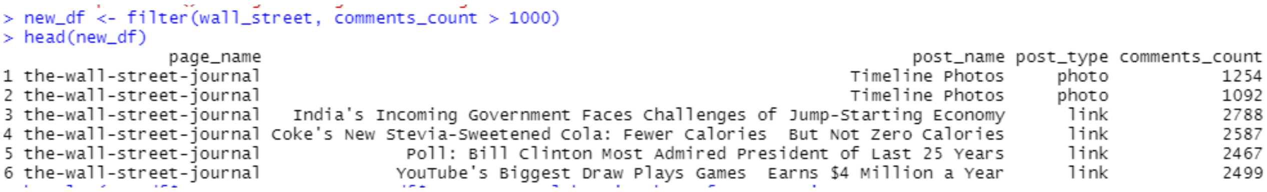


Firstly, from the boxplot we can see that link post type has the most amount of outliers followed by video post type, photo post type, status post type and an event post type. Skewness of majority of box plots are right skew.

Secondly, most engaging type of post type is a status post. It is because the maximum number of comments on the boxplot for status post type is higher than 4 other post types.

Question 8:

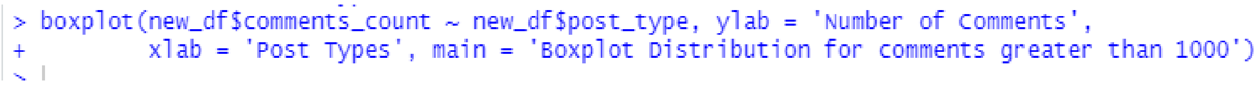
```
new_df <- filter(wall_street, comments_count > 1000)
head(new_df)
```



```
> new_df <- filter(wall_street, comments_count > 1000)
> head(new_df)
                page_name                                                   post_name post_type comments_count
1 the-wall-street-journal                                             Timeline Photos     photo           1254
2 the-wall-street-journal                                             Timeline Photos     photo           1092
3 the-wall-street-journal   India's Incoming Government Faces Challenges of Jump-Starting Economy      link           2788
4 the-wall-street-journal Coke's New Stevia-Sweetened Cola: Fewer Calories  But Not Zero Calories      link           2587
5 the-wall-street-journal           Poll: Bill Clinton Most Admired President of Last 25 Years      link           2467
6 the-wall-street-journal            YouTube's Biggest Draw Plays Games  Earns $4 Million a Year      link           2499
```

I created a new data frame called new_df and stored the data with comments greater than 1000. To view the first 5 rows, I used head function to check.

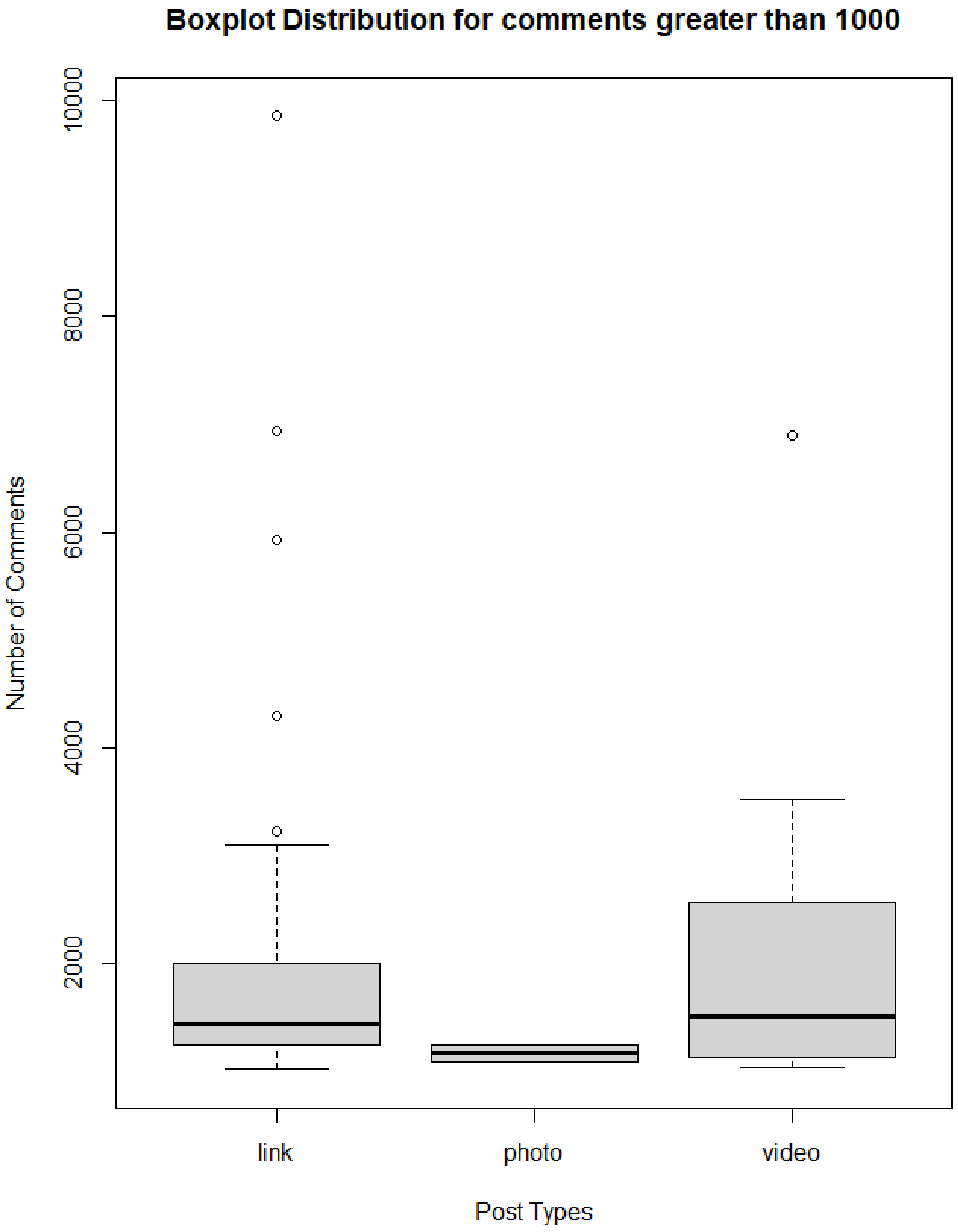
```
boxplot(new_df$comments_count ~ new_df$post_type, ylab = 'Number of
Comments', xlab = 'Post Types', main = 'Boxplot Distribution for
comments greater than 1000')
```



```
> boxplot(new_df$comments_count ~ new_df$post_type, ylab = 'Number of Comments',
+         xlab = 'Post Types', main = 'Boxplot Distribution for comments greater than 1000')
> |
```

I plotted the boxplot for post type with comments greater than 1000 by giving appropriate name to the x and y label as well as the title.

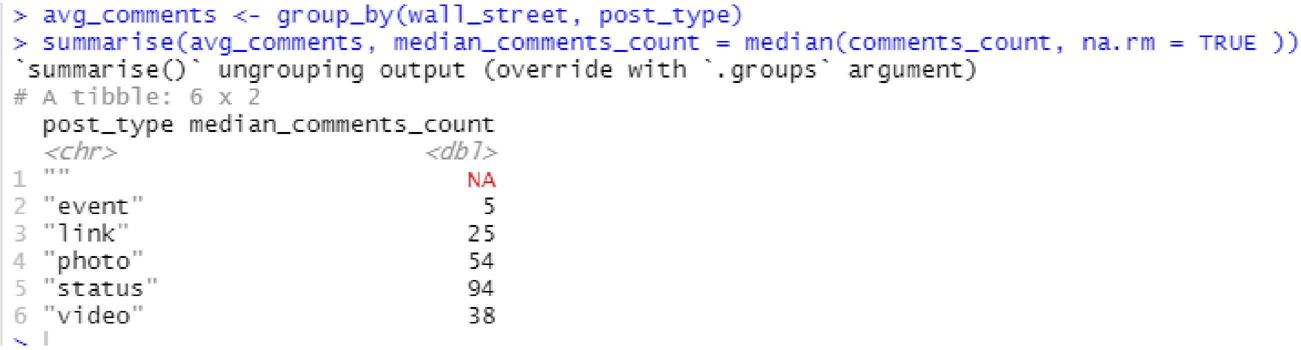
Output:



From the boxplot we can see that link, photo and video type of post have comments more than 1000.

Question 9:

```
avg_comments <- group_by(wall_street, post_type)

summarise(avg_comments, median_comments_count =
median(comments_count, na.rm = TRUE ))
```

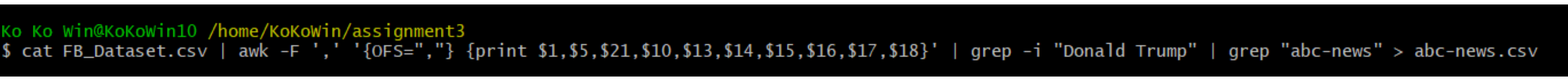


```
> avg_comments <- group_by(wall_street, post_type)
> summarise(avg_comments, median_comments_count = median(comments_count, na.rm = TRUE ))
`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 6 x 2
  post_type median_comments_count
  <chr>                     <dbl>
1 ""                           NA
2 "event"                       5
3 "link"                       25
4 "photo"                      54
5 "status"                     94
6 "video"                      38
>
```

I created a new data object called avg_comments. I used group by function to group the data by the post type accordingly . Furthermore, I used the summarise function to calculate the median comment count by each type of post. **na.rm = TRUE** function is used to remove if there is any empty value. Lastly, as we can see from the output that most effective type of post by the wall street journal is the status post type since it has highest median comment count.

Question 10:

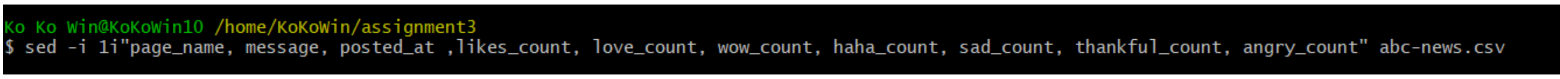
```
cat FB_Dataset.csv | awk -F ',' '{OFS=","} {print
$1,$5,$21,$10,$13,$14,$15,$16,$17,$18}' | grep -i "Donald Trump" | grep
"abc-news" > abc-news.csv
```



```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ cat FB_Dataset.csv | awk -F ',' '{OFS=","} {print $1,$5,$21,$10,$13,$14,$15,$16,$17,$18}' | grep -i "Donald Trump" | grep "abc-news" > abc-news.csv
```

I searched for word ‘Donald Trump’ in message column in **grep -i** function by ignoring cases. {OFS=”,”} is added to change the delimiter to comma so it will be same as the column headings. Secondly, I searched for Donald Trump posted by abc-news along with columns for reaction and time posted and saved it to a csv file name abc-news.csv .

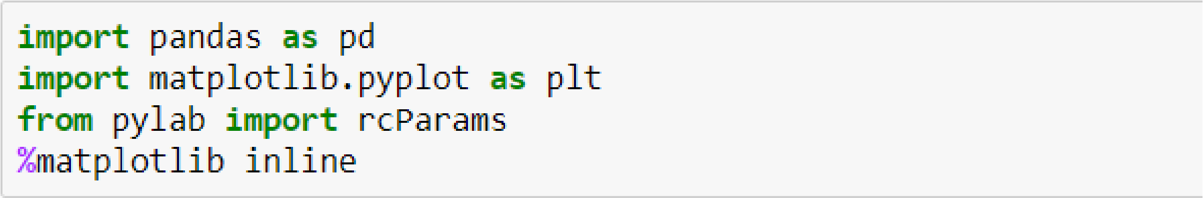
```
sed -i 1i"page_name, message, posted_at ,likes_count, love_count,
wow_count, haha_count, sad_count, thankful_count, angry_count" abc-
news.csv
```



```
Ko Ko Win@KoKoWin10 /home/KoKoWin/assignment3
$ sed -i 1i"page_name, message, posted_at ,likes_count, love_count, wow_count, haha_count, sad_count, thankful_count, angry_count" abc-news.csv
```

I used sed -i 1i function to added the column heading at the first row so it will be readable in R or Python.

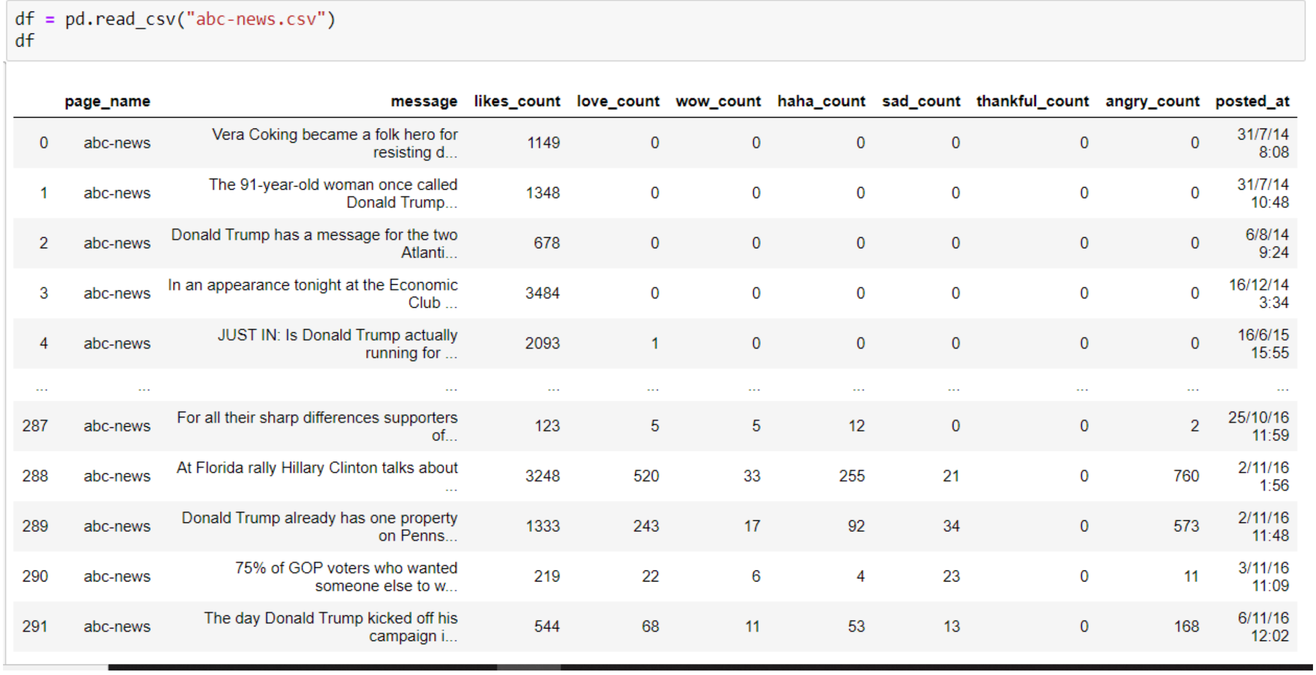
```
import pandas as pd
import matplotlib.pyplot as plt
from pylab import rcParams
%matplotlib inline
```



```
import pandas as pd
import matplotlib.pyplot as plt
from pylab import rcParams
%matplotlib inline
```

Importing necessary libraries.

```
df = pd.read_csv("abc-news.csv")
df
```

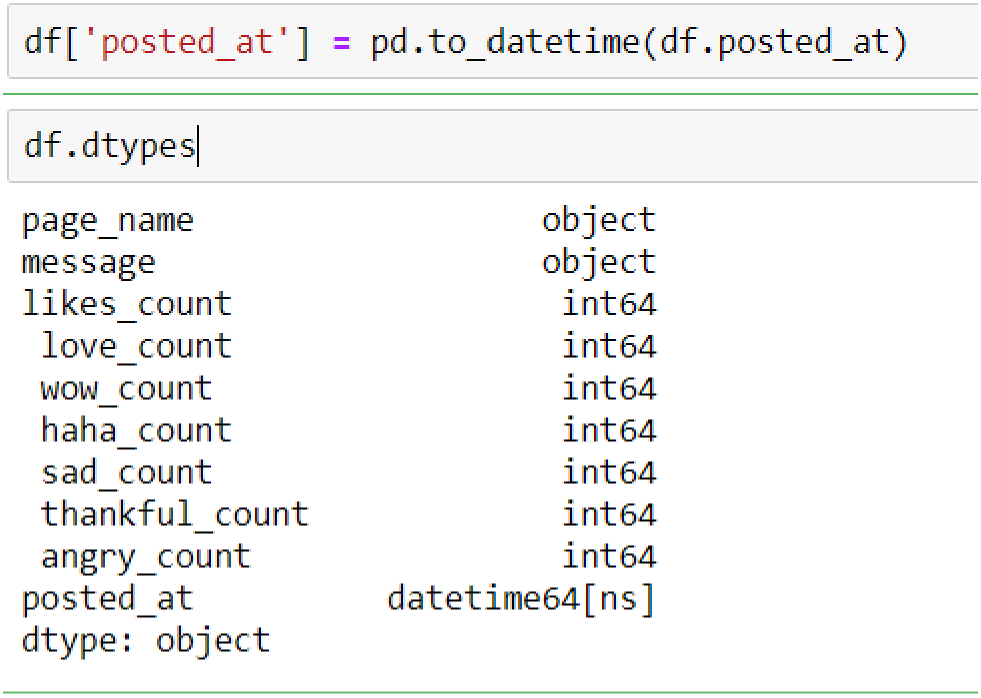


```
df = pd.read_csv("abc-news.csv")
df
```

	page_name	message	likes_count	love_count	wow_count	haha_count	sad_count	thankful_count	angry_count	posted_at
0	abc-news	Vera Coking became a folk hero for resisting d...	1149	0	0	0	0	0	0	31/7/14 8:08
1	abc-news	The 91-year-old woman once called Donald Trump...	1348	0	0	0	0	0	0	31/7/14 10:48
2	abc-news	Donald Trump has a message for the two Atlanti...	678	0	0	0	0	0	0	6/8/14 9:24
3	abc-news	In an appearance tonight at the Economic Club ...	3484	0	0	0	0	0	0	16/12/14 3:34
4	abc-news	JUST IN: Is Donald Trump actually running for ...	2093	1	0	0	0	0	0	16/6/15 15:55
...	...	...	...	...	...	...	...	...	...	...
287	abc-news	For all their sharp differences supporters of...	123	5	5	12	0	0	2	25/10/16 11:59
288	abc-news	At Florida rally Hillary Clinton talks about ...	3248	520	33	255	21	0	760	2/11/16 1:56
289	abc-news	Donald Trump already has one property on Penns...	1333	243	17	92	34	0	573	2/11/16 11:48
290	abc-news	75% of GOP voters who wanted someone else to w...	219	22	6	4	23	0	11	3/11/16 11:09
291	abc-news	The day Donald Trump kicked off his campaign i...	544	68	11	53	13	0	168	6/11/16 12:02

I used panda function to read the csv file containing post messages,reactions and time posted and called it ‘df’ . I chose the above 7 columns as a reaction because as a Facebook user when a certain post is something you like or agreed on I usually will give back a positive reactions which are likes,love,wow,haha and thankful. On the other hand, if the post is something I don’t like or agreed on I would give back a negative reaction which are sad and angry reactions.

```
df['posted_at'] = pd.to_datetime(df.posted_at)
df.dtypes
```



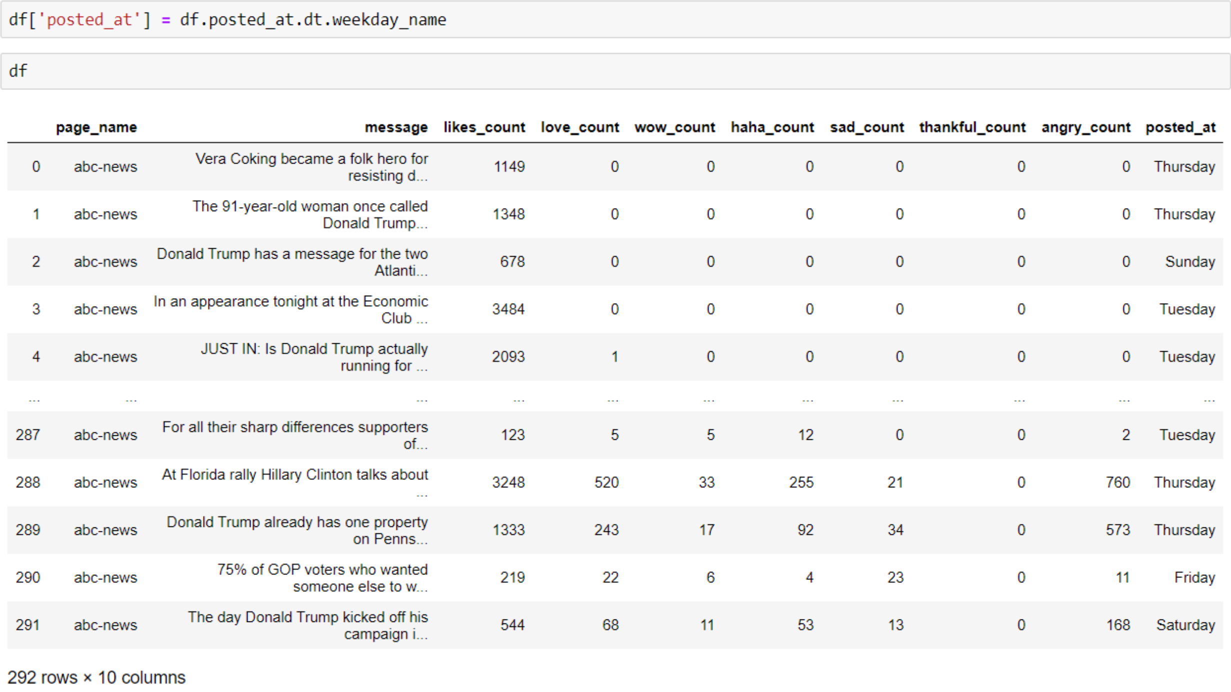
```
df['posted_at'] = pd.to_datetime(df.posted_at)
```

```
df.dtypes
```

```
page_name                  object
message                    object
likes_count                 int64
 love_count                 int64
 wow_count                  int64
 haha_count                 int64
 sad_count                  int64
 thankful_count             int64
 angry_count                int64
posted_at          datetime64[ns]
dtype: object
```

posted_at column contains date and time the post was posted and to change it to the weekday as the requirement of the question I first changed the datatype of the posted_at column into a datetime data type.

```
df['posted_at'] = df.posted_at.dt.weekday_name
df
```



```
df['posted_at'] = df.posted_at.dt.weekday_name
```

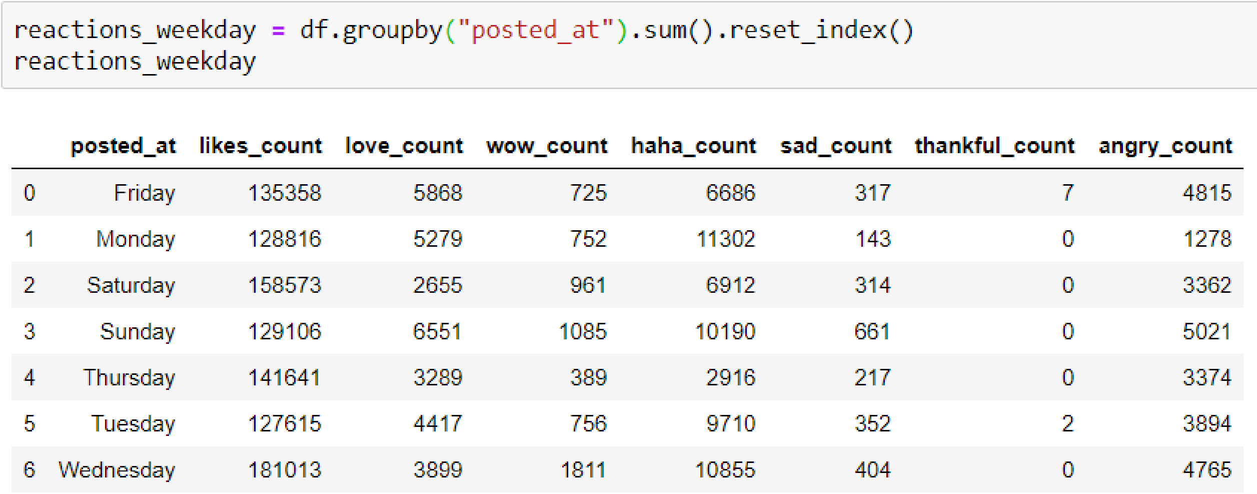
```
df
```

	page_name	message	likes_count	love_count	wow_count	haha_count	sad_count	thankful_count	angry_count	posted_at
0	abc-news	Vera Coking became a folk hero for resisting d...	1149	0	0	0	0	0	0	Thursday
1	abc-news	The 91-year-old woman once called Donald Trump...	1348	0	0	0	0	0	0	Thursday
2	abc-news	Donald Trump has a message for the two Atlanti...	678	0	0	0	0	0	0	Sunday
3	abc-news	In an appearance tonight at the Economic Club ...	3484	0	0	0	0	0	0	Tuesday
4	abc-news	JUST IN: Is Donald Trump actually running for ...	2093	1	0	0	0	0	0	Tuesday
...	...	...	...	...	...	...	...	...	...	...
287	abc-news	For all their sharp differences supporters of...	123	5	5	12	0	0	2	Tuesday
288	abc-news	At Florida rally Hillary Clinton talks about ...	3248	520	33	255	21	0	760	Thursday
289	abc-news	Donald Trump already has one property on Penns...	1333	243	17	92	34	0	573	Thursday
290	abc-news	75% of GOP voters who wanted someone else to w...	219	22	6	4	23	0	11	Friday
291	abc-news	The day Donald Trump kicked off his campaign i...	544	68	11	53	13	0	168	Saturday

292 rows × 10 columns

I used the date time function in python to change the posted_at column into weekdays name.

```
reactions_weekday = df.groupby("posted_at").sum().reset_index()
reactions_weekday
```

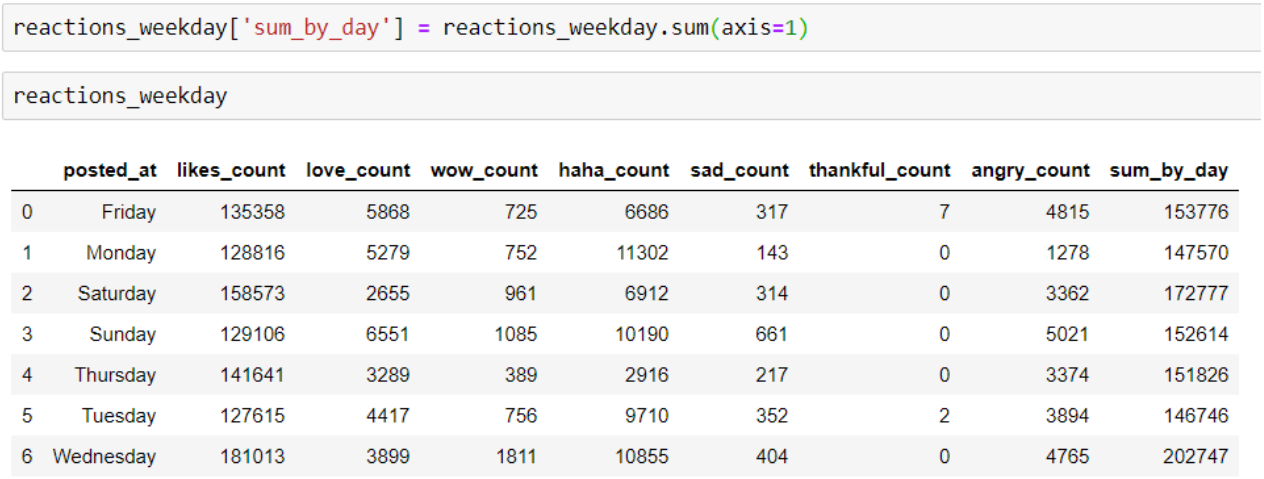


```
reactions_weekday = df.groupby("posted_at").sum().reset_index()
reactions_weekday
```

	posted_at	likes_count	love_count	wow_count	haha_count	sad_count	thankful_count	angry_count
0	Friday	135358	5868	725	6686	317	7	4815
1	Monday	128816	5279	752	11302	143	0	1278
2	Saturday	158573	2655	961	6912	314	0	3362
3	Sunday	129106	6551	1085	10190	661	0	5021
4	Thursday	141641	3289	389	2916	217	0	3374
5	Tuesday	127615	4417	756	9710	352	2	3894
6	Wednesday	181013	3899	1811	10855	404	0	4765

I used the group by function to group the day the post was posted and sum of each individual reaction for that day and saved it to a variable name reactions_weekday.

```
reactions_weekday['sum_by_day'] = reactions_weekday.sum(axis=1)
reactions_weekday
```



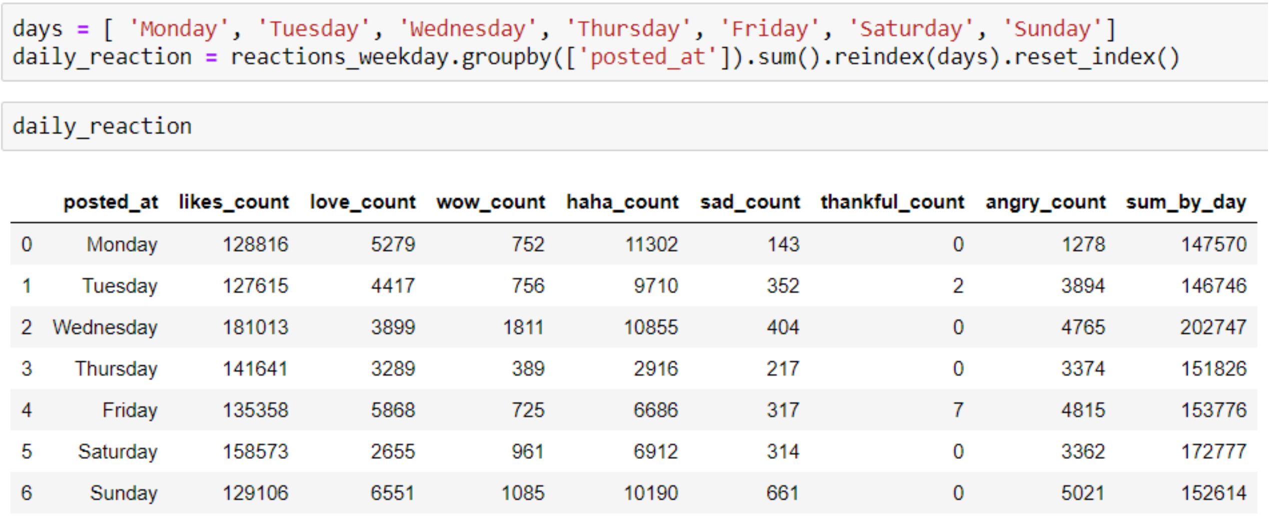
```
reactions_weekday['sum_by_day'] = reactions_weekday.sum(axis=1)
```

```
reactions_weekday
```

	posted_at	likes_count	love_count	wow_count	haha_count	sad_count	thankful_count	angry_count	sum_by_day
0	Friday	135358	5868	725	6686	317	7	4815	153776
1	Monday	128816	5279	752	11302	143	0	1278	147570
2	Saturday	158573	2655	961	6912	314	0	3362	172777
3	Sunday	129106	6551	1085	10190	661	0	5021	152614
4	Thursday	141641	3289	389	2916	217	0	3374	151826
5	Tuesday	127615	4417	756	9710	352	2	3894	146746
6	Wednesday	181013	3899	1811	10855	404	0	4765	202747

I created a new column called sum_by_day where I store the sum of all the reactions for each day.

```
days = [ 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']
daily_reaction = reactions_weekday.groupby(['posted_at']).sum().reindex(days).reset_index()
daily_reaction
```



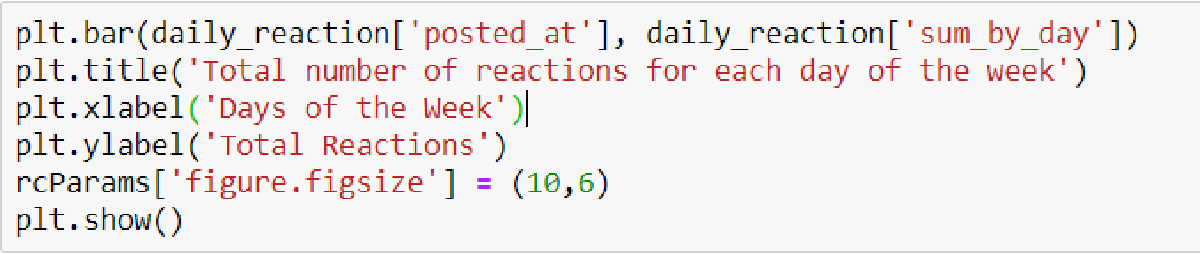
```
days = [ 'Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']
daily_reaction = reactions_weekday.groupby(['posted_at']).sum().reindex(days).reset_index()
```

```
daily_reaction
```

	posted_at	likes_count	love_count	wow_count	haha_count	sad_count	thankful_count	angry_count	sum_by_day
0	Monday	128816	5279	752	11302	143	0	1278	147570
1	Tuesday	127615	4417	756	9710	352	2	3894	146746
2	Wednesday	181013	3899	1811	10855	404	0	4765	202747
3	Thursday	141641	3289	389	2916	217	0	3374	151826
4	Friday	135358	5868	725	6686	317	7	4815	153776
5	Saturday	158573	2655	961	6912	314	0	3362	172777
6	Sunday	129106	6551	1085	10190	661	0	5021	152614

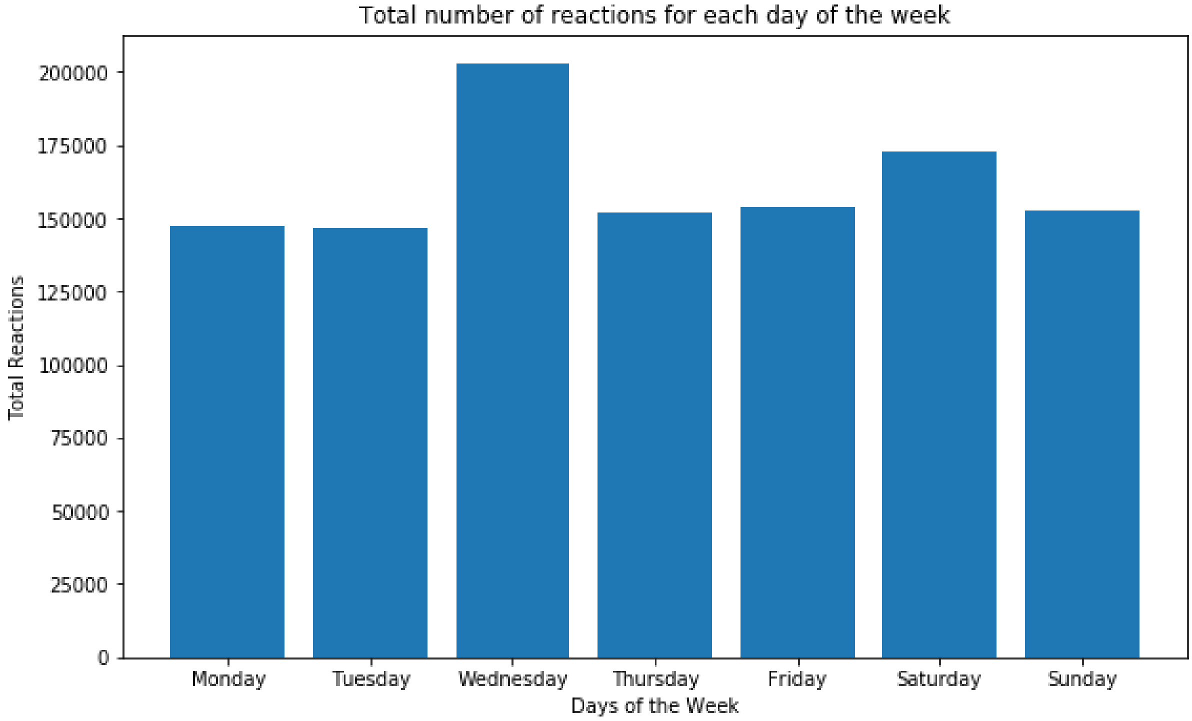
The question asked us to sort the day accordingly starting from Monday to Sunday. I created a new list called day where it contains all the days in a week. I created a new variable called daily_reaction and I group by the column posted_at and used the function reindex to make the days accordingly like the question want.

```
plt.bar(daily_reaction['posted_at'], daily_reaction['sum_by_day'])
plt.title('Total number of reactions for each day of the week')
plt.xlabel('Days of the Week')
plt.ylabel('Total Reactions')
rcParams['figure.figsize'] = (10,6)
plt.show()
```



```
plt.bar(daily_reaction['posted_at'], daily_reaction['sum_by_day'])
plt.title('Total number of reactions for each day of the week')
plt.xlabel('Days of the Week')
plt.ylabel('Total Reactions')
rcParams['figure.figsize'] = (10,6)
plt.show()
```

That is the code for plotting the bar chart. Weekdays will be at the x axis and total reactions will be at the y axis.



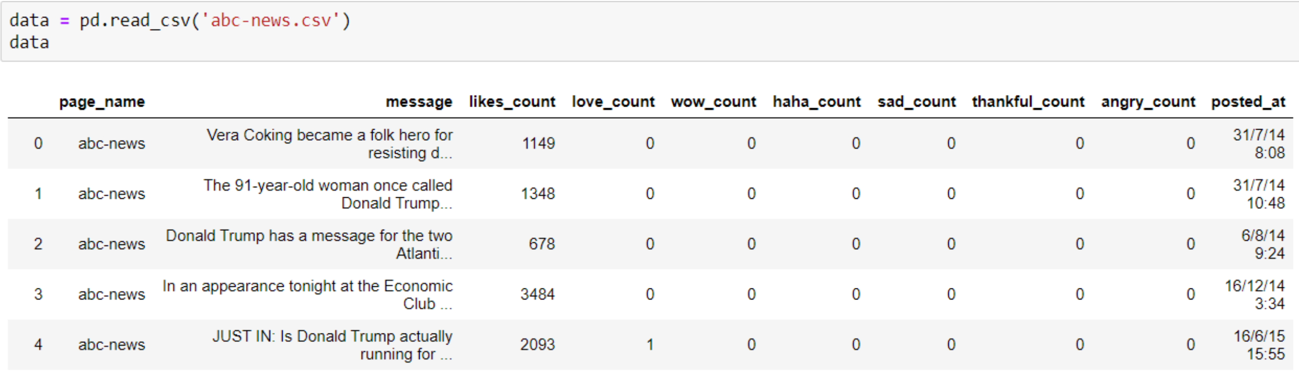
Bar chart for total number of reactions for each day of the week.

Question 11:

Wednesday and Saturday are the two days which the users have shown most reactions to the post at over 200000 and 175000 respectively. On Saturday, number of reactions to the post increases and it has second highest reaction among all the other days. However, it decreases again on Sunday. To conclude, we cannot say that there is a big difference between weekdays and weekends.

Question 12:

```
data = pd.read_csv('abc-news.csv')
data
```

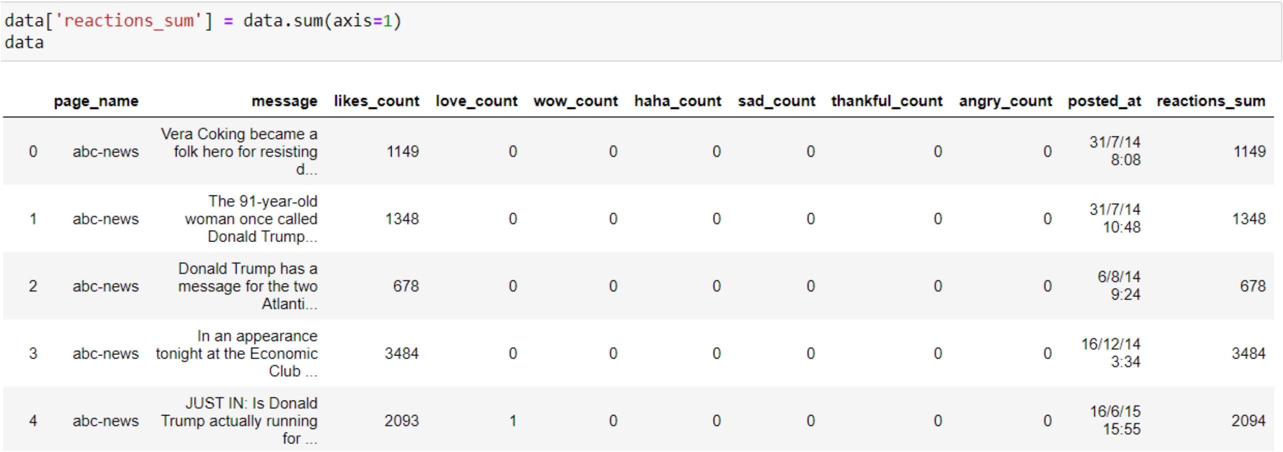


```
data = pd.read_csv('abc-news.csv')
data
```

	page_name	message	likes_count	love_count	wow_count	haha_count	sad_count	thankful_count	angry_count	posted_at
0	abc-news	Vera Coking became a folk hero for resisting d...	1149	0	0	0	0	0	0	31/7/14 8:08
1	abc-news	The 91-year-old woman once called Donald Trump...	1348	0	0	0	0	0	0	31/7/14 10:48
2	abc-news	Donald Trump has a message for the two Atlanti...	678	0	0	0	0	0	0	6/8/14 9:24
3	abc-news	In an appearance tonight at the Economic Club ...	3484	0	0	0	0	0	0	16/12/14 3:34
4	abc-news	JUST IN: Is Donald Trump actually running for ...	2093	1	0	0	0	0	0	16/6/15 15:55

I read the csv file again for question 12 and saved it to a variable name data.

```
data['reactions_sum'] = data.sum(axis=1)
data
```

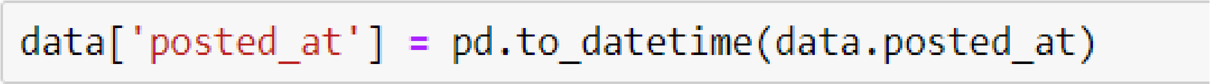


```
data['reactions_sum'] = data.sum(axis=1)
data
```

	page_name	message	likes_count	love_count	wow_count	haha_count	sad_count	thankful_count	angry_count	posted_at	reactions_sum
0	abc-news	Vera Coking became a folk hero for resisting d...	1149	0	0	0	0	0	0	31/7/14 8:08	1149
1	abc-news	The 91-year-old woman once called Donald Trump...	1348	0	0	0	0	0	0	31/7/14 10:48	1348
2	abc-news	Donald Trump has a message for the two Atlanti...	678	0	0	0	0	0	0	6/8/14 9:24	678
3	abc-news	In an appearance tonight at the Economic Club ...	3484	0	0	0	0	0	0	16/12/14 3:34	3484
4	abc-news	JUST IN: Is Donald Trump actually running for ...	2093	1	0	0	0	0	0	16/6/15 15:55	2094

After that I summed all the reactions for each day and saved it in a new column called reactions_sum.

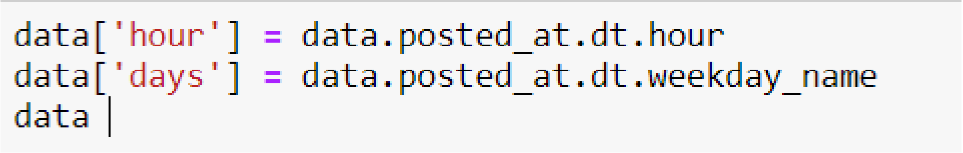
```
data['posted_at'] = pd.to_datetime(data.posted_at)
```



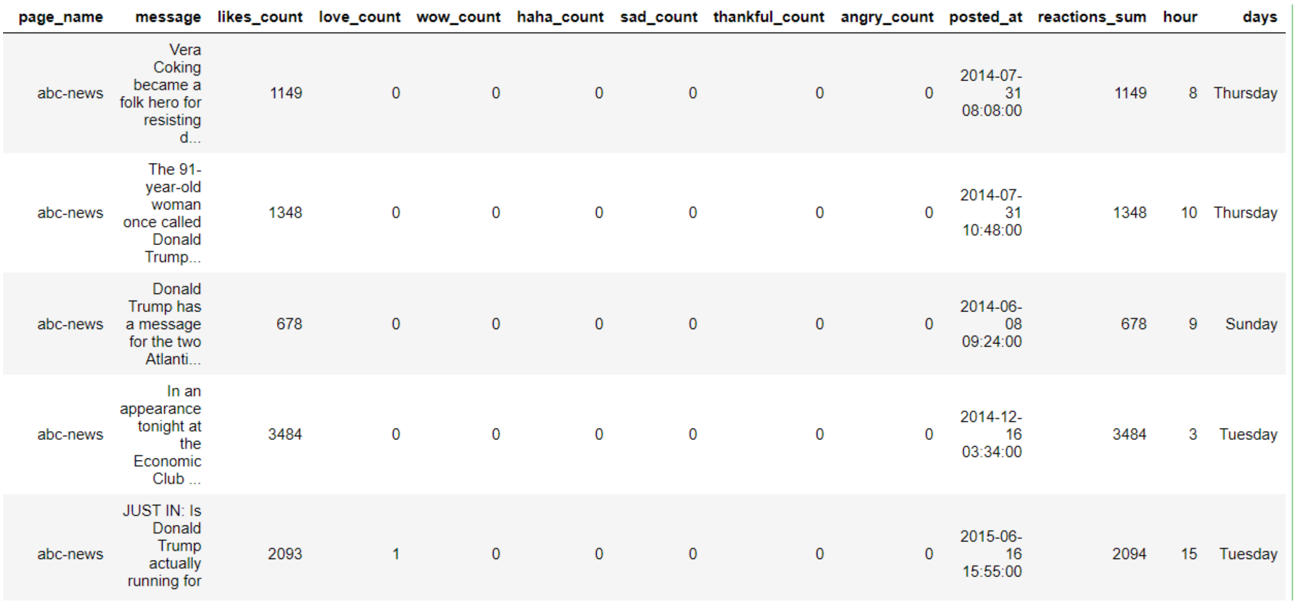
```
data['posted_at'] = pd.to_datetime(data.posted_at)
```

I changed the data type of the column posted_at to a datetime data type.

```
data['hour'] = data.posted_at.dt.hour
data['days'] = data.posted_at.dt.weekday_name
data
```



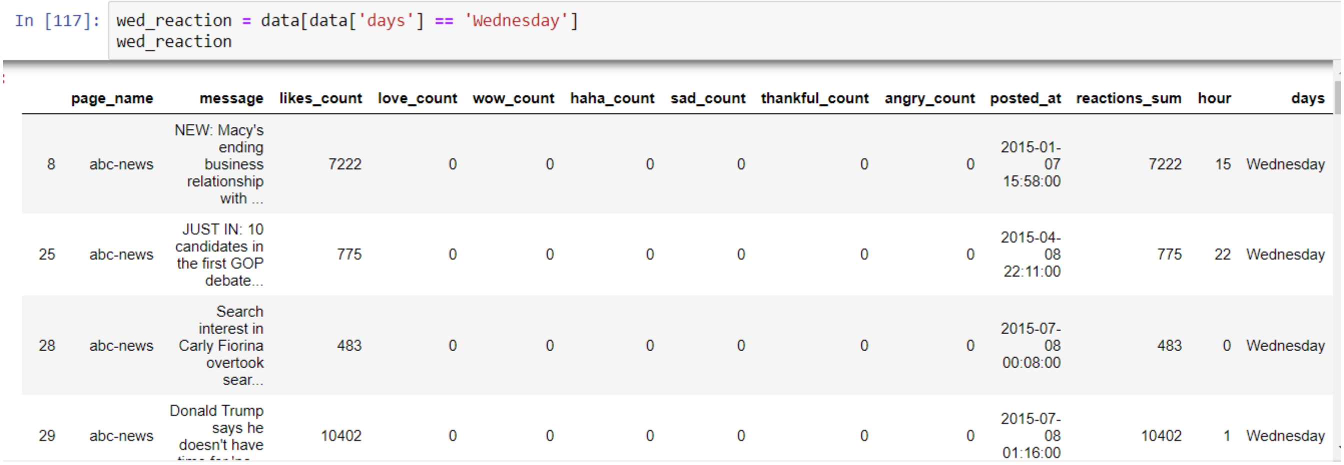
```
data['hour'] = data.posted_at.dt.hour
data['days'] = data.posted_at.dt.weekday_name
data
```



page_name	message	likes_count	love_count	wow_count	haha_count	sad_count	thankful_count	angry_count	posted_at	reactions_sum	hour	days
abc-news	Vera Coking became a folk hero for resisting d...	1149	0	0	0	0	0	0	2014-07-31 08:08:00	1149	8	Thursday
abc-news	The 91-year-old woman once called Donald Trump...	1348	0	0	0	0	0	0	2014-07-31 10:48:00	1348	10	Thursday
abc-news	Donald Trump has a message for the two Atlanti...	678	0	0	0	0	0	0	2014-06-08 09:24:00	678	9	Sunday
abc-news	In an appearance tonight at the Economic Club ...	3484	0	0	0	0	0	0	2014-12-16 03:34:00	3484	3	Tuesday
abc-news	JUST IN: Is Donald Trump actually running for	2093	1	0	0	0	0	0	2015-06-16 15:55:00	2094	15	Tuesday

I created a 2 new column called hour and days. Where I stored the hour and weekday name in those 2 columns.

```
wed_reaction = data[data['days'] == 'Wednesday']
wed_reaction
```

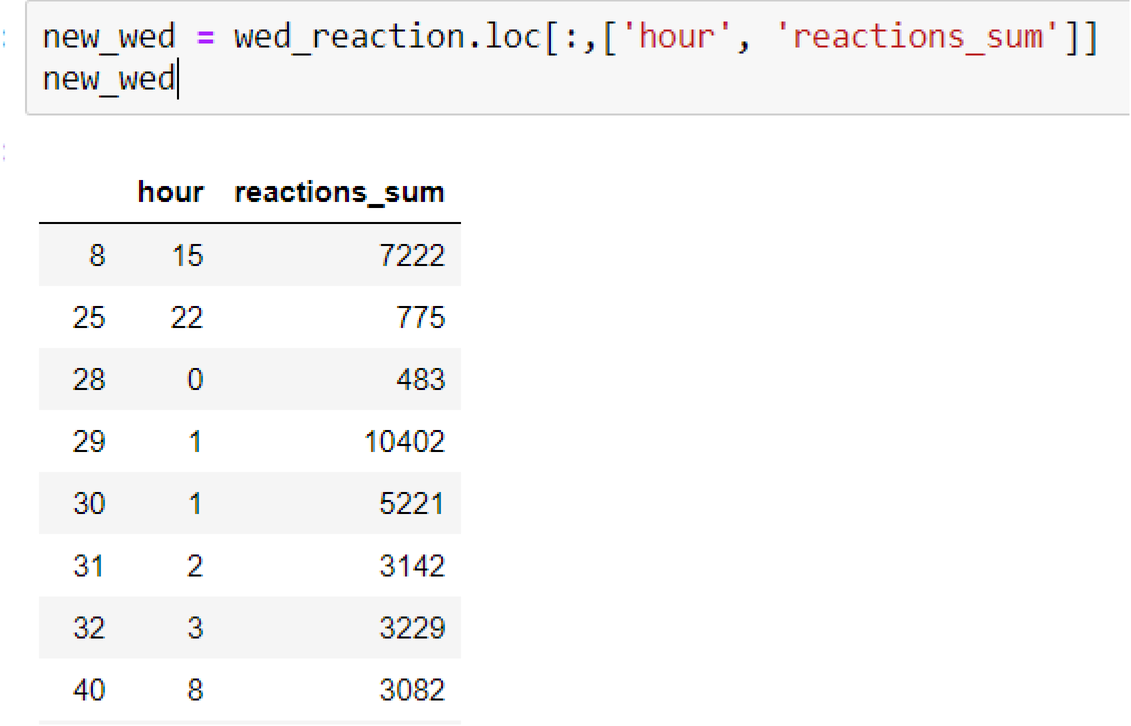


```
In [117]: wed_reaction = data[data['days'] == 'Wednesday']
          wed_reaction
```

	page_name	message	likes_count	love_count	wow_count	haha_count	sad_count	thankful_count	angry_count	posted_at	reactions_sum	hour	days
8	abc-news	NEW: Macy's ending business relationship with ...	7222	0	0	0	0	0	0	2015-01-07 15:58:00	7222	15	Wednesday
25	abc-news	JUST IN: 10 candidates in the first GOP debate...	775	0	0	0	0	0	0	2015-04-08 22:11:00	775	22	Wednesday
28	abc-news	Search interest in Carly Fiorina overtook sear...	483	0	0	0	0	0	0	2015-07-08 00:08:00	483	0	Wednesday
29	abc-news	Donald Trump says he doesn't have	10402	0	0	0	0	0	0	2015-07-08 01:16:00	10402	1	Wednesday

From the original dataframe “data” I extracted data for only Wednesday and saved it to a variable name wed_reaction.

```
new_wed = wed_reaction.loc[:,['hour', 'reactions_sum']]
new_wed
```

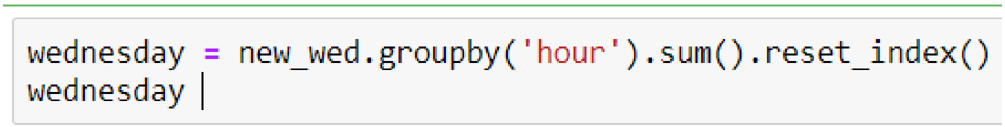


```
new_wed = wed_reaction.loc[:,['hour', 'reactions_sum']]
new_wed
```

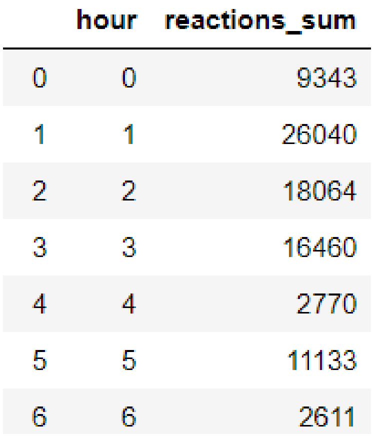
	hour	reactions_sum
8	15	7222
25	22	775
28	0	483
29	1	10402
30	1	5221
31	2	3142
32	3	3229
40	8	3082

Here I used .loc function to only take hour and reactions_sum column from our wed_reaction dataframe and saved it to a variable name new_wed.

```
wednesday = new_wed.groupby('hour').sum().reset_index()
wednesday
```



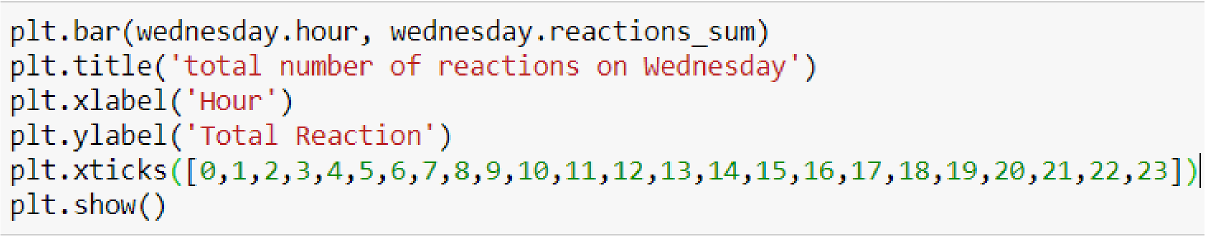
```
wednesday = new_wed.groupby('hour').sum().reset_index()
wednesday
```



	hour	reactions_sum
0	0	9343
1	1	26040
2	2	18064
3	3	16460
4	4	2770
5	5	11133
6	6	2611

I used the groupby function to grouped the same hour and sum the number of reactions for each hour.

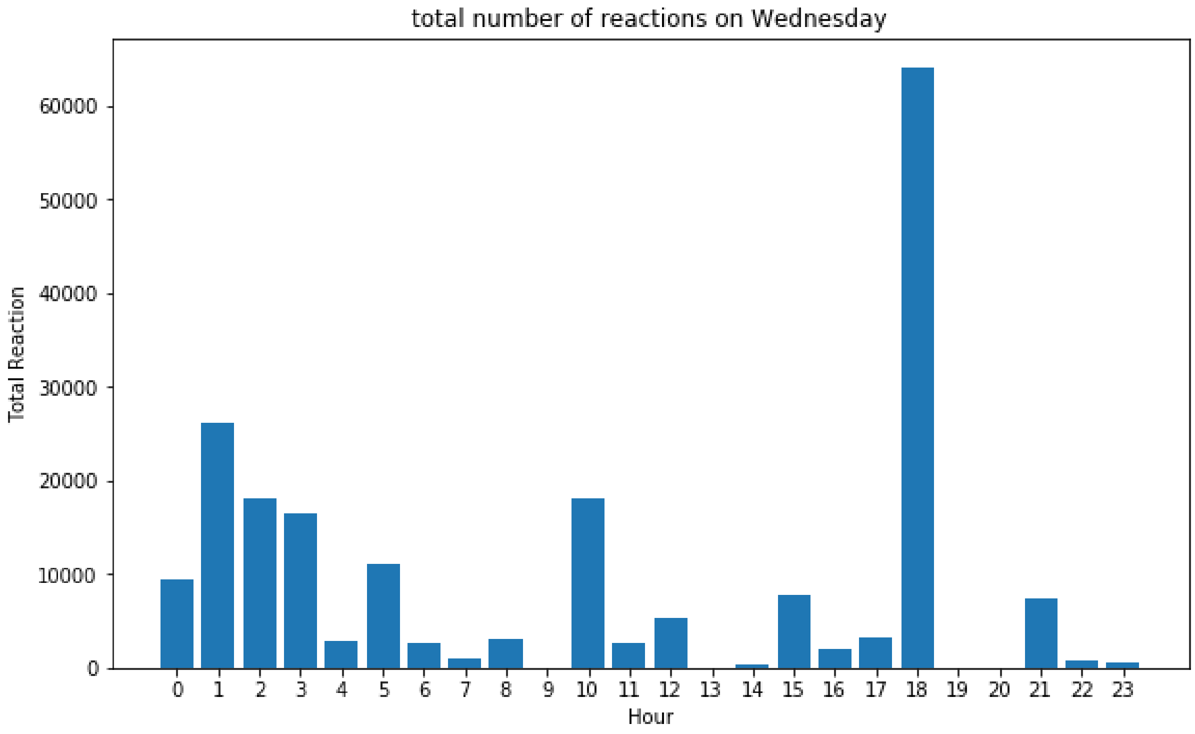
```
plt.bar(wednesday.hour, wednesday.reactions_sum)
plt.title('total number of reactions on Wednesday')
plt.xlabel('Hour')
plt.ylabel('Total Reaction')
plt.xticks([0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23]
)
plt.show()
```



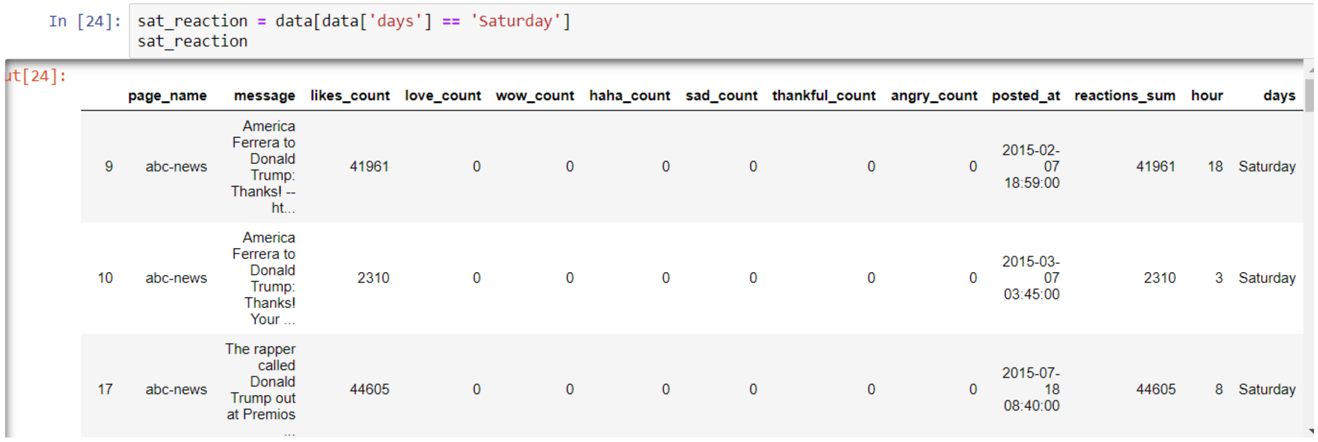
```
plt.bar(wednesday.hour, wednesday.reactions_sum)
plt.title('total number of reactions on Wednesday')
plt.xlabel('Hour')
plt.ylabel('Total Reaction')
plt.xticks([0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23])
plt.show()
```

Above is the code for plotting the bar chart for total reactions against the hour.

Bar chart for total number of reactions on Wednesday.



```
sat_reaction = data[data['days'] == 'Saturday']
sat_reaction
```



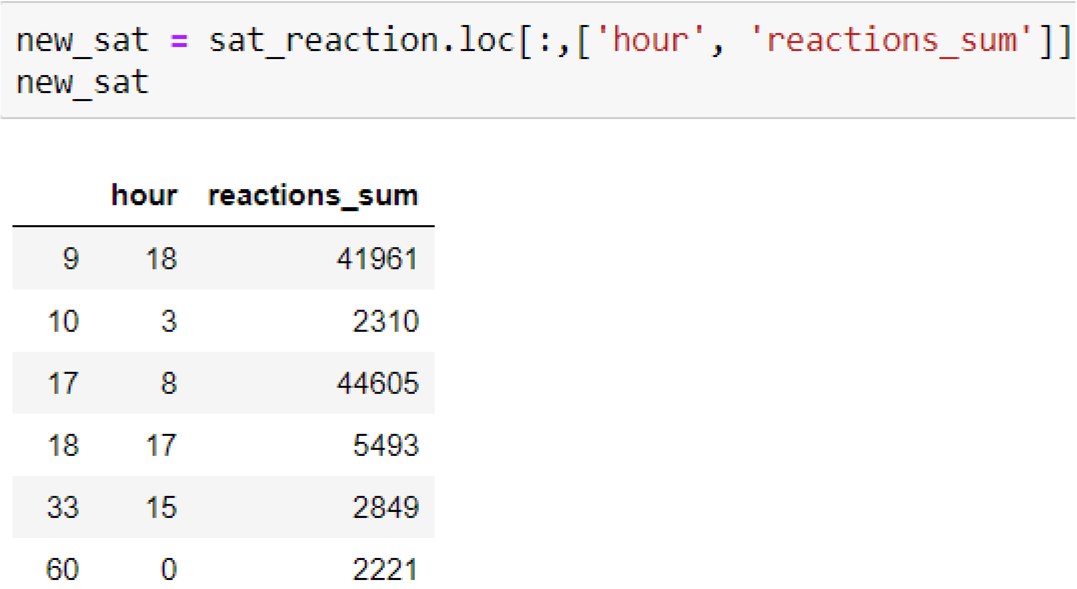
```
In [24]: sat_reaction = data[data['days'] == 'Saturday']
         sat_reaction
```

ut[24]:

	page_name	message	likes_count	love_count	wow_count	haha_count	sad_count	thankful_count	angry_count	posted_at	reactions_sum	hour	days
9	abc-news	America Ferrera to Donald Trump: Thanks! -- ht...	41961	0	0	0	0	0	0	2015-02-07 18:59:00	41961	18	Saturday
10	abc-news	America Ferrera to Donald Trump: Thanks! Your ...	2310	0	0	0	0	0	0	2015-03-07 03:45:00	2310	3	Saturday
17	abc-news	The rapper called Donald Trump out at Premios ...	44605	0	0	0	0	0	0	2015-07-18 08:40:00	44605	8	Saturday

From the original dataframe called data I extracted data for only Saturday and saved it to variable name sat_reaction which contains data for only Saturday.

```
new_sat = sat_reaction.loc[:,['hour', 'reactions_sum']]
new_sat
```

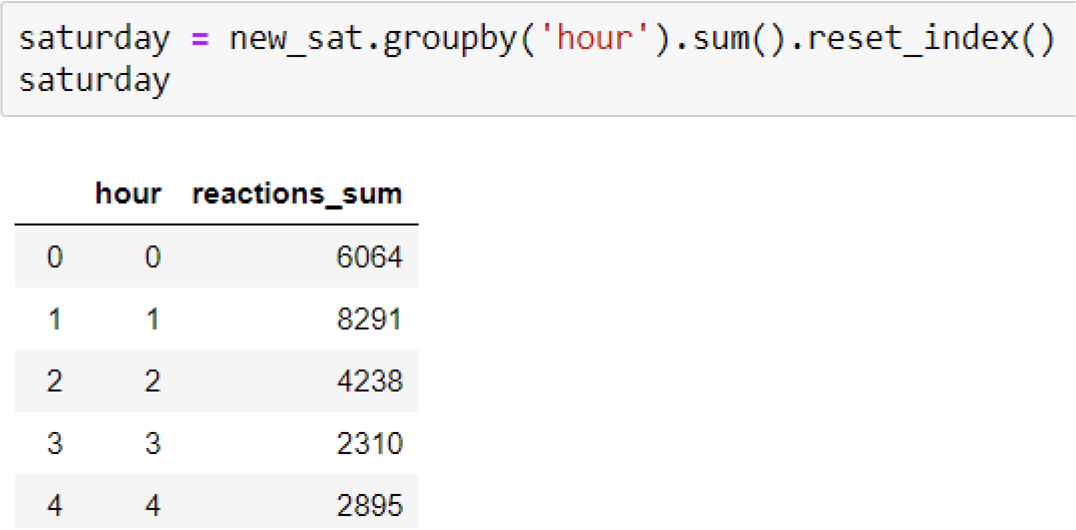


```
new_sat = sat_reaction.loc[:,['hour', 'reactions_sum']]
new_sat
```

	hour	reactions_sum
9	18	41961
10	3	2310
17	8	44605
18	17	5493
33	15	2849
60	0	2221

I used .loc function to only extract 2 columns which are hour and reactions_sum from the sat_reaction column. Lastly, saved it to a variable name new_sat

```
saturday = new_sat.groupby('hour').sum().reset_index()
saturday
```

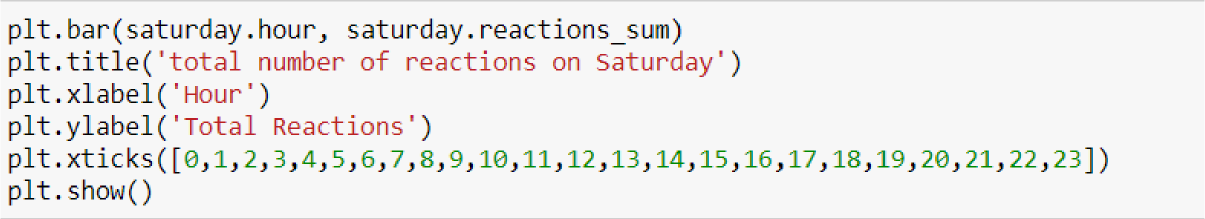


```
saturday = new_sat.groupby('hour').sum().reset_index()
saturday
```

	hour	reactions_sum
0	0	6064
1	1	8291
2	2	4238
3	3	2310
4	4	2895

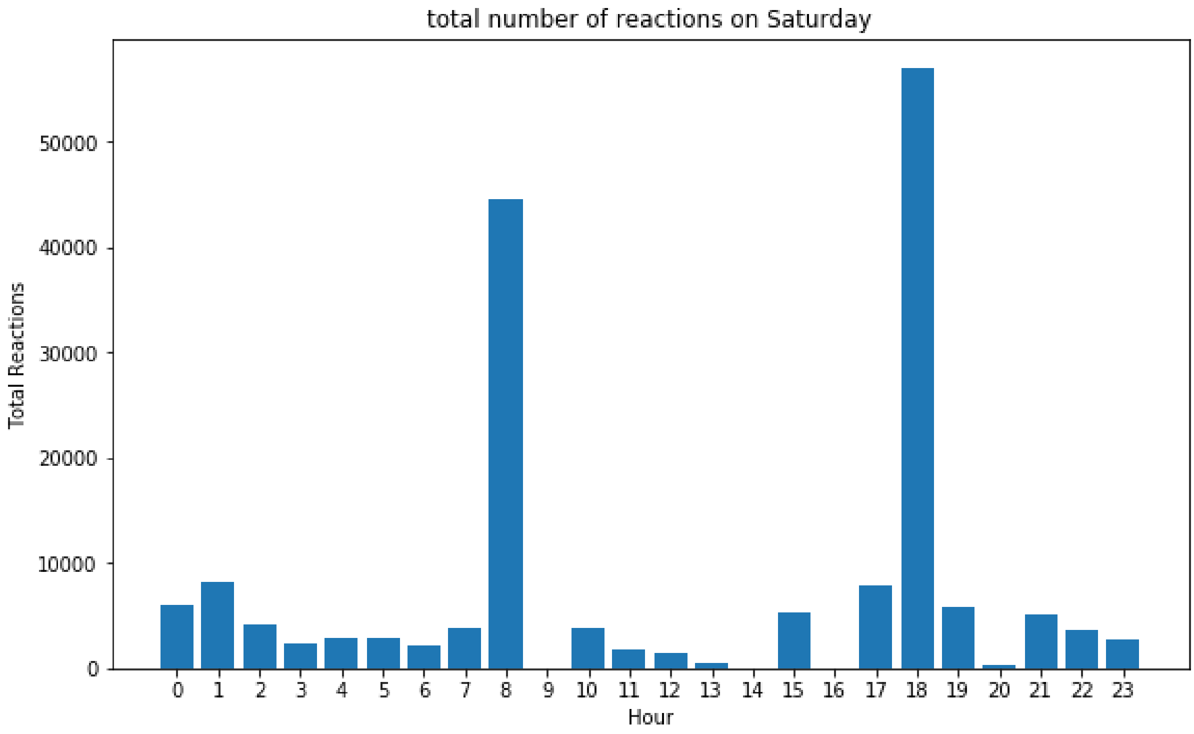
I used the groupby function to grouped the total number of reactions according to its hour and saved it to a variable name saturday.

```
plt.bar(saturday.hour, saturday.reactions_sum)

plt.title('total number of reactions on Saturday')

plt.xlabel('Hour')

plt.ylabel('Total Reactions')

plt.xticks([0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,
22,23])

plt.show()
```



```
plt.bar(saturday.hour, saturday.reactions_sum)
plt.title('total number of reactions on Saturday')
plt.xlabel('Hour')
plt.ylabel('Total Reactions')
plt.xticks([0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23])
plt.show()
```

Above is the code for plotting the bar chart for Total Reactions against the hour for Saturday.



Bar chart for total number of reaction on Saturday against the Hours.

Question 13:

a) According to our bar chart maximum number of reactions on abc-news post for the term Donald Trump was recorded on Wednesday at 18:00 hour. Total number of reactions was 64011.

b) Yes, I think it's a good idea to publish on Wednesday and Saturday as well as the peak hours during those two days which is 18:00 and 08:00 since it gets most amount reactions by the user. It is because majority of users following news pages will be the adults. Therefore, during their time going to work at 08:00 or coming back from work at 18:00, they will browse through the Facebook for any updated news while being on the public transport such as train or a bus. Therefore, posting during the peak hours on Wednesday and Saturday will get the most reactions to the post.