**FIT 1043**

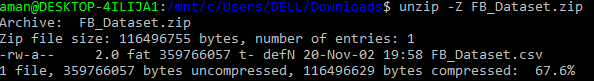
**Assignment 3**

**Task A**

1. Decompress the file. How big is it?

Decompression of the data and its size.

**Unzip –Z FB\_Dataset.zip**

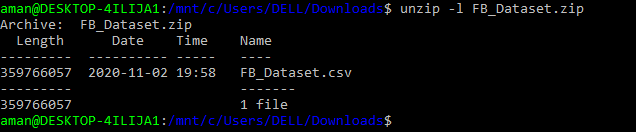


**Explanation:**

The file is in the compressed to decompress it we can simply used the above code and the file will be decompressed the size of file is 116496755 bytes when it’s compressed after decompression it is 116496629 bytes.

There is total 1 dataset in the zipped file and which we can see in the added snippet below to know the size of file we can simply run the code;

**Unzip –l FB\_Dataset.zip**

****

**Explanation:**

In the above snippet we can clearly see that there is one file saved in the zipped folder and the length of the file is 359766057 Bytes which is approx 360 MB.

1. What delimiter is used to separate the columns in the file? The 2nd column is the unique identifier for a Facebook post. Print out the names of other columns in the output.

**Code: unzip -p FacebookNews.zip | awk -F',' '{print NF; exit}'**

**D:\R Programs Practice\Unix Shell Outputs\FB_ data delimiter.PNG**

**Explanation:**

Since the file is a CSV(Comma-separated values) format hence the file is using the delimited ‘,’ to separate the column. And we used the ‘,’ while reading the file. And we can see that there are 21 columns in the processed FB\_dataset file.

**Other Columns:**

**Code: awk '{arr[$5] = 1} END {for (key in arr) {print key}}' FB\_Dataset.csv**

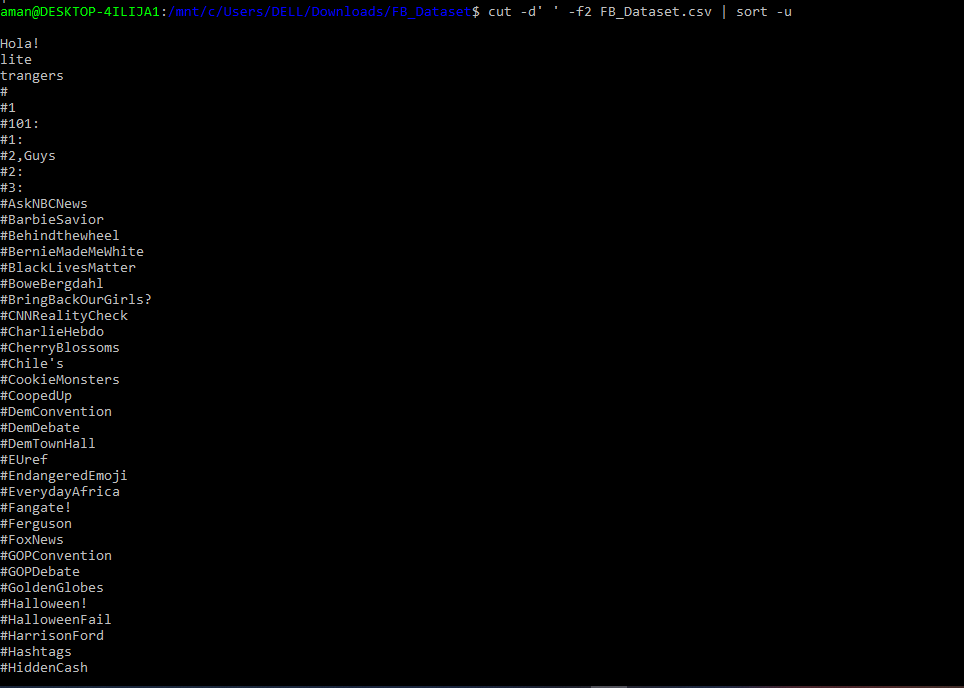
****

**Explanation:**

In the given dataset the first column is for the header of the data which shows the kind of data contained in the row while the other columns contains the data like description, post type, likes count etc. in the given data file.

1. How many unique pages are there?

**Code: cut -d' ' -f2 list.txt | sort –u**



**Explanation:**

To find the unique we simply have to sort the distinct values in the page id column and it will show all the distinct values for example; after running the code the most frequently used distinct hashtag was black\_lives\_matter then another one is Charlie\_Hebdo etc.

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

**Code: sed -E -n "2 {s/.\*,([^ ]\*).\*/\1 - /;h}; $ {s/.\*,([^ ]\*).\*/\1/;H;x;s/\n//;p}" file.csv**

D:\R Programs Practice\Unix Shell Outputs\FB_ data date_range.PNG

**Explanation:**

In the code snippet we can see that the dataset is ranged from the date 1/1/12 to 7/11/16. Hence it’s the data of the four years collected from the different news channels.

1. How many times has the term “Donald Trump” (ignore the case) appeared in the content of post names?When was the first mention of “Donald Trump” (ignore the case) in the post names and what was the post name? Considering different columns which you have for this post, excluding the message and post name, can you say whether people’s reactions were positive/ negative to this post? How many reactions can you see against this post?

**Code: grep –c “Trump” name of .CSV file.**

****

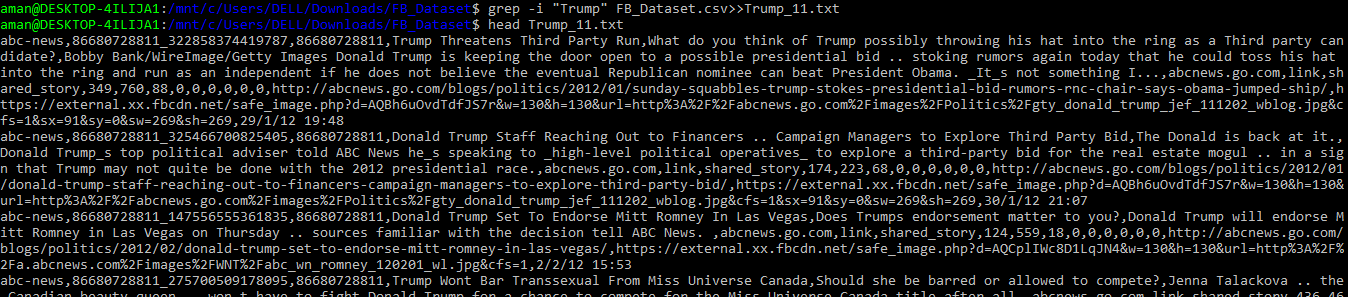
**Explanation:**

In the above code snippet we can see that the term Trump is written 26776 times on social media while the term “Donald Trump” is written 11668 times on the social media.

1. Select the post id and number of likes of posts in which the term “Trump” (Ignore the case) is mentioned in the post content and the number of likes is greater than 100. Then, sort the data based on the like\_count (descending sort) andsave it in a file named as “trump.txt”. (You need to add a screenshot of the output, including the first 5 rows and the column headers in your report).

**Code: grep “Donald Trump” name of .CSV file > Trump.txt**

**head trump.txt**

****

**Explanation:**

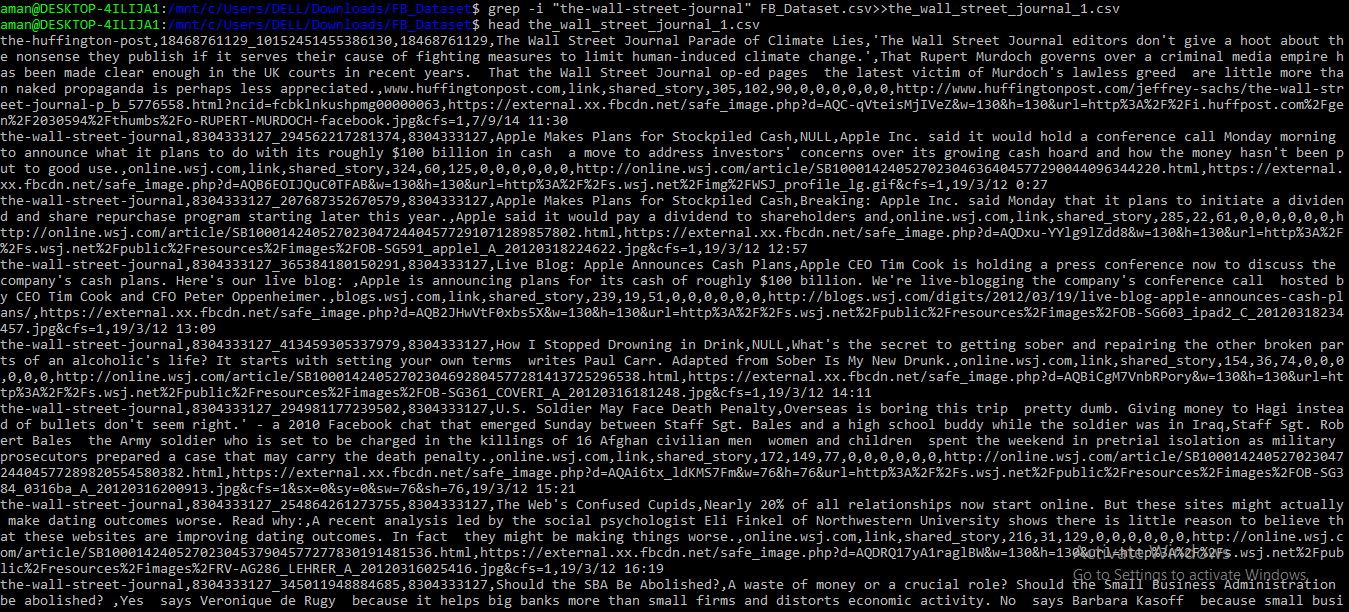
To know the post where Trump is mentioned first we have to run the same code which we have used previously in the case Italian food and then we will run exactly same way the head of trump.txt file.

Then extract the data from root csv file and then count the occurrence of Trump in the file.

**Task B1**

1. “the-wall-street-journal” has asked youto focus on the analysis of the post\_types forwhich the number of comments is less than 4000. You need to read the “the-wall-street-journal.csv” file generated as mentioned above into R, filter the data based on therequirements of “the-wall-street-journal”, and draw boxplots to show the distributionof comments made against each type of post (event, link, photo, status and video). Youneed to present one plot which contains different boxplots for different post types. Whatcan you infer from this plot? Can you detect which one is the most engaging post type?Make sure that your plot has proper labels and a title.

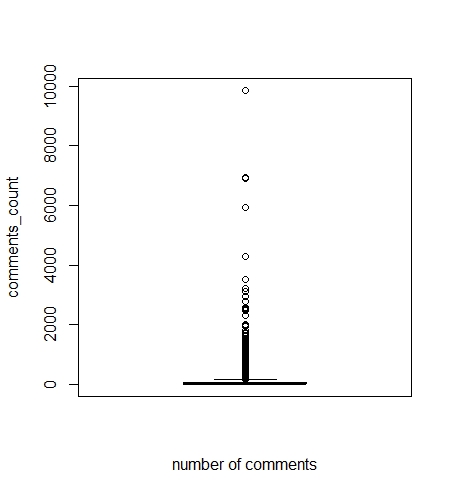
**Code: grep -i "the-wall-street-journal" FB\_Dataset.csv>>the\_wall\_street\_journal\_1.csv**

****

**Explanation:**

In the above code snippet we can clearly see that the data of wall street journal is extracted from the root csv file and then we have created its own separate csv file which is then used in R studio for the visualization of the wall street journals data and its researches.

1. You may have noticed that the presence of outliers affects the readability andinterpretation of the data in the box plots. Redraw the boxplot by filtering out values(comments\_count) greater than 1000.

****

**Fig 1 : outliers in the normal views.**

In the above plot we can see that the comments count in the page of wall street journal have outliers as most of the comments are lying under the range of 1000 which are not the outliers but after applying the following code;

**Code:** **boxplot(the\_wall\_street\_data$comments\_count>1000,ylab = "comments\_count")**

The boxplot will look like below which contains only the outliers of the dataset;

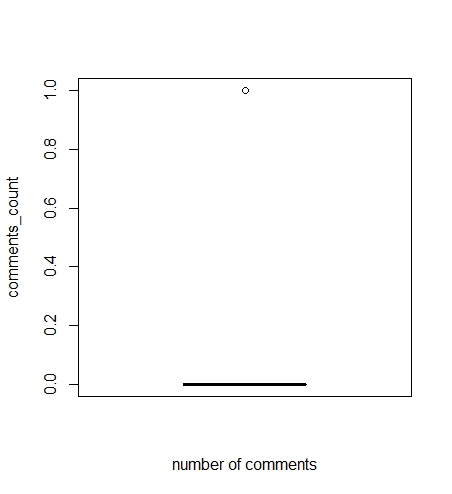


Fig 2: Outliers detected in the dataset using boxplot visualization.

1. Which type of post (event, link, photo, status,or video) has on average been mosteffective for “the-wall-street-journal.csv”? In other words, which post\_type has the highest median comment count?

**Code: median(the\_wall\_street\_data$post\_type, na.rm = FALSE)**

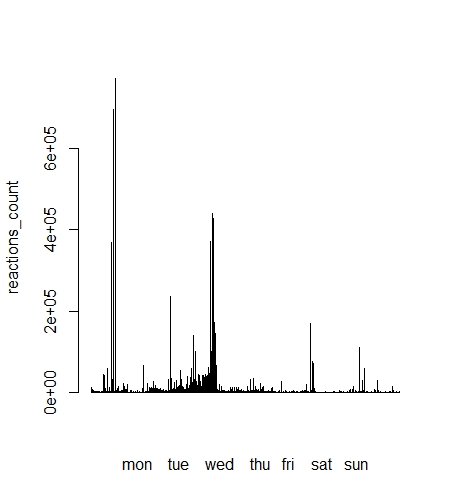
****

**Explanation:**

In the above snippet we can see that the post type which is the highest median of the wall street journal dataset is basically a link of wsj.com and its id is given above hence we can see that the links are the most shared data according to the wall street journals research.

**Task B2**

1. Create a bar chart which shows the total number of reactions to the posts published by“abc-news”, in which the posted message contains the term “Donald Trump”(ignorethe case) for each day of the week ('Monday', 'Tuesday', 'Wednesday', 'Thursday','Friday', 'Saturday', 'Sunday'). Make sure the bar chart is sorted based on the weekdaysas shown in the screenshot below. Understanding what should be considered as areaction is a part of the answer to this question which you can figure out by checkingdifferent columns of your dataset. You need to mention and justify the criterion whichyou choose to define a reaction to a post. (Please pay attention that the plot does notshow the real values and it is created with fake data just to show you how the outputshould be).



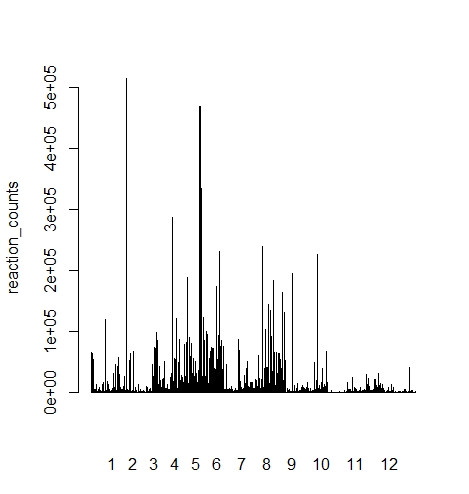
Here in the above plot the bars are going from left to right which is from Mon to Sunday and we can clearly see here that the most of the days reaction is given is the working day while least used on the Saturday.

1. Considering the created bar chart of question 10, name two days in which users have shown the most reactions to the posts. Is there any difference between the number of reactions during the weekdays and at the weekends?

As we can see the above barplot we can say that there is a certain difference between weekdays and weekends we can see here that the term Trump is mostly used on Wednesday and Thursday when most of the public is going to work but if take a look on progression of the bar plot we can see that on weekends people doesn’t use or less used social media so Trump isn’t written on the social media in weekdays, the days when trump is written on the social media are Tuesday, Wednesday and Thursday.

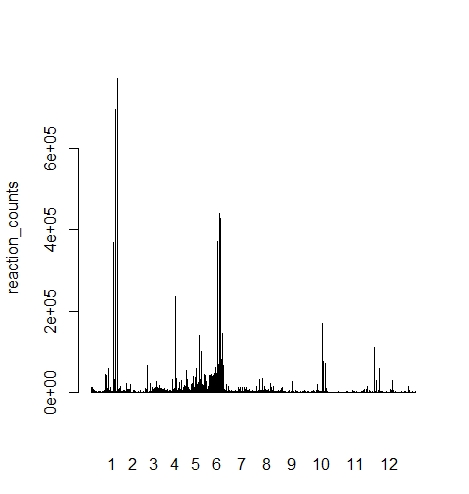
1. We need to take a closer look at the total reactions in the two days which users have shown the most reactions. Create two bar charts to show the hourly total reactions for each of two days. What time did the most reactions happen on each day? Is there any similarity between the numbers of hourly reactions in these two days? (Please pay attention that the sample plot given below does not show the real values and it is created with fake data just to show you how the output should be presented for this question.)

The reaction counts per hour for the two days looks like below, the reaction count on the weekend is look like this;



**Fig 3: Reaction counts in weekday**

While on the other hands on the weekdays or we can say on the working days the reaction counts looks like below;



**Fig 4: Reaction counts in weekend**

If we do the comparison we can clearly see that people use social in working days more than the weekends and hence the reaction counts are more on weekdays than weekends.

1. Considering your exploration about the reactions in different days/times for the term “Donald Trump” in posts by “abc-news”, answer the following questions.
2. What was the day and time which had the maximum number of reactions?

The day and time of the most reactions are usually in the weekdays as we have seen its during the office hours which in between 11:00 AM to 12:00 PM when the most of reactions comes according to our analysis while least reactions of analysis at 5:00 PM in the evening .

1. Do you think it is a good idea to recommend publishing a general post aboutTrump in the days which you found in question 11 and the peak hours which you found in question 12? What is your suggestion based on the analysis whichyou did in this task? Justify your answer

Yes, It will be a good idea to publish the general post of the data analysis of the popularity of the Donald Trump and the time on which he gets most and the least reactions from the social media because in the era of digitization the the social media popularity of the political personalities can be a decision factor.