FIT1043 Assignment3

Name: Lianzheng Xie

Student ID: 32068611

Task A : Investigating User global-scale check-in data collected from Foursquare Data in the Shell

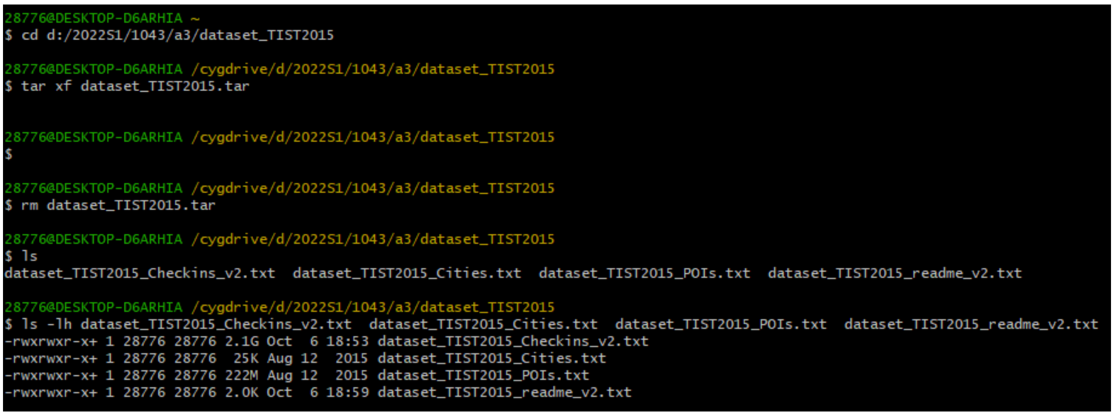
`cd` **open the folder where the dataset resides**

`tar xf dataset_TIST2015.tar` **decompress the compressed file,**

`ls` **get the file,**

`ls -lh` **get each file size.**

1) There are 4 files in the tar file, dataset_TIST2015_Checkins_v2.txt is 2.1G, dataset_TIST2015_readme_v2.txt is 2.0k, dataset_TIST2015_Cities.txt is 222M, dataset_TIST2015_POIs.txt is 25k.

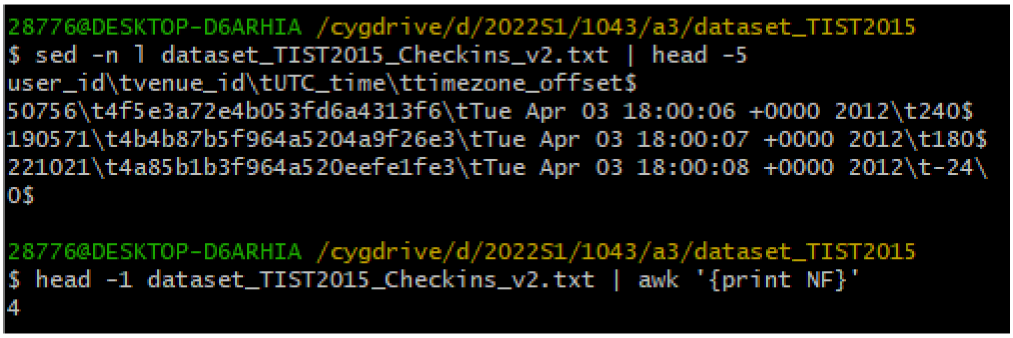


```
28776@DESKTOP-D6ARHIA ~
$ cd d:/2022S1/1043/a3/dataset_TIST2015

28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ tar xf dataset_TIST2015.tar


28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$

28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ rm dataset_TIST2015.tar

28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ ls
dataset_TIST2015_Checkins_v2.txt  dataset_TIST2015_Cities.txt  dataset_TIST2015_POIs.txt  dataset_TIST2015_readme_v2.txt

28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ ls -lh dataset_TIST2015_Checkins_v2.txt  dataset_TIST2015_Cities.txt  dataset_TIST2015_POIs.txt  dataset_TIST2015_readme_v2.txt
-rwxrwxr-x+ 1 28776 28776 2.1G Oct  6 18:53 dataset_TIST2015_Checkins_v2.txt
-rwxrwxr-x+ 1 28776 28776  25K Aug 12  2015 dataset_TIST2015_Cities.txt
-rwxrwxr-x+ 1 28776 28776 222M Aug 12  2015 dataset_TIST2015_POIs.txt
-rwxrwxr-x+ 1 28776 28776 2.0K Oct  6 18:59 dataset_TIST2015_readme_v2.txt
```

2) The delimiter of dataset_TIST2015_Checkins_v2.txt is \t(\<tab>), and this file has 4 columns.

`sed -n l dataset_TIST2015_Checkins_v2.txt | head -5` **Look at the separator in the first five lines of the file**

`head -1 dataset_TIST2015_Checkins_v2.txt | awk '{print NF}'` **get the number of columns of first line of file**



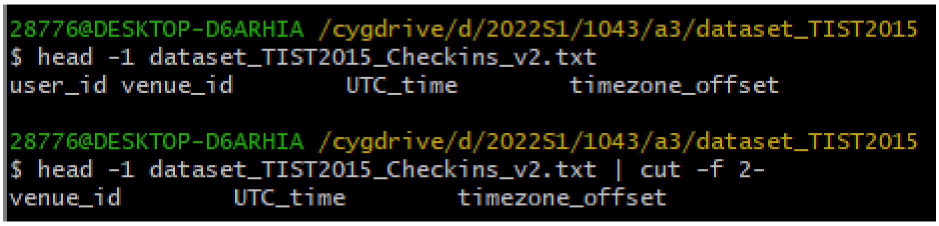
```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ sed -n l dataset_TIST2015_Checkins_v2.txt | head -5
user_id\tvenue_id\tUTC_time\ttimezone_offset$
50756\t4f5e3a72e4b053fd6a4313f6\tTue Apr 03 18:00:06 +0000 2012\t240$
190571\t4b4b87b5f964a5204a9f26e3\tTue Apr 03 18:00:07 +0000 2012\t180$
221021\t4a85b1b3f964a520eefe1fe3\tTue Apr 03 18:00:08 +0000 2012\t-24\
0$

28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ head -1 dataset_TIST2015_Checkins_v2.txt | awk '{print NF}'
4
```

3) Other columns are venue_id, UTC_time and timezone_offset

`head -1 dataset_TIST2015_Checkins_v2.txt` Show the first line of the file.

`head -1 dataset_TIST2015_Checkins_v2.txt | cut -f 2-`

Show all columns in the first row except the first column, cut is cut from the second column.



```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ head -1 dataset_TIST2015_Checkins_v2.txt
user_id venue_id        UTC_time        timezone_offset

28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ head -1 dataset_TIST2015_Checkins_v2.txt | cut -f 2-
venue_id        UTC_time        timezone_offset
```

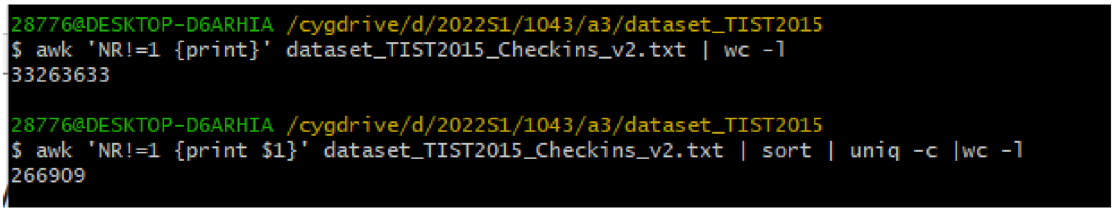
4) There are 33263633 Checkins, and 266909 users in the file.

`awk 'NR!=1 {print}' dataset_TIST2015_Checkins_v2.txt | wc -l`

Get all the data in the file except the first line and then get the number of lines

`awk 'NR!=1 {print $1}' dataset_TIST2015_Checkins_v2.txt|sort| uniq -c |wc -l`

Get the first column in the file except the first line, then sort, remove the repetition and count the number of repetitions, and finally get the number of rows.

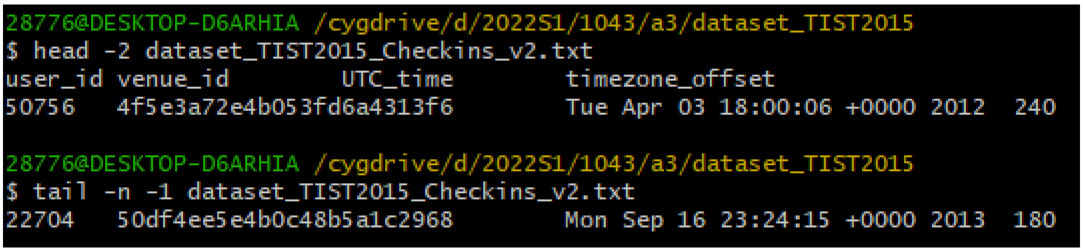


```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ awk 'NR!=1 {print}' dataset_TIST2015_Checkins_v2.txt | wc -l
33263633

28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ awk 'NR!=1 {print $1}' dataset_TIST2015_Checkins_v2.txt | sort | uniq -c |wc -l
266909
```

5) First date is Tue Apr 03 18:00:06, last date is Mon Sep 16 23:24:15.

`head -2 dataset_TIST2015_Checkins_v2.txt` Show the first two line of the file

`tail -n -1 dataset_TIST2015_Checkins_v2.txt` Show the last line of the file



```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ head -2 dataset_TIST2015_Checkins_v2.txt
user_id venue_id        UTC_time        timezone_offset
50756   4f5e3a72e4b053fd6a4313f6        Tue Apr 03 18:00:06 +0000 2012  240

28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ tail -n -1 dataset_TIST2015_Checkins_v2.txt
22704   50df4ee5e4b0c48b5a1c2968        Mon Sep 16 23:24:15 +0000 2013  180
```

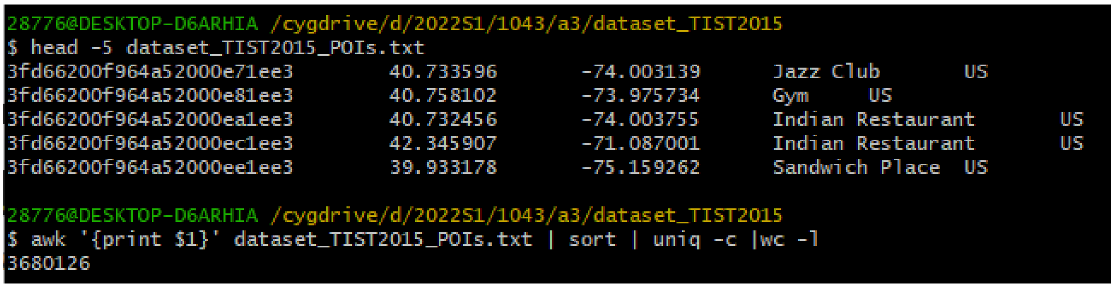
6) There are 3680126 unique venue IDs in the file.

`head -5 dataset_TIST2015_POIs.txt`

Show the first 5 rows to find venue IDs in the first column

`awk '{print $1}' dataset_TIST2015_POIs.txt | sort | uniq -c |wc -l`

Get the first column of data in the file, then sort, remove the repetition and count the number of repetitions, and finally get the number of rows.

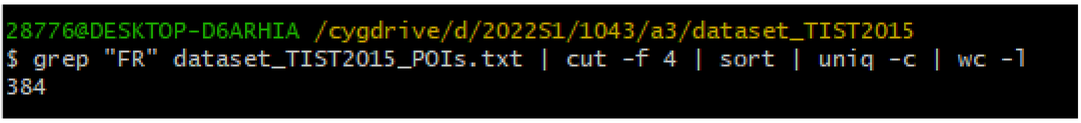


```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ head -5 dataset_TIST2015_POIs.txt
3fd66200f964a52000e71ee3        40.733596       -74.003139      Jazz Club       US
3fd66200f964a52000e81ee3        40.758102       -73.975734      Gym     US
3fd66200f964a52000ea1ee3        40.732456       -74.003755      Indian Restaurant       US
3fd66200f964a52000ec1ee3        42.345907       -71.087001      Indian Restaurant       US
3fd66200f964a52000ee1ee3        39.933178       -75.159262      Sandwich Place  US

28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ awk '{print $1}' dataset_TIST2015_POIs.txt | sort | uniq -c |wc -l
3680126
```

7) France contains 384 unique Venue categories in the file.

`grep "FR" dataset_TIST2015_POIs.txt| cut -f 4 | sort | uniq -c| wc -l`

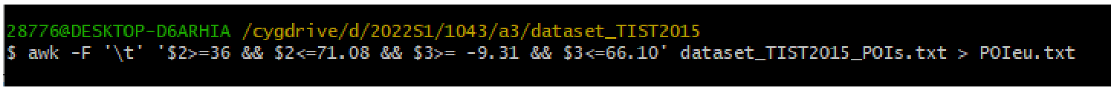
grep finds all the rows containing FR, intercepts the fourth column representing the site category, then sort, remove the repetition and count the number of repetitions, and finally get the number of rows.



```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ grep "FR" dataset_TIST2015_POIs.txt | cut -f 4 | sort | uniq -c | wc -l
384
```

8) A. `awk -F '\t' '$2>=36 && $2<=71.08 && $3>=-9.31 && $3<=66.10' dataset_TIST2015_POIs.txt> POIeu.txt`

According longitude and latitude range of Europe(36,71.08&-9.31,66.10) get the all data that meets the conditions, then print as a txt.

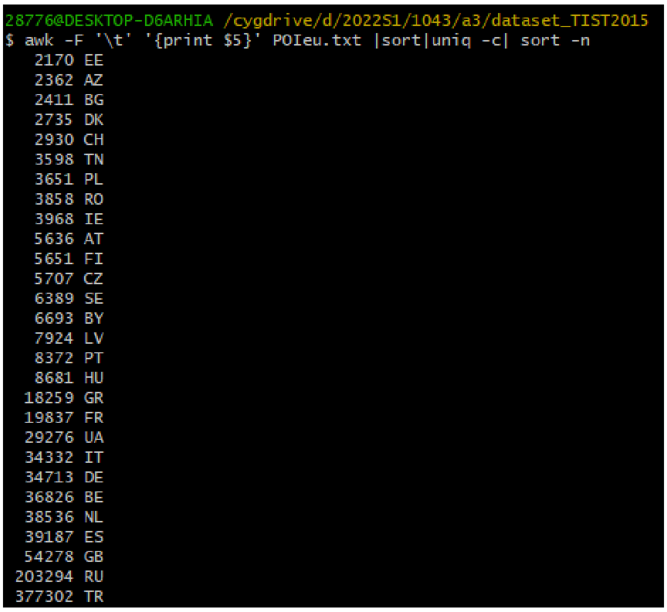


```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ awk -F '\t' '$2>=36 && $2<=71.08 && $3>= -9.31 && $3<=66.10' dataset_TIST2015_POIs.txt > POIeu.txt
```

B. `awk -F '\t' '{print $5}' POIeu.txt |sort | uniq -c| sort -n`

According to ‘\t’ get the fifth column(country) of txt(A8.A), then sort, remove the repetition and count the number of repetitions, and finally get the number of rows.

Most venues is Turkey(TR) with 377302, the least venues is Estonia(EE) with 2170.

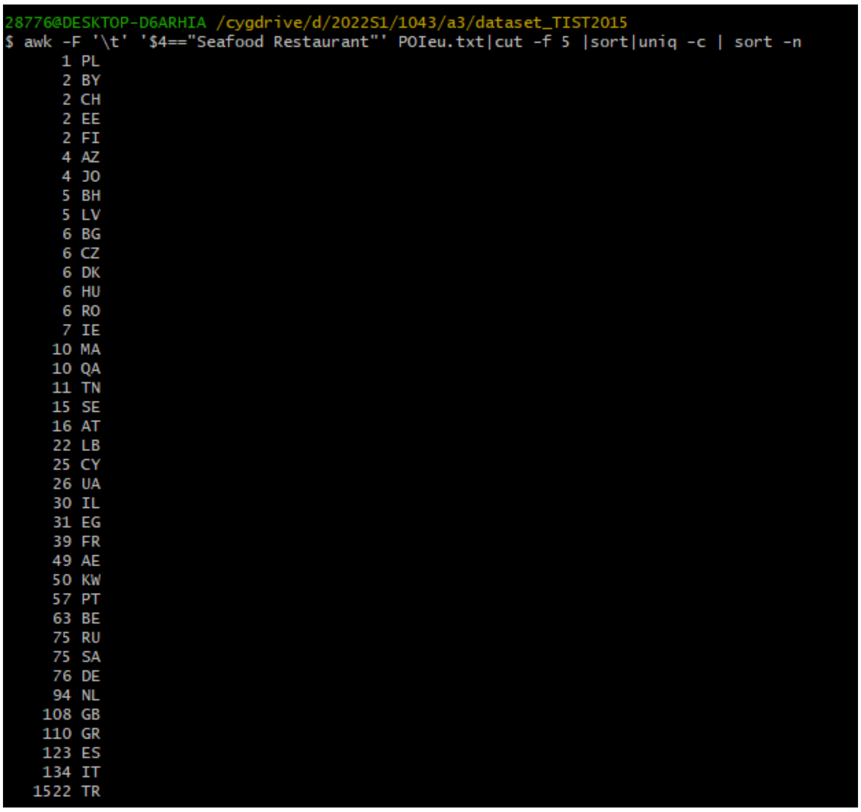


```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ awk -F '\t' '{print $5}' POIeu.txt |sort|uniq -c| sort -n
   2170 EE
   2362 AZ
   2411 BG
   2735 DK
   2930 CH
   3598 TN
   3651 PL
   3858 RO
   3968 IE
   5636 AT
   5651 FI
   5707 CZ
   6389 SE
   6693 BY
   7924 LV
   8372 PT
   8681 HU
  18259 GR
  19837 FR
  29276 UA
  34332 IT
  34713 DE
  36826 BE
  38536 NL
  39187 ES
  54278 GB
 203294 RU
 377302 TR
```

C. `awk -F '\t' '$4=="Seafood Restaurant"' POIeu.txt|cut -f 5 |sort|uniq -c | sort -n`

Put out all the rows in the file whose fourth column is Seafood Restaurant, and the fifth column will be captured, then sort, remove the repetition and count the number of repetitions, and finally get the number of rows.

Turkey(TR) has most Seafood restaurants with 1522.

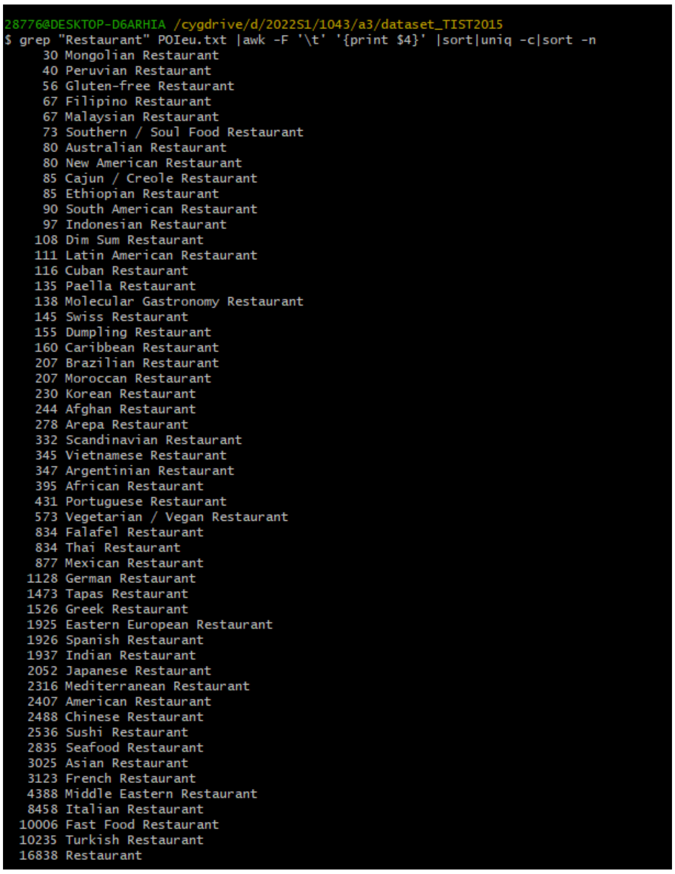


```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ awk -F '\t' '$4=="Seafood Restaurant"' POIeu.txt|cut -f 5 |sort|uniq -c | sort -n
      1 PL
      2 BY
      2 CH
      2 EE
      2 FI
      4 AZ
      4 JO
      5 BH
      5 LV
      6 BG
      6 CZ
      6 DK
      6 HU
      6 RO
      7 IE
     10 MA
     10 QA
     11 TN
     15 SE
     16 AT
     22 LB
     25 CY
     26 UA
     30 IL
     31 EG
     39 FR
     49 AE
     50 KW
     57 PT
     63 BE
     75 RU
     75 SA
     76 DE
     94 NL
    108 GB
    110 GR
    123 ES
    134 IT
   1522 TR
```

D. `grep "Restaurant" POIeu.txt |awk -F '\t' '{print $4}' |sort|uniq -c|sort -n`

Found out all restaurant type and print them out with numbers

“Restaurant” is most common class of restaurant in Europe with 16838.



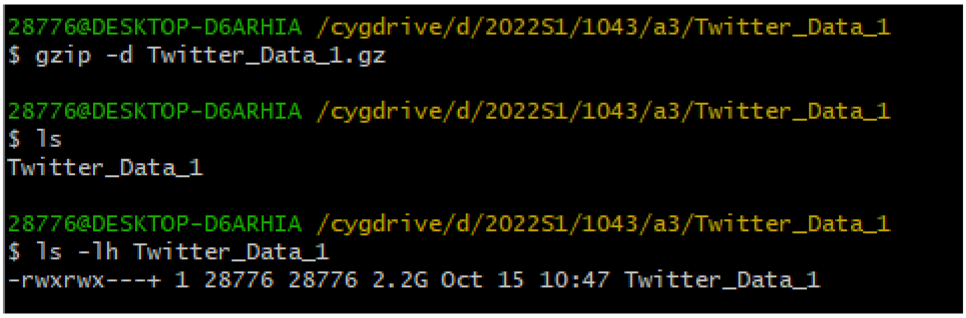
```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ grep "Restaurant" POIeu.txt |awk -F '\t' '{print $4}' |sort|uniq -c|sort -n
     30 Mongolian Restaurant
     40 Peruvian Restaurant
     56 Gluten-free Restaurant
     67 Filipino Restaurant
     67 Malaysian Restaurant
     73 Southern / Soul Food Restaurant
     80 Australian Restaurant
     80 New American Restaurant
     85 Cajun / Creole Restaurant
     85 Ethiopian Restaurant
     90 South American Restaurant
     97 Indonesian Restaurant
    108 Dim Sum Restaurant
    111 Latin American Restaurant
    116 Cuban Restaurant
    135 Paella Restaurant
    138 Molecular Gastronomy Restaurant
    145 Swiss Restaurant
    155 Dumpling Restaurant
    160 Caribbean Restaurant
    207 Brazilian Restaurant
    207 Moroccan Restaurant
    230 Korean Restaurant
    244 Afghan Restaurant
    278 Arepa Restaurant
    332 Scandinavian Restaurant
    345 Vietnamese Restaurant
    347 Argentinian Restaurant
    395 African Restaurant
    431 Portuguese Restaurant
    573 Vegetarian / Vegan Restaurant
    834 Falafel Restaurant
    834 Thai Restaurant
    877 Mexican Restaurant
   1128 German Restaurant
   1473 Tapas Restaurant
   1526 Greek Restaurant
   1925 Eastern European Restaurant
   1926 Spanish Restaurant
   1937 Indian Restaurant
   2052 Japanese Restaurant
   2316 Mediterranean Restaurant
   2407 American Restaurant
   2488 Chinese Restaurant
   2536 Sushi Restaurant
   2835 Seafood Restaurant
   3025 Asian Restaurant
   3123 French Restaurant
   4388 Middle Eastern Restaurant
   8458 Italian Restaurant
  10006 Fast Food Restaurant
  10235 Turkish Restaurant
  16838 Restaurant
```

Task B: Investigating the Twitter Data in the Shell and Graphing in R

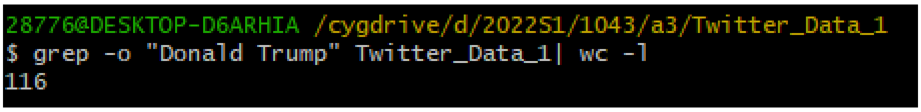
1) It appeared 116 times.

`gzip -d Twitter_Data_1.gz,ls,ls -lh` Unzip the files, look up file, get file size

`grep -o "Donald Trump" Twitter_Data_1 | wc -l` Look for "Donald Trump" in the file, that is, the grep command uses the -o parameter to convert rows into columns, and then statistics.



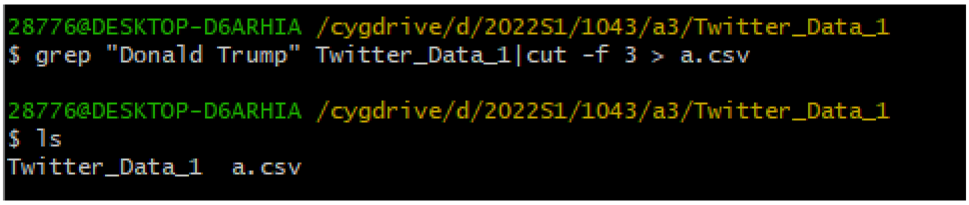
```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/Twitter_Data_1
$ gzip -d Twitter_Data_1.gz

28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/Twitter_Data_1
$ ls
Twitter_Data_1

28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/Twitter_Data_1
$ ls -lh Twitter_Data_1
-rwxrwx---+ 1 28776 28776 2.2G Oct 15 10:47 Twitter_Data_1
```



```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/Twitter_Data_1
$ grep -o "Donald Trump" Twitter_Data_1| wc -l
116
```

2) grep finds the line containing "Donald Trump" in the file, intercepts the third column(timestamps) and exports it to csv.

```
grep "Donald Trump" Twitter_Data_1 | cut -f 3 > a.csv
```

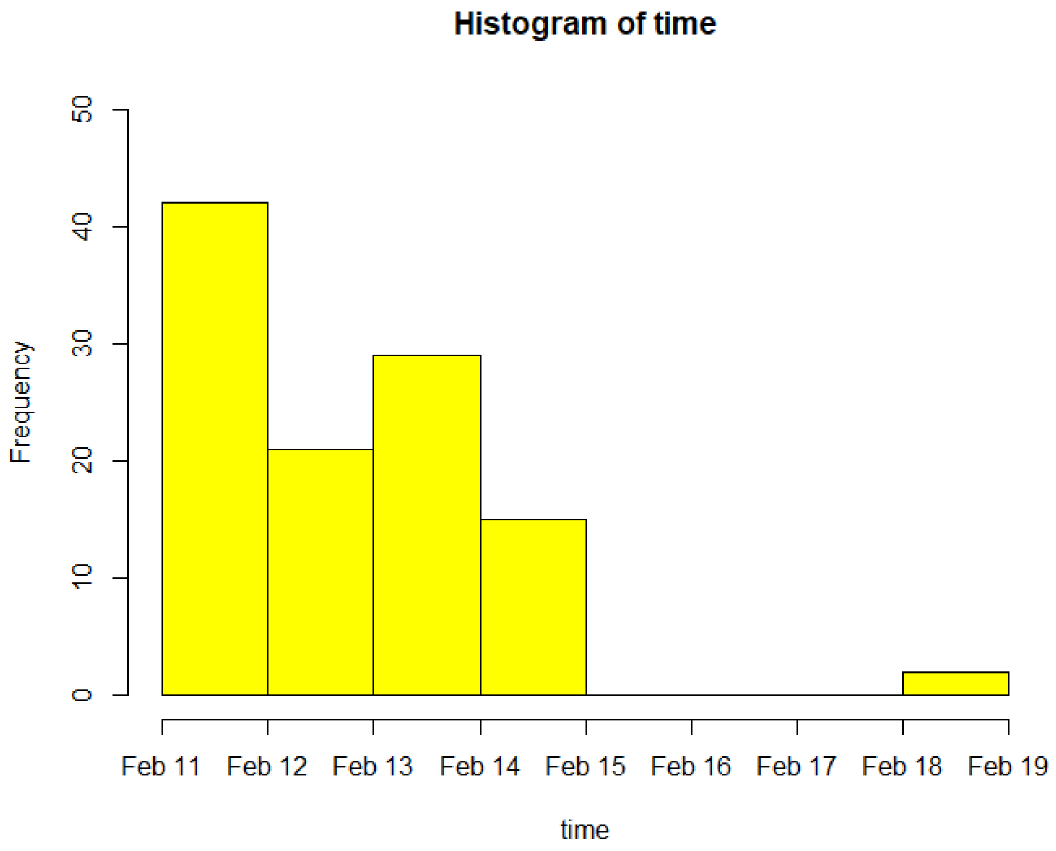


```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/Twitter_Data_1
$ grep "Donald Trump" Twitter_Data_1|cut -f 3 > a.csv

28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/Twitter_Data_1
$ ls
Twitter_Data_1  a.csv
```

Read the data in a.csv with R and change it according to the format of the timestamp.

```
setwd("D:/2022S1/1043/a3/Twitter_Data_1")
Sys.setlocale("LC_TIME", "C")
twitter <- read.csv("a.csv", header = F)
twitter$V1 <- strptime(twitter$V1, format = "%a %b %d %H:%M:%S %z %Y", tz = 'UTC')
```

3) hist(twitter$V1,"days",xlab = "time",col = "yellow",freq = T,ylim = c(0,50))

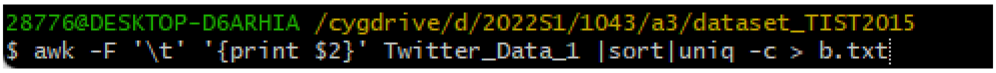


4) It can be seen from the figure (Q3) that the data before February 15 had the largest number of occurrences on February 11 (more than 40 times), followed by February 13 (less than 30 times), February 12 (more than 20 times), February 14 (less than 20 times). There were no tweets about Donald Trump for three days(Feb 15,16,17), and then two more tweets about him on February 18.

5)

Pull out all the user data in the second column of twitter data, then sort, remove the repetition and count the number of repetitions, and finally get the number of rows. Finally, put the data into b.txt.

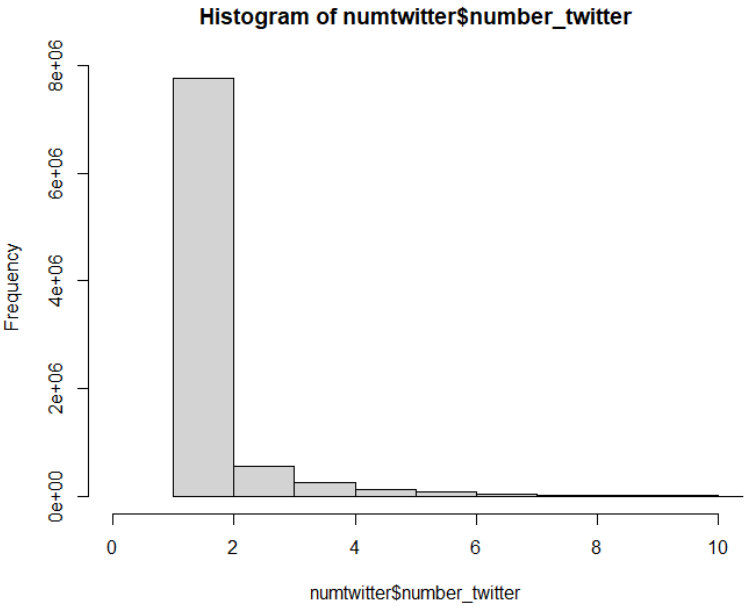
```
awk -F '\t' '{print $2}' Twitter_Data_1 |sort | uniq -c > b.txt
```



```
28776@DESKTOP-D6ARHIA /cygdrive/d/2022S1/1043/a3/dataset_TIST2015
$ awk -F '\t' '{print $2}' Twitter_Data_1 |sort|uniq -c > b.txt
```

```
numtwitter <- read.table("b.txt",fill = TRUE, head = FALSE) # read the txt as a table
names(numtwitter)<- c("number_twitter","id")# rename of each columns
max(numtwitter$number_twitter) # find the max freq
hist(numtwitter$number_twitter,breaks = 243,freq = T, xlim = c(0,10)) # create histogram
```

Change the names of the two columns in the txt to “number-twitter” and “id”, max() get the max number of twitter.


Histogram of numtwitter$number_twitter
Frequency
0e+00
2e+06
4e+06
6e+06
8e+06
numtwitter$number_twitter