

Task A

```
Edward Priyatna@LAPTOP-GG6FAS2B ~/HW3
$ ls -lh
total 4.6G
-rwxr-xr-x 1 Edward Priyatna None 2.3G Oct 18 13:04 dataset_TIST2015.tar
-rwxr-xr-x 1 Edward Priyatna None 2.1G Oct 6 18:53 dataset_TIST2015_Checkins_v2.txt
-rwxr-xr-x 1 Edward Priyatna None 25K Aug 12 2015 dataset_TIST2015_Cities.txt
-rwxr-xr-x 1 Edward Priyatna None 222M Aug 12 2015 dataset_TIST2015_POIs.txt
-rwxr-xr-x 1 Edward Priyatna None 2.0K Oct 6 18:59 dataset_TIST2015_readme_v2.txt
```

1.

First unzip the file

ls-lh #ls list name of file, lh tells size of file

The size of the data files are 2.1GB, 25KB, 222MB and 2.0KB

```
user_id venue_id UTC_time timezone_offset
50756 4f5e3a72e4b053fd6a4313f6 Tue Apr 03 18:00:06 +0000 2012 240
190571 4b4b87b5f964a5204a9f26e3 Tue Apr 03 18:00:07 +0000 2012 180
221021 4a85b1b3f964a520eefe1fe3 Tue Apr 03 18:00:08 +0000 2012 -240
66981 4b4606f2f964a520751426e3 Tue Apr 03 18:00:08 +0000 2012 -300
21010 4c2b4e8a9a559c74832f0de2 Tue Apr 03 18:00:09 +0000 2012 240
28761 4b4bade2f964a520cfa326e3 Tue Apr 03 18:00:09 +0000 2012 -240
39350 49bbd6c0f964a520f4531fe3 Tue Apr 03 18:00:09 +0000 2012 -240
1446 4e88cf4ed22d53877981fdb Tue Apr 03 18:00:09 +0000 2012 -300
82296 4dfc825bc65b31579b2e7679 Tue Apr 03 18:00:11 +0000 2012 180
```

2.

head dataset_TIST2015_Checkins_v2.txt | less #pipe into less to see the file

Then press /tab button

Tabs are used and there are 4 columns

```
Edward Priyatna@LAPTOP-GG6FAS2B ~/HW3
$ head -1 dataset_TIST2015_Checkins_v2.txt
user_id venue_id UTC_time timezone_offset
```

3.

head -1 dataset_TIST2015_Checkins_v2.txt #see the header

the column names are user_id, venue_id, UTC_time and timezone_offset

```
Edward Priyatna@LAPTOP-GG6FAS2B ~/HW3_1
$ sort dataset_TIST2015_Checkins_v2.txt | uniq | wc -l
33253305
```

4.

sort dataset_TIST2015_Checkins_v2.txt | uniq | wc -l #use this because there are some duplicate rows

use uniq to find unique value. Count how many unique value with wc -l.

```
Edward Priyatna@LAPTOP-GG6FAS2B ~/HW3_1
$ cut -f 1 dataset_TIST2015_Checkins_v2.txt | sort | uniq | wc -l
266910
```

cut -f 1 dataset_TIST2015_Checkins_v2.txt | sort | uniq | wc -l

#cut to column 1 user id then sort them then use uniq to find unique value. Count how many unique value with wc -l.

There are 266910 unique user id (including the header), so there are 266909 user id.

```
Edward Priyatna@LAPTOP-GG6FAS2B ~/HW3
$ 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
```

5.

head -2 dataset_TIST2015_Checkins_v2.txt

#see the head which is beginning of the file. -2 because -1 is header.

April 3 2012 is the first date

```
Edward Priyatna@LAPTOP-GG6FAS2B ~/HW3
$ tail -1 dataset_TIST2015_Checkins_v2.txt
22704 50df4ee5e4b0c48b5a1c2968 Mon Sep 16 23:24:15 +0000 2013 180
```

tail -1 dataset_TIST2015_Checkins_v2.txt #see the tail which is end of file

September 16 2013

```
Edward Priyatna@LAPTOP-GG6FAS2B ~/HW3
$ head -1 dataset_TIST2015_POIs.txt
3fd66200f964a5200e71ee3      40.733596      -74.003139      Jazz Club      US

Edward Priyatna@LAPTOP-GG6FAS2B ~/HW3
$ cut -d' ' -f1 dataset_TIST2015_POIs.txt | sort -u | wc -l
3680126
```

6. Head -1 dataset_TIST2015_POIs.txt #see the file so we can know which column is venue id
 Cut -d' ' -f1 dataset_TIST2015_POIs.txt | sort -u | wc -l
 #cut to column 1 then sort the unique value then count how many line are there
 There are 3680126 unique venue ID.

```
Edward Priyatna@LAPTOP-GG6FAS2B ~/HW3_1
$ grep "FR" dataset_TIST2015_POIs.txt|cut -f 4 | sort | uniq | wc -l
384
```

7. grep "FR" dataset_TIST2015_POIs.txt|cut -f 4 | sort | uniq | wc -l
 #search for country code FR, then cut to column 4, then sort the values, then uniq, then count how many lines are there
 There are 384 unique venue categories in France.

8. A. awk -F "\t" '{ if(\$2>=35 && \$2<=72 && \$3>=-25 && \$3<=45) {print}}' dataset_TIST2015_POIs.txt > POIeu.txt
 #use awk, file is delimited by tab. Column 2 between 35 to 72. Column 3 between -25 to 75. Because that is the latitude and longitude of Europe. Then {print} to print all the values. > to write the data to a new txt.

```
Edward Priyatna@LAPTOP-GG6FAS2B ~/HW3_1
$ awk -F '\t' '{print $5}' eu_seafood.txt | sort | uniq -c | sort
1 PL
2 BY
2 CH
2 EE
2 FI
5 LV
6 BG
6 CZ
6 DK
6 HU
6 RO
7 IE
11 TN
15 SE
16 AT
20 CY
26 UA
39 FR
57 PT
63 BE
64 RU
76 DE
94 NL
108 GB
110 GR
123 ES
134 IT
1522 TR
```

- B. awk -F "\t" '{print \$5}' POIeu.txt | sort | uniq -c | sort
 #use awk, file is delimited by tab. Print column 5. Pipe the result to sort, sort the value. Pipe to uniq -c count frequency of each unique value. Then pipe to sort to be sorted.
 Turkey has the most venues

```
Edward Priyatna@LAPTOP-GG6FAS2B ~/HW3_1
$ awk -F '\t' '{print $5}' eu_seafood.txt | sort | uniq -c | sort
1 PL
2 BY
2 CH
2 EE
2 FI
5 LV
6 BG
6 CZ
6 DK
6 HU
6 RO
7 IE
11 TN
15 SE
16 AT
20 CY
26 UA
39 FR
57 PT
63 BE
64 RU
76 DE
94 NL
108 GB
110 GR
123 ES
134 IT
1522 TR
```

C.

grep -w 'Seafood Restaurant' POleu.txt > eu_seafood.txt #grep -w search for 'Seafood Restaurant'. Then > put it to eu_seafood.txt
 awk -F '\t' '{print \$5}' eu_seafood.txt | sort | uniq -c | sort
 # use awk, file is delimited by tab. Print column 5 Pipe the result to sort, sort the value. Pipe to uniq -c count frequency of each unique value. Then pipe to sort to be sorted.

```
Edward.Priyatna@LAPTOP-GG6FAS2B ~/HW3_1
$ awk -F '\t' '{print $4}' eu_restaurant.txt | sort | uniq -c | sort
  23 Filipino Restaurant
  27 Mongolian Restaurant
  37 Peruvian Restaurant
  51 Gluten-free Restaurant
  53 Malaysian Restaurant
  54 New American Restaurant
  57 Southern / Soul Food Restaurant
  67 Australian Restaurant
  67 Indonesian Restaurant
  72 Cajun / Creole Restaurant
  77 Ethiopian Restaurant
  89 South American Restaurant
  95 Cuban Restaurant
  95 Latin American Restaurant
  96 Dim Sum Restaurant
 126 Molecular Gastronomy Restaurant
 129 Dumpling Restaurant
 130 Paella Restaurant
 137 Caribbean Restaurant
 137 Swiss Restaurant
 150 Moroccan Restaurant
 179 Afghan Restaurant
 181 Arepa Restaurant
 185 Brazilian Restaurant
 207 Korean Restaurant
 313 Argentinian Restaurant
 322 African Restaurant
 326 Scandinavian Restaurant
 338 Vietnamese Restaurant
 422 Portuguese Restaurant
 522 Vegetarian / Vegan Restaurant
 580 Falafel Restaurant
 753 Thai Restaurant
 756 Mexican Restaurant
1096 German Restaurant
1288 American Restaurant
1371 Indian Restaurant
1456 Tapas Restaurant
1497 Greek Restaurant
1689 Eastern European Restaurant
1749 Japanese Restaurant
1911 Spanish Restaurant
2072 Mediterranean Restaurant
2122 Sushi Restaurant
2219 Chinese Restaurant
2411 Asian Restaurant
2529 Seafood Restaurant
2713 Middle Eastern Restaurant
2861 French Restaurant
7659 Italian Restaurant
8632 Fast Food Restaurant
10093 Turkish Restaurant
15196 Restaurant
```

D.
 grep -w 'Restaurant' POleu.txt > eu_restaurant.txt # grep -w search for 'Restaurant'. Then > put it to eu_restaurant.txt
 awk -F '\t' '{print \$4}' eu_restaurant.txt | sort | uniq -c | sort
 #Use awk, file is delimited by tab. Print column 4. Then pipe to sort where it sort the values. Pipe to uniq -c counts the frequency of each unique values. Pipe to sort to sort the values.
 Restaurant appears the most. But 'Restaurant' might mean unclassified restaurant. So maybe the correct answer is 'Turkish Restaurant'.

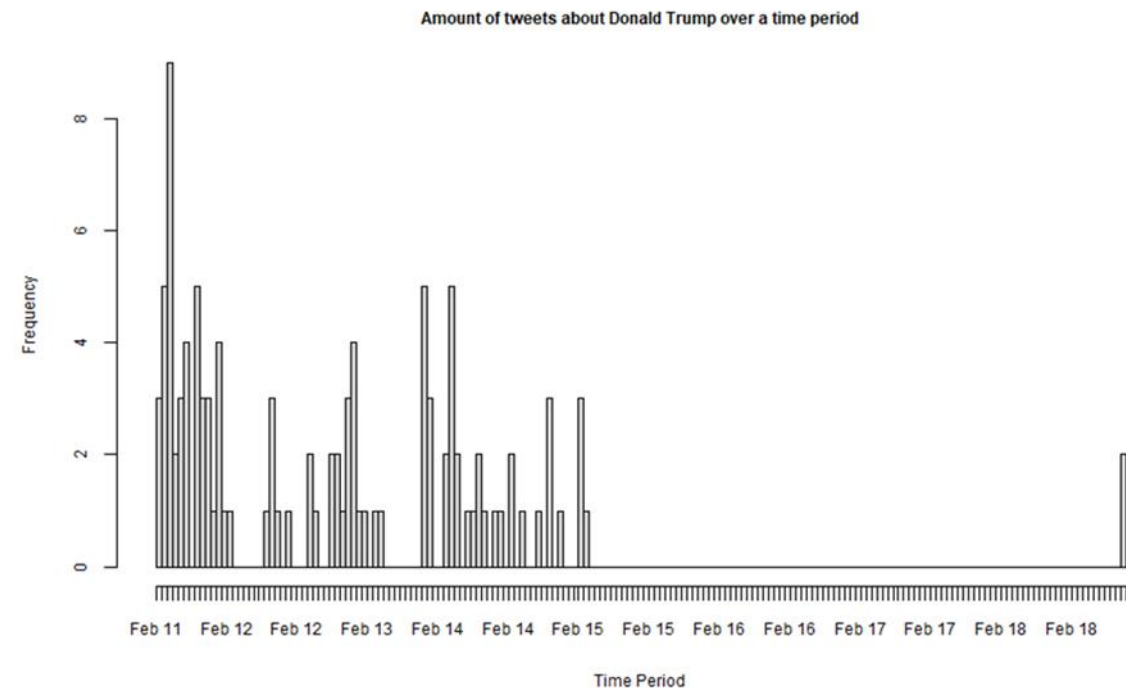
Task B

- ```
Edward.Priyatna@LAPTOP-GG6FAS2B ~/HW3_2
$ grep -o 'Donald Trump' Donald_Trump.txt | wc -l
114
```

grep -w 'Donald Trump' Twitter\_Data\_1 > Donald\_Trump.txt  
 #grep searches for 'Donald Trump' w means only when the whole word matches. > to pass it to new txt Donald\_Trump.txt  
 grep -o 'Donald Trump' Donald\_Trump.txt | wc -l
- taf = read.table("timestamps.txt",header=FALSE,sep=",")  
 Read.table will read the timestamp.csv file into a dataframe and store it in a variable (here is taf), header=FALSE is to indicate that there is no header, sep=',' to represent the delimiter of the csv file

taf[['V1']] <- strptime(taf[['V1']],format="%a %b %d %T %z %Y")

# strptime to convert the timestamps into datetime format. The format is telling the program that the column is following certain format. After converting it, save it back to the original column.



3. `par(mar=c(5, 4, 2, 1))` #setting margin so the graph is nice

```
hist(taf[["V1"]], breaks='hours', main='Amount of tweets about Donald Trump over a time period', freq=TRUE, xlab='Time Period', cex.main=0.75, cex.lab=0.75, cex.axis=0.75)
```

# hist is to create a histogram by taking a column (here is taf[["V1"]]), breaks is to set the number of bins (day means to break by the date of the dataset), main is the title of the histogram. freq is to set the number of frequency to be a integer, xlab is to set the name of x-axis, the last three code is to adjust the size of title, the name of x-axis, and axes.

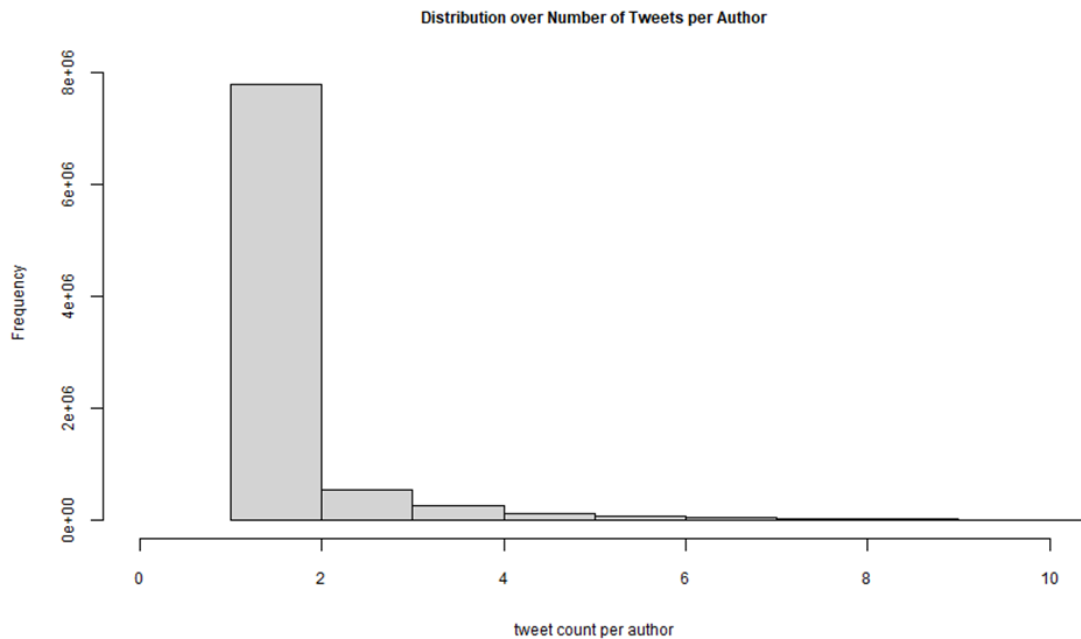
4. 

```
Tue Feb 11 12:28:36 +0000 2014 RT @aadan_smith: Be interesting to see the detail on this one: BBC News - Donald Trump loses offshore wind farm challenge http://t.co/qACG...
Tue Feb 11 12:47:26 +0000 2014 RT @havantaclu07MP: Donald Trump loses legal challenge to windfarm near his Scottish golf resort http://t.co/30QpW/hpA4 via @guardian
Tue Feb 11 12:55:09 +0000 2014 RT @thescottsman: Donald Trump loses Aberdeen Bay wind farm legal challenge: http://t.co/C0232Vend
Tue Feb 11 13:22:29 +0000 2014 RT @businessgreen: Breaking: Donald Trump trumped over offshore wind farm challenge http://t.co/qy12FL55Rn
Tue Feb 11 13:24:31 +0000 2014 Donald Trump vangt bot in windmolenzaak http://t.co/kyw1qei0UW #nederland #noordholland #haarlem
Tue Feb 11 13:34:00 +0000 2014 Breaking: Donald Trump trumped over offshore wind farm challenge: Scottish courts reject petition for judi... http://t.co/nbqghziy04
Tue Feb 11 13:46:21 +0000 2014 RT [http://t.co/y54BVzts3] voir cette histoire: Donald Trump perd sa bataille juridique en Ecosse et de votre parcours sera un parc eolien.
Tue Feb 11 13:59:18 +0000 2014 Ha ha Donald Trump!
Tue Feb 11 14:01:15 +0000 2014 Donald Trump loses legal challenge to windfarm near his Scottish golf resort - http://t.co/cP5pcU003o
Tue Feb 11 14:01:48 +0000 2014 and in other news... Donald Trump may have to golf in Scotland with a view of an experimental windfarm. "Sad face" http://t.co/S8n0EQsDsr
Tue Feb 11 14:10:46 +0000 2014 RT @sunny_hundal: World's smallest violin for Donald Trump - he's lost a legal challenge to a windfarm near his Scottish golf resort http://t.co/LJKJC0d9X5
Tue Feb 11 14:38:52 +0000 2014 Kasus kincir angin, Donald Trump kalah: Pengadilan Amerika Serikat Donald Trump kalah dalam sidang di Skotlandi... http://t.co/PX1bERcrv7
Tue Feb 11 14:40:57 +0000 2014 Kasus kincir angin, Donald Trump kalah http://t.co/0R7K0Gy79o
Tue Feb 11 14:47:08 +0000 2014 Confirmation that Donald Trump is the new owner of the Doonbeg Resort in Co. Clare
Tue Feb 11 14:47:49 +0000 2014 RT @badger5000: HAHAAHA, FUCK YOU WIGGY BELLEND @realdonaldtrump RT @BBCJamesCook: Donald Trump loses wind farm battle: http://t.co/uVvr19F8...
Tue Feb 11 14:58:25 +0000 2014 RT @guardianeco: Donald Trump loses legal challenge to windfarm near his Scottish golf resort http://t.co/LJKJC0d9X5
Tue Feb 11 14:58:52 +0000 2014 Donald Trump acts more and more like a movie super villain every day http://t.co/8e5aPn2VVA
Tue Feb 11 15:05:37 +0000 2014 Aberdeen genius at it's best: site Donald Trump at Balmiedie, then build a wind farm to capture the energy... http://t.co/vuzzhPmeRX
Tue Feb 11 15:36:12 +0000 2014 Green News: Breaking: Donald Trump trumped over offshore wind farm challenge http://t.co/NzkF7n4u6i
Tue Feb 11 16:01:43 +0000 2014 RT @clarefm: CONFIRMED: The Trump Organisation, owned by Donald Trump, has purchased The Lodge at Doonbeg. More to come on @clarefm news at...
Tue Feb 11 16:28:19 +0000 2014 selomerece Donald Trump pierde batalla legal en Escocia http://t.co/VrkGhu12id
Tue Feb 11 16:39:24 +0000 2014 Does anyone feel bad for Donald Trump? http://t.co/3f1lDPmPm
Tue Feb 11 17:03:47 +0000 2014 Donald Trump pierde batalla legal en Escocia http://t.co/1QAmLUtW6t
Tue Feb 11 17:06:30 +0000 2014 IF Donald Trump wanted to stop that wind farm he should have played his... Trump card!!! welcome to Reporting Scotland my name's Jackie Bird
Tue Feb 11 17:36:11 +0000 2014 RT @GerryBriden: Donald Trump's world-famous Irish course, previously Doonbeg, will be now renamed Trump International Golf Links, Irela...
Tue Feb 11 17:37:22 +0000 2014 Donald Trump loses court battle against offshore wind farm, quite right to... http://t.co/Eq3HP83F0t
Tue Feb 11 19:19:47 +0000 2014 RT @newstalkfm: US billionaire Donald Trump buys #Doonbeg resort in Clare http://t.co/PX1pDQ2QVI http://t.co/Cxtuyhdj0E
Tue Feb 11 19:44:00 +0000 2014 Donald Trump loses wind farm ruling http://t.co/A0ebsir359
Tue Feb 11 19:51:21 +0000 2014 When asked if direct sales is a pyramid scheme, my reply is a corporation has only one person at the top. - Donald Trump
Tue Feb 11 19:53:10 +0000 2014 Donald Trump loses wind farm ruling: A legal challenge to a planned offshore wind farm, which could be seen fro... http://t.co/n3zphe6RvU
Tue Feb 11 19:57:09 +0000 2014 Roban miles de dólares a dos visitantes del resort de Donald Trump en Doral http://t.co/V0au2r5xxc
Tue Feb 11 20:11:26 +0000 2014 #NowPlaying - Mac Miller - Donald Trump
Tue Feb 11 20:13:27 +0000 2014 Donald Trump is one of the dumbest people on the planet
Tue Feb 11 20:22:12 +0000 2014 Donald Trump buys Irish golf resort after losing Scotland court battle: US billionaire says he'll divert energ... http://t.co/g5EXip5Gaw
Tue Feb 11 21:11:49 +0000 2014 RT @Golfchannel: From Doral to Doonbeg: Donald Trump purchases Greg Norman-designed Doonbeg Golf Club in Ireland: http://t.co/hvR1R0JHJn
Tue Feb 11 21:24:24 +0000 2014 Donald Trump pierde batalla legal en Escocia http://t.co/0rDwju4RB
Tue Feb 11 21:31:39 +0000 2014 #VOFNUK: Donald Trump has purchased Doonbeg Golf Resort in Ireland, featuring a Greg Norman links ... http://t.co/m2cK2n5J13
Tue Feb 11 22:13:01 +0000 2014 Donald Trump buys Irish golf resort http://t.co/TmBUgMXEX
Tue Feb 11 23:14:38 +0000 2014 As governor, Donald Trump would overhaul Tappan Zee bridge for "peanuts" | Bedford NY Real Estate http://t.co/S8hfQY7QNR
Tue Feb 11 23:19:25 +0000 2014 RT @CelebInsults: Donald Trump vs Hoes http://t.co/dwFqt51da
Tue Feb 11 23:25:27 +0000 2014 RT @grafikmafia: Hahahah, fuck you Donald Trump
Tue Feb 11 23:29:26 +0000 2014 Tegenslag voor Donald Trump in strijd tegen windmolens - Buitenland - AD http://t.co/aMWBndQiv Denkt dat alles voor geld moet wijken
Wed Feb 12 00:00:44 +0000 2014 @TrillWolendezzz So you jankin with yo Donald Trump head shhhh?
Wed Feb 12 01:04:52 +0000 2014 RT @Oireachtas_RX: Donald Trump surveys his new golf course http://t.co/FOck1SEo3P
Wed Feb 12 08:31:32 +0000 2014 RT @magnus11ewellin: As Donald Trump loses his wind farm legal battle our man Camley puts pen to paper . . . http://t.co/MJARwARP5X
Wed Feb 12 09:25:14 +0000 2014 RT @magnus11ewellin: As Donald Trump loses his wind farm legal battle our man Camley puts pen to paper . . . http://t.co/MJARwARP5X
Wed Feb 12 09:41:41 +0000 2014 Donald Trump loses legal battle against offshore wind project - GOOD! http://t.co/t2c4uX2ed via @edie
Wed Feb 12 09:52:51 +0000 2014 Windfarms 1, Donald Trump 0
Wed Feb 12 10:48:24 +0000 2014 RT @magnus11ewellin: As Donald Trump loses his wind farm legal battle our man Camley puts pen to paper . . . http://t.co/MJARwARP5X
Wed Feb 12 12:13:16 +0000 2014 RT @JameyGodley: Defeated Donald Trump turns his back on Scotland - The Scotsman: http://t.co/tQouHz8aHR
```

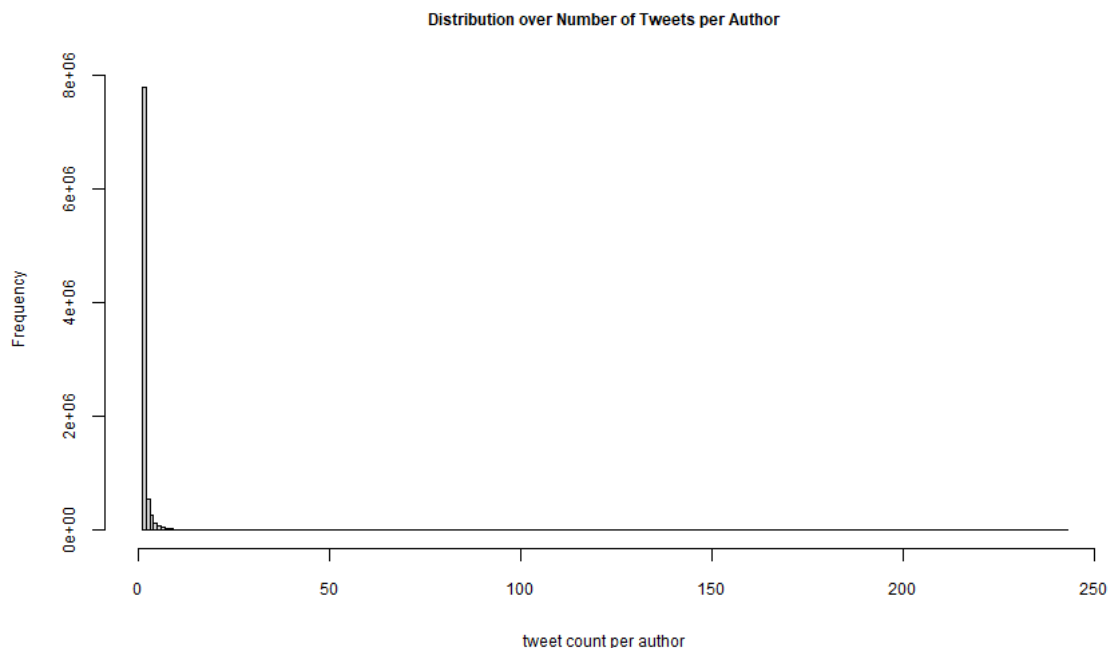
`cut -f 3,4 Donald_Trump.txt | less` #pipe the file to less to see the file then type `/wind farm`

From the image we can see a lot of wind farms before February 15. I searched the internet and found this article: <https://www.theguardian.com/world/2014/feb/11/donald-trump-loses-windfarm-scottish-golf-resort>. It seems that he opposes the building of a wind farm near his golf course. However, he lost the legal battle and a wind farm was build near his golf course.





5. `hist(tweet_count[['V1']],breaks=260,main='Distribution over Number of Tweets per Author',xlab='tweet count per author',freq=TRUE,cex.main=0.75, cex.lab=0.75, cex.axis=0.75, xlim=c(0,10) )`  
 # `hist` is to create a histogram by taking a column (here is `df2[['V1']]`), `breaks` is to set the number of bins, `main` is the title of the histogram. `freq = TRUE` is to set the number of frequency to be a integer, `xlab` is to set the name of x-axis, the last three code is to adjust the size of title, the name of x-axis and axes, and `xlim` limit the output of x-axis.



`hist(tweet_count[['V1']],breaks=260,main='Distribution over Number of Tweets per Author',xlab='tweet count per author',freq=TRUE,cex.main=0.75, cex.lab=0.75, cex.axis=0.75)`  
 # `hist` is to create a histogram by taking a column (here is `tweet_count[['V1']]`), `breaks` is to set the number of bins, `main` is the title of the histogram. `freq = TRUE` is to set the number of frequency to be a integer, `xlab` is to set the name of x-axis, the last three code is to adjust the size of title, the name of x-axis and axes.

This is the full graph. It is hard to see.

It is very hard to find the correct x and y limit of the graph because the graph drops drastically after `x=1`. Most twitter accounts only sends one tweet and never send another.