Project 1. Temporal Exploration of Tweets

Abstract

In this project, the main purpose of this project is to conduct analysis of twitter data and understand the basics of using R, in addition, I tried to do some geospatial analysis related this project.

1. Basic Operation

```
#read the csv files
pu2014 <- read.csv('Twitter Data File_pu2014.csv', header = TRUE)
#selecting specific column
pu2014c01 <- pu2014[, c('epoch', 'weekday', 'hour','month','day','year', 'longitude','latitude'
#run first five row of data
head(pu2014c01, 5)
#conduct a summary
summary(pu2014c01)
#convert epoch time to datetime.
pu2014c01$datetime <- as.POSIXct(pu2014c01$ep, origin = "1970-01-01")
head(pu2014c01, 5)
```

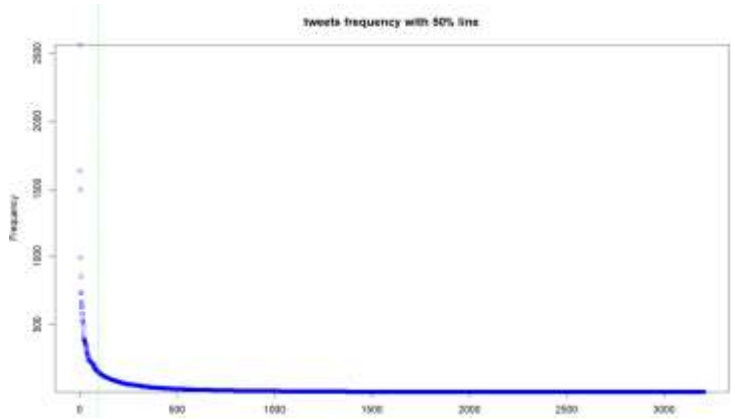
```
Min.   :1.389e+09   Length:68077       Min.   : 0.00   Length:68077       Min.   : 1.00
1st Qu.:1.391e+09   Class :character   1st Qu.:10.00   Class :character   1st Qu.: 9.00
Median :1.394e+09   Mode  :character   Median :15.00   Mode  :character   Median :18.00
Mean   :1.397e+09                      Mean   :14.14                      Mean   :16.58
3rd Qu.:1.398e+09                      3rd Qu.:20.00                      3rd Qu.:25.00
Max.   :1.420e+09                      Max.   :23.00                      Max.   :31.00
     year        longitude         latitude
Min.   :2014   Min.   :-87.00   Min.   :40.40
1st Qu.:2014   1st Qu.:-86.94   1st Qu.:40.43
Median :2014   Median :-86.92   Median :40.43
Mean   :2014   Mean   :-86.93   Mean   :40.44
3rd Qu.:2014   3rd Qu.:-86.92   3rd Qu.:40.45
Max.   :2014   Max.   :-86.89   Max.   :40.49
> #convert epoch time to datetime.
       epoch weekday hour month day year longitude latitude            datetime
1 1388552464     Wed    0   Jan   1 2014 -86.94425 40.47112 2014-01-01 00:01:04
2 1388552467     Wed    0   Jan   1 2014 -86.94266 40.44576 2014-01-01 00:01:07
3 1388552533     Wed    0   Jan   1 2014 -86.93918 40.47966 2014-01-01 00:02:13
4 1388552645     Wed    0   Jan   1 2014 -86.99292 40.45820 2014-01-01 00:04:05
5 1388552648     Wed    0   Jan   1 2014 -86.90060 40.42621 2014-01-01 00:04:08
>
```

2. Basic Statistics

```
#find distinct users in this table
distinct_user <- unique(pu2014c01$user_id)
num_distinct_user <- length(distinct_user)

cat('Distinct Twitter Users in this table:', num_distinct_user,"\n")
```

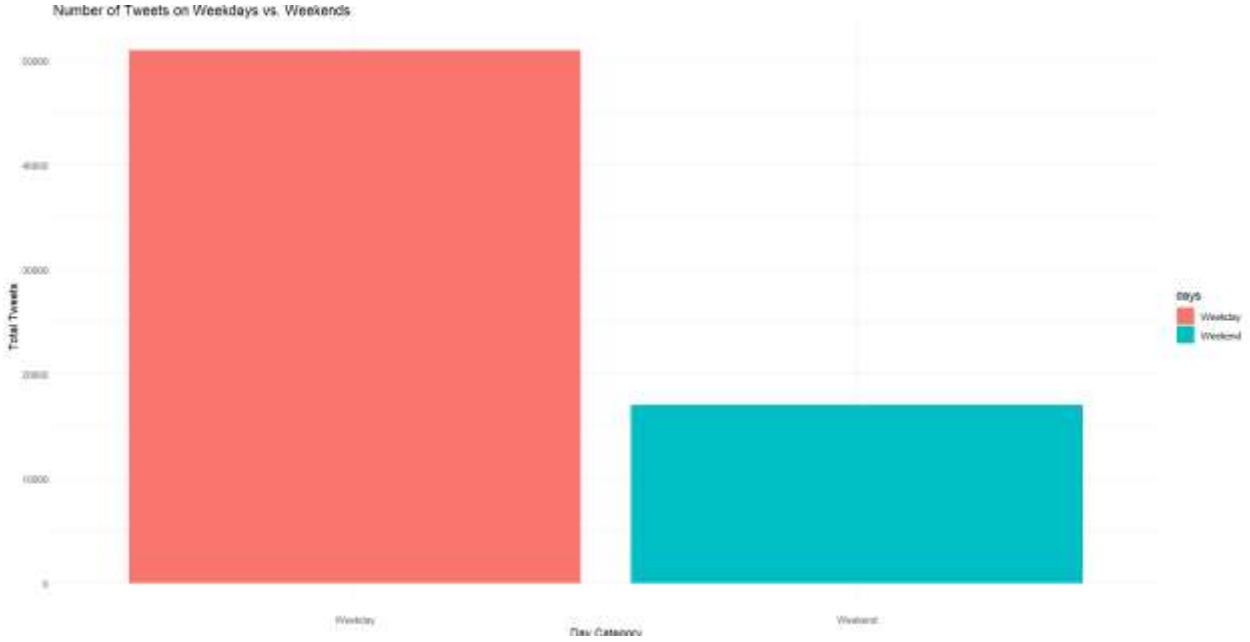
```
>
> cat('Distinct Twitter Users in this table:', num_distinct_user,"\n")
Distinct Twitter Users in this table: 3204
>
```



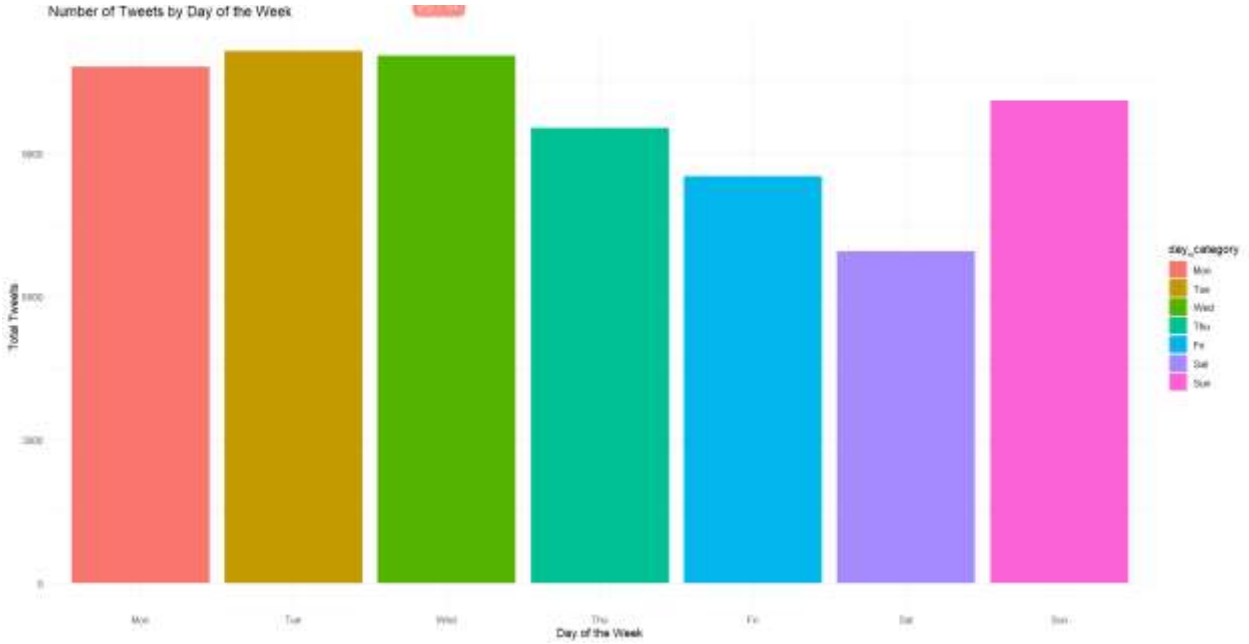
The graph showed the frequency of each user's tweets, in addition, it showed that most of the tweets were generated by a small number of users, the green line split 50 percent of the tweets based on the users. We can see that the line is skewed at the very left, which is close to the less users with massive tweets.

Additional codes are in the reference.

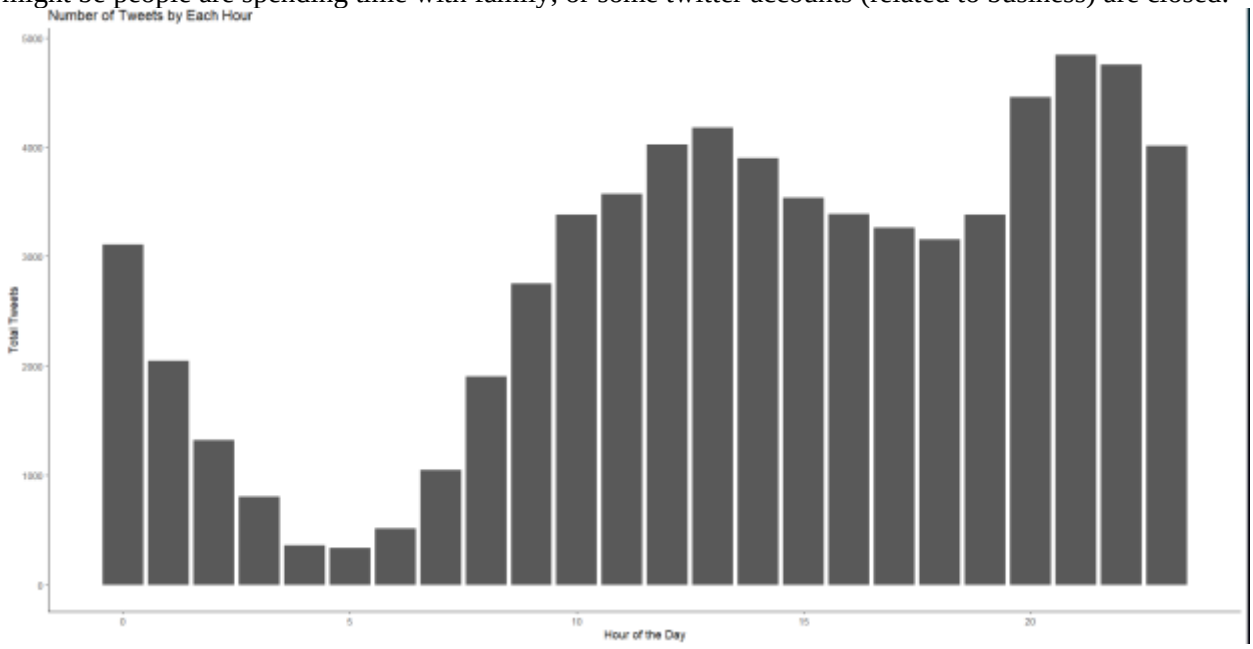
3. Tweeting behavior over time



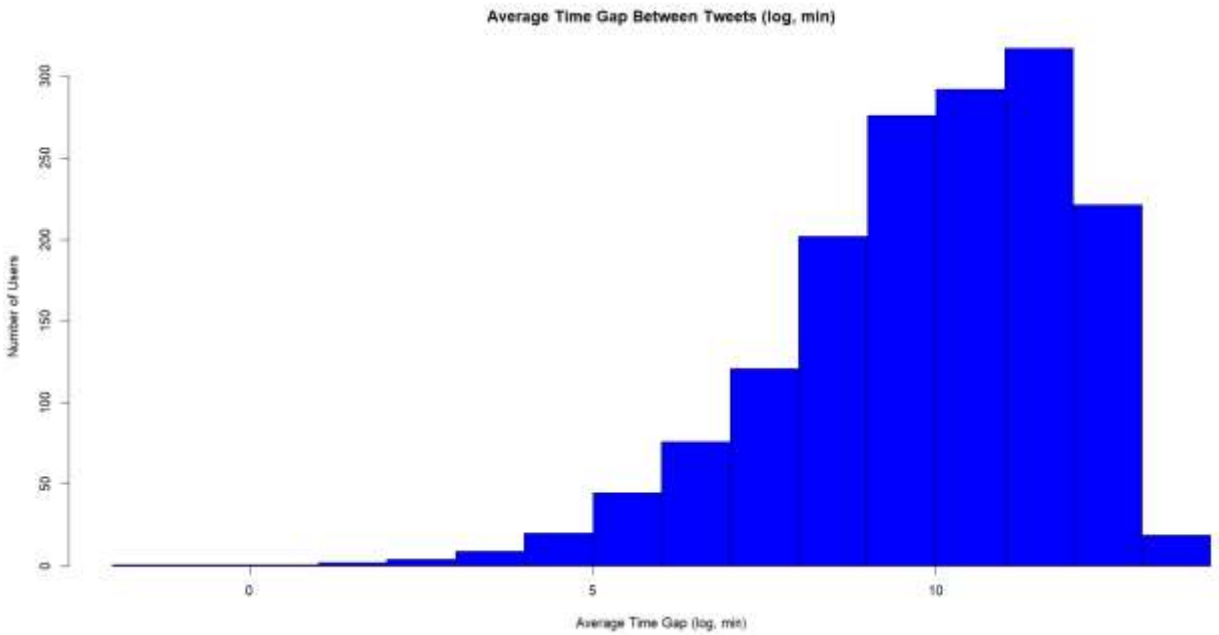
The first bar chart shows that weekdays have more tweets, however, weekdays have five days and weekends have only two days, additional analysis is required.



This graph shows the tweets of each day, it shows that Saturday has the least tweets, one possible reason might be people are spending time with family, or some twitter accounts (related to business) are closed.



Here is the graph showing the tweets for each hour, during the mid night, there are less tweets, also, there are more tweets during lunch time (around 12 pm) than dinner time (18 to 20). Maybe people eat and tweet during lunch time but they spend more time with family during dinner time.

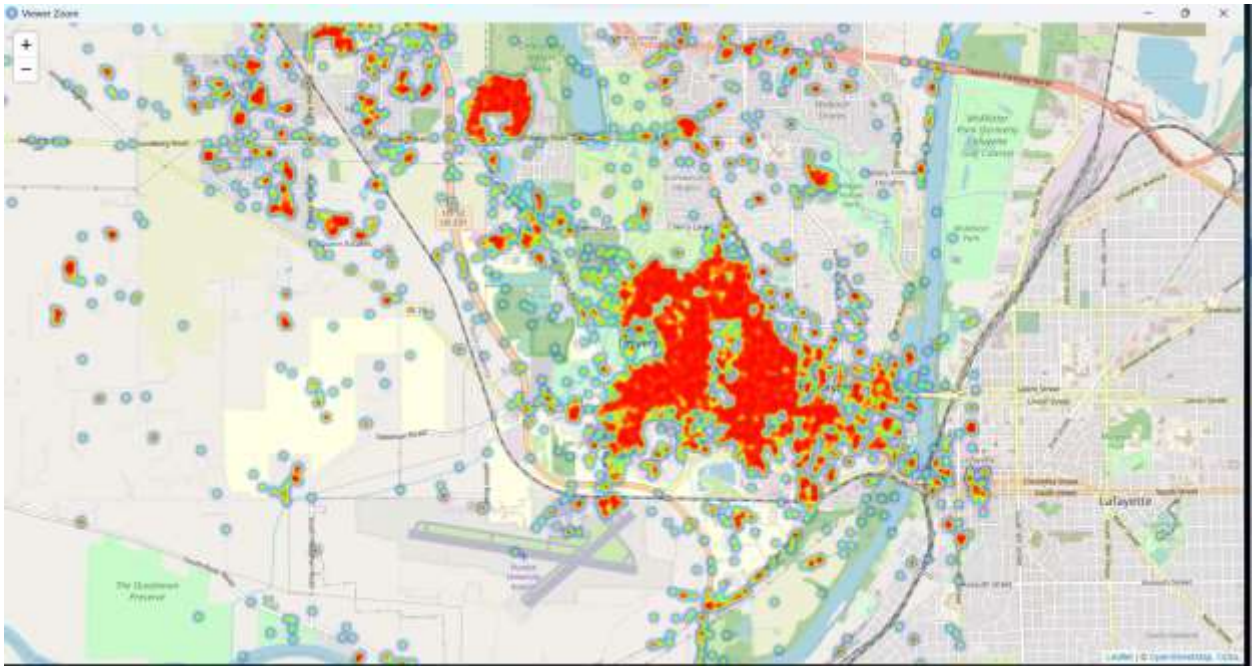


This graph shows the distribution of tweet gap for the users, as most users would have a longer gap between each of their tweets. Only a few are tweeting more frequently.

4. Summary/Conclusion/Concluding Remarks

I tried to do a heatmap for the location of tweets. Google Map would require API, OpenStreetMap was used in this part.

Here is the result graph showing the location where people tweet the most. Probably in the future analysis, more details shall be included within this map.



Acknowledgement

I would like to thank all the classmates who were present in the class on Wednesday which gave me a lot of ideas. In addition, the YouTube channel “thenewboston” has helped me a lot.

References

```
33
34  #show that small amount of people tweet the most tweets.
35  #draw the plot again
36  plot(usertweet_count$freq, col = 'blue', ylab = 'Frequency', xlab = 'user',
37       main = "tweets frequency with 50% line")
38
39  # calculate the cumulative sum of tweet frequencies
40  cumulative_freq <- cumsum(usertweet_count$freq)
41
42  # find the total amount of tweet
43  total_tweets <- sum(usertweet_count$freq)
44  # find the 50 percent of the tweet is at
45  threshold <- total_tweets * 0.5
46  index_50_percent <- which(cumulative_freq >= threshold)[1]
47
48  # Draw a vertical line at the 50% point
49  abline(v = index_50_percent, col = 'green', lty = 2)#dashed lines
50  #the above will show the line at the left side which represents that most tweets
51  #were did by the small amount of users.
```

```
#extra: Create a leaflet map
twitmap <- leaflet() %>%
  setView(lng = mean(pu2014c01$longitude), lat = mean(pu2014c01$latitude), zoom = 5)

#add the OpenStreetMap base layer
twitmap <- twitmap %>%
  addTiles()

#add the heatmap layer
twitmap <- twitmap %>%
  addHeatmap(
    data = pu2014c01,
    lat = ~latitude,
    lng = ~longitude,
    radius = 5,  #set the radius, can be changed in the future
    blur = 5    #set the blur, can be changed in the future
  )

#visualize the result on the map
```

```
#calculate the time gap between consecutive tweets for each user in seconds
pu2014c01$time_gap <- c(0, diff(pu2014c01$datetime))  # Use 0 for the first tweet of each user

#calculate the average time gap for each user in minutes
user_avg_time_gap <- aggregate(time_gap ~ user_id, data = pu2014c01, FUN = function(x) mean(x) / 60)

#take out zero or negative time gaps
user_avg_time_gap <- user_avg_time_gap[user_avg_time_gap$time_gap > 0, ]

#create a histogram to visualize the average time gaps in log scale
hist(log(user_avg_time_gap$time_gap), col = 'blue', xlab = 'Average Time Gap (log, min)',
     ylab = 'Number of Users', main = 'Average Time Gap Between Tweets (log, min)')
```