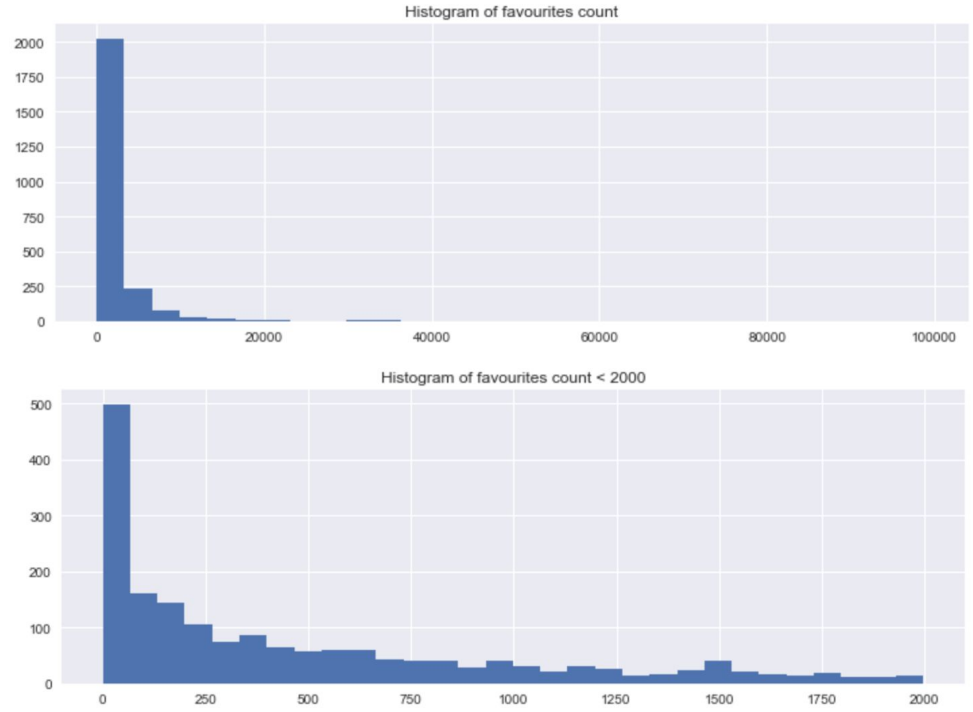
Part 1: Descriptive Tasks

Candidate: Kelly Peng
Date: 12/03/2017

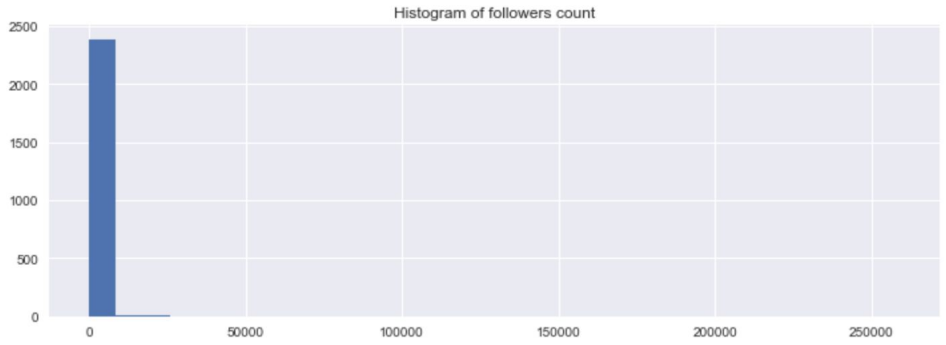
1.Make histograms of followers count, friends count, favorite count, and status count, all of which are in age_profiles.csv.

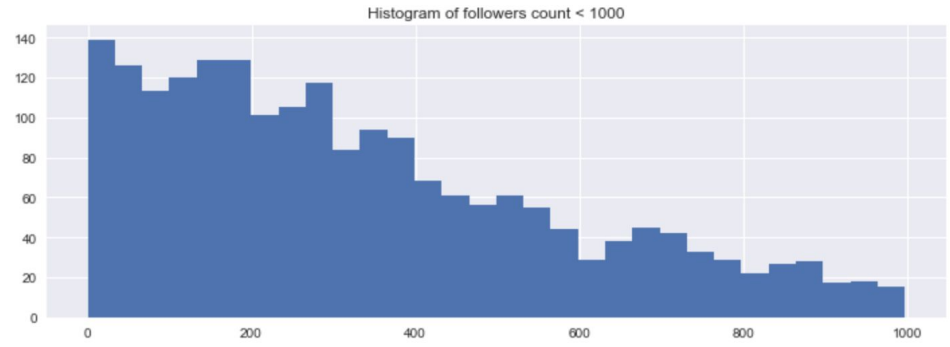
By looking at the description of data, we can see the distribution of followers count, friends count, favorite count, and status count are extremely right skewed. Very small number of user have extremely large number of followers, friends, etc. Thus in the histograms below, I plotted both the original histogram with all outliers included, as well as histograms that only includes around 75% of all data points.

1. Favourite count

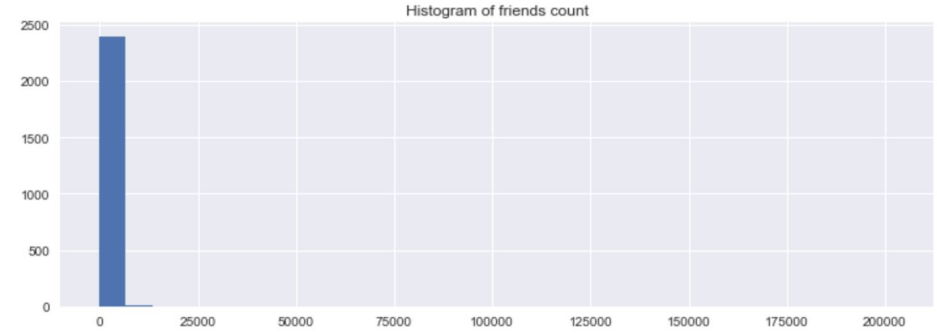


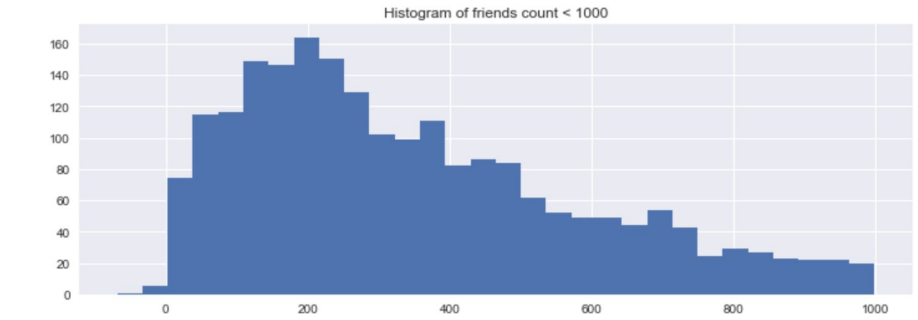
2. Followers count:



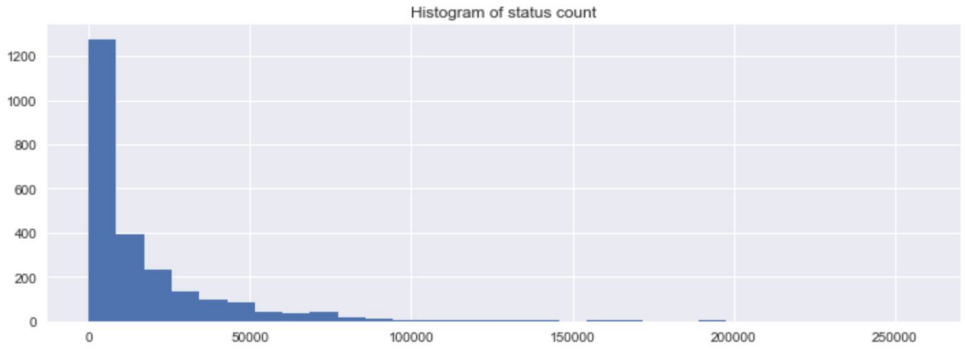


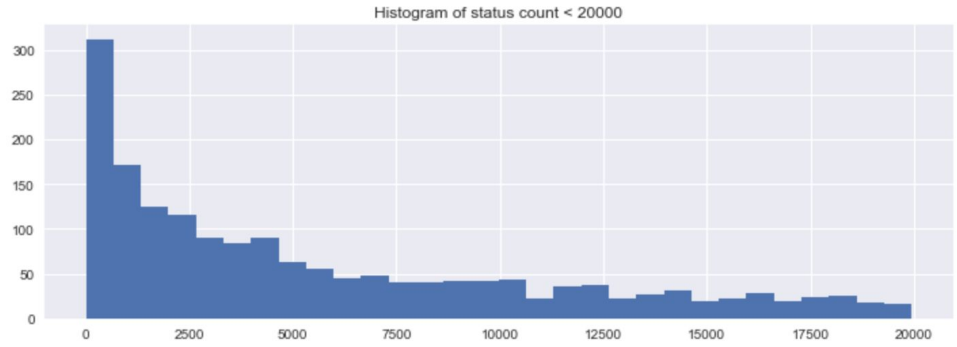
3. Friends count:





4. Statuses count:



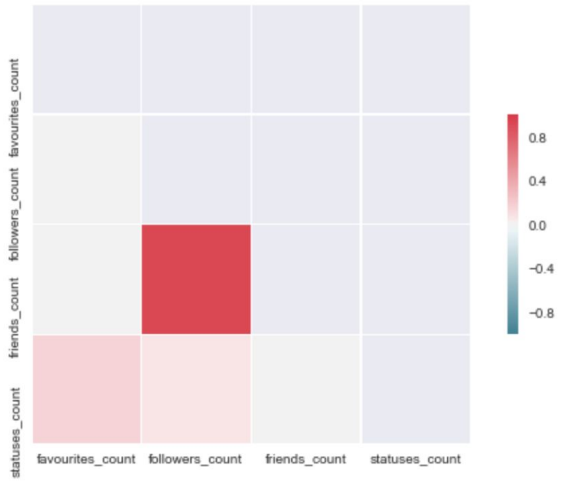


5. **Are friend and follower counts correlated? Are favorite and status counts correlated? What stories can you(speculatively) tell about the relationships between any of these variables, based on your findings?**

   By looking at correlation matrix:

	favourites_count	**followers_count**	**friends_count**	**statuses_count**
favourites_count	1.000000	0.020342	0.002715	0.171091
followers_count	0.020342	1.000000	0.935946	0.077311
friends_count	0.002715	0.935946	1.000000	0.030932
statuses_count	0.171091	0.077311	0.030932	1.000000

   and plot:



We can find insights as follows:

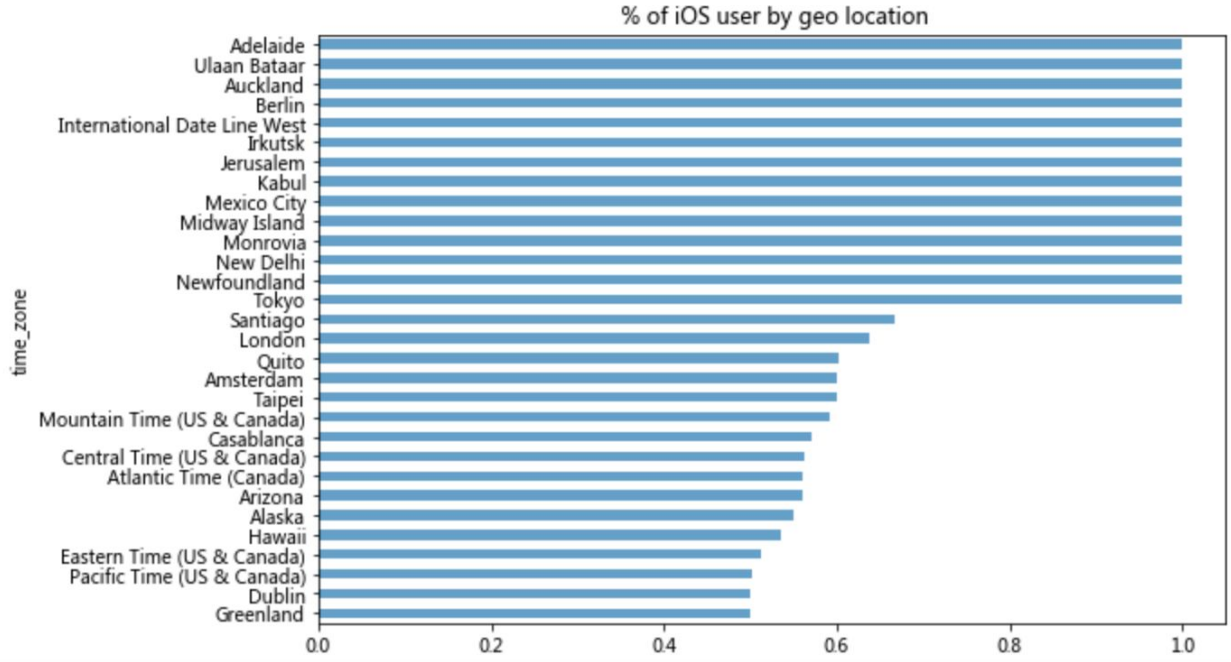
a. Friends and followers are strongly correlated, with a correlation of 0.9359.
b. Statuses_count and favourites_count are extremely weakly correlated, with a correlation of 0.1711.
c. statuses_count and followers_count/friends_count are extremely weakly correlated, with a correlation of 0.0773 and 0.0309.

From the above findings we can speculatively tell stories as follows:

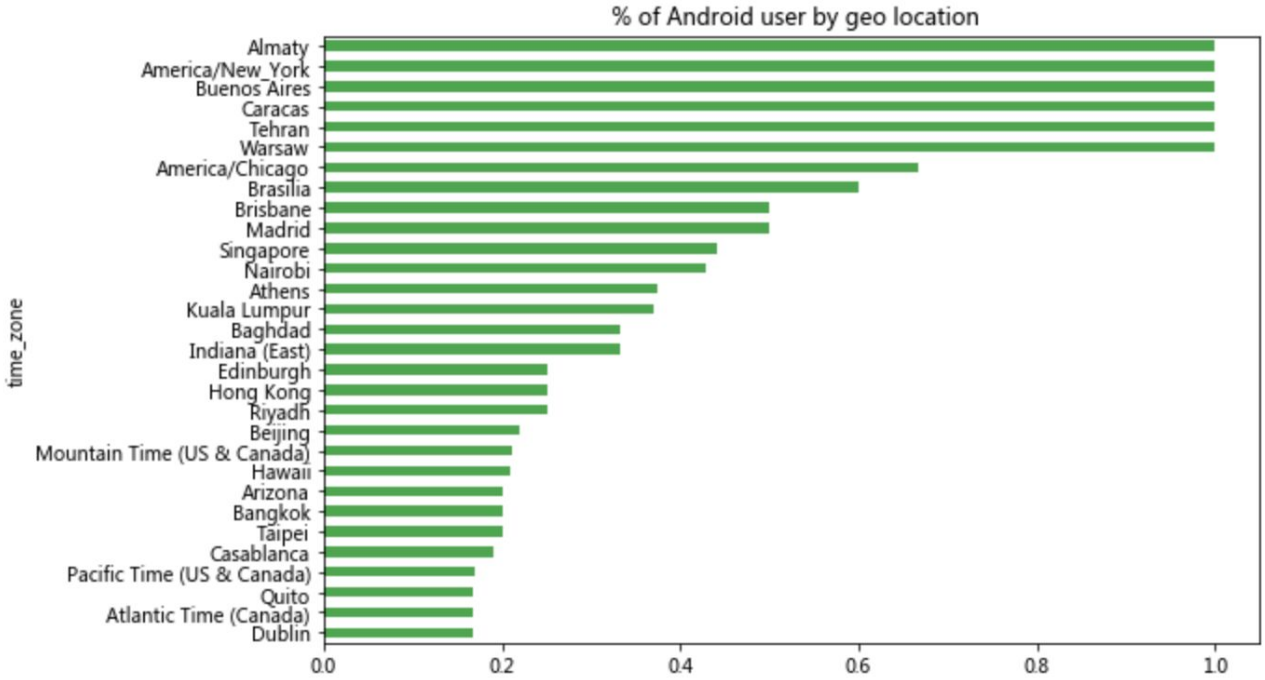
a. If a user has more followers, he/she may have more friends. If a user has more friends, he/she may have more followers. Which means, for most users, friends(mutual followers) constitute most of the followers of a user.
b. If a user has liked many tweets, he/she might have posted more tweets as well, vice versa. But the correlation is weak. This can be due to if a user has posted many tweets, that means he/she is a heavy user of twitter, and a heavy user of twitter might also like more tweets.

2. Which time zone has the highest proportion of know iOS users in age_profiles.csv? Which time zone has the highest proportion of Android users? *(**Note:** There's no file as age_profiles.csv, I guess here it means age_profiles.json)*

a. If we only care about proportion, some time zone only have 1 data points, thus led to the proportion to be 1.
b. Proportion of iOS users:



c. Proportion of Android users:



3. Use the "mentions" data in *mentions.csv* to come up with a list of Twitter handles that were mentioned by more than one user.

Note: *Users can change their twitter handles, but cannot change their twitter ID. Therefore one twitter ID can match >1 twitter handles, but one handle can have and only have one ID. Based on the description of the question, I assume the question is asking for twitter handles instead of twitter IDs.*

Among 10,746 mentioned IDs, 801 are mentioned by more than 1 user. They are included in mentioned_more_than_one.csv.

a. Top 20 handles:

i. Handle and the number of times they got mentioned:

```
MentionedHandle
girlposts          32
Sexualgif          30
SportsCenter       24
YouTube            22
BabyAnimalPics     20
SoDamnTrue         19
RelatableQuote     18
UberFacts          17
CuteEmergency      17
WORIDSTARHlPHOP    17
WorldStarFunny     16
FunnyPicsDepot     16
HornyFacts         15
FIirtationship     15
AboutVirgos        14
TweetLikeAGirI     14
wizkhalifa         14
FreddyAmazin       14
EmWatson           13
Drrake             12
Name: ID, dtype: int64
```

1.

ii. List:

1. ['girlposts',
   'Sexualgif',
   'SportsCenter',
   'YouTube',
   'BabyAnimalPics',
   'SoDamnTrue',
   'RelatableQuote',
   'UberFacts',
   'CuteEmergency',
   'WORIDSTARHIPHOP',
   'WorldStarFunny',
   'FunnyPicsDepot',
   'HornyFacts',
   'FIirtationship',
   'AboutVirgos',
   'TweetLikeAGirl',
   'wizkhalifa',
   'FreddyAmazin',
   'EmWatson',
   'Drrake']

b. Which actor/actress in this top 20 list starred in the Harry Potter movies, and how many unique users mentioned this star's Twitter handle?

i. Emma Watson. 13 unique users have mentioned her in tweets in our data.

4. Break down the sample by age-decade (age 10-20, 20-30, etc). Make a bar chart of age group sample size (x-axis: age group, y-axis: per-group sample size)

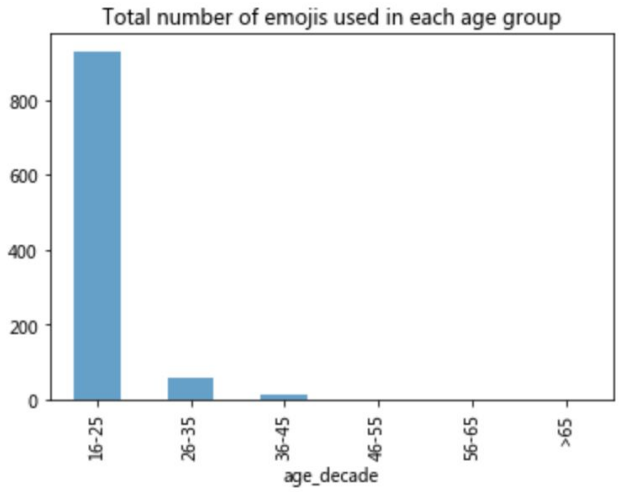
a. Among 1,711 data points, 5 users are aged over 80 (0.2%), 2 are even older than 100. Consider user age older than 0 as outliers.

i.

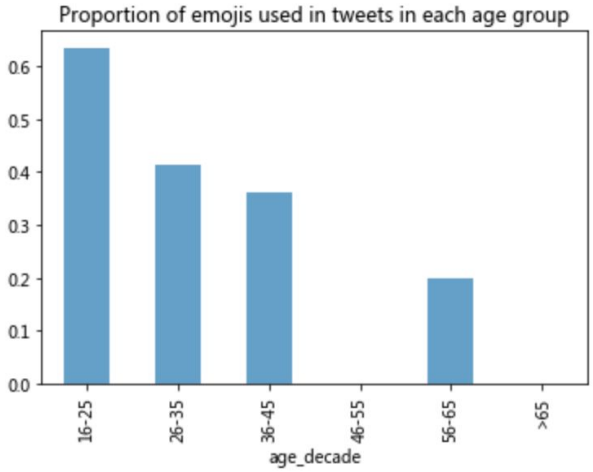
b. **Which age group uses the most emojis in their profile status?

i. There are two ways to consider "most emojis", one way is look at absolute number of emoji usage in each age group, another way is to look at the proportion of emoji usage within each group to find out the group that prefer to use emojis the most.

ii. If look at absolute number of total emoji usage, users age between 16 and 25 use emojis the most, but this can be because users age between 16 and 25 have more tweets:



iii. If we look at relative proportion, users age between 16 and 25 are also more likely to use emojis:



c. Which is the most common emoji?

i. I define the "most common emoji" as the emoji used for the largest number of times. I found 😂 is the most common emoji, with 164 appearances. And the top 5 emojis are: [('😂', 164), ('😍', 63), ('🤩', 53), ('🤬', 50), ('🎉', 48)]