The WeRateGogs Project

An analysis of data wrangled from twitter

By: Malak Alshedokhi

According to the statistics generated by statista website, the number of twitter daily users is increasing by day. In the firs quarter of 2019 it is almost 350 million worldwide. For this project, twitter API was used to analyze the the tweets obtained from WeRateDogs (@dog_rates) account. This is a very poplar account that is used for rating dogs.

In this analysis data were gathered from 3 different sources. The twitter archive of WeRateDogs that was provided by Udacity, which contains basic information about tweets like (tweetID, timestamp, source, text, retweet information, etc). The second file was the tweet images, were they were ran through neural networks to give a prediction about the dog breed. This file which is a type TSV was downloaded using the Python library Requests. Finally, Tweepy Python library was used to retrieve each tweets in a separate JSON.

To begin the analysis, it was a good idea to see numbers and statistics about the twitter archive file. As shown below:

```
In [15]: twitter_archive.describe()
```

Out[15]:

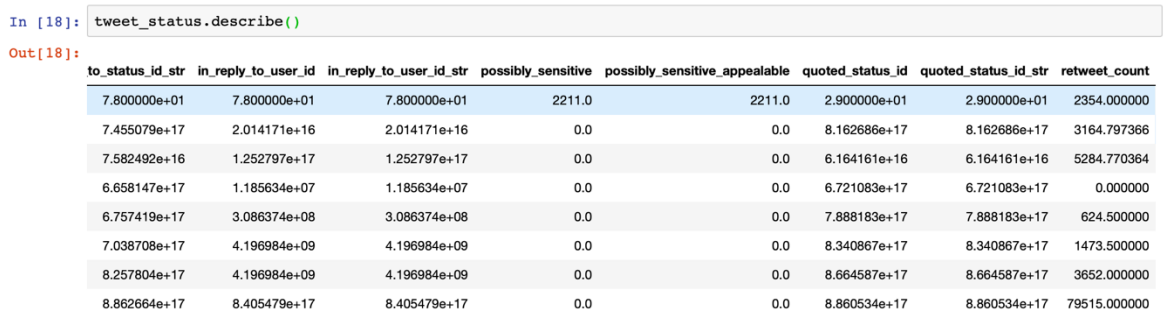
	tweet_id	in_reply_to_status_id	in_reply_to_user_id	retweeted_status_id	retweeted_status_user_id	rating_numerator	rating_denominator
count	2.356000e+03	7.800000e+01	7.800000e+01	1.810000e+02	1.810000e+02	2356.000000	2356.000000
mean	7.427716e+17	7.455079e+17	2.014171e+16	7.720400e+17	1.241698e+16	13.126486	10.455433
std	6.856705e+16	7.582492e+16	1.252797e+17	6.236928e+16	9.599254e+16	45.876648	6.745237
min	6.660209e+17	6.658147e+17	1.185634e+07	6.661041e+17	7.832140e+05	0.000000	0.000000
25%	6.783989e+17	6.757419e+17	3.086374e+08	7.186315e+17	4.196984e+09	10.000000	10.000000
50%	7.196279e+17	7.038708e+17	4.196984e+09	7.804657e+17	4.196984e+09	11.000000	10.000000
75%	7.993373e+17	8.257804e+17	4.196984e+09	8.203146e+17	4.196984e+09	12.000000	10.000000
max	8.924206e+17	8.862664e+17	8.405479e+17	8.874740e+17	7.874618e+17	1776.000000	170.000000

The rating numerator and the rating denominator are the factors used to rate the dogs, as we can see the meant of the numerator is higher than the mean of the denominator.

```
In [18]: tweet_status.describe()
```

Out[18]:

	contributors	coordinates	favorite_count
count	0.0	0.0	2354.000000
mean	NaN	NaN	8080.968564
std	NaN	NaN	11814.771334
min	NaN	NaN	0.000000
25%	NaN	NaN	1415.000000
50%	NaN	NaN	3603.500000
75%	NaN	NaN	10122.250000
max	NaN	NaN	132810.000000


In [18]: tweet_status.describe()

Out[18]:

to_status_id_str	in_reply_to_user_id	in_reply_to_user_id_str	possibly_sensitive	possibly_sensitive_appealable	quoted_status_id	quoted_status_id_str	retweet_count
7.800000e+01	7.800000e+01	7.800000e+01	2211.0	2211.0	2.900000e+01	2.900000e+01	2354.000000
7.455079e+17	2.014171e+16	2.014171e+16	0.0	0.0	8.162686e+17	8.162686e+17	3164.797366
7.582492e+16	1.252797e+17	1.252797e+17	0.0	0.0	6.164161e+16	6.164161e+16	5284.770364
6.658147e+17	1.185634e+07	1.185634e+07	0.0	0.0	6.721083e+17	6.721083e+17	0.000000
6.757419e+17	3.086374e+08	3.086374e+08	0.0	0.0	7.888183e+17	7.888183e+17	624.500000
7.038708e+17	4.196984e+09	4.196984e+09	0.0	0.0	8.340867e+17	8.340867e+17	1473.500000
8.257804e+17	4.196984e+09	4.196984e+09	0.0	0.0	8.664587e+17	8.664587e+17	3652.000000
8.862664e+17	8.405479e+17	8.405479e+17	0.0	0.0	8.860534e+17	8.860534e+17	79515.000000

And the above pictures showing that the favorite and the retweet count are pretty high which is because of the popularity of the website.

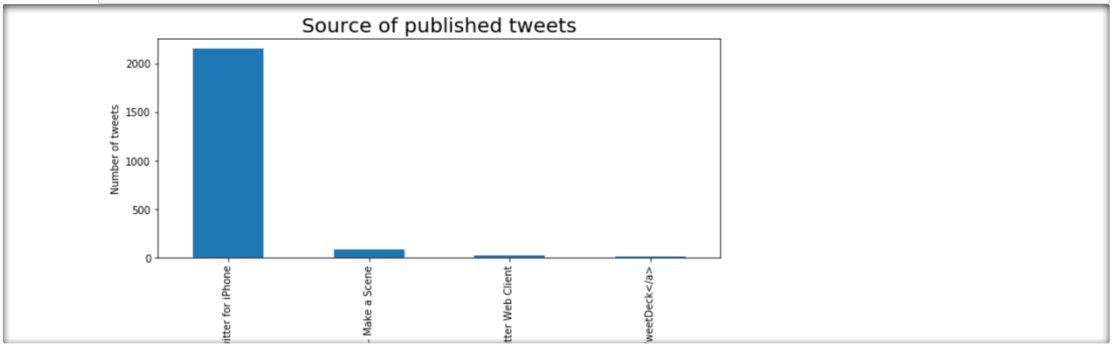
Afterwards, I have started to do some cleaning to the data. Like:

1. Dropping the columns with many missing values
2. Replacing the default Python keyword NaN with the word None.
3. Dropping the columns with dog type and replacing them with one column called Dog_Type to represent the type of the dog in each tweet.
4. Many columns had non-informative headings, those were renamed to be clearer.
5. The ID columns had to be changed to String datatype for easier manipulation in Python.
6. In the column name, some values were ‘a’, ‘an’ and ‘the’. Those values are misleading and do not represents names. They were dropped as well.
7. The source column contained unneeded information such a hyperlink. Those were cleaned to have only the actual source of the tweet.

Some insights:

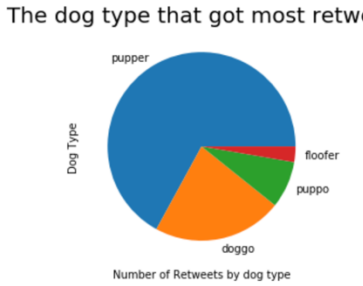
Insights:

```
In [140]: # Plot to show the type of dogs that got the most retweet
fig = plt.figure(figsize=(10,4))
tweeter_df['source_x'].value_counts().plot(kind='bar')
plt.title("Source of published tweets",fontsize=20)
plt.ylabel("Number of tweets")
plt.xlabel("Source");
```



Most users were using Twitter application for iPhone, which indicate that the tweets are by personal users and the popularity of the iPhone device.

```
In [143]: # Plot to show the type of dogs that got the most retweet
fig = plt.figure(figsize=(15,4))
tweeter_df.groupby('Dog_Type').count()['retweet_count'].sort_values(ascending=False).nlargest(4).plot(kind='pie')
plt.title("The dog type that got most retweet",fontsize=20)
plt.ylabel("Dog Type")
plt.xlabel("Number of Retweets by dog type");
```



The dog type pupper had the most retweet which indicated the popularity of this type.

```
In [144]: tweeter_df.describe()
```

Out[144]:

	rating_numerator	rating_denominator	img_num	probability_1_conf	probability_2_conf	probability_3_conf	favorite_count	retweet_count
count	2286.000000	2286.000000	2006.000000	2006.000000	2.006000e+03	2.006000e+03	2284.000000	2284.000000
mean	13.226159	10.455381	1.206381	0.594220	1.341823e-01	6.034463e-02	8229.103765	3217.584063
std	46.556335	6.794867	0.563458	0.272283	1.004650e-01	5.104234e-02	11898.263057	5328.422152
min	0.000000	0.000000	1.000000	0.044333	1.011300e-08	1.740170e-10	0.000000	0.000000
25%	10.000000	10.000000	1.000000	0.362857	5.361625e-02	1.605313e-02	1509.500000	652.750000
50%	11.000000	10.000000	1.000000	0.587440	1.175370e-01	4.919990e-02	3706.500000	1512.000000
75%	12.000000	10.000000	1.000000	0.847216	1.951377e-01	9.241083e-02	10372.750000	3713.000000
max	1776.000000	170.000000	4.000000	1.000000	4.880140e-01	2.734190e-01	132810.000000	79515.000000

- The maximum retweets number that the account got is 81489
- The minimum retweets number that the account got is only 1
- Most tweets were published using twitter for iPhone. That gives an insight about the accounts are most probably personal accounts.

Lastly, the newly cleaned dataset were saved and CSV file called twitter-archive-master.