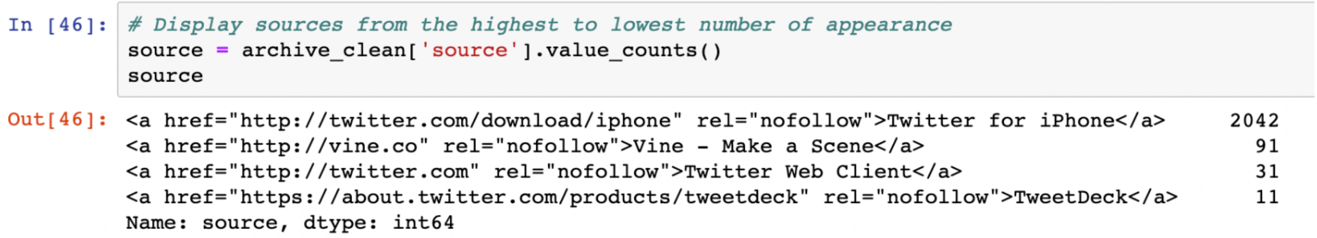
Report for data analyzing and visualizing process in Project 4 of Udacity’s Data Analyst Nanodegree

After cleaning the data and save into the **twitter_archive_master.csv** file, I still used the **archive_clean** table for data analyzing as the csv file requires changing the format of some datatypes again if I open it.

Insight 1: Which is the most common source that gathers data for WeRateDogs

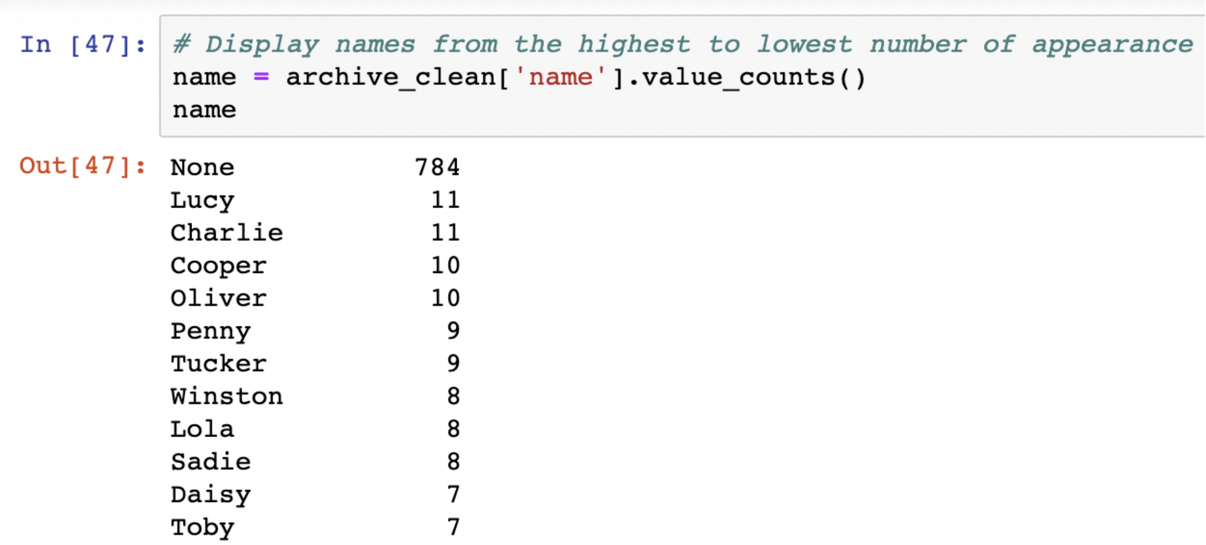


```
In [46]: # Display sources from the highest to lowest number of appearance
         source = archive_clean['source'].value_counts()
         source

Out[46]: <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>     2042
         <a href="http://vine.co" rel="nofollow">Vine - Make a Scene</a>                          91
         <a href="http://twitter.com" rel="nofollow">Twitter Web Client</a>                       31
         <a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a>      11
         Name: source, dtype: int64
```

There are four sources of data and the source containing most of the tweets is Twitter for iPhone with 2042 tweets coming from that.

Insight 2: Which is the most common dog name

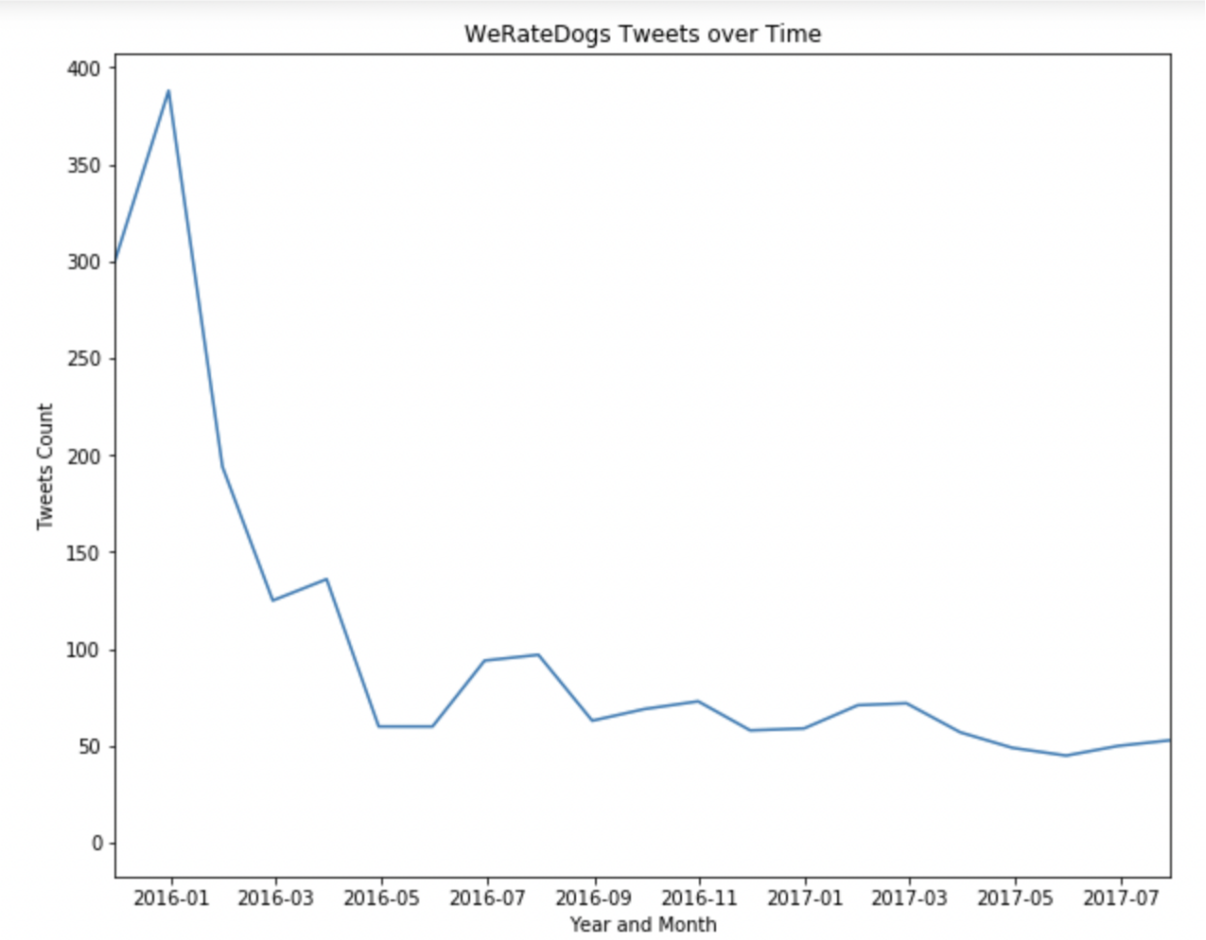


```
In [47]: # Display names from the highest to lowest number of appearance
         name = archive_clean['name'].value_counts()
         name

Out[47]: None          784
         Lucy           11
         Charlie        11
         Cooper         10
         Oliver         10
         Penny           9
         Tucker          9
         Winston         8
         Lola            8
         Sadie           8
         Daisy           7
         Toby            7
```

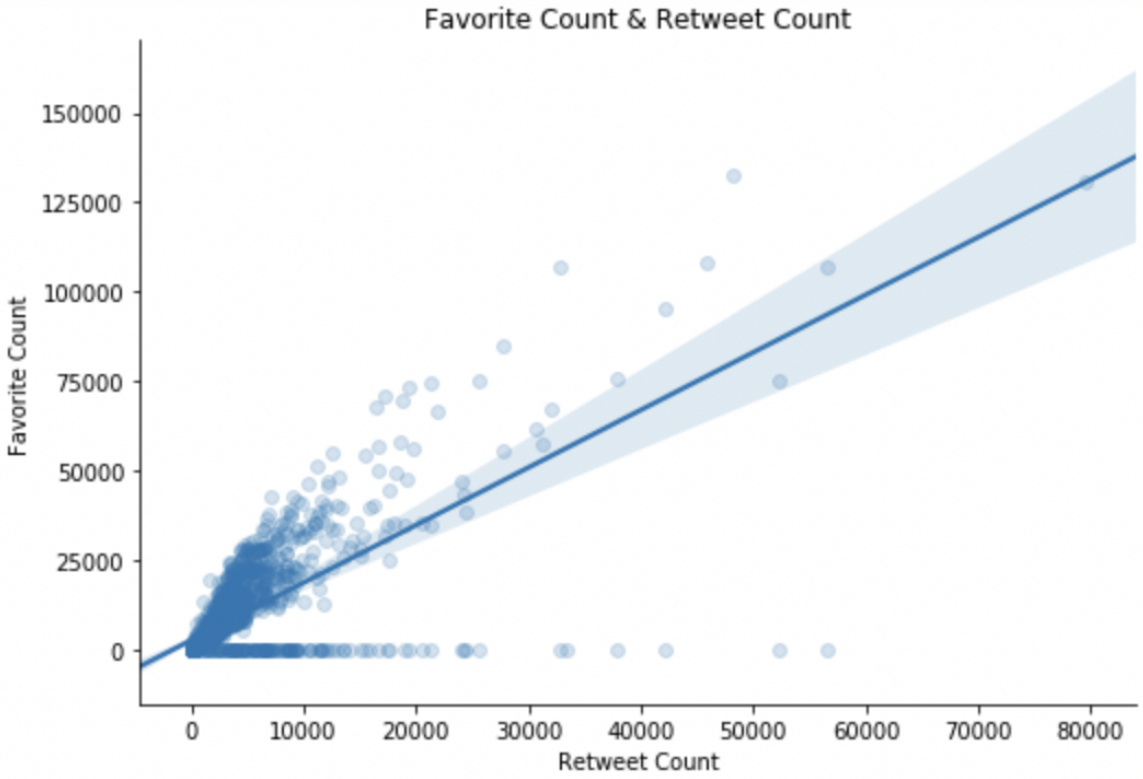
There is a total of 931 different names. And the most common dog name is Lucy and Charlie, with 11 times of appearance for each.

Insight 3: Which month has the highest number of tweet counts



From the graph, we can see tweet counts reached their peak in January 2016, with approximately 400.

Visualization analysis:



The scatter plot is to check the correlation between favorite count and retweet count. Then I use code to check to corr. result and it appears to be 0.714, which shows a positive correlation between favorite and retweet. If you like a tweet (favorite), there is 71.4% of you going to retweet it.