Act Report

goal: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for -worthy analyses and visualizations.

The Data

Enhanced Twitter Archive: The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

Additional Data via the Twitter API: Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API. Well, "anyone" who has access to data for the 3000 most recent tweets, at least. But you, because you have the WeRateDogs Twitter archive and specifically the tweet IDs within it, can gather this data for all 5000+. And guess what? You're going to query Twitter's API to gather this valuable data.

Image Predictions File One more cool thing: I ran every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs*. The results: a table full of image predictions (the top three only) alongside each tweetID,image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

Analysis

I clean and analyze data from this account by show compare the favorite counts & retweet counts ,what the most source for tweet,what is the 6 frequent breed,what the most stage of doges,and number of tweets per month?, I used different types of graphs for display and analysis of matplotlib libray such as countplot,pie chart,bar chart,scatter,and plot.line

Finding

As shown in the table a bove that the mean (retweet:2971.322, favorite:7752.137) for retweet and favorite ,largest number of favorite equal to 107015

```
In [73]: merge_df_clean.describe()
```

Out[73]:

	id	retweet_count	favorite_count	img_num
count	1.518000e+03	1518.000000	1518.000000	2075.000000
mean	7.386500e+17	2971.322134	7752.137681	1.203855
std	6.699942e+16	4867.362390	10966.529752	0.561875
min	6.660293e+17	2.000000	0.000000	1.000000
25%	6.767853e+17	607.250000	1405.250000	1.000000
50%	7.145456e+17	1389.000000	3558.500000	1.000000
75%	7.904331e+17	3493.500000	9629.000000	1.000000
max	8.918152e+17	56625.000000	107015.000000	4.000000

Merged Dataset info :

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2356 entries, 0 to 2355
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   tweet_id        2356 non-null   object
 1   tweet_date      2356 non-null   datetime64[ns, UTC]
 2   text            2356 non-null   object
 3   name            2356 non-null   object
 4   stages_of_dogs  399 non-null    object
 5   id              1518 non-null   float64
 6   source          1518 non-null   object
 7   retweet_count   1518 non-null   float64
 8   favorite_count  1518 non-null   float64
 9   jpg_url         2075 non-null   object
 10  img_num         2075 non-null   float64
 11  breed           1708 non-null   object
dtypes: datetime64[ns, UTC](1), float64(4), object(7)
```

The dataset have 2356 observations,
12 columns and with no null values. The data types of the variables are divided in 4 float, 1 datetime , and 7 object.

Quantities variables :

:

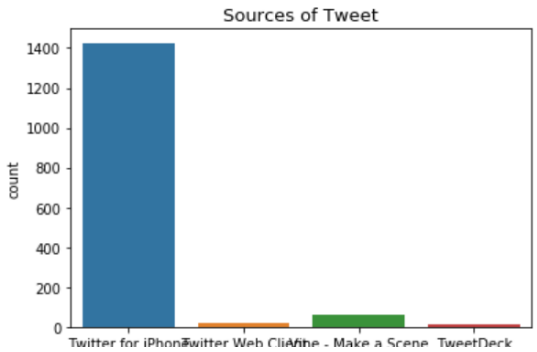
	tweet_id	text	name	stages_of_dogs	source	jpg_url	breed
count	2356	2356	2356	399	1518	2075	1708
unique	2356	2356	957	4	4	2009	116
top	693155686491000832	This is Kreg. He's riding an invisible jet ski...	None	pupper	Twitter for iPhone	https://pbs.twimg.com/media/ChK1tdBWwAQ1flD.jpg	labrador_retriever
freq	1	1	745	265	1426	2	110

Count for name 2356,but unique 957, count stage of dogs 399 ,unique just4

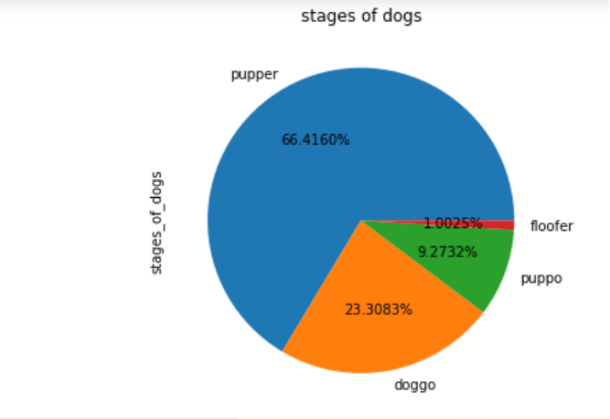
Analyzing and Visualizing Data

1. What is the most used source? Iphone is most used source for tweet

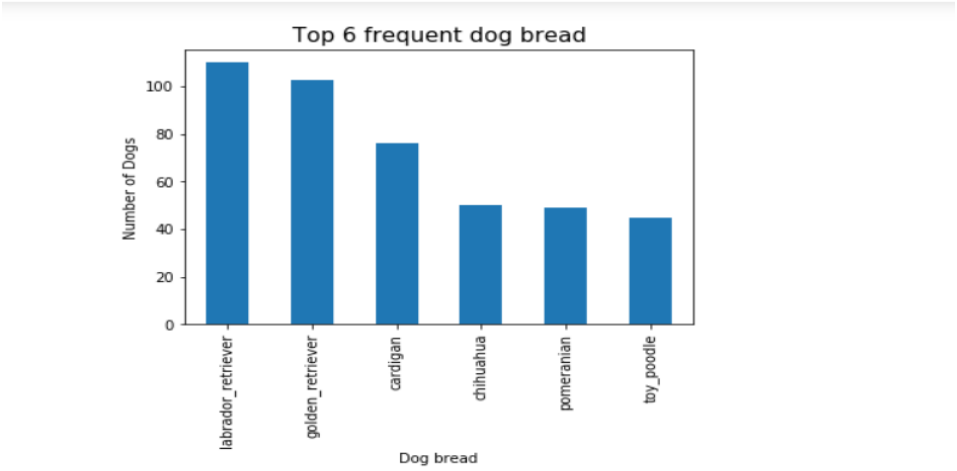
```
<matplotlib.axes._subplots.AxesSubplot at 0x10bfe4a9388>
```



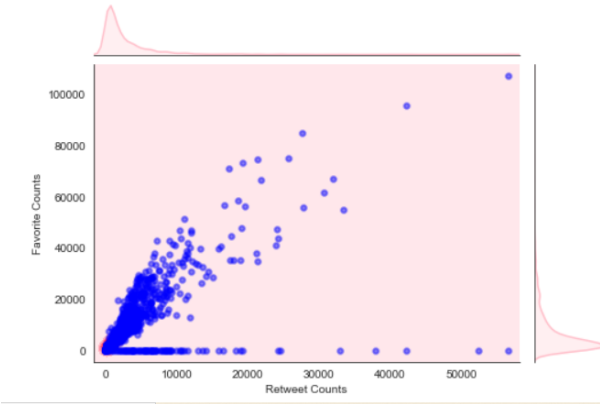
2. What is the most stage of doges?pupper the most stage of doges , doggo ,puppo then floofer



3. What is the 6 frequent bread?
   Labrador_retriever is frequent 110
   golden_retriever is frequent 103
   cardigan is frequent 76
   chihuahua is frequent 50
   pomeranian is frequent 49
   toy_poodle is frequent 45



Visualization compare the favorite counts & retweet counts



4. Number of Tweets per month?
   Notice the most increase in the number ion 12 month ,2015 year

```
Out[81]: Text(0, 0.5, 'Number of Tweets')
```