ACT REPORT

This is the summary of the project from the data wrangling process.

Three dataset was gathered and worked on and they are:

1. **twitter_archive_enhanced.csv**, this dataset was provided for the project and had over 2356 tweets when I downloaded it.
2. **Image_prediction.tsv**, I programmatically downloaded the file with over 2075 predictions of dog breeds classification
3. **Tweet_json_text**, I scrapped the twitter API using python tweepy's library and has 2327 tweets.

Accessing the datasets, quality issues and tidiness issues where found and I made use of several python pandas methods to get them clean using define, code and test method in the cleaning process.

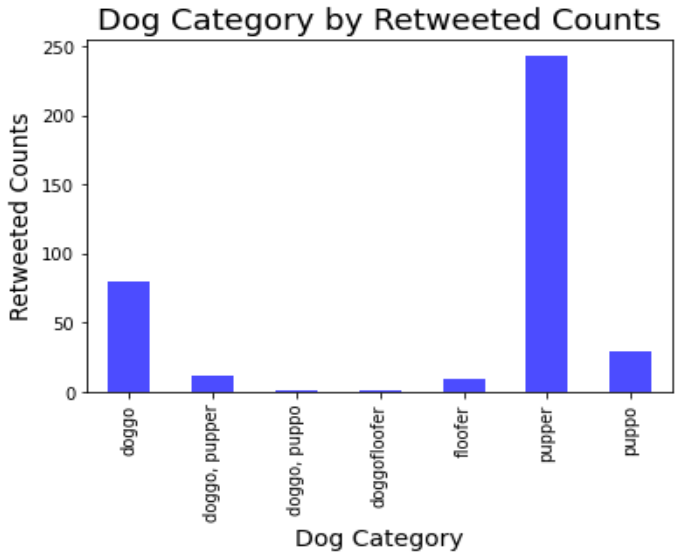
Furthermore, the three datasets was merged and called **twitter_archive_master.csv** and here are some visuals gotten after I loaded it into a pandas dataframe.

Out[84]:

	tweet_id	rating_numerator	rating_denominator	img_num	p1_conf	p2_conf	p3_conf	retweet_count	favorite_count
count	2.356000e+03	2356.000000	2356.000000	2075.000000	2075.000000	2.075000e+03	2.075000e+03	2327.000000	2327.000000
mean	7.427716e+17	13.126486	10.455433	1.203855	0.594548	1.345886e-01	6.032417e-02	2468.171465	7047.888698
std	6.856705e+16	45.876648	6.745237	0.561875	0.271174	1.006657e-01	5.090593e-02	4179.936890	10952.818982
min	6.660209e+17	0.000000	0.000000	1.000000	0.044333	1.011300e-08	1.740170e-10	1.000000	0.000000
25%	6.783989e+17	10.000000	10.000000	1.000000	0.364412	5.388625e-02	1.622240e-02	493.500000	1224.000000
50%	7.196279e+17	11.000000	10.000000	1.000000	0.588230	1.181810e-01	4.944380e-02	1148.000000	3049.000000
75%	7.993373e+17	12.000000	10.000000	1.000000	0.843855	1.955655e-01	9.180755e-02	2858.000000	8596.500000
max	8.924206e+17	1776.000000	170.000000	4.000000	1.000000	4.880140e-01	2.734190e-01	70643.000000	144748.000000

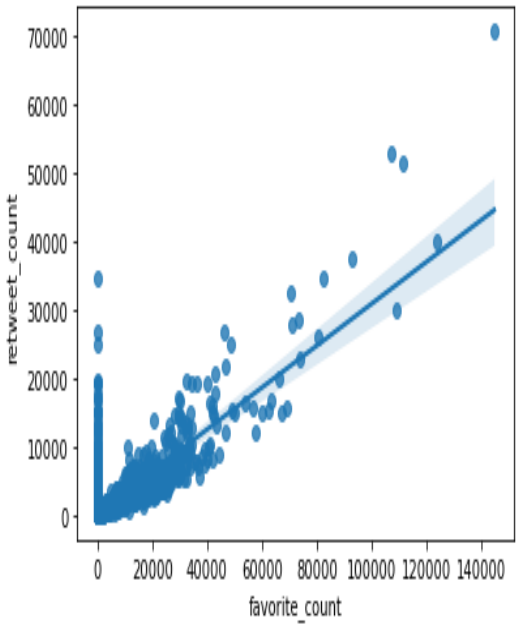
The above shows all the useful descriptive statistics for numerical data of the dataset.

```
Out[93]: Text(0, 0.5, 'Retweeted Counts')
```



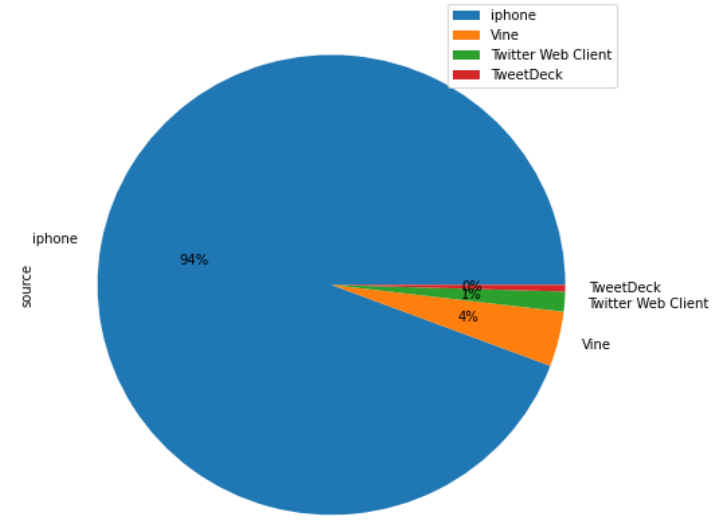
From the bar chart above, the most favorite dog category that was rated according to favorite and retweet count was pupper, followed by doggo and then puppo.

```
Out[99]: <AxesSubplot:xlabel='favorite_count', ylabel='retweet_count'>
```



Looking at the graph above, there is a strong positive linear relationship between the retweeted count and the favorite count showing that the two variables have a direct connection. The higher the likes on a dog post, the higher the retweets.

From the pie chart below, it can be depicted that those making use of iphone as a source appliance had 94%, follwed by Vine with4%, twitter Web Client with 1% and TweetDeck with 0% approximately.


iphone
Vine
Twitter Web Client
TweetDeck
iphone
source
94%
0%
1%
4%
TweetDeck
Twitter Web Client
Vine