ACT REPORT

This is the summary of the Data Analysis process that was undertaken for the data wrangling project.

Three datasets were used for this project.

The first dataset were provided by Udacity, which is a comma separated value (csv) file named twitter_archive_enhanced.csv. It was downloaded manually and contains some information about 2356 tweets.

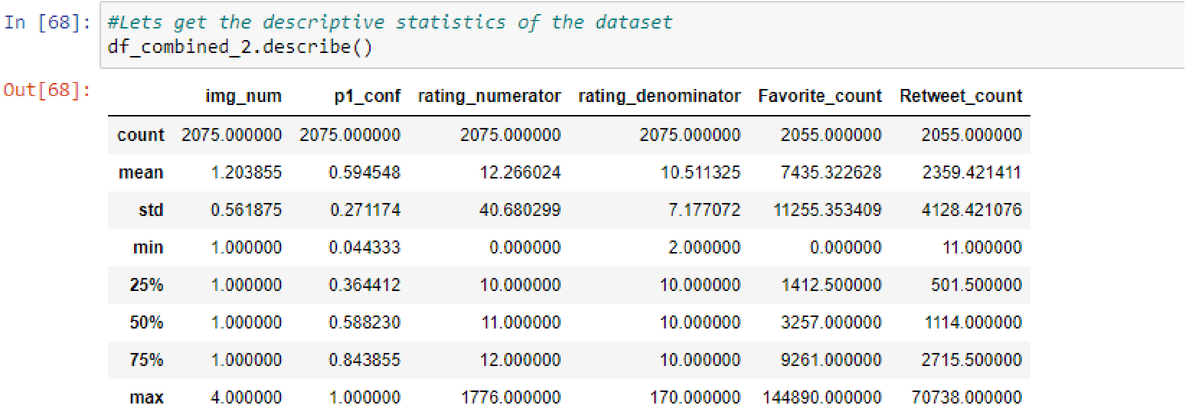
The second dataset was downloaded programmatically using requests library. It was a tsv file named image_prediction.tsv which was hosted on Udacity server. It contains 2075 predictions made by a neural network algorithm that can classify dog breeds.

The third dataset were scrapped from the twitter API using python Tweepy's Library. Retweet count and favorite count were extracted from the json file named "tweet_json_text".

Eight (8) quality issues and four (4) tidiness issues were identified during the assessment stage. It was cleaned programmatically using various python pandas methods.

After the data wrangling processes, the three datasets were merged into a single dataframe. Some insights and visualizations were drawn from the merged dataset as follows:

This is a snapshot of the descriptive statistics from the dataset.



```
In [68]: #Lets get the descriptive statistics of the dataset
         df_combined_2.describe()
```

Out[68]:

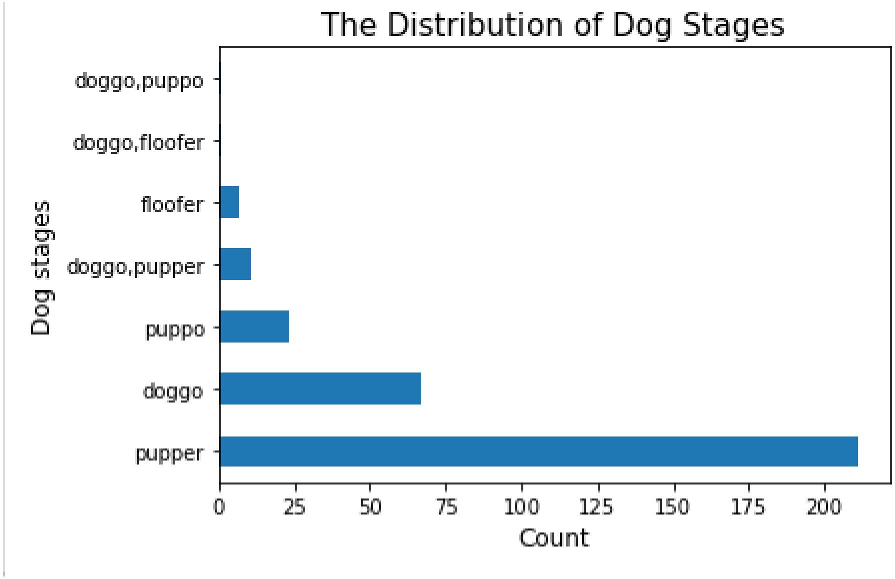
	img_num	p1_conf	rating_numerator	rating_denominator	Favorite_count	Retweet_count
count	2075.000000	2075.000000	2075.000000	2075.000000	2055.000000	2055.000000
mean	1.203855	0.594548	12.266024	10.511325	7435.322628	2359.421411
std	0.561875	0.271174	40.680299	7.177072	11255.353409	4128.421076
min	1.000000	0.044333	0.000000	2.000000	0.000000	11.000000
25%	1.000000	0.364412	10.000000	10.000000	1412.500000	501.500000
50%	1.000000	0.588230	11.000000	10.000000	3257.000000	1114.000000
75%	1.000000	0.843855	12.000000	10.000000	9261.000000	2715.500000
max	4.000000	1.000000	1776.000000	170.000000	144890.000000	70738.000000

INSIGHTS FROM THE ANALYSIS

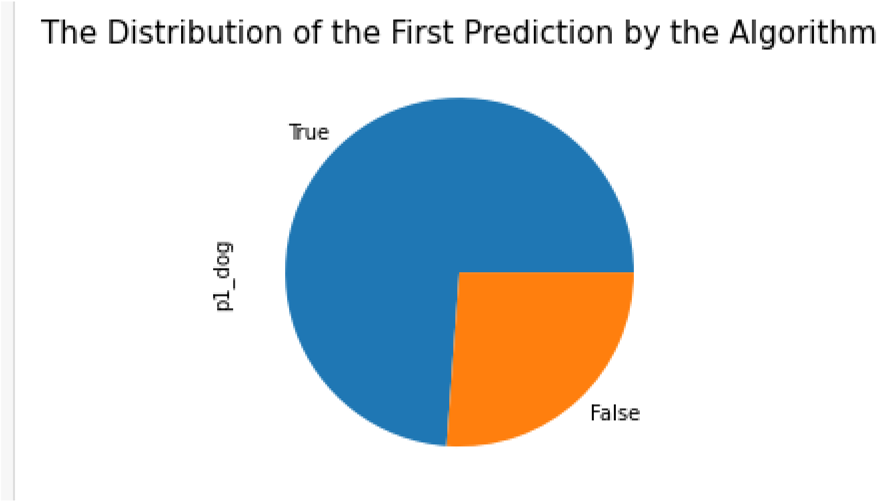
1. The maximum number of prediction images per dog were 4 images while the minimum were 1 image.

2. One (1) image per dog recorded the highest number (1780 dogs), followed by dogs with 2 images (198 dogs)

3. The dog stage with the highest number were pupper (211), followed by doggo (67)
4. The prediction by the algorithm of the image to the type of were were True for about 73.8%
5. Majority of the tweet were from iPhone (98.0%)

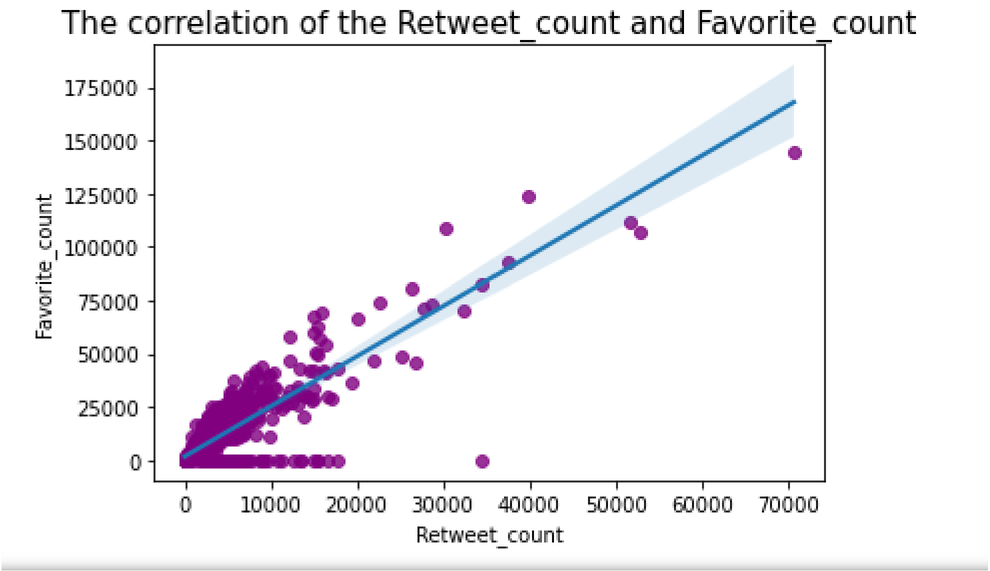


From the barchart above, it can be seen that pupper stage had the highest count, followed by doggo. This could be because they pupper are younger and probably more cute, which explains people having them more when compared to other dog stage.



The Distribution of the First Prediction by the Algorithm

From the pie-chart above, it can be seen that a high percentage of the first predictions by the neural network algorithm were True (73.8%)



The correlation of the Retweet_count and Favorite_count

From the graph above, there is a positive linear relationship between retweet_count and favorite_count.

This is the end of the summary for the project!