Report: act report

This act report includes the summary of the Data Analysis process that was taken for the data wrangling project.

In this project, I worked with three datasets.

Udacity provided the first dataset which is a csv file named twitter_archive_enhanced.csv. It contains basic information about 2354 tweets and was downloaded manually.

The second dataset was a tsv file named image_prediction.tsv which was hosted on udacity server and I programmatically downloaded the file. It contains 2075.

For the third dataset, I tried to scrape from twitter api but I failed so I had to download the json file from Udacity. This third dataset contains information like the rewetet count, favorite count, followers count and friends count each tweet recieved for 2354 tweets in the file "tweet_json_text".

During accessing the data, I found out 10 quality issues and 4 tidiness issues. I used a variety of Pandas methods to clean them up.

Here are some insights and visualizations that I got after I merged the three datasets into a master dataset named twitter_archive_master.csv.

```
[207] # Looking at the statistical description of our master dataset
      data.describe()
```

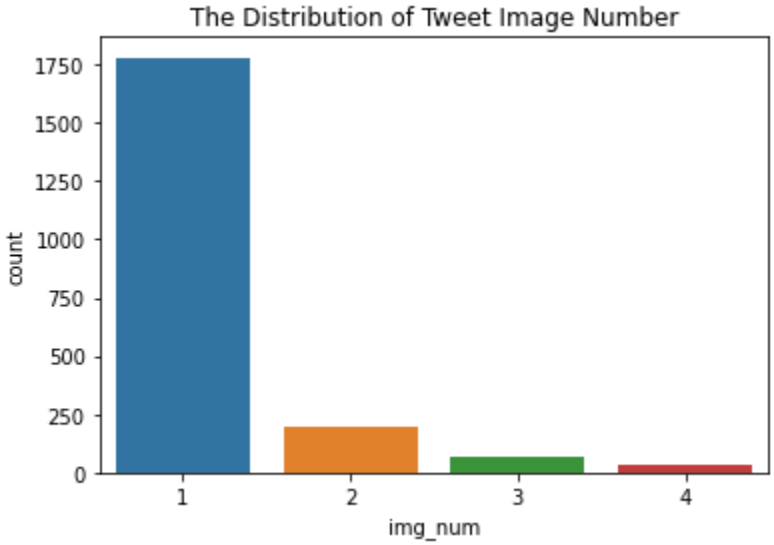
	tweet_id	rating_numerator	rating_denominator	img_num	p1_conf	p2_conf	p3_conf	retweet_count	favorite_count
count	2.073000e+03	2073.000000	2073.000000	2073.000000	2073.000000	2.073000e+03	2.073000e+03	2073.000000	2073.000000
mean	7.383634e+17	12.265798	10.511819	1.203570	0.594532	1.346665e-01	6.034005e-02	2976.089243	8556.718283
std	6.780118e+16	40.699924	7.180517	0.561856	0.271234	1.006830e-01	5.092769e-02	5054.897526	12098.640994
min	6.660209e+17	0.000000	2.000000	1.000000	0.044333	1.011300e-08	1.740170e-10	16.000000	0.000000
25%	6.764706e+17	10.000000	10.000000	1.000000	0.364095	5.390140e-02	1.619920e-02	634.000000	1674.000000
50%	7.119681e+17	11.000000	10.000000	1.000000	0.588230	1.186220e-01	4.947150e-02	1408.000000	3864.000000
75%	7.931959e+17	12.000000	10.000000	1.000000	0.843911	1.955730e-01	9.193000e-02	3443.000000	10937.000000
max	8.924206e+17	1776.000000	170.000000	4.000000	1.000000	4.880140e-01	2.734190e-01	79515.000000	132810.000000

Insights:

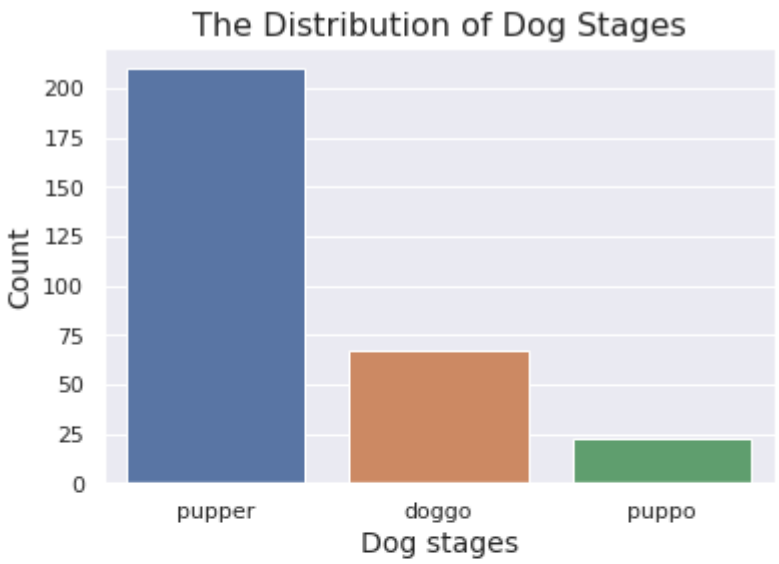
1. The minimum favorite count is 0, mean is 8556 and the maximum is 132810
2. The minimum retweet count is 16, mean is 2976 and the maximum is 79515
3. About 32% of the dogs have no name

Visualization

Question 1: How many image number occurred most for each tweet's most confident image prediction?



Question 2: What is the most popular dog stage according to the neural network's image prediction?




```
[212] data.stage.value_counts()
```

```
pupper           210
doggo             67
puppo             23
doggo,pupper      11
floofer            7
doggo,puppo        1
doggo,floofer      1
Name: stage, dtype: int64
```

But there are a lot of missing data on stage.

Thank you.