Report: act_report

Create a **250-word-minimum written report** called "act_report.pdf" or "act_report.html" that communicates the insights and displays the visualization(s) produced from your wrangles data. This is to be framed as an external document, like a blog post or magazine article, for example.

A good Dog

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 2356 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 predictions made by a neural network that can classify dog breeds.

For the third dataset, I scrapped the twitter API using python Tweepy's Library. This third dataset contains information like the rewetet count, favorite count, followers count and friends count each tweet recieved for 2327 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`.

First I loaded the the master dataset in a pandas dataframe.

In [1]:

```
import pandas as pd
data = pd.read_csv("twitter_archive_master.csv")
```

In [2]:

```
data.describe()
```

Out[2]:

	tweet_id	rating_numerator	rating_denominator	img_num	p1_conf	p2_conf
count	1.986000e+03	1986.000000	1986.000000	1986.000000	1986.000000	1.986000e+03
mean	7.356142e+17	12.281974	10.534240	1.203424	0.593452	1.344853e-01
std	6.740686e+16	41.581180	7.335369	0.561492	0.271961	1.005944e-01
min	6.660209e+17	0.000000	2.000000	1.000000	0.044333	1.011300e-08
25%	6.758214e+17	10.000000	10.000000	1.000000	0.362656	5.407533e-02
50%	7.082494e+17	11.000000	10.000000	1.000000	0.587357	1.175370e-01
75%	7.873791e+17	12.000000	10.000000	1.000000	0.844920	1.951377e-01
max	8.924206e+17	1776.000000	170.000000	4.000000	1.000000	4.880140e-01

Insights

- The minimum favorite count is 66, mean is 7714, and the maximum favorite count is 144955
- The minimum retweet count is 11, mean is 2245, and the maximum retweet count is 70786
- About 32% of the dogs have no name

- Image number 1 is the most prominent (frequent)
- The merged dataset has 21 columns and 1986 rows, all the rows except for the dog stage column are completely filed with no missing value.
- The columns are 'tweet_id', 'timestamp', 'source', 'text', 'expanded_urls', 'rating_numerator', 'rating_denominator', 'name', 'stage', 'retweet_count', 'favorite_count', 'jpg_url', 'img_num', 'p1', 'p1_conf', 'p1_dog', 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog'.
- Nine of the columns are object data type (string), one is datetime, five are integer data types, three are floats, and the remaining three are boolean data types.

Visualizations

1. The most occcuring image number that corresponds to each tweet's most confident prediction is 1.

Distribution of Tweet Image Number

1. The most popular dog stage that were rated by the WeRateDogs Twitter account was pupper, follwed by doggo and then puppo.

Distribution of dog stages

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

A reasonable hypothesis is that the most popular tweets get the highest number of retweet count and favorite count. I tested the correlation between retweet_count and favorite_count and the r^2 is 0.928. That is a high value showing a strong correlation between them.

Linear Correlation bBtween Retweet count and Favorite count

That is the summary of the Data Wrangling process!

Two Best Friends Hugging Because They are Good Dog Brents!