In this document I will briefly discuss my findings regarding Udacity's Twitter Data Wrangling Project.

Insights

Insight #1 Identified dog breeds

What are the 5 most common breeds as identified by the neural network and their absolute frequency?

N.b. Only original tweets and based solely on the 1st prediction.

The two most popular dog breeds are closely related breeds of the same family: retrievers, a breed known for its natural affection and toleration of children. The subsequent three (pembroke, chihuahua, and pug) are all small dog breeds with fairly high meme potential. By not setting any parameters, it becomes visible that (assuming that the neural network is at least fairly accurate) not all photos actually contain dogs, but it would require manual examination to confirm this hypothesis.

Insight #2 Dog stages

What is the relative frequency of (identified) dog stages?

Although nearly 85% of all dogs are uncategorized, the relative frequency of identified dogs are as following (using Lucid Software's explanation of dog-related lingo in their video "What is a Pupper? What is a Doggo?"): The most popular type of floofer (which can be any kind of dog, but usually refers to big dogs with a lot of fur) by a landslide is a smoll doggo, a so-called pupper. A big pupper, commonly known as doggo, is the second most common type of floofer. The intermediate stage between a pupper and a doggo makes up around 1% of all floofers, whereas the authentic fur-heavy floofer makes up less than half a percent of all floofers.

Insight #3 Dog names

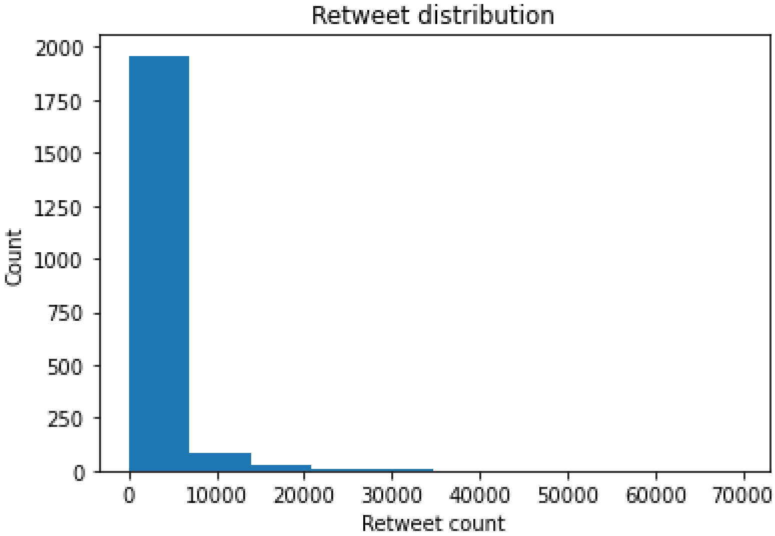
What are the 10 most popular dog names?

Aside from the list of the ten most popular dog names, included are some other statistics that reveal some interesting characteristics: there are (2094-704) 1390 registered dog names in this dataset. 930 of those are unique. Considering that only the 10 most popular names already account for 84 observations, the majority of dog names appear only once in this dataset. Additionally, the most frequently used names, Lucy and Charlie, still make up roughly ((22/2094)*100) 1% of all observations, confirming an incredible diversity in dog names!

Insight #4 Retweet distribution

How is the retweet count distributed?

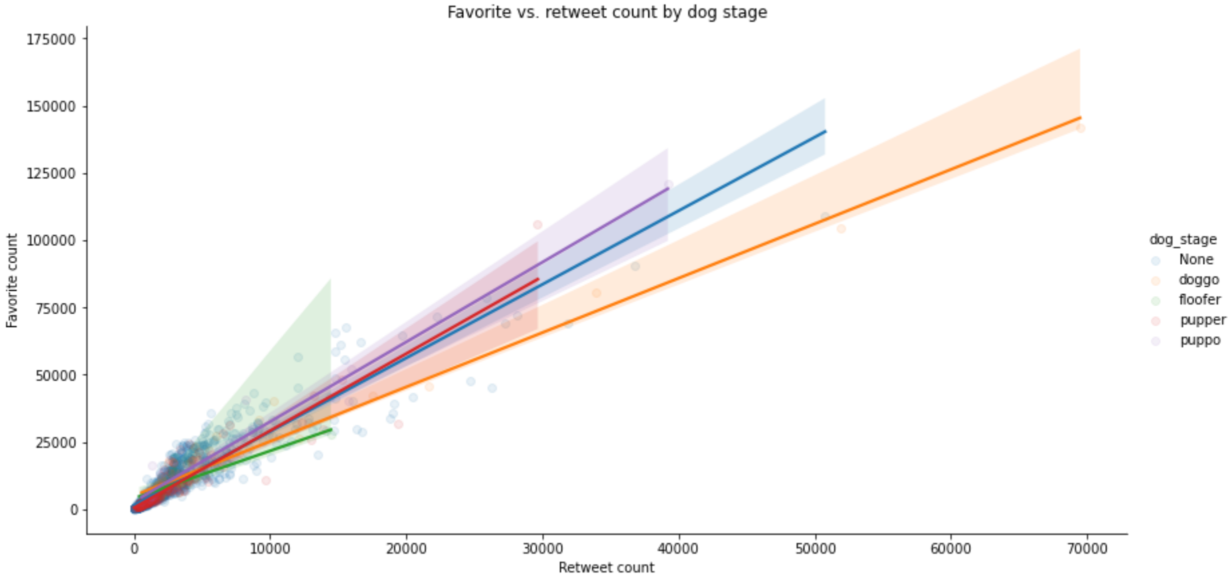
As can be derived from the statistics and figure below, this variable is strongly positively skewed and leptokurtic: Nearly 75% of the data falls below the mean, the max is almost 17 standard deviations above the mean, and the IQR is only half a standard deviation. Translated to reality, out of 2079 original tweets, nearly 2,000 got less than 5,000 retweets. A few tweets went viral, leading to a max of nearly 70,000 retweets.



Visualization #1 Dog stage and tweet virality

Is there a correlation between the amount of times a tweet is favorited and retweeted and does this correlation differ between dog stages?

Juxtaposed with the occurrence of identified dog stages, where pupper was the most frequently occurring, in terms of retweets and favorite count, puppers are outperformed by every other dog stage. Possibly, the owners of the Twitter account WeRateDogs like puppers more than the audience does. Maybe the audience also loves puppers, but do not feel comfortable associating themselves with this type of content. Taking in consideration the dog_stage distribution, perhaps people have grown tired from seeing an overabundance of tweets with photos of puppers.


Favorite vs. retweet count by dog stage
Favorite count
175000
150000
125000
100000
75000
50000
25000
0
0
10000
20000
30000
40000
50000
60000
70000
Retweet count
dog_stage
None
doggo
floofer
pupper
puppo