

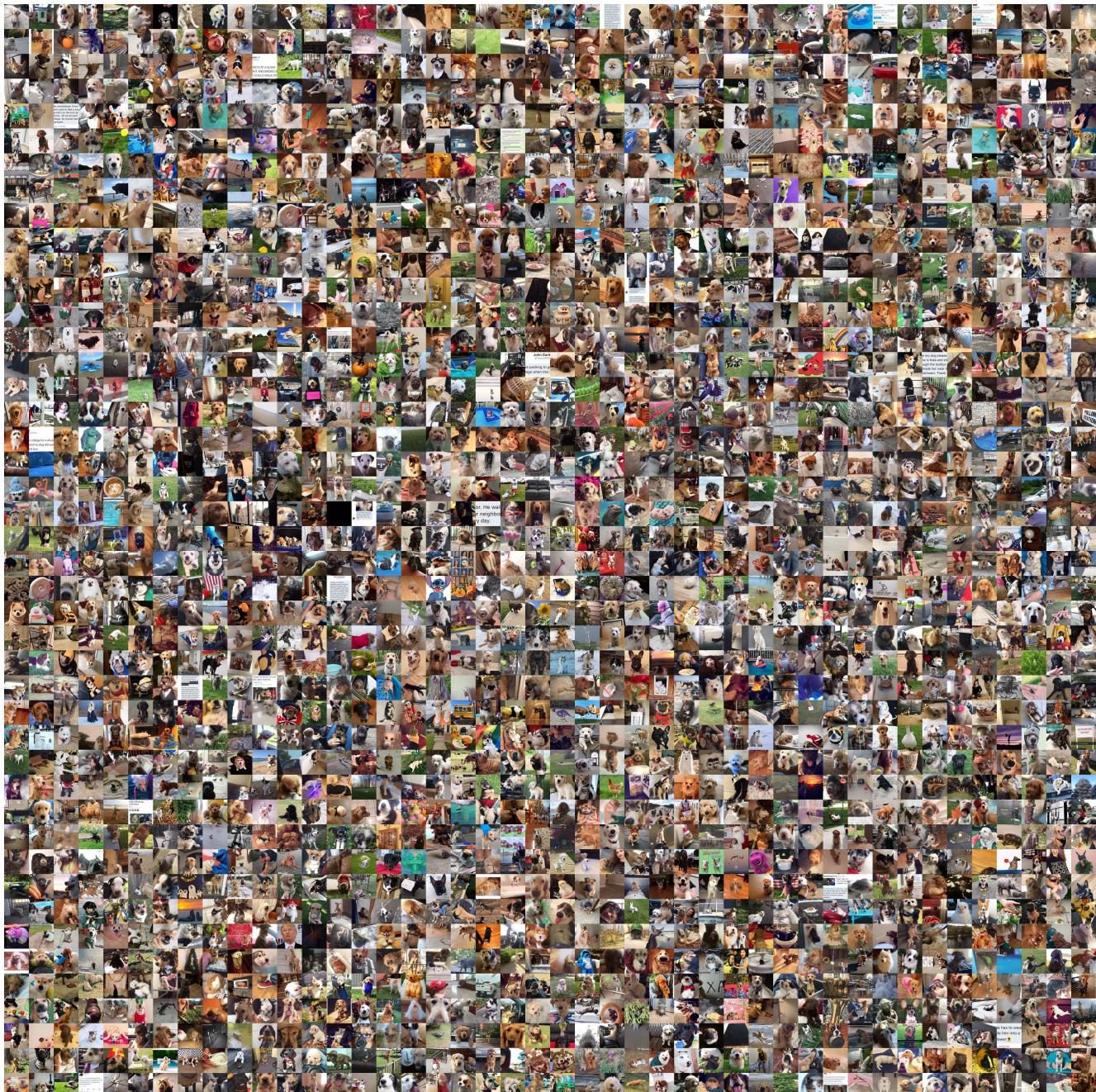
Report - @WeRateDogs on Twitter

Background

This project is an exercise in Data Wrangling - gathering, assessing and cleaning data from the Twitter account @WeRateDogs. In addition to data gathered using the Twitter API and an Enhanced dataset that expanded on data embedded in the tweet text, Udacity provided a dataset that was the result of an Image Classification model that predicted which dog breeds were shown in the images. I found this particular data set the most interesting because it provided insight to how the Machine Learning model works.

In fact, below you can see a sample of the dogs represented in the tweets. If you peer closely, you can see that they aren't all dogs!

These are a sample of the dogs represented in the dataset! If you peer closely, you can see that they aren't all dogs.



I cleaned these 3 datasets and extracted and downloaded the thumbnails associated with the tweets to answer the questions below. Since the data is not 100% reliable -- **There are many items that remained to be cleaned across the datasets. The data itself (ML model predictions) is many times incorrect or perhaps biased** -- Please take these questions and answers with a grain of salt -- or crumb of dog biscuit if you prefer! Hopefully you will find some entertaining observations too, such as which dogs were identified as "Seat belt"!

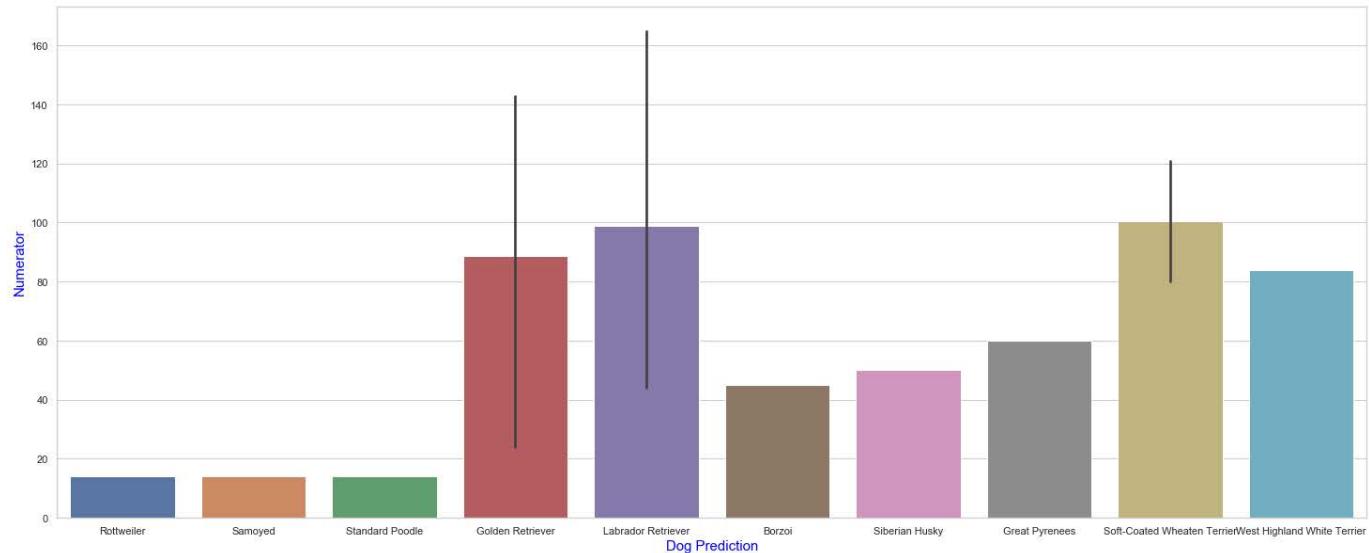
Questions Answered

- What 20 breeds of dogs had the highest ratings? (Numerator / Demominator)
- What is the trend of the retweet and favorite counts over 2015 - 2017?
- What are the most popular categories of dog?
- It looks like this dataset is skewed towards English speakers (Was Google translate was automatically used?)
- What are the top 20 dog names out of a total of 929 names?
- What are the top 20 dog breeds out of the breeds correctly identified as dog breeds?
- Is a cardigan a dog or a sweater? Let's look!
- What were the top 5 incorrect predictions for things that weren't dog breeds? (This could be funny!)
- Let's look and see what pictures were misclassified and perhaps why?

Visualizing

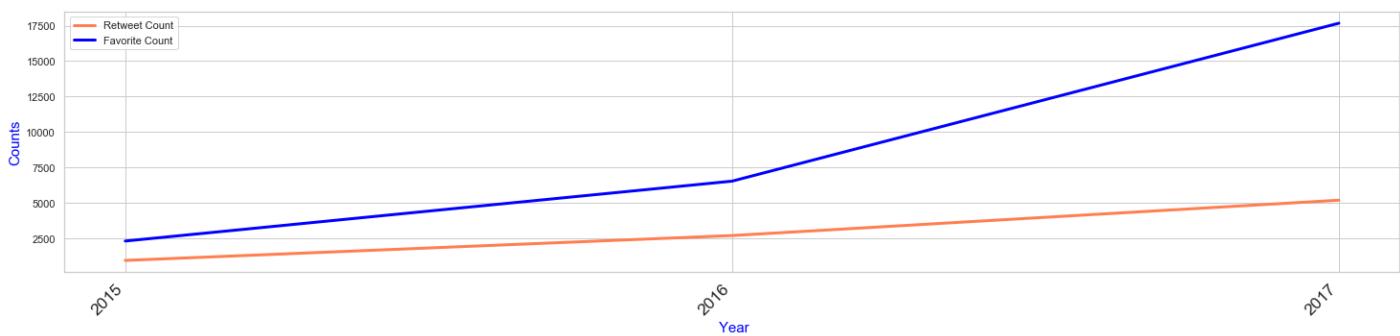
What 10 breeds of dogs had the highest ratings (Numerator)?

Soft-Coated Wheaten Terrier, Labrador Retriever, Golden Retriever



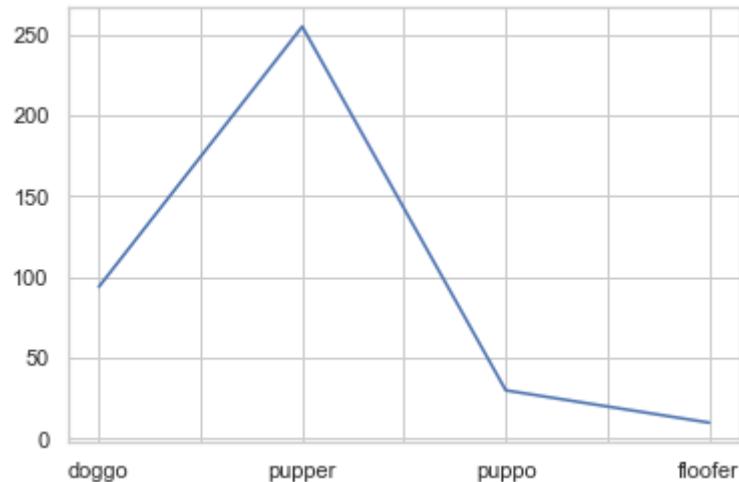
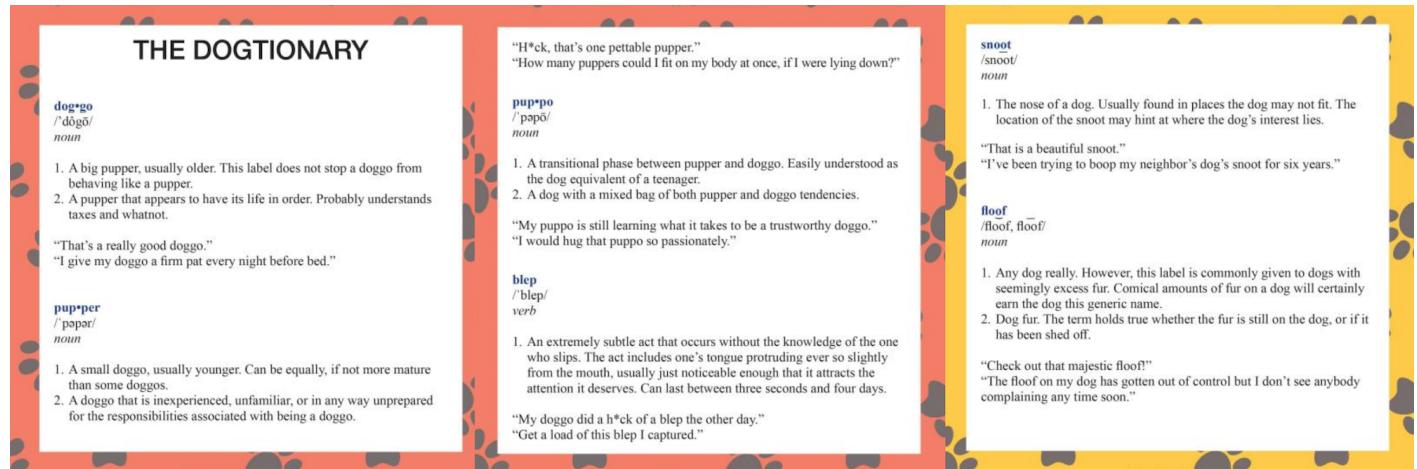
What is the trend of the retweet and favorite counts over 2015 - 2017?

This is trending up especially after 2016 for favorite counts.



What are the most popular category of dog?

Pupper is the most popular. Please refer to the Dogtionary provided by Udacity below :



It looks like this dataset is skewed towards English speakers (Was Google translate was automatically used?)

In [37]: `dfa.language.value_counts()`

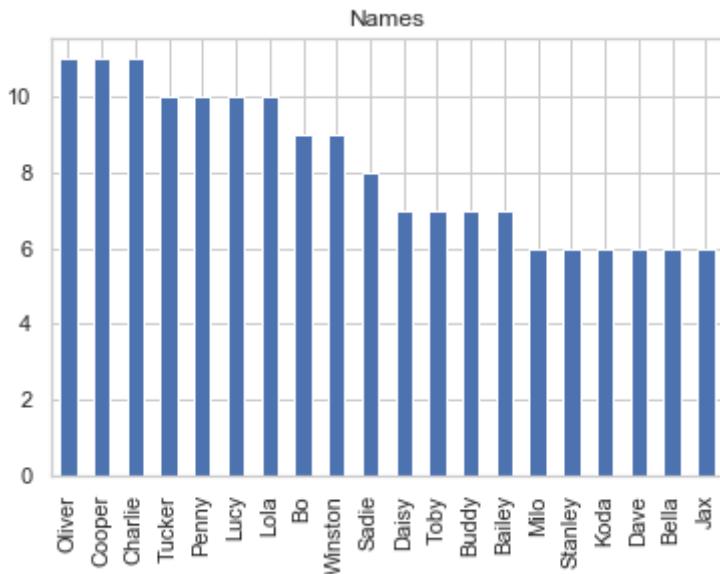
Out[37]:

language	count
English	2313
None	7
Dutch	3
Indonesian	3
Basque	1
Estonian	1
Tagalog	1
Spanish	1
Romanian	1

Name: language, dtype: int64

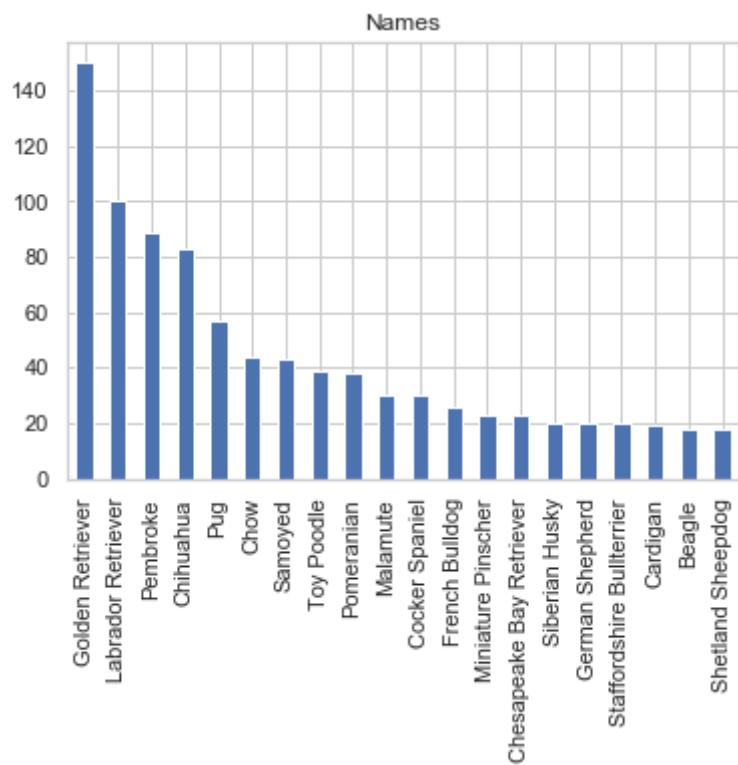
What are the top 20 dog names out of a total of 929 names?

Oliver, Cooper, & Charlie are all really popular in this dataset.



What are the top 20 dog breeds out of the breeds correctly identified as dog breeds?

Golden Retriever is #1



Is a cardigan a dog or a sweater? Let's look!

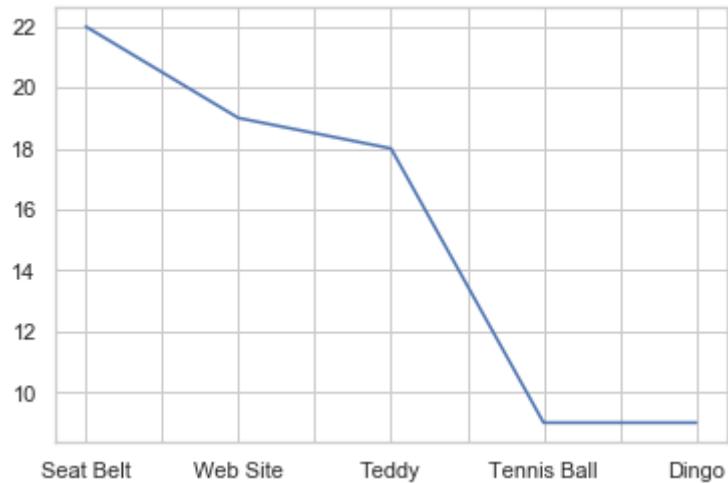
Since I'm not a domain expert, I don't know if a Cardigan is a dog or a mis-classified sweater. Here we find all thumbnails labeled as Cardigan, and glue them together in a single image. Cardigan is a type of corgi breed I learned.

These are mostly corgis but certainly not all!



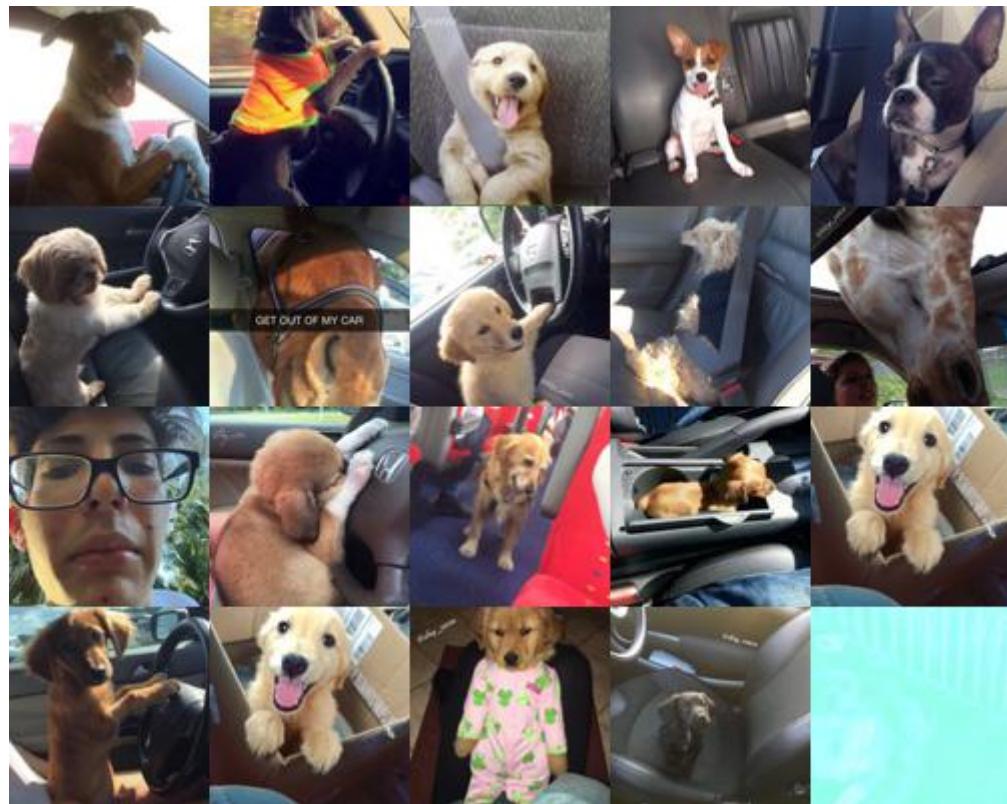
What were the top 5 incorrect predictions for things that weren't dog breeds? (This could be funny!)

Seat belt is very high on the list. Let's look into that.



Let's look and see what pictures were misclassified as seat belt and perhaps why?

We can see how some of these dogs in cars can be identified as seat belt since they are mostly in cars, however, not all reflect the car correlation : Person's face and dog in a box.



Thank you for reading!