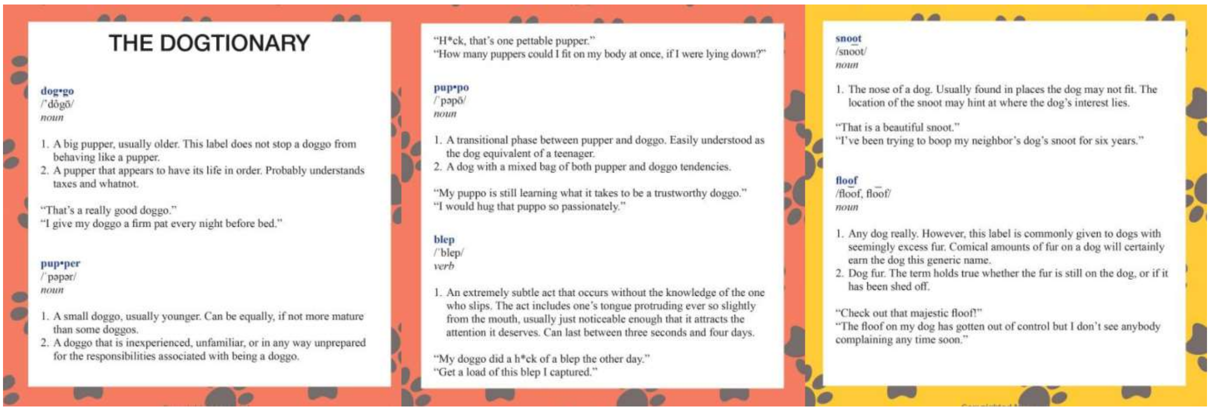
Wrangle Report

Objective

This project, involved the challenge of integrating data from three disparate sources into a single, usable dataset. The final dataset was used to produce meaningful insights of the WeRateDogs Tweets



THE DOGTIONARY

dog•go
/ˈdôgō/
noun

1. A big pupper, usually older. This label does not stop a doggo from behaving like a pupper.
2. A pupper that appears to have its life in order. Probably understands taxes and whatnot.

"That's a really good doggo."
"I give my doggo a firm pat every night before bed."

pup•per
/ˈpəpər/
noun

1. A small doggo, usually younger. Can be equally, if not more mature than some doggos.
2. A doggo that is inexperienced, unfamiliar, or in any way unprepared for the responsibilities associated with being a doggo.

"H*ck, that's one pettable pupper."
"How many puppers could I fit on my body at once, if I were lying down?"

pup•po
/ˈpəpō/
noun

1. A transitional phase between pupper and doggo. Easily understood as the dog equivalent of a teenager.
2. A dog with a mixed bag of both pupper and doggo tendencies.

"My puppo is still learning what it takes to be a trustworthy doggo."
"I would hug that puppo so passionately."

blep
/ˈblep/
verb

1. An extremely subtle act that occurs without the knowledge of the one who slips. The act includes one's tongue protruding ever so slightly from the mouth, usually just noticeable enough that it attracts the attention it deserves. Can last between three seconds and four days.

"My doggo did a h*ck of a blep the other day."
"Get a load of this blep I captured."

snoot
/snoot/
noun

1. The nose of a dog. Usually found in places the dog may not fit. The location of the snoot may hint at where the dog's interest lies.

"That is a beautiful snoot."
"I've been trying to boop my neighbor's dog's snoot for six years."

floof
/floof, floof/
noun

1. Any dog really. However, this label is commonly given to dogs with seemingly excess fur. Comical amounts of fur on a dog will certainly earn the dog this generic name.
2. Dog fur. The term holds true whether the fur is still on the dog, or if it has been shed off.

"Check out that majestic floof!"
"The floof on my dog has gotten out of control but I don't see anybody complaining any time soon."

Fig. 1. Dogtionary

Data gathering

The primary source of information, **twitter_archive**, was obtained through manual download and presented a multitude of quality and tidiness issues upon initial inspection. Two additional datasets, **image_prediction** and **tweet_jsons**, were collected through online methods, with image_prediction utilizing the "requests" library and image_prediction being gathered through the Twitter API (although the provided data was used, as the author's developer account had not yet been approved). While both of these datasets presented fewer problems than twitter_archive, they still required some level of cleaning and refinement.

Data Assessing

Problems with the datasets were identified visually and programmatically. Consistency, validity, accuracy and consistency issues were assessed. The problems were grouped into Quality and Tidiness issues. Ten quality issues, and two tidiness issues were identified.

Data Cleaning

This is perhaps the most challenging part of the wrangling process. Define, code, test was the approach employed to fix the identified issues in the assessment section. The twitter_archive dataframe was quite messy. The first task was to remove records that were irrelevant to the analysis, such as retweets and replies, identified through the "in_reply_to_status_id" and "retweeted_status_id" columns. After this initial culling, missing were inputed by using the twitter profile link, and tweet id. Columns with wrong datatypes were addressed, and inconsistent naming conventions were also treated.

Additionally, four separate columns designated for dog stage were consolidated into a single column. Other minor issues were resolved, as detailed in the code file "wrangle_act.ipynb."

The image_prediction dataset was comparatively straightforward to clean. For image_prediction, only the prediction with the highest confidence was used (as long as it pertained to a dog breed), with all other columns being discarded.

Finally, The "retweet_count" and "favorite_count" columns from tweet_json were extracted and merged with enhanced_archive via the "tweet_id" column. This was in turn merged with the image_prediction dataset to create the final dataset.

Conclusion

While the data was initially far from ideal, a significant amount of effort resulted in a well-formed, usable dataset. Of course, there is always room for further improvement, but the end result of the data wrangling process was a marked improvement over the original sources.