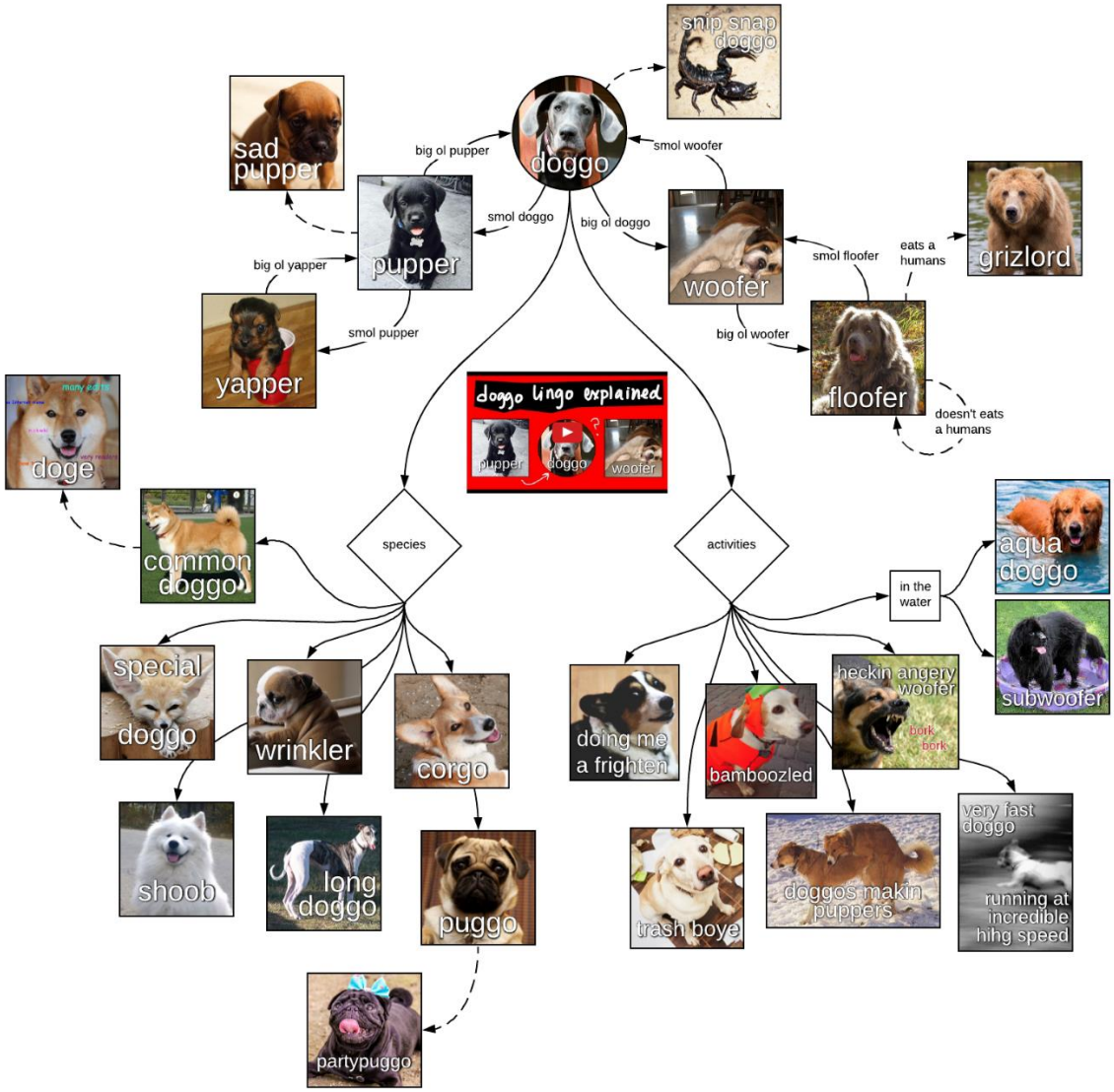
Wrangle Report

INTRODUCTION

This project was harder than I first thought. Not because I don't use twitter, but because we had to use an API to gather the data. After gathering the data the advance of my project became faster. Our task was to gather, asses, clean and analyze data about dogs from given files and the twitter API. The tweets belong to the twitter user @dog_rates. Because I wasn't familiar with the different words used in English for dogs, I made a short investigation. During this project I learned a lot about dog names. I attach here a help image, and on Youtube there is a very good video.

(source: https://lucidchart.zendesk.com/hc/en-us/articles/360001139063-Template-Doggo-Diagram)



DATA GATHERING

The three data sources:

- twitter-archive-enhanced.csv: This file was the easiest to gather, because we was able to simply download it.
- image-predictions.tsv: This was hosted on an Udacity server and we were able to download it programmatically using the Requests library.
- API data: We had to use the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

image-predictions.tsva csv file of twitter data, a dog image prediction file and the twitter API. All sources used the tweet_id of the tweets, so it was possible to merge them into one dataframe.

DATA ASSESSING

It wasn't easy to find 8 quality issues, but I think it would have been easier if I try to find issues in columns which I intended to drop later.

Assessing the data programmatically and visually was useful. For finding issues I needed both approaches.

DATA CLEANING

Cleaning the data had two parts, solving the tidiness issues, and solving the quality issues. I didn't solve missing value issues during this project.

All sources used the tweet id of the tweets, so it was possible to merge them into one dataframe. Working on one dataframe was easier, and this way we can avoid cases when we do the same work twice on two dataframes. After that I made a dataframe simpler by converting four dog type columns to one.

After that I solved some quality issues: deleting retweets, converting timestamp column to datetime, finding invalid dog names, checking the score values, and at the end I deleted the unnecessary columns.

ANALYZE AND VISUALIZE DATA

This was the last part, there were a lot of possibilities to analyze this dataset, but because the main goal of this project was the data wrangling part, I examined only few features of the dataset.