Wrangle Report

Gathering data

> This project was a great learning experience for me as I found the gathering aspect to be very challenging. Gathering the first two datasets was not too difficult because I felt like it was similar to what I have done in past projects (reading csv files). The first dataset, which contains the bulk of the detail about each listed tweet (up until August 1, 2017), was just a simple csv that was provided.

Next, there was the image predictions file. Gathering this was slightly more difficult because it involved pulling from a url, then turning that url into a tsv file and using pandas to read that. Using the “requests.get” ability was something that I thought made clear sense and was usefully discussed in the course material. I thought this information was particularly interesting because the content of photos was predicted by the network.

The last file to gather was to use the archive to store the tweet data in a JSON file. I had an extreme amount of difficulty here. I first tried to go in order of outputting the archive detail immediately following the image file. Upon further investigation, I realized that the tweet_ids from the archive file looked weird, so I dug into this further and noticed the actual ids that I needed to use were embedded in the url. I parsed this in the cleansing phase and then ran the API code through. It was a little frustrating when my code did not work at the beginning, but I was very excited to solve the problem.

This phase was the most challenging aspect of the project, but also the part where I learned the most. I can definitely see myself using these approaches to gather data in the future.

Assessing data

When I was assessing the data, I had to ask myself what I wanted to do and what might keep me from effectively accomplishing this.

First, I considered the quality issues of completeness, uniqueness, timeliness, validity, accuracy, and consistency. I also took into account the project instructions of what I should be evaluating. Finally, I considered aspects of the data that may be untidy. To assess, I looked further into the information of each of the dataframes, made sure I was familiar with the information, and moved on to. begin cleaning. The issues I identified are listed in the notebook.

Cleaning Data

To actually clean the data, I corrected most of the issues I identified in “Assessing data”. I decided which ones based on the data I wanted to analyze for my insights and visualization. At the very beginning, I got rid of anything within the image predictions dataframe that did not have any images. Then, I removed retweeted values. I also removed columns as I went that become obsolete and noisy. I made sure the breeds columns were consistent with all being lowercase. I pulled out the tweet_id from the url and replaced this as the value in the archived dataframe. I also cleaned up the source for this df. I changed the timestamp to be datetime format. I made sure I only had tweet_ids that were in all the dataframes. I renamed columns in the image dataframe to be useful. Next, I only wanted to keep photos when they are of dogs, so I removed all the others. Then, I was ready to draw insights!