Wrangle Report

Gathering Data

Data was gathered from several sources. The twitter_archive_enhanced CSV file was provided with the project. The image_predictions TSV file was programmatically downloaded from Udacity's website within the notebook. The tweet_json file was read in line-by-line into a textfile, using Tweepy to interface with Twitter's API. I used the larger datafile – twitter_archive_enhanced – to supply tweet id's, in order to ensure I had the largest possible dataset to pull from. Only 10 of the tweets were not found, and it took approximately 30 minutes to retrieve the requested data – tweet id, number of retweets, and number of favorites.

Following the initial gathering of the data, all files were read into Pandas dataframes, retaining their unique names for ease of use.

Assessing Data

Data assessment primarily consisted of observations with some drilling-down as was deemed necessary. For each dataset, first I opened it in the notebook and generally examined it for inconsistencies, unclear data, untidy data, or obvious errors that might need fixing. Primarily I find that this initial visual assessment is best for finding patterns, rather than one-off issues, like a misspelling somewhere.

I used the .info() and .value_counts() pandas functions extensively, again to look at both patterns in the data (incorrect dtypes, missing data, data that should be missing but isn't) and to catch smaller errors (one-off inconsistencies, for example where a dog was never actually rated, or the wrong fraction was picked up). While examining the data I took notes for myself offline, and kept a running list of quality and tidiness issues that needed to be fixed. These lists are at the end of the Assessment section.

Cleaning Data

My first step was to copy all of the dataframes from the gathering and assessing phases, with the ultimate goal of producing clean datasets while ensuring the original datasets were also retained. (I also found that when I needed to start over, it was significantly easier to re-create the clean datasets, rather than have to start over entirely from the beginning.)

I tackled the quality issues first, to ensure that it was (relatively) clean data that I was transforming in the tidiness section of this phase. Working through the list, I followed the same loop each time: define what needs to be done to the dataset, programmatically execute it, and test to make sure that the code was correct. Although unique to each issue, generally the description was an action or series of actions, followed by the code, followed by visually (or, rarely, programmatically) assessing the dataset to ensure that changes were correct.

I followed the same define-execute-test loop for the tidiness issues to create a master dataset appropriate for analysis.