WeRateDogs Twitter Data Analysis

Wrangling Report

Brian Shin

Sources

I gathered the data from a given CSV, a website, and Twitter's Tweepy API. I used Tweepy to access the API and gather the JSON data for the tweets. I stored the JSON data in a text file, then loaded what I needed into a pandas dataframe.

Process

I audited the data by checking data types, value counts, lower/uppercase statuses, number of non-null entries, and numeric summaries. I combined (inner-joined) all three tables because each column is a feature of the individual tweet. I reshaped the dog stages (floofer, puppo, pupper, doggo) into a single column rather than multiple columns, and fixed names that were labelled incorrectly (lowercase, in 'text' column, etc). Numerators were fixed to be more within a reasonable range. The sources of each tweet was cleaned up to be easier to understand as well. Lastly, I converted several columns to new data types:

- Dog_stage to 1 column as a category
- Tweet_id, in_reply_to_status_id, in_reply_to_user_id to strings
- Timestamps to datatime objects

For each issue, I defined, coded, and tested each changes to make sure they were made correctly.

Storage

All the data was combined in to one master dataframe.