Wrangle Report

By Chenchen Li

Apirl 24, 2018

Introduction

The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs

This report will describe the data wrangling effort made in the weRateDogs project.

Data wrangling consists of:

- Gathering data
- Assessing data
- Cleaning data

Gather Data

Gathering Data for this project composed of three pieces of data as described below:

- Getting data from an existing file (twitter-archive-enhanced.csv) Reading from csv file using pandas
- Downloading a file from the internet (image-predictions.tsv) Downloading file using requests
- Querying an API (tweet_json.txt) Get JSON object of all the tweet_ids using Tweepy

Assess Data

After gathering data, we will next assess and document them visually and programmatically for quality and tidiness issues.

Quality

archive dataset

- Dataset contains retweets, we only want original ratings (no retweets) that have images
- Some Tweets with no images
- Invalid names i.e 'None', 'a', 'an' in 'name' column
- Erroneous datatypes (timestamp, source, dog stages, tweet_id, in_reply_to_status_id, in_reply_to_user_id)
- In several columns null objects are non-null (None to NaN)
- The numerator and denominator columns have invalid values
- Tweet ID# 810984652412424192 doesn't contain a rating
- Tweet with more than one #/# sometimes have the first occurence erroneously used for the rating numerators and denominators

image_predictions dataset

- Missing values from images dataset (2075 rows instead of 2356)
- Some tweets are have 2 different tweet_id one redirect to the other

Tidiness

- In 'archive', the columns 'retweeted_status_id' 'retweeted_status_user_id' and 'retweeted_status_timestamp' are useless after we get rid of retweets.
- Dog "stage" variable in four columns: doggo, floofer, pupper, puppo
- 'json_tweets' and 'image_predictions' should be joined into 'archive'

Clean Data

Cleaning our data was probably the most time consuming part of the entire project but also again the most helpful. Our process was Define, Code and Test and we were always making a copy of tha dataset even we made the copy in file to test the change before applying to the main dataset.

Store Data

I stored the clean datafile in a CSV file named "tweeter_achive_master.csv".

Remaining issues

- I decided not to change the numerators or denominators for the dog ratings because they probably were not mistakes. They were humorous deviations. If actual analysis is done, then exclude those values.
- Some of the variables did not seem necessary or useful (e.g., in_reply_to_status_id), but I left them in case someone else could make use of them. The cost of storing the additional data is minimal.