Wrangle Report

WeRateDogs Twitter Data
Yeong Heo

This report is intended to describe the wrangling efforts related to completing "WeRateDogs" project. The wrangling process consists of 1) gathering the data, 2) assessing the data, and 3) cleaning the data.

1) Gathering data

This project requires to gather data from three different sources. The first is to download manually and load a csv file. The second is to download a file programmatically using the Requests library with the given URL. The third is to 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. Then read the file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

2) Assessing data

After gathering each datasets from three different sources, they were assessed visually and programmatically for quality and tidiness issues. The datasets were manageable size allowing quick scanning through the rows and filter function for preliminary investigation. Then the datasets were investigated through Python's Pandas functions using various functions for programmatical assessment. There were at least eight (8) quality issues and two (2) tidiness issues in the datasets which listed below and were addressed during the cleaning process.

Quality Issues

1. Convert tweet_id to 'string'
2. "tw_archive_df" has 181 retweets and 78 replies which may not needed in our analysis
3. Convert timestamp to 'datetime' data type
4. Fix incorrect dog names in names column
5. Remove unnecessary number (+0000) from timestamp
6. simplify source text
7. Drop rows with no images ("tw_archive_df" has 2356 rows while the "image_predictions_df" has only 2075 rows. The image predictions do not contain photos beyond August 1st, 2017.)
8. Convert source and dog_type columns to 'category' datatype

Tidiness Issues

1. three dataframes exist, only one dataframe should be enough for this project
2. there are three breed prediction columns - having one column with most confident prediction will sufficient
3. there are three dog stage columns - they should be melted into one column

3. Cleaning data

As the final stage of the wrangling process, each of the documented issues went through the cleaning process which consisted of ‘define, code, and test’ practice. While most of the issues were fixed using various libraries and functions, they were a few things that had to be addressed manually.