Wrangle Report

By: Deepak Singh

The data wrangling project was a great learning experience for me. I learned a lot about the wrangling process and some useful Python libraries. Data wrangling consists of :

- Gathering Data
- Assessing Data
- Cleaning data

Gathering Data

Data for this project came from 3 various sources described below:

1. The WeRateDogs Twitter archive. It was a file on hand. I download this file manually by clicking on the link provided in the project description and saved it as: twitter_archive_enhanced.csv
2. The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv. I wrote a request and response piece of code and downloaded the file in the local directory. After the above I read the tsv file in the pandas dataframe
3. I extracted each tweet's retweet count and favorite ("like") count. Using the tweet IDs in the WeRateDogs Twitter archive, I queried 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 called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame

Assessing Data

After gathering the data in the gather step, I assessed the data both visually and programmatically. Programmatic assessment includes running methods such as info (), describe (), value_counts() etc to check any unusual data that does not follow quality and tidiness guidelines. We document data quality and Tidiness issues. Below are some of the data quality and tidiness issues. There may be more issues present and we can re-iterate the assess process anytime in data wrangling.

Data Quality Issues:

archive dataframe:

Exclude any tweet that is a retweet. Two fields are significant. First, the retweeted_status contains the source tweet (i.e., the tweet that was retweeted). The present or absence of this field can be used to identify tweets that are retweets. Second, the retweet_count is the count of the retweets of the source tweet, not this tweet.I isolate all rows in the retweeted_status column that have a value

and delete it from the archive dataframe. This will remove tweets that are a retweet from the dataframe.

- Some ratings numerator and denominator are incorrectly extracted
- Fix rating numerator that have decimals
- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be integer instead of float
- retweeted_status_timestamp, timestamp should be datetime instead of object (string)
- Names of dogs are miss labelled, mispelled or missing. Cross-reference text data with Names column
- 181 records have a retweeted_status_id, these will need to be exluded from the dataset
- Remove extra characters after '&' in archive['text'].
- In several columns null objects are non-null

image_predictions dataframe:

- Rename column names to more informative (pic_1, pic_2..)

Tidiness Issues:

Some of the structural issues in the involved dataset are captured below.

- Various stages of dogs in columns instead of rows in archive dataset
- keep only tweet_id and jpg_url columns in image_predictins dataframe
- join all three datasets

Cleaning Data

Cleaning the data is the last step in data wrangling. I worked on the issues identified in the assess step. For all the issues identified above, I followed the below step for the cleaning process: Define – define in detail what is needed to fix the issue, Code – write python code to fix the issue, test – test the corrected data visually and programmatically. All the cleaning was done on the copy of the datasets. Cleaning is an iterative step where in case we spot any more issues we can revisit assessment phase and document it and then subsequently clean it. After cleaning the quality and structural issues I created a master dataset which I used to gain insights and create visualisations.