Wrangling data for Wrangling and Analyze Data project

By Jennifer Powell

Gather

The first step to wrangle the data for this project was to gather the 3 files required for this project:

1. WeRateDogs Twitter archive (twitter-archive-enhanced.csv), which I downloaded from udacity, and read into a dataframe.
2. The tweet image predictions file (image_predictions.tsv), which was downloaded programmatically and read into a dataframe
3. The retweet and favorite count data, which was queried from Twitter and acquired programmatically using the Twitter API. The tweet_id from the archive file was used to obtain JSON data from Python's Tweepy library. The resulting data was stored one line at a time in a text file: tweet_json.txt. This was then read into a 3rd dataframe.
4. I decided to obtain a 4th file (language-codes_csv.csv) to use for decoding the 2 char code in the language column of the image_predictions table to a readable text. I downloaded this file programmaticcally from https://datahub.io/core/language-codes and read it into a 4th dataframe.

Assess

*Detect and document at least **eight (8) quality issues** and **two (2) tidiness issues***

The next step was to assess the data in the 4 dataframes. I went through each dataframe and looked at the quality of the data (missing data, garbage data, duplicates). These were the quality issues found in each of the datasets:

Quality

`twitter_df` table

1. erroneous datatypes - id, retweet_count and favorite_count should be int; timestamp should be datetime
2. language abbrev should be full name, not cryptic
3. many columns have only 1 value or no data (contributors, coordinates, favorited, geo, place, possibly_sensitive, possibly_sensitive_appealable)

`img_pred_df` table

1. some of the p1 values look to be garbage (eg crash helmet, water bottle)
2. many records are not dogs (p1_dog=False)
3. some records are retweets

`wrd_df` table

1. Some names look to be garbage (eg a, an) clean up and set to 'None'?
2. some records contain retweets
3. Erroneous datatypes (source column,stage should be category, timestamp should be datetime)
4. a handful of denominators aren't 10
5. missing images
6. source column contains extra info, only 4 values
7. stages of dog:doggo, pupper, puppo, and floof(er) some have multiple ratings for same dog
8. missing floof stage, only checked for floofer

Tidiness

Column headers are values, not variable names

- `wrd_df` table -doggo, floofer, pupper, puppo columns contain only 4 values in 4 columns

A single observational unit is stored in multiple tables.

- The source and text of the tweet are stored in both wrd_df and twitter_df. We only want the twitter_df favorite_count and retweet_count so can combine with wrd_df and drop rest of columns in twitter_clean
- jpg_url is repeated in img_pred_df. Only need p1 and p1_conf, can combine these with wrd_df and drop rest of columns

Clean

1. Language abbreviations in twitter_df were replaced using a lookup table.
2. Determined records with missing images and dropped them from wrd_df.
3. Removed records from wrd_df that are retweets.
4. One character names in wrd_df set to ‘None’.
5. Numerators can be greater than 10, but decided to make denominators all 10 since so few.
6. Created subsets of **twitter_df** : 'id','retweet_count','favorite_count','lang' and **wrd_df**: 'tweet_id', 'source', 'timestamp','expanded_urls', 'text', 'rating_numerator', 'rating_denominator','name','doggo', 'floofer', 'pupper', 'puppo' dropping unused columns.
7. Combined the doggo, floofer, pupper, puppo columns into 1 categorical **stage** column.
8. Convert wrd_df.source to category type with 4 values (iPhone,Vine,WebClient, TweetDeck).
9. Deleted rows which weren’t dogs (ie p1_dog == False).
10. Converted twitter_df.retweet_count and favorite_count to integer.
11. Converted wrd_df.timestamp to datetime.
12. Combined the dataframes into one master.