Wrangle Report

by Stefanie Powazny

1. Gathering Data

   I gathered the first data by manually downloading the twitter-archive-enhanced.csv file that Udacity provided to me.

   The second dataset was programmatically downloaded from Udacity's server using the requests function (URL = https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)

   The third dataset was gathered via the Twitter API by using the Tweepy library and stored as a JSON file. Specifically, I gathered information about the favorite count and retweet count for each tweet as well as the tweet id.

2. Assessing Data

   I assessed the data visually (i.e., using the head() and sample() function) and programmatically to find quality and tidiness issues.

   Quality Issues
   Here is a list of quality issues I found for the three datasets:

   a) DataFrame twitter_archive
   - retweeted_user_id and retweeted_status_id column: there are some retweets
   - expanded_urls column: tweets/ retweets without images
   - timestamp: not datetime format
   - name column: none appears 745 (missing data but not NAN)
   - name column: some names are false (O, a, not..)
   - tweet_id: is int, should be type object as no calculation is needed
   - text and rating_numerator column: tweets that include more than one rating and/or decimal numbers, hence, wrong or missing data in the rating_numerator and rating_denominator column
   - pupper, puppo, floofer and doggo column: For 1976 IDs there are no dog "stage" information.
   - pupper, puppo, floofer and doggo column: There are some IDs with more than one dog "stage" information (two dogs are rated).
   - missing column for the fraction of rating_numerator and rating_denominator

   b) DataFrame predictions
   - p1,p2,p3 columns: dog breeds are not consistently lower or uppercase
   - tweet_id is int, should be type object as no calculation is needed
   - img_num column does not contain new information

   c) DataFrame twitter_add_info
   - tweet_id is int, should be type object as no calculation is needed

Tidiness Issues
Here is a list of tidiness issues I found:

- twitter_archive: 4 columns (dogger, floofer, pupper and puppo) for one variable (dog stage)
- predictions: the dog breed prediction could be packed into one column (breed_pred)
- predictions: the prediction confidence could be packed into one column (pred_confidence)
- predictions: jpg_url, breed_pred and pred_confidence should be joined to twitter_archive DataFrame
- twitter_add_info: favorite_count and retweet_count column should be joined to twitter_archive DataFrame

3. Cleaning Data

Before I started cleaning the datasets I created copies of each dataset. Afterwards, I tried to fix every problem programmatically. First, I fixed the quality issues regarding missing/ wrong data. Some issues were fixed in one cleaning step as they were closely related. In one case I had to change the dog stage information manually. Because the code I wrote does apply in most cases, but it does not regard the order of occurrence of the words (if doggo occurs before pupper the dog stage will be doggo even if pupper is the meant dog stage and doggo is only a part of a word like "didodoggo" for ID 817777686764523521). As there's only one case in this dataset where the function does not work properly I changed the dog_stage for this ID manually to pupper.
After this, I fixed the tidiness issues, merged the datasets, and fixed the remaining quality issues.

4. Storing Data

After cleaning the data, I stored the final cleaned dataset as a csv-File: complete_df_clean.to_csv('twitter_archive_master.csv', encoding='utf-8', index=False)