Wrangle Report

By LaShonda Dickson

Data wrangling consist of three parts:

1.) Gathering Data
2.) Assessing Data
3.) Cleaning Data

1.Gathering Data

Data was gathered from three different sources and loaded into Jupyter Notebook title wrangle_act.ipynb.

- ✓ WeRateDogs Twitter archive already given as **twitter_archive_enhanced.csv** provided by Udacity
- ✓ Second data frame was programmatically downloaded from Udacity's service using the request function (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)
- ✓ To obtain df_tweet dataframe, which contains tweet ID, retweet count, and favorite count I set up Tweepy API to acquire the required Keys via Tweepy library. From there, I was able to fetch each tweet imported into a text file called **tweet_json.txt.**

2.Assessing Data

I visually and programmatically assessed the data within wrangle_act.ipynb by utilizing the following Pandas functions: *.head(), .sample(), .columns, .info(), .value_counts(), .unique()*. Also exported this data into excel to have a surface level look, while keeping in mind "Key Points" stated in the Project Motivation. I was able to detect and document the Quality and Tidiness issues as outlined below:

<u>Quality Issues</u>

twitter_archive

- Retweeted_user_id and retweeted_status_id colums: there are some retweets
- Expanded_urls column: tweets/retweets without images
- Timestamp: not datetime formate
- Name column: none appears 745(missing data but not NaN), 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 incorrect 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

Predictions

- p1,p2,p3 columns: dog breeds 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

twitter_add_info

- tweet_id is int, should be type object as no calculation is needed

Tidiness Issues

twitter_archive table

- 4 columns (dogger, floofer, pupper and puppo) for one variable (dog stage)

Predictions

- the dog breed prediction could be consolidated into one column (breed_pred)
- the prediction confidence could be consolidated into one column (pred_confidence)
- 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

First, I created copies of the dataframes before cleaning by outlining and documenting the define, code, and test steps. I started by removing the missing data and then merging the three data frames into one named df_twitter_archive_clean.

For the most part I used functions of Pandas, loops and defined my own functions. Also cleaned some data manually for incorrect dog ratings. I re-extracted, cleaning, and correcting names, ratings, dog stages, and cleaning the tweets with the non-dog images. This projected challenged me to pay close attention to detail and improve my data wrangling skills.

Finally, the cleaned master data set which will be used in the data analysis stored in a csv file name twitter_archive_master.csv