Udacity’s Data Analytics Nanodegree
Project: Wrangle and Analyze Data from @dog_rates

Binh Nguyen,

November 24, 2019

1. Gathering efforts

Three resources are used to make the final dataset in DataFram format: (1) twitter_archive_enhanced.csv, (2): image_predictions.tsv, and (3) tweet_json.txt . Each data after cleaning and tidying up is merged and saved as ‘twitter_archive_master.csv’.

File #1 and file #2 can be downloaded manually to the local computer, and uploaded to the working space on Udacity or programmatically using requests library. File #3 was downloaded using **Tweepy** library with an Twitter API development account. Alternatively, the json file can be imported directly to the workspace as the file was given.

After three files above were available on the workspace. They are imported to the Jupyter notebook using **pandas** libraries or **read** function with **json** library. The quality check started from completeness, validity, accuracy, consistency. The tidiness check is based on Hadley Wickham’s paper in which a tidy dataset should have one row for one observation, one column for one type of observation, and one table forms one unit of observation.

The tweet data can be retrieved with an API Tweet account. The Python script to get data with Tweepy library was given by the course work. There is approximate 35 minutes to gather the data with tweet_id in the -enhanced.csv. The json file then be extracted to get **tweet_id**, **favorite_count** and **retweet_count.** These three lists are finally form the third DataFrame using pandas.

2. Quality check

completeness:

1. rows in -enhanced.csv, image_predictions, .json files containign different rows (2356, 2075, and 2333)
2. -enhanced.csv contains three groups of columns that has less than 2356 rows (*in_reply_to_*, *expanded_urls*, *retweeted_status_*)

validity:

3. column timestamp, retweeted_status_timestamp is not in datetime format with a trailing of *+0000*
4. a few other columns should be in **category** data type such as *name*, s*ource* columns
5. the source columns contains a *html* tag, and the text should be extracted

accuracy

6. -enhanced.csv contains data of retweets, this may need to clean up

7. in -enhanced file, *name* column contains name of dogs that don't much sense as as *a, an, the*

8. the *rating_denominator* is not all as 10

9. the *rating_numerator* contains maximum value of 1776 with another few outliers

consistency

10. the *rating_demominator* needed clean up so the rating is consistent

11. the name of dogs are in lower case and title, they should be consistent

Tidiness¶

1.Columns doggo, floofer, pupper, puppo should be presented in one columns as the **stage**

2.the image_predictions.tsv has three rounds of prediction. The table should be tidied up to present one type of observation

3. three separate files (or dataframes) are related to **WeRateDogs** tweet data. They should be merged into one dataframe to make a complete observation unit

3. Wrangling

- The three files can be imported and merge to from a master file. These approach created a rather large file with 31 columns. A large number of columns can be tricky to carry out .melt function. Alternatively, each DataFrame can be cleaned up individually and merged later. I selected the second approach.
- Helpful function to explore the datasets including **.dtypes, .head(), .sample(), .info(), .describe()** to see the plain field of data.
- String functions were used to remove trailing “+0000” timestamp or to change the text format more consistency.
- Drop and remove duplicate functions were used remove unused or less relevant columns
- .**melt** or customized function with .**apply**() to combine the same type of observation to one columns