Ainur Artay
Data Analytics NanoDegree
January 2020

DATA WRANGLING REPORT

PREPARATION

I loaded the necessary modules in the Python workspace, including Tweepy, an API from Twitter and set the option to display the full string.

GATHERING THE DATA

I gathered the data from 3 sources in the following ways:

- By manually loading twitter_archive_enhanced.csv file provided by Udacity
- By uploading image-predictions.tsv using the given link
- Programmatically uploading the Tweet JSON data from Twitter

Each set of data was observed on completeness and validity, in order to build an idea on how to clean it further.

CLEANING THE DATA

I cleaned the initial messy data to obtain good quality and tidy data. The following criteria were used to achieve this:
Clean data -- complete, accurate, valid and consistent.
Tidy data -- each variable forms a column and each observation forms a row.
Steps taken to meet these criteria are outlined below:

Quality:

1. Dropped all retweet rows with non-null values in retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp
2. Dropped all reply rows with non-null values in in_reply_to_status_id and in_reply_to_user_id.
3. Converted the timestamp column data type to datetime
4. Replace the source string with the display portion of itself. Extract the string between `<a href="">` and `</a>`
5. Dropped tweets where rating_denominator is not 10
6. Dropped tweets where rating_numerator >= 15
7. Dropped tweets without `expanded_urls`
8. Dropped tweets with missing json_data. Change data type for retweet_count and favorite_count to `int`
9. Replaced lowercase words in the name column with "none"

Tidiness:

1. Dropped all columns exclusively related to retweets

2. Dropped all columns exclusively related to replies
3. Created a categorical variable (/column) *stage* to capture (and replace) the variables (/columns) *doggo, floofer, pupper, puppo*
4. Joined the *retweet_count* and *favorite_count* columns from JSON table to the archive table on *tweet_id*
5. Kept only tweets with images by creating columns *breed* and *confidence* in the predictions table (for quality), then inner-join it to the archive table
6. Dropped the *rating_denominator* column. Renamed the *rating_numerator* column to *rating.*
7. Dropped the *expanded_urls* column. Reordered columns bringing numerical columns to the left.

(detailed steps were noted in the wrangle_act.ipynb file)