5/26/2021
DA Plager
Udacity Data Analyst Nanodegree, Project 4
Data Wrangling Report

I. Project Goal: Wrangle @WeRateDogs Twitter data.

II. Gather

Data was gathered from three sources, which included:

1. A directly available comma-separated flat file of enhanced @WeRateDogs Twitter data (twitter-archive-enhanced.csv).
2. An internet "requested" byte-formatted tab-separated flat file (image-predictions.tsv) containing three predictions of dog breed from a single image respectively associated with each unique tweet within the twitter-archive-enhanced.csv dataset.
3. Respective "extended tweet" JSON strings for each unique tweet_id within the twitter-archive-enhanced.csv dataset, accessed via Twitter API and its access library, tweepy, and subsequently used to extract each tweet's "retweet" and "favorite" counts (saved as retweet_fav_counts.csv).

These data were respectively loaded into three separate Pandas DataFrames, **df_1, df_2, and df_3**.

III. Assess

In general, each of the three DataFrames (df_1, df_2, and df_3) were visually and programmatically assessed for data quality (missing, invalid, inaccurate, or inconsistent content) and data tidiness (column, row, or table structure).

More specifically, because a large portion of potentially missing data (in the form of NaNs and "None" strings) appeared to be due to inappropriately structured data and because it is often best to address these two issues first, missing data and data structure were simultaneously assessed first. Other potential data quality issues (additional missing, invalid, inaccurate, or inconsistent data) were subsequently assessed.

Visual assessment of df_1, df_2, and df_3 was performed before programmatic assessment and involved a combination of viewing each dataset in Jupyter Notebook (e.g., better for 'tweet_id' viewing) and in Excel (e.g., better for full 'text' and 'img_url' viewing). Any potential data issues (labeled v1 through v15) were gathered into a "Visual Assessment Issues" section.

Programmatic assessment was also performed focusing on missing data and data structure issues first and then on other data quality issues by separately assessing numeric columns and non-numeric columns. Any potential data issues (labeled p1 through v19) were gathered into a "Programmatic Assessment Issues" section.

IV. Clean

The three gathered and assessed DataFrames, df_1, df_2, and df_3, were **copied and named df1_clean, df2_clean, and df3_clean**, respectively. Missing data and data structural issues involving all three DataFrames were cleaned first.

Per Udacity's "key points" stating that "only original ratings (no retweets) that have images" and to conform to the rules of tidiness, as part of this initial cleaning:

A. Dropped the 259 reply and retweet rows and dropped all five `reply` and `retweet` columns from df1_clean.
B. Generated a new 'dog_stage' column with a 1-to-1 relationship to the 'tweet_id' column in df1_clean to capture the dog stage information (i.e., doggo, floofer, pupper, and/or puppo) initially in the four dog stage columns (doggo, floofer, pupper, and puppo).
C. Merged all columns of df2_clean (i.e., image prediction) with df1_clean based on 'tweet_id' and renamed as **df1_clean_plus**.
D. Merged the 'retweet_count' and 'favorite_count' columns of df3_clean with df1_clean_plus (eliminating the need for df3_clean).
E. Dropped the residual four dog stage columns (doggo through puppo) from df1_clean_plus.

After cleaning the preceding structural issues and associated missing data issues, a variety of other data quality issues (additional missing, invalid, inaccurate, or inconsistent data) were cleaned.

As alluded to above, the final df1_clean_plus was saved as a comma-separated value file, i.e., **twitter_archive_master.csv**.