Wrangle Report

1 Wrangle Report

1.1 Gather Data

Data is gathered from 3 file resources and saved as 3 Data Frames: *df1*, *df2*, *df3*.

1.1.1 1. Gather data from file on hand

Use pd.read_csv() to read data from existing file *twitter-archive-enhanced.csv* and save it as *df1*.

1.1.2 2. Download file using Requests library and URL

Download file *image_prediction.tsv* programmatically from the Internet and store data in *df2*.

1.1.3 3. Gather data from twitter API using Python’s Tweepy library and store data

NOTE TO REVIEWER: this student had mobile verification issues so the following Twitter API code was sent to this student from a Udacity instructor Tweet IDs for which to gather additional data via Twitter's API.

1.2 Assess Data

1.2.1 Quality

· Tweet_ID in df1 has a missing data.
· Erroneous datatypes in Tweet_ID, in_reply_to_status_id, in_reply_to_user_id and timestamp.
· df1 shouldn't have a retweets because only original ratings is needed. Also missing images in ratings and some ratings are wrong.
· nulls in df1 represented as 'None' in columns 'name', 'doggo', 'floofer', 'pupper','puppo'.
· Also in df1 some dog names are not correct.
· In df2 erroneous dog names there is no column for each dog phase.
· Unuseful columns In df1 should be removed 'retweeted_status_id' 'retweeted_status_user_id' and 'retweeted_status_timestamp' .

1.2.2 Tidiness

· These columns in df1 'doggo', 'floofer', 'pupper','puppo' represent one variable so it should be merged in a one column named 'phase'.
· Rating_numerator and denominator should be compined in one variable rating.
· df3 should be part of df1
· The information about one type of observational unit is spread across three dataframes, So merge all dataframes in order to create one master dataset.

1.3 Clean Data

Copy *df1*, *df2*, *df3* as *df1_clean*, *df2_clean*, *df3_clean*.

- Delete retweets and observations without ID
- Delete unusual columns 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'.
- Not all ratings have images and we only needs rating with images. So Delete observations without image
- Drop columns 'doggo','floofer','pupper','puppo'. Replace 'None' with np.nan
- Create 'phase' column to represent a dog phase of life.
- Join df3 table to df1 table on tweet_id.
- Convert timestamp to datetime data type.
- Convert in_reply_to_status_id and in_reply_to_user_id to string data type.
- Correct column 'name' convert wrong names with np.nan.
- correct the wrong value observations of rating_numerator and rating_denominator, if rating_denominator > 10 and divisible by 10 use the quotient as divisor to divide the rating_numerator, if the numerator is divisible assign the quotient as the rating_numerator, then the rest records if the text column contains any fraction with denominator 10 update the rating_denominator to 10 and update the rating_numerator with the numerator value of this fraction. Create new column rating=rating_numerator/rating_denominator and then drop rating_numerator and rating_denominator also drop oberservations with extreme ratings.
- Create new columns prdct_breed and prdct_conf.

1.4 Store Data

Store the clean data frames df1_clean in a CSV file named twitter_archive_master.csv after merging the data frames, and also store and df2_clean in additional file 'twitter_image_predictions.csv'.