Wrangle Report

Introduction

This report aims to show the effort of wrangling the data for Udacity Data Analysis Nano Degree, the datasets used in this project twitter archive for the @dog_rates account (WeRateDogs). WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. WeRateDogs has over 4 million followers and has received international media coverage.

Project Outline

- Gathering Data from Different sources
- Assessing data
- Cleaning Data

Gathering Data

The datasets used in this project as follow:

- The WeRateDogs Twitter archive. this file to us by Udacity.
- The tweet image predictions, is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using

the Requests library and the following
URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

- Query the Twitter API for each tweet's JSON data using
  Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

Assessing Summary

After gathering each of the above pieces of data, and after I see each dataset

Quality :

- convert the type of id, tweets_id for all datasets into a string

1 - tw_df_enhanced :

- the name column in this dataset have many invalid values and it seems like regular words in lowercase
- should change the type of the timestamp column
- there are 181 retweets
- 78 tweets are replies
- the expanded_urls column has 59 null URLs
- 20 records have rating_denominator above 10.
- Delete columns that won't be used

2 - df_image_predictions :

- will be filtered to shows only if the prediction is true

3 - json_tw :

- there 179 that are retweeted not original tweets
- 29 records are quoted tweets
- 78 tweets are also replies

4- master_df :

- removing rows with null values in name jpg_url img_num p1 p1_conf p1_dog p2 p2_conf p2_dog p3 p3_conf p3_dog life_stage
- after melt function its generate to many dublicated id tweets so i will remove it

Tidiness :

- in tw_df_enhanced dataset there are four columns (doggo, floofer, pupper, and puppo) can be melted in one column
- merging the datasets into one data frame rather than three

Cleaning Data:

first, I take a copy from each dataset in case I made a mistake so I can retake a copy from the original version.

I clean each dataset separately, and I keep only original tweets and removes the retweets, replies. also, there were some issues that I fix it manually like the mistakes of reading the data from the tweet. to make the data frame tidy, I use the melt function to merge 4 columns of the dog stages into one column.
after I meet the points that I mentioned in the assessment section, I merged the datasets into one data frame called master_df based on the id of the tweet.