Wrangle Report

WeRateDogs project is a part of Data Analyst Nanodegree program. This report briefly documents the data wrangling steps of Wrangling and Analyzing process. In this assignment, I used data provided by twitter account @dog_rates to perform analysis and deliver ultimate insights.

The wrangling process is divided into four main sections:

- Gathering
- Assessing
- Cleaning
- Storing/ Visualizing

Gather

The dataset being wrangled is the combination of three separated data sources:

1. Tweet archive dataset of Twitter user @dog_rates, a.k.a WeRateDogs, which contains tweet data for around 2000+ tweets.
2. Retweet and Favorite count dataset gathered by querying from Twitter API.
3. Image prediction dataset that using neuron network to classify breeds of dogs

The Twitter archive dataset could be downloaded manually from a link provided by Udacity. For the second dataset, I used Tweepy library to query Twitter API, this process took around 20 minutes of running time. After querying each tweet ID, I wrote JSON to require the file and then read this file into a pandas data frame. Finally, I download "image_predictions.csv" programmatically by using request library of python.

Assess

After gathering three different datasets, I combined them into one master data for examining both visually and programmatically. Several Quality issues and Tidiness issues have been identified throughout the process, which is plainly outlined as below

Quality

twitter_archive table

- erroneous data type
    - timestamp, retweeted_status_timestamp
    - tweet_id, retweeted_status_id, retweeted_status_user_id
    - retweet_count, favorite_count
- Tweet data contains unoriginal tweets such as retweets and replying tweets.
- Tweets have no images.
- Tweets have inappropriate sources.
- Null values represented as non-null (None) values in name and dog stage columns.
- Name column contains missing/ mislabelled names of dog.
- Missing/ misspelled name of dog where it could be extracted from text column.
- Numerators in decimal numbers are extracted incorrectly.
- Numerators and denominators extracted inaccurately where text record has more than one fractions.

image_predictions table

- erroneous data type: tweet_id
- missing records (2356 instead of 2075)

tweet_data table

- missing records (2345 instead of 2356)

Tidiness

- four columns represent one variable dog stages (doggo, floofer, pupper, puppo) in twitter_archive table
- tweet_count and image_prediction tables should be joined with twitter_archive table

Clean

As the issues have been carefully pointed out, I resolved every single case by three steps Define, Code, Test respectively using abundant pandas methods and library. Most of the cleaning steps are carried out programmatically, manual practices are acceptable but are not preferable.
The cleaned dataset would be used for further steps of analyzing and visualizing to extract value insights afterward.

Analyze and Visualize

The cleaned dataset should be saved into one final file for additional work later on if necessary, I did save it to "df_master.csv" file. On doing analyzing task, I performed some calculations on distinct variables and plot these results simultaneously. Visualization truly helps us acquire a better understanding of data in a matter of a second.