Wrangling WeRateDogs Twitter Data

Data wrangling includes three parts:

1. Gathering Data
2. Assessing Data
3. Cleaning Data

1.Gathering Data

I gathered data from three different sources in a Jupyter Notebook titled **wrangle_act.ipynb**.

- The WeRateDogs Twitter archive were already given as **twitter_archive_enhanced.csv***.* I downloaded it manually and created **df_*twitter_archive*** data frame.
- I downloaded **image_predictions.tsv** file programmatically using the Requests library and the given url and created **df_image_predictions**. This consists tweet image predictions.
- To obtain the **df_tweet** data frame with tweet ID, retweet count, and favorite count I did the following: I set-up tweepy API with the required keys by using Tweepy library. Then I fetched each tweet one by one and write into a file called **tweet_json.txt**. Since it is a long process I put this code into a separate python file called **fetch_twitter_data.py**. Then I read **tweet_json.txt** file to create the **df_tweet** data frame.

2. Assessing Data

I assessed the data in **wrangle_act.ipynb** both visually and programmatically. I used *.head(), .value_counts(), info(), .isnull(), .notnull()* functions of Pandas to assess the data programmatically. I also exported the data into Excel to have deeper look. Then by taking into account the ‘Key Points’ which are stated in the project motivation, I detected and documented quality & tidiness issues as the following:

Quality Issues

df_twitter_archive table

- 181 Retweets
- Missing data in expanded_urls (Tweets without images)
- Not all images are dog images
- The ratings aren't all correct
- Incorrect dog names (a, an, the, etc.)
- Missing values in dog names (represented as None)
- Erroneous data type (tweet_id, timestamp)
- Hard to understand the source from source column
- Dog stage's type to categorical

df_image_predictions table

- Missing records (2075 instead of 2356, some of them are probably the retweets, replies or tweets with no images)
- Image predictions that are not dog breed
- Lowercase breed names in p1, p2, p3 and '_' is used instead of space.

Tidiness Issues

- Merge three data frames.
- Drop unneeded columns.
- Combine dog "stage" columns (i.e. doggo, floofer, pupper, and puppo) into one column.
- Combine rating_numerator and rating_denominator columns into one column.

To detect all the issues that I listed was not easy at the beginning, so after I started cleaning data I discovered more issues and updated my list.

3.Cleaning Data

First, I created copies of the data frames before cleaning data. I cleaned data by documenting the define, code, and test steps of the cleaning process. I started with dealing the missing data and then merged three data frames to work on a single data frame named as **df_twitter_archive_clean.** After that, I tried to solve the other quality and tidiness issues in a logical order.

I mostly used functions of Pandas, loops and defined my own functions couple of times. I also cleaned some data manually for incorrect dog ratings. I spent quite time on re-extracting, cleaning and correcting names, ratings, dog stages, and cleaning the tweets with the non-dog images. Overall, I believe that this project challenged me to improve my data wrangling skills.

Finally the cleaned master data set which will be used in data analysis is stored in a csv file named **twitter_archive_master.csv**.