DATA WRANGLING REPORT

PROJECT OBJECTIVE

Wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "*Wow!*"-worthy analyses and visualizations.

DATA WRANGLING STEPS

STEP 1: GATHER

This phase required gathering data from three different sources:

- Reading data from a given CSV file: For this gathering process, the csv file containing 2037 records of the weRateDogs tweet archive was provided and read into a data frame called df_archive.
- Downloading a TSV file from a given url: The second dataset required us to download the file from a url provided programmatically and load in a dataset. This was stored in the df_image dataset.
- Using Tweepy and Twitter API: For the final dataset, we were required to read the json file returned when we call the twitter api and store in a txt file, and then read into a dataframe. This was stored in the df_json dataset.

STEP 2: ASSESS

After gathering the dataset, both manual and programmatic assessment was done on each of the dataframes. During this process, a number of observations on quality and tidiness were made:

There was an interesting point to note about some of the ratings extracted from the text. I noticed that some ratings had an aggregate of the number of dogs in the picture, for example, similar cases where a rating of 88/80 was given to a group of 8 dogs all together, meaning the individual rating for each dog is actually 11/10. This was corrected during the cleaning phase.

Quality

archive dataset

1. some records are retweets
2. erroneous dog stage classification
3. some records don't have images
4. Inaccurate rating values. Some texts contain dates/values in the same format as the ratings which have then been mistaken for ratings.

5. Some columns have no values

image dataset
1. non-descriptive column names

tweepy api dataset
1. retweet counts and favorite counts are objects instead of integers
2. deleted tweets

Tidiness
1. Retweet count and favorite count should be part of the archive dataset.
2. The image dataset should be part of the archive dataset.
3. three different confidence score predictions and breeds(only one, ié the highest confidence score is required for this analysis)

STEP 3: CLEAN

Most of the cleaning up of the identified problems were done programmatically as can be seen in the attached ipyb file.

STEP 4: STORE

Finally, after thoroughly cleaning the three datasets, they were merged into one master dataset, redundant columns were eliminated as well as columns that were not aiding our analysis. This merged dataset was written into a .csv file.

Fig.1 Gives the final information of our master dataset

```
In [310]: df_archive_clean.info()

<class 'pandas.core.frame.DataFrame'>
Int64Index: 1971 entries, 0 to 1970
Data columns (total 16 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   tweet_id            1971 non-null   int64
 1   img_num             1971 non-null   int64
 2   predicted_breed     1971 non-null   object
 3   confidence_score    1971 non-null   float64
 4   is_dog              1971 non-null   bool
 5   timestamp           1971 non-null   datetime64[ns, UTC]
 6   text                1971 non-null   object
 7   rating_numerator    1971 non-null   int64
 8   rating_denominator  1971 non-null   int64
 9   name                1971 non-null   object
 10  doggo               1971 non-null   object
 11  floofer             1971 non-null   object
 12  pupper              1971 non-null   object
 13  puppo               1971 non-null   object
 14  retweet_count       1971 non-null   int32
 15  favorite_count      1971 non-null   int32
dtypes: bool(1), datetime64[ns, UTC](1), float64(1), int32(2), int64(4), object(7)
memory usage: 297.4+ KB
```

Fig.1 df_archive_clean.info()