Wrangle_report on We Rate Dogs Project

The wrangle Process consists of three important steps:

1. Gather
2. Assessment
3. Cleaning

Gathering Process:

In this project I have been given three files

1. **Enhanced Twitter Archive**, which is manually downloaded
2. **Image Predictions File**, which is programmatically downloaded from Udacity servers using url
3. **Additional Data via the Twitter API**, the data has to be programmatically downloaded from Twitter using Twitter API for the Tweet ID contained in the Enhanced Twitter Archive file downloaded in step 1

Assessment Process:

Each file has been assessed manually and programmatically to find the data quality and data tidiness issues and are summarized below

Data Quality Issues:

Enhanced Twitter Archive File:

1. Tweet ID field should be an object data type, not integers or floats because they are not numeric and aren't intended to perform calculations.
2. **Dog Names:** In the name column, there are several values that are not dog names, like 'a', 'the', 'such', etc. Notice that all of these observations have lowercase characters, an important pattern that could be used to clean up this field. Another way is to drop duplicated values
3. There are lot of missing values such as 'in_reply_to_status_id' and 'in_reply_to_user_id'. This indicates not all tweets have replies. there are only 78 replies
4. data type of 'timestamp' and 'retweeted_status_timestamp' is incorrect. It should be converted to 'datetime'
5. There are 181 retweets such as 'retweeted_status_id', 'retweeted_status_user_id' and 'retweeted_status_timestamp'
6. There are missing urls in 'expanded_urls' column
7. There are 23 'rating_denominator' that are not equal to 10

Image Predictions File

8. Tweet_id is showing as integer. It should be converted to string

9. There are unequal number records between predictions and archive file

Twitter API:

10. There are 26 missing values

Data Tidiness Issues:

Enhanced Twitter Archive File:

1. The 4 different columns doggo, floofer, pupper and puppo, are all relative to the same variable that identifies the stage of dog. So, we can melt these columns into a single column named "dog stage"

Image Predictions File and Json File:

2. Identify the best predicted breed type based on confidence and create two new columns with breed and confidence

3. The df2 and df3 are part of the same observational unit as df1 but there are three separate tables so they should be merged and stored in a file called twitter_archive_master.csv, as per project instructions.

Additional Cleaning:

After merging the above three files the following issues were identified and subsequently needs a cleaning

- retweet_count, favorite_count are converted to float. this should be fixed
- dog_stages to be converted from string to category

Cleaning Process:

Each assessment step regarding the data quality and tidiness issues have been cleaned using Define, Code and Test process

Post cleaning, the datasets have been merged to create a **`twitter_archive_master.csv`**