Data Wrangling Report

Data Gathering:

First, data was gathered from three different sources using different techniques as follows:

- The **WeRateDogs** Twitter archive data (**twitter_archive_enhanced.csv**) was directly downloaded and imported into a data frame called **df_twitter_archives**.
- Use of the **Requests** python library to download the tweet image predictions (**image_predictions.tsv**) from a link provided by Udacity. This data is then written and assigned to a data frame called **df_image_predictions**.
- Use of the **Tweepy** library to query additional data via the Twitter API. After going through and understanding the code and procedure provided by Udacity for using **Tweepy** to extract data from Twitter, I created twitter developer account and ran the code to obtain the required data. This was then stored in a data frame, **df_tweet_status**.

Data Assessment:

After collecting all the required data, I proceeded to assess it both by visual assessment and programmatic assessment techniques. This was a crucial stage of my data wrangling since I was able to test and identify the quality/tidiness issues which were to be cleaned in the next stage. The following lists the actions performed to achieve this goal:

1) Viewing the records of each of the three data frames separately.
2) Using **info()** function to view information on the columns, datatypes and missing values in each of the three data frames.
3) Confirmation of the columns that are duplicated across all the three data frames.
4) Confirmation of the ranges of the rating numerator and rating denominator in the **df_twitter_archives** dataset.
5) Investigated the **doggo, puppo, pupper** and **floofer** columns from the **df_twitter_archives** programmatically to identify number of records with multiple dog stages.
6) Viewed the dog names to identify erroneous names.
7) Investigated the structure of the contents in the source column of the **df_twitter_archives** data frame.
8) Check for any duplicated tweet IDs and using the dog name, confirm possibility of a duplication due to a retweet or tweet replies.
   From this assessment, I identified the below 9 quality issues and 4 tidiness issues:

Quality issues

i. **rating_denominator** has values that are not 10.
ii. **rating_numerator** has lower values (e.g. 0) and higher values (e.g. 1776, 960, 666 etc.) than expected.

iii. Erroneous datatypes (**in_reply_to_status_id, in_reply_to_user_id** and **timestamp** columns) in **df_twitter_archives** table.
iv. Some records show more than one dog stage (**doggo** and **pupper**: 12 records, **doggo** and **puppo**/**floofer**: 1 record).
v. Some dog names are erroneous (e.g. my, not, one, this, very, unacceptable).
vi. The source column has html tags and other information not really needed (e.g. '<a href="http://twitter.com/download/...'").
vii. Out of the 2,356 tweet id's from the **df_twitter_archives** table, only 2,075 have image predictions in **df_image_predictions** table.
viii. Inconsistency using lower/upper case on column **p1** in the **df_image_predictions** table.
ix. The presence of the retweets/replies in the **df_twitter_archives** dataset implies possible record duplicates in the data frame.

Tidiness issues

i. The column **tweet_id** in **df_twitter_archives** duplicated in **df_tweet_status** and **df_image_predictions** tables.
ii. **retweet_count** and **favorite_count** columns from **df_tweet_status** table should be part of the **df_twitter_archives** table.
iii. On **df_image_predictions**, only the columns **p1** and **p1_conf** are necessary since they present breed prediction with highest certainty.
iv. One variable in three different columns in **df_twitter_archives** table.

Data Cleaning

After identifying the quality and tidiness issues, I proceeded to clean the data sets under this section. First, I created a copy for each of the three data frames (**twitter_archives_clean, image_predictions_clean, and tweet_status_clean**). Then, for each issue data cleaning was programmatically performed in three steps of **define**, **code** and **test**. In the process of cleaning, the 4 columns from **twitter_archives_clean** having dog stages was combined into one column, named **stage** whose data type was converted to *categorical*. Also, records containing retweets and tweet replies were removed, as well as their related columns to eliminate any possibility of records duplication.

Storing Data

Having resolved each of the identified issue, I saved the gathered, assessed, and cleaned master dataset to a CSV file named "**twitter_archive_master.csv**".