Data Wrangling Report

The Data Wrangling process consists in four parts:

I. Gathering Data
II. Assessing Data
III. Cleaning Data
IV. Storing, Analyzing and Visualizing Data

I. Gathering Data

These are the three sources to get the required data for this project:

1. **twitter-archive-enhanced.csv**: file provided by Udacity. This file is downloaded manually and save it in the project
2. **image-predictions.tsv**: Download predictions data programmatically from Udacity's server and save data as flat file.
3. **tweet_json.txt**: Get extra data (retweet and favorite counts) for "WeRateDogs" tweets in the archive via Twitter API.

II. Assessing Data

You can find a list of detected issues (programmatically and via Excel) below:

1. Quality Issues:

- df_archive:
    1. There are retweets and replies.
    2. There are some ratings without images.
    3. There are columns that should be removed like 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls'.
    4. There names that are invalid (like a, such). All invalid names start with lowercase.
    5. The rating denominator seems to be inaccurate in some cases because there are rows with a rating denominator different from 10. Remove those entries

and also once all entries have the same rating_denominator, the column can be dropped as well.

6. The rating_numerator seems to be inaccurate too. There are valid cases with this column below 15, but rating_numerator greater than 15 are invalid. Delete rows with a rating_numerator higher than 15.
7. The timestamp column has incorrect format and type.
8. The source column contains htm.

- df_predictions:
  9. Predicted breed names should be normalized. This means that columns p1, p2 and p3 shouldn't contain '_' as separator and all should be capitalized, in order to analyze this information later in an easy way.
- df_extended:
  10. Remove unnecessary columns. We are interested only in 'id', 'favorite_count' and 'retweet_count' columns.

2. Tidiness Issues:

- df_archive:
  1. There are four columns: doggo, floofer, pupper and puppo that should be in a unique Categorical column.
- df_predictions:
  2. Predicted breed names should be in an unique column called 'dog_breed'. Once we have the dog_breed column, the unnecessary columns can be dropped.
- df_extended:
  3. Rename the column 'id' to 'tweet_id' to allow easy merging.
- All dataframes affected
  4. Merge the 3 dataframes

III. Cleaning Data

1. Quality Issues:

1. Filter out rows which have 'retweeted_status_id' or 'in_reply_to_user_id' columns with values.
2. Remove rows whose 'expanded_urls' column is null.

3. Remove those columns from df_archive_clean.
4.
    a. Step 1: Replace names which start with lowercase and which are equal to 'None' to null (np.nan).
    b. Step 2: Replace names which are equal to 'None' to null (np.nan).
5.
    a. Step 1: Remove those entries.
    b. Step 2: Once all entries have the same rating_denominator, the column can be dropped as well.
6.
    a. Step 1: Delete rows from df_archive_clean with a rating_numerator higher than 15.
    b. Step 2**:** Rename rating_numerator column to rating.
7. Fix both issues, removing +0000 and converting into datetime type.
8. Remove html code from 'source' column.
9. Replace '_' from 'p1', 'p2' and 'p3' columns. Capitalize breed names.
10. Drop not needed columns from df_extended_clean.

2. Tidiness Issues

1. Create a Categorical column called 'dog_stage' with the values of doggo, floofer, pupper and puppo.
2. Create a column called 'dog_breed' with the values of p1, p2 and p3 columns. Drop not needed columns.
3. Rename the column 'id' to 'tweet_id' in df_extended_clean.
4. Join 3 dataframes.

IV. Storing Data.

The result of cleaning data was stored as a CSV file format called ‘twitter_archive_master.csv’ using the pandas method to_csv.