Wrangling Report

In this project data from “We Rate Dogs” Twitter account was wrangled in order to produce some insights and make this data useful.

Data wrangling steps were as follows:

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

And here is a brief about the procedure done in every step:

1- Data Gathering:

In this step we had to collect the data from various sources, first for the main piece of data i.e. (The tweets) we were provided an enhanced version of the tweets that contains valuable information, this data was transmitted from the account owner to Udacity and then to us.

We needed extra data regarding the tweets, so we decided to use “TwitterAPI” to download this extra data, since we had tweets id’s so it was possible for us to query using the id the other pieces of data which are: retweet_count and favorite_count.
This step was done successfully and downloaded data was saved to a text file named “tweet-json.txt” afterwards it was saved as a ‘csv’ file.

Next we were provided a file containing image prediction for the tweets, it came as a “.tsv” file, we loaded it into a DataFrame.

2- Data Assessing:

Assessing was done both visually and programmatically in order to know the condition of the data from the “Quality” and “Tidiness” perspectives, the exact procedure is shown in the “wrangle-act.ipynb” file.
The found issues were as follows:
First: Quality issues:

1) `name` contains `Null` values,and stored as None.
2) `in_reply_to_status_id`, `in_reply_to_user_id`,`retweeted_status_id`, and `retweeted_status_user_id` conatains `Null` values.
3) Dog stages columns contains `Null` values,and stored as None.
4) `rating_denominator` is inconsistent it is not always equals to 10 (I know this rating system isn't seriuos but at least let's fix the denominator at 10).
5) `timestamp` is in wrong data type.
6) `source` is stored in `html` format, and wrong data type.
7) `in_reply_to_status_id`, `in_reply_to_user_id`,`retweeted_status_id`, and `retweeted_status_user_id` are of wrong datatype and this lead to losing some digits.
8) Table contains tweets that are retweets or replies.
9) `tweet_id` data type is better to be changed as `int` can lead to data loss.
10) There are tweets with more than one dog_stage.

11) Dog stages columns are of wrong data type.

Second: Tidiness issues:

1) Dog stages values in `df_main` are stored in several columns.
2) Dog stages column headers are values.
3) `expanded_urls` column in `df_main` table contains several values.
4) `expanded_urls` column values should be in another table because it represents another entity.
5) `in_reply_to_status_id`, `in_reply_to_user_id`,`retweeted_status_id`, and `retweeted_status_user_id` columns headers contains values.
6) `df_extend` should be merged with the `df_main` table as they represent one observational unit.

3- Data Cleaning:

Cleaning the data was a lengthy process, but I was done with care and precision in order to produce the best results.
We dealt with the 'Null' values first then studied the "Tidiness" issues, then other quality problem were dealt with.
For the complete details please refer to the "wrangle-act.ipynb" file.
At the end of the process, we were only left with high quality and Tidiness data that is ready for analyzing and visualizations.
Clean data were saved into a new file named "twitter_archive_master.csv".