WeRateDogs Project - Data Wrangling Report

Project Objectives

The main objectives of this project are listed as below:

- Gather data from three different sources, assess the data quality and identify quality and tidiness issues and finally solve those issues by performing cleaning actions.
- Store, analyse and visualize the final clean data.
- Prepare an act report based on the findings from the visualised data.

Gathering Data

In this step, we have gathered data from three different sources:

1. **WeRateDogs twitter archive data:** This data was already provided by Udacity, therefore the action here was to read the data via pandas read_csv function.
2. **Image predictions file:** This data was already hosted in Udacity's servers. Here the action is to download the file programmatically by using the requests library.
3. **Additional data points for tweets in archive:** In this third pillar here, we enhance twitter archive data by accessing Twitter API and query the data points programmatically (retweet and like counts)

Assesing and Cleaning Data

A list of issues have been identified during the data assessment process:

Quality

archive_data table

1. Some values in rating numerator columns seem to be too absurd.
2. There should be only 10 as values in the denominator values column, but we see other numbers.
3. Some rows in this table have non-null values in retweet columns, meaning that these are duplicated rows.
4. Missing values for columns : in_reply_to_status_id, in_reply_to_user_id,retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
5. Timestamp column is a string instead of a datetime
6. Source column contains HTML tag elements (i.e. href)
7. tweet_id is an integer

***image_predictions* table**

8. Inconsistent data: lowercase and uppercase names for p1, p2 and p3 columns

Tidiness

9. Dog types in the archive_data table are separate tables. It should be melted into one column.
10. image predictions table and tweet information are on separate tables.

Result

At the end of this wrangling process, three clean data sources were joint on tweet id key and stored in a csv file called “twitter_archive_master.csv”