Data Wrangle Report

In this project, I have gathered data from the tweet archive of Twitter user dog_rates, including the twitter archive, tweet image predictions, and each tweet's json data. After the data were ready, I assessed the quality and tidiness of the data, and necessary cleaning was processed to the data I have. The cleaned data were stored and analyzed.

Most quality issues have been found in the twitter_archive_enhanced. Below are the issues I observed:

1. Since eventually this data will be used to for prediction of dog breeds, in the data all the reply tweets and retweets by the user are not necessary, and we can just delete those data.
2. Another issue I find is the format of timestamp. Fortunately pandas has built-in function to convert string to date time, and this can be handled easily.
3. The next issue I observed is there are missing expanded urls by using the info function on dataset. Besides, expanded urls contain links which are not twitter links. So here I extract the short twitter link from the text.
4. In the text of tweets, the ratings are in the format of "number/number". However the original data have wrong ratings, so in this part I extract the ratings again.
5. Some dog names are missing or have the wrong name, like 'a', 'the', 'an'. I have checked dog names and change those wrong or missing names to 'unknown'.
6. One part I'm confused is the unclear columns of 'doggo, floofer, pupper, puppo'. I didn't see any proper use for those columns here. I changed each column to either 0 or 1 in this step.
7. Last one is the format of source. The source has html style like format and very hard to read. I extracted the real source name from the old and replaced it.

For tidiness issues, one I observed is from twitter_archive_enhanced and the other is from tweets. For the first one, I didn't see it useful to keep 'doggo, floofer, pupper, puppo' columns, so I just deleted those data. The second one is when I read the data from tweets json file, the user id has a wrong format. I cleaned the user id and it finally displayed as int number.

It is an very interesting project and also a good chance for me to practice all the skills learnt from this part.