Wrangle Report

First, I gathered the necessary data to the project: the twitter_archive dataframe was available in a csv file, so I just read it; The image predictions file was available in a Udacity url, so I requested it and stored it into a tsv file; The tweet counts dataframe was a little more difficult, since I had to create a developer account and access the twitter API to get that information. For that, I also had to deal with exceptions of tweets that had already been deleted and the twitter API rate limit of 15 minutes.

I performed a series of cleaning steps in the project's dataframes, some related to quality issues and other related to tidiness issues observed during the assess phase.

Quality Issues:

- I fixed some data types in the three dataframes (tweet_id from int or float to string, timestamp from string to datetime), consisting of three different issues at least;
- I removed undesirable characters present in the source column of the twitter_archive dataframe;
- I adjusted outliers that were causing distortion to the rating numerator and denominator columns;
- I removed retweets from all the dataframes (other three issues) by using the retweeted_status_id column of the twitter_archive dataframe;
- I removed tweets that did not refer to dogs.

Tidiness Issues:

- In the twitter archive dataframe, I transformed the dog stages columns in one column in which the rows contained the dog stages. I was careful to remove all duplicates and to prioritize records filled with a stage instead of 'None' by using an auxiliar method in which I attributed lower numbers to the stages and a higher number to the 'None's.
- I resumed the image prediction model columns to only two colums containing the best model in terms of confidence interval and the breed predicted by that model.
- I removed unnecessary columns, like the retweet columns in the twitter archive dataframe (once the retweets were remove, they had no more use), the other models information in the image prediction dataframe and auxiliar columns created only to perform some part of the cleaning.
- As a final tidiness issue, I merged the three dataframes using the tweet_id columns. After that, I removed the rows that had no images reported and stored the resulting dataframe in a csv file.