Wrangle Report

These three datasets had a lot of observable issues so I set about to remove around 10 quality issues and 4 tidiness issues that I thought would streamline the datasets for analysis. Since I knew I would have to eventually merge the datasets, I did it in the beginning so I don't have to refer to distinct datasets while writing my code. For this I first had to rename the column ‘id’ to ‘tweet_id’ in tweet_df to match the same column in img_df and consequently merge them. After merging the three I deleted retweets and tweets that had no images associated with them.

I made multiple copies of the data frame throughout this document to make sure I have a backup of my progress and in cause I want to compare to an earlier version. I then combine the different dog states into a singular column names ‘dog_stages’. I then noticed that we had duplicate tweet ids and decided to drop those. I converted in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id and retweeted_status_user_id to object data type so that if we perform a .describe on the data then we don't get additional columns that cloud our main data. I then convert the timestamp data to data time format. This was probably one of the most useful metric as now we can see trends over time.

I then needed to make the ratings given out into useful data so I first saw that there was a lot of inaccurate numerators beyond the range of what WeRateDogs usually gives. This was probably entered incorrectly so I added a filter and removed all entries with numerators below 10 or above 15. I then converted the numerator and denominator columns to float and created a new column, ‘rating_numeric’, that divided them to get a numeric value of the rating. I also converted them to string and did the same in a string column, ‘rating’, preserve the visual aesthetic on the WeRateDogs twitter page. I then capitalised all dog names in p1,p2,p3 to make them consistent and finally removed the columns that I didn't need for analysis. I finally saved this file and began to preform analysis on the cleaned WeRateDogs data.