Wrangle report

This is a wrangling report about the we rate dogs project. I defined 8 quality issues and 2 tidiness issues as follows:

Quality issues:

1. In the main data frame: Nondescriptive column: We will replace the links in the source column with the actual app they tweeted with in words since it's nicer this way.

The sources column contains an html tagged text that has the link of the app they used to send the tweet. What I did is replace those tags and the text with the name of the app name.

2. In the main data frame: Convert the timestamp column to a daytime object.

I converted the time stamp column to a daytime object because it allows me to manage the time more, even though I didn't actually mess with the time in this project.

3. In the main data frame: Change rating_numerator to 13 if it was more than 13

I'm not actually familiar with this account, but I read in the project page that it's what sets it unique is the fact that it gives ratings more than the denominator. But even though I found some very high ratings, impossible ratings lets say. So I thought I should max out the ratings to 13 if it was more than 13. And 13 is still > than the denominator.

4. In the main data frame: Change the rating_denominator to 10 if it was more than 10

Here I fixed the rating denominator to be 10 if it was more than 10.

5. In the main data frame: Remove all retweets

I removed all the retweets because we don’t need them.

6. In the main data frame: Remove all tweets that are replys

I removed all the replies the we rate dogs account sent because we just need tweets.

7. In the twitter_data data frame: Rename favorite_count to likes_count, because twittter has changed that awhile back.

8. In the images data frame: Nondescriptive column headers: We will change the column headers for p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog to names that are more descriptive.

We changed the names to prodiction1, prodiction1_confedence, prodiction1_dog and so on, up to number 3.

Tidiness issues:

1. Combine the 4 columns for dog stages into 1 column.

I did that because it reduces the columns in the dataset and it makes the information available in one column so it can make more sense.

2. Combine the twitter data data frame with the main data frame, that’s enhanced_df. Because they’re about the same thing really.