Memo Re: Data Wrangling
Date: 12 April 2019
To: Director of Data

As requested, I wrangled the data required for analysis of the WeRateDogs tweets. This work included the following datasets:

1. **Image predictions** (dog breed predictions based on images associated with tweets)
2. **Twitter archive data** (historical data from tweets such as date, tweet id, text etc.)
3. **Tweet json data** (streamed from twitter API using tweet ids from the above. Includes likes & retweets)

The following actions were performed to wrangle the data. They are divided between data quality issues, and data tidiness issues. Quality issues

Quality Issues

Images Table:

- This table contains predicted breeds that are general objects, not breeds (i.e. dish towel, coat hanger). There is a boolean value column next to the breed prediction which is populate True or False. In the instance that the prediction is not a real breed, the column contains False. Therefore I removed all rows with the Boolean column containing false so that are analysis would focus strictly on dog breeds, not objects.
- I also dropped the True/False columns following the above, as they are not relevant to my analysis.

Tweets Table:

- The timestamp data was not necessary for my analysis, so in order to simplifly I dropped the time portion, so that I was left only with the date.
- There were numerous types with the rating numerator as they were far greater than 10. Given the nature of the WeRateDog ratings (often 11/10, 12/10 etc), I kept all ratings up to a maximum of 15, after which I assume they were an input error. Ratings over 15 were dropped in order to avoid outliers scewing the dataset.
- Certain tweets contained denominators other than 10 for ratings. These were dropped to ensure we could compare ‘apples to apples’
- The tweets table has 181 tweets that contain a value for 'retweet_status_id'. I assumed that these are retweets that should be excluded from the dataset. After excluding these tweets I dropped any re-tweet-related columns as they were no longer necessary
- Typos were present in the dog names column, as they contained values such as 'a', 'the', 'an' etc. These values were removed and replaced with ‘none’
- Several tweets did not have a related image, meaning we didn’t have the breed prediction data for analysis. I removed this tweets by joining the tweets data onto

the image data, meaning any records without an match from the images dataset were dropped.

Tidiness Issues

- The information for analysis was contained in three separate tables (image info table, twitter archive, & tweets table). I merged the tree tables into one, left joining the twitter archive and tweets table onto the image table, so that all tweets had image data.
- The dog type variable (doggo, floofer, pupper, puppo) was stored in 4 different columns. Using the melt function I created one column containing this variable.