Wrangling Act

Data for this project was gathered from the following three sources:

- The Twitter archive of WeRateDogs containing basic information about each tweet until August 2017
- A programmatically downloaded image predictions file containing the top three dog breed predictions for each tweet containing an image
- A json file containing information about favorite-, retweet-, and follower count for each tweet id. This dataset was gathered using Twitters API tweepy.

For the final master DataFrame, only tweets which included an image and were neither retweets nor replies were considered.

The following quality and tidiness issues were addressed:

For the archive DataFrame:

- Retweets and replies were removed from the twitter archive.
- Values in the dog stages columns which contained the string 'None' instead of the proper value NaN were reassigned.
- The rating nominator and denominator were checked for erroneous values and all rows were which contained values contradictory to WeRateDogs rating guideline were removed from the DataFrame.
- The 'Name' column contained values other than nouns which is a result of the algorithm not always querying the desired part of the tweet text. Those values were set to NaN to get an accurate prediction of the most common names.
- The hyperlink at the end of each tweets text was removed.
- The columns 'Doggo', 'floofer', 'pupper' and 'puppo' were all grouped into one column named 'dog_stage'. Tweets mentioning several dog stages at once were named 'multiple'.
- Erroneous datatypes (for columns: timestamp, tweet_id, dog_stage, rating_denominator and rating_numerator) were corrected.
- Columns of no use for this project were dropped.

For the predictions DataFrame:

- The DataFrame was reshaped to only contain one column for each category and not multiple columns containing values.
- Erroneous datatypes were corrected.

For the tweepy DataFrame:

- The DataFrame was joined with the archive DataFrame as DataFrames containing similar information should be combined into one table.

The result were 2 clean master DataFrames:

- The master archive DataFrame which contains 1656 rows and 12 columns
- The predictions master DataFrame which contains 4611 rows (3 rows for each tweet id due to 3 predictions for each image) and 7 columns.