Wrangle Report

For the wrangling process I worked with three different datasets. First I investigated the datasets to identify any missing data, incorrect data, or redudant data. I utilized the pandas and missingo libraries to assist with these efforts.

In the first dataset, which contained the twitter archive data, I identified the following quality issues

- Timestamp was of object datatype and needed to be a datetime datatype.

  The column was changed to a datetime datatype

- The source column contained a long and redundant URL which needed to be shortened

  The source column was cleaned so that the urls were replaced with shortened and easily identifiable values

- There are 59 missing entries in the expanded_url column.

  The missing entries were dropped from the dataframe

- There are entries that are retweets and need to be removed

  The retweets were dropped from the dataframe

- The name column contains entries that are not names, like "a" and "by", all lower case names entries appear to be non-names

  The names that were lower case were identified and replaced with None values

In the second dataset which contained the prediction data for the images I found the following quality issues:

- The prediction dataframe contains entries that are not dogs.

  No cleaning was performed for this issue, just the quality issue noted

- The prediction dataframe columns need descriptive column names.

  The column names were renamed to better identifiable names

- Some of the predictions that are "false" for a dog are indeed pictures of dogs.

  No cleaning was performed for this issue, just the quality issue noted

- The img_num column is not needed and can be removed.

  The column was removed

In the third dataset tweet_json which contains the json data I found the following quality issues:

- The tweet_json data has entries with 0's in the retweet_count and favorite_count, that may be missing data which needs to be removed.

  The 0's were identified and those entries removed

For tidiness, there were a few issues noted that needed to be completed:

- The twitter archive dataset had redundant columns that needed to be combined.

  These columns were combined into one

- All three datasets needed to be combined into one master dataset

  All dataset were merged into one using the tweet_id column

An additional issue noted was the difference in records between all three datasets. They did not each have the same amount of records, which needed to be fixed so accurate visualizations could be created. *The combined master dataset was cleaned to remove missing values so all columns had equal records*

In [ ]: