Wrangle Report by Tyler Sanders

Assessment of the data

Data Source	Details
twitter_archive_enhanced	The following columns all contained missing values: 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeed_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'
twitter_archive_enhanced	The 'text' contains the body of the tweet, but then ends each tweet with a hyperlink. The hyperlink is already represented in the 'expanded_urls' column, so having it shown twice is redundant information and is not really part of the text of the tweet.
twitter_archive_enhanced	The columns 'doggo', 'floofer', 'pupper', and 'puppo' have the string "None" in many of their rows. These columns seem to be categorical variables and would benefit by melting them into a single categorical column.
twitter_archive_enhanced	After accessing the 'rating_denomiator' field, I found that 2333 of the 2356 observations have a denominator of ten. - 3 tweets have a denominator less than ten. - 20 tweets have denominators greater than ten.
twitter_archive_enhanced	For the denominators that had values less than ten - Row 313 has a rating of 960/00 but was then given the valid rating of 13/10 - Row 516 was one of the few times where WeRateDogs did not give a rating. - The tweet in row 2335 misrepresented the rating in error when a fraction of 1/2 was included in the text.
twitter_archive_enhanced	The rows with denominator that are greater than 10 are valid, but to be able to compare them with the rest, I will need to standardize them to the value of 10.
twitter_archive_enhanced	Additionally - Tweets with decimal points show incorrect ratings than what they were given.
image_prediction	'p1', 'p2', and 'p3' are all in snake case format and some of the categorical values are capitalized while others are not.
tweet_json	In the **tweet_json** file 'retweet count' has an inconsistent format by not having an underscore to replace the space between words. It is a duplicate column to 'reteet_count' and has all NaN values.
tweet_json	**tweet_json** contains some duplicated rows that will need to be dropped

Cleaning the data

Data Source	Details
twitter_archive_enhanced, tweet_json	I dropped the columns that contained all NaN entries.
tweet_json	Tweet_json had 140 rows that were duplicated, after they were dropped 2026 rows remained.
twitter_archive_enhanced	Some rows were showing incorrect ratings and to be updated, such as row 342 and 516 which had no rating given. For row 342 I dropped the entire row, since there was no dog or photo, but for row 516 I replaced the rating with the median value for all ratings.
twitter_archive_enhanced	I Normalized all the fractional ratings to have a denominator of 10 to a standardize the data to facilitate comparisons
twitter_archive_enhanced	After 'rating_denominator' values were set to 10, I dropped the column and renamed 'rating_numerator' to the general 'rating'
twitter_archive_enhanced	I corrected Ratings containing decimal places, Using Regular Expression to loop through each text and pull out any digits with decimal places.
twitter_archive_enhanced	After all the decimal places were accounted for, I round to the nearest integer, to keep the ratings on a discrete scale.
image_prediction	Columns **p1**, **p2**, and **p3** are all in snake case format and some of the categorical values are capitalized while others are not. To clean this up I removed the underscores of each row and put the classifications in title format.
twitter_archive_enhanced	The 'timestamp' column is formatted down to the millisecond. This is accurate and thorough information, but for future timeseries analysis, I will convert the time in a date format of month, day year
twitter_archive_enhanced, tweet_json	To finish I did some small fixes to the data, setting the 'rating_numerator' and 'retweet_count' to integer datatypes.
twitter_archive_enhanced, image_prediction, tweet_json	I Merged **twitter_archive_enhanced**, **tweet_json**, and **image_prediction** Dataframes into one Dataframe **twitter_archive_master.csv**