Wrangling Report

To successfully achieve the tasks mentioned in the Rubric, after importing the libraries I will use, I first loaded twitter-archive-enhanced.csv into my notebook which I ran locally with Visual Studio. Then I got the image-predicitions.tsv with the Requests library, and finally with the use of the tweet ID from image-predictions.tsv I gathered the posts and tweet count info by using the Twitter API called Tweepy and saved it in a separate data frame.
After assessing each data frame visually (for typos, looking for ways to combine columns that it makes sense, etc) and programmatically (checking for missing values, duplicated ones, datatypes in each column, I concluded these quality and tidiness issues:

Quality Issues

Archived data frame
- `timestamp` column is in string format, it should be datetime
- `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id` and `retweeted_status_user_id` are float type, they should be string
- `tweet_id` is an integer format, should be string
- `source` column's values include html code which is not needed, only need url
- `rate_numerator` and `rate_denominator` should be in one column and without decimals
- Some dog names start with lowercase and are not dog names. These need to be removed.
- Remove retweets(as per Key Points)

Image predictions data frame
- p1, p2, p3 names in columns are not consistent in upper or lower case (should all be lowercase)
- `tweet_id` is an integer format, should be string

Tweets data frame
- `tweet_id` is an integer format, should be string
- Retweet and Favorite: `retweet_count` and `favorite_count` should be integers, not floats. There is no such thing as half of a retweet or favorite.
- Remove retweets(as per Key Points)

Tidiness issues

- doggo, floofer, puppo, pupper should not be separate columns but one column and it should state what category the dog is in
- need to combine datasets to a master dataset for analysing: `df_archived`, `df_tweets`, `df_image_predictions`
- removing all non relevant columns from the master dataset

I went through and successfully cleaned all of these in my notebook and added notes to each line of code to make it super clean.