Wrangling Report

Below I’m listing tidiness and quality issues, that were found in the dataset, and the wrangling actions taken to fix each issue.

Tidiness Issues and Actions

1. All three dataframes belong to a single observational unit and they should be merged into a single dataframe., **Action:** merged into a single dataframe “twitter_archive_master”.

2. The 4 different columns doggo, floofer, pupper and puppo, are all representation of the same variable that identifies the stage of dog. Therefore they should be merged in a single column. **Action:** merged into a single column “dog_stage”

Quality Issues and Actions

1. tweet id columns have wrong dtype. They should be str. **Action:** fixed
2. Two unrelated columns: (Unamed: 17 and Unamed:18). **Action:** dropped.
3. Unrelative columns that start with "retweet_" since the dataset doesn't include retweets. **Action:** dropped.
4. non-informative column such as rating_denominator. **Action:** dropped.
5. wrong entries in dog_stage column. **Action:** converted to NaN.
6. in_reply_to_status_id and in_reply_to_user_id coulmns full of missing values. **Action:** dropped.
7. replace tweet_id column values (wrong enteries) in twitter_archive_master_df with values from id column in tweet_df . **Action:** fixed.
8. rating_numerator column has several quality issues: There are values that are not correctly extracted, out of range values and wrong data type.

   Action: correct values from the tweets are extracted correctly and replaced the old ones. Rows with out of range values are dropped. “str” data type is converted to “float”.