Wrangling process:

After taking a closer look at the data from all three sources, here are the issues that I found:

Quality Issues:

1. I noticed that timestamps columns from the twitter archive are of type string when they should be of type Date, I fixed them by converting these columns into Date columns.
2. When I viewed the data I noticed strange dog names which are most likely to be extracted incorrectly, occurrences such as names with only the letter “a” or “O”, I replaced these rows with null which is in my opinion the most appropriate fix.
3. Another issues with names that there are names with the value “None” when they should be null, I fixed that by converting these values to null.
4. Also, the dog stages columns have the same pattern of having the “None” value opposed to null, fixed it by converting these rows to nulls.
5. As has been noted on the project retweets are not considered an original rating of the dogs, so I removed the retweeted rows.
6. I noticed some inconsistency in the ranking rating_denominator values, most of the data have the value 10 but there are some occurrences other than that, sometime even the value zero which will cause calculations errors later if not fixed. I fixed this by keeping rows with the 10 value only.
7. There are some unnecessary columns in the archieve dataframe such as source and expanded_url we can view tweets using the tweet_id so these columns are not really beneficial to our data set, I also dropped other columns for same reasons.
8. I couldn’t really decide right away with this issue, the dogs stage columns can be great for analysis, but the issue is that most of our data have these values as null so dropping null values can make our data set very small, so I decided to just drop these. If I decided to keep them I should melt them In one columns since all four columns represent the same dog stage variable.
9. Lastly, I wanted to keep dogs that we know their names, so I dropped rows with null names

Tidiness issues:

1. Having two columns for the rating is not really tidy, So I made a new column with the results calculated from as rating_numerator / rating_denominator and after that I dropped these two columns.
2. Currently our data is spread out across many dataframes so I must join them together, I started by merging the “archieve_clean” and “predictions_clean” and merged the result of these two with the “tweets_clean” dataframe

After all this wrangling I stored the result as csv file.