### Quality issues

1. Change the following to datatypes:
   1. df\_archive\_clean:
      * tweet\_id: string, timestamp: datetime, source: category, retweeted\_status\_id: string, retweeted\_status\_user\_id: string, retweeted\_status\_timestamp: datetime
   2. df\_predict\_clean:
      * tweet\_id: string
   3. df\_tweet\_json
      * tweet\_id: string
2. retweeted\_status\_id's could be self-retweets. This means that any of these 181 records could represent duplicate data. It appears we may be able to identify 'self-retweets' by comparing retweeted\_status\_user\_id
3. The rating scale appears to be inconsistent.
4. The rating scale appears as though it will likely have outliers that will need to be removed for analysis/visualization. Merge into one column as float. Some division by 0 in the denominators
5. There are tweets in \*\*twitter-archive-enhanced\*\* that do not have corresponding records in \*\*image-predictions\*\* and vice versa.
6. \*\*source\*\* column contains interesting data about devices but is not readable.

#### image-predictions

1. There appear to be 66 duplicate images/image predictions. Since image\_predictions tweet IDs are a subset of twitter-archive-enhanced tweet IDs, the 66 duplicates in image-predictions will corrospond to tweets in the later set.
2. There are 324 records where no dogs were detected whatsoever.
3. There are 3 separate predictions for each ‘dog’ photo.

Take the most likely of p1, p2, p3, move it to a new column predicted\_breed and drop the others.

### Tidiness issues

#### df\_archive\_en / twitter-archive-enhanced

1. Merge df\_image\_predict, df\_tweet\_json, and df\_archive\_en into one master clean dataframe

2. doggo, floofer, pupper, puppo columns are categorical values of the same variable.

#### df\_image\_predict / image-predictions