We rate dog’s data quality issues

Archive

1. Re-tweets are included in the dataset these are identified by “RT @” Remove them.
2. There are 3 separate tables. – We fix this by merging our data into one table using “twitter\_id”
3. Replies are included in the dataset – Identified in reply\_to\_status\_id remove all.
4. Dog classification names are “NONE” should be nan
5. Some names are missing in the name column - This is not worth fixing noticing that we may build inaccurate data. Looking at the text field we can see names involved however there is too high a risk of getting inaccurate data if we were to mine it.
6. Missing fields for dog type – I thought to mine the data from text field but the risk of mining inaccurate data is too high.
7. Only tweets with images – solved by querying data where
8. Redundant columns (doggo, puppo etc) – Place values into a new one dimensional column and drop unneeded columns.

Predictions

1. Not very human readable columns – Change and make understandable

Tidiness

1. Drop unnecessary tables
2. Fix column headings to make everything more readable to a human
3. Lower case everything.