Wrangle report

one archive table and imagePrediction in the en

Archive table:

Quality:

1. timestamp variable's type should be time format rather than string.
2. Expanded_urls,in_reply_to_status_id,in_reply_to_user_id only has 78 non-null value.
3. retweeted_status_id,retweeted_status_user_id and retweeted_status_timestamp have 2175 missing value → delete the rows populated with values under these columns and remove these columns
4. Tweet_ID should be string rather than numeric value
5. Expanded_urls have 59 url with missing value → fill these missing expanded_urls
6. in_reply_to_status_id,in_reply_to_user_idretweeted_status_id,retweeted_status_user_id and retweeted_status_timestamp have a lot of missing value --> examine more in programmatical portion.
7. Timestamp variable has the suffix '+0000" → change it to time type. Instead of string
8. Source column's value has "<a href=" and ".r" around text. some rows have the same source url.
9. Some name variable has 'a', 'an' and 'the' → replace ‘a’, ‘an’, ‘the’, ‘Non’ with nan value.
10. The value inside 'text' is incomplete.
11. tweet_id=883482846933004288, the text states the rating should be 13.5/10, but the rating numerator is 5 → Etracting data from text.

Tidiness

1. the doggo, floofer, pupper, and puppo column should be merged into one column
2. merge tweet_json table with Twitter_Archive Table by tweet_id

imagePrediction Table

Quality:

12. P1 , P2, P3 columns → Some values have 1st letter capitalized, and others are not, → Convert them all into all lower case character.
13. p1_conf, p2_conf and p3_conf display as 6 digit decimal value. → change to 4 digit
14. ‘newfoundland’,‘laptop’,‘sea_lion’,‘lhasa’,‘mitten’,‘feather_boa’ inp2andp1 are not a breed of dog.

Tidiness

1. The column names in imagePrediction are very random → rename with more meaningful names.