List of Data Issues

Analysis after processing the dataframes¶

For df_twitter_archive

- Wrong data existed in the following columns:
- 1.in_reply_to_status_id
- 2.in_reply_to_user_id
- 3.retweeted_status_id
- 4.retweeted_status_user_id
- 5.timestamp
- 6.retweeted_status_timestamp
- Missing data in the following: incorrect rating_numerator and rating_denominator for row numbers : 1069, 1166, 2336
- Few retweets named as tweets

For df_image_predictions¶

- Missing Data (2075 records instead of 2356)

For new_df¶

- Missing Data (2347 records instead of 2356)
- All dataframes exist as individual dataframes but they contain data for same tweet ids so it should be a single dataframe.

Stepwise List of Wrangling Issues Handling (Commented in Code)

1. Names to be cleaned
2. Timestamps to be fixed
3. URL to be expanded
4. Rectifying incorrect ratings
5. Formatting numerator, denominator, text
6. Removing Redundant Entry
7. Rectifying Missing Values - Pupper
8. Rectifying Retweets
9. Getting Final Rating in Decimal Form
10. Rectifying Text
11. Checking for accuracy
12. Handling missing data
13. Merging to create final data
14. Rectifying dog_stage
15. Dropping Redundant Columns
16. Dropping retweets Column
17. Readying final DF
18. Storing Cleaned Data
19. Creating VIsualizations