The analysis compare of three dataset

1.1- Twitter_archive_enhanced .csv read as df_tw

• *Twitter_id should be a string value not a int and don't duplicated as this column is a unique value*

• *Time stamp have incorrect type string and should be*

• *Doggo , floofer , puupper , puppo should be in column which represent the dog age stage (tidy issue)*

• *url_expanded have missing value .*

• *suggessting to substitue rating_numerator, rating_denominator with column called rating*

1.2- df_image_pred

• *Tweet_id also should be a string and unique value*

• *i suggest that P1_conf , p2_conf , p3_conf substituted by p max which represent highest value of p and take theequivalent value from P column (*

• *P1_dog , p2_dog , p3_dog should be one column and if all hav and pe false value rows , this rows should be deleted (tidy and quality)*

• *Check jpg_ul for duplication and remove this duplication in cleaning*

1.3- Twitter_df

• *Some empty column should be removed like (contributors , coordinates)*

• *Column (in_reply_to_screen_name, n_reply_to_status_id, in_reply_to_status_id_str , in_reply_to_user_id,in_reply_to_user_id_str)*

$assessingfristdataframe(archivedataframe=df_tw$

2- cleaning Summary

2.1 df_tw_clean

- Removing the last 6 digit of time stamp and changing format of date and time.
- remove expanded url which empty
- remove un-necessary column

2.2- df_image_pred_clean

- Remove duplicated url
- Substitute p_conf with p_max and choosing cross ponding p1
- Also substitute p_dog with one column

2.3-twitter_id_clean

- Remove column which don't make sense to analysis (geo, entities, coordinates, contribution, place)

After assessing and cleaning 3 data set, merging datasets to one, making iterative assessing and cleaning again

Assessing summary 2

- Merging 3 data set with tweet_id

Cleaning summery 2

- Channing tweet_id to string datatype is a best practice for id s.
- Remove row which didn't have jpg_ulr as this tweet without picture will not make any sense
- Changing the ratin_numerator and rating_denominator to float as
- substituting p_conf s column with maximum values column and give the name cross ponding to this values
- Substituting rating denominator and nominator with rating values which is the denominator divide by nominator