Wrangle Report

Gathering Data:

In this project, the data is gathered from three different sources. The first one is the Weratedogs twitter archive by manual download. The tweet image predictions are downloaded programmatically using requests library through an url. The last one is the additional data like favorite count,retweet count.This data is obtained using tweepy library.

Assessing Data:

Quality Issues:

1.Few "name" column values contains names which are not correct names(a,an,the).

2."name" column contains values with none.

3.Html tags in the source column.

4."Timestamp" column should be changed to datetime from object.

5.We need only orginal tweets. There are values in retweets and missing values in expanded_urls column.

6.We can clearly observe that,few values are wrongly extracted for the rating_denominator column as there are other integer values in the text.In few instances,there are multiple scores because of multiple dogs in an image, so values of the denominators are summed up.

7.We can observe that,few values are wrongly extracted for the rating_numerator column as there are other integer values in the text. In few instances, there are multiple dogs in an image, so values of the numerators are summed up.

8.Some values in the predictions columns(ex:P1) are starting with uppercase letter and some with lowercase.

9.There are duplicate images in the predictions dataframe, which means there are retweets here also.While merging with df_twitter_archive these rows will be dropped.

10.There are images, which are not dogs.

Tidiness Issues:

1.There are 4 columns for dog stage category("Doggo","floofer","pupper" and "puppo"). 4 columns can be reduced to 1 column.

2.Predictions should be a part of df_twitter_archive table.

3.There are 3 columns for predictions. It can be reduced to one column

4.There are 3 columns for confidence. It can also be reduced to one column.

5.Should merge this table with df_twitter_archive,because they are from same observation unit.

Cleaning Data:

Before cleaning, I have created copies of all the three data frames. I have removed retweets and replies, as we want only original tweets. Removed rows, which have images, which are not dogs. Removed p1_dog,p2_dog,p3_dog columns as all are dogs .Dropped p2,p2_conf,p3,p3_conf as there are no false predictions. Renaming p1 and p1_conf to prediction and confidence. Creating a new column 'Dog_stage' for dog category instead of 4 columns. Merge df_predictions_clean with df_twitter_archive_clean Merge df_api_clean with final_df .The name, source,timestamp column issues are also fixed. Finally, the cleaned data is stored in a csv file.