Gather - – The image prediction spreadsheet was downloaded from the server and imported into dataframes.

– The twitter archive was imported into the dataframe.

– The twitter api data was collected by querying the twitter api with the required twitter tweet ids.

Assess -

Quality :-

–The image prediction data conatined a lot of predictions which were not relevant to dogs. The p1 column does not contain all dog names.

– The p2 column similarly does not conatin all dog names.

– The p3 column also conatins many non dog predictions.

– In the twitter archive table many rows were not having the names. The values were None.

– In the twitter archive table some rows were having names as 'a' or 'an'.

– The twitter archive file had columns like in_reply_to_status_id, in_reply_to_user_id which were Nan.

– The twitter archive file has many values NaN for the columns retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.

-- The twitter api data had their tweet status as false even though there tweet count was more than 0.

Tidyness :-

– The extended entities column of the twitter api table has multiple variables in it. The id and id_str. They need to be split up or the column needs to be removed as the dataframe already has id as column.

– The twitter archive and the twitter api spreadsheet should be merged based on the tweet id as they both describe the tweet details.

Clean

- Taking copy of the original datat-sets we start fixing the issues.

– Filter the image prediction datatset on the true value of p1_dog, p2_dog and p3_dog. Kepp them if they are true.

– Remove the rows from the twitter archive file which have names as None or a or an.

– Make the retweeted value as true for the rows which have a retweet count more than 0.

– Fill the NaN values of the twitter archive file with None.

–Dropping the column extended entities column of the twitter api table, as there is already an id column, thus did not split the values.

– Merged the twitter archive datatframe and the twitter api dataframe on the basis of twitter tweet id.