Reporting: wrangle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

- The columns such 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id','retweeted_status_user_id', 'retweeted_status_timestamp' had large proprortion of missing values (these features had over 92% missing values) were dropped,in addition some of these columns are redundant.

- Also The column 'expanded_urls' has 2.50% missing values in twitter-archive-enhanced data however this feature had some limitations on handling missing values since it was difficult to decide on imputing urls and dropping may also lead to data loss.

- The columns such the name, 'doggo', 'floofer', 'pupper', 'puppo' in twitter-archive-enhanced data set have object name 'None' (this is could be missing value which have been replace with a value None) thus they will not appear as missing but as object. These values were replaced with np.nan so that they can appear as null values in python

- Some features such as text column had mixed lower and upper case strings, while other rows have upper case strings only and others lower cases trings only within a given column in the twitter-archive-enhanced data set. All the strings were converted lower case to ensure consistency and uniformity in the data.

- The data type for Timestamp column in twitter-archive-enhanced dataset appears as object instead of datetime dtype. The datetime dtype has been parsed on timestamp to enhanced data quality.

- The text column contain multiple variables such as html, url links within each single row in twitter-archive-enhanced data set . All the ending url links were removed the html ampersand code was replaced with & and the newlinesymbols can be removed using replace and regex functions in text column of cleaned twitter-archive-enhanced data.

- The text column contain some white spaces in twitter-archive-enhanced data set. The leading and trailing white spaces from a string was removed using strip() function.

- The dypes for tweet_id is integer and should be object in twitter-archive-enhanced and image prediction dataset. The tweet_id column was converted to strings in cleaned twitter-archive-enhanced and cleaned image prediction dataset.

- The name column contain some uncommon values (dog names) such as a , an,very, this , quite, unacceptable which could be misplaced strings which in twitter-archive-enhanced dataset. The ideal name of dogs were checked by looking at the text. The dogs with such uncommon (weird) names were replaced with assigned value None.

- The source column is a bit dirty with HTML format with a and \a tags surrounding the text (The column looks redundant). The source column is a bit messy and may not necessary for our annalysis and so it was dropped using drop() function

- The retweeted_status_timestamp column in twitter_archive dataset depicts that the there 181 retweets which may not be neccessary for analysing dogs images. This column is not necessary and should be dropped using drop() function, however had been taken care of since it had been dropped together with other features with missing values.

- Since 'doggo', 'floofer', 'pupper', 'puppo' columns are stages of dogs they should in a same column name (one variable) and not separate columns. We Melt the 'doggo', 'floofer', 'pupper', 'puppo' columns to a *dog_stage* and *dn_stages* columns. Then we drop the Immediate *dn_stages* column.
- Since the image_predictions dataset has common column (tweet_id) with twitter_archive dataset thus the two tables need to be merged also the twitter json data has tweet_id. Generally the table should be maerged into one. We merged darchive_clean and df_image_clean tables using merge() function from pandas then assign it as merge_twitter. Then and merge again with jtweet_clean into one table using the merge function and assign it as tweet_master using tweet_id which is common identifier.
- Checking and Handling the duplicates of merged dataset. 5836 duplicates were removed

- Droping columns with one unique values since they are not useful for analysis. The friends column was droppped since it had only one unique value and may not have beeen relevant for our analysis

- Gathered, assessed, and cleaned master dataset was saved in CSV file named "twitter_archive_master.csv".

In [ ]:

In [ ]: