Wrangle Report

In this project, I had to gather the data using twitter API then assess the data and clean the data to get relevant insights out of the data. Finally, after cleaning the data I was able to draw out various meaningful insights out of the data.

Data:

There were altogether 3 dataset:

- **df_tweet**- Dataframe with retweet and favorite count
- **df_image_predictions-** Dataframe with image prediction data
- **df_archive** - Dataframe with archived tweets data

Assessing and cleaning all those data:

The data we gathered was not perfect and for the analysis I had to assess the data and clean the data.

The issues that I found in these data are:

df_archive

Issues:

- There were names such as "a","an","the" which is very unlikely to be a dog's name:

  To deal with this I filtered out the data with the "a", "an" and "the" then replaced it with the name if there was one in the tweet or else replaced it with "none"

- Datatype correction was needed (timestamp is shown as object):

  I converted the datatype of timestamp which was "object" to "timestamp" which would help when analyzing based on time.

- Some entries were retweets (we need to avoid those):

  I filtered out the data that were retweets and dropped them.

- Some Missing values:

  Some missing values and those that are not required are dropped.

df_image_predictions

Issues:

- There are some missing data as df_archive has in total 2356 rows but df_image_predictions only has 2075 rows:

  I joined the dataset and only kept the one which was present in df_image_predictions.

- There also seems to be an inconsistency with the casing for p1,p1,p3 (some are titlecase, some lowercase, etc):

  I changed the case for all data to lowercase to maintain consistency.

Tidiness:

- Four columns (doggo, floofer, pupper, and puppo) could be made into one:

  I combined all the four columns into one column named dog_stage. If there were more than one entry, I joined it using hyphen.

df_tweet

Issues:

- There are some missing data as df_archive have total 2356 rows but df_tweet only has 2331 rows:

  I joined the dataset and only kept the one which was present in df_tweet.

- Changing the column id to tweet_id.

Merged Data:

Tidiness:

- In the merged data the column related to retweets are dropped as we are not considering retweets.