Gathering the data:

1. Reading the CSV file Twitter Archive
2. Reading the TSV file Image Predictions
3. Reading using twitter API

Assessing:

Quality issues:

twitter archive dataframe:

1) The colums (in_reply_to_status_id , in_reply_to_user_id, retweeted_status_id , retweeted_status_user_id, retweeted_status_timestamp) have a lot of missing values/

2) All IDs data type should be String not integer.

3) In column name, some hase None value

4) Time stamp column hase (+0000) in its values, that made it lengthy and messy

5) In column name, some names are not accurate, like "a" , "an" and “the”.

image prediction dataframe:

6) column names are not informative and descriptive.

7) tweet_id should be string data type

tweet_status dataframe:

8) Id columns are integers not strings

Tidiness issues:

1) The nature of the data source provided us with 3 data frames, while we might merge them to get one clean table.

2) The columns (doggo , floofer , pupper , puppo) could be as values of one column named type.

Cleaning the data:

1. Make copy of the dataframes.

2. Delete the retweets and keep the original tweets. The column in_reply_to_status_id will be used.

3. Deleting columns with missing values

4. Replacing NaN with word none

5. Add a new column called Dog_Type to define whether it is "doggo,floofer , pu pper ,puppo". if it is not defined it will be Nan. Then will drop the 4 columns.

6. Renaming columns in image predictions data frame to have more informative headings.

7. Changing the data type of ID columns to be String instead of int.

8. Column name has the words ‘a’, “an” and “the”, those rows will be dropped

9. The timestamp column, will separate the date from the time each in a separate column.

10. The source column, has the source of the tweet along with a link, this will be cl eaned to contain only the source.

11. Dropping the columns : 'retweeted_status_user_id', 'source_y'.