Wrangle report

Gather Data

- The archived and predictions file was already provided
- 3rd file was extracted by tweet_id in json format and stored in a file named tweet_id and converted it to a dataframe.

Assess Data

- The following issues were detected:

1. Null values in the columns(in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp,expanded_urls)
2. incorrect names in names column('a','quite','the','an')
3. Erroranous datatypes(tweet_id,timestamp)
4. erroneous datatypes(tweet_id)

Image_predictions table

1. erroneous datatypes(tweet_id)

Json_data table

- Null values in columns(contributors,coordinates,in_reply_to_screen_name,in_reply_to_status_id,extended_entities,retweeted_status,quoted_status_permalink,quoted_status_id_str,quoted_status_id,quoted_status,possibly_sensitive,possibly_sensitive_appealable)
- erroneous datatypes(created_at,id,id_str,possibly_sensitive,possibly_sensitive_appealable,quoted_status_id)
- Invalid letters in lang column
- Massive change in the values of favorite_count column.

- Nulls in (contributers,coordinates,geo,in_reply_to_screen_name,in_reply_to_status_id_str,place,)

Tidiness

- Dictionaries given in coulmns(entities,extended_entities, user,quoted_status_permalink,retweeted_status).
- retweets in archieved and df table not required.
- There are 4 seperate columns of dog types instead of 1.

Clean Data

- Create copies of data frame of 3 datasets
- Run a loop to replace the null values with None in the archieves table
- Converted the the data types of tweet_id and timestamp column
- Replaced the single letters in the names column to None
- Converted the datatype of tweet_id column from integer to object
- Replaced the null values with None in the columns
- Converted the datatypes of Created_id to datetime, id and id_str to object
- Replaced the None values to empty strings and added the 4 columns of doggo,pupper,puppo,floofer into a single column called dog_category. Later I dropped the 4 columns.
- Dropped the following columns in archieves and predictions table. [retweeted_status_user_id','in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_timestamp'], Predictions-table- ['p1','p1_conf','p1_dog','p2','p2_conf','p2_dog','p3','p3_conf','p3_dog']