Reporting: wragle_report

wrangle report by Olugbenga Felix Ajiga

I began by collecting data from different sources like flat file containing WeRateDogs archived tweets. Also using twitter API to gather tweets from twitter real time but I got some challeges as regards this so I wont use the API. At the end these three processes were performed WeRateDogs

1. Data gathering;
2. Data assessment;
3. Data cleaning.

Quality issues:

```
IN predictions_df: the prediction number needs to have data type integer
One row has total confidence greater than 1 which is not normal
The values in p1, p2, and p3 should be properly named
Change 'None' to empty cell in doggo,floofer, pupper, puppo then delete
There are duplicated jpg_url

IN twitter_archive_df: some tweets in archive have missing data for retweet_count or favorite_count
alot of 0 in the favorite_count
IN twitter_archive_df: the dog_stage columns (doggo, floofer, pupper, and puppo) should have a categorical data type
using data type float for rating so that the column can be uniform
All dogs should have a name with proper letter casing
dropping timestamp column not needed
```

Tidiness issues

```
IN predictions_df table
For predictions_df: columns [ prediction 1, prediction 2, and prediction 3] are untidy

twitter_archive_df: stage columns [doggo, floofer, pupper, and puppo] are untidy
jpg_url variable should be in twitter_archive_df table to satisfy tidiness definition
twitter_archive_df and predictions_df tables form two different obseravations units and will be kept seperately

IN twitter_archive_df table:
doggo, floofer, pupper and puppo will be merged into one column
tables twitter_archive_df and tweet_in_json_df will be under one observational unit
There should have been a dataframe for each of these observational units [tweet data, dog data, and image predictions]
```

Assessment process

Tidiness issues are predictions_df: columns [prediction 1, prediction 2, and prediction 3] are untidy twitter_archive_df: stage columns [doggo, floofer, pupper, and puppo] are untidy There should have been a data frame for each of these observational units [tweet data, dog data, and image predictions]

Quality Assessment: predictions_df: the prediction number was changed to integer data type The values in p1, p2, and p3 were properly named 'None'in the names were changed to an empty cell in doggo,floofer, pupper, puppo then delete There are duplicated jpg_url used data float data type for rating so that the column can be uniform All dogs names were properly named with the right alphabet casing

The cleaning process

Cleaning was performed to take care of the following: removing retweets and duplicate values in jpg_url column. removing duplicate values in jpg_url column Removing pictures that are not of dogs when at least one value of p1_dog, p2_dog, and p3_dog is true with confidence greater than 1. Remove any invalid names and text related to lowercase names that do not have the phrase "We only rate dogs" timestamp column was deleted

IN predictions_df: the prediction number was changed integer One row has total confidence greater than 1 which is not normal The values in p1, p2, and p3 was properly named Change 'None' to empty cell in doggo,floofer, pupper, puppo then delete

IN twitter_archive_df: some tweets in archive have missing data for retweet_count or favorite_count alot of 0 in the favorite_count so retweet_count and favorite_count was removed from the datafame.

rating number column was converted to float so that the column can be uniform All dog names were converted to proper letter casing dropping timestamp column because it was not needed

My take home on the project

By the time I was done with the wrangling process, I can say I now have the confident to perform the same process in real life. Thank you Udacity and mastschool for your guidance and support. I must confess this project is really very challenging. I spent hours trying to figure out a lot of things but at the end it was really worth it.