Twitter Data Wrangling Project

1. Gathering data

this project starts with gathering data from multiple source

1. csv file
2. tsv file downloaded using `requests` library
3. accessing twitter site using `tweepy` library and foring text file.
   contains json object representing retweetting

the output of these steps **3 `pandas DataFrames`**

2. Assessing Data

data assessment is the process of evaluating data in order to determine whether they meet the quality required for projects or business processes

and usually, it did in two steps 1. Visual assessment 2. programmatic assessment

1. Visual assessment

in WeRateDogs wrangling project I used a spreadsheet and text editor to examine and assess some random data

2. programmatic assessment

for programmatic assessment I used pandas liberary and jupyter notebook like
df.header(), df.info(),df.sample(), df.value_counts(), df.unique(), df[mask], df.describe()

here is the list of data quality issue founded in WeRateDogs project

1. `timestamp` and `retweeted_status_timestamp` data type is object
2. some tweets don't have image
3. some tweets are not originals `in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp`
4. `doggo, floofer, pupper, puppo` null data is None not Nan
5. column `name` some time have incorrect name as "a, an, or number"
6. some columns at this point would have no use in analysis `in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp`
7. The `rating_numerator` column should of type float and also it should be correctly extracted from `text` column
8. columns name not descriptive `p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog`
9. some images do not belong to original tweets

founded tidiness issues

1. merge the 3 datasets to `archive_df` DataFrame
2. one variable in four columns `doggo, floofer, pupper, puppo`
3. columns doggo, floofer, pupper, puppo have no use

3. Cleaning Data

Data cleaning refers to preparing data for analysis by removing or modifying data that is incomplete, irrelevant, duplicated, or improperly formatted.

this process structured in three steps as

- Define
- Code
- Test

for each of the previous quality and tidiness issue, I followed this structure in cleaning the data these are the solutions for the quality and tidiness issue

Quality

1. change timestamp and retweeted_status_timestamp data type from object to datetime
2. select rows that have images from archive_clean_df
3. delete null rows from `in_reply_to_status_id, retweeted_status_id` columns
4. set None value in `doggo, floofer, pupper, puppo` columns to null
5. drop columns that will have no use for analysis process `in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp`
6. clear incorrect column `name` values which have incorrect name like "a, an, or number, ..."
7. The `rating_numerator` column should of type float and also it should be correctly extracted from `text` column
8. rename columns `p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog` name to descriptive names
9. drop images that do not belong to original tweets

Tidiness

1. merge the 3 datasets to `archive_df` DataFrame
2. add new column called classification to hold value of doggo, floofer, pupper, puppo column
3. drop columns doggo, floofer, pupper, puppo