WeRateDogs project
(wrangle report)

Data Gathering:

In the project I needed to gather 3 files in 3 different formats , every file will make a sigle dataframe , and that's how I handle every format and gather it to my wrangle_act.ipynb file

1- twitter-archive-enhanced.csv (CSV file) :

I used pandas library to read (twitter-archive-enhanced.csv) file so I can make my first dataframe df_archive.

2- image_predictions.tsv (downloaded programmatically) :

I used requests library to get given url in the classroom and split file name from url as tsv file (image_predictions.tsv file) and wrote it in the operating system then I used pandas library to read tsv file so I can make my second dataframe df_image_pred.

3- twitter api for additional data (json file):

I can't use tweepy library to get the json file because I couldn't get access to developer account on twitter , i followed the instructions from twitter but I didn’t get answer so I had to use (tweet_json.txt) file directly , i made the code which It was meant to be used if I could use the tweepy library.

I don’t use lines directly from json file as they are strings so I use code json.loades(line) to convert lines to tweets in dictionary type which by it I can split tweet_id , retweet_count , fav_count and user_count from tweet dictionary.

I made empty df_list to put on it (tweet_id, retweet_count, fav_count, user_count) from tweet_json.txt file then I convert them to the third dataframe df_api .

Data Assessment:

Visual assessment:

1. only original tweets required no retweets , replies or tweets without images .
2. missing values of in reply to status id,in reply to user id,retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp in archived dataframe .
3. unavailable dognames (none, a, an , the).
4. tiddness (df_api + df_archive + df_api) .
5. dogs stage type are columns values shouldn't be columns names .
6. p2 & p3 columns not needed , the first prediction is higher probability .
7. names of p1 & p1_dog & p1_conf is not Better representation .
8. different predictions for dog images .
9. some predictions are not dogs .

programmatic assessment:

1. tweet id,in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id are strings not floats .
2. time stamp , retweeted_status_timestamp are datetimes not strings.
3. dog stages are category not string .
4. unlogical rating values .
5. tweet_id is string .
6. img_num is category .

assessment summary

A- Missing data (completeness issue) :

1. missing values of in reply to status id,in reply to user id,retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp in archived dataframe .

B- Tidiness issues:

1. requiered one dataframe .
2. dogs stage type are columns values shouldn't be columns names .

3. **p2 & p3 columns not needed , the first prediction is higher probabilty .**
4. **names of p1 & p1_dog & p1_conf is not Better representation .**

C- Quality issues:

Validity:

1. **only original tweets required not retweets or replies or tweets without images .**
2. **tweet id is string not floats .**
3. **time stamp is datetimes not strings .**
4. **dog stages are category not string .**
5. **img_num is category .**

Accuracy:

1. **unavailable dognames (none, a, an , the).**

Consistency:

1. **i will uniform the rating method so the rating_numerator from 5 to 15 and rating_denominator must be 10 .**
2. **columns arrange in not the best way.**

Data Cleaning:

1- Quality issues:

	issue	solution
1	only original tweets required not retweets or replies or tweets without images .	1- from image df take only tweet ids of tweets with images to the archived df then take the df of orginal tweets no replies or retweets. 2- also for the image_cleaned dataframe , drop the retweet and replies.
2	missing values of in reply to status id , in reply to user id , retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp in archived dataframe .	drop all columns of retweet and replies as it is not needed for my cleaned dataframe , the reqiered is orginal tweets.

3	unavailable dognames (none, a, an , the).	convert the weird names to NaN
4	tweet id is string not floats .	Use as type method
5	time stamp is datetimes not strings .	Use as type method
6	dog stages are category not string .	Use as type method
7	img_num is category .	Use as type method
8	rating_numerators and rating_denominators are incorrect .	1- make the denominator to be 10 or a number is divisble by 10 when the tweet has more than 1 dog in image. 2- according to the old tweets and the rating method used in the beginning of weratedogs twitter account when the numenator is less the denominator, i will suppose the minimum numerator is 5 . 3- i will use the dataframe when the ratio between rating_numerator and rating_denominator is less than 1.5 , the logical ratio used in weratedogs account is less than 15/10. 4- before all , deal with the rating_numerator as a float extracted from text.
9	columns arrange in not the best way.	1 – I will add new column (rating ratio) 2- rearrang columns

2- Tidiness issues:

	issue	solution
1	p2 & p3 columns not needed , the first prediction is higher probabilty.	drop columns of p2 & p3.
2	names of p1 & p1_dog & p1_conf is not Better representation	1- rename p1 to breed. 2-rename p1_conf to accuracy. 3-rename p1_dog to is_dog.
3	dogs stage type are columns values shouldn't be columns names	add all floofer, pupper, puppo columns to one column dog_stage.
4	requiered one dataframe	Merge the 3 dataframes after cleaning.