Wrangle report

My wrangling process contain 3 main phases:

- **Gathering**
  - **Read file csv**
  - **Request file from URL**
  - **Use twitter API**
- **Assessing**
  - **Visual assessing**
  - **Programmatically assessing**
  - **Resalt => quality problems & tideness problems**
- **Cleaning**
  - **Solve problems in assessing phase**

Phase 1: Gathering

1. **For first part of data archive twitter file, I use pandas function (read_csv) to read the file and add this data in archive_df dataframe.**

2. **For the second part of data Image predictions file, I use the requests function requests.get(url) to get the data from URL then I use os function path.join to extract the from this data.**

3. **For the third part of data API, I use tweepy library and json to extract the data from twitter and save tweets in a json format in txt file tweet_json.txt , then I load data form json file row by row, then save them in tweets list; after that I looped at tweets list to extract specific columns (tweet_id , retweet_count, favorite_count) and save them in lists , then concatenate them in dataframe called api_df.**

Phase 2: Assessing

I use two strategies to assess the data (visual , programmatically assess), by two types (quality, tideness)

By visual strategy I find:

In the archive dataframe:

1. values of name is called 'none' quality
2. values of name is called 'a' and 'the' and 'an' quality
3. incorrected values at (in_reply_to_status_id , in_reply_to_user_id) quality
4. values of (doggo,floofer, pupper, puppo) is called 'None' quality

in the Image predictions dataframe:

1. name of variable not clear quality

By programmatically strategy using function (head, info, describe, tail, duplicated, isnull, value_counts)

I find:

In the archive dataframe:

1. - missing data at (in_reply_to_status_id ,in_reply_to_user_id ,retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) quality
2. - missing data at (expanded_urls) quality
3. - (timestamp) is string quality
4. - (rating_numerator and rating_denominator) are strings quality
5. - (retweeted_status_timestamp) is string quality
6. - rate not valid(rating_denominator= 0 ,>10 and rating_numerator= 0 ,>14) quality
7. - archive table not have one object tideness
8. - column headers are values not variable name (doggo floofer pupper puppo) tideness
9. - multiple variables are stored in one variable(text) tideness

Phase 3: Cleaning

in this phase I try to solve problems that faced me at phase 2

so I will get all problems then solve them.

In the archive dataframe:

1. values of name is called 'a' and ‘the’ and ‘an’ quality
   convert 'a' , ‘the’ and ‘an’ to None
2. incorrected values at (in_reply_to_status_id , in_reply_to_user_id) quality
   drop columns (in_reply_to_status_id , in_reply_to_user_id)
3. values of (doggo,floofer, pupper, puppo) is called 'None' quality
   put doges data in new dataframe and drop rows that have None at all columns

in the Image predictions dataframe:

4. name of variable not clear quality

   change name of variables to clear names (prediction1, prediction2, prediction3, prediction1_confidence, prediction2_confidence, prediction2_confidence)

By programmatically strategy using function (head, info, describe, tail, duplicated, isnull, value_counts)

I find:

In the archive dataframe:

1. - missing data at (in_reply_to_status_id ,in_reply_to_user_id ,retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) quality
   drop columns (in_reply_to_status_id ,in_reply_to_user_id ,retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)

2. - missing data at (expanded_urls) quality

Drop rows that have null values in expanded_url

3. - (timestamp) is string quality
   convert datatype to datetime
4. - (rating_numerator and rating_denominator) are strings quality
   Convert datatype to float
5. - (retweeted_status_timestamp) is string quality
   we drop them in 2nd problem
6. - rate not valid(rating_denominator= 0 ,>10 and rating_numerator= 0 ,>14) quality
   drop rows that have rating_denominator= 0 ,>10 or rating_numerator= 0 ,>14
7. - archive table not have one object tideness
8. convert header of variables from doggo, floofer, pupper and puppo to one variable stadge_of_dog and store values in stadge_of_dogs dataframe then drop name,doggo, floofer, pupper and puppo from archive_copy dataframe.
9. - column headers are values not variable name (doggo floofer pupper puppo) tideness

   convert header of variables from doggo, floofer, pupper and puppo to one variable stadge_of_dog and store values in stadge_of_dogs dataframe then drop name,doggo, floofer, pupper and puppo from archive_copy dataframe.

10. - multiple variables are stored in one variable(text) tideness
    extract link from text and put in column 'link' and add text to 'Text' column then drop text column.

After finishing cleaning phase I merge all data in one dataframe called ‘master_dataframe’ then store them in csv file called ‘twitter_archive_master.csv’

And store dataframe of doges that called ‘dog_rate in another file ‘dogs.csv’.