wrangle_report

1) Gathering data

I gathered Data from different sources

- I read twitter_archived from csv file
- And read image_predictions from a tsv file using url after programmatically downloading it
- Finally I read a json file in a DataFrame called api_df

2) Assessing data

I have started assessing data file by file programmatically and visually

I have assessed twitter_archived in the first visually using [.sample(5) & .head()]

After that I assessed it programmatically using some functions | methods like [.info & .describe & .count &value_counts &duplicated]

While assessing the data I have discovered some issues

- **Quality issues**

  1- tweet_id must be string instead of int

  2- timestamp must be date_time instead of string

  3- 'None' values in name

  4- 'None' values in dog_stage

5. **unnecessary rating_denominator column**
6. **unnecessary columns like (in_reply_to_status_id ,in_reply_to_status_id, retweeted_status_id , retweeted_status_user_id)**
7. **missing values in expanded_urls**

- **Tidiness issues**

1. **dogs stage have 4 columns despite it is 1 variable deserve only 1 col**

Secondly I have assessed image_predictions table first visually using [.head()]
After that I assessed it programmatically using
some functions | methods like
[.info & duplicated]

While assessing the data I have discovered some issues

- **Quality issues**
  1. **tweet_id must be string instead of int**
  2. **drop unnecessary all false results in p1_dog & p2_dog & p3_dog**

finally , I have assessed api_df table first visually using [head()]

After that I assessed it programmatically using some functions | methods like

[info & describe]

While assessing the data I have discovered some issues

- **Quality issues**

  tweet_id data type must be object instead of int

- **Tidiness issues**

  all data must be combined in one table not separated

3) Cleaning data

In the beginning I started with copying the original data before cleaning using copy method

I have started cleaning issue by issue and started from **quality** issues

1- I have converted tweet_id from int to string in image_predictions_clean table

2- I have converted tweet_id from int to string in twitter_archived_clean table

3- I have converted tweet_id from int to string in api_df_clean table
4- I have converted timestamp to date_time in twitter_archived_clean
5- I have dropped the unnecessary columns for retweet and replies in twitter_archived_clean table
6- I have replaced None with nan values
7- I have dropped rating dominator column
8- I have removed rows for missing values in expanded urls column
9- I have dropped unnecessary all false results in p1_dog & p2_dog & p3_dog

Secondly , I have cleaned **tidiness** issues

1- I have made a one col for one variable instead of 4 columns
2- I have combined all data in one table

4) storing data

After wrangling I saved the data in a csv file using to_csv method