Wrangling Act Report

Gathering Data

1. Data from twitter achieve → normally load using pd.read_csv
2. Prediction data downloaded programmatically via requests.get and saved into the file under name 'prediction.csv'
3. Twitter data is downloaded through API
    a. Because my internet download speed is slow , I have to split the twitter_id into 2 part in order to not causes and error.
    b. And as the internet download speed is slow, I can't go back and revise the coding.. therefore I decide to redownload the NaN data

<u>Data issue I investigate into Is as follow</u>

Assessing Data

Issue in main:

- Quality :Correct the wrong data in rating_numerator
- Quality correct the data type; change rating numerator and denominator to float, change timestamp to datatime

Issue in prediction:

- Quality issue: if not dog drop it

Issue in TweetData:

- Quality issue: correct the data type (twitterdata_a on tweet_id); this was done in order to merge the data properly

Merge data

- Tidy issue: Prediction combine with Tweeter Data = Data A
- Tidy issue: combine Data A with main as Data B
- Quality issue: Eliminate the one without tweeter data retreivable (mean post might be delete therefore no favorite and retweet count)
- Quality issue: Take one with highest % prediction as a dog type
- Tidy Issue: Eliminate unnecessary column
- Quality issue : Fix if the highest confidence is not dog
- Quality issue: Rating of 1 dog = 0 --> drop this dog out because it is actually a plagarism post not about the dog rate itself

Cleaning

Cleaning

Issue in main:

- Fixed: Quality :Correct the wrong data in rating_numerator using extract function and replace that number with the old one
- Fixed: Quality correct the data type
  Clean Use astype to change the datatype to a correct type, change rating numerator and denominator to float, change timestamp to datatime

Issue in prediction:

- fixed: Quality issue: if not dog drop it,
  cleaning process: If not dog drop it, use query to detect if no field from p1_dog,p2_dog,p3_dog mean the data is not a dog

Issue in TweetData:

- Fixed Tidy issue: correct the data type (twitterdata_a on tweet_id); this was done in order to merge the data properly

  Cleaning process: astype to change datatype to correct type

Merge data

- Fixed: Tidy issue: Prediction combine with Tweeter Data = Data A

  Clean use merge data

- Fixed: Tidy issue: combine Data A with main as Data B

  Clean using merge data

- Fixed: Quality issue: Eliminate the one without tweeter data retreivable (mean post might be delete therefore no favorite and retweet count)

  Clean using Query

- Fixed: Quality issue: Take one with highest % prediction as a dog type

  clean using query and apply method

- Fiexd: Tidy Issue: Eliminate unnecessary column

  clean using drop

- Fixed: Quality issue : Fix if the highest confidence is not dog

  clean using query

- Fixed: Quality issue: Rating of 1 dog = 0 --> drop this dog out because it is actually a plagarism post not about the dog rate itself

  clean using query

Reason why most of the part were clean after the data was merge is because I detect these problem when I did my analysis part, so I decide to clean in merge data set in order to not cause a error in a previous algorithm.

I save the result as “cleaned_dograte.csv”
To use in analysis part