WRANGLE DATA

Wrangling data divided into 3 main sections

1. Gathering data
2. Assessing data
3. Cleaning data

Will discuss each section individually

1) Gathering data: gather data from 3 different sources
   i) Csv file : comma separated values file , which contain data of WeRateDogs twitter archive and this file download it manually then read the file in code use pd.read_csv('file_path') then data in file will be loaded in dataframe and ready to work in it
   ii) Download tsv file: first tsv means tab separated values file. second download this file from requests library then open by writing it to tsv file then load it in dataframe using use pd.read_csv('file_path',sep='\t')
   iii) Download tweet_json.txt: read file by pd.read_json() and assign values in dataframe called tweet_info

2) Assessing data: assess data divided to 2 main issues
   a) Quality issue: dirty, content issue
   And in Quality issue solve many problems as :

   - some missing values in columns 'doggo' , 'floofer' , 'pupper' ,'puppo' will extract it from 'text' column by extract('Regex')
   - replace caps to lowercase values in columns 'doggo' , 'floofer' , 'pupper' ,'puppo' after extracting from text
   - replace None values to null to be easy later (use isnull) in columns 'doggo' , 'floofer' , 'pupper' ,'puppo'
   - replace None values to null to be easy later (use isnull) in name column
   - change datatype of 'timestamp' , 'retweeted_status_timestamp' columns from string/object to datetime
   - change all 'rating_denominator' column values to 10
   - change all 'rating_numerator' column values to be large than 10 so the first number large than 10 is 11 , so will assign all values less than 10 to 11
   - incorrect dog names so extract it from text column by extract('Regex')
   - make breed column which contain name of breed of dog according to p1_dog which true
   - rename 'id' column in dataframe tweet_info into 'tweet_id'

   b) Tidiness issue:messy, structural issue
   - add new column called stage to save in it the dog stage which divided in 4 column 'doggo' , 'floofer' , 'pupper' ,'puppo' so merge them all in 1 colmn stage

   - add jpg_url' , 'type' columns to df_enhanced dataframe
   - add 'retweet_count' , 'favorite_count' columns to df_enhanced dataframe

3) Cleaning data : this also divided into 3 sections
   After assessing data and have some issues so take one by one issue and solve it
   Solve it by writing 3 sections of cleaning
   a) Define : which define in few words the issue and the way will solve it
   b) Code : write the running code
   c) Test : test the result the come from the running coding