Wrangle Report

1. Objective:
    - Find the top N WeRateDogs favorite dogs
    - Find the top N retweet count dogs
    - Find the top N favorite count dogs

When display the dogs, should show the dog's basic information, including 'text', 'name', 'stage', 'breed', 'rating_numerator' or 'retweet_count' or 'favorite_count', 'image'.

2. Read the existing data file in
    - Read in twitter-archive-enhanced.csv file as dog_rates table
    - Read in image-predictions.tsv as dog_rates_image table

3. Gather data and create the dog_rates_add dataframe

    In order to realize our objective, we need to collect more data except what we already have. Firstly, we need to register one twitter account. In order to get the twitter data easily, then tweepy is installed. Since we will often access to twitter web, we need one proxy. After everything is done, we can begin our gathering data travel.
    - Download the json data based on the tweet_id in dog_rates table
    - Save the json data using json.dump
    - Read the json file to one list using json.loads
    - Understand the json file
    - Create the dog_rates_add dataframe, including the information, e.g., 'tweet_id', 'retweet_count', 'favorite_count', 'followers_count'. It's very easy to get the data, just regard json file as one dict.

4. Assess the data

    We want to combine all the necessary data to one table
    - Quality
        - ✓ timestamp should be datetime
        - ✓ source, expanded url seems not that useful, can drop
        - ✓ some items are duplicate, e.g., 'This is Charlie. He fell asleep on a heating vent...' All duplicate are from the items that author retweets his own tweet. no matter, we will drop all the in_reply and retweet ones this time
        - ✓ for the first time, we just focus on the tweet created directly by author, so drop in_reply and retweet onesretweet ones. And drop the items and related columns
        - ✓ rating_denominator min/max value is not reasonable, because several dogs in one picture
        - ✓ the rating_numerator is not extracted correctly, the text may include two xx/xx string, should get the score one, e.g, '...ok jomny I know

you're excited but 960/00 isn't a valid rating, 13/10 is tho...'

- ✓ some xx/xx strings are not rating value, e.g., '...She smiles 24/7 &amp...', its tweet_id is 810984652412424192
- ✓ rating for tweet_id 676957860086095872 is wrong, extract not correctly
- ✓ name is not extracted correctly, 'This is by far', should not always extract the word after 'This is'. remove the wrong names, e.g., 'a', 'an', 'the'.etc
- ✓ (not done this version) name "jack is xxx", some names can not be extracted correctly
- ✓ some stage items are not right, e.g., 'doggopupper', 'doggofloofer'.etc
- ✓ for the dog bread, just keep the dog breed that confidience value bigger than p_value. rename p1 to breed
- ✓ for name, stage, breed. make all the none to '', to display more clearly
- ✓ add another two columns, retweet_probability, favorite_probability, float, then recalculate by deviding follower_counts
- ✓ drop the unuseful column, rating_denominator

- Tidiness
  - ✓ image url, p1, p1_conf, p1_dog in dog_rates_image table should be integrated to dog_rates table
  - ✓ all the information in dog_rates_add table should be integrated to dog_rates table
  - ✓ doggo, floofer, pupper, puppo in dog_rates table should be one column, because all describe the stage of dog.

5. Clean the data

   Before the cleaning travel, we need to make a copy for the original dog_rates dataframe, name it dog_rates_clean. We will use the dog_rates_clean table later. Actually, some of the quality issues above are found in clean phase, then add them in. We usually can find the quality issues when we are in cleaning phase, that's ok, just add them to the quality items and do the cleaning.

   - ✓ In reply and retweet items are dropped this time, however, actually, some replies revise the rating value and some retweets are also useful for our analysis. In next version, will just drop the retweets that author retweet his own.
   - ✓ Rating value revision should be one big task in this cleaning part, we use re package to extract what we want. Moreover, revision is needed for the multiple dogs image.
   - ✓ We add the retweet probability, favorite probability, however, since there's no big difference for follower count. The result is the same with just using count.

6. Display what we want to show

   This part will be wrote in another act_report.pdf