Gathering Data

First thing I have done is gathering data from twitter-archive-enhanced.csv by download from resources on Udacity then read it using “pd.read_csv“

Then download programmatically from URL: (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv) By using requests.get(url) then read it by pd.read_csv('image-predictions.tsv', sep='\t')

Then by using Tweepy to query Twitter's API for data included in the WeRateDogs Twitter archive. From this data gets retweet count and favorite count

First I gain the access from twitter developer get the (API key is consumer_key, API secret key is consumer_secret, Access Token is access_token, Access Token Secret is access_secret)

Then

```
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)

auth.set_access_token(access_token, access_secret)
```

After getting IDs most of them get without errors but number id with errors equal 25

Then write its JSON data to a tweet_json.txt file with each tweet's JSON data on its own line

```
#storing data as tweet_json.txt

with open('tweet_json.txt', 'w') as f:

 json.dump(Tweets, f)
```

Then read the file 'tweet_json.txt' by using pd.read_json('tweet_json.txt')

Then convert it to Data Frame by using panda

Assessing Data for this Project

First thing assessing all gathered data by print it and then check for duplicated values then get info() of each data frame by searching manually and checking values from 3 tables to find quality issues and Tidiness issues then write it down as follow:

Quality issues

1. duplicates in dfjson (tweet_json)
2. duplicates in ip.jpg_url (image)
3. fix types in column in_reply_to_status_id , in_reply_to_user_id,retweeted_status_id, retweeted_status_user_id

4. source column has HTML
5. rating_denominator should all to be 10
6. timestamp to datetime.
7. null values
8. rating= rating_numerator/rating_denominator
9. types of tweet id should be object not int
10. delete unneeded columns

11. select two types of dogs at same stage
12. p1 p2 p3 lower cases

13. p1 p2 p3 replace _ with space

Tidiness issues

1. Doggo, floofer, pupper and puppo columns to be in one column.
2. Name of column id in table dfjson need to change to tweet_id.

Steps:

1. Make copy of all data frame to clean it and keep source as same
2. Remove duplicates in dfjson
3. Remove duplicates image
4. Fix types of columns in_reply_to_status_id , in_reply_to_user_id,retweeted_status_id, retweeted_status_user_id

5. Fix type in tweet id in every data frame
6. Remove Html from source and then change to type category
7. Make all denominator to be equal 10
8. Then check all of denominators by duplicates
9. Fix type of timestamp to datetime
10. Calculate Rate by dividing num / den
11. Check values of rate
12. Change Ip name in column by using title() method
13. Then replace _ by space
14. Then checking when I checked I found (–) used as spacing
15. Then replace - by space
16. Then delete unneeded columns like: ['in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp','expanded_urls']
17. Change none value to NaN in columns (['doggo', 'floofer', 'pupper', 'puppo')
18. Then create another column contain types merge by each other
19. Check if there any problem in these method
20. Then remove columns ['doggo', 'floofer', 'pupper', 'puppo']
21. Then rename column in dfjson from id to tweet_id
22. Then merge all three data frames in one data frame
23. Then calculate number of null values
24. Then remove it
25. Then test
26. Then storing df to twitter_archive_master.csv