wrangle_act

November 9, 2022

1 Project: Wrangling and Analyze Data

1.1 Data Gathering

In the cell below, gather **all** three pieces of data for this project and load them in the notebook. **Note:** the methods required to gather each data are different. 1. Directly download the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv)

```
In [88]: import pandas as pd
         import numpy as np
         import requests
         import os
         import tweepy
         from tweepy import OAuthHandler
         import json
         from timeit import default_timer as timer
         import warnings
         warnings.filterwarnings('ignore')
         import seaborn as sb
         import matplotlib.pyplot as plt
         %matplotlib inline

In [89]: #Read the csv file into the dataframe
         df_tweet= pd.read_csv('twitter-archive-enhanced.csv')

In [90]: list(df_tweet)

Out[90]: ['tweet_id',
          'in_reply_to_status_id',
          'in_reply_to_user_id',
          'timestamp',
          'source',
          'text',
          'retweeted_status_id',
          'retweeted_status_user_id',
          'retweeted_status_timestamp',
          'expanded_urls',
          'rating_numerator',
```

```
 'rating_denominator',
 'name',
 'doggo',
 'floofer',
 'pupper',
 'puppo']
```

2. Use the Requests library to download the tweet image prediction (image_predictions.tsv)

```
In [91]: #this is the folder where the file for image_predictions.tsv will be downloaded
         url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predicti
         response = requests.get(url)
         response

Out[91]: <Response [200]>

In [92]: with open('image-prediction.tsv', 'wb') as file:
             file.write(response.content)

In [93]: #To read the image predictions table into a dataframe
         image = pd.read_csv('image-predictions.tsv', sep='\t')

In [94]: os.listdir()

Out[94]: ['image-prediction.tsv',
          'twitter-archive-enhanced.csv',
          'image',
          'act_report.ipynb',
          'tweet_json.txt',
          'image-predictions.tsv',
          '.ipynb_checkpoints',
          'wrangle_report.ipynb',
          'wrangle_act.ipynb']
```

3. Use the Tweepy library to query additional data via the Twitter API (tweet_json.txt)

```
In [9]: # Query Twitter API for each tweet in the Twitter archive and save JSON in a text file
        # These are hidden to comply with Twitter's API terms and conditions
        consumer_key = 'HIDDEN'
        consumer_secret ='HIDDEN'
        access_token = 'HIDDEN'
        access_secret = 'HIDDEN'
        auth = OAuthHandler(consumer_key, consumer_secret)
        auth.set_access_token(access_token, access_secret)

        api = tweepy.API(auth, wait_on_rate_limit=True)

        # Tweet IDs for which to gather additional data via Twitter's API
        tweet_ids = df_tweet.tweet_id.values
```

```
        len(tweet_ids)

        # Query Twitter's API for JSON data for each tweet ID in the Twitter archive
        count = 0
        fails_dict = {}
        start = timer()
        # Save each tweet's returned JSON as a new line in a .txt file
        with open('tweet_json.txt', 'w') as outfile:
            # This loop will likely take 20-30 minutes to run because of Twitter's rate limit
            for tweet_id in tweet_ids:
                count += 1
                print(str(count) + ": " + str(tweet_id))
                try:
                    tweet = api.get_status(tweet_id, tweet_mode='extended')
                    print("Success")
                    json.dump(tweet._json, outfile)
                    outfile.write('\n')
                except tweepy.TweepError as e:
                    print("Fail")
                    fails_dict[tweet_id] = e
                    pass
        end = timer()
        print(end - start)
        print(fails_dict)
```

```
1: 892420643555336193
Success
2: 892177421306343426
Success
3: 891815181378084864
Success
4: 891689557279858688
Success
5: 891327558926688256
Success
6: 891087950875897856
Success
7: 890971913173991426
Success
8: 890729181411237888
Success
9: 890609185150312448
Success
10: 890240255349198849
Success
11: 890006608113172480
Success
12: 889880896479866881
```

```
Success
13: 889665388333682689
Success
14: 889638837579907072
Success
15: 889531135344209921
Success
16: 889278841981685760
Success
17: 888917238123831296
Success
18: 888804989199671297
Success
19: 888554962724278272
Success
20: 888202515573088257
Fail
21: 888078434458587136
Success
22: 887705289381826560
Success
23: 887517139158093824
Success
24: 887473957103951883
Success
25: 887343217045368832
Success
26: 887101392804085760
Success
27: 886983233522544640
Success
28: 886736880519319552
Success
29: 886680336477933568
Success
30: 886366144734445568
Success
31: 886267009285017600
Success
32: 886258384151887873
Success
33: 886054160059072513
Success
34: 885984800019947520
Success
35: 885528943205470208
Success
36: 885518971528720385
```

```
Success
37: 885311592912609280
Success
38: 885167619883638784
Success
39: 884925521741709313
Success
40: 884876753390489601
Success
41: 884562892145688576
Success
42: 884441805382717440
Success
43: 884247878851493888
Success
44: 884162670584377345
Success
45: 883838122936631299
Success
46: 883482846933004288
Success
47: 883360690899218434
Success
48: 883117836046086144
Success
49: 882992080364220416
Success
50: 882762694511734784
Success
51: 882627270321602560
Success
52: 882268110199369728
Success
53: 882045870035918850
Success
54: 881906580714921986
Success
55: 881666595344535552
Success
56: 881633300179243008
Success
57: 881536004380872706
Success
58: 881268444196462592
Success
59: 880935762899988482
Success
60: 880872448815771648
```

```
Success
61: 880465832366813184
Success
62: 880221127280381952
Success
63: 880095782870896641
Success
64: 879862464715927552
Success
65: 879674319642796034
Success
66: 879492040517615616
Success
67: 879415818425184262
Success
68: 879376492567855104
Success
69: 879130579576475649
Success
70: 879050749262655488
Success
71: 879008229531029506
Success
72: 878776093423087618
Success
73: 878604707211726852
Success
74: 878404777348136964
Success
75: 878316110768087041
Success
76: 878281511006478336
Success
77: 878057613040115712
Success
78: 877736472329191424
Success
79: 877611172832227328
Success
80: 877556246731214848
Success
81: 877316821321428993
Success
82: 877201837425926144
Success
83: 876838120628539392
Success
84: 876537666061221889
```

```
Success
85: 876484053909872640
Success
86: 876120275196170240
Success
87: 875747767867523072
Success
88: 875144289856114688
Success
89: 875097192612077568
Success
90: 875021211251597312
Success
91: 874680097055178752
Success
92: 874434818259525634
Success
93: 874296783580663808
Success
94: 874057562936811520
Success
95: 874012996292530176
Success
96: 873697596434513921
Fail
97: 873580283840344065
Success
98: 873337748698140672
Success
99: 873213775632977920
Success
100: 872967104147763200
Success
101: 872820683541237760
Success
102: 872668790621863937
Fail
103: 872620804844003328
Success
104: 872486979161796608
Success
105: 872261713294495745
Fail
106: 872122724285648897
Success
107: 871879754684805121
Success
108: 871762521631449091
```

```
Success
109: 871515927908634625
Success
110: 871166179821445120
Success
111: 871102520638267392
Success
112: 871032628920680449
Success
113: 870804317367881728
Success
114: 870726314365509632
Success
115: 870656317836468226
Success
116: 870374049280663552
Success
117: 870308999962521604
Success
118: 870063196459192321
Success
119: 869988702071779329
Fail
120: 869772420881756160
Success
121: 869702957897576449
Success
122: 869596645499047938
Success
123: 869227993411051520
Success
124: 868880397819494401
Success
125: 868639477480148993
Success
126: 868622495443632128
Success
127: 868552278524837888
Success
128: 867900495410671616
Success
129: 867774946302451713
Success
130: 867421006826221569
Success
131: 867072653475098625
Success
132: 867051520902168576
```

```
Success
133: 866816280283807744
Fail
134: 866720684873056260
Success
135: 866686824827068416
Success
136: 866450705531457537
Success
137: 866334964761202691
Success
138: 866094527597207552
Success
139: 865718153858494464
Success
140: 865359393868664832
Success
141: 865006731092295680
Success
142: 864873206498414592
Success
143: 864279568663928832
Success
144: 864197398364647424
Success
145: 863907417377173506
Success
146: 863553081350529029
Success
147: 863471782782697472
Success
148: 863432100342583297
Success
149: 863427515083354112
Success
150: 863079547188785154
Success
151: 863062471531167744
Success
152: 862831371563274240
Success
153: 862722525377298433
Success
154: 862457590147678208
Success
155: 862096992088072192
Success
156: 861769973181624320
```

```
Fail
157: 861383897657036800
Success
158: 861288531465048066
Success
159: 861005113778896900
Success
160: 860981674716409858
Success
161: 860924035999428608
Success
162: 860563773140209665
Success
163: 860524505164394496
Success
164: 860276583193509888
Success
165: 860184849394610176
Success
166: 860177593139703809
Success
167: 859924526012018688
Success
168: 859851578198683649
Success
169: 859607811541651456
Success
170: 859196978902773760
Success
171: 859074603037188101
Success
172: 858860390427611136
Success
173: 858843525470990336
Success
174: 858471635011153920
Success
175: 858107933456039936
Success
176: 857989990357356544
Success
177: 857746408056729600
Success
178: 857393404942143489
Success
179: 857263160327368704
Success
180: 857214891891077121
```