We Rate Dogs Data Wrangling poroject.

Gathering Data

I started my project by downloading the ‘twitter-archive-enhanced.csv’ file manually. Then, created a folder named 'image_predictions' before I downloaded ‘image-predictions.tsv’ programmatically from Udacity's server using the requests library. Next, I wrote it into image_predictions.tsv

‘twitter_data’ was created by accessing and downloading Twitter’s JSON data using the tweepy library. Firstly, I extracted a list of tweet ID from the ‘twitter-archive-enhanced.csv’ file, then looped through each ID and query Twitter’s API with the ID to get each tweet’s JSON data. Subsequently, I recorded the data in a text file named ‘tweet-json.txt’, with each tweet’s data written in a new line. After the query was completed and all the data was written in the text file, I read the text file line by line, obtained each tweet’s information (tweet ID, retweet count, favorite count, and followers count) using the json library, and appended the information into an empty list. Finally, I convert the list of dictionaries to a pandas DataFrame and saved it into ‘twitter_data’

I started my assignment by manually downloading the file 'twitter-archive-enhanced.csv'. Then I created a folder named 'image_predictions' before downloading 'image-predictions.tsv' programmatically from the Udacity server using the requests library. Then I wrote it to 'image_predictions.tsv'.

twitter_data' was created by accessing and downloading JSON data from udacity, I’ve requested access to twitter API and my request was not approved. Firstly, I extracted a list of tweet IDs from the 'twitter-archive-enhanced.csv' file, then went through each ID and queried the Twitter API with the ID to get the JSON data for each tweet. I then saved the data to a text file named 'tweet-json.txt', with the data for each tweet written to a new line. Once the query had been completed and all the data had been written to the text file, I read the text file line by line, obtained the information for each tweet (tweet ID, number of retweets, number of favourites and number of followers) using the json library, and added the information to an empty list. Finally, I converted the dictionary list into a panda DataFrame and saved it in 'twitter_data'.

Assessing and Cleaning

Several quality and cleanliness problems were identified for all three tables. Further details of the problems identified, and the solutions are in the table below:

Quality issue

In [ ]:
```
##Reference : https://stackoverflow.com/questions/64506283/create-a-pandas-table
```

Twitter archive table

In [ ]:
```
```

In [26]:
```
import pandas as pd
>>> d = {
...     'ISSUES': ['Keep original ratings that have images',
                   'Erroneous datatypes in these columns ',
                   'Missing values in name and dog','Some records have more than on dog
'Source column is in HTML-formatted string not a normal string',
'Error in dog names',
'Some values in rating_numerator not showing proper float values',
'Text column includes a text and a short link'],
...     'SOLUTIONS': ['Delete retweets by filtering the NaN of retweeted_status_user_id
                      'Separate the dog stages to know which records have more than one
'Extract HTML values from source',
'Change error name in dog name to None',
'Spot those records and confirm changes made',
'Remove hyperlinks in tweets']
... }
>>> df = pd.DataFrame(data=d)
>>> df
```

Out[26]:

	ISSUES	SOLUTIONS
0	Keep original ratings that have images	Delete retweets by filtering the NaN of retwee...
1	Erroneous datatypes in these columns	Convert timestamp to datetime
2	Missing values in name and dog	Change missing values in dog name to unnamed
3	Some records have more than on dog stage	Separate the dog stages to know which records ...
4	Source column is in HTML-formatted string not ...	Extract HTML values from source
5	Error in dog names	Change error name in dog name to None
6	Some values in rating_numerator not showing pr...	Spot those records and confirm changes made
7	Text column includes a text and a short link	Remove hyperlinks in tweets

Image prediction table

In [40]:
```
import pandas as pd
>>> d = {
...     'ISSUES': ['Erroneous datatype (tweet_id)',
                   'Missing images (only 2075 counts out of possible 2356 '],
...     'SOLUTIONS': ['Convert tweet id to string',
                      'Drop rows with missing images']
... }
>>> df = pd.DataFrame(data=d)
>>> df
```

Out[40]:

	ISSUES	SOLUTIONS
0	Erroneous datatype (tweet_id)	Convert tweet id to string
1	Missing images (only 2075 counts out of possib...	Drop rows with missing images

TWITTER API TABLE

In [44]:
```
>>> d = {
...     'ISSUES': ['Erroneous datatype (tweet_id)'],
...     'SOLUTIONS': ['Convert tweet id to string']
... }
>>> df = pd.DataFrame(data=d)
>>> df
```

Out[44]:

	ISSUES	SOLUTIONS
0	Erroneous datatype (tweet_id)	Convert tweet id to string

TIDINESS

TWITTER ARCHIVE TABLE

In [ ]:
```
##Differents column to be merged in dog_stage column: doggo, floofer, pupper and puppo
```

In [52]:
```
>>> d = {
...     'ISSUES': ['different columns in twitter_archive table should be merged into on
...     'SOLUTIONS': ['Merge columns into one column named "dog_stage"']
... }
>>> df = pd.DataFrame(data=d)
>>> df
```

Out[52]:

	ISSUES	SOLUTIONS
0	different columns in twitter_archive table sho...	Merge columns into one column named "dog_stage"

TWITTER API TABLE

In [56]:
```
>>> d = {
...     'ISSUES': ['twitter api table columns'],
...     'SOLUTIONS': ['Merge table with twitter archive table']
... }
>>> df = pd.DataFrame(data=d)
>>> df
```

Out[56]:

	ISSUES	SOLUTIONS
0	twitter api table columns	Merge table with twitter archive table

IMAGE PREDICTION TABLE

In [54]:
```
>>> d = {
...     'ISSUES': ['Image predictions table should be added to twitter archive table'],
...     'SOLUTIONS': ['Merge table with twitter archive table']
... }
>>> df = pd.DataFrame(data=d)
>>> df
```

Out[54]:

	ISSUES	SOLUTIONS
0	Image predictions table should be added to twi...	Merge table with twitter archive table

Storing Cleaned Data

So now the data set is clean and ready for analysis. I saved the main table as twitter_archive_master.csv.

Then I began my data analysis.

In [ ]: