Data Wrangling Report

Emmanuel Gutierrez

Purpose: To wrangle/cleanup data from 3 sources and to provide insights and visualization from the insights

Wrangle Sources: The following files needed cleaning and restructuring as part of this assignment:

- **twitter-archive-enhanced.csv :** CSV File containing archival information of the tweets sent by the WeRateDogs account.
- **image_predictions.tsv:** File including results of an image prediction algorithm to detect the breed of the dogs
- **tweet_json.txt:** Query results from Twitter API primarily including Twitter-specific metrics such as favorite counts and retweets

Wrangle Process: The wrangle process was divided into 3 steps: Gathering, Assessing, and Cleaning the data.

Gathering: On the gathering process, I imported all the libraries that I needed:

```
#Here I import all of the libraries that I will be using
import pandas
import numpy
import requests
import tweepy
import json
import timeit
import matplotlib.pyplot as matp
```

Next, I imported the 3 data sources using different methods based on the source of the data:

Twitter Archive:

```
#Here I read the twitter archive enhanced CSV with a pandas function and create
ta = pandas.read_csv('twitter-archive-enhanced.csv')
#here I validate the new dataframe has the content
ta.head()
```

Image Prediction:

```
#Here I query the file and create a dataframe titled ip
open('image_predictions.tsv','wb') as sheet:
    sheet.write(r.content)
ip = pandas.read_table('image_predictions.tsv')
ip.head(10)
```

Twitter API Query:

```
#Here I create a dataframe titled "td" to populate the queried twitter data
td = pandas.read_json('tweet_json.txt', lines = True,encoding='utf-8')
td.head()
```

Assessing: To assess the data, I used a combination of info, shape, and print to view the data and its structure from a different lens:

```
#Here I check how many rows, columns the dataframe has
print("twitter-archive-enhanced.csv Row Count, Column Count:")
ta.shape
```

```
twitter-archive-enhanced.csv Row Count, Column Count:

(2356, 17)
```

```
#Here I check the column structures of the dataframe
print("twitter-archive-enhanced.csvv dataframe info:")
ta.info()
```

```
tweet_id                    2356 non-null int64
in_reply_to_status_id       78 non-null float64
in_reply_to_user_id         78 non-null float64
timestamp                   2356 non-null object
source                      2356 non-null object
```

```
#Here I checked the first 10 rows of the dataframe
print("twitter-archive-enhanced.csv First 10 Rows:")
ta.head(10)
```

```
twitter-archive-enhanced.csv First 10 Rows:
```

I also utilized the power of excel and its pivot table function, as well as databricks to further dive into the data insights

Once the assessment was complete, I identified the following **8** quality opportunities and **3** tidiness opportunities:

Quality issues

Twitter Archive Enhanced CSV Issues:

- 2176 are valid tweets, the remaining 181 should be filtered out as they are retweets (based on the non-blank field on retweeted_status_id)
- Through manual scanning, Identified 5 tweets that had the wrong numerator/denominator as the script grabbed a date mentioned in the text field instead of the score
- Identified a tweet that was categorized as both "doggo" and "floofer".
- 55 tweets were populated with "a" as a name
- Source column shows link instead of a label (Phone, Web,etc...)

Image Predictions TSV issues:

- 325 rows had FALSE on all 3 p(x)_dog fields (p1,p2,p3). Indicating that with a high degree of certainty, these rows do not have a dog in the image
- Names should be made a lower case to keep consistency in case through each character

Tweet_json.txt Issues:

- Remove any entries with a retweeted_status that is not false to ensure only original tweets remain
- Change the column titled "ID" to "tweet_id" to fit the naming structure of the other data frames

Tidiness Issues:

Twitter Archive Enhanced CSV Issues:

- Remove the following columns from twitter archive enhanced:

(in_reply_to_status_id, in_reply_to_user_id,text, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)

Image Predictions TSV issues:

- Multiply p1_conf,p2_conf,p3_conf by 100 and round to 2 decimals to make percentages more readable

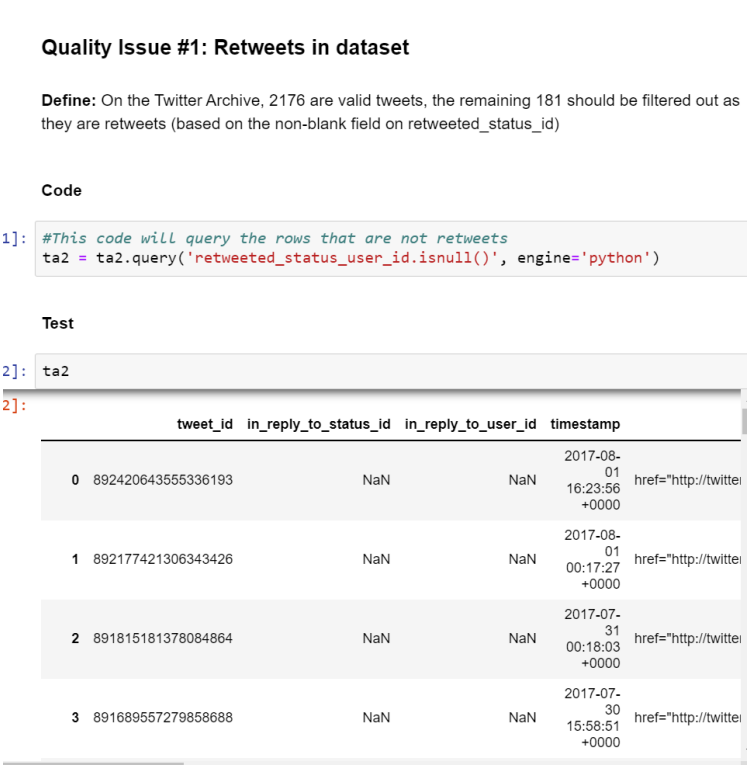
Tweet_json.txt Issues:

- Remove all columns minus tweet_id, favorite_count, retweet_count from tweet_json.txt

All Tables:

- Join the datasets and create one master dataset.

Cleaning: On the cleaning process, I addressed every item on quality and tidiness, and used the various functionalities available on python to clean the data (Examples below):



Quality Issue #1: Retweets in dataset

Define: On the Twitter Archive, 2176 are valid tweets, the remaining 181 should be filtered out as they are retweets (based on the non-blank field on retweeted_status_id)

Code

```
1]: #This code will query the rows that are not retweets
    ta2 = ta2.query('retweeted_status_user_id.isnull()', engine='python')
```

Test

```
2]: ta2
```

2]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	href="http://twitte
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	href="http://twitte
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	href="http://twitte
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	href="http://twitte

Tidiness Issue #1: Retweets in dataset

Define: Remove the following columns from twitter archive enhanced: (in_reply_to_status_id, in_reply_to_user_id,text, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)

Code

```
#Here I remove the columns that I will not need for my analysis from Twitter Arc
ta2 = ta2.drop(columns=['in_reply_to_status_id', 'in_reply_to_user_id', 'text','
```

Test

```
#Here I preview the list to ensure the columns mentioned were dropped
ta2
```

	tweet_id	timestamp	source	expanded_urls	ra
0	892420643555336193	2017-08-01 16:23:56 +0000	Twitter for Iphone	https://twitter.com/dog_rates/status/892420643...	
1	892177421306343426	2017-08-01 00:17:27 +0000	Twitter for Iphone	https://twitter.com/dog_rates/status/892177421...	
2	891815181378084864	2017-07-31 00:18:03 +0000	Twitter for Iphone	https://twitter.com/dog_rates/status/891815181...	
3	891689557279858688	2017-07-30	Twitter for	https://twitter.com/dog_rates/status/891689557	

Conclusion: At the end, I merge all the tables into one to make it easier for analysis. This concluded my wrangling effort which allowed me to begin my analysis phase.