DATA WRANGLING REPORT

Using Python and its libraries, I gathered data from the tweet archive of Twitter user @dog_rates, also known as WeRateDogs as provided by Udacity for this project. These datasets were assessed/wrangled for quality and tidiness, afterwhich it were cleaned, analyzed and visualized.

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings have a denominator of 10, however, many of the numerators are greater than 10, i.e 11/10, 12/10, 13/10, etc. These gave improper fraction or simply akward ratings. This unique rating system is a big part of the popularity of WeRateDogs and such the ratings does not need to be cleaned.

My wrangling efforts on the 'WeRateDogs' dataset project were in the following step:

Step 1:Data gathering using the provided datasets

Step 2: Assessing and cleaning the datasets

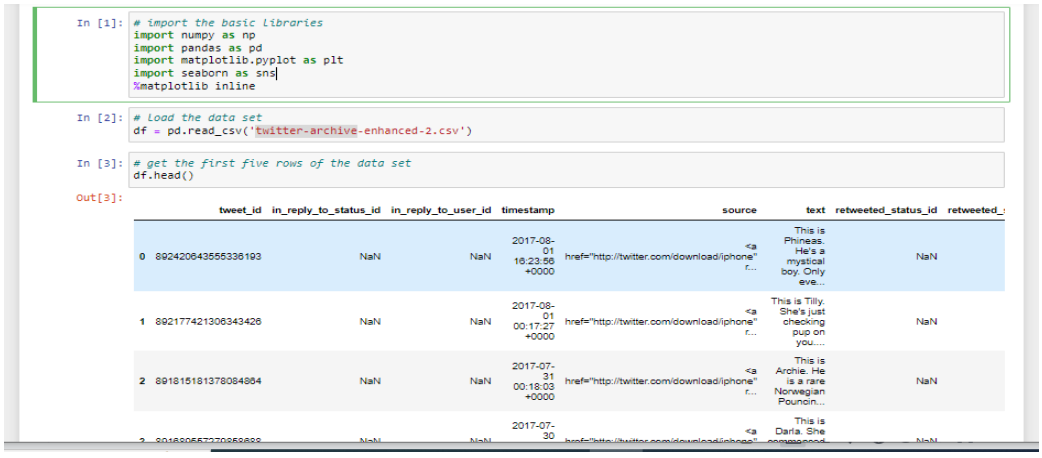
Step 3: Storing the data

Step 4: Conducting exploratory data analysis and creating visualizations

Step 5: Insights and Reporting

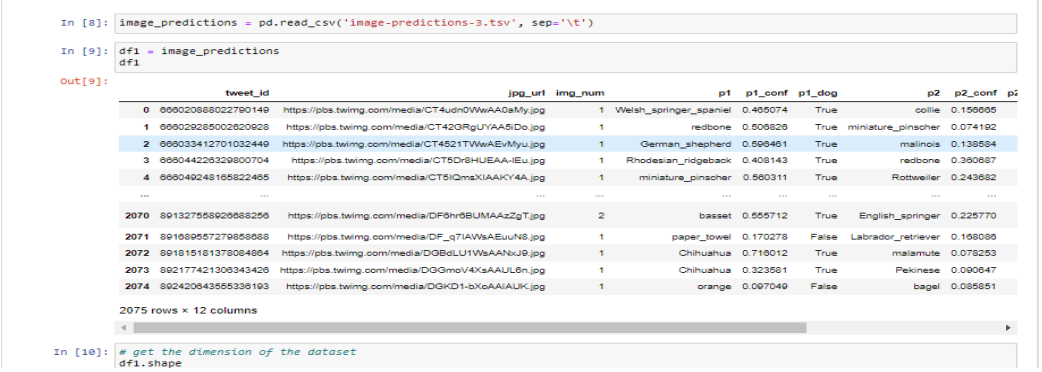
Step 1: Data gathering using the provided datasets

The three datasets (Twitter Archive, JSON file and Image predictions) for this project were loaded in the notebook using Python libraries as shown below:



```
In [1]: # import the basic libraries
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        %matplotlib inline

In [2]: # load the data set
        df = pd.read_csv('twitter-archive-enhanced-2.csv')

In [3]: # get the first five rows of the data set
        df.head()

Out[3]:
```

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	retweeted_status_id	retweeted_
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="http://twitter.com/download/iphone" r...	This is Phineas. He's a mystical boy. Only eve...	NaN	
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="http://twitter.com/download/iphone" r...	This is Tilly. She's just checking pup on you....	NaN	
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="http://twitter.com/download/iphone" r...	This is Archie. He is a rare Norwegian Pouncin...	NaN	
3	891689557279858688	NaN	NaN	2017-07-30	<a	This is Darla. She	NaN	

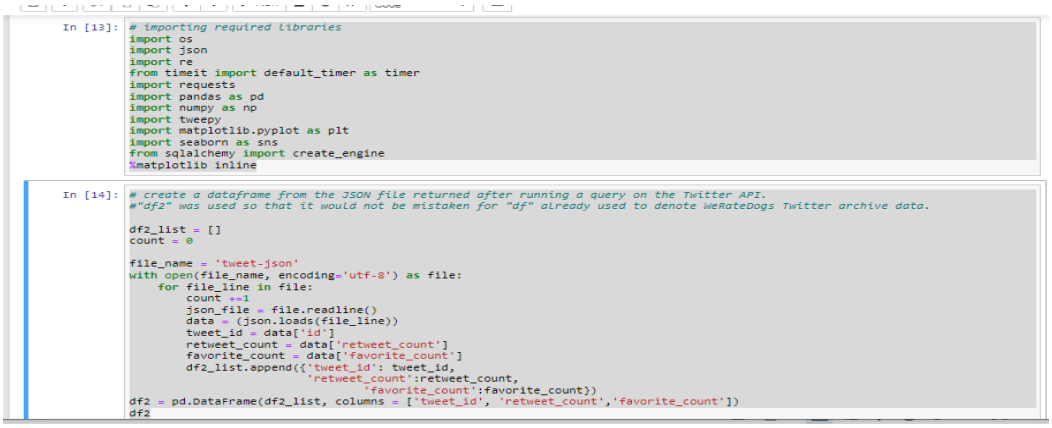


```
In [8]: image_predictions = pd.read_csv('image-predictions-3.tsv', sep='\t')

In [9]: df1 = image_predictions
        df1

Out[9]:
```

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springer_spaniel	0.465074	True	collie	0.156665	
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	redbone	0.506826	True	miniature_pinscher	0.074192	
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_shepherd	0.596461	True	malinois	0.138584	
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_ridgeback	0.408143	True	redbone	0.360687	
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_pinscher	0.560311	True	Rottweiler	0.243682	
...	...	...	...	...	...	...	...	...	...
2070	891327558926688256	https://pbs.twimg.com/media/DF6hr6BUMAAzZgT.jpg	2	basset	0.555712	True	English_springer	0.225770	
2071	891689557279858688	https://pbs.twimg.com/media/DF_q7IAWsAEuuN8.jpg	1	paper_towel	0.170278	False	Labrador_retriever	0.168086	
2072	891815181378084864	https://pbs.twimg.com/media/DGBdLU1WsAANxJ9.jpg	1	Chihuahua	0.716012	True	malamute	0.078253	
2073	892177421306343426	https://pbs.twimg.com/media/DGGmoV4XsAAUL6n.jpg	1	Chihuahua	0.323581	True	Pekinese	0.090647	
2074	892420643555336193	https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg	1	orange	0.097049	False	bagel	0.085851	

2075 rows × 12 columns

```
In [10]: # get the dimension of the dataset
         df1.shape
```



```
In [13]: # importing required libraries
import os
import json
import re
from timeit import default_timer as timer
import requests
import pandas as pd
import numpy as np
import tweepy
import matplotlib.pyplot as plt
import seaborn as sns
from sqlalchemy import create_engine
%matplotlib inline
```

```
In [14]: # create a dataframe from the JSON file returned after running a query on the Twitter API.
#"df2" was used so that it would not be mistaken for "df" already used to denote WeRateDogs Twitter archive data.

df2_list = []
count = 0

file_name = 'tweet-json'
with open(file_name, encoding='utf-8') as file:
    for file_line in file:
        count +=1
        json_file = file.readline()
        data = (json.loads(file_line))
        tweet_id = data['id']
        retweet_count = data['retweet_count']
        favorite_count = data['favorite_count']
        df2_list.append({'tweet_id': tweet_id,
                         'retweet_count':retweet_count,
                         'favorite_count':favorite_count})
df2 = pd.DataFrame(df2_list, columns = ['tweet_id', 'retweet_count','favorite_count'])
df2
```

The dimensions of each of the datasets were gotten to give information about individual datasets.

Step 2: Assessing and cleaning the datasets

In this section, the datasets were investigated for data quality and tidiness (structural) issues. Each of these were sorted and code applied to clean/sort the issues. I documented eight (8) quality issues and two (2) tidiness issues which were:

Quality issues

1. The Dog names were not consistent in the written form. Some of the first letter of the dog names were written in upper case while others were written in lower case. All the names were cleaned to be first letter capital.
2. Since it is only original ratings (and not retweets) that have images that are needed for the wrangling exercise, retweets columns would be dropped. The 'retweeted_status_id retweeted_status_user_id retweeted_status_timestamp' columns in the Twitter Archive are not required and were dropped.
3. The 'source' column of the twitter archive dataset can be used to extract the type of device used to post on twitter. If it is twitter by iphone, web, or other means.
4. The dog numenator ratings were not equal or less than the denominator (10) to be a proper fraction. There are many instances of higher than 10 numerator ratings. These are improper fractions.
5. There are "False" prediction in the dog, which is not possible, since it is a dog dataset that is under review and supposed to be "True" prediction/statement.
6. There were missing values in the twitter archive dataset columns. The columns (doggo, floofer, pupper, puppo) were given wrong representation. Represented as "None" instead of "NaN".

7. The values in ''tweet_id', in_reply_to_status_id', 'in_reply_to_user_id' columns were represented as 'float' data type, instead of 'string (object) '.
8. The 'in_reply_to_status_id', 'in_reply_to_user_id' columns in the twitter archive dataset have too many with Null value / 'NaN'.

Tidiness issue

1. The Dog categories (doggo, floofer, pupper, puppo) are categorical variables which should be in a single column, instead of the four columns they were shown. Being in a single column makes the representation neater and straight forward.
2. There were three datasets which were to be accessed at the initial stage, out of which two datasets were later joined to form a Table. This should have been mergerd into a single dataset from the onset for students to analyse. This led to repeated tasks that would have been focused on a single dataset from the begining.

Step 3: Storing the data

The three (3) datasets were merged into a single dataset. This was stored as Twitter_archive_master.csv.

Step 4 and 5: Conducting exploratory data analysis, creating visualizations and Insights

Visualization were done after the exploration. The insights from the analysis were:
1. Dog Posts made with iPhone (Twitter for iPhone) had the highest favourite counts (35,088,047,291).
2. The Dog Type, Pupper had the highest favourite counts (3,620,860,858)
3.The minimum and maximun numerator dog ratings were 1 and 165 respectively. The minimum and maximun denominator dog ratings were 2 and 150 respectively. The ratings many times had greater numerators then denominators. This unique rating system is a big part of the popularity of WeRateDogs.