Introduction

In this project I am trying to wrangle WeRateDogs Twitter data, passing by the processes of data wrangling and then create a clear visualization , Using Python and its libraries, I gathered data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. The dataset that I am wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. You can see the work I did in wrangling in wrangle_act.ipynb

I divide the project into the following :

- Gathering data
- Assessing data
- Cleaning data
- Storing, Analyzing, and Visualizing
- Limitations
- Conclusion

After I import the required libraries and functions such as :

- pandas
- NumPy
- requests
- tweepy
- json

Gathering Data

I Gather the data as required from three sources :

- **1st source : Twitter Archive**

The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students. WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

- **2nd Source : The tweet image predictions**

The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

- **3rd Source : Twitter API**

Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, I query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data written to its own line. Then I read this .txt file line by line into a pandas DataFrame .

I requested twitter to set up my own Twitter applications so I access the Twitter API in order to complete a Data Wrangling project using Tweepy to query Twitter's AP for data included in the WeRateDogs Twitter archive. I received a Success message that proved my new Twitter developer account , and then I used the Keys I need it in the following codes.

Assessing Data

After gathering each of the above pieces of data, I assess them visually and programmatically for quality and tidiness issues.

- ***You only want original ratings (no retweets) that have images ,Though there are 5000+ tweets in the dataset, not all are dog ratings and Some are retweets.***
- ***The requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.***

Quality

A) Twitter Archieved (TAE)

1. Remove retweet and replies datasets
2. Change the variables datatypes :
   - tweet_id ____ strings instead of integer
   - in_reply_to_status_id ____ strings instead of float.
   - in_reply_to_user_id _ strings instead of float.
   - retweeted_status_id ___ strings instead of float.
   - retweeted_status_user_id ____ strings instead of float.
   - retweeted_status_timestamp ___ datetime instead of object (string).
3. Keep 'name' column with values that start with capital words
4. Change any values that 'Null' to 'None'
5. Delete the link from the text column

B) The tweet image predictions (TIP)

1. drop the jpg_url that are duplicate.
   - tweet_id, strings instead of integer
2. change underscore to space in p1,p2 and p3
3. make all the values in p1,p2,p3 is starting with capital letter
4. Rename coloumns
5. Change any values that 'Nan' to 'None'
6. The names for the heading columns isn't clear , I will change
   - p1 to "First prediction"
   - p2 to "Second prediction"
   - p3 to "Third prediction"

- p1_conf to "First confidence prediction"
- p2_conf to "Second confidence prediction"
- p3_conf to "Third confidence prediction"

7. Change any values that 'Nan' to 'None'

C) Twitter API (TAP)

1. Change the variables datatypes in Twitter API data
   - Tweet_id ____ strings instead of integer
   - Date_time ____ datetime instead of object (string).

Tidiness

1. Merge Dog types in one column insted of four columns (doggo, floofer, pupper, puppo), no need for that .
2. Merge TAE,TIP which is first and second dataframes(Twitter Archieved + tweet image predictions)to be T12 based on ID.
3. Combine the datasets together T12 and TAP (Twitter Archieved + The tweet image predictions + Twitter API) based on ID.

Cleaning Data

I Cleaned each of the issues I documented while assessing. I Performed this cleaning in wrangle_act.ipynb as well. After the last cleaning task which is merge the datasets together, we can move to the next step which is storing , analyzing and visualization step.

Storing, Analysing, and Visualizing

I store the data in csv file which is 'twitter_archive_master.csv' then I do some analysis and visualization which is present in the wrangle_act.ipynb and the other report .

Limitations

I analyzed the dataset but some columns I couldn’t understand it very well because I didn’t collect the data by myself such as image predictition data , Also some skills I couldn’t be perfect with because I didn't use it before in analyzing data. Also Documentation was not clear for me because it is not famous in my country to adopt a dog , not many people in my country having a dog but I tried to understand as much as I can

Conclusions

The dataset that I was wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.
WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.
If you are looking to adopt a dog , Most popular type is pupper.

By Engr.Ahmed Alshehri
Alshehriahmedh@gmail.com
Udacity student