DATA ANALYST NANODEGREE

Project # 4

DATA WRANGLING

(@WeRateDogs)

(Report 1- Wrangle report)

Himanshu Saini
Delhi-NCR, India

Introduction

Using Python and its libraries, you will gather 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 I will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**

Project details

The tasks of this project are as follows:

- Data wrangling, which consists of:
  - Gathering data
  - Assessing data
  - Cleaning data
- Visualization
- Reporting on
  - our data wrangling efforts (Report 1- **Wrangle report**)
  - our data analyses and visualizations (Report 2- **act-report**)

Gathering data:

The data for this project consist on three different dataset that were obtained as following:

1. The twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.

2. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information

3. By using twitter developer account got an access to download @dog_rates Twitter archive.
   Then I query the Twitter API for each tweet's JSON data using Python tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file.

Assessing data:

After the data was gathered, Dataframes consists:

Twitter archive file as twitter_df
Shape: (2356, 17)

The tweet image predictions as image_pred
Shape: (2075, 12)

Twitter API & JSON as tweet_api
Shape: (2331, 4)

After visual and programmatic assessments of datasets I have come up with following quality and tidiness issues:

Quality Assessment.

Twitter archive data:

1) Timestamp is an 'object' type.

2) Looking programmatically, some names are inaccurate such as "a", "an", "the", "very", "by", etc.

3) Name has values that are the string "None" instead of NaN

4) In 2365 only 23 cases where the denominator of rating is not equal to 10. These entries will be removed.

5) Calculating Ratings of the Dog

6) There is no duplicated tweetids found.

Image Predication Data:

1) The "p1" and "p1_conf" columns will be renamed with more explanatory titles.(i mage Predications)

2) Drop 66 jpg_url duplicated

3) There is no duplicated tweet ids found in Data set.

Tweet API _Json Data:

1) There is no duplicated tweet ids found in Data set.

Tidiness Assessments

1) Change columns "doggo", "floofer", "pupper", and "puppo" from wide to long format.

2) Have to extract the url from text column.

After all these assessment :

- Drop columns that won't be used for analysis in al dataset
- Merge all the data into "twitter_archive_master.csv"

Cleaning Data:

- First and very helpful step was to create a copy of the three original data frames. I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original.

- Other interesting cleaning code was to melt the dog stages in one column instead of four columns as original presented in twitter archive.

.

Conclusion

Finally cleaning above quality and tidiness issues, twitter_archive_master.csv is the combined and cleaned data which consists (1976, 11) Rows and Columns.

Sources:

https://stackoverflow.com/questions/28384588/twitter-api-get-tweets-with-specific-id

https://www.tutorialspoint.com/python_text_processing/python_extract_url_from_text.htm

https://jakevdp.github.io/PythonDataScienceHandbook/03.07-merge-and-join.html

https://knowledge.udacity.com/questions/45245