Wrangle_Report

Introduction

Data Wrangling is the process that we gather data from a different kind of sources and in a variety of formats, assess its quality and tidiness, then clean it. In this project, I will use the skills that I have learned in the data wrangling class to data wrangling the dataset from Twitter. The dataset that I use is the tweet archive of Twitter user@dog_rates, which known as WeRateDogs.

In this dataset, it includes the data where the WeRateDogs(Twitter account) people's dog with a humorous comment about the dog. This archive contains some basic tweet data, such as tweet ID, text, timestamp and text for all 5000 + of the tweets since August 1, 2017.

Project details:

The report will contain the following

- Data wrangling
  - Gathering data (WeRateDogs Twitter archive)
  - Assessing data (Detect and document **eight (8) quality issues** and **two (2) tidiness issues)**
  - Cleaning data
- Storing, analyzing, and visualizing my wrangled data (3 insights and a least 1 visualization)
- Reporting on my data wrangling efforts and my data analyses and visualizations

The libraries:
pandas, Numpy, tweepy, json will be used to analysis the data.

Gathering Data:

This project contains three dataset and they are listed below:

1) Twitter archive File : twitter_archive enhanced.csv file that was provided by Udacity and downloaded manually.
2) Image_predictions.tsv: a file that is hosted on Udacity's servers and can be downloaded programmatically using the Requests library.
3) tweet_json.txt : The pandas dataframe will store the tweet ID, retweet count, and favorite count in this txt file and later we will read it line by line.To be more specific, each tweet's JSON data will be written to its own line. The tweet IDS in the weRateDogs twitter archive will also be used to archive, query its Twitter API for each tweet's JSON dara using the Pythons's tweepy library and then store each of these tweet's entire set of JSON data in tweet_json.txt

Accessing Data:

First, I checked the data in these three datasets visually. I printed the details information in these three dataset in the Juypter notebook. I checked theses data by using the method, such as info, value counts, duplicated values as well as sample etc. So that way I can have a basic ideas on how these dataset look like and find out what I want to do in the Data Cleaning Steps. This process also helps me to look for the quality and tidiness issues.

Cleaning Data:

For the first dataset twitter_archive , I created a copy of it and then deleted the retweets by filtering any retweeted_status_user_id that contains NaN. Then I print it and count the values.

After that, I deleted the columns in clean_twitter_archive that I don't need for the data analysis process, such as
'source',
'in_reply_to_status_id',
'in_reply_to_user_id',
'retweeted_status_id',
'retweeted_status_user_id',
'retweeted_status_timestamp',
'expanded_urls'

and printed out the list and check the header of the columns.

Then I considered to make my timestamp data look cleaner so I separated the timestamp in today, month and year. And then convert the 'rating_numerator and 'rating_denominator" to float type. Then update the numerators and check them.

And then, I separated the dog types to doggo, floofer, pupper and puppo. I updated the denominator and numerators in the dataset as there are five tweets their denominator are not equal to 10. I also considered to delete five tweets that do not have act ratings. Then convert all integer in rating to float.

For the second dataset, I made a copy of the dataset tweet_json, checked its info. Then counted how many "Original tweet" in the "retweeted_status" are.

Then I made a copy of the image_prediction dataset and then dropped any jpg_ url that are duplicated, Then double checked to make sure any duplicated jpg_url are dropped. After that, I created the lists to store the first prediction and confidence level. The first prediction means the first true that we found in the data for the type of dog. As the original data in the table contins 3 predictions and confidence levels. Doing this filtering, I could create a column for dog type and a column for confidence level. Then I also dropped the type of dog that contains “unknown” as these are not very helpful for us.

After this, I dropped any columns that I don’t need for the analysis in the clean_image_prediction dataset.

'p1',
'p1_conf',
'p1_dog',
'p2',
'p2_conf',
'p2_dog',
'p3',
'p3_conf',
'p3_dog'

The two tideness issues that I found which needed to be fixed were convert the tweet_id in the clean_tweet_json dataset to integer and then merge them with the other table. Then we would merge them into one dataset so that it could be easy to look at the information later.

Conclusion

In this project, I have learned a lot in Data Wrangling and use Python programming languages to obtain the information I need. I have learned how to gather data via JSON, assess data with different libraries and clean the data. These are the useful tools that data analyst use when they need to data wrangle tables.