WRANGLING REPORT

Authored by: Abhishek Pandey

.com/abhishekpandeyit https://.com/PandeyJii_

INTRODUCTION

This project contains the wrangling and analysis or Archived chats of Twitter. The dataset that I will be 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. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 6 million followers and has received international media coverage.

The Analysis of my visualisation can be found in the **act_report** file.

DATA GATHERING

Data is successfully gathered From at least the three different sources and in three differ ent file formats on the Project Details page.

- Each piece of data is stored into a separate pandas DataFrame.

Following work has been done during gathering data:

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

- **Twitter archive file:** The twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.
- **The tweet image predictions**,: The file image_predictions.tsv was hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information
- **Twitter API & JSON:** Using the tweet IDs, keys and Tokens in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored the JSON data in a file called tweet_json.txt file. I read this .txt file into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

DATA ASSESING:

Once the three tables were obtained I assessed the data as following:

- Visually, I used two tools. One was by printing the three entire dataframes separate in Jupyter Notebook and two by checking the csv file0s in Excel.
- After the data was gathered, assessment was performed Programmatically, by using different methods
  - .head()
  - .sample()
  - .info()
  - .value_counts()
- Then I separated the issues encountered in quality issues and tidiness issues. Key points to keep in mind for this process was that original ratings with images Were wanted.

Tidiness Issue.

WRANGLING REPORT

Authored by: Abhishek Pandey

.com/abhishekpandeyit https://.com/PandeyJii_

- Combining all dataframes together as they all contained information about the same tweets
- Combining 4 variables about dog type into 1 column “dog_stage”

Quality Issue:

- Data contained retweets
- Tweet id was the incorrect data type
- Timestamp was the incorrect datatype
- Name contained the string “None” instead of a NaN
- Name contained various inaccuracies which were regular lowercase words
- The name O’Malley was incorrectly extracted as “O”
- Rating numerators which contained decimals were incorrected exported
- Ratings are unstandardized
- Undesired columns present

DATA CLEANING

- First and very helpful step was to create a copy of the three original dataframes. I wrote the codes to manipulate the copies.
- There were a couple of cleaning steps that were very challenging. One of them was in the image prediction table.
- Other cleaning code was to melt the dog stages in one column instead of four columns as original presented in twitter archive.
- One very challenging cleaning step was when I had to correct some numerators that were actual decimals. This issue was brought to my attention after the first Udacity review. Using Excel and visual assessment was not sufficient to verify those decimals. Therefore, I had to run a code in order to check those actual tweets (decimals numerators).
- I used following methods/functions of python libraries for data cleaning.

  1 merge(), 2 reduce(), 3 .extract(), 4 .drop() ,5 .isnan , 6 .astype(),

  7. .to_datetime(), 8 .islower(), 9 .replace(), 10 .rename(), 11 set_option(),

  12 .loc[], 13 value_counts(), 14 .info(), 15 .head(), 16 Loops, 17 Regular expressions .

CONCLUSION

Data wrangling is a core skill that whoever handles data should be familiar with. I used Python and its some Libraries. The advantages of this tool :

• For gathering data there are several packages that help scraping data off the web, that help using APIs to collect data (Tweepy for Twitter) or to communicate with SQL databases.

• It is strong in dealing with huge data (compared to Excel). It can also deal with a large variety of unstructured data like JSON or structured data like ERP/SQL databases.