UDACITY PROJECT
WRANGLE AND ANALYZE DATA

18.06.2020

OVERVIEW

1. Project Background and Description

Real-world data rarely comes clean. 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. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that you 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 4 million followers and has received international media coverage.

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. More on this soon.

2. Software Requirements

The entirety of this project can be completed inside the Udacity classroom on the **Project Workspace: Complete and Submit Project** page using the Jupyter Notebook provided there. (Note: This Project Workspace may not be available in all versions of this project, in which case you should follow the directions below.)

If you want to work outside of the Udacity classroom, the following software requirements apply:

- You need to be able to work in a Jupyter Notebook on your computer. Please revisit our Jupyter Notebook and Anaconda tutorials earlier in the Nanodegree program for installation instructions.
- The following packages (libraries) need to be installed. You can install these packages via conda or pip. Please revisit our Anaconda tutorial earlier in the Nanodegree program for package installation instructions.
    - pandas
    - NumPy
    - requests
    - tweepy
    - json
- You need to be able to create written documents that contain images and you need to be able to export these documents as PDF files. This task can be done in a Jupyter Notebook, but you might prefer to use a word processor like Google Docs, which is free, or Microsoft Word.
- A text editor, like Sublime, which is free, will be useful but is not required

3. Wrangle Report

The tweet archive of Twitter user @dog rates, also known as WeRateDogs, is the dataset wrangled within the project. WeRateDogs is a Twitter account rating dogs of people with a humorous comment about the dog.
The aims of the WeRateDogs Twitter project included:

• Wrangling the twitter data in the following processes:

- Gather Data
- Assess Data
- Clean Data

• Store, analyze and visualize the data you have wrangled
• Data wrangling documentation and data interpretation and visualization

GATHERING DATA:

Data was collected from 3 different sources:

1) The twitter archive file was supplied and downloaded manually. The file contains different variables for each tweet, including tweet Id, timestamp, text, rating numerator and denominator, name, etc.

2) Additional data were obtained using the Twitter API, including favorite count and retweet count.

3) The tweet image predictions file was programmatically downloaded from Udacity 's servers using the Requests library.

ASSESSING DATA:

After the data was gathered, assessment was performed using the following methods:

- info()
- value_counts()
- head()
- sample()

Quality Issues

Quality issues that were cleaned:

1. Remove Retweets from our data.
2. Delete columns that no longer needed
3. Change 'tweet_id' from an integer to a string
4. Change the timestamp to correct datetime format
5. Issues with names correcting them.
6. Dog ratings inaccurate
7. Unstandardized ratings

Tidiness Issues

• Combining all dataframes together as they all contained information about the same tweets

twitter_archive:

- The last four columns all relate to the same variable (dogoo, floofer, pupper, puppo)

Images_prediction:

- this data set is part of the same observational unit as the data in the archive - one table with all basic information about the dog ratings

twitter:

- this data set is also part of the same observational unit - one table with all basic information about the dog ratings

CLEANING DATA:

After the assessment, I cleaned the data through the following means:
Define, Code and Test
The issues found during the assessment process were cleaned and tested using the following methods and techniques:

- merge()
- reduce()
- extract()
- drop()
- isnan
- astype()
- to_datetime()
