Udacity Nano Degree project 5

Wrangle and Analyze Data

Wrangling Report

Introduction

This project requires us to use python to gather data from a variety of sources and in a variety of formats such as(csv, tsv, json etc) and the assess its quality and tidiness, then clean it. This is called data wrangling. I documented my wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python and its libraries

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. There are 3 steps involved in the wrangling processes including: data gathering, data accessing and data cleaning,

STEP1 Data Gathering

The WeRateDogs Twitter archive which consists of 5000+ tweets I download this file manually by clicking the link provided by Udacity and get the following data “**twitter_archive_enhanced.csv”.**

The tweet image predictions which shows the prediction of top3 breed of dog (in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and I downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Twitter API contains each tweet's retweet count and favorite ("like") count which are not shown in the archive data. I used the tweet IDs in the WeRateDogs Twitter archive, 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 should be written to its own line. Then I read this .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

STEP 2 Data Assessing

In this step, I try to identify quality and tidiness issues of the dataset provided.
So low-quality data are considered as dirty data which has incompleteness and inaccuracy issues.
And untidy data are regarded as messy data with structural issues

I used two types of assessment:

- Visual assessment which I just review them through Google Sheets to see the overall of the data
- Programmatic assessment which I use the method such as describe, sample, info, value_counts etc. from pandas library to check the data

Here are the notes and issues I identified during this process:

Quality Issues:

- Tweet_id, timestamp, sources, img_num and dog_stages are not in the right data type
- Remove redundant word for source definition
- Some dog's name is not right
- 183 retweets needs to be deleted
- Some denominators are not equal 10
- Some numerators are less than 10
- Some tweets don't include images
- There is inconsistency regarding letter cases for some breeds in p1, p2, and p3 from Image Prediction File

Tidiness Issues:

- Merge the three dataframes into one by using tweet_id
- Combine 4 dog stages into a single column

STEP3 Data Cleaning

After the assessing data and issues I found, I use the programmatic method(python code) to clean the data so that it could be analyzed in a succinct and clean way. The step consists 3 substeps: 1) define the issue with own language; 2) write the code based on our definition; 3) test the data sets and make sure it is cleaned through previous step.

Conclusion:

Finally, I get following documents as the deliverables of my project:

`wrangle_act.ipynb:` code for data wrangling, analyzing and visuslizing.

`wrangle_report.pdf:` briefly describes your wrangling efforts.

`act_report.pdf` : communicates the insights and displays the visualization(s) produced from your wrangled data.