Wrangle Report

by Adaobi Onyeakagbu

Introduction

This report is in partial fulfilment of the Udacity Data Wrangling project for reporting my wrangling efforts. This project involves wrangling, analyzing and visualizing 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.

My Data Wrangling efforts include:

- Gathering data
- Assessing data
- Cleaning data

Gathering Data ¶

In this section I used the following processes for gathering the data that would be used for the wrangling and analysis:

1. The WeRateDogs Twitter archive file (twitter_archive_enhanced.csv) which was a csv file that I downloaded manually
2. Tweet image predictions file (image_predictions.tsv) which is hosted on Udacity servers and I downloaded programmatically using the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)
3. JSON data file which I got by querying the Twitter API I created to get access to each tweet's JSON data using Python's Tweepy library and storing them in the tweet_json.txt file. I only queried tweets available in the tweet archive. I then read this .txt file line by line into a pandas dataframe but saved only the tweet ID, favorite count, retweet count.

Assessing Data

In this section, I visually and programmatically assessed the data that I gathered, for further information and to see if there were any cleanliness and tidiness issues. The following are the processes and tools I used for assessing:

1. I used Pandas to read the files and visually assessed all three in the dataframe.
2. I investigated further details of the dataframes using info(), describe(), value_count(), duplicated(), and other programmatic functions.
3. After assessing, I found 10 issues with the cleanliness of the tables and 2 issues with the tidiness of the tables which I would then fix in the next section. These cleanliness and tidiness issues were:
    - Archive data contains some retweets
    - Expanded URLs missing for 50+ entries
    - Errors in rating entry (denominators should be 10 or not rating at all or another value was taken or error in ratings with decimals)
    - Missing data for dog stage classification
    - Incorrect names of dogs
    - Wrong datatype for Timestamp column, which also needs to be split into day, month and year
    - Unnecessary columns for analysis
    - Duplicate images
    - Some incorrect dog breeds from prediction (e.g the)
    - Unnecessary columns for analysis
    - Columns for dog stage classification are separated instead of in one
    - The three tables should be joined to one on tweet id

Cleaning Data

In this section, I defined the problem I wanted to address, programmatically fixed it and then tested it. The end goal of this whole process in the project was to get one table that has quality information for each non-duplicate tweet that contains an image. Hence these are the steps I took to achieve this:

1. First, making a copy of the tables helped to leave room for errors in the cleaning process. It didn't make changes to the main tables.
2. I had to use Manual cleaning for the some of the issues. This is in particular when correcting rating entries with decimals erronously entered, another value was taken and those that were not ratings at all. I discovered these error while programmatically assessing the ratings that were relatively too low and too high.
3. Most of the other cleaning processes were programmatic
4. Eventually after fixing the cleanliness issues, I made one column for dog stage and then combined the three tables to get a comprehensive table for each tweet. I also drop rows that do not have images and JSON data.

Conclusion

In the wrangling, I was able to use test my knowledge in dealing with APIs and also leverage the most common cleaning functions and methods in the pandas library to clean the data which I gathered.