Wrangle Report

Udacity Data Analyst NanoDegree: Wrangle and Analyze Data Project

By: Bijay Shrestha

Introduction

The dataset that has been wrangled in this project is sourced from the twitter archive associated with tweets from various twitter users @dog_rates, also known as WeRateDogs. This dataset consists of 2356 tweets in total that range from Nov 2015 to Aug 2017. WeRateDogs is a twitter account that has received a huge amount of international media attention as users rate dogs with a humorous comment. After loading the dataset, it has been assessed and cleaned utilizing various analytics methodologies of pandas framework. Additionally, visualizations have been created which can be found in the act_report.pdf included in the submission.

Gathering Data

Data was gathered from three different sources:

- The first dataset was downloaded manually as twitter-archive-enhanced-2.csv from Udacity
- The second dataset was programmatically downloaded from Udacity's server using the requests function
    - url = "https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv"
    - response = requests.get(url)
- Additionally the third set of twitter data was downloaded querying the twitter API using tweepy library. The tweet_id, favorite_count, retweet_count were programmatically extracted from tweet_json.txt file.

Assessing Data

After the data was gathered, it was assessed using the following methods:

- .head()
- .info()
- .sample()
- .value_counts()
- .describe()

There were some quality and tidiness issues. Below are some of the findings:

Quality Issues

- need to clean the retweets data as we only want the original tweets
- has quite a few unnecessary columns, should be dropped for data readability; in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be should be dropped
- timestamp, retweeted_status_timestamp should be datetime instead of objects
- in several columns null objects are not null (NAN instead of None)
- name column has inconsistent values such as 'None', 'a', 'an', not', 'this', 'the', 'one', 'my', 'mad', 'all', 'quite', 'by'

- rating_numerator, rating_denominator have inconsistent values, denominator should be standardized with a value of 10
- contents of 'source' column should be in a more readable format to display only the source
- there are 5 numerator ratings that have extremely high values (>14 out of 10) which will skew the data ;the rows containing these values should be dropped.

Tidiness Issues

- need to combine four columns doggo, floofer, pupper, puppo into one as dog_stage
- need to join image predictions and twitter api data to twitter archive data

Cleaning Data

A new table was created to address these quality and tidiness issues. All three datasets were merged into this new table. Four columns were merged into one column named as ‘dog_stage’. Unnecessary columns were dropped for analytics purposes. Incorrect data types were converted into the correct ones. Inconsistent values in specific columns were either replaced or dropped using codes. Some of the contents were shortened to make them more readable. For each issue, cleaning was performed programmatically using three steps below:

- Define - The issue was defined
- Code - The issue was solved programmatically
- Test - The solution was tested programmatically

Some of the methods used for cleaning data are listed below:

- .merge()
- .astype()
- .islower()
- .to_datetime()
- .info()
- .extract()
- .head()
- .drop()

Storing data

After cleaning data it should be stored so that it is accessible for analysts to do some analytics. The cleaned data was stored as twitter_archive_master.csv file for analytical purposes.

Conclusion

The data we gather is never perfect and doesn’t always come from a single source. We need to collect it from various sources, combine and clean it before we do some valuable analysis on it. Data wrangling is one of the important steps in data analytics and professionals should be able to utilize various methodologies and libraries to perform the task.