Wrangle report – Udacity data analyst nanodegree

By Elad Shahar

Sept 2018

Intro:

I found the data wrangling project of the data analyst nanodegree the most challenging of the entire program. Extensive code writing and learning new libraries together with a wide variety of cleaning techniques I needed to research. It was a very hard but satisfying journey.

Gather stage

During the data gathering stage I gathered data from 3 distinct sources

1. Weratedogs twitter archive provided to udacity for the purpose of this project. The archive provides the basic tweet data including tweet id, name, text, classification of dog to dog type and timestamp.
2. Image predictions tsv file downloaded programmatically via Python request library
   The file holds the result of running the image of each dog via a convolutional neural network to identify the dogs breed.
3. Tweet JSON data downloaded via the Tweepy library for twitter API, this I used primarily for the retweet count and favorite count per each tweet.

Assess stage

Assessment stage started with a programmatic examination of the structure and data types of each of the three data frames as well as visual assessment.

1. The first area I focused on was tidiness issues, combining dog type columns into one column named dog_type (‘floofer’, ‘doggo’, ’pupper’ and ‘puppo’), matching the tweet_id index in naming among all data frames and casting it as a string. The last tidiness issue I addressed is how best merge together the three dataframes into one single dataframe inner joined by tweet id

2. The second area of focus was data quality

Erroneous datatypes which require re casting

Removing columns which we are not using

Removing retweets rows from the dataframe

Extracting gender from the text column

Correcting names where possible and assigning to none when not possible

Assessing rating numerator and denominator values and their relation. Enough issues there to lead me to creation of a new ratio column.

Assessing outlier values in the rating_ratio, correcting when possible and removing rows when impossible

Clean stage

3.The last step of the data wrangling process was the cleaning stage, each point to be cleaned clearly defined, cleaned and then tested.

The data type recasting and column/row removal were relatively straight forward, while the gender extraction and dog type conversion to a single column took a lot of research and trial and error.

Analysis and visualization

After cleaning the data I saved the master dataframe to a csv file and proceeded to add an analysis of the clean data and some visualizations.

To summarize, this was a very difficult but rewarding project, I have learned a lot in a short time and I am sure I will be using these skills extensively in the future.

Thank you for the great challenge!