We Rate Dogs Wrangle Report

A.J. Lozano

3/13/2019

The We Rate Dogs Project required gathering data from three different sources. The first source came from a csv file named twitter-archive-enhanced. I turned the csv into a data frame using Pandas. The second source was gathered programmatically via the requests library and the image-predictions tab separated file. The final source required importing tweepy and Twitter's API function. This return JSON data for the tweets. The required information was turned into a third data frame.

With the data frames in hand, I was able to start a visual and programmatic assessment of the data within each data frame. Eight quality issues and two tidiness issues need to be corrected per project guidelines.

Visual assessments on larger data frames are generally better using spreadsheet software. Regardless, I looked at each data frame in Python for issues with completeness, and consistency. The Twitter data frame had several issues that could be seen visually. Nan values were a common theme within many of the columns and rows. The name column was missing data. There is a tidiness issue regarding dog classifications listed in four different columns. The classifications would be better off condensed and within one column. The image data frame had potential issues with picture classification. This could be better assessed programmatically. The API data frame did not appear to have any issues at first glance.

After both types of assessments, I was able to compile a list of potential fixes. For example, some of the data formats were incorrect and needed to be changed. Denominators were not all a value of 10 and needed correction. Several dog names were invalid. I looked for names like a, an, old, the and worked on correction. Names were not all uniform. I used the Title function to correct name issues. I merged several columns into one condensed easier to read column.

Each and every issue's inconsistency was identified, coding to make the change was implemented and finally tested to ensure corrections were completed.

The three separate data frames were corrected, condensed and finally joined on the primary key(tweet_id) to create one final clean data frame.