wrangle_report

April 23, 2019

0.1 Tweet Archive of " We Rate Dogs" Twitter Account

0.1.1 Wrangling Report

Reem Alashhab In this project, I went through the three phases of data wrangling:

Gathering Data

In this phase, I gathered the data of three separate data sets. For the Twitter Archive Data set, I imported the file directly from the project directory that was added after downloading the file from Udacity. Next, I downloaded the file for Image Predictions Data Set using Requests Library. Lastly, for the Additional Twitter Data Data Set, I used the twitter API to import data directly to the project work space. For not found tweets, I loaded their id numbers to a text file.

Assessing Data

After gathering the required data, I loaded each data set in a separate data frame to assess each of them and inspect quality and tidiness issues. Quality issues are related to the content issues in the data sets. And tidiness issues are related to the structural issues found in these data sets. I assessed the three data sets (Twitter Archive Data set, Image Predections Data Set, Additional Twitter Data Data Set) using pandas functions such as info(), describe(), head(), tail(), sample(), query(), value_counts(), duplicated(), sort_values(). I found many quality issues such as the wrong dog names (ex: a, the , one, his) in Twitter Archive Data Set, the wrong data type for tweet id in all the three data sets, and the 66 duplicates in jpg_url column in Image Predictions Data Set. For the tidness issues, I found seprate four dog stages columns in Twitter Archive Data set that need to be combined to only one column. In addition, I found the created_at column that includes useful datetime elements such as day, month, year that is best to extract to separate them for further analysis.

It's important to mention that I needed an external work space to view clearly some data that I was not able to view it fully such as the tweet text. I used the excel sheet as it helped me to find tweets that include two dog ratings, and discover the problem of the different ways of writing a dog name. It helped me to also found some tweets that don't include the dog name nor the dog stage.

Cleaning Data

Before going deep in the cleaning process of data, I dropped un needed columns and rows from each data set to make it faster to clean the remaining columns and rows. For outliers found, I gathered them in a list, searched for them using for loop , and then dropped them using the drop function from pandas library (ex: rating_numerator, rating_denominator). For wrong values, I used the loc function to locate them and then fix them such as the dog_name, and dog_stage. To convert wrong data types , I used the astype function to convert them to the right type. For rows with none values, I converted them to NaN using np.nan from numpy library. For tidiness issues such as the four seprate dog stages, I used combine_first function to combine them into a one column , and then dropped them using the drop fucntion. In the final step of the cleaning phase, I merged the three data sets in one master data frame and then save it in a csv file for further investigation and analysis.

For the full information about each data wrangling phase, please refer to wrangle_act file. It's important to mention that there are still some quality and tidiness issues that are unfixed due to the project time limitation.