Data Wrangling Report

By: Gregory Gardner

Introduction

The dataset that was wrangled is the WeRateDogs Twitter archive which contains 2,356 basic tweet data for From November, 2015 thru August, 2017. This project required us to gather the appropriate data, asses the gathered data, then clean the gathered data based on our assessments. Finally, we stored, analyzed, and visualized our wrangled data.

Gathering Data

We gathered data from three different sources in three different manners. First, we manually downloaded a .csv file containing archived tweets that was located on Udacity's server and saved it to the same folder as our jupyter workspace. The data was then imported into our jupyter notebook with pandas read_csv. Secondly, we programmatically downloaded a complimentary data set called image-predictions.tsv which contained tweet image predictions from an https hosted on Udacity's servers and saved it to the main file directory where we were then able to import it into our jupyter notebook. Finally, we extracted data directly from Twitter's API using python's tweepy library. Using the tweet id's from our archived tweets, we were able to query the Twitter API for each tweet's JSON data and stored the entire JSON data set in a .txt file. We then imported the JSON objects from the .txt file into a list of dictionaries which we then used to create our third data frame.

Assessing Data

First, we performed a visual assessment using .head() on all three data frames. Next, we performed a programmatic assessment on all three data frames using .info(), value_counts(), and sort_values(). After performing these assessments on the data frames we discovered both quality and tidiness issues that would need to be addresses before we could properly analyze and visualize the data. The following tidiness issues were found:

> The 'id' column in the third data frame was differently named than the other two data frames.

All three data frames needed to be joined into a single data frame.

The dog stages are values in the column names and the four columns needed to be condensed into a single column filled with the dog stages values (doggo, pupper, puppo, floof(er).

The following quality issues were found and ultimately addressed:

The timestamp column was in the wrong format and needed to be changed to date time format.

The name column had many erroneous values to fix.

The values in the ratings_denominator column needed to all equal 10.

Certain columns that were not pertinent to our analysis needed to be removed.

The tweet id columns needed to be changed from integers to strings.

The values in columns p1, p2, and p3 needed to be uniform in style and changed to begin with a capital letter.

Cleaning Data

After we assessed our data for quality and tidy issues we began cleaning our data both visually and programmatically. This process entailed first defining the issue and the proper steps needed to fix the issue. Then we would run code that would fix the issue. Then finally we would test each issue to ensure that any issues or problems were fixed. There were a few instances that occurred where in the process of cleaning and testing, we discovered additional issues which we would then have to address.

Storing, Analyzing and Visualizing

Finally, we stored our cleaned data to a .csv file. Then we were able to make some analysis and conclusions by accessing our clean data and plotting our data so that we could visually make better conclusions.