Wrangle Report

by Javier Soto

Introduction

The purpose of this project is to practice data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset wrangled is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

This report briefly describes my wrangling efforts:

- Gathering data
- Assessing data
- Cleaning data

Gathering data

The data for this project comprises three different datasets that were obtained per the following:

- **Twitter archive file:** the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.
- **The tweet image predictions**: this file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information
- **Twitter API & JSON:** using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe.

Assessing data

Once the three data files were generated, I programmatically assessed the data using python code and jupyter notebooks. I also manually inspected the .csv files using Excel.

Then I listed the quality and tidiness issues and addressed them step-by-step.

Cleaning data

This part of the data wrangling was divided into three parts: Define, code and test the code. These three steps were executed on each of the issues described in the assess section (this included ‘melting’ the dog stages into one column instead of four columns as originally presented in the twitter archive file).

Conclusion

Using the data analysis skills, I learned at udavity.com, I was able to wrangle the WeRateDogs Twitter data to create an interesting and trustworthy analysis and visualization. Although the data in the twitter archive file was good, it was not robust enough for a full analysis and only contained very basic tweet information. Additional gathering, with assessing and cleaning was required.