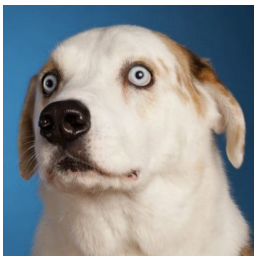
Wrangle Report: WeRateDogs

By: Allison Darrington



Introduction

Real world data is not clean. Unclean data consists of quality issues like missing data, incorrect data types, invalid records or inconsistent data. Data can also be untidy. Untidy data has structural issues like multiple variables in a column. Before data can be analyzed, it must be gathered, assessed, then cleaned. Clean data results in higher quality and more accurate analysis'.

This project will be analyzing data from the Twitter account, WeRateDogs. Data is collected from multiple sources, then cleaned and assessed. Once the data is clean and tidy, an analysis will be performed.

The Data

Data for this project will be collected from 3 sources. The data will be gathered, cleaned and analyzed with Python.

Enhanced Twitter Archive:

An archive of tweets from WeRateDogs as they stood on August 1, 2017. This contains basic tweet data (name of dog, rating etc) from 5000+ tweets. This does not include retweets or favorites.

Additional Data via the Twitter API:

This data fills in some gaps of data we didn't get from the Twitter Archive. Using Twitter's API, retweets and tweet counts for each WeRateDog tweet can be added.

Image Prediction File:

Every image in the Enhanced Twitter Archive was run through a neutral network that can classify breeds of dogs. This results in a data set full of breed predictions alongside tweet ID and image URL.

Gathering Data

The Enhanced Archived Twitter was provided by Udacity. This file is in .csv format. The file was downloaded and saved to the Udacity Project Workspace. This file was read using the pandas read_csv function.

The Image Prediction File was also provided by Udacity which is hosted on their servers. It was downloaded programmatically. The URL was used with the Requests library, and then read with pandas read_csv function.

Twitter API & JSON File was created using Twitters' API and Python's Tweepy Library. Using the Tweet IDs from the archive, the Twitter API was queried for each tweet's JSON data using Tweepy. The Data was stored in a JSON.txt file. The Tweet ID, retweets and favorite information was extracted line by line and put into a data frame.

Assessing Data

The data was visibly and programmatically accessed using Jupyter Notebook.

Visual Assessment consists of becoming acquainted with the data set by viewing a portion of its contents with .head function. Columns are inspected for potential entry errors.

Programmatic Assessment consists of using code to dive deeper into the data. Completeness, validity, accuracy and consistency are the main data quality dimensions. Many functions are used in the Programmatic Assessment including .head, .tail, .sample, .info, .describe. Value_counts, .duplicated, .isnull.

Quality and tidiness issues were documented in markdown cells within the Jupyter Notebook.

Cleaning Data

The cleaning data process first begins by making a copy of the original dataset. Each data issue will follow the three step process of Define, Code and Test.

Quality issues like wrong data types, incorrect names and incorrect ratings were corrected. Tidiness issues like too many columns for one variable, and separating the data frames into two separate ones to comply with the rules of tidy data.

Conclusion

Data Wrangling is a challenging but rewarding skill for all data analysts. Since much of the world's data is not clean, it is an important skill to hone. Data wrangling is known to take up the majority of a data analysts' time, but without proper wrangling, assessing and cleaning, the insights and conclusions we gather could be incorrect. Therefore, data analysts must remain diligent and thorough when wrangling, assessing and cleaning data.