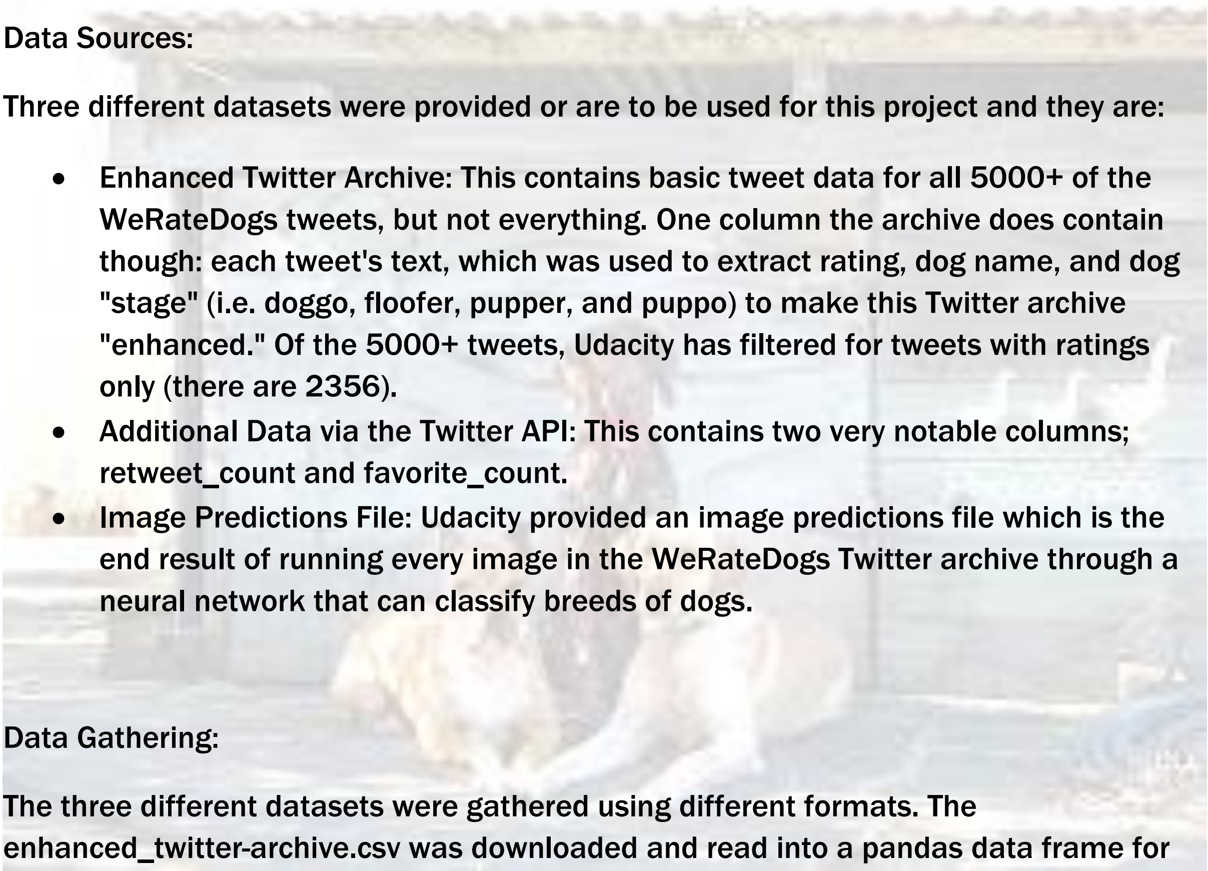
WeRateDogs Udacity Project

Data Wrangling Project by Theresa Sunday

Data wrangling is an important aspect of any data analysis process. Final analyses might be affected if data is not properly wrangled. A data scientist or analyst spends most of his time wrangling data. Data hardly ever comes clean.

For this project, data wrangling is split into three parts namely-: Data gathering, Data assessing and Data Cleaning.



Data Sources:

Three different datasets were provided or are to be used for this project and they are:

- Enhanced Twitter Archive: This contains basic tweet data for all 5000+ of the WeRateDogs tweets, but not everything. One column the archive does contain though: each tweet's text, which was used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, Udacity has filtered for tweets with ratings only (there are 2356).
- Additional Data via the Twitter API: This contains two very notable columns; retweet_count and favorite_count.
- Image Predictions File: Udacity provided an image predictions file which is the end result of running every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs.

Data Gathering:

The three different datasets were gathered using different formats. The enhanced_twitter-archive.csv was downloaded and read into a pandas data frame for assessment. The image predictions file was downloaded programmatically using the requests library. The third dataset should have been queried from twitter but due to issues with getting a twitter developer account, I read a downloaded json file containing data on retweet and favorite count into a data frame.

Data Wrangling Report

Data Assessing:

The next stage after gathering the data was to assess for tidiness and quality. I had to use visual and programmatic assessment that was taught during the course to correctly identify issues with the data's quality and tidiness. I had to use the .info(),.head(),.sample() and so on to identify at least 8 quality issues and 2 tidiness issues.

Data Cleaning:

This had to be the hardest part of the process for me, I had to learn new pandas functions to manipulate data and clean data. Melting 4 columns into one had to be the greatest challenge for this project. All highlighted issues were cleaned perfectly in the jupyter notebook .

The cleaned dataset was stored in a new csv file which would be used for analysis.