Data Wrangling Project

Wrangling Process Overview

This report goes through the process of gathering, assessing, cleaning and analysing a data set, particularly the - WeRateDogs account on twitter. This twitter account takes user submitted pictures of dogs and rates them with a humorous comment for each picture. Dogs are rated on a scale of one to ten, but are invariably given ratings in excess of the maximum, such as ”13/10”.

Data Gathering

The data set was gathered from three different sources:

- twitter_archive_enhanced.csv: A comma separated values(CSV) file containing information regarding tweets: mainly the rating and stages of the dogs in the tweet, along with the timestamp and the text of the tweets. This file contains 2356 rows and 17 columns.
- tweet_json.json: A JSON file which contains detailed information about each tweet, mainly the retweet count and the favourite count associated to each tweet. This data was downloaded using the Tweepy python library. This file contains 2332 rows and 32 columns.
- image_predictions.csv: A tab separated values (TSV) file containing predictions based on each image. This data set was downloaded using the requests python library. It contains the image URL associated to each tweet, along with the predictions of the contents of the image, which are derived using a neural network. This file contains 2075 rows and 12 columns.

Data Assessment

The assessment phase involved the process of finding issues with data set which hinder our analysis phase. The most prominent issues identified are listed here:

- Right off the bat, we notice that the number of rows in each data set don’t match. This means that we do not have the data for each tweet present in the archival data set.
- Lots of the columns in the archival data set and the scraped data set have null values. Some columns also had redundant data.
- Categorical data was split across multiple columns, thus creating these sparse columns with a lot of null values.

Data Cleaning

The cleaning phase involved the process of rectifying the issues described in the previous assessment phase. This was done by using various programmatic techniques accompanied by the occasional manual search-sort techniques. The initial data set had 6763 rows (all three sources combined). After applying various data cleaning methods and normalizing the data set, we are left with 1811 rows.