wrangle_report

June 21, 2018

In the world of Twitter, WeRateDogs is a popular account known for posting fun pictures of dogs with humerous appraisals with scores that are almost always about 100%. For this project, I used data from WeRateDogs to practice wrangling and analyzing.

Three files were involved with this process: - ‘twitter-archive-enhanced.csv’ was easy to import, as it was provided manually. - ‘image_predictions.csv’ was a little trickier, as it needed to be programmatically imported to the project. - ‘tweet_json.txt’ was the hardest to import, as it involved connecting to the Twitter API and programmatically grabbing extra metadata for each tweet we had in the archive.

Importing this data was informative, as each file has a unique method of retrieval and parsing. Considering how much can go wrong when recreating an environment, it is important to make sure as much of your process is programmatic and reproducible.

There was a lot to be cleaned in the data, as with most data sets in the wild. In no particular order, a number of things to look out for in any data set: - Make sure your data types are correct. Strings should be objects, numbers should be an appropriate numeric type if they are operable and objects if they are not (such as if they are IDs, zip codes, etc). I have noticed that pandas does not play nice with datetimes, and they need to be fixed after import. - Make sure you are encoding correctly! This one can save you major headaches and is very simple to do. - Fields with human input will always be messy. There are a number of standard things to check for: bad encoding, leftover symbol characters (ie ‘ “&” ’), capitalization, spelling mistakes, and different number forms (1 or one) to name a few. - Check for fields than can be melted down. This seems to happen more often that I would have thought and can be caught with a visual overview.

I found that there was some churn in this project, as I expect there will be in future projects. By churn I mean that after I cleaned the data, I moved on to find insights only to discover that I need to go back and fix something new. In some cases, I simply wanted to go back and further modify data that was already correct, but not yet useful for an analysis I wanted to run.