WeRateDogs

A Data-Wrangling project by Emmanuel Bett

Wrangle Report

This project makes part of the coursework leading to Udacity's Data Analysis Nanodegree. Its objective is to demonstrate skills for the Data Wrangling phase of the Data Analysis Process. Data Wrangling is a key part of analytics and is typically described as the one where analysts and data scientists spend the most part of their time. The project gathers data from different sources related to the WeRateDogs Twitter account, which posts and rates photos of followers' dogs. After assessing and cleaning the data, reports are written to communicate the results of the initial analysis. I thought Data Wrangling was going to be the easiest and fastest of all lessons in the DAND. I swiftly listened to every video and assumed I already knew almost all of it. However, when I started the project, I quickly realized that I needed to put much more work into it. Already in the gathering phase, specifically querying Twitter's API, it was clear that I needed to go over the material again, take notes, and practice in my own Jupyter Notebook. I found functional the structure (define, code, test) proposed for the cleaning process, but in my case, it was clear that I needed flexibility and iteration. I had to go back to cleaning more, even until the last chart was produced. I believe sticking to the process is a must, especially when dealing with multiple sources of messy and untidy data. Although it was recommended to start by solving tidiness issues, in this case removing first the unnecessary rows with retweets and replies, followed by the organization in tables of observational units was important. When cleaning the datasets, I felt like I was able to put into practice many of the skills I have been learning over the last months. Extracting HTML contents from a tag within the column of a pandas data frame using BeautifulSoup was a nice accomplishment. I have been working with relational databases for a while now, and I have already developed the sense to understand when observational units are not in their own tables. I ended up keeping separate tables for tweets, dogs, and predictions. What I wasn't sure about was whether a master dataset made sense because having different row counts makes the result different from what you expected.

The key is to keep in mind whether you want to keep dogs or tweets as the main organizational unit in a master dataset. When creating charts for the last section of the project, I really missed R's ggplot. I started using matplotlib but was quickly drawn to look for other options, which brought me to the Altair library. I read its documentation and plotted my visualizations on it. Although I found Altair intuitive to use, when I looked under the hood I realized that every chart makes a copy of the whole dataset in JSON format, which makes it less than ideal when working with large datasets or when performing EDA with tons of visualizations. Lastly, I used Jupyter Lab for the first time with this project, which I found a nice evolution from the traditional Jupyter Notebook. Being able to read text files directly in the workspace, was a good add