Rebecca Lewis

November 16, 2019

DSC 540

Final Project

For the final project, I connected to an API from the cocktaildb.com, an “open crowd-sourced” online database of drink recipes. [1] Their free JSON API allows users to connect to the database and search for drinks by attributes like name, ingredient, or category. For this project, I performed two separate API calls. The first was to get a list of all the drinks that included Vodka. After evaluating the data that was returned, I realized that this was only a high-level view of the recipes. In order to get the ingredients in each drink, I had to perform another API call using the drink id of each recipe. Once I retrieved a dataset that contained all of the ingredients, I joined it with the overview information and loaded it into a SQLite database.

My cleaning and transformation steps included:

1. I cleaned up the column names to remove hyphens to play nicely with SQLite.
2. The ingredients data included many columns that I did not need, such as foreign language translations for instructions and ingredients. These were excluded from the dataframe.
3. The ingredients and amounts needed were in different columns. After I read the data into a dictionary format, the columns were unordered making the recipe hard to understand. When filtering data into only the columns I needed, I reordered them so the measures and ingredients were together.
4. Once I cleaned the ingredients data, I joined it with the general recipe dataset and loaded the combined data into the database.

I encountered the following challenges:

1. My biggest challenge was time. I started on this project much later than I should have so the stress of completing it in time interfered with my ability to think clearly. Additionally, I may have performed more elaborate cleanup steps.
2. I spent far too much time trying to get dictionaries and lists to work. Once I changed my approach to using dataframes, I was able to perform the steps to retrieve and combine the data in a way that was intuitive for me.
3. When creating the dataframe, I explicitly set the order of columns. However, when I was inserting each row into the database, the ‘to\_dict’ method was unordering the data again. I found a solution online that converted each row to an ordered dictionary using the order of the dataframe columns. This allowed me to maintain the structure.

The data wrangling process is a crucial part of the data science life cycle. Before a data scientist can derive meaning from data, he/she must ensure they are retrieving all the data needed and that it is in a format that can most effectively be used for analysis. An important lesson I learned with this project is that there is not one “correct” way to perform analysis or cleaning steps. For each project, the best tool for the data and wrangler should be used. Of course, this will very based on the individuals strengths and weaknesses.

[1] TheCocktailDB.com. Retrieved November 16, 2019 from <https://www.thecocktaildb.com/>