In bracino quando sumus

**AN ETL PROJECT TO DEMONSTRATE THE SKILLS WE HAVE LEARNED ABOUT RELATIONAL AND NON-RELATIONAL DATABASES. SUBMITTED 16 NOVEMBER 2019 BY AMY WHITE & HUY NGUYEN.**

[WHEN WE ARE AT THE BREWERY]

|  |  |  |
| --- | --- | --- |
|  | | |
|  | INTRODUCTION In getting to know each other and exploring mutual interests, we discovered that both of us enjoyed a nice glass of cold beer at the end of a long day. This led to us searching to see if anyone had made public datasets analyzing craft beer, as one of Amy’s college friends keeps a spreadsheet ranking every beer that she has ever tried. They were a-plenty, so we decided to make this our project topic.  Throughout this project, we struggled to compare datasets from different sources as each had their own list of beer ids, as well as brewery ids. This led us to question a bit of our methodology, and whether we may have gotten exactly what we wanted had we scraped the data from the internet ourselves. However, we decided to stick with the data we had, pulling as much useful information as possible. We gained a bit of respect for those Data Scientists who do this daily, as it seems pretty monotonous to be constantly updating data so that the reference labels stay the same. |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  | EXTRACT We found our first dataset, which was later abandoned for better, more thorough data, when searching through Kaggle.com for popular datasets. Our two main datasets both came from data.world, where fellow developers had put together a database of spreadsheets analyzing beer content (ABV: Alcohol by Volume, IBU: International Bitterness units, & SRM: Standard Reference Method, which classifies beers by color) against styles, categories, and the breweries that did the real work, as well as a database with a spreadsheet containing reviews on both beers and breweries.  One of the most interesting datasets that we found was an API, openbreweryDB.org, that someone had stylized into a python library. While we were unable to gain access in the limited amount of time to the untappd API, or BreweryDB’s API, this provided a solid set of data that we would recommend most to craft beer aficionados. |  | TRANSFORM Amy worked with the first dataset, called Beer-Curious. In this dataset, there were 6 spreadsheets: a primary sheet with the list of beers, which also referenced 3 other spreadsheets, styles, categories, and breweries. Upon further review of the data in the beers and styles spreadsheets, Amy recognized that many of the SRM values were absent, and much of the IBU data appeared to be off, so she decided to eliminate those columns in PANDAS. Additionally, she reordered the data in each spreadsheet to be listed in order of the ‘id’, often the primary key, rather than the index number.  Having abandoned the Kaggle dataset, recognizing that the Beer-Curious had the same data but in a more thorough matter, Huy went to work on the review dataset, called beeradvocate. Within this set, he extracted the different reviews of each brewery and created a new column that would provide us with the average review of each brewery.  Amy and Huy both worked with the data from openbreweryDB API. Huy initially cleaned the dataset, eliminating unnecessary columns (mostly those with [] or NaN values), and Amy divided the information provided in that dataset into 4 worksheets: one for brewery type, another for brewery contact, a third for brewery address and the last for brewery coordinates. The fourth was the most difficult, as we were unsure how to prepare this data to be loaded into a SQL database while maintaining the accuracy of the Latitude and Longitude columns. |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | NUMBER  OF  CRAFT BREWERIES IN AMERICA THROUGH THE  YEARS | A close up of a map  Description automatically generatedA picture containing text, map  Description automatically generated |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | LOAD Amy created the initial schema using quickdatabasediagrams, as she found the ERDs generated helped her the most when navigating SQL datasets. Later, she went back to create a second schema using the tables that she had created using the data from openbreweryDB, as she was unsure without the second dataset referencing any additional tables whether it would be considered to have used enough sources.  Huy then took this schema, shared over a google drive folder that Amy had created to house the project work, and uploaded the datasets to pg Admin 4 using SQLAlchemy. He also consolidated the multiple cleaning datasets that Amy wrote into a single Jupyter Notebook file that also contained his work. Additionally, he created a new folder housing only the information and resources which were used by the project.  Amy put the report together in Microsoft word, and cleaned up the formatting to make the project look presentable. She also edited the file created by Huy so that it would include the additional work she had done with the openbreweryDB dataset, as she wanted to present all of the work that she had put into the project.  Our resources for each step are listed in the diagram below.  For any additional questions, please feel free to reach out to either of us via Slack. | | | | | | | | | | | |  |
|  |  |  |  |  | |  |  |  | |  |  |  |  |
|  |  | | | | A circuit board  Description automatically generated | | | |  | | | |  |
|  | EXTRACT   * openbreweryDB.org * data.world/   mathiasburton/  beer-curious   * data.world/   socialmediadata/  beeradvocate | | | | TRANSFORM   * Much like the transformation of hops into beer throughout the brewing process, we used PANDAS, a Python library, to sort and clean our data. | | | | LOAD   * app.quickdatabase   diagrams.com   * SQLAlchemy * pg Admin 4 (PostGres) | | | |  |