ELT Project - 2019 Summary

Team Members: Brickey LeQuire and Alex Koynoff

Team Name: Drunken Zip Codes

Summary: Following ETL procedures, we gathered information containing count of bars in the USA with zip code information. We obtained additional information by zip code to bring in various demographics for analysis as appropriate. We reviewed the data sources, and removed columns that would not be needed for analysis. We then scrubbed the information further and performed cleaning, such as removing NaN values and default values that the Census API brings in such as default placeholder values of -666666666.0.

After the data was reviewed and cleaned, we created tables in MySQL that will be used for loading of the cleaned data.

The loaded data in MySQL could be used for analyzing the concentration of bars in zip codes, cities, states. Also, census information could be used to calculate bar count per capita for example. The code could also be updated easily to pull additional census information using its API, depending on the business need.

Detailed Process:

Extract: Gather the data from the sources in csv and API.

1. Dataset with 17,000+ bars.
2. Dataset with zip code information to bring in states and city name.
3. Census API to pull in various demographics.

Transform:

- Review the datasets and select the columns to load.
- Clean the datasets by dropping blank values as appropriate and any bad default values.
- Perform merging of data as appropriate.
- Translated zip codes for missing leading zeros
- Clean the zip codes for all data sets to make them consistent (XXXXX vs XXXXX-XXXX)
- Derive calculated values (e.g. count of bars by zip code).
- Create a database and tables before loading:
    - Database: **Bars_db**
        - Tables:
            - **bars_count**: List of bars count per zip code
            - **census**: Demographics and zip codes from Census
            - **zip_info**: Zip code information (city, state)

- **bars_values**: Merged data frames of bars, Calculated columns(done via Python) and zip codes

Load: Load data to a relational database (MySQL).

- Once the data was loaded, we ran into the below situations/possible enhancements:
  - The “if_exists” function behaved a bit different that we thought. “Replace” removes all records from the table if any exist, and it re-inserts the values again. This worked ok for our database due to its small size, but it might be an issue for very large databases. The “truncate” function was done as a somewhat of a workaround before loading anything into the database. It functions similar to “replace”. Ideally, a function using the “if exists()” would need to be implemented to only load unique values. It is a complex function that would require more time to implement after this project.
  - When running a query to get the sum of bars per state, the query timed out. It was looping through one bar at a time and 40,000+ zip codes, going in a loop over and over. We indexed the zipcode columns and that made the query finish within 2 seconds.
  - If given more time, we wanted to create the MySQL tables via MySQLAlchemy to fully automate the T&L portion of the ETL process. We created classes to create the tables and column names, and got it to work once, but when trying to iterate through the dictionaries to upload the information to MySQL, we needed more time to finish the code and fully test it. A reference jupyter notebook with this code is uploaded to GitHub. It is named: **DrunkenZipCodes.SQLAlchemy** for reference.

Sources:

Bars info: https://data.world/datafiniti/breweries-brew-pubs-in-the-usa

Zip Code info:

- https://www.unitedstateszipcodes.org/zip-code-database/

Census info:

- Used Census API

MySQL tables and queries: refer to the MySQL file