ETL Project

CWRU Data Analytics Boot Camp

Spring 2020

Technical Report

Randy Dettmer, Evan Kamis, Ali Rizvi, Sumeet Dhawan

**Impact of Brewery Trend on DUI**

**Background**

It's a great time to be a craft beer fan in the U.S.! There are a ton of beer styles and brands to choose from and breweries have become very successful in the last several years. Breweries owe it all to beer lovers around the world! However, the other side of the story is that DUI’s are also increasing and we would like to see if there is a correlation of rise of breweries in US with the amount of DUI’s in the 50 states in the US.

**Extraction**

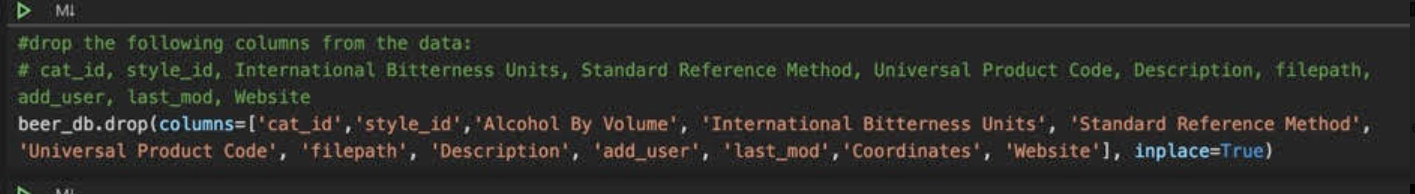
We explored datasets from the public platform Kaggle and Open Beer Dataset. Upon reviewing the datasets, we realized that the Open Beer Dataset contained over 5000 entries of breweries of which ~4500 were from the United States and were broken down by States and hence, we selected the dataset. Additionally, the DUI dataset from Kaggle that contains DUI related arrests and deaths in the United States was also broken down by States was also selected. All the datasets were from the year 2015 as that was the most recent we could find to study our hypothesis. The final sources of our datasets are as follows:

* DUI data from Kaggle – CSV File
* Open Beer Dataset – CSV file which also could be called through an API ([https://data.opendatasoft.com/explore/dataset/open-beer-database%40public-us/table/](about:blank))

**Transformation**

Our first steps in cleaning up the datasets was to use Pandas and figuring out which variables were not relevant. After multiple parser error and corrupt file errors, we managed to

open csv file first in Excel and then save as a csv uft-8 file. After this step, we dropped multiple columns that did not add value to our study. Figure 1 illustrates the datframe after dropping the non-relevant columns and the NaN values. Further, we realized another problem in the dataset which required significant cleaning and attention. Some states had repeat entries due to the way they were represented, for example, there were data entries with the State “California” and data entries with state “CA” which are the same but could not be grouped together. We had to use the *replace* function in Pandas to overcome the problem. Dataset also contained entries from outside USA which we needed to filter using the *iloc* function.



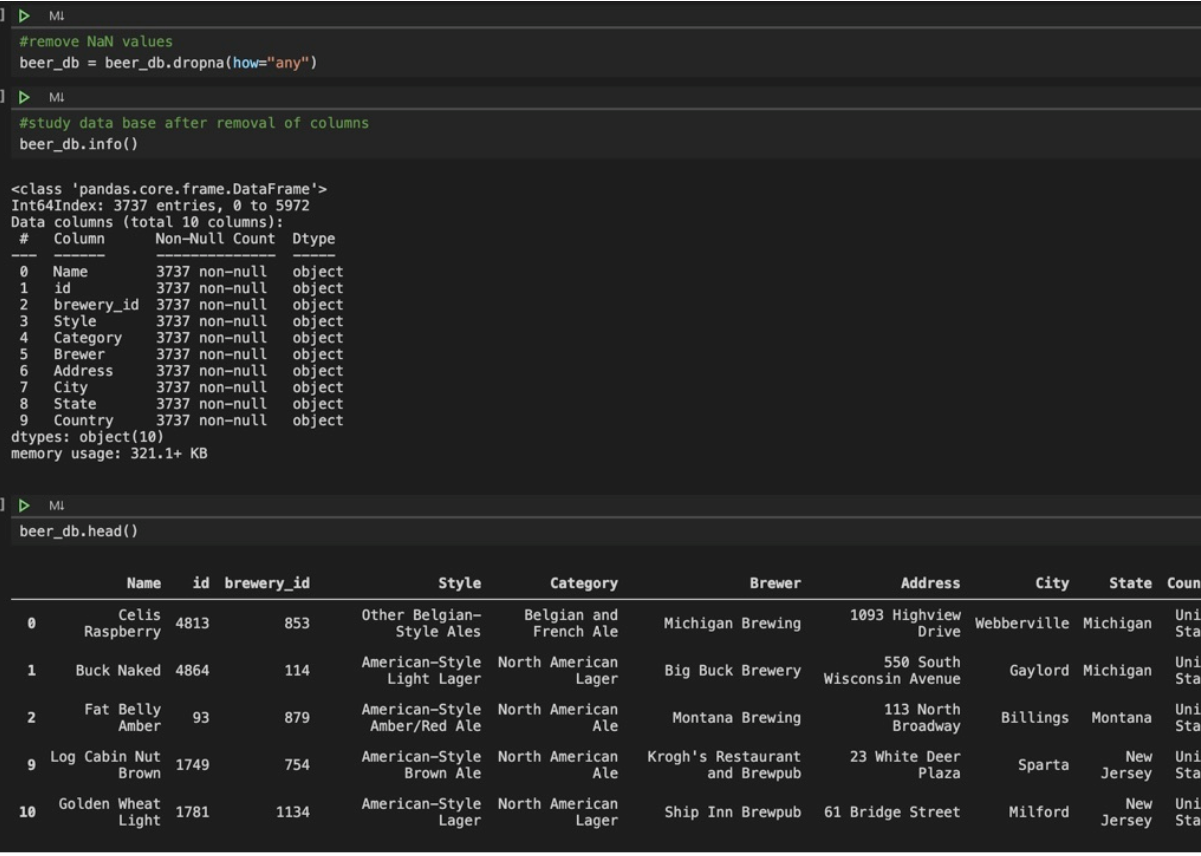


Figure 1

After some initial cleaning, we used Quickdb visualization to understand the data columns and the relation. The below figure 2. Illustrates the 5 datasets and their relationship between each other. Figure 3 illustrated the one to many relationship between the final two datasets we selected.

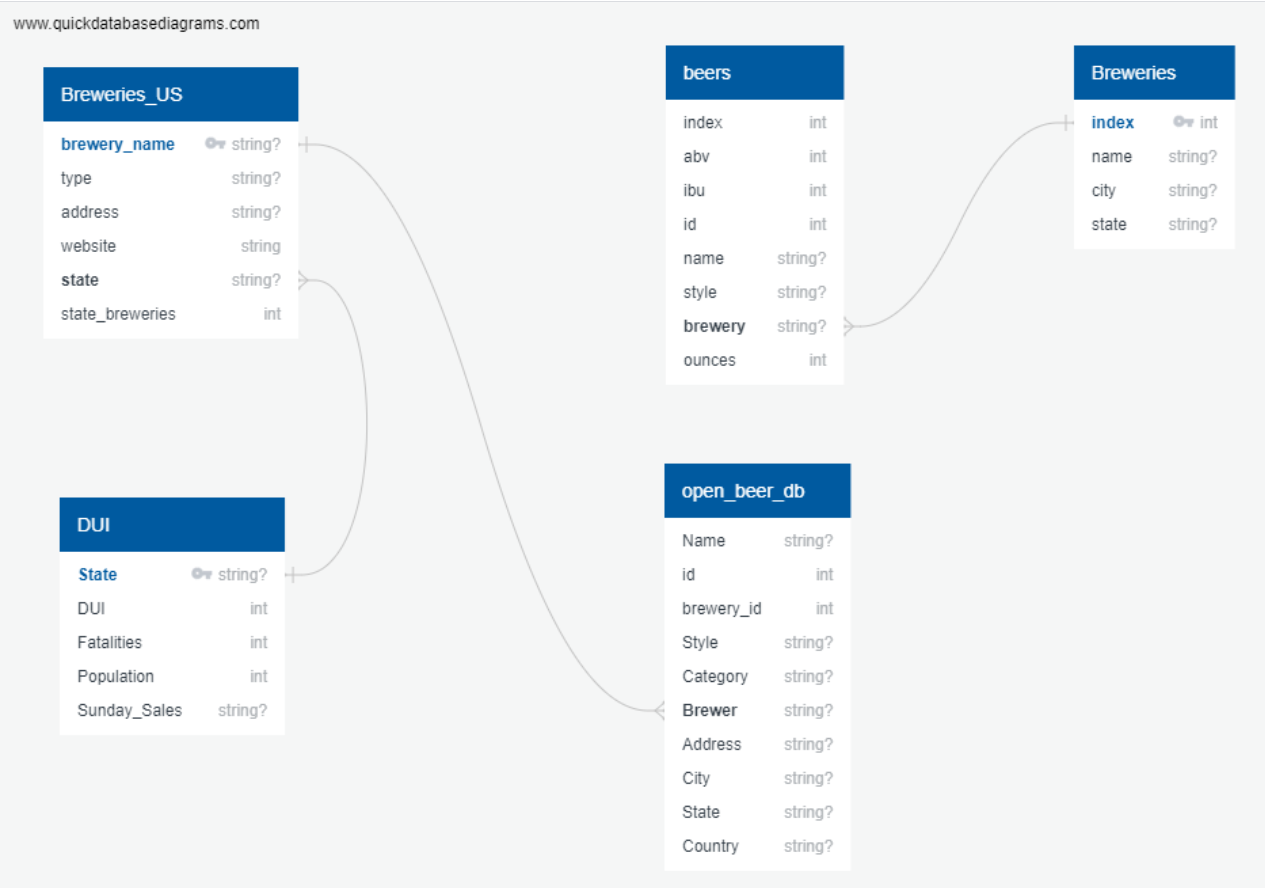


Figure 2



Figure 3

The dataset was grouped by state (Figure 4) and then merged with the DUI data (Figure 5)

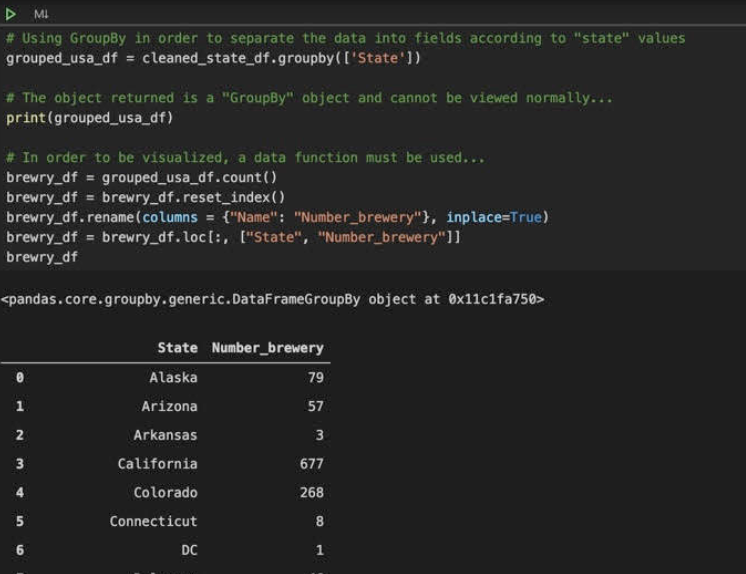


Figure 4

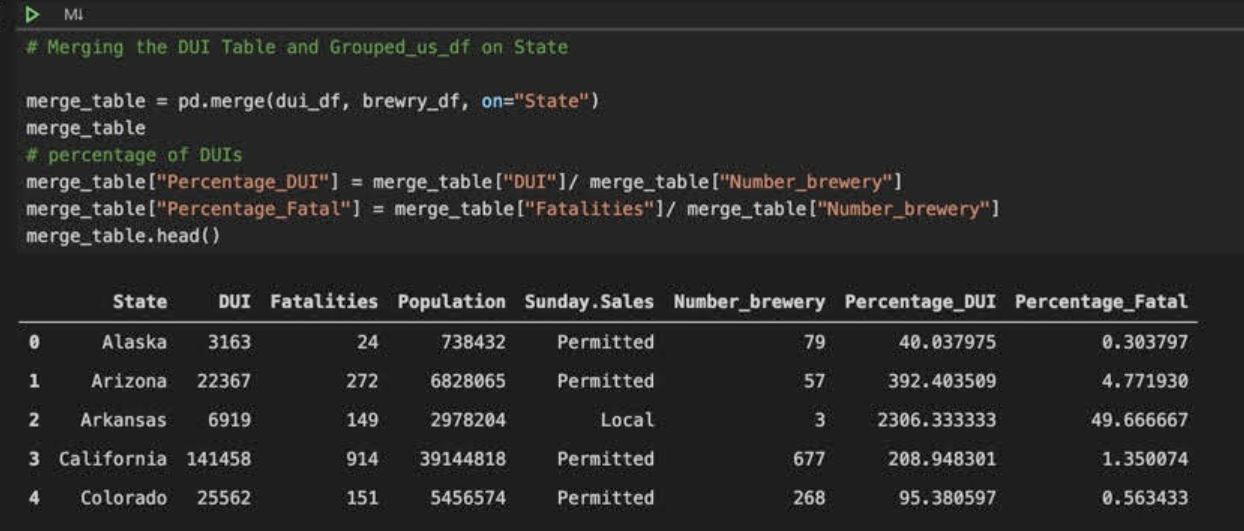
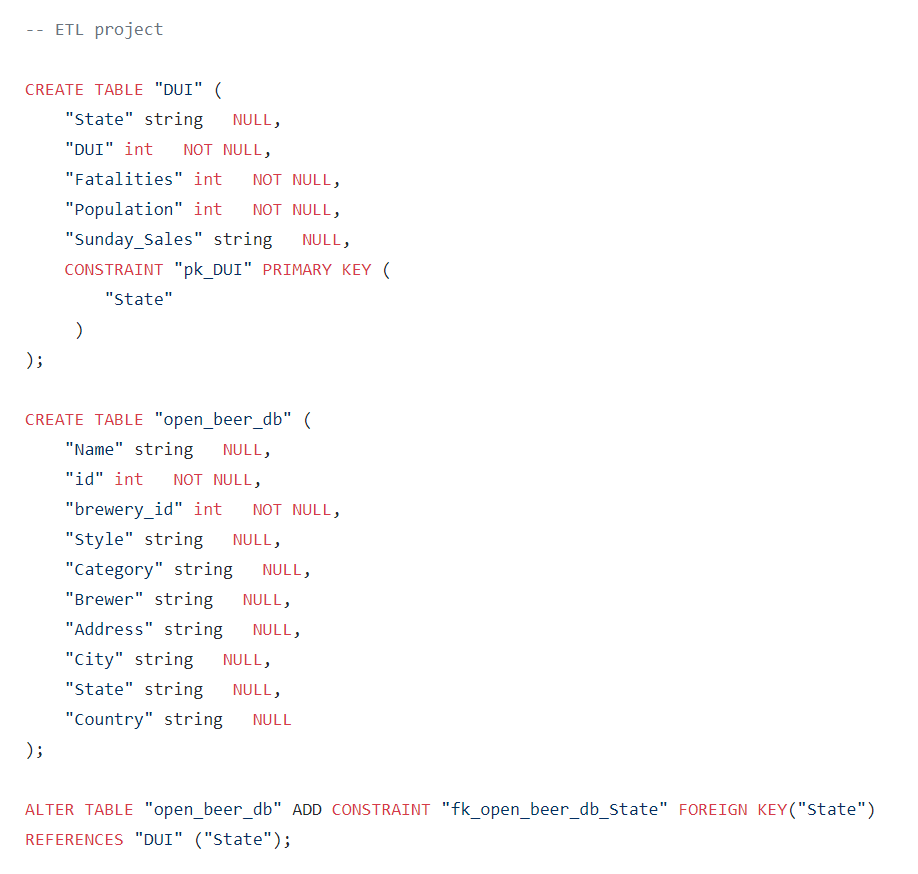


Figure 5

**Load**

The last step was to transfer our final output into a DataBase. The following query was used to establish the database.



Not done yet and not sure

We created a database and respective tables/collections to match the columns from the final Panda’s Data Frame using MYSQL and then connected to the database using SQLAlchemy and loaded the result. Here we were able to perform multiple queries to suit a desired criterion.

Finally, we pushed the dataset into MySQL database and joined the two tables using State

#### ETL Proposal.docx

* Word document submitted prior to starting project - containing initial project statement, goals, data sources