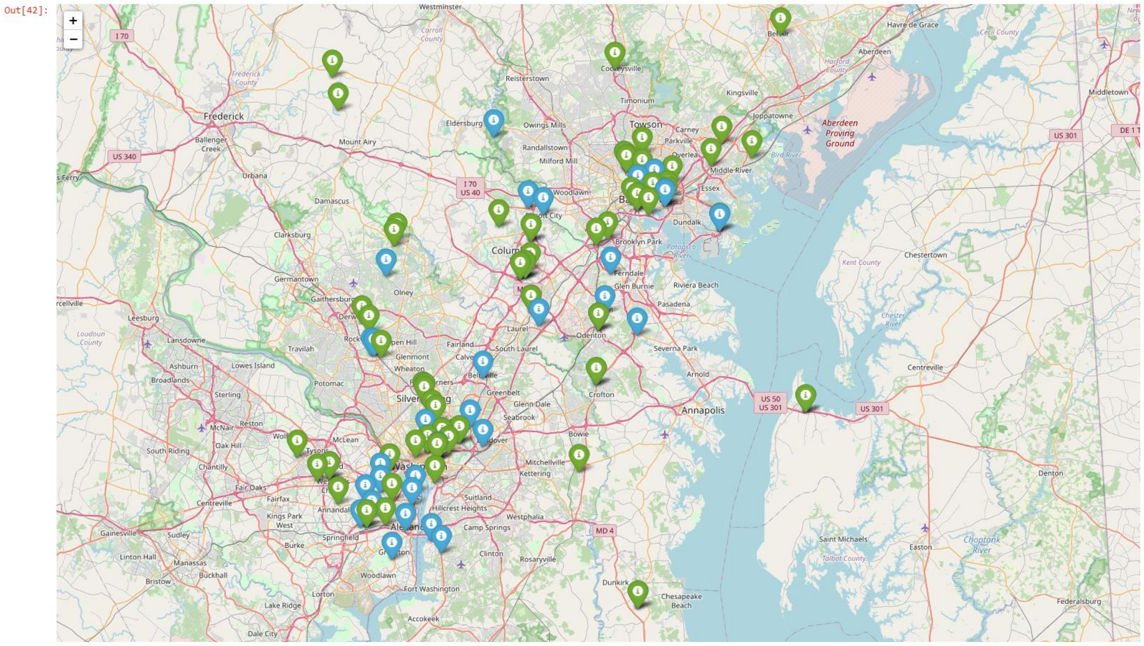

Finding a Spot to Build a Brewery

FINAL PROJECT

Alex Shaw | Applied Data Science Capstone | June 8, 2020

1. Introduction

1.1 BACKGROUND

For anyone that enjoys a visit to a brewery, it is likely that the thought of one day opening a brewery of one's own is tantalizing. There are certainly many decisions that would need to be made when pursuing this idea. One of the simplest and most straightforward is trying to decide on a location. There are a few assumptions that I have made for this project. First, is that it would be desirable to join a brewery district - which assumes that people that like breweries are already drawn to a an area that has breweries and it is easier to lure them a couple of blocks than to travel to a completely different area. Additionally, breweries tend draw locals more frequently than visitors and locals with more money are likely to spend more money.

1.2 PROBLEM

Within 50 miles of my location, there are over 80 breweries and there is no straightforward way to identify these self-defined features without collecting data and analyzing it. Additionally, providing a second metric to evaluate those areas would normally be quite time consuming.

1.3 INTEREST

There is a very straightforward case for the business case for being able to find an ideal location for opening a new location based on available data. For this specific circumstance, discovering brewery clusters could appeal to both the casual beer connoisseur planning their trip, as well as the entrepreneur planning their next move.

2. Data acquisition and cleaning

2.1 DATA SOURCES

Locating geographic data for breweries was obtained from both the Google Maps Places API as well as the Foursquare API. Housing data was obtained from Zillow on their research page in a .csv file.

2.2 DATA CLEANING

Combining the data from multiple sources was straightforward. Data was imported into a pandas data frame. Cleaning the duplicates out of the data was a bit challenging since the different sources had slightly different names for places. Utilizing the FuzzyWuzzy package, I was able to develop a likelihood score for the probability that any one entry matched any entry on the opposite list. For this to work properly common words needed to be removed – for this case the words brewery, brewing, company, and co were all removed before a matching score was generated.

2.3 FEATURE SELECTION

For the purpose of this project – the only data that was critical was the location data. The analysis was performed on a derivative data frame that only carried on only the name and latitude and longitude of the breweries.

3. Exploratory Data Analysis

3.1 MAPPING THE DATA

As this project was about finding geo-spatial relationships. It was critical in the process to ensure that the data was able to be mapped – which helped to verify that it was correct and that there was enough of it to performing the clustering in the next step. This is shown in Figure 1 – and demonstrates the importance of having multiple data sources. The pins in Blue are from Foursquare and the pins in green are from the Google Maps Places API. This data is shown again in Figure 2.

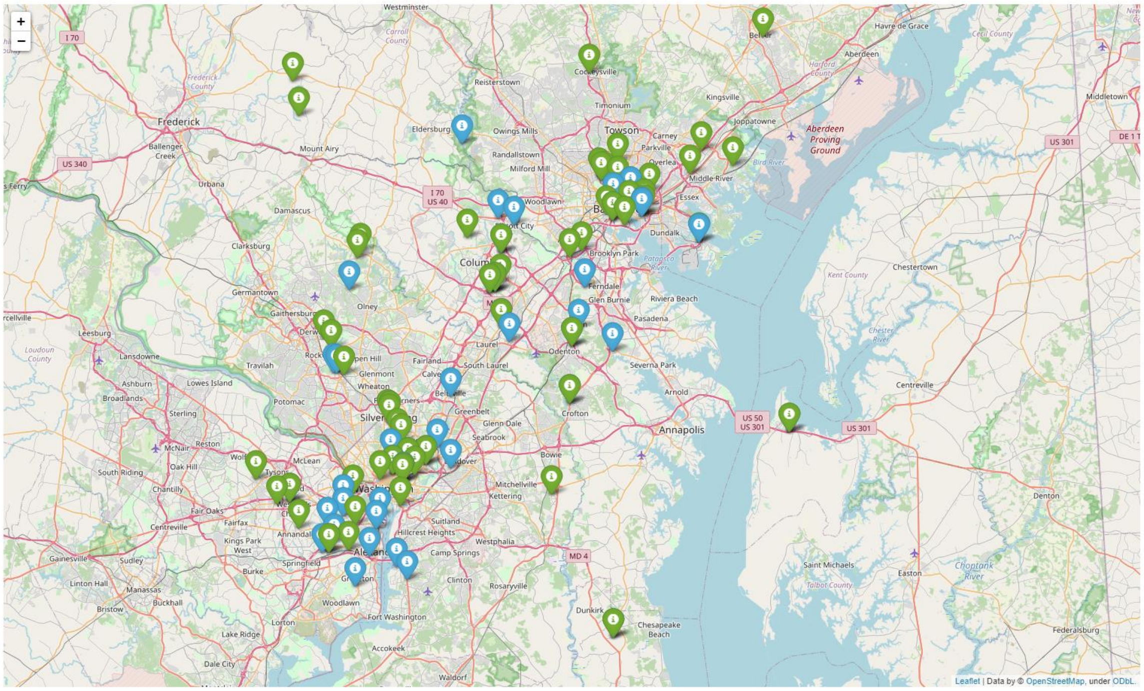

Figure 1 – Mapping the Data

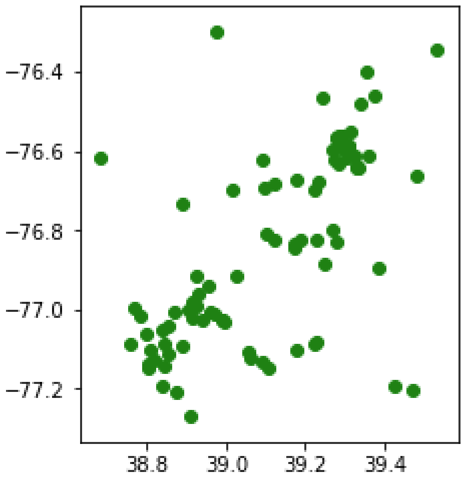

Figure 2 - Another depiction of the map data

It appears that there are at least several areas that would likely meet the constraints identified in the problem.

4. Modeling

4.1 MODEL SELECTION

DBSCAN was chosen for its ability to cluster based on meaningful geo-spatial parameters, in our case, I was able to select 2 kilometers as the maximum distance that two breweries could be from each other and be clustered together.

The results of this algorithm are depicted in Figure 3. One important note is that if the do not meet the criteria to be in a cluster – they are left un-clustered. In Figure 3 the un-clustered breweries are given the label “-1”.

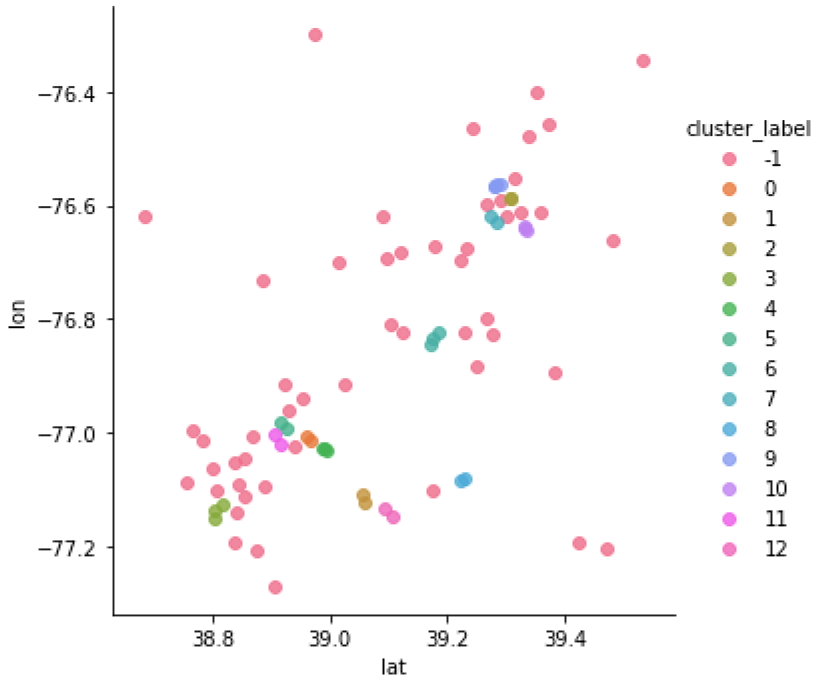

Figure 3 - Results of DBSCAN clustering

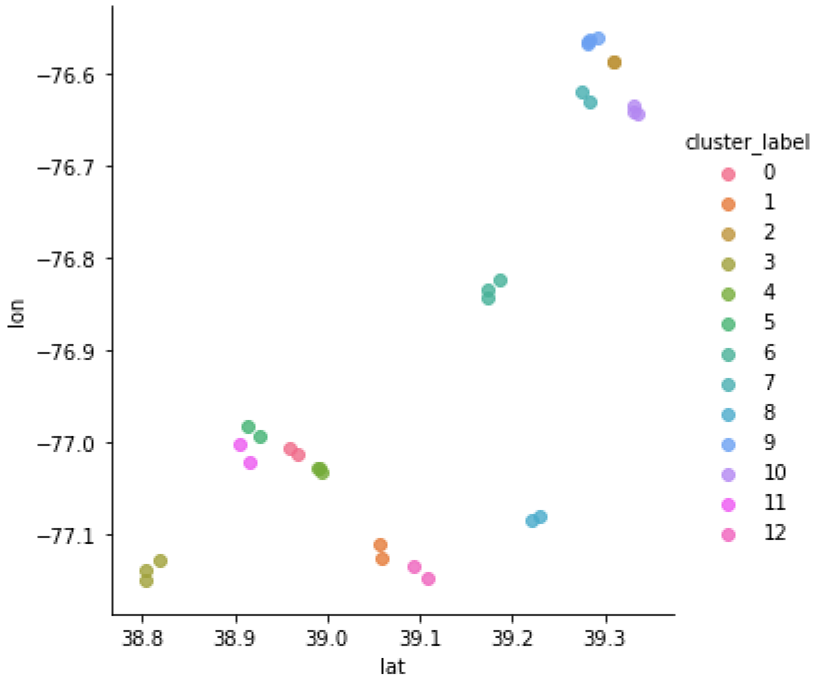

Figure 4 - Clusters with the un-clustered breweries removed

5. Conclusions

This methodology provides a rapid way to identify clusters and rank them based on the median housing values of the neighborhoods that they are located. The data that it returns is very robust based on the real-world assumptions that drove the algorithm.

If these parameters were truly all a brewery proprietor was looking to focus on – the answer for that person's problem would be crystal clear.

6. Future directions

It is obvious that there is a lot more to an entertainment than just the property values. Being able to compare the quality of these establishments would also provide an important additional metric.

Beyond simply comparing quality – being able to prove or disprove the underlying assumptions would be another exciting avenue of exploration – do breweries do better when clustered? Do higher median home values really translate into more revenue at breweries?