Applied Data Science Capstone Report

Introduction/Business Problem

Known for its beautiful coastline, summertime outdoor activities, seafood cuisine, forested interior, and many lighthouses, Maine is a heavily trafficked tourist destination. With over 37 million people visiting and spending $6.2 billion in 2018, it has a competitive and profitable hospitality industry. Your friend has been thinking of a career change and would like to join the hospitality industry in Maine by opening a brewery. Their brewery would provide both a unique social location and delicious, local beers to the many tourists.

Problem: Having never visited, they are at a loss on what area they should open a business in. Using the Foursquare API, we will identify the best location for their future brewery.

Data

For this project the following data is required:

1. Maine data (including city and town names and counties), scraped from Wikipedia https://en.wikipedia.org/wiki/List_of_towns_in_Maine, https://en.wikipedia.org/wiki/List_of_cities_in_Maine
2. Latitude and Longitude of the towns above, scraped from Maps of the World https://www.mapsofworld.com/usa/states/maine/lat-long.html
3. Venue data related to the towns above, Foursquare API

Methodology

The first requirement is to set-up the environment before beginning any work. This is done by importing the various libraries to be used throughout the notebook. These include numpy, pandas, matplotlib, requests, folium, and sklearn. The next set-up step is importing the necessary data. First, a list of cities and towns in Maine is needed. These can be scraped from Wikipedia, but with two different pages listing the needed details they need to be imported separately. Once imported they should be formatted to remove extra characters and/or spaces and combined. Since the Foursquare API uses longitude and latitude, this information is also required. These coordinates can also be scraped via the website Maps of the World. All of the data should then be joined to create a dataframe including the town/city name, country name, longitude, and latitude. A map can be generated to view the town/city data clustered by county and noted by various colors.

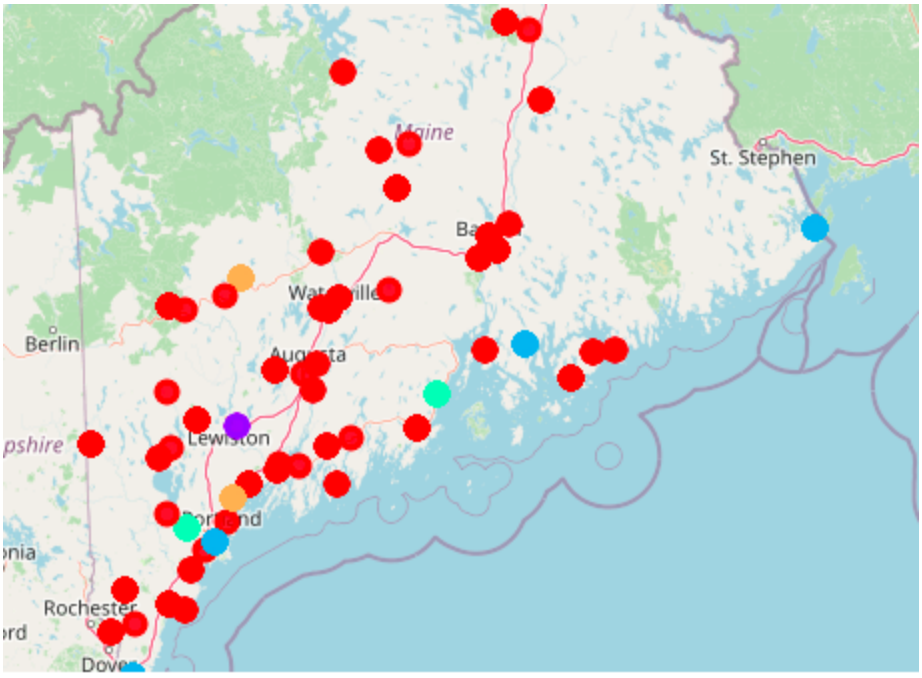
With the dataframe in good order, the Foursquare API can be called and a count of venues by Maine towns generated. The data should then be modified with and grouped by one hot encoded to allow for an easier to use binary system. The data can then be clustered using kmeans. The cluster labels are then applied to the dataframe for a complete view of the data.

Using Folium, a map is generated to represent the five clusters, each repeated by a different color. Finally, each of the “hits” for breweries within each cluster were invested to determine how many were in each and review additional details.

Results

The results can be interpreted one of two ways, each outlined below.

1. Due to the fact that the majority of breweries in Maine are found in the Portland/Lewiston area, one can assume that this is a popular destination with the economy to support the industry. Because this area is over a large enough square mileage, 35 miles between the cities, it can also be assumed that the area can support and would welcome an additional brewery.
2. If your friend is interested in opening a brewery in an area without competition, the data tells us that there are multiple clusters of towns, around large anchor cities, without active breweries. These anchor cities include Augusta, Waterville, Bangor, and Mount Desert Island.



Discussion

Unfortunately, with the unreliability of the geocoding library, geocoder, I had to find a dataset of Maine coordinates through another source. I was able to access a dataset that could be scraped via the web, but it was not as comprehensive as the list scraped from Wikipedia. This means that some smaller towns were excluded from the dataset. If they had nearby breweries, they were excluded from the clusters and may have skewed the results.

Another complication I came across was the lack of data available via Foursquare in Maine. After reviewing the data, it was found that no venues were logged for Portland, the largest city in the state.

Conclusion

Given the opportunity to redo this project, I would cluster by county and would then suggest a county to open a business within. With the sparse data, this would help return a more conclusive suggestion. If there were multiple counties with no businesses in the area of interest, I would do outside research on common tourist destinations to better help narrow down the results. To give additional help, I would select a more common business, such as seafood restaurants. With more data it would be easier to pinpoint the area of need.