A Distillery in the Finger Lakes

Introduction: Business Problem

An alcoholic beverage distiller wants to open a new distillery in the Finger Lakes region of New York. The Finger Lakes region of Western New York is a very popular destination for tourism - especially for visiting vineyards, breweries and natural geographic features, such as gorges and waterfalls. The Finger Lakes region is very large and mostly rural. The stakeholders want to find a location that will be close to similar businesses like wineries, and breweries, but also be located in an area that already gets a good amount of tourism activity. The stakeholders would also prefer to locate within an area where there are no existing distilleries if possible.

Data science methodology and tools will be used to identify and rank potential locations within the Finger Lakes region that are close to the most popular attractions while also being close to vineyards, wineries and breweries

Data Requirements and Sources

In order to identify the most desirable areas, based upon the distiller's preferences, we will need to analyze:

- the location of vineyards, wineries, breweries, and tourist attractions
- the density of preferred venue types within each area
- the popularity ranking of venues within each area, to be derived from Foursquare user activity

The Foursquare API will be used to provide venue information, location data, and venue popularity data. To support this analysis we will need the following data from the Foursquare API:

- name
- categories
- location (latitude, longitude, town)
- number of check-ins
- number of tips

Data available from the Foursquare API is incomplete related to licensed wineries, breweries and distilleries in the New York, Finger Lakes Region. Therefore, we will also use location data extracted from the https://data.ny.gov webpage which provides information regarding the name, address, type, and location of all New York licenses for alcohol production. The venues that are not listed by Foursquare will not have data to support popularity analysis. The following data from the New York State Liquor Authority will be used in this analysis to show clustering of similar alcohol producing businesses:

- name
- category (license type)
- location (latitude, longitude, town)

We'll also use Mapquest's Geo-coding API to get latitude and longitude values for the town's to be used in clustering analysis and visualization.

Methodology

In order to support a decision as to the placement of a new distilling business with the Finger Lakes region, we'll need to provide the stakeholders with an analysis of the location and distribution of other alcoholic beverage producing businesses along with an estimate of the level of consumer traffic that each potential area can expect. In order to accomplish this, we'll plot some basic maps of the data we get from Foursquare's API, the State of New York, and the Mapquest Geo-coding API. The Finger Lakes region is well known for the number of wineries located there. Maybe lesser known is the fact that there are also a good number of breweries in the region and distilleries are on the uptick. There are many attractions, other than wine, beer and liquor, that pull a good amount of tourist traffic into the Finger Lakes region – including natural features such as the lakes themselves,

gorges and waterfalls, and nature preserves. We'll use the Foursquare API data to create a traffic frequency estimate to add weight to our location data. We will then group the venue and traffic data by town and use k-Means clustering to analyze our location and traffic data to find similarities within the region.

Limitations of the Foursquare Data

To begin the analysis, we'll have a look at what the Foursquare API is giving us. We are going to look venues from the Foursquare API that are within a 50 mile radius of Ovid, NY. Ovid is roughly at the geographic center of the part of the Finger Lakes region that we are interested in – that is, the area with the assumed highest concentration of alcoholic beverage producing businesses and tourism.

Here's what the Foursquare API has for us:

- there is data for 100 venues within a 50 mile radius of Ovid, NY
- Foursquare groups these 100 venues into 46 categories
- out of the 100 venues Foursquare has data for 3 Wineries, 6 Vineyards, and 7 Breweries
- there is no data for distilleries available from Foursquare
- the categories related to tourist attractions that we will use in our analysis are – 'FARM', 'TRAIL', 'PARK', 'SCENIC LOOKOUT', 'WATERFALL'
- all other venue categories will be grouped together into a general category called "other" as they are not necessarily desirable based on stakeholder requirements but may impact traffic frequency

Traffic Frequency

Given the sparsity of the data available from the Foursquare API for the Finger Lakes region, we will only use this data to an overview of traffic data based on user activity. Licensing data from the New York State Liquor Authority will be used to show density of existing distilleries, wineries, and breweries. We need to come up with a way to derive a traffic frequency value from what the Foursquare API can provide. The easy answer would be to use the count value for checkins at a venue – but Foursquare has removed access to that data from the API. We could use the count value for "Likes" but anyone can like a venue in Foursquare and so it doesn't necessarily mean the user actually visited the location. So what we will use is the sum of counts for photos taken of a location and tips provided for a location. The count data for photos and tips are only available through detail calls for a single location – these are premium calls to the Foursquare API. Since our sandbox account on Foursquare is limited to 50 premium calls per day, we'll grab this data 50 rows at a time and load them into a CSV file.

Once the traffic data frame was been created, the traffic values were normalized. Here is a summary of the traffic data frame:

	traffic
count	99.000000
mean	0.135133
std	0.155953
min	0.004912
25%	0.038802
50%	0.083497
75%	0.154715
max	1.000000

Location Data from the New York State Liquor Authority

We need a more complete picture of where the wineries, breweries and distilleries are located so that's the data we grab from the New York State Liquor Authority. Here's what the Liquor Authority gives us:

- there are 216 rows of data for licensed wineries, breweries, and distilleries within 50 miles of Ovid, NY
- after removing duplicates and invalid rows, the New York State data contains 96 wineries, 38 breweries, and 14 distilleries
- many of these are missing location data and must be updated in order to be used in our analysis
- some of the rows provided are missing a name value, we'll use the DBA name for the row as the name value for these cases
- for rows with missing latitude and longitude data, we'll get the lat, long coordinates of the "city" value from the Mapquest Geo-coding API

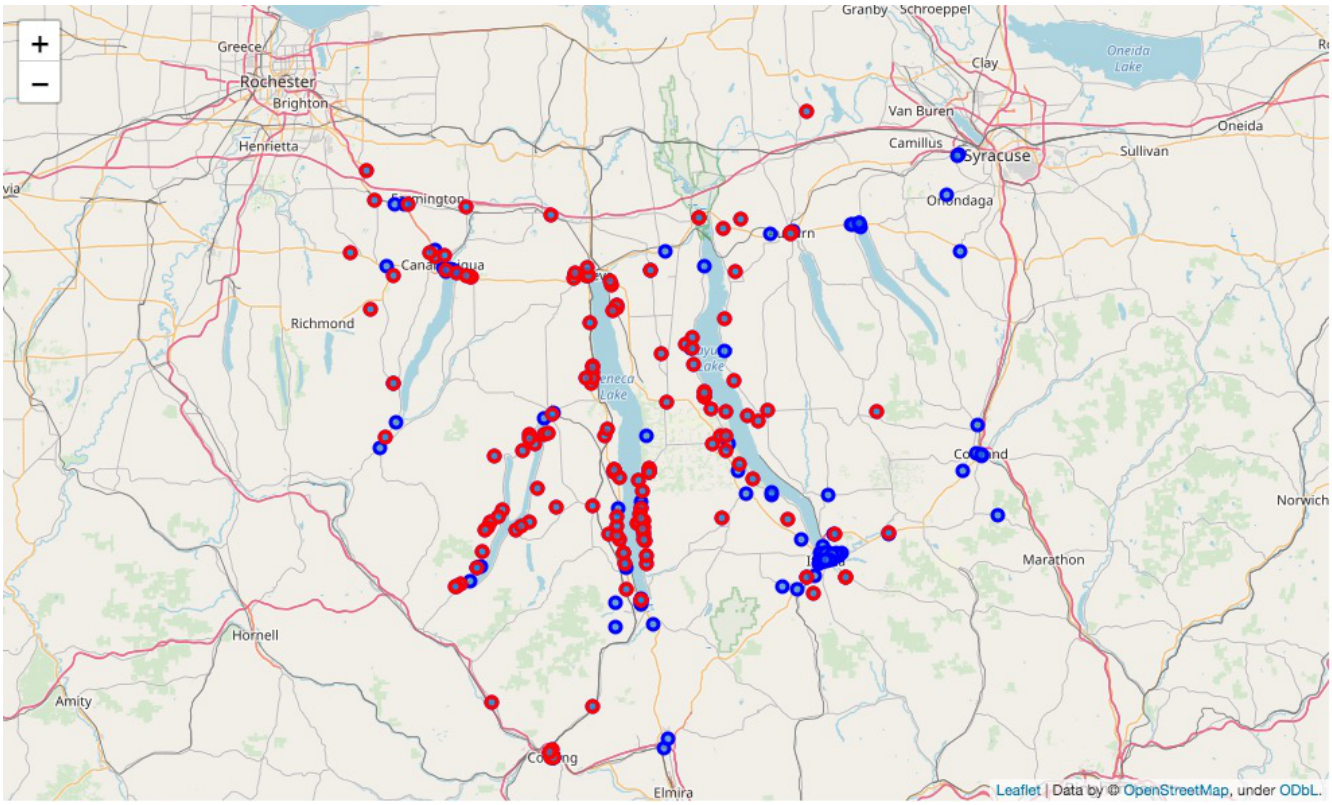
The data taken from the New York State Liquor authority does not contain information that can be used to predict traffic; so all of the traffic predictions will be made based upon what we can get from the Foursquare API. The Foursquare data is quite sparse when compared to the actual number of venues that located within our area of interest for this analysis. For the purpose of this analysis we will assign a traffic frequency to all of the locations obtained from the New York State Liquor Authority that is derived from the minimum traffic value associated with locations in the Foursquare data. The reasons that we will use the minimum traffic value are as follows:

- venues with the highest traffic values within the Foursquare API are not considered high value for the purposes of this analysis
- leaving traffic values for locations from the Liquor Authority data set to zero is unrealistic (just because there is no data from Foursquare related to a venue doesn't mean the location gets no traffic)
- using a mean value for traffic frequency skews the frequency data towards highly populated areas and this goes contrary to the requirements set out by the stakeholders to find locations in rural areas in close proximity to wineries and scenic attractions

The normalized traffic value that will be assigned to the NY State data is 0.004912.

Combining the data obtained from the Foursquare API and the NY State Liquor Authority gives us a data frame containing 106 wineries, 45 breweries, 13 distilleries, 11 scenic venues, and 72 other venues after we finished cleaning the data.

Here’s a map of the Finger Lakes region showing all locations available for use in our analysis from both the Foursquare API and the New York State Liquor Authority.



All locations from the Foursquare API (Blue) and NY Liquor Authority (Red)

Using k-Means Clustering to Analyze Location Data

Once we had built a data frame to provide location and traffic data for our analysis, we could now use k-Means clustering to group similar locations and provide insight into which areas of the Finger Lakes shared similar frequencies of the types of venues our stakeholders are interested in along with similar levels of traffic. After the cluster analysis is complete we’ll need to visualize the clusters on a map – so the cluster analysis will be based on the towns provided within our data frame for each venue.

After being grouped by town, the data to be used for clustering was normalized. In order to determine the most valid number of clusters for this analysis, the k-Means clustering was run multiple times with differing values of K (number of clusters) – clustering was run with K being set to 4, 5, 6, and 7. Running k-Means with k=5 provided clustering that best reflects similarities within our data. After running k-Means clustering against our data we get latitude and longitude data for the towns in our data frame and merge it with the k-Means labels

Cluster Group Counts

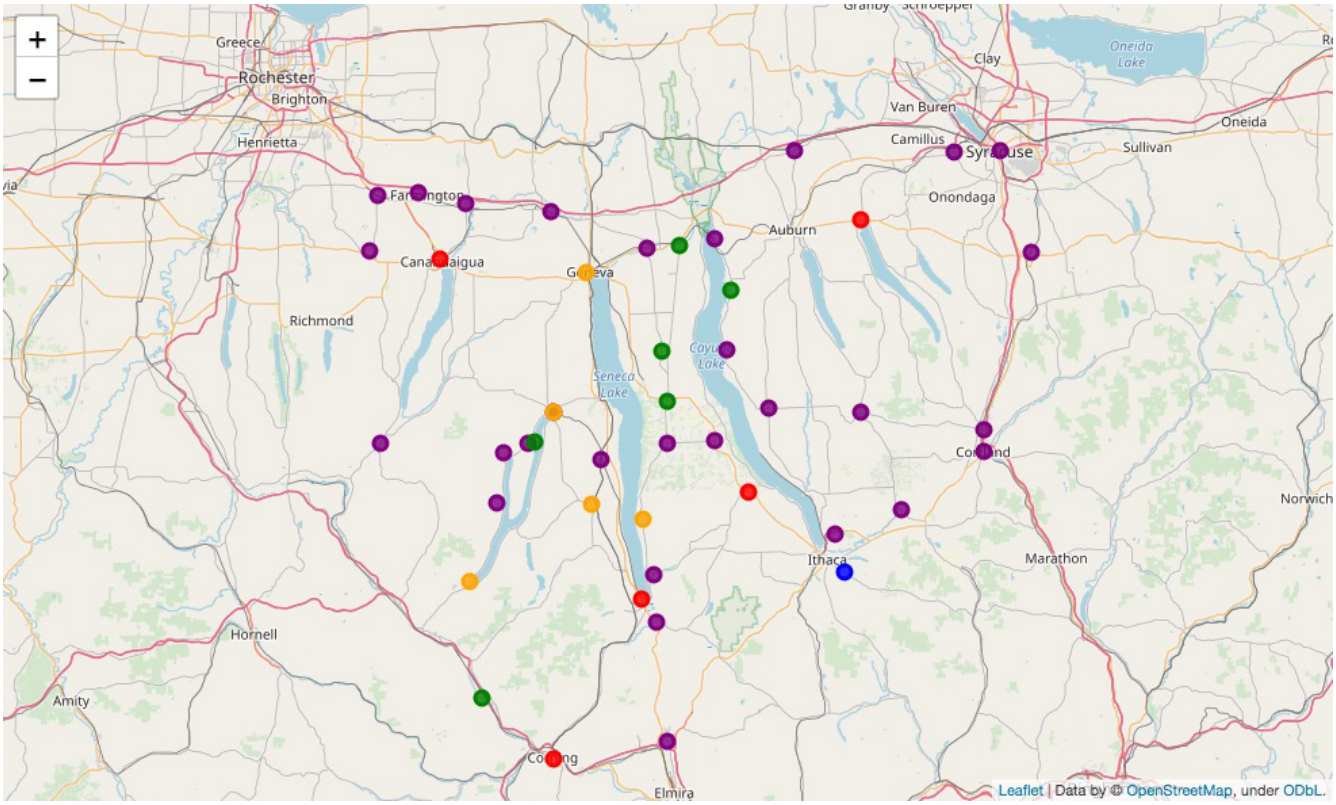
The follow table shows counts for venue category and traffic frequency for each of the clusters generated by the k-Means clustering analysis.

Cluster	Distillery	Winery	Brewery	Scenic	Other	Traffic
0	1	10	13	3	24	3.7407
1	1	2	4	6	20	5.2014
2	4	45	12	0	5	0.8861
3	7	13	1	0	2	0.1660
4	0	36	15	2	21	2.1179

K-Means clustering results

Mapping Cluster Analysis Results

The following map shows clusters of towns/areas with similar category counts and traffic frequency.



Cluster Results: Red = Cluster 0, Blue = Cluster 1, Orange = Cluster 2, Green = Cluster 3, Purple = Cluster 4

Cluster Data by Town

The following table show the cluster membership and venue category makeup for each town.

Cluster	Town	Distilleries	Wineries	Breweries	Scenic	Other	Traffic
0	Auburn	0	1	3	0	3	0.4529
0	Canandaigua	1	2	3	0	6	0.3998
0	Corning	0	0	5	0	3	0.2760
0	Skaneateles	0	0	0	0	7	0.5157
0	Trumansburg	0	4	0	1	2	0.6709
0	Watkins Glen	0	3	2	2	3	1.4253
1	Ithaca	1	2	4	6	20	5.2014
2	Dundee	1	7	1	0	1	0.2161
2	Geneva	1	7	5	0	1	0.2682
2	Hammondsport	1	9	2	0	1	0.1316
2	Hector	1	11	2	0	0	0.1218
2	Penn Yan	0	11	2	0	2	0.1483
3	Campbell	1	0	0	0	0	0.0049
3	Keuka Park	1	1	0	0	0	0.0098
3	Ovid	1	5	1	0	0	0.0344
3	Romulus	2	5	0	0	0	0.0344
3	Seneca Falls	1	1	0	0	2	0.0727
3	Union Springs	1	1	0	0	0	0.0098
4	Aurora	0	1	0	0	1	0.0806
4	Bloomfield	0	2	1	0	0	0.0147
4	Bluff Point	0	1	0	0	0	0.0049
4	Branchport	0	1	0	0	0	0.0049
4	Burdett	0	4	1	0	0	0.1336
4	Cayuga	0	2	0	0	0	0.0098
4	Cortland	0	0	0	0	6	0.3212
4	Fairmount	0	0	0	0	1	0.0108
4	Farmington	0	0	1	0	1	0.0344
4	Freeville	0	0	2	0	0	0.0864
4	Himrod	0	1	0	0	0	0.0049
4	Homer	0	0	0	0	1	0.0383
4	Horseheads	0	0	0	0	2	0.0422
4	Interlaken	0	4	1	0	1	0.2593
4	King Ferry	0	2	1	0	0	0.0147
4	Lafayette	0	0	0	1	0	0.1552
4	Lansing	0	0	0	0	1	0.0678
4	Locke	0	1	0	0	0	0.0049
4	Lodi	0	5	1	0	0	0.1051
4	Manchester	0	0	1	0	0	0.0049
4	Milo	0	1	0	0	0	0.0049
4	Montour Falls	0	0	0	1	0	0.0894
4	Naples	0	2	1	0	2	0.0894
4	Phelps	0	0	1	0	0	0.0049
4	Pulteney	0	1	0	0	0	0.0049
4	Rock Stream	0	5	1	0	1	0.0344
4	Sterling	0	1	0	0	0	0.0049
4	Syracuse	0	0	0	0	2	0.1228
4	Victor	0	0	2	0	1	0.0422
4	Village Of Bloomfield	0	0	1	0	0	0.2790
4	Waterloo	0	1	0	0	1	0.0373
4	Weedsport	0	1	0	0	0	0.0049

Cluster details with category counts and traffic by town

Results

From the k-Means analysis, it seems that locations within cluster 0 and cluster 4 best match the requirements identified by the stakeholders. Towns within cluster 4 are the only locations which meet all stakeholder requirements.

Cluster	Distillery	Winery	Brewery	Scenic	Other	Traffic
0	1	10	13	3	24	3.7407
1	1	2	4	6	20	5.2014
2	4	45	12	0	5	0.8861
3	7	13	1	0	2	0.1660
4	0	36	15	2	21	2.1179

Characteristics of cluster 4:

- contains 32 towns
- contains no other existing distilleries
- contains the second highest concentration of wineries at 36
- contains the highest concentration of breweries at 15
- contains the third highest number of scenic attractions at 2
- has the third highest traffic frequency at 2.1179

Characteristics of cluster 0:

- contains 6 towns
- contains 1 other existing distillery
- contains the second lowest concentration of wineries at 10
- contains the second highest concentration of breweries at 13
- contains the second highest number of scenic attractions at 3
- has the second highest traffic frequency at 3.7407

If we look at the details of towns within cluster 0 and cluster 4, it becomes clear that there are suitable locations within both clusters. The following table highlights possible locations that best meet stakeholder requirements from cluster 0 and cluster 4.

Cluster	Town	Distilleries	Wineries	Breweries	Scenic	Other	Traffic
0	Trumansburg	0	4	0	1	2	0.6709
0	Watkins Glen	0	3	2	2	3	1.4253
4	Burdett	0	4	1	0	0	0.1336
4	Interlaken	0	4	1	0	1	0.2593
4	Lodi	0	5	1	0	0	0.1051
4	Rock Stream	0	5	1	0	1	0.0344

Towns that best meet stakeholder requirements

Discussion

One of the requirements for this assignment was to use data available from the Foursquare API to solve a business problem. Given the sparsity of data available from Foursquare for the Finger Lakes region, this study would most likely benefit from some additional traffic statistics. However, even given the limitations of the Foursquare data, it appears that we have a good starting point for locating a new distillery in the Finger Lakes. We at the very least have a good understanding of the distribution of wineries, breweries, and distilleries and a basic view of how much consumer traffic each area enjoys. One additional datapoint that would be necessary in order to realistically make a good decision on this would be the availability of suitable properties for sale of

lease within the region. The result of using k-Means clustering to analyze the occurrence of wineries, breweries, and distilleries within the Finger Lakes Region, enhanced by traffic a frequency obtained via the Foursquare API, provide a short list of interesting locations to satisfy stakeholder requirements. We went from a total of 216 locations to analyze down to a list of 6 locations most suited to stakeholder needs, aided by clustering analysis.

Conclusion

The Finger Lakes region of New York is a great place to visit for tourists seeking to enjoy wine, beer, and spirits at their source given the prevalence of natural beauty and the density of wineries, breweries, and distilleries in the region. The very things that make it desirable, however, make it more challenging to decipher if you are interested in establishing an alcohol production related business in the region. The model developed in the process of this analysis provides a good starting point on which to base that decision. Enhanced with further data, this model could be very useful in selecting good candidate locations for a distillery.

References

Foursquare API - https://developer.foursquare.com/

New York State Data - https://data.ny.gov/

Mapquest Geo-coding API - https://developer.mapquest.com/