CAPSTONE BREWING

Leveraging data science to identify optimum brewery locations in Canada

John Fowler
28 July 2021

CAPSTONE BREWING

Leveraging data science to identify optimum brewery locations in Canada

John Fowler
28 July 2021

Introduction

The work presented here details my Capstone Project for the IBM Data Science Professional Certificate hosted by Coursera. This project requires a business problem to be identified that leverages location data from Foursquare and involves the clustering of neighbourhood data in its analysis. To that end a fictional company, "Capstone Brewing", is identified as the sponsor of this work. The company's management team hopes to leverage data science as a means of informing their new brewpub business on a path to success.

Background

The craft brewing industry is growing rapidly in Canada. Beer Canada, a volunteer trade association, reports that the number of breweries has been increasing at more than 8% per year over the last five years with most being small local businesses [1]. The Conference Board of Canada was funded to examine the economic impact of the brewing industry in Canada. The resulting report [2] highlights the local nature of the business with 85% of the beer brewed in Canada being consumed locally. In addition, it is reported than Canada's most popular alcoholic beverage contributes 5.7 billion dollars in annual tax revenue to the government, 13.6 billion dollars to GDP, and provides 149,000 direct jobs. Indirect jobs provided by the industry are difficult to categorize but Canadian farmers certainly benefit with over 300,000 tonnes of malting barley sourced from Canada.

The craft brewing industry generally follows the definitions laid out by the Brewers Association where micro breweries are characterized as those having less than 15000 barrels of production with over 75% sold offsite and brewpubs having a similar production but with more than 25% of product sold onsite[3].

The marketing and sales of craft beer typically builds on consumer demand for local hand-crafted products. This relationship was explored by Long et al [4] through a survey conducted at an Oklahoma craft beer festival in 2016. This work evaluated consumers' perception of locality and used data science methodologies to build the profile of various craft beer consumers. The study found that US consumers viewed beer produced in the same state as local and it seems likely that this would translate well to Canadian provinces and territories. The authors applied K-means cluster analysis on lifestyle and demographic information to identify five clusters of beer consumers. The 'learners' cluster was further identified as the most likely to prefer and consume craft beer and as an appropriate target for smaller local breweries. Generally speaking, this cluster was shown to have a statistically significant desire to support the local economy, trying something new, and consume something local. Demographically the cluster tended to be younger (21-40), highly educate (undergraduate or post-graduate studies), and within the higher ranges of annual household income (over 60,000 USD).

Business problem

Given the large and increasing number of competitors careful consideration is required in selecting a location for a new craft beer venue. Potential regions must be evaluated to have the demographic and neighbourhood characteristics required to support a successful business. To that end the business problem may be expressed as follows:

> *Identify regions within Canada that are underserved by the craft brewing industry and would support a new microbrewery or brewpub*

Stakeholders

This report is targeted at new entrants, such as our fictional client Capstone Brewing, or existing businesses in the industry looking to expand. Stakeholders would include the brewery management teams, those providing financing to the business, and community business development organizations. The findings could also be used by various levels of government to develop programs in support of job creation and increasing tax revenue. Barley farmers or trade associations may also be interested in the growth potential of the industry.

Proposed analytical approach

Information on existing breweries will be evaluated to determine their location and the nature of surrounding venues within walking distance. The number and distance of other breweries within a region will be determined as well as the demographic characteristics of these regions. The correlation of these features to various performance metrics will be explored.

Clustering (K-means) will be applied to characterize the current regions/neighbourhoods for breweries. If the performance metrics are shown to be significant, classification (decision tree) will be used with these labels to highlight a subset of the more successful breweries. Should this not prove to be the case the consumer profile provided in the background material will be applied. Finally, these outcomes will be applied to determine regions which share the desired characteristics that would support the addition of a new microbrewery or brewpub.

Data

The data requirements, sources, collection, and cleaning will be discussed in the following section.

Requirements

A variety of data is required to address the business problem. This will include the following;

- A list of Canadian brewers with their geolocation and venue information
- A list of venues within walking distance of these locations and their characteristics such as venue category
- The number and distance between breweries within regions from which they can expect to draw onsite customers
- Demographic information for regions from which they can expect to draw onsite customers
- Performance metrics for breweries such as number of stars, likes, ratings, and longevity

Sources

No one source of data can meet the requirements outlined above. While this work will draw heavily on the Foursquare several other sources will be exploited including the website RateBeer.com and Statistics Canada open data.

Foursquare

Foursquare is a "search-and-discover mobile app" built on a crowd sourced database of location data. The database information can be accessed through a developer API. The latitude and longitude of venues is linked to various characteristics and user reported ratings. The API will provide the record for a specific venue or collect and provide the data of various venues within a specified distance of a point of interest.

Rate Beer

The initial exploration of foursquare breweries and brewpub data indicated that users had labelled several venues as meeting these categories when, in fact, the businesses did not actually brew their own beer. To address this deficiency a comprehensive list on the website ratebeer.com was scraped to obtain a more accurate listing of Canadian brewers.

Statistics Canada Census Data

Statistics Canada gathers and provides open datasets of Canadian census data. The latest census available is from 2016 and this demographic data is made available for various geographic areas. One of these regional subsets is Forward Sortation Areas (FSA). These FSA are defined by the first three characters of Canadian postal codes. The census data for these areas is accessed though Statistics Canada's Web Data Service for which a resource URL is specified and a JSON response can be obtained.

Statistics Canada Geo Data

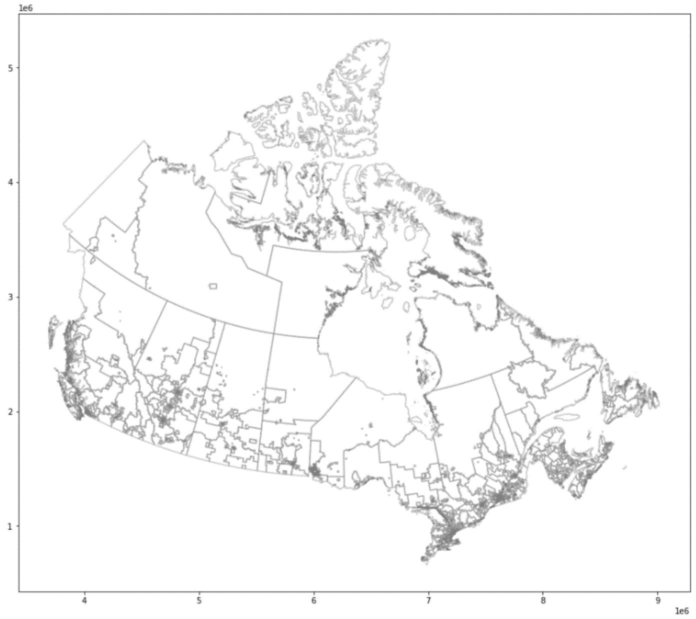
To visualize the FSA geodata is required describing their boundaries. These boundaries are available through Statistics Canada and in this instance were downloaded from the Statistics Canada website as a shapefile suitable for plotting using various Python libraries.

Collection and cleaning

The collection and cleaning of the required data is described below.

FSA boundaries

The boundaries of the FSA have been obtained as shape files from Statistics Canada Geo Data through their open data portal. In addition to the geometry(‘geometry’) each region includes the FSA identifier (‘CFUSID’) and the name of the province it is associated with (‘PRNAME’). The boundaries cover the entire country and are plotted below as a shape file using Python. Note the coordinate reference frame is based in meters (EPSG:3347) rather than degrees (EPSG:4326) such as used by folium. This is easily changed using Python libraries depending on the required operations. Care must be taken to track the current reference frame while using the data. No cleaning was required for this dataset.



List of Canadian breweries

The list of Canadian breweries was scraped from Ratebeer.com using beautiful soup. The resulting dataset included the establishment name(‘name’), type(’type’), community(‘community’), the date established(‘established’), and the number of beers that had been rated for the business(‘brews’). The brewery ‘type’ variable was used to filter the results to microbreweries and brewpubs removing large commercial brewers and others business types not relevant to this work. The resulting dataset included 1125 breweries which is consistent with literature from Beer Canada [1]. Two minor spelling errors were identified in reviewing the data and were corrected. The date the brewery was established was used to determine the current years in operation for each venue (‘years_operating’)

Foursquare data for Canadian breweries

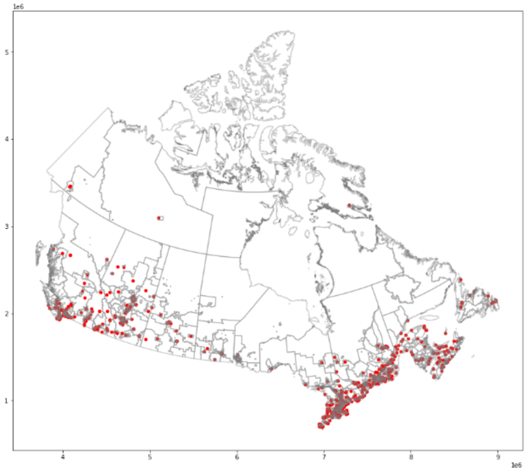
The Foursquare API was then used to search for venue information on the list of Canadian breweries gathered above. From the 1125 breweries on the list data was returned for 1064 venues. An examination of the establishments that were not found revealed that many were permanently closed, were chains with no community address, did not have data in Foursquare, or had no user information in Foursquare. For the purposes of this work only the venues with entries in Foursquare were considered. The features of the venue data are summarized with their variable name in the table below.

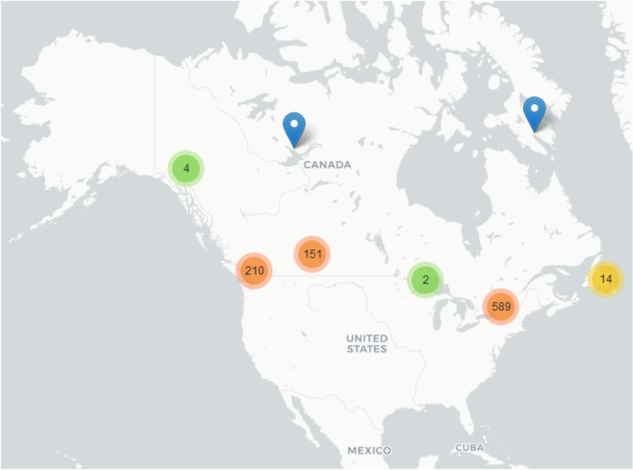
Variable	**Description**
fs_id	The foursquare ID for the venue
fs_name	The name listed in foursquare
fs_lat	The locations latitude
fs_lng	The locations longitude
fs_postalcode	The postal code of the venue
fs_city	The venues' city or town
fs_province	The venues' province

From the venues found in Foursquare a number of duplicate entries were identified which were dropped resulting in 977 records. Finally, a review of the brewery location using 'fs_province' identified a few odd US breweries had been included from California, Vermont, New York, Michigan, and Washington state. These venues were also dropped from the dataset with 972 records remaining.

Recall the FSA were defined by the first three characters of the postal code. From the 972 records 145 were found to have incomplete, invalid, or no entry for their postal code. Where available the FSA ('FSA') was collected from the postal code. To address the remaining records a shapely point was defined from the latitude and longitude provided by foursquare for the venues. The shapefile geometries for each of the FSA were then cycled through for each point to determine its associated FSA.

From this dataset every brewery under consideration has been plotted using shapely and as an interactive folium map with clustering labels. The red points in the plot on the left below are each brewery and a snapshot of the dynamic folium map is shown on the right.





Foursquare data for venues within walking distance of Canadian breweries

The Foursquare API was then used to search for information on venues in the neighbourhood surrounding Canadian breweries. The data from up to 50 venues were gathered around each brewery. Walking distance is typically considered to be 400 to 500 meters [5]. Based on this definition, venues within 500 meters of each brewery were characterized. This identified 12362 venues associated with the 972 breweries. The data gathered includes the Venue name ('Venue'), latitude ('Venue Latitude'), longitude ('Venue Longitude'), and category ('Venue Category'). In 81 instances there were no surrounding venues identified by foursquare. This is not overly surprising as it is not uncommon for breweries to be found in light industrial areas which likely have few foursquare venues. The surrounding venues in this instance were coded with a venue category of "Nothing" rather than NaN so as not to lose this information. Overall, 421 unique venue categories were identified. One hot encoding was applied to the resulting dataset and the top ten venue categories for each brewery were identified.

Canadian breweries nearest neighbouring competitors

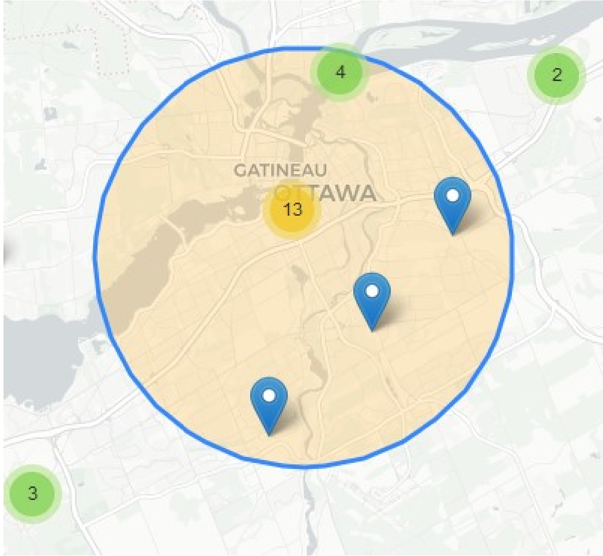
The region of interest (ROI) for each brewery with regards to competitors was chosen to be a 10 km radius representing an approximate 15-minute drive to the venue. Breweries within the ROI were determined for each individual brewery and stored as the feature 'NN' with the name and distance stored in a list. This information was them compiled to provide the number of surrounding breweries ('num_NN'), the average distance to these breweries ('average_NN'), and the median distance ('median_NN') for each of the breweries.

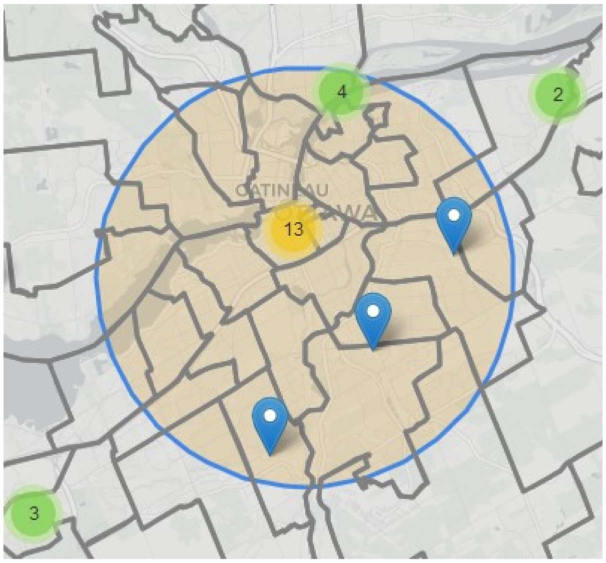
Demographic features

Census data was gathered for each of the 1620 FSA. Only the demographic data related to the customer profile described in the Introduction above were gathered. The feature list is provided below;

Feature name	**Description**
GEO_ID	Forward Sortation Area (FSA)
20 to 84 year olds	Number of 20 to 84 year olds in FSA
20 to 29 year olds	Number of 20 to 29 year olds in FSA
30 to 49 year olds	Number of 30 to 49 year olds in FSA
50 to 64 year olds	Number of 50 to 64 year olds in FSA
65 to 84 year olds	Number of 65 to 84 year olds in FSA
20 to 49 year olds	Number of 20 to 49 year olds in FSA
median age	Median age in FSA
single	Number of single people over the age of 16 in FSA
married no kids	Number of married couples with no kids
Income 35-	Number of households with income under 35k
Income 35-59	Number of households with income between 35 and 59k
Income 60-79	Number of households with income between 60 and 79k
Income 80-99	Number of households with income between 80 and 99k
Income 100+	Number of households with income over 100k
median income	Median income in FSA
households over median income	Number of households over 2016 national median income
high school or less	Number of people with education of high school or less
college	Number of people with college education
university	Number of people with university education

Having determined demographic data for each FSA, a method must be developed to determine results for a ROI surrounding individual breweries. Similar to looking for neighbouring breweries, a 10 km radius is considered around each brewery as the area from which it will draw customers.





Statistics Canada defines urban regions as those having population densities in excess of 400 persons per square kilometer [6]. From this information each FSA is evaluated as urban or rural. Rural FSA tend to be larger and have clusters of population while urban FSA tend to have a more uniform distributions of population. Place theory [7] suggests rural populations are also willing to travel further to access higher value-added businesses. Based on these considerations the following methodology is applied. When urban FSA are within a ROI the proportion of population attributed to the ROI is based on the percentage of the FSA area within the ROI. For rural FSA the entire population is attributed to the ROI for FSA within or partially within the ROI. For median or average values associated with FSA the value for the ROI is determined by multiplying the FSA values by the population of the FSA associated with the ROI, summing all of the results, and dividing by the total population attributed to the ROI.

Brewery performance metrics

Detailed calls to the Foursquare API were used to determine performance metrics for each of the breweries. This includes the following features;

Feature	**Description**
fs_tip_count	The number of tips provided by Foursquare users
fs_price_tier	The average price tier reported by Foursquare users
fs_likes	The number of Foursquare users likes
fs_rating	The average rating of Foursquare users for the brewery
fs_total_ratings	The number of Foursquare users that provided a rating for the brewery

The ‘years_operating’ feature calculated above when the list of Canadian breweries was gathered is also considered as a potential performance metric. While breweries which have failed to thrive may have closed and are not included in the dataset their successful competitors can be identified by the number of years they have been operating.

References

[1] Beer Canada , https://industry.beercanada.com/statistics , accessed 28 July 2021.

[2] Hermus, G., “Brewing Up Benefits: The Economic Footprint of Canada’s Beer Economy”, The Conference Board of Canada, Ottawa, Jan 2018.

[3] Brewers Association, https://www.brewersassociation.org/, accessed 28 July 2021.

[4] Long, J., et al, “Craft Beer Consumers’ Lifestyles and Perceptions of Locality”, International Journal of Hospitality Beverage Management, Vol 2, No 1, Article 1., doi https://dx.doi.org/10.34051/j/2019.5.

[5] Bankrate, https://www.bankrate.com/glossary/w/walking-distance/, accessed 28 July 2021.

[6] Statistics Canada, https://www.statcan.gc.ca/eng/subjects/standard/pcrac/2016/introduction, accessed 28 July 2021.

[7] Steif, Ken. 2013. "Why Do Certain Retail Stores Cluster Together?", Planetizen, Oct 2013.