Capstone Report:

Using machine learning to decide where to start a Brewery & Organic Farm in British Columbia

Introduction

Life in Vancouver, BC, Canada is beautiful, close to many man-made and naturally occuring amenities. However, real estate prices are the most unaffordable in the Country (alongside Toronto). It is a great place to live, but not a good place to start a new family; with one parent on parental leave it is very difficult to support a family on a single income for much time. In addition, living spaces can be cramped and access to private land where you can cultivate your own garden is very expensive to come by.

My husband and I have lived in Vancouver for the past 5 years. I work for a startup company and he works for the city of Port Coquitlam as an arborist. We have recently had a son and would like to consider moving to an area in BC that is more affordable where our son would be able to interact with nature more frequently on our property.

We are both business minded and are interested in starting up a business with friends to run an organic farm and set up a brewery. Our friends have 2 years experience running their own organic plot on an acre of land and a number of years working for the UBC farm prior to that. Collectively we have 0 experience in brewing but we have some great business experience to rely on and are all hard workers.

Our biggest business problem is choosing the best city to build these businesses in that also agrees with our lifestyle. This decision is critical and could be the difference between success and failure. We would need land that is appropriate for farming, and ideally would give us access to a long growing season. As for a brewery, we would need a population that would support it AND we would not want to be competing with many other breweries in the area. Most practically, we need to be able to afford purchasing a piece of land that would support these activities. Lastly, lifestyle is extremely important to us and we would like to have access to hiking trails.

The goal of this project is to secure a list of cities where a farming/brewery business would be opportunistic alongside reasonable real estate prices.

Data

The final result is a list of 5 cities for consideration based on the following:

1. Housing costs
2. Population size
3. Access to trails (important lifestyle attribute for us)
4. Number of breweries, farms and if they have a farmers market in the city (All these are searchable 'Venue Categories' in Foursquare based on https://developer.foursquare.com/docs/build-with-foursquare/categories/)

Data Sources

In order to solve this problem I required 4 main datasets

1. Housing prices in BC
   Source: Statistics Canada and downloaded a .csv containing median housing price for cities in BC, ON, and NS. [1]
2. Population sizes in BC
   Source: Government of BC website, and downloaded two .csv files containing population sizes for rural and metropolitan areas respectively. [2]
3. Number of breweries, farms, farmers markets, and trails in BC
   Source: Foursquare
4. Growing season length in BC
   Source: Scrape websites for average first day of frost [3], and average last day of frost [4]

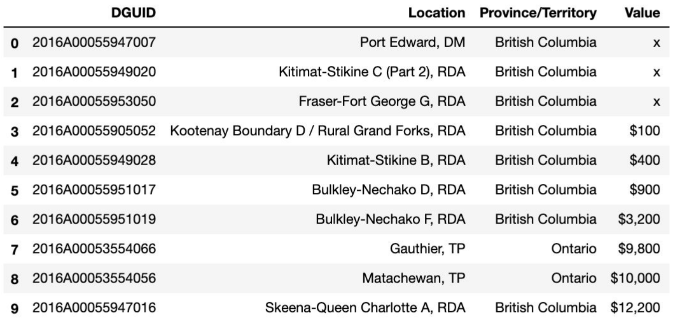
Data Cleaning

There was a lot of data cleaning required in order to complete this project because the data was taken from many different sources and was in many different formats. Data cleaning, reorganizing and merging was necessary for all 4 data sources.

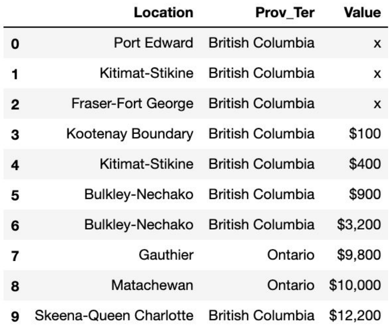
Median Housing Prices in BC

Some common issues with the data is that it was very important to have the city name accurate and standardized because this was going to be the value that was used to merge datasets together. This meant splitting cells with city names that had commas or dashes or combined two cities together.

An example of this was when parsing through the data extracted from the housing price .csv, commas, and spaces were used to split cells because it was common for a region initial to be included alongside a city name within the same cell. Once the cells were split, a loop was created to determine which cells had initials for regions (ie string value of 1 or less) or they were a legitimate second word to the city name (ie. North Vancouver).



	DGUID	Location	Province/Territory	Value
0	2016A00055947007	Port Edward, DM	British Columbia	x
1	2016A00055949020	Kitimat-Stikine C (Part 2), RDA	British Columbia	x
2	2016A00055953050	Fraser-Fort George G, RDA	British Columbia	x
3	2016A00055905052	Kootenay Boundary D / Rural Grand Forks, RDA	British Columbia	$100
4	2016A00055949028	Kitimat-Stikine B, RDA	British Columbia	$400
5	2016A00055951017	Bulkley-Nechako D, RDA	British Columbia	$900
6	2016A00055951019	Bulkley-Nechako F, RDA	British Columbia	$3,200
7	2016A00053554066	Gauthier, TP	Ontario	$9,800
8	2016A00053554056	Matachewan, TP	Ontario	$10,000
9	2016A00055947016	Skeena-Queen Charlotte A, RDA	British Columbia	$12,200



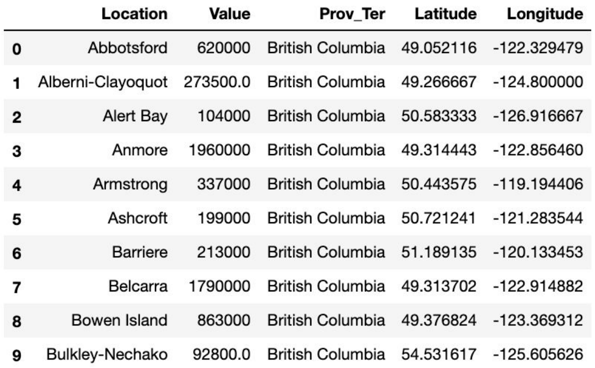
	Location	Prov_Ter	Value
0	Port Edward	British Columbia	x
1	Kitimat-Stikine	British Columbia	x
2	Fraser-Fort George	British Columbia	x
3	Kootenay Boundary	British Columbia	$100
4	Kitimat-Stikine	British Columbia	$400
5	Bulkley-Nechako	British Columbia	$900
6	Bulkley-Nechako	British Columbia	$3,200
7	Gauthier	Ontario	$9,800
8	Matachewan	Ontario	$10,000
9	Skeena-Queen Charlotte	British Columbia	$12,200

Rows with no information regarding housing prices were deleted and then the dataset was sliced by 61% to only include cities in BC.

The second challenge the housing price data presented was that there were many duplicates represented in the dataset with different median housing prices for varying nuances within the region/city. It was decided to extract duplicates, compute the median house price within regions with duplicates, and then re-add these median prices to the original dataset so there were no duplicates. Since the duplicates represented 41% of the overall dataset, it was important to deal with them.

The next issue that came up was when utilizing the geocoder function to procure latitude and longitude values for the cities (to be utilized in Foursquare API call later). Some of the names were region names and not cities, these would throw an error. In addition, the geocoder API call itself was very shaky and would time out quite a bit. A loop was built such that the data could be stored in increments until through the entire city list.

The final dataset from housing price data was organized as shown below:



	Location	Value	Prov_Ter	Latitude	Longitude
0	Abbotsford	620000	British Columbia	49.052116	-122.329479
1	Alberni-Clayoquot	273500.0	British Columbia	49.266667	-124.800000
2	Alert Bay	104000	British Columbia	50.583333	-126.916667
3	Anmore	1960000	British Columbia	49.314443	-122.856460
4	Armstrong	337000	British Columbia	50.443575	-119.194406
5	Ashcroft	199000	British Columbia	50.721241	-121.283544
6	Barriere	213000	British Columbia	51.189135	-120.133453
7	Belcarra	1790000	British Columbia	49.313702	-122.914882
8	Bowen Island	863000	British Columbia	49.376824	-123.369312
9	Bulkley-Nechako	92800.0	British Columbia	54.531617	-125.605626

Population Sizes in BC

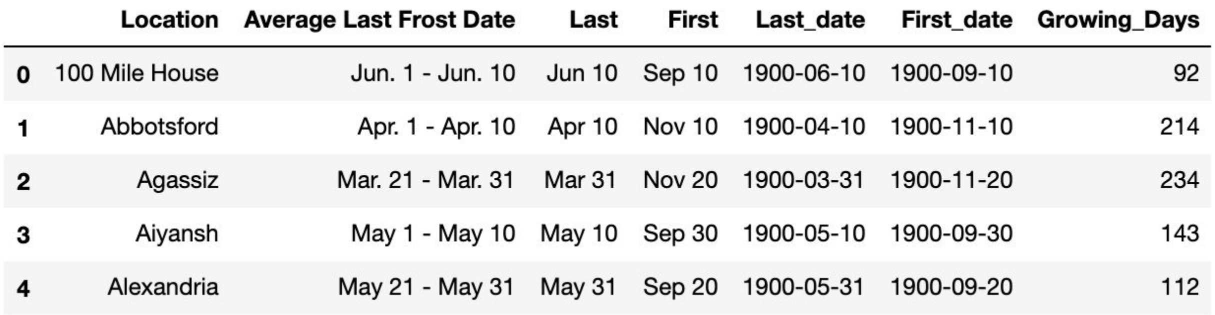
This data set was quite a bit more clean than the previous and all that was required was to merge the municipal and metro centre data sets together, and remove duplicates.. The only issue that was encountered were preceding spaces within the population dataset on the city names that did not perform the merge properly because the location names did not match. Once the preceding spaces were stripped from the city name, the merge worked. This data set contained population counts from 2011 to 2019 and the population changes in 198 BC cities.

Growing Season Length

This was a datapoint that was important to the analysis but not easy to come by accurate information or a large set of data.

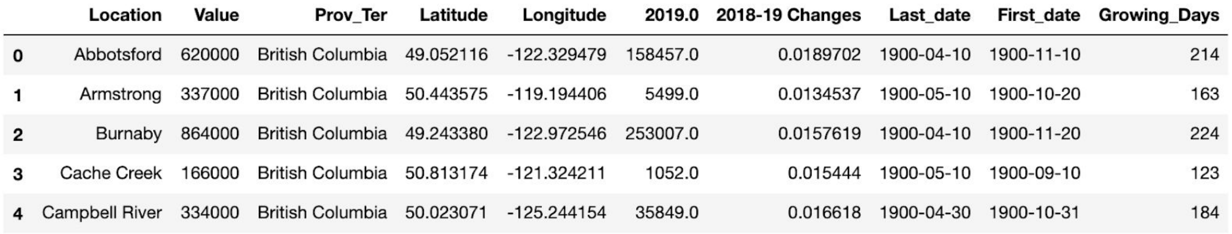
Ultimately a website containing a range of dates for the first and last days of frost in various BC cities was found. The Pandas read_html function was utilized in order to scrape this website data and store it in a dataframe. The last step was reformatting the string containing the date to a time date that could be used to calculate the difference in days between the first day of frost and the previous last day of frost, otherwise known as the growing season.

The following dataset was stored containing the growing season information for 311 BC cities.



	Location	Average Last Frost Date	Last	First	Last_date	First_date	Growing_Days
0	100 Mile House	Jun. 1 - Jun. 10	Jun 10	Sep 10	1900-06-10	1900-09-10	92
1	Abbotsford	Apr. 1 - Apr. 10	Apr 10	Nov 10	1900-04-10	1900-11-10	214
2	Agassiz	Mar. 21 - Mar. 31	Mar 31	Nov 20	1900-03-31	1900-11-20	234
3	Aiyansh	May 1 - May 10	May 10	Sep 30	1900-05-10	1900-09-30	143
4	Alexandria	May 21 - May 31	May 31	Sep 20	1900-05-31	1900-09-20	112

To reduce the number of API calls that were made to Foursquare, the House Price, Population and Growing Season datasets were all merged together first. In total, 108 cities in BC contained information from all three of aforementioned data sets.

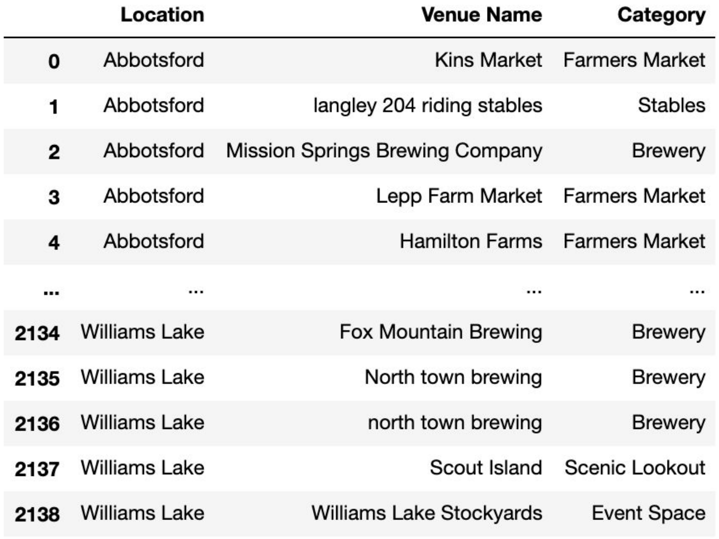


	Location	Value	Prov_Ter	Latitude	Longitude	2019.0	2018-19 Changes	Last_date	First_date	Growing_Days
0	Abbotsford	620000	British Columbia	49.052116	-122.329479	158457.0	0.0189702	1900-04-10	1900-11-10	214
1	Armstrong	337000	British Columbia	50.443575	-119.194406	5499.0	0.0134537	1900-05-10	1900-10-20	163
2	Burnaby	864000	British Columbia	49.243380	-122.972546	253007.0	0.0157619	1900-04-10	1900-11-20	224
3	Cache Creek	166000	British Columbia	50.813174	-121.324211	1052.0	0.015444	1900-05-10	1900-09-10	123
4	Campbell River	334000	British Columbia	50.023071	-125.244154	35849.0	0.016618	1900-04-30	1900-10-31	184

Foursquare Data

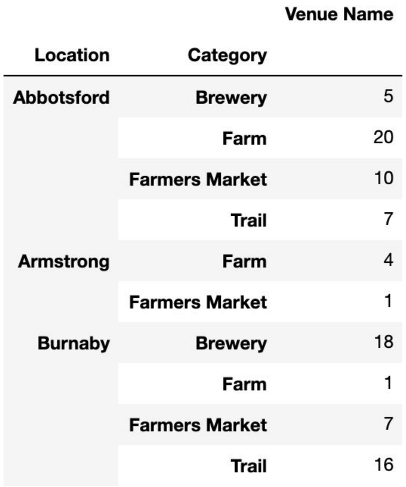
Acquiring this data and reformatting it presented some challenges. The first challenge was properly reading the resulting .json file into a usable dataframe that contained the city name, venue name, and venue category.

The resulting dataframe contained a lot of information, some not useful to the purposes of this project. This dataframe is shown below:

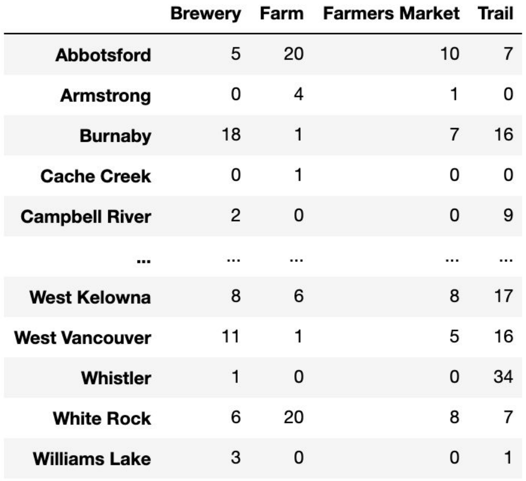


	Location	Venue Name	Category
0	Abbotsford	Kins Market	Farmers Market
1	Abbotsford	langley 204 riding stables	Stables
2	Abbotsford	Mission Springs Brewing Company	Brewery
3	Abbotsford	Lepp Farm Market	Farmers Market
4	Abbotsford	Hamilton Farms	Farmers Market
...	...	...	...
2134	Williams Lake	Fox Mountain Brewing	Brewery
2135	Williams Lake	North town brewing	Brewery
2136	Williams Lake	north town brewing	Brewery
2137	Williams Lake	Scout Island	Scenic Lookout
2138	Williams Lake	Williams Lake Stockyards	Event Space

Any venue category that was not explicitly labelled as a Trail, Brewery, Farm, or Farmers Market was excluded from further analysis. In addition, the data had to be counted and significantly reformatted in order to create one row for each city with the number of venues for each category desired.



		Venue Name
Location	**Category**	
Abbotsford	**Brewery**	5
	Farm	20
	Farmers Market	10
	Trail	7
Armstrong	**Farm**	4
	Farmers Market	1
Burnaby	**Brewery**	18
	Farm	1
	Farmers Market	7
	Trail	16



	Brewery	Farm	Farmers Market	Trail
Abbotsford	5	20	10	7
Armstrong	0	4	1	0
Burnaby	18	1	7	16
Cache Creek	0	1	0	0
Campbell River	2	0	0	9
...	...	...	...	...
West Kelowna	8	6	8	17
West Vancouver	11	1	5	16
Whistler	1	0	0	34
White Rock	6	20	8	7
Williams Lake	3	0	0	1

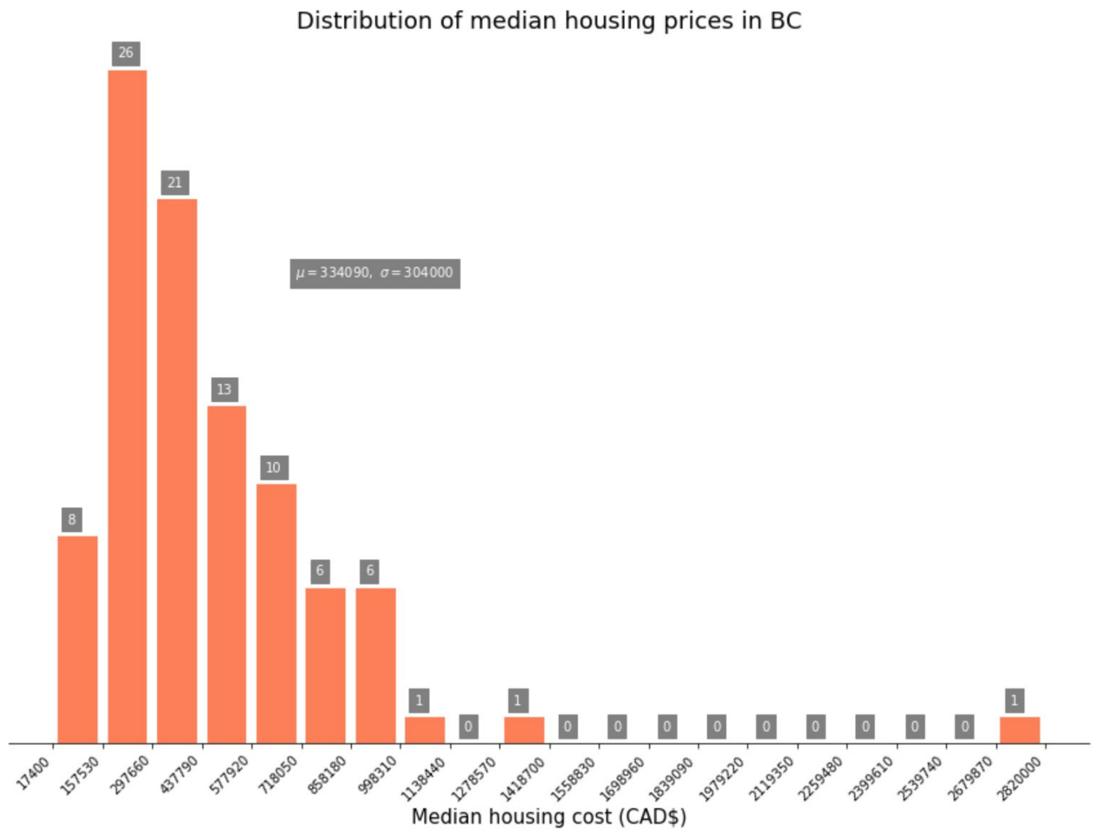
Finally, this dataset was merged together with the dataset containing housing prices, population size and growing season length. The final merged dataframe contained no null data points for 93 cities.

Methodology

Descriptive Statistics

A quick summary of the problem we are trying to solve: provide list of 5 cities that would be opportunistic and affordable to open a brewery and organic farm.

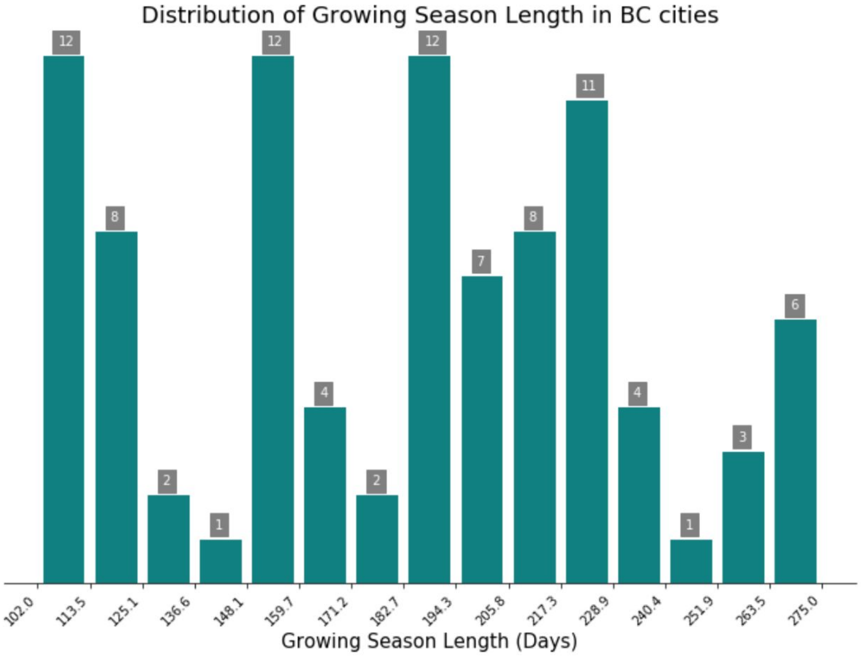
The most important element to this problem is the price of real estate in each city.   Naturally, the first step to visualize the dataset was determine how housing prices were distributed by creating a histogram of the median housing prices for each city in the database.



This shows that most housing prices in BC cities are between 160-600 thousand dollars.

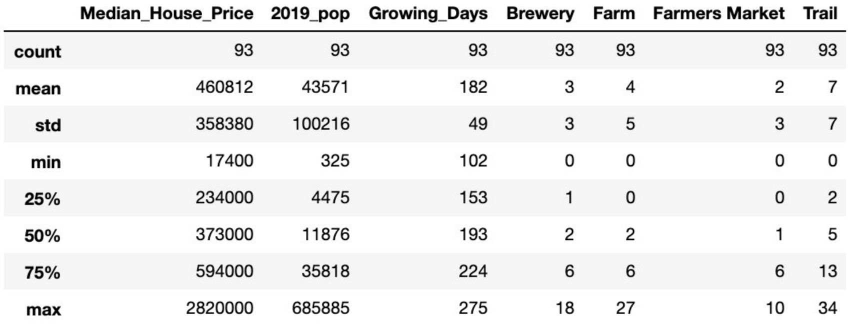
This is good news as this price range is within reach to start up a business/place of residence in contrast to those in Vancouver,BC which represents the median house price between 1.3 and 1.4 million on this chart.

The next variable of interest was the growing season length. This was important to understanding the viability of starting an organic farm or potentially growing grain(hops) for the brewery. A plot of the distribution of number of breweries in BC cities indicates that the data is not normally distributed. This may be a result of the extremely varied climates, geology and soil conditions that exist within BC. There seems to be quite a large variance of low, medium and long growing seasons in BC cities. There seems to be a good number of cities with a growing season between 190 to 230. This is good news as this kind of growing season could provide an environment where a vegetable farm could succeed.



The descriptive statistics on the full dataset is provided below:

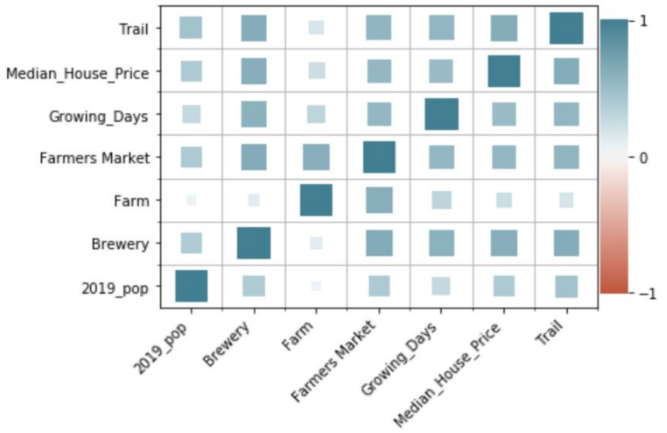
Descriptive Statistics on full dataset

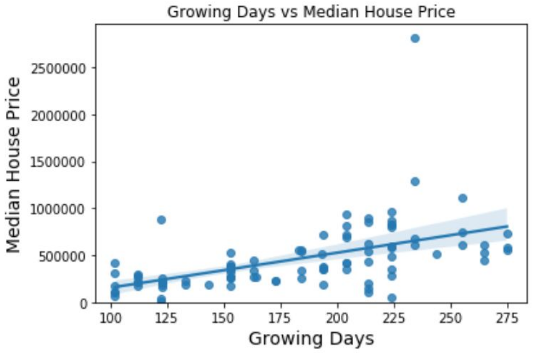


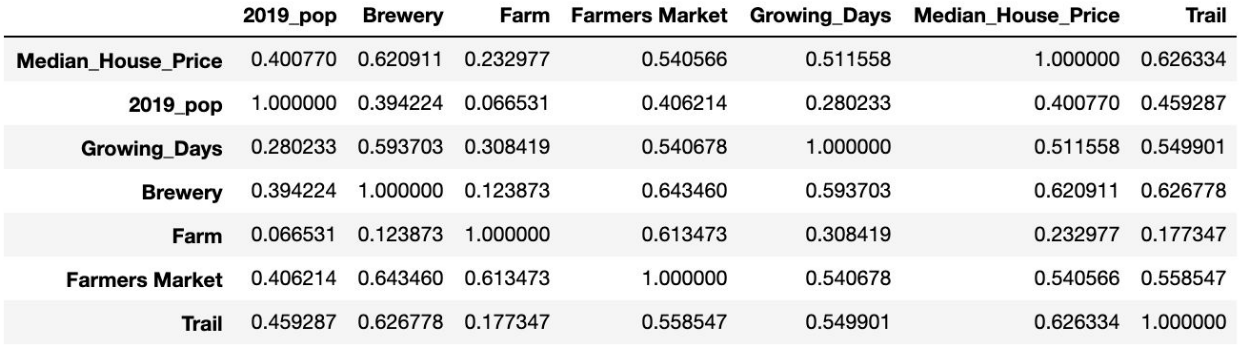
	Median_House_Price	2019_pop	Growing_Days	Brewery	Farm	Farmers Market	Trail
count	93	93	93	93	93	93	93
mean	460812	43571	182	3	4	2	7
std	358380	100216	49	3	5	3	7
min	17400	325	102	0	0	0	0
25%	234000	4475	153	1	0	0	2
50%	373000	11876	193	2	2	1	5
75%	594000	35818	224	6	6	6	13
max	2820000	685885	275	18	27	10	34

Another promising sign is that the average number of breweries in a BC city is 3. Of course it depends on the population size to understand if this is supportable but there seems to be room to add breweries in BC.

Another useful tool is to complete a correlation table for the variables to discover if there are any interesting correlations between features in this dataset. The original correlation heatmap was tough to read, so a customized correlation chart was created to depict higher correlations with larger tiles alongside colour. From the correlation plot below, there seems to be moderate correlation between the number of breweries and trails in a city with median housing price. The correlation table below the correlation heatmap confirms this and also shows that growing season length appears to be moderately correlated with median housing price. This was expected but means it would be useful to cluster cities using machine learning in order to try and group cities together by more complex means in order to find value where it is tough to see with this initial kind of analysis.

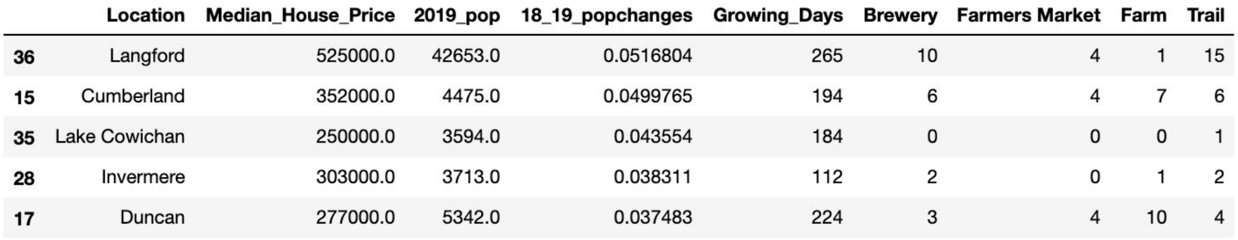






	2019_pop	Brewery	Farm	Farmers Market	Growing_Days	Median_House_Price	Trail
Median_House_Price	0.400770	0.620911	0.232977	0.540566	0.511558	1.000000	0.626334
2019_pop	1.000000	0.394224	0.066531	0.406214	0.280233	0.400770	0.459287
Growing_Days	0.280233	0.593703	0.308419	0.540678	1.000000	0.511558	0.549901
Brewery	0.394224	1.000000	0.123873	0.643460	0.593703	0.620911	0.626778
Farm	0.066531	0.123873	1.000000	0.613473	0.308419	0.232977	0.177347
Farmers Market	0.406214	0.643460	0.613473	1.000000	0.540678	0.540566	0.558547
Trail	0.459287	0.626778	0.177347	0.558547	0.549901	0.626334	1.000000

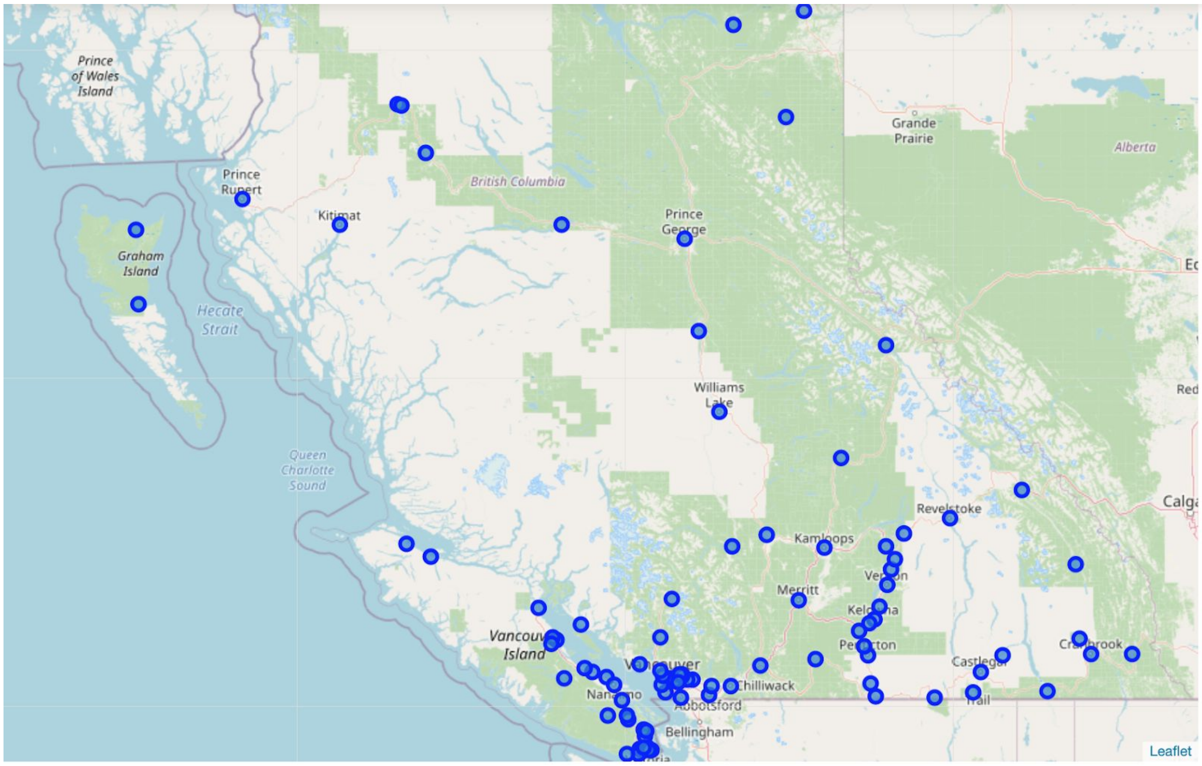
It was also of interest to note the fastest growing cities in 2019 which could be a sound real estate investment and expect those prices to climb into the future. The top 5 fastest growing cities in BC in 2019 are shown in the table below.



	Location	Median_House_Price	2019_pop	18_19_popchanges	Growing_Days	Brewery	Farmers Market	Farm	Trail
36	Langford	525000.0	42653.0	0.0516804	265	10	4	1	15
15	Cumberland	352000.0	4475.0	0.0499765	194	6	4	7	6
35	Lake Cowichan	250000.0	3594.0	0.043554	184	0	0	0	1
28	Invermere	303000.0	3713.0	0.038311	112	2	0	1	2
17	Duncan	277000.0	5342.0	0.037483	224	3	4	10	4

These insights will be returned to once the clustering analysis has been completed.

The full dataset of BC cities was plotted on a map below, with most of the cities residing in the southern quarter of the province.



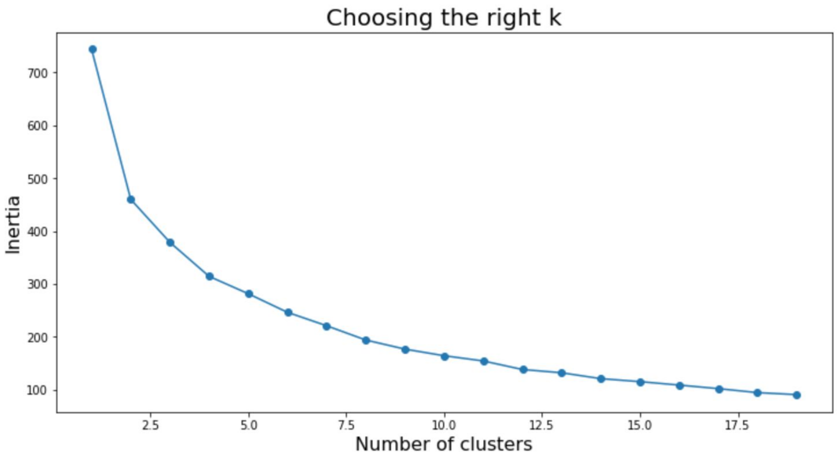
Machine Learning

Due to the number of parameters of interest (# farms, #breweries, #farmers markets, #trails, population size, growing season length, median housing price) it was useful to utilize machine learning to cluster cities into groups for further analysis if necessary.

This problem requires unsupervised machine learning because there are no labels in the dataset. It is desired to summarize key aspects of clusters for the dataset and perform some additional descriptive analysis on clusters that represent desirable traits. It is also desirable to find outliers on both sides to determine if there are immediate suboptimal or optimal cities. Due to these requirements and the small nature of the dataset, K means was the algorithm chosen to perform the cluster analysis.

KMeans Clustering

When using Kmeans to cluster data, the first step is to determine the number of clusters required. This was done by creating a loop that plotted inertia between centroids versus the number of clusters. It is desired to choose the K value at the “elbow” of this graph. In this case, there is not a clear elbow but it appears that a value between 5 and 7 could be interpreted as an elbow.

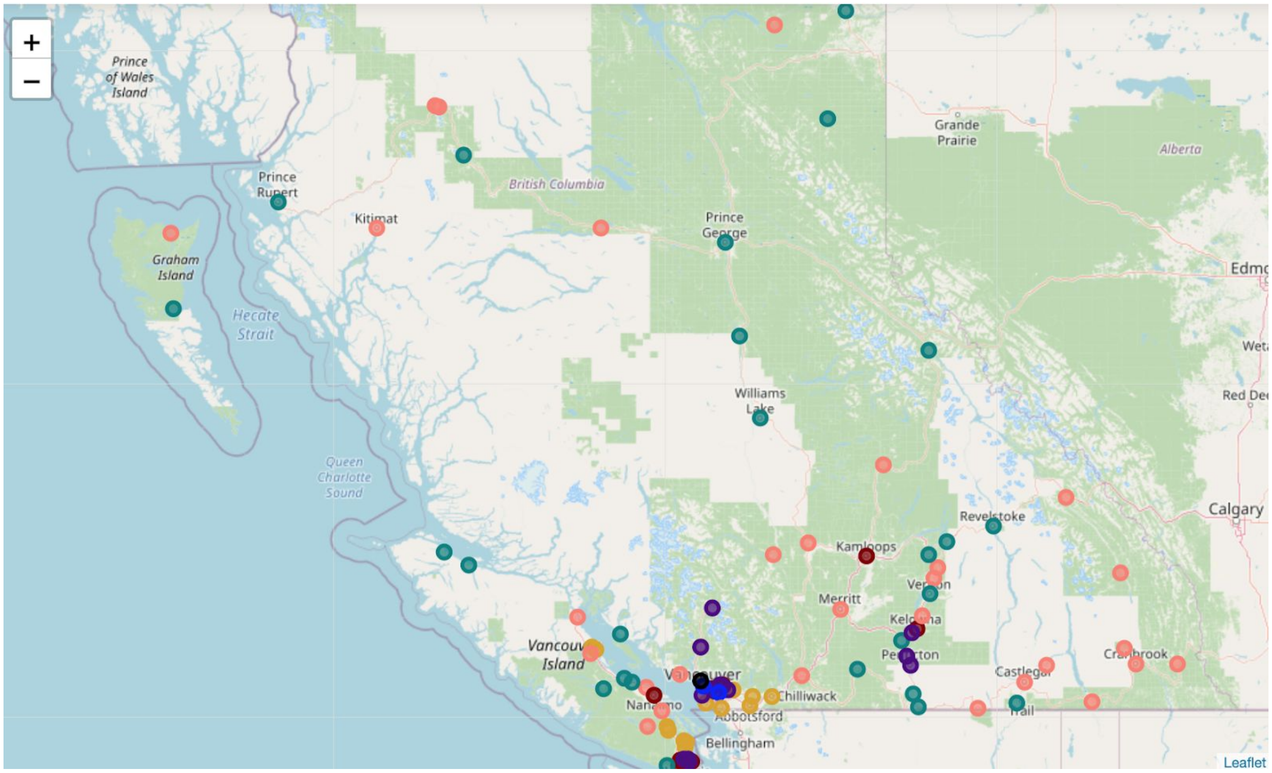


Results

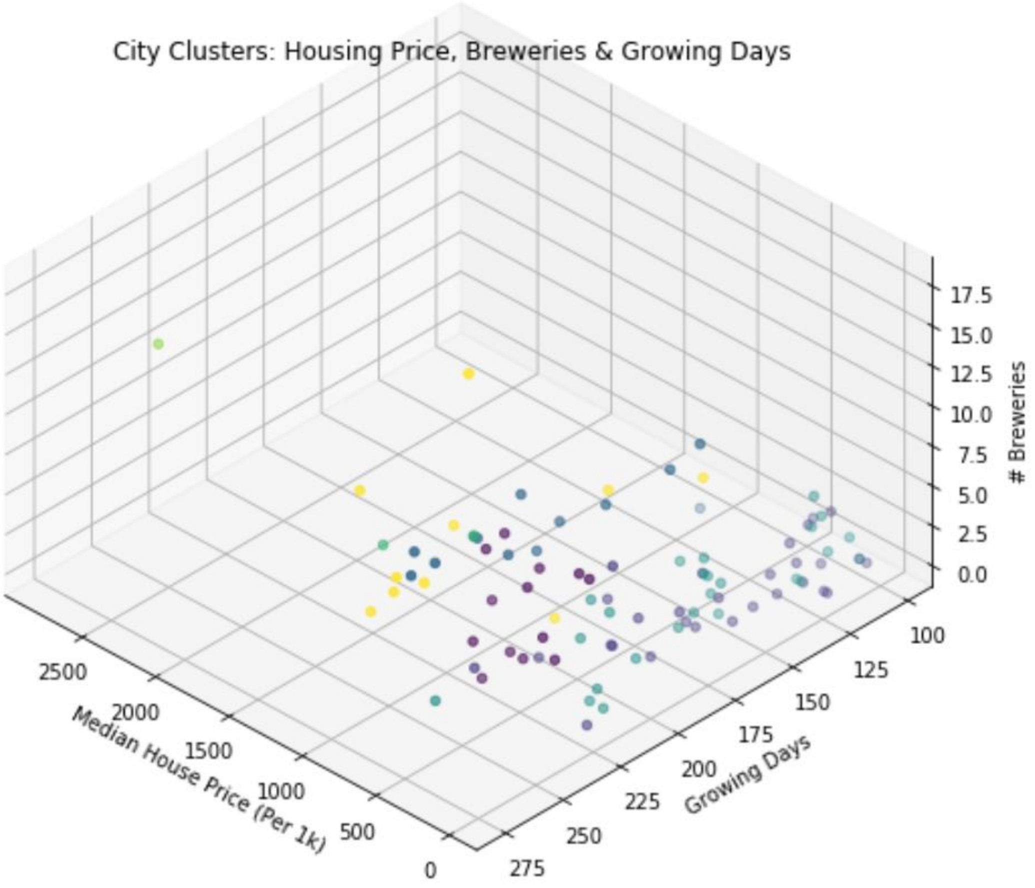
Cluster Map

A map of the cluster analysis is shown below.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
Colour	Goldenrod	Coral	Purple	Teal	Blue	Black	Maroon



Clusters can be visualized below using a 3D plot of the most critical features in the dataset.



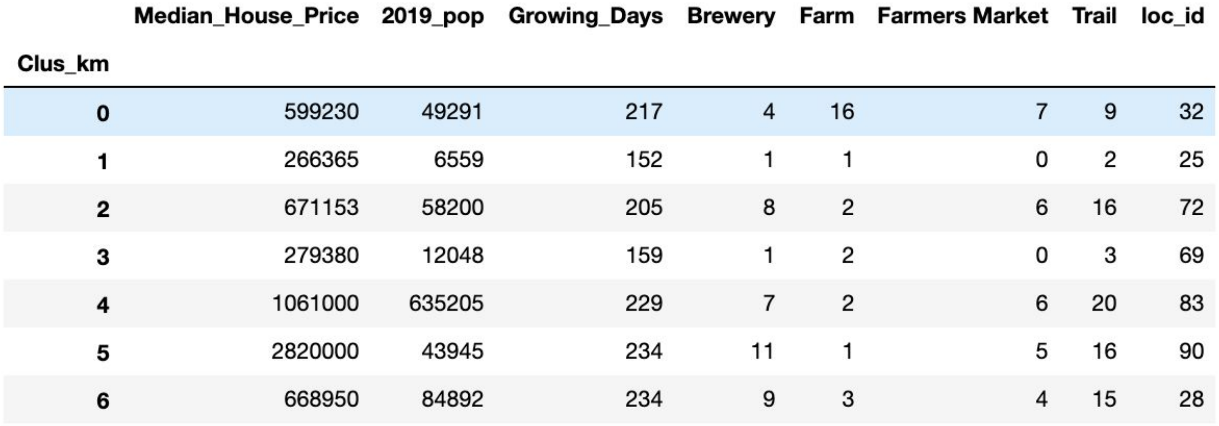
This helps visually see how the clusters are distributed for the three main data features of interest (number of breweries, median house price, and length of growing season in days). However, it is difficult to assess the differences between clusters that look close together on these axes due to the fact that other information is not being represented here - including number of farms and farmers markets, the number of trails close by and the population size.

Perhaps the simplest method to distinguish between clusters is the centroid position of each cluster.

Discussion

Cluster Centroid Position

The position of cluster centroids can give an idea on how the cities have been grouped in the table below:



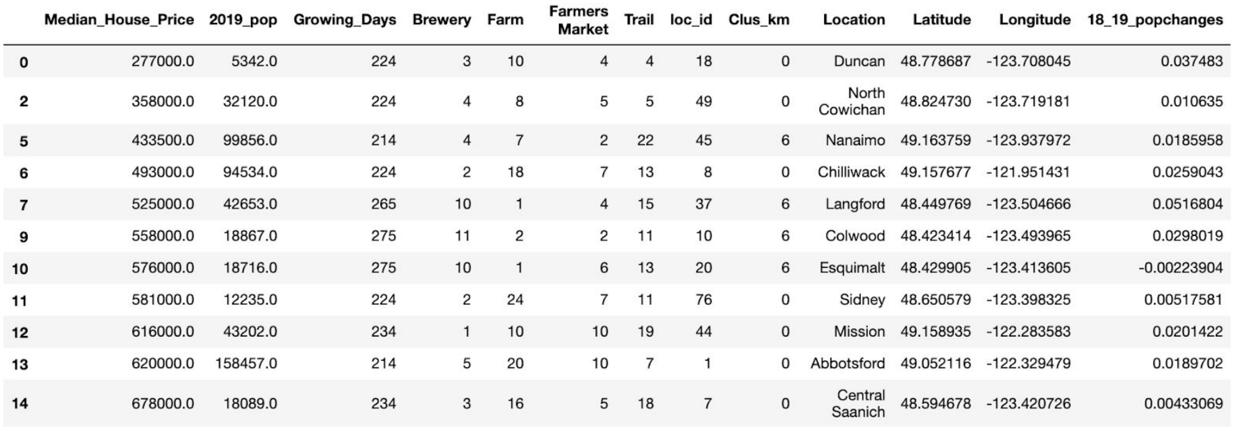
	Median_House_Price	2019_pop	Growing_Days	Brewery	Farm	Farmers Market	Trail	loc_id
Clus_km								
0	599230	49291	217	4	16	7	9	32
1	266365	6559	152	1	1	0	2	25
2	671153	58200	205	8	2	6	16	72
3	279380	12048	159	1	2	0	3	69
4	1061000	635205	229	7	2	6	20	83
5	2820000	43945	234	11	1	5	16	90
6	668950	84892	234	9	3	4	15	28

For example, it is easy to immediately eliminate cities within cluster 4 and 5 due to the high real estate prices. Additionally, cluster 1 and 3 can be eliminated due to short growing seasons. That leaves cluster's 0, 2 and 6. The differences between the 3 seem to be population size, number of breweries and number of farms. The early candidate that appears to be ideally suited to solve the business problem is cluster 0 as it has a lower number of breweries, a high number of farms and a growing season above average. However, cluster 6 also seems ideal, with a growing season at the 75th percentile, more breweries, but almost double the population to support them (compared with cluster 0). Real estate prices are slightly higher in cluster 6 (~$70 000 higher), but overall the two are comparable.

	Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
# Cities	13	29	13	25	2	1	10

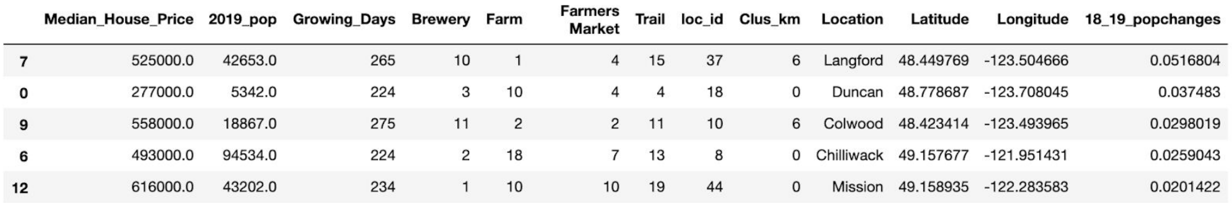
Therefore, if we take the cities within clusters 0 and 6, we have reduced our list from 93 to 23.

We will create a new dataset with these two clusters for our final list and then slim the list down to only include cities with a growing season longer than 210 days, more than 3 trails close by, and median house value less than $700 000. This brought the final list of cities down to 11.



	Median_House_Price	2019_pop	Growing_Days	Brewery	Farm	Farmers Market	Trail	loc_id	Clus_km	Location	Latitude	Longitude	18_19_popchanges
0	277000.0	5342.0	224	3	10	4	4	18	0	Duncan	48.778687	-123.708045	0.037483
2	358000.0	32120.0	224	4	8	5	5	49	0	North Cowichan	48.824730	-123.719181	0.010635
5	433500.0	99856.0	214	4	7	2	22	45	6	Nanaimo	49.163759	-123.937972	0.0185958
6	493000.0	94534.0	224	2	18	7	13	8	0	Chilliwack	49.157677	-121.951431	0.0259043
7	525000.0	42653.0	265	10	1	4	15	37	6	Langford	48.449769	-123.504666	0.0516804
9	558000.0	18867.0	275	11	2	2	11	10	6	Colwood	48.423414	-123.493965	0.0298019
10	576000.0	18716.0	275	10	1	6	13	20	6	Esquimalt	48.429905	-123.413605	-0.00223904
11	581000.0	12235.0	224	2	24	7	11	76	0	Sidney	48.650579	-123.398325	0.00517581
12	616000.0	43202.0	234	1	10	10	19	44	0	Mission	49.158935	-122.283583	0.0201422
13	620000.0	158457.0	214	5	20	10	7	1	0	Abbotsford	49.052116	-122.329479	0.0189702
14	678000.0	18089.0	234	3	16	5	18	7	0	Central Saanich	48.594678	-123.420726	0.00433069

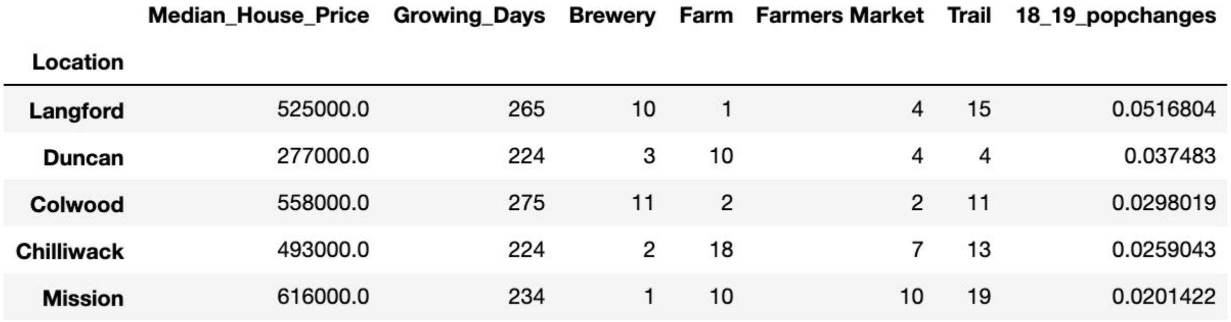
The top 5 list is generated by taking the cities with the greatest population growth to indicate a healthy real estate market.



	Median_House_Price	2019_pop	Growing_Days	Brewery	Farm	Farmers Market	Trail	loc_id	Clus_km	Location	Latitude	Longitude	18_19_popchanges
7	525000.0	42653.0	265	10	1	4	15	37	6	Langford	48.449769	-123.504666	0.0516804
0	277000.0	5342.0	224	3	10	4	4	18	0	Duncan	48.778687	-123.708045	0.037483
9	558000.0	18867.0	275	11	2	2	11	10	6	Colwood	48.423414	-123.493965	0.0298019
6	493000.0	94534.0	224	2	18	7	13	8	0	Chilliwack	49.157677	-121.951431	0.0259043
12	616000.0	43202.0	234	1	10	10	19	44	0	Mission	49.158935	-122.283583	0.0201422

Conclusions

Using machine learning cluster analysis and further refining the list of BC cities based on access to trails, real estate prices, growing season length, and population growth rate in 2019, the top 5 cities in BC that one could affordably move to and start a brewery and organic farm are:



	Median_House_Price	Growing_Days	Brewery	Farm	Farmers Market	Trail	18_19_popchanges
Location							
Langford	525000.0	265	10	1	4	15	0.0516804
Duncan	277000.0	224	3	10	4	4	0.037483
Colwood	558000.0	275	11	2	2	11	0.0298019
Chilliwack	493000.0	224	2	18	7	13	0.0259043
Mission	616000.0	234	1	10	10	19	0.0201422

This information is incredibly valuable and provides a list of opportunistic cities in British Columbia to open a brewery and organic farm where real estate is somewhat affordable.

References

[1] Housing Prices in BC from Statistics Canada
https://www150.statcan.gc.ca/n1/en/subjects/housing/housing_costs_and_affordability

[2] Population sizes of rural and metropolitan areas, Government of BC
https://www2.gov.bc.ca/gov/content/data/statistics/people-population-community/population/population-estimates

[3] Range of dates for the last day of frost for cities in BC
https://www.plantmaps.com/interactive-british-columbia-first-frost-date-map.php

[4] Range of dates for the last day of frost for cities in BC
https://www.plantmaps.com/interactive-british-columbia-last-frost-date-map.php