Capstone Project - Evaluating the best Town in Ontario, Canada to immigrate to

Shaun Diplock

22nd March 2021

1. Introduction

1.1. Background

The problem / challenge addressed in this report and study is a very personal one; the target audience is myself and my future wife and family. Therefore, this 'business problem' is an unconventional one, as I performed this study for myself as the main stakeholder. I have attempted to keep personal details and information minimal throughout this report, however as a natural consequence of the subject matter I occasionally reference items of personal interest.

I met my fiancé over 5 years ago in Ottawa, Canada, whilst I was working in the area on a business trip. I live in England, and work as a system development manager and engineer, and frequently must travel abroad for client site visits and business meetings. Our long-distance relationship has developed to the point where we are now engaged, and I am actively trying to immigrate to Canada so we can seriously start our life together.

Moving to another country is a daunting prospect, and one that merits time and research into identifying the best area for both of us. Naturally, I am very anxious (but also excited) for what the future holds - this project represents a very real, genuine attempt to evaluate some areas that will maximise the chance of our move and choice being a success.

1.2. Problem

The minimum criteria that the area must meet is as follows:

1. Within a 60-minute drive (approximately 70 km) of Smiths Falls, Ontario (this is where my fiancé currently works).
2. We do not want to live in Quebec.
3. We cannot live in the United States (Smiths Falls is relatively close to the border).
4. We do not want to live in a very small town - the town / city must contain more than 2,500 residents.
5. We do not want to live in a major city or urban area - the town / city must contain less than 50,000 residents.

Providing the above minimum criteria are met, towns / cities / areas are ranked using the following attributes:

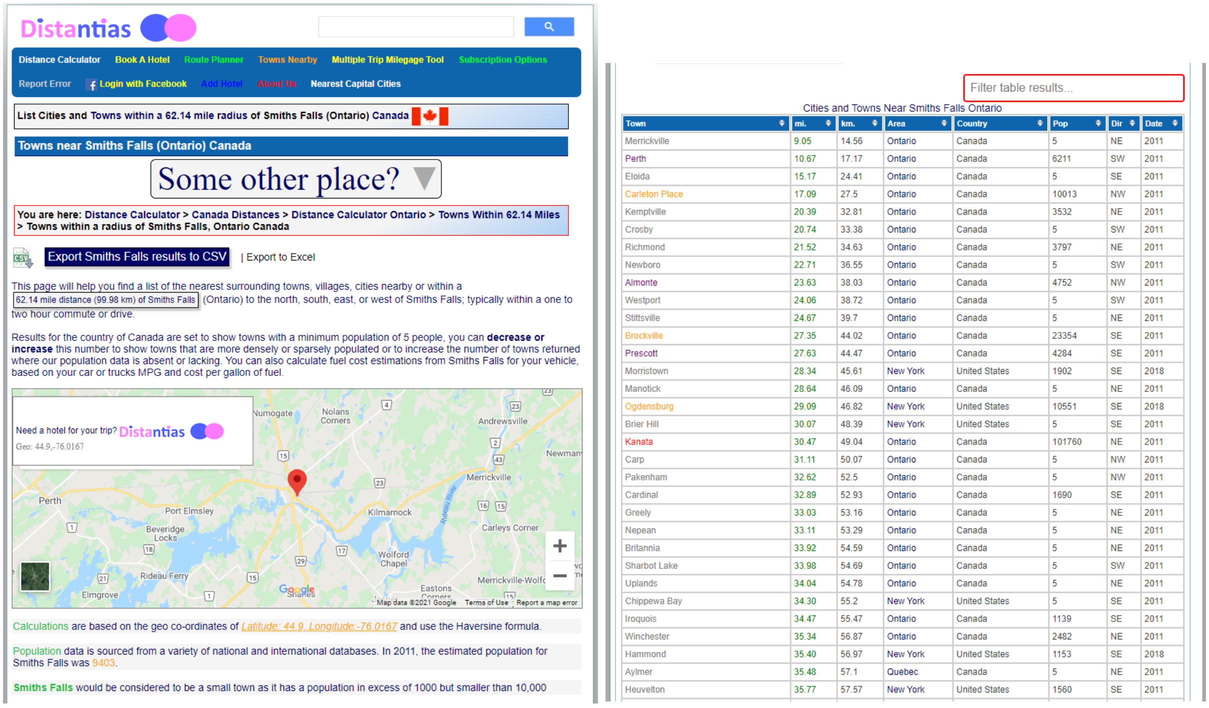
1. The amount of restaurants in the town / city (excluding fast food restaurants) - high priority.
2. The amount of bars in the town / city - high priority.
3. The amount of gyms / fitness studios in the town / city - medium priority.
4. The amount of entertainment venues in the town / city - medium priority.
5. The amount of outdoor spaces, such as parks and trails in the town / city - medium priority.
6. The amount of shopping outlets and retail stores - low priority.
7. The distance from Smiths Falls, Ontario - high priority.

With the above criteria evaluated, I hoped to be able to identify some suitable areas for us to move to when I immigrate, to maximise the chance of our future life together being happy and successful.

2. Data Acquisition and Preparation

2.1. Data Sources

The source data for this problem was acquired from Distantias's location proximity tool. This website provides an easy-to-use tool which can quickly search for all towns, cities, and habituated places in proximity to another town or place. It then provides the location data in a convenient .csv file which was easy to process for the study.



List Cities and Towns within a 62.14 mile radius of Smiths Falls (Ontario) Canada

Towns near Smiths Falls (Ontario) Canada

You are here: Distance Calculator > Canada Distances > Distance Calculator Ontario > Towns Within 62.14 Miles > Towns within a radius of Smiths Falls, Ontario Canada

Export Smiths Falls results to CSV | Export to Excel

This page will help you find a list of the nearest surrounding towns, villages, cities nearby or within a 62.14 mile distance (99.98 km) of Smiths Falls (Ontario) to the north, south, east, or west of Smiths Falls; typically within a one to two hour commute or drive.

Results for the country of Canada are set to show towns with a minimum population of 5 people, you can **decrease or increase** this number to show towns that are more densely or sparsely populated or to increase the number of towns returned where our population data is absent or lacking. You can also calculate fuel cost estimations from Smiths Falls for your vehicle, based on your car or trucks MPG and cost per gallon of fuel.

Need a hotel for your trip? Distantias

Geo: 44.9,-76.0167

Calculations are based on the geo co-ordinates of *Latitude: 44.9, Longitude:-76.0167* and use the Haversine formula.

Population data is sourced from a variety of national and international databases. In 2011, the estimated population for Smiths Falls was 9403.

Smiths Falls would be considered to be a small town as it has a population in excess of 1000 but smaller than 10,000

Cities and Towns Near Smiths Falls Ontario

Town	mi.	km.	Area	Country	Pop	Dir	Date
Merrickville	9.05	14.56	Ontario	Canada	5	NE	2011
Perth	10.67	17.17	Ontario	Canada	6211	SW	2011
Eloida	15.17	24.41	Ontario	Canada	5	SE	2011
Carleton Place	17.09	27.5	Ontario	Canada	10013	NW	2011
Kemptville	20.39	32.81	Ontario	Canada	3532	NE	2011
Crosby	20.74	33.38	Ontario	Canada	5	SW	2011
Richmond	21.52	34.63	Ontario	Canada	3797	NE	2011
Newboro	22.71	36.55	Ontario	Canada	5	SW	2011
Almonte	23.63	38.03	Ontario	Canada	4752	NW	2011
Westport	24.06	38.72	Ontario	Canada	5	SW	2011
Stittsville	24.67	39.7	Ontario	Canada	5	NE	2011
Brockville	27.35	44.02	Ontario	Canada	23354	SE	2011
Prescott	27.63	44.47	Ontario	Canada	4284	SE	2011
Morristown	28.34	45.61	New York	United States	1902	SE	2018
Manotick	28.64	46.09	Ontario	Canada	5	NE	2011
Ogdensburg	29.09	46.82	New York	United States	10551	SE	2018
Brier Hill	30.07	48.39	New York	United States	5	SE	2011
Kanata	30.47	49.04	Ontario	Canada	101760	NE	2011
Carp	31.11	50.07	Ontario	Canada	5	NW	2011
Pakenham	32.62	52.5	Ontario	Canada	5	NW	2011
Cardinal	32.89	52.93	Ontario	Canada	1690	SE	2011
Greely	33.03	53.16	Ontario	Canada	5	NE	2011
Nepean	33.11	53.29	Ontario	Canada	5	NE	2011
Britannia	33.92	54.59	Ontario	Canada	5	NE	2011
Sharbot Lake	33.98	54.69	Ontario	Canada	5	SW	2011
Uplands	34.04	54.78	Ontario	Canada	5	NE	2011
Chippewa Bay	34.30	55.2	New York	United States	5	SE	2011
Iroquois	34.47	55.47	Ontario	Canada	1139	SE	2011
Winchester	35.34	56.87	Ontario	Canada	2482	NE	2011
Hammond	35.40	56.97	New York	United States	1153	SE	2018
Aylmer	35.48	57.1	Quebec	Canada	5	NE	2011
Heuvelton	35.77	57.57	New York	United States	1560	SE	2011

Figure 1: The Distantias user interface and example data output

Unfortunately, this tool charges a small fee to use. Despite a thorough search for equivalent libraries / functions and data sources, the ease of which this tool provides outweighed the negative aspect of a small extra charge to access the data.

I used this tool to obtain data for all Towns and Cities with a population greater than 5 people, within a radius of 100km centred on Smiths Falls, Ontario, Canada.

2.2. Data Cleaning

The reliability of this data seems reasonably good, with transparency about some population data that may be missing from the returned query:

> *'We don't have data for every town and city in Canada and we specify this with NA in our data table. Population data is sourced from a variety of national and international databases some of which are more current than others. The oldest data set is from 2011 but we do make ongoing updates as new census data is released'.*

A provisional review of my queried data did indeed contain hamlets and settlements with no population data; evaluating these manually showed that the settlements are so tiny that no census data has ever been collected from them. Therefore, these were simply dropped from the data as they do not fulfil criterium 4 as detailed in the introduction.

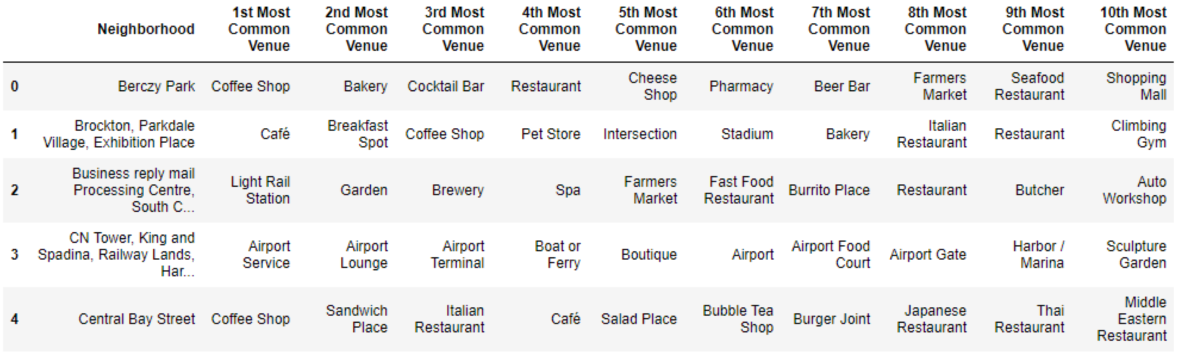
After dropping any NA rows, the data was then processed in turn according to the criteria discussed above: it was first filtered to remove areas too far away, in undesired regions, and with too low or too high a population.

One observation that reduces my confidence in the Distantias data is that almost all of it has a source data of 2011, making it now 10 years out of date. Therefore, I am conscious that this study will not perfectly reflect the current- status of the Towns investigated in the study, especially for areas that have had significant development and gentrification in the last 10 years.

2.3. Feature Extraction

This data will then be leveraged using Foursquare in order to evaluate the areas, towns, cities and neighbourhoods that meet the minimum acceptable criteria. This Foursquare location and venues data was then used to evaluate and rank Towns based on the target attributes.

The Foursquare API can be called to provide lots of meaningful and relevant information - for instance it can be used to examine and cluster the frequency of various amenities in an area, as shown by the following results from a previous and related exercise:



	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Berczy Park	Coffee Shop	Bakery	Cocktail Bar	Restaurant	Cheese Shop	Pharmacy	Beer Bar	Farmers Market	Seafood Restaurant	Shopping Mall
1	Brockton, Parkdale Village, Exhibition Place	Café	Breakfast Spot	Coffee Shop	Pet Store	Intersection	Stadium	Bakery	Italian Restaurant	Restaurant	Climbing Gym
2	Business reply mail Processing Centre, South C...	Light Rail Station	Garden	Brewery	Spa	Farmers Market	Fast Food Restaurant	Burrito Place	Restaurant	Butcher	Auto Workshop
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Airport Terminal	Boat or Ferry	Boutique	Airport	Airport Food Court	Airport Gate	Harbor / Marina	Sculpture Garden
4	Central Bay Street	Coffee Shop	Sandwich Place	Italian Restaurant	Café	Salad Place	Bubble Tea Shop	Burger Joint	Japanese Restaurant	Thai Restaurant	Middle Eastern Restaurant

Figure 2: The Foursquare data can be used to provide a variety of meaningful insights

I used the Foursquare data to extract information pertaining to our key attributes: the number of restaurants, bars, fitness venues, entertainment venues, outdoor spaces, and shopping outlets. This (in addition to the distance data present in the Distantias source data) allowed me to evaluate and rank the Towns based on the key attributes.

3. Methodology

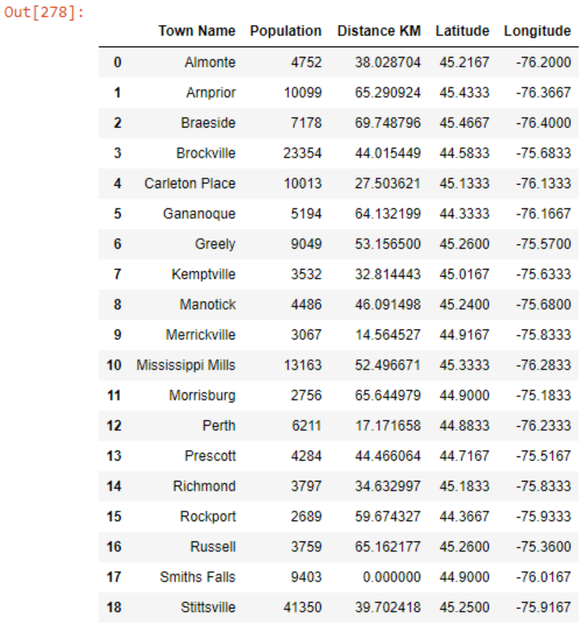
3.1. Exploratory Data Analysis

After the Distantias data was initially imported into the notebook and data frame, I first examined the data using the pandas head, dtypes and shape functions. This initial data frame contained 16 columns (listing data such as the Town Name, distance from Smiths Falls, Latitude & Longitude, Population and other related but very useful data), and 134 rows – each pertaining to a different city within 100km of Smiths Falls.

This data was then processed using conditional data frame slices to remove Towns and Cities that did not meet the study criteria, as follows:

1. Filtered on Towns only within 70km of Smiths Falls, Ontario, Canada.
2. Filtered to remove Towns that are in the Province of Quebec, Canada.
3. Filtered to remove Towns in America.
4. Filtered to remove Towns with less than 2,500 inhabitants.
5. Filtered to remove Cities with more than 50,000 inhabitants.

This resulted in a data frame containing only 19 rows / Towns that meet the above criteria. However, this data frame still contained a large amount of unnecessary data, so it was cleaned further by removing all columns except the 'Town Name', 'Population', 'Distance KM', 'latitude' and 'Longitude' columns.



Out[278]:

	Town Name	Population	Distance KM	Latitude	Longitude
0	Almonte	4752	38.028704	45.2167	-76.2000
1	Arnprior	10099	65.290924	45.4333	-76.3667
2	Braeside	7178	69.748796	45.4667	-76.4000
3	Brockville	23354	44.015449	44.5833	-75.6833
4	Carleton Place	10013	27.503621	45.1333	-76.1333
5	Gananoque	5194	64.132199	44.3333	-76.1667
6	Greely	9049	53.156500	45.2600	-75.5700
7	Kemptville	3532	32.814443	45.0167	-75.6333
8	Manotick	4486	46.091498	45.2400	-75.6800
9	Merrickville	3067	14.564527	44.9167	-75.8333
10	Mississippi Mills	13163	52.496671	45.3333	-76.2833
11	Morrisburg	2756	65.644979	44.9000	-75.1833
12	Perth	6211	17.171658	44.8833	-76.2333
13	Prescott	4284	44.466064	44.7167	-75.5167
14	Richmond	3797	34.632997	45.1833	-75.8333
15	Rockport	2689	59.674327	44.3667	-75.9333
16	Russell	3759	65.162177	45.2600	-75.3600
17	Smiths Falls	9403	0.000000	44.9000	-76.0167
18	Stittsville	41350	39.702418	45.2500	-75.9167

Figure 3: The results of the initial filtering – the 'shortlist' of Towns that meet the minimum criteria.

This data frame represented the final processed, cleaned, and sorted data frame, ready to explore further and evaluate using the Foursquare API. To explore this data further I imported folium and some other required libraries, and generated a map based on the given coordinates of the Towns:

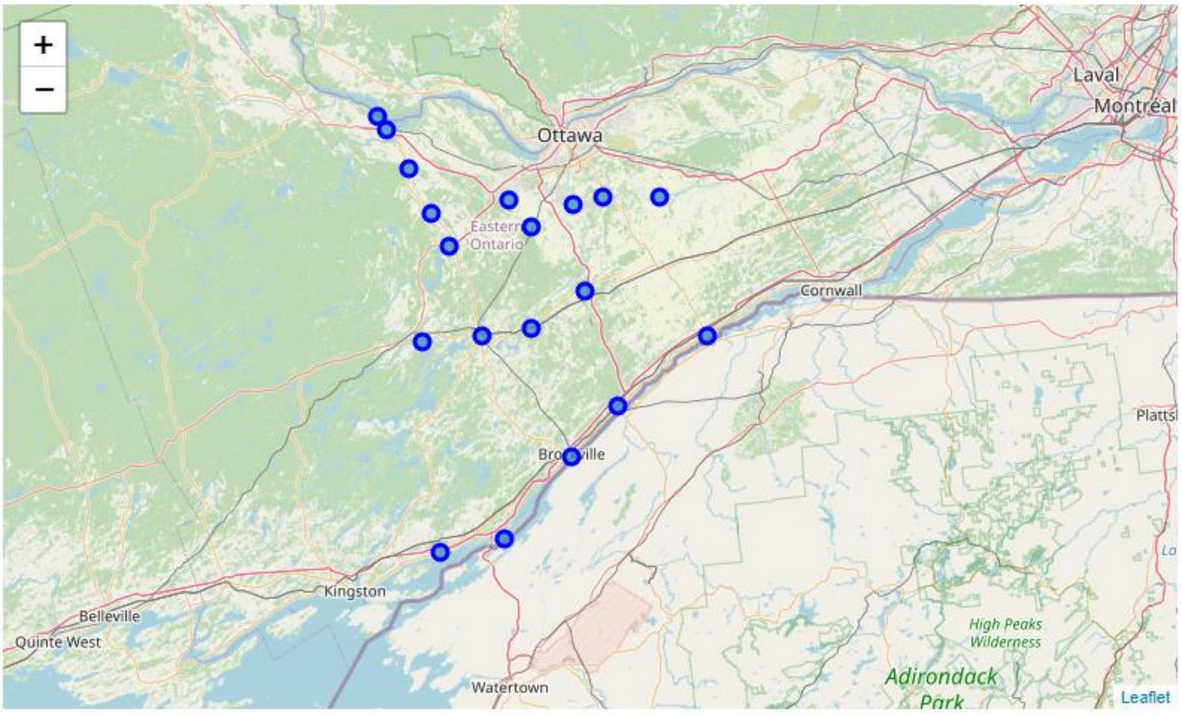
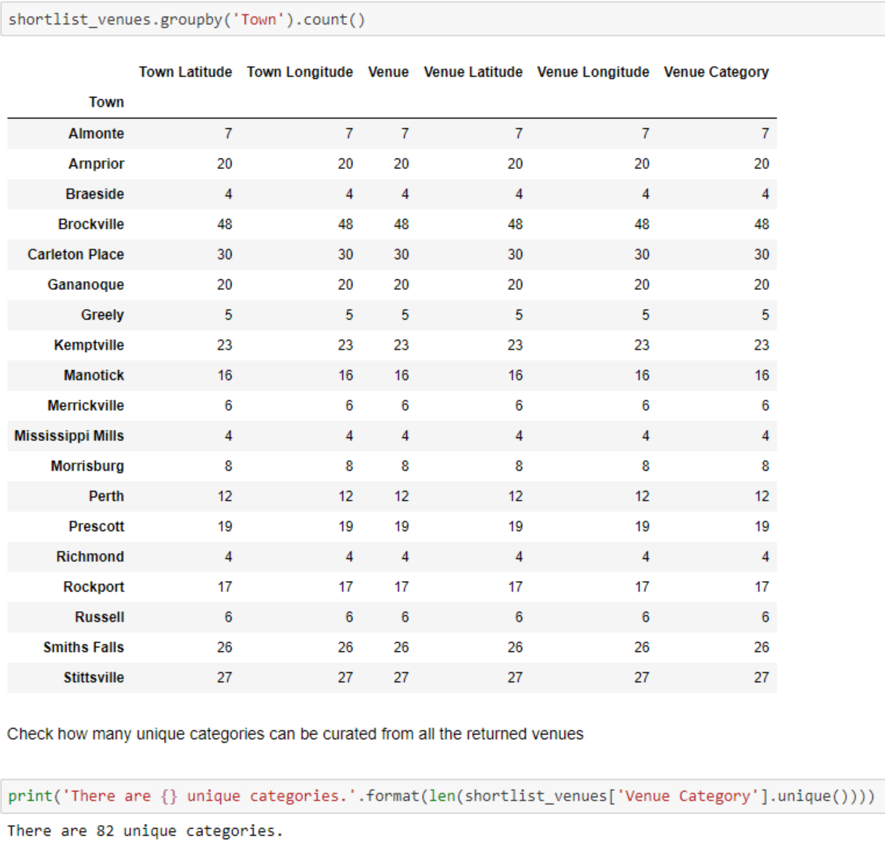


Figure 4: The folium map was useful in visualising the distribution of Towns that meet the minimum criteria for the study. It was reassuring to see that the Towns were fairly evenly distributed rather than all being clustered in one locale.

3.2. Leveraging and exploring the Foursquare data

A function was first defined to iteratively enable me to perform repeated GET requests using the Foursquare API. The shortlist function was then used directly to perform the request for each Town, and place the results in a new data frame called 'shortlist_venues'. The Foursquare request used a radius of 3km from the centre of each Town – this value was settled on after some initial experimentation and was deemed to be a good radial distance (any less and many results were excluded, any further and any unapplicable results from outside the grounds of the city may be erroneously included).

This resulted in 302 venues being returned, representing all the venues present on Foursquare within 3km of all the 19 shortlisted Towns. To explore this data further, the pandas 'groupby' function was used to see how many venues were returned for each Town, and the 'unique' function used to evaluate how many venue categories were listed:



```
shortlist_venues.groupby('Town').count()
```

Town	Town Latitude	Town Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Almonte	7	7	7	7	7	7
Arnprior	20	20	20	20	20	20
Braeside	4	4	4	4	4	4
Brockville	48	48	48	48	48	48
Carleton Place	30	30	30	30	30	30
Gananoque	20	20	20	20	20	20
Greely	5	5	5	5	5	5
Kemptville	23	23	23	23	23	23
Manotick	16	16	16	16	16	16
Merrickville	6	6	6	6	6	6
Mississippi Mills	4	4	4	4	4	4
Morrisburg	8	8	8	8	8	8
Perth	12	12	12	12	12	12
Prescott	19	19	19	19	19	19
Richmond	4	4	4	4	4	4
Rockport	17	17	17	17	17	17
Russell	6	6	6	6	6	6
Smiths Falls	26	26	26	26	26	26
Stittsville	27	27	27	27	27	27

Check how many unique categories can be curated from all the returned venues

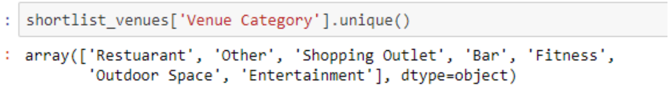
```
print('There are {} unique categories.'.format(len(shortlist_venues['Venue Category'].unique())))
```

```
There are 82 unique categories.
```

Figure 5: Results of some initial exploratory Analysis on the Foursquare results

3.3. Processing and evaluating the Foursquare data

However, the venue categories returned were very specific (for example, 'pizza place' and 'wings joint', rather than being simply labelled as 'restaurant'). To evaluate my six key categories, I re-classified these specific venue types into general ones to be able to evaluate the amount that each Town contains. This was achieved by first defining lists for all of the specific venues, and then using the 'string replace' pandas function to update the categories in the shortlist_venues data frame.



```
: shortlist_venues['Venue Category'].unique()
```

```
: array(['Restuarant', 'Other', 'Shopping Outlet', 'Bar', 'Fitness',
         'Outdoor Space', 'Entertainment'], dtype=object)
```

Figure 6: After updating the Foresquare results, the number of unique categories reduced from 82 specific types down to 7 general categories; ready to be evaluated further.

Once the venue data was generalised, I used the pandas 'groupby' functions to count the total amount of each venues type in each town and sort them into individual data frames. These data frames were then merged to form a final 'scoring' data frame that contained the Town names and the counts for all categories.

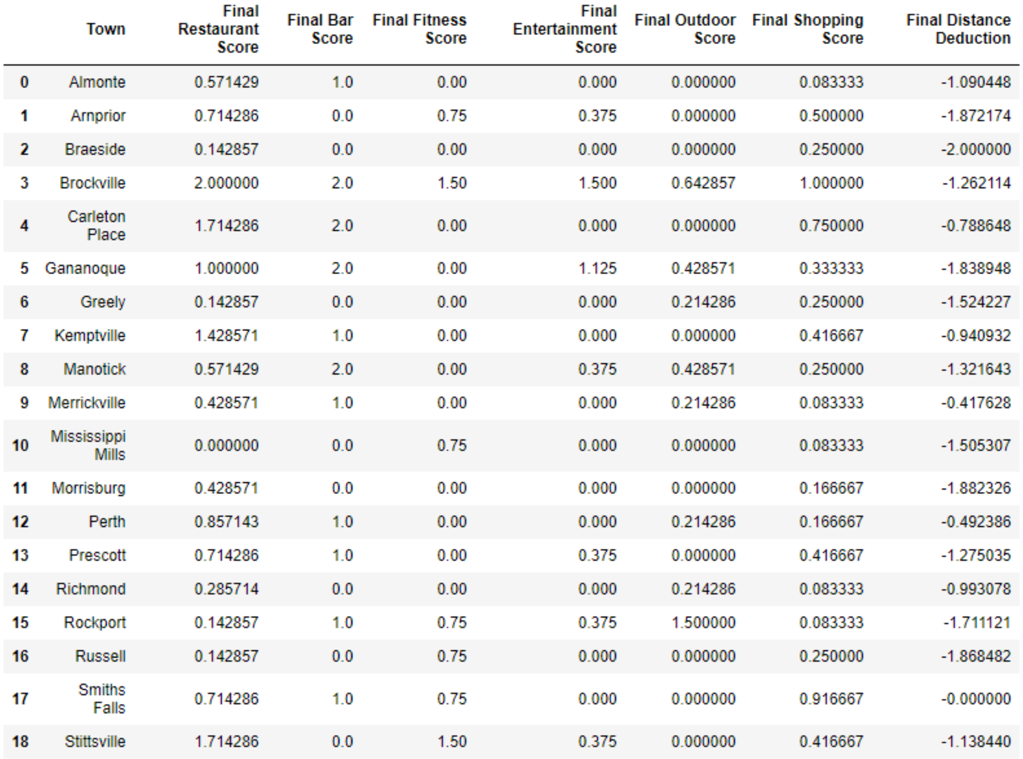
However, the raw counts of each venue type alone are not great values to score each town on, as the amount of venue types in each category can vary wildly. For instance, many of the towns contain over 10 restaurants, but none of them contain more than 2 bars. Therefore, using the raw count would massively bias and weight the results towards towns with high number of restaurants, rather than score the number of bars with a similar weight. To address this and convert them into reliable scores, the counts were normalised by dividing by the maximum value in each column. This converted the counts into scores between 0 and 1 for each of the Towns.

Finally, I multiplied each attribute score by a weighting factor, to represent the priority of importance as outlined initially, as follows:

1. Restaurants - high priority.
2. Bars - high priority.
3. Fitness - medium priority.
4. Entertainment - medium priority.
5. Outdoor - medium priority.
6. Shopping Outlets - low priority.
7. Distance from Smiths Falls- high priority - this is an inverse / negative metric (the further away, the worse)

The weighting factors that were used to multiply the scores by were as follows:

- High priority: Weighting of 2
- Medium priority: Weighting of 1.5
- Low priority: Weighting of 1 (i.e. no change)

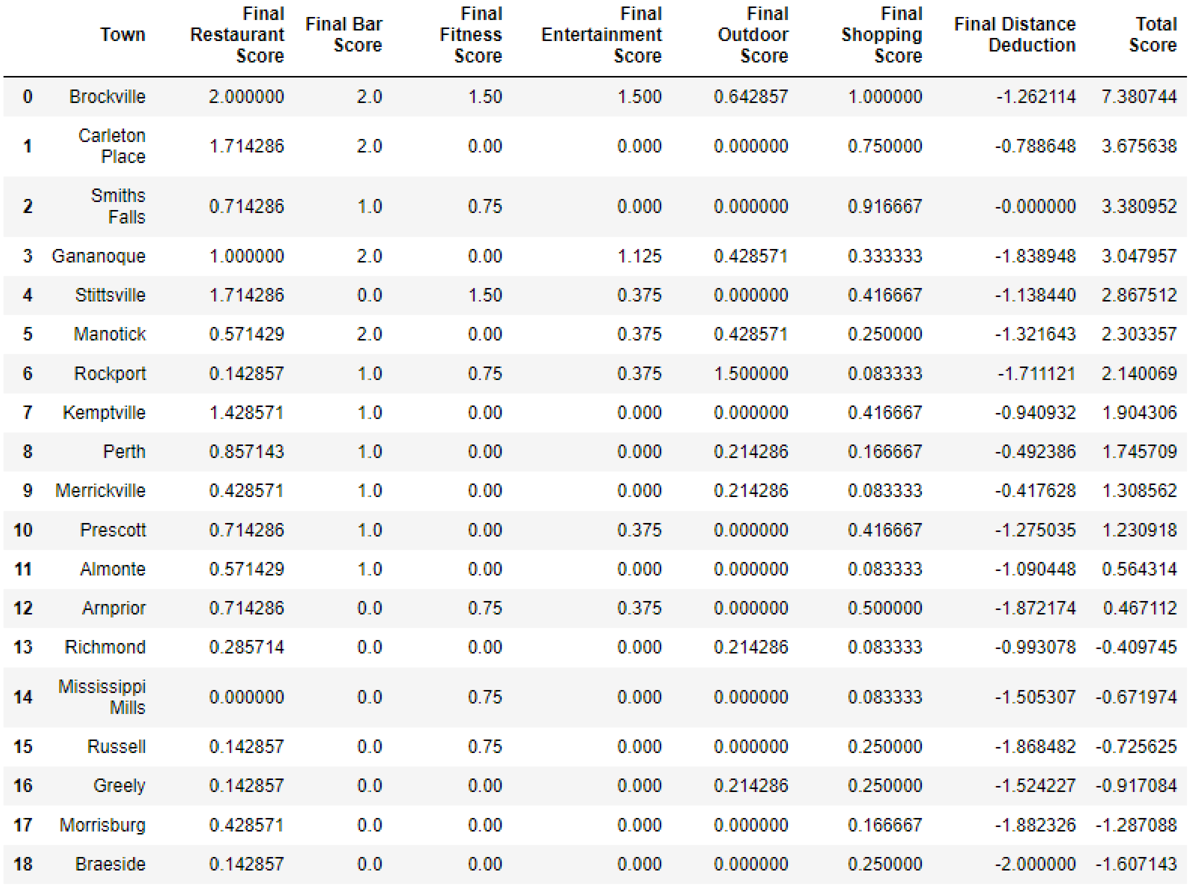


	Town	Final Restaurant Score	Final Bar Score	Final Fitness Score	Final Entertainment Score	Final Outdoor Score	Final Shopping Score	Final Distance Deduction
0	Almonte	0.571429	1.0	0.00	0.000	0.000000	0.083333	-1.090448
1	Arnprior	0.714286	0.0	0.75	0.375	0.000000	0.500000	-1.872174
2	Braeside	0.142857	0.0	0.00	0.000	0.000000	0.250000	-2.000000
3	Brockville	2.000000	2.0	1.50	1.500	0.642857	1.000000	-1.262114
4	Carleton Place	1.714286	2.0	0.00	0.000	0.000000	0.750000	-0.788648
5	Gananoque	1.000000	2.0	0.00	1.125	0.428571	0.333333	-1.838948
6	Greely	0.142857	0.0	0.00	0.000	0.214286	0.250000	-1.524227
7	Kemptville	1.428571	1.0	0.00	0.000	0.000000	0.416667	-0.940932
8	Manotick	0.571429	2.0	0.00	0.375	0.428571	0.250000	-1.321643
9	Merrickville	0.428571	1.0	0.00	0.000	0.214286	0.083333	-0.417628
10	Mississippi Mills	0.000000	0.0	0.75	0.000	0.000000	0.083333	-1.505307
11	Morrisburg	0.428571	0.0	0.00	0.000	0.000000	0.166667	-1.882326
12	Perth	0.857143	1.0	0.00	0.000	0.214286	0.166667	-0.492386
13	Prescott	0.714286	1.0	0.00	0.375	0.000000	0.416667	-1.275035
14	Richmond	0.285714	0.0	0.00	0.000	0.214286	0.083333	-0.993078
15	Rockport	0.142857	1.0	0.75	0.375	1.500000	0.083333	-1.711121
16	Russell	0.142857	0.0	0.75	0.000	0.000000	0.250000	-1.868482
17	Smiths Falls	0.714286	1.0	0.75	0.000	0.000000	0.916667	-0.000000
18	Stittsville	1.714286	0.0	1.50	0.375	0.000000	0.416667	-1.138440

Figure 6: The final weighted scores for each Town across all venue categories

4. Results

To evaluate the final results, I added a new 'Overall Score' column to the data frame and summed each row to give the overall score for all of the 19 Towns. This was then sorted in descending order using pandas to show the final ranking order for each Town:



	Town	Final Restaurant Score	Final Bar Score	Final Fitness Score	Final Entertainment Score	Final Outdoor Score	Final Shopping Score	Final Distance Deduction	Total Score
0	Brockville	2.000000	2.0	1.50	1.500	0.642857	1.000000	-1.262114	7.380744
1	Carleton Place	1.714286	2.0	0.00	0.000	0.000000	0.750000	-0.788648	3.675638
2	Smiths Falls	0.714286	1.0	0.75	0.000	0.000000	0.916667	-0.000000	3.380952
3	Gananoque	1.000000	2.0	0.00	1.125	0.428571	0.333333	-1.838948	3.047957
4	Stittsville	1.714286	0.0	1.50	0.375	0.000000	0.416667	-1.138440	2.867512
5	Manotick	0.571429	2.0	0.00	0.375	0.428571	0.250000	-1.321643	2.303357
6	Rockport	0.142857	1.0	0.75	0.375	1.500000	0.083333	-1.711121	2.140069
7	Kemptville	1.428571	1.0	0.00	0.000	0.000000	0.416667	-0.940932	1.904306
8	Perth	0.857143	1.0	0.00	0.000	0.214286	0.166667	-0.492386	1.745709
9	Merrickville	0.428571	1.0	0.00	0.000	0.214286	0.083333	-0.417628	1.308562
10	Prescott	0.714286	1.0	0.00	0.375	0.000000	0.416667	-1.275035	1.230918
11	Almonte	0.571429	1.0	0.00	0.000	0.000000	0.083333	-1.090448	0.564314
12	Arnprior	0.714286	0.0	0.75	0.375	0.000000	0.500000	-1.872174	0.467112
13	Richmond	0.285714	0.0	0.00	0.000	0.214286	0.083333	-0.993078	-0.409745
14	Mississippi Mills	0.000000	0.0	0.75	0.000	0.000000	0.083333	-1.505307	-0.671974
15	Russell	0.142857	0.0	0.75	0.000	0.000000	0.250000	-1.868482	-0.725625
16	Greely	0.142857	0.0	0.00	0.000	0.214286	0.250000	-1.524227	-0.917084
17	Morrisburg	0.428571	0.0	0.00	0.000	0.000000	0.166667	-1.882326	-1.287088
18	Braeside	0.142857	0.0	0.00	0.000	0.000000	0.250000	-2.000000	-1.607143

Figure 7: The final results of the study

5. Discussion

The results show that the Town of Brockville is the clear favourite based on our ranking criteria, with Carleton Place and Smiths Falls coming in as the second and third highest ranked Towns to consider immigrating to. Brockville scores highly across all categories, despite being approximately 45km away from Smiths Falls. This is very likely due to it being a considerably 'larger' Town than most of the others – with a population of over 23,000 it is over double the size of the second and third-highest ranked Towns (Carleton Place and Smiths Falls, respectively). This makes sense – larger towns generally have more amenities!

Interestingly, at the end of the study when I spoke to my fiancé and explained the details of the course and what I had chosen to investigate, she was able to predict the top three Towns without seeing a line of code or data!

However, the data and consequential results do not perfectly represent the most up-to-date status of the Towns in question. As identified during my exploratory analysis, the data provided by Distantias has a source date of 2011, making the population data now 10 years out-of-date.

Also, I am not entirely confident in the Foursquare API results, as I have visited some of these places several times and can confirm that there are lots of amenities that are missing in the Foursquare data entirely. For instance, in December 2020 I visited Perth with my fiance and we dined at a popular Italian restaurant in the centre of the Town called Bistro 52. This restaurant, and many other bars and shops I have encountered in the area are simply missing from the Foursquare results. If I were to perform the study again I would use a different, more reliable, data source (such as google API).

The study itself is essentially a very limited study. Whilst I am happy with the project and the results, there are many more factors that one should consider when moving to a new area, not just the amount of amneties that are present. For instance, this study does not take into consideration housing prices, ratings of venues or the proximity to other services (hospitals, schools, train stations etc). I have also not considered other important socio-economic factors we may want to look at when conducting a more forensic study, such as literary rate or even the community-aspects of the various Towns and areas.

6. Conclusion

In this study I evaluated some good candidates to consider moving to when I immigrate to Canada. I obtained the source data from the Distantias location proximity search tool and analysed the applicable towns in this dataset using the Foursquare API service. The criteria used for scoring each town was the number of restaurants, bars, fitness venues, outdoor spaces and shopping outlets present in each Town. The distance from Smiths Falls, Ontario, Canada was also used as a way of deducting points for Towns further away (this is important to consider as my fiancé must commute to Smiths Falls for her profession).

The study was performed predominantly using the pandas, numpy and folium libraries, functions and packages. The three top rated Towns for us to consider moving to are Brockville, Carleton Place and Smiths Falls, however more research will be needed on these specific areas to give a better overall 'human' assessment of the suitability of each area.