Table of Content

A. Introduction

A.1. Description

Imagine that you are attempting a new clothing business in Canada. You have a variety of cloth types with very different prices. In addition, you produce for all four seasons. You, may be, have some questions like: Which type of cloths should I sale in each region? Or, perhaps, you are, simply, a new immigrant who is curious about different region's weather to begin a new life. In the case of weather, most common quote is that «Canada is cold», which is not entirely true. First, not all regions are cold. And second, sometimes, it is really hard to bear the hot and humid summer of some cities.

But the question is not just about the weather. To live or to begin a business in a city, it is essential to be well informed about so many subjects. Some of them are: general life style in the city, development likelihood of population, real state rates and forecasts, assurance costs, culture, and accessibility of services.

In this report our concern is about two aspects:

- **Climate:** we are going to understand, if the weather is, somehow, predictable by historical weather information and geographical coordinates.
- **Life style:** we will use statistics of different type of venues in a 1000 meter radius of downtown to determine an estimator index to have an idea about how life is comfortable in each city. Besides, we could have a general idea about cultural similarity of different cities.

A.2. Data description

The weather data, used in this project, comes from four tables of two sites. The tables are easily accessible; however, the registered data in these tables need a considerable manipulation.

- The weather information comes from three tables in «***Wikipedia***» [1].
    - First Table contains the weather information related to «average temperature» for two months of January and July, respectively, as representative of cold and hot months.
    - Second table covers the weather information related to « Heat, cold and frost averages »
    - Finally, the third table shows the weather information related to «extreme temperature» for two months of January and July, respectively, as representative of cold and hot months.
- For the last table, list and the coordinate of Canadian cities are extracted from «***SimpleMaps***» [2].

- Also, we will use «***FourSquare***» as our sources for venues. We will use the coordinate of cities in last table to find our indicator venue occurrence around a fix radius of 1000 m for all cities. [3]
- And finally , the coordinate of downtown are manually extracted from ***google.map*** [4]

B. Methodology

B.1 Data preparation, methods and tools

As the weather information in Wikipedia tables are well organized, gathering data for our analyses was not a complicated process. The manipulation part, on the other hand, was a real challenge. The main problem was mixed information in tables' cells. Brief, after arranging data, they all have sent to ***GitHub*** for further execution of python codes.

Another challenge was the coordinate of downtown. The coordinate which came from ***SimpleMap*** and Wikipedia are for airport stations. This means, there is no venues in a very big radius for these coordinates. As many of Canadian cities have some lakes very closed to their downtown, using the downtown coordinate with a radius of 1000 meters could easily bias information from venues extracted from ***FourSquare***, because it is possible that a big part of radius covers the water, without any business inside, of course. As a consequence, I had to extract the downtown coordinates, manually, by using google map for all ***30*** selected cities, to be able to choose a good point of downtown in order to avoid empty surfaces as lakes, rail roads, airports and parks.

Next problem was the different outcome for different radius of the same coordinate. It forced me to get venues from different radius of 100, 200, 300, 400, 500 and 1000 meters for each altitude, then concatenate the information and finally remove the duplicates to have a more reliable data. Even so, I could not be sure that the results cover all venues in the radius of 1000 meters. However, we could assume that as the problem is for all cities, somehow, the error is proportional. Hopefully, it would not have a considerable effect as a bias on our general analyze.

For visualization, the clustering has done by ***K-Means*** method for ***exact best k*** found by ***elbow*** method. Then the clusters have shown on map by using ***Folium*** library.

To check, if weather is predictable by historical information, I used ***multiple regressions*** and finally for the same weather prediction by altitude, ***simple regression*** was practical.

B.2 In practice: apply methods and tools to analyze data

Step 1: Integrity

After cleaning data as mentioned above, the integrity of data has been checked too. This section contained unifying columns name and type.

Step 2: Descriptive statistics

As our source, already, include the preprocessed data, the information in tables *are* descriptive statistics. Some details have shown in *Figure1.a*, *Figure1.b* and *Figure1.c*:

Figure 1.a Descriptive statistics: *Basic data frame as the source of climate information*

	Location	Region	Weather station	Latitude (N)	Longitude (W)	Elevation (m)	January DegreeC - LowRecord	January DegreeC - LowAverage	January DegreeC - HighAverage	January DegreeC - HighRecord	...
0	Baker Lake	NU	YBK	64.29889	-96.07778	19	-50.6	-34.8	-27.7	-27.7	...
1	Calgary	AB	YYC	51.11389	-114.02028	1084	-44.4	-13.2	-0.9	17.6	...
2	Charlottetown	PE	YYG	46.28861	-63.12861	49	-30.5	-12.1	-3.4	15.1	...
3	Churchill	MB	YYQ	58.73917	-94.06639	29	-45.0	-30.1	-21.9	1.7	...
4	Dawson Creek	BC	YDA	64.04306	-139.12778	370	-53.8	-30.1	-21.8	9.7	...

Figure 1.b Descriptive statistics: *Cities' boxplot of climate information*

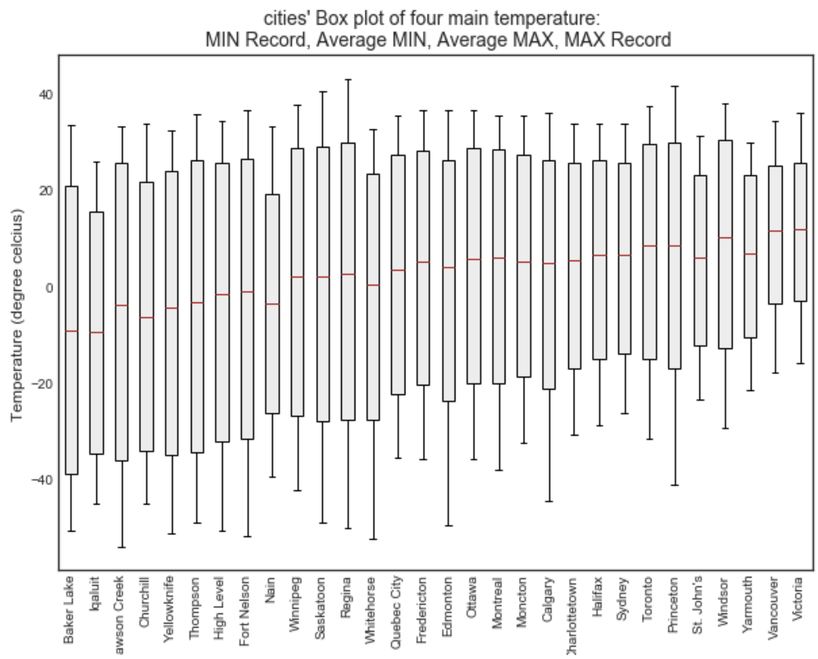
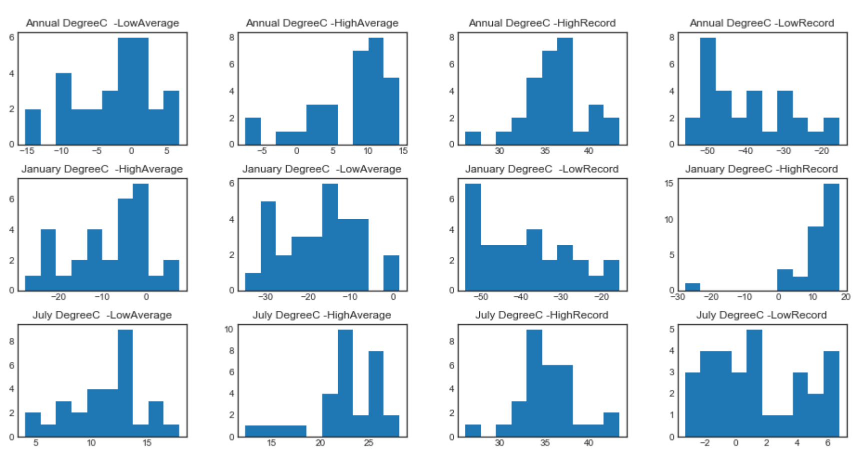


Figure 1.c Descriptive statistics: *Histogram of main climate information*



Step 3: Explore the problem by raw data visualization

Considering our data, talking about temperature in Canada brings us to three questions:

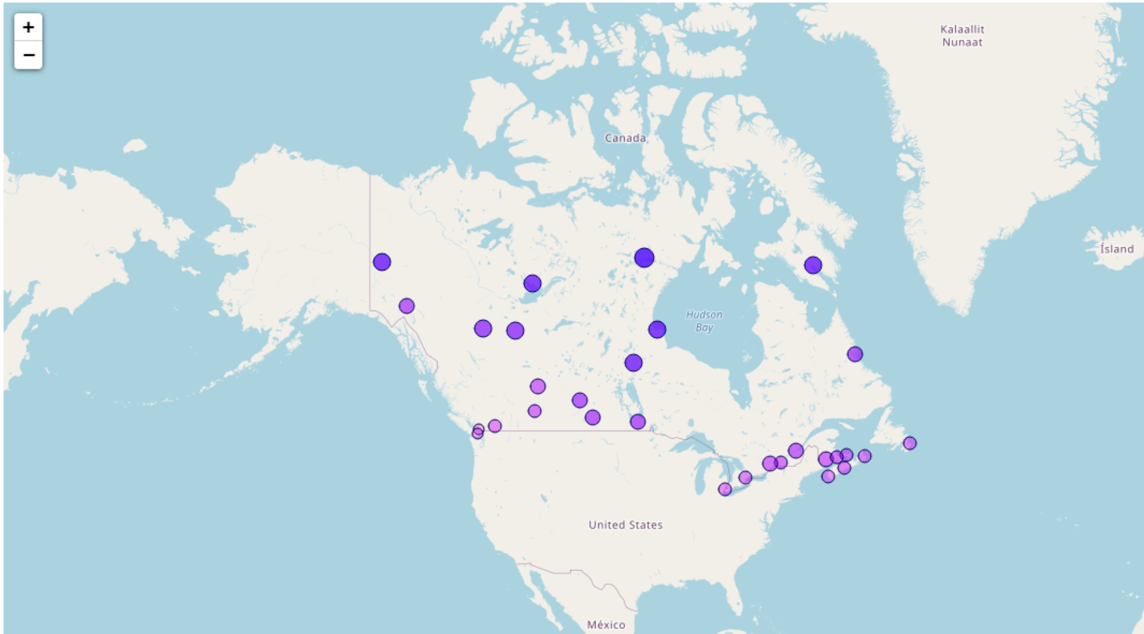
1- What is the average low temperature of each city?
2- What is the average high temperature of each city?
3- How many days per year I could go out without being worried about frosting?

For the first and the third question, *Figure.2* could help us. In this map, the diameters of markers are proportional to number of frost-free days of each city. It means bigger circle shows less frosty days. On the other hand, the density of red color shows that how much the city is warm at July.

Figure.2 Visualizing map of 30 Canadian cities based on summer temperature and number of non-frosty days

For second question, in *Figure.3,* we traced the January Low temperature as an index of intensity of cold weather. The diameter and the color, both, are proportional measure of low temperature. Bigger the circle is, colder the weather would be. The density of blue color shows the same. (More intense = colder)

Figure.3 Visualizing map of 30 Canadian cities based on winter temperature



Know we could have our primary insights:

1- Center and north of Canada are colder than east and west, at winter and at summer
2- East of Canada has about the same high average as west , in July
3- West of Canada is warmer than east, in January

Step 4: Correlation matrix

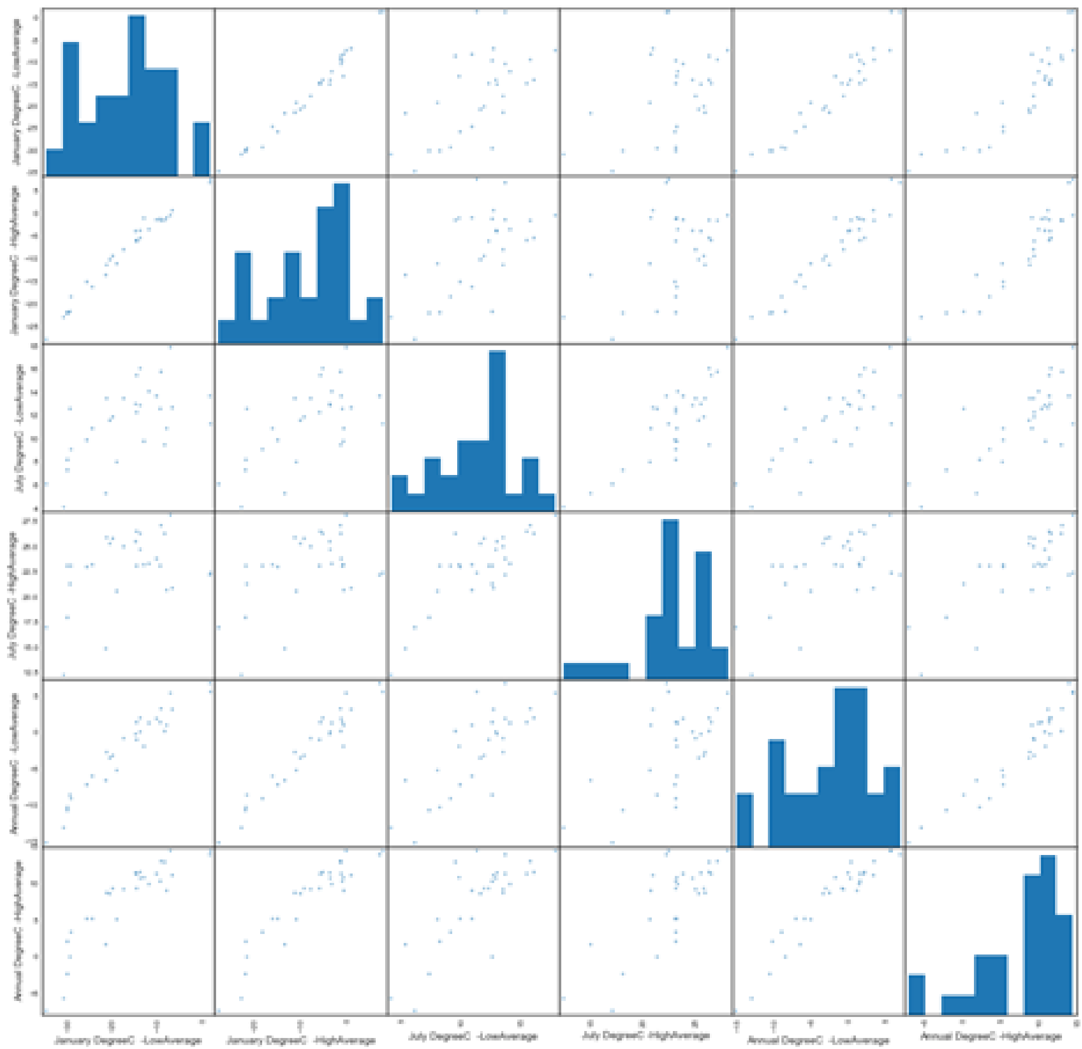
In fact, the objective is to find out if, somehow, the weather is predictable by using our historic data. The question is what are our dependent and our independent variables? Or simply, when we want predict the weather which features, exactly, need to be predicted?

First of all, we are looking in numeric data, so some feature as cities; weather station etc. will be eliminated automatically from equation. Secondly, looking at table, we have three independent variables: latitude, longitude and elevation. All the other columns, which are included weather information, are dependent variables. But do we need to execute a regression model holding all these variables? Let's check the correlation between dependent variables in *Figure.4 and Figure.5*

Figure.4 Correlation matrix for climate dataset

	January DegreeC - LowAverage	January DegreeC - HighAverage	July DegreeC - LowAverage	July DegreeC - HighAverage	Annual DegreeC - LowAverage	Annual DegreeC - HighAverage
January DegreeC - LowAverage	1.000000	0.986341	0.598462	0.432188	0.959537	0.887340
January DegreeC - HighAverage	0.986341	1.000000	0.626303	0.499479	0.965610	0.928835
July DegreeC - LowAverage	0.598462	0.626303	1.000000	0.820351	0.786183	0.780423
July DegreeC - HighAverage	0.432188	0.499479	0.820351	1.000000	0.610975	0.769182
Annual DegreeC - LowAverage	0.959537	0.965610	0.786183	0.610975	1.000000	0.952283
Annual DegreeC - HighAverage	0.887340	0.928835	0.780423	0.769182	0.952283	1.000000

Figure.5 Scatter plots matrix for climate dataset



As we see in correlation matrix all the temperature indexes (beside of frost free days which is not included in matrix) are well correlated. In fact, all of them have a very strong correlation with «Annual DegreeC -LowAverage» and «Annual DegreeC -HighAverage». So, considering the correlation indexes, we continue with «Annual DegreeC -LowAverage» as our first dependent variable. Also we add the «Frost-free days» to our analysis. with the help of ***sklearn*** library,

- We split data to train and test subsets.
- Then, we normalized data.
- And finally, we fit our normalized data to a regression model.

Here are the results:

```
Coefficients:  [[-0.97244301          -0.47387296              -0.46361122]]
Residual sum of squares: 0.24
Variance score: 0.76
r2_score:  0.7451418653829405
```

As the scores are not good enough, we could not conclude that «Annual DegreeC -LowAverage» and «Frost-free days» are predictable by geographical coordinates (latitude, longitude, elevation)

We repeat the regression, this time a simple linear regression instead of multiple one for all possible pairs of dependent-independent variables. Here is the final result for (latitude vs low average):

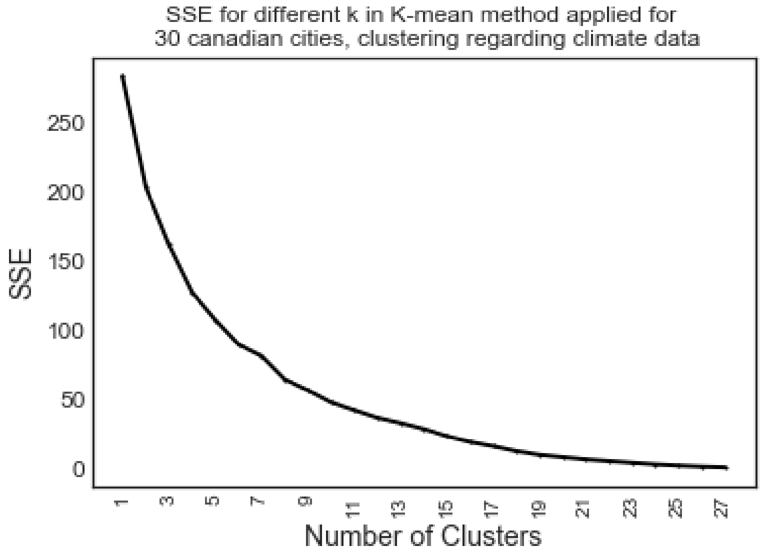
```
Coefficients:  [[-0.89941192]] with intercept of [0.01760437]
Residual sum of squares: 0.09
Variance score: 0.89
r2_score: 0.8735988297984665
```

This time the scores are good and we could suggest predicting the low temperature by latitude, but the problem is that is not a really useful information, because, most of major cities of Canada, are situated in a thin band of south of country.

Step 5: Clustering cities, based on weather information

The purpose of this section is to cluster cities regarding weather information. The result, which we expect, will be a sort of gathered cities with the same climatic characteristics. The applied method is K-Mean. We find the best k for this method by executing the model for different k and calculation of SST for each k. By tracing the plot of K vs SST, estimation of the best k could be realized, visually, by elbow method on graph. *Figure.6* shows the mentioned graph for our dataset. Also, the Kneelocator of ***kneed*** library could be referred, as a more reliable method, to find an exact best k. Here, we used Kneelocator to find exact best k which is 8.

Figure.6 Elbow plot to find best number of cluster based on climate dataset



We assign the cluster calculated for k=8, then, we visualize the result as in *Figure.7*

Figure.7 Clustering cities, based on weather information

As you could see on the map, there is a clear pattern in cities' behavior with two interesting exceptions. First, it is pair of Calgary and Edmonton, which is in center of Canada, in the same cluster as Montreal, Quebec city and Ottawa. And second, it is Princeton in BC which has a cold weather as Winnipeg and Regina. This one, probably, is a consequence of Princeton's high elevation.

Step 6: Define a life style (or service) indicator

The general idea is that the number and variety of venues in a city could reflect the style, wealth and culture of society. Unfortunately, because of cost and time limit, we could not cover all the venue of these 30 cities. So we try to get venues information as much as possible for a 1000 meter radius of each downtown. Foursquare gave us the venues in **246 unique categories**. The ten first lines of summarized result are as following in *figure.8*. We will use a normalized version of this Venue Score as our life style index.

Figure.8 Summarize table of the number of venues inside different radius

Radius Category	100	200	300	400	500	1000	VenueScore
Location							
Montreal	7.0	21.0	20.0	23.0	35.0	68.0	174.0
Toronto	21.0	23.0	22.0	33.0	17.0	56.0	172.0
Vancouver	12.0	17.0	22.0	15.0	30.0	55.0	151.0
Edmonton	NaN	34.0	27.0	39.0	12.0	35.0	147.0
Ottawa	8.0	10.0	29.0	17.0	23.0	59.0	146.0
Quebec City	10.0	12.0	31.0	7.0	20.0	64.0	144.0
Halifax	10.0	24.0	23.0	38.0	19.0	26.0	140.0
Victoria	6.0	25.0	50.0	26.0	15.0	6.0	128.0
Winnipeg	2.0	9.0	24.0	17.0	18.0	55.0	125.0
Calgary	6.0	2.0	4.0	10.0	19.0	67.0	108.0

Step 7: Clustering cities, based on top venues for each city

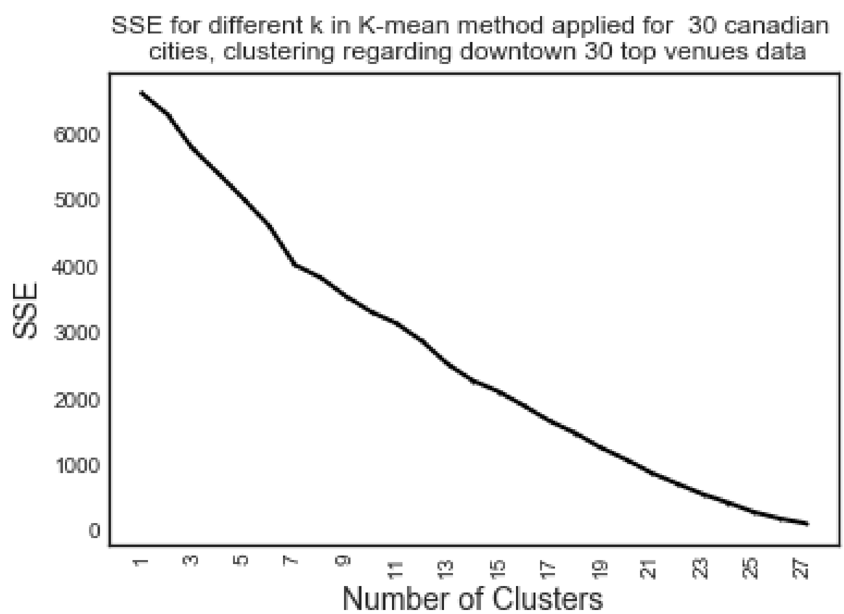
After arranging data, we choose 30 first important venues of each city. The final table would be like ***Figure.9***

Figure.9 top 30 venues in downtown for each cities' downtown

	Location	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	...
0	Baker Lake	Hotel	Fast Food Restaurant	Yoga Studio	Event Space	Gaming Cafe	Furniture / Home Store	Frozen Yogurt Shop	Fried Chicken Joint	French Restaurant	...
1	Calgary	Hotel	Coffee Shop	Restaurant	Pub	Steakhouse	Bar	Cocktail Bar	Café	Brazilian Restaurant	...
2	Charlottetown	Coffee Shop	Hotel	Seafood Restaurant	Pub	Restaurant	History Museum	Sushi Restaurant	Ice Cream Shop	Gastropub	...
3	Churchill	Inn	Hotel	Gastropub	Harbor / Marina	Bed & Breakfast	Coffee Shop	Train Station	Restaurant	Beach	...
4	Dawson Creek	Coffee Shop	Café	Historic Site	Burger Joint	Grocery Store	Convenience Store	Restaurant	Sandwich Place	Fast Food Restaurant	...

We apply K-Means method, the real best k return by calculation give us 14 cluster, which is difficult to visualize and make decision in real word (14 cluster on 30 cities is too much). So we try to estimate a more realistic k by looking for another good k on k-SSE graph as bellow, ***Figure.10.***

Figure.10 Elbow plot to find best number of cluster based on downtown venues

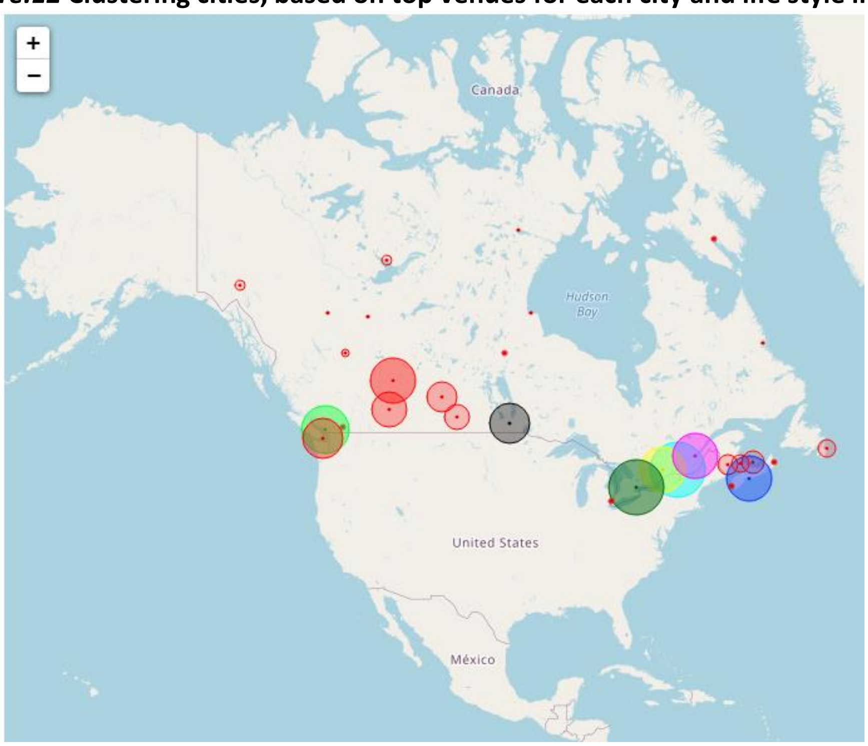


It seems k=8 too is another break point similar to elbow. So we choose k=8. The result of clustering will be as illustrated in ***Figure 11***

Figure 11. **Clustering cities, based on top venues for each city**

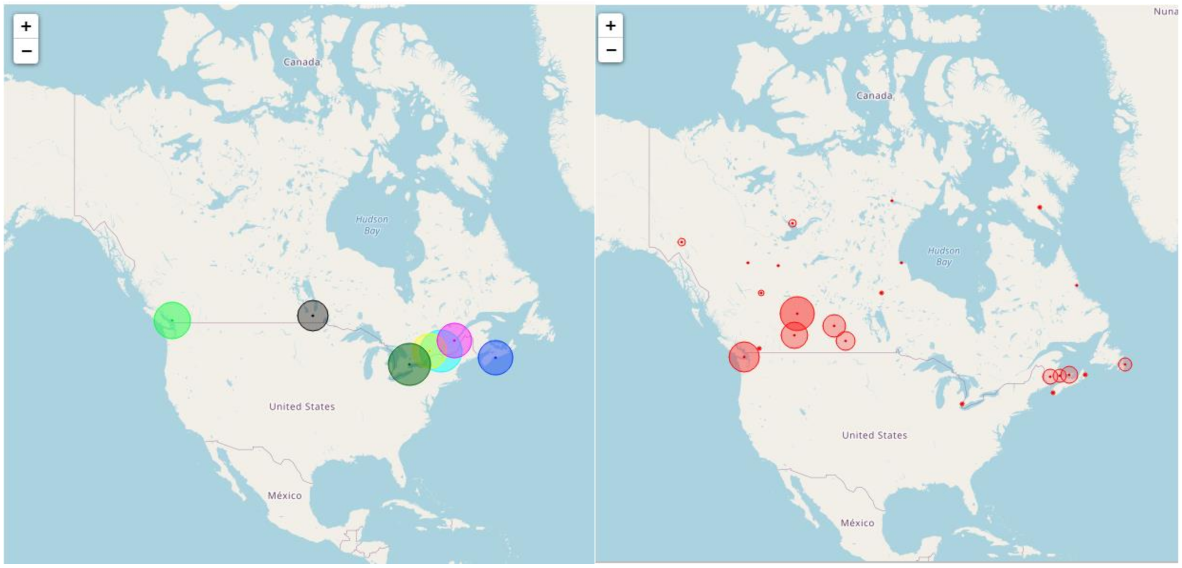
We have 23 cities in the same cluster and 7 another cluster which each one contains just one city. To have a better visualization of situation, we are going to retrace the same map, but this time, the radius of marker will be proportional to the normalized life style index. (See ***Figure 12***)

Figure.12 **Clustering cities, based on top venues for each city and life style index**



Again, to have a better visibility we separate cluster 1 from clusters 0, 2, 3, 4, 5, 6 and 7.

Figure.13 **Clustering cities, based on top venues for each city and life style index**
Separate cluster 1 (right map) from clusters 0, 2, 3, 4, 5, 6 and 7 (left map)



Now we could observe that 7 out of 12 main cities of Canada are clustered alone. For next 5 most comfort cities of Calgary, Edmonton, Regina, Saskatoon and Victoria, Perhaps we could consider the possibility that these cities are more similar to other cities of Canada regarding Canadian culture it means, there is less number of immigrants. Perhaps the multicultural nature of seven first cities makes them uni-cluster. But what is the difference between them? That is a question for another project.

C. Results

At this point that we finished our analyses, we could summarize some helpful results in a final table. So we put all following together, in a unique table, as the outcome of this project:

- The basic information
- Cluster codes: for Clustering cities, based on top venues for each city
- Cluster codes: for Clustering cities, based on weather information
- Life style (or service) indicator

Table.1 **Summarized result table: basic information, Cluster codes and life style indicator**

Location	Region	Downtown Latitude (N)	Downtown Longitude (W)	Climate Cluster	Cultural similarity Cluster	Number of Venues in downtown (r=1000m)	Service accessibility index
Montreal	QC	45.506276	-73.565902	2	2	174	100%
Toronto	ON	43.654097	-79.379946	5	3	172	99%
Vancouver	BC	49.278877	-123.115975	4	6	151	87%
Edmonton	AB	53.541762	-113.496502	2	1	147	84%
Ottawa	ON	45.421401	-75.699699	2	5	146	84%
Quebec City	QC	46.812671	-71.213642	2	4	144	83%
Halifax	NS	44.646946	-63.575615	7	0	140	80%
Victoria	BC	48.427104	-123.366767	4	1	128	74%
Winnipeg	MB	49.892412	-97.140338	0	7	125	72%
Calgary	AB	51.048967	-114.067123	2	1	108	62%
Saskatoon	SK	52.127357	-106.664435	0	1	93	53%
Regina	SK	50.450650	-104.612339	0	1	78	45%
Charlottetown	PE	46.234212	-63.127499	7	1	68	39%
Fredericton	NB	45.962307	-66.642358	2	1	63	36%
St. John's	NL	47.559473	-52.710136	7	1	57	33%
Moncton	NB	46.089910	-64.780679	2	1	57	33%
Whitehorse	YT	60.720308	-135.055446	1	1	33	19%
Yellowknife	NT	62.452935	-114.386882	1	1	29	17%
Dawson Creek	BC	55.757065	-120.234002	1	1	23	13%
Iqaluit	NU	63.749418	-68.521182	6	1	19	11%
Sydney	NS	46.139310	-60.172719	7	1	17	10%
Windsor	ON	42.301646	-82.997735	5	1	17	10%
Yarmouth	NS	43.836799	-66.118090	7	1	17	10%
Thompson	MB	55.743680	-97.855411	1	1	14	8%
Princeton	BC	49.457305	-120.511413	0	1	13	7%
Churchill	MB	58.770010	-94.165489	3	1	10	6%
High Level	AB	58.514318	-117.136904	1	1	8	5%
Fort Nelson	BC	58.803779	-122.696529	1	1	7	4%
Baker Lake	NU	64.319224	-96.029856	3	1	4	2%
Nain	NL	56.542536	-61.694467	6	1	3	2%

D. Conclusion

Objective of this project was to give a big picture of reginal climate and culture in some important Canadian cities. It would be helpful for someone who has plans to move to a Canadian city or begin a business.

We checked data from different aspects to find out how we could achieve our objective:

- We tried to understand if the climate is predictable by geographical altitudes. The results on regression showed us that is not a very reliable strategy. So we don't use anymore a regression model.
- We did cities clustering by climate data. It worked well and we found that Canadian major cities could be clustered in 8 groups. So, one of our suggested methods is to refer to provided cluster map.
- Then, as the climate is not the only important parameter, getting help of foursquare, we added a life comfort indicator. This indicator reflects the accessibility of services in each city. However, it is not suggested to combine it with our climate clustering because of special importance of this index. If we want to integrate them it would better to prepare a decision matrix which is out of subject in this project.
- And finally, we clustered cities by considering 30 most important venues of each city. This clustering, in some ways, could lead us to a general picture of social, cultural and even financial resemblance of cities.

So we showed three principal aspects of this subject:

- Similar cities by climate situation
- Service accessibility index for each city
- Cultural, social and financial resemblance of cities

For a future development, this subject could be expand by adding analysis on real state subject and assurances (automobile, house, and health) etc. Also, it would be a good idea if we have some quantified information about the quality of services that are mentioned in the current project. Another missing part of this project, as mentioned in 2nd previous section is comparing the major cities to understand what makes them different from each other (why they are not in same cluster).

E. References

- [1] ***https://en.wikipedia.org/wiki/Temperature_in_Canada***
- [2] ***https://simplemaps.com/data/canada-cities***
- [3] ***https://foursquare.com/***
- [4] ***https://www.google.ca/maps/place/Canada***