Estimating clusters of French people in Montreal based on borough top venues

Michael Erblang

September 1st, 2020

I. Introduction

1. Overview

I am a French man currently living in Paris. I have lived in Los Angeles for two years and I have loved discovering new culture, new country and my experience there was amazing. But pretty quickly I met French people living around. We started hanging out and shared our French lifestyle with our american friends.

From what I heard and what I experienced, people working abroad tend to connect to their own culture as it may be a way to feel more confortable.

What if immigrants all over the world tend to cluster and live surrounded by their peers when they live in a foreign country ? If so would not be easier to target those people who will be delighted to find some products from home (cheese!) or educational institution, place of worship, etc.?

We can wonder where immigrants people tend to live and why in this place. It may be because there is a company, a school, embassy, etc. And maybe these clusters already brought some businesses such as restaurant, shop, etc. The best example of businesses development based on immigration cluster is Chinatown in New York. But some clusters abroad are less obvious and can be estimated analyzing the neighborhood landscape.
And so we can define a **borough identity based on the popular venues** of this borough.

2. Approach

I will estimate those clusters in the city of Montreal, Quebec in eastern Canada.
Montreal is known to be a cosmopolitan city influenced by both the French and the North American cultures. It makes it the perfect target to estimate French population clusters.

- First I need the **top venues from Montreal boroughs.**
- Then I need to define what is "French culture" in top venues. For that, I will create my **French database based on the top venues of Paris**, the capital of France and the biggest representation of French culture abroad.
- Finally I will build a model to estimate those clusters of French people.

II. Data acquisition and cleaning

1. Data sources

I need to explore the popular venues of Paris boroughs and Montreal boroughs on Foursquare, plus the list of Paris boroughs and the list of Montreal boroughs on Wikipedia.

Geocoder will give us the GPS coordinates of each borough and a Geojson file of Montreal will allows us to visualize the clusters.

Finally an official census report of Montreal would allow us to validate our model.

2. Data acquisition

	Borough	Name	Latitude	Longitude	Area(km²)
0	1er	Louvre	48.862509	2.338211	1.83
1	2e	Bourse	48.866535	2.340165	0.99
2	3e	Temple	48.863706	2.361094	1.17
3	4e	Hôtel-de-Ville	48.854596	2.362937	1.60
4	5e	Panthéon	48.845461	2.344518	2.54
5	6e	Luxembourg	48.850537	2.332774	2.15
6	7e	Palais-Bourbon	48.857980	2.315100	4.09
7	8e	Élysée	48.877786	2.316351	3.88
8	9e	Opéra[note 1]	48.872831	2.340419	2.18
9	10e	Entrepôt, anciennement Enclos Saint-Laurent	48.871737	2.357201	2.89
10	11e	Popincourt	48.858736	2.378556	3.67
11	12e	Reuilly (hors bois de Vincennes)	48.840792	2.388745	6.37
12	13e	Gobelins	48.831916	2.355855	7.15
13	14e	Observatoire	48.832764	2.324742	5.64
14	15e	Vaugirard	48.842120	2.299240	8.48
15	16e	Passy (hors bois de Boulogne)	48.863950	2.277365	7.91
16	17e	Batignolles-Monceau	48.883283	2.319248	5.67
17	18e	Buttes-Montmartre	48.892381	2.344977	6.01
18	19e	Buttes-Chaumont	48.887130	2.382775	6.79
19	20e	Ménilmontant	48.864767	2.398592	5.98

Table 1: Paris boroughs

	Borough	Latitude	Longitude	Area(km²)
0	Ahuntsic-Cartierville	45.541892	-73.680319	24.2
1	Anjou	45.604898	-73.546672	13.7
2	Cote-des-Neiges-Notre-Dame-de-Grace	45.483609	-73.626970	21.4
3	Lachine	45.448676	-73.711204	17.7
4	LaSalle	45.432514	-73.629267	16.3
5	Le Plateau-Mont-Royal	45.521836	-73.582173	8.1
6	Le Sud-Ouest	45.467991	-73.588561	15.7
7	L'Ile-Bizard-Sainte-Genevieve	45.495042	-73.903445	23.6
8	Mercier-Hochelaga-Maisonneuve	45.574099	-73.525838	25.4
9	Montreal-Nord	45.593899	-73.637606	11.1
10	Outremont	45.518617	-73.606886	3.9
11	Pierrefonds-Roxboro	45.495508	-73.847175	27.1
12	Riviere-des-Prairies-Pointe-aux-Trembles	45.659235	-73.531247	42.3
13	Rosemont-La Petite-Patrie	45.550747	-73.582290	15.9
14	Saint-Laurent	45.508877	-73.687519	42.8
15	Saint-Leonard	45.586710	-73.596949	13.5
16	Verdun	45.459078	-73.573177	9.7
17	Ville-Marie	45.499805	-73.571169	16.5
18	Villeray-Saint-Michel-Parc-Extension	45.537006	-73.625796	16.5

Table 2: Montreal boroughs

First let's introduce Paris districts and Montreal boroughs.

Above there are the name of the boroughs, their latitude and longitude plus the area of each. I convert the areas in km² to enhance the way to get the top venues in every boroughs considering borough as (√Area)x(√Area) square so I can locally search for top venues.

Also Foursquare allows us to get only 100 of the top venues from a defined region. So I decided searching for top 100 venues in every 20 Paris boroughs. It allows me mining 100x20 = 2000 top venues in Paris.

3. Data cleaning

I am interested by the categories name of the top venues as it is a classifying process already made by Foursquare. Counting the categories occurences will help me assign a weight to categories and so I can obtain my Paris identity weighted-matrix which is also my French culture weighted-matrix. Then I normalized the values.

	Paris
French Restaurant	286
Hotel	112
Italian Restaurant	75
Bar	68
Bistro	53
Bakery	53
Japanese Restaurant	46
Wine Bar	45
Plaza	44
Coffee Shop	42
Café	41
Restaurant	37
Pizza Place	32
Cocktail Bar	28
Vietnamese Restaurant	26
Bookstore	25
Thai Restaurant	24
Garden	22
Korean Restaurant	21
Park	21

	Paris
French Restaurant	1.000000
Hotel	0.385965
Italian Restaurant	0.263158
Bar	0.252632
Bistro	0.192982
Bakery	0.175439
Japanese Restaurant	0.161404
Wine Bar	0.154386
Plaza	0.150877
Café	0.143860
Coffee Shop	0.143860
Restaurant	0.129825
Pizza Place	0.112281
Cocktail Bar	0.094737
Thai Restaurant	0.087719
Vietnamese Restaurant	0.087719
Garden	0.084211
Bookstore	0.084211
Park	0.080702
Seafood Restaurant	0.070175

Table 3: top venues occurence and weighting process – French matrix

We are processing the same for Montreal borough top venues and so we obtained a Montreal borough weighted-series for every borough, resulting in the Montreal identity matrix.

At this part we observe the French matrix contains more categories than the Montreal one as both cities have differents top venues. We reassigned the non-assigned categories to existing one when it is possible. Otherwise I dropped the categories.

For example 'Brasserie' is only present in the French dataframe. A Brasserie is a kind of bar/ restaurant where you can find mostly typical French food and beverage. So I can reassign it to the 'French Restaurant' category renaming it and suming it to the existing 'French Restaurant' value. We reassign some others categories following the same pipeline.

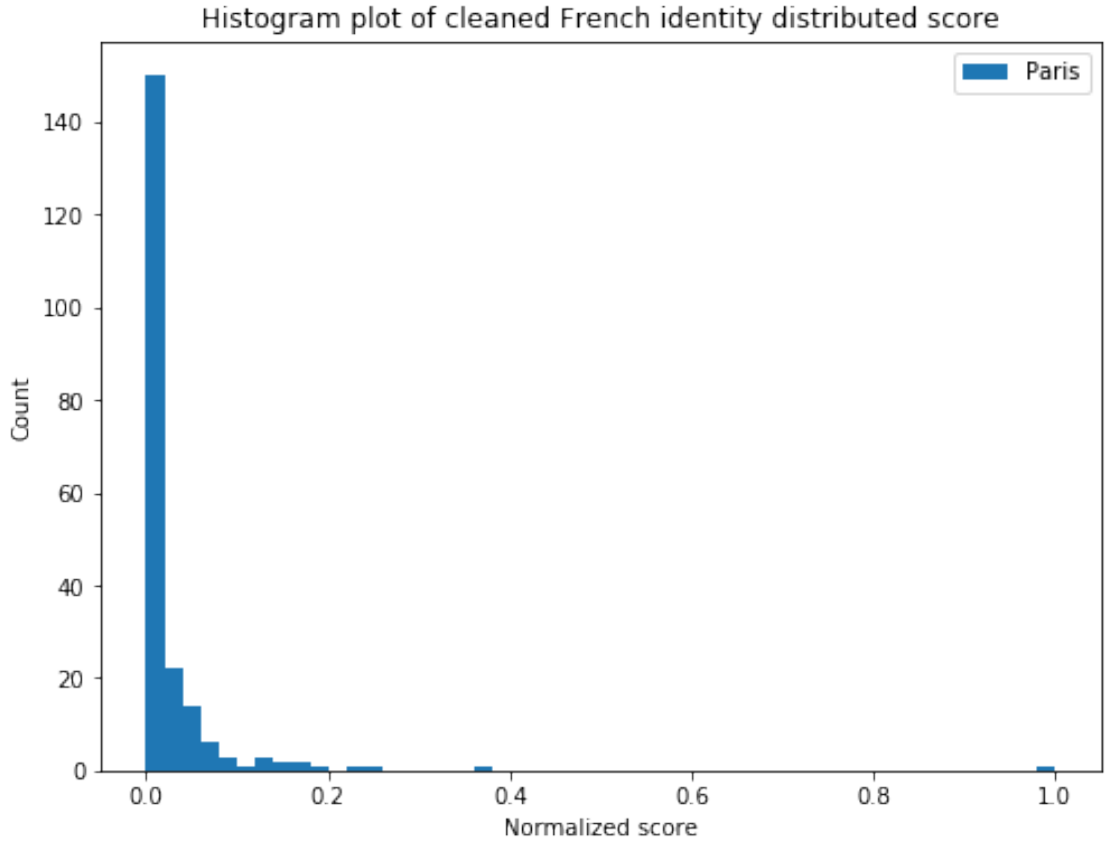
Reassigned categories	New category assignment	Reason of reassignment	Dropped categories
Brasserie	*French Restaurant*	Type of French restaurant	*Lebanese Restaurant*
Wine Shop	*Wine bar*	Closest match to wine	*Moroccan Restaurant*
History Museum	*Museum*	Mother category	*African Restaurant*
Pedestrian Plaza	*Plaza*	Mother category	*Noodle House*
Bed & Breakfast	*Hostel*	Mother category	*Argentinian Restaurant*
Fountain	*Plaza*	Mother category	*Ramen Restaurant*
Science Museum	*Museum*	Mother category	*Udon Restaurant*
Basque Restaurant	*French Restaurant*	Southern France food	*Turkish Restaurant*
Corsican Restaurant	*French Restaurant*	Southern Frence food	*Bubble Tea Shop*
Multiplex	*Movie Theater*	Mother category	*Candy Store*

Table 4: Simple feature selection during data cleaning.

III. Exploratory Data Analysis

1. French identity weighted matrix – Data Analysis

First let's see how the weight is distributed among the categories.

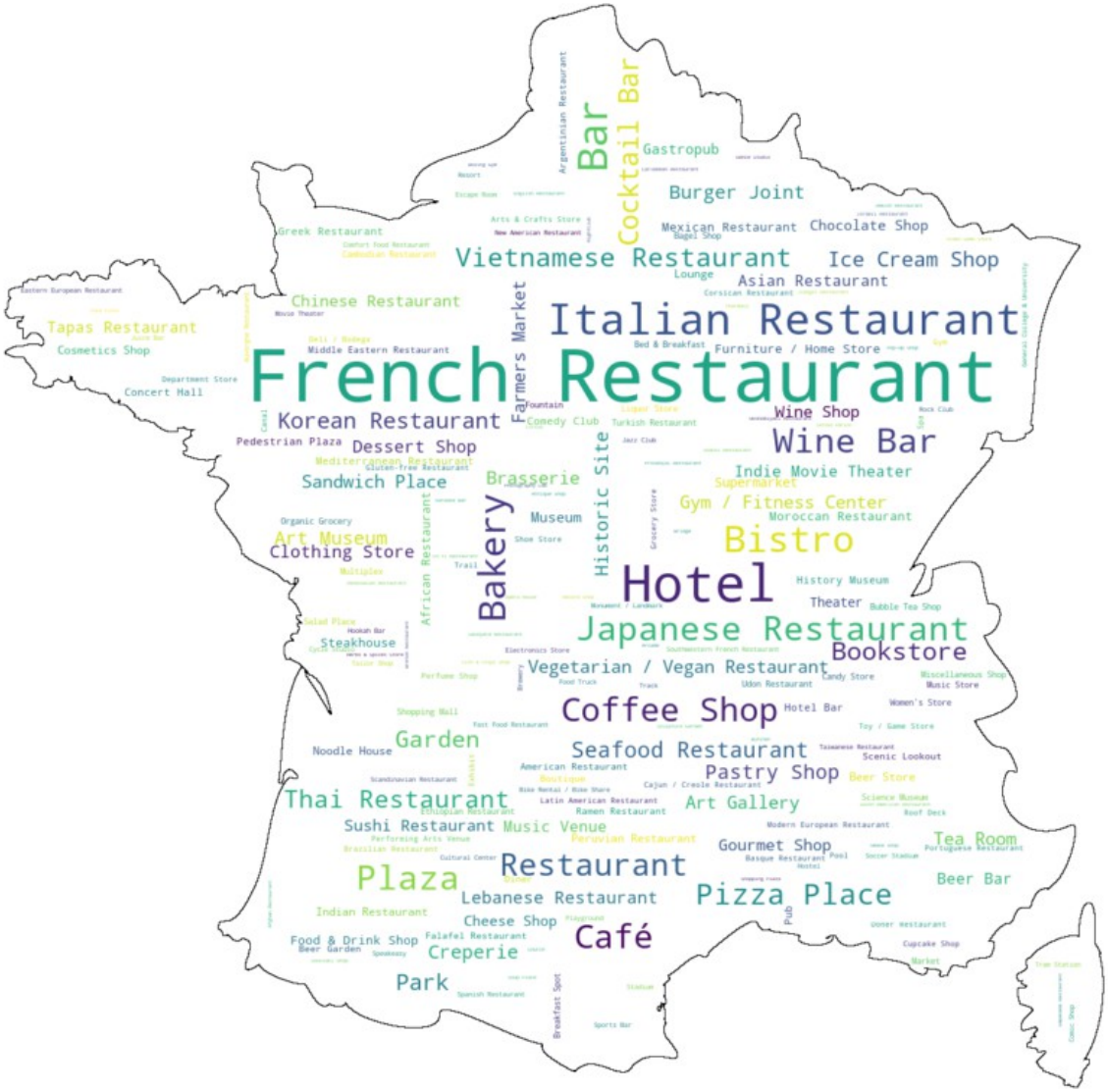


Histogram 1: Weight of the categories in the French Identity matrix

The 2000 venues are classified in 200+ different categories. Most of them are weighted zero (non existing category in Paris) or very low (Rare category in Paris). Most of those data have a very small impact on defining the French identity matrix, but they are still part of making it unique.

Categories with a high number of hits will have an higher weight and so an higher score.

Let's display a word cloud map to have a quick idea of the most prominent terms.

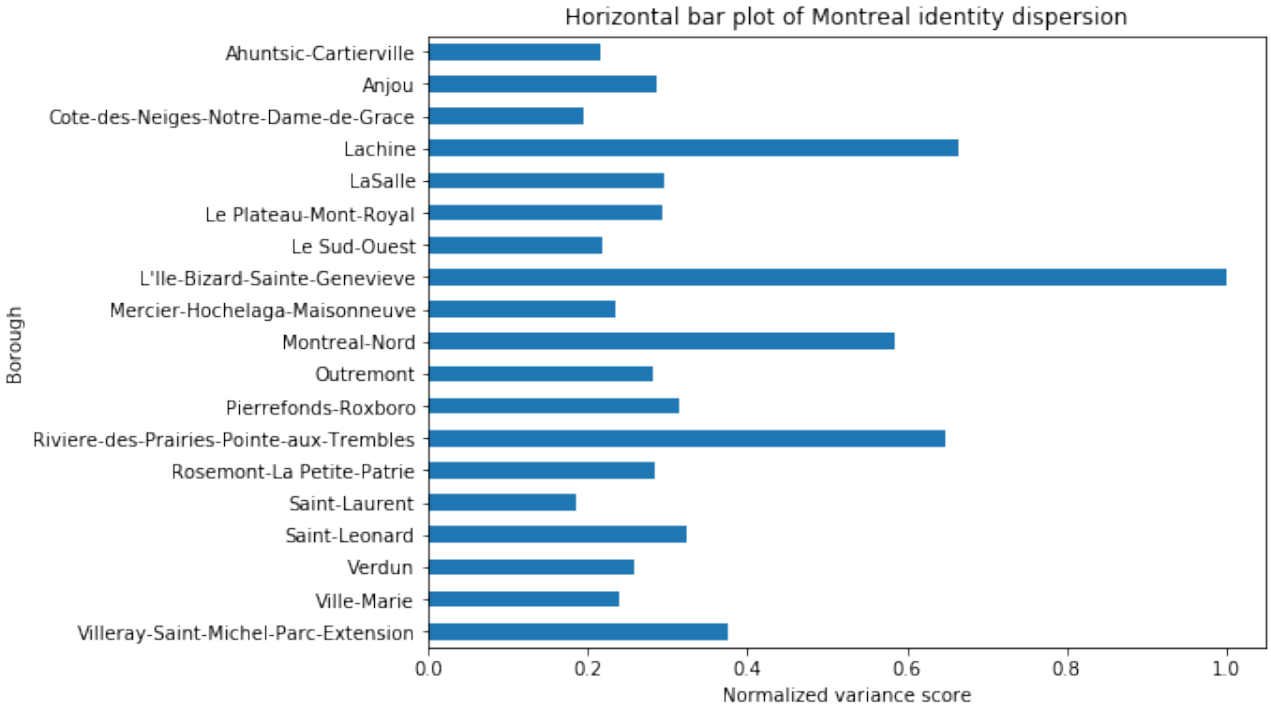


WordCloud 1: Paris top venues categories

2. Montreal boroughs identity weighted matrix – Data Analysis

The Montreal weighted matrix is containing weighted series for every boroughs. But most of the series values are set to zero as each Montreal borough is different (different venues).

What is interesting to analyze is the dispersion of each borough categories. As our previous analyze of the French matrix, the dispersion measures the diversity of the venues in a place.



HorizontalBarChart 1: Measurement of the Montreal identity with top venues dispersion

The representation of the variance in every boroughs is a good indicator of its diversity.

A less dense area will not have many venues and so not many categories. So its weight will be higher as we normalized dividing by the total number of occurences. High weight among zero weight (for the categories non existing in the area) will result in a higher variance.

For example L'Ile-Bizard has the highest variance score. And indeed the island is mostly contained with Parcs and Golf courses.

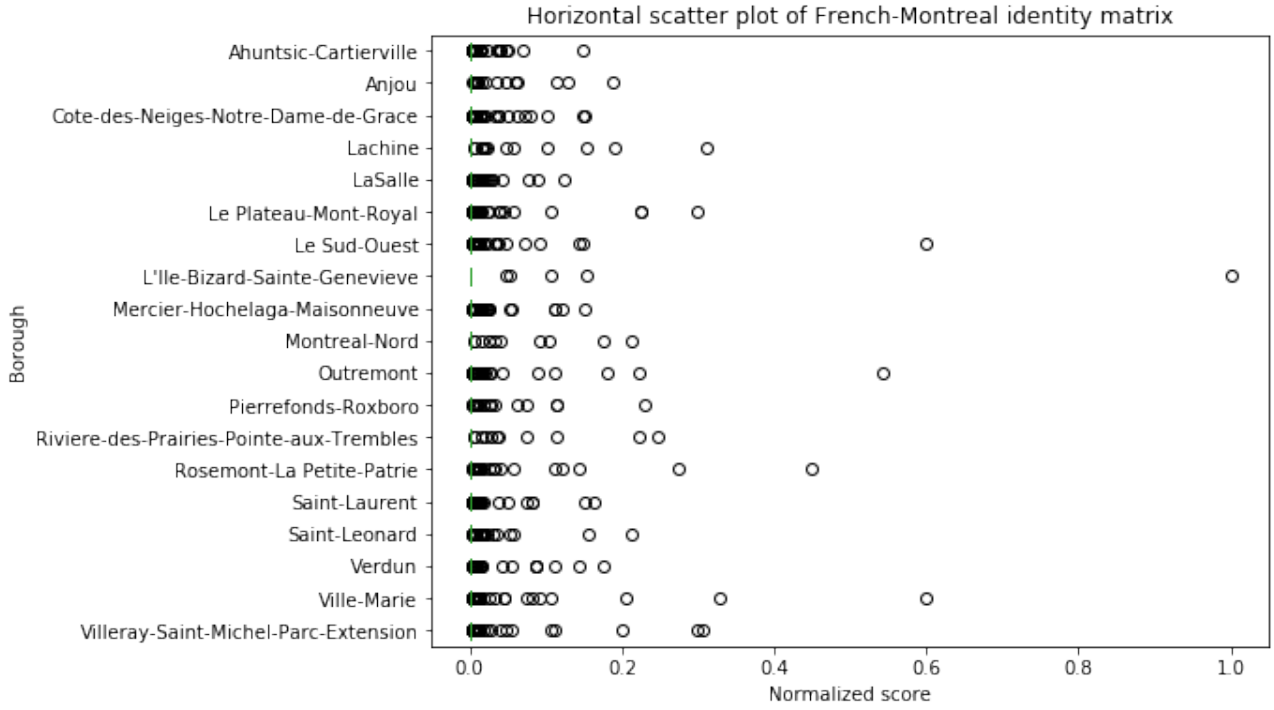
WordCloud 2: Montreal top venues categories

3. New Data matrix : French identity filter applied on Montreal boroughs matrix.

Now we have both **the French identity matrix** and **the Montreal borough identity matrix,** we can calculate our Montreal borough matrix of French culture.

I reweight **the Montreal borough matrix** multiplying every borough (columns) by the **French matrix (the one in Table 3).**

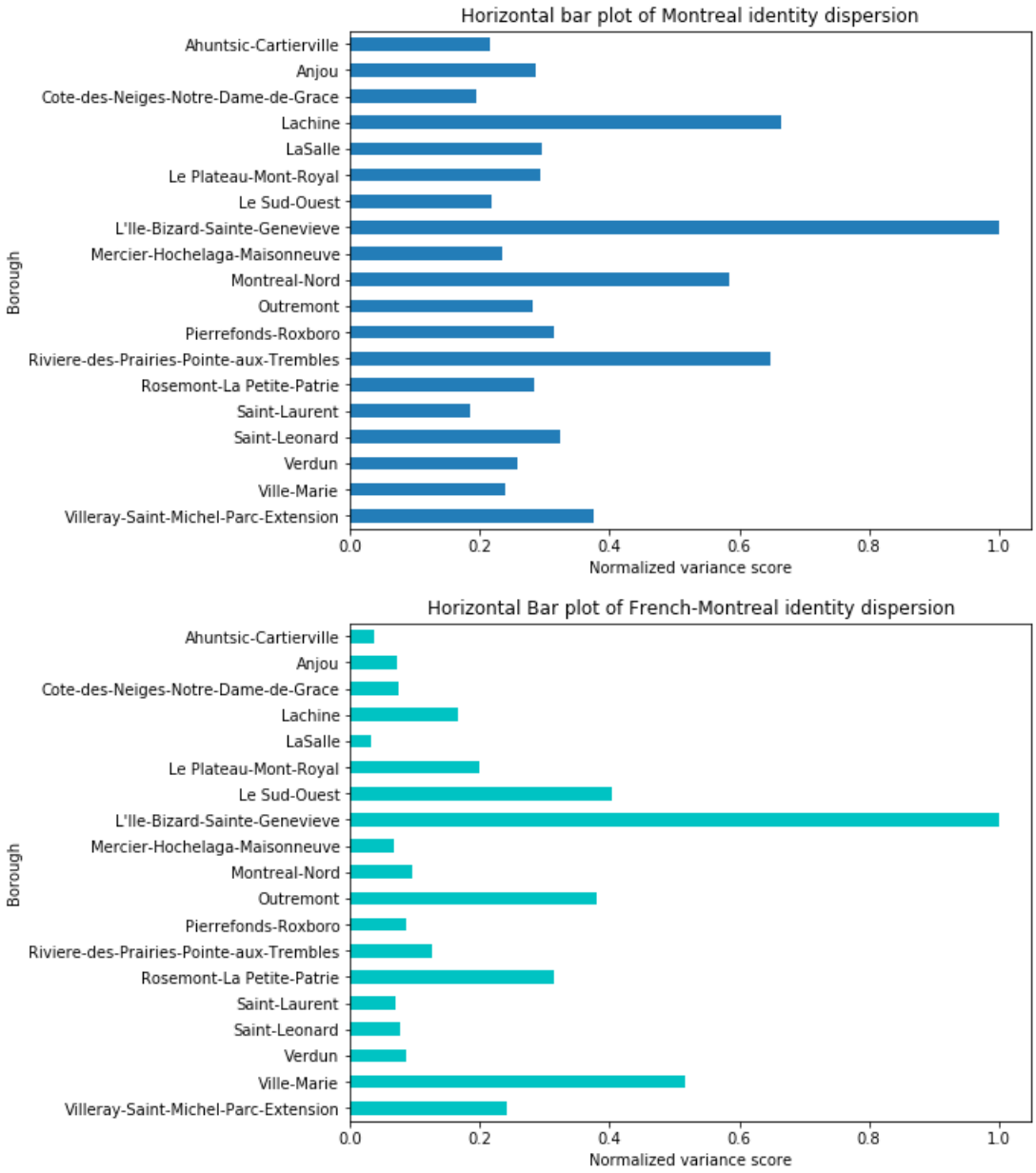
Let's take a look at the resulting matrix.



HorizontalScatterPlot 1: French-Montreal identity matrix

The concept there is to consider the French identity matrix as a filter for French top venues. We applied this filter to Montreal and we can now spot the French top venues in Montreal.

Let's analyze the top venues dispersion of this new filtered Montreal borough identity matrix comparing it with the top venues dispersion of the non-filtered one.



HorizontalBarChart 2: Comparison between old Montreal matrix vs new French-Montreal matrix

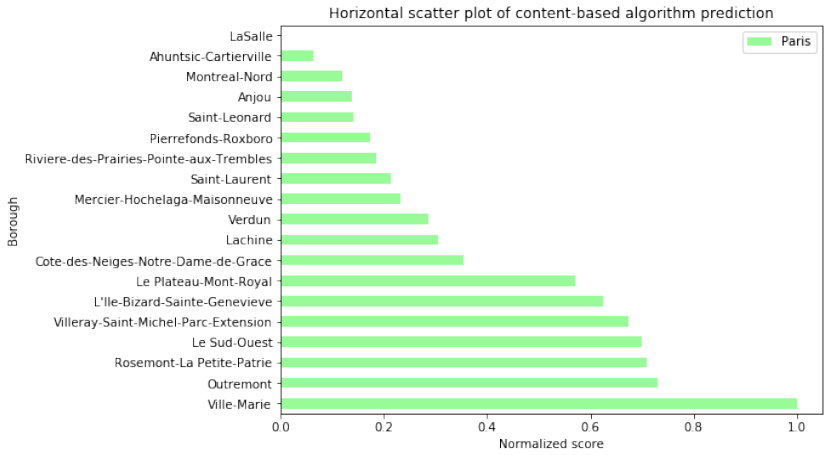
The French filter I applied maximized the variance between the groups (inter groups variance) so we can almost already spot clusters. One cluster of low variance, one of mid variance and one of high variance with L'Ile-Bizard.

IV. Building Estimation models of French clusters

1. Content-based Recommender inspired by Netflix recommendation algorithms

1.1 Applying content-based algorithm

I sum all categories for each borough and sort the result in descending order.



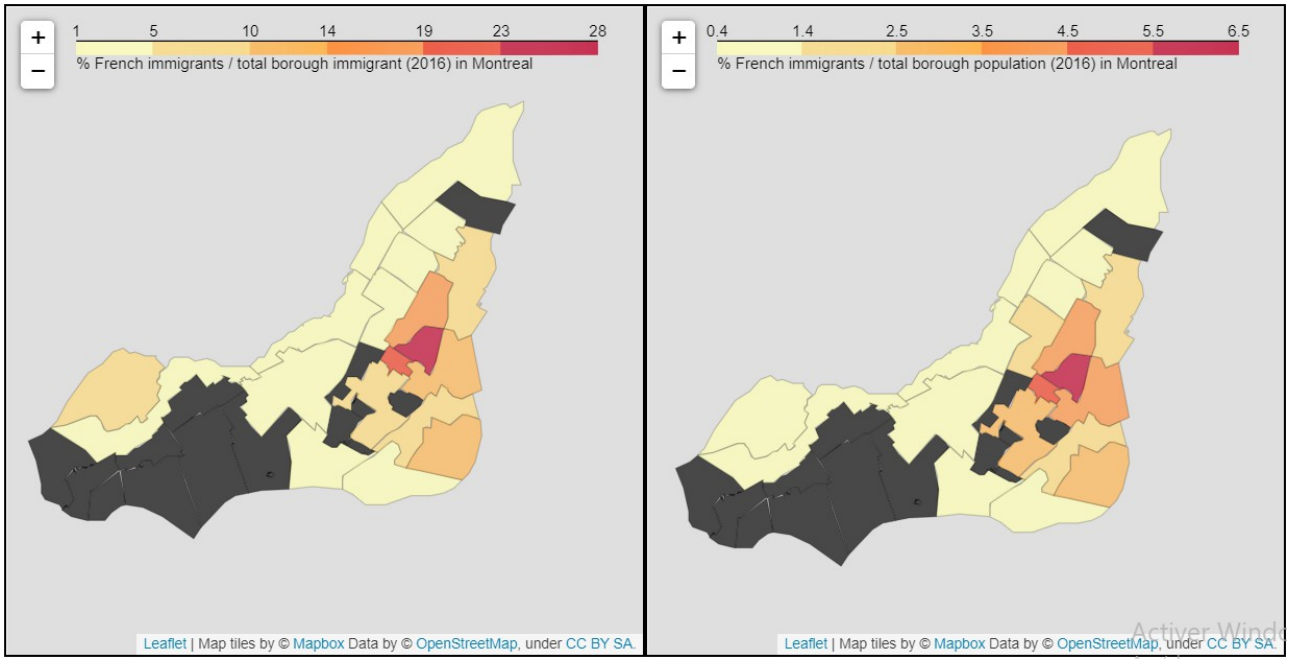
HorizontalBarChart 3: Content-based recommender Matrix - Best match with French Identity

The highest the score is, the best the match is.

We can see **Ville-Marie** is the best match for French cluster estimation. Ville-Marie till Le Plateaux-Mont-Royal are the top scores. Let's verify it comparing with official census data.

1.2 Model validation from official census reports

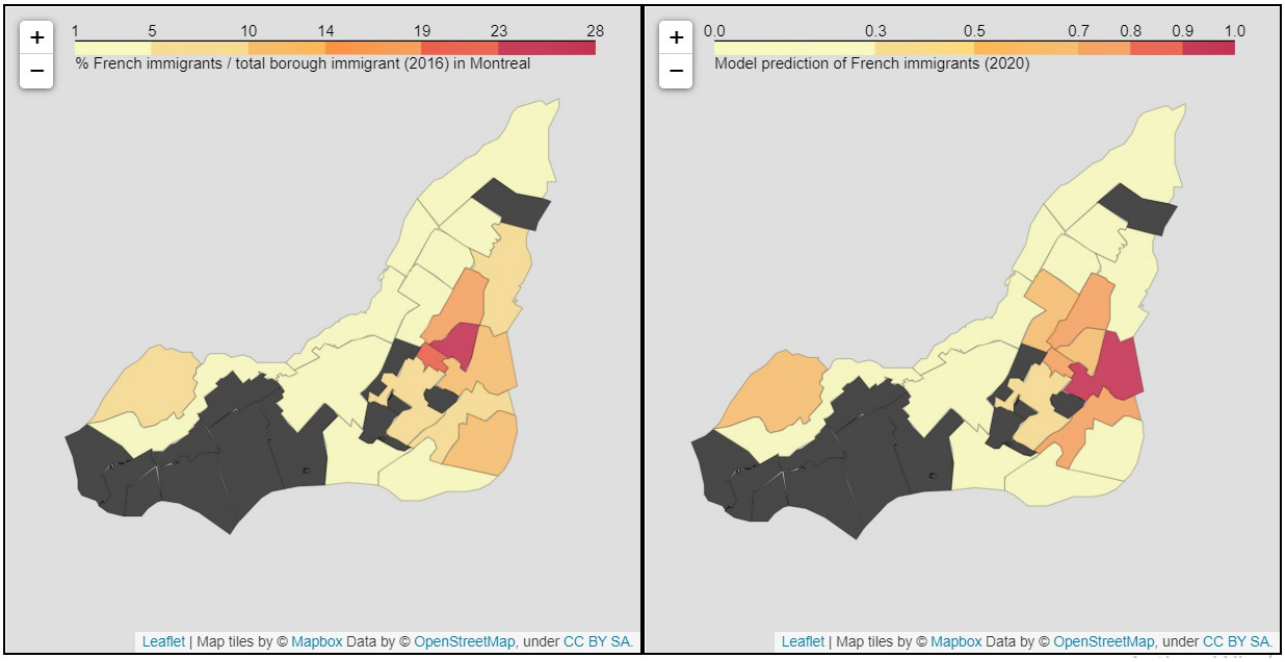
Both maps are from official 2016 census data.



Map 1: Distribution of French immigrant population in Montreal (2016)

I am especially interested in the first map as it represents the weight of French immigrants among others countries immigrants and so it defines better the French influence.

Now let's compare the first map with our model results.



Map 2: Distribution of the French population comparing census data (left) with model prediction (right)

We can see French people are most likely living in the east part of Montreal. This could be explained by the presence of the Francophone university of Montreal or The French embassy for example.

1.2 Quantifying the model efficiency

A way to evaluate this model would be to calcute the RMSE of the borough ranking model (cf. HorizontalBarChart 3) VS borough ranking census.

Root-mean-square error (RMSD)	**Normalized Root-mean-square error (NRMSD)**
3.66	0.203 → 20.3%

Table 5: Performance of content-based model.

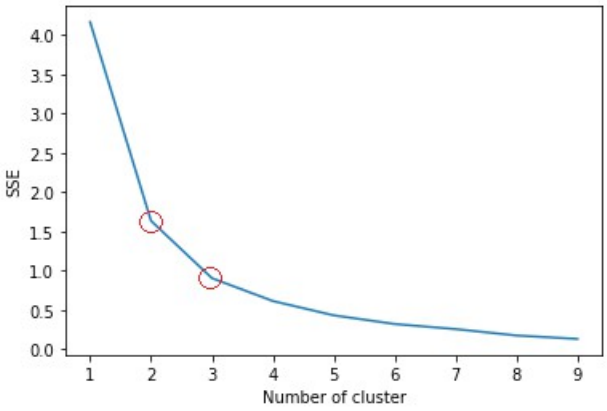
This model gives us a good idea of the area where French people may live but the predicted ranking of the top borough is not so accurate : it represents around 20% of error in the sorted ranking.

We can try to optimize the RMSD making a better match in the ranking. Clustering the model borough and the census data would assign fewer labels than ranking labels : That way we would detect labeled 'group of boroughs' and not specific labeled borough.

2. K-means clustering

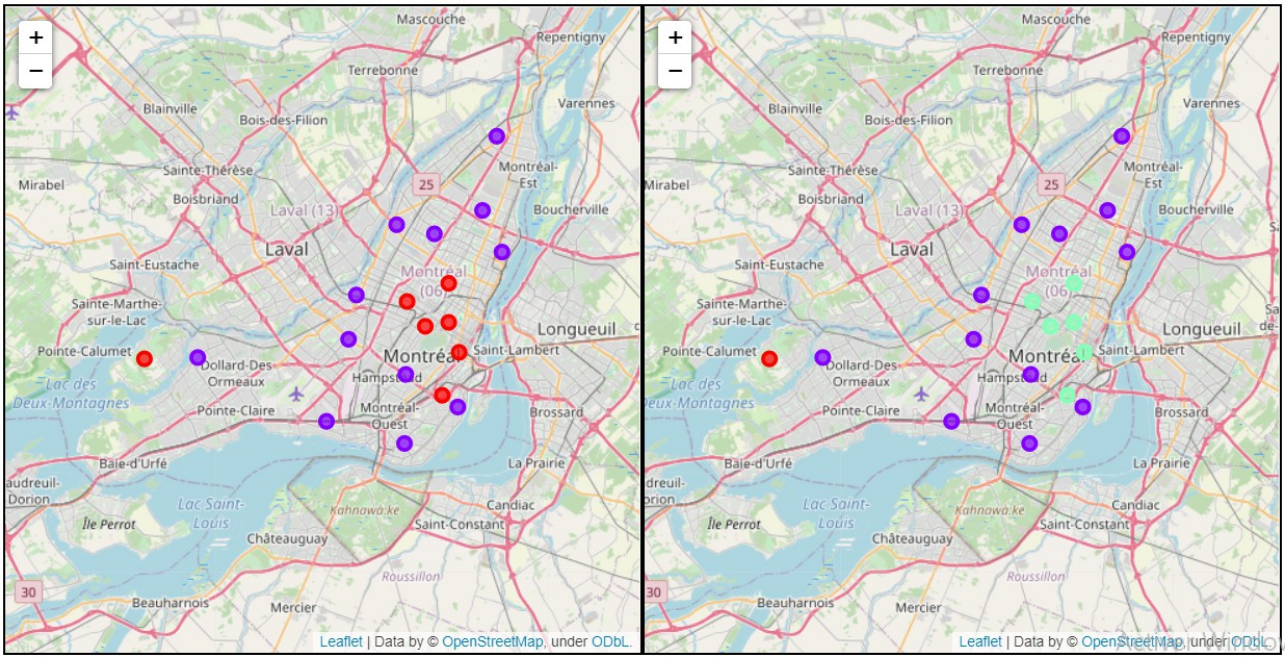
2.1 Finding and applying best K clusters

We apply K-means algorithm to our Montreal borough French-identity weighted matrix.
But first let's try to find the best K number of clusters spotting the elbow keypoint.



Linechart 1: SSE score for K number of cluster

We find best K for K=2 clusters. But K=3 can be as well interesting to study.
Let's take a look at the cluster map of Montreal.

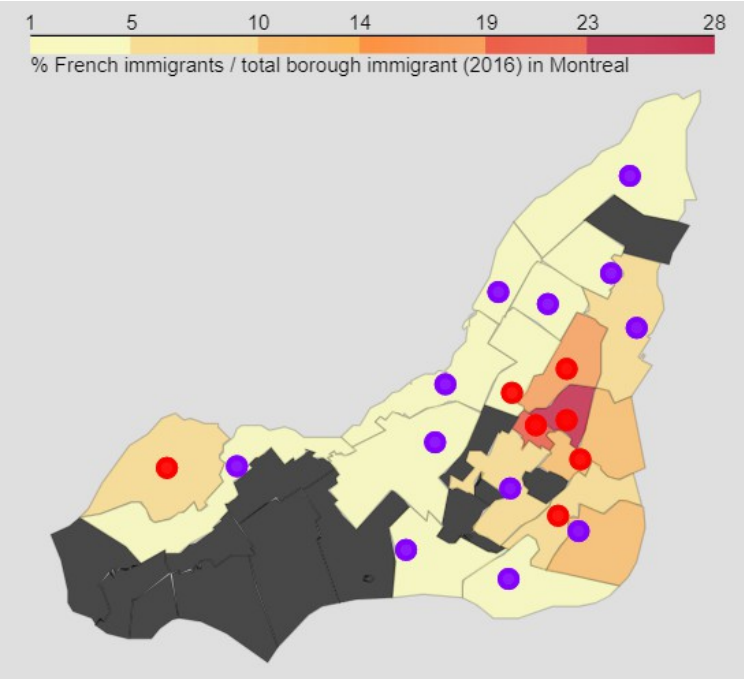


Map 3: K-means clusters for K=2 (left) and K=3 (right)

The difference between K=2 and K=3 clusters remains in the split of L'Ile-Bizard in another cluster. Therefore I will continue considering 2 mains clusters (K=2) but the case of L'Ile-Bizard will be discuss in the result section and later.

2.2 Model validation comparing to official census reports

Now let's compare our kmeans model to official census report.



Map 4: Model clusters vs Choropleth 2016 census map

Red and blue dots are representing French clusters and non-French clusters estimated by my model, respectively. The choropleth map represent **actual** clusters of French people.

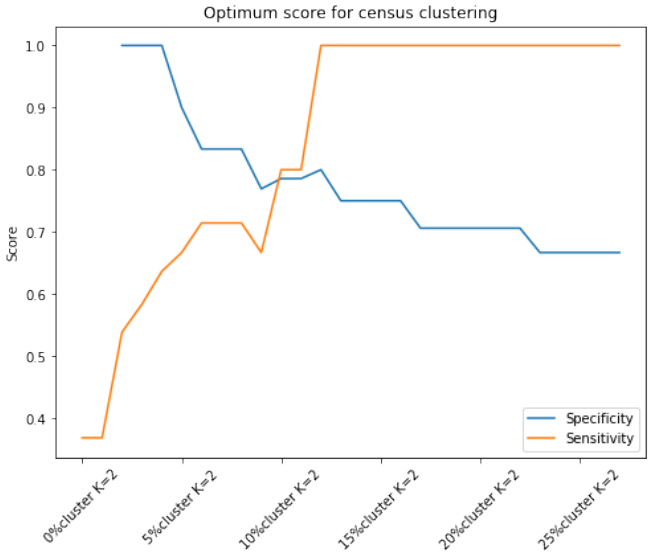
The model gives overall a good estimate. We are going to try to quantify it.

First we need to cluster the census report. We can start stating:

- from >1% acual French immigrants let's label it 'French cluster'. And 'non-French cluster' for <1%. Then calculate different performance scores.
- Then we slide the limit with a 1% step till 28% : >28% French labeled 'French cluster' and <28% labeled 'non-French cluster'.

This iterative system will allows us to find the optimum x% for best performance score.

A good way to spot the optimum x% is to plot the ROC curve but in my opinion the sensitivity vs specificity line chart was more meaningful.

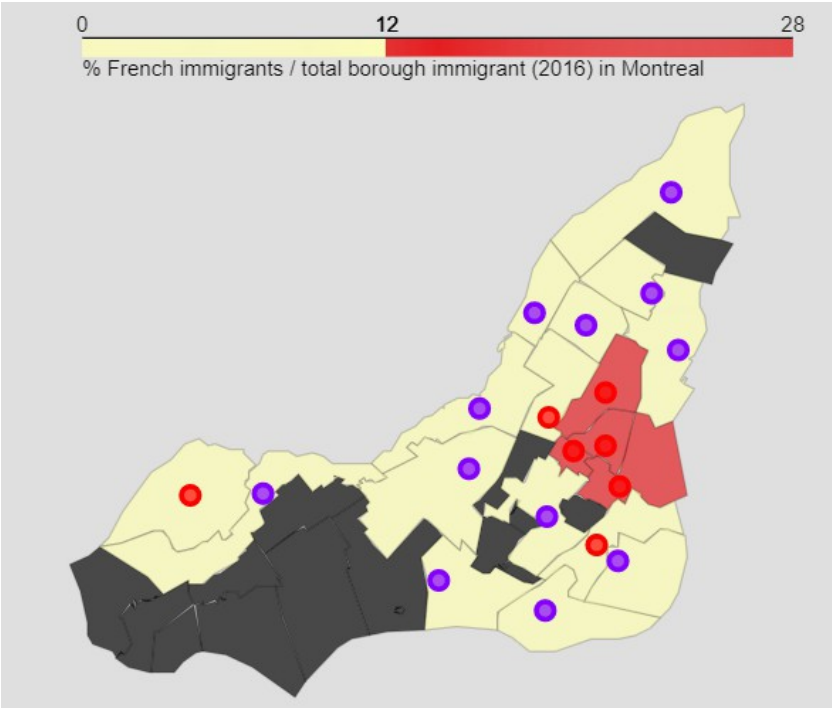


Linechart 2: Specificity vs Sensitivity for optimum census clustering

The best match for sensitivity and specificity is around the cross point where it maximizes both values. '12%' is the best value.

We can now consider a French cluster significant if at least 12% of the total borough immigrant population is French. We apply it to the 2016 census data.

This way we can cluster our choropleth map.



Map 5: Model clustering VS Choropleth census data

2.3 Quantifying the model efficiency

FN	FP	TN	TP	F-1 Score	Jaccard Score	Specificity	Sensitivity	Accuracy	NRMSE
0	3	12	4	0.727	0.429	0.800	1	0.842	0.158

We optimized the RMSE and these values represents the best match we can calculate estimating French cluster .

V. Results section

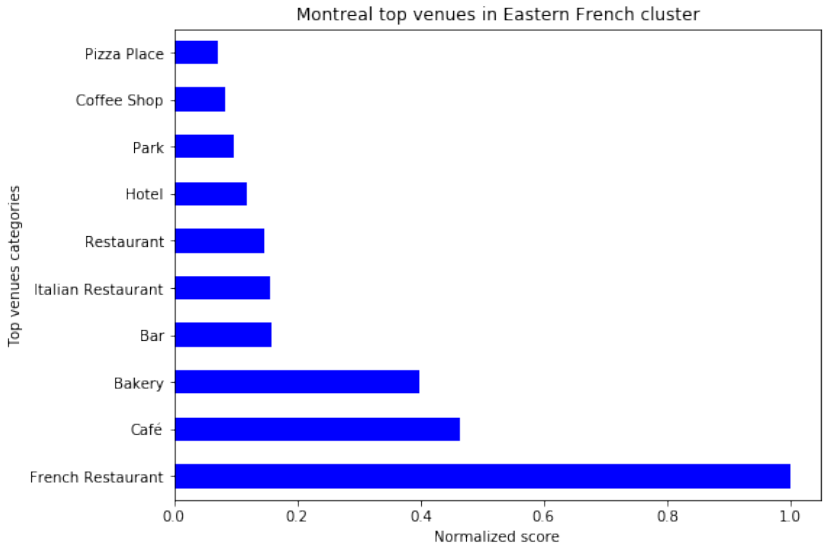
First of all we notice 2 things :

- •Our model gives us a good idea of the presence of French people in Montreal : The East coast is well represented.
- • There are some false clusters estimation around the main cluster and especially on the west side.

1. East coast cluster : - French culture at its best

Le Plateau-Mont-Royal and around are the boroughs with the biggest rate of French people according to the 2016 census.

Let's take a look at the distribution of the top venues in the cluster eastern Montreal.

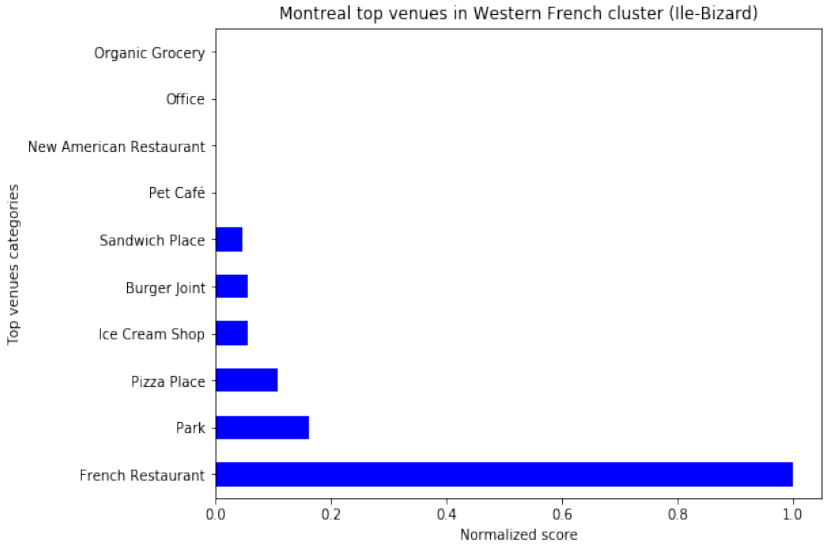


HorizontalBarChart 3: Montreal top venues in Eastern cluster (French cluster)

French Restaurant, Café and Bakery are the top 3 venues this cluster has in common with Paris. They are indeed deeply rooted in France landscape or at least in the Paris one.

2. West coast cluster : - French restaurants and Nature.

This western Montreal cluster is the island L'Ile-Bizard-Sainte-Genevieve.



HorizontalBarChart 4: Montreal top venues in Western cluster (Ile-Bizard)

We already proved L'Ile-Bizard has few top venues looking at it normalized variance (cf. Horizontal BarChart 2) because it is a wide and nature area. We can confirm here as this cluster of French-identity in L'Ile-Bizard is only defined by 6 different categories

The borough is known for its golf course. But Paris does not contained any golf course so it explains why we do not see it in the graph. It has few venues and some of these are the French restaurants, probably in a smaller number than in another Montreal borough, but there they weight big compared to others categories.

VI. Discussion

1. A good idea of the living zone of French people.

Indeed we can recognize the areas spaesely populated by French people are not showing up either in our model. The Choropleth map shows a good estimate of these low-rate places.

Also the models are giving us an interesting cluster on the east coast where around 6 boroughs are sticking out. These boroughs are actually neighborhoods as they are close to each other. It corroborates the concept of immigrant population tending to live geographicaly close to each other. We can validate this trend with the French population in Montreal thanks to the 2016 official census data.

But some parameters should be carefully considered. The east part of Montreal is the most densely populated part of Montreal. Our model is based on Paris top venues which is a very dense city as well. So it may explain why we found the same kind of trending places. It may not be directly link to a concept of culture.

2. The specific case of L'Ile-Bizard-Sainte-Genevieve.

Thanks to our previous calculation (cf. HorizotalBarChart 2; HorizontalScatterPlot 1; HorizontalBarChart 4), we can see the specificity of this borough is to contain very few top venues categories. Meaning his categories matrix is highly weighted and so any match with the French matrix increases drastically its matching score. That is how L'Ile-Bizard score the maximum weight for French restaurant : it may not contained many of these but there are part of the top venues of the island.

First possible explanation could be the town landscape of this region. It is an wide island not that much urbanized except for the towns along the coast. Meaning the borough is not a dense place and so the top venues are quite limited.

Second explanation could be the way to search top venues in a borough : we searched in a circle area from the middle of the borough with a given radius that is the root square of the area of the borough. As stated before **the borough of L'Ile-Bizard-Sainte-Genevieve is a vast area with its main residential zones along the east coast of the island which could lead to a miscalculation of the top venues from this borough particularly.** The algorithm may not include the totality of those towns in the top venues research.

VII. Conclusion

Every country has his own culture. Defining a trend based on people culture sounds possible and quite interesting. We started stating someone living abroad will tend to live in a place closer to his initial culture and will tend to live closer to people from his home country. And these communities can have an impact on the landscape of their neighborhood.

I was able to spot the main cluster of French people in Montreal based only on the top venues of the city. It is a low investment model to cluster immigrants culture expression in a city and we obtian interesting results. Therefore it can be improved.

First in this situation, we define our French model on the city of Paris only which represents 3.1% of the total population in France. It gives us an imprecise idea of the French culture which is for sure more complex. A database compiling proportionally the top venues of every France cities could be a closer match.

Then we were searching a top 100 venues by borough resulting in a 2000 venues for Paris and 1395 venues for Montreal. It is not enough and we can consider searching for top 100 venues in smaller areas, then merge all those areas to obtain an even truer identity of a city and a more populated database.
We can as well rethink the way to weight data. Giving more importance to the top categories of the city may give us a better 'filter'.

Finally this study is about Montreal, a cosmopolitan city from a democratic country. It is not true everywhere abroad and it will not necessary work in a city/ country with a different immigration policy.
To definitely validate my model I will have to test it with many others cities around the world and with others cultures to cluster.