IBM Data Science- Capstone Project Report
The battle of neighborhoods

Classification of Montreal neighborhoods

Introduction

Montreal is the second most-populous city in Canada, and the size and the shape of the island of Montreal makes daily transportations a challenge.

A good classification of the different neighborhoods by types of venues could help planning the transportation resources in the different times of the day. For instance, an area with many bars and night-clubs will not require transportation resources (bus, taxi…) at the same time of the day as for an area with mainly offices or commercial venues.

A taxi company could use a classification analysis on the neighborhood to plan a better dispatch of the taxi float at each time of the day. Also, the city transportation system could also use this input along with all the statistics it is able to collect. It can also help for an expansion of the metro network which is likely to happen in Montreal in a close future.

Data

We will use the Foursquare API to find data on the venues in each neighborhood. To use Foursquare we need geographic coordinates (latitude, longitude) for each neighborhood.

Neighborhoods data

We start by scraping the Wikipedia page of the postal codes of Montreal:

	0	1	2	3	4	5	6	7	8
0	H0ANon assigné	H1APointe-aux-Trembles	H2ASaint-MichelEst	H3ACentre-ville de MontréalNord (Université Mc...	H4ANotre-Dame-de-GrâceNord-est	H5APlace Bonaventure	H7ADuvernay-Est	H8ANon assigné	H9ADollard-Des-OrmeauxNord-ouest
1	H0BNon assigné	H1BMontréal-Est	H2BAhuntsicNord	H3BCentre-ville de MontréalEst	H4BNotre-Dame-de-GrâceSud-ouest	H5BComplexe Desjardins	H7BSaint-François	H8BNon assigné	H9BDollard-Des-OrmeauxEst
2	H0CNon assigné	H1CRivière-des-PrairiesNord-est	H2CAhuntsicCentre	H3CGriffintown(Incluant Île Notre-Dame & Île S...	H4CSaint-Henri	H5CNon assigné	H7CSaint-Vincent-de-Paul	H8CNon assigné	H9CL'Île-BizardNord-est
3	H0ENon assigné	H1ERivière-des-PrairiesSud-ouest	H2EVillerayNord-est	H3EÎle des Sœurs	H4EVille Émard	H5ENon assigné	H7EDuvernay	H8ENon assigné	H9EL'Île-BizardSud-ouest
4	H0GNon assigné	H1GMontréal-NordNord	H2GPetite-PatrieNord-est	H3GCentre-ville de MontréalSud-est (Université...	H4GVerdunNord	H5GNon assigné	H7GPont-Viau	H8GNon assigné	H9GDollard-Des-OrmeauxSud-ouest

Note: *The tables displayed in the reports show only the first five lines of the table*

We a bit of work on the DataFrame we obtain this table:

	Postal Code	Neighborhood
0	H1A	Pointe-aux-Trembles
1	H2A	Saint-MichelEst
2	H3A	Centre-ville de MontréalNord (Université McGill)
3	H4A	Notre-Dame-de-GrâceNord-est
4	H5A	Place Bonaventure

Now we need to obtain the geographic coordinates of the neighborhoods. After many unsuccessful tries with geopy geocode which is not reliable with incomplete addresses, I was lucky to find the coordinates for most of the postal codes in a table on the web:

	Unnamed: 0	Place	Code	Country	Admin1	Admin2	Admin3
0	1.0	Mont-Joli	G5H	Canada	Quebec	Bas-Saint-Laurent	Mont-Joli
1	NaN	48.584/-68.192	48.584/-68.192	48.584/-68.192	48.584/-68.192	48.584/-68.192	48.584/-68.192
2	2.0	Duvernay-Est	H7A	Canada	Quebec	NaN	NaN
3	NaN	45.674/-73.592	45.674/-73.592	45.674/-73.592	45.674/-73.592	45.674/-73.592	45.674/-73.592
4	3.0	Saint-Vincent-de-Paul	H7C	Canada	Quebec	NaN	NaN

Here again we have to reformat:

	Code	Place	Latitude	Longitude
0	G5H	Mont-Joli	48.584	-68.192
1	H7A	Duvernay-Est	45.674	-73.592
2	H7C	Saint-Vincent-de-Paul	45.617	-73.649
3	H7E	Duvernay	45.623	-73.695
4	H7G	Pont-Viau	45.577	-73.687

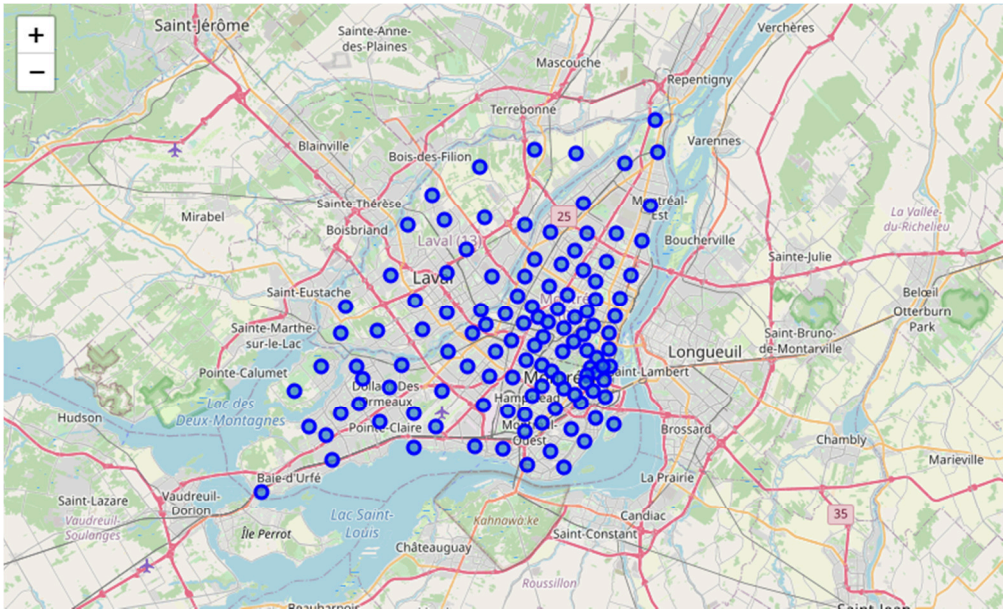
Now we can merge Wikipedia data with the coordinate's dataset:

Code	Neighborhood	Latitude	Longitude
H1A	Pointe-aux-Trembles	NaN	NaN
H2A	Saint-MichelEst	NaN	NaN
H3A	Centre-ville de MontréalNord (Université McGill)	45.504	-73.575
H4A	Notre-Dame-de-GrâceNord-est	45.472	-73.615
H5A	Place Bonaventure	NaN	NaN

Some of the coordinates were missing in the table found on the web, and to get the missing coordinates, the only solution is to search them manually on google map, we will also use google map to modify inaccurate coordinates. Now the table is complete:

Code	Neighborhood	Latitude	Longitude
H1A	Pointe-aux-Trembles	45.674145	-73.500435
H2A	Saint-MichelEst	45.561809	-73.601338
H3A	Centre-ville de MontréalNord (Université McGill)	45.504000	-73.575000
H4A	Notre-Dame-de-GrâceNord-est	45.472000	-73.615000
H5A	Place Bonaventure	45.499840	-73.565970

And we can display the neighborhoods on a map:



We will limit the scope of the study to the island of Montreal so we get rid of the Neighborhoods North of Riviere-des-prairies.

Foursquare data: venues and categories

We will use Foursquare API to explore venue categories in each neighborhood. Venues can be categorized as residential, professional, shopping or leisure. We need to know what the venue categories are in the Foursquare database.

With a request to the API, we obtain the following categories with their identification code (needed to request the venues for one particular category):

```
Arts & Entertainment           4d4b7104d754a06370d81259
College & University           4d4b7105d754a06372d81259
Event                          4d4b7105d754a06373d81259
Food                           4d4b7105d754a06374d81259
Nightlife Spot                 4d4b7105d754a06376d81259
Outdoors & Recreation          4d4b7105d754a06377d81259
Professional & Other Places    4d4b7105d754a06375d81259
Residence                      4e67e38e036454776db1fb3a
Shop & Service                 4d4b7105d754a06378d81259
Travel & Transport             4d4b7105d754a06379d81259
```

There are 10 top categories that we will use to classify the neighborhoods.

▪ Methodology

For each neighborhood, we request to the Foursquare API the number of venues in each category and add the results in the neighborhoods dataframe:

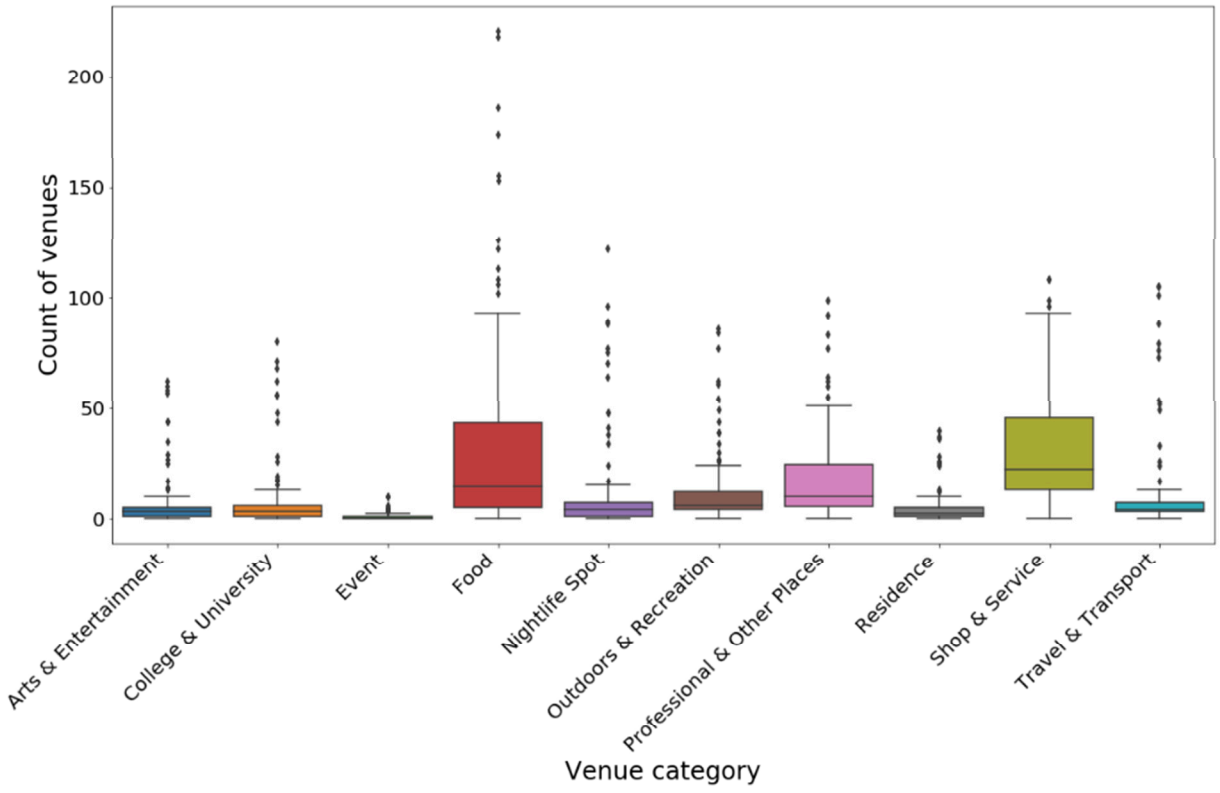
Code	Neighborhood	Latitude	Longitude	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
H1A	Pointe-aux-Trembles	45.674145	-73.500435	1	1	0	4	0	5	2	0	4	5
H2A	Saint-MichelEst	45.561809	-73.601338	3	2	1	19	8	6	13	2	28	5
H3A	Centre-ville de MontréalNord (Université McGill)	45.504000	-73.575000	57	80	4	186	88	77	83	36	108	88
H4A	Notre-Dame-de-GrâceNord-est	45.472000	-73.615000	17	1	0	84	6	9	25	2	51	6
H9A	Dollard-Des-OrmeauxNord-ouest	45.495801	-73.832858	0	2	0	5	2	4	4	0	2	2

We can calculate the total number of venues for each neighborhood:

Code	Neighborhood	Latitude	Longitude	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport	Total venues
H1A	Pointe-aux-Trembles	45.674145	-73.500435	1	1	0	4	0	5	2	0	4	5	22
H2A	Saint-MichelEst	45.561809	-73.601338	3	2	1	19	8	6	13	2	28	5	87
H3A	Centre-ville de MontréalNord (Université McGill)	45.504000	-73.575000	57	80	4	186	88	77	83	36	108	88	807
H4A	Notre-Dame-de-GrâceNord-est	45.472000	-73.615000	17	1	0	84	6	9	25	2	51	6	201
H9A	Dollard-Des-OrmeauxNord-ouest	45.495801	-73.832858	0	2	0	5	2	4	4	0	2	2	21

Data exploration

Let's print a boxplot for the number of venues in each category to understand more the structure of the data:



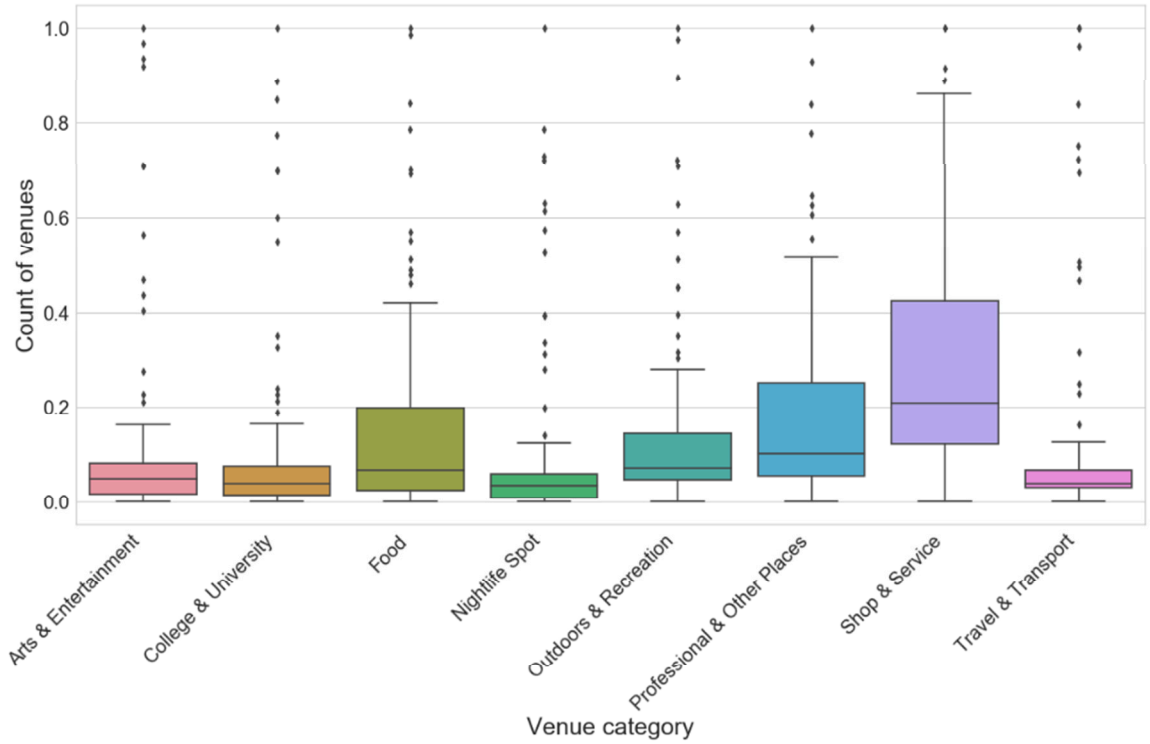
We see that the event & residence categories has very few numbers in the different areas, we can drop these categories as it won't make any big differences (Also most of residences are not in Foursquare data so it is not relevant to use this category)

Data preprocessing

We normalize the data with MinMaxScaler so each category has an equal impact on the study. Otherwise food places and shops would have a bigger impact due to their higher figures:

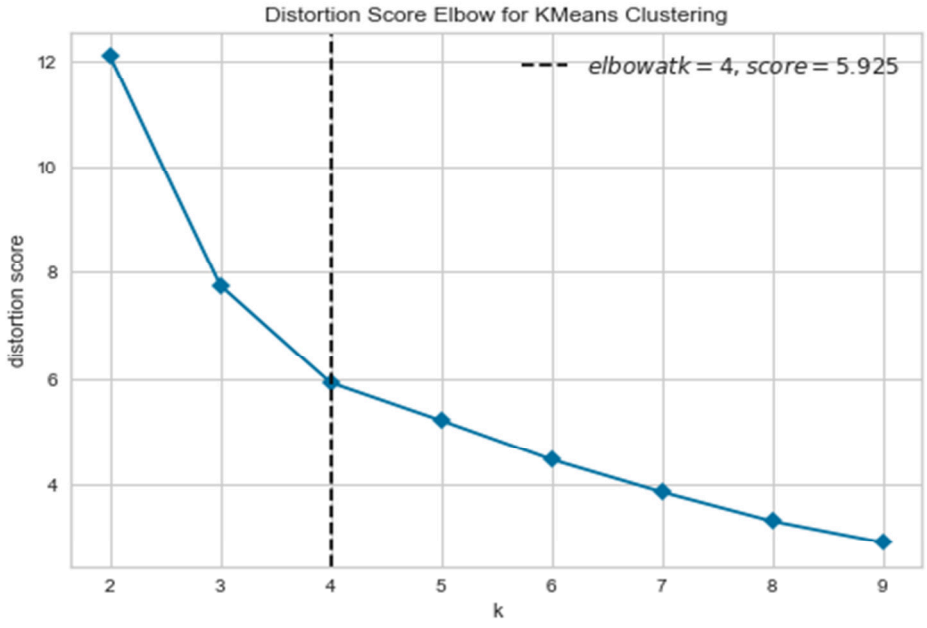
	Arts & Entertainment	College & University	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Shop & Service	Travel & Transport
0	0.016129	0.0125	0.018100	0.000000	0.058140	0.020202	0.037037	0.047619
1	0.048387	0.0250	0.085973	0.065574	0.069767	0.131313	0.259259	0.047619
2	0.919355	1.0000	0.841629	0.721311	0.895349	0.838384	1.000000	0.838095
3	0.274194	0.0125	0.380090	0.049180	0.104651	0.252525	0.472222	0.057143
4	0.000000	0.0250	0.022624	0.016393	0.046512	0.040404	0.018519	0.019048

Let's plot again the normalized data:



Clustering

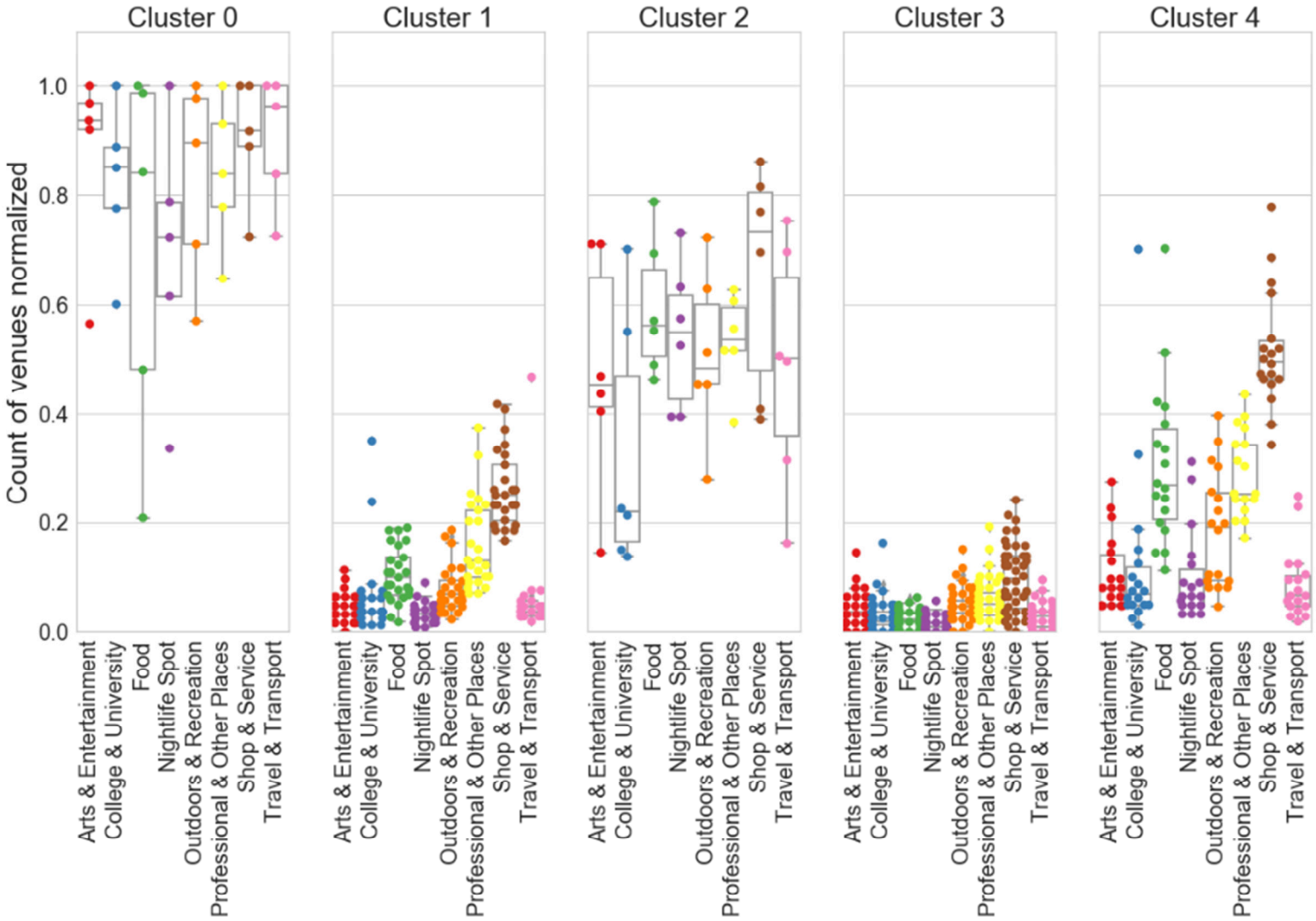
By testing the k-means clustering algorithm with different number of clusters, we start from 2 and increasing up to 5 adds good and easy to interpret clustering but beyond 5 the clustering becomes difficult to interpret so we choose to keep a number of 5 clusters for the study. To validate this empirical interpretation, we use the elbow method:



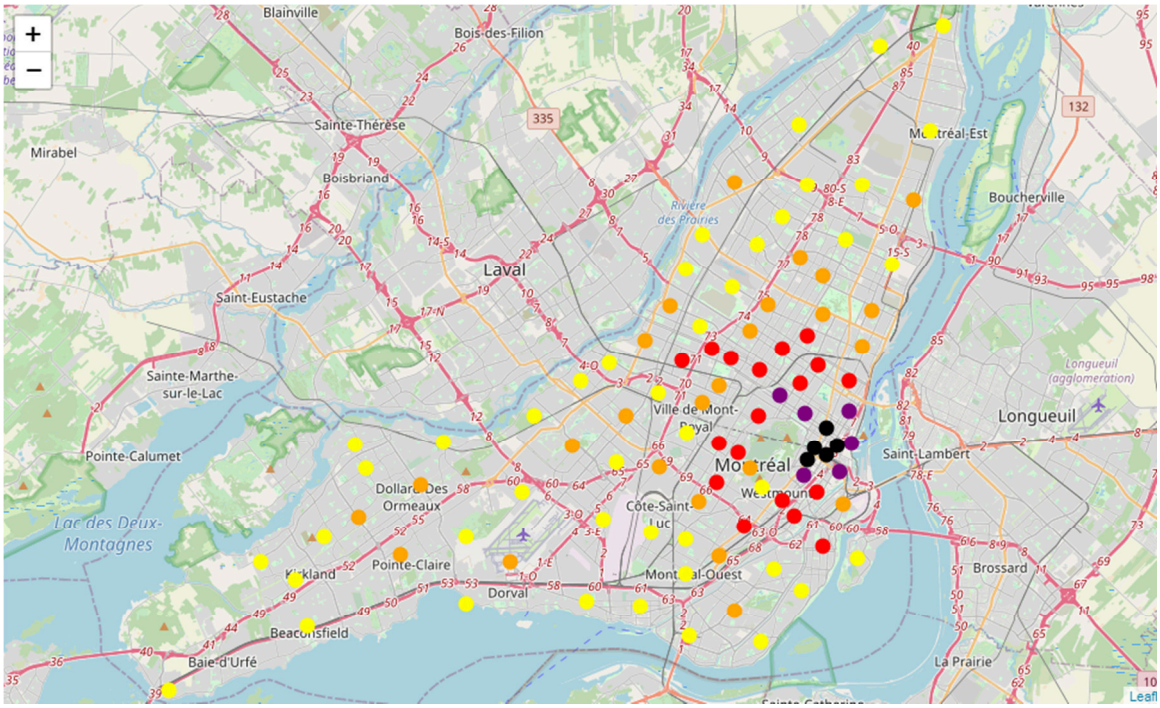
The elbow method (inflexion point of the distortion score of the clusters) gives 4 as an optimal number of clusters...

Since we still find a way to interpret the clustering with 5 clusters as we will see in the following plot we stick with 5

We can use a boxplot in each cluster to help the interpretation:



Let's see the result on the map:



Results

By looking at the boxplots we can characterize the clusters:

- Cluster 0 (Black) has the highest scores for all venue categories. These neighborhood are the economic and cultural center of Montreal
- Cluster 2 (Purple) has high score for all venue categories
- Cluster 4 (Red) has lower scores with best scores in Shops, Professional places & Food. These areas are mostly commercials
- Cluster 1 (Orange) has low marks with better scores for Shops and Professional places. This cluster correspond to commercial & residential suburban areas
- Cluster 3 (Yellow) is mostly residential, with very low scores everywhere

After viewing the map:

- Cluster 0 correspond to downtown Montreal
- Cluster 2 is the downtown immediate periphery.
- Cluster 4 corresponds to secondary cultural and economic centres where the first big waves of immigration (Italians, Greeks…) settled. These area are in a process of gentrification.

- Clusters 1 and 3 aren't so clearly geographically distributed but most of the orange points (cluster 1) are located close to downtown

Discussion

This clustering study relies only on the Foursquare Data. We cannot be certain that the data is complete since some venues could be missing. For instance, small shops or professional places data is very likely to be incomplete since Foursquare main purpose is to give tips on food places and cultural/touristic venues.

Also the study doesn't consider the size or importance of the venues and one venue such as a train station surely has a really bigger impact on the attractivity of an area than a restaurant.

Conclusion

The clustering method proved to be a good way to understand the economic and cultural dynamic of a city. The more various, complete and precise is the data, the more the clusters will give clear insights and the more we will give credit to the results.