Where to open a Movie Theater in Montreal.

1. Introduction

In this project, we are going to look for an optimal location to open a movie theater. Specifically, this report can provide a reference for stakeholders who are interested in **opening a movie theater in Montreal, Quebec, Canada**.

Montreal is the second-largest city in Canada and the largest city in the province of Quebec, located along the Saint Lawrence River at its junction with the Ottawa River. It sits on an island. In this report, we will focus on all areas on the Montreal island. There are many movie theaters on Montreal island, we will **conclude where are the existing movie theaters**. Then we will use a clustering model to **find similar areas** on the island considering demographic data of each borough and region. The preferred area shall be **distant from existing movie theaters**.

We will use data science tools to fetch the raw data, visualize it then **generate a few most promising areas based on the above criteria**. In the meanwhile, we will also explain the advantage and traits for the candidates, so that **stakeholders can make the final decision** base on the analysis.

2. Data

Based on the definition of our problem, factors that may impact our decision are:

- Demographic information, e.g. population, density, education, age, income.
- Number of existing shopping malls in the neighborhood and nearby.
- Number of existing movie theatres in the neighborhood and nearby.

We decided to use a regularly spaced grid of locations all around the whole Montreal island, to define our neighborhoods. Concretely, we will use popular hexagon honeycomb to define our neighborhoods.

In this project, we will fetch or extract data from the following data sources:

- Montreal census information of the 2016 year.
- Centers of hexagon neighborhoods will be generated algorithmically and approximately addresses of centers of those areas will be obtained using Google Geocoding API.
- Shopping malls and movie theaters data in every neighborhood will be obtained using Foursquare API.
- Coordinate of Montreal center will be obtained using Google Geocoding API of well-known Montreal location.
- Montreal borough shapefile is obtained from Carto.

Montreal Island Shape File

To show the Montreal island boundary in the **folium** map, we need a **geojson** definition file for Montreal island. We downloaded this shapefile from the Carto website.

Folium

It's not difficult to use `folium`, just required a few lines of code to show Montreal island with boundary data.

Figure 1: Montreal Island in Folium map.

Next step, we want to generate candidate cells in the map, more specifically, only within Montreal island. It's popular to use the honeycomb hexagon grid when dealing with problems related to the map. Unlike circle, there is no spacing among hexagons which make sure no missing area. Furthermore, the distance between any two adjacent hexagons is the same.

Unfortunately, Folium doesn't provide native support to draw hexagon in the map view, we have to write some code to support this feature.

We write a method to calculate the hexagon vertices' coordinates by giving centroids coordinates and length of the side. After that, we generate a honeycomb hexagon grid throughout the island.

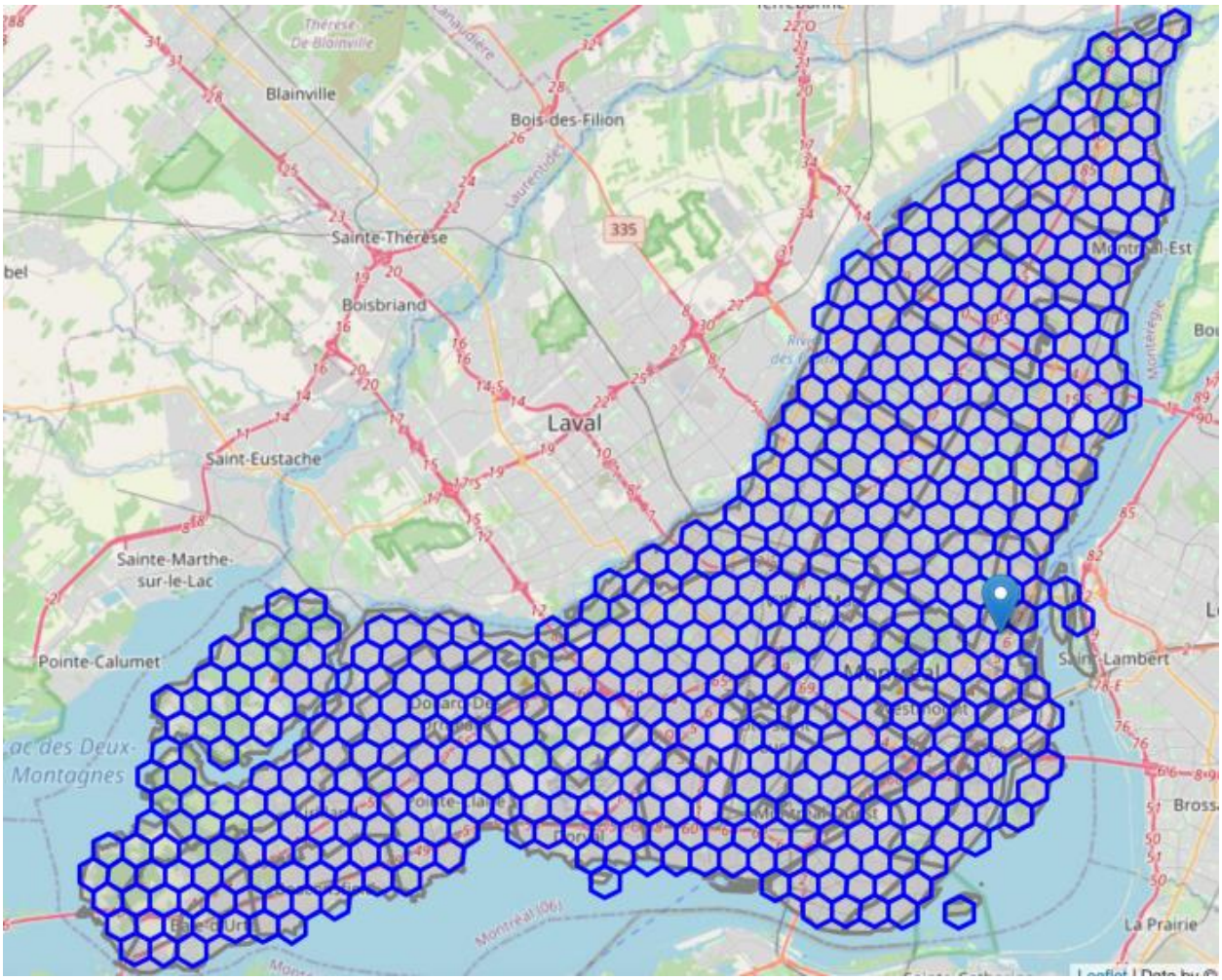

Figure 2: Honeycomb hexagon grid in the Montreal map.

So far we created a honeycomb grid on the island and we generated the center coordinates for each hexagon. We will use Google Geocoding API to reversely lookup the address accordingly. It requires a Google API key to use this set of APIs. It can be applied from Google Developer Console.

Let's put all the data in a Pandas Dataframe, and show the first 10 items. Each row contains the center address of a hexagon and corresponding latitude and longitude degrees which are in WGS84 spherical coordinate system, X/Y columns are in UTM Cartesian coordinate system which uses the common metric unit — meter or kilometer.

	Address	Latitude	Longitude	X	Y	Distance from downtown
0	2 Rue Forbes, Sainte-Anne-de-Bellevue, QC H9X 1W8	45.409692	-73.950681	582110.849218	5.028999e+06	32359.175834
1	17 Rue East Cottages, Sainte-Anne-de-Bellevue,...	45.409574	-73.937904	583110.849218	5.028999e+06	31426.340867
2	714 Rue Victoria, Baie-d'Urfé, QC H9X 2K7	45.409455	-73.925127	584110.849218	5.028999e+06	30497.762869
3	20 Avenue Morningside, Senneville, QC H9X 1A9	45.417545	-73.956926	581610.849218	5.029865e+06	32534.078157
4	21123 Ch Ste-Marie, Sainte-Anne-de-Bellevue, Q...	45.417427	-73.944147	582610.849218	5.029865e+06	31590.582160
5	Rond-point Clark-Graham, Baie-d'Urfé, QC H9X 4B6	45.417308	-73.931368	583610.849218	5.029865e+06	30650.669175
6	90 Rue Morgan, Baie-d'Urfé, QC H9X 3A8	45.417188	-73.918589	584610.849218	5.029865e+06	29714.679210
7	207 Rue Calais, Baie-d'Urfé, QC H9X 2L6	45.417066	-73.905811	585610.849218	5.029865e+06	28782.994984
8	Beaconsfield / Redfern, Beaconsfield, QC H9W 4M9	45.416943	-73.893032	586610.849218	5.029865e+06	27856.048534
9	458 Rue Lakeshore, Beaconsfield, QC H9W 4J6	45.416819	-73.880254	587610.849218	5.029865e+06	26934.329017

Figure 3: Dataframe of candidate hexagons.

Foursquare API

The Foursquare Places API offers real-time access to Foursquare's global database of rich venue data and user content to power your location-based experiences in your app or website.

Now we generated all the candidate neighborhoods on Montreal island, we will get all movie theaters information using Foursquare API. From Foursquare API documentation, we can find the corresponding movie theater category in Venue Categories. The corresponding ID of **Movie Theater** in Foursquare API is `4bf58dd8d48988d17f941735` which is under **Arts & Entertainment** main category. It contains several sub-categories:

- Drive-in Theater, id: `56aa371be4b08b9a8d5734de`
- Indie Movie Theater, id: `4bf58dd8d48988d17e941735`
- `Multiplex, id: 4bf58dd8d48988d180941735`

Unlike coffee shops, restaurants everywhere, there aren't lots of movie theaters in the region, it also makes sense since we don't expect movie theater in every neighborhood.

Let's fetch all the movie theaters on Montreal island first. To do so, we will fetch movie theaters data in each borough and municipality. From the response of Foursquare APIs, there are a total of 44 movie theaters on Montreal island. Let's plot it in a map view.

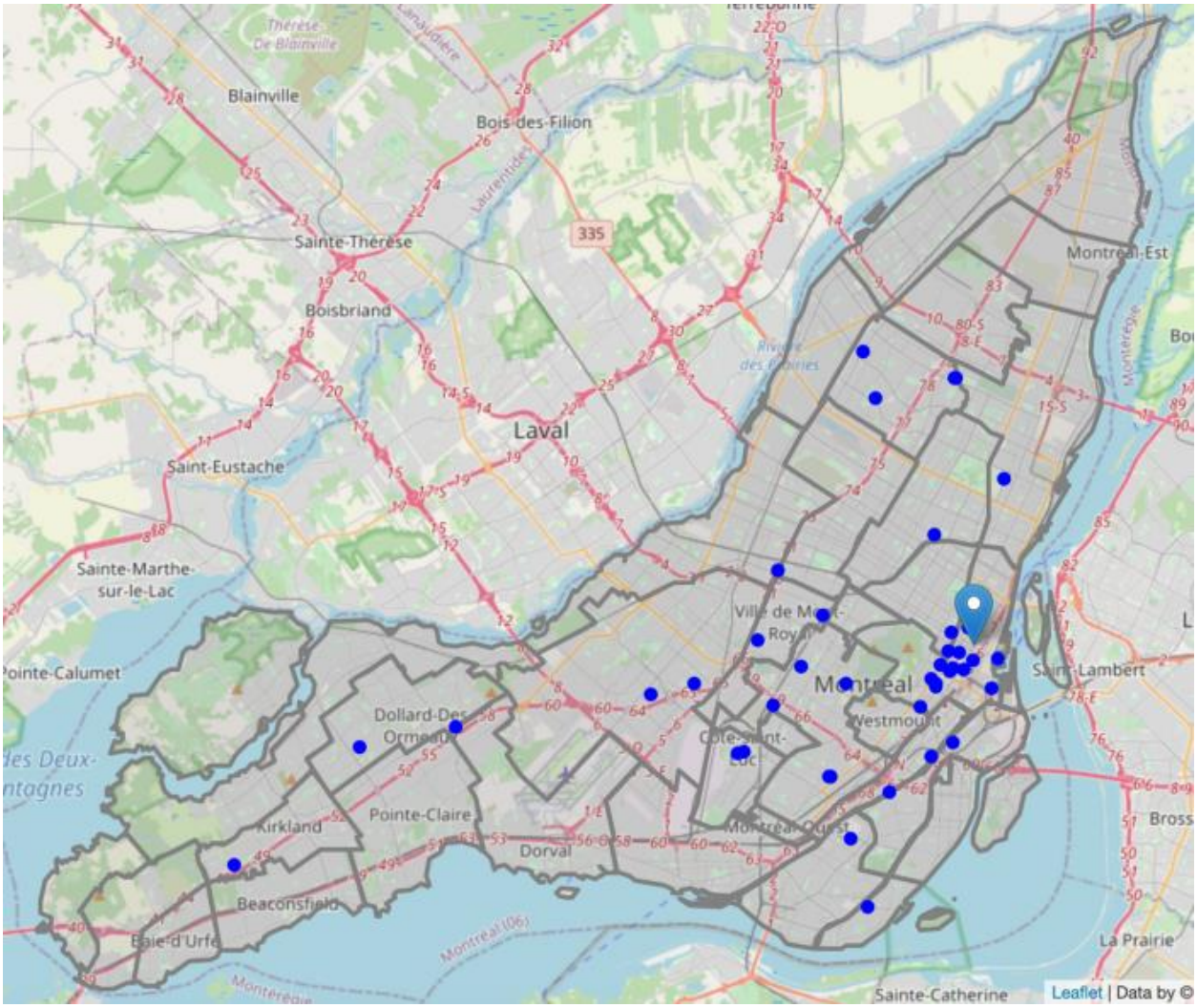

Figure 4: Movie theaters on Montreal island.

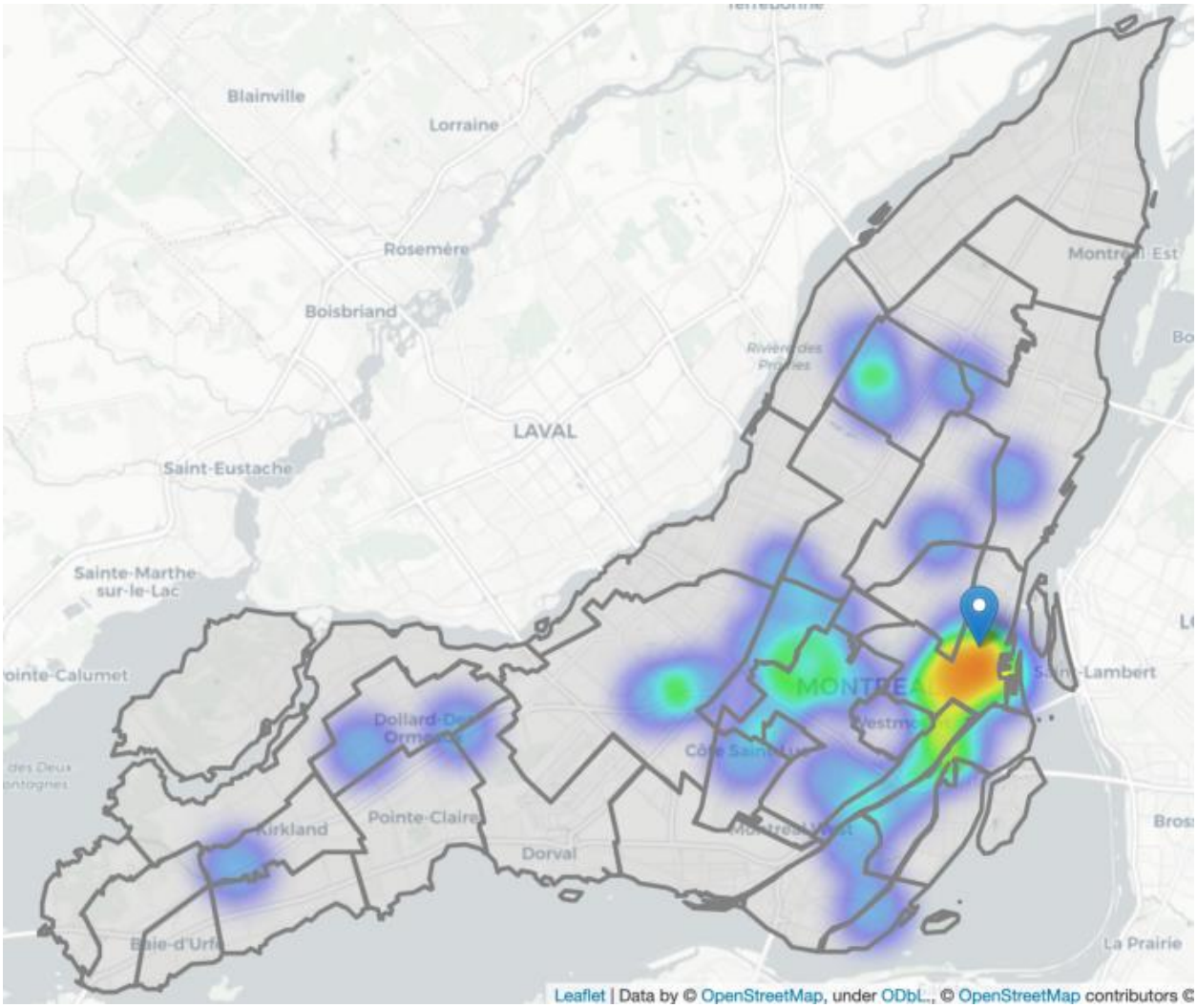

Figure 5: Heatmap of movie theater distribution on the island.

From heatmap, we can see the movie theaters are mainly concentrated in downtown areas and the center of the island. Usually, there are also a lot of shopping malls nearby, let's pull out the shopping centers data on Montreal island using Foursquare APIs.

From Foursquare API documentation, there are several categories related to shopping malls or shopping centers.

We will fetch all shopping malls data in the above categories and show them on the map with movie theaters data.

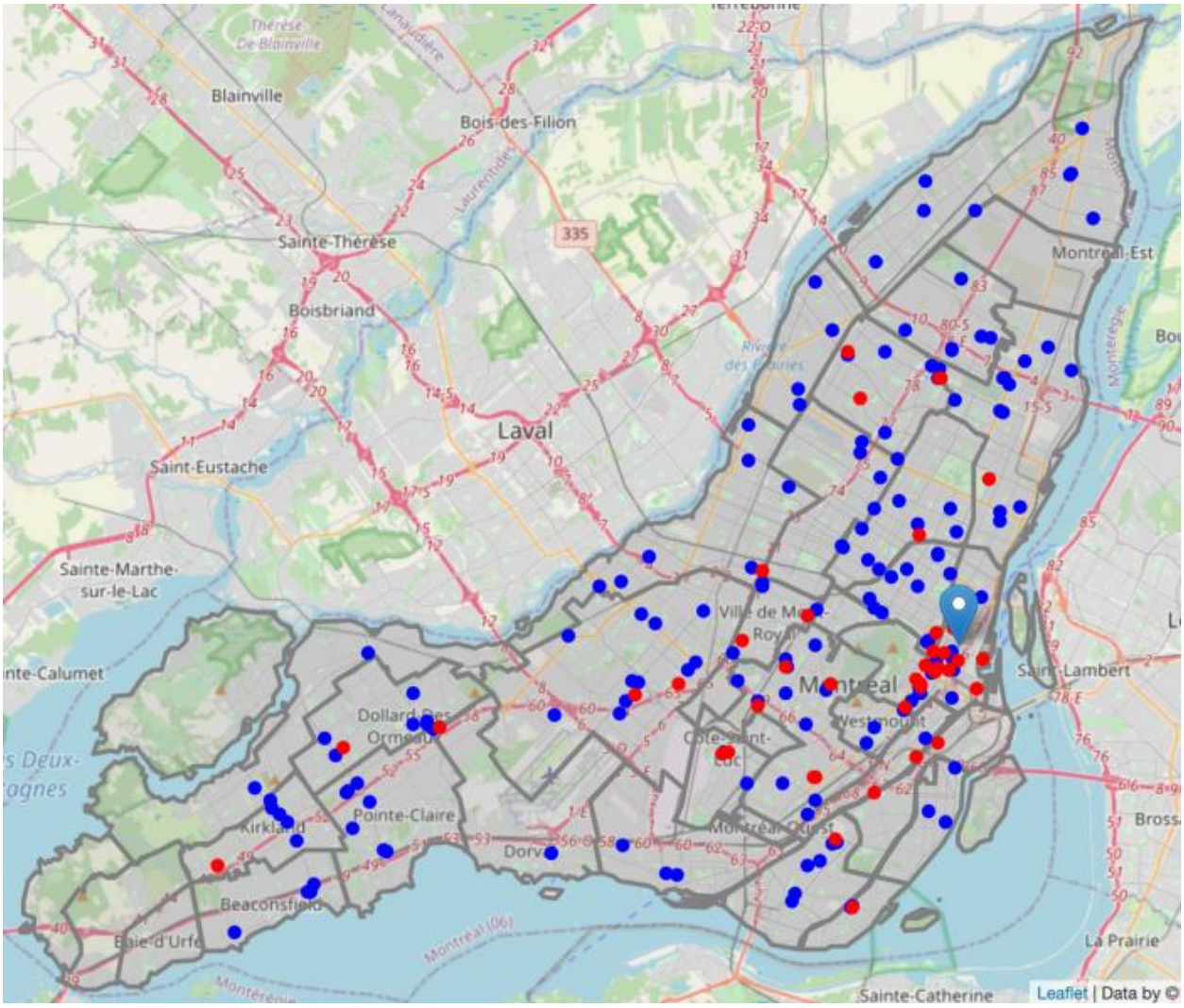

Figure 6: The shopping mall in blue, movie theater in the red.

From the map view, we can see movie theater is located near shopping malls in most scenarios. Before that, we need to cluster all the candidate hexagons based on certain information, in this project, we pull out census data as major features for clustering.

Montreal Census information

Now we will fetch census information of each borough or municipalities on Montreal island. The latest data was collected in 2016. We can get it from the [Montreal city official website].

It's a pretty big excel file containing a lot of data, I modified some sheets a bit to extract data easier into Pandas Dataframe. We only focus on several basic census information: `Population`, `Density`, `Age`, `Education` and `Income`.

Borough	Population	Area	Density	Average Age	Average Education	Average Income
Ahuntsic-Cartierville	134245	24.160	5556.498344	39.9	1.799020	29181
Anjou	42796	13.680	3128.362573	43.8	1.503555	31478
Côte-des-Neiges–Notre-Dame-de-Grâce	166520	21.440	7766.791045	36.1	2.131304	24715
Lachine	44489	17.720	2510.665914	40.2	1.546935	31374
LaSalle	76853	16.270	4723.601721	41.6	1.500000	28358
Le Plateau-Mont-Royal	104000	8.130	12792.127921	33.6	2.419321	30361
Le Sud-Ouest	78151	15.680	4984.119898	35.7	1.809057	29041
L'Île-Bizard–Sainte-Geneviève	18413	23.600	780.211864	43.1	1.796028	36583
Mercier–Hochelaga-Maisonneuve	136024	25.410	5353.168044	38.5	1.586892	29857
Montréal-Nord	84234	11.050	7622.986425	40.0	1.122727	23474
Outremont	23954	3.850	6221.818182	36.1	2.511962	44537
Pierrefonds-Roxboro	69297	27.060	2560.864745	41.0	1.737113	31235
Rivière-des-Prairies–Pointe-aux-Trembles	106743	42.280	2524.668874	42.8	1.261464	32311

Figure 7: Census Dataframe after pre-processing.

Next, we will show census data distribution on a choropleth map. *A Choropleth Map is a map composed of colored polygons. It is used to represent spatial variations of a quantity.* We also show shopping centers and movie theaters' locations on the same map.

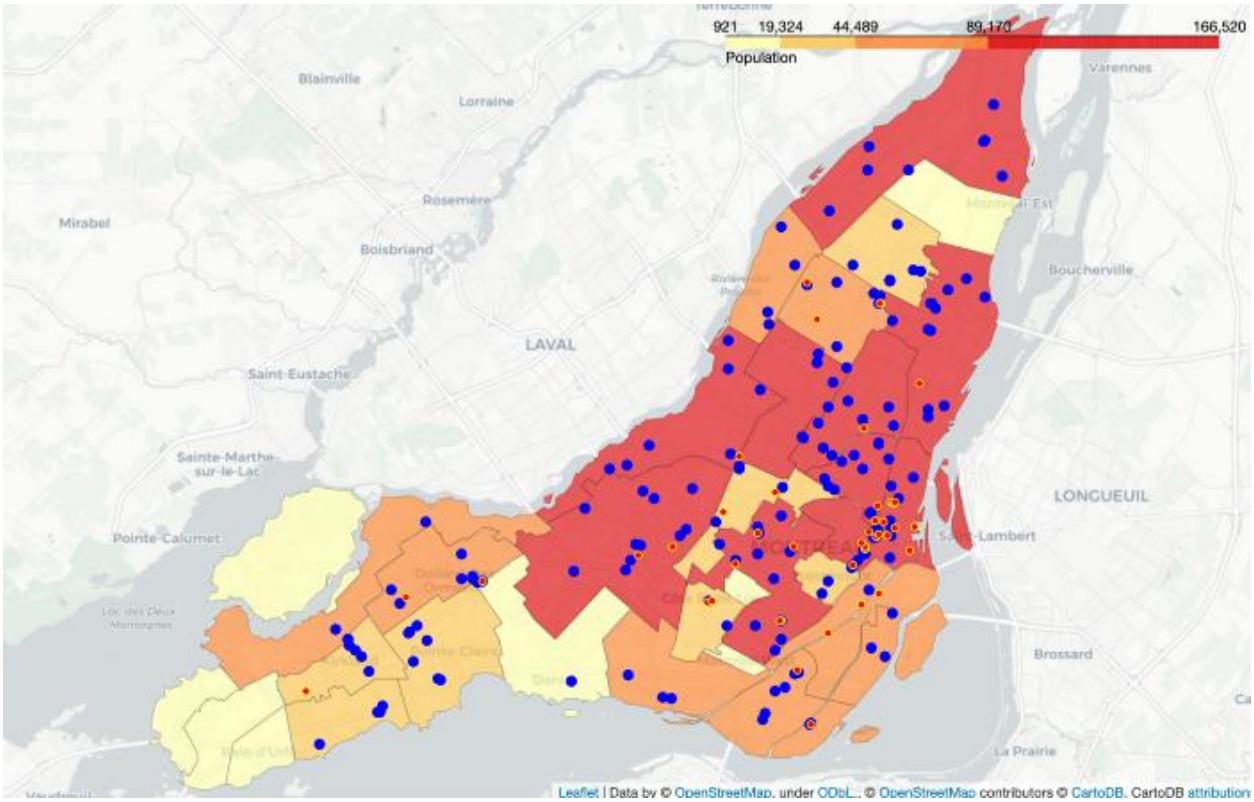

Figure 8: Population distribution by boroughs.

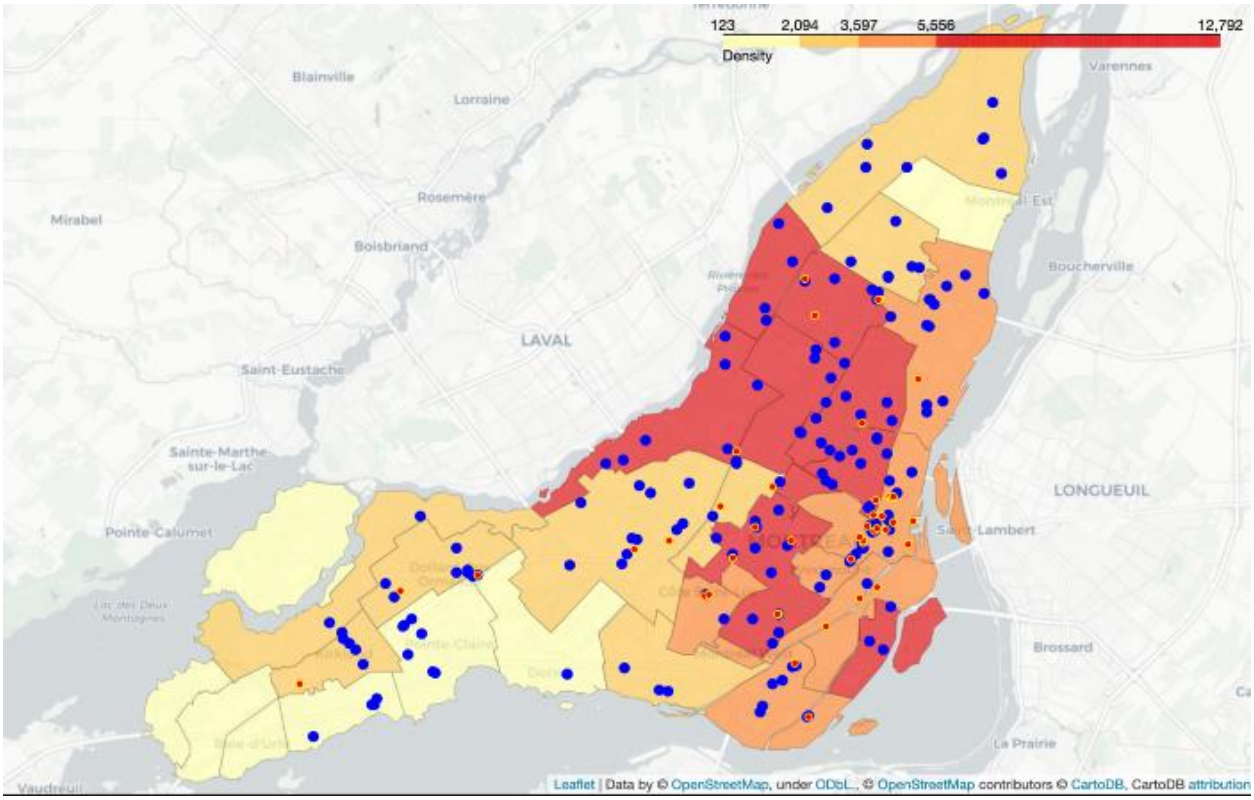

Figure 9: Density distribution by boroughs.

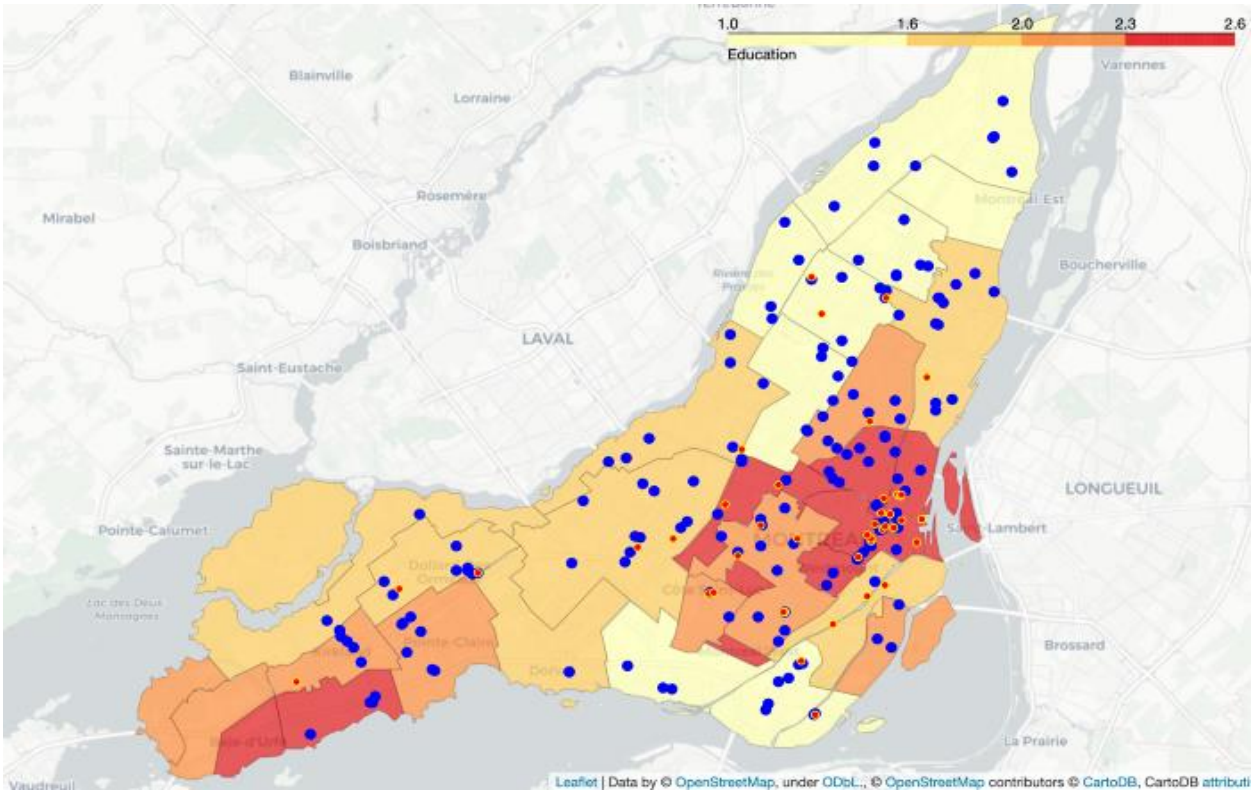

Figure 10: Education distribution by boroughs.

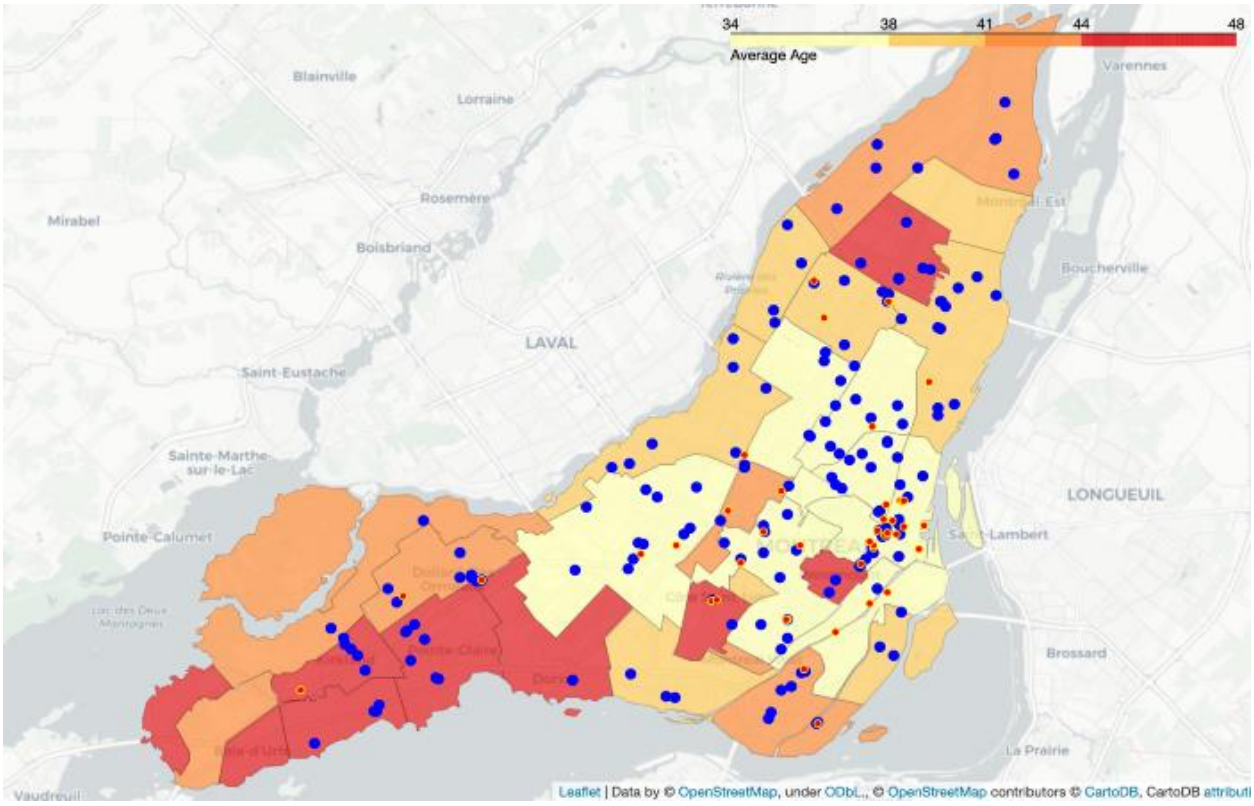

Figure 11: Age distribution by boroughs.

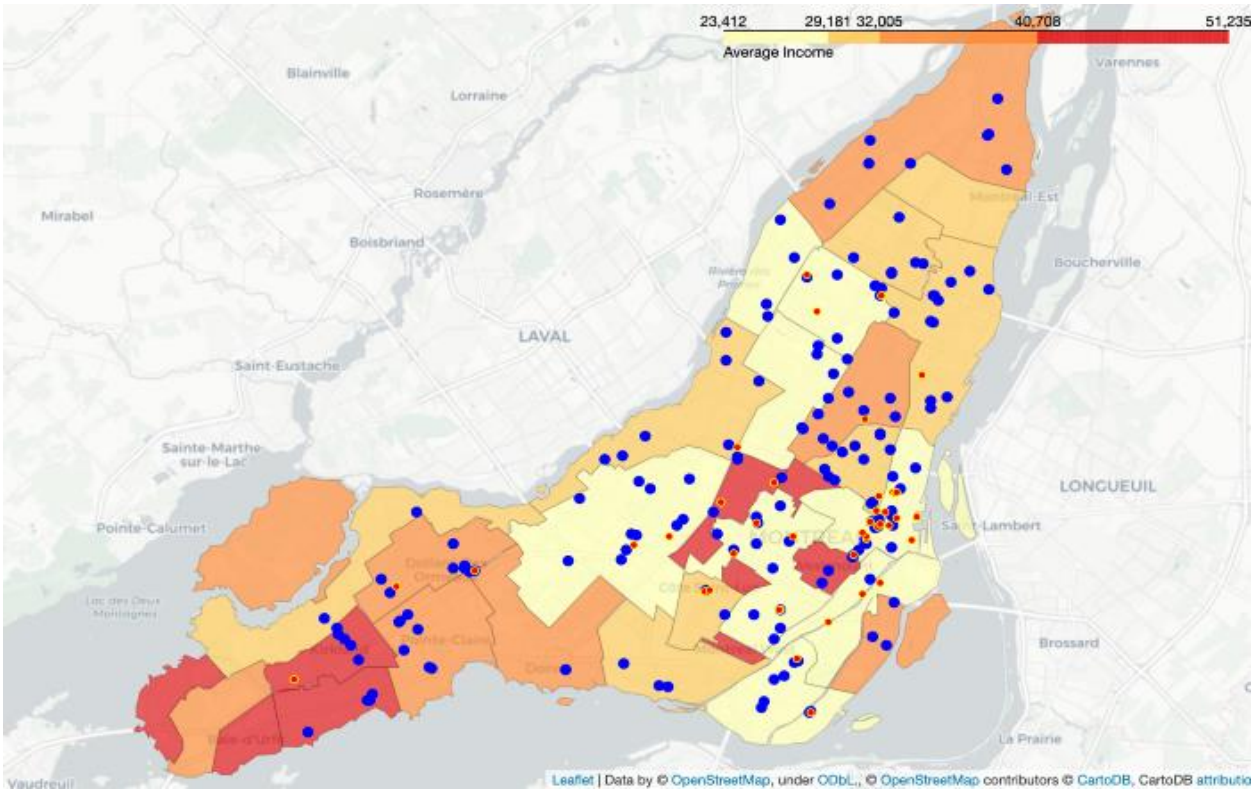

Figure 12: Income distribution by boroughs.

From the above choropleth maps, we can see movie theaters are mostly located in areas with a higher population. Same for shopping centers' locations. Moreover, most movie theaters locate in the area with lower revenue. Regions with higher revenue have fewer shopping centers and movie theaters. So far, we retrieved all the necessary raw data we needed and visualized them. In the following steps, we will manipulate these datasets, extract data, and generate new features for the machine learning algorithm. Finally, we will find out the best suitable place to open a movie theater on Montreal island.

3. Methodology

The business purpose of this project is to find a suitable place on Montreal island to open a movie theater.

Now we retrieved the following data:

1. All movie theaters data on Montreal island
2. All shopping centers data on Montreal island
3. 2016 Montreal census data for each borough, concretely, population, density, age, education and income data for each borough or municipality within Montreal island.
4. Boundary data of each borough and municipality on Montreal island.

We also generated a honeycomb hexagons grid throughout the whole Montreal Island. Based on the above raw data, we will try to generate new features accordingly, e.g. **census information for each candidate cell**, and the **number of movie theaters and shopping malls in local and nearby**.

In the final step, we will focus on the most promising areas with more shopping malls and fewer movie theaters. And we will also present the candidate hexagon cells in the map view for stakeholders to make the final decision.

4. Analysis

We got the basis census information of each borough and municipality. We want to get the census information for each candidate hexagon cell accordingly, we calculate those census information based on borough and municipality which intersects with the cell.
If a hexagon is in one borough completely, we will use the borough's census info as hexagon's one. So it means for all hexagons inside one borough, we will treat them the same for census feature.
Accordingly, if a hexagon has a 50% intersection with two boroughs respectively, we will generate the census data of this hexagon, 50% ratio from these two boroughs respectively. Based on this rule, we can calculate the census for all hexagons. Let's merge this data frame with the previous location data frame and generate a new one: **candidates_df** which contains basic information on each hexagon. We print several rows of this data frame.

	Address	Latitude	Longitude	X	Y	Distance from downtown	Population	Density	Average Age	Average Education	Average Income	Boundary
200	7751 Rue Hervé Saint-Martin, Saint-Laurent, QC...	45.477701	-73.738307	598610.849218	5.036793e+06	14238.592694	98651.955544	2307.597594	38.112567	1.882084	26936.311359	POLYGON ((-73.73819062519662 45.48289688374208...
201	Unnamed Road, Dorval, QC H9P 1A2	45.477559	-73.725515	599610.849218	5.036793e+06	13275.396852	39308.281785	1266.828924	42.348853	1.819408	33446.070382	POLYGON ((-73.7253979492433945.48275483267659...
202	Autoroute Chomedey, Saint-Laurent, QC H4T	45.477416	-73.712724	600610.849218	5.036793e+06	12318.068082	72329.460624	1845.953980	39.991615	1.854283	29823.781715	POLYGON ((-73.71260536925523 45.48261134929982...
203	7700 Autoroute Côte-de-Liesse, Saint-Laurent, ...	45.477271	-73.699932	601610.849218	5.036793e+06	11368.088712	98828.000000	2310.685060	38.100000	1.882270	26917.000000	POLYGON ((-73.69981288619405 45.48246643364013...
204	Canadian National Montreal, Taschereau Yard, S...	45.477125	-73.687141	602610.849218	5.036793e+06	10427.467604	98828.000000	2310.685060	38.100000	1.882270	26917.000000	POLYGON ((-73.68702050102164 45.48232008572624...
205	6762 Chemin Wallenberg, Côte Saint-Luc, QC H4W...	45.476977	-73.674350	603610.849218	5.036793e+06	9498.985226	38665.548038	4454.331230	43.810090	2.049003	30661.150161	POLYGON ((-73.67422821469975 45.48217230558711...
206	5757 Boul Cavendish, Côte Saint-Luc, QC H4W 2W8	45.476828	-73.661559	604610.849218	5.036793e+06	8586.580227	39886.419244	4845.859689	43.906458	2.072130	30774.458000	POLYGON ((-73.66143602819015 45.48202309325193...

Figure 13: Census info of hexagons.

Looking good. Now we have census information in each hexagon area. Then we will calculate the shopping center and movie theaters related information for each hexagon area. We will calculate the following features for shopping malls and movie theaters:

1. The number of shopping malls and movie theaters within the current hexagon cell.
2. The number of shopping malls and movie theaters within 1 km away from the center of the hexagon cell.
3. A number of shopping malls and movie theaters within 3 km away from the center of the hexagon cell.

Now we prepared all the data we need, we can use the **K-Means clustering algorithm** to group the similar candidate hexagon areas into clusters.

K-Means Clustering

We pick up census features and the number of shopping malls and the number of movie theaters as input features.

	Population	Density	Average Age	Average Education	Average Income	Cinemas in cell	Cinemas in 1km	Cinemas in 3km	Malls in cell	Malls in 1km	Malls in 3km
200	98651.955544	2307.597594	38.112567	1.882084	26936.311359	0	0	0	0	0	3
201	39308.281785	1266.828924	42.348853	1.819408	33446.070382	0	0	1	0	0	5
202	72329.460624	1845.953980	39.991615	1.854283	29823.781715	0	0	1	0	0	5
203	98828.000000	2310.685060	38.100000	1.882270	26917.000000	0	0	4	0	0	5
204	98828.000000	2310.685060	38.100000	1.882270	26917.000000	0	0	4	0	0	7
205	38665.548038	4454.331230	43.810090	2.049003	30661.150161	0	2	3	0	1	5
206	39886.419244	4845.859689	43.906458	2.072130	30774.458000	2	0	4	1	0	6
207	21145.107694	4307.018336	42.003744	2.276651	38871.266579	0	0	7	0	0	10
208	99044.319762	6133.245911	38.045434	2.254198	33156.914002	0	0	8	0	0	13
209	166520.000000	7766.791045	36.100000	2.131304	24715.000000	0	3	3	0	1	10

Figure 14: Selected features as input parameters for the K-Means Clustering Algorithm

We will run an evaluation step first to select the best **K** which is the number of categories in the algorithm.

We use the **Sum of Squared Distance** and **Silhouette Score** two methods to evaluate the K-Means algorithm for different **K**.

Sum of Squared Distance measures error between data points and their assigned clusters' centroids. Smaller means better.

Silhouette Score focuses on minimizing the sum of squared distance inside the cluster as well, meanwhile, it also tries to maximize the distance between its neighborhoods. From its definition, the bigger the value is, the better K is.

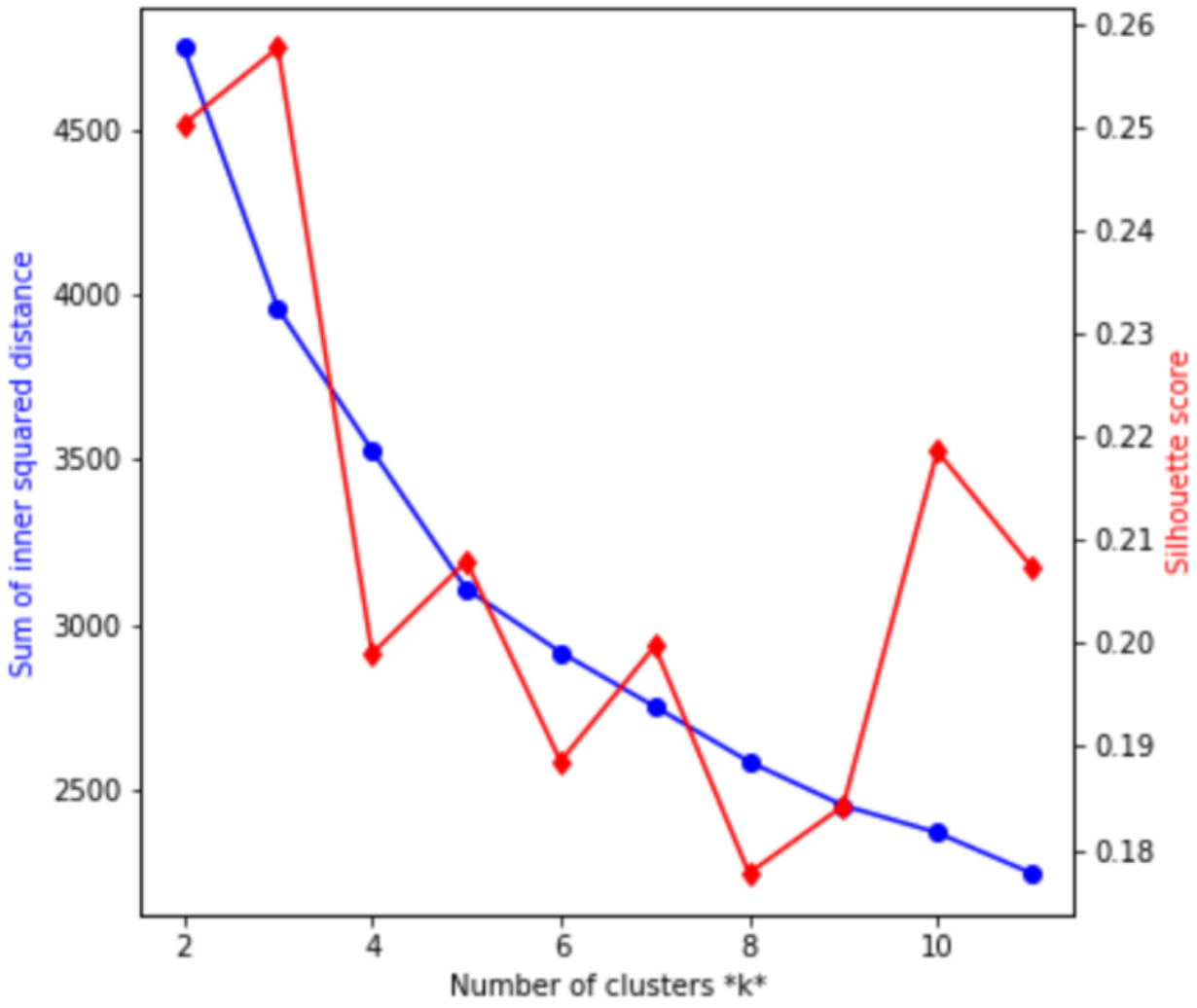

Figure 15: K selection for K-Means Clustering

From the figure, we can see Sum of Squared Distance going down when K becomes bigger. When K=2,3, Silhouette Score is higher, but SSE is still high at that time, we

choose K=10 for this project, it's a balanced number for both Sum of Squared Distance and Silhouette Score. Let's run the K-Means algorithm again with **k=10**.
Let's visualize clustering results with a different color in the map view.

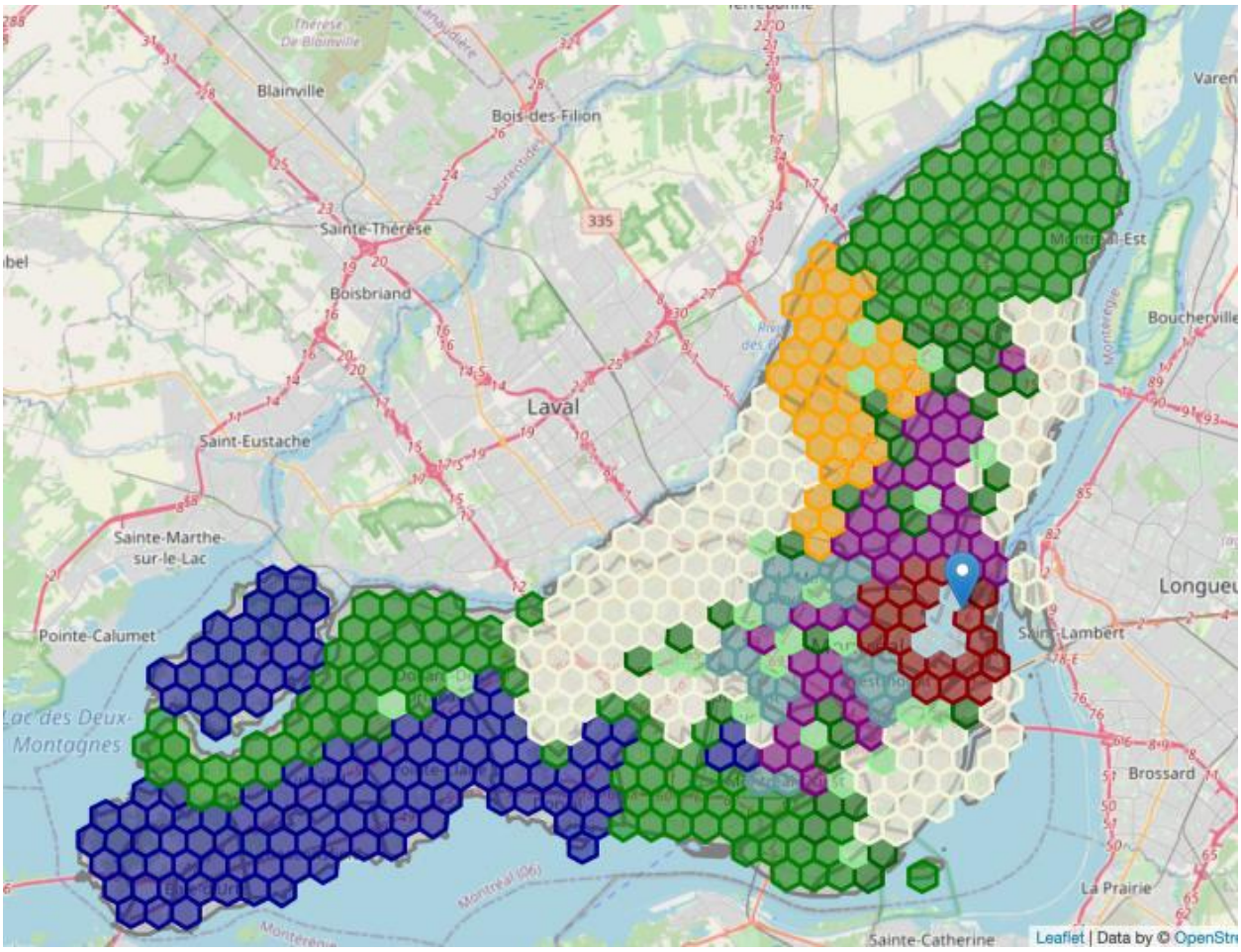

Figure 16: 10 clusters of candidate hexagons

Let's put everything together on one map view:
1. Clusters in colors for hexagons
2. Shopping malls in blue point
3. Movie theaters in redpoint with yellow ring.

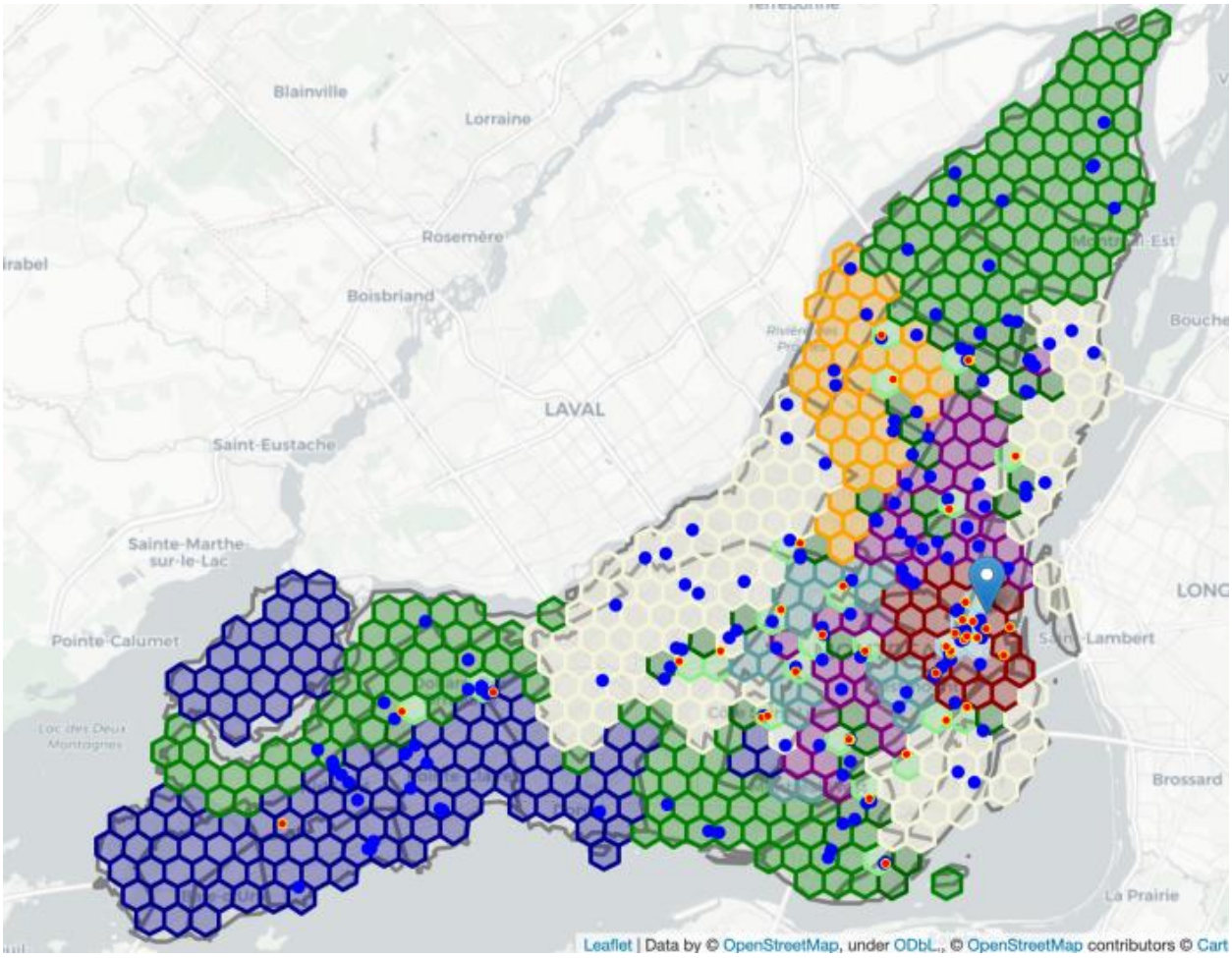

Figure 17: Clusters with Shopping and Movie theaters

From the cluster plot in the above map view, we can see there is one cluster in light blue composed of 4 hexagons in downtown, there are full of movie theaters and shopping malls in this cluster.

The purple cluster contains the area with a lot of shopping malls. The light green cluster contains more shopping malls and movie theaters except for the downtown cluster.

Let's assign weights to all three movie theaters related features and combine them into one feature. Same for shopping malls. It's easier for sorting.

We will calculate weighted Mall Score and weighted Cinema Score, then generate a new Score feature for sorting.

The higher final score is, it means there are more shopping malls and fewer movie theaters.

Cluster 7 have the highest score, it has more shopping malls and fewer movie theaters. Let's explore more characteristics of cluster 7.

	Distance from downtown	Population	Density	Average Age	Average Education	Average Income	Cinemas in cell	Cinemas in 1km	Cinemas in 3km	Malls in cell	Malls in 1km	Malls in 3km	Cluster	Mall Score	Cinema Score	Score
count	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.0	40.000000	40.000000	40.00000	40.000000	40.000000	40.0	40.000000	40.000000	40.000000
mean	5419.586758	132613.112910	8251.628033	36.502441	2.074046	29123.672362	0.0	0.100000	5.050000	0.77500	1.075000	14.500000	7.0	21.600000	5.350000	16.250000
std	1854.333169	22909.968441	1789.093565	1.431799	0.201771	2911.562095	0.0	0.303822	2.767539	0.80024	1.071484	3.954874	0.0	7.389528	2.931264	8.113774
min	1952.670347	89170.000000	5353.168044	33.600000	1.586892	24715.000000	0.0	0.000000	1.000000	0.00000	0.000000	7.000000	7.0	13.000000	1.000000	3.000000
25%	4085.078312	113829.451571	7440.062372	36.050423	1.952450	25985.847854	0.0	0.000000	2.000000	0.00000	0.000000	11.750000	7.0	16.000000	2.000000	11.000000
50%	5264.889535	139052.644630	7819.535410	36.531560	2.131147	30124.081923	0.0	0.000000	5.000000	1.00000	1.000000	14.500000	7.0	19.000000	6.000000	13.000000
75%	6763.023333	140789.752715	8806.940063	37.400000	2.209832	31660.322973	0.0	0.000000	7.000000	1.00000	2.000000	17.000000	7.0	27.000000	8.000000	25.250000
max	9458.115449	166520.000000	12792.127921	39.434653	2.426965	34222.429660	0.0	1.000000	10.000000	3.00000	4.000000	24.000000	7.0	39.000000	10.000000	31.000000

Figure 18: Statistics of cluster 7

There are **40 hexagons** in Cluster 7 with an average of 0.77 *Malls in local* and 0.0 *Cinemas in local*. Let's plot all clusters for comparison of each feature in a bar chart using **matplotlib.pyplot** library. We highlight Cluster 7 which is our target cluster.

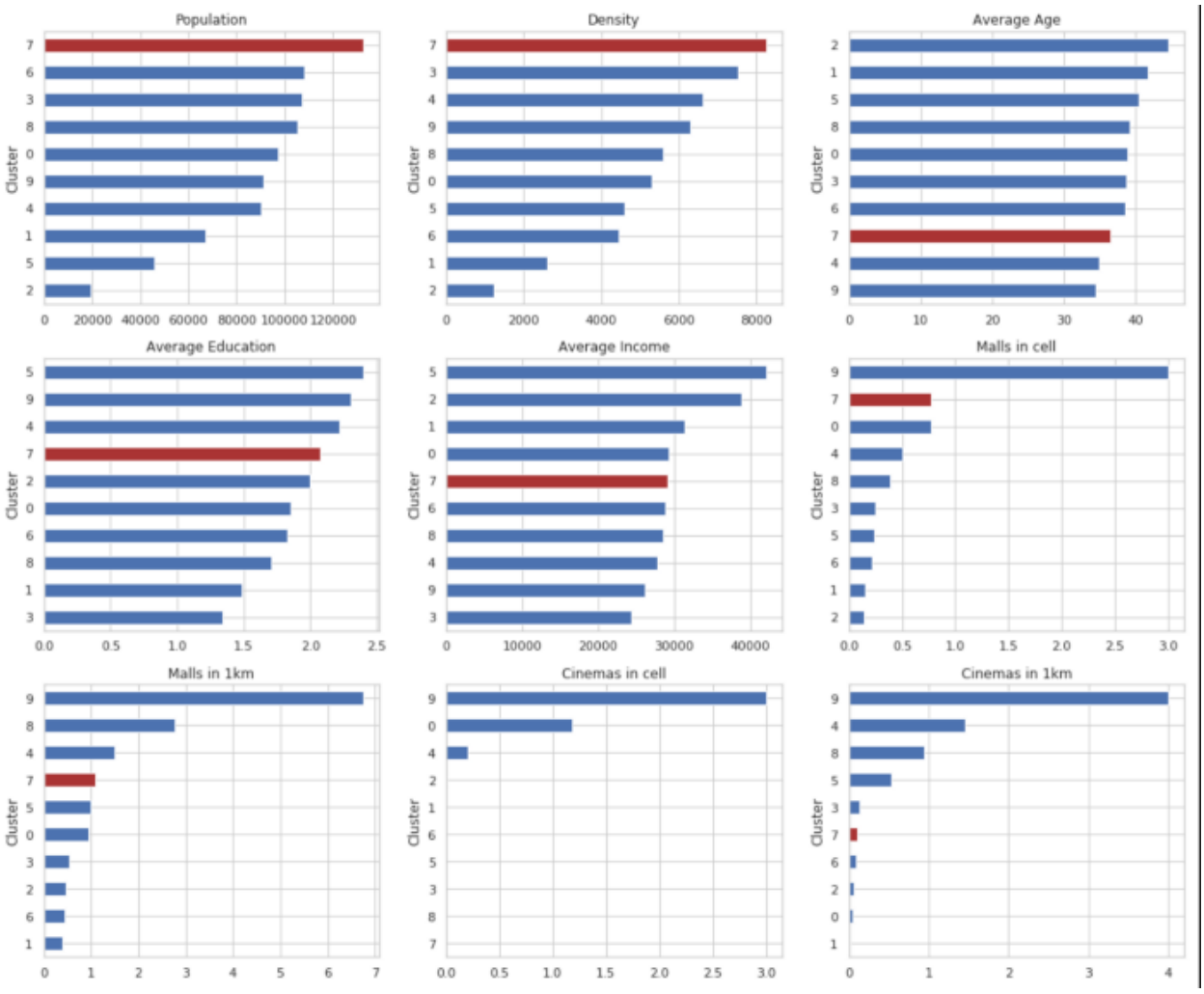

Figure 19: Feature comparison of clusters.

From the bar chart, we can see that **Cluster 7** has the most population and density among all the clusters. Furthermore, it has fairly more shopping centers in the hexagon area or nearby and relatively fewer movie theaters.
Next, we sort all hexagons in Cluster 7 by Score in descending order and pick the first 5 hexagons. They will be our first choice position to open a movie theater.

	Population	Density	Average Age	Average Education	Average Income	Malls in cell	Malls in 1km	Malls in 3km	Cinemas in cell	Cinemas in 1km	Cinemas in 3km	Mall Score	Cinema Score	Score
377	104000.000000	12792.127921	33.600000	2.419321	30361.000000	1	4	21	0	0	7	38	7	31
388	139590.000000	8806.940063	37.400000	1.955852	32008.000000	1	3	17	0	0	1	31	1	30
360	97394.898948	12249.970165	33.806291	2.426965	31530.751299	3	1	21	0	0	10	39	10	29
376	107878.011953	12357.887793	34.014061	2.368820	30540.462930	2	2	19	0	0	6	35	6	29
387	139857.436623	8801.714957	37.312172	1.928706	31468.735348	2	2	14	0	0	2	30	2	28

As the above statistics information, there are 1~3 shopping malls in local and more shopping malls nearby, but without any movie theater within 1 km. Looks quite good selections.
Let's plot Cluster 7 hexagons in the map view, gray out the other clusters and highlight our 5 choices as well.

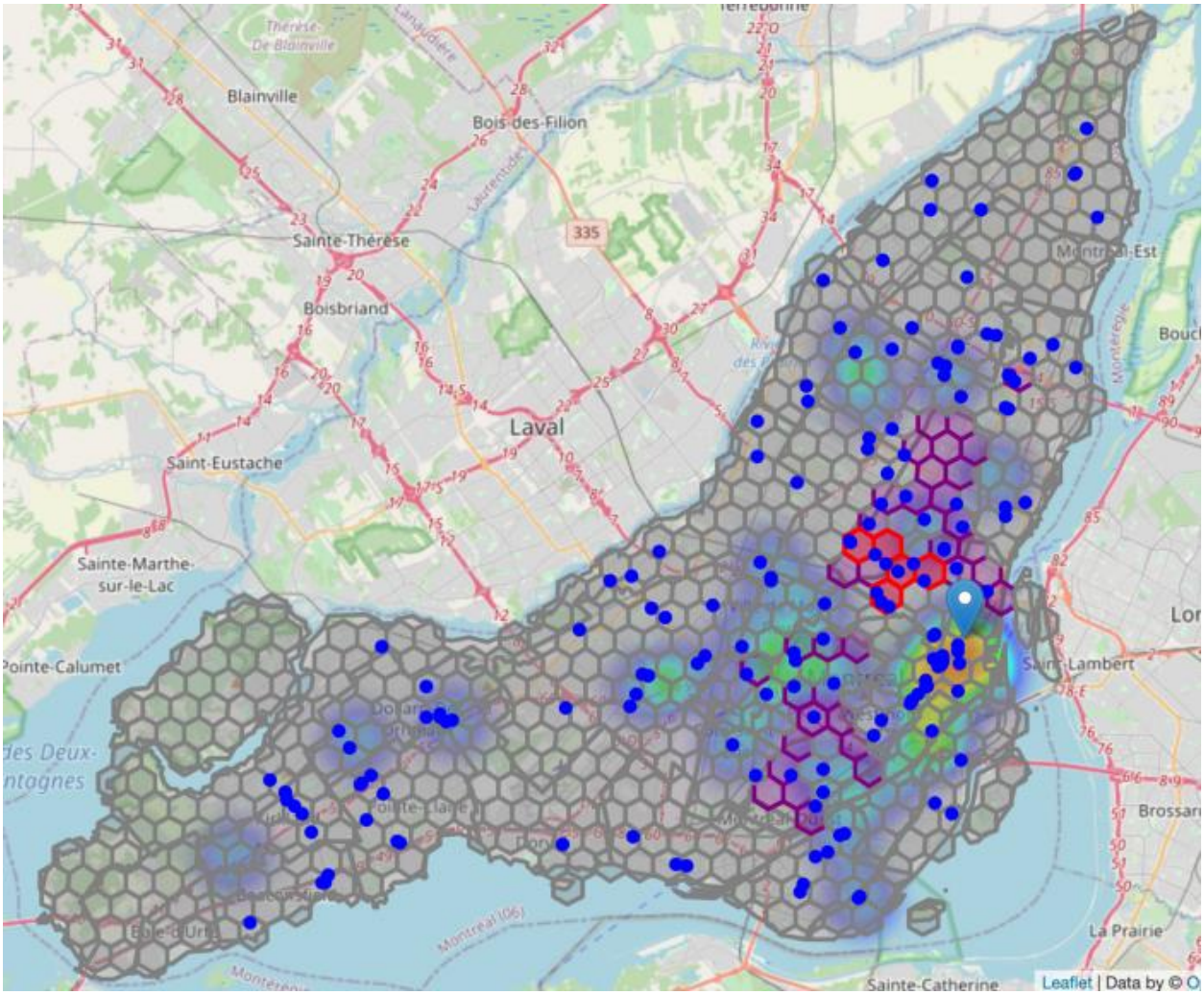

Figure 20: Five most promising candidate areas.

This concludes our analysis. We have found out 5 most promising zones with more shopping malls nearby and fewer movie theaters around the area. Each zone is in regular hexagon shape which is popular in map view. The zones in the cluster have the most population and density comparing with other clusters.

5. Result and Discussion.

We generated hexagon areas all over Montreal island. And we group them into 10 clusters according to census data information including population, density, age, education, and income. Shopping center information and existing movie theaters information are also considered when running the clustering algorithm.

From data analysis and visualization, we can see movie theaters are always located near shopping malls usually, which inspired us to find out the area with more shopping malls and fewer movie theaters.

After the K-Means Clustering machine learning algorithm, we got the cluster with most shopping malls nearby and fewer movie theaters on average. We also discovered the other characteristics of the cluster. It shows the cluster has the most population and density which implies the highest traffic among all the clusters.

There are 40 hexagon areas in this cluster, we sort all these hexagon areas by shopping malls and movie theaters info in descending order which targets to cover more shopping malls and fewer movie theaters in the local cell or nearby.

We draw our conclusion with the 5 most promising hexagon areas satisfying all our conditions. These recommended zones shall be a good starting point for further analysis. There are also other factors which could be taken into account, e.g. real traffic data and the revenue of every movie theater, parking lots nearby. They will be helpful to find more accurate results.

6. Conclusion

The purpose of this project is to find an area on Montreal island to open a movie theater. After fetching data from several data sources and process them into a clean data frame, applying the K-Means clustering algorithm, we picked the cluster with more shopping malls and fewer movie theaters on average. By sorting all candidate areas in the cluster, we get the most 5 promising zones which are used as starting points for final exploration by stakeholders.

The final decision on optimal movie theater's location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into consideration additional factors like the parking lot of each location, traffic of existing movie theaters in the cluster, and current revenue of them, etc.