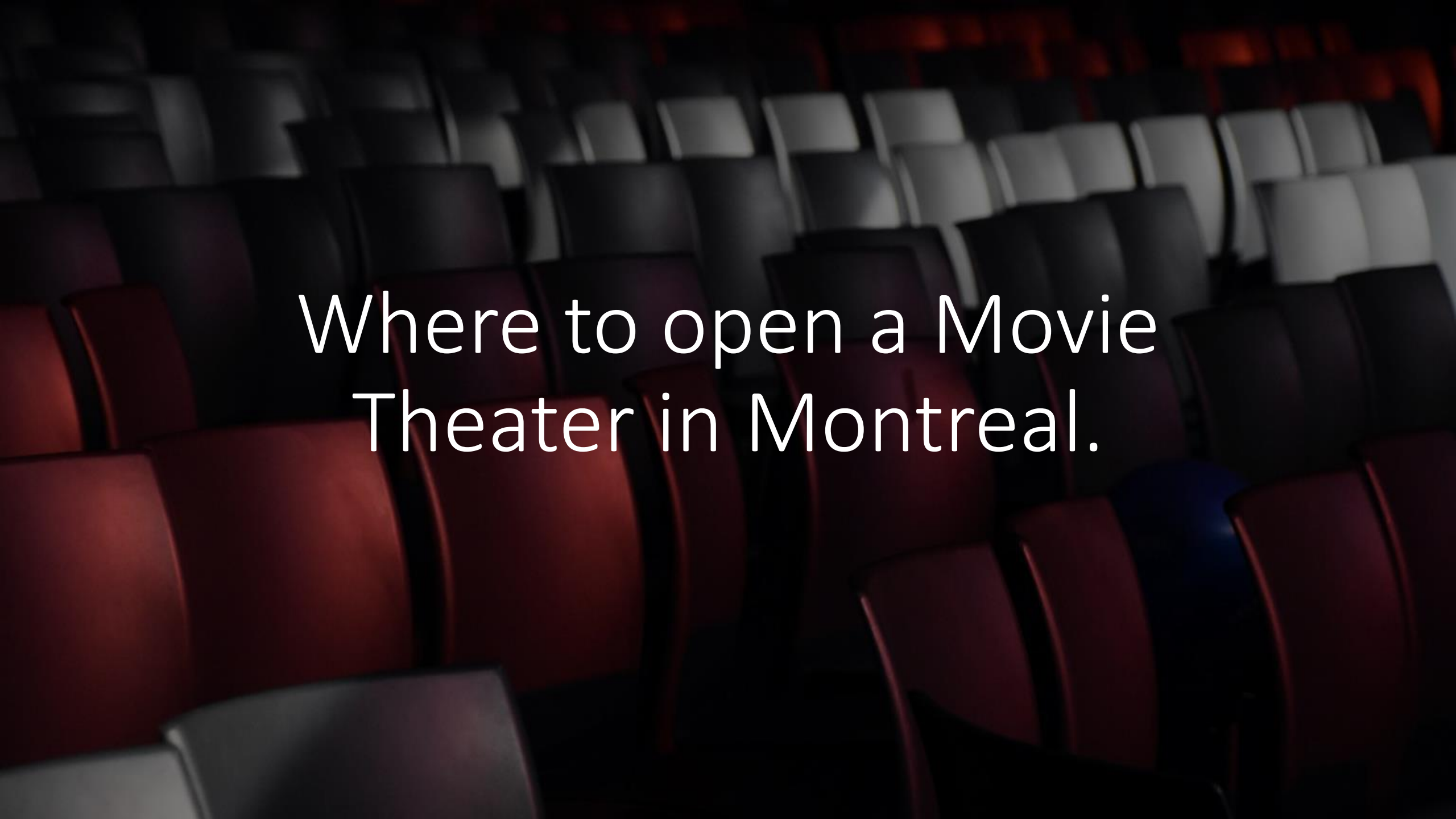

Where to open a Movie
Theater in Montreal.

Introduction

- 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.

- 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**.

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.
- 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.

Methodology

- The business purpose of this project is to find a suitable place on Montreal island to open a movie theater.
- The following datasets were retrieved.
  - All movie theaters data on Montreal island
  - All shopping centers data on Montreal island
  - 2016 Montreal census data for each borough, concretely, population, density, age, education and income data for each borough or municipality within Montreal island.
  - Boundary data of each borough and municipality on Montreal island.
- We also generated a honeycomb hexagons grid throughout the whole Montreal Island.
- 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.

Analysis

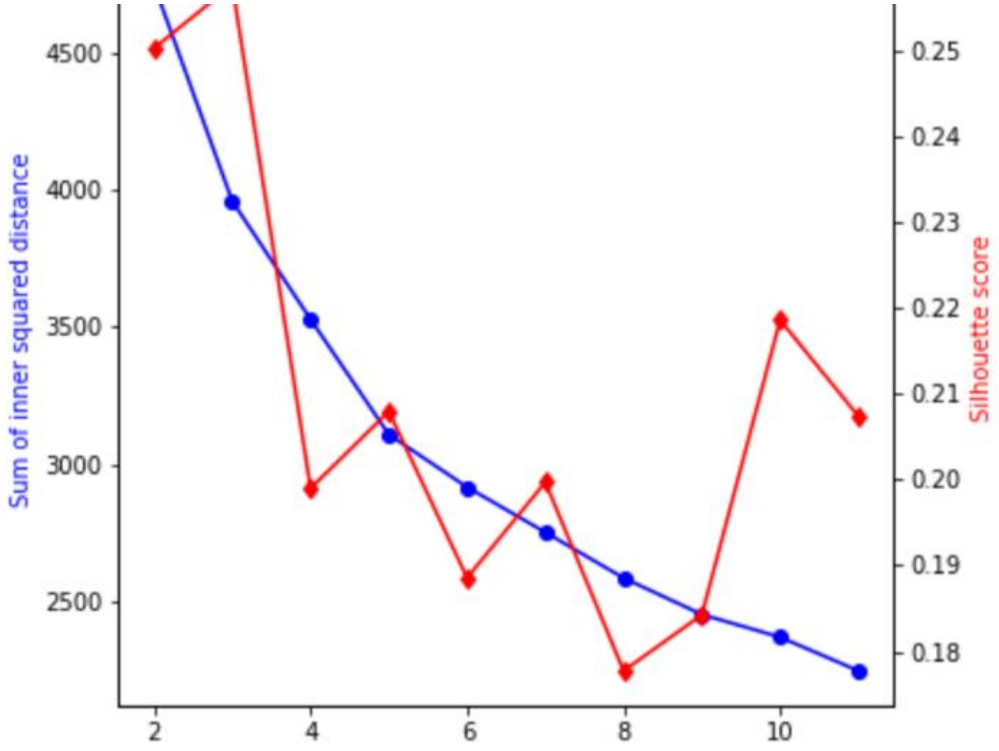
- We got the basis census information of each borough and municipality.
- We want to get the census information for each candidate hexagon cell accordingly, we calculate those census information based on borough and municipality which intersects with the cell.
- If a hexagon is in one borough completely, 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.

K-Means Clustering

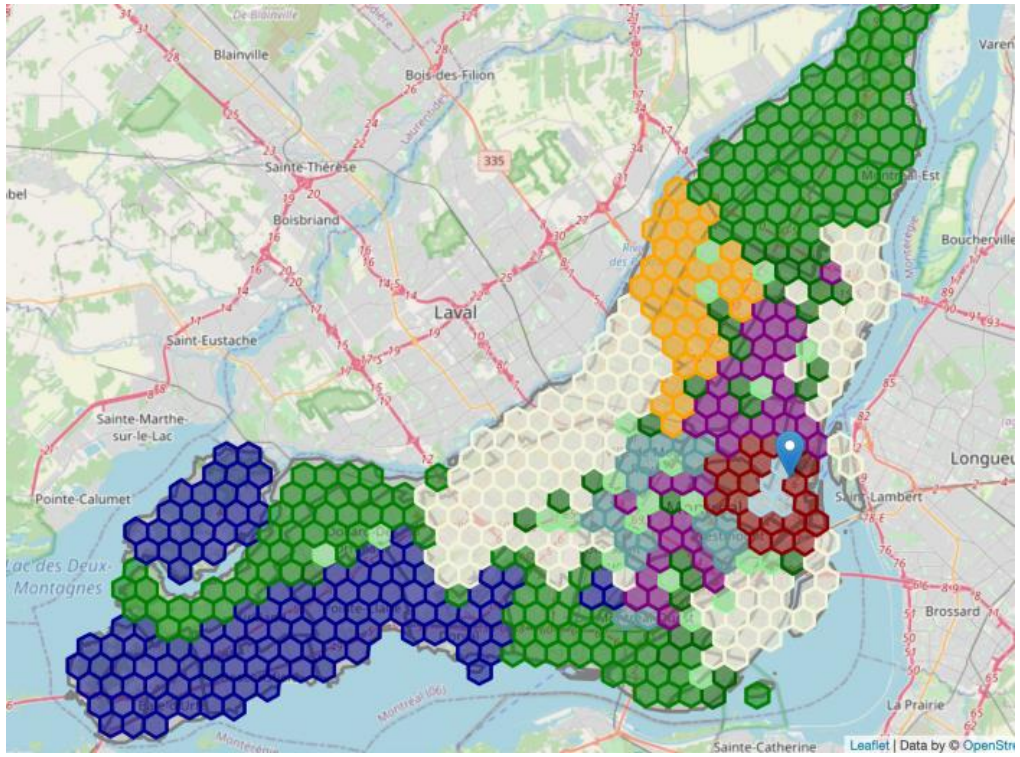
- We pick up census features and the number of shopping malls and the number of movie theaters as input features.
- We will run an evaluation step first to select the best **K** which is the number of categories in the algorithm.
- We use the **Sum of Squared Distance** and **Silhouette Score** two methods 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.

K selection for K-Means Clustering

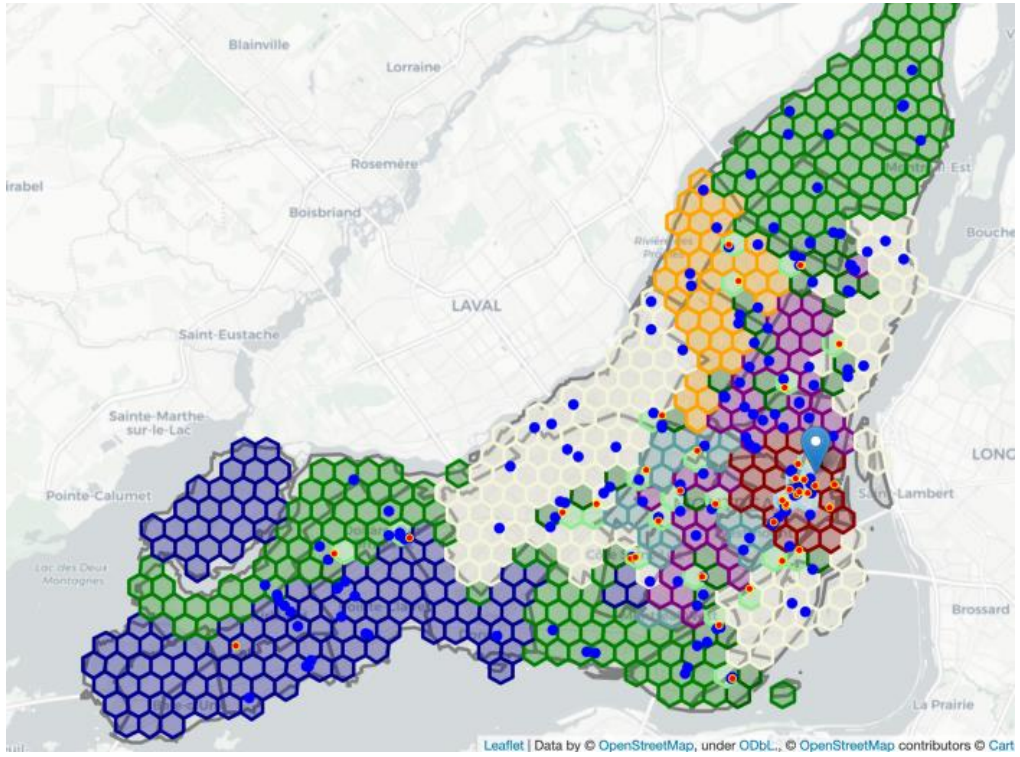
- From the figure, we can see Sum of Squared Distance going down when K becomes bigger.
- When K=2,3, Silhouette Score is higher, but SSE is still high at that time, we choose K=10 for this project, it's a balanced number for both Sum of Squared Distance and Silhouette Score.



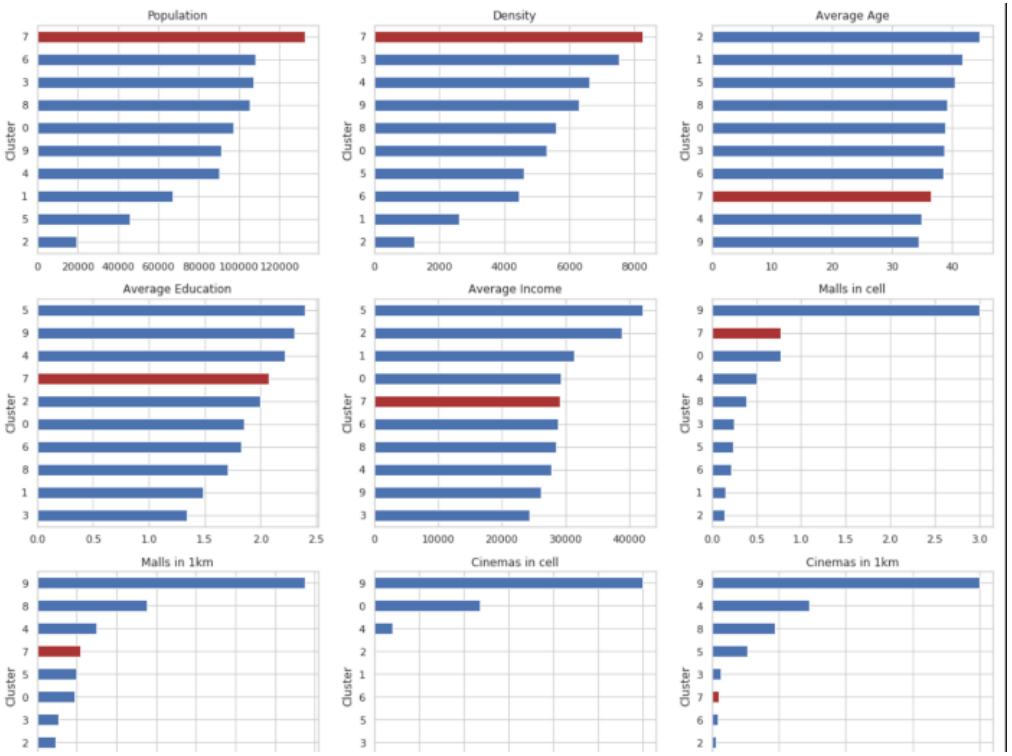
- with **k=10**.
- Let’s visualize clustering results with a different color in the map view.



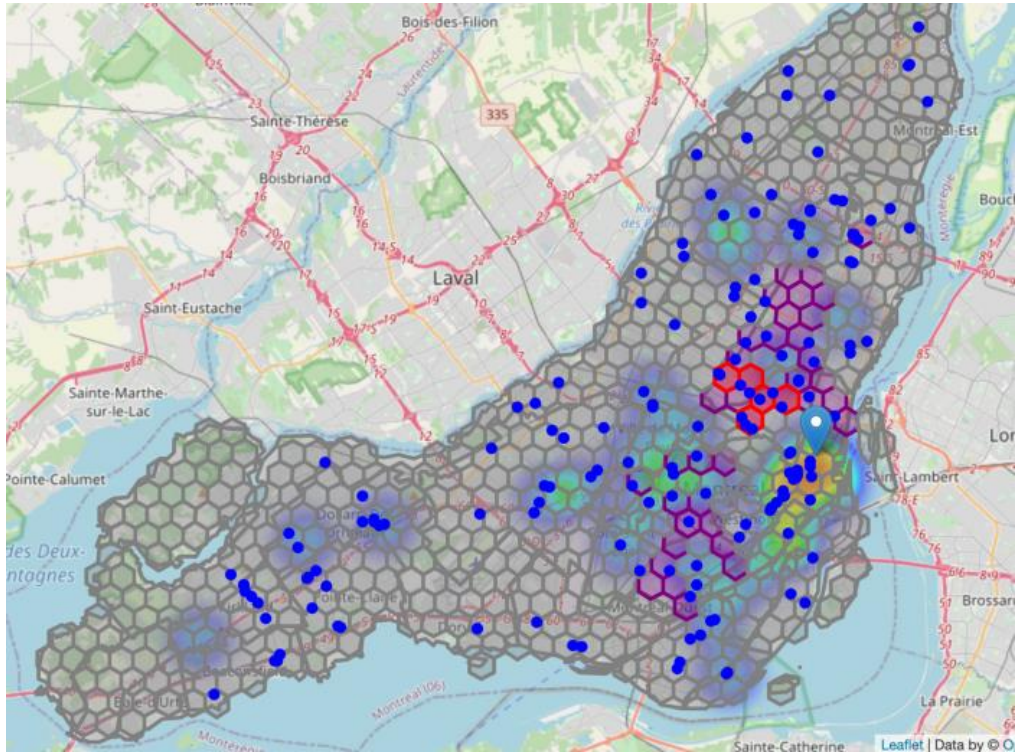
- Let’s put everything together on one map view:
- 1. Clusters in colors for hexagons
- 2. Shopping malls in blue point
- 3. Movie theaters in redpoint with yellow ring.



- 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.
- 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.



- 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.



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.

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.