Battle of neighborhoods

Live in Montreal

Capstone Project — IBM Data Science

1. Introduction

Montreal, the 2nd most populated city in Canada has more than 4 millions population with a slow growth of 0.73% average every year. Montreal is also known as the 2nd largest economy in Canada by having a variety of businesses implementing themselves out there. As the nest of opportunities, many big tech companies started considering the city to have new offices - Google, Facebook, Microsoft to name a few.

In the case where an individual had to relocate for a job opportunity, what would be the best locations we could suggest to him? The purpose of this report is to identify what would be the best options through data driven research. We will identify amenities and venues based on their ratings from which we will offer options based on the relocator preferences.

This project targets mostly individuals that are not familiar with the city and that are searching for a convenient borough where they can live. It will also bring options that fit the individual interest. As an example, an individual in the need of relocating, who likes parks, we expect him willing to be close to that type of venue.

2. Data

Montreal city boroughs names with their coordinates (latitude and longitude)

- Data pulled from Wikipedia with the BeautifulSoup library.
- Will be used with Foursquare API data to define the best venues of each borough.
- We will use Folium to visualize the different boroughs within Montreal.

Common venues with their type(e.g : Restaurants, Bars, Malls, Parks, etc.) and their location (latitude and longitude).

- For each 32 neighborhoods of Montreal.
- Clustering process with K-Mean algorithm to define more precisely where good venues and amenities are.
- Data will be visualized on a folium generated map.
- Will use the panda library to analyze and organize the data.

Criminality data of Montreal (the last 6 months of 2020)

- Pulled from the Montreal City website.
- Includes types of criminal events(For more insights).
- Data will be visualized on a folium generated map.
- Correlation matrix will be analysed against other datasets available.
- Will use the panda library to analyze and organize the data.

Demographic census of Montreal

- Data covering population, density and real estate prices.
- While the census available is stale(data from 2016), we can correlate and put in perspective the results.
- Data will be visualized on a folium generated map.
- Correlation matrix will be analysed against datasets available.

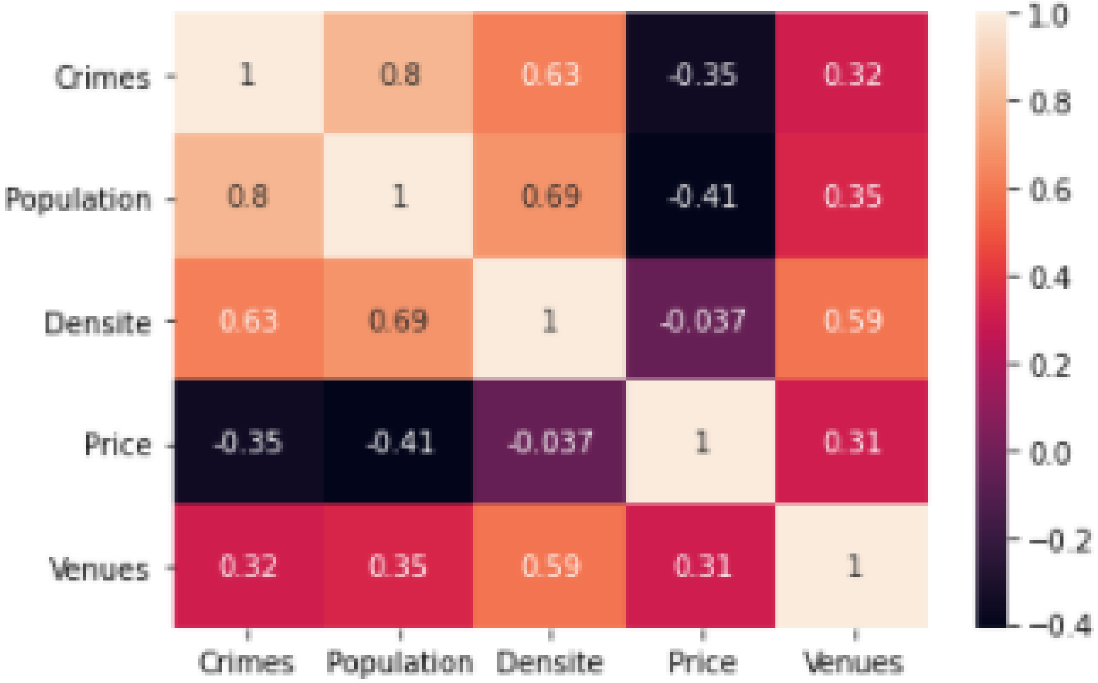
3. Methodology

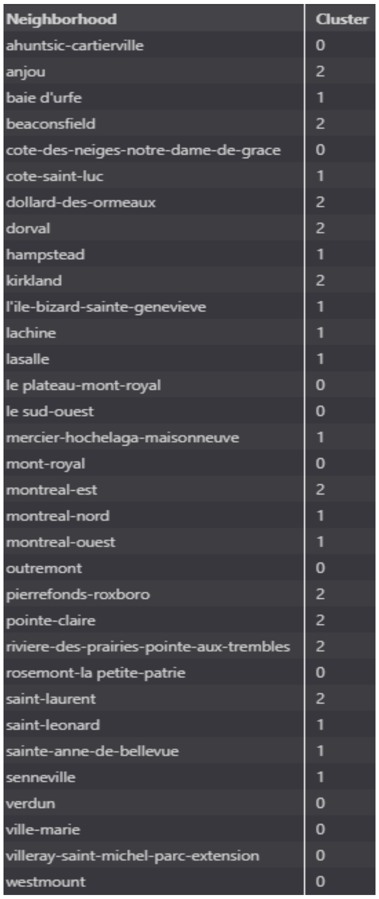
Several platforms and techniques will be used during this report.

- Python as the interpreter language. As R, extensively used in the Data analytic field. Useful for the diversity of libraries.
- Geocoders to convert addresses into coordinates.
    - A reverse will also be used to convert coordinates into addresses, used to pull addresses with criminal logs coordinates.
- Pandas for dataframe manipulation.
    - Core of our data engineering. Most datasets require filtering and cleaning.
- Seaborn for heatmap(correlation matrix) and boxploting
    - Used for properties correlation analysis and cluster vs real estate price, population and criminality dataset analysis.
- Folium for map visualization with our point of interest(Neighborhoods and venues).
    - Criminal, population/density, real estate average prices and venues datasets have been displayed in the map representing Montreal, with a coloration segmenting each neighborhood and various counts.
- Foursquare offers an API giving access to a wide range of data related to locations.
    - Data retrieved has been used in various data frames to define, for example, the most common type of venues within a neighborhood.
- As a clustering algorithm, K-Mean will be used to define ideal locations. Sklearn library helps for that purpose.
    - The clustering process has been used against our top common venues dataset to regroup similar clusters

4. Result

We've gone through various datasets, which provided insights of Montreal as a city. It has been noticed that the proximity to venues correlates the population and their density.



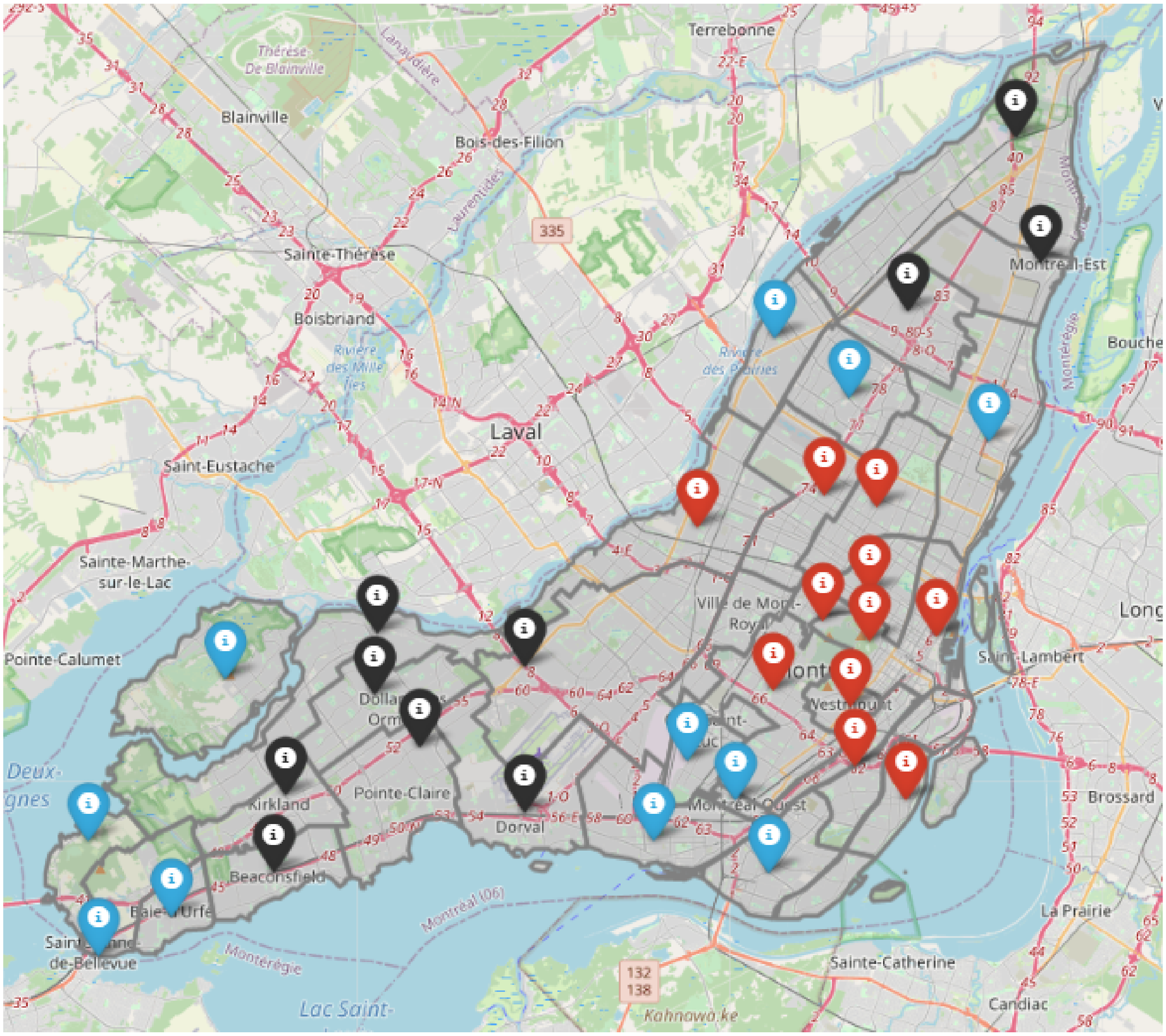


Neighborhood	Cluster
ahuntsic-cartierville	0
anjou	2
baie d'urfe	1
beaconsfield	2
cote-des-neiges-notre-dame-de-grace	0
cote-saint-luc	1
dollard-des-ormeaux	2
dorval	2
hampstead	1
kirkland	2
l'ile-bizard-sainte-genevieve	1
lachine	1
lasalle	1
le plateau-mont-royal	0
le sud-ouest	0
mercier-hochelaga-maisonneuve	1
mont-royal	0
montreal-est	2
montreal-nord	1
montreal-ouest	1
outremont	0
pierrefonds-roxboro	2
pointe-claire	2
riviere-des-prairies-pointe-aux-trembles	2
rosemont-la petite-patrie	0
saint-laurent	2
saint-leonard	1
sainte-anne-de-bellevue	1
senneville	1
verdun	0
ville-marie	0
villeray-saint-michel-parc-extension	0
westmount	0

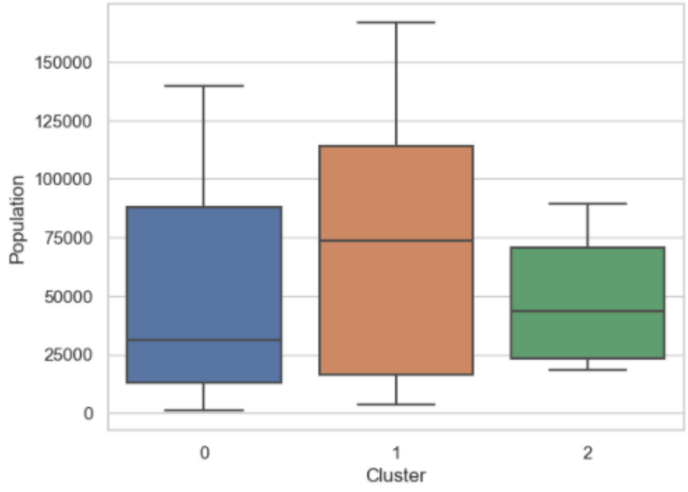
After observing the correlation matrix, we have done a KMeans clustering on our neighborhoods venues data.

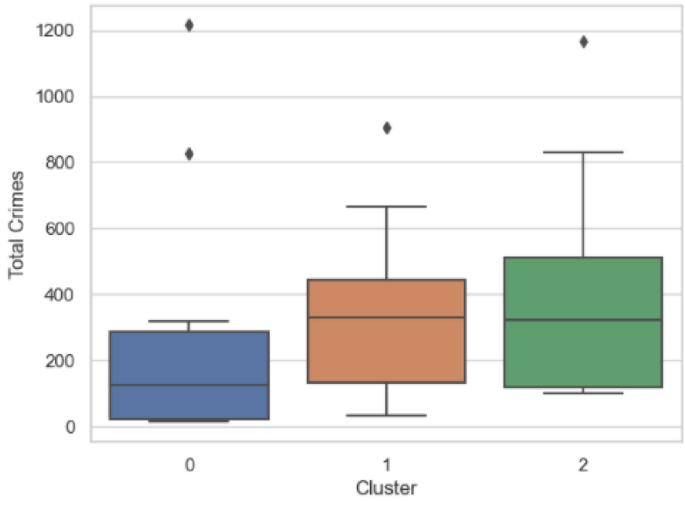
Cluster 0 = Red
Cluster 1 = Blue
Cluster 2 = Black

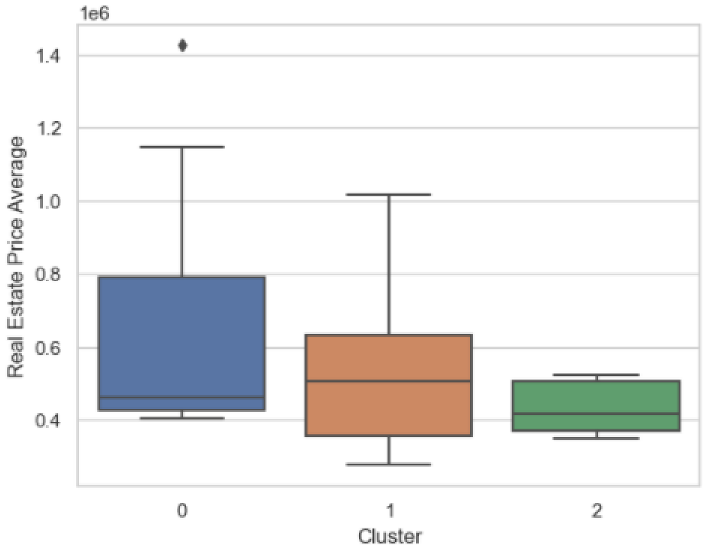
This map puts in perspective the repartition of the cluster's neighborhood. We can observe some proximity patterns, explainable by the venue's data pulled based on radius. Some venues overlap from one neighborhood to another.



The following 3 box plots put in perspective the range of real estate average price, criminal activities count and population count.







What we observe in the above boxplots

Clusters are based on venue repartition and occurrences. We've done so to have it as a baseline of our analysis. Here, we see the boxplots of our clusters, covering real estate price average, total crimes and population count. **Observation#1** points on a **low and close repartition** of real estate prices within the **cluster 2**. **Observation#2** points out on a low and fairly spreaded crime counts across **cluster1**(a few exceptions). **Observation#3** analyze the correlation between population and crimes which are seeing less through our clustering.

5.Conclusion

The analysis we've done observed multiple correlations and possible trade-offs required when selecting a neighborhood or even a cluster(of neighborhoods).

We've observed a clustering excessively relative to the proximity of the neighborhoods. This is explained by the same venues listed per Neighborhoods since they are within a certain radius. This is a good thing to leverage the proximity as part of our clustering. Definitely applicable for our **Cluster#0**, the other clusters are segmented either in 2 or 3 groups.

As low cost option, anywhere within **Cluster#2** is cheap and well surrounded by various venues. While this same cluster occupies a range of criminal activities, Beaconsfield is the neighborhood with the least.

Westmount identified as part of **Cluster#0** makes the difference with an average real estate cost near 1.5 million. The cluster has a median barely reaching the semi million. The criminality within the cluster is also low for having a median below 100 criminal activities within 6 months (note with a few exceptional neighborhoods).

What about **Cluster#1** ? Well... It is a cluster which obviously regroups neighborhoods with varied venues(Café/Coffee Shops, restaurants and other convenience stores). However, the neighborhoods from the west island are not sharing the same criminal records compared to the others. As an example, Baie d'Urfe, Senneville, Sainte-Anne de Bellevue are the neighborhoods with the least criminal activities.

As a result, our clustering process helped on grouping neighborhoods similar in terms of venues and we've used the boxploting on verifying the pattern with our other datasets. The decision is yours and depends on how much you can afford and how much you rate the risk of criminality within the neighborhood.