Capstone Battle of Neighborhoods

-

The most appealing district for international students

By Othmane Laoufir

Table of Contents

Disclaimer

The following content has been produced for the sole purpose of education. The data used in this document has been obtained freely for a free personal use. It should not be disposed in any sort of commercial use and the author cannot be accountable for a lack of data accuracy nor for any loss or damage encountered by using this document.
The software, data, computations and results are provided “as is” without warranty of any kind, express or implied, including but not limited to the warranties of merchantability, fitness for a particular purpose and noninfringement.

Introduction

North American universities are considered the best institutions concerning higher education. Each year these universities welcome thousands of students coming from abroad. In particular, Montreal ranks in the top ten destinations for studies according to international students.

Nevertheless, when deciding to study abroad, there are plethora of elements to prepare. And one of the most difficult parts of this preparation is related to housing. Whether it is for buying or renting a property, one must undertake several checks prior to taking any decision. Visiting the property is a crucial step that international students cannot obviously do.

The project, as presented by its owner consists on providing a turnkey housing to international students so everything is buttoned up a way that the client e.g. the student has nothing else to do than accessing the property when he arrives and get the best of his studying experience.

The Business Problem

Choosing the right property to invest in is crucial to the project owner. He must insure his return on investment by avoiding unoccupied properties on the one hand, and limit the initial investment amount on the other hand to keep the project sustainable.
Limiting the risk of unoccupied properties can be achieved through property selection : the property should be sufficiently appealing to students to insure that the offer will meet its demand. Since the project owner targets students, the first matter will be distance from universities. The second one, will be the offered environment : properties with more nearby activities should be more attractive than isolated ones.
Limiting the property’s total cost of ownership should also be more suitable by limiting the undertaken risk in case of project failure.

Thus, the question to address in this study could be summarized as follow:

(1) : *“What is the most reliable district to invest in, to launch a rental project addressed to international students?”*

This question itself must be divided into subsidiary issues:

(A) : *“Which housing features are relevant to fit a given student expectations ?”*

(B) : *“Which kind of nearby activities are the most likely to raise the property’s market value?”*

The Data Section

In this section, we discuss the elements that we will be able to leverage to address the business problems.
Ideally, one should access Nation-wide Real Estate data to conduct the study, but for the sake of keeping things simple and quickly achievable, having a sample of consistent housing data would be enough.
Then the collected data must be geocoded to query the Foursquare API to collect venues. A venue is an object that contains information about a given place.
Under the ***assumption*** that *distance to universities is likely to positively influence the price of a given property*, it could be relevant to measure this distance. This operation will also require geocoding.
To achieve data processing two developer accounts are required. Since Google geocoding service requires to pay, free alternatives had to be found.

- The Nokia HERE API has been used to geocode elements.
- The Foursquare API has been used to find nearby venues.

After collecting data, the inputs must be properly cleaned and formatted. Then each observable gets computed prior to any sort of analysis.

Methodology

Finding out which Neighborhood is the most suitable to run a rental niche business is equivalent to answering the question :

> "Is a given neighborhood suitable to attract international students ?"

Whenever the answer is yes :

> "To what extent is the given neighborhood suitable compared to the rest of neighborhoods"

Collecting data

Rental offers data has been obtained through web scraping. This step returned 589 entries.
Duplicated rows removal deleted 5 rows. This step returned 584 entries.
There is no intuitive way to control last step's output. Querying Nokia HERE returned 584 responses among which many were empty. In fact, deciding to keep an entry or not based on the API's capabilities was a quite good classifier since it prevents us from the hassle of checking each entry and deduce how to effectively patch it to produce a consistent dataset.
Regarding to data comprehensiveness in non-null responses. It could be deemed that any provided address passing the HERE API query filter is reliable. 370 items were returned out of 584 queries.
Then categories having Nonetype district values turned out to be garbage data, with very few exceptions that have been handled manually.
For the rest of entries having awkward values, the initial list of official neighborhoods helps confronting venues dataset's column data to a predefined model. This step ends up with 313 valid entries belonging to 19 distinct neighborhoods.
Each entry is then put into the Foursquare API to find out nearby venues. The API queries returned 8132 venues in Montreal.

Cleaning data

Basic cleanings

To properly cope with data, several transformations had to be made on the dataset prior to being able to make any computation.

The data collection processes has output two different datasets. A dataset containing rental offers in Montreal, and another one containing venues at Montreal.

After making sure that each entry of the first dataset corresponds to an actual district of the city, each row is then transformed so its stored price is computable.

Data geocoding

The current dataset consisting on a collection of labeled addresses or at list indicative information about the exact property location.

On querying the HERE API, the response object can either:

- Be empty in case the API fails to serve the related query information
- An object containing :
  - ✔ The full address of the query information.
  - ✔ The GPS coordinates of the query information.

For each query, the response object is handled in a way to collect :

- the Address label
- the belonging District
- the Latitude coordinate
- the Longitude coordinate

Whenever the response object is empty, the query index is identified, and the corresponding instance in the offers dataset is deleted. At the end of this step there are 370 observable in the offers dataset.

On the next step the dataset has been filtered on the empty District criterion. It turned out that all of the filtered data was irrelevant, being addresses in other cities to other continents in the most extreme cases. Except for one value that has been manually corrected, all of the rest has been deleted. At the end of this stage, the dataset contains 313 observable.

In this model, an offer is an object characterizing a property for rent in the area of Montreal, QC and defined by the following features :

- an Address label
- a belonging District
- a Latitude coordinate
- a Longitude coordinate
- a rental Price

Each offer has then been used to explore activity-related locations on the Foursquare API. A venue is basically any non-residential place which main features are :

- Venue Location
- Venue Category

Each query on the Foursquare API consisted on finding up to 50 venues located 500 meters around the provided Latitude and Longitude coordinates. The retained elements to constitute the Venues dataset were :

- an Name label
- abelonging Category
- a belonging District
- a Latitude coordinate
- a Longitude coordinate

Computing distances

As stated in the introductory part, having an idea of the distance between a given property and local universities could bear valuable information. To keep track of this measure, since geocoding data is available, the distance between two points a (x , y) and b (x' , y') can be computed through calculating the Euclidian distance between two points as stated in the equation below:

$$E=\sqrt{(x-x')^2+(y-y')^2}$$

But as this computation is achieved through the whole Rental Offers dataset, further transformations, namely grouping and clustering could eventually break the relevance of such indicator.
As data is aggregated, the distance column content is made of a certain number of districts, each containing a given number of observable from which a Euclidian distance has been computed.

$$S=\sum_{i=1}^{N} e_i$$

Where e_i is the Euclidian distance computed for a given observable.

Since multiple locations are summarized by their belonging district feature, during the grouping process, each neighborhood in town gets allocated the mean geocoding value corresponding to the mean of latitudes and longitudes of the group of observable. On the same time the distance column also gets summarized by its mean value which keeps features consistent.

Grouping properties

Grouping observable by Neighborhood helped transforming the shape of the dataset from 313 entries to 18 entries.

Each entry of the grouped offers view is characterized as below :

- a Neighborhood name label
- a Latitude coordinate
- a Longitude coordinate
- an Average Distance to universities
- a property Count

Counting properties

Each observable belonging to a unique District, that we imposed to be not empty, counting the number of offers by Neighborhood is simply achieved through summing every observable having the same belonging District.

Venues categorization

Another thorough task consisted on transforming the venues dataset by aggregating the different venues categories. The dataset initially contained 306 different categories. The aggregation process consisted on relabeling each category to a more standardized categorization.

The retained categories were defined as follow:

Essentials : Necessary services as supermarkets, banks, pharmacies and so on.

Leisure : Restaurants, parks, and cultural places

Casual Food : Nonessential food.

Entertainment : Nightlife, shows and other related activities.

Retail : Shops, boutique, and other sales-related activities.

Sport : sport-related infrastructures, facilities and clubs

Transportation : Public transportation services

Other : Anything else.

Relabeling values

Once that the dataset is split into eight categories, since each category is mutually exclusive with one another, an observable belonging to a category X will necessarily be excluded from $\overline{X}$ meaning any other category.

Thus, the venue–category column content can be re-encoded into eight distinct columns bearing a Boolean value v encoded by the scalar $\{ 0 , 1 \}$

Analyzing data

The first thing to discuss here is data consistency. Responses that have been collected are a sample corresponding to the best Marketing fit to the querying user profile. Since the venues dataset is generated from the rental offers dataset baseline, it inherits the same biases than its predecessor.
Since the rental offers dataset observations are not only targeting students, using this information could lead to biased results. Nonetheless, performing an analysis without any control variable could end up with inaccurate results.
The preferred approach consists on clustering neighborhoods on the sole basis of venues, and then control the output by leveraging the rental offers dataset.

Working datasets

Two distinct datasets were prepared for different purpose.

As the Rental Offers dataset is potentially biased by the abnormal distribution of offers geography, the subsequent venues dataset will necessarily reproduce the distribution features and thus shape the sample.

Moreover, Rental Offers dataset describes the local rental market as a whole, and not as offers targeted to our use case random client profile.

To cope with this issue, using one of two datasets as a testing sample is quite relevant. Under the ***assumption*** that :

"*The existence of a venue at a given localization is likely to influence nearby properties' rental price*"

Since the two working datasets are bound by the ***relationship*** stating that :

"*Each venue is located at less than 500 meters of a Rental Offer.*"

Rental Offers dataset features, namely :

- Price,
- Local Count of Offers,
- Average Distance from universities

...can be leveraged to control the model coherence. As soon as those indicators are not input in the modeling process, it can be used to stress-test the model results.

Choice of algorithm

Preliminary considerations

The pre-computation limits raised so far are:

- Sample-wide abnormal distribution of geographical observable
- No intuitive independent variable to train the model

Thus any supervised learning approach would poorly address the questions raised in the Business Problem section.

The principle of unsupervised approaches consists on letting the software

- compare each obsevable to every other observable of the considered class, feature by feature
- split the dataset into consistent groups based on the previous comparison
- classify each observable into one of the groups previously defined

The right algorithm

The chosen algorithm to perform the classification is K-Means.

The K-Means clustering consists on computing unlabeled data belonging group – or cluster – based on the calculation of the distance between each datapoint and the center of the corresponding cluster.

Each row of the unlabeled data matrix F consisted on the sample frequencies of the eight features corresponding to general venues categories.

For two distinct points in the sample, the first is a random datapoint that is not a cluster center, the second is any other point that is a cluster center. The first datapoint belongs to the second's cluster as soon as the Euclidian distance between the two points is relatively inferior to the same computation applied to other clusters center.

The matching process described above is applied to the eight features' relative frequencies. A cluster is a category containing datapoints sharing similar features within the group and being as dissimilar as possible to other clusters' features. For any given observable, the sum of frequencies should return 1. Every datapoint should fall into exactly one cluster.

Challenging the output model

Back to the data cleaning step Neighborhood coordinates were defined as the couple { x , y } having x and y corresponding to the respective mean values of Latitude and Longitude within a subgroup of the dataset defined as a District-wide subset of the Rental Offers dataset.

The Euclidian distance between each datapoint and its District coordinates is equivalent to computing the mean squared error (MSE) of the grouped subset.

Regarding to the District column consistency, having District instances categorized by their nearby activities, one can control the modeling step output by exploring the market data by cluster to check the grouping accuracy.

The offers Count can serve as a ranking criterion. This indicator provides information about housing local scarcity, and local attractiveness for the rental activity as well.

The Average Price can also serve as a ranking criterion. This indicator evaluates the offer value by district. Assuming that Rental market structure does not allow the demand-side to negotiate the price, offer values are deemed to be market values.

The Average Distance to universities can be useful to measure the impact of proximity on rental price in a supervised approach. Nevertheless, using this measure ex-post in a clustering approach will enable the analyst to assess the overall model relevance. Having :

H_0: "The average distance to universities impacts the observable classification"

The advantage of doing so will confirm the relation if the clustering outputs observable do fall within the same distance bin. Alternatively, it will invalidate the hypothesis if H_0 the average distance to universities feature is not shared by observable belonging to the same given cluster.

Results

The clustering process with K = 4 returned four groups indexed from 0 to 3. For the sake of simplicity each index value is incremented by 1 so the clusters are labeled from 1 to 4.

Output description

Clusters distribution

Out of 18 inputs, clusters 1 and 4 population turned out to be outliers, and clusters 2 and 3 contained comparable values :

Cluster 1 contains 1 District
Cluster 2 contains 2 Districts
Cluster 3 contains 14 Districts
Cluster 4 contains 1 District

- ✔ Cluster 1 distance to universities is nearly the triple of the sample mean distance value. At the contrary Cluster 4 distance to universities is about 33% the sample mean distance, but 100% of nearby venues fall into the Other categories meaning that the neighborhood is not attractive.

- ✔ Cluster 2 contains two observable. The mean distance to universities is dissimilar among the cluster's population. But interestingly the market rental price of both neighborhoods is quite similar, ranging from $950 to $960.

- ✔ With more than 77% of entries falling into cluster 3, paying a special attention to what is in it turns out to be crucial. 8 observable out of 14 are above the subset mean distance to universities. There are 7 observable out of 14 above the subset mean rental price. On the same time, although the subset median price is about $200 below the mean subset price, every district belonging to the five first deciles is below both the median and mean subset price.

Segments description

Cluster 1 describes peripheral Districts which properties are relatively far from local universities and having mainly access to sport-related infrastructures.

Cluster 2 describes downtown districts containing properties falling into the same pricing bin – Around $950.

Cluster 3 describes downtown districts offering a quite high level of nearby activities, each providing balanced types of activities.

Cluster 4 describes Districts offering only out-of-scope nearby activities.

Output evaluation

Data summary

Cluster	Neighborhood	Count - Price	Average - Price	Average - AvgDist
1	Anjou	1	$2,400.00	0.106670813
2	Ahuntsic-Cartierville	9	$960.56	0.046252543
	St-Léonard	1	$950.00	0.096266200
3	Côte-des-Neiges-Notre-Dame-de-Grâce	66	$1,205.98	0.024546990
	La Salle	8	$1,767.50	0.059574432
	Lachine	11	$1,000.91	0.056319707
	Le Plateau-Mont-Royal	56	$1,812.48	0.027594900
	Le Sud-Ouest	21	$2,098.57	0.027565790
	Mercier-Hochelaga-Maisonneuve	9	$991.28	0.050314554
	Montréal-Nord	6	$721.50	0.107675236
	Outremont	13	$2,189.62	0.026172039
	Pointe-aux-Trembles	7	$968.57	0.180830855
	Rosemont-la Petite-Patrie	23	$1,776.09	0.047049119
	St-Laurent	12	$1,382.75	0.019638902
	Verdun	13	$2,641.15	0.040085516
	Ville-Marie	36	$2,061.31	0.019099551
	Villeray-St-Michel-Parc-Extension	20	$1,431.50	0.050645270
4	Pierrefonds	1	$1,100.00	0.012993497
Total Result		**313**	**$1,618.27**	**0.037434330**

Table 1: District rental features summary.

		CasualFood	Entertainment	Essentials	Leisure	Other	Retail	Sport	Transportation
1	Anjou	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.58%	0.00%
2	Ahuntsic-Cartierville	0.28%	0.00%	2.35%	0.51%	0.00%	2.73%	1.75%	6.35%
	St-Léonard	0.00%	0.00%	0.29%	0.03%	0.00%	0.11%	0.00%	0.00%
3	Côte-des-Neiges-Notre-Dame-de-Grâce	12.39%	4.68%	21.65%	16.26%	2.17%	12.30%	14.91%	41.27%
	La Salle	0.77%	0.00%	0.88%	0.82%	0.00%	3.64%	2.05%	3.17%
	Lachine	0.42%	0.51%	1.27%	1.08%	2.17%	0.80%	1.46%	0.00%
	Le Plateau-Mont-Royal	38.99%	42.66%	27.52%	25.59%	28.26%	33.83%	38.60%	6.35%
	Le Sud-Ouest	6.83%	4.81%	7.35%	6.21%	14.13%	5.69%	11.70%	9.52%
	Mercier-Hochelaga-Maisonneuve	1.20%	1.14%	1.76%	1.76%	2.17%	2.16%	0.88%	15.87%
	Montréal-Nord	0.07%	0.13%	1.18%	0.60%	1.09%	0.68%	0.29%	0.00%
	Outremont	6.12%	2.66%	7.44%	7.26%	1.09%	4.56%	1.75%	0.00%
	Pointe-aux-Trembles	0.28%	0.00%	0.29%	0.17%	4.35%	0.68%	0.00%	3.17%
	Rosemont-la Petite-Patrie	8.66%	8.99%	8.81%	7.80%	0.00%	11.16%	5.56%	1.59%
	St-Laurent	0.49%	0.13%	1.08%	1.13%	1.09%	2.51%	0.88%	6.35%
	Verdun	1.06%	1.77%	3.62%	2.84%	3.26%	1.71%	2.05%	0.00%
	Ville-Marie	18.30%	31.39%	10.19%	24.17%	36.96%	12.07%	9.36%	3.17%
	Villeray-St-Michel-Parc-Extension	4.15%	1.14%	4.31%	3.77%	2.17%	5.35%	8.19%	3.17%
4	Pierrefonds	0.00%	0.00%	0.00%	0.00%	1.09%	0.00%	0.00%	0.00%
Total		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

Table 2: Downtown activities distribution among Districts

When it comes to cluster analysis, it seems that clusters 1 and 4 are considered outliers as soon as all of their activities fall into a single category.

Cluster 2 data is more balanced in terms of activities available nearby.

And finally cluster 3 data is best described by its leisure nearby activities. Indeed leisure activities frequency is quite high for merely every observable. Essentials activities are frequently represented as well among the subset data. Then comes Casual Food and Retail basically. Thus, cluster 3 locations show quite intensive activity. So far, this could result in higher attractiveness for this area.

		CasualFood	Entertainment	Essentials	Leisure	Other	Retail	Sport	Transportation	Total
1	Anjou	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	100.00%
2	Ahuntsic-Cartierville	5.00%	0.00%	30.00%	22.50%	0.00%	30.00%	7.50%	5.00%	100.00%
	St-Léonard	0.00%	0.00%	60.00%	20.00%	0.00%	20.00%	0.00%	0.00%	100.00%
3	Côte-des-Neiges-Notre-Dame-de-Grâce	14.74%	3.10%	18.51%	47.99%	0.17%	9.05%	4.27%	2.18%	100.00%
	La Salle	12.22%	0.00%	10.00%	32.22%	0.00%	35.56%	7.78%	2.22%	100.00%
	Lachine	8.00%	5.33%	17.33%	50.67%	2.67%	9.33%	6.67%	0.00%	100.00%
	Le Plateau-Mont-Royal	21.87%	13.30%	11.09%	35.61%	1.03%	11.73%	5.21%	0.16%	100.00%
	Le Sud-Ouest	18.03%	7.06%	13.94%	40.71%	2.42%	9.29%	7.43%	1.12%	100.00%
	Mercier-Hochelaga-Maisonneuve	12.14%	6.43%	12.86%	44.29%	1.43%	13.57%	2.14%	7.14%	100.00%
	Montréal-Nord	2.33%	2.33%	27.91%	48.84%	2.33%	13.95%	2.33%	0.00%	100.00%
	Outremont	17.86%	4.31%	15.61%	52.57%	0.21%	8.21%	1.23%	0.00%	100.00%
	Pointe-aux-Trembles	16.00%	0.00%	12.00%	24.00%	16.00%	24.00%	0.00%	8.00%	100.00%
	Rosemont-la Petite-Patrie	18.17%	10.49%	13.29%	40.62%	0.00%	14.48%	2.81%	0.15%	100.00%
	St-Laurent	7.87%	1.12%	12.36%	44.94%	1.12%	24.72%	3.37%	4.49%	100.00%
	Verdun	7.85%	7.33%	19.37%	52.36%	1.57%	7.85%	3.66%	0.00%	100.00%
	Ville-Marie	15.87%	15.14%	6.35%	52.01%	2.08%	6.47%	1.95%	0.12%	100.00%
	Villeray-St-Michel-Parc-Extension	18.21%	2.78%	13.58%	41.05%	0.62%	14.51%	8.64%	0.62%	100.00%
4	Pierrefonds	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	100.00%

Table 3: Activities distribution by District

The second most expansive District, namely Ville-Marie also offers one of the highest levels of activity. About 20% of the collected Venues are actually located nearby Ville-Marie, for an average rental price of $2061,31.

At the opposite, Mercier-Hochelaga-Maisonneuve is the second cheapest District in cluster 3. It offers 140 nearby Venues over 8132 collected which corresponds to about 1,72% of the city activities. Rental price overthere is around $991,28.

Having its average rental price around $200 above the sample mean rental price, the District of Plateau-Mont-Royal offers the highest level of activity in the city. On top of every distribution of activity features, the neighborhood represents 31% of sample's activities.
There are also numerous rental offers, which indicates that the district is potentially on high demand. The neighborhood's average distance to universities is also 30% below the sample mean distance from universities which makes it a good choice.

Discussion

Regarding to the clusters distribution, one can wonder if in-cluster consistency is achieved. The outliers are correctly identified by the algorithm. Nonetheless, the potential lack of homogeneity among cluster 3 observations raises some limitations towards using a fully unsupervised approach.

A further step could be undertaken to extend the current study findings by

- measuring the level of significance for the location factor : to what extent having a property located in a district rather than one another can lead to a shift in its pricing.
- measuring the impact of each venue category on the overall property valuation on the rental market

But moving onto the supervised realm implies having a proper controlling feature. The most intuitive one is the property rental place.

To optimize the modeling process, the Rental Offers dataset should be purged from non-compliant properties. Every property that is not suitable to international students must be excluded from the study sample.

Meanwhile the number of observations, should be raised to compensate the shrink in the sample by an extended range of compliant observations.

To do so, the analyst must have access to a client profile dataset containing consumer preferences to constitute the control variable dataset and guaranteeing its independence from the model input factors. Such dataset should eventually contain

- activity-related categories the client prefers
- expectations from a short-term rental e.g less than 3 years in terms of
  - Property size
  - Indoor design
  - Property features
  - Opinion on appartment-sharing
  - Target price

Constituting such dataset could be very helpful as a test sample while other factors leveraged in the current study could serve as predictors to regress to find out the best fit to both the project owner’s constraints and the international student’s expectations.

Conclusion

To conclude, choosing the right area to launch a rental business is not as straightforward as one could imagine. Depending on the project positioning, the so-called “right” Neighborhood to select will change.

For a very high value added offer, positioned on the luxury segment, the project owner can choose to rent appartments in Ville-Marie. The district is very active, and host several cultural events which should attract students. This choice comes at a quite high price since properties are very expansive in this area. Thus, making this choice, though risky, is likely to offer a good payoff.

For a standard offer targeting functional housing, choosing Mercier-Hochelaga-Maisonneuve is the cost-killing solution. Properties are affordable for a first investment. The neighborhood, although far less active than the previous one, still offers a balanced level of nearby activities for the daily student life.

Finally, the study reveals that the best deal achievable is around Le Plateau-Mont-Royal. With a very intensive level of activities, an acceptable distance from local universities, and prices ranging near the central tendency measure of the dataset, this district is affordable and attractive.