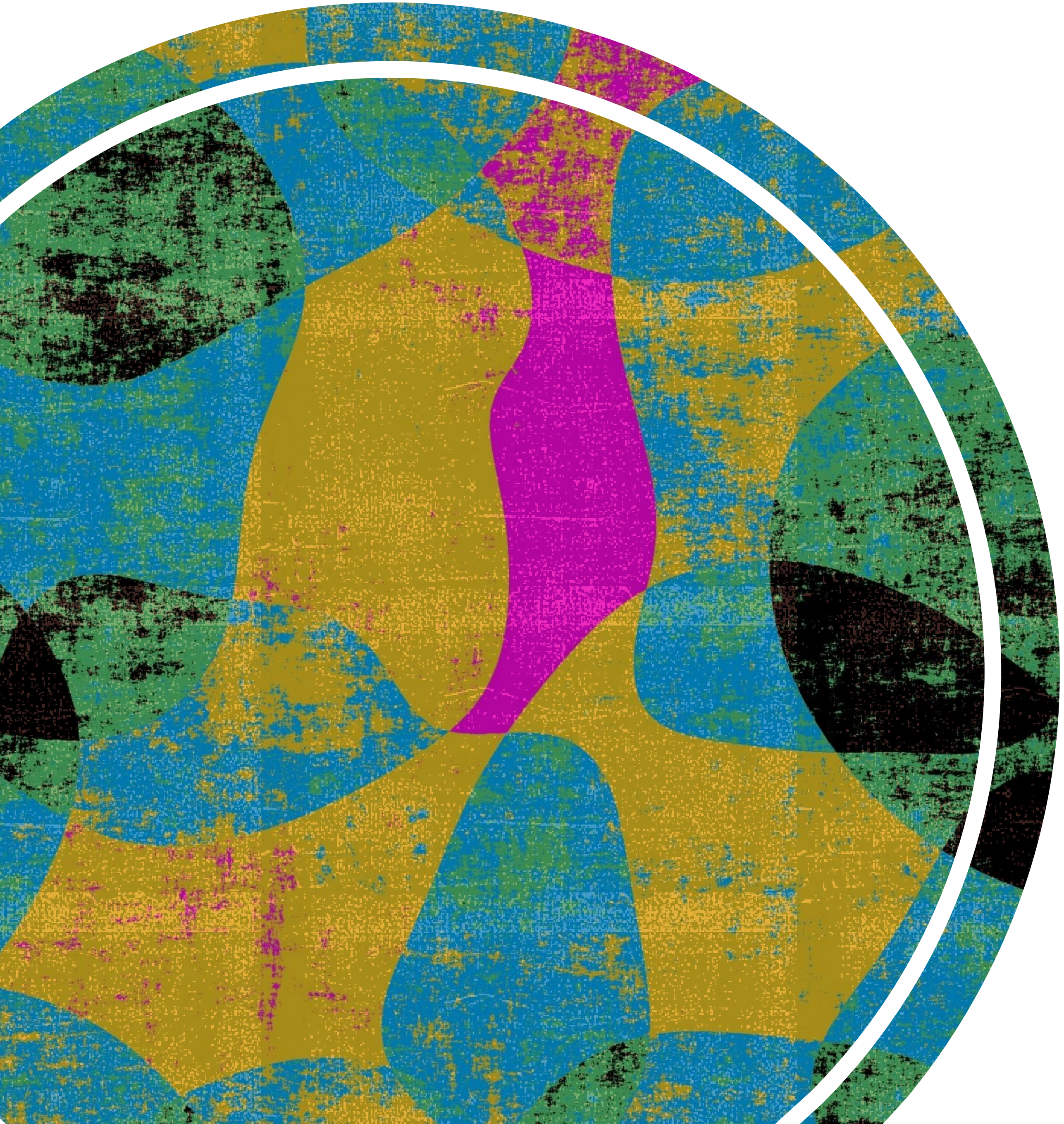


INTERCITY NEIGHBORHOOD CLUSTERING

Efren Mora

2020-09-02

CASE

A small chain business owner stablished in the city of Toronto is looking to expand its business to another big city in Canada and is particularly interested in Montreal. The idea of expanding is new, and she wants to be thorough. She is quite familiar with the city of Toronto and its neighborhoods, but not so much with the city of Montreal. So, to be better oriented and get a better feel of this new city, she has asked for help identifying and mapping equivalent neighborhoods in these two cities and to carry a comparative analysis among these in terms of venue composition and census features.

DATA SOURCES

Foursquare API:

- Main tool. It was used to virtually explore Toronto and Montreal Neighborhoods.
- When making ‘GET Venues’ requests to the Foursquare API, I used the ‘search’ endpoint to pick-up venues within specific category IDs.
- I used all the header (or main) category IDs from the Foursquare category tree.
  - **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

DATA SOURCES

Wikipedia:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_H

- To obtain the addresses for Canadian neighborhoods, which are the first 3 digits of a Canadian postal code, formally referred to as a forward sortation area or FSA for short.

Geocoder (Powered by Bing search engine):

- To convert FSAs to geographical coordinates

FSA Boundary File from Census 2016: https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2016-eng.cfm

- To show the FSA neighborhoods boundaries superimposed on a map of the city of Toronto, and Montreal.

DATA SOURCES

Population by FSA from Census 2016: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/comprehensive.cfm

- To calculate and include population density as a feature for neighborhood clustering.

Median total income of households per FSA from Census Profile, 2016 Census: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/download-telecharger/comp/page_dl-tc.cfm?Lang=E

- To include median income as a feature for neighborhood clustering.

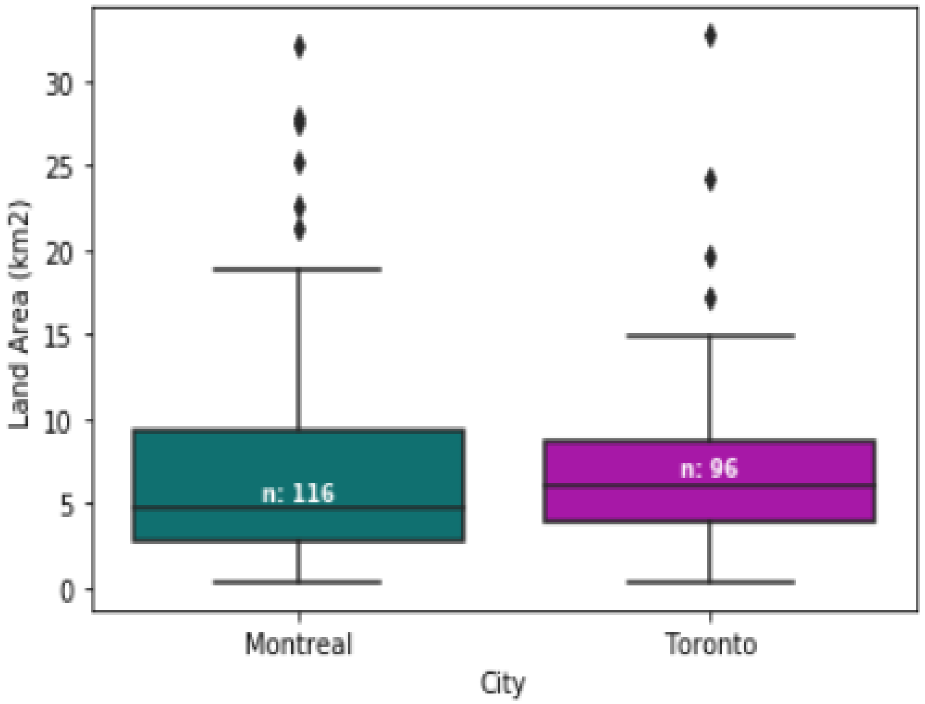
METHODOLOGY

- The goal was to identify and group into clusters neighborhoods that share similar venues composition and certain census attributes such as population density and median income of households. By venues composition we mean the most frequent venue category types found in a neighborhood or neighborhood cluster.
- To accomplish this task, the unsupervised learning K-means clustering algorithm was used. This is a versatile algorithm that can be used to identify unknown groups in complex data sets.
- In essence, I fed the algorithm an unlabeled dataset containing all the Toronto's and Montreal's neighborhoods and their respective features.
- The features were: the mean of the frequency of each venue category found in each neighborhood, and each neighborhood's normalized population density (pop per Km2), median income of households, and venues density (venues per km2).
- The elbow method helped identify the value for K, that is into how many groups or clusters our pool of neighborhoods can be divided. The neighborhoods that end up in a given cluster were similar to each other and dissimilar to neighborhoods in another cluster.

EXPLORATORY DATA ANALYSIS

Demographic Variables Distribution

Land area (km2) per neigh

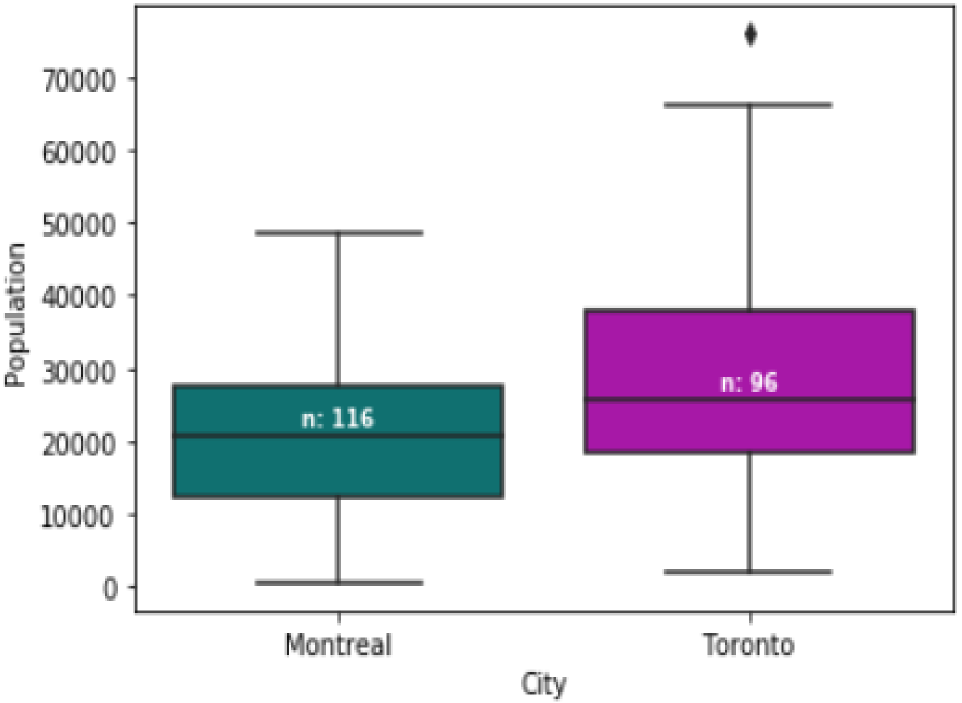


	Montreal	Toronto
# of Neighborhoods	116	96
Surface Area (km2)	775.75	663.36
Population	2,365,019	2,732,094
Pop Density (# of people / km2)	3,048	4,118

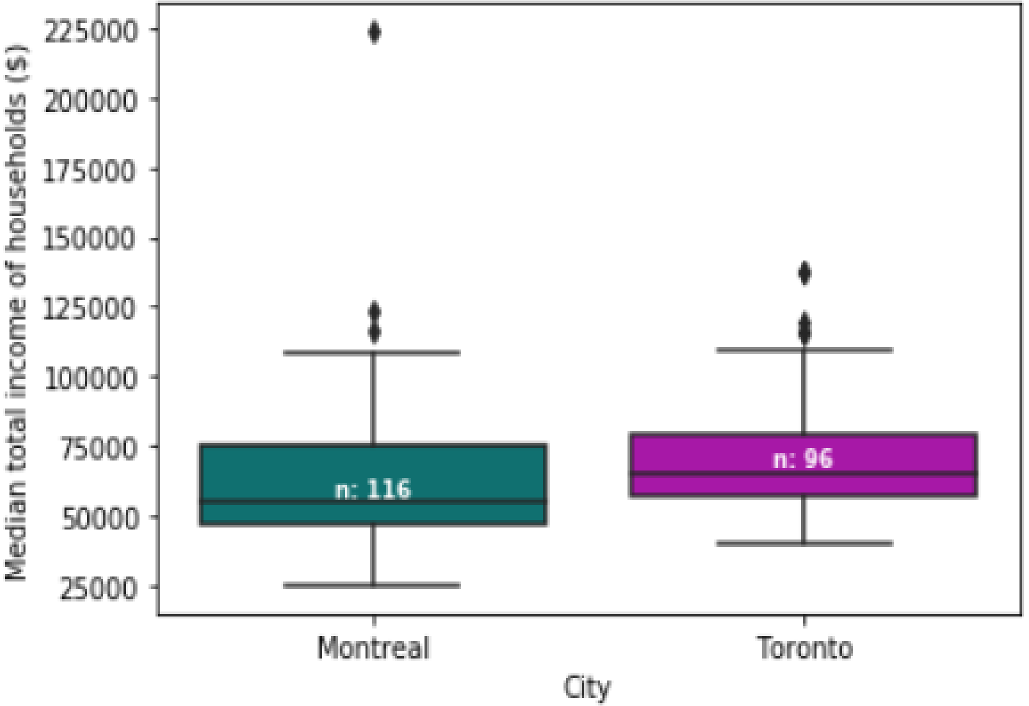
EXPLORATORY DATA ANALYSIS

Demographic Variables Distribution

Toronto's neighborhoods tend to be more populous than those of Montreal



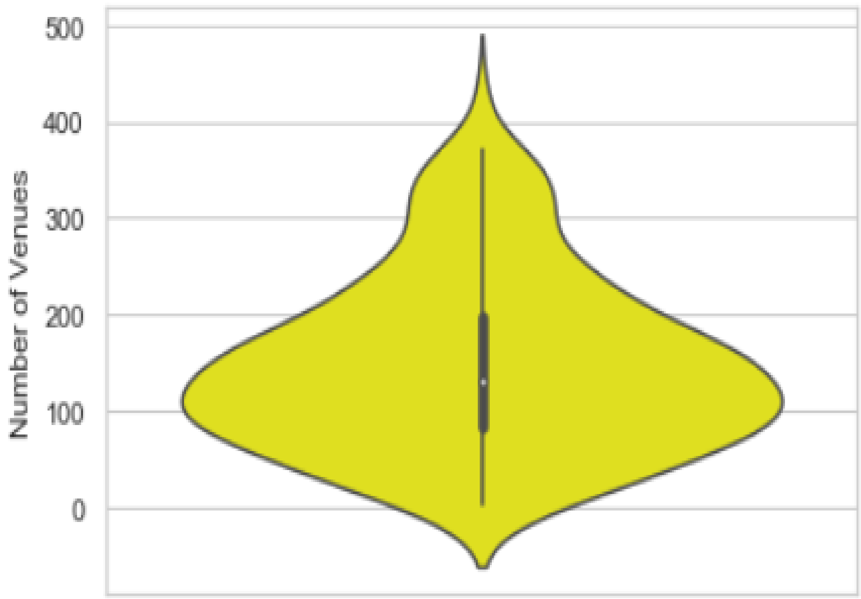
Montreal has a bigger proportion of poor neighborhoods



EXPLORATORY DATA ANALYSIS

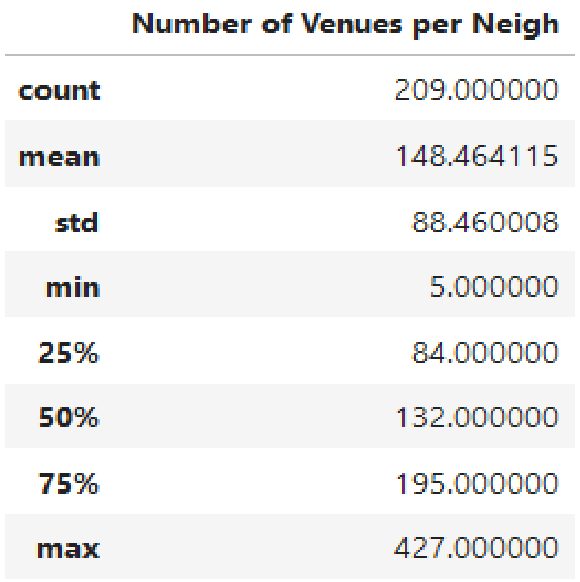
Venues Distribution

- The top half neighborhoods have a bigger differences in the number of venues per neighborhood compared to the lower half.



# of Venues per Neighborhood Distribution

- After data cleaning 209 neighborhoods (from 212) and 31,018 venues (from 31,035) remained to be analyzed.



	Number of Venues per Neigh
count	209.000000
mean	148.464115
std	88.460008
min	5.000000
25%	84.000000
50%	132.000000
75%	195.000000
max	427.000000

K-MEANS CLUSTERING ANALYSIS

Feature Transformation

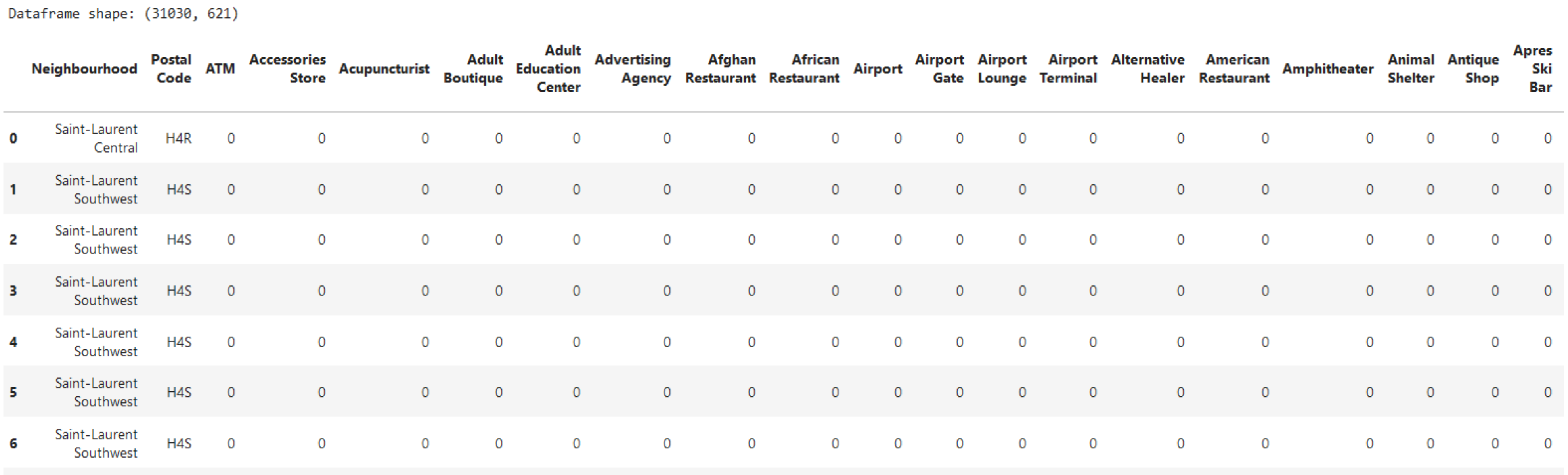
Before we can feed our data to the algorithm, we need to transform and normalize our features.

One Hot Encoding

- K-means clustering cannot work with categorical data. We need to encode the *venue categories* to a numerical form.
- This procedure requires that the categorical values be mapped to integer values.
- Then, each integer value is represented as a binary vector, that is all zero values except for the index of the integer, which is marked with a 1.

K-MEANS CLUSTERING ANALYSIS

Feature Transformation



```
Dataframe shape: (31030, 621)
```

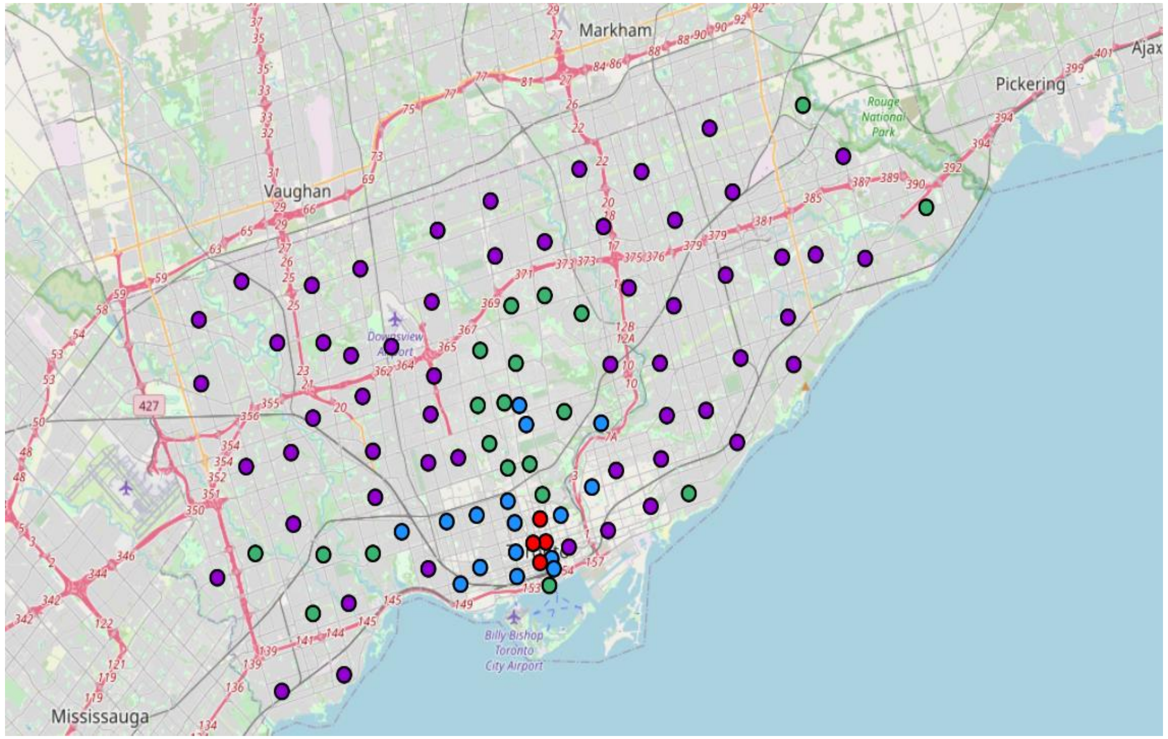
	Neighbourhood	Postal Code	ATM	Accessories Store	Acupuncturist	Adult Boutique	Adult Education Center	Advertising Agency	Afghan Restaurant	African Restaurant	Airport	Airport Gate	Airport Lounge	Airport Terminal	Alternative Healer	American Restaurant	Amphitheater	Animal Shelter	Antique Shop	Apres Ski Bar
0	Saint-Laurent Central	H4R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	Saint-Laurent Southwest	H4S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	Saint-Laurent Southwest	H4S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	Saint-Laurent Southwest	H4S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	Saint-Laurent Southwest	H4S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Saint-Laurent Southwest	H4S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	Saint-Laurent Southwest	H4S	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- Each row represents 1 of the 31,018 venues found.
- The first column holds the name of the neighborhood in which the venue was found.
- Each of the remainder columns represent one of the 612 unique venue categories.
- Then, for a particular row, there is a value of 1 under the category name of the venue found, and a value of zero under the rest

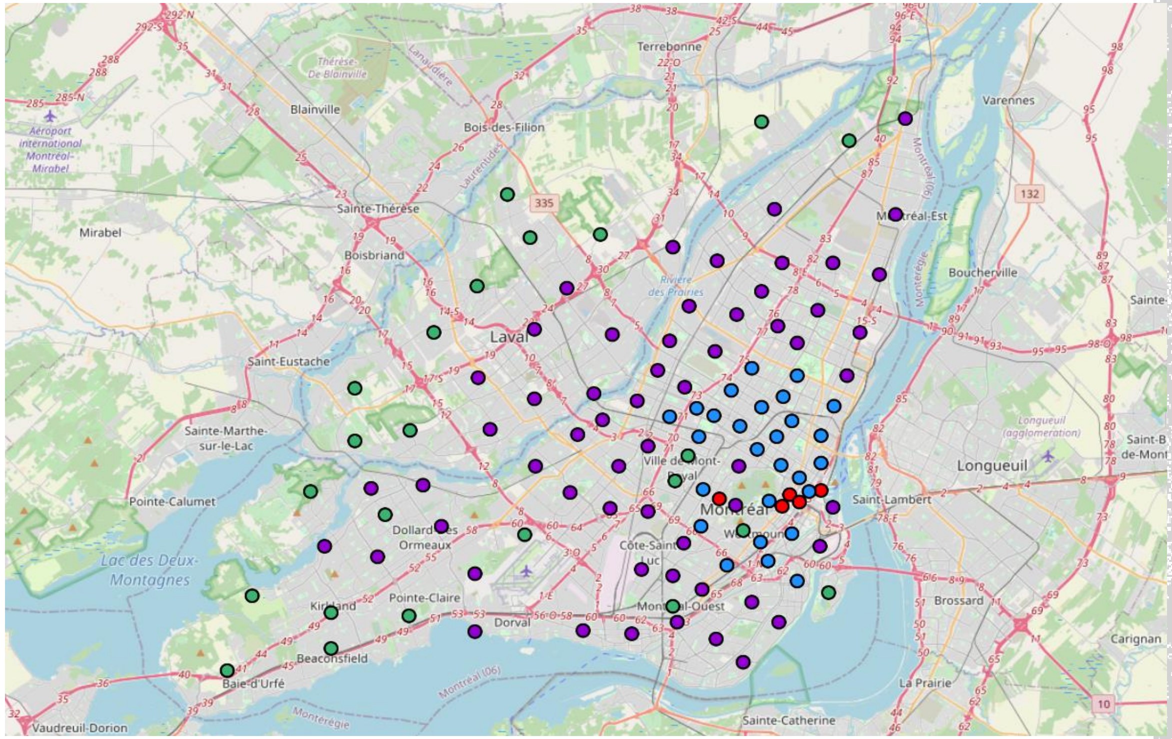
RESULTS

Neighborhood Clusters

Metropolitan Toronto



Metropolitan Montreal

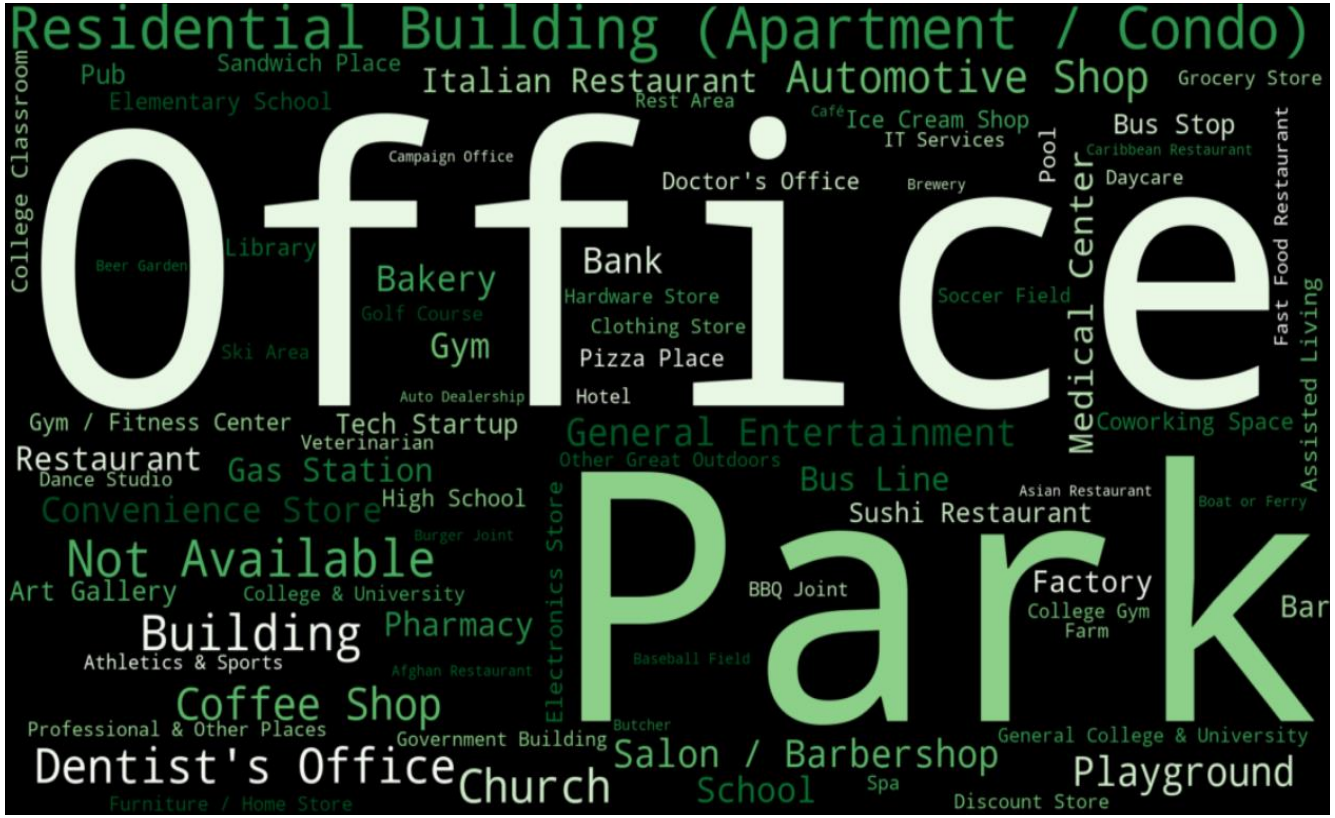


| Cluster 1 = **Green** | Cluster 2 = Blue | Cluster 3 = **Red** | Cluster 4 = **Purple** |

NEIGHBORHOOD CLUSTERS

Cluster 1 [Green] – Residential and Commercial Areas | Mid-Town

Most Popular Venue Categories



Characteristics

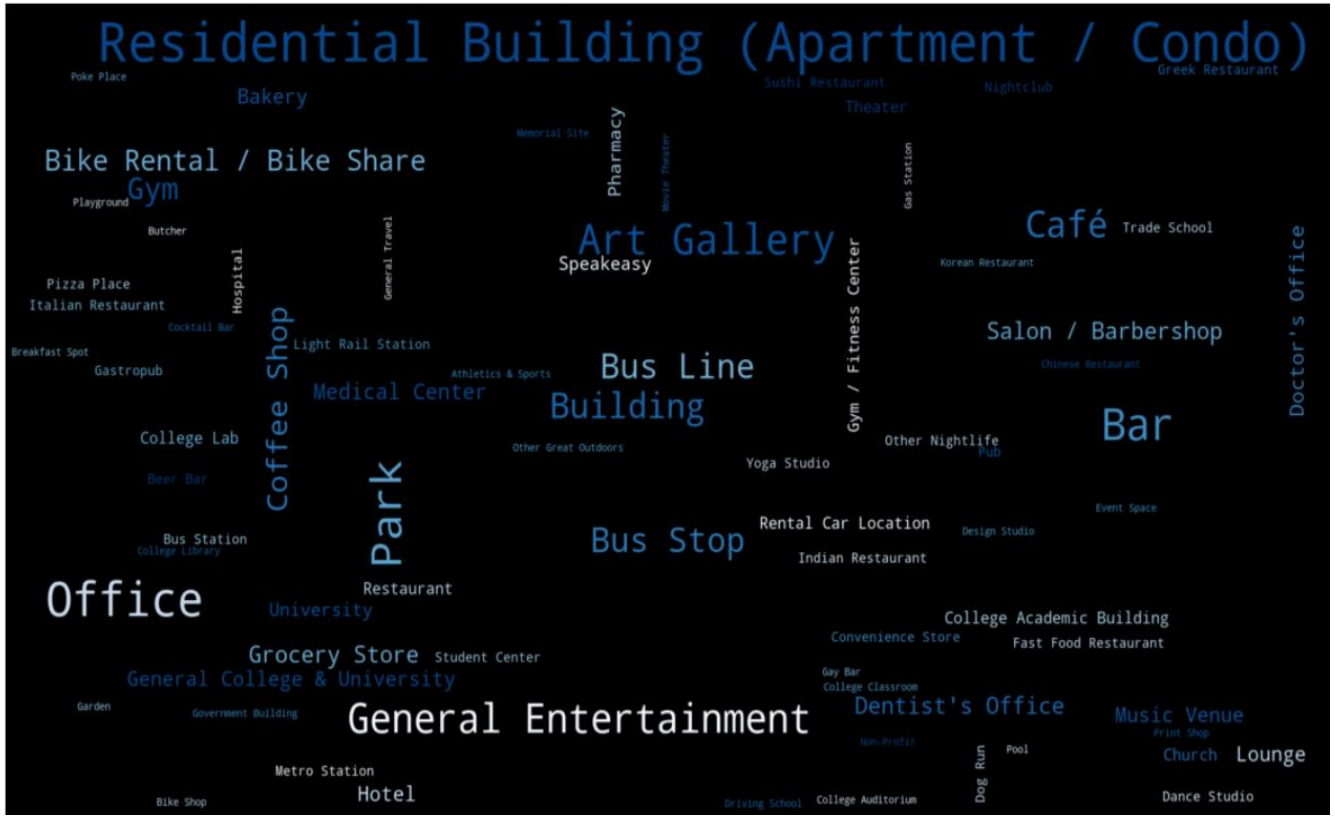
# of Neighborhoods	**Population Density**
Total: 43 Montreal: 23 Toronto: 20	Low
Venues Density	**Median Income**
Low	High
Popular Venue Categories	
• Parks • Offices • Residential Buildings • Playgrounds	• Bakeries • Italian Restaurants • Bus lines • Gyms • and more…

* Refer to supporting graphs on slides 17 and 18

NEIGHBORHOOD CLUSTERS

Cluster 2 [Blue] – Residential and Commercial Areas | Downtown

Most Popular Venue Categories



Characteristics

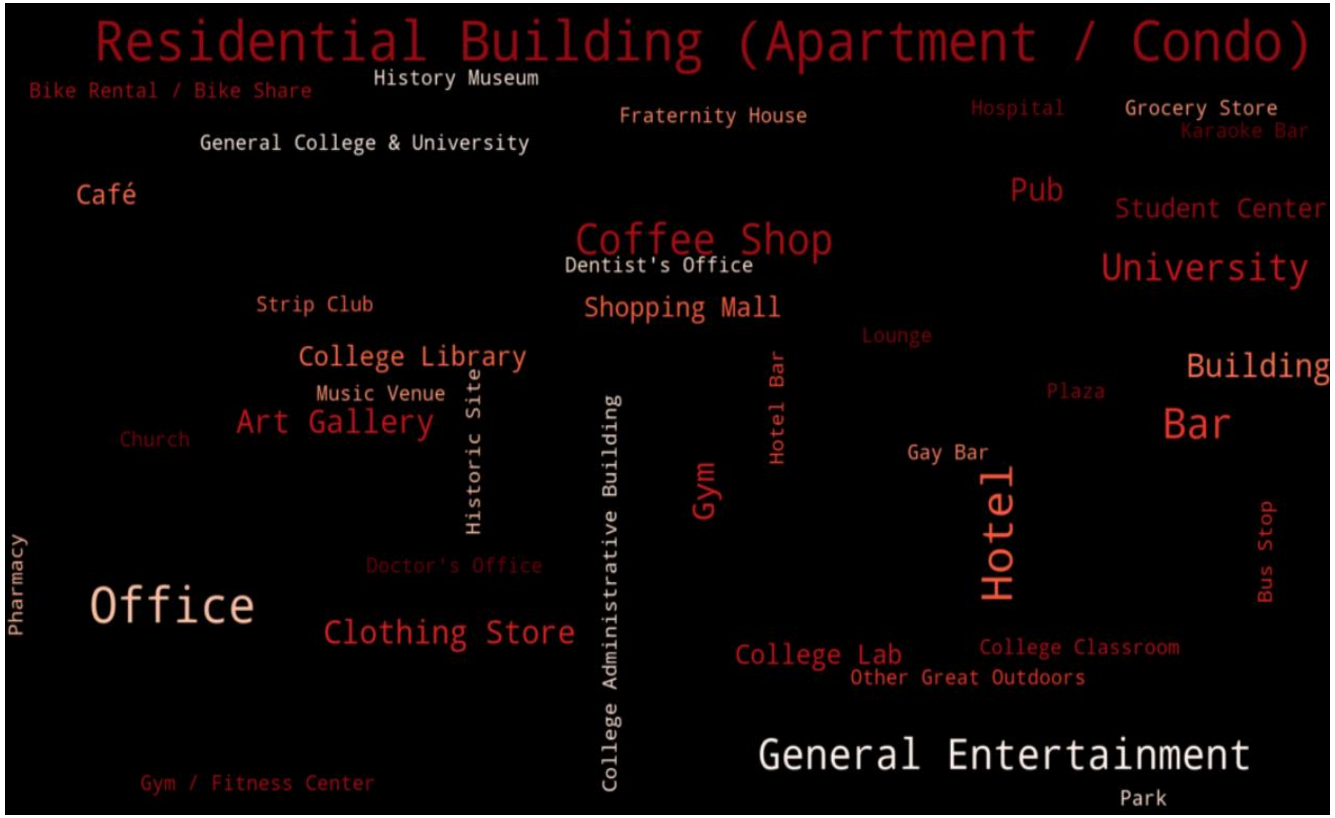
# of Neighborhoods	**Population Density**
Total: 43 Montreal: 27 Toronto: 16	Med-to-High
Venues Density	**Median Income**
Med	Med-to-High
Popular Venue Categories	
• Entertainment Venues • Bars • Cafes • Restaurants	• Art Galleries • Condos & Apartments • Bus lines, Metro Stn • and more…

* Refer to supporting graphs on slides 17 and 18

NEIGHBORHOOD CLUSTERS

Cluster 3 [Red] – Mostly Commercial Areas | Downtown

Most Popular Venue Categories



Characteristics

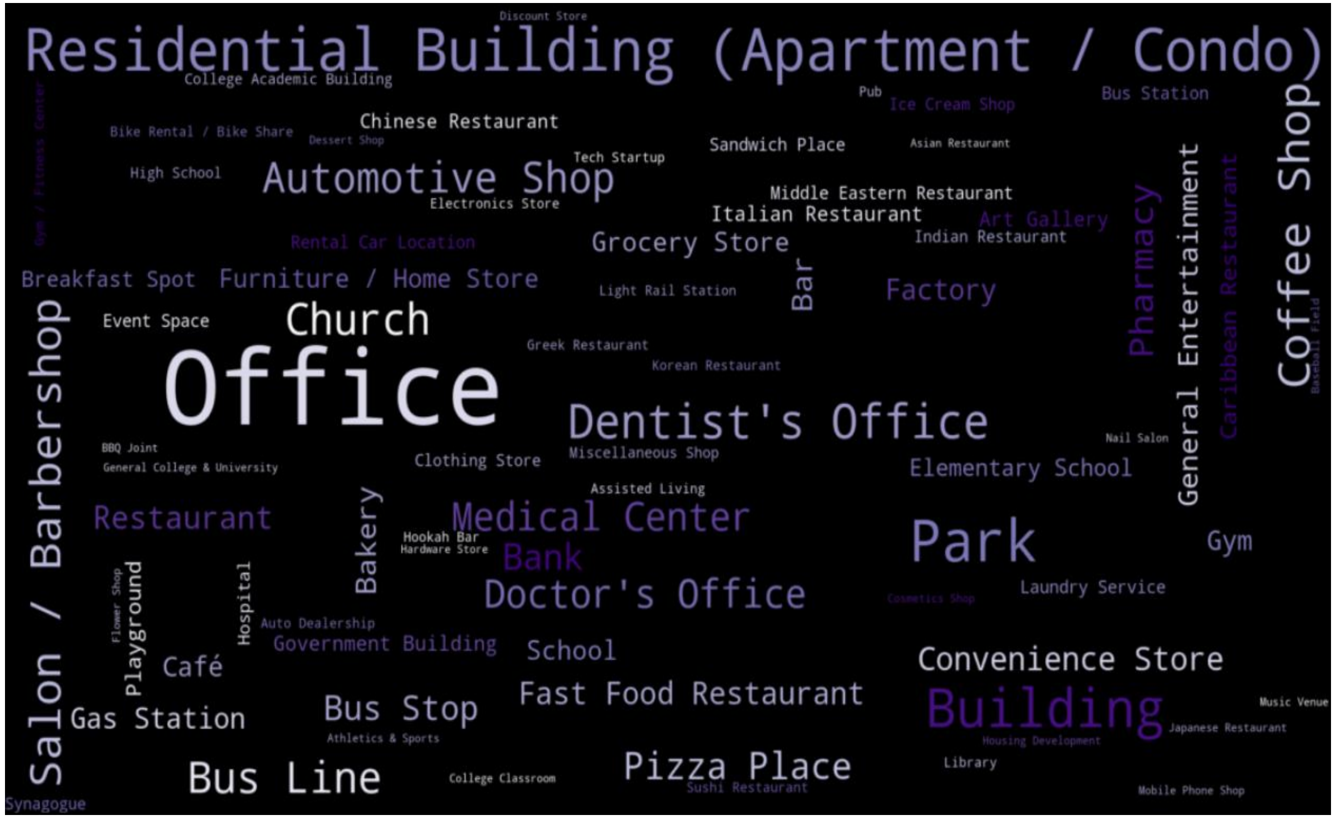
# of Neighborhoods	Population Density
Total: 9 Montreal: 5 Toronto: 4	Low-to-High
Venues Density	**Median Income**
High	Low-to-Med
Popular Venue Categories	
• Entertainment Venues • Bars, Pubs, Cofee Shops, Cafes • Shopping Malls • Historic Sites	• University & College Buildings • Condos & Apartments • Bus lines, Metro Stn • and more…

* Refer to supporting graphs on slides 17 and 18

NEIGHBORHOOD CLUSTERS

Cluster 4 [Purple] – Residential and some Industrial Areas | City Suburbs

Most Popular Venue Categories



Characteristics

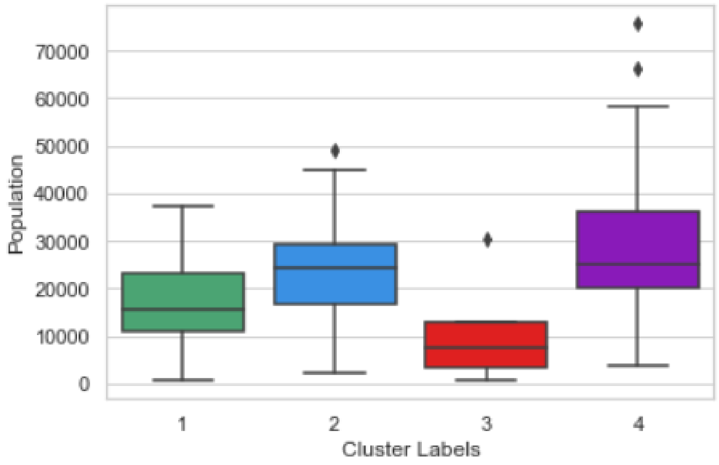
# of Neighborhoods	Population Density
Total: 114 Montreal: 58 Toronto: 56	Low-to-Med
Venues Density	**Median Income**
Low	Med
Popular Venue Categories	
• Offices • Healthcare Venues • Grocery Stores • Factories	• Educational Centers • Condos & Apartments • Assorted Ethnic Restaurants.

* Refer to supporting graphs on slides 17 and 18

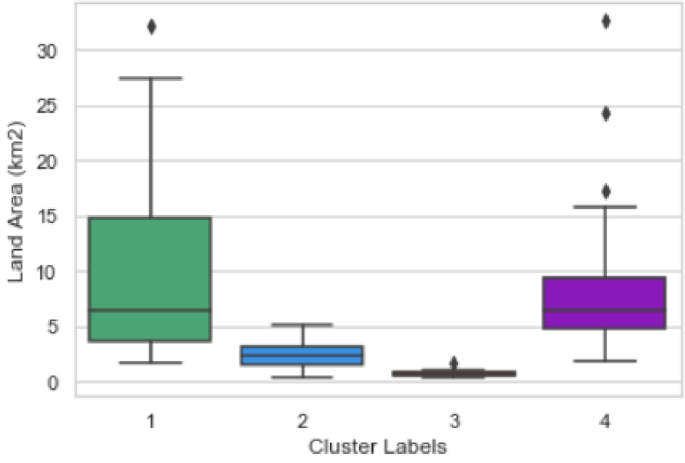
SUPPORTING GRAPHS

Neighborhood Clusters

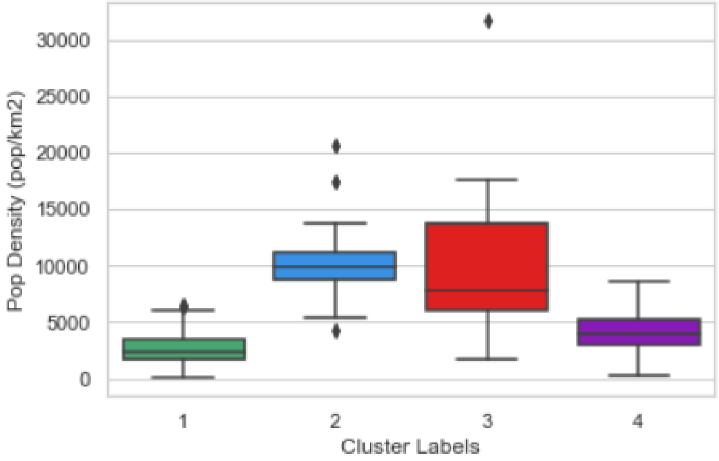
Population Distribution



Land Area (km2) Distribution



Population Density Distribution

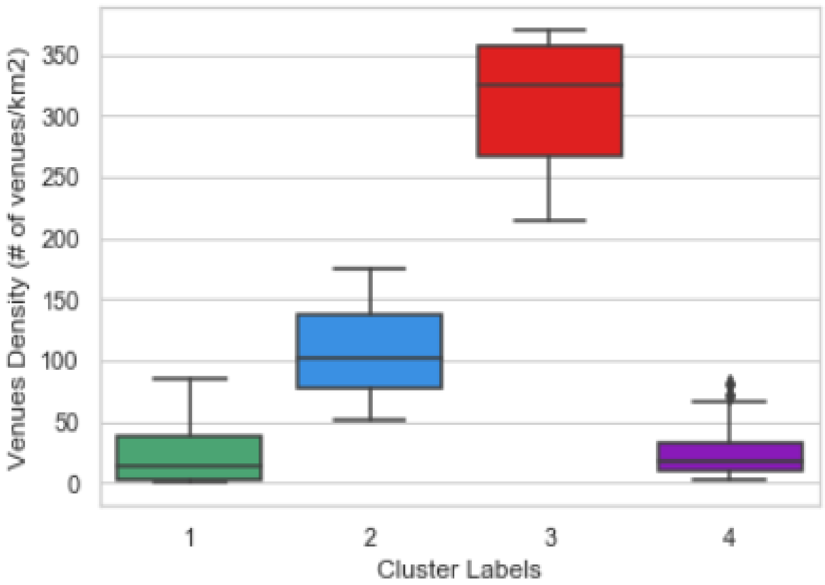


|Cluster 1 = **Green** | Cluster 2 = Blue | Cluster 3 = **Red** | Cluster 4 = **Purple**|

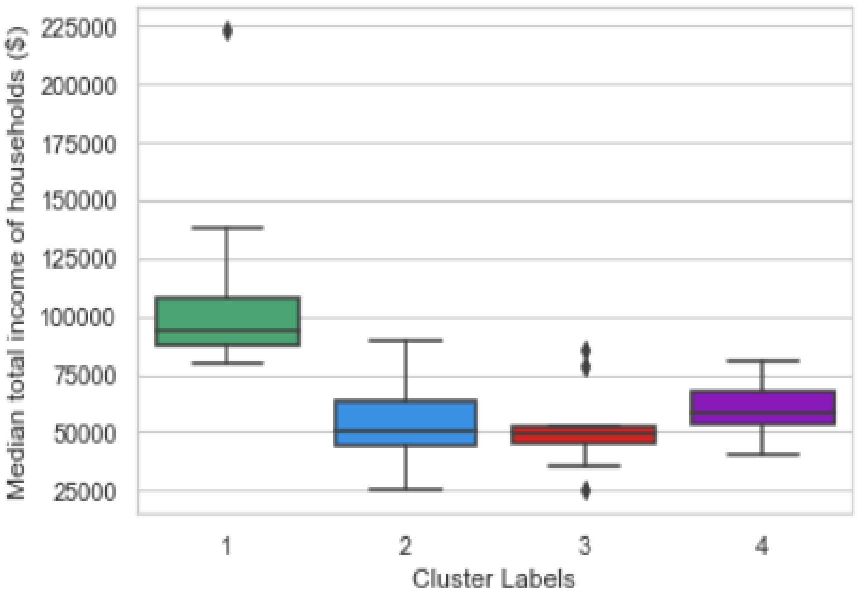
SUPPORTING GRAPHS

Neighborhood Clusters

Venues Density Distribution



Median Income of Households Distribution



| Cluster 1 = **Green** | Cluster 2 = Blue | Cluster 3 = **Red** | Cluster 4 = **Purple** |

CONCLUSION

- In this project I used the k-means clustering algorithm.
- This method helped me identify and group into clusters neighborhoods, from the city of Toronto and the City of Montreal, that share similar venues composition and certain census attributes such as population density, and median income of households.
- I found that grouping neighborhoods into 4 clusters explains most of the variation observed in these two cities.
- Based on my findings, I categorized my clusters into cluster 1 (green) residential and commercial areas mid-town, cluster 2 (blue) residential and commercial areas downtown, cluster 3 (red) commercial area downtown, cluster 4 (purple) residential and some industrial areas city suburbs.
- Finally, there are many ways in which I could have approach this problem. But my results show the power of thinking outside of the box with unsupervised methods like k-means.

FUTURE DIRECTIONS

- This study is meant to be carried out on the initial phases when considering a new location for your business.
- At a subsequent stage, a more business-specific analysis will have to be done.
- This report, however, will help make better-informed decisions in the final stages of your search.
- For example:
  - You want to expand your chain of vegan restaurants. After unfolding a more business-specific (vegan restaurants) analysis, you have located several pocket areas of low restaurant density (also areas of low vegan-restaurant density) across the target city (e.g. Montreal) that are at an adequate distance from the city center.
  - With the results of my study on hand, stakeholders can now zero-in on that optimal location for your vegan restaurant.
  - They can use this information to evaluate each of the tentative locations based on the attractiveness of the neighborhoods these are in; proximity to parks, water, or major roads; presence of venues that trigger the use of vegan restaurants such as a variety of entertainment venues, shopping malls, gyms, and more; real estate availability; social and economic dynamics of every neighborhood, which may include population density, median total income of households, and more.

THANK YOU