The Comparison of New York City and Toronto

Amanda Volkamer

1. Introduction

Which city is the best city to live in? This project will compare two cities, Toronto and New York City. Both cities are very diverse and are the financial capitals of their respective countries. When the two are compared, which one will be considered the best city to live in? For someone who is looking between New York City and Toronto for a job relocation, for example, this is an important question. Some things to consider for finding a new home are safety, housing market, and amenities such as restaurants, parks, and entertainment. And within both cities, which neighborhood is the best to live in? This project aims to provide insight and answer these questions based on the parameters

This project will take into account crime, housing, and venues when comparing the similarities between the two cities. These are important factors to consider when making the decision where to move. Safety and affordability are the vital decision makers for relocation. Do you feel safe in this neighborhood? Can you afford to live in this neighborhood? And how similar will it be to where I live now? Or I really want to live near a park or by lots of restaurants, where is the best place for me?

2. Data

In order to compare New York City and Toronto, I need to pull the data such as the geographical data that includes neighborhoods and boroughs, crime data, housing data, as well as venue data collected from Foursquare. The geographical data will need to include neighborhoods, boroughs, and postal codes for both cities in order to use maps and segment into neighborhoods for the rest of the data for this project.

Sources for Geographical Data

- New York City (JSON): https://cocl.us/new_york_dataset
- Toronto: https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=945633050

The crime data needs to be comparable to one another and because New York City and Toronto will have different sources, I want to make sure the crime data for both cities include similar crimes and are from the year 2019. New York City has a comprehensive master dataset that categorize the crimes in subsets. Toronto, on the other hand, has separate datasets by crime category so I will have to combine the Homicide and MCI data into one dataset in order to compare to the New York City Crime Data. I am choosing the year 2019 because that is the most recent year with the complete data which can later be comparable to the housing data and venues, which are current or recent data.

Sources for Crime Data

- New York City: https://data.cityofnewyork.us/Public-Safety/NYC-crime/qb7u-rbmr
- New York City GeoJSON: https://raw.githubusercontent.com/dwillis/nyc-maps/master/police_precincts.geojson
- Toronto: https://data.torontopolice.on.ca/
- Toronto GeoJSON: https://raw.githubusercontent.com/jasonicarter/toronto-geojson/master/toronto_crs84.geojson

The housing data also has to be comparable to each other. The housing data has to have housing prices or values and location information. For the New York City Housing Data, I will scrape the information from the city-data site and use the uszipcode library to pull the zip codes for the dataframe. The Toronto Housing Data is comprised from a .csv file that includes the list prices of properties. To make the Toronto Housing Data comparable to New York City Housing Data, I will set the minimum and maximum values after converting the minimum and maximum of the New York City Housing Data from the US Dollar to the Canadian Dollar. Also, only the data for the Toronto proper will be used.

Sources for Housing Data

- New York City: http://www.city-data.com/zipmaps/New-York-New-York.html
- Toronto: https://www.kaggle.com/mnabaee/ontarioproperties

Venues Data will be pulled using the Foursquare API. The Foursquare API provides numerous venue categories. It has one of the largest databases of over 105 million places.

2.1 Data Cleaning

All null values will be excluded from the data used for the analysis. I will my discretion for determining unnecessary data to exclude from my analysis based on the usefulness and purpose of the data and this project.

3. Methodology

3.1 Maps

3.1.1 Maps of the Neighborhoods in Manhattan and Toronto

For the geographical data, I will use Folium maps to show the neighborhoods in Manhattan, New York City and in Downtown Toronto, Toronto. To do so, I will load and clean both New York City and Toronto data before joining both data together in a combined data.

3.1.2 Choropleth Maps

To showcase the crime data for New York City and Toronto, I will use choropleth maps using Folium to show the density of reported crimes by neighborhoods in New York City and Toronto. To start, I will count the number of occurrences of crime that occurred in each neighborhood to show the density of the crimes by precinct number or neighborhood in New York City and Toronto.

3.2 K Means Clustering

I will use the K Means Clustering approach due to its simplicity and ability to use similarity to compare. K Means is a clustering algorithm which searches clusters within the data and aim to minimize the data dispersion for each cluster. Therefore, each cluster represents a set of data with a similar pattern.

The Housing and Venues Data will undergo this unsupervised learning to find similarities between the two cities. To prepare each section, I will need to one hot encode the categorical values in both New York City and Toronto Housing and Venues data. Then I will use the Elbow method. The Elbow method is a chart that compares error vs number of clusters is done. The elbow of the line indicates the optimal value of k. The value of k is the number of clusters that is needed to accurately represent the data. Once, the value of k is determined, I will then run the cluster analysis.

3.2.1 Housing Data

The clusters for both cities will represent a range of housing values and affordability. I will then label each cluster in a range from low to high budget.

3.2.2 Venue Data

To compare the venues in the neighborhoods of Manhattan and Downtown Toronto, I will use K Means clustering. To do this, I will use the GET function to pull all the venues based in each borough from the Foursquare API and append them into the merged dataframe I will use for the geographical data for Manhattan and downtown Toronto. Next, I will analyze each neighborhood in each borough. Then I will group the neighborhoods by taking the mean of frequency of occurrence for venue type and append them into a new dataframe to display the top 10 venues for each neighborhood.

3.3 Combine and Overlap the Data

I intend to overlay the K Means Cluster Maps of the Housing Data on the Crime Choropleth Maps for both New York City and Toronto. This will show some interesting insights about the characteristics of the neighborhoods and potentially some correlational relationships between crime and housing prices.

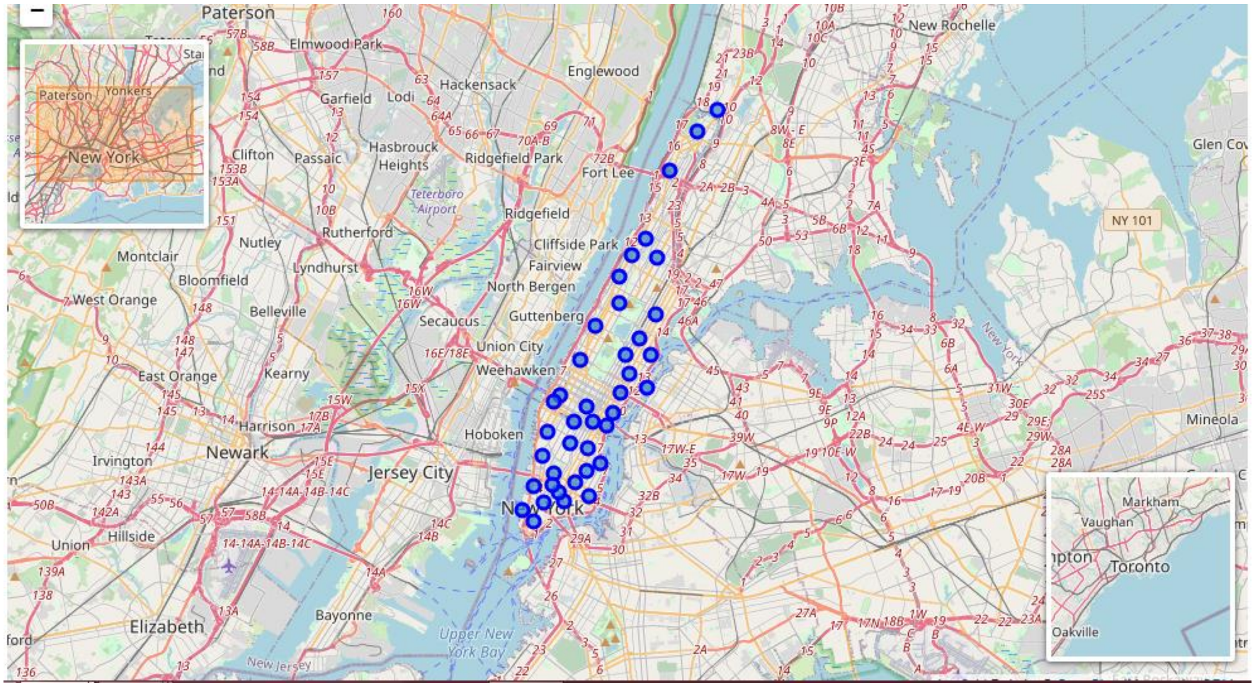
4. Results

4.1 Maps

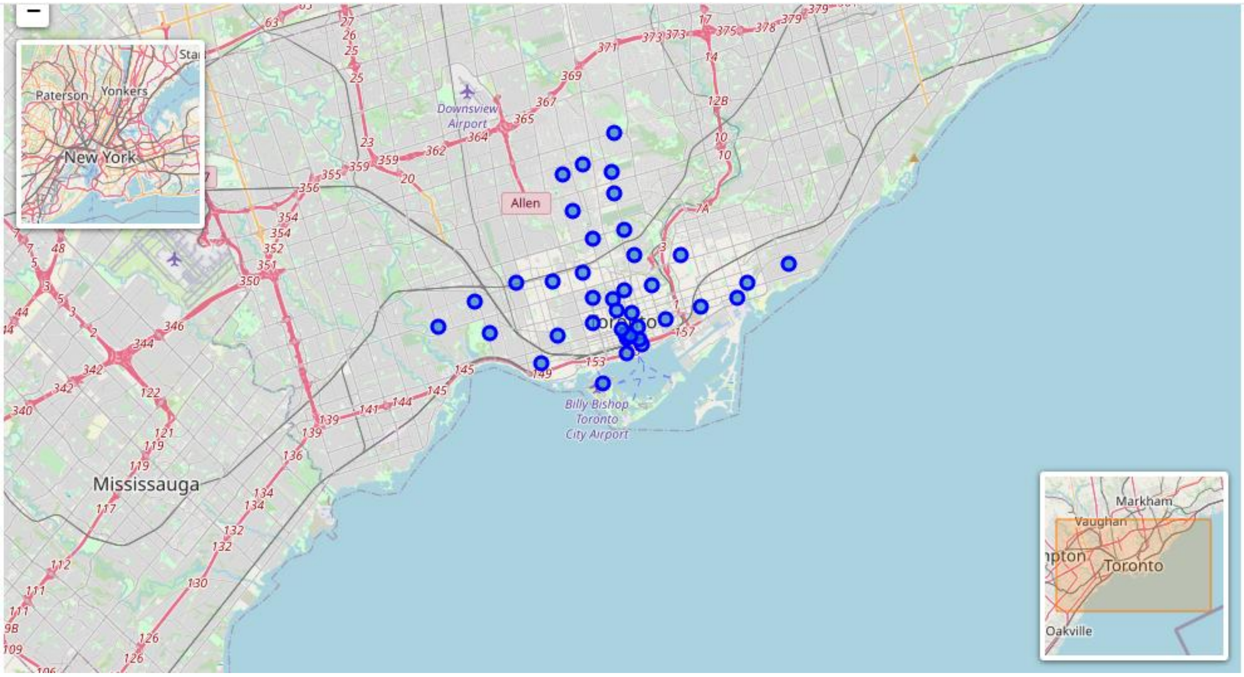
4.1.1 Maps of the Neighborhoods of Manhattan and Toronto

To begin this project, it's necessary to look at what the neighborhoods look like in New York City and Toronto.

Manhattan, New York City has 40 neighborhoods.



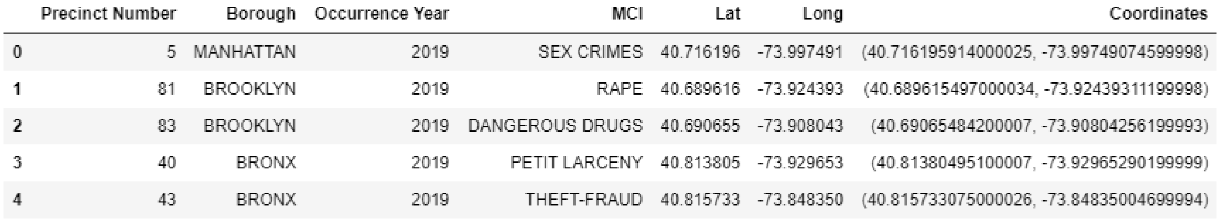
The city of Toronto has 39 neighborhoods.



4.1.2 Choropleth Heat Maps - Crime Data

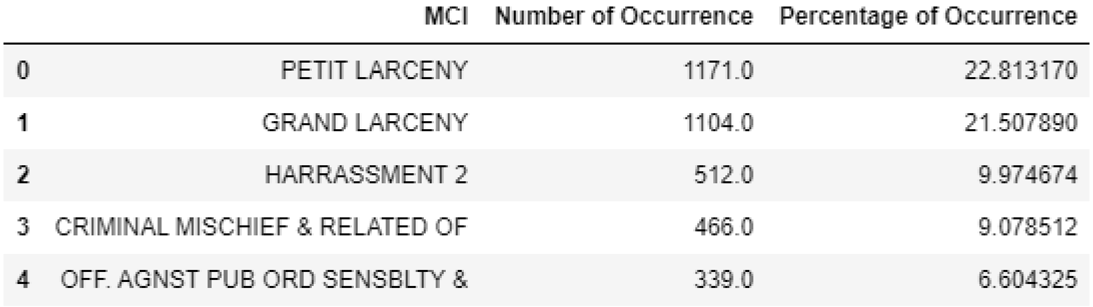
4.1.2.1 New York City Crime

The New York City Crime Data .csv file includes every complaint incident recorded with the NYPD dating back consistently to 2000. The file includes 57 crime descriptions such as assault, theft, and murder. After preprocessing the data and pulling only the data for the year 2019, this is what my dataframe frame looked like.



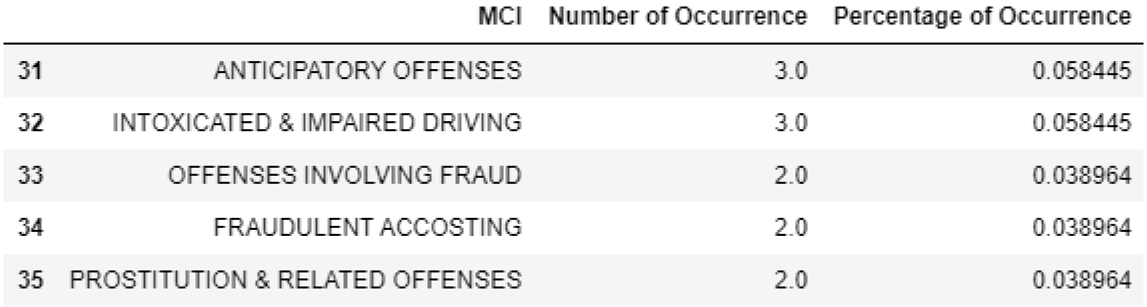
	Precinct Number	Borough	Occurrence Year	MCI	Lat	Long	Coordinates
0	5	MANHATTAN	2019	SEX CRIMES	40.716196	-73.997491	(40.716195914000025, -73.99749074599998)
1	81	BROOKLYN	2019	RAPE	40.689616	-73.924393	(40.689615497000034, -73.92439311199998)
2	83	BROOKLYN	2019	DANGEROUS DRUGS	40.690655	-73.908043	(40.69065484200007, -73.90804256199993)
3	40	BRONX	2019	PETIT LARCENY	40.813805	-73.929653	(40.81380495100007, -73.92965290199999)
4	43	BRONX	2019	THEFT-FRAUD	40.815733	-73.848350	(40.815733075000026, -73.84835004699994)

I thought it would be interesting to see what the most occurring crime is. So, I calculated the number of incidences by crime and took the percentage from the total incidences reported in the year of 2019 in New York City. The first image shows the five highest percentages.



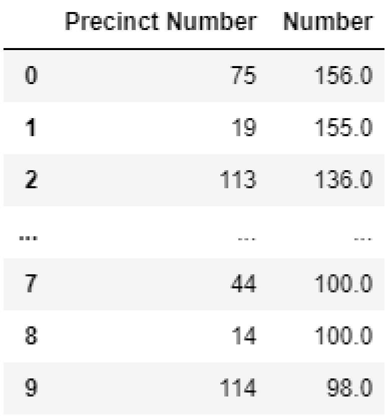
	MCI	Number of Occurrence	Percentage of Occurrence
0	PETIT LARCENY	1171.0	22.813170
1	GRAND LARCENY	1104.0	21.507890
2	HARRASSMENT 2	512.0	9.974674
3	CRIMINAL MISCHIEF & RELATED OF	466.0	9.078512
4	OFF. AGNST PUB ORD SENSBLTY &	339.0	6.604325

And the next image shows the fives lowest percentages.



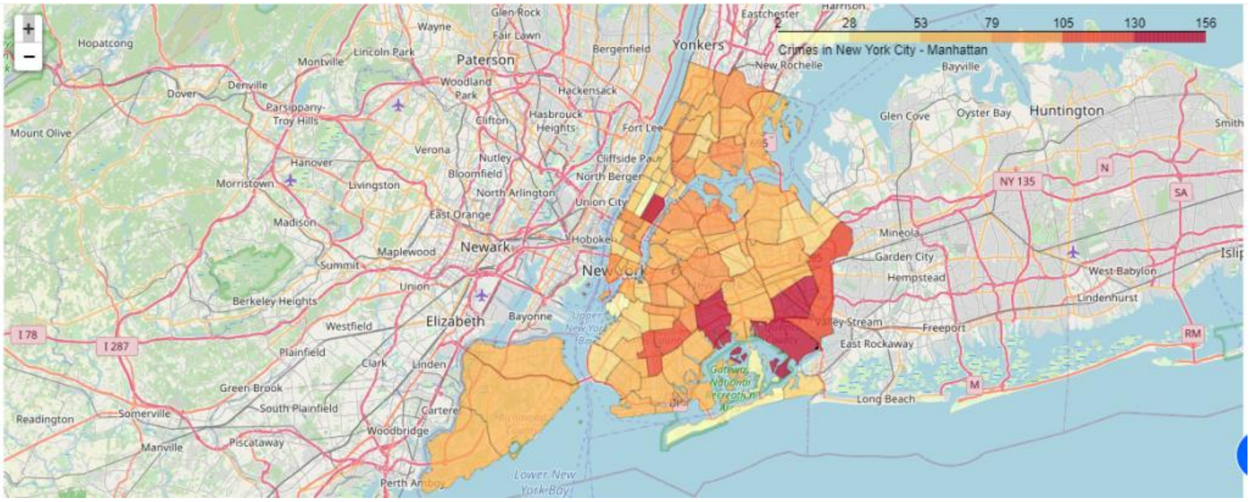
	MCI	Number of Occurrence	Percentage of Occurrence
31	ANTICIPATORY OFFENSES	3.0	0.058445
32	INTOXICATED & IMPAIRED DRIVING	3.0	0.058445
33	OFFENSES INVOLVING FRAUD	2.0	0.038964
34	FRAUDULENT ACCOSTING	2.0	0.038964
35	PROSTITUTION & RELATED OFFENSES	2.0	0.038964

In order to look at how the crime is distributed in New York City, I counted the number of incidences for each Precinct Number.



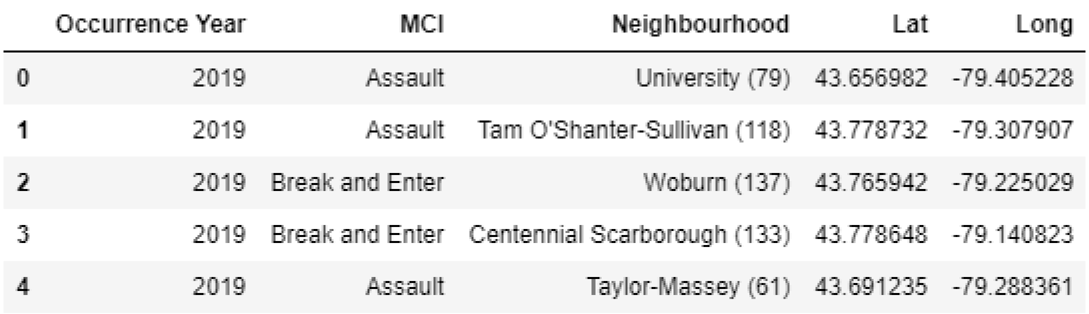
	Precinct Number	Number
0	75	156.0
1	19	155.0
2	113	136.0
...	...	...
7	44	100.0
8	14	100.0
9	114	98.0

I used the frequency count to create a Choropleth Map. The lighter colors indicate a lower frequency of crime and the darker colors indicate a higher frequency of crime. This map shows which neighborhoods have higher reported crimes in the year 2019 in New York City.



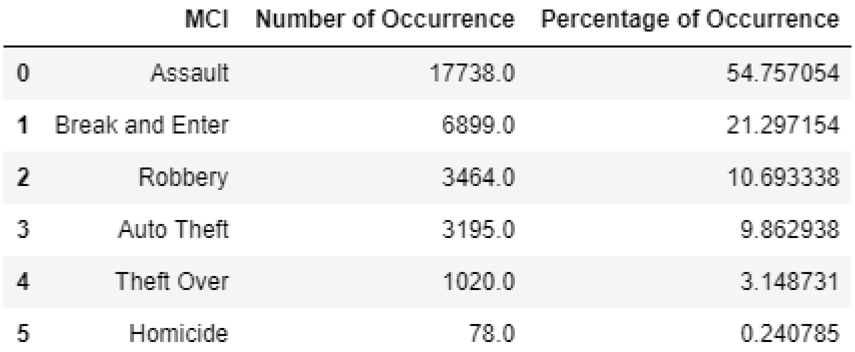
4.1.2.2 Toronto Crime

The Toronto Homicide Data and the Toronto MCI Data were preprocessed and then combined into a joint dataframe as seen below. The Toronto Crime Data includes 8 crime categories. This dataframe includes only the data from the year 2019.



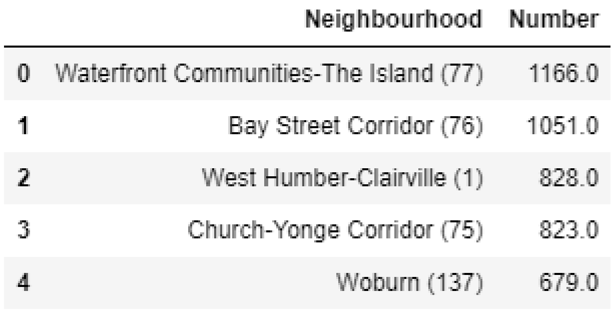
	Occurrence Year	MCI	Neighbourhood	Lat	Long
0	2019	Assault	University (79)	43.656982	-79.405228
1	2019	Assault	Tam O'Shanter-Sullivan (118)	43.778732	-79.307907
2	2019	Break and Enter	Woburn (137)	43.765942	-79.225029
3	2019	Break and Enter	Centennial Scarborough (133)	43.778648	-79.140823
4	2019	Assault	Taylor-Massey (61)	43.691235	-79.288361

I also wanted to see what the most occurring crime in Toronto was. I calculated the number of incidences by crime and took the percentage from the total incidences reported in the year of 2019 in Toronto.



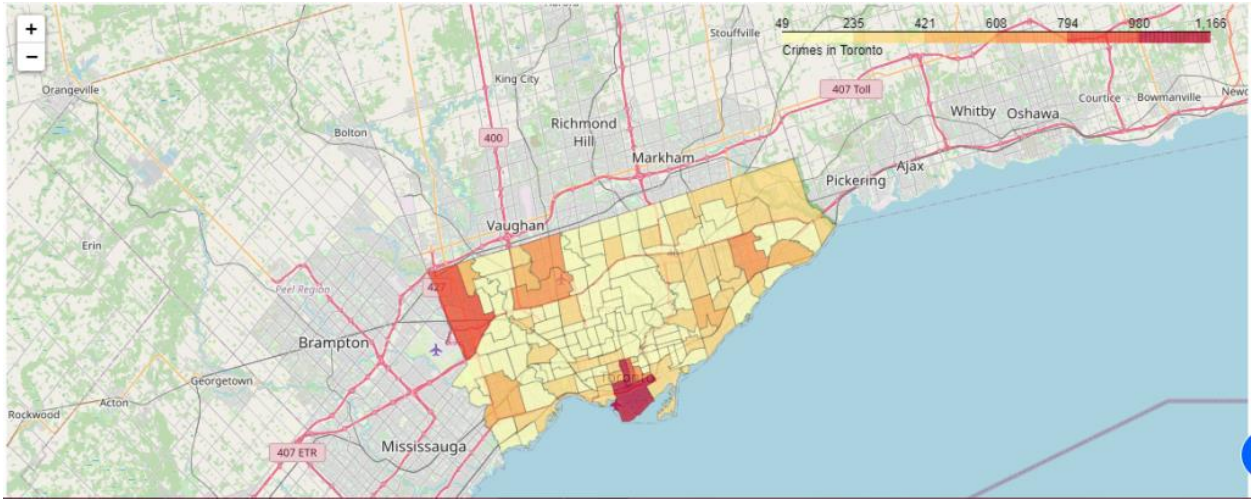
	MCI	Number of Occurrence	Percentage of Occurrence
0	Assault	17738.0	54.757054
1	Break and Enter	6899.0	21.297154
2	Robbery	3464.0	10.693338
3	Auto Theft	3195.0	9.862938
4	Theft Over	1020.0	3.148731
5	Homicide	78.0	0.240785

Like the New York Crime Data, I counted the number of incidences that occurred for each neighborhood in Toronto.



	Neighbourhood	Number
0	Waterfront Communities-The Island (77)	1166.0
1	Bay Street Corridor (76)	1051.0
2	West Humber-Clairville (1)	828.0
3	Church-Yonge Corridor (75)	823.0
4	Woburn (137)	679.0

I used the frequency count to create a Toronto Crime Choropleth Map. The lighter colors indicate a lower frequency of crime and the darker colors indicate a higher frequency of crime. This map shows which neighborhoods have higher reported crimes in the year 2019 in Toronto. As seen here, there is only one neighborhood that has the highest density of crime as compared to the other neighborhoods with lower density.

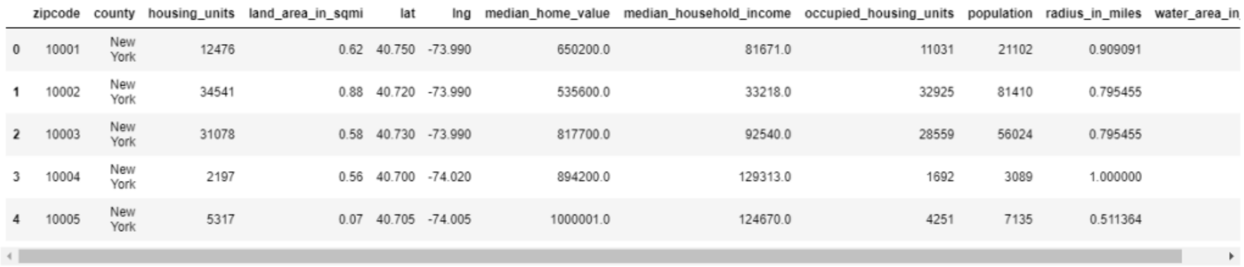


4.2 K Means Clustering

4.2.1 Housing Data

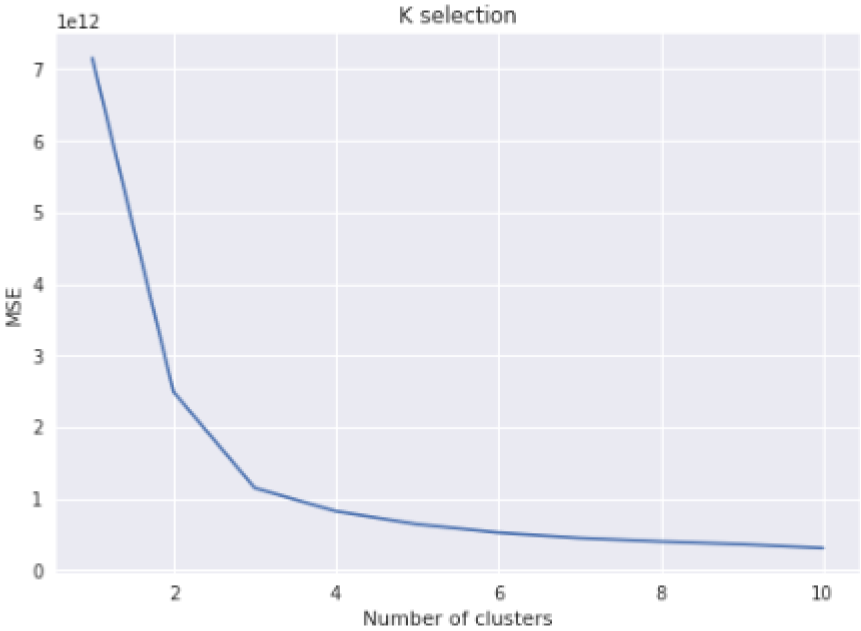
4.2.1.1 New York City

I pulled the housing data for New York City by scraping it from this link. Then, I pulled further information about the zip codes by using the uszipcodes library. After preprocessing the data, I got this dataframe.

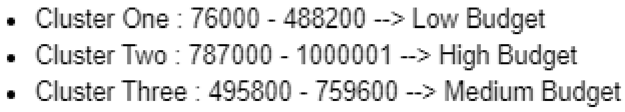


	zipcode	county	housing_units	land_area_in_sqmi	lat	lng	median_home_value	median_household_income	occupied_housing_units	population	radius_in_miles	water_area_in
0	10001	New York	12476	0.62	40.750	-73.990	650200.0	81671.0	11031	21102	0.909091	
1	10002	New York	34541	0.88	40.720	-73.990	535600.0	33218.0	32925	81410	0.795455	
2	10003	New York	31078	0.58	40.730	-73.990	817700.0	92540.0	28559	56024	0.795455	
3	10004	New York	2197	0.56	40.700	-74.020	894200.0	129313.0	1692	3089	1.000000	
4	10005	New York	5317	0.07	40.705	-74.005	1000001.0	124670.0	4251	7135	0.511364	

Now, I was ready to start the cluster analysis. In order to determine the best value of k, I used the elbow method. The mean squared error (MSE) is plotted with the number of clusters. The elbow in this graph indicates that the optimal value of k is 3.

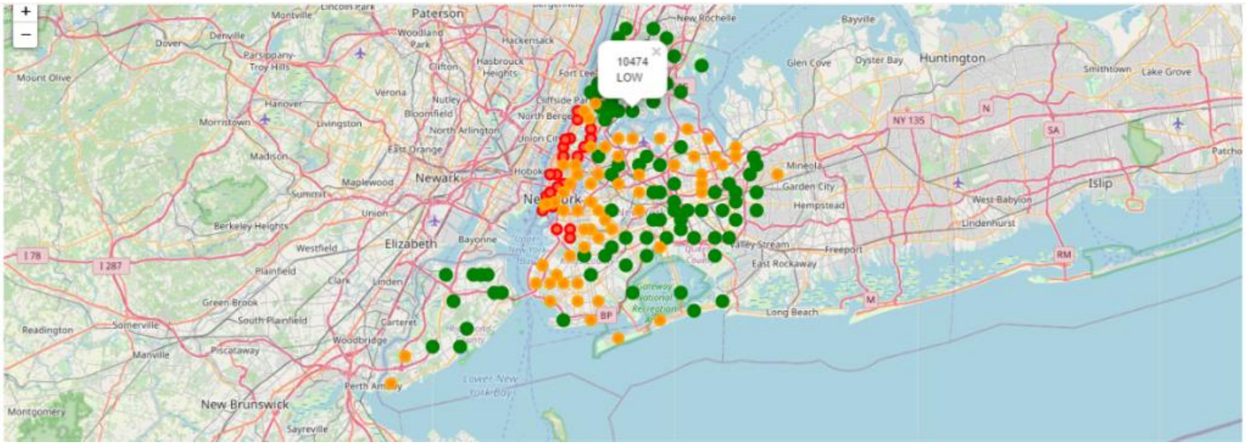


After running the K Means cluster with the k value of 3, I was able to see how the housing values in New York City is grouped. So, I labeled each cluster as Low, Medium, or High Budget in terms of comparable affordability.

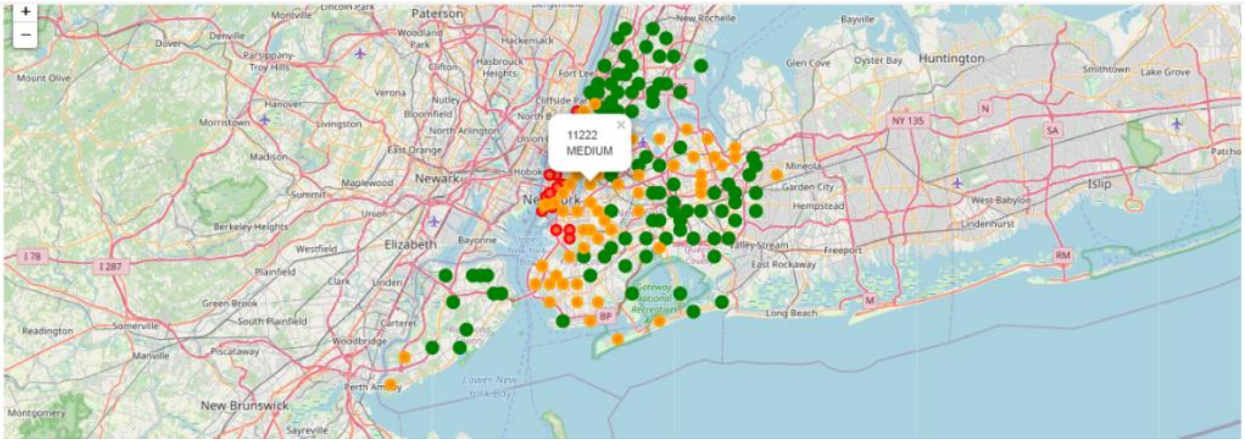


- Cluster One : 76000 - 488200 --> Low Budget
- Cluster Two : 787000 - 1000001 --> High Budget
- Cluster Three : 495800 - 759600 --> Medium Budget

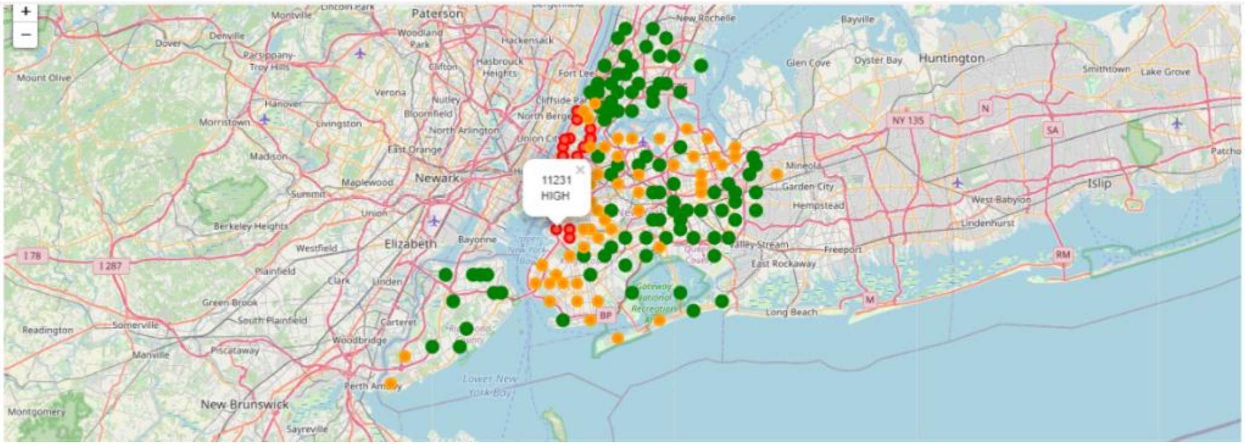
To better visualize the clusters of the housing values in New York City, I created a cluster map. In this image, you can see how the different housing clusters are placed in New York City. The green markers represent the low budget housing.



The yellow markers represent the medium budget housing.



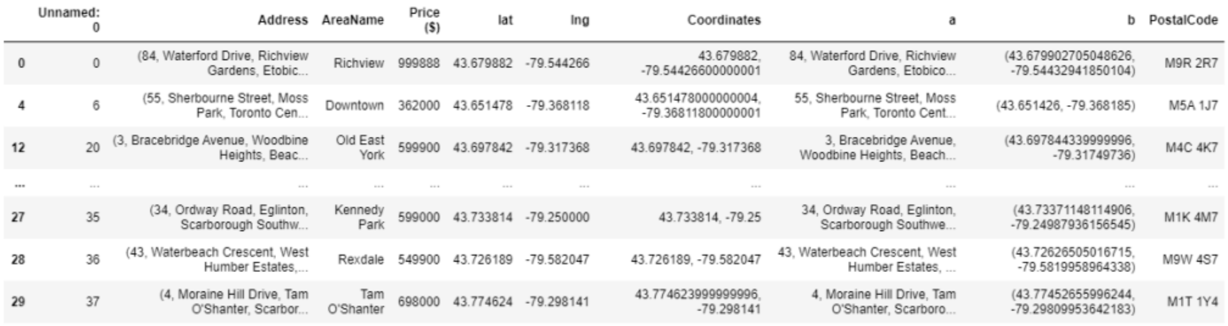
And the red markers represent the high budget housing.



4.2.1.2 Toronto

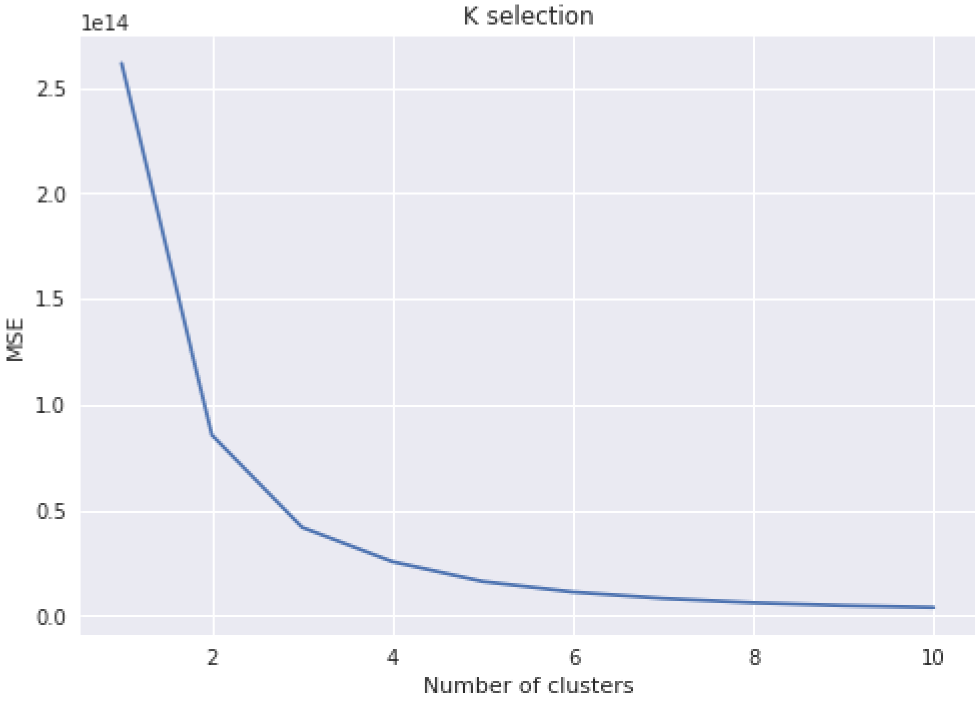
The Toronto Housing data comes from a .csv file. The file included properties values, latitude, and longitude. During my process of cleaning the data, I had to pull only the data within a certain limit for the property values for two reasons. The first reason was some of the property values seemed disproportionate and seemed to be for commercial properties, with is not in the scope of this project. And the second, is that this data set seems to use a different currency, the Canadian Dollar, as compared to the New York City Housing Data that uses the US Dollar. To remedy these complications, I used the minimum and maximum housing values from the New York Data as the range after converting the values to the Canadian Dollar currency.

The addresses included surrounding areas of Toronto, so I pulled only the data for Toronto. Also, the addresses did not provide postal codes, so I reverse geocoded the latitude and longitude values using Nominatim and RateLimiter. After the preprocessing, this is what my dataframe looked like.

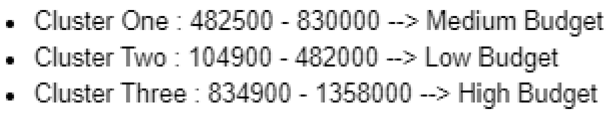


	Unnamed: 0	Address	AreaName	Price ($)	lat	lng	Coordinates	a	b	PostalCode
0	0	(84, Waterford Drive, Richview Gardens, Etobic...	Richview	999888	43.679882	-79.544266	43.679882, -79.54426600000001	84, Waterford Drive, Richview Gardens, Etobico...	(43.679902705048626, -79.54432941850104)	M9R 2R7
4	6	(55, Sherbourne Street, Moss Park, Toronto Cen...	Downtown	362000	43.651478	-79.368118	43.651478000000004, -79.36811800000001	55, Sherbourne Street, Moss Park, Toronto Cent...	(43.651426, -79.368185)	M5A 1J7
12	20	(3, Bracebridge Avenue, Woodbine Heights, Beac...	Old East York	599900	43.697842	-79.317368	43.697842, -79.317368	3, Bracebridge Avenue, Woodbine Heights, Beach...	(43.697844339999996, -79.31749736)	M4C 4K7
...	...	...	...	...	...	...	...	...	...	...
27	35	(34, Ordway Road, Eglinton, Scarborough Southw...	Kennedy Park	599000	43.733814	-79.250000	43.733814, -79.25	34, Ordway Road, Eglinton, Scarborough Southwe...	(43.73371148114906, -79.24987936156545)	M1K 4M7
28	36	(43, Waterbeach Crescent, West Humber Estates,...	Rexdale	549900	43.726189	-79.582047	43.726189, -79.582047	43, Waterbeach Crescent, West Humber Estates, ...	(43.72626505016715, -79.5819958964338)	M9W 4S7
29	37	(4, Moraine Hill Drive, Tam O'Shanter, Scarbor...	Tam O'Shanter	698000	43.774624	-79.298141	43.774623999999996, -79.298141	4, Moraine Hill Drive, Tam O'Shanter, Scarboro...	(43.77452655996244, -79.29809953642183)	M1T 1Y4

Next, I was ready to start the cluster analysis. In order to determine the best value of k, I used the elbow method. The mean squared error (MSE) is plotted with the number of clusters. The elbow in this graph indicates that the optimal value of k is 3.

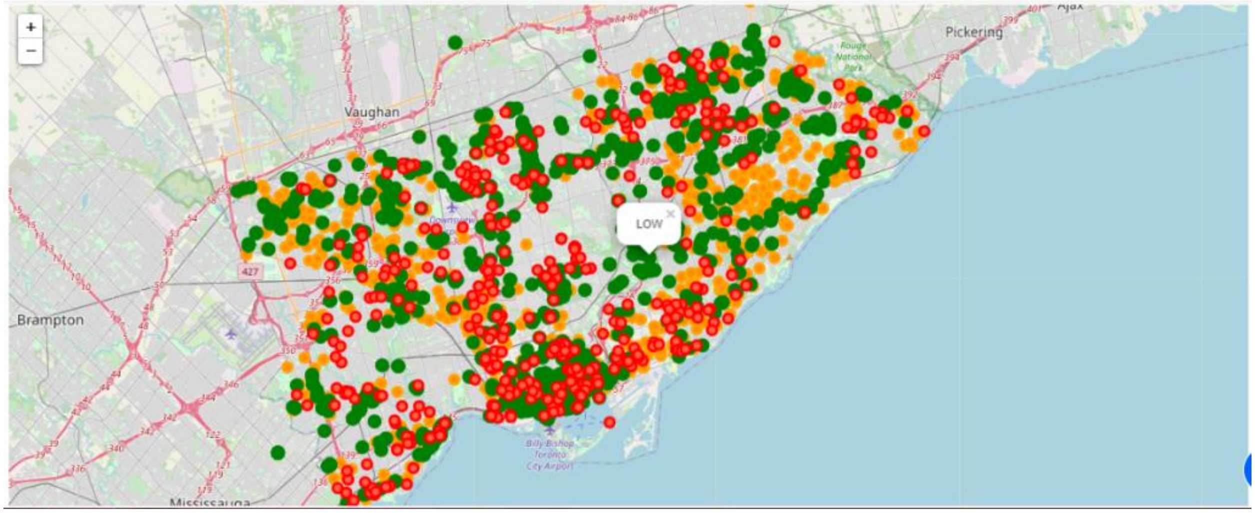


After running the K Means cluster with the k value of 3, I was able to see how the housing values in Toronto is grouped. So, I labeled each cluster as Low, Medium, or High Budget in terms of comparable affordability. Keep in mind the housing values are as Canadian Dollars, not US Dollars in the New York City data.

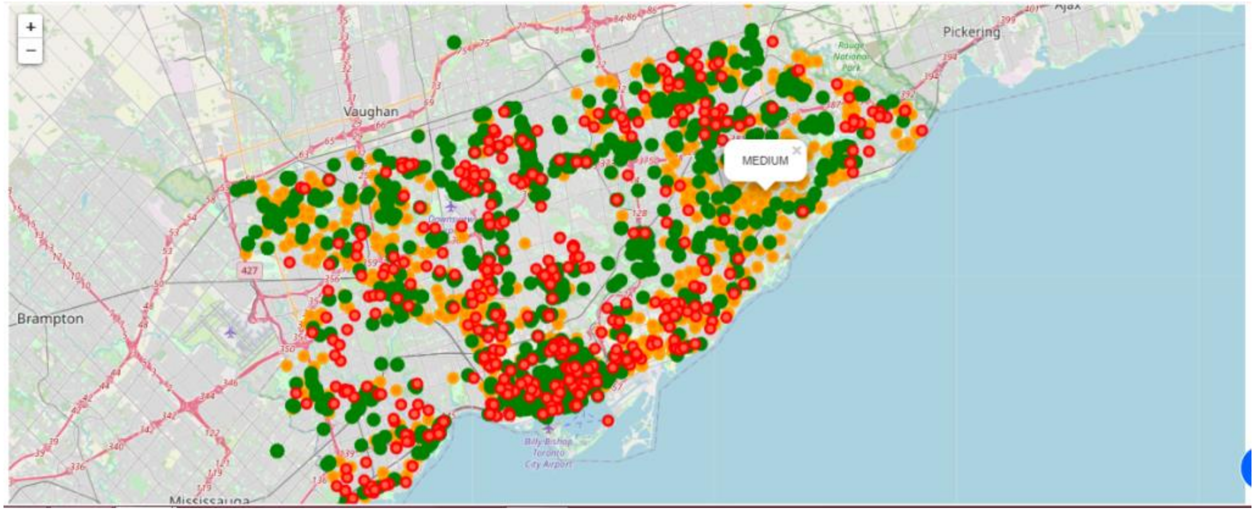


- Cluster One : 482500 - 830000 --> Medium Budget
- Cluster Two : 104900 - 482000 --> Low Budget
- Cluster Three : 834900 - 1358000 --> High Budget

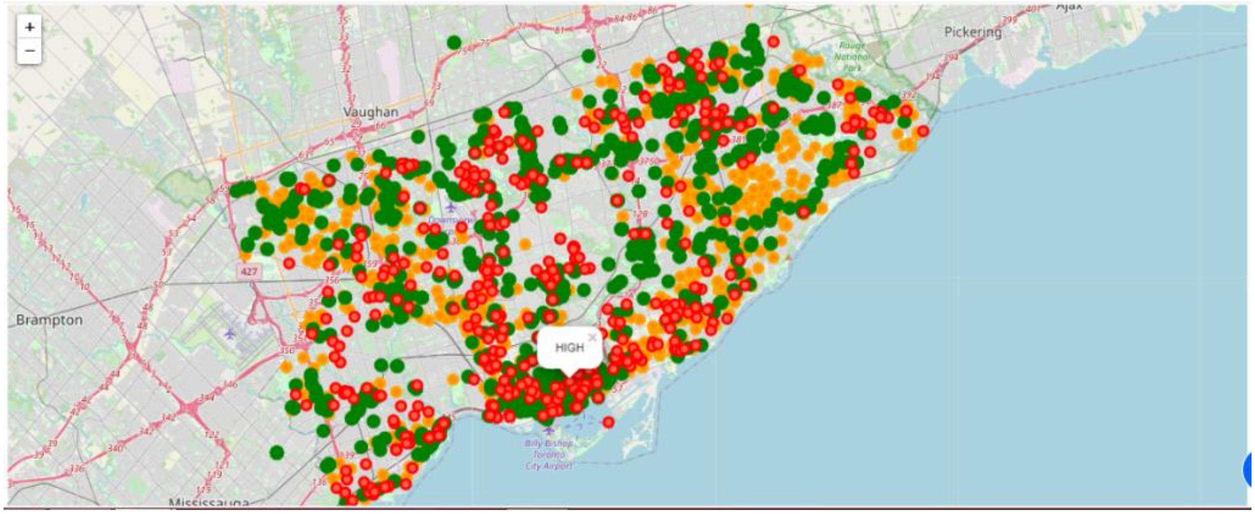
To better visualize the clusters of the housing values in Toronto, I created a cluster map. In this image, you can see how the different housing clusters are placed in Toronto. The green markers represent the low budget housing.



The yellow markers represent the medium budget housing.

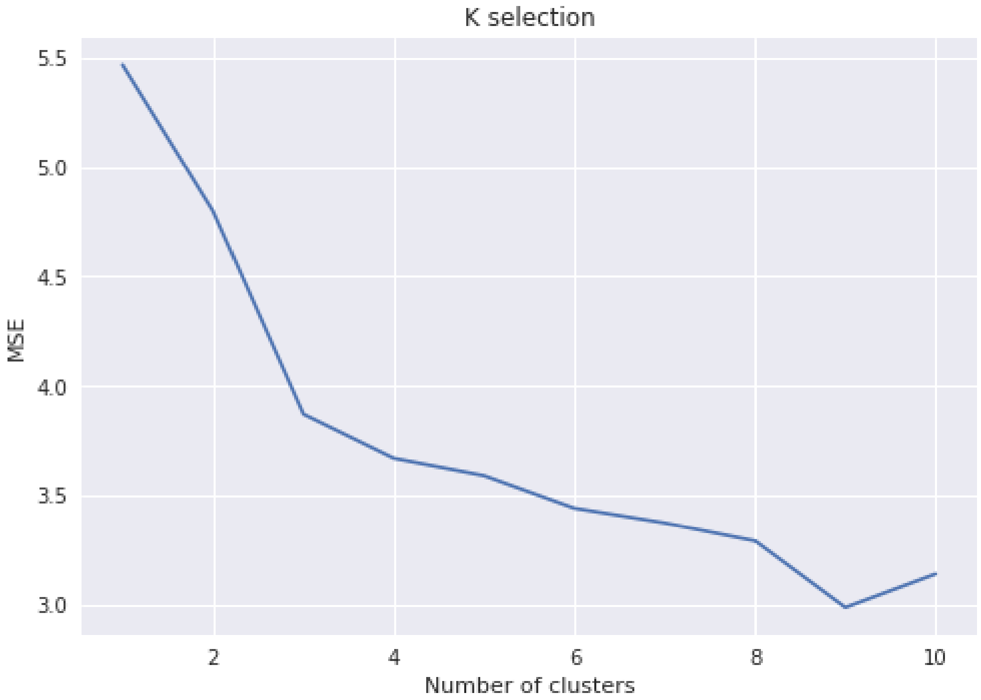


And the red markers represent the high budget housing.

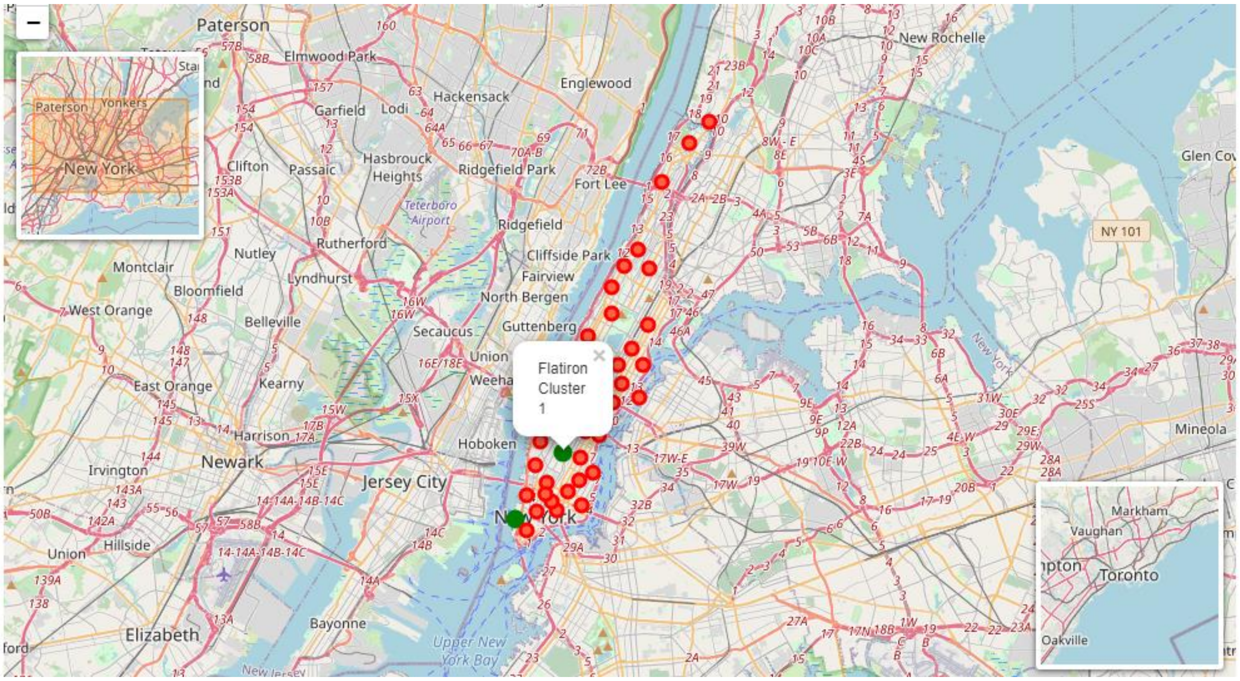


4.2.2 Venue Data

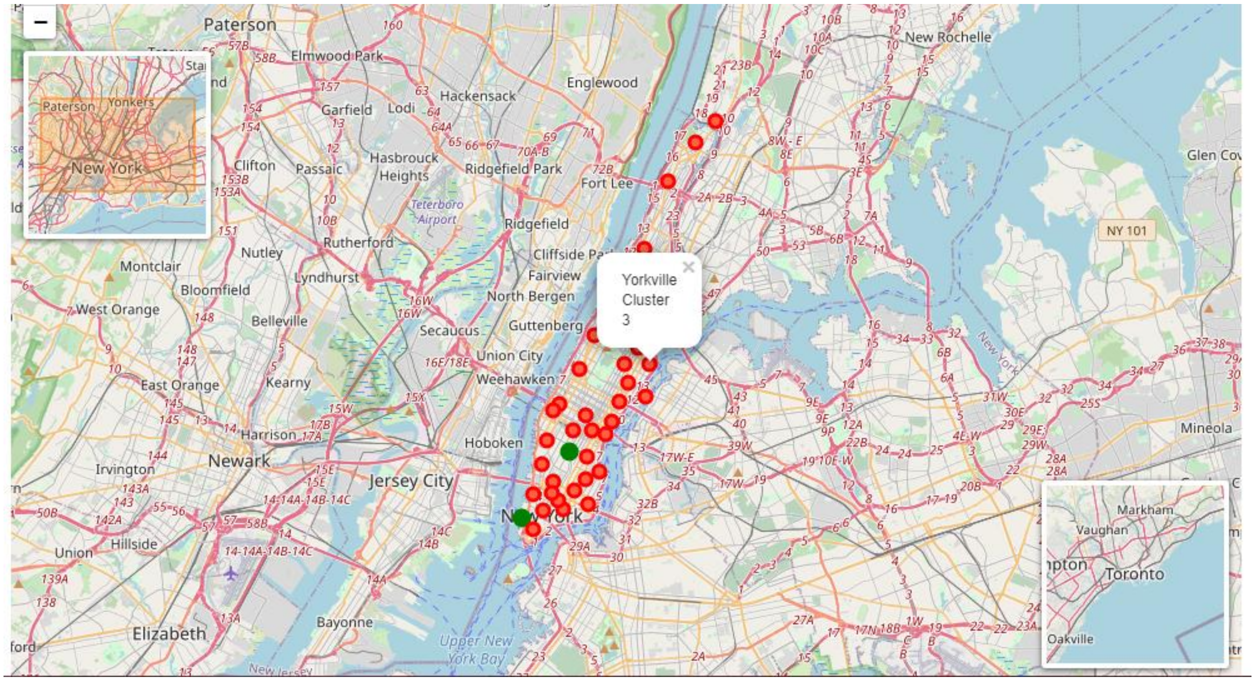
I pulled the venue data for both New York City and Toronto from Foursquare into a combined dataframe. After preprocessing the data, I started the cluster analysis. I used the elbow method to determine the value of k to use. The elbow in this graph indicates that the optimal value of k is 3.



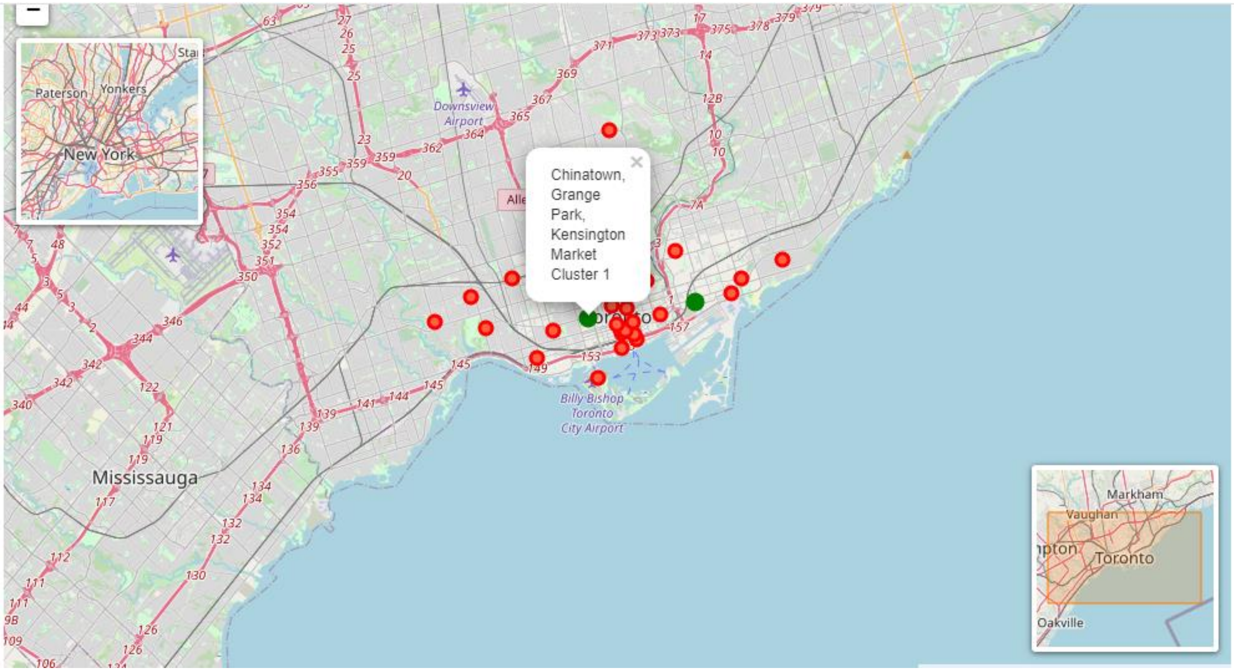
To visualize the venue clusters, I made cluster maps for both Manhattan, NY and Toronto. Here you can see the map of Manhattan with Cluster One pointed out and represented with two dots.



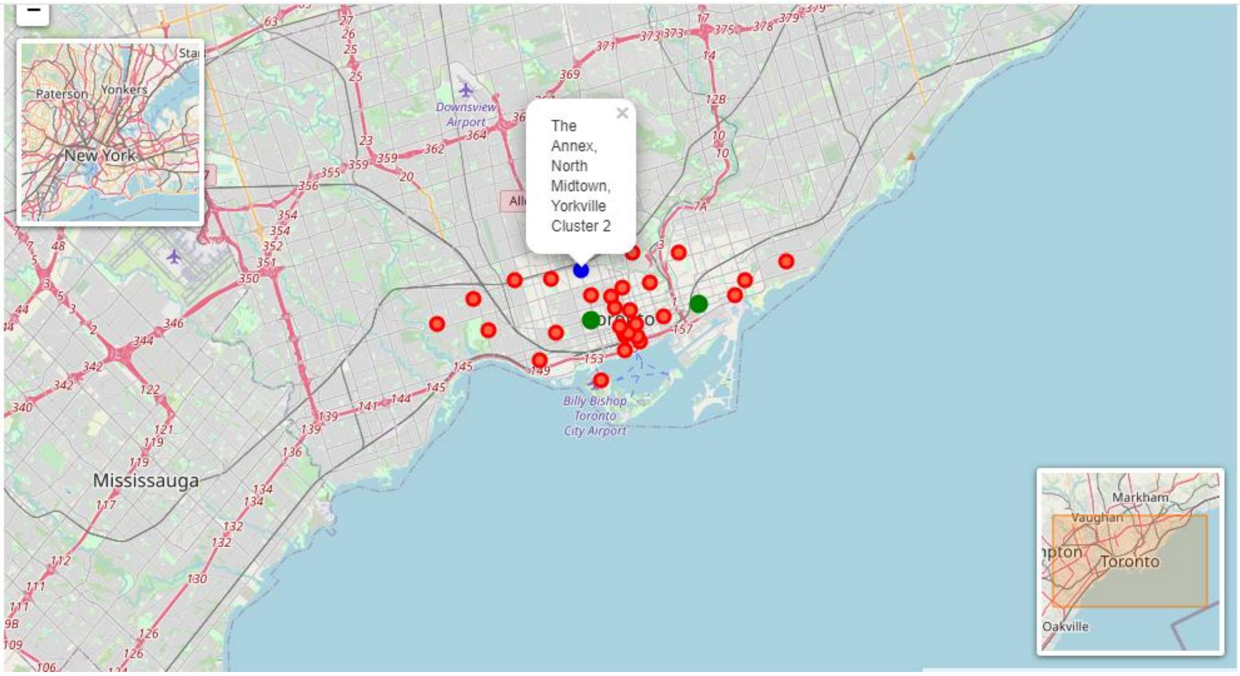
And the Cluster Three accounts for the majority of Manhattan.



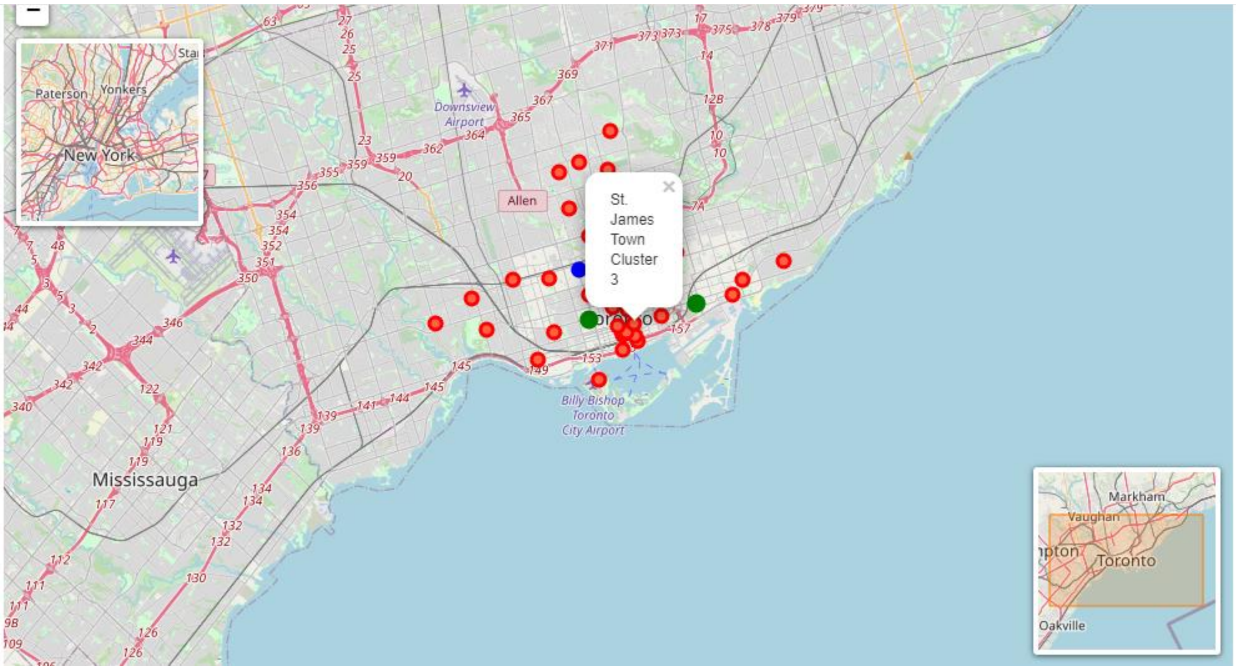
The images below show a map of Toronto. There are two dots for Cluster One in Toronto.


Chinatown, Grange Park, Kensington Market Cluster 1
Downsview Airport
Billy Bishop Toronto City Airport
Mississauga
Toronto
New York
Paterson
Yonkers
Markham
Vaughan
Oakville

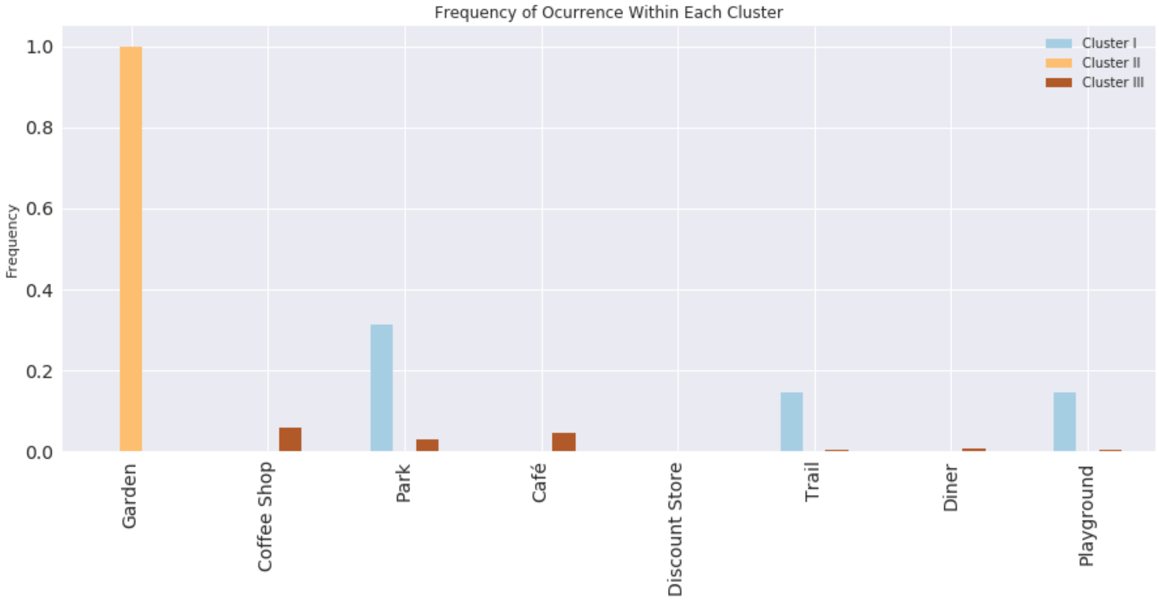
There is only one dot among the maps of New York City and Toronto that belongs to Cluster Two. It seems there is something unique in this spot in Toronto.



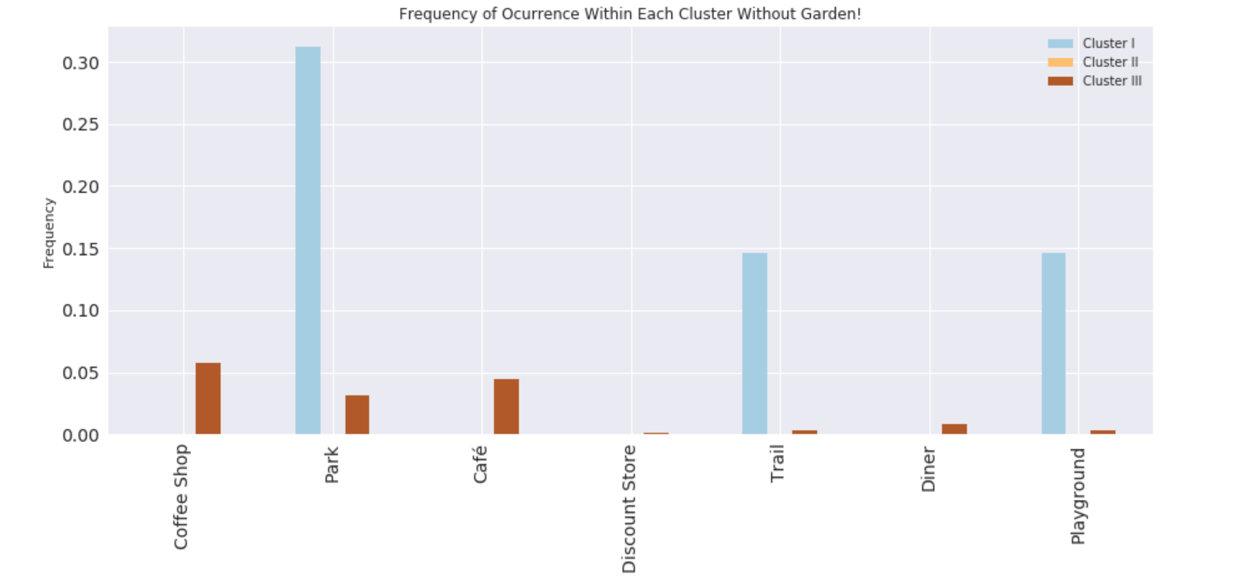
The majority of the dots belong to Cluster Three.



I wanted to see how the clusters were categorized. The bar chart shows the features with higher frequency in the k centroid of the K Means cluster analysis. As seen below, Cluster Two is overwhelmingly representative of the Garden category.



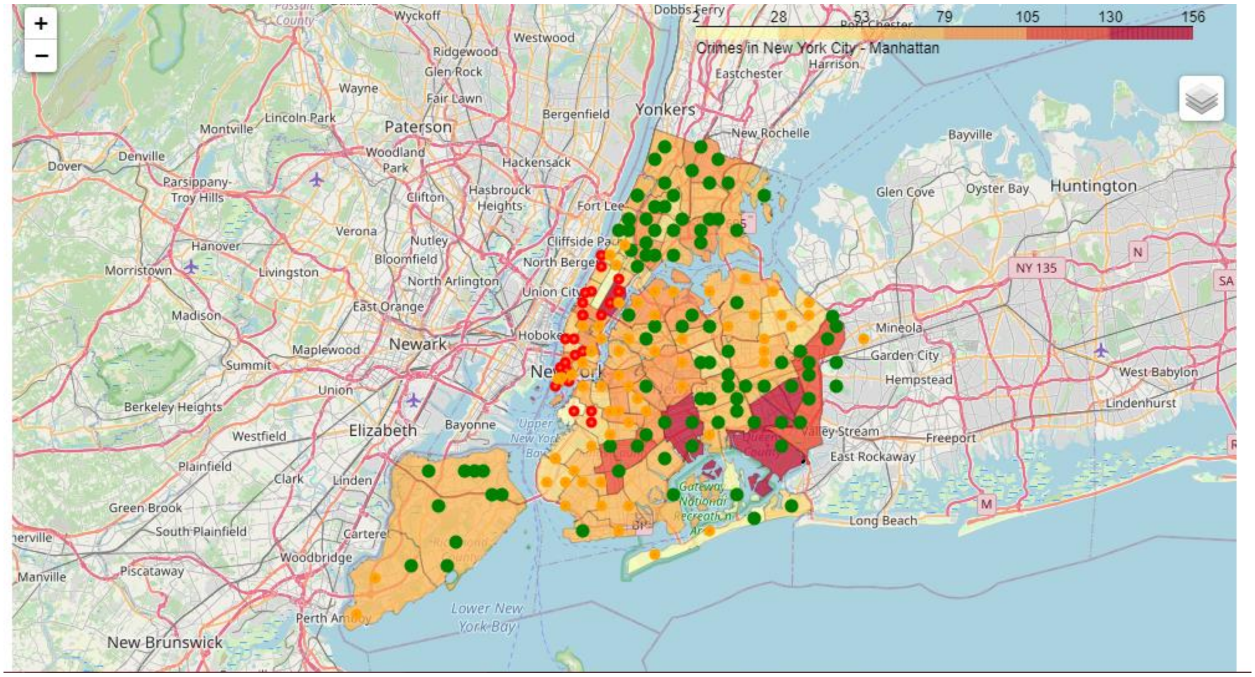
Since Cluster Two does not give us any more information, I removed it from the bar chart to better analyze the other clusters. Cluster One seems to feature the outdoor related categories, such as, Parks, Trials, and Playground. Cluster Three features food and drinks categories, like Coffee Shop, Cafe, and Diner.



4.3 Combine and Overlay the Data

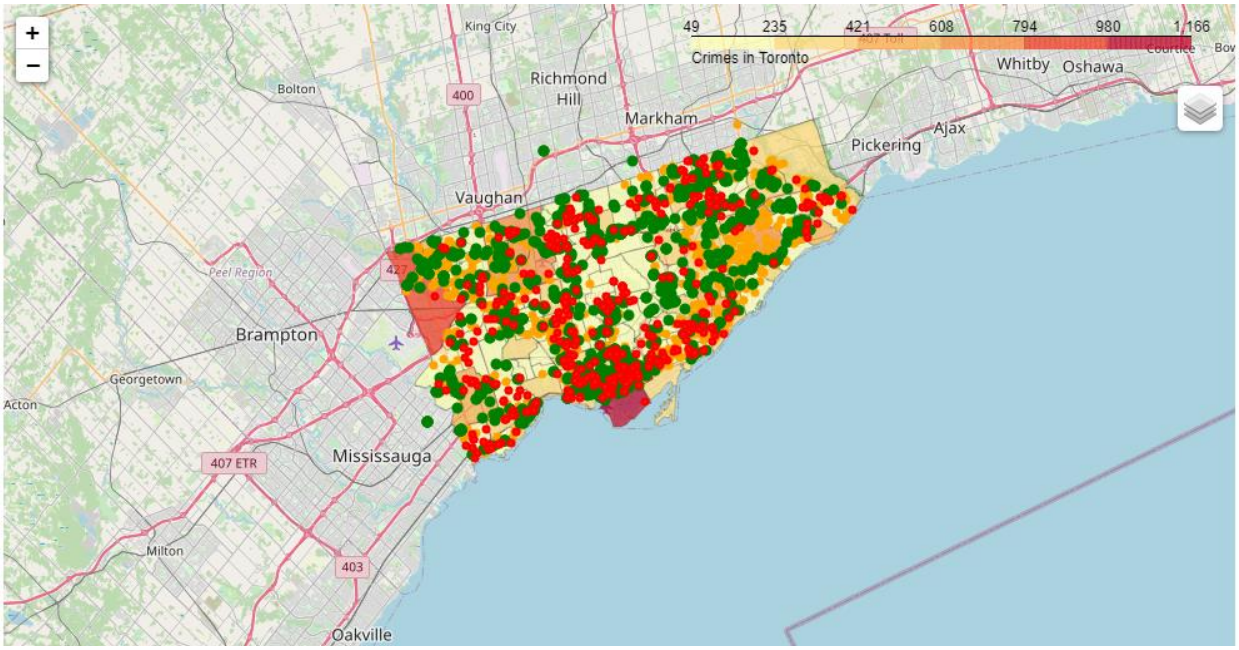
4.3.1 New York City Crime and Housing

To visualize the potential relationship of crime and housing values, I overlapped the Housing Map over the Crime Choropleth Map. Here, the New York City Map show both the crime density in the neighborhoods and the housing values distributed in the city.



4.3.2 Toronto Crime and Housing

Here, the Toronto Map show both the crime density in the neighborhoods and the housing values distributed in the city.



5. Discussion

The crime section of this project worked well keeping the New York City Data separate from the Toronto Data. This is because the records were organized differently. The NYPD kept the crime record under one dataset and there was a wide variety of crime categories, whereas the Toronto law enforcement kept their crime records separate in different datasets by crime category. It was at my discretion to use only the Toronto Homicide and MCI datasets for this project, but I encourage the use of more if not all, of the datasets available for further analysis.

The Housing section of this project came with a few complications. Again, pulling data from various sources came with the complication of what is available and how it was organized. Also, comparing two global cities came with the complication of different currencies and differences in cost of living due to differing cultures and laws.

This project is only useful to a very specific audience. Those who live in New York City or Toronto may be the only ones who would find the information helpful. The information on venues has a narrower focus to those in Manhattan due to the limited ability in using the Lite version of the Foursquare API. I would recommend broadening the scope of the venue information to the city level.

Within the venue section, Cluster Two was only representative of the garden category, thus indicating the cluster did not segment the data correctly and the centroid is exactly located in that neighborhood in Toronto. This had a high frequency of garden places and since there was no other cluster with similar venues around, I can say the Venues K Means did well.

I was curious to see visually how the crime and the housing values compared within each city. I think it brings up some interesting potential for further research. On the New York City Map, the area with the highest density of crime in 2019 seemed to have a grouping of low budget housing, however, there are other areas, both high, medium, and low crime density that had a mix of housing clusters dispersed across. So, it cannot be concluded there is a correlation between crime occurrence and housing values. This would need to be further analyzed.

6. Conclusion

After performing exploratory and cluster analysis, I have determined the percentages of crime occurred, housing values, and types of venues in both New York City and Toronto. I have built interactive maps to visualize these analyses to help better determine the ideal neighborhood to move to.