# A Search for better Residential Rental investment in

# Toronto neighborhood

*Finding needle in Haystack*

## Preface

The below Data analysis project is developed as a part of IBM DataScience Capstone project. This is a final project of a 9 course professional certificate that aimed at providing intermediate knowledge of Data Science. The scope of the project is to frame a business question around any city in the world , use the public data available for the city . Transform the data as required and further must utilize Foursquare client /API to explore the various location data in the selected city to answer the Business query .

Foursquare client API is a social location service that provides real time access to foursquare client global database. The database has rich information about venue and user ratings.

For my project , I have created a hypothetical business query for Toronto city below as suggested in the course . I will be getting all the relavent neighbourhood data from public datasets , Transform and prepare this data to further explore and understand the city's neighbourhoods to answer my business questions.

## Understanding the Business

### *Introduction*

Toronto is not only the financial hub and largest populated city of Canada, It is also one of cities which is vouched for real estate investments in the world. Toronto real-estate market rise around 20% every year.

Many large builders and also mid-sized investors prefer to set up their real estate businesses in neighbourhoods of Toronto.

One of the most popular investment strategy for mid- size and small investors is to buy residential properties and rent out. In last few years Toronto real estate market has seen sharp uprise and is highly inflated. Investing on rental properties in such volatile market can be risky and tricky at the same time.

The choice of neighborhood is critical as it can impact the demand and also timeline as in how quick the property is occupied. The Return on Investment depends on these two factors. It is observed that in some of the popular neighborhoods , buying homes prove economical than renting. With low mortgage interest rate and low downpayments, Residents prefer to pay Mortgage interests rather than paying rent and own the property after amortization period.

In such sxenario, the best bet would be on temporary residents like students ,temporary workers or trainees. The probablity of this group of people renting home is more than buying. Hence investing in residential rentals in the neighbourhoods with large and popular academic institutions make sense.

### *Requirement*

Mr A is a mid-size real estate investor and is looking to invest in residential properties to rent in Toronto.He is mainly looking into single family homes or Condo properties .

From his experience he suggests that temporary tenants prefer economical condos or small houses in main city rather than extended outskirts with well connected public transit. He thinks such tenants prefers comfortable and maintainable small spaces in busy hub over big luxurious homes in natrure rich outskirts. But at the same time they would look for easy accessblity to places and nearness to their university or schols for their children.

He wants us the Data scientists team ,to explore various neighbourhoods of Toronto city and suggest which neighbourhood is more suitable for his type of investment.

The main concern of Mr A is that , the properties that he would invest on should not be vacant for more than 3 months any time of the year. Also, he preferes tenants who dont live more than 3-4 years in a stretch in same property .He believes in carrying out maintainence changes himself every few years .With tentants living in property it gets difficult for him.

With this given criteria from Mr A , we would target neighbourhoods with universities and school districts . Most of the temporary tenants are either university/college students or single families who come to city for work purposes. Such population would prefer to live near their locations of interest and be well connected even without their own travel options.

## Analytic Approach

From our business understanding , we can summarize our business query to below questions.

1. Which neighbourhood has universities ,schools and other colleges that would attract floating population to rent houses in that area?
2. What is the public travel accessiblity rating of the neighbourhood? How far is it from bus terminal, train stop or others?

To answer the above queries our approach would be as below :

* Get the list of neighbourhoods of Toronto city.
* Consider only those neighbourhoods which have university/colleges/schools within a specific radius of distance.
* Consider the various public transit options available in the locality.
* From the result of data cleaning and merging, we will choose the best list of neighbourhoods based on their aggregated score which measures the availablity of interested locations and travel accessblities.
* Further we will cluster the above neighbourhoods list and recommend the cluster that is most suitable for Mr A can to invest on.

In the entire Data analysis process , we plan to use python and pandas for retrieving and preparing the dataset.

For clustering we will be using popular k-means algorithm. For generating visualization maps we will use Folium. Also , we are using Four square API ( public location service) to get the venues.

## Data Requirements and Collection technique

1. The primary datasource that will be used for getting Toronto neighbourhoods is

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

We will extract the details by web scrape technique from the wikipedia page The information contains postal codes, Boroughs and associated Neighborhoods.

1. The latitude and longitude data is used from IBM assignment csv.

<http://cocl.us/Geospatial_data>

1. We will be using [Foursquare](https://foursquare.com/) client to get all the near by venues for the updated list of neighbourhoods of Toronto that we get after data cleaning . We use Foursquare developer account .

Connection to Foursquare database happens through its Restful API using a URI (Univeral Resource Identifier). There are additional input parameters that can be passed in to retrieve different kind of information for a venue.

As our Business intends to search the rental properties only in and around neighborhoods having Universities and schools with some options of public transit , We have used only below categories of venues from Foursquare venue categories list to keep the data limited to our specification .

| **CategoriesId** | **Venue Categories** |
| --- | --- |
| 4d4b7105d754a06372d81259 | college and university |
| 4bf58dd8d48988d13b941735 | school |
| 4bf58dd8d48988d1ed931735 | airport |
| 52f2ab2ebcbc57f1066b8b4f | bustop |
| 4bf58dd8d48988d129951735 | Train station |
| 52f2ab2ebcbc57f1066b8b51 | Tram station |

Note: These are main categories. This also includes several subcategories which will be filtered out later based on our requirement.

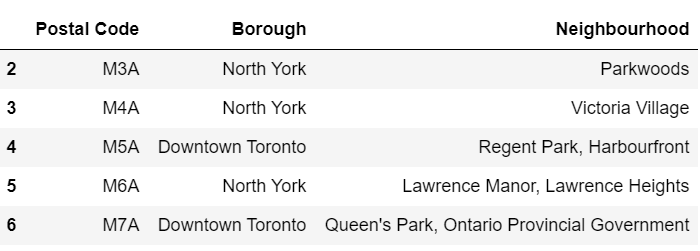
## Methodology

### A. Preparing the Data for Analysis

This section involves information on how we prepared our data source for analysis.

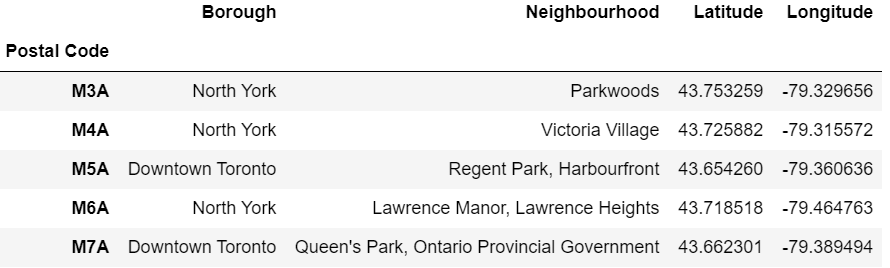
1. Required libraries installed . This includes panda, numpy , json, geocoder, matplotlib, k-means and Folium are primary ones.

1. The Toronto neighborhood data is obtained from the public wikipedia page by web scraping technique. The dataset we will create would consists of information in columns Postal Codes, Boroughs and Neighborhoods. The raw data is cleaned to filter out all the Boroughs that are unassigned for our analysis. Also, we have confirmed that the dataset does not have unassigned neighborhoods. The snippet of resulting dataset .



picture 1 : Web scraped dataset

3. The latitude and longitude data is obtained into dataframe and is merged with above data . The join is per formed on postal\_code column as index.



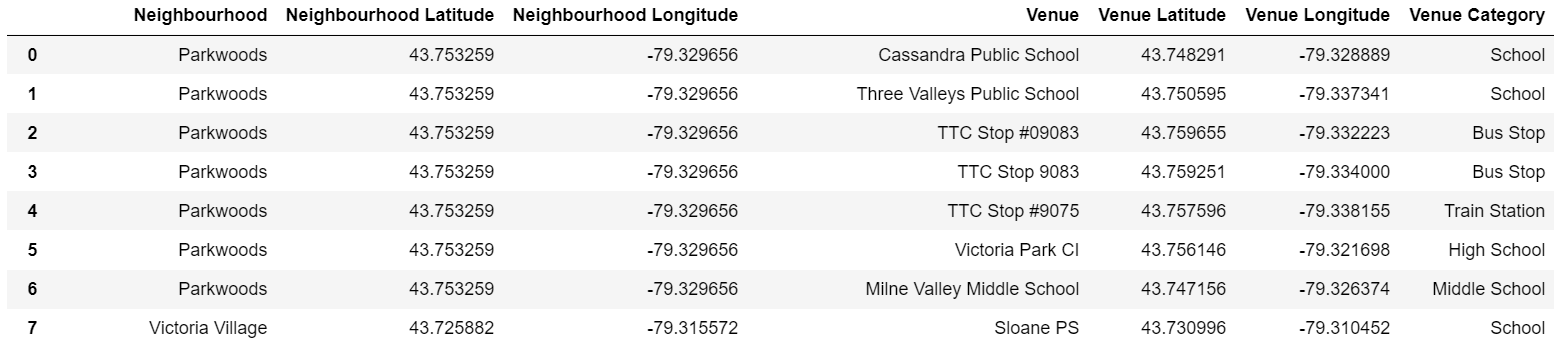
picture 2: Merged Dataset

The dataset at this point have 10 Boroughs and 103 unique Neighborhoods. This dataset forms the basis for further data analysis.

4. Next step is to log into the Foursquare client . This will need the Client ID and Client Secret that we would have obtained when a developer account is created.

5. We will run the Foursquare function to retrieve the list of different Venue category types. Also , we would define the Foursquare function to extract the nearby Venues for all of our neighborhood . In this function we will input our required categories and the radius to be 1km . This is done in request URI .

6. Toronto coordinates are obtained and using the coordinates above function are run to obtain the nearby venues for each neighborhood.



picture 3: Snippet of Nearby venues of each neighborhood

We obtain a total of 2154 venues for 103 neighborhoods . We have 87 unique categories in the result.

All these 87 categories are sub categories of the major categories that we have included in our request.

7. We have then created One Hot encoding on categories for each neighborhood in the above dataset.

This is done to convert the categorical values i.e venues to numerical (**boolean**). This is necessity for applying any machine learning algorithms on the dataset as they can only perform analysis on numerical values.

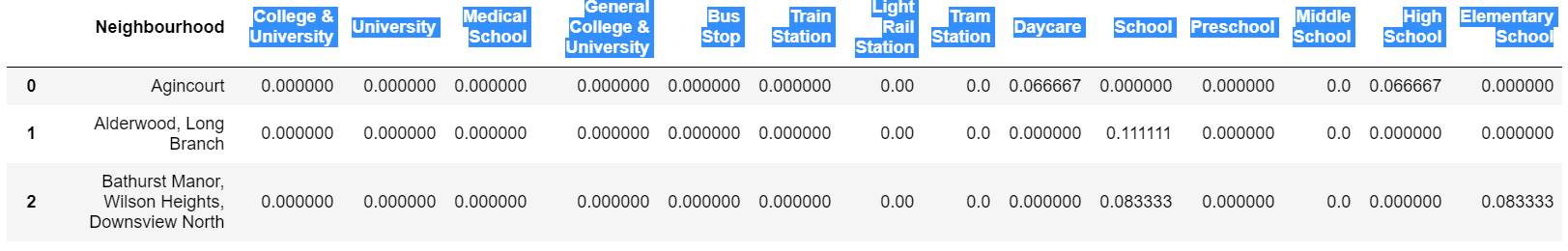
8. Further on the encoded dataset we calculate the mean function for each neighborhood. This gives us the frequency of occurence of each Venue category out of total venue categories in each neighborhood.

For example , a value of **Day Care= 0.066667** in Agincourt neighborhood indicates that the ratio of occurence of Day care in proportion with other venue categories is 0.066667.

9. At this step , we further cleaned our dataset to filter out unnecessary venue categories. The initial categories that we included has several sub categories. Out of those some categories are redundant and some dont need to be considered in our analysis. For example ‘**College ground**’ category that is part of ‘**College and university**’ category can be removed. In our scenario we would consider the nearby college and we dont need to consider grounds.

10. After the above transformation our dataset has only 17 venue categories for 103 neihborhoods.The categories included are

* College & University
* University
* General College & University
* Bus Stop
* Train Station
* Light Rail Station
* Tram Station
* Day Care
* School
* Pre-School
* Middle School
* High School
* Elementary School



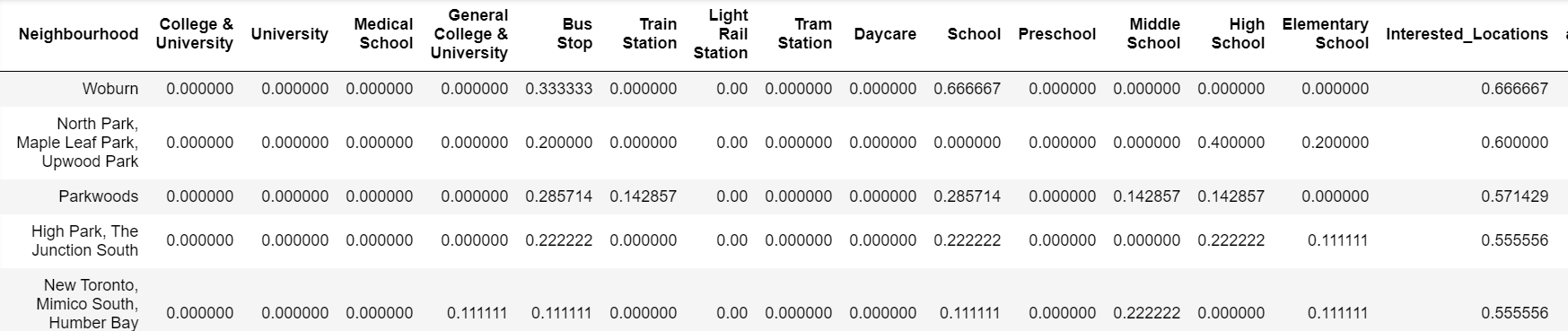
Picture 4: Snippet of Hot encoded and cleaned dataset

11. Adding new columns ‘Interested\_Locations’ and ’ Accessblity\_score’ to the dataset.

* Maximum Number of occurences of venues that are of interest to business in each neighborhood is calculated as ‘Interested\_Locations’.For example , the sum of number of occurence of schools, unversity,college in the specific neighborhood.
* Maximum number of occurences of various public transport modes are calculated as ‘Accessiblity\_score’.Example , the sum of number of occurence of Bus stop, train station etc. in the specific neighborhood.

12. **‘Interested\_Locations’ and ’ Accessblity\_score’** will be used as measures to rate the suitablity for our Business . We have considered the neighborhoods with higher the value of entity interested\_location meaning maximum number of summation of locations like universities, schools and colleges and higher accessblity\_score meaning more options of public transit available.

13. We have filtered only those neighborhoods that have positive values for both ‘Interested\_Locations’ and ’ Accessblity\_score’ . With this our source dataset now only contains 31 neighborhoods in descending order. Top choice is at the top.



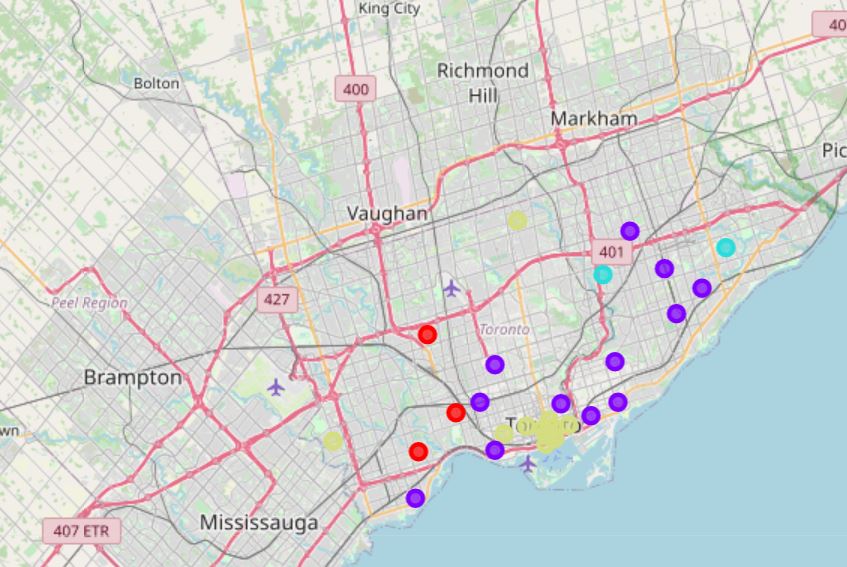
picture 5: Snippet of top list of neighborhoods dataset

. **B. Clustering the dataset**

1. We have used k-Means Clustering algorithm to group the resultant neighborhood list from our analysis.

2. k-means clustering is unsupervised machine learning algorithm used for the unlabeled data.It forms k groups within the available data points.

3. Best value of k= 4 is identified by elbow method using euclidean distance.



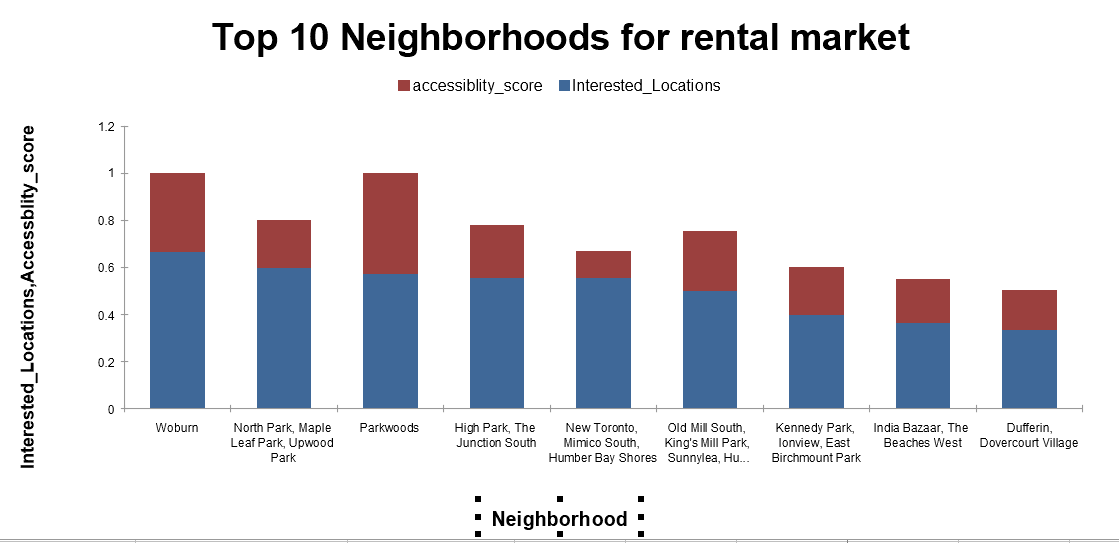
Picture 6: Toronto Map with 4 different clusters superimposed.

### C. Analysis of Clusters

* Cluster 0 (Represented in red color in Map) has only schools in most common venues and maximum number of train stations out of other clusters.
* Cluster 1 (Represented in purple color in Map) has a good mix of schools and colleges , but no bigger universities. However it is well connected with all modes of transit.
* Cluster 2 (Represented in turquoise color in Map)has only schools with only bus stop in top common venues. Also it comprises of only 2 neighborhoods.
* Cluster 3 (Represented in Yellow color in Map) has well connected neighborhoods with several universities and also have good number of schools and colleges. It has best bus transit facity out of the four.

## Results

As a result of our Data analysis , we have shortlisted 31 neighborhoods for our client. Below represents the bar chart of top 10 neighborhoods that have many universities/schools or colleges and is also well connected with public transit.



Picture 7: Bar chart listing the Top 10 neighborhoods for investment

Further we have also created clusters for giving a group of neighborhoods as an option rather than one single neighborhood. This is important because this data analysis has not taken into consideration of availablity of properties and their prices into account. From our analysis , we recommend our Business to invest in below mentioned priority:

* Cluster 3 has all of the locations of interest to the business. All the neighbourhoods are close to each other and can be accessed by residents in any of these neighborhood. Rental properties in Cluster 3 would giver better ROI
* Cluster 1 will be second choice. It has good number of schools and colleges with good transit options.

## Discussion

The neighborhoods considered for analysis are limited to Toronto city and includes no suburb information for simplicity of model. This data analysis is carried out solely based on the number of venue categories available from Foursquare. It is not prescribed to reccommend the options just based on the calculation of neighboorhood and venue dataset.

Data analysis could be significantly improved by considering the average home prices , Expected rent and options like crime rate in that location. Since I could not get the public data on these information , I based my analysis on the available relavent data.

k-means algorithm is most suitable algorithm for clustering in such categorical dataset. However , we did get few outliers even with the best value of k. This could not be avoided in unsupervised model of algorithm.

Clusters is not perfectly disjoint in this case as all the records of data belong to similar Venue categories in our requirement.So labelling data is difficult and consists of ambiguity.

## Conclusion

Toronto is no doubt a hot market for investing in real estate properties. But as discussed earlier , the neighborhood data should be carefully studied to ensure better Return of Investment. The volatile inancial market and fluctuating Bank interest rates just add to the problem. But if invested in right neighborhood for right type of tenant , Business could make huge passive income with significant profits.

## References

1. Toronto neighborhoods

<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

2. Toronto neighborhoods- Coordinates information

<http://cocl.us/Geospatial_data>

3. FourSquare Venue Information

[Venue Categories | Build with Foursquare](https://developer.foursquare.com/docs/build-with-foursquare/categories/)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |