Capstone Project – The Battle of the Neighborhoods

1. Introduction/Business Problem:

1.1 Background:

Canada has one of the hottest real estate markets in the world. Cities like Vancouver and Toronto lead the charts in terms of expensive housing in the country. Toronto attracts a lot of people from different part of the country and as well as the world as its technological hub is constantly growing and thriving. With ever increasing population in the Greater Toronto Area, the real estate market is having a hard time meeting the demand. This has resulted in the housing prices skyrocketing in the last few years.

1.2 Business Problem:

Given the limited supply of housing and ever-increasing prices, it is imperative that people make an informed decision while buying a property.

This capstone project will explore the average prices of different neighborhoods in Toronto, analyze different venues in each of those neighborhoods and determine which venue categories, such as restaurants, schools, public transport and so on, have the most impact on housing prices. At the end, the top five categories that seem to affect housing prices will be summarized.

1.3 Target Audience:

The results from this project should be very helpful to the general public that includes buyers and real estate agents who could use it as a potential tool to decide which neighborhood would be good to purchase a property in.

2. Data Section:

For this capstone project, we need to gather data from few different sources. The sources are listed below:

1. We need to gather all the postal code information for the city of Toronto. Data for this will be scraped from this link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

2. The data for average prices in Toronto based on postal codes can be obtained from this link: https://housepricehub.com/cities/city/Toronto. The sample, for example, will look like below:

All Postal Areas in Toronto

Show 100 entries Search:

Postal Code	City	Average Price	Average Price Per Frontage Ft	Total listings	Average Price Trend
M2L	Toronto	$6,307,644	N/A	112	
M4W	Toronto	$5,731,778	N/A	28	
M5P	Toronto	$5,459,942	N/A	37	
M3B	Toronto	$5,102,447	N/A	63	
M3C	Toronto	$4,821,856	N/A	19	
M2P	Toronto	$4,344,466	$76,333	36	
M4V	Toronto	$4,126,787	$85,428	33	
M4N	Toronto	$4,108,118	N/A	63	
M5R	Toronto	$3,575,950	N/A	42	
M5N	Toronto	$3,223,840	N/A	32	

3. We will also need to gather data about different venues, categories for each of the neighborhoods. This can be achieved by using the **Foursquare API data**. The data retrieved from Foursquare API will include neighborhood, latitude, longitude, name of the venue, category, etc. A sample of data obtained from Foursquare is shown below:

Out[113]:

	name	categories	lat	lng
0	Jimmy's Coffee	Coffee Shop	43.658421	-79.385613
1	Tim Hortons	Coffee Shop	43.658570	-79.385123
2	Somethin' 2 Talk About	Middle Eastern Restaurant	43.658395	-79.385338
3	Hailed Coffee	Coffee Shop	43.658833	-79.383684
4	Neo Coffee Bar	Coffee Shop	43.660140	-79.385870

3. Methodology:

As mentioned in the introduction section, Toronto is one of the hottest real estate markets in the world. So, in order to make an informed decision in such a market, it is important to have access to as much as detailed information as possible. In this study, we will carry out a detailed analysis of different neighborhoods in Toronto and see what kind of impact different venues and number of venues have on house prices.

In the first step, we collected postal code, borough and neighborhood for different areas in Toronto from Wikipedia. We also collected the average house prices for each of those postal codes from the “housepricehub” website. After importing the required data, we cleaned all the data and matched all house prices data with postal code, boroughs and neighborhoods data.

The resulting dataset is shown below:

[11]:

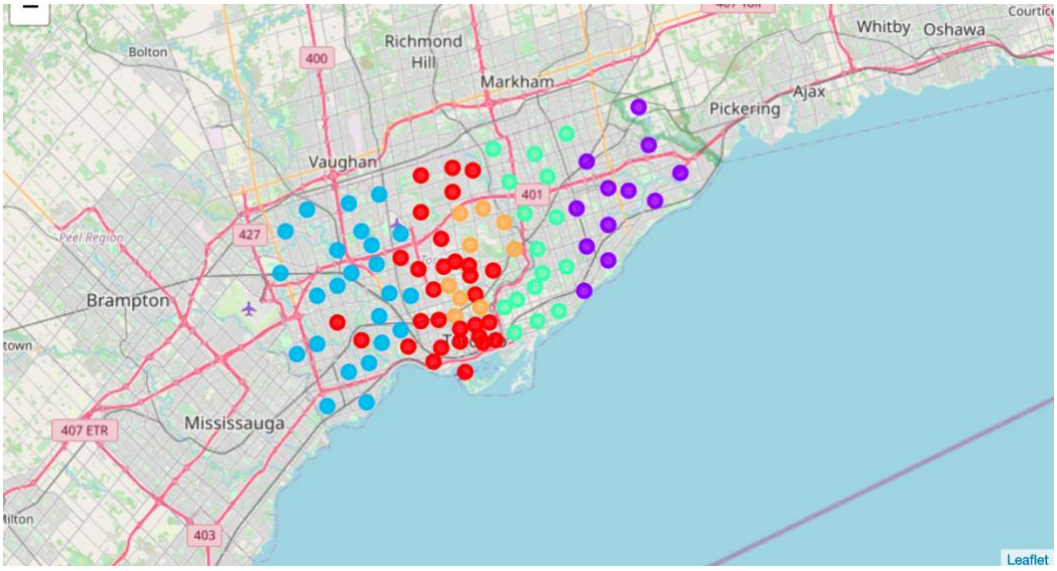
	Postalcode	Borough	Neighborhood	Latitude	Longitude	Average Price
0	M3A	North York	Parkwoods	43.753259	-79.329656	1739779
1	M4A	North York	Victoria Village	43.725882	-79.315572	1162631
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636	2215326
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763	1714409
4	M9A	Etobicoke	Islington Avenue, Humber Valley Village	43.667856	-79.532242	2698405

In the second step, we used *k-means clustering* to compare different neighborhoods and create clusters for those neighborhoods. This clustering process should give us an indication of what different clusters look like and how the average house prices in each of those clusters look like. We used a k-means cluster value of 5. The average house prices for each of those clusters are below:

6]:

Labels	Latitude	Longitude	Average Price
0	43.693884	-79.415403	2.164773e+06
1	43.764024	-79.228896	1.049548e+06
2	43.691742	-79.517243	1.211150e+06
3	43.730659	-79.319442	1.320387e+06
4	43.716190	-79.383470	4.834319e+06

The results from the clustering algorithm is shown below:



After running the k-means cluster algorithm, we looked at all the average prices data for different boroughs and neighborhoods in the city of Toronto. Based on available data, we decided to pick Toronto borough for our data analysis. We specifically picked Toronto borough as it contained the greatest number of neighborhoods amongst all boroughs (East York, North York, Scarborough, Toronto and Etobicoke). Toronto dataset also contained average house prices that varied between 1.5 Million to 3 Million Canadian Dollars. Having such a vast difference in house prices will be very helpful as it be would be easier to analyze and differentiate different neighborhoods.

In the third and final step, we used Foursquare API to gather different venues' information for each of the neighborhoods. The data from Foursquare included venue name, venue location (latitude and longitude), venue category, etc. The field we are interested in this study is venue category. The category assigned for each entry was from the lowest hierarchal level for category assignment. We were able to obtain the category hierarchal level from the Foursquare website (https://developer.foursquare.com/docs/build-with-foursquare/categories). There are over 100+ entries for entire hierarchal level. In order to understand the data, we decided to use the only the top level categories for all entries. The top level categories are: Arts & Entertainment, College & University, Events, Food, Nightlife Spot, Outdoor & Recreation, Professional & Other Places, Residence, Shop & Service and Travel & Transport. In order to carry out this process, we imported category data from Foursquare website and then assigned main category to each of the entries. This what the dataset looks like after assigning the main category:

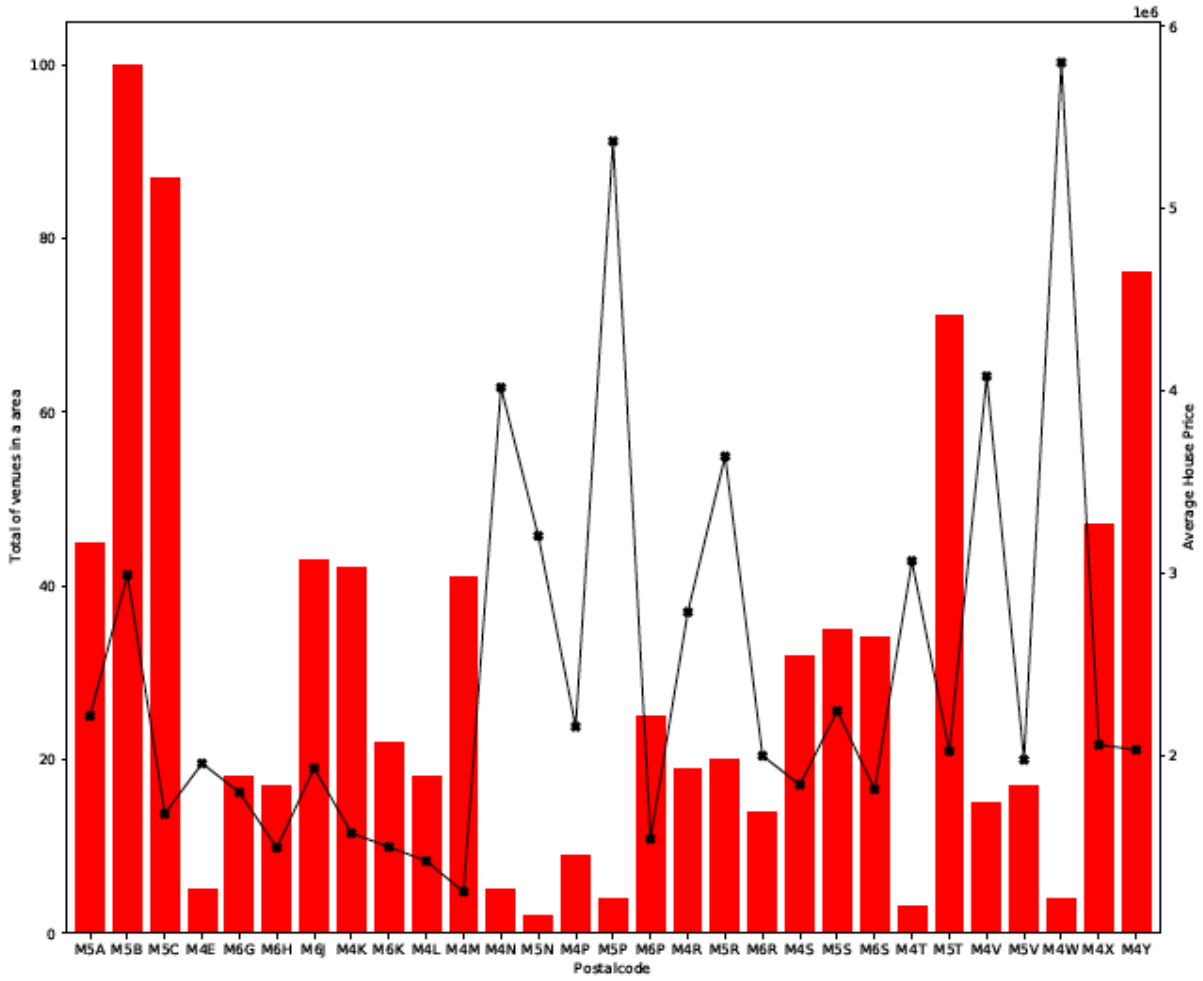
[42]:

	name	categories	lat	lng	Main_category
0	Storm Crow Manor	Theme Restaurant	43.666840	-79.381593	Food
1	DanceLifeX Centre	Dance Studio	43.666956	-79.385297	Arts & Entertainment
2	The Alley	Bubble Tea Shop	43.665922	-79.385567	Food
3	Glad Day Bookshop	Bookstore	43.665271	-79.380785	Shop & Service
4	Smith	Breakfast Spot	43.666927	-79.381421	Food
5	Fabarnak	Restaurant	43.666377	-79.380964	Food
6	Bar Volo	Beer Bar	43.665462	-79.385692	Nightlife Spot
7	Sansotei Ramen 三草亭	Ramen Restaurant	43.666735	-79.385353	Food
8	Como En Casa	Mexican Restaurant	43.665160	-79.384796	Food
9	Barbara Hall Park	Park	43.666879	-79.381068	Outdoors & Recreation
10	Ho's Team Barber Shop	Salon / Barbershop	43.665630	-79.381359	Shop & Service
11	FUEL+	Juice Bar	43.664399	-79.380427	Food
12	Coach House Restaurant	Diner	43.664991	-79.384814	Food
13	Starbucks	Coffee Shop	43.664980	-79.380510	Food
14	T-Swirl Crepe	Creperie	43.663452	-79.384125	Food
15	The Men's Room	Men's Store	43.664446	-79.380067	Shop & Service

After assigning main category to all entries, the next step was to count each category type for each of the neighborhoods. This information can then be used to plot data along with average house prices to see the impact of different venue categories, number of venues has on house prices in a certain neighborhood.

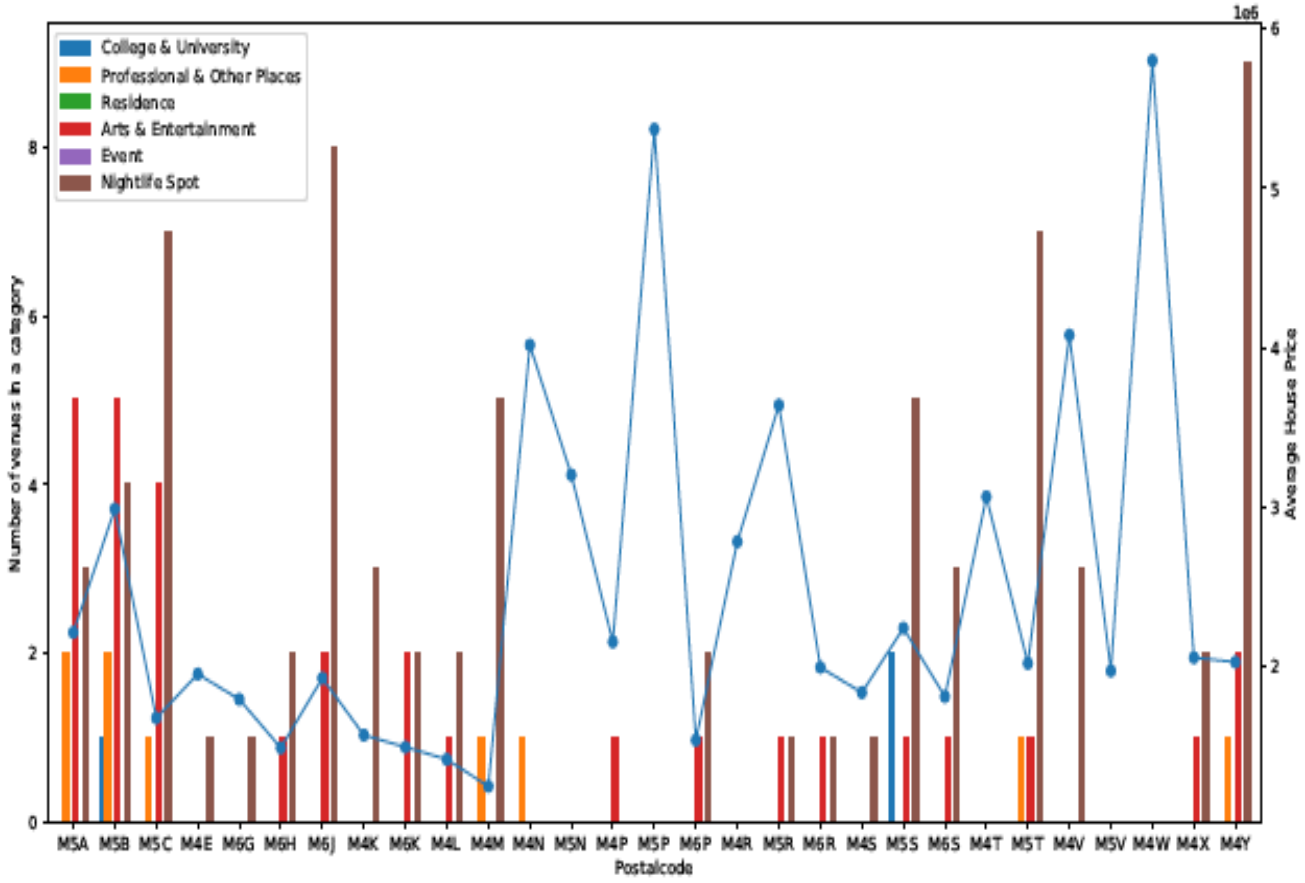
4. Results:

After carrying out all required data analysis and assigning main category, we used bar charts to compare different neighborhoods that contain venue category count information and average house prices. The first plot below looks at the overall venue count for each of the neighborhoods based on postal codes.



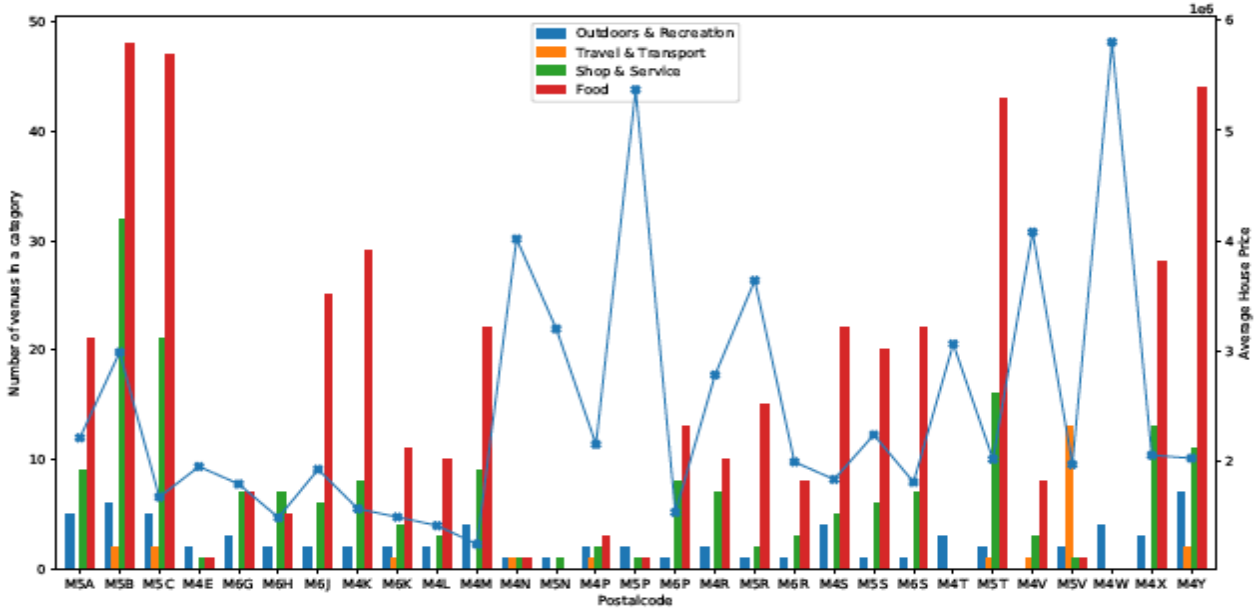
In the above plot, the bars represent the number of venues in each of the areas (postal code). The line represents the average house prices in each of the areas (secondary Y-axis). Based on the plot, there seems to be no direct impact of total venues on house prices. There does seem to be few scenarios where the number of venues seems to have an opposite effect on house prices i.e. lower the number of venues in an area, the higher the house prices. However, there are also scenarios where the higher number of venues in an area, the lower the house prices.

The second plot below shows the number of venues in each of the area for the following categories: College & University, Professional & Other Places, Residence, Arts & Entertainment, Events and Nightlife Spot. The line plot represents average house prices in each area (secondary Y-axis).



The “Nightlife Spot” category has a clear trend in the plot shown above. Areas with higher count of nightlife spots seem to have lower average house prices. The areas with highest house prices don’t have any or very few nightlife spots in the above plot.

The plot below shows the remaining categories. The remaining categories include Outdoors & Recreation, Travel & Transport, Shop & Service and Food. The line plot shows the average house prices in each area. One category that seems to have a trend is Food. For the most part, the higher the count for food venues in an area, the lower the house prices. The top 3-4 areas that have the highest prices have a very low count of food venues.



5. Discussion:

Based upon the findings in the results section, it is evident that there exists a co-relation between number of venues in an area, number of different venue categories and average house prices. The study looked at 10 different categories that covered almost all of the major categories that people would be interested in. Findings from this study can be used to make an informed decision while trying to decide which neighborhood would be the best to buy a house in. For example, if a person likes having lots of nightlife spots, outdoor & recreation while having a smaller number of college & university and want the house rice in such a location to be reasonable they can use these plots/study to arrive at such a decision. Even a real estate agent can use results from this study to help their clients find an ideal location based on their preferences. This study can also be applied to a different borough in Toronto or to a different city.

This study can be further extended in the future by including data that contains information about schools, their rankings and proximity, etc. I believe that will add another valuable layer to this project.

6. Conclusion:

In this project, we analyzed the house prices in various neighborhoods in Toronto and studied the impact of various venue categories on house prices. This study can be used by the general public such as buyers and real estate agents to make informed decisions. By making use of various techniques and methods such as webpage data scrapping, clustering data using k-means clustering, obtaining data from Foursquare API, cleansing data and making use of bar/line plots I feel this study nicely encapsulates the entire Applied Data Science Certification.