Best Place to Live in Toronto

Alan KM Wong

2020-05-01

Chapter 1 - Introduction / Business Problem

Consider being someone who would like to move to Toronto or invest in real estate within the Greater Toronto Area, it is often a challenge to identify where the best place to buy is. There are just so many factors to consider such as:

- Are there shops nearby where I can buy my daily groceries
- When needed, are there medical facilities where I can see a doctor?
- Are there recreation facilities such as parks, or nature where I can take a walk?
- How much does an average unit cost in the neighbourhood?
- Will the value of my property rise in the future if I buy today?

All these questions weighs in an affects how we shape our decision. Needless to say, purchasing real estate is often a big investment and takes up a considerable amount of ones savings. A lot of exploration and research is needed to be able to make an informed decision. Within our pursuit to find a reasonable answer, we consider two major factors; the quality of living within the neighbourhood as well as its' housing price “attractive"-ness..

Looking into these two major facets, we will attempt to use Data Science techniques to help answer our question “Where is the best place to live in Toronto” with our audience being people who are interested in moving there.

Toronto Neighbourhoods with the Best Quality of Life vs Housing Price Aggregate¶

"Quality of Life" is an abstract term. For different people, this can mean different things. In our evaluation, we will take the number of venues within a neighbourhood to determine a "Quality of Life" index. Within this index, we will consider the number of essential venues within walking distance (taken as 1 kilometre walking distance) of the neighbourhood.

These venues are defined as:
- Medical Centre (4bf58dd8d48988d104941735) - These include such medical facilities as: Clinics, Pharmacies, Hospitals, Dentists, etc.
- Shop & Service (4d4b7105d754a06378d81259) - This broad category covers anything from ATM machines, Auto Garages, Clothing Stores, etc.
- Restaurant/Food (4d4b7105d754a06374d81259) - Covers anything ranging from restaurants of various cuisines, to Supermarkets.
- Outdoor & Recreation (4d4b7105d754a06377d81259) - These include Parks, baseball fields, tennis courts, etc.
- Arts & Entertainment (4d4b7104d754a06370d81259) - Venues here can range from Museums, Art galleries, mini-golf courses, etc.

"Housing price" can vary depending on various factors. This could range from its' size, type of housing, number of bedrooms, whether it has a garden, etc. As a result, housing prices between neighbourhoods can vary greatly based on the type of housing built. This can pose serious challenges towards how we can compare the housing price between neighbourhoods. To get a better appreciation of how "Attractive" a neighbourhood is, we attempt to compile what we call the Housing Price Aggregate. This is another index that takes into consideration how quickly a listing is sold, how often listings are sold above its’ asking price, and the average price of housing within that particular neighbourhood.

By cross referencing the "Quality of Life" against the Housing Price Aggregate, we can better understand and recommend where the best place to live in Toronto is.

Chapter 2 - Data

In order to support our investigations, various data sources need to be pooled together. When determining our Quality of Life index, we will mainly leverage on GPS Coordinates of different Toronto neighbourhoods together with Four Square API. As for our Housing Price Aggregate, this will be based on data from Zolo; a real estate listing company in Canada.

Geolocation Data¶

Four Square
Data from Foursquare provides the venue details based on location data. We will be determining the number of venues of a specific type within a certain distance.

GPS coordinates of location, place and city
https://www.gps-latitude-longitude.com/address-to-longitude-latitude-gps-coordinates
Provides useful tools to determine the GPS coordinates of neighbourhoods within the Greater Toronto area. The extraction of latitude and longitude information will support us when querying data from Four Square.

Toronto Housing Price¶
Zolo Hottest Toronto Neighbourhoods
https://www.zolo.ca/toronto-real-estate/neighbourhoods
Zolo listing provides several key data for each neighborhood in Toronto in the past 28 days such as: Listing sold under 10 days, listings sold above asking price, average sale price, number of active listings, etc.

Chapter 3 - Methodology

Data Acquisition, Cleaning, and Exploration
As a first step in performing our analysis, we need to first gather our required data. Within Toronto, the zoning and classification of neighbourhoods have changed over many years. In the past, it was very common to segregate neighbourhoods based on the Canadian Postal code that starts with “M”. Later in 2011, the Toronto Real Estate Board (TREB) came up with a new listing system that introduced the concept of “Districts”. All these different ways to identify a “Neighbourhood” makes it considerably challenging for communication amongst local residents, but more importantly making sense of different data sources. Therefore in our report, we will mainly be using the neighbourhood profiles as designated by the City of Toronto (https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/)

3.1 Housing Price Aggregate
In order to build our Housing Price Aggregate, we will be web scraping data available from Zolo; a respected real estate listings company in Canada. Zolo maintains a relatively up-to-date tabulation of real estate activity for each neighbourhood. From this tabulation, we are able to isolate the the percentage of listings sold above asking price, percentage of listings sold within 10 days of listing, and the average price sold.

After our initial web scraping, we are able to extract the details into a Panda Dataframe:

	Neighbourhood (# Rank out of 143)	Sold under 10d	Sold above asking	Average sale price	Active listings
0	Agincourt North	70%	81%	$714K	16
1	Agincourt South-Malvern West	66%	77%	$642K	21
2	Alderwood	64%	46%	$969K	24
3	Annex	51%	42%	$1.8M	88
4	Banbury-Don Mills	56%	47%	$1.1M	50

Looking closely, we quickly want to make some adjustments to clean up our data. First of all, we would like to alter both “Sold under 10 days” and “Sold above asking” from a percentage into a numerical/float value.

Next we also observed that the “Average sale price” is a string with abbreviated numbers representing K = 1,000 and M = 1,000,000. Hence, we need to clean up these values in order to better explore our data frame.

last but not least, we want to drop any neighbourhoods that have missing data as this would skew our analysis later on. After cleaning the data, we end up with a Dataframe as below:

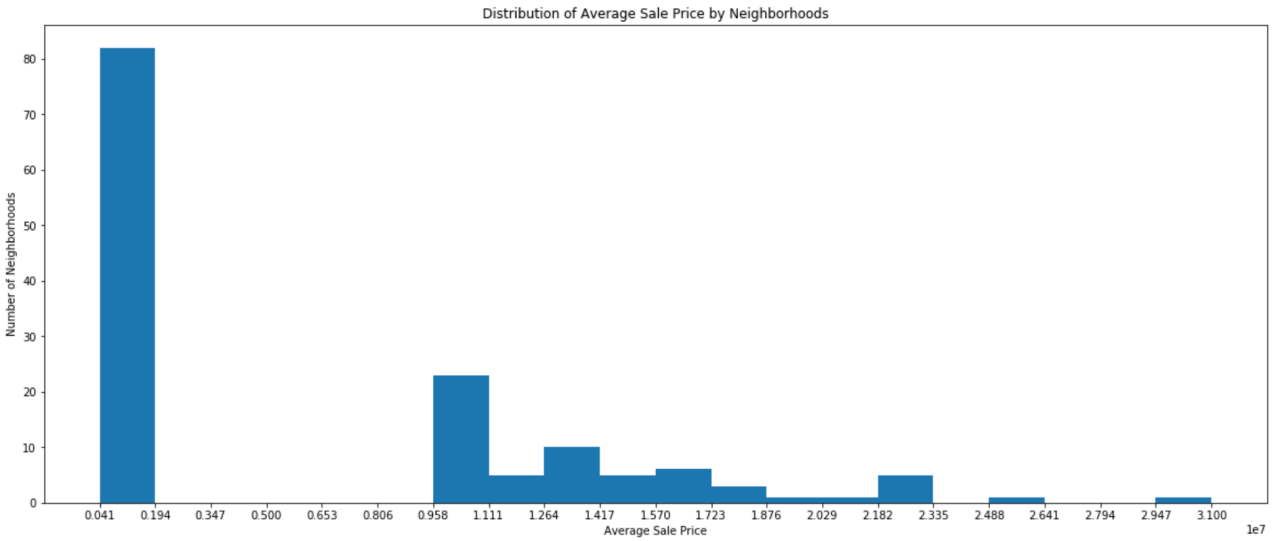
	Neighbourhood (# Rank out of 143)	Sold under 10d	Sold above asking	Average sale price	Active listings
0	Agincourt North	70	81	714000	16
1	Agincourt South-Malvern West	66	77	642000	21
2	Alderwood	64	46	969000	24
3	Annex	51	42	18000000	88
4	Banbury-Don Mills	56	47	11000000	50

With our clean Dataframe, we can now explore our data to gain some insights.

Highest Average Sale Price
First we look at which neighbourhoods command the highest average sale price. We sort our data, list the top 5 most expensive neighbourhoods.

	Neighbourhood (# Rank out of 143)	Sold under 10d	Sold above asking	Average sale price	Active listings	RankAvgSalePrice
0	Bridle Path-Sunnybrook-York Mills	45	33	31000000	57	0
1	Kingsway South	60	40	25000000	18	1
2	Forest Hill North	80	73	23000000	5	2
3	Forest Hill South	52	29	23000000	28	3
4	Lawrence Park South	67	39	22000000	21	4

In addition, also plot a histogram to understand the distribution of Average Sale Price amongst all neighbourhoods.

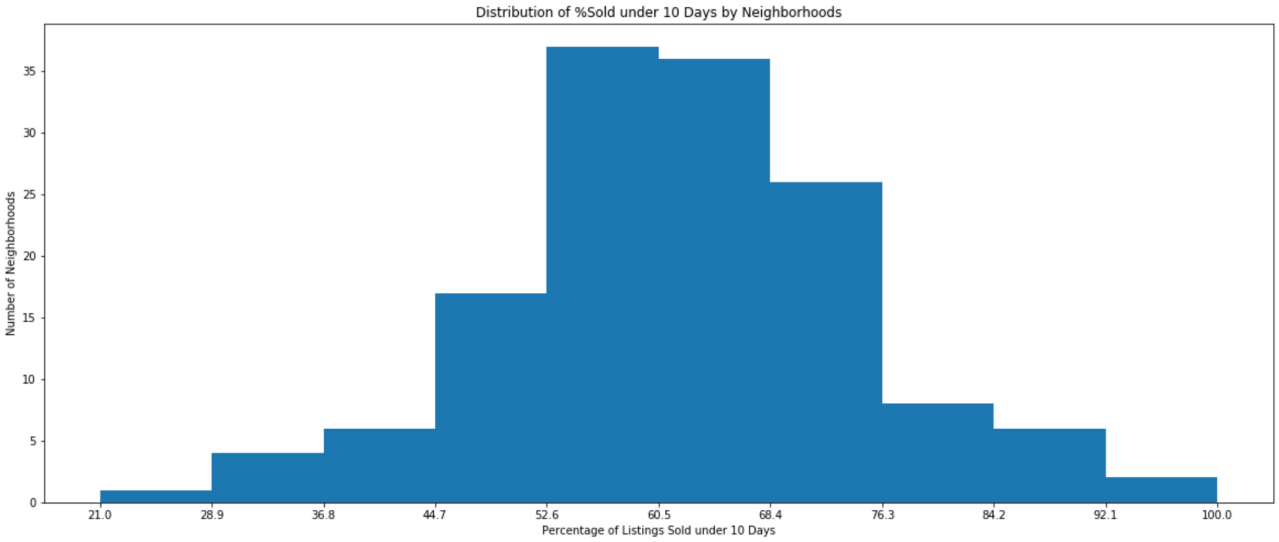


One interesting observation we see here is around 80+ neighbourhoods with an average sale price under 2M CDN and then a large price gap until we start to see the luxury market that starts around 10M CDN. At the same time, we want to save the ranking of each neighbourhood. Those with a higher average sale price (more expensive) will be given a low rank, whilst those more affordable are given a higher rank.

Sold Under 10 Days

We next proceed to explore the neighbourhoods based on the percentage of listings sold within 10 days of initial listing. Again, we are interested to see the Top 5 neighbourhoods as well as the distribution of by “Sold under 10 days” by neighbourhood.

	Neighbourhood (# Rank out of 143)	Sold under 10d	Sold above asking	Average sale price	Active listings	RankAvgSalePrice	RankSoldUnder10d
138	Playter Estates-Danforth	90	80	16000000	1	17	138
139	Woodbine-Lumsden	91	82	11000000	9	51	139
140	Danforth	92	92	13000000	6	29	140
141	North St. James Town	100	71	766000	4	102	141
142	Beechborough-Greenbrook	100	67	826000	4	87	142

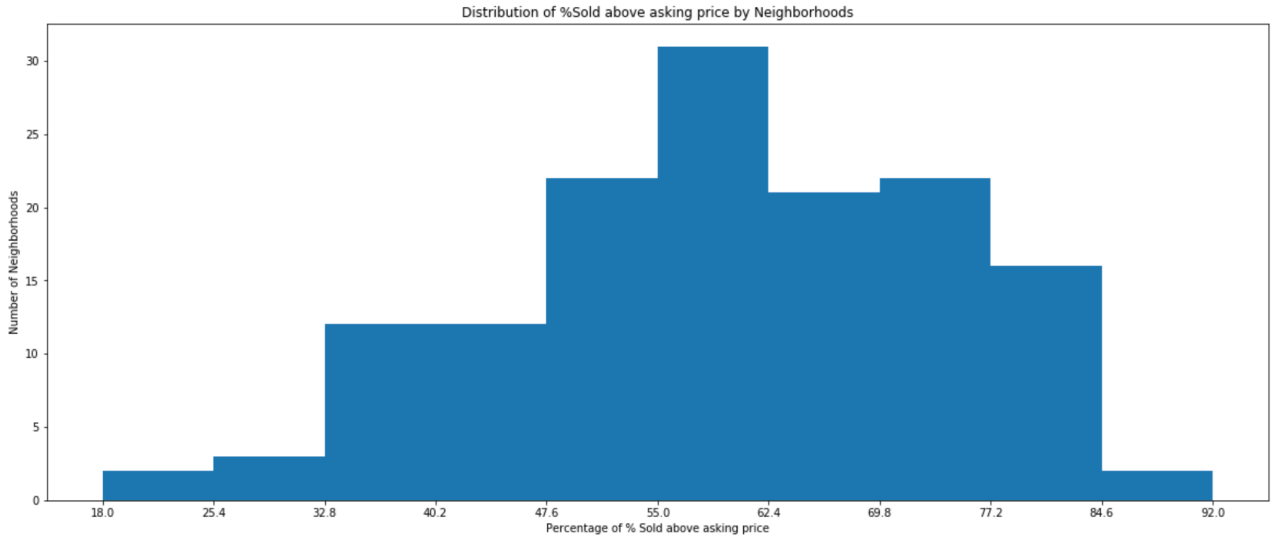


What we can observe here is the average percentage sold under 10 days to be well over 50%. This gives an indication the real estate market in Toronto is very vibrant and active. A good sign for anyone interested to invest in the property market. Similar to before, we would like to give a ranking to each neighbourhood where a higher percentage sold under 10 days will have a higher rank. In contrast, neighbourhoods with slower turnarounds will be given a lower rank.

Sold Above Asking Price

The last dimension we are interested to explore is the percentage of listing sold above the asking price in each neighbourhood. Which neighbourhoods would rank in the Top 5, and what is the distribution.

	Neighbourhood (# Rank out of 143)	Sold under 10d	Sold above asking	Average sale price	Active listings	RankAvgSalePrice	RankSoldUnder10d	RankSoldAboveAsk
138	South Riverdale	82	82	10000000	29	60	130	138
139	Rustic	82	82	11000000	3	52	129	139
140	Woodbine-Lumsden	91	82	11000000	9	51	139	140
141	Ionview	67	87	686000	2	120	92	141
142	Danforth	92	92	13000000	6	29	140	142

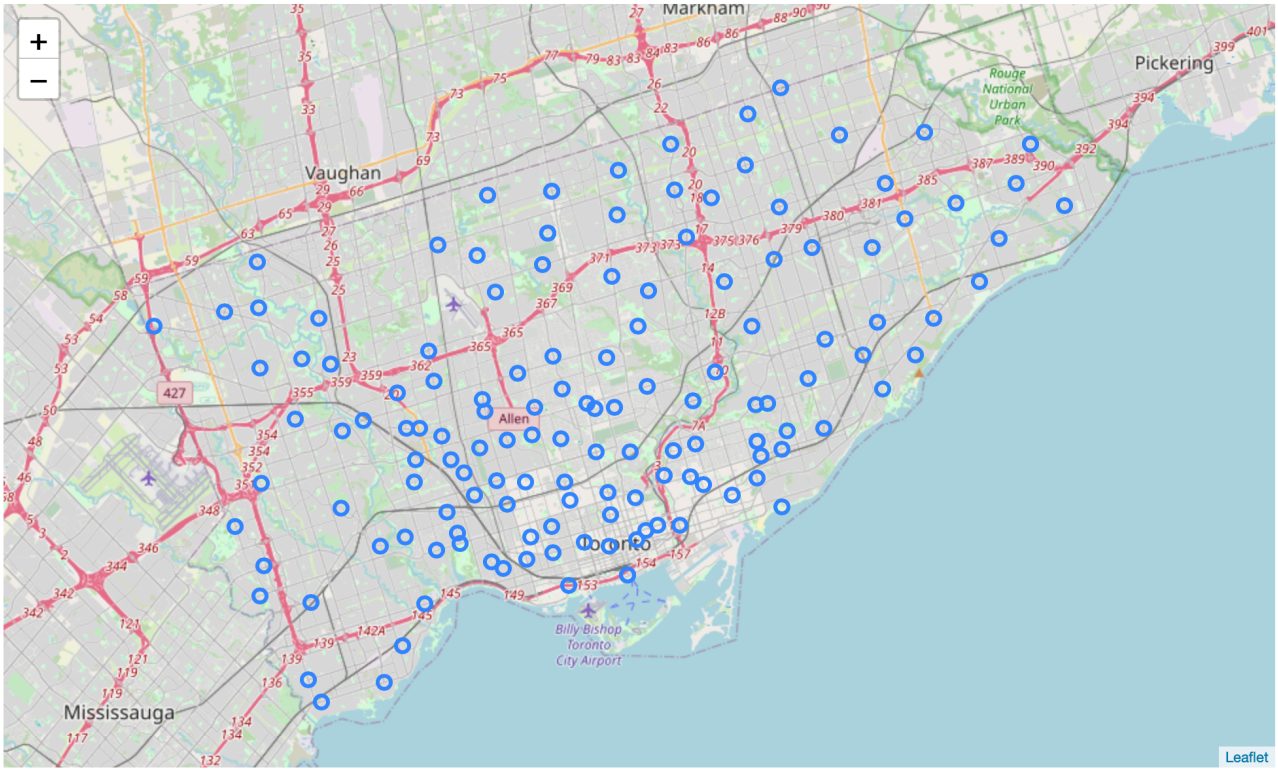


As expected, with a vibrant real estate market, we observed that more than half of all neighbourhoods had 50%+ of their listings sold above the asking price. Our Top 5 neighbourhoods all had 80%+ of their listings sold above asking price. Neighbourhoods with a higher Solve above asking price percentage is ranked higher, and those with fewer sold above asking price are given a lower rank.

After exploring the three features that will constitute our Housing Price Aggregate, we sum up the Rankings each neighbourhood achieved in each category to calculate the final score. We chose to use the rankings rather than the actual percentage of prices, as this offers us the opportunity to normalise our data. In order to achieve a high score in our Housing Price Aggregate, a neighbourhood would need to have:

$$LowAvgSalePrice + HighPercentSoldUnder10Days + HighPercentSoldAboveAskingPrice$$

With the list of Neighbourhoods, we further process this through www.gps-latitude-longitude.com to find the corresponding GPS coordinates. As not everyone will be familiar with all the neighbourhoods within Toronto, it is always a good practice to visualise our neighbourhood listing against a map of Toronto. To do this, we make use of Folium maps.



3.2 Quality of Living Index

As we described before, once we have our list of neighbourhoods and their corresponding GPS coordinates, we are then able leverage FourSquare's Places API. Using the Explore endpoint we query the number of essential venues within one kilometre of each neighbourhood. By specifying the types of venues we are interested in, we are able to count the number of venues by categories outlined in Chapter 1.

Some basic data cleansing needed to be performed on our DataFrame as some neighbourhoods did not contain venues corresponding to our five categories.

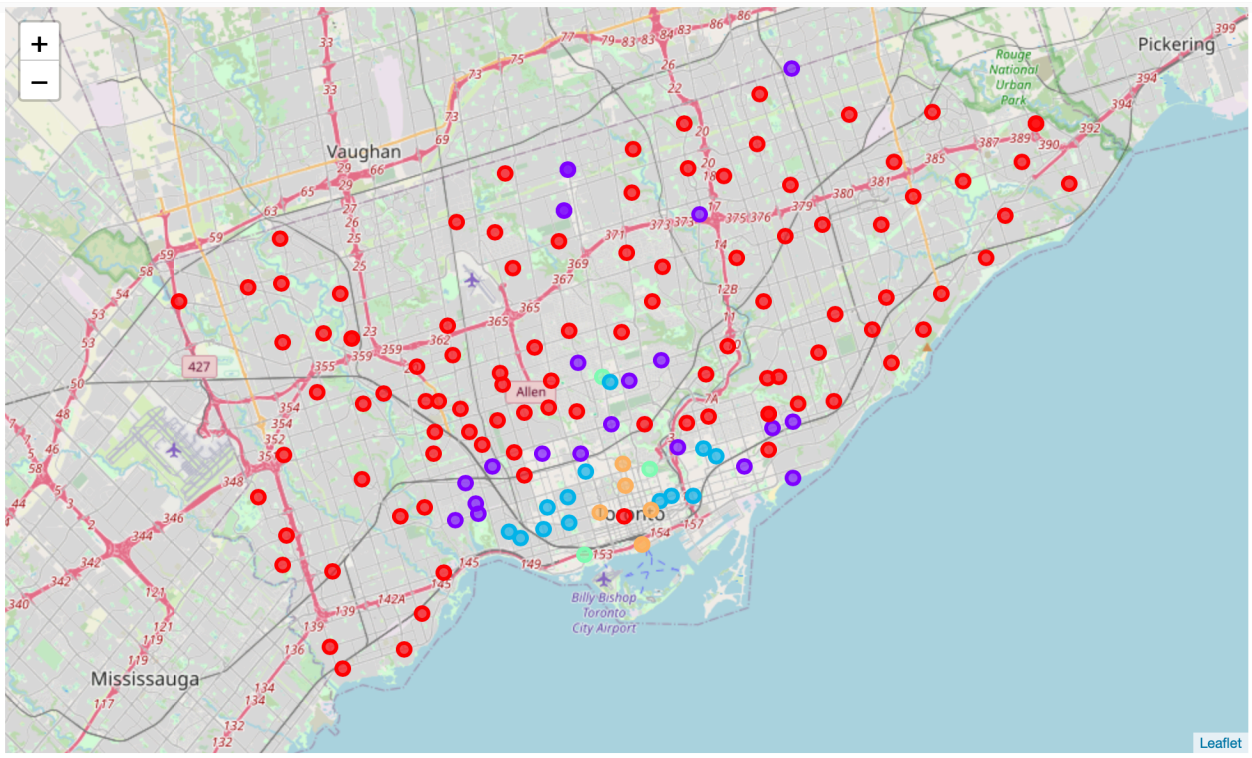
For simplicity, we sum up the number of each type of venue to calculate a final "Quality of Life" index (QOL).

	Neighborhood	Medical	Shops	Food	Recreation	Arts	QOL
0	Bridle Path-Sunnybrook-York Mills	4	1	6	2	1	14
1	Kingsway South	6	22	24	5	7	64
2	Lawrence Park South	7	34	38	18	1	98
3	Casa Loma	14	37	50	20	5	126
4	Forest Hill North	8	17	26	10	3	64

Now with our Quality of Life index calculated for each neighbourhood, want to cluster them into five different groups depicting a range between "Many Venues" to "Few Venues". In order to do this, we take the calculate five bins of equal size and classify which bin a neighbourhood belonged to.

	Neighborhood	Medical	Shops	Food	Recreation	Arts	QOL	Cluster
0	Bridle Path-Sunnybrook-York Mills	4	1	6	2	1	14	0
1	Kingsway South	6	22	24	5	7	64	0
2	Lawrence Park South	7	34	38	18	1	98	1
3	Casa Loma	14	37	50	20	5	126	1
4	Forest Hill North	8	17	26	10	3	64	0

Together with our clusters, we once again map our QOL clusters against the backdrop of the map of Toronto to better understand the distribution venues by neighbourhood.



From the above map, we make several observations. As we would normally expect, neighbourhoods within Downtown Toronto contained the highest number of essential venues. As we move towards the outskirts of Toronto, the number of venues start to decline (depicted by Red markers), with only few notable exceptions.

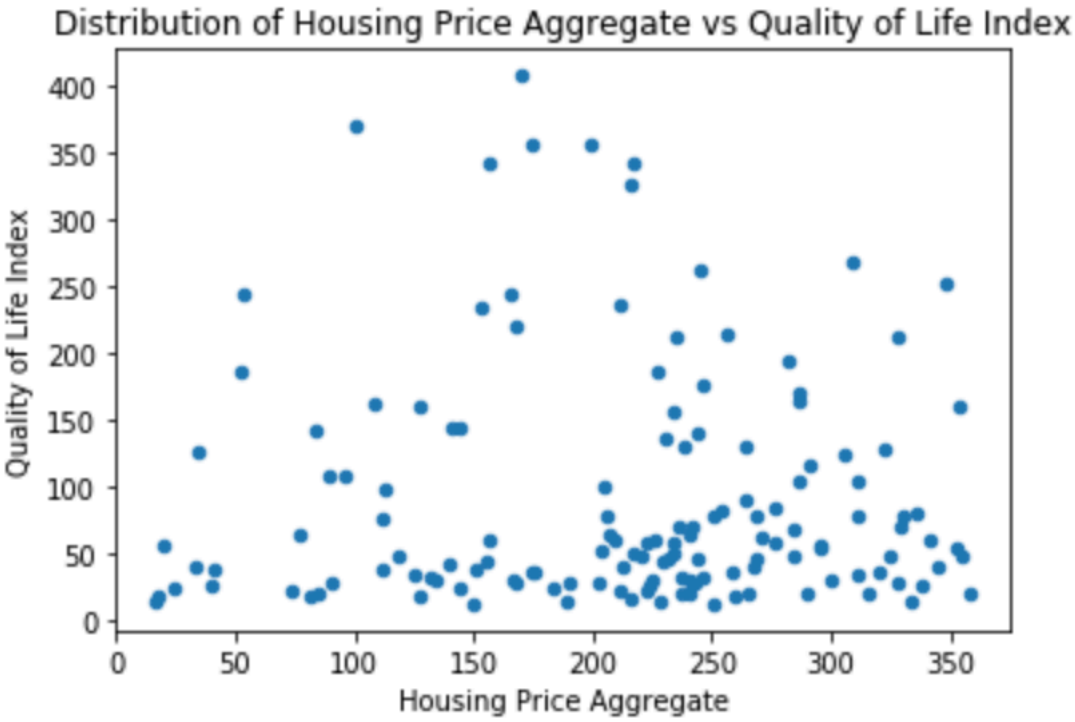
At this stage, we have finally cleansed and explored our data into a single and am able to merge everything into a single DataFrame. As a recap, this Dataframe now consists of:

- A list of Neighbourhoods in Toronto
- *Housing Price Aggregate* - Housing price information of each Neighbourhood (e.g. Average Sale Price)
- GPS coordinates of each Neighbourhood in Toronto
- *Quality of Life Index* - A count of the number of Medical, Shops, Food, Recreation and Arts venues within 1km of each Neighbourhood

Chapter 4 - Results

As we stated at the very beginning, in order to find the “Best place to live in Toronto”, we want now correlate our Quality of Life Index with our Housing Price Aggregate and be able to finally come to a recommendation.

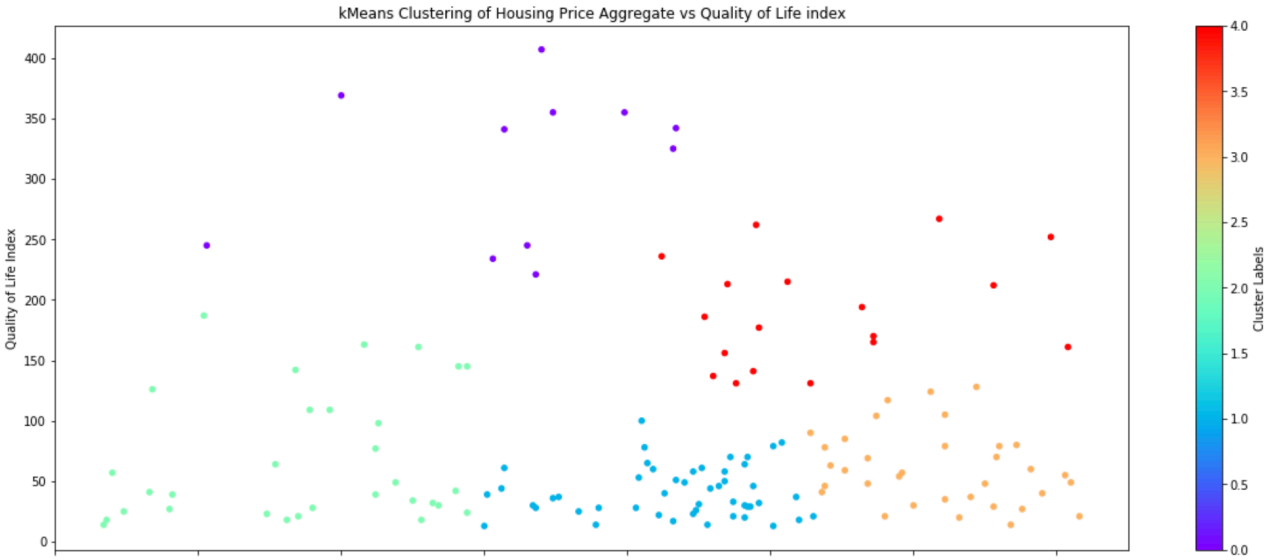
We do so by first correlating the two leading factors and plot a scatter plot to understand the distribution of Neighbourhoods.



Initial observations of our scatter plot shows there is not a direct nor linear relationship between our Housing Price Aggregate and Quality of Life index. In fact, we see a near wide range of Housing Price Aggregate all with low Quality of Life index. Meanwhile, neighbourhoods that scored highest in our Quality of Life index is not necessarily the most expensive.

Machine Learning Model
Without any clear direct correlation, we decide not to proceed with supervised machine learning models such as Linear/Polynomial/Logistic Regression. Also as we are not attempting to classify our neighbourhoods, other supervised machine learning models such as SVM or Gradient Descent also does not suit our purposes. Instead, we opt to use the unsupervised learning model - kMeans clustering to group our neighbourhoods into different clusters based on our two indices.

We run kMeans clustering with a cluster size of five to try to group our neighbourhoods into five different clusters. We plot once again our scatter plot to better understand our distribution.



Next, let's look into the different clusters to understand the results.

Cluster 0 can be summarised as those neighbourhoods with:
- Highest Quality of Living (Good)
- Modest Housing Price Aggregate (Balanced)

	Aggregate	QOL	Cluster Labels	Neighborhood	Latitude	Longitude
80	217	342	0	Moss Park	43.654789	-79.372602
82	216	325	0	Kensington-Chinatown	43.653943	-79.400403
92	199	355	0	Waterfront Communities C8	43.641648	-79.377921
98	174	355	0	Waterfront Communities C1	43.641648	-79.377921
99	170	407	0	University	43.652420	-79.387101

Cluster 1 represents the neighbourhoods with the:
- Lowest Quality of Living index (Bad)
- Modest House Pricing Aggregate (Balanced)

	Aggregate	QOL	Cluster Labels	Neighborhood	Latitude	Longitude
40	265	21	1	Humbermede	43.738994	-79.539408
43	260	18	1	Willowridge-Martingrove-Richview	43.676158	-79.569572
44	259	37	1	Rexdale-Kipling	43.719857	-79.570600
46	254	82	1	Corso Italia-Davenport	43.677370	-79.446128
47	251	13	1	West Humber-Clairville	43.735649	-79.625844

Cluster 2 represents neighbourhoods with:
- Low Quality of Living (Bad)
- Poor Housing Price Aggregate (Bad)

	Aggregate	QOL	Cluster Labels	Neighborhood	Latitude	Longitude
110	144	145	2	Willowdale East	43.771171	-79.419750
111	144	24	2	Black Creek	43.697015	-79.486948
112	141	145	2	Willowdale West	43.771171	-79.419750
113	140	42	2	Downsview-Roding-CFB	43.726699	-79.482061
114	134	30	2	Hillcrest Village	43.804877	-79.354690

Cluster 3 represents neighborhoods that have a:
- Relatively lower Quality of Living (Bad)
- High Housing Price Aggregate (Good)

	Aggregate	QOL	Cluster Labels	Neighborhood	Latitude	Longitude
0	358	21	3	Humberlea-Pelmo Park W5	43.721319	-79.533217
1	355	49	3	Agincourt North	43.808053	-79.266502
3	353	55	3	Ionview	43.730824	-79.273900
5	345	40	3	Woburn	43.776470	-79.231728
6	341	60	3	Agincourt South-Malvern West	43.789964	-79.242296

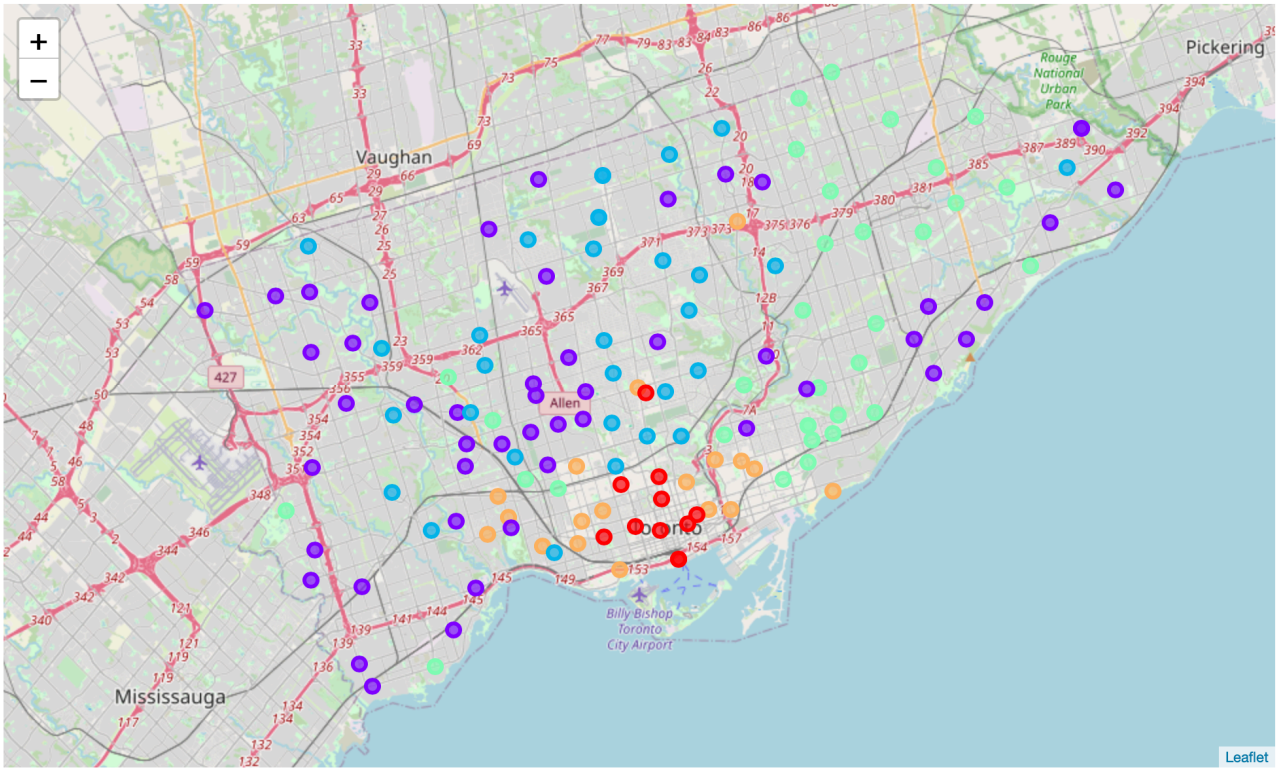
and finally Cluster 4 represents those neighborhoods that have a:
- Modest Quality of Life Index (Balanced)
- High Housing Price Aggregate (Good)

	Aggregate	QOL	Cluster Labels	Neighborhood	Latitude	Longitude
2	354	161	4	Henry Farm	43.769583	-79.346524
4	348	252	4	North St. James Town	43.670867	-79.373306
12	328	212	4	South Riverdale	43.660603	-79.350340
21	309	267	4	Niagara	43.637383	-79.408493
29	286	170	4	Roncesvalles	43.646317	-79.449068

Summarising our results, we can label our Clusters as below and narrow down our recommendation to those Clusters that ranked favourably:

Cluster	Quality of Life Index	Housing Price Aggregate	Recommended?
Cluster 0	Good	Balanced	Yes
Cluster 1	Bad	Balanced	No
Cluster 2	Bad	Bad	No
Cluster 3	Bad	Good	No
Cluster 4	Balanced	Good	Yes

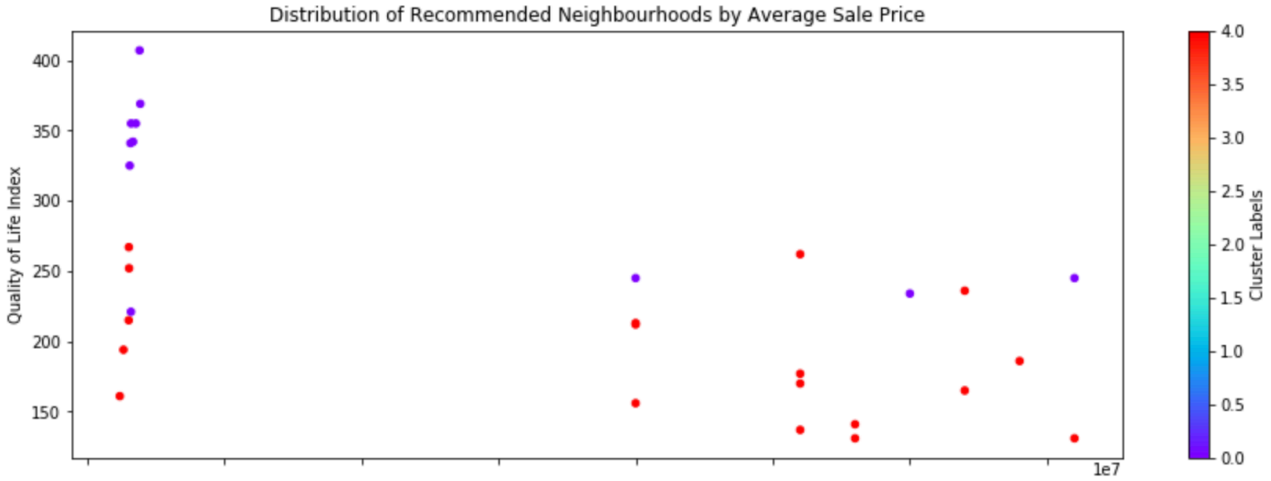
With our final Clustering achieved, we want to see which neighbourhoods we would recommend and which we wouldn't on the map.



Cluster 0 - Red - Recommend
Cluster 1 - Purple,
Cluster 2 - Blue,
Cluster 3 - Green,
Cluster 4 - Orange - Recommend

Chapter 5 - Discussion

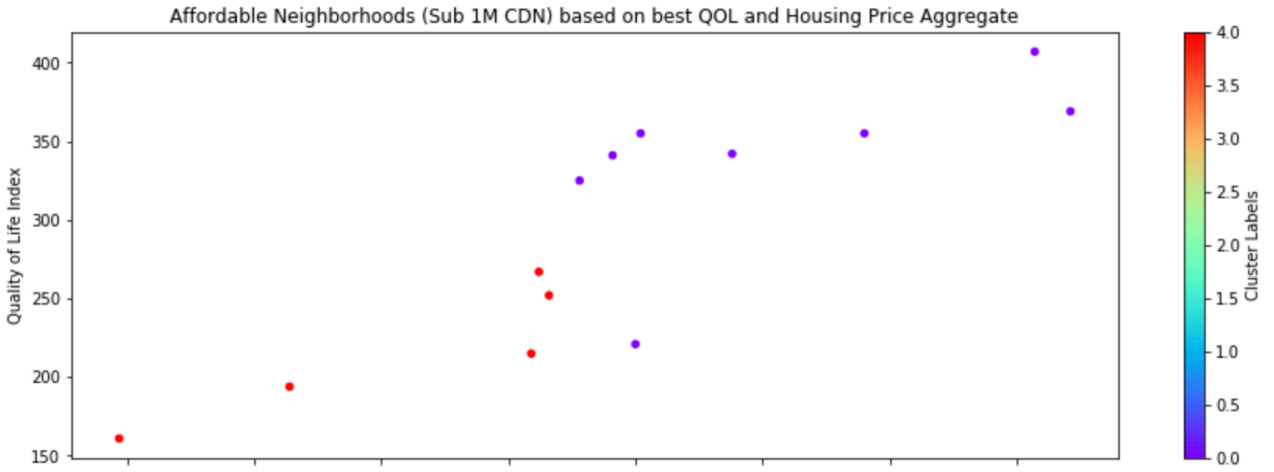
With our recommended Neighbourhoods narrowed down to those belonging in Cluster0 and Cluster 4, we can take a deeper look to further refine where is the best place to live in Toronto. While having a good Quality of Life, and having real estate that has a strong price increase potential is great, the last question that hasn't been answered yet is "Can you afford it?". In our discussion here, we want to dig deeper into this final factor before coming to our top recommendations of neighbourhoods.



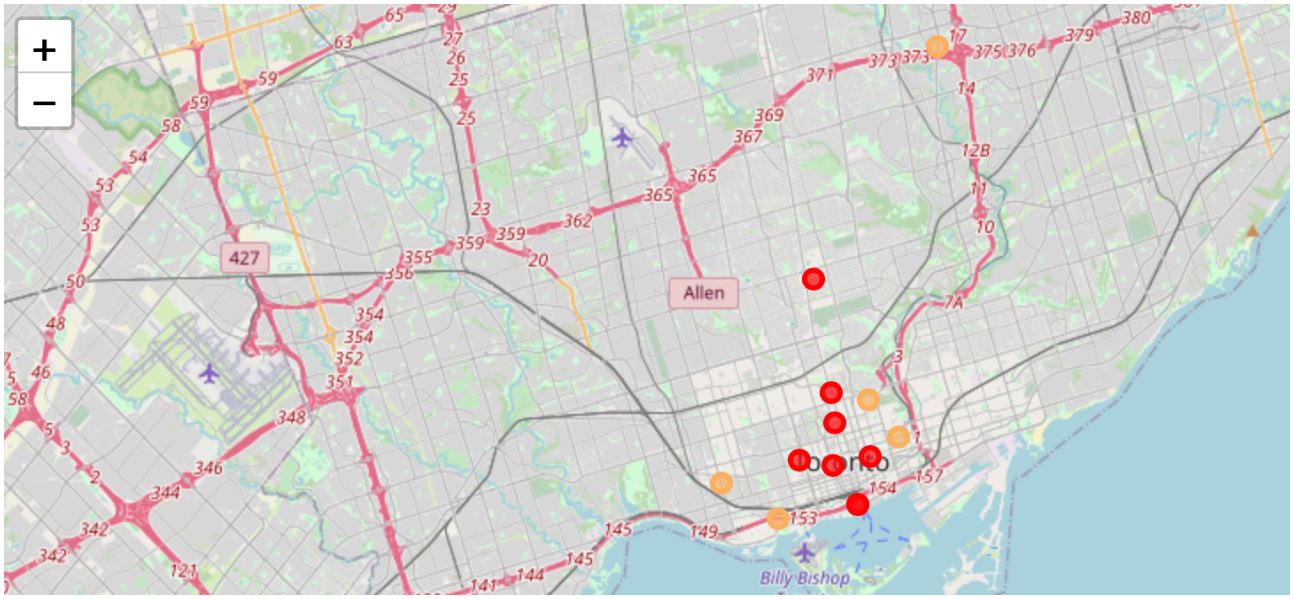
Our scatter plot shows a very interesting result. There is once again a large disparity between neighbourhoods with an average sale price below 1M CDN, and those that are above 10M CDN. This seems to suggest that roughly half of the neighbourhoods we would normally recommend have been well recognised as some of the best places to live in Toronto, reflected by the high average prices. As most people could not easily afford housing that costs 10M+ CDN, we shall refer to these as the luxury market.

Even more interesting are the remaining neighbourhoods that offer similar (if not better) Quality of Life but available at a much lower price range (sub 1M CDN). These seem to be neighbourhoods comparatively undervalued.

Isolating the “Sub 1M CDN” market, we see a linear relationship between the average sale price and the Quality of Life of a neighbourhood.



Let's further plot on a map where our shortlisted Neighbourhoods reside.



Our final recommendation to the best place to live in Toronto narrows down to a decision based on affordability.

Recommendation 1 Sub 1M Neighborhoods with highest QOL & Housing price aggregate

	Neighbourhood (# Rank out of 143)	Average sale price
92	Waterfront Communities C8	890000
99	University	957000
124	Bay Street Corridor	971000

Recommendation 2 Most affordable Sub 1M Neighborgoods with highest QOL & Housing price aggregate

	Neighbourhood (# Rank out of 143)	Average sale price
2	Henry Farm	597000
33	Regent Park	664000
45	Little Portugal	759000

Chapter 6 - Conclusion & Next Steps

Based on our analysis the best Neighbourhoods to settle down on depends on what you are looking for. Our analysis results recommend Neighbourhoods that fall under our "Cluster 0" or "Cluster 4" classifications. In general, they have a net positive rating in terms of both Quality of Life and scored considerably well in our Housing Price Aggregate.

It is worthwhile to highlight here again that our Housing Price Aggregate takes into consideration of how fast listings are being sold, and how many listings were sold above the asking price. This gives us a measure into how attractive a neighbourhood is in general and serves as an indication of a good investment. The average price of the neighbourhood is also factored into this aggregate where we favoured more affordable housing over the luxury market.

Meanwhile, our analysis also interestingly described where not to invest. This could be classified as those neighbourhoods under Cluster 2 and 3. These are often neighbourhoods where we either have very few essential venues or did not comparatively fair well in terms of our housing price aggregate. This does not mean they are not good neighbourhoods to live in, but reflective of only a comparative measure based on how we have defined our indices.

To better our analysis, it would have been worthwhile to further dive into how we should define the "Quality of Life" index. In our analysis, we made a very crude approximation based on the number of essential venues within a one kilometre radius from the centre of each neighbourhood. This made many approximations that may have biased our analysis:

- Neighbourhoods are seldom circular
- The centre of some neighbourhoods are very close to each other leading to overlaps between the 1km radius
- Many people in Canada drive a cars. Is 1km a meaningful radius given how accessible other parts of a city are?

Furthermore, different people have different life styles. Some prefer to live within the heart of the city where the lifestyle is vibrant with many venues and attractions. Meanwhile others prefer to live in a quiet neighbourhood with a close proximity to parks and nature. As a result, dependant on your personal life choices the recommendation to be made can vary. Instead of making assumptions and a generic recommendation, the analysis could be better tailored by taking input from the reader to adjust weights given to different features.

Lastly to better improve our analysis, further datasets could be used to better understand how additional features could influence our results. The City of Toronto via it's Open Data Portal offers many details of each neighbourhood that could be further factored in (https://open.toronto.ca/dataset/neighbourhood-profiles/) This includes for instance:

- Housing profiles within each neighbourhood (number of condos, number of semi-detached houses, etc.)
- Ethnicity and language spoken most in each neighbourhood
- Marital status and family size
- Education level and breakdown
- Income level