Data Science Capstone - Report

Exploring the Relationship between Housing Prices in Toronto and Venues in the Neighborhood

ANNA CYBULSKY

JUNE 2020

Introduction/Business Problem

The target audience of the problem addressed in this project are prospective house buyers in Toronto, who would like to understand the factors that impact housing prices, and in particular with a focus on what kind of venues are available in the neighborhood. We will explore whether available nearby venues may affect the price of a house in a given neighborhood.

This problem is of interest because the housing market in Toronto is very competitive and is clearly a sellers market (even with the effects of the Covid-19 crisis). Houses with multiple offers often sell for significantly above the asking price. Therefore, individuals with a set budget who are looking to buy a house may have to make some compromises, for example with regards to the location and the venues available in close proximity.

It will be important to first explore how the price of houses may be correlated with regards to key characteristics such as the surface area or the number of bedrooms, bathrooms, and type of house. Then we will be able to better evaluate the impact of the neighborhood venues on housing prices.

Data

The first dataset required is the list of Toronto Neighborhoods with their coordinates, created in the previous Capstone course assignment.

The second dataset is a list of houses in Toronto with their selling price and key characteristics, such as the address, the living area, number of bedrooms, and so on. Using the address, the coordinates will be added to the dataset and then each house will be assigned a cluster corresponding to the Toronto neighbourhood.
The dataset was handcrafted in another project on GitHub by searching through a real estate website's data and extracting it into a dataframe.
The dataset was created by slavaspirin:
https://github.com/slavaspirin/Toronto-housing-price-prediction/blob/master/

It includes:

- House sale price
- House listing price
- Address
- Number of bedrooms
- Number of bathrooms
- Number of parking spots
- Type (detached, semi-detached…)
- Latitude & longitude coordinates
- Neighborhood

The third dataset is the Foursquare venues data, which will be used to characterise the venues in close proximity to the house / its neighborhood. The work will group the venues into a smaller

number of categories, such as restaurants, bars, schools, parks, shops, grocery stores and so on, which may add to the value of the house if in close proximity.
It will include:

- Venue name
- Venue category
- Venue coordinates

Methodology

In the first notebook (Part 1), I clean the dataset and then explore the data to see if any correlations can be visually observed.

Cleaning
First, I removed condos data from the dataset to focus only on houses.

Second, I cleaned up the row data to make analysis easier. For example, the column “bedrooms” had strings in it along with numbers, so the strings were removed. It also contained half-bedrooms, so these were converted to floating values.

Before cleaning:

	final_price	list_price	bedrooms	bathrooms	sqft	parking	type	full_address	lat	long	city_district
0	855,000	$870,000	2 + 1 beds	2 baths	800–899 sq. ft.	1 parking	Condo Apt	38 Grenville St, Toronto	NaN	NaN	NaN
1	885,000	$898,000	3 beds	2 baths	N/A sq. ft.	6 parking	Semi-Detached	2 Cabot Crt, Toronto	NaN	NaN	NaN
2	550,000	$549,900	1 beds	1 baths	500–599 sq. ft.	no parking	Condo Apt	30 Roehampton Ave, Toronto	NaN	NaN	NaN
3	665,000	$600,000	1 + 1 beds	1 baths	600–699 sq. ft.	1 parking	Condo Apt	65 East Liberty St, Toronto	NaN	NaN	NaN
4	825,513	$839,000	2 beds	2 baths	N/A sq. ft.	1 parking	Detached	61 Twelfth St, Toronto	NaN	NaN	NaN

After cleaning:

	final_price	list_price	bedrooms	bathrooms	sqft	parking	type	full_address	lat	long	city_district
1	885000.0	898000.0	3	2	0	6	Semi-Detached	2 Cabot Crt, Toronto	NaN	NaN	NaN
4	825513.0	839000.0	2	2	0	1	Detached	61 Twelfth St, Toronto	NaN	NaN	NaN
6	2700000.0	2798000.0	4	5	2500–3000	2	Detached	110 Albertus Ave, Toronto	NaN	NaN	NaN
8	975000.0	954900.0	2	2	1100–1500	1	Duplex	182 Broadway Ave, Toronto	NaN	NaN	NaN
10	1057000.0	1079000.0	3.51	2	0	1	Semi-Detached	342 Indian Road Cres, Toronto	NaN	NaN	NaN

Unfortunately the house square foot data was too often unavailable, which means it could not be used later on to normalise the pricing data with respect to the house surface area / living space.

Exploring

The following plots show the data exploration that was done to visually see whether any of the features seemed to have a strong correlation with the house price.

The following features were explored:

- bedrooms
- bathrooms
- square-feet (for available data)
- parking
- house type

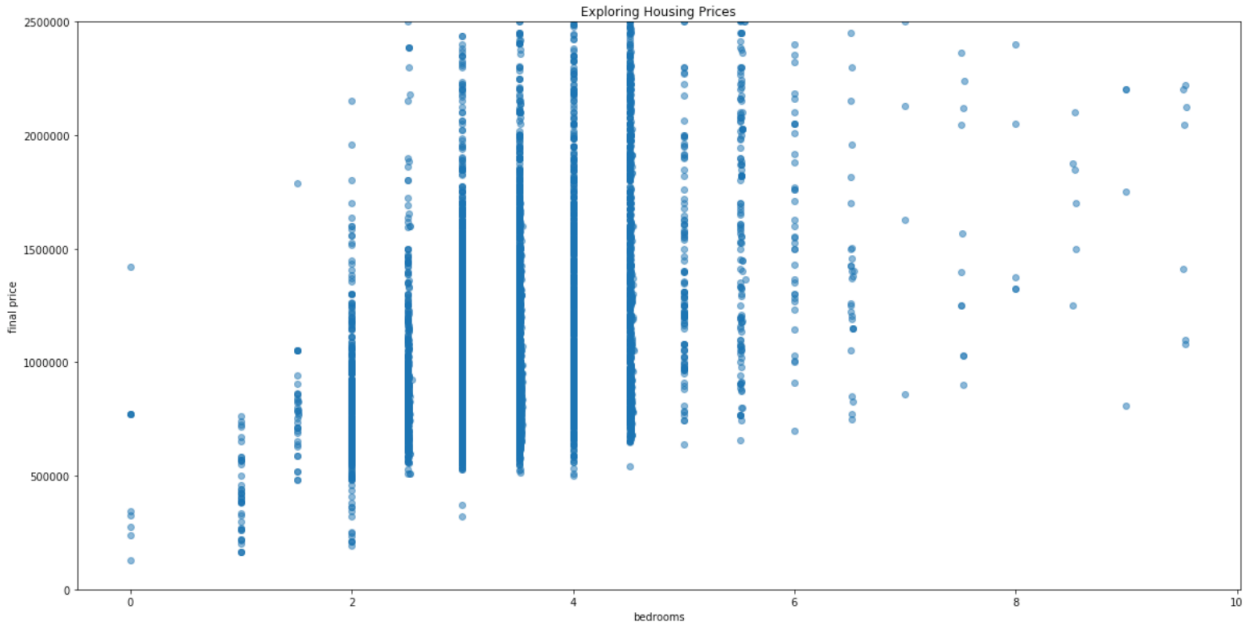
Note that a y-axis limit was set on the house price for most of the plots, since some of the data maximum outliers go even beyond 1 billion dollars.

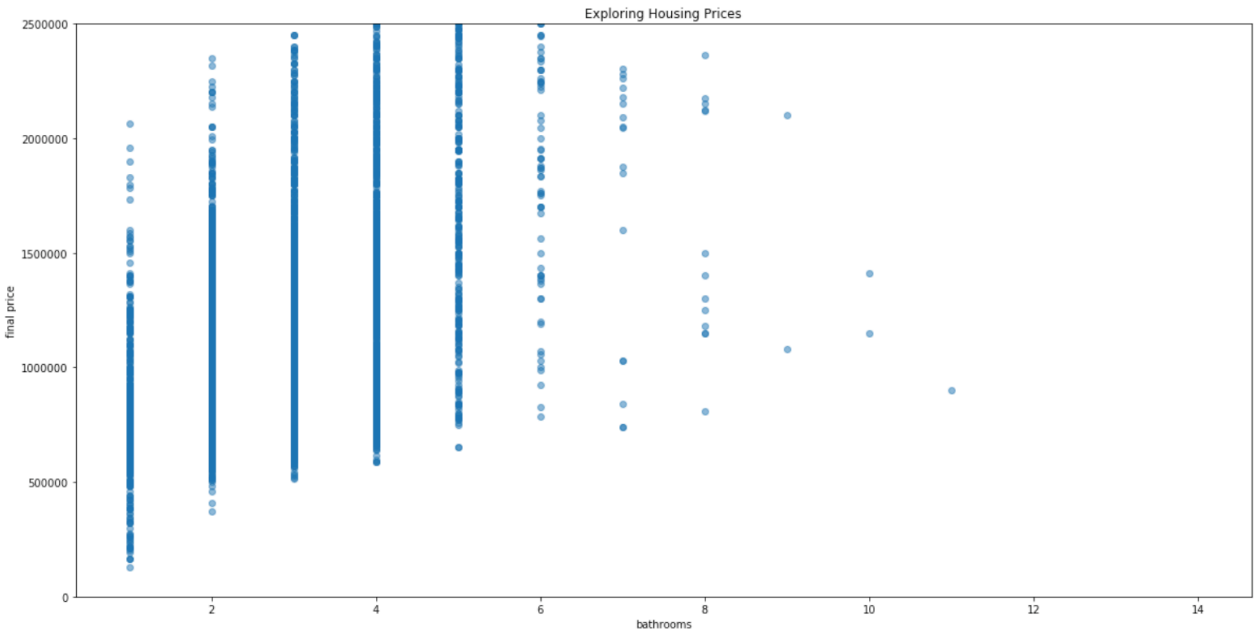
The strongest correlation appears to be for living area (squarefeet), then bathrooms and bedrooms. However, for bedrooms, the relationship only looks linearly proportional in the beginning and then the curve seems to flatten out with a very wide price range.

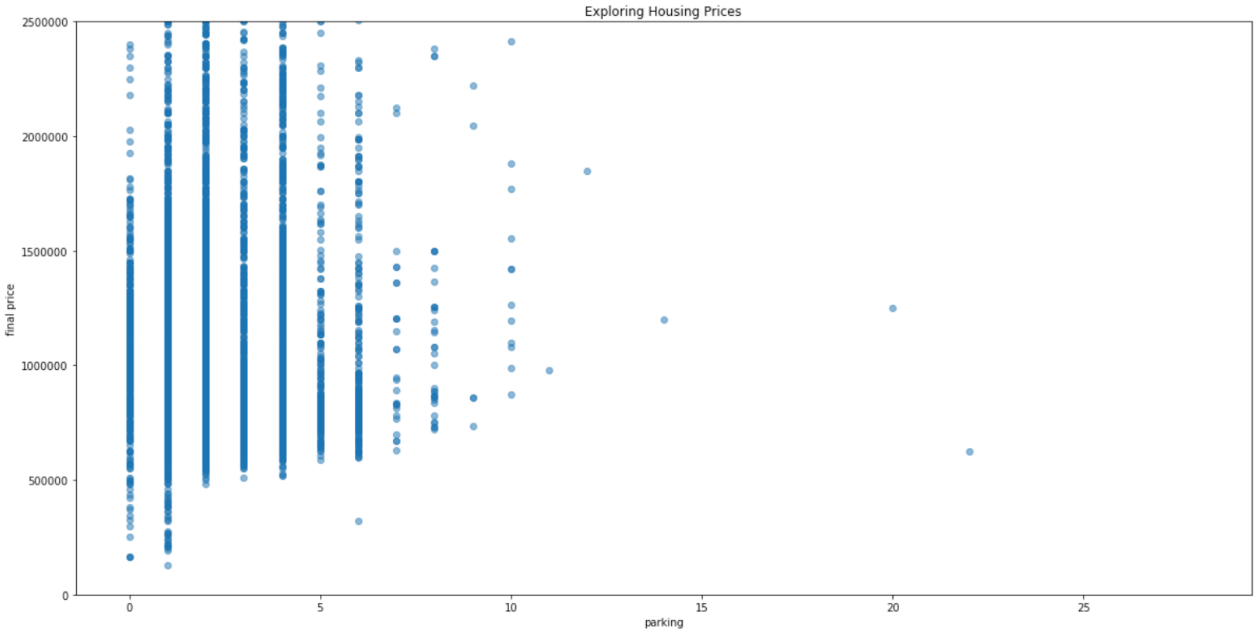
With regards to house type, differences can be noticed, but due to the low number of houses in many of the categories, only the three largest categories were selected for further analysis. This categorical variable is then converted to an integer type to ease further analysis, according to the legend below:

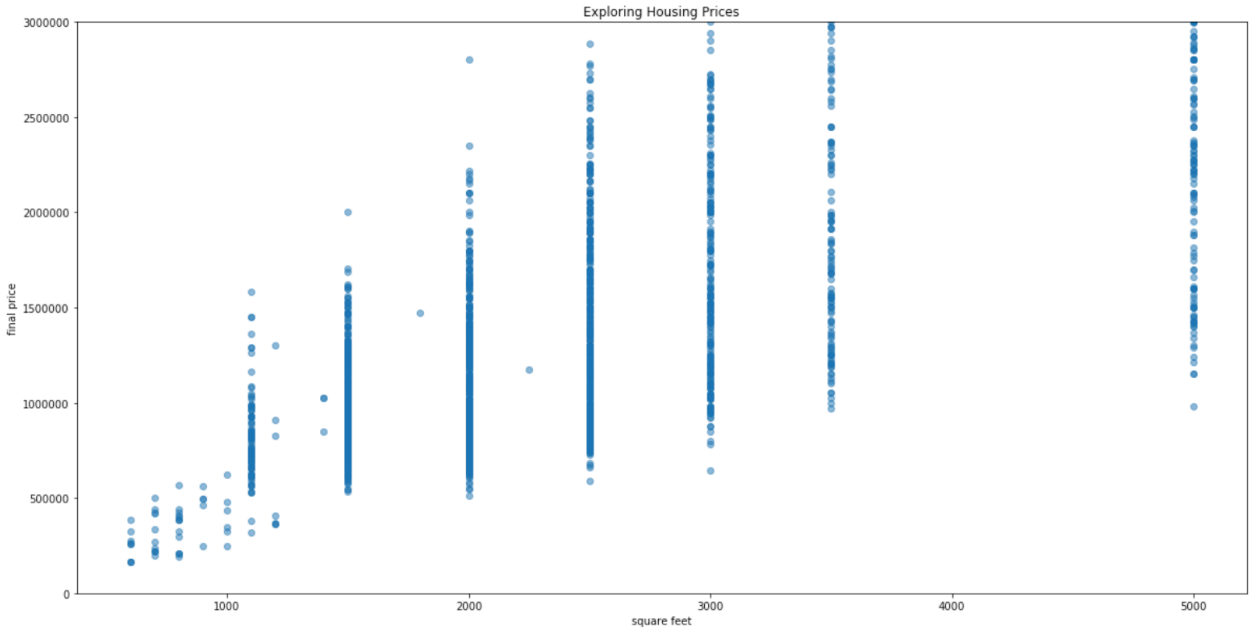
- Att/Row/Twnhouse = 1
- Semi-Detached = 2
- Detached = 3

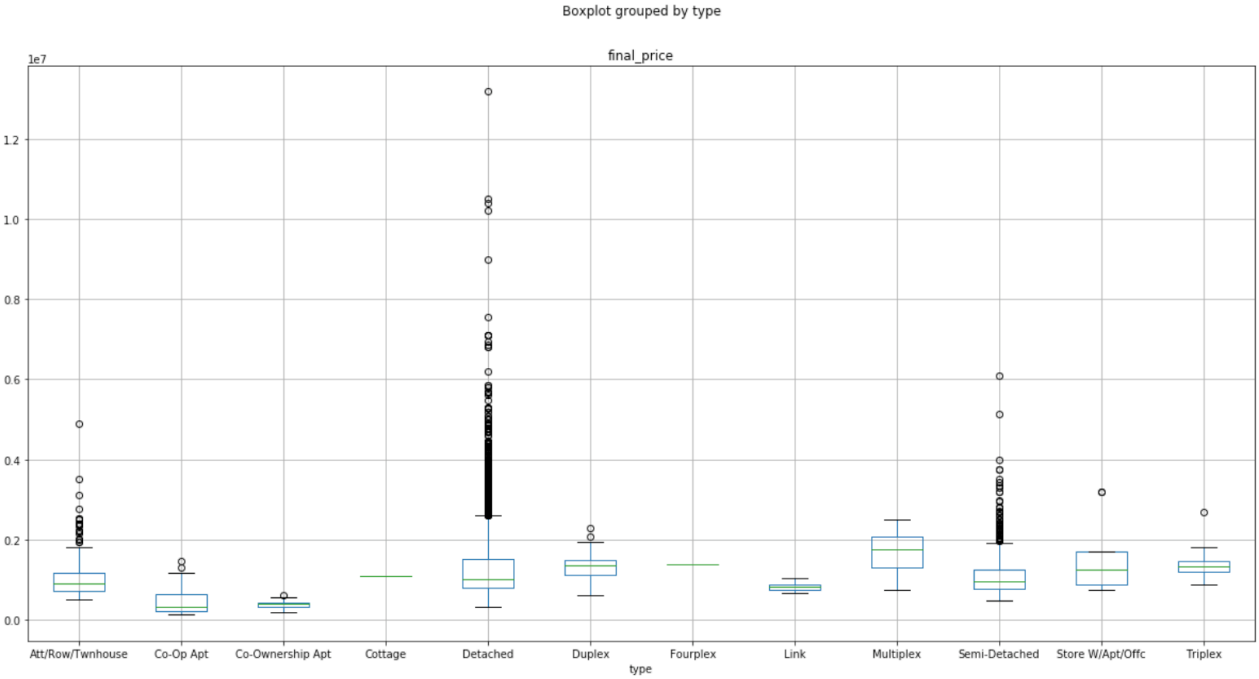
Finally the cleaned dataset is exported to begin further analysis in the Part 2 Notebook.

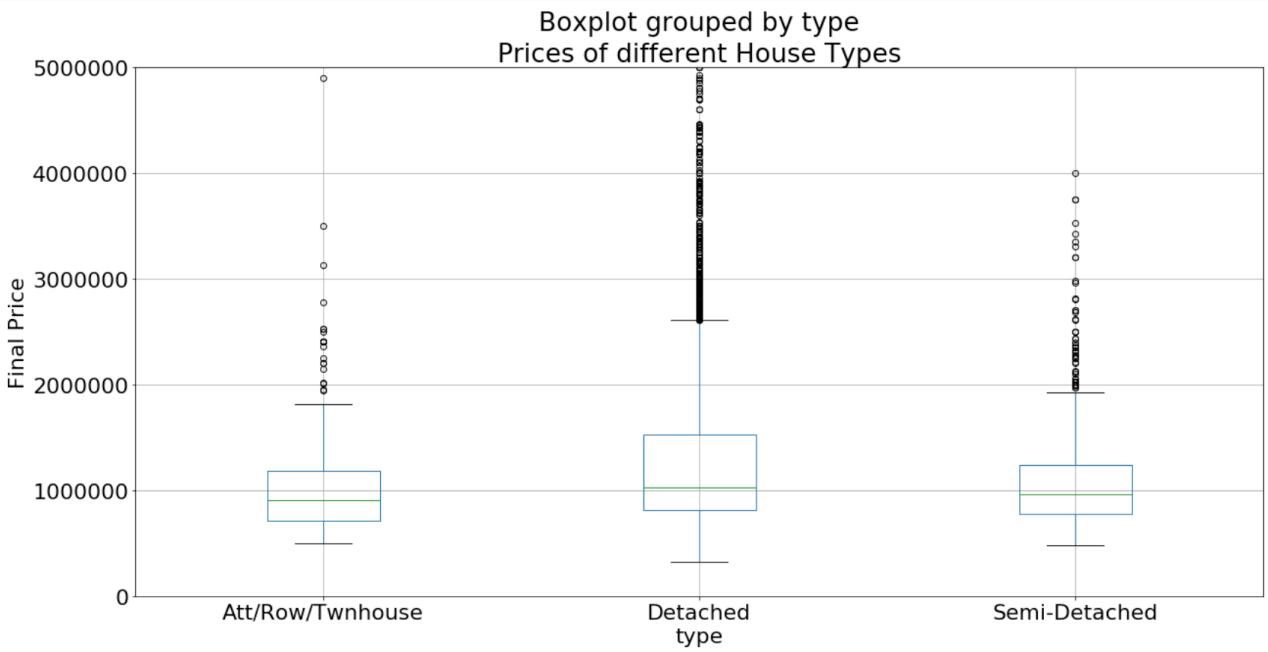



Exploring Housing Prices
final price
bathrooms
0
500000
1000000
1500000
2000000
2500000
2
4
6
8
10
12
14


Exploring Housing Prices
final price
parking
0
500000
1000000
1500000
2000000
2500000
0
5
10
15
20
25


Exploring Housing Prices
final price
square feet
0
500000
1000000
1500000
2000000
2500000
3000000
1000
2000
3000
4000
5000


Boxplot grouped by type
final_price
1e7
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Att/Row/Twnhouse
Co-Op Apt
Co-Ownership Apt
Cottage
Detached
Duplex
Fourplex
Link
Multiplex
Semi-Detached
Store W/Apt/Offc
Triplex
type

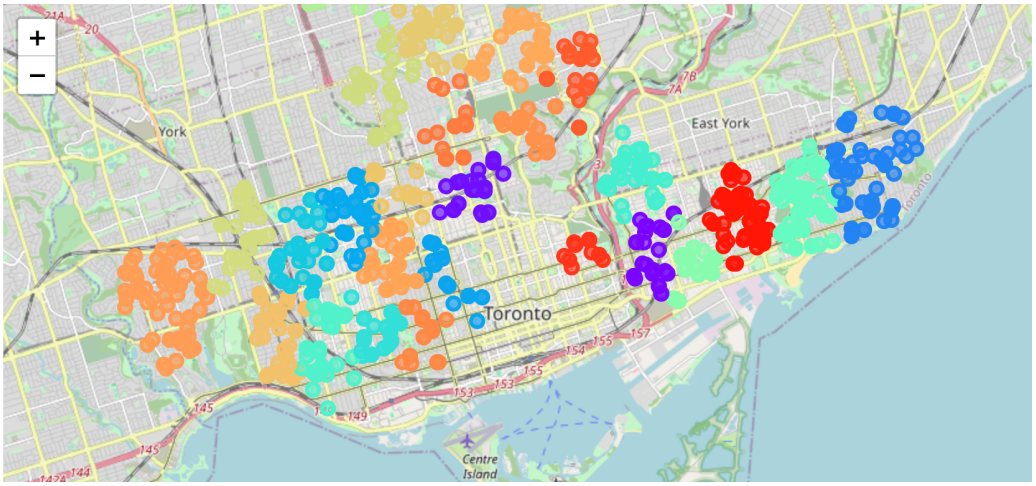


Analysis / Clustering

In the second notebook (Part 2), the following steps are performed:

1. Cluster the houses into neighborhoods
2. Explore price data
3. Use Foursquare API to get venues for each neighborhood
4. Significantly reduce the number of venue categories
5. Cluster the neighborhoods based on venues
6. Cluster the neighborhoods based on housing data
7. Explore results

The houses were clustered into neighborhoods using the K Means algorithm, and using the neighborhood latitude and longitude coordinates as the centroids for the algorithm. Once the houses were clustered, the neighborhoods that contained data for less than 30 houses were dropped from the analysis to ensure that the results would be statistically significant. A map was created to visualise the houses' location with their colored neighborhood cluster to check that the algorithm was set up correctly.



Next, it was checked whether there were any strong correlations between the housing data and the house characteristics to see whether normalizing the data with respect to one or two features made sense, before looking for a correlation with regards to the neighborhood venues.

	final_price	list_price	bedrooms	bathrooms	parking	sqft	type
final_price	1.000000	0.988715	0.520764	0.714395	0.434566	0.038079	0.361581
list_price	0.988715	1.000000	0.508658	0.716424	0.442706	0.035461	0.371733
bedrooms	0.520764	0.508658	1.000000	0.576177	0.299215	0.034831	0.269378
bathrooms	0.714395	0.716424	0.576177	1.000000	0.408849	0.057405	0.278976
parking	0.434566	0.442706	0.299215	0.408849	1.000000	0.032546	0.337480
sqft	0.038079	0.035461	0.034831	0.057405	0.032546	1.000000	-0.030106
type	0.361581	0.371733	0.269378	0.278976	0.337480	-0.030106	1.000000

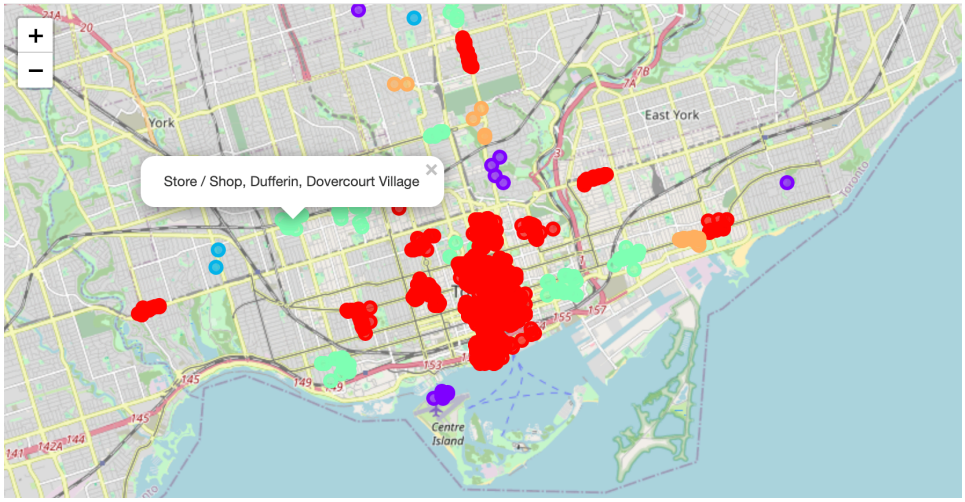
However, not having found any strong correlations (pearson correlation coefficient > 0.75), the dataset was left as is for clustering.

For the next step, the Foursquare API was used to get a list of venues for each neighborhood in Toronto (excluding neighborhoods that are not in the Toronto Borough).

Then the venue categories are transformed to be much more general, since in the default data the number of unique venue categories is more than 200. For example, the specific kind of cuisine a restaurant has to offer is not important - it is simply interesting whether there are restaurants in the neighborhood. The number of categories was reduced to about a dozen, including an “Other” category for venues that were in “rare” categories.

Next, the neighborhoods were clustered using K Means based on the type of venues available. Unfortunately, as was also seen in the previous assignment (using the default specific venue categories), many of the neighborhoods are very similar - offering restaurants, bars, shops, etc - creating one cluster with most neighborhoods. One of the clusters seems to be an anomaly from lack of data, as there are two neighborhoods containing only parks.

A partial map was created to visualize the clustering results. Each dot represents a venue and the colors represent the different neighborhood clusters. The label indicates the venue type (general category) and the neighborhood.



Afterwards the neighborhoods are clustered using K Means based on the housing data, excluding surface area since this data was unavailable for many houses.

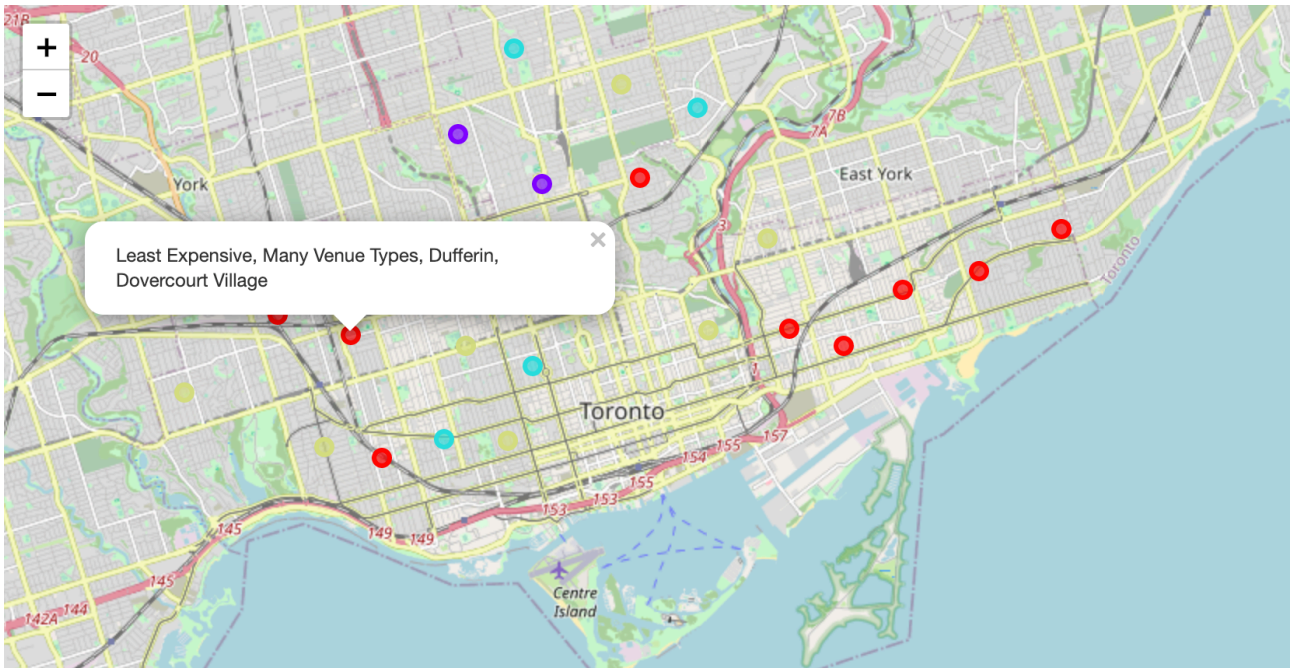
Finally, the venue cluster and housing cluster results for each neighborhood are combined into a single dataset, and the correlation between the House Price Clusters and the Neighborhood Venue Clusters is calculated. The results are also visualized on a map with colors and labels to observe whether any correlation is evident.

Results

The results indicate that there is a low correlation ~0.31 between the house (price) cluster and venue cluster for Toronto neighborhoods. The final map in the notebook allows to visualise the house clusters with the different colors, and the label indicates the venue cluster.

For example, we see the Dufferin / Dovercourt Village neighborhood is in the least expensive house cluster, and is also in the neighborhood cluster with many venue types.
We can see that there are also “Medium Expensive” neighborhoods in the “Few Venues” cluster.

Therefore, the relationship between housing price and venues in the neighborhood is weak. However, this analysis can help prospective house buyers choose a neighborhood in their price range, while considering the kind of venues that they would like to have nearby.



Discussion

It is important to note that the venue-based clusters are not distributed very evenly - meaning that there is one cluster that contains over half the neighborhoods and three clusters with a small number of neighborhoods (see image below). Compared to the previous assignment, where the clustering produced one cluster containing nearly all the neighborhoods, this is an improvement. This result was obtained by vastly reducing the number of venue categories - from the initial 211 to 14 categories (see image below). Unfortunately, there were still 152 venues that fell into the “Other” category, when the original venue categories were too specific/unique to easily generalise them, and therefore we lose these venues as a feature of differentiation.

Cluster	Neighborhood
0	20
1	4
2	2
3	8
4	3

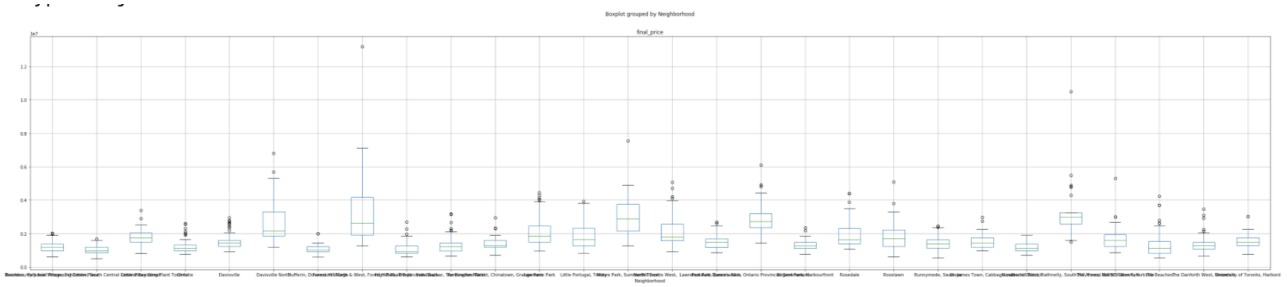
```
venues['General Category'].value_counts()
```

```
Restaurant       458
Store / Shop     281
Other            152
Bar / Pub        116
Café / Tea        85
Gym               36
Hotel             27
Bakery            25
Park              22
Deli / Bodega     16
Bank              13
Art Gallery        8
Concert Hall       7
Pharmacy           6
Name: General Category, dtype: int64
```

Since the mean of the venues available per neighborhood is used as the input to the algorithm, the clustering result is indicative of the venues distribution in that neighborhood (and not the number of venues). The clustering results indicate that many of the neighborhoods of Toronto are very similar to one another in terms of the type of venues they offer. The alternative explanation is that the data is not sufficiently complete, especially for some neighborhoods (for example when only a park is listed) or it is not well described/classified for a K Means algorithm in its features - since the venue category contains so many unique values, making it difficult to identify neighborhoods that are distinctly similar to one another and different from others. An alternative test for K Means would be to cluster the neighborhoods based on the number of venues available (instead of the mean) - however this depends on the data being sufficiently representative.

On the other hand, the clusters created based on the housing data were somewhat more evenly dispersed. The two clusters containing the least houses are the two most expensive neighborhood clusters, so it is logical that these clusters would contain less neighborhoods. The boxplot below was created to check the dispersion of prices by neighborhood, where clear differences in prices can be observed.

Cluster	Neighborhood
0	9
1	2
2	7
3	7
4	3



Looking at the mean data of the clusters, we can see a clear price difference between the different clusters, which also tends to corresponds with the other characteristics - number of bedrooms and bathrooms, parking spots, and the type of house. (Type 2 is semi-detached, type 3 is detached.)

Cluster	final_price	list_price	bedrooms	bathrooms	sqft	parking	type	lat	long
0	1.180765e+06	1.087356e+06	3.344595	2.488230	2344.266791	1.254647	2.202435	43.668372	-79.374079
1	3.377111e+06	3.444834e+06	4.576012	4.248750	2354.032051	2.786250	2.865000	43.693723	-79.410901
2	1.914446e+06	1.911297e+06	3.974598	3.338233	2336.863058	1.954450	2.545672	43.693533	-79.408579
3	1.475984e+06	1.370776e+06	3.561166	2.703712	2347.744500	1.267718	2.206300	43.667108	-79.408415
4	2.838587e+06	2.874491e+06	4.221083	4.110937	2474.472456	2.055625	2.622004	43.697963	-79.390201

Conclusions

We can conclude that with the datasets we have, there is a weak correlation (0.31) between a neighborhoods' house prices and the venues available in the neighborhood. The neighborhood clusters & maps created can guide house-searchers to know where they may get a better value for their money, if they are looking to live near certain types of venues. One limitation of this analysis is that the venue data is probably not sufficiently complete for the clustering results to be accurate. Another point of improvement is that the house prices were not normalized with respect to other characteristics beforehand, to try to isolate the impact of neighborhood venues on the house price. Furthermore, other characteristics may need to be taken into account, for example the proximity to a subway station - although as only downtown Toronto was considered, the majority of neighborhoods are close to the subway. Finally, performing this analysis by neighborhood may be too broad, and rather the availability of nearby venues should be done individually for a given radius around each house.