Exploring Housing Prices in Toronto & Venues in the Neighborhood

Anna Cybulsky
June 2020

Business Problem

- Audience: Prospective house buyers in Toronto
- Problem: how are housing prices correlated to key characteristics, and whether they are correlated to venues available in the neighbourhood

Data

	Dataset	Description	Source
1	Toronto Neighborhoods	Neighbourhoods in Toronto with their coordinates	Previous assignment
2	Toronto Property Sales	Property Sales Data including price & characteristics	slavaspirin's GitHub
3	Toronto Venues	Venues available in Toronto neighbourhoods with their category	Foursquare

Methodology - Part 1

- Step 1: Dataset Cleaning

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

After cleaning Convert to float, Remove strings, Filter out condos

	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

Methodology - Part 1

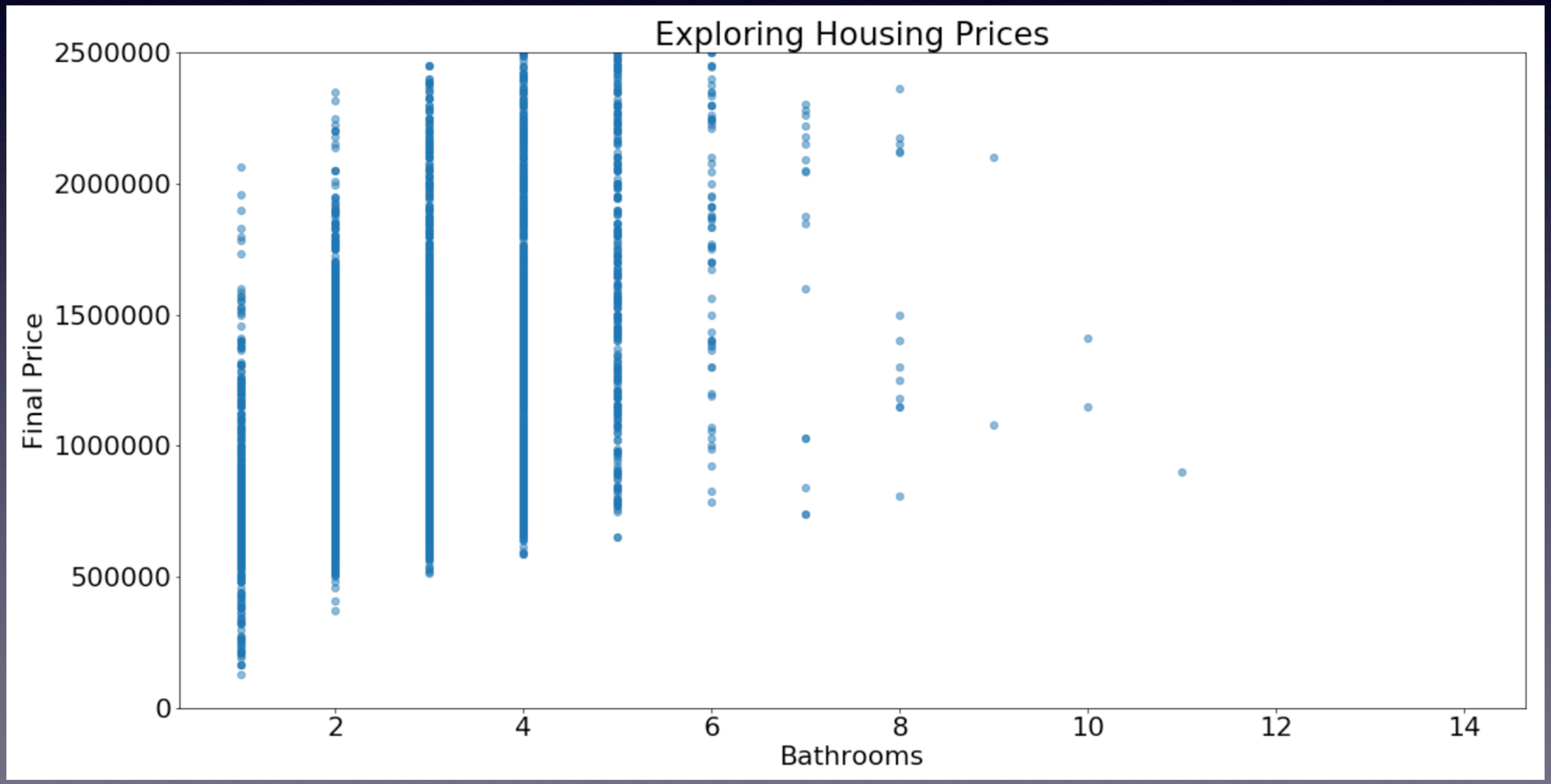
- Step 2: Exploring Data

Final Price vs. Number of Bedrooms

Methodology - Part 1

- Step 2: Exploring

Final Price vs. Number of Bathrooms

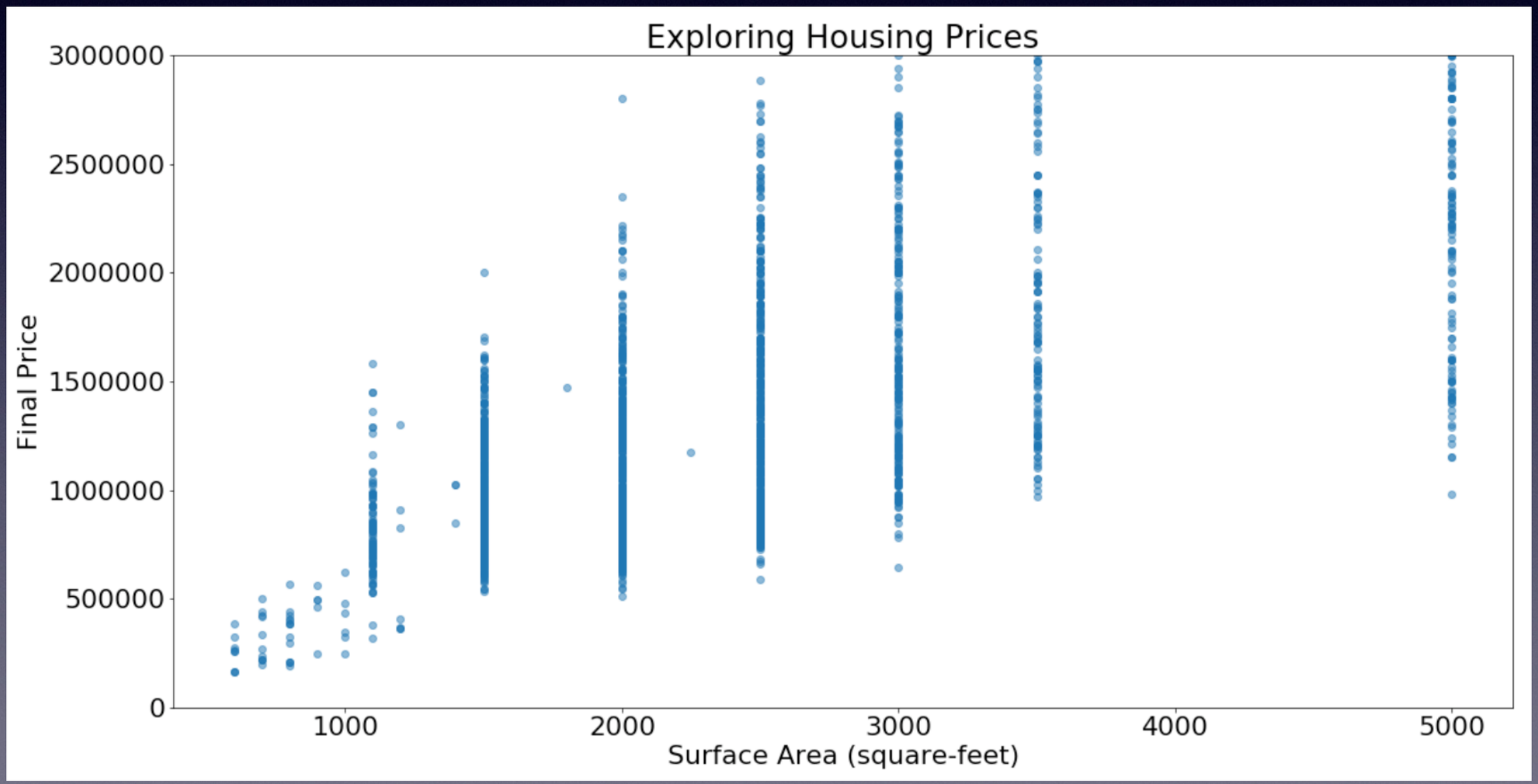


Methodology - Part 1

- Step 2: Exploring

Note: many houses don't have this data available so it cannot be used for further analysis

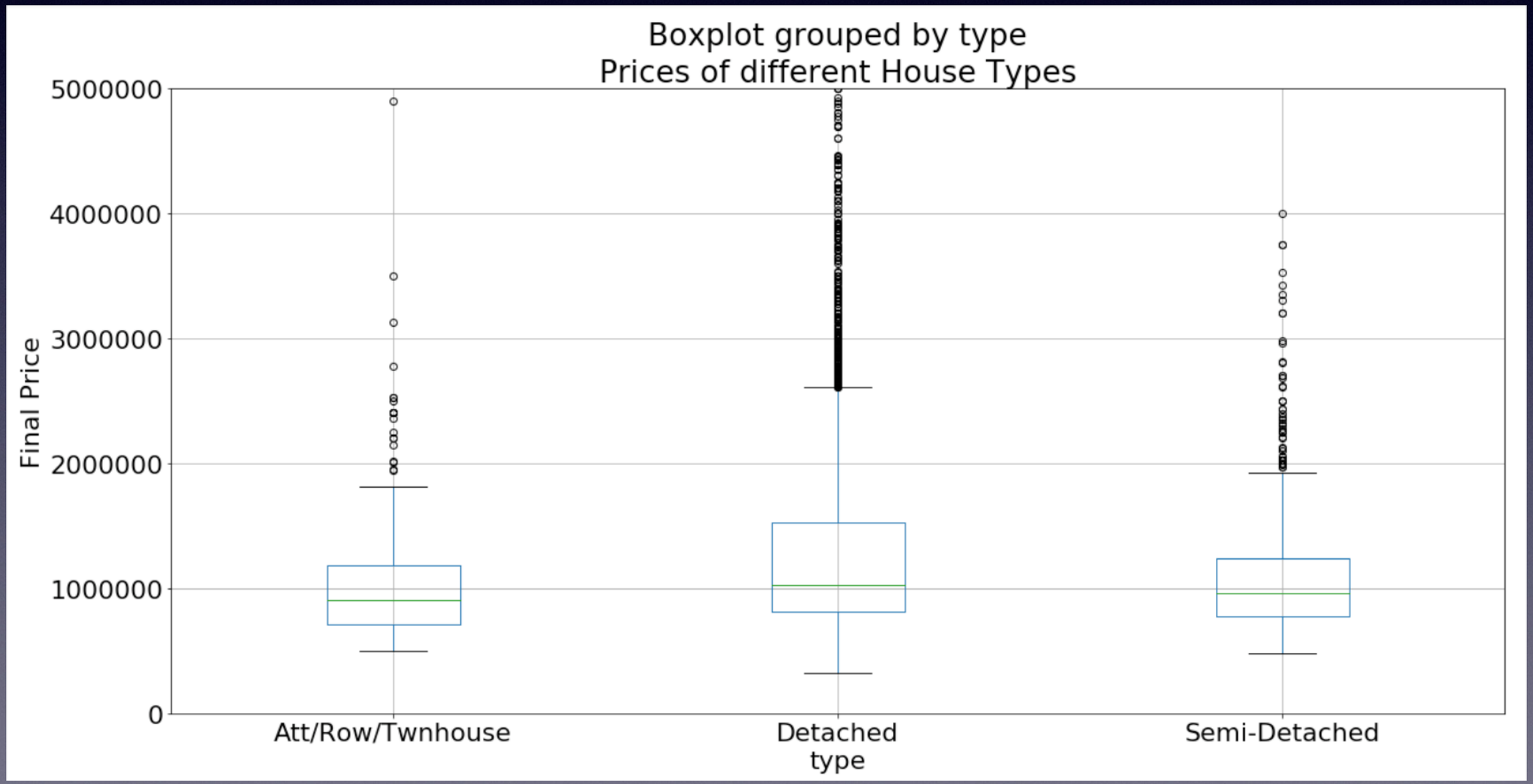
Final Price vs. Square Feet



Methodology - Part 1

- Step 2: Exploring **Dataset filtered on three most common house types**

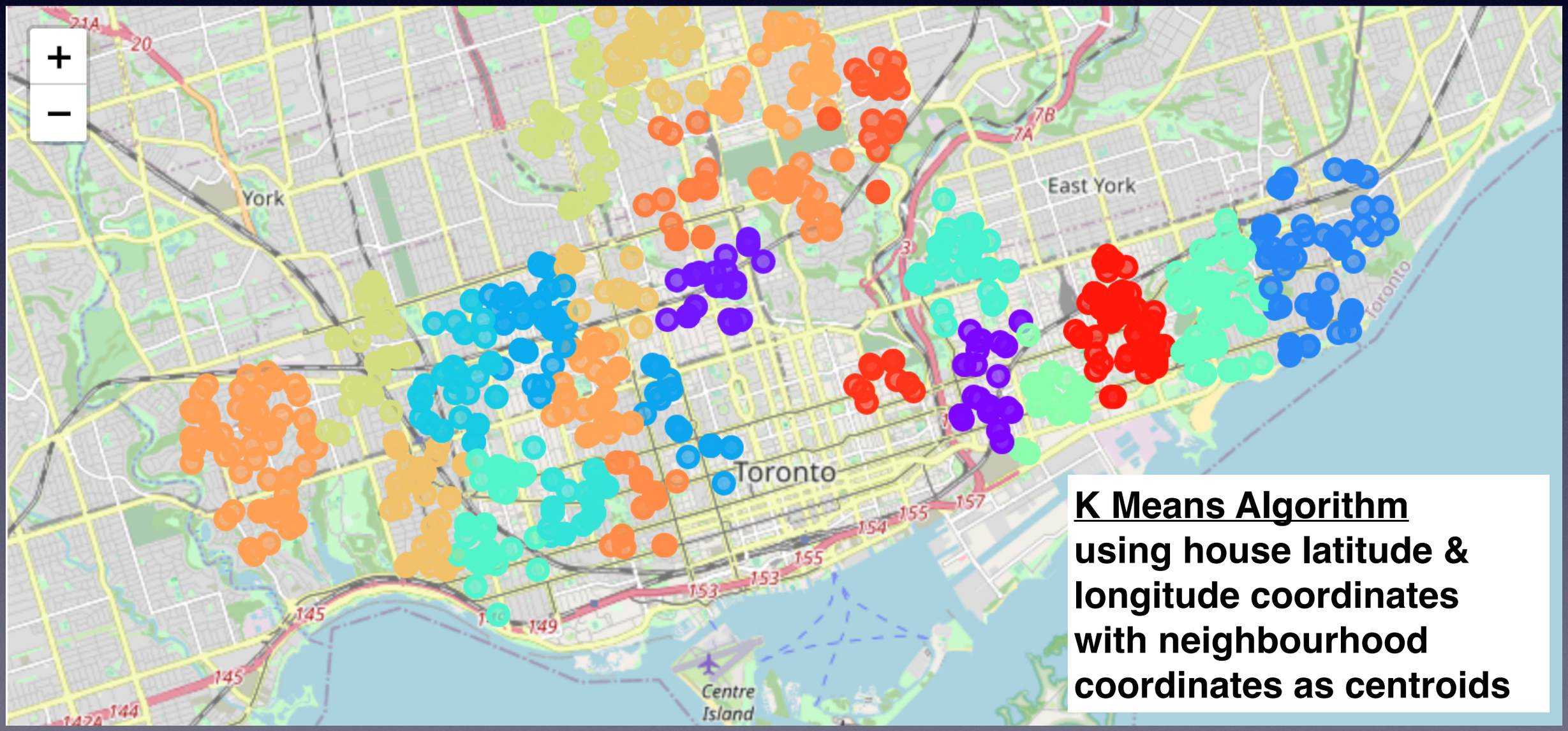
Final Price vs. House Type



Methodology - Part 2

- Step 1: Clustering Houses into Neighborhoods

Map of House-Neighbourhood Clusters



Methodology - Part 2

- Step 2: Exploring correlations between house price & its features

Pearson Correlation

	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

Strongest correlation observed with number of bathrooms & then bedrooms

Methodology - Part 2

- Step 3: Use Foursquare API to retrieve Toronto venues data
- Step 4:
  - Generalise and reduce venue categories
  - Use one-hot encoding to convert venue category into boolean type
  - Then group by neighbhoord and take the mean
  - Cluster the neighbourhoods by venue types using K Means

	Neighborhood	Art Gallery	Bakery	Bank	Bar / Pub	Café / Tea	Concert Hall	Deli / Bodega	Gym	Hotel	Other	Park	Pharmacy	Restaurant	Store / Shop
0	Regent Park, Harbourfront	0	1	0	0	0	0	0	0	0	0	0	0	0	0
1	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	0	0	0	1	0
3	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	1	0	0	0	0
4	Regent Park, Harbourfront	0	0	0	0	0	0	0	0	0	1	0	0	0	0

One-Hot Encoding for Clustering

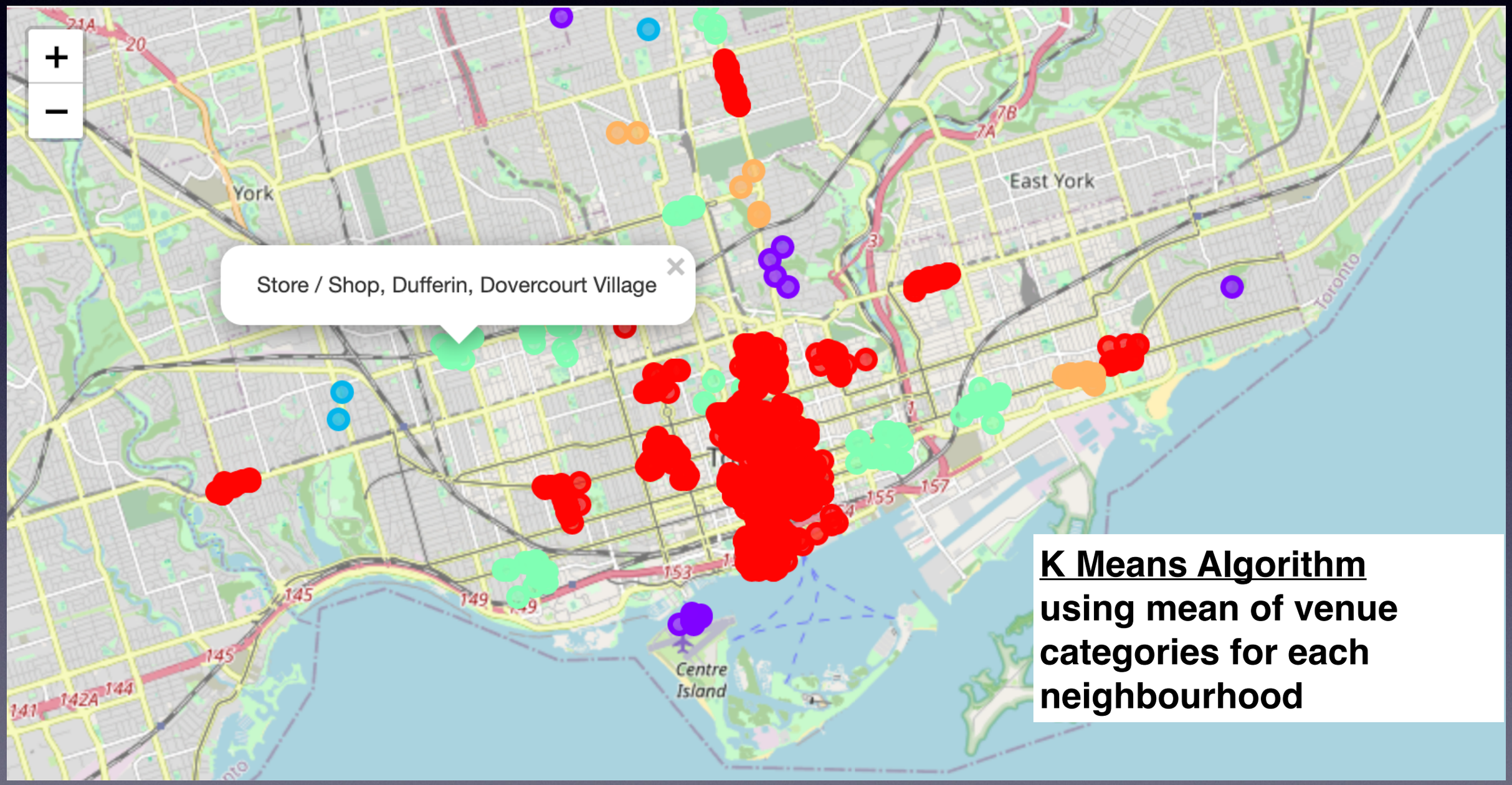
```
venues['General Category'].value_counts()
```

```
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
```

Generalised Venue Categories

Methodology - Part 2

Map of Venue-Neighborhood Clusters



K Means Algorithm
using mean of venue categories for each neighbourhood

Many neighbourhoods fall under the same cluster (Red) - with many venue types

Methodology - Part 2

- Step 5: Cluster the neighbourhoods based on housing data

House-Neighborhood Clusters with Mean Characteristics

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

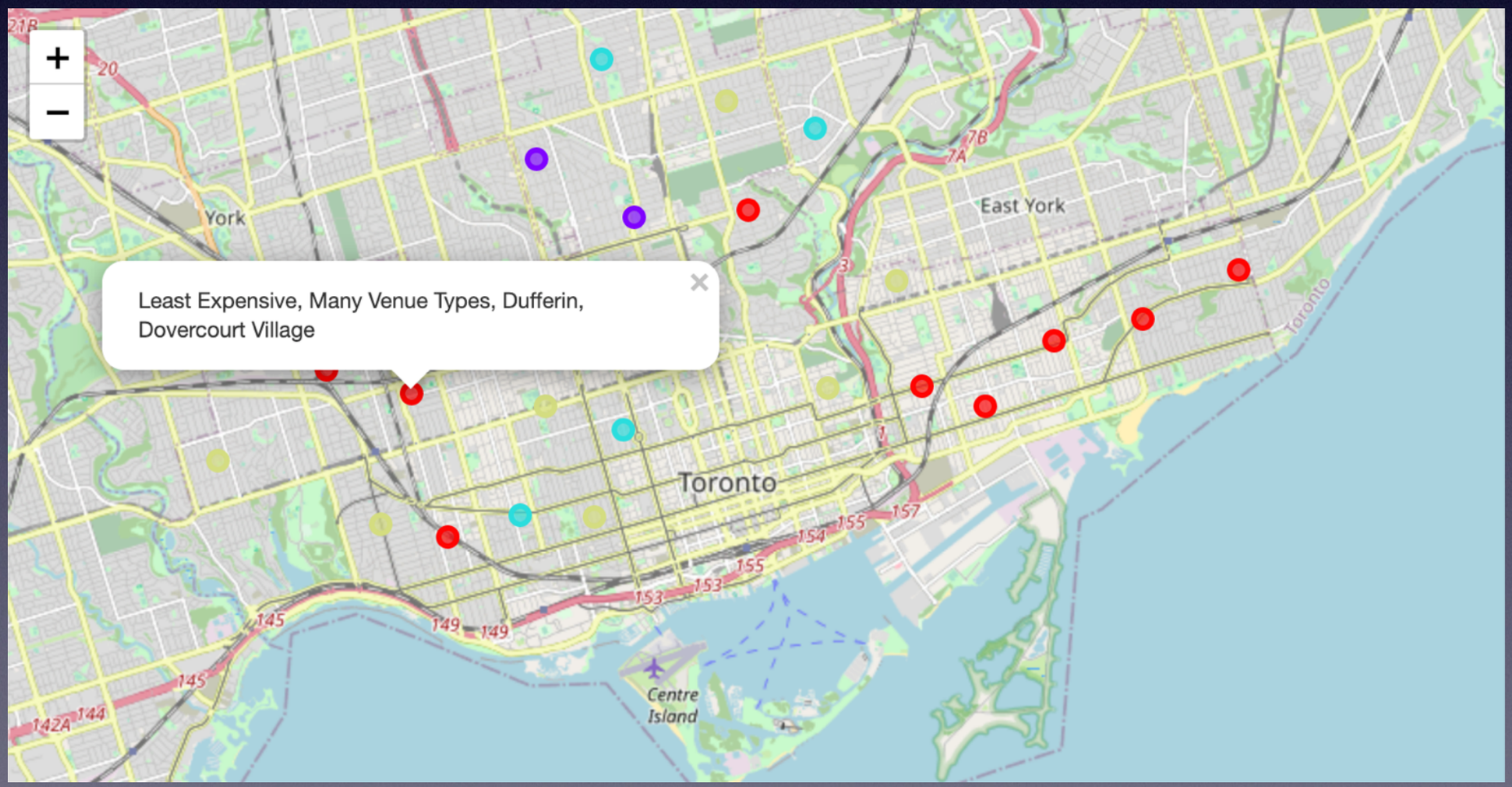
Clusters
0: Least Expensive
1: Most Expensive
2: Medium Expensive
3: Low-Medium Expensive
4: Very Expensive

K Means Algorithm
using mean of house characteristics for each neighbourhood (excluding sqft, lat, & long)

Results

- Compare housing clusters to venue clusters
  - Pearson Correlation of **0.31** calculated —> **weak correlation**

Map of House-Neighbourhood Clusters with both Cluster Labels



Example: Least Expensive Neighborhood with Many Venue Types

Observations

- Venue-based clusters are not evenly distributed
  - First cluster contains more than half the neighbourhoods
- Since mean of venue types per neighbourhood is used, clustering result is indicative of venues distribution (and not the number of venues)
- Result indicates that many neighbourhoods are very similar to each other in terms of available venue types
  - Possible limitations: the data is not sufficiently representative, or the venue categories are not described at the right level of differentiation (too specific vs. too general)

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

Venue-based Cluster Count

Conclusions

- **Weak correlation** between neighborhoods' house prices and venues available
- **Clusters & maps** can guide prospective house buyers with regards to **neighborhoods' value**
- **Limitation:** venue data may not be sufficiently complete, especially for some neighbourhoods
- **Improvement:** normalize housing prices with respect to house characteristics first
  - consider impact of proximity to subway
  - perform analysis as per house location, neighbourhood clusters may be too broad