Using Clustering to Target Best Choices for a New Store-Front

W Duquaine

Introduction

- Most new retail businesses fail within their first 12 months due to either poor location or insufficient funding, or both.

- This investigation uses Data Science techniques to help locate the best locations to start a new upscale restaurant.

- Unsupervised Machine Learning algorithms are applied to locating the best initial set of locations to start a new store-front.

Data Used

- Toronto, Canada was the target city used in this investigation. Has 140 neighborhoods.

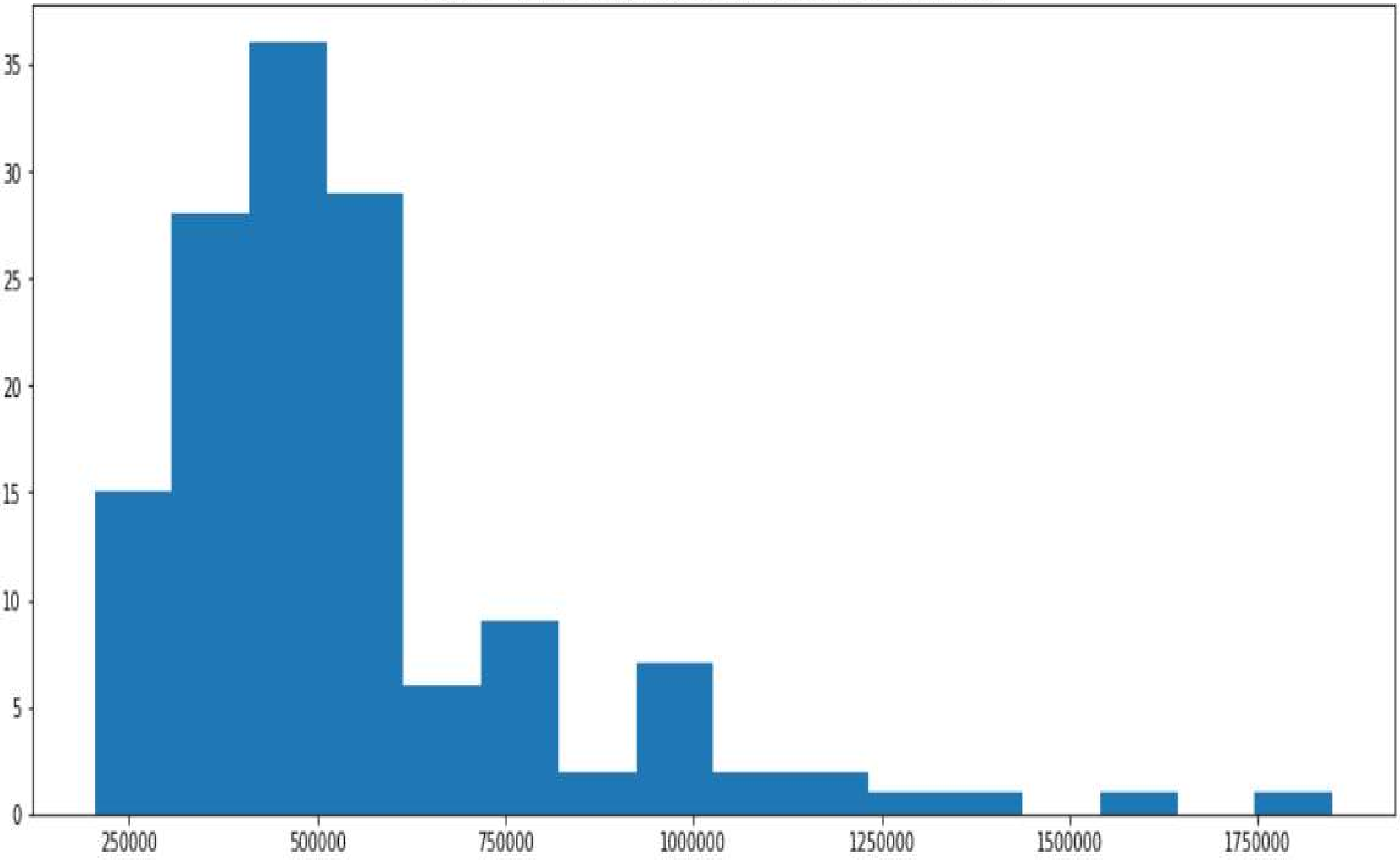
- Primary Data Sources:
  - City of Toronto “Open Data Portal” for neighborhood names and demographic data.
  - Nominatim network (REST) API to extract latitude and longitude information for each neighborhoop
  - Foursqaure Labs network API to extract venue information about each neighborhood – types, counts, etc

Methodology

- Neighborhood and demographic data from Toronto Open Data portal was retrieved, cleaned and slimmed down. Latitude and longitude data was added, and used to lookup the venues for each of the neighborhoods.
- Exploratory analysis of both Real Estate data and Venues data showed a marked clustering orientation of the data.
- Kmeans clustering was used to extract a list of top 24 potential neighborhoods.
- Ranking was used to comb through the 24 candidates and create a “Top 10” list of recommended neighborhoods for the new upscale Steakhouse store-front.

Neighborhood Real Estate Distribution

Figure 1 - Toronto Neighbourhood Real Estate Price Distribution



Mean value is 548K. For an upscale Steakhouse, we are focusing on areas above that number.

Scatter Plot = Clustering

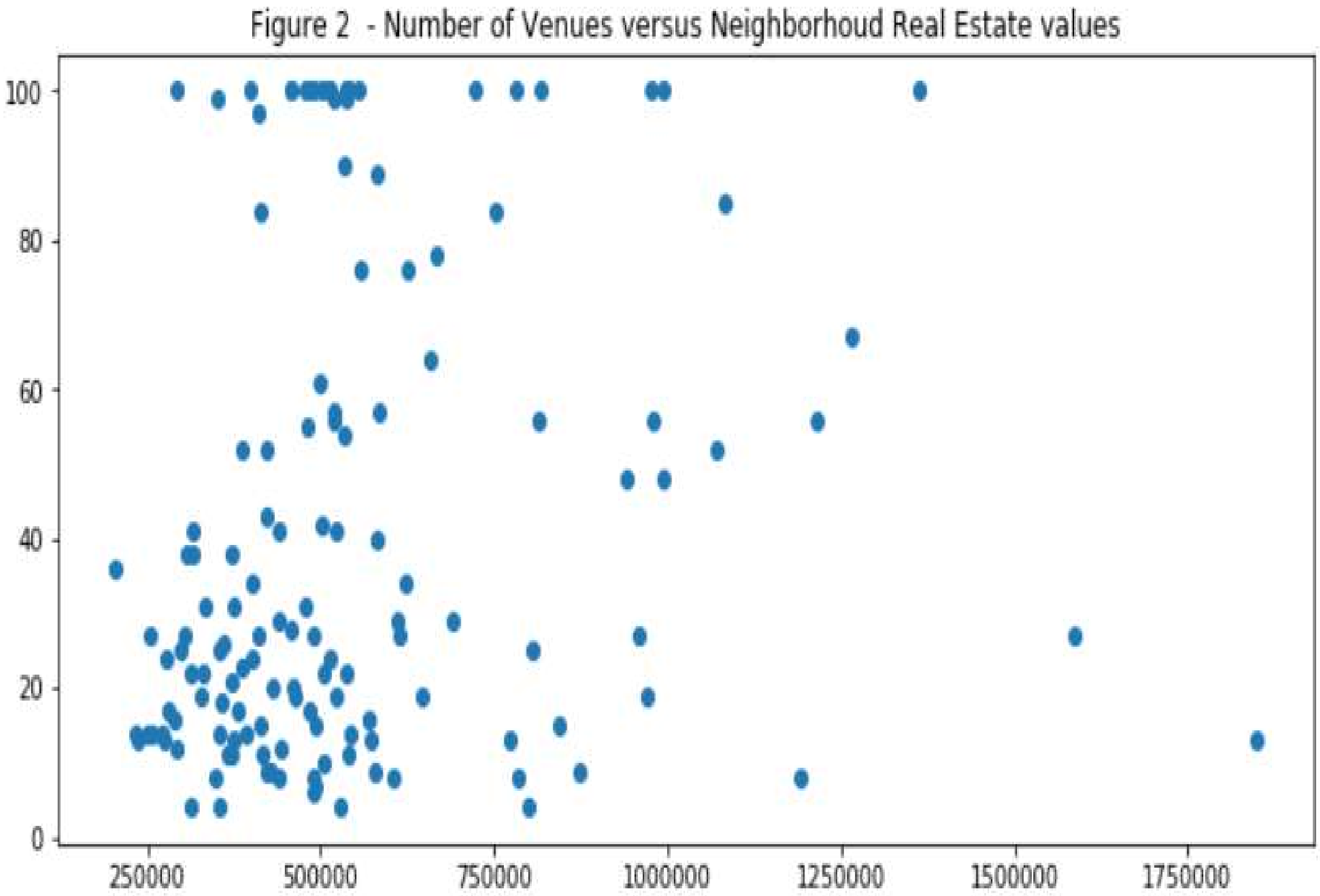
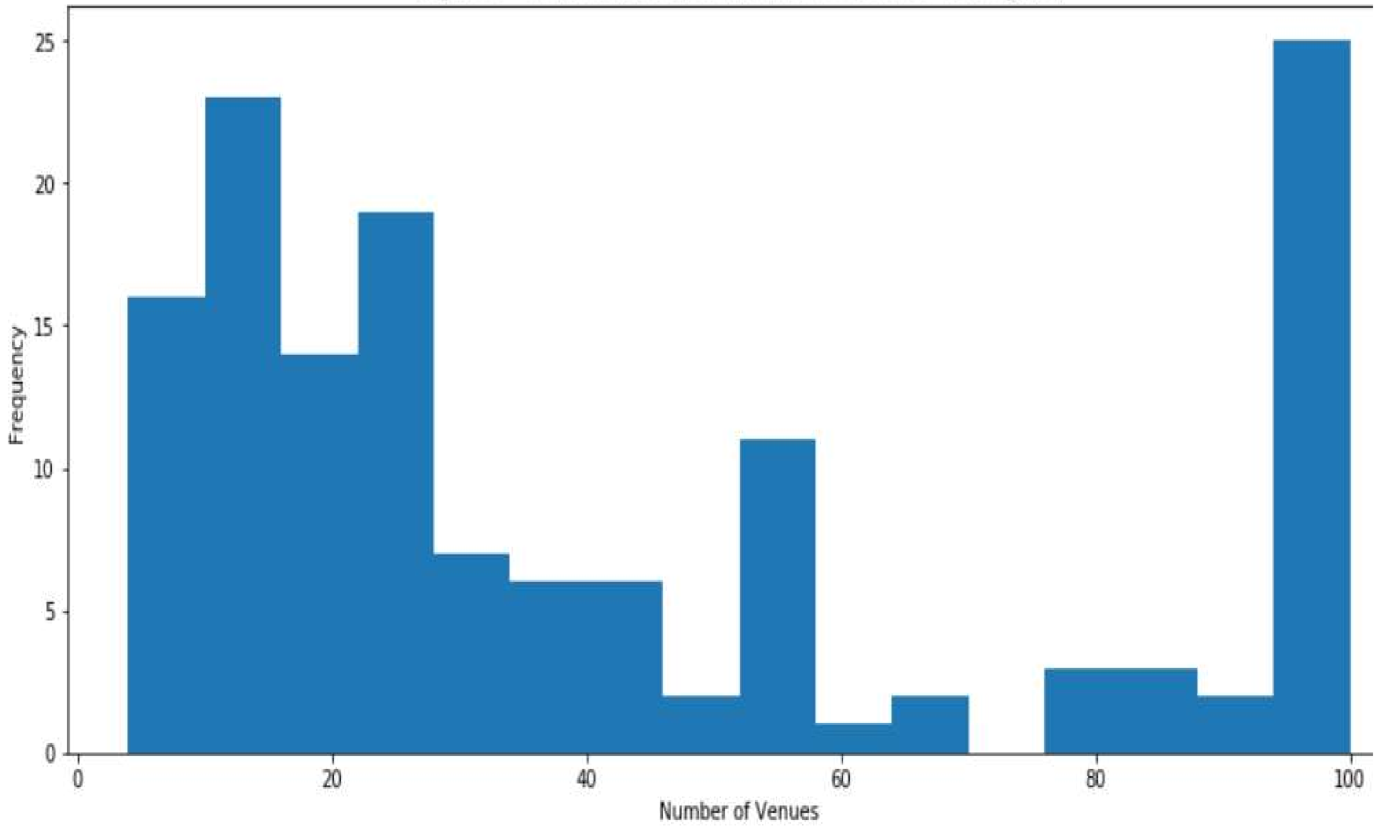


Chart shows this problem has a clustering orientation, not a linear regression orientation. We want to focus on the upper band when targeting for a high-end, upscale venue.

Venues Distribution Across Neighborhoods

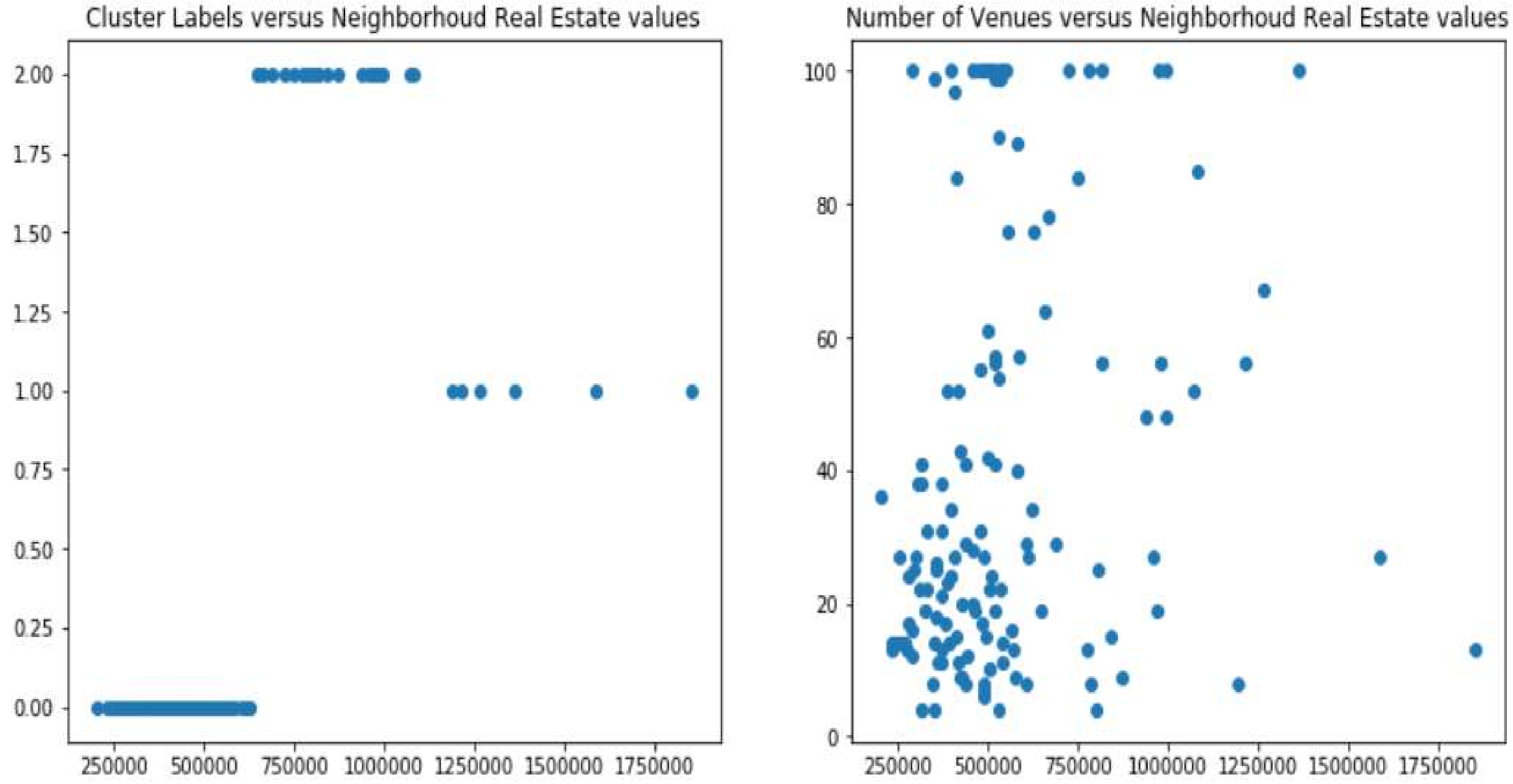
Figure 3 - Distribution of Total Number of Venues - Histogram



Highly skewed at both ends.
We want to target the high-end (right side) with highest number of venues/magnets

Results of Kmeans Clustering

Figure 4



Cluster Labels versus Neighborhoud Real Estate values

Number of Venues versus Neighborhoud Real Estate values

Left chart shows results of Kmeans clustering. Right chart compares to previous clustering observations.

The results from the top band are selected and ranked for best candidates.

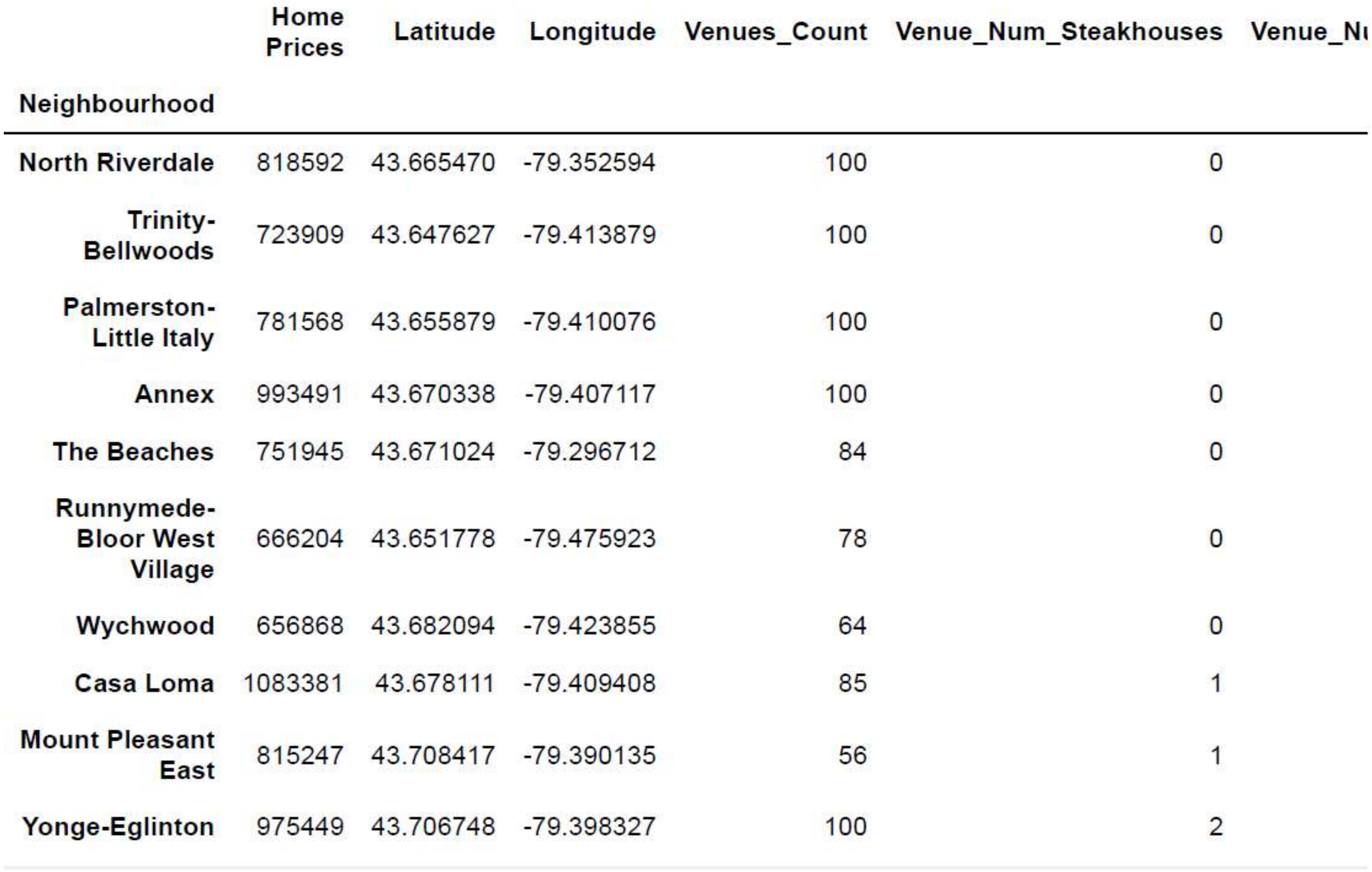
Results

- Data science Clustering techniques were a good fit for the problem.

- Results produced a “Top 10” List of neighborhoods

- Noticed a “sweet spot” in the data that marked the key characteristics of the best neighborhoods for a new upscale venue.

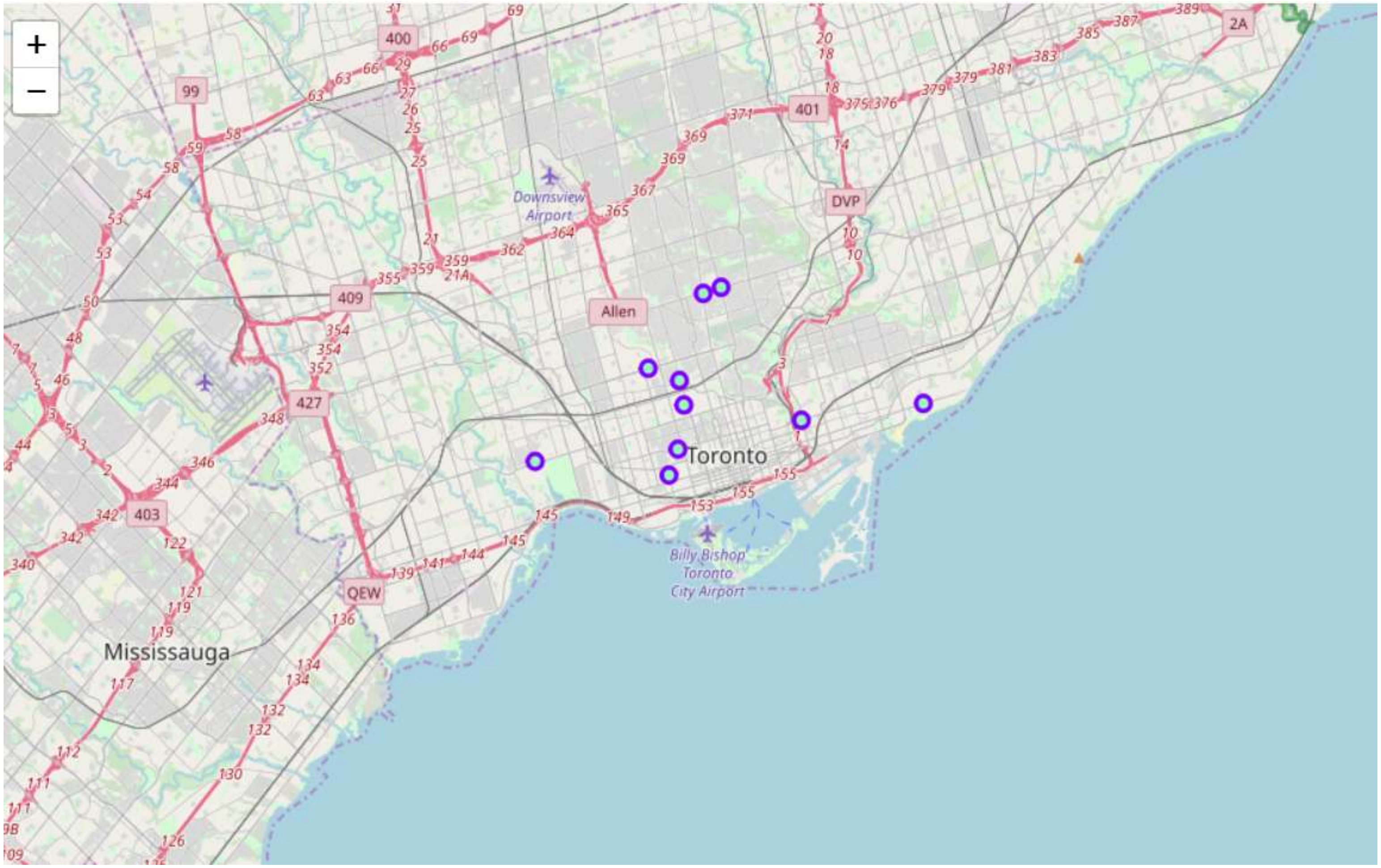
Top 10 Candidate Neighborhoods



	Home Prices	Latitude	Longitude	Venues_Count	Venue_Num_Steakhouses	Venue_Nu
Neighbourhood						
North Riverdale	818592	43.665470	-79.352594	100	0	
Trinity-Bellwoods	723909	43.647627	-79.413879	100	0	
Palmerston-Little Italy	781568	43.655879	-79.410076	100	0	
Annex	993491	43.670338	-79.407117	100	0	
The Beaches	751945	43.671024	-79.296712	84	0	
Runnymede-Bloor West Village	666204	43.651778	-79.475923	78	0	
Wychwood	656868	43.682094	-79.423855	64	0	
Casa Loma	1083381	43.678111	-79.409408	85	1	
Mount Pleasant East	815247	43.708417	-79.390135	56	1	
Yonge-Eglinton	975449	43.706748	-79.398327	100	2	

Notice the most popular areas (venue-wise) are in the 723-993K price region.
Very high end regions (areas above 1,200K) are not in the list (too exclusive and/or zoning restrictions).

Map of Top 10 Candidate Areas



Marked in blue. Notice how most are near the center of the city, toward the lake

Discussion

- Emerging city “open data” portals are getting to be good sources for data mining applications.

- Unsupervised learning (such as Kmeans Clustering) can be used to extract useful results.

- Network based REST type APIs offered by new search and geocoding engines can be leveraged to strongly assist in mining public data.

- A lot of public data is still aggregated at the city level rather than at the neighborhood level. This currently limits the number of cities that can be mined for this type of problem.

Conclusion

- Large cities with good neighborhood demographic data can be mined for locating good set of initial choice locations for starting a store-front business

- Data Science techniques can be used to save time and provide a strong starting point for owners or investors who want to open new store-front business in large urban areas,