# Evaluating Neighborhoods

In New York and Toronto

*John Shubin*
*20 April 2020*

## 1 Introduction

### 1.1 Background

Each city has unique characteristics, and acquires a distinct personality. Municipalities are differentiated by their size, the density of housing, the mix and locations of commercial and industrial sites, and by the demographics of the city's population. Drawing a distinction between cities is a challenge that can be attempted by an analysis of any of these attributes. The question is how much of the difference derives from a unique culture within a city, or are an outgrowth of the physical amenities provided within a municipality. This study will analyze the foursquare dataset for two cities, New York and Toronto, using venue categories and number of different types of venues to define the nature of a neighborhood.

### 1.2 Business Problem

I have just received an offer from a Toronto-based firm that requires relocation. The position is comparable and salary and benefits are attractive, so the primary considerations for me are lifestyle changes. I would like to evaluate the new city in comparison to my hometown New York, and determine if the two cities are substantially similar or significantly different. Second, I would like to find the neighborhood that most closely matches the one where I currently reside in Brooklyn.

While the physical size of Toronto and New York is comparable, the population of New York is almost three times that of Toronto, implying higher population density, which might lead to higher numbers of venues in a given neighborhood's area. Toronto covers an area of 243.3 square miles, or 630 square kilometers and has a population of 2.93 million residents. New York City, as a whole, covers 302.6 $mi^2$ (784 $km^2$) and has a population of 8.623 million residents.

The availability of arts, entertainment, restaurants, and nightlife are the primary features that will be used to evaluate neighborhoods in a community.

# 2 Data

Data to be used for characterizing a neighborhood will be venue information from Foursquare. Specifically, I will be using the 'search' endpoint with the parameter 'intent' set to 'browse', a seach that appears to yield the broadest results. Foursquare attempts to provide a comprehensive list of venues within a metropolitan area. There are a few challenges encountered when using this data. The first is that the limits on the quantity of returned data points often yields a cluster of venues very close to the search latitude and longitude (the 'center' of the search), and sparse data at the fringes of the neighborhood (Figure 1).

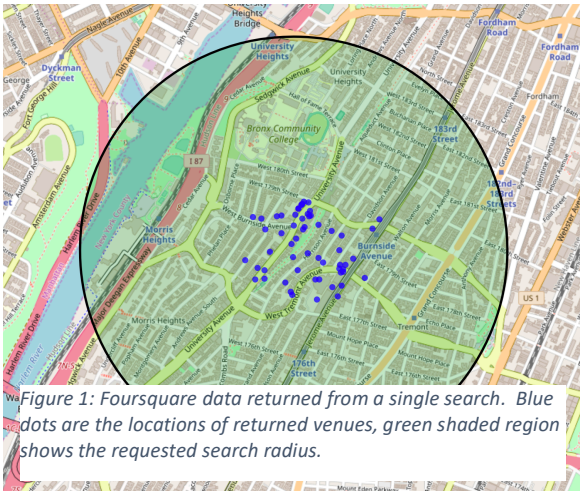



*Figure 1: Foursquare data returned from a single search. Blue dots are the locations of returned venues, green shaded region shows the requested search radius.*

This can be overcome, somewhat, by varying the search coordinates slightly, and adding unique results found for searches at these nearby points to the venue data for a give neighborhood.

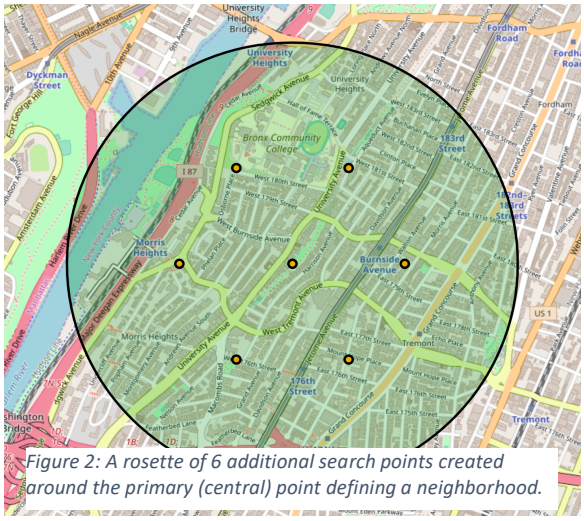



*Figure 2: A rosette of 6 additional search points created around the primary (central) point defining a neighborhood.*

The decision to harvest a larger set of venues for each neighborhood is supported by searching a rosette pattern of locations surrounding the center of the neighborhood (Figure 2). A decision to include 7 points to search for each neighborhood was based on the tradeoff of additional searches required (Foursquare limits the number of queries allowed per day from a single API key) and the harvesting of unnecessary duplicate venues from Foursquare.

Using multiple points provided a much better set of venues for the test neighborhood. When the duplicates are eliminated, and venues that fall outside the neighborhood were dropped, the resulting set of venues harvested provided adequate coverage for the entire neighborhood (Figure 3).

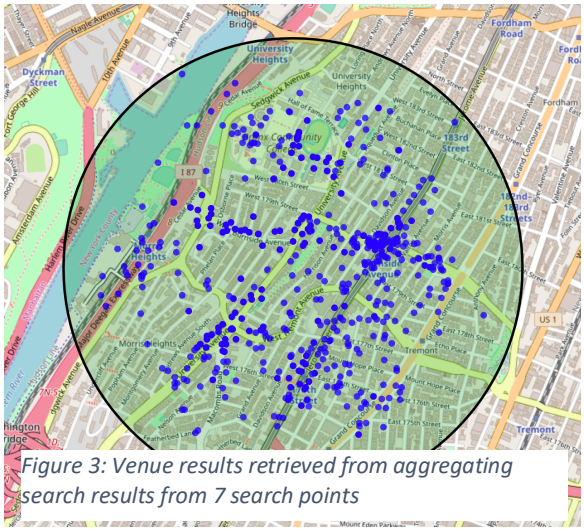



*Figure 3: Venue results retrieved from aggregating search results from 7 search points*

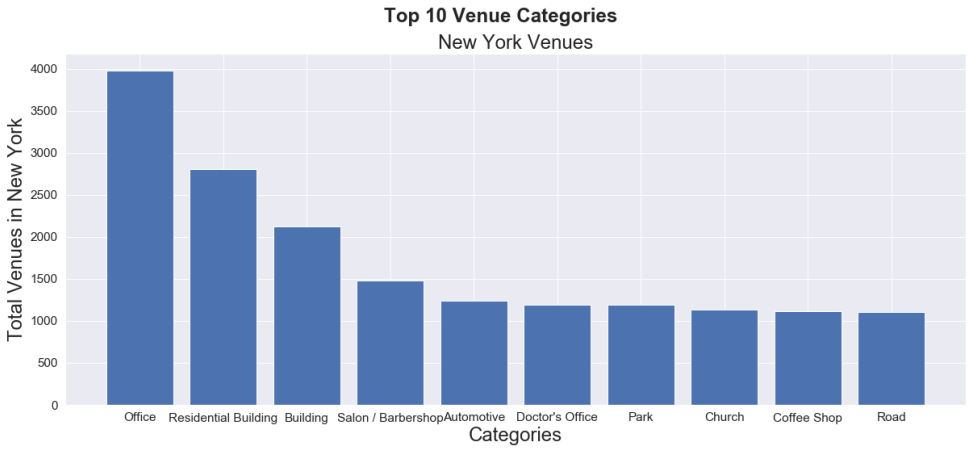
## 2.1 Data Exploration

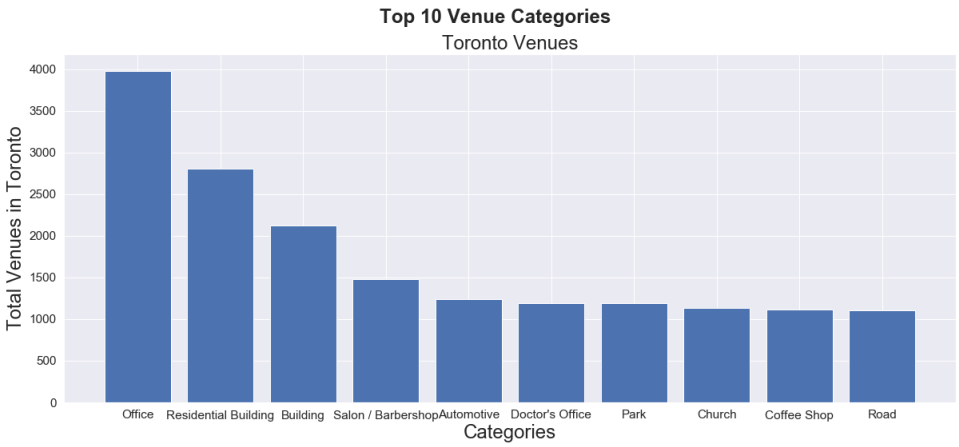
Venues are described by a single attribute, called the venue category. This category provides a single description that attempts to convey the principal nature of the venue.

The taxonomy of descriptions found in the 'Category' field in the foursquare database is not fully documented, and searches often yield unique category counts that exceed 50% of the total number of venues returned. Increasing the number of venues included in each neighborhood would provide a better representation of the 'flavor' of a neighborhood.
An alternative approach would be to restrict search categories and use the results of multiple searches to provide a list of venues that are actively considered more important to the researcher.

Collecting data from each neighborhood yielded a broad array of distinct categories. New York neighborhoods contained venues with 656 unique categories, and Toronto Neighborhoods had 628 unique categories.

Exploring the top 10 categories found collectively in New York and Toronto yielded similar results:



These plots suggest that venues in the two cities are at least similar, but these top category summaries don't contain enough detail to provide an accurate comparison of the cities, and doesn't provide any information about specific neighborhoods. The problem with analyzing neighborhoods with category data, perhaps by counting the number of each specific venue category within each neighborhood, is the large number of distinct categories. The hundreds of unique categories found in each city would create a feature space with too many dimensions to be analyzed in context of the relatively small number neighborhoods under consideration.

## 2.2 Feature Engineering

We can limit the number of features by aggregating similar features into more general categories. Foursquare also has a set of general categories, that collect the 936 specific categories into 10 general categories. These parent categories may provide aggregation that creates a more consistent and comparable model for the venues in a neighborhood, but are not returned by the foursquare endpoints, so creating a database that links the specific categories returned by a search to their more general (parent) groups was required. The category listing includes 10 top level categories, some of which hold up to four levels of subcategories within them. The top-level category with the most subcategories is, unsurprisingly, 'food' which includes restaurants, grocery stores, and specialty food shops.

## 2.3 Data Cleaning

A number of venues did not have categories attached included in their records. These venues were dropped from search results. Because the category type descriptor in the Foursquare database included subcategories of varying depth, aggregating venues to a top-level descriptor required linking each category individually to its ultimate parent.

When the data was aggregated to top level categories, the distribution of each category within the collected data is found in Figure 4.

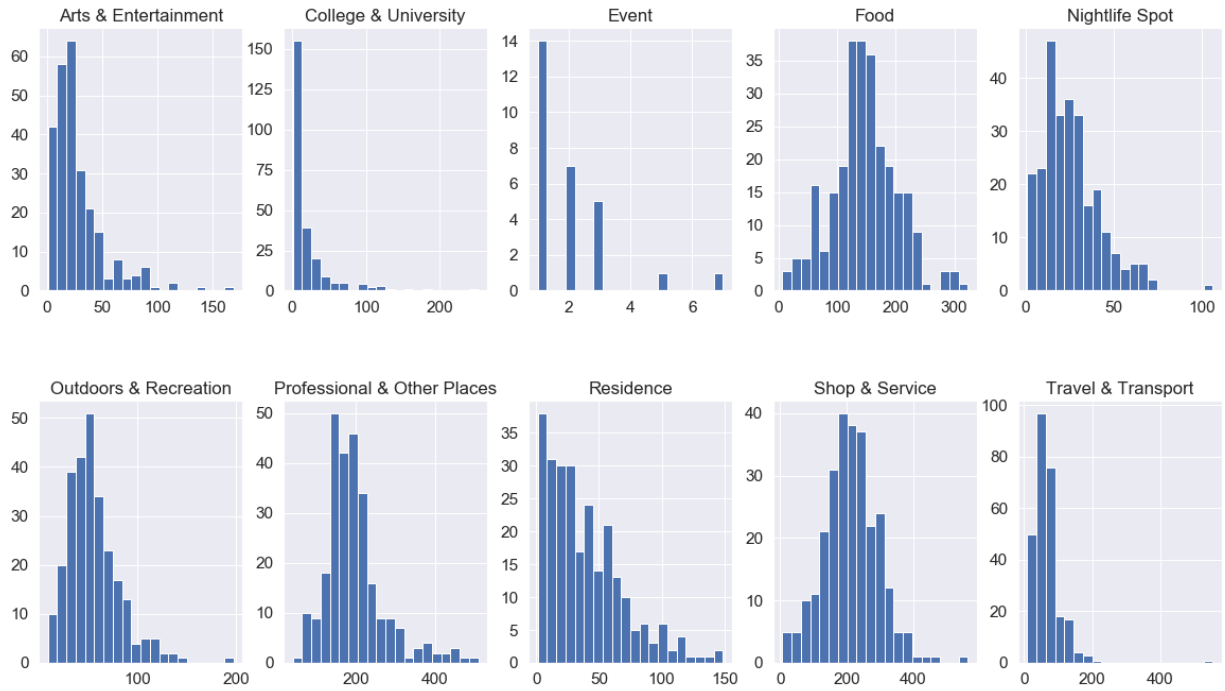



*Figure 4: Distribution of venue category counts within a neighborhood*

One of the ten top level categories, 'Events', was missing from the majority of neighborhoods in the search results, with 242 of 270 neighborhoods including none of this type of venue. Additional inspection of the details of these venues led to the assumption that the 'Events' category is simply not used consistently. The decision was made to drop this feature prior to any additional analysis.

Additionally, neighborhood features that did not have any venues of a specific top-level category (e.g. higher education), ended up with null values, and were coded to NaN by pandas. These NaN values were replaced with zero, which is the correct count of these venues in these neighborhoods.

## 2.4 Exploratory Analysis

An unsupervised clustering model was created to separate the neighborhoods in both cities into two distinct groups, with the results shown in Table 1. Although this modeling is not conclusive, results where neighborhoods from each city were placed into opposite groups would suggest an overall difference in the makeup of neighborhoods in each city. Alternatively, a result that allocated similar percentages of each city's neighborhoods to each cluster would suggest that taken as a whole, the cities were similar in composition in a neighborhood-by-neighborhood comparison.

Ratio of neighborhoods in each cluster, by city

| | New York | Toronto |
|---|---|---|
| Cluster A | 110 | 83 |
| Cluster B | 61 | 16 |

| | New York | Toronto |
|---|---|---|
| Cluster A | 64.3% | 83.8% |
| Cluster B | 35.7% | 16.2% |

*Table 1*

At first glance, the results are inconclusive. Although both cities tended to have more neighborhoods placed into cluster A, this tendency was much stronger for Toronto. Testing the data with a $X^2$-test yields a value of 4.8, giving a p-value of 0.0285 at a significance level of 0.05, indicating we should reject the hypothesis that the neighborhoods from the two cities are similar. Although finding the exact nature of the difference between the cities' neighborhoods is beyond the scope of this problem, it seems reasonable to assume that either differences in venue density or differences in venue mix are likely suspects. The difference does suggest that finding a neighborhood similar to my current neighborhood is less likely to happen by chance, and supports using a data-driven approach to help with the problem.

Lastly, the Hopkins ratio was calculated for the aggregated data for the neighborhoods of New York and Toronto combined. The resulting index, 0.77, suggests the categorized neighborhood data is a valid candidate for cluster analysis.

### 2.4.1 Determining Candidates for the number of clusters, K

The data was analyzed using k-means for various k, and plotting the cluster inertia for each value of k. The knee plot (Figure 5) suggests that the optimal value for k is four.

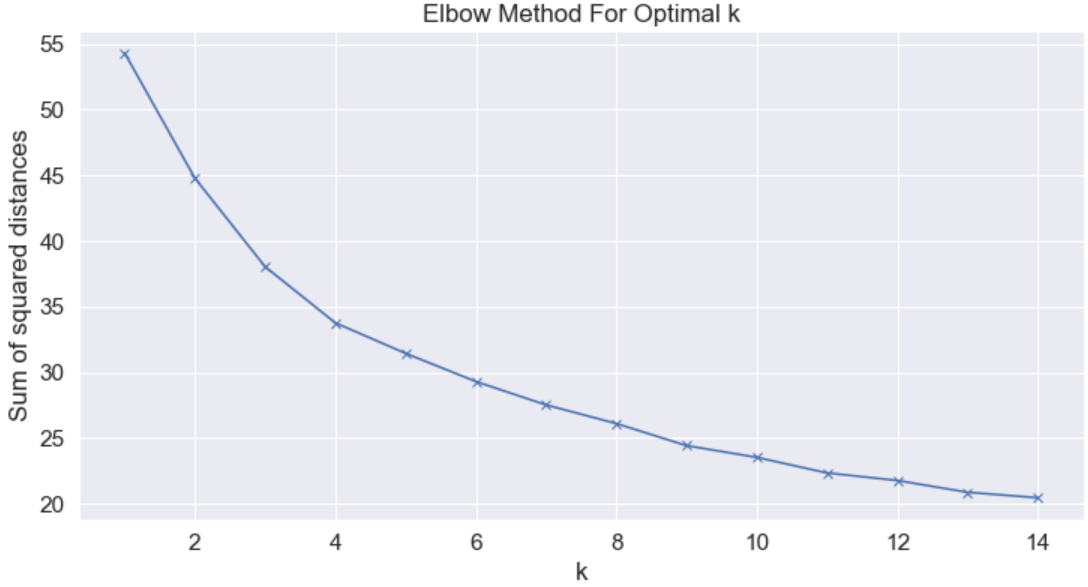



*Figure 5: Plot used with elbow method*

For hierarchical clustering the data was analyzed by distance criteria with the scipy.cluster.hierarchy library to produce a dendrogram (Figure 6). The dendrogram suggests by setting the distance threshold to 15, we are left with 5 clusters.

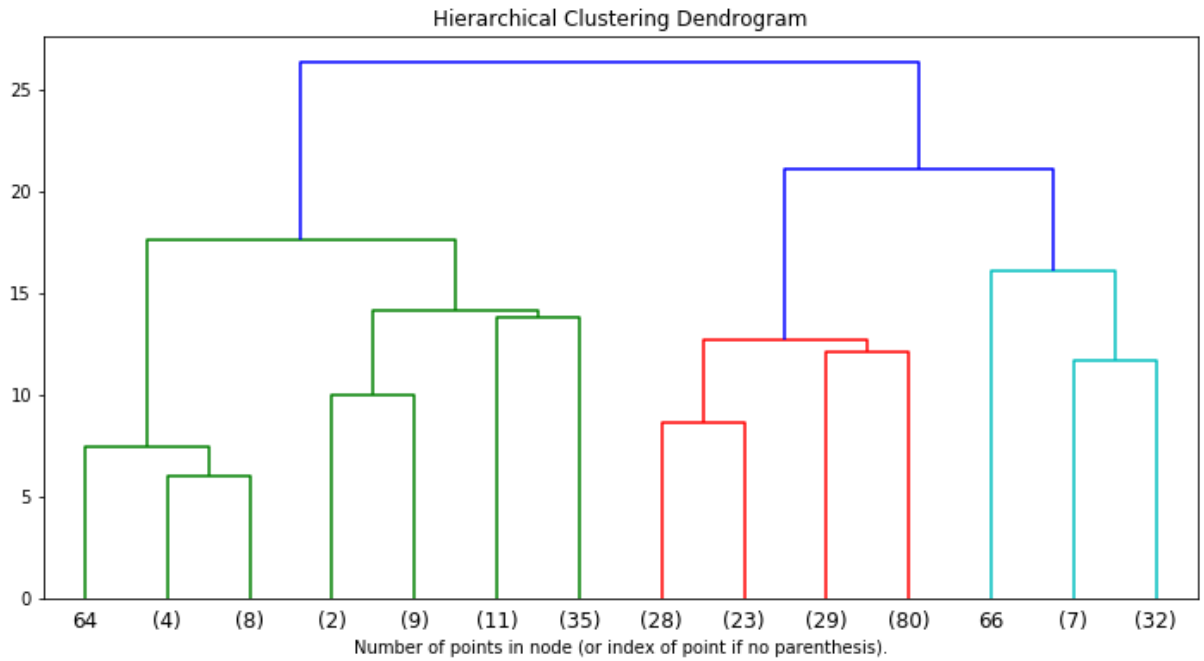



*Figure 6: Distance Dendrogram of Agglomerated Clusters*

# 3 Analysis of Neighborhood Data

Analysis of neighborhood data is conducted to produce candidates for relocation that match my current neighborhood in Brooklyn, in ZIPcode11104.

## 3.1 Cluster Analysis

We can build a set of neighborhoods that would be similar to mine by clustering the venue data set, grouping together similar neighborhoods. Candidate neighborhoods would be those neighborhoods that clustered with my current home in zip code 11104. For visualization of the resulting clusters, data was projected into two dimensions using principle component analysis. Individual feature contributions to the two components were plotted (**Error! Reference source not found.**).

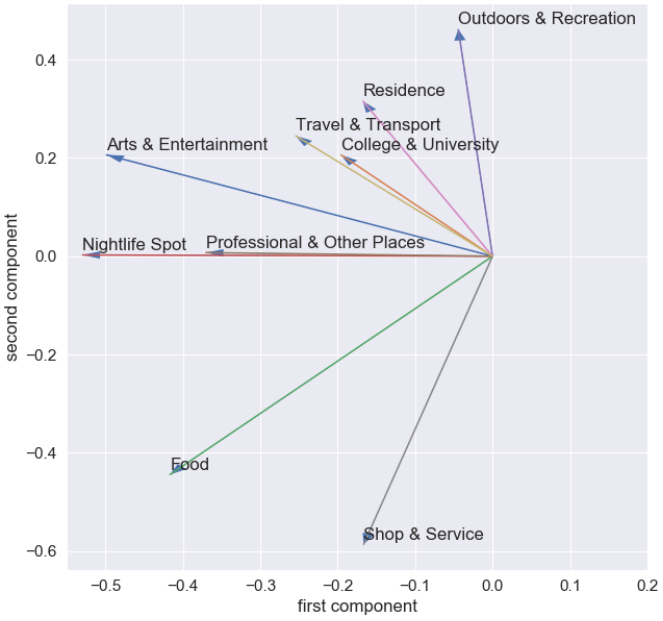


*Figure 7: Feature Contributions to PCA components*

Cluster Plots for k=3, k=4, k=5 are compared in Figure 8. The cluster of interest, containing the neighborhood 11104, in all cases is plotted in red.

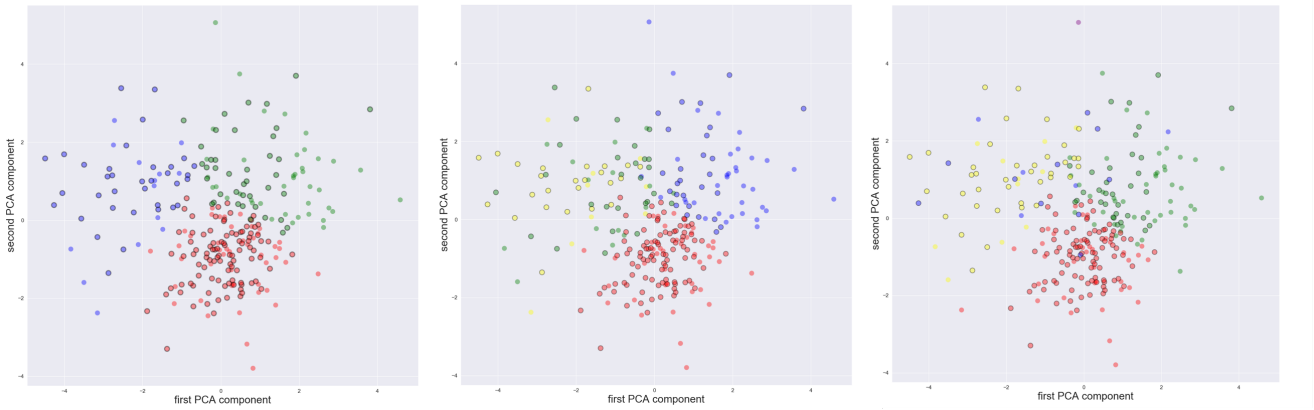



*Figure 8: from left to right: k=3, k=4, k=5*

In all cases, neighborhood 11104 is in the largest cluster. We can map the neighborhoods of Toronto, highlighting those in the same cluster as, or clustered with, ZIPCode 11104 (Figure 9).

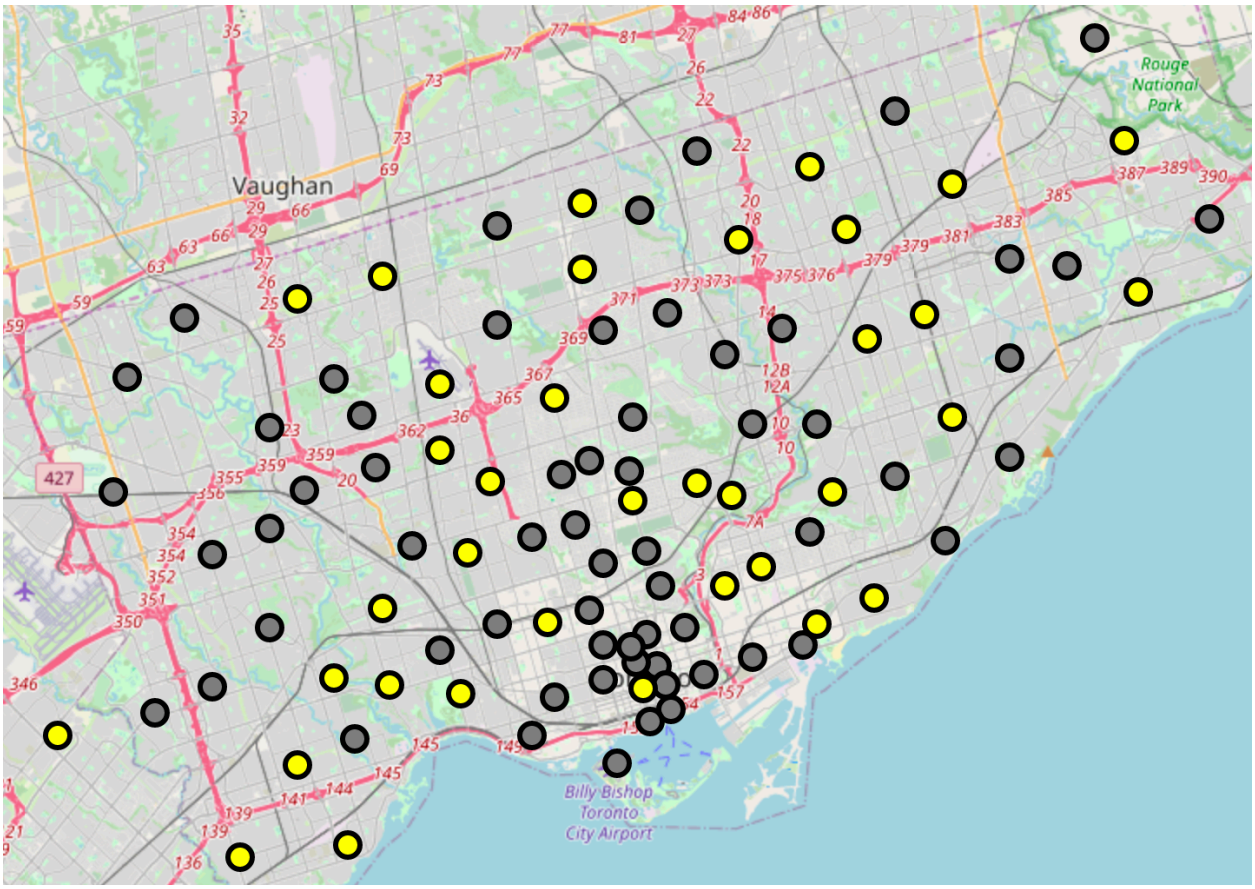



*Figure 9: Neighborhoods Clustered with ZIPCode 11104*

Results of this visualization show neighborhoods (in yellow) similar to 11104 scattered throughout the Toronto municipality. Although this analysis limits the number of neighborhoods to consider, it still leaves 36 (or 36.3%) of the neighborhoods as candidates. This presents limited value as it does not streamline the decision-making process enough.

## 3.2 Performance of K-Means Clustering on neighborhood data.

We can use the Silhouette coefficient to evaluate the effectiveness of the k-means clustering used on our neighborhood data. The results shown in Figure 10.

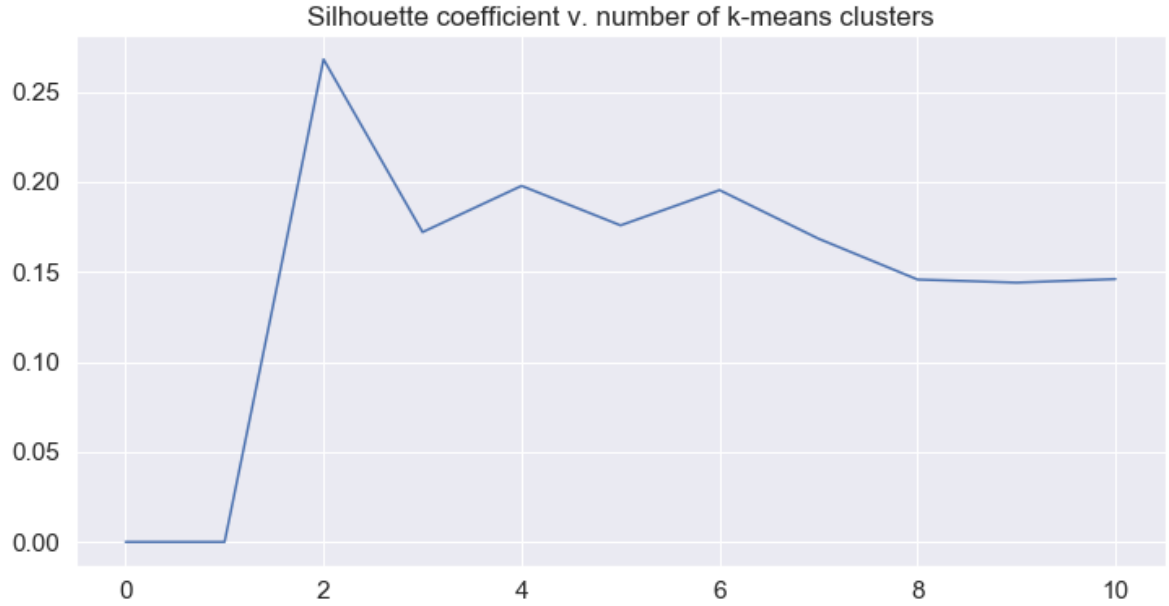



*Figure 10: Silhouette coefficient v. different values of k*

The low values for each clustering scenario, with the value never exceeding 0.25 for K ≧ 3 suggest there is little underlying structure driving the cluster results, so the resulting clusters may be somewhat arbitrary or artificial.

## 3.3 Distance Analysis

A more direct approach to determine which neighborhoods are similar to my home neighborhood in Brooklyn would be to measure the feature distance between the neighborhoods based on the venues found through Foursquare. Using normalized data, we can use the n-dimensional distance formula to calculate the distance between the Toronto Neighborhoods and ZipCode 11104. For a home neighborhood *h* and another neighborhood *n,* each with *a* attributes:

*Equation 1:* $Distance^2 = \sum_{i=1}^{a}(x_{ni} - x_{hi})^2$

Calculating the feature distance for each neighborhood relative to my neighborhood gives a us a way to determine which neighborhoods have feature sets that most closely match the features of my current home.

Initial results are not exceptionally promising, as shown in the histogram in Figure 11. Few neighborhoods are very similar, and most fall in a narrow range, suggesting most neighborhoods have a similar range of dissimilarities.

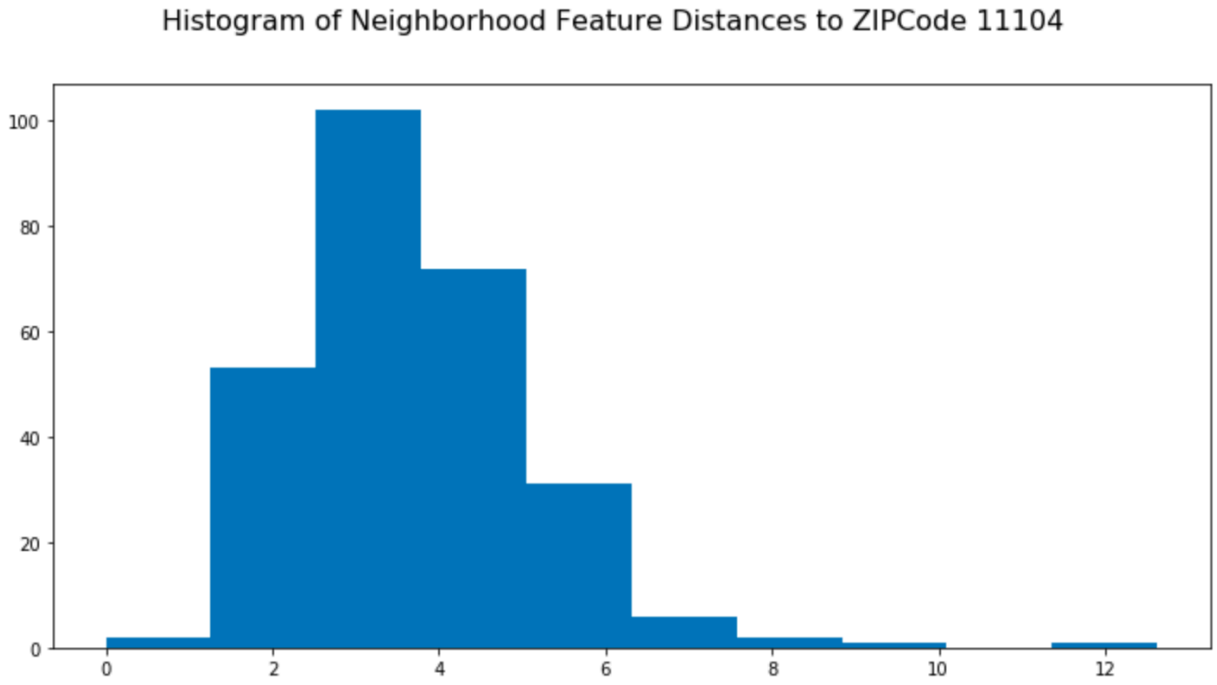



*Figure 11: Histogram of Feature Distance Relative to ZIPCode 11104*

It might be possible to limit the number of features we examined, for example reducing the examined categories to just 'Food', 'Arts and Entertainment', and 'Night Life'. This lower dimension feature space would be smaller, and feature distances should be closer to our home neighborhood.

## 3.4 Map of Feature Distances

When we plot each neighborhood's feature distance from ZIPCode 11104, we get the visualization shown in Figure 12. In this mapping we get more we get a few candidates for exploration, for example Postal Codes M6N and M6B are plotted as bright green dots, indicating lower feature distance. This might suggest a starting point for exploring housing options in Toronto.

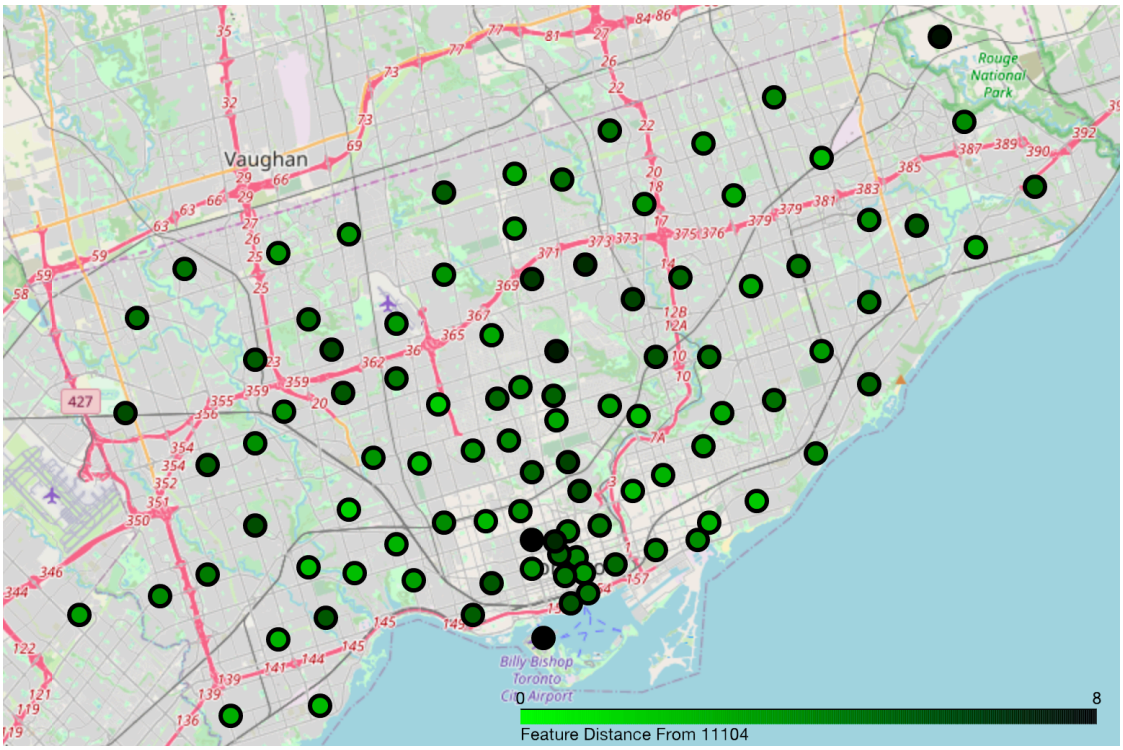



*Figure 12: Map of feature distance from ZIPCode 11104 for neighborhoods in Toronto*

# 4 Conclusions

Using a data-driven approach to evaluating a decision, like the neighborhood in a new city that most closely matches your current one, offers clear advantages to the alternative of conducting a broad survey of the new location. Analysis provides a tool for limiting the extent of a search, potentially saving time while improving outcomes.

The approach is only limited by the quality of the data that is used in the analysis. In the case of using venue data from foursquare we were able to get a reasonable recommendation for investigation. There are several limitations apparent in the dataset, however. Differences in the venues found in the FourSquare dataset are coded with a 'category' field that defines the nature of an establishment at a location. The classification of each venue has an uncertain level of quality and relies on a significant amount of subjectivity.