Applied Data Science Capstone Project –

Battle of the Neighborhoods –

Company Expansion Site

Johnathan Boyce

9/19/2019

# Introduction

## Background

Swaggy Intelligent Canines and Kitties (SICK Co.) is expanding its company footprint. As the world’s fastest growing manufacturer of doohickeys, thingamabobs, and (most importantly!) shrubberies, it is extremely important for SICK Co. to establish a corporate location in the Toronto metropolitan area.

## Problem

As a consultant to SICK Co., our objective is to identify the best location for SICK Co. to expand within the Toronto metropolitan area. Since this expansion is critical to the Company’s success, SICK Co. created an Expansion Committee (EC) solely for the purpose of spearheading the proper selection of the site locale.

Per instruction of the EC, areas for expansion would be based on the types of venues (e.g, shops, entertainment centers, and various other points of interest). In addition, the EC devised a scoring/weighting system based on the type of venue. Venue score, as well as the prevalence of crime in the area, will play a material part in the final decision on where to add the new location.

# Data

## Data Sources

The data sources used for this analysis came from a variety of websites and data tables, including the following:

* FourSquare Places API (<https://developer.foursquare.com/docs/api>)
* Wikipedia List of Postal Codes (<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>)
* Toronto Police Service Public Safety Data Portal (<http://data.torontopolice.on.ca/datasets/mci-2014-to-2018/geoservice>)
* Scores / weightings for venue categories provided by the EC for use in this expansion project (toronto\_venues\_scores\_only.xlsx)

## Data Collection

The FourSquare data includes features relevant to the goal of scoring and identifying Toronto areas for the corporate expansion project. Some of the available features are Postal Code, latitude, longitude, and venue category. Data extraction from FourSquare entails querying its API using protocols found within their documentation. Once extracted, we will merge and/or compare the data with our other sources.

Data found on the Wikipedia List of Postal Codes includes Postal Codes, as well as the corresponding Boroughs and Neighborhoods.

Data found on the Toronto Police Service Public Safety Data Portal includes Major Crime Indicators (MCI) and can be easily downloaded from the website (or extracted using a webscraping tool). MCI categories include Assault, Break and Enter, Auto Theft, Robbery and Theft Over (excluding Sexual Assaults). In addition to MCI category, features include the location (at the nearest intersection), Building Type, Neighborhood, Police Division, and Occurrence Year/Month/Day/Hour.

## Data Understanding

The data, once compiled, will be used to find the ideal location to expand the corporate footprint of SICK Co. By using the total scoring/weighting provided by the EC, we are able to derive the total score for each Neighborhood by multiplying the venue category score x the number of venues in that specific category in that Neighborhood (for example, Art Museums are given a 10 score; so, a venue that has two Art Museums will have a score of 10 x 2 = 20).

While the EC determined the scoring/weighting system, an additional layer of analysis will be the crime in that location. Neighborhoods that are deemed to have below average crime will be looked at more favorably when compared to Neighborhoods that have above average crime rates. The combination of Neighborhood scores and crime rates will be used to determine the ideal areas for potential expansion for SICK Co.

## Data Preparation and Exploratory Analysis

### FourSquare

In order to extract targeted features from the FourSquare API, we constructed a URL to send a request to FourSquare’s API for venue names, venue categories, and geographic location information such as latitude and longitude. **Table 1** shows a sample of the features obtained using the FourSquare API.

Table 1 – Venue Features Extracted from FourSquare (Random Row Sample)

|  |  |  |  |
| --- | --- | --- | --- |
| Venue | Venue  Latitude | Venue  Longitude | Venue  Category |
| Ed's Real Scoop | 43.660656 | -79.342019 | Ice Cream Shop |
| The Tulip Steakhouse | 43.666348 | -79.316854 | Steakhouse |
| Shisha&Co | 43.656748 | -79.374337 | Smoke Shop |
| Pizza Pizza | 43.706138 | -79.389292 | Pizza Place |
| Union Food Court | 43.644596 | -79.3812 | Food Court |
| Starbucks | 43.64099 | -79.376264 | Coffee Shop |
| I Deal Coffee | 43.655058 | -79.403254 | Coffee Shop |
| Old Navy | 43.77799 | -79.344091 | Clothing Store |
| Wish | 43.668759 | -79.385694 | Restaurant |
| Hudson's Bay | 43.65204 | -79.380391 | Department Store |

### Wikipedia

In order to extract from the Wikipedia URL the list of Toronto Postal Codes, Boroughs, and Neighborhoods, we used the Pandas library to extract the targeted table and create a Pandas dataframe. There were 77 Boroughs that were “Not Assigned.” In those instances, we removed those rows. When evaluating Neighborhoods, we set “Not Assigned” neighborhoods to “Queen’s Park.” In order to flatten the dataframe, we combined the Neighborhoods for each Postal Code (as shown in **Table 2** below showing the first ten rows).

Table 2 – Toronto Postal Codes with Boroughs and Neighborhoods (First Ten Rows)

|  |  |  |
| --- | --- | --- |
| Postal Code | Borough | Neighborhood |
| M1B | Scarborough | Rouge, Malvern |
| M1C | Scarborough | Highland Creek, Rouge Hill, Port Union |
| M1E | Scarborough | Guildwood, Morningside, West Hill |
| M1G | Scarborough | Woburn |
| M1H | Scarborough | Cedarbrae |
| M1J | Scarborough | Scarborough Village |
| M1K | Scarborough | East Birchmount Park, Ionview, Kennedy Park |
| M1L | Scarborough | Clairlea, Golden Mile, Oakridge |
| M1M | Scarborough | Cliffcrest, Cliffside, Scarborough Village West |
| M1N | Scarborough | Birch Cliff, Cliffside West |

Using the geospatial coordinates for each Postal Code, we can add latitude and longitude to our dataframe. **Table 3** shows the first five rows of our merged dataframe, which contains 103 unique Postal Codes and 274 venue categories.

Table 3 – Toronto Postal Codes with Latitude and Longitude (First Five Rows)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| PostalCode | Borough | Neighborhood | Latitude | Longitude |
| M1B | Scarborough | Rouge, Malvern | 43.806686 | -79.194353 |
| M1C | Scarborough | Highland Creek, Rouge Hill, Port Union | 43.784535 | -79.160497 |
| M1E | Scarborough | Guildwood, Morningside, West Hill | 43.763573 | -79.188711 |
| M1G | Scarborough | Woburn | 43.770992 | -79.216917 |
| M1H | Scarborough | Cedarbrae | 43.773136 | -79.239476 |

After importing the venue scores/weightings as provided by the EC, we can evaluate the Postal Codes that have the highest total score. Below, in **Figure 1**, is the bar plot of the Postal Codes with the top 20 scores. Then, **Figure 2**shows a bar plot of the Postal Codes with top five scores. The top five Postal Codes are M5J, M5K, M5X, M5H, and M5W.

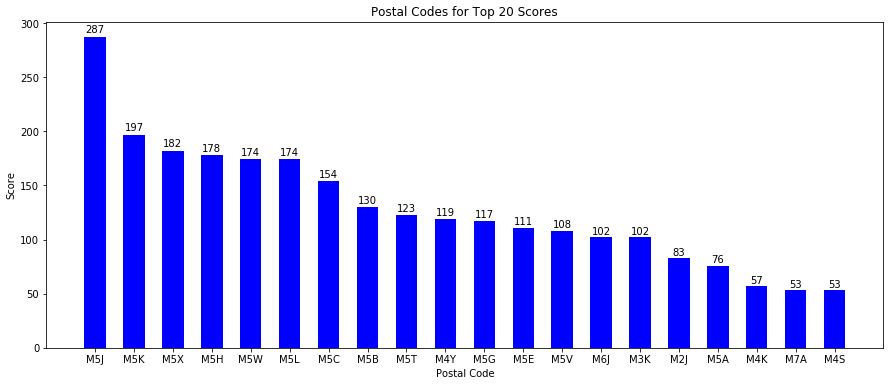


Figure 1 – Postal Codes with Top 20 Scores

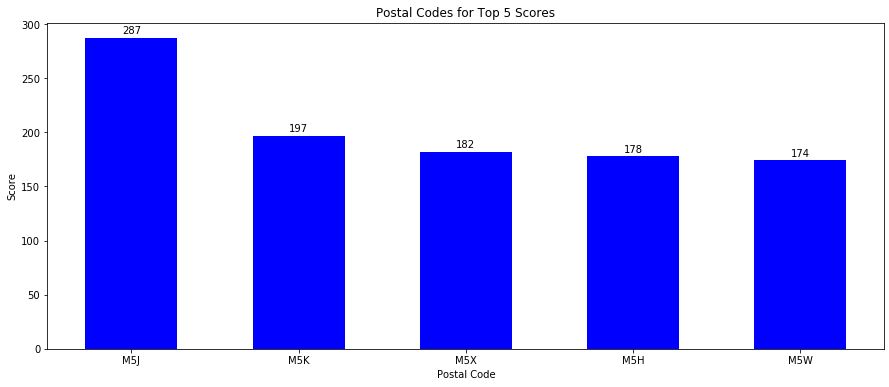


Figure 2 – Postal Codes with Top Five Scores

Let’s also take a look at composition of the major contributors to the score of each Postal Code. First, in **Figure 3**, we can see the major contributors to the score for Postal Code M5J (the Postal Code with the highest score) were Aquariums, Baseball Stadiums, Hotels, Basketball Stadiums and Scenic Lookouts.

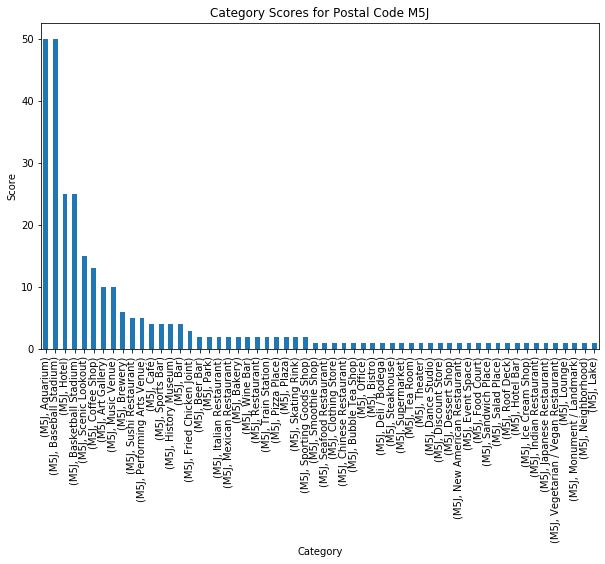


Figure 3 – M5J Postal Code Bar Plot with Scores By Category

Next, in **Figure 4**, major contributors to score for Postal Code M5K were Hotels, Basketball Stadiums, American Restaurants, Coffee Shops, Art Galleries, and Concert Halls.

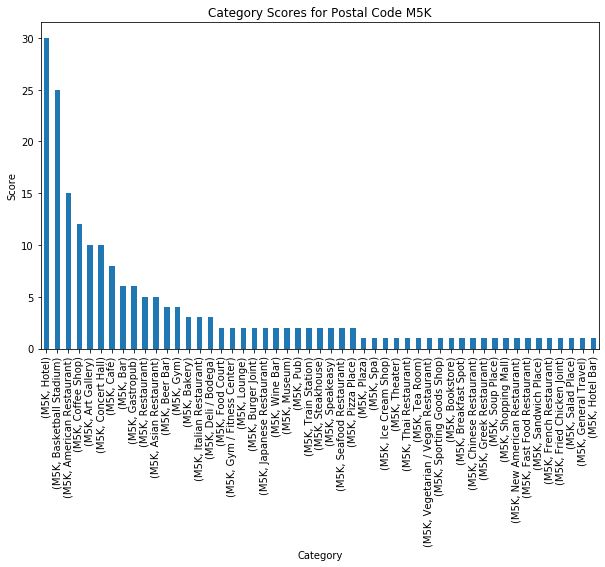


Figure 4 – M5K Postal Code Bar Plot with Scores by Category

A quick scan of **Figure 5** shows that the categories with the highest scores within Postal Code M5X are Hotels, American Restaurants, Asian Restaurants, Art Galleries, and Concert Halls.

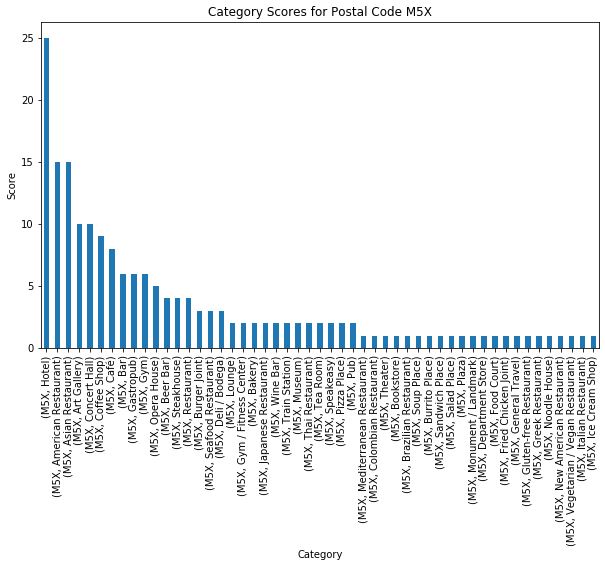


Figure 5 – M5X Postal Code Bar Plot with Scores by Category

**Figure 6** shows that the primary drivers for the total score found in Postal Code M5H are Hotels, American Restaurants, Sushi Restaurants, Art Galleries, and Art Museums.

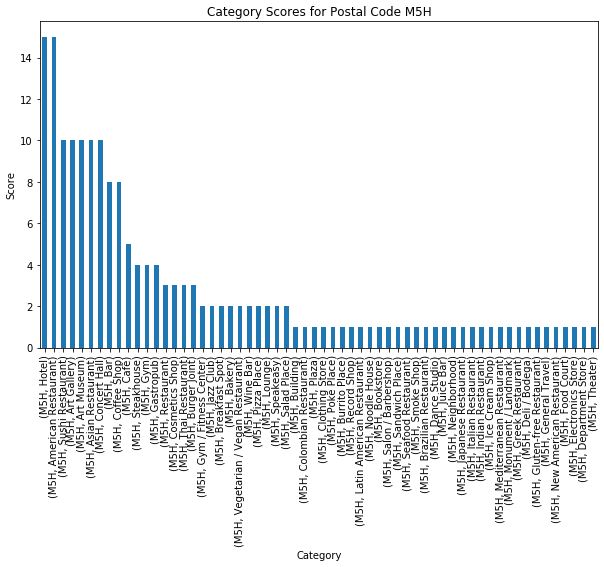


Figure 6 – M5H Postal Code Bar Plot with Scores by Category

Continuing on with the fifth highest scoring Postal Code, **Figure 7** shows that significant impacts to scoring for Postal Code M5W came from Basketball Stadiums, Art Galleries, Hotels, Coffee Shops, Cocktail Bars, and Beer Bars.

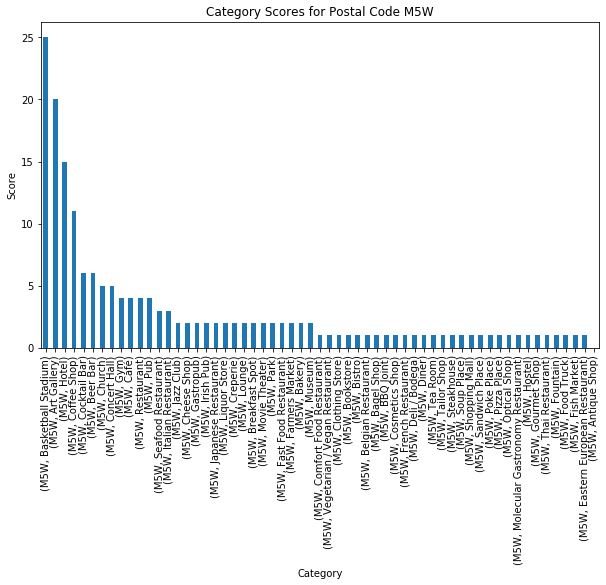


Figure 7 – M5W Postal Code Bar Plot with Scores by Category

### Toronto Police Service Public Safety Data Portal

Data found on the Toronto Police Service Public Safety Data Portal includes Major Crime Indicators (MCI). MCI categories include Assault, Break and Enter, Auto Theft, Robbery and Theft Over (excluding Sexual Assaults). In addition to MCI category, features include the location (nearest intersection), Building Type, Neighborhood, Police Division, and Occurrence Year/Month/Day/Hour. It should be noted that occurrence locations found in the data portal have been deliberately moved to the nearest road intersection, with the goal of protecting the privacy of the parties involved in the occurrence.

After creating a dataframe from the crime data (df\_cd) found on the Toronto Police data portal, we check for any data issues. There are 167,525 rows and 29 features (columns). **Table 4** shows the data types for the features after creating df\_cd shows the features included in the MCI data available on the Toronto Police Data Portal.

Table 4 – Data Types of Available Features in Toronto Crime Dataframe

|  |  |
| --- | --- |
| Feature | Object Type |
| X | float64 |
| Y | float64 |
| Index\_ | int64 |
| event\_unique\_id | object |
| occurrencedate | object |
| reporteddate | object |
| premisetype | object |
| ucr\_code | int64 |
| ucr\_ext | int64 |
| offence | object |
| reportedyear | int64 |
| reportedmonth | object |
| reportedday | int64 |
| reporteddayofyear | int64 |
| reporteddayofweek | object |
| reportedhour | int64 |
| occurrenceyear | float64 |
| occurrencemonth | object |
| occurrenceday | float64 |
| occurrencedayofyear | float64 |
| occurrencedayofweek | object |
| occurrencehour | int64 |
| MCI | object |
| Division | object |
| Hood\_ID | int64 |
| Neighbourhood | object |
| Lat | float64 |
| Long | float64 |
| ObjectId | int64 |

As we proceed with the data cleaning and preparation process, we notice that there are duplicate event\_unique\_id’s. After dropping these duplicate records, we are left with 145,817 rows. Further inspection of df\_cd shows 40 records that have NULL values, which is 0.03% of all records. **Table 5** shows the number of NULL values by feature, as well as the NULL % of total NULL rows by feature. After removing these NULL records, we are left with 145,777 rows.

Table 5 – Number of NULL Rows by Feature

|  |  |  |
| --- | --- | --- |
| Feature | Count of NULL Rows | NULL % of Total Rows |
| X | 0 | 0.00% |
| Y | 0 | 0.00% |
| Index\_ | 0 | 0.00% |
| event\_unique\_id | 0 | 0.00% |
| occurrencedate | 0 | 0.00% |
| reporteddate | 0 | 0.00% |
| premisetype | 0 | 0.00% |
| ucr\_code | 0 | 0.00% |
| ucr\_ext | 0 | 0.00% |
| offence | 0 | 0.00% |
| reportedyear | 0 | 0.00% |
| reportedmonth | 0 | 0.00% |
| reportedday | 0 | 0.00% |
| reporteddayofyear | 0 | 0.00% |
| reporteddayofweek | 0 | 0.00% |
| reportedhour | 0 | 0.00% |
| occurrenceyear | 40 | 0.03% |
| occurrencemonth | 40 | 0.03% |
| occurrenceday | 40 | 0.03% |
| occurrencedayofyear | 40 | 0.03% |
| occurrencedayofweek | 40 | 0.03% |
| occurrencehour | 0 | 0.00% |
| MCI | 0 | 0.00% |
| Division | 0 | 0.00% |
| Hood\_ID | 0 | 0.00% |
| Neighbourhood | 0 | 0.00% |
| Lat | 0 | 0.00% |
| Long | 0 | 0.00% |
| ObjectId | 0 | 0.00% |

A quick view of the available years shows that our area of focus should be years 2014 through 2018, since years prior to 2014 have limited data points (see **Table 6**). We will drop all records prior to year 2014.

Table 6 – Number of Crime Occurrences by Year

|  |  |
| --- | --- |
| Occurrence Year | Number of Occurrences |
| 2000 | 13 |
| 2001 | 10 |
| 2002 | 7 |
| 2003 | 8 |
| 2004 | 9 |
| 2005 | 8 |
| 2006 | 7 |
| 2007 | 16 |
| 2008 | 23 |
| 2009 | 28 |
| 2010 | 49 |
| 2011 | 66 |
| 2012 | 117 |
| 2013 | 452 |
| 2014 | 27,829 |
| 2015 | 28,045 |
| 2016 | 28,274 |
| 2017 | 29,746 |
| 2018 | 31,070 |

Now that we have cleaned and prepared our data, all systems are “go” for us to dive into deeper analysis and discover the story within the crime statistics. **Figure 8** shows the Total crimes by year for the remaining records in our dataframe.

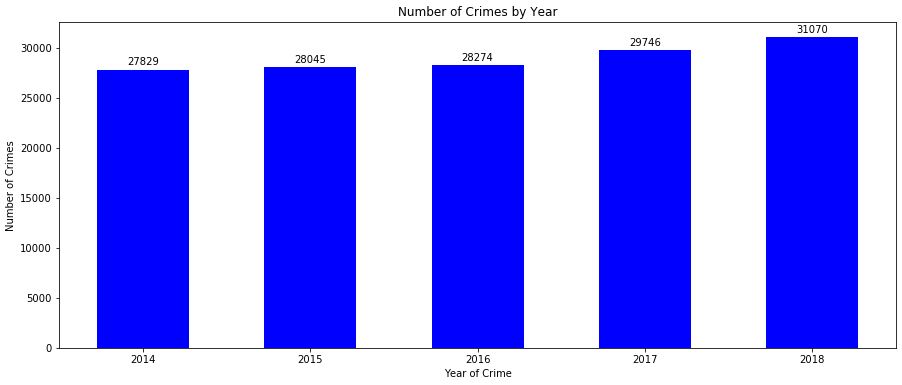


Figure 8 – Number of Crimes by Year in the Toronto Area

# Methodology

## K-Means Clustering Analysis

Our goal is to discover which Postal Codes have worse crime activity compared to other regions in Toronto. In order to help us with this goal (which also helps us determine where the Postal Codes with top five scores rank compared to each other), we will implement a k-means clustering analysis to uncover any overall crime trends for those areas.

K-means clustering analysis is a simple but useful tool to quickly discover insights from our unlabeled data set. But first, we have to normalize the data so it fits into our k-means clustering analysis applications. Normalization is a statistical method that helps mathematical-based algorithms (like k-means) interpret features with different magnitudes and distributions equally (which prevents the undesirable effects of features exhibiting a disproportionate contribution on the overall results).

When we group the crime data as-is, it looks like the information in **Figure 9**. Notice that MCI is a categorical variable, which is an unusable structure for the k-means algorithm because it uses the Euclidean distance function. In order to make the MCI variable usable, we’ll convert it using one hot encoding.



Figure 9 - Transforming Categorical Variable Using OneHot Encoding

## Elbow Plot Analysis

Once the data is normalized and we temporarily remove the Neighborhood feature (since it’s categorical/qualitative), we check for the optimal number of clusters, k, to use by performing an “Elbow” Plot Analysis. The Elbow method analyzes the number of k-means clusters, k, by plotting and examining the sum of squared distances from the cluster center (SSD) for a given k. When plotted, we can visually identify the optimal number of clusters, which should clearly show a point on the graph where reduction in SSD diminishes for increases in k (see **Figure 10**). This point is referred to as the "elbow." When observing the elbow plot in **Figure 10**, the optimal k clusters appears to be near two. We will use k=2 for our k-means analysis.

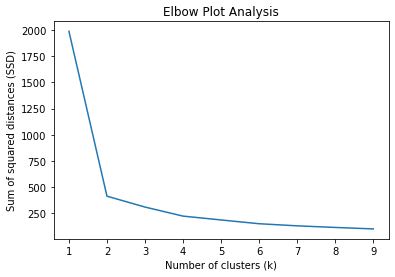


Figure 10 - Elbow Plot Analysis for OneHot MCI features

# Results

After using k=2 for our k-means clustering analysis, we can gain further insights by grouping the data based on the cluster labels and taking the mean of the normalized features yields **Table 7**. We can see that Label 0 has 56 Neighborhoods, while Label 1 has 85 Neighborhoods. When looking at the normalized data that was used for the analysis, we can use the following interpretation of the numbers:

* The negative values mean “lower than most” and positive values mean “higher than most.”
* Label 0 is the not so lucky label that has Neighborhoods higher than average MCI counts (i.e., more crime than average).
* Label 1 has Neighborhoods with lower than average MCI counts (e.g., assault, auto theft, break and enter, robbery and theft over). So, it has less crime (on average) than other cohorts.

Table 7 – K-means Labeling Results for Normalized Toronto Crime Occurrences

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Labels | Assault | Auto\_Theft | Break\_and\_Enter | Robbery | Theft\_Over |
| 0 | 0.266645 | 0.312836 | 0.200536 | 0.25538 | 0.172422 |
| 1 | -0.175672 | -0.206103 | -0.132118 | -0.16825 | -0.113595 |

Now, we have to make the connection between the Neighborhoods provided by the Toronto Police Service Data Portal and the Postal Codes we identified using FourSquare. After visually inspecting the location of the top five Postal Code scores, it was determined that they're located in the following Toronto PD Neighborhoods:

* Waterfront Communities-The Island (77)
* Bay Street Corridor (76)

Let's see how these two Neighborhoods were labeled using k-means. Table 8 shows the k-means Labels. Based on the Labels, Bay Street Corridor (Label = 1) has lower than average MCI counts. However, Waterfront Communities-The Island (Label = 0) has higher than average MCI counts. In light of this fact, Postal Codes that are in the top 5 scores and are located in Bay Street Corridor would be more desirable for the expansion location due to lower relative rates of crime.

Table 8 – Neighborhood Scores and K-Means Labels

|  |  |  |  |
| --- | --- | --- | --- |
| Postal Code | Toronto Police Department Neighborhood | Score | Label |
| **M5J** | Waterfront Communities-The Island (77) | 287 | 0 |
| M5K | Bay Street Corridor (76) | 197 | 1 |
| M5X | Bay Street Corridor (76) | 182 | 1 |
| M5H | Bay Street Corridor (76) | 178 | 1 |
| **M5W** | Waterfront Communities-The Island (77) | 174 | 0 |

# Discussion

Our recommendation is to choose the new SICK Co. corporate to be located in the Bay Street Corridor Neighborhood (Postal Code M5K). Ultimately, employee safety should be one of the most important factors when deciding where to expand corporate locations. In this case, while the Waterfront Communities-The Island Neighborhood has the highest total score (287 based on the weightings provided by the EC), it resides in an area that has higher crime activity (based on our k-means analysis) when compared to the Bay Street Corridor Neighborhood.

Table 8 – Neighborhood Scores and K-Means Labels

|  |  |  |  |
| --- | --- | --- | --- |
| Postal Code | Toronto Police Department Neighborhood | Score | Label |
| **M5J** | Waterfront Communities-The Island (77) | 287 | 0 |
| M5K | Bay Street Corridor (76) | 197 | 1 |
| M5X | Bay Street Corridor (76) | 182 | 1 |
| M5H | Bay Street Corridor (76) | 178 | 1 |
| **M5W** | Waterfront Communities-The Island (77) | 174 | 0 |

# Conclusion

Choosing the expansion location in the Bay Street Corridor Neighborhood (Postal Code M5K) is consistent with SICK Co.’s goal of selecting an area that has one of the highest scores as well as entering a relatively safe area with relatively lower crime occurrences when compared to the top five scores. These findings are supported by the k-means clustering algorithm which supplied insights for our unlabeled data set. Final expansion steps should commence at the earliest convenience of the EC.