# Introduction

## Background

Supreme/Successful/Smart/Swag Intelligent/Ingenuity/Iffy Craft/Crafty/Candy/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 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. This expansion is critical to the success of SICK Co. As such, an 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 point system for 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 site.

# Data

where you describe the data that will be used to solve the problem and the source of the 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 (see Table [XX]). Data extraction from FourSquare entailed querying its API using protocols found within their documentation. Once extracted, we will merge and/or compare the data with our other sources.

Merged the data between

The Wikipedia List of Postal Codes includes Postal Codes, as well as 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 (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 potential 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

### FourSquare

ADD .HEAD() TABLE OF FOURSQUARE DATA

ADD .HEAD() TABLE OF WIKIPEDIA DATA

### 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 1** below showing the first ten rows).

**Table 1 -- 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 2** shows the first five rows of our merged dataframe, which contains 103 unique Postal Codes and 274 venue categories.

**Table 2 – 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 box plot of the Postal Codes with the top 20 scores. Then, **Figure 2**shows a box plot of the Postal Codes with top five scores. The top five Postal Codes are M5J, M5K, M5X, M5H, and M5W.

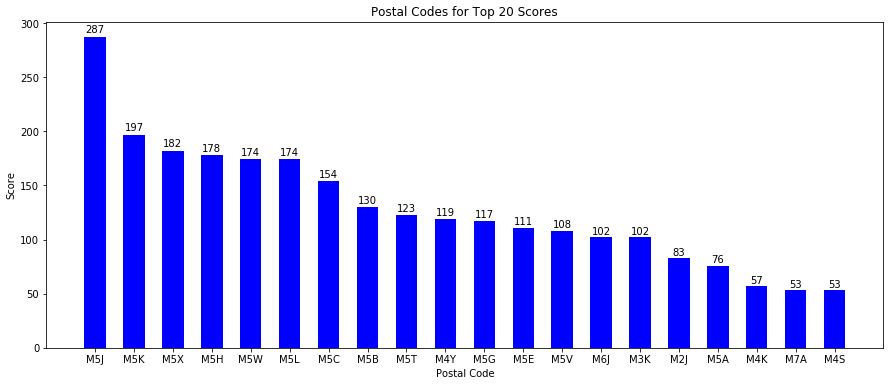


Figure 1. Postal Codes with Top 20 Scores

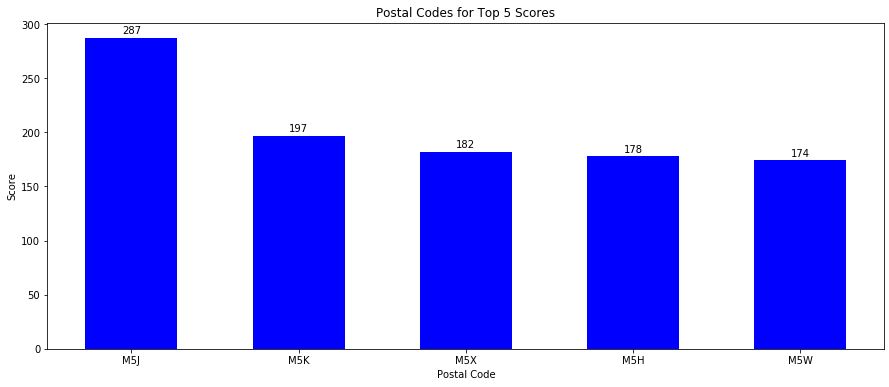


Figure 2. Postal Codes with Top Five Scores

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.

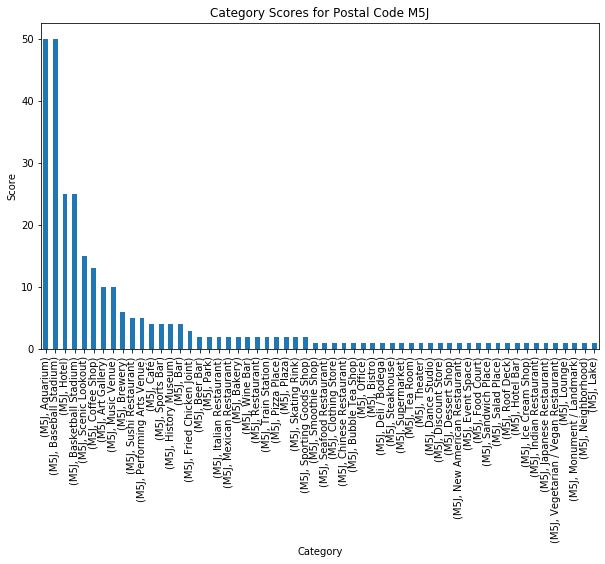


Figure 3. M5J Postal Code Box Plot with Scores By Category

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.

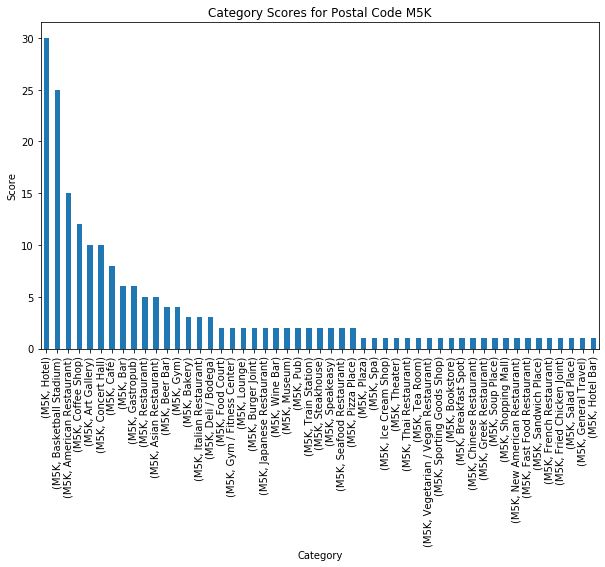


Figure 4. M5K Postal Code Box Plot with Scores by Category

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.

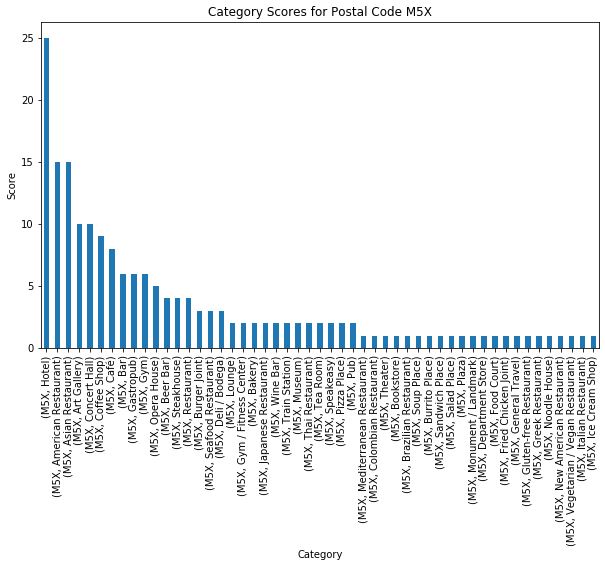


Figure 5. M5X Postal Code Box Plot with Scores by Category

**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.

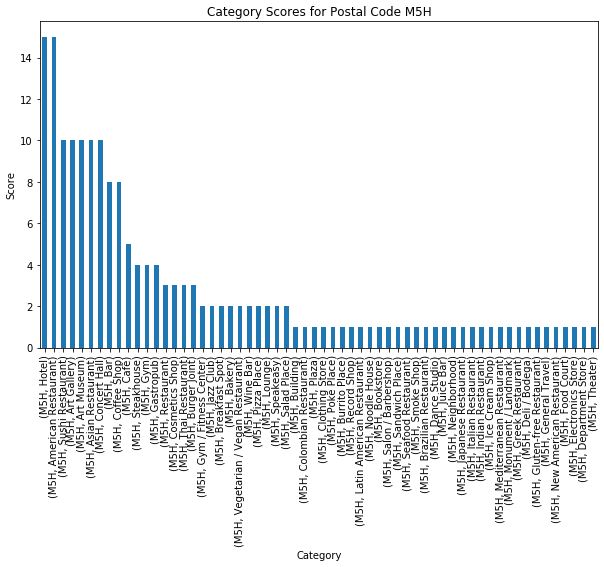


Figure 6. M5H Postal Code Box Plot with Scores by Category

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.

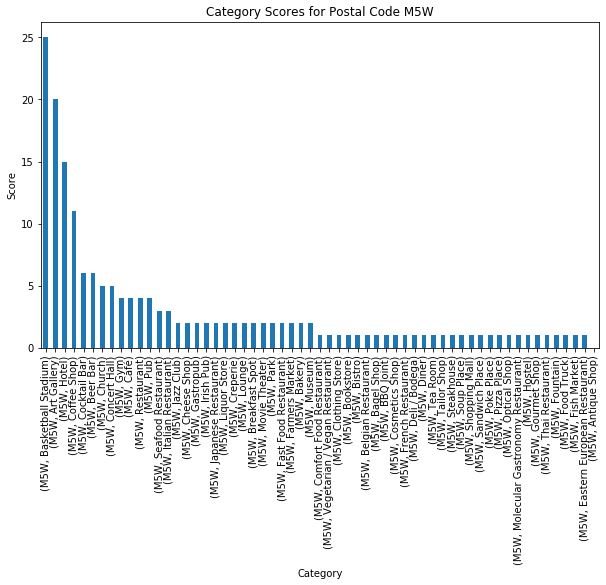


Figure 7. M5W Postal Code Box Plot with Scores by Category

### 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 found on the Toronto Police data portal, we check for any data issues. There are 167,525 rows and 29 features (columns). **Table 3** shows the data types for the features after creating the dataframe shows the features included in the MCI data available on the Toronto Police Data Portal.

**Table 3 – 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 |

There are duplicate event\_unique\_id’s. After dropping these records, we are left with 145,817 rows. Further inspection shows 40 records that have NULL

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 4**). We will drop all records prior to year 2014.

**Table 4 – Number of 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 |

TPD – CONTINUE FROM NOTEBOOK LINE [110]

CRIME DATA MULTIPLE IDS…DELETED DUPLICATES

MISSING DATA IN 0.03% OF DATA…DELETED ROWS

ONEHOT

ADD .HEAD() TABLE OF TORONTO PD CRIME DATA

ADD BARPLOTS OF TORONTO PD CRIME DATA

FORMAT NEEDED FOR K-MEANS (NORMALIZATION)

# Methodology

section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.

Discuss Elbow Analysis

"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 number of clusters (k). When plotted, we can visually identify the optimal number of clusters, which should clearly show a point on the graph where reduction in SSE diminishes for increases in k. This point is referred to as the "elbow."

ADD K-MEANS ELBOW PLOT

# Results

section where you discuss the results.

Top 20 postal codes

Top 5 postal codes

k-means unveils higher than average crime areas and lower than average crime areas

# Discussion

section where you discuss any observations you noted and any recommendations you can make based on the results.

# Conclusion

section where you conclude the report.