Café B2D

Toronto, ON

Created by: David Unrau

Date: April 27, 2020

Version: 1.0

Contents

[2 Introduction 2](#_Toc38896704)

[3 Data 2](#_Toc38896707)

[4 Methodology 3](#_Toc38896711)

[5 Results 7](#_Toc38896716)

[6 Discussion 7](#_Toc38896717)

[7 Conclusion 8](#_Toc38896718)

# Introduction

## Background

**Café B2D** (Bean 2 Drip) is an up-and-coming café specializing in high end coffee related beverages, exceptional pastries and made-to-order items.  The client who is opening the new café comes with over 20 years in the industry and has become a leader in the field. They have selected Toronto as the starting point for their venture as it is their hometown and Toronto has the market to support this new take on a café.

## Problem

With the market already overly saturated with coffee shops and cafés within Toronto, it has proven to be challenging to locate the right neighbourhood in which to set up shop. This project intends to provide appropriate neighbourhood selections based on the following criteria:

1. Average yearly income within the neighbourhood must be greater than or equal to $150,000
2. Location to not be within an area that is overly saturated with other venues
3. The population within the neighbourhood (including surrounding areas) needs to be in excess of 7,500

Out of scope analysis for this portion of the project (but will need to be reviewed in the future prior to final decision) includes:

1. Real estate availability
2. Zoning restrictions
3. Overall attractiveness of the area (does it fit with the brand)

# Data

## Description

The data that is required for this project will need to include the demographic data of the Toronto neighbourhoods, the geographic data including the mapping as well the venue information. This data will be able to be used to generate the necessary results and conclusions for the client.

## Data Sources

The data for the demographic and geographic details (excluding the actual geojson data which will be found elsewhere) of the neighborhoods can be found on the Toronto open data platform. There is a lot of information located in these tables so we will be trimming them down to the information that we need to accomplish our goals. The demographic data can be located at <https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/ef0239b1-832b-4d0b-a1f3-4153e53b189e?format=csv> and the geographic data can be located at <https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/a083c865-6d60-4d1d-b6c6-b0c8a85f9c15?format=csv&projection=4326>.

The geojson data (the data used for created stylized maps) is located on GitHub at <https://raw.githubusercontent.com/jasonicarter/toronto-geojson/master/toronto_crs84.geojson>. This data is required in order to create the choropleth maps which will visually show the areas that fit the criteria.

The other source of data that we require is the venue data to help us determine the saturation of the market within each of the neighbourhoods. This data will be retrieved from foursquare.com utilizing their api. In order to get all of the data that we require, we will be searching based on *coffee*, *Hortons*, *café* and *second cup*. The purpose of the multiple searches is to ensure that all of the data is captured.

## Cleaning

Within the cleaning stage of the project, we will be taking the data that is retrieved and transforming it into workable data. A lot of the columns will not be required (i.e. the demographic data has over 2000 columns that had to be transposed to rows and then majority of the data needed to be dropped).

Column names are another piece to get cleaned in order to make the data make more sense for the project. Column data types will also be updated to ensure that some of the data can be used for calculations and further analysis.

Additional calculated columns will be generated and appended to the dataframes in order to ensure that we have all of the data that will be required in order to make informed decisions.

# Methodology

## Data Review

With the data gathered and cleaned, we can start to see what it is showing us. We can see that the data shows that there are five neighbourhoods that fit the criteria. Those neighbourhoods are along with their attributes can be found below in Table 1.

Table 1: Candidate Neighbourhoods:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **ID Number** | **Population** | **Land Area** | **Average Income** | **Radius** | **Link** | **Longitude** | **Latitude** |
| Bridle Path-Sunnybrook-York Mills | 41 | 9266 | 8.91 | 308010 | 1492 | Bridle Path-Sunnybrook-York Mills (41) | -79.378904 | 43.731013 |
| Casa Loma | 96 | 10968 | 1.93 | 165047 | 695 | Casa Loma (96) | -79.408007 | 43.681852 |
| Forest Hill South | 101 | 10732 | 2.45 | 204521 | 783 | Forest Hill South (101) | -79.414318 | 43.694526 |
| Lawrence Park South | 103 | 15179 | 3.24 | 169203 | 900 | Lawrence Park South (103) | -79.406039 | 43.717212 |
| Rosedale-Moore Park | 98 | 20923 | 4.65 | 207903 | 1078 | Rosedale-Moore Park (98) | -79.379669 | 43.68282 |

The data is also showing that the neighbourhood average yearly income ranges from $165k to just over $300k per year, making these 5 neighbourhoods excellent candidates as per the client’s requirements.

In addition the average yearly income, the population sizes range from just over 9,000 to over 20,000 making these neighbourhoods large enough to sustain the new venue.

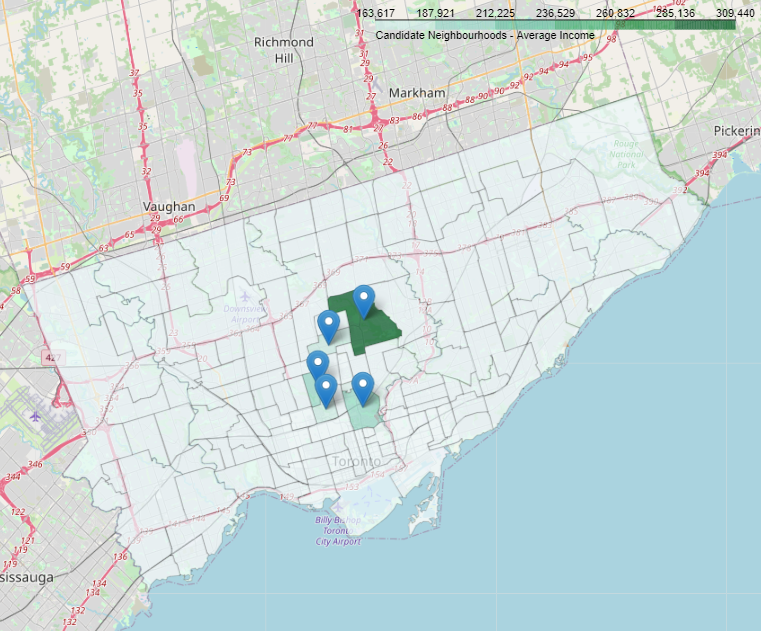
We have also gathered the Foursquare data and will utilize this information into further saturation calculations, maps and plot analysis.

## Maps

With the data retrieved and cleaned, now is a good time to visually represent the locations of the neighbourhoods and the venues.

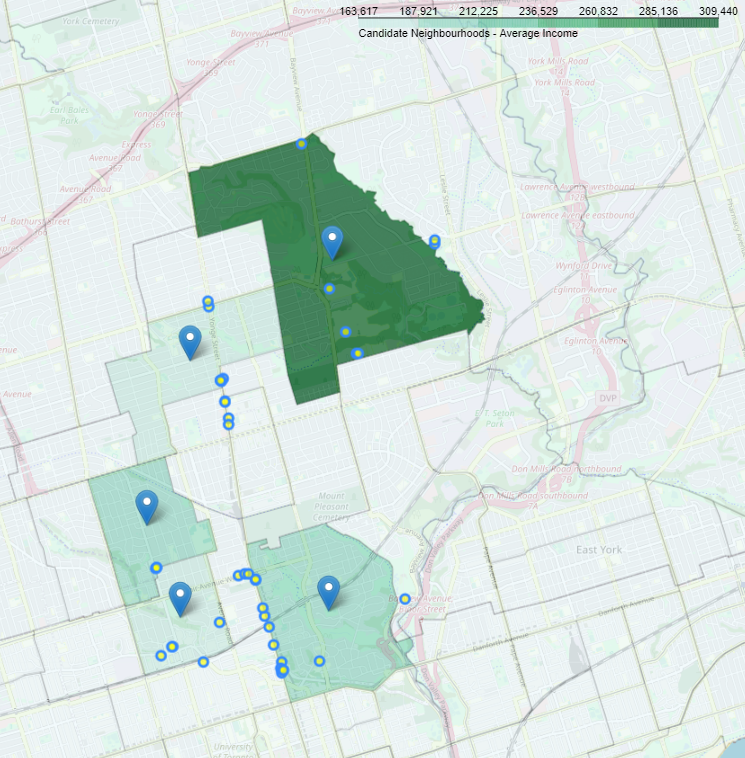
First, we will start with just the five selected neighbourhoods. As you can see in Figure 1, the neighbourhoods are all within close proximity to each other and are located centrally within the whole of Toronto. The darker the colour of green, the higher the yearly average income is within the neighbourhood.

Figure 1: Selected Neighbourhoods



Now we will overlay the venues and you will start to see which neighbourhoods are more saturated with other venues than others. As it was in Figure 1, the darker the colour within Figure 2 refers to the higher the yearly average salary. You will also notice the smaller blue dots with yellow centers, this makes up the venues within each of the given neighbourhoods.

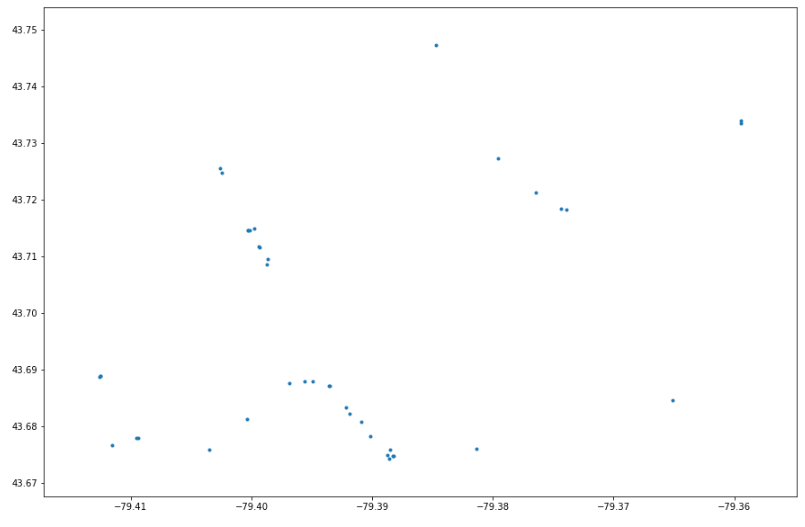
Figure 2: Neighbourhoods with Venues



## Scatter Plot

A scatter plot can be a good representative of how all the points relate to each other. In Figure 3, it displays the locations of all of the venues according to their longitude (x axis) and latitude (y axis). In the next section, we will show the clustering and how each of the neighbourhoods relate to the venues.

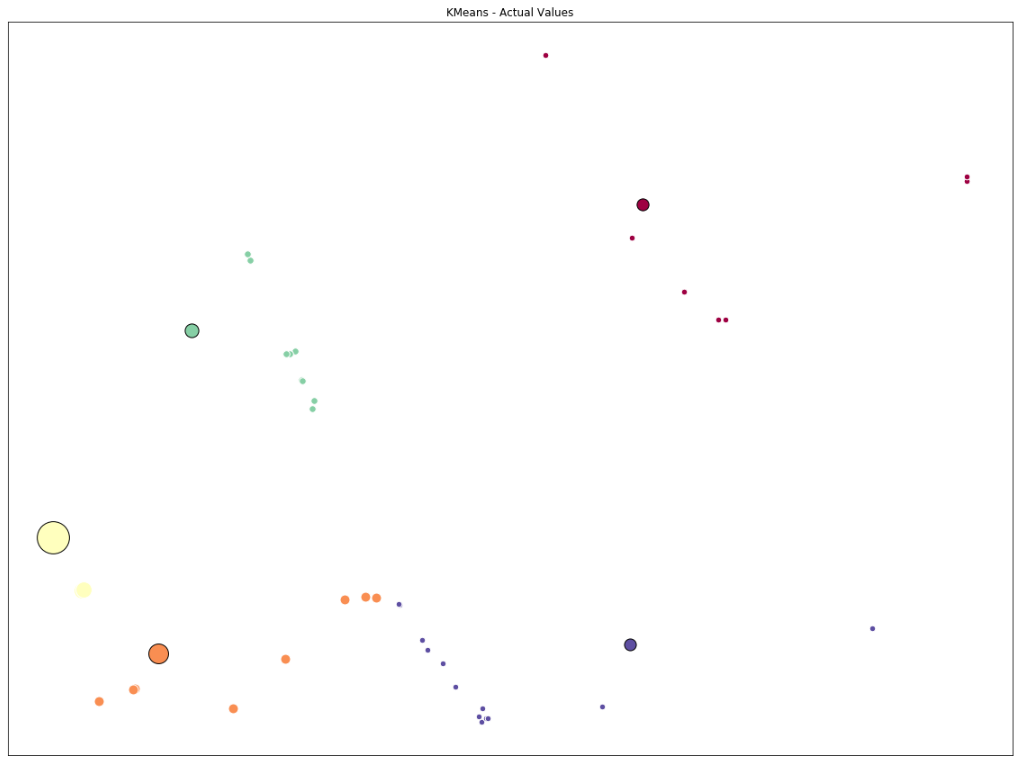
Figure 3: Scatter Plot



## Relationship of Venues to Neighbourhood (k-means)

With the utilization of k-means, we are able to create a cluster diagram and show the relationship of the neighbourhood to all of the venues located within. Figure 4 shows the clustering of each of the neighbourhoods and their venues. The larger coloured circles in each cluster represents the longitude and latitude marker of the neighbourhood and the smaller coloured circles represent the venues. The larger the neighbourhood marker is, the less saturated the marker is making those neighbourhoods a more probable candidate.

Figure 4: K-means Cluster Diagram



## Calculated Values

We are at the point that we have enough data that we can calculate the venue saturation within each of the neighbourhoods, found in Table 2 below. As it shows, the highest average yearly income neighbourhood is not the best candidate.

Table 2: Calculated Saturation

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **ID Number** | **Population** | **Land Area** | **Average Income** | **Radius** | **Venue Count** | **Population per Venue** |
| Bridle Path-Sunnybrook-York Mills | 41 | 9266 | 8.91 | 308010 | 1492 | 7 | 1324 |
| Casa Loma | 96 | 10968 | 1.93 | 165047 | 695 | 5 | 2194 |
| Forest Hill South | 101 | 10732 | 2.45 | 204521 | 783 | 3 | 3577 |
| Lawrence Park South | 103 | 15179 | 3.24 | 169203 | 900 | 10 | 1518 |
| Rosedale-Moore Park | 98 | 20923 | 4.65 | 207903 | 1078 | 16 | 1308 |

# Results

With the data that was gathered, and the methodologies that we employed, we can safely recommend the following candidate neighbourhoods as outlined within Table 3 below.

Table 3: Final Recommendation

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Name** | **ID Number** | **Population** | **Land Area** | **Average Income** | **Radius** | **Venue Count** | **Population per Venue** |
| Forest Hill South | 101 | 10732 | 2.45 | 204521 | 783 | 3 | 3577 |
| Casa Loma | 96 | 10968 | 1.93 | 165047 | 695 | 5 | 2194 |
| Lawrence Park South | 103 | 15179 | 3.24 | 169203 | 900 | 10 | 1518 |

The population per venue for Forest Hill South shows that the area is not highly saturated with other venues as there are only three listed for it. This puts the population per venue at 3,577 which is almost 1,400 more than the next closest location.

# Discussion

With the data that is displayed above in Table 3, we can see that there are 3 potential candidates that could support the client’s new venue.

In addition to the data that is listed in the table, we can also refer back to the Figure 2 and the close proximity of all the neighbourhoods. Forest Hill South is the darker green section located in the lower left quadrant of the image.

Casa Loma is located directly south and Lawrence Park South is located directly north of Forest Hill South. This makes Forest Hill South stand out even further as it could draw in people from the surrounding neighbourhoods.

It would be our recommendation to review Forest Hill South in a more central to central-west location to find any potential locations to set up shop.

# Conclusion

The purpose of this project was to find suitable neighbourhoods for our client to opent their new shop and we were able to provide three potential neighbourhoods that are directly linked to each other making the location even more attractive. By calculating the population per current venue, we were able to determine the venue saturation within each of the selected neighbourhoods.

Other characteristics will be used by the client in the final decision of location which may include real estate availability, zoning restrictions and overall attractiveness of the area (does it fit the brand).