Selecting a Restaurant Location in Toronto

Table of Contents

Introduction

Background

Restaurants are a common form of small business that provides many benefits. The owner of a restaurant has a potential revenue source and creative outlet. A restaurant's presence also provides its community with an alternative source of food and a potential social meeting place. The location of a restaurant is a significant factor in its success.

Problem

Understanding the revenue potential for a new restaurant location is foundational piece of information when deciding where to open a location. This project aims to provide insight in to the revenue potential of restaurant locations in Toronto.

Interest

Entrepreneurs looking to start their first business or corporations looking for a new location will be interested in understanding the revenue potential of a particular location. This analysis could also provide value for anyone looking to perform small business planning, such a government agencies looking to change policies.

Data acquisition and processing

Data sources

The following information is required to perform analysis: spatial breakdown, financial attributes associated with the spatial breakdown, and venues associated with the spatial breakdown.

Spatial information will be based on the Canadian postal code breakdown for the city of Toronto and secondary geospatial data set. The postal code data set is available on Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. Geospatial data for the city of Toronto is available from the following source: https://cocl.us/Geospatial_data

Financial information will be acquired from the Canadian census data sets: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/download-telecharger/comp/page_dl-tc.cfm?Lang=E

Existing venue information will be acquired from the Foursquare APIs. Since this data set is crowd sourced it is dynamic. Analysis using this data set may yield different results depending on when the analysis is performed.

Data processing

When correlating multiple data sets they will not always match perfectly. There will be mismatches between the data sets and decisions need to be made about how to address these anomalies.

The information from the Canadian census had multiple data sets available. Some of these sets were not viable because they required access to a paid data set. Other data sets did not contain the financial information required. The Census_2016_Forward_Sortation_Area.zip file was identified as being able to link postal codes with income information. The size of this data set prevented it from being loaded directly for querying. Use of the accompanying index file was required to identify the subsections in the complete data set that were associated with a particular postal code. The sections for a particular postal code needed to be read and then further processing could be done.

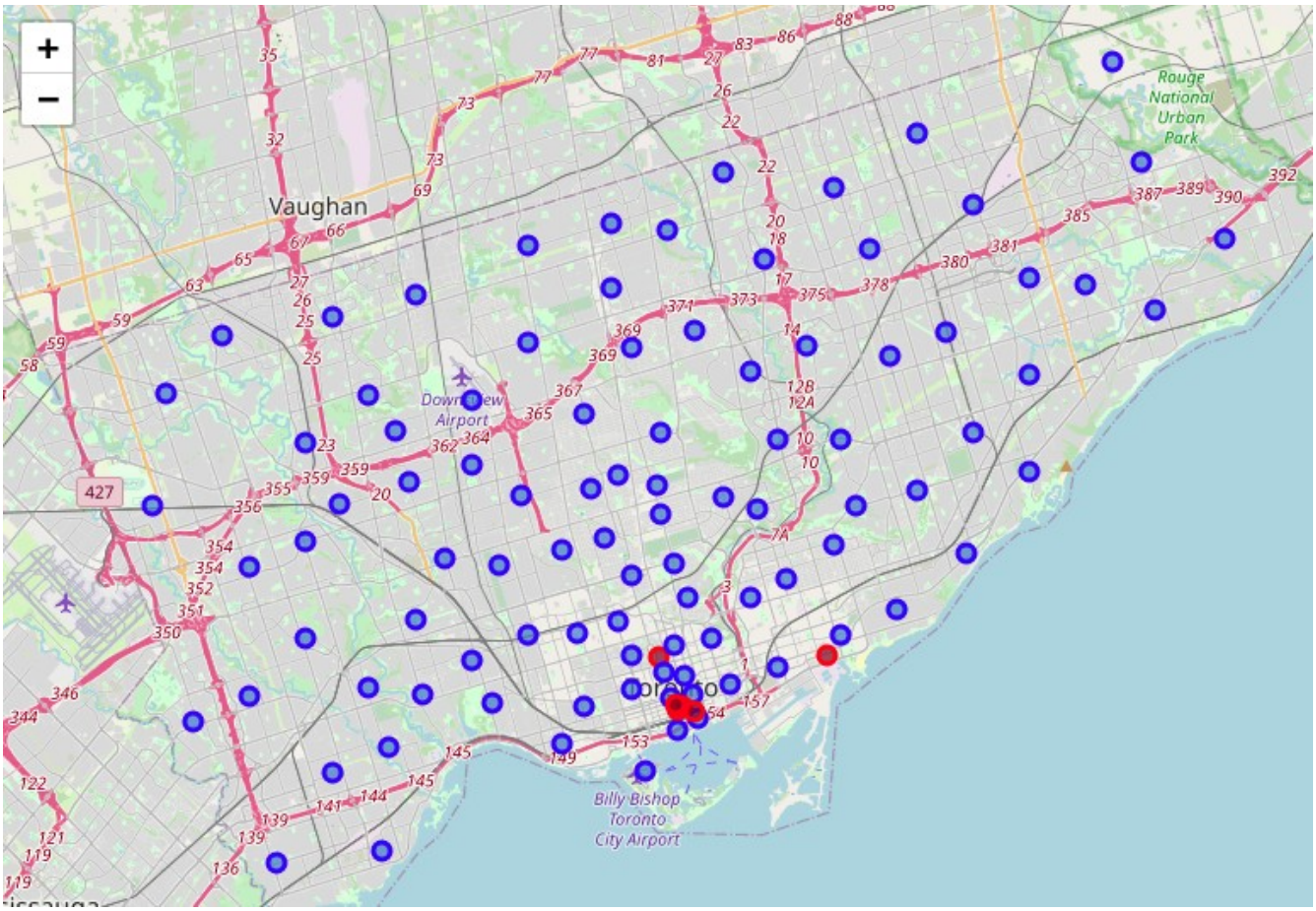
There were multiple adjustments required using the 2016 Census data:

- Since the census data was from 2016 there appear to have been changes to the available postal codes. Any postal codes that could not be matched to the census data were dropped since income data is required.
- Not all of the postal codes had income data. These postal codes were removed from the data set. The viability of this decision was confirmed with exploratory analysis.

Exploratory Data Analysis

While reviewing the data sets there were some interesting points that needed to be reconciled.

The Canadian Census information contained multiple types of income information. After reviewing the different types of income information the value for average income was chosen over median income because it represented the potential disposable income for a given postal code.



Some of the income records in the Canadian Census data were invalid values and could not be parsed. The postal codes associated with these missing records were rendered on a map (red circles). The postal codes with missing income information are in close proximity to other postal codes with income. The venue search radius (1km) large enough that the postal codes with missing information can easily be aliased with neighboring postal codes. As such, the postal codes without income data will not be included in the analysis.

The venue information returned from the Foursquare API identified over 300 unique venue categories. Not all of these venues would be considered competition for a new restaurant so the list needed to be reviewed and a subset of these venue categories identified as competing establishments. The substrings ‘Restaurant’, ‘Steakhouse’, and ‘Bistro’ were selected as keywords to identify competing establishments.

Analysis

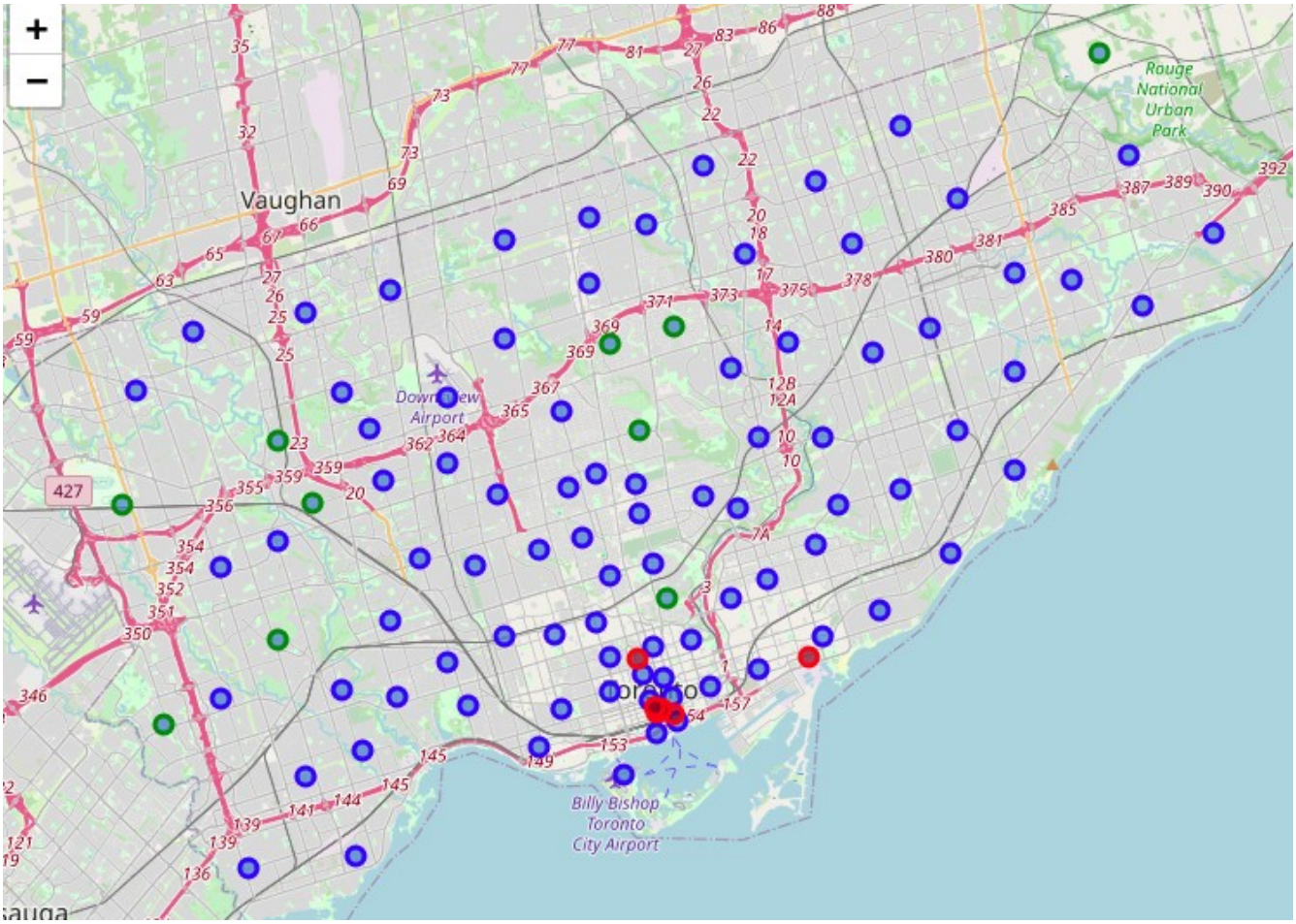
The postal code data set was extended by the following series of operations:

- Appending columns with lat/lon information for each postal code.
- Culling based on the available census information (postal code M7R did not exist in the census data and was removed from the postal code data set).
- Appending an ‘income’ column based on the census data. When census data was not available the the sentinel value NaN was assigned to income.
- Appending a column that contained the number of existing competing venues identified for each postal code.
- Appending a column that represented the relative revenue (income / (competing venue count + 1)) to represent the relative revenue that a new restaurant could expect. For postal codes with no income information the relative revenue was set to zero.

The resulting data set was sorted by relative revenue and the top ten locations were identified:

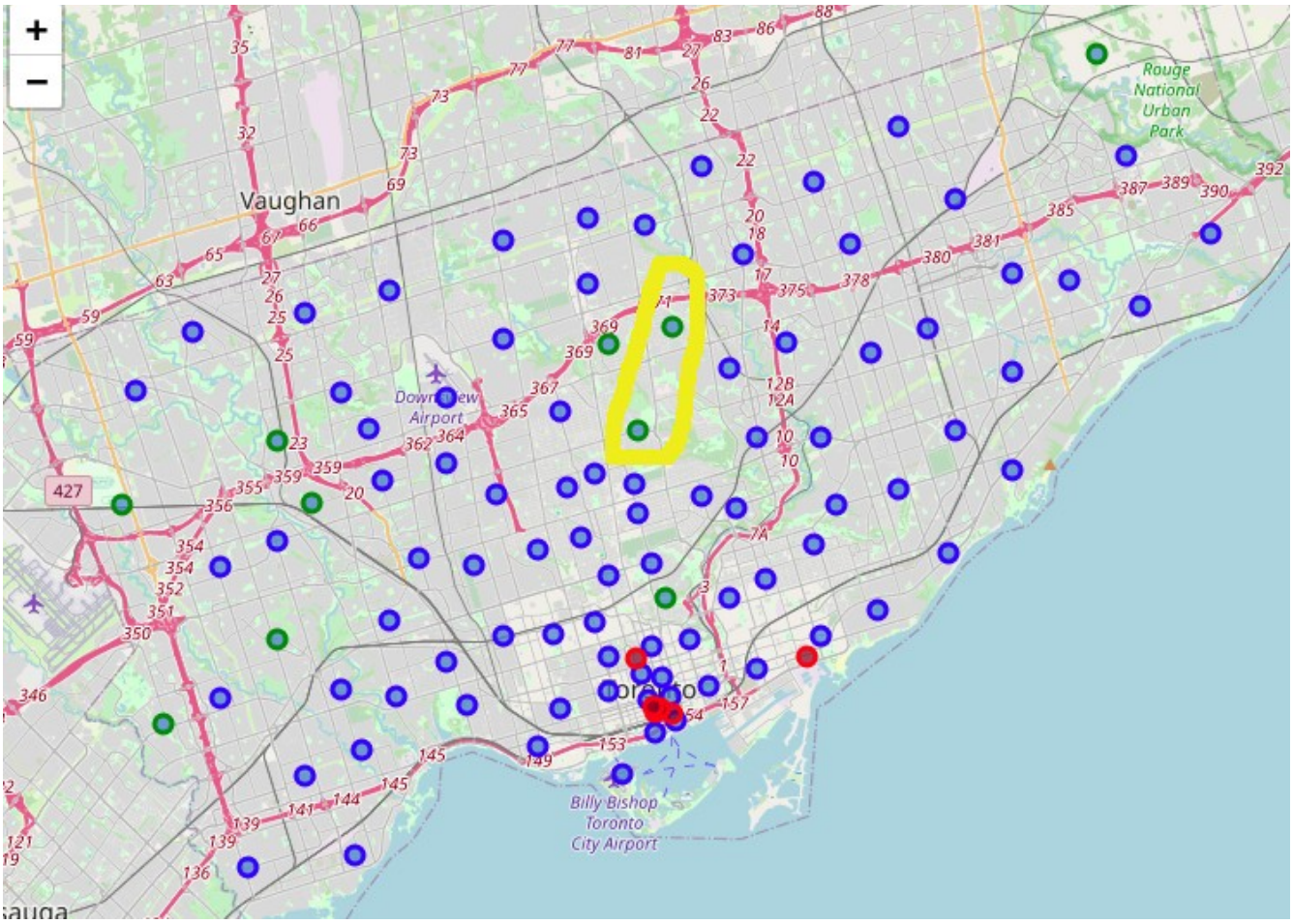
Postcode	Relative Revenue	Competing Venue Count	Average Income
M2L	$306,301.00	0	$306,301.00
M4N	$203,739.00	1	$407,478.00
M9A	$160,481.00	0	$160,481.00
M1X	$105,913.00	0	$105,913.00
M9C	$98,891.00	0	$98,891.00
M4W	$89,832.75	3	$359,331.00
M9W	$77,220.00	0	$77,220.00
M9M	$73,319.00	0	$73,319.00
M2P	$67,243.00	3	$268,974.00
M9N	$65,571.00	0	$65,571.00

The following map shows the chosen locations in green:



There is a fairly even distribution of locations across the city which is an indication that the method of the analysis are valid.

The top locations are located near the center of the city, but not in the dense downtown area (region highlighted in yellow):



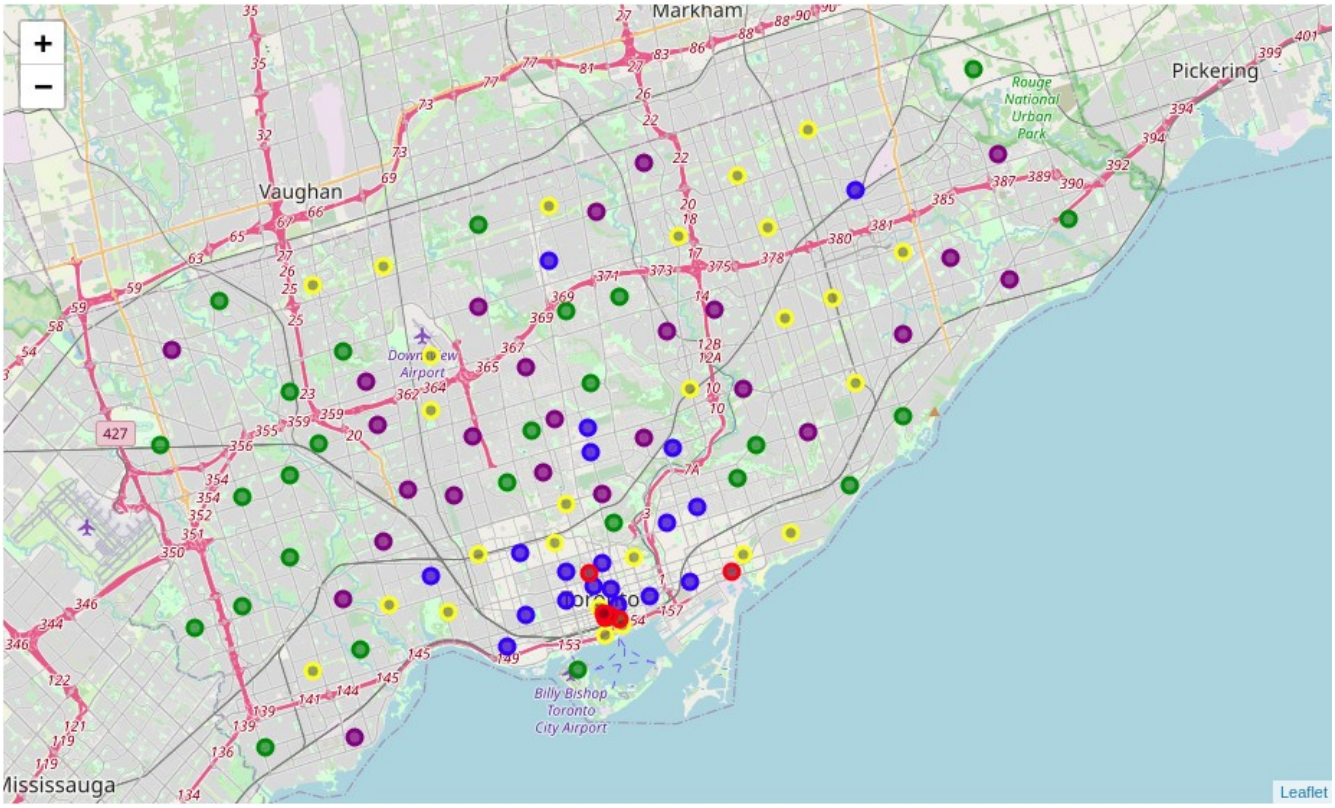
The following map classifies the locations in to quartiles based on the relative revenue potential. The locations without income information are noted in red. The quartiles are defined as follows:

GREEN – top 25% (most relative revenue potential)

PURPLE – 26-50%

YELLOW – 75-51%

BLUE – 100-76% (bottom 25%, least relative revenue potential)



The distributions for the quartiles are relatively evenly distributed which indicates a that the method of analysis is valid.

It should be noted that the core downtown area near the waterfront appears to represent the area of worst relative revenue potential. This likely due to a high number of competing venues in close proximity (additional analysis would be required to confirm).

Conclusions

This analysis was able to combine income data with geospatial data to determine which areas in Toronto may present areas of the city where a restaurant could generate large relative revenue. All of the locations in the city were classified and the entirety of the data set can be used for other analysis (such as governmental business development planning). Since the locations identified are not in close proximity to postal codes with missing income information, it is reasonable to assume that discarding that missing data was a viable decision.

Future Directions

Analysis of the results revealed that many of the identified locations have no competing restaurants. Understanding why certain postal codes have no competing venues could impact the viability of actually opening a restaurant in that location (e.g. that postal code does not have a commercial district that could support a restaurant). Also, the Census data set contains significant demographic information for each postal code. This demographic information could be further leveraged to refine the analysis.