Analyzing Toronto Venues and Neighbourhoods

1. Introduction

1.1 Background

Toronto, Canada is one of largest cities in North America and the largest in Canada. A financial and business hub in North America, it is central and key to the economy and people of Canada. While Toronto boasts an extremely high GDP ($330 billion in 2013), it lacks in late night activities for residence to leisure. Each year there are a number of new bars and clubs situated in Toronto. With the current economic upturn, there is an opportunity for new bars to open. The success of these venues are dependent on the neighbourhoods they are located in. Therefore, it is advantageous for potential investors to accurately predict how successful a bar will be in any given neighbourhood.

1.2 Problem

Data that may contribute to determining the success of a bar might include the average income of the neighbourhood, population density within the neighbourhood, the number of bars in the neighbourhood, and the number of restaurants within the neighbourhood. This project will aim to predict which neighbourhood will bring the most success for opening a potential bar.

1.3 Interest

Potential investors will be the most interested in the accurate prediction of success. Other people who are interested in opening their own venue may also be interested.

2. Data Scraping and Scrubbing

2.1 Data Sources

The data for different neighbourhoods in Toronto and demographics of Toronto neighbourhoods can be found here and here, respectfully. I then used the Foursquare API to gather different types of venues based on the location of the neighbourhoods.

2.2 Data Scrubbing

The neighbourhood data were combined into one table with all NaN values being dropped. The venues data were gathered based on the first 100 venues close to the centre of a certain neighbourhood. After this data was concatenated, it removed any NaN values as well as duplicate values. Both data sets used the Postal Code as the link.

There were some issues with the scrubbing and cleaning process. First, multiple venues were being shown for the same venue and there was no way to automate the removal of duplicates.

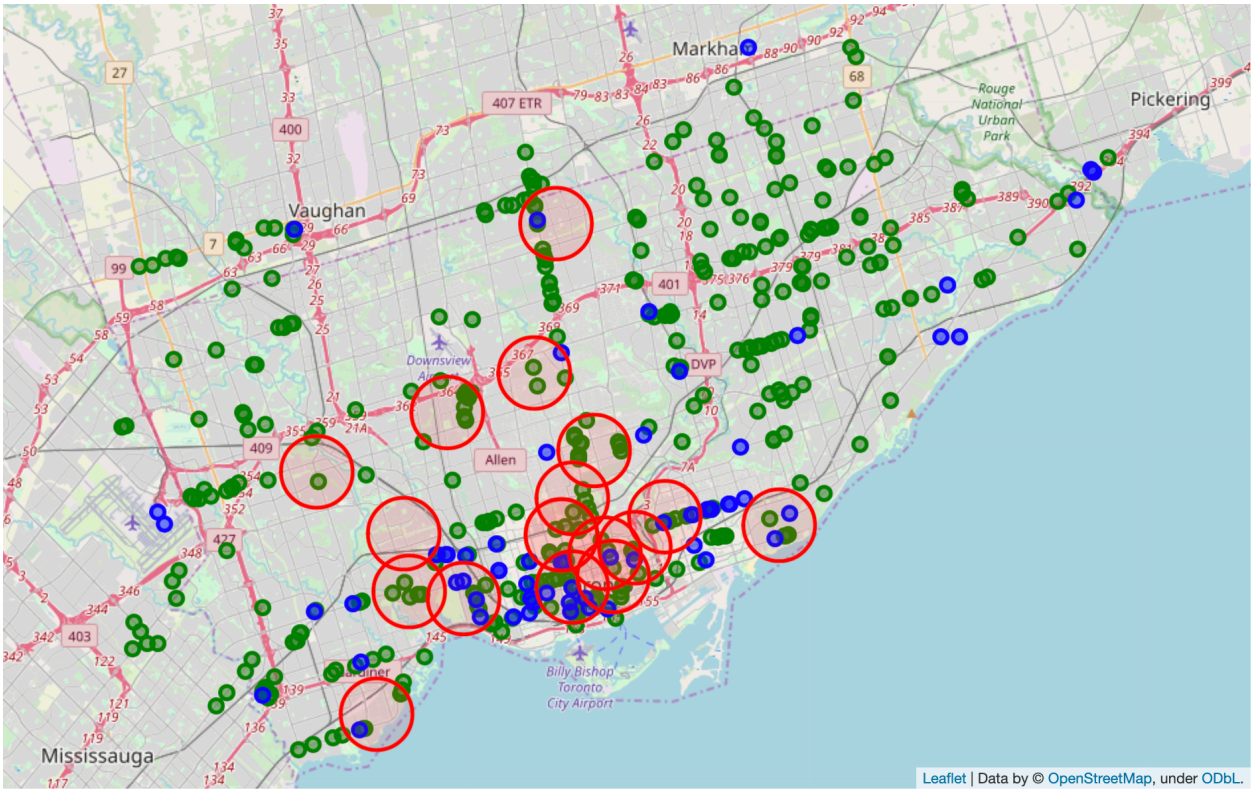
Second, many NaN values appeared for Postal Code, which is unfortunate because the Postal Code was used to link the two dataframes. These rows were dropped as there was no way around this problem.

Third, the neighbourhoods are not built in circles but Foursquare would retrieve information for venues based on the distance of the venue to the centre of the neighbourhood.

3. Methodology

3.1 Mapping Venues and Neighbourhoods

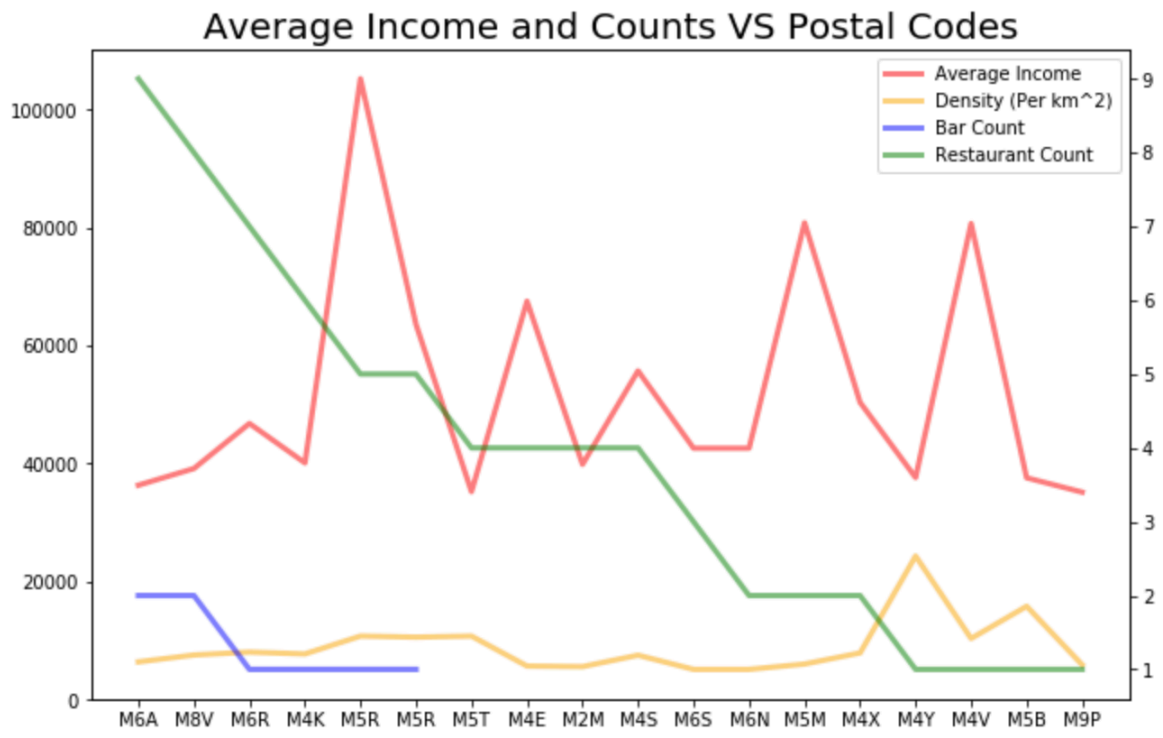
Using the gathered data, the first step was to use folium to visually map the neighbourhoods that fit the following requirements: average income above $35,000 and population density (people/km^2) over 5,000. The red represents the different neighbourhoods that fit the requirements. We can see that there are a limited number of locations available. The next step is to compare the number of bars and restaurants within these neighbourhoods to find a location with some potential.



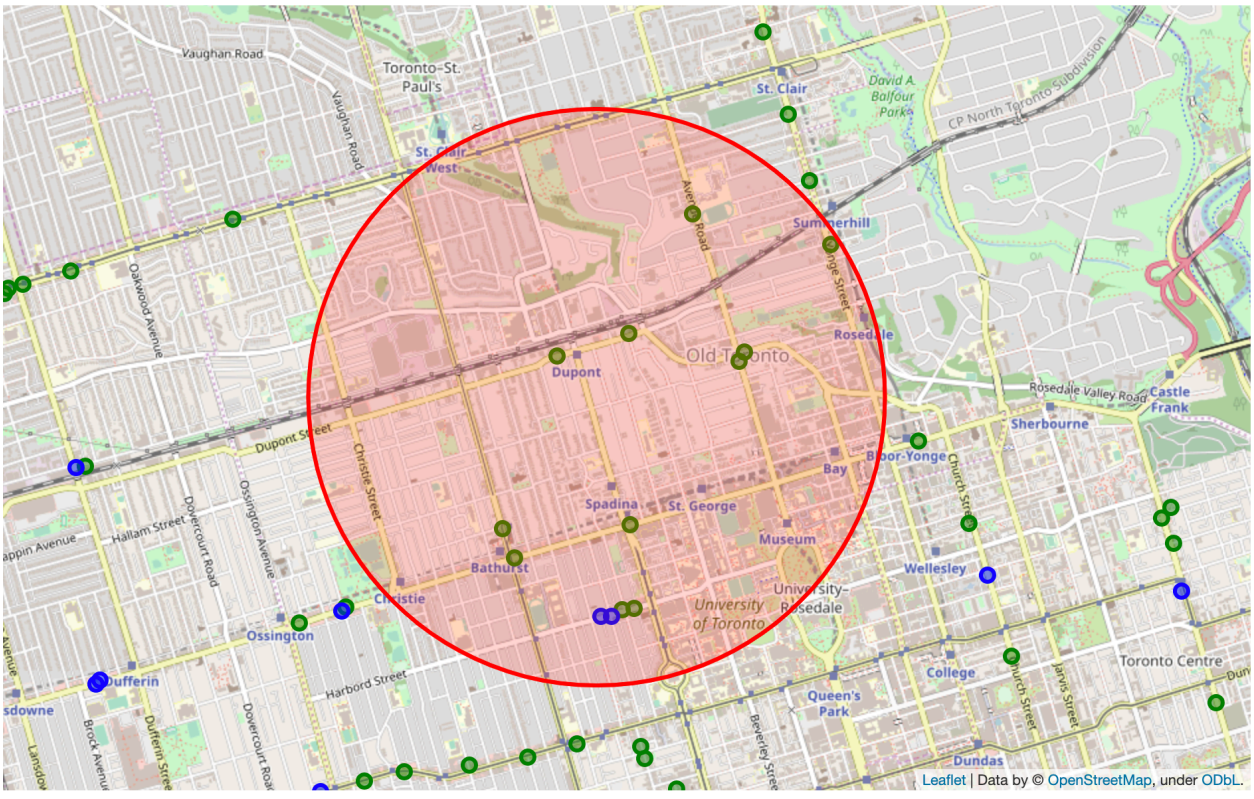
4. Analysis

4.1 Graphing

Using the filtered data, we found out the different number of bars and restaurants within a certain neighbourhood by using specific Postal Codes that linked the two data sets. Here is a graph displaying the number of bars and restaurants. The Average Income and Density values are shown on the left y-axis and the Bar and Restaurant values are shown on the right y-axis.



From the line graph, you can see that the Postal Code M5R has the highest average income, an average density count, an average amount of restaurants and a small number of bars located in the area. To verify this finding, let's take a look at the map.



5. Results and Discussion

The results show that there are a number of neighbourhoods in Toronto suitable for the potential opening of a bar/club. However, the one neighbourhood that did stand out above the rest was The Annex. It had one of the highest average incomes, high a high population density, a high number of restaurants, and a low number of bars. The Annex, by far, is the best choice when analyzing this data with those metrics.

As enjoyable as this was, there were a lot of mistakes and factors overlooked while going through this project. First, what makes a suitable location for a restaurant? It was automatically assumed that those four factors are correlated with the success of a restaurant. Among other things, if this project were to be done again, much more effort would be put into research and diving deeper into figuring out what the actual problem is rather than just assuming.

Secondly, a lot of the Postal Code that Foursquare provided was incomplete. Therefore, as we continued to narrow down results and filter them, data was omitted due to the incompleteness. Once again, the data is the most important, so if done again, I would spend a lot more time gathering data using similar API's such as google maps, etc.

Thirdly, the classification of bars and restaurants were based off keywords that I simply thought of on the spot. Had I known that lack of classification by Foursquare, I would have taken a different route to classify bars and restaurants.

Lastly, the analysis, due to my lack of Python knowledge was quite simple and did not show any advanced analytical techniques. If I were to repeat this project, I would want to add a polynomial regression that shows the correlation between multiple factors.

6. Conclusion

The project was to find a suitable location for a bar in Toronto by looking at four metrics in order to aid potential investors. By scraping data off the web and also using Foursquare, the data was scrubbed, merged, mapped, and then finally analyzed. Add the end, a suitable location that met the criteria for all 4 metrics was found in The Annex neighbourhood in downtown Toronto.

There are many more optics to be looked at when deciding on the bar location. Other factors to take into consideration may include but not limited to: demographics of neighbourhoods, noise level, crime activity, accessibility via public transportation, rent price.

Overall, this project was quite interesting and it definitely helped me to expand my python knowledge as well as problem solving skills. However, the most important aspect for me was that it was enjoyable. I hope to increase my data science skills further and also pursue a career in this field.