Toronto Neighbourhood Analysis

Stan Richards

Coursera Capstone Project

June 2021

Contents

1 Introduction

1.1 Business Understanding

For this project the business problem in focus is ideal property locations in Toronto, Canada. The results of the project will be targeted at stakeholders who want to open a new location for their coffee chain.

The chain targets an alternative market to typical coffee shops, preferring to operate in locations that are not typically identified as suitable locations for other chains that focus on transport hubs or business areas.

The key indicator that for a suitable location would be lack of coffee shops in the most common venues, other features like parks or other public spaces is a bonus.

The results we hope to achieve are clear indications of neighbourhoods that are in high population areas with very few coffee shops in the location and some of the bonus features.

2 Data

2.1 Data Requirements

As we are focusing on footfall and competition we will look to collect data on the following:

- How many coffee shops are there in the neighbourhood
- How close to high population areas is the neighbourhood
- Are there public spaces likely to generate footfall (parks, public places etc)

The following data sources will be needed to extract/generate the required information:

centers of neighbourhoods latlongs have come in a preloaded csv due to the geocoder module failing to collect them

- Top 10 most common venues
- Corresponding population of areas
- Are there public spaces likely to generate footfall (parks, public places etc)

With this data collected for each neighbourhood we will produce a list of neighbourhoods that do not feature coffee shops in their top 10 locations and do feature public spaces.

3 Methodology

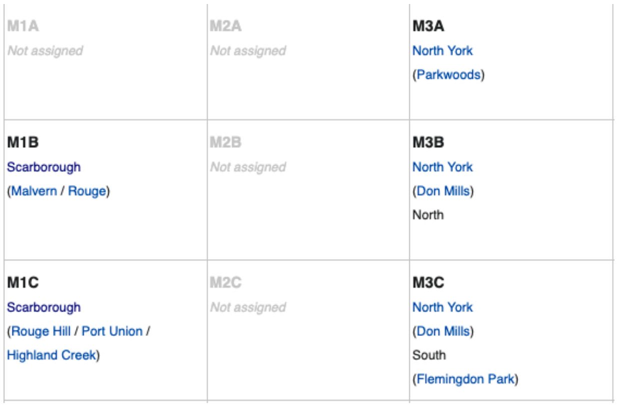
3.1 Data Collection

For this analysis we divide the city of Toronto into neighbourhoods as defined by source [1]

This source was parsed using the BeautifulSoup package and translated into a pandas dataframe for further analysis.

Structure of the produced dataframe:

Postal Code	Borough	Neighbourhood
M3A	North York	Parkwoods



M1A *Not assigned*	**M2A** *Not assigned*	**M3A** North York (Parkwoods)
M1B Scarborough (Malvern / Rouge)	**M2B** *Not assigned*	**M3B** North York (Don Mills) North
M1C Scarborough (Rouge Hill / Port Union / Highland Creek)	**M2C** *Not assigned*	**M3C** North York (Don Mills) South (Flemingdon Park)

Figure 1: Neighbourhoods list in [1]

As can be seen in the dataframe, parsing this link provides a list of neighbourhoods with their corresponding boroughs and postal codes. These postal codes will be used to obtain latitude and longitude values, which is the required input to the Foursquare API to return location data.

To obtain the lat/long information the postal codes of the neighbourhoods are passed into the Geocoder package. (In the code example on the github profile a backup csv is included as the geocoder module often fails to return the correct information)

This extends our dataframe with the addition of the columns:

Latitude	Longitude
43.753259	-79.329656

This latitude and longitude data allows use of the foursquare API to fetch known locations around the latlong coordinates. These known locations will then be prepared to indicate what the most common venues in an area are.
In addition to the locations, we fetched the population information for the neighbourhoods from the 2016 census [2] to compare with our neighbourhood locations. This will be used to generate a chlorepleth map to overlay with our filtered neighbourhoods.

3.2 Data Preparation

The data was collected and prepared in [4]. After initially visualising the locations of all of the neighbourhoods in Figure 2

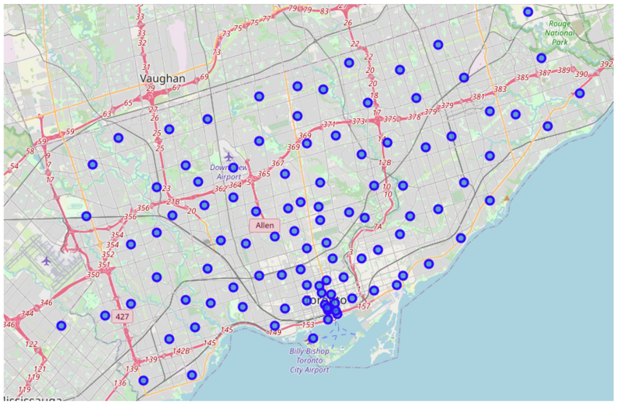


Figure 2: All of the locations in our dataframe visualised in folium.

Latitude and Longitude information was passed into the foursquare to fetch all venues associated with a neighbourhood. This returned 257 unique venue categories. This data was then transformed to present a dataframe containing the neighbourhoods with their corresponding top 10 most common venues to give an idea of the landscape of an area. This format was used to filter the locations in our modelling stage. The census information in [2] contained a significant amount of data, for the purpose of this visualisation it was parsed to collect simply the boundary information of the neighbourhoods and the corresponding populations. This will be added to our folium visualisations in the modelling section.

3.3 Modelling

For our modelling we initially analysed the previous experiment of k-fold analysis in [5]. One of the clusters analysed here showed that one cluster did not feature any coffee shops and featured a lot of public spaces like parks and skate parks. This seemed good starting point for our analysis. The dataframe was then filtered to drop all neighbourhoods containing coffee shops in its top 10 venues and subsequently drop any of the remaining neighbourhoods that did not feature at least one example of a public space.

The population data from the census was then converted to a chlorepleth map to allow imaging of the remaining neighbourhoods with population data.

4 Results

The results of the above analyis are shown in Figure 3.

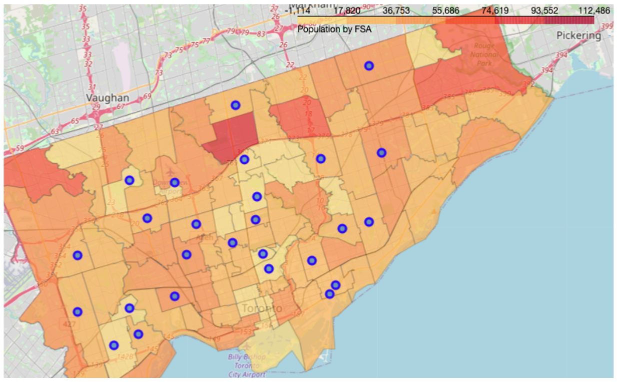


Figure 3: Filtered locations from our analsys overlayed with chlorepleth of population data.

There are multiple suitable locations distributed around roughly uniformly around the city. This warrants further discussion of some particular standout loctions.

5 Discussion

In the resulting filtering some particular locations where highlighted to show ideal locations from the perspective of population and corresponding venues. The population density is not as high as some of the other locations and the neighbouring venues do not fit the brief as well as some other examples. These locations are reasonable candidates but not the primary recommendation.

In Figure 4 the locations corresponding to the highest population areas are shown.

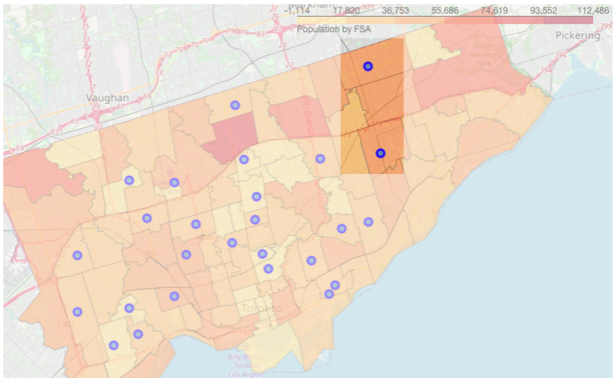


Figure 4: Filtered locations corresponding to the highest population errors

In Figure 5 the two locations with the neighbouring venues that most closely fit the business requirements are shown. Both locations are close to eachother and feature multiple examples of public spaces like parks, art exhibitions and skate parks in their most common venues. These locations are therefore recommended for the new chain opening as they also exist in a high population density area.

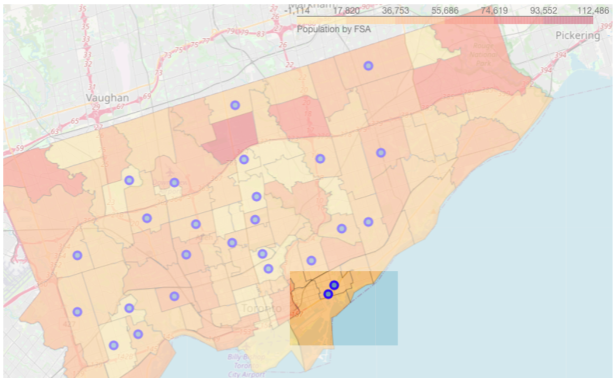


Figure 5: Filtered locations corresponding to the best neighbouring venues

6 Conclusion

This investigation utilised multiple data source to identify which locations from a list of neighbourhoods are suitable for the business requirements of the chain in questions. Some specific recommendations were highlighted based on specifics of the business requirements.

References

[1] https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M.

[2] https://www.toronto.ca/city-government/data-research-maps/neighbourhoods-communities/neighbourhood-profiles/

[3] https://github.com/ag2816/Visualizations/blob/master/OntarioF

[4] https://github.com/stan-git-phys/Coursera_Capstone/blob/main/TorontoAnalysisFinal.ipynb

[5] https://github.com/stan-git-phys/Coursera_Capstone/blob/main/TorontoScraping.ipynb