Detecting Potential Business Improvement Areas in Toronto with Machine Learning

Diego Caballero Loza

July 2020

1 Introduction

1.1 Background

Across the world, urban centres utilize Business Improvement Areas (BIA) to improve, promote, and revitalize their local neighbourhoods and businesses. Organization and management of BIAs vary from country to country, but City of Toronto defines a BIA as an "association of commercial property owners and tenants within a defined area who work in partnership with the City to create thriving, competitive, and safe business areas that attract shoppers, diners, tourists, and new businesses" [1]. Local businesses that participate in the BIA program have an increased capacity to improve the quality of life in their local neighbourhood and the city as a whole.

Currently there are 80 BIAs across the city, which is the largest number of BIAs of any urban centre in the world. In conjunction, they generate more than $34 million in funding towards civic improvement. From 2001 to 2018 the city has increased its number of BIAs by 38, which marks the success of the BIA program and its continued growth.

Toronto's BIA program has a proven track record of revitalizing and enhancing local neighbourhoods and businesses since 1970 [1]. As part of a BIA, businesses benefit from an enhanced business climate as a result of a more attractive and marketable image of their area. Therefore, it is advantageous for business owners to have a tool that can help them make decisions concerning their involvement in a BIA.

1.2 Business Problem

The City of Toronto has multiple data sets with information regarding the pedestrian volume in a neighbourhood, and the geographical boundaries of neighbourhoods and BIAs. This, in addition with venue data provided by Foursquare Places, can be utilized to define areas of the city where groups of venues exist in close proximity of each other.

The goal of this project is to leverage geographical and venue data to aid commercial property owners and tenants with the process of defining new BIAs or extending existing ones based on the location and type of businesses in the city.

2 Data

For this project, three data sets from the City of Toronto's Open Data Portal were used to leverage the Foursquare Places venue data. The first data set contains information about pedestrian and vehicle volume across intersections throughout the city. The second data set contains geographical data for neighbourhood profiles in the city. Lastly, the third data set contains the geographical data for the Business Improvement Areas in the city.

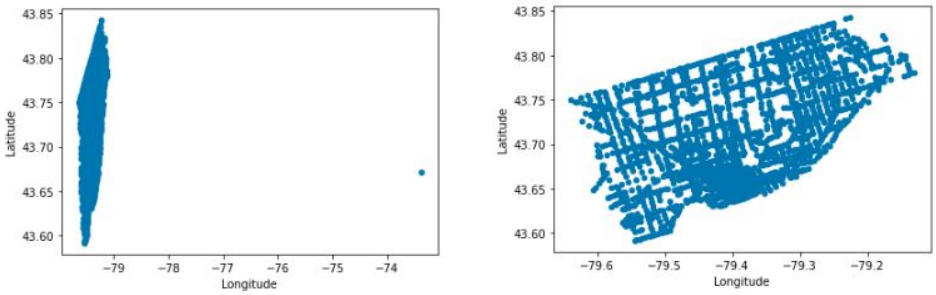
2.1 Pedestrian Volume in the City

2.1.1 Source

This data set contains the most recent eight-hour peak pedestrian volume counts collected throughout intersections in the city that have traffic signals. The data was typically collected between the hours of 7:30 a.m. and 6:00 p.m. The data was obtained from the City of Toronto's Open Data Portal.

2.1.2 Cleaning

There are intersections that have zero pedestrian volume which are likely to be highway ramps or other intersections were pedestrians have no access. These entries were removed since I am only interested in areas of high pedestrian volume. Furthermore, one of the intersections with a high pedestrian volume had a corrupt set of coordinates. Figure 1a shows a scatter plot of the latitude and longitude values of each intersection. This scatter plot should resemble the shape of the city, but you can see the corrupt point on the right of the plot. I fixed this issue by finding the correct location and updating the entry accordingly. Figure 1b shows the scatter plot after correcting the data point.



(a) Corrupt Intersection Location (b) Corrected Intersection Location

Figure 1: Cleaning the Pedestrian Volume Data

2.1.3 Feature Selection

The vehicle volume was dropped given that people driving in the city are not likely to stop at a venue while commuting, especially in the busiest areas of the city. The opposite is more likely for people walking: if they see a store or a cafe, they might stop to visit that venue. The vehicle volume would skew the process of selecting the busiest neighbourhoods in the context of what is the likelihood of someone walking into a business.

2.2 Neighbourhood Boundaries

2.2.1 Source

Neighbourhood boundary data was used in conjunction with pedestrian volume counts to determine the busiest area of the city. I was interested in finding such area since it is the area with the most venues and highest pedestrian volume. The data can be found on the City of Toronto's Open Data Portal and figure 2 displays the boundaries of the neighbourhoods over a map of the city.

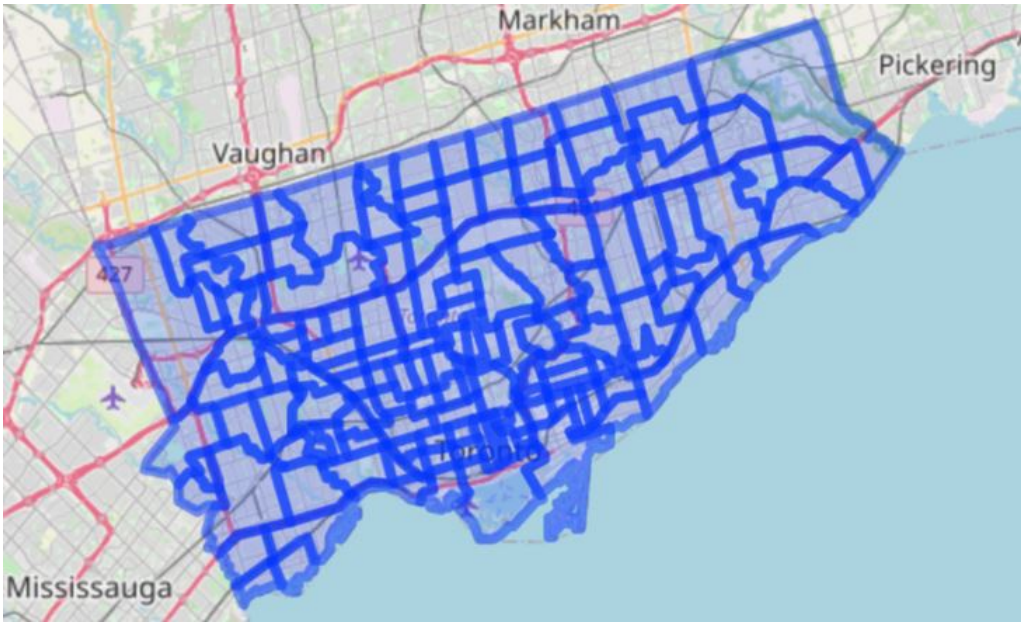


Figure 2: Neighbourhood Boundaries

2.3 BIA Boundaries

2.3.1 Source

The BIA boundary data was used to modify the area given by the busiest neighbourhoods to include the BIAs that exist in the area. Furthermore, it was used to assign venues to their corresponding BIAs which allowed for the calculation of aggregate metrics on each BIA. The data can be found on the City of Toronto's BIA information portal. Figure 3 displays the BIA boundaries over a map of the city.

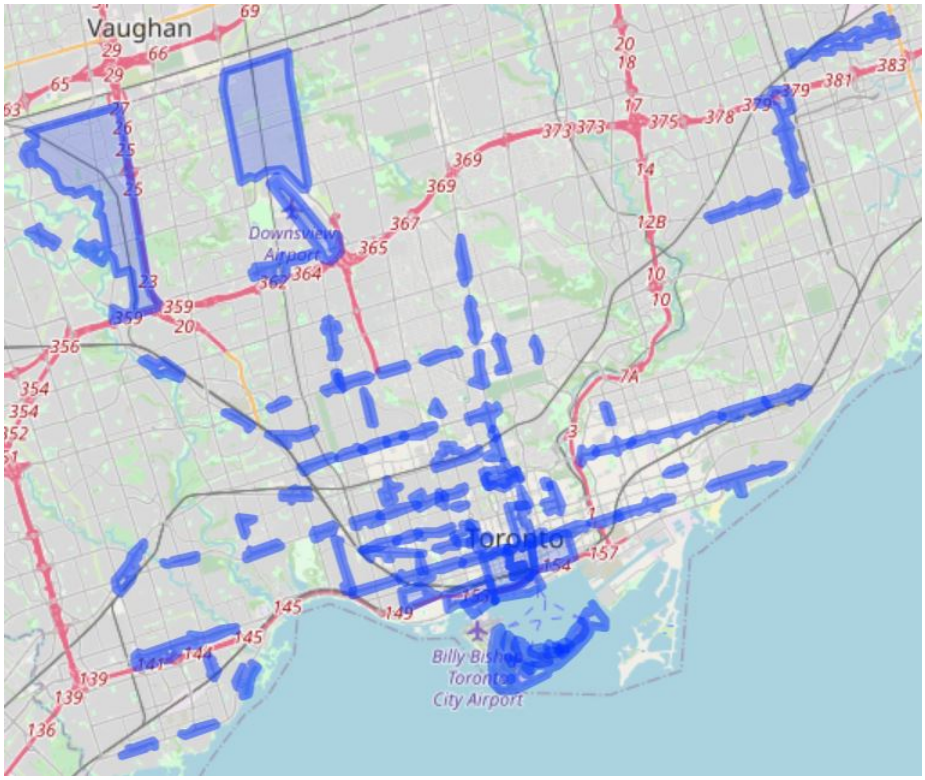


Figure 3: BIA Boundaries

2.4 Foursquare Location Data

2.4.1 Source

The Foursquare Places API allows real-time access to a global database of venue data. The API was used to retrieve venue data across the city. More information about the API and its uses can be found on their docs page.

2.4.2 Cleaning

Through the process data acquisition, the API returned duplicate venues, empty call requests, and error messages. The API calls retrieve data for a specific geographical areas and duplicates arise when these areas overlap. Areas where the API returns no venues (e.g. residential neighbourhoods, parks, the lake, etc.) create entries with missing data. During server maintenance, the API sends a warning message and no venue data is returned. In all the aforementioned cases the corresponding entries were removed. In the case of duplicates only one entry was kept.

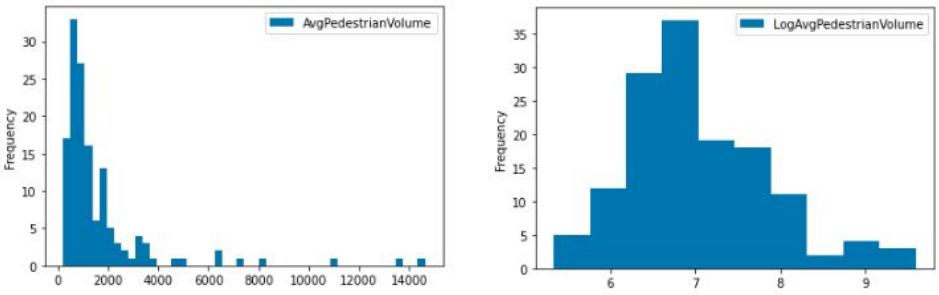
2.4.3 Feature Selection

The venue data contains the venue name, location, and category (e.g. coffee shop or clothing store). The list of categories is very granular which resulted in over 583 unique categories. Venue categories also have a parent category, children categories, or both. Therefore, to reduce category granularity I created an algorithm to assign the parent category to each venue which resulted in less than 150 unique categories. Additionally, I retrieved the uppermost parent category of each venue to group them into six unique categories.

3 Methodology

3.1 Defining the Busiest Area in the City

The pedestrian volume and neighbourhood boundary were used to determine the busiest neighbourhoods in the city. I designed an algorithm that takes a groups of points and polygons, and assigns the points to the polygons they belong to. The algorithm was used to allocate each intersection to their corresponding neighbourhood. This allowed for the calculation of the average pedestrian volume in each neighbourhood.



(a) Average Pedestrian Volume (b) Log Average Pedestrian Volume

Figure 4: Transforming the Average Pedestrian Volume Across Toronto Neighbourhoods

Figure 4a shows the distribution of the average pedestrian volume across all neighbourhoods. The distribution has a long tail, and most of the values are concentrated below 2,000. To better visualize the busiest neighbourhoods on a map (Figure 5), the log of the average pedestrian volume was used instead (Figure 4b).

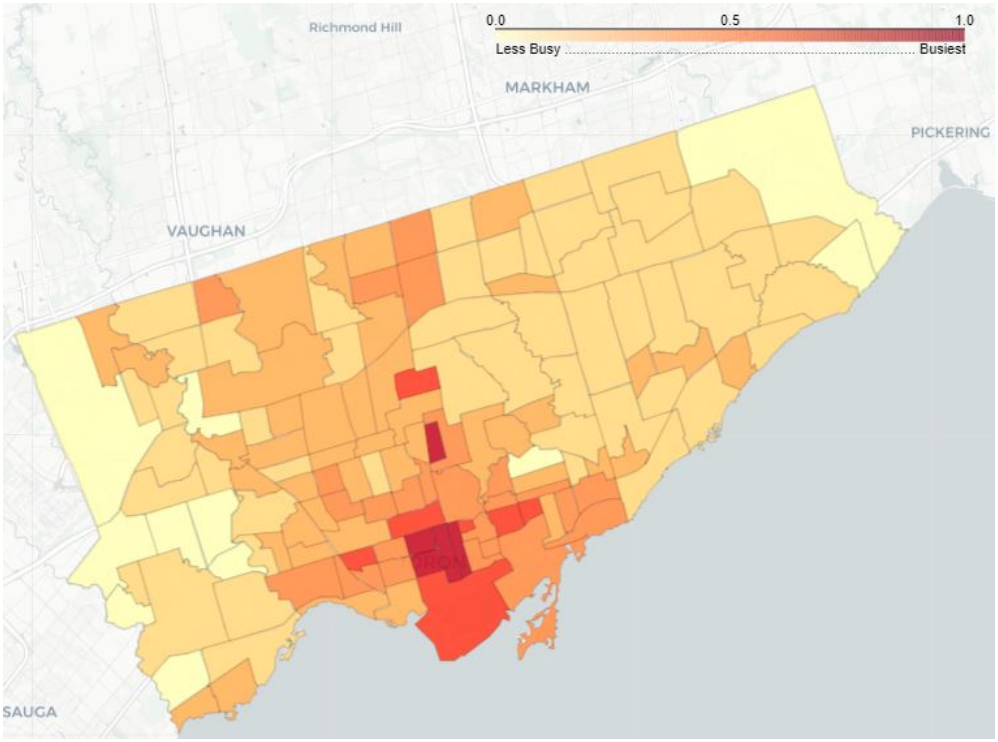


Figure 5: Busiest Toronto Neighbourhoods by Pedestrian Volume

You can see that the busiest neighbourhoods are located near the south of the city (Figure 5). Given that this is the downtown area of the city where the financial district is located, it makes sense that these neighbourhoods have the highest pedestrian volume. Based on this information, I chose the ten busiest neighbourhoods and defined a perimeter around the outer most boundaries. Following this step, I used the BIA geographical data to extend the perimeter to fully contain the BIAs that cross its boundary. Additionally, I decided to remove the southernmost BIA (The Waterfront BIA) since only a small portion crosses the perimeter. Lastly, the perimeter was modified to fully enclose all the BIA boundaries. The resulting perimeter is shown by figure 6.

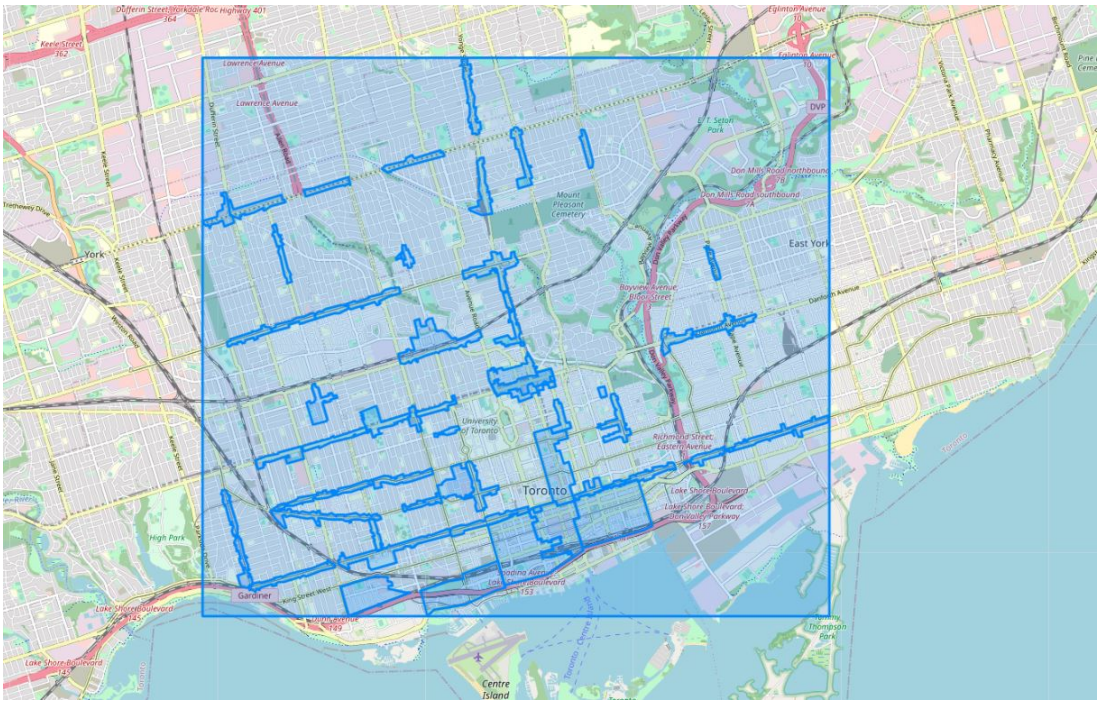


Figure 6: Perimeter with BIAs

3.2 Using *k*-Center Clustering to Retrieve Venue Data

Given a radius r, the k-Center Clustering method finds k points that define the centres of k circles of fixed radius r such that the distances between circles are minimized. I implemented this algorithm from section 4.2 of the book Geometric Approximation Algorithms by Sariel Har-Peled of the University of Illinois.

The algorithm requires iteration through defined points inside the polygon in order to find the k centres of the circles. To meet this requirement, I covered the polygon with random, uniformly distributed points (Figure 7).

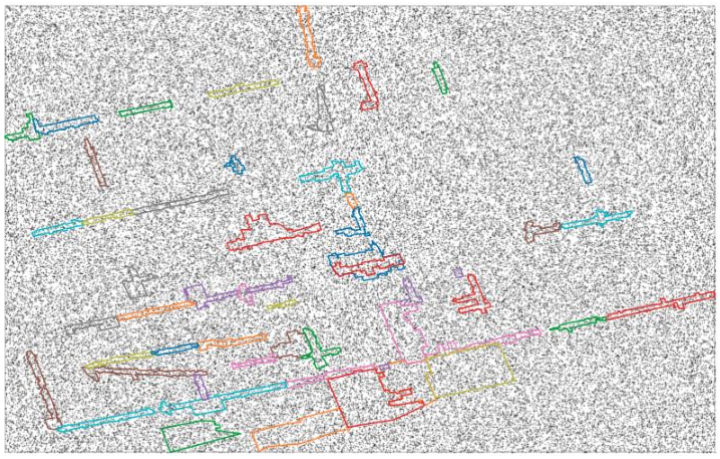


Figure 7: Area of Interest with BIAs and Uniformly Distributed Points

The API limits the number of requests to 5,000 per hour so to reduce the amount the total run time of the solution, the radius was chosen to produce at most 5,000 circles. A radius of 109 metres gave the best results with approximately 4,900 circles that covered the vast majority of the perimeter and returned a satisfactory number of venues. Lastly, venues that aren't captured by one API call might get captured by the API call of an adjacent, overlapping circle. Figure 8 shows the *Trinity-Bellwoods* neighbourhood with its corresponding circles. There are clear gaps between the circles but this is a consequence of the cap of API calls.

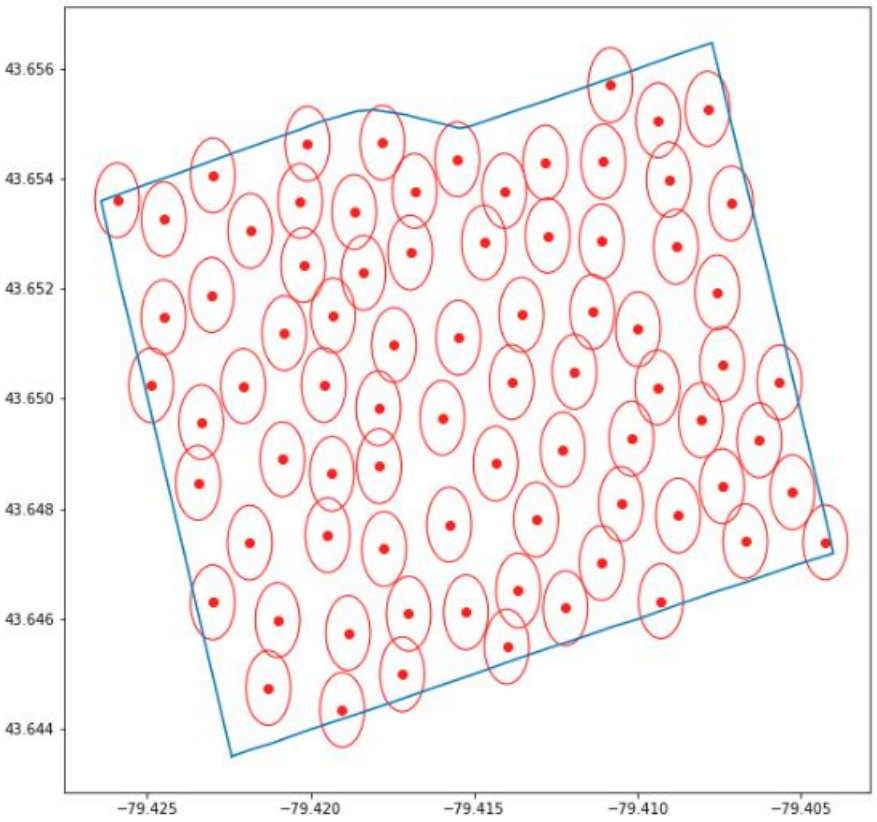


Figure 8: Visualization of the k-Center Clustering Algorithm

3.3 Foursquare Venue Data

With the circles defined, an API request was done for each of them. Each API call takes just under half a second but with almost 5,000 API calls in total, it would have taken about 41 minutes to get all the data. To overcome this challenge, I parallelized the API calls and acquired the data in less than a minute.

After cleaning up and processing the data, I had over 32,000 venues within the perimeter. The next step was to assign each venue to its corresponding BIA (if it is inside a BIA) in order to compare the venue distributions inside and outside BIAs. Figure 9 shows the distributions of venues by category, inside and outside BIAs. From the figure it is clear that there is a difference in their distributions.

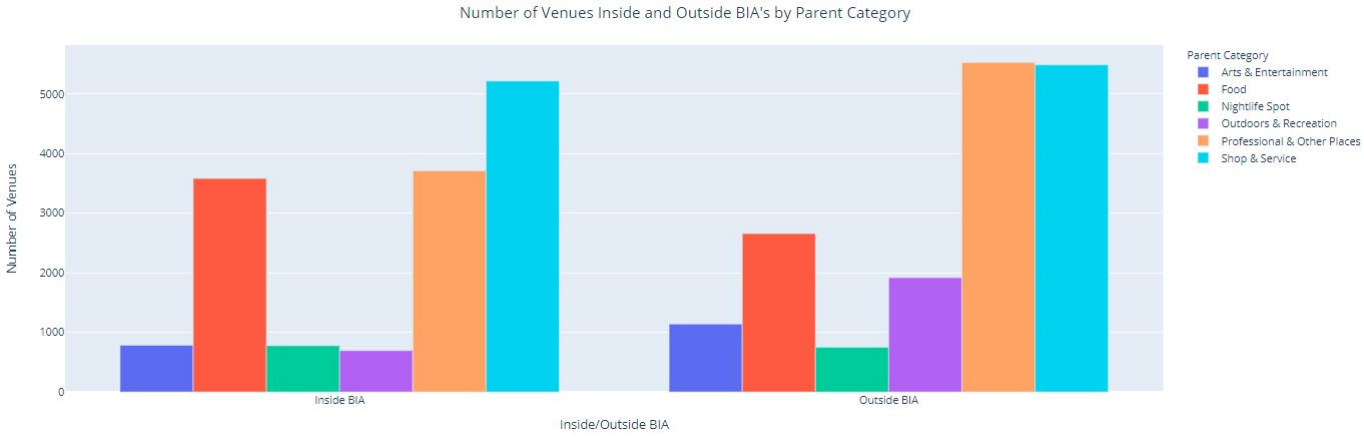


Figure 9: Distribution of Venues by Category Inside and Outside BIAs

The figure above shows some interesting differences in the number of venues inside and outside BIAs. First of all, it is worth noting that there are 15.55% more venues outside of BIAs than inside of them. There is a similar number of 'Shop & Service' venues inside and outside BIAs; but, since there is a higher number of venues outside the BIAs this is an indication that BIAs weight the importance of these types of venues more. The same applies to venues in the 'Nightlife Spot' category. This case is more pronounced the 'Food' category since there are more food venues inside BIAs than outside of them. It is interesting to see that even though BIAs cover a lower percentage of the area of interest, there are more food venues inside them than outside of them.

3.4 Clustering Venues Outside BIAs Using DBSCAN Clustering

To detect whether a group of venues could be part of a BIA, I used the DBSCAN clustering algorithm with the Haversine distance to cluster venues outside BIAs. Creating the clusters allowed for comparison against venues inside existing BIAs.

The DBSCAN algorithm has a hyper-parameter *eps* of which its optimal value was found using a k Nearest Neighbour algorithm. After performing the DBSCAN on the set of venues, I filtered out the clusters that didn't have a significant number of venues in them (Figure 10).

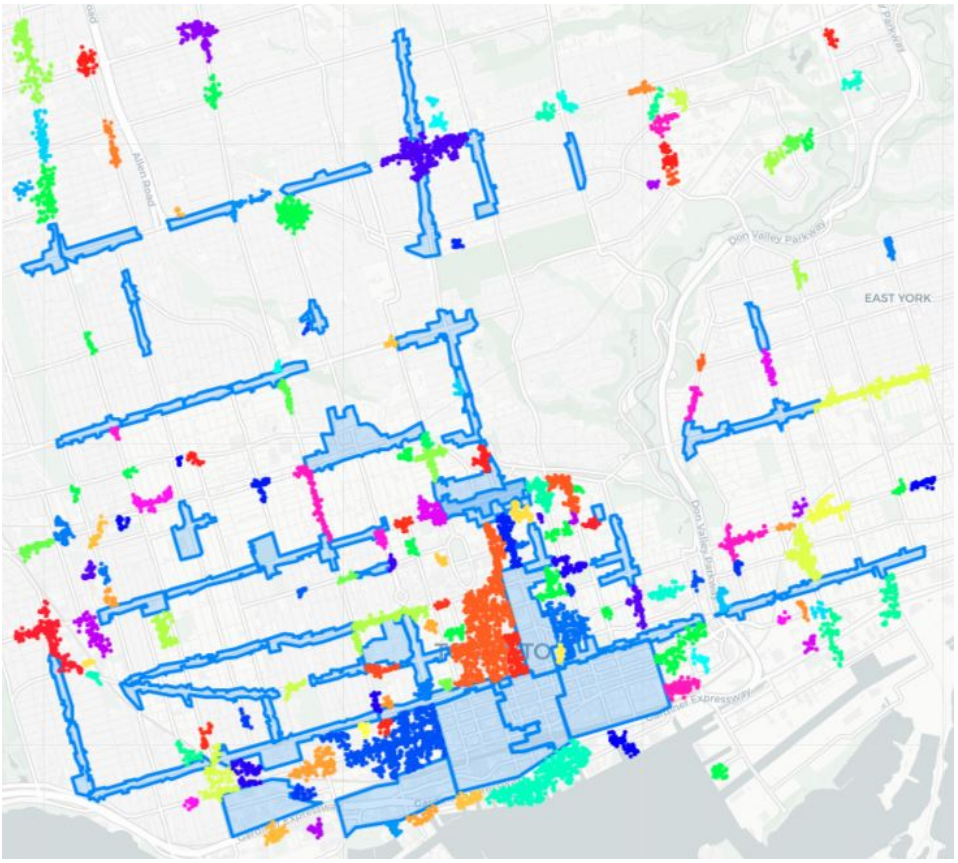


Figure 10: Clusters of Venues Outside BIAs

These are the clusters that will be passed through the trained Autoencoder to determine which ones are most similar to existing BIAs. In the next section, I show the results of using the Autoencoder to achieve this goal.

3.5 Autoencoder to Detect Potential BIAs

The set of existing BIAs was used to train an Autoencoder. The Autoencoder has two hidden layers: one for encoding and one for decoding the data. The activation function used in each node was the ReLu function. Eight features were used as inputs: the log average pedestrian volume, the percentage of venues by category (six different categories), and the total number of venues inside the BIA. The encoder layer compressed these features to six new, encoded features. A MinMax scaler was used to normalize the data, the Adam optimization algorithm was used as the optimizer, and the Mean Standard Loss was used to calculate the reconstruction loss during training.

The algorithm was trained for 120 epochs until the training loss converged (Figure 11).

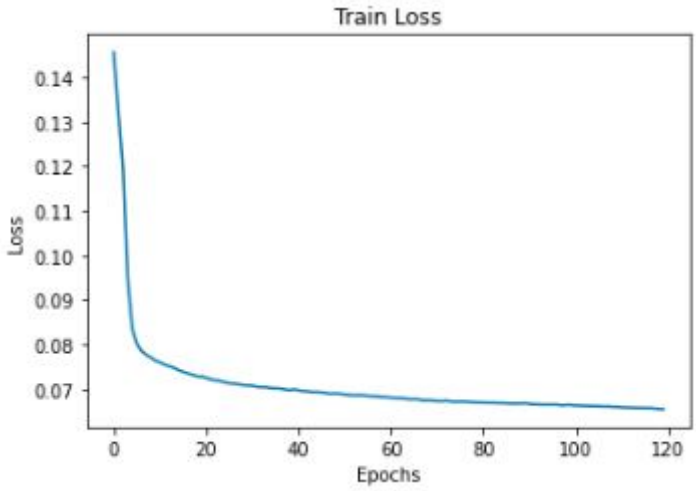


Figure 11: Epochs vs Reconstruction Loss

After training the Autoencoder, the set of clusters found using DBSCAN can be passed through the network and their reconstruction losses calculated. In the next section I show the reconstruction loss distribution for all the clusters, as well as a map containing the ten clusters with the lowest reconstruction loss.

4 Results

A low reconstruction loss signifies that the cluster is similar to the set of existing BIAs, indicating that the set of venues could exist within a BIA. Figure 12 shows the distribution of the reconstruction loss across all clusters. The plot demonstrates that multiple clusters have a low reconstruction loss while some have a higher loss and may not be suitable to be in a BIA. Figure 13 shows the map of city with the existing BIAs as well as the ten clusters with the lowest reconstruction loss.

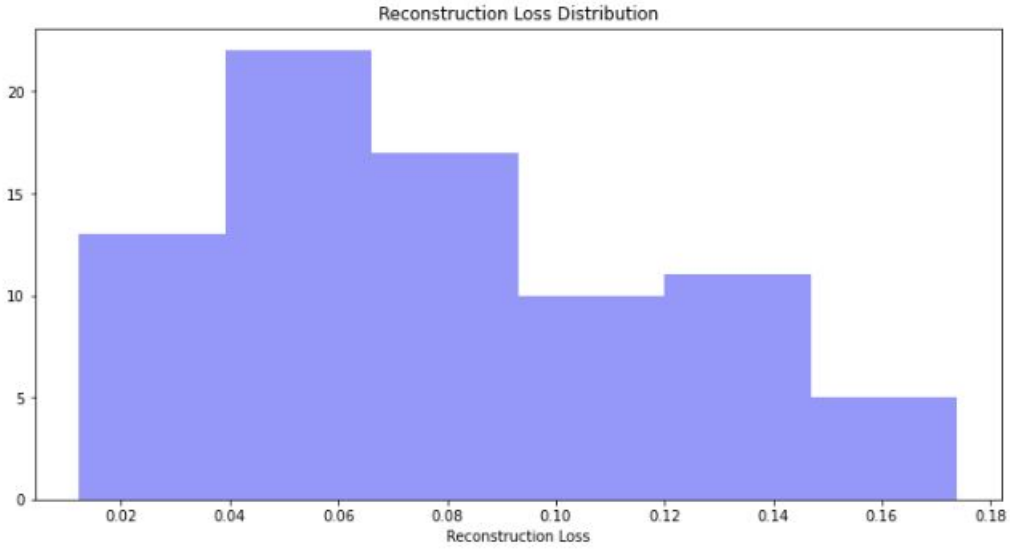


Figure 12: Reconstruction Loss Distribution

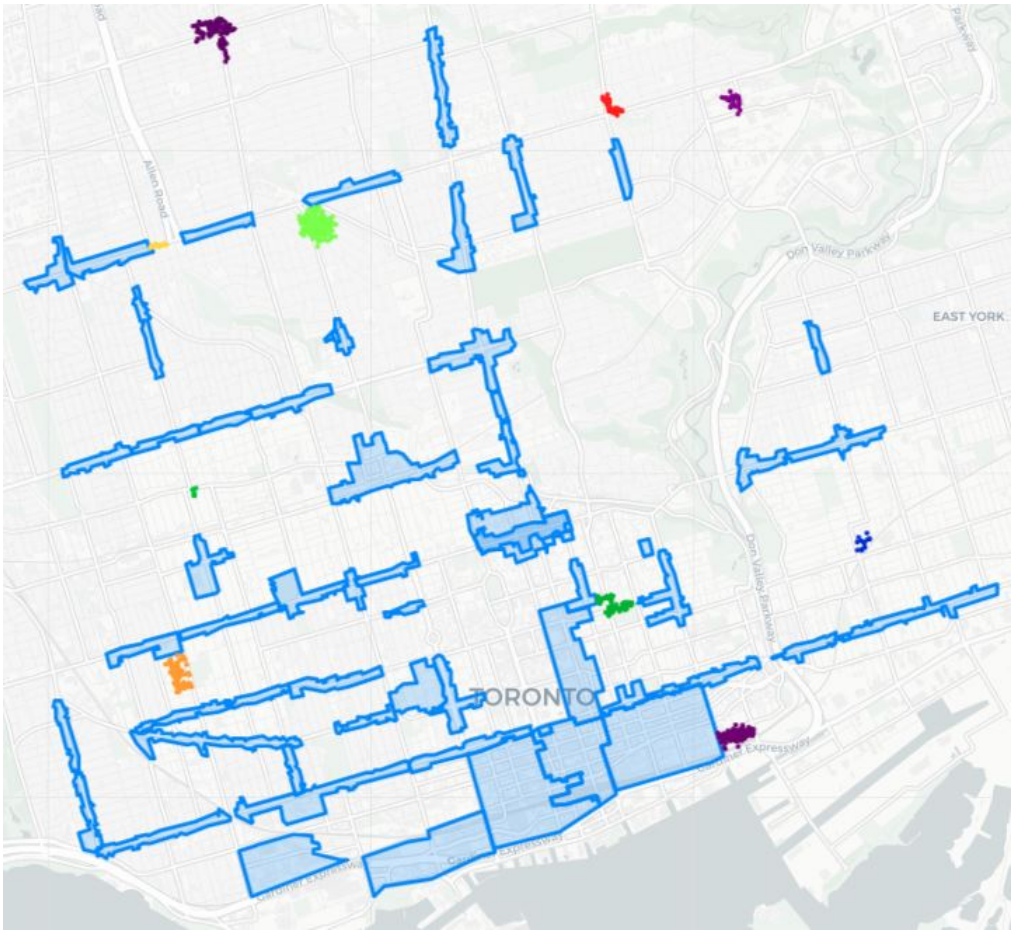


Figure 13: Reconstruction Loss Distribution

From the map, it can be seen that some of the clusters are not surrounded by existing BIAs while some are adjacent to them. The ones that are adjacent to existing BIAs could potentially become additions to them while the standalone ones could be part of their own BIA.

5 Discussion

In this project, I was able to explore the different types of venues within the BIAs of Toronto. A particular piece of information that I found interesting was the fact that even though BIAs cover less area, they contain almost the same number of venues dedicated to shopping and nightlife and more food venues than areas outside BIAs.

Trying to detect venues with the potential of inclusion in a BIA was both ambitious and challenging, and using an Autoencoder to achieve that goal came with some limitations. The first limitation I encountered was the lack of a sufficient number of samples to train the algorithm on. This limitation came from the cap of API calls which limited the number of BIAs I could include reasonably with respect to run time. The second limitation was the absence of negative data (that is, data that I could confidently flag as not a BIA) which transformed my problem into a one-class classification problem. One-class classification isn't as well developed as other classification methods with much ongoing research on this field. I spent multiple hours reading through papers trying to determine the best course of action. With the inclusion of negative data I could turn this into a binary classification problem and use more traditional and accessible methods.

6 Conclusion

The Business Improvement Area program in the City of Toronto has proven to be successful in improving neighbourhoods in the city and helping new and existing businesses thrive. In this study I analyzed the BIAs in the busiest area of the city and used geographical and venue data to determine if other neighbourhoods in the city could host new BIAs. I used multiple clustering algorithms to retrieve data using the Foursquare Places API and to cluster such data into groups of venues. Finally, I implemented an Autoencoder to see whether groups of venues could be included within a BIA and detected the ten areas with the highest potential of inclusion in a new or existing BIA. I believe this solution can help business owners in their decisions of extending existing BIAs or creating new ones to further improve the city and its neighbourhoods.

7 Future Improvements

There is still a lot of work to do in order to improve the performance of the Autoencoder. For example, it would be useful to extend the solution to include all the venues and BIAs in the city. Furthermore, creating negative data (i.e. areas in the city that are guaranteed to not be a BIA) from residential areas, suburbs, parks, etc. would be useful to increase the number of training samples. Creating negative data would also allow me to transform the problem from a one-class classification problem to a binary classification problem. Then I could use more common and better developed model performance metrics. I hope to come back and improve this solution.