**Final Capstone Project – City Clusters of Toronto Metropolitan Area**

# **Introduction**

I am an employee of a foreign development and investment firm that is seeking different opportunities in Toronto area. There is no limit to what the firm invests as long as it's profitable opportunity. We are looking to enter the Toronto market because it is a big city and there's a lot of traffic.

However, with us being a foreign company and having little local knowledge, we need to do some exploratory analysis of the city before we decide on our entry strategy and projects. I am tasked to do some research to have a high-level understanding of the metropolitan area and present the findings to my supervisor and investors.

# **Business Problem**

In this analysis, I am trying to get a high-level understanding of the city. Since we are hoping to invest in various types of opportunities, we will be looking at different venues in different neighborhoods. There are several neighborhoods in Toronto area. It will be less efficient to examine each neighborhood for a high-level analysis like this. Therefore, I am planning to cluster these neighborhoods into clusters based on their venue category counts. After determining the optimal number of clusters and successfully clustering neighborhoods, I can check within each cluster to see what the majority of the venues are in each cluster. The result of this observation can give me a high-level understanding of each cluster, namely, I can say if cluster/district 1 is a food district and district 2 is a financial district, etc.

To summarize, my business problem is to determine the optimal number of clusters of neighborhoods in Toronto, and then to hopefully examine each cluster to determine what kind of district it is to inform our investment strategy.

# **Data**

## Data Sources

In order to solve my business problem, I'll need geological information for all neighborhoods in Toronto, especially latitudes and longitudes. I'll also need the categories of the venues to determine the type of district they are located in. For example, if a cluster has a lot of venues categorized as restaurant or coffee shop, we can say this cluster is more likely a food district.

Fortunately, all the data can be obtained from the internet with a quick parse and download. The venue data can be obtained from the Foursquare data center. I don't need additional data sources. The data collection process is shown below. Data regarding Toronto neighborhoods can be obtained from the link <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M> and data regarding coordinates of each neighborhood can be obtained using the link <http://cocl.us/Geospatial_data>. Venue data can be obtained from Foursquare database with its API.

## Data Cleaning

Data needs to be parsed from the web using pandas’ beautiful soup library. After parsing, the first version of the dataset has “n\” at the end of each entry due to the website coding. Operations were conducted to remove this redundancy. It is also noticed that some boroughs have a value of “Not assigned”. These boroughs were discarded. Coordinates data was obtained directly from a csv file on the internet. Two tables were combined to create one data frame of postal codes, boroughs, neighborhoods, latitudes, and longitudes, as shown in the figure below.

Graphical user interface, text, application

Description automatically generated

Figure 1. Geological data of Toronto Neighborhoods

Data regarding venues near each neighborhood is obtained directly from Foursquare. A column of neighborhood is attached to the data frame as shown below.

Graphical user interface, text, application

Description automatically generated

Figure 2. Venue data

Data was further processed and grouped by neighborhood, summing up the number of same venues, to reflect the venue situation of each neighborhood. This data table will be used as the basis of the k-mean cluster analysis to slice Toronto Metropolitan area.

## Feature Selection

No feature selection was conducted in this analysis as the goal is to have a comprehensive understanding of the city. Each venue category contributes to the understanding of the city as we are a foreign firm without local knowledge. There is also no way for us to know the weighted importance of each category, so we are unable to estimate the effect of excluding certain categories from the dataset.

# **Exploratory Data Analysis**

## Visualize the Neighborhoods

I started by mapping out the neighborhoods of Toronto on a map as shown below.

Map

Description automatically generated

Figure 3. Toronto Neighborhoods

Then a dummy function was called to convert categorical data about venues into binary data, which then can help us calculate the number of venue categories in each neighborhood. Part of the processed data is shown below.

Table

Description automatically generated

Figure 4. Processed venue category data.

## Determining the Optimal Number of K for K-mean Clustering

I’m trying to conduct a K-mean cluster analysis. This analysis clusters data points into groups based on their features. The K, or number of clusters, is predetermined before the analysis and can be any value. Therefore, evaluating K is one of the most important tasks in K-mean clustering. One of the most common methods used for evaluation is referred to as elbow analysis, where a series of Ks is used for analysis. Then the square error is recorded. As the magnitude of K increases, the error will decrease, but to a certain level, as error won’t reduce to 0. The K where the sharpest drop in decreasing rate occurs is the optimal K. Using this algorithm and test for K between 1 and 15, an elbow plot is created below.

Chart, line chart

Description automatically generated

Figure 5. Elbow plot

Using a calculation algorithm, the elbow point can be calculated as when K is equal to 3. Therefore, 3 will be used as the number of clusters.

# **Data Analysis**

## K-Mean Clustering

K-mean clustering was then conducted with K value of 3. All neighborhoods are analyzed and clustered together based on their venue category counts. A map plot is shown below to represent the clusters.

Map

Description automatically generated

Figure 6. Cluster result

As can be seen from the figure, the shapes of the clusters are quite irregular. The main reason for it is that the clustering did not take geographical distance into consideration. If coordinates were to be used, they were not normalized with venue category counts and they would be weighted significantly more, rendering the other features irrelevant.

After clustering, neighborhoods are divided into three sets. The first cluster has 16 neighborhoods, including neighborhoods such as Berczy Park, Brockton, Church and Wellesley. The second cluster includes 9 neighborhoods, including Central Bay Street, Commerce Court, Victoria Hotel, and St. James Town. The last cluster includes 71 neighborhoods, including Agincourt, Christie, and Davisville North.

## Venue Categories Distribution

Let’s first look at the overall distribution of venues in all three clusters. A line graph is shown below.

Chart

Description automatically generated

Figure 7. Venue distribution

It can be seen from the figure that the overall trend of venue distributions in all 3 clusters are very similar. It is reasonable because Toronto is a relatively developed city with comprehensive infrastructure and facilities throughout. It is reasonable that all neighborhoods have all kinds of venues including American restaurants, coffee shops, etc. If we look at the top 10 venues in each cluster, there are significant discrepancies as shown below.

Graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

Figure 8. Top 10 venues in each cluster

Coffee shop consists of the most venues in each cluster, which makes sense because people love to drink coffee. The majority of the venues in all clusters is related to some sort of restaurants, may it be pizza, sushi, Italian, or Japanese. It can be noticed that cluster 2 and 3 have more variety in the top 10 than cluster 1, with venues such as gym, clothing store, park, and pharmacy.

If we include the top 20 venue categories, we will see more variety and differences as shown below.

Graphical user interface, application

Description automatically generatedA picture containing graphical user interface, application

Description automatically generatedGraphical user interface, application

Description automatically generated

Figure 9. Top 20 venues

For all 3 clusters, besides the most common venue being coffee shop and café for cluster 1 and 2, no ranking is the same for each cluster. However, restaurants still reign supreme as the most majority of venues.

Although the overall trends are very similar, there are some discrepancies spotted in the graph. After examining it with the data after clustering, several observations can be made. Firstly, cluster 3 has fewer venues overall. At the same time, cluster 3 has more airport-related venues, pizza places and parks than the other 2 clusters. Secondly, cluster 2 has significantly more hotels than the other clusters, as well as culture-related venues including art galleries, art museums, aquariums, among others. It also has the most sports stadiums including baseball stadium and basketball stadiums.

# **Discussion of Result**

From the data analyzed above, several remarks about each cluster can be made. I’ll start with cluster 3.

Cluster 3 has relatively the fewest venues in all three clusters. It has the most amount parks and pharmacy, as well as grocery stores and banks. This distribution and observation lead me to believe that cluster 3 represents a more rural and outside-of-downtown district. This is a district that is not busy with work and may be less populated. Cluster 3 may be more suitable for family living with all the parks for children and grocery stores for everyday living. A healthy amount of restaurants and liquor places can spice up the area. There are also airport-related venues in cluster 3, which suggests that some of the neighborhoods are near the airport. It will be reasonable to classify cluster 3 as rural living district.

On the opposite side of the spectrum, cluster 2 represents the downtown area well. It has 95 coffee shops, which is the most amongst the three. It has 36 hotels which suggests that it is a tourism center. The more culture related venues such as museums and galleries also support the statement that a large amount of tourists visit this cluster. The number of stadiums also suggest that cluster 2 is the business and downtown center of the city as most cities have their stadiums located near downtown. It also has more department stores, shopping centers, and shopping malls, which also suggest downtown and being more populated. It would be safe to classify cluster 2 as the downtown/business center district.

In the middle lies cluster 1, which acts the transition area between downtown and rural area. It likely attracts young working population who often reside in apartment rentals and enjoy the livid life of downtown. It has all types of restaurants and drinking places. It is less family-life than cluster 3 with fewer parks and living-related venues, but more modern facilities like yoga studios and pubs. On the other hand, it has more living vibe to it than cluster 1, with more restaurants and drinking places. Living in cluster 1 will be easier and cheaper than living in cluster 2. It is safe to classify cluster 1 as the urban living district.

# **Conclusion**

In this analysis, I’m trying to get a high-level understanding of the neighborhoods in Toronto metropolitan area. To achieve this, I obtained neighborhood and venue data from the internet and different data sources. After processing the data and performing the k-mean cluster analysis, I successfully clustered the neighborhoods into three districts: urban living, downtown/business center, and rural living. I’ve also obtained a list of neighborhoods within each cluster.

These results will help greatly with my business problem. With these findings, we, as a foreign company, have a starting point to consider our strategy in several ways. Firstly, we have a basic understanding of our potential customer base. Cluster 1 provides customers that have not settled and younger, who may prefer new and innovative offerings more. It will also be great to invest in multi-family apartments. People in cluster 2 are more business oriented and may prefer functional offerings. People in cluster 3 will appreciate safety and offerings helping their life, child education, among others. Secondly, it informs us what types of venues are already prevalent. For example, it will be very competitive if we decide to open another coffee shop in cluster 2, as there are already 95 coffee shops there. However, surprisingly, there are not many banks in district 2, which may be an opportunity for us to invest or partner. Another example will be that although cluster 1 is urban living, there are not many department stores to satisfy the needs of young adults, which can be another opportunity.

In conclusion, in this fairly simple data analysis, we revealed a high-level profile of Toronto and get a basic understanding of its city clusters, which in turn help us make business decisions more wisely.

# **Future Directions**

There are several directions that future analysis can go. Firstly, it is possible to first aggregate some of the venue categories to have a more generalized understanding. Sometimes, it is overfitting to differentiate between Japanese restaurant and American restaurant. A larger category of restaurant may be more suitable for high-level analysis. Secondly, it may be better to somehow include geographical distances between neighborhoods to create clusters of neighborhoods that not only have similar venues, but physically closer to one another. Thirdly, it is possible to drill in to one specific category to create sub-clusters within larger clusters. For example, it is possible to get information on Japanese-related venues, which may lead to the cluster of a Japan town.