|  |
| --- |
| The Battle of Neighborhoods |
| Using Python Machine Learning to understand the city structure of Toronto and predict the venue development of suburban Toronto area |

|  |
| --- |
| Zixing Wang  5-16-2020  This report serves as a part of Coursera IBM Data Science Certification Capstone Project. |

The Battle of Neighborhoods

Using Python Machine Learning to understand the city structure of Toronto and predict the future venue development of suburban Toronto area

# Table of Contents

Contents

[Table of Contents 1](#_Toc40626484)

[Abstract 1](#_Toc40626485)

[Introduction 2](#_Toc40626486)

[Methodology 2](#_Toc40626487)

[Results 3](#_Toc40626488)

[Section1: Data Cleaning and Understanding 3](#_Toc40626489)

[Section 2: Categorize venues in Toronto and cluster neighborhoods using density-based clustering 5](#_Toc40626490)

[Section 3: Predict the development of postal code M9R by comparing with inner city Toronto area through collaborative filtering 9](#_Toc40626491)

[Discussion and Conclusion 9](#_Toc40626492)

[References 9](#_Toc40626493)

[Acknowledgement 9](#_Toc40626494)

# Abstract

Understanding city structure is crucial for future development of the city and predicting the growth of suburban areas. In this project, the structure of Toronto has been studied with data science. We have found the downtown Toronto venues are mostly composed of restaurants and tea & coffee places. While the further out areas are for activities. The more suburban areas are more diverse in terms of venue categories. Attempts have been made to predict the future development of postal M9R by collaborative filtering, and found to be not successful. Indicating the venue structure goes through dramatic change during development and the city and suburban areas are not similar.

# Introduction

With the development of human society and the commute system, the cities are growing ever faster. As people build more infrastructures and opening more shops, a clear understanding of the city structure is needed. This helps maximizes the use of new venues, reduces resource consumption, and thus improves profit. For example, a bus station should be built in a location where nearby residents’ walking distance is minimal, and transportation time is the shortest between destinations. A coffee shop should be built where customers are most likely to stop by for a cup of coffee without too much walk from their offices or apartments. However, all the above-mentioned strategies cannot be achieved without a clear understanding of the city structure. Data science serves as a handy tool in achieving its purpose. In this study, venue information in the city of Toronto has been collected, categorized, and clustered to understand human activities within the cities. Attempts have been made to predict the future development of the suburban Toronto area, based on similar regions within the city of Toronto with collaborative filtering.

# Methodology

The whole project is divided into three sections. The first section focuses on data collection, cleaning, and understanding. The second section future categorizes all venues into ten major categories, and then cluster them into three groups by category frequency. The third section attempts to predict a suburban postal code M9R’s future development by comparing it with Toronto's inner-city areas.

First, the Toronto area is sectioned based on postal codes. All postal codes in Toronto will be scraped from the Wikipedia page [1] using BeautifulSoup. The data will be cleaned by eliminating postal codes without assigned neighborhoods. Then latitude and longitude of each postal code are obtained through geocoder and added to the Toronto data frame. We further narrow down the region from the city of Toronto to Boroughs only containing the word “Toronto”. Then the top 100 venues, with their latitude, longitude, and venue category, within 500 m radius of each postal code is collected using Foursquare API. Data understand was attempted by clustering the postal codes by venue category frequencies into 5 clusters using k-means. However, there are too many venue categories to help gain a clear understanding of the neighborhoods.

In section 2, all venue categories are future divided into ten general categories, namely: restaurants, tea& coffee, hotel, gym & sports, amenities, shopping, art & culture, bars, grocery, and others. All venue categories containing the word “Restaurant” will be automatically be put under category “restaurants”. Then the rest of the top 100 most popular venue categories are divided into the above-mentioned ten categories. Then the means of each of the ten categories are calculated for each postal code, and the postal codes are clustered by density-based clustering based on the mean of each category.

In section 3, venue categories of postal code M9R have been extracted using Foursquare API and averaged. Then it is compared to postal codes with “Toronto” in their Borough names. Toronto venue data frame will be rearranged to 4 columns: postal code, venue categories, mean and category ID, respectively. Venue categories with zero means will be deleted from the data frame to save space, leaving a total of 843 category venues from each postal code. Then the new Toronto venues will be grouped by postal codes. Category ID will be assigned to each category. Attempts will be made to find the postal codes most similar to M9R and new venues will be predicted for the M9R region.

# Results

## Section1: Data Cleaning and Understanding

180 postal codes are extracted from the Wikipedia website. 103 of the postal codes have assigned Borough. 39 postal codes contain “Toronto” in its Borough name. A total of 1581 venues are extracted within 500 m for each of 39 postal codes. Though the target is 100 venues for each postal code, a number less than 3900 indicates many postal codes have much less than 100 venues in its 500 m radius.

The final master dataset looks like this:

A screenshot of a cell phone

Description automatically generated

Using one hot encoding, groupby function, and calculate the mean, I obtained the average number of each venue category for each postal code respectively.

A screenshot of a cell phone

Description automatically generated

Among the 1581 venues, there are 220 types of unique venue categories. “Coffee shop” and “café” are the most popular venues totaling 222 stores in our list. The popularity is followed by restaurants and hotels. However, the venue categories returned by Foursquare API is too specific, which hinders our ability in judging the venues correctly. For example, coffee shop and café are similar and should be put under the same category. There are entries “Restaurant”, “Asian Restaurant” and “Japanese Restaurant”, which each should be considered a general category of the later, and should not be compared in parallel. This issue will be solved in in section 2, where the venue categories are further categorized.

A screenshot of a cell phone

Description automatically generated

When clustering the neighborhoods based on the specific venue categories using k-means. The k-means clustering results in a distribution of different types of neighborhoods in Figure 1.

A close up of a map

Description automatically generated

Figure 1: Distribution of 5 clusters generated by k-means clustering based on 220 types of venue categories.

Ten most common venues for each clusters are:

Cluster 1:

A screenshot of a computer

Description automatically generated

Cluster 2:

A screenshot of a cell phone

Description automatically generated

Cluster 3:

A screenshot of a cell phone

Description automatically generated

Cluster 4:

A screenshot of a cell phone

Description automatically generated

Cluster 5:

A screenshot of a cell phone

Description automatically generated

However, no obvious similarity can be observed among neighborhoods within the same cluster.

## Section 2: Categorize venues in Toronto and cluster neighborhoods using density-based clustering

To improve the clustering technique, we will future categorize 220 venue clusters into 10 general categories and cluster using density-based clustering. All venue categories containing word “Restaurant” will be put into “Restaurant” list. Then the rest top 100 venue categories are manually put into different categories. From general categories, we observe the most popular kind of venue are restaurants, different from coffee shop & café from section 1. 539 out of 1581 venues are restaurants, taking more than 1/3 of total venue numbers. Others category takes the second and Tea & Coffee is the third most popular category with 283 venues and bars take the 4th position totaling 101 venues.

A screen shot of a social media post

Description automatically generated

Table: popular categories from the most to the least.

Density based clustering with epsilon=0.25 and minimum samples to be 2 returns 3 clusters, excluding the outliers. The top 5 most common categories of each cluster are shown below:

Cluster 1:

A screenshot of a cell phone

Description automatically generated

The most popular venues fall into other category, meaning the venues in these neighborhood are more diverse. Restaurant, shopping, tea & Coffee also take a heavy share in their composition.

Cluster 2:

A picture containing food

Description automatically generated

The top three most common venues in cluster 2 are restaurants, tea & coffee and others. Meaning these areas are popular for either busy people having no time to cook and drinks coffee often.

Cluster 3:

A screenshot of a cell phone

Description automatically generated

The second most popular category are still restaurants. However, gym & sports and amenities also take a heavy share of the venue composition, meaning this is likely to be places where people carry leisure activities.

The outliers:

A screenshot of a cell phone

Description automatically generated

By plotting each cluster onto the Toronto map, we can see cluster 2 marked in purple are aggregated towards the center of Toronto city. Cluster 1 marked in red are closer to the costal area. Judging by its venues category, they are likely to be airports that provides a variety of services. Cluster 3 are closer to the inner coast. The outliers lie in the most inner coast regions.

A close up of a map

Description automatically generated

A screenshot of a cell phone

Description automatically generated

## Section 3: Predict the development of postal code M9R by comparing with inner city Toronto area through collaborative filtering

We extract 8 venues from postal code M9R within 500 m radius, meaning region M9R is not populated. After comparing Pearson correlation between M9R and other postal codes, it is found that the correlation between them are very weak. The highest Pearson correlation coefficient found is 1.09E-15, very close to zero.

A screenshot of a cell phone

Description automatically generated

# Discussion

By comparing the clustering results from 220 venue categories and 10 general categories, a clearer trend can be observed with 10 general categories. Further putting 220 category venues into 10 makes it easier to understand the Toronto city structure and should be preferred.

It is found that the most common venues in the center of Toronto are restaurants, then coffee & tea. This makes sense since we expect inner city to be business heavy. With more working people, more restaurants are built to provide food for busy working class. Tea & coffee places provides a good leisure spot, or provide them coffee before a busy day. Next to the city center, gyms and sports venues are more common. This also makes sense since people are likely to stop by a gym on their way to or back from work for workouts. The further inner coast regions are more diverse in terms of venue categories.

# However, little similarity can be found between a further inland region like postal code M9R, to the Toronto city. The little similarity between the two indicates the venue composition changes significantly during city development. The need for people is different from the city and the suburban areas. Conclusion

From this project we can conclude that less category numbers can give a better overview of the neighborhood structure in Toronto. Using k-means clustering on ten categories, we see the downtown is more popular for food and drinks, while the outer skirt are more diverse. Downtown is closer to the coast. People like to do activities between home and work. However, the venue structure of suburban area is hard to predict using collaborative filtering. Collaborative filtering is not the best algorithm for prediction of area development, first because the data points are too little from postal code M9R. Second, the venue categories used for collaborative filtering are too specific, which makes it difficult to find similar postal codes.

# References

1. <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>, postal codes, Borough, and Neighborhood in Toronto, Canada.

# Acknowledgement

I acknowledge Coursera for providing necessary materials to complete this project.