Identify Neighborhoods that are suitable for raising families in Toronto

1.Introduction

1.1 Project Background:

The goal of the doing this analysis is to identify Neighborhoods in Toronto, Canada that are suitable for raising families with children.

1.2 Interest

The audience of this business problem will be residential real estate developers as they would like to develop some residential housings where small families would like to live in.

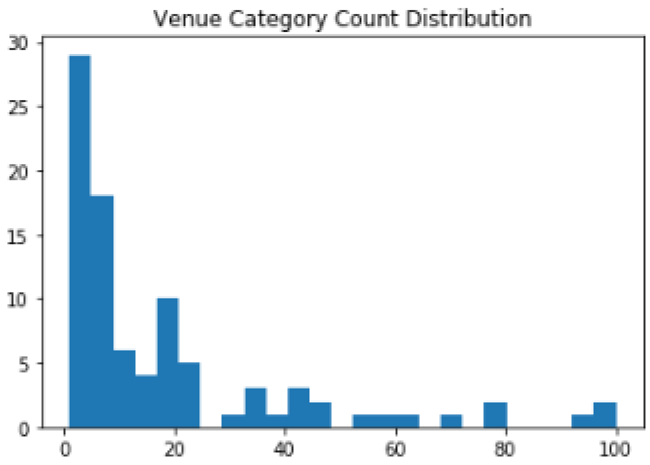
2.Data Acquisition and Cleaning:

- Toronto city data that contains Boroughs, Neighborhoods along with their Latitude and Longitude.
  - Toronto Neighborhood Data Source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
  - Latitude and Longitude Data Source: http://cocl.us/Geospatial_data
  - We will combine merge the above data source to get the exact Geo location for each neighborhood in Toronto, and will explore characteristics of each neighborhood using the combined data.

- Elementary & Secondary Schools, Restaurants, Parks in each neighborhood
  - Data Source: Foursquare API
  - We will use API information together with Toronto Neighborhood's location to identify which neighborhood has the best schools, restaurants, and parks combination.

3.Explore Neighborhoods in Toronto

We see that many postal codes have only a few venue categories (the leftmost bar on the diagram is much higher than the others). These postal codes have too little data to make a meaningful analysis, therefore we will exclude them from the dataset.



4.Analyze each neighborhood

In this part, we will do one hot encoding to pivot category values into columns of the data frame.

There is one observation that we have to be careful about: one of the category values is Neighborhood. After one hot encoding, this value will become a column name. We are already using the column Neighborhood to represent the neighborhood name. To avoid confusing these columns, we will rename the column that comes from one hot encoding as Neighborhood Category.

5. Cluster Neighborhoods

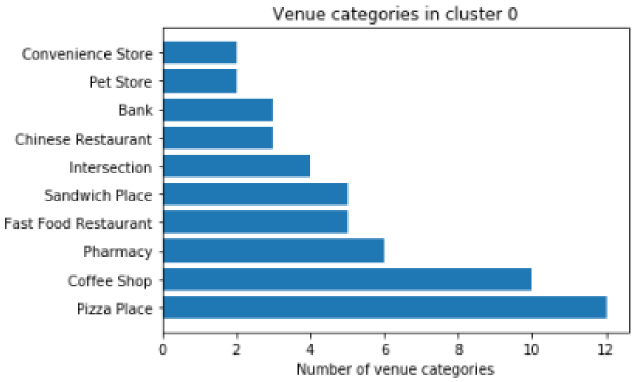
In this part, we Run k-means to cluster the neighborhood into 5 clusters. And Create a new data frame that includes the cluster as well as the top 10 venues for each neighborhood.

6. Examine Clusters

In this section, we will examine each cluster and determine the venue categories that distinguish each other. Based on the defining categories, we can then assign a name to each cluster.

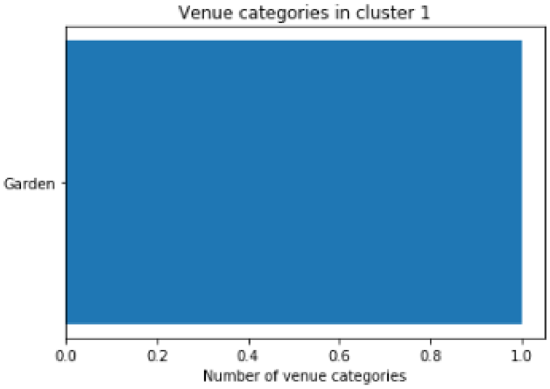
6.1 Cluster 0: Residential Area/Downtown

Cluster 0 conclusion: Venue categories in this cluster appear to be predominantly restaurants and Coffee Shop, suggesting places that are found in residential areas and Downtown.



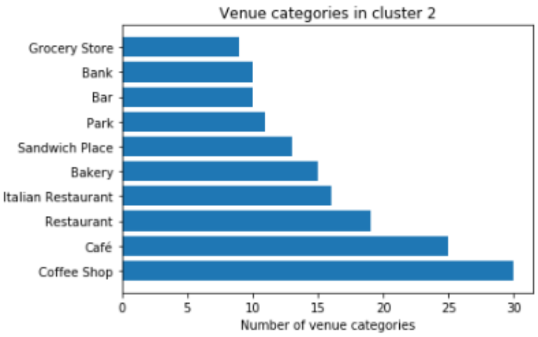
6.2 Cluster 1: Garden

Cluster 1 conclusion: Venue categories in this cluster appear to be predominantly Garden.



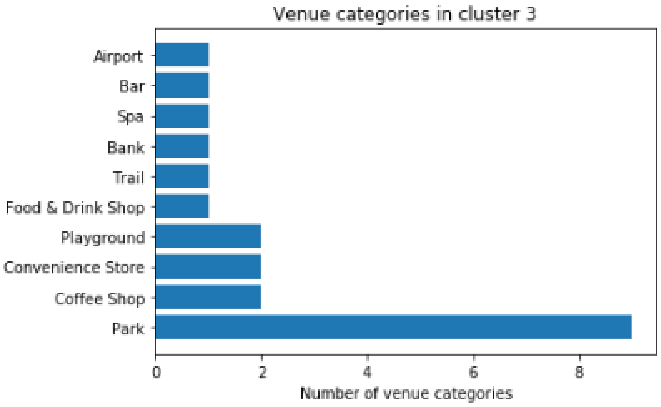
6.3 Cluster 2: Residential Areas/Downtown

Cluster 2 conclusion: Venue categories in this cluster appear to be predominantly Cafes and Coffee Shop, suggesting places that are found in residential areas and Downtown.



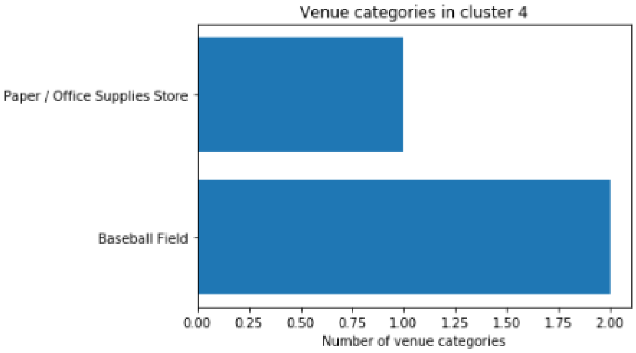
6.4 Cluster 3: Park

Cluster 3 conclusion: This is a cluster of neighborhoods with parks, could be potentially a residential area.



6.5 Cluster 4: Baseball Field

Cluster 4 conclusion: This is a cluster of neighborhoods with Baseball Field and Office Supply Stores.



7.Conclusion

Among the above clusters it would appear that the following clusters are best suited for families with children: Cluster 0 and 2: Venue categories in these 2 clusters are predominantly coffee shops, restaurants, cafes and some shops which is all suitable for families