Coursera Capstone Project

Introduction

For many investors, Toronto is a paradise. It's the biggest city in Canada. It's also one of largest economy, traffic center in the city. So, opening a shopping center may be a good way to make money. Shopping mall is very popular nowadays. It can attract many customers spending money in it, like taking food in restaurant, or buying clothes. Of course, many investors have come out of this idea and many shopping mall has been built. So the location of shopping center is much more important because a bad location may bring overlap between two shopping malls' influence radius.

Business problem:

This capstone project is aim to choose a proper location for shopping mall in Toronto to maximize stakeholders' benefits. Through using data science methodology and machine learning, we try to solve this problem that where is the best place to create a new shopping center.

Data

We need following data to accomplish the task:

1. Toronto Neighborhoods data
2. Coordinates of Neighborhoods, including longitude and latitude of Neighborhoods
3. Venue data, especially related data to shopping mall.

Source of Data and methods to extract them:

The Wikipedia contains a Toronto neighborhoods data set: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M which contains a list of neighborhoods in Toronto. Through using Python and the beautifulsoup package, I will scrape the table list in the Wikipedia. Then, we can get coordinates of the neighborhoods through Python Geocoder package or directly, we have obtained the longitude and latitude of the neighborhoods in the following link: http://cocl.us/Geospatial_data

After that, we will get the venue data by using API of Foursquare. Foursquare City Guide, commonly known as Foursquare, is a local search-and-discovery mobile app developed by Foursquare labs inc. The app provides personalized recommendations of places to go near a user's current location based on users' previous browsing history and check-in history. This is a project containing different kinds of data science skills, from web scraping techniques, machine learning skills, working with APIs,

basic data cleaning, data wrangling to map visualization.

Methodology

At the beginning of the project, we get the list of Toronto neighborhoods and their coordinate through scraping the data from the Wikipedia page(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). We use python and import the beautifulsoup package to extract the neighborhoods. In the process, we examine the html text of the certain page, then find the location of the table and scrape it by python. We save the list in a dataframe provided by pandas which is also a package of python. Next, latitude and longitude of the neighborhoods are necessary to locate places in Foursquare. We can use Geocoder package as a way, but in this project, the data is provided in the former lessons. We can download it directly from http://cocl.us/Geospatial_data. Then, we merge the dataframe readed from the link by read_csv command with the origin dataframe from the Wikipedia.

In the second step, we will use Foursquare API to get top 100 venues that are within a radius of meters. We can obtain Foursquare ID and secret key by registering a Foursquare Developer Account. Then we can make API calls to Foursquare to get the coordinate of the venue data in Jason format and we can extract the venue name, venue category,

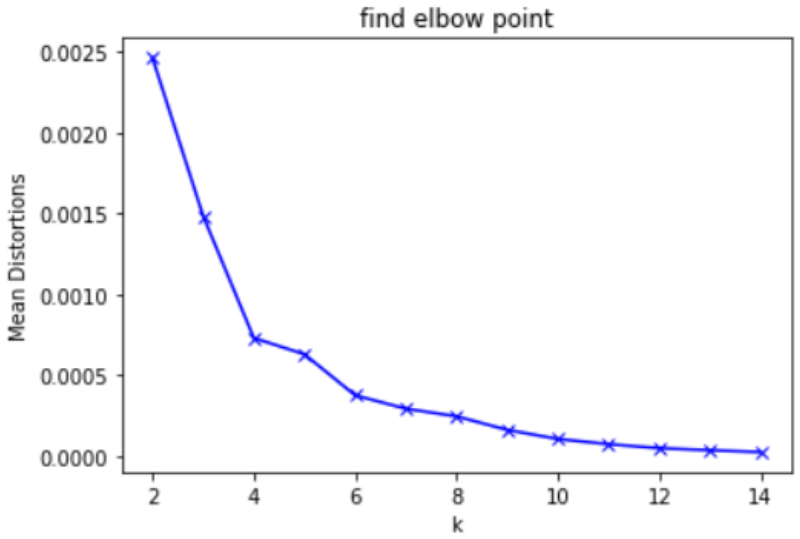
longitude and latitude. Then we can analyze the neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of the occurrence of each venue category. By doing so, we can filter the Shopping mall through the keyword Shopping Mall in venue category.

Finally, we perform clustering on the data by using the K-means clustering algorithm. In the process, we use the Kmeans object from the skit-learn package. K means is a kind of unsupervised machine learning algorithm and is a good way to cluster the unlabeled data which can meet our need this project. In the process of clustering, we use a loop to try different number of clusters, calculating the mean distortion of the final clustering results, and plot them by Matplotlib package. Through observing the picture, we find the elbow point at the number of 4. So we apply the 4 cluster numbers to the final clustering. We visualize the clustering result on the Toronto map which can help us find the best location of the a new shopping mall.

Results

1. Through the elbow rule, choose the K=4 in the clustering algorithm.

```
[31]: Text(0.5, 1.0, 'find elbow point')
```



2. The result of the K-means algorithm cluster the shopping mall into 4 categories:

Cluster 0: Neighborhoods with none existence or low number of shopping mall

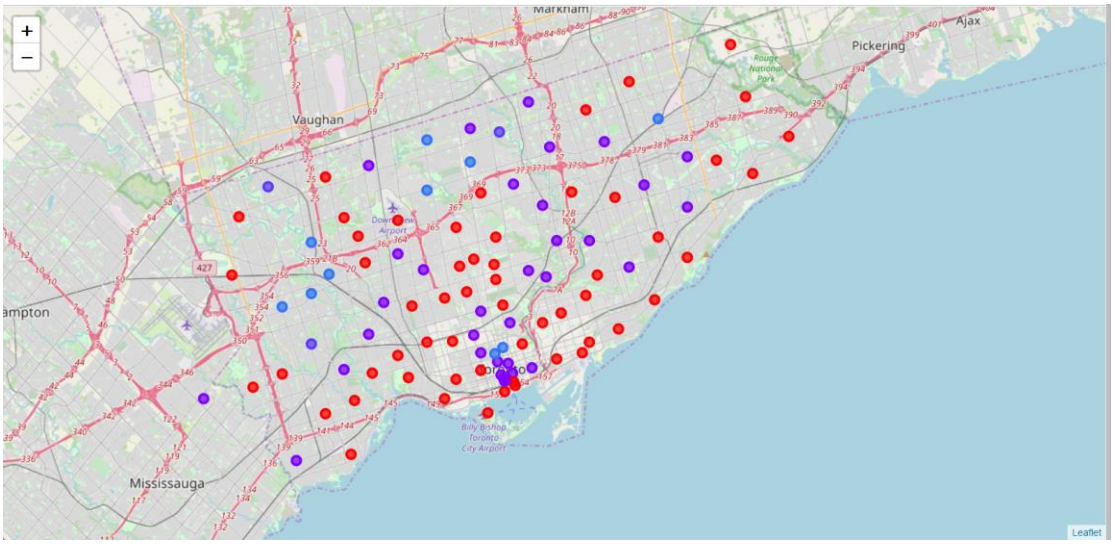
Cluster 1: Neighborhoods with low number of shopping mall

Cluster 2: Neighborhoods with high concentration of shopping mall

Cluster 3: Neighborhoods with moderate number of shopping mall

In the picture, cluster 0: red, cluster 1: purple, cluster 2: blue

cluster 3:wathet



Discussion

As observations noted from the map in the Results section, most of the shopping mall are concentrated in the central area of Toronto. On the other hand, cluster 0 which has very low shopping mall is distributed and surround the Toronto center. In this way, stakeholders have a strong chance to invest the shopping mall in places around Toronto. Some cluster 1 points locate in the Toronto, stakeholders can consider investing in such places since cluster 1 points have low number of shopping malls. The number of cluster 2 is very rare, which is also a good news for investors since cluster 2 have a strong concentration of shopping mall. So, investors should avoid to build a new shopping mall in neighborhoods in cluster 2. Finally, the cluster 3 is also distributed which have a moderate number of shopping mall. Stakeholders can also take them into consideration, since a bit concentration of shopping mall means many people may live there.

Future Research

Although we consider the occurrence frequency of shopping mall in clustering them, it is not enough. Many other factors may also have a strong impact on the choice of location, like the population density. Many cluster 0 points appear in the map, which imply there may exist a low density in certain place. So, in the future research, we should take more factors related to the shopping mall into the model.