The best place to live in Greater Toronto Area (GTA)

Angell -- Toronto, Canada

Introduction

The Greater Toronto Area (GTA) is the most populous metropolitan area in Canada. It includes the City of Toronto and the regional municipalities of Durham, Halton, Peel, and York. In total, the region contains 25 urban, suburban, and rural municipalities.[6] The Greater Toronto Area begins in Burlington in Halton Region, and extends along Lake Ontario past downtown Toronto eastward to Clarington in Durham Region. Durham Region holds the northernmost municipality in the GTA, although a number of municipalities in the northern periphery of Greater Toronto are situated in York Region.

GTA area is like the heart of Ontario in which has higher property price, more job opportunity and better and more facilities. However, living cost and living efficiency vary from areas to areas, so quality of living would also vary from people to people. Choosing the best area based on different people's income, habit and needs would be a rather good choice. Therefore we need to find out the similarity and also differences between and amongst neighborhoods so that different people with different needs and wants can find the best place to live.

Business Problems

The questions I aim to answer in this project are the following:

1. List and visualize all major parts of GTA area with top existing infrastructure.
2. What are the best locations in GTA are as per infrastructure?
3. Which areas have the potential for the development of infrastructure of different kinds?
4. Which all area lake the infrastructure facilities?
5. What is the best places to stay within a city for all vital infrastructure facilities?

Target Audience

The purpose of this project is to help people in exploring better facilities around their neighborhood. It will help people making a smart and efficient decision on selecting great neighborhoods out of numbers of other postal areas in GTA area.

Lots of people are migrating from various cities of Ontario, Canada and needed lots of research for good housing prices, new business, and reputed professional places for their children. This project is for those who are looking for better neighborhoods and businesses.

It will help people to get the awareness of the area and neighborhood before moving to a new city, state, country, or place for their work or to start a new fresh life.

Data Description

My data for this project would only cover the GTA area which including but not limited to Toronto. It is a more geographic concept rather than a political concept.

The data source is mainly comes from Google Website
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

However, as substitute, I will also use Foursquare API to get some information and data to better analyze and demonstrate the project.

Using this data will allow exploration and examination to answer the questions. This is a project that will make use of many data science skills, from web scraping Wikipedia, working with Foursquare API, data cleaning, data wrangling and map visualization and to machine learning (K-means clustering).

Methodology

Data Exploration

Firstly, we need to get the list of neighborhoods in GTA area, Fortunately, the list is available on the web page. We have to do web scraping using Python request to extract the list of neighborhood data. However, this is just a list of pin codes, and cities.

Data Geocoding

We need to get the geographical coordinates in the form of latitude and longitude in order to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert the address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas data frames.

Data Visualization

Visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinate's data returned by Geocoder are correctly plotted in the GTA area.

Data Wrangling

We are also preparing the data for use in selection. Based on the occurrence of infrastructures in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new infrastructures and which neighborhoods are most suitable to visitors to stay.

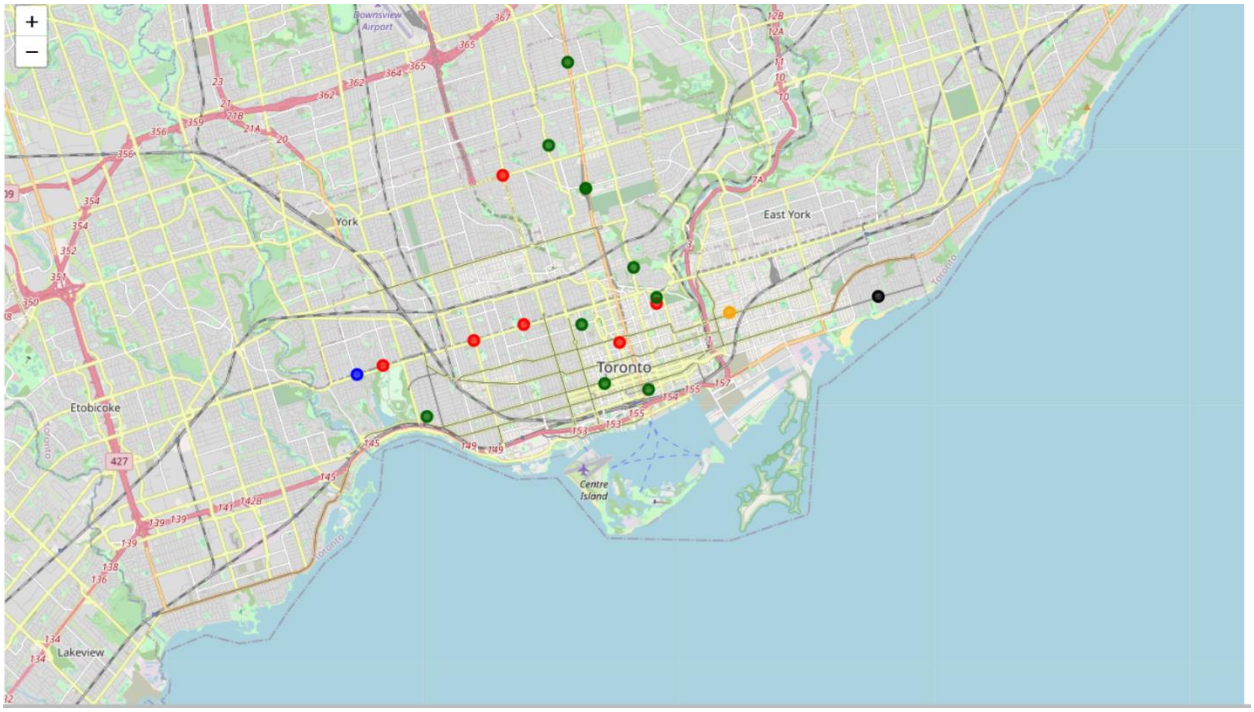
Data Clustering

Finally, we will perform clustering on the data by using K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.it is one of the simples and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project.

We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “no. of existing infrastructures”. The results will allow us to identify which neighborhoods have higher, medium and lower concentration of infrastructures. Based on the occurrence of infrastructures in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new infrastructures.

Results

Clusters



The results from K-means clustering show that we can categorize Toronto neighborhoods into 5 clusters based on food and drinks.

Cluster 0: the most common venue is Beach and the Second is food services.

Cluster 1: There are many food and drinks like restaurant and bar or coffee shop.

Cluster 2: The most common Venue is coffee shop as well as the second most common Venue.

Cluster 3: The most common venue is a restaurant.

Cluster 4: There are so many food and drinks around this neighborhoods.

Discussion

(Based to constraint on API calls and search radius, the true result might vary.)

Most of the infrastructures are concentrated in cluster 4. On the other hand, Cluster 0 and cluster 3 has a very low number of infrastructures in the neighborhoods. This represents a great opportunity and high potential areas to open new infrastructures as it is very little to no competition from existing varied infrastructures.

Conclusion

In this project, I have gone through the process of identifying the business problems, specifying the data required, extracting and preparing the data, visualizing the results, performing machine learing by clustering the data into 3 clusters based on their frequency similarities, tacking and reaching to a definitive solution to business problems (mentioned in results). Lastly, the project is providing recommendation to the relevant stakeholders regarding the best location to move in. The project also provides visitors and immigrants to the city regarding postal office areas for growth and living prosperously.