IBM data Science Capstone Project

Opening a new shopping mall in Toronto, Canada

By: Milind Hingane

Introduction

- The aim of the project is to identify venues in Toronto, Canada based on their latitude and Longitude. In this notebook, we will identify various venues in the city of Toronto, Canada using Foursquare API to help property developers to select the location that suit them the best. So, to decide which location is suitable for the shopping mall we required to identify factors such as location cost, it will be profitable or not and so many.

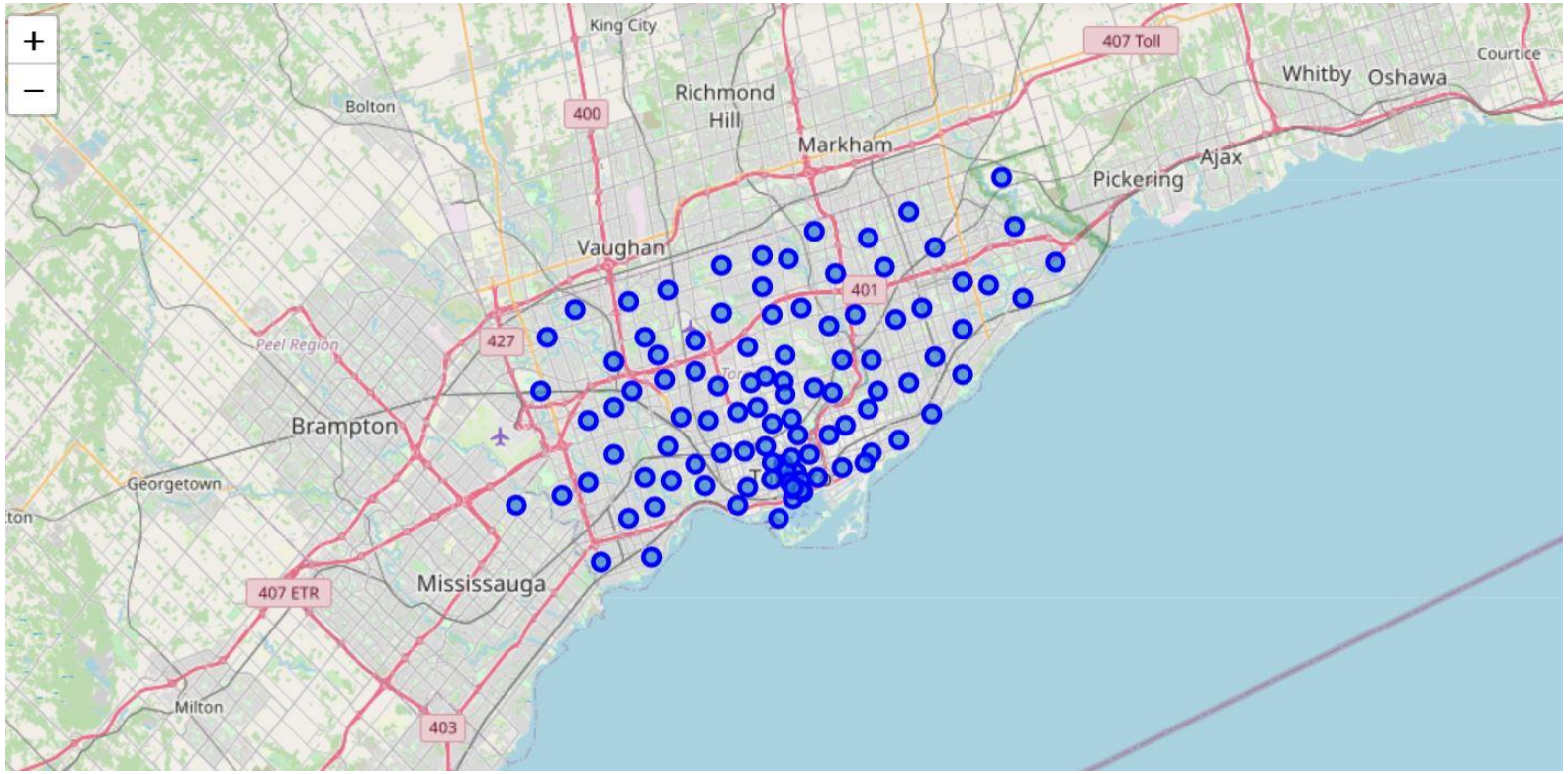
Business Problem: -

- The objective of this IBM data science capstone project is to analyze and select the best location for opening a shopping mall in city of Toronto, Canada. So, for that we are using the data science methodology and machine learning technique such as clustering, this project provide solution to the business question.
- Business question :- In the city of Toronto, Canada if a property developer is lolling to open a new shopping mall, where would you recommend that they open it?

Data Discussion:

- Data Required :
  - list of neighborhoods in city of Toronto
  - Longitude and latitude coordinates of the neighborhood.
  - Venue data related to shopping malls.
- Data Collection:
  - Geographical coordinates using CSV file (http://cocl.us/Geospatial_data)

Creating a map of Toronto by latitude and longitude values using map folium function.

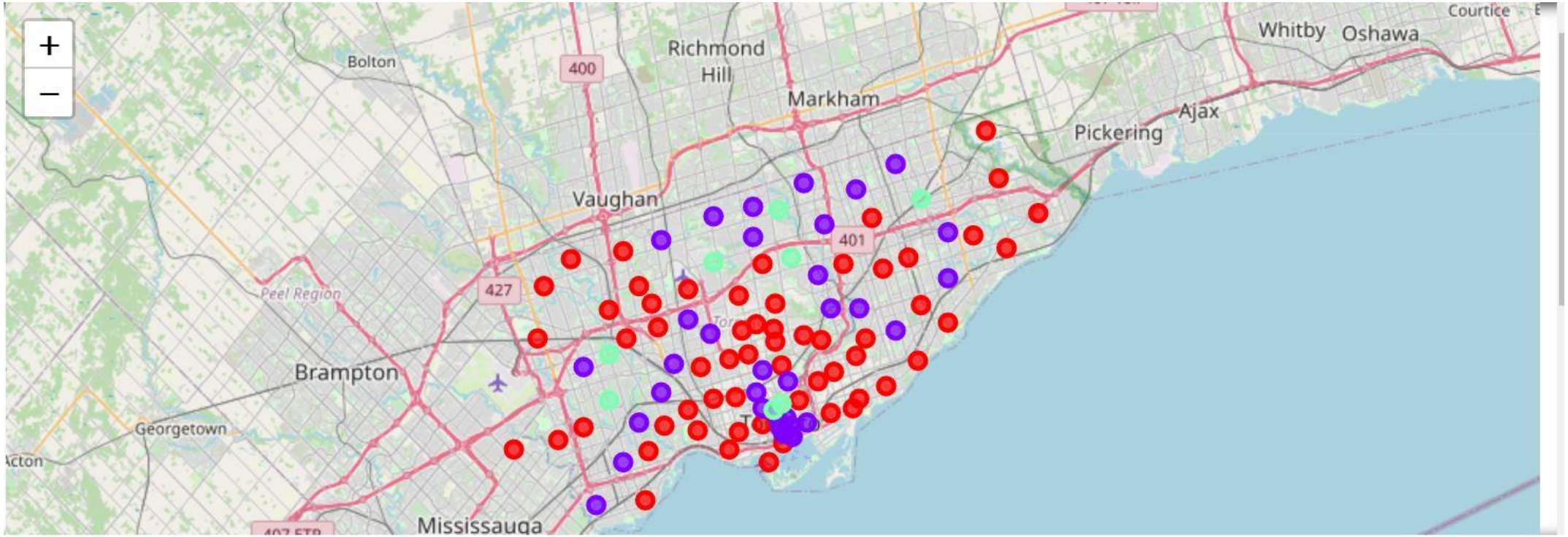


Methodology:

- Firstly, we will collect the list of neighborhoods in the city of Toronto using the Wikipedia page(https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M). Then we will do web scrapping using the beautiful soup and python request to extract the list of data. So, this is just a name along with the postal code. We need the geographic data such as longitude and latitude of respective data's so that we used CSV file that has the geographic coordinates of each postal code and Neighborhood (http://cocl.us/Geospatial_data). Then we consolidated two data, respectively. After collecting all the data, we can visualize the map of Toronto using the folium package.
- Secondly, we will use the foursquare API to get the 100 venues in the radius of 2000meters in Toronto city. For that, we need to make API calls to foursquare passing in the geographic coordinates of the neighborhoods in a loop. Foursquare API will return all the data in JSON format. We need to extract the Venue Name, Venue Latitude, Venue Longitude, Venue Category. With this data, we can also check how many unique categories received. Then we will analyze each neighborhood by grouping the rows and taking the mean frequency occurrence of each category and preparing this data for the clustering. As we want to analyze the shopping mall data, we will filter the venue like a shopping mall in neighborhoods.
- Lastly, we will perform the clustering of the data by using k-means clustering. It is one of the most popular and simplest unsupervised machine learning techniques suitable for this project. We will now cluster neighborhoods into three clusters based on their frequency of occurrence for the shopping mall. The result will give clarity in which neighborhood has the highest concentration of shopping malls and which have a lower number of a shopping mall. Based on the occurrence of the neighborhood it will help us to answer the question as to which neighborhood are most suitable for opening the shopping malls.

Results

- Cluster 0 (Red Color) : Neighborhood has no existence of the shopping mall.
- Cluster 1 (Purple Color) : Neighborhood has moderate number of shopping mall.
- Cluster 2 (Mint Green): Neighborhood has high number of shopping mall.



Conclusion

- The project recommended the builder to open a new shopping mall in cluster 0.
- This project help the builder or investor to open and invest in a good location to get a maximum amount of profit with less amount of competition.

Thank you