The Battle of the Neighborhoods

Coursera Applied Data Science Capstone Final Project

Introduction

The purpose of this project is to help identify better neighborhoods for people who are looking for good neighboorhoods in Scarborough, Toronto.

The audience of this project is the people who are seeking good neighborhoods to live in Toronto. They will care about this problem because this project will help them select a good neighborhood for them.

This project will help people understand the neighborhood and help them select the best one for their purposes.

The goal of this project is to analyze a variety of features of Scarborough and the neighborhoods around this area and compare them. The variables being compared include school ratings, median housing price, crime rates, weather conditions, and resources.

Scarborough is a neighborhood that is growing in popularity, and is gaining many residents. This neighborhood is especially popular among immigrants.

Main purpose of this project:

The main purpose of this project is to suggest a better neighborhood in a new city for the people interested.

Comparing infrastructure

Sorted list of house in terms of housing prices in order

Sorted list of schools in terms of location, fees, rating and reviews

Data

The project will use Foursquare API because it is an efficient and detailed data collecting source

Data Link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Will use the same Scarborough dataset that we scrapped from wikipedia on Week 3.

The dataset consists of latitude and longitude, zip codes.

In order to gain data about different venues in different neighborhoods of specific borough, we will use Foursquare location information. Foursquare API will be the only data source used because it provides information such as venue names, locations, menus, and photos.

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 100 meter.

The data retrieved from Foursquare contains information of venues within a specified distance of the longitude and latitude of the postcodes.

Information obtained per venue:

Neighborhood, Neighborhood Latitude, Neighborhood Longitude, Venue, Name of the venue e.g. the name of a store or restaurant, Venue Latitude, Venue Longitude, Venue Category

Methodology

Clustering using K-means clustering approach:

To compare the similarities of two neighborhoods, we explore, segment, and group them into clusters to find similar neighborhoods in a big city like New York and Toronto. To do that, we need to cluster data using k-means clustering algorithm
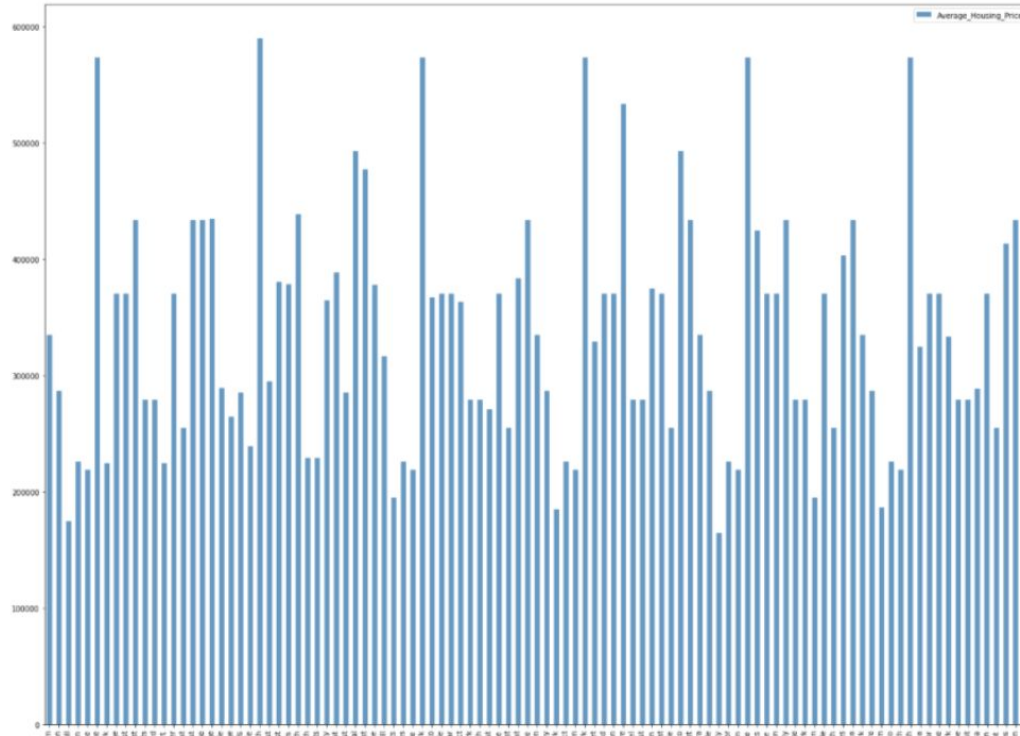
Work flow:

Due to http request limitations in Foursquare API, the number of places per neighborhood parameter would be set to 100 and the radius parameter would be set to 500.
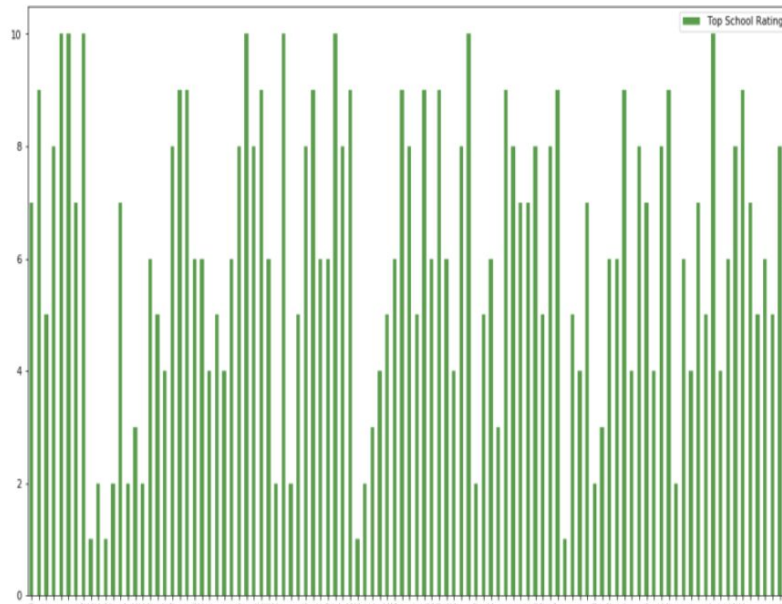
Results - Map of Clusters in Scarborough
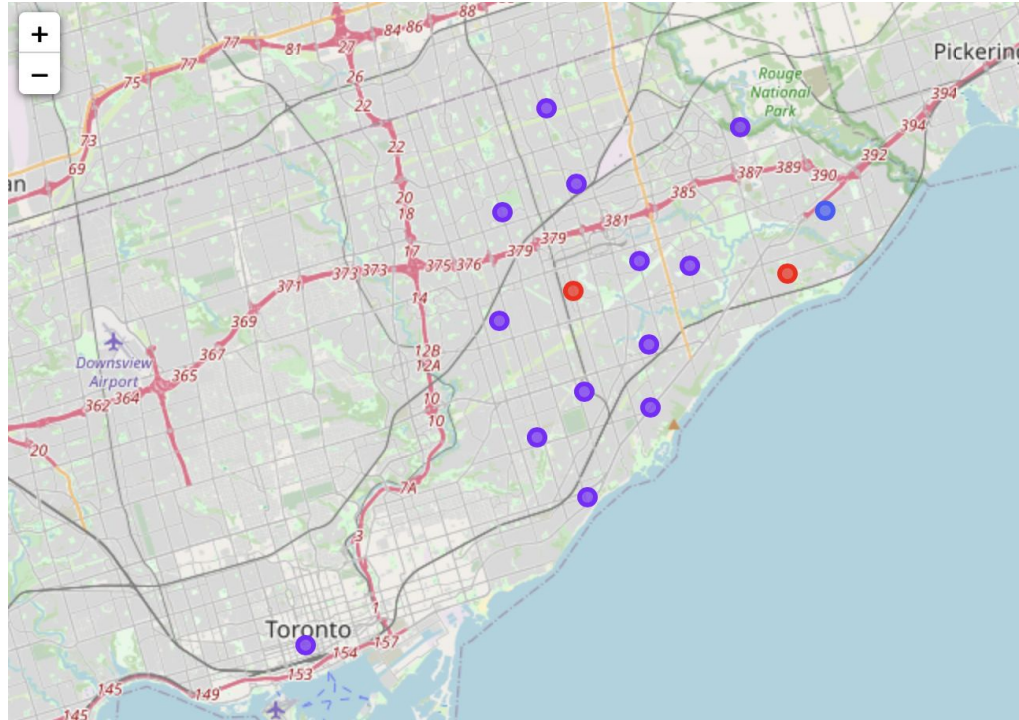
Results - Average housing price by clusters in Scarborough

Results - School ratings by clusters in Scarborough

Location

Scarborough is a neighborhood that is growing in popularity, and is gaining many residents. This neighborhood is especially popular among immigrants.

Foursquare API

In order to gain data about different venues in different neighborhoods of specific borough, we will use Foursquare location information. Foursquare API will be the only data source used because it provides information such as venue names, locations, menus, and photos.

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 100 meter.

The data retrieved from Foursquare contains information of venues within a specified distance of the longitude and latitude of the postcodes.

Discussion

The problem we are trying to solve is that many people have trouble selecting a good neighborhood that fits their needs.

The main purpose of this project is to suggest a better neighborhood in a new city for the people interested.

Comparing infrastructure

Sorted list of house in terms of housing prices in order

Sorted list of schools in terms of location, fees, rating and reviews

Conclusion

In this project, the neighborhoods were separated into 10 different clusters for 103 different latitude and longigute values from the dataset, which have very similar neighborhoods around them. The charts above results present a particular neighborhood based on average house prices and school ratings.

This project was created for the final project of the applied data science capstone Coursera class. Many data science tools were used. Folium is a very useful technique to consolidate information and make the analysis and decision on the map.

This project can be further developed to help find best house in Scarborough, in terms of all required things (daily needs/things needed to live a better life) around and also in terms of cost effectiveness.

Conclusion part 2

The libraries that were used:

Pandas: For creating and manipulating dataframes.

Folium: Python visualization library would be used to visualize the neighborhoods cluster distribution of using interactive leaflet map.

Scikit Learn: For importing k-means clustering.

JSON: Library to handle JSON files.

XML: To separate data from presentation and XML stores data in plain text format.

Geocoder: To retrieve Location Data.

Beautiful Soup and Requests: To scrap and library to handle http requests.

Matplotlib: Python Plotting Module.