Final Report | Capstone Project – The Battle of Neighbourhoods Finding a Better Place in Scarborough, Toronto

1.Introduction:

This project aims to analyse the various amenities within a particular neighbourhood and then give a comparative analysis of the features across various neighbourhoods. The features include median housing price, school rankings according to their ratings, crime rates of the particular area, road connectivity, weather conditions, efficiency of emergency services , availability of water etc , among others.

It will serve as a guideline for people aiming to move to Scarborough from other places and help them select the most appropriate neighbourhood according to their needs.

2.Data Section

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

I have used the same data that we scrapped from Wikipedia for our week 3 assignment

Due to some issue in my notebook, beautiful soup was not working and hence I have used the csv file provided to us in week 3 to get the location data.

This csv file has the latitude and the longitude of the various neighbourhoods of Scarborough and is stored in a data frame .

The data scraped from the link is used to create a data frame and then the data frame is merged with the latitude data frame to create a complete data set for the purposes of the project.

Foursquare API Data:

We will need data about different venues in different neighbourhoods of that specific borough. In order to gain that information we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the required information can be obtained through the API.

After finding the list of neighbourhoods, we then connect to the Foursquare API to gather information about venues within each neighbourhood. For each neighbourhood, I have chosen the radius to be 100 metres.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighbourhood

2. Neighbourhood Latitude

3. Neighbourhood Longitude

4. Venue

5. Name of the venue e.g. the name of a store or restaurant

6. Venue Latitude

7. Venue Longitude

8. Venue Category

3. Methodology Section

Clustering Approach:

To compare the similarities of two cities, we decided to explore neighbourhoods, segment them, and group them into clusters to find similar neighbourhoods in a big city like New York and Toronto. To be able to do that, we need to cluster data which is a form of unsupervised machine learning: k-means clustering algorithm.

Work Flow:

Using credentials of Foursquare API features of near-by places of the neighbourhoods would be mined. Due to http request limitations the number of places per neighbourhood parameter would reasonably be set to 100 and the radius parameter would be set to 500.

4. Conclusion Section

In this project, I used k-means cluster algorithm to separate the neighbourhood into 10(Ten) different clusters which have very-similar neighbourhoods around them. I used this data to then find out the most occurring venues in various neighbourhoods.

Besides this, I charted the average house prices for each neighbourhood and school ratings for the same neighbourhoods.

I was able to apply all the tools that were taught to me over this 9-course journey to create this capstone project and I am very grateful to IBM and Coursera for this opportunity. I hope to continue doing more projects like this to improve my skills and knowledge.

Future Works:

This project can be further improved by including further analysis for affordable housing according to the users’ budget.

It could also be improved upon to include financial data about the current residents’ of the various neighbourhoods, thereby giving the user a better idea about the relative affluence of the neighbourhoods.

Libraries Used to Develop the Project:

Pandas: For creating and manipulating data-frames.

Folium: Python visualization library: used to visualize the neighbourhoods cluster distribution of neighbourhoods.

Scikit Learn: For importing k-means clustering model.

JSON: Library to handle JSON files.

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

Matplotlib: Python Plotting Module.

CSV: The location data file included in the week 3 module was used to include the latitude and longitude