Coursera capstone

IBM applied data science capstone

Choosing a neighbourhood to live in Toronto, Canada

By: Belal Ibrahim

Introduction:

Finding a suitable neighbourhood to live in according to your needs can be a very hard task especially in a very big city as Toronto, which is the economic center of Canada. This city has more than 103 neighbourhoods which makes comparing all of them manually very hard.

Business problem:

The aim of this project is to cluster similar neighbourhoods in the city of Toronto together so that people can easily choose where to live based on their interests, lifestyle, and the data of each neighbourhood.

Target audience of this project:

This tool can also be used by real estate agents to find the suitable location for the customer based on their interests and lifestyle. It can also be used by individuals to choose the neighbourhood to live in

Data used:

To make this project I needed the following data:

- Neighbourhood names
- Neighbourhood locations
- Venue data for each venue

I used data from a wikipedia table to get the names of the neighbourhoods of Toronto. I also used the data from an online dataset to get the coordinates of the neighbourhoods based on their postal code. This is the link to the datasets that I used:
"https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M"
"https://cocl.us/Geospatial_data"
Then I used data from the foursquare API to get the number and type of venues at each neighbourhood, then used it to cluster the venues together based on their number.

Methodology:

First I got the list of neighbourhood names from the wikipedia page, then I scraped and cleaned it using the pandas module. Second I needed to get the locations of the neighbourhoods so I used this online data set: "https://cocl.us/Geospatial_data" to get it. Then I used the Foursquare API to get data about the venues around each neighbourhood like their names, category and location. After that, I used the kmeans algorithm to cluster the neighbourhoods according to the number and type of venues they each one has. The neighbourhoods were clustered into 5 clusters with similar numbers and types of venues. Finally I used the Folium module to draw a map Toronto with all the neighbourhoods superimposed on top of it.

Results:

The results of this project is a dataframe with all the neighbourhoods of Toronto, their coordinates, number of venues, most common venue type, borough, postal code, and a cluster label indicating the cluster they belong to.The dataframe was used to draw a map with the neighbourhoods positioned on top of it and colour coded according to their cluster:

Red: cluster 0

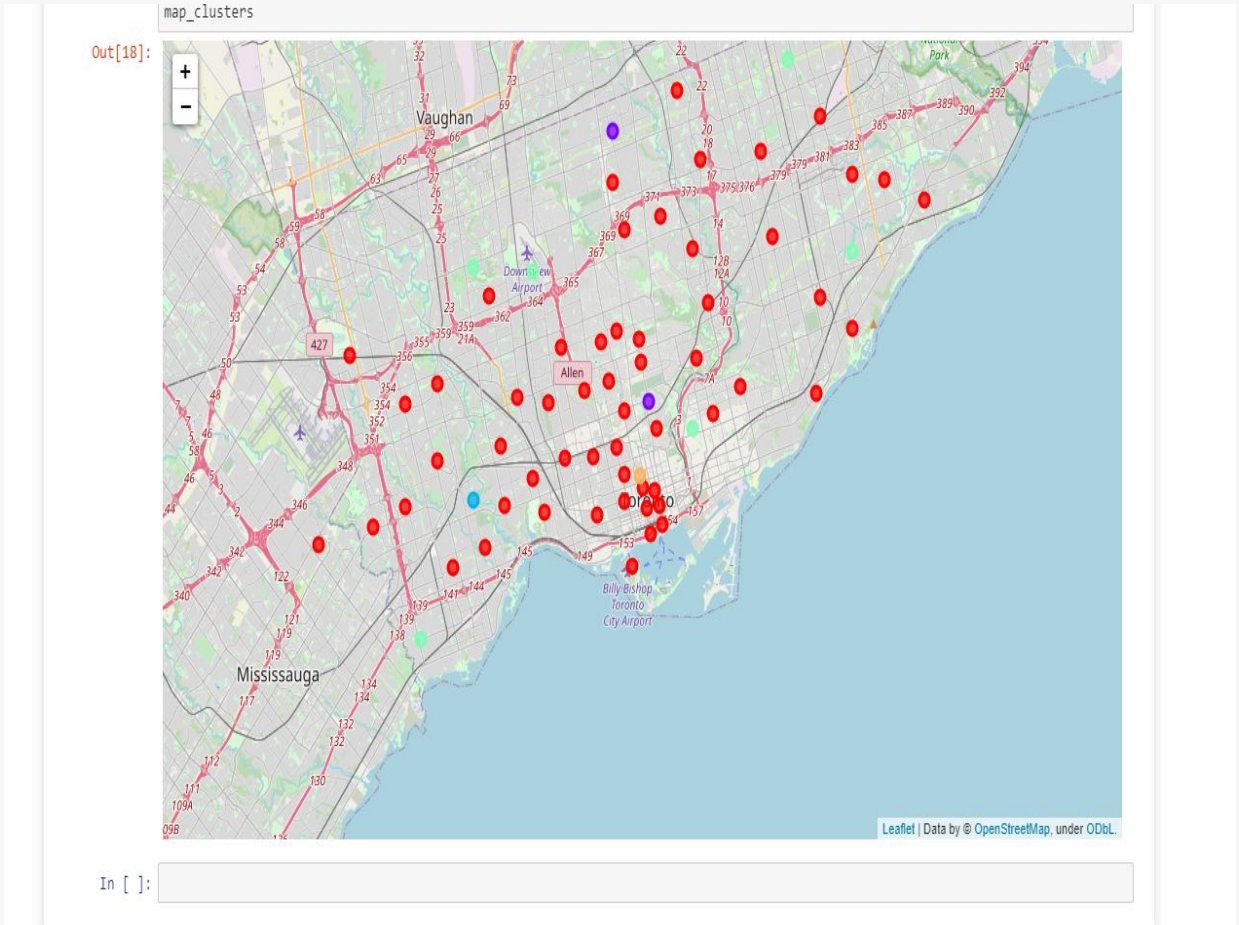
Purple: cluster 1

Light blue: cluster 2

Green: cluster 3

Orange: cluster 4

This is how the data looks when visualized on the map:



Discussion:

The dataframe has some neighbourhoods in the same row because they have the same postal code as the data set I used depended on postal codes to get the coordinates of the neighbourhood. That is why sometimes some neighbourhoods are represented by the same point on the map.
It is also noticed that most of the neighbourhoods were in cluster 0 which means that most neighbourhoods are similar to each other in the number and type of venues they have.

Conclusion:

In conclusion, this tool can be very useful to real estate agents in recommending neighbourhoods to customers or even individuals who want to find a suitable neighbourhood to live in.