Final Project of Applied Data Science Capstone
- Find a new living area in Toronto

1. Introduction

It is quite common to move to a new place for various purposes, such as get a new job, go to college, or even for retirement. However, it is difficult to foresee the living situation in the new places, especially for a totally strange place that you have never been to or even in another country. Although there is a lot of information on the internet that you can access to evaluate the city or county that you are moving to, you cannot collect all the details in an easy way since the information is mostly fragmented.

Essentially, the quality of life (QoL) can be determined by many factors. For example, how far the supermarket is? Is there a hospital near me? Will it be loud in the night due to the prosperous bar business? Do I have the access to any green area like parks or river for jogging? All these things can play a role in influencing the QoL when people search for a place to live. Nevertheless, the apartment or the house itself is also important, but that is out of the scope of this project.

Therefore, a quick and thorough way to gather all the necessary information and then present it to the people with the need before they move would be a great help in this context. The aim of this project is to preliminarily present a simple system to help the users choose their future residential area based on their subjective requirement. Based on their requirement, the system will go through the internet to retrieve all the locational information and categorize them into clusters so to recommend some suitable areas to the users.

To make the project closer to the realistic scenario, we assume that we are a real estate agency and providing the service to help our clients find their ideal residence in a brand-new area. Here is requirement from our client, David, who just got a new job in Toronto and will move there within 2 months. Therefore, he contracted us with this task to find an ideal place and an apartment for him.

David is a software engineer. He works mostly from home and only needs to go to the office for meetings sometimes. Therefore, the distance between his home and office is not that important to him. However, the quality of life is the thing he cares the most in the new neighborhood.

During the interview with David, he told us that he does not like to cook because he hates the cooking smell inside the house, so the most critical thing for him is to have access to any restaurants right down the street within 200 meters. In addition, he likes drinking fresh coffee from the coffee shops because that really opens his day. Besides, due to his job, he mostly

sits in front of the computer during the day, so he likes to work out in a gym after the long working time. These two factors are equally important for him. However, David is allergic to many things since he was a little kid, so it would be the best if he can avoid living in an area close to parks, flower stores, pet stores, or something similar.

So, this is the requirement from our client and now it is time to design a system and solve this problem.

2. Data for this project

To help David find the ideal place to live, we will need the locational data from Toronto, which we will retrieve from Foursquare. After retrieving the data, we will perform a clustering process to filter the top 3 neighborhoods for David to choose and that will be the end results of this project.

3. Methods

In this project, we used python to implement our system and the detailed steps are described in the following sections.

3.1. Retrieve necessary location data from Wiki and Foursquare.

First of all, we imported libraries, including the pandas, numpy, bs4, sklearn, matplotlib, folium, geopy.geocoders and json. Then we set up the credentials for retrieving data from Foursquare. Next, we got the borough data of Toronto for searching the location data from Foursquare. There is a table listing all the boroughs, post codes and neighborhoods from Wikipedia.

	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government

Figure 1. Dataframe that includes all the data, including the postal code, borough, and the neighborhood of Toronto.

3.2. Clean up data.

After completing the previous section, we created a dataframe with the data from wiki and used the goecoders to get the latitudes and longitudes of each neighborhood to attach back to the full dataframe.

Next, we sent the request for location data via Foursquare API for our neighborhoods. To make it easier, we made a function to loop through the neighborhoods and the function returned the venues under the radius of 200 meters from each neighborhood in Toronto that we passed in. After the data was retrieved, we appended it to the previous dataframe.

So far, we have collected all the real-time data of the venue information from the date May 1st, 2021. To better screen the suitable neighborhoods to match the requirement of the client, we categorized the shops into bigger categories, such as "All restaurant", "All coffee", "All sports", "All bars", and "All allergies".

	Neighborhood	All restaurants	All sports	All coffee	All allergies	All bars	Accessories Store	Adult Boutique	Art Gallery	Arts & Crafts Store	Auto Workshop	BBQ Joint	Bank	Bike Shop	Bookstore	Boutique	Bubble Tea Shop	Burrito Place	Cheese Shop	Chocolate Shop	Clot S
0	Alderwood, Long Branch	0.333333	0.0	0.333333	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Bathurst Manor, Wilson Heights, Downsview North	0.714286	0.0	0.142857	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Bedford Park, Lawrence Manor East	0.666667	0.0	0.111111	0.0	0.111111	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Birch Cliff, Cliffside West	0.000000	0.0	1.000000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Brockton, Parkdale Village	0.000000	0.0	0.500000	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 2. Categorized venue data and their proportion in that neighborhood.

3.3. Perform analysis, clustering, and visualization.

After all the data was ready, we performed K-means clustering to see which neighborhoods are closer to the other based on the venues within their area. After finishing this step, we sorted the dataframe with the top 5 venues and appended it back.

	Cluster Labels	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	0	Alderwood, Long Branch	All restaurants	All coffee	Pharmacy	Wings Joint	Organic Grocery
1	2	Bathurst Manor, Wilson Heights, Downsview North	All restaurants	All coffee	Deli / Bodega	Jewelry Store	Opera House
2	0	Bedford Park, Lawrence Manor East	All restaurants	All coffee	All bars	Cupcake Shop	Kids Store
3	4	Birch Cliff, Cliffside West	All coffee	All restaurants	Jewelry Store	Opera House	Music Venue
4	4	Brockton, Parkdale Village, Exhibition Place	All coffee	Playground	All restaurants	Jewelry Store	Opera House

Figure 3. Neighborhood with their respective cluster labels

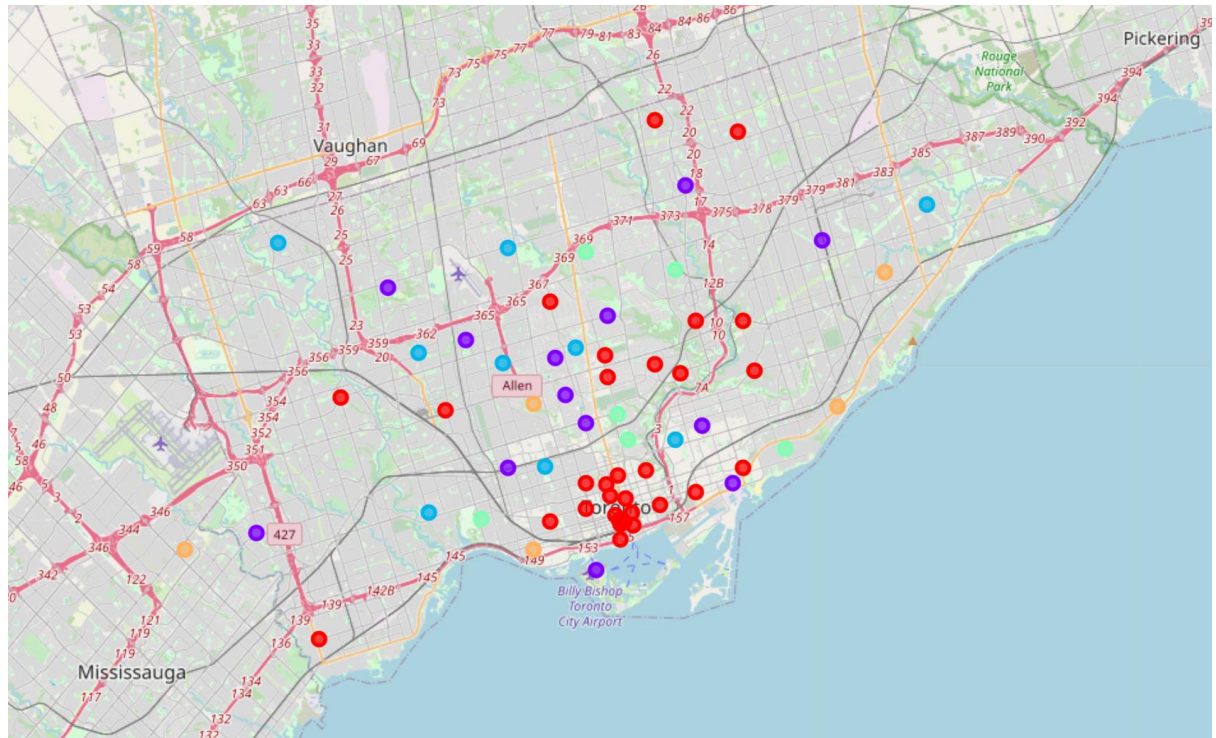We then plotted the clustered neighborhoods on the map showing their locations in Toronto.

Figure 4. The distribution of all the clustered neighborhoods differentiated by color.

4. Results

According to David's requirement, the category "All restaurants" must be the first one, and "All coffee" and "All sports" must be either 2nd and 3rd places (the order does not matter). The most important criteria is the category "All allergies" must not appear in the top 5 list for his own good. Therefore, we could easily identify which neighborhoods are the best choice for David.

	Borough	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
70	Etobicoke	0	All restaurants	All coffee	Jewelry Store	Opera House	Music Venue
79	Central Toronto	0	All restaurants	All coffee	Dessert Shop	Toy / Game Store	Bookstore
80	Downtown Toronto	0	All sports	All coffee	BBQ Joint	All restaurants	Trail
84	Downtown Toronto	0	All restaurants	All coffee	All bars	Arts & Crafts Store	Dessert Shop
90	Scarborough	0	All restaurants	Bank	Pharmacy	All sports	All coffee
92	Downtown Toronto Stn A	0	All restaurants	All bars	All coffee	All allergies	Hotel
93	Etobicoke	0	All restaurants	All coffee	Pharmacy	Wings Joint	Organic Grocery
96	Downtown Toronto	0	All restaurants	All coffee	All allergies	All bars	General Entertainment
97	Downtown Toronto	0	All restaurants	All coffee	All sports	Deli / Bodega	All bars
99	Downtown Toronto	0	All restaurants	Burrito Place	Martial Arts School	All allergies	All bars

Figure 5. The potential neighborhoods that may fulfill the requirement of the client.

	PostalCode	Borough	Neighborhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
97	M5X	Downtown Toronto	First Canadian Place, Underground city	43.648429	-79.38228	0	All restaurants	All coffee	All sports	Deli / Bodega	All bars

Figure 6. In this project case, the No.97 with Postal code "M5X" should be the suitable choice.
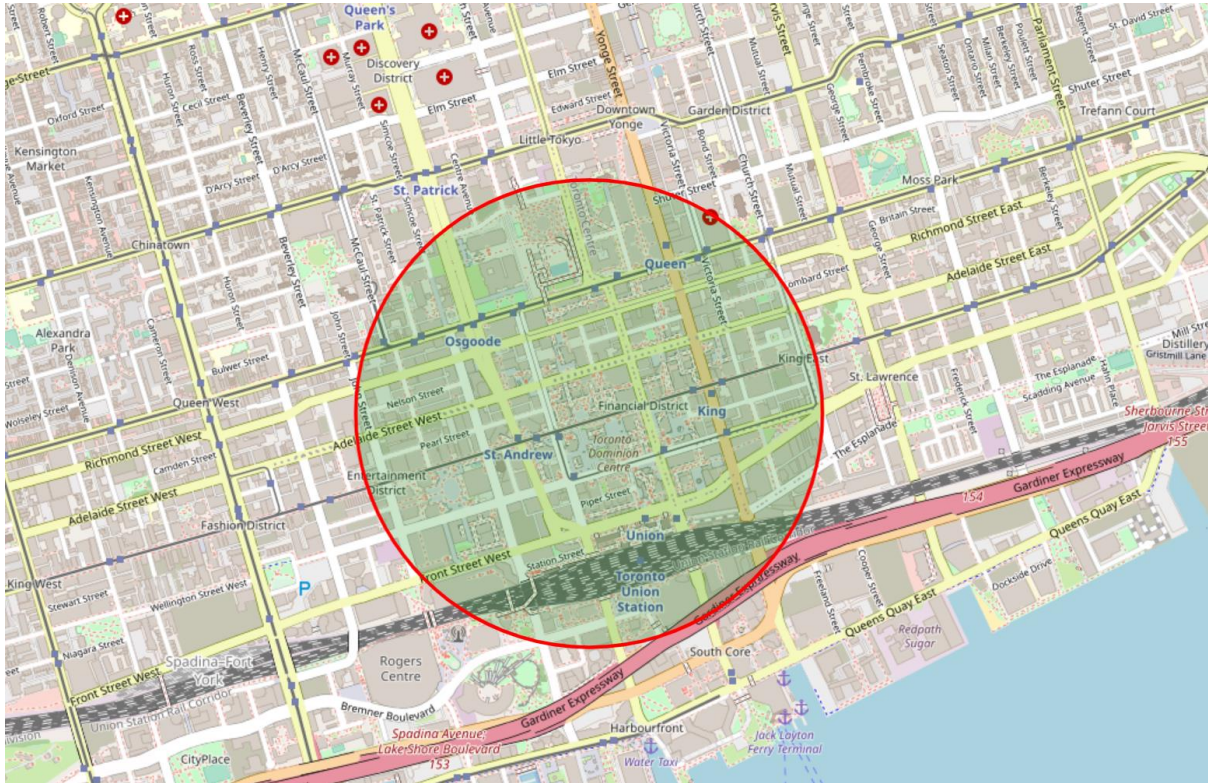
Figure 7. The rough location of the selected neighborhood on the map.

5. Discussion and conclusion

After the data analysis, we found out that the most suitable areas for David to live in Toronto would be the neighborhood of First Canadian Place or Underground city, which are both inside Downtown Toronto. The reason is that the restaurants here are the most common venues and are followed by coffee shops or cafe and sports related facilities. Most important of all, there are no venues that could cause allergy to David. Therefore, these two neighborhoods are the end results of this project.

For the future steps, a more thorough investigation of the price of apartments located in these two neighborhoods should be conducted, so that David can find the apartment that suits him the most. However, this will require more information from him because there are also lots of criteria that can influence the price of an apartment, such as the size of the apartment, the room numbers, the interior design, and so on.