**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.

1. **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.

1. **Methods**

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

* 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.

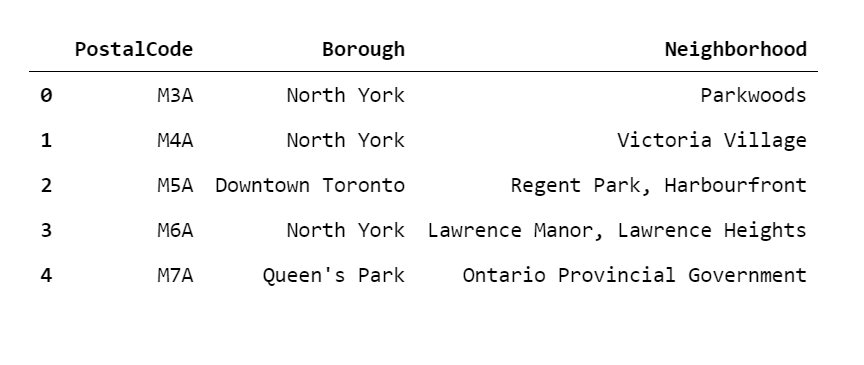


Figure 1. Dataframe that includes all the data, including the postal code, borough, and the neighborhood of Toronto.

* 1. 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”.

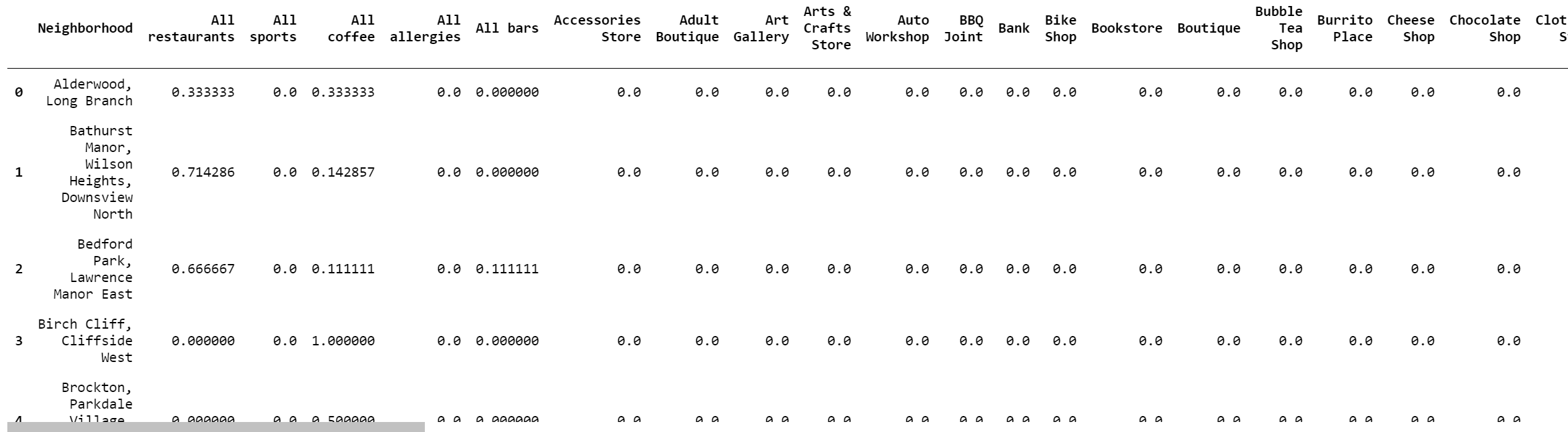
****

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

* 1. 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.

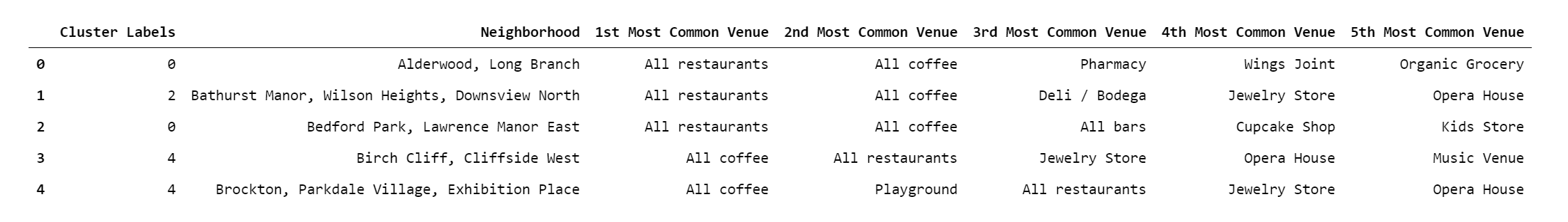


Figure 3. Neighborhood with their respective cluster labels

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

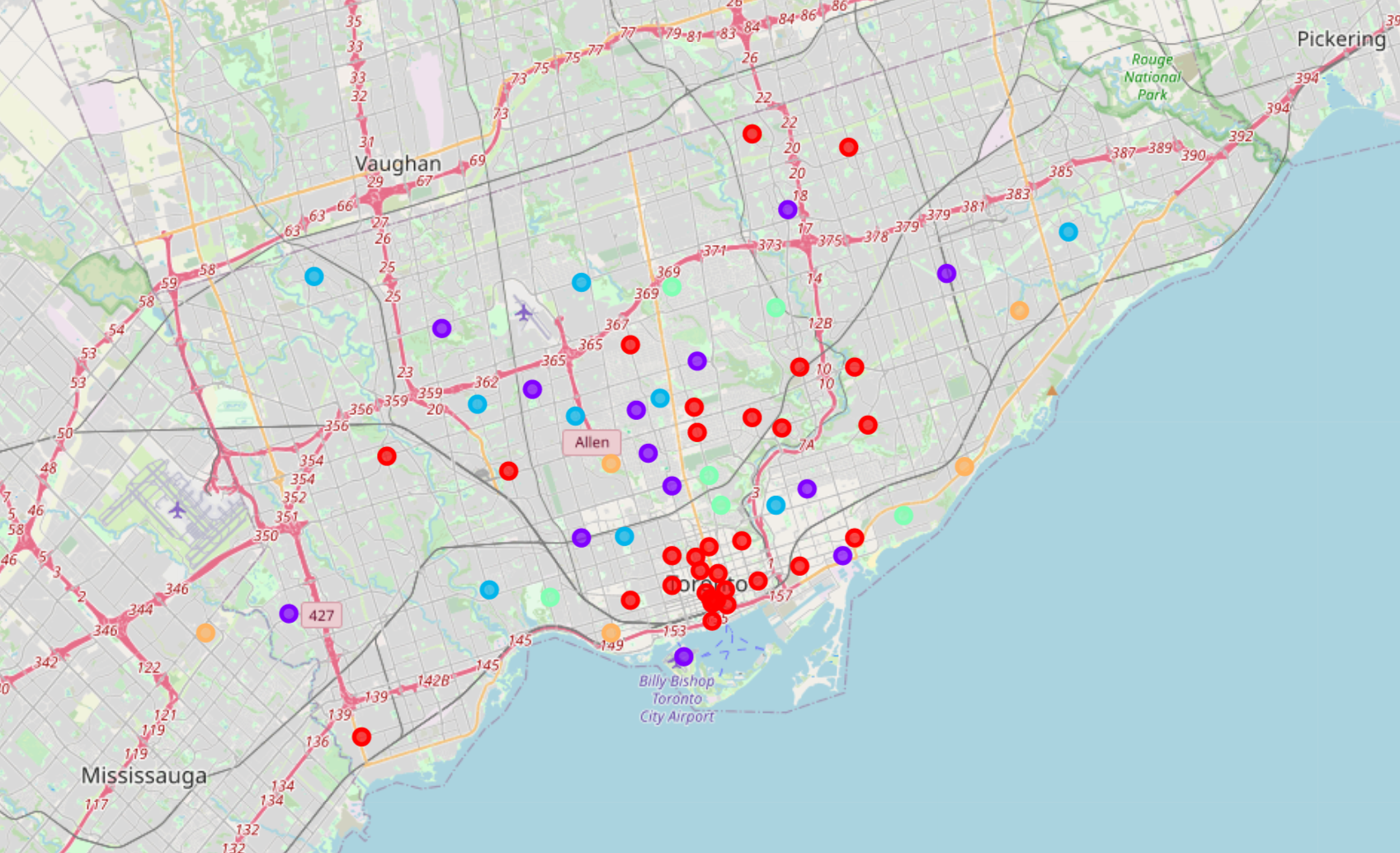


Figure 4. The distribution of all the clustered neighborhoods differentiated by color.

1. **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.



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

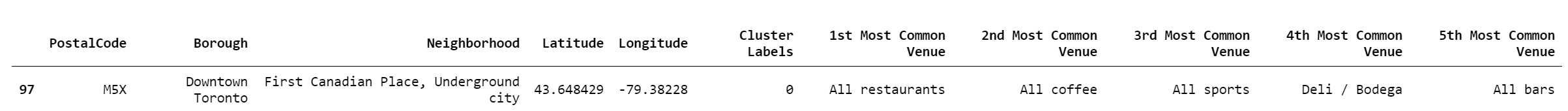


Figure 6. In this project case, the No.97 with Postal code “M5X” should be the suitable choice.

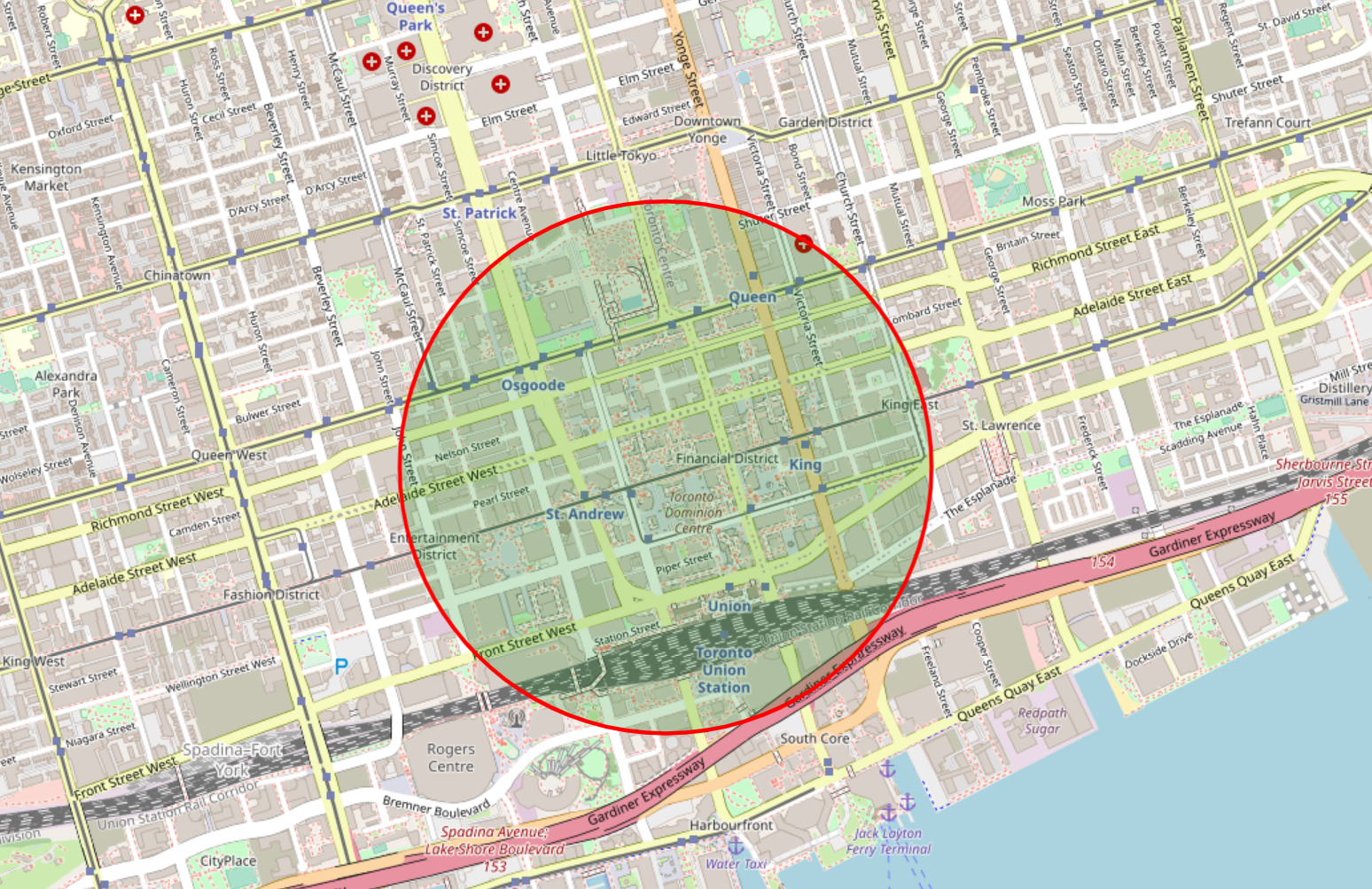


Figure 7. The rough location of the selected neighborhood on the map.

1. **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.