Predicting the location for a new hostel

1. Background

Travelling is one of the most amazing experience anyone can have. The travel industry is one of the largest service industries in the world, an increasingly important one in the modern age. In the last decade traveling for tourism has been on the rise among the youth. Therefore, finding an affordable accommodation in a lively neighborhood is extremely important for anyone visiting a new country for a short stay. That's why hostels now a days have been taking over the accommodation scene.

2. Problem description

In this report we will be studying the neighborhoods of Toronto city in Canada to find the best location for a new hostel. The challenge here is to find the geographical data for the different venues in each neighborhood in the city, then start analyzing this data to find the locations characterized by having venues that are most important for any traveler visiting a new country for a short stay, such as, coffee shops, restaurants, bars, rental car services, etc.. within a walking distance from the target location. It is also important to reduce competition, so we should aim at locations with less number of hotels.

3. Data

To accomplish this analysis, I'm going to use the following data as follows:

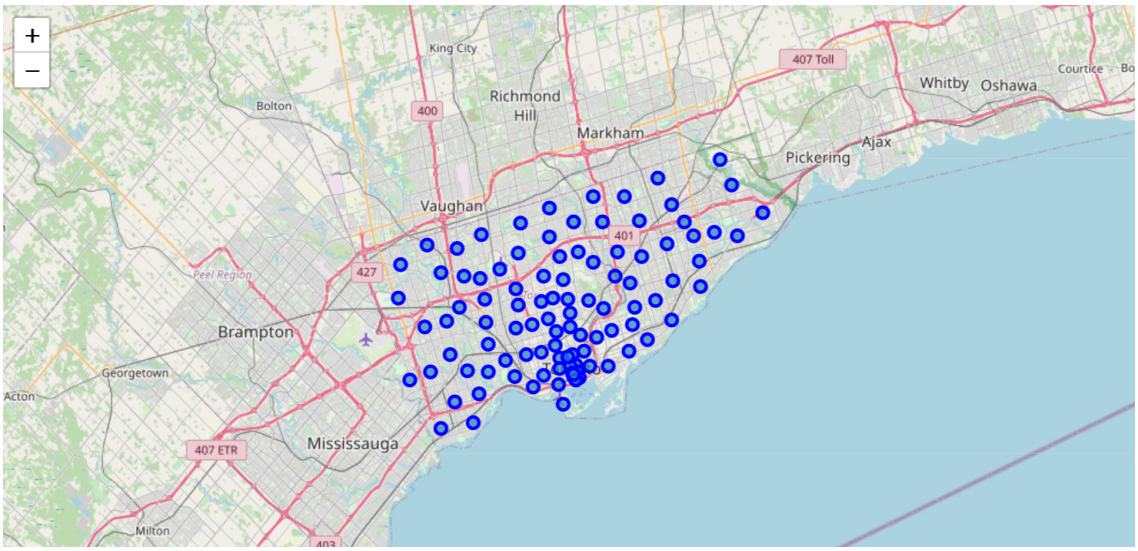
a. I will scrap the neighborhoods postal codes information from Wikipedia.
b. Using this postal codes, I will then use geopy library to get the geographical coordinates of each postal code.
c. Then using Foursquare API, I will gather the venues information within a radius of 500 meters around each location.
d. After transforming this information to a data frame, it will then be used to classify the neighborhoods into clusters using sklearn library.
e. The clusters with the most suitable characteristics for the location of a new hostel is going to be mapped to the neighborhoods with least number of hotels.

4. Methodology

To select the most suitable neighborhood for opening a new hostel, I will be looking for neighborhoods in Toronto city with the lowest number of existing hotels and the highest number of travelers-friendly venues such as restaurants, coffee shops, bars, etc.. within a walking distance of 500 meters.

First, I need to get the postal codes for the neighborhoods in Toronto. To do so, I scraped this information from Wikipedia https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M using BeautifulSoup library.

Then I had to clean the data by removing any postal codes with no borough names and assign the neighborhoods that has no names, the same borough name. After that, I used the pgeocode library to get the longitude and latitude for each postal code. Plotting them on the map to get an idea of their distribution as shown below.



The data revealed that Toronto has 102 neighborhoods divided over 9 boroughs. For each neighborhood, I will use foursquare API to gather information about all venues that lie within a radius of 500 meters, which will contain the location of the venue, its name and its category.

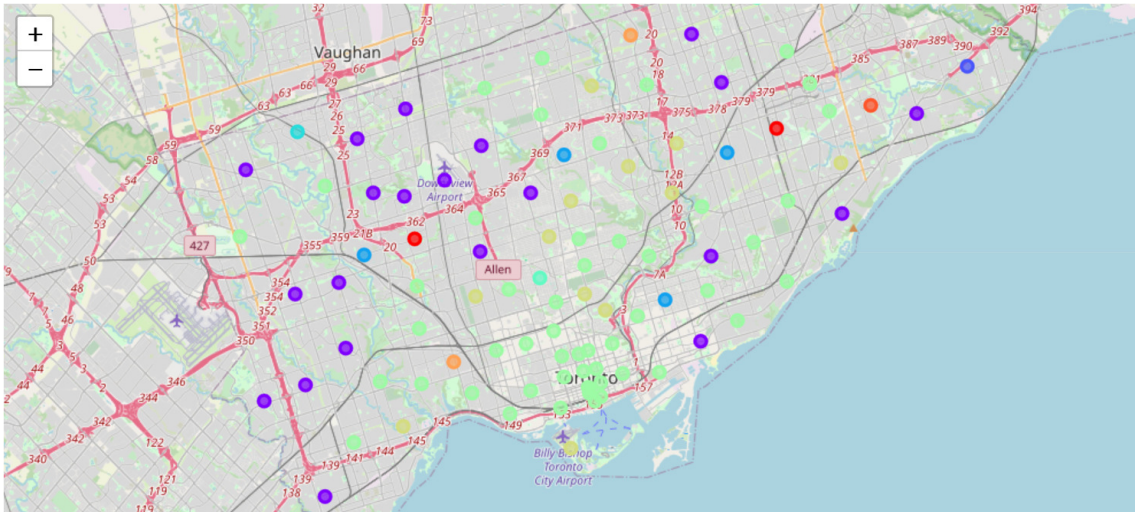
The following table is a sample of the output from this step:

neighborhood	neighborhood_latitude	neighborhood_longitude	venue	venue_latitude	venue_longitude	venue_category
Rouge Hill / Port Union / Highland Creek	43.7878	-79.1564	Scarborough Historical Society	43.788755	-79.162438	History Museum
Guildwood / Morningside / West Hill	43.7678	-79.1866	Chick-N-Joy	43.768752	-79.187982	Fried Chicken Joint
Guildwood / Morningside / West Hill	43.7678	-79.1866	Little Caesars Pizza	43.769046	-79.184386	Pizza Place
Guildwood / Morningside / West Hill	43.7678	-79.1866	LCBO	43.771462	-79.184384	Liquor Store
Guildwood / Morningside / West Hill	43.7678	-79.1866	Bulk Barn	43.771342	-79.184341	Food & Drink Shop

Now I will start classifying these neighborhoods into clusters using Kmeans method. This is the most suitable clustering technique in our situation, since the date we are dealing with is unlabeled data.

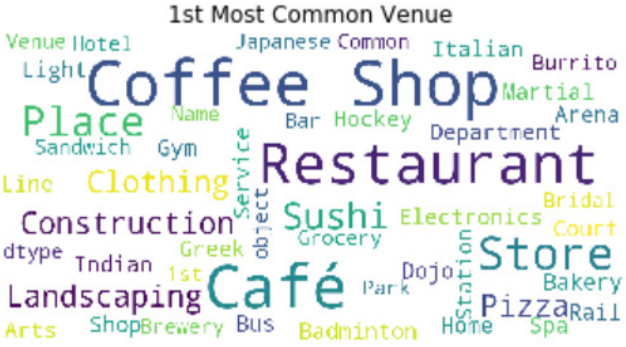
Before fitting my data, I need to get the mean of the frequency of occurrence of each category for and I will only consider the top 5 most common venues.

I segmented the neighborhoods into 10 clusters. I tried 5 clusters first, but 10 clusters gave better classification results. The following map shows the distribution neighborhoods classified into 10 clusters:



5. Results

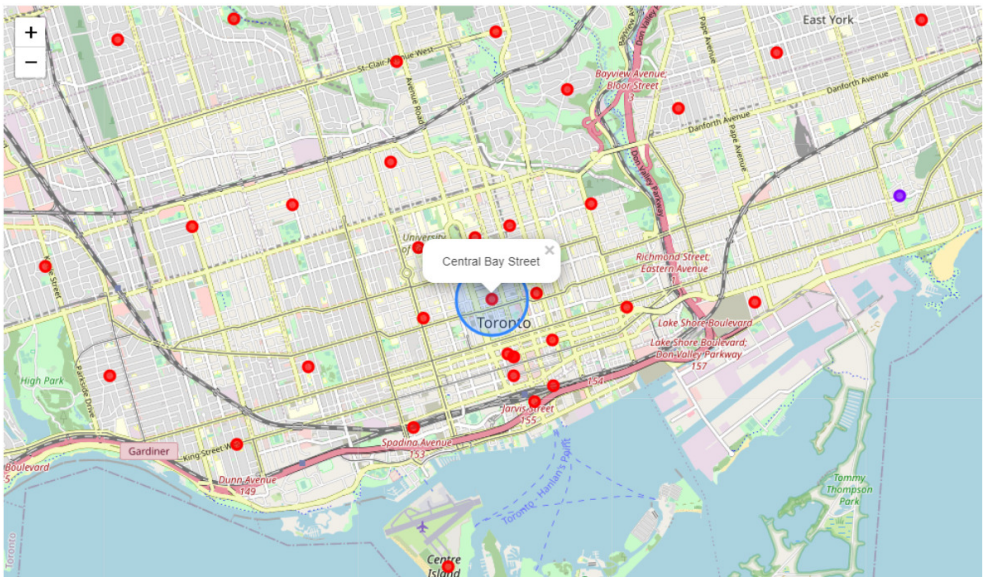
After exploring the resulting clusters, I found that cluster 6 is the liveliest cluster. It has all the venue categories that are suitable for our business proposition. As can be seen in the following word cloud plot, for the top five most common venues.



The following table shows the number of hotels in each neighborhood, we can see that we have 3 neighborhoods with only 1 hotel, so I picked Central Bay Street to examine further.

	neighborhood	venue_category	count
0	Toronto Dominion Centre / Design Exchange	Hotel	9
1	Commerce Court / Victoria Hotel	Hotel	7
2	First Canadian Place / Underground city	Hotel	7
3	Berczy Park	Hotel	5
4	Stn A PO Boxes	Hotel	4
5	Richmond / Adelaide / King	Hotel	3
6	Church and Wellesley	Hotel	2
7	Garden District, Ryerson	Hotel	2
8	Northwest	Hotel	2
9	St. James Town	Hotel	2
10	CN Tower / King and Spadina / Railway Lands / ...	Hotel	1
11	Central Bay Street	Hotel	1
12	Willowdale	Hotel	1

This is the location of the neighborhood on the map with a highlighted circle of 500 meters radius.



These are the top 10 venue types located within a radius of 500 meters around the center of the Central bay street neighborhood.

	neighborhood	venue_category	count
0	Central Bay Street	Coffee Shop	14
1	Central Bay Street	Italian Restaurant	3
2	Central Bay Street	Café	3
3	Central Bay Street	Sandwich Place	2
4	Central Bay Street	Breakfast Spot	2
5	Central Bay Street	Bubble Tea Shop	2
6	Central Bay Street	Middle Eastern Restaurant	2
7	Central Bay Street	Restaurant	2
8	Central Bay Street	Neighborhood	1
9	Central Bay Street	Park	1

6. Discussion

It has been observed that Toronto can be mainly classified to 3 main clusters, when using 10 clusters, 53% of the neighborhoods where grouped under cluster 6, which's characterized by dinning venues, like restaurants and coffee shops. Neighborhoods of the city center belong to that cluster. The second largest cluster is cluster 1, 21.43% of the neighborhoods fall in that cluster and this one forms the city edges where most of the shopping stores are located. The 3rd cluster is number 7. Neighborhoods that are mainly outdoor areas and parks which comprises around 12.2% of the total belong to that cluster.

7. Conclusion

In my study I have concluded that neighborhoods which belong to cluster 6 are the most suitable locations for a new hostel. I have also excluded the neighborhoods which are already crowded by hotels, so that left me with 3 choices, CN Tower, Central Bay Street, and Willodale. The three of them equally have the lowest number of hotel venues. And I finally picked Central Bay Street.