Selecting Location for an Apartment Building in Toronto

1. Introduction:

This project aims to find the ideal location for Apartments20 to build an apartment building in the city of Toronto, Canada. Apartments20 specializes in building apartments geared towards millennials. The company has had success in moderately sized cities across Canada. Now, the company is looking to take its business to the next level by expanding to the most populous city in the country. With nearly one and a half million millennials currently residing in the region (almost 20% of the total millennial population in Canada),[1] there is no shortage of potential tenants. That being said, the company has had trouble deciding on the ideal location of their breakthrough building.

After doing some additional research on millenials' preferences when choosing a residence, the company determined that millennials place a high degree of value on the neighborhood in which they live. So, the question becomes: what type of neighborhood is ideal? Apartments20 knows that millennials are a very-health conscious generation and prefer to be within walking distance of a gym or other trendy health studio (e. g. yoga studio, crossfit gym, spin classes). They also enjoy their nightlife and love to have multiple options for dining. Apartments20's goal is to find a location that will attract millennials by allowing them easy access to their favorite activities and venues.

2. Business Problem:

Apartments20 is looking to find neighborhoods in the city of Toronto that would be the most suitable for their new apartment building. Based on the research mentioned above, the ideal neighborhood for their new building will have close vicinity to health studios, a variety of restaurants, and nightlife destinations.

3. Data:

3 a. Data Sources

To find the best neighborhood for Apartments20 to build their new apartment building, I will use the Foursquare public dataset. I can use this data to find venues (e.g. Gym, Pizza Place, Hotspot Restaurant, etc.) in the neighborhoods of Toronto. More specifically, I can use the Foursquare API to get a list of all the venues within a given radius of a pair of geographical coordinates. To access the latitude and longitude of neighborhoods in Toronto I will be scraping the Wikipedia page entitled "List of postal codes of Canada: M." This will give me a list of the various neighborhoods and the postal codes they correspond to. Then I will be using data from the csv file (located here: http://cocl.us/Geospatial_data) to associate these neighborhoods with geographical coordinates.

3 b. Data scraping and pre-processing:

The data had to be pre-processed before I could analyze it and apply a Machine Learning algorithm. I began by scraping the Postal code, borough and neighborhood data from the Wikipedia page using the pandas library. Then I removed any missing data by dropping rows where the borough and/or neighborhood was “Not Assigned.” I aggregated the data by grouping neighborhoods with the same postal code together in the same row. I then checked to make sure there were no null values in the dataframe before adding the coordinate information.

To add the coordinates, I used the pandas library to read data from the CSV file into a pandas dataframe. Next, I merged the two dataframes so that the “Latitude” and “Longitude” columns were appended to the end of the dataframe discussed in the previous paragraph.

Now, I had to extract the venue data from Foursquare. To do this, I defined a function that would loop through all of the postal codes in my dataframe, find the venues within a specified radius, and then return a dataframe with every venue that was found and what postal code it belonged to. I then converted the venue data from categorical to numerical using one-hot encoding.

4. Methodology

4 a. Exploratory Data Analysis

Before executing our clustering algorithm, I wanted to explore what some of the most prevalent venues in various neighborhoods were. In order to do this, I created a function to rank the top 10 most common venues in each neighborhood, and output the results as a dataframe (see figure 1 below).

	Postal Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	M1B	Fast Food Restaurant	Trail	Coffee Shop	Bank	Restaurant	Chinese Restaurant	Bakery	Paper / Office Supplies Store	Caribbean Restaurant	Greek Restaurant
1	M1C	Italian Restaurant	Breakfast Spot	Burger Joint	Park	Playground	Zoo	Electronics Store	Elementary School	Empanada Restaurant	Ethiopian Restaurant
2	M1E	Pizza Place	Fast Food Restaurant	Bank	Coffee Shop	Burger Joint	Greek Restaurant	Liquor Store	Sandwich Place	Supermarket	Juice Bar
3	M1G	Park	Coffee Shop	Mobile Phone Shop	Indian Restaurant	Fast Food Restaurant	Pharmacy	Chinese Restaurant	Ethiopian Restaurant	Dumpling Restaurant	Eastern European Restaurant
4	M1H	Coffee Shop	Bakery	Gas Station	Bank	Indian Restaurant	Athletics & Sports	Chinese Restaurant	Thai Restaurant	Fried Chicken Joint	Grocery Store

Figure1

Above we see the first five rows of this dataframe, which shows the most popular venue types in the vicinity of the first five postal codes. As shown in the figure, there are a variety of restaurants in all of these neighborhoods shown. If this is an indication of the data as a whole, then the key will be to find neighborhoods that have access to gyms/health and nightlife options in addition to a variety of restaurants. I will segment the neighborhoods of Toronto into different clusters, and then evaluate which clusters most closely align with these attributes.

4 b. Modeling

In order to segment the neighborhoods of Toronto into different groups, I utilized the K-Means clustering algorithm. This will cluster the neighborhoods of Toronto together into groups that share similar characteristics. Once I have the neighborhoods clustered into groups, I will evaluate which cluster has characteristics that most closely align with the “ideal” neighborhood for Apartments20 to invest in a new apartment building. The cluster that has neighborhoods that best reflect the preferences of millennials (access to a variety of restaurants, nightlife options, gyms) will be the recommended cluster for Apartments20.

5. Results

I elected to use a k-value of 8, because it most evenly distributed the data points into different clusters. The vast majority the data points ended up in cluster 0, 3, or 5. (With other k-values, most of the data points were sorted into only one or two clusters). Now, I will examine the attributes of these three clusters and determine which best aligns with Apartments20’s “ideal” location.

<u>Cluster 0:</u>

Cluster 0 has a lot of food venues, but there is not too much variety. For example, “Fast Food Restaurant,” “Pizza Place” and “Coffee Shop” are seen several times each. In addition, no gyms or fitness studios are shown in this snapshot of the cluster. (See figure two below)

	Cluster Labels	Postal Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
2	0	M1E	Pizza Place	Fast Food Restaurant	Bank	Coffee Shop	Burger Joint	Greek Restaurant	Liquor Store	Sandwich Place	Supermarket	Juice Bar
3	0	M1G	Park	Coffee Shop	Mobile Phone Shop	Indian Restaurant	Fast Food Restaurant	Pharmacy	Chinese Restaurant	Ethiopian Restaurant	Dumpling Restaurant	Eastern European Restaurant
5	0	M1J	Ice Cream Shop	Convenience Store	Coffee Shop	Sandwich Place	Fast Food Restaurant	Pizza Place	Bowling Alley	Restaurant	Grocery Store	Train Station
7	0	M1L	Intersection	Coffee Shop	Bus Line	Convenience Store	Bakery	Park	Mexican Restaurant	Fast Food Restaurant	Sandwich Place	Beer Store
8	0	M1M	Pizza Place	Ice Cream Shop	Beach	Sports Bar	Cajun / Creole Restaurant	Burger Joint	Park	Hardware Store	Electronics Store	Elementary School

Figure2

<u>Cluster 3:</u>

In cluster 3, like in cluster 0, we see a number of food venues, but there is more variety in cluster 3. As opposed to seeing a bunch of Fast Food Restaurants and Pizza Places, the types of restaurants in this cluster include “Thai Restaurant,” “Turkish Restaurant,” “Chinese Restaurant,” “Italian Restaurant,” “Japanese Restaurant,” “Greek Restaurant,” and more.

Cluster 3 also includes a number of gyms in its most popular venues. For example, gym is the 9th and 10th most popular venue in postal codes M1N and M3K, respectively. It is also the 4th most common venue in M4N and 5th most common in M4P.

In terms of nightlife, neighborhood M4E stands out as having "pub" as its most common venue and "bar" as its 8th most common. Other neighborhoods also include bar, beer bar, brewery and/or pub in their 10 most common venues (e. g. M4J, M4M, M4K, M4L).

	Cluster Labels	Postal Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
9	3	M1N	Park	Restaurant	Café	Skating Rink	Thai Restaurant	General Entertainment	Diner	Dessert Shop	Gym	Photography Studio
29	3	M3K	Turkish Restaurant	Coffee Shop	Other Repair Shop	Sandwich Place	Chinese Restaurant	Liquor Store	Electronics Store	Italian Restaurant	Park	Gym
36	3	M4E	Pub	Coffee Shop	Beach	Pizza Place	Japanese Restaurant	Breakfast Spot	Caribbean Restaurant	Bar	Tea Room	Bakery
39	3	M4J	Coffee Shop	Café	Greek Restaurant	Pizza Place	Convenience Store	Park	Bar	Beer Bar	Fast Food Restaurant	Ethiopian Restaurant
40	3	M4K	Greek Restaurant	Coffee Shop	Café	Pub	Italian Restaurant	Pizza Place	Fast Food Restaurant	Furniture / Home Store	Ramen Restaurant	Bookstore
41	3	M4L	Indian Restaurant	Coffee Shop	Beach	Grocery Store	Café	Brewery	Park	Burrito Place	Bakery	Harbor / Marina
42	3	M4M	Coffee Shop	Bar	Café	Diner	Vietnamese Restaurant	Bakery	Brewery	American Restaurant	Italian Restaurant	French Restaurant
43	3	M4N	Park	Bookstore	Trail	Gym / Fitness Center	Coffee Shop	College Gym	Café	College Quad	Fast Food Restaurant	Farmers Market
44	3	M4P	Coffee Shop	Italian Restaurant	Dessert Shop	Café	Gym	Pizza Place	Pharmacy	Sushi Restaurant	Supermarket	Restaurant
45	3	M4R	Coffee Shop	Italian Restaurant	Skating Rink	Diner	Mexican Restaurant	Café	Park	Sushi Restaurant	Bakery	Jazz Club

Figure 3

Cluster 5:

In cluster 5 there is a moderate amount of variety to the food venues, with "Chinese Restaurant," "Fast Food Restaurant," "Asian Restaurant," and "Indian Restaurant" all appearing multiple times. Coffee Shop is also very common.

No gyms appear in this snapshot of the cluster and there are bar only appears as the 8th most common venue in neighborhood M1R.

	Cluster Labels	Postal Code	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	5	M1B	Fast Food Restaurant	Trail	Coffee Shop	Bank	Restaurant	Chinese Restaurant	Bakery	Paper / Office Supplies Store	Caribbean Restaurant	Greek Restaurant
4	5	M1H	Coffee Shop	Bakery	Gas Station	Bank	Indian Restaurant	Athletics & Sports	Chinese Restaurant	Thai Restaurant	Fried Chicken Joint	Grocery Store
6	5	M1K	Chinese Restaurant	Coffee Shop	Fast Food Restaurant	Discount Store	Grocery Store	Bank	Asian Restaurant	Light Rail Station	Sandwich Place	Pharmacy
10	5	M1P	Restaurant	Coffee Shop	Pharmacy	Electronics Store	Chinese Restaurant	Furniture / Home Store	Fast Food Restaurant	Bakery	Asian Restaurant	Indian Restaurant
11	5	M1R	Pizza Place	Middle Eastern Restaurant	Intersection	Grocery Store	Burger Joint	Furniture / Home Store	Restaurant	Bar	Coffee Shop	Korean Restaurant

6. Discussion

The k-means clustering algorithm resulted in three main clusters, each with their own unique attributes. Cluster 0 had a lot of standard American food. It was lacking nightlife and access to health studios. Cluster 3 contained a wide variety of food venues, a number of nightlife options, and a fair amount of gyms. Cluster 5 consisted of many Asian food options, limited nightlife, and no gyms or health studios.

Cluster 3 most closely aligns with the preferences of millennials, and therefore would be the recommended cluster to build the apartment building based off of these results. Of course, there are a number of other factors that Apartments20 would want to consider before deciding on a location, such as the price of real estate and competition from other buildings.

7. Conclusion

This report has demonstrated how clustering can be a useful tool in when trying to segment data points together based on similarities. In this case, I used K-means clustering to group neighborhoods of Toronto together. Then, I evaluated the attributes of those clusters to see how they lined up with the preferences of millennials, so that Apartments20 could find a profitable location to build a new apartment building.

Based on the results of the clustering algorithm, Apartments20 should consider choosing a neighborhood in cluster 3 to build their new apartment building because the venues of this cluster most closely align with the preferences of their customer base.

Works Cited:

1. Clinkard, John. “Canada's Millennials – Where Are They Now and Where Are They Moving? - Constructconnect.com.” *Daily Commercial News*, 25 Apr. 2019, canada.constructconnect.com/dcn/news/economic/2019/04/canadas-millennials-now-moving.