Coursera Capstone Project: The Battle of Neighborhoods

Leandro Leberon

June 15, 2021

Business Problem

After searching google for Indian Restaurants in Toronto, I found out that there are just a few of them. Given that are just a few Indian Restaurants in Toronto, it is a great idea to open one, but the great question is where to open it? In this project I will try to analyze Toronto's neighborhoods and find out which ones would be a great choice to open an Indian Restaurant.

Target Audience

The target audience for this project is people who are looking to open an Indian Restaurant in Toronto, developers who are interested in machine learning with python, and people living in Toronto and looking for Indian Restaurants.

Data Sources

For this project I used data from three different sources:

- Wikipedia: I used Wikipedia to web scrape a list of Toronto's boroughs, neighborhoods, and postal codes.
- Coursera: Coursera provided me a CSV that contains the geographical coordinates for each neighborhood in Toronto.
- Foursquare: I used Foursquare to obtain near venues for each neighborhood in Toronto.

Methodology

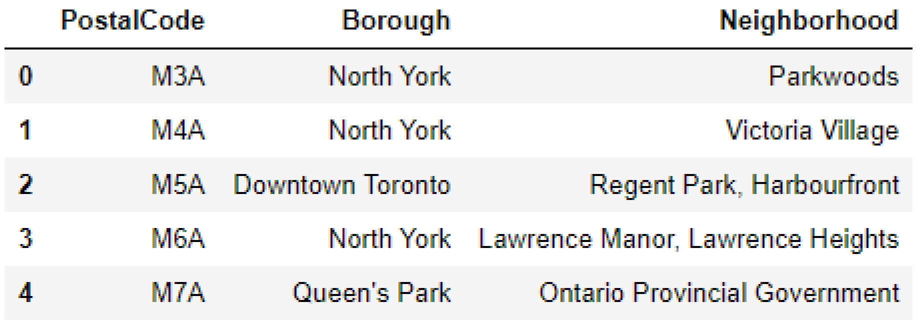
In this project I used data science techniques such as:

- Web scraping
- Data transforming
- K-mean clustering
- Data visualization
- Data cleaning

During the entire project, the following conditions were considered:

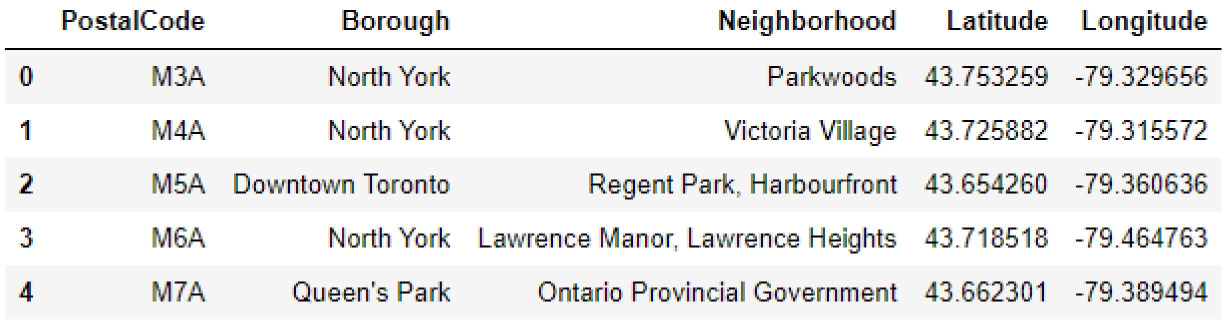
- Only process the cells that have an assigned borough. Ignore cells with a borough that is Not assigned.
- More than one neighborhood can exist in one postal code area. For example, in the table on the Wikipedia page, you will notice that M5A is listed twice and has two neighborhoods: Harbourfront and Regent Park. These two rows will be combined into one row with the neighborhoods separated with a comma as shown in row 11 in the above table.
- If a cell has a borough but a Not assigned neighborhood, then the neighborhood will be the same as the borough.

After installing all required libraries for the project, I started by web scraping Wikipedia to obtain Toronto's boroughs, neighborhoods, and postal codes. For this process I used Beautiful Soup, a great tool for web scraping. I created a Pandas data frame with the information I got.



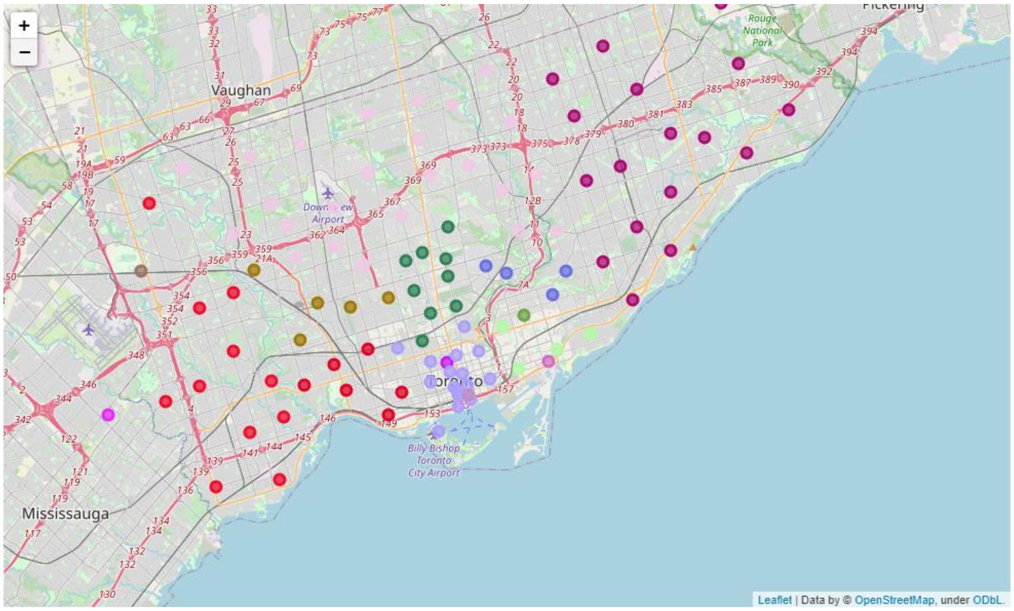
	PostalCode	Borough	Neighborhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Queen's Park	Ontario Provincial Government

To visualize this data in a map I needed the coordinates for each neighborhood, for which I used a CSV that Coursera provided. I created a separate data frame with the geocoordinates and the merged both data frames into one.

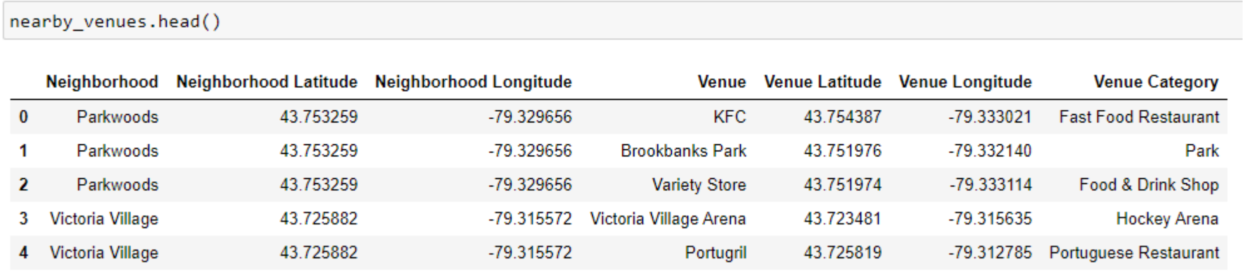


	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.753259	-79.329656
1	M4A	North York	Victoria Village	43.725882	-79.315572
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	M7A	Queen's Park	Ontario Provincial Government	43.662301	-79.389494

With this data frame it was possible for me to create a map with a marker for each neighborhood in Toronto. For this process I used Folium, a great tool for visualizing geo-spatial data.


+
−
Pickering
Rouge National Park
Vaughan
Downsview Airport
Toronto
Billy Bishop Toronto City Airport
Mississauga
Leaflet | Data by © OpenStreetMap, under ODbL.

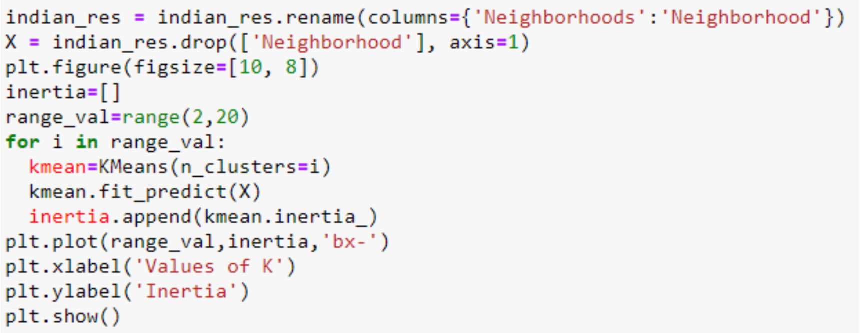
I then decided to use Foursquare's API to obtain a list of venues in Toronto. These venues include schools, restaurants, parks, shops, etc. After obtaining this information from Foursquare I created a Pandas data frame to store it.



```
nearby_venues.head()
```

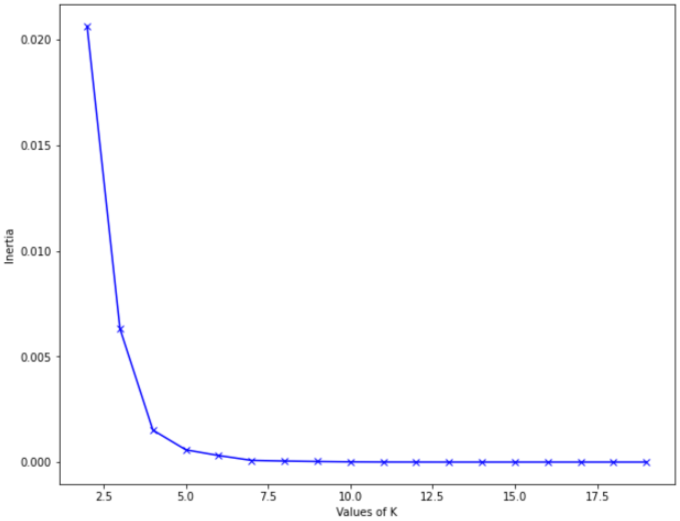
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Parkwoods	43.753259	-79.329656	KFC	43.754387	-79.333021	Fast Food Restaurant
1	Parkwoods	43.753259	-79.329656	Brookbanks Park	43.751976	-79.332140	Park
2	Parkwoods	43.753259	-79.329656	Variety Store	43.751974	-79.333114	Food & Drink Shop
3	Victoria Village	43.725882	-79.315572	Victoria Village Arena	43.723481	-79.315635	Hockey Arena
4	Victoria Village	43.725882	-79.315572	Portugril	43.725819	-79.312785	Portuguese Restaurant

Since I am only interested in Indian Restaurants, I used a process called K-mean clustering to classify the neighborhoods based on how many Indian Restaurants were present in each one of them. To obtain the K-value I decided to use the Elbow Method.

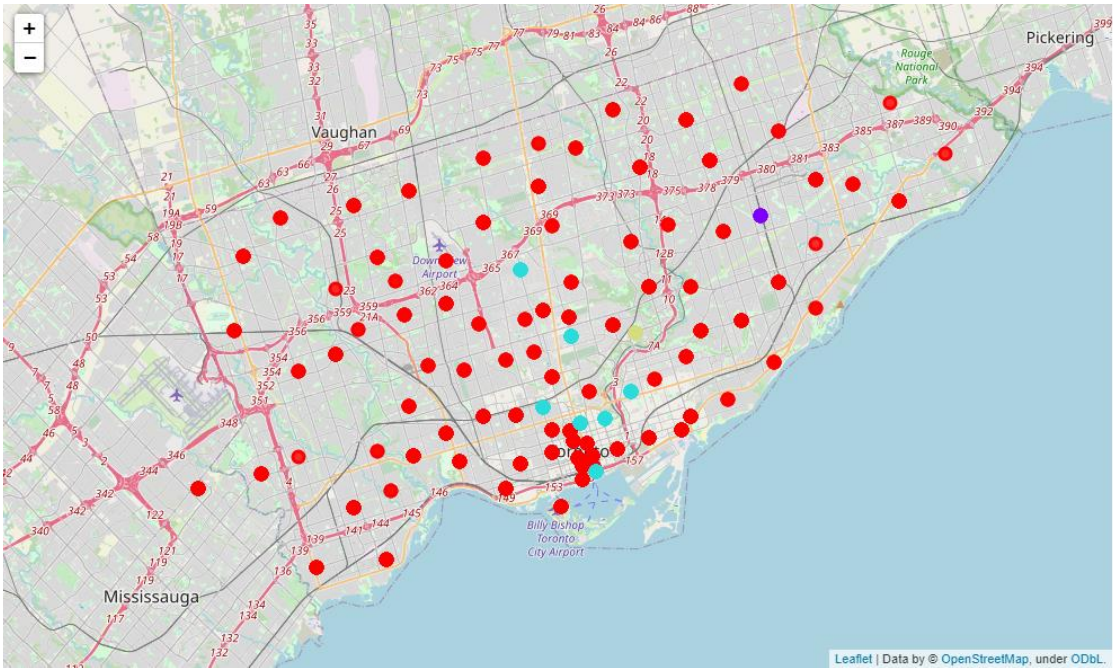


```
indian_res = indian_res.rename(columns={'Neighborhoods':'Neighborhood'})
X = indian_res.drop(['Neighborhood'], axis=1)
plt.figure(figsize=[10, 8])
inertia=[]
range_val=range(2,20)
for i in range_val:
  kmean=KMeans(n_clusters=i)
  kmean.fit_predict(X)
  inertia.append(kmean.inertia_)
plt.plot(range_val,inertia,'bx-')
plt.xlabel('Values of K')
plt.ylabel('Inertia')
plt.show()
```

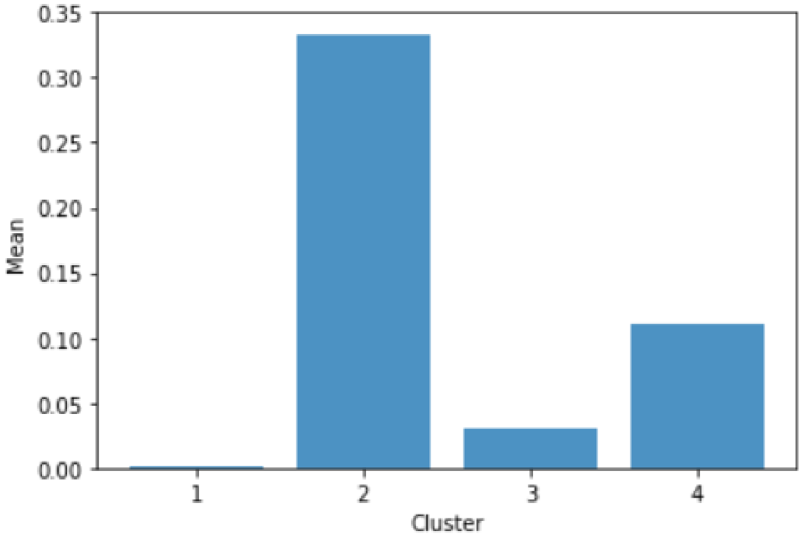
Which resulted in the following graph:



Based on the results, the best value of K is 4, this means that I should classify the data into 4 different clusters. After classifying each neighborhood into a cluster, I created a Folium map to visualize the results.


+
−
Vaughan
Pickering
Rouge National Park
Downsview Airport
Toronto
Billy Bishop Toronto City Airport
Mississauga
Leaflet | Data by © OpenStreetMap, under ODbL.

Finally, I created a Pandas data frame for each cluster based on how many Indian Restaurant are present. After doing this, I was able to visualize the average number of Indian Restaurants by clusters.



Results

We can see that cluster 2 has the highest number of Indian Restaurants, followed by cluster 4, cluster 3, and finally cluster 1. Based on my hypothesis I would safely assume that the best cluster to open an Indian Restaurant is cluster 1 since it contains the greatest number of neighborhoods and the smaller number of Indian Restaurants. Theoretically, an Indian Restaurant in cluster 1 would have less competition and more demand than any other cluster.

Discussion

Based on the results on this project I can say that Python is a great tool for manipulating data. It is very useful to visualize the results of this project, even if I never go to Toronto, there is a great chance that I will be using these skills in the future. I would recommend entrepreneurs to use this approach to identify potential locations to open any type of venue in any part of the world.

Conclusion

I really enjoyed this course and learned a lot. Even though this is no bachelor’s degree, I can demonstrate to any employer the skills I got from this course, which at the end of the day is what matters the most. I hope my project helps any other self-proclaimed Data Scientists :). glhf.