Opening a new Pizza Place in Toronto

Gurjyot Singh

July 26, 2019

1. Introduction

1.1.Background

Toronto is a big city and abundant with people. It is the most populated city in Canada. Food places in Canada are famous for having a variety of cuisines. The most popular type of food places are Pizza Places. Pizza is loved by everyone and is affordable too. There are more than 1500 pizzerias in the city of Toronto. To start a new pizza place is a difficult task and requires some analysis of the places first. The main obstacle is predicting if a new pizza place will be popular or not. To do so, one needs to do some analysis of neighbourhoods and the type of food places there.

1.2. Problem

Data used in the analysis will contain different neighbourhoods, the types of restaurants there and the average number of pedestrians moving from that place. The project aims to finding a suitable place in Toronto where there are few pizza places nearby and a good amount of pedestrians pass by.

1.3. Interests

People looking to open a new Pizza Place in the city of Toronto will be very much interested in this analysis as it will help them pin point specific locations where they have promising future of a Pizza Place. This project can be modified to work for any kind of business, so anyone looking to open a new business would be very much interested in this project.

2. Data acquisition and cleaning

2.1.Data Sources

There is no single source of data which could give the types of restaurants and the number of people passing by from locations. We will need to combine data from different sources. The neighbourhoods in Toronto can be scraped by Wikipedia page. We will fetch the restaurants in a neighbourhood using Foursquare API. The dataset for volume of pedestrians passing by at various intersections in Toronto is made available on the website of Toronto. The pedestrian volume data however was last updated in 2018, being not so old, it should work fine.

2.2.Data Cleaning

The data was downloaded from website of Toronto. The data was very clean and didn't require much processing. The dataset contained coordinates of intersections in the city of Toronto and people passing by those intersections each day. I used Nominatim geocoder from geopy.geocoder. Nominatim geocoder has a method called reverse() used for reverse geocoding coordinates. I found the neighbourhoods of coordinates using this method. There were some locations for which the package couldn't determine neighbourhood, so I used the city district for those locations instead. After doing so, I noticed there were still some locations with no value in the neighbourhood column. To fill in those missing values, I used the main street name on the intersection.

Next, I grouped the dataframe by neighbourhoods and used average of all the coordinates of intersections in the neighbourhood. I summed up all the pedestrians crossing the intersections in neighbourhood. By doing so, a dataframe with neighbourhoods, pedestrian volumes, and coordinates of the neighbourhoods was achieved.

This dataframe was then fed into foursquare to fetch all the venues in the neighbourhood. Then the venues were filtered out to get only the type of potential competitors with our said Pizza Place. The competitors include burger joints, taco places, and many more fast food joints.

Next, the competitors were counted according to the neighbourhoods and were joined with dataframe containing pedestrian volumes, giving us the final dataframe used for recommendation.

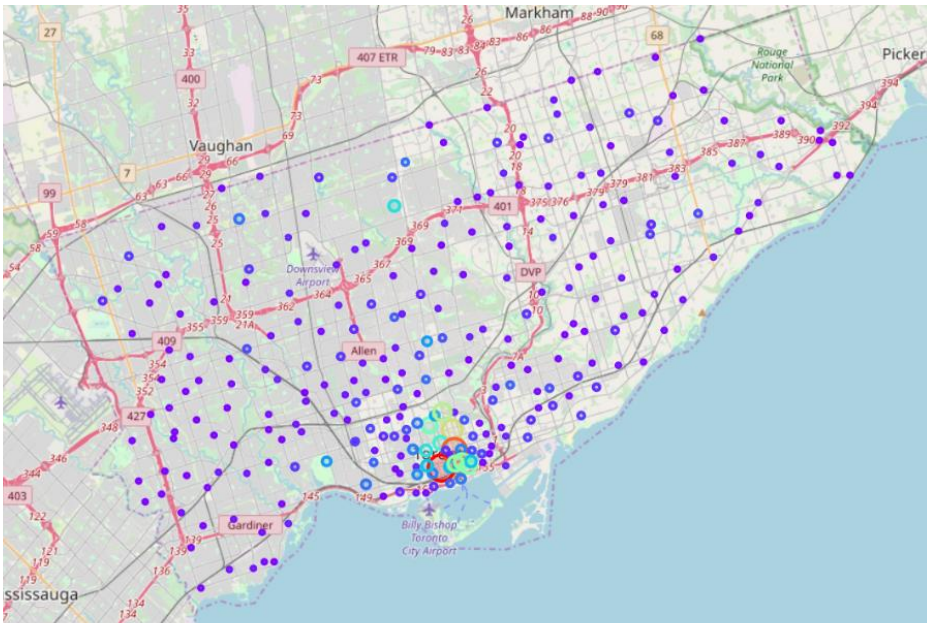
2.3.Feature Selection

The dataset didn't contain much of redundant features, but we still had to drop features like TCS#, Main, Midblock Route, Side 1 Route, Side 2 Route, Activation Date, Count Date and 8 Peak Hr Vehicle Volume simply because they weren't relevant to our problem statement.

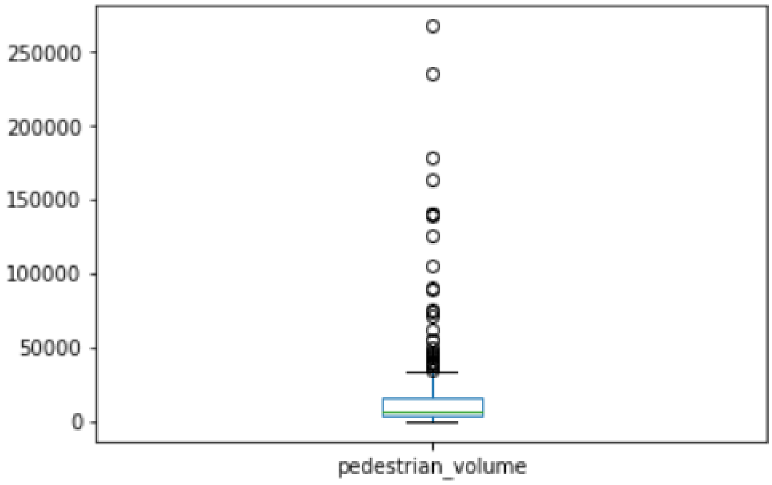
The features selected in final dataset were latitude, longitude, 8 Peak Hr Pedestrian Volume. Rest of the features in final dataframe were neighbourhood name and number of competitors which were calculated in the project itself.

3. Methodology

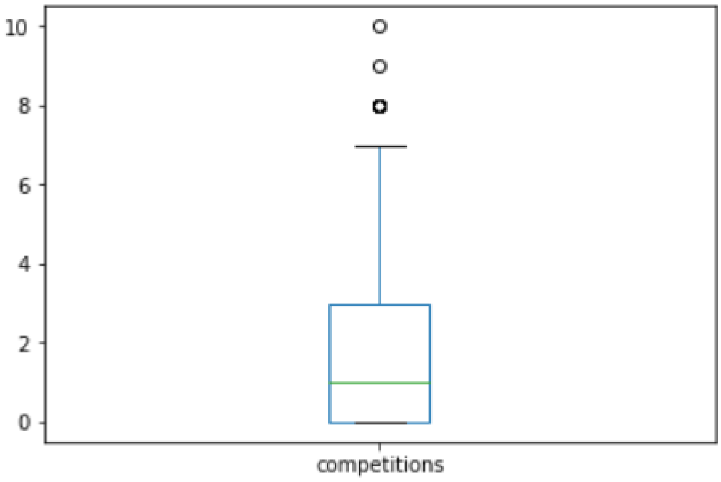
After collecting the data on neighbourhoods and pedestrian volumes, I did a geospatial mapping of the neighbourhoods along with the number of pedestrians passing. Red to violet represents most number of pedestrian to least. Same is with the size of markers.



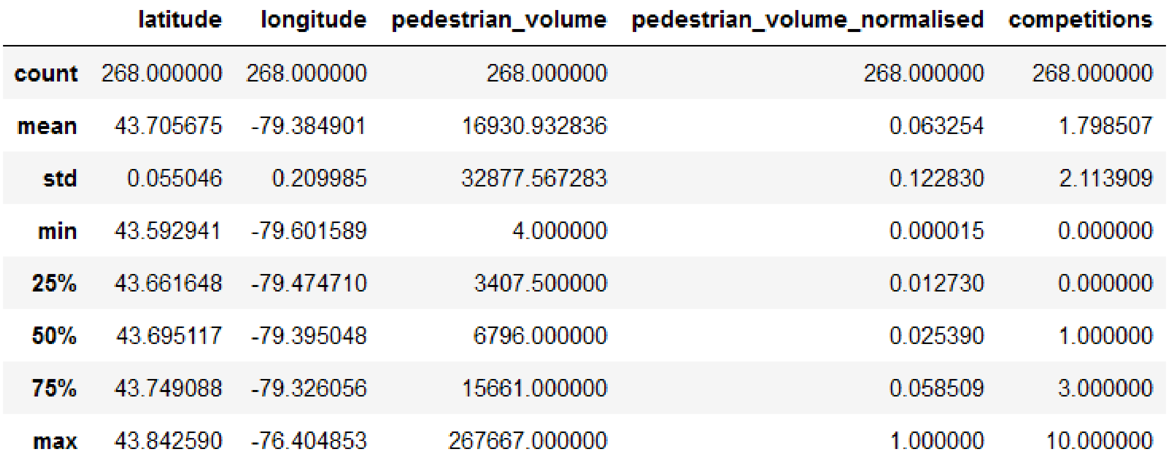
The analysis on data was done during cleaning and gathering data side by side. I had to check the data on almost every step for its consistency. After collecting and cleaning the full data, I used box plots to understand the skewness of data. It helped me identify outliers and the interquartile ranges. Outliers couldn’t be ignored, but I couldn’t remove the outliers because of the fact that these were the most popular places and if they were to be cast out, it needed to be from the criteria I set in recommendation and not before that.



The box plot for number of competitors in neighbourhood gave me an idea to set the limits for choosing number of minimum competitors. I chose the threshold at 75% of data, which gave me 3 competitors or less. Similarly, then I chose the threshold for pedestrian volume as 75% of data, which gave me as 15,661 or more. I chose to set the limit at 15,000.



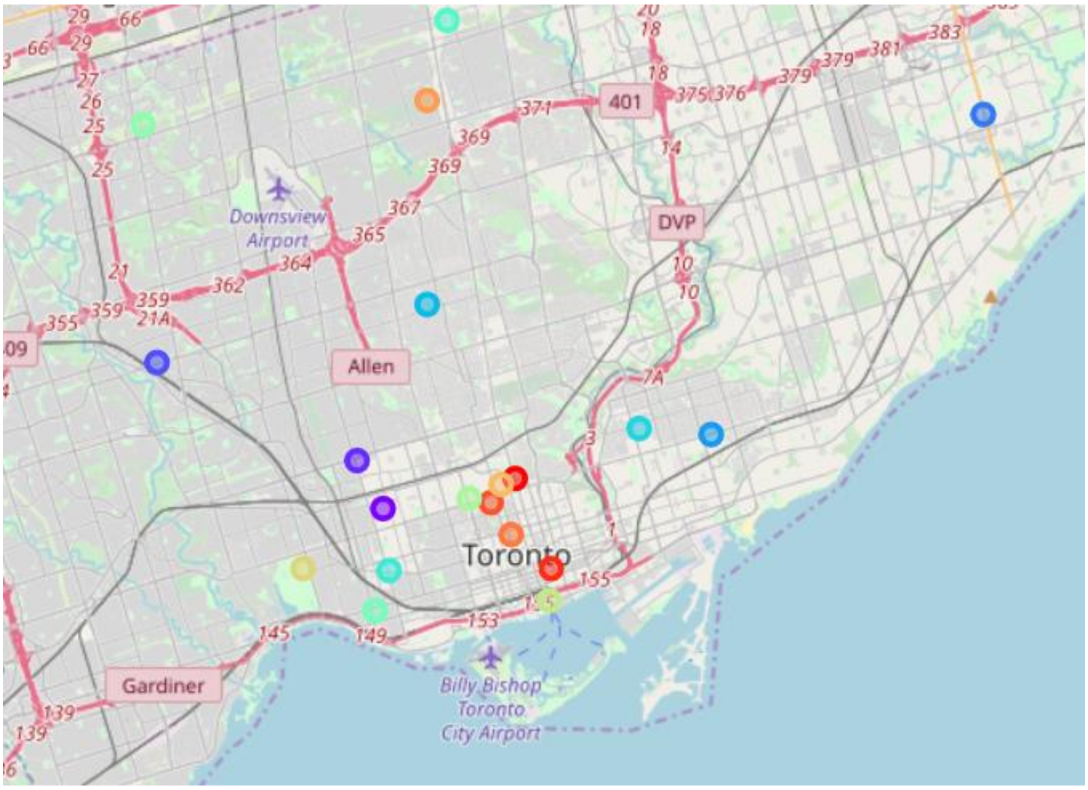
The describe method of dataframe gave me a very accurate insight into the data. The final values of limits were set after consideration with the description of dataframe.



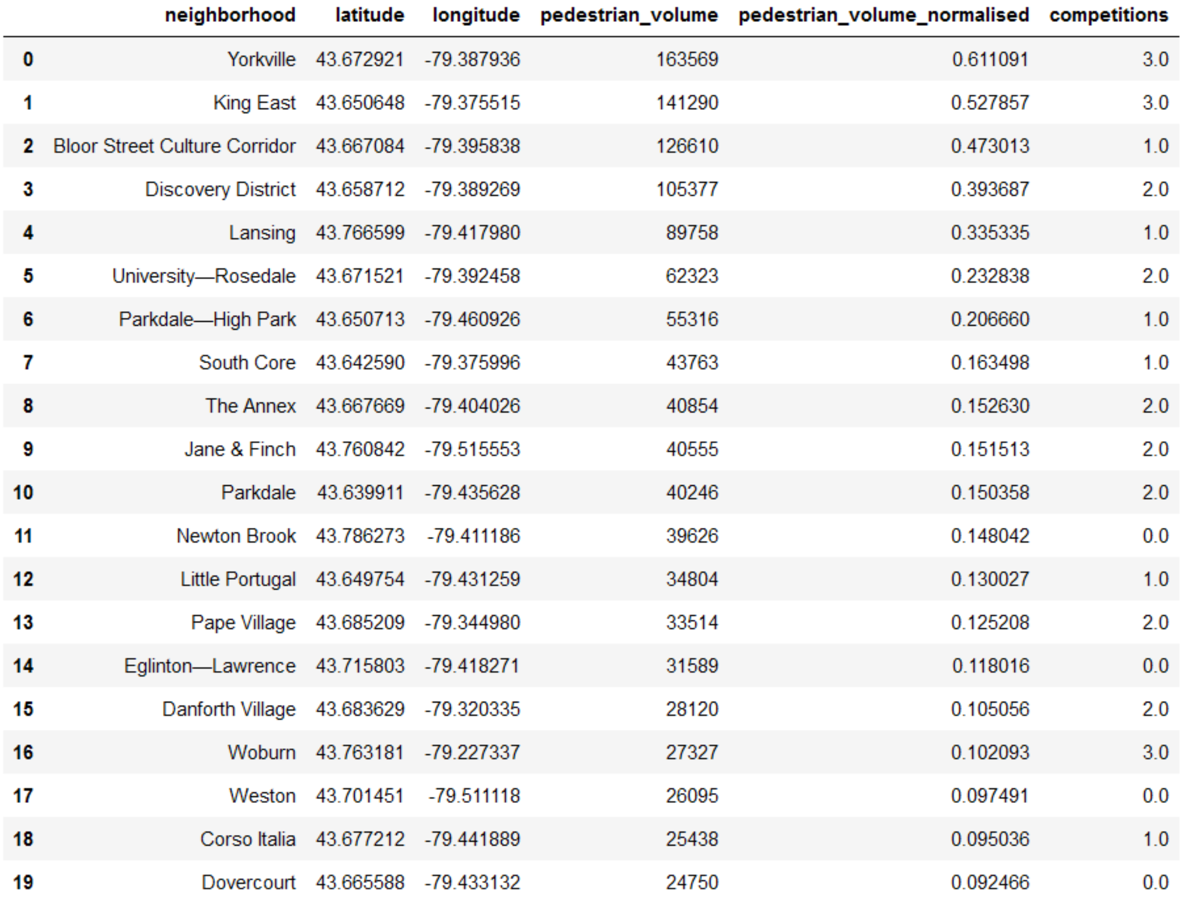
	latitude	longitude	pedestrian_volume	pedestrian_volume_normalised	competitions
count	268.000000	268.000000	268.000000	268.000000	268.000000
mean	43.705675	-79.384901	16930.932836	0.063254	1.798507
std	0.055046	0.209985	32877.567283	0.122830	2.113909
min	43.592941	-79.601589	4.000000	0.000015	0.000000
25%	43.661648	-79.474710	3407.500000	0.012730	0.000000
50%	43.695117	-79.395048	6796.000000	0.025390	1.000000
75%	43.749088	-79.326056	15661.000000	0.058509	3.000000
max	43.842590	-76.404853	267667.000000	1.000000	10.000000

4. Results

The final recommendation consisted of 20 such places with less than 3 competitors and more than 15,000 pedestrians passing by every day. This recommendation will allow the new businessmen to jump start their business. The recommendation was made by considering the fact that target business is a Pizza Place. The project can be easily morphed to be working for any kind of business. Here are the top 20 recommendations, Red being most recommended and violet being least one.



The result set contains all the necessary information for the client including neighbourhood name, latitude and longitude, number of pedestrians and the number of competitors.



	neighborhood	latitude	longitude	pedestrian_volume	pedestrian_volume_normalised	competitions
0	Yorkville	43.672921	-79.387936	163569	0.611091	3.0
1	King East	43.650648	-79.375515	141290	0.527857	3.0
2	Bloor Street Culture Corridor	43.667084	-79.395838	126610	0.473013	1.0
3	Discovery District	43.658712	-79.389269	105377	0.393687	2.0
4	Lansing	43.766599	-79.417980	89758	0.335335	1.0
5	University—Rosedale	43.671521	-79.392458	62323	0.232838	2.0
6	Parkdale—High Park	43.650713	-79.460926	55316	0.206660	1.0
7	South Core	43.642590	-79.375996	43763	0.163498	1.0
8	The Annex	43.667669	-79.404026	40854	0.152630	2.0
9	Jane & Finch	43.760842	-79.515553	40555	0.151513	2.0
10	Parkdale	43.639911	-79.435628	40246	0.150358	2.0
11	Newton Brook	43.786273	-79.411186	39626	0.148042	0.0
12	Little Portugal	43.649754	-79.431259	34804	0.130027	1.0
13	Pape Village	43.685209	-79.344980	33514	0.125208	2.0
14	Eglinton—Lawrence	43.715803	-79.418271	31589	0.118016	0.0
15	Danforth Village	43.683629	-79.320335	28120	0.105056	2.0
16	Woburn	43.763181	-79.227337	27327	0.102093	3.0
17	Weston	43.701451	-79.511118	26095	0.097491	0.0
18	Corso Italia	43.677212	-79.441889	25438	0.095036	1.0
19	Dovercourt	43.665588	-79.433132	24750	0.092466	0.0

It is recommended that a place within 400-500 meters of the coordinates would hold true for the business recommendation.

5. Discussion

The project was aimed at opening a new Pizza Place in the city of Toronto, Canada. But the process for recommending a place for any other kind of business in Toronto is the same. The project can be modified to work for any kind of business in Toronto. For other cities, if the data of pedestrian volumes can be determined, the same project will work.

The project is not aimed on at Pizza Places in Toronto, but can be aimed at any kind of business in any city in to world provided correct and updated data can be found.

I observed the recommendation is slightly bias to recommend in Downtown Toronto. This is because of the fact once there is an area in Downtown Toronto with a few competitors, it will be most recommended due to the fact that there is the most number of pedestrians in Downtown. But opening a new Pizza Place Uptown, where there is not much competition can also become a success. This issue can probably be resolved by applying some more analysis techniques that I am currently unaware of.

6. Conclusion

Top 20 places in Toronto for opening a Pizza Place were recommended to the client. The client is recommended to look within 400-500 meters of the given coordinates.

Concluding this project, I can say that I have learnt a lot in the Data Science Professional Certificate offered by IBM on Coursera. I am very much confident in Data Science now than ever.