Clustering solution for recommendation to new residents

1. Introduction

Let's assume that a person needs a recommendation about moving to a new city in the province of Ontario. Multiple constraints are in play, from job availability to school rating for children availability of other type leisure activity and important facilities. Foursquare location data will provide the required in terms of location, ratings, and available facilities. This information is extremely important for someone to move especially if he is new to the area. A comparison with multiple cities will be added for bench-marking purposes.

Essentially, we will be using data from different sources, this includes for example employment status for different neighborhood, available facilities and their rating (available from foursquare). We can also add safety features such as crime rate in a specific area and education quality.

2. Data acquisition and cleaning

2.1 Data sources

The data will be acquired from two sources. The first being the data set of crime rate by Neighbourhood as csv file (Neighbourhood_Crime_Rates_(Boundary_File)_.csv) from the toronto police service website (https://data.torontopolice.on.ca/). The second one is the API of Foursquare developer portal.

2.2 Data cleaning

Data are downloaded or scraped from the sources mentioned above. There were a lot of data that needs to be removed as it is out of the scope of this project that needs to be dropped. For example, instead handling each type of crime apart, I took only the average per neighborhood since it provide enough information about the overall rate.

3. Exploratory Data Analysis

Here we are trying to provide analysis about the information we obtained after cleaning the data. We start by making sure we having the correct neighborhoods in the area as seen in Fig.1

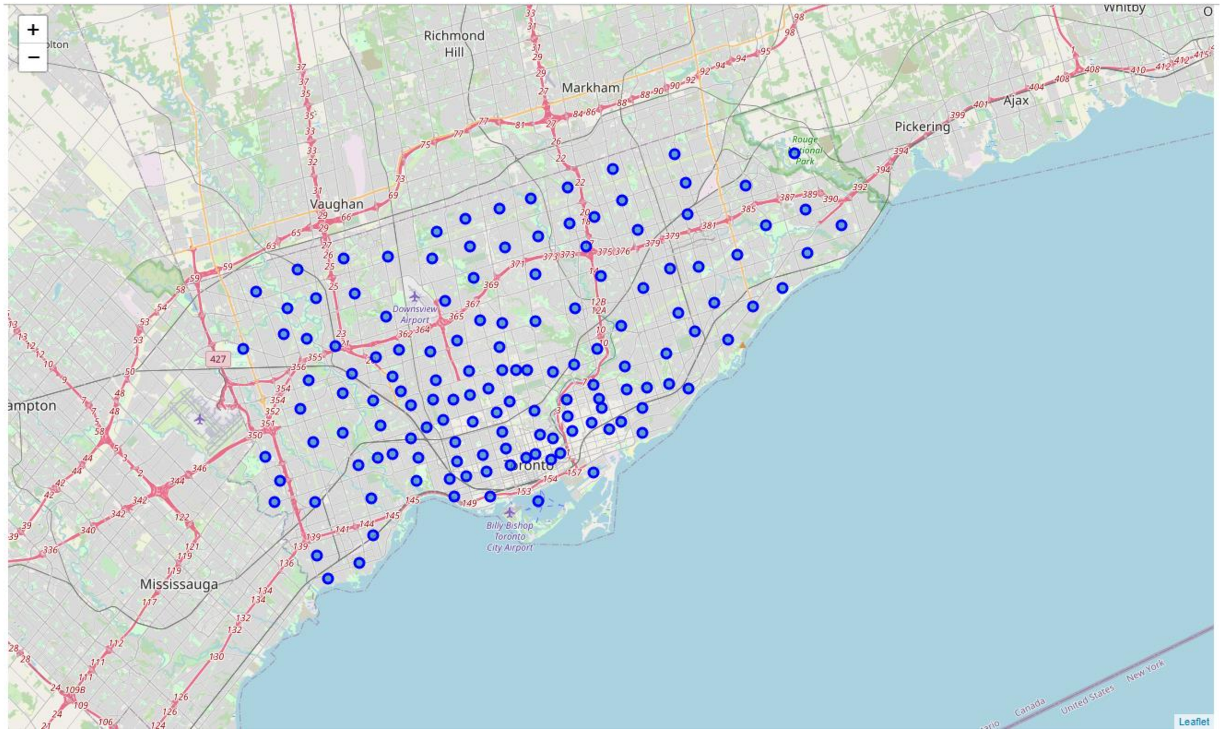

Figure 1. distribution of the neighborhoods in the Toronto area

Next, we grab all the available venues and attractions from the Foursquare website, and we matched with total crime rate. We will also merge both tables to try account for both attractions and crime rate at each neighborhood. We will after that to understand the correlation between different attraction and the crime rate. As shown in Fig. 2, where we see that there is definitely some correlations between the features. Lets focus on crime as it seems here that mostly it has low correlation in general but this because we are treating the average of all types of crimes and the correlation can be lost after the averaging procedure. By zooming, we find that there are attractions that have relatively high correlation with the crime rate as shown in Fig. 3. Lets here to get the correlation between the specific column of crime rate and the remaining venues types. It is seen here that multiple venues have a correlation rate higher than 0.2.

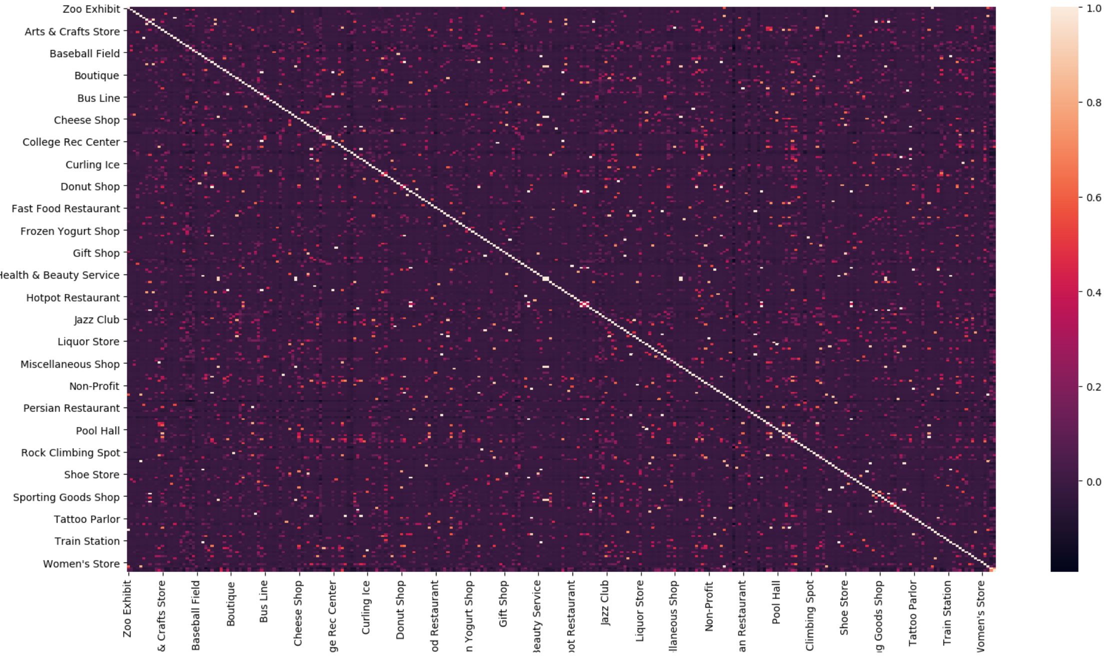

Figure 2. Correlation heatmap

Figure 3. Heatmap after zoom in

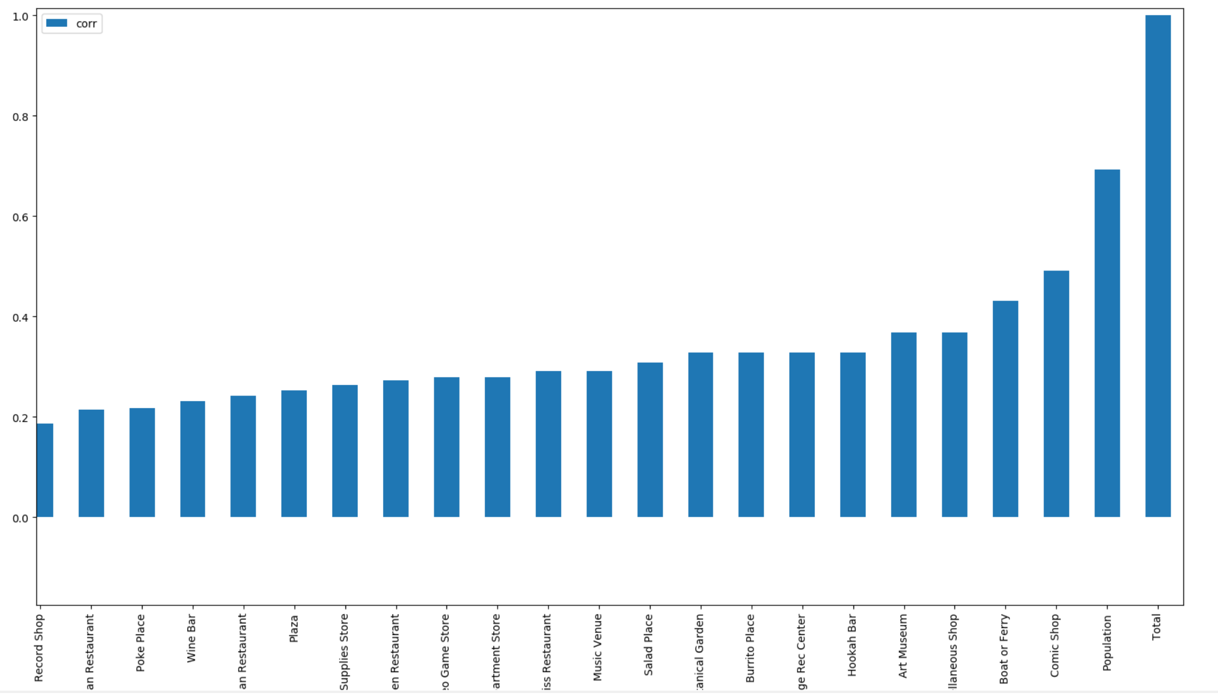

Figure 4 correlation with crime rate

4. Predictive Modeling

Here we can use an unsupervised clustering model to see how the crime rate is associated with the most visited venues. We use a k-mean clustering model to see where the venues are more related with crimes in specific neighborhood. With 10 clusters. Next, we plot the cluster on the map.

We obtain the following result:

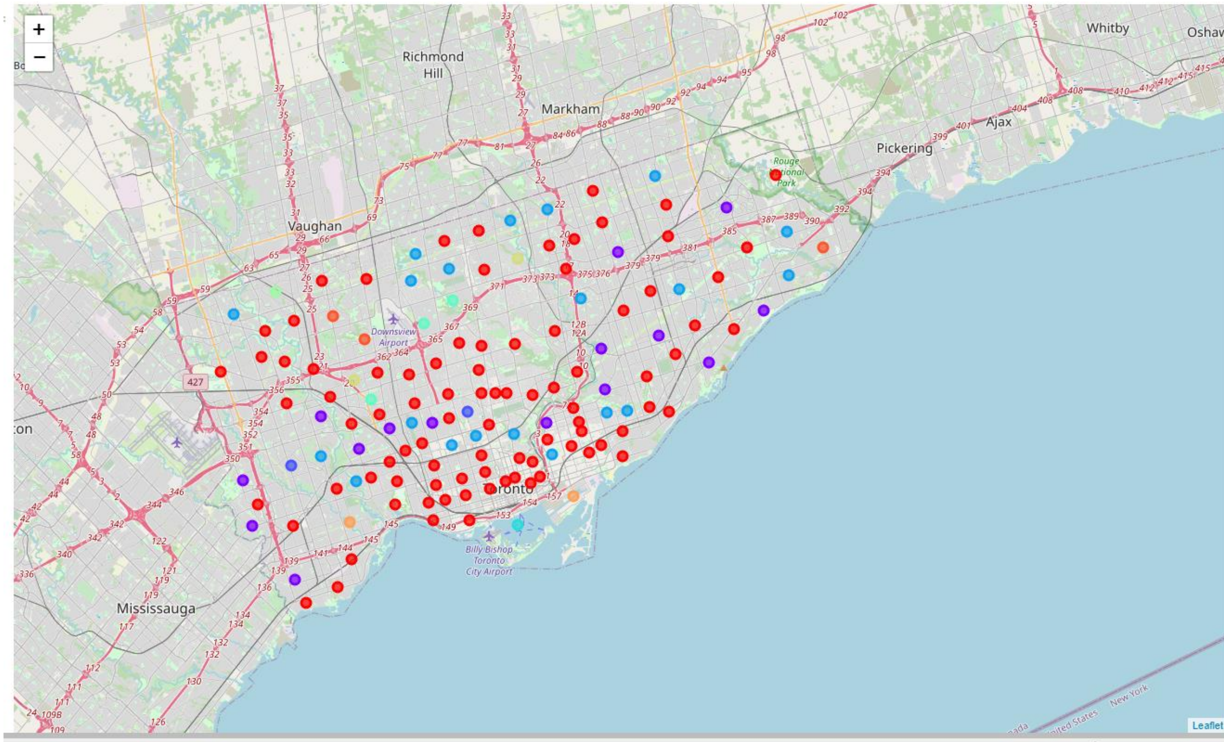

Figure 5. clustering method

The algorithm tries to provide clustering while accounting for crime rate. By taking a closer look at the cluster we have the following tables:

```
dfinalShrinked.loc[dfinalShrinked['Cluster Labels'] == 0, dfinalShrinked.columns[[1] + list(range(5, dfinalShrinked.shape[1]))]]
```

	LATITUDE	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	43.687859	Coffee Shop	Italian Restaurant	Sushi Restaurant	Pizza Place	Thai Restaurant	Pub	Bank	Café	Pharmacy
1	43.765736	Massage Studio	Japanese Restaurant	Pizza Place	Coffee Shop	Caribbean Restaurant	Fast Food Restaurant	Metro Station	Sushi Restaurant	Furniture / Home Store
3	43.714672	Fast Food Restaurant	Restaurant	Greek Restaurant	Bowling Alley	Seafood Restaurant	Café	Bookstore	Fried Chicken Joint	Sandwich Place
6	43.671050	Japanese Restaurant	Coffee Shop	Park	Pub	Bar	Thai Restaurant	Bakery	BBQ Joint	Pizza Place
7	43.737988	Indian Restaurant	Caribbean Restaurant	Pizza Place	ATM	Pharmacy	Coffee Shop	Supermarket	Thai Restaurant	Bank
...	...	...	...	...	...	...	...	...	...	...
133	43.676773	Bar	Baseball Field	Grocery Store	Café	Total	Ethiopian Restaurant	Egyptian Restaurant	Electronics Store	Elementary School
134	43.791536	Bridal Shop	Café	Total	History Museum	Dog Run	Field	Fast Food Restaurant	Farmers Market	Farm
136	43.786982	Breakfast Spot	Japanese Restaurant	Convenience Store	Pharmacy	Park	Restaurant	Shopping Mall	Skating Rink	Burger Joint
138	43.703797	Coffee Shop	Brewery	Shopping Mall	Bike Shop	Fish & Chips Shop	Sporting Goods Shop	Burger Joint	Skating Rink	Mexican Restaurant
139	43.699024	Coffee Shop	Grocery Store	Hostel	Bank	Antique Shop	Pharmacy	Argentinian Restaurant	Discount Store	Bike Shop

89 rows × 10 columns

Figure 6. cluster referring to red dots

Here we observe no high link between the crime rate and the available venues. However, the orange dots, seems to have some kind relation between the crime rate with essential venues like playground and event spaces:

```
[182]: dfinalShrinked.loc[dfinalShrinked['Cluster Labels'] == 2, dfinalShrinked.columns[[1] + list(range(5, dfinalShrinked.shape[1]))]]
```

[182]:

	LATITUDE	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
30	43.666051	Playground	Construction & Landscaping	Total	American Restaurant	Falafel Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Elementary School
54	43.694526	Playground	Total	African Restaurant	Event Space	Dumpling Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Elementary School

Figure 7. second cluster related to orange dots

The third class seems to feature more crime rate related to the available venues as the total rate of crimes comes in second and third place.

	LATITUDE	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
18	43.694998	Theater	Park	Total	American Restaurant	Event Space	Dumpling Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store
31	43.671995	Café	Park	Dog Run	Pool	Total	Animal Shelter	Donut Shop	Fish & Chips Shop	Filipino Restaurant
40	43.767490	Park	Construction & Landscaping	Gym / Fitness Center	Total	Amphitheater	Falafel Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store
42	43.778813	American Restaurant	Gym / Fitness Center	Park	Total	Amphitheater	Falafel Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store
45	43.796802	Basketball Court	Dog Run	Park	Total	Amphitheater	Falafel Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store
50	43.681852	Park	History Museum	Bistro	Historic Site	Modern European Restaurant	Museum	Castle	Steakhouse	Total
57	43.688569	Women's Store	Park	Total	African Restaurant	Event Space	Dumpling Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store
61	43.802988	Residential Building (Apartment / Condo)	Park	Total	American Restaurant	Event Space	Dumpling Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store
64	43.670886	Park	Indian Restaurant	Fast Food Restaurant	Total	Amphitheater	Falafel Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store
71	43.790775	Home Service	Park	Bus Station	Construction & Landscaping	Total	Wine Shop	Elementary School	Dog Run	Donut Shop
74	43.682820	Playground	Tennis Court	Park	Candy Store	Total	Electronics Store	Donut Shop	Dumpling Restaurant	Eastern European Restaurant
85	43.764813	Baseball Field	Park	Playground	Convenience Store	Total	Animal Shelter	Donut Shop	Fish & Chips Shop	Filipino Restaurant
88	43.760366	Park	Convenience Store	Greek Restaurant	Total	Amphitheater	Falafel Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store
92	43.657420	Playground	Garden	Park	River	Total	Electronics Store	Donut Shop	Dumpling Restaurant	Eastern European Restaurant
96	43.771210	Park	Mobile Phone Shop	Total	American Restaurant	Falafel Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store	Elementary School
98	43.746868	Coffee Shop	Japanese Restaurant	Park	Total	Amphitheater	Falafel Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store
107	43.755033	Construction & Landscaping	Food & Drink Shop	Park	Total	Amphitheater	Falafel Restaurant	Eastern European Restaurant	Egyptian Restaurant	Electronics Store

Figure 8. third class related to the blue dots

5. Conclusion:

In this study, I analyzed the relationship between crime rate and the available venues per neighborhood to make a decision about the most suited neighborhood for new people that wants to settle in the region of Toronto. K-means algorithm is able to spot the place where there is little correlation with certain public places with crime rate while in other neighborhood crime rate seems to have high correlation with specific venues

6. Future directions

It is possible to increase the accuracy of the algorithm by investigating each type of crime separately. In such case we will have a higher correlation with certain type crime and certain types venues. As it is plausible for example to have higher robbery crimes near ATMs for example.