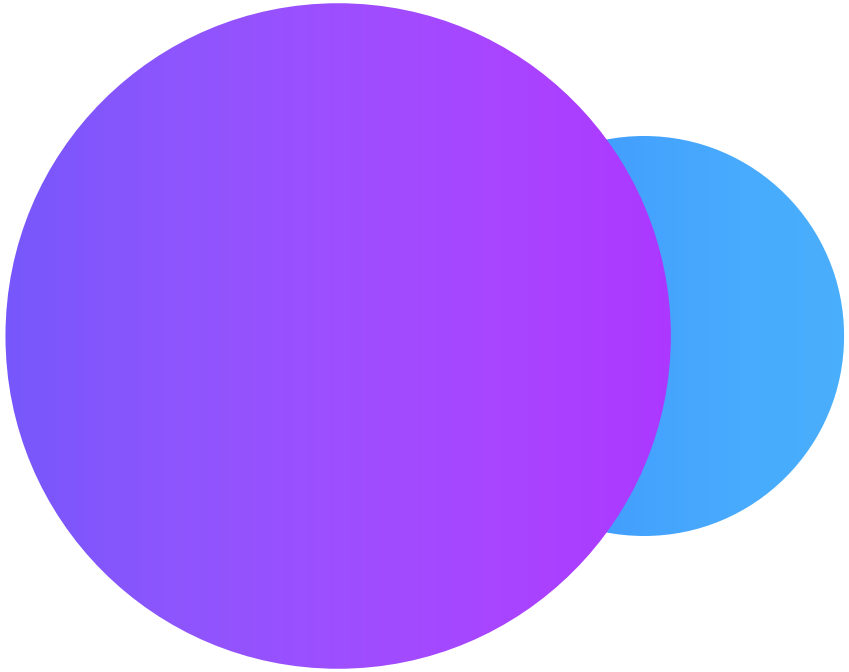
2020

IBM Applied Data Science Capstone Project

Segmenting and Clustering Venues in Toronto

Daisy Z.

Table of Content

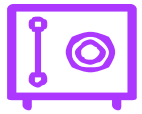


PART / 01

Project Background

Project Background

Potential audience



potential real estate investors

potential renters

potential real estate buyers

instructors and peer graders

In the city of Toronto, someone would like to open a new restaurant.
We are going to pick a best location for this business purpose.
Many factors affect popularity of a restaurant, such as taste, location, special menu etc.
Here, we are going to focus on location first.

Data Preparation

Geospatial Coordinate

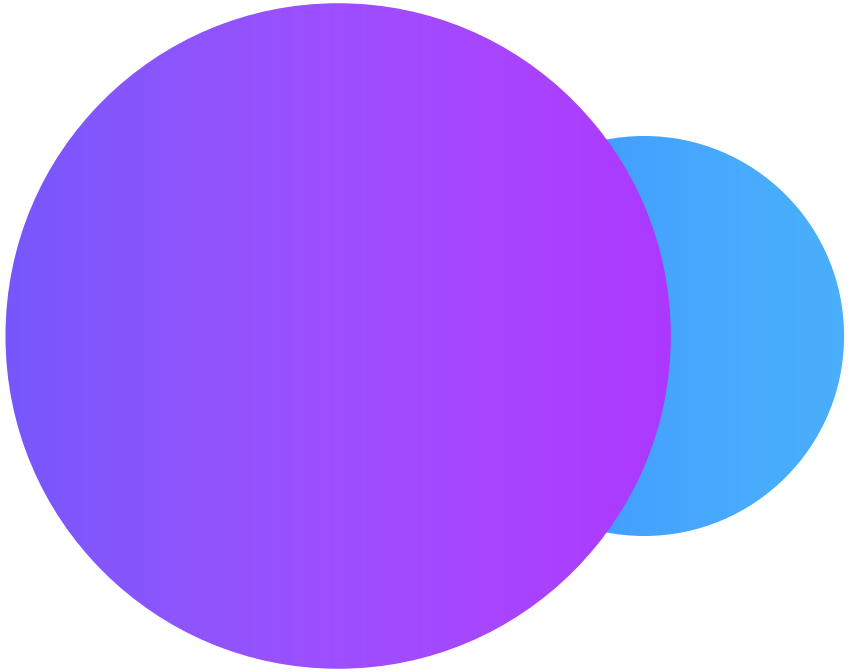
Other relational data: a csv file containing geospatial coordinate data. We can join coordinate data with venue data.

Foursquare API

The Foursquare API: geographical data with related information will be access via Python scraping to get most venues for each neighborhood in the city of Toronto. By doing so, we can visualize geographical data on the map and clearly see how venues are distributed in neighbors.

Data Wrangling

We need to check missing values, duplicate values, and to select useful columns that we are going to use for building models.



PART / 02

Methodology

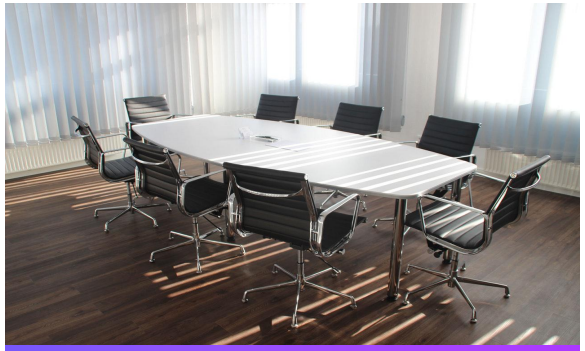
Methodology

EDA

Exploratory data analysis to see frequency of venues by neighbor

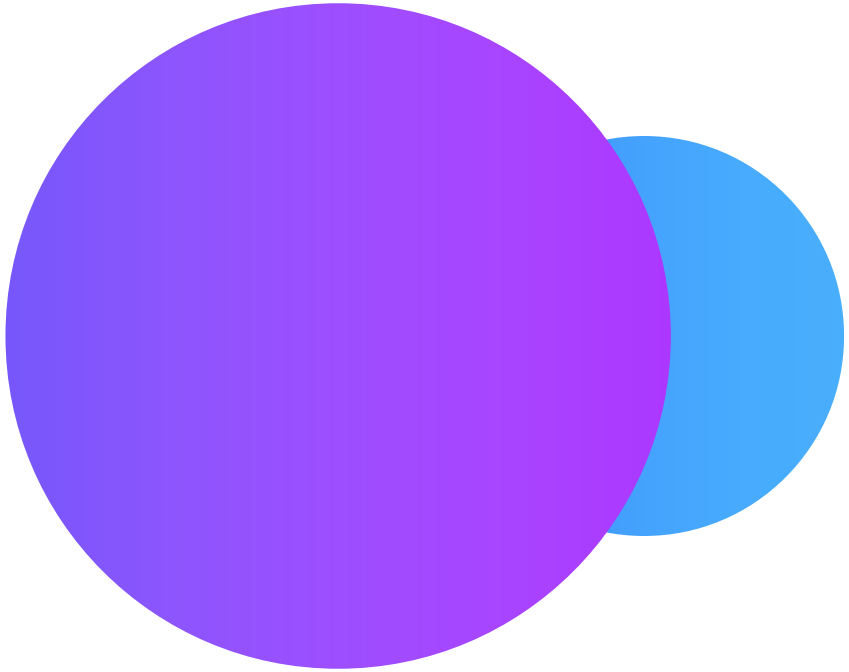
One-hot coding

feature engineering



K-means & Elbow method

K-means is a simple and quick way to do so. And we have created 15 features, including 10 most common venues, postcode, geospatial coordinates, borough etc.



PART / 03

Modeling

Modeling

	Borough	Clusters	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	North York	0	Intersection	Coffee Shop	Pizza Place	Hockey Arena	Portuguese Restaurant
8	East York	0	Pizza Place	Fast Food Restaurant	Gastropub	Café	Athletics & Sports
10	North York	0	Park	Pub	Pizza Place	Japanese Restaurant	Distribution Center
50	North York	0	Empanada Restaurant	Pizza Place	Dog Run	Department Store	Dessert Shop
63	York	0	Pizza Place	Bus Line	Caribbean Restaurant	Brewery	Women's Store
70	Etobicoke	0	Pizza Place	Middle Eastern Restaurant	Chinese Restaurant	Coffee Shop	Discount Store
77	Etobicoke	0	Park	Bus Line	Pizza Place	Sandwich Place	Discount Store
93	Etobicoke	0	Pizza Place	Coffee Shop	Pharmacy	Sandwich Place	Skating Rink

	Borough	Clusters	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
2	Downtown Toronto	4	Coffee Shop	Park	Pub	Bakery	Theater
3	North York	4	Clothing Store	Furniture / Home Store	Women's Store	Coffee Shop	Miscellaneous Shop
4	Downtown Toronto	4	Coffee Shop	Yoga Studio	Distribution Center	Mexican Restaurant	Bank
7	North York	4	Café	Baseball Field	Gym / Fitness Center	Caribbean Restaurant	Japanese Restaurant
9	Downtown Toronto	4	Clothing Store	Coffee Shop	Café	Cosmetics Shop	Japanese Restaurant
13	North York	4	Beer Store	Gym	Restaurant	Coffee Shop	Italian Restaurant
14	East York	4	Skating Rink	Dance Studio	Spa	Diner	Curling Ice

	Borough	Clusters	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	North York	6	Park	Bus Stop	Food & Drink Shop	Distribution Center	Dessert Shop
21	York	6	Park	Market	Women's Store	Gluten-free Restaurant	Gift Shop
35	East York	6	Park	Convenience Store	Coffee Shop	Dessert Shop	Dim Sum Restaurant
40	North York	6	Park	Airport	Doner Restaurant	Dessert Shop	Dim Sum Restaurant
49	North York	6	Park	Bakery	Construction & Landscaping	Doner Restaurant	Dim Sum Restaurant
64	York	6	Park	Convenience Store	Empanada Restaurant	Electronics Store	Eastern European Restaurant
66	North York	6	Park	Bank	Convenience Store	Bar	Women's Store
85	Scarborough	6	Park	Playground	Doner Restaurant	Dessert Shop	Dim Sum Restaurant
91	Downtown Toronto	6	Park	Playground	Trail	Eastern European Restaurant	Dumpling Restaurant

```
: Clusters
0       8
1       1
2       1
3       1
4      74
5       1
6      11
7       1
Name: Neighbourhood, dtype: int64
```

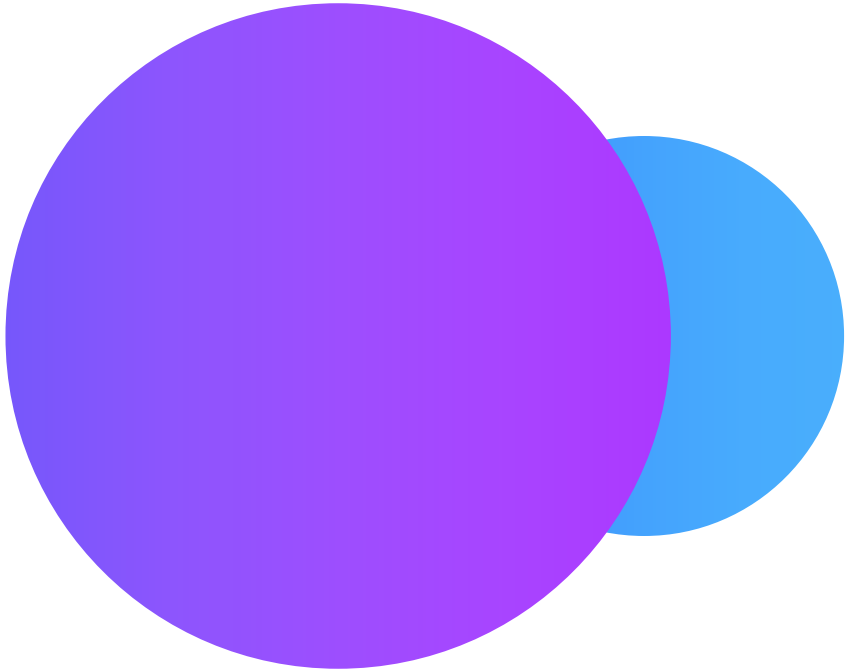
We can see from tables above that clusters being 0, 4 and 6 have number of venues much more than others. Then we begin to explore three clusters deeply to see which cluster of neighbor are we going to choose.

By interpreting results of each cluster, we summarized features and named each cluster

Cluster 0 -- fast food, coffee shop, restaurant etc

Cluster 4 -- convenient store, coffee shop, area that suitable for women

Cluster 6 -- park area, transportation area, and area that suitable for sports people



PART / 04

Result and Conclusion

Result

As the result, we decide to put the location in the area of cluster 0, where lots of restaurant, coffee shop around here, which means that flow of people are huge around here.

However, other cluster having only 1 venue does not mean in reality there is only one venue located around that area. We have a very bias result here.

Furthermore, when choosing location of a new venue, especially restaurant, we not only consider the number of people around the area, but also their power of consumption, which means their income level. If we open three Michelin star restaurant surrounding by fast food restaurant, we can simply conclude that people who eat fast food daily wouldn't have buying power of dining at high-end restaurant. So for further exploration, we can add income level of local people, marital status, and even more stats about people so that we can draw more precise and sound conclusion.

THANK YOU