Cognitive Class.ai Cousera

CAPSTONE

Report

Clustering of Neighbourhoods in Toronto to Aid Selection of Residence

Author :

Oladayo K. Ogunnoiki

Reviewed by :

Peer Reviewed

April 29, 2020

Contents

1 Introduction 2
1.1 Background . . . 2
1.2 Problem Statement . . . 2
1.3 Interest . . . 2

2 Data 3
2.1 Data Sources . . . 3
2.2 Data Cleaning . . . 3
2.3 Data Exploration and Visualization . . . 4

3 Methodology 7
3.1 K-Means . . . 7
3.2 HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) . . . 9

4 Results 9
4.1 K-Means Map . . . 9
4.2 HDBSCAN Map . . . 10

5 Discussion 11

6 Conclusion 11

1 Introduction

1.1 Background

A neighborhood is a local community within a city or town. In a multi-ethnic city like Toronto, there are many diverse neighborhoods, a total of 140 communities. Communities play a major role in the lives of people. It influences the beliefs and nature of children and youths. It also influences the decisions of adults and political leaders. There are multiple roles a community plays in the lives of people, which chooses a community of residence a non-trivial matter. As people transition from one stage of life to another, their choices and taste are subject to change—including choice of community. For example, as young couples transition to young families, there is a need for a community suitable for raising a family. As most life transitions, there is a need to make a choice based on predetermined criteria, which is subject to the individual or group. Considering the myriad of factors that mold the nature of a community, it is, therefore, advantageous to have a system that simplifies and aids in deciding on the next resident neighborhood.

1.2 Problem Statement

It is been said that to make an informed decision, knowledge, and understanding of the past are required, in short data. In this case, to make an informed decision on which neighborhood to live in, data on the different neighborhoods in the city of choice is needed. As highlighted above, there are myriad of features in the data describing a neighborhood. In light of this challenge, the objective of this project is to Segment and Cluster all the neighborhoods in Toronto.

1.3 Interest

The results from this project will prove useful to individuals who are searching for a new neighborhood of residence. It will also prove useful to businesses looking to appeal to a different residential clientele. It proved useful to urban/community planners who are looking to transform their community into a state similar to another community.

2 Data

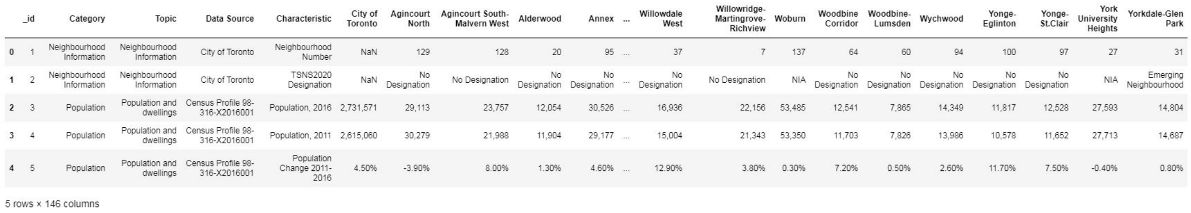
In developing a non-trivial solution, the data required for this project had to be rich. The data consists of environmental structures, such as top venues in the neighborhood, and non-environmental structures, such as the family types and average income in the neighborhood. They will be classified into census and non-census data.

2.1 Data Sources

Foursquare location data will be leveraged in retrieving non-census data. Foursquare provides a rich dataset on the environmental structures in a neighborhood, which will be streamlined to narrow down the features. This data can be accessed using a Foursquare developer account. The census data will be collected from the City of Toronto website. The City of Toronto website has publicly available data on all the neighborhoods from the Census in 2016. This data has a large number of features, which will not be streamlined due to the relevance of the features. The data is downloaded from the website as a CSV document, `https://open.toronto.ca/dataset/neighbourhood-profiles/`.

2.2 Data Cleaning

The census data retrieved from the City of Toronto website contains different categories, topics, and characteristics of data. For example, the Categories column has entries such as Neighbourhood, income, and mobility.



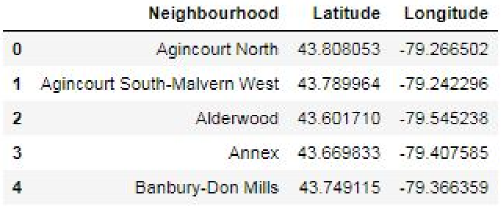
	_id	Category	Topic	Data Source	Characteristic	City of Toronto	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	...	Willowdale West	Willowridge-Martingrove-Richview	Woburn	Woodbine Corridor	Woodbine-Lumsden	Wychwood	Yonge-Eglinton	Yonge-St.Clair	York University Heights	Yorkdale-Glen Park
0	1	Neighbourhood Information	Neighbourhood Information	City of Toronto	Neighbourhood Number	NaN	129	128	20	95	...	37	7	137	64	60	94	100	97	27	31
1	2	Neighbourhood Information	Neighbourhood Information	City of Toronto	TSNS2020 Designation	NaN	No Designation	No Designation	No Designation	No Designation	...	No Designation	No Designation	NIA	No Designation	No Designation	No Designation	No Designation	No Designation	NIA	Emerging Neighbourhood
2	3	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2016	2,731,571	29,113	23,757	12,054	30,526	...	16,936	22,156	53,485	12,541	7,865	14,349	11,817	12,528	27,593	14,804
3	4	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2011	2,615,060	30,279	21,988	11,904	29,177	...	15,004	21,343	53,350	11,703	7,826	13,986	10,578	11,652	27,713	14,687
4	5	Population	Population and dwellings	Census Profile 98-316-X2016001	Population Change 2011-2016	4.50%	-3.90%	8.00%	1.30%	4.60%	...	12.90%	3.80%	0.30%	7.20%	0.50%	2.60%	11.70%	7.50%	-0.40%	0.80%

5 rows × 146 columns

Figure 1: First 5 rows of Census table

Figure 1 above highlights is a sample of the columns and rows of the census data. The Categories, Topics, and Data sources are removed because they are not essential for this project. The table Characteristics column is set as the header and the Neighbourhood columns are set as the row for this project.

GoogleMaps API is used to retrieve the longitude and latitude of the different neighborhoods due to the inconsistency of the GeoPy service.

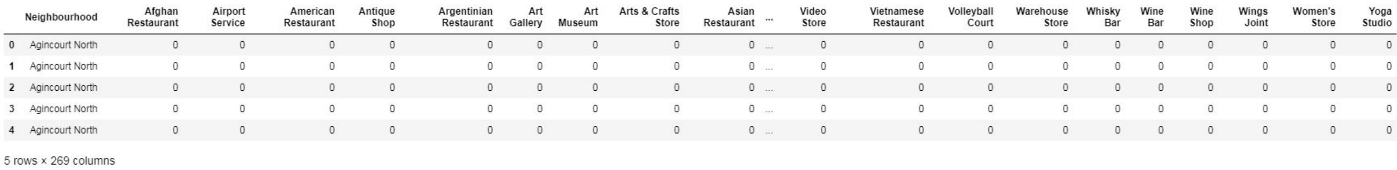


	Neighbourhood	Latitude	Longitude
0	Agincourt North	43.808053	-79.266502
1	Agincourt South-Malvern West	43.789964	-79.242296
2	Alderwood	43.601710	-79.545238
3	Annex	43.669833	-79.407585
4	Banbury-Don Mills	43.749115	-79.366359

Figure 2: First 5 rows of the latitude and longitude of the neighborhoods using the GoogleMaps API

Figure 2 above is a sample of the longitude and latitude of some of the neighborhoods.

Using the FourSquare API explore feature, the top 100 popular venues in each neighborhood are retrieved. Figure 3 below is a sample of the venues retrieved using FourSquare.

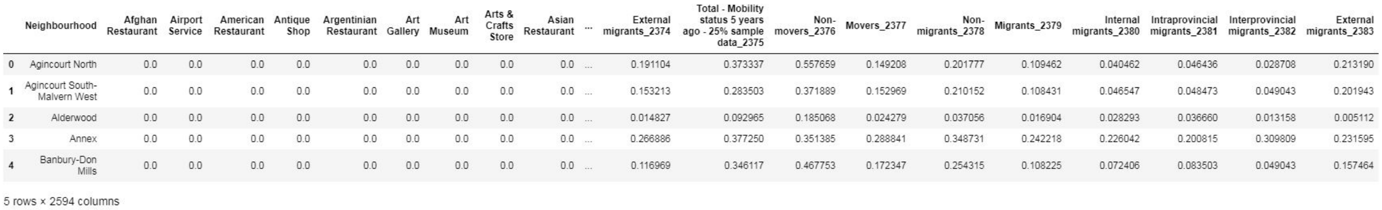


	Neighbourhood	Afghan Restaurant	Airport Service	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	...	Video Store	Vietnamese Restaurant	Volleyball Court	Warehouse Store	Whisky Bar	Wine Bar	Wine Shop	Wings Joint	Women's Store	Yoga Studio
0	Agincourt North	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	Agincourt North	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	Agincourt North	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	Agincourt North	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	Agincourt North	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 269 columns

Figure 3: A sample of the venue information retrieved using the FourSquare API

Once the census and non-census have been retrieved they are combined into one table. The table below is a sample of the combined data.



	Neighbourhood	Afghan Restaurant	Airport Service	American Restaurant	Antique Shop	Argentinian Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	...	External migrants_2374	Total - Mobility status 5 years ago - 25% sample data_2375	Non-movers_2376	Movers_2377	Non-migrants_2378	Migrants_2379	Internal migrants_2380	Intraprovincial migrants_2381	Interprovincial migrants_2382	External migrants_2383
0	Agincourt North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.191104	0.373337	0.557659	0.149208	0.201777	0.109462	0.040462	0.046436	0.028708	0.213190
1	Agincourt South-Malvern West	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.153213	0.283503	0.371889	0.152969	0.210152	0.108431	0.046547	0.048473	0.049043	0.201943
2	Alderwood	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.014827	0.092965	0.185068	0.024279	0.037056	0.016904	0.028293	0.036660	0.013158	0.005112
3	Annex	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.266886	0.377250	0.351385	0.288841	0.348731	0.242218	0.226042	0.200815	0.309809	0.231595
4	Banbury-Don Mills	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.116969	0.346117	0.467753	0.172347	0.254315	0.108225	0.072406	0.083503	0.049043	0.157464

5 rows × 2594 columns

Figure 4: A sample of the combination of the census and non-census data

2.3 Data Exploration and Visualization

In this section, a subset of the features is plotted in represent to the neighborhoods. These features are the Population in 2011, Land area in square kilometers, Pre-retirement (55-64 years), and Youth (15-24 years).

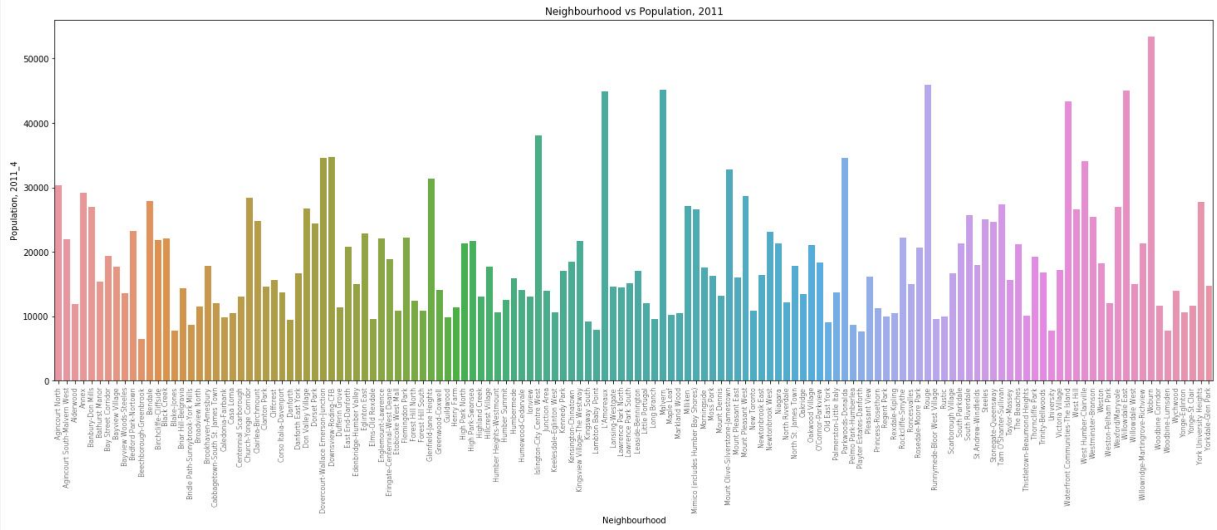


Figure 5: A plot highlighting the Population in 2011 across the different neighborhoods

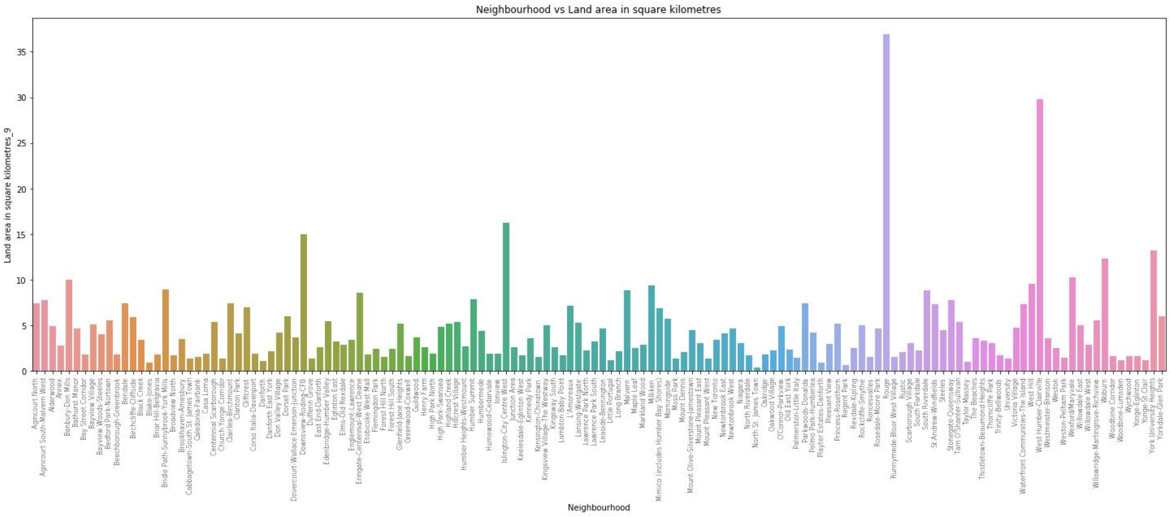


Figure 6: A plot highlighting the Land area in square kilometers across the different neighborhoods

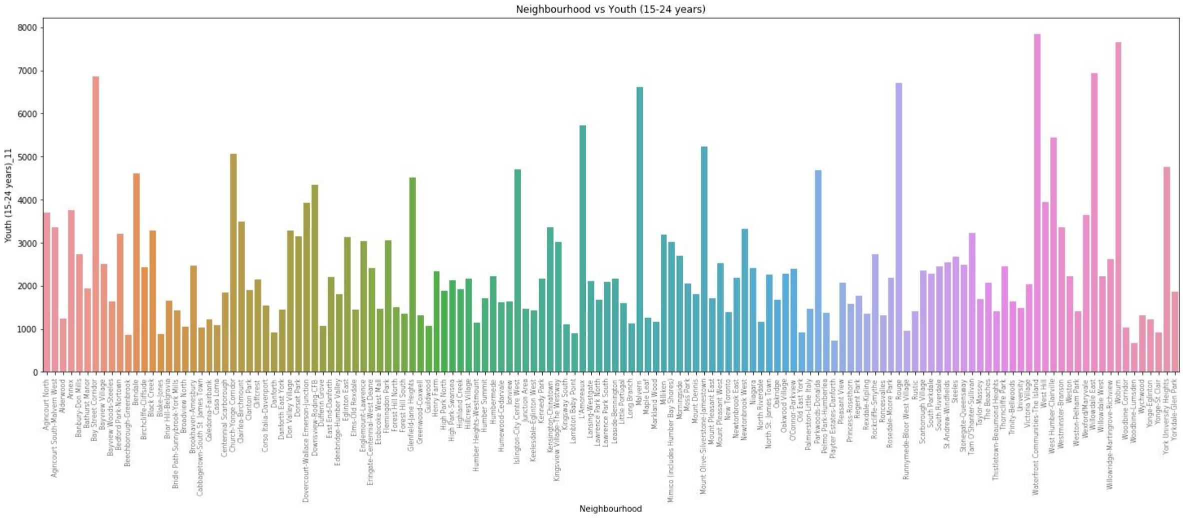


Figure 7: A plot highlighting the Youth (15-24 years) across the different neighborhoods

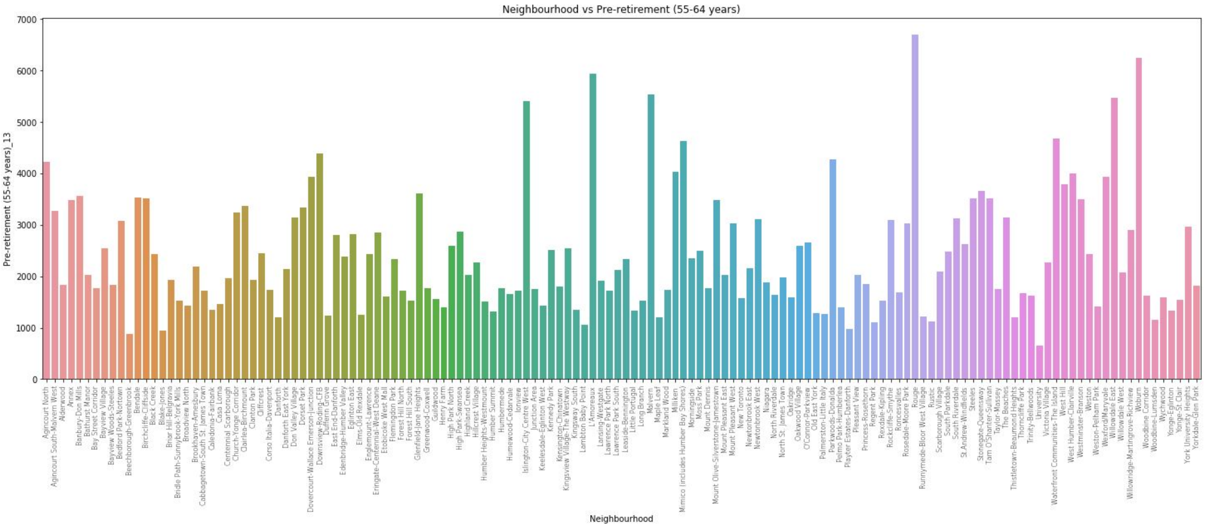


Figure 8: A plot highlighting the Pre-retirement (55-64 years) across the different neighborhoods

3 Methodology

This project is to get an optimal group of clusters for the neighborhoods based on the features. This requires gaining insight into underlying patterns in the data, a typical unsupervised learning problem. For this project, two different unsupervised machine learning algorithms are used: K-means clustering and Hdbscan.

3.1 K-Means

K-means clustering requires starting with an initial guess for the number of clusters. K-means using the euclidean distance then determine the right cluster centers through an iterative process. To get the right number of clusters, the Elbow method is used with a given metric. The number of clusters is changed and assessed to find the optimal number. For this project, Distortion and Inertia are used as metrics. Distortion is the average of the euclidean squared distance from the centroid of the respective clusters. Inertia is the sum of squared distances of the data points to their cluster center. Clusters from 1-30 are passed through the algorithm. There are a couple of assumptions behind the K-means algorithm:

1. Round and spherical clusters
2. Equally sized and dense clusters
3. Clusters with high density at the center
4. Absence of noise and outliers in the data

The assumptions above have to hold for the algorithm to work effectively and efficiently. If the data doesn't meet the above requirements, a density-based or hierarchical algorithm will be required.
Below is a graph showing the change in Distortion and Inertia over the different K clusters.

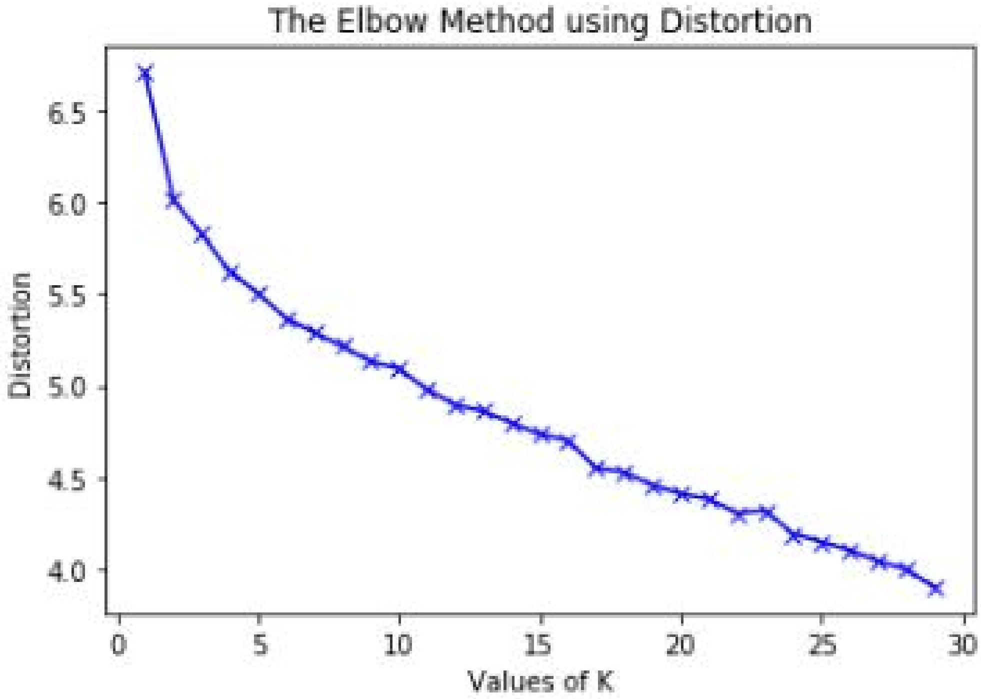


Figure 9: Figure of K clusters vs the Distortion

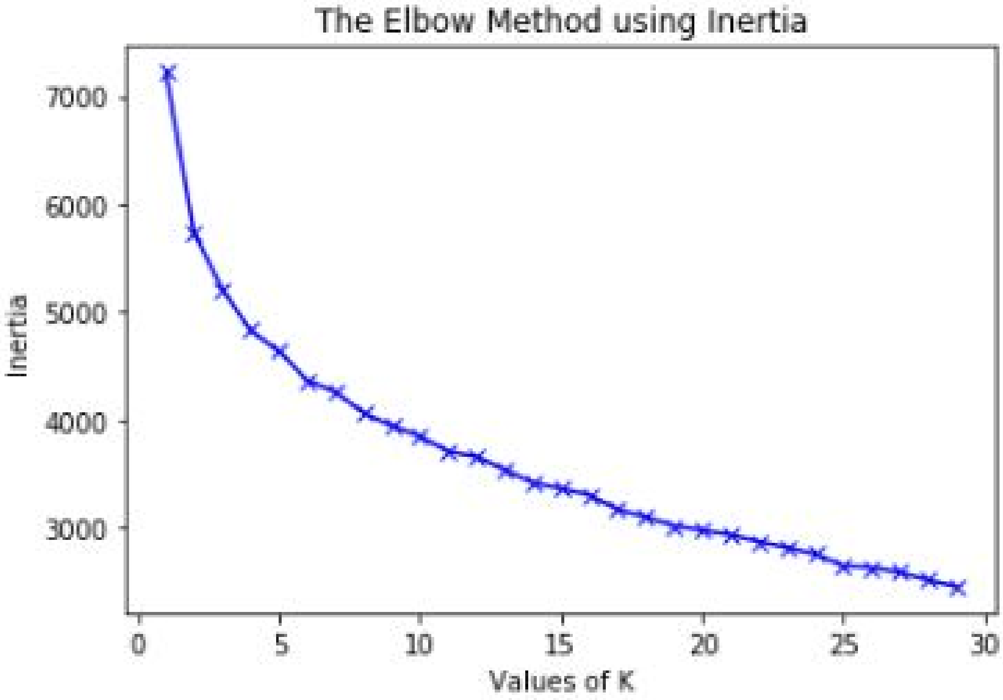


Figure 10: Figure of K clusters vs the Inertia

3.2 HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise)

Hdbscan is a Hierarchical density-based method of clustering. It starts by assigning all data points as its cluster. It builds a hierarchy by merging the two nearest data points. Hbscan works well when the assumptions for a K-means algorithm do not hold. In the case of this project, the assumptions do not hold. In Figures 9 and 10, the optimal number of clusters is not clear as there is no clear elbow in the plots.

4 Results

4.1 K-Means Map

Using the Elbow method, there was no clear optimal number of clusters as there are multiple points that seemed to be the elbow. 10 was the number of clusters chosen. Below is a map, showing all the neighborhoods and color-coded to identify their clusters.

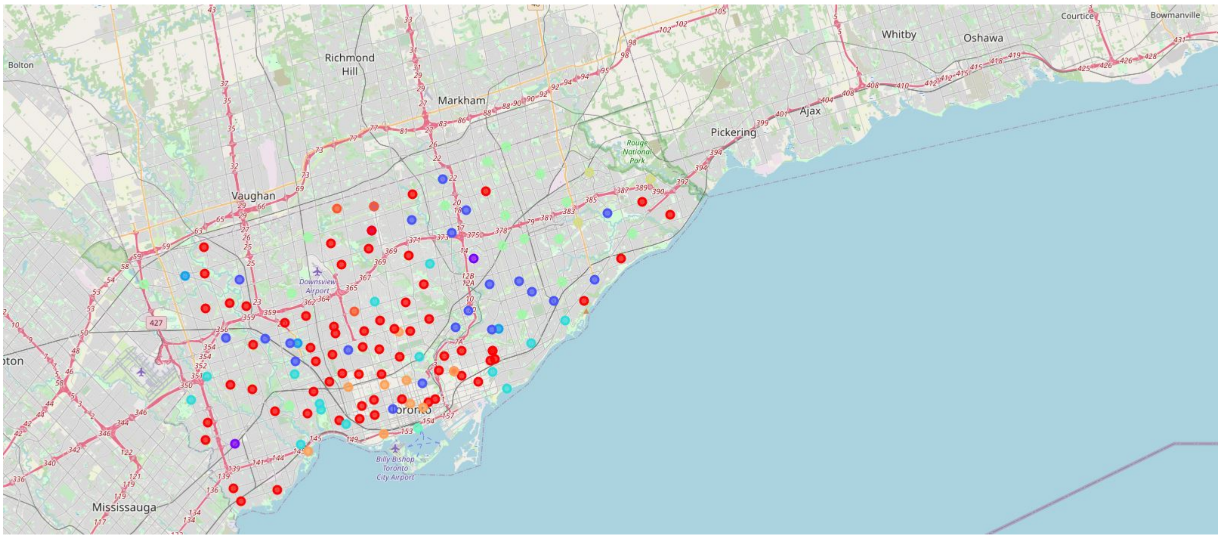


Figure 11: A map displaying the different neighborhoods color-coded across the 10 clusters

4.2 HDBSCAN Map

Using Hdbscan the result is 6 clusters. The Hdbscan solution is chosen for this experiment due to the assumptions not met using the K-means algorithm. Below is a map, showing all the neighborhoods and their clusters by color.

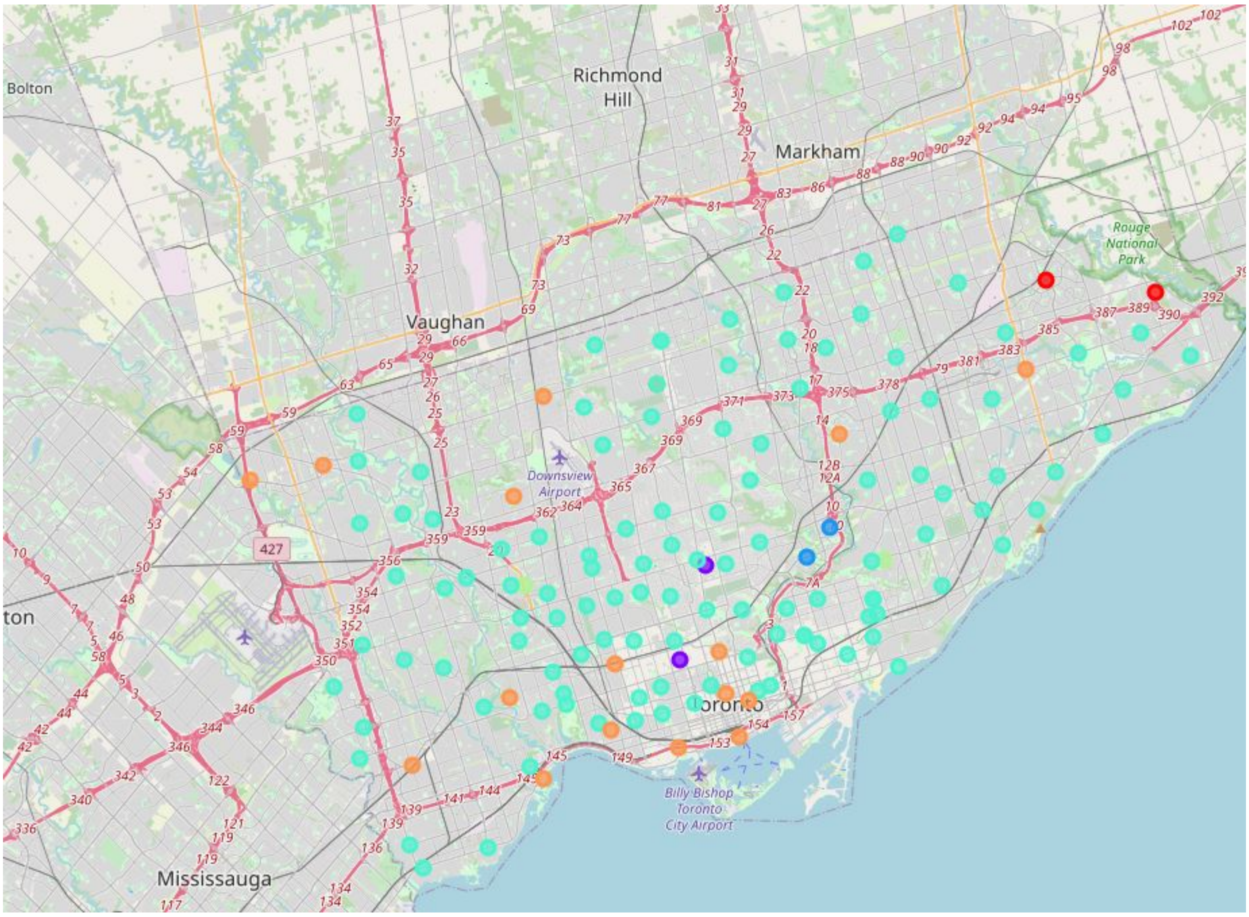
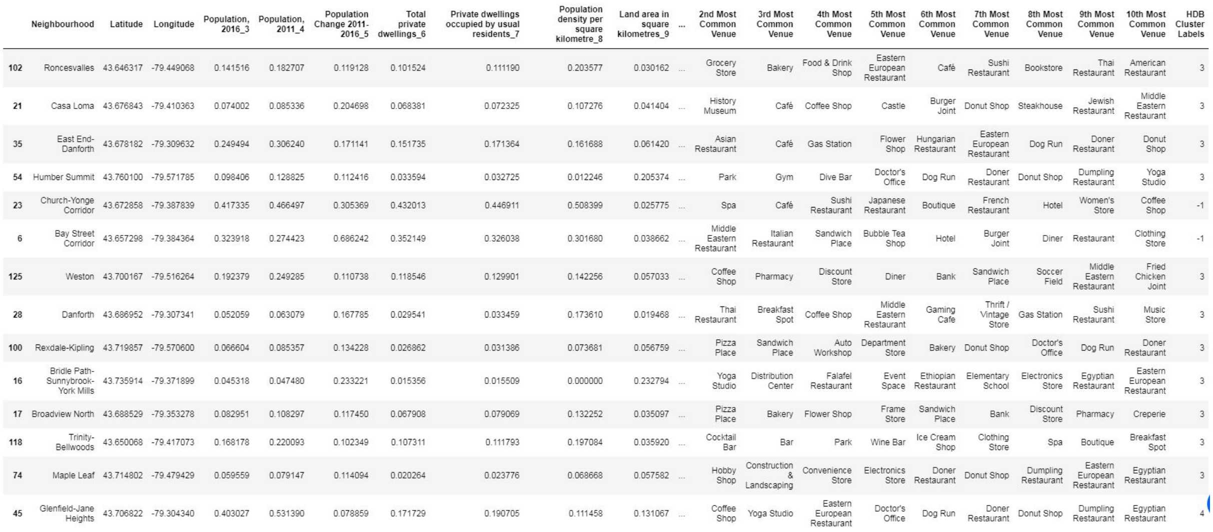


Figure 12: A map displaying the different neighborhoods color-coded across the 6 clusters

5 Discussion

In this project, using the preferred clustering algorithm, the neighborhoods in Toronto were grouped into 6 different clusters. Hdbscan was the preferred clustering algorithm due to the assumptions not met by the data to be optimal for a K-means algorithm. The neighborhoods were grouped based on census and non-census data. Depending on the institution or individual, they can use the results to identify similar neighborhoods. Also, certain institutions or individuals might desire the clustering to be done solely either the census or non-census data. The program can be modified for such situations by choosing the data of choice and removing the process that merges the data. A sample of the neighborhoods in their different is displayed below:



	Neighbourhood	Latitude	Longitude	Population, 2016_3	Population, 2011_4	Population Change 2011-2016_5	Total private dwellings_6	Private dwellings occupied by usual residents_7	Population density per square kilometre_8	Land area in square kilometres_9	...	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	HDB Cluster Labels
102	Roncesvalles	43.646317	-79.449068	0.141516	0.182707	0.119128	0.101524	0.111190	0.203577	0.030162	...	Grocery Store	Bakery	Food & Drink Shop	Eastern European Restaurant	Café	Sushi Restaurant	Bookstore	Thai Restaurant	American Restaurant	3
21	Casa Loma	43.676843	-79.410363	0.074002	0.085336	0.204698	0.068381	0.072325	0.107276	0.041404	...	History Museum	Café	Coffee Shop	Castle	Burger Joint	Donut Shop	Steakhouse	Jewish Restaurant	Middle Eastern Restaurant	3
35	East End-Danforth	43.678182	-79.309632	0.249494	0.306240	0.171141	0.151735	0.171364	0.161688	0.061420	...	Asian Restaurant	Café	Gas Station	Flower Shop	Hungarian Restaurant	Eastern European Restaurant	Dog Run	Doner Restaurant	Donut Shop	3
54	Humber Summit	43.760100	-79.571785	0.098406	0.128825	0.112416	0.033594	0.032725	0.012246	0.205374	...	Park	Gym	Dive Bar	Doctor's Office	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant	Yoga Studio	3
23	Church-Yonge Corridor	43.672858	-79.387839	0.417335	0.466497	0.305369	0.432013	0.446911	0.508399	0.025775	...	Spa	Café	Sushi Restaurant	Japanese Restaurant	Boutique	French Restaurant	Hotel	Women's Store	Coffee Shop	-1
6	Bay Street Corridor	43.657298	-79.384364	0.323918	0.274423	0.686242	0.352149	0.326038	0.301680	0.038662	...	Middle Eastern Restaurant	Italian Restaurant	Sandwich Place	Bubble Tea Shop	Hotel	Burger Joint	Diner	Restaurant	Clothing Store	-1
125	Weston	43.700167	-79.516264	0.192379	0.249285	0.110738	0.118546	0.129901	0.142256	0.057033	...	Coffee Shop	Pharmacy	Discount Store	Diner	Bank	Sandwich Place	Soccer Field	Middle Eastern Restaurant	Fried Chicken Joint	3
28	Danforth	43.686952	-79.307341	0.052059	0.063079	0.167785	0.029541	0.033459	0.173610	0.019468	...	Thai Restaurant	Breakfast Spot	Coffee Shop	Middle Eastern Restaurant	Gaming Cafe	Thrift / Vintage Store	Gas Station	Sushi Restaurant	Music Store	3
100	Rexdale-Kipling	43.719857	-79.570600	0.066604	0.085357	0.134228	0.026862	0.031386	0.073681	0.056759	...	Pizza Place	Sandwich Place	Auto Workshop	Department Store	Bakery	Donut Shop	Doctor's Office	Dog Run	Doner Restaurant	3
16	Bridle Path-Sunnybrook-York Mills	43.735914	-79.371899	0.045318	0.047480	0.233221	0.015356	0.015509	0.000000	0.232794	...	Yoga Studio	Distribution Center	Falafel Restaurant	Event Space	Ethiopian Restaurant	Elementary School	Electronics Store	Egyptian Restaurant	Eastern European Restaurant	3
17	Broadview North	43.688529	-79.353278	0.082951	0.108297	0.117450	0.067908	0.079069	0.132252	0.035097	...	Pizza Place	Bakery	Flower Shop	Frame Store	Sandwich Place	Bank	Discount Store	Pharmacy	Creperie	3
118	Trinity-Bellwoods	43.650068	-79.417073	0.168178	0.220093	0.102349	0.107311	0.111793	0.197084	0.035920	...	Cocktail Bar	Bar	Park	Wine Bar	Ice Cream Shop	Clothing Store	Spa	Boutique	Breakfast Spot	3
74	Maple Leaf	43.714802	-79.479429	0.059559	0.079147	0.114094	0.020264	0.023776	0.068668	0.057582	...	Hobby Shop	Construction & Landscaping	Convenience Store	Electronics Store	Doner Restaurant	Donut Shop	Dumpling Restaurant	Eastern European Restaurant	Egyptian Restaurant	3
45	Glenfield-Jane Heights	43.706822	-79.304340	0.403027	0.531390	0.078859	0.171729	0.190705	0.111458	0.131067	...	Coffee Shop	Yoga Studio	Eastern European Restaurant	Doctor's Office	Dog Run	Doner Restaurant	Donut Shop	Dumpling Restaurant	Egyptian Restaurant	4

Figure 13: A map displaying the different neighborhoods color-coded across the 6 clusters

6 Conclusion

The objective of this project is to Segment and Cluster all the neighborhoods in Toronto. The data consists of the census and non-census types. The census data was retrieved from the City of Toronto website and the non-census data was leveraging the Foursquare API. GoogleMaps API was used to the longitude and latitude data each of the neighborhoods. Two different clustering algorithms were used: K-means and Hdbscan. For K-means to work effectively, the data must meet certain assumptions; but in the case of this project, the assumptions were not met. Hdbscan was used and became the algorithm of choice for this project: clustering the data into 6 groups based on census and non-census data.