



DATA SCIENCE PROFESSIONAL CERTIFICATION



K-means & DBSCAN Clustering for Desirable Living Locale

January 28, 2019

Wade Hooper

K-means & DBSCAN Clustering for Desirable Living Locale

Revision Version	Date	Comments
<i>BOTN-REP-WH-001</i>	01/20/2019	Added Introduction and Data sections
<i>BOTN-REP-WH-002</i>	01/28/2019	Added remaining sections of the report

Abstract

Data analysis was conducted on popular venues and crimes statistics within Ottawa, Ontario. Specifically, the University of Ottawa and Carleton University were chosen as reference points for user impact. DBSCAN and k-means machine learning algorithms were used on crime and Foursquare venue data, respectively. DBSCAN clustered coordinates and total crime reports of all neighbourhoods; dominant clusters were present around the University of Ottawa and Carleton University. The Ottawa downtown core showcased high levels of criminal activity, also given that large social venues are present therein. The neighbourhoods nearby Carleton University reported a fifty percent decrease in criminal activity.

K-means classified clusters amongst the top five popular venues within each neighbourhood; this again returned clusters around university campuses. Eastern Ottawa reported largely popular grocery and clothing stores, while crime levels were similar to those surround Carleton University. The downtown core is popularised by coffee shops and high capacity venues such as parks, concert halls, and cafes, while Carleton is denoted by coffee shops and bakeries.

Given the distribution of popular venues surrounding each campus (and neighbouring areas), crime data illustrates potential residents are more at risk around the downtown area, and half as much in danger at any other location in Ottawa.

K-means & DBSCAN Clustering for Desirable Living Locale

Table of Contents

1.0	Introduction/Business Problem	1
2.0	Data.....	1
3.0	Methodology	2
4.0	Results.....	3
5.0	Discussion	5
6.0	Conclusion	5
7.0	References.....	6

List of Figures

Figure 1: K-means Elbow Plot	3
Figure 2: DBSCAN Map Clusters	3
Figure 3: K-means Map Clusters	4

List of Tables

Table 1: DBSCAN Inputs.....	2
Table 2: DBSCAN University of Ottawa, Carleton University, and Eastern Ottawa Clusters.....	3
Table 3: K-means University of Ottawa, Carleton University, and Eastern Ottawa Clusters.....	4

Abbreviations

DBSCAN	Density-based spatial clustering of applications with noise
--------	---

K-means & DBSCAN Clustering for Desirable Living Locale

1.0 Introduction/Business Problem

Moving for many people can be a complicated and stressful ordeal. In a large city such as Ottawa, there may be different requirements when moving to a new neighbourhood; whether these are comforts similar to home (foods, shops, etc.) or safety (such as a similar lack of high crime levels in the surrounding area).

This paper focused on providing an analysis of neighbourhood safety as compared to available services and stores. Simply put, if a person had certain preferences to shops or restaurants in a new neighbourhood, would that same neighbourhood be one where they might feel safe to live, commute, and work? This paper will provide a benefit to anyone in the Ottawa National Capital Region, searching for a new home or apartment; however, focus was given to new students or residents in the Ottawa area who may have little familiarity with neighbourhood crime levels. Please note, this report will *not* incorporate any cost-of-living into the datasets (such as rent, utilities, etc.), but instead focus on crime and popular venue data.

2.0 Data

To visualize the distribution of nearby restaurants, stores, or shops, the Foursquare API was used to retrieve location data for each venue from nearby neighbourhoods [1]. Crime incidents were visualized via data provided by the Ottawa Police Service as compiled by Carleton University [2]; this dataset lists incident occurrences for specific crimes (assault, theft, etc.) for the 2015 year within each neighbourhood ward (or borough). While the most recent crime dataset was listed in 2015, this report assumes stagnate or increasing trend in criminal activity, to ere on the side of caution for potential new residents.

The Foursquare data was used to cluster neighbourhoods across Ottawa. As the clusters would highlight trends, doing so would gain insights into popular venues of each neighbourhood; such as if neighbourhoods in Ottawa's downtown core have a large variation in available restaurant cuisine (e.g. Asian, French, etc.), or if the neighbourhoods further from the city center have nearby fitness facilities. The resulting clusters will be displayed on an overhead map of the Ottawa region.

In addition to neighbourhoods clustered by popular venues, crime data was incorporated and displayed on the overhead map; an example being a higher incident reports in the inner city or large neighbourhoods. While the crime dataset included many incidents, assault (both non-sexual and sexual) and robbery amounts were isolated to provide a simplified description of each neighbourhood for potential new residents.

K-means & DBSCAN Clustering for Desirable Living Locale

3.0 Methodology

While the crime dataset lists individual incidents (and a grand total) for each neighbourhood, coordinate values were not provided for each report; this lead to assigning total incidents to the neighbourhood center, an example being a single point for all 319 incidents in the Orleans neighbourhood. Due to the dense nature and non-spherical distribution of the Ottawa neighbourhoods, a DBSCAN algorithm was used to initially cluster the neighbourhoods based on the attributes and parameters detailed in Table 1 below:

Epsilon and minimum sample parameters were determined by empirical testing. Errors in mislabelling neighbourhoods as outliers was attempted to be minimized; this was done by using the Rideau Canal as a separation point for primary clusters on the east and west.

Table 1: DBSCAN Inputs

Data Attribute	Description
Latitude	Neighbourhood coordinate value
Longitude	Neighbourhood coordinate value
Grand Total	Incident total per neighbourhood
DBSCAN Parameter	
Epsilon	0.75
Minimum samples	2
Algorithm	K-Nearest Neighbours Auto

After the initial clustering, the top five venues for each neighbourhood were retrieved with use of the Foursquare API with a search radius of 1200 meters, and subsequently clustered with a k-means algorithm. The parameters used with k-means were fifty initial runs, random state of one, and five clusters. As k-means has been used in the past to categorize neighbourhoods in New York and Toronto based on popular venues, the Ottawa market could benefit from a similar analysis.

K-means & DBSCAN Clustering for Desirable Living Locale

4.0 Results

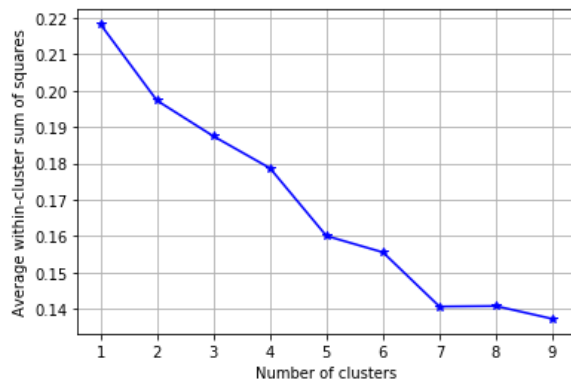


Figure 1: K-means Elbow Plot

The “ideal” number of clusters was selected with the Elbow plot, as shown in Figure 1 at right. However, the plot shows a linear slope and no discernable “elbow” to pinpoint cluster efficiency. Such a trend may suggest the k-means approach may not be suitable in classifying this dataset, or that alternative preprocessing techniques must be used.

DBSCAN clusters are shown in Figure 2, at right. Locations of notable student campuses are shown as green, red, and blue icons for Algonquin College, Carleton University, and University of Ottawa respectively. The size of each cluster represents a scale of total reported incidents relative to the max value (605 incidents); where the large central blue cluster is the dominant crime area in the Ottawa region (due to the downtown location). Please refer to the Jupyter notebook linked on the submission page to explore the full map of the Ottawa region (be careful to use the NB viewer link [shown here](#) to properly view map outputs) [3].

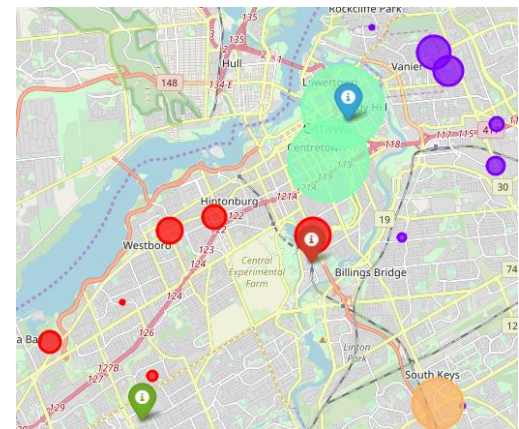


Figure 2: DBSCAN Map Clusters

Data frame results are listed in Table 2 for the University of Ottawa, Carleton University, and Eastern Ottawa areas. These tables highlight individual incident totals for each neighbourhood, as well as the total represented on the overhead map.

Table 2: DBSCAN University of Ottawa, Carleton University, and Eastern Ottawa Clusters

Neighbourhood	Assaults	Sexual Assault	Robbery	Grand Total (all incidents)	Cluster
Univ. of Ottawa	385	52	69	605	3 (blue)
Centertown	310	72	60	593	
Carleton University	120	57	23	255	0 (red)
Westborough	106	17	15	184	
Civic Hospital	106	15	11	167	
Vanier	144	22	35	250	1 (purple)
Overbrook	110	15	31	209	
Riverview	61	14	16	133	
Beacon Hill	41	16	15	107	

K-means & DBSCAN Clustering for Desirable Living Locale

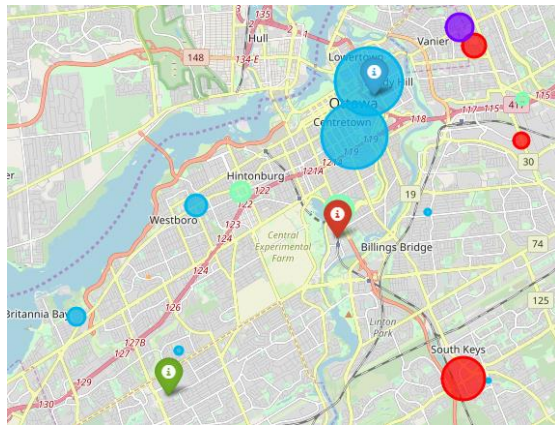


Figure 3: K-means Map Clusters

K-means clusters are shown in Figure 3, at left. The dark blue is referred to as the University of Ottawa cluster, while the light green is referred to as the Carleton University cluster. Eastern Ottawa is represented by the red cluster. Remaining clusters are best viewable on the shared Jupyter notebook.

Data frame results are listed in Table 3 below for the University of Ottawa, Carleton University, and Eastern Ottawa areas. This table highlights the top five venues for neighbourhoods within each cluster.

Please note that k-means is subject to variation based on initial parameters, the dataset showcased in the Jupyter notebook and below was deemed the best result from empirical testing and mapping.

Table 3: K-means University of Ottawa, Carleton University, and Eastern Ottawa Clusters

Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	Cluster
Univ. of Ottawa	Coffee Shop	Café	Hotel	Restaurant	Park	2 (dark blue)
Centertown	Coffee Shop	Café	Restaurant	Vegetarian / Vegan Restaurant	Concert Hall	
Carleton University	Coffee Shop	Pub	Bakery	Park	Diner	3 (light green)
Civic Hospital	Coffee Shop	Italian Restaurant	Bakery	Sandwich Place	Tapas Restaurant	
Meadowlands	Snack Place	Shopping Mall	Supermarket	Pool Hall	Sandwich Place	
Heron Gate	Grocery Store	Coffee Shop	Clothing Store	Department Store	Fast Food Restaurant	0 (red)
Overbrook	Pizza Place	Fast Food Restaurant	Grocery Store	Coffee Shop	Indian Restaurant	
Riverview	Coffee Shop	Grocery Store	Fast Food Restaurant	Sandwich Place	Seafood Restaurant	

K-means & DBSCAN Clustering for Desirable Living Locale

5.0 Discussion

The lack of a reliable k-value from the Elbow plot (refer to Figure 1, pg. 3) would suggest the venues dataset as not sufficiently clustered or spherical, and required an additional method to reveal patterns within another aspect of the data (such as crime levels) [4]. For a future investigation, precise coordinate values of each popular venue may be useful in obtaining more insights from Foursquare data.

Results from the DBSCAN clustering indicate higher crime levels nearby the University of Ottawa than the residential areas near Carleton University (refer to Table 2, pg. 3). Specifically a near fifty percent decrease in total crimes reported relative to the downtown core. Eastern Ottawa shows a similar magnitude to Carleton University; therefore, Carleton University students may consider certain area of Eastern Ottawa at a comparable or greater risk to safety due to a longer commute.

When coupled with k-means cluster results shown in Figure 3, and Table 3, both major university campuses share a similar preference for coffee shops; this is to be expected as coffee shops are fairly ubiquitous within large cities. Investigating the *specific* coffee shops for a future analysis may assist in helping pinpoint desirable brand-name locations (Starbucks, Bridgehead, etc.) that pertain to a student's tastes. Of note, the neighbourhoods surrounding Carleton University are popularized by bakeries, again where additional research may be conducted into specific storefronts. While downtown Ottawa provides more social spaces and access to large entertainment locations (concert halls, parks, stadiums, etc.), Eastern Ottawa is highlighted by grocery and clothing stores, as well as Indian cuisine. The chances are, if a student is searching for a decent food shopping experience (or Indian cuisine), they will find themselves in the "Hunt Club" area of Ottawa.

Another topic of exploration would be inclusion of rental and utility costs in the Ottawa area; this may give yet another level of insight into the ideal location for students and new residents.

6.0 Conclusion

Students in the Ottawa area are at a greater risk to crime in the downtown core, particularly those living in close proximity (mere blocks, or on campus) to the University of Ottawa. Concerned students of Carleton University may not see Eastern Ottawa as desirable, due to comparable crime levels as near campus, with an additional lengthy commute.

Coffee shops are the most popular venue surrounding university campuses in Ottawa, while the Eastern districts are represented by everyday needs, such as grocery and clothing stores. Prospective residents will need to weigh the cost of their travel time against the proximity to neighbourhoods with high levels of reported crime.

K-means & DBSCAN Clustering for Desirable Living Locale

7.0 References

- [1] Ottawa Police Service. Ottawa Postal Codes. [Online].
] <https://www.ottawapolice.ca/en/contact-us/resources/ottawa-postal-codes.pdf>
- [2] Carleton University. (2016) Ottawa Police Service Crime Data. [Online].
] <https://library.carleton.ca/find/gis/geospatial-data/ottawa-police-service-crime-data>
- [3] Wade Hooper. (2019, January) Battle of The Neighbourhoods. [Online].
] <https://nbviewer.jupyter.org/github/WadeHooper/IBMDSCapstone/blob/master/BattleOfTheNeighborhoods.ipynb>
- [4] Yellowbrick. (2016) scikit-yellowbrick. [Online]. <http://www.scikit-yb.org/en/latest/api/cluster/elbow.html>