# LOCATION SELECTION FOR TORONTO PUBLIC LIBRARY

Thripurasundari Venugopalan

COURSERA CAPSTONE

Aug 2021

# Contents

# Executive Summary

The objective of this project is to identify a suitable location for a library in Toronto, considering location specific criteria. The project gathers a list of possible locations; their proximity to existing libraries and later groups them into different clusters depending on their similarities. The similarities are based on the most common venues near by and distance to the nearest library.

Data is collected from various sources, such as Wikipedia, Foursquare API, Google GeoCoding API. K-Means clustering methodology has been used to identify the clusters.

The recommended solution can be obtained by analysing the clusters based on their near by venues and the distance to the existing libraries.

## Introduction (Business Problem)

The Corporation of the City of Toronto's Education Department wishes to build a free library in Toronto, called Toronto Central Library, at a suitable location, such that it can be used by various students (especially financially weak) across Toronto to enhance the education, research, and development in the city.

The library under discussion needs to be at an appropriate location with proximity to the following:

1) Bus/Train Station (supporting ease of local or public transportation facilities)
2) Coffee shop/Café/ Restaurant (as a hangout between study/research hours)
3) Any recreational area such as Playground or park (optional)

It should be in an area where a library does not exist in proximity.

# Literature review

In the popular Data Science journal, Towards Data Science, **Dr. Michael J. Garbade, talks about** K-means clustering as an unsupervised machine learning methodology (https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1). The K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids

It halts creating and optimizing clusters when either:

- The centroids have stabilized — there is no change in their values because the clustering has been successful.
- The defined number of iterations has been achieved.

However, its performance is usually not as competitive as those of the other sophisticated clustering techniques because slight variations in the data could lead to high variance.

Furthermore, clusters are assumed to be spherical and evenly sized, something which may reduce the accuracy of the K-means clustering Python results.

# Methodology

To identify a suitable location for a library, in Toronto, a group of neighborhoods in Toronto region were selected. The data was obtained by "Web Scraping", a methodology of extracting data from web pages, using Beautiful Soup API Library in Python

### Data Requirement

The following factors were considered to select appropriate location for library in Toronto:

1. Different neighborhoods in Toronto with information containing Postal Code, Borough Name for each neighborhood (Fig. 1)
2. Geographic Coordinates of the Neighborhoods (Fig. 2)

3. Common Venues with proximity to the neighborhoods (Fig. 3)
4. List of Existing libraries in the region under consideration (Toronto) (Fig. 4)
5. Distance from the neighborhoods to the closest existing library (Fig. 5)

## Data Collection

The data was obtained from various sources as follows:

1. List of Neighborhoods (beginning with 'M') scraped from Wikipedia webpage https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada
2. Coordinates of Neighborhoods using **Google Maps GeoCoding API**
3. Coordinates of the Libraries in the vicinity of Toronto central location using **Foursquare API – Search Venues**
4. Top 10 most common venues in every candidate neighborhood using **Foursquare API – Explorer Venues**
5. Distance between each candidate Neighborhood location and the nearest Library using Haversine formula

## Data Understanding

1. Neighborhood candidates

| | PostalCode | Borough | Neighborhood |
|---|---|---|---|
| 0 | M3A | North York | Parkwoods |
| 1 | M4A | North York | Victoria Village |
| 2 | M5A | Downtown Toronto | Regent Park, Harbourfront |
| 3 | M6A | North York | Lawrence Manor, Lawrence Heights |
| 4 | M7A | Queen's Park | Ontario Provincial Government |
| 5 | M9A | Etobicoke | Islington Avenue |
| 6 | M1B | Scarborough | Malvern, Rouge |
| 7 | M3B | North York | Don MillsNorth |

Fig. 1 Neighborhood candidates

2. Geographic coordinates of candidate neighborhoods

| | PostalCode | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|---|
| 0 | M3A | North York | Parkwoods | 43.752777 | -79.326440 |
| 1 | M4A | North York | Victoria Village | 43.735735 | -79.312418 |
| 2 | M5A | Downtown Toronto | Regent Park, Harbourfront | 43.660323 | -79.362044 |
| 3 | M6A | North York | Lawrence Manor, Lawrence Heights | 43.725924 | -79.436320 |
| 4 | M7A | Queen's Park | Ontario Provincial Government | 43.662278 | -79.391527 |
| 5 | M9A | Etobicoke | Islington Avenue | 43.682778 | -79.540297 |
| 6 | M1B | Scarborough | Malvern, Rouge | 43.809160 | -79.221690 |
| 7 | M3B | North York | Don MillsNorth | 43.744847 | -79.340923 |
| 8 | M4B | East York | Parkview Hill, Woodbine Gardens | 43.706139 | -79.325415 |
| 9 | M5B | Downtown Toronto | Garden District, Ryerson | 43.657658 | -79.378802 |
| 10 | M6B | North York | Glencairn | 43.711705 | -79.431075 |
| 11 | M9B | Etobicoke | West Deane Park, Princess Gardens, Martin Grov... | 43.661359 | -79.558161 |
| 12 | M1C | Scarborough | Rouge Hill, Port Union, Highland Creek | 43.774869 | -79.157587 |

Fig. 2 Geographic coordinates of candidate neighborhoods

The following is a map of Canada with the candidate neighborhood locations

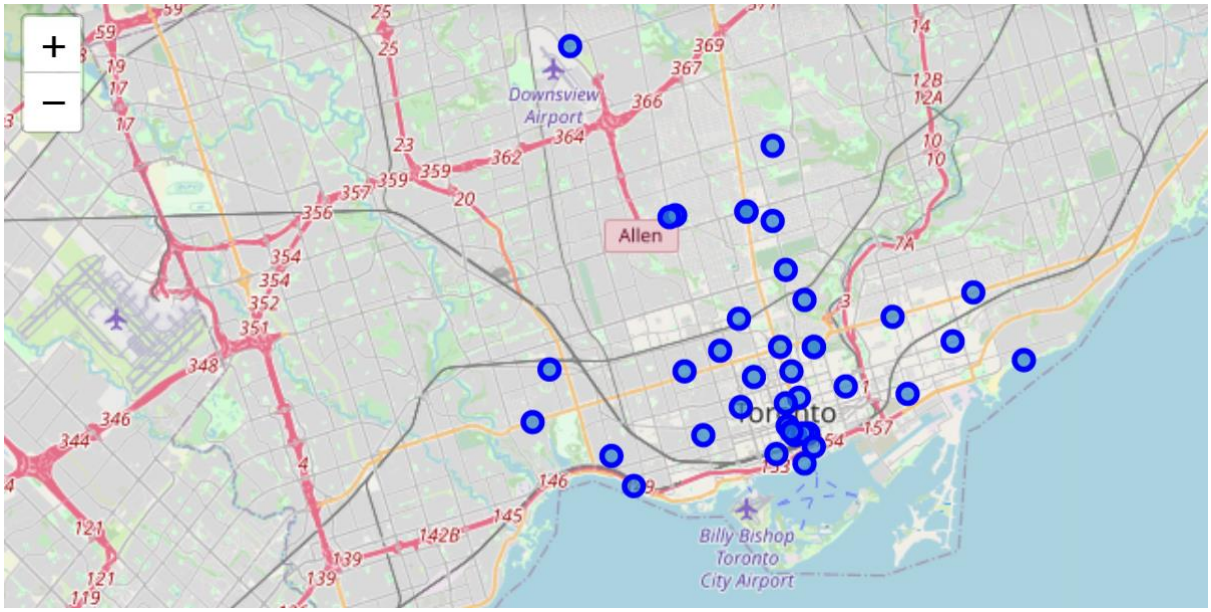


Fig. 2.1 Map of Canada with candidate neighborhoods marked using coordinate locations

3. Common Venues with proximity to the neighborhoods

| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Regent Park Aquatic Centre | 43.660600 | -79.361392 | Pool |
| 1 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Sumach Espresso | 43.658135 | -79.359515 | Coffee Shop |
| 2 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Daniels Spectrum | 43.660137 | -79.361808 | Performing Arts Venue |
| 3 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Sukhothai | 43.658444 | -79.365681 | Thai Restaurant |
| 4 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Paintbox Bistro | 43.660050 | -79.362855 | Restaurant |
| 5 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Thai To Go | 43.663418 | -79.360710 | Thai Restaurant |
| 6 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Qi sushi | 43.662552 | -79.364258 | Sushi Restaurant |
| 7 | Regent Park, Harbourfront | 43.660323 | -79.362044 | I Love Churros | 43.658364 | -79.365583 | Food Truck |
| 8 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Dominion Pub and Kitchen | 43.656919 | -79.358967 | Pub |
| 9 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Regent Park | 43.660383 | -79.361810 | Park |
| 10 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Mercedes-Benz Downtown | 43.661431 | -79.356585 | Auto Dealership |
| 11 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Vistek | 43.657046 | -79.359667 | Electronics Store |
| 12 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Shoppers Drug Mart | 43.660204 | -79.361332 | Pharmacy |

Fig. 3 Venue of candidate neighborhoods

4. List of Existing libraries in the region under consideration (Toronto)

| | Name | Latitude | Longitude |
|---|---|---|---|
| 0 | Toronto Public Library - Toronto Reference Lib... | 43.671795 | -79.386944 |
| 1 | Japan Foundation | 43.670609 | -79.386011 |
| 2 | Toronto Public Library - Lillian H. Smith Branch | 43.658137 | -79.398372 |
| 3 | Toronto Public Library (Fort York Branch) | 43.639172 | -79.400445 |
| 4 | Toronto Vegetarian Association | 43.655953 | -79.392854 |
| 5 | Riverdale Library | 43.665780 | -79.353175 |
| 6 | Toronto Public Library - Bloor Gladstone Branch | 43.660097 | -79.434173 |
| 7 | Toronto Public Library - Pape/Danforth Branch | 43.678603 | -79.344443 |
| 8 | Toronto Public Library - Deer Park Branch | 43.688710 | -79.392603 |
| 9 | Toronto Public Library (St. James Town) | 43.668790 | -79.374998 |
| 10 | Balzac's Coffee | 43.671726 | -79.386952 |
| 11 | Toronto Public Library - Northern District Branch | 43.708481 | -79.400241 |
| 12 | Toronto Public Library (Sanderson Branch) | 43.652165 | -79.405754 |
| 13 | Toronto Public Library | 43.652631 | -79.383295 |
| 14 | Indigo | 43.653515 | -79.380696 |
| 15 | Starbucks | 43.670340 | -79.388262 |
| 16 | Toronto Public Library - Palmerston Branch | 43.665074 | -79.413978 |
| 17 | Royal Ontario Museum | 43.668367 | -79.394813 |
| 18 | Library Bar | 43.645500 | -79.381602 |
| 19 | Toronto Public Library - Wychwood Branch | 43.682076 | -79.417748 |
| 20 | Yorkville Public Library | 43.671841 | -79.388659 |

Fig. 4 Coordinates of libraries in Toronto

5. Distance from the neighborhoods to the closest library (in km)

| | PostalCode | DistanceToLib | Borough | Neighborhood | Latitude | Longitude |
|---|---|---|---|---|---|---|
| 0 | M5A | 2.756569 | Downtown Toronto | Regent Park, Harbourfront | 43.660323 | -79.362044 |
| 1 | M5B | 1.492597 | Downtown Toronto | Garden District, Ryerson | 43.657658 | -79.378802 |
| 2 | M4E | 5.210063 | East Toronto | The Beaches | 43.667348 | -79.296693 |
| 3 | M5E | 2.681541 | Downtown Toronto | Berczy Park | 43.648100 | -79.375200 |
| 4 | M5G | 1.545173 | Downtown Toronto | Central Bay Street | 43.656101 | -79.383866 |
| 5 | M6G | 2.460269 | Downtown Toronto | Christie | 43.664589 | -79.420675 |
| 6 | M5H | 1.835027 | Downtown Toronto | Richmond, Adelaide, King | 43.649929 | -79.383248 |
| 7 | M6H | 10.499267 | West Toronto | Dufferin, Dovercourt Village | 43.750428 | -79.462787 |
| 8 | M4J | 3.188303 | East York/East Toronto | The Danforth East | 43.685043 | -79.315192 |
| 9 | M5J | 3.031953 | Downtown Toronto | Harbourfront East, Union Station, Toronto Islands | 43.640189 | -79.376650 |
| 10 | M6J | 2.016365 | West Toronto | Little Portugal, Trinity | 43.647283 | -79.413725 |
| 11 | M4K | 0.065263 | East Toronto | The Danforth West, Riverdale | 43.678985 | -79.344910 |

Fig. 5 Distance of candidate neighborhoods to nearest library

## Assumptions

For the sake of the project, some requirements have been simplified. However, the project can be extended to achieve the complex requirements.

1. Neighborhoods in Toronto with Borough name starting with M and Neighborhood name having ‘Toronto’ alone have been considered
2. Distance between Neighborhood and nearest library is calculated based on Aerial Distance (Bird flying distance, not road distance)

## Data Preparation and Clean-up

The Foursquare API provides details about locations, venues nearby these locations by distance, category, ratings, etc.

The Foursquare API was then used to get the list of libraries already existing in Toronto using the “caregoryid” of a library (as per Foursquare API category Id specification). Libraries containing the words College or University were excluded, as they may not be open to public and only libraries containing the word Toronto were considered.

The distance between each of the neighbourhood locations and the closest library was calculated using Haversine formula, a method to calculate the aerial distance between locations in km. As Haversine method is aerial distance calculation, and not the driving distance, for all practical applications, driving distance would be more suitable. This would require the use of Google Maps API, (or similar) to calculate the distance between coordinates. However, for the sake of simplicity, Aerial distance was considered.

The following graph shows the approximate distance to the nearest library for each Neighborhood

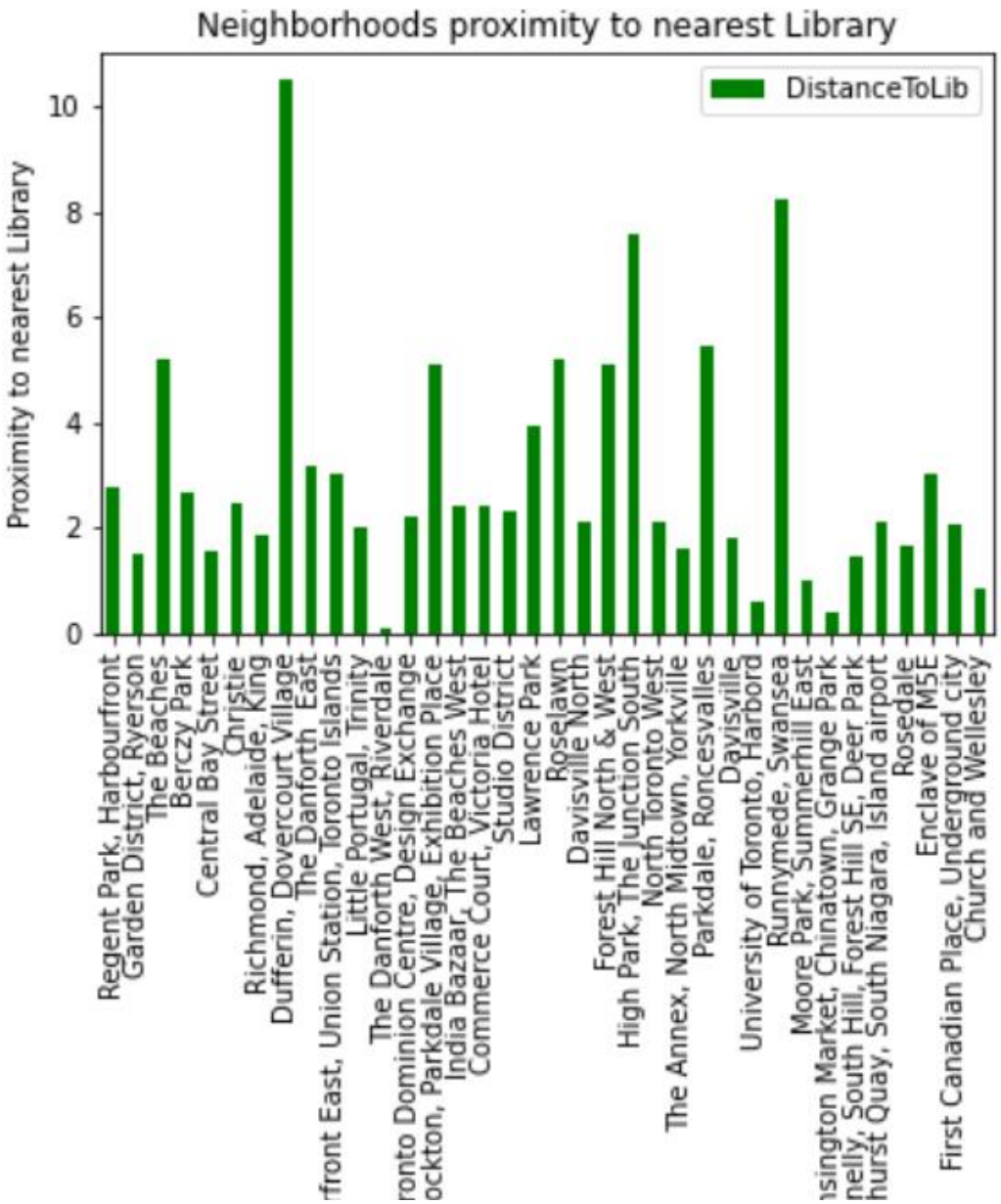


Fig. 6 View of Distance of candidate neighborhoods to nearest library

The farthest location is Dufferin, Dovercourt Village, with distance of 10 km to the nearest library, whereas Enclave of M4L has a library in proximity.

As a next step, the list of most common venues near each of the neighborhood locations was obtained, along with the Venue Category for each venue. If the Venue category consists of the caregory id of a library (as per Foursquare API category Id specification), it was removed from the data set as the objective was to identify a location where the library facility was not available nearby.

Following is a glimpse of top 10 most common venue locations close to the neighborhoods

| DistanceToLib | Neighborhood | Latitude | Longitude | 1st Most Common Venue | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2.778660 | Regent Park, Harbourfront | 43.660323 | -79.362044 | Coffee Shop | Thai Restaurant | Restaurant | Pharmacy | Pub | Fast Food Restaurant | Food Truck | Sushi Restaurant | Auto Dealership | Indian Restaurant |
| 1.628925 | Garden District, Ryerson | 43.657658 | -79.378802 | Coffee Shop | Clothing Store | Café | Cosmetics Shop | Hotel | Middle Eastern Restaurant | Japanese Restaurant | Sandwich Place | Bubble Tea Shop | Pizza Place |
| 9.455135 | The Beaches | 43.667348 | -79.296693 | Beach | Park | Japanese Restaurant | Bakery | Café | Bar | Sandwich Place | Shoe Store | Salon / Barbershop | Pharmacy |
| 2.750241 | Berczy Park | 43.648100 | -79.375200 | Coffee Shop | Restaurant | Café | Italian Restaurant | Seafood Restaurant | Bakery | Hotel | Japanese Restaurant | Cocktail Bar | Gastropub |
| 1.629061 | Central Bay Street | 43.656101 | -79.383866 | Coffee Shop | Clothing Store | Hotel | Department Store | Diner | Bookstore | Middle Eastern Restaurant | Sushi Restaurant | Movie Theater | Bubble Tea Shop |
| 2.460269 | Christie | 43.664589 | -79.420675 | Korean Restaurant | Café | Grocery Store | Cocktail Bar | Mexican Restaurant | Indian Restaurant | Coffee Shop | Pub | Karaoke Bar | Ice Cream Shop |

Fig. 7 Sample data of 10 most common venues

## Normalization

Once all the data collected so far, were consolidated, the data was normalized using One-Hot methodology for Venue data to convert categorical data into numerical data, and the Distance to the nearest library was normalizing using a Standard Scaler, from scikit learn libraries.

Post normalization the data looks as shown below:

| | DistanceToLib | Yoga Studio | American Restaurant | Antique Shop | Aquarium | Arepa Restaurant | Art Gallery | Art Museum | [illegible] & Crafts Store | Asian Restaurant | Athletics & Sports | Auto Dealership | BBQ Joint | Bagel Shop | Bakery |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -0.101685 | 0.00 | 0.00 | 0.00 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.045455 | 0.00 | 0.00 | 0.000000 |
| 1 | -0.667467 | 0.00 | 0.00 | 0.00 | 0.0 | 0.0 | 0.010000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.00 | 0.00 | 0.010000 |
| 2 | 0.996554 | 0.00 | 0.00 | 0.00 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.00 | 0.00 | 0.041667 |
| 3 | -0.135269 | 0.01 | 0.01 | 0.01 | 0.0 | 0.0 | 0.020000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.01 | 0.01 | 0.040000 |
| 4 | -0.643933 | 0.00 | 0.00 | 0.00 | 0.0 | 0.0 | 0.011364 | 0.011364 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.00 | 0.00 | 0.000000 |

Fig. 8 Data sample post normalization

## Modeling

K-Means Clustering Machine Learning algorithm is suitable for scenarios involving clustering of data. In order to identify the optimal number of clusters, Elbow method is used. The maximum bend is at 3

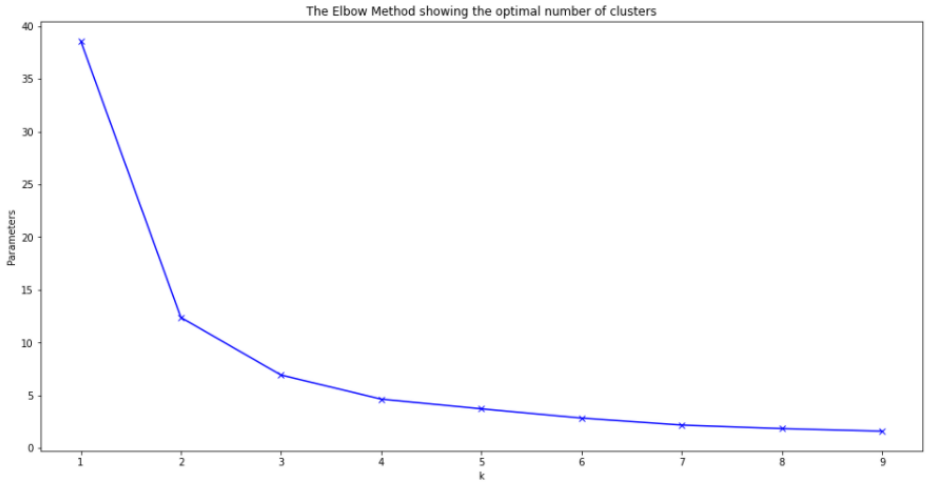


Fig. 9 Optimal K using Elbow method

Hence K-Means Clustering algorithm with 3 clusters was applied, to analyse the groups of similar locations.

The map of the candidate neighborhoods classified based on resulting clusters is as follows:

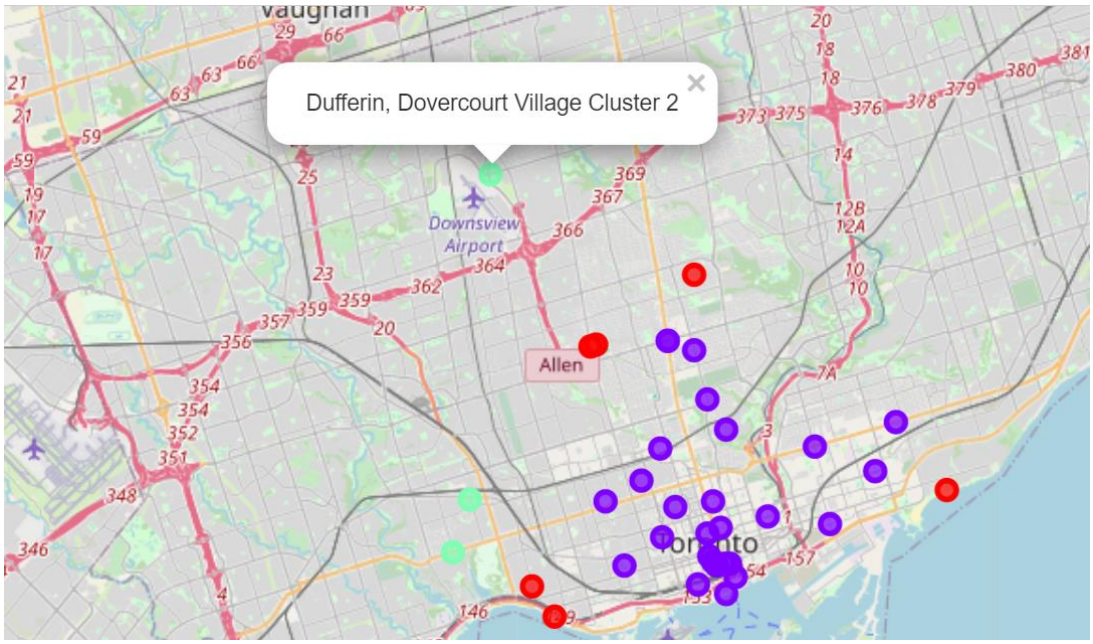


Fig. 10 Map of candidate neighborhoods clusters

## Evaluation

On analysis, the Clusters can be easily categorized based following features:

| Cluster | Distance to nearest library | Most common features |
|---|---|---|

| | | |
|---|---|---|
| 0 | 0 to 3 km | Coffee shop, café, bar |
| 1 | 3.5 to 6 km | Yoga Studio, American Restaurant, Moroccan restaurant |
| 2 | 7.5 to 10.5 km | Restaurant, Home Store |

# Results

Once the group of clusters were identified, each cluster was analysed for the nearby venue locations and the distance to the nearest library.

The locations where the distance to the nearest library was the high, with proximity to public commutation facilities such as Bus Station, Metro, etc., to enable easy of reach to the library for the common public, could be identified

# Discussion

The clusters have been classified by the similarity in venue locations as well as the distance to the nearest library.

# Conclusion

Based on the identified clusters, Dufferin, Dovercourt Village was found to be the most suitable location for Toronto Library, as it has longer distance to the nearest library and all public transport facilities such as Bus Station, Metro Station along with a few restaurants nearby.

| | DistanceToLib | Latitude | Longitude | 1st Most Common Venue | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 10.499267 | 43.750428 | -79.462787 | Metro Station | Playground | Bus Station | Escape Room | French Restaurant | Gym Pool | Gas Station | Men's Store | Furniture / Home Store | Ethiopian Restaurant |

The map of Canada shows the different neighborhood locations classified into their clusters marked by similar coloured marker. Dufferin, Dovercourt Village which is identified as the most suitable location is selected.

# Acknowledgement

This project has been implemented as practical experience of Data Science project implementation through the Coursera Capstone Project, for IBM Data Science Professional Certificate.

# References

Coursera – IBM Data Science Professional Certificate Courses

Towards Data Science (https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1)