Finding the most suitable location to open a new Indian Restaurant in Toronto, Canada.

Project Report

Submitted in fulfillment
for the degree of Bachelor of
Technology
In
Computer Science and Engineering and Information Technology

By

Kritika Bajaj (171466)

Under the supervision of Dr Jagpreet Sidhu

To

Department of computer science and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

Certificate

Candidate's Declaration

I hereby declare that the work presented in this report entitled "**Finding the most suitable location to open a new Indian Restaurant in Toronto, Canada.**" in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering/Information Technology submitted in the department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from July 2018 to May 2019 under the supervision of Dr. Jagpreet Sidhu (Assistant Professor, Senior Grade, Computer Science &Engineering Department).The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student Signature)
Kritika Bajaj (171466)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

(Supervisor Signature) Supervisor
Name: Dr. Jagpreet Sidhu
Designation: Assistant Professor (Senior Grade)
Department name: Computer Science and Engineering and Information Technology Dated:

ACKNOWLEDGEMENT

I want to take this chance to thank almighty for blessing me with his grace and taking my job to a successful culmination. We owe our profound gratitude to our project supervisor, Dr. Jagpreet Sidhu who took keen interest and guided us all along in my project work titled - "**Finding the most suitable location to open a new Indian Restaurant in Toronto, Canada.**" till the completion of our project by providing all the necessary information for developing the project. The project development helped us in research and we got to know a lot of new things in our domain. We are really thankful to him.

Table of content

Table of content

LIST OF FIGURES

List of graphs

ABSTRACT

I am creating a hypothetical situation for a idea that there won't be enough Indian Restaurants in Toronto Area. Therefore it is probably a outstanding opportunity for an entrepreneur who's primarily based totally in Canada. As the Indian food is famous amongst Asian community, so this entrepreneur may think about beginning its commercial enterprise in regions wherein asian community resides. With the cause in mind, locating the area to open such a restaurant is one of the maximum critical choices for this entrepreneur and this assignment is developed to assist him locate the maximum appropriate area.

CHAPTER 1: INTRODUCTION

1. Introduction

Toronto is one of the maximum densely populated regions in Canada. Being the land of possibility, it brings in loads of human beings from one-of-a-kind ethnic backgrounds to the center town of Canada, Toronto. Being the biggest town in Canada with an envisioned populace of over 6 million, there's absolute confidence approximately the diversity of the populace. Multiculturalism is visible thru the numerous neighbourhoods including; Chinatown, Corso Italia, Little India, Kensington Market, Little Italy, Koreatown and lots of more. Downtown Toronto being the hub of interactions among ethnicities brings many possibilities for marketers to begin or develop their enterprise. It is an area wherein human beings can attempt the first-class of each culture, both even as they paintings or simply passing thru. Toronto is properly known for its high-quality meals. Indian ethnic and organic meals has huge call for in Tornoto and is witnessing true growth. But there may now no longer be sufficient Indian Restaurants in Toronto Area. Therefore it is probably a high-quality possibility for an entrepreneur who's primarily based totally in Canada. As the Indian meals is famous amongst there, so this entrepreneur would possibly consider beginning its enterprise in regions wherein Asian community resides.

The goal of this challenge is to use Foursquare vicinity information and nearby clustering of venue data to decide what is probably the 'first-class' neighbourhood in Toronto to open a restaurant. We want to find places(Neighbourhood) which have a probably unfulfilled call for for Indian Restaurant. Also, we want places that have low opposition and aren't already crowded. We might additionally decide upon vicinity as near famous town Neighbourhood, assuming the first situations are met. We will use our information technological know-how powers to generate a few maximum promising neighbourhoods primarily based totally in this standards. Advantages of each region will then be actually Expressed in order that first-class feasible very last vicinity can be selected with the aid of using stakeholders. As vicinity choice is a multi-standards decision trouble and has a strategic significance for plenty restaurants.

Fig 1.1 Restaurant

The key problem for the eating place is a way to choose the ideal area due to the fact a desire made has several consequences at the inns destiny business. Therefore, the determination of the geographical web website online wherein a motel is to be placed is avery vital task for the investor due to the fact if the choice is appropriate, it'll store the prices of a relocation or reconfiguration and produce a myriad of advantages, such as: the shorter payback of the invested resources, a larger marketplace share, a better traveller satisfaction (which ends up in their loyalty, allows the operations of the motel, etc.). All of the foregoing factors out the reality that the area choice is a very vital strategic selection that calls for the buyers complete attention and careful making plans due to the fact destiny operation relies upon in this first step. So Through this project, we can discover the maximum appropriate area for an entrepreneur to open a brand new Indian eating place in Toronto, Canada.

1.2 Motivation

One of my uncle lives in Canada and he was thinking of opening an Indian restaurant there. But he is unable to decide the appropriate location for the project. So this gives me an idea of this project. Therefore, with this project we would like to develop a system in which we can predict the correct and appropriate location where this Indian restaurant can be build. With our project an enterpreneur can find the suitable location to open an Indian restaurant

1.3 Problem Statement

The objective of this capstone project is to find the most suitable location for the entrepreneur to open a new Indian Restaurant in Toronto,Canada. By using data science methods and tools along with machine learning algorithms such as clustering, this project aims to provide solutions to answer the business question : In Toronto, if an entrepreneur wants to open an Indian Restaurant, where should they consider opening it?

1.4 Objectives

The aim of our project are as follows:

1.Scrapping of data

1. Knowledge of machine learning

3. Develop a plan for project.

4. Create a developer account in folium for plotting map.

5. To identify the best machine learning algorithm to analyse.

6. Run the algorithms from scratch

1.5 Methodology

1.5.1 Scrapping of Toronto neighborhoods through Wikipedia.

1.5.2 Getting Latitude and Longitude information of those neighborhoods

1.5.3 Using Foursquare API to get venue information associated with these neighborhoods.

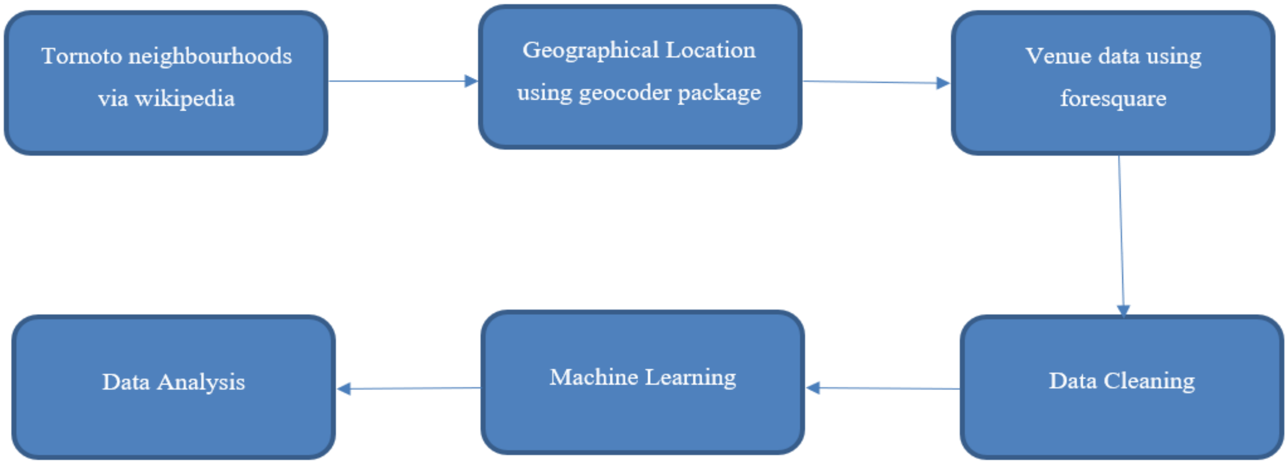
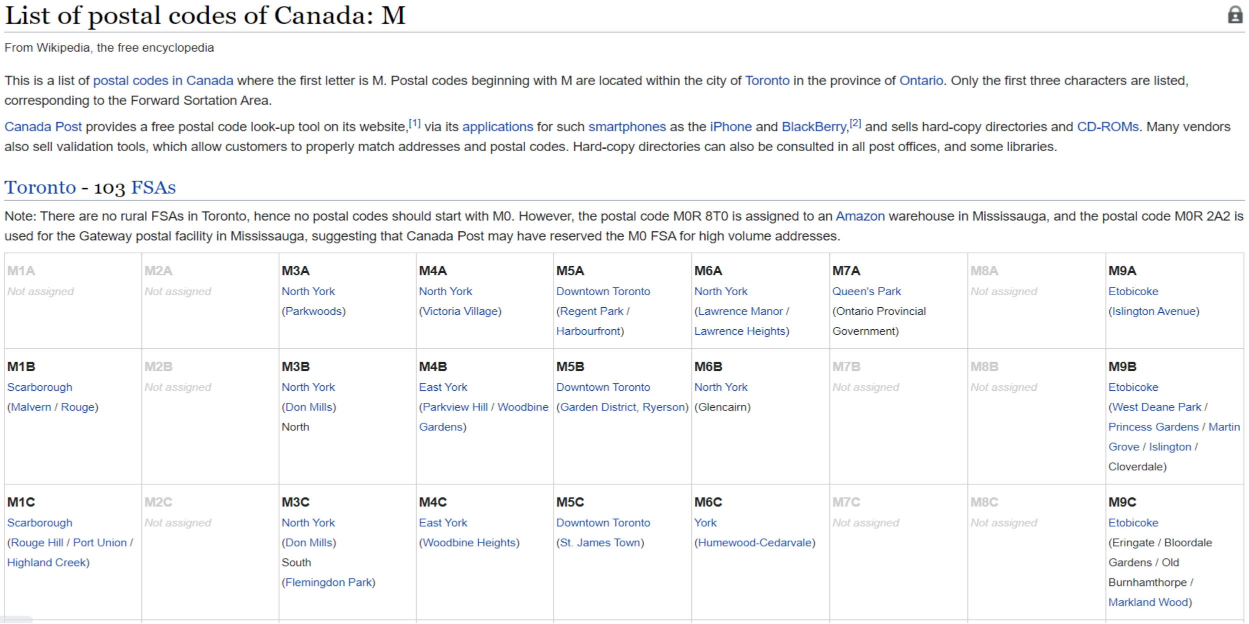

Tornoto neighbourhoods via wikipedia
Geographical Location using geocoder package
Venue data using foresquare
Data Cleaning
Machine Learning
Data Analysis

Fig 1.2 Flowchart for the process

1.5.1 Data acquisition

The Wikipedia incorporates a listing of postal codes in Canada where the primary letter is M. Postal codes starting with M are placed inside the town of Toronto. This Wikipedia site shown supplied nearly all the facts approximately the neighbourhoods. It blanketed the postal code, borough and the call of the neighbourhoods found in Toronto.



List of postal codes of Canada: M

From Wikipedia, the free encyclopedia

This is a list of postal codes in Canada where the first letter is M. Postal codes beginning with M are located within the city of Toronto in the province of Ontario. Only the first three characters are listed, corresponding to the Forward Sortation Area.

Canada Post provides a free postal code look-up tool on its website,[1] via its applications for such smartphones as the iPhone and BlackBerry,[2] and sells hard-copy directories and CD-ROMs. Many vendors also sell validation tools, which allow customers to properly match addresses and postal codes. Hard-copy directories can also be consulted in all post offices, and some libraries.

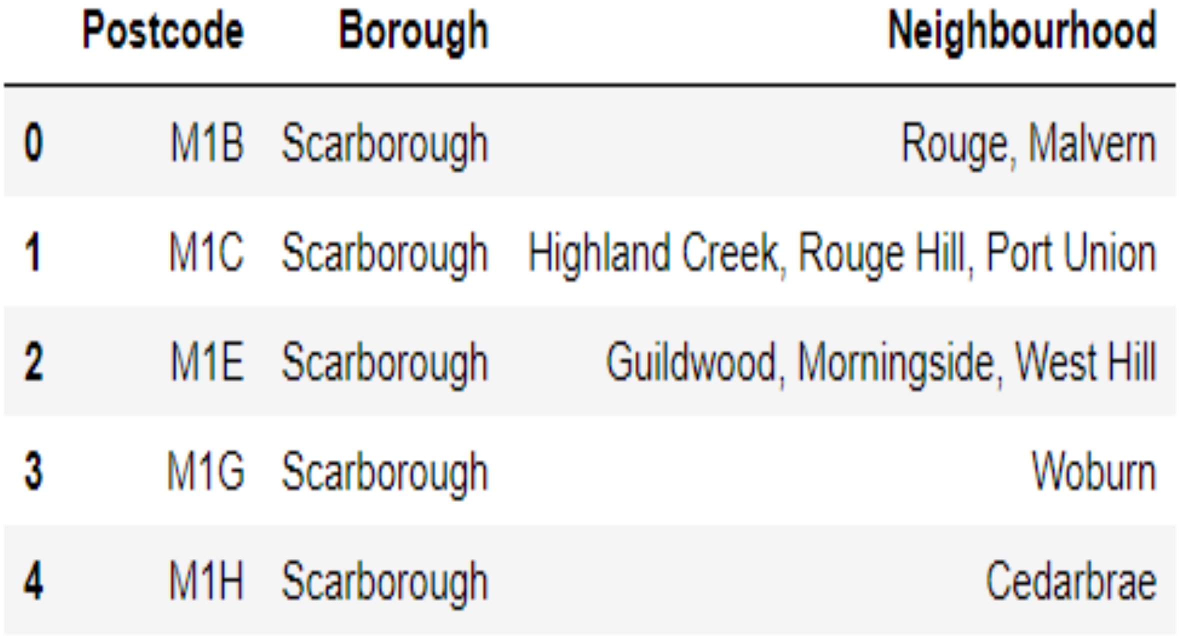
Toronto - 103 FSAs

Note: There are no rural FSAs in Toronto, hence no postal codes should start with M0. However, the postal code M0R 8T0 is assigned to an Amazon warehouse in Mississauga, and the postal code M0R 2A2 is used for the Gateway postal facility in Mississauga, suggesting that Canada Post may have reserved the M0 FSA for high volume addresses.

M1A *Not assigned*	M2A *Not assigned*	**M3A** North York (Parkwoods)	**M4A** North York (Victoria Village)	**M5A** Downtown Toronto (Regent Park / Harbourfront)	**M6A** North York (Lawrence Manor / Lawrence Heights)	**M7A** Queen's Park (Ontario Provincial Government)	M8A *Not assigned*	**M9A** Etobicoke (Islington Avenue)
M1B Scarborough (Malvern / Rouge)	M2B *Not assigned*	**M3B** North York (Don Mills) North	**M4B** East York (Parkview Hill / Woodbine Gardens)	**M5B** Downtown Toronto (Garden District, Ryerson)	**M6B** North York (Glencairn)	M7B *Not assigned*	M8B *Not assigned*	**M9B** Etobicoke (West Deane Park / Princess Gardens / Martin Grove / Islington / Cloverdale)
M1C Scarborough (Rouge Hill / Port Union / Highland Creek)	M2C *Not assigned*	**M3C** North York (Don Mills) South (Flemingdon Park)	**M4C** East York (Woodbine Heights)	**M5C** Downtown Toronto (St. James Town)	**M6C** York (Humewood-Cedarvale)	M7C *Not assigned*	M8C *Not assigned*	**M9C** Etobicoke (Eringate / Bloordale Gardens / Old Burnhamthorpe / Markland Wood)

Figure 1.3: Wikipedia Page showing List of Neighborhoods in Toronto with respective Postal Codes

Because the information was not in a suitable design for analysis, the information was scraped from this website.



	Postcode	Borough	Neighbourhood
0	M1B	Scarborough	Rouge, Malvern
1	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

Figure 1.4: Data scraped from website and put into Pandas data frame

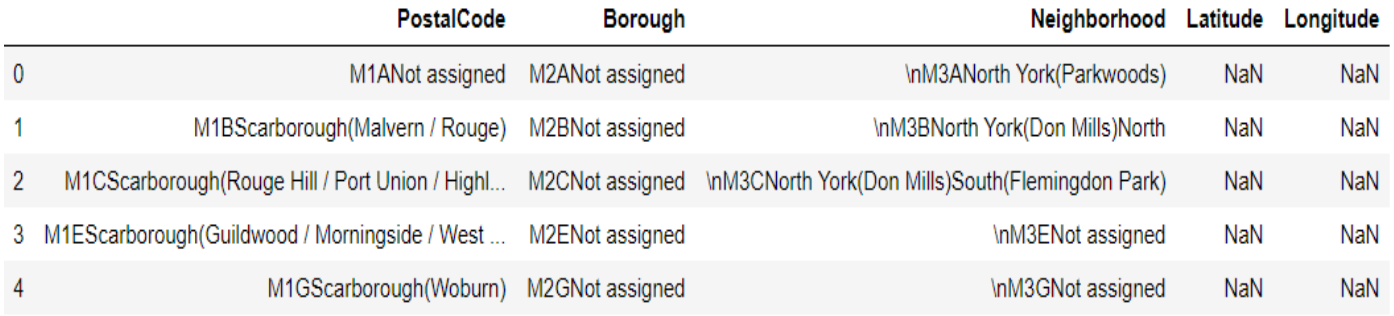
1.5.2 Getting Latitude and Longitude data of these neighborhoods

The second data source gave us the geographical coordinates of the districts with the respective postcodes. The file was in CSV format so we fixed it to a pandas data frame.

CSV File: Refers to the Comma Separated Value file, where each row or row has multiple fields separated by commas. We can classify each row as a row and each field as a column. You can read the csv files as given below:

coordinates = pd.read_csv("Geospatial_Coordinates.csv")

coordinates.head()



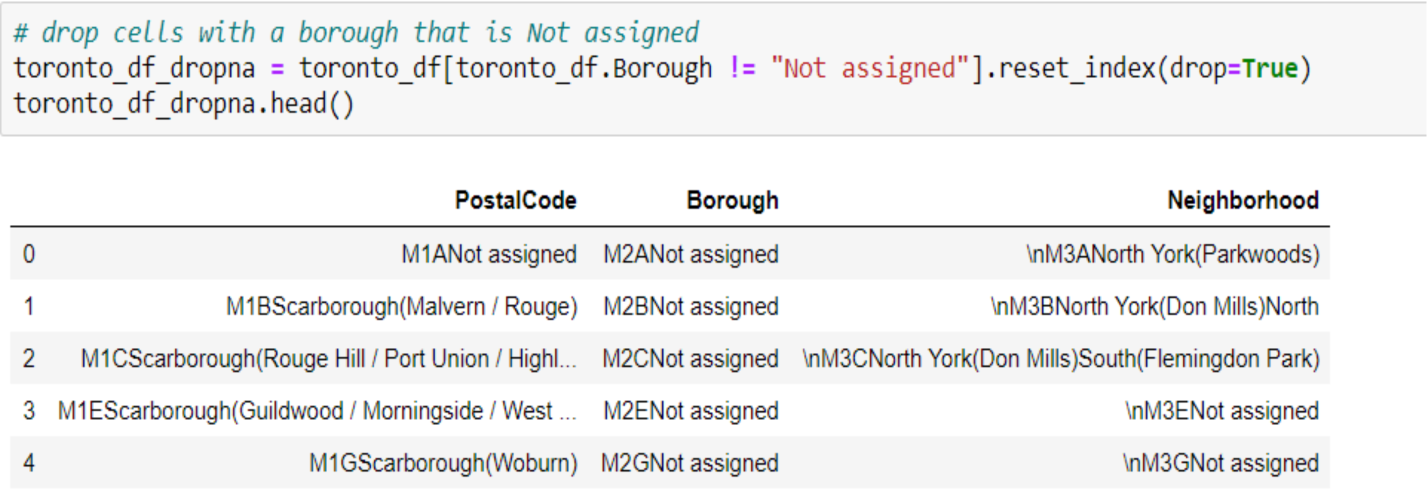
	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1ANot assigned	M2ANot assigned	\nM3ANorth York(Parkwoods)	NaN	NaN
1	M1BScarborough(Malvern / Rouge)	M2BNot assigned	\nM3BNorth York(Don Mills)North	NaN	NaN
2	M1CScarborough(Rouge Hill / Port Union / Highl...	M2CNot assigned	\nM3CNorth York(Don Mills)South(Flemingdon Park)	NaN	NaN
3	M1EScarborough(Guildwood / Morningside / West ...	M2ENot assigned	\nM3ENot assigned	NaN	NaN
4	M1GScarborough(Woburn)	M2GNot assigned	\nM3GNot assigned	NaN	NaN

Fig 1.5: Venue data pulled from Foresquare explore API

1.5.3 Data Cleansing

After all the information has been gathered and placed in data frames, it was necessary to cleanse and merge the data in order to start the analysis process.

1. Only cells that are assigned a Borough are processed and the left ones are deleted.

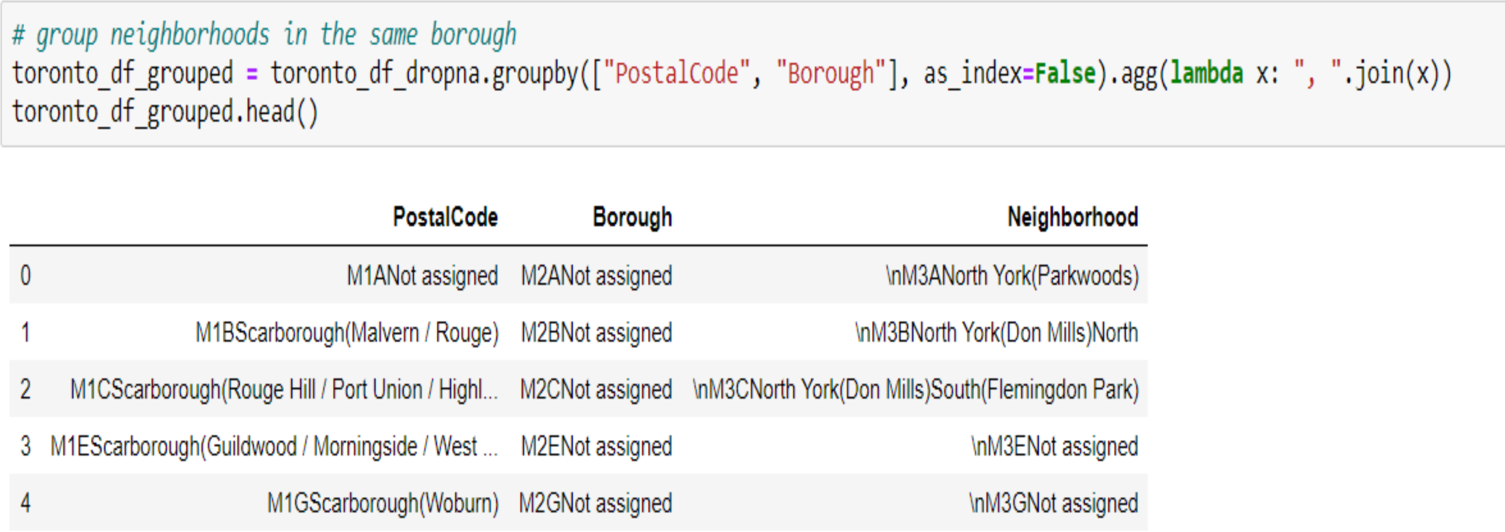


```
# drop cells with a borough that is Not assigned
toronto_df_dropna = toronto_df[toronto_df.Borough != "Not assigned"].reset_index(drop=True)
toronto_df_dropna.head()
```

	PostalCode	Borough	Neighborhood
0	M1ANot assigned	M2ANot assigned	\nM3ANorth York(Parkwoods)
1	M1BScarborough(Malvern / Rouge)	M2BNot assigned	\nM3BNorth York(Don Mills)North
2	M1CScarborough(Rouge Hill / Port Union / Highl...	M2CNot assigned	\nM3CNorth York(Don Mills)South(Flemingdon Park)
3	M1EScarborough(Guildwood / Morningside / West ...	M2ENot assigned	\nM3ENot assigned
4	M1GScarborough(Woburn)	M2GNot assigned	\nM3GNot assigned

Fig 1.6 Boroughs not assigned are deleted

2. There can be more than one neighborhood in a zip code area. These two lines are merged into one line, with the neighborhoods separated by a comma.

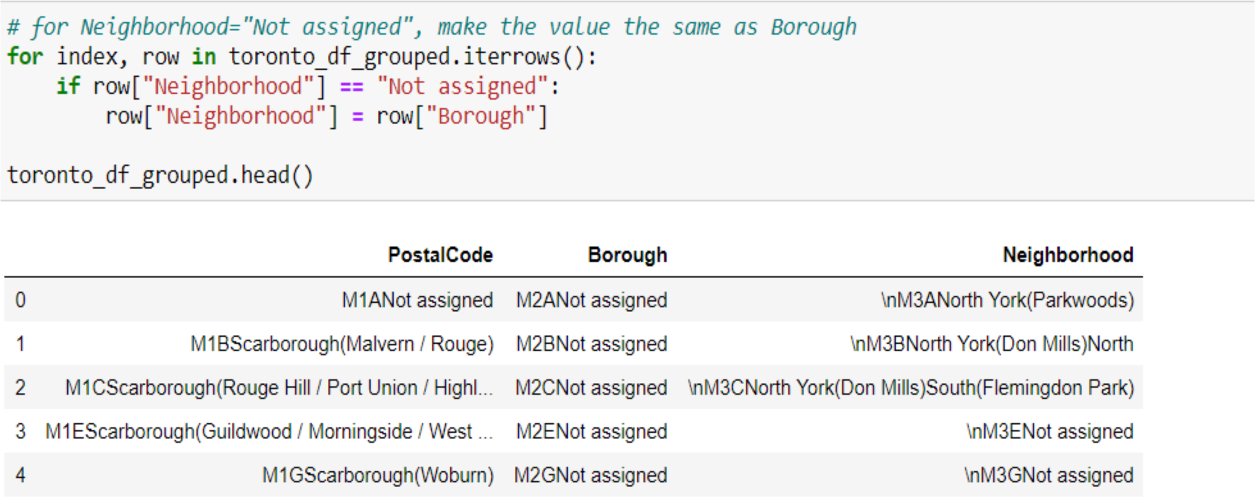


```
# group neighborhoods in the same borough
toronto_df_grouped = toronto_df_dropna.groupby(["PostalCode", "Borough"], as_index=False).agg(lambda x: ", ".join(x))
toronto_df_grouped.head()
```

	PostalCode	Borough	Neighborhood
0	M1ANot assigned	M2ANot assigned	\nM3ANorth York(Parkwoods)
1	M1BScarborough(Malvern / Rouge)	M2BNot assigned	\nM3BNorth York(Don Mills)North
2	M1CScarborough(Rouge Hill / Port Union / Highl...	M2CNot assigned	\nM3CNorth York(Don Mills)South(Flemingdon Park)
3	M1EScarborough(Guildwood / Morningside / West ...	M2ENot assigned	\nM3ENot assigned
4	M1GScarborough(Woburn)	M2GNot assigned	\nM3GNot assigned

Figure 1.7 lines having same neighbourhood are combined together

3. If a cell has a borough but an unassigned neighborhood, the neighborhood is the same as the borough.



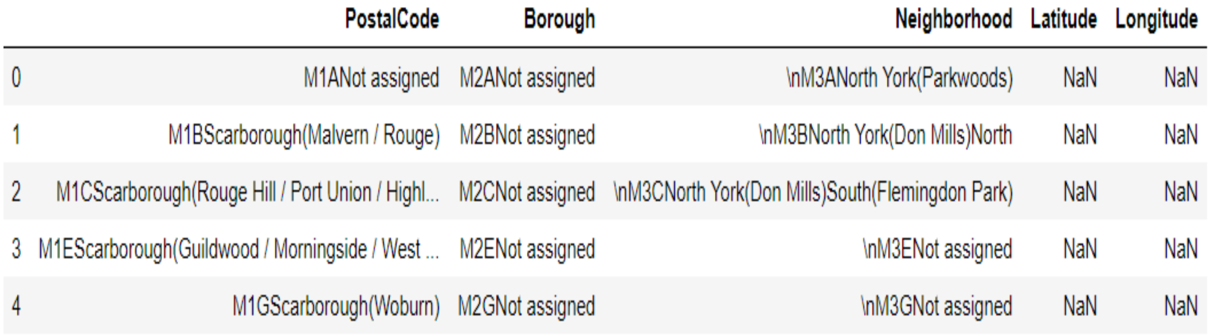
```
# for Neighborhood="Not assigned", make the value the same as Borough
for index, row in toronto_df_grouped.iterrows():
    if row["Neighborhood"] == "Not assigned":
        row["Neighborhood"] = row["Borough"]

toronto_df_grouped.head()
```

	PostalCode	Borough	Neighborhood
0	M1ANot assigned	M2ANot assigned	\nM3ANorth York(Parkwoods)
1	M1BScarborough(Malvern / Rouge)	M2BNot assigned	\nM3BNorth York(Don Mills)North
2	M1CScarborough(Rouge Hill / Port Union / Highl...	M2CNot assigned	\nM3CNorth York(Don Mills)South(Flemingdon Park)
3	M1EScarborough(Guildwood / Morningside / West ...	M2ENot assigned	\nM3ENot assigned
4	M1GScarborough(Woburn)	M2GNot assigned	\nM3GNot assigned

Fig 1.8 making unassigned neighbourhood which are same as borough

After the execution of the subsequent assumptions, the rows have been grouped primarily based totally on the borough as proven we merged the 2 tables collectively primarily based totally on Postal Code.

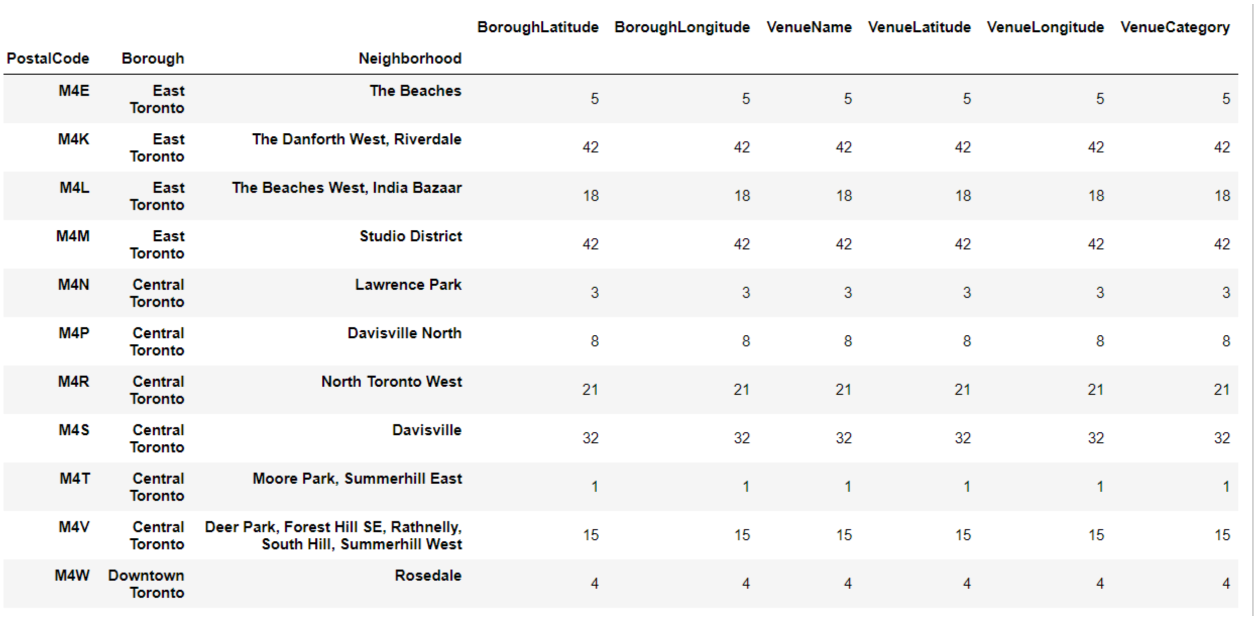


	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1ANot assigned	M2ANot assigned	\nM3ANorth York(Parkwoods)	NaN	NaN
1	M1BScarborough(Malvern / Rouge)	M2BNot assigned	\nM3BNorth York(Don Mills)North	NaN	NaN
2	M1CScarborough(Rouge Hill / Port Union / Highl...	M2CNot assigned	\nM3CNorth York(Don Mills)South(Flemingdon Park)	NaN	NaN
3	M1EScarborough(Guildwood / Morningside / West ...	M2ENot assigned	\nM3ENot assigned	NaN	NaN
4	M1GScarborough(Woburn)	M2GNot assigned	\nM3GNot assigned	NaN	NaN

Fig 1.9 Merging the two tables together

1.5.4 Using Foursquare API to get venue data related to these neighborhoods

Now venue facts is pulled from foresquare API which Schools, Parks, Café Shops,hotels, Restaurants etc. Getting this facts changed into essential to reading the variety of Indian Restaurants all over Toronto.

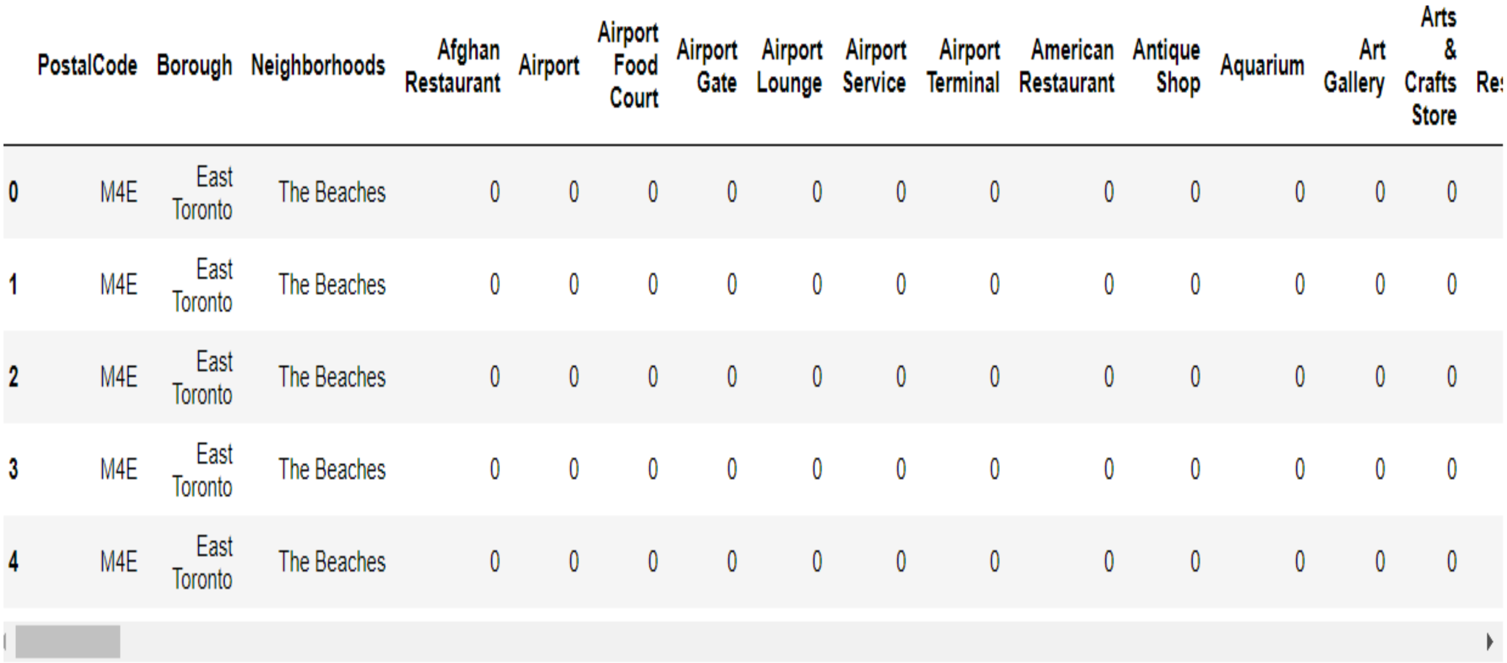


			BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
PostalCode	Borough	Neighborhood						
M4E	East Toronto	The Beaches	5	5	5	5	5	5
M4K	East Toronto	The Danforth West, Riverdale	42	42	42	42	42	42
M4L	East Toronto	The Beaches West, India Bazaar	18	18	18	18	18	18
M4M	East Toronto	Studio District	42	42	42	42	42	42
M4N	Central Toronto	Lawrence Park	3	3	3	3	3	3
M4P	Central Toronto	Davisville North	8	8	8	8	8	8
M4R	Central Toronto	North Toronto West	21	21	21	21	21	21
M4S	Central Toronto	Davisville	32	32	32	32	32	32
M4T	Central Toronto	Moore Park, Summerhill East	1	1	1	1	1	1
M4V	Central Toronto	Deer Park, Forest Hill SE, Rathnelly, South Hill, Summerhill West	15	15	15	15	15	15
M4W	Downtown Toronto	Rosedale	4	4	4	4	4	4

Fig 1.10 Venues near respective neighbourhoods

1.5.5 Machine Learning

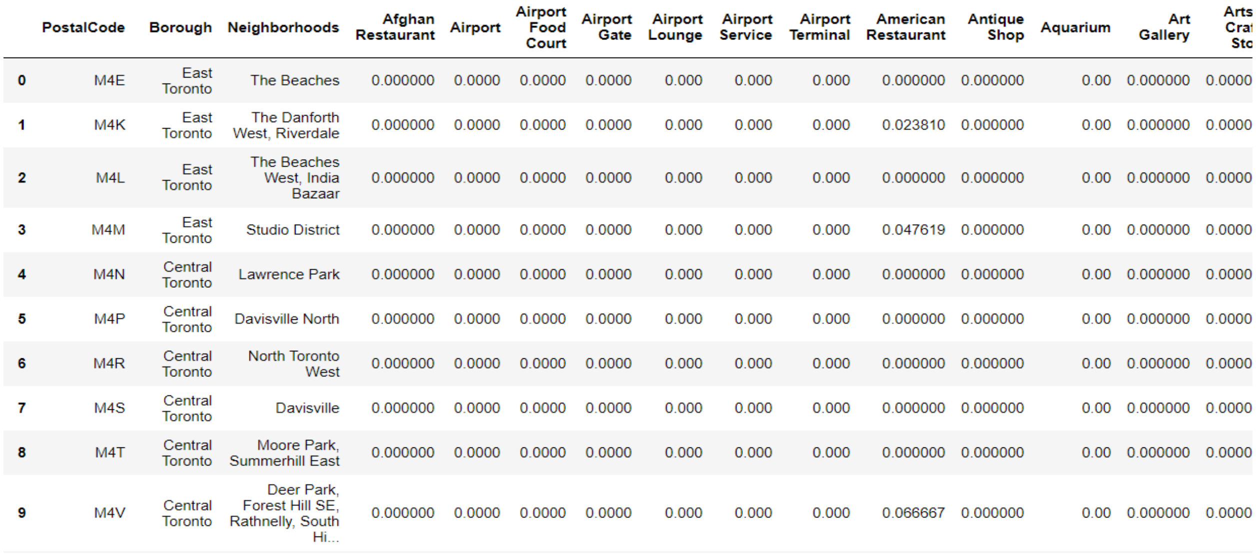
At that point to explore the realities we achieved a technique wherein Categorical Data is changed over into Numerical Data for Machine Learning calculations. This strategy is called One hot encoding. For everything about neighbourhoods, singular settings had been changed into the recurrence at what number of the ones Venues had been situated around there.



	PostalCode	Borough	Neighborhoods	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Arts & Crafts Store	Re
0	M4E	East Toronto	The Beaches	0	0	0	0	0	0	0	0	0	0	0	0	
1	M4E	East Toronto	The Beaches	0	0	0	0	0	0	0	0	0	0	0	0	
2	M4E	East Toronto	The Beaches	0	0	0	0	0	0	0	0	0	0	0	0	
3	M4E	East Toronto	The Beaches	0	0	0	0	0	0	0	0	0	0	0	0	
4	M4E	East Toronto	The Beaches	0	0	0	0	0	0	0	0	0	0	0	0	

Fig 1.11 One Hot Encoding

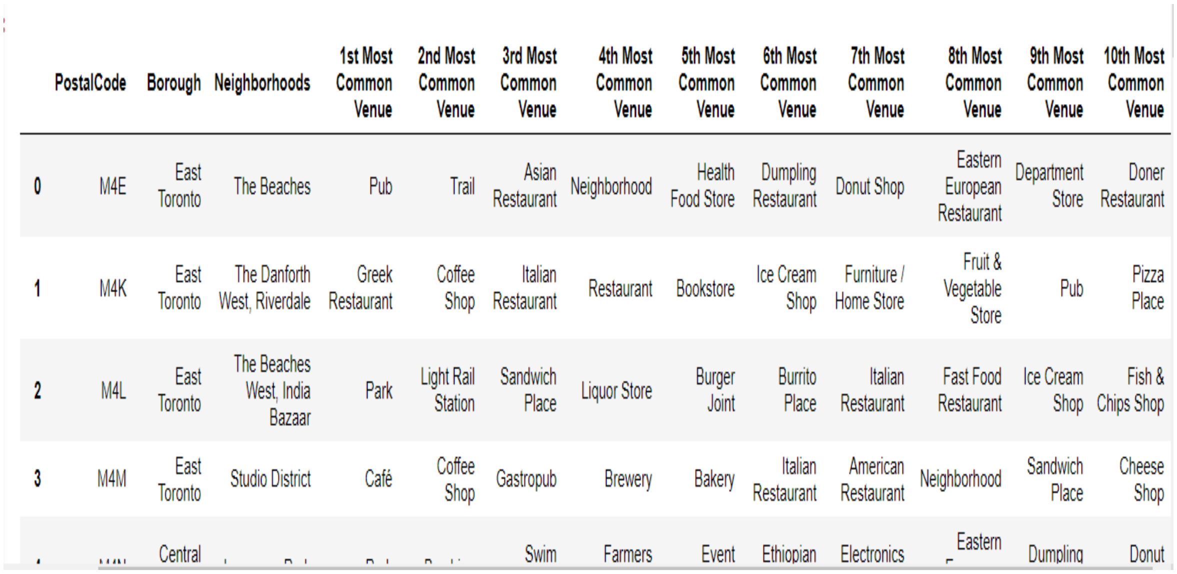
After this the rows are grouped by Neighborhood and by taking the average of the frequency of occurrence of each Venue Category.



	PostalCode	Borough	Neighborhoods	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Arts Cra Sto
0	M4E	East Toronto	The Beaches	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	0.00	0.000000	0.0000
1	M4K	East Toronto	The Danforth West, Riverdale	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.023810	0.000000	0.00	0.000000	0.0000
2	M4L	East Toronto	The Beaches West, India Bazaar	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	0.00	0.000000	0.0000
3	M4M	East Toronto	Studio District	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.047619	0.000000	0.00	0.000000	0.0000
4	M4N	Central Toronto	Lawrence Park	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	0.00	0.000000	0.0000
5	M4P	Central Toronto	Davisville North	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	0.00	0.000000	0.0000
6	M4R	Central Toronto	North Toronto West	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	0.00	0.000000	0.0000
7	M4S	Central Toronto	Davisville	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	0.00	0.000000	0.0000
8	M4T	Central Toronto	Moore Park, Summerhill East	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.000000	0.000000	0.00	0.000000	0.0000
9	M4V	Central Toronto	Deer Park, Forest Hill SE, Rathnelly, South Hi...	0.000000	0.0000	0.0000	0.0000	0.000	0.000	0.000	0.066667	0.000000	0.00	0.000000	0.0000

Fig 1.12 Grouped neighbourhoods by taking average of frequency of each venue

After this a new data frame has been created which save the Neighborhood names in addition to the imply number of Indian Restaurants in that Neighborhood. This acknowledge the information to be sum up primarily based totally on every single Neighborhood and made the information lots easier to examine.



	PostalCode	Borough	Neighborhoods	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
0	M4E	East Toronto	The Beaches	Pub	Trail	Asian Restaurant	Neighborhood	Health Food Store	Dumpling Restaurant	Donut Shop	Eastern European Restaurant	Department Store	Doner Restaurant
1	M4K	East Toronto	The Danforth West, Riverdale	Greek Restaurant	Coffee Shop	Italian Restaurant	Restaurant	Bookstore	Ice Cream Shop	Furniture / Home Store	Fruit & Vegetable Store	Pub	Pizza Place
2	M4L	East Toronto	The Beaches West, India Bazaar	Park	Light Rail Station	Sandwich Place	Liquor Store	Burger Joint	Burrito Place	Italian Restaurant	Fast Food Restaurant	Ice Cream Shop	Fish & Chips Shop
3	M4M	East Toronto	Studio District	Café	Coffee Shop	Gastropub	Brewery	Bakery	Italian Restaurant	American Restaurant	Neighborhood	Sandwich Place	Cheese Shop
		Central				Swim	Farmers	Event	Ethiopian	Electronics	Eastern	Dumpling	Donut

Fig 1.13 Different common venues in each neighbourhoods

1.5.6 K-means clustering algorithm

Clustering is one of the maximum not unusual place exploratory statistics evaluation method used to get an instinct approximately the shape of the statistics. It may be described because the undertaking of figuring out subgroups withinside the statistics such that statistics factors withinside the equal subgroup (cluster) are very comparable even as statistics factors in special clusters are exceptional. In various words, we endeavor to find homogeneous subgroups in the measurements to such an extent that insights factors in each bunch are pretty much as tantamount as practical in accordance with a similitude degree, for example, euclidean-essentially based absolutely distance or connection principally based thoroughly distance. The decision of which comparability degree to apply is application-explicit.

Clustering evaluation may be completed on the idea of functions wherein we attempt to discover subdivision of samples primarily based totally on functions or on the idea of samples wherein we try and discover subgroups of functions primarily based totally on samples. Unlike supervised gaining knowledge of, clustering is taken into consideration an unmonitored gaining knowledge of approach due to the fact we don't have the floor reality to evaluate the output of the clustering set of rules to the actual labels to assess its performance.

Kmeans set of rules is an iterative arrangement of decides that endeavors to segment the dataset into Kpre-depicted marvelous non-covering subgroups (bunches) wherein each measurement factor has a place with best one gathering. It endeavors to make the intra-bunch measurements factors as tantamount as achievable even as also keeping the groups as unique (far) as attainable. It doles out measurements components to a group with the end goal that the amount of the squared distance among the insights factors and the bunch's centroid (science recommend of the entirety of the measurements factors that have a place with that group) is on the base. The significantly less form we've inside bunches, the extra homogeneous (similar) the measurements factors are in the equivalent group.

The way kmeans set of rules works is as per the following:

1. Determine amount of groups K.

2. Introduce centroids with the guide of utilizing first rearranging the dataset after which haphazardly choosing K insights factors for the centroids with out substitution.

3. Continue to emphasize till there might be no substitute to the centroids. i.e challenge of insights elements to groups isn't evolving.

4 Compute the amount of the squared distance among measurements components and all centroids.

5 Assign each measurement factor to the closest group (centroid).

6 Compute the centroids for the groups with the guide of utilizing taking the normal of the all insights factors that have a place with each bunch.

The strategy kmeans follows to cure the problem is called Expectation-Maximization. The E-step is relegating the insights components to the closest bunch. The M-venture is registering the centroid of each bunch.

The objective function is:

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2 \tag{1}$$

where wik = 1 for the information point xi on the off chance that it has a place with bunch k; in any case wik = 0. What's more, μk is the focal point of mass of the gathering of xi. It's a two section minimization issue. First we limit J w.wik and treat μk as fixed. At that point we limit J w.μk and treat fixed wik. In fact we recognize J w.wik first and update the bunch mappings (step E). At that point we separate J w.μk and recalculate the focuses of gravity as indicated by the group tasks from the past advance (venture M). Thus, step E is:

$$\begin{aligned} \frac{\partial J}{\partial w_{ik}} &= \sum_{i=1}^{m} \sum_{k=1}^{K} \|x^i - \mu_k\|^2 \\ &\Rightarrow w_{ik} = \begin{cases} 1 & \text{if } k = argmin_j \|x^i - \mu_j\|^2 \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2}$$

At the end of the day, allocate the information guide xi toward the nearest bunch decided by its amount of squared separation from group's centroid.

And M-step is:

$$\frac{\partial J}{\partial \mu_k} = 2\sum_{i=1}^{m} w_{ik}(x^i - \mu_k) = 0$$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{m} w_{ik}x^i}{\sum_{i=1}^{m} w_{ik}} \tag{3}$$

Which interprets to recomputing the centroid of every cluster to reflect the brand new assignments.

1. Since clustering algorithms consisting of kmeans use distance-primarily based totally measurements to decide the similarity among facts points, it's advocated to standardize the facts to have an average of 0 and a trendy deviation of 1 on the grounds that almost continually the functions in any dataset could have one-of-a-kind gadgets of measurements inclusive of age vs income.
2. Given kmeans iterative nature and the random initialization of centroids on the begin the set of rules, one-of-a-kind initializations might also additionally result in one-of-a-kind clusters on the grounds that kmeans set of rules might also additionally caught in a nearby most useful and might not converge to international most useful. Therefore, it's advocated to run the set of rules the usage of one-of-a-kind initializations of centroids and choose the effects of the run that that yielded the decrease sum of squared distance.
3. Allocatement of examples isn't converting is the identical issue as no alternate in within-cluster variation:

$$\frac{1}{m_k}\sum_{i=1}^{m_k} \left\|x^i - \mu_{c^k}\right\|^2 \tag{4}$$

K-Means Clustering

To make the investigation seriously intriguing, we needed to bunch the areas dependent on the neighborhoods that have comparable Italian cafés by and large. To do this, we utilize the gathering of K-reserves. We utilized the ElbowPoint strategy to get our ideal K-esteem, which was neither over-nor mistakenly adjusted to the model. In this method, we run a test on an alternate number of K qualities, measure the precision, and afterward pick the best one.K-esteem The best K-esteem is picked at where the line shows the most honed bend. Also, K-implies gathering puts protests that are comparable dependent on a specific variable into a similar gathering. The Indian cafés were partitioned into 4 gatherings. Every one of these gatherings has been named from 0 to 3 on the grounds that the names are recorded beginning with 0 rather than 1.

Then, we consolidated the area's information with the table above and made another table that frames the reason for breaking down new freedoms to open another Indian café in Toronto. Next we made a guide utilizing the Folium bundle in Python and every area was hued dependent on the group.

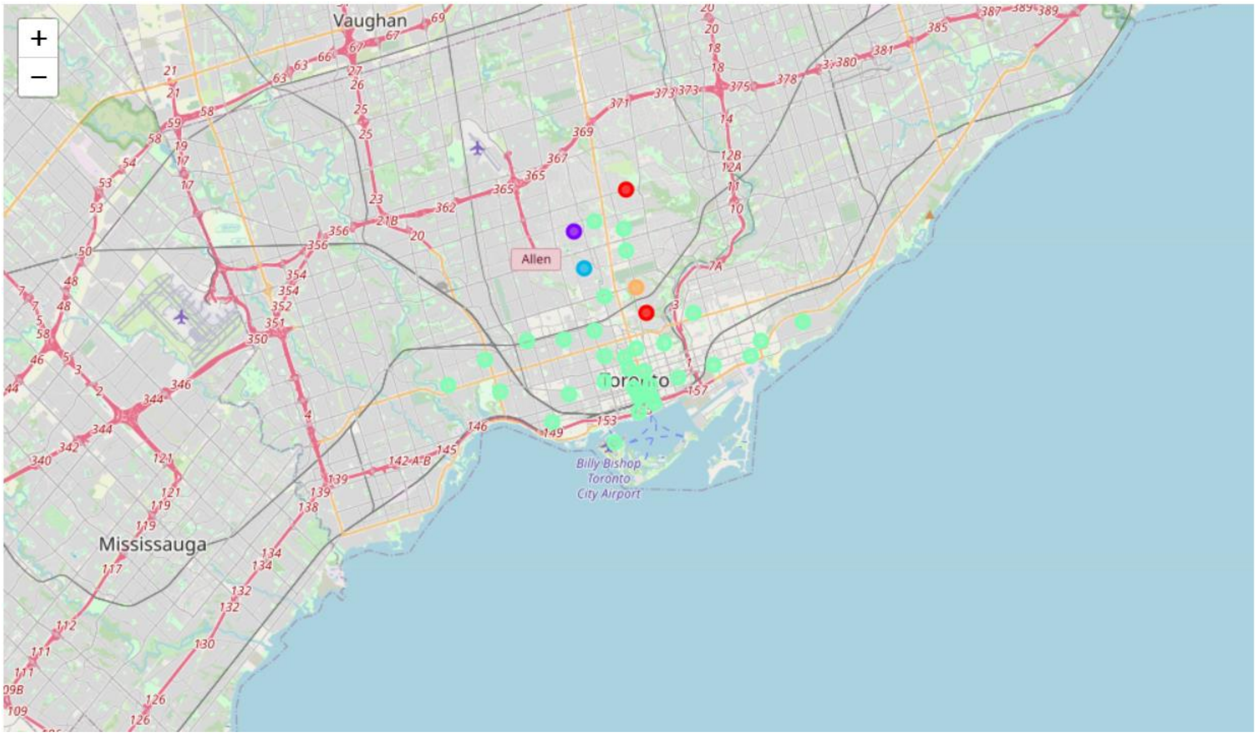


Fig 1.14 Map representing different clusters

1.6 Implementation of Different Algorithm from Scratch

Some common techniques which is applied to almost all algorithm in their implementation:

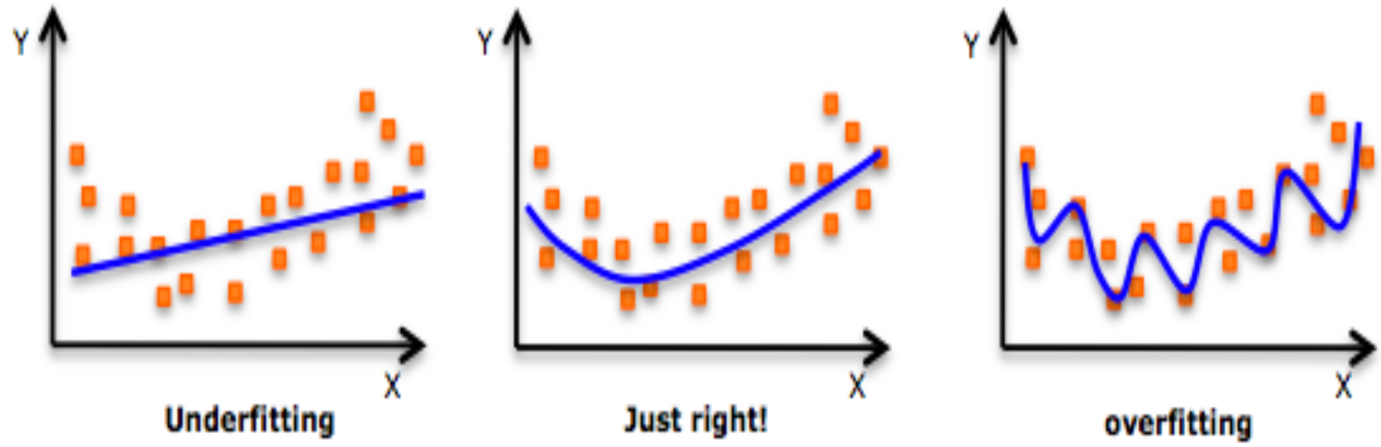
Feature selection on Dataset In some cases it is possible that there are a large number of features in our dataset and we need to delete some of irrelevant features. But random deletion of any important feature can lead to high probability of errors in our final result thus decreasing the accuracy of our algorithms .So, we use this method of feature selection which selects the best feature according to their contribution in prediction.

SelectKBest is an in-built feature of Skicit Learn used to sort the features according to the score generated from the most important feature to the least important feature.

Our data-set can be sometimes classified as:

Under-fitted Model: This means that the data is sufficient under the training data set and is often missed in the data set used, so we are not able to get the correct results.

Over-fitted Model: This means that the data is trained really well and is fitted accurately according to the training data set that it will not give accurate result for the test data set as it is trained too closely to train data set.



Graph 1.1: Mode

1.6 Organization

In Chapter 2, literature review is discussed, that contains the key terms, value and functionality of algorithms and their types. We have discussed various algorithms and their pros and cons so that we can identify the best algorithms for a health dataset.

In Chapter 3, we improved the system architecture and enter all the program requisites so that we can use the algorithms and these were tested in the environment to analyze the information for improved results.

In Chapter 4, all the algorithms were discussed ,used and the statistics or composition behind these algorithms to understand better and considered how those algorithms are used. We also considered the parameter we compared algorithms.

In Chapter 5, we then concluded with the report and included it in the project reading list. We have also discussed the scope and future outcome of the project.

Chapter-2

Literature Review

2.1 Terminologies

This section discusses the various terminologies published in literature.

2.1.1 Data Science: Data Science =Data+Science

The fields of bringing out insights from data using scientific techniques is called Data Science. Any scientific method applied to data to extract benefits would be part of data science.

2.1.1.2 Spectrum of Buissness analysis

MIS: MIS are used to track what is happening in an organisation. Benefit of MIS to an organisation is typically low.

Detective Analysis: Compared to MIS Detective Analysis is more complex but it also adds more value to an organisation.

Dashboarding: Dashboards are created in real time. They are used to answer what is happening in a buissness.

Predictive Modelling: To predict what is likely to happen in a granular level. For example. In case of bank which person is likely to get default in next 30 days.

Big Data: Complexity of handling the data goes beyond the traditional systems. This could be because of increased Volume, Variety and Velocity.

2.1.2 Why Data Science?

There will be 50 Bn devices connected to Internet by 2020. These devices are laptops , smartphones, watches and smart devices. What used to be cost millions of dollars in 80. Now cost a few scents we can store this data cheaply. The computational cost is falling. In essence-we are creating and storing data at humongous scale and can run computation on it in very low costs.

Example:

Recent store amazon go.

No lines, No checkout.

Swipe phone while entering store. Pick out what you want and then just go. As soon as you walk off the store you will be charged with respective items on your account. With the help of data science , Deep Learning and computer Vision.

2.1.3 Different prerequisites for Data Science:

Forecasting: is a process of predicting or estimating the future based on past and present data. For example: How many passangers can we expect in a given flight?

How many customer call can we expect in next hours?

Predictive Modelling: use to perform prediction more granular like “who are the customers who are likely to buy a product in next month? And then act accordingly.

Machine Learning: Method of teaching machines to learn things and improve predictions/behaviour based on data on their own.

Example: Create an algo which can power google search.

Amazon recommendation system.

2.1.4 Machine Learning: Machine learning explores the algorithm that learn or built models from the historical data and these models can be used for different tasks. Example prediction, decision making .Which helps the algorithms to predict output when they receive the input of the same domain. It basically learns the similarity pattern between the inputs by which the algorithm is trained and imply a output from the test dataset’s input.

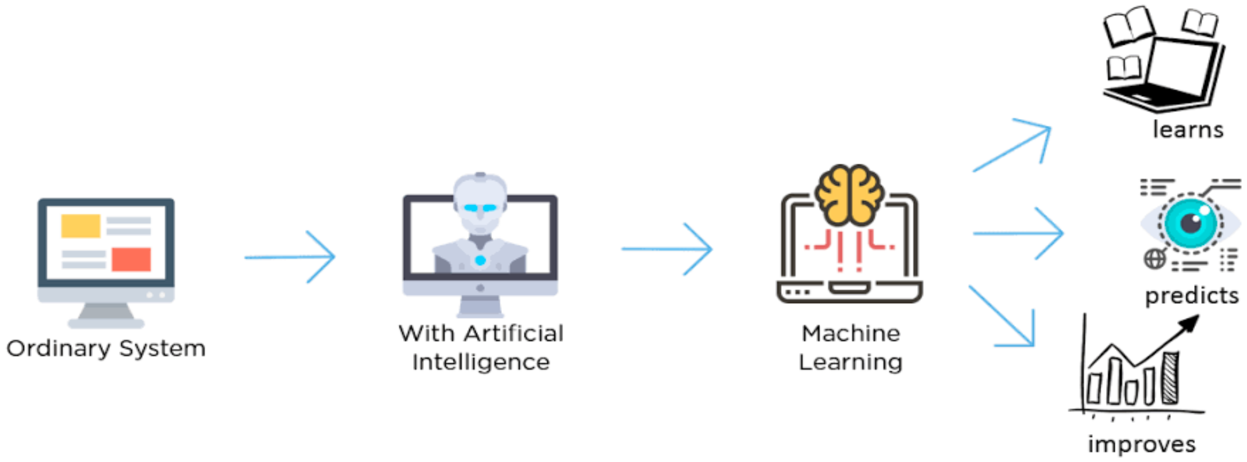


Fig 2.1: Machine Learning

2.1.5 Working of Machine learning?

The machine learning procedure starts from the gathering of information from a dataset of a selected table type. In addition, there are many algorithms we can use. Therefore, the next step is to choose an algorithm to use. We then split the dataset into a specific two-dimensional scale: Rail and Test (70% Train details and 30% test data). Then we train the algorithm and train data and keep it ready for use. The last step is to provide the input we want to test the algorithm and the algorithm will give us the answer in sequence.

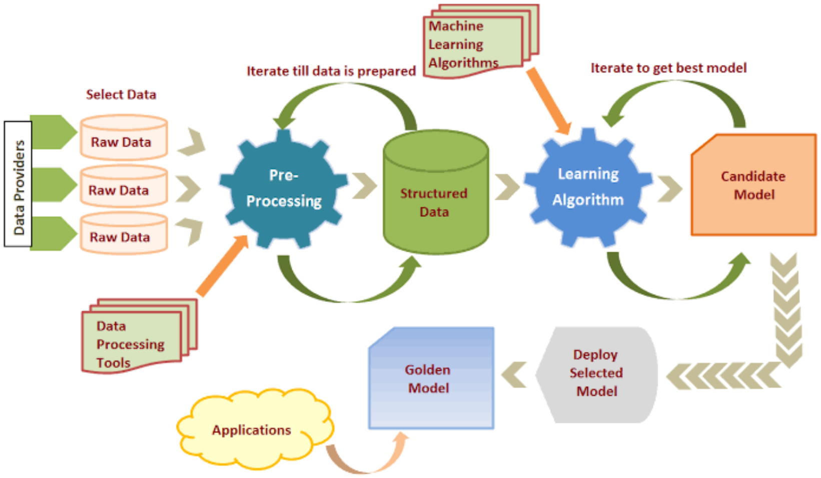


Fig 2.2: Working of machine learning

2.1.6 Types of machine learning

1.Supervised Machine Learning: This kind of computation applies to a dataset already trained by past results and previous results using labeled data to predict the effect of latest information. In this instance the known database is analyzed, the algorithm simply makes a linear expression which helps in predicting the outcomes of the new information. Also it can analyze data and results and differentiate past saved information to detect errors and be able to do modification and train the model appropriately.

2. Unsupervised machine learning: this sort of algorithm varies from supervised machine learning algorithm as these algorithms are used when the model isn't trained before neither it's classified nor it's checked. Unsupervised learning algorithms make the system a hidden structure or pattern within the unlabelled dataset and predict possible results with the utilization of such patterns while removing the outliers.

4. **Reinforcement Machine Learning:** the most idea of reinforcement learning is that it's reward based training during which the model interacts with the environment by doing actions and discovering errors or rewards. the foremost relevant characteristics of reinforcement learning are the trial and error and delayed reward. during this case the model learns from its mistakes or errors and therefore the new model is formed to interact with the machines to automatically determine the result and the ideal behaviour to reinforce the working and for performance optimization.

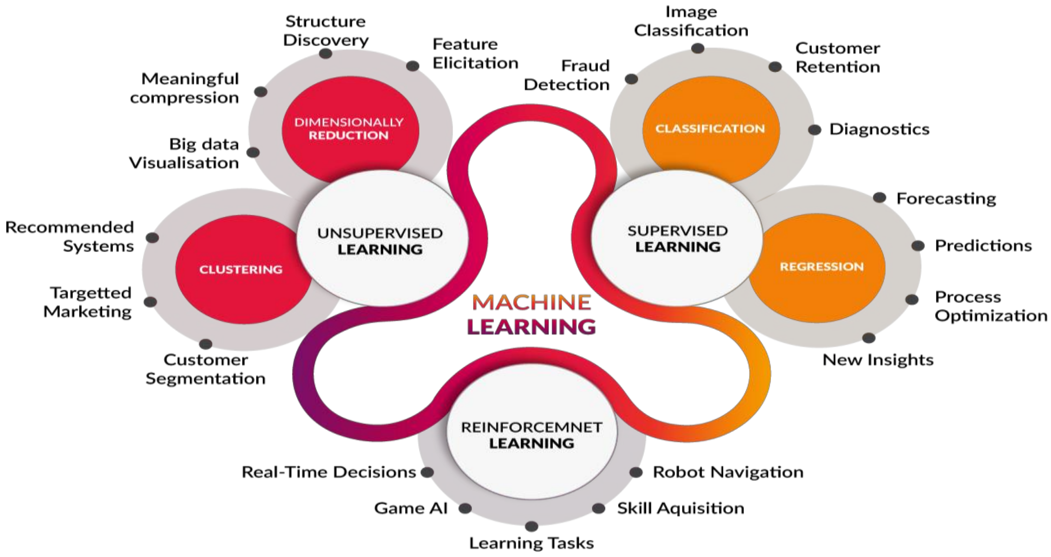


Fig 2.3: Types of machine learning

Chapter-3

System Development

3.1 System Requirements

The algorithms utilized in this project require some standard programming **because it** requires the processing of algorithms.

- Windows 10 (64-bit)
- Jupyter Notebook
- Python
- 4 GB RAM
- Intel (R) Core (TM) i3 processor

3.2 Why Python?

Python is a programming language that features a sizable amount of viewers and is precise and precise. In addition, python offers an assortment of bundles that make the worst of tasks or tasks difficult. Python has libraries in the records used for example - working with images, working with content or working with audio records. In any case, when working with another OS, python is perfect. Python is a great network that makes it easy to look for help and tips and tricks.

3.3 Scikit Learn

Scikit read Python library is frequently used for machine learning and is in a position to include various returns, classifications and compilation algorithms.

3.4 The Pandas

Pandas is an open-source, easy-to-use data structures and data analysis tools for the Python programing language . Python with Pandas is employed during a wide selection of fields including academic and commercial domains.

3.5 Numpy

NumPy is a Python package that stands for 'Numerical Python'. It is a key library of computer science, containing an arr-n-dimensional object, providing tools for integrating C, C ++ etc. Also useful in algebraic order, random number power etc. is a container of various sizes of standard data.

3.6 Anaconda

Anaconda is a standard Python data science platform, leading to an open source machine learning environment.

3.7 Matplotlib

Matplotlib is a Python programming library that provides an API-based application for embedding sites into applications. It's very similar to MATLAB embedded in Python programming language. Histogram, bar plots, streaming plots, pie plot area, Matplotlib can show a wide range of observations. With a little effort and a tint of visual skills, with Matplotlib, we can create any observation

3.8 Jupyter Notebook

Jupyter Notebook is a natural computational environment for building Jupyter scripts. It is a document that follows a modified structure, and contains an ordered list of input / output cells that can contain code, text, statistics, sites and rich media. It usually ends with the ".ipynb" extension.

Chapter-4

Performance analysis

4. Data Analysis

We have a total of 5 clusters (0,1,2,3,4).

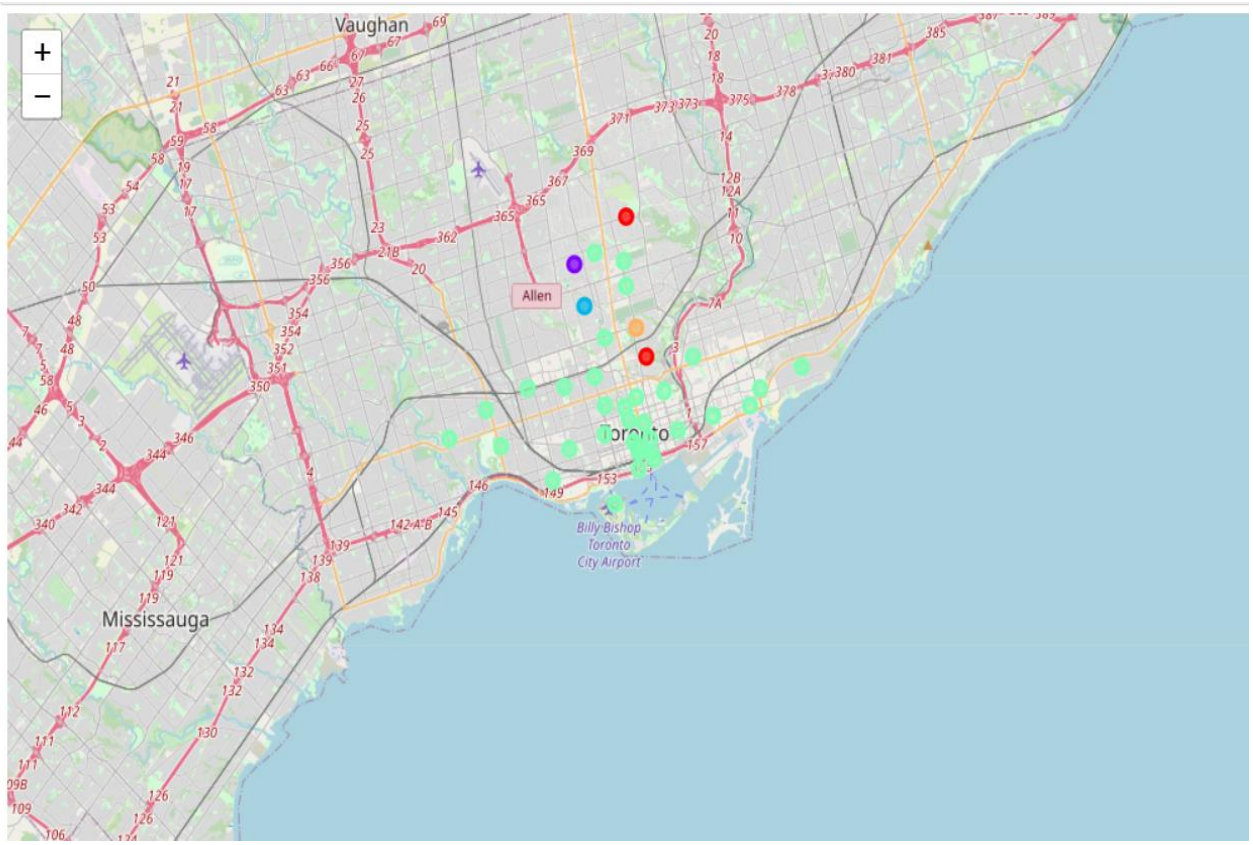
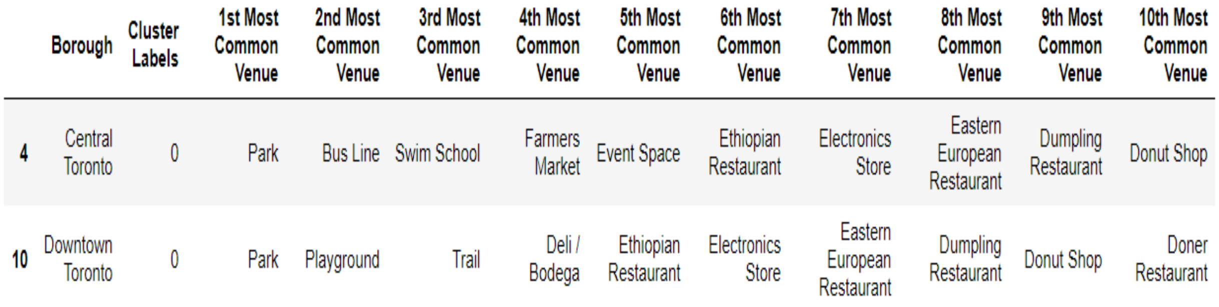


Fig 4.1 Map representing different clusters

A map is made utilizing the Folium package in Python and every neighbourhood was shaded dependent on the group mark. The map shows the various groups that had a comparative mean recurrence of Indian eateries.

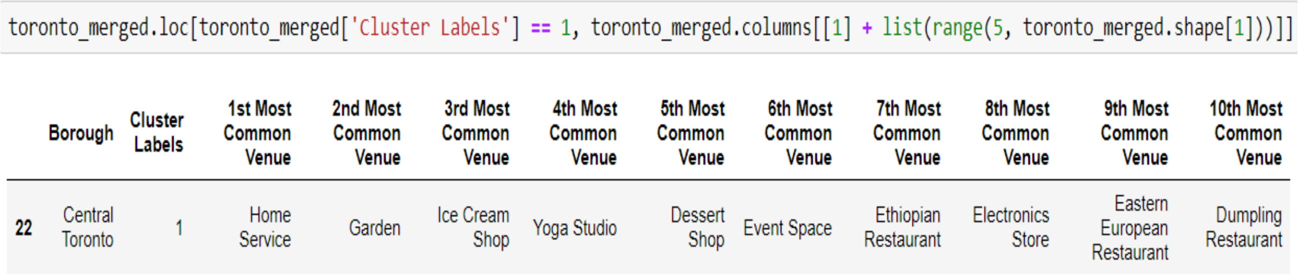
Cluster 1



	Borough	Cluster Labels	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
4	Central Toronto	0	Park	Bus Line	Swim School	Farmers Market	Event Space	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop
10	Downtown Toronto	0	Park	Playground	Trail	Deli / Bodega	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant

Cluster 1 was in the Central Tornoto and Downtown Tornoto area. Cluster 1 which are mostly business areas with Park, Bus Line, Swim School etc.

Cluster 2

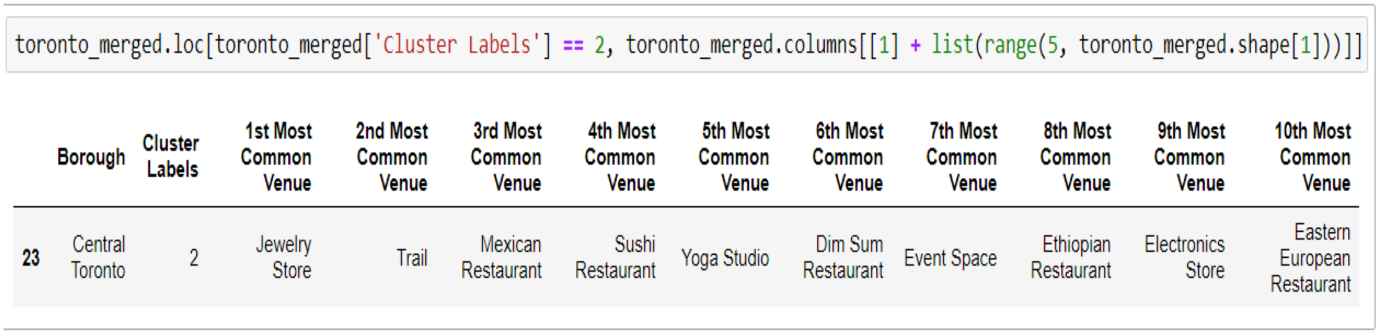


```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 1, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1]))]]
```

	Borough	Cluster Labels	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
22	Central Toronto	1	Home Service	Garden	Ice Cream Shop	Yoga Studio	Dessert Shop	Event Space	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant

Cluster 2 was in the Central Tornoto area. Cluster 2 which are mostly business areas with Yoga Studio, restaurants, supermarkets etc.

Cluster 3

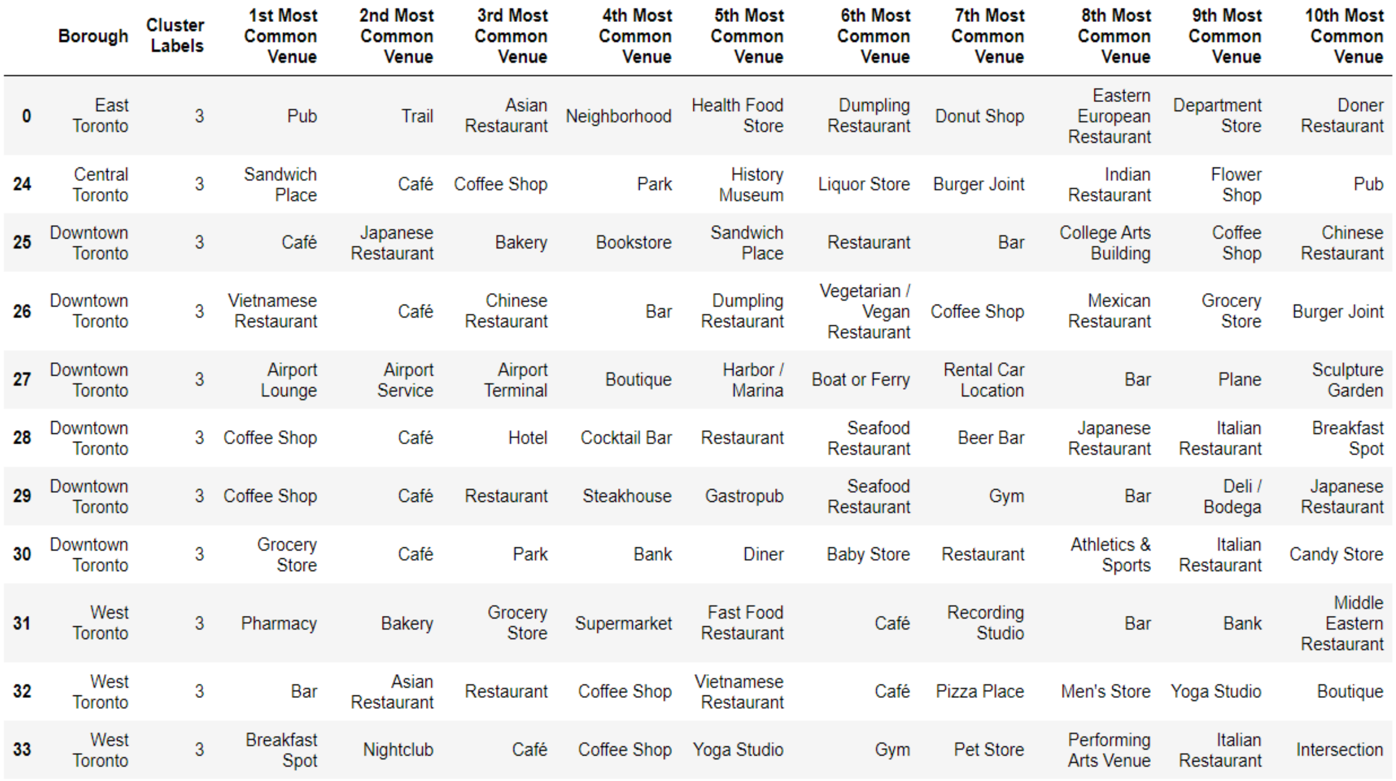


```
toronto_merged.loc[toronto_merged['Cluster Labels'] == 2, toronto_merged.columns[[1] + list(range(5, toronto_merged.shape[1]))]]
```

	Borough	Cluster Labels	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
23	Central Toronto	2	Jewelry Store	Trail	Mexican Restaurant	Sushi Restaurant	Yoga Studio	Dim Sum Restaurant	Event Space	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant

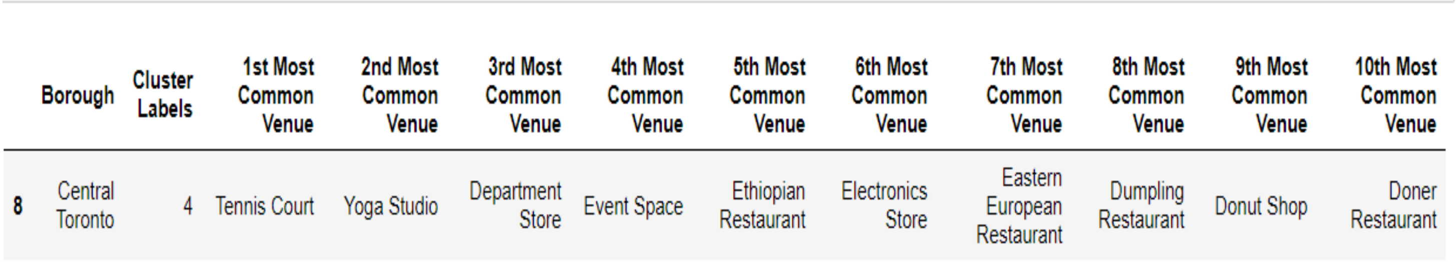
Cluster 4

Cluster 4 was in the Central ,East, west, DownTown Tornoto area. Cluster 4 which are mostly business areas with Hostel, Bakery, café, supermarkets etc.



	Borough	Cluster Labels	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
0	East Toronto	3	Pub	Trail	Asian Restaurant	Neighborhood	Health Food Store	Dumpling Restaurant	Donut Shop	Eastern European Restaurant	Department Store	Doner Restaurant
24	Central Toronto	3	Sandwich Place	Café	Coffee Shop	Park	History Museum	Liquor Store	Burger Joint	Indian Restaurant	Flower Shop	Pub
25	Downtown Toronto	3	Café	Japanese Restaurant	Bakery	Bookstore	Sandwich Place	Restaurant	Bar	College Arts Building	Coffee Shop	Chinese Restaurant
26	Downtown Toronto	3	Vietnamese Restaurant	Café	Chinese Restaurant	Bar	Dumpling Restaurant	Vegetarian / Vegan Restaurant	Coffee Shop	Mexican Restaurant	Grocery Store	Burger Joint
27	Downtown Toronto	3	Airport Lounge	Airport Service	Airport Terminal	Boutique	Harbor / Marina	Boat or Ferry	Rental Car Location	Bar	Plane	Sculpture Garden
28	Downtown Toronto	3	Coffee Shop	Café	Hotel	Cocktail Bar	Restaurant	Seafood Restaurant	Beer Bar	Japanese Restaurant	Italian Restaurant	Breakfast Spot
29	Downtown Toronto	3	Coffee Shop	Café	Restaurant	Steakhouse	Gastropub	Seafood Restaurant	Gym	Bar	Deli / Bodega	Japanese Restaurant
30	Downtown Toronto	3	Grocery Store	Café	Park	Bank	Diner	Baby Store	Restaurant	Athletics & Sports	Italian Restaurant	Candy Store
31	West Toronto	3	Pharmacy	Bakery	Grocery Store	Supermarket	Fast Food Restaurant	Café	Recording Studio	Bar	Bank	Middle Eastern Restaurant
32	West Toronto	3	Bar	Asian Restaurant	Restaurant	Coffee Shop	Vietnamese Restaurant	Café	Pizza Place	Men's Store	Yoga Studio	Boutique
33	West Toronto	3	Breakfast Spot	Nightclub	Café	Coffee Shop	Yoga Studio	Gym	Pet Store	Performing Arts Venue	Italian Restaurant	Intersection

Cluster 5



	Borough	Cluster Labels	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
8	Central Toronto	4	Tennis Court	Yoga Studio	Department Store	Event Space	Ethiopian Restaurant	Electronics Store	Eastern European Restaurant	Dumpling Restaurant	Donut Shop	Doner Restaurant

Cluster 5 was in the Central Tornoto area. Cluster 5 which are mostly business areas with Tennis Court, Restaurants, Event Space, stores etc.

A large portion of the Indian Restaurants are in group 2 addressed by the violet clusters. The Neighborhoods situated in the North York region that have the most elevated normal of Indian Restaurants are Dundas Street and Gerrad street. There are countless Neighborhoods in group 4 and Indian cafés moreover. We see that in the west Toronto territory (group 3) has the subsequent last normal of Indian Restaurants. Taking a gander at the close by settings, the ideal spot to place another Indian Restaurant in Downtown Toronto as there are numerous Neighborhoods in the space yet practically zero Indian Restaurants, subsequently, killing any opposition. The second-best Neighborhoods that have an extraordinary chance would be in regions like Boat or Ferry, Coffeeshop, and so on which is in Cluster 4. Having approx. 70 neighborhoods nearby with no Indian Restaurants offers a decent chance for opening another eatery.

So restauarant should be open in bunch 4 It had greatest number of neighborhoods with no Indian Restaurants. A portion of the downsides of this investigation are — the grouping is totally founded on information acquired from the Foursquare API. Additionally, the investigation doesn't contemplate of the Indian populace across neighborhoods as they play a tremendous factor while picking which spot to open another Indian café. This closes the ideal discoveries for this task and prescribes the business visionary to open a real Indian eatery in these areas with practically no opposition.

Chapter 5
Conclusion

With everything taken into account, to end off this assignment, we got an opportunity on a business issue, and it was taken care of to such an extent that it resembled how a bona fide data scientist would do. We utilized different Python libraries to bring the information, control the substance and isolated and imagine those datasets. We have utilized Foursquare API to analyze the settings in neighborhoods of Toronto, get an uncommon extent of data from Website which we scratched with the Beautifulsoup Web scratching Library. We in like manner imagined utilizing different plots present in seaborn and Matplotlib libraries. Likewise, we applied AI framework to expect the mix-up given the information and utilized Folium to picture it on a guide.

Chapter 6

Future Scope

Spots that have opportunity to get better or certain disadvantages give us that this venture can be better with the help of more data and unmistakable Machine Learning methodologies. Moreover, we can use this dare to examine any circumstance, for instance, opening a substitute food or opening of a Movie Complex, etc. In a perfect world, this undertaking goes about as an underlying heading to handle more mind boggling genuine issues utilizing information science.

References

[1] https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a#:~:text=Kmeans%20algorithm%20is%20an%20iterative,belongs%20to%20only%20one%20group.&text=Keep%20iterating%20until%20there%20is,to%20clusters%20isn't%20changing.

[2] https://www.simplilearn.com/tutorials/data-science-tutorial/what-is-data-science

[3] https://github.com/ayush5499/TorontoRestaurantLocationHunt

[4] https://towardsdatascience.com/battle-of-the-neighborhoods-b72ec0dc76b5

[5] https://towardsdatascience.com/battle-of-the-neighborhoods-b72ec0dc76b5

[6] http://www.zinkohlaing.com/data-science/using-machine-learning-to-find-locations-to-open-a-burmese-restaurant-in-toronto-ibm-capstone-project/

[7] https://pkmudaliar9669.medium.com/ibm-data-science-capstone-final-project-a8abe80e6f75

[8] http://www.zinkohlaing.com/data-science/using-machine-learning-to-find-locations-to-open-a-burmese-restaurant-in-toronto-ibm-capstone-project/

[9] https://towardsdatascience.com/strategic-location-for-establishing-an-asian-restaurant-c3aecf2496b1

[10] https://medium.com/@shaswatd673/restaurant-location-recommender-using-k-means-6b3a54f27e64

[11] https://www.linkedin.com/pulse/data-science-capstone-best-neighborhood-san-francisco-davis-nix/?articleId=6612389638839644160