# APPLIED DATA SCIENCE – CAPSTONE IBM

# Project: TORONTO ENTREPRENEURIAL For INDIAN BUSINESSMEN

---

## CONTENT



For detailed data reports refer report.pptx

## INTRODUCTION

The city of Toronto has approached our company to help them develop a service that helps the **Indian entrepreneurs** who want to establish new businesses in the city of Toronto select an ideal business location based on the Indian ethnic communities they want to be a part of.

This service will help find an ideal location for a new business based on such factors as business venue, population density in the area, the demographics in the area, average income, proximity to other business venues.

## PROBLEM STATEMENT

The success and failure of any Business depends on a vast spectrum of economics and demographics factors. Entrepreneurs may want to find an optimal venue and geographic location for their new business venture so that they get maximum profit from it. Such an optimal venue/place selection process has to consider various indicators that may deliver long and prosperous existence for any new business.

As we know that there is vast population of Indian community residing in City of Toronto.So if an Indian businessman wants to start a business considering the indian population than he must take into consideration the trends and patterns of it.

Successful businesses help the economy grow, lower the unemployment, and reduce crime. The multicultural city of Toronto wants to offer such an online service where the Indian entrepreneurs can receive all the necessary information that will help them in picking the location for their new ventures based on their desire to support the Indian ethnic community of Toronto.

### AUDIENCE

The target audience for this service: Indian Business Entrepreneurs seeking to establish new businesses in the city of Toronto Having a tool that can help the entrepreneurs to choose the right location for their business will assure long and prosperous existence for such businesses, serving the communities and helping them grow.

### PROPOSED SOLUTION

The solution will leverage the Foursquare location data as well as the demographics open dataset available from Wikipedia. In order to advise the entrepreneurs on a good location, we will consider the density (frequency) of similar business venues in various parts of Toronto that cater to preferred ethnic area/neighbourhoods, average income, population, population density, population growth rate, spoken languages in the same area.

## DATA SOURCES

To solve the problem our service will rely on open datasets generated from the following sources:

•Wikipedia -- Toronto Boroughs/Neighbourhoods https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M -- Canada Census - Toronto Demographics https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

•GeoCoder/Google Geolocation APIs

•Foursquare APIs

**Toronto Boroughs/Neighbourhoods**: 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

**Canada Census - Toronto Demographics**: a list of demographic data on each Toronto neighbourhood as taken from the Canadian Census.

**GeoCoder/Google Geolocation APIs**: converts addresses into geographic coordinates

**Foursquare APIs**: offers rich location-based experiences and enables access to millions of up to date business venues, tips, photos and many other helpful tips

# METHODOLOGY

In order to perform statistical inference, and apply the machine learning algorithms, the data must be acquired and pre-processed based on the rules derived from the preliminary data analysis

**DATA ANALYSIS** - Identify the significant informational indicators to use in inferential statistics and machine learning algorithm [Unsupervised: K-Means]

**DATA ANALYSIS** - Statistical Validation: The datasets underwent statistical analysis and cross referencing in order to determine the data validity and proper distribution, mean and standard deviations, outlier identification.

**MACHINE LEARNING** - UNSUPERVISED MACHINE LEARNING K-MEANS: In order to cluster various regions of the city based on the business analysis requirements the solution utilizes the unsupervised machine learning algorithm **K-MEANS**

**DATA RESULTS** - Present the finding to the stakeholders

**DISCUSSION** - discuss data investigative findings based on the results

**CONCLUSION** - report conclusions

# MACHINE LEARNING

Our data analysis shows lack of proper data labeling in the datasets used by the solution based on the data analysis and the solution requirements we suggest using an unsupervised machine learning approach we suggest using k-means unsupervised machine learning algorithm to identify geo clusters in the city of toronto that are most suitable for opening new small businesses in the city of toronto in order to perform accurate geo clustering our algorithm relies on google geo locations and foursquare apis

## SUGGESTED NEIGHBOURHOODS

• Thorncliffe Park

•Rouge

•Oakridge

•Malvern

•Scarborough Village

•Morningside

•Dorset Park

•Woburn

•Thistletown

•Flemingdon Park

•Highland Creek

•Victoria Village

•West Hill

•Cliffcrest

**Total Number of venues: 93 Total Number of unique categories: 53**
**Total Number of clusters: 8**

**Cluster 0:** Scarborough-Tamil Recommendations for Cluster 0:

| American Restaurant | Warehouse Store | Italian Restaurant | Japanese Restaurant | Korean Restaurant | Latin American Restaurant | Liquor Store | Medical Center | Mexican Restaurant | Motel |
|---|---|---|---|---|---|---|---|---|---|

**Cluster 1:** Scarborough-Tamil Recommendations for Cluster 1:

| American Restaurant | Italian Restaurant | Japanese Restaurant | Korean Restaurant | Latin American Restaurant | Liquor Store | Medical Center | Metro Station | Mexican Restaurant | Motel |
|---|---|---|---|---|---|---|---|---|---|

**Cluster 2:** East York-Urdu Recommendations for Cluster 2:

| American Restaurant | Intersection | Italian Restaurant | Japanese Restaurant | Korean Restaurant | Latin American Restaurant | Medical Center | Metro Station | Mexican Restaurant | Hockey Arena |
|---|---|---|---|---|---|---|---|---|---|

**Cluster 3:** Scarborough-Tamil Recommendations for Cluster 3:

| American Restaurant | Italian Restaurant | Japanese Restaurant | Korean Restaurant | Latin American Restaurant | Liquor Store | Medical Center | Metro Station | Mexican Restaurant | Motel |
|---|---|---|---|---|---|---|---|---|---|

**Cluster 4:** Etobicoke-Punjabi Recommendations for Cluster 4:

American Restaurant Intersection Italian Restaurant Japanese Restaurant Korean Restaurant Latin American Restaurant Liquor Store Medical Center Metro Station Mexican Restaurant

**Cluster 5:** Scarborough-Tamil Recommendations for Cluster 5:

| American Restaurant | Indian Restaurant | Warehouse Store | Italian Restaurant | Japanese Restaurant | Korean Restaurant | Latin American Restaurant | Liquor Store | Metro Station | Mote |
|---|---|---|---|---|---|---|---|---|---|

**Cluster 6:**Scarborough-Tamil Recommendations for Cluster 6:

| Insurance Office | Italian Restaurant | Japanese Restaurant | Korean Restaurant | Latin American Restaurant | Liquor Store | Medical Center | Metro Station | Mexican Restaurant | Par |
|---|---|---|---|---|---|---|---|---|---|

**Cluster 7:** Scarborough-Tamil Recommendations for Cluster 7:

American Restaurant Italian Restaurant Japanese Restaurant Korean Restaurant L atin American Restaurant Liquor Store Medical Center Metro Station Mexican Restaurant Motel

# DISCUSSION

There are very interesting trends showing up n the data analysis that suggest that it is possible to recommend new locations to indian businesses that want to expand or new businesses looking for the first location.

There are multiple statistical methodologies that can be employed to formulate a sound business hypothesis. Such formulated hypothesis do require validation via gathering and processing the supporting evidence.

Such supporting evidence can be produced by employing one or more machine learning algorithms.

There is additional potential in employing a recommender system algorithm to further improve the recommendations report based on a number of business requirements.

# CONCLUSION

Given enough relevant data it is possible to generate sufficient amount of supporting evidence in order to recommend with a high level of precision geo locations for new or growing businesses.

The current project demonstrates that a new location can be selected based on a list of indicators derived via inferential statistics and the results processed with k-means clustering machine learning algorithm.

Being a business with close ties to various ethnic communities in toronto we definitely concur that the findings presented in this report have strong merits.