




# **COURSERA CAPSTONE APPLIED DATA SCIENCE PROJECT**

Presented by –  
PRANJAL SHRIVASTAVA



# INTRODUCTION :

- ❑ **Toronto city**, capital of the province of Ontario, southeastern Canada. It is the most populous city in Canada and the country's financial and commercial center. The demographics of Toronto make Toronto one of the most multicultural and multiracial cities in the world. More than half of the entire Indian-Canadian population resides in Toronto, people from India love food and I love to eat, thus Toronto is one of the best places to start an INDIAN RESTAURANT.

# BUSINESS PROBLEM :

- ❑ In this capstone project, we will analyze the neighborhoods in Toronto to identify the most profitable neighborhood for opening an Indian Restaurant, by using Web Scraping, Data Pre-processing, Machine learning algorithms like K-Means clustering algorithm, and Foursquare API Service.

# TARGET AUDIENCE :

- ☐ The business owner who wants to invest or open a start-up company or restaurant.
- ☐ The freelancer who loves to have their own small company or restaurant as a side business.
- ☐ Indian crowd who wants to find neighborhoods with lots of options for Indian restaurants.
- ☐ Tourists who want to eat Indian food.

# DATA SOURCE :

- ❑ Toronto City Neighbourhoods Data – [https://en.wikipedia.org/wiki/List\\_of\\_postal\\_codes\\_of\\_Canada:\\_M](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
- ❑ Geographical Coordinates of the Neighbourhoods – [https://cocl.us/Geospatial\\_data](https://cocl.us/Geospatial_data)
- ❑ Location Data of Neighbourhood – Foursquare API Services

# METHODOLOGY :

- ❑ First, web scraping for data using the 'BeautifulSoup' package is done.
- ❑ Then a dataframe is created which contains Postal Code, Borough, Neighbourhood.
- ❑ Then further processing of the dataframe is done, like removing unassigned values, merging different neighborhoods with the same borough.
- ❑ The next step is to add geographical coordinates, for that the Geospatial\_data.csv is used.
- ❑ Then merging is done based on Neighbourhood.
- ❑ Finally, the dataframe is modified, in which Borough contains 'Toronto'.
- ❑ Then using Foursquare API developer services, 100 venues were explored under the radius of 500m. A Foursquare developer account to obtain a Client ID and Client Secret key to pull the data. From Foursquare, the names, categories, latitude, and longitude of the venues were pulled.
- ❑ Then one-hot encoding was performed, for each of the neighborhood's venues were turned into the frequency at how many of those Venues were located in each neighborhood. Then, the values in each venue category were grouped by the average of the frequency of each venue category.

# METHODOLOGY :

- ❑ Then, a new dataframe was created, which stored Neighbourhood and Indian Restaurant column. It was created to make further operations simpler.
- ❑ At last, clustering of Indian Restaurant was done, it was based on K-Means Clustering Algorithm. Here k-value was taken as 4, then the model was fitted. And finally, the clustered map was displayed.

# RESULTS :

- ❖ Cluster 0 – Green Colour
  - ❖ Cluster 1 – Blue Colour
  - ❖ Cluster 2 – Purple Colour
  - ❖ Cluster 3 – Brown Colour
- 
- Cluster 0 - Depicts the least frequency of Indian Restaurants among the neighborhoods.
  - Cluster 2(The Annex, North Midtown, Yorkville) - Depicts the maximum frequency of Indian Restaurants among the neighborhoods.

