**Capstone Project - Recommending Toronto Neighbourhoods to Immigrants**

**Introduction :**

There is a large scale migration from India to Canada every year and Toronto is one of the major destination of these new migrants because of its financial activity and cosmopolitan culture.

The new landed immigrants have to first decide, which Neighbourhood they should choose to stay based on their preferences of venues.

A project that helps to create a source of information, which can recommend to them the best Neighbourhoods for them based on their preferences of venues and also provide an overall information of the venues available across the Neighbourhoods and some clustering of similar Neighbourhoods basis the venues will be helpful to the new immigrants.

**Business problem :**

New immigrants to Toronto, Canada are constrained in their choices regarding Neighbourhoods to stay due to lack of readily available information on relevant Neighbourhood choices.

1. Immigrants do not have a source of information to recommend to them some Neighbourhoods to stay based on their venue preferences.

2. Immigrants do not have information. On Top 10 venues per Neighbourhood

3. Immigrants do not have information on similar neighbourhoods basis the

venues

**Data (Sources ):**

1. Toronto Neighbourhoods Post Codes

<https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=890001695>

| **Postcode** | **Borough** | **Neighbourhood** |
| --- | --- | --- |

This data provides the list of Neighbourhoods in Toronto along with their postal codes.

1. Co-Ordinates Data ( Geospatial\_Coordinates.csv )

| **Postal Code** | **Latitude** | **Longitude** |
| --- | --- | --- |

This data provides the longitude – latitude for the Neighbourhoods obtained from the first data source.

1. Four Square Data on Toronto Neighbourhoods/ Venues

'[https://api.foursquare.com/v2/venues/explore?&client\_id={}&client\_secret={}&v={}&ll={},{}&radius={}&limit={}'.format(](https://api.foursquare.com/v2/venues/explore?&client_id=%7B%7D&client_secret=%7B%7D&v=%7B%7D&ll=%7B%7D,%7B%7D&radius=%7B%7D&limit=%7B%7D%27.format() CLIENT\_ID, CLIENT\_SECRET, VERSION, Neighbourhood\_latitude, Neighbourhood\_longitude, radius, LIMIT)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Neighbourhood | Neighbourhood Latitude | Neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |

This data obtained from Four Square provides the Venues/Venue Category along with other location details for the Venues.

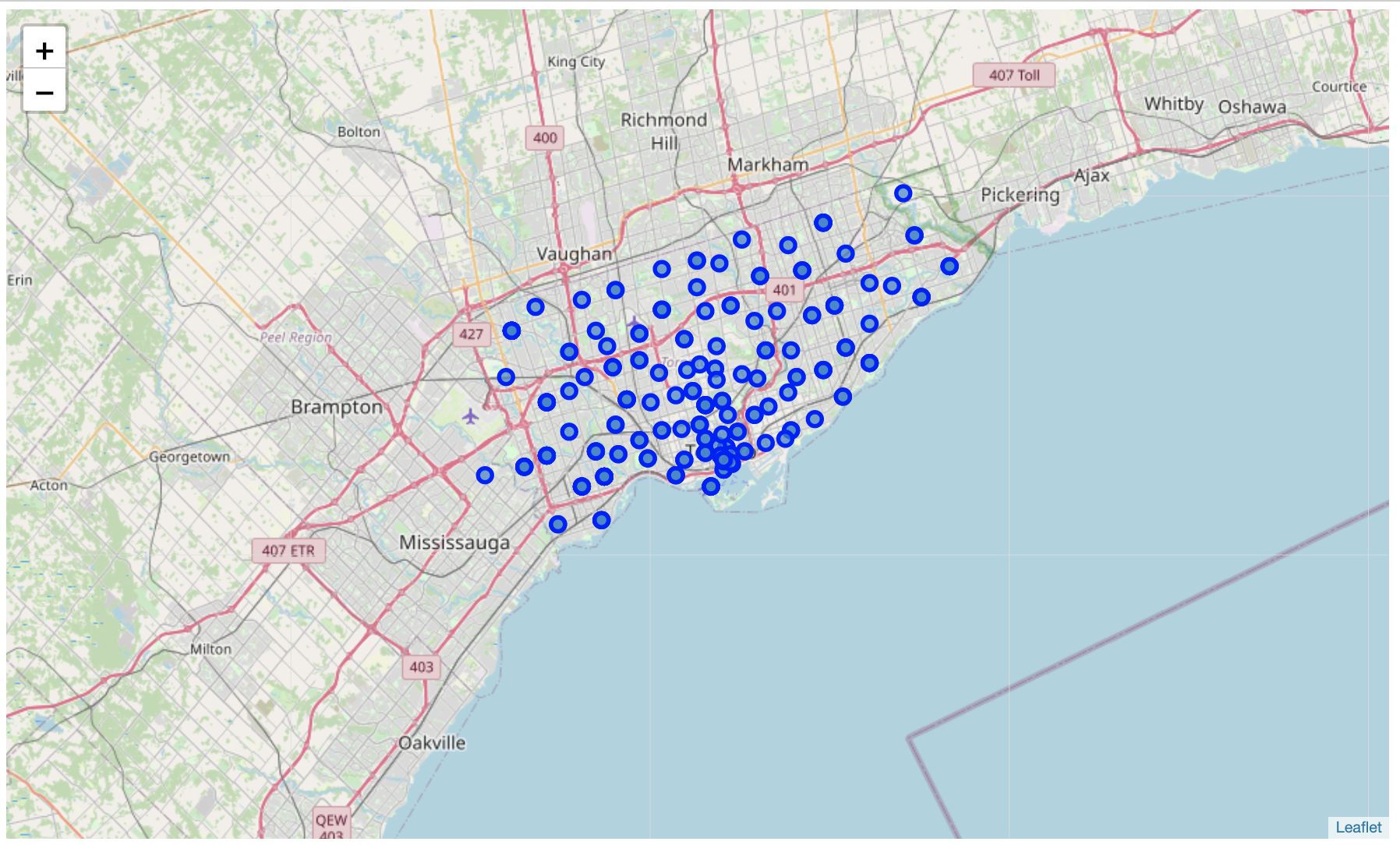
Above database will be used to provide the required analysis and build a Neighbourhoods recommendation system for the new immigrants

**Methodology:**

Including (1) Data cleaning and organising (2) Exploratory Data Analysis (3) Solution Development

1. **Data cleaning and organising :**

* The original Neighbourhoods database had the shape (288, 3) that was obtained from Wikipedia. (Data source 1)
* Missing NA entries were first dropped and then the ‘Not assigned’ Neighbourhood entries were replaced with the ‘Borough’ entries.
* The data was then grouped ‘Postcode’ – wise and the data took the shape (103, 3).
* The Neighbourhoods with same ‘Postcode’ were joined to a single row.
* The co-ordinates data, from Data source 2, was then merged with the earlier data to update the location data of the Neighbourhoods.
* This Neighbourhoods data was then visualized on the Toronto Map.



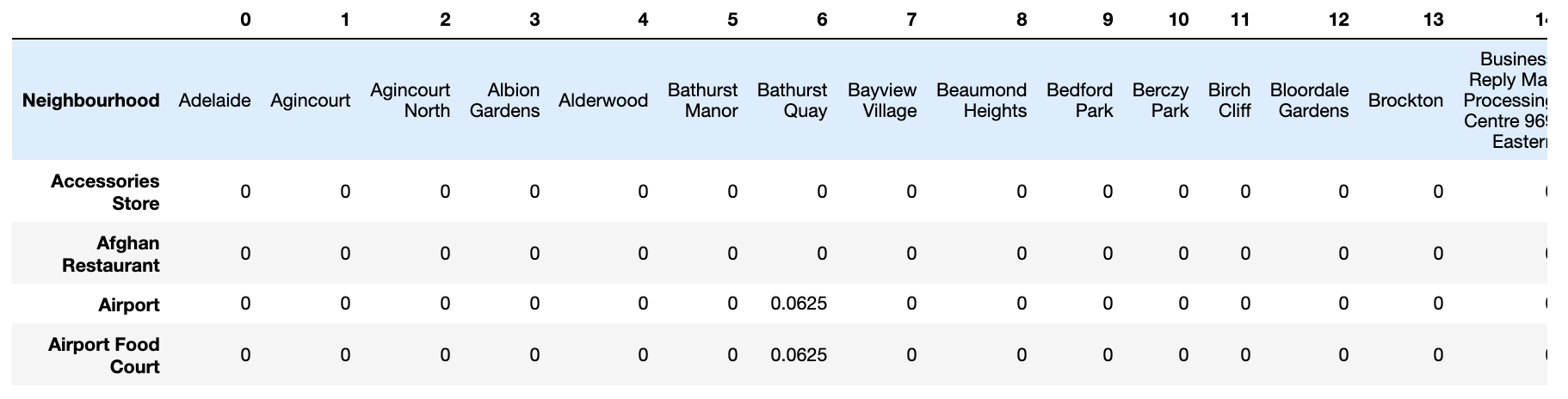
* Next the Venues/ Venue categories and their locations data were obtained from Four Square for the Neighbourhoods.
* The Venues data was then merged with the Neighbourhoods data to obtained the following database.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Neighbourhood | Neighbourhood Latitude | Neighbourhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |

* The above database had the shape (4201, 7)

1. **Exploratory Data Analysis :**

* Dummy encoding was then done on ‘Venue Category’ to create a separate database of dummy encoded venue category labels.
* Neighbourhoods column was then added to create a (204, 271) database.
* From the above database a matrix database of Neighbourhoods – Venue Categories was created

****

1. **Solution Development :**

* From the above database top 10 venues/Neighbourhood were generated by sorting and looping through the data-frame.
* Next, K-Means clustering was used to create cluster of similar Neighbourhoods basis the venues.
* A total of 5 clusters were created and the same were visualized on the map.
* Next, a recommendation system was developed for the immigrants,
  + Immigrant would first make a choice of his top 5 venue requirements and give a score to those out of 5.
  + A vector of shape (204,1) of immigrant preferences was created from the above choices.
  + A dot product of this vector was then taken with the earlier (204, 271) matrix of Neighbourhoods – Venue Categories to obtain the score of the Neighbourhoods for the specific immigrant.
  + These scores were then sorted to obtain the 10 top Neighbourhoods as per the immigrants preferences.

**Results :**

* The recommended Neighbourhoods with their scores is as below for the specific immigrant.

Neighbourhood

CFB Toronto 0.9999

Downsview East 0.9999

Guildwood 0.5716

West Hill 0.5716

Morningside 0.5716

Railway Lands 0.1764

Bathurst Quay 0.1764

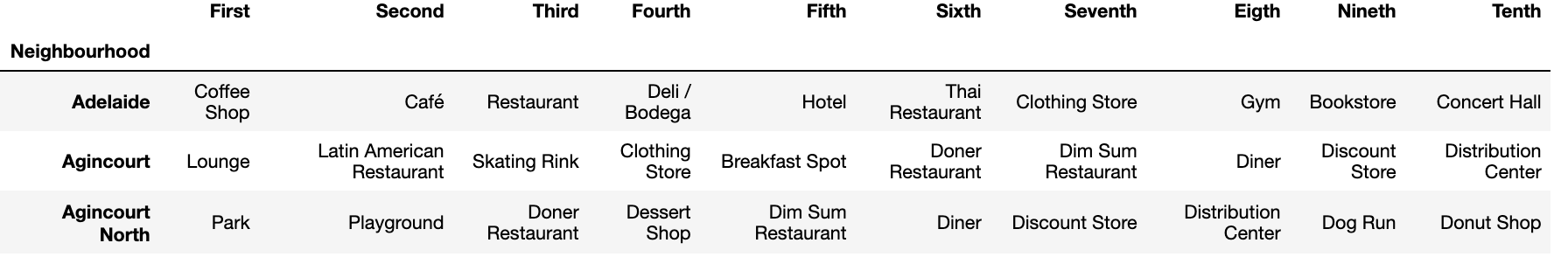
CN Tower 0.1764

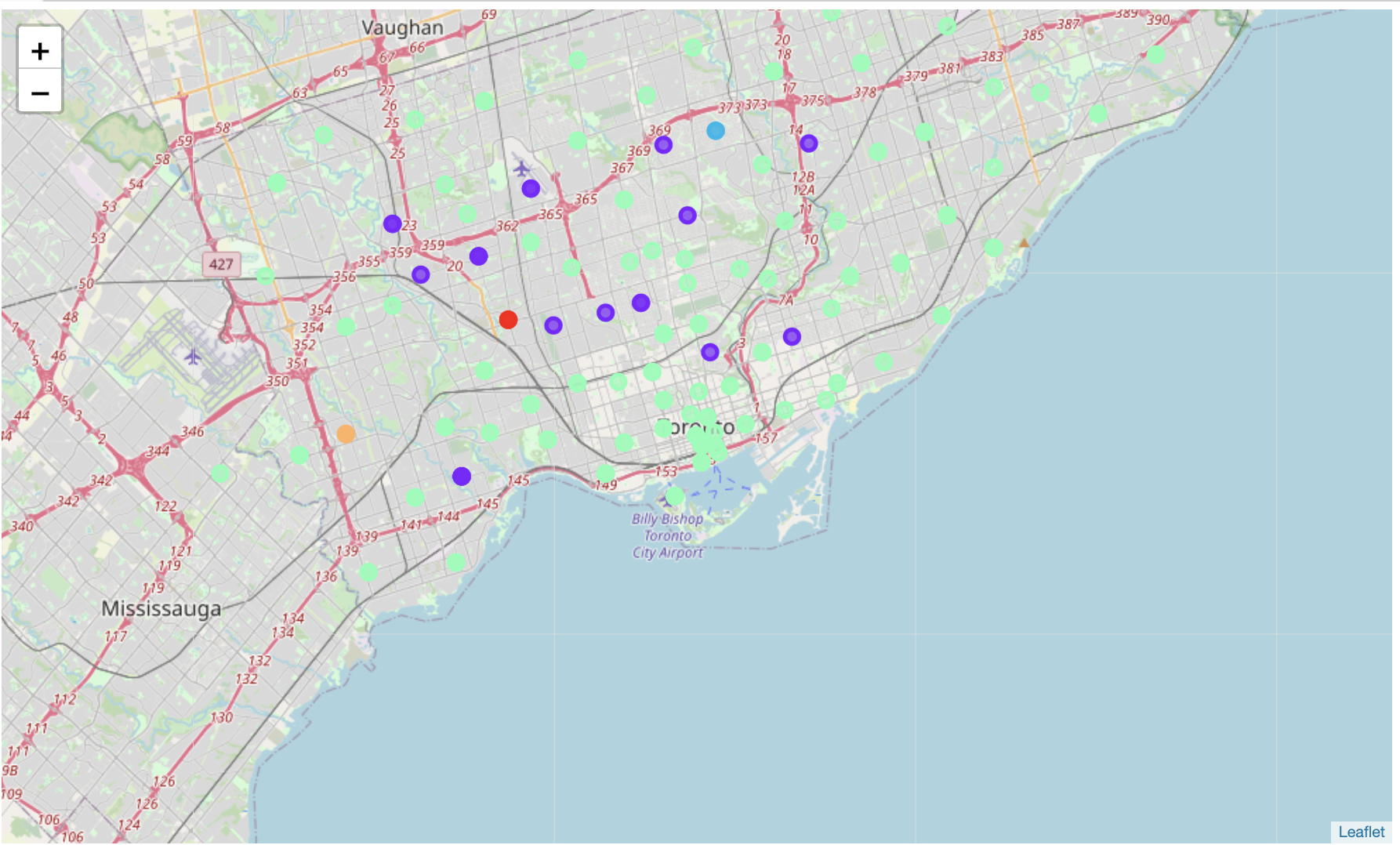
Island airport 0.1764

King and Spadina 0.1764

Name: score, dtype: float64

* The top 10 Venue categories for the Neighbourhoods was then obtained.



* 5 Similar Neighbourhoods clusters were obtained and plotted on the Toronto Map for easy visualization.

**Conclusion :**

The solution developed through the project aims at providing the required information to the new immigrants to Toronto, Canada and recommending the best suited Neighbourhoods to the immigrants as per their preferences.

From a Data Science perspective this project has involved utilization of several techniques like,

* Data cleaning and organizing
* Database creation
* Data visualization
* Machine learning K-Means Clustering
* Machine learning Recommendation System

The approach can now be extended to other geographies or to projects having similar requirements.