LOCATION RECOMMENDATION FOR A NEW BUSINESS UNIT

THE BATTLE OF NEIGHBORHOODS

DATA SCIENCE CAPSTONE PROJECT

PRESENTED BY

BHAVANI MADDULA

Business Problem

- Statement:
  - Identify a neighborhood in Toronto to open an 'Indian Arts, Dance and Cultural Centre' and with restaurants in the vicinity

- Solution:
  - Find a neighborhood inhabited by people with Indian origin and with good number of restaurants

Data Section

1. Data Sources
   1. Toronto neighborhoods:
      https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

Example: Few rows in the extracted dataframe:

Postcode	Borough	Neighbourhood
M1A	Not assigned	Not assigned
M2A	Not assigned	Not assigned
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Harbourfront

2. Demographics of Toronto neighborhoods:

https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighborhoods

Example: Few rows in the extracted dataframe:

```
[3]: Demographics.head()
```

[3]:

	Name	FM	Census Tracts	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters	Second most common language (after English) by name	Second most common language (after English) by percentage	Map
0	Toronto CMA Average	NaN	All	5113149	5903.63	866	9.0	40704	10.6	11.4	NaN	NaN	NaN
1	Agincourt	S	0377.01, 0377.02, 0377.03, 0377.04, 0378.02, 0...	44577	12.45	3580	4.6	25750	11.1	5.9	Cantonese (19.3%)	19.3% Cantonese	NaN
2	Alderwood	E	0211.00, 0212.00	11656	4.94	2360	-4.0	35239	8.8	8.5	Polish (6.2%)	06.2% Polish	NaN
3	Alexandra Park	OCoT	0039.00	4355	0.32	13609	0.0	19687	13.8	28.0	Cantonese (17.9%)	17.9% Cantonese	NaN
4	Allenby	OCoT	0140.00	2513	0.58	4333	-1.0	245592	5.2	3.4	Russian	01.4%	NaN

3. Geocoder/Google geolocation API/Geo spatial coordinates csv file:

- Finding the geospatial coordinates of a location - Geocoding

- The following information was returned for the coordinates of 'Downtown Toronto'

- The geographical coordinates of Downtown Toronto are 43.6541737, -79.3808116451341

4. Foursquare API:

- Foursquare - one of the most popular Location Based Social Network (LBSN) in recent times
- Provides personalized recommendations of places to go to near a user's current location based on users' previous browsing history, purchases, or check-in history
- Example: Venues returned by Foursquare for 'Parkwoods' neighborhood

|	name	categories	lat	lng
0	Brookbanks Park	Park	43.751976	-79.332140
1	KFC	Fast Food Restaurant	43.754387	-79.333021
2	Variety Store	Food & Drink Shop	43.751974	-79.333114

```
print('{} venues were returned by Foursquare.'.format(nearby_venues.shape[0]))
3 venues were returned by Foursquare.
```

Methodology

- Merge the three dataframes related to Demographics of Toronto, Neighborhoods of Toronto, Geospatial coordinates for neighborhoods of Toronto

- Use Foursquare credentials to get the top 100 venues in each neighborhood

- 222 unique categories were identified

- Create a dataframe listing the top 10 venues for each neighborhood - data set for performing clustering at a later stage
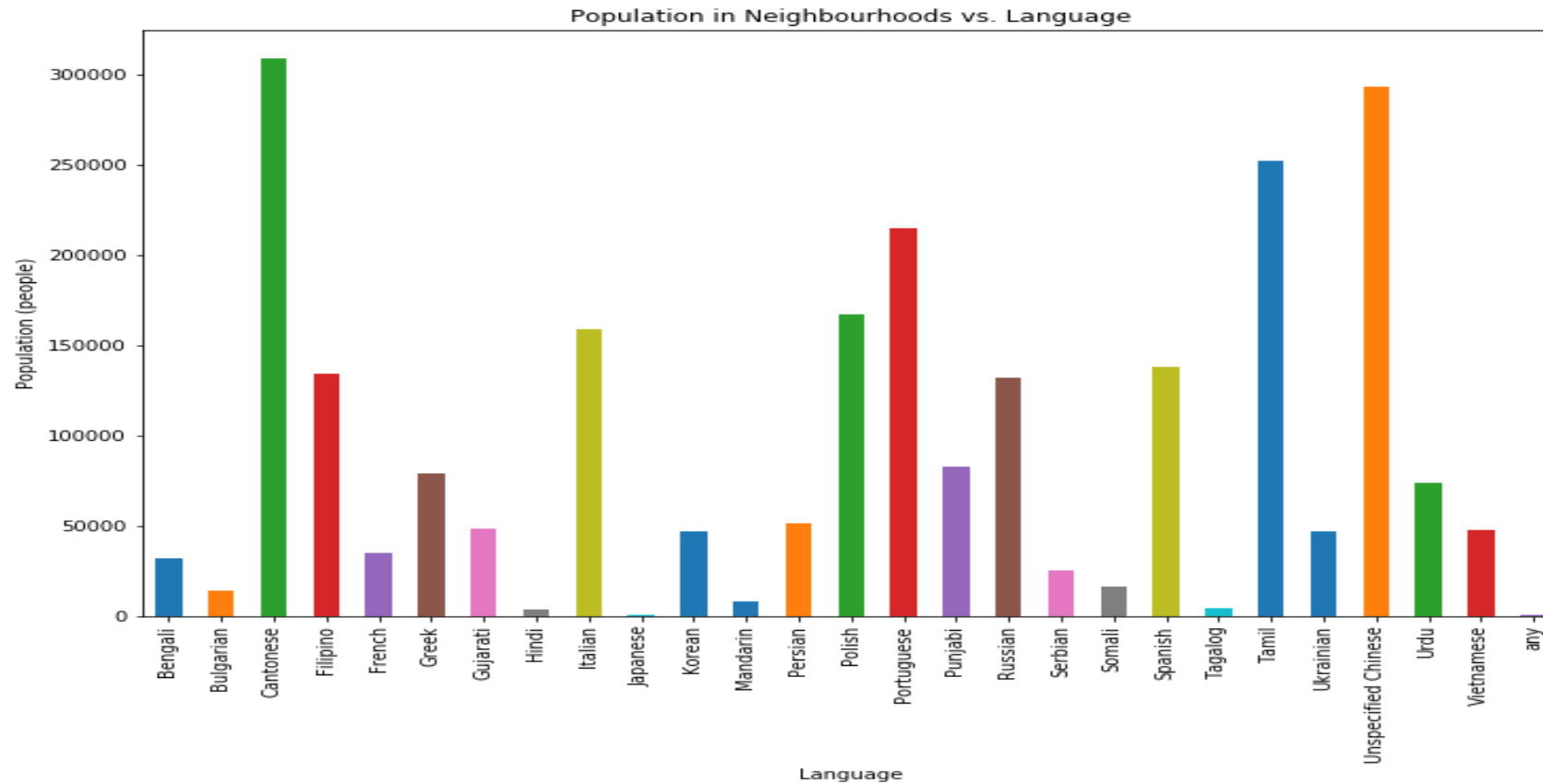
Exploratory Data Analysis

Top 10 most populated neighborhoods

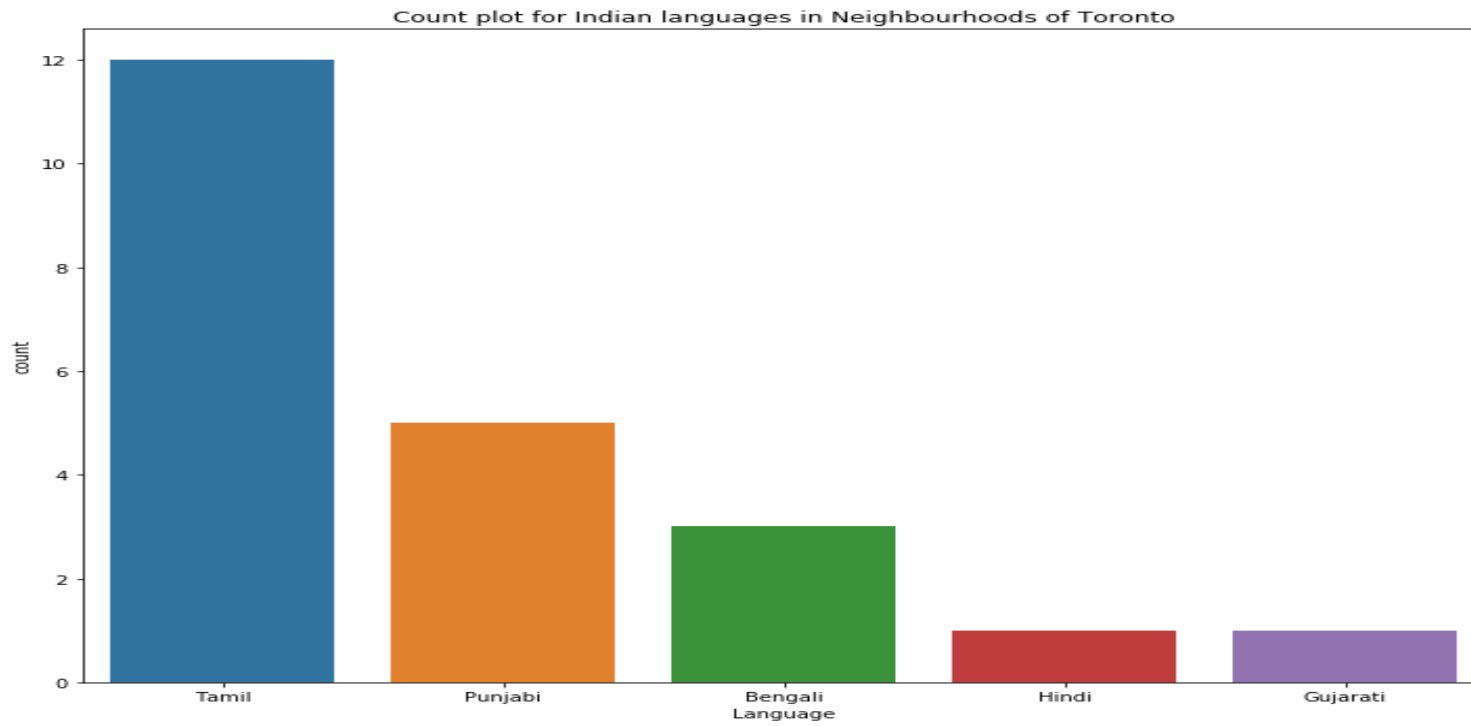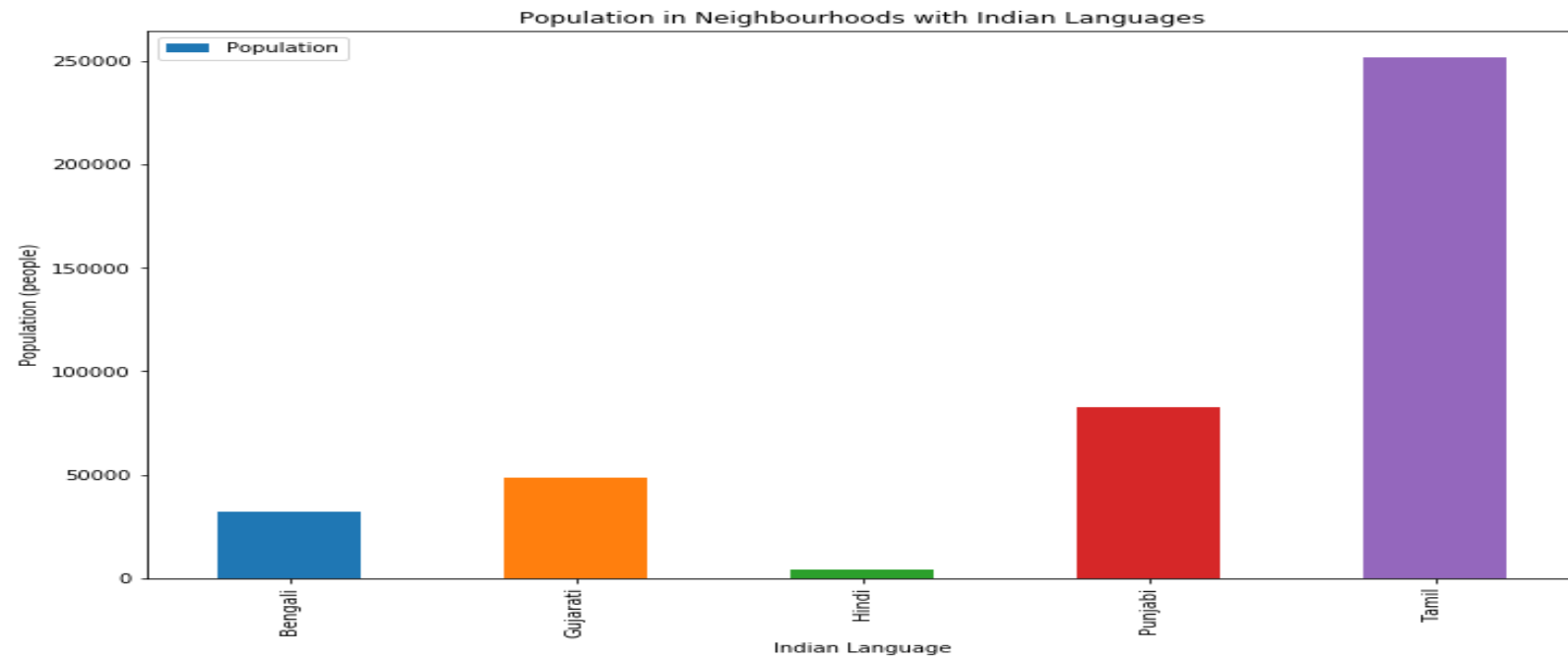
Top 10 Most Populated Neighbourhoods
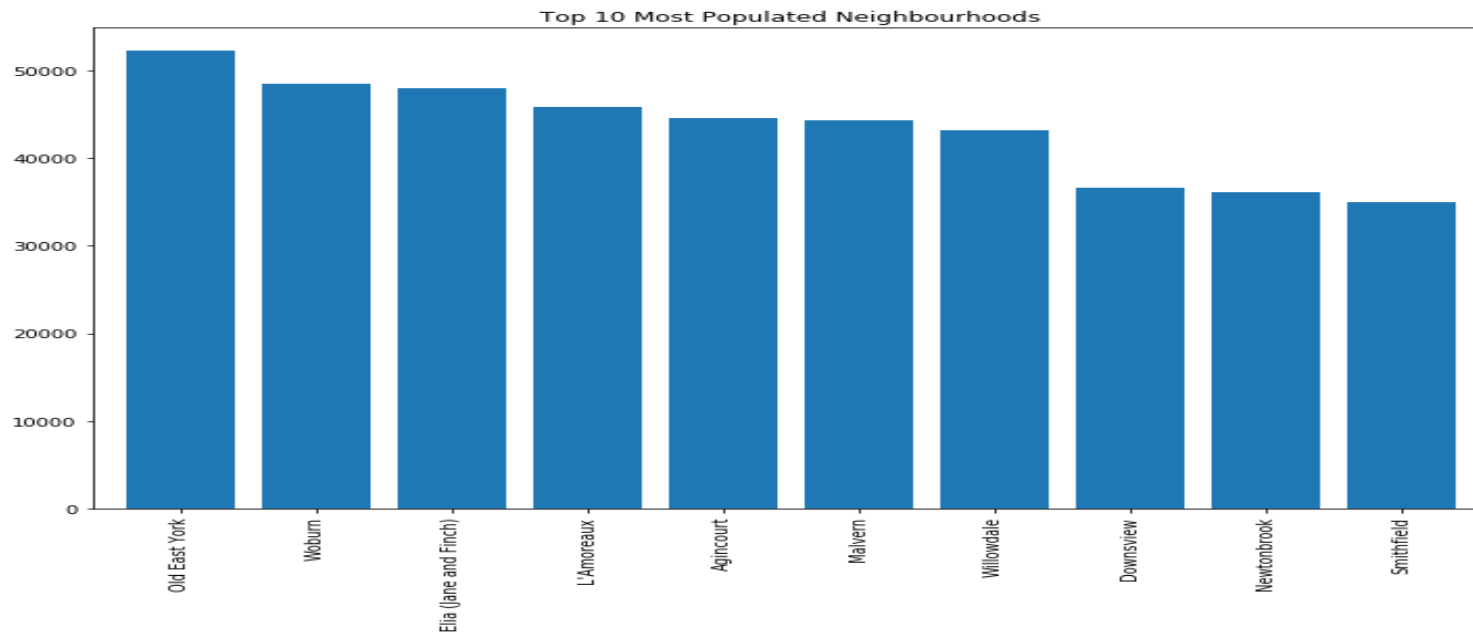
Neighborhood population versus the second most popular language (after English) in the neighborhood
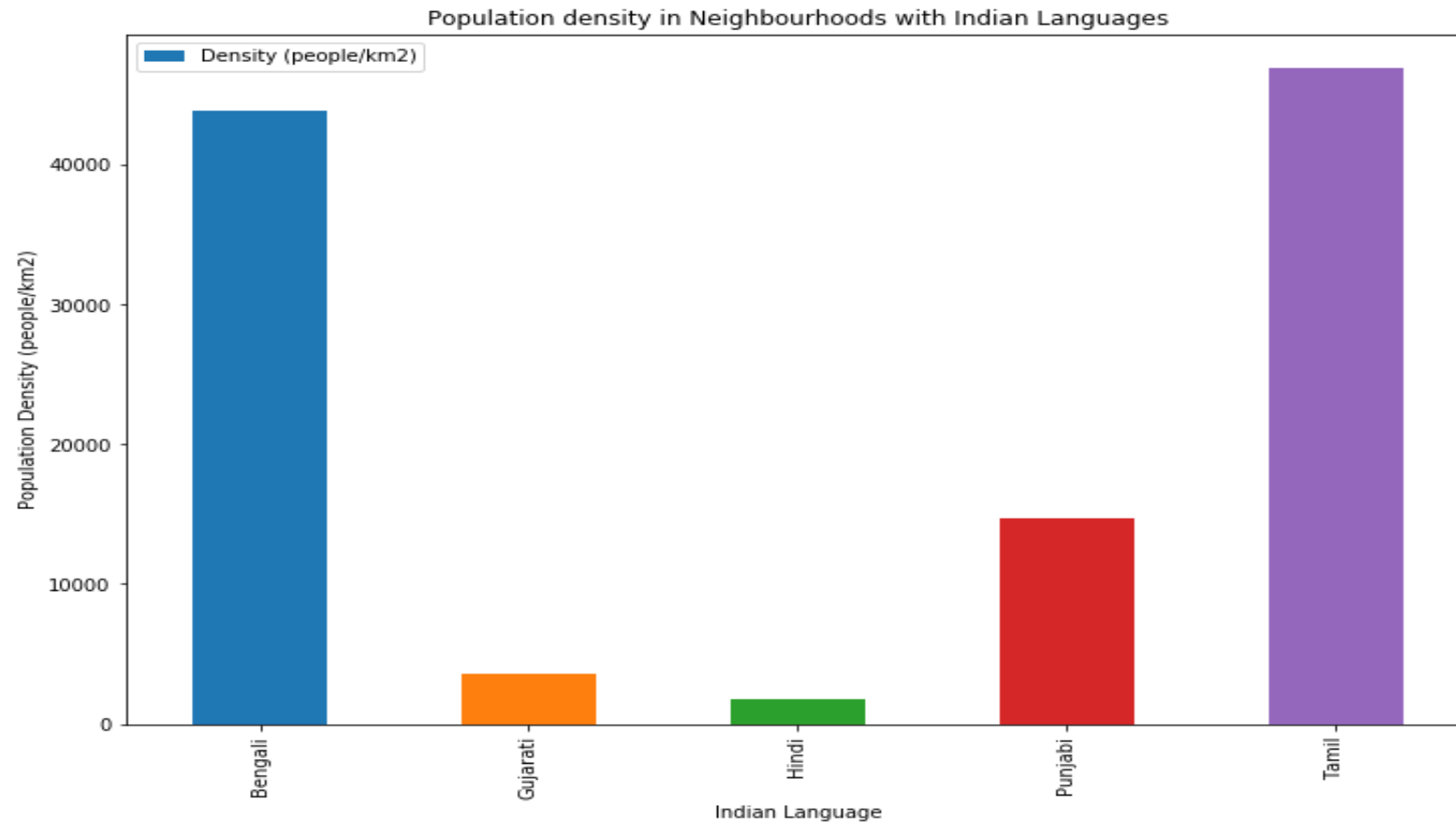
Indian languages in different neighborhoods of Toronto

Population in neighborhoods with Indian languages

# Population density in neighborhoods with Indian languages

Inference

- Identify a neighborhood
  - where Tamil is the most popular language (after English) and
  - with good number of restaurants

Machine Learning Algorithm

Goal: Identify a suitable neighborhood among several neighborhoods

- An unsupervised machine learning problem
- Clustering techniques can be used

- K means clustering
  - Simple and inexpensive
  - Efficient when working with large data sets

- Elbow method:
  - find the optimum value for the number of clusters, 'k'

Elbow Method for Finding Optimal Number of Clusters (k)
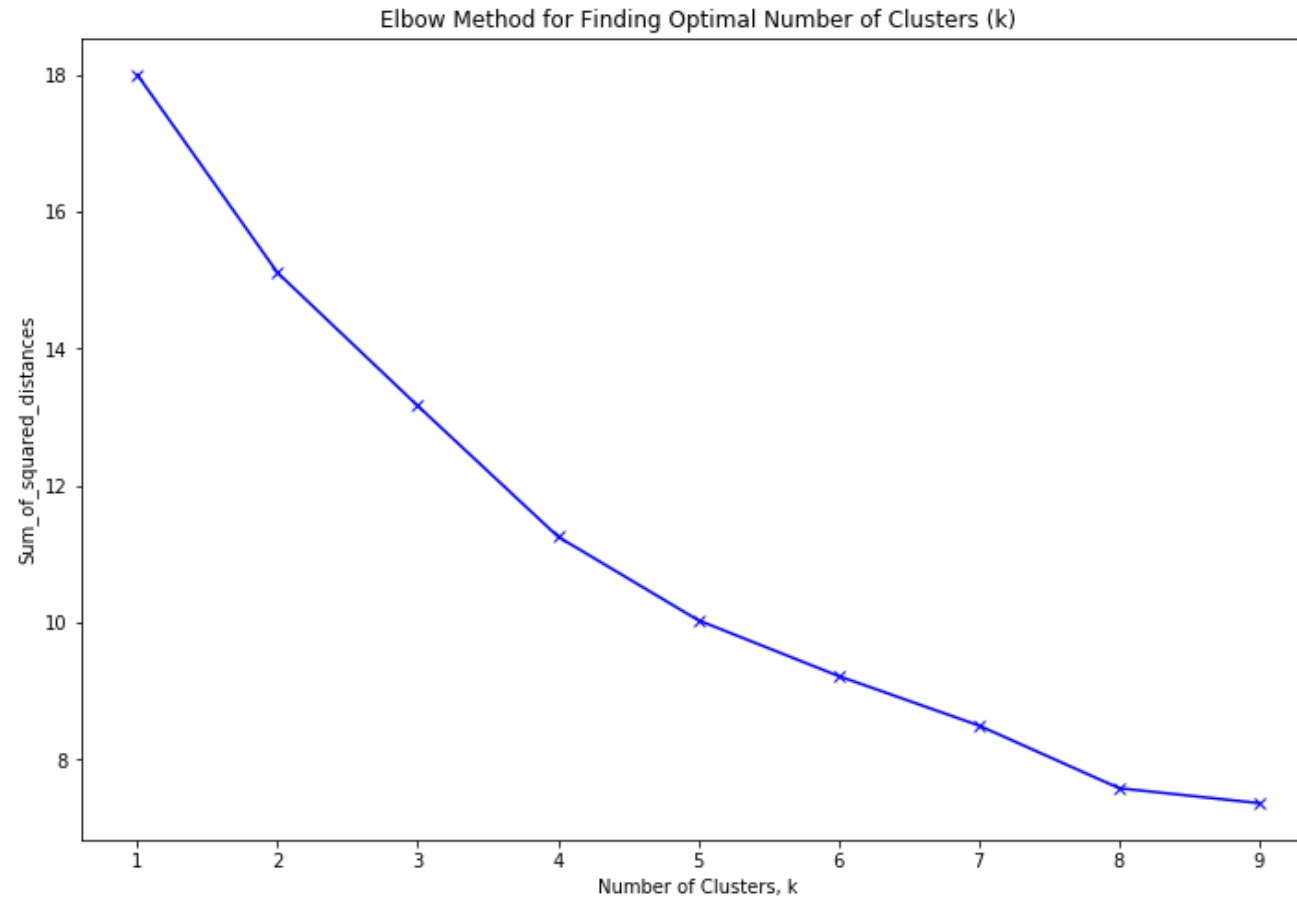
Results

- Each cluster was analysed to identify neighborhoods with Indian language and restaurants in the vicinity

- Few neighborhoods meeting the language requirement were observed, but do not include many restaurants

- The results show that 'Dorset Park' is the best choice
  - a neighborhood with Tamil language
  - includes relatively more number of restaurants

Neighborhoods of Toronto

Clusters of Neighbourhoods (k = 5)
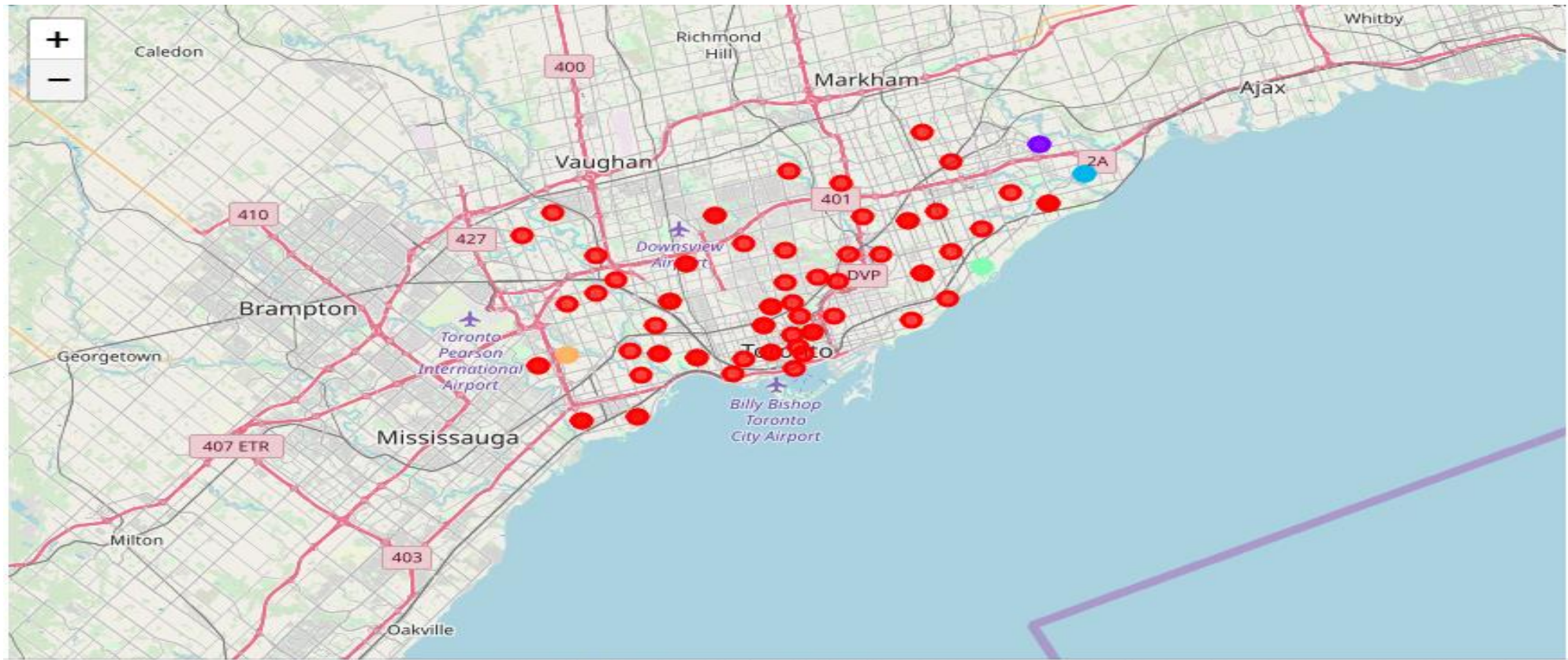
10 Most Common Venues in 'Dorset Park' Neighborhood

Results corresponding to 'Dorset Park' Neighbourhood

```
cluster_0[cluster_0['Neighbourhood']=='Dorset Park']
```

	Density (people/km2)	Language	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
9	3331	Tamil	0	Indian Restaurant	Latin American Restaurant	Vietnamese Restaurant	Pet Store	Chinese Restaurant	Yoga Studio	Food Court	Food & Drink Shop	Food	Fish Market

Discussion

- Optimum k value is 5
- First cluster
  - very dense
  - includes 67 neighborhoods
  - several restaurants, pizza points and coffee shops
- Second cluster
  - 2 identical neighborhoods
- Third cluster
  - 3 similar neighborhoods
- Fourth and Fifth clusters
  - 2 neighborhoods each

Conclusion

- Data visualization - powerful in drawing insights from data and towards addressing the business problem at hand

- Location Based Social Networks - provide the flexibility to solve several interesting problems

- In this capstone project, exploratory data analysis and Foursquare API data were used to identify a neighborhood for opening an 'Indian Arts, Dance and Cultural Centre'

- Innovative business solutions can be obtained by performing similar analysis on many problems related to several fields