Best Toronto Neighbourhoods to open a Persian Restaurant

Introduction

Arash Khosravi 28 Apr 2020

To be awarded a Data Science professional certificate from IBM, the candidate must demonstrate the knowledge on the real-world data to solve a business problem, me as a candidate decided to work on the real dataset related to Toronto area for opening a new business by analysing and combining several data sets. In this project I will gather the data from the web (Wikipedia) and by using the Foursquare location data will find out which neighborhood is/are best to open the restaurant.

First Step: Description of the Business Problem

Problem Statement: Prospects of opening a Persian Restaurant in Toronto, Canada.

Toronto, the biggest city of the province of Ontario, and the most populous Canadian city. This city is multicultural city with population of more than 6 million, each area has its own culture and nationality who live there, some of the known neighborhoods are, Greektown, Little Portugal, Little Italy, Koreatown, Little India, Persian Place and etc.
One of the most immigrant-friendly cities in North America with more than half of the entire Persian Canadian population residing in Toronto it is one of the best places to start a Persian restaurant.
In this project I will analyze the neighborhoods in Toronto to identify the most profitable area since the success of the restaurant depends on the people and ambience. we already know that Toronto shelter a greater number of Persians than any other city in Canada, it is a good idea to start the restaurant here, but we just need to make sure whether it is a profitable idea or not. If so, where we can place it, so it yields more profit to the owner.
This analysis will be a comprehensive guide to start or expand restaurants targeting the Persian crowd. This analysis will give an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business.

Who will be interested in this project:

- Business owners
- Business Analyst
- Persian People
- Persian food lovers

Step two
Data acquisition and cleaning

Data Sources

- I'm using wiki page to get the information related to each Postal Code and related neighborhood in Toronto. The data can be retrieved from the link https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

- I'm using "https://cocl.us/Geospatial_data" csv file to get all the geographical coordinates of the neighbourhoods.

- I'm going to identify the neighbourhoods which are densely populated with Persians, to get information about the distribution of population by their ethnicity I'm using "Demographics of Toronto" (https://en.m.wikipedia.org/wiki/Demographics_of_Toronto#Ethnic_diversity)

- To get location and other information about various venues in Toronto I'm using Foursquare's explore API. Using the Foursquare's explore API (which gives venues recommendations), I'm fetching details about the venues up present in Toronto and collected their names, categories and locations (latitude and longitude).

From Foursquare API (https://developer.foursquare.com/docs), I retrieved the following for each venue:

- **Name**: The name of the venue.
- **Category**: The category type as defined by the API.
- **Latitude**: The latitude value of the venue.
- **Longitude**: The longitude value of the venue.

Data Cleaning

Scraping Toronto Neighbourhoods Table from Wikipedia

Scraped the following Wikipedia page, "*List of Postal code of Canada: M*" in order to obtain the data about the Toronto & the Neighbourhoods in it.

- Assumptions made to attain the below DataFrame:

Dataframe will consist of three columns:

- PostalCode,
- Borough,
- Neighbourhood

- Only the cells that have an assigned borough will be processed. Borough that is not assigned are ignored.
- More than one neighbourhood can exist in one postal code area. =
- If a cell has a borough but a Not assigned neighbourhood, then the neighbourhood will be the same as the borough.

Wikipedia — package is used to scrape the data from wiki.

```
import wikipedia as wp
html = wp.page("List of postal codes of Canada: M").html().encode("UTF-8")
df = pd.read_html(html, header = 0)[0]
df = pd.DataFrame(df)
df
```

	Postal code	Borough	Neighborhood
0	M1A	Not assigned	NaN
1	M2A	Not assigned	NaN
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park / Harbourfront
...	...	...	...
175	M5Z	Not assigned	NaN
176	M6Z	Not assigned	NaN
177	M7Z	Not assigned	NaN
178	M8Z	Etobicoke	Mimico NW / The Queensway West / South of Bloo...
179	M9Z	Not assigned	NaN

180 rows × 3 columns

After some cleaning we got the proper DataFrame with the Postal code, Borough & Neighborhood information.

	Borough	PostalCode	Neighborhood
0	Central Toronto	M4N	Lawrence Park
1	Central Toronto	M4P	Davisville North
2	Central Toronto	M4R	North Toronto West
3	Central Toronto	M4S	Davisville
4	Central Toronto	M4T	Moore Park / Summerhill East

Adding geographical coordinates to the neighborhoods

Next important step is adding the geographical coordinates to these neighborhoods. To do so I'm extracting the data present in the Geospatial Data csv file and I'm combining it with the existing neighborhood DataFrame by merging them both based on the postal code.

	PostalCode	Latitude	Longitude
0	M1B	43.806686	-79.194353
1	M1C	43.784535	-79.160497
2	M1E	43.763573	-79.188711
3	M1G	43.770992	-79.216917
4	M1H	43.773136	-79.239476

I'm renaming the columns to match the existing DataFrame formed from 'List of Postal code of Canada: M' wiki page. After that I'm merging both the DataFrame into one by merging on the postal code.

	Borough	PostalCode	Neighborhood	Latitude	Longitude
0	Central Toronto	M4N	Lawrence Park	43.728020	-79.388790
1	Central Toronto	M4P	Davisville North	43.712751	-79.390197
2	Central Toronto	M4R	North Toronto West	43.715383	-79.405678
3	Central Toronto	M4S	Davisville	43.704324	-79.388790
4	Central Toronto	M4T	Moore Park / Summerhill East	43.689574	-79.383160

```
print('The dataframe has {} boroughs and {} neighborhoods.'.format(
        len(toronto_DF['Borough'].unique()),
        toronto_DF.shape[0]
    )
)
```

```
The dataframe has 10 boroughs and 103 neighborhoods.
```

Scrap the distribution of population from Wikipedia

Another factor that can help us in deciding which neighborhood would be best option to open a restaurant is, the distribution of population based on the ethnic diversity for each neighborhood. As this helps us in identifying the neighborhoods which are densely populated with Persian crowd since that neighborhood would be an ideal place to open a Persian restaurant.

Scraped the following Wikipedia page, "***Demographics of Toronto***" in order to obtain the data about the Toronto & the Neighborhoods in it. Compared to all the neighborhoods in Toronto below given neighborhoods only had considerable amount of Persian crowd. We are examining those neighborhood's population to identify the densely populated neighborhoods with Persian population.

```
#overall population distribution
html = wp.page("Demographics of Toronto").html().encode("UTF-8")
```

There were only six neighborhoods in Toronto which Persian population spread across, so we are gathering the population, it's percentage in each riding in those neighborhoods.

	Riding	Population	Ethnic Origin #1	Ethnic Origin 1 in %	Ethnic Origin #2	Ethnic Origin 2 in %	Ethnic Origin #3	Ethnic Origin 3 in %	Ethnic Origin #4	Ethnic Origin 4 in %	Ethnic Origin #5	Ethnic Origin 5 in %	Ethnic Origin #6	Ethnic Origin 6 in %	Ethnic Origin #7	Ethnic Origin 7 in %	Ethnic Origin #8	Ethnic Origin 8 in %	Ethnic Origin #9	Ethnic Origin 9 in %
0	Spadina-Fort York	114315	English	16.4	Chinese	16.0	Irish	14.6	Canadian	14.0	Scottish	13.2	French	7.70	German	7.6	NaN	NaN	NaN	NaN
1	Beaches-East York	108435	English	24.2	Irish	19.9	Canadian	19.7	Scottish	18.9	French	8.7	German	8.40	NaN	NaN	NaN	NaN	NaN	NaN
2	Davenport	107395	Portuguese	22.7	English	13.6	Canadian	12.8	Irish	11.5	Italian	11.1	Scottish	11.00	NaN	NaN	NaN	NaN	NaN	NaN
3	Parkdale-High Park	106445	English	22.3	Irish	20.0	Scottish	18.7	Canadian	16.1	German	9.8	French	8.88	Polish	8.5	NaN	NaN	NaN	NaN
4	Toronto-Danforth	105395	English	22.9	Irish	19.5	Scottish	18.7	Canadian	18.4	Chinese	13.8	French	8.86	German	8.8	Greek	7.3	NaN	NaN
5	Toronto-St. Paul's	104940	English	18.5	Canadian	16.1	Irish	15.2	Scottish	14.8	Polish	10.3	German	7.90	Russian	7.7	Italian	7.3	French	7.2
6	University-Rosedale	100520	English	20.6	Irish	16.6	Scottish	16.3	Canadian	15.2	Chinese	14.7	German	8.70	French	7.7	Italian	7.4	NaN	NaN
7	Toronto Centre	99590	English	15.7	Canadian	13.7	Irish	13.4	Scottish	12.6	Chinese	12.5	French	7.20	NaN	NaN	NaN	NaN	NaN	NaN

Get location data using Foursquare

Foursquare API is very useful online application used my many developers & other applications like Uber etc. In this project I have used it to retrieve information about the places present in the neighborhoods of Toronto. The API returns a JSON file and we need to turn that into a data-frame. Here I've chosen 100 popular spots for each neighborhood within a radius of 1km.

```
In [32]: toronto_venues.head(10)
```

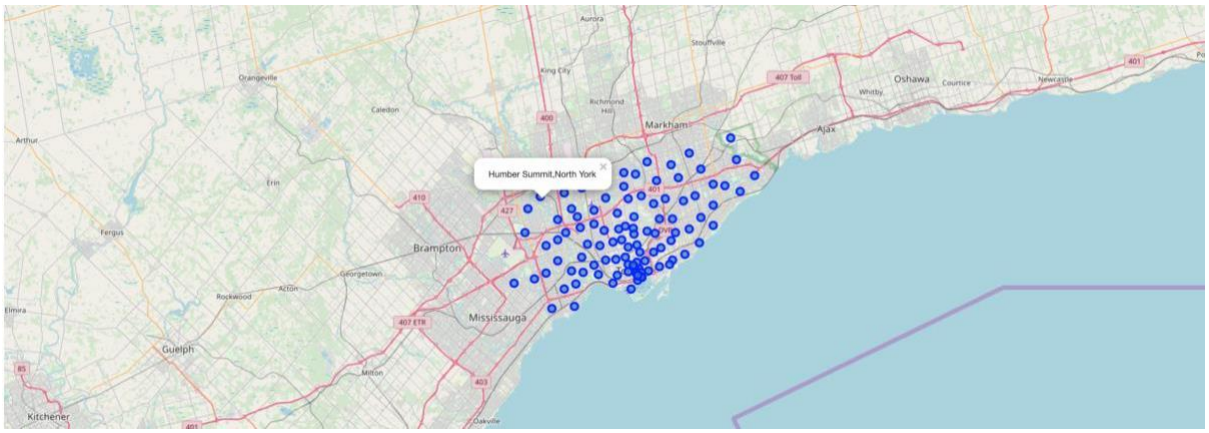
Out[32]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Lawrence Park	43.728020	-79.388790	Lawrence Park Ravine	43.726963	-79.394382	Park
1	Lawrence Park	43.728020	-79.388790	Zodiac Swim School	43.728532	-79.382860	Swim School
2	Lawrence Park	43.728020	-79.388790	TTC Bus #162 - Lawrence-Donway	43.728026	-79.382805	Bus Line
3	Davisville North	43.712751	-79.390197	Homeway Restaurant & Brunch	43.712641	-79.391557	Breakfast Spot
4	Davisville North	43.712751	-79.390197	Summerhill Market North	43.715499	-79.392881	Food & Drink Shop
5	Davisville North	43.712751	-79.390197	Sherwood Park	43.716551	-79.387776	Park
6	Davisville North	43.712751	-79.390197	Winners	43.713236	-79.393873	Clothing Store
7	Davisville North	43.712751	-79.390197	Best Western Roehampton Hotel & Suites	43.708878	-79.390880	Hotel
8	Davisville North	43.712751	-79.390197	Subway	43.708378	-79.390473	Sandwich Place
9	Davisville North	43.712751	-79.390197	Gym	43.713126	-79.393537	Gym

Exploratory Data Analysis:

Folium Library and Leaflet Map

Folium is a python library; I'm using it to draw an interactive leaflet map using coordinate data.



Relationship between neighbourhood and Persian Restaurant

First, we will extract the Neighbourhood and Persian Restaurant column from the above Toronto DataFrame for further analysis:

	Neighborhood	Yoga Studio	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store	Asian Restaurant	Athletics & Sports	Auto Garage
0	Adelaide, King, Richmond	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.030000	0.000000	0.0	0.010000	0.010000	0.000000	0.03	0.000000	0.0
1	Agincourt	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00	0.000000	0.0
2	Agincourt North, L'Amoreaux East, Milliken, St...	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00	0.000000	0.0
3	Albion Gardens, Beaumond Heights, Humbergate, ...	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000	0.00	0.000000	0.0

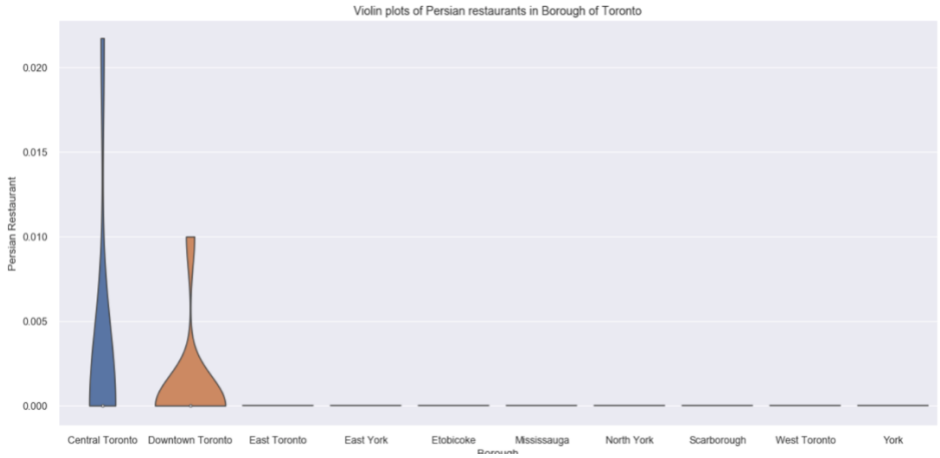
After performing pandas one hot encoding for the venue categories, let us merge this DataFrame with the Toronto DataFrame with latitude & longitude information on neighbourhood. Finally extract just the Persian restaurant values along with neighbourhood information.

```
toronto_merged = pd.merge(toronto_DF, toronto_part, on='Neighborhood')
toronto_merged
```

]:

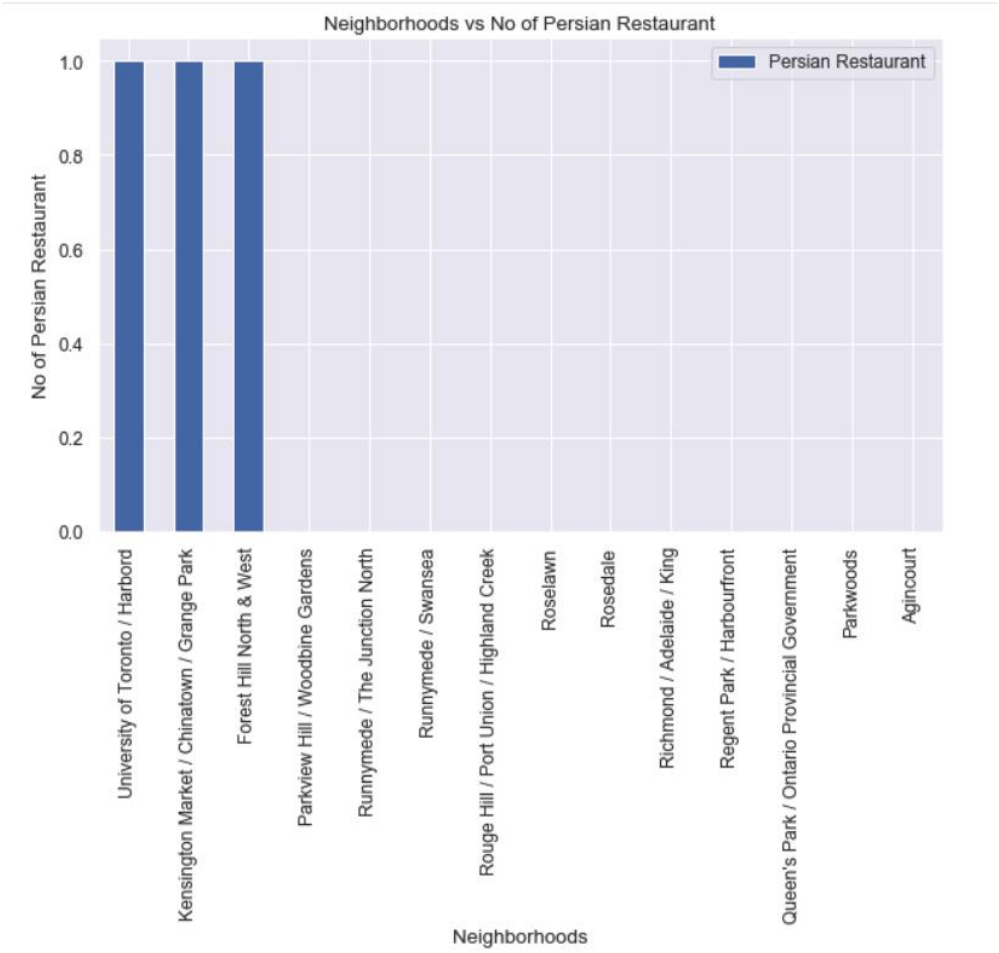
	Borough	Postalcode	Neighborhood	Latitude	Longitude	Cluster Labels	Indian Restaurant
0	Central Toronto	M4N	Lawrence Park	43.728020	-79.388790	0	0.000000
1	Central Toronto	M4P	Davisville North	43.712751	-79.390197	0	0.000000
2	Central Toronto	M4R	North Toronto West	43.715383	-79.405678	0	0.000000
3	Central Toronto	M4S	Davisville	43.704324	-79.388790	4	0.028571
4	Central Toronto	M5N	Roselawn	43.711695	-79.416936	0	0.000000
5	Downtown Toronto	M4W	Rosedale	43.679563	-79.377529	0	0.000000
6	Downtown Toronto	M4Y	Church and Wellesley	43.665860	-79.383160	5	0.011765
7	Downtown Toronto	M5C	St. James Town	43.651494	-79.375418	5	0.010000
8	Downtown Toronto	M5E	Berczy Park	43.644771	-79.373306	4	0.017544
9	Downtown Toronto	M5G	Central Bay Street	43.657952	-79.387383	5	0.011364
10	Downtown Toronto	M5W	Stn A PO Boxes 25 The Esplanade	43.646435	-79.374846	5	0.010101

Let’s try to draw some plot using the above DataFrame:



With the help of this violin plots we can identify the boroughs with densely populated Persian restaurants. It is drawn using seaborn library to show the distribution of Persian restaurants in different boroughs.

Let’s also visualize the neighbourhood with Persian Restaurants:



Relationship between neighbourhood and Persian population

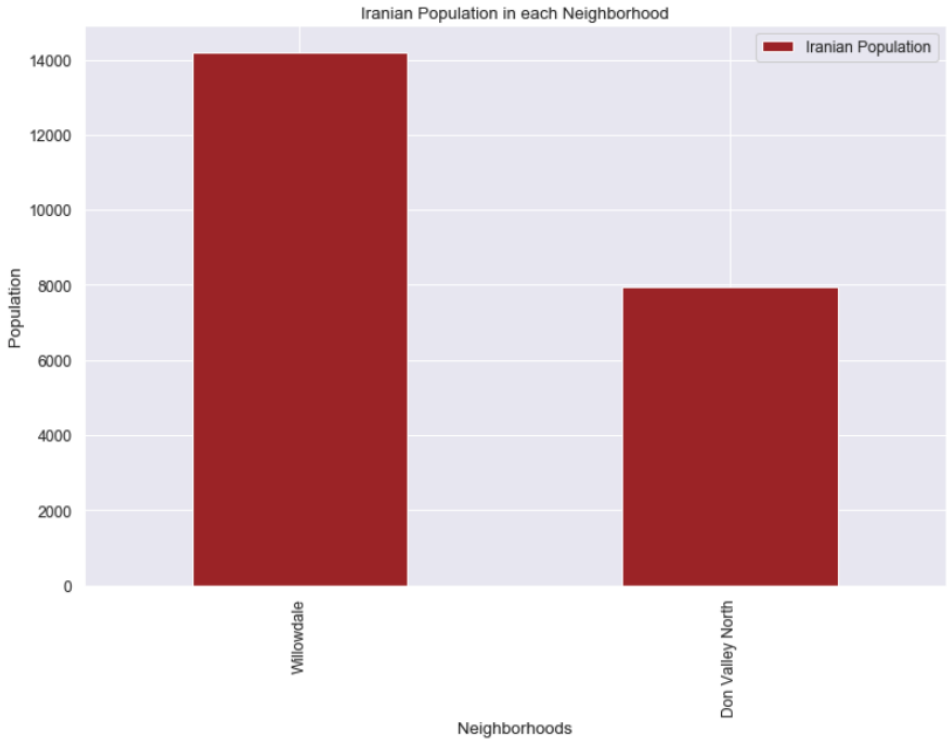
Another key feature is the distribution of Persian crowd in each neighbourhood. Let us analyse the neighbourhoods and identify the neighbourhoods with highest number of Persian populations.

To achieve that we are joining all the neighbourhood's DataFrame from using the wiki page with ethnic population and in that we are extracting just the Persian population for each neighbourhood.

```
pop_persian_df
```

	Riding	Population	Ethnic Origin #1	Ethnic Origin 1 in %	Ethnic Origin #2	Ethnic Origin 2 in %	Ethnic Origin #3	Ethnic Origin 3 in %	Ethnic Origin #4	Ethnic Origin 4 in %	Ethnic Origin #5	Ethnic Origin 5 in %	Ethnic Origin #6	Ethnic Origin 6 in %	Ethnic Origin #7	Ethnic Origin 7 in %	Ethnic Origin #8	Ethnic Origin 8 in %
0	Willowdale	117405	Chinese	25.9	Iranian	12.1	Korean	10.6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	Don Valley North	109060	Chinese	32.4	East Indian	7.3	Iranian	7.3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Let's draw a graph to visualize the population spread in neighborhoods:



This analysis & visualization of the relationship between neighbourhoods & Persian population present in those neighbourhoods helps us in identifying the highly populated Persian neighbourhoods. Once we identify those neighbourhoods it helps us in deciding where to place the new Persian restaurant. Persian restaurant placed in a densely populated Persian neighbourhood is more likely to get more Persian customers than a restaurant placed in a neighbourhood with less or no Persian population. Thus, this analysis helps in the determining the success of the new Persian restaurant.

Relationship between Persian population and Persian restaurant

After performing the data cleaning & data analysis we couldn't identify a big relationship established between densely populated Persian neighborhoods & number of Persian restaurants.

This might be because of the missing in data as this an area which can improved in future analysis to get a more insight about the business problem.

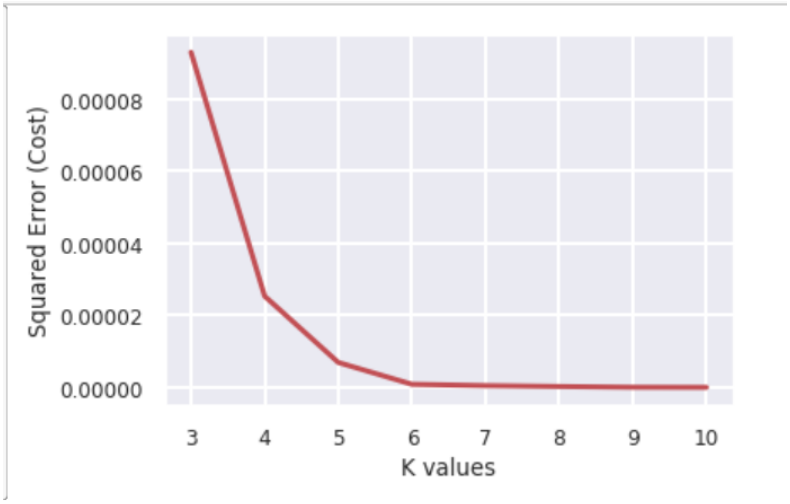
]:

	Iranian Population	Neighborhood
0	7961.38	Henry Farm
1	7961.38	Bayview Village
2	7961.38	Bayview Woods-Steeles
3	7961.38	Hillcrest Village
4	7961.38	Don Valley Village
5	7961.38	and Pleasant View

Predictive Modelling

Clustering Neighbourhoods of Toronto

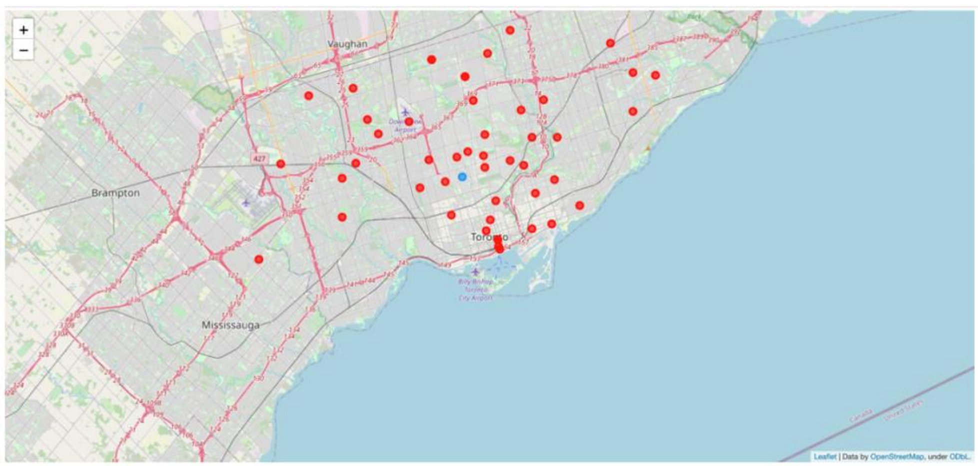
First step in K-means clustering is to identify best K value meaning the number of clusters in a given dataset. To do so we are going to use the elbow method on the Toronto dataset with Persian restaurant percentage.



After analyzing using elbow method using distortion score & Squared error for each K value, looks like K = 6 is the best value.

```
array([2, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0], dtype=int32)
```

After assigning the clusters to the dataset we can plot the map to see the clusters



Examine the Clusters

We have total of 6 clusters such as 0,1,2,3,4,5. Let us examine one after the other.

Cluster 0 contains all the neighbourhoods which has least number of Persian restaurants. It is shown in red colour in the map

```
#Cluster 0
toronto_merged.loc[toronto_merged['Cluster Labels'] == 0]
```

	Borough	PostalCode	Neighborhood	Latitude	Longitude	Cluster Labels	Persian Restaurant
0	Central Toronto	M4N	Lawrence Park	43.728020	-79.388790	0.0	0.0
1	Central Toronto	M4P	Davisville North	43.712751	-79.390197	0.0	0.0
2	Central Toronto	M4R	North Toronto West	43.715383	-79.405678	0.0	0.0
3	Central Toronto	M4S	Davisville	43.704324	-79.388790	0.0	0.0
4	Central Toronto	M5N	Roselawn	43.711695	-79.416936	0.0	0.0
6	Downtown Toronto	M4W	Rosedale	43.679563	-79.377529	0.0	0.0
7	Downtown Toronto	M4Y	Church and Wellesley	43.665860	-79.383160	0.0	0.0
8	Downtown Toronto	M5C	St. James Town	43.651494	-79.375418	0.0	0.0
9	Downtown Toronto	M5C	St. James Town	43.651494	-79.375418	0.0	0.0
10	Downtown Toronto	M5E	Berczy Park	43.644771	-79.373306	0.0	0.0
11	Downtown Toronto	M5G	Central Bay Street	43.657952	-79.387383	0.0	0.0
12	Downtown Toronto	M5W	Stn A PO Boxes	43.646435	-79.374846	0.0	0.0
13	Downtown Toronto	M6G	Christie	43.669542	-79.422564	0.0	0.0
14	East Toronto	M4E	The Beaches	43.676357	-79.293031	0.0	0.0
15	East Toronto	M4M	Studio District	43.659526	-79.340923	0.0	0.0
16	East Toronto	M7Y	Business reply mail Processing CentrE	43.662744	-79.321558	0.0	0.0
17	East York	M4C	Woodbine Heights	43.695344	-79.318389	0.0	0.0
18	East York	M4G	Leaside	43.709060	-79.363452	0.0	0.0
19	East York	M4H	Thorncliffe Park	43.705369	-79.349372	0.0	0.0
20	East York	M4J	East Toronto	43.685347	-79.338106	0.0	0.0
21	Etobicoke	M9A	Islington Avenue	43.667856	-79.532242	0.0	0.0

Cluster 2 contains the neighborhood which is sparsely populated with Persian restaurants. It is shown in purple color in the map.

```
#Cluster 2
toronto_merged.loc[toronto_merged['Cluster Labels'] == 2]
```

	Borough	PostalCode	Neighborhood	Latitude	Longitude	Cluster Labels	Persian Restaurant
5	Central Toronto	M5P	Forest Hill North & West	43.696948	-79.411307	2.0	0.020833

Cluster 1& 3 & 4 & 5 has no rows meaning no data points or no neighborhood was near to these centroids.

```
#Cluster 3
toronto_merged.loc[toronto_merged['Cluster Labels'] == 3]
```

Borough	PostalCode	Neighborhood	Latitude	Longitude	Cluster Labels	Persian Restaurant

Cluster 4 has no rows meaning no data points or neighborhood was near to this centroid.

```
#Cluster 4
toronto_merged.loc[toronto_merged['Cluster Labels'] == 4]
```

Borough	PostalCode	Neighborhood	Latitude	Longitude	Cluster Labels	Persian Restaurant

Results and Discussion

Results

We have reached the end of the analysis; in this section we will document all the findings from above clustering & visualization of the dataset. In this project, we started off with the business problem of identifying a good neighbourhood to open a new Persian restaurant. To achieve that we looked into all the neighbourhoods in Toronto, analysed the Persian population in each neighbourhood & number of Persian restaurants in those neighbourhoods to come to conclusion about which neighbourhood would be a better spot. We have used variety of data sources to set up a very realistic data-analysis scenario. We have found out that —

• In those 11 boroughs we identified that only **Central Toronto, Downtown Toronto, North York** have high number of Persian restaurants with the help of Violin plots between Number of Persian restaurants in Borough of Toronto.

• In all the ridings, Willowdale, Don Valley North are the densely populated with Persian crowd ridings.

• With the help of clusters examining & violin plots looks like Downtown Toronto, Central Toronto, East York are already densely populated with Persian restaurants. So, it is better idea to leave those boroughs out and consider only, East Toronto & North York for the new restaurant's location.

• After careful consideration it is a good idea to open a new Persian restaurant in Willowdale since it has high number of Persian populations which gives a higher number of customers possibility and lower competition since very less Persian restaurants in the neighbourhoods.

Discussion

According to this analysis, Willowdale borough will provide the least competition for the new upcoming Persian restaurant as there is very little Persian restaurants spread or no Persian restaurants in few neighbourhoods. Also looking at the population distribution looks like it is densely populated with Persian crowd which helps the new restaurant by providing high customer visit possibility. So, definitely this region could potentially be a perfect place for starting a quality Persian restaurant. Some of the drawbacks of this analysis are — the clustering is completely based only on data obtained from Foursquare API and the data about the Persian population distribution in each neighbourhood is also based on the 2016 census which is not up-to date. Thus, there is huge gap of 3 years in the population distribution data. Even Though there are lots of areas where it can be improved yet this analysis has certainly provided us with some good insights, preliminary information on possibilities & a head start into this business problem by setting the step stones properly.

Conclusion

Finally, to conclude this project, we have got a chance to on a business problem like how a real like data scientists would do. We have used many python libraries to fetch the data, to manipulate the contents & to analyse and visualize those datasets. We have made use of Foursquare API to explore the venues in neighbourhoods of Toronto, then get good amount of data from Wikipedia which we scraped with help of Wikipedia python library and visualized using various plots present in seaborn & matplotlib. We also applied machine learning technique to predict the output given the data and used Folium to visualize it on a map.
Some of the drawbacks or areas of improvements shows us that this analysis can be further improved with the help of more data and different machine learning technique. Similarly, we can use this project to analysis any scenario such as opening a different cuisine restaurant or opening of a new gym and etc. Hopefully, this project helps acts as initial guidance to take more complex real-life challenges using data-science.