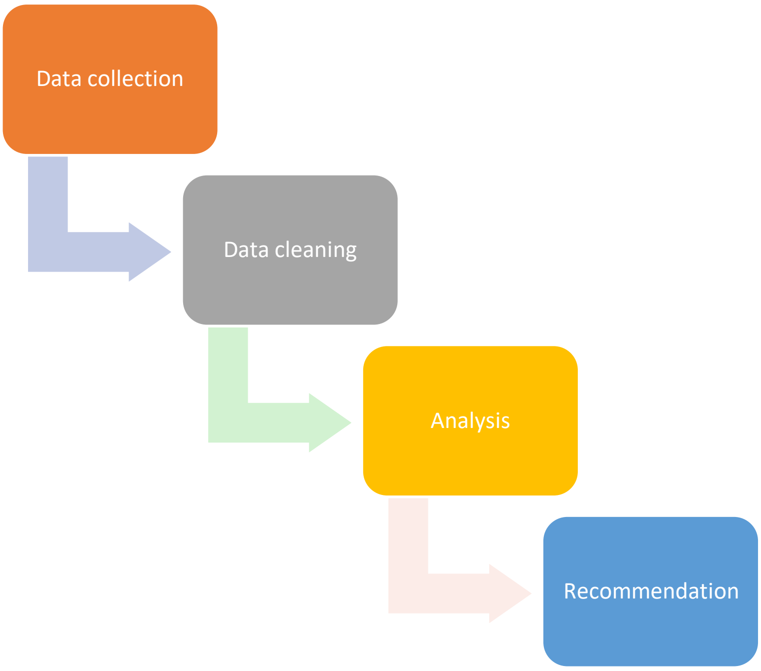
BUSINESS PROBLEM

In the given problem we're trying to find the best location for a restaurant in Toronto. The solution derived from all the data present, will help everyone who could be looking for a place in the city for a new venture. In such big cities finding the optimum location is a daunting task, but with the help of data analysis of various locations, best possible locations could be pinned which could bring monetary benefits to the owners. Population of neighbourhood will help in determining, the sites where highest no. of people would come to visit. Also, places with highly crowded restaurants have to skipped to decrease the entry barrier of a new player. Vicinity to the business centres will also play a key role.

Data science will be used to determine the neighbourhoods with the best quality of footfalls in the restaurant. All the stakeholders shall be benefitted from this location.

We'll use simple macro steps to reach at our target.


Data collection

Data cleaning

Analysis

Recommendation

DATA

Our problem clearly requires a certain type of data set to come upon the best possible outcomes.

- Wikipedia will be used to get the postal codes of various locations

```
from urllib.request import urlopen
wiki =
"https://en.wikipedia.org/wiki/List_of_postal_codes_of_Ca
nada:_M"
```

- Foursquares API will be used to get exact locations, no. of restaurants and type of restaurants.

We are getting following data from these credible sources:

- Postal codes
- Longitudes and latitudes
- Distance from centre of the city
- Age group of people
- Income of people
- Existing restaurants in the neighbourhood

Set of data obtained

	Postal_Code	**Borough**	**Neighborhood**
0	M3A\n	North York\n	Parkwoods\n
1	M4A\n	North York\n	Victoria Village\n
2	M5A\n	Downtown Toronto\n	Regent Park, Harbourfront\n
3	M6A\n	North York\n	Lawrence Manor, Lawrence Heights\n
4	M7A\n	Downtown Toronto\n	Queen's Park, Ontario Provincial Government\n
5	M9A\n	Etobicoke\n	Islington Avenue, Humber Valley Village\n
6	M1B\n	Scarborough\n	Malvern, Rouge\n
7	M3B\n	North York\n	Don Mills\n
8	M4B\n	East York\n	Parkview Hill, Woodbine Gardens\n
9	M5B\n	Downtown Toronto\n	Garden District, Ryerson\n
10	M6B\n	North York\n	Glencairn\n
11	M9B\n	Etobicoke\n	West Deane Park, Princess Gardens, Martin Grov...
12	M1C\n	Scarborough\n	Rouge Hill, Port Union, Highland Creek\n
	Postal_Code	Borough	Neighborhood
0	M3A\n	North York\n	Parkwoods\n
1	M4A\n	North York\n	Victoria Village\n
2	M5A\n	Downtown Toronto\n	Regent Park, Harbourfront\n
3	M6A\n	North York\n	Lawrence Manor, Lawrence Heights\n
4	M7A\n	Downtown Toronto\n	Queen's Park, Ontario Provincial Government\n
5	M9A\n	Etobicoke\n	Islington Avenue, Humber Valley Village\n
6	M1B\n	Scarborough\n	Malvern, Rouge\n
7	M3B\n	North York\n	Don Mills\n
8	M4B\n	East York\n	Parkview Hill, Woodbine Gardens\n
9	M5B\n	Downtown Toronto\n	Garden District, Ryerson\n
10	M6B\n	North York\n	Glencairn\n
11	M9B\n	Etobicoke\n	West Deane Park, Princess Gardens, Martin Grov...
12	M1C\n	Scarborough\n	Rouge Hill, Port Union, Highland Creek\n

(Goes further up to 102 postal codes)

METHODOLOGY

We have taken following steps:

- We have collected data from aforementioned sources.
- Location (in terms of longitude and latitude) along with the category of restaurant.
- We have identified restaurants' density in various areas.
- We'll use K-mean to cluster areas with promising potential.
- Chosen clusters are made to be close to less than a km away from centre of the city.
- Map using Folium will be displayed to give a practical outlook of the chosen locations.

ANALYSIS

1. **Identification and cleaning**
   We have to identify and capture the data from all mentioned sources. Some portion of the data is missing, so we need to clean that portion out of our dataset.
   Raw data obtained from Wikipedia:

	Postal_Code	Borough	Neighborhood
0	M1A\n	Not assigned\n	Not assigned\n
1	M2A\n	Not assigned\n	Not assigned\n
2	M3A\n	North York\n	Parkwoods\n
3	M4A\n	North York\n	Victoria Village\n
4	M5A\n	Downtown Toronto\n	Regent Park, Harbourfront\n
5	M6A\n	North York\n	Lawrence Manor, Lawrence Heights\n
6	M7A\n	Downtown Toronto\n	Queen's Park, Ontario Provincial Government\n
7	M8A\n	Not assigned\n	Not assigned\n
8	M9A\n	Etobicoke\n	Islington Avenue, Humber Valley Village\n
9	M1B\n	Scarborough\n	Malvern, Rouge\n
10	M2B\n	Not assigned\n	Not assigned\n
11	M3B\n	North York\n	Don Mills\n
12	M4B\n	East York\n	Parkview Hill, Woodbine Gardens\n

2. **Combining data sources**
   With postal address and longitude-latitude from different data source present with us, we need to combine all of them.

	Postal_Code	Borough	Neighborhood
0	M5A\n	Downtown Toronto\n	Regent Park, Harbourfront\n
1	M7A\n	Downtown Toronto\n	Queen's Park, Ontario Provincial Government\n

2	M5B\n	Downtown Toronto\n	Garden District, Ryerson\n
3	M5C\n	Downtown Toronto\n	St. James Town\n
4	M4E\n	East Toronto\n	The Beaches\n
5	M5E\n	Downtown Toronto\n	Berczy Park\n
6	M5G\n	Downtown Toronto\n	Central Bay Street\n
7	M6G\n	Downtown Toronto\n	Christie\n
8	M5H\n	Downtown Toronto\n	Richmond, Adelaide, King\n
9	M6H\n	West Toronto\n	Dufferin, Dovercourt Village\n
10	M5J\n	Downtown Toronto\n	Harbourfront East, Union Station, Toronto Isla...
11	M6J\n	West Toronto\n	Little Portugal, Trinity\n
12	M4K\n	East Toronto\n	The Danforth West, Riverdale\n

3. **Sorting neighbourhood on the basis of latitude and longitude**
   From previous step, a resulting data set will contain data about neighbourhood, Its postal code and latitude-longitudes
4. **Clustering**
   We have used K-cluster algorithm to cluster various neighbourhoods. Each cluster is analysed on the basis of distinguishing features. No. of restaurants and types are our determining variables.
5. **Frequency of visits**
   Obtain the frequency of visits in the restaurants and rank them

Subset of obtained set of frequencies

```
----Berczy Park
----
                venue  freq
0         Coffee Shop  0.08
1          Restaurant  0.05
2                Café  0.05
3  Seafood Restaurant  0.04
4               Hotel  0.04


----Christie
----
               venue  freq
0  Korean Restaurant  0.22
1        Coffee Shop  0.05
2                Pub  0.03
3      Grocery Store  0.03
4     Ice Cream Shop  0.03


----Davisville
```

```
----
                venue  freq
0     Sushi Restaurant  0.10
1   Italian Restaurant  0.10
2          Coffee Shop  0.07
3                 Park  0.05
4    Convenience Store  0.05


----Davisville North
----
                venue  freq
0     Sushi Restaurant  0.10
1   Italian Restaurant  0.10
2          Coffee Shop  0.07
3                 Park  0.05
4    Convenience Store  0.05
```

RESULT & CONCLUSION

In our result we have found 65 neighbourhoods inside the geographical co-ordinates obtained using Foursqaure API. Out of the 5 clusters ,1 cluster shows the perfect density for opening of the restaurant.

St James Town comes out to be the tight choice to open restaurant as it's neither the densest area creating an entry barrier nor low crowded area which might make the project economically viable.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Category
St. James Town\n	43.669403	-79.372704	Mr. Jerk	Caribbean Restaurant
St. James Town\n	43.669403	-79.372704	Cranberries	Diner
St. James Town\n	43.669403	-79.372704	Murgatroid	Restaurant
St. James Town\n	43.669403	-79.372704	F'Amelia	Italian Restaurant
St. James Town\n	43.669403	-79.372704	Cabbagetown Brew	43.666923