1. **Data acquisition and cleaning**

- **Data sources**

  First we need all the neighborhoods with its postal code , we took them from Wikipedia https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

  I scraped the data and add it to a data frame

[277]:

	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

  A csv file for the latitude and longitude with the postal codes of Canada https://cocl.us/Geospatial_data also downloaded into a data frame

[280]:

	Postal Code	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

  Also we need the statistics of the population of Canada and its neighborhoods with the

  Average salary and second most common language which may help us to decide what kind of restaurant we need to open this also we found in Wikipedia

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

after scraping the data and downloaded into a data frame

[108]:

	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
5	Amesbury	NY	0280.00, 0281.01, 0281.02	17318	3.51	4934	1.1	27546	16.4	19.7	Spanish (6.1%)	06.1% Spanish	NaN
6	Armour Heights	NY	0298.00	4384	2.29	1914	2.0	116651	10.8	16.1	Russian (9.4%)	09.4% Russian	NaN
7	Banbury	NY	0267.00	6641	2.72	2442	5.0	92319	6.1	4.8	Unspecified Chinese (5.1%)	05.1% Unspecified Chinese	NaN
8	Bathurst Manor	NY	0297.01, 0310.01, 0310.02	14945	4.69	3187	12.3	34169	13.4	18.6	Russian (9.5%)	09.5% Russian	NaN
10	Bayview Village	NY	0305.01, 305.02	12280	4.14	2966	41.6	46752	14.4	15.6	Cantonese (8.4%)	08.4% Cantonese	NaN

- **Data Cleaning**

  Data downloaded from the csv and scraped from the Wikipedia source for the neighborhoods were combined into one table , first I removed the rows in the first data frame when the BOROUGH is not assigned as it as no value to me , then I merged any rows which has the same postal codes , after that I created my data frame after including only York neighborhoods and exclude all others

[282]:

	Postal Code	Borough	Neighborhood	Latitude	Longitude
17	M2H	North York	Hillcrest Village	43.803762	-79.363452
18	M2J	North York	Fairview, Henry Farm, Oriole	43.778517	-79.346556
19	M2K	North York	Bayview Village	43.786947	-79.385975
20	M2L	North York	York Mills, Silver Hills	43.757490	-79.374714
21	M2M	North York	Willowdale, Newtonbrook	43.789053	-79.408493

After the data frame is ready I used the foursquare API in order to get all the venues for York region using the previous data frame and printing the venues were returned by Foursquare then I added the categories of the venues

[273]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Hillcrest Village	43.803762	-79.363452	Eagle's Nest Golf Club	43.805455	-79.364186	Golf Course
1	Hillcrest Village	43.803762	-79.363452	AY Jackson Pool	43.804515	-79.366138	Pool
3	Hillcrest Village	43.803762	-79.363452	Duncan Creek Park	43.805539	-79.360695	Dog Run
4	Fairview, Henry Farm, Oriole	43.778517	-79.346556	The LEGO Store	43.778207	-79.343483	Toy / Game Store
5	Fairview, Henry Farm, Oriole	43.778517	-79.346556	DAVIDsTEA	43.777593	-79.345089	Tea Room

After I got the data frame of York venues with its categories I filtered it out and created my data frame with restaurant categories in order to use it in clustering and statistics later , also another data frame created which has all venues except the restaurant categories

The Restaurant venue data frame :

1]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
2	Hillcrest Village	43.803762	-79.363452	Villa Madina	43.801685	-79.363938	Mediterranean Restaurant
11	Fairview, Henry Farm, Oriole	43.778517	-79.346556	New York Fries - Fairview Mall	43.778605	-79.343577	Restaurant
17	Fairview, Henry Farm, Oriole	43.778517	-79.346556	Moxie's Classic Grill	43.777779	-79.343185	American Restaurant
24	Fairview, Henry Farm, Oriole	43.778517	-79.346556	Thai Express	43.777990	-79.344091	Restaurant
30	Fairview, Henry Farm, Oriole	43.778517	-79.346556	Heart Sushi	43.777203	-79.343805	Japanese Restaurant

The all Other Venues data Frame except restaurants:

:

	Neighborhood	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Hillcrest Village	-79.363452	Eagle's Nest Golf Club	43.805455	-79.364186	Golf Course
1	Hillcrest Village	-79.363452	AY Jackson Pool	43.804515	-79.366138	Pool
3	Hillcrest Village	-79.363452	Duncan Creek Park	43.805539	-79.360695	Dog Run
4	Fairview, Henry Farm, Oriole	-79.346556	The LEGO Store	43.778207	-79.343483	Toy / Game Store
5	Fairview, Henry Farm, Oriole	-79.346556	DAVIDsTEA	43.777593	-79.345089	Tea Room

Also for the population data frame after scraping the data from Wikipedia , it must be cleaned and all the not necessary columns should be removed in my case **Census Tracts , Land area (km2)**, **% Change in Population since 2001**, **% Renters** and **Map** , the other cleaning process is the column FM sould be filtered in order to distinguish the regions I queried only the values FM=="NY" or FM=="Y" or FM=="EY for York Region and set the index as Name of Neighborhood

Now I have cleared data frame as below

3]:

	Name	FM	Population	Density (people/km2)	Average Income	Transit Commuting %	Second most common language (after English) by name	Second most common language (after English) by percentage
8	Bathurst Manor	NY	14945	3187	34169	13.4	Russian (9.5%)	09.5% Russian
40	Don Mills	NY	21372	2377	47515	10.8	Unspecified Chinese (3.9%)	03.9% Unspecified Chinese
44	Downsview	NY	36613	2270	26751	14.4	Italian (11.7%)	11.7% Italian
67	Henry Farm	NY	2790	3066	56395	15.6	Mandarin (3.9%)	03.9% Mandarin
91	Leaside	EY	13876	4938	82670	9.7	Bulgarian (0.4%)	00.4% Bulgarian
168	Willowdale	NY	43144	5618	39895	15.6	Cantonese (7.9%)	07.9% Cantonese
169	Wilson Heights	NY	13732	3317	37978	15.9	Filipino (6.2%)	06.2% Filipino