2 Data cleaning and feature selection

2.1 Data sourse

For this project neighborhoods data was scraped from wikipedia page using BeautifulSoup library (for more information link here).This data includes postal code for each borough.

For exploring neighborhoods in this project was use Foursquare API. Foursquare is a technology company that built a massive dataset of location data. They actually crowd-sourced their data and had people use their app to build their dataset and add venues and complete any missing information they had in their dataset. Currently its location data is the most comprehensive out there, and quite accurate that it powers location data for many popular services like Apple Maps, Uber, Snapchat, Twitter and many others, and is currently being used by over 100,000 developers, and this number is only growing.

2.2 Data cleaning

Original data set have 217 neighborhoods. (Figure 1)

	Postal Code	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park
3	M5A	Downtown Toronto	Harbourfront
4	M6A	North York	Lawrence Manor
...	...	...	...
212	M8Z	Etobicoke	Mimico NW
213	M8Z	Etobicoke	The Queensway West
214	M8Z	Etobicoke	South of Bloor
215	M8Z	Etobicoke	Kingsway Park South West
216	M8Z	Etobicoke	Royal York South West

217 rows × 3 columns

Figure 1. Original Dataset.

Further, get the coordinates for every neighborhood using Nominatim search engine and join this data with original dataset.

However, for some neighborhoods engine not found coordinates, so this row was dropped. After that, clean data includes 198 neighborhoods.(Figure 2, Figure 3)

	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.761124	-79.324059
1	M4A	North York	Victoria Village	43.732658	-79.311189
2	M5A	Downtown Toronto	Regent Park	43.660706	-79.360457
3	M5A	Downtown Toronto	Harbourfront	43.640080	-79.380150
4	M6A	North York	Lawrence Manor	43.722079	-79.437507
...	...	...	...	...	...
193	M8Y	Etobicoke	Royal York South East	43.648183	-79.511296
194	M8Z	Etobicoke	Mimico NW	43.616677	-79.496805
195	M8Z	Etobicoke	The Queensway West	43.623618	-79.514764
196	M8Z	Etobicoke	South of Bloor	43.670489	-79.386465
197	M8Z	Etobicoke	Royal York South West	43.648183	-79.511296

198 rows × 5 columns

Figure 2. Data with coordinates after cleaning.

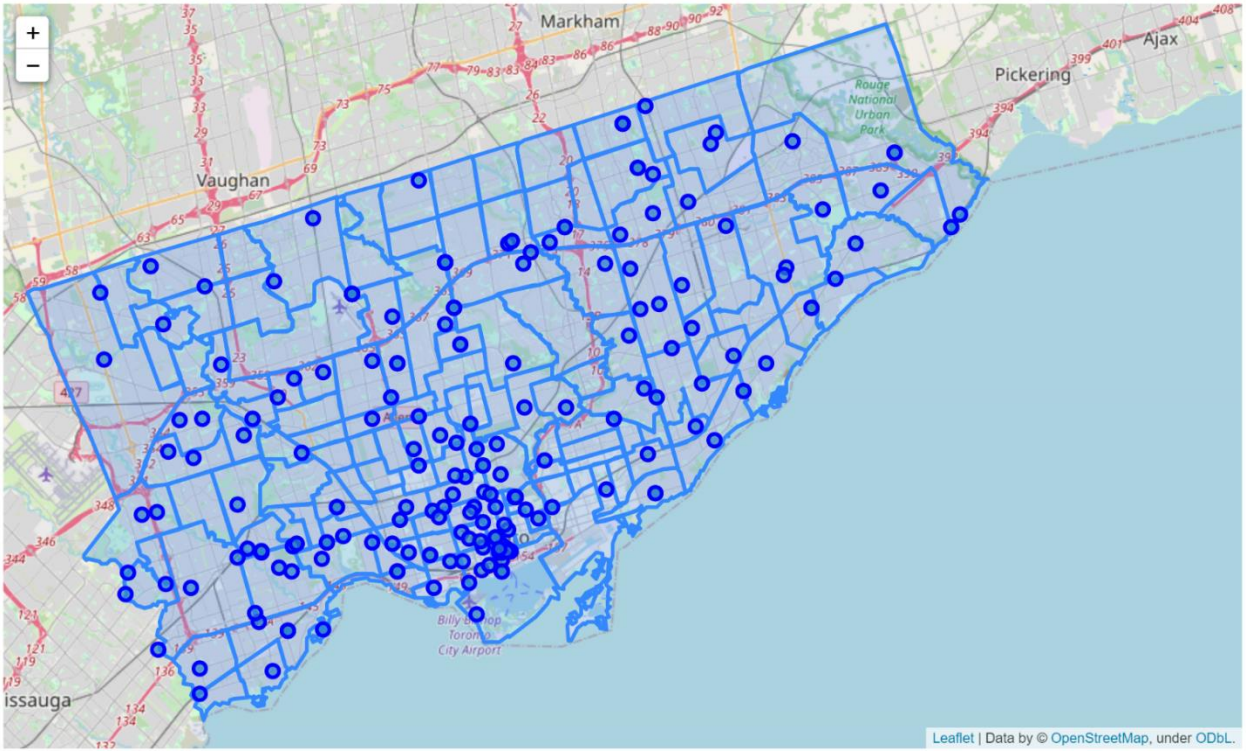

Figure 3. Neighborhoods in Toronto map. Straight line denotes districts borders.

According to map above (Figure 3), we see that some points located in a distance between others. So, for analyze neighborhoods needs calculate maximum distance between closest points, and this distance is about 3 kilometers.