# 2. Data acquisition and cleaning

## 2.1 Data sources

Neighborhood in New York City has a total of 5 boroughs and 306 neighborhoods. this dataset exists for free on the web, here is the link to the dataset: [here](https://geo.nyu.edu/catalog/nyu_2451_34572).

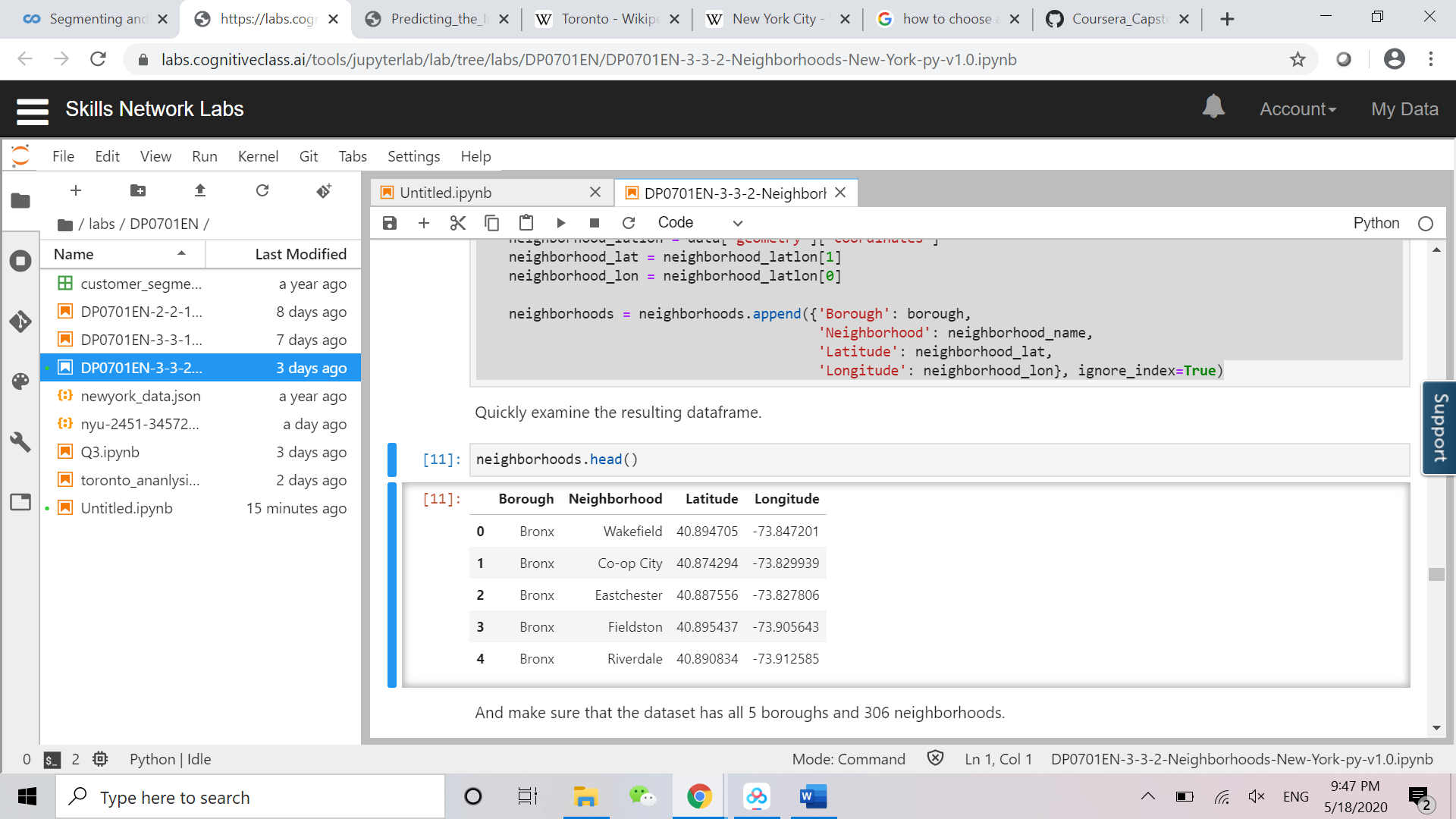
In Toronto ,I use [here](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M,) in order to obtain the data that is in the table of postal codes and [here](http://cocl.us/Geospatial_data)  is a link to a csv file that has the geographical coordinates of each postal code.

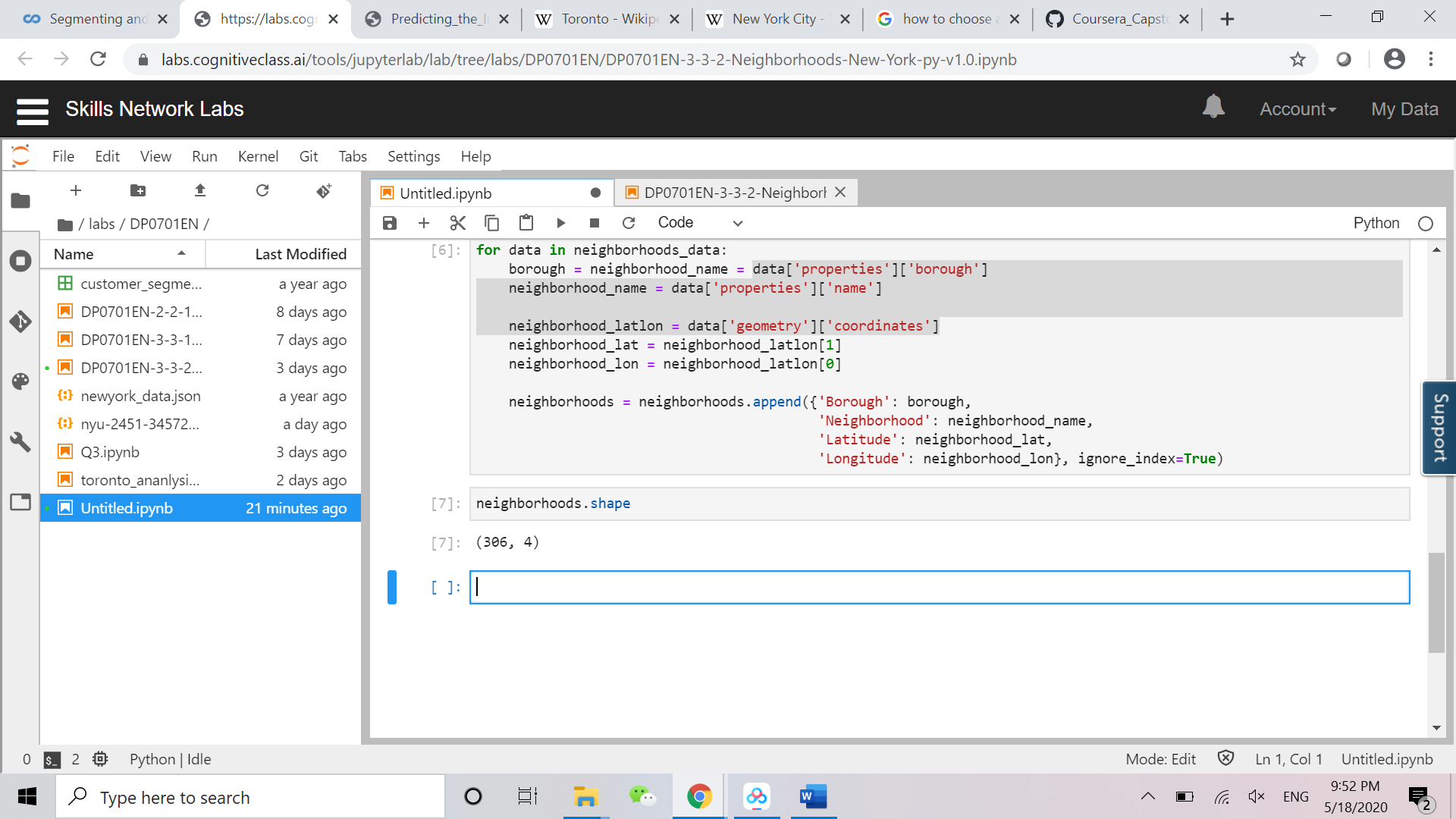
In addition, the document about city size and population density in New York City is [here](https://www.macrotrends.net/cities/23083/new-york-city/population), and in Toronto is [here](https://worldpopulationreview.com/world-cities/toronto-population/).

## 2.2 Data cleaning and Feature selection

Neighborhood in New York city

The data of neighborhood in New York city has a total of 5 boroughs and 306 neighborhoods. In order to segment the neighborhoods and explore them, it is essential to need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. all the data I need is in the features key, which is a list of the neighborhoods, so it is necessary to transform data[‘feature’][‘properties’][‘borough’] , data[‘feature’] [‘properties’][‘name’] and data[‘feature’] ['geometry']['coordinates'] into a pandas data frame. The result is as show below:



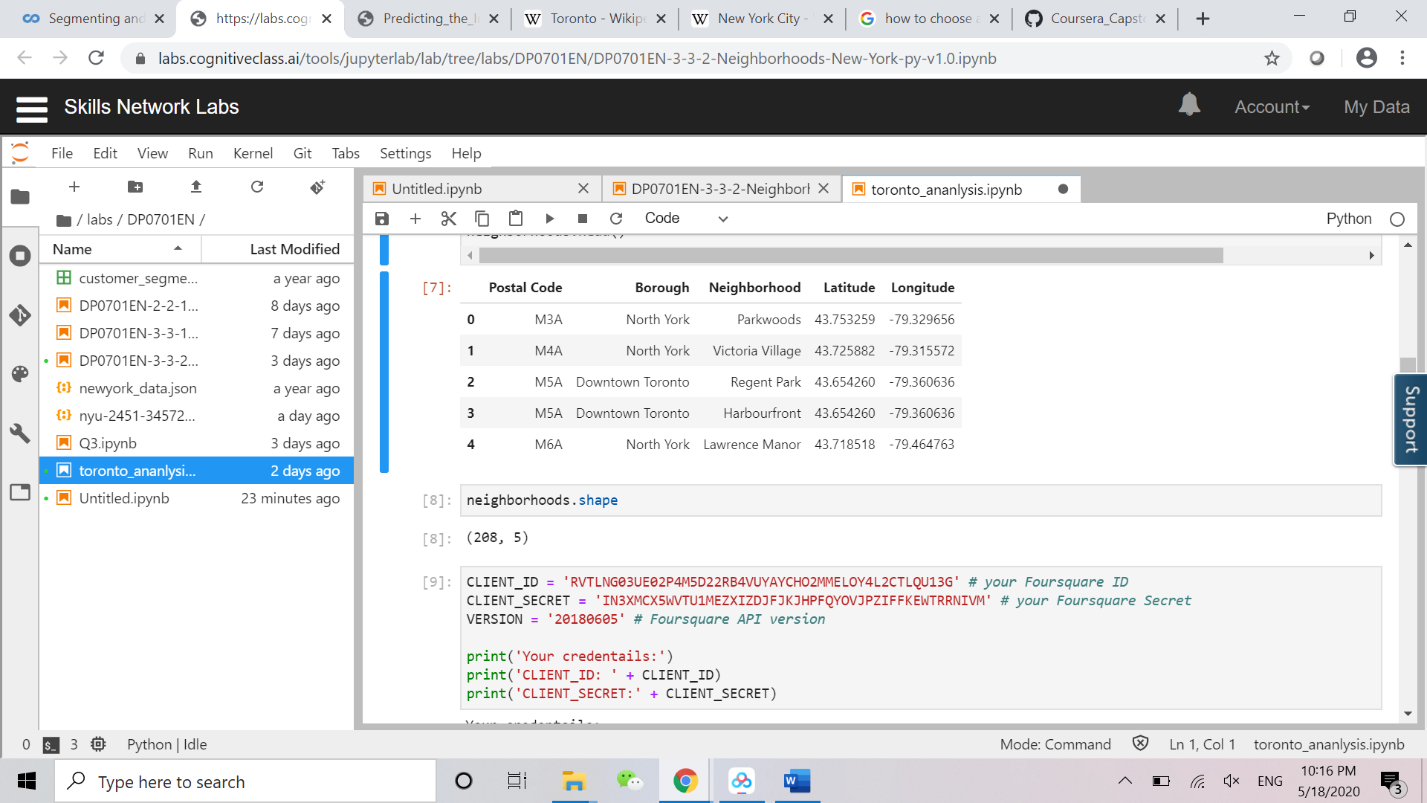


Then, we can use this data frame to explore all venues in New York city by utilizing the Foursquare API.

Neighborhood in Toronto

It is essential to ignore which borough is Not assigned. And if more than one neighborhood can exist in one postal code area (combined into one row with the neighborhoods separated with a comma), I separate them in two rows. What is more, if a cell has a borough but a Not assigned neighborhood, the neighborhood will be the same as the borough.

In order to utilize the Foursquare location data, it is necessary to get the latitude and the longitude coordinates of each neighborhood. There has a csv file that has the geographical coordinates of each postal code and I combined two files together by using the same postal code. The result as below:



Then, we can use this data frame to explore all venues in Toronto by utilizing the Foursquare API.

New York Population

we use 1960-2020 to forecast the future 10 years’ population in New York city.

| **Year** | **Population** | **Growth Rate** |
| --- | --- | --- |
| 2020 | 18,804,000 | -0.01% |
| 2015 | 18,648,000 | 0.31% |
| 2010 | 18,365,000 | 0.31% |
| 2005 | 18,087,000 | 0.31% |
| 2000 | 17,813,000 | 0.86% |
| 1995 | 16,943,000 | 1.04% |
| 1990 | 16,086,000 | 0.48% |
| 1985 | 15,827,000 | 0.29% |
| 1980 | 15,601,000 | -0.22% |
| 1975 | 15,880,000 | -0.39% |
| 1970 | 16,191,000 | 0.94% |
| 1965 | 15,177,000 | 1.39% |
| 1960 | 14,164,000 | 1.40% |

Population of Toronto

we use 1960-2020 to forecast the future 10 years’ population in Toronto.

| **Year** | **Population** | **Growth Rate** |
| --- | --- | --- |
| 2020 | 6196731 | 0.011 |
| 2015 | 5867292 | 0.013 |
| 2010 | 5499233 | 0.0178 |
| 2005 | 5035232 | 0.0179 |
| 2000 | 4607142 | 0.0188 |
| 1995 | 4197157 | 0.0197 |
| 1990 | 3806957 | 0.0256 |
| 1985 | 3355459 | 0.0221 |
| 1980 | 3008032 | 0.0166 |
| 1975 | 2770072 | 0.0179 |
| 1970 | 2534788 | 0.0391 |
| 1965 | 2092902 | 0.0371 |
| 1960 | 1744328 | 0.0503 |