**DATA**

For this project, I worked on getting a list of all the Neighborhoods in the city of Toronto. I utilized the Beautiful Soup web scraper to parse the list of zip codes and neighborhoods from Wikipedia. Once I got the data into a list, I utilized the panda’s library to put that data in a Data Frame.

I then performed data cleaning operations, to filter unnecessary information, fill missing values and complete the data frame.

Here’s how the Data frame looks like:

|  | **PostalCode** | **Borough** | **Neighborhood** |
| --- | --- | --- | --- |
| **2** | M3A | North York | Parkwoods |
| **3** | M4A | North York | Victoria Village |
| **4** | M5A | Downtown Toronto | Harbourfront |
| **5** | M5A | Downtown Toronto | Regent Park |
| **6** | M6A | North York | Lawrence Heights |

Since FourSquare requires the latitude and longitude, I used a CSV file provided by cocl.us which contains the location coordinates for all the postal codes.

|  | **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 |

Now we have two sets of data frames one with location data and one with neighborhood information. I merged both the data frames to have a single unified list.

|  | **PostalCode** | **Borough** | **Neighborhood** | **Latitude** | **Longitude** |
| --- | --- | --- | --- | --- | --- |
| **0** | M1B | Scarborough | Rouge, Malvern | 43.806686 | -79.194353 |
| **1** | M1C | Scarborough | Highland Creek, Rouge Hill, Port Union | 43.784535 | -79.160497 |
| **2** | M1E | Scarborough | Guildwood, Morningside, West Hill | 43.763573 | -79.188711 |
| **3** | M1G | Scarborough | Woburn | 43.770992 | -79.216917 |
| **4** | M1H | Scarborough | Cedarbrae | 43.773136 | -79.239476 |

At this stage, since I want to focus only on the Borough’s in Toronto, I filter the data. My final data frame looks as follows:

|  | **PostalCode** | **Borough** | **Neighborhood** | **Latitude** | **Longitude** |
| --- | --- | --- | --- | --- | --- |
| **37** | M4E | East Toronto | The Beaches | 43.676357 | -79.293031 |
| **41** | M4K | East Toronto | The Danforth West, Riverdale | 43.679557 | -79.352188 |
| **42** | M4L | East Toronto | The Beaches West, India Bazaar | 43.668999 | -79.315572 |
| **43** | M4M | East Toronto | Studio District | 43.659526 | -79.340923 |
| **44** | M4N | Central Toronto | Lawrence Park | 43.728020 | -79.388790 |

Now I have a list of all the Neighborhood’s with a Borough in Toronto. This is my dataset for the analysis. Using Folium a visualization library, I can plot all the Neighborhoods on a Map and prep the data for clustering.

Using FourSquare API’s I will perform exploratory analysis to identify the features that are most suitable for an apartment complex and then work on clustering and segmentation to identify potential zones.