## ****Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods****

**Introduction/Business Problem**

**I have been approached by a group of professional couples (Pharmacists and Engineers) who intent to emigrate to the United States or Canada from Nigeria. They intend to live in Brooklyn, New York or Toronto, Canada.**

**The Electrical Engineer believes he will be comfortable finding a job in any of the locations but the pharmacists requested that I compare between the two locations and their neighborhood in order to find a suitable location for the pharmacy outlet while maintaining a relatively calm family life.**

**Our terms of reference are as follows:**

1. **To compare between Toronto and Brooklyn New York City in order to decide on what city is preferred for an emigrant from Nigeria in terms of cultural diversity and tolerance, proximity to relaxation and recreational centers, and relatively serene neighborhoods.**
2. **After choosing between Toronto and New York, I will then decide on best neighborhood to situate a pharmacy outlet using data analysis. My choice of neighborhoods to situate a pharmacy outlet will be based on neighborhoods with already high “most common venues visited” as pharmacy as that will mean that there is room for competition in the market share in such neighborhoods**

**Data section**

**To achieve the above business problem, I will be using data from the following links:**

1. <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

**This Wikipedia site contains an updated neighborhoods and boroughs in the Toronto metropolis. It is in an unstructured data format and hence I used BeautifulSoup to extract it**

1. <http://cocl.us/Geospatial_data>

**This is an already extracted gps coordinates of neighborhood in Toronto**

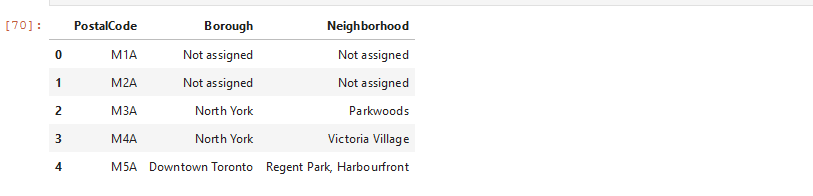
1. **Foursquare (www.foursquare.com): This is a technology company that has a large dataset of accurate location data. It is widely used by top companies such as Apple, Uber, Twitter, and Snapchat. Their API is currently being used by over 100,000 developers**
2. **newyork\_data.json from** <https://cocl.us/new_york_dataset>**: This data set gives us the new York neighborhood inclusive of their gps coordinates**

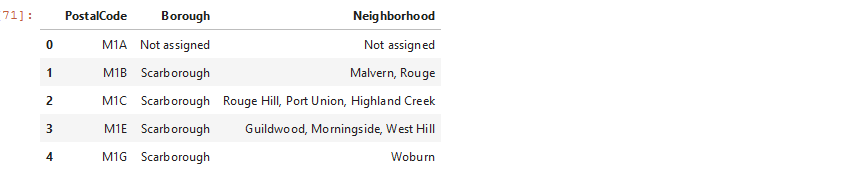
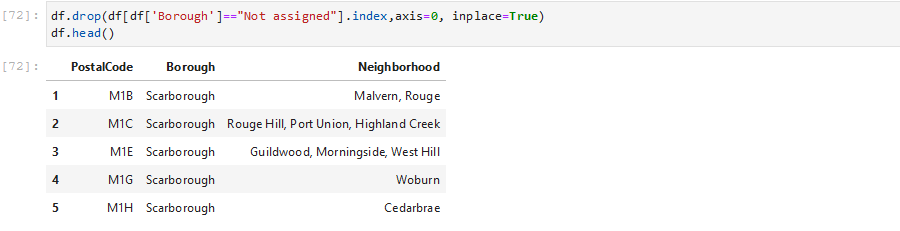
**Methodology**

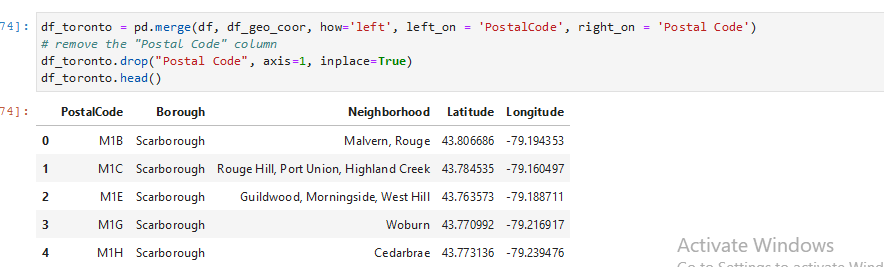
**Data Preparation for Toronto Data**

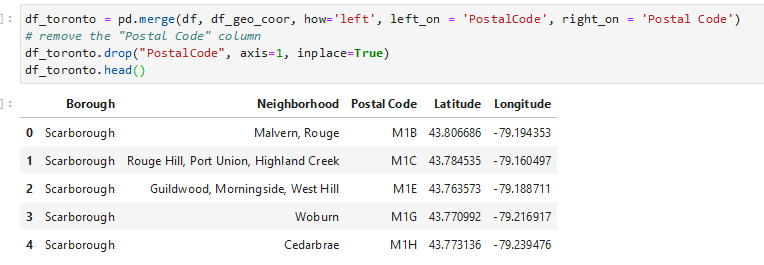
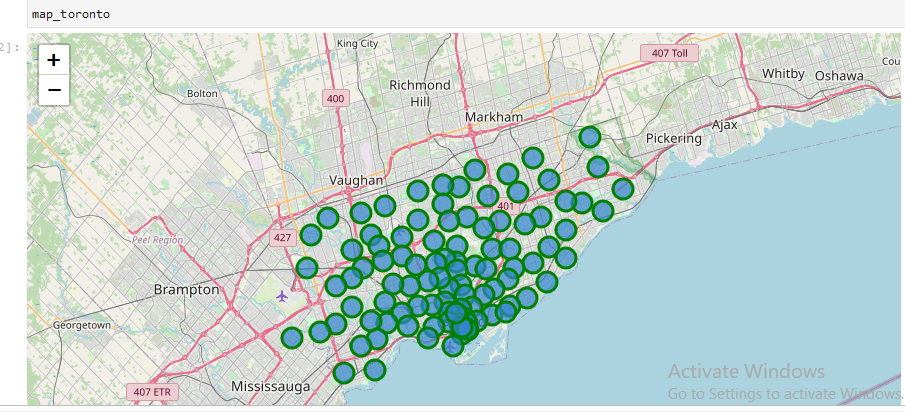
**At this stage, libraries useful for the data manipulation was imported. This includes: ProgressBar, BeautifulSoup, numpy, pandas, matplotlib, requests, Nominatim, matplotlib, KMeans # import k-means, folium and lxml.**

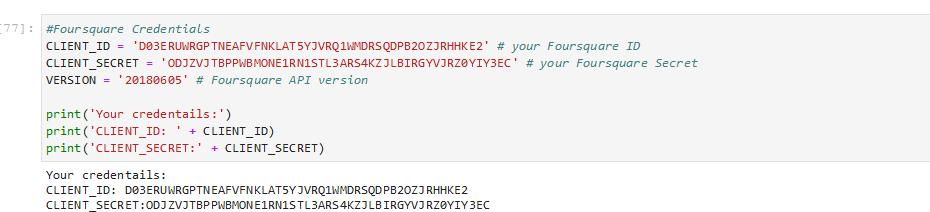
1. **We begin manipulation with our Toronto and New York data by converting it to dataframe.**

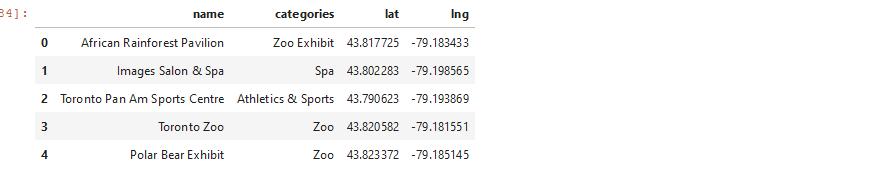


1. **We then arranged the data by postal code** 
2. **The table shows that there are missing values “not assigned” hence the entire rows were removed**
3. **We now include gps coordinates from the table** <http://cocl.us/Geospatial_data>



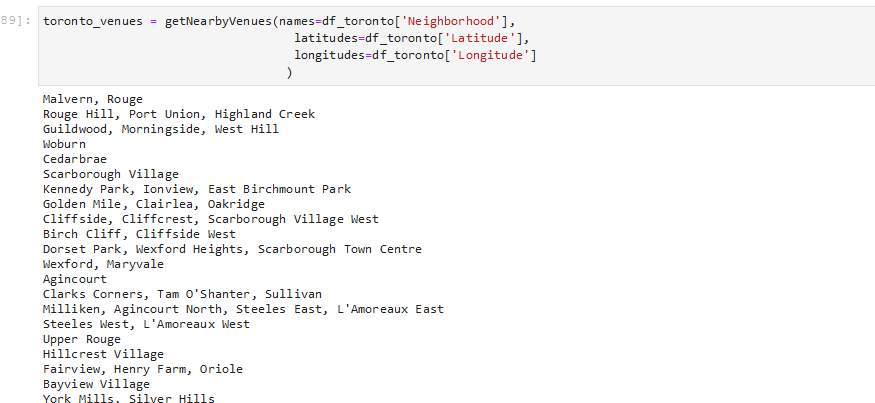
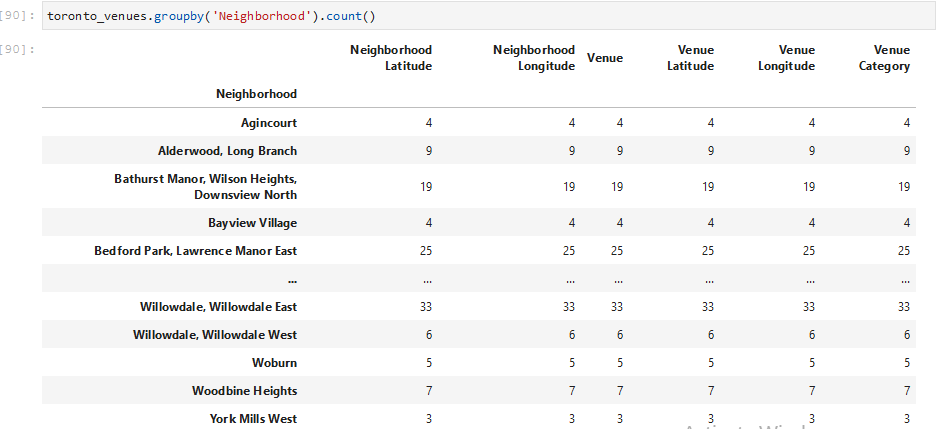
1. **At this stage, the gps coordinate was merged with the dataframe and postal code column was dropped** 
2. Nominatim which is a geolocator was used to attached coordinates to locations on a grapgh and then graph was plotted using folium 
3. Foursquare is used at this stage for analysis

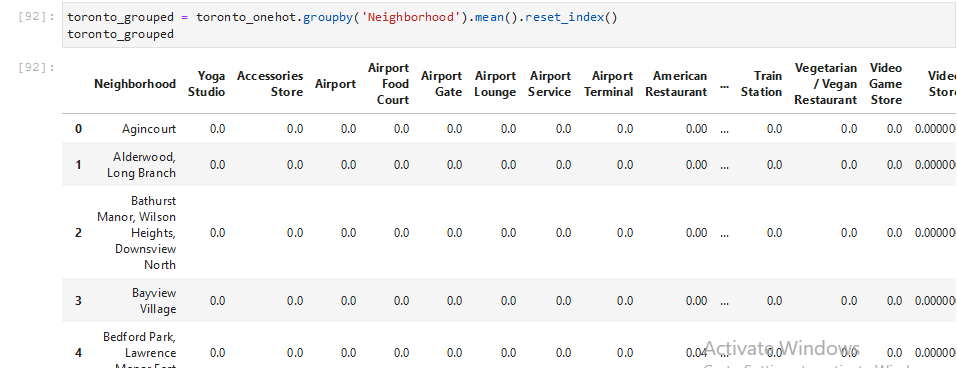


1. We download all venues at a radius of 2000 around Malvern, Rouge and put it in a dataframe for manipulation. 

**Data Modelling for Toronto**

**At this stage, we want to give meaningful insight to data by manipulation and analysis. I want to reveal patterns and structure within the data and hence get some insights from it.**

1. **Get list of venues around Toronto** 
2. **Group by neighborhoods in Toronto** 
3. **Analyse each neighborhoods in Toronto to get the frequency of visits to the various location.**



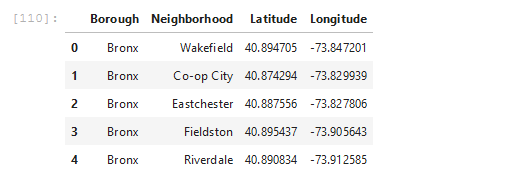
1. **Rank the venues according to how busy they are.** 

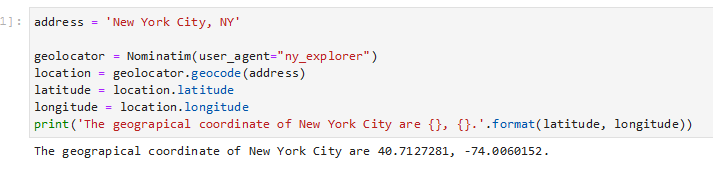
**Results for Toronto**

**Neighborhoods having venues with very busy Pharmacy outlets can be recommended as good location for situating a pharmacy outlet as that means a large market share.** 

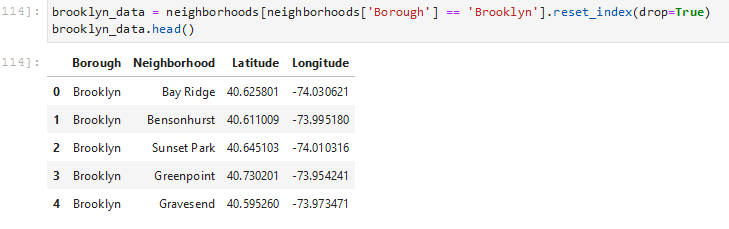
**NEW YORK CITY**

**Following similar data analysis steps for Brooklyn, Newyork**

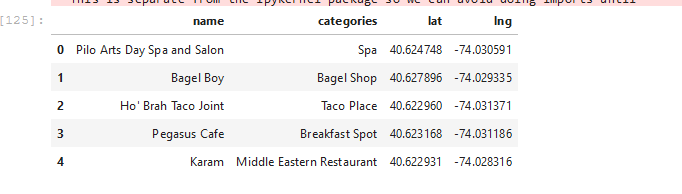
1. **Tranform the data into a pandas dataframe** 
2. Nominatim which is a geolocator was used to attached coordinates to locations on a grapgh and then graph was plotted using folium

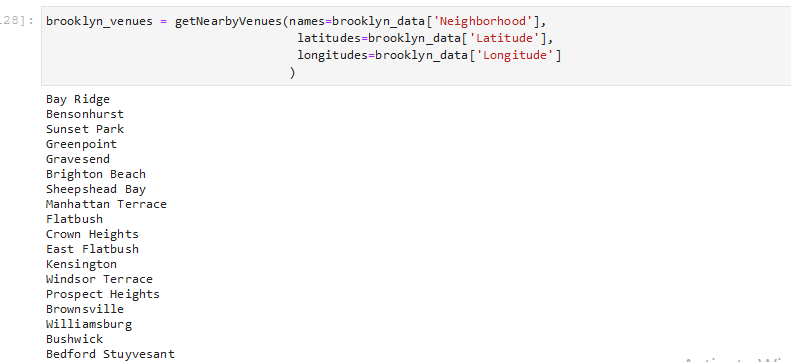
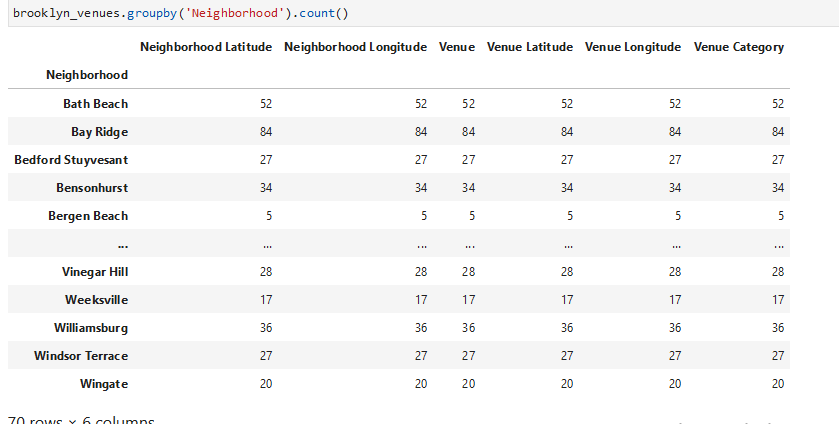
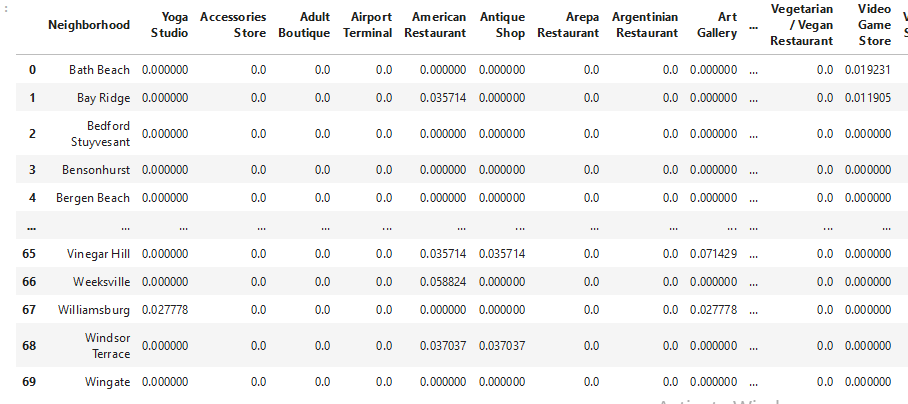


1. Let's slice the original dataframe and create a new dataframe of the Brooklyn data.



1. **Using Foursquare to get venues around the neighborhood** 
2. Get the Categories of the Venue

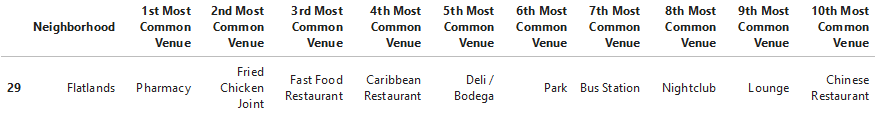


1. Explore Neighborhoods in Brooklyn 
2. Analyse the frequency of patronage of various venues 
3. Analyze each Brooklyn Neighborhood to show the frequency of use 
4. **Get top ten most commonly used venues**



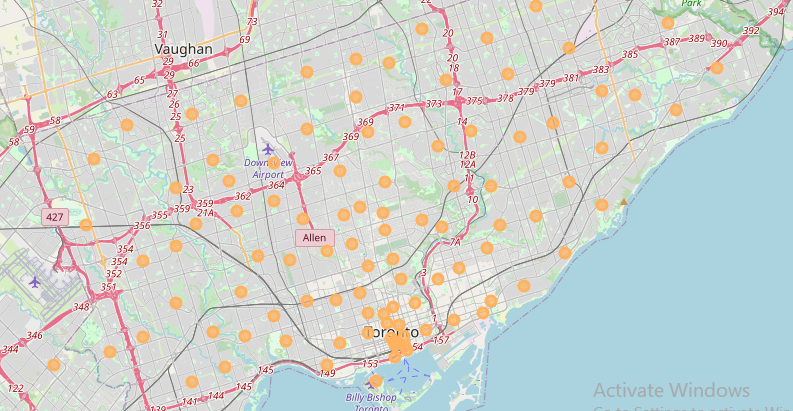
**Results for Brooklyn New York**

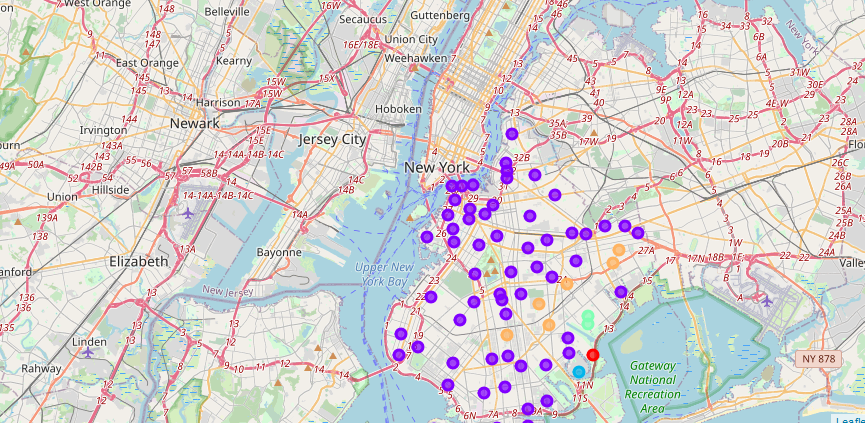
**Neighborhoods having venues with very busy Pharmacy outlets can be recommended as good location for situating a pharmacy outlet as that means a large market share**



**EVALUATION**

**Comparing between Toronto and New York using K Clusters**





**Discussion**

**From the above analysis, results and evaluation, I will make the following recommendation to my client:**

1. **Between Newyork and Toronto, with respect to serenity of the neighborhoods and relative calm, I will suggest they emigrate to Toronto Canada**
2. **Within the Toronto neighborhood, it will be better to situate a pharmacy in an already where there is high frequency to Pharmacy outlets as that signifies a large market share and good competition**



**Conclusion**

**We can safely conclude that with the help of data science, strategic and business decisions can be more accurate.**