Capstone Project Final Report - Karsten Kraak

**Clustering of Boroughs in Toronto and New York City to Identify Similar Boroughs**

*21 October 2019*

## Introduction to Problem and Discussion of Background

As a person with experience moving between different countries and cities, I have found that there are some cities that I really enjoyed living in and some cities that I really did not like. This depended on the borough and venues around me. The following project will focus on a hypothetical scenario in which I am planning on relocating to New York City from Toronto. Every year thousands of people move to NYC, and therefore I am planning on building a tool in which a person can identify similar boroughs in Toronto and NYC based on the surrounding venues. The tool will be using Foursquare API venue data and use clustering techniques to highlight similar areas. These similar areas will then be displayed on a map of NYC and Toronto to assist the user with their decision. Please see the ‘Description of Data’ section for a more detailed explanation of the data that will be used.

## Description of Data

The location data will be collected from 2 separate data sources. The Toronto data will be scraped from the following webpage: <https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>

This data will then be combined with the Location Data for NYC, which will be taken from the following:

<https://cocl.us/new_york_dataset>

Next the Foursquare API will be used to explore 1000 venues within a 1000-metre radius of the centre of each borough, considering the scope of this project is focusing on boroughs it was deemed more appropriate to use a larger radius and more venues than the examples seen previously on the course. The top 10 most common venues in each borough will then be used to assign clusters to each borough using K-means clustering.

## Methodology

The data for the boroughs in Toronto was scraped from the Wikipedia webpage as mentioned above. The code used the “Beautiful Soup” scraping method and then a data frame was created to store only borough name data within the online table. Once this had been created all rows where the borough was listed as “not assigned” were removed and only unique boroughs were selected. Next the code used the geocoder to obtain the latitude and longitude of each borough and placed them within the data frame. Another column for “City” was then added to the data frame to be able to identify which borough belonged to which city when the New York and Toronto data frames were combined. Finally, the boroughs that didn’t contain the word “Toronto” were dropped, this was due to the Foursquare API being unable to source any information for these other boroughs.

The data for the boroughs in New York City was taken using the json data from the example previously displayed on the course. A ‘for loop’ was used to extract the borough names and the data was then added into a data frame. Once again, a column for “City” was added to identify the city that each borough belonged to and only the unique borough names were saved into the data frame. The geocoder method was used to find the latitude and longitude values for each borough to create the finalised New York data frame.

Next the data frames for Toronto and New York were concatenated to produce a combined data frame. This data frame and the Foursquare API were then used to find the nearest 1000 venues in a 1000-metre radius of each borough location, and a data frame was then produced to reflect each venue returned by the query, its location and its category/venue type. Next, the ‘one-hot’ encoding method was used to summarise the venue results, and the mean of these values was taken. The top 10 venues in each borough was then listed by sorting the venue categories by their mean values. These top 10 venues were then inserted into a combined data frame to allow for a comparison between most common venues in each borough.

K-means clustering was then used to cluster all the boroughs based on the 10 most common venues using the data frame mentioned above. 5 clusters were created as any more clusters would over-specify the data and therefore there would be no matches, any less clusters would mean under-specifying and thereby there would be too many matches between boroughs that aren’t similar enough.

A data frame was produced to show the clusters and their most common venues for a side-by-side analysis in a tabular format. Finally, the application will produce two folium maps that will display similar boroughs in NYC and Toronto based on the results of the clustering.

## Results and Discussion

Figure 1 is the Folium map for the clusters within New York City. An initial observation from the clustering shows that four of the boroughs (Staten Island, Manhattan, Brooklyn and Queens are categorised as Cluster 1 and one cluster (Bronx) has been classified to be Cluster 4. By examining the map, it can be seen that the four boroughs in Cluster 1 seem to all be located more centrally within the city than the Bronx, and this could therefore explain why this borough has been clustered separately.

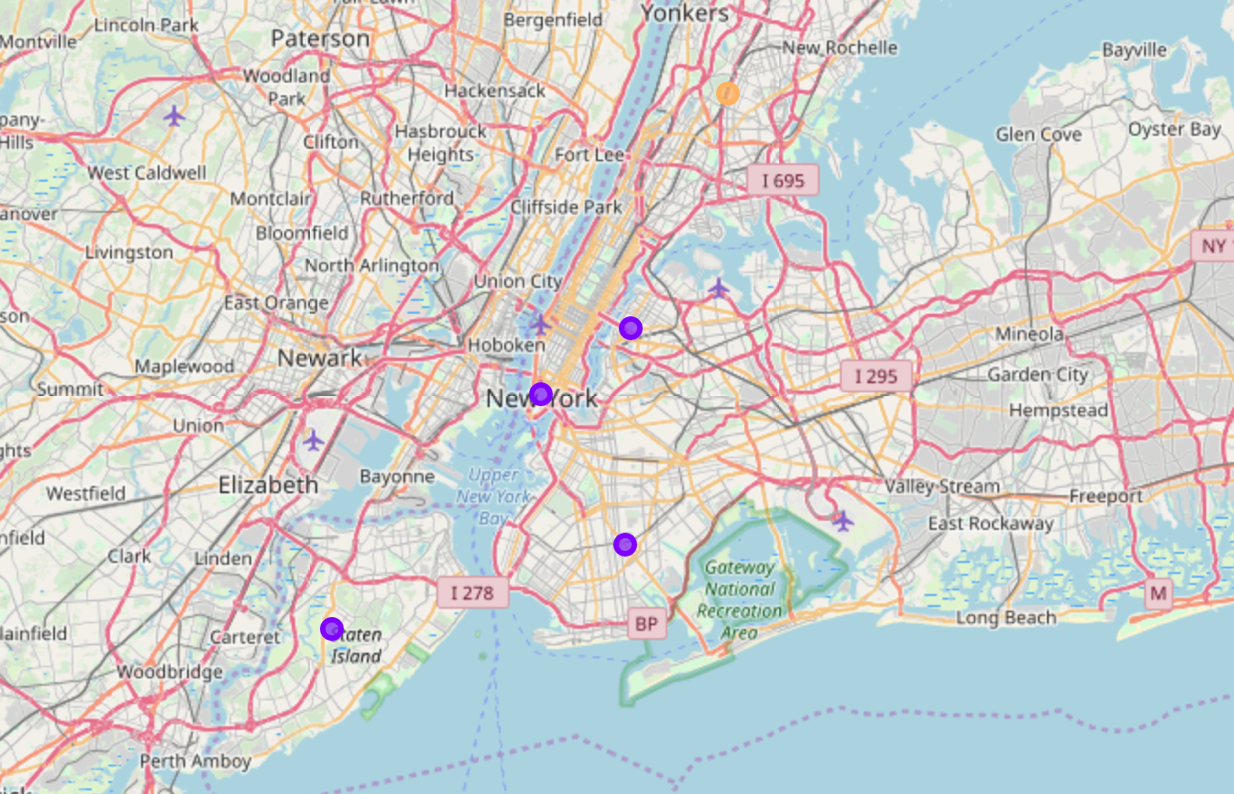


Figure 1: New York Folium Map with Clusters

Figure 2 below is the Folium map for the clusters within Toronto. Initial observations show that each one of the boroughs in Toronto has been clustered separately. When examining the Folium map, it can be seen that all of these boroughs are spread far apart, and when looking at the terrain it can also be seen that the terrain varies greatly between each borough. This would therefore likely have a large impact on the categories of venue present in each venue and would lead to the boroughs being clustered separately.

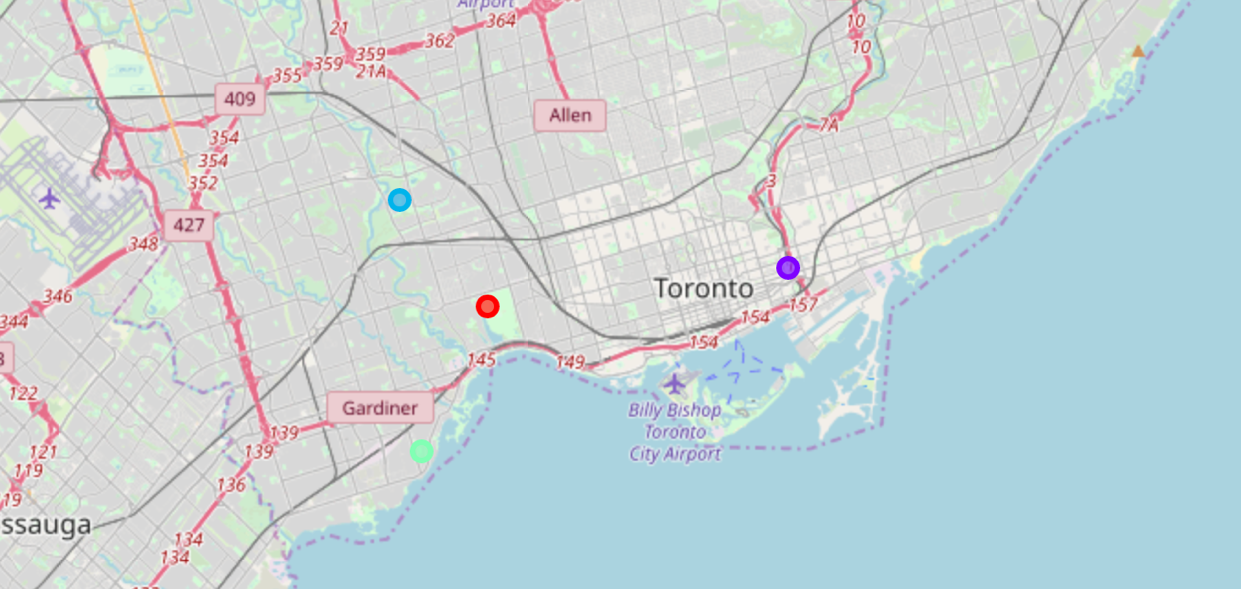


Figure 2: Toronto Folium Map with Clusters

When examining both Figure 1 and 2 simultaneously, it can immediately be seen that only one of the boroughs in Toronto has been categorised to be in the same cluster (Cluster 1) as the boroughs in New York City – this borough was Downtown Toronto. Upon an examination of the location and terrain of Downtown Toronto it could be seen that it has similarities to the other boroughs in Cluster 1; it is located very centrally within the city, and it appears to be in a more populated area of Toronto. An analysis of the most common venues in each borough as presented in Figure 3 could be conducted to examine if the types of venues are similar for the Cluster 1 boroughs.



Figure 3: Combined Data Frame Showing Each Borough and Respective Cluster

An examination of Figure 3 indicates that there are some very interesting insights. Firstly, looking at the boroughs within Cluster 1, it clear why each of these boroughs has been clustered together. A comparison of Manhattan to Downtown Toronto show some similarities as there are 2 direct matches between the top 10 common venues, however many of the other categories are similar, for example in Toronto ‘Pizza place’ is the 3rd most common venue and in Manhattan the 2nd most common venue is ‘Italian Restaurants.’ A further examination of Downtown Toronto show that there are many more similarities between the most common venues in the remaining boroughs in Cluster 1. The next step is to analyse Figure 3 to determine why the remaining boroughs in New York and Toronto have been clustered separately:

* Comparing the borough in Cluster 0 (West Toronto) to the remaining boroughs revealed that the most ‘most common venues’ with other boroughs was 4 direct matches, and this was with Downtown Toronto. This therefore does not provide a valuable insight as to why these two boroughs have not been clustered together.
* Comparing the borough in Cluster 2 (East Toronto) to the remaining boroughs revealed that the most ‘most common venues’ with other boroughs was 4 direct matches with Central Toronto. Once again this does not provide a valuable insight as it would be expected that these would be clustered together considering many of the other venues in the list are similar.
* Comparing the borough in Cluster 3 (Central Toronto) to the remaining boroughs revealed that the most ‘most common venues’ with other boroughs was 4 direct matches with East Toronto, as already mentioned above.
* Comparing the borough in Cluster 4 (Bronx) to the remaining boroughs revealed that the most ‘most common venues’ with other boroughs was 2 direct matches with Downtown Toronto, and therefore it is clearer in this case why it would be categorised to be in another cluster.

## Conclusion

A critical evaluation of the results would conclude that analysis of venue data would not be appropriate for comparing and clustering boroughs in separate cities, as only Downtown Toronto was placed in the same cluster as some of the boroughs in New York City.

This is perhaps due to the size of each borough being too large, and therefore each borough is likely to have thousands of varying venues. It would be more valuable if a comparison of each neighbourhood were to be produced instead as this would allow for a more in-depth analysis and would show more accurate matches between matching neighbourhoods in clusters. This could also show more variation between the clustering in New York City, and would allow for more localised clusters and a more in-depth analysis of these.