An intercomparison of two cities: Toronto, Canada and London, United Kingdom

Benjamin Allen

March 27, 2021

**1. Introduction**

“To what extent are the cities of Toronto and London similar to each other?” is the proposed question for the Data Science Capstone project. Both Toronto, Canada, and London, United Kingdom, are the capital cities of their respective countries, comprising populations well into the millions. Whilst both cities may be classed as capital cities, the difference in history and continental divide could present stark differences in composition with respect to small businesses, and the presence of large international corporations. Furthermore, due to the difference in geographic make-up with respect to street layout, it would be interesting to see whether there are significant differences in this area.

Therefore, I propose to leverage Foursquare data to map and intercompare the composition of both respective cities in terms of business type.

**2. Data**

It was my intent to scrape the postcode locations from the respective cities Wikipedia entry page for London [1], and Toronto [2]. As Wikipedia provides tabulated entries of postcodes for both cities, it makes sense to utilise techniques obtained earlier in the course to leverage the outcome we seek.

With regards to co-ordinate data for the post-codes, for Toronto, the co-ordinate data was obtained from the Google Maps Geocoding API. However, for our London dataset, conveniently; British Ordnance-Survey (OSBNG) National grid co-ordinates [3] were provided for each borough. For those unfamiliar – OSBNG is a system of geographic grid references that divides up the British Isles into 100km squares, which can then be divided up further for greater accuracy. This will be utilised to convert to latitude/longitude.

Once processed and wrangled, I intend to cross-reference the locations of businesses in both cities with Foursquare data. Lastly, I will then cluster the results in order to determine composition.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighbourhood : Name of the Neighbourhood
2. Neighbourhood Latitude : Latitude of the Neighbourhood
3. Neighbourhood Longitude : Longitude of the Neighbourhood
4. Venue : Name of the Venue
5. Venue Latitude : Latitude of Venue
6. Venue Longitude : Longitude of Venue
7. Venue Category : Category of Venue

Based on all the information collected for both London and Toronto, we have sufficient data to build our model. We cluster the neighbourhoods together based on similar venue categories. We then present our observations and findings. Using this data, our stakeholders can take the necessary decision.

**3. Methodology**

We will be creating our model utilizing the Python programming language, with the following packages:

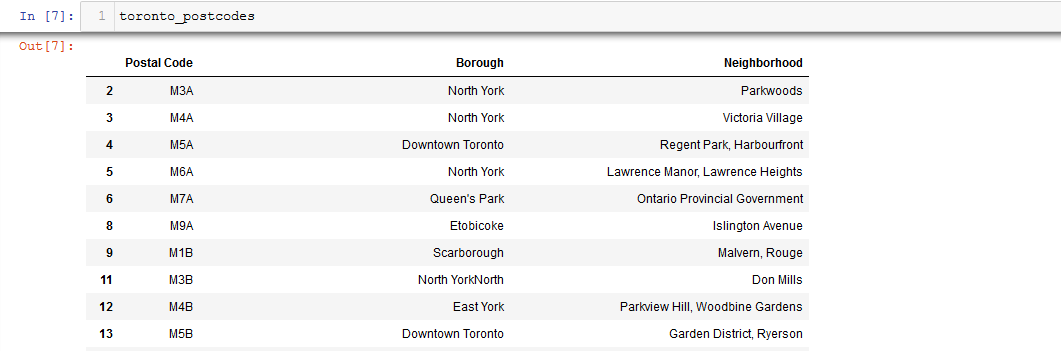
I. Pandas: To collect and manipulate data in JSON and HTMl and then data analysis  
II. Requests: Handling http requests  
III. NumPy – Handling array computation  
IV. Matrixplot Library: Detailing our generated maps  
V. Folium: Generating maps of London and Toronto  
VI. Sklearn: To import the KMeans machine learning model.

The chosen approach is to explore each of the cities individually, plot neighbourhoods being considered on a general map to showcase the neighbourhoods being considered and then construct our model by neighbourhood clustering. Finally, plot the new map with the clustered neighbourhoods, which will allow us to draw insights and compare our findings between the two cities.

**3.1 Data acquisition**

**Toronto**

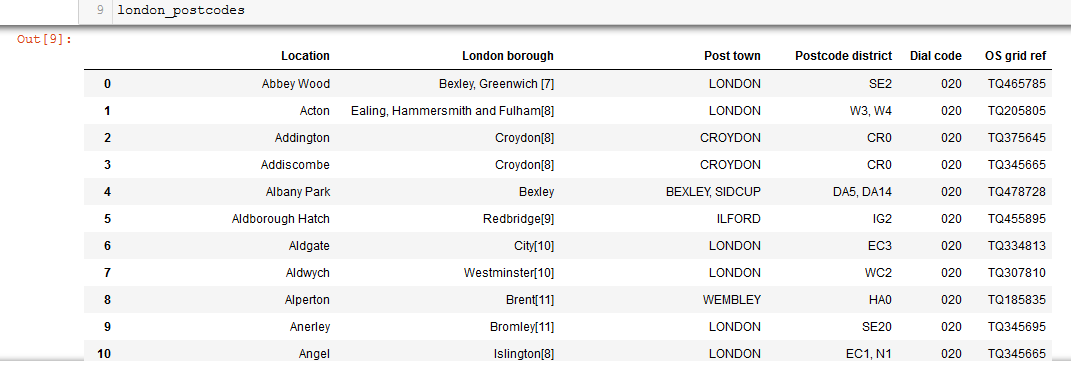
Grabbing our data from the Wikipedia page on Tornonto’s postcodes, and parsing the HTML code resulted in the following table:



*Fig 1. Toronto Postcodes parsed successfully into a dataframe.*

**London**

Very similar as we see for Toronto; we were able to grab the postcode data from the Wikipedia entry. Here is the tabulated dataframe:

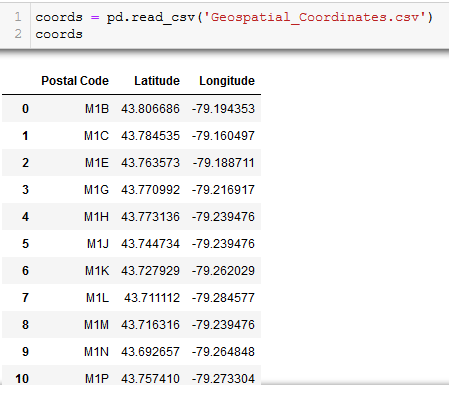


*Fig 2. The dataframe of London neighbourhoods, with the accompanying OS Grid reference which we will utilise to convert to lat/long.*

**3.2 Handling Geographic Coordinates**

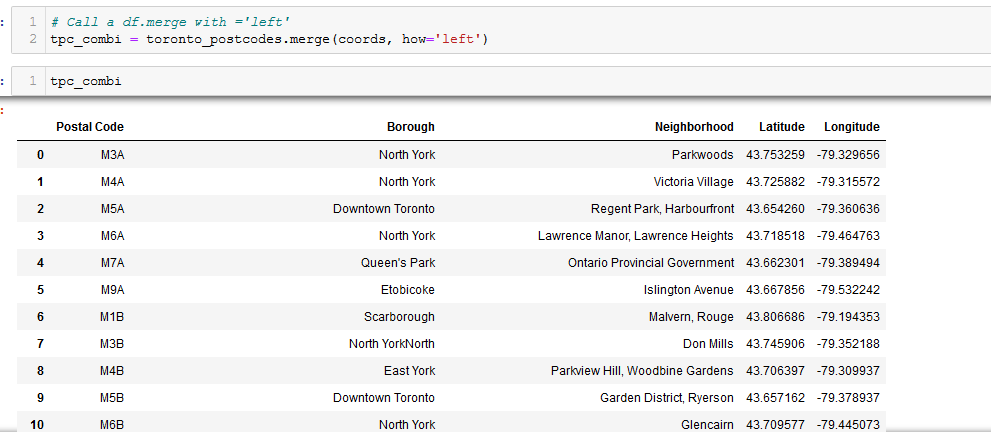
**Toronto**

The latitude and longitude of Toronto neighborhood data can be accessed via the web, or by parsing through the coordinate data held within a CSV file obtained from earlier on in the course. I chose the latter.



*Fig 3. Toronto borough co-ordinate data parsed through from the CSV.*

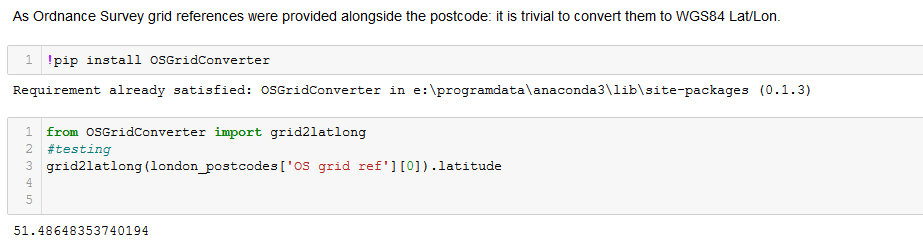
And then merging was trivial:



*Fig 4. Our combined dataframe encompassing the Lat/long for each postcode.*

**London**

To facilitate the conversion OSBNG references to lat/long, it is necessary to install a third-party package, otherwise, the user would be required to derive the calculation mathematically.



*Fig 5. Installation of OSGridConverter and testing on our dataframe.*

Iterating over the dataframe and appending the new lat/lon columns gives us our dataframe ready for clustering.

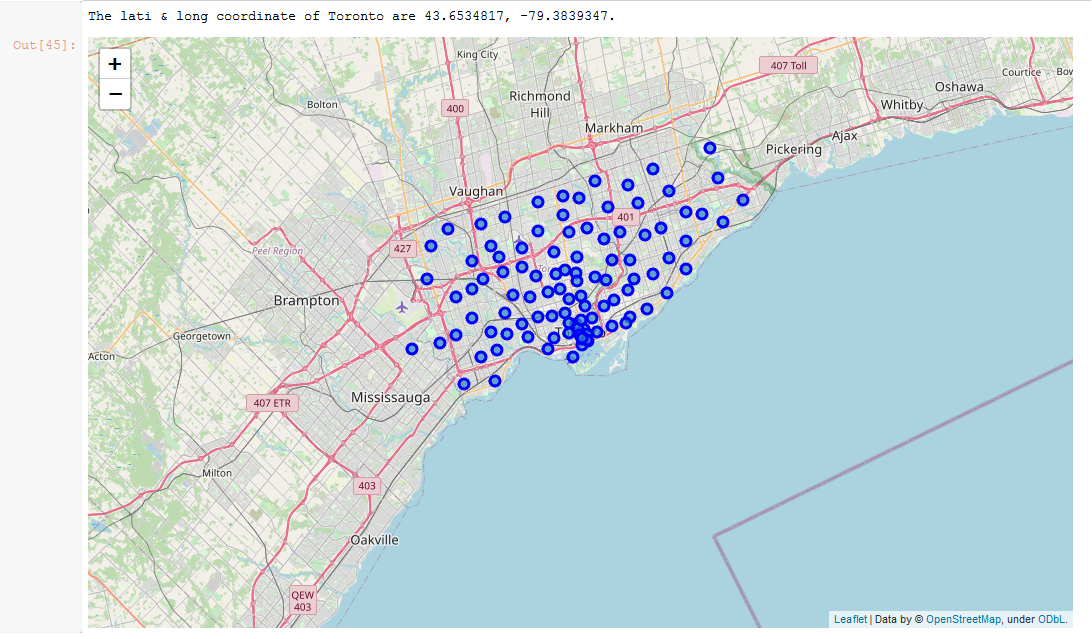


*Fig 6. Our dataframe containing the Lat/lon values alongside their respective grid references and accompanying postcodes.*

**3.3 Foursquare API calling**

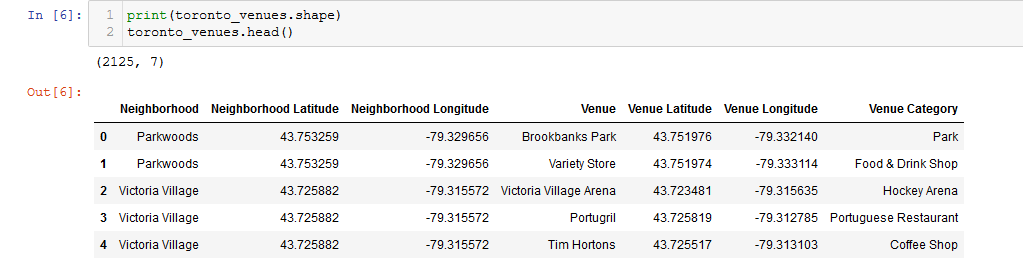
**Toronto**

We first wanted to display the neighbourhood locations using folium and a package called Geopy’s Geolocator, superimposed over a Google Maps image of the city.



*Fig 7. The Folium result – Toronto neighbourhoods superimposed.*

We then called the foursquare API, which resulted in 2125 successful calls for our neighbourhoods, returning venue data that were then appended to the dataframe.



*Fig 8. The dataframe containing venue information for each of our neighbourhoods.*

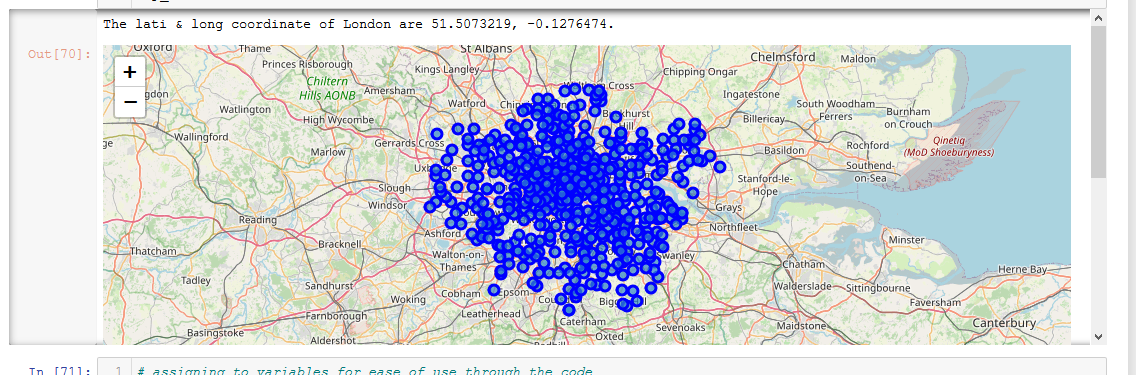
Which, after further cleaning and encoding, we were able to derive the most common venues for each neighborhood.



*Fig 9. Dataframe containing the most frequent Toronto neighborhood venues types.*

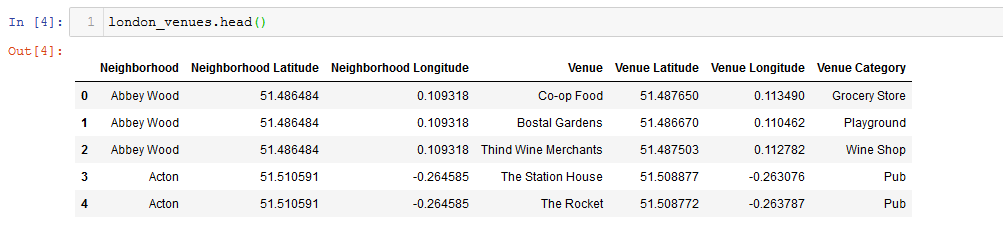
**London**

Similar for Toronto; we overplotted the neigbborhood information scraped from the Wikipedia article:



*Fig 10. Neighborhoods overplotted. All residing withinside the M25 circular motorway as to be expected.*

With the lat/long data now added to our dataframe:



*Fig 11. London neighbourhood data with lat/lon included.*

And lastly, the most common venues for London as completed for Toronto venues:

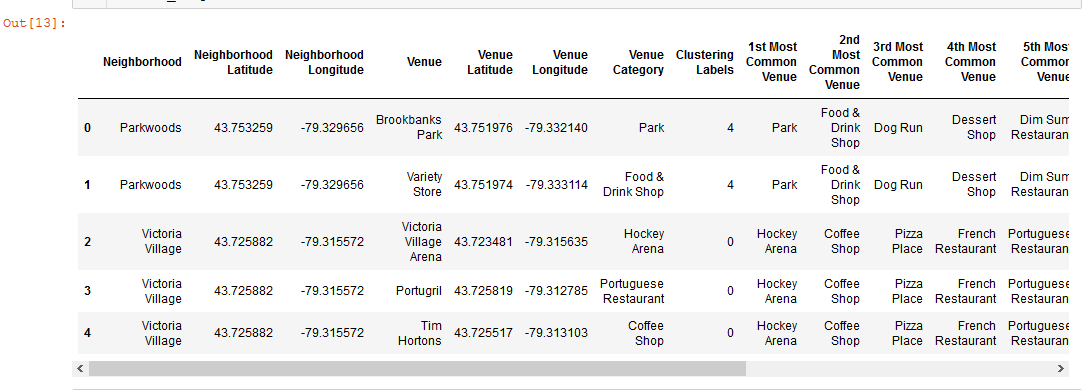


*Fig 12. Most common venues for the top 5 results of our dataframe.*

**3.4 K-means Clustering**

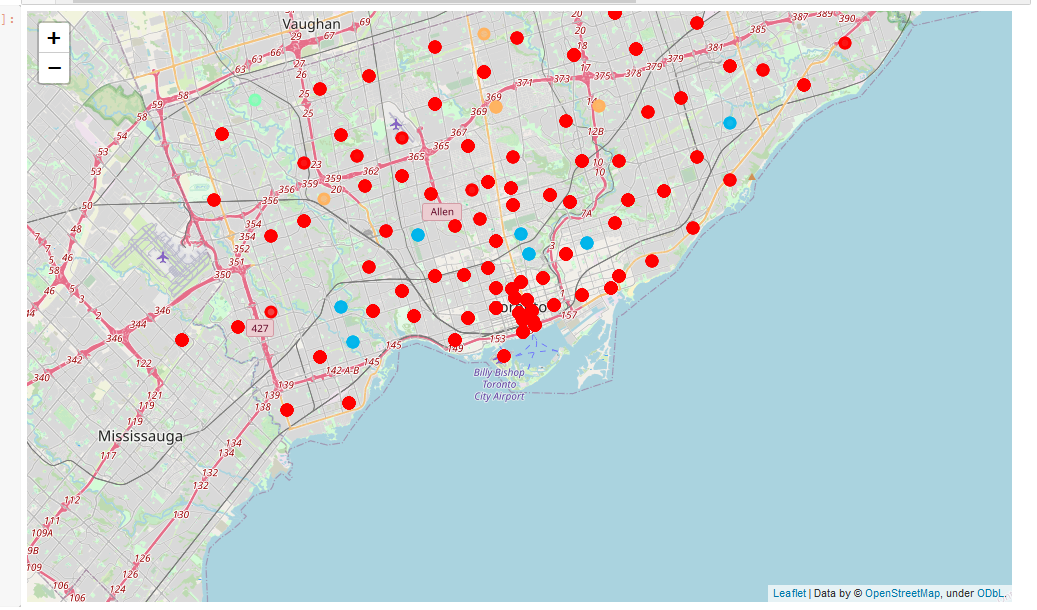
**Toronto**

Merging our dataframes:



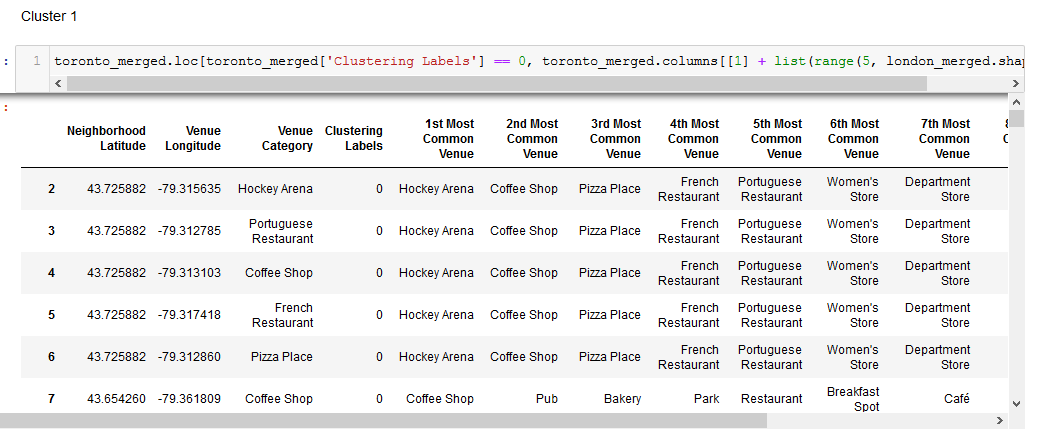
*Fig. 13 Merging co-ordinate data with our clustering labels dataframe.*

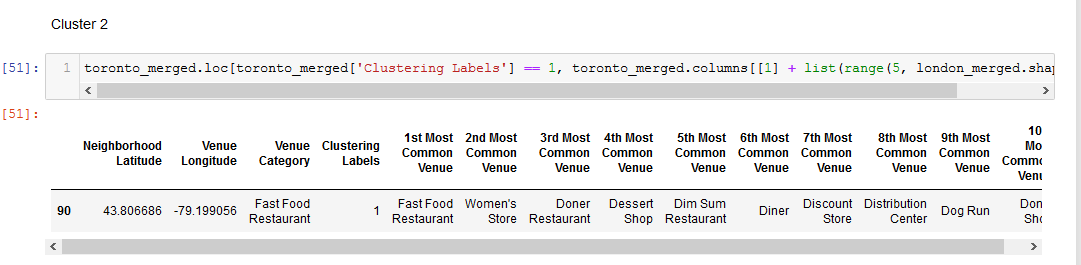
Showcasing our differing neighborhoods on a folium map.

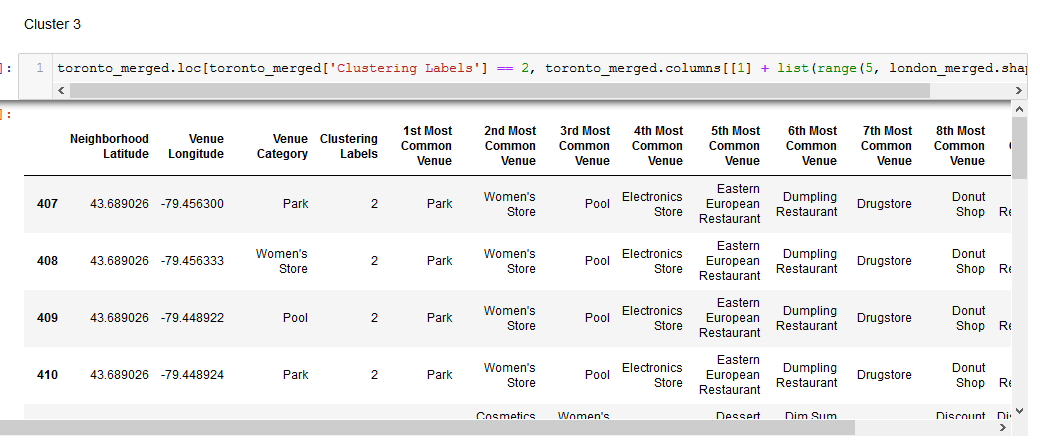


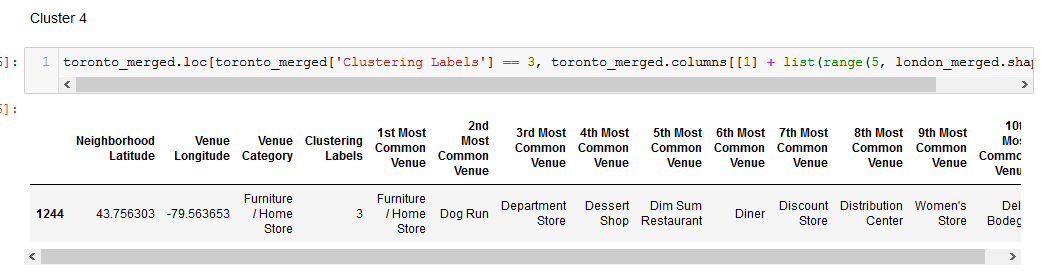
*Fig.14 Our neighbourhood clusters visualised for Toronto.*

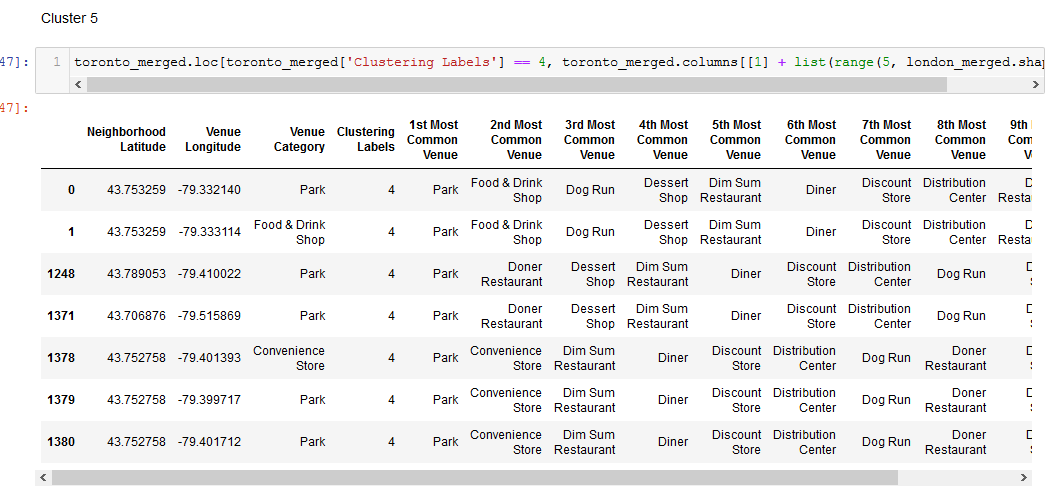
And finally – our clusters showcased below.







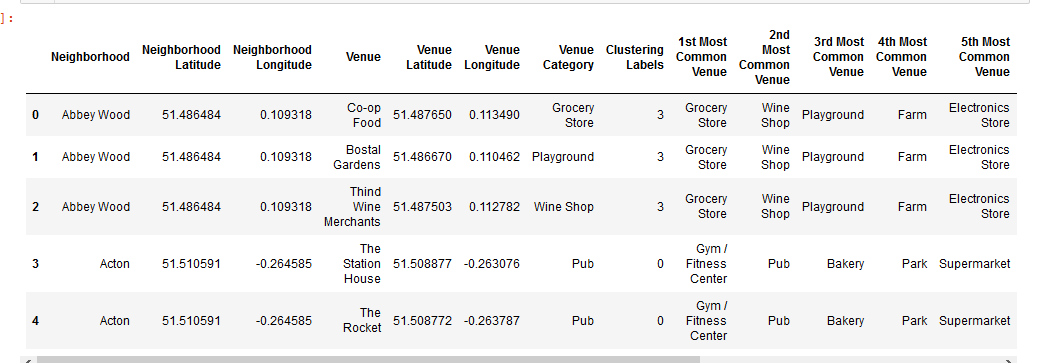




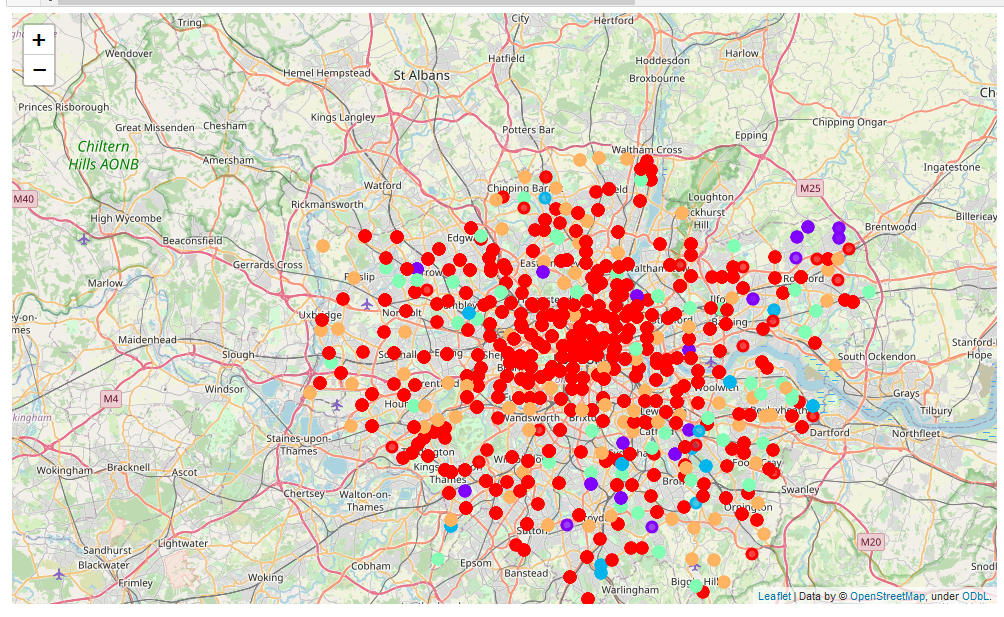
*Fig. 15 – Toronto clusters as derived from the K-Means algorithm.*

**London**

As similar for Toronto, we merged our dataframes to encompass both venue data and the neighborhood lat/long.

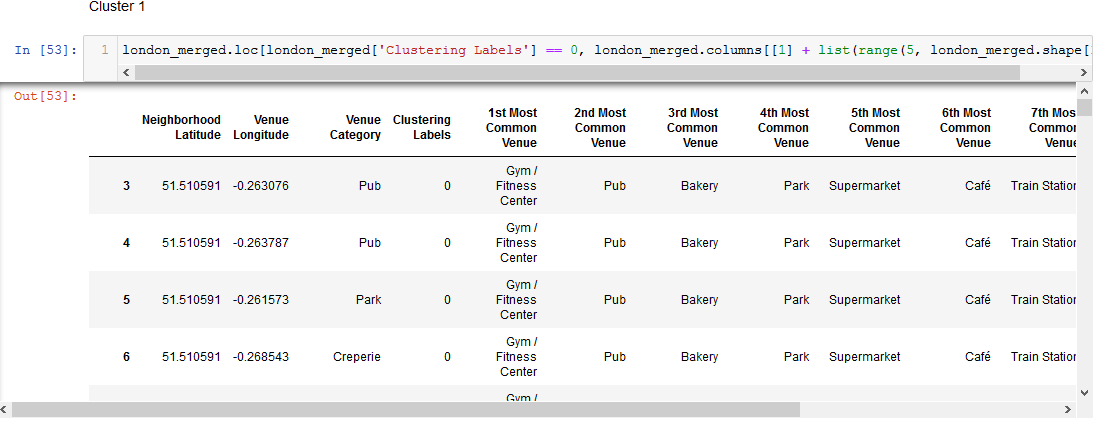


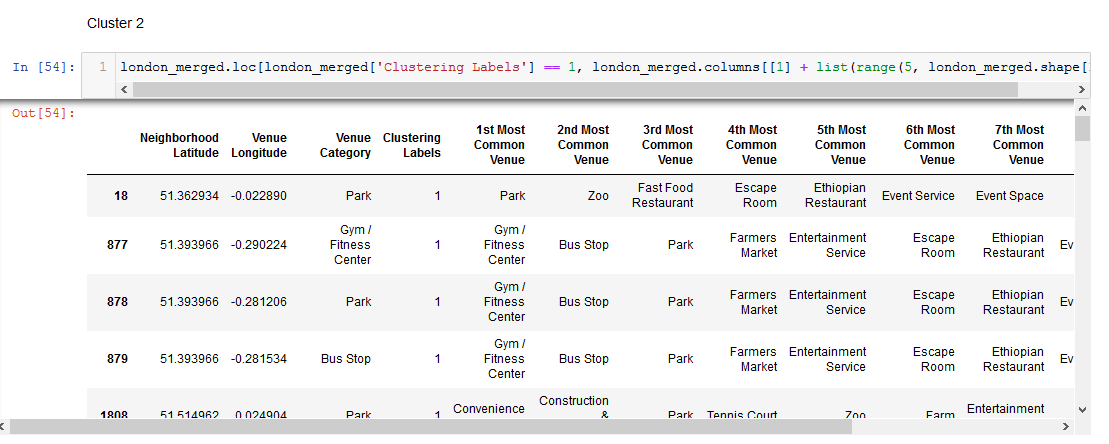
*Fig 16. Our merged dataframe for London.*

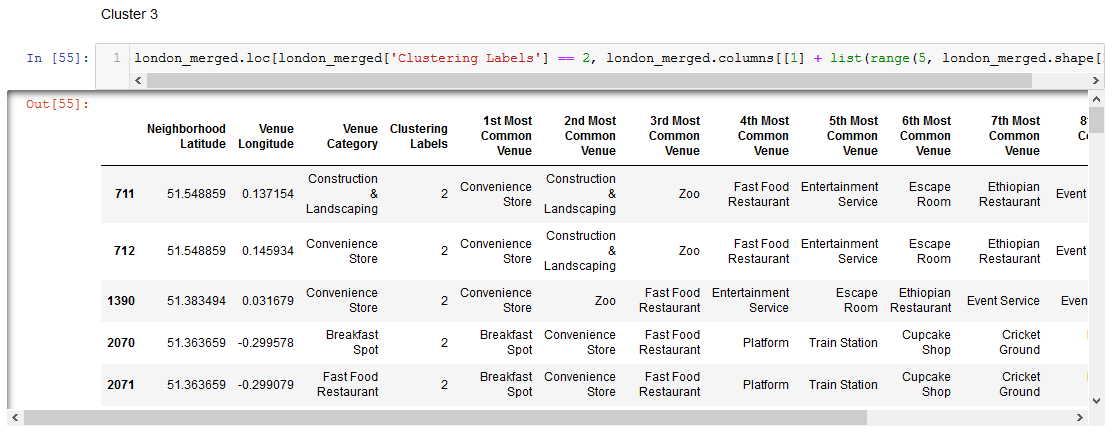


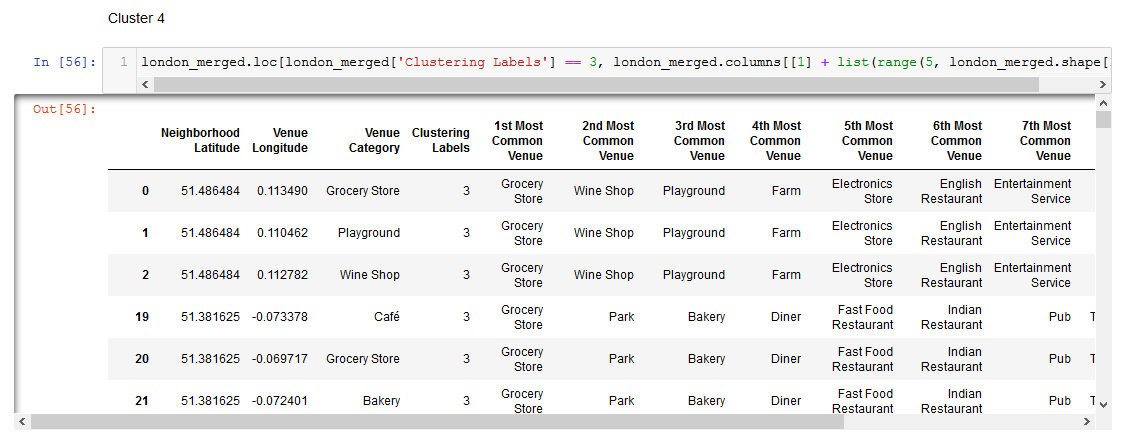
*Fig 17. London neighborhoods clustered visually shown.*

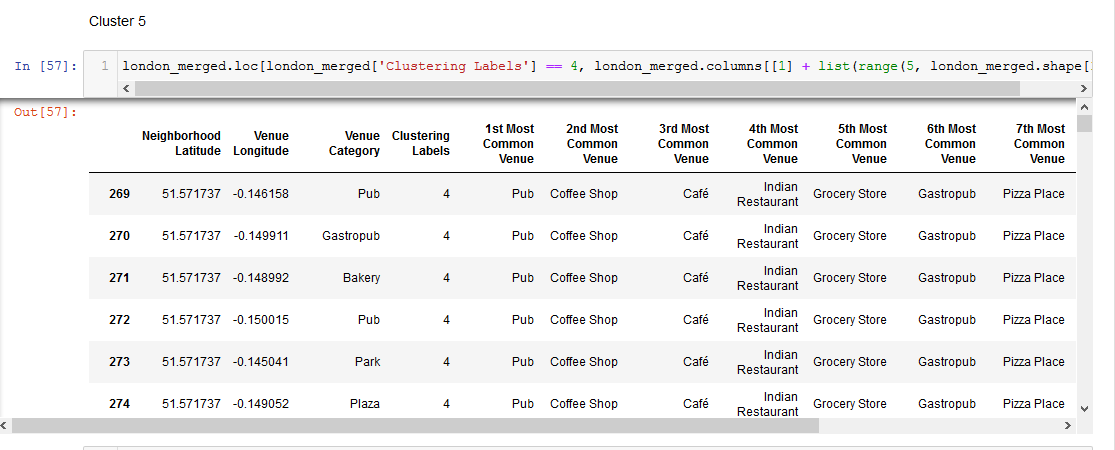
And as per Toronto – our five clusters:











*Fig. 18 – London derived cluster analysis*

**4. Results & Discussion**

From the cluster analysis of Toronto, we can see definitive local tastes indicative of cultural favourites – most notably, the presence of large numbers of hockey arenas in the city. In the other clusters, we see the presence of large numbers of parks and open green spaces, which indicates that Canadians largely favour outdoor activities, coupled with sports. Furthermore, as expected of a capital city; it is highly multicultural, with large numbers of Portuguese and French restaurants making appearances in the clustering results.

Comparably, the neighbourhoods of London are also highly cosmopolitan. There are numerous different cuisine options including Indian, Italian, Turkish and Chinese. Interestingly, there is a reasonable presence of African restaurants, which is understandable when considering the historical imperial context of the city. Furthermore, it has a lot of shopping options too with that of the electronics stores, flower shops, fish markets, and clothing stores. The main modes of transport seem to be buses and trains, with the clustering algorithm picking those up. For leisure, the neighbourhoods are set up to have lots of parks, gyms and historic sites. Most notably, the large number of pubs are indicative of local and historical preferences, which continues into the current day. With respect to London especially, the K-Means algorithm has located a definitive cluster of neighbourhoods towards the east and south of the city - these areas are typically home to emigratory communities, which has been interesting to see this noticed by the algorithm.

**5. Conclusion**

The aim of this project was to explore the cities of London and Toronto and see to what extent the differences of the two cities were. We explored both the cities based on their postal codes and then extrapolated the common venues present in each of the neighborhoods, finally concluding with clustering similar neighborhoods together.

Both cities encompass high levels of multi culturalism, whilst retaining the presence of cultural favorites – in the case of Londoners this is the presence of pubs, and in Toronto – Hockey pitches and arenas. The ability of the K-Means algorithm to identify clusters in such densely packed cities is a testament to its capability and robustness.

**References**

[1] https://en.wikipedia.org/wiki/List\_of\_areas\_of\_London

[2] https://en.wikipedia.org/wiki/List\_of\_postal\_codes\_of\_Canada:\_M

[3] https://www.ordnancesurvey.co.uk/about