The Battle of the Neighbourhoods –

A Comparison of Neighbourhood Features in Major Financial Capitals

Samuel Lee

October 9th, 2019

Project github repository**:** <https://github.com/Sam-Lee1/Coursera_Capstone>

**1. Introduction**

**1.1 Background**

We will analyse and compare neighbourhoods in major financial capitals around the world; London, New York, Frankfurt and Toronto. All these cities are diverse, bustling metropolises; containing major stock exchanges, leading businesses and tourist destinations. As such a lot of people visit them or move to them, often not knowing much about the area – we will use Foursquare venue data about neighbourhoods in these cities to help solve this problem, by using it’s venue data to characterise what these neighbourhoods are like, and to categorise similar neighbourhoods, helping people make more informed decisions about what areas to explore moving to or visiting. Residents looking to move elsewhere in the city could similarly leverage this information. We can further examine if any features stand out about these neighbourhoods, ie if they’d particularly suit young families, tourists or professionals. We will then represent this data by mapping the categorised neighbourhoods on the city maps. Further we shall analyse the distribution of neighbourhood types in each city, to compare the degree of similarity between city pairs.

So, we have a few different potential target audiences for our data and analysis - our primary audience being people relocating, within or into these cities, along with people visiting these cities for business or pleasure; to see whether and where they'd like to go. Further we shall target business owners, either looking to relocate their offices or looking for gaps in the market that their business might fill; seeing which areas would suit them best.

**2. Data**

**2.1 Data Sources and Acquisition**

There are four types of data we shall need to gather:

Post/Zip code area data – we shall scrape this data as needed from [Mapawi](http://zip-code.en.mapawi.com/germany/10/kreisfreie-stadt-frankfurt-am-main/2/165/frankfurt-am-main/60311/5404/%20) for Frankfurt, from [Doogal](https://www.doogal.co.uk/london_postcodes.php%20) for London and from [Wikipedia](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) for Toronto, and define these as our neighbourhoods or districts for each area. For New York we shall query the geopy python library using neighbourhood coordinates from [the NYU spatial data repository](https://cocl.us/new_york_dataset) and extract the zip codes.

Coordinates – we shall query the geopy python library to get the latitude and longitude for each post/zip code area. For New York, we shall instead use coordinate data from the NYU spatial data depository. Neighbourhood data is already listed in our Wikipedia source for Toronto and on the NYU spatial data repository for New York.

Neighbourhood data – for London and Frankfurt we shall use our coordinate data to query the geopy python library, and from the returned address data we shall extract the names of the neighbourhoods that best fit our post/zip code areas.

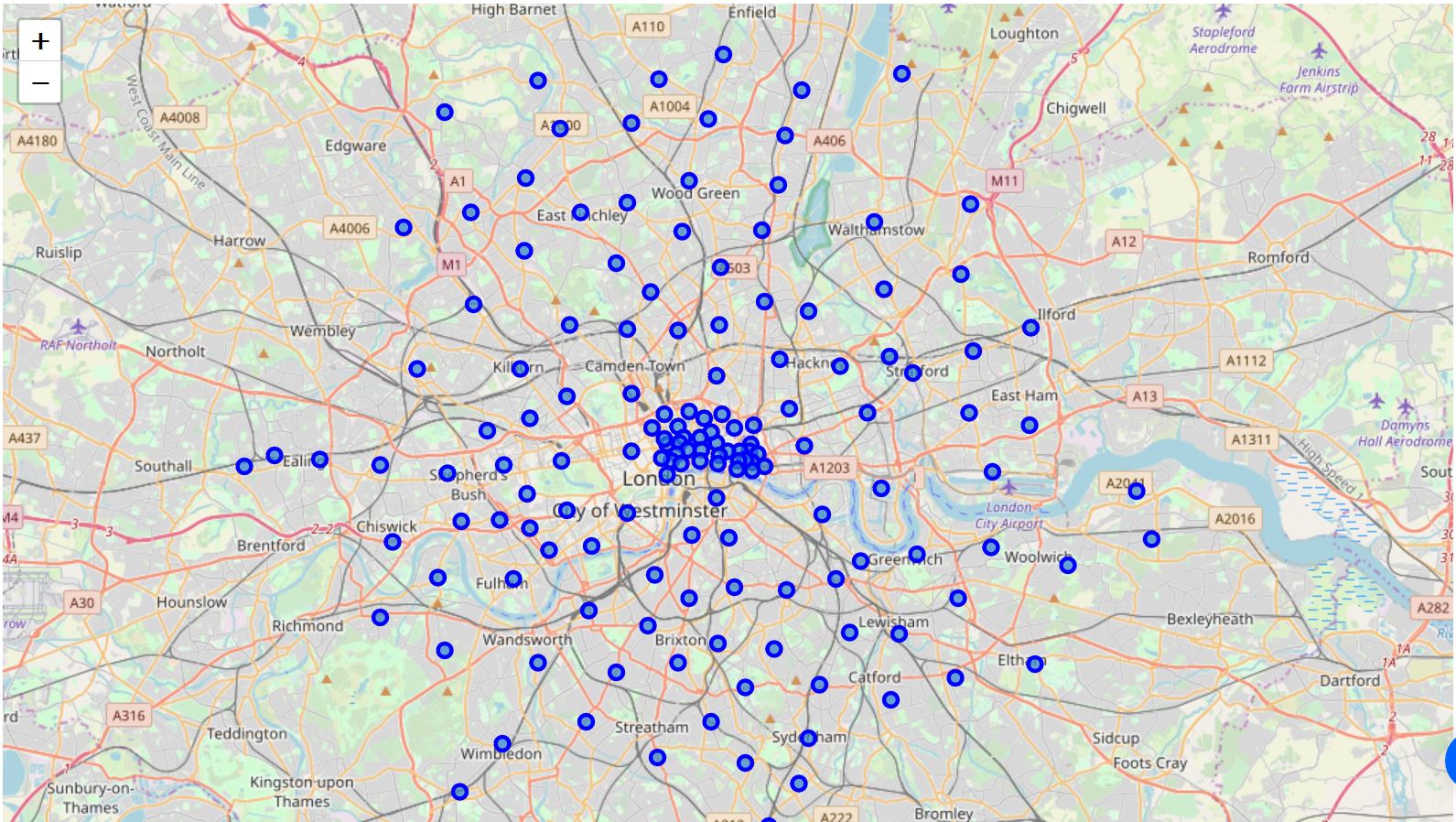
Venue data – we shall make calls to the Foursquare API using these coordinates, to obtain information about nearby venues in the catchment area, and to categorise these venues. We will clean and process this categorised venue data to obtain the relative frequency of different venue categories for each neighbourhood, and therefore the most prevalent venue types for each neighbourhood; ie that Central Harlem in New York has African restaurants as its most frequent venue type, whereas for North York in Toronto it is coffee shops.

So in summary, we shall gather post/zip code area data and combine it with coordinate data, to define and locate our neighbourhoods. Having done this, we shall use the foursquare API to gather nearby venues and categorise them, then further processing this data, to obtain the ranked relative frequency of each venue category for each neighbourhood. It is this categorised venue frequency data that we shall analyse and extract features from. Using k-means cluster analysis we will categorise similar types of neighbourhoods, based on their nearby venues. We shall then examine and characterise these neighbourhood groupings, extracting features such as neighbourhoods that have many facilities for children, which would suit young families. We shall represent this data on labelled navigable maps and finally we shall use it to compare the overall relative similarity between our cities.

**2.2 Data Cleaning**

Having gathered the post code and neighbourhood data as discussed above I obtained separate data frames for each city and formatted them such that they included post/zip code, neighbourhood and coordinates for each neighbourhood. Querying some post/zip code data produced redundant coordinates, 11 of our 156 London postcodes and 8 of our 32 Frankfurt zip codes. These redundant coordinates were dropped from the dataset, leaving 583 neighbourhoods across the four cities to analyse. This neighbourhood location data was then combined into an inter-city data frame and rendered as a folium map, with a labelled node for each neighbourhood, as shown below:

The inter-city dataset rendered as a folium map, screenshotted centred on London:

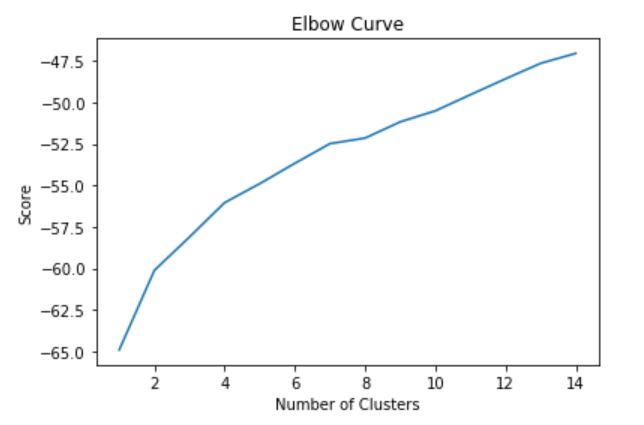
****

Next the foursquare API was queried for each neighbourhood in each cities data frame in turn. 5 neighbourhoods didn’t return venue data from foursquare and were dropped from the dataset, leaving us with 578 neighbourhoods. Neighbourhood venue data was used to create one hot encoded data frames of the mean occurrence of each venue category for each neighbourhood, ready for analysis. The mean category occurrence tables were again combined into another inter-city data frame.

**3. Methodology**

Having gathered and prepared the data it was now ready for analysis. K means clustering was chosen as the primary method of analysis, since the returned clusters will be able to be directly used by potential users to make judgements about what neighbourhoods might be like to visit or work in and make informed decisions about what options to explore. In order to use K means cluster analysis I first had to decide what the best number of clusters to use was, I chose to analyse the range of 1 – 15 clusters, since a larger number would lead to too many small clusters for a data set of this size and have an impractical computational overhead. So, the one hot encoded inter-city data set was clustered using 1 through 15 clusters in turn, and the score method was used to evaluate how far the points were from the cluster centroids. This data was then plotted onto the below elbow, to analyse were the marginal decrease in total distance from the centroids tended to a limit and therefore the increase in accuracy of our cluster model became negligible with the addition of additional clusters.

Elbow curve analysis of K means clustering on the one hot encoded inter-city data set:



On examining the elbow curve, it was apparent that there wasn’t a classic sharp elbow point within the chosen range. Examining higher numbers of clusters was discounted as a possibility for the reasons discussed earlier, there is a significant decrease in marginal improvement at 7 clusters and this is a good number of clusters to represent our data, so 7 clusters was chosen for the analysis.

K means cluster analysis with 7 clusters was performed again on the data set and broken down for further analysis. This returned 7 clusters of size 24, 149, 242, 62, 13, 1 and 87. A cluster of size 1 wouldn’t be helpful for making comparisons and can be considered an outlier. I deferred deciding how to handle the outlier until labelling the clusters, where it became apparent that this neighbourhood was a good fit for one of the other clusters and it was amalgamated into cluster 4, as you can see in the results section. The now 6 clusters were labelled in a way that characterised them based on their most frequently occurring venues and colour coded for addition to the map.

Next the distribution of these clusters in each city was analysed, and of the inter-city dataset. The total neighbourhoods of each cluster type were converted to a relative frequency for each, then the sum of the mean square difference between each cluster type was evaluated between each city pairing and between each city and the overall dataset.

**4. Results**

**4.1 Cluster Breakdown**

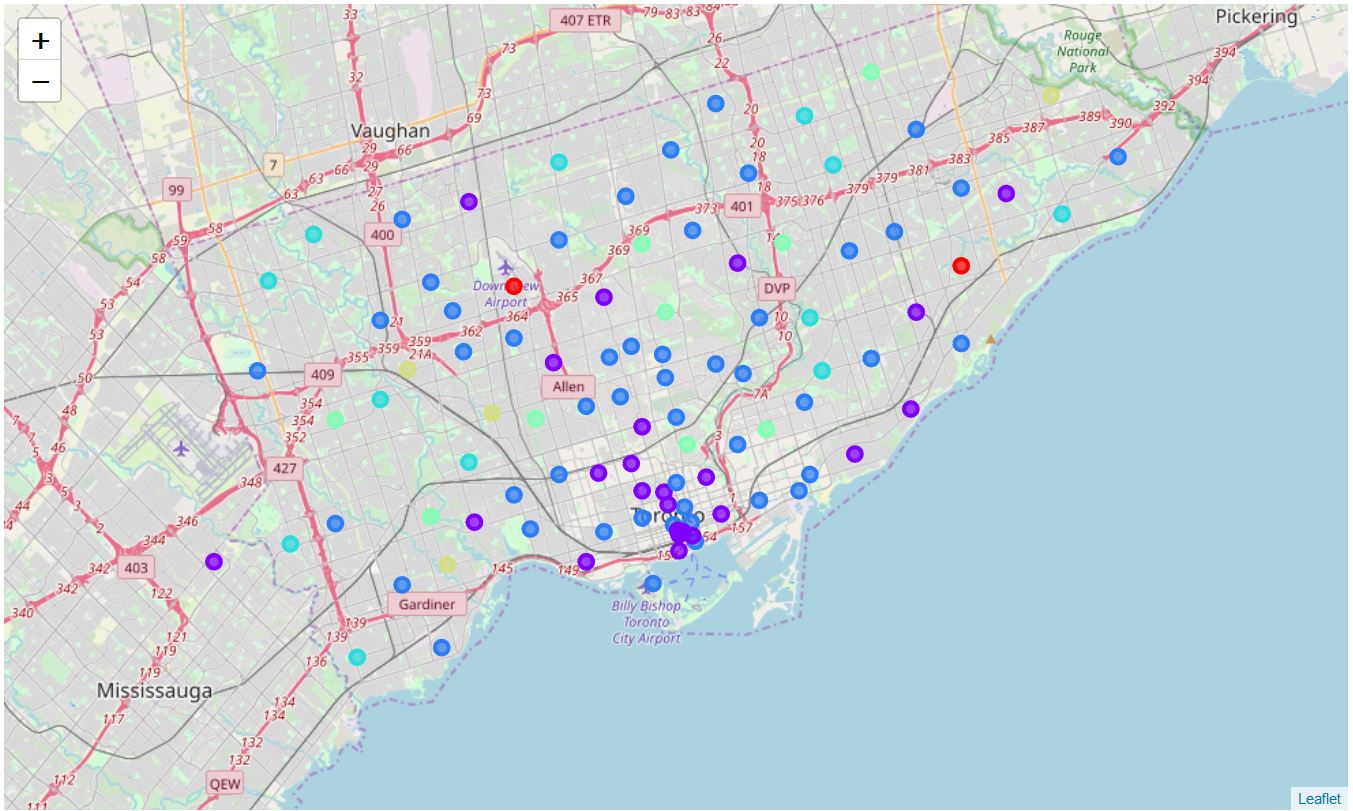
The outlier cluster was deemed a good fit for cluster 4 and amalgamated into it, cluster 6 was then relabelled cluster 5 such that the clusters were still consecutive numerically. The 6 clusters:

* Cluster 0, labelled ‘Theme Parks and Recreation’ due to frequency of Theme parks and other forms of recreation near these neighbourhoods, coloured red, 24 neighbourhoods
* Cluster 1, labelled ‘Coffee Shops, Cafes and Pubs’ due to high frequency of each in neighbourhoods, coloured purple, 149 neighbourhoods
* Cluster 2, labelled ‘World Foods’ due to high frequency of restaurants serving cuisine from around the world, coloured blue, 242 neighbourhoods
* Cluster 3, labelled ‘Pizza Places and Fast Food’ due to high frequency of both in neighbourhoods, coloured turquoise, 62 neighbourhoods
* Cluster 4, labelled ‘Green and Peaceful Spaces’ due to high frequency of both in neighbourhoods, coloured green, 14 neighbourhoods
* Cluster 5, labelled ‘Convenience Stores’ due to high frequency of various shops, especially convenience stores in neighbourhoods, coloured yellow, 87 neighbourhoods

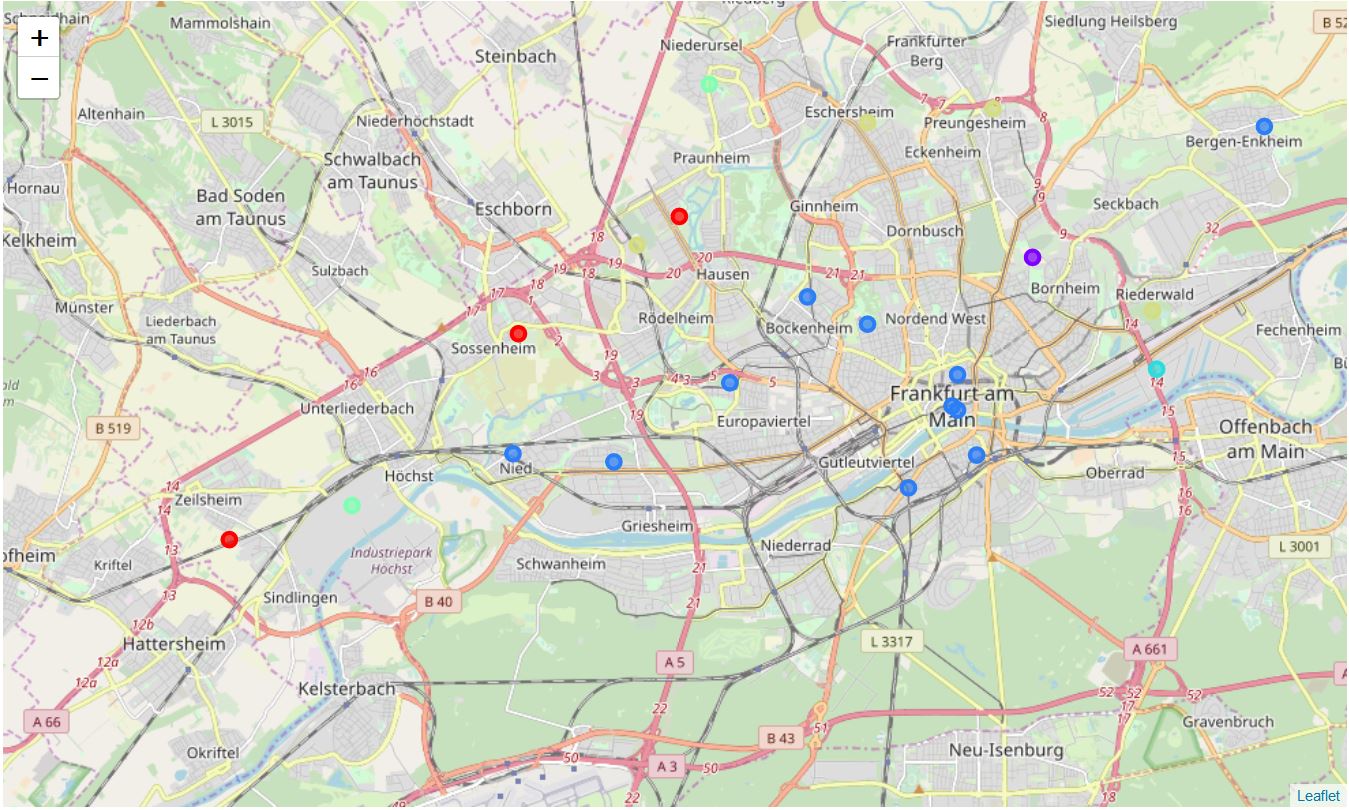
**4.2 Cluster Map**

These clusters were then used to colour code and label our folium map. Our map, shown below centred on each of our cities in turn, had a node added for each neighbourhood; labelled with the cluster name, neighbourhood and post/zip code. Labelled interactive folium cluster map available here: <https://nbviewer.jupyter.org/github/Sam-Lee1/Coursera_Capstone/blob/master/Interactive%20Map.ipynb>

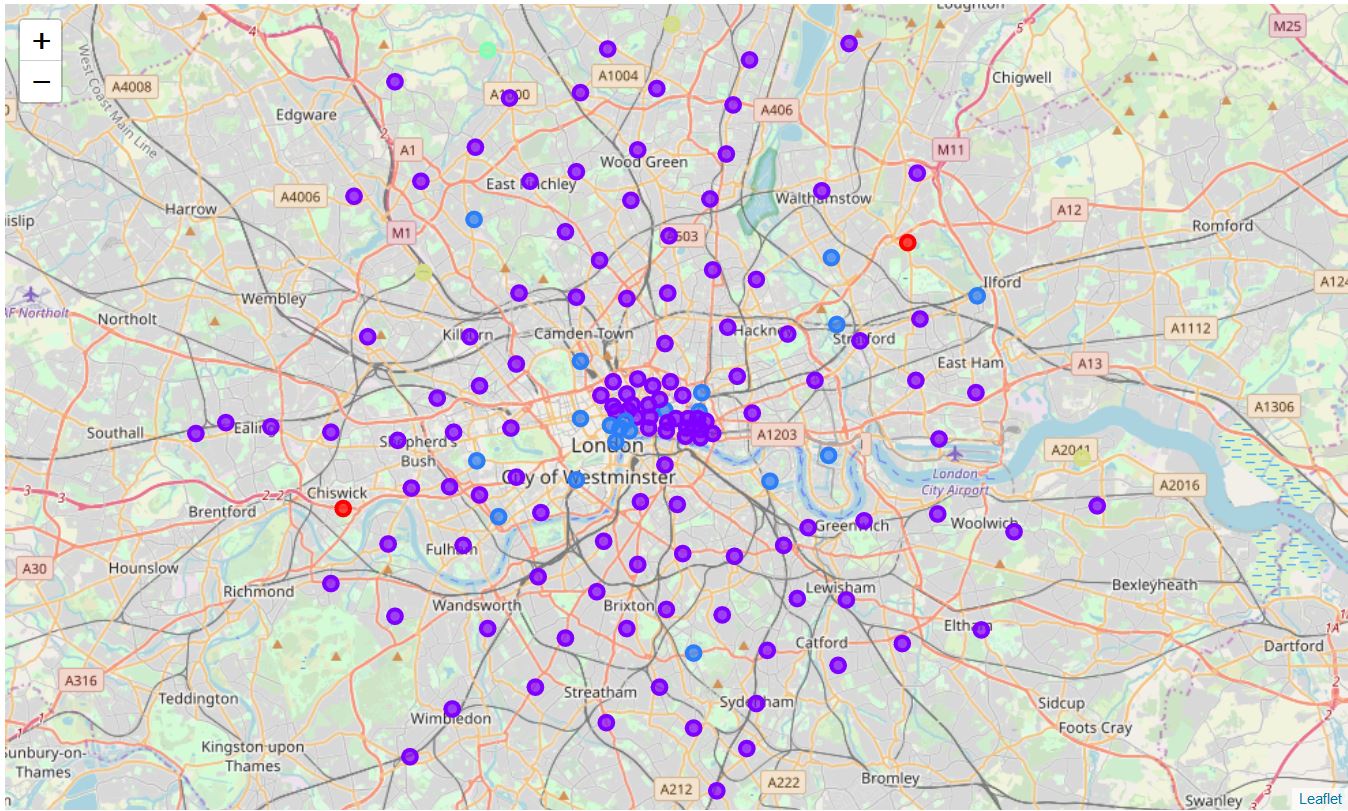
Centred on Toronto:

****

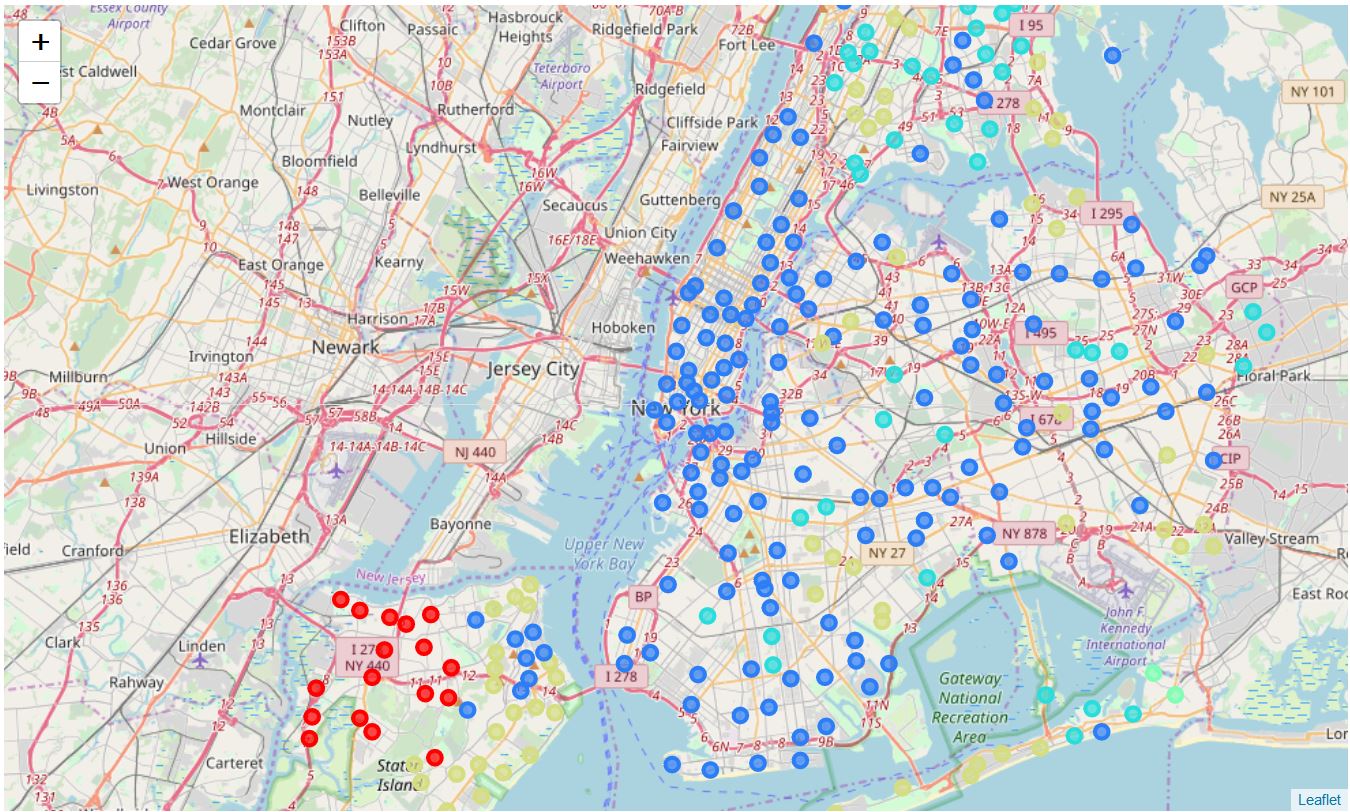
Centred on Frankfurt:

****

Centred on London:

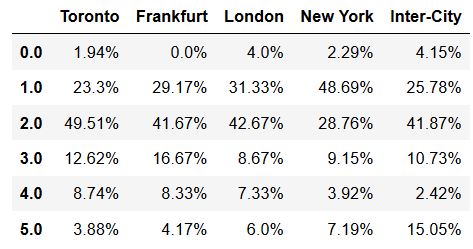
****

Centred on New York:

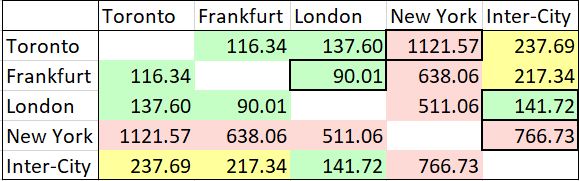
****

**4.3 Mean Square Difference of Relative Neighbourhood Cluster Frequencies**

In order to evaluate the relative similarity of the different cities, further analysis was performed on the cluster types of neighbourhoods in each city. First the relative frequency of each cluster type in each city and the inter-city dataset was evaluated:

 Relative frequency of neighbourhoods of each cluster in each city

This data was used to find the sum of the mean squared differences in cluster distribution between each pair of cities, and between each city and the inter-city dataset. More similar pairs are coloured green, to less similar red, and the most similar and dissimilar city pairs are highlighted, along with the cities most similar and dissimilar to the overall dataset:

****

**5. Discussion**

So, 578 neighbourhoods across four cities have been clustered and characterised. There is some scope to extract useful features from the characterising of the clusters at the current level of detail, it is reasonable to surmise neighbourhoods in the ‘Theme Parks and Recreation’ cluster or the ‘Green and Peaceful Spaces’ cluster might be more suitable than average for young families, with the greater availability for facilities for young children. Our other clusters, such as ‘World Foods’ might be of greater interest to those who particularly enjoy those facilities but to extract more features that might suit certain demographics will require examining the venue data in more detail.

A quick glance at the maps can reveal some information about the types of venues most frequently occurring in neighbourhoods in our cities. Particularly, London’s prevalence of neighbourhoods in the ‘Coffee Shops, Cafes and Pubs’ cluster and New York’s prevalence of neighbourhoods in the ‘World Foods’ cluster.

The analysis of the sum of mean squared differences in relative cluster frequency between each city and the overall dataset generates a few main takeaways. Firstly that the most similar neighbourhood distribution between our cities was that of London and Frankfurt, and that London had the most similar distribution to the overall dataset. Further that New York and Toronto were the most dissimilar, and New York was the most dissimilar city to our overall dataset.

**6. Conclusion**

**6.1 Possible Future Directions**

In future we could scale up the concept, to include more cities across more countries or take a more thorough look at an individual country. Another potentially fruitful direction to take would be deepen the analysis of each neighbourhood. The data could be combined with house or rent price data or with employment opportunity data to get a more complete picture of what living in an area might be like.

The true potential of this methodology could be unlocked by building an application based on the data – features such as choosing your neighbourhood and finding the most like it in an area or searching for neighbourhoods based on desired features could provide easy access to and interpretation of the data for an end user. Such an application would be more effective the more data had been made processed.

**6.2 Conclusion**

To summarise, analysis of nearby venue data for 578 neighbourhoods in 4 cities has yielded location, name and characterising information for each. This information has been tabulated and rendered as navigable, labelled maps. Further analysis has yielded a general measure of similarity between each city.

This information could be used by an interested party to prioritise what areas to explore establishing a business in, moving to or visiting. The true potential for easy access to and interpretation of this data would be enabled by building an application on top of it, and also by expanding the scope of data collection.