Data Science Specialization Capstone Project

Comparison of Neighborhoods of Bengaluru, Seoul, Vancouver and San Francisco

Introduction

The problem I am considering is to compare the neighborhoods of four cities in four different countries. The countries I have selected are Bengaluru in India, Seoul in South Korea, Vancouver in Canada and San Francisco in USA. Attempt will be made to check which neighborhoods of the 4 cities are similar.

The target audience for this project is the owners of a restaurant chain which might already have their franchises set up in Vancouver and San Francisco and who want to enter new markets in Asia. They might consider other prominent tech cities in Asia since their target customer is the tech community. Since Seoul and Bengaluru have many MNCs and have a large tech community, they are the target of this project.

Data Sources

The neighborhood data of the four cities is taken from Wikipedia pages.

https://en.wikipedia.org/wiki/List_of_wards_in_Bangalore

https://en.wikipedia.org/wiki/List_of_districts_of_Seoul

https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Vancouver

https://en.wikipedia.org/wiki/List_of_neighborhoods_in_San_Francisco

Bengaluru has 199 neighborhoods, Seoul has 25 neighborhoods, Vancouver has 33 and San Francisco has 114 neighborhoods.

Importing Data

Neighborhoods lists of Cities

We already have the links of the Wikipedia pages from which we can get the list of neighborhoods in each city. Beautiful Soup library is used to extract the information from the wiki tables in the pages. This data is stored in a pandas data frame. Along with the neighborhoods, the city name, the state name and the country name are stored.

Geolocation of the Neighborhoods

The geopy library is used to get the location data of the Neighborhoods. Now in geopy library, Nominatim service is used. For using the free service of Nominatim, there is a restriction of 1call per sec to the service. To avoid 'timeout' error, there needs to be at least 1 sec gap in each call even if a for loop is used. To provide a sufficient gap to accommodate network delay, a gap of 2 sec is provided. The gap is provided by calling the sleep function.

Timeout Errors

Even after providing a 2 sec gap in calls, there are timeout errors. So, to handle these errors is simple. Simply call the Nominatim service again for these locations after checking for network connectivity.

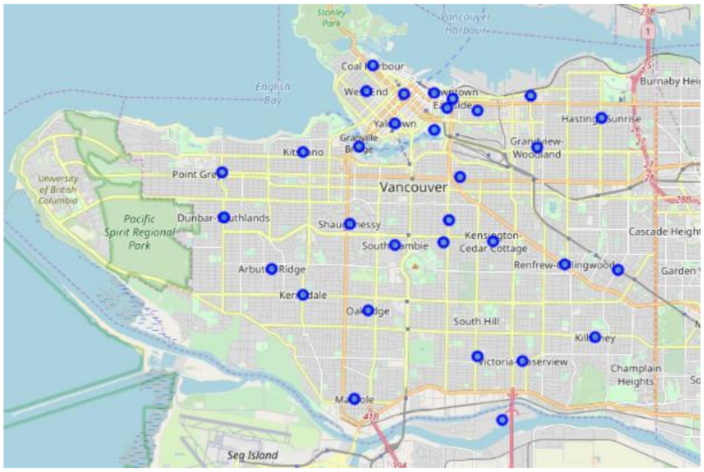
Missing/No Coordinates

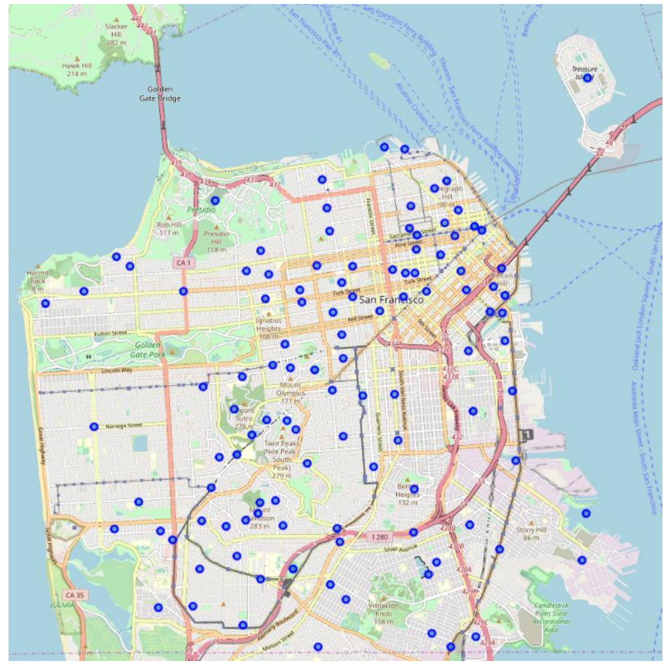
Some locations will not resolve into coordinates. This can happen because some locations may have different spellings. These can be rectified by using different spellings. Some locations will not resolve despite that. Then that data is procured manually searching on Google Maps.

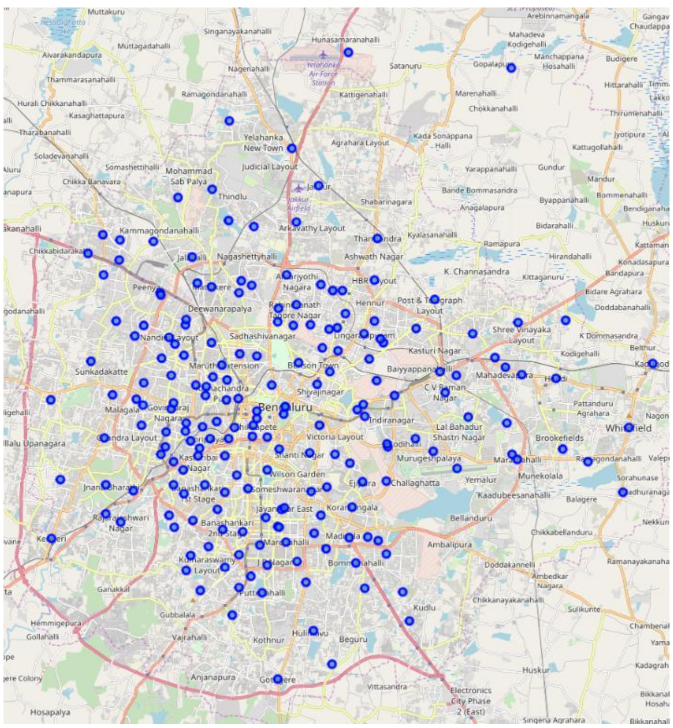
Maps

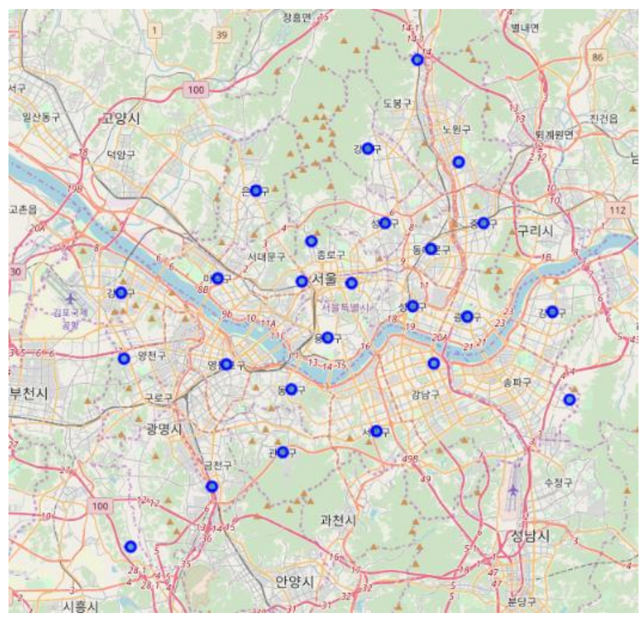
Maps are generated for each city with neighborhoods shown as markers. But before that a world map is created with the cites as markers. Below is the world map.

Maps of the cities are below.









Venues

Getting Venues

City	Area (km^2)	No. of Neighborhoods	Avg Neighbourhood Radius Considered (km)
Bengaluru	709	195	1000
Seoul	605.2	25	2500
Vancouver	115	33	500
San Francisco	121.4	114	500

Individual Clustering Results

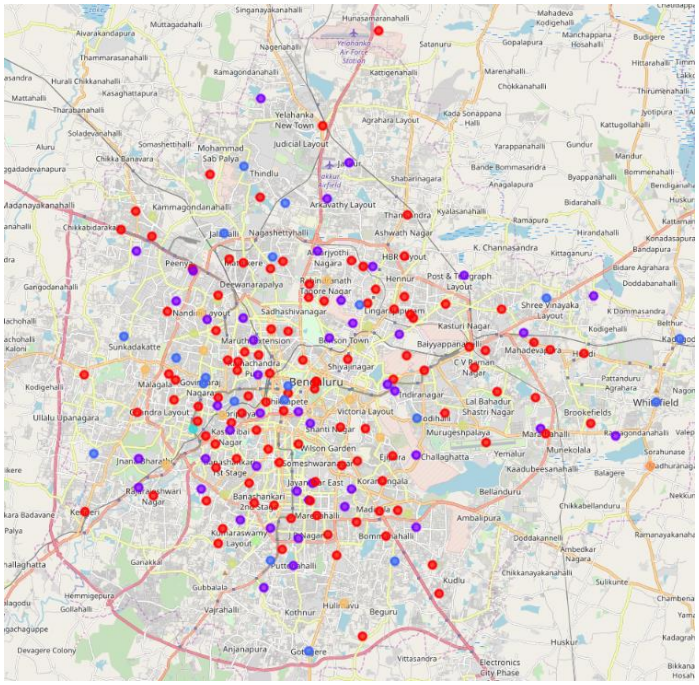
Individual Clustering will help understand how the individual locations can be clustered. To be consistent with all the individual location clustering and the complete clustering, there are going to be 8 clusters.

It must be noted that cluster labels are not the same across different locations. So, Cluster 0 in Bengaluru is not the same as Cluster 0 in San Francisco.

Bengaluru

Looking at the clustering of the neighborhoods in Bengaluru, there are 5 clusters with 3 possible outliers, with majority of the neighborhoods in cluster 0. The neighborhoods of all clusters look equally distributed on the map.

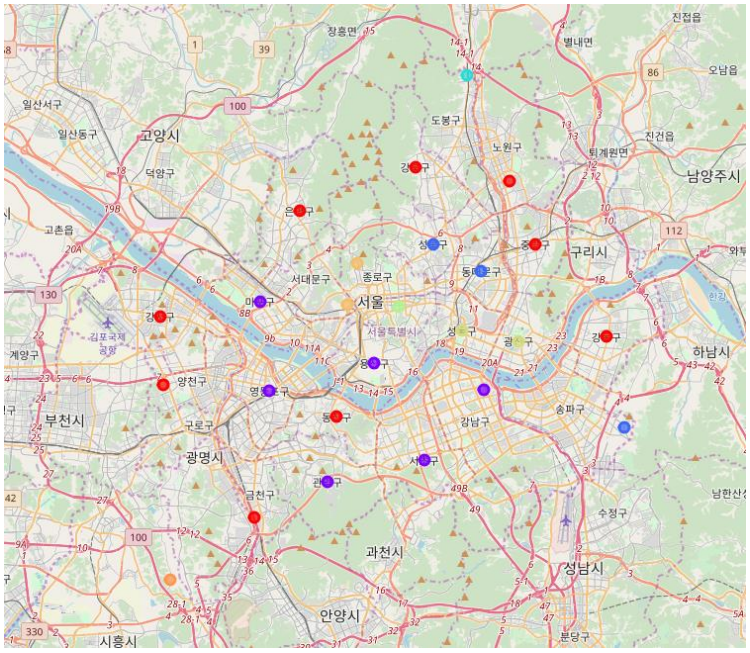
Cluster Labels	Neighbourhood
0	108
1	9
2	1
3	1
4	45
5	10
6	1
7	20



Seoul

In Seoul, there are only 25 neighborhoods, so it's difficult to tell if the clusters with only one neighborhood are outliers or not. Looking at map, neighborhoods in cluster 0 are at the edges of the city.

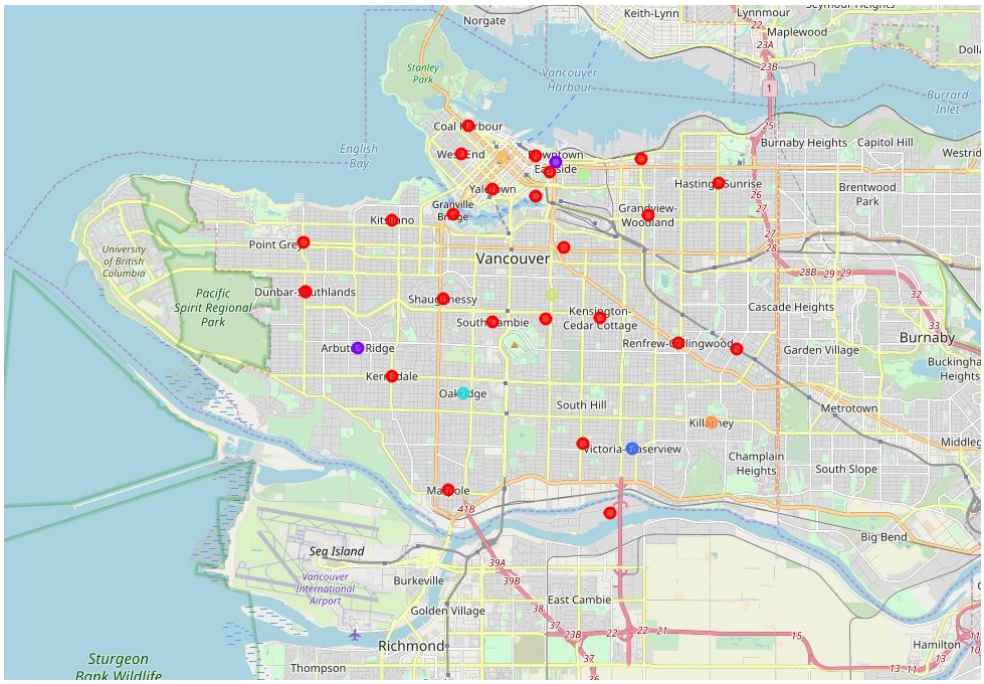
Cluster Labels	Neighbourhood
0	9
1	2
2	2
3	3
4	1
5	6
6	1
7	1



Vancouver

There are 33 neighborhoods, so it's difficult to tell if the clusters with only one neighborhood is outlier or not. Most of the neighborhoods are in cluster 1.

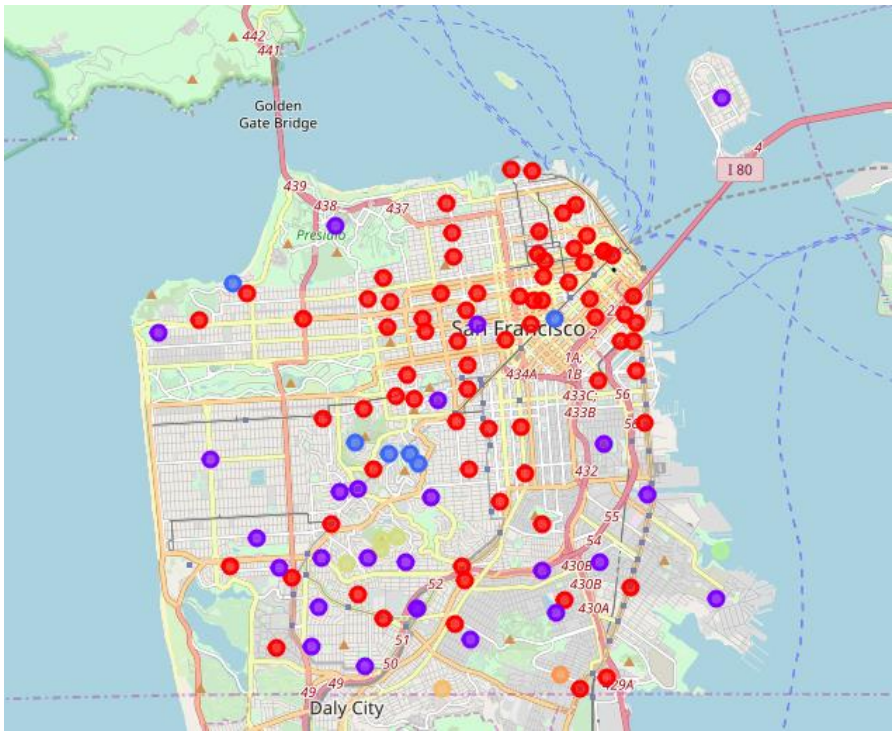
Cluster Labels	Neighbourhood
0	1
1	25
2	1
3	1
4	2
5	1
6	1
7	1



San Francisco

The neighborhoods can be divided into 4 clusters and possible 4 possible outliers.

Cluster Labels	Neighbourhood
0	1
1	1
2	1
3	6
4	4
5	74
6	1
7	26



Complete Clustering Results

Cluster Labels	City	Neighbourhood
0	San Francisco	1
1	Bengaluru	12
	San Francisco	6
2	Bengaluru	15
	San Francisco	6
	Seoul	1
	Vancouver	2
3	Bengaluru	1
4	Bengaluru	1
5	Bengaluru	64
	San Francisco	35
	Seoul	15
	Vancouver	17
6	Bengaluru	102
	San Francisco	65
	Seoul	9
	Vancouver	14
7	San Francisco	1

The complete clustering gives some interesting results. Seoul and Vancouver can be divided into three clusters which the other locations also have. Bengaluru and San Francisco have a lot of common clusters. There are some possible outliers in San Francisco and Bengaluru which don't fit in any common clusters.

The big takeaway from this is that there are three clusters with neighborhoods from all the locations. So, these neighborhoods can be considered similar based on the venues present in them.

Discussion

The objective of this analysis was that if there are is a restaurant franchise in both Vancouver and San Francisco and they want to open a new franchise in Bengaluru and San Francisco then in which neighborhood of the cities they should open. Based on Complete Clustering neighborhoods in clusters 2,5 and 6 are similar neighborhoods. So, if the restaurant in the neighborhoods of these clusters in Vancouver and San Francisco then a new franchise can be opened in the neighborhoods of the same clusters in Bengaluru and Seoul. Since majority of the neighborhoods of all the locations are in these three clusters then there is a good probability of finding a match.

Conclusion

The result showed that the restaurant franchise can be opened in Bengaluru and Seoul though more data and analysis is needed. More data like the customer rating and pricing details will help but with the free Foursquare API there is limited access to the required data.