Capstone project: Finding the closeness between cities based on venue categories

---

Muhammed Zahit Gök

08.01.2020

---

1. Introduction

1.1 Description and Discussion of the background

The capital cities especially having big population are the important indicator to understand the trend and similarity between the countries. In this study, we used the venue data of three large citites namely New York, Toronto and Berlin. New York is one of the biggest city of the USA. It contains of 166 neighbourhoods comprised of various venue. Toronto is the capital city of the economy and business centers. Although this city is the center of important business, it does not contain as much neighbourhood and venue as New York. The last city Berlin is the capital and multicultural city of Germany. The population of this city is about 4 million and includes 41 neighbourhood. In this study, we aimed at identifying the closeness between cities based on the venue category. We compared the Toronto and Berlin with New York and found which one is closer to New York.

1.2 Data Collection

- All cities' data related to Borough and Neighbourhoods was collected from the web side Wikipedia
- The longitude and latitude were obtained from the geocoding library of python
- Regarding the venues and venue category, we used the foursquare API

2. Methodology

The data obtained from wikipedia and the data of the latitude-longitude(LL) values were transmitted to the csv file so that we do not repeat the download process repetitively. After combining two tables containing wikipedia and LL data, we have tables including the columns as Borough, Neighbourhood, Latitude, Longitude. We used Four sqaure API with the values radius as 100 and limit as 100 and got the venue and venue category in the neighbourhood. In order to show the point on the map the library folium was used. The total venues each city New York, Toronto and Berlin is shown the figures below.

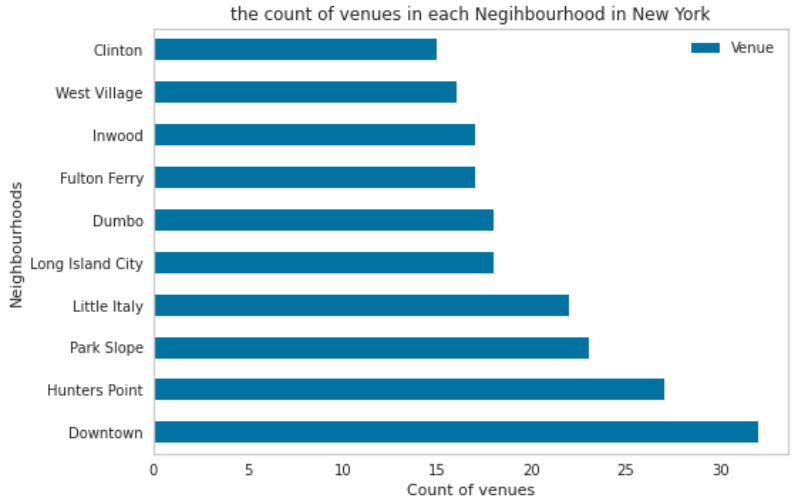

Figure 1: Count of Venues in New York.

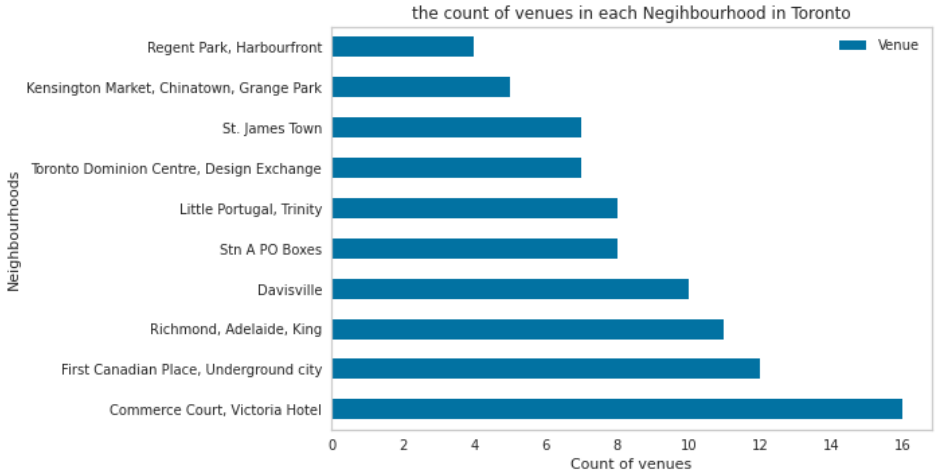

Figure 2: Count of Venues in Toronto

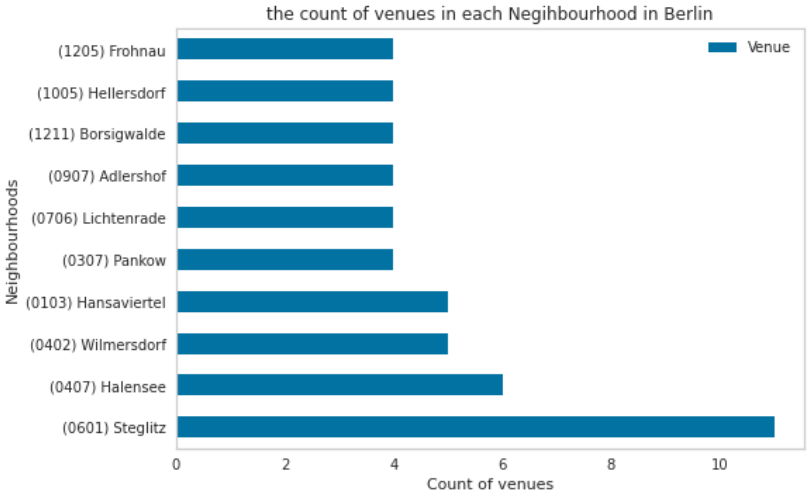

Figure 3: Count of Venues in Berlin.

In the neighbourhoods of New York, Downtown has the biggest number of venue and in Toronto and

Berlin, Commerce Court,Victory Hotel and Stiglitz neighbourhood have much more venues. In order to determine which neighbourhood in Newyork and Toronto are similar to each other, we used k-means clustering based on 10 categories as Café, Italian Restaurant, Coffee Shop, Bakery, Supermarket, Sushi Restaurant, Spa, Restaurant, Asian Restaurant, Burger Joint. Before starting the clustering process, we concatenate the data of New York and Toronto. After that we transfom the table such that the data is converted into as 1 or 0 acording to the existence of the venue category in the neighbourhood. Moreover, we grouped each neighbourhood then got the mean of each venue category. To start the process of k-mean clustering, we wanted to find the best k for clustering algorithm. we used a special library named as KElbowVisualizer performing elbow method. For the data Newyork and Toronto, we found the best k value 9 as shown the figure below.

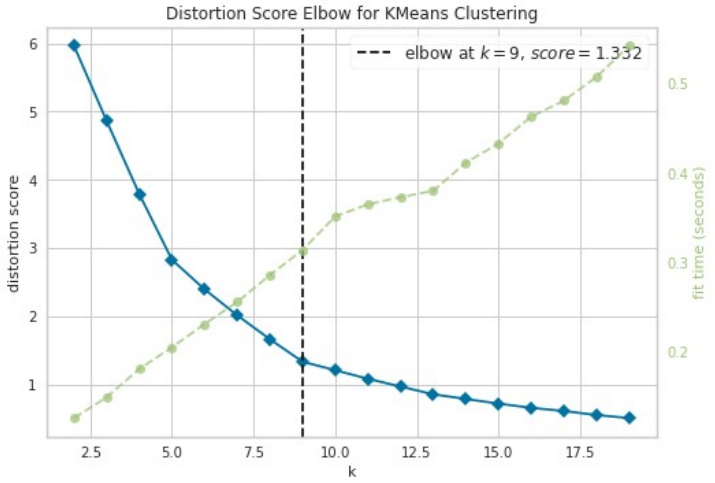

Figure 4: Finding the Best k value for the data Newyork and Toronto

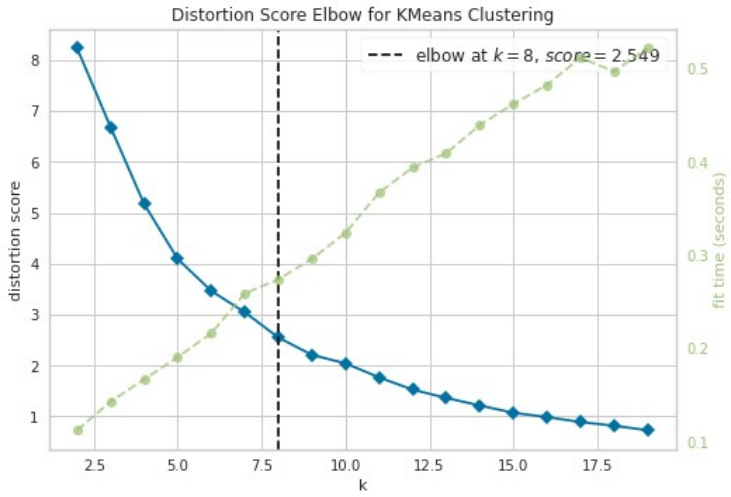

Figure 5: Finding the Best k value for the data Newyork and Berlin

After performing k-means clustering, we got some neighbourhoods of different counties similar to each other. Visualization of which cluster contains of neighbourhood of the cities as observed in figure below.

	Cluster_0	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5	Cluster_6	Cluster_7	Cluster_8
New York	566	2	3	0	3	120	10	6	94
Toronto	45	11	0	1	0	11	15	3	19

Figure 6: the count of neighbourhood in the clusters based on the cities Newyork and Toronto

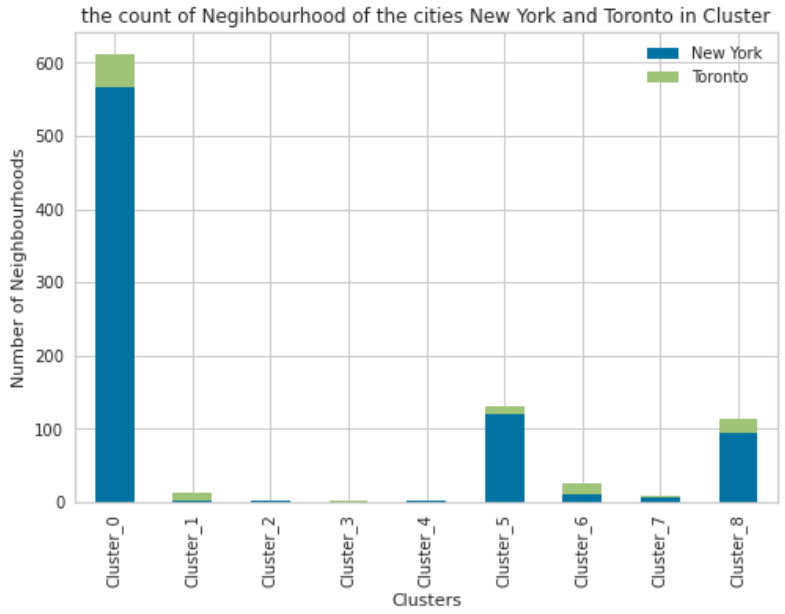

Figure 7: Bar plot of the count of neighbourhood in the clusters based on the cities Newyork and Toronto

We also performed the same process for the data which is the union of New York and Berlin. We found the best k as 8. We got the statistics about the cluster as shown below.

	Cluster_0	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5	Cluster_6	Cluster_7
New York	94	569	3	0	0	122	3	13
Berlin	10	51	3	2	9	5	4	16

Figure 8: the count of neighbourhood in the clusters based on the cities Newyork and Berlin

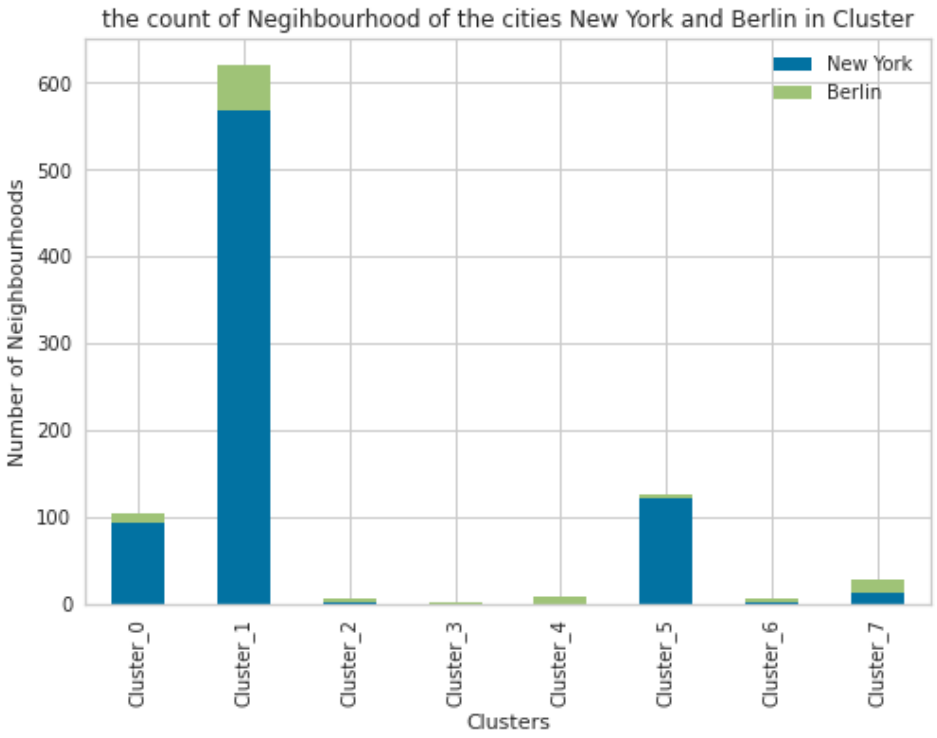

Figure 9: Bar plot of the count of neighbourhood in the clusters based on the cities Newyork and Berlin

This kind of result gives us some information about the closeness between neighbourhoods and cities. However, we need to find the result numerically to understand whether Toronto or Berlin is close to New York. We took advantage of the cosine similarity metric by using the data containing of the distribution of the venue category of each city. We got the sum of each category in each city. As an example of the distribution of new york is shown below. Moreover, After performing cosine similarity metric to all cities, we got the table including similarity score as shown with heat map.

	Value	Count_2
0	Café	15
1	Italian Restaurant	21
2	Coffee Shop	25
3	Bakery	10
4	Supermarket	4
5	Sushi Restaurant	7
6	Spa	4
7	Restaurant	6
8	Asian Restaurant	5
9	Burger Joint	7

Figure 10: the count of venue category based on the cities Newyork

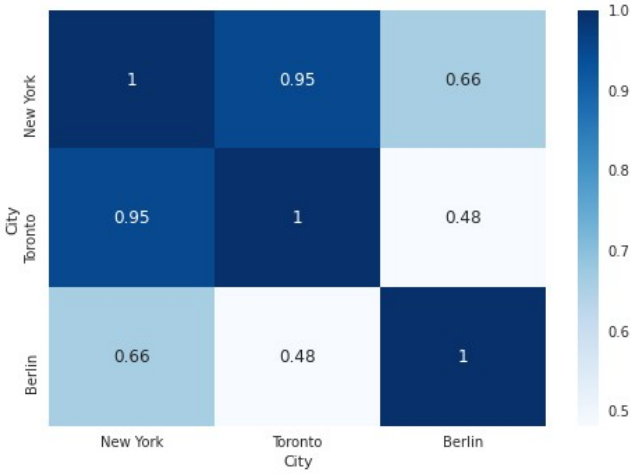

Figure 11: Total count of venue category based on the cities Newyork, Toronto and Berlin

we can conclude that Toronto seems more similar to the city New York The similarity score between the mentioned cities is 0.95. After performing these techniques, we want to show the statistics about the total venue categories in all cities with tree map.

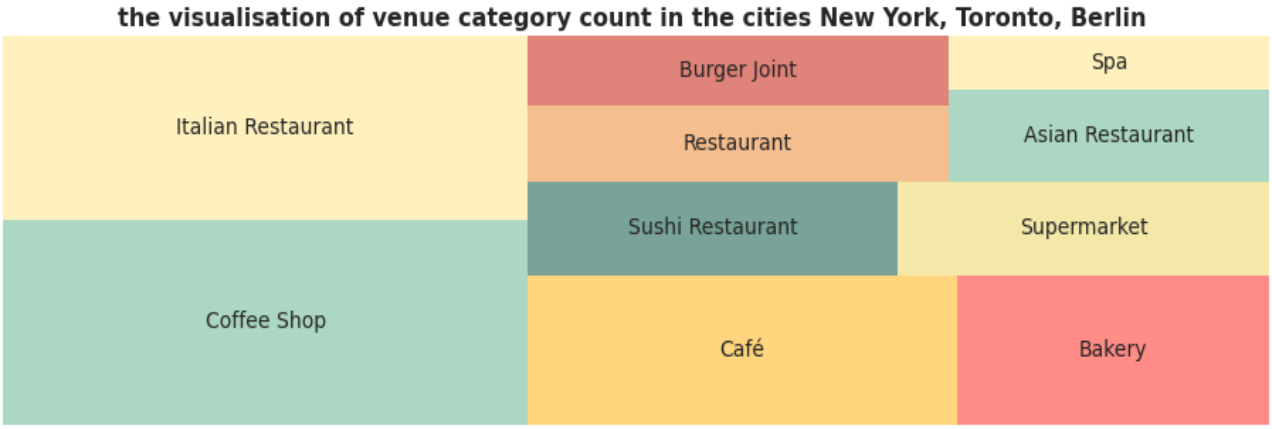

Figure 12: Total count of venue category based on the cities Newyork, Toronto and Berlin

It is imported to note that we did not put the folium figures in the report. The reason is that the physical distance between cities does not enable to show the neighbourhoods similarity in one figure. In this case, the folium figure exists in the jupyter notebook web site so that you can analyse the mentioned cities with the world map.

3. Discussion

These large and important cities can be a sample to show the trends of the country. In this study, we found the similarity between cities by using k-means algorithm and cosine similarity metric. We used a certain number of venue category in our approaches. If we increase the category number, the result may be better. Moreover, radius which determines the area of search can be increased. Thus, we may find better similarity. We also gather some statistics about cities, neighbourhoods and the clusters. At the end of our study, we got the statistical result from the data consisting of three cities.

4. Conclusion

By performing statistical approaches on the big cities, we can find similarity and solve the problems in the identical cities with these similarities. The big cities are the good sample for analysing the countries.