Comparing international cities for their similarities and differences

Patrick Bosch

April 2020

1 Introduction

1.1 Background

Since centuries urbanization and concentration of industry around centralized locations led to the evolution of ever-growing cities. In combination with globalization, this led to large and international cities that have millions of inhabitants. Usually, countries and even cities have their specific properties that are different from other cities and countries. International cities have an increase in the diversity of inhabitants, mostly due to immigration and ex-pats, either temporarily or permanently. This change in population might change the properties of the cities. Around the world, many such large cities formed, and we want to compare them according to their specific properties.

1.2 Problem

In this report, we will focus on three such cities, more concretely, New York, London, and Paris. We want to answer the question, what the similarities and differences of these cities are. We explore them in terms of social and shopping places, as well as how neighborhoods are structured.

1.3 Interest

Two main groups are interested in this report. First, the people who plan to move to such a large city and need to make a choice which one might be the best fit for them. Second, companies who want to move a branch to such a city and want to make sure their workers feel comfortable in a similar environment as their original location.

2 Data acquisition and cleaning

2.1 Data sources

We use three data sources for our analysis. First, we use Wikipedia to get data for neighborhoods and their boroughs for each city. Usually, a city is partitioned into boroughs or districts which are further partitioned into neighborhoods. We are interested in the neighborhoods and their properties so that we can cluster and compare them later on. Second, we use location data from OpenStreetMaps and Google to get coordinates for each neighborhood. The location data for each neighborhood is necessary so that we can get additional information (venue data) for each neighborhood. Third, we use Foursquare to get information about venues for each neighborhood. For each neighborhood, we get types of venues, as well as how many of each type exist. The clustering algorithm will use these venues to decide which neighborhoods are similar to each other.

2.2 Data cleaning and processing

2.2.1 Neighborhood and location data

While Wikipedia provides borough and neighborhood data for each city, we need to clean and process it differently for each of them.

A table provides the data for London readily[1]. Still, it does provide information for the surrounding areas as well. Therefore, we need to first filter out the data according to the *post town*. After that, we can drop additional information such as post town, postcode district, the dial code, or the OS grid ref. We do need to clean the borough information, though, as it contains references and is not fully comma separated. Similarly, the neighborhood section also contains alternate names for some of the neighborhoods, which we remove so that getting location coordinates later on works properly. Additionally, we correct one neighborhood that is misspelled: Somerstown instead of Somers Town. We also detected that several pairs of neighborhoods have the same name but a different borough. We need to take this into account later on when we aggregate data for each neighborhood. OpenStreetMap does not always provide location data for a neighborhood, so we additionally use the Google API. We also make sure with a sanity check that all of the neighborhoods are within 30km of the city center with one exception that is 30.3 km away and is still accepted.

The data for New York is also available in a singular table[2]. Similar to London, it also has additional not needed information that we remove. Specifically, the area, population, and population per area are not of interest to us. The data is structured according to the community board with aggregated neighborhoods. So we split it into one neighborhood per row while removing information about community boards and keeping only the borough information. Similar to London, we use OpenStreetMaps and the Google API to get location coordinates

[1] `https://en.wikipedia.org/wiki/List_of_areas_of_London`

[2] `https://en.wikipedia.org/wiki/Neighborhoods_in_New_York_City`

and use a sanity check to confirm all coordinates are within range of the city center. Again, there is one exception with 30.4 km distance.

The data for Seoul is also available in one table[3], although one district is missing information about its neighborhoods. We need to merge that information from the district website into our table[4]. We remove the data about general information of a district, or points of interests so that we end up with neighborhoods and boroughs only. The neighborhoods are aggregated by district, so we split them into single rows. Again, OpenStreetMaps and Google were used to acquire location information, and we perform a sanity check to make sure the coordinates are close to the city center.

2.2.2 Venue data

We acquire the venue data for each neighborhood in each city through Foursquare. For each neighborhood, we request 200 venues in a radius of 750 m around the center of it, but the API seems to limit it to 100 venues. The data does not require further cleaning but requires preprocessing. We use one-hot encoding to mark each venue and then group them by neighborhood and borough and take the mean value to get weights for each category. Grouping by neighborhood and borough prevents the merging of neighborhoods with the same name. With this, the data is ready for further processing.

3 Exploratory data analysis

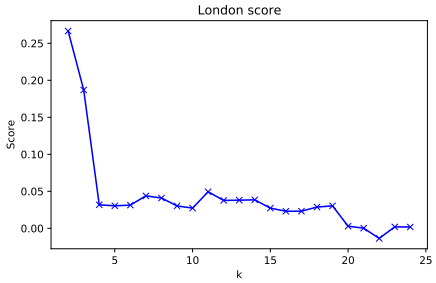
In this section, we first present our general methodology and also show initial analysis.

3.1 Methodology

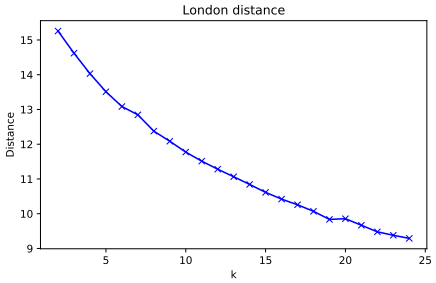
We base our methodology on the venue data that we have acquired from Foursquare. We do not use any other data, but only the amount of each venue type in a neighborhood. We use the weight of each venue type in each neighborhood to identify possible similarities between neighborhoods and also to rank the most common venues for each neighborhood.

For finding similarities between neighborhoods, we use k means clustering, which takes the weighted venues per neighborhood for each city as input. An important aspect of k means is the correct value for k. For this purpose, we use both the elbow and silhouette method to acquire a practical and optimal value for k. We used the Euclidean distance as a metric and chose a fixed random state so that our results are replicable. Choosing the correct value for k is not straightforward, though, as neither of the two methods in use gives a precise value for k. Figures 1, 2, and 3 show the different scores for London, New York, and Seoul, respectively. While two would be a good value in most cases, it is far

[3] `https://en.wikipedia.org/wiki/List_of_districts_of_Seoul`
[4] `https://en.wikipedia.org/wiki/Jung_District,_Seoul`

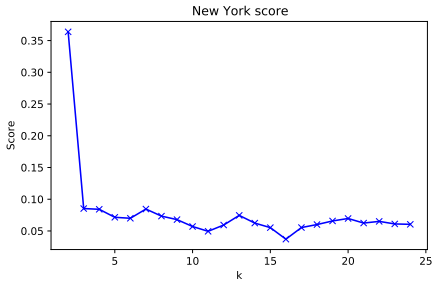


(a) Silhouette score.

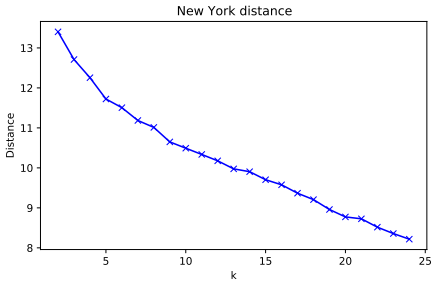


(b) Distance score.

Figure 1: Silhouette and distance score for London. While two is not a sensible choice, seven seems to be the most optimal besides two.



(a) Silhouette score.



(b) Distance score.

Figure 2: Silhouette and distance score for New York. While two is not a sensible choice, seven seems to be the most optimal besides two.

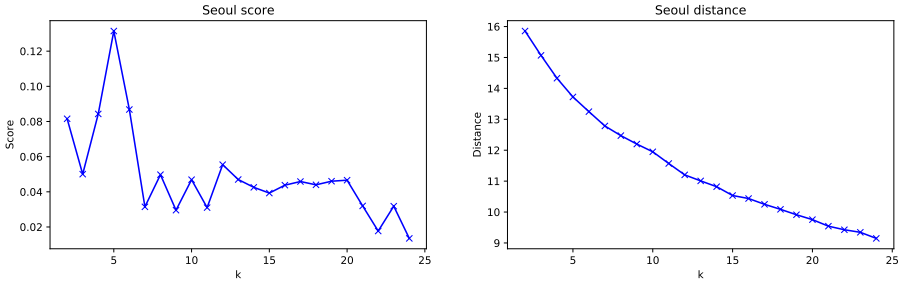
from practical as a lot of information would be lost. Cities usually have more than two types of neighborhoods. Looking further, we can see seven as a fitting value for London and New York, while five is a good value for Seoul. These values are more sensible, and we will use them in our further analysis.

Later on, we will use the most popular venues for each cluster of neighborhoods in combination with their locations to determine a category for each cluster.

For our initial analysis in the next section, we take the most common venue of each neighborhood and add them up to make an initial comparison between the cities. We do this exercise to find out if we can find initial similarities or differences.

3.2 Initial analysis

Our initial analysis of the most common venues in neighborhoods shows two expected and one unexpected result. London is famous for having pubs as a



(a) Silhouette score.

(b) Distance score.

Figure 3: Silhouette and distance score for Seoul. Five seems to be the most optimal choice.

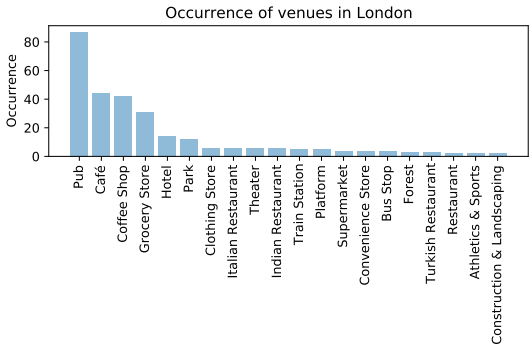


Figure 4: Occurrences of venues in London showing pubs as the most common venue in neighborhoods.

common sight and even attraction for tourists. And indeed, the data shows that the most common venue of most neighborhoods is pubs, as can be seen in Figure x. Cafes, coffee shops, and interestingly, grocery stores follow with some distance. Hotels and park are less common, but still high in the ranking.

New York also shows a more or less expected result, with pizza places being the most common venue of neighborhoods (Figure x). Coffee shops, delis and bodegas, and restaurants follow it. But also common enough are parks, bars, and golf courses, with the latter being comparatively large.

Slightly unexpected are the results for Seoul in Figure x. The most common venue is coffee shops, shortly followed by Korean restaurants. These restaurants likely also include small food to go shops, similar to delis in New York. Korea though, is quite famous for developing a large coffee culture, which puts the results into perspective[5]. Interesting though is the large difference to the next

[5] `https://en.wikipedia.org/wiki/Coffee_in_Korea`

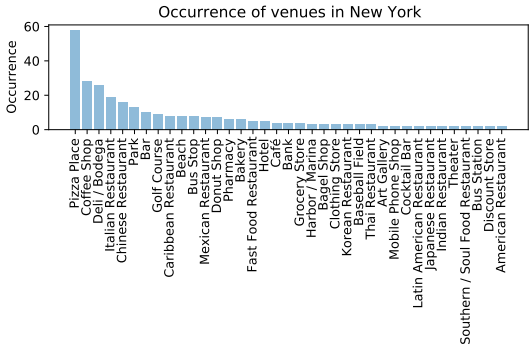


Figure 5: Occurrences of venues in New York showing pizza places as the most common venue in neighborhoods.

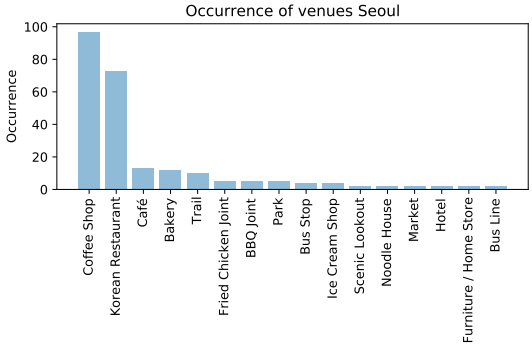


Figure 6: Occurrences of venues in Seoul showing coffee shops as the most common venue in neighborhoods.

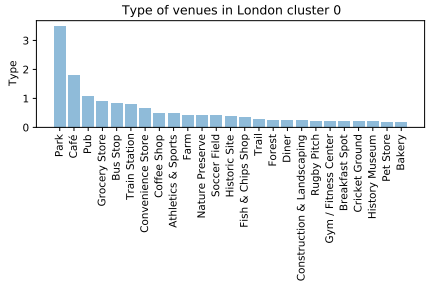
common venue. Both New York and London have a smoother slope.

4 Results

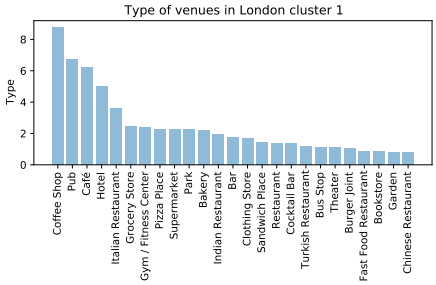
In this section, we present the results for the different cities.

4.1 London

The first cluster of London (Figure 7a) has as its most popular venues parks, followed by Cafes and pubs. This priority hints at neighborhoods with high recreational value. Still, the high occurrence of grocery stores hints more at residential areas. We can cross-check with the map in the annex, and indeed the neighborhoods are more outside of the city. We can, therefore, categorize it as recreational/residential.

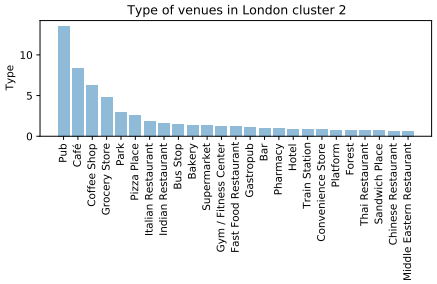


(a) First cluster.

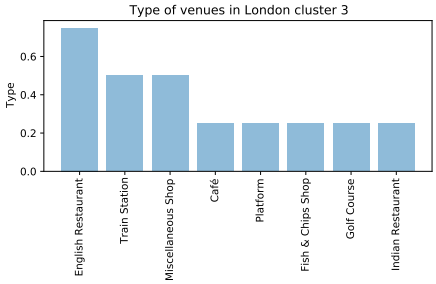


(b) Second cluster.

Figure 7: Types of venues in Londons first and second cluster.



(a) Third cluster.



(b) Fourth cluster.

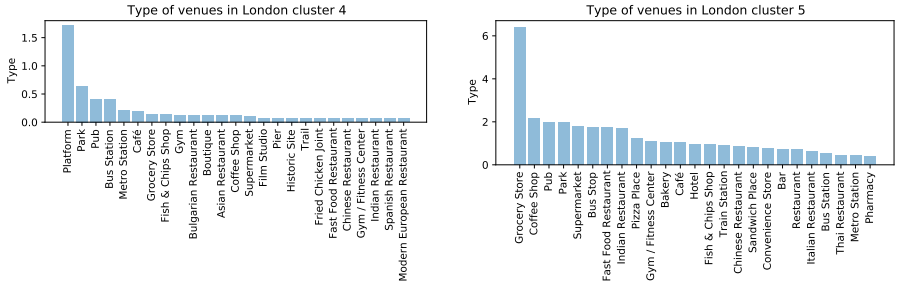
Figure 8: Types of venues in Londons third and fourth cluster.

The second cluster (Figure 7b) has a high amount of coffee shops, pubs, and cafes, shortly followed by hotels. This cluster seems to describe a touristic area, but not necessarily for sightseeing. If we compare it with the map in the annex, it includes the entire city center. We will categorize it as a tourist residence/city center.

The third cluster (Figure 8a) is not entirely clear. It consists mostly of pubs, cafes, coffee shops, and other restaurants, but it also contains grocery stores and supermarkets. This order seems partially residential, combined with a high amount of entertainment. The location of the neighborhood outside of the city center enforces that category.

The fourth cluster (Figure 8b) is a tiny one that is hard to categorize. It has English restaurants, a train station, but also a golf course. This may be an older residential area or a more wealthy one. It also lies reasonably far away from the city center. The main venue of the fifth one are platforms, which are the stations for the tram. These neighborhoods also lie outside of the city center and have parks and pubs. This cluster very likely contains purely residential areas.

The sixth cluster (Figure 9b) has a very high amount of grocery stores. It



(a) Fifth cluster.

(b) Sixth cluster.

Figure 9: Types of venues in Londons fifth and sixth cluster.

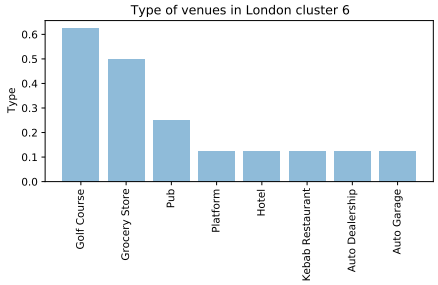


Figure 10: Types of venues in Londons seventh cluster.

seems to be residential areas more outside the city.

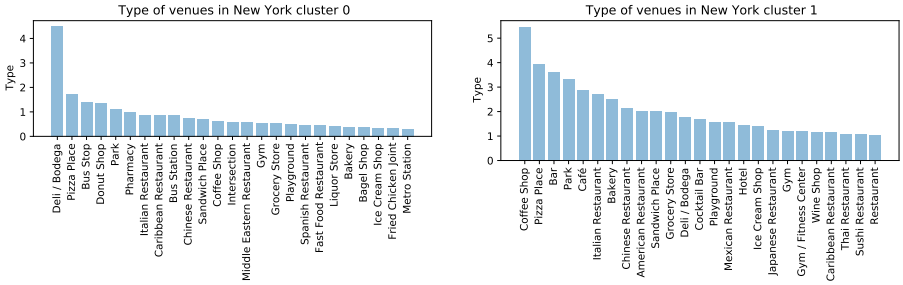
The seventh (Figure 10) cluster is again tiny and consists of two neighborhoods beside each other. The main defining venue is golf courses, which likely define the whole neighborhood and fit into a recreational area.

To summarize, London has many residential/recreational neighborhoods, defined by either grocery stores and supermarkets or parks and such. The cluster that encompasses the city center is different from that and hosts tourism and other activities. Interestingly enough, grocery shops can be found nearly everywhere. At the same time, entertainment and social activities are mainly focused on pubs and cafes/coffee shops.

4.2 New York

The first cluster of New York (Figure 11a) is dominated by Delis and Bodegas, which prepare food or are mostly grocery stores. While there are other venues, such as bus stops, parks, or pharmacies present as well, the vast majority is food preparation in this cluster. Therefore, we categorize it is recreational/residential/food area. These areas are also not part of the city center.

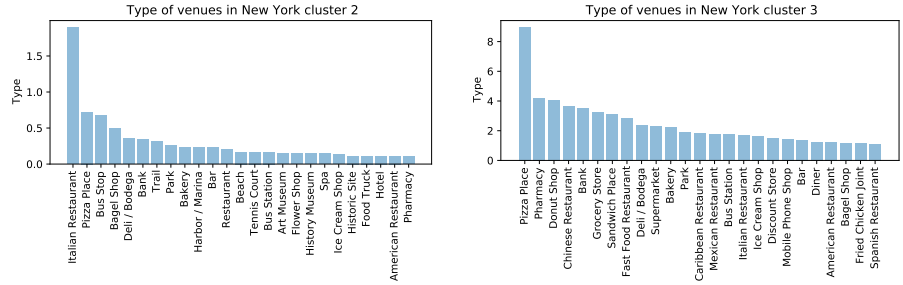
The second cluster (Figure 11b) is a mix of places, ranging from coffee shops



(a) First cluster.

(b) Second cluster.

Figure 11: Types of venues in New Yorks first and second cluster.



(a) Third cluster.

(b) Fourth cluster.

Figure 12: Types of venues in New Yorks third and fourth cluster.

to pizza places, bars, parks, and so on. Combined with the map in the annex, we can see that this encompasses the city center, especially most of Manhatten. A significant difference with the London city center is the relative absence of hotels in this area.

The third cluster (Figure 12a) is dominated by Italian restaurants and lies mostly at the edge of the city. These seem Italian dominated neighborhoods.

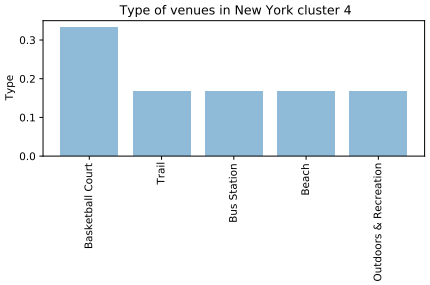
While the fourth cluster (Figure 12b) is dominated by pizza places, in the fourth position are banks. These neighborhoods stretch over a large part of New York and seem mostly residential areas. It is not clear why there are so many banks, though.

The fifth cluster (Figure 13a) is tiny and seems to be composed of pure recreational areas.

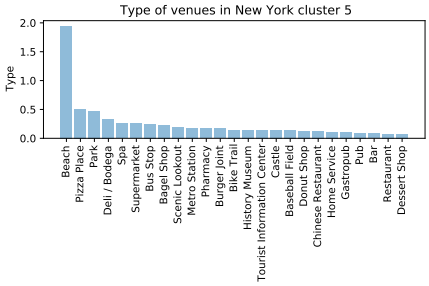
The sixth cluster (Figure 13b) encompasses all beaches and the surrounding neighborhoods and seems mostly recreational.

Similarly, the seventh cluster (Figure 14) seems recreational as well with golf courses, the harbor, and other activities. Notably, it is small but also contains tourist information.

To summarize, New Yorks clusters are fairly different from each other. There



(a) Fifth cluster.



(b) Sixth cluster.

Figure 13: Types of venues in New Yorks fifth and sixth cluster.

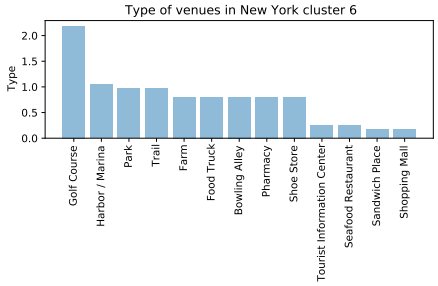


Figure 14: Types of venues in New Yorks seventh cluster.

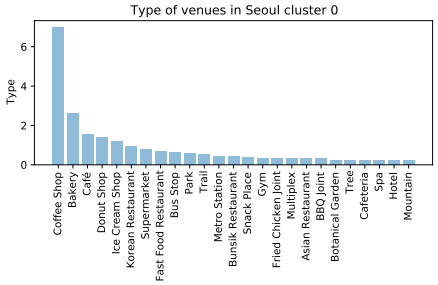
are mainly Italian neighborhoods, while grocery stores/bodegas are mainly found in a single cluster. There are clear recreational areas with beaches, but also a large number of residential areas. Tourism, while expected, seems to not manifest as strong with hotels not being a very defining feature. Social and recreational activities seem to include bars, beaches, but also food, such as pizza places.

4.3 Seoul

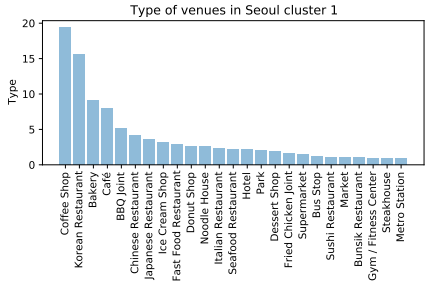
Coffee shops mainly dominate the first cluster of Seoul (Figure 15a). These neighborhoods are at the edge of the city and contain other culinary venues, but also some entertainment. These neighborhoods seem more residential and social.

The second cluster (Figure 15b) composes most of the city center. The most common venue is again coffee shops. Still, a significant amount of restaurants are present as well, of which many offer local food. While partially similar to the city center of New York, the presence of many hotels also shows similarities with London.

Korean restaurants entirely dominate the third cluster (Figure 16a). It cov-

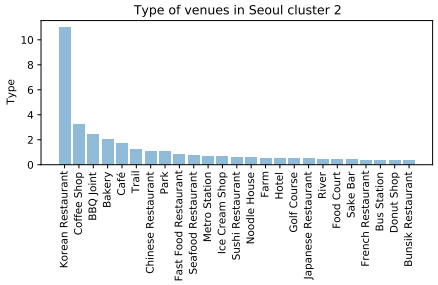


(a) First cluster.

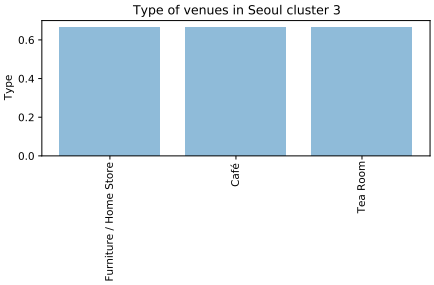


(b) Second cluster.

Figure 15: Types of venues in Seouls first and second cluster.



(a) Third cluster.



(b) Fourth cluster.

Figure 16: Types of venues in Seouls third and fourth cluster.

ers part of the city center, but neighborhoods can also be found outside. Many of these restaurants are likely similar to Delis in New York, which prepare and sell food to go.

The fourth cluster (Figure 16b) is the smallest and consists of one neighborhood. There is no precise categorization possible as it has a furniture shop, a cafe, and a tea room. Additionally, it is close to the airport.

The fifth cluster (Figure 17) is dominated by bakeries and parks, but also additional restaurants. These areas seem more residential and recreational.

To summarize, Seoul has a strong coffee environment; these shops dominate two entire clusters. Additionally, restaurants, especially local ones, are very common, which suggests that many people eat outside often instead of preparing food. The lack of grocery stores hints at that as well. Tourism seems to be a strong part as well, with hotels being a defining feature of the city center. While recreational venues being partially present, much of social life seems to revolve around restaurants.

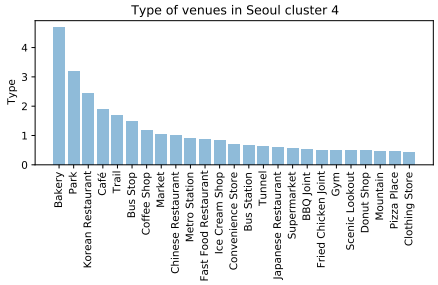


Figure 17: Types of venues in Seouls fifth cluster.

5 Discussion

We want to find the similarities and differences between the cities in this report. From the previous result section, we can see that both do exist. We will use the general cluster composition, but also differences in venue types.

First, we compare the city centers. The main similarity is the presence of a significant amount of venues, especially for food and drink. A common one in all is coffee shops. There are differences though as well. London tends to have more pubs, which we saw in our exploratory analysis. Similar New York has pizza places and bars, while Seoul has restaurants, mostly local ones. Hotels in the city center are typical in London and Seoul, but not New York.

Recreational areas differ between New York and the other cities most. The city of New York is the only one that has a beach, while people in New York and London seem to be fond of golf. Parks, on the other hand, are common in all three cities and are part of many neighborhoods.

All three cities use a similar mix of transportation. It consists of metros and buses and is most common in the inner city, but also found in the outer skirts of the city.

The most significant difference with the most impact on daily life is how residential areas are set up. While all cities have large areas of residential areas, the cities differ in what exists in residential areas. London, for example, has a large number of grocery stores and supermarkets, while having little restaurants. This distribution indicates that people prepare more food or at least eat more at home, with restaurants being less frequented. New York seems to have more places that also sell already prepared food, but as delis and bodegas are put together, it is difficult to say. While the first sell prepared food, the latter is more similar to a grocery store. But New York also seems more segregated as Italian neighborhoods are identifiable. Seoul differs significantly from the two other cities. There are many more restaurants, especially Korean ones, in residential areas, indicating two things. First, it is more common to eat prepared food, and second, eating outside is also a social event. Compared to New York and London, the number of restaurants compared to other activities is much higher.

To summarize, the core city and how things work there are similar, while how it materializes (pubs vs. pizza places) is different. Living outside the core differs though in terms of different social activities and behavior.

6 Conclusion

We compared the differences and similarities between three cities, London, New York, and Seoul. We used data of neighborhoods and venues to analyze and cluster neighborhoods together so that we can compare the cities. We found that while the core city is similar, they can differ in what type of venues are present. Living though can be very different in the cities, similar to recreational activities.

7 Future considerations

In future reports, we can include more information about the population, but can also consider the cultural background. Although the latter already expresses itself partially in the distribution of venues. Additionally, workplaces, their distribution, and locations can be considered as well.

Annex

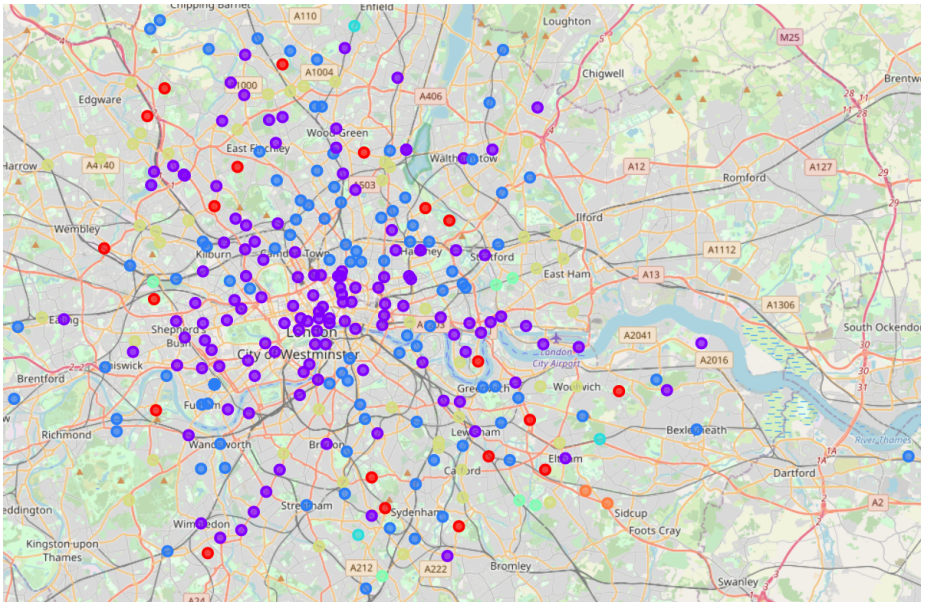


Figure 18: The map of London with each neighborhood in a cluster. Colors are as follows: 0 = red, 1 = purple, 2 = blue, 3 = turquoise, 4 = green, 5 = yellow, 6 = orange.

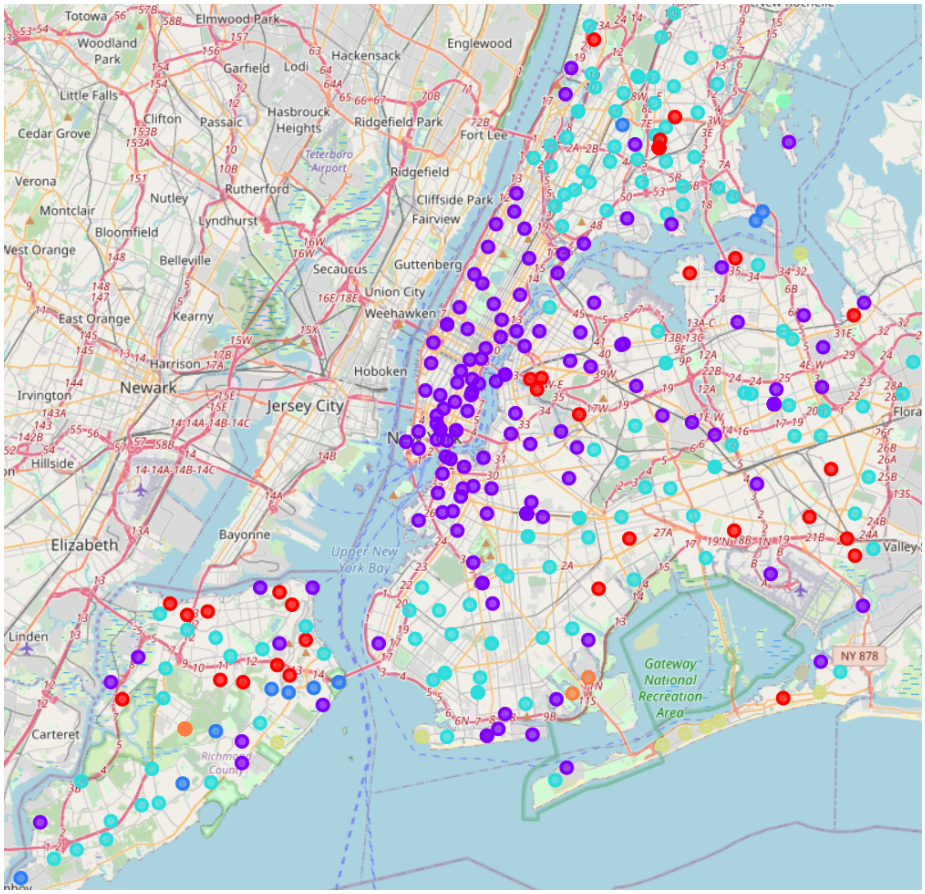


Figure 19: The map of New York with each neighborhood in a cluster. Colors are as follows: 0 = red, 1 = purple, 2 = blue, 3 = turquoise, 4 = green, 5 = yellow, 6 = orange.

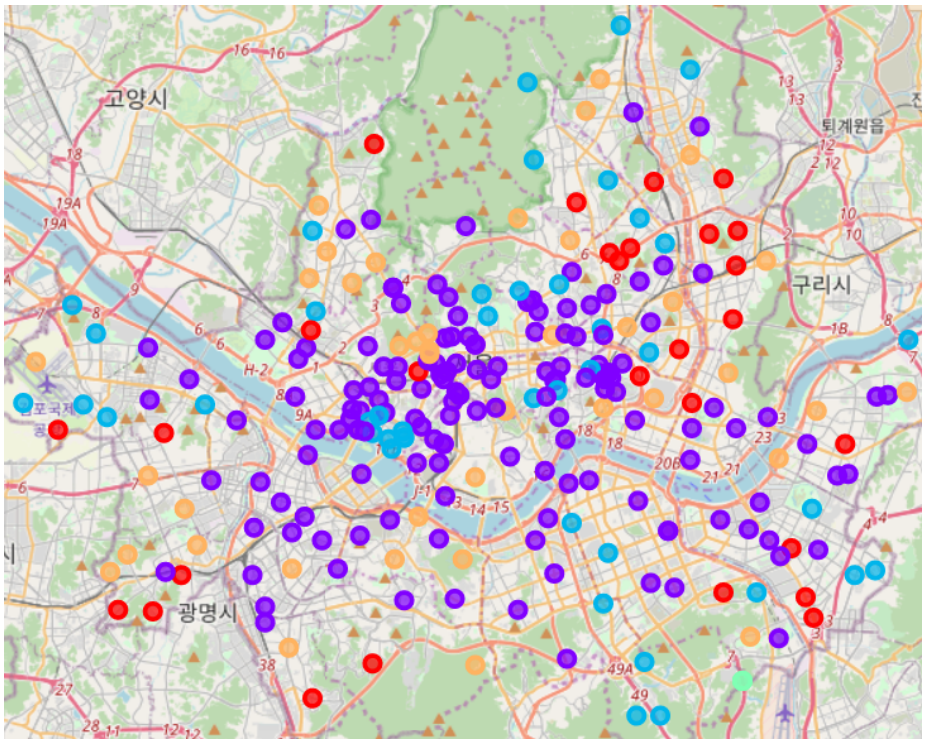


Figure 20: The map of Seoul with each neighborhood in a cluster. Colors are as follows: 0 = red, 1 = purple, 2 = blue, 3 = green, 4 = orange.