Coursera Applied Data Science Capstone

Xavier Legaz

September 2020

Contents

List of Figures

1 Introduction

Moving out is always a stressful situation, specially when it is to a different country. One of the major problems in this situation, is to chose in which neighborhood in an unknown city to live, given the scarce knowledge people have about the cultural and economical geography of foreign countries.

It is very common that the migrant leaves the city for reasons different than the city itself, being them work, family, etc., therefore, he/she wants to find a similar place to establish his/her new home. Or even if he/she chooses a new kind of neighborhood to live, it will be similar to one he/she is familiar with.

Of course, no two neighborhoods are equal, specially in different countries, but an array of parameters can be found to determine similarities and differences among them. Those parameters have to be taken from the major interests and necessities of the average migrant, and have to be comparable among the cultural differences and the other dissimilarities that exist between the cities.

The first parameter, which is imperative to include in any kind of comparison, and a necessity for all the interested parties, is the main economical one. A person can´t live in a place he/she can't afford, therefore the cost of rent is an obligatory first parameter. Given that the there is variance in this values, and having into account what could be found researching, the average rent per neighborhood is sufficient for this purpose.

Once all the neighborhoods that can be afforded are found, a more subjective matter arises, what kind of neighborhood the person likes. There is, of course, many ways to characterize a neighborhood, none complete, all with their pros and cons. For this project, considering the reach and objective of it, the selected method is based on the kind of venues that neighborhood has, and the proportion of them. This can give a general sense of the function or purpose the neighborhood has in that city, and the tastes and necessities of the people who live there.

For this particular project, only two-bedrooms apartments will be selected, for the sake of simplicity but can be extended to any kind of residence. For the same purpose and with the same disclaimer, only the neighborhood of Inner London will be compared.

The final objective of this project is, given a particular neighborhood, to provide a list of similar ones in a different city so the migrant can have a better understanding of the city and where he wishes or can afford to move to.

Of course, this information has to be presented also in a map, because the geographical location of the house depends on a wide range of reasons, like proximity to a place of work, hobby, etc, so the user can chose the one that is more convenient to him/her.

2 Data

2.1 Data Requirements

As stated, the data has to be from New York City, and Inner London.

Two sets of data will be needed for each city, one with the price of the average rent for two-bedroom apartments, and the other with the venues per neighborhood.

The data from both cities must be from comparable periods. Because of the relative small inflation in both countries, the rates can be from periods separated by more than a year, but a year will be used as the limit for precision sake. Given the tendency and speed of change of neighborhood character and identity, the data has to be no older than 2017, so the results are still valid at the moment, and can be considered contemporary between the different data sets.

The rates will all be presented in United States dollars, so the appropriate exchange rate will be used to convert the pound sterling. An average during the relevant period will be used to have a more accurate result.

The values will be rounded to have a more clean and clear presentation and can be easier to the user to interpret.

Since the interest is in the neighborhood as a cultural and economic center, the geographical specific limits are not important, since those things tend to change gradually across the borders, therefore the neighborhoods will be represented by a circle of 500 mts. of radius centered on the geographical center. The venues inside that circle will be used as representative of all the venues in the neighborhood.

To this end, a list of the geographical coordinates of all the centers of the neighborhoods have to be gathered.

2.2 Data Collection

The data from the venues will be gather from Foursquare because it provides a wide selection of venues, not only places of business, but also recreational non-profit areas like parks, to have a more complete understanding of the neighborhood.

The ten most common venues of each neighborhood will be used to compare them, since the intention is not to have a comprehensive list, but a general "feeling" of the character. Any venue with very few instances would be not only irrelevant, it can be prejudicial for this project, since it can give false positives. An example would be a pet store in a heavily business oriented neighborhood, or a single disco in a residential area.

The ten selected venues in each neighborhood will be inspected to determine if it's a valid instance. For example there had been cases of the actual neighborhood appearing as a venue.

For the data on the rent rates of London, the official information provided by the Government of the UK through the Valuation Office Agency will be used (Valuation Office Agency). A yearly inform is released with all the data needed. It is already presented in an Excel file table, it is free and reliable, therefore is the best option. Only the average rates of the neighborhoods will be used, discarding the rest of the data on the table. A simple multiplication of the values by the exchange rate for the period used, taken from Google, converts the rates to US dollars.

To find the coordinates of the neighborhoods, a dataset provided by the website Doogal.co.uk, which provides the postcodes of the UK is going to be used. The dataset is quite comprehensive, and it also includes the subdivitions of the districts, so multiple latitudes and longitudes are presented for each district. A simple average of all the coordinates gives an acceptable result, given that not much importance is given to the hard limits of the districts, as explained in the introduction.

In the case of New York city rent rates, no official records were found, therefore private information will have to be used. The New York real estate marketplace Street Easy provides downloadable monthly data in csv form about the rental and buy of apartments in the city in the period between from 2010 to July of 2020. Since our data from London is from the season 2018/2019, we will make an average of the the same period. As always, the data that is not useful to us, will be discarded (e.g., Borough).

For the coordinates of the Manhattan neighborhoods, a dataset from NYC Open Data, a public data source from the New York City Government, will be used. The latitude and longitud values are expressed in the same column, so they have to be separated to be used.

The common problem among both cities is that the datasets of coordinates and rent rates don't have the same neighborhoods or postocode districts in them. So, the decision was made to prioritize the location data, only using the neighborhoods in the location tables. That created the situation where a neighborhood can be chosen with no information about the rent rates.

3 Methodology

3.1 First Considerations

All the analysis work is going to be made using Pandas dataframes. Because all the data is obtained from *.csv* and *Excel* files, the Pandas tools for downloading them into dataframes are the most practical and easier options. All the data manipulation is going to be made using the diverse Pandas tools.

The Foursquare database has a limit of requests per day, for that reason, a maximum of venues can be consulted for each neighborhood, this can make that every run of the program, small differences in the values can appear. This being said, the general results shouldn't change in a significant way, so all the graphs and figures apprearing in this report are considered representative.

3.2 Exploratory Data Analysis

The first condition the model has to satisfy is that the rent rates have to be comparable. If the rent in New York is much more expensive than in London, then there is no reason to compare them and try to find similarities.

For this, a histogram is created with the rent rates per neighborhood, it has to be taken into consideration that in the dataset are significantly more London neighborhoods than New York ones, so the absolute values are not important.

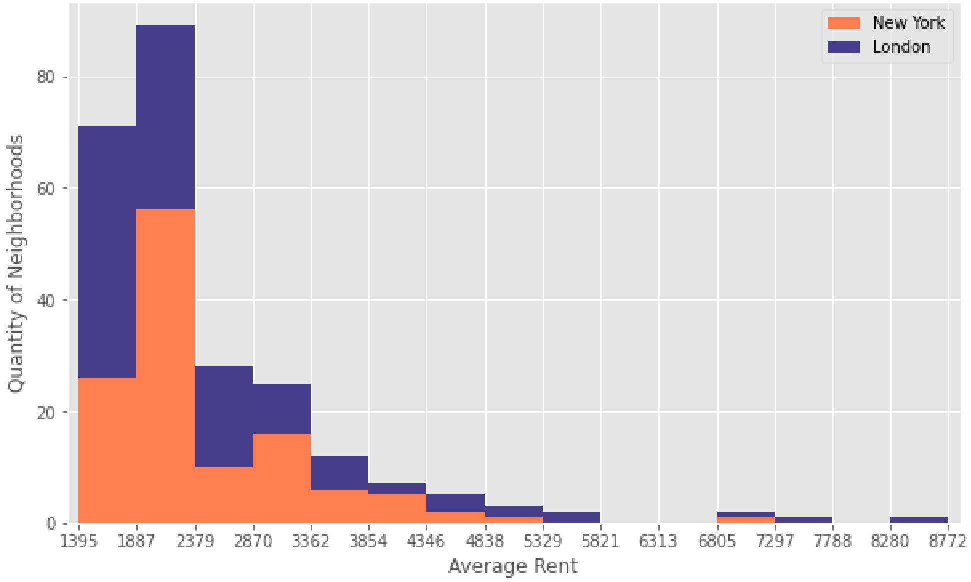

Figure 3.1: Histogram of Rent Averages for New York and Inner London per Neighborhood

Another way to present this information is by using a boxplot, which will show us the median, and lower and upper quartile. The outliers are not included, since are not important for the generality of the analysis.

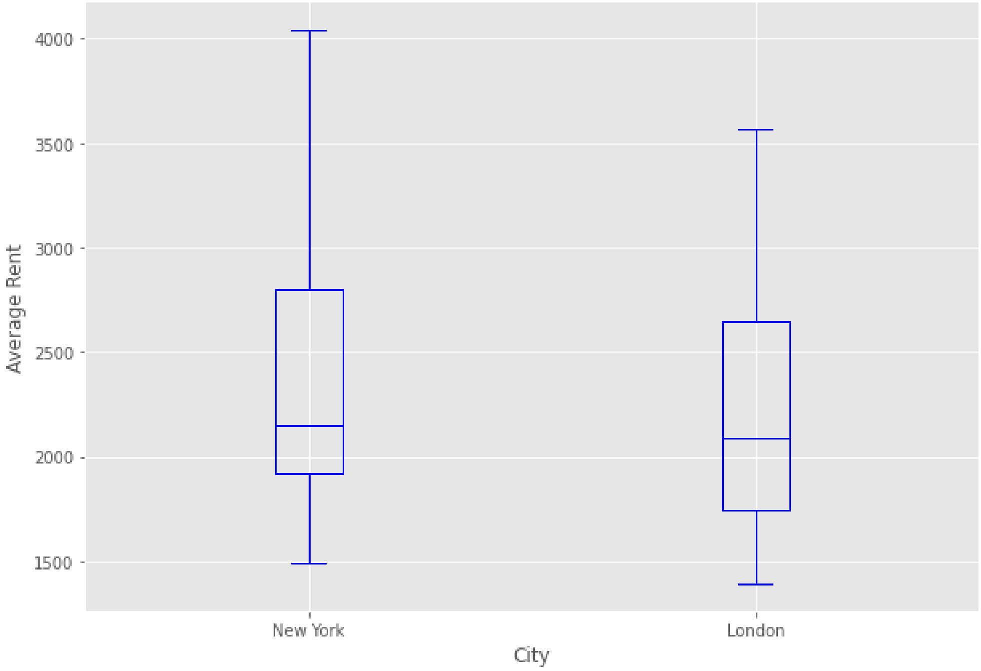

Figure 3.2: Boxplotof Rent Averages for New York and Inner London per Neighborhood

Is clear that, not only the rent rates are comparable, but they follow the same pattern. Meaning that both cities are very similar in what concerns to the rent market. That means that the results of our study are going to be valid and relevant.

Once we have all the data of all cities in one dataframe, a serie of explore requests are made to the Foursquare API, asking for all the venues 500 mts. around the coordinates of each neighborhood.

It has to be checked that the selection of a limit of 100 venues per neighborhood is justified, since it can be leaving out too many representative venues.

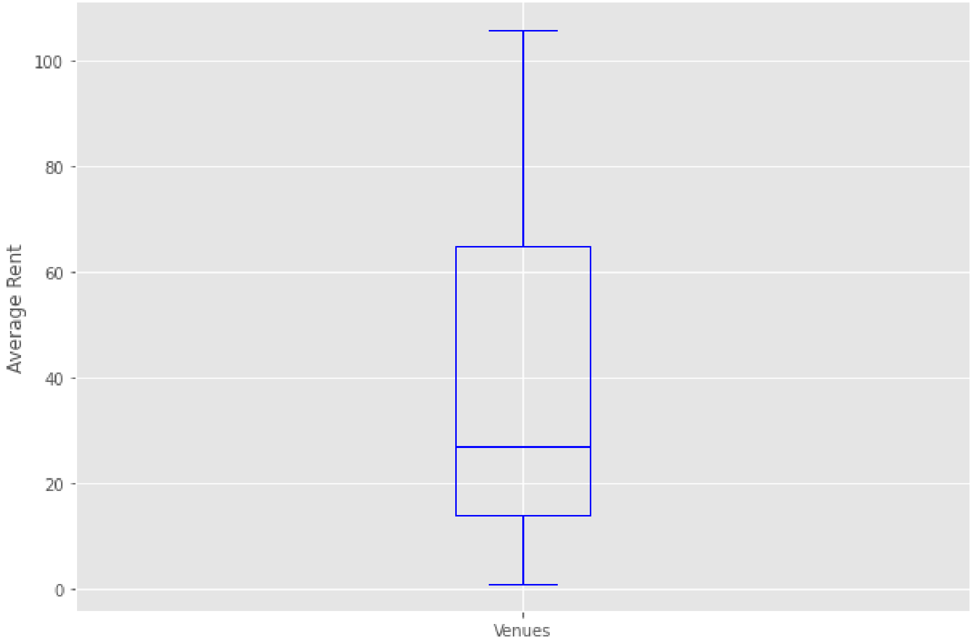

Figure 3.3: Boxplot of of Neighborhoods in New York and Inner London per quantity of Venues

The boxplot shows that most of the neighborhoods have less than 100 venues, therefore, the limit was correct.

Since the data about the venues categories is diverse and not quantitative, a qualitative approach was selected to analyse it. A word map can show what categories are more common on both cities, to understand if a comparison between them is not only possible, but sensible.

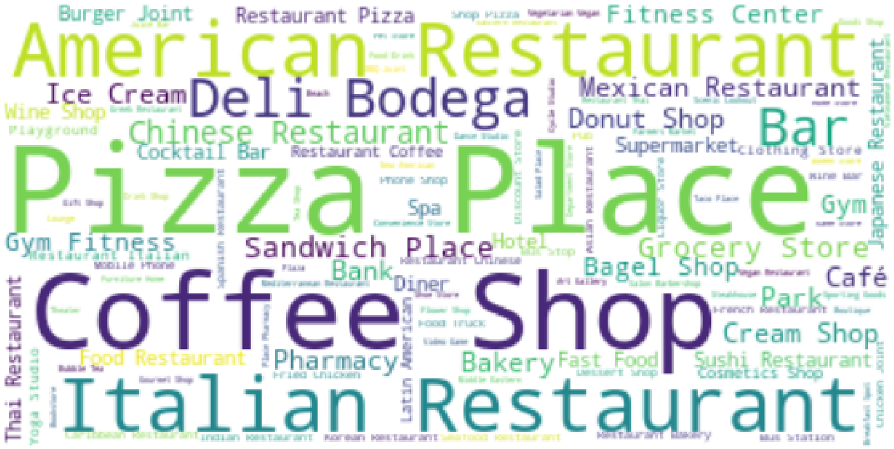

Figure 3.4: Word Map of Venue Categories in New York

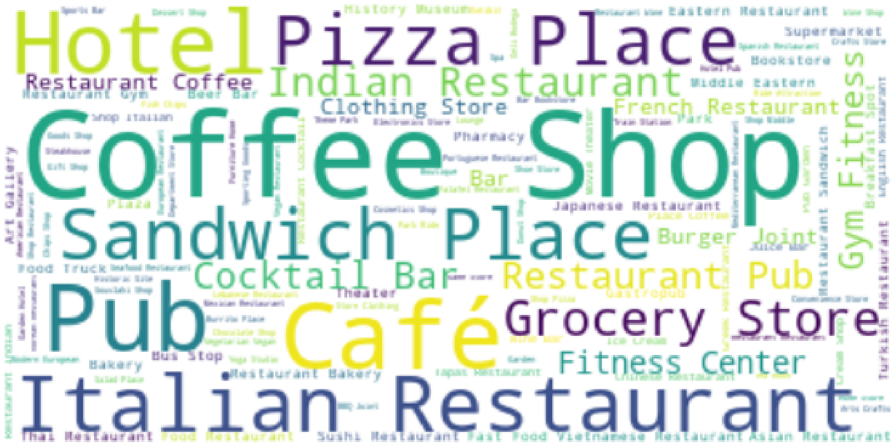

Figure 3.5: Word Map of Venue Categories in London

In both cities, categories like **Coffe Shop**, **Pizza Place**, **Italian Restaurant**, **Sandwich Place** are among the most common ones. Of course, every city has its own idiosyncrasy, therefore, it is expectable that **Pub** would be a very popular category in London, and **American Restaurant** in New York. But, it can be said that they are comparable in this aspect.

3.3 Procedure

Since the objective of this project is not recommending neighborhoods to the user based on his/her likes or dislikes, but to show him/her which ones are similar to another in other city, a recommendation algorithm is not going to be used.

A clustering algorithm is the option that returns the best results towards the intended goal, since it is intended to segmentate big amounts of data into different clusters, each element inside the same cluster is similar to each other, and different to other ones. That is exactly what the project is about, finding a cluster that contains the input neighborhood and at least another one of the other city, would mean those neighborhoods are similar, by definition.

Given not only the needs of the project, but also the graphical aspect of it (geographical identification of the clusters), k-means clustering proves to be the ideal choice. A map can be displayed showing each neighborhood with a different color, identifying each cluster.

If one cluster has only neighborhoods from one city, then if the input is one of this neighborhoods, no similar one from the other city would be found, and the output would be empty. To avoid this situation, every cluster should have at least one neighborhood from each city, so that is the parameter that has to be maximised.

For this, the algorithm has to be iterated for several amounts of clusters (k), and the one that has the best probability of finding a similar neighborhood from the other city, given a certain input, is chosen. Each cluster is inspected for composition, the ones that have neighborhoods from both cities have their quantity of elements added to an auxiliary variable. At the end, that variable is divided by the total of neighborhoods, and the probability is obtained.

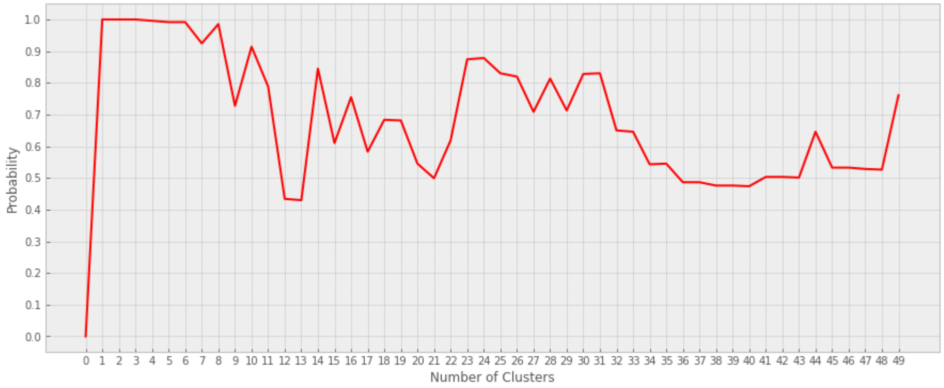

Figure 3.6: Probability of finding a similar neighborhood of different city for a number of clusters

When the number of clusters is very low, the probability is very high, but that provides very little information. It is easy to see this when a single cluster is taken, all the neighborhoods will be contained by it, therefore, the probability would be 1, but that obviously doesn't provides any information at all. So, the number of clusters have to be higher. It was decided that the number of clusters have to be larger than 20. The maximum, given this restrictions, is 24, which gives a probability of 87.84%.

Once the number of clusters is decided, the algorithm is applied and the neighborhoods are divided into each cluster. When the user inputs a neighborhood, a list of neighborhoods that belong to the same cluster, but different city is created. And the matter of rent rates appears.

Since the project is oriented at people who want to find a place to live, the rent rate has to be equal or lesser than the one in the input neighborhood. For that, the list of neighborhoods is sorted in an ascending order according to the rent average, and the input neighborhood is inserted in the place it corresponds. That way, the previous element is the closest neighborhood rent-wise.

Of course, if there are not any neighborhoods with a smaller rent, the value of the neighborhood with the next larger rent will be given.

4 Results

After applying the k-means clustering, according to what was discussed in the previous section, the program returned the list of neighborhoods separated in 24 clusters. The map of the neighborhoods was modified by adding a color corresponding to each cluster.

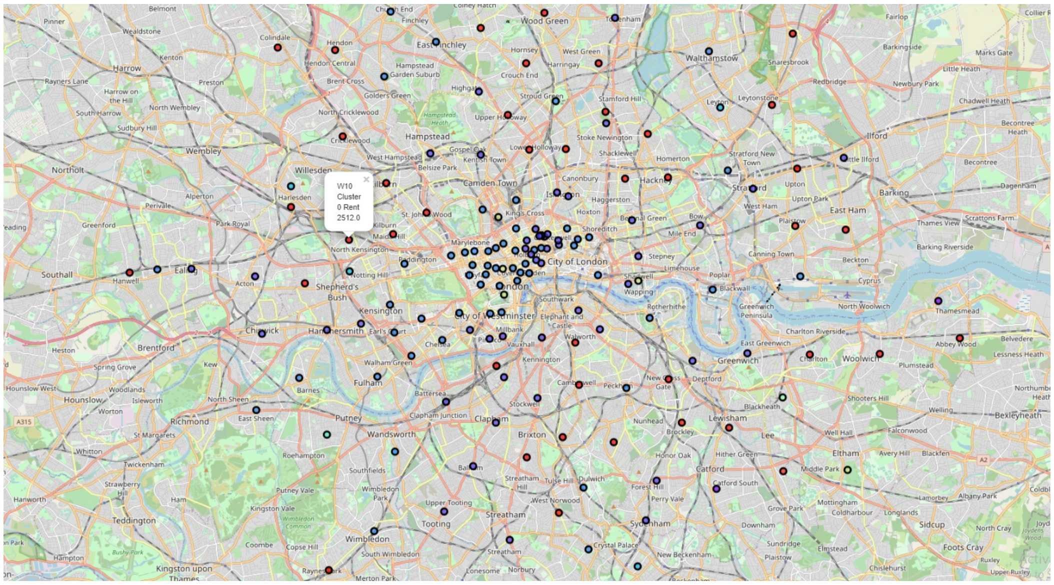

Figure 4.1: Map of Neighborhoods of London Separated in Clusters

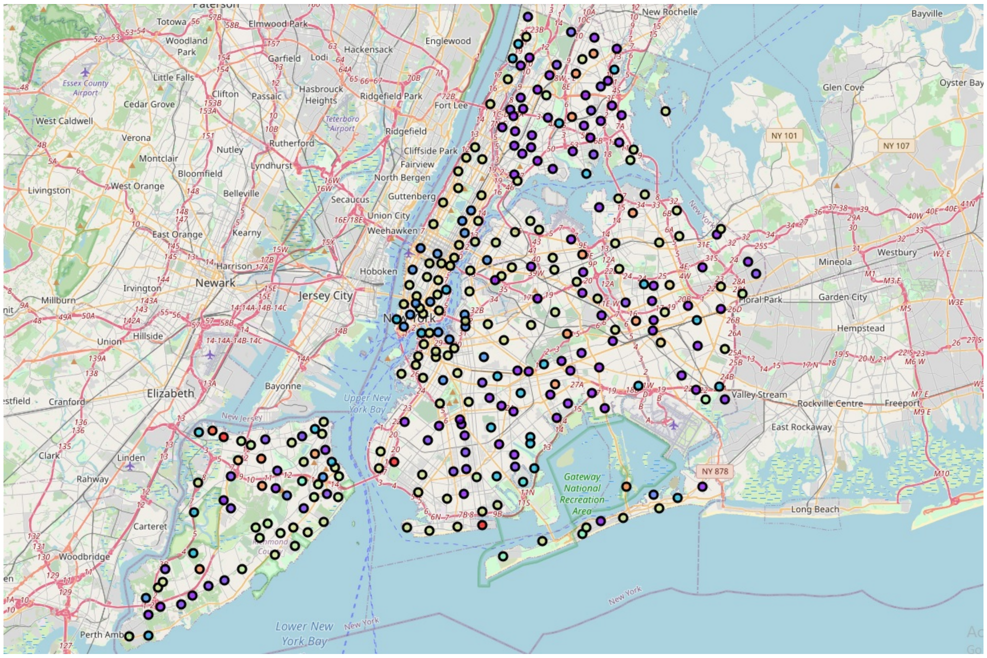

Figure 4.2: Map of Neighborhoods of New York Separated in Clusters

The probability of finding one neighborhood of the other city in one cluster was set to be 87.84%, so an easy way to see if this complies, is with a bar chart.

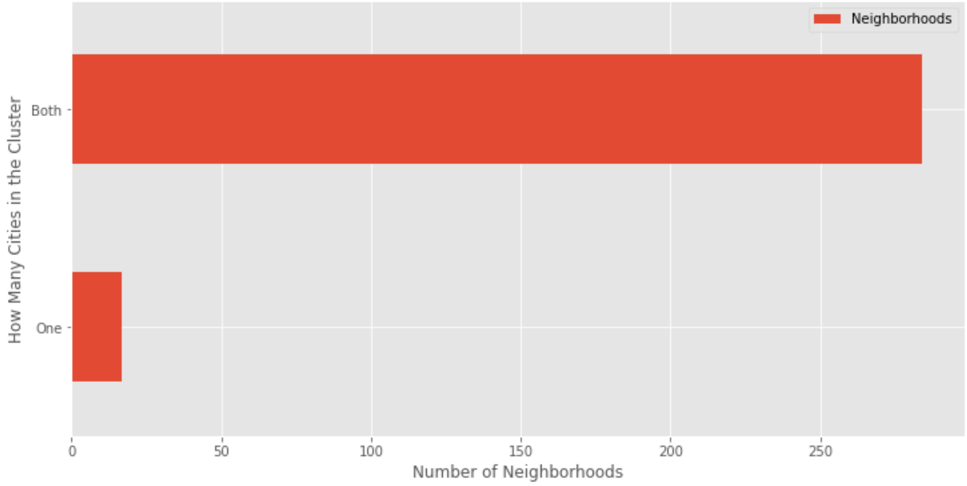

Figure 4.3: Neighborhoods in Clusters with one City and with Both

The result is clear, there are very few instances when a similar neighborhood can't be found in another city.

For example, when the neighborhood of New York named *Belmont*, with an average rent of $2.030, the result was the London neighborhood in the postcode district N21, with an average rent of $1.723, affordable for someone living in Belmont. The ten most common venues in each of those neighborhoods are:

	Neighborhood		
	Order	**Belmont**	**N21**
Venues	**1st**	Italian Restaurant	Italian Restaurant
	2nd	Pizza Place	Bar
	3rd	Deli / Bodega	Bus Stop
	4th	Bakery	Deli / Bodega
	5th	Dessert Shop	Pub
	6th	Donut Shop	Supermarket
	7th	Bank	Bistro
	8th	Food & Drink Shop	Middle Eastern Restaurant
	9th	Spanish Restaurant	Train Station
	10th	Café	Coffee Shop

Table 1: Ten Most Common Venues of the Similar Neighborhoods

The location of both neighborhoods can give information about its similarity.

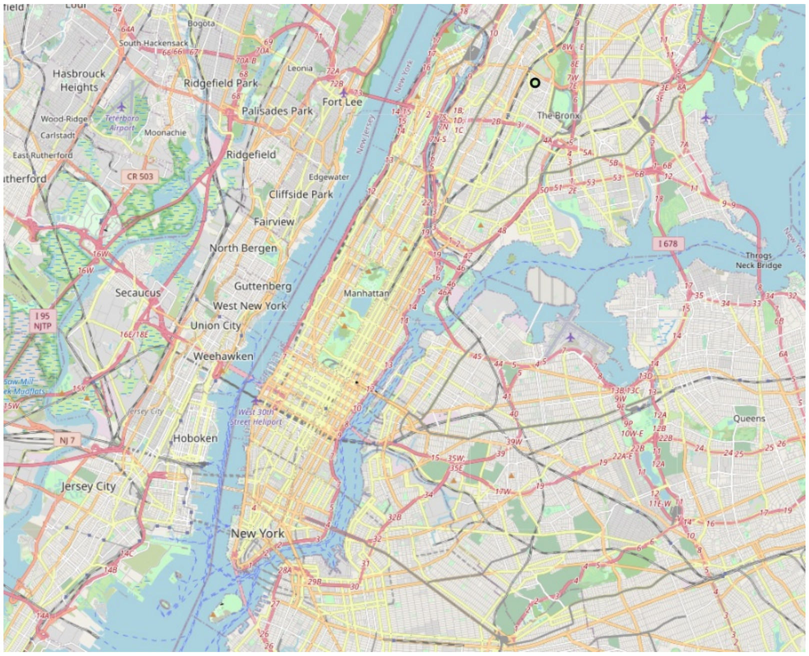

Figure 4.4: Location of Belmont, NYC

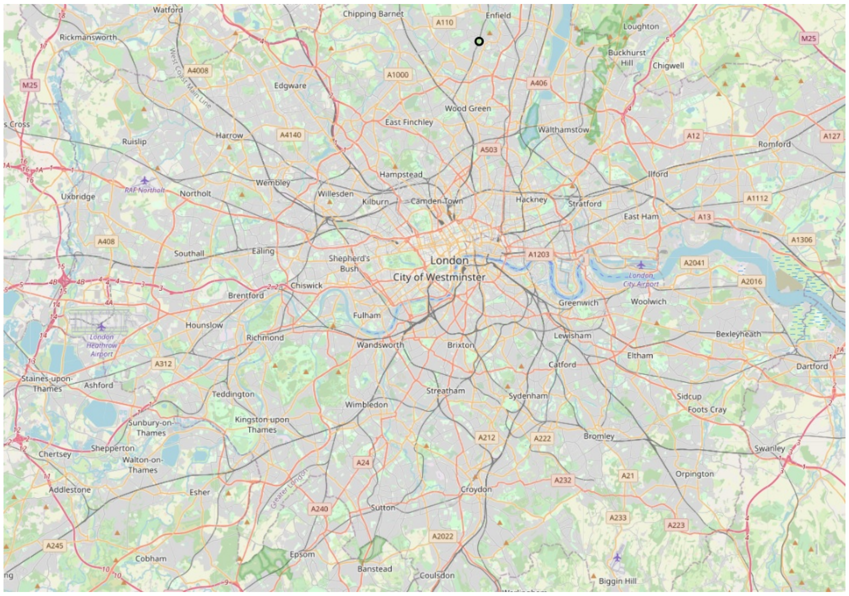

Figure 4.5: Location of N21, London

Both neighborhoods are far North from what is considered Downtown, near to the limits of their respective city.

5 Discussion

Inspecting the venues obtained in the previous example, shows that not only the most common venue is the same one, *Italian Restaurant*, but both places look very similar in the style of the venues they have. *Deli/Bodegas* in both, *Cafés* or *Coffee Shops*, foreign restaurants. It can be said that both neighborhoods have a similar "spirit". The geographical relative position of both neighborhoods in their respective city reinforces this idea.

There are some neighborhoods that could not be paired up, but that is expected, since two cities in different countries can never be exactly the same, the history of the places (e.g., immigration process) can change some neighborhoods in a great manner. But, the important thing is to keep those lonely neighborhoods to a minimum. In this case, about 12%.

Some neighborhoods are very close together, specially in the zones closer to the Downtown area of both cities, that resulted in many neighborhoods of those areas going to the same cluster. A more dynamic value of the radius that determines the neighborhood limits could be implemented to have better results.

Better results could be obtained using all the venues per neighborhood, without the limitation of 100, but it was shown that those were the minority of the cases, so this results can be considered as valid.

It is important to notice that the algorithm does not take as similar types of venues that could be considered the same, or very alike, like *Café* and *Coffee Shop*, or that some *Italian Restaurant* could be a *Pizza Place*. A better management of the database, with more comprehensive terms could get better and more accurate results.

Further considerations and a deeper study should be made before using the same tools and algorithms to compare two cities with bigger cultural differences, but that is outside the reach of this study.

6 Conclusion

The hypothesis in which this project grounds itself, is that we can find two similar neighborhoods in two different cities using machine learning tools to process the venues and the rent rates of all the neighborhoods in those cities. This has many uses, the one we prioritize is the real state possibilities.

The tests showed very promising results in regards of proving it. It was shown that most neighborhoods will be paired with at least one other neighborhood from the other city, and that they both have similar types of venues. Also, and this was not in the original intention, they also showed similar relative geographical positions in their cities. This new dimension clearly reinforces the idea that using venues is enough to obtain similarities.

This methods could be used in many other areas, besides real state market, and could help to have a better understanding of each city and find better solutions for different problems they could have.