FINAL REPORT COURSERA CAPSTONE ASSIGNMENT

Analysis on similarities and dissimilarities of two european cities: Madrid and Berlin

Gonzalo Herrero

INDEX

1. Introduction, business problem. First approach to the problem

1.1. How battle of the neighborhoods problem can be useful

This final project for the Applied Data Science Capstone course will be performing a study on different neighborhoods based on the "Battle of the neighborhoods" model. This model is useful to search for similarities and dissimilarities between several neighborhoods and analyze the parameters that make possible this studies.

Designed as a final project to unify and work on every concept studied in the previous courses, the choice of comparing how similar or dissimilar different neighborhoods are is perfect to:

- Practice the ways of searching for data in datasets on the internet
- Analyze the data using different techniques and tools
- Practice with the use of visualization tools
- Use every concept involving the data science evaluation process to extract information and conclusions about the subject studied, in this case the similarities and dissimilarities between two diferent neighborhoods and, with a wider point of view, the similarities and dissimilarities of major cities in different countries as well.

1.2. How to study the different components of the neighborhoods

To analyze the different aspects that make two multicultural, diverse and inclusive cities similar it is important to establish the main parameters this study will be based on. This is crucial since the different parameters will need different datasets and diverse approaches on how to search for data, collect data and, finally, analyze the data.

Every major city shares aspects of daily life of its citizens because of globalization, multiculturalism and population, but, since these cities are located, not only in different countries, but also in different continents and share different values and ways of life it is interesting to search for those things many cities around the globe have in common. With that in mind, this project will evaluate:

- Number of bars and restaurants
- Number of locals dedicated to night life
- Number of museums, theaters or other cultural spaces
- Number of parks and green areas
- Every other aspect considered important in the proceding of this study

2. Real examples used for this Project

For this type of projects it is important to use real life data and examples in order to use all the knowledge adquired on the previous courses for a real life based experiment. And to provide the data that will be analyzed the two cities whose neighborhoods will be compared are: Madrid and Berlin.

I have decided to compare these two cities for several reasons:

- Since they are two major global cities, it will be useful to analyze globalization and how it has affected different cities around the globe
- Being both cities european, it is also interesting to analyze how two different cities in Europe have changed throughout the years and how similar and dissimilar they have become in different aspects shared by all major cities

3. Conclusions on problem introduction and first approach

We have analyzed how we will proceed with the 'battle of the neighborhoods' assignment. But, although this is a mere students project, this kind of studies of different neighborhoods in a city are important to companies operating in those countries in order to perform market analysis and search for the best places to install their businesses.

These studies give a wide view on what are the preferences of the citizens living in the different cities and neighborhoods, what are their ways of buying, the kind of entertainment they look for and, in conclusion, everything that give the company the possibilities of obtaining benefits from the future customers.

Since this study gives a general view on how the different neighborhoods are structured, the target audiences of this study are very wide, including:

- Owners of restaurants of bars that may want to open new establishments and want to search for a location with a lower number of them
- Shops owners that may want to rent a local in a more central or shopping/tourist-likely parts of the city
- Investigators performing different social studies
- Even public institutions as local governments/city councils that want to improve services, improve the touristic appeal and make life more confortable for their citizens

In conclusion, this study will cover a wide range of aspects involving neighborhoods and cities and it is very useful also for a wide range of actors and targeted audiences.

4. Comparison on raw, general numbers

Before working with maps and the locations of all places analyzed, it has been considered important to study the raw, general numbers in order to have a first wide visión of the data scraped.

With the help of Wikipedia, I have obtained the data about the different neighborhoods and the geographical división of each city from the división in localities (Berlin) and districts (Madrid):

Districts Madrid:

Centro	*Chamartín*	*Moncloa-Aravaca*	*Puente de Vallecas*	*Villaverde*	*Barajas*
Arganzuela	*Tetuán*	*Latina*	*Moratalaz*	*Villa de Vallecas*	
Retiro	*Chamberí*	*Carabanchel*	*Ciudad Lineal*	*Vicálvaro*	
Salamanca	*Fuencarral-El Pardo*	*Usera*	*Hortaleza*	*San Blas-Canillejas*	

Localities Berlin:

Mitte	*Spandau*	*Treptow-Köpenick*
Friedrichshain-Kreuzberg	*Steglitz-Zehlendorf*	*Marzahn-Hellersdorf*
Pankow	*Tempelhof-Schöneberg*	*Lichtenberg*
Charlottenburg-Wilmersdorf	*Neukölln*	*Reinickendorf*

Using the major divisions of Madrid and Berlin, it has been posible to gather information about a more precise división in the neighborhoods that will be used to analyze the data and conclude the study.

Having gathered the administrative divisions of each city, with the help of Foursquare, it has been posible to obtain the geographical position of each neighborhood. With the geolocator API, I have obtained the coordinates, in terms of latitude and longitude, of each neighborhood.

Having obtained the coordinates of each neighborhood, with the use of Foursquare, I have been able to gathered a quite complete table of data about the different venues in a city (in each neighborhood) to, afterwards, compare both cities.

	Neighborhood	Latitude	Longitude
0	Adlershof	52.437790	13.54778
1	Afrikanisches Viertel	52.558269	13.33389
2	Alt-Hohenschönhausen	52.547060	13.50055
3	Alt-Treptow	52.493500	13.45711
4	Altglienicke	52.420060	13.53969

	Neighborhood	Latitude	Longitude
0	Arganzuela	40.40021	-3.69618
1	Barajas	40.49181	-3.56948
2	Carabanchel	40.39094	-3.72420
3	Centro	41.62812	-4.72705
4	Chamartín	40.45000	-3.70000

Latitude and longitude of neighborhoods in Berlin and Madrid

With these complete dataset, it is now posible to perform a first analysis on the number of venues, of significante places in a city, and compare them individually and generally.

Firstly, we shall analyze the main aspects that every major city has:

- Restaurants:

  Analyzing the different types and number of restaurants from the data gathered from Foursquare, I have compared the numbers and percentages of this type of venue in Madrid and Berlin.

  The number of restaurants in the administrative region of Madrid is a 40.54 % higher than the numbers of Berlin.

- Cultural spaces:

  The data studied in Foursquare shows a bigger concentration of cultural spaces of all types (theaters, museums, open-air expos, etc.) in the city of Madrid that in Berlin. Specifically, 52.53 % more cultural venues in Madrid.

- Hotels:

  In relation to hotels, hostels, motels and every other building of this type, Berlin shows a higher number of them tan Madrid.

  With a 30.31 % more of hotels, Berlin situates itself as an easier place to establish a hotel or to search for one in a tourist or work travel destination.

- Green spots:

  The number of green spots comparison between the two cities shows the biggest different in every aspecto and venue considered.

  In this case, Berlin wins widely with up to a 181.65 % difference between green places in the city (parks, rivers, forests, mountains, etc.).

- Shops:

  The number of shops in both cities shows a difference of a 68 % higher number in Berlin than in Madrid considering all available venues in Foursquare.

There are many more aspects that can be treated and studied to gain a wider view of the similarities and dissimilarities between these two cities, but, since these aspects are the most common ones and can be found in almost every major city in the world, it seems unnecessary in this study to elaborate a more complex analysis (if we were to make a report, a study for a big company, with, not only more time, but also more resources, to study the habits of its citizens, every single aspecto should be covered, including: Banks, companies, specific shops and stores, parkings, pedestrian streets, etc.)

When studying these concepts, it should also be mentioned the área covered by both cities. With 606 and 892 squared kilometers in Madrid and Berlin, shows a bigger concentration of number of venues in Madrid.

Another aspecto to be considered when analyzing this two cities is the difference between neighborhoods and the concentration of venues per each one. For instance, if most of the venues are concentrated in a reduced number of neighborhoods (as it seems to happen more in Berlin), this general analysis on number of venues in the cities is insufficient.

To solve this problema, it also has been studied the number of venues per neighborhood also with the help of pint maps to see the distance and concentration of these pints (number of venues) in each neighborhood in Madrid and Berlin.

5. Maps and other helpful tools

5.1. Importance of clustering data and finding relevant data individually

On the previous point, we have made clear how a wider view of the data gathered can be helpful as a first approach to analyzing the data scraped from Wikipedia and Foursquare. But, since we pretend to find similarities and dissimilarities between two cities, we Will need a more thorough investigation.

Specifically, we will need to find more specific information on neighborhoods in an individual way. Every neighborhood has its own specificities and, even though we are working with two large cities as a "one", we need to point out some things.

To search for these similarities between neighborhoods in a city, I have separated the different venues (restaurants, shops, museums, etc.) for each neighborhood to compare them.

With these comparisons and maps to show in a more clear way the location of the different neighborhoods and their distances from the city center, we will be able to conclude in a more complete way the study.

5.2. Differences between neighborhoods in the city of Madrid

The city of Madrid covers around 606 km2 of área and, as almost every city in the world, it has been expanding from the city center to outwards.

This way, the most highly populated neighborhoods are those in the city center, whereas those outside the city are mainly residential.

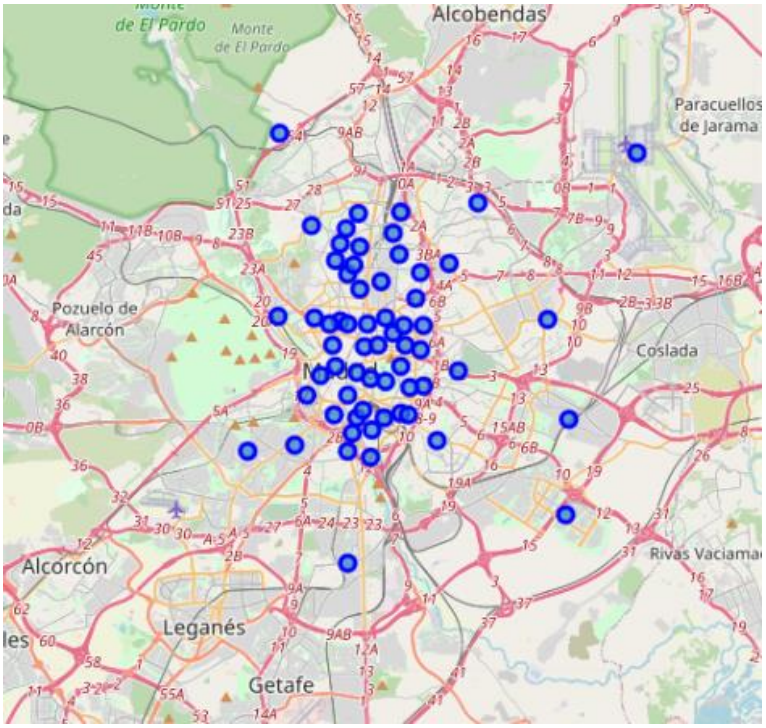

Image of the neighborhoods from which I gathered the number of venues in Madrid

The map displayed above shows how clearly there are neighborhoods at an important distance from the center of Madrid and how most of the neighborhoods are at a close distance from the city centre. This is mainly due to a higher population of the most centrical áreas.

Performing data analysis on the important venues and their numbers for each neighborhood, I have obtained the following data:

The data gathered shows perfectly what it has been discused about the distance from the city center and the differences between neighborhoods.

For example, studying the data about the restaurants in the city, the data shows that neighborhoods such as 'Centro', 'Retiro' (right in the city centre) or 'Latina' (a quite touristic location) show a higher concentration of restaurants than other neighborhoods and clearly higher than the mean.

Other neighborhoods such as 'Carabanchel', 'Moratalaz', 'Villa de Vallecas' or 'Vicálvaro', eventhough they are not as centrical or as touristic as the others mentioned show a high concentration of restaurants. This can be perfectly explained as they are neighborhoods a bit separated from the city center and, because of that, créate neighborhoods which are kind of little cities inside the main city with their own concentration of restaurants, shops, etc.

In relation to other neighborhoods, the lower density of population and the fact that many of the neighborhoods are more residential than those closer to the city center.

5.3. Differences between neighborhoods in the city of Berlin

As the study conducted individually to the neighborhoods in Madrid, I have studied those neighborhoods in Berlin.

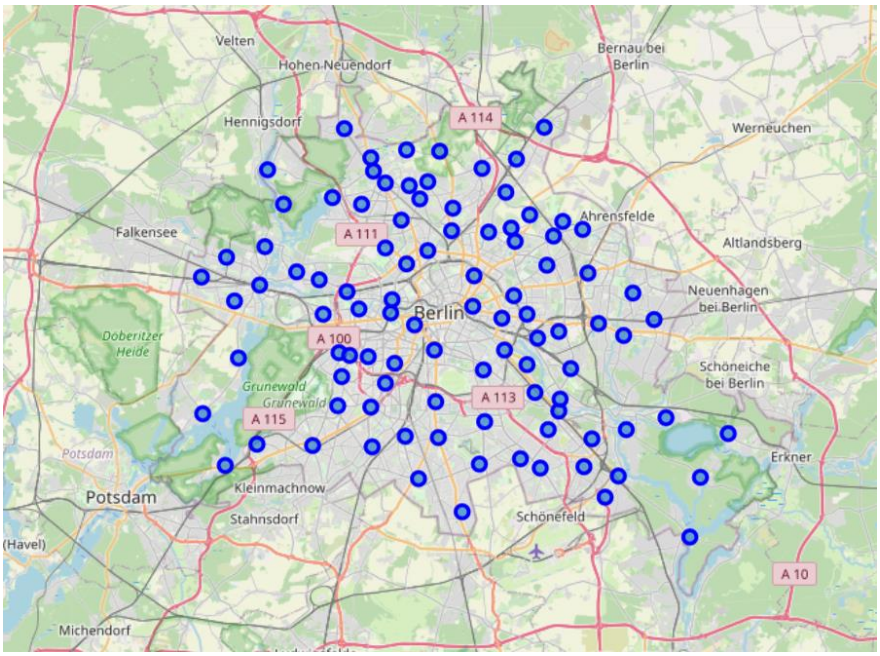

Image of neighborhoods from which I gathered the number of venues in Berlin

Analyzing, in the same way as done before, the boroughs of Berlin and the number of venues obtained from Foursquare marked for Berlin:

- Restaurants:

  In this category there are five boroughs that have the highest concentrations on restaurants in whole Berlin: 'Charlottenburg-Wilmersdorf', 'Friedrichshain-Kreuzberg', 'Steglitz-Zehlendorf', 'Treptow-Köpenick' and 'Lichtenberg'.

- Cultural spaces:

  On the number of cultural spaces the boroughs of 'Charlottenburg-Wilmersdorf', 'Mitte', 'Marzahn-Hellersdorf' and 'Lichtenberg' raise to the top positions on the number of this type of venues.

- Hotels:

  On hostels, motels, hostals, etc. the top boroughs of Berlin are: 'Charlottenburg-Wilmersdorf', 'Mitte' and 'Marzahn-Hellersdorf'.

Image of the localities/boroughs of Berlin

- Green spots:

  As mentioned previously, Berlin almost doubles the number of green spots venues according to Foursquare to the ones in Madrid. And the boroughs that complete the top positions regarding these venues are: 'Treptow-Köpenick', 'Steglitz-Zehlendorf', 'Mitte' and 'Pankow'.

Regarding to the different commerces, shops, stores and cultural spaces within the city of Berlin, the boroughs that place a higher concentration on these venues are those in Central and Northern Berlin as shown in the map showing the different boroughs.

The green spots venues are mainly located North and South of Berlin.

5.4. Conclusions on maps analysis

These two map analysis on Madrid and Berlin show a greater similarity than the study performed with the raw numbers preiously.

For instance,distance from the cities centers is an important matter to have in mind when studying the neighborhoods of different cities. This way, those neighborhoods closer to the city center usually have more restaurants, shops, stores, hotel, etc. whereas more outside neighborhoods have a lower concentration on these types of venues (opposition between more touristic and financially important boroughs and more residential ones).

Also, the further we stray from the centre, the more green spots and places we can find and the closer the cities neighborhoods are to nature.

6. Conclusion of the Final Project Report

This study, this Final Project for the Data Science IBM course has been useful to apply many of the things learned during the following courses. It has been helpful to use every tool learned to search for data (data scraping with Foursquare), apply different visualization tools and, finally, analyze all the data gathered.

The main goal of this study was to compare two european cities, Madrid and Berlin, and find the similarities and dissimilarities these two examples of major cities in Europe have.

6.1. General numbers

With the connection of almost every part of the globe and the globalization as a new global phenomenon that affects every big city in Europe, it has been interesting to face, in a general and wide way, and compare the global number of venues.

Restaurants	The city of Madrid has a higher 40.54 % of restaurants concentration in the average of all neighbourhoods studied than those in the city of Berlin
Cultural sites	As in the number of restaurants, in Madrid the cultural sites represent a higher percentage over the total sum of number of venues studied with Foursquare. This way, Madrid has a higher 52.5 % of cultural sites proportion overall than Berlin
Hotels	A good parameter to compare two cities can also be the proportion of hotels over the total number of venues. In this case Berlin wins by a 30.31 %
Green spaces	Studying green spaces both cities have, Berlin clearly outputs Madrid with almost a 182 % higher concentration of Parks, forests, rivers, canals, etc. than Madrid
Shops and stores	Berlins also outputs Madrid when talking about shops, stores and other locals to buy products with a 68 % higher proportion of these types of venues

The data gathered in the table above shows the proportion of each type of venue over the total number of venues. Because of this, we are able to work and analyze the data in a more individual way, not only for each neighborhood, but also for each type of venue obtained from Foursquare.

Even though it does not seem to be many similarities between both cities, the biggest similarity we can obtain from this data is how big cities, although they vary from each other in some

aspects, have grown to offer every possible aspect to cover their citizens need and how almost every venue that figures in Foursquare is covered and available in Madrid and Berlin.

6.2. Similarities and dissimilarities through maps data

Working with the data that has enabled me to build the maps used for this study and watching the administrative divisions in districts (Madrid) and boroughs (Berlin) one aspecto has drawn my attention: correlation between distance from the city centre and number of venues.

In Madrid, the most populated districts, located in the city centre, have the higher ratio of number of venues proportion. And in Berlin, the proportion of number of venues increases in the localities located Central and North of Berlin.

This distance factor also affects the green zones in both cities. Most neighborhoods that have a higher proportion of 'green venues' are a bit outside Berlin and Madrid and are clearly less populated than the others.

6.3. Final statements

It is true that Madrid and Berlin have their differences, something normal considering they are two cities in different countries with a different history and population. These differences are reflected mainly on the different proportion of venues per neighborhood and in total. But, as european cities also share several similarities worth pointing out.

Both cities share a particularly interesting relation between more crowded, centrical and touristic neighborhoods and a wider proportion of comercial locals and a clear difference between the proportion of specific venues on these neighborhoods and more residential or low populated ones.

But there is a global phenomenon affecting in many positive ways big cities over the world and, being both cities european, obviously Madrid and Berlin. The phenomenon of globalization puts together people, traditions, cultures all over the world, but also companies that find the global connection as a Brand new opportunity to reach more people.

Globalization has changed every country, in many ways as a positive matter, and this clearly affects Madrid and Berlin, where the number of venues studied is enormous, including restaurants of all types and countries, companies from all over the world or cultural and artistic spaces to bring together diverse people.

All of these aspects make this study an interesting project for me (and in a more complete, studied project for a big company to develop) but also for any company trying to sell new goods in a different city in a different country, for anyone deciding to move to another country, to any investigation project developed based on the citizens, etc.

To sum up, and interesting opportunity to use great tools as Foursquare and the libraries implemented in Python and continue the path of Data Analysis started in this course.