Battle of the neighborhoods

Berlin vs Hamburg

Alexandros Chrysidis

15/05/2021

1. Introduction

In this project we will try to find similarities between the cities of Berlin and Hamburg, Germany. Specifically, this report will be targeted to:

- Citizens from either city, interested in moving to the other.
- Companies interested in relocating or expanding in either city.

Since there are lots of venues in Berlin and Hamburg, we will try to detect neighbourhoods that are similar to each other, by analysing the most common venues.

We will use our data science powers to generate a few most promising neighbourhoods based on this criterion. Advantages of each area will then be clearly expressed so that best possible final neighbourhood can be chosen by the interested parties.

2. Data

2.1 Data source

For this analysis we will need 2 types of data. The postal codes and coordinates of the cities, which can be found on GitHub, here. Moreover, we are going to need venue location and information, in order to compare the neighborhoods. This is achieved by scraping data from the Foursquare API.

2.2 Data cleaning

First step was to clean the ZIP-Code data obtained, since the original document had multiple columns of irrelevant data and many major cities which were not needed. After cleaning the original dataset, we were left with 296 rows and 4 columns of data for the cities of Berlin and Hamburg.

Next, the venues were scrapped, normalized, and put into a dataframe in order to prepare them for further analysis.

In order to achieve that, we had to scrap venues for each area code of each city, clean the data and then merge into one table as shown below.

	Zipcode	State	Zipcode Latitude	Zipcode Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	10115	Berlin	52.5323	13.3846	Hotel i31	52.531107	13.384270	Hotel
1	10115	Berlin	52.5323	13.3846	Hotel ULTRA Concept Store	52.529362	13.396969	Furniture / Home Store
2	10115	Berlin	52.5323	13.3846	Deutsches Theater	52.523970	13.382001	Theater
3	10115	Berlin	52.5323	13.3846	Ackerstadtpalast	52.529721	13.396777	Performing Arts Venue
4	10115	Berlin	52.5323	13.3846	Sammlung Boros	52.523352	13.384213	Art Gallery
5	10115	Berlin	52.5323	13.3846	Factory Kitchen	52.537449	13.394714	Restaurant
6	10115	Berlin	52.5323	13.3846	Hamburger Bahnhof – Museum für Gegenwart	52.528513	13.372067	Art Museum
7	10115	Berlin	52.5323	13.3846	Du Bonheur	52.536310	13.397558	Pastry Shop
8	10115	Berlin	52.5323	13.3846	Bandol sur Mer	52.528992	13.395436	French Restaurant
9	10115	Berlin	52.5323	13.3846	H Gedenkstätte Berliner Mauer	52.535750	13.390708	Tram Station

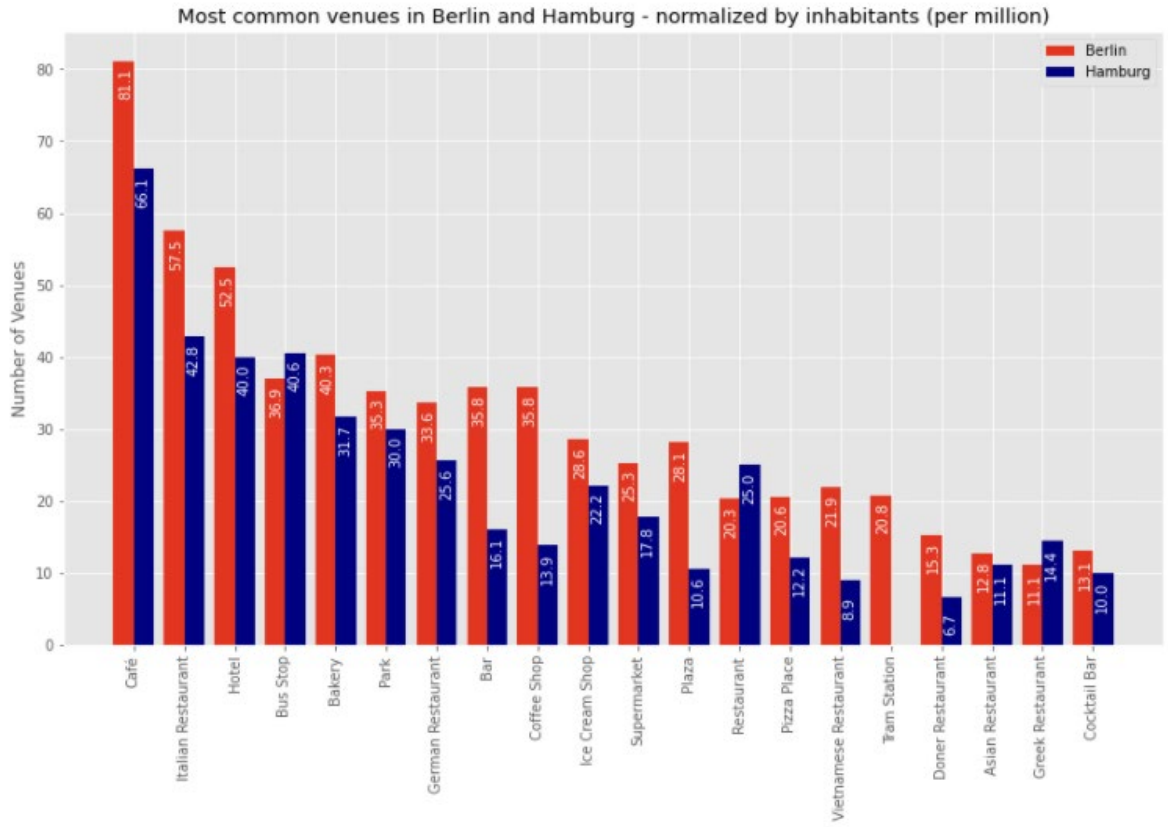
The data is scraped, aggregated and summarized. Next step is to start analyzing each neighborhood seperately in order to prepare forclustering.

3. Data analysis

3.1 Calculations

In order to analyze the neighborhoods of the two cities, we had to define our parameters. Firstly, we needed to calculate the top venue categories for each area code and sort them in descending order, to create the parameter with which we were going to cluster the neighborhoods.

Since the two cities have a larger gap in the amount of residents, we then normalized the results in order to show venue categories per capita (per mil.). Lastly, we visualized the results to get a better understanding of the shape of our dataset.



3.2 k-Means clustering

Now that we created our top 20 categories, we needed to cluster the area codes according to this dataset. k-Means is a type of partition-based clustering in order to partitioning the data base into groups of individuals with similar characteristics. It divides data into non-overlapping subsets (clusters) without any cluster-internal structure. k-Means tries to minimize intra-cluster distances (e.g. Euclidean or other methods for measuring of distance) and maximize inter-cluster distances. It is an iterative algorithm, but the results depend on the initial defined number of clusters. In turn, this means that results (i.e. clusters) are guaranteed, but may be the optimum. Therefore, the algorithm will be run several times with different amount of initially defined clusters. The algorithm returns inertia, or cost, which can be recognized as a measure of how internally coherent clusters are.

In order to find the best k (the number of clusters with the minimum inter-cluster-distance) we assumed a range from 1 to 25 and ran the algorithm to calculate the cost of each k.

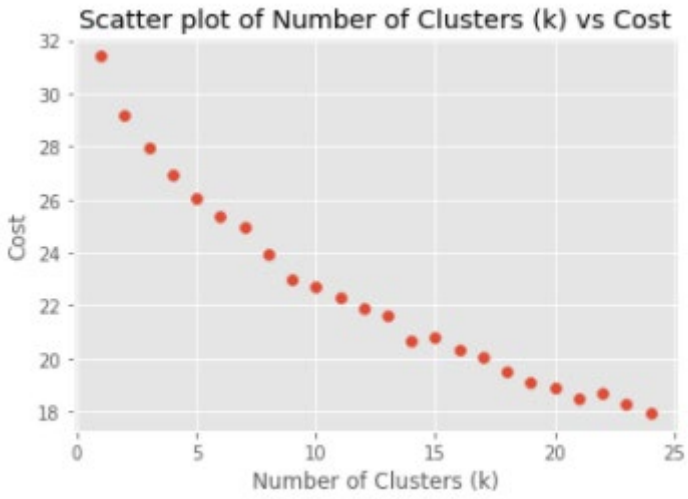
	k	Cost
0	1	31.411739
1	2	29.175120
2	3	27.933183
3	4	26.953626
4	5	26.063333
5	6	25.354706
6	7	24.973040
7	8	23.927760
8	9	22.992647
9	10	22.708074
10	11	22.285820
11	12	21.902745
12	13	21.597440
13	14	20.662899
14	15	20.795345
15	16	20.311052
16	17	20.086581
17	18	19.529389
18	19	19.074907
19	20	18.919807
20	21	18.477108
21	22	18.706107
22	23	18.302873
23	24	17.953517

But what became clear was, that the cost always dropped with an increasing amount of clusters.

We had to find a method to pick out the **best k.** The method we used is known as the ***elbow method.***

The elbow method uses the visualized data to determine, at which point the cost value sharply shifts to form the **tip of the elbow.**

In the graph below it becomes clear, that the abovementioned elbow is achieved wen k=15.



3.3 Mapping

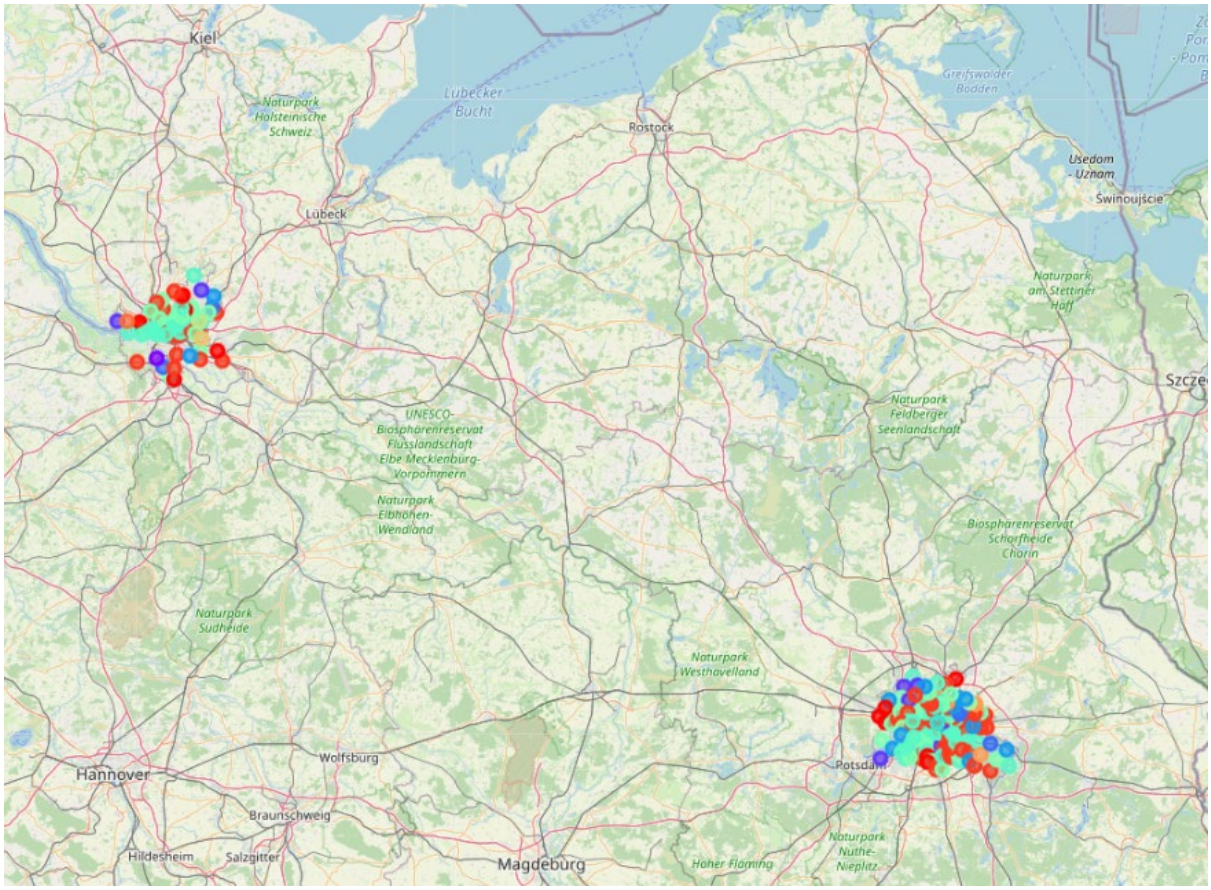
We created our clusters, our top 20 categories and our cleaned area codes. Next step was to merge the data and create a map, showing the different clusters in both cities.

Since 20 categories would only clutter the dataset, we decided to continue with the top 10 most common venue categories per neighborhood and combine them with their respective clusters.

The resulting dataset consisted of 267 rows and 15 columns.

	zipcode	state	latitude	longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	10115	Berlin	52.5323	13.3846	7	Café	Restaurant	Coffee Shop	Hotel	Art Gallery	Pastry Shop
1	10117	Berlin	52.5170	13.3872	14	Hotel	History Museum	Plaza	Chocolate Shop	Exhibit	Monument / Landmark
2	10119	Berlin	52.5305	13.4053	7	Coffee Shop	Italian Restaurant	Ice Cream Shop	Bakery	Café	Bookstore
3	10178	Berlin	52.5213	13.4096	14	Hotel	Plaza	Café	Bookstore	Gym / Fitness Center	Scenic Lookout
4	10179	Berlin	52.5122	13.4164	14	Hotel	Nightclub	Bakery	Café	Thai Restaurant	Event Space

The results were visualized on a map to create a better understanding of the clusters.



4. Conclusion

It became certainly clear, that there are some clusters with more ZIP-Codes and others that only had 1.

It was a result that could have been expected, since some neighbourhoods shared the same structure (e.g. predominantly housing, office etc.). Those neighbourhoods fell into the same cluster, as they shared a similar venue structure.

Purple, red, yellow and dark blue were some of the clusters with many neighbourhoods. Areas that populated cluster 1 (purple) are observed around the city centres, while others (dark blue, yellow) were more peripheral, indicating -probably heavily weighted- housing areas.

5. Discussion

This tool certainly makes it easier, to some extent, to analyze neighborhoods and the similarities between them.

One point of discussion could be, that only the venue structure is not a great amount of information to conclude that two areas are similar to each other. If we wanted to take this analysis a

step further, we could also analyze the climate of the cities/areas and even the location in relation to commute times and economic indexes.

Nevertheless, it could be used as an important step for an individual or corporations who are thinking of moving or expanding to another city, which areas share some similarities to their existing one.

Thank you for reading.