COMPARATIVE ANALYSIS OF GERMAN CITIES

MUHAMMED AYDIN

1. Introduction

1.1 Background

Germany has a decentralized federal government system. Each state has its own constitution, assembly, and government. Moreover, each state has its own legitimate history, that's why the borders between them are not artificial, each border has its own reason. Within the states, the degree of decentralization, as the power of local state organizations and government, is also high. So that each city in Germany may have several unique characteristic features.

On the other hand, after the Second World War, many immigrants from Europe came to cities in Germany as workers. They have also made many contributions to the cities where they live. So that analyzing similarities between cities in Germany may give different insights about Germany and German culture.

1.2 Problem

Each city around the world has its own unique features, such as; Paris with Eiffel Tower. However, they share many similarities, such as restaurants and coffee shops. Hence, the aim of this project is to determine common features between some cities in Germany and predict how much they are similar to each other.

1.3 Interest

Germany is not only an important country in the world but also a famous tourist destination. For this reason, an analysis among cities in Germany and a similarity metric might be very interesting for the researchers and people who may want to visit Germany.

2. Data acquisition and cleaning

2.1 Data source

In order to obtain valid city names, I scraped https://en.wikipedia.org/wiki/List_of_cities_in_Germany_by_population, then I selected the cities having more than five hundred thousand residents. I collected data from Foursquare to analyze the features of cities by using its API. Basically, I used the explore function of Foursquare to option popular spots in the cities. Thereafter, I made a dataset that consists of latitude, longitude, category, and distance variables.

2.2 Data cleaning

Data downloaded and scarped from different sources were combined into one dataset. The scraped data was used only for the selection of cities, and population data was already available. The dataset was obtained via API provides one hundred for each selected city, and it contains several features about cities. It was saved at first as a JSON format, and after determining related data, it was turned into a data frame. Having quite a clean dataset provided flexibility to prepare the dataset for the next step.

2.3 Feature selection

The reason for the selection of five hundred thousand residents threshold is to maximize representability and generalizability of findings. Due to limited data availability at Foursquare about the city Hannover, it was eliminated although it has more than five hundred thousand residents Distance values are used to define how popular spots distribute. In order to cluster their distribution among cities, longitude and latitude were used. The characteristic features of cities were determined according to categories of stores in the cities, and similarity analysis is based on them as well.

3. Exploratory Data Analysis

3.1 Target variables

Our model is based on the share of types of popular spots among cities in Germany. So that we may measure the type of cities. For example, a city might have more restaurants than museums, which shows the character of the city. Hence, the similarity among cities has been measured by the correlation among categories of popular spots in cities. For instance, it is expected that if two cities have more art galleries than theaters, the correlation between that will be higher.

Secondly, the distribution of popular spots location is also analyzed based on their longitudes and latitudes. After applying cluster methods, some cities are divided into four clusters while some are divided into three clusters. Moreover, the visualization of clusters explains many things, such as how grouped cities in terms of popular spots.

3.2 Distance Analysis of popular spots

As it is seen in Table 5, each city has different types of distribution in terms of the distances among popular spots. Some of them have normal distribution while others show different tendencies. Moreover, the range of distributions changes, for instance; the range for Esses and München is different, as from 200 to 1000, and from 100 to 600 respectively. According to Table 6, it may also assume that the degree of centralization for each city has different tendencies, such as that those cities having more than one pick value might have more than one center. Additionally, the shape and range of distributions might give some insights into their city plans.

3.3 Location Analysis of popular spots

According to the longitudes and latitudes of popular spots, Table 6, they can be mapped in order to see their locations. Since the location data is based on popular spots of the city, it, therefore, provides much information about the city; firstly, how the city is planned.

For instance, the following assumption for Frankfurt am Main might be claimed, it is a big city in terms of size(km2), there are more than one city centers, finally, there are parallel roads at the city centers.

3.4 Analysis of Common Categories

Table 1 shows that the most common three location categories among cities are hotel, Cafe, and Plaza. Taken into account the used data source and the features of the facilities, it is reasonable. Moreover, Italian Restaurant is in sixth place, it also may show people's average attitude toward the world cuisines. In Table 3 the share of the ten most common variables for each city is seen. Each category has a different composition of cities, and it shows some characteristics of cities. The Clothing Store can be a good example. As it is known that Düsseldorf is like a Moda center for Germany, that having the biggest share in the clothing store category is understandable. Restaurant types in the city may also show some insights about the culture, such as that Cologne (Köln) has the highest number in the Italian Restaurant category.

Table 2 and Table 4 indicate the most common categories per city. Hotels, Cafes, Coffee shops, or a type of restaurant are the three most common categories among almost every city. It is appropriate with the dataset. Furthermore, each city has its own unique distribution of categories. Berlin might be a valid instance with its popularity in the field of Art. So that History Museums, Art Galleries, and Exhibits are the three most common categories after Hotels in Berlin. Some cities also have some unique features, e.g; München Bavarian has Restaurant as the most common feature after Cafe and Plaza. The distribution of categories might also demonstrate characteristic features of cities. If the distribution is not normal, and few dominant categories have the most share in a city, it might be interpreted as an indicator of a non-colorful lifestyle in that city.

4. Deployed Models

4.1 Clustering models

In order to see how popular spots distribute among cities, scatter plots were used (see Table 6). Because of having an unlabeled dataset, the K Means method was deployed to cluster cities based on their location data of popular spots. By implementing K Means method, it is aimed to observe the similarities among cities. As it is indicated in Table 7, some cities have three clusters while others have four. In addition to cluster numbers, the scatter plots of clusters are also shown in the Table, it is therefore observed some similarities between cities, such as München and Berlin, Leipzig and Bremen.

4.2 Similarity Correlation models

To analyze the similarity between cities, a new Data Frame was made based on the grouped categories per each city. Thereafter, the correlation table is made from these numbers of categories for each city, it demonstrates the degree of similarity between cities. According to Table 8, the most similar city to Berlin is Dresden. As Table 4 shows, the distribution of categories for each city is similar, as having Hotels and Art Museums and others. On the other hand, the most similar city to Dresden is not Berlin, it is Bremen. Finally, the most similar two cities are München and Bremen with a 0.71 score. Regarding Table 4, it is a valid conclusion.

5. Conclusions

Foursquare provides a wonderful dataset to analyze cities. Category data, especially, may give many insights about cities by exploring their distribution, dominancy, and frequency. It is shown that there are several unique characteristic features for each city, in terms of distribution of distances between spots, their locations. After deploying K Means and correlation methods, it is tried to measure the degree of similarity between cities regarding popular spots locations and their categories.As having decentralized government system in Germany, its cities have many unique features, besides their similarities

6. Future directions

Since the study is conducted on cities having more than 500000 population, the number of cities in the study is limited. That’s why it can be increased. Moreover, the number of categories for each city might be affected by population size. Thus, it can be normalized. Lastly, the clusters are based on only location values, so there can be more variables in order to make more appropriate clusters.

APPENDIX

TABLE 1:

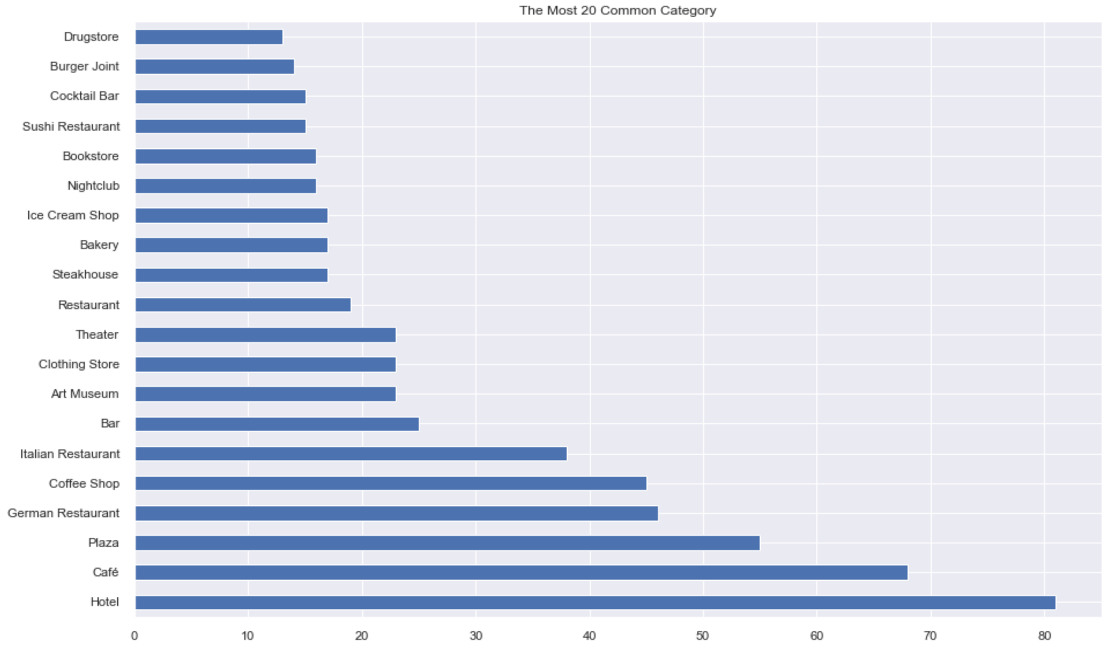


TABLE 2:

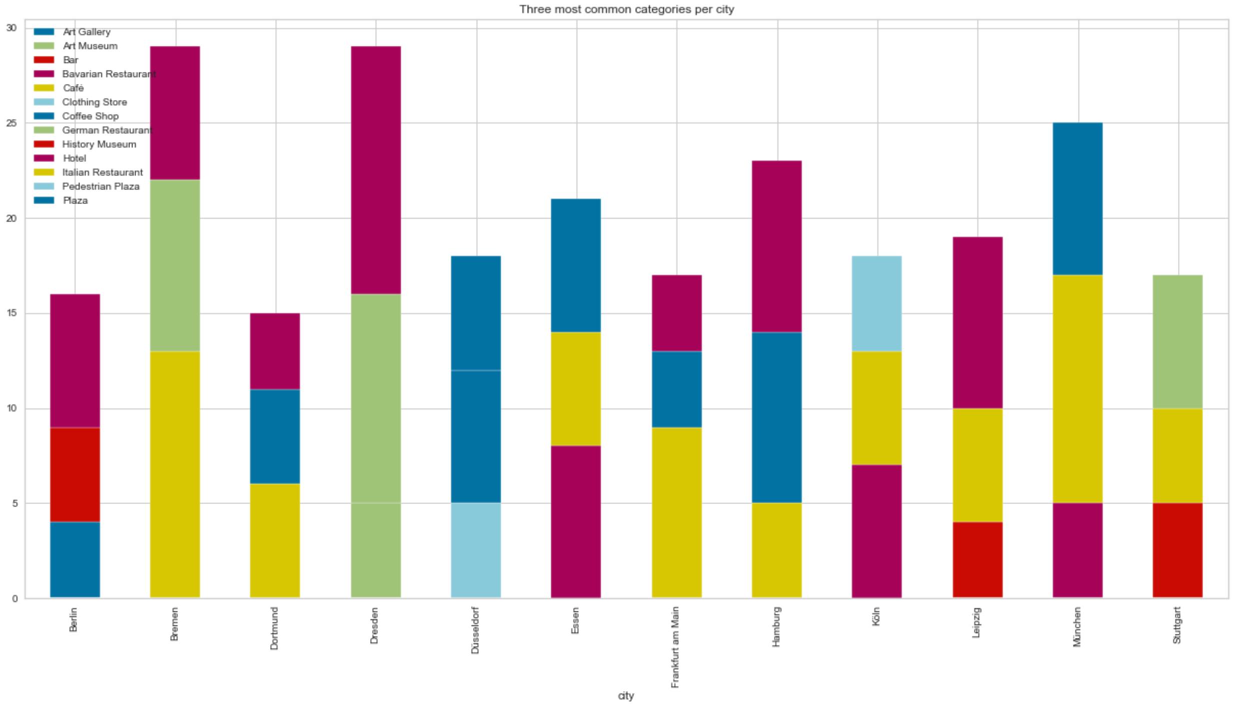


TABLE 3:

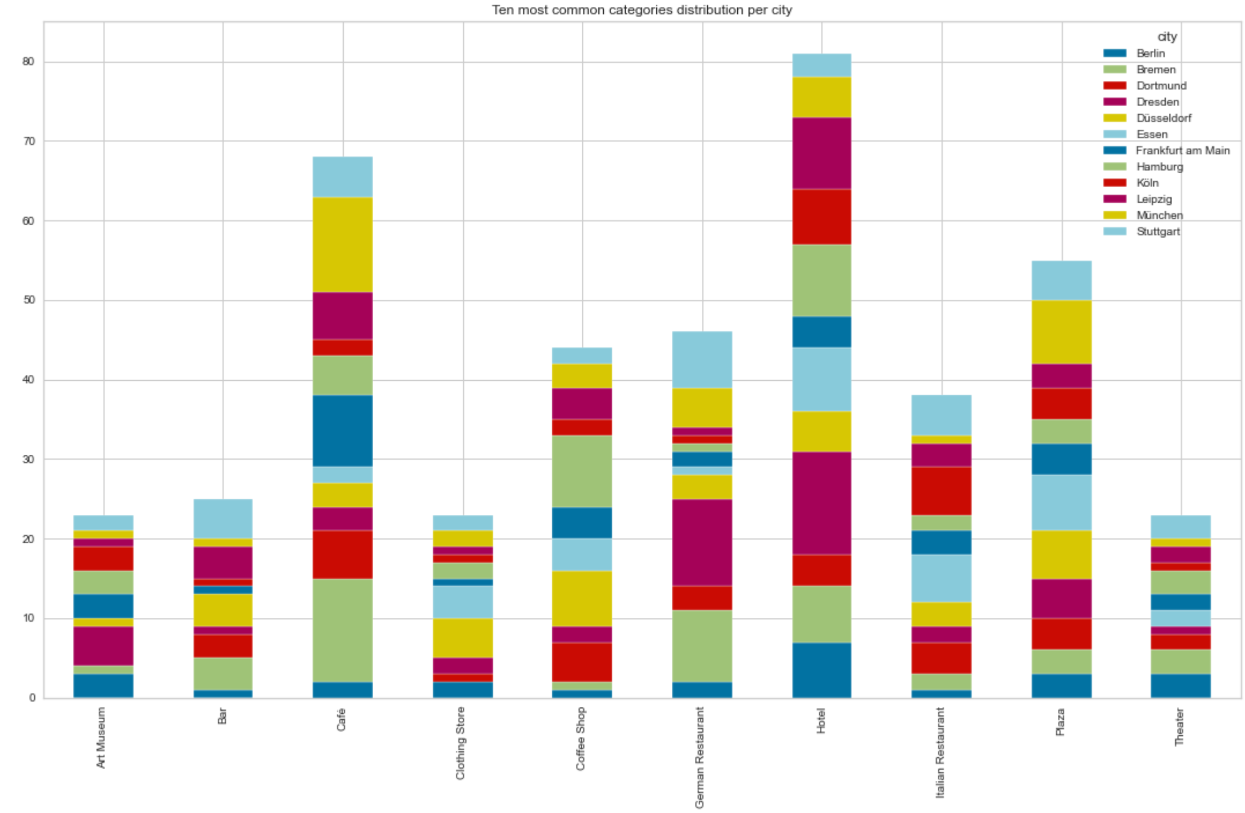

Ten most common categories distribution per city
city
Berlin
Bremen
Dortmund
Dresden
Düsseldorf
Essen
Frankfurt am Main
Hamburg
Köln
Leipzig
München
Stuttgart
0
10
20
30
40
50
60
70
80
Art Museum
Bar
Café
Clothing Store
Coffee Shop
German Restaurant
Hotel
Italian Restaurant
Plaza
Theater

TABLE 4: CATEGORIES PER CITY

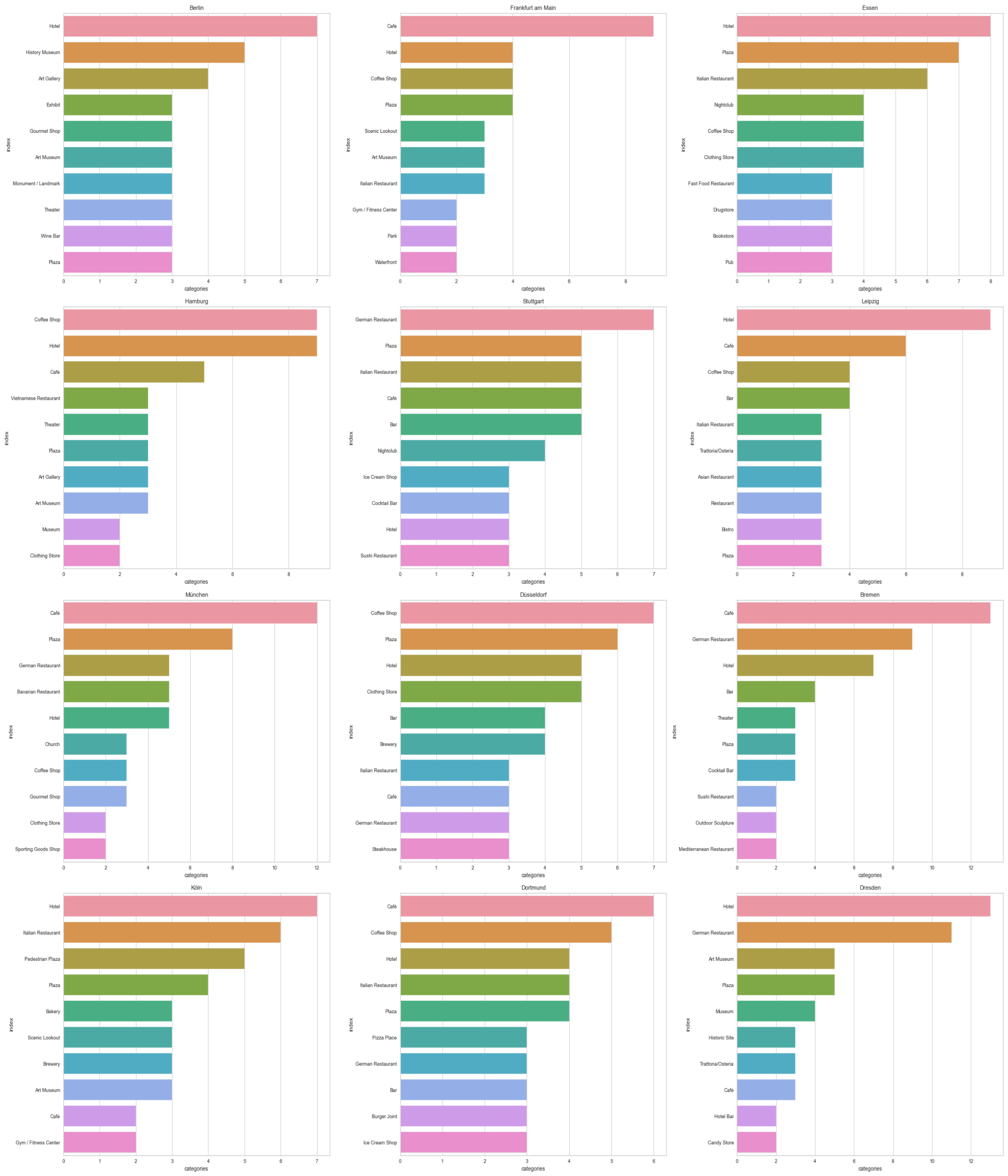


TABLE 5: DISTRIBUTION OF POPULAR SPOTS PER CITY

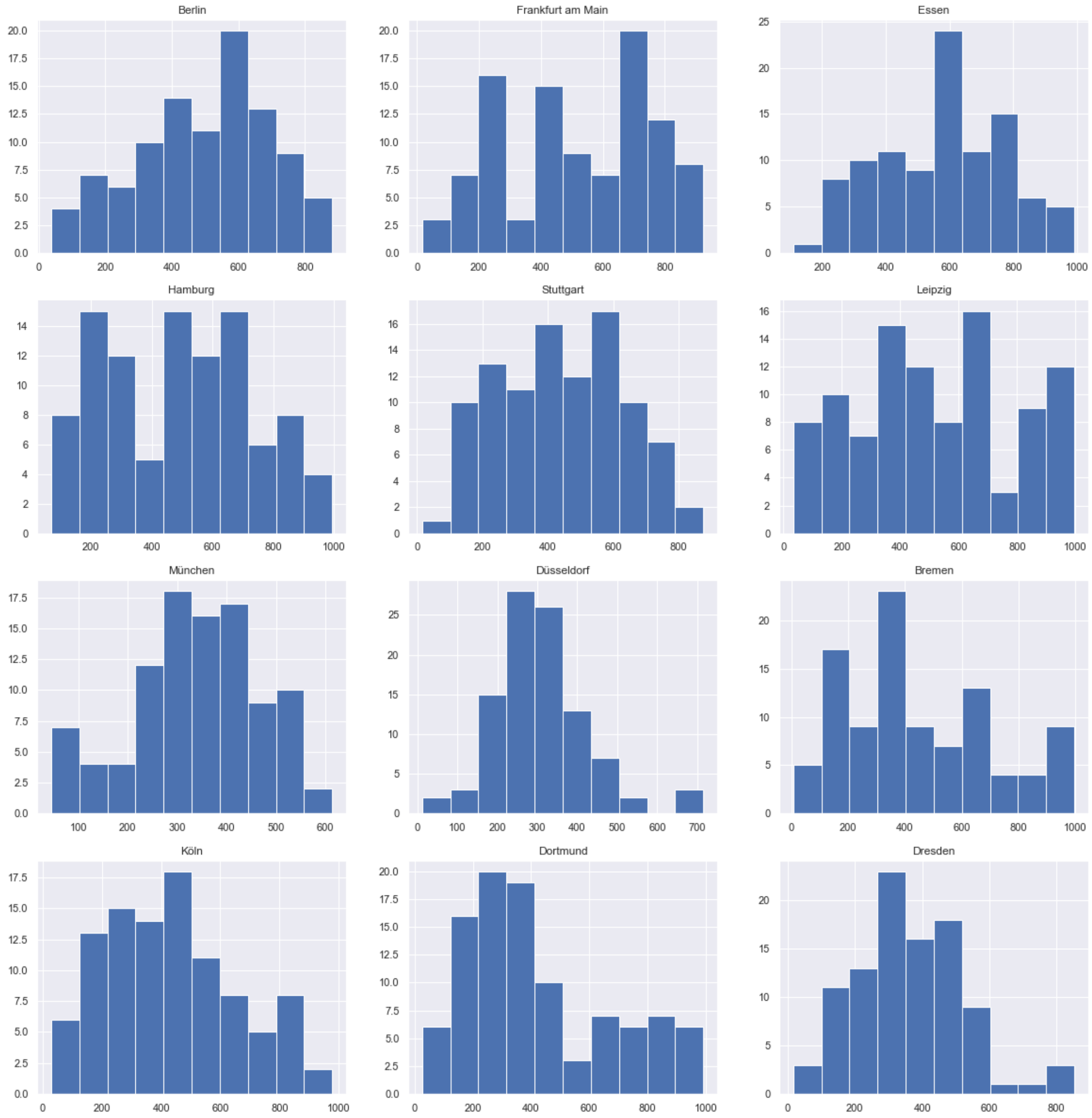


TABLE 6: SCATTER PLOT OF DISTRIBUTION PER CITIES

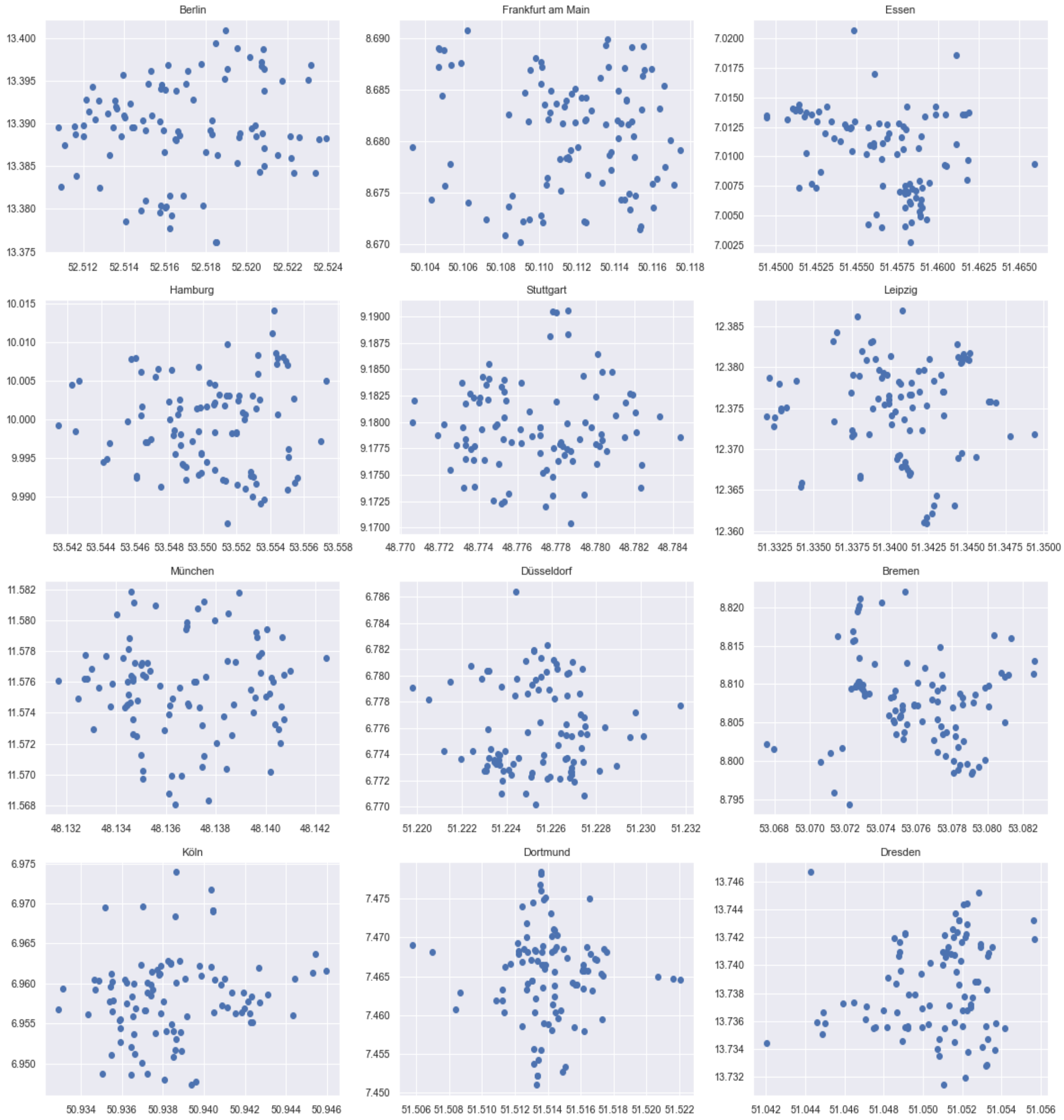


TABLE 7: CLUSTERED SCATTER OF DISTRIBUTION PER CITY

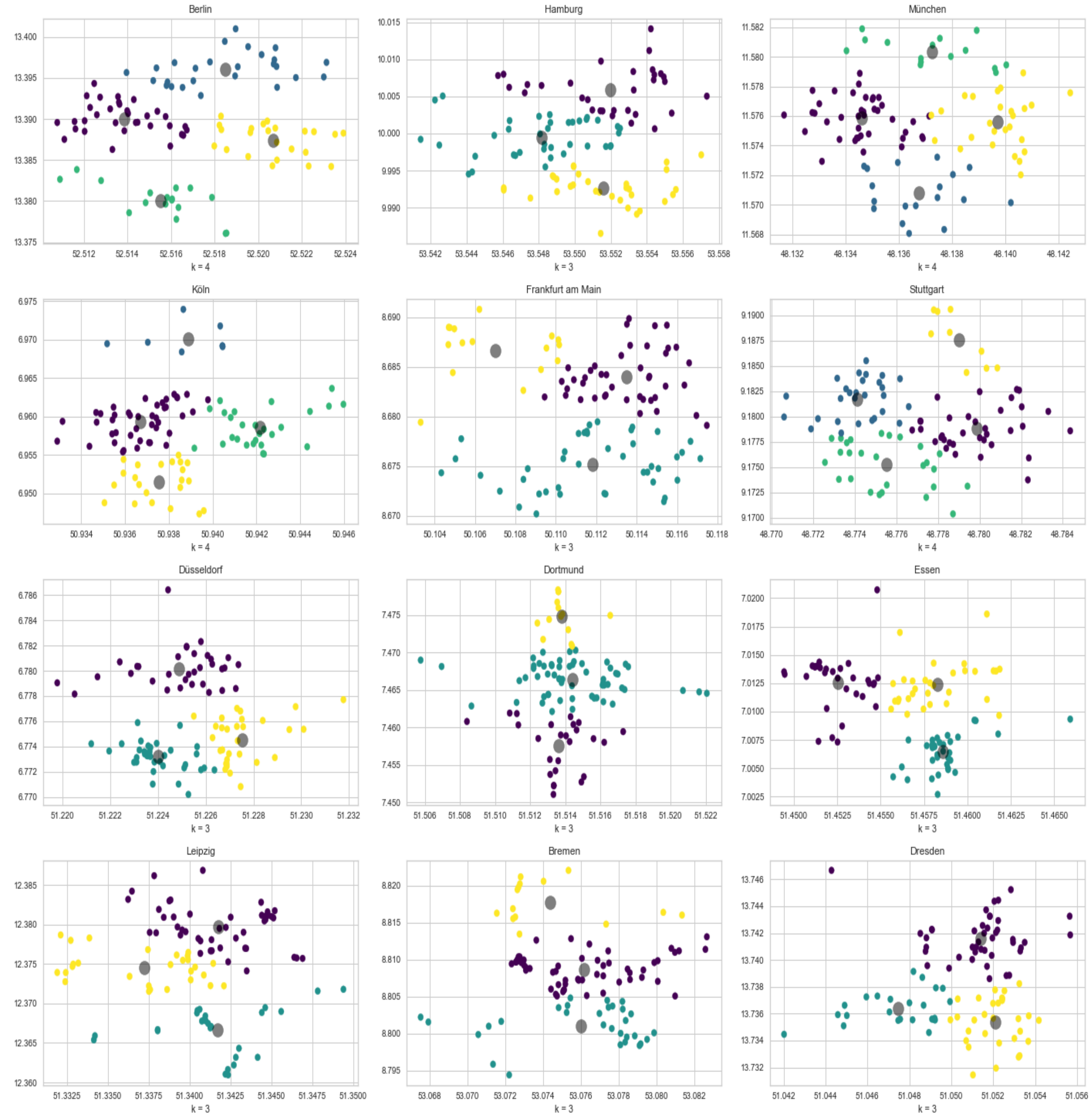
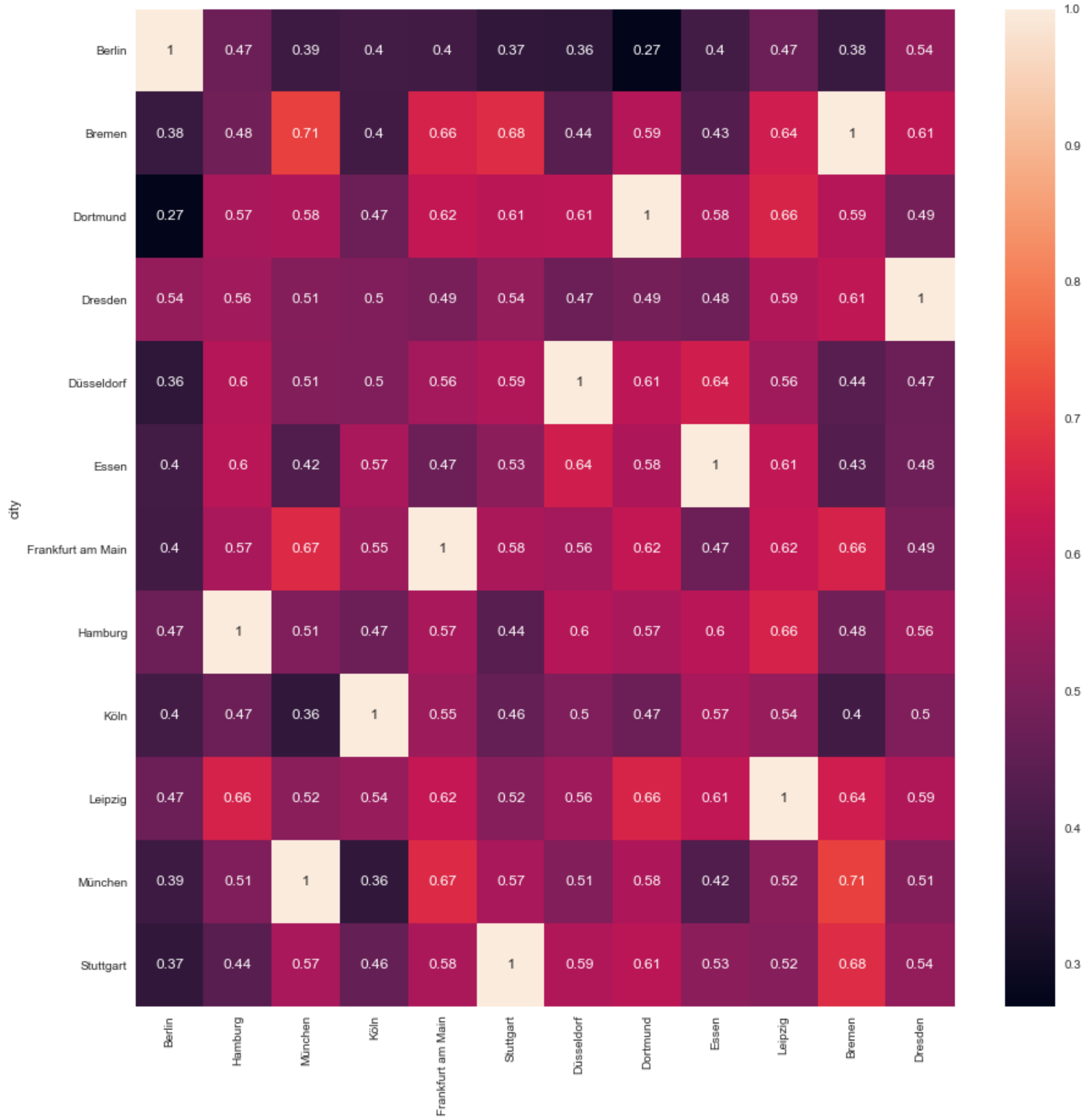


TABLE 8: HEATMAP FOR SIMILARITIES AMONG CITIES



city	Berlin	Hamburg	München	Köln	Frankfurt am Main	Stuttgart	Düsseldorf	Dortmund	Essen	Leipzig	Bremen	Dresden
Berlin	1	0.47	0.39	0.4	0.4	0.37	0.36	0.27	0.4	0.47	0.38	0.54
Bremen	0.38	0.48	0.71	0.4	0.66	0.68	0.44	0.59	0.43	0.64	1	0.61
Dortmund	0.27	0.57	0.58	0.47	0.62	0.61	0.61	1	0.58	0.66	0.59	0.49
Dresden	0.54	0.56	0.51	0.5	0.49	0.54	0.47	0.49	0.48	0.59	0.61	1
Düsseldorf	0.36	0.6	0.51	0.5	0.56	0.59	1	0.61	0.64	0.56	0.44	0.47
Essen	0.4	0.6	0.42	0.57	0.47	0.53	0.64	0.58	1	0.61	0.43	0.48
Frankfurt am Main	0.4	0.57	0.67	0.55	1	0.58	0.56	0.62	0.47	0.62	0.66	0.49
Hamburg	0.47	1	0.51	0.47	0.57	0.44	0.6	0.57	0.6	0.66	0.48	0.56
Köln	0.4	0.47	0.36	1	0.55	0.46	0.5	0.47	0.57	0.54	0.4	0.5
Leipzig	0.47	0.66	0.52	0.54	0.62	0.52	0.56	0.66	0.61	1	0.64	0.59
München	0.39	0.51	1	0.36	0.67	0.57	0.51	0.58	0.42	0.52	0.71	0.51
Stuttgart	0.37	0.44	0.57	0.46	0.58	1	0.59	0.61	0.53	0.52	0.68	0.54