Brexit & Venue Data Analysis of London/Frankfurt

Leonard Hußke

March 18th, 2020

This is a submission to the Final Assignment of the 'Applied Data Science'-Course provided by Coursera and hosted by IBM.

Introduction

Since Great Britain left the European Union on the 31st of January this year, the employees of 30 banks and 20 financial service providers are going to move to Frankfurt. At least 1500 new jobs were created in the banking sector in Frankfurt. Since a good connection to restaurants, green spaces and cultural offers is more and more important for townspeople it is important to provide information about the different boroughs and areas of the financial capital of Germany. Employers are interested in a smooth relocation of employees so that they can return to work as soon as possible.

The goal is to compare the areas of London with the boroughs of Frankfurt, using data science techniques and machine learning. The employers and employees can use the information to find places in Frankfurt that are quite similar (e.g. same diversity of venues) to their home in London.

Acknowledgments

This repo uses the *Foursquare API* to obtain information about top venues of different boroughs of Frankfurt and London.

I also used *Folium* to create interactive maps and *geopy* to obtain coordinates.

Scikit-learn was used to create a classification model.

My thanks to Alex Aklson for leading the 'Applied Data Science'-Course and for teaching different tools and techniques to make this exciting work possible!

Structure

The structure of this report is the following. The report has five parts, the first is Introduction, with an overview of the background and a short description of the main interest of the project. The second section deals with the Data, where I describe the data that will be used to solve the problem and the source of the data. The third part is Methodology, which is the main component of the report where I discuss and describe the exploratory data analysis. This is followed by the Results and Discussion section, where I discuss the results and observations. The final fifth section is the Conclusion, where I conclude the report.

Data

This section gives an overview of the data used to solve the problem and describes shortly how to access and use the data to contribute to solve the problem.

Data Sources

Two different data sources where used in this project. The first data source is Wikipedia to get a list of the different boroughs and areas of London and Frankfurt. The second source is the Foursquare API to get in-depth information about the areas of both cities.

Geospatial Data

At the first step I visualized the different areas of Frankfurt and London to get an overview. To visualize the cities, I scraped the data from Wikipedia and used the *geopy* library to get latitude and longitude for the different areas.

To compare the areas/boroughs of London/Frankfurt we have to get to know both cities a little bit better. London has a population of approximately 9 million people, covers 1572 km^2 and is organized in city of London & 32 boroughs.

Frankfurt on the other side has a population of approximately 0,8 million people and covers only 248,31 km^2 . It is organized in 46 boroughs, but those boroughs are significantly smaller than the boroughs of London. Because of that it is not useful to compare the boroughs.

I decided to compare the London areas with the boroughs of Frankfurt because the area and the population is more comparable. In total we got 46 boroughs of Frankfurt and around 518 areas of London.

E.g.

London area: Barnes, area: 4,50 km^2 , population: 21.218

Frankfurt area: Ostend, area: 5,56 km^2 , population: 29.171

	Borough	Latitude	Longitude
0	Altstadt	50.1104	8.6829
1	Innenstadt	50.113	8.67434
2	Bahnhofsviertel	50.1077	8.66868
3	Westend-Süd	50.1152	8.66227
4	Westend-Nord	50.1264	8.66792

...

Tab. 1: Frankfurt boroughs and coordinates

	Area	Latitude	Longitude
0	Abbey Wood	51.4876	0.11405
1	Acton	51.5081	-0.273261
2	Addington	51.3586	-0.0316347
3	Addiscombe	51.3797	-0.0742821
4	Albany Park	51.4354	0.125965

...

Tab. 2: London areas and coordinates

Based on the generated data by the *geopy* library I created *folium* maps to visualize the boroughs/areas I am working with. Both maps have the same scale.

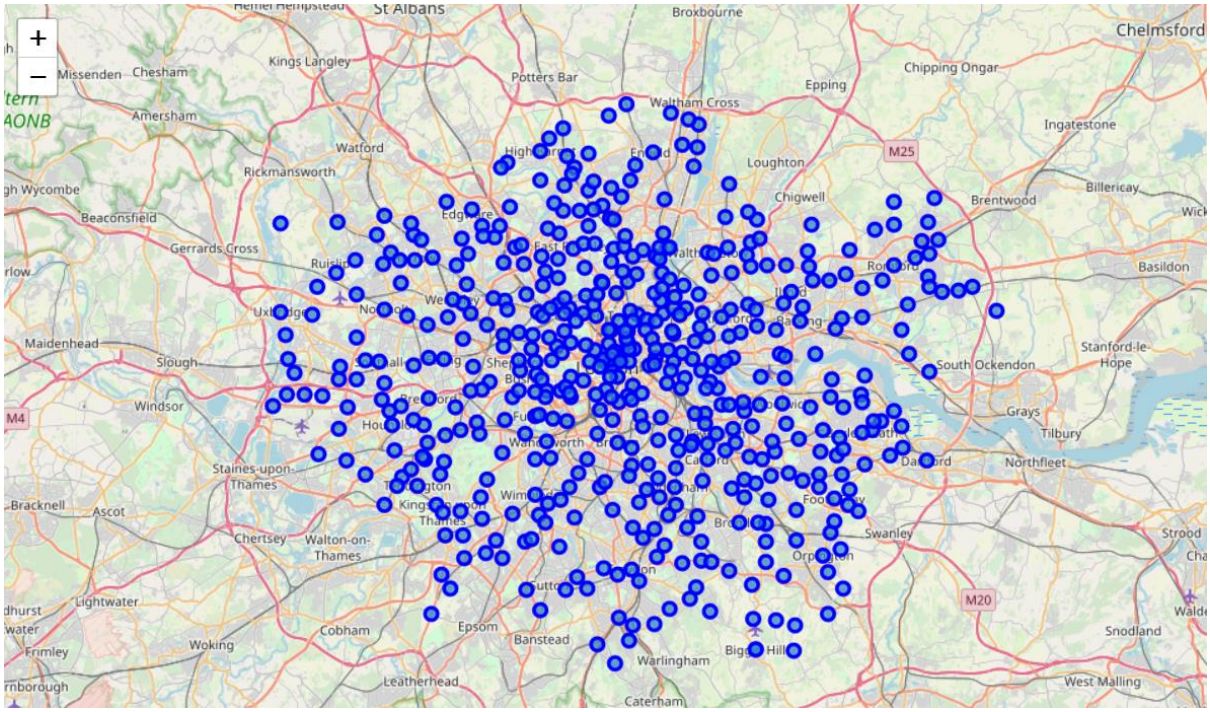

Fig. 1: Areas of London visualized with folium and geopy

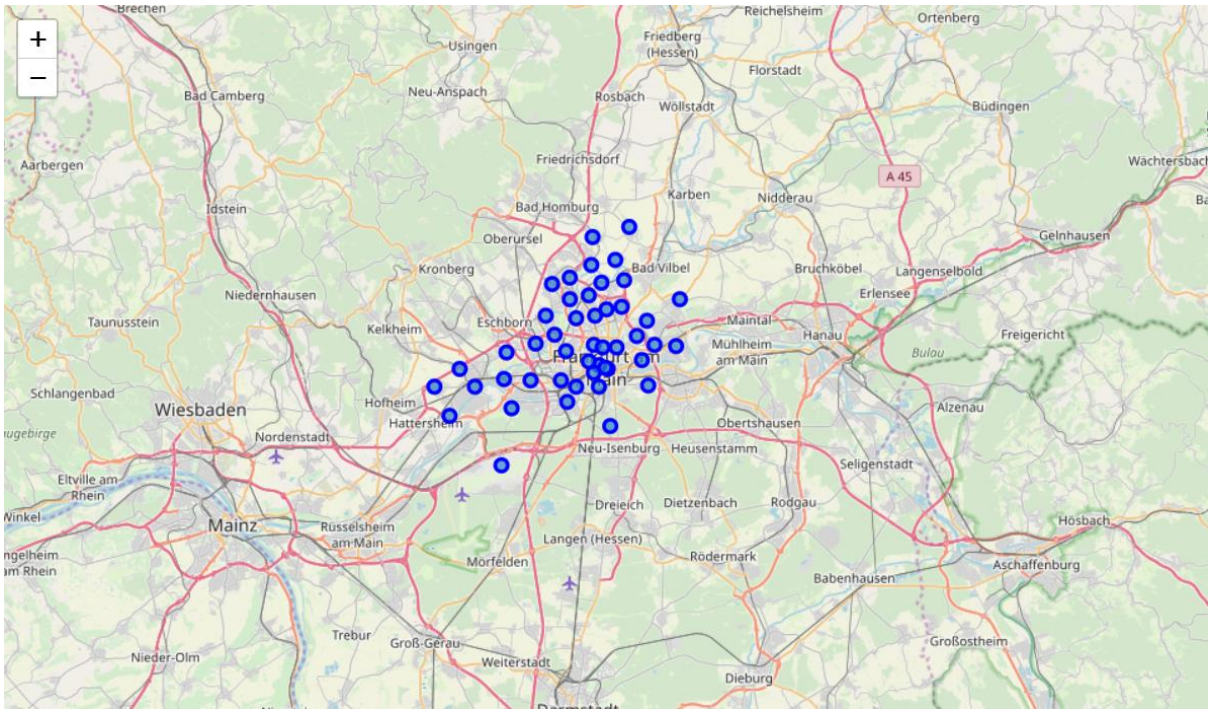

Fig. 2: Boroughs of Frankfurt visualized with folium and geopy

Venue Data

To fetch the venue data, I had to think of a radius for each borough/area. I opted 500 meters because it is enough to cover most of the venues. And boroughs/areas which are close to each other do not overlap that much.

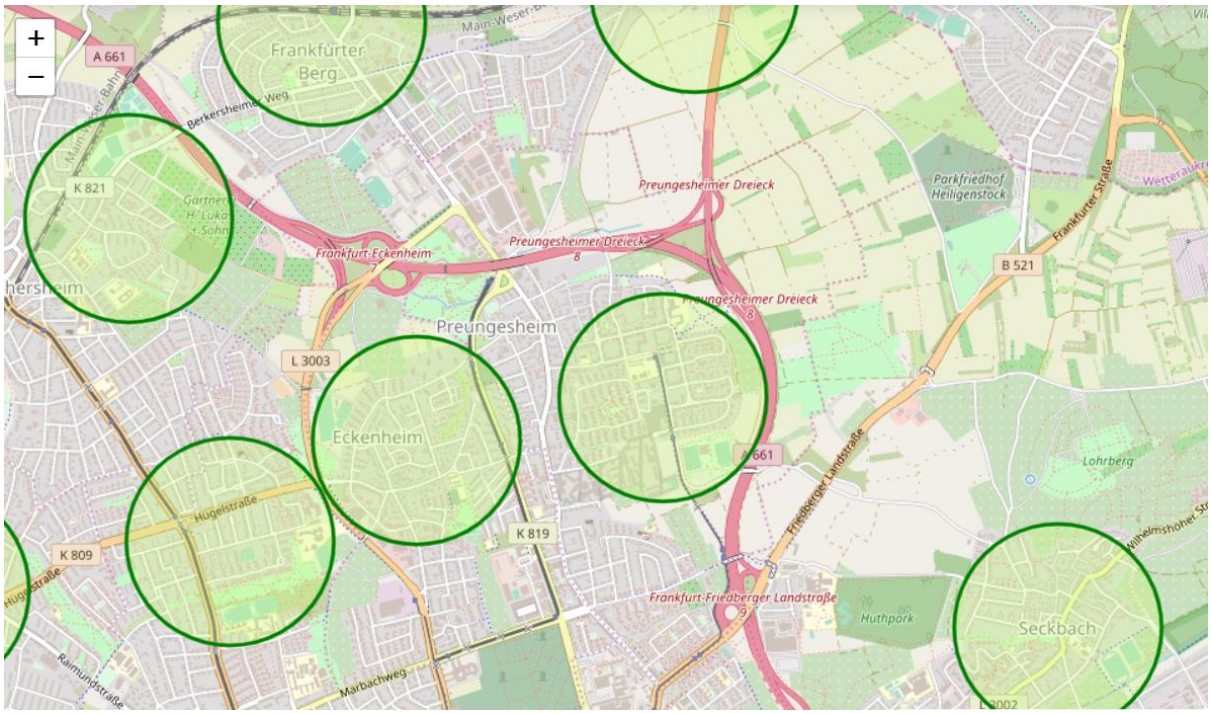

Fig. 3: 500m radius of boroughs in Frankfurt

The Foursquare API was used to obtain information about the top hundred venues of each borough/area. I chose hundred venues because it provides enough information to classify the areas and it is in the range of about 90.000 free calls of the Foursquare API. To get further information of each borough/area you could increase the number of venues fetched or you could increase the radius of each borough/area.

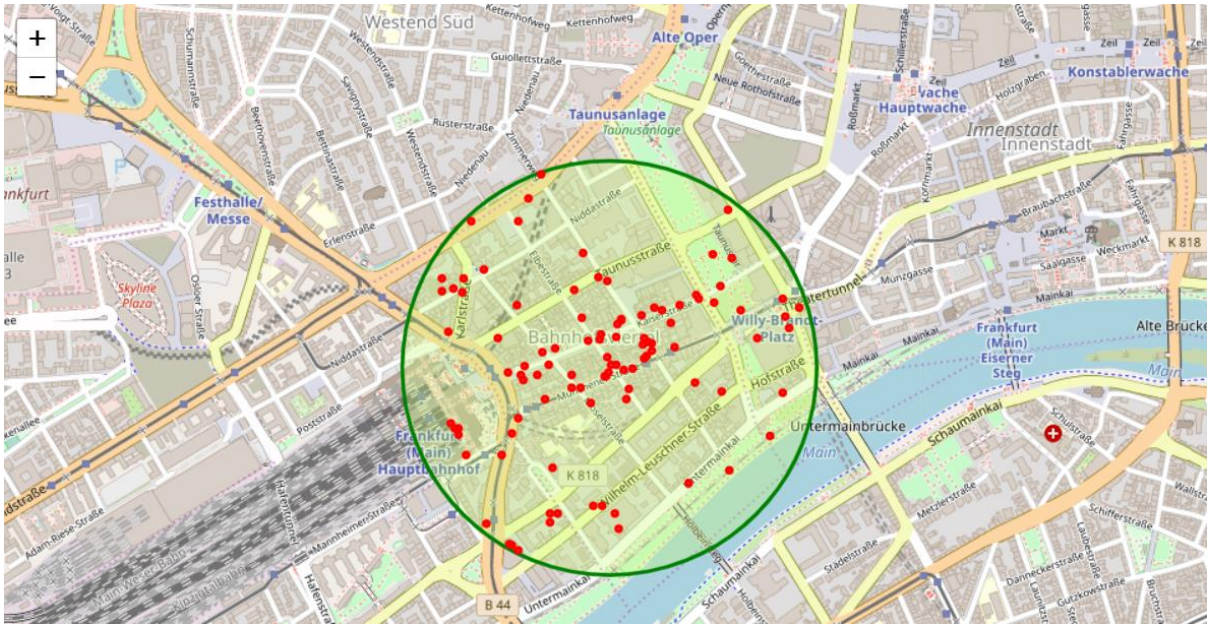

Fig. 4: Location of venues, represented with red dots, of "Bahnhofsviertel" (borough of Frankfurt)

As you can see in the figure above, the top hundred venues of a borough with a radius of 500m are displayed. If you create the figure Fig. 4 in a *jupyter notebook*, you get an interactive map and you can select the red dots for further information about the venue.

In the following section I describe how I used the data.

Methodology

As described in the introduction, I want to assign each area of London to a borough of Frankfurt.

I got all the location data I need to obtain all the areas/boroughs of both cities and I fetched all the venues with their specific coordinates, category and name.

First step of the analysis will be the calculation and exploration of the venues across the different areas/boroughs. Here I will use a heat map to compare areas/boroughs of both cities with a high number of venues.

In the second and final step I will focus on classifying the areas of London. I will have to prepare the features for the logistic regression model and I need to train it. I will present a map of London with markers labeling the different areas and which borough of Frankfurt fit best.

Exploratory Data Analysis

To get a rough picture of the boroughs/areas I am working with I checked the distribution of the venues per area. As already described, we have 46 boroughs in Frankfurt and 518 areas in London. The mean distribution in Frankfurt is 17 venues per borough and in London it is 24 venues per area. If we compare the distribution of both cities we can see that the distribution has similarities. Both cities have many boroughs/areas with less than 10 venues in a radius of 500 meters and then some “hotspots” with up to hundred venues.

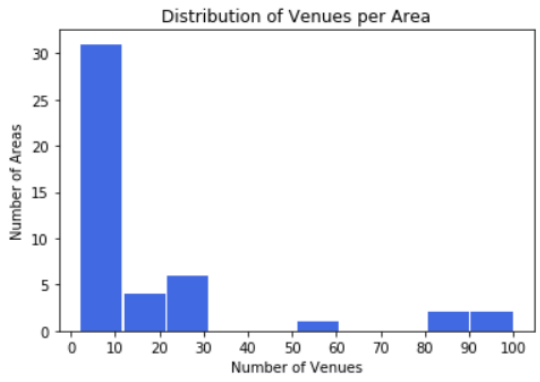

Fig. 5: Distribution of Venues per Area in Frankfurt

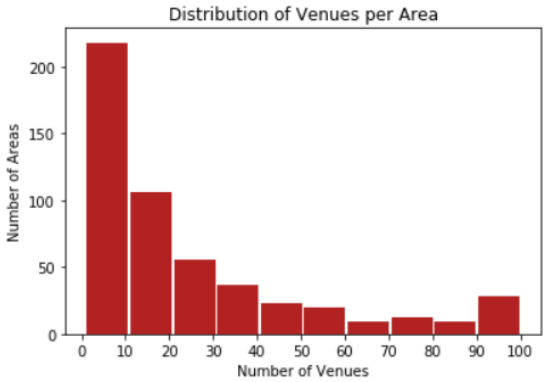

Fig. 6: Distribution of Venues per Area in London

To visualize the insights I got, I decided to create heat maps for both cities.

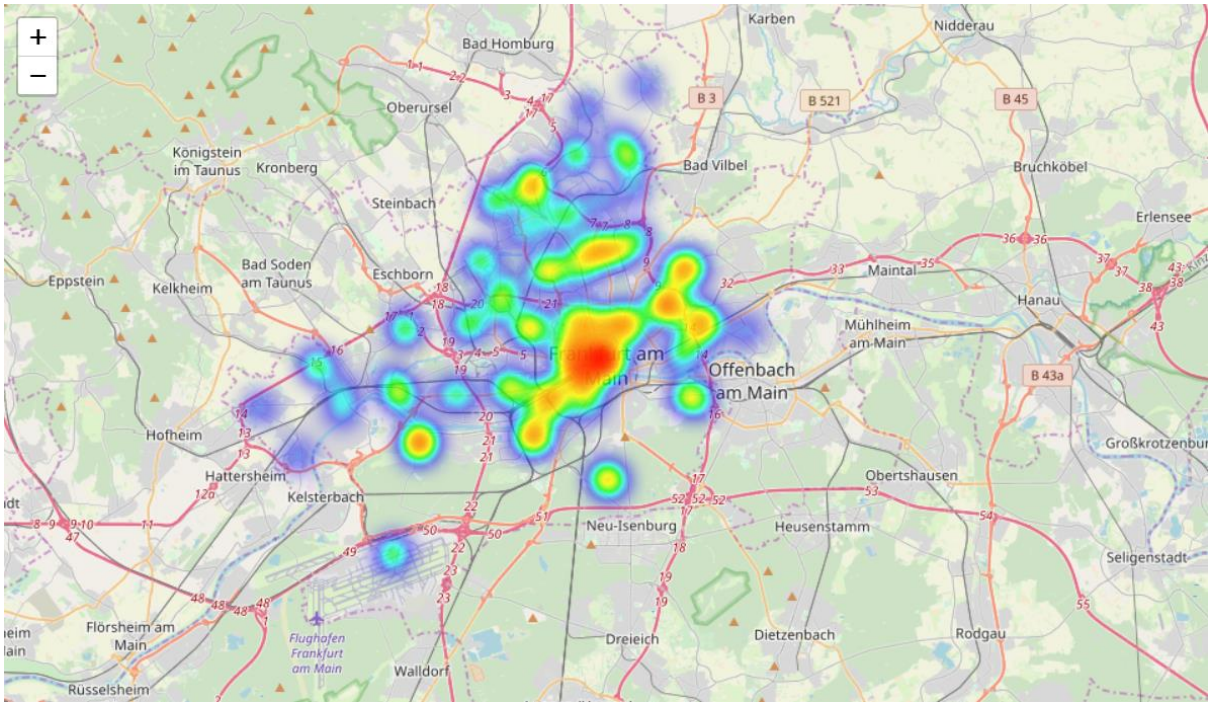

Fig. 7: Heat map of venue location in Frankfurt

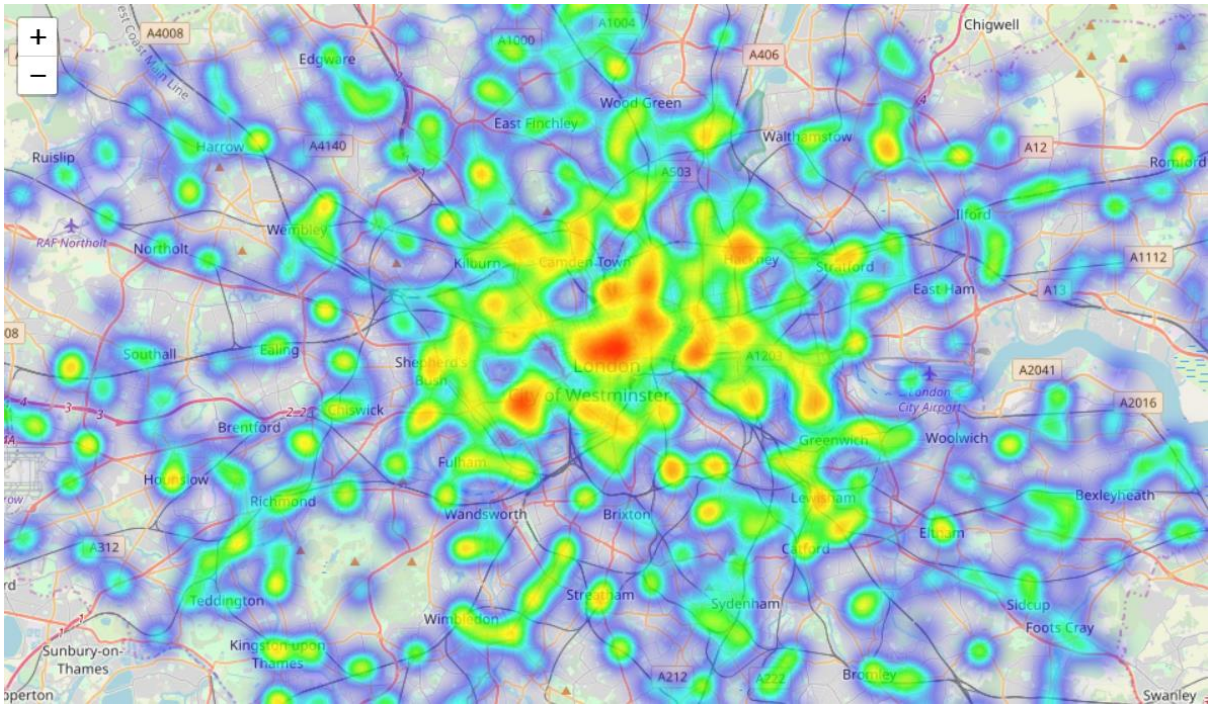

Fig. 8: Heat map of venue location in London

As expected, the closer you are to the city center, the more venues you can find.

Now I will explore the different venues. Frankfurt has 180 and London has 412 unique venue categories. But the main difference is the number of venues in total. Frankfurt has about 800 venues and London has about 12500 venues.

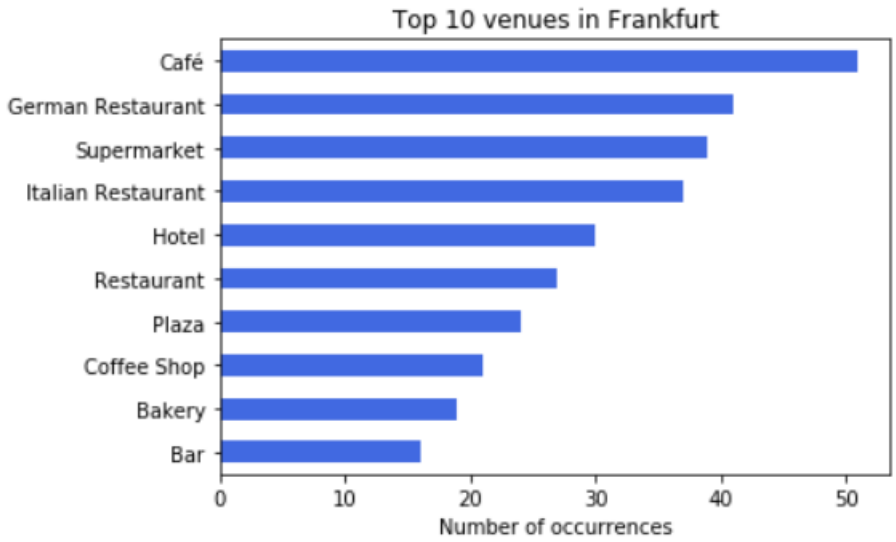

Fig. 9: Top 10 venues in Frankfurt

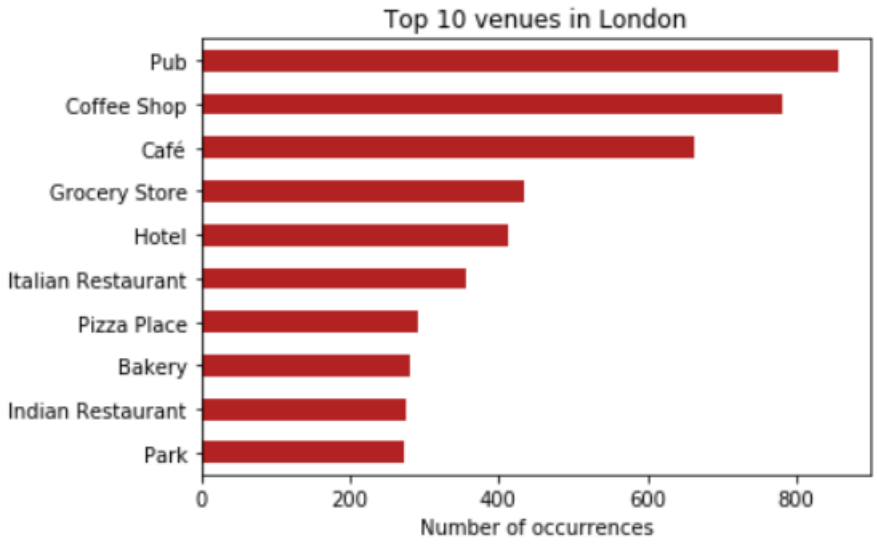

Fig. 10: Top 10 venues in London

As we can see, the top ten venues of both cities are quite similar. Coffee/Café is by far the most common venue in Frankfurt and London. In both cities are many Italian Restaurants/Pizza Places. One of the main differences is the high number of German Restaurants in Frankfurt and the high number of Indian Restaurants in London. This difference is due to varying population of both cities. There are way more Indian people living in London than in Frankfurt. Apart from that, there are no different venues. Only the number of those venues is different in both cities.

Based on these information, we can say that the main venues in both cities are quite similar. That is a good sign for any employees whom have to move to Frankfurt.

Venue Data

To compare the different boroughs and areas, I have to map the Boroughs/Areas with their occurring venue categories. The different categories are categorical data, so I use one-hot-encoding

to generate binary features. Now I group them by borough/area to get the mean frequency of occurrence of each category. These features can be used for machine learning.

	Area	African Restaurant	Airport Lounge	Airport Service	American Restaurant	Apple Wine Pub	Art Gallery	Art Museum	Asian Restaurant	Athletics & Sports	...	Train Station	Tram Station	Transportation Service	Trattoria/Osteria	Turkish Restaurant
0	Altstadt	0.00	0.0	0.0	0.0	0.0	0.0	0.043478	0.00	0.0	...	0.0	0.0	0.0	0.0	0.00
1	Bahnhofsviertel	0.01	0.0	0.0	0.0	0.0	0.0	0.010000	0.01	0.0	...	0.0	0.0	0.0	0.0	0.00
2	Berkersheim	0.00	0.0	0.0	0.0	0.0	0.0	0.000000	0.00	0.0	...	0.0	0.0	0.0	0.0	0.00
3	Bockenheim	0.00	0.0	0.0	0.0	0.0	0.0	0.000000	0.12	0.0	...	0.0	0.0	0.0	0.0	0.04
4	Bonames	0.00	0.0	0.0	0.0	0.0	0.0	0.000000	0.00	0.0	...	0.0	0.0	0.0	0.0	0.00

...

Tab. 3: Mean frequency of each category in every borough of Frankfurt

To obtain further information I show the ten most common categories in each borough/area.

	Borough	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Altstadt	Café	Plaza	Restaurant	German Restaurant	Art Museum	Burger Joint	Coffee Shop	Bistro	History Museum	Italian Restaurant
1	Bahnhofsviertel	Hotel	Indian Restaurant	Bar	Bakery	Restaurant	Café	Chinese Restaurant	Seafood Restaurant	Vietnamese Restaurant	Burger Joint
2	Berkersheim	German Restaurant	Construction & Landscaping	Farmers Market	Garden	Furniture / Home Store	French Restaurant	Fountain	Food Court	Food & Drink Shop	Food
3	Bockenheim	Café	Asian Restaurant	Ice Cream Shop	Supermarket	Hookah Bar	Gym	Greek Restaurant	Metro Station	Middle Eastern Restaurant	Department Store
4	Bonames	Diner	Doner Restaurant	Café	Metro Station	Bakery	German Restaurant	Ice Cream Shop	Italian Restaurant	Electronics Store	Food Court

...

Tab. 4: Boroughs merged with the ten most common venues in Frankfurt

If you have a closer look you can find many of the most common Venues of the cities in the most common venues of each borough/area as well. Seem like I did it right.

I am not able to start the classification algorithm now, because, as I mentioned earlier, the feature size of the Frankfurt dataset and the London dataset is not equal. We got 180 categories/features in Frankfurt and 412 categories/features in London. Since a machine learning model needs preprocessed data, it is important to reduce the feature size of the London dataset. I adjusted both datasets. I deleted missing values and a some duplicated columns.

Now I got a Frankfurt dataset with 46 rows, each row for a borough, and 164 columns, each column for a category except the one column with the borough/area name. And a London dataset with 518 rows for each area and also 164 columns. I dropped the “area”-column of the London dataset to get only the categories of London without labeling.

I split the Frankfurt dataset into the “X_train”-set which includes all the categories and the “y_train”-set which contains only the label of the borough. These are the sets to train the logistic regression model.

Now I classified the boroughs of Frankfurt for each area of London by using the London dataset as test set.

	Numerical Labels	Area	Latitude	Longitude	Classification 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	7th Most Common Venue	8th Most Common Venue	
0	40	Abbey Wood	51.487621	0.114050	Sossenheim	Convenience Store	Grocery Store	Playground	Campground	Zoo Exhibit	Filipino Restaurant	Event Service	Event Space	
1	40	Acton	51.508140	-0.273261	Sossenheim	Pub	Gym / Fitness Center	Hotel	Cocktail Bar	Coffee Shop	Fast Food Restaurant	Grocery Store	Chinese Restaurant	I
2	14	Addington	51.358636	-0.031635	Griesheim	English Restaurant	Tram Station	Gas Station	Bus Station	Film Studio	Event Service	Event Space	Exhibit	Re
3	40	Addiscombe	51.379692	-0.074282	Sossenheim	Park	Grocery Store	Café	Pub	Cosmetics Shop	Chinese Restaurant	Bakery	Fast Food Restaurant	
4	40	Albany Park	51.435384	0.125965	Sossenheim	Pub	Indian Restaurant	Train Station	Grocery Store	Zoo Exhibit	Filipino Restaurant	Ethiopian Restaurant	Event Service	

Tab. 5: The London dataset with added "Classification Labels" and "Numerical Labels"

In the table above I added the output of the machine learning model and merged it with the London dataset which includes latitude and longitude.

Now I got all data merged together in one dataset to generate a *folium* map of the London areas which are labeled with the best fitting borough of Frankfurt.

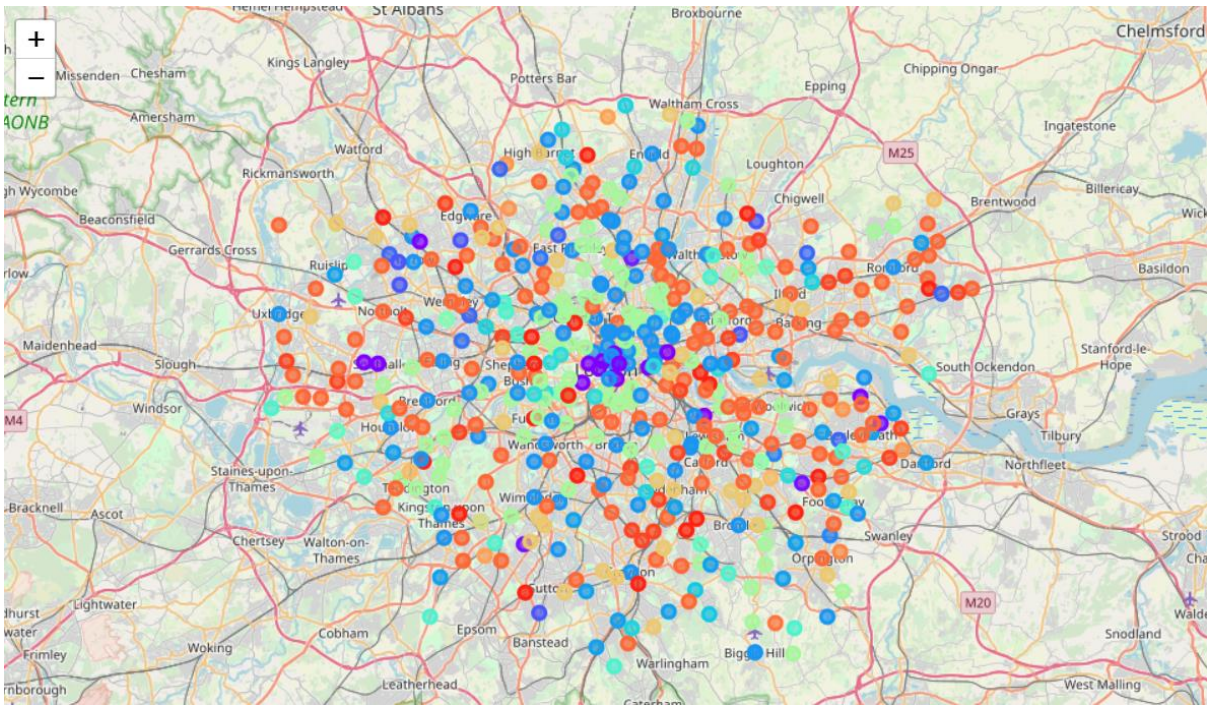

Fig. 11: London map categorized by Frankfurt boroughs

Each color represents a different borough of Frankfurt. If you generate Fig. 11 in a jupyter notebook you obtain an interactive map and if you clicked one of the colorful circles, a HTML-Marker will appear with the name of the area and the fitting borough of Frankfurt.

The purple label stands for the "Bahnhofsviertel" in Frankfurt. A borough with a lot of venues in the center of Frankfurt. That's why there are also many purple labels in the center of London. The city centers seem to have similarities.

This concludes the analysis.

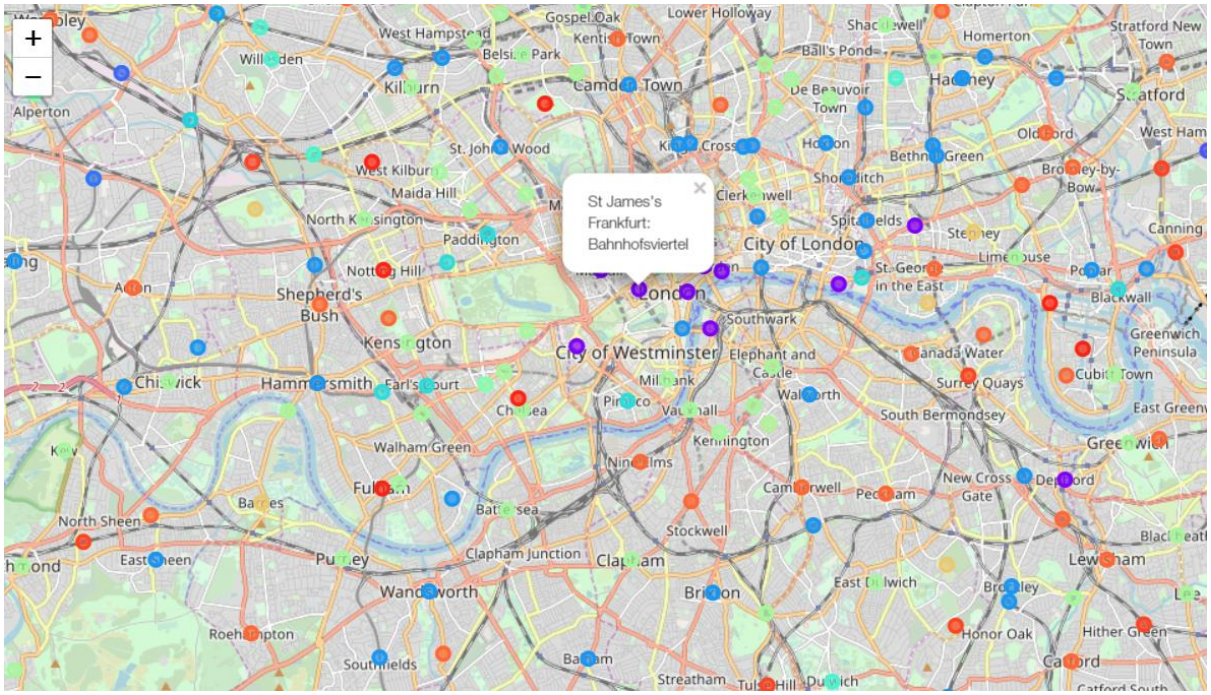

Fig. 12: HTML-Marker shows the fitting borough of Frankfurt

Result and Discussion

After fetching the venue data for London and Frankfurt, I could show that you can appropriately compare the boroughs of Frankfurt with the areas of London. Both datasets contained quite similar categories. To improve the logistic regression model, it is important to preprocess the data. I could spent more time comparing and selecting the right features to get a better result. For example merging two fairly similar columns to one column (e.g. Café and Coffee Shop). Another option would be choosing a different radius or using GIS data to obtain the real borders of each borough/area.

Result of all this is the colorful map shown in Fig. 12. Each dot contains the best potential borough for the selected area. This, of course, does not imply that those boroughs are actually optimal locations for a Relocation. Purpose of this analysis was only to provide information about areas of London with quite similar venues to the boroughs of Frankfurt. Recommended boroughs should therefore be considered only as a starting point for a more detailed analysis.

Conclusion

Purpose of this project was to identify boroughs of Frankfurt which have a similar distribution of venues like areas in London.

As a result employers and employees from London are able to explore the map of London to find matching boroughs of Frankfurt. Not only bank and financial service providers can use the information. Also city managers and people who would like to move to Frankfurt could use it. Even people who live in Frankfurt and want to move to London can use the information as well.