Final Capstone Project Report

Clustering the neighborhoods in Frankfurt

Filippo Claps

1. Introduction

Business problem

Frankfurt is the main financial center in Germany, and one of the most important in Europe. Almost one million people live in the city, and many more work there and commute every day. The city has a wide offer of restaurants from all over the world, cultural points-of-interest, and outdoor activities. Furthermore, being the city relatively small, the density of activities in each district is very high.

The purpose of this modeling activity is to provide people interested in moving to Frankfurt with an additional tool to evaluate the possible neighborhoods where to move in, and whether those neighborhoods are suited for their needs. For example, a student will probably prefer a district relatively close to the University, with a vibrant nightlife and many restaurants, while a worker with family might prefer a quieter area with good connection to public transportation.

Data

To solve this problem, I will use a list of the main districts in the city of Frankfurt (Stadtteile) and extract the main venues for each district using the Foursquare API. Given that the city is relatively small, the venue data returned from Foursquare, accounting for the limits of the free account, should give a complete picture of all the points-of-interest that characterize each neighborhood.

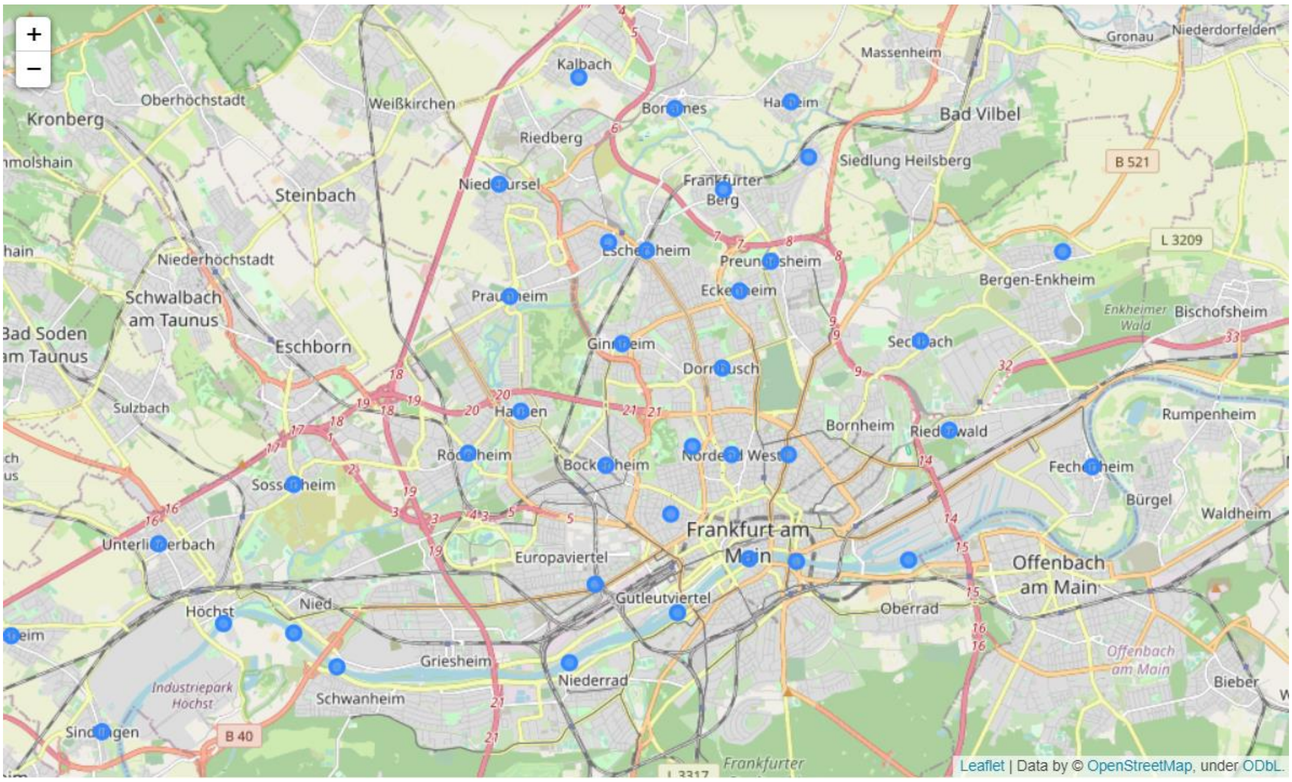

Figure 1 - Map of Frankfurt and Districts

Solution

It is possible to identify the typology of a neighborhood by using the venue data collected with Foursquare and to transform them in such a way that they can be “fed” to a model. Since there are no explicit labels or numerical value to be predicted, classification and regression models will be ruled out. A clustering algorithm, the K-Means algorithm in particular, seems the most appropriate choice to solve this problem.

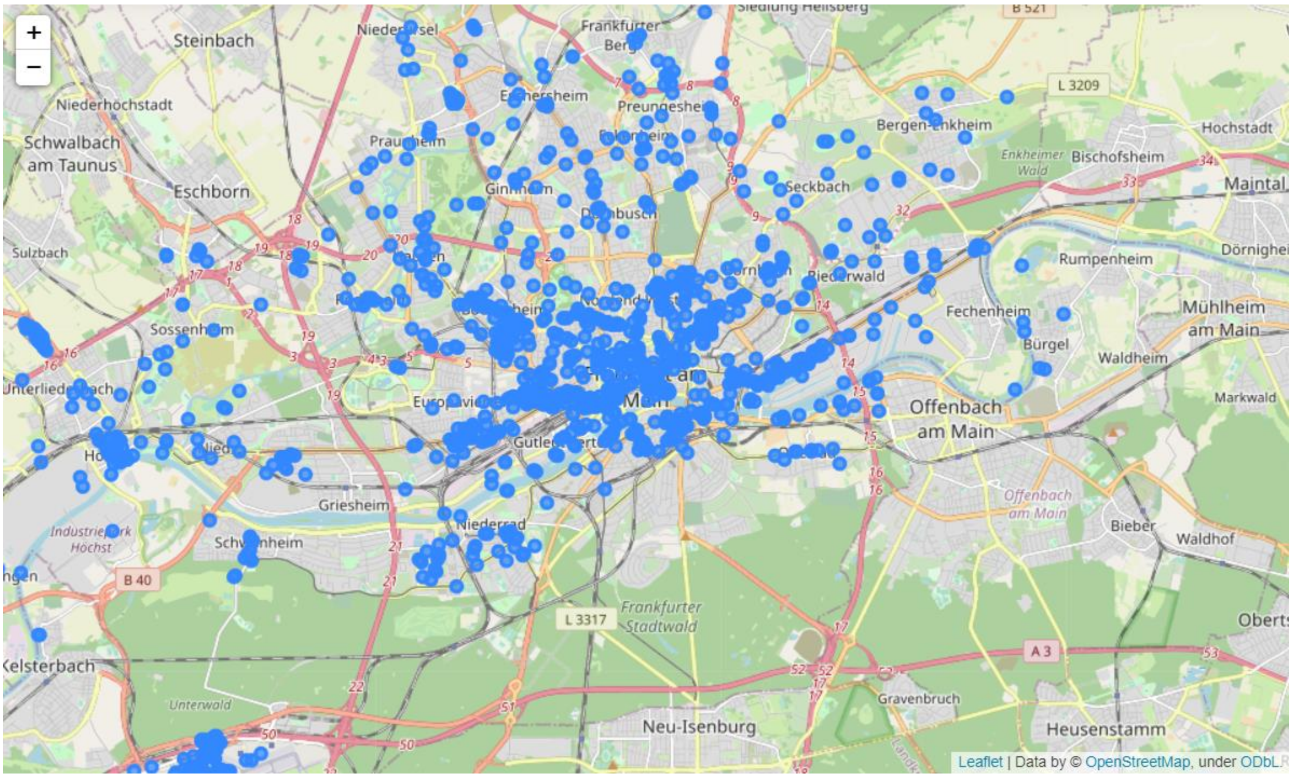

Figure 2 - Map of all the locations extracted with the Foursquare API

2. Methodology

In this step, the neighborhood data have been explored in depth. The top five most common venues are printed for each district, after the table and its venue categories were one hot encoded and grouped by neighborhhod name. This step is necessary, because the K-means algorithm can only read numerical data. The result are visible in more depth in the attached Jupyter notebook. The number of clusters must be set before training the model. After various attempt, I have found that the most appropriate number of clusters for this problem is 5. The resulting cluser labels are then stored in the dataframe and visualized in the interactive map.

	Neighborhood	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	Apple Wine Pub	German Restaurant	Plaza	Bar	Scenic Lookout	Thai Restaurant	Falafel Restaurant	Electronics Store	Italian Restaurant
1	Bergen-Enkheim	Trail	Taverna	Water Park	Plaza	Ice Cream Shop	Paper / Office Supplies Store	Italian Restaurant	German Restaurant	Wine Shop	Donut Shop
2	Berkersheim	Bakery	Train Station	Hotel	Soccer Field	Light Rail Station	Pharmacy	Bus Stop	Supermarket	German Restaurant	Duty-free Shop
3	Bockenheim	Italian Restaurant	Café	Asian Restaurant	Botanical Garden	Bakery	Wine Bar	Bar	Spanish Restaurant	Pizza Place	Japanese Restaurant
4	Bonames	Café	Italian Restaurant	Metro Station	Event Service	Electronics Store	Burger Joint	Garden Center	Doner Restaurant	Athletics & Sports	Golf Course

Figure 3 - Overview of the 10 most common venues in each neighborood

3. Results and Conclusion

The best way to visualize the results of the clustering, is to plot the data on a folium dynamical map, which has been attached below for reference. Intuitively, the outcome makes sense when compared to anecdotal experience. The neighborhoods marked in green belong to Cluster 0, and their most popular venues are cafè, ethnical restaurants and clubs. In the clusters marked in purple instead, the most popular venues are hotels, German restaurants, and public transportation stations. The other clusters are way smaller, and their related neighborhoods contain unique characteristics that distinguish them (i.e: the airport). This project has shown that with a simple algorithm, it is possible to provide end user with a powerful support tool for decision making.

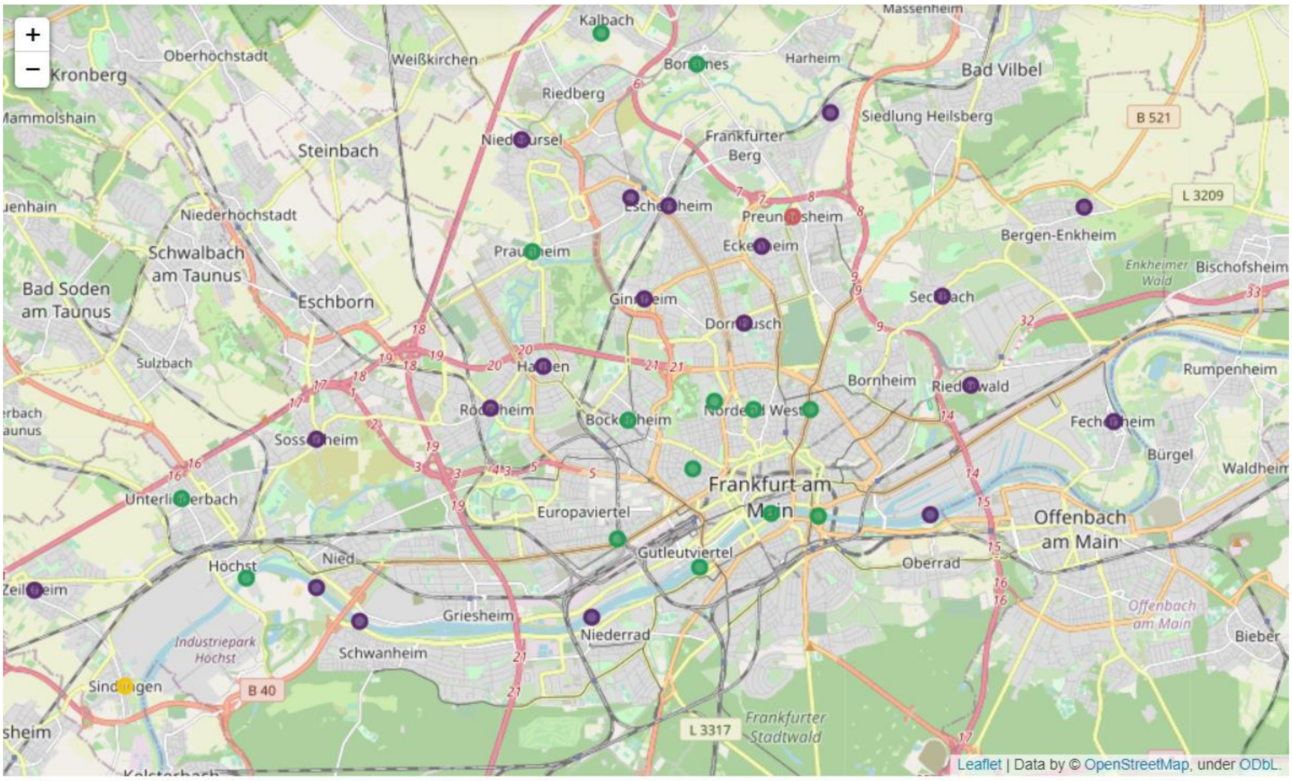

Figure 4 - Clustered Frankfurt neighborhoods