Finding the best location for opening a beer hall in Berlin

Introduction

1.1 Background

The city of Berlin is the capital of Germany, with a population exceeding 4 million inhabitants and continues to grow dynamically. It is known for a bustling cultural scene, as well as it's tech industry and advanced science environment. Due to the international and mostly young population, a multitude of gastronomic venues is present, ranging from fine local dining to diverse street food. On the same line, bars and pubs generate a lot of money, with great outlooks as the wealth of the region is growing disproportionally strong.

1.2 Problem

Despite the multitude of choices for drinking venues, a rather surprising fact becomes obvious to the international visitor trying to find authentic German experiences: There is not a single, folksy beer hall in the city. Therefore, in this report, an analysis of the Berlin neighbourhoods is performed with the goal of finding out the optimal spot for constructing such a venue. Here, optimal is defined as usual in business decisions, as the spot with the highest potential of creating high revenue.

1.3 Stakeholders

The highest interest in this project naturally is held by internationals, both tourists, students and long term residents. Furthermore, the local population is seen as another important factor, especially younger people as they tend to spend more money on beer than older persons.

Data

2.1 Sources

The location data is obtained from Foursquare:

https://www.foursquare.com/

further data points are obtained from Wikipedia:
https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin

and the statistics provider Statista:
https://de.statista.com/statistik/daten/studie/259905/umfrage/mietpreise-in-berlin-nach-bezirken/

2.2 Processing

The borough data was extracted from the Wikipedia website by using the Python BeautifulSoup4 package. The borough table and associated information such as inhabitants, area and density were stored in a Pandas DataFrame. Headers and empty lines were removed and numbers which were stored as strings were converted into floats.

Population data from Wikipedia/Statista was combined with the location data into a single data frame.

As direct income data for each borough is not available, a proxy is used, which in this case is the average income tax payed per borough, data which is available and is approximately linearly proportional to the income.

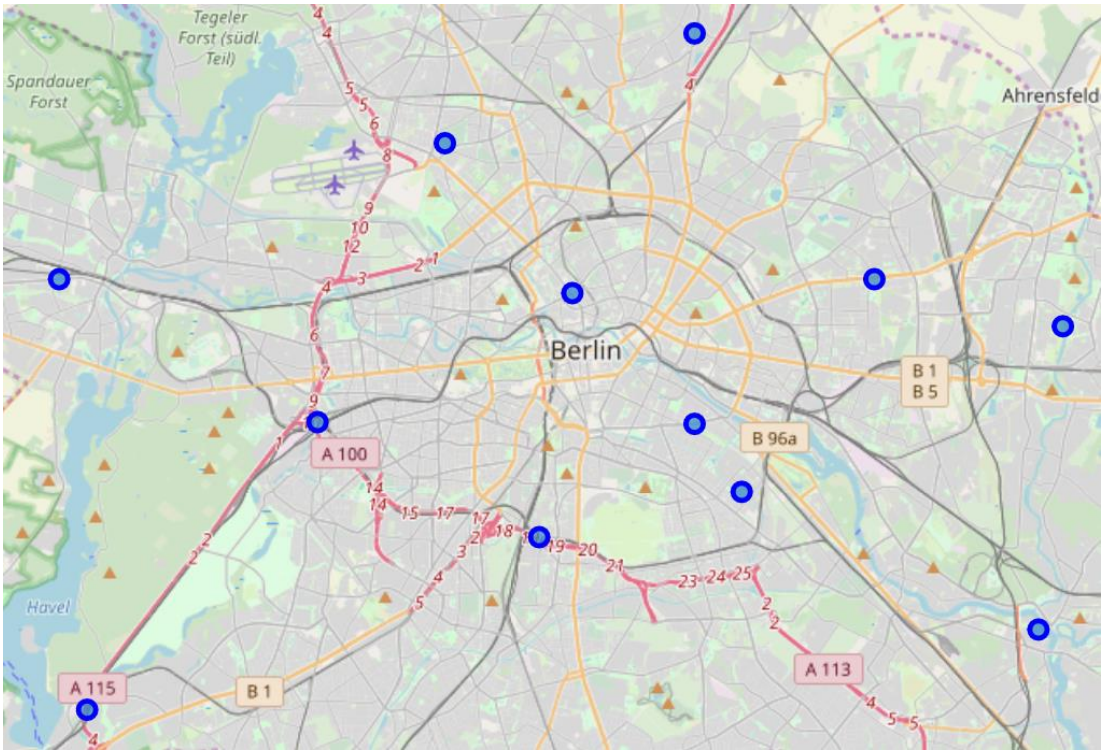
Location data visualization and analysis

The table of sorted and processed Berlin neighbourhoods was first inspected visually. Here it already becomes apparent, that the neighbourhoods differ heavily both in total population and area, leading to an up to ten fold difference in population density among the neighbourhoods, as depicted in tab.1.

Borough	Population	Area	Density	Latitude	Longitude
Charlottenburg-Wilmersdorf	319628	64.72	4878	52.500000	13.283333
Friedrichshain-Kreuzberg	268225	20.16	13187	52.499567	13.431419
Lichtenberg	259881	52.29	4952	52.534306	13.502326
Marzahn-Hellersdorf	248264	61.74	4046	52.522935	13.576597
Mitte	332919	39.47	8272	52.530644	13.383068
Neukölln	310283	44.93	6804	52.483333	13.450000
Pankow	366441	103.01	3476	52.592879	13.431700
Reinickendorf	240454	89.46	2712	52.566667	13.333333
Spandau	223962	91.91	2441	52.534080	13.181716
Steglitz-Zehlendorf	293989	102.5	2818	52.430884	13.192662
Tempelhof-Schöneberg	335060	53.09	6256	52.472160	13.370287
Treptow-Köpenick	241335	168.42	1406	52.450000	13.566667

Tab. 1: Overview over the Berlin neighbourhoods, geographic location as well as population data

The geographic position is best understood using a map (map1).



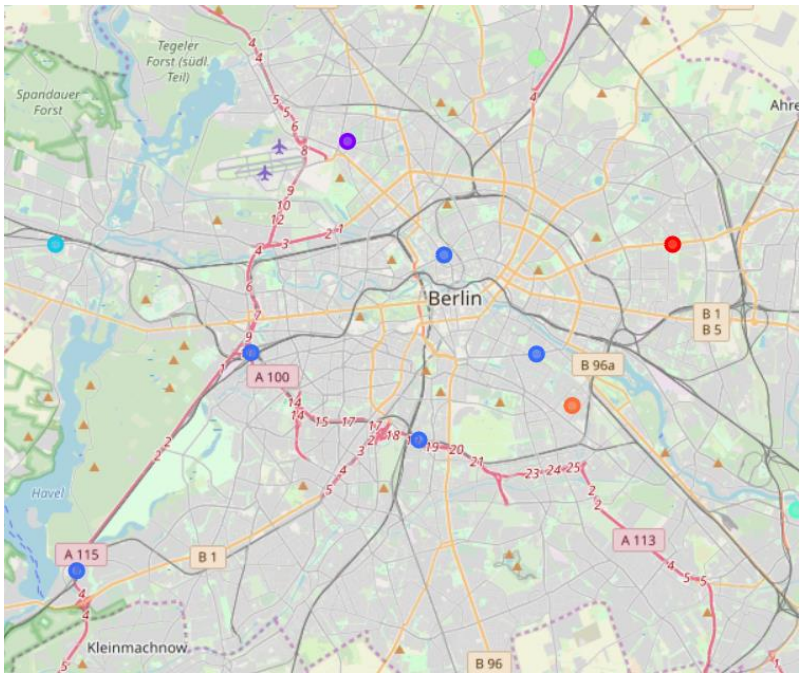
Map1: An overview over the location of the neighbourhoods

The neighbourhood data was subsequently enriched with data from Foursquare, leading to a dataset of venues of interest in a given neighbourhood (tab2). This dataset groups categories of venues for each neighbourhood into the ten most common occurances.

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
Charlottenburg-Wilmersdorf	Automotive Shop	Pet Store	Intersection	Food & Drink Shop	Rest Area	Light Rail Station	Hotel	Steakhouse	Scenic Lookout	Park
Friedrichshain-Kreuzberg	Bar	Café	German Restaurant	Vietnamese Restaurant	Turkish Restaurant	Cocktail Bar	Italian Restaurant	Breakfast Spot	Korean Restaurant	Middle Eastern Restaurant
Lichtenberg	Fast Food Restaurant	Tram Station	Furniture / Home Store	Pool Hall	Supermarket	Hardware Store	Soccer Field	Drugstore	Falafel Restaurant	Farmers Market
Marzahn-Hellersdorf	Stadium	Playground	Athletics & Sports	Soccer Field	Furniture / Home Store	Doner Restaurant	Drugstore	Eastern European Restaurant	Falafel Restaurant	Farmers Market
Mitte	Hotel	Coffee Shop	Café	Italian Restaurant	Bakery	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Drugstore	Plaza	Gym / Fitness Center

Tab2: An overview over most common venues in each neighbourhood

This dataset is then used in a k-means classification algorithm with k = 5 to find clusters of similar neighbourhoods. By this, we can sort neighbourhoods into similar spots. By then further analyzing the cluster of the neighbourhood with the highest possible revenue for a beer hall, one can find low cost but high revenue areas.



Map2: Clusters of similar neighbourhoods based on venue popularity

Prediction

As became obvious from the above analysis, Friedrichshain-Kreuzberg is the venue with bars, cafés and German restaurants are the three most common venues, indicating a population very willing to pay money for such services. Therefore, a beer hall could be successful in this area and similar neighbourhoods. So the cluster of Friedrichshain-Kreuzberg is further analyzed. For this task, the population structure and rent prices are analyzed, creating a model predicting the most promising neighbourhood as described for this project.

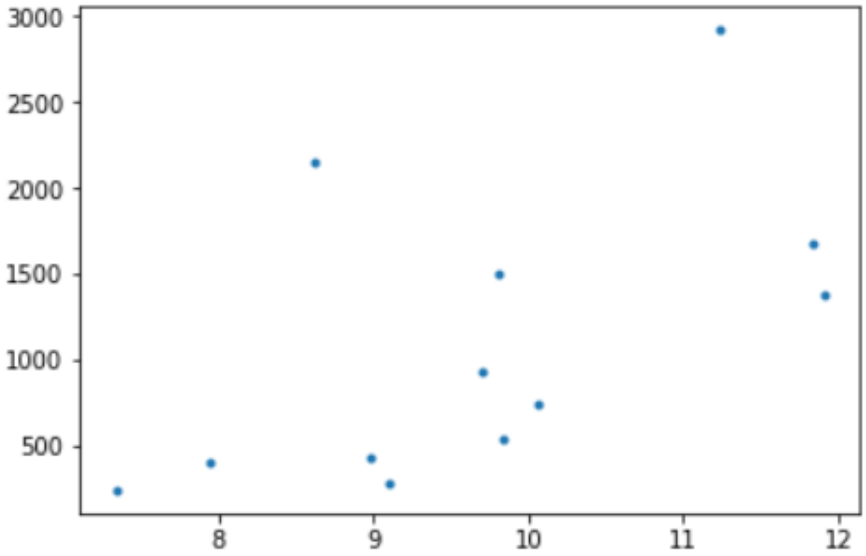


Fig.1: Taxes plotted over average rent prices

When looking at the taxes and the rent prices, one can see an increase in prices with increasing taxes to pay, and therefore income. As a linear relationship does not look likely in this case, an exponential association is assumed. The data is fitted to a simple exponential function, leading to a reasonably good fit for this project (fig.2).

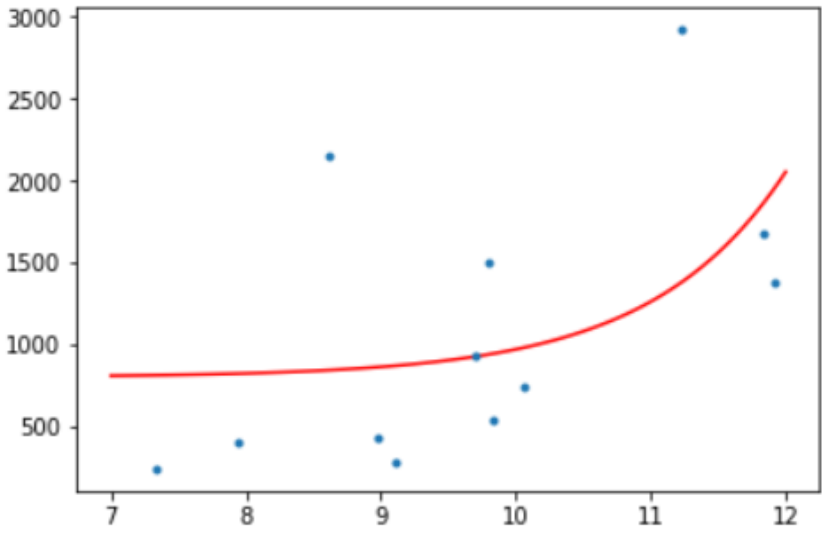


Fig.2: An exponential fit of rent prices over taxes

By looking at this fit, it becomes clear that there are especially two outliers above the fit line, which mean disproportionately high income taxes payed versus rent prices. That means however, that the income of those boroughs seems to be disproportionately high as well. Looking at the raw data, it is obvious that the two boroughs in question are Charlottenburg-Wilmersdorf and Reinickendorf. Comparing this result with the result of the clustering, we see that reinickendorf does not belong in the same cluster as Friedrichshain-Kreuzberg, therefore being not an optimal location for a beer hall. Charlottenburg-Wilmersdorf however belongs to the same cluster as Friedrichshain-Kreuzberg, making it an optimal location for the projects objective.

Conclusion and Discussion

In this project, it was attempted to cluster the boroughs of Berlin, Germany according to their bar, café and restaurant infrastructure using Foursquare data. It became clear that the most popular location for bars and similar venues was Friedrichshain-Kreuzberg. To pick out the most profitable borough from the cluster of boroughs similar to Friedrichshain-Kreuzberg, average rent prices of venues and average income taxes of inhabitants were analyzed. The average income taxes payed were used as a proxy for understanding the general average income structure of a specific borough. By analyzing the data, it became clear that an

exponential function might be the best to describe the relationship between the two variables. After fitting the aforementioned variables to the generalized exponential function, it became clear that the boroughs above the fit line represent the boroughs in which the income is disproportionally high compared to the average rent prices payed for a given venue. The two most clearly observable outliers were Reinickendorf and Charlottenburg-Wilmersdorf, therefore representing good target locations for the planned beer hall. However, as Reinickendorf does not belong to the same cluster as Friedrichshain-Kreuzberg, it can safely be assumed that a bar venue is not too popular in Reinickendorf and therefore this location is not the optimal choice for opening a beer hall. In contrast to that, Charlottenburg-Wilmersdorf is part of the same cluster as Friedrichshain-Kreuzberg, therefore representing the best location of the given boroughs in Berlin for opening a beer hall.

Future analyses might be optimized by including for example the age structure of different boroughs, as can be assumed that certain age ranges are more likely to visit beer halls. If data is available concerning the alcohol habits of different boroughs, this would represent invaluable data to be included in a potential further amelioration of the predictive capacities of the model.