Where to open a bar in Berlin?

Karsten Poddig

October 22, 2019

1. Introduction

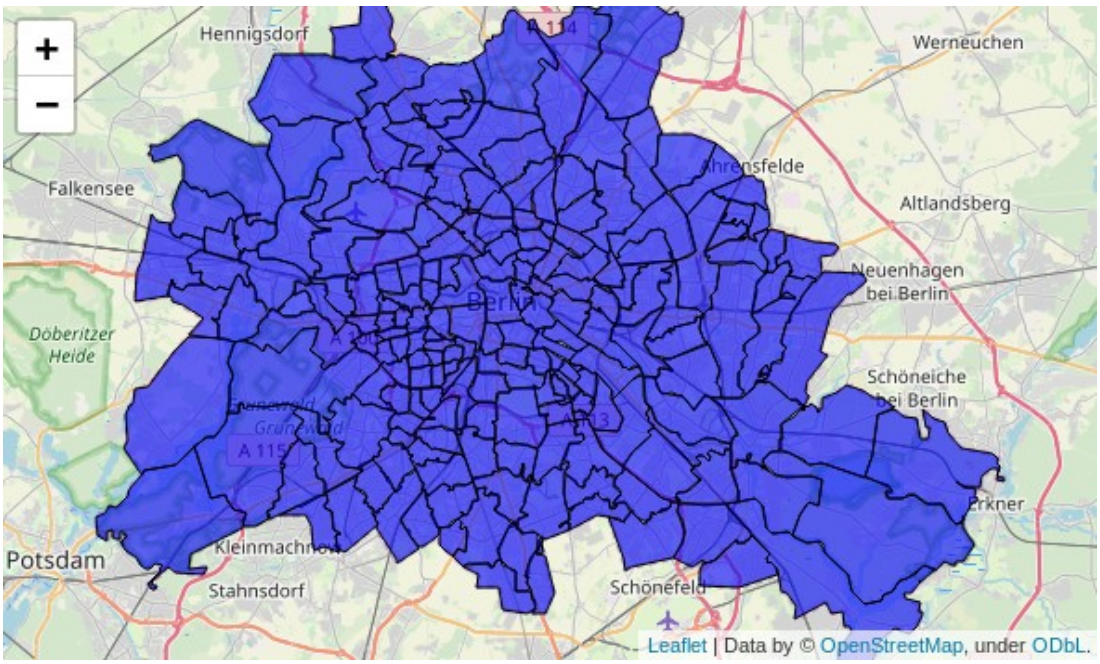
Berlin is probably the biggest city in Europe when it comes to its night life. This analysis deals with the task to find an attractive region in Berlin for opening a bar. Therefore the relationship of bars and venues of other categories is investigated. The motivation comes from the following thoughts: If one wants open a bar in a certain on the hand there shouldn’t be to many bars already. On the other hand it’s of course important to find a region with appropriate other venues in order to make sure there is enough clientele for this bar.

In the next chapters I will develop a model to analyze the relationship of bars and other venues categories. Afterwards the predicted number of bars in each region will be computed. The difference of the predicted and real number of bars shows whether the model would expect more or less bars. The regions where the model expects more bars than there are right now are exactly the regions which could be of interest. Of course if someone want's to open a bar there are a lot of very important factors, which are not inherited in this model - i.e. rent, location (more concret than just the postal code region) and many others. But in the situation if the general region is not determined, this analysis could give a hint which region could be fruitful.

2. Data

2.1. Preparation of Geo Data

Above I simply mentioned regions of Berlin. More concretely these are the postal code regions of Berlin. In a first step I downloaded ‘https://data.technologiestiftung-berlin.de/data/plz/plz.geojson’. This is a geojson file which inherits the information of the postal code regions in Berlin. They are visualized in the following plot.



2.2. Collection of venues

In a next step I will search for venues in Berlin using the Foursquare API. Unfortunately one cannot search all venues with a single query. Therefore for each of the 190 regions a center point is determined. This center point is the mean of all edge coordinates of the boundaries of the regions. These points are used as center for each query. Afterwards all the venues (query results) are combined in a single DataFrame.

Since some center points of the regions are spatially very close some venues are possibly listed multiple times at this stage. Furthermore some venues don't belong to the region of the center point in the regarding region. Hence I deleted all the venues which don't belong to the same postal code region as the center point from the query, which found the venue. This solves both problems. The result is a DataFrame with some thousands of venues and their information, like name, venue category and the region.

One more data cleaning step is the redefinition of some venue categories. Some bars which were found are not listed simply as 'Bar', but as a more specific type of bar, for example 'Wine Bar'. Since I don't want to make a distinction between certain sub categories I'll relabel the categories of some venues as 'Bar'. This is done for venues in one of the categories on the right (besides of 'Salon / Barber Shop'):

Category
Bar
Cocktail Bar
Wine Bar
Hookah Bar
Beer Bar
Gay Bar
Dive Bar
Hotel Bar
Sports Bar
Whisky Bar
Salon / Barbershop
Beach Bar
Piano Bar
Juice Bar
Karaoke Bar

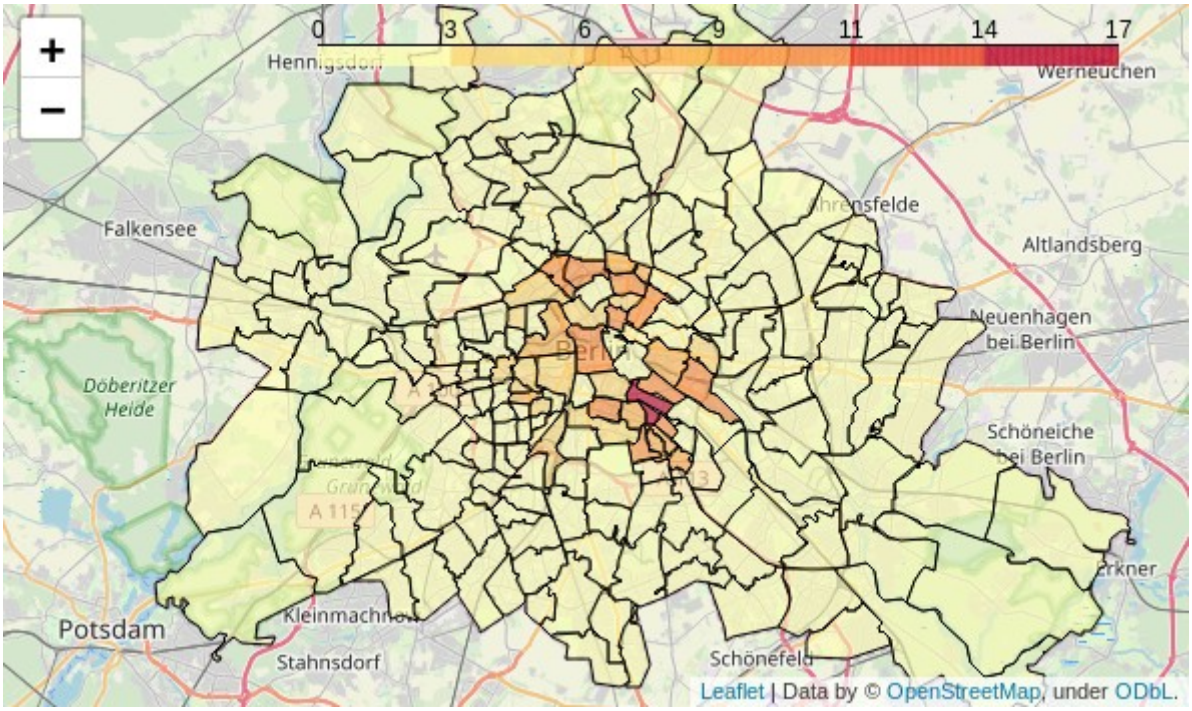
3. Methodology

3.1. Visualization of Query Results

Since it's the aim to find a good model to predict the number of bars in the postal code regions, several plots of this variable are provided.

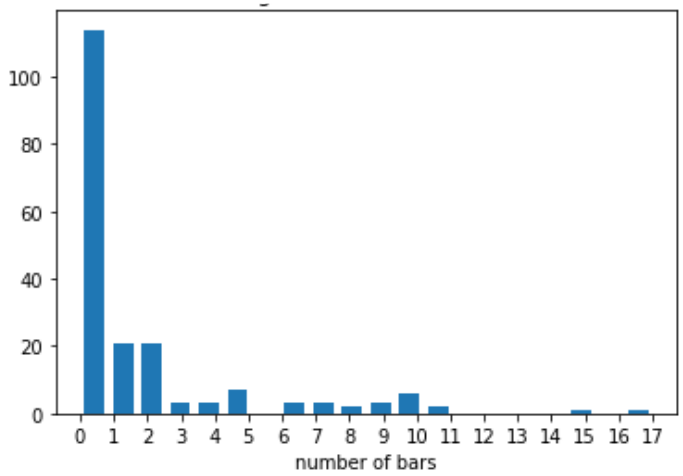
First the number of bars in each region is computed and printed (in descending order):

The following spatial plot shows the number of bars in each region:



As one would expect most of the bars are located in the inner city in Berlin.

The next plot provides a histogram of number of bars per region.



3.2. Application of Neural Network

In this chapter it comes to building the regression model to predict the number of bars in each region. Since it's the intention to predict this entity it will serve as independent variable in the model. The number of venues of other categories will be used as dependent variable, i.e. the number of Restaurants, Cafés, Plazas, etc. At this stage there still about a few hundred different categories contained in the DataFrame. Since there are only 190 samples (one for each region) the amount of different categories cannot be to large. I'll use only the top 20 most frequent categories of venues in Berlin. These categories are listed on the right (in the blue frame).

Category	
Supermarket	322
Café	291
Italian Restaurant	226
Hotel	162
Bakery	159
Bar	154
Coffee Shop	124
Ice Cream Shop	117
German Restaurant	115
Bus Stop	113
Park	106
Vietnamese Restaurant	103
Drugstore	89
Plaza	79
Restaurant	78
Pizza Place	71
Gym / Fitness Center	64
Doner Restaurant	60
Asian Restaurant	57
Cocktail Bar	54
Organic Grocery	51
Tram Station	49
Greek Restaurant	48

In a next step one counts the number of Bars and venues of the categories on the left in each region. The model which I am using is a Neural Network with one layer and the rectified linear unit (relu) as activation function. This model is very similar to a linear model. The only difference is the maximization to zero after the application of the first and only layer.

The model is trained with the independent and dependent variable. Furthermore I used the Adams optimizer, a learning rate of 0.04 and 80 epochs.

4. Results

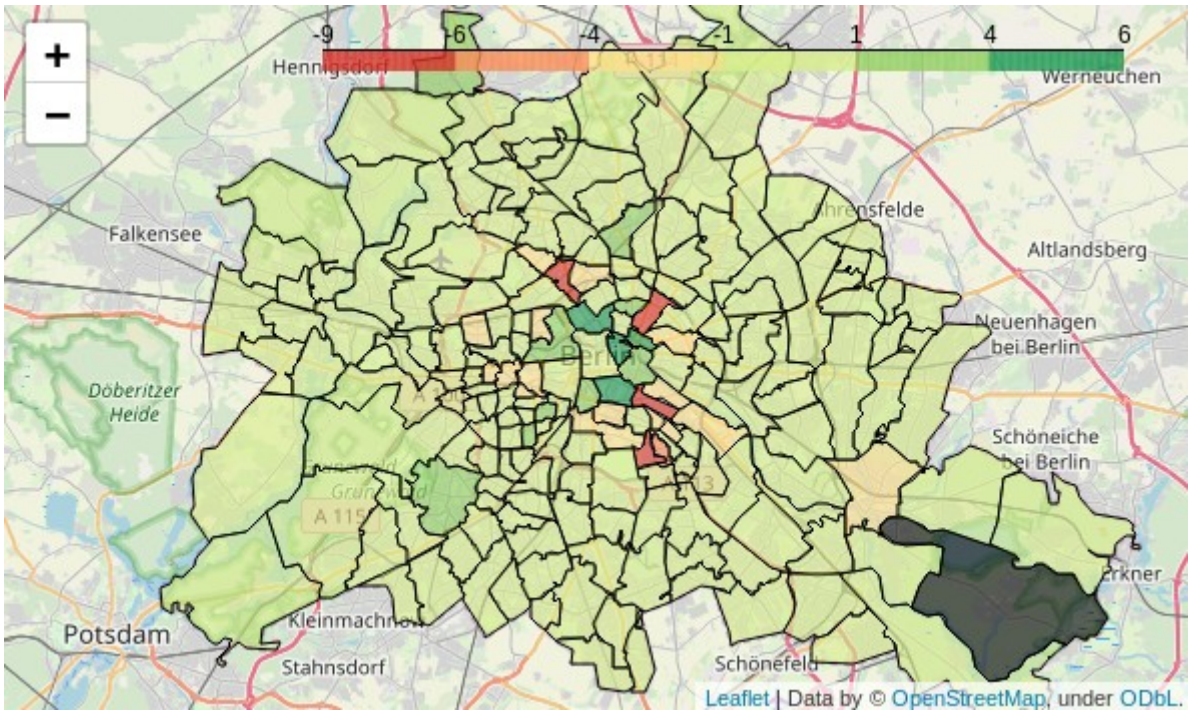
The model in the above setting achieves a mean absolute error (MAE) of 0.74.

Let’s take a look at the calibrated weights of the model:

Regarding the model positive coefficients have an increasing effec on the number of bars. This holds for Cafés, Coffee Shops, Hotels and Plazas. While Bakeries, Asian Restaurants and Gyms seem to have a decreasing effect on the number of bars.

venue_category	coef
Asian Restaurant	-0.471258
Bakery	-0.175684
Bus Stop	-2.578593
Café	1.059286
Coffee Shop	1.648937
Doner Restaurant	0.522624
Drugstore	-0.222207
German Restaurant	0.013230
Gym / Fitness Center	-0.432473
Hotel	0.151093
Ice Cream Shop	-0.113445
Italian Restaurant	0.179842
Park	0.631692
Pizza Place	0.016916
Plaza	0.992595
Restaurant	-0.263432
Supermarket	-0.782066
Vietnamese Restaurant	-0.535420

Finally the next plot provides the graphical result: It shows the difference of the predicted and the real number of bars in each region. As we can there are some regions in the inner city where the regression model predicts up to 6 bars more than there actually are.



The regions with the highest difference are:

postal_code	y	y_pred_nn	y_diff
10115	5	11.183410	6.183410
10178	0	4.687673	4.687673
10969	3	6.689893	3.689893
10435	5	8.190502	3.190502
10963	1	4.188982	3.188982

The next table shows the number of venues in each category for these regions.

Indeed it is striking that in these regions the real number of bars is relatively small compared to the number of other venues. In particular the postal code region 10115 has 3 Cafés, 7 Coffee Shops, multiple Restaurants and only 5 bars.

postal_code	10115	10178	10969	10435	10963
Asian Restaurant	0	1	0	0	1
Bakery	1	0	2	1	1
Bar	5	0	3	5	1
Bus Stop	0	0	0	0	0
Café	3	1	2	5	3
Coffee Shop	7	4	3	2	1
Doner Restaurant	0	0	0	2	0
Drugstore	1	0	0	0	0
German Restaurant	1	1	1	3	0
Gym / Fitness Center	0	1	1	1	1
Hotel	4	3	3	1	4
Ice Cream Shop	2	2	0	1	0
Italian Restaurant	3	0	2	0	0
Park	0	0	2	0	2
Pizza Place	0	0	1	3	0
Plaza	0	1	0	1	0
Restaurant	4	2	0	1	1
Supermarket	1	0	0	0	0
Vietnamese Restaurant	2	3	1	1	0

5. Discussion

t is worth to mention that this analysis still has some weaknesses. The first one is that even in Berlin the list of venues which is found by the Foursquare API is not close to being complete. A lot of venues are simply not found. Of course other APIs with more venues provide more data and therefore would arguably yield more stable results. In particular for the above model there are 20 parameters fitted to just 190 observations. This is a rate of 190/20 roughly 10 observations per parameters, which is still quite low. Therefore there is good chance that the above model is overfitted. The most important action to eliminate this overfitting and build a better model simply more data is needed. Since we are restricted to the Foursquare API we remain which this conclusion in this analysis.

6. Conclusion

As it was described in the introduction I built a regression model which predicts the number of bars for each postal code region in Berlin and identified the regions with the highest difference of predicted and real number of bars. In chapter 4 the top five regions are listed. These regions have indeed a low number of bars compared to the number of other venues.