The perfect location for

Berlins next climbing hall

Data Science specialization capstone project

Fabian Friese

18.10.2019

1 Introduction

1.1 Background

Climbing is getting more and more popular as a mass sport. In Germany the national climbing association, the Deutscher Alpenverein (DAV), is the fifth biggest German sports association in terms of members. 1.054.000 Germans are currently members of the DAV, which is about 1,3 % of the population of Germany. That means that in Germany climbing sports are more popular than the quite popular German Handball or track and field athletics. While in the past the climbing sport was only popular among some passionate mountaineers, nowadays climbing is prevalent among the crowd, including businessmen, families, students, senior citizens and even children.

Of course many climbers live near mountains and pursue their sport in the outdoors. In the south of Germany the are mountain regions like the Alps and many climbers are organized in climbing clubs in the south of Germany. However, in order to allow also Germans living in the northern parts of Germany, which consist of rather flat geography, to get access to the climbing sport, indoor climbing walls in climbing halls are build.

In the last few years a significant increase in the number of climbers could be observed. In the last decade the number of memberships in the DAV increased by 47,8 %, while the population in Germany remained more or less unchanged. The increasing trend of climbers results in a need for the construction of new climbing halls, especially in the flat parts of Germany, like in Germanys capital Berlin.

1.2 Business problem

Constructing and maintaining a climbing hall cause substantial costs. Therefore entrance fees are significantly higher than for regular gyms, swimming pools or other sports venues. The choice of an appropriate position for a new climbing hall is of tremendous importance, because a lack of customers could lead to bankruptcy in short time. This project aims to find a methodology to find an appropriate location for a new climbing hall in Berlin.

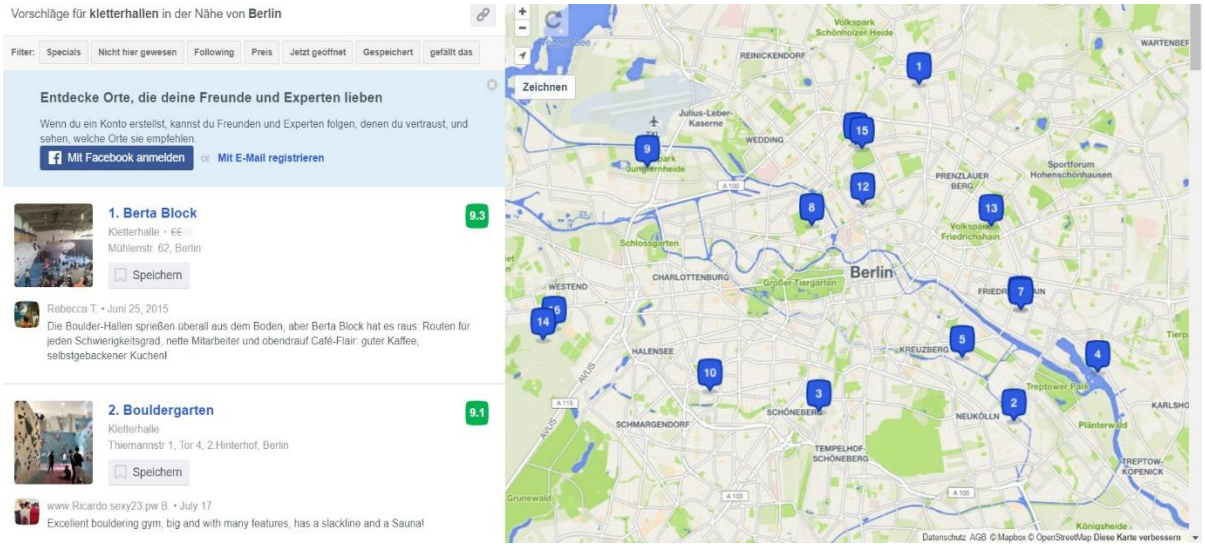
1.3 Interest

Start-ups and businesspeople interested in investing in a climbing hall are faced with the problem to find an appropriate location for the new sports venue. Very passionate climbers might also be interested to learn about climbing venue information when thinking about relocation.

2 Datasources

2.1 Foursquare-Data

The most important data source is going to be Foursquare. Foursquare will provide the study with all the necessary data for both climbing halls and other venues of interest, which could influence the success of a potentially constructed climbing hall. The Foursquare-API is used to retrieve the needed data. Figure 1 shows an exemplary screenshot of search results regarding existing climbing halls in and around Berlin.



2.2 Additional Data

Apart from the Foursquare data other sources of information might be used. Information about the districts and neighbourhoods in Berlin might be helpful to predict potential customers for different locations. Interesting information including data about the population, income, purchasing power, educational level (measured in terms of proximity to universities and highschools) might be taken into account. Those data might be downloaded from the internet or extracted using methods of webscraping, for example scraping data from Wikipedia.

3 Data acquisition and exploration

3.1 Getting the neighbourhoods from Wikipedia

We aim to find good locations for climbing halls in Berlin. The question to be answered is which neighbourhood is suitable for opening a new climbing hall. Therefore, first of all we need information about the neighbourhoods in Berlin. Wikipedia provides us with all the necessary information. https://de.wikipedia.org/wiki/Liste_der_Bezirke_und_Ortsteile_Berlins#Ortsteile shows a table with all neighbourhoods of Berlin as well as information about the area and inhabitants for each neighbourhood. This additional information might be useful in subsequent steps.

We extract the table from Wikipedia applying the web scraping methodologies from Beautiful Soup. As a result, we get lists with all the information from Wikipedia which we now can use in Python.

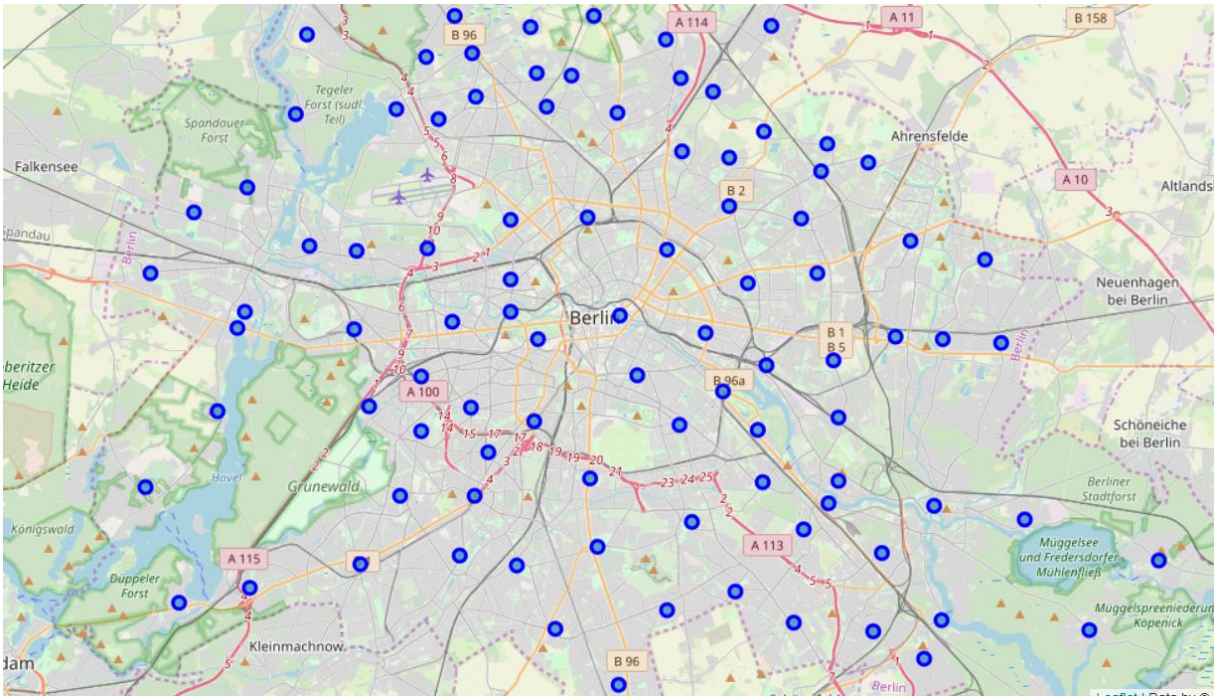
3.2 Finding the coordinates for all neighbourhoods

In the next step we use geocoder to find the corresponding coordinates to all the neighbourhoods. We apply the Nominatim method from geopy.geocoders to get both the longitude and the latitude for all neighbourhoods. The resulting lists are then used to construct a Pandas dataframe with all the information. The resulting dataframe contains 96 rows and 8 columns, which indicates that the we have information in 8 categories for 96 neighbourhoods. The following table shows the first entries of this dataframe.

	Number	Neighbourhood	Borough	Area	Inhabitants	Inhabitants per sqkm	Longitude	Latitude
0	0101	Mitte	Mitte	10,7	99.998	9346	13.402376	52.517690
1	0102	Moabit	Mitte	7,72	78.491	10.167	13.342542	52.530102
2	0103	Hansaviertel	Mitte	0,53	5.831	11.002	13.341872	52.519123
3	0104	Tiergarten	Mitte	5,17	14.529	2810	13.357260	52.509778
4	0105	Wedding	Mitte	9,23	86.468	9368	13.341970	52.550123
5	0106	Gesundbrunnen	Mitte	6,13	94.293	15.382	13.384846	52.550920
6	0201	Friedrichshain	Friedrichshain-Kreuzberg	9,78	131.953	13.492	13.450290	52.512215
7	0202	Kreuzberg	Friedrichshain-Kreuzberg	10,4	154.010	14.809	13.411914	52.497644

3.3 Visualizing the neighbourhoods of Berlin

With the information we have so far, we are now able to construct a map of Berlin showing all the neighbourhoods. For the generation of the map we use the folium package. The following figure shows this map with all the neighbourhoods of Berlin.



3.4 Getting information for venues in Berlin

Now we use the Foursquare API to get information about the venues in each Neighbourhood. This is a step where we have to be extremely careful, as the parameters we choose for this task will highly influence the quality of results. The most important parameter is the radius from the centre of the neighbourhoods, in which we search for venues. If we choose this parameter too small there might be venues which are not assigned to any neighbourhoods, while a too big radius might lead to venues that are assigned to multiple neighbourhoods.

First we choose a quite high radius of 10000 m to make sure that we find all climbing halls. The following figure shows the first entries of the resulting dataframe.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Wedding	52.550123	13.341970	Magic Mountain Kletterhalle	52.548613	13.381902	Climbing Gym
Wedding	52.550123	13.341970	Waldhochseilgarten Jungfernheide	52.543001	13.290644	Rock Climbing Spot
Gesundbrunnen	52.550920	13.384846	Berta Block	52.563548	13.409754	Climbing Gym
Weißensee	52.554619	13.463002	Berta Block	52.563548	13.409754	Climbing Gym
Blankenburg	52.593211	13.454182	Berta Block	52.563548	13.409754	Climbing Gym
Heinersdorf	52.572825	13.437015	Berta Block	52.563548	13.409754	Climbing Gym
Stadtrandsiedlung Malchow	52.571019	13.463285	Berta Block	52.563548	13.409754	Climbing Gym
Pankow	52.597663	13.436351	Berta Block	52.563548	13.409754	Climbing Gym
Niederschönhausen	52.585806	13.401397	Berta Block	52.563548	13.409754	Climbing Gym
Niederschönhausen	52.585806	13.401397	Magic Mountain Kletterhalle	52.548613	13.381902	Climbing Gym
Rosenthal	52.598319	13.375519	Berta Block	52.563548	13.409754	Climbing Gym
Wilhelmsruh	52.588012	13.362206	Berta Block	52.563548	13.409754	Climbing Gym

First of all we realize, that some climbing halls are categorized as “Climbing Gym” while others are found in the category “Rock Climbing Spot”. Therefore we have to include both this categories when we search for climbing halls.

We also see, that the high radius led to a situation where for example the climbing hall “Berta Block” was assigned to many neighborhoods. However, this step allowed us to identify all the climbing halls listed in fourquare by applying the function “unique”. There are 9 climbing halls listed in Foursquare.

Now we try to get a dataframe where each climbing hall is assigned to only one neighborhood. Therefore we try to set the radius to 500m. The following figure shows the results.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Gesundbrunnen	52.550920	13.384846	Magic Mountain Kletterhalle	52.548613	13.381902	Climbing Gym
Wilmersdorf	52.487115	13.320330	Boulderworx	52.486881	13.318215	Climbing Gym
Charlottenburg-Nord	52.540525	13.296266	Waldhochseilgarten Jungfernheide	52.543001	13.290644	Rock Climbing Spot
Hellersdorf	52.536854	13.604774	BergWerk	52.537724	13.603606	Rock Climbing Spot

It is easy to see that a radius of 500m was too small to find all climbing halls. After trying different options for that parameter, we figured out that a radius of 1750 m is a good choice as the following figure illustrates.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Moabit	52.530102	13.342542	DAV Kletterzentrum Berlin	52.528419	13.362783	Climbing Gym
Gesundbrunnen	52.550920	13.384846	Magic Mountain Kletterhalle	52.548613	13.381902	Climbing Gym
Friedrichshain	52.512215	13.450290	Der Kegel	52.507287	13.454390	Climbing Gym
Charlottenburg-Nord	52.540525	13.296266	Waldhochseilgarten Jungfernheide	52.543001	13.290644	Rock Climbing Spot
Schöneberg	52.482157	13.355190	Bright Site	52.481732	13.365768	Climbing Gym
Mariendorf	52.440080	13.390028	Südbloc Boulderhalle	52.439969	13.379566	Climbing Gym
Alt-Treptow	52.492563	13.459874	Bouldergarten	52.479470	13.451368	Climbing Gym
Plänterwald	52.479544	13.478808	Ostbloc Boulderhalle	52.491496	13.487964	Climbing Gym
Hellersdorf	52.536854	13.604774	BergWerk	52.537724	13.603606	Rock Climbing Spot
Rummelsburg	52.501370	13.483514	Ostbloc Boulderhalle	52.491496	13.487964	Climbing Gym

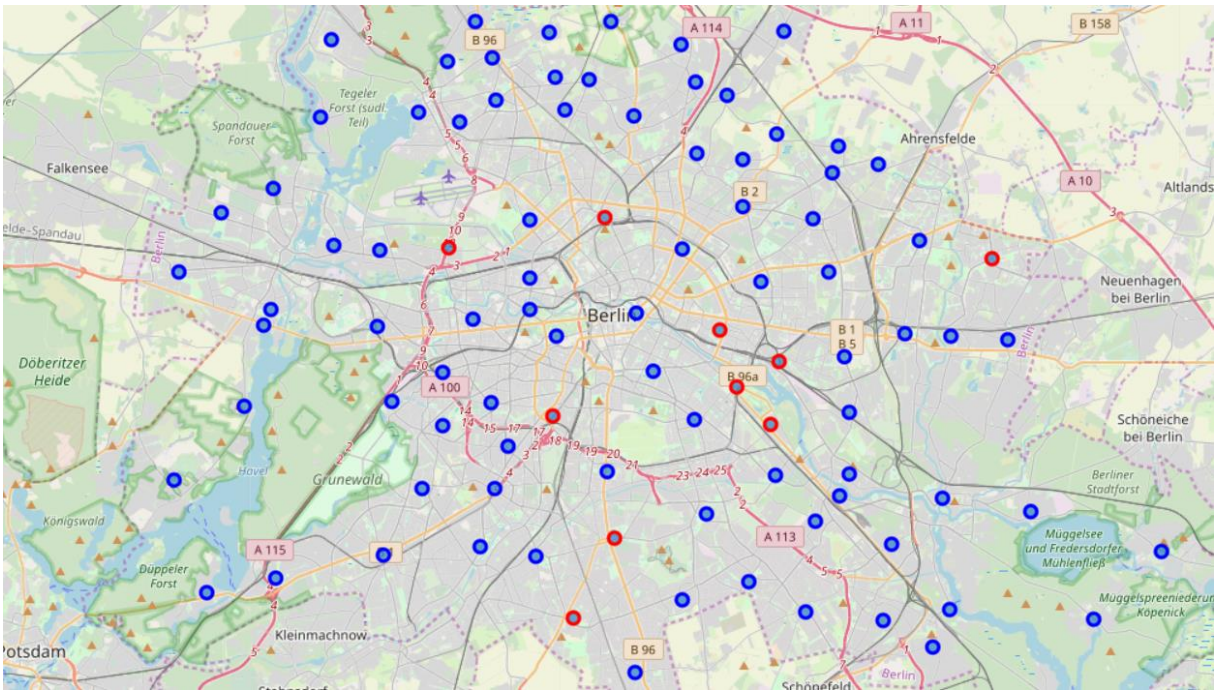
With this parameter choice we were able to find all the climbing halls. Only one venue, “Ostbloc Boulderhalle”, was listed twice. We therefore cleaned the data manually by comparing the coordinates of the venue and the two assigned neighbourhoods. As “Ostcbloc Boulderhalle” is much closer to “Rummelsburg” than to “Plänterwald” we deleted the duplicate entry assigning it to Plänterwald. The following figure illustrates this.

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Moabit	52.530102	13.342542	DAV Kletterzentrum Berlin	52.528419	13.362783	Climbing Gym
Gesundbrunnen	52.550920	13.384846	Magic Mountain Kletterhalle	52.548613	13.381902	Climbing Gym
Friedrichshain	52.512215	13.450290	Der Kegel	52.507287	13.454390	Climbing Gym
Charlottenburg-Nord	52.540525	13.296266	Waldhochseilgarten Jungfernheide	52.543001	13.290644	Rock Climbing Spot
Schöneberg	52.482157	13.355190	Bright Site	52.481732	13.365768	Climbing Gym
Mariendorf	52.440080	13.390028	Südbloc Boulderhalle	52.439969	13.379566	Climbing Gym
Alt-Treptow	52.492563	13.459874	Bouldergarten	52.479470	13.451368	Climbing Gym
Plänterwald	52.479544	13.478808	Ostbloc Boulderhalle	52.491496	13.487964	Climbing Gym
Hellersdorf	52.536854	13.604774	BergWerk	52.537724	13.603606	Rock Climbing Spot
Rummelsburg	52.501370	13.483514	Ostbloc Boulderhalle	52.491496	13.487964	Climbing Gym

We are also interested in inspecting the total results of Foursquare enquiry. Getting the shape of the resulting dataframe shows, that in total 4722 venues could be found.

3.5.Finding the neighbourhoods with existing climbing halls

By filtering the dataframe with all the venues we can create a list of the names of neighbourhoods with existing climbing halls. This allows us to update the map of the neighbourhoods in Berlin. The following map shows neighbourhoods with existing climbing halls in red while neighbourhoods without climbing halls remain blue.



4 Finding suitable locations

Now we want to answer the question, which neighbourhoods are suitable for opening a new climbing hall. In order to answer this question, we use three different approaches. The first one is quite simple as it only reports the biggest neighbourhoods in terms of venues. For the second approach we are going to find out which venues are usually close to climbing halls. The third approach is the most elaborate one as it applies machine learning algorithms in order to decide where to build new climbing halls.

4.1 Finding the biggest neighbourhoods

Although this approach is quite simple it has some practical relevance. Here we just find the neighbourhoods with the highest density of existing venues of all categories. The idea behind this approach is, that thereby we can identify areas, where people dwell in their leisure time. These are often areas with commercial centres or sports complexes. Opening a climbing hall in these areas might be a smart choice as it is highly possible that customers will combine the visit of the climbing hall with a visit of a café to chat to their climbing buddies afterwards. It might be an appropriate approach to identify area with high purchasing powers as well.

In order to find those biggest neighbourhoods, we just group the dataframe by neighbourhood and count the number of entries. The following abstract of the

resulting table shows, that there are 19 neighbourhoods where the limit of 100 venues per neighbourhood was binding.

Neighborhood	
Gesundbrunnen	100
Tiergarten	100
Halensee	100
Hansaviertel	100
Rummelsburg	100
Prenzlauer Berg	100
Friedrichshain	100
Friedenau	100
Steglitz	100
Fennpfuhl	100
Charlottenburg	100
Neukölln	100
Wedding	100
Moabit	100
Alt-Treptow	100
Mitte	100
Wilmersdorf	100
Kreuzberg	100
Schöneberg	100
Tempelhof	98
Borsigwalde	82
Westend	79

Therefore, we tried to refine the results by increasing the maximum possible number of found venues per neighbourhood. However, this enquiry exceeded the limits of the Foursquare free plan and therefore we were not able to conduct this study.

Using the results shown in the table we compared the locations of the current climbing halls and found out, that 6 of the 9 climbing halls are located in one of those Top 20 neighbourhoods. This study can be used to narrow down the list of possible candidates of neighbourhoods for a new climbing hall to reduce the effort for computational more expensive studies.

4.2 Finding the venues close to climbing halls

In this approach we are going to find out, which venues are usually close to climbing halls and which in particular are not. In order to achieve this, we split our dataframe to get a subset which contains all the venues which are located in a neighbourhood with an existing climbing hall.

Now we want to get the 10 most common venues in both the neighbourhoods with climbing halls only and in the dataset of all the neighbourhoods, including both neighbourhoods with and without climbing halls. The following table shows on the left side the frequencies of existence for the different venue categories for neighbourhoods with climbing halls, while the right table shows the corresponding frequencies when all neighbourhoods are taken into account.

	index	0
36	Café	0.059355
192	Supermarket	0.046452
14	Bar	0.037419
45	Coffee Shop	0.029677
141	Park	0.024516
136	Nightclub	0.023226
12	Bakery	0.023226
150	Pizza Place	0.020645
112	Italian Restaurant	0.020645
106	Ice Cream Shop	0.018065

	index	0
310	Supermarket	0.086616
59	Café	0.047226
176	Italian Restaurant	0.036213
163	Hotel	0.025413
231	Park	0.025413
140	German Restaurant	0.023931
24	Bakery	0.023719
105	Drugstore	0.022660
54	Bus Stop	0.020330
166	Ice Cream Shop	0.020119

Comparing these tables we see, that in neighbourhoods with climbing halls venues that are visited in the leisure time, such as bars, coffee shops, pizza places and even nightclubs are more present than in the overall dataset, where venues for the everyday life such as hotels, parks, bakerys and drugstores are more relevant.

However, we also see that many venues, such as supermarkets and cafés, are very common in both datasets. In order to use this data for our analysis we are rather interested in the information, which venues are more frequent than on average in neighbourhoods with climbing halls, and which venues are less frequent than the average frequency.

In order to get this information, we need to combine both dataframes with the frequencies to get the difference between both subsets for each venue category. As the subset with only the neighbourhoods with climbing halls do not include all venue categories we therefore perform an outer join and fill all the produced NaNs with zeroes.

Now we can get the differences to see special features of the neighbourhoods with climbing halls. In the following figure the left table contains the Top 10 venue categories that are more common in neighbourhoods with climbing halls, while the right table contains the venues that are less common.

Bar	0.017577	0.037419	0.019842
Nightclub	0.005294	0.023226	0.017931
Café	0.047226	0.059355	0.012129
Coffee Shop	0.019060	0.029677	0.010618
Climbing Gym	0.001694	0.010323	0.008628
Rock Club	0.001694	0.010323	0.008628
Vegetarian / Vegan Restaurant	0.005930	0.012903	0.006974
Pizza Place	0.013765	0.020645	0.006880
Airport Service	0.001271	0.007742	0.006471
Wine Bar	0.004024	0.010323	0.006299

Gas Station	0.007624	0.002581	-0.005043
Plaza	0.011436	0.005161	-0.006275
Trattoria/Osteria	0.012071	0.005161	-0.006910
Greek Restaurant	0.010801	0.003871	-0.006930
Tram Station	0.012706	0.005161	-0.007545
German Restaurant	0.023931	0.012903	-0.011027
Drugstore	0.022660	0.010323	-0.012337
Hotel	0.025413	0.010323	-0.015090
Italian Restaurant	0.036213	0.020645	-0.015568
Supermarket	0.086616	0.046452	-0.040164

We can clearly see another confirmation, that climbing halls are located in areas designated to leisure time, as there are many bars, nightclubs, coffee shops, and rock clubs. The data shows that climbing gyms are also more present in neighbourhoods with climbing halls. How reassuring is this?! In contrast there are much less supermarkets, drugstores or gas stations in that areas.

Now we can use this information about the Top 10 and Bottom 10 venues, as we call them from now on, to calculate a score which allows us to find the areas which have the venues typical for neighbourhoods with climbing halls. We calculate it as follows

$$\begin{aligned} & Score \\ = \; & \sum Frequencies\ of\ Top\ 10\ venues \\ & - \\ & \sum Frequencies\ of\ Bottom\ 10\ venues \\ = \; & Top\ Score - Neg\ Score \end{aligned}$$

Again, we decide to report the Top 20 Neighbourhoods with the highest scores.

Neighbourhood	Pos Score	Neg Score	Tot Score
Friedrichshain	0.330000	0.050000	0.280000
Gesundbrunnen	0.320000	0.040000	0.280000
Neukölln	0.350000	0.070000	0.280000
Alt-Treptow	0.280000	0.040000	0.240000
Rummelsburg	0.300000	0.080000	0.220000
Kreuzberg	0.280000	0.060000	0.220000
Wedding	0.280000	0.110000	0.170000
Blankenfelde	0.142857	0.000000	0.142857
Schöneberg	0.230000	0.100000	0.130000
Prenzlauer Berg	0.190000	0.070000	0.120000
Moabit	0.190000	0.110000	0.080000
Fennpfuhl	0.250000	0.180000	0.070000
Steglitz	0.140000	0.160000	-0.020000
Charlottenburg-Nord	0.157143	0.185714	-0.028571
Friedenau	0.160000	0.210000	-0.050000
Plänterwald	0.134328	0.223881	-0.089552
Mitte	0.070000	0.160000	-0.090000
Hansaviertel	0.100000	0.190000	-0.090000
Dahlem	0.127273	0.218182	-0.090909
Tiergarten	0.080000	0.190000	-0.110000

Six out of nine climbing halls are located in those Top 20 neighbourhoods.

4.3 Classifying neighbourhoods with logistic regression

This time we are use machine learning to find out where we should open a new climbing hall. We use the classification algorithm of logistic regression to choose where to open a new climbing hall without determining any rules that can be applied to guess whether a neighbourhood is suitable for a climbing hall or not.

There is one challenge in applying logistic regression to this problem: We **do** know which neighbourhoods **are** suitable for climbing halls as we **do** know the locations of the existing halls. However, we **do not** know which neighbourhoods are inherently **not** suitable for climbing halls. All neighbourhoods where no climbing halls exist could be either a neighbourhood suitable for a climbing hall, where just nobody constructed a climbing hall yet, or a neighbourhood inherently not suitable. There is no way to tell. Therefore, we don´t have any pristine training data set.

However, we can assume that the market for climbing halls is nearly saturated and there are only few neighbourhoods suitable for climbing halls left, that do not actually locate a climbing hall. Therefore, we can mark all neighbourhoods without climbing halls as unsuitable, knowing that this will only lead to small errors. We can then train the logistic regression using the whole dataset as training data. After this we can classify all neighbourhoods again applying this fitted logistic regression. The results should classify nearly all neighbourhoods with climbing halls as suitable and most of the neighbourhoods without climbing halls as unsuitable. Those neighbourhoods, that do not locate a climbing hall yet but now are classified as suitable, are those we are looking for.

In order to determine those neighbourhoods, we first have to group our dataframe containing all the venues by neighbourhoods and calculate the frequencies of the venue categories. We then add another column stating whether a climbing hall is located in the corresponding neighbourhood or not.

This dataframe is split in X_train and Y_train. X_train contains all the frequencies for all venue categories, while Y_train just contains the information whether there is a climbing hall or not.

After normalizing X_train we can use this data to perform the logistic regression. The results are used to classify all the neighbourhoods once again and to determine the probability that each neighbourhood is suitable for a climbing hall or not.

When we now inspect this classification, we first of all notice, that all the neighbourhoods where climbing halls are already located are classified as suitable. However, some neighbourhoods could be classified as suitable, although there is no climbing hall located there yet. This is good news.

By comparing the classifications with the real data we can calculate the jaccard similarity, the f1-score and the log loss as performance metrics. The following table shows the results.

Jaccard similiarity	0,9375
F1-Score	0,9436
Log Loss	0,6273

The results show, that the classification was quite close to the original data. The deviation from a perfect fit just results from the neighbourhoods that were classified as suitable, although they do not locate any climbing hall yet. This deviation from a perfect fit therefore is desirable.

As the last step we want to know which neighbourhoods should be selected for opening a new climbing hall. We therefore filter the dataframe and get the following 6 neighbourhoods.

```
In [32]: new = yhat - Y_train
         new_neighbourhoods = berlin_grouped["Neighbourhood"].loc[new == 1]
         new_neighbourhoods

Out[32]: 18    Fennpfuhl
         44    Kreuzberg
         56        Mitte
         61     Neukölln
         83    Tempelhof
         88      Wedding
```

4.4 Comparing the results

When we compare the results from the three methodologies we applied, we observe that the logistic regression gives us the most relevant information, as it was able to classify each neighbourhood with existing climbing halls as such. The other two methods are less informative. The following table shows 5 categories of neighbourhoods with different meanings:

Blue: Neighbourhoods with already existing climbing halls. Opening another one might not be interesting.

Red: There is no climbing hall yet but only method 1 or 2 classified this neighbourhood as suitable. Not recommended to open a climbing hall here.

Orange: There is no climbing hall yet and methods 1 and 2 both classified it as suitable, but the logistic regression did not. Low recommendation to open a climbing hall here.

Light green: There is no climbing hall yet and the logistic regression and one other method identified this neighbourhood as suitable. Neighbourhood recommended for opening a climbing hall.

Dark green: There is no climbing hall yet and all methods classified the neighbourhood as suitable. High recommendation to open a climbing hall here.

Neighbourhood	Biggest neighbourhoods	Top 10 / Bottom 10 Venues	Logistic regression	Exist climbinghall
Alt-Treptow	X	X	X	X
Blankenfelde		X		
Charlottenburg	X			
Charlottenburg-Nord		X	X	X
Dahlem		X		
Fennpfuhl	X	X	X	
Friedenau	X	X		
Friedrichshain	X	X	X	X
Gesundbrunnen	X	X	X	X
Halensee	X			
Hansaviertel	X	X		
Hellersdorf			X	X
Kreuzberg	X	X	X	
Mariendorf			X	X
Marienfelde			X	X
Mitte	X	X	X	
Moabit	X	X		
Neukölln	X	X	X	
Plänterwald		X	X	
Prenzlauer Berg	X	X		
Rummelsburg	X	X	X	X
Schöneberg	X	X	X	X
Steglitz	X	X		
Tempelhof	X		X	
Tiergarten	X	X		
Wedding	X	X	X	
Wilmersdorf	X			

The table shows, that neighbourhoods with the highest recommendation to open a climbing hall are Fennpfuhl, Kreuzberg, Mitte, Neukölln and Wedding.

5 Conclusion

The study illustrated the enormous power of quantitative analysis. With only little effort and a few lines of programming code we could apply the methods from data science and machine learning to identify potential candidates for a new climbing hall. The simplicity of this study made quite some simplifying assumptions necessary. We assumed, that the success of a climbing hall depends on the surrounding venues. While it is highly plausible that this assumption also holds in reality, practical examples show that sports venues can still be extraordinary successful, although located in quite remote areas. Nevertheless, especially in the first few month or years the proximity to certain venues might be helpful. A decision taker should not blindly trust in the results of this quantitative study, but it can be used to narrow down the possible locations of a new climbing hall to be constructed.

6 Future directions

All those studies still have quite some potential for improvement. The first approach could be refined by choosing a higher limit of potentially found venues per neighbourhood. However, this would imply upgrading the plan for the Foursquare API, as the free plan limits the maximum number of enquiries.

For the second approach we could enhance the validity of the score be considering more categories than only the Top 10 and Bottom 10. It might be useful to also find some elaborate strategy to determine weights for the impact of venue categories with different importance.

The machine learning approach can be refined by also trying different methods like random forest or ANNs. Potentially also parameter tuning for the existing logistic regression might lead to better results. However, the presented results are quite significant and helpful, therefore we see no real importance of refining this study.

To take a real decision on where to open the climbing hall the expertise of a marketing specialist is also needed. However, this would lead too far in a data scientists practical report.

7 Acknowledgements

I´d like to thank all the teachers of the data science specialization as well as all the peer reviewers :).