Finding the most similar neighborhood to the particular location to build a new branch of successful sport center in Berlin

Narmin Ghaffari Laleh

May, 2020

1. Introduction

1.1 Background

Berlin is the capital and largest city of Germany by both area and population. It is divided into 12 Boroughs and it includes 192 total codes. As one the most crowded cities, there are various entertainment places and running a successful business based on the peoples favor and interests is not a easy task. People are working for long hours in this metropolis city and it is really important to build a comfortable entertainment environment for the rest of their weekly days. One of very successful sport centers in Berlin is Sport-Club Charlottenburg (SCC Berlin) which has a high amount of professional athletes and with high rate of monthly new members.

1.2 Problem

The owner of this sport center is trying to build the second branch of this club in other neighborhood of Berlin. Based on the features of current location, he wants to find out the most similar areas to the current location.

1.3 Interest

This can be interesting for every person which is looking for similar areas based on the available entertainment places in Berlin. It is usually a common question for startup businesses that where is the best location for the project based on the current venues in that location.

2. Data acquisition and cleaning

2.1 Data sources

The data set which is required for this project is obtained from website geonames.org. This data set includes all the postal codes of Berlin with their corresponding latitude and longitude information. For further investigations of the similarities between the different neighborhoods, their area and population can be obtained from postal-codes.cybo.com.

2.2 Data cleaning

The updated data from geonames.org website to the python pandas has several problems and need to be cleaned and organized. Fig 2.1 shows the uploaded data set to the python environment. Based on the original data set in the website, we know, that we need the code from the first row and then Latitude/ Longitude of that postal code in the following row. This manner should be followed for all the rows in the data set. The result of this manipulation is shown in Fig 2.2.

named: 0	Place	Code	Country	Admin1	Admin2	Admin3	Admin4
1.0	Berlin	10117	Germany	Berlin	NaN	Berlin, Stadt	Berlin
NaN	52.517/13.387	52.517/13.387	52.517/13.387	52.517/13.387	52.517/13.387	52.517/13.387	52.517/13.387
2.0	Berlin	10115	Germany	Berlin	NaN	Berlin, Stadt	Berlin
NaN	52.532/13.385	52.532/13.385	52.532/13.385	52.532/13.385	52.532/13.385	52.532/13.385	52.532/13.385
3.0	Berlin	10119	Germany	Berlin	NaN	Berlin, Stadt	Berlin
...	...	...	...	...	...	...	...
194.0	Berlin	14131	Germany	Berlin	NaN	Berlin, Stadt	Berlin
NaN	52.517/13.4	52.517/13.4	52.517/13.4	52.517/13.4	52.517/13.4	52.517/13.4	52.517/13.4
195.0	Reinickendorf	13047	Germany	Berlin	NaN	Berlin, Stadt	Berlin
NaN	52.567/13.333	52.567/13.333	52.567/13.333	52.567/13.333	52.567/13.333	52.567/13.333	52.567/13.333
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fig 2.1 - The uploaded data set in python

	PostalCode	Latitude	Longitude
0	10117	52.517	13.387
1	10115	52.532	13.385
2	10119	52.53	13.405
3	10178	52.521	13.41
4	10179	52.512	13.416

Fig 2.2 - Cleaned data set in the python data frame

The next step is to remove the duplicate postal codes from the data set. Then it is the best time to plot each postal code on the map of Berlin to see their distribution on the map. Fig 2.3 shows the map of Berlin with the corresponding 195 postal codes which are indicator of different neighborhoods in this city.

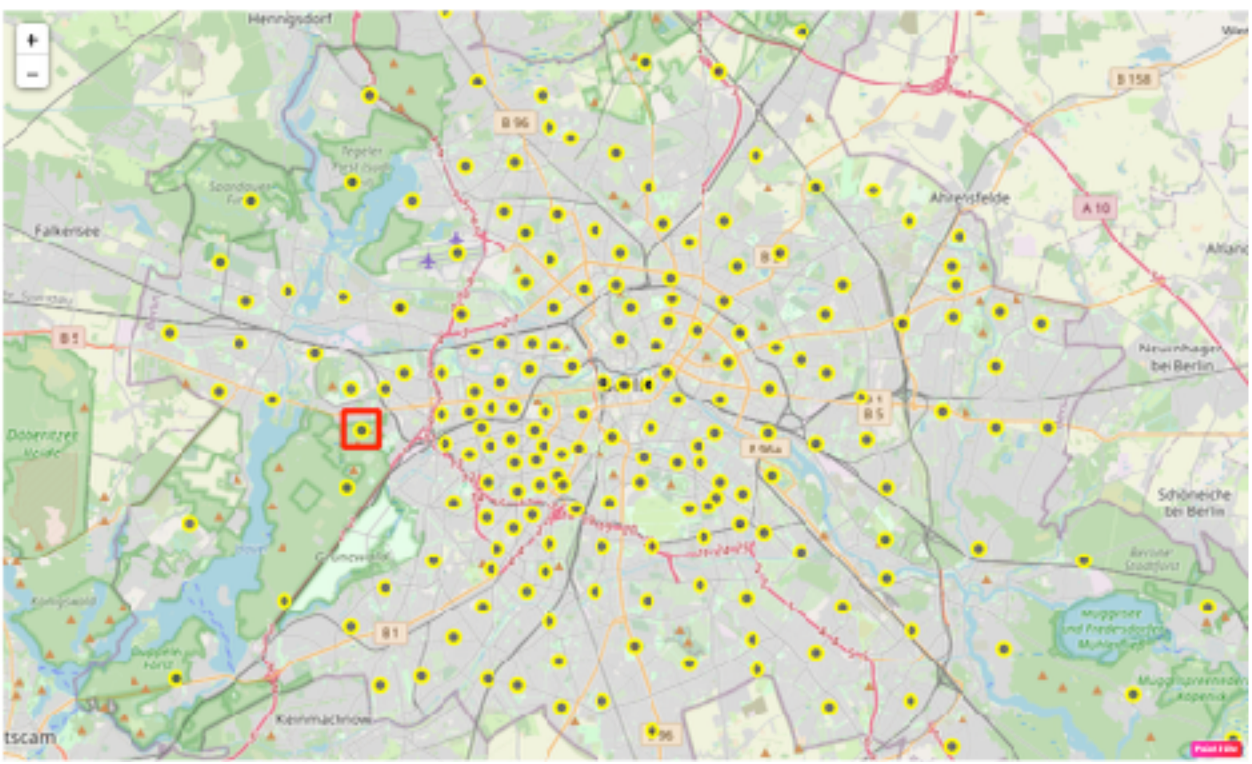

Fig 2.3 - The scatter plot of all the postal codes within the map of Berlin. The location of the SCC is highlighted with the red rectangle.

As it is highlighted with red rectangle in the Fig 2.3, the SCC is located in the neighborhood of Charlottenburg with the postal code of 14055. We will search for the similar neighborhoods to this location for the second branch of the SCC.

2.3 Feature selection

After data cleaning, it is the time to ask Foursquare for the venues in the neighborhood of each postal code. These venues will be the features of each neighborhood and they will characterize each neighborhood. For example Fig 2.4 shows three important venues in the postal code 14055 which the main branch of SCC is located. As it is clear from this figure, we see that mountain, rock climbing spot and rest area are 3 main venues in this neighborhood which of course have high effect on the success of this sport center.

	name	categories	lat	lng
0	Drachenberg	Mountain	52.502594	13.249834
1	Kletterturm Teufelsberg	Rock Climbing Spot	52.499728	13.245146
2	Parkplatz Drachenfliegerweg	Rest Area	52.501568	13.250645

Fig 2.4 - Three main venues in the neighborhood of SCC

The same procedure is taken for all the postal codes in Berlin and their most popular venues are used as a futures for the evaluation of data set.

3. Predictive model

As it has been discussed before, the aim of this project is to find the similar neighborhoods to the neighborhood of main branch of SCC as the most successful sport center in Berlin to establish the second branch. For this reason, clustering is the most useful model. Different neighborhoods of Berlin will be clustered based on their venues and then among the neighborhoods which fall in the cluster of SCC location the most similar one will be chosen as the location for the second branch of SCC. It is important step to convert all the features in the data set to one hot encoding system before applying the clustering model.

Using elbow method helps us to find out the optimal number of clusters for K-means clustering. Using this method gives us the approximate optimal number of 20 clusters for the all of the postal codes in the Berlin based on their venues. The first location of the SCC got the cluster label of 12 during this experiment and all the other data points which have the same clustering label show the familiar features and they would be good option for the location of second branch of SCC.

	PostalCode	Latitude	Longitude	ClusterLabels	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	10117	52.517	13.387	5.0	Wine Bar	Gourmet Shop	Cosmetics Shop	Bookstore	Clothing Store	Exhibit	Italian Restaurant	Salad Place
1	10115	52.532	13.385	3.0	Coffee Shop	Hotel	Café	Trattoria/Osteria	Organic Grocery	Schnitzel Restaurant	Beer Bar	Science Museum
2	10119	52.53	13.405	5.0	Italian Restaurant	Bakery	Ice Cream Shop	Café	Park	Salon / Barbershop	Beer Bar	Beer Garden
3	10178	52.521	13.41	5.0	Coffee Shop	Clothing Store	Hotel	Vietnamese Restaurant	Tour Provider	Historic Site	Science Museum	Optical Shop
4	10179	52.512	13.416	6.0	Nightclub	Bakery	Hotel	History Museum	Beer Garden	Russian Restaurant	Bar	Tourist Information Center
...	...	...	...	...	...	...	...	...	...	...	...	...
182	13583	52.544	13.182	16.0	Bar	Italian Restaurant	Bakery	Supermarket	Flea Market	Falafel Restaurant	Farm	Farmers Market
183	15537	52.4	13.717	12.0	IT Services	Shopping Mall	Miscellaneous Shop	Business Service	Big Box Store	Locksmith	Pet Store	Film Studio
184	13159	52.623	13.368	11.0	Clothing Store	Zoo Exhibit	Fabric Shop	Falafel Restaurant	Farm	Farmers Market	Fast Food Restaurant	Filipino Restaurant
185	14131	52.517	13.4	6.0	History Museum	Hotel	Theater	Museum	Plaza	Art Gallery	Art Museum	Roof Deck
186	13047	52.567	13.333	16.0	Supermarket	Bank	Restaurant	Bakery	Big Box Store	Trattoria/Osteria	Drugstore	Hotel

Fig 3.1 - Data Set with cluster labels

So the new column named ClusterLabels can be added to the data set which shows the cluster number of each postal code. Extracting the postal codes which belong to the same cluster with out initial point results in the Fig 3.2.

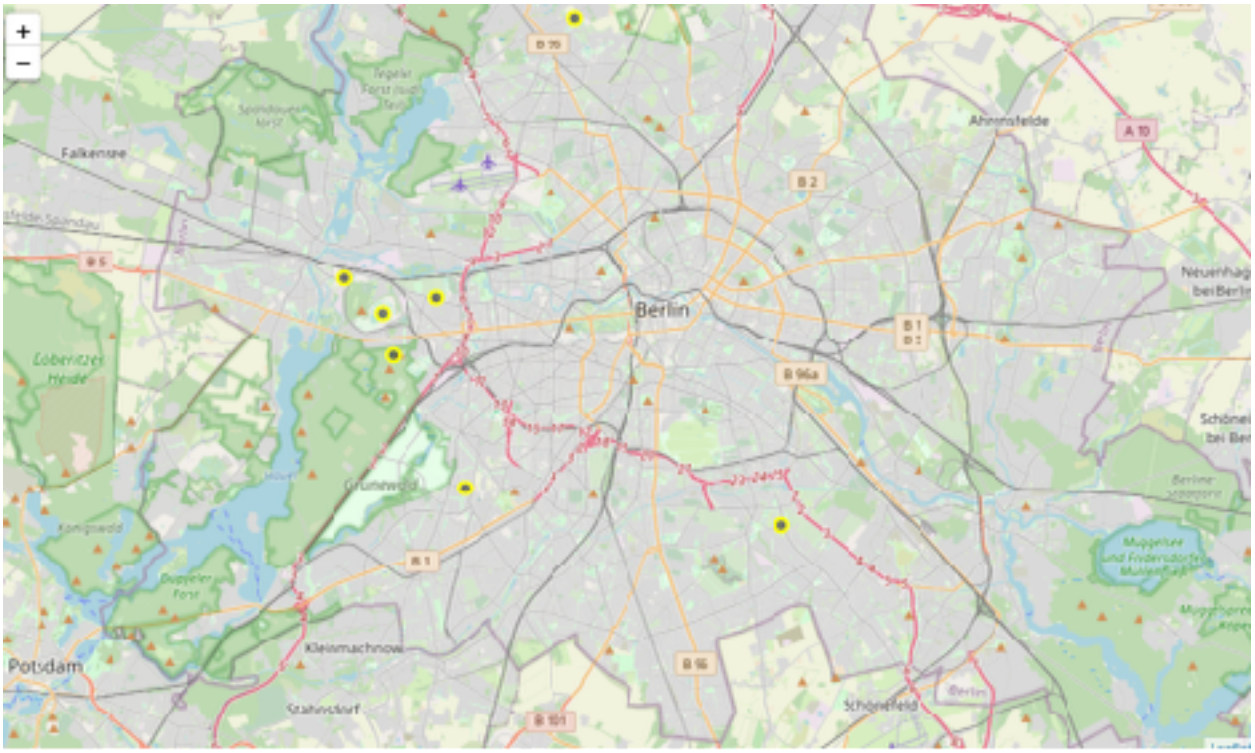

Fig 3.2 - The map of Berlin containing the data points with same characteristics with the initial location of SCC.

4. Conclusion

The aim of project was to find the best location for the second branch of SCC which is the most successful sport center in Berlin. Using the venues in each neighborhood of Berlin as a features, we constructed the clustering method. All the points which are falling to the same cluster with the initial location of SCC would be a good option to build the second branch. 6 different neighborhoods fell in to the cluster of initial SCC. As it is clear in Fig 3.2 three of the data points are very close to the initial location which is a proof that they are not a good location for the future project. However two of the points, one in ht north of Berlin with the postal code of 13469 and one in the south with postal code of 12359 are better candidates for the second successful branch of SCC.

5. Future Direction

During the investigation, it became clear that the features are not enough for this clustering. For more precise results, it is better to include the population and area of the each neighborhood and provide more features for the clustering model.