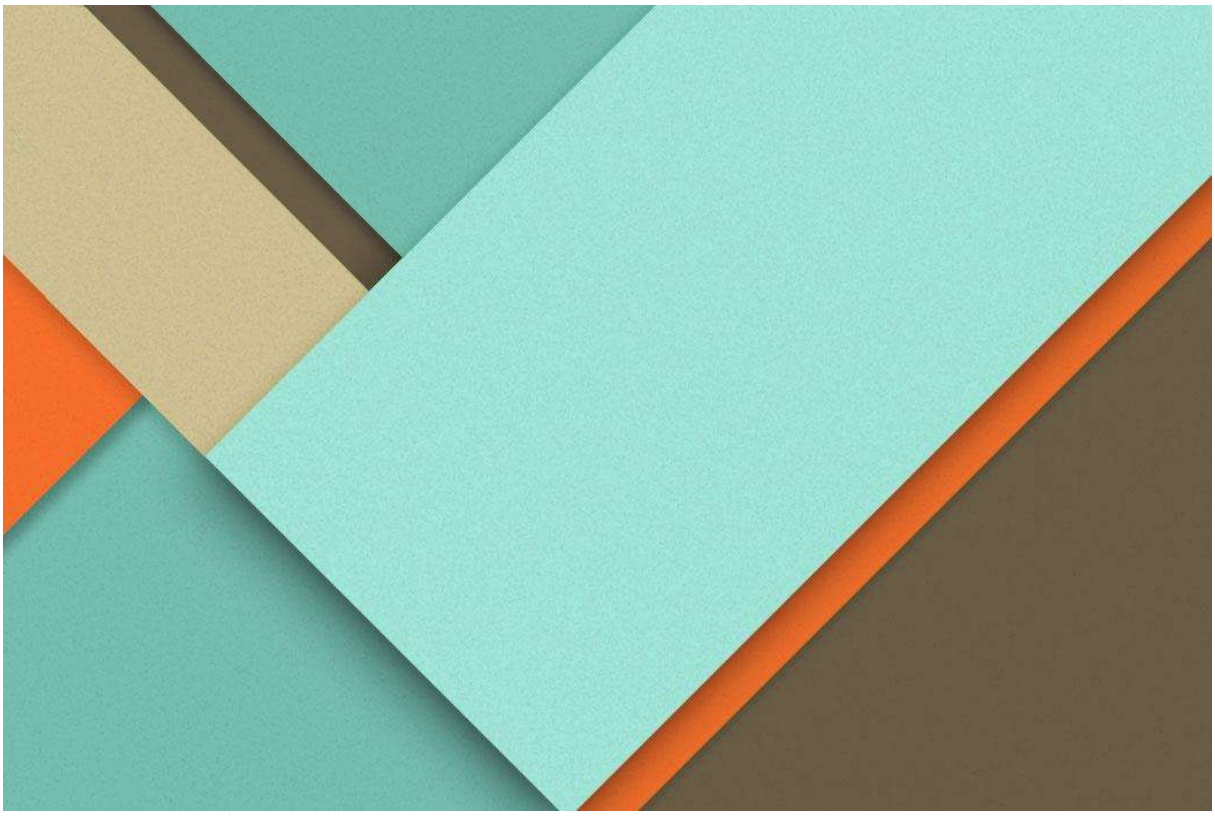


Coursera Capstone Project

11.11.2020

Zied Brahem

Business Understanding

Berlin is the cultural capital of Europe. It is a unique city with a very diverse population and comprises a lot of boroughs that are strikingly dissimilar from one another. Berlin also attracts a large number of tourists from Europe and around the world. In 2018 Berlin ranked third in the number of bednights with a staggering 32.87 million bednights, behind just London and Paris. A good portion of these tourists are mostly looking forward to discovering Berlin's famous nightlife attractions, namely nightclubs. In this context the goal of this project is to determine first of all the concentration of clubs in Berlin's different neighborhoods and then make a recommendation to a tourist regarding the area to book an accommodation in, where Berlin's coolest nightclubs are.

Goals

1. Determine the concentration of clubs in Berlin's neighborhoods
2. Build club clusters using the k_means clustering method.
3. Make a recommendation regarding the area to book an accommodation for tourists that want to visit Berlin's coolest clubs.

Analytical approach

In this project I am going to start by collecting data online about Berlin's different neighborhoods. A good source for this would be the **geopy** library. Geospatial coordinates of every neighborhood can be retrieved through this GitHub link.

Datasets on bars' and clubs' locations and ratings can be collected from the **Foursquare** platform.

Using the k-means clustering method, clusters can be built where the highest concentration of clubs are located.

Depending on the location of the cluster centroids a recommendation of a number of thorough in which to book an accommodation will be offered.

Data acquisition

The needed data for this project consist of the following

- Geospatial coordinates of Berlin's center: these can be retrieved with the geopy library. These will be needed to build a map of Berlin using Folium.
- Geospatial coordinates and names of Berlin's boroughs: after an internet search a GitHub link was found.
- Datasets of Berlin's clubs including name, popularity, geospatial coordinates. These can be collected from the Foursquare platform. Since the limit number of venues retrievable by Foursquare is 100, this will be the number of considered nightclubs. Foursquare gives in this case a recommendation of the 100 most popular venues inside a given radius depending on the venue category. The venue category will be "nightclub" and we will be looking for the best 100 venues inside a radius of 5km for city center.

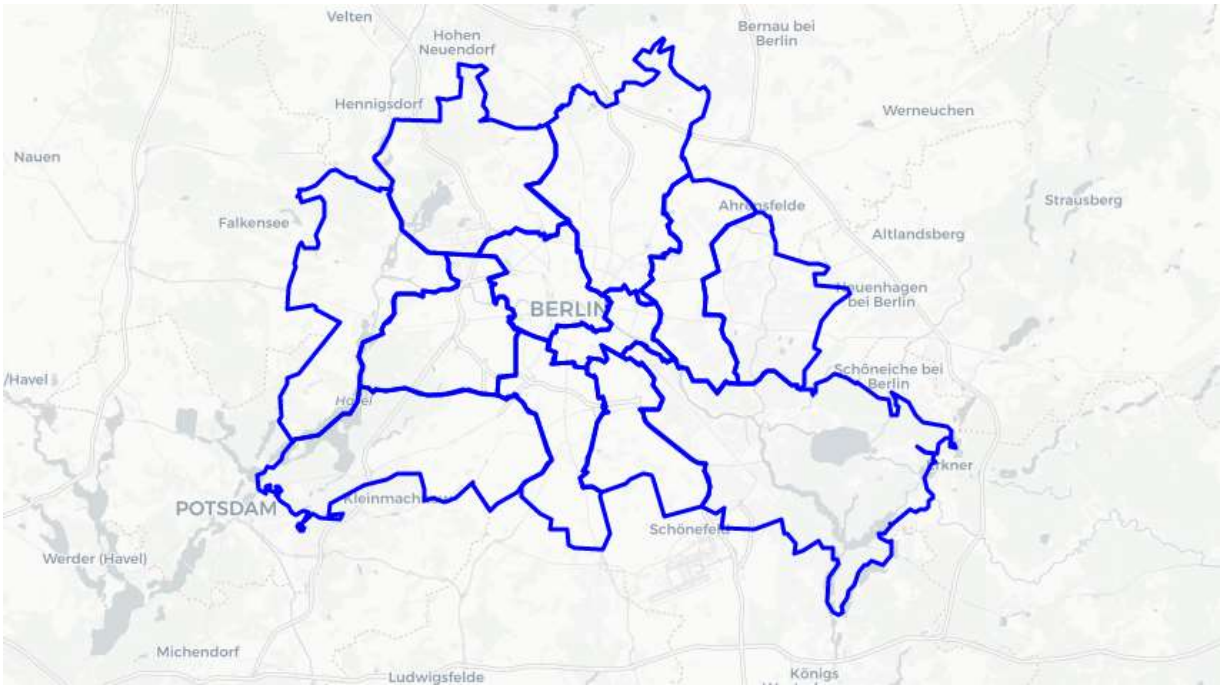
Methodology

We will start by creating a map of Berlin using the geopy library spatial coordinates of Berlin's center and Folium.

The geographical coordinate of Berlin are 52.5015217, 13.4025498.

In the next step we will retrieve through a geojson file found on GitHub the data concerning the different boroughs in Berlin. This is the GitHub link to the file in question: https://raw.githubusercontent.com/m-hoerz/berlin-shapes/master/berliner-bezirke.geojson

After normalizing the data in the geojson file, we used it to plot the areas of every borough on the Folium map. This the result.



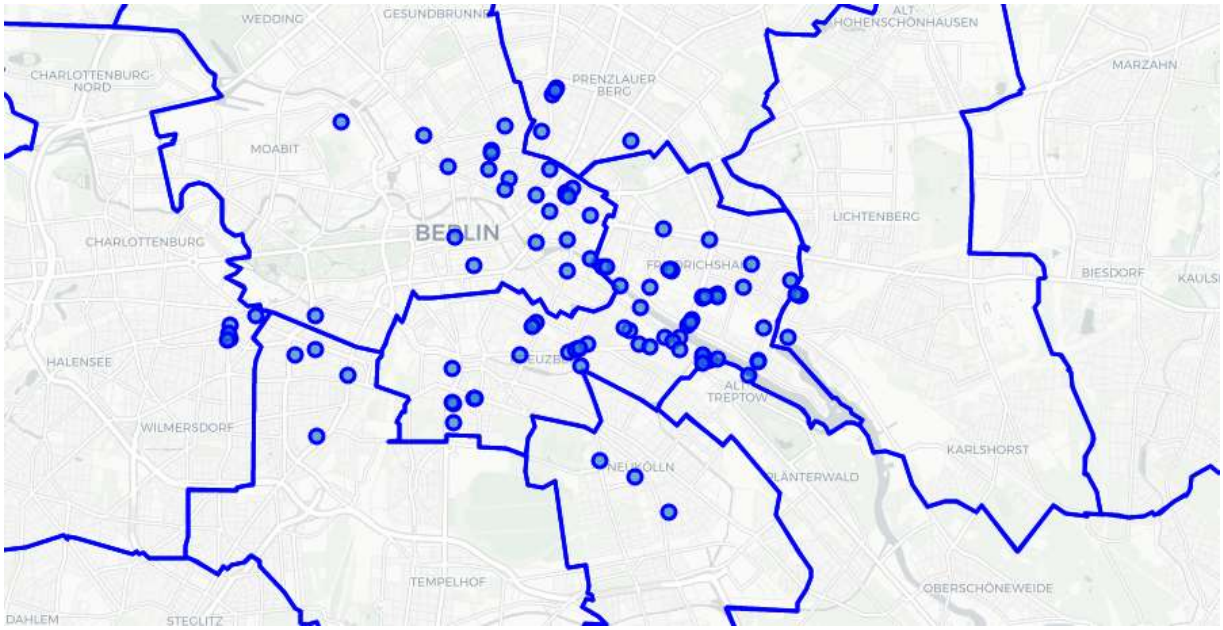
In the next step we will turn our attention to the Foursquare platform. We are going to extract information about nightclubs in Berlin using the Foursquare API and assign every venue to specific borough. Then we are going to evaluate the results in order to choose the borough with the most venues.

We're interested in venues in the 'nightclub' category. So that the tourist stays in a relatively central area, we are going to limit our search to 5km to city center. Those will be measured from the city center retrieved from the geopy library.

The retrieved information is saved into a pandas dataframe with the following columns:

- Name
- Category
- Latitude
- Longitude

Let’s try to visualize these results by plotting the geospatial coordinates of the resulting venues in our Berlin Folium map.



In the next step we tried to assign every data point to a borough. In order to accomplish this we used the shapely library. Using the Polygon function we were able to check if each data point belongs to the polygon created using the geospatial coordinates of every borough's borders. The results were added to the venues dataframe under a column named "borough".

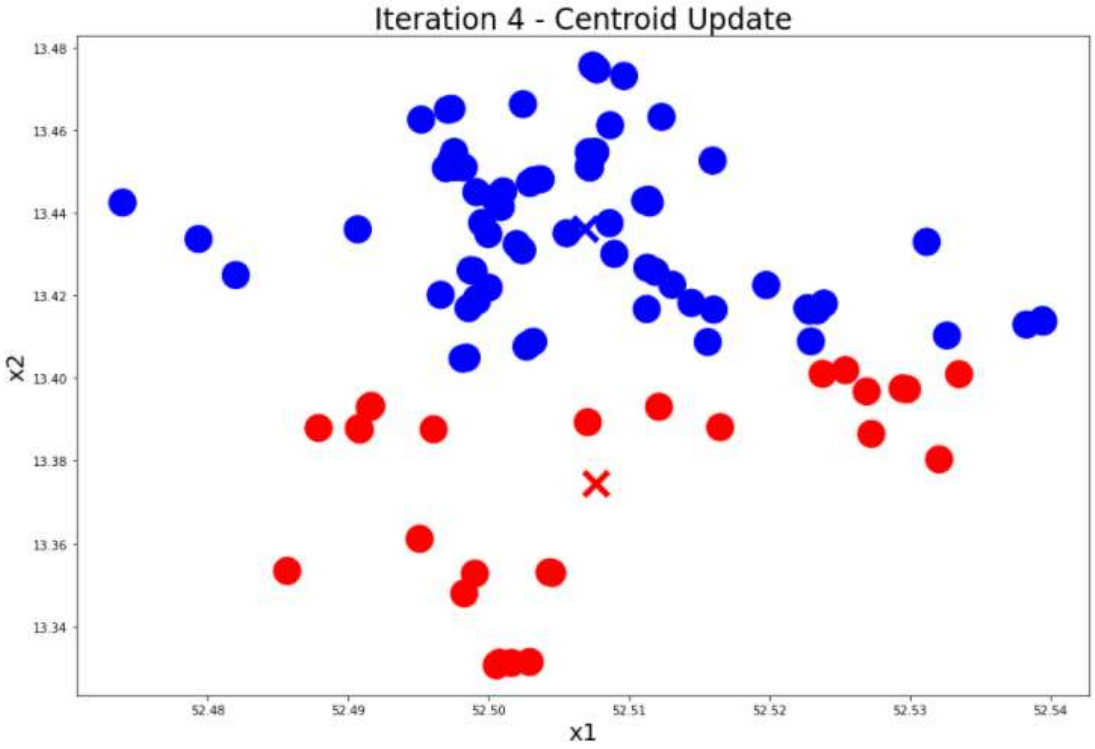
After grouping the rows of the venues data frame by boroughs we have the following results.

borough	name	categories	lat	lng
Charlottenburg-Wilmersdorf	4	4	4	4
Friedrichshain-Kreuzberg	54	54	54	54
Lichtenberg	3	3	3	3
Mitte	23	23	23	23
Neukölln	4	4	4	4
Pankow	5	5	5	5
Tempelhof-Schöneberg	4	4	4	4
Treptow-Köpenick	3	3	3	3

It looks like **Friedrichshain-Kreuzberg** is by a significant margin the borough in Berlin with the most clubs with 54 nightclubs, followed by **Mitte** with 23.

Nevertheless there is a possibility that a lot of clubs are on the border of one boroughs and therefore it would make sense to recommend another borough for booking an accommodation. We would like to further analyze the results by performing a clustering analysis using the k-means clustering method to determine in which borough the centroids of the data point clusters would be.

For this analysis we chose to perform 4 iterations and define two centroids.



Now that the centroids of the two clusters have been found, the idea is to assign a borough to each centroid. To do this we used again the Polygon function from the shapely library.

It turns out the first centroid belongs to the borough **Friedrichshain-Kreuzberg**, while the second was found in **Mitte**.

Results

We found out that **Friedrichshain-Kreuzberg** has the highest number of nightclubs in Berlin with 54. In the second place was **Mitte** with roughly half as many with 23 nightclubs. After performing a clustering analysis with k_means and 4 iterations we found out that the centroids of two resulting clusters belong to different boroughs, namely one to **Friedrichshain-Kreuzberg** and one to **Mitte**.

Conclusion

This concludes our analysis. We will recommend both boroughs to book an accommodation, while giving a slight edge to **Friedrichshain-Kreuzberg**, since it had roughly twice as much nightclubs as **Mitte**. One way to further develop this service is to include music genres in the choice of the venue and use more access to Foursquare to call more than 100 venues at a time.