Berlin Venues Project

Simge Mullaoğlu

1. Introduction

1.1 Background

Berlin is the largest city in Germany and becoming more popular day by day. Berlin has more than 3.6 million population and 55 percent of the population is younger than 45 years of age, the average age was 42.7. [2] So we can say that Berlin is a young city.

Berlin has 12 boroughs and all of them have some characteristic places. Some of them are becoming more popular than others because of these characteristic places. [1] As we can see 80000 jobs created by start-ups in Berlin in 2020. (Figure 1) Also we can see that almost 35% of all German finch start-ups locate in Berlin. (Figure 2)

Berlin is a living city and has so many different places like shopping centers, restaurants, coffee shops, offices etc. All of these are some reasons for attracting people including expats and newbie startups.

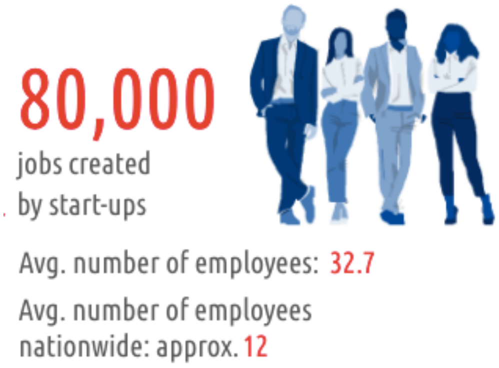


Figure 1 - Jobs Created by Start-ups in Berlin [3]

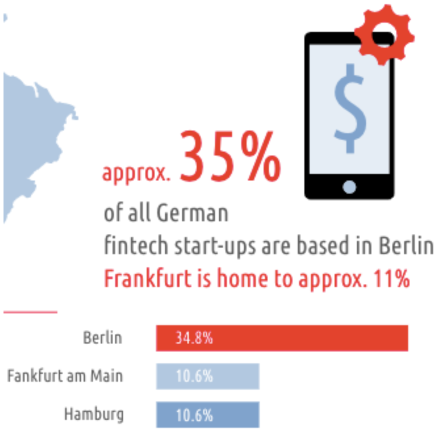


Figure 2 - Fintech start-ups with location percentage in Germany [3]

1.2 Business Problem

As I mentioned, there are a lot of demands for living in Berlin. It has so many places to attract young people, job seekers, startups.

If you are looking for a place to set up your office, popularity is not enough to decide and you need to explore the real world data. We always hear some complaints from people about the location of the office, or even in reviews on sites like Glassdoor, people consider the location of offices a minus or plus. Sometimes because of the distance or lack of transportation, they even quit their job. So this particular decision, with other plus features of your small company, of course, might attract more people to work with.

In this project, we will be answering below questions:

1 - Which neigborhoods have more social places like restaurants, Coffee shops and bars?

2 - In which neighborhoods the offices/work places are more popular?

3 - Which neighborhoods have more popular transportation centers?

4 - Which places are more attractive for people who are working in offices?

By answering these questions and combining them, we will be finding our final question:

Where should we open our office of tech startup to attract new workers?

We will be exploring boroughs and venues of Berlin to find out where is the best place to locate your new office. At the end of the project, we will be giving suggestion of locations to CEOs or Managers of startups for setting up their new offices.

2. Data Description

In this project, the data will be related with Berlin location. I have three different data sets. I will clean the data, make some preparations and then combine all of them to get the final data set.

First data set extracted online from geonames.org and form of csv file. [4] It is composed of the information of Berlin Postal Codes, and Location (Latitude/Longitude). In this data, each postal code has their unique coordinate data in Location column. Location column will be dividing two different columns in data preparation phase, since the column contains two different data and we will be using them separately in the project. Postal Codes are numeric data with five figures(e.g. 10115), and Location column is Varchar data type with latitude and longitude information (e.g. '52.532/13.385').

Second data set also extracted online from geonames.org. [4] It is in csv file format, and contains Postal Codes and Boroughs of Berlin. In data each Postal Code corresponds to a borough, so we are expecting that some postal codes are in the same borough. Postal codes are numeric (integer) with five figures like 10115 and Borough column is varchar data type (e.g. 'Berlin-Mitte'). This data set will be combining with others later based on Postal Code column.

Third data set is extracting from foursquare by using **Foursquare API** for getting the most popular venues in Berlin. The API is returning us most popular venues in Berlin, neighborhoods, latitude/longitude information, venue names and venue categories. After the getting and combining data sets, I will be using Folium library for map rendering. I will give details in methodology section.

By combining these three data sets I will be exploring, clustering, and make data analysis based on combined data.

3. Methodology

Data sets are downloaded and scraped from multiple sources and combined into one final data set. Null or meaningless values are checked and luckily there are not so many null or meaningless values. I decided to drop them from the dataset. Also, some transformations are made to specific columns like the Location column. It is divided into two different columns called Latitude and Longitude and converted to float data types, then the original column is removed since it is not going to be used.

The third data set scraped by using Foursquare API - Venues Explore endpoint. I used this endpoint by giving client id, client secret, latitude and longitude of Berlin, version of the Foursquare API (in this project '20180605' is used), radius, and the limit (default Foursquare API limit value 100 is used). After making get request, I got the name and category of the venue, latitude/longitude information of the venue. For making get request, the requests library of Python should be imported before.

After getting the venues in Berlin from the API call, put the data into a data frame. 7 columns and 3189 rows are obtained. Then we perform one hot encoding by calling the get_dummies function. This helps us to work with integers (0,1) rather than strings. This function will return a new dataframe with a column for every level of rating that presents. As result, we got 3189 rows and 319 columns that represent venues' categories. After grouping by Neighborhood and taking mean of them, we got values, categories, and neighborhoods. By sorting the values in our data frame and putting them in the newly created data frame that contains new columns like 1st, 2nd Most Common Venue, we got our final data frame.

At this point, I used the **K-means clustering** method from the scikit-learn library. Based on the data and after trying different cluster numbers, I choose 5 as a cluster number. I run k-means clustering by the number of 5 clusters and used the fit function. Also, the cluster labels are added. Then I merged this data set with the previous data set that contains postal code, neighborhood, latitude, and longitude by taking the neighborhood column as a reference. By using "**geolocator.geocode**(address)" I got latitude and longitude information of Berlin. I used the **folium** library for map rendering and visualized my clusters on the map in the Berlin location. (Figure 3) After that, I analyzed the data inside my clusters.

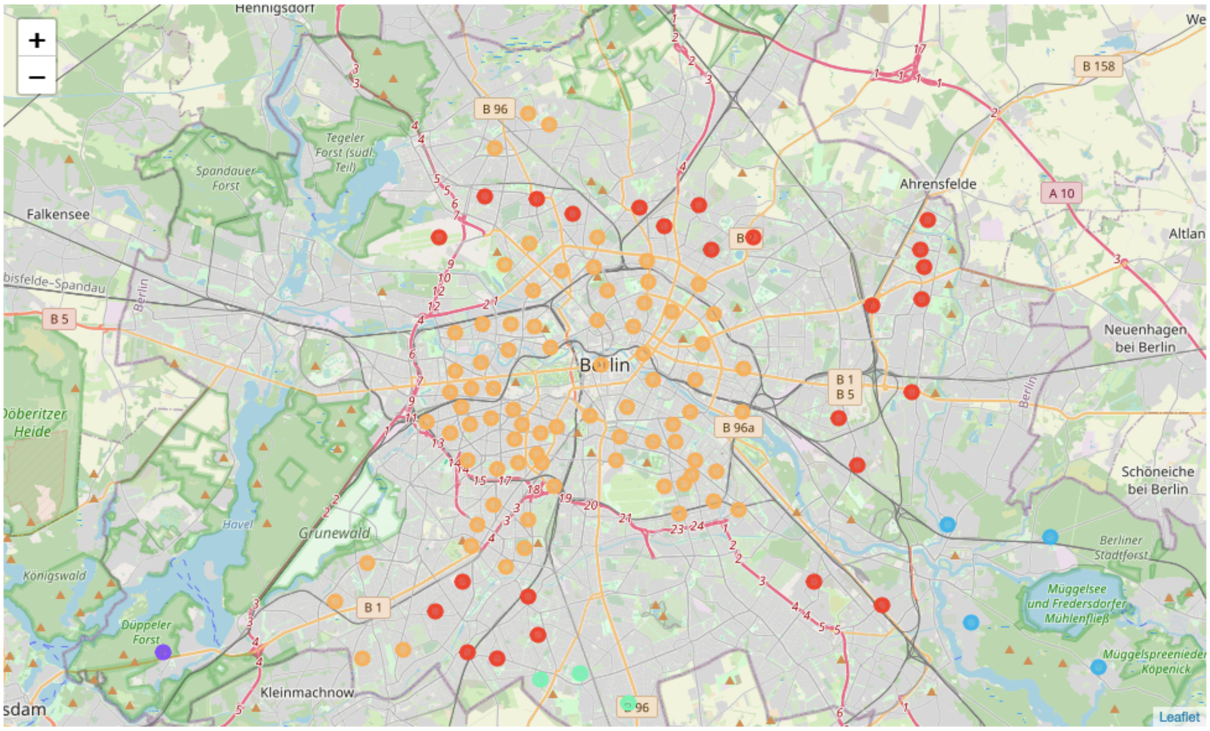


Figure 3 -Clusters in the Map

4. Analysis

After getting the data from Foursquare by using Foursquare API, I analyzed the data. By getting the count of the venues according to the neighborhoods, we can see that the most venues are located in Neukölln (421), Kreuzberg (363), Berlin-Mitte (342), and Charlottenburg (317). (Figure 4)

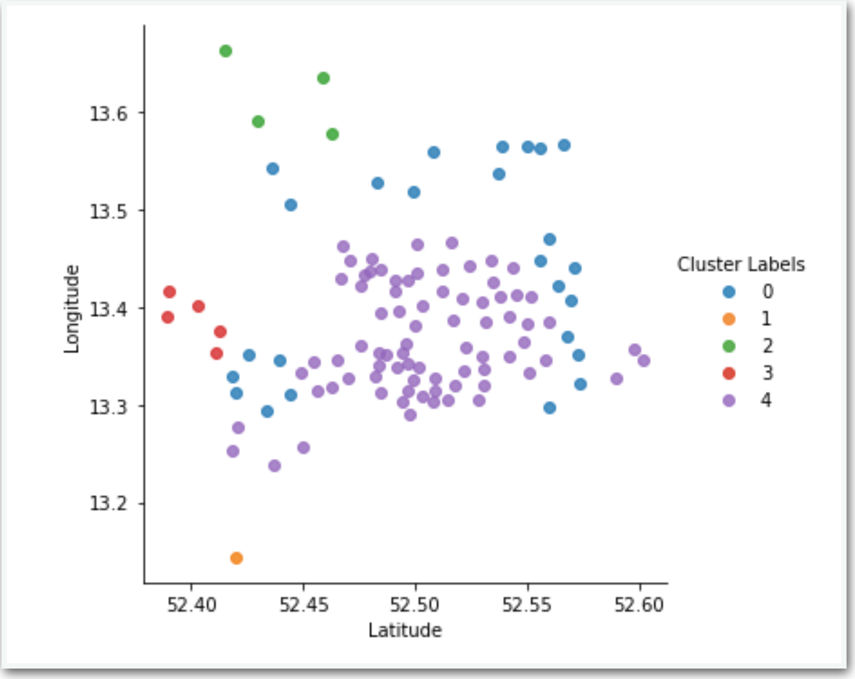


Figure 4 - Clusters based on Location (Latitude/Logitude)

As I mentioned before, I used K-Means clustering and I have 5 different clusters. I analyzed every cluster one by one.

In Cluster 1 there 25 records. It contains Lichtenberg, Lankwitz, Adlershof, Marzahn, Weißensee, Pankow and Reinickendorf neighborhoods. Generally, the most popular venue in this cluster is Supermarket, the second most popular venues are Tram Station, Light Rail Station, and BBQ Joint. The third most popular venues are Park and Plaza.

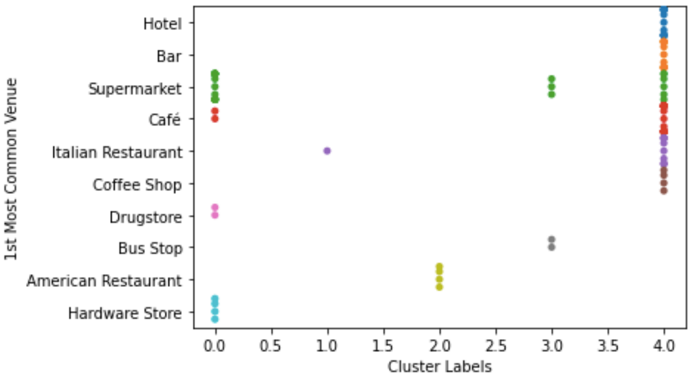
In Cluster 2, we can see that there is only one record and it is the Wannsee neighborhood. The most popular venue is Italian Restaurant, the second most popular is farmers market, etc.

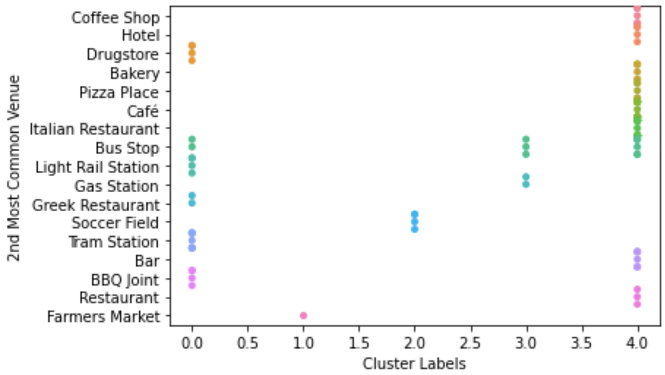
In Cluster 3, Köpenick is the only neighborhood and its most popular venue is American Restaurant, the second venue is Soccer Field, the third Gastropub, etc.

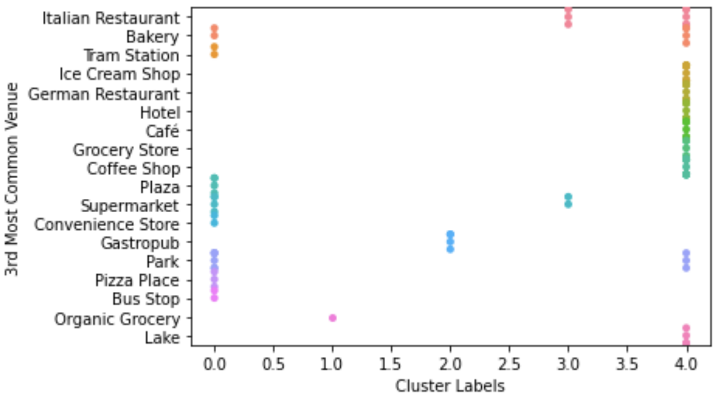
In Cluster 4, Marienfelde and Lichtenrade are the only neighborhoods. The most popular venues are Supermarket and Bus Stop, then Gas Station, Italian Restaurant, Tennis court and Mexican Restaurant, Park, etc.

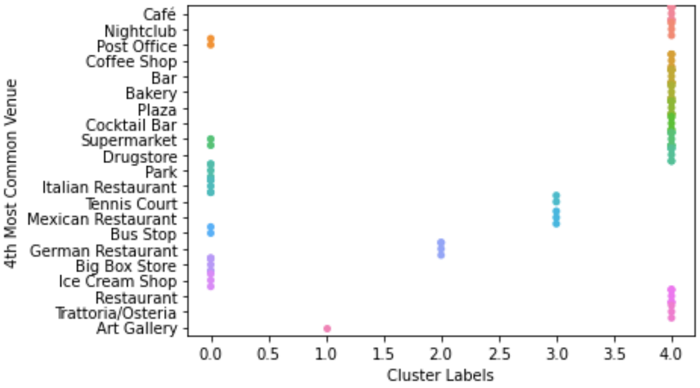
In Cluster 5, there are 78 records. Berlin-Mitte, Friedrichshain, Prenzlauer Berg, Tiergarten, Charlottenburg, Wilmersdorf, Tempelhof, Schöneberg, Kreuzberg, Neukölln, Steglitz, Wedding, Wittenau, and Zehlendorf are in this cluster. We can see that the most popular venues here are Cafe, Italian Restaurant, Hotel, and Bar. The second most popular places are Bakery, Bus Stop, and Restaurant. The third most popular venues are German Restaurant, Ice Cream Shop, Coffee Shop, Lake and Park. The fourth most popular venues are Plaza, Drugstore, Nightclub, and Cocktail Bar. The Fifth most popular places are Pool and Playground, etc.

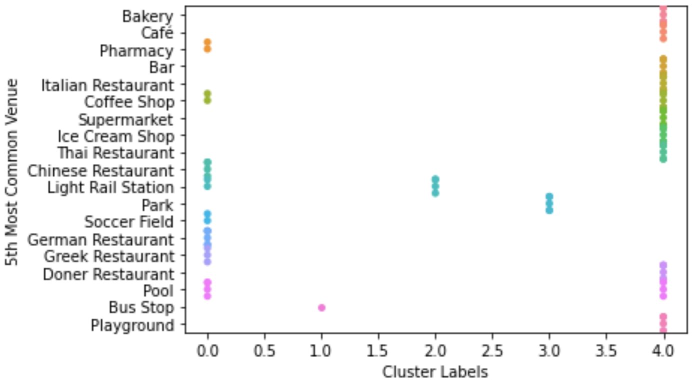
You can see the detailed graphics below.

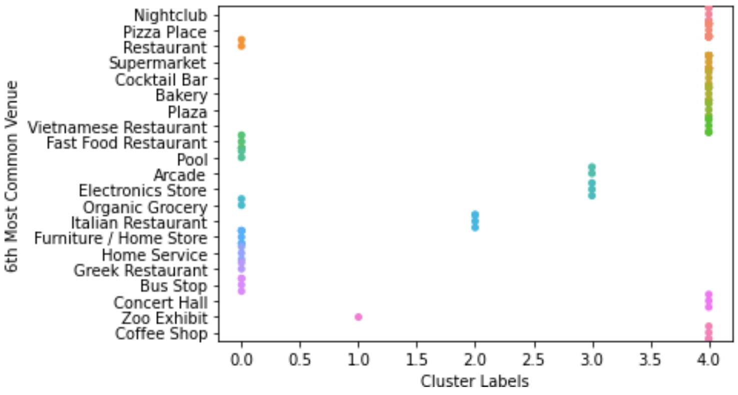












5. Results

As we can see from the analysis, the cluster divided based on some characteristics. For example, in cluster 1 Plaza, Supermarket, Restaurants, Rail Station, and Stores are popular. Also in Cluster 5, Restaurants, Plaza and Coffee shops are popular.

When some neighborhoods have some specific popular places, other neighborhoods have others. For example, we can see that nightclubs or cocktail bars are more popular in Cluster 5. Soccer Field is more popular in Cluster 3. Moreover, transportation points are more popular in Cluster 1. So we can say that if we are looking for a specific category, we can find the place where it is more popular on the clusters.

6. Discussion

From the observations, we can say that Cluster 1 and Cluster 5 have Plaza as popular venues. Plaza itself is not enough for locating your office in that neighborhood. So there need to be restaurants, coffee shops and some places that your worker can be socialized. The place that has almost all of them will be a better choice for opening an office.

For example, locating your office in Köpenick might not be a good idea, since we can see that Soccer Field and American Restaurant are popular places. Wilmersdorf might be a good option, since it has already Plaza as a popular place, and it has Cafe, Restaurants, Hotel, Bakery, Supermarket, even Boutique. Tempelhof might be another good option since again it has already Plaza as a popular venue, and it has popular Restaurants, Cafe, Bar, and Supermarket. Zehlendorf and Schöneberg are also might be good options since, besides Plaza, it has Restaurants and Cafe. Also, Zehlendorf has Lake and Park where people can go for relaxing. Lichterfelde is another option for locating your office, since it has Restaurants, Supermarket, Rail Station and Bus Stop. Other neighborhoods in Berlin might not be a good option for locating your office.

7. Conclusion

In this project, I examined the neighborhood and venues in Berlin city. I was looking for an answer for people like CEOs or business owners, co-founders where they should locate their offices. After performing analysis for each cluster and neighborhood, I got my final results. Although all neighborhoods in Berlin have some other popular venues, not every one of them is well suitable for a start-up or office.

If you are looking for a place to set up your office and looking for the best place for your employees, you need to meet up with their expectations. Having restaurants, coffee shops, public transportation, park, etc. make the place more attractive for people. Based on my analysis, I can say that locating your offices in Wilmersdorf, Tempelhof, Zehlendorf, Lichterfelde, and maybe in Schöneberg and Steglitz will attract your current or future employees.

References

[1] https://theculturetrip.com/europe/germany/berlin/articles/the-10-coolest-neighbourhoods-in-berlin/

[2] https://www.businesslocationcenter.de/en/business-location/berlin-at-a-glance/demographic-data/#:~:text=The%20population%20of%20Berlin&text=With%20its%20roughly%203.77%20million,the%20average%20age%20was%2042.7.

[3] https://www.businesslocationcenter.de/en/startup-capital-berlin/

[4] https://www.geonames.org/postalcode-search.html?q=14169&country=DE