**Selecting the ideal neighbourhood for a foreign student moving to Berlin**

Isaac von Kaufmann

19/11/2020

**1. Introduction**

**1.1 Background**

Berlin is the capital of Germany. Its rich culture, diverse music scene, and broad prospectus of food and drink locations make the city highly popular with young people and students. Every year over 40 million 20-29-year olds visit Berlin of which nearly 200,000 are foreign students, attracted by the prospect of studying in the most multicultural city in Germany with its 175 museums and over 100 vegan friendly restaurants among other attractions. However, Berlin also has the highest crime rate of any German region with 13,746 per 100,000 people in 2019 - more than double the German average. Thus, for students contemplating studying abroad in Berlin, careful consideration of housing location is important to ensure safety.

**1.2 Problem**

This project will consider the scenario in which a student has decided to study abroad in Berlin for a year and is trying to determine the best neighbourhood to live in during this time. Primarily, the student is concerned with their safety while living abroad and so choosing a district with historically low crime rates is crucial to their decision. Once they have selected a desired district, the student then wishes to select a neighbourhood based on criteria including the availability of various venues (food, drink, music etc.) nearby.

**2. Data acquisition and cleaning**

**2.1 Data acquisition**

The data was acquired from three sources. Firstly, data was collected from a [Berlin crime dataset](https://www.kaggle.com/danilzyryanov/crime-in-berlin-2012-2019) from Kaggle showing the frequency of various crimes in each neighbourhood of Berlin from 2012-2019. The following columns were included:

* **Year**: Year of statistics (2012 - 2019)
* **District**: Name of district
* **Code**: Neighbourhood ID
* **Location**: Neighbourhood
* **Robbery**: Robbery not on street
* **Street\_robbery**: Robbery on street
* **Injury**: Injury from assault
* **Agg**\_**assault**: Aggravated assault
* **Threat**: Deprivation of liberty, coercion, threat, persecution
* **Theft**: Larceny
* **Car**: Car theft
* **From**\_**car**: Theft from car
* **Bike**: Bike theft
* **Burglary**: Burglary
* **Fire**: Using the fire with damage without intention
* **Arson**: Using the fire with damage with intention
* **Damage**: Property damage
* **Graffiti**: Property damage due the graffiti
* **Drugs**: Crimes connected with drugs
* **Local**: Crime is close to the living place of criminal

Secondly, data was scraped from a [Wikipedia page](https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin) containing a list of the twelve Berlin districts. Additional columns include:

* **Borough**: Names of the twelve boroughs/districts
* **Population**: The population of each borough from 2010
* **Area**: Area of each borough in km^2
* **Density**: Population density of each borough in km^2

Finally, the names of neighbourhoods within the district of Steglitz-Zehlendorf were taken from the [Steglitz-Zehlendorf Wikipedia page](https://en.wikipedia.org/wiki/Steglitz-Zehlendorf). The dataset was created using the following columns:

* **Neighbourhood**: Names of each neighbourhood in the district
* **District**: Name of the relevant district
* **Latitude**: Latitude of the neighbourhoods
* **Longitude**: Longitude of the neighbourhoods

**2.2 Data cleaning**

From the Berlin crime data, only those crimes committed within the most recent year (2019) are selected. The dataset is also grouped by district and the neighbourhood column is dropped. Finally, a ‘total’ column is added, summing all crimes committed in each district.

Table

Description automatically generated

Figure Berlin crime data after pre-processing

Additional tabular data concerning the twelve districts is scraped from Wikipedia using the Beautiful Soup python library. Due to the presence of a map in column 5, row 1 of the Wikipedia table, Data from Charlottenburg-Wilmersdorf had to be extracted separately in order to drop the map column, before appending to the rest of the dataset. Once appended, the ‘Area’ and ‘Density’ columns were dropped as they were not required for analysis. Further cleaning of the ‘Population’ column involved removing the strings ‘\n’ and ‘,’ and converting the population data from string to integer, and the ‘Borough’ column was renamed to ‘District’ in preparation of merging the two datasets.

Table

Description automatically generated

Figure List of Berlin districts and populations

Graphical user interface, table

Description automatically generatedThe two datasets are merged on the district names to combine necessary information into one dataset. Then, total crime in each district was divided by the respective population to create new column showing the crime rate per person in each district.

Figure Berlin district crime

Once the crime data has been visualised, we can identify the safest district with the lowest crime rate and select this as our chosen district for further investigation. The final data was sourced from the list of neighbourhoods on the Wikipedia page of the safest district and was created from scratch, filling both the ‘Neighbourhood’ and ‘District’ columns but leaving the ‘Latitude’ and ‘Longitude’ columns empty.

Table

Description automatically generated

Figure Neighbourhoods in the safest district

Coordinates of the neighbourhoods were obtained using Google Maps API geocoding to obtain the final dataset.

**Table

Description automatically generated**

Figure Neighbourhoods in the safest district

This dataset is then used to identify the 10 most common venues for each neighbourhood using Foursquare’s API, before using the K-means clustering algorithm to cluster similar neighbourhoods together.

**3. Methodology**

**3.1 Exploratory Data Analysis**

**3.1.1 Districts with the highest crime rates**

Obtaining and visualising the five districts with the highest crime rate in 2019, it can be seen that the least safe district is Mitte followed by Friedrich-Kreuzberg, Charlottenburg-Wilmersdorf, Neukölln and Telpelhof-Schöneberg.

**Chart, bar chart

Description automatically generated**

Figure Districts with the highest crime rates

**3.1.2 Districts with the lowest crime rates**

**Chart, bar chart

Description automatically generated**Obtaining and visualising the five districts with the highest crime rate in 2019, it can be seen that the safest district is Steglitz-Zehlendorf followed by Marzahn-Hellersdorf, Reinickendorf, Spandau and Pankow.

Figure Districts with the lowest crime rates

Chart, bar chart

Description automatically generatedGiven we established the primary determinant of the students living location to be safety, Steglitz-Zehlendorf was chosen as the desired district and for further investigation. The visualisation below demonstrates the relative frequency of various crimes within the district in 2019.

Figure Frequency of different crimes in Steglitz-Zehlendorf

**3.1.3 Neighbourhoods in Steglitz-Zehlendorf**

There are seven neighbourhoods in the Steglitz-Zehlendorf district, which are visualised on the map below using python’s folium library.

Map

Description automatically generated

Figure Neighbourhoods in Steglitz-Zehlendorf

**3.2 Modelling**

Table

Description automatically generatedThe dataset containing latitudinal and longitudinal data for each neighbourhood was used to identify all venues within a 500m radius of each neighbourhood by connecting to Foursquare’s API. This returns a json file with all the retrieved venues in each neighbourhood which is then converted to a pandas data frame.

Figure Venue data from each neighbourhood

One hot encoding is then applied to the venues data to convert categorical variables into a form suitable for machine learning algorithms. Venues data is grouped by neighbourhood and venue means are calculated before the ten most common venues within each neighbourhood are identified.

The K-means clustering unsupervised machine learning algorithm is used to cluster data based on a predefined cluster size and is used in this scenario to cluster similar neighbourhoods based on the similarity of neighbourhood venues. A K value of 4 will be used to cluster the seven neighbourhoods into four clusters. This should help a student select a neighbourhood based on their specific venue interests.

**4. Results**

Once the algorithm has been run, we can access each cluster to see which neighbourhoods were assigned to each.

This first cluster (label 0) has only one neighbourhood, Wannsee, implying that the neighbourhood’s venues are suitably different from the other neighbourhoods in Steglitz-Zehlendorf. The most common venues are supermarkets, harbours, restaurants and post offices.

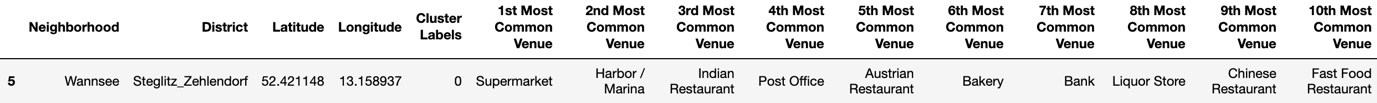


Figure Cluster label 0

This second cluster (label 1) includes four neighbourhoods, Lankwitz, Lichterfelde, Steglitz and Zehlendorf. These neighbourhoods all have similar common venues, including restaurants, café’s, bakeries and various stores.

Graphical user interface, application

Description automatically generated

Figure Cluster label 1

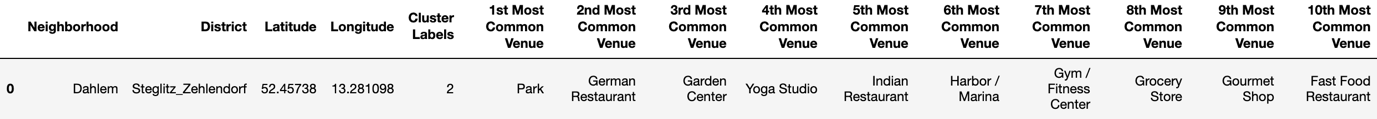
The third cluster (label 2) includes one neighbourhood, Dahlem, and features common venues including parks, restaurants, garden centres and yoga studios.

Figure Cluster label 2

The fourth and final cluster (label 3) includes just one neighbourhood again, Nikolassee, and features common venues such as trails, supermarkets, lakes, and parks.

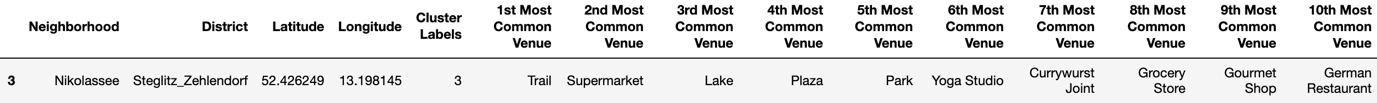


Figure Cluster label 3

Map

Description automatically generatedThe four clusters are visualised on a map using python’s folium library.

Figure Clustered neighbourhoods in Steglitz-Zehlendorf

It can be seen from the visualisation that the second cluster (label 1) is displayed with purple dots and that the four neighbourhoods are all directly neighbouring each other and closer to Berlin’s city centre (North-East of map). The other three clusters each featuring one neighbourhood each are displayed with red, green, and yellow dots (label 0, 2 and 3 respectively) and neighbour bodies of water and green space.

**5. Discussion**

This project aimed to help a student identify the safest borough to relocate to in Berlin and to help them identify the ideal neighbourhoods to consider based on their specific set of preferences. From this analysis, cluster label 1 (purple) appears to meet the students’ needs most closely, and upon closer inspection, Steglitz in particular. Steglitz offers many favourable activities for young people with many different restaurants, movie theatres, fitness centres and being the closest neighbourhood to the centre of Berlin it offers the greatest connectivity too. However, for people less concerned with these factors and seeking a more peaceful location, Dahlem and Nikolassee offer various parks, lakes and garden centres.

**6. Conclusion**

This project usefully enables individuals the chance to filter and identify locations based on their safety and selection of venues, however, it could be adapted to account for any number of features such as including the consideration of house prices in each area should budget be an issue. Further analysis may also consider the crime rates within each district in order to more accurately make a decision on the preferred neighbourhood location. Additionally, further use might be made of the specific breakdown of crimes in each area as a student may deem themselves more or less at risk to a certain selection of factors than an older individual e.g. car theft would likely be less of a contributing factor for a foreign student.