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 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 containing a list of the twelve Berlin districts. Additional columns include:

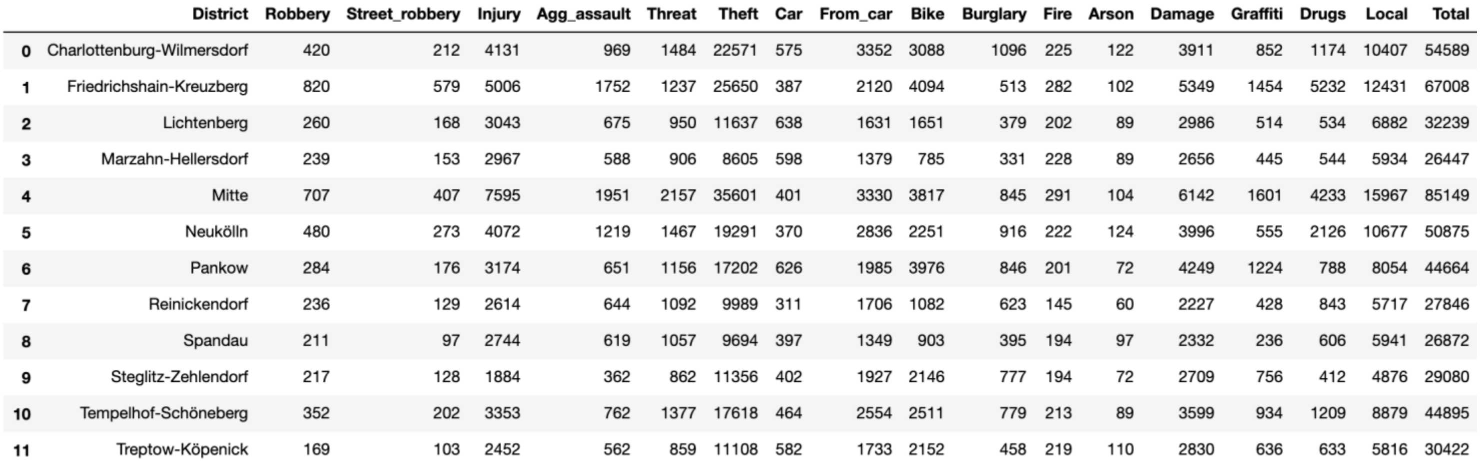
- **Borough**: Names of the twelve boroughs/districts
- **Population**: The population of each borough from 2010
- **Area**: Area of each borough in km^2
- **Density**: Population density of each borough in km^2

Finally, the names of neighbourhoods within the district of Steglitz-Zehlendorf were taken from the Steglitz-Zehlendorf Wikipedia page. 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

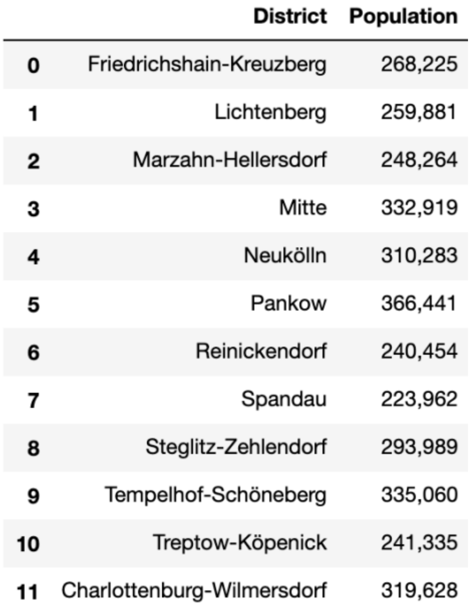
From the Berlin crime data, only those crimes committed within the most recent year (2019) are selected. The dataset is also grouped by district and the neighbourhood column is dropped. Finally, a ‘total’ column is added, summing all crimes committed in each district.



	District	Robbery	Street_robbery	Injury	Agg_assault	Threat	Theft	Car	From_car	Bike	Burglary	Fire	Arson	Damage	Graffiti	Drugs	Local	Total
0	Charlottenburg-Wilmersdorf	420	212	4131	969	1484	22571	575	3352	3088	1096	225	122	3911	852	1174	10407	54589
1	Friedrichshain-Kreuzberg	820	579	5006	1752	1237	25650	387	2120	4094	513	282	102	5349	1454	5232	12431	67008
2	Lichtenberg	260	168	3043	675	950	11637	638	1631	1651	379	202	89	2986	514	534	6882	32239
3	Marzahn-Hellersdorf	239	153	2967	588	906	8605	598	1379	785	331	228	89	2656	445	544	5934	26447
4	Mitte	707	407	7595	1951	2157	35601	401	3330	3817	845	291	104	6142	1601	4233	15967	85149
5	Neukölln	480	273	4072	1219	1467	19291	370	2836	2251	916	222	124	3996	555	2126	10677	50875
6	Pankow	284	176	3174	651	1156	17202	626	1985	3976	846	201	72	4249	1224	788	8054	44664
7	Reinickendorf	236	129	2614	644	1092	9989	311	1706	1082	623	145	60	2227	428	843	5717	27846
8	Spandau	211	97	2744	619	1057	9694	397	1349	903	395	194	97	2332	236	606	5941	26872
9	Steglitz-Zehlendorf	217	128	1884	362	862	11356	402	1927	2146	777	194	72	2709	756	412	4876	29080
10	Tempelhof-Schöneberg	352	202	3353	762	1377	17618	464	2554	2511	779	213	89	3599	934	1209	8879	44895
11	Treptow-Köpenick	169	103	2452	562	859	11108	582	1733	2152	458	219	110	2830	636	633	5816	30422

Figure 1 Berlin crime data after pre-processing

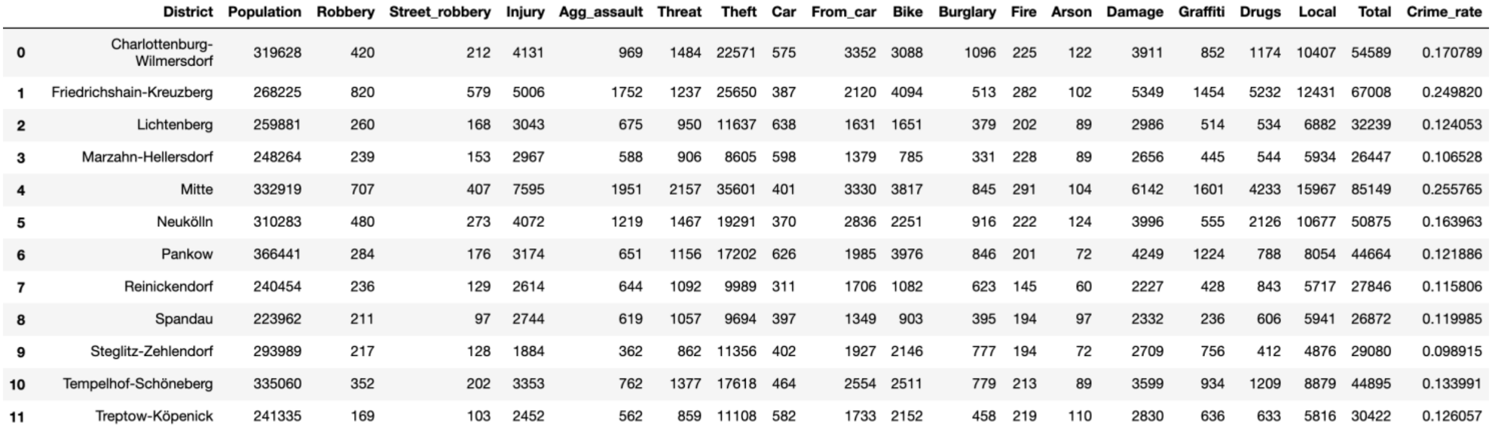
Additional tabular data concerning the twelve districts is scraped from Wikipedia using the Beautiful Soup python library. Due to the presence of a map in column 5, row 1 of the Wikipedia table, Data from Charlottenburg-Wilmersdorf had to be extracted separately in order to drop the map column, before 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.



	District	Population
0	Friedrichshain-Kreuzberg	268,225
1	Lichtenberg	259,881
2	Marzahn-Hellersdorf	248,264
3	Mitte	332,919
4	Neukölln	310,283
5	Pankow	366,441
6	Reinickendorf	240,454
7	Spandau	223,962
8	Steglitz-Zehlendorf	293,989
9	Tempelhof-Schöneberg	335,060
10	Treptow-Köpenick	241,335
11	Charlottenburg-Wilmersdorf	319,628

Figure 2 List of Berlin districts and populations

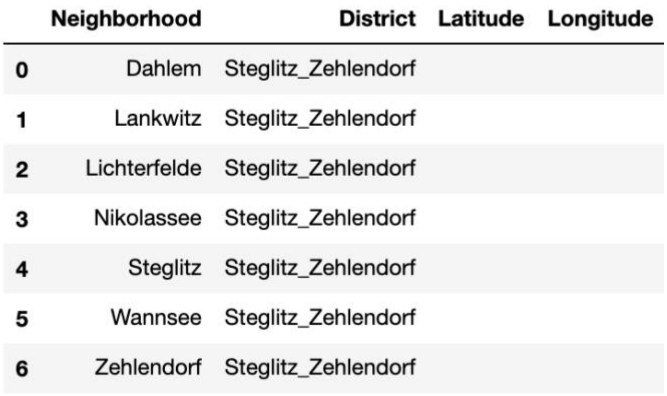
The two datasets are merged on the district names to combine necessary information into one dataset. Then, total crime in each district was divided by the respective population to create new column showing the crime rate per person in each district.



	District	Population	Robbery	Street_robbery	Injury	Agg_assault	Threat	Theft	Car	From_car	Bike	Burglary	Fire	Arson	Damage	Graffiti	Drugs	Local	Total	Crime_rate
0	Charlottenburg-Wilmersdorf	319628	420	212	4131	969	1484	22571	575	3352	3088	1096	225	122	3911	852	1174	10407	54589	0.170789
1	Friedrichshain-Kreuzberg	268225	820	579	5006	1752	1237	25650	387	2120	4094	513	282	102	5349	1454	5232	12431	67008	0.249820
2	Lichtenberg	259881	260	168	3043	675	950	11637	638	1631	1651	379	202	89	2986	514	534	6882	32239	0.124053
3	Marzahn-Hellersdorf	248264	239	153	2967	588	906	8605	598	1379	785	331	228	89	2656	445	544	5934	26447	0.106528
4	Mitte	332919	707	407	7595	1951	2157	35601	401	3330	3817	845	291	104	6142	1601	4233	15967	85149	0.255765
5	Neukölln	310283	480	273	4072	1219	1467	19291	370	2836	2251	916	222	124	3996	555	2126	10677	50875	0.163963
6	Pankow	366441	284	176	3174	651	1156	17202	626	1985	3976	846	201	72	4249	1224	788	8054	44664	0.121886
7	Reinickendorf	240454	236	129	2614	644	1092	9989	311	1706	1082	623	145	60	2227	428	843	5717	27846	0.115806
8	Spandau	223962	211	97	2744	619	1057	9694	397	1349	903	395	194	97	2332	236	606	5941	26872	0.119985
9	Steglitz-Zehlendorf	293989	217	128	1884	362	862	11356	402	1927	2146	777	194	72	2709	756	412	4876	29080	0.098915
10	Tempelhof-Schöneberg	335060	352	202	3353	762	1377	17618	464	2554	2511	779	213	89	3599	934	1209	8879	44895	0.133991
11	Treptow-Köpenick	241335	169	103	2452	562	859	11108	582	1733	2152	458	219	110	2830	636	633	5816	30422	0.126057

Figure 3 Berlin district crime

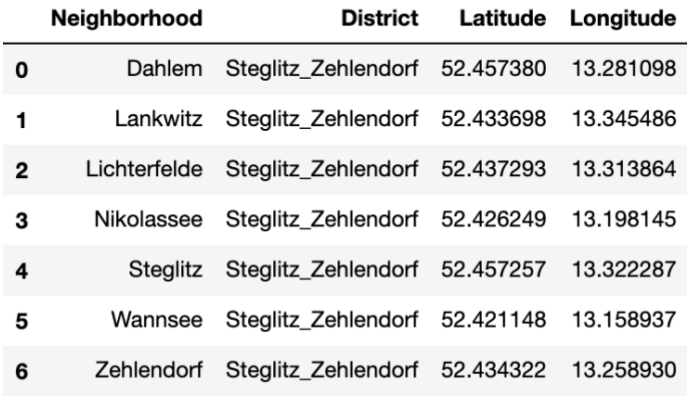
Once the crime data has been visualised, we can identify the safest district with the lowest crime rate and select this as our chosen district for further investigation. The final data was sourced from the list of neighbourhoods on the Wikipedia page of the safest district and was created from scratch,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



	Neighborhood	District	Latitude	Longitude
0	Dahlem	Steglitz_Zehlendorf		
1	Lankwitz	Steglitz_Zehlendorf		
2	Lichterfelde	Steglitz_Zehlendorf		
3	Nikolassee	Steglitz_Zehlendorf		
4	Steglitz	Steglitz_Zehlendorf		
5	Wannsee	Steglitz_Zehlendorf		
6	Zehlendorf	Steglitz_Zehlendorf		

Figure 4 Neighbourhoods in the safest district

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



	Neighborhood	District	Latitude	Longitude
0	Dahlem	Steglitz_Zehlendorf	52.457380	13.281098
1	Lankwitz	Steglitz_Zehlendorf	52.433698	13.345486
2	Lichterfelde	Steglitz_Zehlendorf	52.437293	13.313864
3	Nikolassee	Steglitz_Zehlendorf	52.426249	13.198145
4	Steglitz	Steglitz_Zehlendorf	52.457257	13.322287
5	Wannsee	Steglitz_Zehlendorf	52.421148	13.158937
6	Zehlendorf	Steglitz_Zehlendorf	52.434322	13.258930

Figure 5 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.

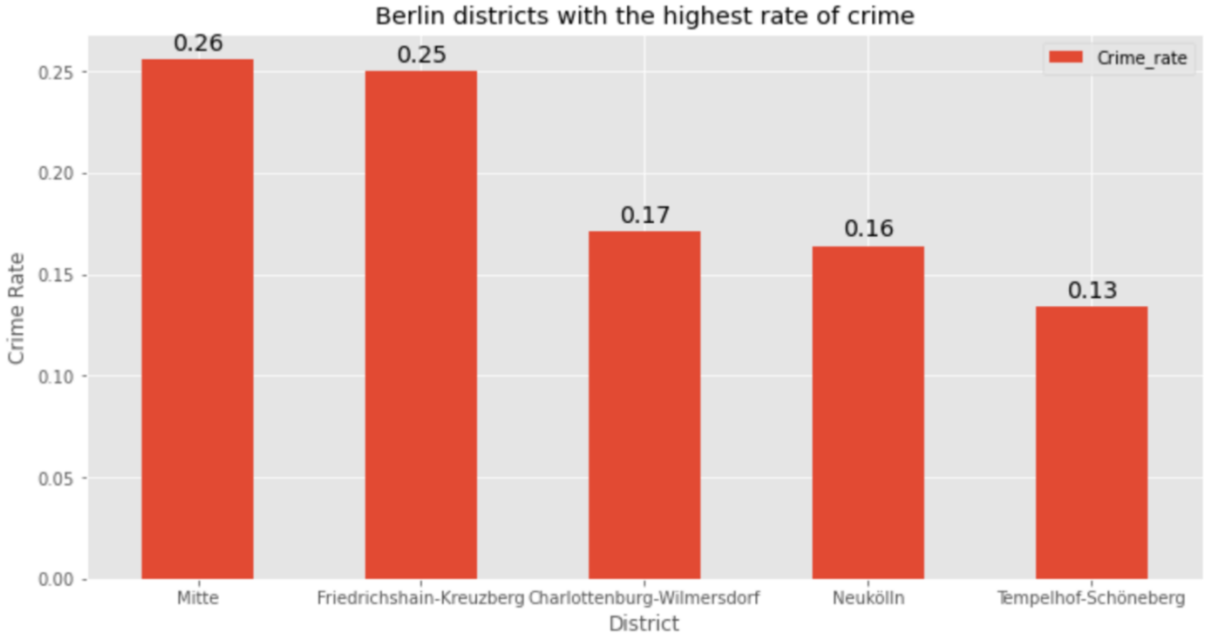


Figure 6 Districts with the highest crime rates

3.1.2 Districts with the lowest crime rates

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.

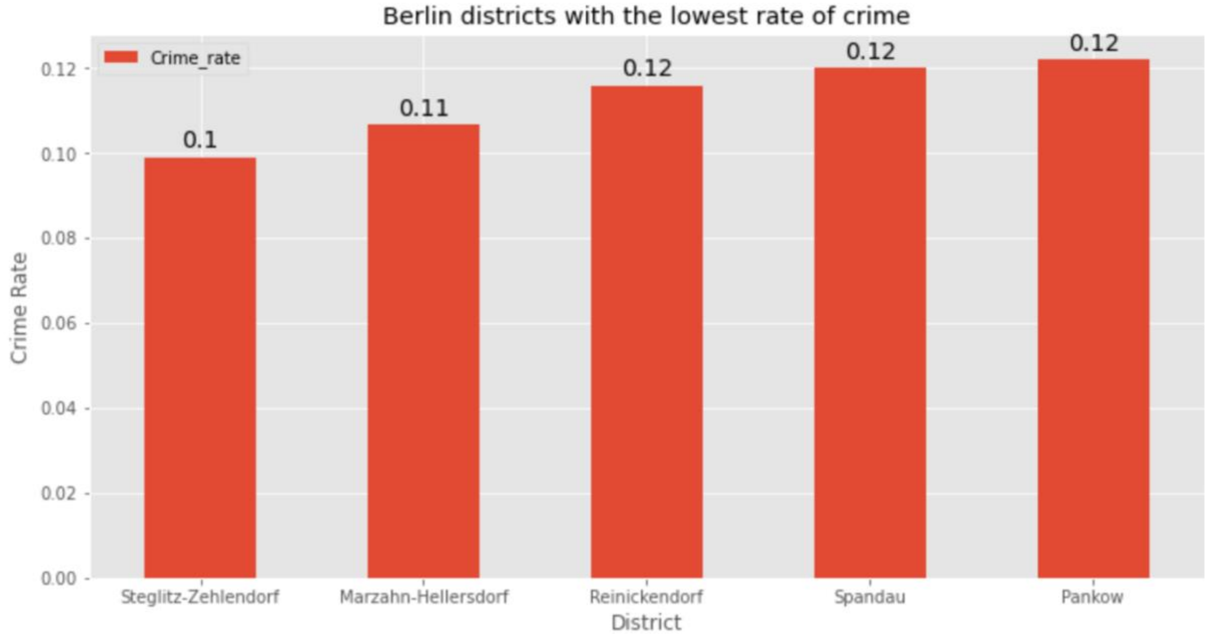


Figure 7 Districts with the lowest crime rates

Given 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.

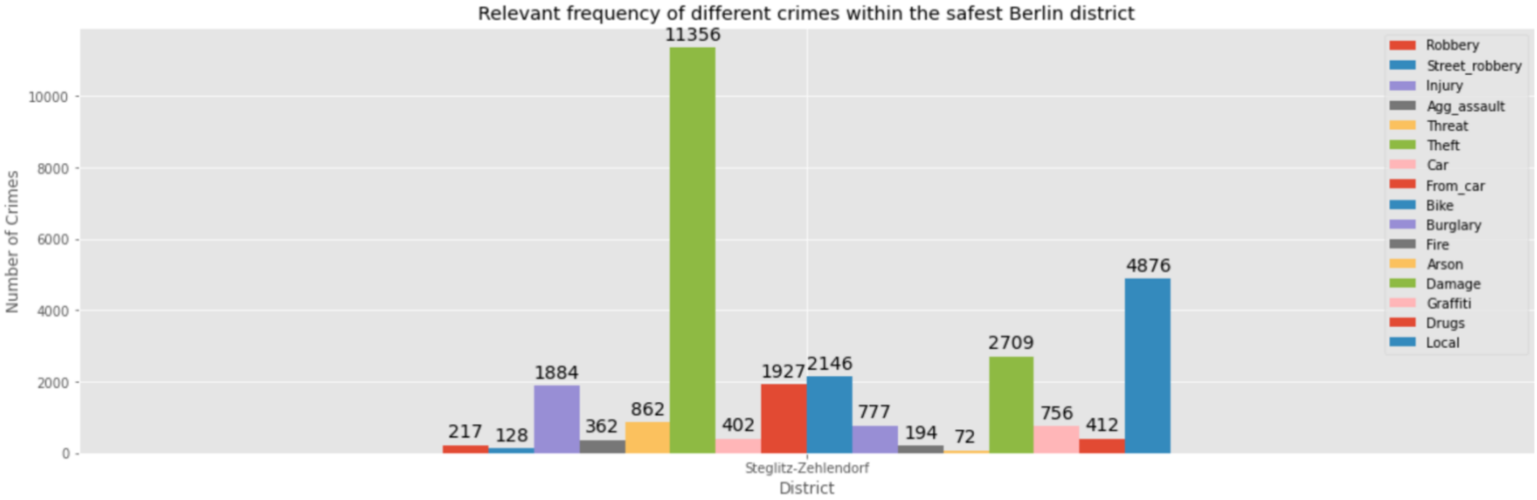


Figure 8 Frequency of different crimes in Steglitz-Zehlendorf

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.

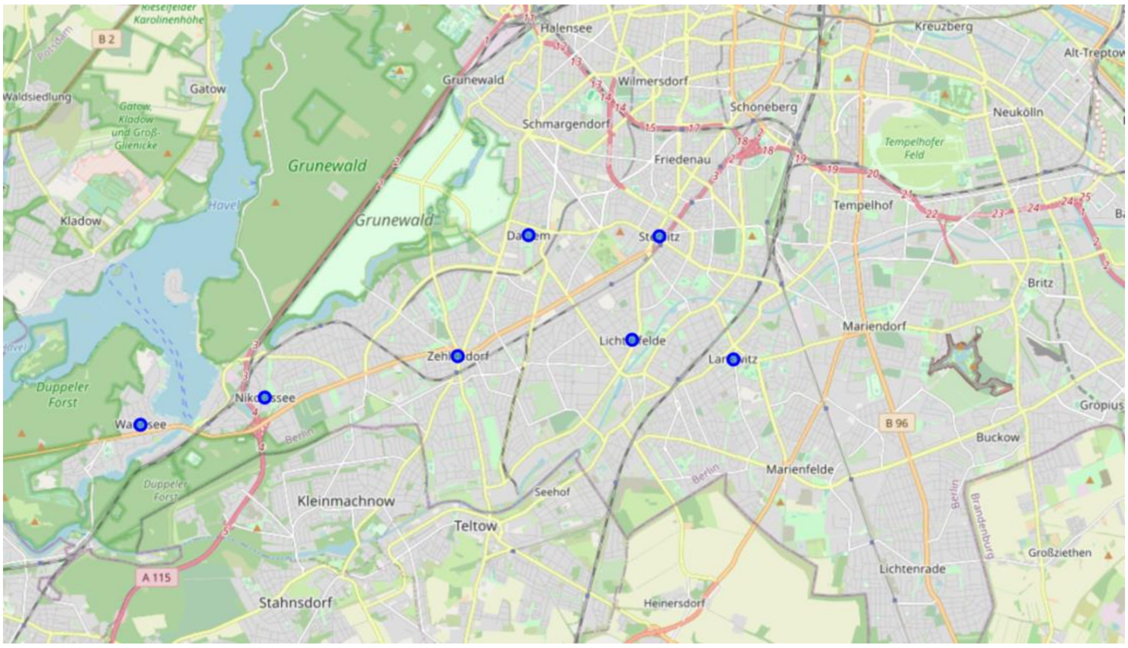
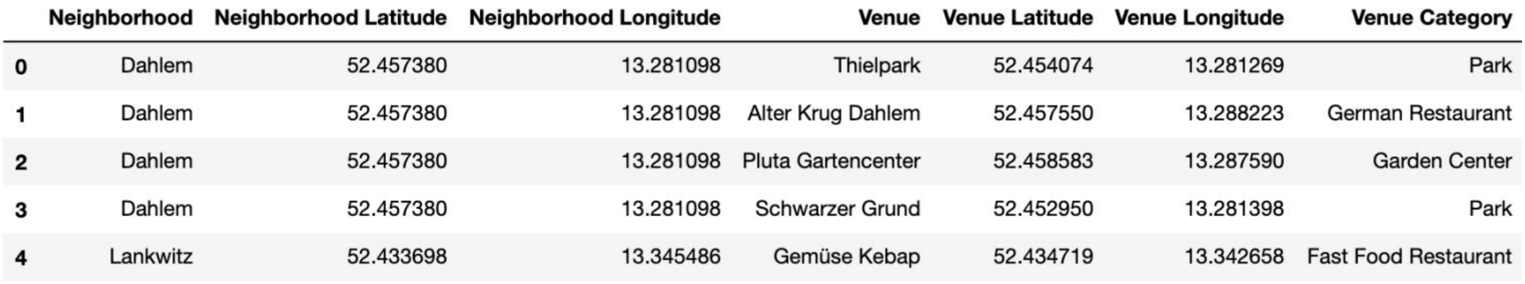


Figure 9 Neighbourhoods in Steglitz-Zehlendorf

3.2 Modelling

The 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.



	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Dahlem	52.457380	13.281098	Thielpark	52.454074	13.281269	Park
1	Dahlem	52.457380	13.281098	Alter Krug Dahlem	52.457550	13.288223	German Restaurant
2	Dahlem	52.457380	13.281098	Pluta Gartencenter	52.458583	13.287590	Garden Center
3	Dahlem	52.457380	13.281098	Schwarzer Grund	52.452950	13.281398	Park
4	Lankwitz	52.433698	13.345486	Gemüse Kebap	52.434719	13.342658	Fast Food Restaurant

Figure 10 Venue data from each neighbourhood

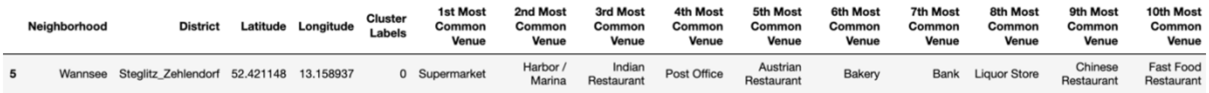
One hot encoding is then applied to the venues data to convert categorical variables into a form suitable for machine learning algorithms. Venues data is grouped by neighbourhood and venue means are calculated before the ten most common venues within each neighbourhood are identified.

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.

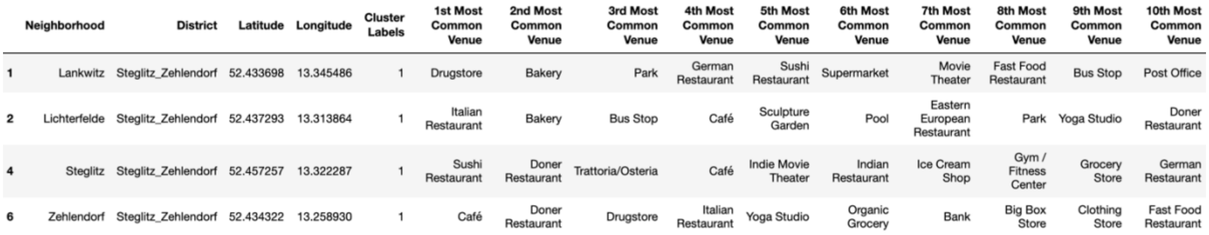
This first cluster (label 0) has only one neighbourhood, Wannsee, implying that the neighbourhood's venues are suitably different from the other neighbourhoods in Steglitz-Zehlendorf. The most common venues are supermarkets, harbours, restaurants and post offices.



	Neighborhood	District	Latitude	Longitude	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
5	Wannsee	Steglitz_Zehlendorf	52.421148	13.158937	0	Supermarket	Harbor / Marina	Indian Restaurant	Post Office	Austrian Restaurant	Bakery	Bank	Liquor Store	Chinese Restaurant	Fast Food Restaurant

Figure 11 Cluster label 0

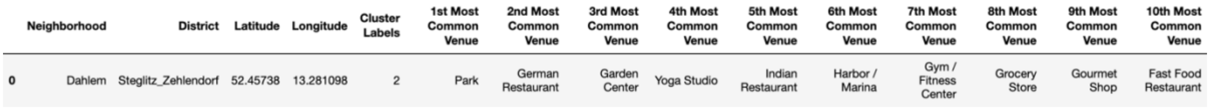
This second cluster (label 1) includes four neighbourhoods, Lankwitz, Lichterfelde, Steglitz and Zehlendorf. These neighbourhoods all have similar common venues, including restaurants, café's, bakeries and various stores.



	Neighborhood	District	Latitude	Longitude	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
1	Lankwitz	Steglitz_Zehlendorf	52.433698	13.345486	1	Drugstore	Bakery	Park	German Restaurant	Sushi Restaurant	Supermarket	Movie Theater	Fast Food Restaurant	Bus Stop	Post Office
2	Lichterfelde	Steglitz_Zehlendorf	52.437293	13.313864	1	Italian Restaurant	Bakery	Bus Stop	Café	Sculpture Garden	Pool	Eastern European Restaurant	Park	Yoga Studio	Doner Restaurant
4	Steglitz	Steglitz_Zehlendorf	52.457257	13.322287	1	Sushi Restaurant	Doner Restaurant	Trattoria/Osteria	Café	Indie Movie Theater	Indian Restaurant	Ice Cream Shop	Gym / Fitness Center	Grocery Store	German Restaurant
6	Zehlendorf	Steglitz_Zehlendorf	52.434322	13.258930	1	Café	Doner Restaurant	Drugstore	Italian Restaurant	Yoga Studio	Organic Grocery	Bank	Big Box Store	Clothing Store	Fast Food Restaurant

Figure 12 Cluster label 1

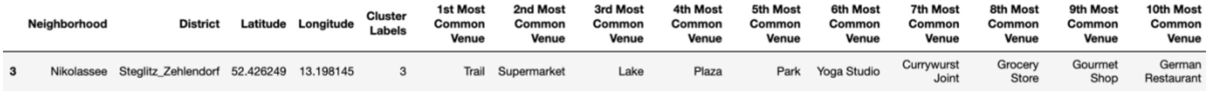
The third cluster (label 2) includes one neighbourhood, Dahlem, and features common venues including parks, restaurants, garden centres and yoga studios.



	Neighborhood	District	Latitude	Longitude	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
0	Dahlem	Steglitz_Zehlendorf	52.45738	13.281098	2	Park	German Restaurant	Garden Center	Yoga Studio	Indian Restaurant	Harbor / Marina	Gym / Fitness Center	Grocery Store	Gourmet Shop	Fast Food Restaurant

Figure 13 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.



	Neighborhood	District	Latitude	Longitude	Cluster Labels	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	9th Most Common Venue	10th Most Common Venue
3	Nikolassee	Steglitz_Zehlendorf	52.426249	13.198145	3	Trail	Supermarket	Lake	Plaza	Park	Yoga Studio	Currywurst Joint	Grocery Store	Gourmet Shop	German Restaurant

Figure 14 Cluster label 3

The four clusters are visualised on a map using python's folium library.

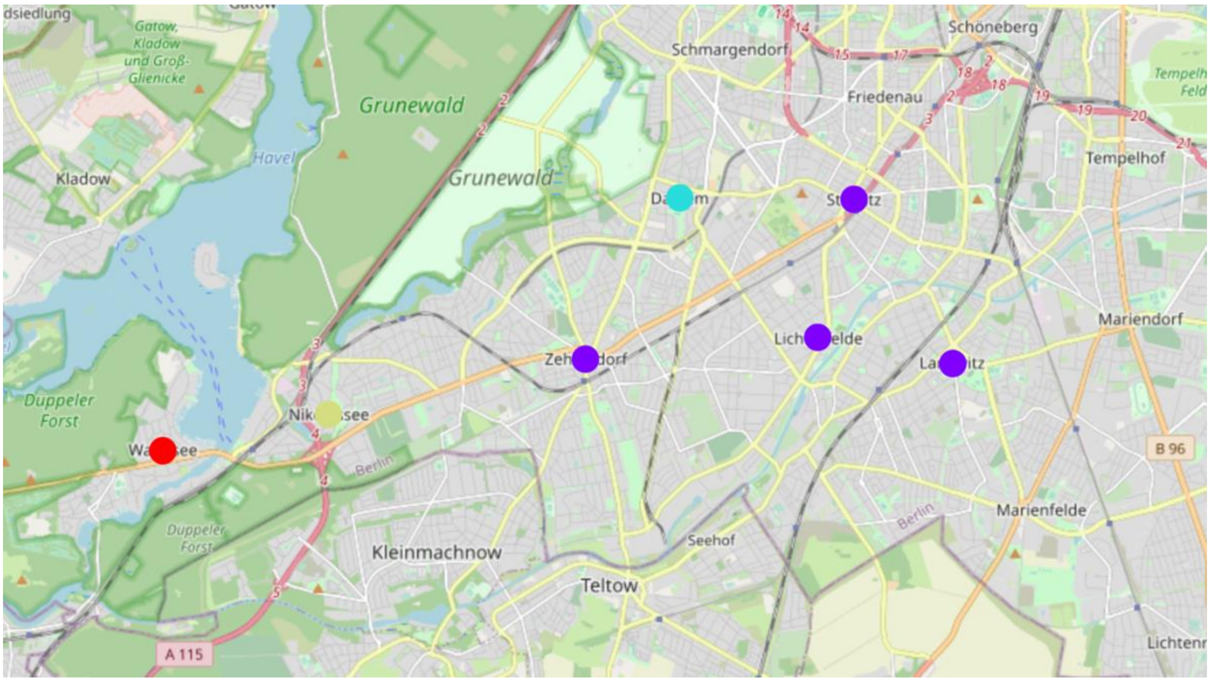


Figure 15 Clustered neighbourhoods in Steglitz-Zehlendorf

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.