COURSERA CAPSTONE PROJECT PRESENTATION

ISAAC VON KAUFMANN

19/11/2020

1. INTRODUCTION

- **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. However, Berlin also has the highest crime rate of any German region with 13,746 per 100,000 people in 2019. Thus, for students contemplating studying abroad in Berlin, careful consideration of housing location is important to ensure safety.
- **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

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.
- Secondly, data was scraped from a Wikipedia page containing a list of the twelve Berlin districts.
- Finally, the names of neighbourhoods within the district of Steglitz-Zehlendorf were taken from the Steglitz-Zehlendorf Wikipedia page.

Data Cleaning: The three data sources were cleaned separately:

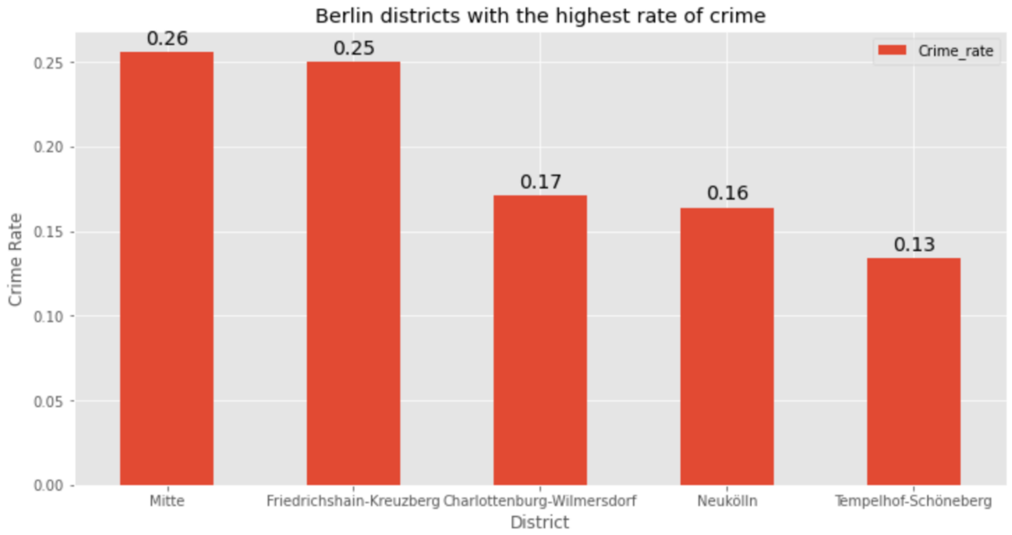
- From the Berlin crime data, only those crimes committed within the most recent year (2019) are selected.
- Additional tabular data concerning the twelve districts is scraped from Wikipedia using the Beautiful Soup python library.
- The two datasets are merged on the district names to combine necessary information into one dataset, and the crime rate per person is calculated.
- 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
- Coordinates of the neighbourhoods were obtained using Google Maps API geocoding to obtain the final dataset.
- 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

Exploratory Data Analysis:

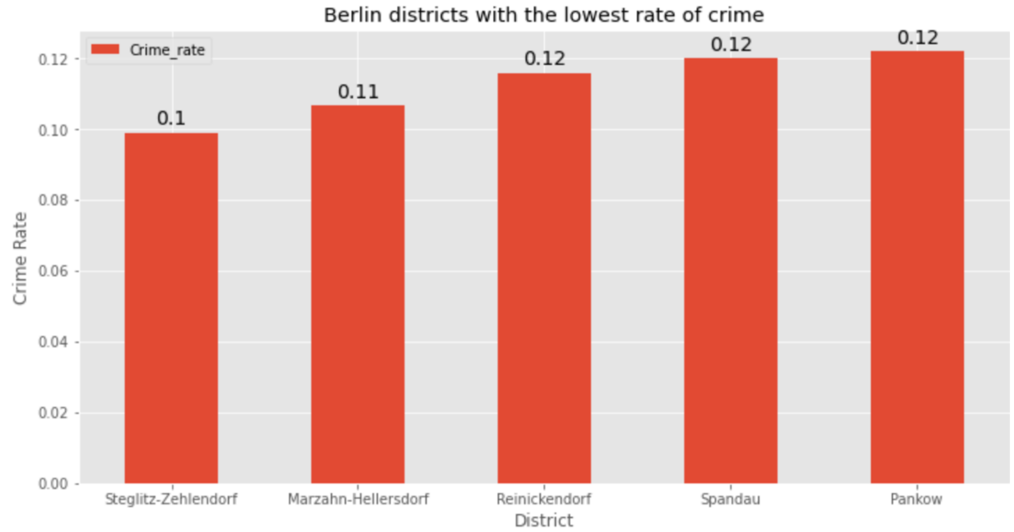
District with the highest crime rate

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



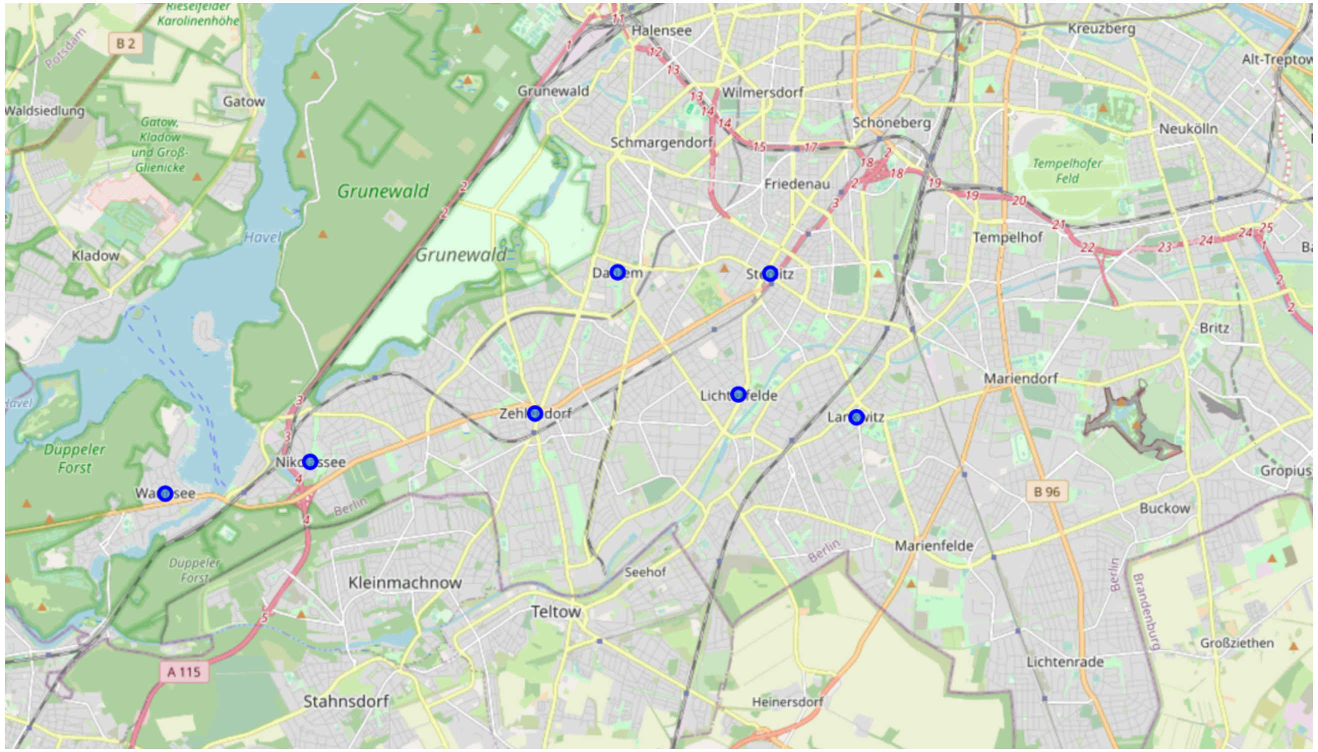
District with the lowest crime rate

- Obtaining and visualising the five districts with the highest crime rate in 2019, the safest district is Steglitz-Zehlendorf followed by Marzahn-Hellersdorf, Reinickendorf, Spandau and Pankow.



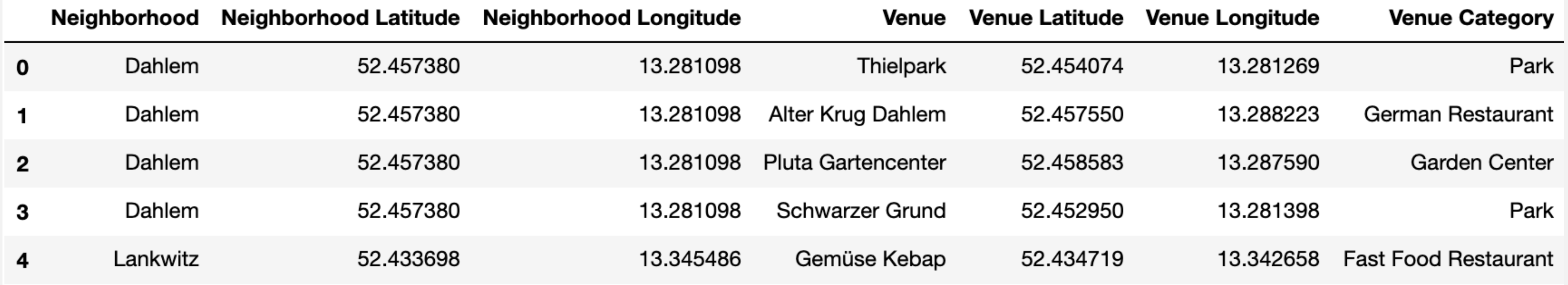
Neighbourhoods in Steglitz-Zehlendorf

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



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.

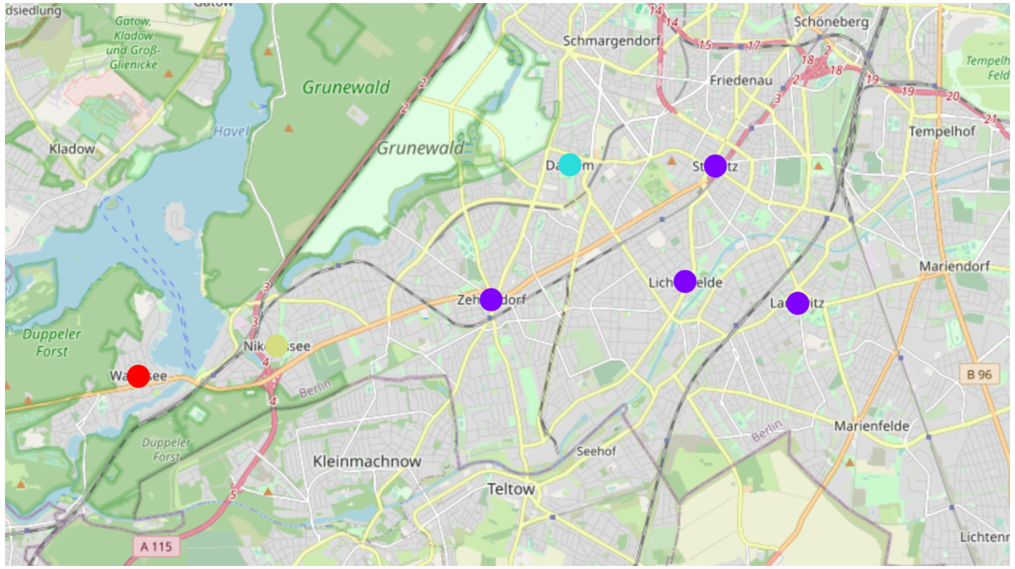


	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Dahlem	52.457380	13.281098	Thielpark	52.454074	13.281269	Park
1	Dahlem	52.457380	13.281098	Alter Krug Dahlem	52.457550	13.288223	German Restaurant
2	Dahlem	52.457380	13.281098	Pluta Gartencenter	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

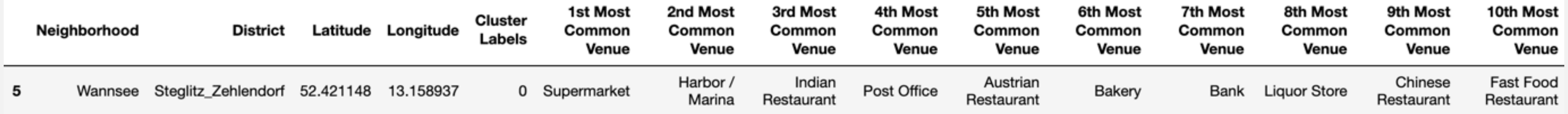
- One hot encoding is then applied to the venues data which is then grouped by neighbourhood and venue means are calculated before the ten most common venues within each neighbourhood are identified.
- K-means clustering is used to cluster data based on a predefined cluster size and is used in this scenario 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.

4. RESULTS

- Once the algorithm has been run, we can access each cluster to see which neighbourhoods were assigned to each.
- It can be seen from the visualisation that the second cluster (label 1) is displayed with purple dots. The other three clusters featuring one neighbourhood each are displayed with red, green, and yellow dots (label 0, 2 and 3 respectively).



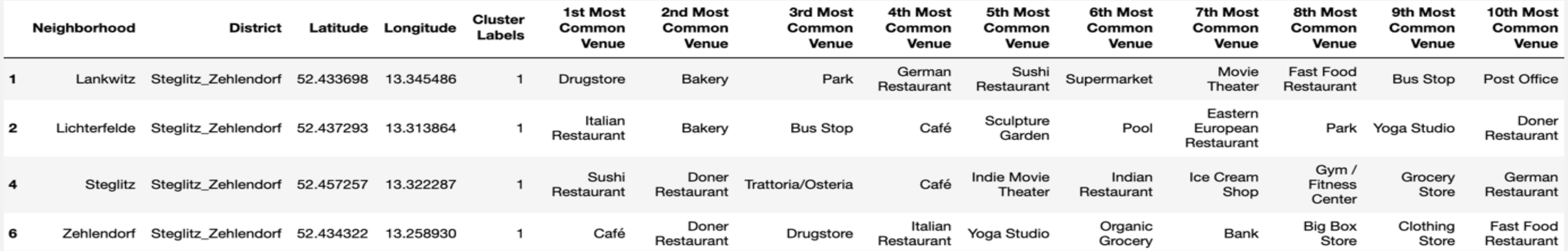
Cluster 1 (red):



	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

- 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.

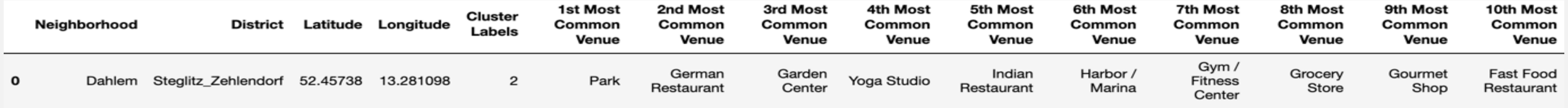
Cluster 2 (purple):



	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

- 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.

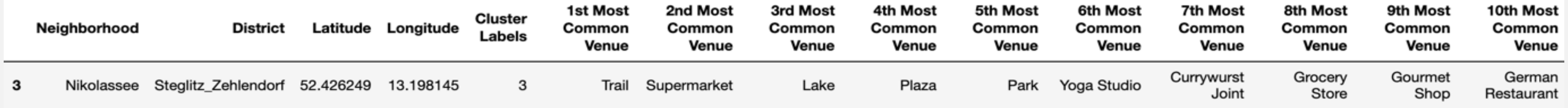
Cluster 3 (green):



	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

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

Cluster 4 (yellow):



	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

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

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.
- 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 decide 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 at risk to a different selection of factors than an older individual e.g. car theft would likely be less of a contributing factor for a foreign student.