The Battle of the Neighborhoods

Determine a Good Neighborhood for a New Student Apartment with Data Science Approach

Xiaodi Liu

1 Introduction

1.1 Background

Germany is now becoming a popular destination for international students, until 2018, Germany encompasses almost 400,000 international students, most of them are enrolled in institutions and universities fora higher education. It is also reported that from 2014 to 2019, international students numbers have increased for 30,9 %. Students are attracted to start a study life in Germany mainly due to these reasons: low tuition, low cost of living, high working possibilities during study and numerous scholarship possibilities. This reality has raised a high demand for student apartment.

1.2 Business Problem

In this project, the objective is to find a solution of choosing proper location of the new-build student apartment in Berlin based on demands of the client. The client wants to have the apartment settled in a safe neighborhood with a rental price under average level in the city. Since the apartment is targeted for students who usually travelled around with public transit, there should better be groceries and coffee shops nearby.

1.3 Interest

Reasonably this solution shall be attractive to people in real estate business, those are considering about investigating on a student apartment projects in Berlin.

2 Data Description

In this chapter, we are going to introduce the data sets used in the project and some necessary processing and cleansing work carried out in those data.

2.1 Data Source

As mentioned in the Business problem section, the relevant data in this project include the crimes data in Berlin gathered from 2012 to 2019, the rental price of boroughs in Berlin from year 2019 and the neighborhoods in all Berlin's boroughs as well as the venues data, by using Foursquare API.

Crimes data in Berlin are found available in the Kaggle public source: https://www.kaggle.com/danilzyryanov/crime-in-berlin-2012-2019?select=Berlin_crimes.csv

A detailed list of all city parts in Berlin and the corresponding postal codes: https://www.statistik-berlin-brandenburg.de/produkte/verzeichnisse/ZuordnungderBezirkezuPostleitzahlen.pdf

The rental price in all districts in Berlin :https://www.statista.com/statistics/800580/rent-expenditure-apartments-berlin-germany-by-district/ The relevant venues information like groceries and coffee shops are explored by FourSquare request.

2.2 Data Pre-processing

2.2.1 Berlin Crimes Data

The Berlin crimes data set contains all sorts of crimes cases from robbery, injury, threat and etc. of overall 150 locations in Berlin for eight years from 2012 to 2019, as shown in figure 1.

	Year	District	Code	Location	Robbery	Street_robbery	Injury	Agg_assault	Threat	Theft	Car	From_car	Bike	Burglary	Fire	Arson	Damage	Graffiti	Drugs	Local
0	2012	Mitte	10111	Tiergarten Süd	70	46	586	194	118	2263	18	328	120	68	16	4	273	26	171	1032
1	2012	Mitte	10112	Regierungsviertel	65	29	474	123	142	3203	10	307	170	37	10	4	380	124	98	870
2	2012	Mitte	10113	Alexanderplatz	242	136	1541	454	304	8988	81	792	822	275	49	27	1538	522	435	3108
3	2012	Mitte	10114	Brunnenstraße Süd	52	25	254	60	66	1916	86	192	396	131	14	5	428	122	213	752
4	2012	Mitte	10221	Moabit West	130	51	629	185	199	2470	94	410	325	161	42	22	516	64	259	1403

Fig. 1: Crimes in Berlin from 2012 to 2019.

Then some pre-processing work are carried out to adjust the data to a required structure. I followed these steps.

- I have modified column names *District*, *Code* into *Borough* and *Postal Code*
- Then I calculate the averaged crime numbers over the 8 years in each district by grouping the data on columns year and district.
- Add a new column *Total Cases* to save number of all crimes in that neighborhood.

The data set is looking like the following figure after the processing.

	Borough	Postal Code	Location	Robbery	Street_robbery	Injury	Agg_assault	Threat	Theft	Car	From_car	Bike	Burglary	Fire	Arson	Damage	Graffiti	Drugs	Total Cases
0	Mitte	10111	Tiergarten Süd	71.0	49.0	483.0	131.0	120.0	2526.0	18.0	293.0	236.0	57.0	13.0	4.0	277.0	45.0	175.0	326.0
1	Mitte	10112	Regierungsviertel	56.0	28.0	478.0	104.0	132.0	4349.0	14.0	244.0	313.0	44.0	8.0	2.0	380.0	129.0	95.0	511.0
2	Mitte	10113	Alexanderplatz	181.0	106.0	1541.0	411.0	333.0	10783.0	79.0	716.0	989.0	209.0	42.0	16.0	1223.0	380.0	684.0	1619.0
3	Mitte	10114	Brunnenstraße Süd	36.0	19.0	240.0	52.0	69.0	2208.0	61.0	220.0	417.0	98.0	12.0	5.0	401.0	157.0	94.0	563.0
4	Mitte	10221	Moabit West	90.0	43.0	637.0	176.0	199.0	2612.0	58.0	446.0	383.0	118.0	34.0	14.0	529.0	81.0	417.0	624.0

Fig. 2: Crimes in Berlin from 2012 to 2019.

2.2.2 Berlin Postal Codes Data

When we take a deeper look into the crimes data, one find out the postal code are somehow wrongly collected, not only that some postal codes are missing, but some of the recorded postal codes are wrongly given. So I search online around to find a more reliable and compete data set of Berlin districts and postal codes.

The loaded data frame is like following:

	Borough	PostalCode
0	Mitte	10115;10117;10119;10178;10179;10435;10551;1055...
1	Friedrichshain-Kreuzberg	10179;10243;10245;10247;10249;10367;10785;1096...
2	Pankow	10119;10247;10249;10405;10407;10409;10435;1043...
3	Charlottenburg-Wilmersdorf	10553;10585;10587;10589;10623;10625;10627;1062...
4	Spandau	13581;13583;13585;13587;13589;13591;13593;1359...

Fig. 3: All Boroughs in Berlin and their corresponding Postal Codes.

Based on figure 3, each borough is leading to multiple neighborhoods or areas, in this project, we want all the postal codes to get an exact geographical coordinates via *geopy package*, therefore the data set needs to be further transformed in the proper format as shown in figure 4.

2.2.3 Berlin Rental Price

Similarly, the rental price in Berlin from year 2019 is imported as data frame from the given web page. The data contains 9 borough names and the related rental price per quart-meter. Due to the data protection rules, up to three boroughs rental price are not open to a non-premium account. In the figure 5, we can take a look at the data in a clear view.

	Borough	PostalCode
0	Mitte	10115
1	Mitte	10117
2	Mitte	10119
3	Mitte	10178
4	Mitte	10179
...	...	...
249	Reinickendorf	13507
250	Reinickendorf	13509
251	Reinickendorf	13599
252	Reinickendorf	13629

Fig. 4: All Postal Codes in Berlin and their corresponding Boroughs.

	Borough	Rental Price per Qm
0	Mitte	13.42
1	Friedrichshain-Kreuzberg	13.00
2	Charlottenburg-Wilmersdorf	12.65
3	Pankow	10.94
4	Steglitz-Zehlendorf	10.67
5	Tempelhof-Schöneberg	10.52
6	Berlin total	10.44
7	Neukölln	10.06
8	Treptow-Köpenick	9.92
9	Lichtenberg	9.26

Fig. 5: All Boroughs in Berlin and their corresponding rental price from year 2019.

2.2.4 Venues in Berlin's Neighborhoods

The venues data are achieved via FourSquare API, one first get the latitude and longitude to each postal code area by appling the *geopy.geocoder* package in python. Then I get connected to the API with account credentials.

```
CLIENT_ID = 'UJ50JPVA2FQOF5GD0YRACXEV4UMDA0N1EANKDITSRAQKJTIH'
CLIENT_SECRET = 'B0OD3WNABMHYB5JZRTOIGQA02GGT3QQR2PJPHFYS2EUQMS4H'
VERSION = '20180605'
LIMIT = 50 # limit of number of venues returned by Foursquare API
radius = 750
print('Your credentails:')
print('CLIENT_ID: ' + CLIENT_ID)
print('CLIENT_SECRET:' + CLIENT_SECRET)
...
berlin_venues = getNearbyVenues(names=berlin_df['Borough'],
                                latitudes=berlin_df['Latitude'],
                                longitudes=berlin_df['Longitude'
    ])
```

Listing 1: Connect to FourSquare API and get nearby venues

Afterwards, write a function to get nearby venues with given latitudes and longitudes as well as a user defined radius. The obtained venues data frame is presented in figure 6.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Charlottenburg-Wilmersdorf	52.49962	13.32316	Mövenpick Weinkeller	52.499556	13.320365	Wine Shop
1	Charlottenburg-Wilmersdorf	52.49962	13.32316	BERLINRODEO interior concepts GmbH	52.501163	13.325814	Furniture / Home Store
2	Charlottenburg-Wilmersdorf	52.49962	13.32316	Sweet2go	52.497776	13.322638	Dessert Shop
3	Charlottenburg-Wilmersdorf	52.49962	13.32316	Piccola Taormina	52.501231	13.324884	Italian Restaurant
4	Charlottenburg-Wilmersdorf	52.49962	13.32316	EDEKA	52.500454	13.322082	Supermarket

Fig. 6: All venues nearby with the name and category information.

3 Methodology

The purpose of this project is to use machine learning algorithm to find a proper location for an apartment based on certain demands on the surroundings. To achieve that, one first collect all necessary data input. The next step is to determine the approach that is suited for the issue. Before applying any methods, an exploratory analysis on the pre-processed data is taken place first to grab a deeper understanding of the data and our goal.Thereafter, I used a k-means cluster to cluster all the relevant venues in berlin's neighborhoods and try to find out whether suitable spots are achieved within those clusters.

3.1 Exploratory Analysis

3.1.1 Import all relevant libraries

```
#dataframes
import pandas as pd
import numpy as np
import random

#handle requests and json files
import json # deal with json files
```

```
import requests

# Matplotlib and associated plotting modules
from matplotlib import pyplot as plt
import matplotlib.cm as cm
import matplotlib.colors as colors
import seaborn as sns
!pip install folium
import folium
from bs4 import BeautifulSoup
# module to convert an address into latitude and longitude values
!pip install geopy
!pip install geocoder
from geopy.geocoders import Nominatim
import geocoder
#tranforming json file into a pandas dataframe library
from pandas.io.json import json_normalize
#Library for k-mean algoritm
from sklearn.cluster import KMeans
print('Libraries imported.')
```

Listing 2: libraries import

3.1.2 Safe Boroughs in Berlin

According to the client's demand, the apartment is settled in a safe place with lower crime rates. So based on the given data set, I first select the *Total Cases* for each borough and plot the result out in figure 7. As I am actually inter-

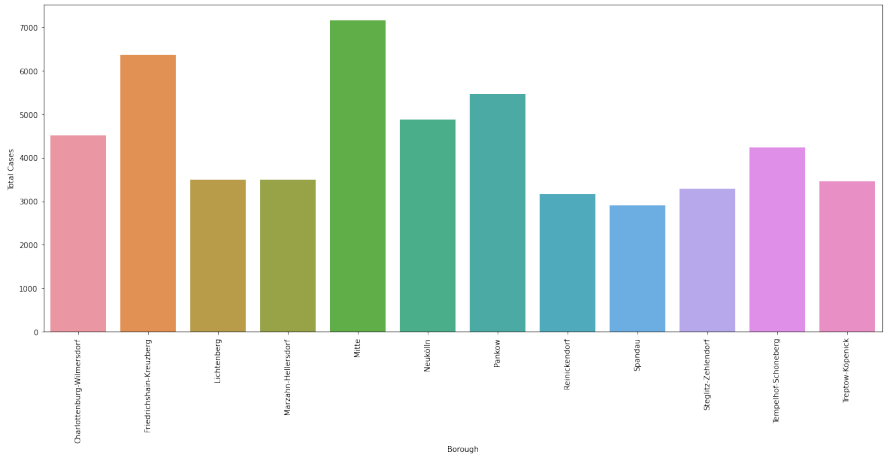


Fig. 7: Total crimes in each boroughs of Berlin

ested in the safety in each postal code area rather than the entire borough, I simply divided the total crime cases evenly to each postal code area according to the entire number of allocated postal codes in that borough. As the crime data transferred in to picture 8, I merge the rental price column at the end

as well. According to the averaged crime cases, I find out all boroughs with

	Borough	PostalCode	Crime Cases	Rental Price per Qm
0	Charlottenburg-Wilmersdorf	10553	128.942857	12.65
1	Charlottenburg-Wilmersdorf	10585	128.942857	12.65
2	Charlottenburg-Wilmersdorf	10587	128.942857	12.65
3	Charlottenburg-Wilmersdorf	10589	128.942857	12.65
4	Charlottenburg-Wilmersdorf	10623	128.942857	12.65

Fig. 8: Averaged crime cases in each boroughs of Berlin

crime cases under mean level and claimed those districts as safe areas. There are three of them: **Charlottenburg-Wilmersdorf**, **Steglitz-Zehlendorf** and **Tempelhof-Schoeneberg**. And their crime numbers are shown in bar graph **??**

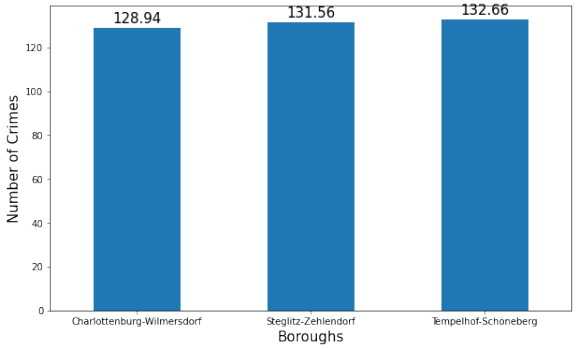


Fig. 9: Safe areas in Berlin with crime cases under average level.

3.1.3 Boroughs with lower Rental Price

The next step is to observe the boroughs with rental price, similar as for crime cases, I further filter out the boroughs with Rental Price lower than the mean level. Two boroughs **Steglitz-Zehlendorf** and **Tempelhof-Schoeneberg** are achieved while the other borough **Charlottenburg-Wilmersdorf** is left out due to its high rental price. Let's take a look at the results. According to the analysis results, the choices of Boroughs are narrowed down to tow options. I further plot the neighborhoods in Berlin and distinguish those in **Steglitz-Zehlendorf** and **Tempelhof-Schoeneberg** from other boroughs. It is shown in map 11. The neighborhoods chosen as candidates for new student apartment are marked in green while the remained areas where either not safe or expensive in living are marked in blue. We can conclude that the potential good spots for the apartment are in the south-west side of Berlin.

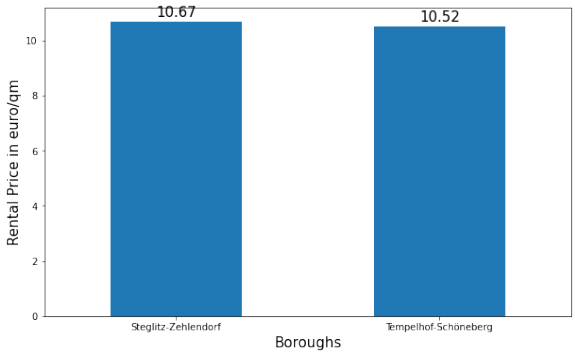


Fig. 10: Boroughs in Berlin with rental price under average level.

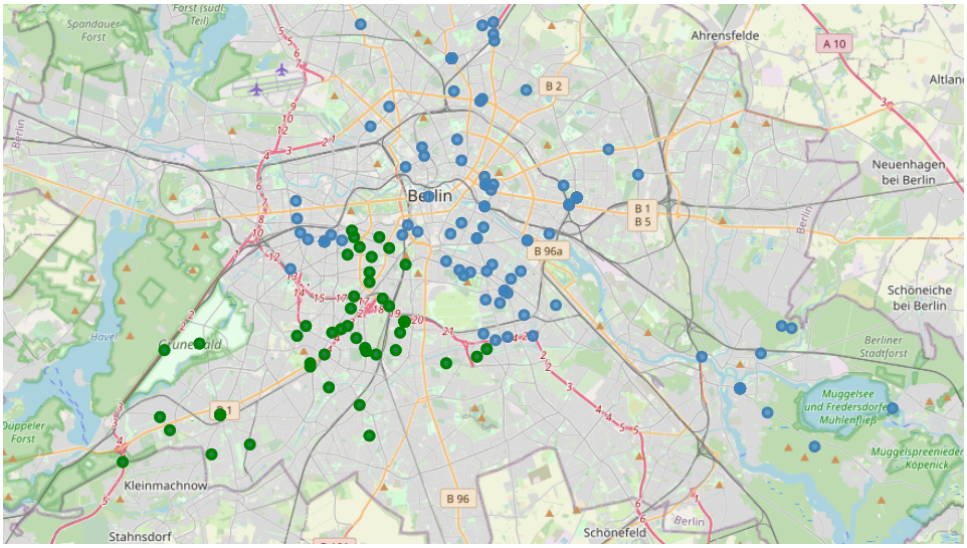


Fig. 11: Neighborhoods in Berlin, Green are the candidates neighborhoods where Blue are the filtered out areas.

3.1.4 Groceries and Coffee Shops nearby

I further explore the venues data according to the client's demand on the surroundings. Based on the requirements, I selected out all groceries (in orange) and coffee shops (in purple) in the boroughs candidates and visualize the results as in figure 12 Based on this map one can suggest, there are plenty coffee shops and groceries in the provided candidates, so there is highly possible that a good spot to be decided.

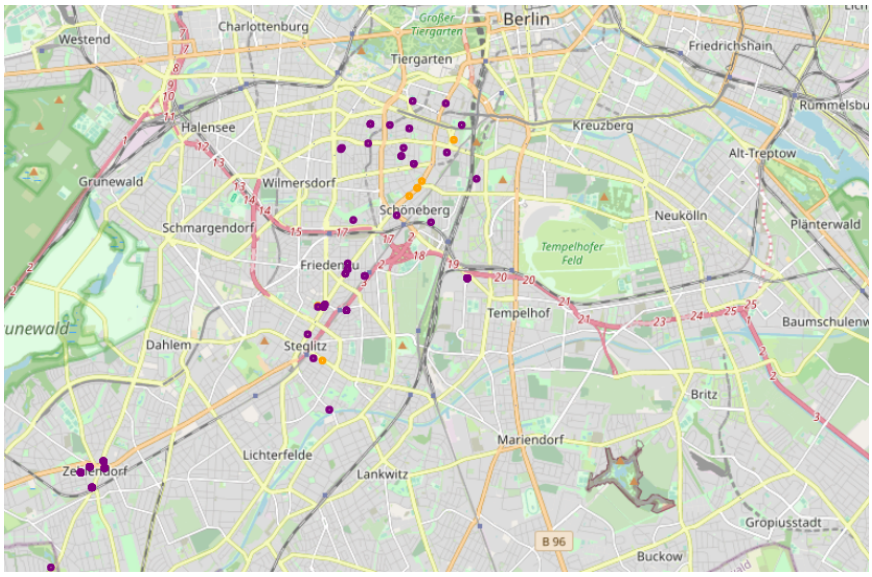


Fig. 12: Groceries and Coffee Shops near the neighborhood candidates.

3.2 K-Means Clustering

In this section, I present the main component of applying a k-means cluster on the venues data set to see if a good cluster can be found for the neighborhoods of new-build student apartment. I followed these steps:

- One hot encoding of the venues
- Generate a k-means cluster, set up k values
- Run clustering of the venues
- Allocate labels to data frame and merge data frame together.

First of all, to handle with categorical data, one hot encoding is carried out for all venue categories and get the data ready for a clustering. Then I set the generate a k-means cluster with up to seven clusters and then fit the data with it. After I merge the rental price and crime cases data together with the venues, one finally achieve a data set as in table 13.

```
#generate a k-means cluster
kclusters = 7
#ignore neiborhood column
berlin_venues_grouped_clustering = berlin_venues_grouped.drop(['
    Neighborhood'], axis=1)
#run the clustering
kmeans = KMeans(n_clusters=kclusters, random_state=0).fit(
    berlin_venues_grouped_clustering)
```

Listing 3: libraries import

	Neighborhood	PostalCode	Crime Cases	Rental Price per Qm	Latitude	Longitude	Cluster Labels	ATM	Accessories Store	Adult Boutique	...	Volleyball Court
0	Charlottenburg-Wilmersdorf	10553	128.942857	12.65	52.499620	13.323160	5	0.0	0.0	0.0	...	0.0
1	Charlottenburg-Wilmersdorf	10585	128.942857	12.65	52.514816	13.304804	5	0.0	0.0	0.0	...	0.0
2	Charlottenburg-Wilmersdorf	10587	128.942857	12.65	52.499620	13.323160	5	0.0	0.0	0.0	...	0.0
3	Charlottenburg-Wilmersdorf	10589	128.942857	12.65	52.499620	13.323160	5	0.0	0.0	0.0	...	0.0
4	Charlottenburg-Wilmersdorf	10623	128.942857	12.65	52.499620	13.323160	5	0.0	0.0	0.0	...	0.0

5 rows × 310 columns

Fig. 13: Clustered venues data with labels, as well as the rental price and crime cases.

A visualization via *folium map* is also presented for a intuitive observation.

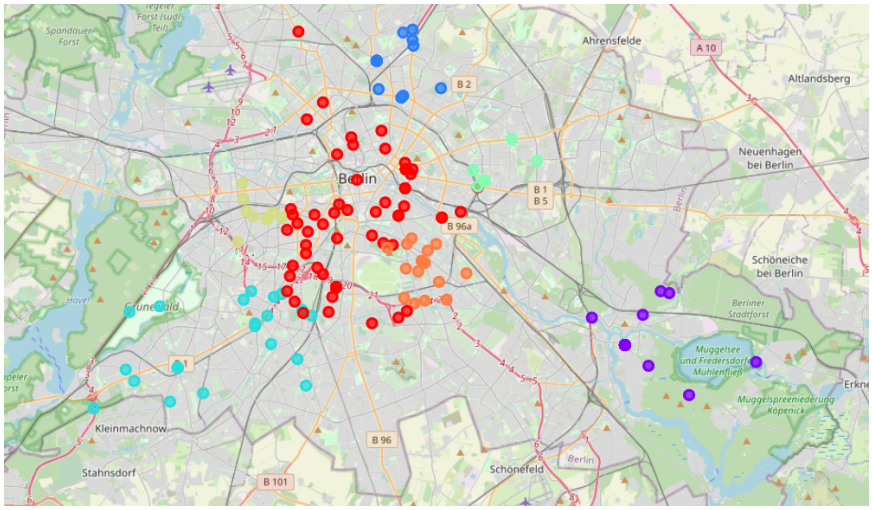
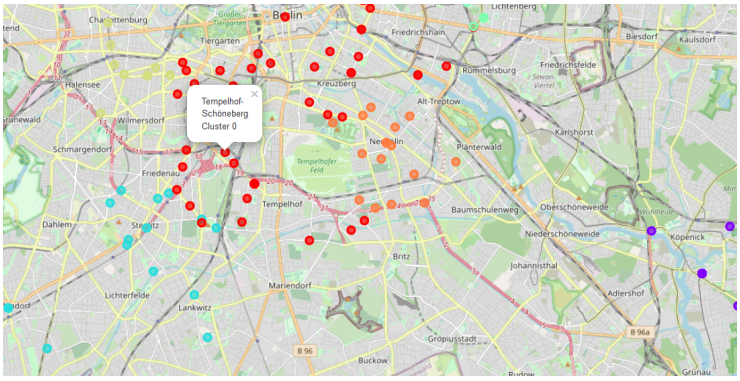


Fig. 14: Clustered venues data in 7 colors.

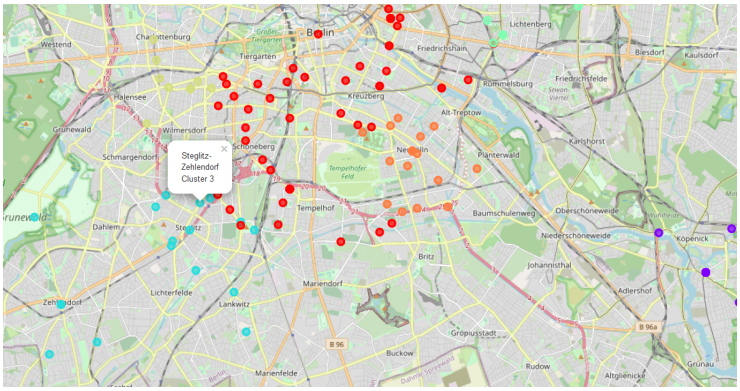
4 Results

In this part, I would discuss on the results based on the k-means clustering and compare it with the previous assumption of the candidates via exploratory analysis. Seven clusters are generated after the clustering and let's take a deeper look at the last map with labeled venues.

It is great to find out that Cluster 1 and Cluster 4 match our candidates for the neighborhoods and two Boroughs **Tempelhof-Schoeneberg** and **Steglitz-Zehlendorf** are determined to be a good selection for neighborhoods of the new student apartment. One can also examine the clustered data set to prove the result:



(a) Cluster 1 contains Tempelhof-Schoeneberg



(b) Cluster 4 contains Steglitz-Zehlendorf

Fig. 15: Cluster 1 stands for neighborhoods in borough Tempelhof-Schoeneberg and Cluster 4 stands for neighborhoods in borough Steglitz-Zehlendorf

	Neighborhood	PostalCode	Longitude	Latitude	Cluster Labels	Crime Cases	Rental Price per Qm
0	Steglitz-Zehlendorf	12249	13.351344	52.425315	3	131.56	10.67
1	Steglitz-Zehlendorf	14193	13.222906	52.458019	3	131.56	10.67
2	Steglitz-Zehlendorf	14197	13.311042	52.467170	3	131.56	10.67
3	Steglitz-Zehlendorf	12161	13.333121	52.466244	3	131.56	10.67
4	Steglitz-Zehlendorf	14167	13.276418	52.422015	3	131.56	10.67

Fig. 16: Cluster 4 contains neighborhoods in Steglitz-Zehlendorf with low rental price and safe surroundings.

5 Conclusion

In this case study, I have explored up to three data sets in Berlin to find a solution for a client, who wish to find a good location for a new student apartment. The requirements include the safety in surrounding areas and the low rental price. Additionally, the client also want the apartment to be close to venues like coffee shops and groceries, both are often considered by students. After a thorough analysis over the data set, we have determine the most likely boroughs and narrow the choices down to 2. Then we go through a k-means clustering algorithm to clustered the venues in Berlin. Based on this result, one can eventually select out the suited neighborhood in this case. Though the entire analysis is still humble and not taking all possible factors into consideration like the land-price, the distance to the schools and public transits, it is still a good example of applying data science to solve real life problems.