Berlin crime rate by district vs. Berlin foursquare venue clusters?

Brief introduction to the project

Subject of this work will be to find out relation or even possible correlation betwen Berlin crime rate in each district and Berlin district clusters, created by machine learning clustering with K-means clustering algorithm of the venues returned by the foursquare API venues search.

Idea is to find out what is the best place to live in Berlin from the standpoint of crime and from the standpoint of different venues that are categorized by the foursquare application. Ideally we would like to live as close as possible to the venues we have interest in (in my case for example theaters, music pubs possible with live music, and areas for leisure like parks) and also to choose district that is as safe as possible from the standpoint of crime.

After clustering I would like to check venues I am personally interested in like theaters, live music pubs and green areas like parks. Maps for these venues will be created and compared with clusters map, crime rate values and locations of the venues I am interested in.

I hope there are some interesting conclusion that could be found in this project.

Describing data that will be used in the project

First data that will be used is the excel table available on the web page of Berlin police https://www.berlin.de/polizei/service/kriminalitaetsatlas/. Table is not suitable for use directly in Python pandas manipulation so first we have to create table with the data that will be used in our project.

There are 7 sheets with the crimes data by the district and areas of the district. Only total data for each disttict is used. From these 7 excel sheets for each year one total sheet 2012-2018 is created. That table is used in our analysis.

Interesting columns that will be used are:

- sum of crimes in the district (Sum of Straftaten -insgesamt-) as Total_sum
- sum of body injures (Sum of Körper-verletzungen -insgesamt-) as Total_body_injuries
- sum of house burglaries (Sum of Wohnraum -einbruch) as Total_thefts
- sum of thefts (Sum of Diebstahl -insgesamt-) as Total_burglaries

Second data that is used is scrapped table from the wikipedia web page for Berlin districts. I wanted to check are name and number of the districts same as the number and names of the district in the excel table used for crime numbers and apart of that total population column is extracted from the same web page.

Wikipedia web page from which second data is extracted is
https://en.wikipedia.org/wiki/Boroughs_and_neighborhoods_of_Berlin

Third data information that will be used for the project are data from the foursqare venue search. This data are available with the special search that is created. Result of the search are venues in the areas that are searched. These venues are categorized by the name, type of the venue (bar, bus station, hotel etc) latitude, longitude and some other categories that we do not need.

From this foursquare data clusters will be created with the help of machine learning algorithm K-means clustering.

Other data that will be used are particular search for the type of the venus I am interested in. Venues that will be searched are theater, pub (it is not possible to find pubs with the live music so search is just for pub) and green area or park. From maps that are created from this data we can see location of the venues we have interest in and visually compare location of this particular venues with the location of districts.

Techniques that are used in the project

Different techniques are used in completing this project.

Excell table for crime data is prepared for the use in Python environment. This is done in excell and PowerBI. Table was structured in a way that it was not possible to use it directly in pandas and fastest and most effective way to prepare table for use in pandas dataframe was in PowerBI.

District table was extracted from wikipedia web page with the help of the web scrapping library BeautifulSoup and code for extracting table from the web page.

Visualization is done in matplotlib as an easiest way to perform simple but effective visualization for our project.

Clustering is done with machine learning algorithm K-means clustering. This is one of the easiest for use and most effective algorithms for clustering.

Venues are inspected with the help of foursquare api requests for venues and type of the venues.

Results of the techniques that are used in the project

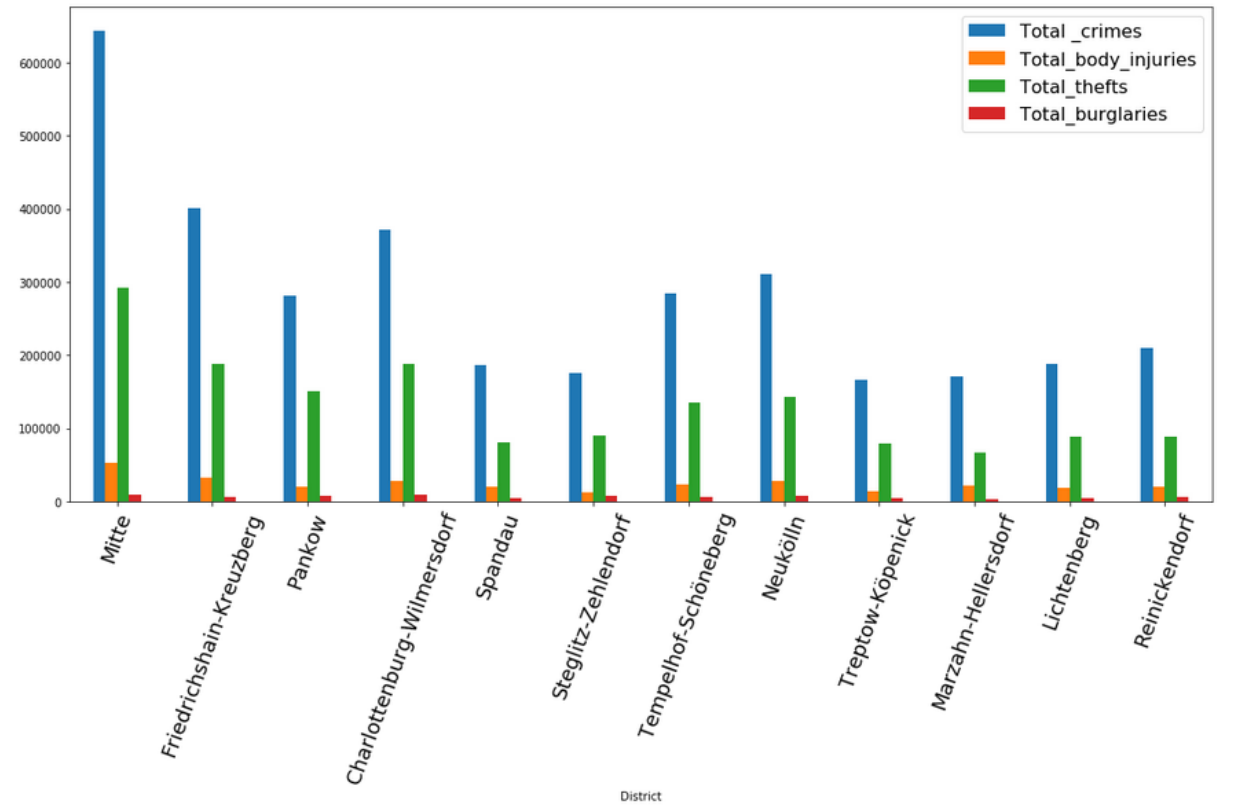
Visualization of the crime data in districts and comparison with the location of the districts

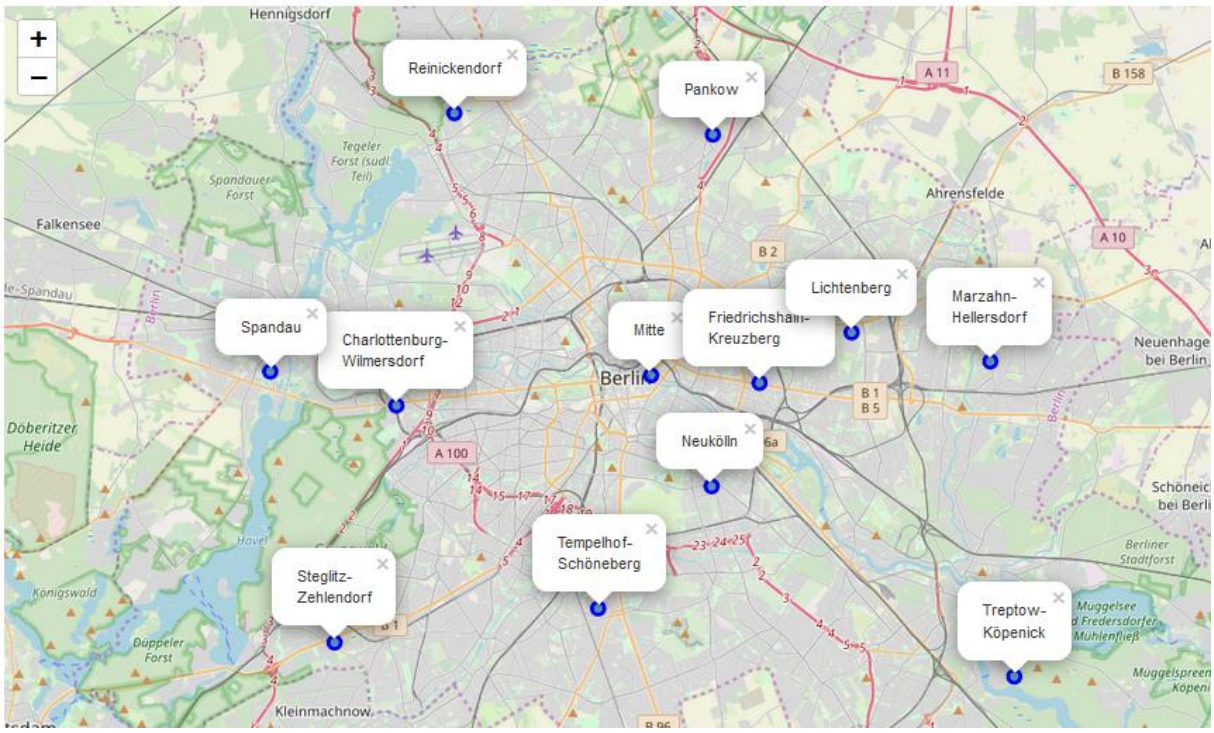
After the table of the crime data is prepared for loading to the pandas dataframe first thing that has to be done is visualization of the data. Visualization is done for each year to check is there any differences between crimes data in each neighborhood year by year.

Since the number of the crimes year by year in each neighborhood is almost the same aggregation of the data is used for further analysis.

Here is the final chart of the crime data:

```
#visualization for the total number of crimes
berlin_crimes_2012_2018.iloc[:,[0,2,3,4,5]].groupby('District', sort=False).sum().plot(kind='bar',
                                                                                       figsize=(18,8))
plt.xticks(fontsize=18, rotation=70)
plt.legend(fontsize=16)
plt.show()
```



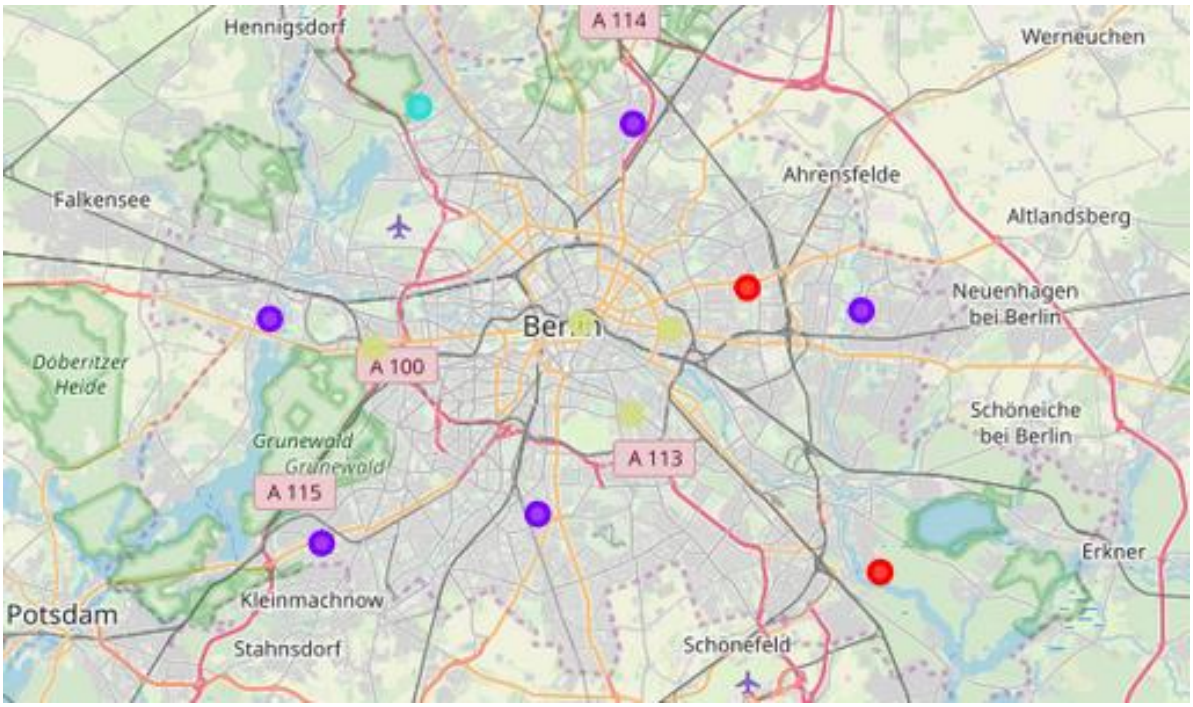


We can see comparing the chart and the map that districts with the most crimes cumulative for the years 2012-2018 are: central Berlin – Mitte, Freidrichshain-Kreuzberg, Charlottenburg-Wilmersdorf and Neukoln. All these districts are centraly located.

Districts with the least amount of crimes are Spandau, Stegllitz_Zehlendorf and Treptow-Kopenick. These are all peripheral Districts of Berlin.

Clustering of the Berlin district according to type of the venues

Machine learning algorithm K-means clustering is used for this part of the project



Cluster 0 - red color - Lichtenberg and Treptow-Kopenick

Cluster 1 - purple color - Marzahn-Hellersdorf, Pankow, Spandau, Steglitz-Zehlendorf, Tempelhof-Schöneberg

Cluster 2 - light blue – Reinickendorf

Cluster 3 – green/yellow color - Charlottenburg-Wilmersdorf, Friedrichshain-Kreuzberg, Mitte, Neukoln.

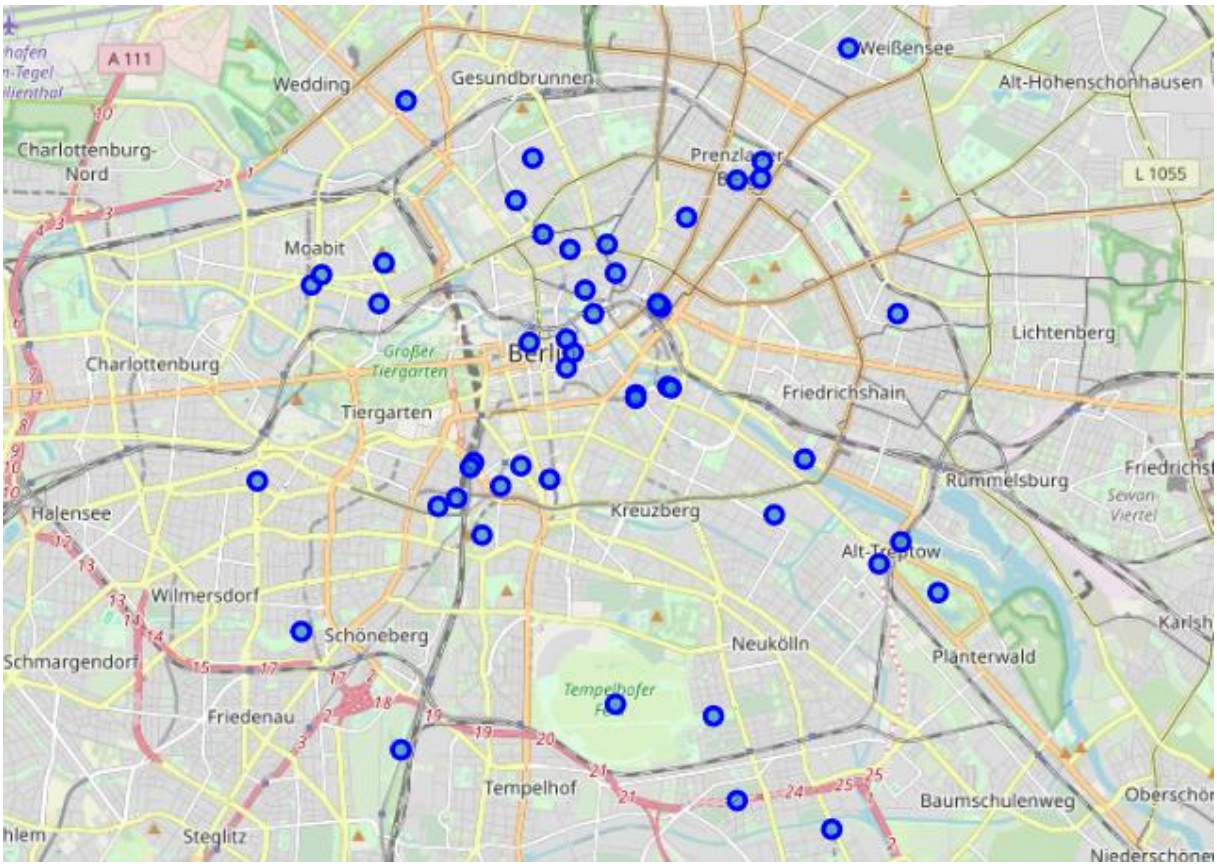
We can see from the clustering that clusters with most of the crimes are all located in the same cluster in this case cluster nr. 3.

Inspection of the particular Berlin venues

Map of the location of theaters

Map of the location of music pubs

Map of the locations of the parks and green areas



Conclusion

As we can see from the clusters map central districts are all located in the cluster 3 and we know from the data crime visualization chart that central districts have more crimes than peripheral districts.

For example Marzahn-Hellersdorf, most eastern part of the city has almost 4 times less crimes than central Berlin or almost 2,5 times less crime that most of the central parts of the city.

It is not perfectly clear what are the distinct features of other clusters. It could be for example that Reinickendorf is separated in distinct cluster due to the most common venue in Reinickendorf and that is hostel. Also distinct feature of the Cluster 1 - purple color one with Marzahn-Hellersdorf, Pankow, Spandau, Steglitz-Zehlendorf, Tempelhof-Schöneberg is probably location of the many supermarkets and commercial activities in that area.

General conclusion is that central, most vibrant parts of the city that means parts with the biggest number of the interesting venues, are at the same time parts of the city with the highest crime rate and they are all located in the same cluster, so there is for sure correlation between number of crimes and number of venues in the area. What is the strength of the correlation is hard to say only from this data.

From the maps of the interesting venues we can see that most of the activities are happening, as expected, in the central part of the town. There are bit more music pubs and theaters in the wider western area of the city and bit more parks and green areas in the souuth-eastern part of the city but in general most of this venues are centrally located.

Regarding our question it seems there is a clear trade-off between security of the place and number of ineresting venues. Most vibrant parts of the city are also parts with the biggest crime rate. Peripheral parts of the town are parts with smalest crime rate and also less activities and less interesting venues.

Choice of the city district to live according to this data is not an easy one and there are no clear winner, but I personally would choose eastern part of the city Lichtenberg and Marzahn-Hellersdorf. They are as close as possible to the center and also in the area with the least crime rates. Additional important thing in this choice is vicinity of the green areas that are located mostly in the south-eastern parts of the city.