Crime and Neighborhoods of Tallinn, Estonia

Applied Data Science Capstone Project

Introduction and description of the problem

Tallinn is the capital and most populous city of Estonia. Located in the northern part of the country, on the shore of the Gulf of Finland of the Baltic Sea, it has a population of ca 430 thousand. Tallinn is the main financial, industrial and cultural centre of Estonia. It is located 80 kilometres south of Helsinki, Finland, 320 kilometres west of Saint Petersburg, Russia, 301 kilometres north of Riga, Latvia, and 380 kilometres east of Stockholm, Sweden. As Tallinn's Old Town is one of the best preserved medieval cities in Europe - it is listed as a UNESCO World Heritage Site - the whole city boasts at the same time with vibrant cultural and economic life. During the last 15 or so years, Tallinn has grown into being one of the best cities for start-up businesses, which sets new standards to the safety and security of the city to attract foreign talent. One of the safety measures of a city is its crime rate. Although the number of crimes in Tallinn has declined considerably over the years, it is still worth to investigate, how the different districts compare to each other in this respect. So I will seek to answer the following questions:

1. What districts have the highest crime rate?
2. Which neighborhood data correlates to crime level?
3. Using Foursquare data, what venues are most common in the selected neighborhoods?

By definition, Tallinn is divided into eight districts which are further divided into 84 subdistricts or neighborhoods. My aim is to conduct the analysis on the neighborhood level.

Data

Necessary data will be obtained from the following sources:

1. Estonian Open Data (crime data): https://avaandmed.eesti.ee
2. Tallinn in numbers (population data): https://www.tallinn.ee/eng/g2677s126569
3. Tallinn Geospatial Data (neighborhood geometry): https://www.tallinn.ee/est/geoportaal/Andmed
4. Statistics Estonia (income statistics): https://www.stat.ee/en
5. Foursquare Developers Access to venue data: https://foursquare.com/

Also, as local geospatial data is presented in Lambert coordinates, it is necessary convert them to WGS84 coordinates which shall be separately done with Estonian Land Board's application (https://www.maaamet.ee/rr/geo-lest/).

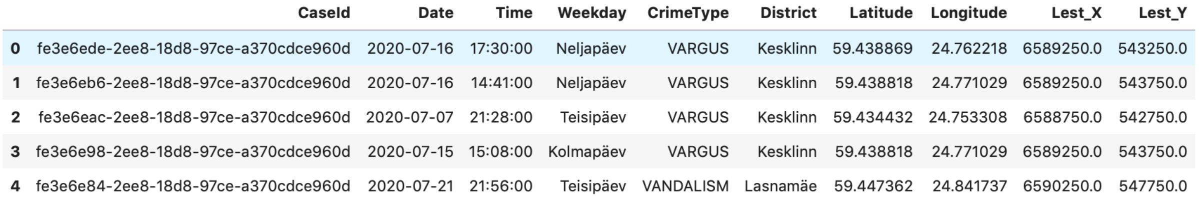
Libraries to be used: pandas, geopandas, numpy, scipy, sklearn, matplotlib, seaborn, folium, geopy.

Methodology - data loading, cleaning and wrangling

Data for the particular analysis was obtained from several sources as seen above, but will fall into the following main categories:

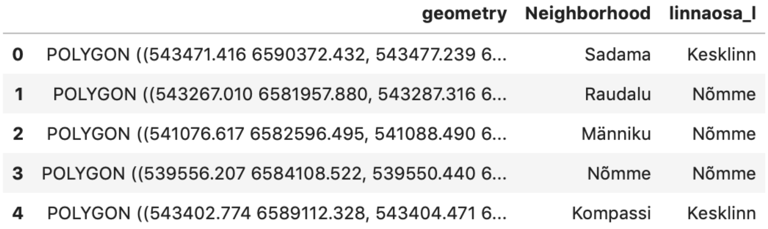
1. Crime data, including information on date, crime type, district and geolocation of the crime. Data source is Estonian Open Data, where the csv files on crimes committed in public space and crimes against property are available. These two datasets were merged initially in Excel

(including some operations like geocoordinates conversion and correction of words including vocals with "umlaut"s) and the loaded to this notebook.



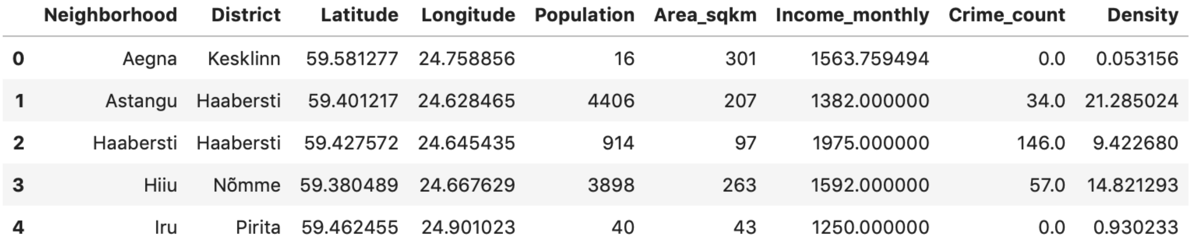
	CaseId	Date	Time	Weekday	CrimeType	District	Latitude	Longitude	Lest_X	Lest_Y
0	fe3e6ede-2ee8-18d8-97ce-a370cdce960d	2020-07-16	17:30:00	Neljapäev	VARGUS	Kesklinn	59.438869	24.762218	6589250.0	543250.0
1	fe3e6eb6-2ee8-18d8-97ce-a370cdce960d	2020-07-16	14:41:00	Neljapäev	VARGUS	Kesklinn	59.438818	24.771029	6589250.0	543750.0
2	fe3e6eac-2ee8-18d8-97ce-a370cdce960d	2020-07-07	21:28:00	Teisipäev	VARGUS	Kesklinn	59.434432	24.753308	6588750.0	542750.0
3	fe3e6e98-2ee8-18d8-97ce-a370cdce960d	2020-07-15	15:08:00	Kolmapäev	VARGUS	Kesklinn	59.438818	24.771029	6589250.0	543750.0
4	fe3e6e84-2ee8-18d8-97ce-a370cdce960d	2020-07-21	21:56:00	Teisipäev	VANDALISM	Lasnamäe	59.447362	24.841737	6590250.0	547750.0

2. Neighborhood geometries, which were obtained from Tallinn geospatial open data.



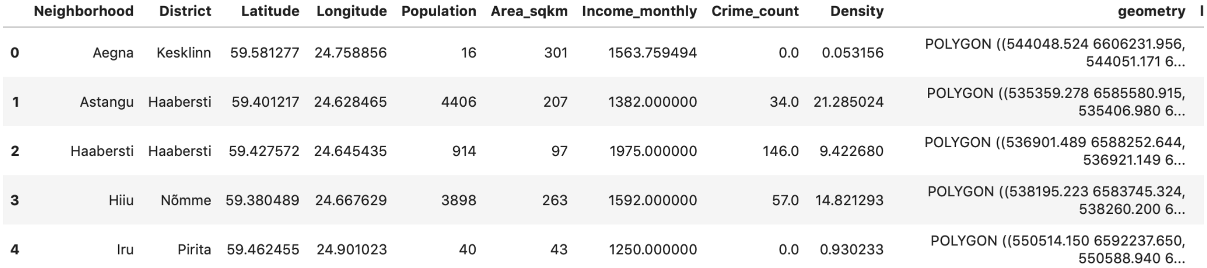
	geometry	Neighborhood	linnaosa_l
0	POLYGON ((543471.416 6590372.432, 543477.239 6...	Sadama	Kesklinn
1	POLYGON ((543267.010 6581957.880, 543287.316 6...	Raudalu	Nõmme
2	POLYGON ((541076.617 6582596.495, 541088.490 6...	Männiku	Nõmme
3	POLYGON ((539556.207 6584108.522, 539550.440 6...	Nõmme	Nõmme
4	POLYGON ((543402.774 6589112.328, 543404.471 6...	Kompassi	Kesklinn

3. Neighborhood data, including information on population and size of the neighborhoods (from Tallinn Open Data) and monthly income per neighborhood (from Estonian Statistics).



	Neighborhood	District	Latitude	Longitude	Population	Area_sqkm	Income_monthly	Crime_count	Density
0	Aegna	Kesklinn	59.581277	24.758856	16	301	1563.759494	0.0	0.053156
1	Astangu	Haabersti	59.401217	24.628465	4406	207	1382.000000	34.0	21.285024
2	Haabersti	Haabersti	59.427572	24.645435	914	97	1975.000000	146.0	9.422680
3	Hiiu	Nõmme	59.380489	24.667629	3898	263	1592.000000	57.0	14.821293
4	Iru	Pirita	59.462455	24.901023	40	43	1250.000000	0.0	0.930233

Main aim of data preparation was to associate the crimes with respective neighborhoods. So with the last step of data preparation, the merged dataframe was created by using the spatial join method.



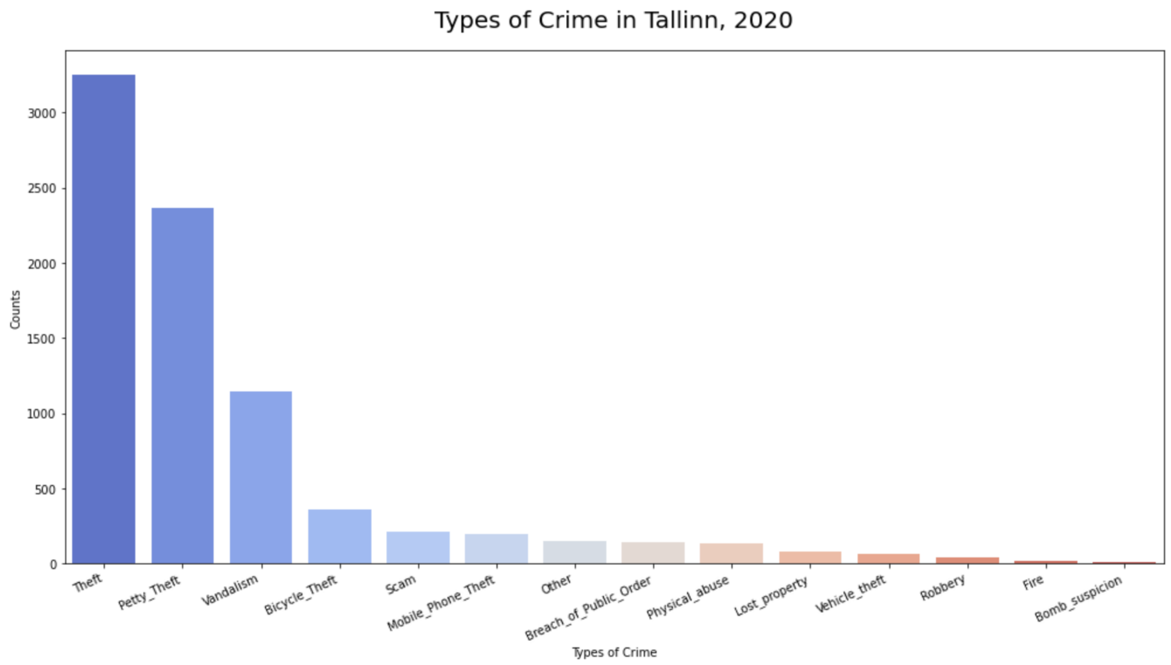
	Neighborhood	District	Latitude	Longitude	Population	Area_sqkm	Income_monthly	Crime_count	Density	geometry	l
0	Aegna	Kesklinn	59.581277	24.758856	16	301	1563.759494	0.0	0.053156	POLYGON ((544048.524 6606231.956, 544051.171 6...	
1	Astangu	Haabersti	59.401217	24.628465	4406	207	1382.000000	34.0	21.285024	POLYGON ((535359.278 6585580.915, 535406.980 6...	
2	Haabersti	Haabersti	59.427572	24.645435	914	97	1975.000000	146.0	9.422680	POLYGON ((536901.489 6588252.644, 536921.149 6...	
3	Hiiu	Nõmme	59.380489	24.667629	3898	263	1592.000000	57.0	14.821293	POLYGON ((538195.223 6583745.324, 538260.200 6...	
4	Iru	Pirita	59.462455	24.901023	40	43	1250.000000	0.0	0.930233	POLYGON ((550514.150 6592237.650, 550588.940 6...	

Now I was able to carry out the planned map visualisation and analytics.

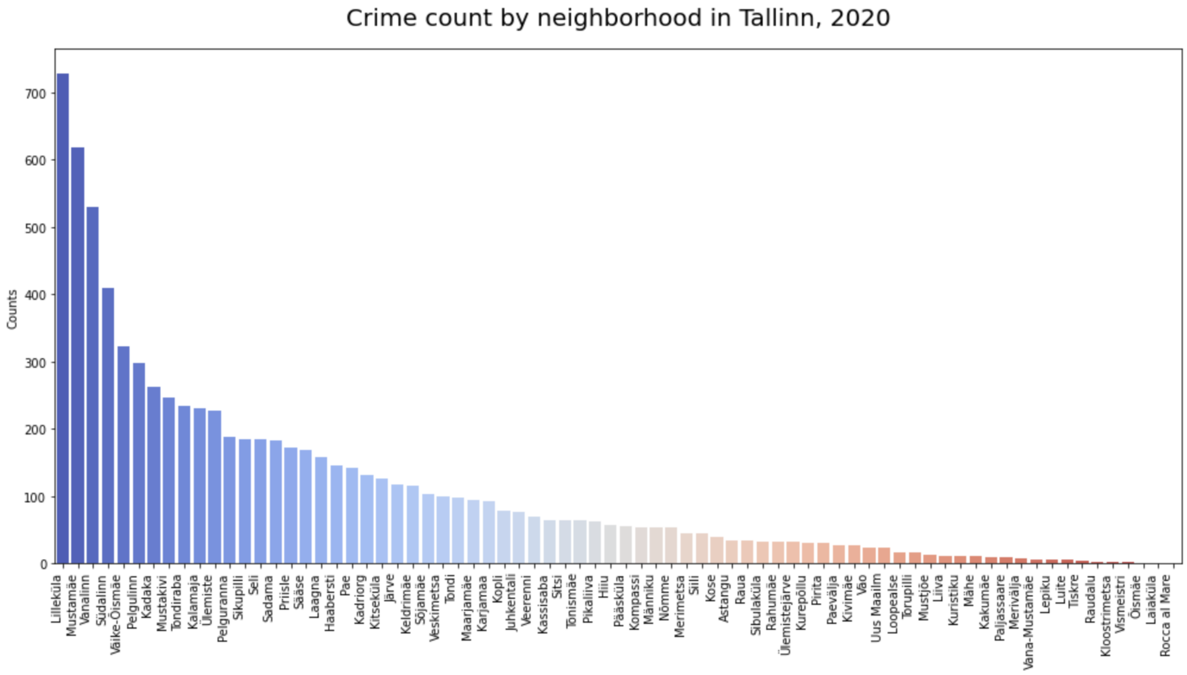
Data analysis – visualisation and clustering

1. **Visualisations**

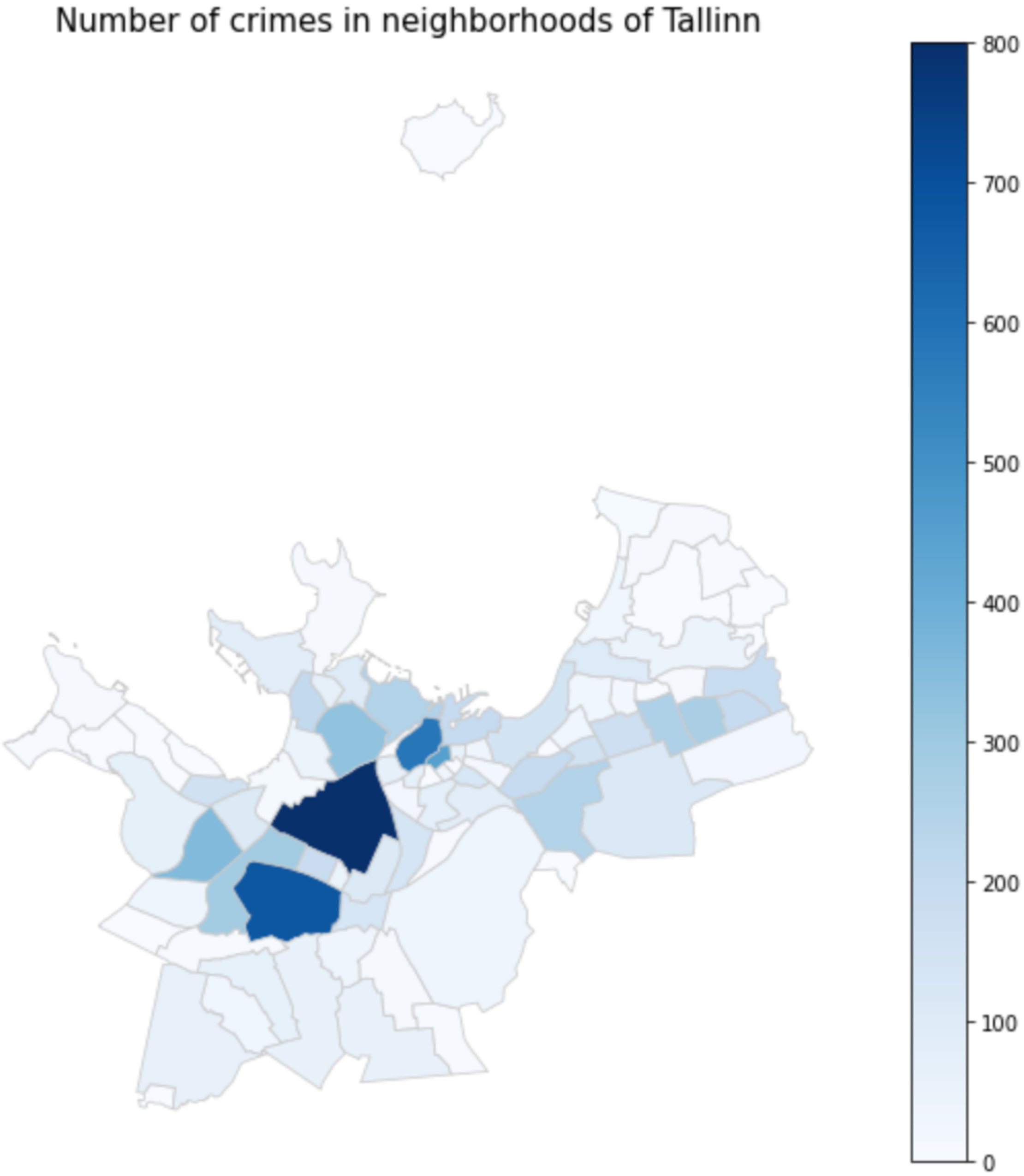
Under this subsection I performed the first step of data analysis by visualising the crime data of Tallinn neighborhoods. From the crime dataframe we know that the total number of crimes committed in 2020 in public spaces and anti-property crimes in Tallinn was 8171. Using seaborn countplot, the distribution of crimes was the following ...



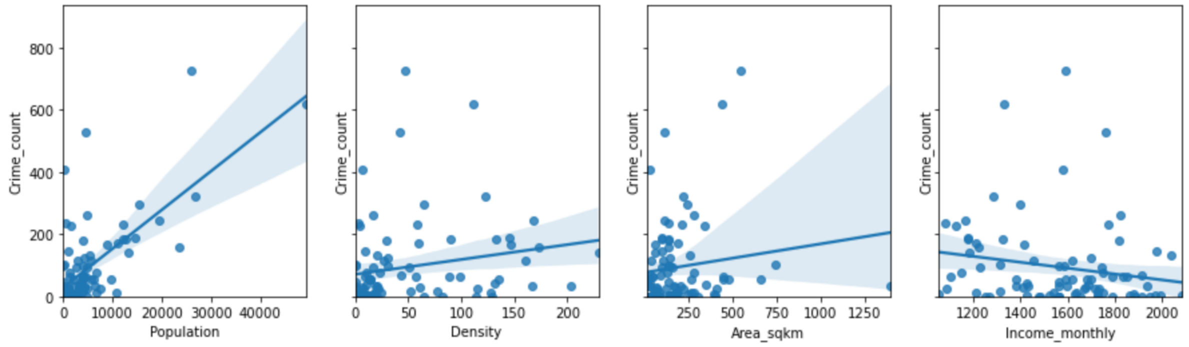
... whereas the distribution between the neighborhoods was:



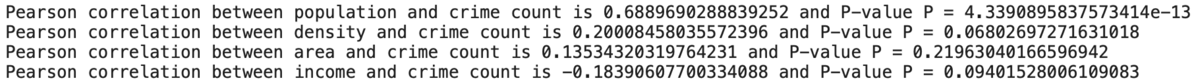
To get a better grasp how the crimes per neighborhood are distributed geospatially, a choropleth map was created – the darker the colour, the larger the crime count:



As a next step it was important to decide on which neighborhoods to concentrate on while continuing with analysis, ie to find out the most popular venues of a neighborhood. For that a correlation analysis was done between the crime count and the available general data of neighborhoods – population, area, population density, and income. The results can be seen from the following scatterplots…



… and the respective calculations:

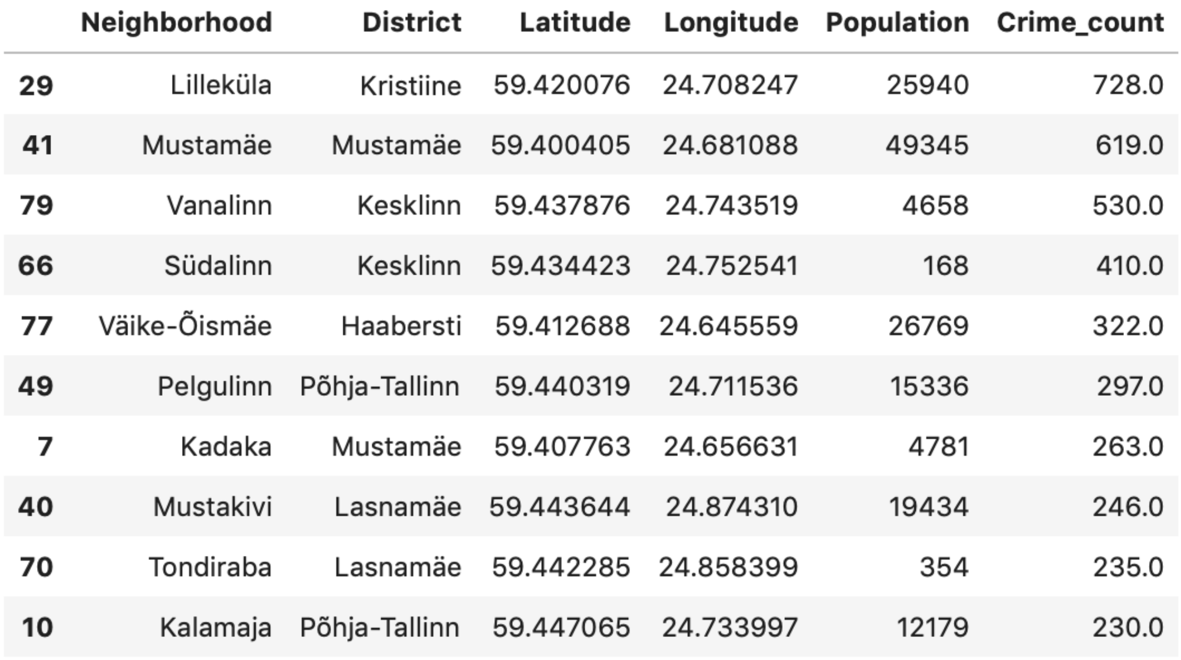


```
Pearson correlation between population and crime count is 0.6889690288839252 and P-value P = 4.3390895837573414e-13
Pearson correlation between density and crime count is 0.20008458035572396 and P-value P = 0.06802697271631018
Pearson correlation between area and crime count is 0.13534320319764231 and P-value P = 0.21963040166596942
Pearson correlation between income and crime count is -0.18390607700334088 and P-value P = 0.09401528006109083
```

As we can see, the only strong and meaningful link is between the crime count and population size of a neighborhood. Therefore it was decided that in the next step we'll proceed with the analysis of top 10 neighborhoods with the highest crime count and see, what illustrates these neighborhoods. Assumption being that higher population count describes better what drives crime.

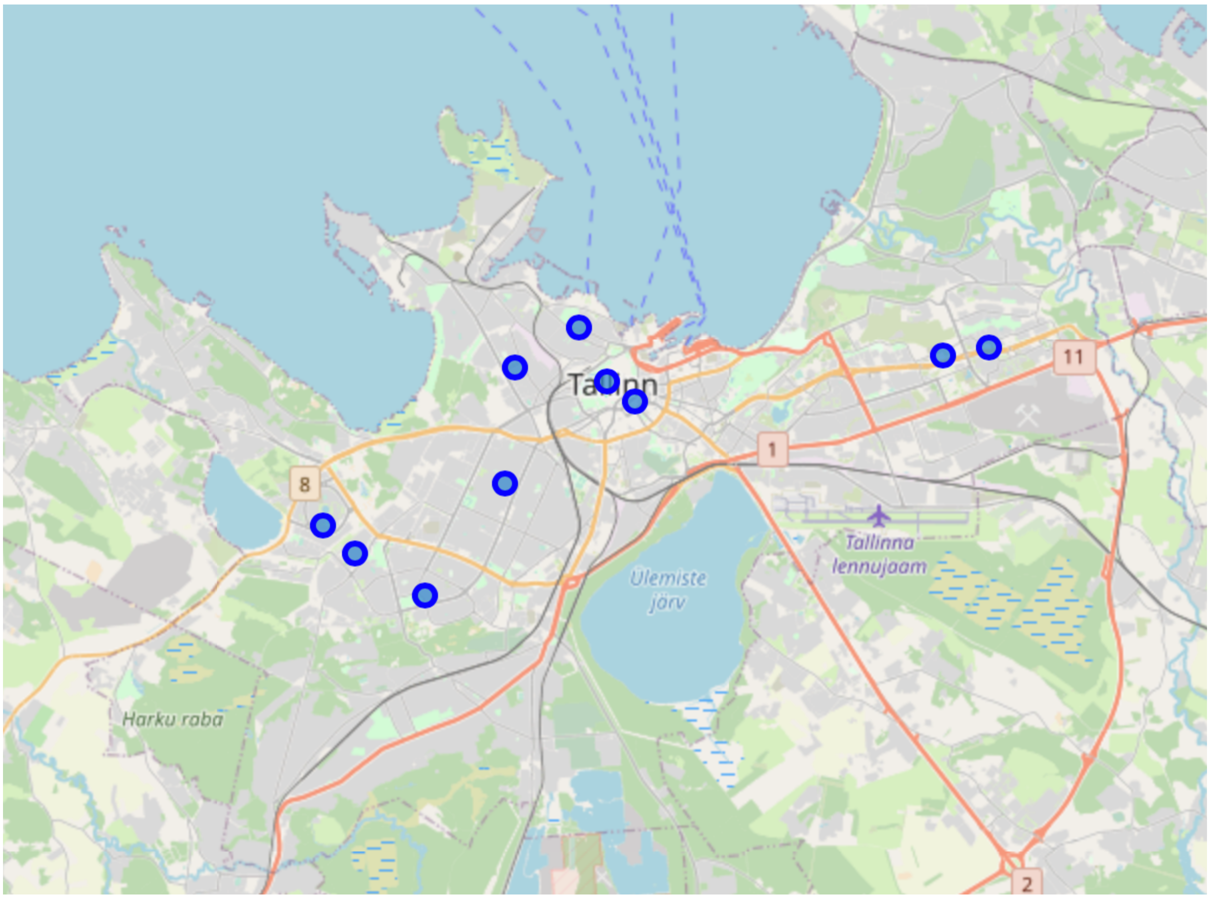
2. Clustering

Now a separate dataframe was created with the data of ten neighborhoods having the highest crime count…

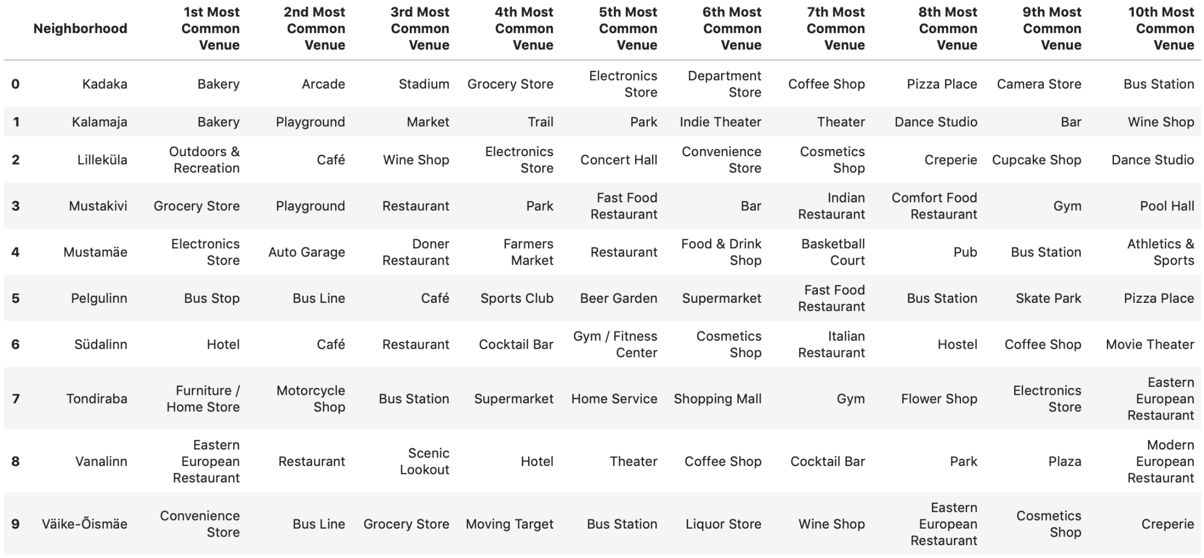


	Neighborhood	District	Latitude	Longitude	Population	Crime_count
29	Lilleküla	Kristiine	59.420076	24.708247	25940	728.0
41	Mustamäe	Mustamäe	59.400405	24.681088	49345	619.0
79	Vanalinn	Kesklinn	59.437876	24.743519	4658	530.0
66	Südalinn	Kesklinn	59.434423	24.752541	168	410.0
77	Väike-Õismäe	Haabersti	59.412688	24.645559	26769	322.0
49	Pelgulinn	Põhja-Tallinn	59.440319	24.711536	15336	297.0
7	Kadaka	Mustamäe	59.407763	24.656631	4781	263.0
40	Mustakivi	Lasnamäe	59.443644	24.874310	19434	246.0
70	Tondiraba	Lasnamäe	59.442285	24.858399	354	235.0
10	Kalamaja	Põhja-Tallinn	59.447065	24.733997	12179	230.0

… with their locations on Tallinn map by superimposing the neighborhoods on the map of Tallinn using geopy:

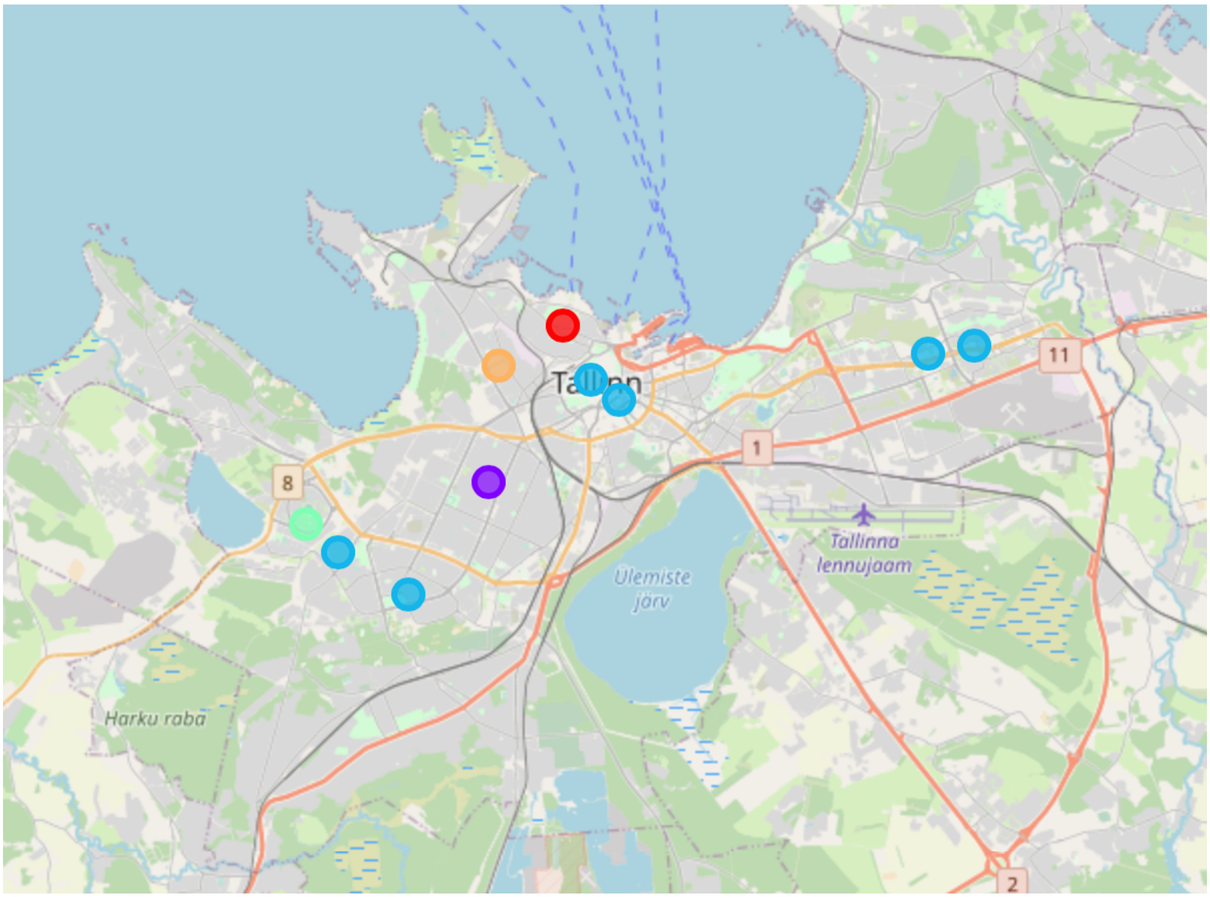


After visualizing our targeted neighborhoods I created a dataframe consisting of these neighborhoods and then leveraging foursquare API I found out the top surrounding venues within 500 meters radius. This results in Json file displaying all the nearby venues which was further cleaned to form a dataframe showing all the top venues neighborhood has to offer with their location coordinates. Then after forming the dataframe the values were one hot encoded to deploy unsupervised machine learning model of K-means clustering to categorize neighborhoods into different clusters based on their similarity.

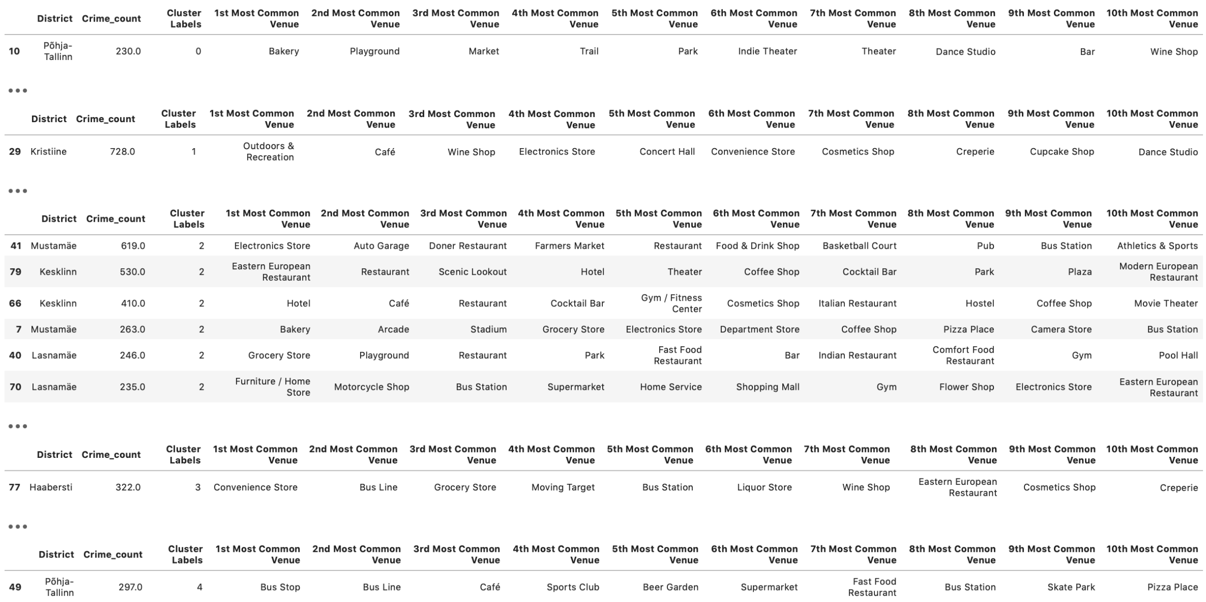


	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Kadaka	Bakery	Arcade	Stadium	Grocery Store	Electronics Store	Department Store	Coffee Shop	Pizza Place	Camera Store	Bus Station
1	Kalamaja	Bakery	Playground	Market	Trail	Park	Indie Theater	Theater	Dance Studio	Bar	Wine Shop
2	Lilleküla	Outdoors & Recreation	Café	Wine Shop	Electronics Store	Concert Hall	Convenience Store	Cosmetics Shop	Creperie	Cupcake Shop	Dance Studio
3	Mustakivi	Grocery Store	Playground	Restaurant	Park	Fast Food Restaurant	Bar	Indian Restaurant	Comfort Food Restaurant	Gym	Pool Hall
4	Mustamäe	Electronics Store	Auto Garage	Doner Restaurant	Farmers Market	Restaurant	Food & Drink Shop	Basketball Court	Pub	Bus Station	Athletics & Sports
5	Pelgulinn	Bus Stop	Bus Line	Café	Sports Club	Beer Garden	Supermarket	Fast Food Restaurant	Bus Station	Skate Park	Pizza Place
6	Südalinn	Hotel	Café	Restaurant	Cocktail Bar	Gym / Fitness Center	Cosmetics Shop	Italian Restaurant	Hostel	Coffee Shop	Movie Theater
7	Tondiraba	Furniture / Home Store	Motorcycle Shop	Bus Station	Supermarket	Home Service	Shopping Mall	Gym	Flower Shop	Electronics Store	Eastern European Restaurant
8	Vanalinn	Eastern European Restaurant	Restaurant	Scenic Lookout	Hotel	Theater	Coffee Shop	Cocktail Bar	Park	Plaza	Modern European Restaurant
9	Väike-Õismäe	Convenience Store	Bus Line	Grocery Store	Moving Target	Bus Station	Liquor Store	Wine Shop	Eastern European Restaurant	Cosmetics Shop	Creperie

Based on the analysis using K-means clustering with the help of foursquare data, I was able to narrow down the hunt for neighborhoods with highest crime count in Tallinn clustered based on what each neighborhood has to offer. The different color represents the different types of neighborhoods based on their similarity of venues city has to offer.



3. Examining the clusters



	District	Crime_count	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
10	Põhja-Tallinn	230.0	0	Bakery	Playground	Market	Trail	Park	Indie Theater	Theater	Dance Studio	Bar	Wine Shop

•••

	District	Crime_count	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
29	Kristiine	728.0	1	Outdoors & Recreation	Café	Wine Shop	Electronics Store	Concert Hall	Convenience Store	Cosmetics Shop	Creperie	Cupcake Shop	Dance Studio

•••

	District	Crime_count	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
41	Mustamäe	619.0	2	Electronics Store	Auto Garage	Doner Restaurant	Farmers Market	Restaurant	Food & Drink Shop	Basketball Court	Pub	Bus Station	Athletics & Sports
79	Kesklinn	530.0	2	Eastern European Restaurant	Restaurant	Scenic Lookout	Hotel	Theater	Coffee Shop	Cocktail Bar	Park	Plaza	Modern European Restaurant
66	Kesklinn	410.0	2	Hotel	Café	Restaurant	Cocktail Bar	Gym / Fitness Center	Cosmetics Shop	Italian Restaurant	Hostel	Coffee Shop	Movie Theater
7	Mustamäe	263.0	2	Bakery	Arcade	Stadium	Grocery Store	Electronics Store	Department Store	Coffee Shop	Pizza Place	Camera Store	Bus Station
40	Lasnamäe	246.0	2	Grocery Store	Playground	Restaurant	Park	Fast Food Restaurant	Bar	Indian Restaurant	Comfort Food Restaurant	Gym	Pool Hall
70	Lasnamäe	235.0	2	Furniture / Home Store	Motorcycle Shop	Bus Station	Supermarket	Home Service	Shopping Mall	Gym	Flower Shop	Electronics Store	Eastern European Restaurant

•••

	District	Crime_count	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
77	Haabersti	322.0	3	Convenience Store	Bus Line	Grocery Store	Moving Target	Bus Station	Liquor Store	Wine Shop	Eastern European Restaurant	Cosmetics Shop	Creperie

•••

	District	Crime_count	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
49	Põhja-Tallinn	297.0	4	Bus Stop	Bus Line	Café	Sports Club	Beer Garden	Supermarket	Fast Food Restaurant	Bus Station	Skate Park	Pizza Place

The description of clusters is as follows:

- Cluster 0 (red dot on map) – a lot of public space and go-to places in your free time. Shopping is not dominant in this cluster, but chilling out is;
- Cluster 1 (purple dot) – a cluster close to city center, firstly a green region, but shopping venues are important;
- Cluster 2 (light blue dots) – services and shopping for the population. What stands out – it combines the neighborhoods located in "bedroom suburbs" (farther out from the center) as well city center and old city neighborhood, where typically tourism concentrates to (not that it was a case in 2020);
- Cluster 3 (light green) – number one venue is a suburb shopping center with definitely attracts crime, mainly theft;
- Cluster 4 (yellow dot) – also a lot of public space, but not as "hippy" as cluster 0.

Discussion

The methodology of the project is based on the neighborhood data of city of Tallinn using foursquare API and integrating it with crime count. As Tallinn is willing to become more and more attractive as a living environment not only for its current residents but to the prospective incomers, it is worth knowing how neighborhoods rank is respect of crime statistics. And as opposed to general focus of finding the safest neighborhoods, it was decided to explore what describes the top ten neighborhoods with highest crime count.

For this project I explored the crime statistics of neighborhoods of Tallinn, joined (or merged) it with neighborhoods' geospatial data and general data like population, area and income. Thereafter I conducted exploratory data analysis and visualisation on the different types of crime committed in different neighborhoods and also the analysis correlating the crime count to other neighborhood data. Based on this data I decided to narrow down to explore further the top ten neighborhoods with highest crime rate as it with reasonable degree of confidence described the trends of most populous neighborhoods of the city.

Keeping in mind that the main crime type was theft and assuming that "the bigger the neighborhood, the more crime", I went on clustering the neighborhoods around their most popular venues and was able to conclude that these neighborhoods are whether:

- A large neighborhood population wise with a lot of public space with pastime amenities (bakeries, cafes etc) and/or shopping facilities for assumingly local people (as they are lower ranked in popularity); or
- A large neighborhood with significant amount of shopping, attracting therefore assumingly crowds also from other neighborhoods; or
- City centre / Old Town which are always and typically attractive places for shops and services for general public.

Conclusion

Using a combination of datasets from the City of Tallinn and Estonian Open Data and Foursquare venue data I was able to analyse, discover and describe neighborhoods, crime, population data and statistically describe quantitatively venues by locations of interest.

I am not sure whether the findings here are detailed enough to help anyone to choose his or her living place – average price of its square meter is always an argument as well for example, or the distance to the closest school – but hopefully I was able to show that if you keep your eye on your belongings you will be safe anywhere in this nice city.