Charles University
Faculty of Social Sciences
Institute of Economic Studies

Data Processing in Python
SS 2022/2023

Project Report: Examination of Criminality in the Czech Republic based on Socio-Economic Determinants

Tomáš Barhoň, Radim Plško

August 2023

1 Introduction

This report provides an overview of a project conducted by Tomáš Barhoň and Radim Plško for the Data Processing in Python class in the summer semester 2022/23. The project aims to study the impact of socio-economic indicators on the level of economic criminality in different regions of the Czech Republic ("obce s rozšířenou působností").

2 Data Source

The primary data source for this project is the https://kriminalita.policie.cz/ API, which provides information about various types of crimes and their precise geographical locations ("points of crimes"). Therefore, in order to be able to work with crimes on the level of regions ("ORP"), we needed to fit these points of crime to the geographical locations of ORPs.

The socio-economic data were obtained from various open-data sources, but mainly from PAQ research (https://www.datapaq.cz/) where all data were recorded on the level of ORP from the beginning which was the most convenient for us.

The study focused on the following four socio-economic indicators:

1. Lidé v exekuci (2021) - The percentage of people with foreclosure.
2. Podíl lidí bez středního vzdělání (2021) - The percentage of people without completed high school education.
3. Domácnosti čerpající přídavek na živobytí (2020) - The percentage of households receiving social benefits.
4. Propadání (průměr 2015–2021) - The percentage of children that obtain a grade of 5 from any subject at the end of the summer semester.

In order to be able to conduct such analysis we made some assumptionts to our data. As there was not enough data for this specific granuality we had to take data from not exactly matching time periods. Our assumption is that the socio-economical variables that we have chosen are relatively stable over time thus we only tried to get the data from the correct period to some plausible extent.

3 Crime Data

The crime data was subset to meet specific conditions. The crimes included in the study are illegal, verifiable, and of an economic nature, such as thefts and burglaries - to be most relevant to the case of chosen specific socio-economic variables. The data was analyzed for the period from 2021 to June 2023, yielding about 500,000 criminal records. An import step in the analysis is that we did divide the ammounts in each ORP by its population as not ORPs do have the population which would leed to incorrect conclusions. It might be suggested that the data should be also for instance divided by the number of months in which it was analyzed but we did not make this choice at the time of our analysis as

it does not affect our research. However, if anyone would like to compare it with different analysis this change would have to be considered.

4 Code Overview

In our code we utilized multiple visualization and data manipulation packages such as numpy, pandas, seaborn, matplotlib, folium, geopandas, pytest, shapely and two our modules visualizer, data API downloader.

We used jupyter notebookes for most of our visualizations as the final output of our project. However, we wanted to make them as clean as possible and wanted to help others with similar research ideas. Thus we created two our own modules that do have classes which can perform lot of tasks useful even outside of our project. We used numpy style documentation to document them. Our main output is a set of geographical visualization which employ a combination of tabular data with geojson created from a geographical shapefile downloaded from the "ČZUK"website.

The main notebook where you can find all our analysis for this report is app/main.ipynb but there are 3 others starting with the name how_to_*.ipynb that are basically tutorials how to use the modules and their classes in your own analysis.

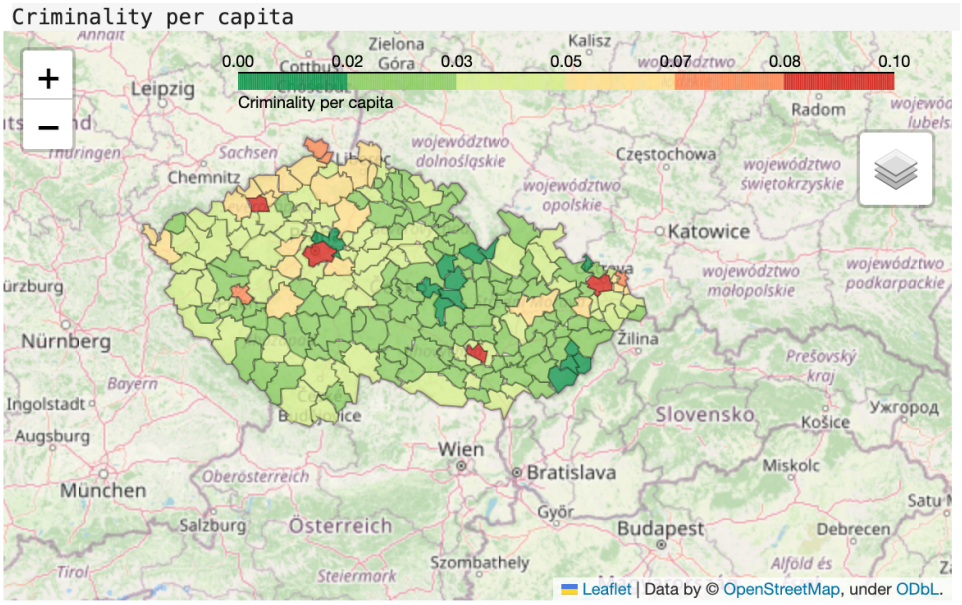
All of the visual outputs can be found in the Project_Report folder.

5 Findings

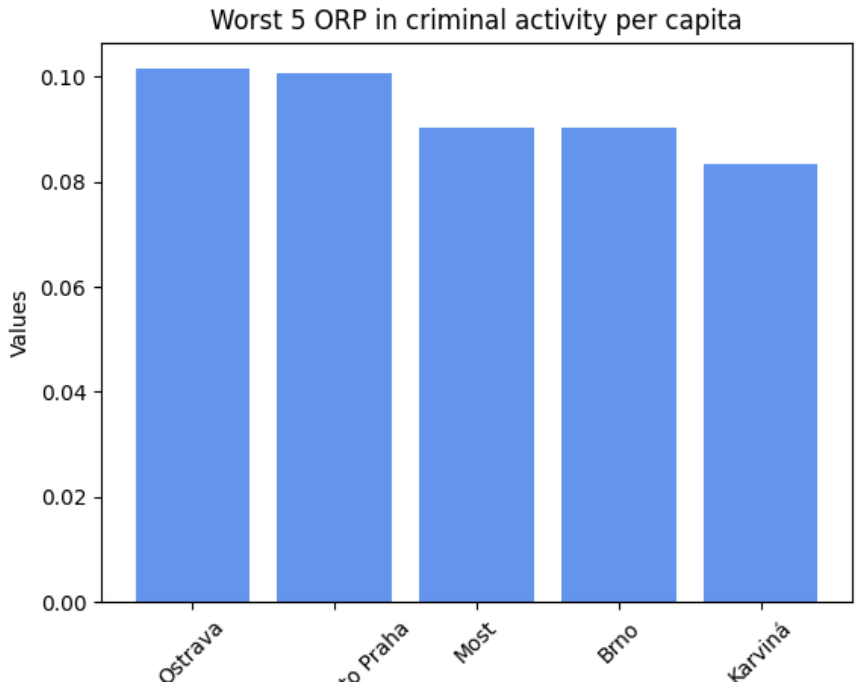
After our first try to plot criminality with folium in each of the ORPs we immediately realized that we forgot to divide the data by the population of the regions. To our suprise this did not extremely transform the view we have seen before.

5.1 Criminality per capita

One of the most interesting findings of the whole project was that the population density seems to be quite a strong determinant of the criminal activity. There is a lot of reasoning we can find for that. For example there is a much wealthier population in cities, that can lure a lot of the robbers. Also the more crowded places will hide quite a lot of pickpocketing. The last but quite important note is that a lot of companies do have their head office or at least leagally they live in Prague or other big cities. Thus this is also another factor that does play quite an important role.



Here we can see that except the 3. and 5. position we see cities that are not really often in the worst 5 in all of the other measures. Thus we can see quite a strong ommited variable bias which is not captured in the independent varibles.



5.2 Note for the following indicators

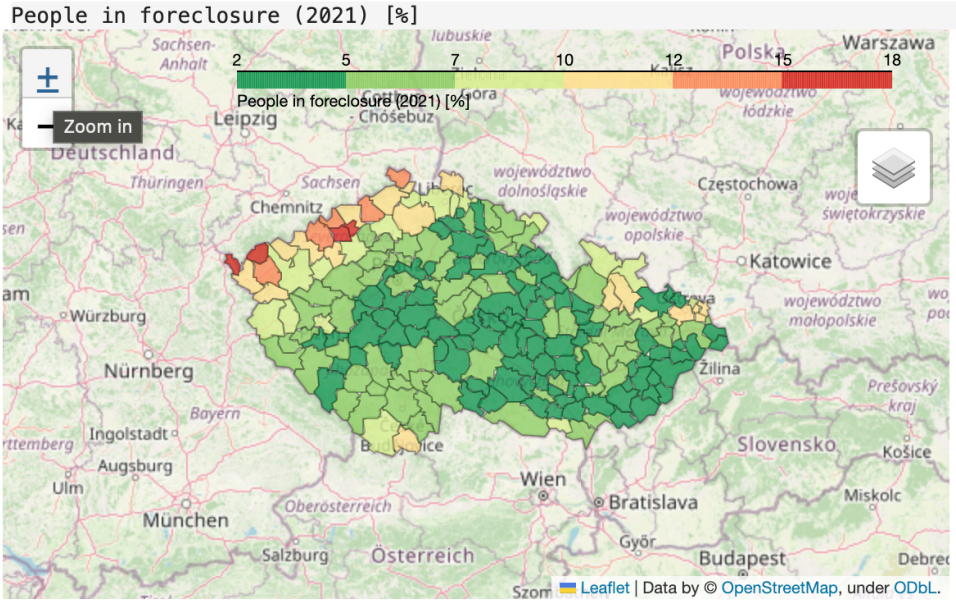
Secondly, we plotted already mentioned socio-economic indicators and their geographical structure across the whole Czech Republic. For each of them we again show the worst regions where the lawmakers should focus in the future.

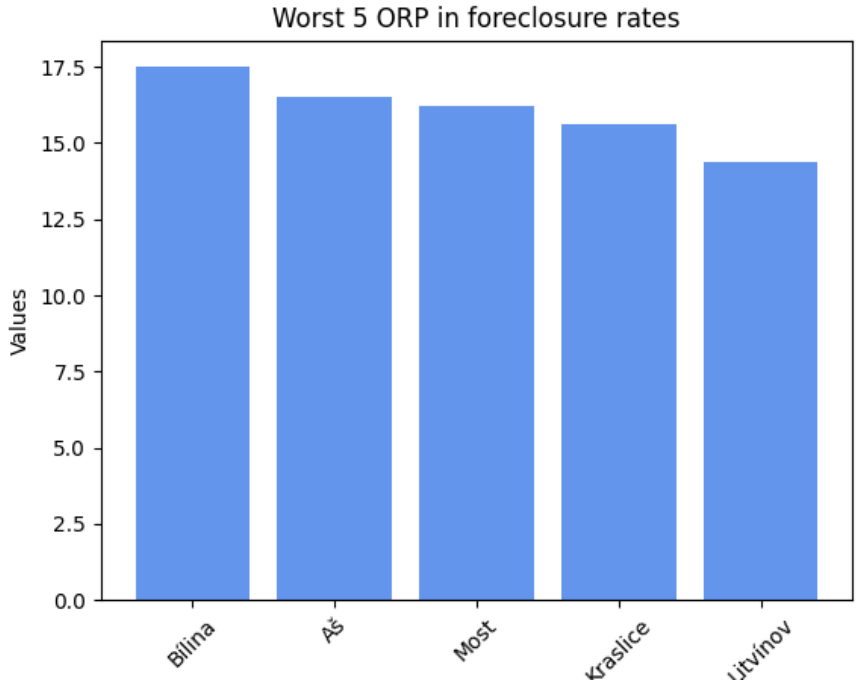
Our main goal is to uncover geographically more detailed picture of the problematic indicators. The reason for that is that in the public debate there is usually mentioned that the problematic regions are "Ústecký kraj","Moravskoslezký kraj"and "Karlovarský kraj"but our data shows that this is quite a strong simplification and there are vast differences between different ORPs. This might help to find a more economic and more targeted treatment of those socio-economical problematics.

It is also important to note that the results in each parameter should not be mistaken with the criminality itself. As only the criminality per capita are the real data of criminality. Our parameters are just realtively strongly correlated with the response variable and improving them will definitely help there is still a lot of ommited variable bias. For example, as we can see that a lot of the crime happens in cities that do have quite high standart of life just increasing the level of public lighting migh be more effective there which is not shown by our variables.

5.3 People in foreclose

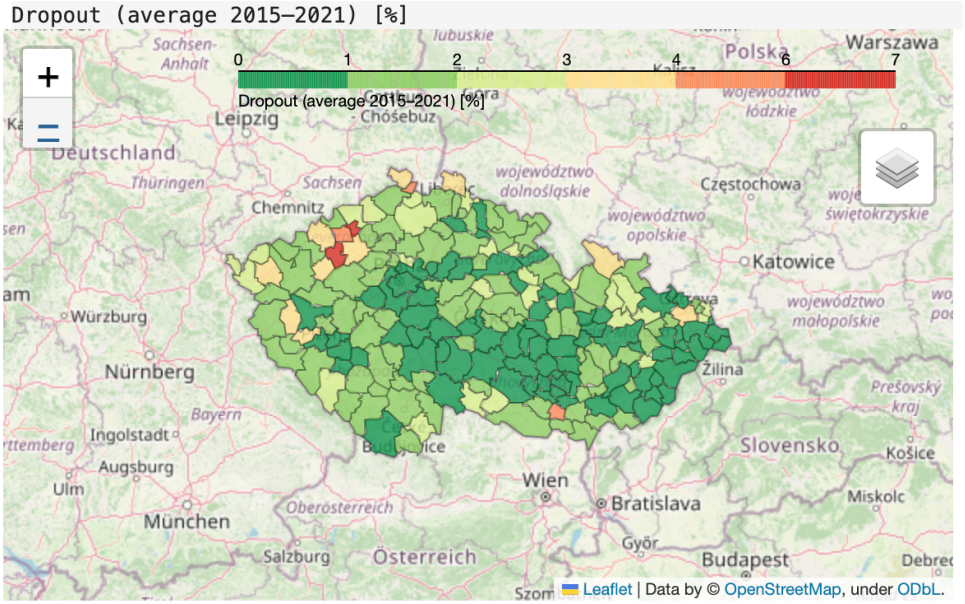
In the foreclosure we can relatively clearly observe quite a strong pattern where the problem seems to be much higher in the ORPs of the former "Sudety"region. This will be the case also for other variables but here it is quite an obvious one.

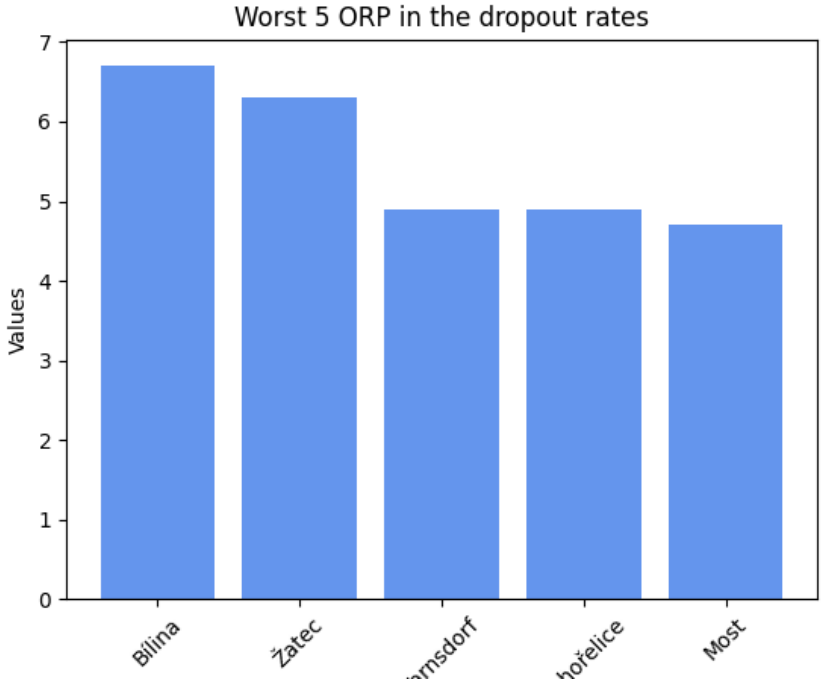




5.4 Dropout

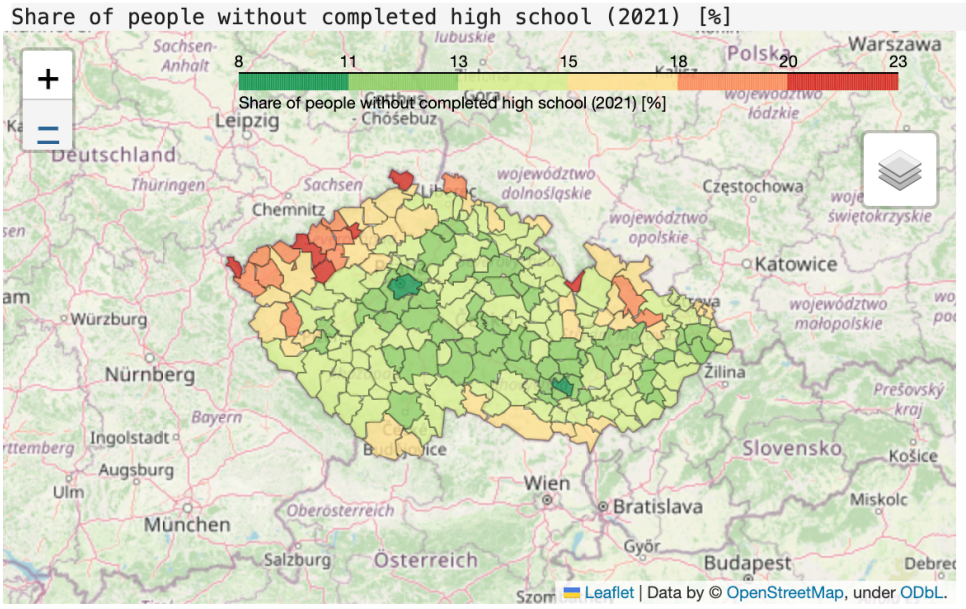
The dropout map is relatively suprising as we see some new regions and pattern seems to be relatively weak. Of course there is the well known center around Most although that itself is not included. But there are some suprising regions such as Varnsdorf and Pohořelice where especially for the latter it would be interesting to study the detailes behind these disturbing results.

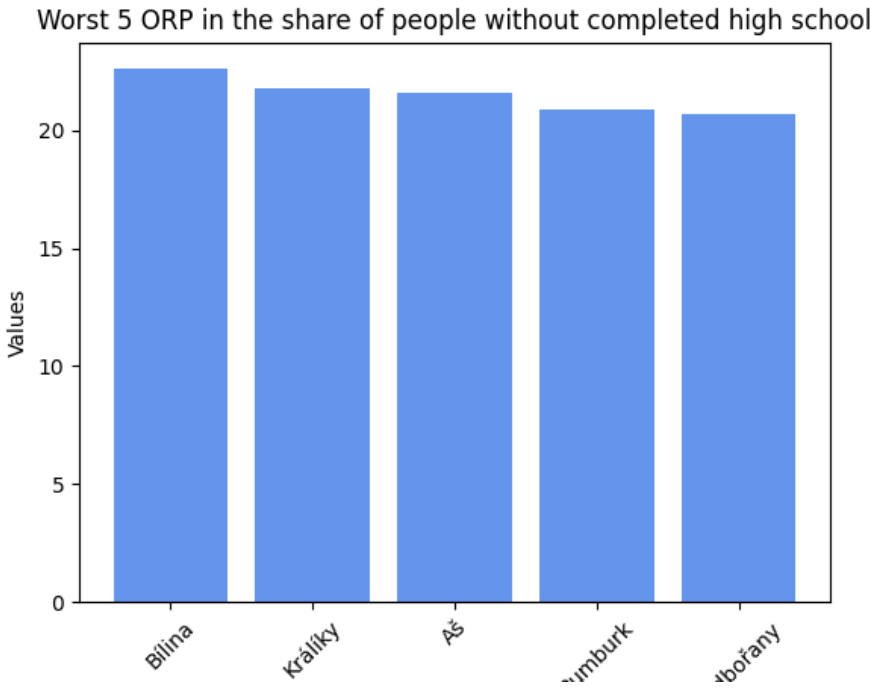




5.5 Share of people without completed high school

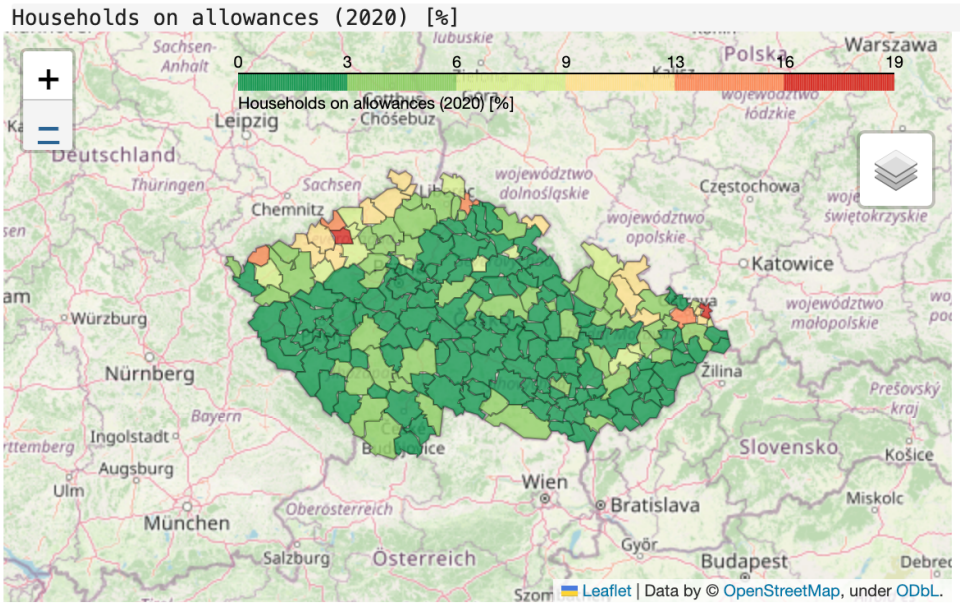
To our suprise there is significant difference between dropout and the ammount of people without high school which we really did not expect. Only the leading ORP "Bílina"is part of both bar plots but otherwise it seems to be relatively uncorrelated. In this graph we again need to point out the "Sudety"pattern which is striking. The outcome of this map has a realtively clear reasoning as people with higher education tend to move from the countryside to big cities for better work position whereas the people with poor education stay in these regions. This map relatively follows the population density map with a small exception in the north-west where the problems are most likely much deeper.

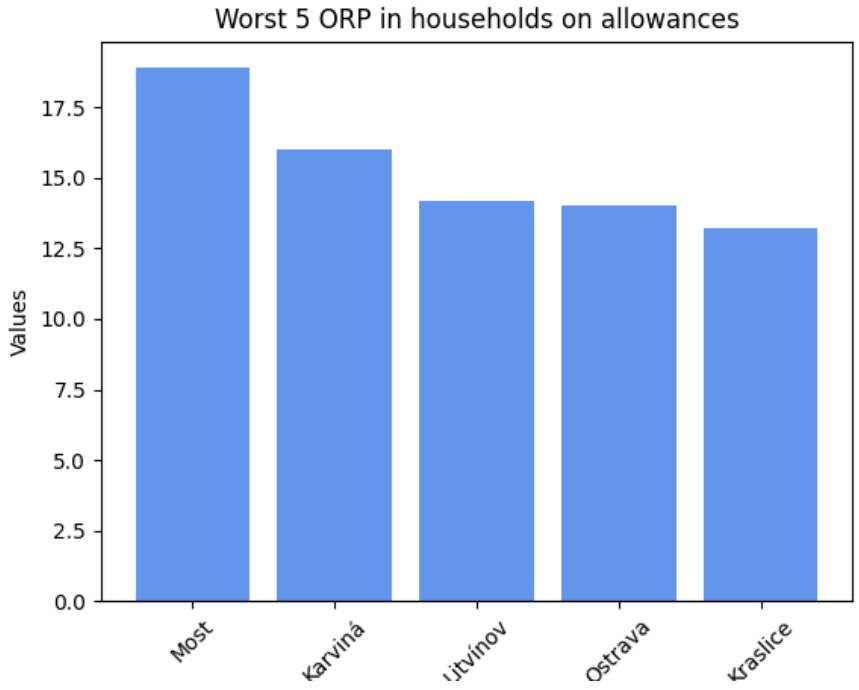




5.6 Households on allowances

This map is the least informative in our view as it almost perfectly follow the allowances map which makes quite a lot of sense. For us the most suprising member here are Kraslice but again they are relatively similar to other ORPs in their region where the problem is strong.





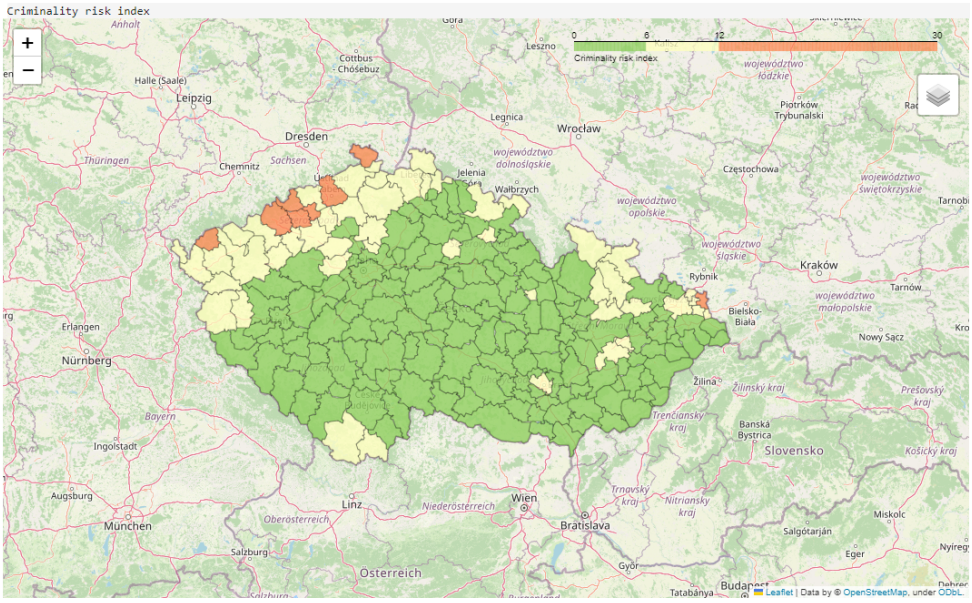
5.7 Criminality risk index

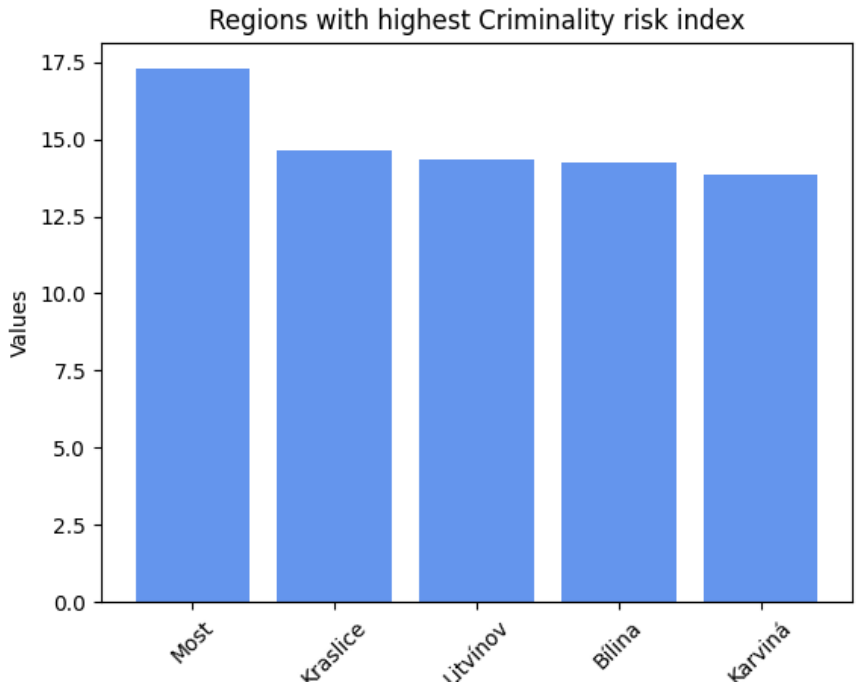
Our main goal was to define an index which will uncover 5 regions, where there are quite strong prerequisities for high criminality in the future in order to be able to focus on that problem in those regions. Our former idea was that the index will be weighted by the correlations of each variable but after a lot of tweaking we found out that it is not an optimal choice as the two least correlated variables tend to make the index less correlated. That is why we decided after a lot of trial and error that the index will be only weighted summ of 0.6*foreclosure + 0.4*households on allowances. This new measure has achieved a correlation of 0.64 which is not statically different from the correlation of foreclosure itself but it does have a stronger reasoning and is much more informative. With that we seperated the ORPs in 3 levels.

1. Low level of risk
2. Middle level of risk
3. High level of risk

With this we pinpointed 8 regions that are in a high-risk of economical criminal activities. Those are the regions where special treatment is required and should be analyzed in much more detailed level on their own.

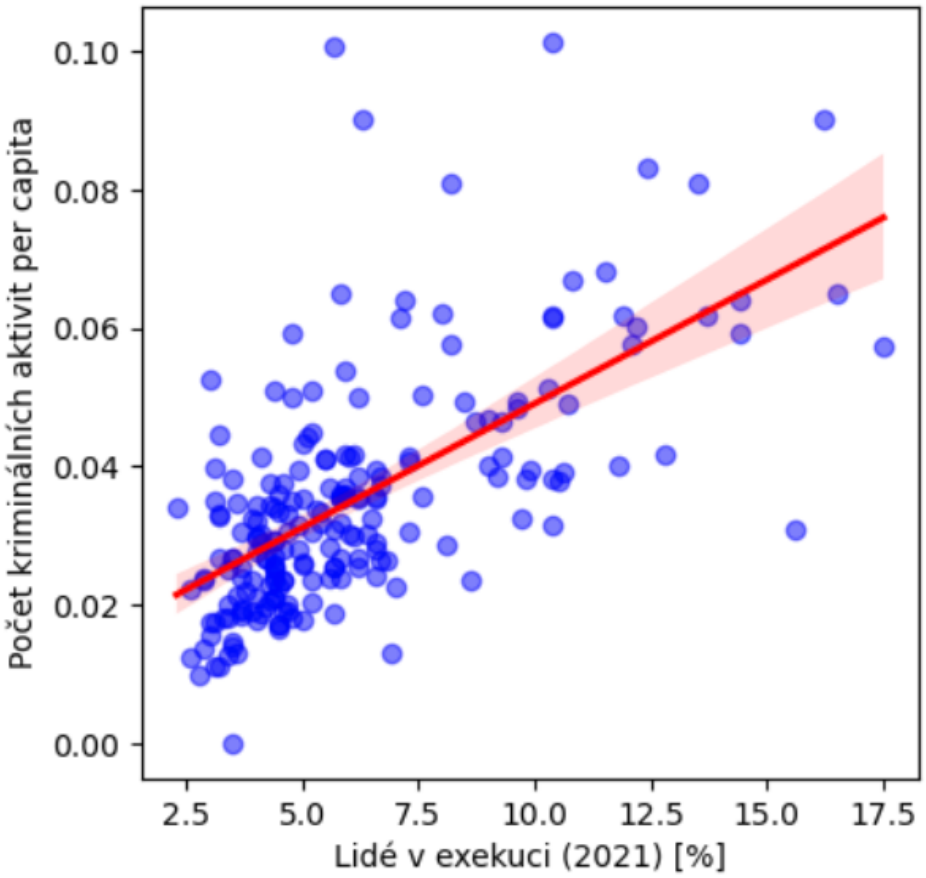
We do believe that our analysis could be encouraging for other researchers that could follow with more detailed analysis of the problematic regions. Underneath you can observe the exact correlations and our other visualizations that we mentioned in the text.

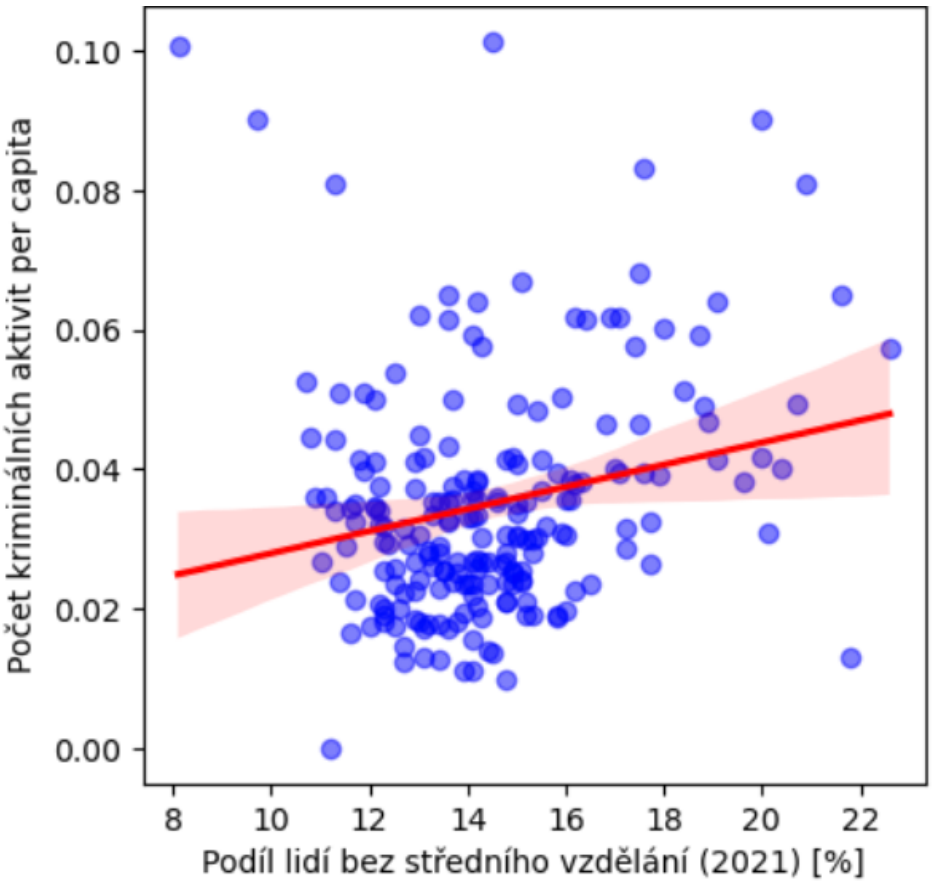


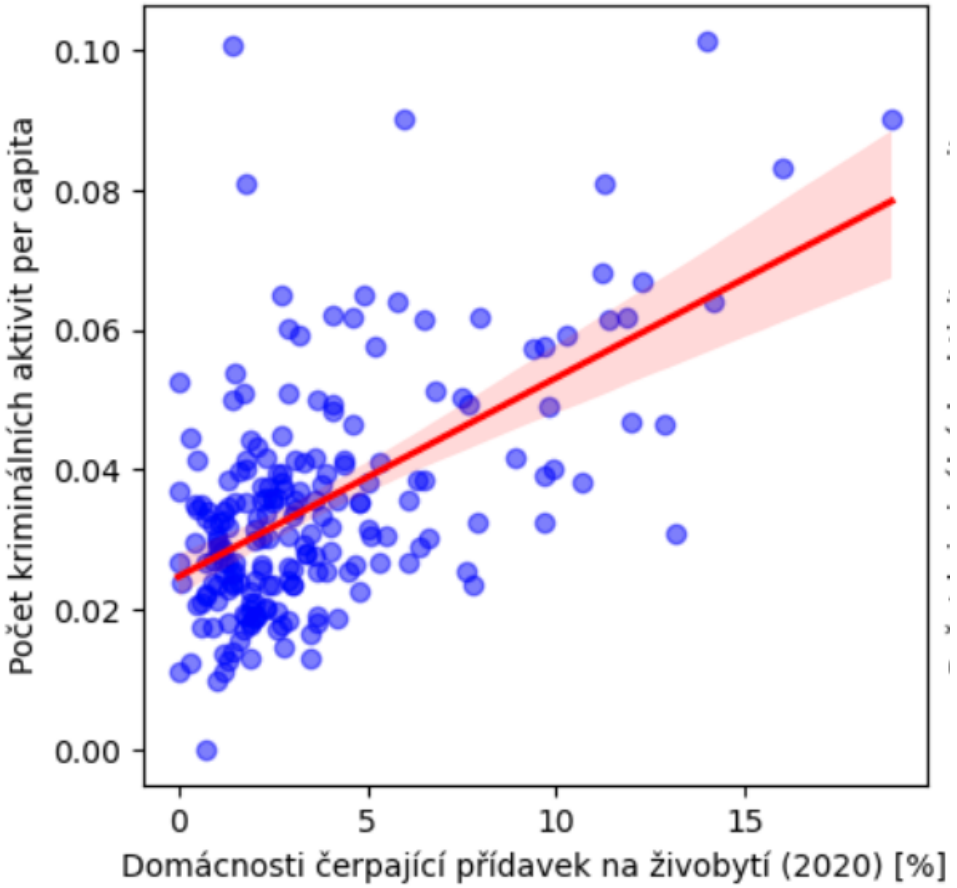


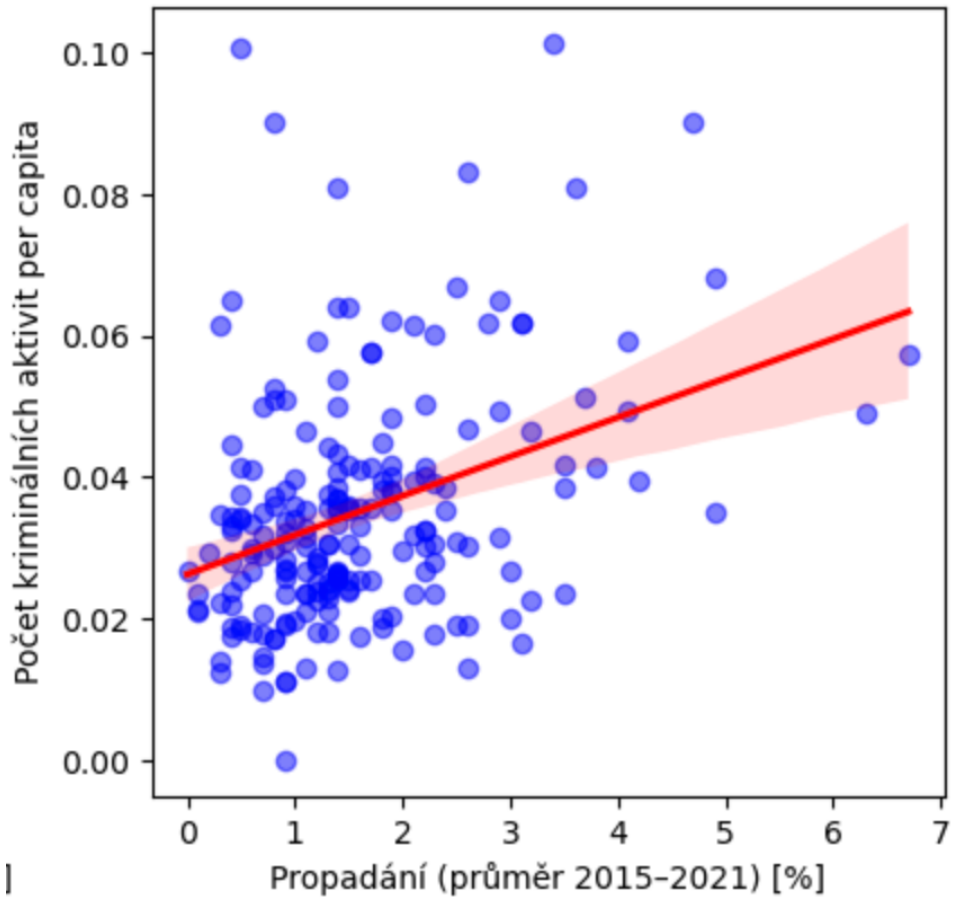
5.8 Correlation scatter plots

As a further analysis, we plot individual indicators on the correlation graph to really see how much these indicators correlate with the number of criminal activities per capita. You can see the graphs below:

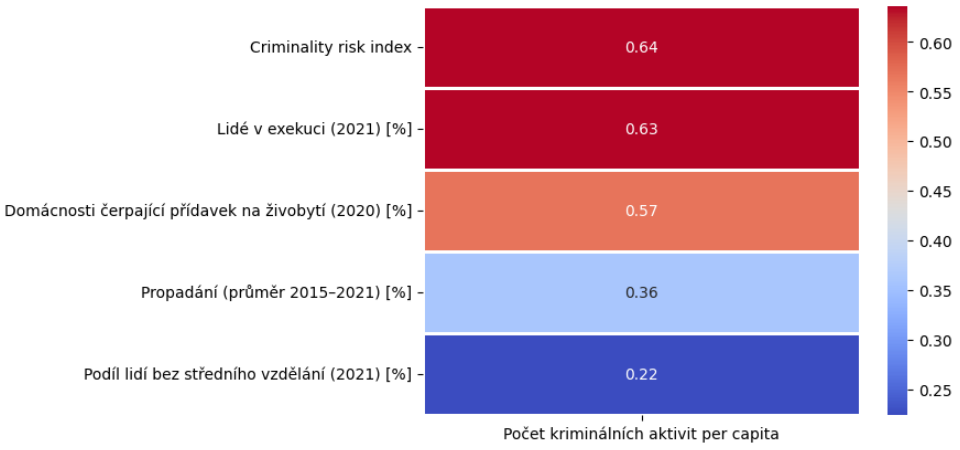



Počet kriminálních aktivit per capita
0.10
0.08
0.06
0.04
0.02
0.00
8
10
12
14
16
18
20
22
Podíl lidí bez středního vzdělání (2021) [%]


Počet kriminálních aktivit per capita
0.10
0.08
0.06
0.04
0.02
0.00
0
5
10
15
Domácnosti čerpající přídavek na živobytí (2020) [%]



As you can see, the most correlated with the number of criminal activities per capita is "The percentage of people with foreclosure"(0.63) and "The percentage of households receiving social benefits"(0.57) which are almost of the same correlation. On the other hand, the least correlated is "The percentage of people without completed high school education"(0.22) which does not influence the number of criminality in the regions as much - it also has the biggest deviation of the all mentioned indicators (which makes sense). Exact correlation numbers are visualized below on the correlation scale.



6 References

The data used in this project, that we acquired from PAQ research, were obtained from various sources including the Czech Statistical Office, the Agency for Social Inclusion, the Ministry of Labour and Social Affairs, the Chamber of Executors of the Czech Republic, and the Czech Household Panel Study.

The records of crime acts are exclusively from the Police of the Czech Republic which as the only one has the resources for it.

1. https://www.datapaq.cz/
2. PAQ data endpoints:
3. Domácnosti čerpající přídavek na živobytí (2020) po ORP - Agentura pro sociální začleňování, MPSV
4. Podíl lidí bez středního vzdělání - ČSÚ, SLDB 2021
5. Propadání (2015-2021)- ČŠI
6. Lidé v exekuci (2021)- Exekutorská komora ČR, ČSÚ, Czech Household Panel Study
7. https://kriminalita.policie.cz/
8. https://www.czso.cz/csu/xs/obyvatelstvo-xs Czech Statistical office - the data on the population in each ORP as the data was in quite a messy Excel file, we had to transform it manually and the new table is now to your disposal in our repository (app/počet_obyvatel_ORP.xlsx) and can be used in other projects with similar nature. It makes it easier to share the project with others.