

SAN - Final Assignment

Kriminalita v okresech ČR v závislosti na počtu cizinců

David Mikulík, Michaela Pejšová, Jakub Švec, Hoang Nam Tran

1. ledna, 2025

Obsah

Úvod.....	3
Výzkumná otázka.....	3
Zdroje a zpracování dat.....	3
Data o kriminalitě.....	3
Data o cizincích.....	4
Geografická data.....	5
Počet obyvatel.....	5
Integrace dat.....	6
Vizualizace.....	6
Datová analýza.....	11
Explorační analýza dat.....	11
Data o kriminalitě.....	11
Data o cizincích.....	14
Socioekonomické ukazatele.....	17
Korelační analýza.....	18
Regresní a GLM modely.....	21
Podíl cizinců.....	21
Kombinace podílu cizinců a nezaměstnanosti.....	21
Sociální podpora a podíl cizinců.....	22
Generalizované lineární modely (GLM).....	22
Ověření hypotéz.....	23
Diskuze a Závěr.....	23
Výkaz práce.....	24
Odvedená práce.....	24

Úvod

Kriminalita je dlouhodobě sledovaným společenským jevem, který má zásadní dopad na kvalitu života obyvatel a stabilitu regionů. Současné diskuse o vztahu mezi přítomností cizinců a kriminalitou vzbuzují pozornost nejen mezi odborníky, ale také ve veřejném prostoru. Tento projekt se zaměřuje na analýzu tohoto vztahu na úrovni okresů v České republice. Cílem je zjistit, zda existuje vztah mezi počtem cizinců a kriminalitou, a analyzovat, jak socioekonomické faktory mohou tento vztah ovlivňovat.

K analýze jsme využili dostupná data o kriminalitě, demografii a geografii jednotlivých okresů. Použité metody zahrnují průzkumnou analýzu dat (EDA), regresní modelování a statistické testy. Tento report prezentuje klíčová zjištění a diskutuje jejich dopady.

Výzkumná otázka

Hlavní výzkumná otázka: Jak ovlivňuje přítomnost cizinců míru kriminality na úrovni okresů v České republice?

Dílčí otázky:

1. Existuje korelace mezi podílem cizinců v populaci a celkovou mírou kriminality?
2. Jaký vliv na kriminalitu mají faktory, jako je národnost a pohlaví cizinců?
3. Jsou různé typy trestné činnosti (např. násilné, majetkové, drogové) ovlivněny odlišně v závislosti na podílu cizinců v populaci

Zdroje a zpracování dat

Data o kriminalitě

Zdroj: Mapa kriminality – Policie ČR

Data o kriminalitě pochází z projektu Mapa kriminality, který provozuje Policie ČR. Používáme data z roku 2023, která byla stažena a zpracována po jednotlivých měsících. Dataset obsahuje podrobné záznamy o všech incidentech, které Policie ČR přijala a řešila. Každý trestný čin nebo přestupek je uveden jako samostatný záznam s následujícími atributy:

- **id** – Unikátní identifikátor případu. Každý případ může být uveden na více řádcích, protože může mít více typů.
- **x, y** – Souřadnice místa incidentu.
- **mp** – Nejsme si jisti, tento sloupec nepoužíváme. Vzhledem k booleanovské hodnotě a názvu by se mohlo jednat o indikátor, zda se případ nějakým způsobem dotýkal městské policie.
- **date** – Datum a čas události.
- **state** – Stav vyřízení (objasněný, neobjasněný, skutek se nestal, skutek není trestným činem).
- **relevance** – Určuje, jaký vztah má skutek k danému místu (místo spáchání, nahlášení, místo následku, zahraniční lokalita, chybná lokalita, neurčeno).
- **types** – Typ trestné činnosti (např. drogová činnost, autonehoda, násilná trestná činnost, ...).

Z datasetu byly odstraněny záznamy o případech, které byly označeny jako "skutek se nestal" nebo "skutek není trestným činem", jelikož se nejedná o reálnou kriminalitu. Stejně tak byly vyřazeny záznamy s relevancí "Zahraniční lokalita" a "Chybná lokalita" z důvodu nemožnosti přesného zařazení do okresů.

Pro zajištění přehledu o kriminalitě na úrovni okresů jsme trestné činy seskupili podle souřadnic do příslušných okresů a spočítali počet případů pro každou hlavní kategorii kriminality. Kromě toho jsme stanovili celkový počet unikátních trestných činů, přičemž tento počet není součtem jednotlivých kategorií, protože některé případy mohou spadat do více kategorií zároveň.

Sloupce ve finálním datasetu (sloupce obsahují počty daných skutků):

- **unique_crimes** (unikátní trestné činy)
- **violent** (násilná trestná činnost)
- **fire_explosion** (požáry, výbuchy)
- **burglary** (vloupání)
- **theft** (krádež)

- fraud (podvod)
 - other_property_related (ostatní majetková trestná činnost)
 - general_endangerment (obecné ohrožení)
 - traffic_accident (dopravní nehoda)
 - drugs (drogová trestná činnost)
 - weapons (zbraně)
 - extremism (extremismus)
 - offense (ostatní přestupky)
-

Data o cizincích

Zdroj: ČSÚ – Výroční zpráva Ministerstva vnitra

Data o cizincích pocházejí z výroční zprávy Ministerstva vnitra za rok 2023, publikované ČSÚ. Dataset obsahuje přehled o počtu cizinců přihlášených k pobytu v jednotlivých okresech ČR. Kromě celkového počtu cizinců zahrnuje data o rozložení podle pohlaví, původu (EU nebo mimo EU) a u vybraných zemí s vysokým počtem cizinců i podrobnější rozpis.

Sloupce ve finálním datasetu:

- name (název okresu)
 - foreigners_total (celkový počet cizinců)
 - men (muži)
 - women (ženy)
 - EU_total (celkový počet cizinců pocházejících ze zemí EU)
 - Slovakia (Slováci)
 - Germany (Němci)
 - Poland (Poláci)
 - outside_EU_total (celkový počet cizinců pocházejících ze zemí mimo EU)
 - Ukraine (Ukrajinci)
 - Vietnam (Vietnamci)
 - Russia (Rusové)
-

Geografická data

Zdroj: OpenStreetMap a RÚIAN

Geodata, tedy hranice jednotlivých okresů, byla stažena z OpenStreetMap a Registru územní identifikace, adres a nemovitostí (RÚIAN). Tato data byla využita pro geografické zařazení trestných činů do okresů, vypočtení plochy okresů a tvorbu vizualizací kriminality na mapě.

Sloupce ve finálním datasetu:

- name (název okresu)
 - id (id okresu)
 - geometry (vymezení okresu)
 - area_km2 (rozloha okresu)
-

Počet obyvatel

Zdroj: ČSÚ

Abychom mohli spočítat hustotu zalidnění přepočítat kriminalitu na obyvatele a normalizovat tak data, obstarali jsme ještě údaje o počtech obyvatel v okresech.

Související sloupce ve finálním datasetu:

- name (název okresu)
 - inhabitants (počet obyvatel)
 - population_density (hustota zalidnění)
-

Pro doplnění lineárních modelů o potenciální confounder variables jsme přidali ještě různé další faktory, které podle nás mohou ovlivňovat kriminalitu. Jedná se o data týkající se exekucí, sociálních dávek, zaměstnanosti, a další, abychom mohli pozorovat, zda prediktory týkající se cizinců v jejich konkurenci stojí jako relevantní.

Exekuce

Zdroj: Mapa zadlužení, Exekutorská komora

Související sloupce ve finálním datasetu:

- distraint (procento obyvatel v exekuci)
-

Nezaměstnanost

Zdroj: MPSV

Související sloupce ve finálním datasetu:

- unemployment_rate (procento nezaměstnaných)
 - employable_unemployed (počet nezaměstnaných, kteří mohou okamžitě vstoupit do pracovního poměru)
-

Sociální dávky

Zdroj: ČSÚ

Zde se jedná o počet vyplacených dávek.

Související sloupce ve finálním datasetu:

- total_number_of_social_support_allowance (celkový počet dávek)
 - child_allowance (přídavek na dítě)
 - housing_allowance (příspěvek na bydlení)
 - parental_allowance (rodičovský příspěvek)
 - birth_grant (porodné)
 - funeral_grant (pohřebné)
 - foster_care_allowance (dávky péstounské péče)
-

Integrace dat

Všechna data jsme na základě jména nebo id okresu spojili do jedné tabulky obsahující všechny zmíněné sloupce a řádek pro každý z okresů v ČR, celkem tedy 77 řádků.

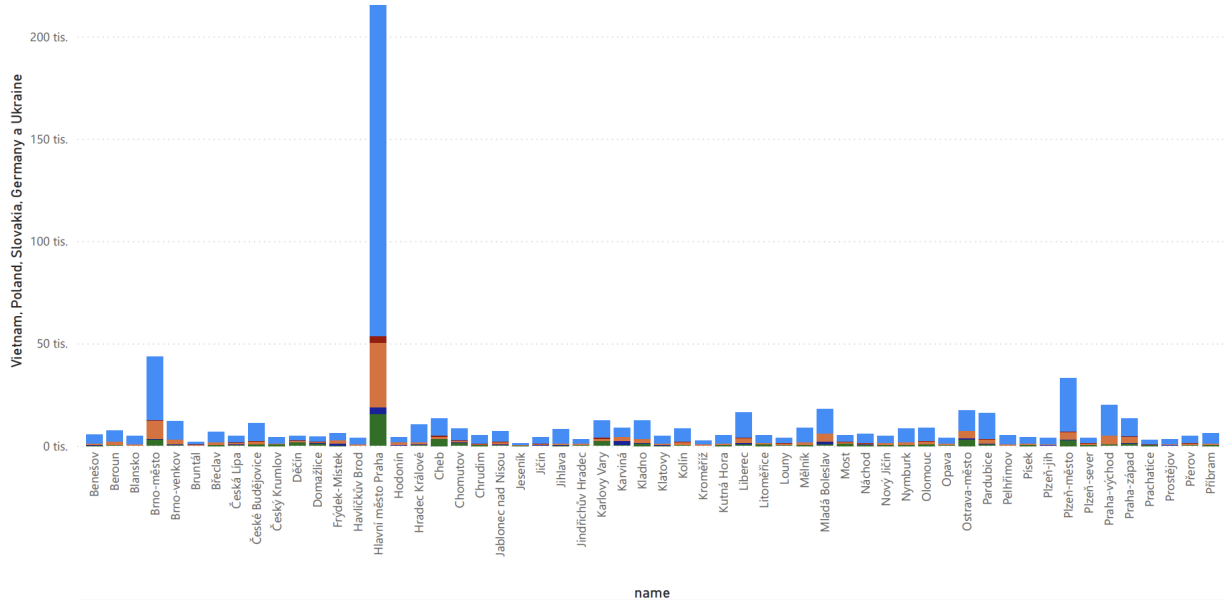
Vytvořili jsme ještě jednu variantu této tabulky, a to sice tabulku obsahující místo absolutních hodnot relativní - přepočítané na obyvatele. Výhodou je, že takto vynormovaná data jsou porovnatelná mezi okresy. Tato data považujeme za výsledek preprocessingu a výchozí bod další analýzy.

Vizualizace

Počet a poměr Cizinců v jednotlivých městech:

Vietnam, Poland, Slovakia, Germany a Ukraine podle kategorie name

● Vietnam ● Poland ● Slovakia ● Germany ● Ukraine

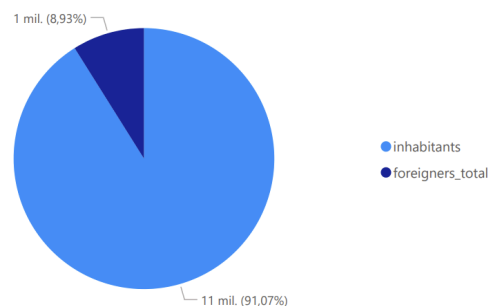
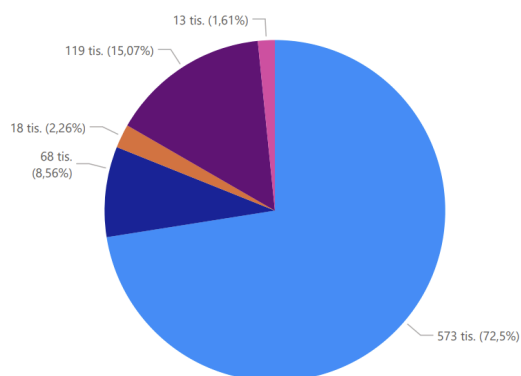


Z grafu je patrné, že nejvíce cizinců na území ČR je lidí z Ukrajiny, tento fakt je důsledkem války na Ukrajině. Druhou nejčetnější skupinou jsou lidé ze Slovenska. Opět je největší četnost spojena s velikostí města.

Složení cizinců v ČR podle národnosti a jejich poměr vzhledem k počtu občanů ČR:

Ukraine, Vietnam, Poland, Slovakia a Germany

inhabitants a foreigners_total

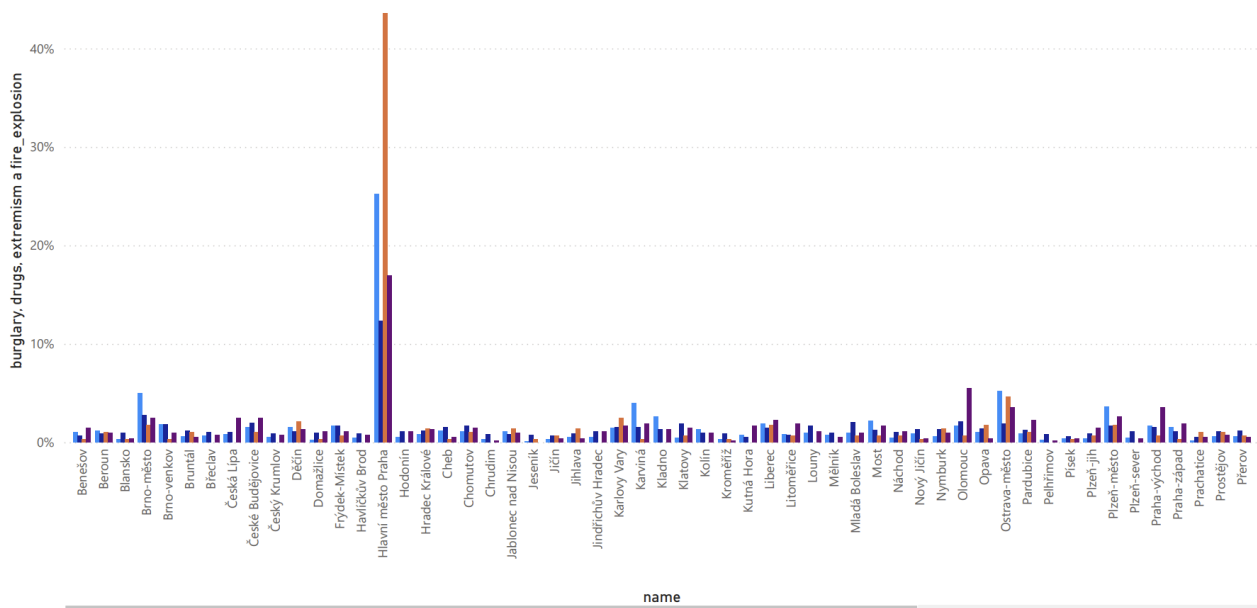


Počet a poměr Kriminálních činů v jednotlivých městech:

Pro lepší přehlednost jsou tyto grafy rozděleny do 3 grafů.

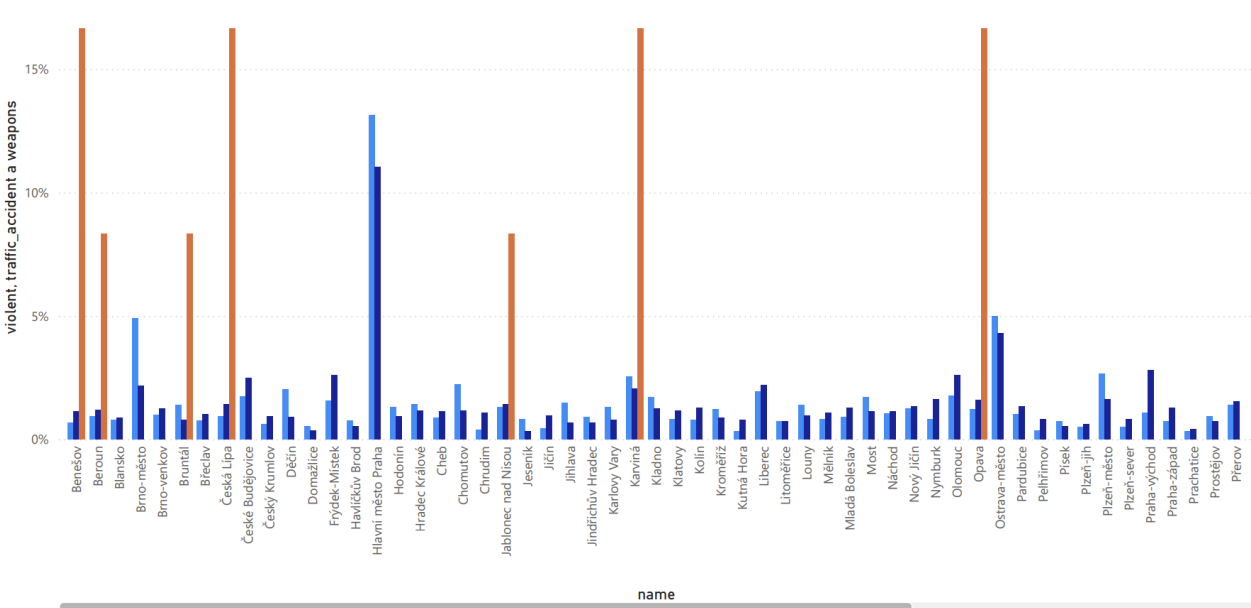
burglary, drugs, extremism a fire_explosion podle kategorie name

burglary drugs extremism fire_explosion

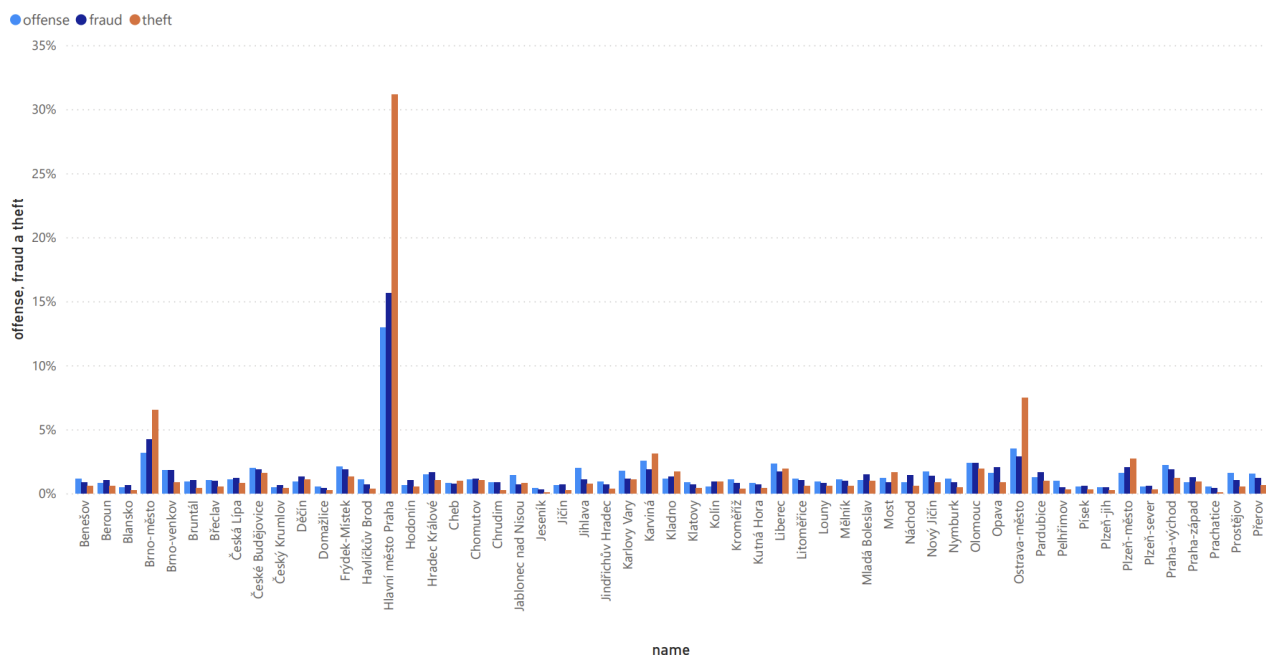


violent, traffic_accident a weapons podle kategorie name

violent traffic_accident weapons



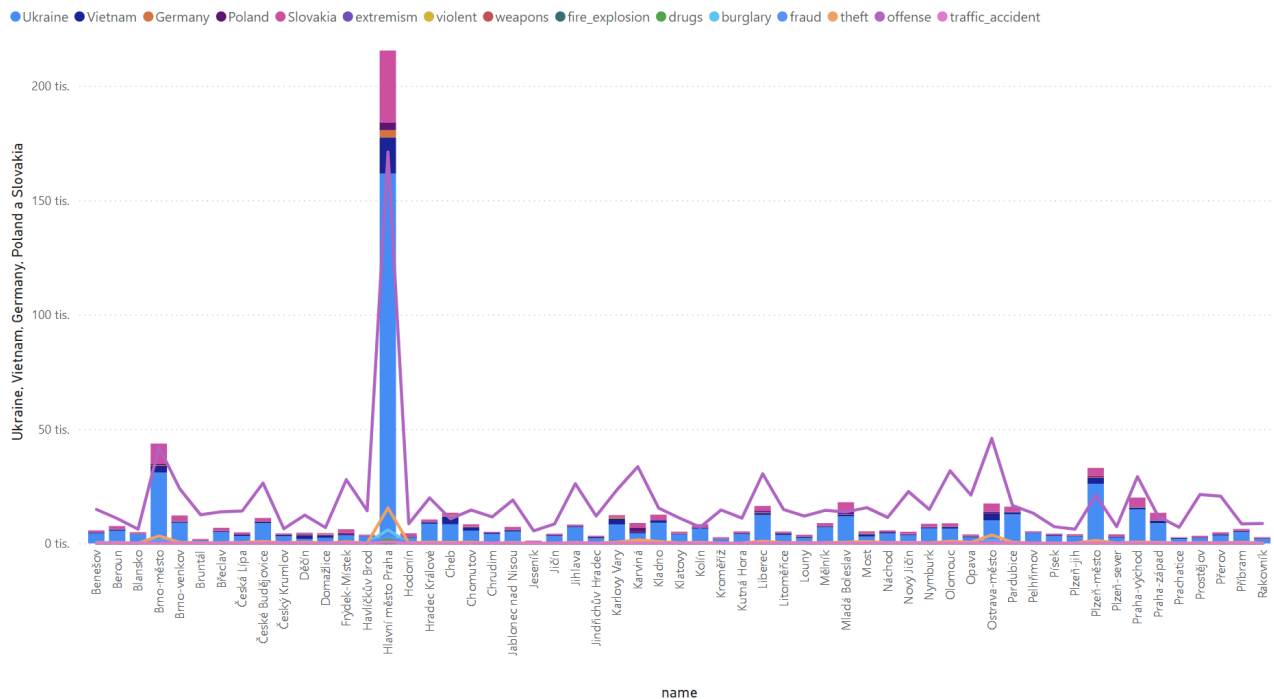
offense, fraud a theft podle kategorie name



Z grafů je patrné, že nejvíce extrémní kriminální činnosti se odehrávají v hlavním městě. Kriminální činnosti spojené s nelegálním držení zbraní, “Weapons”, se narozdíl od ostatních činností nedějí ve všech okresech, nebo se dějí pouze ve velmi malém množství. V některých okresech je ale jejich četnost poměrně výrazná oproti jiným.

Počet a poměr kriminálních činů a cizinců v jednotlivých městech:

Ukraine, Vietnam, Germany, Poland, Slovakia, extremism, violent, weapons, fire_explosion, drugs, burglary, fraud, theft, offense a traffic_accident podle kategorie name

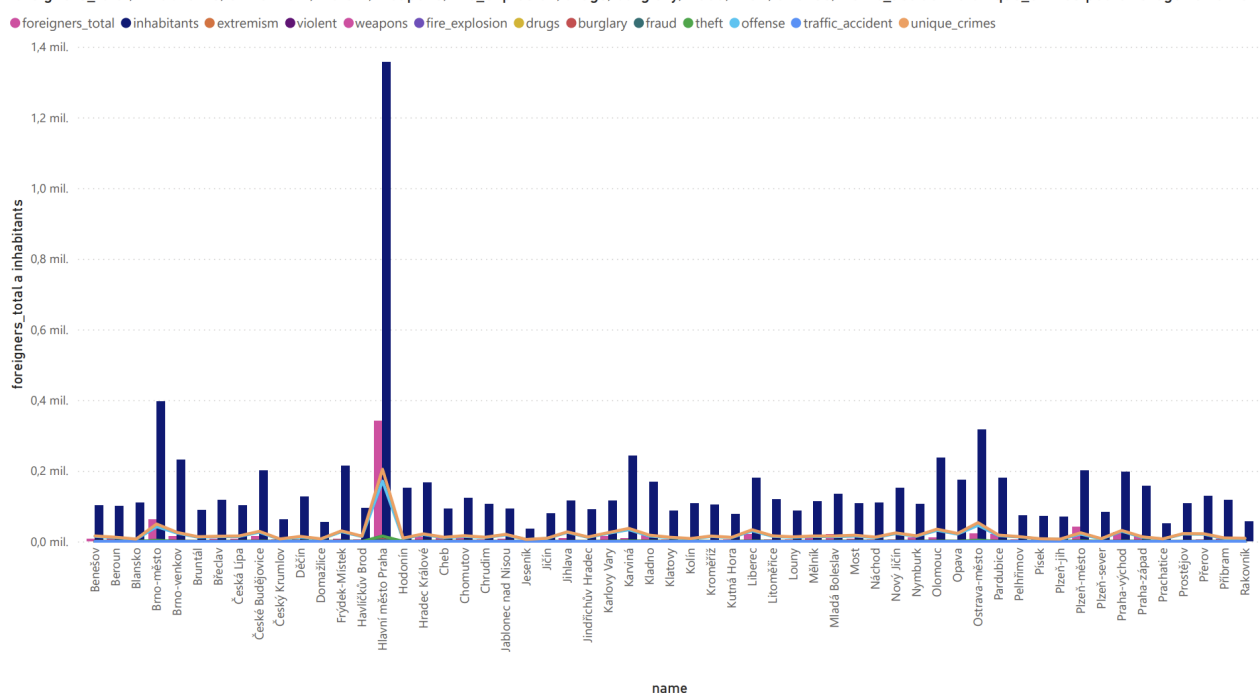


Snaha o zjištění zda existuje nějaká vizuální korelace mezi počtem cizinců a kriminální činností. Dle grafu to nelze přímo potvrdit.

Některé oblasti, mají větší kriminalitu z důvodu jejich geografické polohy, např. Karviná, která má vyšší kriminalitu v tomto může být způsobeno svým umístěním a dalšími faktory spojené s historií této oblasti. Podobné souvislosti mohou být v Přerově, Jihlavě atd.

Počet a poměr kriminálních činů vzhledem k cizincům a občanům v jednotlivých městech:

foreigners_total, inhabitants, extremism, violent, weapons, fire_explosion, drugs, burglary, fraud, theft, offense, traffic_accident a unique_crimes podle kategorie name

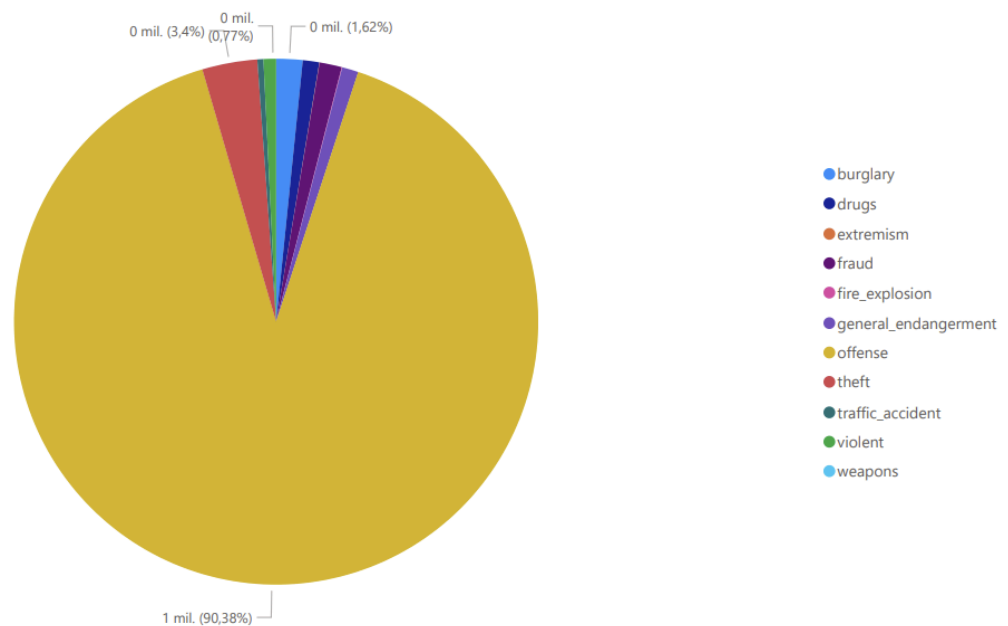


Snaha o zjištění zda existuje nějaká vizuální korelace mezi součtem cizinců nebo počtem obyvatel a kriminální činnosti. Dle základního grafu opět nelze potvrdit, vzhledem k tomu, že žádná vizuální korelace zde není přímo viditelná.

Z grafu je znatelné, že ve velkých městech žije nejen více českých obyvatel, ale také více migrantů.

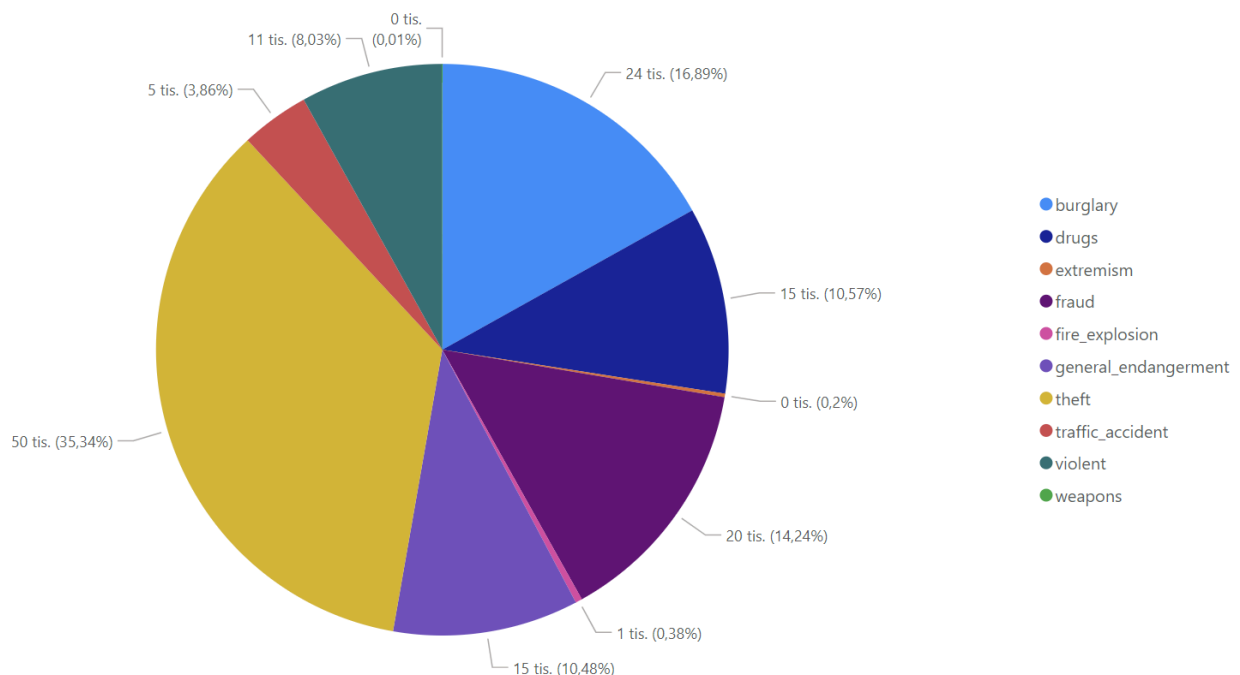
Poměr jednotlivých trestných činů:

burglary, drugs, extremism, fraud, fire_explosion, general_endangerment, offense, theft, traffic_accident, violent a weapons



Nejvíce četná kriminální činnost je za přestupky.

V případě ostatních trestných činů je poměr následující:



Datová analýza

Explorační analýza dat

Abychom lépe porozuměli datům, se kterými pracujeme, provedli jsme průzkumnou datovou analýzu. Protože ale pracujeme s velkým množstvím prediktorů, je analýza pouze stručná. Pro přehlednější vizualizaci v mapě používáme data přepočtená na počet obyvatel.

Data o kriminalitě

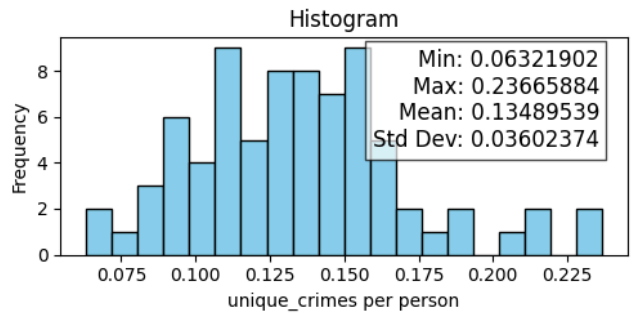
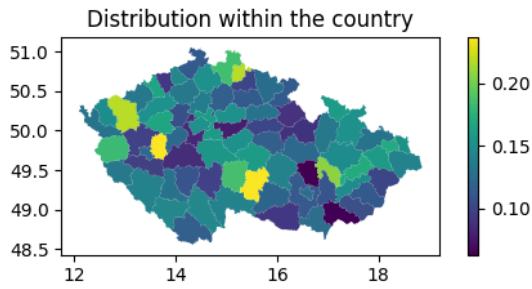
Data o počtu libovolných trestných činů jsou v rámci okresů rozložena přibližně normálně, není v nich viditelná závislost na geografické lokalitě. Podobně rozložená jsou ještě data o přestupcích, podvodech a o dopravních nehodách.

Další skupinou jsou činy typu násilí, obecné ohrožení, vloupání, krádež, a další majetková činnost. Hustěji se vyskytují v konkrétních regionech, hlavně na severozápadě Čech a severu Moravy - v rámci ČR se jedná o chudší oblasti. Nebo se jedná o větší města. Nejsou rozděleny zcela normálně, křivka je vychýlená.

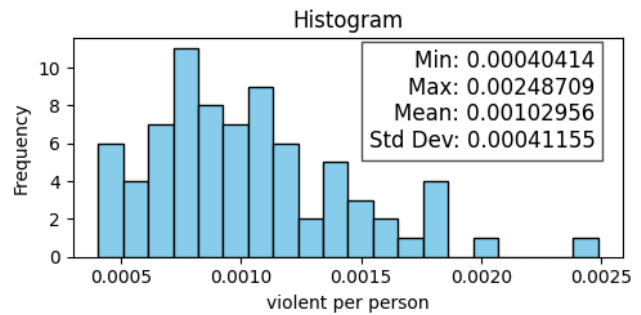
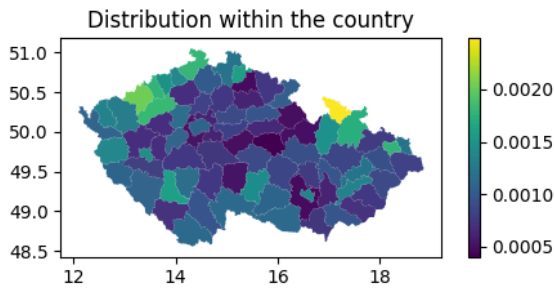
Drogová kriminalita je častější v příhraničních oblastech.

U méně častých trestných činech, jako je žhářství či extrémismus, není možné vypořádat žádný vzor, protože se jedná o ojedinělé události. Je možné, že pokud bychom pracovali s daty sesbíranými za více let, že by se v datech nějaký vzor objevil.

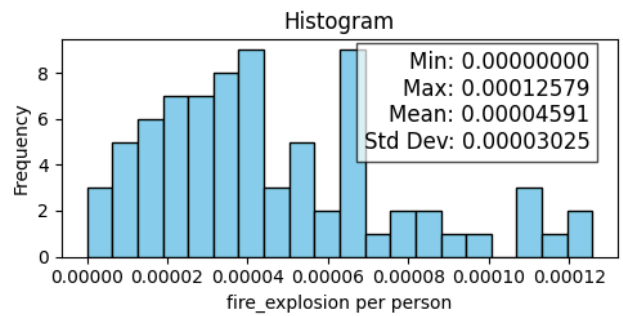
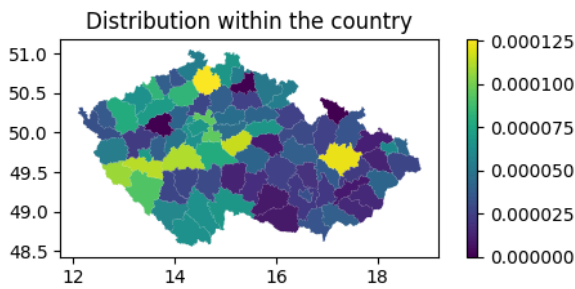
Unique_crimes per person



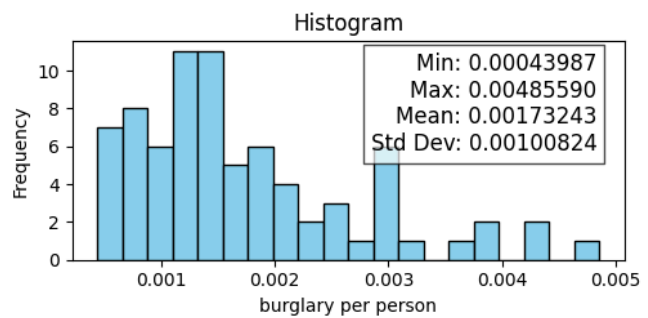
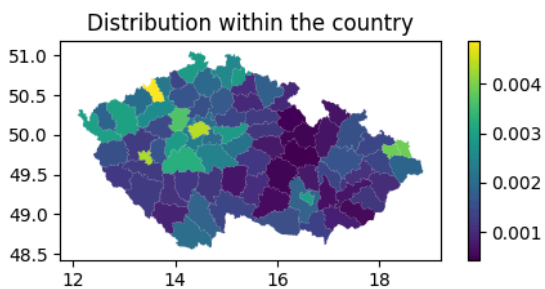
Violent per person



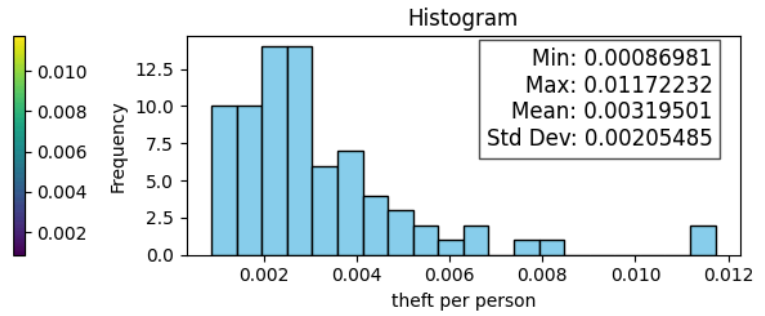
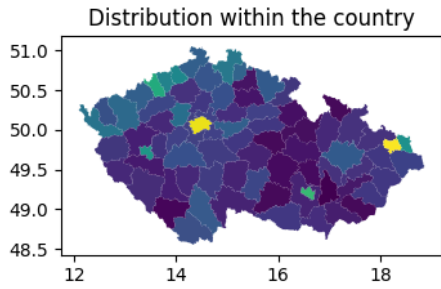
Fire_explosion per person



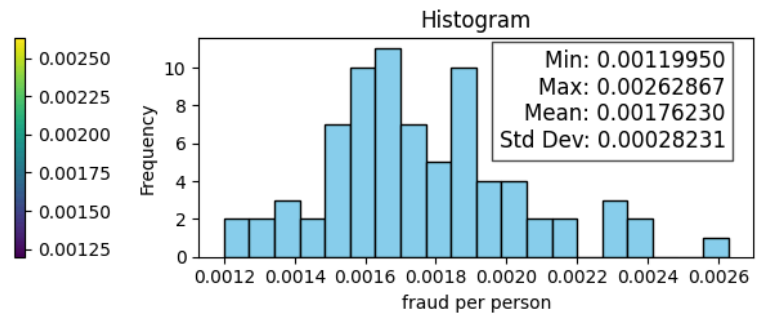
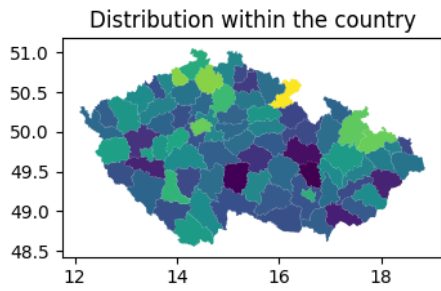
Burglary per person



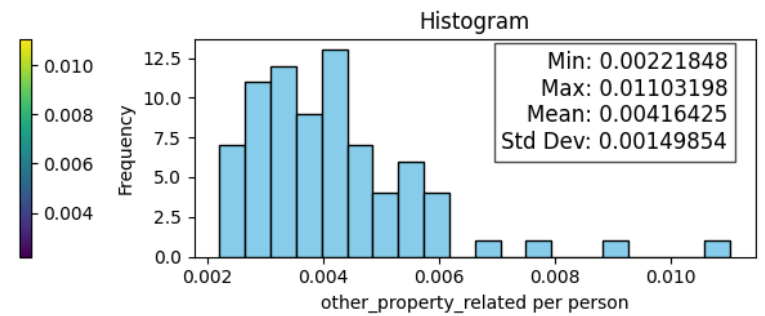
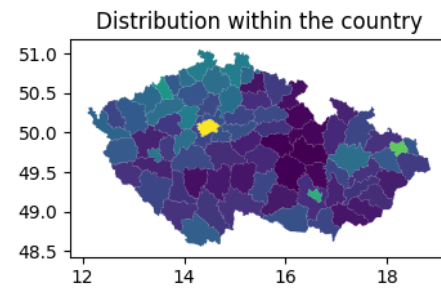
Theft per person



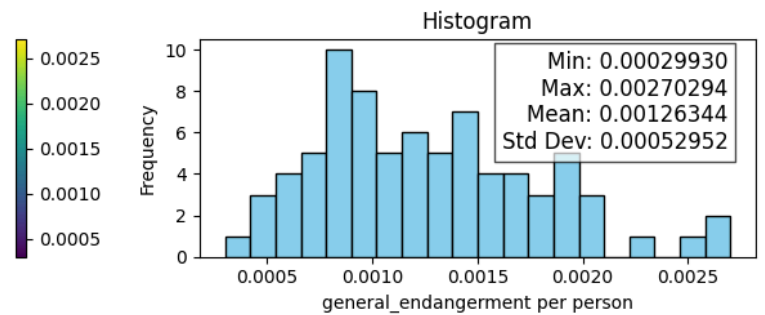
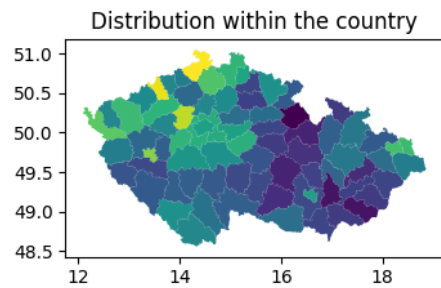
Fraud per person



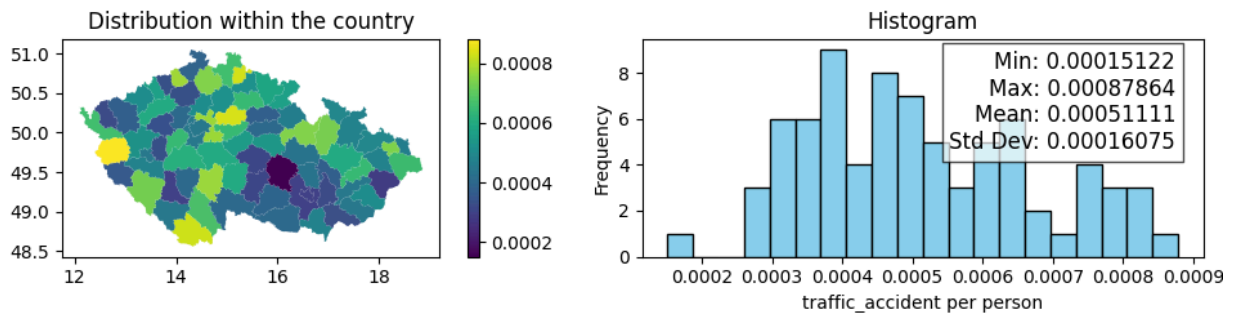
Other_property_related per person



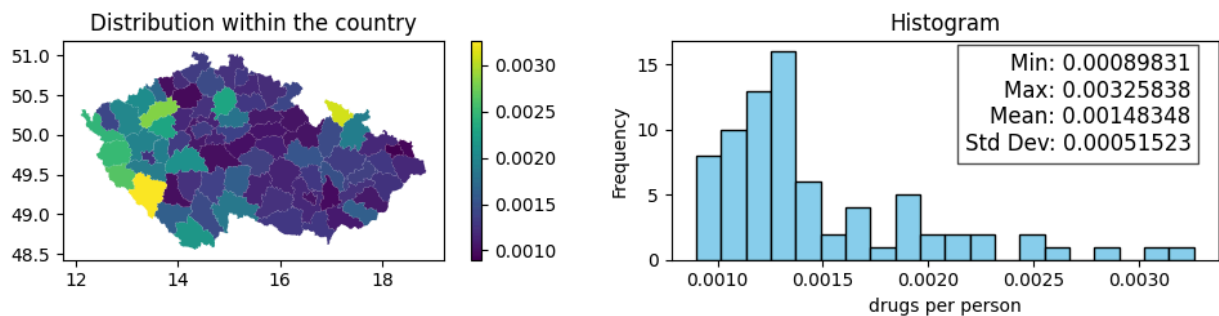
General_endangerment per person



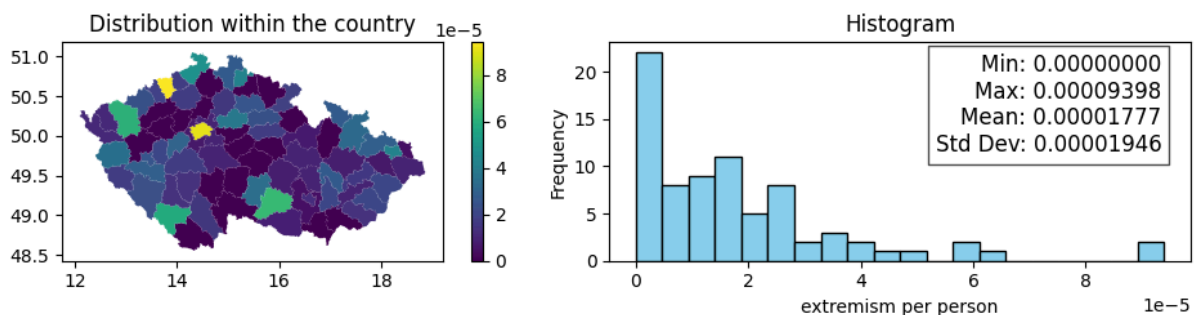
Traffic_accident per person



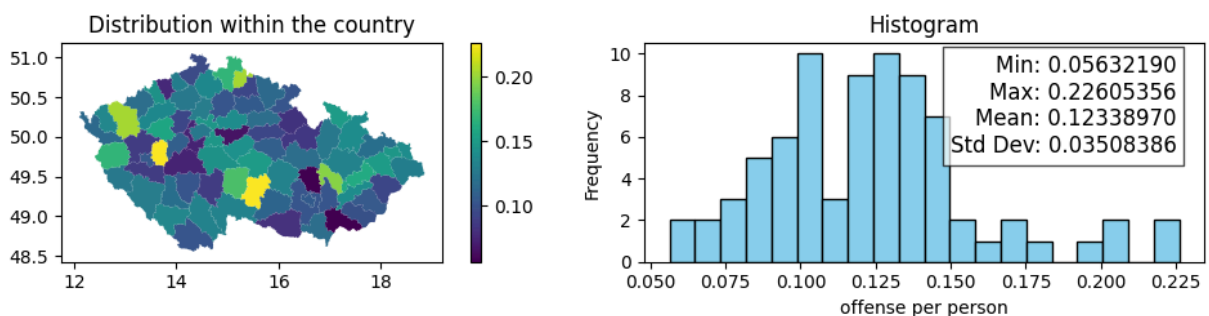
Drugs per person



Extremism per person



Offense per person



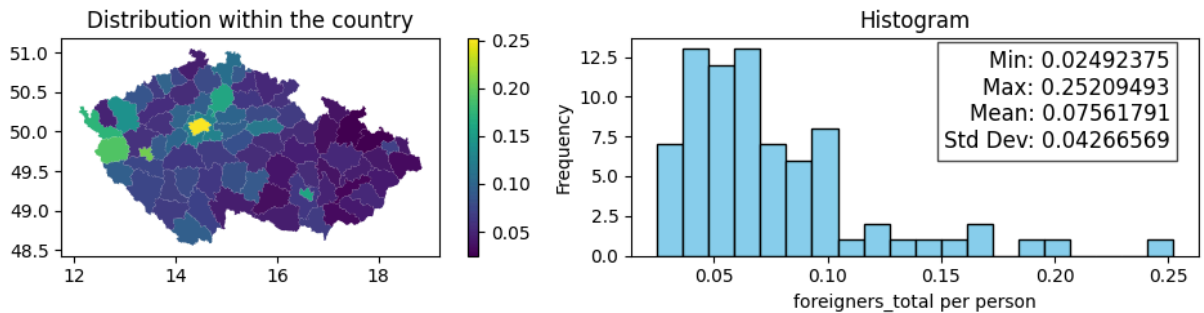
Data o cizincích

Je vidět, že celkový podíl cizinců je geograficky rozložen stejně jako celkový podíl cizinců-mužů a cizinek. Pro účely pozdější regrese by bylo užitečné upravit tyto proměnné tak, abychom z nich vyčetli více informací. Rozložení každopádně normální není, je hodně vychýlené. Geograficky souvisí s velkými městy nebo příhraničními oblastmi.

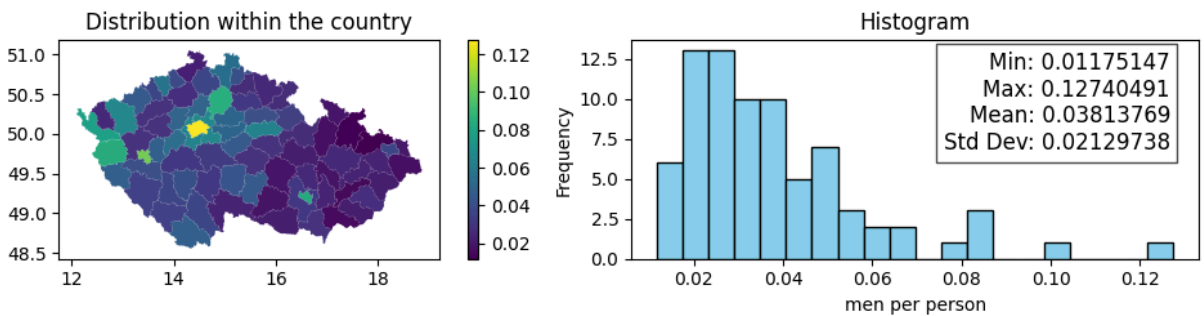
Zajímavé je geografické rozložení cizinců jednotlivých národností - zatímco Poláci a Němci žijí často v příhraničních oblastech odpovídajících států, u Slováků toto rozložení neplatí -

protože se do ČR stěhují často za lepším vzděláním či pracovním uplatněním, přitahují je větší města. Rusové jsou silně soustředěni v Praze, a v okolí Karlových Varů. Občané Vietnamu žijí často v příhraničních oblastech na západě ČR. Ukrajinců je v ČR mnoho a také se soustředí spíše na větší města, ale ve větším množství žijí ve více oblastech a jejich rozložení je od normálního vychýleno poměrně málo.

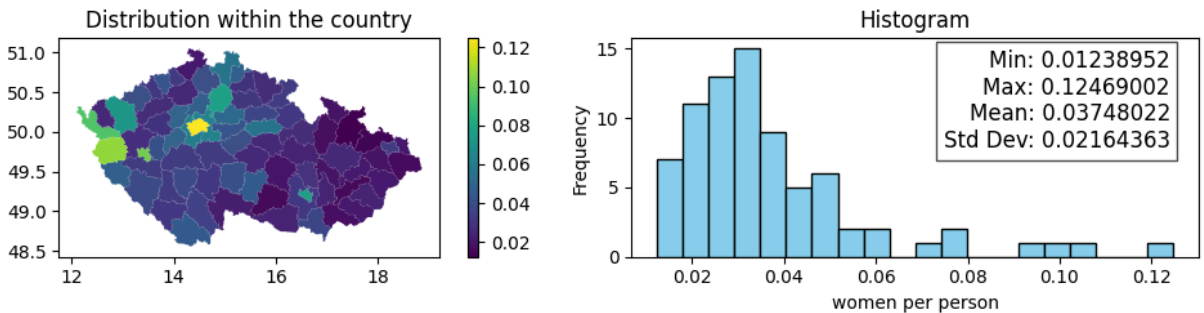
Foreigners_total per person



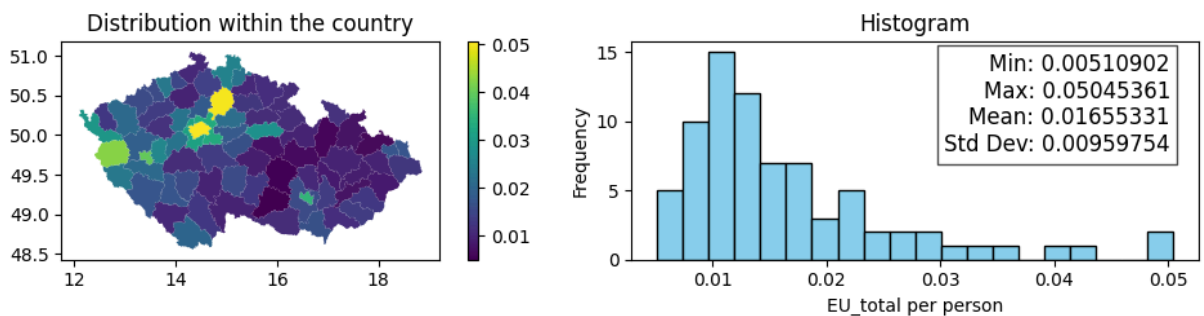
Men per person



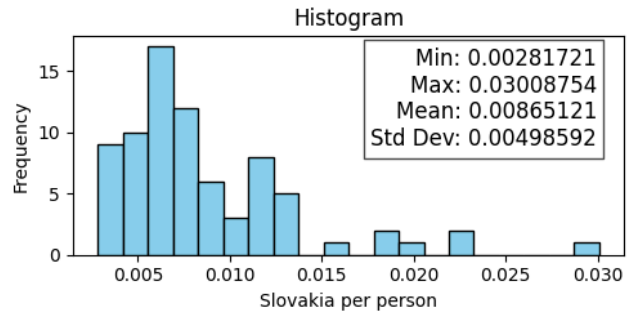
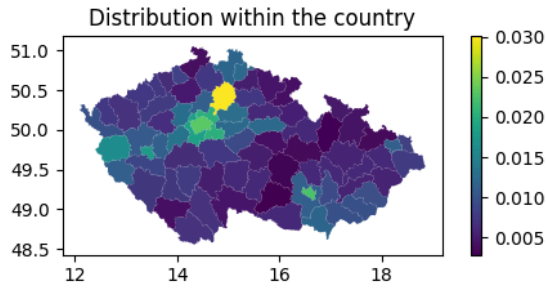
Women per person



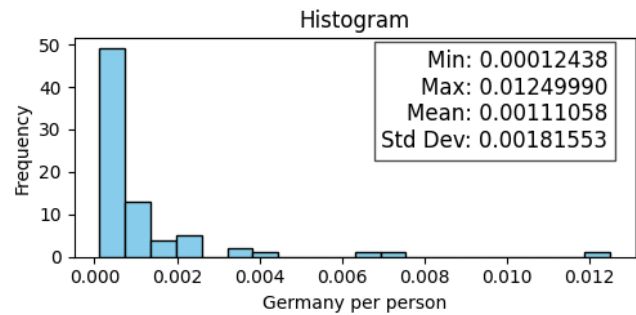
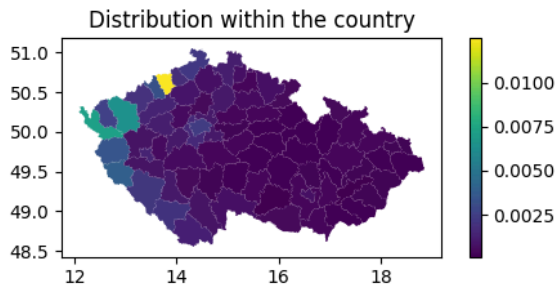
Eu_total per person



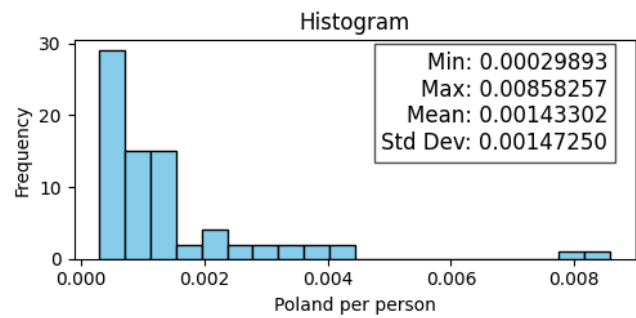
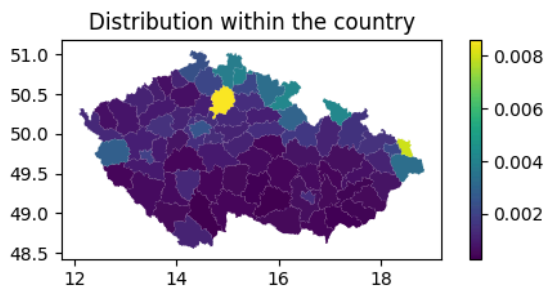
Slovakia per person



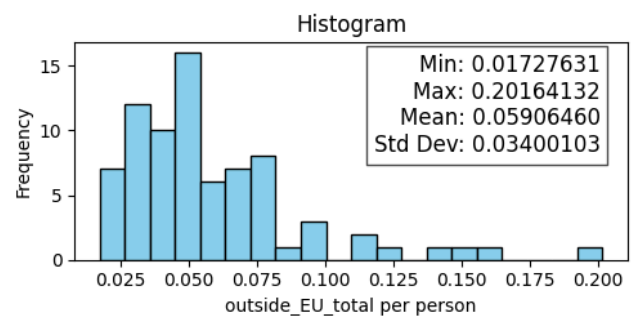
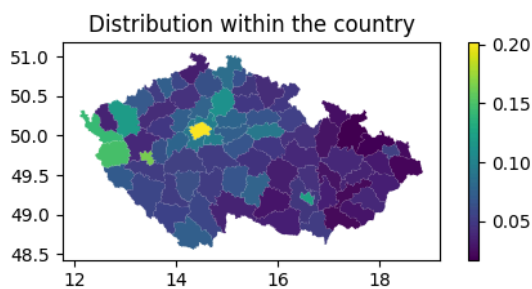
Germany per person



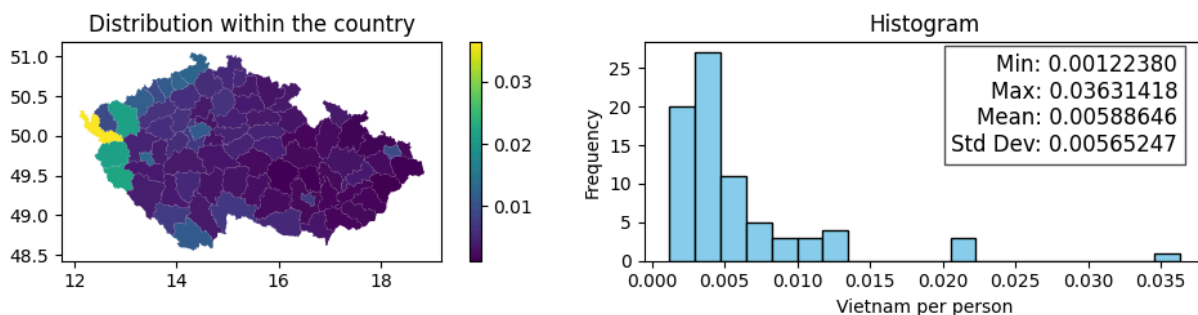
Poland per person



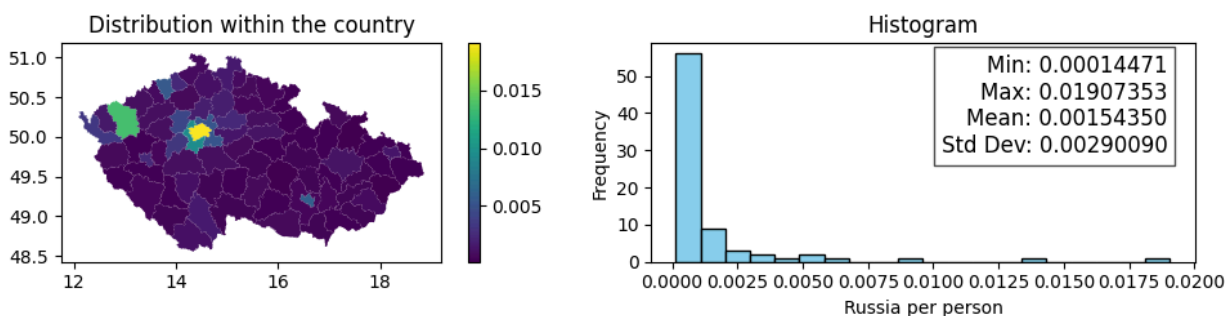
Outside_eu_total per person



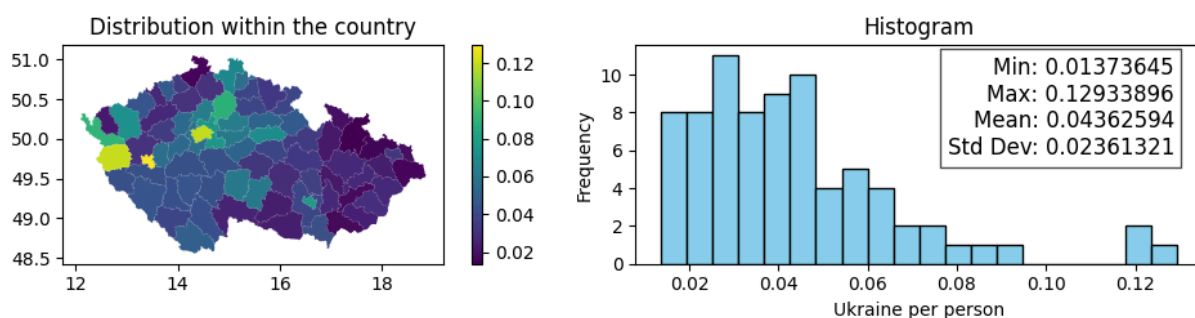
Vietnam per person



Russia per person



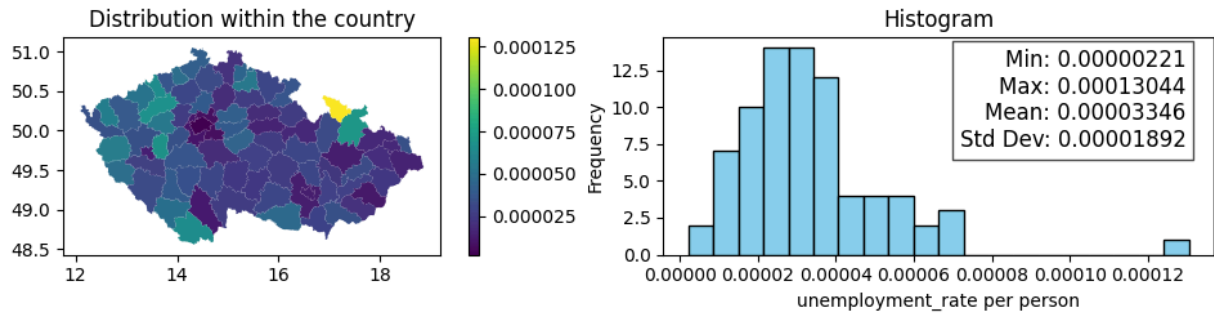
Ukraine per person



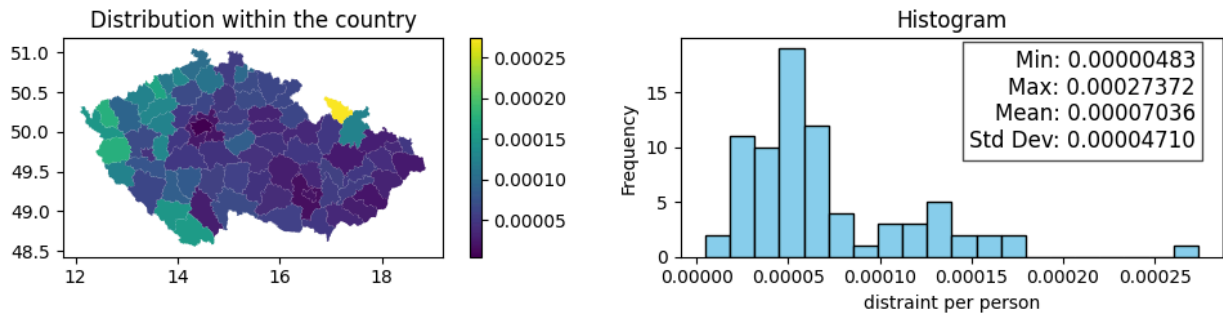
Socioekonomické ukazatele

Některé ukazatele, jako nezaměstnanost, procento obyvatel v exekuci, či pobírání sociálních dávek libovolného druhu na osobu, spolu zcela jednoznačně souvisí. Zvýšené hodnoty se vyskytují ve stejných oblastech, ale souvisí spolu už jen z principu věci - v oblastech s vyšší nezaměstnaností bude pravděpodobně více lidí pobírat příspěvky na živobytí od státu. S ekonomickou situací v regionu vlastně přímo nesouvisí jen pobírání porodného a pohřebného.

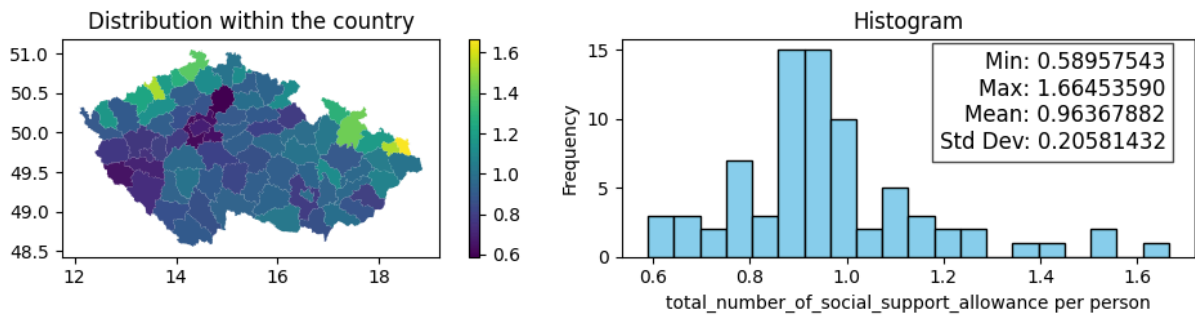
Unemployment_rate per person



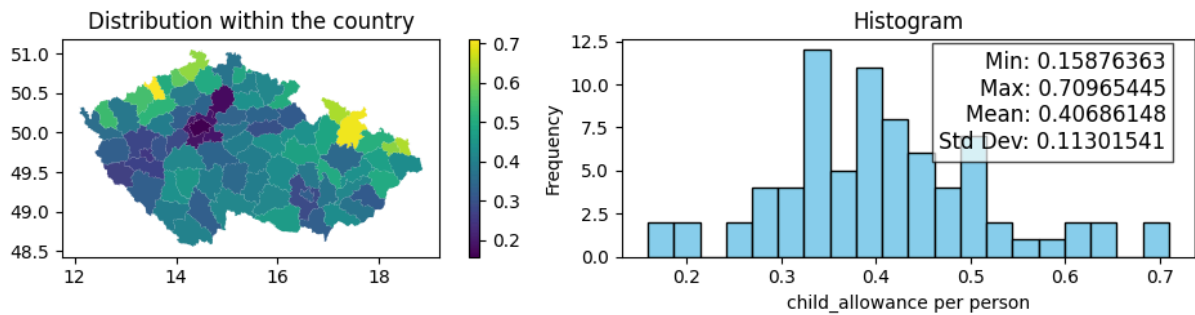
Distrant per person



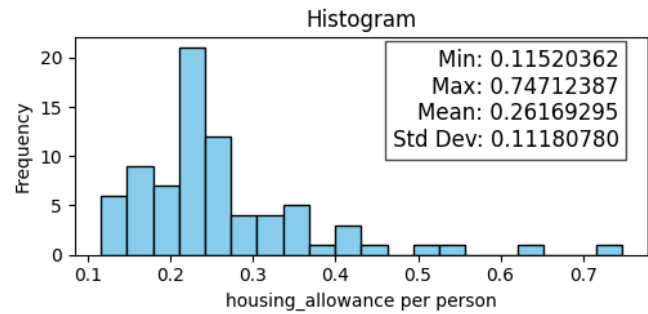
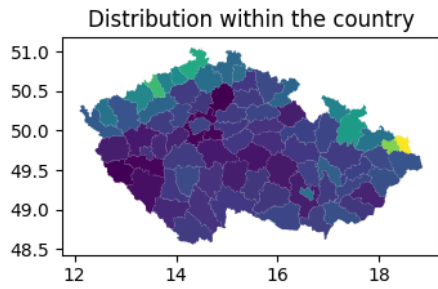
Total_number_of_social_support_allowance per person



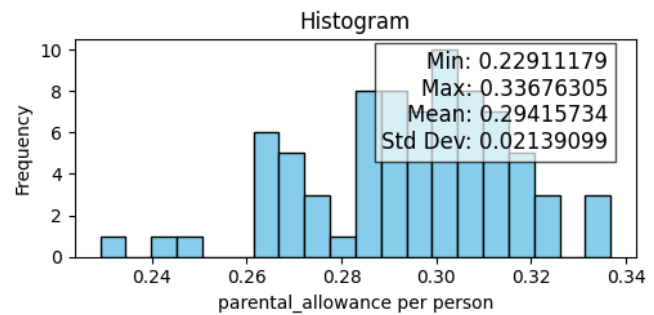
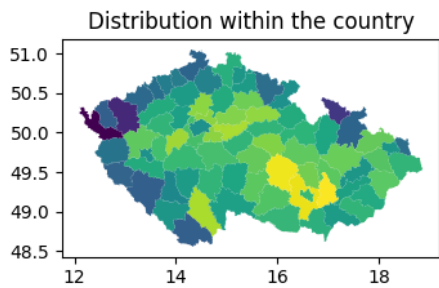
Child_allowance per person



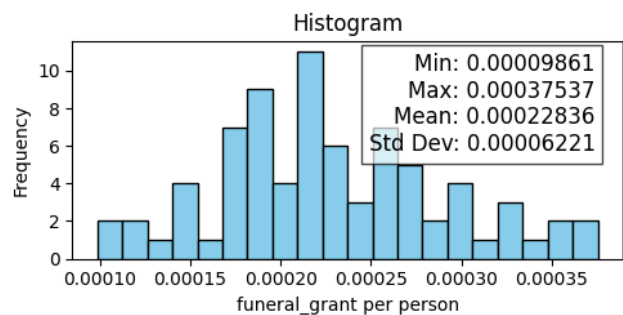
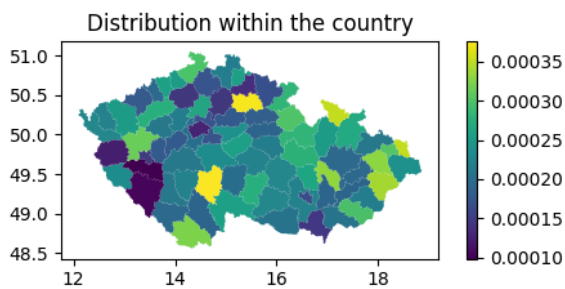
Housing_allowance per person



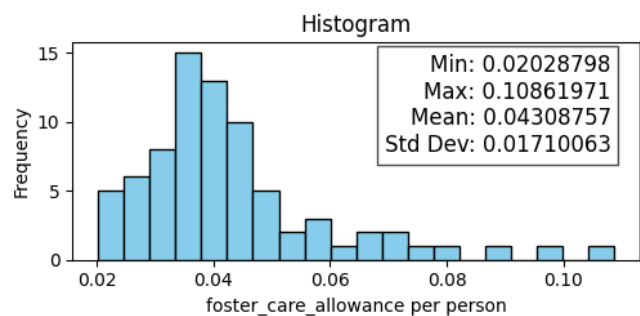
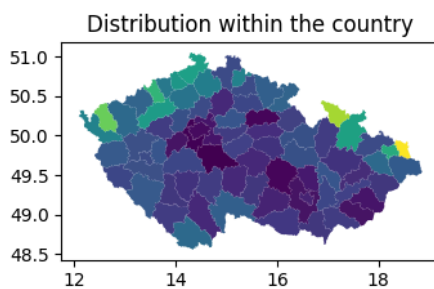
Parental_allowance per person



Funeral_grant per person

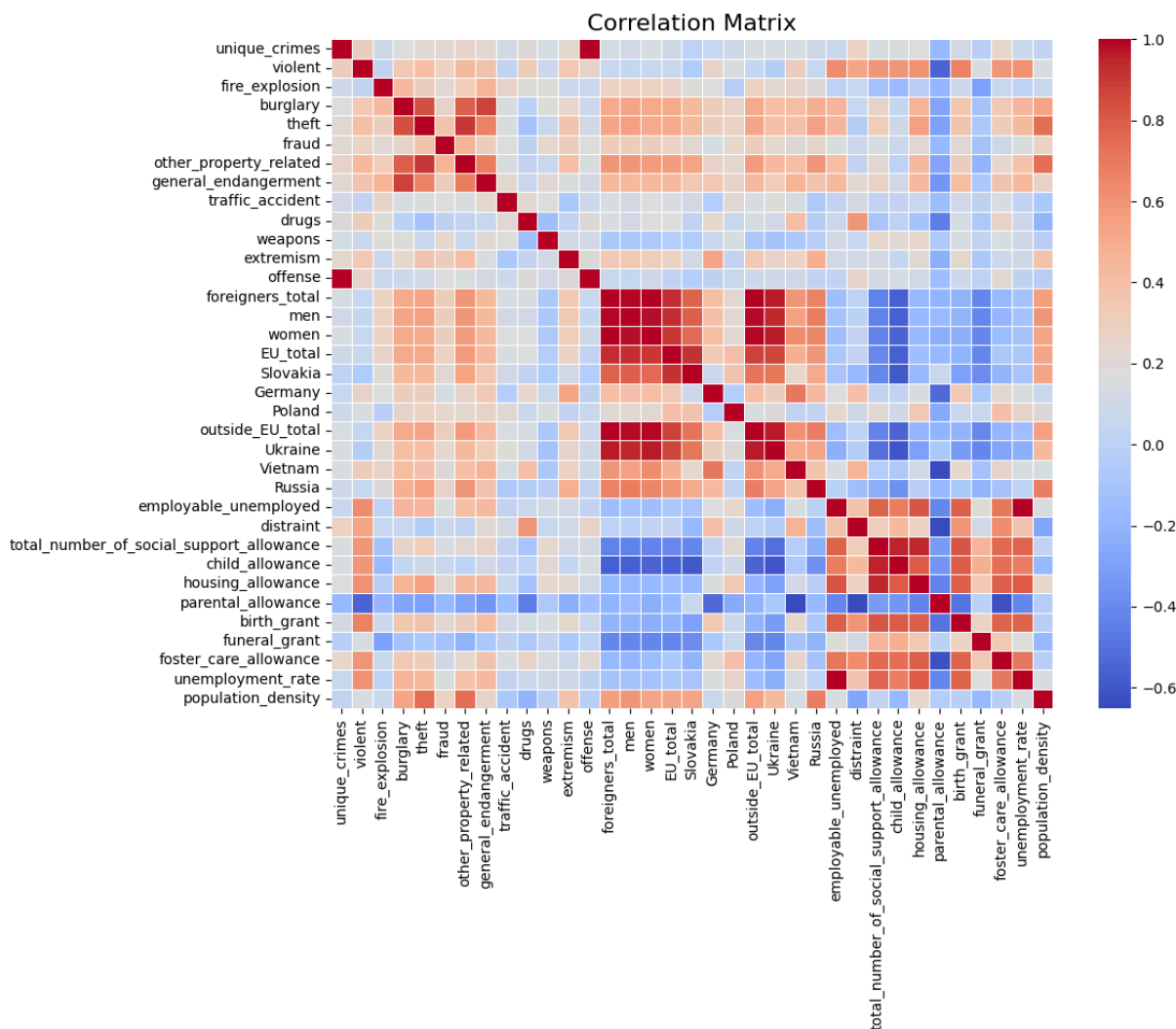


Foster_care_allowance per person



Korelační analýza

Nad datasetem jsme vytvořili korelační matici, abychom vyšetřili vztahy mezi jednotlivými proměnnými a případně odhalili ty problematické, které by mohly způsobovat problémy.



Vysoké hodnoty korelací

Z korelační matice je patrné, že v datech existuje poměrně mnoho korelovaných párů.

- Vysoká pozitivní korelace proměnných unique crimes a offense - pravděpodobně způsobeno tím, že velká část deliktů, se kterými pracujeme, je zařazena mezi přestupky.
- Vyšší pozitivní korelace mezi proměnnými burglary, theft, other_property_related a general_endangerment - může být způsobeno faktem, že tyto trestné činy jsou často páchaný podobným typem pachatelů, nebo může jeden čin být často klasifikován do více z těchto kategorií najednou.
- Velmi vysoká pozitivní korelace u foreigners_total, men a women - z toho důvodu, že srovnáváme podíl cizinců (všech) v populaci a podíl cizinců (mužů) v populaci či podíl cizinek (žen) v populaci. Tyto proměnné spolu jednoznačně souvisí, protože čím více všech cizinců je v populaci, tím více bude v populaci cizinců-mužů a tím více to bude cizinek. Pro další využití těchto proměnných by možná bylo lepší

přepočítat poměr žen a mužů na počet cizinců, nikoli počet obyvatel. Podobný problém nastává u EU_total a outside_EU_total.

- Zajímavá je středně vysoká pozitivní korelace proměnné violent a různých prediktorů týkajících se sociálních dávek, která naznačuje, že pachatelé násilí pravděpodobně často pocházejí z horších sociálních podmínek. Zajímavé také je, že u loupeží a krádeží tato korelace není tak významná.
- Obecně negativní korelace je u prediktorů týkajících se sociálních dávek vs. prediktorů týkajících se cizinců, což by se dalo vysvětlit tak, že cizinci se většinou stěhují do bohatších okresů ČR a nepobírají sociální dávky.
- Vyšší pozitivní korelace se vyskytuje i mezi většinou prediktorů týkajících se sociálních dávek - pravděpodobně z důvodu, že rodiny v horší ekonomické situaci pobírají často více dávek najednou.

Abychom mohli správně provést lineární regresi, je vždy potřeba vyšetřit, jestli nepracujeme s některými z těchto vysoce korelovaných párů a případně tuto situaci vhodně ošetřit, např. pomocí VIF.

Korelace mezi prediktory týkajícími se kriminality a cizinců

Mezi celkovým počtem trestných činů a prediktory, které se týkají cizinců, se vyšší hodnoty korelace nevyskytují, zajímavější jsou ale výsledky u jednotlivých typů trestné činnosti.

Pozitivní korelace se vyskytuje (namátkově, páru je mnoho):

- Mezi celkovým poměrem cizinců a majetkovou trestnou činností vč. loupeží a krádeží, i u většiny národností cizinců kromě Němců a Poláků.
- Mezi poměrem Vietnamců a drogovou činností; u ostatních národností se pozitivní korelace ve vztahu k drogám nevyskytuje.

Negativní korelace se vyskytuje např. mezi poměrem cizinců v populaci a kriminalitou, která se týká střelných zbraní.

Na druhou stranu, prediktory týkající se cizinců rozhodně nejsou jediné korelující s kriminalitou. Např. hustota zalidnění vysoce koreluje s majetkovou trestnou činností, a také s poměrem cizinců, proto si nemůžeme rovnou domýšlet kauzalitu.

Regresní a GLM modely

Regresní a generalizované lineární modely (GLM) byly využity k analýze vztahů mezi kriminalitou a vybranými socioekonomickými faktory v okresech České republiky, kdy jsme použili počet trestných činů na osobu v okrese jako závislou proměnnou y . Níže uvádíme detailní analýzu výsledků jednotlivých modelů a jejich interpretaci, zaměřenou na srozumitelné vysvětlení.

Podíl cizinců

Samostatná analýza podílu cizinců ukázala, že větší podíl cizinců v okresech je spojen s vyšší kriminalitou. Tento vztah byl statisticky významný ($p < 0.001$), ale podíl cizinců vysvětlil pouze 28 % variance kriminality ($R^2 = 0.279$).

Zlepšení modelu bylo dosaženo použitím logaritmické transformace kriminality, která zlepšila normalitu reziduí ($p = 0.353$) a odstranila problém s heteroskedasticitou. To naznačuje, že podíl cizinců má vliv na kriminalitu, ale pravděpodobně působí ve spojení s dalšími faktory.

Kombinace podílu cizinců a nezaměstnanosti

Kombinace podílu cizinců a nezaměstnanosti vytvořila robustnější model, který vysvětlil 63 % variance kriminality ($R^2 = 0.628$). Obě proměnné byly statisticky významné ($p < 0.001$), přičemž podíl cizinců měl silnější vliv než nezaměstnanost.

Model byl statisticky robustní, protože rezidua vykazovala normalitu a nebyla přítomna heteroskedasticita. Tento výsledek naznačuje, že kombinace těchto dvou faktorů lépe vysvětluje rozdíly v kriminalitě mezi okresy.

Sociální podpora a podíl cizinců

Model zahrnující podíl cizinců a úroveň sociální podpory vysvětlil 72 % variance kriminality ($R^2 = 0.721$). Sociální podpora měla negativní vliv na kriminalitu, což znamená, že vyšší úroveň sociální podpory byla spojena s nižší kriminalitou. Naopak podíl cizinců zůstal pozitivně spojen s nárůstem kriminality.

Tento model byl nejrobustnější, protože diagnostika ukázala, že rezidua byla normálně rozložena ($p = 0.951$) a model nebyl ovlivněn heteroskedasticitou.

Generalizované lineární modely (GLM)

GLM byly použity ke zlepšení predikční schopnosti modelů, kdy jsme vzali gamma modely s logaritmickou linkovací funkcí. Tyto modely umožnily lépe zachytit nelineární vztahy a lépe se vypořádaly s problémy diagnostiky.

1. GLM s podílem cizinců a nezaměstnaností:

- Tento model potenciálně vysvětluje 81 % variance kriminality ($R^2 = 0.811$).

- Podíl cizinců i nezaměstnanost měly pozitivní a významný vliv na kriminalitu ($p < 0.001$).
- Jeho AIC bylo -728.7317

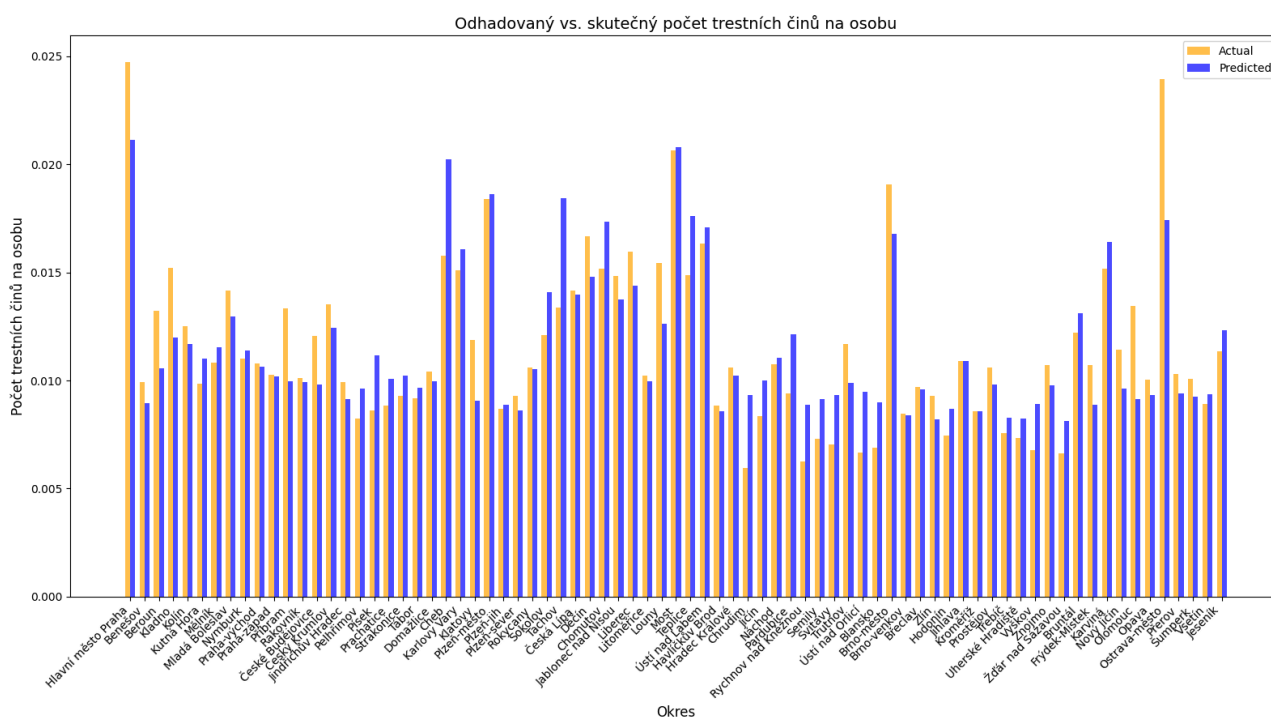
2. GLM zahrnující sociální podporu, podíl cizinců a nezaměstnanost:

- Tento GLM model potenciálně vysvětluje 89 % variance kriminality ($R^2 = 0.8977$).
- Sociální podpora měla ochranný vliv, zatímco podíl cizinců a nezaměstnanost měly pozitivní vliv.
- Tento model poskytl nejlepší statistické vlastnosti a ukázal komplexní vztahy mezi zkoumanými faktory.
- Jeho AIC bylo -741.3077

3. GLM zahrnující sociální podporu, podíl cizinců a procento v exekuci:

- Tento model byl o trochu lepší než předchozí ve vysvětlování variability, vysvětluje totiž potenciálně 90 % variance ($R^2 = 0.9069$) a zároveň při přidání všech tří socioekonomických ukazatelů došlo k ztrátě významu nezaměstnanosti, tedy byl výběr mezi tímto a předchozím modelem, jako nejlepším.
- Rozhodnutí vybrat tento jako nejlepší utvrdilo i AIC, které je -744.1834 a je tedy nejnížší ze všech tří. A zároveň jeho složitost je stejná jako druhý nejlepší model.
- Všechny tři znaky modelu měli pozitivní vliv na kriminalitu a zároveň jejich p-hodnoty byly velmi nízké (největší $p < 0.002$), což potvrzuje jejich významnost.
- Vzorec tohoto modelu je

$$\log(y) = -5.6562 + 0.0470(\text{cizinci}) + 0.0305(\text{exekuce}) + 0.6012(\text{dávky})$$
- Obrázek níže je predikce tohoto modelu na datech



Ověření hypotéz

S ohledem na vybraný model nám vyšlo, že podíl cizinců má pozitivní vliv na kriminalitu, kdy při nárůstu o procento dojde k 4,8 % nárůstu páchané kriminality na osobu v okrese, ale jedná se jen o obecnou informaci. Proto jsme vyzkoušeli další modely, které zkoumaly, zdali je i důležité pohlaví cizince či národnost. Dále jsme chtěli otestovat hypotézu, kterou jsme měli z korelační analýzy, a to zda cizinci ovlivňují, jaký typ kriminality je páchán a jestli skutečně mají cizinci pozitivní vliv na burglary, theft, other_property_related a general_endangerment.

Hypotéza o pohlaví

K ověření této hypotézy jsme vzali náš nejlepší model a vyměnili procento cizinců za dva znaky - procenta mužů cizinců a žen cizinek v populaci. Došlo k mírnému zhoršení modelu, kdy nám porostlo AIC a kleslo i pseudo R², ale výsledkem bylo, že pohlaví má vliv jen v případě mužů, u žen to má jen minimální záporný a nevýznamný vliv ke kriminalitě ($p_{hodnota}(ženy) = 0.957$), zbylé dva socioekonomické ukazatele zůstaly statisticky významné.

Rovnice modelu:

$$\log(y) = -5.6628 + 0.0963(muži) - 0.0022(ženy) + 0.0305(exekuce) + 0.6036(dávky)$$

Hypotéza o národnosti

K tomuto účelu jsme vytvořili GLM model, který byl prvotně složen z nejpočetnějších národnostních menšin (Slovensko, Rusko, Vietnam, Německo, Polsko, Ukrajina) a zbytek jsme rozdělili do národností z EU a mimo EU. Vyšel nám model, kde několik národností nemělo silný vliv na model, tak jsme provedli regularizaci Lasso pro vhodný výběr znaků, po které nám zbyli Vietnamci, Poláci, Ukrajinci, Slováci, a menšiny z nečlenských zemí EU. Všechny tyto menšiny mají pozitivní vliv na nárůst kriminality na osobu, kromě Ukrajinců, to by se mohlo vysvětlit tím, že ukrajinskou menšinu tvoří spíše ženy, kdy jsme v hypotéze o pohlaví odvodili sice nevýznamný ale negativní vliv žen na kriminalitu.

Rovnice zjednodušeného modelu zaokrouhlená na dvě desetinná místa pro úsporu:

$$\log(y) = -4.78 + 0.26(vn) + 0.21(other) + 0.56(pl) + 0.07(sk) - 0.04(ua)$$

Hypotéza o druhu páchané kriminality

Pro nedostatek dat byly vynechány extremistické trestné činy a trestné činy související se zbraněmi. Po užití GLM modelů k vysvětlení variance spáchaných druhů trestných činů, nám vyšlo na hladině významnosti, že procento cizinců ovlivňuje počet majetkových trestných činů na osobu typu krádež, vloupání a podvodů. Dále byl významný u obecného ohrožení, ale ne na jiných násilných činech. To souhlasí s prvotním průzkumem dat a korelační analýzou.

Diskuze a závěr

Hlavním cílem bylo zjistit, zda má podíl cizinců v populaci vliv na kriminalitu. K tomuto účelu jsme vytvořili nejdříve modely lineární regrese, kdy jsme začali nejjednodušším modelem, který obsahoval pouze podíl cizinců. Tento model jsme postupně vylepšovali přidáváním a kombinováním jiných socioekonomických znaků jako confounder variables, konkrétně jsme vybrali nezaměstnanost, exekuce a sociální dávky. Nakonec jsme přešli ke GLM modelům, kde nám vyšel nejlepší model zahrnující podíl cizinců, exekuce a sociální dávky. Podíl cizinců celou dobu zůstal relevantním, statisticky významným prediktorem.

Pokračovali jsme testováním dalších hypotéz - zda nějaké charakteristiky cizinců mají vliv na kriminalitu. Zaměřili jsme se na pohlaví a národnost, dále jsme chtěli otestovat naši nalezenou korelaci, a to zda i podíl cizinců ovlivňuje druh páchané kriminality. K pohlaví jsme zjistili, že to má statisticky významný vliv jen u mužů, u žen to nemá podstatný význam. Dále jsme vyzkoušeli GLM model rozdělující cizince do národností, kdy jsme ubíráním znaků pomocí LASSO regularizace, došli pouze k několika národnostem. Nakonec jsme otestovali náš poznatek z korelační analýzy, který se potvrdil, a to, že cizinci mají vliv na majetkové trestné činy a obecné ohrožení.

Závěrem lze říci, že naše modely potvrzují souvislost mezi nárůstem počtu cizinců v okrese a vyšší kriminalitou na osobu. Nelze vyloučit existenci jiné, dosud neodhalené matoucí proměnné, která by mohla tento vztah ovlivnit. Proměnné týkající se cizinců si však udržely statistickou významnost i po zahrnutí relevantních socioekonomických faktorů spojených s kriminalitou. Zjistili jsme také, že kriminalitu zvyšuje přítomnost mužů mezi cizinci a že vliv na růst kriminality mají pouze některé národnosti.

Výkaz práce

Odvedená práce

M. Pejšová (50 %):

- výběr tématu - 2 hod
- shromáždění vhodných dat – 3 hod
- stanovení tématu, úvodní dokument - 4 hod
- zpracování dat, spojení datasetů – 5 hod
- zpracování geografických dat, příprava pro vizualizace – 4 hod
- datová analýza - EDA - 3 hod
- datová analýza - korelační analýza - 2 hod
- závěrečný report – 5 hod

D. Mikulík (10 %):

- Úvodní dokument a stanovení tématu – 4h
- datová analýza - korelační analýza - 1h
- Závěrečný report – 5h

J. Švec (20 %):

- Úvodní dokument a stanovení tématu – 4h
- vizualizace a úprava dat – 10h
- analýza a třídění výstupů – 2h
- Závěrečný report – 5h

H. Tran (20 %):

- Úvodní dokument a stanovení tématu – 4h
- zpracování confounder dat - 4h
- datová analýza - lineární regrese a GLM - 11h
- report - 4h