Skóre rizikovosti a sčítání lidu - co jsme zjistili

Michal Půlpán

08. 01. 2023

1 Úvod

V poslední několika letech jsme v České republice mohli zaznamenat několik zajímavých mikroekonomických událostí. Výrazně narostla kupní síla, obyvatelstvo více utrácelo, ale i se znatelně zadlužilo. To se mimo jiné projevilo i na růstu e-commerce a maloobchodu. Další nevídanou situací byl i znatelný růst cen nemovitostí napříč českými městy i suburbii způsobený mimojiné nezastavitelnou poptávkou po novém bydlení, investicím a nebo jen touze získat více místa či zahradu.

To vše, vzhledem k negativnímu vývoji ekonomické situace, má vliv na rekordní zadlužení obyvatelstva. Každá instituce poskytující úvěry fyzickým osobám se tak musí chovat opatrněji a věnovat více pozornosti, komu peníze poskytuje. V tomto projektu se podíváme na to, jak lze využít moderních technologií, věřejně dostupných dat a data science ke svému prospěchu.

Pro finanční instituace poskytující úvěry, či obchodníky nabízející své zboží nebo služby s odloženou splatností, by tak mohlo být zajímavé, jaké faktory ovlivňují schopnost splácení klientů. To by samo o sobě mohlo napomoci k větší obezřetnosti prodejce, přesnějšímu marketingovému cílení a nebo naopak k uzpůsobení služeb jednotlivým skupinám obyvatelstva.

Podíváme se tak na vztah mezi agregovanými daty ze sčítání lidu a podílu exekucí a bankrotů v jednotlivých obcích a pokusíme se z nich odvodit, jaké sociodemografických údaje ovlivňují, zda jsou obyvatelé schopni splácet své dluhy či nikoli.

1.1 Co od dat očekáváme

Klást budeme důraz na to, jaké sociodemografické údaje ovlivňují, zda jsou lidé schopni splácet své dluhy. Přesněji řečeno, jaký je vztah mezi podílem exekucí a bankrotů s různými údaji o obyvatelstvu.

1.1.1 Očekávání a překvapení

Již při prvním pohledu na data (resp. jejich strukturu a obsah) člověk začne mít nějaká první očekávání - například, že by se mohlo ukázat, že větší podíl exekucí a bankrotů má vliv na věkovou strukturu obyvatelstva, nebo na vzdělání. To by bylo překvapení, protože by to znamenalo, že lidé s vyšším vzděláním jsou schopni splácet své dluhy lépe než ostatní. Je tomu tak nebo se ukáže, že věková struktura a vzdělání nemají žádný vliv na schopnost splácet dluhy? Nemůže se ukázat, že porovnávat věkovou strukturu a vzdělání nedává příliš smysl neboť věková struktura a vzdělání mohou záviset i na faktu, že míra dosaženého vzdělání postupně roste? Očekávání tak musíme ověřit a překvapení najít v průběhu bádání.

Výsledkem tohoto projektu by mělo být zobrazení vztahu mezi sociodemografickými údaji a podílem exekucí a bankrotů v jednotlivých obcích. Výsledný model by mohl být schopen na základě několika

sociodemografických údajů o populaci odhadnout míru její zadluženosti.

2 Data

Využijeme vybrané výsledky ze sčítání lidu z roku 2011 a 2021 a data o exekucích a bankrotech za jednotlivé obce České republiky.

2.1 Původ dat

Celá datová sada za jednotlivé obce České republiky obsahuje:

- vybrané výsledky sčítání lidu 2011 a 2021
- počet obyvatel celkem v jednotlivých sociodemografických skupinách
- výskyt vybraných trestných činů
- podíl osob v bankrotu či exekuci

Byla stažena ze stránek nextcloud.profinit.eu, pravděpodobně je ale původ neagregovaných datasetů od Českého statistického úřadu.

2.2 Stručný popis

Pojďme si představit používané datasety a jejich strukturu. Oba datasety jsou díky dělení na obce velmi granulované a tím i relativně dlouhé.

2.2.1 Sčítání lidu (Cenzus)

Data ze sčítání lidu z let 2011 a 2021 (`geodata/census11_21/data_obce_vyhl_nevyhl.csv`) jsou obsáhlá v jednom souboru a obsahují informace o jednotlivých obcích. Každá obec má svůj řádek a sloupce obsahují informace o jednotlivých sociodemografických skupinách. Výsledky jsou však agregované, takže neobsahují informace o jednotlivých obyvatelích, ale o celkovém počtu obyvatel v daných skupinách. V datech tak máme například pro každou obec v České republice informaci o počtu dětí, vzdělání, věku, rodinného stavu, občanství, ale i o vlastníkovi domů (družstvo, fyz. osoba, obec/stát, podílové) či stavebního materiálu domu, ve kterém obyvatelé žijí a mnoho více.

Dataset je složený z 6246 pozorování o 160 proměnných. Díky tomu je dataset velmi široký a obsahuje velikou škálu sociodemografikcých údajů. Každé pozorování je vytvořeno pro obec, kterých je celkem 6246.

2.2.2 Exekuce a bankroty

Data o exekucích a bankrotech (`geodata/score/obce_skore_rizikovosti.csv`) jsou dělená dle obcí České republiky a obsahují informace o podílu exekucí (klouzavý průměr za roky 2018-2021) a bankrotů (klouzavý průměr za orky 2018-2021) v jednotlivých obcích. Výsledky jsou agregované, takže neobsahují informace o jednotlivých obyvatelích, ale o celkovém počtu exekucí a bankrotů v dané obci. Mimojiné obsahují i tzv. skóre obce, které je výsledkem normovaného průměru počtu exekucí a bankrotů v obci. Toto skóre je v rozmezí 0 až 1, kde 0 znamená, že v obci nebyly žádné exekuce ani bankroty a 1 znamená, že v obci byly exekuce a bankroty v průměru v každém roce.

Dataset je složený z 6254 pozorování o 10 proměnných. Každé porování je vytvořeno pro obec, kterých je celkem 6254.

2.3 Kvalita dat

Data ze sčítání lidu z nějakého důvodu neobsahují data ze 14 obcí. Dle struktury území ČR mezi roky 2013 a 2022 od ČSÚ je v České republice (od 1.1.2016) celkem 6260 obcí. V našich datech je však "pouze" 6246 resp. 6254 obcí pro dataset exekucí. Těžko říct, zda je to chyba v našich datech, nebo v původních datech od ČSÚ, ale i mezi datasety je rozdíl v 8 obcích, které ve výsledcích ze sčítání lidu chybějí. Konkrétně se jedná o obce:

- Krhová v okrese Vsetín (500062)
  - 2024 obyvatel
- Poličná v okrese Vsetín (500071)
  - 1745 obyvatel
- Bražec v okrese Karlovy Vary (500101)
  - 221 obyvatel
- Doupovské Hradiště v okrese Karlovy Vary (500127)
  - 160 obyvatel
- Kozlov v okrese Olomouc (500135)
  - 270 obyvatel
- Luboměř pod Strážnou v okrese Přerov (500151)
  - 122 obyvatel
- Město Libavá v okrese Olomouc (500160)
  - 0 obyvatel
- Polná na Šumavě v okrese Český Krumlov (500194)
  - 202 obyvatel

Ač se jedná o obce (pro nás) v relativně zajímavém místě (s předpokládanou horší ekonomickou situací), tak se jedná o relativně malé obce, takže bychom je nemuseli brát v úvahu.

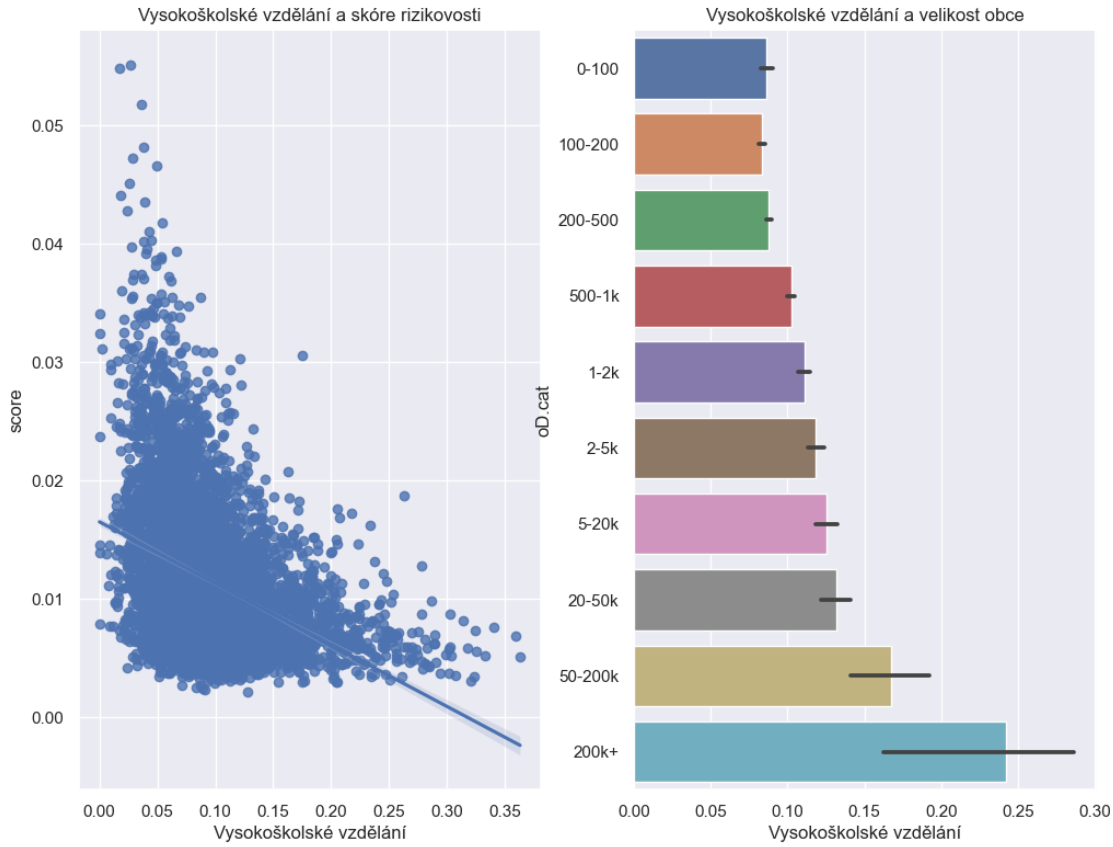
Dle výstupů z nástroje `pandas-profiling` žádný ze zmíněných datasetů nemá problém s chybějícími daty ve sloupcích. Jediné, co je vhodné upravit jsou názvy sloupců, ale ty jsou vždy vysvětleny v přiložených textových souborech.

3 Příběhy

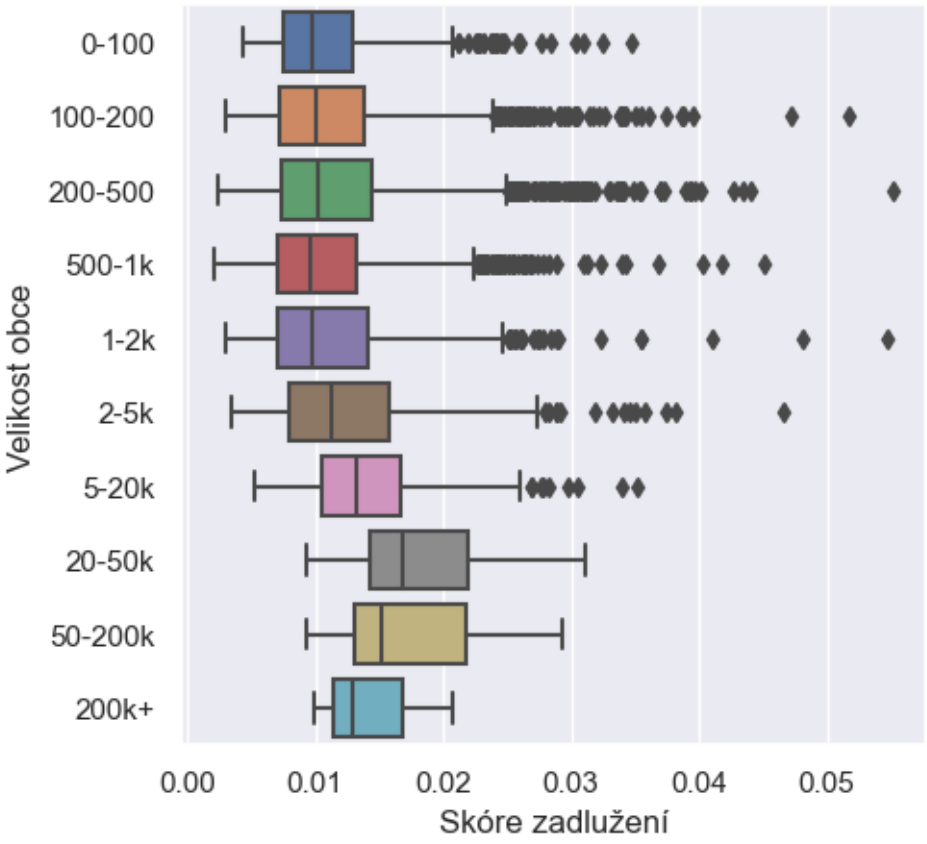
Poskytnuté datasety jsou velmi obsáhlé a obsahují veliké množství informací. Na jejich základě by mělo být možné vytvořit (a podložit) veliké množství příběhů. Ukážeme si jeden z nich.

3.1 Vzdělání a dluhy

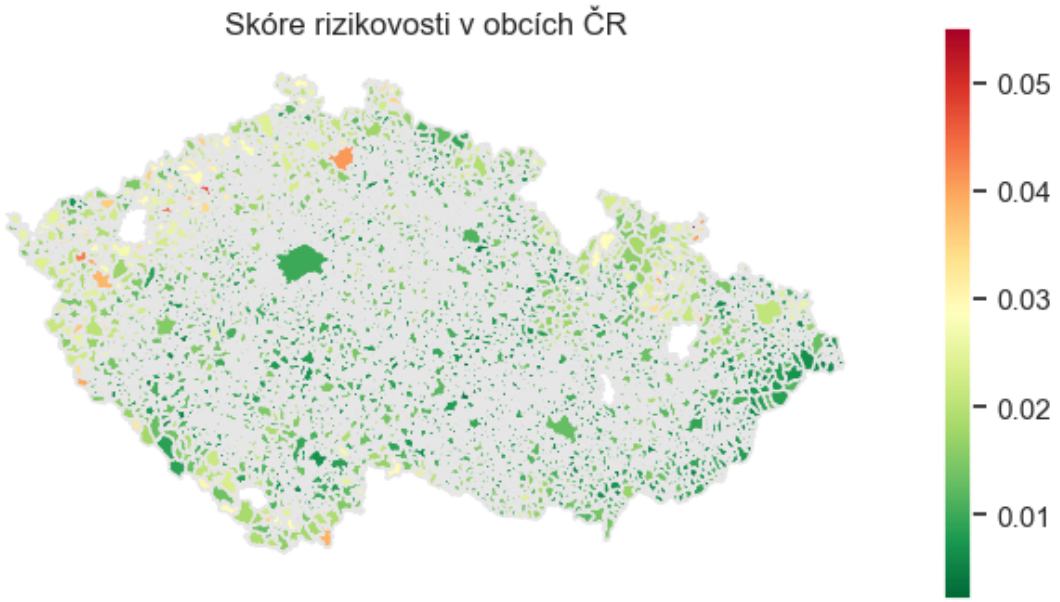
Této hypotéze se podíváme na vztah vzdělání a zadlužení obyvatel českých obcí. Můžeme nějak využít informace o počtu vysokoškoláků zastoupených v obci k tomu abychom určili míru její zadluženosti?

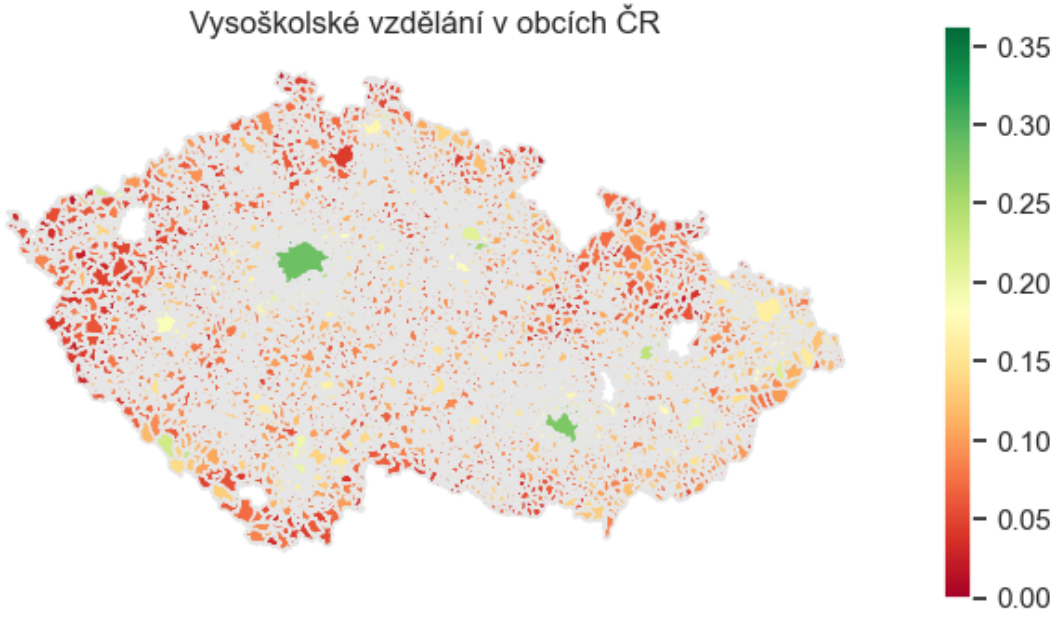


Dle grafů výše se zdá, že s přibývajícím počtem vysokoškoláků v obci klesá i skóre zadlužení obce. To znamená, že obce s vysokoškoláky mají méně dluhů. To je ale jen první pohled na data. Co když se podíváme na poměr mezi počtem vysokoškoláků a počtem obyvatel v obci? Jak můžeme vidět, s velikostí obce nám roste i počet vysokoškoláků. Jaký vztah tedy má velikost a zadlužení obce?



Pro větší přehlednost a možnou představu (s demografickou a geografickou znalosti ČR) by nám mohla pomoci mapa ČR. Vytvoříme si tedy mapu ČR, kde budeme moci vidět, jak jsou vysokoškoláci a skóre rizikovosti rozděleni po jednotlivých obcích.

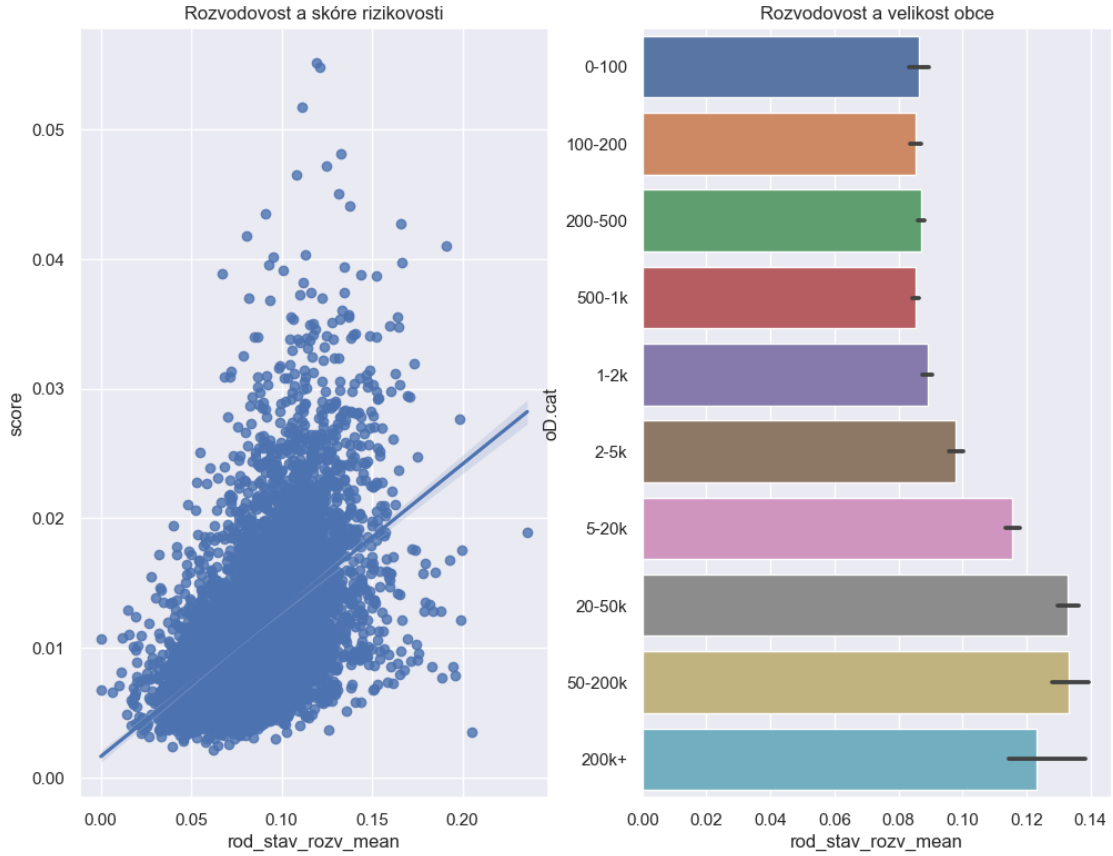




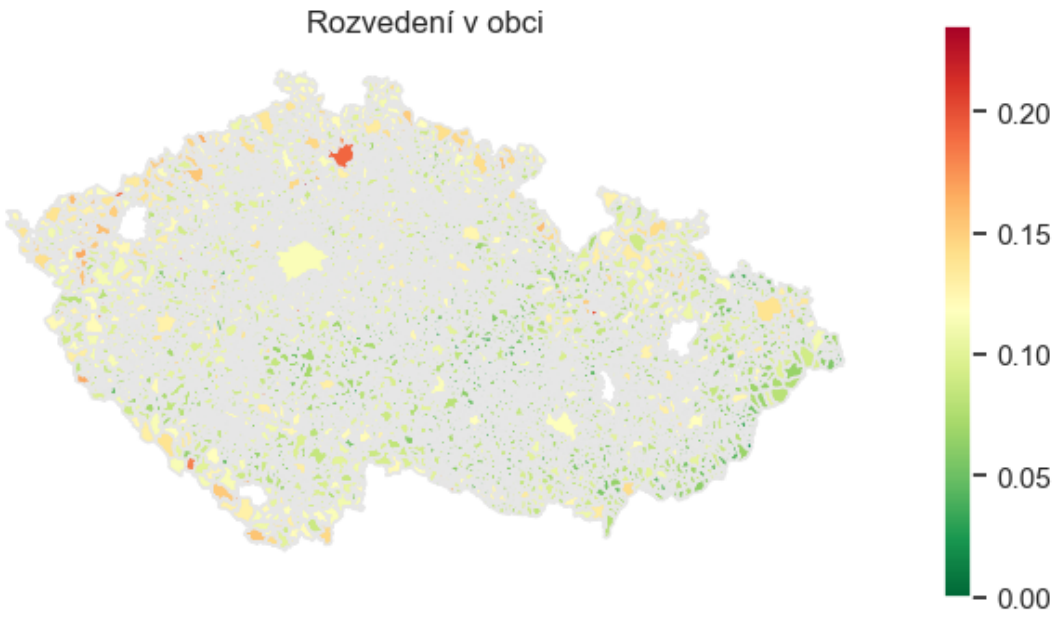
Dle výše uvedených map můžeme poměrně hezky vidět, že by teoreticky nějaký vztah mezi vzdělaností a rizikovostí obce opravdu existovat mohl. Dobrým příkladem by mohl být Karlovarský a Ústecký kraj a celkově většina pohraničí (dokonce prakticky kopírující hranici bývalých Sudet).

3.2 Rozvodovost a dluhy

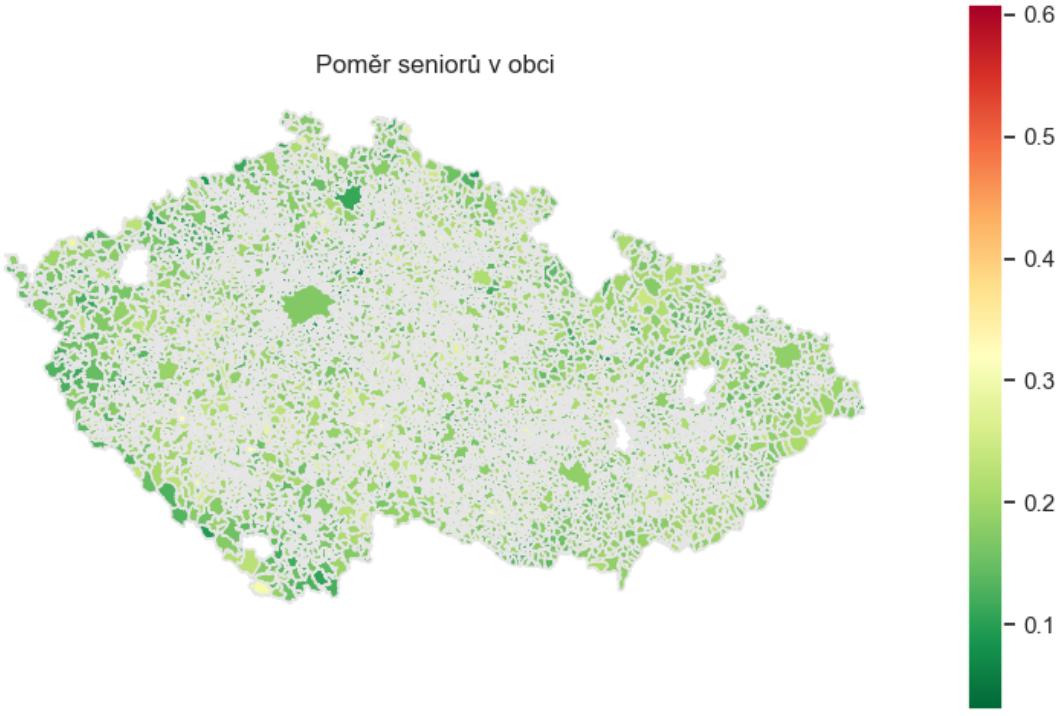
Pojďme si představit druhý příběh, který by mohl člověka napadnout v kontextu dluhů. A to ve vztahu rozvodovosti obyvatel českých obcí. Můžeme nějak využít informace o počtu rozvedených obyvatel v obci k tomu abychom určili míru její zadluženosti? Pojďme se podívat na data.



Opět se může zdát, že nějaký vztah mezi rozvodovostí a zadlužeností obce existuje. Pokud si rozvedené v obcích znázorníme na mapě ČR, můžeme vidět, že nám opět rozvodovost mírně koreluje se skóre rizikovostí. Rozvodovost ale nemůžeme brát jako jediný faktor neboť sami můžeme vidět, jak je nízká rozvodovost v ČR neobvyklá. Jediné oblasti ČR s rozvodovostí pod 5% jsou na hranici se Slovenskem (tedy oblast od Jablůnkova po Uherský Brod a poté oblast vymezená Litomyšlí a Jihlavou (tedy převážně oblast Železných hor a Žďárských vrchů)).



V tuto chvíli by stálo za to se pozastavit a rozvodovost více porovnat s poměrně důležitým faktorem - věkem.



Můžeme opět náhlednout, že oblasti s nižším zastoupením seniorů opět mírně korelují s vyšším

skóré rizikovosti.

Zajímavého úkazu, kterého si můžeme na mapách všimnout je pokaždé výrazné Ralsko (město cca mezi Mělníkem a Libercem). Zprvu jsem ho považoval za chybu v datech, ale po zkontrolování a zjištění faktu, že se jedná o obec známou především díky vojenskému prostoru jenž opustila sovětská armáda se zdají data být správná.

Ač je zkoumání rozvodovosti a dluhů poměrně zajímavé, dále tuto hypotézu rozvíjet nebudeme a bude sloužit jen jako příklad dalšího možného vztahu mezi daty.

4 Modelování

V předchozí sekci jsme si ukázali dvě možné hypotézy v kontextu dluhů. V této sekci se pokusíme o vytvoření modelu, který by nám umožnil předpovědět skóre rizikovosti obce na základě vzdělání a rozvodovosti obyvatel. Vytvoříme si tedy model, který na základě vstupních proměnných (sociodemografické údaje) o obci bude schopen předpovídat její skóre rizikovosti.

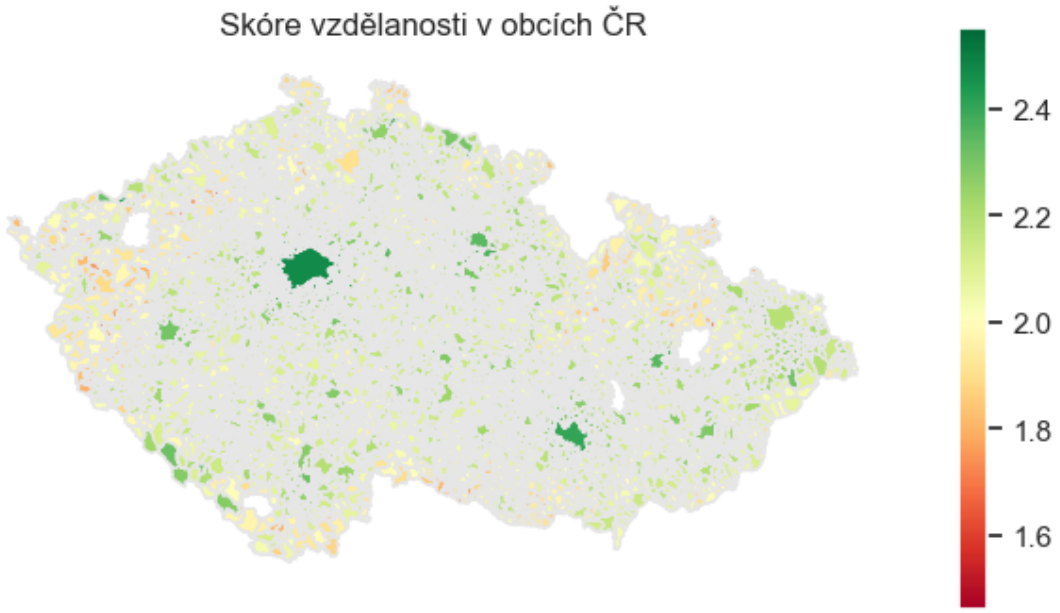
V hypotéze diskutující vztah mezi vzděláním a dluhy jsme se snažili zjistit, zda-li je možné nějak využít informace o počtu vysokoškoláků zastoupených v obci k tomu abychom určili míru její zadluženosti. Výsledky nám ukázaly, že vztah mezi vzděláním a zadlužeností obce existuje (nebo to tak alespoň vypadá). Většina obcí s vysokoškoláky má nižší skóre rizikovosti než obce bez vysokoškoláků. Je nutno ale podoknout jednu poměrně důležitou věc. Úroveň vzdělání (alespoň tzv. "na papíře") v České republice roste. Pokud bychom tedy porovnali standard generace narozené v meziválečném období, velmi pravděpodobně by jako dostačující míru dosaženého vzdělenání jejich doby považovali základní školu. Naopak dnes podobným standardem může být dokončení střední školy s maturitou. Za tímto účelem byl vytvořený *vzdelani_skore* sloupec, který je vypočítáván jako

$$vzdelani\_skore = \frac{\sum_{\text{koeficient vzdělání}} \text{koeficient vzdělání} * hodnota(\text{úroveň vzdělání})}{\sum_{\text{koeficient vzdělání}} \text{koeficient vzdělání}}$$

. Kde *koeficient vzdělání* definujeme jako: - 1 pro základní vzdělání - 2 pro střední vzdělání bez maturity - 2.5 pro střední vzdělání s maturitou - 2.75 pro vyšší odborné vzdělání - 3 pro vysokoškolské vzdělání

Toto nám umožní rozumně dobře srovnat na podobnou úroveň středoškolské bez maturity, s maturitou a vyšší odborné vzdělání a vyjádřit trochu detailněji vzdělanost obce a nezaměřovat se pouze na vysokoškoláky.

Tímto způsobem jsme vytvořili sloupec, který nám ukazuje, jak moc je vzdělání obce vyšší než standardní vzdělání generace narozené v meziválečném období. Pokud bychom tedy vytvořili mapu ČR, kde by bylo vidět, jak je vzdělání obce vyšší než standardní vzdělání generace narozené v meziválečném období, mohli bychom zjistit, zda-li je vztah mezi vzděláním a zadlužeností obce opravdu existovat mohl.



Nyní můžeme na mapě výše vidět, jak se nám vzdělanost lépe rozprostřela na oblasti s převažujícím základním vzděláním a oblastím jejichž průměr se blíží středoškolskému vzdělání s maturitou.

```
ROC AUC by cval:  [0.4447384  0.39692716 0.45426124 0.46667524 0.44861826]
ROC AUC by test:  0.441499996477385
```

4.1 Model

Nyní se můžeme vrátit k modelování. Vytvoříme si tedy model, který na základě vstupních proměnných o obci bude schopen předpovídat její skóre rizikovosti. V úvahu vezmeme následující proměnné: - *vzdelani_skore* - skóre vzdělání obce - *pocet_obyvatel* - počet obyvatel obce - *rozvedeni* - míru zastoupení rozvedených obyvatel - *dospeli* - míru zastoupení dospělých obyvatel

Budeme postupovat podle následujícího postupu:

1. vybereme pouze sloupce s proměnnými, které chceme použít

```
df_model = map_df[
    [
        'vzdelani_score',
        'oD.cat',
        'score',
        'rod_stav_rozv_mean',
        'dospeli',
        'kod_obec',
        'nazev_obec'
    ]
]
```

2. přejmenujeme sloupce

```
df_model.rename(columns={
    'oD.cat': 'pocet_obyvatel',
    'rod_stav_rozv_mean': 'rozvedeni',
}, inplace=True)
```

3. One-hot encoding pro kategorické proměnné

```
df_model['pocet_obyvatel_orig'] = df_model['pocet_obyvatel']
df_model = pd.get_dummies(df_model, columns=['pocet_obyvatel'], drop_first=True)
```

4. Rozdělení dat na trénovací a testovací

```
from sklearn.model_selection import train_test_split
# splitting the data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)
```

5. Modelování

```
X = df_model[[
    'pocet_obyvatel_100-200',
    'pocet_obyvatel_200-500',
    'pocet_obyvatel_500-1k',
    'pocet_obyvatel_1-2k',
    'pocet_obyvatel_2-5k',
    'pocet_obyvatel_5-20k',
    'pocet_obyvatel_20-50k',
    'pocet_obyvatel_50-200k',
    'pocet_obyvatel_200k+',
    'rozvedeni',
    'dospeli',
    'vzdelani_score'
]]
y = df_model['score']

# create model and fit
LR = LinearRegression()
modelA = LR.fit(X_train, y_train)
```

4.2 Vyhodnocení výkonu modelu

K vyhodnocení výkonu modelu použijeme cross-validaci a porovnáme výsledky s testovacím datasetem.

```
from sklearn.model_selection import cross_val_score
from sklearn.metrics import r2_score
# cross-validace
scores = cross_val_score(LR, X_train, y_train, scoring='r2', cv=5)
print('ROC AUC by cval: ', scores)

# výsledky na testovacím datasetu
y_prediction =  modelA.predict(X_test)
```

```
from sklearn.metrics import r2_score
print('ROC AUC by test: ', r2_score(y_test, y_prediction))
```

Výsledek ROC AUC na testovacím datasetu je zhruba 0.44.

5 Závěr

Z dat se nám povedlo vyextrahovat širokou škálu zajímavých informací. Podívali jsme se na vztah mezi poměrem lidí s vystudovanou vysokou školou a skóre (finanční) rizikovosti obce. Nicméně, dokončené vysokoškolské studium jsme zobecnili na parametr ukazující průměrnou úroveň vzdělání v dané obci. Všimli jsme si, že nějaký vztah mezi zmíněnými proměnnými by existovat mohl. Dále jsme se podívali na podobný vztah i v kontextu rozvodovosti, která by mohla též fungovat jako relativně spolehlivý ukazatel určitého sociodemografického trendu. Nicméně, ukázalo se i jako rozumné zahrnout do úvahy poměr věkových skupin obyvatel obce.

Následně jsme se pustili na modelování. Vytvořili jsme model, který na základě vstupních proměnných o obci bude schopen předpovídat její skóre rizikovosti. Výsledky modelu jsou však poměrně slabé (0.44) a je třeba se zamyslet nad dalšími možnostmi zlepšení. Jednou z možností je zahrnout do modelu i další proměnné, které by mohly být pro predikci skóre rizikovosti obce relevantní. Další možností by mohlo být využití více sofistikovaných modelovacích techniky. Nicméně, s těmi autor tohoto textu má nulového znalosti (i z teoretického hlediska) a i takto je to pro něj dostatečně zajímavý projekt, protože se jednalo o první takovýto pokus i s lineární regresí.