Fakulta matematiky, fyziky a informatiky Univerzity Komenského, Bratislava

Projekt z manažmentu dát

Analýza demokratického indexu vo vzťahu k rôznym štatistikám štátov

Andrej Špitalský, 2DAV

Obsah

0 Predstavenie témy

V mojom projekte sa budem venovať tzv. demokratickému indexu spoločnosti *Economist Group*. Jedná sa o každoročnú štatistiku o štátoch, ktorá zhodnocuje napríklad volebný proces, politickú kultúru či občianske práva v 167 krajinách. Výsledný údaj je číslo na škále od 0 do 10, kde 10 je najlepší možný výsledok. Pozrieme sa na to, či vieme nájsť koreláciu medzi inými štatistikami o krajinách s ich demokratickým indexom medzi rokmi 2010 až 2023.

1 Dáta

1.1 Predstavenie dát

Budeme hľadať koreláciu demokratického indexu so 6 rôznymi indikátormi získanými z portálu World Bank – data. V projekte sa často budú vyskytovať nasledovné skratky.

EDU percento žiakov, ktoré po prvostupňovom vzdelaní pokračujú na druhostupňové

GDP hrubý domáci produkt na obyvateľa v amerických dolároch

GNI Gini index – označuje nerovnomernosť v rozložení príjmov medzi obyvateľmi, hodnota 0 značí perfektnú rovnomernosť, hodnota 100 značí perfektnú nerovnomernosť (jeden obyvateľ zarába, zvyšní nie)

LEX očakávaná stredná dĺžka života v rokoch v momente narodenia

MIE percento z HDP investovaného do armády a zbrojenia

RDE percento z HDP investovaného do vedy a výskumu

Pre demokratický index budeme používať skratku **DCI**.
V tejto tabuľke je uvedených niekoľko jednoduchých štatistík o získaných dátach.

Skratka	**Priemer**	**Štandardná odchýlka**	**Minimum**	**Maximum**
DCI	5.427	2.232	0.26	9.93
EDU	91.907	10.402	52.482	100
GDP	14470.2	20801.8	216.83	133711.8
GNI	40.599	11.145	23.2	63.4
LEX	71.408	8.141	45.596	85.533
MIE	1.975	1.899	0.035	33.547
RDE	0.847	0.858	0.01	5.706

Zdroje a pôvodné metadáta sú v súboroch `metadata_world_bank_indicators.csv` a `metadata_democracy_index.csv`.

1.2 Čistenie dát

Chýbajúce hodnoty v dátach boli nahradené nasledovnou logikou:

- ak daný štát mal pre daný indikátor aspoň jednu nechýbajúcu hodnotu, chýbajúca hodnota bola nahradená hodnotou indikátora v roku čo najbližšie k chýbajúcej hodnote
- inak, chýbajúca hodnota bola nahradená priemerom hodnôt indikátora všetkých ostatných krajín v daný rok (jedine pre Gini index boli nahradené maximom, keďže v dátach bolo veľmi veľa chýbajúcich hodnôt a po preštudovaní to boli zväčša krajiny, ktoré mali podľa iných dát tento index vyšší)

2 Analýza dát

2.1 Lineárna regresia

Jedným z klasických spôsobov, ako zistiť, či a ako dáta závisia od iných dát, je lineárna regresia. Vieme pomocou nej zistiť, či sa dá demokratický index aproximovať pomocou lineárnej kombinácie zvyšných 6 indikátorov. Následne môžeme zistiť „dobrosť“ takejto aproximácie (tzv. R^2 koeficient, hodnota bežne medzi 0 a 1, čím bližšie k 1, tým je aproximácia považovaná za lepšiu, lebo zachytáva viac variability závislej premennej) a či od jednotlivých indikátorov závisí demokratický index pozitívne alebo negatívne.

R^2	**EDU**	**GDP**	**GNI**	**LEX**	**MIE**	**RDE**
0.4995	0.003	0.000022	-0.02	0.103	-0.33	0.346

Podľa hodnoty R^2 koeficientu vidíme, že lineárna aproximácia demokratického indexu zachytáva 50% jeho variability. Takisto vidíme pozitívnu koreláciu medzi stúpajúcim indikátorom pre školstvo, hrubý domáci produktu, strednú dĺžku života a výdavkov na vedu a výskum (so stúpajúcou hodnotou týchto indikátorov stúpa aj demokratický index). Negatívnu koreláciu vidíme pri výdavkoch na armádu a zbrojenie a pri Gini indexe (so stúpajúcou hodnotou týchto indikátorov klesá demokratický index).

Z týchto dát však nevieme určiť, že aký podiel vo výslednej aproximácii demokratického indexu jednotlivé indikátory zohrávajú. Znormalizujeme preto dáta nasledovne – od každej hodnoty v jednotlivých stĺpcoch (vektor dát pre jeden indikátor) odpočítame priemer dát v tomto stĺpci a predelíme ich štandardnou odchýlkou. Týmto získame také dáta, že priemer každého indikátora je v 0 a jeho štandardná odchýlka je 1. Teraz už koeficienty lineárnej regresie budú relevantné aj svojou hodnotou, nie iba znamienkom.

R^2	**EDU**	**GDP**	**GNI**	**LEX**	**MIE**	**RDE**
0.4995	0.013	0.204	-0.1	0.376	-0.28	0.133

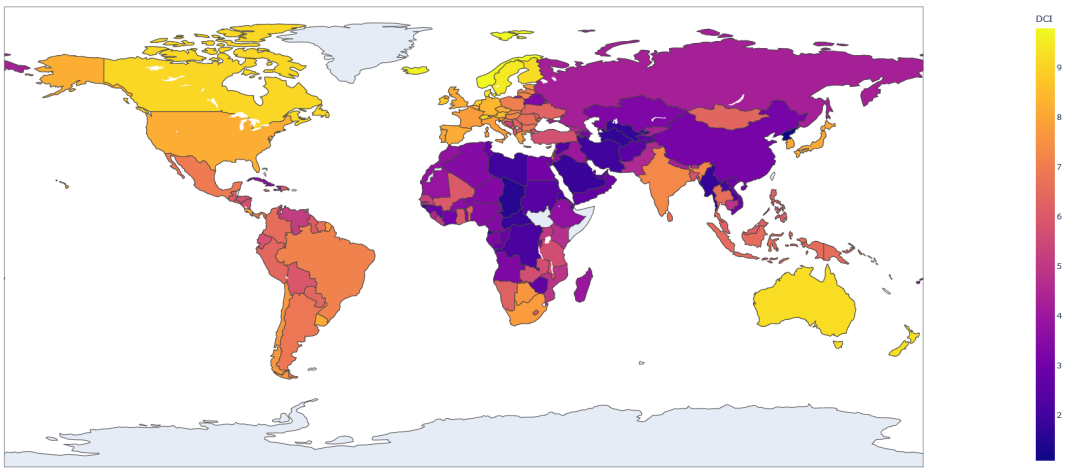
Z týchto dát už môžeme vyčítať, že pri aproximácii demokratického indexu pomocou lineárnej regresie má najväčší pozitívny vplyv stredná dĺžka života a najväčší negatívny vplyv investovanie do armády a zbrojenia.

2.2 Demokratický index podľa geografickej polohy

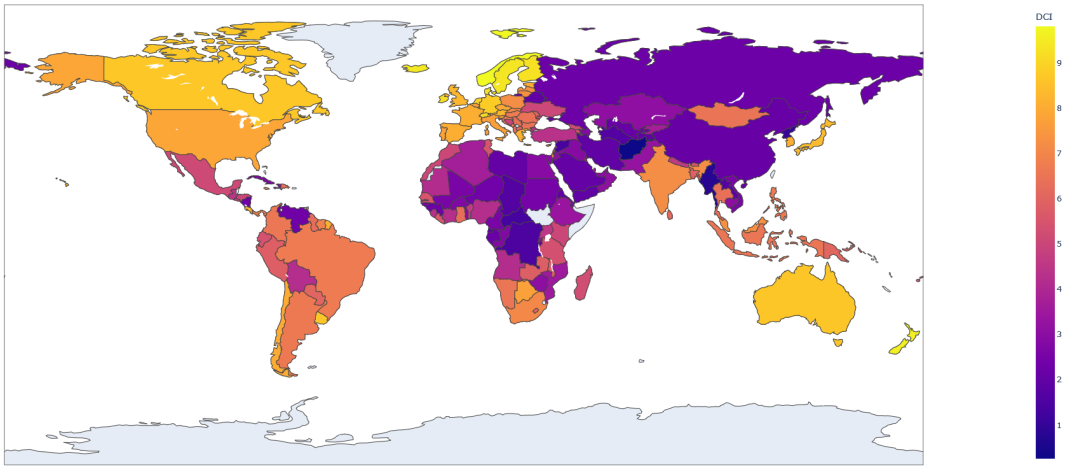
Vykreslíme mapu sveta tak, že jednotlivé krajiny budú zafarbené podľa hodnoty ich demokratického indexu (modrá farba značí hodnotu 0, žltá 10). Spravíme tak pre roky 2010 a 2023.

Môžeme si všimnúť, že vysoké hodnoty dedmokratického indexu nájdeme pre oba roky hlavne v Škandinávii, Kanade a Austrálii. Naopak nízke hodnoty sú konzistentne v centrálnej a severnej Afrike a na Blízkom východe. Mierny regionálny prepad medzi rokmi 2010 a 2023 môžeme vidieť pre strednú Ameriku. V Európe, na západe vidíme skôr mierne stúpajúci trend a na východe mierne klesajúci.

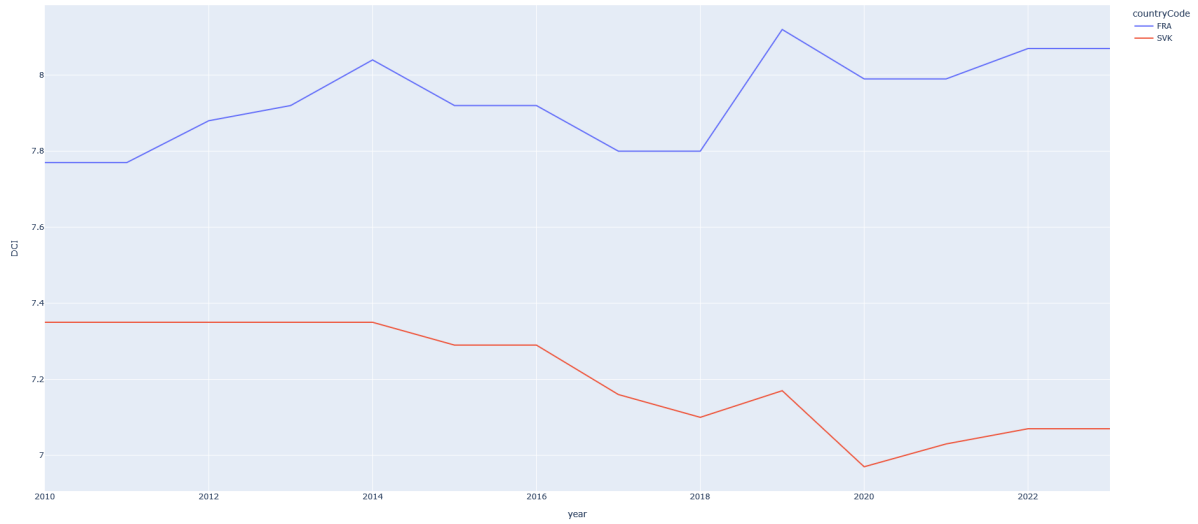
Democratic index of countries in 2010

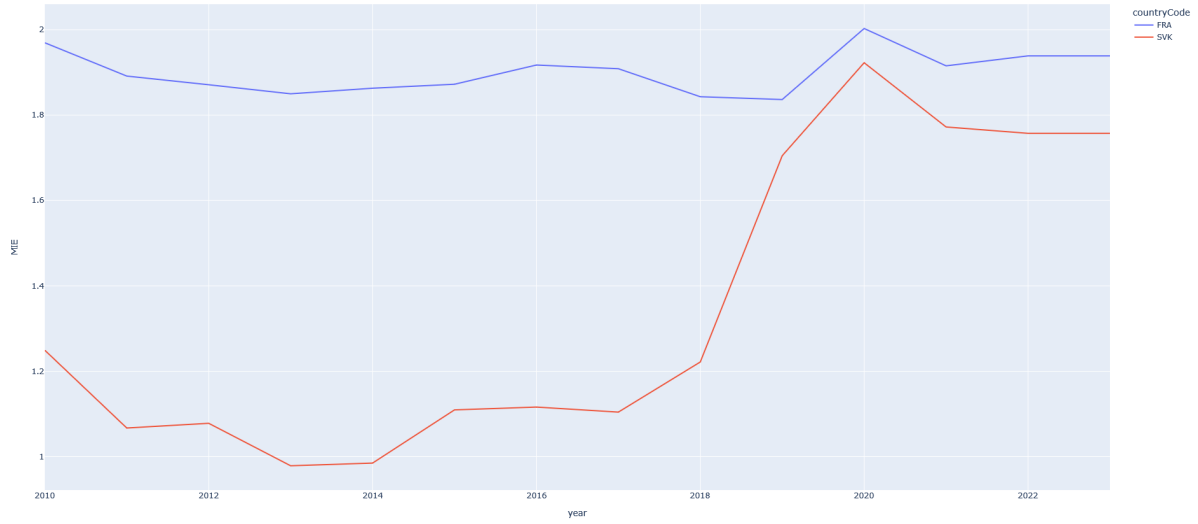


Democratic index of countries in 2023



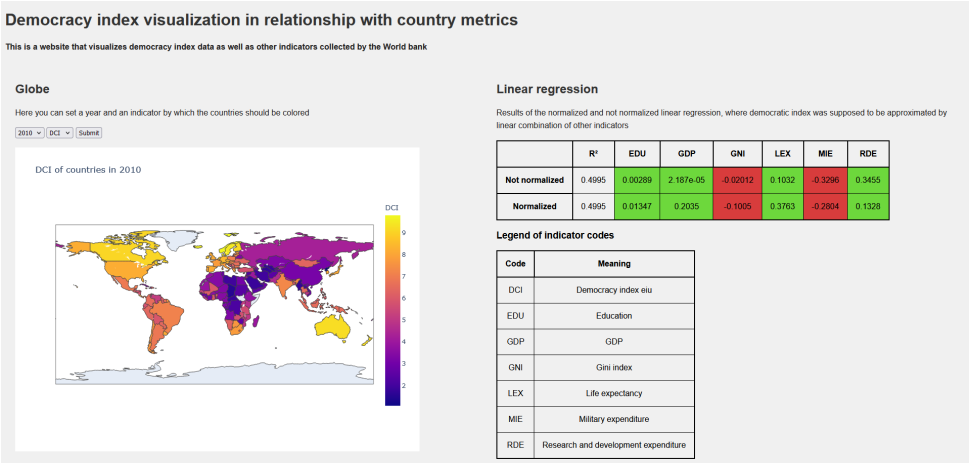
Skúsime porovnať dáta pre Slovenskú republiku a Francúzsko a či sa naše výsledky zhodujú s reálnymi trendmi v dátach. Zhodu vidno napríklad pri investíciách do armády a zbrojenia. Vidíme, že výdavky Slovenskej republiky sa zvýšili, zatiaľ čo výdavky Francúzske viacmenej stagnovali. Tento trend súhlasí s výsledkami lineárnej regresie, ktorá implikovala negatívnu koreláciu medzi týmito dvoma indikátormi. Jedná sa však o špecificky vybraný príklad v dátach, z ktorého nemožno robiť väčšie závery.

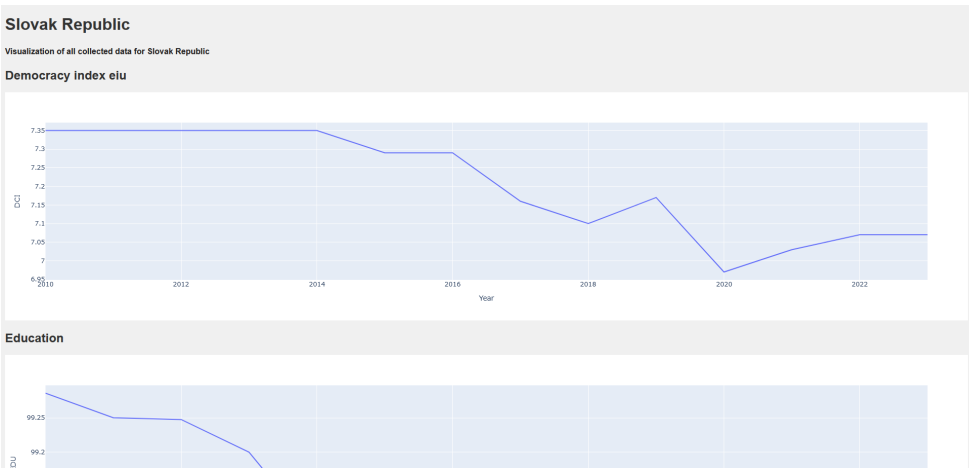




3 Flask aplikácia

Pre lepšiu vizualizáciu dát a vzťahov medzi nimi bola vytvorená aplikácia vo `Flask`-u. Je možné si rozkliknúť napríklad jednotlivé krajiny a vidieť priebeh jednotlivých indikátorov medzi rokmi 2010 až 2023. Takisto je k dispozícii interaktívna mapa, kde sa dá zvoliť rok a indikátor, podľa ktorého sa krajiny zafarbia. Nakoniec, naprogramovaná je aj samostatná stránka pre interaktívne PCA vizualizácie, kde sa dajú zvoliť indikátory (mimo demokratického indexu) a rok a následne pre tieto dáta sa spočíta PCA. Vo výslednej vizualizácii sú jednotlivé body zafarbené podľa ich demokratického indexu a dá sa pozorovať, aké kombinácie stĺpcov majú za následok napríklad formovanie zhlukov podobnej farby či nejaký gradient.





4 Zhrnutie a diskusia

V mojom projekte som sa venoval demokratickému indexu a jeho korelovaniu so štatistikami ako stredná dĺžka života a Gini index. Analyzovali sme výsledky lineárnej regresie a overili jej korektnosť na jednoduchom príklade. Pre lepšiu vizualizáciu dát som naprogramoval `Flask` aplikáciu, pomocou ktorej sa dajú prehľadne vizualizovať získané dáta a hľadať ďalšie vzťahy medzi nimi.

V retrospektíve by som najmä zmenil prístup k čisteniu a ukladaniu dát. Myslím si, že by nebolo zložité vytvoriť štruktúru, ktorá by bola výrazne viac flexibilná v zmysle, že by som vedel jednoducho prídavať dátové súbory z portálu World bank a zvyšok kódu by na to zareagoval. Momentálne sú bohužial použité indikátory natvrdo uložené v programe. Takisto by kódu pomohlo rozdelenie do viacerých tried pre prehľadnosť a tiež lepšie zadefinovaná komunikácia medzi triedami.

Takisto si myslím, že môj proces čistenia dát vie v niektorých prípadoch dosť meniť trendy v dátach. Myslím si, že by si to zaslúžilo revíziu. Takisto by som mohol v budúcnosti pridať spracovanie regiónov, keďže to by pomohlo vizualizovať trendy napríklad v Európe.

Na druhej strane ma prekvapilo, aká priamočiara je komunikácia webstránky a kódu pomocou `Flask`-u. Projekt ma naučil spracovávať formuláre, vybudoval vo mne rešpekt pred čistením dát a zorientoval som sa v komunikácii medzi knižnicami `pandas`, `numpy`, `scikit-learn`, `sqlite` a `plotly`.