České vysoké učení technické v Praze

Fakulta elektrotechnická

Závěrečný úkol z předmětu

B4M36SAN -- Statistical Data Analysis

Globální analýza socioekonomických a zdravotních indikátorů a jejich vliv na úmrtnost novorozenců

Autoři: Svitlana Hrynchenko, Ivana Klikarová, Klára Zinková, Marina Radić

Praha 2024

1. **Výzkumná otázka:**

"Jaký je vztah mezi úmrtností novorozenců a vybranými sociodemografickými a zdravotními faktory v různých zemích ve vybraném období?"

Výzkum se zaměřuje na analýzu socioekonomických indikátorů a indikátorů, které souvisí se zdravotní péčí v různých zemích s cílem identifikovat vztahy mezi nimi a zjistit jejich vliv na úmrtnost novorozenců (věk 0 - 11 měsíců).

2. **Data:**

Zdroje jednotlivých datasetů a všech indikátorů jsou uvedeny v přiložené excelovém souboru Zdroje_datasety_faktory.xlsx.

3. **Postup práce:**

- **Exploratory Data Analysis(EDA):** aplikovaná popisná analýza dat, abychom pochopili strukturu dat, identifikovali chybějící hodnoty a sledovali distribuce různých proměnných.

- **Čištění dat a příprava:** Sjednocení různých datasetů (provedeno přímo v excelu) a předem vybraných faktorů, které mohou mohou mít vliv při řešení výzkumné otázky. Dále zpracování chybějících hodnot a duplicitních řádků, aby byla zajištěna kvalita dat a jejich možná analýza.

- **Statistická analýza:** provedení statistické analýzy, například výpočet korelačních matic, abychom pochopili vztahy mezi různými proměnnými.

- **Statistické Modelování**: provedení regresní analýzy (lineární a polynomiální), GLM Poissonova modelu a binární klasifikace (logistická regrese, LDA, QDA) k určení vztahu mezi socioekonomickými faktory a mírou úmrtnosti novorozenců.

- **Metody Shlukování:** využití metod jako K-Means, Hierarchical Clustering, DBSCAN a Gaussian Mixture Models k segmentaci zemí na základě různých indikátorů a rozdělení zemí s podobnými faktory ovlivňující úmrtnost novorozenců.

- **Redukce Dimenze:** použití metod jako PCA, multidimenzionální škálování, Isomap, t-SNE a Locally Linear Embedding k redukci dimenzí a vizualizaci vztahů mezi zeměmi a porovnání výsledků s metodami shlukování.

- **Vizualizace:** k vizuálnímu znázornění dat a zjištění používat grafy, grafy rozptylu, krabicové grafy a choropletové mapy.

- **Interpretace a vyhodnocení:** každý krok analýzy bude doprovázen interpretací a závěry, které zdůrazní klíčová zjištění a jejich důsledky pro míru úmrtnosti novorozenců. Používáme metriky, jako silhouette score, Calinski-Harabasz index, and Davies-Bouldin index pro ověření shluků.

I. Počáteční hodnocení a čištění dat.

Datová sada, se kterou pracujeme, zahrnuje řadu socioekonomických proměnných a proměnných souvisejících se zdravotní péčí v různých zemích. Abychom mohli zahájit analýzu, musíme nejprve načíst a vyhodnotit datovou sadu. Zkontrolujeme jak jsou počáteční data uspořádaná v řádcích a sloupcích a zjistíme, kde se nachází chybějící hodnoty a duplicity. Toto počáteční vyhodnocení bude sloužit jako podklad pro naše další kroky při čištění a analýze dat. Začněme načtením a prozkoumáním datové sady.

a) Počáteční prozkoumání a vyhodnocení datové sady

Datová sada obsahuje různé socioekonomické ukazatele a ukazatele související se zdravotní péčí pro různé země. Zde je jejich přehled:

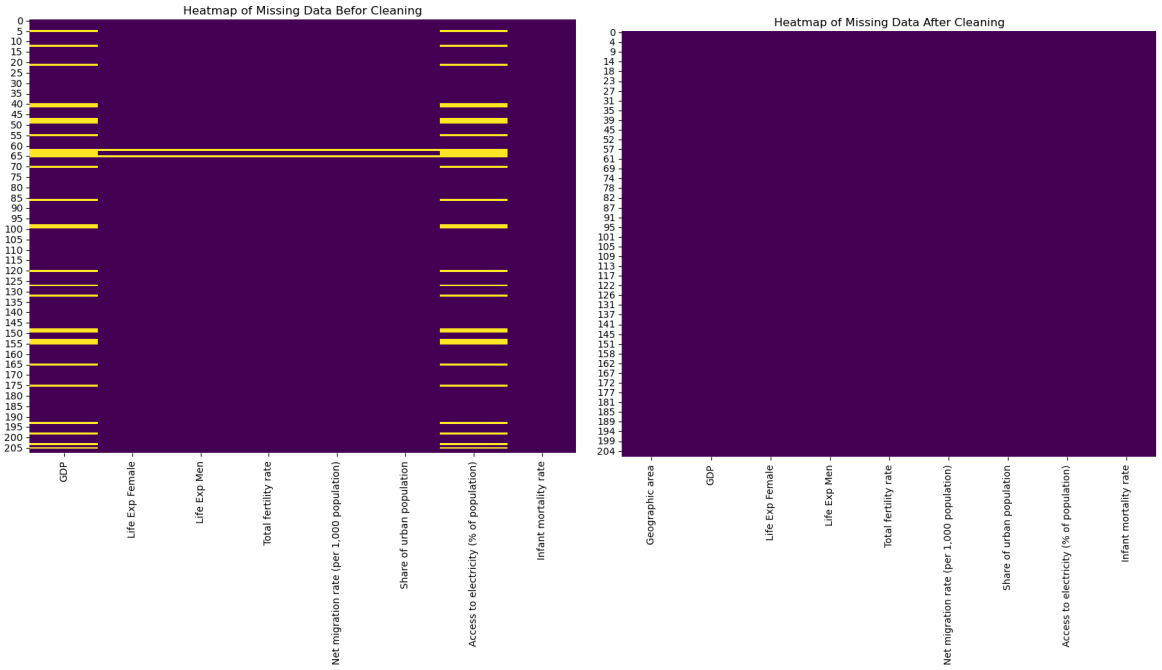
Geographic Area	Názvy zemí (např. Afghánistán, Albánie).
GDP	Hodnoty hrubého domácího produktu.
Life Expectancy	Délka života: oddělené údaje pro ženy a muže.
Total Fertility Rate	Průměrný počet dětí narozených během života jedné ženy
Net Migration Rate	Rozdíl mezi počtem osob, které do země během roku přišli a které ji opustili, na 1 000 osob.
Share of Urban Population	Podíl obyvatelstva žijícího ve městech.
Access to Electricity	Podíl obyvatelstva s přístupem k elektřině.
Infant Mortality Rate	Počet úmrtí narozených dětí do jednoho roku na 1000 živě narozených dětí

b) Shrnutí čištění dat

Proces čištění dat vedl k následujícím úpravám datového souboru:

- Odstraněné řádky: Řádky s chybějícími hodnotami byly odstraněny. Duplicitní záznamy v souboru dat byly odstraněny, aby byla zachována integrita dat.

Výsledkem těchto kroků je, že **datový soubor nyní obsahuje 175 řádků a 9 sloupců,** což představuje zmenšení oproti původní velikosti v důsledku odstranění řádků s chybějícími kritickými údaji a duplicit.



Na prvním grafu s názvem "Heatmap of Missing Data Before Cleaning" vidíme několik žlutých vodorovných čar různé délky, které představují chybějící data v souboru. Chybějící údaje jsou zřetelné zejména ve sloupcích GDP a Access to electricity (% of population), ojediněle chybějí i v dalších sloupcích.

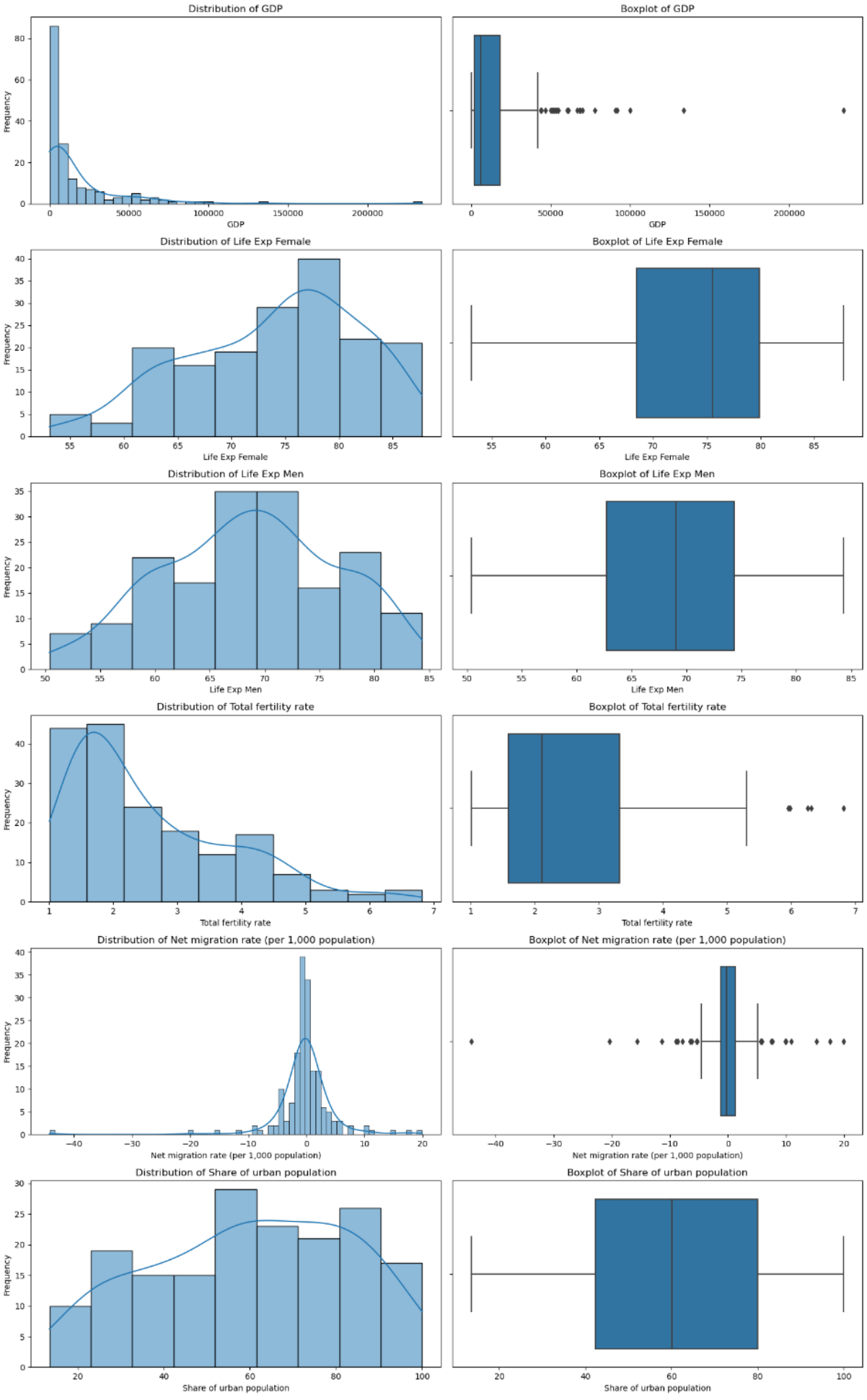
Naproti tomu druhý graf s názvem "Heatmap of Missing Data After Cleaning" ukazuje plnou barvu bez žlutých čar, v souboru nejsou žádné chybějící hodnoty, čištění proběhlo úspěšně.

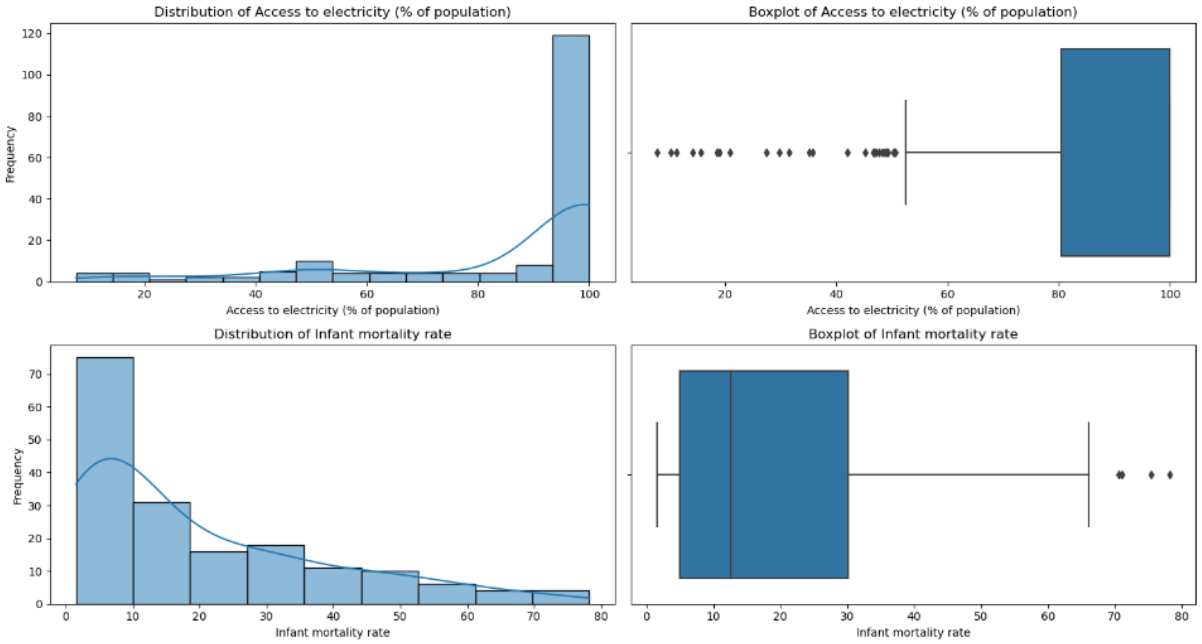
II. EDA.

a) Distribuce jednotlivých indikátorů

K vizualizaci a pochopení rozložení jednotlivých indikátorů, identifikaci outlierů a zkoumání vztahů mezi indikátory, zejména se zaměřením na úmrtnost novorozenců, jsme provedli EDA.

Zobrazená data představují distribuci i boxploty pro každý číselný indikátor v našem souboru dat:


Distribution of GDP
Boxplot of GDP
Frequency
GDP
Distribution of Life Exp Female
Boxplot of Life Exp Female
Life Exp Female
Distribution of Life Exp Men
Boxplot of Life Exp Men
Life Exp Men
Distribution of Total fertility rate
Boxplot of Total fertility rate
Total fertility rate
Distribution of Net migration rate (per 1,000 population)
Boxplot of Net migration rate (per 1,000 population)
Net migration rate (per 1,000 population)
Distribution of Share of urban population
Boxplot of Share of urban population
Share of urban population



- GDP: Tato proměnná vykazuje doprava vychýlené rozdělení, což naznačuje, že velký počet zemí má nižší HDP a méně zemí má výrazně vyšší HDP. Boxplot odhaluje několik outlierů na horním konci.

- Life Expectancy (Female and Male): Oba indikátory průměrné délky života vykazují poněkud normální rozdělení, ale s mírným vychýlením doleva. To naznačuje, že většina zemí má vysokou střední délku života, přičemž méně zemí má výrazně nižší střední délku života.

- Total Fertility Rate: Rozdělení je mírně vychýleno doprava, což naznačuje, že většina zemí má nižší míru plodnosti, s některými výjimkami na horním konci.

- Net Migration Rate: Tento indikátor má různorodé rozdělení s několika outliery na obou koncích, což naznačuje značné rozdíly ve vzorcích migrace v jednotlivých zemích.

- Share of Urban Population: Rozdělení je poměrně rovnoměrné, s mírným vychýlením doleva. Většina zemí má vysoký podíl obyvatel žijících ve městech.

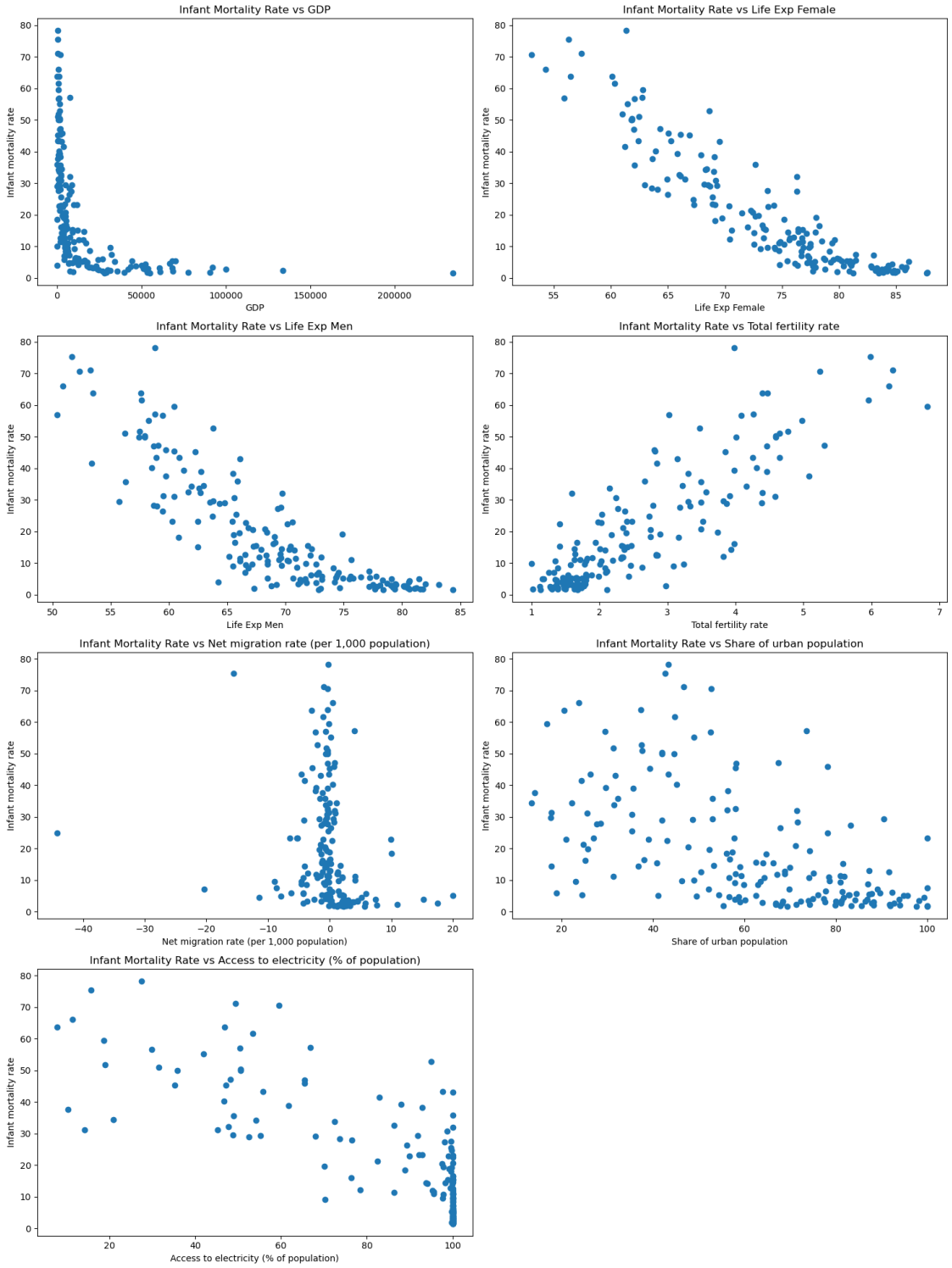
- Access to Electricity: Většina zemí má vysoký přístup k elektřině, což naznačuje rozložení s levým sklonem. Existuje několik zemí s výrazně nižší dostupností.

- Infant Mortality Rate: Tato pro naši analýzu kritická proměnná je pravotočivá, což znamená, že většina zemí má nižší míru kojenecké úmrtnosti, přičemž v některých zemích je tato míra vyšší.

b) Analýza scatter plotů

Z grafů jsou patrné zřetelné trendy mezi úmrtností novorozenců a různými indikátory. Vyšší HDP a lepší přístup k elektřině korelují s nižší kojeneckou úmrtností, což naznačuje, že hospodářský rozvoj a infrastruktura hrají zásadní roli při zlepšování přežití novorozenců. Naopak vyšší porodnosti se zpravidla shoduje s vyšší kojeneckou úmrtností. Střední délka života vykazuje silný negativní vztah s

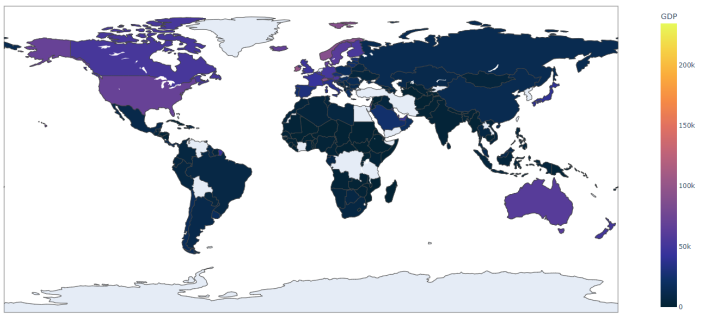
kojeneckou úmrtností, což ukazuje na celkový zdravotní stav jako rozhodující faktor. Tyto poznatky budou podkladem pro budoucí shlukování zemí.



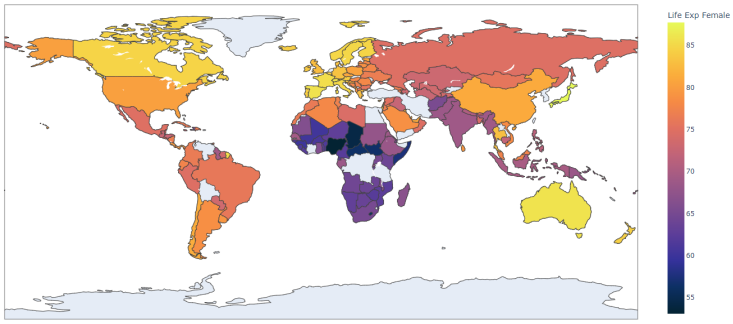
c) Geoprostorová analýza indikátorů

Choropletové mapy vytvořené pomocí funkce **px.choropleth** (knihovna Plotly) znázorňují globální rozdíly v indikátorech. Mapy ukazují vyšší ekonomický a zdravotní stav v rozvinutých regionech a zároveň zdůrazňují problémy v některých částech Afriky a Asie. Tyto vizualizace mají zásadní význam pro určování regionů, které vyžadují cílenou rozvojovou pomoc a zásahy v oblasti zdravotní péče.

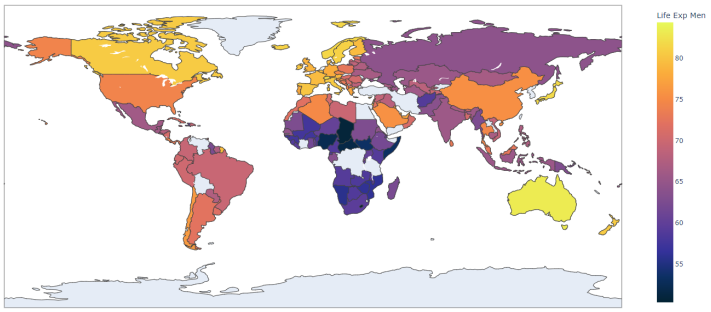
Countries by GDP



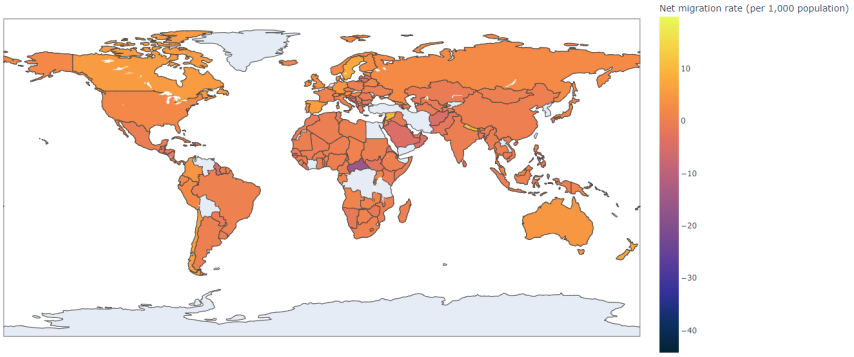
Countries by Life Exp Female



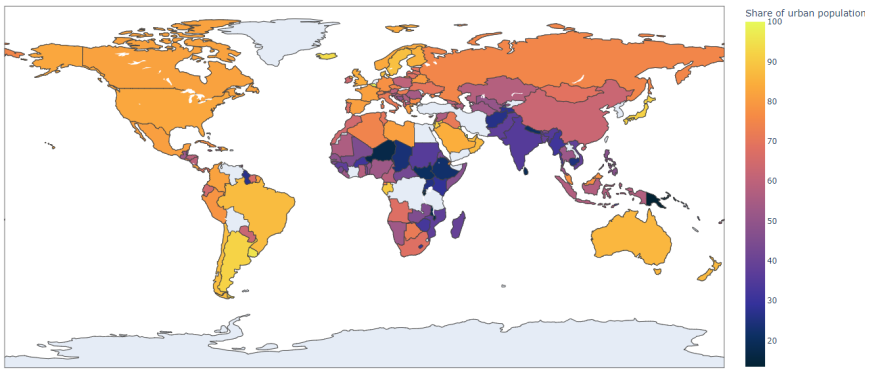
Countries by Life Exp Men

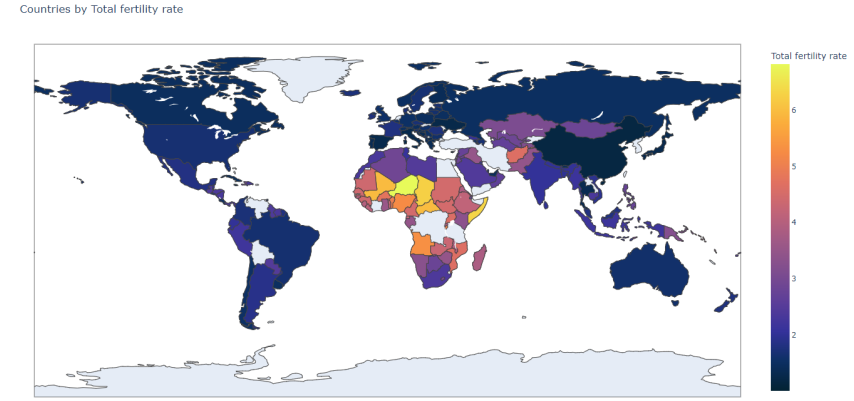


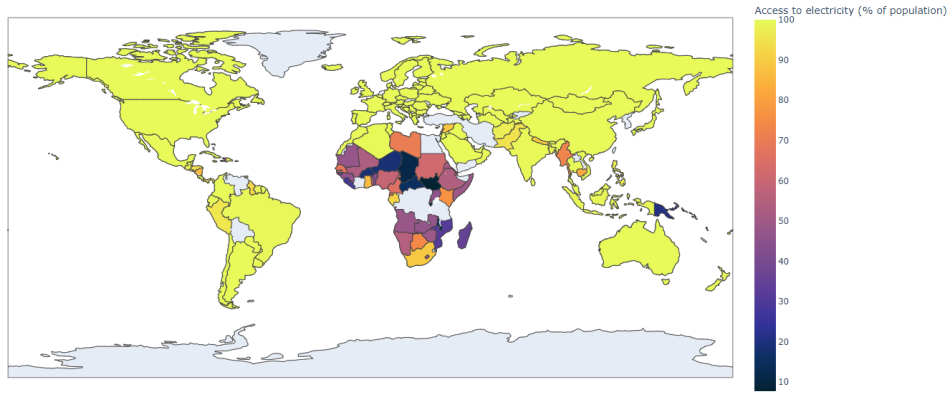
Countries by Net migration rate (per 1,000 population)



Countries by Share of urban population



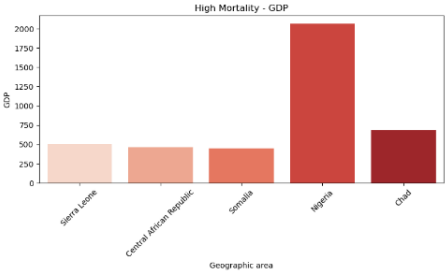


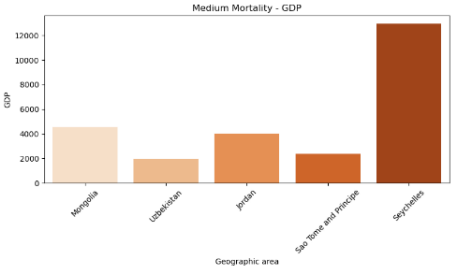


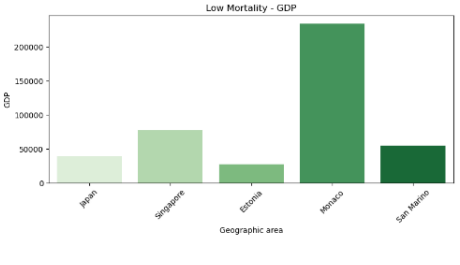
Poslední mapa ukazuje vyšší míru úmrtnosti novorozenců v regionech s nižším HDP a nižší dostupností zdravotnických služeb, zejména v Africe.

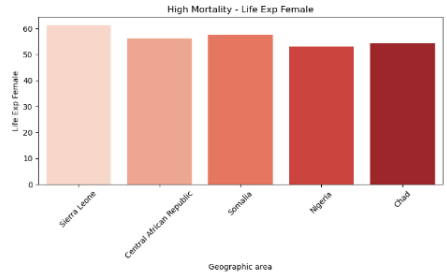
d) Srovnávací analýza vysoké, střední a nízké úmrtnosti novorozenců v různých zemích.

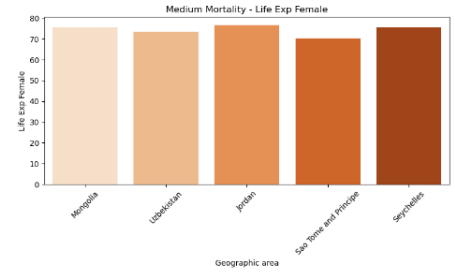
Sloupcové grafy představují srovnávací analýzu vysoké, střední a nízké kojenecké úmrtnosti v různých zemích. Tyto vizualizace, vytvořené pomocí **matplotlib** a **seaborn**, zdůrazňují souvislost mezi mírou úmrtnosti novorozenců a socioekonomickými faktory.

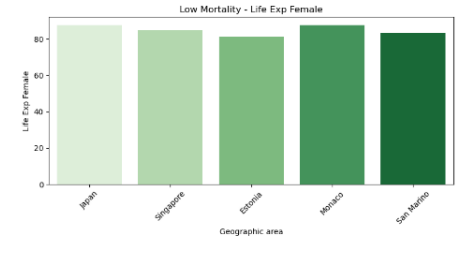

High Mortality - GDP
GDP
Sierra Leone
Central African Republic
Somalia
Nigeria
Chad
Geographic area

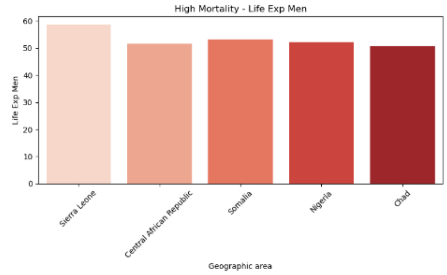

Medium Mortality - GDP
GDP
Mongolia
Uzbekistan
Jordan
Sao Tome and Principe
Seychelles
Geographic area

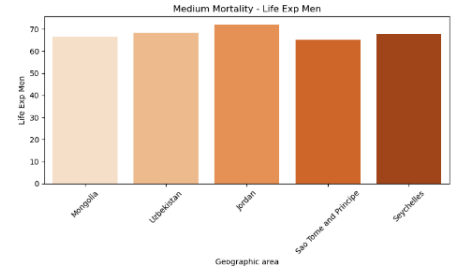

Low Mortality - GDP
GDP
Japan
Singapore
Estonia
Monaco
San Marino
Geographic area

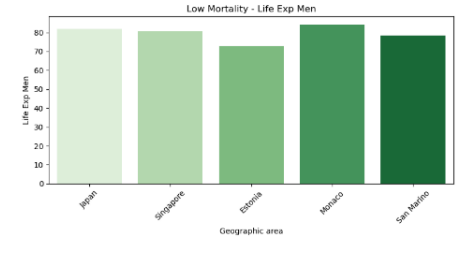

High Mortality - Life Exp Female
Life Exp Female
Sierra Leone
Central African Republic
Somalia
Nigeria
Chad
Geographic area

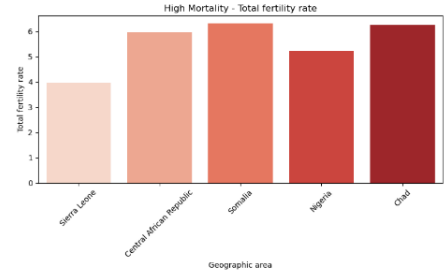

Medium Mortality - Life Exp Female
Life Exp Female
Mongolia
Uzbekistan
Jordan
Sao Tome and Principe
Seychelles
Geographic area

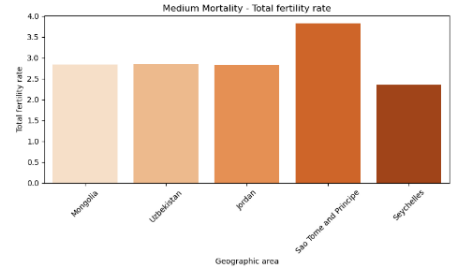

Low Mortality - Life Exp Female
Life Exp Female
Japan
Singapore
Estonia
Monaco
San Marino
Geographic area

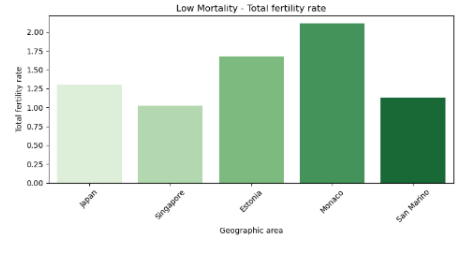

High Mortality - Life Exp Men
Life Exp Men
Sierra Leone
Central African Republic
Somalia
Nigeria
Chad
Geographic area

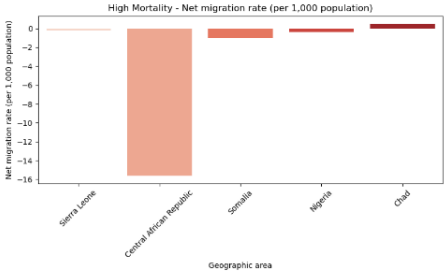

Medium Mortality - Life Exp Men
Life Exp Men
Mongolia
Uzbekistan
Jordan
Sao Tome and Principe
Seychelles
Geographic area

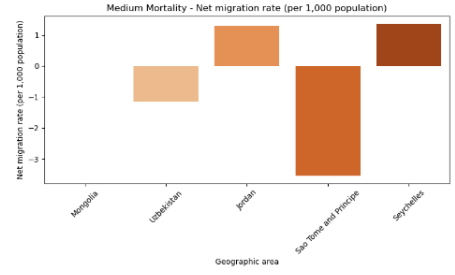

Low Mortality - Life Exp Men
Life Exp Men
Japan
Singapore
Estonia
Monaco
San Marino
Geographic area

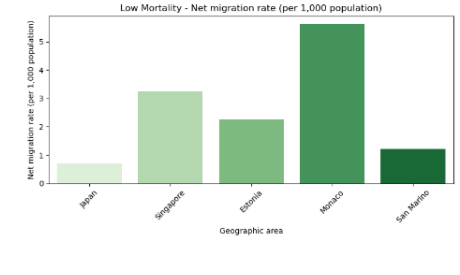

High Mortality - Total fertility rate
Total fertility rate
Sierra Leone
Central African Republic
Somalia
Nigeria
Chad
Geographic area

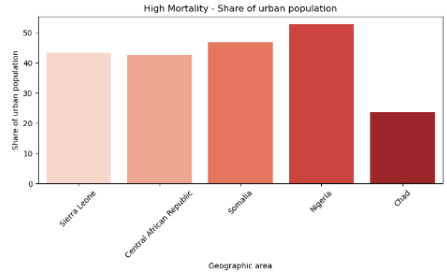

Medium Mortality - Total fertility rate
Total fertility rate
Mongolia
Uzbekistan
Jordan
Sao Tome and Principe
Seychelles
Geographic area

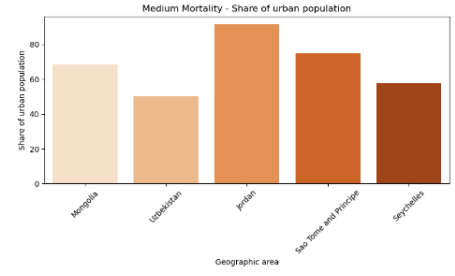

Low Mortality - Total fertility rate
Total fertility rate
Japan
Singapore
Estonia
Monaco
San Marino
Geographic area

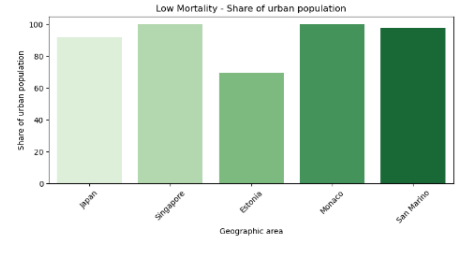

High Mortality - Net migration rate (per 1,000 population)
Net migration rate (per 1,000 population)
Sierra Leone
Central African Republic
Somalia
Nigeria
Chad
Geographic area

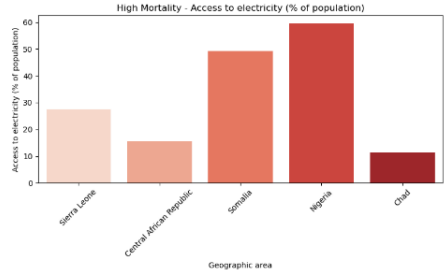

Medium Mortality - Net migration rate (per 1,000 population)
Net migration rate (per 1,000 population)
Mongolia
Uzbekistan
Jordan
Sao Tome and Principe
Seychelles
Geographic area

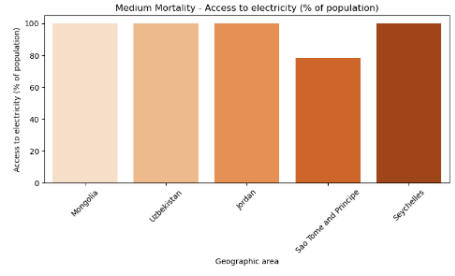

Low Mortality - Net migration rate (per 1,000 population)
Net migration rate (per 1,000 population)
Japan
Singapore
Estonia
Monaco
San Marino
Geographic area

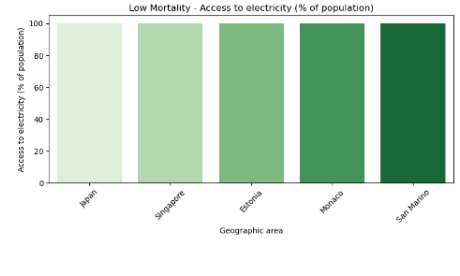

High Mortality - Share of urban population
Share of urban population
Sierra Leone
Central African Republic
Somalia
Nigeria
Chad
Geographic area


Medium Mortality - Share of urban population
Share of urban population
Mongolia
Uzbekistan
Jordan
Sao Tome and Principe
Seychelles
Geographic area


Low Mortality - Share of urban population
Share of urban population
Japan
Singapore
Estonia
Monaco
San Marino
Geographic area


High Mortality - Access to electricity (% of population)
Access to electricity (% of population)
Sierra Leone
Central African Republic
Somalia
Nigeria
Chad
Geographic area

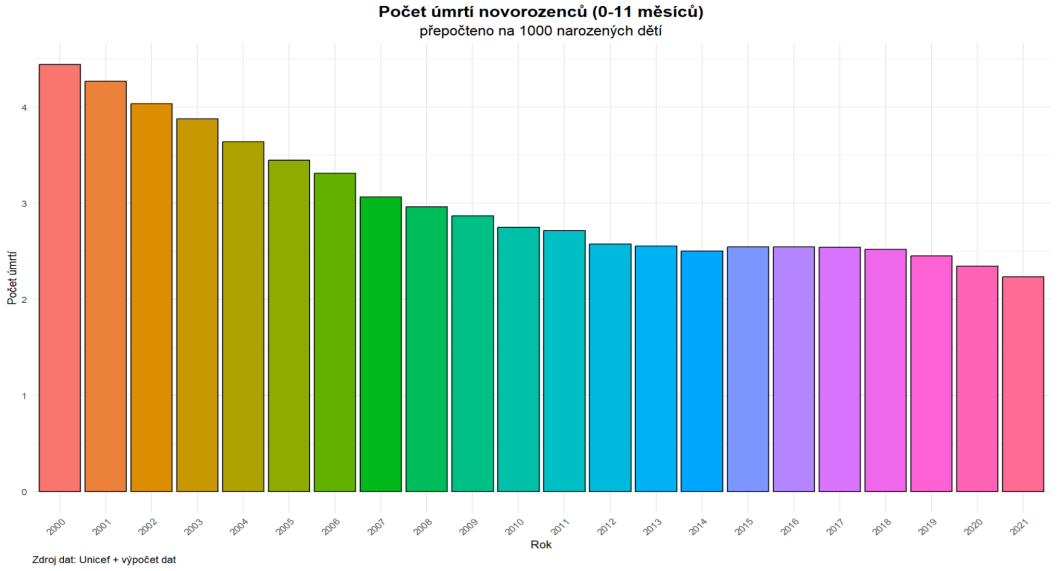

Medium Mortality - Access to electricity (% of population)
Access to electricity (% of population)
Mongolia
Uzbekistan
Jordan
Sao Tome and Principe
Seychelles
Geographic area

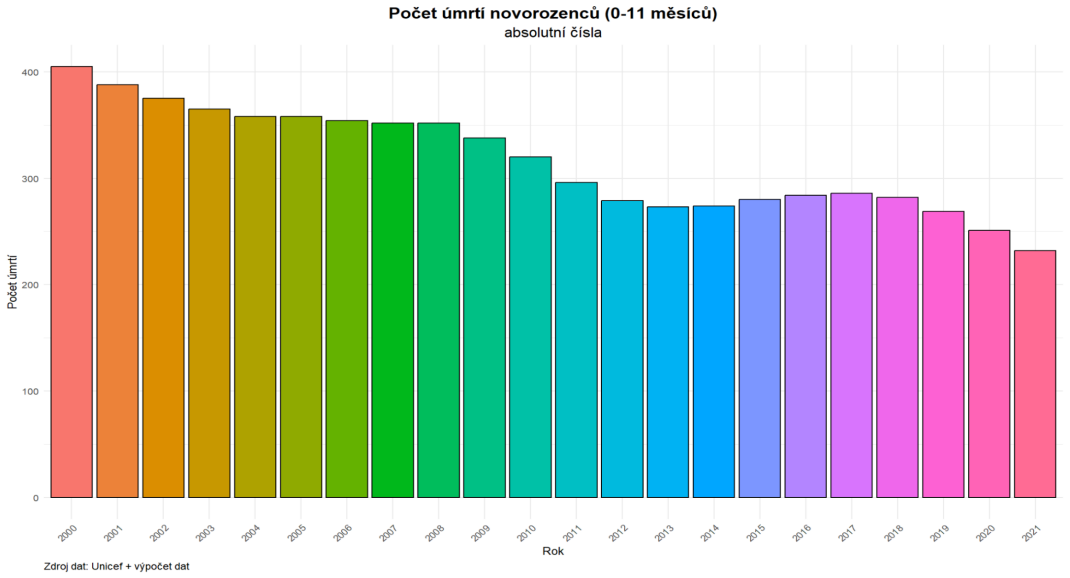

Low Mortality - Access to electricity (% of population)
Access to electricity (% of population)
Japan
Singapore
Estonia
Monaco
San Marino
Geographic area

Česká republika

Navíc pro Českou republiku jsou zde zobrazeny časové grafy zobrazující úmrtnost novorozenců na 1000 narozených dětí a absolutní počty úmrtí novorozenců v letech 2000 - 2021

- tyto časové grafy byly vytvořeny v programu R
- z grafů vidíme, že od roku 2000 je trend v úmrtnosti novorozenců téměř stále klesající. Mezi lety 2015 - 2017 je vidět mírný nárůst, jak v absolutních tak v přepočtených číslech





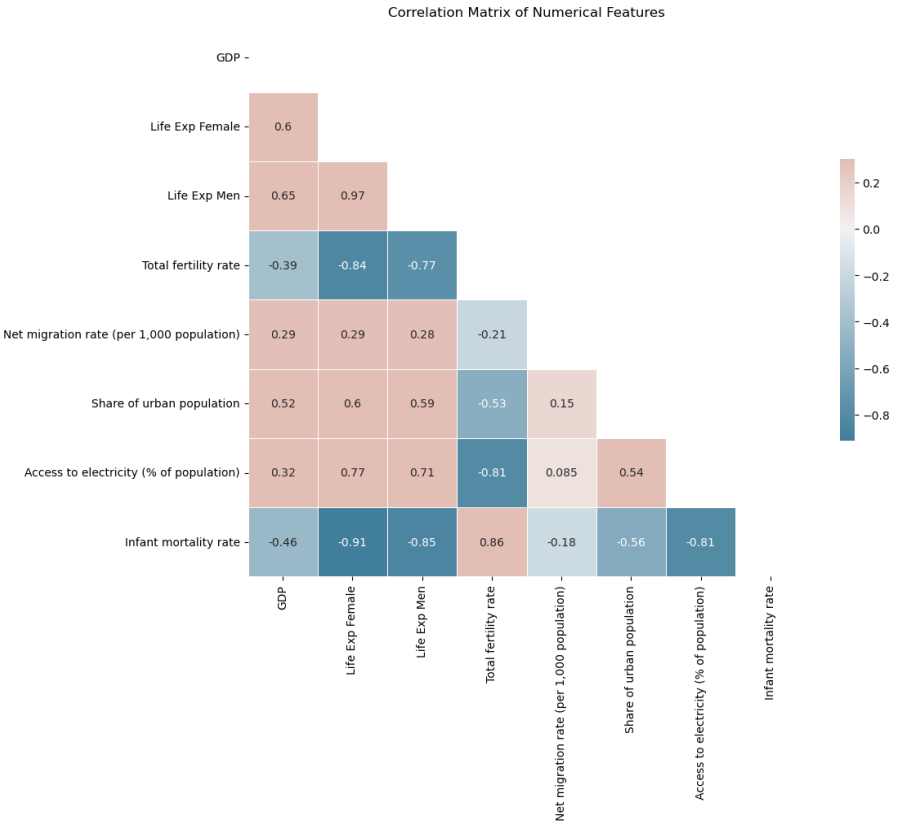
e) Korelační matice pro číselné indikátory

Heatmap,vytvořená pomocí funkce heatmap knihovny **seaborn**, zobrazuje korelační matici pro číselné indikátory, která ilustruje vzájemné vztahy mezi nimi:

- Silná negativní korelace: Úmrtnost novorozenců vykazuje silnou negativní korelaci s HDP, střední délkou života (žen i mužů) a přístupem k elektřině, což naznačuje, že vyšší hodnoty těchto ukazatelů jsou spojeny s lepší mírou přežití novorozenců.

- Silná pozitivní korelace: Očekávaná délka života žen a mužů je silně korelovaná, tedy očekávaná délka života mají podobný dopad na obě pohlaví.

- Silná negativní korelace mezi mírou porodnosti a střední délkou života a pozitivní korelace mezi mírou porodnosti a úmrtností novorozenců naznačuje, že vyšší porodnost může být spojena s nižší střední délkou života a vyšší úmrtností novorozenců.

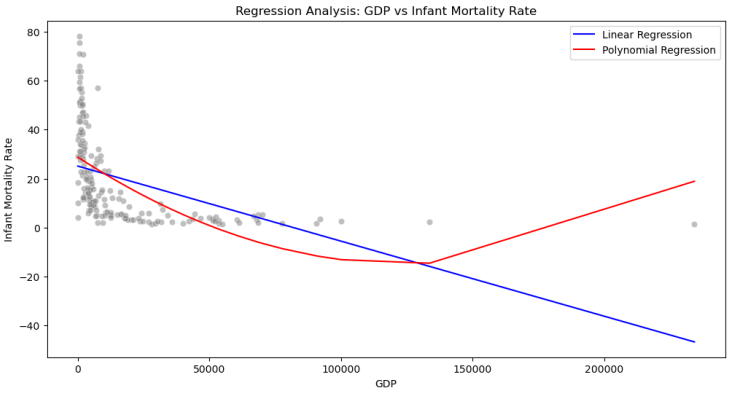


SHRNUTÍ VÝSLEDKŮ EDA ANALÝZY :

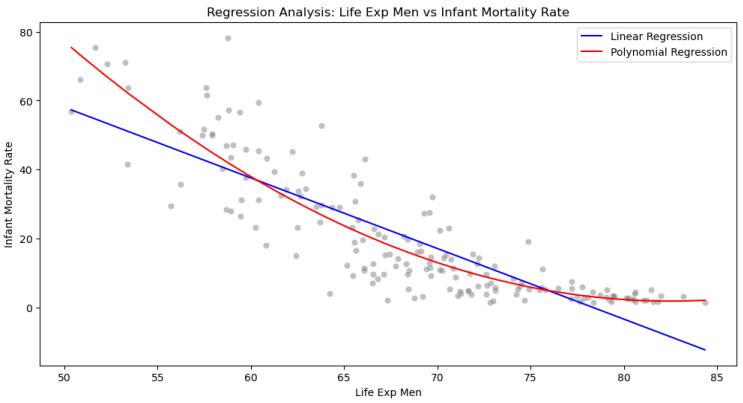
Vyšší HDP, střední délka života a přístup k elektřině jsou silně spojeny s nižší úmrtností novorozenců. Vysoká porodnost koreluje se zvýšenou kojeneckou úmrtností, což pravděpodobně ukazuje na zatížení zdravotní péče. Choropletové mapy a sloupcové grafy zobrazily výrazné regionální rozdíly, přičemž rozvinuté země vykazují příznivé zdravotní a ekonomické ukazatele, zatímco mnohé rozvojové země zaostávají. Scatter ploty tyto vzorce potvrdily a naznačily, že cílená zdravotní a ekonomická politika by mohla přispět k lepšímu zdravotnímu stavu obyvatelstva a nižší úmrtnosti novorozenců.

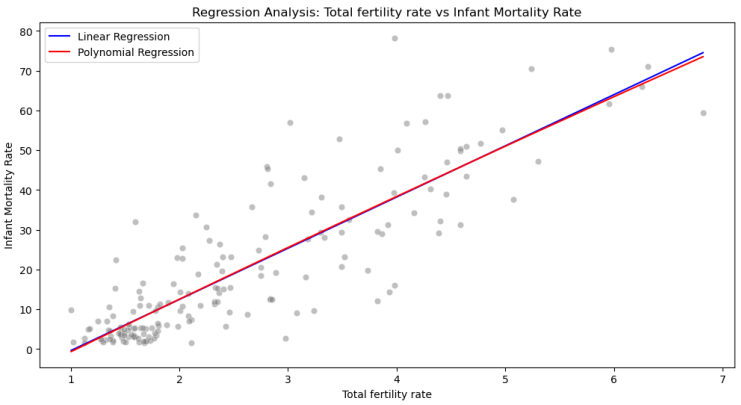
III. Regresní analýza: Lineární vs. Polynomiální regrese

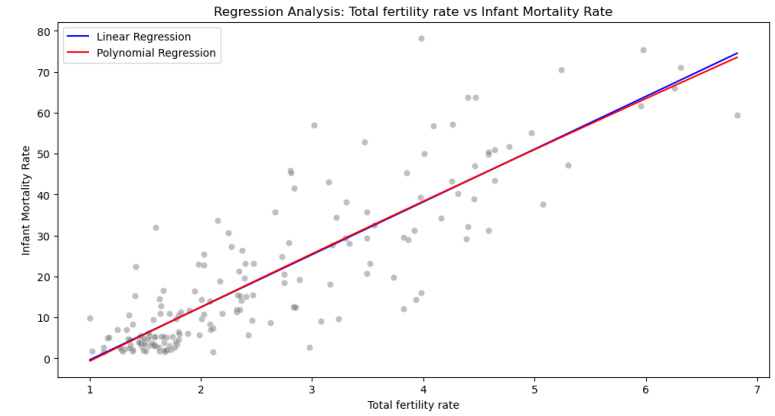
V Pythonu jsme použili lineární a polynomiální regresi pro posouzení vztahu mezi socioekonomickými faktory a mírou úmrtnosti novorozenců.

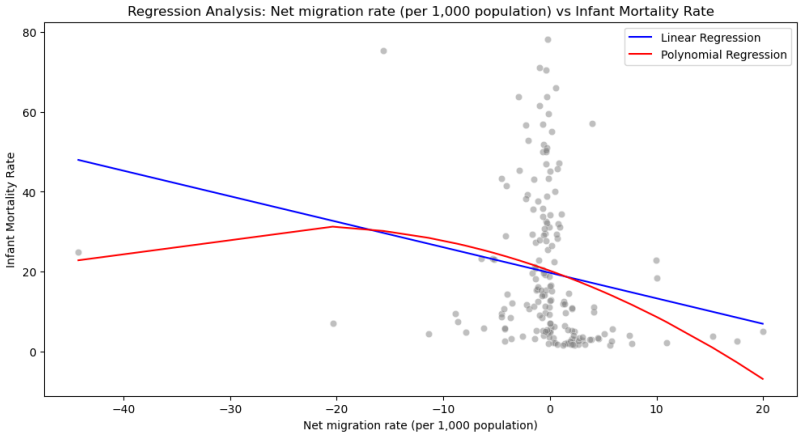

Regression Analysis: GDP vs Infant Mortality Rate
Linear Regression
Polynomial Regression
Infant Mortality Rate
GDP
80
60
40
20
0
−20
−40
0
50000
100000
150000
200000

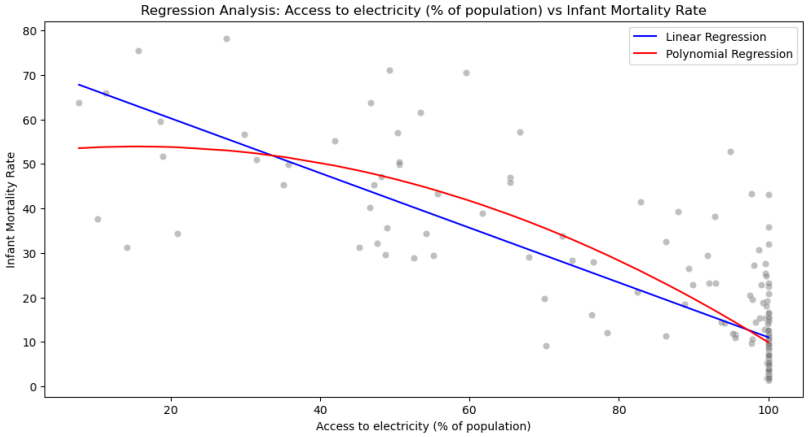

Regression Analysis: Life Exp Female vs Infant Mortality Rate
Linear Regression
Polynomial Regression
Infant Mortality Rate
Life Exp Female
80
60
40
20
0
55
60
65
70
75
80
85


Regression Analysis: Life Exp Men vs Infant Mortality Rate
Linear Regression
Polynomial Regression
Infant Mortality Rate
Life Exp Men
80
60
40
20
0
50
55
60
65
70
75
80
85


Regression Analysis: Total fertility rate vs Infant Mortality Rate
Linear Regression
Polynomial Regression
Infant Mortality Rate
Total fertility rate
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7







- **GDP:** Lineární model s R-kvadrátem 0.207930 oproti polynomiálnímu modelu s R-kvadrátem 0.329812 ukazuje lepší výkon polynomiálního modelu.

- **Life Exp Female, Life Exp Men:** Polynomiální modely pro obě pohlaví indikují silnější nelineární vztahy než lineární modely.

- **Total fertility rate:** Lineární regrese s R-kvadrátem 0.735668; polynomiální regrese s R-kvadrátem 0.735751.

- **Net migration rate (per 1,000 population):** Lineární regrese s R-kvadrátem 0.034149; polynomiální regrese s R-kvadrátem 0.054412.

- **Share of urban population:** Lineární regrese s R-kvadrátem 0.314340; polynomiální regrese s R-kvadrátem 0.316237.

- **Access to electricity (% of population):** Lineární regrese s R-kvadrátem 0.648542; polynomiální regrese s R-kvadrátem 0.679937.

	Indicator	Model	MSE	RMSE	MAE	Explained Variance	R-squared
0	GDP	Linear	279.267896	16.711310	13.262785	0.207930	0.207930
1	GDP	Polynomial	236.294926	15.371888	12.299432	0.329812	0.329812
2	Life Exp Female	Linear	59.672426	7.724793	6.188762	0.830755	0.830755
3	Life Exp Female	Polynomial	44.679854	6.684299	4.703024	0.873277	0.873277
4	Life Exp Men	Linear	97.644336	9.881515	7.790136	0.723058	0.723058
5	Life Exp Men	Polynomial	73.739336	8.587161	5.935529	0.790858	0.790858
6	Total fertility rate	Linear	93.198025	9.653912	6.926900	0.735668	0.735668
7	Total fertility rate	Polynomial	93.169063	9.652412	6.926184	0.735751	0.735751
8	Net migration rate (per 1,000 population)	Linear	340.539691	18.453718	15.039770	0.034149	0.034149
9	Net migration rate (per 1,000 population)	Polynomial	333.395335	18.259116	14.946234	0.054412	0.054412
10	Share of urban population	Linear	241.749798	15.548305	11.608996	0.314340	0.314340
11	Share of urban population	Polynomial	241.081286	15.526793	11.686288	0.316237	0.316237
12	Access to electricity (% of population)	Linear	123.917121	11.131807	8.445207	0.648542	0.648542
13	Access to electricity (% of population)	Polynomial	112.847693	10.622979	8.100849	0.679937	0.679937

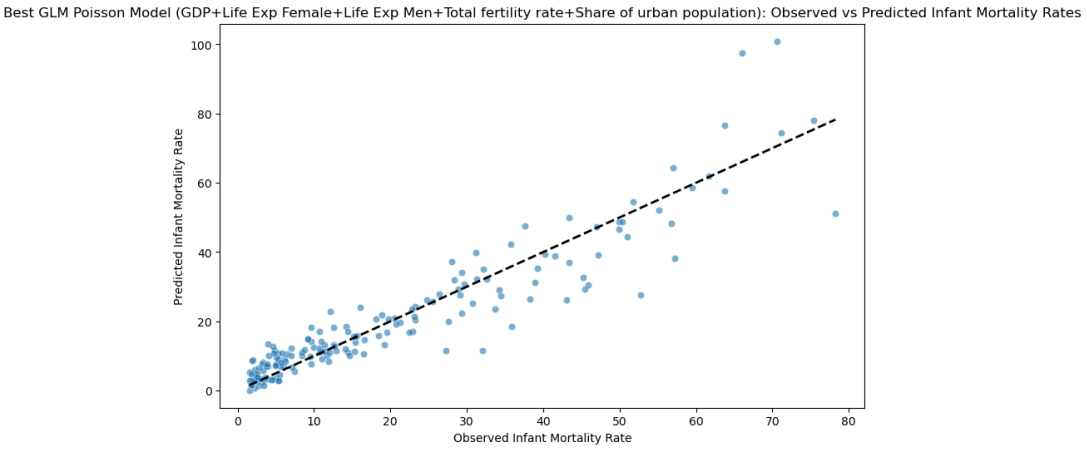
IV. GLM Poissonův model

Pro analýzu vlivu socioekonomických faktorů na míru úmrtnosti novorozenců jsme použili Generalized Linear Model s Poissonovou distribucí .Nejlepší kombinace prediktorů byla vybrána na základě nejnižší hodnoty Akaikeho informačního kritéria (AIC), které měří relativní kvalitu statistických modelů pro soubor dat. Model s nejnižší hodnotou AIC (1118.574650) zahrnoval **GDP, Life Exp Female, Life Exp Men, Total fertility rate a Share of urban population.**

	Combination	AIC
0	GDP, Life Exp Female, Life Exp Men, Total fert...	1118.574650
1	GDP, Life Exp Female, Life Exp Men, Total fert...	1118.891123
2	GDP, Life Exp Female, Life Exp Men, Total fert...	1119.361424
3	GDP, Life Exp Female, Life Exp Men, Total fert...	1119.424907
4	GDP, Life Exp Female, Life Exp Men, Total fert...	1120.288049
...	...	...
122	GDP	2128.185407
123	GDP, Net migration rate (per 1,000 population)	2128.819566
124	Net migration rate (per 1,000 population), Sha...	2622.410695
125	Share of urban population	2670.875103
126	Net migration rate (per 1,000 population)	3546.517431

127 rows × 2 columns

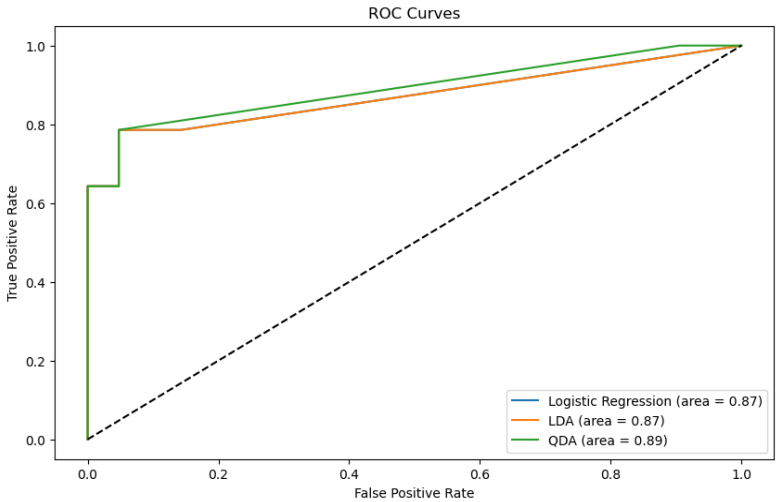
Z grafu pozorovaných vs předpovězených hodnot je zřejmé, že vybraný GLM model efektivně zachycuje trend v našem souboru dat.

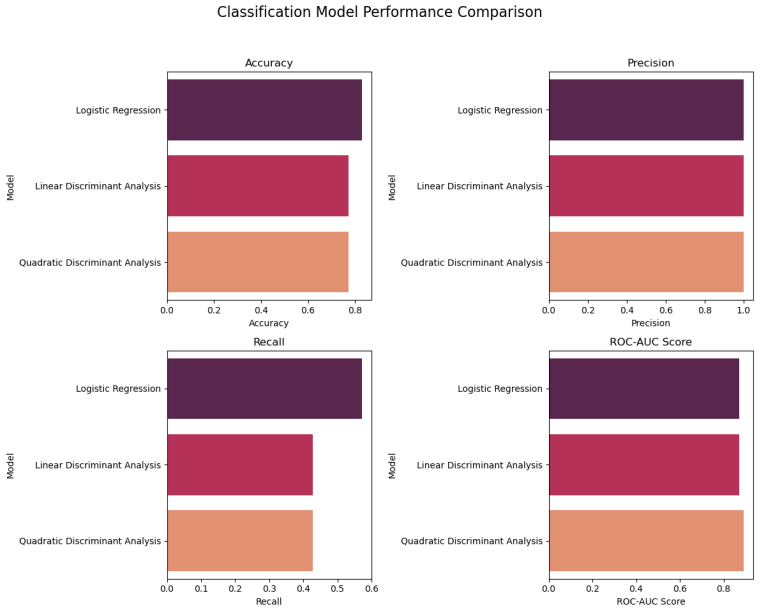


V. Logistická regrese vs LDA vs QDA

Provedli jsme binární klasifikaci míry úmrtnosti novorozenců pomocí logistické regrese, lineární diskriminační analýzy (LDA) a kvadratické diskriminační analýzy (QDA). Rozdělili jsme data na tréninkovou a testovací sadu a získali jsme následující výsledky:

- **Accuracy:** QDA měla nejvyšší přesnost, což ukazuje, že byla nejlepší při správném klasifikování zemí do kategorií vysoké a nízké úmrtnosti.
- **Precision:** Logistická regrese vykázala nejvyšší přesnost, což naznačuje, že byla nejlepší v minimalizaci falešných pozitiv.
- **Recall:** LDA měla nejvyšší citlivost, což naznačuje, že byla nejlepší v identifikaci pravých pozitiv.
- **ROC-AUC Score:** QDA měla nejvyšší skóre ROC-AUC, což naznačuje, že má nejlepší kompromis mezi pravděpodobností pravých pozitiv a falešných pozitiv ve srovnání s ostatními modely.





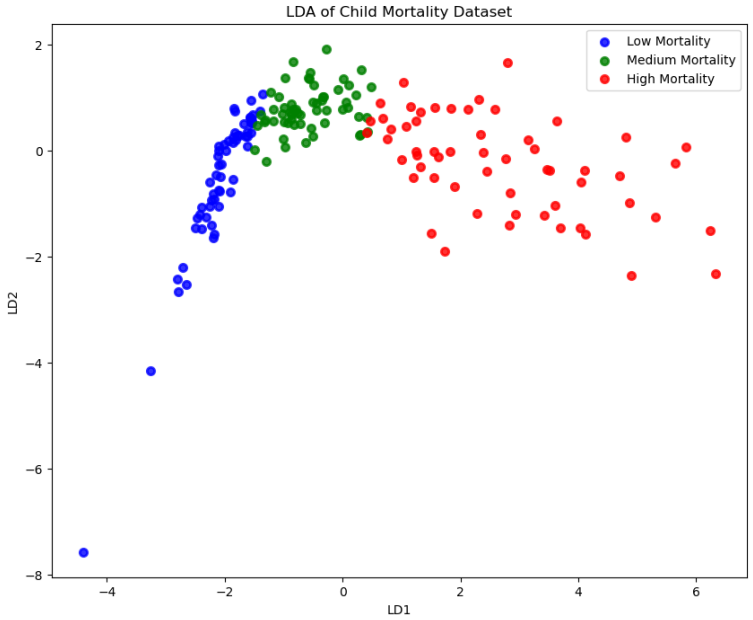
Model	Accuracy	Precision	Recall	ROC-AUC Score
Logistic Regression	0.828571	1.0	0.571429	0.870748
Linear Discriminant Analysis	0.771429	1.0	0.428571	0.870748
Quadratic Discriminant Analysis	0.771429	1.0	0.428571	0.891156

Vizualizace pomocí LDA a QDA ukázala **jasné rozdělení mezi různými třídami úmrtnosti.** Země byly podle předpokládané úmrtnosti rozděleny do tří kategorií: nízká, střední a vysoká úmrtnost. Příklady zahrnují Andorru a Austrálii s nízkou úmrtností, zatímco Afghánistán a Angola s vysokou úmrtností. Tyto klasifikace mohou být porovnány s výsledky metod shlukování (clustering) při další analýze.

- **Class 1:** Země s předpokládanou nízkou úmrtností podle QDA:
  'Andorra', 'Antigua and Barbuda', 'Argentina', 'Australia', 'Austria', 'Bahrain', 'Belarus', 'Belgium', 'Bosnia and Herzegovina', 'Brunei Darussalam', 'Canada', 'Chile', 'China', 'Costa Rica', 'Croatia', 'Cuba', 'Cyprus', 'Czechia', 'Denmark', 'Estonia', 'Finland', 'France', 'Germany', 'Greece', 'Hungary', 'Iceland', 'Ireland', 'Israel', 'Italy', 'Japan', 'Kuwait', 'Latvia', 'Lebanon', 'Lithuania', 'Luxembourg', 'Malaysia', 'Maldives', 'Malta', 'Monaco', 'New Zealand', 'North America', 'North Macedonia', 'Norway', 'Poland', 'Portugal', 'Qatar', 'Romania', 'Russian Federation', 'San Marino', 'Saudi Arabia', 'Serbia', 'Singapore', 'Slovenia', 'Spain', 'Sri Lanka', 'Sweden', 'Switzerland', 'Thailand', 'Turks and Caicos Islands', 'Ukraine', 'United Arab Emirates', 'United Kingdom', 'United States', 'Uruguay'

- **Class 2:** Země s předpokládanou střední úmrtností podle QDA:
  'Albania', 'Algeria', 'Armenia', 'Azerbaijan', 'Bangladesh', 'Barbados', 'Belize', 'Bhutan', 'Brazil', 'British Virgin Islands', 'Bulgaria', 'Cabo Verde', 'Colombia', 'Ecuador', 'El Salvador', 'Fiji', 'Georgia', 'Grenada', 'Guatemala', 'Guyana', 'Honduras', 'Indonesia', 'Iraq', 'Jamaica', 'Jordan', 'Kazakhstan', 'Mauritius', 'Mexico', 'Mongolia', 'Montenegro', 'Morocco', 'Nauru', 'Nepal', 'Nicaragua', 'Oman', 'Palau', 'Panama', 'Paraguay', 'Peru', 'Philippines', 'Samoa', 'Sao Tome and Principe', 'Seychelles', 'Solomon Islands', 'Suriname', 'Syrian Arab Republic', 'Tonga', 'Trinidad and Tobago', 'Tunisia', 'Tuvalu', 'Uzbekistan', 'Viet Nam'

- **Class 3:** Země s předpokládanou vysokou úmrtností podle QDA:
  'Afghanistan', 'Angola', 'Benin', 'Botswana', 'Burkina Faso', 'Burundi', 'Cambodia', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Djibouti', 'Dominica', 'Dominican Republic', 'Equatorial Guinea', 'Eritrea', 'Eswatini', 'Ethiopia', 'Gabon', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Haiti', 'India', 'Kenya', 'Kiribati', 'Lesotho', 'Liberia', 'Libya', 'Madagascar', 'Malawi', 'Mali', 'Marshall Islands', 'Mauritania', 'Mozambique', 'Myanmar', 'Namibia', 'Niger', 'Nigeria', 'Pakistan', 'Papua New Guinea', 'Rwanda', 'Senegal', 'Sierra Leone', 'Somalia', 'South Africa', 'South Asia', 'South Sudan', 'Sub-Saharan Africa', 'Sub-Saharan Africa', 'Sudan', 'Tajikistan', 'Timor-Leste', 'Togo', 'Turkmenistan', 'Uganda', 'Vanuatu', 'Zambia', 'Zimbabwe'



- **Class 1:** Země s předpokládanou nízkou úmrtností podle LDA:
  [Andorra', 'Antigua and Barbuda', 'Argentina', 'Australia', 'Austria', 'Bahrain', 'Belarus', 'Belgium', 'Bosnia and Herzegovina', 'Bulgaria', 'Canada', 'Chile', 'China', 'Croatia', 'Cuba', 'Cyprus', 'Czechia', 'Denmark', 'Estonia', 'Finland', 'France', 'Germany', 'Greece', 'Hungary', 'Iceland', 'Ireland', 'Israel', 'Italy', 'Japan', 'Latvia', 'Lithuania', 'Luxembourg', 'Maldives', 'Malta', 'Monaco', 'Montenegro', 'New Zealand', 'North America', 'North Macedonia', 'Norway', 'Poland', 'Portugal', 'Qatar', 'Romania', 'Russian Federation', 'San Marino', 'Saudi Arabia', 'Serbia', 'Singapore', 'Slovenia', 'Spain', 'Sri Lanka', 'Sweden', 'Switzerland', 'Turks and Caicos Islands', 'United Arab Emirates', 'United Kingdom', 'United States', 'Uruguay'

- **Class 2:** Země s předpokládanou střední úmrtností podle LDA:
  'Albania', 'Algeria', 'Armenia', 'Azerbaijan', 'Bangladesh', 'Barbados', 'Belize', 'Bhutan', 'Brazil', 'British Virgin Islands', 'Brunei Darussalam', 'Cabo Verde', 'Cambodia', 'Colombia', 'Costa Rica', 'Ecuador', 'El Salvador', 'Georgia', 'Grenada', 'Guatemala', 'Honduras', 'Indonesia', 'Iraq', 'Jamaica', 'Jordan', 'Kazakhstan', 'Kuwait', 'Lebanon', 'Libya', 'Malaysia', 'Mauritius', 'Mexico', 'Mongolia', 'Morocco', 'Nauru', 'Nepal', 'Nicaragua', 'Oman', 'Palau', 'Panama', 'Paraguay', 'Peru', 'Philippines', 'Samoa', 'Sao Tome and Principe', 'Seychelles', 'Solomon Islands', 'Suriname', 'Syrian Arab Republic', 'Thailand', 'Tonga', 'Trinidad and Tobago', 'Tunisia', 'Tuvalu', 'Ukraine', 'Uzbekistan', 'Vanuatu', 'Viet Nam'

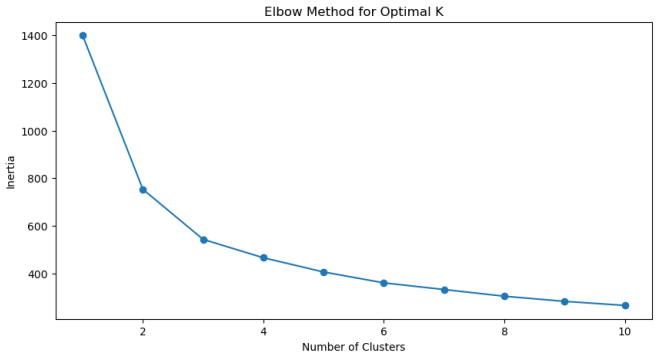
- **Class 3:** Země s předpokládanou vysokou úmrtností podle LDA:
  'Afghanistan', 'Angola', 'Benin', 'Botswana', 'Burkina Faso', 'Burundi', 'Cameroon', 'Central African Republic', 'Chad', 'Comoros', 'Djibouti', 'Dominica', 'Dominican Republic', 'Equatorial Guinea', 'Eritrea', 'Eswatini', 'Ethiopia', 'Fiji', 'Gabon', 'Ghana', 'Guinea', 'Guinea-Bissau', 'Guyana', 'Haiti', 'India', 'Kenya', 'Kiribati', 'Lesotho', 'Liberia', 'Madagascar', 'Malawi', 'Mali', 'Marshall Islands', 'Mauritania', 'Mozambique', 'Myanmar', 'Namibia', 'Niger', 'Nigeria', 'Pakistan', 'Papua New Guinea', 'Rwanda', 'Senegal', 'Sierra Leone', 'Somalia', 'South Africa', 'South Asia', 'South Sudan', 'Sub-Saharan Africa', 'Sub-Saharan Africa', 'Sudan', 'Tajikistan', 'Timor-Leste', 'Togo', 'Turkmenistan', 'Uganda', 'Zambia', 'Zimbabwe'

VI. K-Means shlukování.

Naše analýza využívající shlukování K-Means poskytla strukturovaný přístup ke kategorizaci zemí na základě jejich socioekonomických faktorů a dostupnosti zdravotní péče, aby bylo možné pochopit jejich vliv na míru úmrtnosti novorozenců.

a) Loketní metoda (Elbow Method) pro optimální K:

Loketní graf sestrojený s různými počty shluků ukázal optimální počet shluků při K=3. Tento bod představuje rovnováhu, kdy další shluky nepřinášejí významný přínos při snižování rozptylu uvnitř shluku.



b) Vytváření shluků:

Při použití metody K-Means s **k_optimal=3** na standardizovaný soubor dat jsme identifikovali tři odlišné shluky, které zachycují různé úrovně socioekonomického rozvoje a dostupnosti zdravotní péče:

- **Cluster 0:** Představuje země s vysokým HDP, vynikajícím přístupem k elektřině a nejnižší úmrtností novorozenců, což naznačuje robustní systémy zdravotní péče a ekonomickou stabilitu

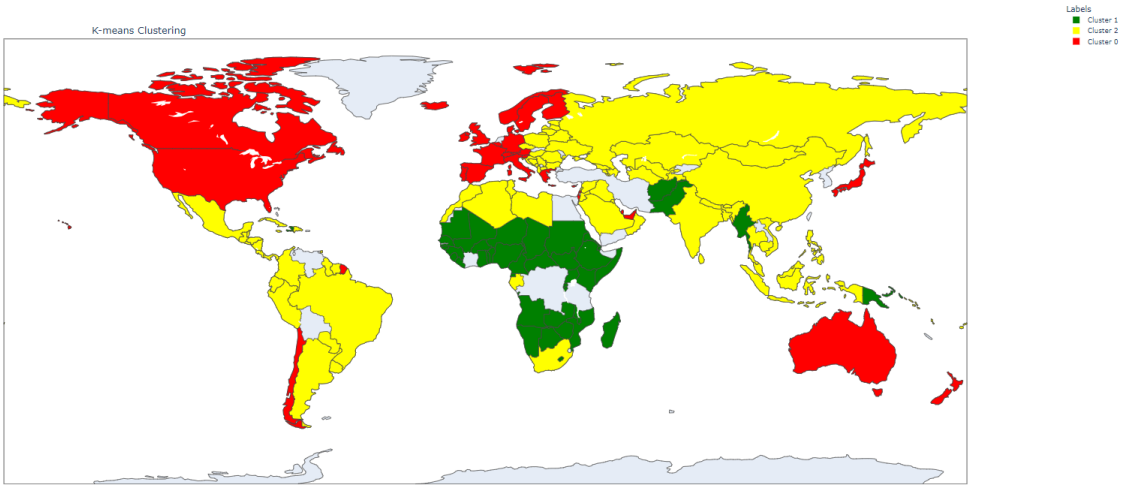
  Země: Andorra, Australia, Austria, Belgium, Canada, Chile, Cyprus, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Japan, Luxembourg, Malta, Monaco, New Zealand, North America, Norway, Portugal, Qatar, San Marino, Singapore, Slovenia, Spain, Sweden, Switzerland, Turks and Caicos Islands, United Arab Emirates, United Kingdom, United States

- **Cluster 1:** Zahrnuje země s nejnižším HDP, špatným přístupem k elektřině a nejvyšší úmrtností novorozenců.

  Země: Afghanistan, Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Comoros, Djibouti, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Ghana, Guinea, Guinea-Bissau, Haiti, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Myanmar, Namibia, Niger, Nigeria, Pakistan, Papua New Guinea, Rwanda, Senegal, Sierra Leone, Somalia, South Sudan, Sub-Saharan Africa, Sub-Saharan Africa, Sudan, Togo, Uganda, Zambia, Zimbabwe

- **Cluster 2:** Obsahuje země se středním HDP a různým přístupem k elektřině, s mírou úmrtnosti vyšší než ve shluku 0, ale výrazně nižší než ve shluku 1.

  Země: Albania, Algeria, Antigua and Barbuda, Argentina, Armenia, Azerbaijan, Bahrain, Bangladesh, Barbados, Belarus, Belize, Bhutan, Bosnia and Herzegovina, Brazil, British Virgin Islands, Brunei Darussalam, Bulgaria, Cabo Verde, Cambodia, China, Colombia, Costa Rica, Croatia, Cuba, Czechia, Dominica, Dominican Republic, Ecuador, El Salvador, Estonia, Fiji, Gabon, Georgia, Grenada, Guatemala, Guyana, Honduras, Hungary, India, Indonesia, Iraq, Jamaica, Jordan, Kazakhstan, Kiribati, Kuwait, Latvia, Lebanon, Libya, Lithuania, Malaysia, Maldives, Marshall Islands, Mauritius, Mexico, Mongolia, Montenegro, Morocco, Nauru, Nepal, Nicaragua, North Macedonia, Oman, Palau, Panama, Paraguay, Peru, Philippines, Poland, Romania, Russian Federation, Samoa, Sao Tome and Principe, Saudi Arabia, Serbia, Seychelles, Solomon Islands, South Africa, South Asia, Sri Lanka, Suriname, Syrian Arab Republic, Tajikistan, Thailand, Timor-Leste, Tonga, Trinidad and Tobago, Tunisia, Turkmenistan, Tuvalu, Ukraine, Uruguay, Uzbekistan, Vanuatu, Viet Nam



c) Metriky hodnocení shluků

Pro posouzení kvality shlukování K-Means jsme vypočítali tři klíčové metriky:

- Silhouette Coefficient (0,37): měří, jak je daný objekt podobný svému shluku ve srovnání s ostatními shluky. Vyšší skóre siluety naznačuje, že objekty se dobře shodují s vlastním shlukem a špatně se shodují se sousedními shluky. V našem případě skóre naznačuje mírnou separaci mezi shluky.

- Calinski-Harabasz Index (135,57): je vyšší, když jsou shluky husté a dobře oddělené, náš výsledek naznačuje, že shluky jsou rozumně odlišné a dobře oddělené.

- Davies-Bouldin Index (0,93): označuje průměrnou "podobnost" mezi shluky,skóre blízké 0 znamená lepší rozdělení. Naše skóre je poměrně nízké, což naznačuje dobré oddělení shluků, ale je zde prostor pro zlepšení.

```
[143]: # Recalculating evaluation metrics
silhouette_avg = silhouette_score(X_standardized, clusters)
calinski_harabasz = calinski_harabasz_score(X_standardized, clusters)
davies_bouldin = davies_bouldin_score(X_standardized, clusters)

silhouette_avg, calinski_harabasz, davies_bouldin

[143]: (0.3738255346118453, 135.56848615742015, 0.9305087312919161)
```

Tyto metriky společně naznačují, že naše shlukování si při seskupování zemí do odlišných kategorií na základě uvažovaných znaků vedlo vhodně.

- V naší průzkumné analýze dat jsme rozšířili Bayesovské informační kritérium (BIC = 1210,75) na algoritmus K-Means, což je netradiční, protože BIC se obvykle používá u pravděpodobnostních modelů. Tato adaptace byla motivována potřebou stanovit srovnávací metriku napříč shlukovacími technikami, včetně modelů Gaussian Mixture Models (GMM), které ze své podstaty používají BIC pro výběr modelu.Obvykle se BIC vypočítá pro několik různých modelů (s různým počtem shluků nebo **různými typy shlukování**) a poté **se tato skóre porovnají.** *Čím nižší je BIC, tím lepší je kompromis mezi vhodností modelu a jeho složitostí.*

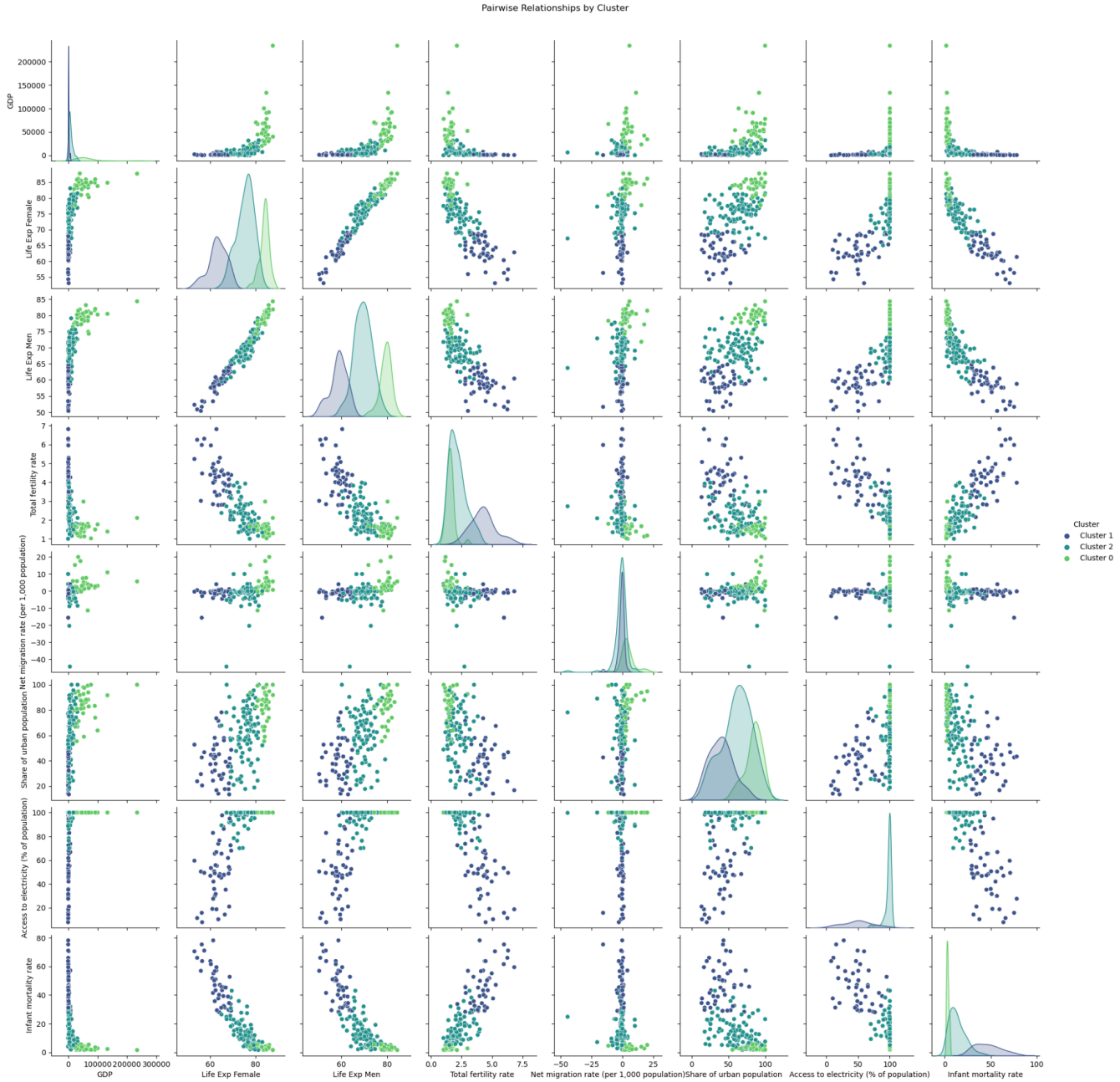
Naše vlastní skóre podobné BIC pro K-Means, odvozené ze součtu čtverců (inertia) uvnitř shluku.

```python
n_samples = X_standardized.shape[0]  # number of samples
n_features = X_standardized.shape[1]  # number of features
n_clusters = k_optimal  # number of clusters
inertia = kmeans.inertia_  # within-cluster sum of squares from KMeans model

# Number of parameters: number of clusters * number of features per cluster centroid
n_parameters = n_clusters * n_features

# BIC-like score for K-Means
bic_like = np.log(n_samples) * n_parameters - 2 * (-inertia)

print(f"BIC-like score for K-Means with {n_clusters} clusters: {bic_like}")
```

```
BIC-like score for K-Means with 3 clusters: 1210.7520165980043
```

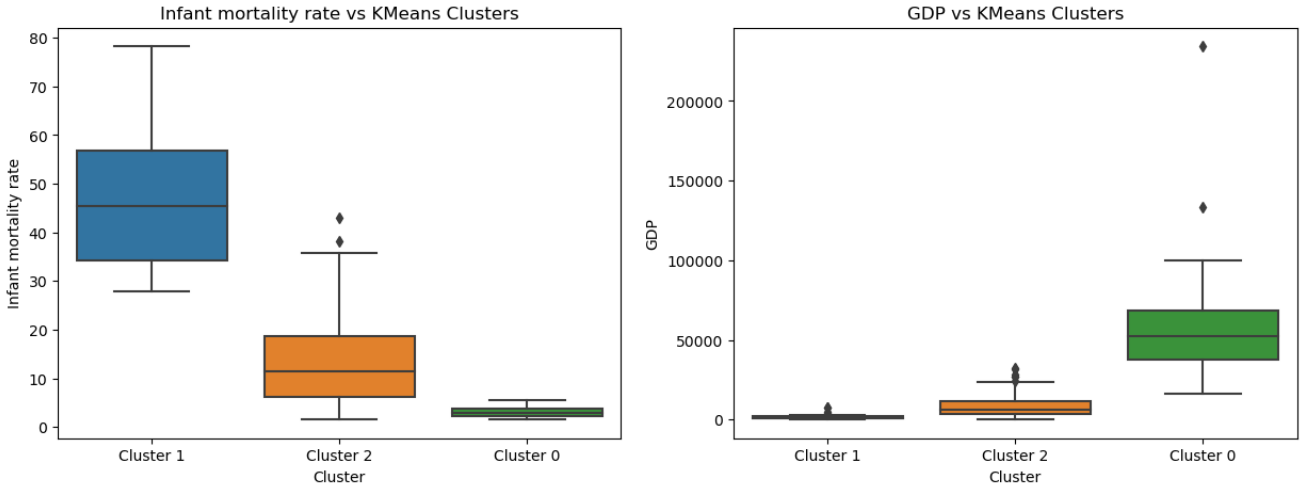
d) Párové vztahy podle shluků:

Párový graf poskytl vizuální srovnání napříč shluky a znázornil, jak spolu jednotlivé dvojice indikátorů v rámci každého shluku interagují.



e) Boxploty GDP, Infant Mortality Rate vs Shluky

Vizuální srovnání zdůrazňují silný nepřímý vztah mezi bohatstvím země a mírou úmrtnosti novorozenců.



III. Gaussian Mixture Model (GMM) shlukování.

Využitím GMM s optimalizovaným počtem komponent (k_optimal=3) jsme získali následující poznatky:

a) Vytváření shluků:

Při použití metody GMM s **k_optimal=3** jsme identifikovali tři odlišné shluky:

- **Cluster 0:** země s mírným HDP a průměrným přístupem k elektřině a zdravotní péči.

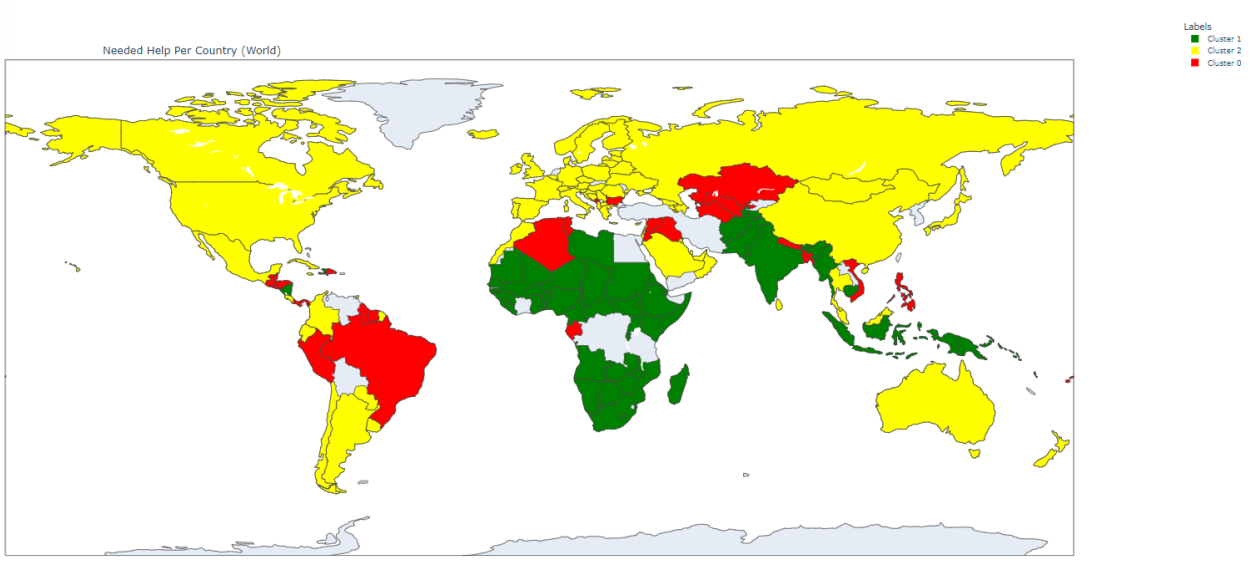
  Země: Algeria, Bangladesh, Belize, Brazil, Bulgaria, Cabo Verde, Dominica, Dominican Republic, El Salvador, Fiji, Gabon, Grenada, Guatemala, Guyana, Honduras, Iraq, Jordan, Kazakhstan, Marshall Islands, Mauritius, Montenegro, Nauru, Nepal, Panama, Peru, Philippines, Samoa, Suriname, Syrian Arab Republic, Tonga, Tunisia, Turkmenistan, Tuvalu, Uzbekistan, Viet Nam

- **Cluster 1:** země s nízkým HDP a vysokou úmrtností dětí, což ukazuje na akutní potřebu zlepšení zdravotní péče a sociální podpory.

  Země: Afghanistan, Angola, Benin, Bhutan, Botswana, Burkina Faso, Burundi, Cambodia, Cameroon, Central African Republic, Chad, Comoros, Djibouti, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Ghana, Guinea, Guinea-Bissau, Haiti, India, Indonesia, Kenya, Kiribati, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mozambique, Myanmar, Namibia, Nicaragua, Niger, Nigeria, Pakistan, Papua New Guinea, Rwanda, Sao Tome and Principe, Senegal, Sierra Leone, Solomon Islands, Somalia, South Africa, South Asia, South Sudan, Sub-Saharan Africa, Sub-Saharan Africa, Sudan, Tajikistan, Timor-Leste, Togo, Uganda, Vanuatu, Zambia, Zimbabwe

- **Cluster 2:** země s vysokým HDP a nízkou úmrtností dětí, což odráží lepší socioekonomické podmínky a dostupnost zdravotní péče

  Země: Albania, Andorra, Antigua and Barbuda, Argentina, Armenia, Australia, Austria, Azerbaijan, Bahrain, Barbados, Belarus, Belgium, Bosnia and Herzegovina, British Virgin Islands, Brunei Darussalam, Canada, Chile, China, Colombia, Costa Rica, Croatia, Cuba, Cyprus, Czechia, Denmark, Ecuador, Estonia, Finland, France, Georgia, Germany, Greece, Hungary, Iceland, Ireland, Israel, Italy, Jamaica, Japan, Kuwait, Latvia, Lebanon, Lithuania, Luxembourg, Malaysia, Maldives, Malta, Mexico, Monaco, Mongolia, Morocco, New Zealand, North America, North Macedonia, Norway, Oman, Palau, Paraguay, Poland, Portugal, Qatar, Romania, Russian Federation, San Marino, Saudi Arabia, Serbia, Seychelles, Singapore, Slovenia, Spain, Sri Lanka, Sweden, Switzerland, Thailand, Trinidad and Tobago, Turks and Caicos Islands, Ukraine, United Arab Emirates, United Kingdom, United States, Uruguay



b) Metriky hodnocení shluků

Pro posouzení kvality shlukování GMM jsme vypočítali BIS kriterium:

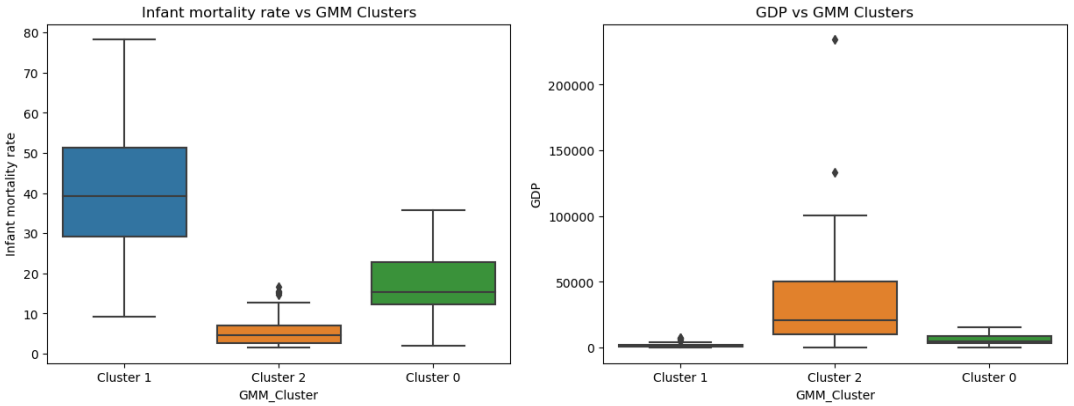
- Hodnota **BIC pro GMM byla 1248.51,** což naznačuje dobré přizpůsobení modelu datům bez přetrénování.

```
# Calculate and print the Bayesian Information Criterion (BIC)
bic = gmm.bic(X_standardized)
print("BIC for GMM:", bic)
```

```
BIC for GMM: 1248.5068463835019
```

c) Boxploty GDP, Infant Mortality Rate vs Šhluky

Vizuální srovnání zdůrazňují silný nepřímý vztah mezi bohatstvím země a mírou úmrtnosti novorozenců.

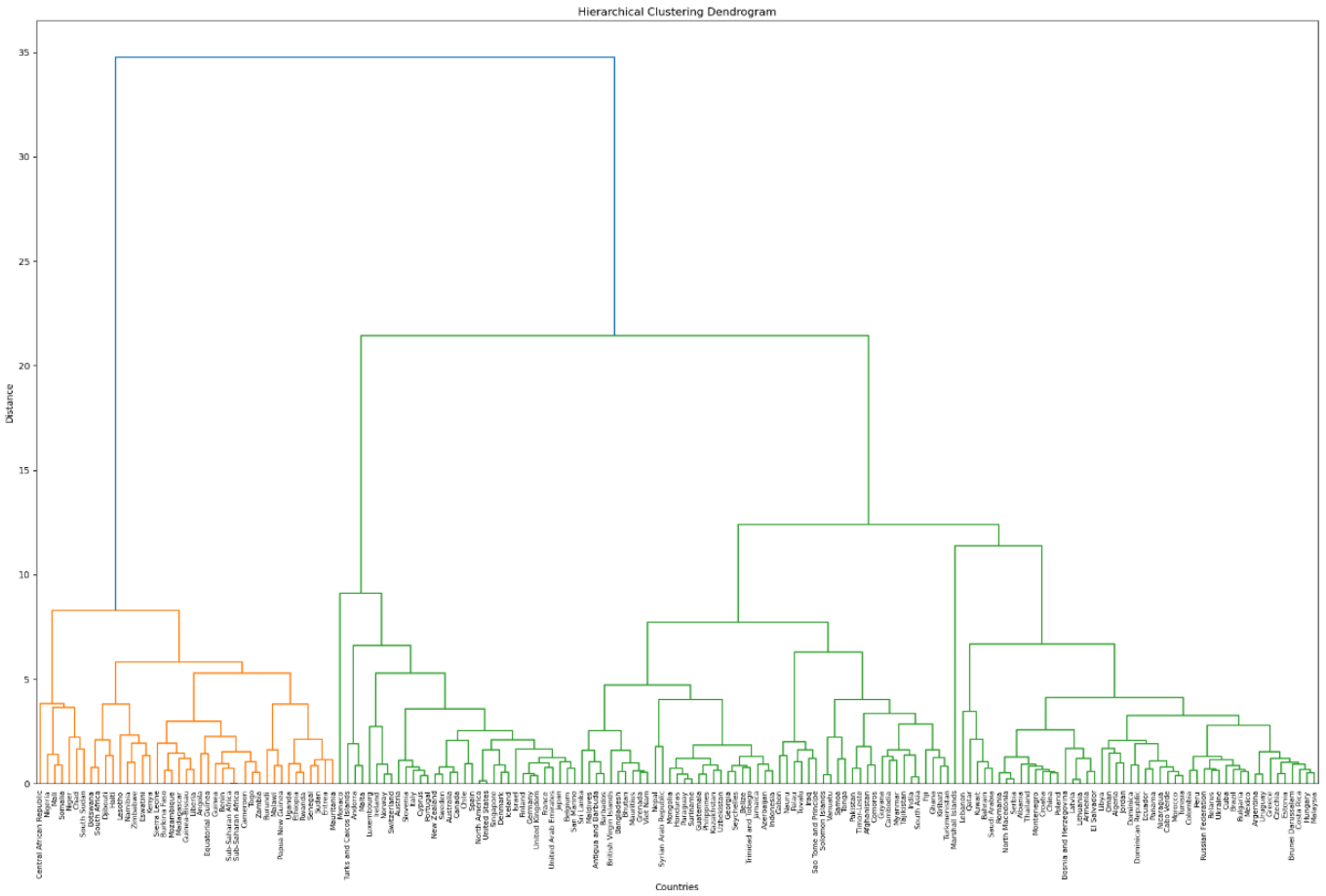


IV. Hierarchical shlukování.

V této části se zaměříme na analýzu hierarchického shlukování v rámci našeho datasetu. Hierarchické shlukování je zvoleno pro jeho schopnost vizualizovat vztahy mezi různými zeměmi a identifikovat možné skupiny s podobnými charakteristikami.

a) Dendrogram

Dendrogram ukazuje vzdálenosti mezi zeměmi podle jejich socioekonomických a zdravotních charakteristik. Země jsou rozděleny do shluků, které odrážejí jejich podobnost. Z dendrogramu možné identifikovat několik hlavních skupin, které reprezentují různé úrovně socioekonomického rozvoje a zdravotní péče.



b) Vytváření shluků:

Bylo aplikováno **aglomerativní hierarchické shlukování** s využitím metody "**ward**", která minimalizuje varianci uvnitř shluků (**agg_cluster = AgglomerativeClustering(n_clusters=3, affinity='euclidean', linkage='ward'**)). Počet shluků byl nastaven na 3:

- **Cluster 0:** země tohoto shluku jsou geograficky rozptýleny, s koncentracemi v Latinské Americe, Asii a východní Evropě.

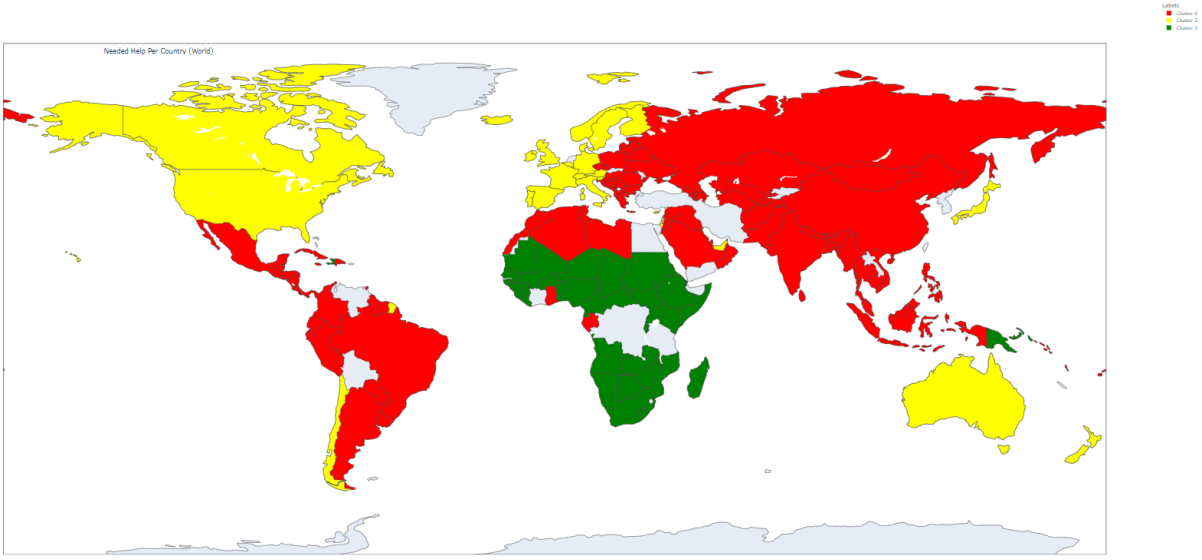
  Země: Afghanistan, Albania, Algeria, Antigua and Barbuda, Argentina, Armenia, Azerbaijan, Bahrain, Bangladesh, Barbados, Belarus, Belize, Bhutan, Bosnia and Herzegovina, Brazil, British Virgin Islands, Brunei Darussalam, Bulgaria, Cabo Verde, Cambodia, China, Colombia, Comoros, Costa Rica, Croatia, Cuba, Czechia, Dominica, Dominican Republic, Ecuador, El Salvador, Estonia, Fiji, Gabon, Georgia, Ghana, Greece, Grenada, Guatemala, Guyana, Honduras, Hungary, India, Indonesia, Iraq, Jamaica, Jordan, Kazakhstan, Kiribati, Kuwait, Latvia, Lebanon, Libya, Lithuania, Malaysia, Maldives, Marshall Islands, Mauritius, Mexico, Mongolia, Montenegro, Morocco, Myanmar, Nauru, Nepal, Nicaragua, North Macedonia, Oman, Pakistan, Palau, Panama, Paraguay, Peru, Philippines, Poland, Qatar, Romania, Russian Federation, Samoa, Sao Tome and Principe, Saudi Arabia, Serbia, Seychelles, Solomon Islands, South Asia, Sri Lanka, Suriname, Syrian Arab Republic, Tajikistan, Thailand, Timor-Leste, Tonga, Trinidad and Tobago, Tunisia, Turkmenistan, Tuvalu, Ukraine, Uruguay, Uzbekistan, Vanuatu, Viet Nam

- **Cluster 1:** skládá se převážně ze subsaharských afrických zemí, které trpí vysokou úmrtností dětí a nižším HDP. To ukazuje na významný vliv socioekonomických faktorů a přístupu ke zdravotní péči na dětskou úmrtnost.

  Země: Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Central African Republic, Chad, Djibouti, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Guinea, Guinea-Bissau, Haiti, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mozambique, Namibia, Niger, Nigeria, Papua New Guinea, Rwanda, Senegal, Sierra Leone, Somalia, South Africa, South Sudan, Sub-Saharan Africa, Sub-Saharan Africa, Sudan, Togo, Uganda, Zambia, Zimbabwe

- **Cluster 2:** vyspělé země s nízkou úmrtností dětí a vysokým HDP, jako jsou Spojené státy, Japonsko a země západní Evropy

  Země: Andorra, Australia, Austria, Belgium, Canada, Chile, Cyprus, Denmark, Finland, France, Germany, Iceland, Ireland, Israel, Italy, Japan, Luxembourg, Malta, Monaco, New Zealand, North America, Norway, Portugal, San Marino, Singapore, Slovenia, Spain, Sweden, Switzerland, Turks and Caicos Islands, United Arab Emirates, United Kingdom, United States



c) Metriky hodnocení shluků

Pro posouzení kvality shlukování K-Means jsme vypočítali tři klíčové metriky:

- Silhouette Coefficient (0.36): Ukazuje na poměrně dobrou kvalitu shluků, ale stále existuje prostor pro zlepšení.

- Calinski-Harabasz Index (126.60): Naznačuje, že shluky jsou dobře oddělené a kohezivní.

- Davies-Bouldin Index(0.92): Nižší hodnota by byla ideální, ale tato hodnota stále naznačuje, že shluky jsou rozumně dobře definované.

- V naší průzkumné analýze dat jsme rozšířili Bayesovské informační kritérium (BIC = 1246.33) na algoritmus Hierarchical Clustering, což je netradiční, protože BIC se obvykle používá u pravděpodobnostních modelů. Tato adaptace byla motivována potřebou stanovit srovnávací metriku napříč shlukovacími technikami, včetně modelů Gaussian Mixture Models (GMM), které ze své podstaty používají BIC pro výběr modelu.

```python
# Determine the number of clusters from hierarchical clustering
n_clusters = len(np.unique(clusters_agg))

# Fit a Gaussian Mixture Model with the number of clusters found by hierarchical clustering
gmm = GaussianMixture(n_components=n_clusters, covariance_type='full')
gmm.fit(X_standardized)

# Calculate the Bayesian Information Criterion
bic = gmm.bic(X_standardized)

print(f"The Bayesian Information Criterion for Hierarchical model is: {bic}")
```

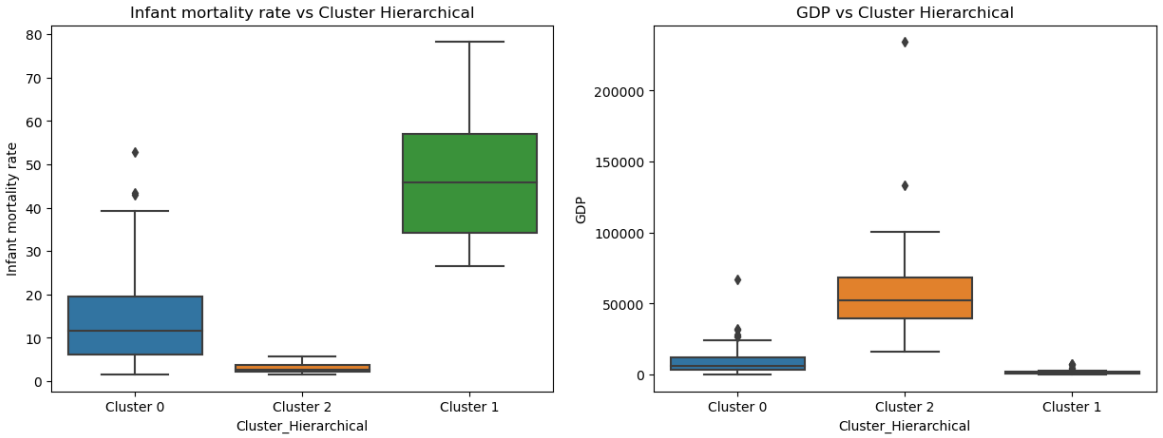
```
The Bayesian Information Criterion for Hierarchical model is: 1248.5068463835014
```

d) Boxploty GDP, Infant Mortality Rate vs Šhluky

Boxploty zobrazují rozložení klíčových proměnných - dětské úmrtnosti a HDP - mezi různými shluky:

- Dětská úmrtnost: Shluk 1 vykazuje výrazně vyšší míru dětské úmrtnosti ve srovnání se shlukem 0 a 2, což naznačuje, že země v tomto shluku čelí větším výzvám v oblasti zdravotní péče a mohou vyžadovat zvýšené zdravotnické zásahy.

- HDP: Shluk 2 má výrazně vyšší HDP ve srovnání s ostatními shluky, což ukazuje na silný vztah mezi ekonomickým bohatstvím a nižší dětskou úmrtností.



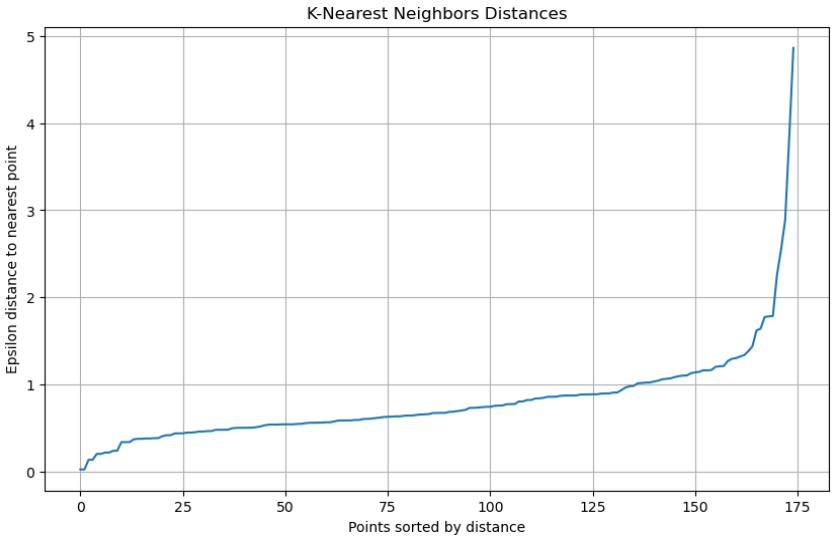
V. DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) je metoda shlukování, která identifikuje shluky jako oblasti s vysokou hustotou bodů, je robustní na práce s outliery/šumem. Dostali jsme dost zajímavé výsledky.

Model 1.

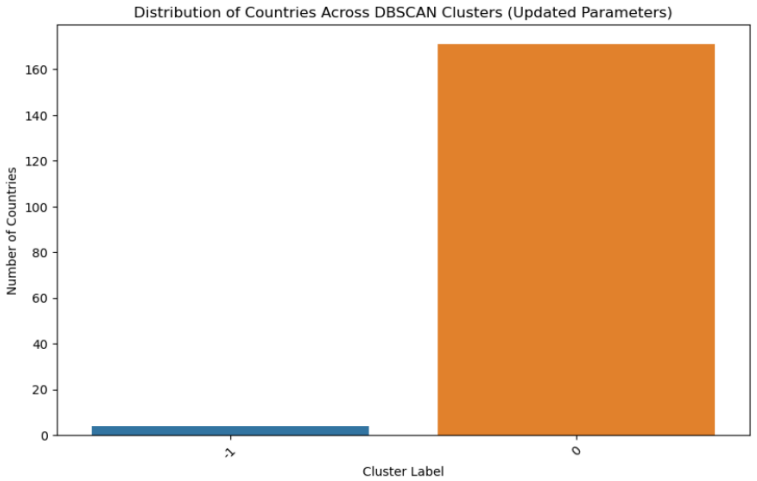
a) Volba parametrů

Pro určení optimální hodnoty epsilonu (eps) byly použity metody k-nejbližších sousedů (KNN). Graf KNN vzdáleností naznačuje, že hodnota eps 2.5 je vhodným odhadem, kde křivka začíná prudce stoupat, což je znakem vhodné hodnoty pro vytvoření shluků.



b) Výsledky DBSCAN

Použitím DBSCAN s původními parametry (eps=0.5, min_samples=5) bylo nalezeno 171 zemí v jednom hlavním shluku a 4 země byly klasifikovány jako šum. Předpokládáme, že se to stalo, protože naše data jsou téměř rovnoměrně rozložená, takže tento algoritmus nemohl rozdělit země do více než jednoho shluku.

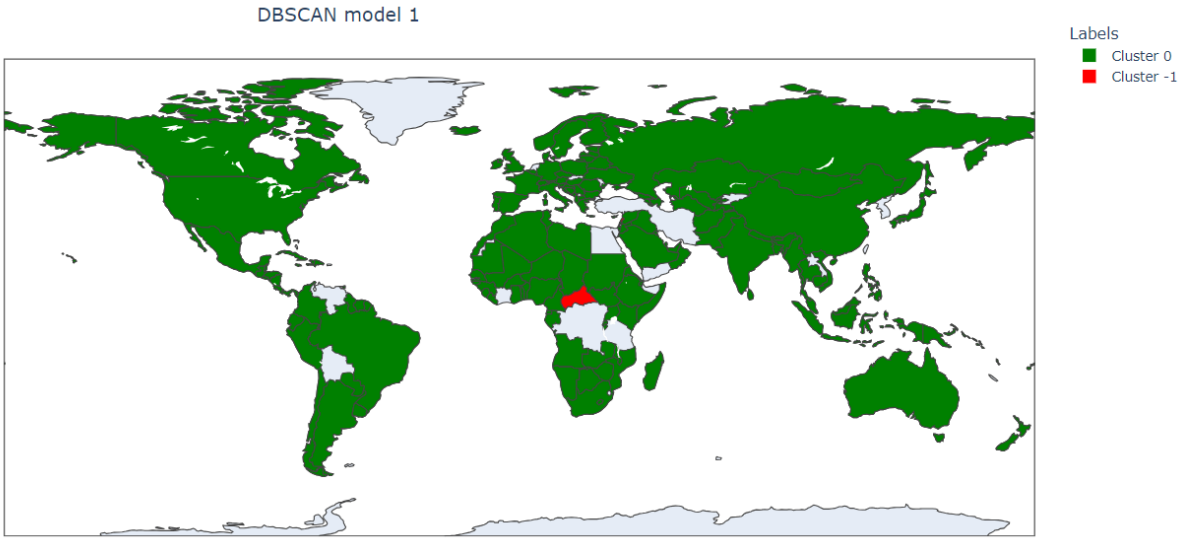



```
[220]: DBSCAN_Cluster
        0    171
       -1      4
       Name: count, dtype: int64
```

Konkrétně země Středoafrická republika, Libanon, Marshallovy ostrovy a Monako byly identifikovány jako odlehlé body mimo hlavní shluky.

```
# Identifying the countries classified as noise by DBSCAN
noise_countries = data[data['DBSCAN_Cluster'] == -1]['Geographic area']

# Displaying the countries classified as noise
noise_countries.tolist()
```

```
['Central African Republic', 'Lebanon', 'Marshall Islands', 'Monaco']
```

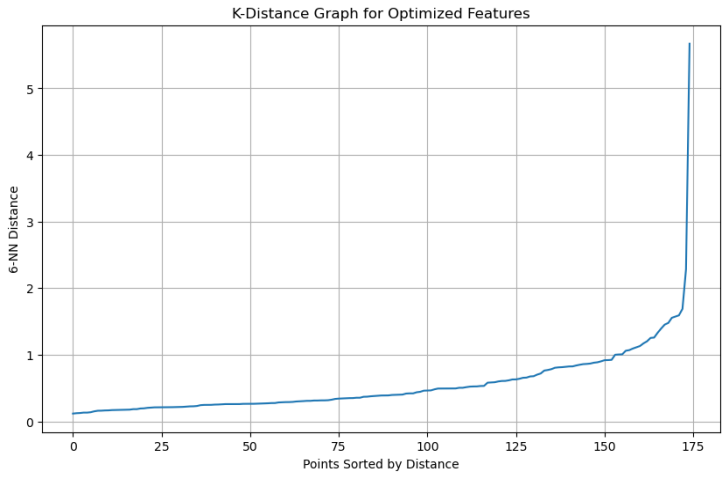


Model 2.

a) Optimalizace vlastností

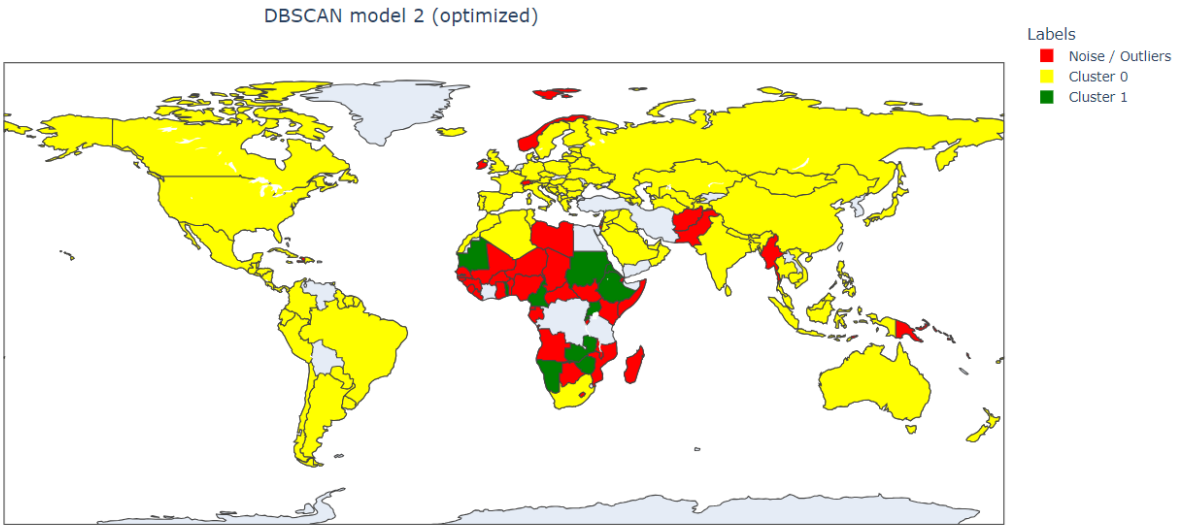
Byla provedena další optimalizace výběru indicátoru pro DBSCAN, která se 'Access to electricity (% of population)', 'Total fertility rate', 'Infant mortality rate', 'GDP'

Graf 6-NN vzdáleností naznačuje, že eps 0.5 je vhodnou hodnotou pro tyto optimalizované rysy.



b) Shluky a interpretace

DBSCAN identifikoval dva hlavní shluky a několik zemí jako šum po optimalizaci vlastností. Hlavní shluk (label 0) obsahoval 115 zemí, zatímco druhý shluk (label 1) obsahoval 12 zemí. Šum (label -1) obsahoval 48 zemí, což naznačuje země s atypickými socioekonomickými a zdravotními charakteristikami ve srovnání s ostatními.



Shrnutím, výhodou použití DBSCAN je jeho schopnost pracovat s daty s outliery/šumem a odhalovat přirozené shluky bez nutnosti předem specifikovat jejich počet. Však pro náš soubor DBSCAN moc užitečný nebyl.

SHRNUTÍ VÝSLEDKŮ METOD SHLUKOVÁNÍ :

Po důkladné analýze dat pomocí různých metod shlukování - K-Means, GMM (Gaussian Mixture Models), Hierarchical Clustering a DBSCAN - můžeme říct:

1. EDA je zásadní krok, který poskytuje základní pochopení distribuce a vztahů mezi různými proměnnými. Provedením deskriptivní statistiky, korelační analýzy a vizualizace byly odhaleny klíčové indikátory na míru úmrtnosti novorozenců.

2. K-Means je jednoduchý a interpretovatelný algoritmus, který dobře fungoval pro identifikaci skupin zemí s podobnými charakteristikámi. Jeho hlavní nevýhodou je nutnost předem určit počet shluků a citlivost na outliery a šum.

3. GMM poskytl hlubší vhled do překrývání a nejistoty ve shlucích.

4. Hierarchical Clustering byl užitečný pro vizualizaci a interpretací vztahů mezi zeměmi, protože poskytl podrobný pohled na vztahy mezi zeměmi díky vizualizaci dendrogramu

5. DBSCAN vynikl při detekci outlierů a identifikaci shluků založených na hustotě, což bylo užitečné pro odhalení atypických zemí.

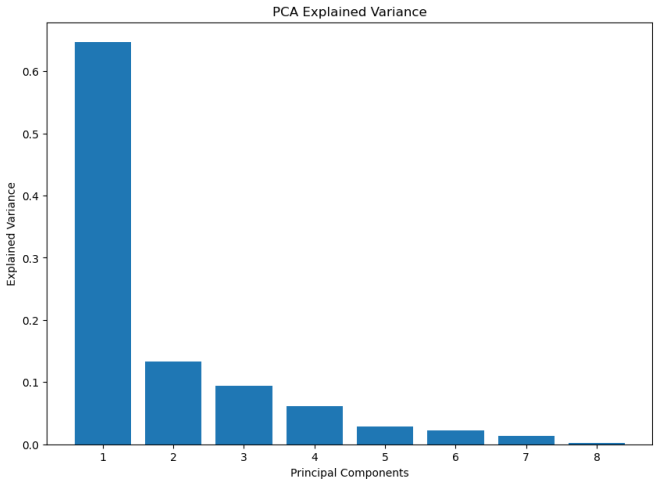
Vybrat "nejlepší" metodu však nakonec závisí na konkrétním cíli analýzy

V. Redukce dimenze

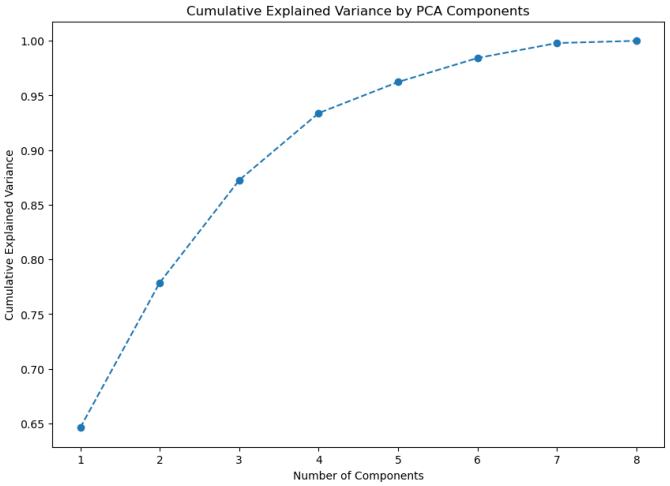
Pro redukci dimenze byly použity a srovnány metody distance preserving approaches (multidimensional scaling, Isomap, locally linear embedding, tSNE) a vzniklé grafy byly obarveny dle rozdělení klastrů pomocí **hierarchical clustering**.

a) PCA

Tento sloupcový graf ukazuje rozptyl vysvětlený každou hlavní komponentou. První komponenta vysvětluje významnou část rozptylu.



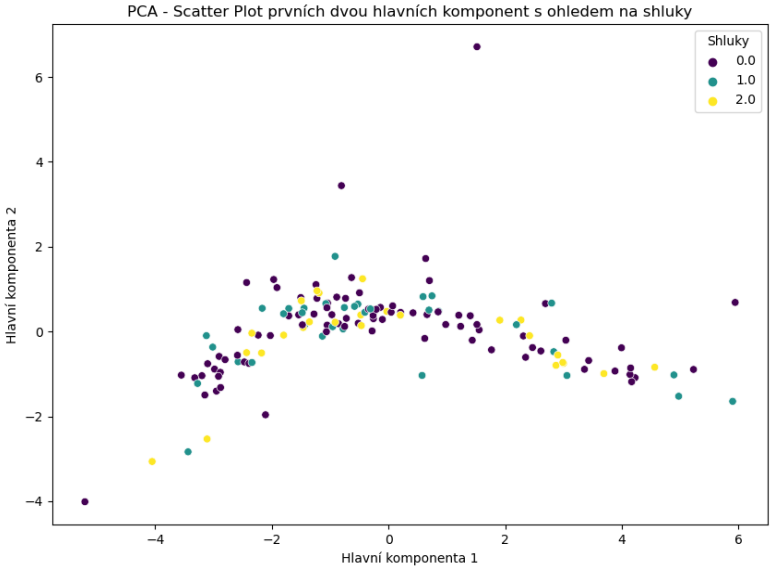
Tento graf pomáhá určit počet komponent, které je třeba zachovat. Křivka se po třetí složce zplošťuje, což naznačuje, že další složky přispívají k vysvětlení rozptylu minimálně. Proto bychom mohli zvážit snížení dimenzionality na tři složky.



Graf ukazuje příspěvek každé původní proměnné k hlavním komponentám. Indikátory blíže ke středu mají menší vliv na rozlišení dat podél PC1 a PC2, vzdálenější indikátory mají naopak větší vliv.

- Body jsou rozloženy hlavně podél první hlavní komponenty a méně podél druhé hlavní komponenty- **první hlavní komponenta zachycuje větší množství variance** v datech.
- Barevně naznačujeme tři různé shluky (označené jako 0.0, 1.0 a 2.0), které představují skupiny zemí s podobnými charakteristikami. Nicméně, **shluky se značně překrývají**, nejsou jasně oddělené, což znamená, že vybrané vlastnosti pro PCA mají překrývající se variance.

- Několik bodů vypadá jako **outliery**, zvláště podél první hlavní komponenty. Tyto body mohou představovat země s unikátními socioekonomickými a zdravotnickými profily (tohle jsme viděli i u clustering metod).



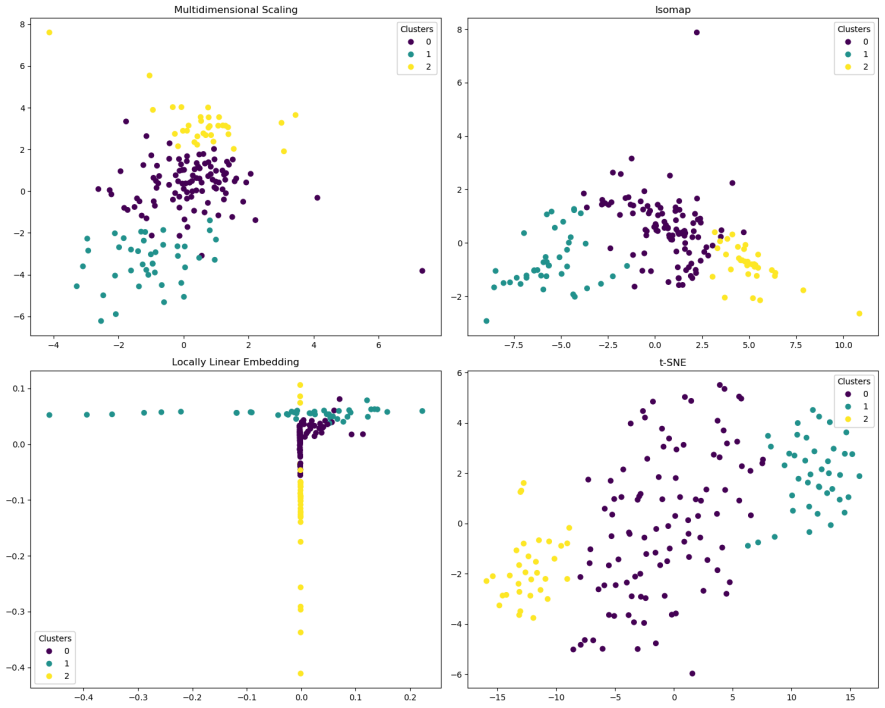
b) Multidimenzionální škálování, Isomap, t-SNE, Locally Linear Embedding

- **Multidimensionální škálování:** poskytuje vizualizaci podobnosti nebo rozdílů mezi různými zeměmi. Ukazuje tři shluky s určitým oddělením, ale s překryvem mezi shluky 0 a 1.

- **Isomap:** používá se pro odhad geodetických vzdáleností mezi body, což je užitečné pro zachování lokálních a globálních struktur. Na grafu shluky jsou více promíchané, což naznačuje menší rozlišení mezi nimi.

- **Locally Linear Embedding:** zaměřuje se na zachování pouze lokálních vzdáleností, může být vhodnější pro komplexní struktury. V grafu vidíme velmi hustý shluk 2 a více rozptýlený shluk 0, zatímco shluk 1 není dobře definován, rozprostírá mezi ostatními.

- **t-SNE:** efektivní pro vizualizaci vysokodimenzionálních dat v prostoru o dvou nebo třech dimenzích. Poskytuje nejlepší separaci shluků, i když mezi shluky 0 a 1 je stále určitý překryv.



SHRNUTÍ VÝSLEDKŮ REDUKCE DIMENZE : Jak je vidět z grafů shluky odpovídají pozicím bodů, hlavně metody Multidimensional scaling, isomap a t-SNE. Metody PCA a Locally Linear Embedding nerozdělily jednotlivé země tak dobře, pro PCA by bylo vhodnější využít nelineární kernel a tudíž byly brány v úvahu jen metody odpovídající clustering methods. Výsledná vizualizace ukázala, že t-SNE může být nejlepší volbou pro identifikaci skupin zemí s podobnými charakteristikami pro naš dataset.

VI. Porovnání výsledků s dostupnými studiemi

- **Socioeconomic inequalities in child mortality: comparisons across nine developing countries** https://pubmed.ncbi.nlm.nih.gov/10686730/

Tento článek se zaměřuje na 9 zemí, které jsou řazeny mezi rozvojové. Na rozdíl od naší statistiky vychází z dat mezi lety 1987-96 a bere v potaz děti do 5 let věku. My jsme brali v úvahu jen data dětí mezo 0 - 11 měsíci věku. Studie narazila na jeden stejný problém jako my a to, že pro země, které jsou řazeny k nejchudším jsou data buď úplně nedostupná nebo jen odhadnutá, což znemožňuje porovnání mezi zeměmi.

- **The Effects of GDP Per Capita on Infant Mortality Rates** https://scholars.fhsu.edu/cgi/viewcontent.cgi?article=1258&context=sacad#:~:text=In%20general%2C%20countries%20with%20higher,and%20well%2Dbeing%20of%20children

Výsledky této studie ukázaly, že HDP (GDP) je statisticky významnou proměnnou, pokud jde o míru kojenecké úmrtnosti. Výsledky ukázaly vyšší významnost při pohledu na málo rozvinuté země. Naše studie měla stejné výsledky v případě HDP, a to při globálním pohledu.

- **Poverty, urban-rural classification and term infant mortality: a population-based multilevel analysis** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6343321/

Tato studie byla provedena v USA, jejím objektem byli kojenci, jejichž data pocházejí z roku 2013. Cílem bylo odhadnout vliv chudoby a klasifikace města či venkova na kojeneckou úmrtnost. Bylo prokázáno, že vysoká chudoba a velmi venkovské okresy jsou spojeny s úmrtností donošených dětí nezávisle na jednotlivých sociodemografických, zdravotních a porodnických faktorech matek.

- **Comparing socioeconomic inequalities between early neonatal mortality and facility delivery: Cross-sectional data from 72 low- and middle-income countries**
  https://www.nature.com/articles/s41598-019-45148-5

Jedna z mála studií, která srovnávala socioekonomické faktory s využitím globálního pohledu: 72 zemí s nízkými a středními příjmy. Většina souborů dat (64 ze 72) byla shromážděna v roce 2000 nebo později. Jednou z částí této studie je kvantifikace asociací vyššího bohatství domácnosti a vyššího vzdělání matky s rizikem předčasného úmrtí novorozence, a to pro srovnání s postneonatální úmrtností kojenců. Studie zjistila malé socioekonomické nerovnosti v časné novorozenecké úmrtnosti, země v různých oblastech světa se chovaly odlišně, což vede k myšlence, že je třeba vzít v úvahu přesnější faktory. Na rozdíl od malých socioekonomických nerovností v časné novorozenecké úmrtnosti byly ve většině zemí velké rozdíly v porodnosti v zařízeních v souvislosti s bohatstvím domácností a vzděláním matek. Nerovnosti související se vzděláním byly ve srovnání s nerovnostmi souvisejícími s bohatstvím o něco větší.

- **Subnational variations in electricity access and infant mortality: Evidence from Ghana**
  https://www.sciencedirect.com/science/article/pii/S2590229621000289#s0035

Studie zkoumala vztah mezi přístupem k elektřině a kojeneckou úmrtností na nižší než národní úrovni v Ghaně, přičemž kontrolovala korelační faktory, jako je interval mezi porody, děti žijící s oběma rodiči, vzdělání žen a rozdělení příjmů. Výsledky ukazují, že v regionech s nízkým výskytem kojenecké úmrtnosti snižuje 10% zlepšení přístupu k elektřině kojeneckou úmrtnost o 11,8 na 1 000 živě narozených dětí, zatímco v regionech s vysokou úmrtností nemá zlepšení přístupu k elektřině na kojeneckou úmrtnost žádný vliv. Rozdělení bohatství nemá na kojeneckou úmrtnost v regionech s nízkou úmrtností žádný vliv, ale v regionech s vysokou úmrtností došlo jak u nejbohatších, tak u nejchudších obyvatel k výraznému poklesu kojenecké úmrtnosti.

VI. Shrnutí celé práce:

Během čištění dat byl odstraněny řádky s chybějícími hodnotami a duplicity, což zmenšilo datovou sadu na 175 řádků a 9 sloupců. To bylo provedeno, protože počáteční heatmapy ukázaly významné chybějící hodnoty v některých sloupcích, zejména v GDP a Access to electricity.

Dál jsme provedli **EDA** s cílem vizualizovat distribucí a boxplotů pro indikátory. Byly identifikovány výrazné korelace, např. mezi vyšším HDP a nižší úmrtností novorozenců.

Při porovnání **lineární a polynomiální regresních modelů** obecně jsme dostali lepší výsledky pro polynomiální, například u GDP s R-kvadrátem 0.329 oproti 0.208 u lineárního modelu.

Při využití **Poissonova modelu** nejlepší model s nejnižší hodnotou AIC (1118.57) zahrnoval proměnné jako GDP, Life Expectancy a Total Fertility Rate. Model efektivně zachytil trend v datové sadě.

Logistická regrese, LDA a QDA byly použity pro binární klasifikaci úmrtnosti novorozenců. QDA dosáhla nejvyšší přesnosti a ROC-AUC skóre, zatímco logistická regrese měla nejvyšší přesnost.

Použili jsme **K-Means, GMM, Hierarchical Clustering a DBSCAN.**

- K-Means: Optimalizace pomocí Loketní metody identifikovala 3 shluky.
- GMM: Upraveno na 3 shluky, s hodnotou BIC 1248.51.
- Hierarchické shlukování: Použití aglomerativní metody s 3 shluky, s hodnotou BIC 1246.33.
- DBSCAN: Identifikoval 2 hlavní shluky a několik zemí jako šum.

DBSCAN byl méně účinný kvůli rozložení dat. Choropletové mapy ukázaly vysoký ekonomický a zdravotní stav ve vyspělých regionech a problémy v některých oblastech Afriky a Asie.

Nakonec jsme použili metody jako **PCA, Multidimensional Scaling, Isomap, t-SNE a Locally Linear Embedding.** PCA ukázala, že první tři komponenty vysvětlují většinu variance. t-SNE poskytla nejlepší separaci shluků.

Naše analýza odhalila významné korelace mezi socioekonomickými a zdravotními indikátory a kojeneckou úmrtností. **Výsledky ukázaly, že ekonomický rozvoj, délka života a přístup k elektřině jsou klíčové faktory ovlivňující kojeneckou úmrtnost.** Metody jako K-Means, GMM a Hierarchical Clustering pomohly identifikovat různé skupiny zemí, zatímco DBSCAN byl užitečný pro identifikaci outlierů.

VII. Prohlášení o příspěvku:

Na práci jsme spolupracovali celá skupina. Každá z nás udělala návrhy na výzkumnou otázku a následně jsme po společné konzultaci jednu vybraly. Následovala společná diskuze o postupu a vyhledávání vhodných datasetů. Jednotlivé části práce jsme tvořily přibližně takto:

- vyhledávání vhodných datasetů: Svitlana, Marina, Ivana, Klára
- zvolení postupu k vyhodnocení výzkumné otázky - příprava a konzultace workplanu: Svitlana, Marina, Ivana, Klára
- cleaning dat, příprava a spojování datasetů: Ivana, Klára
- programování kódu a vizualizace Python: Svitlana, Ivana
- vizualizace v R: Klára
- porovnání výsledků se studiemi: Ivana, Marina
- sepsání reportu: Svitlana, Marina, Ivana, Klára

Celkové rozložení práce v procentech:

- Svitlana 25 %
- Ivana 25 %
- Klára 25 %
- Marina 25 %