Term project - Guns vs Butter

Bence Szabo (2508584)

Research question, hypothesis

While investigating the potential negative association between military expansion and the funding of inequality-reducing social welfare programs - specifically healthcare, does the prioritization of military expenditure crowd out public healthcare investments, thereby limiting the state's capacity to reduce social inequality?

My hypothesis is that increased military spending has a crowd-out effect on public healthcare investment, therefore increasing social inequality.

1. Introduction

Access to healthcare is a very important factor that determines income and social inequality. Public healthcare spending is one of the major mechanisms governments use to redistribute wealth and provide a safety net for lower-income populations. Government have limited budget and high military spending (variable x) often competes directly with social/healthcare spending (variable y). Therefore, by analyzing if military spending reduces healthcare spending, we are analyzing a structural driver of inequality.

While inequality is often measured by income gaps, it is maintained by the lack of access to essential public services. Public healthcare expenditure is a key proxy for a state's commitment to reducing social inequality. My analysis investigates whether higher military expenditure creates a crowding-out effect, effectively reducing the resources available for health programs reducing inequality.

By exploring the negative relationship that might exist between military expansion spending and the allocation of funds that can reduce inequality, with particular emphasis on healthcare, does military spending crowd out government spending on healthcare, therefore weakening the options of states in reducing inequality?

2. Data

I have downloaded data from the World Bank website, selecting the following series:

- Dependent variable (y): Current health expenditure (% of GDP),
- Main explanatory variable (x): Military expenditure (% of GDP),
- Development indicators: GDP growth (annual %), GDP per capita (current US$),
- Population indicator: Population ages 65 and above (% of total population),
- Fiscal policy indicator: Government revenue, excluding grants (% of GDP),
- Urbanization indicator: Urban population (% of total population).

All data was from 2022 and I have also downloaded the same data for 2014 for external validation.

3. Model

3.1. Feature engineering

After importing the data I have converted the missing data (denoted with value "." in the .csv) to null values, converted all numbers to float type and dropped observations where the y or the x variable value was missing. This left me with 298 observations. After this I created dummy variables from the population % aged 65 and above (0-5%, 5-10%, 10-20%, 20+%)

and the urban population % (0-20%, 20-40%, 40-60%, 60-80%, 80+%). For the GDP per capita I created income groups of Low, Lower-Mid, Upper-Mid and High by splitting the data into 4 equal-sized groups using quartiles which ensured that I had enough observations in every category. I also created their respective categories in string variables **(table 2)**.

After this I split the dataset into two datasets, one for 2022 (main) and one for 2014 (for external validation).

3.2. Analysis of the numeric explanatory variables

First I looked at the histograms of the numeric variables **(figure 1)**. The health expenditure is centered around 6.7% with two extremes above 14% - the USA and Timor-Leste. Military expenditure is right-skewed and centered around 1.8% with most of the countries spending less than 3% of their GDP on the military, Qatar and Saudi Arabia being the extremes with more than 6%. The distribution of the population aged 65+ years has two peaks, 4% and 20%, most probably because of the least developed and most developed countries, respectively.

3.3. Analysis of the relationship with the target variable

After this I looked at the relationship of the variables with the target variable by creating scatterplots **(figure 2)** in order to answer our initial research question. Looking at the scatterplot of health expenditure and military spending, a slight negative association can be seen but we cannot expect a strong statistical relationship between the two variables. However, we can expect a strong statistical relationship between healthcare spending and the proportion of the population aged 65+ years. In the case of urbanization, there is also a slight positive correlation between the proportion of the population living in urban areas and healthcare spending.

Next I checked the categorical variables with boxplots **(figure 3)**. In case of urbanization and population age, we can see the same pattern as before from the scatterplots. Regarding the income category of countries we can see that countries with the highest income spend considerably more on healthcare than the rest of them.

Lastly I created a correlation matrix of numeric variables **(figure 4)**, including the dummy variables. The main observations regarding the relationship with the dependent variable are the following:

1. The key variable (military expenditure) has a low negative correlation with health expenditure so we cannot expect strong evidence in the regression for the original hypothesis that increased military spending increases social inequality by reducing the available budget for healthcare expenditure.
2. High government revenue countries have higher healthcare expenditure (53% correlation).
3. Countries with less than 5% of population aged higher than 65 years have lower healthcare expenditure (-37% correlation).
4. Countries with higher proportion (10%+) of population aged higher than 65 years have higher healthcare expenditure (34% correlation).
5. Where the GDP per capita is high, healthcare expenditure is also high (38% correlation).

I also checked correlation between the control variables. Higher correlation can be seen in the following cases:

6. Countries with high percent of the population living in urban areas have higher GDP per capita (56% correlation).

7. Low income countries have low percent of population aged higher than 65 years (58% correlation).

3.4. Regression models

I ran 7 regressions starting from the simplest to the more complex.

In the first regression I looked at the relationship between health expenditure and military expenditure. As expected, there is a slight negative association between them, but this is not statistically significant. The intercept for this model is 6.939, which means that if there is no military expenditure (0% of GDP), the expected expenditure on healthcare would be about 6.94% of GDP. The slope is -0.130, which means that for each 1 percentage point increase in military expenditure, health expenditure is predicted to decrease by 0.13 percentage points on average, so there is a slight crowd-out effect but it is not statistically significant.

							Dependent variable: health_expenditure_pct_gdp
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	6.939^{***}	7.514^{***}	6.303^{***}	5.759^{***}	5.408^{***}	4.316^{***}	3.786^{***}
	(0.407)	(0.457)	(0.614)	(0.567)	(1.099)	(0.895)	(0.753)
gdp_growth_pct_annual		-0.111^{***}					
		(0.038)					
gov_revenue_gdp_pct						0.098^{**}	0.105^{**}
						(0.048)	(0.042)
high_income			2.808^{***}	-0.667	-1.601^{*}	-0.426	
			(0.653)	(0.736)	(0.923)	(1.236)	
lower_middle_income			0.021	-0.459	-0.682	-0.677	
			(0.652)	(0.654)	(0.777)	(0.913)	
military_expenditure_pct_gdp	-0.130	-0.150	-0.279	-0.040	-0.094	-0.032	-0.033
	(0.179)	(0.181)	(0.196)	(0.141)	(0.149)	(0.217)	(0.216)
old_pop_pct_10_20				3.792^{***}	3.792^{***}	2.396^{***}	1.629^{**}
				(0.640)	(0.622)	(0.897)	(0.750)
old_pop_pct_20_plus				4.615^{***}	4.745^{***}	2.257^{**}	1.595^{*}
				(0.603)	(0.618)	(1.022)	(0.950)
old_pop_pct_5_10				0.844	0.782	0.792	
				(0.539)	(0.557)	(0.819)	
upper_middle_income			0.597	-1.813^{***}	-2.371^{***}	-1.967^{*}	
			(0.611)	(0.621)	(0.770)	(1.027)	
urban_pop_pct_20_40					0.343		
					(1.283)		
urban_pop_pct_40_60					0.601		
					(1.237)		
urban_pop_pct_60_80					0.937		
					(1.304)		
urban_pop_pct_80_100					1.648		
					(1.322)		
Observations	147	145	147	147	147	96	96
R^2	0.003	0.063	0.164	0.342	0.358	0.406	0.344
Adjusted R^2	-0.004	0.050	0.140	0.309	0.305	0.351	0.315
Residual Std. Error	2.894 (df=145)	2.830 (df=142)	2.679 (df=142)	2.401 (df=139)	2.408 (df=135)	2.242 (df=87)	2.304 (df=91)
F Statistic	0.521 (df=1; 145)	4.548^{**} (df=2; 142)	8.325^{***} (df=4; 142)	16.954^{***} (df=7; 139)	11.665^{***} (df=11; 135)	10.104^{***} (df=8; 87)	17.353^{***} (df=4; 91)
Note:							$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 1 – Regression table 2022

In the rest of the regressions I added more control variables. The conclusions are the following:

1. First I added the income indicators (regressions 2 & 3), GDP growth % and GDP per capita income groups. We can see that in case of model 3 the explanatory power of the model increased to 16% and the direction of the parameters are as expected. The estimated parameter of 2.808 for high-income countries means that compared to the baseline low-income countries with high income, on average, spend almost 3 percentage point more of their GDP on healthcare.

2. In regression 4 I added the age-related dummy variables. As expected, the parameters of age are statistically significant - if we look at countries with higher percent (10-20%, 20+%) of older population we can expect that they spend, on average, 3.792 and 4.615 percentage points more on healthcare than countries with 0-5% old population (baseline). The explanatory power of the model also increased to about 30%.
3. Adding the urban population dummy variables in regression 5 only improved the explanatory power by 1 percentage point, while the coefficients are not significant (as was expected from the scatterplot). Therefore I will not include it as a control variable in the next model due to the low number of observations.
4. Next I added the government revenue indicator. As expected, governments with higher revenue are expected to spend more on healthcare, for example in case of welfare states.
5. In the last regression I took out those variables which were not significant and the explanatory power of the module remained around 35%.

In conclusion, I would choose model 6 or model 7.

- Model 6: It has all the important drivers, the explanatory power is the highest with 40.6%, however due to the high number of control variables there is a risk of overfitting with only 96 observations.
- Model 7: It has lower explanatory power with 34.4% but by dropping some controls there is a lower risk of overfitting.

Additionally, I ran the same regressions **(table 3)** only on the upper-middle and high income countries (77 observations). On 10% and 5% significance levels, the beta coefficient of military expenditure becomes statistically significant, which indicates that our initial hypothesis - that military spending leads to a crowd-out effect, lowering public healthcare spending - is possibly true in countries with higher income.

However, looking at regression models 4-7 is not meaningful due overfitting on a low number of observations (62).

4. External validity

To assert the external validity of the previous results, I ran the 7 regressions on another year **(table 4)**. I chose year 2014 to allow enough time for policy changes. For regressions 1-3 the parameter of military expenditure is statistically significant, but after including age group dummy variables there is no evidence that the higher spending on military leads to lower spending on health and the explanatory power of the models did not change substantially.

I also ran the regressions on the 2014 data with only upper-middle and high income countries **(table 5)**, which yielded similar results as the regressions on the 2022 data filtered for upper-middle and high income countries.

5. Summary

To summarize the findings, analyzing the 2022 data I could not find statistical evidence that confirms my hypothesis that increased military expenditure leads to lower healthcare spending. Only looking at the countries with higher income the military expenditure had a slight crowd-out effect on healthcare spending on 10% significance level. I found that the main driver of the healthcare spending is the high percent of people older than 65 years in the population. In order to achieve more robust results, moving beyond 2022 to use time series cross-sectional data would be the next step.

Appendix

skimpy summary

Data Summary

Dataframe	Values
Number of rows	147
Number of columns	30

Data Types

Column Type	Count
int64	15
float64	10
string	4
category	1

Categories

Categorical Variables
income_group

number

column	NA	NA %	mean	sd	p0	p25	p50	p75	p100	hist
index	0	0	224	42.58	151	187.5	224	260.5	297	
year	0	0	2022	0	2022	2022	2022	2022	2022	
health_expenditure_pct_gdp	0	0	6.706	2.889	1.818	4.282	6.659	8.825	16.5	
health_expenditure_percap_usd	0	0	1386	2230	16.32	89.34	421	1570	12430	
gdp_growth_pct_annual	2	1.3605442176870748	4.735	6.371	-20.54	2.673	4.242	5.868	63.33	
gdp_percap_usd	2	1.3605442176870748	17640	24480	250.6	2323	6730	21330	123700	
military_expenditure_pct_gdp	0	0	1.797	1.26	0.06874	0.9561	1.452	2.238	6.962	
military_expenditure_usd	0	0	14580000000	75750000000	4627000	235900000	924800000	5255000000	860700000000	
pop_age_above_65_pct_total_pop	0	0	9.871	7.125	1.391	3.511	7.186	16.48	29.4	
total_pop	0	0	51110000	169900000	119900	4223000	11230000	35170000	1425000000	
gov_revenue_gdp_pct	51	34.693877551020411	26.62	10.26	5.056	17.76	26.69	33.42	56.29	
urban_pop_pct_total_pop	0	0	61.45	22.07	13.58	45.6	63.56	79.59	100	
old_pop_pct_0_5	0	0	0.3741	0.4856	0	0	0	1	1	
old_pop_pct_5_10	0	0	0.2381	0.4274	0	0	0	0	1	
old_pop_pct_10_20	0	0	0.2585	0.4393	0	0	0	1	1	
old_pop_pct_20_plus	0	0	0.1293	0.3366	0	0	0	0	1	
urban_pop_pct_0_20	0	0	0.04082	0.1985	0	0	0	0	1	

urban_pop_pct_20_40	0	0	0.1565	0.3645	0	0	0	0	1	
urban_pop_pct_40_60	0	0	0.2789	0.45	0	0	0	1	1	
urban_pop_pct_60_80	0	0	0.2721	0.4466	0	0	0	1	1	
urban_pop_pct_80_100	0	0	0.2517	0.4355	0	0	0	0.5	1	
low_income	0	0	0.2177	0.4141	0	0	0	0	1	
lower_middle_income	0	0	0.2449	0.4315	0	0	0	0	1	
upper_middle_income	0	0	0.2585	0.4393	0	0	0	1	1	
high_income	0	0	0.2653	0.443	0	0	0	1	1	

category

column	NA	NA %	ordered	unique
income_group	2	1.3605442176870748	True	5

string

column	NA	NA %	shortest	longest	min	max	chars per row	words per row	total words
country_name	0	0	Chad	Central African Republic	Albania	Zimbabwe	8.59	1.3	184
country_code	0	0	ALB	ALB	AGO	ZWE	3	1	147
old_pop_cat_string	0	0	1: 0-5%	3: 10-20%	1: 0-5%	4: 20+%	7.76	2	294
urban_pop_cat_string	0	0	5: 80+%	4: 60-80%	1: 0-20%	5: 80+%	8.46	2	294

End

Table 2 – summary table of dataset 2022 after feature engineering

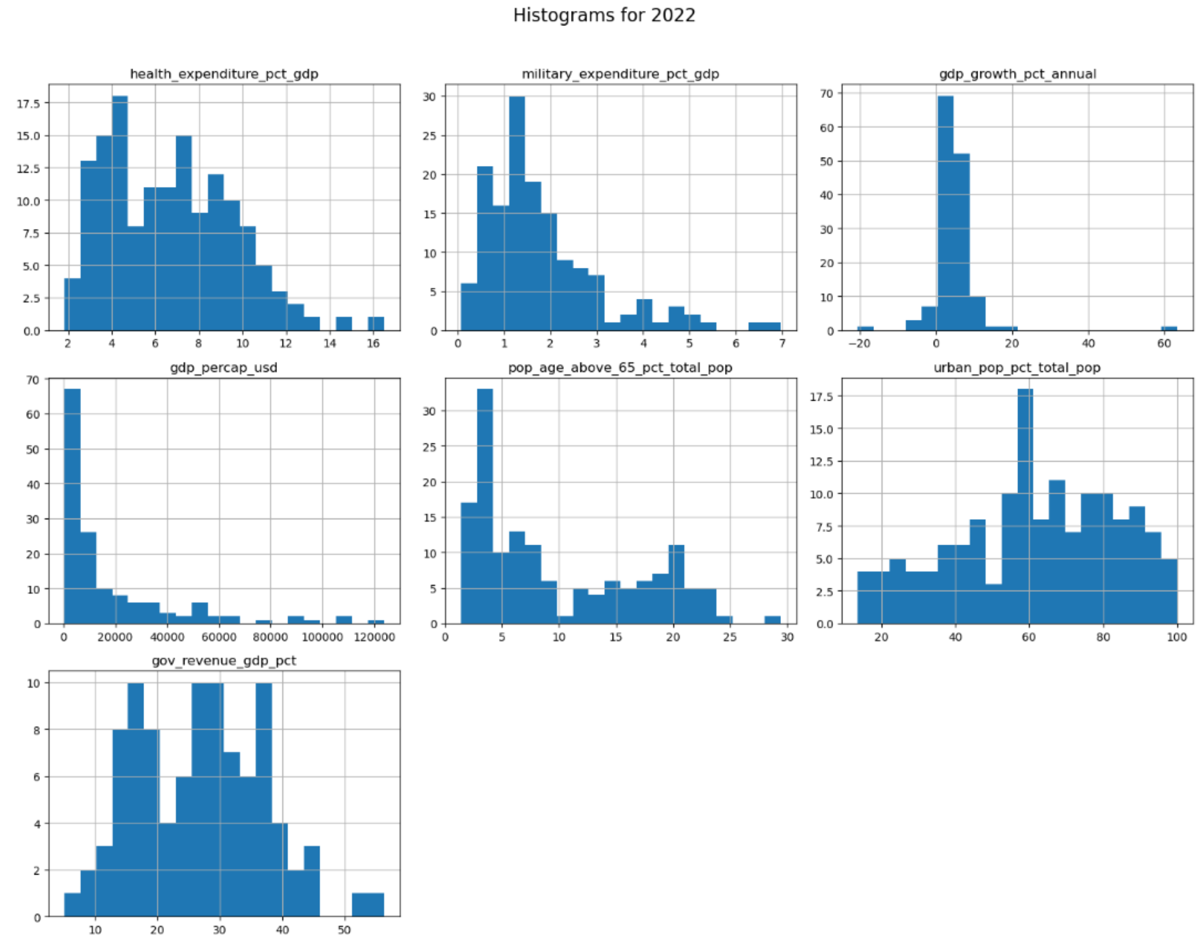
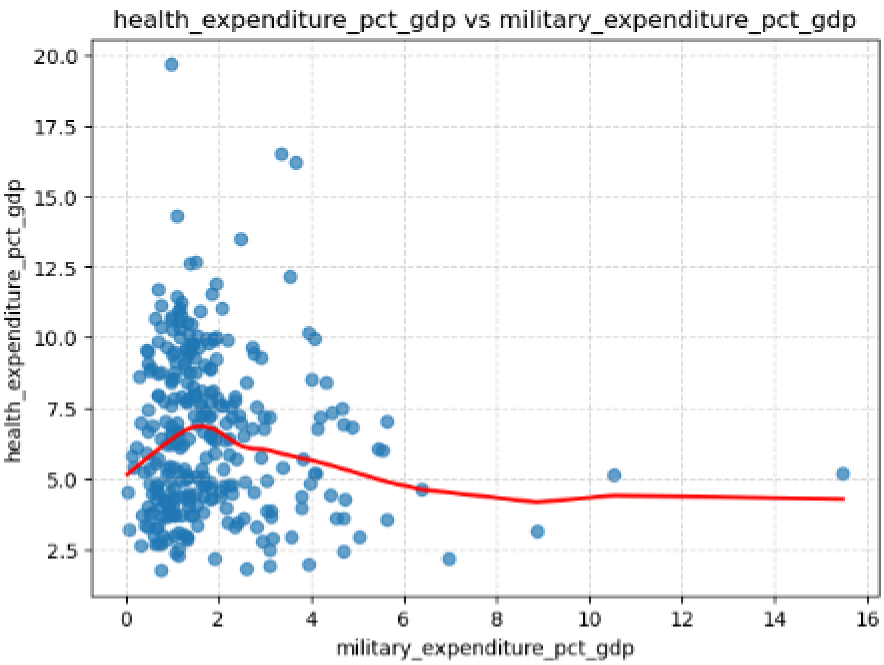
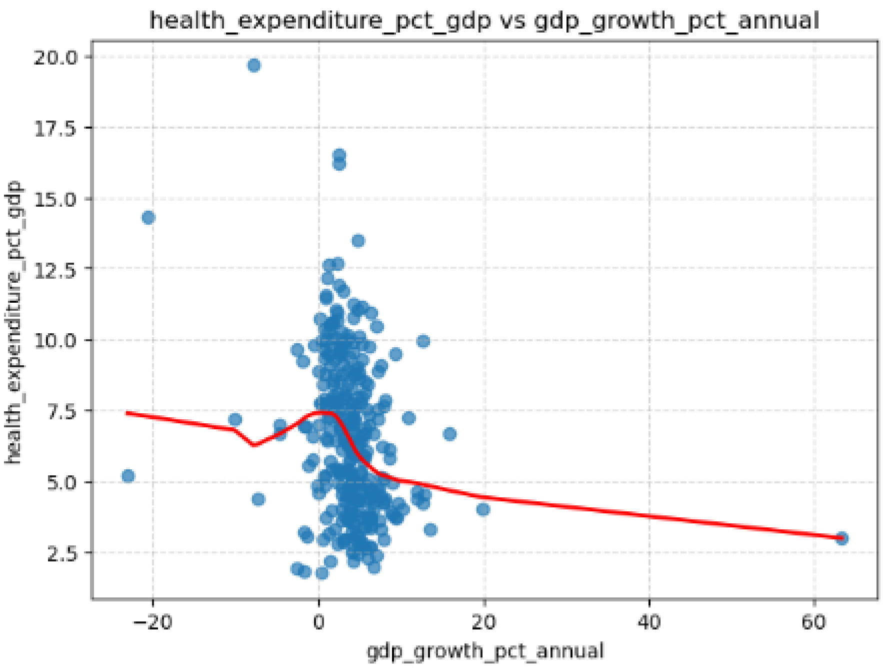
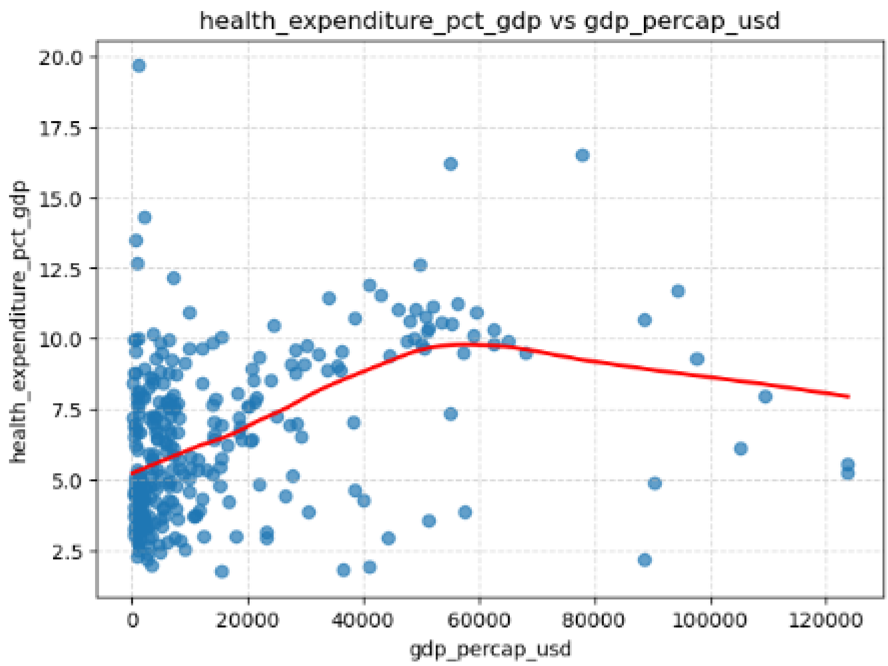
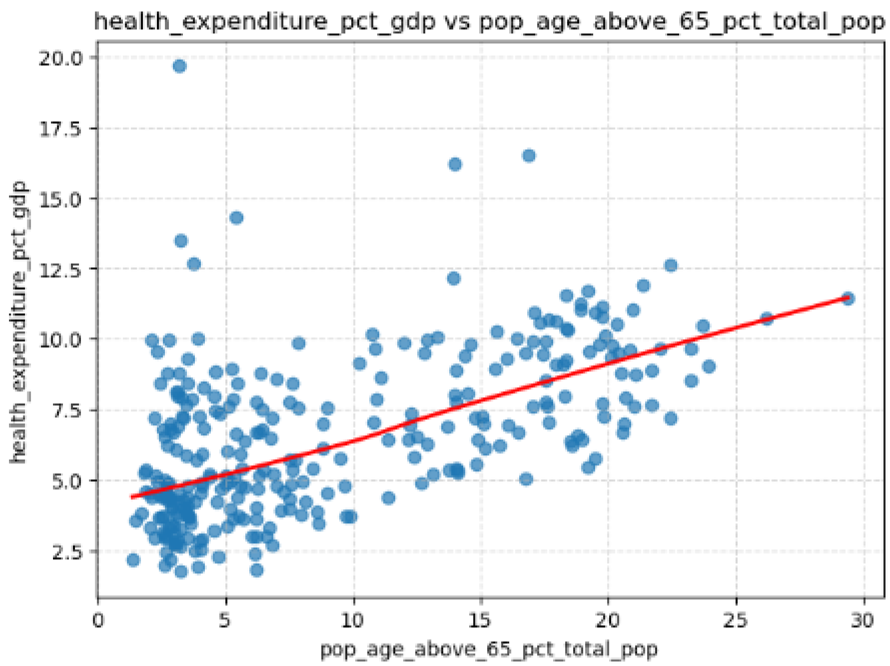


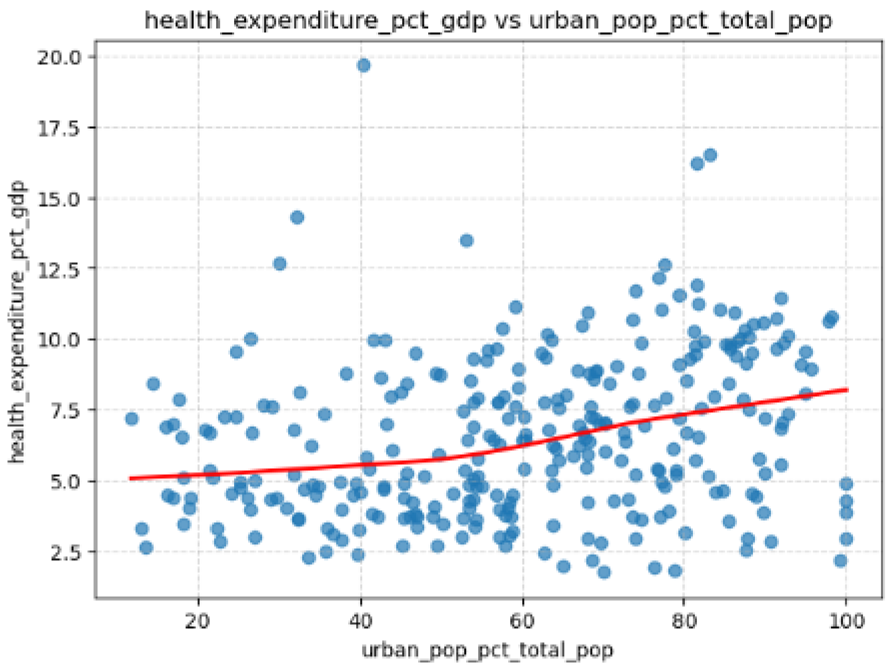
Figure 1 – Histograms of the numeric explanatory drivers


health_expenditure_pct_gdp vs military_expenditure_pct_gdp
health_expenditure_pct_gdp
military_expenditure_pct_gdp
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
2
4
6
8
10
12
14
16


health_expenditure_pct_gdp vs gdp_growth_pct_annual
health_expenditure_pct_gdp
gdp_growth_pct_annual
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
−20
0
20
40
60


health_expenditure_pct_gdp vs gdp_percap_usd
health_expenditure_pct_gdp
gdp_percap_usd
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
20000
40000
60000
80000
100000
120000


health_expenditure_pct_gdp vs pop_age_above_65_pct_total_pop
health_expenditure_pct_gdp
pop_age_above_65_pct_total_pop
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
5
10
15
20
25
30



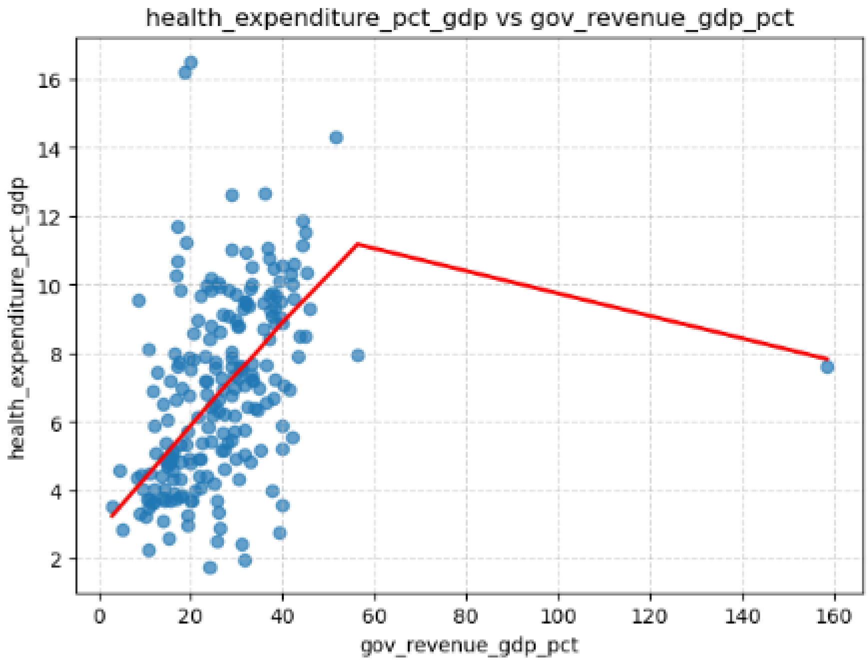


Figure 2 – Scatterplots of the numeric explanatory variables

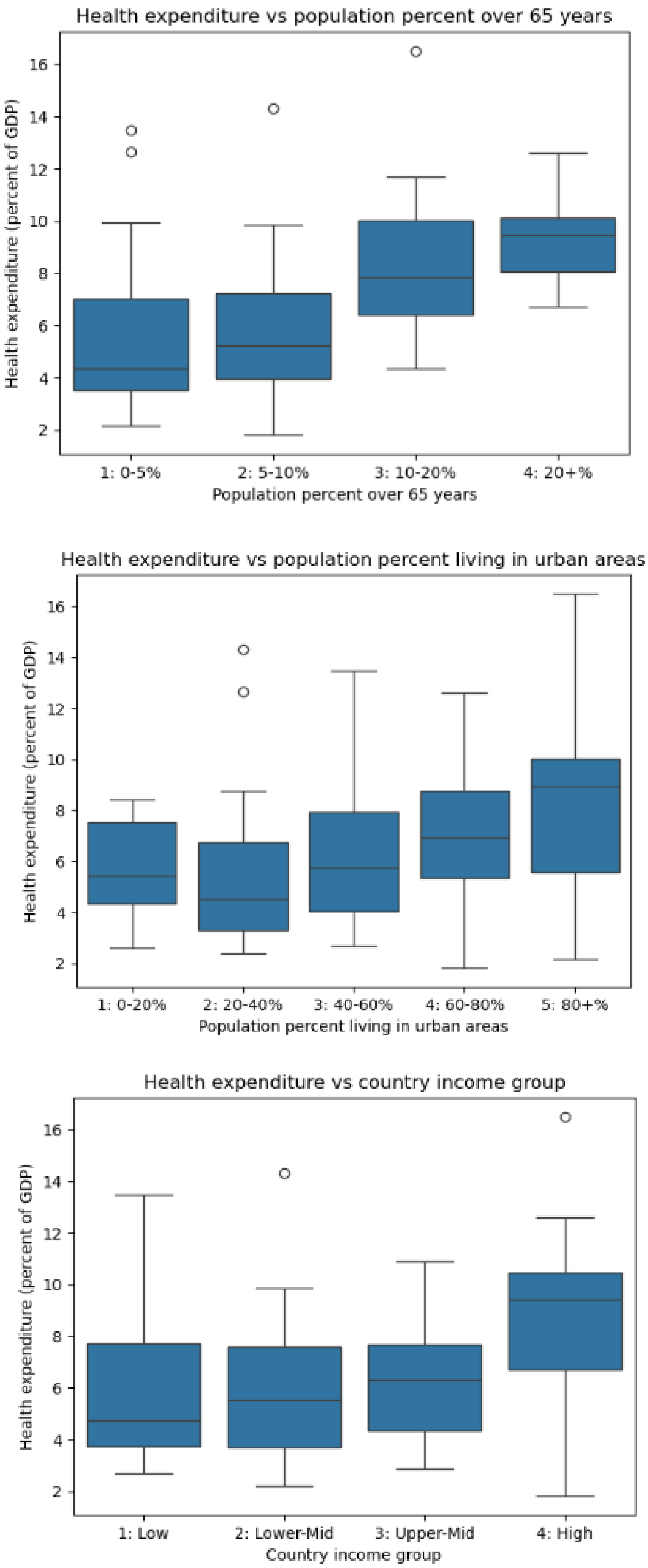


Figure 3 – Boxplots of the categorical explanatory variables

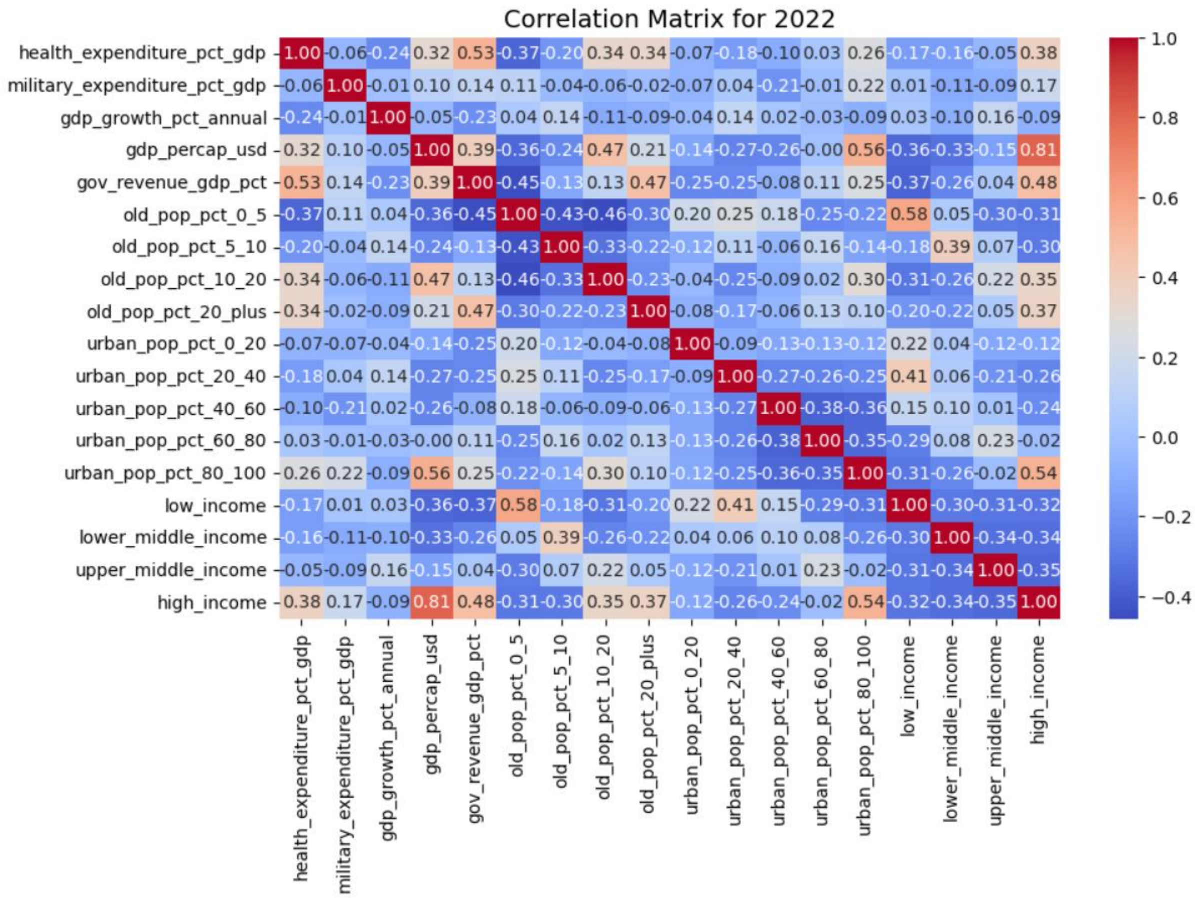


Figure 4 – Correlation matrix

	Dependent variable: health_expenditure_pct_gdp						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	8.270***	8.780***	5.736***	2.806***	2.042***	3.326***	4.614***
	(0.551)	(0.559)	(0.338)	(0.490)	(0.468)	(0.999)	(1.043)
gdp_growth_pct_annual		-0.094***					
		(0.021)					
gov_revenue_gdp_pct						-0.013	0.016
						(0.050)	(0.046)
high_income			4.060***	1.978***	1.355***	2.616***	
			(0.351)	(0.360)	(0.383)	(0.688)	
lower_middle_income			0.000***	0.000	-0.000**	-0.000	
			(0.000)	(0.000)	(0.000)	(0.000)	
military_expenditure_pct_gdp	-0.412*	-0.428*	-0.597**	-0.125	-0.202	0.033	0.011
	(0.246)	(0.239)	(0.259)	(0.196)	(0.197)	(0.281)	(0.263)
old_pop_pct_10_20				4.435***	4.480***	3.750***	3.419***
				(0.659)	(0.725)	(1.114)	(0.821)
old_pop_pct_20_plus				5.070***	5.310***	4.265***	3.951***
				(0.581)	(0.681)	(1.173)	(0.932)
old_pop_pct_5_10				1.207	1.429	1.469	
				(0.873)	(0.885)	(1.246)	
upper_middle_income			1.676***	0.829**	0.687**	0.709	
			(0.326)	(0.350)	(0.337)	(0.480)	
urban_pop_pct_20_40					-1.030**		
					(0.450)		
urban_pop_pct_40_60					0.333		
					(0.366)		
urban_pop_pct_60_80					0.904**		
					(0.416)		
urban_pop_pct_80_100					1.836***		
					(0.324)		
Observations	77	77	77	77	77	62	62
R^2	0.036	0.094	0.201	0.542	0.582	0.435	0.323
Adjusted R^2	0.023	0.070	0.179	0.510	0.533	0.373	0.275
Residual Std. Error	2.857 (df=75)	2.788 (df=74)	2.618 (df=74)	2.023 (df=71)	1.976 (df=68)	2.040 (df=55)	2.194 (df=57)
F Statistic	2.809* (df=1; 75)	11.457*** (df=2; 74)	236.192*** (df=2; 74)	252.947*** (df=5; 71)	110.675*** (df=8; 68)	214.082*** (df=6; 55)	11.357*** (df=4; 57)
Note:							*p<0.1; **p<0.05; ***p<0.01

Table 3 – Regression table 2022 for higher income countries

	Dependent variable: health_expenditure_pct_gdp						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	6.702***	8.288***	6.067***	5.623***	5.413***	5.166***	4.819***
	(0.298)	(0.585)	(0.506)	(0.521)	(0.567)	(0.440)	(0.324)
gdp_growth_pct_annual		-0.330***					
		(0.092)					
gov_revenue_gdp_pct						0.016*	0.019**
						(0.010)	(0.008)
high_income			2.758***	-0.474	-1.815	0.552	
			(0.653)	(0.687)	(1.153)	(0.656)	
lower_middle_income			-0.068	-1.147*	-1.715*	-0.617	
			(0.574)	(0.589)	(0.998)	(0.546)	
military_expenditure_pct_gdp	-0.171*	-0.435***	-0.245**	-0.006	-0.035	-0.073	-0.041
	(0.090)	(0.152)	(0.109)	(0.086)	(0.095)	(0.136)	(0.129)
old_pop_pct_10_20				3.896***	3.915***	3.115***	3.081***
				(0.550)	(0.558)	(0.596)	(0.447)
old_pop_pct_20_plus				4.432***	4.805***	3.084***	3.818***
				(0.787)	(0.797)	(0.879)	(0.684)
old_pop_pct_5_10				0.832*	0.943*	0.556	
				(0.480)	(0.483)	(0.530)	
upper_middle_income			0.656	-1.672***	-2.558**	-1.198*	
			(0.583)	(0.645)	(1.107)	(0.612)	
urban_pop_pct_20_40					0.046		
					(0.658)		
urban_pop_pct_40_60					0.702		
					(0.954)		
urban_pop_pct_60_80					1.051		
					(0.932)		
urban_pop_pct_80_100					1.882*		
					(1.056)		
Observations	151	151	151	151	151	118	118
R^2	0.013	0.169	0.172	0.371	0.388	0.484	0.427
Adjusted R^2	0.006	0.158	0.149	0.340	0.340	0.447	0.407
Residual Std. Error	2.797 (df=149)	2.575 (df=148)	2.588 (df=146)	2.278 (df=143)	2.279 (df=139)	1.912 (df=109)	1.980 (df=113)
F Statistic	3.619* (df=1; 149)	7.738*** (df=2; 148)	8.493*** (df=4; 146)	19.033*** (df=7; 143)	14.065*** (df=11; 139)	15.549*** (df=8; 109)	29.008*** (df=4; 113)
Note:							*p<0.1; **p<0.05; ***p<0.01

Table 4 – Regression table 2014

							Dependent variable: health_expenditure_pct_gdp
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	7.849***	8.570***	5.269***	1.841***	1.351***	2.901***	5.226***
	(0.390)	(0.529)	(0.244)	(0.608)	(0.513)	(0.695)	(0.906)
gdp_growth_pct_annual		-0.181**					
		(0.079)					
gov_revenue_gdp_pct						-0.016	-0.009
						(0.038)	(0.039)
high_income			3.692***	1.642***	1.208***	2.432***	
			(0.344)	(0.352)	(0.334)	(0.485)	
lower_middle_income			-0.000	-0.000***	0.000***	-0.000	
			(0.000)	(0.000)	(0.000)	(0.000)	
military_expenditure_pct_gdp	-0.286***	-0.437***	-0.304**	0.132	0.113	-0.015	-0.011
	(0.102)	(0.129)	(0.136)	(0.115)	(0.119)	(0.173)	(0.160)
old_pop_pct_10_20				5.517***	5.633***	4.541***	3.611***
				(0.898)	(0.908)	(1.152)	(0.820)
old_pop_pct_20_plus				5.903***	6.228***	4.582***	4.414***
				(0.953)	(0.981)	(1.268)	(0.959)
old_pop_pct_5_10				2.915***	3.217***	2.200**	
				(0.964)	(0.994)	(1.054)	
upper_middle_income			1.578***	0.199	0.144	0.469	
			(0.308)	(0.457)	(0.396)	(0.418)	
urban_pop_pct_20_40					0.000***		
					(0.000)		
urban_pop_pct_40_60					-0.280		
					(0.440)		
urban_pop_pct_60_80					0.498		
					(0.347)		
urban_pop_pct_80_100					1.133***		
					(0.330)		
Observations	71	71	71	71	71	62	62
R^2	0.060	0.100	0.202	0.568	0.592	0.541	0.414
Adjusted R^2	0.046	0.074	0.178	0.535	0.547	0.491	0.373
Residual Std. Error	2.752 (df=69)	2.711 (df=68)	2.554 (df=68)	1.921 (df=65)	1.896 (df=63)	1.924 (df=55)	2.135 (df=57)
F Statistic	7.888*** (df=1; 69)	5.759*** (df=2; 68)	216.686*** (df=2; 68)	205.188*** (df=5; 65)	166.191*** (df=7; 63)	177.962*** (df=6; 55)	15.686*** (df=4; 57)
Note:							*p<0.1; **p<0.05; ***p<0.01

Table 5 – Regression table 2014 for higher income countries