Brandeis University

International Business School

Final Report

Bus 211A-1 - Foundations of Data Analytics

Linoy Noikas, Salman Haider, Sana Ijaz, Kyle Allsopp

November 8th, 2023

Table of Contents

Part 1: Business Proposition

Our consultancy specializes in public health and health care data analysis. Our goal is to understand health care and educate the public on the health care situation across the globe. Our aim is to provide recommendations to improve healthcare outcomes and improve population well being at large through our data exploration and analysis. We optimize in resource allocation, shaping policy development, and addressing global disparities.

Our consultancy wanted to focus on an important social and wellbeing issue and knowing that a common interest was Healthcare Analytics, we decided to focus on this credible dataset from the World Bank. This source is widely recognized and respected in the global health area and has an international reputation that we can trust. We can be rest assured that this data is legitimate and captures most countries of the world. We have extracted several databases from the World Bank such as Life Expectancy by Countries, and Nominal GDP by Countries and Access to Clean Water. Using this data from the World Bank will provide us with a more accurate, and well trusted output.

We will be investigating human life expectancy by country for the European Union for our final presentation. The European Union is a partnership of 27 European countries from across Europe who work together to promote shared values and prosperity. Our data is from the years 2001-2019 and looking at different factors that affect it (Prevelence of Undernourishment, CO2, Health Expenditure %, Education Expenditure %, Unemployment, GDP, Injuries, Access to Clean Water, Communicable, NonCommunicable). Using our data we will identify key variables contributing to life expectancy in Europe, specifically the European Union countries. The key variables will help us propose strategic recommendations for the EU to allocate funding and policy efforts to increase life expectancy. Our consultancy will be mainly using Tableau Prep to clean our data, and produce visualization analysis. In order to provide a better understanding of the data, our consultancy will also be using Python to generate histograms to understand the data better. We will be implementing different queries to produce specific outputs using MySQL, and providing different graphs, and visualizations to show our analysis using R.

Part 2: Focus and Growth

2.1 Audience

Our potential clients include international organizations such as the World Bank or WHO, non-governmental organizations, charities, or even just general businesses looking for the next place to expand their business. This set of potential clients will all be able to gain useful insights out of our data to help determine where to distribute their resources or operate as a business. We are also hoping to work with Moderna and Pfizer to see if we can provide specific interventions to regions or specific countries such as medicine and vaccines. With our analyses we hope to assist these organizations in optimizing resource allocation, shaping policy development, and addressing disparities among countries.

What sets us apart is our commitment to utilizing reputable data sources, such as the World Bank, ensuring the legitimacy and reliability of our analyses. Our ability to generate actionable recommendations and provide clear, data-driven insights empowers our clients to make informed decisions, optimize resource allocation, and contribute to improved healthcare and population well-being on a global scale.

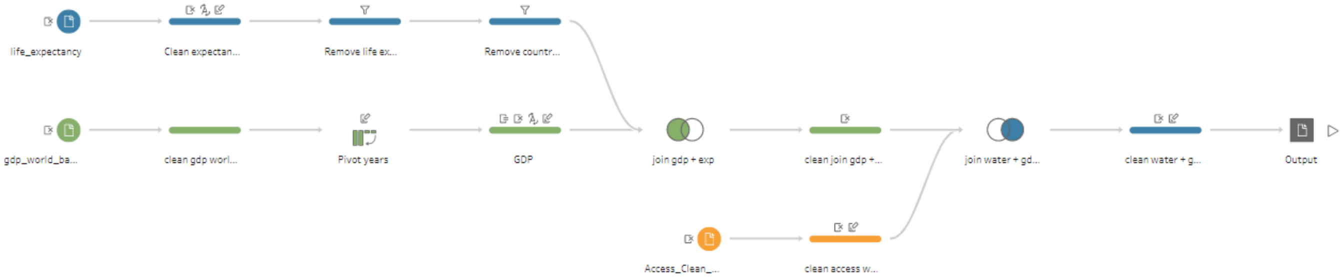
2.2 Potential Areas of Growth

Our consultancy project, specializing in public health and healthcare data analysis, has significant growth potential in serving a diverse range of clients and organizations. These potential clients include healthcare providers, insurers, pharmaceutical companies, government health agencies, NGOs, health tech startups, academic institutions, corporate wellness programs, global health organizations, epidemiological research organizations, and environmental agencies. These entities would be interested in collaborating with our consultancy to leverage our data insights for various purposes, including improving patient care, enhancing insurance offerings, informing policy decisions, advancing research, and addressing public health challenges. By tailoring our services to meet the unique needs of each client group, we can position our consultancy as a valuable partner in improving healthcare outcomes and public well-being.

In addition to the previously mentioned growth avenues, our consultancy can further expand its services by leveraging data on unemployment rates, access to clean water, GDP, and education expenditure. This data enables us to offer labor market analysis, economic impact assessments, and support for water and sanitation initiatives. Moreover, we can specialize in evaluating education policies, align with Sustainable Development Goals, aid regional development strategies, and assess social impacts for investors and philanthropic organizations. Our expertise can also assist global businesses in expansion efforts, partner with community development programs, and facilitate sustainability reporting for companies, demonstrating the breadth of our consultancy's capabilities and its potential to make a meaningful impact on diverse sectors.

Going forward with this project we will continue to work on refining and expanding our dataset to better achieve these goals. We look to incorporate data on areas such as crime, birth rates, pollution, and potentially mental health among others. Adding additional data could help us to refine general recommendations for countries, or specific recommendations for clients our consultancy is working with at a given time. This could also help to address issues we have encountered with a few columns of our data that needed to be removed due to having too many nulls. We will continue to analyze our data as we begin working more with My SQL and make additional changes to our dataset as needed

Part 3: Database Quality



The data preparation process for our life expectancy data begins in Tableau Prep. Here after running our data through a data interpreter a set of clean steps began to prepare our data to be used throughout the project. Our first step was to do an initial clean step which involved changing the name for each of our fields to a cleaner, simpler set of names that will make working with our data down the line easier. Additionally, we set our year field to numeric from string. The next clean step was dedicated to filtering our data to exclude territories of other countries being counted as independent countries in our data. This helped to eliminate a number of variables with nulls as these territories lacked the values the countries had in fields other than life expectancy. Further nulls were eliminated with the removal of the corruption and sanitation fields which both had more than 50% null values. Following these clean steps, we then created a join step with another set of data that included the missing countries from our original data set. This set of data is not as in depth as our primary data but will help us fill in the gaps in our geographical visualizations.

To complement our life expectancy from the World Bank we decided to incorporate GDP data from the same organization to help us better understand the effects of income on life expectancy. We added this to our Prep as a separate input with the first clean step used just to view our data. Then we created a step to pivot years from columns to rows to keep our GDP data in line with our life expectancy data. Our next clean step involves basic clean up such as renaming our fields to make them easier to work with, and removing fields we will not need once the two sets of data are joined. The next step was to join the data in a right join type, this ensures all of our data from life expectancy is preserved and only the data that will be useful to us from GDP is carried over into the new set of data. Finally we undertook a similar join process for a final set of data we added to our overall flow focusing on access to clean water. This data came already formatted correctly from the world bank and simply required a join step.

The final portion of our data preparation process was to deal with the null values in our dataset in a way that ensures accurate analysis and reporting. For the remaining columns with null values, such as "Prevalence of Undernourishment," "CO2," "Health Expenditure %," "Education Expenditure %," "Unemployment," and "Access to Water," we needed a more nuanced approach than just deleting columns or removing all rows with nulls. We grouped the data by 'Country Name' in Python using the Pandas library in Jupyter Notebook and calculated the mean of each variable for each country. For example, for "CO2," we calculated the mean CO2 value for each country. We then filled in the missing values for each country with the calculated mean. This approach ensured that we didn't lose data and accounted for variations between countries. The same process was used for subsequent data sets joined in our Tableau Prep flow to ensure consistency throughout. Some countries however had no data for a given variable for all years in our data set. In such cases, calculating the average would result in zero. These values were left empty as they accurately represented the missing data. For example, if Argentina had no data for "Access to Clean Water," the mean for Argentina in that category remained empty.

Part 4: Correlation and Regression

We conducted multivariate regression analysis to test the significance of our variables and determine how things are affecting life expectancy. In doing so we came up with the following multivariate regression equation:

Life Expectancy = β0 + β1 * **GDP in Billions** + β2 * **Health Expenditure** + β3 * **Education Expenditure** + β4 * **Prevalence of Undernourishment** + β5 * **Communicable Diseases** + β6 * Non-Communicable Diseases + β7 * **Access to Clean Water**

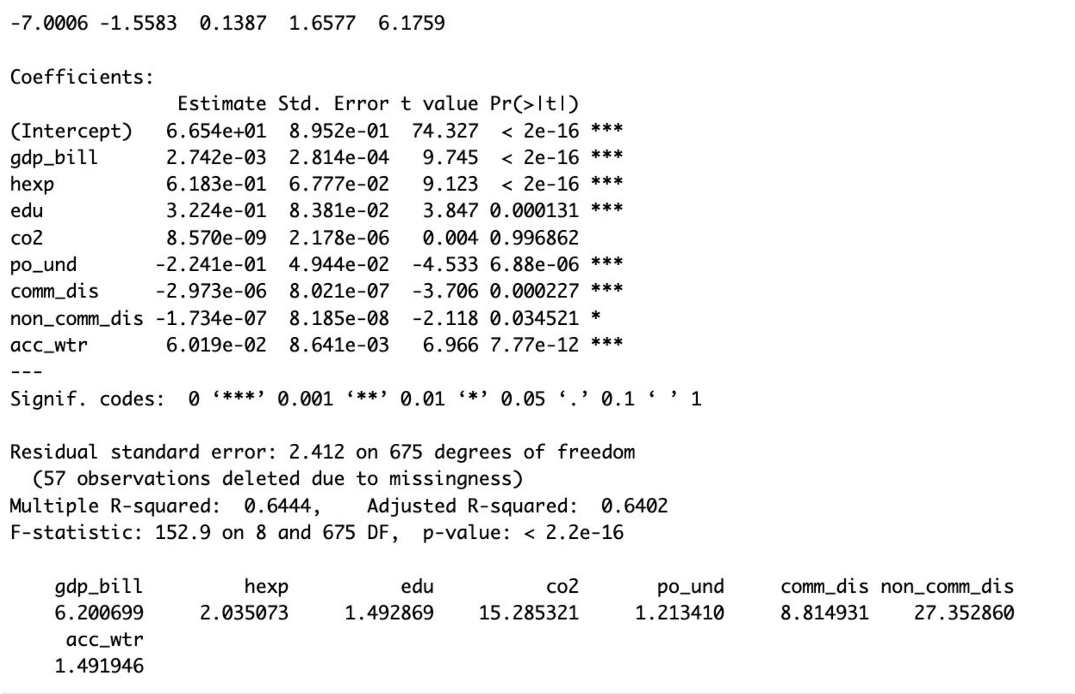
Here, Life Expectancy is the dependent variable and the independent variables in the equation are GDP in Billions, Health Expenditure, Education Expenditure, Prevalence of Undernourishment, Communicable Diseases, Non-Communicable Diseases and Access to Clean Water. We did a multivariate regression instead of individual regressions as in the real world life expectancy is affected by multiple variables so doing a multivariate analysis makes the most sense. B0 in the equation is the y intercept.

We got the following results when we ran our regression:

- Adjusted R-Squared of 64%
- A very high F-statistic (152.9) and a very low p-value (< 2.2e-16) indicate that the model as a whole is statistically significant.

The adjusted R-Square of 64% tells us that 64% of variation in life expectancy is explained by the variables that we have, which is significant. There are other factors that can affect life expectancy too beyond what we have in our current dataset. These could be things like crime rate, employment rate and birth rate etc. These variables could be added to our dataset in the future to get a better understanding of factors that affect life expectancy.

The R output from our regression analysis is given below:

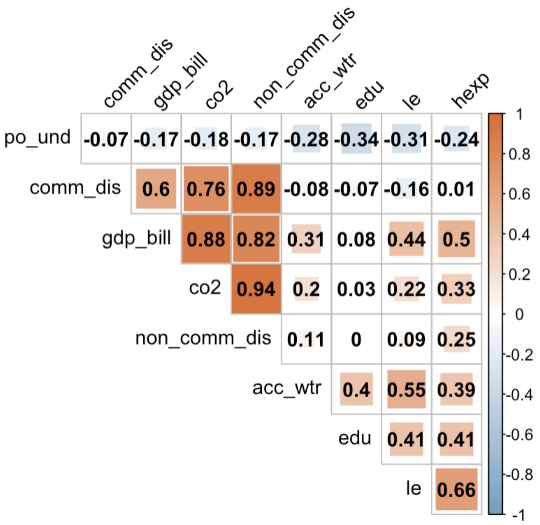


```
-7.0006 -1.5583  0.1387  1.6577  6.1759

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6.654e+01  8.952e-01  74.327  < 2e-16 ***
gdp_bill      2.742e-03  2.814e-04   9.745  < 2e-16 ***
hexp          6.183e-01  6.777e-02   9.123  < 2e-16 ***
edu           3.224e-01  8.381e-02   3.847 0.000131 ***
co2           8.570e-09  2.178e-06   0.004 0.996862    
po_und       -2.241e-01  4.944e-02  -4.533 6.88e-06 ***
comm_dis     -2.973e-06  8.021e-07  -3.706 0.000227 ***
non_comm_dis -1.734e-07  8.185e-08  -2.118 0.034521 *  
acc_wtr       6.019e-02  8.641e-03   6.966 7.77e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.412 on 675 degrees of freedom
  (57 observations deleted due to missingness)
Multiple R-squared:  0.6444,	Adjusted R-squared:  0.6402 
F-statistic: 152.9 on 8 and 675 DF,  p-value: < 2.2e-16

    gdp_bill         hexp          edu          co2       po_und     comm_dis non_comm_dis 
    6.200699     2.035073     1.492869    15.285321     1.213410     8.814931    27.352860 
     acc_wtr 
    1.491946 
```

We can see that our model is statistically significant give the low p value and high F statistic and our individual variables are also all mostly significant at the 3 stars level. Co2 is shown to not be significant. We also checked for multicollinearity in our regression equation using the VIF function in R and noticed that Co2 has high multicollinearity.

The highlighted variables in the regression equation are the most statistically significant variables affecting Life Expectancy.

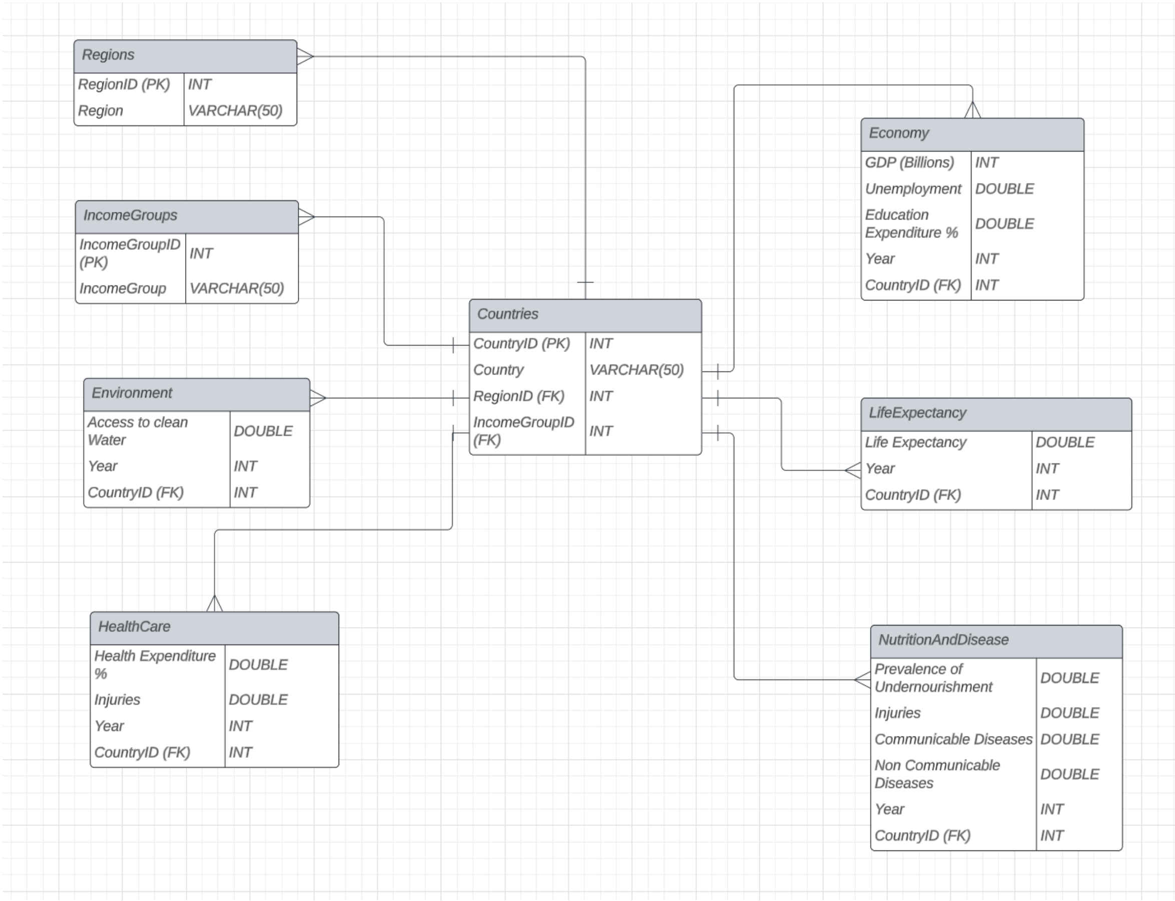
We also did a correlation analysis to see the correlation between the different variables. We can see that life expectancy and prevalence of undernourishment have a -0.31 correlation, which is significant. We can also see that there is a strong positive correlation between life expectancy and GDP in Billions, life expectancy and Access to Clean Water and Life Expectancy and Education Expenditure.



Overall, our regression analysis and correlation analysis told us what our most significant variables were so we could later focus our visualizations and recommendations based on this and continue to develop our analysis. The regression analysis also ensured the significance of our model.

Part 5: Database Design

Below is our EER diagram that we created using LucidChart.



Regions	
RegionID (PK)	INT
Region	VARCHAR(50)

IncomeGroups	
IncomeGroupID (PK)	INT
IncomeGroup	VARCHAR(50)

Environment	
Access to clean Water	DOUBLE
Year	INT
CountryID (FK)	INT

HealthCare	
Health Expenditure %	DOUBLE
Injuries	DOUBLE
Year	INT
CountryID (FK)	INT

Countries	
CountryID (PK)	INT
Country	VARCHAR(50)
RegionID (FK)	INT
IncomeGroupID (FK)	INT

Economy	
GDP (Billions)	INT
Unemployment	DOUBLE
Education Expenditure %	DOUBLE
Year	INT
CountryID (FK)	INT

LifeExpectancy	
Life Expectancy	DOUBLE
Year	INT
CountryID (FK)	INT

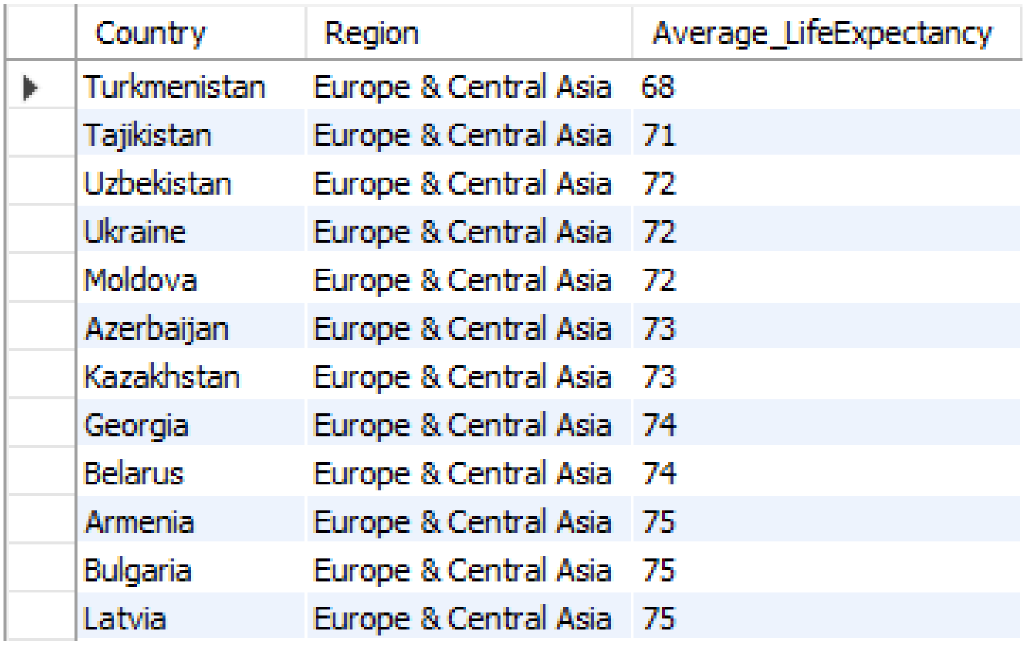
NutritionAndDisease	
Prevalence of Undernourishment	DOUBLE
Injuries	DOUBLE
Communicable Diseases	DOUBLE
Non Communicable Diseases	DOUBLE
Year	INT
CountryID (FK)	INT

Above we have the CountryID as the main Primary code, and the region ID and Income Group ID as the foreign keys. CountryID and Year act together as composite keys in other tables. The Countries table has a one to many relationship between all the other tables. This relationship will make joining tables and creating different queries much easier on SQL.

The Year attribute appears in multiple tables such as Environment, Economy, Healthcare, NutritionAndDisease and LifeExpectancy. This design allows us to capture annual variations in these different attributes. Our years range from 2001 to 2019. We are able to see how various indicators change over time for each country. By including Year alongside CountryID as part of a composite key, the database can store multiple records for the same country, differentiated by the year the data was recorded. This feature will prove to be crucial for us if we choose to perform any time-series analysis in the future and as we look for trends over years.

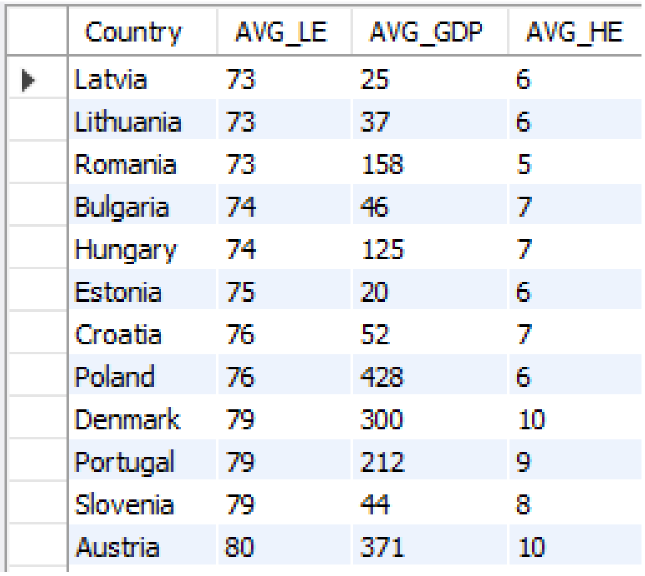
The Countries table is at the center of our database, connecting to every other table. Each country is identified uniquely by CountryID, which is used as a Foreign Key in a lot of tables. With this design choice we are able to gather data points related to a specific country, which is crucial for our on-going analysis for our stakeholders. Our goal is to provide region specific reports and interventions and this design will prove vital for that. Countries table will be pivotal in joining all the other tables and constructing complex SQL queries.

Moving forward from our EER diagram we tried out many different queries and clauses on SQL. We used multiple different clauses to show different output focusing on specific countries, regions, and year. Since our consultancy focuses on the European region we decided to take a closer look at the average life expectancy in the Europe & Central Asia countries. We used JOIN to join the Countries, Regions, and Life Expectancy tables and the WINDOW Function to find the average life expectancy for each country from the years 2001-2019. Below is our SQL output.



Country	Region	Average_LifeExpectancy
Turkmenistan	Europe & Central Asia	68
Tajikistan	Europe & Central Asia	71
Uzbekistan	Europe & Central Asia	72
Ukraine	Europe & Central Asia	72
Moldova	Europe & Central Asia	72
Azerbaijan	Europe & Central Asia	73
Kazakhstan	Europe & Central Asia	73
Georgia	Europe & Central Asia	74
Belarus	Europe & Central Asia	74
Armenia	Europe & Central Asia	75
Bulgaria	Europe & Central Asia	75
Latvia	Europe & Central Asia	75

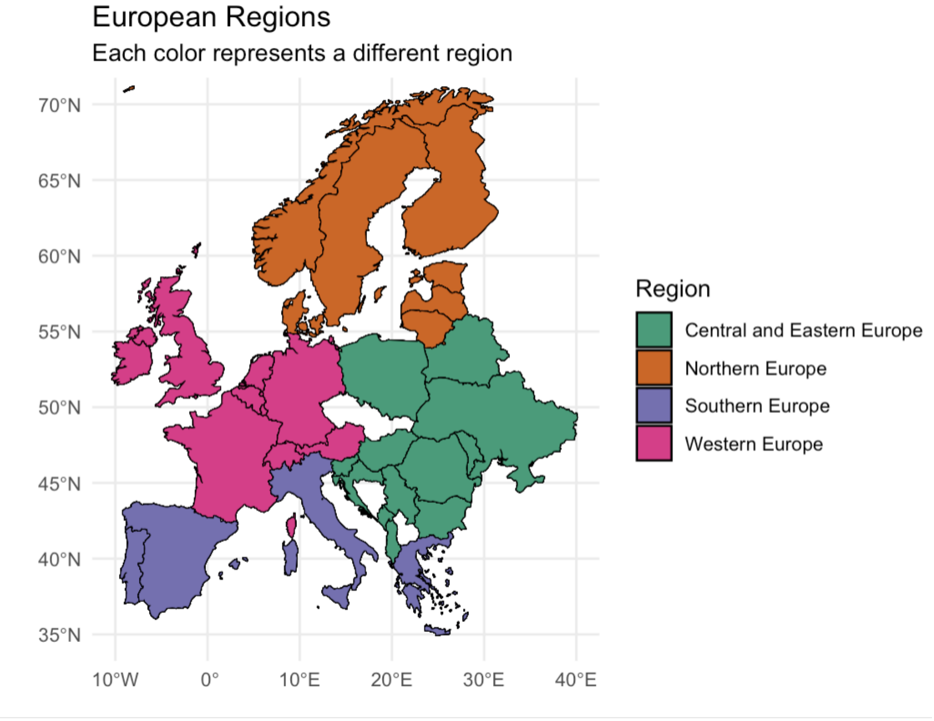
Since our data does not have specifically the Europe region as a separate region from Central Asia, we decided to create a VIEW with just all the 27 countries in the European Union. We used this View and the WINDOW function to find the average life expectancy, average GDP, and average health expenditure. We wanted to see if there is any correlation between these variables, and to see if lower life expectancy means low GDP and lower health expenditure.



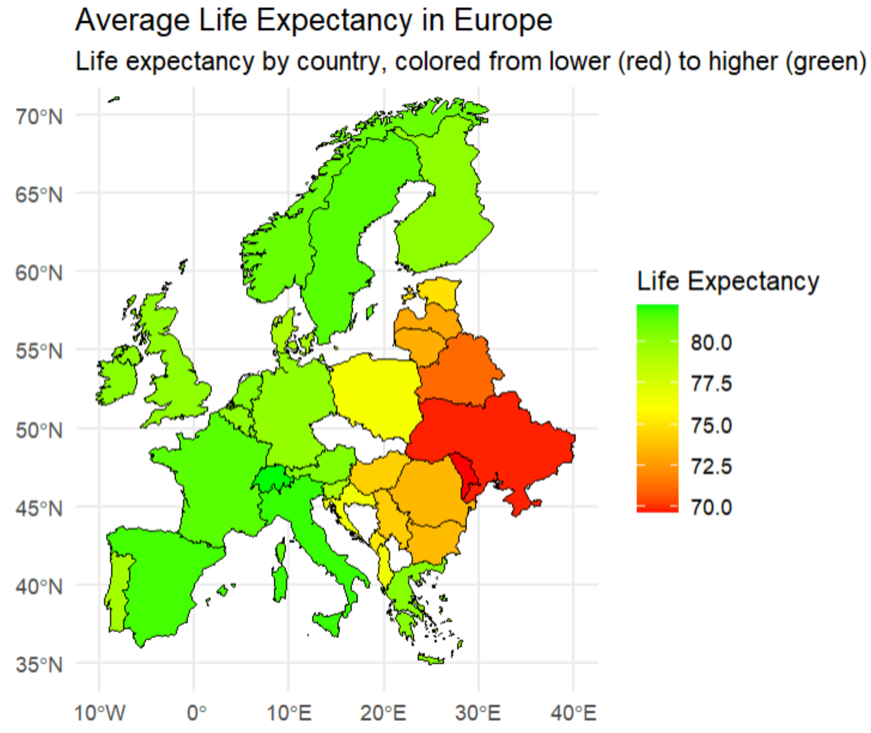
Country	AVG_LE	AVG_GDP	AVG_HE
Latvia	73	25	6
Lithuania	73	37	6
Romania	73	158	5
Bulgaria	74	46	7
Hungary	74	125	7
Estonia	75	20	6
Croatia	76	52	7
Poland	76	428	6
Denmark	79	300	10
Portugal	79	212	9
Slovenia	79	44	8
Austria	80	371	10

Part 6: Visualizations

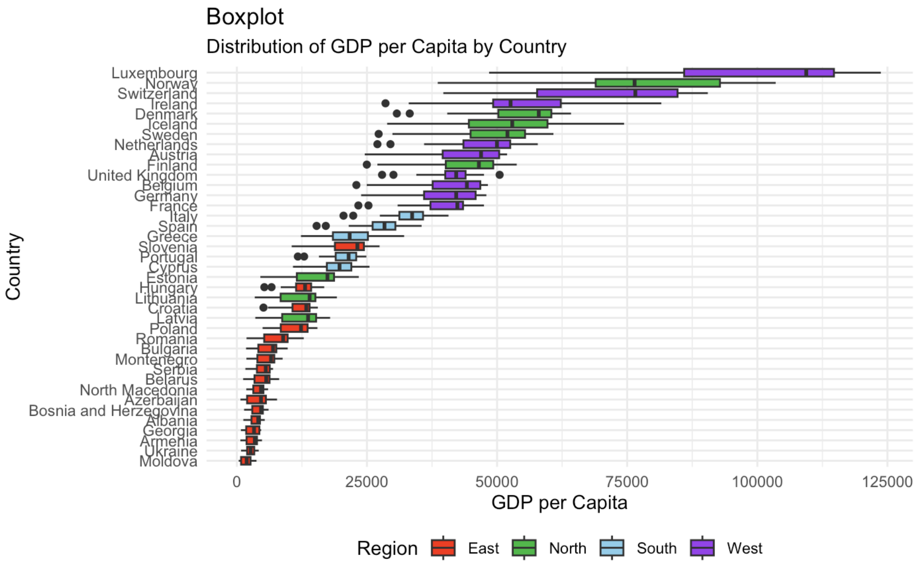
For the visualization portion of our report we created a number of ggplots using R. Our visuals look specifically at the European region as this is what our final presentation focused on. Most of these visuals are using maps of the European continent with the other few focusing on the individual components that affect life expectancy.

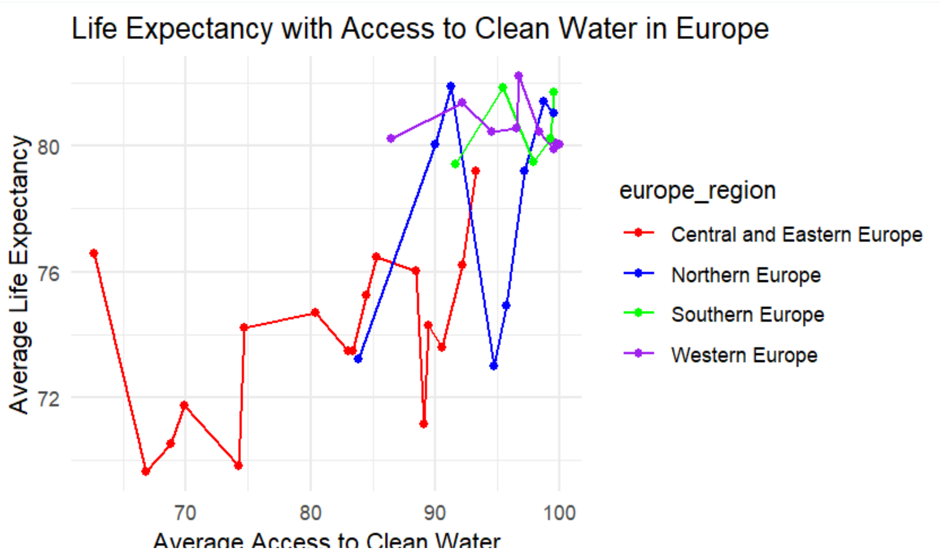


Our first visual here is just a representation of the different regions of Europe. The purpose of this visual is to introduce our audience to the different regions of the continent in a clear and easy manner. This will be important as we highlight the areas of Europe with a lower level of life expectancy.

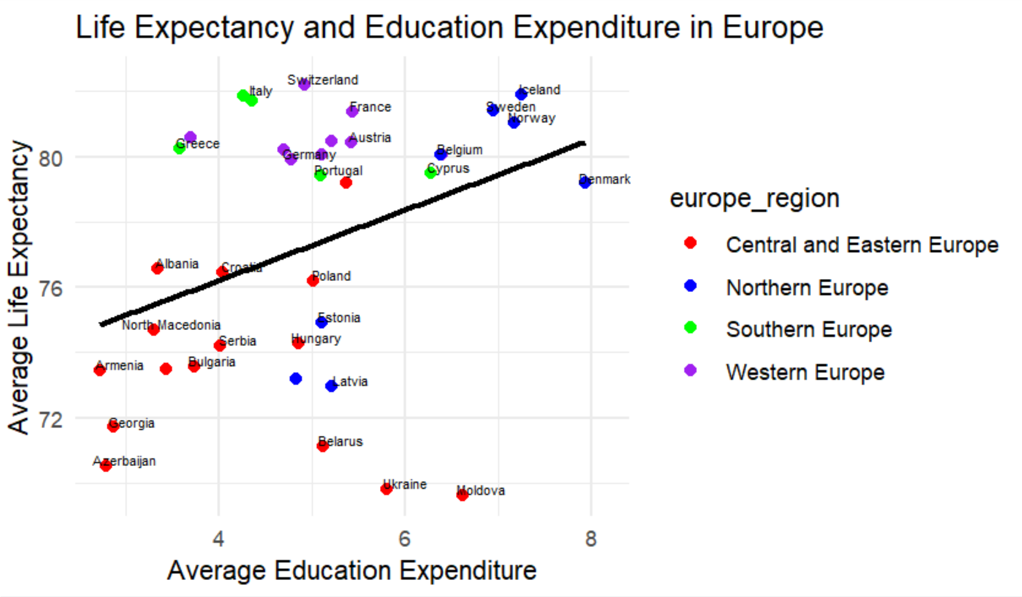

Average Life Expectancy in Europe
Life expectancy by country, colored from lower (red) to higher (green)
70°N
65°N
60°N
55°N
50°N
45°N
40°N
35°N
10°W
0°
10°E
20°E
30°E
40°E
Life Expectancy
80.0
77.5
75.0
72.5
70.0

For this map we look at the life expectancy across all of the countries. When placed back to back with the regions it is easy for the audience to see that Eastern Europe is far below the average in life expectancy when compared to the other regions of the continent. This is where we are looking to make policy recommendations with our analysis.



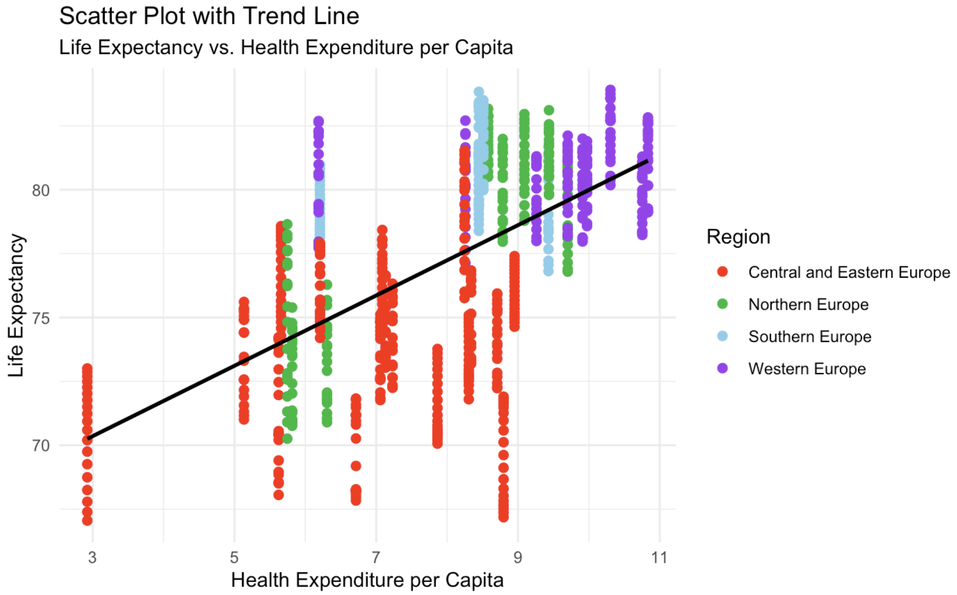


The line graph illustrates the correlation between average life expectancy and access to clean water across four distinct regions in Europe. Notably, the red line depicting Central and Eastern Europe stands out, indicating a lower level of access to clean water, which directly correlates with a reduced average life expectancy in this region compared to the other three. This stark visual contrast highlights the significant impact that clean water accessibility can have on life expectancy, emphasizing the importance of addressing disparities in resource availability for improved public health outcomes across different European regions.

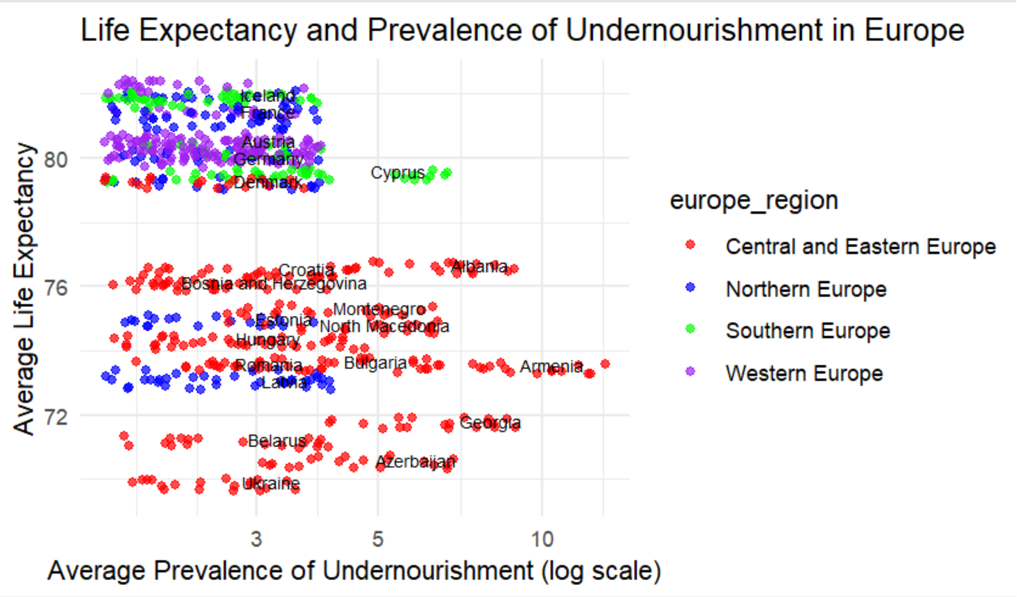


The scatter plot effectively captures the correlation between average life expectancy and average education expenditure across four distinct regions. Notably, the red points corresponding to Eastern and Central Europe are prominent and predominantly situated below the trend line.

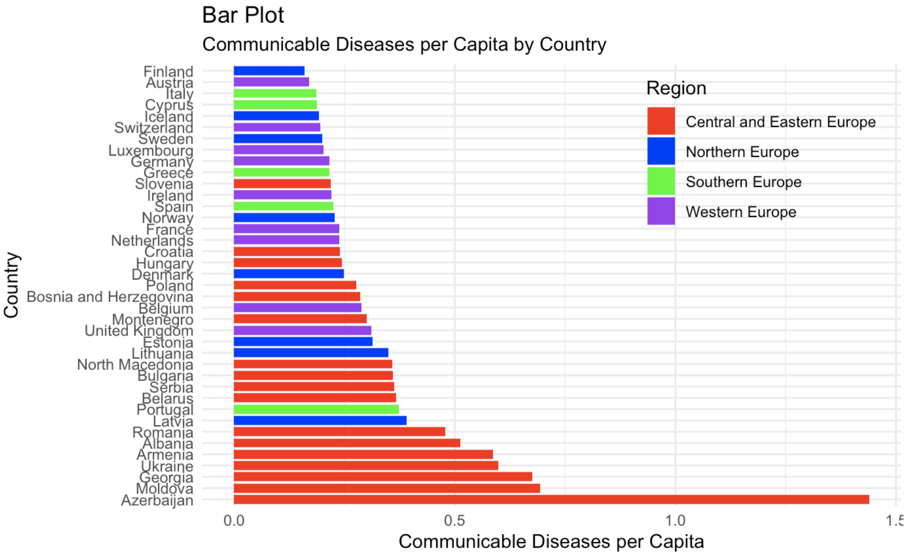
This distinct positioning suggests that, on average, countries in this region tend to exhibit lower life expectancies in relation to their education expenditures when compared to the overall trend observed across the four regions.



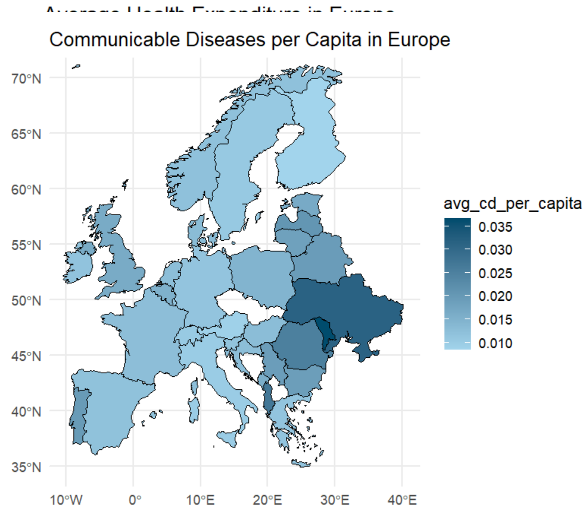
The scatter plot featuring a trendline for life expectancy against health expenditure per capita provides representation of the relationship between these variables. This visual pattern indicates that in Central and Eastern Europe, there is a tendency for countries to have lower life expectancies relative to their health expenditure per capita. The clustering of these red points below the trendline suggests that, on average, nations in this region allocate less health expenditure, and this is associated with a corresponding decrease in life expectancy.

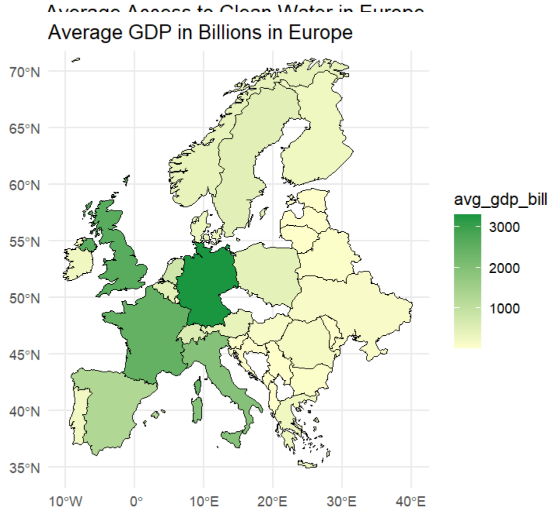


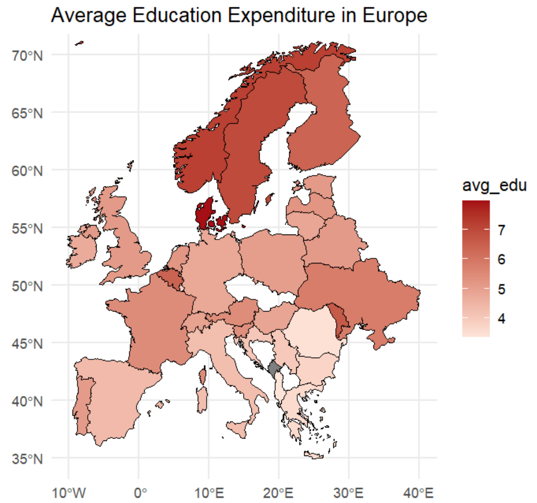
This scatter plot depicts life expectancy against the prevalence of undernourishment. The distinct red points, scattered and predominantly situated in the lower portion of the plot, signify Central and Eastern Europe. This distribution pattern suggests a higher prevalence of undernourishment in this region, correlating with lower life expectancy.

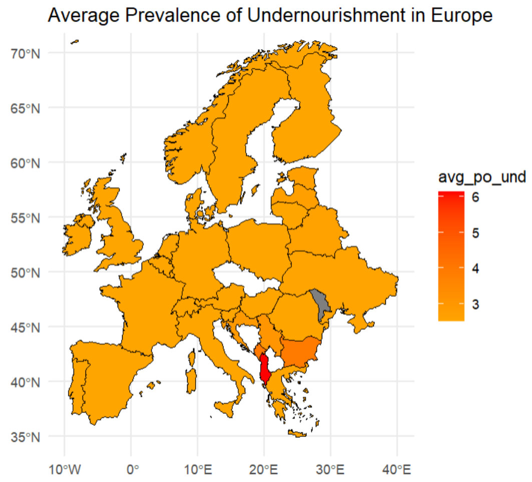


This group of visuals focus on the key factors of life expectancy that we discovered have the largest impact through our regression analysis. We look at these different factors such as: GDP per capita, Access to clean water, Education/Health expenditure, prevalence of undernourishment, and communicable diseases per capita. We highlight the different countries and regions and how this factors into Europe as a whole. This further reinforces our initial hypothesis that Eastern Europe is in need of support in order to catch up to the other three regions of the continent.

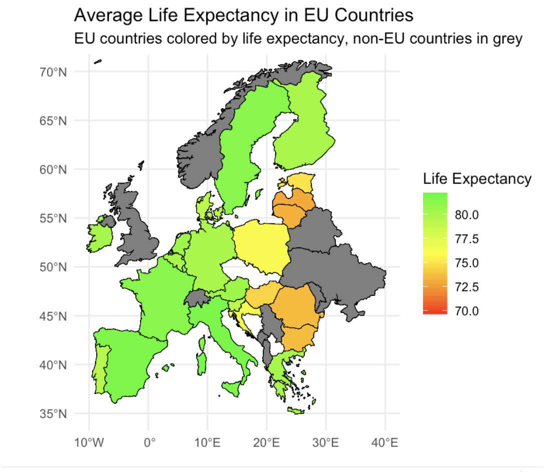


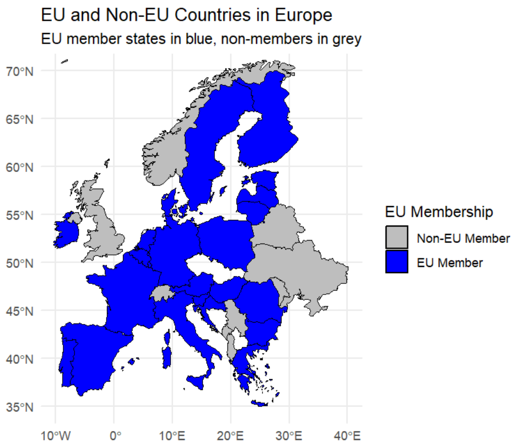






The maps show the effect of each of these variables in each country across Europe. We can see with most of these maps that the Eastern European region is in need of assistance in these areas. We used these maps to highlight the areas in most need urgently and used these in our recommendations below.


Average Life Expectancy in EU Countries
EU countries colored by life expectancy, non-EU countries in grey
70°N
65°N
60°N
55°N
50°N
45°N
40°N
35°N
10°W
0°
10°E
20°E
30°E
40°E
Life Expectancy
80.0
77.5
75.0
72.5
70.0



Our final two visuals focus on the European Union, our sample client for this project. We have seen how these variables affect Europe as a whole but we can see even in the European Union these issues still exist. The organization works to share money and resources to regions in need but will still need to actively work to pull the nations in Eastern Europe in line with those in the other regions of the continent.

Part 7: Recommendations and Learnings

7.1 Economic Policy (GDP)

We recommend different economic policies to boost GDP in the European Union. One suggestion is increasing cooperation and support among member states can help address regional disparities within the European Union. The single market reduces trade barriers and increases overall external trade. Increasing the shared monetary and fiscal policies could help spread the wealth in overall Europe. An additional suggestion can be incorporating more member states into the Eurozone which can help to drive up the GDP value of Eastern European countries, since they have a weaker currency.

7.2 Social Policy (Health and Education)

At our consultancy, health and education is a priority, therefore we have come up with some suggestions to help increase healthcare and education expenditure. One option can be providing grants for struggling countries with low health and education expenditure. The EU can also set minimum education and healthcare expenditure percentages for member countries. In addition, increasing public awareness of diseases and available healthcare and education options to those in underprivileged regions.

7.3 Social Policy (Access to Clean Water)

Access to Clean Water is an important value to increasing life expectancy, especially with climate change and population growth. One way to increase access can be allocating more resources towards the development and maintenance of clean water infrastructure such as modernizing water treatment facilities, improving sewage systems, and regular checks on water sources. Some Western European countries have successfully implemented decentralized water management systems, where local communities play a significant role in managing their water resources. This approach can be more responsive to local needs and challenges.

7.4 Learnings from the journey

Throughout the project journey, we gained valuable insights and experiences. Securing a dataset aligned with our interests proved pivotal, offering versatility for manipulation and visualization. Embracing a hybrid working schedule, which seamlessly integrated both virtual meetings over Zoom and in-person collaborations, enabled us to effectively navigate through varying timelines. Our commitment to consistent communication and feedback, both with our professor and within the group, ensured that we remained on track. A significant learning aspect involved honing our skills in visualizing information through the use of R and proficiently executing SQL queries. These collective learnings have not only improved our technical capabilities but also enhanced our collaborative and communicative prowess.