Brandeis University

International Business School

Group 3: Final Report

Bus 240 F-1 - Information Visualization

Salman Haider, Sana Ijaz, Maria Navarro Castillo, Erick Amezcua, Kyle Allsopp

October 19th, 2023

Table of Contents

Part 1: About Viz-Health Insights

1.1 Context and Market of Consultancy

Viz-Health Insights specializes in public health and healthcare data analysis with the goal of offering recommendations for enhancing healthcare outcomes and overall population well-being. While our potential clients include international organizations such as the World Bank, WHO, and the United Nations, our services are not limited to these entities. Non-governmental organizations, charities, or even businesses in general will be able to gain useful insights out of our data to help determine where to distribute their resources or operate as a business. With our analyses we hope to assist these organizations in optimizing resource allocation, shaping policy development, and addressing disparities among countries.

1.2 Data Source

For our data we turned to the World Bank, a widely recognized and respected source in the global health and finance areas. The World Bank has an impartial and international reputation that we can trust. We can be rest assured that this data is legitimate and captures most countries of the world accurately and thoroughly. We have extracted several databases from the World Bank such as Life Expectancy by Countries, Nominal GDP by Countries, and Access to Clean Water. Using this data from the World Bank will provide us with an accurate, and well trusted output, as well as a number of areas by which we can draw conclusions about the effects of life expectancy across the globe.

1.3 Focus of the Consultancy

Our analytical focus encompasses variables such as life expectancy, healthcare expenditure, education spending, and more to explore their interrelationships. We will delve into correlations and potential causal links to understand the significant disparities in life expectancy among countries.

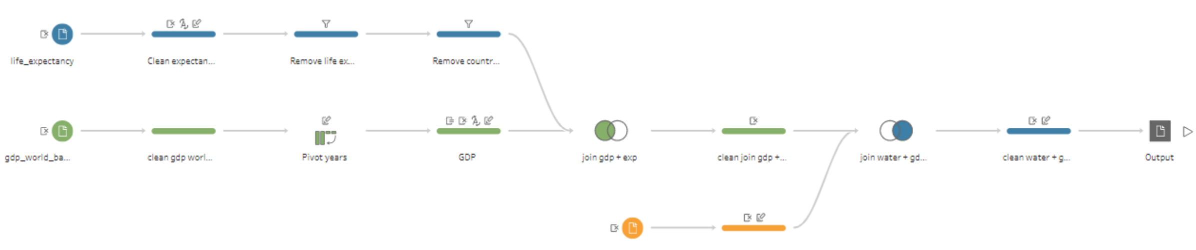
The insights derived from our analysis hold substantial value for consulting firms that specialize in advising international development organizations, healthcare institutions, and governmental bodies. As consultants in our firm, we envision collaborating with entities committed to improving the well-being of people worldwide. These organizations seek to comprehend the intricate relationship between various socioeconomic factors and life expectancy across countries, ultimately aiming to allocate their resources effectively for maximum impact.

1.4 Areas of Potential Growth

As we look ahead to the future, VizHealth is eyeing many key areas of growth that will excite our clients and stakeholders and be very beneficial to them. We want to make sure we retain our clients which would be international development organizations, healthcare institutions and governmental bodies. One area we are looking to grow in is diving into regional customization. We are looking to specialize in regional reports for areas such as Sub-Saharan Africa, South Asia, Latin America and others. We want to start with these regions as they have the most potential to improve their life expectancy and there is a lot of work to be done. We can provide our clients with specialized reports that will allow them to tailor their interventions based on specific local needs. We are also looking to offer predictive analytics–a game-changer for anyone who wants to be ahead of the curve in public health trends. We hope to one day utilize regression and other machine learning models to predict future healthcare trends based on variables like healthcare spending, access to clean water, etc. We are also planning on adding additional data sources incrementally such as crime data, pollution indices, and social media trends. Additional variables we have considered adding in the future are nutritional information such as average caloric intake. We have also considered adding Crime Rates and Income Inequality Metrics which both most likely have an impact on Life Expectancy. Real-time dashboards are on the horizon too, which could be perfect for quick, actionable data and insights. As we add these offerings to our consultancy, this will keep our stakeholders interested in using our insights and tools and continue to provide them more ways to target interventions and improve outcomes in their respective regions.

Part 2: Data Quality Assurance

2.1 Data Preparation



In our data preparation process, we used a combination of Tableau Prep and Jupyter Notebook to better understand our dataset. We initiated this process by establishing a connection to our Life Expectancy dataset and cleaning it with a Data Interpreter to ensure Tableau Prep read the data correctly regardless of the table format.

We noticed that a significant number of fields had null values, which was also corroborated by our analysis in the Jupyter Notebook. As you can see in the screenshot below, Education Expenditure %, Corruption, and Sanitation had a large number of nulls.

```
In [8]: # Count missing values in each column

missing_values = data.isnull().sum()

# Display the count of missing values for each column
print("Missing Values in Each Column:")
print(missing_values)
```

```
Missing Values in Each Column:
Country Name                        0
Country Code                        0
Region                              0
IncomeGroup                         0
Year                                0
Life Expectancy World Bank        188
Prevelance of Undernourishment    684
CO2                               152
Health Expenditure %              180
Education Expenditure %          1090
Unemployment                      304
Corruption                       2331
Sanitation                       1247
Injuries                            0
Communicable                        0
NonCommunicable                     0
dtype: int64
```

This step consisted of ten changes to our dataset. Due to their considerable number of nulls, we removed the Corruption and Sanitation fields. To avoid redundancy, fix typos, and ensure clarity, we renamed seven fields. Additionally, we checked all of the data types and changed the data role of the Country field to Country/Region.

2.2 Removing Life Expectancy Nulls

Upon exploring the null values in the Life Expectancy field, we realized that these came from ten countries that had null values every year or almost every year. Since Life Expectancy is the variable we want to understand more clearly, we excluded these countries, which removed all null values from this field.

2.3 Removing countries with 4 or more fields with null values every year

Using the Jupyter Notebook, we looped through each country and checked for columns with null values for every year it had too much missing information that we cannot impute. These are the countries we excluded along with the respective columns that guided our analysis:

Country Name	**Columns with all null values**
Bermuda	Prevalence of Undernourishment, CO2, Health Expenditure %, Unemployment

Greenland	Prevalence of Undernourishment, CO_2, Health Expenditure %, Unemployment
Guam	Prevalence of Undernourishment, CO_2, Health Expenditure %, Education Expenditure %

Lastly, we checked the Country column and excluded any territories, such as Puerto Rico, that were incorrectly included in this dataset as countries.

2.4 Incorporating GDP and Access to Clean Water Datasets

To further enrich our analysis, we decided to add two other datasets into our flow containing information on the GDPs and Access to Clean Water for each country from 2001 - 2019. For the GDP dataset, after establishing a connection and conducting a preliminary cleaning with the Data Interpreter, we realized that the dataset had a wide format. Each year had its own column, so we pivoted the year columns, which resulted in the creation of the year and GDP columns.

2.5 Cleaning GDP

After pivoting our data, we conducted more in-depth cleaning. We renamed the pivot values column as GDP to ensure it is clear what they represent. We utilized the calculated field feature to divide the GDP values by a billion and rounded the result to three decimal places, and renamed the calculation to GDP (Billions). We removed the old GDP column and any extra columns generated from the pivot step to ensure our dataset only included the necessary data. Anticipating the next step in our cleaning process involving a join, we changed the data type of Year from string to number (whole)

2.6 Joining Datasets

After ensuring both of our datasets were clean, we created a left-join on country and year to return all records from the Life Expectancy table and the matched records from the GDP table. Lastly, we removed any extra columns automatically generated from the join and checked the resulting table for any mismatches or incongruencies.

We then joined our main dataset to Access to Clean Water. We created a right-join on country and year to return all records from the joined Life Expectancy + GDP table and the matched records from the Water table. Lastly, we removed any extra columns automatically generated from the join and checked the resulting table for any mismatches or incongruencies.

2.7 Final Cleaning Steps and Handling Nulls

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 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.

2.8 Final Data Quality

As we progressed through Tableau Prep, we generated a refined output file, providing us with a single, clean dataset for our future work. This final dataset encompasses 16 columns and 19 years of data for 160 countries. It includes details about the country's region as a text string and classifies its income level into bins. At the core of our analysis for this project lies data pertaining to life expectancy. Subsequent columns in the dataset will serve as essential variables to explore and comprehend the factors influencing a country's life expectancy. These variables encompass elements such as CO2 emissions, healthcare and education expenditures, unemployment rates, injury statistics, rates of communicable and non-communicable diseases, GDP figures, and the level of access to clean water.

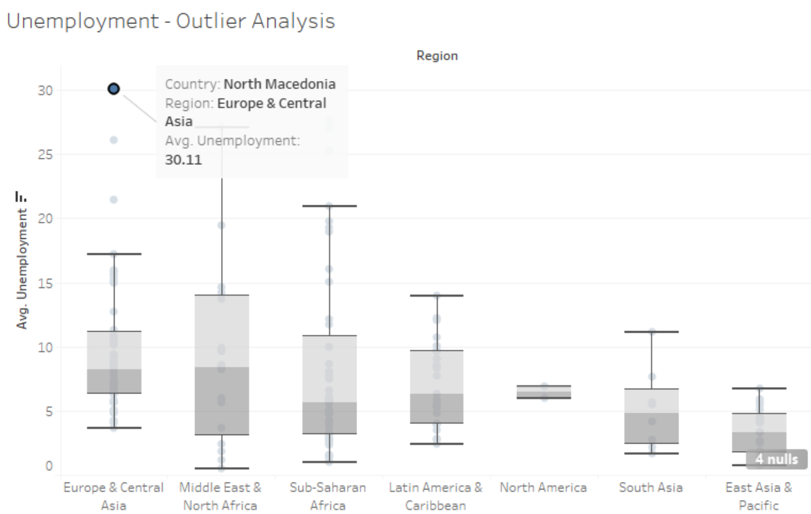
Our data cleaning process and imputation method, implemented in Python using the Pandas library, effectively reduced the presence of null values in our dataset by replacing them with mean values. We made sure to count the nulls before we filled them in and after we filled them in. The remaining nulls might require us to exclude some countries from certain components of our analysis to ensure the integrity of our data visualizations. After completing our data cleaning process, we will be well-equipped to effectively employ this information in discerning the common factors contributing to variations in life expectancy, whether higher or lower.

The plot also highlights that economic development, as measured by GDP per capita, is associated with improved life expectancy, particularly in the middle-income group. This suggests that efforts to improve economic conditions in these countries may lead to better health outcomes and longer life expectancies for their populations.

Part 3: Removing the Outliers

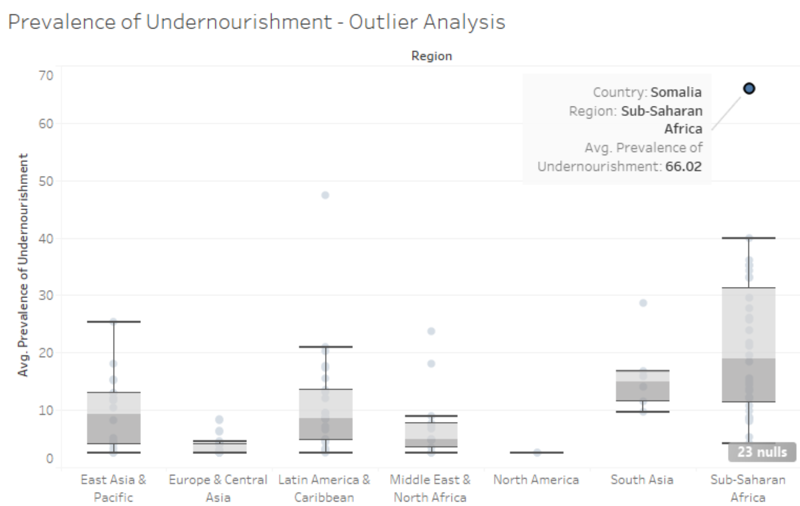
In our extensive examination of life expectancy and its correlated variables, we identified remarkable outliers - one for each key variable. These outliers signify countries that stand apart from the norm within specific aspects related to life expectancy. To gain deeper insights, we conducted thorough analyses for each of these exceptional countries, aiming to understand the unique factors contributing to their outlier status.

3.1 Unemployment



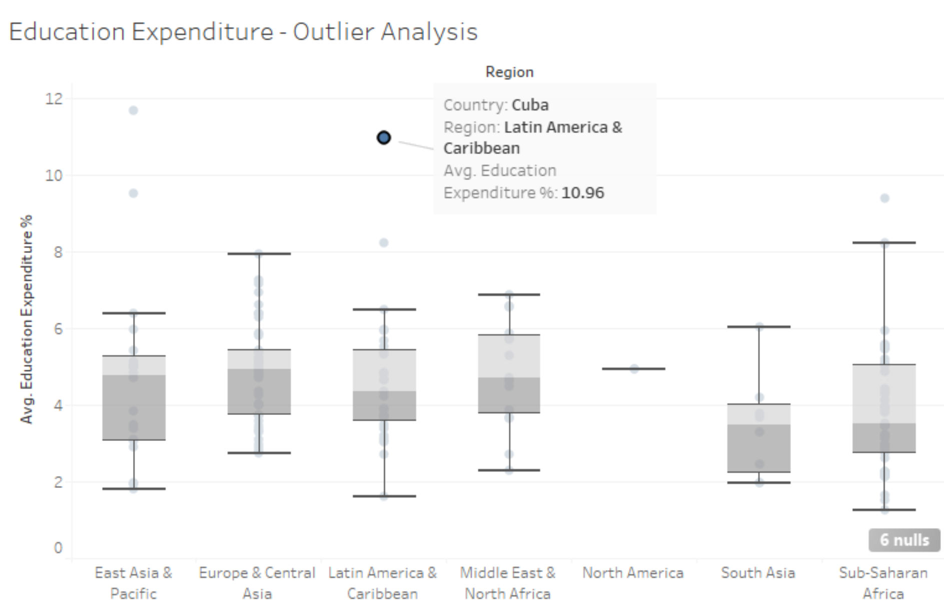
Our outlier analysis spotlights North Macedonia due to its exceptionally high unemployment rate within our dataset. This anomaly is rooted in several factors, including low labor market activity, employment rates, and a persistent skill-employer mismatch. Historical transitions from socialism to a market-oriented economy exacerbated this issue, with the private sector struggling to absorb job losses from state-owned enterprise restructuring. Additionally, the growth of the informal economy, coupled with issues like nepotism and cronyism, has deterred young people, leading to emigration. High youth unemployment persists due to inadequate qualifications, a disconnect between education and market needs, and a demand for prior work experience. These complexities contribute to North Macedonia's outlier status, highlighting the need for a nuanced approach to address the nation's unemployment challenges.

3.2 Prevalence of Undernourishment



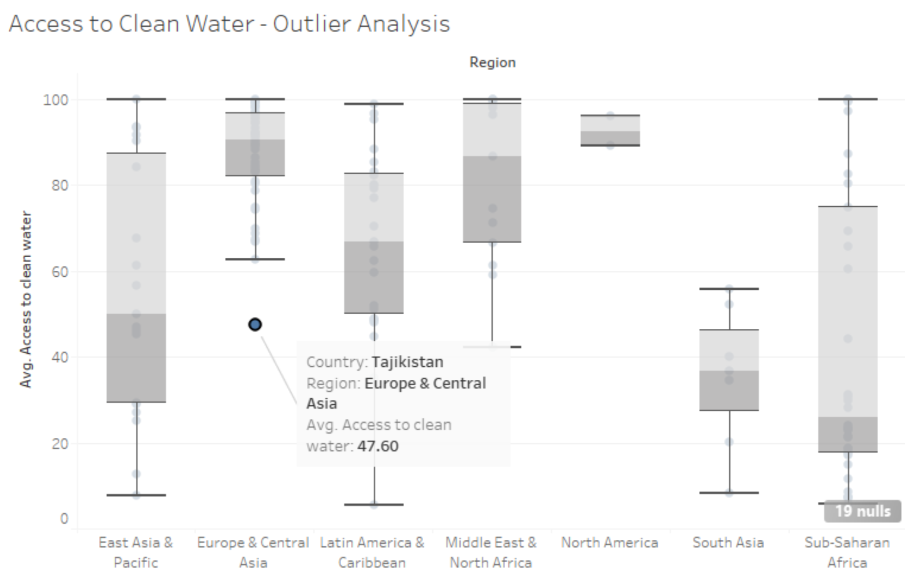
In our outlier analysis, Somalia emerged as a significant outlier for the prevalence of undernourishment in 2018, affecting 1.2 million children and 6.5 million adults. This alarming trend is exacerbated by a 15-year civil war, political instability, and extreme drought. To address this complex issue, a multifaceted approach is required, involving increased humanitarian aid, food imports, and education on fertility and nutrition. These steps are crucial in tackling Somalia's undernourishment crisis and guiding the country towards a more stable future.

3.3 Education Expenditure



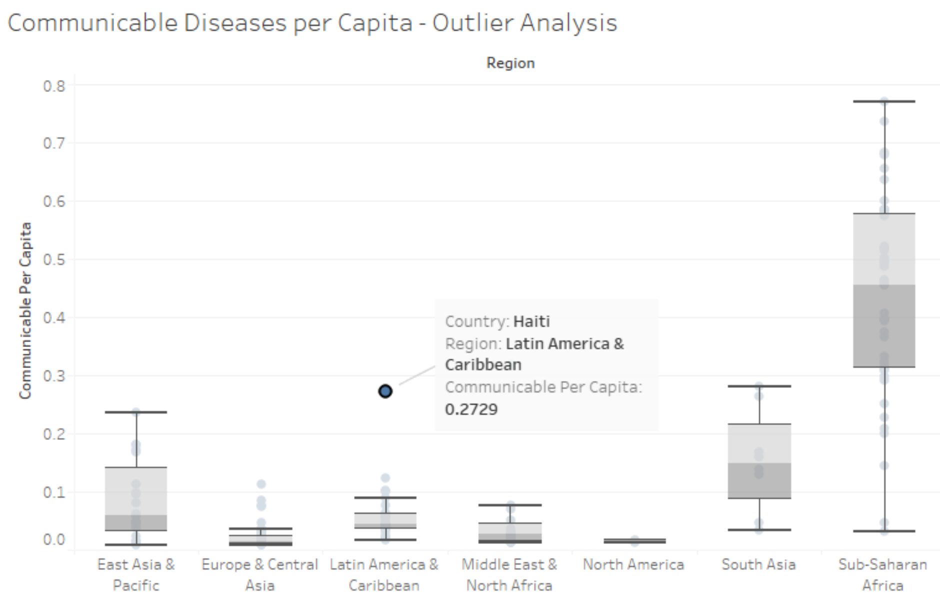
Cuba boasts one of the world's highest education expenditures, a testament to its historical drive to combat illiteracy. Fidel Castro's regime made education universal by dissolving private universities, although opportunities often favored those with political connections. Today, Cuba continues to invest heavily in education, particularly in producing medical graduates and maintaining high literacy rates. However, this sometimes leads to overqualified individuals for limited job opportunities due to economic constraints and restricted freedoms. Addressing this challenge may involve economic investment and greater freedom to retain skilled professionals and reduce brain drain. While education isn't strongly correlated with life expectancy, it remains vital, necessitating a multifaceted perspective.

3.4 Access to Clean Water



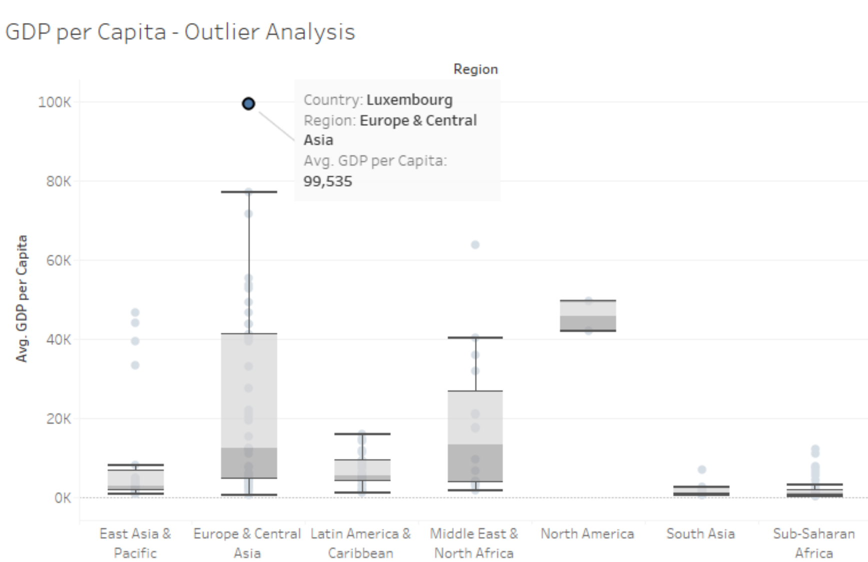
In our clean water access analysis, Tajikistan emerges as a significant outlier in Europe and Central Asia with just 51.4% of its population having access to safe drinking water. This is surprising given the country's natural resources, such as mountainous terrain and glacier-fed lakes. Clean water provision remains a considerable challenge, particularly in rural areas, disproportionately affecting women and girls who must fetch water frequently. Factors contributing to this issue include rapid population growth, water pricing, and meter installation problems. Addressing this requires infrastructure projects, expanded water meter installations, effective pricing mechanisms, and sanitation system investments.

3.5 Communicable Diseases per Capita



Haiti is our largest outlier for communicable diseases. This appears to be the case due to the relatively bad standard of living within Haiti compared to the other countries in Latin America. In fact, Haiti has the lowest overall standard of living of any country in Latin America which could be a large factor in why communicable diseases are so high in the country. This extreme level of poverty results in low healthcare and very poor living conditions when compared to the much more developed countries in the rest of Latin America is primarily responsible for the level of communicable diseases causing Haiti to be such a large outlier.

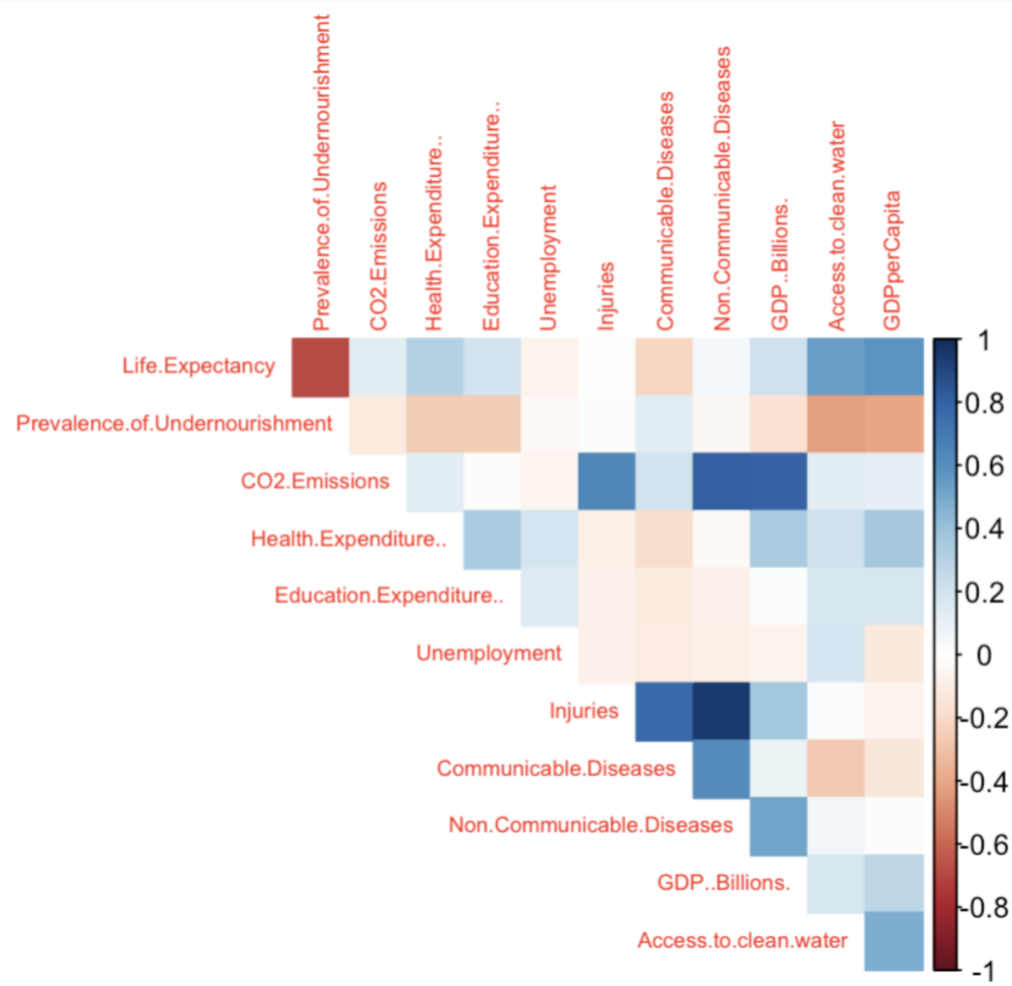
3.6 GDP per Capita



In our GDP per capita outlier analysis, Luxembourg shines as a significant outlier with the world's highest GDP. This peculiarity arises primarily from the country's heavy reliance on the banking and steel sectors, with its large banking sector playing a pivotal role in driving prosperity. The small size and population of Luxembourg result in a remarkably high GDP per capita, fostering relatively equitable wealth distribution and better living conditions compared to some larger European nations. Another factor contributing to Luxembourg's high GDP per capita is its comparatively generous wages, which often approach U.S. levels, thanks to a business-friendly environment and relaxed regulations that have fostered significant bank sector growth. This has further cemented Luxembourg's reputation as a tax haven for international investors.

Part 4: Correlation Matrix and Regression

4.1 Correlation Matrix



Before conducting our regression analysis, we constructed a correlation matrix to assess the relationships among all the variables in our dataset. We paid particular attention to the association between each independent variable and our dependent variable, life expectancy. The top row of our correlation matrix displays these specific relationships. Notably, we observed a strong correlation between life expectancy and the prevalence of undernourishment, GDP per capita, access to clean water, and, to a lesser extent, communicable diseases and health expenditures.

There is a strong negative correlation between life expectancy and the prevalence of undernourishment. The higher the percentage of undernourishment within the population, the lower the life expectancy will be. A weaker negative correlation is observed in relation to the disability-adjusted life years (DALYs) resulting from communicable diseases, where more lost years of healthy life are associated with lower life expectancy.

Additionally, there is a strong positive correlation between life expectancy and GDP per capita, access to clean water, and health expenditures. Wealthier countries, those with a higher percentage of access to clean water, and those who allocate a higher percentage of their overall spending to health expenditures, exhibit higher life expectancy. Understanding these correlations is crucial for policymakers and public health officials when crafting strategies to improve life expectancy, as it highlights the key factors that can be targeted to enhance the overall well-being and longevity of a population.

4.2 Univariate Regression Models with Outliers

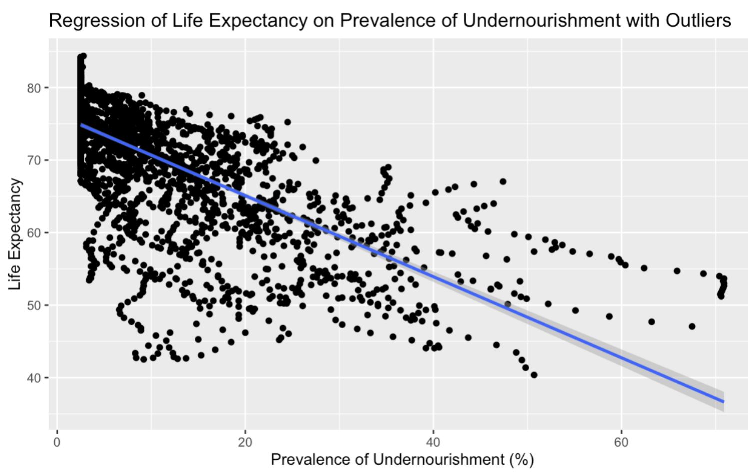
Our analysis explored the relationship between Life Expectancy (dependent variable) and multiple variables that affect it (independent variables). Two variables, "Prevalence of Undernourishment" (Adj. R-squared: 0.4714, p-value: $< 2.2e-16$) and "GDP per Capita" (Adj. R-squared: 0.3445, p-value: $< 2.2e-16$), emerged as particularly significant. Both variables showed statistical significance with very low p-values ($P < 0.005$), hinting that their relationship with Life Expectancy is not due to random chance.

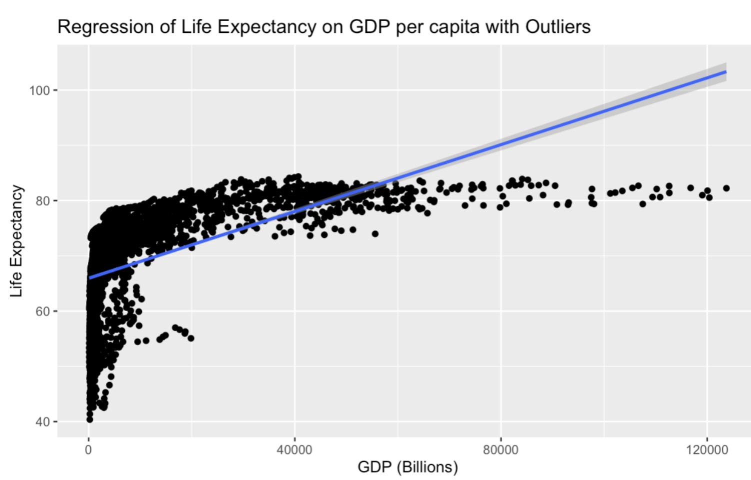
Other variables like "Health Expenditure" and "Access to Clean Water" also showed significance but to a lesser extent than GDP Per Capita and Prevalence of Undernourishment. Both were positively correlated with Life Expectancy, suggesting better water access and higher health expenditure are linked to longer life spans. This comes as no surprise as water in general is essential for sustaining life. Access to Clean water is something that our stakeholders such as NGOs, International Organizations and governments can work towards.

"Communicable Diseases" was negatively associated with Life Expectancy (Adj. R-squared: 0.04597, p-value: < 2.2e-16), indicating that higher rates of such diseases could lead to lower life expectancy.

Finally, variables like "CO2 Emissions" and "Unemployment" had low explanatory power. "Injuries" was a non-significant variable, offering almost no variation in life expectancy in a univariate regression model. This is also a variable that we would not suggest our stakeholders to focus too much on especially if they are strapped for budget. However, countries that have higher budgets and technological innovations, can still try to focus on CO2 Emissions and Unemployment numbers to improve their life expectancy.

Below are regression plots for Prevalence of Undernourishment with Outliers and GDP Per Capita with Outliers:





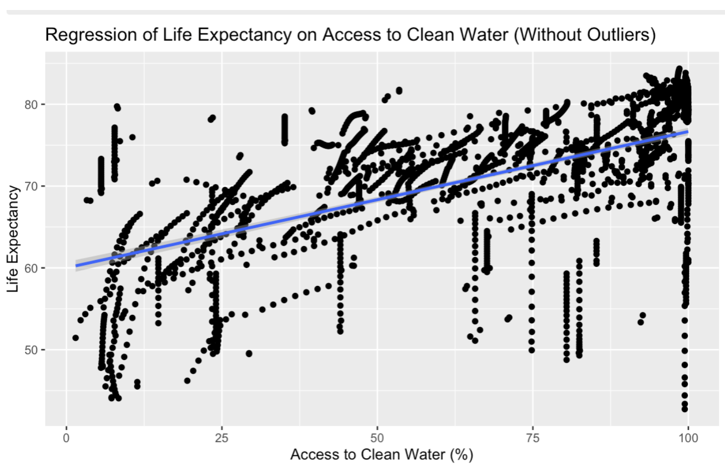
4.3 Outliers and Their Impact

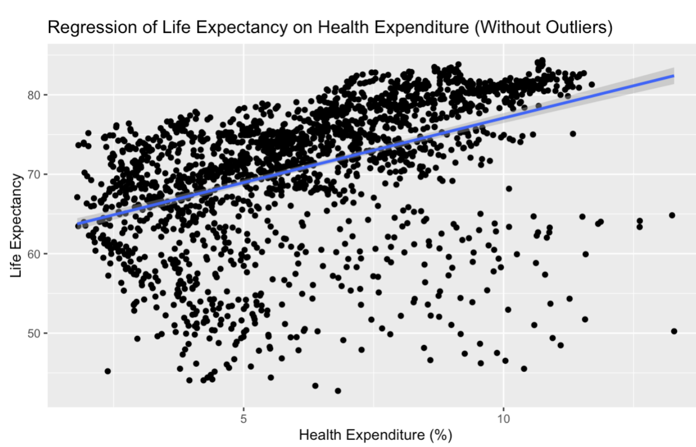
We used a Z-score method to identify and remove outliers, enhancing the reliability of our results. After removing outliers, variables like "Prevalence of Undernourishment" and "Health Expenditure" maintained their significance, however small changes were noticed in the Adjusted R-Squared values.

4.4 Univariate Regression Models without Outliers

We reran our univariate models after identifying and removing outliers via the Z-score method. The significance levels of most variables remained consistent however there were some changes in the Adjusted R-Squared values overall.

Below are plots for some of our regressions. Life Expectancy on Access to Clean Water (Without Outliers) and Life Expectancy on Health Expenditure % (Without Outliers):





4.5 Multivariate Regression Model with Outliers

```
Call:
lm(formula = Life.Expectancy ~ Prevalence.of.Undernourishment +
    CO2.Emissions + Health.Expenditure.. + Education.Expenditure.. +
    Unemployment + Injuries + Communicable.Diseases + Non.Communicable.Diseases +
    Access.to.clean.water + GDPperCapita, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-25.9515  -2.2840   0.6883   3.4637  15.5447

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     7.097e+01  5.627e-01 126.130  < 2e-16 ***
Prevalence.of.Undernourishment -4.189e-01  1.288e-02 -32.532  < 2e-16 ***
CO2.Emissions                  -2.504e-06  3.786e-07  -6.613 4.70e-11 ***
Health.Expenditure..            4.722e-01  5.417e-02   8.718  < 2e-16 ***
Education.Expenditure..        -3.939e-01  7.233e-02  -5.445 5.75e-08 ***
Unemployment                   -2.027e-01  2.081e-02  -9.738  < 2e-16 ***
Injuries                        1.687e-07  1.411e-07   1.196  0.23179
Communicable.Diseases          -2.206e-07  2.017e-08 -10.939  < 2e-16 ***
Non.Communicable.Diseases       7.527e-08  2.693e-08   2.795  0.00523 **
Access.to.clean.water           5.455e-02  4.786e-03  11.398  < 2e-16 ***
GDPperCapita                    9.929e-05  7.962e-06  12.469  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.165 on 2212 degrees of freedom
  (817 observations deleted due to missingness)
Multiple R-squared:  0.682,     Adjusted R-squared:  0.6805
F-statistic: 474.3 on 10 and 2212 DF,  p-value: < 2.2e-16
```

Life Expectancy is a complex dependent variable in our analysis and its a variable that is affected by many factors at once. Thus we wanted to analyze the effect of the independent variables on Life Expectancy in a multivariate regression mode. Thus we built a multivariate model incorporating all variables. The Adjusted R-Squared value came out to 0.6805, suggesting that the model accounts for roughly 68% of the variation in Life Expectancy. Every variable was observed to be statistically significant as shown by the R-output.

4.6 Multivariate Regression Model without Outliers

```
Call:
lm(formula = Life.Expectancy ~ Prevalence.of.Undernourishment +
    CO2.Emissions + Health.Expenditure.. + Education.Expenditure.. +
    Unemployment + Injuries + Communicable.Diseases + Non.Communicable.Diseases +
    Access.to.clean.water + GDPperCapita, data = data_no_outliers)

Residuals:
     Min       1Q   Median       3Q      Max
-26.1840  -2.2868   0.4611   3.2265  13.3725

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     6.970e+01  5.747e-01 121.269  < 2e-16 ***
Prevalence.of.Undernourishment -3.948e-01  1.503e-02 -26.266  < 2e-16 ***
CO2.Emissions                  -6.716e-06  1.157e-06  -5.806 7.42e-09 ***
Health.Expenditure..            4.558e-01  5.989e-02   7.612 4.11e-14 ***
Education.Expenditure..        -1.756e-01  7.759e-02  -2.263   0.0237 *
Unemployment                   -1.328e-01  2.403e-02  -5.528 3.65e-08 ***
Injuries                        3.244e-07  1.567e-07   2.070   0.0386 *
Communicable.Diseases          -3.294e-07  2.344e-08 -14.056  < 2e-16 ***
Non.Communicable.Diseases       2.541e-07  3.865e-08   6.575 6.16e-11 ***
Access.to.clean.water           3.755e-02  4.771e-03   7.870 5.72e-15 ***
GDPperCapita                    1.475e-04  1.117e-05  13.206  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.792 on 2035 degrees of freedom
  (693 observations deleted due to missingness)
Multiple R-squared:  0.6924,    Adjusted R-squared:  0.6908
F-statistic:   458 on 10 and 2035 DF,  p-value: < 2.2e-16
```

After removing outliers, the model's Adjusted R-Squared improved slightly to 0.6908. This enhancement suggests a more accurate and robust model. Despite the outlier removal, all variables remained statistically significant, with "Prevalence of Undernourishment," "Health Expenditure," and "GDP per Capita" being particularly significant.

We ran this model again after we ran the vif() function on our 'multivar_model_without_outliers' model. The Variance Inflation Factor (VIF) gives us an idea of how much the variance of an estimated regression coefficient increases when the predictors are correlated. VIF quantifies how much a variable is inflating the standard errors due to multicollinearity. When we ran the vif() function, we saw that all of our variables were < 5 but "Non.Communicable.Diseases" was 50.298376 and "Injuries" was 27.786934. We ran the multivariate regression again without these variables and while all variables were significant to 3 stars ($P<0.001$), the Adj. R-Squared went down to 0.6685. This is interesting because we imagined that the R-Squared value would go up or at least stay at 0.6908. This may require further investigation in the future. For now, the best model is the multivariate regression model that includes all variables without outliers.

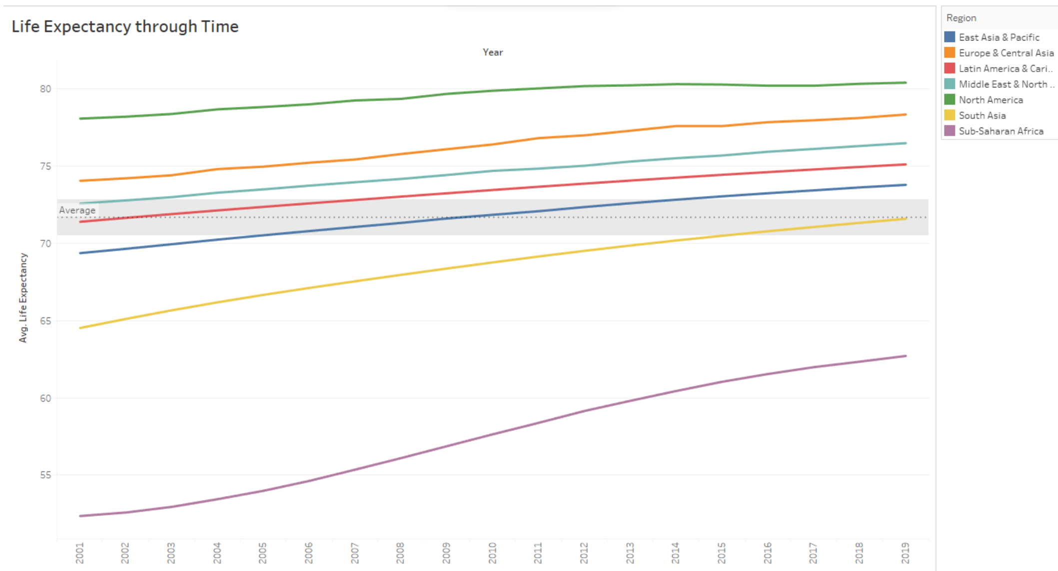
4.7 Conclusions based on Regression

Based on our regression analysis, we would strongly recommend that our stakeholders (NGOs, governments, and international organizations) focus their efforts on addressing undernourishment, GDP and healthcare expenditure as immediate priorities to improve life expectancy. Our multivariate model without outliers, which accounts for a variety of factors simultaneously, shows that the prevalence of undernourishment, GDP Per Capita and health expenditure have highly significant impacts on life expectancy. This implies that reducing undernourishment and investing in healthcare could result in significant improvements in public health. Additionally, efforts to improve access to clean water also prove beneficial, showing a strong positive correlation with life expectancy. Given that our best model explains approximately 69% of the variance in life expectancy, these targeted interventions could have the largest impact. Attention should also be given to reducing CO_2 emissions and unemployment which still show significant negative associations with life expectancy.

Part 5: Visualizations, Interpretations and Insights

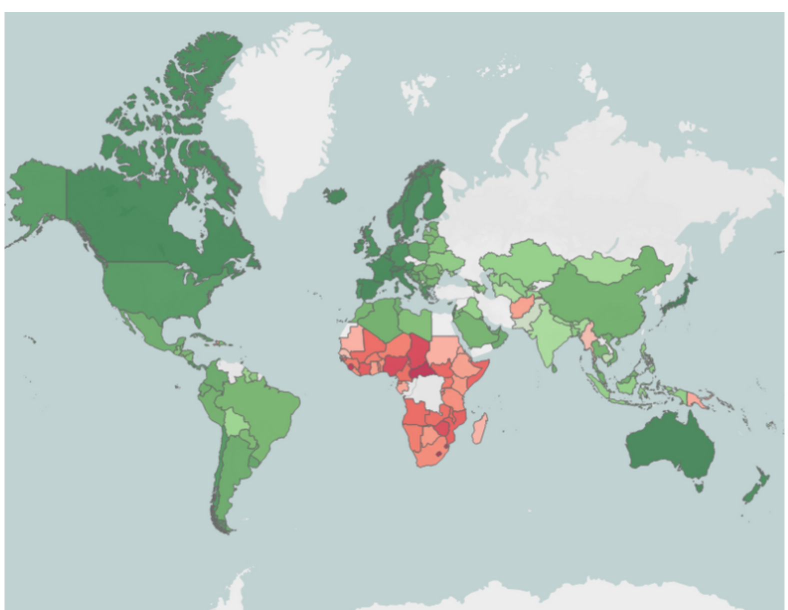
5.1 Life Expectancy through Time - Line Graph

Building upon the outcomes of our regression analysis, we embarked on a comprehensive exploration of the most influential variables, aiming to understand their behavior over time within different income groups and geographic regions.



This line graph illustrates the life expectancy trend spanning from 2001 to 2019 across seven distinct geographic regions. Our observations reveal a consistent upward trajectory in life expectancy across all regions. Notably, North America claims the top position with the highest average life expectancy, closely trailed by Europe and Central Asia. On the other hand, South Asia and Sub-Saharan Africa both exhibit life expectancies below the global average, which is about 72 years, with Sub-Saharan Africa registering a notably lower figure.

5.2 Global Life Expectancy - Map



This global life expectancy map offers a panoramic perspective on how countries compare concerning life expectancy, presenting the data using a color scale. The verdant shades indicate higher life expectancies, while the redder hues signify lower ones. This map reinforces our earlier insights from the line graph. Notably, a significant concentration of countries with the lowest life expectancy can be observed in Africa, with a distinct divergence between Northern African nations, prominently shaded in green, and the remaining African countries, which face more significant challenges.

It's worth noting that there are uncolored regions on the map, which denotes either insufficient or absent data. In our pursuit of data accuracy and consistency, we regrettably opted to exclude these areas from our analysis. However, we acknowledge the need to expand our data collection efforts in these regions to gain a more comprehensive understanding of all countries and, in turn, enhance our ability to serve a broader range of clients in the future.

5.3 Life Expectancy by Year and Region - Highlight Table

Highlight Table

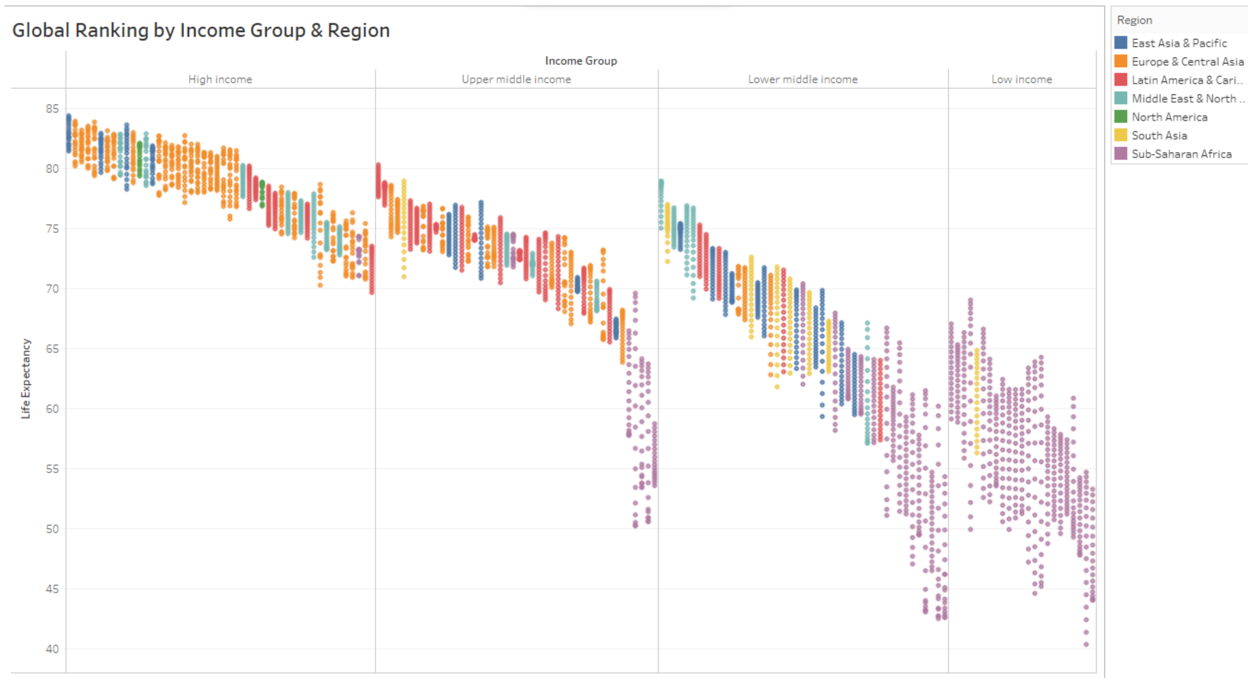
AVG(Life Expectancy)

52.35 80.42

	Region						
Year	North America	Europe & Central Asia	Middle East & North Africa	Latin America & Caribbean	East Asia & Pacific	South Asia	Sub-Saharan Africa
2019	80.42	78.36	76.50	75.12	73.80	71.60	62.72
2018	80.34	78.13	76.32	74.96	73.63	71.34	62.35
2017	80.22	77.98	76.12	74.80	73.44	71.07	61.99
2016	80.22	77.86	75.94	74.63	73.26	70.79	61.55
2015	80.30	77.61	75.70	74.45	73.06	70.50	61.05
2014	80.32	77.61	75.52	74.27	72.84	70.20	60.45
2013	80.25	77.31	75.31	74.08	72.61	69.87	59.81
2012	80.20	77.01	75.04	73.88	72.36	69.53	59.15
2011	80.05	76.83	74.85	73.68	72.09	69.16	58.39
2010	79.89	76.42	74.70	73.47	71.86	68.78	57.64
2009	79.69	76.11	74.44	73.25	71.62	68.38	56.87
2008	79.37	75.80	74.18	73.04	71.33	67.97	56.10
2007	79.27	75.44	73.97	72.82	71.07	67.55	55.35
2006	79.02	75.23	73.75	72.60	70.80	67.12	54.62
2005	78.84	74.98	73.51	72.37	70.53	66.67	53.98
2004	78.69	74.82	73.29	72.14	70.25	66.19	53.44
2003	78.39	74.42	73.00	71.90	69.95	65.67	52.94
2002	78.21	74.22	72.79	71.66	69.65	65.11	52.57
2001	78.09	74.06	72.58	71.40	69.37	64.52	52.35

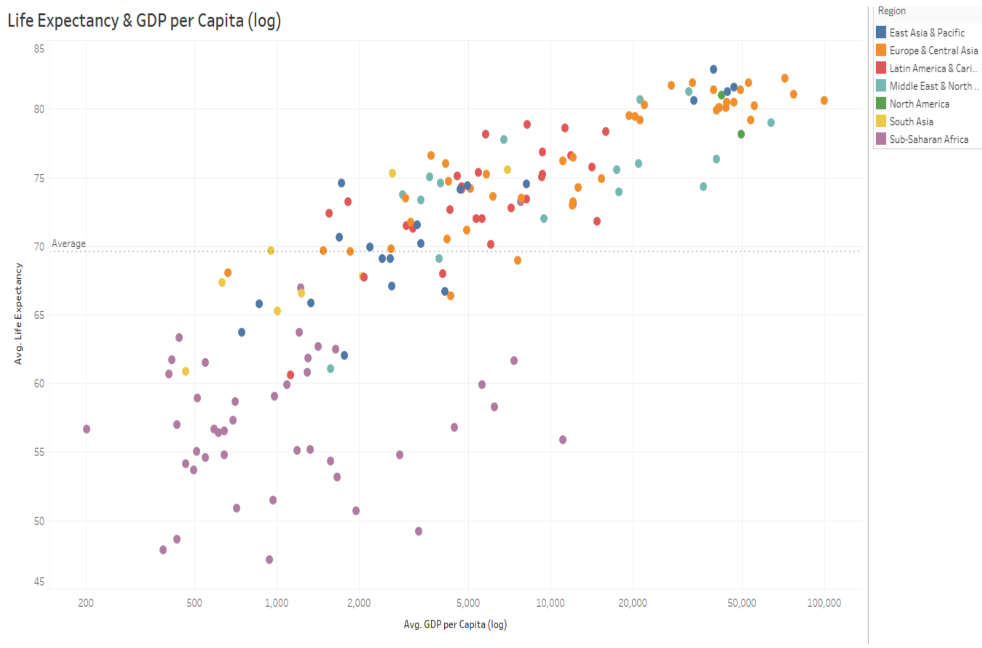
This highlight table examines life expectancy across different regions over the years. The table presents a chronological perspective on life expectancy for each region. Notably, North America claims the highest life expectancy in this specific table, which is intriguing, given that it comprises only two countries, the United States and Canada. Equally noteworthy is the East Asia & Pacific region, which, despite including nations with notably high life expectancies such as Japan, Australia, New Zealand, and Singapore, still reports a comparatively lower average. This phenomenon can be observed in various regions, where some countries within a region perform exceptionally well, while others face challenges and conflicts that impact their life expectancies negatively. Nevertheless, the overarching trend remains positive, as the global average life expectancy steadily rises year after year, with an expectation of continued improvement. This underlines the collective efforts toward better health and well-being worldwide.

5.4 Global Ranking by Income and Region - Scatter Plot



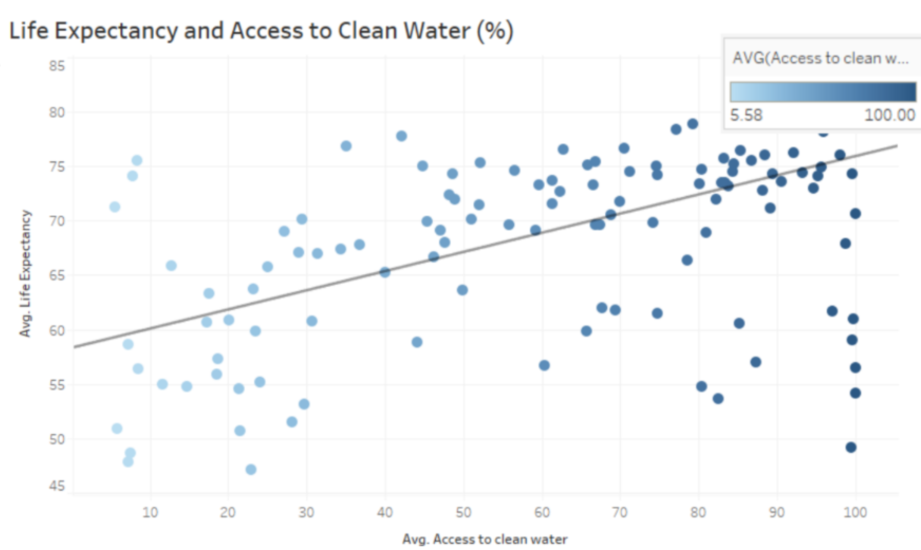
This plot visually presents the global ranking categorized by income group and region. Each vertical sequence of circles corresponds to a specific country, and each individual circle within represents a particular year. One overarching trend becomes evident: a general decline in life expectancy as income decreases. Among the high-income countries, a significant proportion hails from Europe and Central Asia, with some notable exceptions in East Asian and Pacific regions, exemplified by Japan, which ranks exceptionally high in terms of life expectancy. As we move into the upper-middle and lower-middle income brackets, there's greater diversity in the regions represented. However, as we approach the lower end of the income scale, Sub-Saharan African countries dominate, bearing both low income and life expectancy, with their life expectancy figures significantly trailing behind the top-ranking countries.It's worth noting that the arrangement of dots for each country offers intriguing insights. The Sub-Saharan countries exhibit a wider dispersion of dots, indicating substantial variations in life expectancy changes over the past two decades. This underscores the dynamism and challenges experienced in these nations during this period.

5.5 Life Expectancy and GDP per Capita - Scatter Plot



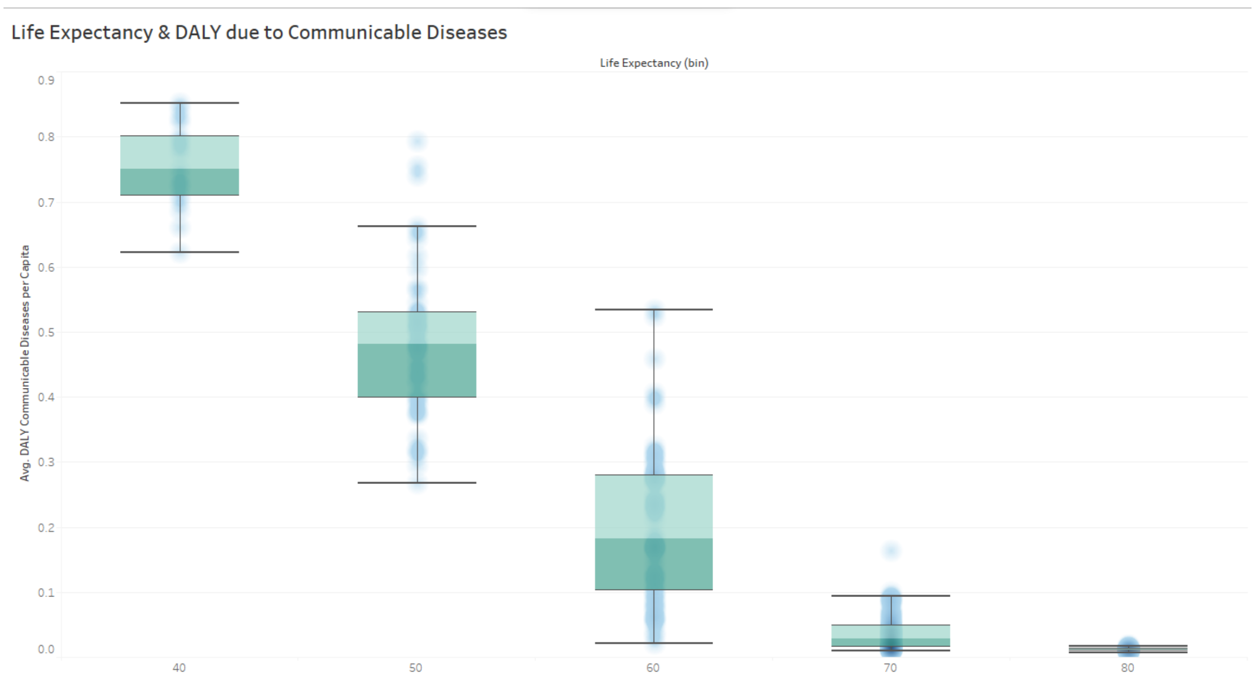
This scatter plot visually depicts the correlation between life expectancy and GDP per capita, employing a color scale to distinguish various regions. For the sake of data analysis, we applied a log transformation to GDP per capita values, while retaining the original values on the x-axis to ensure easier interpretation. The plot reveals a robust positive association between the average life expectancy and the average GDP per capita. This strong correlation shows that as a country's GDP per capita increases, its life expectancy tends to rise in tandem. This insight underscores the substantial impact of economic prosperity on a nation's overall well-being and healthcare outcomes.

5.6 Life Expectancy and Access to Clean Water - Scatter Plot



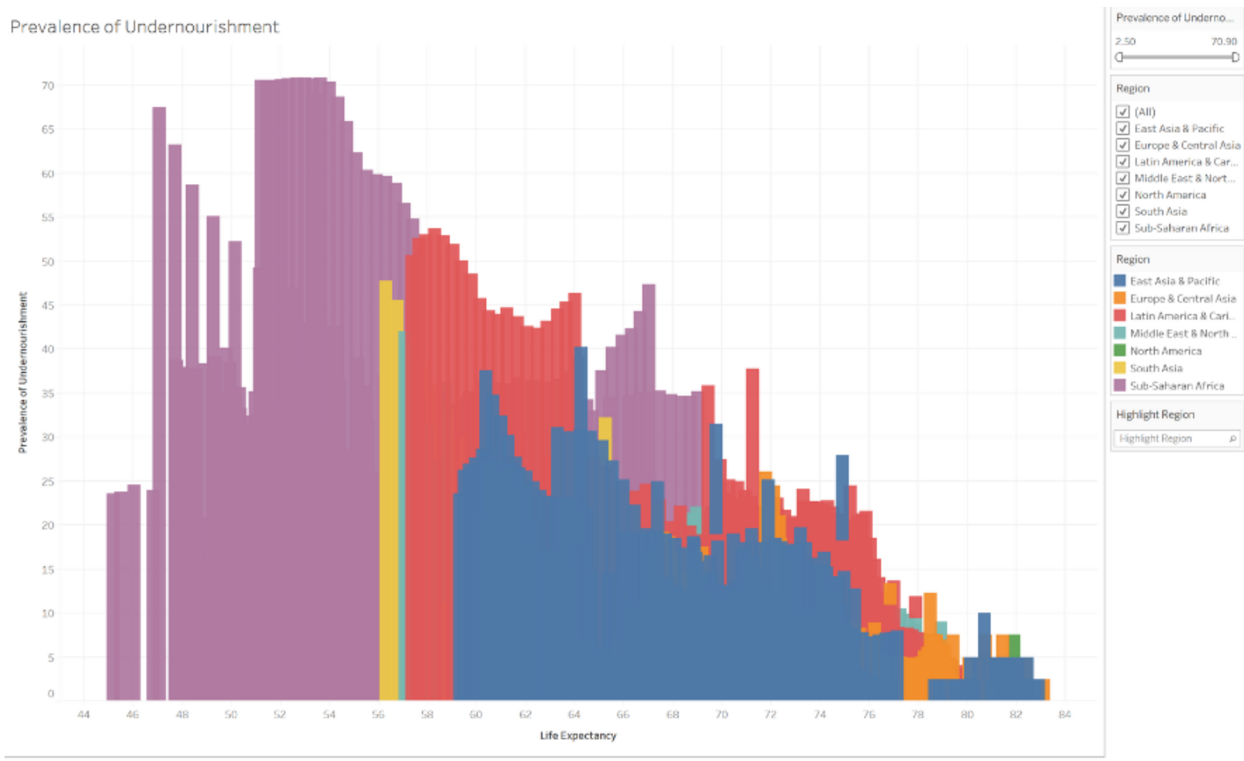
This scatter plot illustrates the connection between life expectancy and the percentage of a population with access to clean water. The shade of the circles deepening into darker blue signifies countries with higher rates of access to clean water. Within this plot, a discernible moderate positive correlation emerges, suggesting that as the percentage of the population with access to clean water increases, there's a corresponding uptick in life expectancy. This underlines the crucial relationship between access to clean water and the overall health and longevity of a population. Access to clean water is vital because it serves as a fundamental cornerstone of public health and well-being. Clean water access contributes to the reduction of waterborne diseases, improves overall health, and enhances the quality of life by promoting sanitation and hygiene, making it an indispensable factor in fostering healthier and more prosperous communities.

5.7 Life Expectancy and Communicable Diseases - Box and Whisker Plot



This box and whisker plot provides insight into the distribution of disability-adjusted life years (DALY) attributed to communicable diseases, encompassing infectious conditions such as HIV, malaria, and tuberculosis. DALY represents the cumulative impact of years of healthy life lost due to premature mortality and years lived with a disability linked to communicable diseases. In this visualization, we've categorized countries into five distinct life expectancy groups and further calculated DALY per capita to ensure a more equitable comparison across all nations. In a general observation, it becomes evident that countries with higher life expectancies tend to exhibit lower DALY per capita figures. Conversely, countries with the lowest life expectancies feature a median DALY per capita around 0.75, implying that the individuals in these countries endure a loss of over half a year in full health due to communicable diseases. Notably, countries within the 70 and 80-year life expectancy bins showcase narrower box plots, indicating reduced variation in DALY per capita. At the upper ranks of life expectancy, individuals experience almost no years of full health lost, underscoring the tremendous effectiveness of healthcare systems and disease control measures in these regions. This highlights the pivotal role of healthcare access and intervention strategies in preserving and extending the overall quality of life for populations.

5.8 Life Expectancy and Prevalence of Undernourishment - Histogram



In this plot, we're examining the prevalence of undernourishment, and what truly stands out is the stark contrast in undernourishment rates for Sub-Saharan African countries depicted in purple, in comparison to the rest of the world. This striking discrepancy points to a considerably higher prevalence of undernourishment in this region, underscoring a critical global concern. Equally intriguing is the dispersion of undernourishment within the East Asia and Pacific region. Notably, undernourishment decreases significantly in countries with the highest life expectancies. This observation serves as a reminder of the interconnectedness between health and nutrition. It indicates that as life expectancy increases, there's a corresponding reduction in undernourishment, emphasizing the crucial role that access to proper nutrition plays in enhancing overall health and well-being.

5.9 Top/Bottom 5 Countries Analysis

To further understand the complexities within life expectancy, we decided to perform further analysis but this time comparing the top and bottom 5 countries for life expectancy. We believe that by comparing these countries, it might give us further insights on what may be a big predictor of life expectancy.

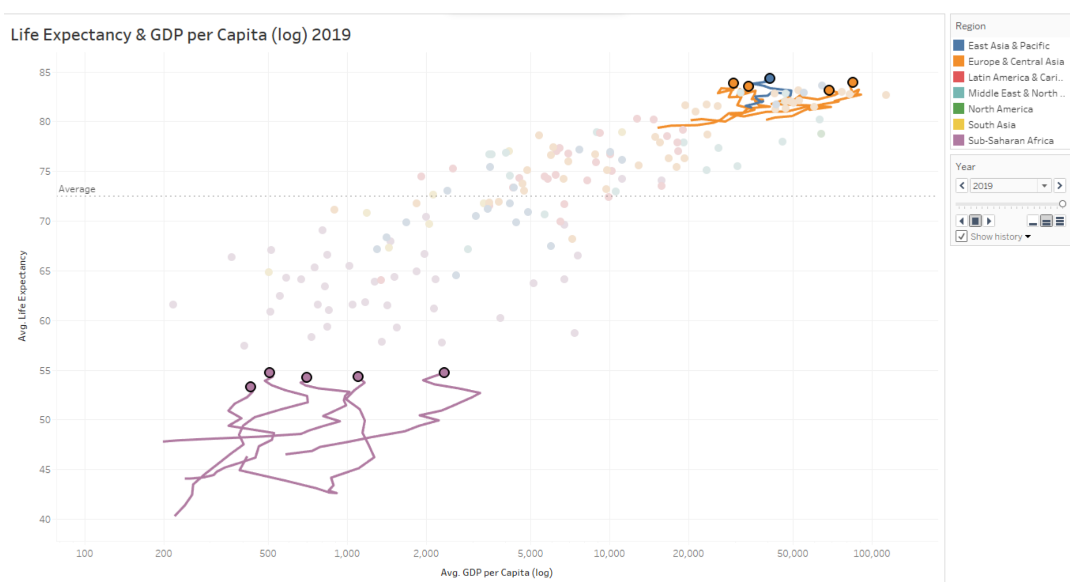
Top 5:

- Japan
- Switzerland
- Spain
- Italy
- Iceland

Bottom 5:

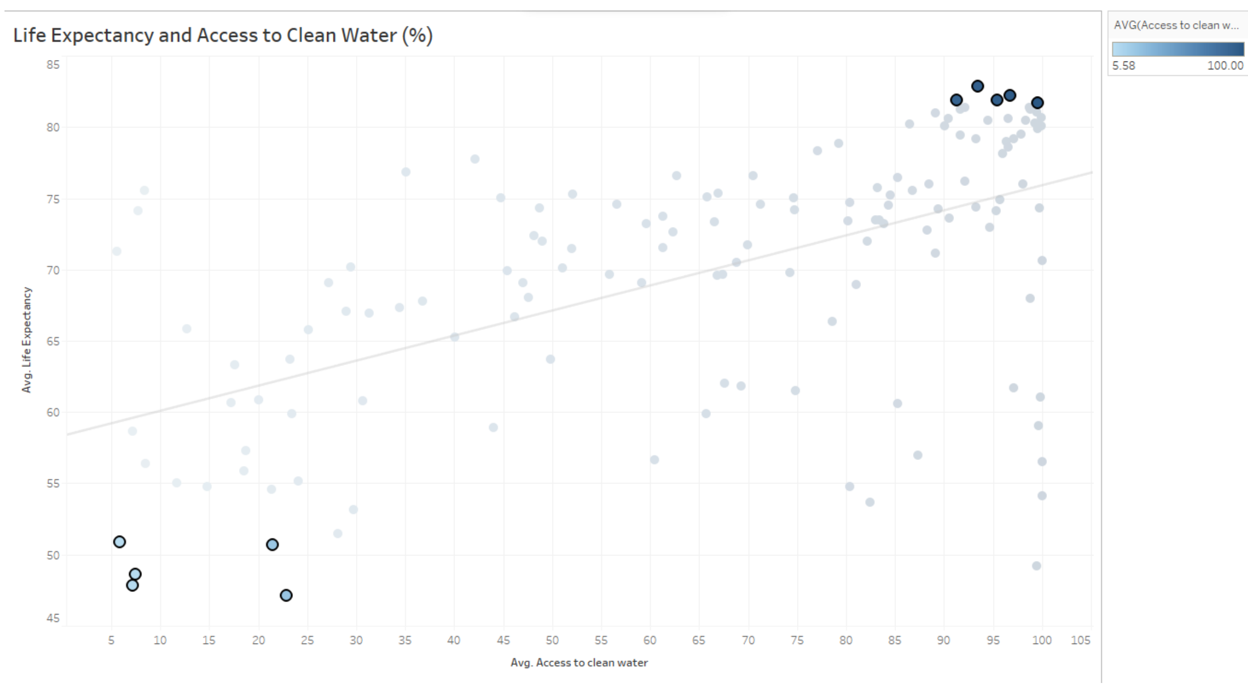
- Eswatini
- Central African Republic
- Sierra Leone
- Lesotho
- Nigeria

5.10 Life Expectancy and GDP per Capita - Scatter Plot



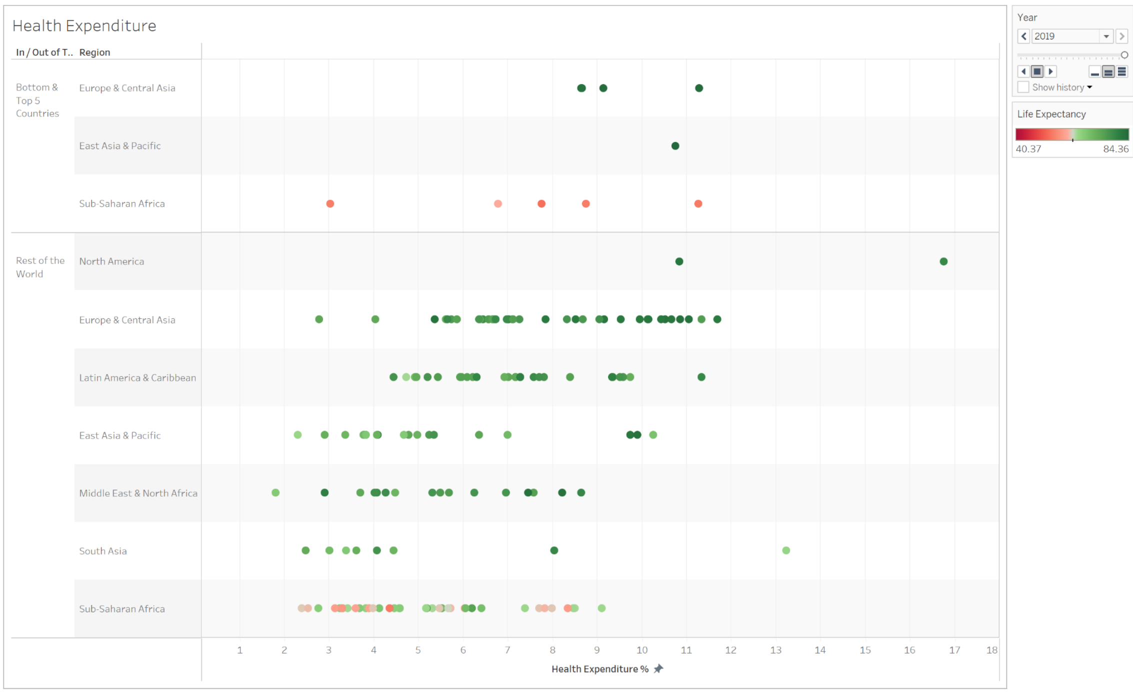
In this graph, there are multiple insights that can be acquired from it. The line following each of the top/bottom countries means its path throughout the years. This means that throughout the years both groups of countries have increased their life expectancies, which is a good thing. Furthermore, it seems that the lower your life expectancy is, the bigger your jumps are going to be when improving. It seems that when you reach a specific level of life expectancy your percentage increases would not be as high. Furthermore, the GDP per capita seems to play a role in these countries. For the bottom 5, the GDP per capita combined between all 5 in 2019 was about $7,703 USD. In comparison the GDP per capita in 2019 for only Japan was about $42,600 USD which is 553% more than the combined GDP per capita of the bottom 5 countries. Then, if the top 5 group GDP per capita from 2019 is combined then it would yield about $243,000 USD. This yields about 3,155% more GDP per capita. As it is clearly seen, there are huge differences in the GDP per capita between the top and bottom countries which might be contributing to life expectancy.

5.11 Life Expectancy and Access to Clean Water - Scatter Plot



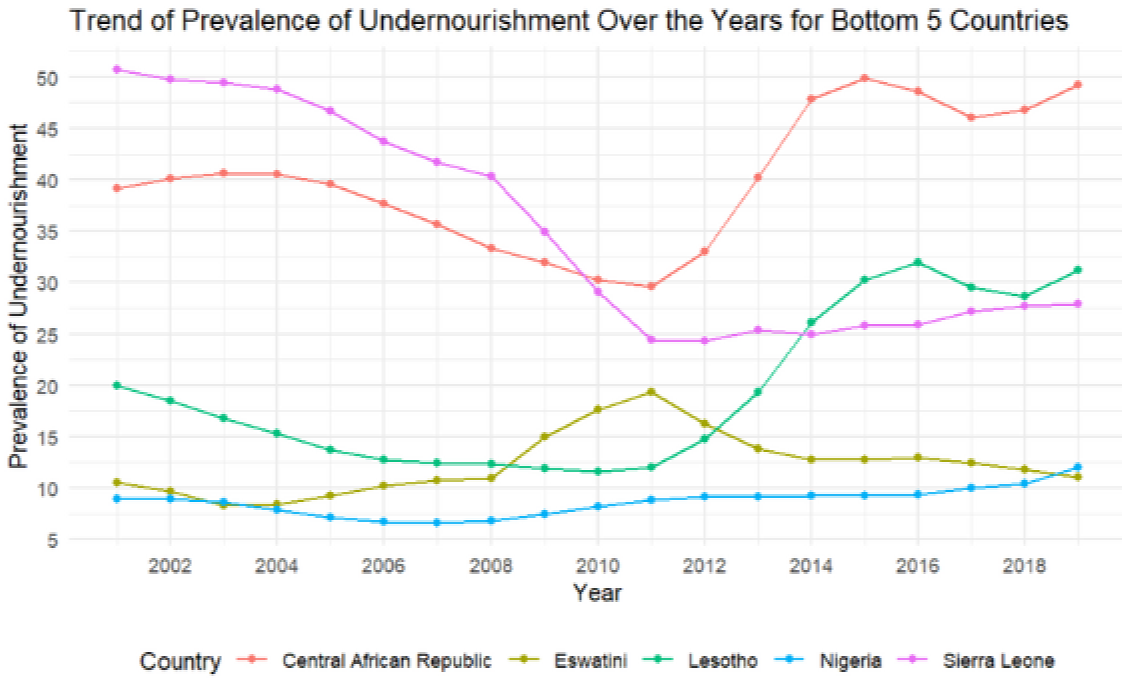
When it comes to access to clean water there are also stark differences between the two groups. It seems that the bottom group has had a lot of issues regarding access to clean water. The bottom group has between 5-25% access to clean water while the top group is in the 90-100% range. This itself tells us that the lower your access to clean water, the more likely you are susceptible to lower life expectancy. There are multiple reasons why these bottom countries have struggled to provide. Many of these countries are experiencing rapid population growth while at the same time having economic and infrastructure issues. When a country has a growing population and it cannot keep up with it, it will experience problems such as not being able to provide access to clean water to almost everyone. Other reasons can be attributed to internal conflicts, education, climate change, and much more. Then, with the top countries, they tend to have great infrastructure, public awareness, strict regulations on water and much more.

5.12 Life Expectancy and Health Expenditure - Strip Plot



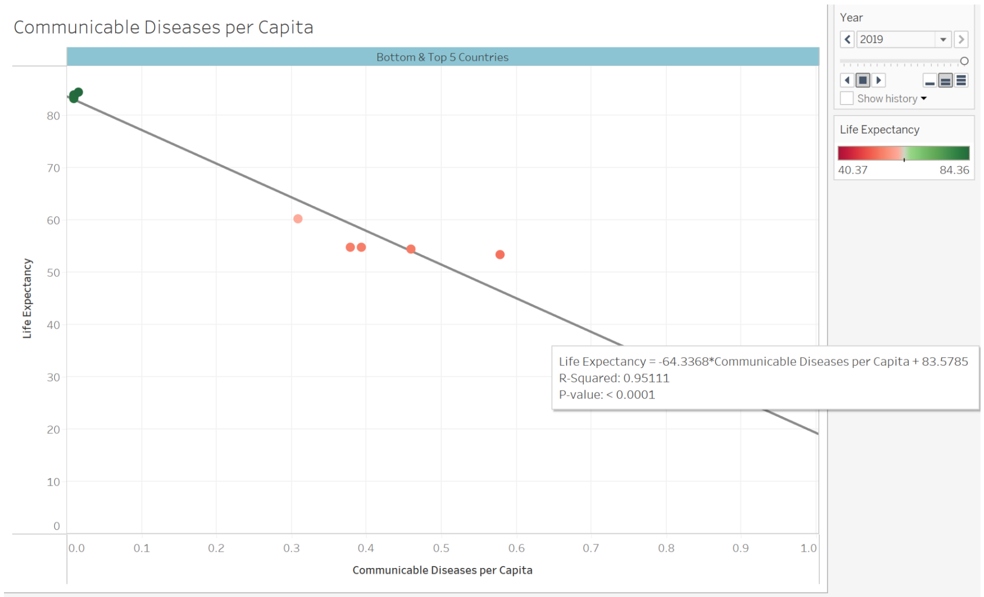
Looking at the graph provided above, the top/bottom 5 countries do not differ much from the general population. Both groups of countries seem to be very close to the world range. While this may be a little paradoxical, a lot of this seems to be in the allocation of resources and different needs. For example, even though the health expenditure % might be lower between a country in Europe and Africa, the actual money spent will probably be significantly larger in Europe instead of Africa. There have been efforts by the bottom group to improve their health expenditure, but different challenges have made it extremely difficult for these countries. Many of these countries do not have adequate infrastructure, they are still in conflict or recovering from conflicts, and many prefer to allocate their resources on only urban areas. In contrast, the top group has better technological health innovations and infrastructure. They also have more money to spend to allocate their resources more efficiently.

5.13 Prevalence of Undernourishment and Life Expectancy - Line Graph



Prevalence of undernourishment has been unstable in the bottom 5 countries throughout the last few years. Some of them have seen a more or less downward trend, while some other ones have seen an upward trend. For example, Central African Republic is the one with the highest prevalence of undernourishment from these countries. In 2013, it went into a conflict that has been ongoing even up until today. As is it apparent in the graph, the biggest spikes/increases happened between 2012-2014. Then, when it comes to Lesotho, it appears that it has been facing high political instability throughout the years too. In 2014 there was an attempted coup. Overall, having political instability affects the country as a whole, because the allocation of resources is not distributed properly. This along with other factors such as geographical location, infrastructure, environmental challenges and more make these countries experience high prevalence of undernourishment. On the other hand, countries such Eswatini have been doing a good job in decreasing the prevalence of undernourishment. They have been able to do this by supporting small-scale farmers, increasing nutrition education, collaborating with international organizations, and implementing food security and safety net programs. In regards to the top group, they provide education programs, food banks for people in need, and more.

5.14 Communicable Diseases per Capita and Life Expectancy - Scatter Plot

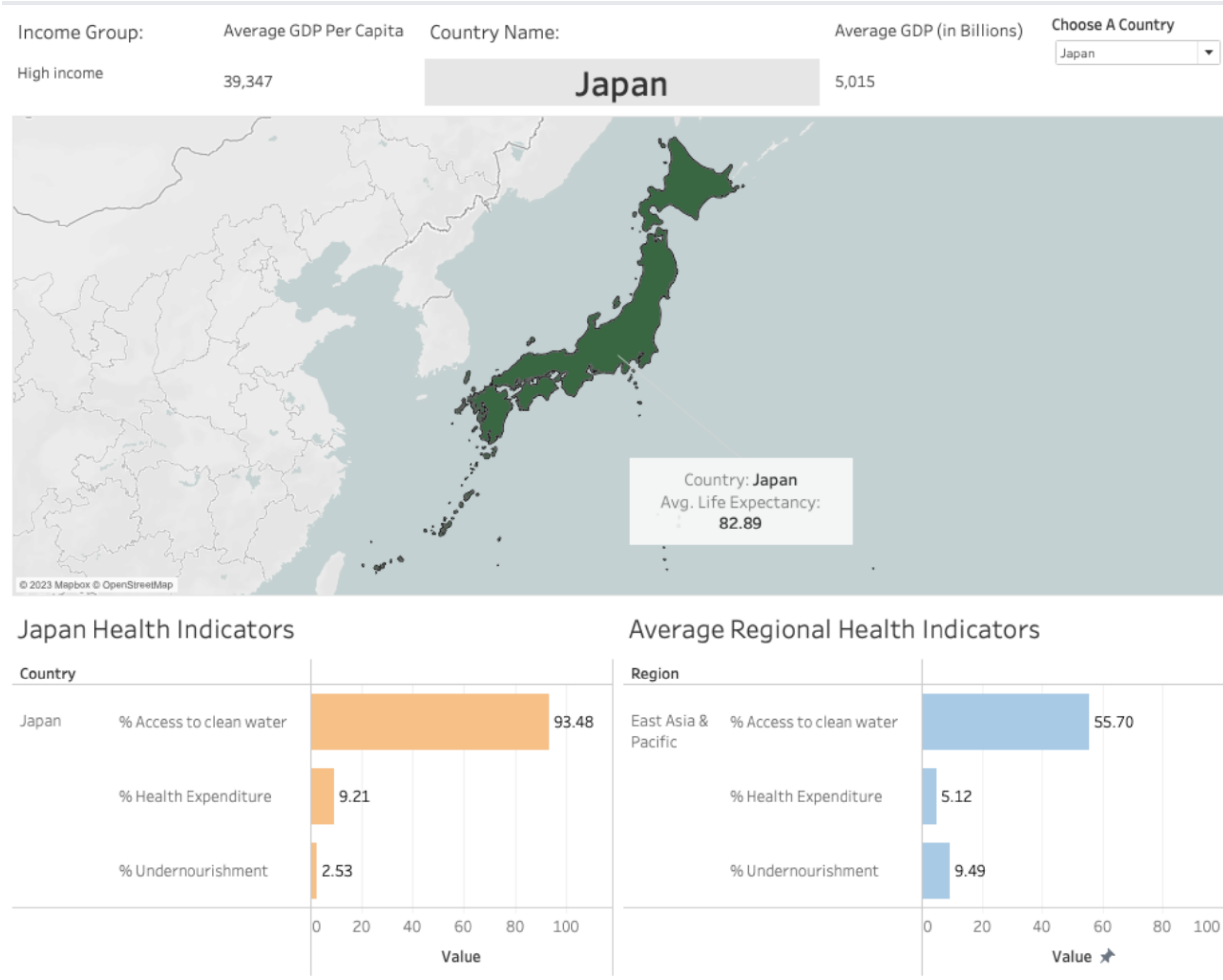


This particular dataset explores communicable diseases and DALY (explained earlier). As is apparent, there seems to be a high correlation, at least in 2019, between communicable diseases per capita and life expectancy. The higher the DALY attributed to communicable diseases, the lower the life expectancy is going to be. The top 5 countries seem to have very few numbers per capita while the bottom 5 seem to be about 3 times higher. This could be caused by a very high variety of factors but many can include better allocation of resources in regards with health expenditure, better education on diseases, and more. Overall, there seems to be different disparities that affect life expectancy as a whole. Of course, it is important to recognize that everything interacts with one another so it is a multi-faceted problem.

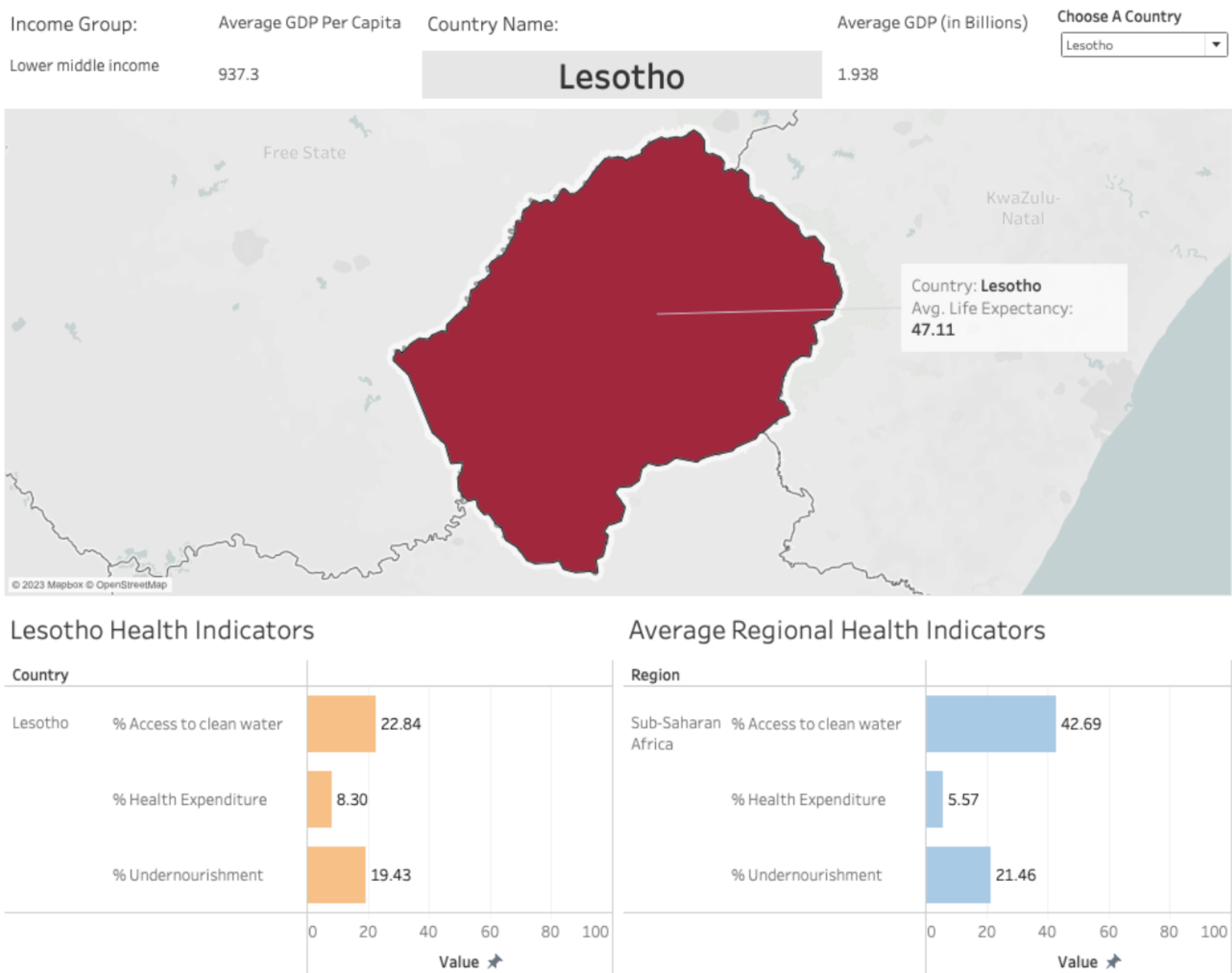
Part 6: Tableau Dashboard

6.1 Country Dashboard

When it came time to put together our presentation we decided to create a set of dashboards to talk about some examples from our research. In making these dashboards we focused on our 5 biggest variables in determining life expectancy based on our regression analysis, Prevalence of Undernourishment, Health Expenditure, Communicable Diseases, GDP per Capita, and Access to Clean Water. By placing these variables in one place it made it easier to walk the class through an example of how these variables affect life expectancy and helped us pivot our presentation into our recommendations for each of these variables.



Income Group:
High income
Average GDP Per Capita
39,347
Country Name:
Japan
Average GDP (in Billions)
5,015
Choose A Country
Japan
Country: Japan
Avg. Life Expectancy:
82.89
© 2023 Mapbox © OpenStreetMap
Japan Health Indicators
Country
Japan
% Access to clean water
93.48
% Health Expenditure
9.21
% Undernourishment
2.53
0 20 40 60 80 100
Value
Average Regional Health Indicators
Region
East Asia & Pacific
% Access to clean water
55.70
% Health Expenditure
5.12
% Undernourishment
9.49
0 20 40 60 80 100
Value



The above two dashboards are examples of our country dashboard. This allows us to select a specific country and see some of the main information important about the country. We can see from the top row of information the difference in income between the two countries. Japan is a high income country with a large GDP and GDP per capita, compared to Lesotho which is a lower income country with a very small GDP and GDP per capita. Below the visual of the country you can see the Health information of the country selected and of the overall region it belongs to. Here we see that Japan not only has high health indicators itself, but it beats all indicators of the region it falls in. In comparison, Lesotho has much lower health indicators and falls under the average for the region it belongs to.

6.2 Region Dashboard

Average GDP (in Billions)

403.1

Average GDP Per Capita

24,281

Region Name:

Europe & Central Asia

Average Communicable Disease Per Capita

0.02179

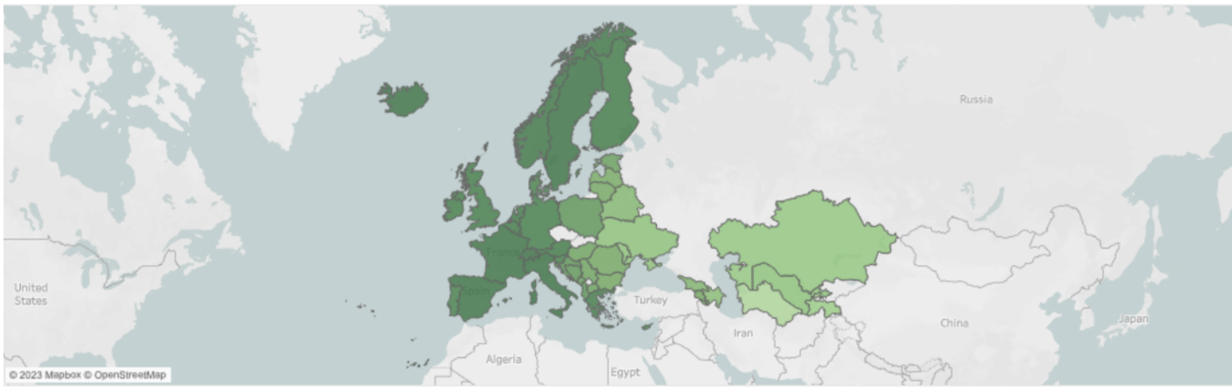
Region

Europe & Central Asia

AVG(Life Expectancy)

47.11 82.89

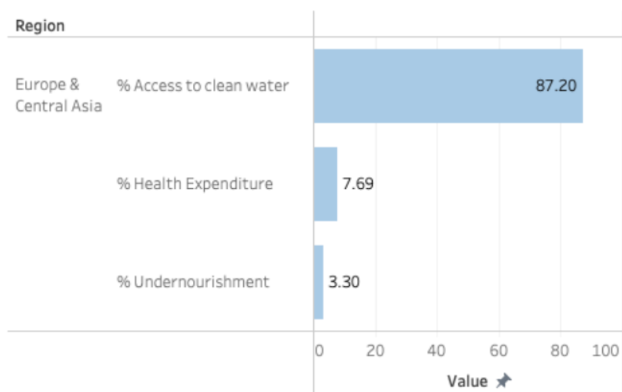
Region Map

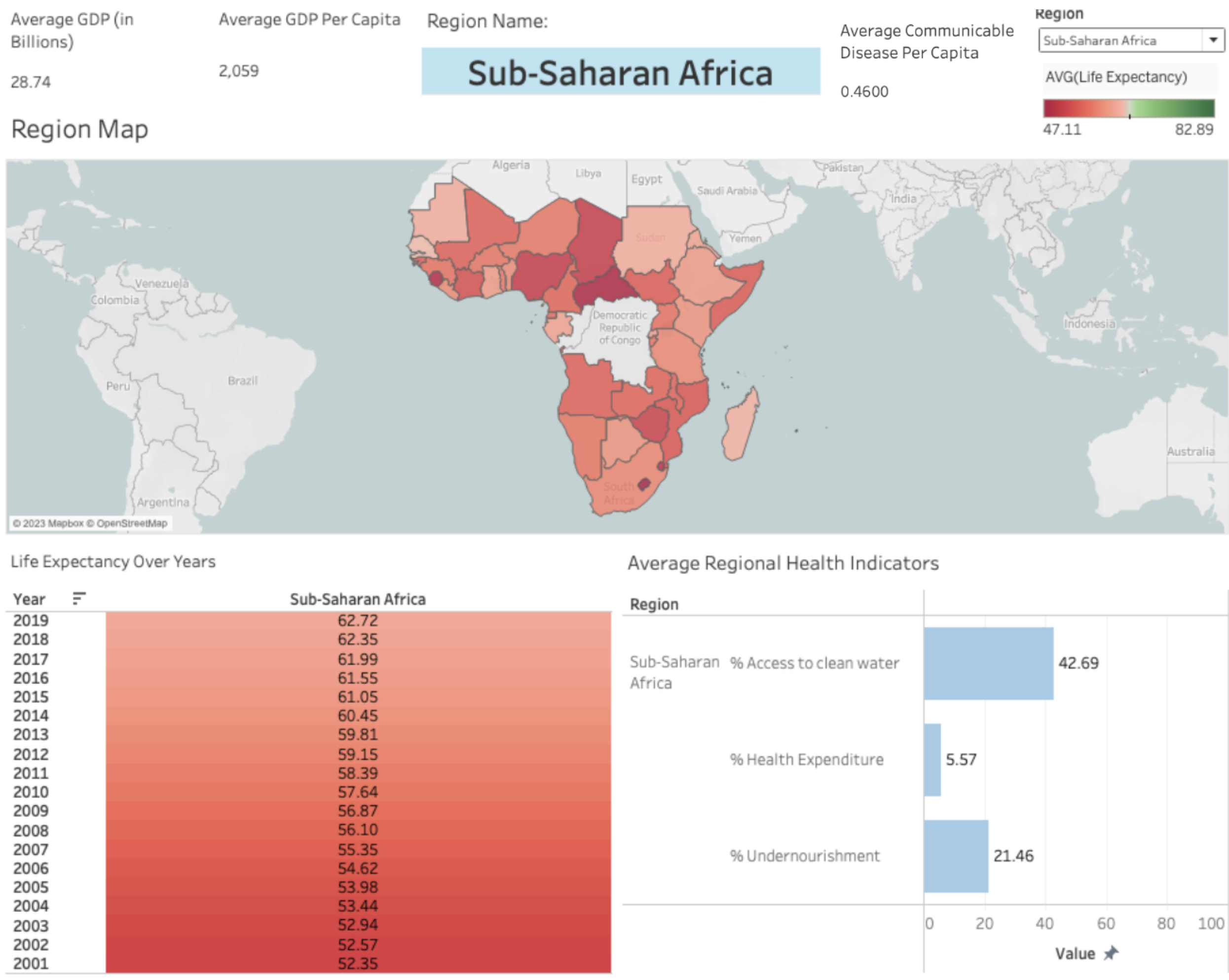


Life Expectancy Over Years

Year	Europe & Central Asia
2019	78.355
2018	78.132
2017	77.982
2016	77.859
2015	77.606
2014	77.609
2013	77.305
2012	77.006
2011	76.827
2010	76.417
2009	76.108
2008	75.796
2007	75.443
2006	75.232
2005	74.978
2004	74.818
2003	74.419
2002	74.222
2001	74.060

Average Regional Health Indicators





The above dashboard are examples of our region dashboard. In these two examples we compare the highest life expectancy region Europe to the lowest life expectancy region Sub-Saharan Africa. We can see from the top corner the vast difference in GDP and GDP per capita between the regions. African countries are much poorer than their counterparts in Europe. We also see the addition of the average communicable diseases per capita. This variable tells how many years of life are lost by living in a region based solely on disabilities. Looking at these variables we can see that people living in Sub-Saharan Africa lose on average a half year of their life due to the prevalence of disabilities in the region, while those living in Europe lose approximately 7 days of their life on average to disabilities. We can also see that health indicators for the region are much higher in Europe than in Sub-Saharan Africa. On the positive side for Sub-Saharan African we can see a larger growth in life expectancy over the years, rising 10 years from 2001-2019 which is a larger increase than in Europe. This gives us hope that the life expectancy in the region could continue to rise closer to that of the rest of the world if positive progress is made to the major variables we have referenced throughout our project.

Part 7: Recommendations

In our comprehensive analysis of world life expectancy data, we employed advanced statistical techniques, including linear and multivariate regression, to unravel the key determinants of life expectancy. This rigorous examination unveiled five variables that overwhelmingly define the variations in life expectancy across different regions and countries. Our recommendations, thus, are centered around these pivotal variables, which have the most substantial impact on the length and quality of human life. These recommendations aim to guide policymakers, healthcare providers, and global organizations in fostering sustainable improvements in life expectancy and overall well-being.

7.1 Gross Domestic Product per capita

Gross Domestic Product (GDP) per capita, a fundamental indicator of economic well-being, emerged as one of the most influential variables in our analysis. Understanding and enhancing GDP is a multifaceted endeavor, and as such, there is no one-size-fits-all solution. Rather, GDP growth hinges on a myriad of adjustments within fiscal and monetary policies. A spectrum of measures can be taken, including stimulating production across sectors, implementing strategic tax cuts, easing regulatory burdens, and investing in vital infrastructure projects. Importantly, the impact of these adjustments varies across countries, depending on their stage of economic development. Notably, for regions like Sub-Saharan Africa, where lower life expectancies are prevalent, the emphasis should be on infrastructure development. It is clear from our analysis that investing in infrastructure represents the foremost driver to propel GDP growth, ultimately contributing to improved life expectancy by enhancing living standards and healthcare accessibility.

7.2 Health Expenditure

Healthcare expenditure, our second crucial variable, plays a pivotal role in shaping life expectancy outcomes worldwide. To enhance this variable, we propose a multifaceted approach that encompasses strengthening primary healthcare services, enhancing healthcare infrastructure, and making healthcare more accessible and affordable for all citizens. Furthermore, implementing effective disease control programs for prevalent health issues is paramount, as is promoting health literacy and fostering healthy living habits within communities. The World Health Organization (WHO) recommends allocating at least 6% of GDP to healthcare, providing a benchmark for nations to strive towards.

An example of the efficacy of these recommendations can be found in the remarkable transformation of Rwanda's healthcare system. In 1994, Rwanda faced a life expectancy of just 49 years, largely attributed to the devastating impact of the genocide. However, through a deliberate focus on strengthening its primary healthcare services, improving infrastructure, and implementing disease control programs, Rwanda managed to increase its life expectancy to 67 years by 2019. Notably, Rwanda's commitment to making healthcare affordable and accessible to its citizens played a pivotal role in this success story, demonstrating that these recommended strategies can indeed yield substantial positive outcomes and enhance overall well-being.

7.3 Communicable Diseases

The reduction of communicable diseases is paramount for extending life expectancy, and countries can take several measures to achieve this goal. These measures include strengthening healthcare infrastructure, implementing effective vaccination programs, promoting public health education, and improving water and sanitation systems to reduce the transmission of waterborne diseases. An illustrative example of these strategies in action can be found in Cuba, where a comprehensive healthcare system, community-based preventive services, and public health education initiatives have played a pivotal role in controlling diseases like tuberculosis and HIV/AIDS. Such a multifaceted approach to combating communicable diseases can lead to significant increases in life expectancy and overall well-being.

7.4 Access to Clean Water

Access to clean water, a vital variable in our analysis, significantly impacts life expectancy. Recommendations to improve access to clean water encompass awareness campaigns, aligning water quality with WHO standards, and investing in water infrastructure. In Japan, stringent regulations, technological innovations, and community involvement ensure clean water access, contributing to the nation's high life expectancy. Additionally, countries with limited resources can adopt cost-effective purification technologies and decentralized water management to address local needs, making strides in providing clean water access even in resource-constrained settings. These measures can substantially enhance life expectancy by addressing this fundamental determinant of public health.

7.5 Prevalence of Undernourishment

Addressing the prevalence of undernourishment, a key variable impacting life expectancy, requires a comprehensive strategy. Priorities include enhancing food access and quality, establishing nutritional rehabilitation centers, and promoting food fortification programs. A noteworthy example of effective policy can be found in Ethiopia's Productive Safety Net Program, which targets individuals with high food insecurity, providing cash or food assistance while encouraging community development and participation in public works projects. This program not only bolsters livelihoods but also empowers women and strengthens local governance. By implementing such strategies, nations can substantially mitigate undernourishment, leading to increased life expectancy and improved overall well-being.

Part 8: Reflections about the journey

Taking on the Life Expectancy project has been an enlightening and rewarding experience for our team. The journey from raw data to data wrangling and meaningful insights had its fair share of hurdles and successes.

8.1 Challenges & Learnings

One of the biggest challenges of our project was dealing with the intricacies of multivariate regression analysis. The large number of variables in our dataset required strong attention to detail. It was important for us to not only understand the statistical significance but also interpret these findings in a real-world context. Thus we first conducted univariate regression analysis with and without outliers and then conducted multivariate regression analyses. This led us to appreciate the complexity of factors that contribute to Life Expectancy.

We also faced challenges with handling outliers. The decision to remove or keep outliers had a noticeable impact on our regression models. This experience taught us the importance of robust data cleaning and validation techniques. In dealing with outliers, our team was also able to make use of our skills in the Python programming language and the Pandas and NumPy libraries specifically. We were able to find a good method for addressing nulls by calculating mean by country for our variables and filling in the missing values with country-wise means.

As a team we also faced the challenge of differing schedules so we found a hybrid working schedule. Our team met over Zoom at times and in person at other times. We spent a lot of time just looking at our dataset and thinking about ways we could tackle it to gain meaningful insights. We were all open to sharing our ideas, and we were all receptive to each other's opinions.

Something our team did very well was reach out to our Professor whenever we needed some extra guidance or clarification. We set up meetings with our Professor to go over expectations and ask for their help when we needed it. One such example was the use of the VIF function in our regression to check for multicollinearity. Our team had not thought about checking for that so our professor was able to guide us through it. This helped us gain a deeper understanding of our models and of the issues multicollinearity can cause in regression models.

8.2 Future Directions

While our multivariate regression model accounts for approximately 69% of the variance in Life Expectancy, there's room for improvement. We have thought about getting additional variables to explain variation in Life Expectancy such as quality of healthcare, measures of social stability, corruption data, crime data and birth rate data. All of these variables could be incorporated to provide a more comprehensive understanding of Life Expectancy and it could make our model more robust and predictive.

Something else we have thought about doing is incorporating the use of Ridge or Lasso regression techniques for regularization. These methods could help us manage multicollinearity among variables and provide a better fit, especially when working with a large number of predictors for Life Expectancy. They can fine-tune the model by penalizing certain coefficients, potentially leading to a more robust model.

We could also delve deeper into variables that showed high levels of multicollinearity in the VIF tests. While these variables were retained in the current model, understanding their relationship could offer a more nuanced interpretation. One such example is the high correlation between Injuries and Non-Communicable diseases and Injuries and Communicable diseases.

Additionally, more sophisticated statistical methods such as machine learning algorithms like Random Forest or Gradient Boosting could be employed. These algorithms could capture more complex, non-linear relationships between variables that our linear model may not fully capture. Finally, Time Series Analysis can also be conducted on our data as it spans multiple years.

Part 9: References

World Bank. "GDP (Current US$) | Data." *The World Bank*, 2023, data.worldbank.org/indicator/NY.GDP.MKTP.CD.

World Bank. "Life Expectancy & Socio-Economic (World Bank)." *www.kaggle.com*, www.kaggle.com/datasets/mjshri23/life-expectancy-and-socio-economic-world-bank.

World Bank. "People Using Safely Managed Drinking Water Services (% of Population) | Data." *Worldbank.org*, 2019, data.worldbank.org/indicator/SH.H2O.SMDW.ZS.

Population. (n.d.). Our World in Data.

https://ourworldindata.org/grapher/population

Karimova, F. (2021, May 10). Tajikistan's desperate thirst for clean drinking water. The Third Pole. https://www.thethirdpole.net/en/pollution/thirst-for-clean-water-continues-in-tajikistan/

YouthWiki. (n.d.). Retrieved from https://national-policies.eacea.ec.europa.eu/youthwiki/chapters/republic-of-north-macedonia/31-general-context