Assignment I

A Comprehensive Exploration Of Global Health and Demographic Dataset

Rhianne Gonsalves | Maria Hallak | Anna Parshikova | Irene SUNNY – *Team 'Tableau Titans'*

Part 1: Data Description

1.1 Motivation

The chosen topic of interest delves into the complex interplay between health expenditures, immunization rates, socioeconomic factors, and their collective influence on life expectancy across the globe. This subject is compelling and crucial due to its direct correlation with national development, individual well-being, and the overall health of populations. It offers critical insights into how different nations prioritize health and education within their socio-economic framework and the tangible outcomes of such investments. The topic is of particular relevance in the contemporary context, where global health disparities are stark, and the quest for enhancing life expectancy is more pressing than ever. In addition to its immediate implications for public health policy and practice, this research topic holds broader implications for global development agendas and international cooperation efforts. By addressing the root causes of health disparities and fostering collaboration between nations, we can work towards achieving more inclusive and sustainable development outcomes. Furthermore, this research can shed light on the importance of investing in preventative healthcare measures and social determinants of health, such as education and income inequality, which have long-lasting effects on population health and well-being. Ultimately, by examining the complex interconnections between health expenditures, immunization rates, and socioeconomic factors, we can contribute to a more comprehensive understanding of the factors shaping health outcomes worldwide and pave the way for more effective interventions to improve population health and life expectancy.

1.2 Description of the Dataset

1.2.1 Overview

The dataset under study is titled "Health and Demographics Dataset," a rich repository of global health and demographic data that stands out for its depth and breadth. It encapsulates a wide array of indicators from countries worldwide, offering a multi-year snapshot that enables the study of trends over time. This dataset is particularly relevant to our topic as it includes direct measures of health expenditures, educational attainment, disease prevalence, and immunization rates—each a critical piece in the puzzle of life expectancy.

The Dataset, spanning from 2000 to 2015, provides a critical historical perspective on global health trends, serving as a benchmark for evaluating the progress and long-term effects of health policies and interventions. Its historical breadth provides a solid baseline for comparison with current data, allowing for retrospective analyses and the development of predictive models. Despite not being the most current, it encompasses a crucial period that can help inform and shape future public health strategies and educational initiatives.

1.2.2 Source

The Health and Demographics dataset can be accessed and downloaded from Kaggle. Users will need to create a Kaggle account if they don't already have one, after which they can download the dataset directly from the dataset page.

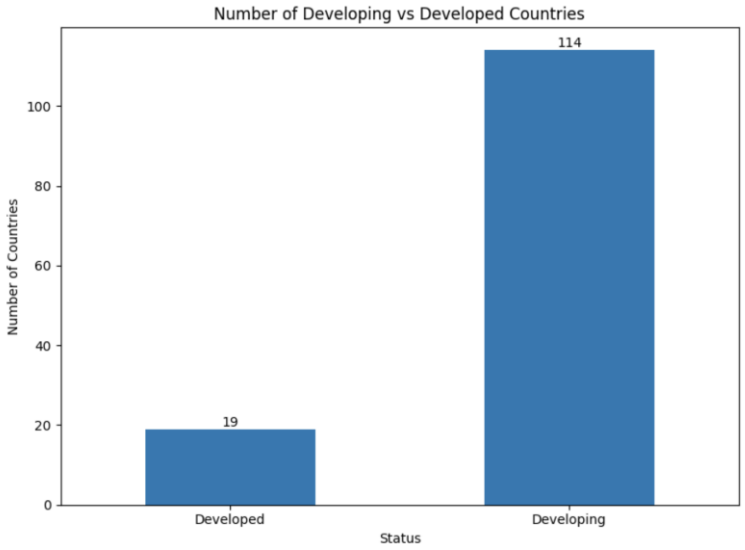
1.2.3 Data Collection

While the dataset compilation methods are not explicitly detailed, it is likely that the data was meticulously aggregated from various reputable international sources, such as the World Health Organization (WHO), The World Bank, and United Nations databases. These sources typically employ standardized data collection methods, including surveys, censuses, administrative records, and official national and international statistics, ensuring the reliability and accuracy of the information. The collation of such data into a single dataset enables a comprehensive analysis that

would otherwise require accessing multiple disparate sources, underscoring the dataset's value to researchers and policymakers.

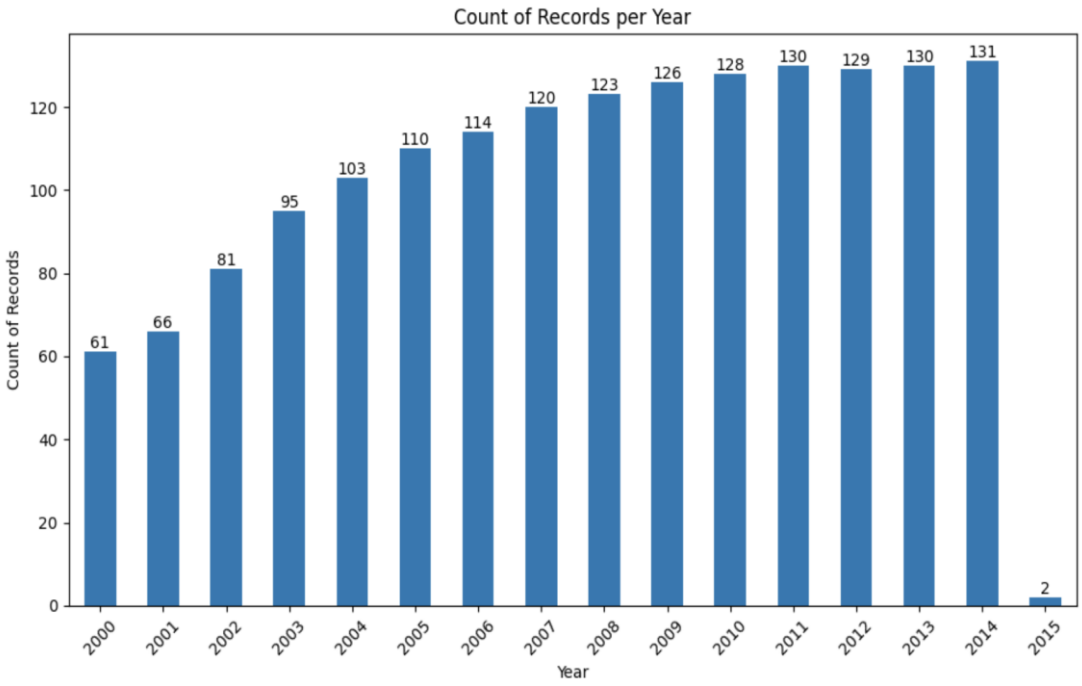
1.2.4 Attributes of The Dataset

The dataset captures the essence of multiple facets of public health and demographics, organized into 22 columns representing both numerical (19) and categorical variables (3) (Refer Appendix 4.1) and contains 1649 instances. Further analysis of the dataset showed us that there are no null values in the data. We observed that the split between developing countries and developed countries is 114 and 19 respectively (Refer Graph 1 as per below).



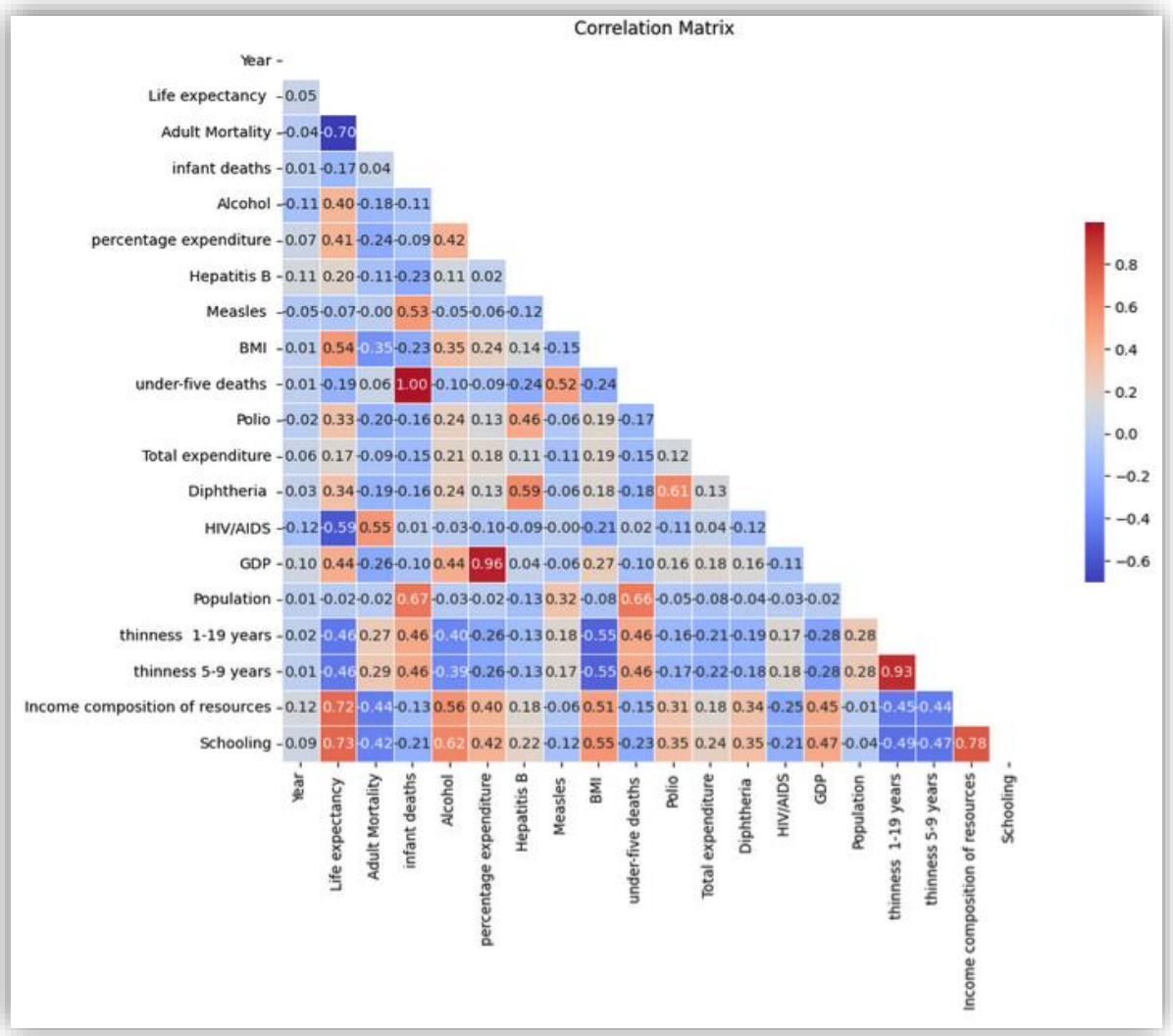
Graph 1 - Number of developing vs developed countries

The dataset demonstrates a consistent increase in the number of records for each year until 2014. This consistency is vital for longitudinal studies and analyses requiring reliable data availability across years to accurately evaluate trends and changes in life expectancy and associated health indicators. For the analysis we exclude data from 2015 as there were only 2 entries.



Graph 2 - Count of records per year

Correlation Matrix



Graph 3 - Correlation Matrix

Correlation matrix uncovered complex interplay between various socioeconomic factors and health outcomes.

1. Firstly, a robust positive correlation of 0.72 was observed between Life Expectancy and Income Composition of Resources. This indicates that countries with higher income composition of resources, serving as a proxy for socioeconomic status, tend to exhibit higher life expectancy rates.
2. Secondly, we found a similarly strong positive correlation of 0.73 between Life Expectancy and Schooling. This underscores the pivotal role of education in enhancing health and longevity, as higher levels of schooling correspond to increased life expectancy.
3. Interestingly, a positive correlation of 0.40 was noted between Life Expectancy and Alcohol Consumption. While this may seem counterintuitive at first glance, it could be influenced by other socio-economic factors such as wealth and healthcare access, particularly in regions where moderate alcohol consumption is prevalent.
4. Conversely, a striking negative correlation of - 0.59 was uncovered between HIV/AIDS prevalence and Life Expectancy. This starkly highlights the detrimental impact of the HIV/AIDS epidemic on public health, with higher prevalence rates associated with lower life expectancy.
5. Finally, a positive correlation of 0.44 was observed between GDP and Life Expectancy, indicating that countries with higher GDP per capita tend to experience greater life expectancy. This underscores the intrinsic link between a nation's economic prosperity and the health outcomes of its population.

These findings underscore the importance of considering socio-economic determinants when examining health outcomes and designing effective public health interventions. As we proceed with our visualizations for this project, these correlations will serve as crucial insights guiding our analysis and interpretation of the data.

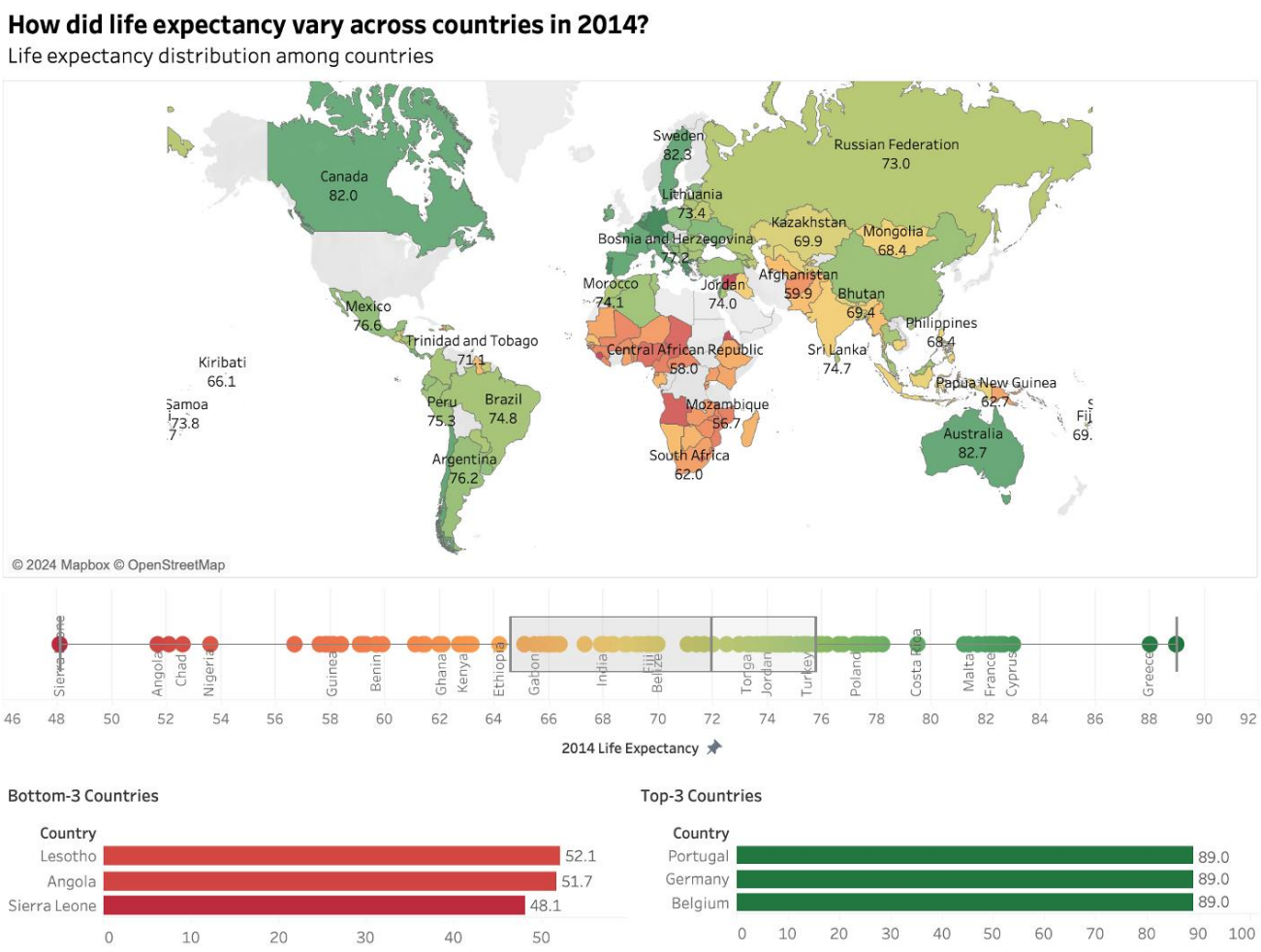
1.2.5 Outlier Detection and Data Transformation Methods

During our analysis, when calculating health indicators across a time period we only included countries that had data throughout the time period. As we observe that there are only 2 instances for the year 2015, we excluded that during the analysis.

Part 2: Exploratory Analysis by Members

2.1 Life expectancy variation across countries by Maria Hallak

Research Question: How did life expectancy vary across countries in 2014?



Information Takeaway

The visualization communicates the global distribution of life expectancy, showcasing the geographical and statistical spread of life expectancies among countries for the year 2014. The world map is the primary narrative tool, with a color-coded representation of life expectancy. It is supplemented by a horizontal box plot providing a statistical summary of the distribution and a display of the top and bottom countries to draw attention to the extremes.

Design Rationale and Visual Encoding

A continuous color scale of green to red is used as the color encoding with darker greens indicating higher life expectancy and darker reds indicating lower life expectancy. The green color is chosen because it is associated with a positive characteristic, here vitality and wellness, while the red color is associated with a warning or a concern, here the shorter life expectancies. The continuous scale is chosen because life expectancy is a quantitative continuous

variable with a natural ordering, and the continuous gradient smoothly communicates incremental differences without dividing the data into categories. This choice avoids the implication of distinct groupings, thus supporting the objective of viewing the global distribution rather than doing direct comparisons between individual countries.
It is important to note that the same color encoding is used across all the aspects of the visualization: the map, the boxplot and the ranking, ensuring visual coherence and facilitating any cross-reference.

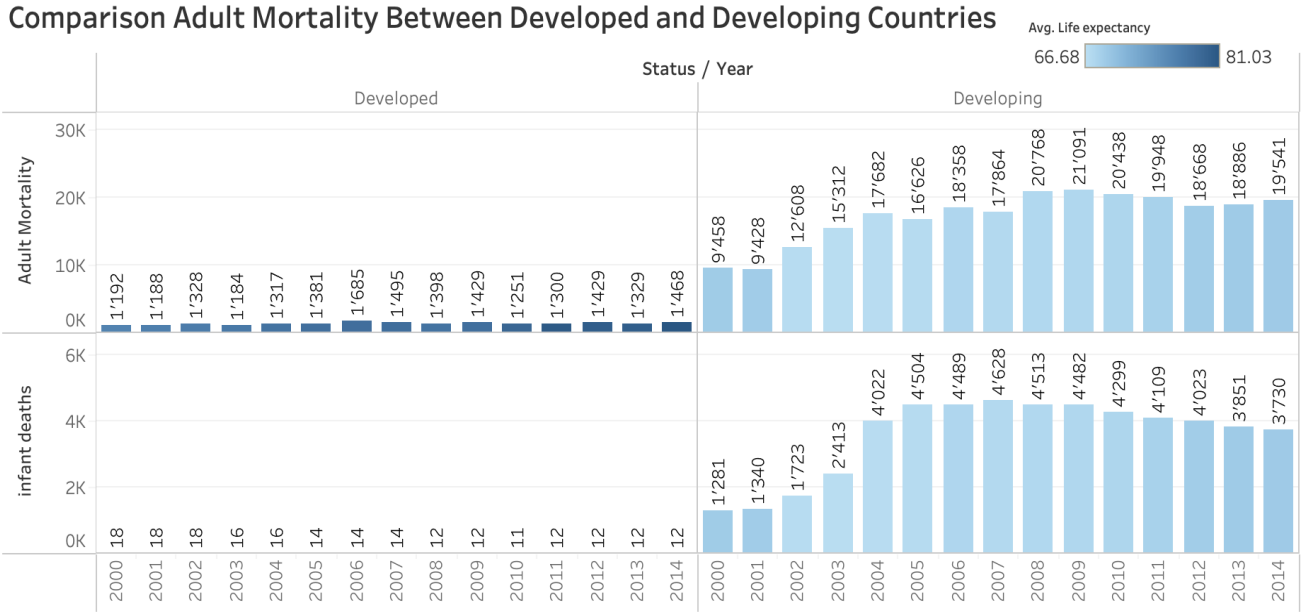
The use of the map seems intuitive to communicate the geographical distribution of life expectancies. It is appropriate because it utilizes the viewer's inherent geographical knowledge to facilitate immediate recognition of location and comparative size of countries, where the position, shape and area of each country on the map act as the visual channels. The geographical display also allows for spatial relationships and regional trends, such as the observation of lower life expectancies predominantly in African countries.

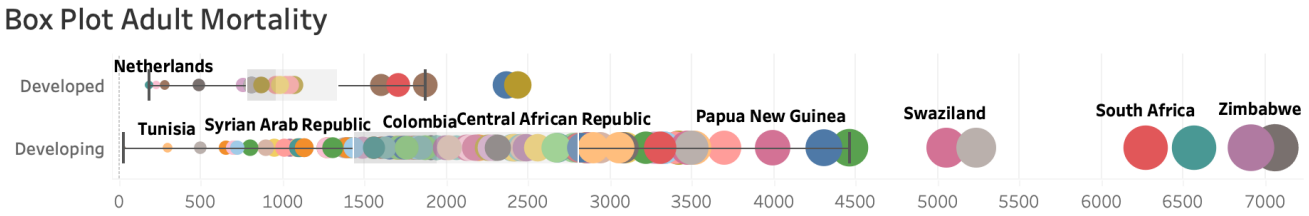
The box plot, with points as marks and positions as the visual channel, is used to show the statistical distribution of life expectancy across all the countries. The same color encoding serves as a visual channel too. The choice of a box plot is fit because it summarizes the data's distribution characteristics (central tendency and variability), allowing the viewers to understand the overall spread and to quickly locate the median, quartiles, and extremes.

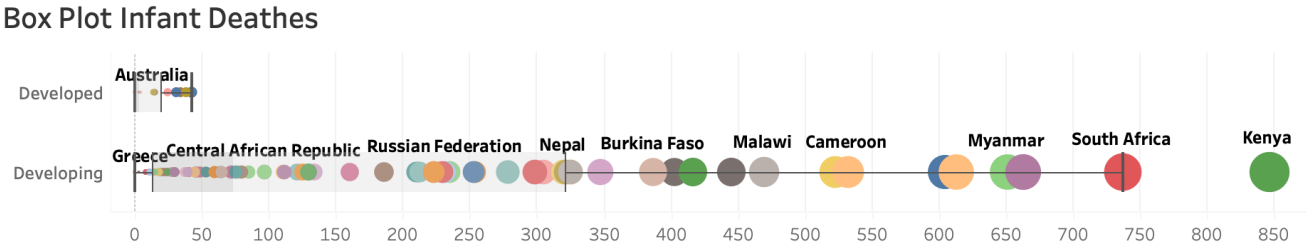
The bar chart for the top and bottom ranking countries in terms of life expectancy provide a focused view on the data's extremes, which may be of interest to the viewer. Position, length and color serve as the visual channels.

2.2 Analyzing the Trends in Adult and Infant Mortality Rates by Anna Parshikova

Research question: What is the trend difference associated with adult and infant mortality rates in countries with different statuses?







Description of the Figure

In this exploratory visualization, I aimed to examine the correlation between adult and infant mortality rates from 2000 to 2014 across different country statuses: developed and developing. The visualization clearly indicates that mortality rates in developing nations notably exceed those observed in developed counterparts for both adults and infants. Additionally, we observe wider variations in mortality rates for developing countries compared to the relatively stable rates in developed countries, suggesting minimal changes over the 14-year period (stagnation). Analysing the shading of bars representing life expectancy averages provides further insight, revealing a similar trend between developed and developing countries, with developing nations showing a shorter life expectancy with slight improvements over time. Conversely, developed countries exhibit higher life expectancy, suggesting a relationship between life expectancy and mortality rates. Further analysis using a box plot highlights differences in mortality rate fluctuations between infants and adults, as well as instances where certain developed countries show mortality rates

akin to those in developing nations. This underscores the need for personalized approaches to address health challenges, transcending simplistic country classifications and considering individual outliers.

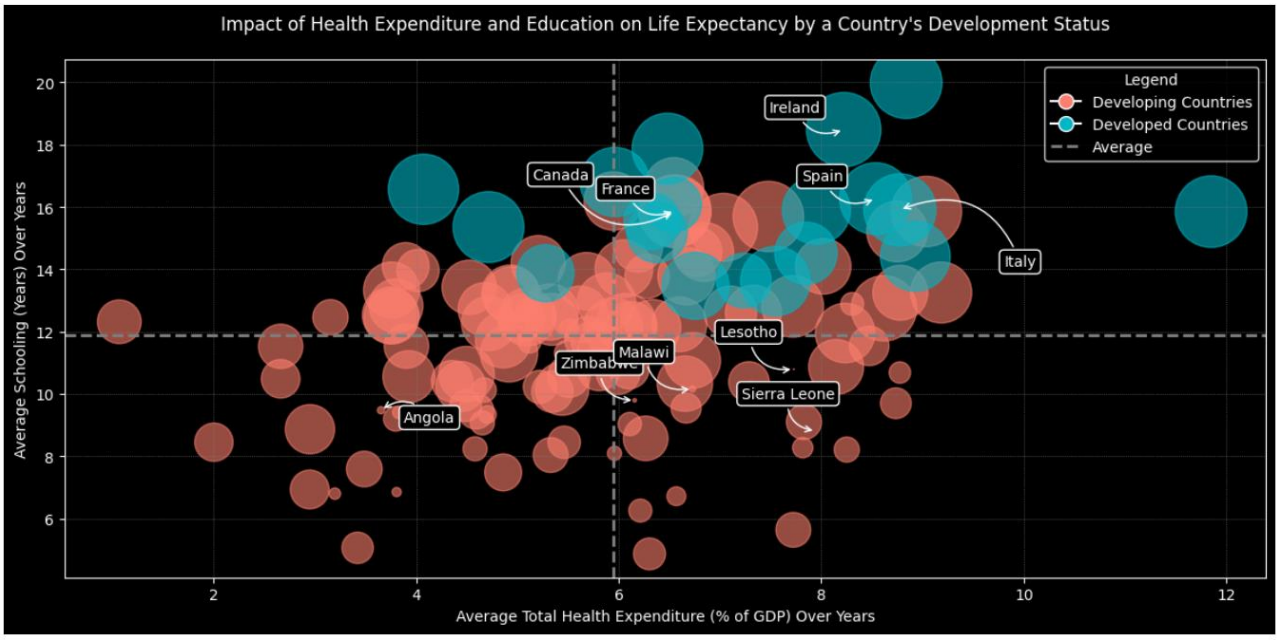
Design Rationale for the Visualization Design

To address my research question, I chose to visualize the data primarily using bar charts, color coding, and box plots. These selections were made based on their suitability for our data and specific questions, as they effectively convey both quantitative and qualitative information while facilitating comparisons and trend analysis. Beginning with bar charts, I depicted adult and infant mortality rates over time, along with country status, enabling us to observe the relationships between variables and trends, with the length of the bars directly representing the magnitude of mortality rates. To enhance insights, I made several modifications: I excluded data from the year 2015 due to incomplete records, potentially skewing trend analysis; introduced color coding to incorporate average life expectancy, providing a new perspective on the visualization and supporting the analysis of disparities. As my exploration progressed, I integrated box plots to illustrate the spread and distribution of mortality rates within both developing and developed countries, addressing insights missing from previous visualization methods. This encoding provided a comprehensive summary of the data, encompassing measures of central tendency, variability, and outliers, facilitating a nuanced understanding of mortality rate distribution and aiding in the identification of extreme values and trends. Finally, I adjusted color shades, bar chart widths, and labels, added annotations and legends, and renamed columns to enhance clarity and facilitate the usage of the visualization.

2.3 Average Life Expectancy vs Social Indicators by Irene Sunny

Research Question: How does the combination of health expenditure and education levels impact life expectancy across different development statuses?

Through the below visualisation, I attempt to present a compelling visual narrative of how a country's average total health expenditure (% of GDP) and average schooling years, which are key determinants of a country's developmental status, correlate with average life expectancy. To ensure easiness, it is the average of the three metrics, over the years, for each country is plotted instead of plotting the data against all countries for all years.



Takeaway

We can infer that, countries with higher investments in health and education (in the top right quadrant) tend to have a higher life expectancy, suggesting that these expenditures are not merely costs but critical investments in human capital that can yield substantial returns in the form of a healthier, longer-living population.

An important takeaway is that developed countries and the developing countries with better life expectancy than the rest predominantly have above-average schooling years, suggesting a strong emphasis on education within these nations. However, they are not uniformly above the average for health expenditure, which reveals that a high standard of education does not necessarily correlate with proportional investment in health. This observation may suggest that developed countries have optimized their health spending or achieved efficiencies in their healthcare systems that allow them to spend less while still maintaining high life expectancies.

The value addition of this visualization lies in its capacity to distil complex, multi-dimensional data into a coherent story that can be immediately grasped by policymakers, researchers, and the public. It provides empirical backing to the argument that sustainable development is multifaceted, requiring balanced growth in both the health and education sectors.
Note: Sierra Leone's life expectancy is so small that the bubble is almost invisible and seems hidden

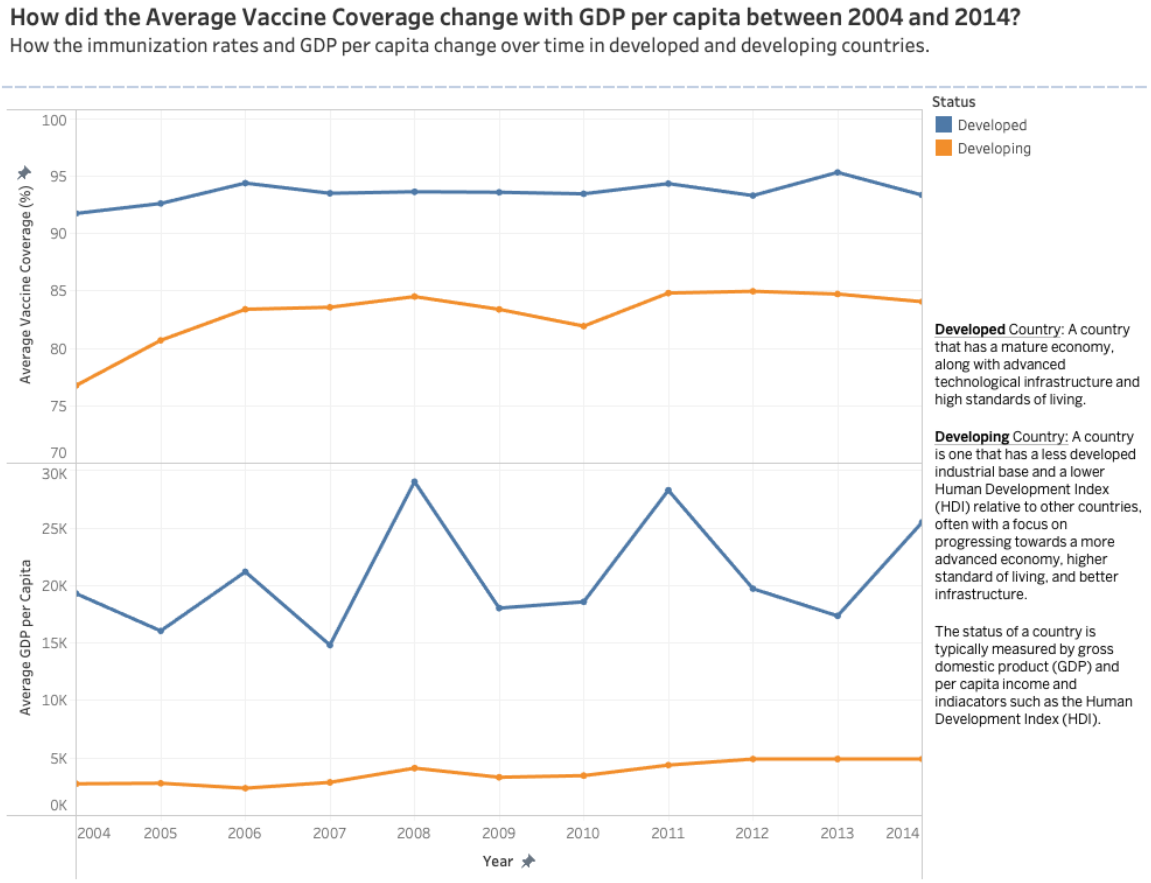
Design Rationale and Visual Encoding
By plotting these variables in a bubble chart where each bubble represents a country, the visualization highlights the intricate balance between economic investment in health and the educational status of a nation as they collectively impact the longevity of its people.
The visual encodings used include color, size, and position on two axes.

- To ensure inclusivity and accessibility for those with colour vision deficiencies, the colours not on the red-green spectrum, were chosen. ‘Cyan’ for developed countries and ‘Dark Salmon’ for developing countries provide an immediate visual cue to the viewer, adhering to colour theory principles that suggest these hues have sufficient contrast to be distinguishable yet are subdued enough not to overwhelm.
- The size of the bubbles corresponds to the life expectancy, allowing for an immediate visual grasp of how countries fare against each other within the depicted dimensions.
- The inclusion of average lines for health expenditure and schooling offers a reference point against which to assess individual country performance, further enriching the analysis.

A deliberate design choice to avoid cluttering was made by annotating only the top 5 and bottom 5 countries by life expectancy. It draws attention to those significantly over or underperforming relative to their development status.

2.4 Effect of change in GDP on average immunization coverage from 2004 to 2014 by Rhianne Gonsalves

Research question: How did the Average Vaccine Coverage change with GDP per capita between 2004 and 2014?



Description of the Figure

The chart is a line graph that compares the average vaccine coverage and average GDP per capita between developed and developing countries from 2004 to 2014. There are two y-axes: the left y-axis scales the average vaccine coverage percentage, and the right y-axis scales the average GDP per capita. The x-axis represents the years from 2004 to 2014.

This graph is useful because it shows us how small changes in the average GDP of a developing country has a lagged effect on the average vaccine coverage in the following year, while changes in the GDP of developed countries do not affect the average vaccine coverage in developed countries.

Two lines are plotted on the graph: the blue line represents the average vaccine coverage in developed countries, which appears to be consistently high with slight fluctuations. The orange line indicates the average vaccine coverage in developing countries, showing higher variability and generally lower coverage compared to the developing countries' line.

I used a text box to provide definitions on how a country is defined as 'Developed' and 'Developing'.
The title of the chart explains the question about the relationship between average vaccine coverage and GDP per capita over time in developed and developing countries.

Design Rationale for the Visualization Design

- **Color Differentiation**: The two lines, each representing a different group of countries, developed (orange) and developing (blue) are distinct enough to differentiate between the two groups easily.
- **Clear Axes**: Both the left and right y-axes are clearly labelled with their respective metrics, "Average Vaccine Coverage %" and "Avg GDP per capita," which allows for an immediate understanding of what each axis represents.
- **Time Series**: The x-axis is used for the time series, running from 2004 to 2014, and is marked at consistent intervals to show the passage of time clearly.
- **Gridlines**: Horizontal gridlines are used to aid in tracking the data points across the chart.
- **Comparative Layout:** The chart effectively uses a dual-axis layout to compare two different metrics: average vaccine coverage and GDP per capita.
- **Informational Text Boxes**: The chart includes a text box that defines the criterion used to set the status of a country as 'Developed Country' or 'Developing Country', This adds context to the data and helps viewers understand the criteria for classification.

Part 3: Findings

The comprehensive exploration of the dataset, through the use of visualizations, reveals insights into the complex dynamics between healthcare factors (expenditures and immunization rates), socioeconomic factors, and their overall influence on life expectancy across the globe.

- All the produced visualizations underscore the evident disparities in life expectancy and health outcomes between developed and developing nations. Developed countries generally exhibit higher life expectancies, lower mortality rates for both adults and infants, attributed by better health indicators and healthcare policies. The visualizations not only quantify these disparities but also contextualize them, making the data comprehensible.
- The strong positive association between life expectancy and factors such as national expenditures in healthcare and education highlights the significance of socioeconomic factors in shaping health outcomes. More importantly, the analysis emphasizes the pivotal role of governmental policies in improving health outcomes and thus contributing to overall national development.
- Examining both adult and infant mortality rates, a pronounced contrast is revealed between developed and developing countries. This highlights the critical challenges faced by developing nations and the urgent need for targeted health interventions and policies.
- Focusing on immunization rates in relation to GDP, it can be seen that developing nations often achieve relatively high immunization rates despite significantly lower economic performance. This suggests a prioritized allocation of funds on vaccinations in these countries most possibly as a preventive measure.

The importance of exploratory analysis through the use of visual tools cannot be overstressed. These visualizations go beyond just providing a numerical analysis to providing a narrative that contextualizes the data. They allow the recognition of patterns, trends, and outliers that might not be apparent through traditional statistical methods, facilitating a more intuitive understanding, analysis and synthesis of information. The visualizations play an important role in the synthesis of actionable decisions for many readers.

4. Appendix

4.1 Data Description

Column Name	Description	Data Type	Mean
Country	Name of the country (133 countries)	object	-
Year	Year the data was recorded	int64	2007.84
Status	Development status ('Developed' or 'Developing')	object	-
Life Expectancy	Average number of years a person is expected to live	float64	69.30
Adult Mortality	Adult mortality rates (probability of dying between 15 and 60 years per 1000 population)	int64	168.22
Infant Deaths	Number of Infant Deaths per 1000 population	int64	-
Alcohol	Alcohol consumption measured as litres per capita	float64	4.53
Percentage Expenditure	Expenditure on health as a percentage of GDP	float64	698.97
Hepatitis B	Hepatitis B immunization coverage	int64	80.94
Measles	Number of reported Measles cases per 1000 population	int64	-
BMI	Average Body Mass Index	float64	38.13
Under-Five Deaths	Number of under-five deaths per 1000 population	int64	-
Polio	Polio immunization coverage percentage	int64	82.55
Total Expenditure	General government expenditure on health as a percentage of total government expenditure	float64	5.96
Diphtheria	DTP3 immunization coverage percentage	int64	82.32
HIV/AIDS	Deaths per 1000 live births HIV/AIDS (0-4 years)	float64	1.98
GDP	Gross Domestic Product per capita	float64	5566.03
Population	Population of the country	float64	1.465e+07
Thinness 1-19 years	Prevalence of thinness among children and adolescents for 1-19 years	float64	4.85
Thinness 5-9 years	Prevalence of thinness among children for 5-9 years	float64	4.90
Income Composition of Resources	Human Development Index in terms of income composition of resources	float64	0.63
Schooling	Number of years of Schooling	float64	12.12