Global Life Expectancy Analysis Project Report

Caleb Damron
*Department of Electrical Engineering
and Computer Science*
The University of Tennessee,
Knoxville Tennessee
cdamron2@vols.utk.edu

Clayton Tucker
*Department of Electrical Engineering
and Computer Science*
The University of Tennessee,
Knoxville Tennessee
ctucke24@vols.utk.edu

Cole Price
*Department of Electrical Engineering
and Computer Science*
The University of Tennessee,
Knoxville Tennessee
yhb368@vols.utk.edu

Danyil Chuprynov
*Department of Electrical Engineering
and Computer Science*
The University of Tennessee,
Knoxville Tennessee
dchupryn@vols.utk.edu

Dhruv Patel
*Department of Electrical Engineering
and Computer Science*
The University of Tennessee,
Knoxville Tennessee
dpate122@vols.utk.edu

Abstract— **This project investigated how social, economic, and environmental factors related to life expectancy across 193 countries. We combined multiple pre-2020 publicly available datasets into a single standardized master table containing each country's life expectancy, GDP, access to safe drinking water, health expenditure, air quality, happiness score, freedom score, region, and income category. Using descriptive statistics and correlation analysis, we found that access to clean water and happiness scores had the strongest positive correlations with life expectancy (both around 0.69), followed by political freedom. GDP and health expenditure as a share of GDP showed weaker positive relationships, while worse air quality was modestly associated with lower life expectancy. Clear gaps appeared between income groups and regions, and several countries stood out as outliers; either achieving relatively high life expectancy at moderate income levels or underperforming despite high-income status. This work demonstrated how integrating heterogeneous public datasets could reveal global health patterns and provided a foundation for interactive visual exploration through a sortable leaderboard and graphs.**

I. Objective

The objective of this project was to identify and visualize broad patterns in how social, economic, and environmental indicators were associated with national life expectancy. Specifically, we asked: Which factors, among income, GDP, healthcare spending, happiness, air quality, political freedom, and access to safe water, were most strongly associated with countries' life expectancy prior to the COVID-19 pandemic?

We focused on pre-2020 data. We did this to avoid any distortions caused by the COVID-19 pandemic and to maximize country coverage, since many nations did not yet have reliable statistics for the most recent years. In practice, we used indicators from roughly a five-year window before 2020 and treated them as a baseline "normal" period.

II. Data

A. Data Sources and Collection

All indicators came from publicly available sources and were chosen to predate 2020 so that they reflected a pre-pandemic baseline. Below are the cleaned file names in the format: Factor (Source):

- Life Expectancy (World Bank).csv
- Income Group and Region (World Bank).csv
- GDP (World Bank).csv
- Safe Drinking Water Access (World Bank).csv
- Health Expenditure % GDP (World Bank).csv
- Happiness Index (Kaggle).csv
- Freedom Index (World Population Review).csv
- Air Quality (Wikipedia).csv

B. Data Processing and Integration

We began with eight separate CSV files and wrote a script to load, clean, and merge them into a single master dataset. The merge key was country name, so standardizing names was a major task. We created a manual mapping to unify naming variations such as "Republic of Korea" vs. "South Korea" and "Russian Federation" vs. "Russia."

We merged all datasets onto the life expectancy table. Countries missing specific values were either filled using alternative reputable sources (e.g., WHO, additional World Bank series) or left blank. When left blank, those countries were excluded only from analyses requiring that indicator.

The final master dataset contained 193 countries and the following columns:
- Country
- Life Expectancy (2019)
- GDP (2019)
- Clean Water Access % (2019)

- Health Expenditure % GDP (2019)
- Air Quality Score
- Happiness Score (2015)
- Freedom Score (0-100)
- Region
- Income Category

We validated the dataset via spot-checking countries against original sources, verifying unique country entries, and ensuring that numerical values fell within expected ranges.

III. MODELS AND ANALYSIS METHODS

Our analysis consisted of four main components. First, we computed descriptive statistics for all indicators. Second, we performed Pearson correlation analysis between life expectancy and each variable. Third, we grouped countries by income category and region to compare average life expectancy across structural classifications. Finally, we identified outliers whose life expectancy differed substantially from what their socioeconomic characteristics would normally predict.

In addition to these statistical analyses, we implemented an interactive visualization interface to make the results easier to explore. The interface was a locally hosted web page built with HTML, CSS, and JavaScript, and it loaded the cleaned master dataset directly from a CSV file. We used Chart.js to create all of the charts on the site, which allowed us to render scatter plots, summary charts, and other visualizations directly in the browser without needing a backend server.

The main page of the website featured a sortable table listing all 193 countries along with their life expectancy, GDP, clean water access, health expenditure, happiness score, air quality score, freedom score, region, and income category. Users could click on any column header to sort the table by that metric, making it easy to identify, for example, the highest life expectancy countries, the lowest clean water access countries, or countries with high happiness scores relative to income.

We also included a set of interactive scatter plots with life expectancy on the y-axis and a selected factor on the x-axis. Users could switch between variables such as GDP, clean water access, health expenditure, air quality score, happiness score, and freedom score. These charts made it possible to visually confirm the trends we observed in the correlation analysis, as well as spot clusters and outliers. We also generated summary visualizations such as a correlation heatmap and plots grouped by income category, which provided a compact view of how all variables related to one another and how life expectancy differed across income levels. Together, these visual components turned the static master dataset into an exploratory tool that complemented and reinforced the quantitative results presented in this report.

Figure 1. Interactive table showing all countries and factors.

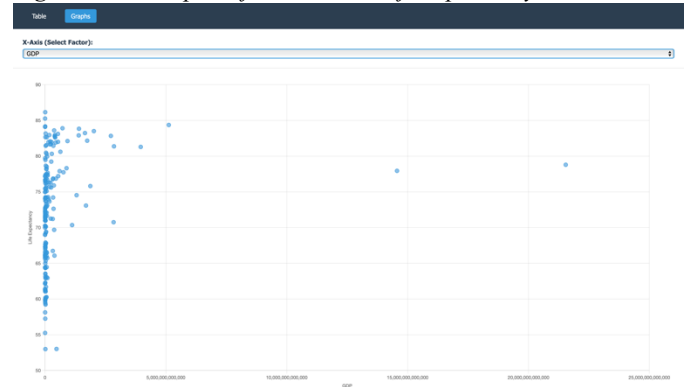Figure 2. Scatterplot of GDP versus life expectancy.

Figure 3. Scatterplot of clean water access versus life expectancy.
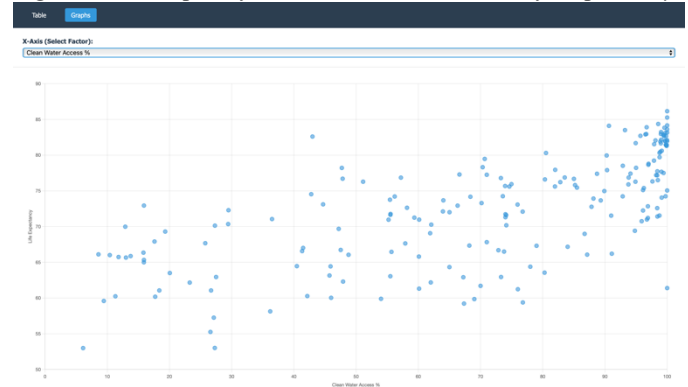

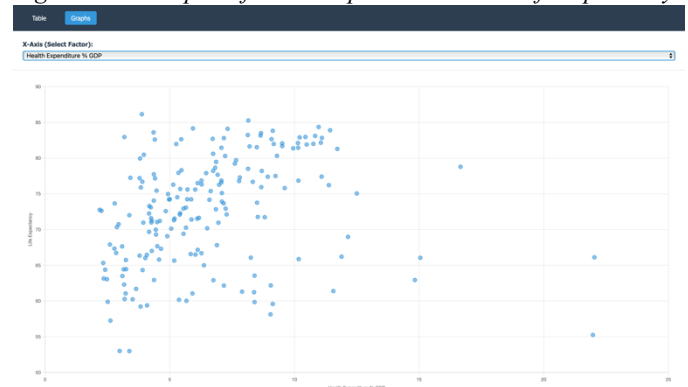Figure 4. Scatterplot of health expenditure versus life expectancy.

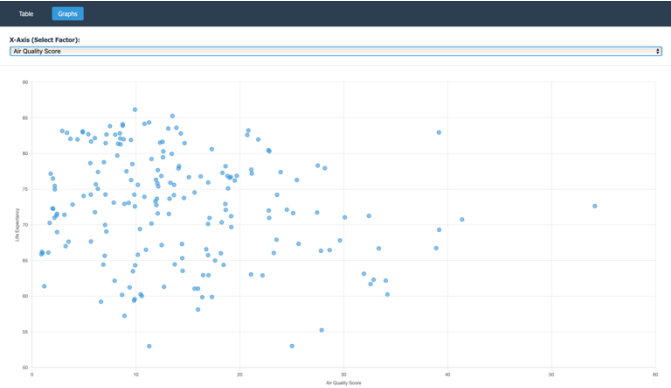*Figure 5. Scatterplot of air quality score versus life expectancy.*

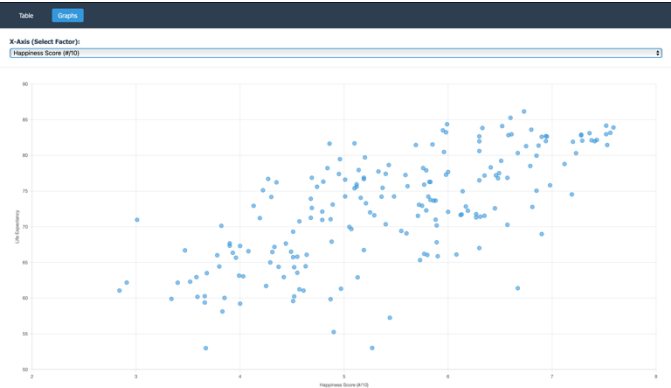*Figure 6. Scatterplot of happiness score versus life expectancy.*

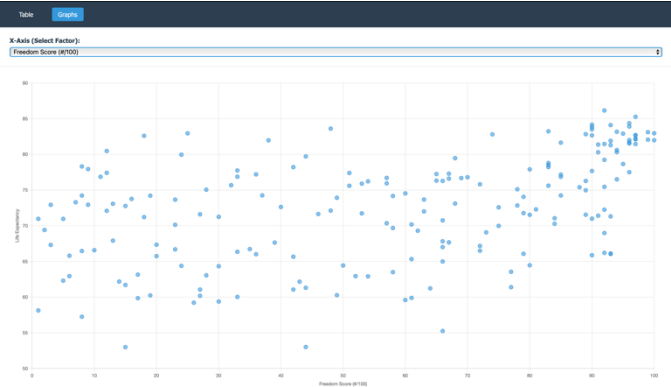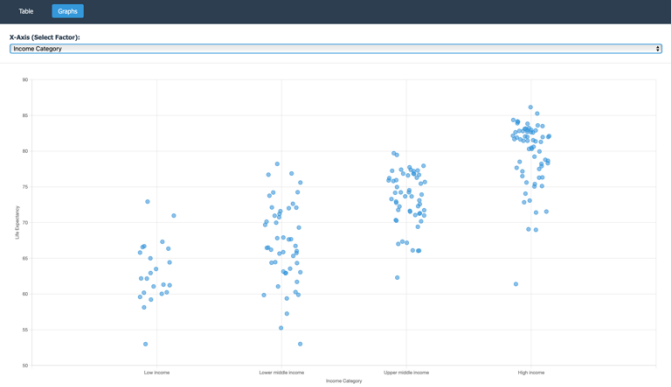*Figure 7. Scatterplot of freedom score versus life expectancy.*

Figure 8. Scatterplot of income levels versus life expectancy.

IV. RESULTS

A. Overall Life Expectancy Patterns

Across all 193 countries in the dataset, the average life expectancy was approximately 72.4 years, with values ranging from roughly 31.5 years at the low end to about 86.2 years at the high end. These extremes highlighted the substantial variation in global health outcomes. When countries were grouped by the World Bank's income categories, clear differences emerged. High-income countries averaged around 79.6 years, whereas upper-middle-income nations averaged about 73.3 years. Lower-middle-income countries averaged roughly 67.0 years, and low-income countries were the lowest at approximately 61.8 years. Although these averages formed a clear upward trend with rising income levels, significant variation still existed within each category, showing that income alone could not fully explain global differences in longevity.

A similar pattern appeared when examining life expectancy by region. Regions such as North America and Europe & Central Asia had the highest average life expectancies, approximately 80.5 and 78.7 years respectively. In contrast, Sub-Saharan Africa had the lowest regional average, roughly 62.5 years, reflecting persistent challenges in infrastructure, healthcare access, and environmental conditions. Other regions, including Latin America, South Asia, and East Asia & Pacific, generally fell in the low-to-mid seventies. These findings indicated that geography and structural regional characteristics also played important roles in shaping health outcomes, although, like income, they did not fully determine a country's position.

B. Correlations with Life Expectancy

Correlation analysis provided further insight into how specific economic, social, and environmental indicators were associated with life expectancy. Access to safe drinking water and the national happiness score both showed strong positive correlations with life expectancy, each around +0.69. These relationships suggested that basic infrastructure and broader measures of societal well-being were closely tied to population health. Political freedom was moderately correlated with life expectancy, with a coefficient of approximately +0.48, indicating that civil liberties and governance quality tended to coincide with better health outcomes.

Economic indicators displayed weaker relationships. Total GDP had a correlation of roughly +0.18, while health expenditure as a percentage of GDP correlated at about +0.20. These findings implied that simply having a larger economy or spending more money on healthcare did not necessarily translate directly into longer lives. Instead, how resources were allocated, and whether they translated into effective services, appeared to matter more. Environmental quality also played a

role. Air quality, defined so that higher scores represented worse conditions, had a correlation of about -0.20 with life expectancy. Although this association was modest, it aligned with expectations: countries with more polluted air tended to experience shorter average lifespans.

Taken together, the correlation results indicated that variables tied to living conditions, infrastructure, and societal well-being were more strongly linked to life expectancy than pure economic inputs. Countries with clean water access, higher happiness levels, and greater freedom generally had longer-lived populations, while the effects of GDP and health spending were present but relatively weak.

C. Outlier Countries

While general patterns held across most countries, several stood out as outliers whose life expectancy levels differed markedly from what their economic classifications might predict. A few lower-middle-income countries, including Jordan, Lebanon, Sri Lanka, and Tunisia, achieved life expectancies in the mid-to-upper seventies, rivaling or surpassing those of some high-income nations. Their performance suggested that targeted public health strategies, strong primary care systems, or favorable cultural and social factors could compensate for lower national income levels.

Conversely, some high-income countries exhibited unexpectedly low life expectancies. Examples included nations such as Nauru, Guyana, the Bahamas, Trinidad and Tobago, Seychelles, and Russia, all of which fell noticeably below the average high-income life expectancy of roughly 79.6 years. In these cases, factors other than income, such as inequality, violence, epidemiological burdens, or environmental risks, may have contributed to the comparatively lower longevity. These deviations from expected patterns highlight the importance of examining multiple indicators rather than relying solely on GDP or income classifications when assessing global health outcomes.

V. Primary Issues Encountered

The project faced several data-related challenges. Many datasets used inconsistent country naming conventions, requiring extensive manual standardization before merging. Discrepancies such as "Republic of Korea" versus "South Korea" prevented automated alignment and had to be corrected by hand. We also encountered missing values across multiple indicators. In some cases, reliable alternative sources such as WHO databases or additional World Bank series allowed us to fill in gaps; in other cases, values had to remain missing, and those countries were excluded only from analyses involving the absent variable. Another limitation came from mismatched reporting years among datasets. Happiness scores were taken from 2015, while most economic and health indicators were

from 2019. We resolved this issue by treating the entire dataset as pre-2020 baseline data, acknowledging that this approach sacrificed year-to-year precision but still allowed for a meaningful cross-sectional comparison. We also encountered minor implementation-related issues during visualization development, such as configuring chart options and aligning layouts, which we resolved through iterative testing.

VI. Future Work

Although our analysis provided a clear snapshot of major factors related to global life expectancy, several natural extensions remain. Future work could incorporate GDP per capita and measures of economic inequality, which would better capture how economic resources translate into individual well-being. Time-series analysis would also be valuable, especially for comparing pre- and post-COVID-19 trends to understand how different countries' underlying conditions influenced their resilience during the pandemic. More detailed environmental and health-related variables, such as PM2.5 concentrations, obesity rates, tobacco use, and chronic disease prevalence, could help explain cases where countries deviated significantly from expected patterns. Finally, predictive modeling using multiple regression or machine learning approaches could be used to estimate life expectancy based on combinations of factors and identify which variables remain significant when controlling others. These analyses could be integrated into an expanded interactive dashboard with richer filtering tools and more sophisticated visual analytics.

VII. Organization And Timeline

Throughout the project, all team members contributed to the initial data collection phase by identifying and downloading datasets from the World Bank, Kaggle, and other public sources. Dhruv Patel and Caleb Damron played the primary role in data cleaning and integration, which included merging datasets, standardizing country names, resolving naming inconsistencies, and handling missing values. Statistical analysis was carried out mainly by Dhruv Patel, Clayton Tucker, and Cole Price, who computed descriptive statistics, generated correlation results, and interpreted the analytical findings. Visualization development was led by Caleb Damron, Danyil Chuprynov, and Cole Price, who implemented the interactive leaderboard, scatter plots, and other graphical elements. The front-end structure and user interface were designed and built by Danyil Chuprynov, Dhruv Patel, and Cole Price. Project documentation and report drafting were handled collaboratively by all members. All members also participated in testing, debugging, and refining the website and final presentation materials.

The project timeline closely followed the schedule established in the original proposal with only minor and negligible deviations. During the first two weeks, the team identified and downloaded the raw datasets needed for the

analysis. By mid-October, the data cleaning and integration stage was completed, including standardization of country names and consolidation of all indicators into the master dataset. Data validation and initial quality checks were finalized shortly thereafter. Statistical analysis and preliminary findings were completed by late October. Development of the interactive leaderboard began at the end of October, followed by the implementation of dynamic scatter plots and additional visualizations in early November. The website frontend was designed and integrated after the visual components were completed. The remaining project period was dedicated to integration testing, debugging, report writing, and preparation of the final presentation.

VIII. CONCLUSION

This project used multiple pre-2020 public datasets to examine how economic, social, and environmental indicators related to life expectancy across 193 countries. After consolidating eight individual datasets into a single standardized master table, we identified clear disparities by region and income group and found strong positive associations between life expectancy and factors such as clean water access, happiness, and political freedom. GDP and health expenditure exhibited only modest relationships with life expectancy, while air quality showed a negative association. We also identified several outlier countries whose life expectancies significantly exceeded or fell below expected levels based on income alone. The project provided hands-on experience with data cleaning, integration, and exploratory analysis and created a foundation for future extensions involving predictive modeling, additional indicators, and a more advanced interactive dashboard.