Socioeconomic Determinants of Life Expectancy: A Global Analysis of Economic, Health, and Educational Factors

Ron Nitzan[1,†] **and Zelzer Adam**[2,†]

[1]*I.D. 215451709*
[2]*I.D. 328489166*
[†]*These authors contributed equally to this work*

September 30, 2024

Abstract

Life expectancy is a critical indicator of a population's health and well-being, influenced by various socioeconomic and lifestyle factors. This study explores the relationships between life expectancy and several key variables, including population size, GDP per capita, BMI, alcohol consumption, and schooling levels across countries and continents over time. Using a dataset that spans multiple years and diverse regions, we analyze how economic development, health-related behaviors, and access to education impact life expectancy. Our results reveal strong correlations between higher GDP per capita, increased years of schooling, and greater life expectancy. Surprisingly, higher BMI and alcohol consumption also show positive associations. The analysis provides insights into how different socioeconomic factors contribute to health outcomes globally, highlighting the complex interplay between wealth, education, and health behaviors. These findings offer valuable implications for public health policies aimed at improving life expectancy, especially in developing regions, where targeted interventions could mitigate adverse health impacts and foster healthier, longer lives.

Contents

1 Introduction 1

2 Results 1

2.1 Correlations . . . 1
2.2 Statistical Tests . . . 2
GDP • BMI • Alcohol Consumption • Schooling
2.3 Regression Analysis . . . 3

3 Methods 3

3.1 Transformations . . . 3
Log Transformation • One-hot Encoding
3.2 Correlations . . . 3
3.3 Statistical Tests . . . 4
Normality Tests • Variance Homogeneity Test: Levene's Test • Independent Two-Sample t-Test • Independent Mann-Whitney U Rank Test • Bonferroni Correction
3.4 Regression Analysis . . . 6

4 Codes 6

4.1 Imports . . . 6
4.2 Normality Tests . . . 6
Shapiro-Wilk Test • Kolmogorov-Smirnov Test
4.3 Variance Homogeneity Test: Levene's Test . . . 7
4.4 Independent Two-Sample T-Test . . . 7
4.5 Independent Mann-Whitney U Rank Test . . . 7

5 Discussion 7

5.1 Conclusions . . . 7
5.2 Limitations of Analysis . . . 7

1. Introduction

Life expectancy is one of the most important measures of a population's overall health and quality of life. It reflects not only the state of healthcare systems but also the broader socioeconomic conditions that affect individuals' well-being. While medical advances have significantly extended life spans in many parts of the world, disparities in life expectancy between countries and regions remain stark. These variations are often driven by a range of socioeconomic factors, including income levels, access to education, and lifestyle behaviors.

Research has shown that wealthier nations tend to have higher life expectancies, largely due to better access to healthcare, healthier diets, and improved living conditions. However, economic development alone does not guarantee longer lives. Other factors, such as education, which can influence health literacy and lifestyle choices, also play a critical role. Meanwhile, health-related behaviors like alcohol consumption and body mass index (BMI) introduce additional layers of complexity in understanding how socioeconomic factors affect life expectancy.

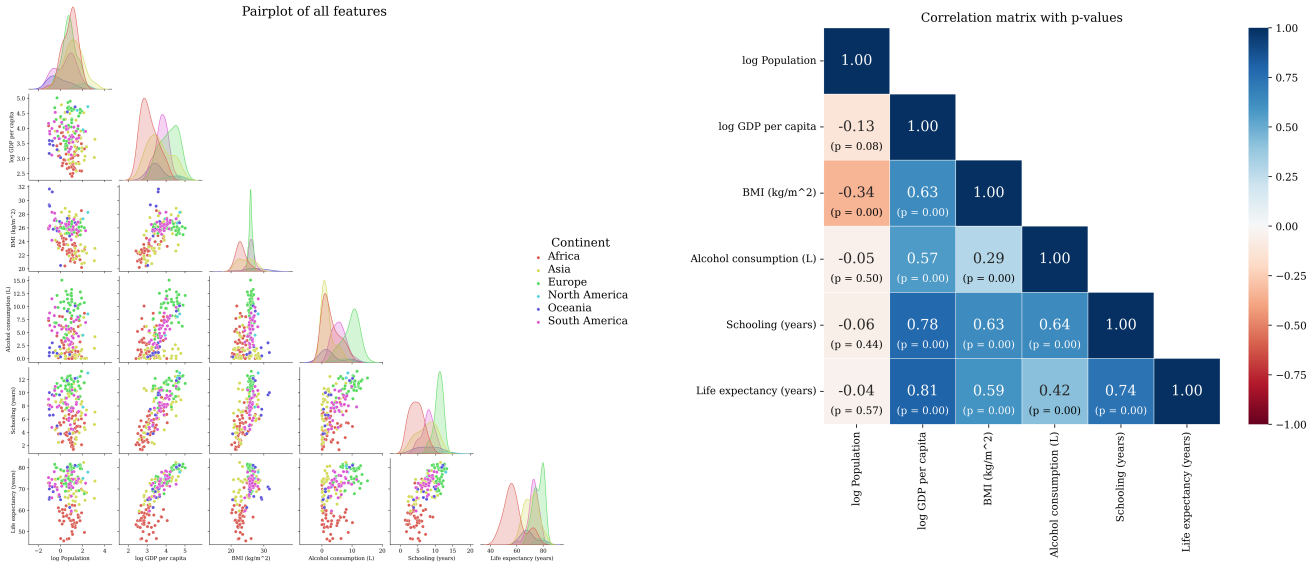
This study aims to investigate the relationship between life expectancy and several key socioeconomic variables, including GDP per capita, BMI, alcohol consumption, schooling, and population size. By analyzing data across multiple countries and continents over time, we seek to identify which factors are most strongly associated with longer or shorter life spans and how these relationships vary globally. The findings will offer insights into the complex interplay between economic conditions, educational attainment, and health behaviors, contributing to a deeper understanding of the determinants of life expectancy and informing public health policies aimed at addressing disparities.

2. Results

2.1. Correlations

To explore the relationships between life expectancy and the socioeconomic factors in our dataset, we calculated a correlation matrix to quantify the strength and direction of associations between the variables. The matrix (1b) reveals several notable correlations. **log GDP per capita** and **schooling years** exhibit strong positive correlations with life expectancy, suggesting that wealthier countries with higher education levels tend to have longer lifespans. In addition, **BMI** and **alcohol consumption** also show positive correlations with life expectancy, albeit weaker. Conversely, population size doesn't seem to relate to life expectancy.

The correlation matrix also highlights the interconnections between other factors; for example, log GDP per capita is positively correlated with schooling, reflecting how economic development often coincides with better access to education. These findings suggest that economic and educational improvements are key drivers of increased life expectancy, while lifestyle-related factors like BMI and alcohol consumption introduce nuanced and region-dependent effects on health outcomes.



(a) Pairplot of features

(b) Correlation matrix with p-values

Figure 1. Correlation between features in the dataset.

2.2. Statistical Tests

In this section, we show the results of statistical tests performed on the distribution of life expectancy in relation to different features across 4 years: 2000, 2005, 2010, and 2015. Each feature was divided into several categories, and then we tested for statistical differences between the means and medians of life expectancy distributions. All results can be seen in table 1.

2.2.1. GDP

The results across all years (2) can suggest that nations with higher **GDP per capita** tend to have longer life expectancies in all tested years. The independent two-sample t-test and Mann-Whitney U test confirmed our hypothesis that higher **GDP per capita** years is associated with longer life expectancy. This is not surprising, as GDP per capita is a great measure for economic growth, which influences public health greatly.

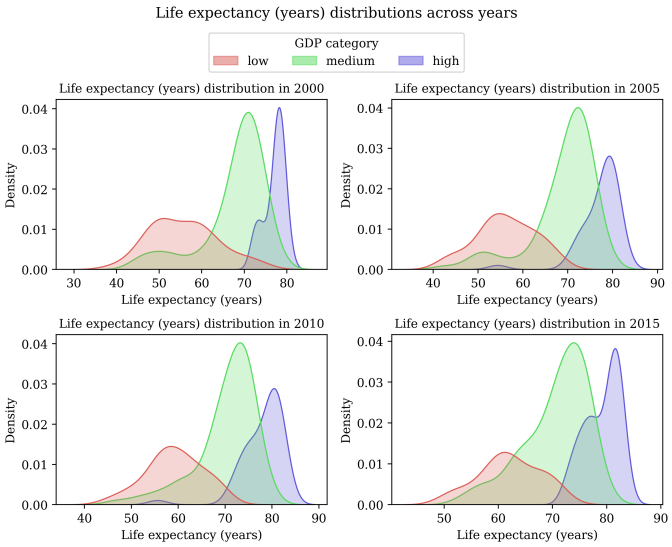


Figure 2. Life expectancy distribution in different GDP categories.

2.2.2. BMI

The results across all years (3) suggest that individuals with a "high" BMI tend to live longer than those with a "low" BMI. The Mann-Whitney U test confirmed significant differences in median **Life expectancy** for all tested years. Therefore, it can be concluded that higher BMI is associated with longer **Life expectancy** across all years. This is quite surprising, as high BMI (above 25) is related to obesity, and yet nations with higher BMI tend to have longer life expectancies.

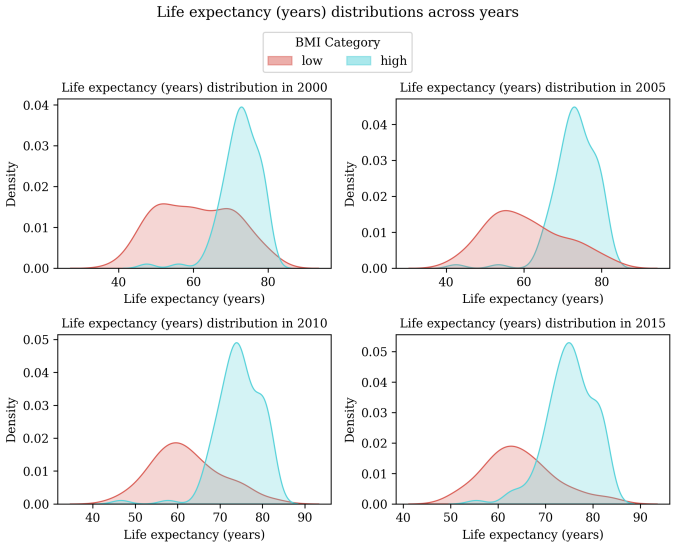


Figure 3. Life expectancy distribution in different BMI categories.

2.2.3. Alcohol Consumption

The results across all years (4) can't suggest that individuals with a "low" **Alcohol consumption** tend to live significantly longer than those with a "high" **Alcohol consumption** in all tested years. The independent two-sample t-test and Mann-Whitney U test failed to confirm our hypothesis that lower alcohol consumption is associated with longer life expectancy. This is quite surprising, as alcohol is known for having adverse health effects on humans.

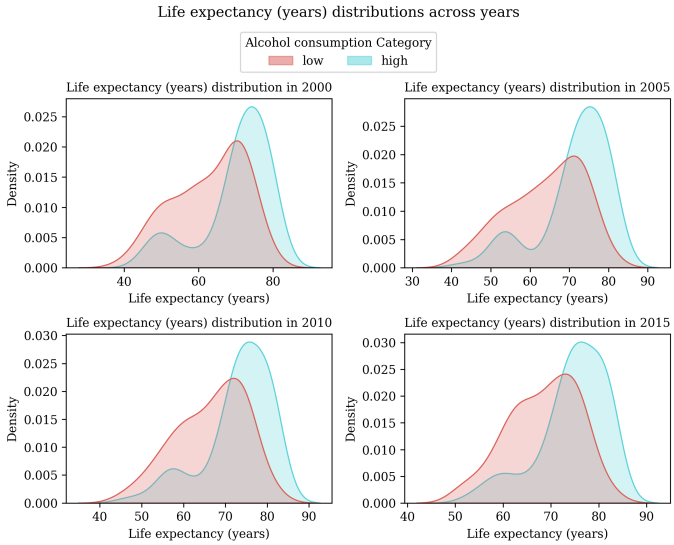


Figure 4. Life expectancy distribution in different Alcohol consumption categories.

2.2.4. Schooling

The results across all years (5) can suggest that individuals with a higher number of **Schooling** years tend to live significantly longer in all tested years. The independent two-sample t-test and Mann-Whitney U test confirmed our hypothesis that a higher number of **Schooling** years is associated with longer life expectancy. This is not surprising, as education is often related to economic growth, which influences public health greatly.

Figure 5. Life expectancy distribution in different Schooling categories.

2.3. Regression Analysis

The regression analysis (2) provides valuable insights into how socioeconomic and lifestyle factors influence life expectancy across different countries and regions. With an **R-squared** value of **0.803**, the model explains approximately 80% of the variance in life expectancy, suggesting a strong fit. Key predictors include **GDP per capita** and **schooling years**, both of which show significant positive associations with life expectancy, indicating that wealthier nations with better educational outcomes tend to have longer life spans. **BMI** and **alcohol consumption** have mixed effects, with alcohol consumption showing a notable negative impact, aligning with existing research on its adverse health effects. The categorical variable for **continent** also has a strong influence, highlighting the geographical disparities in life expectancy. Overall, this regression model confirms that economic development, education, and health-related behaviors are critical determinants of life expectancy, with potential for targeted interventions to improve health outcomes in lower-income regions.

3. Methods

3.1. Transformations

3.1.1. Log Transformation

Log transformations are commonly used in data analysis to normalize skewed distributions, making them more symmetrical and closer to a normal distribution. This technique is especially useful when data exhibits positive skewness, with a long tail on the right, as seen in variables like GDP per capita (6) or population size. By applying a log transformation, large values are compressed while smaller values are spread out, reducing the influence of extreme outliers and stabilizing variance. This transformation can improve the performance of statistical models that assume normality, such as linear regression, by making the data more suitable for analysis.

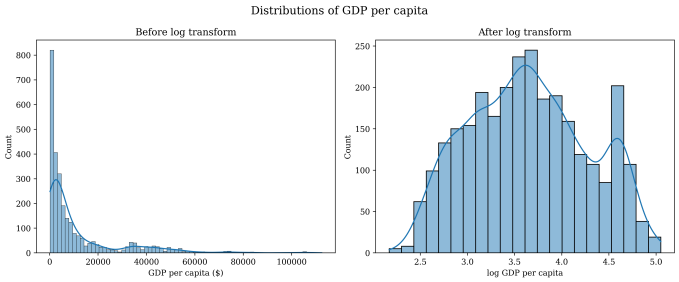


Figure 6. Distributions of **GDP per capita** before and after log transformation.

3.1.2. One-hot Encoding

One-hot encoding is a method used to convert categorical variables into a numerical format that machine learning algorithms can interpret. It involves creating binary columns for each category, where a value of 1 indicates the presence of the category and 0 indicates its absence. One-hot encoding was used to encode continents, where one of the columns ("Africa") was dropped to avoid multicollinearity.

3.2. Correlations

To explore the relationships between the variables in the dataset, a correlation matrix was generated using **Pearson's correlation coefficient**, which measures the linear association between two continuous variables:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- ρ: Pearson's correlation coefficient.
- x_i: i'th sample from first group.
- y_i: i'th sample from second group.
- $\bar{x}$: Mean of first group.
- $\bar{y}$: Mean of second group.
- n: Sample size.

The values of the correlation coefficient range from -1 to 1, where a value closer to 1 indicates a strong positive correlation, and a value closer to -1 represents a strong negative correlation. To assess the statistical significance of these correlations, p-values were calculated for each pairwise comparison. The p-values indicate the probability

Feature	Year	Category	Norm.	Var. h.	H_0 (Null Hypothesis)	T-value	U-value	p-value*	Reject H_0?
GDP per capita ($)	2000	low	Yes	No	$H_0^1 : \text{med}_{\text{medium GDP}} \leq \text{med}_{\text{low GDP}}$ $H_0^2 : \text{med}_{\text{high GDP}} \leq \text{med}_{\text{medium GDP}}$ $H_0^3 : \text{med}_{\text{high GDP}} \leq \text{med}_{\text{low GDP}}$	-	3953.5	0	Yes
		medium	No				3272.5	0	Yes
		high	Yes				1931.0	0	Yes
	2005	low	Yes	Yes		-	3723.0	0	Yes
		medium	No				3543.5	0	Yes
		high	Yes				1909.0	0	Yes
	2010	low	Yes	Yes	$H_0^1 : \mu_{\text{medium GDP}} \leq \mu_{\text{low GDP}}$ $H_0^2 : \mu_{\text{high GDP}} \leq \mu_{\text{medium GDP}}$ $H_0^3 : \mu_{\text{high GDP}} \leq \mu_{\text{low GDP}}$	9.586		0	Yes
		medium	Yes			7.454	-	0	Yes
		high	Yes			16.849		0	Yes
	2015	low	Yes	No	$H_0^1 : \text{med}_{\text{medium GDP}} \leq \text{med}_{\text{low GDP}}$ $H_0^2 : \text{med}_{\text{high GDP}} \leq \text{med}_{\text{med GDP}}$ $H_0^3 : \text{med}_{\text{high GDP}} \leq \text{med}_{\text{low GDP}}$	-	2764.5	0	Yes
		medium	Yes				4408.5	0	Yes
		high	Yes				1820.0	0	Yes
BMI (kg/m²)	2000	low	Yes	No	$H_0^1 : \text{med}_{\text{high BMI}} \leq \text{med}_{\text{low BMI}}$	-	6708.5	0	Yes
		high	Yes						
	2005	low	Yes	No		-	6635.0	0	Yes
		high	Yes						
	2010	low	Yes	No		-	6685.5	0	Yes
		high	Yes						
	2015	low	Yes	No		-	6383.5	0	Yes
		high	Yes						
Alcohol consumption (l)	2000	low	Yes	Yes	$H_0^1 : \text{med}_{\text{low alc}} \leq \text{med}_{\text{high alc}}$	-	2159.5	1.0	No
		high	No						
	2005	low	Yes	Yes		-	2091.0	1.0	No
		high	No						
	2010	low	Yes	Yes	$H_0^1 : \mu_{\text{low alc}} \leq \mu_{\text{high alc}}$	-5.312	-	1.0	No
		high	Yes						
	2015	low	Yes	Yes		-5.678	-	1.0	No
		high	Yes						
Schooling (years)	2000	low	Yes	No		-	4225.5	0	Yes
		medium	No				1928.0	0.004	Yes
		high	Yes				2349.0	0	Yes
	2005	low	Yes	Yes	$H_0^1 : \text{med}_{\text{medium sch}} \leq \text{med}_{\text{low sch}}$ $H_0^2 : \text{med}_{\text{high sch}} \leq \text{med}_{\text{medium sch}}$ $H_0^3 : \text{med}_{\text{high sch}} \leq \text{med}_{\text{low sch}}$	-	3450.5	0	Yes
		medium	No				2666.0	0	Yes
		high	Yes				2671.0	0	Yes
	2010	low	Yes	Yes		-	2733.5	0	Yes
		medium	No				3005.0	0	Yes
		high	Yes				2893.0	0	Yes
	2015	low	Yes	Yes	$H_0^1 : \mu_{\text{medium sch}} \leq \mu_{\text{low sch}}$ $H_0^2 : \mu_{\text{high sch}} \leq \mu_{\text{medium sch}}$ $H_0^3 : \mu_{\text{high sch}} \leq \mu_{\text{low sch}}$	6.074		0	Yes
		medium	Yes			5.940	-	0	Yes
		high	Yes			13.774		0	Yes

Table 1. Table summarizing all results.

* one-sided.

of observing the calculated correlation under the null hypothesis that there is no association between the variables:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

A p-value less than 0.05 was considered statistically significant, suggesting that the observed correlation is unlikely to have occurred by chance. This combination of correlation coefficients and p-values provides a comprehensive view of both the strength and the significance of the relationships within the dataset.

3.3. Statistical Tests

3.3.1. Normality Tests

To assess whether life expectancy follows a normal distribution across different features such as continent, and within specific years, we conducted two normality tests, with the choice of test based on the sample size. Usually, when a null hypothesis is not rejected, it is not accepted as statistically true. However, in these tests, it is widely accepted to assume normality if the null hypothesis is not rejected.

The hypotheses for the normality tests were as follows:

H_0 : The data is normally distributed.
H_1 : The data is not normally distributed.

Shapiro-Wilk Test: The Shapiro-Wilk test was used for datasets with a sample size less than or equal to a threshold of 30. This test is effective for small datasets and is more sensitive to detecting deviations from normality. The test produces the **W statistic**, which measures how well the data of a sample $X_1, ..., X_n$ correspond to a normal distribution, calculated as follows:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where:

- W: Shapiro-Wilk test statistic.
- $x_{(i)}$: Ordered sample values.
- $\bar{x}$: Sample mean.
- a_i: Weights derived from the covariance matrix of the ordered

sample.
- n: Sample size.

A value of W close to 1 indicates normality, while lower values suggest a deviation from normality.

Kolmogorov-Smirnov Test: For larger datasets, the Kolmogorov-Smirnov (K-S) test was applied. This non-parametric test compares the empirical cumulative distribution function (CDF) of the sample data to the CDF of a reference normal distribution. The test produces the **D statistic**, which measures the supremum difference between the empirical and reference CDFs, calculated as follows:

$$D = \sup_x |F_n(x) - F(x)|$$

where:

- D: Kolmogorov-Smirnov test statistic.
- $F_n(x)$: Empirical CDF of the sample.
- $F(x)$: CDF of the reference normal distribution.

A larger value of D indicates a greater deviation from normality.

As all sample sizes were sufficiently large (>30), the Kolmogorov-Smirnov test was exclusively used for assessing normality across all feature groupings.

3.3.2. Variance Homogeneity Test: Levene's Test

To assess the equality of variances across different groups, we applied **Levene's test**. This test evaluates the assumption of homogeneity of variances, which is a requirement for parametric tests like the independent t-test. Levene's test is less sensitive to deviations from normality than other tests for variance equality. Usually, when a null hypothesis is not rejected, it is not accepted as statistically true. However, in this test, it is widely accepted to assume variance homogeneity if the null hypothesis is not rejected. The null and alternative hypotheses are as follows:

$$H_0 : \sigma_1^2 = ... = \sigma_n^2$$
$$H_1 : \exists i \neq j : \sigma_i^2 \neq \sigma_j^2$$

Where $1, ... n$ are the different groups tested.

Levene's test produces the W statistic, which is calculated as follows:

$$W = \frac{(N-k)}{(k-1)} \cdot \frac{\sum_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Z_{i.})^2} \quad (1)$$

where:

- W: Levene's test statistic.
- N: Total number of observations across all groups.
- k: Number of groups.
- N_i: Number of observations in the i-th group.
- Z_{ij}: Absolute deviation of the j-th observation in the i-th group from the mean of that group.
- $Z_{i.}$: Mean of the absolute deviations for the i-th group.
- $Z_{..}$: Overall mean of the absolute deviations.

3.3.3. Independent Two-Sample t-Test

The independent two-sample t-test was employed to compare the means of life expectancy between two independent groups. This test is appropriate when the assumptions of normality and homogeneity of variances are met. Given that we did not know the standard deviation of the samples, the t-test was chosen over the z-test. Furthermore, we employed the version of the t-test that does not assume equal sample sizes, ensuring robustness regardless of sample size discrepancies. The null and alternative hypotheses are as follows:

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

Where μ_1 is the mean of the first group, and μ_2 is the mean of the second group.

The test statistic t for the one-sided t-test is calculated using the formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (2)$$

where s_p, the pooled standard deviation, is computed as:

$$s_p^2 = \frac{(n_1 - 1)s_{X_1}^2 + (n_2 - 1)s_{X_2}^2}{n_1 + n_2 - 2} \quad (3)$$

where:

- t: The t-test statistic.
- s_p: The pooled standard deviation of the two groups.
- $\bar{X}_i$: Mean of group i.
- $s_{X_i}^2$: Sample variance of group i.
- n_i: Number of observations in group i.

The pooled standard deviation of the two groups s_p is used as it provides an unbiased estimator of the common variance, whether or not the population means are equal.

This test was applied following verification of the assumptions through the Kolmogorov-Smirnov test for normality and Levene's

Table 2. OLS Regression Results

Variable	Coefficient	Std. Error	t-Statistic	P-value
Intercept	-0.7080	0.020	-35.041	0.000
Year	0.1140	0.009	13.179	0.000
log Population	0.0571	0.010	5.791	0.000
log GDP per capita	0.5343	0.014	36.945	0.000
BMI (kg/m²)	0.0133	0.014	0.981	0.327
Alcohol consumption (l)	-0.1835	0.014	-12.894	0.000
Schooling (years)	0.0844	0.017	4.924	0.000
Asia	0.7964	0.028	28.072	0.000
Europe	1.2062	0.037	32.905	0.000
North America	0.9178	0.075	12.315	0.000
South America	0.7623	0.043	17.852	0.000
Oceania	1.0414	0.031	34.131	0.000
R-squared		0.803		
Adjusted R-squared		0.802		
F-statistic		1055		
Prob (F-statistic)		0.000		

test for variance homogeneity. Moreover, we assumed independence between the groups (samples), as we assumed each country is mostly independent of the others. Additionally, given our a priori hypothesis regarding the relationship between the selected feature and **Life expectancy**, a one-sided t-test was employed to evaluate the anticipated direction of the effect. If the p-value is less than 0.05, we reject the null hypothesis, indicating a significant difference between the means of the two groups.

3.3.4. Independent Mann-Whitney U Rank Test

The Mann-Whitney U rank test, a non-parametric alternative to the independent two-sample t-test, was utilized to compare the distributions of **Life expectancy** between two independent groups. The independent test was employed as the features being compared were independent of each other, ensuring that the analysis accurately reflects differences between distinct groups without interference. This test is employed when the assumptions of the t-test are not met, particularly when the data do not follow a normal distribution or when the variances are unequal. The null and alternative hypotheses are as follows:

$$H_0 : \text{median}_1 \leq \text{median}_2$$
$$H_1 : \text{median}_1 > \text{median}_2$$

Where median_1 is the median of the first group, and median_2 is the median of the second group. The Mann-Whitney U statistic is calculated using the formula:

$$U = \min(U_1, U_2) \tag{4}$$

where:

- U_1 is the test statistic for the first group, computed as:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - W_1 \tag{5}$$

- U_2 is the test statistic for the second group, computed as:

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - W_2 \tag{6}$$

- W_1 and W_2 are the sums of the ranks for the first and second groups, respectively,
- n_1 and n_2 are the sample sizes of the two groups.

We used the Independent Mann-Whitney U rank test because the assumptions of the t-test—namely, normal distribution of the data and homogeneity of variances—were not met. A one-sided test was applied based on a priori assumptions regarding the relationship between the selected feature and **Life expectancy**. If the p-value is less than 0.05, we reject the null hypothesis, indicating a significant difference between the distributions of the two groups.

3.3.5. Bonferroni Correction

When conducting multiple statistical tests, the probability of incorrectly rejecting at least one null hypothesis (Type I error) increases with the number of tests performed. To address this issue, we applied the Bonferroni correction, a widely used method to control the family-wise error rate (FWER).

The Bonferroni correction adjusts the significance level for multiple comparisons to reduce the likelihood of Type I errors. It is a conservative method that involves dividing the desired overall significance level (α) by the number of tests (m) conducted. The corrected significance level (α_{adj}) for each individual test is calculated as follows:

$$\alpha_{\text{adj}} = \frac{\alpha}{m} \tag{7}$$

For each hypothesis test conducted, the p-value obtained was compared against the Bonferroni-adjusted significance level. Only those tests where the p-value was less than α_{adj} were considered statistically significant, thus allowing for more reliable conclusions while accounting for the multiplicity of tests performed.

3.4. Regression Analysis

In this analysis, we utilize the Ordinary Least Squares (OLS) method to estimate the relationship between socioeconomic factors and life expectancy. OLS assumes that the dependent variable y is normally distributed around the linear combination of predictor variables X with coefficients β and variance σ^2, i.e., $y \sim \mathcal{N}(X\beta, \sigma^2 I)$. The OLS estimator for the coefficient vector β is given by the Maximum Likelihood Estimator (MLE):

$$\hat{\beta} = (X^T X)^{-1} X^T y \sim \mathcal{N}\left(\beta, \sigma^2 (X^T X)^{-1}\right),$$

where $X^T X$ represents the matrix of predictor variables. To conduct hypothesis testing on the significance of each coefficient, we calculate the t-statistics for each estimated coefficient $\hat{\beta}_j$ using the formula:

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} S_j} \sim t_{n-p},$$

where:

- $S_j^2 = (X^T X)^{-1}_{jj}$: Variance of the estimated coefficient.
- $\hat{\sigma}^2$: Estimated variance of the residuals.
- $n - p$: Degrees of freedom.

The sample variance $\hat{\sigma}^2$ is computed as:

$$\hat{\sigma}^2 = \frac{1}{n-p}(y - \hat{y})^T (y - \hat{y}),$$

where:

- y: Vector of observed values.
- $\hat{y} = X\hat{\beta}$: Predicted values from the model.

This procedure allows for testing the null hypothesis:

$$H_0 : \beta_j = 0$$

The resulting p-values provide insights into the statistical significance of each socioeconomic factor in predicting life expectancy.

The model's goodness-of-fit was evaluated using the **R-squared** statistic, which represents the proportion of variance in the dependent variable explained by the predictors:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the sample mean of the observed values.

4. Codes

4.1. Imports

To perform the statistical tests, we first import the necessary libraries:

```
import scipy.stats as stats
import numpy as np
```

4.2. Normality Tests

4.2.1. Shapiro-Wilk Test

To perform the Shapiro-Wilk test for normality on datasets with sample sizes less than or equal to 30, use the following code:

```
statistic, p_value = stats.shapiro(data_to_test)
```

4.2.2. Kolmogorov-Smirnov Test

For larger datasets, the Kolmogorov-Smirnov test can be applied using:

```
statistic, p_value = stats.kstest(
                       data_to_test,
                       'norm',
                       args=(
                       np.mean(data_to_test),
                       np.std(data_to_test)
                       )
                     )
```

4.3. Variance Homogeneity Test: Levene's Test

To assess the homogeneity of variances using Levene's test, use the following code:

```
W_value, p_value = stats.levene(*samples)
```

4.4. Independent Two-Sample T-Test

For comparing the means of **Life expectancy** between two independent groups, use the following code:

```
t_value, p_value = stats.ttest_ind(
                     data_to_test1,
                     data_to_test2,
                     alternative="greater"
                   )
```

4.5. Independent Mann-Whitney U Rank Test

For non-parametric comparison of two independent samples, use:

```
U_value, p_value = stats.mannwhitneyu(
                     data_to_test1,
                     data_to_test2,
                     alternative="greater"
                   )
```

5. Discussion

5.1. Conclusions

The study revealed significant insights into the socioeconomic factors influencing life expectancy across different regions. A clear positive association was found between GDP per capita, years of schooling, and life expectancy, indicating that wealthier countries with better educational systems tend to experience longer lifespans. These findings underscore the critical role of economic development and education in enhancing public health outcomes. Surprisingly, higher BMI was also associated with increased life expectancy, despite being traditionally linked to negative health outcomes such as obesity. This counterintuitive result suggests that the relationship between BMI and life expectancy may vary across regions, potentially reflecting differences in healthcare systems or lifestyle factors. On the other hand, alcohol consumption showed a generally negative effect on life expectancy, though statistical significance was not consistently confirmed, suggesting that its impact may be more complex or context-dependent. Additionally, the study highlighted significant geographical disparities, emphasizing global inequalities in health outcomes.

5.2. Limitations of Analysis

The study faced several limitations. First, the availability and consistency of data across countries and time periods could have impacted the robustness of the findings, as some regions may lack accurate or reliable data. Moreover, while the study identified correlations between socioeconomic factors and life expectancy, it did not establish direct causal relationships. Variables such as GDP and schooling may serve as proxies for more nuanced factors like healthcare access or policy differences. Unmeasured variables, such as healthcare quality, environmental conditions, and cultural influences, may also significantly affect life expectancy but were not captured in the analysis. In addition, the study only analyzed data across 16 years (2000-2015), which may impact the robustness of the results shown. Finally, the unexpected results regarding BMI and the mixed findings on alcohol consumption indicate the potential influence of other unmeasured lifestyle or genetic factors, complicating the interpretation of these variables. These limitations suggest that further research, incorporating a broader range of variables and more granular data, is necessary to fully understand the determinants of life expectancy.

References

[1] NCD Risk Factor Collaboration, "National Adult Body-Mass Index", https://www.ncdrisc.org/data-downloads-adiposity.html.

[2] Our World in Data, "Mean years of schooling (long run)", https://ourworldindata.org/grapher/mean-years-of-schooling-long-run.

[3] Source GitHub repository, https://github.com/AdamZlr/Statistical-Theory.

[4] World Health Organization, "Alcohol recorded per capita (15+) consumption (in litres of pure alcohol)", https://www.who.int/data/gho/data/indicators/indicator-details/GHO/alcohol-recorded-per-capita-(15-)-consumption-(in-litres-of-pure-alcohol).

[5] World Health Organization, "Life expectancy at birth (years)", https://www.who.int/data/gho/data/indicators/indicator-details/GHO/life-expectancy-at-birth-(years).

[6] World Bank, "GDP per capita (current USD)", https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?most_recent_year_desc=true.

[7] World Bank, "Total population", https://data.worldbank.org/indicator/SP.POP.TOTL?most_recent_year_desc=true.