Analysis of Economic and Social Indicators Across Countries

Eloy Celaya López

Cis 331 - Winter 2025

Table of Contents

1. Introduction

1.1. Motivation & Dataset Description

Understanding the factors that drive development is crucial for shaping effective policies and reducing global inequality. In a world where access to health, education, and economic resources varies dramatically between countries, analyzing development indicators can help uncover the root causes of these disparities. By exploring data-driven insights, this project aims to contribute to the broader goal of promoting sustainable and inclusive global development. This project uses data from the World Bank's World Development Indicators dataset. This website allows you to choose from 1496 different indicators, every country in the world and more than 60 years, making each created dataset unique. The created dataset includes 23 indicators such as GDP per capita, health and education expenditure, school enrollment, access to basic water services, infant mortality rate, and life expectancy. These indicators span over 200 countries between the years 2000 and 2021, allowing for a broad analysis of global trends in development.

1.2. Project Goals

The main goals of the project are to explore disparities in development across regions, understand how different social and economic features relate to one another, and identify which variables are most informative when it comes to characterizing, predicting or modeling a country's developmental status. Key questions guiding this work include:

- How are education, health, and income levels connected across countries?
- What factors appear most important in explaining life expectancy differences?
- Can certain indicators provide early signals of a country's economic standing?

Based on the research questions and domain knowledge, this study focuses on exploring several key relationships between development indicators. These include examining the correlation between GDP per capita and life expectancy, and analyzing how education levels influence both employment outcomes and income distribution. The project also looks at the relationship between healthcare expenditure and overall health, particularly in terms of life expectancy and infant mortality. In addition, it explores how access to basic resources like clean water connects to broader patterns of economic development. Finally, a key goal of the study is to identify which indicators are most useful in predicting how developed a country is, as well as which ones are the best predictors of life expectancy.

1.3. Real-World Applications

Understanding the relationships between development indicators is not only useful for academic research, it has important real-world applications. By identifying which factors are most strongly associated with better health outcomes or higher economic performance, this analysis can help inform more effective policy decisions, international aid strategies, and resource allocation. For example, insights from the data can support governments in prioritizing investments in areas like education, infrastructure, or public health. Additionally, using predictive models to assess a country's level of development or forecast life expectancy based on socioeconomic features can provide valuable tools for anticipating needs, measuring progress, and shaping long-term development goals.

2. Data preparation

2.1. Loading and Analyzing Data

The first and one of the most important steps to properly analyze a data set is data preprocessing or preparation. With it, we can prepare our data so it can be correctly analyzed later.

After importing my data set and displaying, the format was not the best to work with. At first there were 26 columns: Country Name, Country Code, Series Name, Series Code and one column for every year (2000-2021) with the value for every Country-serie (indicator) combination. Also, one of the first problems I found was the format of each Year's title, which was "YYYY ['YR'YYYY]", a really unusual format. The first step I took was changing the format to a more simple "YYYY", which made the dataset cleaner.

Since the data set was downloaded from a website, the missing values were displayed as "..", so they wouldn't appear as missing. I changed all the ".." for "NaN" to correctly display missing values. The next thing I needed was for every indicator value to take a numeric format, since all the indicators were numeric, so I converted every value starting from the 5th column to numeric.

After considering that for some early years or some really late ones there could be lots of missing values, I decided to check it before continuing and, if for any year there were more than 50% missing values drop the year.

Although all the values are moderately high, they don't differ much from each other and none of them are above 50%. Therefore, none of the years were removed.

After this quick check, the next step I took was to begin changing the format of the data set. Firstly, I reshaped the data from wide to long format, by creating two now columns: Year, which would hold every year repeatedly for every country and Value, which would display the correct indicator value for every Country, Year and Series (Indicator) combination. I also changed the format of the Year variable to numeric (int) to properly use it later. Since the Series Code column is just a duplicate from Series, I decided to drop it. The final step to have the data in a correct and easy to use format was to display each indicator as a separate variable and column, instead of having all of them in the same column. Now, the dataset had a Country Name, Country Code, Year, and one column for every indicator, with one different value for every Country-Year combination.

I then realized there was a missing variable that could be really important for regional analysis, the region for each country. Luckily, the World's Bank included downloadable metadata, which had every country and their regions listed. After importing it to my notebook I merged the Region column with the data set by Country Code, which appeared in both data sets. After checking for missing values in the new added column, there weren't any, which was great.

Now that the dataset had a clearer format, I could begin to analyze it. There are 4774 rows and 29 columns.

After displaying the metadata, I learnt that there are 3 object variables: Country Name, Country Code and Region, a Year variable, which is numeric and 25 numeric variables, each of one indicator.

After a quick check of the distribution for each variable and the main statistics, there seemed to be no major errors with the values, all the executed transformations looked good.

2.2. Handling Missing Values

The next step was to check for missing values and act accordingly. After calculating missing values and percentages, I created a table to display the information and easily check missingness. It is important to highlight that the variables Country Name, Country Code, Year, Population, total and Region have no missing values.

In this data set, missing values are a big problem, since almost every variable has some, with some having more than 50% of the values missing. This presents a big challenge, especially because both of the inequality indicators, Gini index and Income share held by the highest 10% have more than 66% of their values missing, which removes the mean or median imputation option, since there are too many missing values. Estimating variables with many missing values could greatly increase the noise of the data set, so it is not a great option. Therefore, I will be dropping every indicator with more than 35% missing values. This includes Government expenditure on education, total (% of government expenditure), Income share held by highest 10%, Literacy rate, adult total (% of people ages 15 and above), Physicians (per 1,000 people), School enrollment, primary (% net) and School enrollment, secondary (% net).

It is possible that for some Country- Year combinations, most of the indicators could be missing. We are also going to drop every row that has more than 50% of its values missing, if the total number is not too high. After checking how many rows have more than 50% of their values missing, we get a total number of 82, which represents only 1.7% of all the rows. Therefore, they were dropped. Now that the variables and rows with most columns were dropped, we are left with a dataset consisting of 4692 rows and 22 columns, and, most importantly, noise won't increase as much when executing imputations for the rest of the missing values.

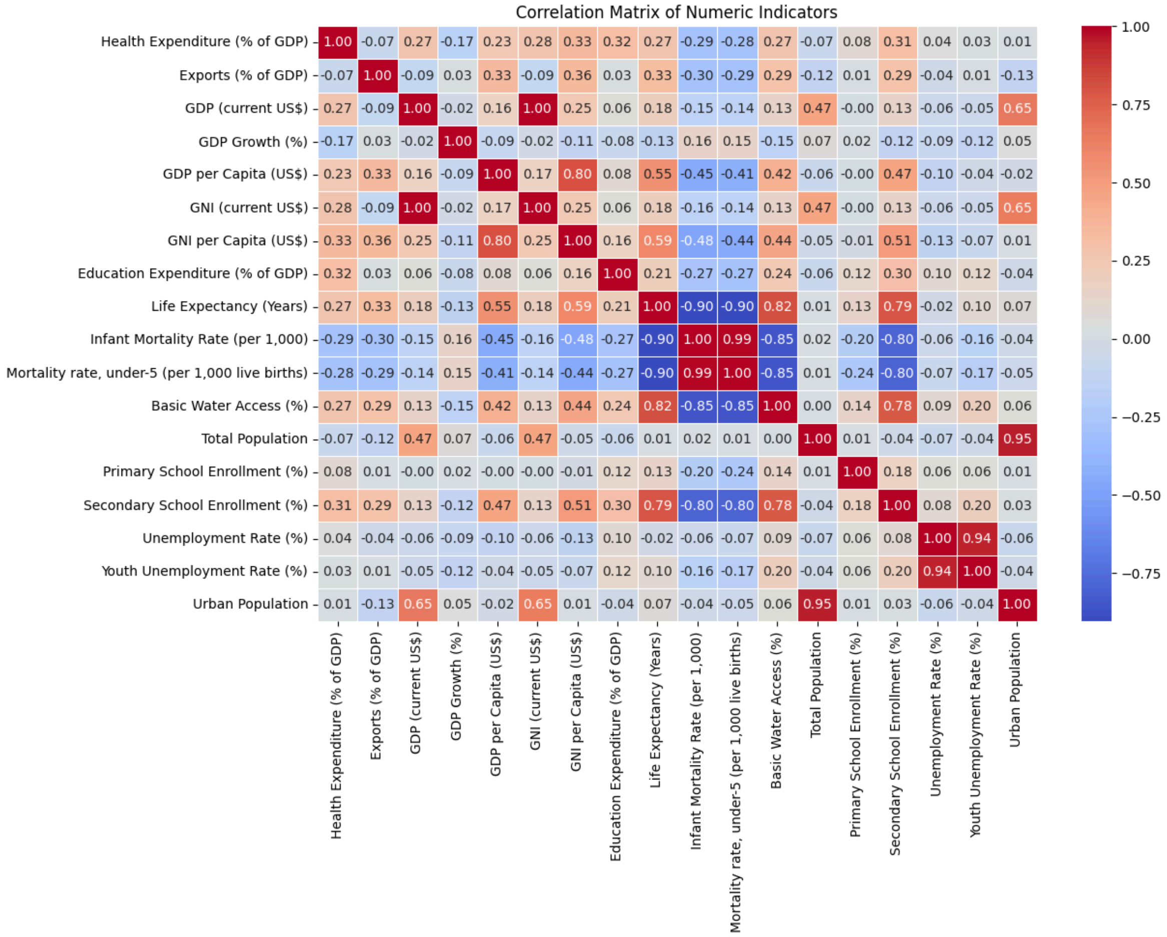
Since all the indicators are numeric, I will use the median imputation method to estimate them, but instead of using the median of each variable, I will group the data by Region and apply the method with the median of each region. Using the median instead of the mean ensures that the value used to predict missing values isn't biased due to outliers, and using the region medians ensures a more accurate prediction, since countries in the same regions tend to be similar to each other. All the medians were checked to see if they had reasonable values and, when doing this I realized the medians for School enrollment, primary (% gross) had some regional values just above 100%, which is strange. After looking more into it and checking the official metadata, I learnt that this variable is calculated by dividing the total number of students enrolled in primary education, regardless of age, by the population of the age group which officially corresponds to primary education and multiplying by 100.

To end with missing values for now, I checked again the data set for missing values and the only variable that still had some was the Gini Index, which hasn't been handled yet. The data set is now clean and ready to use for the EDA. The last small step I took was to change the names of each indicator, to keep them short and improve readability in every visualization.

3. Exploratory Data Analysis (EDA)

3.1. Correlation Matrix

The first step I took in the Exploratory Data Analysis is to create the correlation coefficients matrix and check the correlation between every variable. Even though the number of variables isn't extremely high, this will also help with dimensionality and multicollinearity reduction. Checking the correlation between every variable could help to eliminate attributes that have high correlation with each other but keeping the ones that have high correlation with some factors we will later predict, such as the Gini index or Life expectancy. Bellow is the correlation matrix:

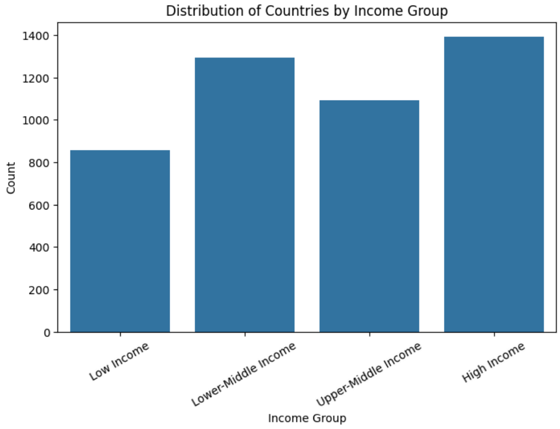


There is a perfect positive correlation between GDP total and GNI total, so the second variable will be dropped, since it doesn't have any high correlation with any of the variables we want to predict. There is another high correlation between GDP per capita and GNI per capita, but both variables will be kept since they could be used for different things. There is also an almost perfect positive correlation between population and urban population, which is normal, but both will be kept since they can be used to show different things and can even be combined to show urban population %. Mortality rate under 5 and Mortality rate infant have again an almost perfect correlation, as we could have predicted. We are going to drop Mortality rate under 5, and the infant Mortality rate will be used for the analysis. Mortality rate and people using at least basic drinking services both have a strong correlation with Life expectancy and with each other, but both will be kept since Life expectancy will be used as a response variable and both have high correlation with it. Finally, Unemployment and Youth Unemployment also have a very high positive correlation, but we will keep both variables for now.

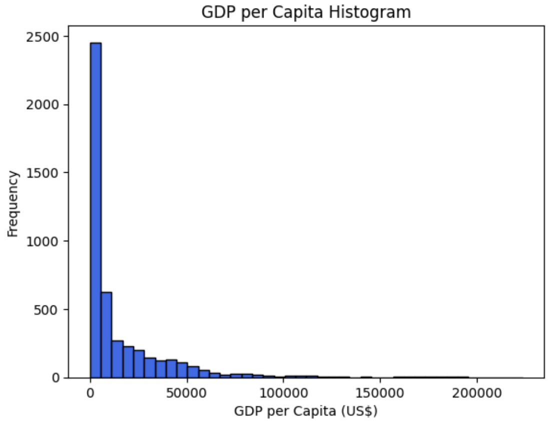
Now that the variables GNI total and Mortality rate under 5 have been dropped, let's add a new variable. This new variable is the "Income Group", that will be a categorical variable that classifies countries based on their GNI per capita into four categories: Low income (<$1000), Lower-Middle Income ($1000-$3999), Upper-Middle Income ($4000-$11999) and High Income (>$12000).

3.2. Univariate EDA

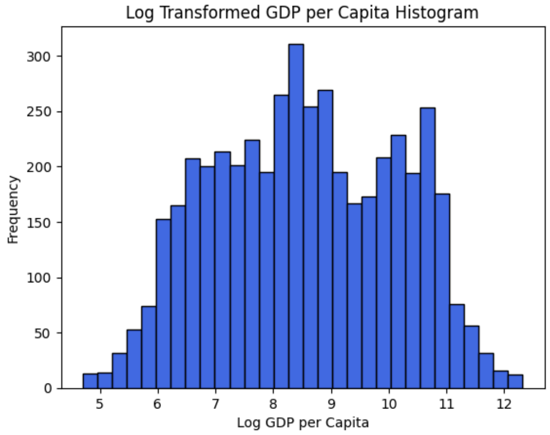
Let’s start by analyzing the newly created Income Group variable:



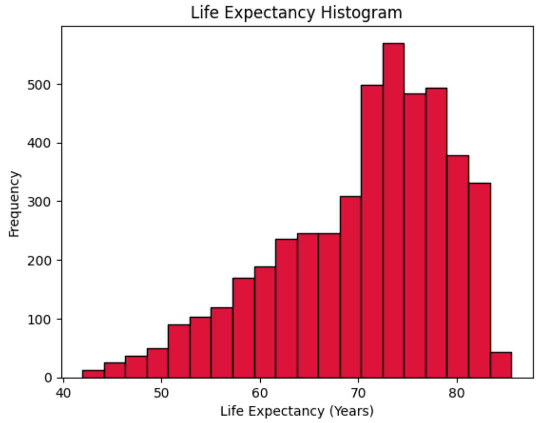
On the left we can see a bar chart that shows the distribution of Countries by Income Group, showing the total count for each Income Group. The Group with the least number of countries is Low Income, with a count of 857, and the group with the highest count is High Income, with 1391 countries.



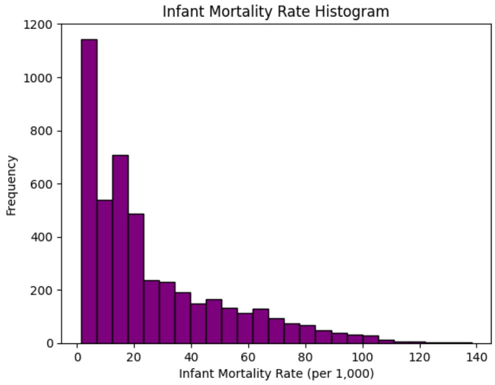
On the left we can see the Histogram for GDP per capita. There is a huge peak in the first bin (almost 2500 countries), meaning lots of countries have really low GDP per Capita. We can see a really high right-skewed distribution, so a log transformation should be applied, since it will compress higher values and remove the high concentration of countries in the first bin.



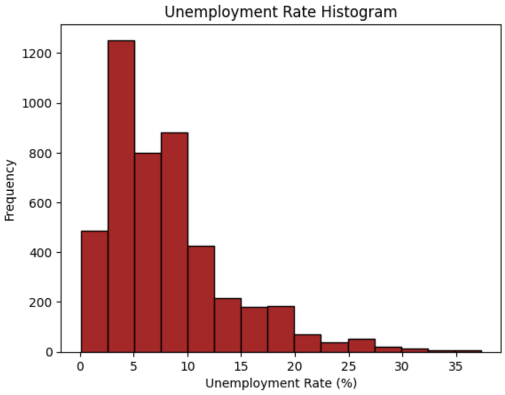
This figure shows the variable GDP per Capita after applying the log transformation to normalize the distribution. This new variable follows a normal distribution, which will help in later analysis since it works better for modeling.



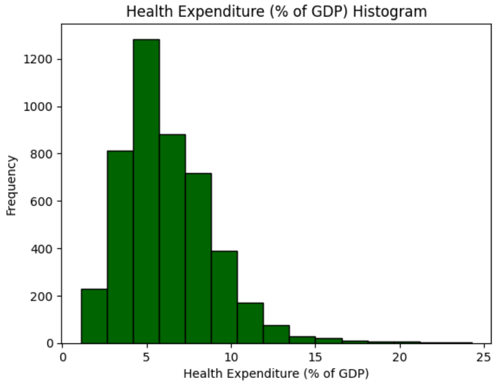
On the left, the histogram shows the distribution for Life Expectancy. Although it is a bit left skewed, no transformation is needed. The peak is around 72 years, with only a few countries with more than 85 and less than 50 years of Life Expectancy,



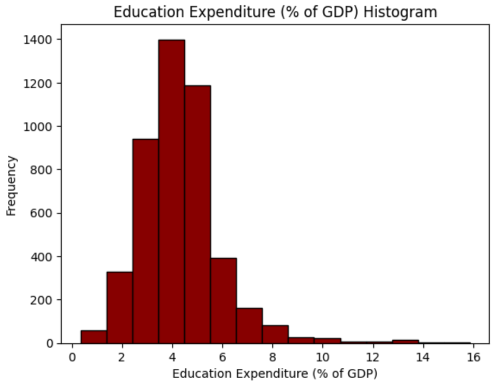
The distribution for Infant Mortality Rate is again right skewed, with most of the data concentrated in low mortality values. Again, the best option is to apply a log transformation to normalize the variable.



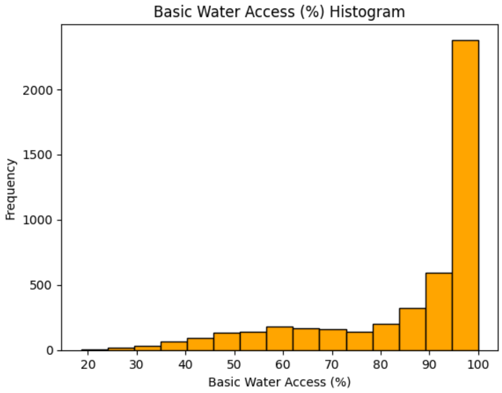
The variable Unemployment Rate follows a right skewed distribution, with a peak around 4% with more than 1200 values. A log transformation will be again applied to this variable.



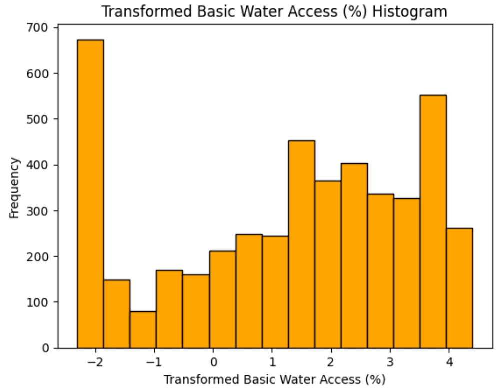
We can see how most countries spend between 3 and 8% of GDP in Health, with the peak at around 5% with more than 1200 countries. Some countries have a really high Health Expenditure, making the histogram a bit right skewed.



Most countries spend between 3 and 5% on Education, with the peak being 4% with about 1400 countries. Again, there is some visible right skewness on the distribution.

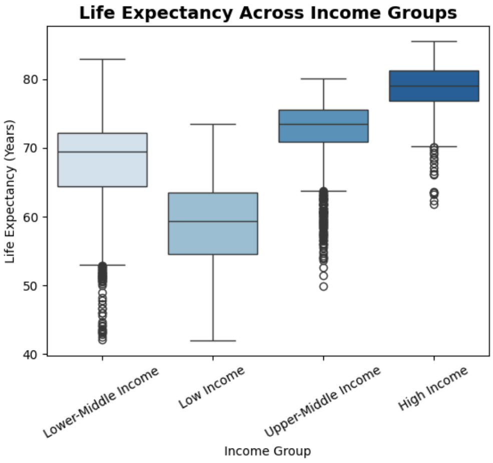


Regarding basic water access, we can see how the variable's distribution is really skewed left, with most countries having nearly 100% water access. Since it is extremely skewed, I applied an inverse log transformation.

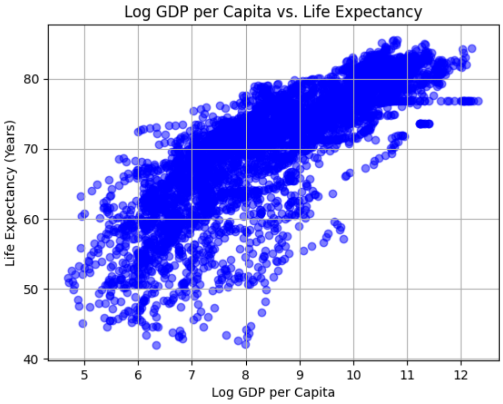


After transforming the variable, now low values mean high water access. We can see a peak at extremely low values, which represent countries with perfect basic water access, but the peak is not as high as before applying the transformation. There is much less skewness.

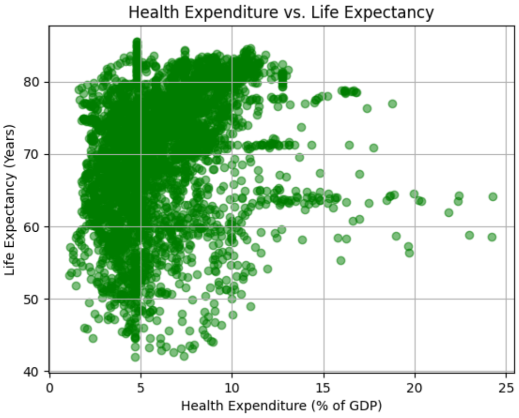
3.3. Bivariate EDA



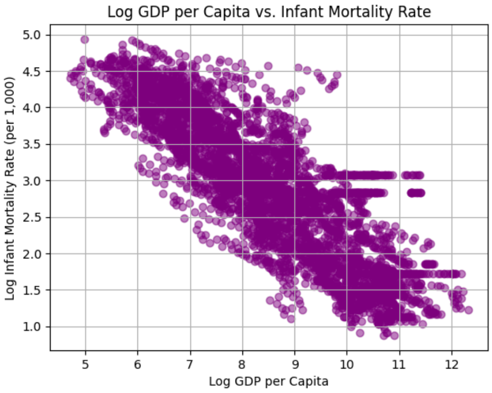
Income Group and Life Expectancy:
This visualization shows 4 side by side boxplots, one for each Income Group. As we could think before plotting the data, the higher the Income is for a group, the higher its Life Expectancy is. Almost every group has lots of outliers and we can conclude that these variables have a positive relationship that will be later studied in detail.



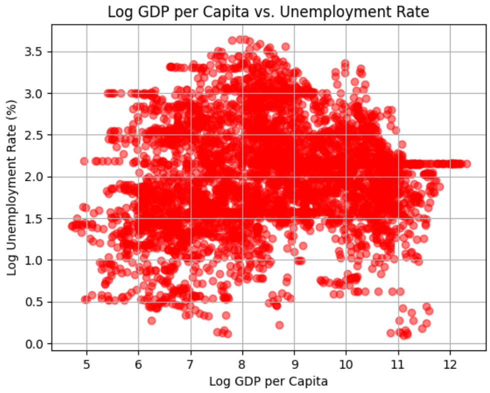
We are also going to check the relationship between Life Expectancy and the transformed variable Log GDP per Capita with a scatter plot.
We can see that there is a strong positive relationship between the two variables, meaning that as Life Expectancy increases GDP per Capita increases. The spread at low GDP Log GDP per Capita means that other factors also influence Life Expectancy, but, overall, it looks like GDP per Capita has a great influence on Life Expectancy.



Let’s also check if countries that spend more on Healthcare have higher Life Expectancy:
There is no clear relationship between the two variables and there is a bug cluster of points around the 5%, which most countries spend on Health. Some countries spend a lot but their Life Expectancy is not as high.



Does higher income reduce Infant Mortality Rates?
There clearly is a strong negative relationship, meaning that as GDP per capita increases, Mortality Rate decreases, although there is higher variability in lower GDP countries.



Do wealthier nations have lower unemployment?
We can see how the spread of points is even, so there is no clear relationship between GDP per Capita and Unemployment Rate. Therefore, we can not say that wealthier nations have lower unemployment rates.

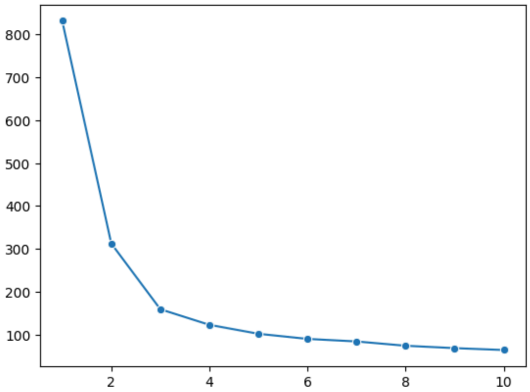
4. Model Building and Evaluation

4.1. Unsupervised Learning - Clustering

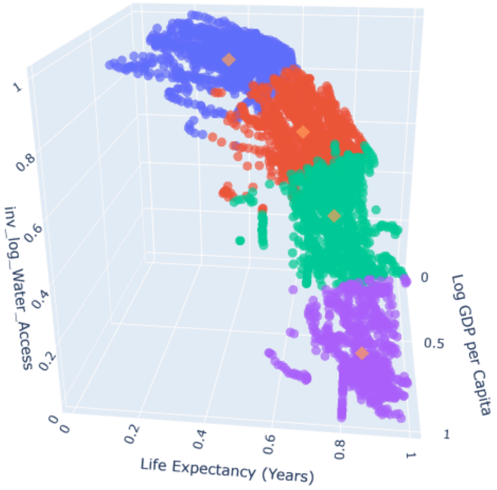
A k-means clustering model was applied as the unsupervised learning model for the variables Log GDP per Capita, Life Expectancy (Years) and Transformed Basic Water Access.

The objective of this unsupervised learning model was to identify meaningful groups of countries based on development indicators without relying on any predefined labels. This approach allows us to group countries with similar profiles, such as GDP per capita, life expectancy, and access to basic water services, and to better understand how these indicators interact in shaping global development outcomes.

K-means was selected due to its simplicity, efficiency and ability to partition countries into distinct, non-overlapping clusters based on Euclidean distance. Since the goal was to identify groups of countries with similar development profiles, K-Means provided an effective way to do this using continuous, scaled features.



To determine the optimal number of clusters, I used the elbow method, plotting the within-cluster sum of squares (WCSS) for different values of k (as shown in the left). 4 clusters were selected, as the WCSS curve noticeably flattened at that point, indicating diminishing returns from adding more clusters.



Different plots were created to visualize the clusters, with one of them being an interactive 3D graph, which can be seen on the left. The interactive plot can be easily accessed using this link. In the graph, the 4 clusters can be easily appreciated. One cluster (purple) represents countries with high GDP per capita, high Life Expectancy and almost 100% basic water access. Both the red and green clusters represent countries with moderate values for all three indicators, and the blue cluster represents the most undeveloped countries, with low values for every variable.

Clustering Evaluation & Conclusions

To evaluate the clustering quality, I examined the Within-Cluster Sum of Squares (WCSS), which measures how tightly grouped the points are around their assigned centroids. The final K-means model with 4 clusters achieved a WCSS of approximately 122.8, indicating relatively compact clusters. However, when inspecting the 3D visualization, it becomes clear that the data does not exhibit strongly separated or clearly predefined groupings — most points form a continuous cloud without sharp boundaries. This is reflected in the Silhouette Score, which was approximately 0.40, suggesting that while the model captured general trends in the data, the clusters are not strongly distinct.

Still, the clusters provide a useful approximation of global development patterns, grouping countries based on shared characteristics and enabling broad comparisons across income, health, and infrastructure. The analysis showed how unsupervised learning can reveal patterns without labeled data and emphasized the value of transformations and visualizations in interpreting complex datasets.

4.2. Supervised Learning - Regression

A linear regression model was developed to predict a country's Life Expectancy based on several socioeconomic and infrastructure-related indicators. The motivation behind this task was to understand how development factors such as income, education, and access to basic services impact population health and longevity, as well as creating a model that can correctly predict Life Expectancy when we don't have this attribute, but only other predictors. Linear regression was chosen because Life Expectancy is a continuous variable and the model provides high efficiency, ease of implementation, and ability to provide direct insight into variable relationships through coefficients.

After performing correlation analysis on all available variables, we selected a subset of features that showed moderate to strong correlation with the target variable. These included: Log GDP per Capita, Health Expenditure (% of GDP), Exports (% of GDP), Education Expenditure (% of GDP), Basic Water Access (%), Log Infant Mortality Rate (per 1,000), Secondary School Enrollment (%), GNI per Capita (US$), inv_log_Water_Access. The target variable was Life Expectancy (Years).

A matrix X was created with the predictor attributes and a vector y with the target, and then the data was split horizontally into training and testing sets using a 80/20 split. A standard linear regression model from scikit-learn was then trained on the training data. The main advantage of linear regression is its transparency and low variance, meaning it's less prone to overfitting compared to more complex models. It also runs efficiently on large datasets and is straightforward to evaluate using metrics like R^2 and RMSE. However, linear regression assumes a linear relationship between predictors and the response variable, which may not fully capture more complex patterns in the data.

Additionally, in this analysis, the input data was not standardized, which affects the interpretability of the model coefficients. While the model still performs well in terms of predictive accuracy, the raw coefficient values should not be directly compared or used to determine variable importance without accounting for the differing units and scales of the features.

Regression Model Evaluation & Conclusions

The linear regression model was evaluated using Root Mean Squared Error (RMSE) and R-squared (R^2). On the training set, the model achieved an RMSE of 3.82 and an R^2 of 0.81, while on the test set, the RMSE was 3.68 with an R^2 of 0.83. These results indicate that the model generalizes well, with no signs of overfitting, as its performance on new data is consistent with the training data.

In the context of this problem, an RMSE of 3.68 means that, on average, the model's predictions for life expectancy differ from the actual values by approximately 3.7 years. Given that global life expectancy ranges roughly between 40 and 85 years, this level of error is relatively small and suggests that the model captures key patterns in the data. The R^2 value of 0.83 further indicates that the model explains 83% of the variance in life expectancy, making it a strong baseline for predictive performance. Therefore, we can conclude that our model works great for predicting Life Expectancy.

4.3. Supervised Learning - Classification

The goal of this classification task was to predict a country's Income Group (Low, Lower-Middle, Upper-Middle, or High) using a set of development indicators. Two different models were applied to compare performance: a Decision Tree Classifier and a K-Nearest Neighbors (KNN) model. The objective was not only to evaluate their predictive performance but also to understand how each model handles the structure of socioeconomic data. Each model used five key features: Log GDP per Capita, Health Expenditure (% of GDP), Education Expenditure (% of GDP), Basic Water Access (%), and Secondary School Enrollment (%).

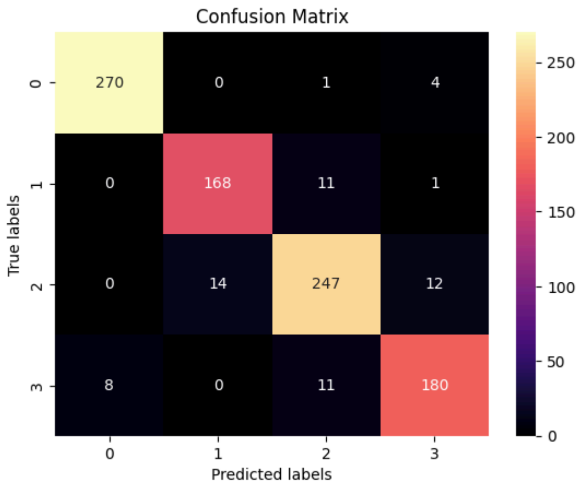
4.3.1. Decision Tree

The Decision Tree Classifier was chosen for its clarity and interpretability. This model works by recursively splitting the data based on feature thresholds, making it easy to understand how specific variables contribute to classification decisions. The model used the entropy criterion, and one big advantage is that the data doesn't have to be scaled, the model can be applied to different scaled variables. However, a notable drawback is their tendency to overfit the training data, particularly when the tree is allowed to grow too deep.

Decision Tree Evaluation

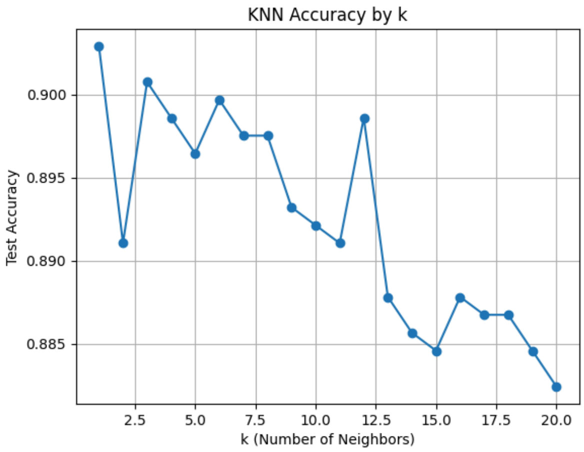
The Decision Tree model performed very well in classifying countries into income groups. On the training data, the model reached 100% accuracy, which means it perfectly classified all training examples. While this might seem ideal at first glance, it usually suggests that the model has likely overfit the training data — in other words, it may have learned patterns too specifically and might not generalize as well to new data. That said, the model still achieved a 93% accuracy on the testing set, which is a strong result. The confusion matrix shows that most predictions across all four income classes were correct. For example, Class 0 had 270 out of 275 cases correctly predicted, and the other classes had similarly high accuracy with only a few misclassifications.

The classification report also showed consistent results, with precision, recall, and F1-scores all above 0.90 for every class. This indicates that the model didn't just do well on average, but also treated each class fairly evenly. To test how well the model generalizes beyond a single test set, I ran 6-fold cross-validation, which gave an average accuracy of about 88%. This drop from the test accuracy is expected and suggests that while the model does generalize decently, its performance can vary slightly depending on the data split. In summary, the Decision Tree provided strong and balanced predictions, and while it showed signs of overfitting during training, it still performed reliably on unseen data.



4.3.2. KNN

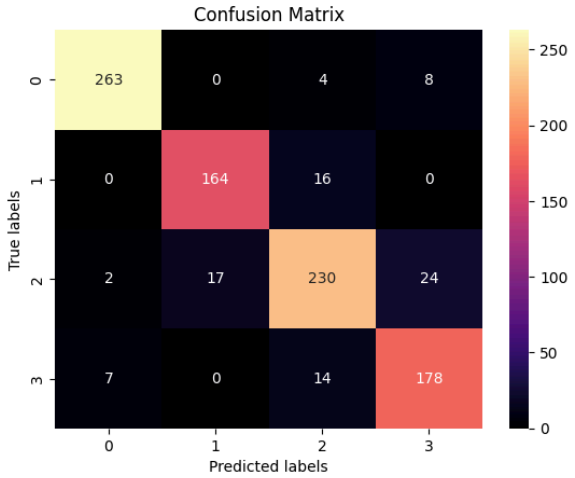
The K-Nearest Neighbors (KNN) model was selected as a complementary approach due to its simplicity and its ability to model nonlinear relationships without requiring assumptions about the data distribution. KNN classifies data points based on the majority class of their closest neighbors in the feature space, making it particularly effective when class boundaries are not easily captured by rule-based splits. One of its main strengths is its flexibility, since with proper tuning (such as choosing an optimal value for k) and feature scaling, it can perform competitively across many datasets. However, KNN has important limitations: it is sensitive to the scale of the data, requiring standardization, and it can become computationally expensive for large datasets.



Lots of different k values were tried, to choose one with high accuracy while trying to avoid overfitting. We can see the accuracy for some different values of K on the graph just to the left. K = 5 was chosen after considering different options. After this, a pipeline with scaling and cross-validation was used to ensure proper generalization and to avoid data leakage.

KNN Model Evaluation

The K-Nearest Neighbors (KNN) model showed strong overall performance, though it was slightly less accurate than the Decision Tree in certain areas. On the training data, the model achieved about 95.6% accuracy, and on the testing set, it scored 90.1%, which is quite solid. This small drop between training and testing performance suggests that the model generalizes well without heavily overfitting the training data.



Looking at the confusion matrix, most of the classifications were correct across the four income group classes. However, there were a few more misclassifications compared to the decision tree, particularly in Class 2 and Class 3. For example, 24 countries from Class 2 were incorrectly classified as Class 3. Despite that, the model maintained consistent predictions overall.

The classification report further supports this, with precision, recall, and F1-scores all at 0.90 on average. Class 0 again performed the best (F1 = 0.96), while Class 2 and Class 3 had slightly lower F1-scores (0.86 and 0.87 respectively), indicating a bit more difficulty distinguishing between the middle-income groups.

To better evaluate the model's performance across different data splits, a 6-fold cross-validation was also performed using a pipeline that included scaling. The average accuracy across all folds was 81%, which is lower than the test set accuracy. This drop suggests that while the model performs well on the current test split, its generalization may be slightly more limited depending on the data variation. However, this is not uncommon for KNN, especially when some classes have overlapping features.

Overall, the KNN model performed very well, especially considering its simplicity. While slightly less consistent than the decision tree, it showed solid results with good accuracy and balanced performance across all classes. With proper scaling and tuning (as you applied), it proves to be a reliable model for this kind of classification task.

4.3.3. Classification Conclusions

Both the Decision Tree and K-Nearest Neighbors (KNN) models were able to predict countries' income groups with high accuracy using development indicators such as GDP per capita, education, health expenditure, and access to water. While both models performed well, the Decision Tree outperformed KNN in terms of accuracy and consistency across classes. The Decision Tree achieved a 93% test accuracy and showed strong performance across all income levels, with precision and recall consistently above 0.90. It also had a near-perfect result on the training data (100% accuracy), which suggests some overfitting, but the cross-validation score (88%) confirmed that it generalized well to new data.

The KNN model, while slightly behind, still achieved a solid 90% test accuracy, and also demonstrated balanced performance across all classes. It had a smaller gap between training and testing accuracy, indicating less overfitting, but its cross-validation score averaged 81%, which suggests that it was more sensitive to how the data was split. This is a known limitation of KNN, especially in cases where classes are less clearly separated.

In the context of this project, accuracy reflects how often the model correctly classifies a country into the right income group. Since these predictions could be used to guide development aid or policy planning, high accuracy and consistent classification are essential. Misclassifying a low-income country as high-income, for example, could result in that country being overlooked for support. In this sense, the Decision Tree's more stable performance makes it the more reliable option for this application.

Overall, both models were effective, but the Decision Tree model is the better choice due to its higher generalization performance, strong metrics, and ease of interpretation for real-world decision-making.

5. Reflection

During this project, I encountered several challenges that required careful problem-solving and adaptability. One of the most notable difficulties was dealing with missing or inconsistent data. In some cases, entire variables had to be dropped, while in others, I had to impute values using region-based medians or decide whether transformations were necessary to reduce skewness. Another challenge was selecting the most relevant features without overloading the models. Balancing interpretability and predictive performance often required revisiting earlier stages of the workflow.

To overcome these challenges, I used a mix of documentation, class notes, and trial-and-error. For example, I initially had issues applying transformations to variables like water access, but by visualizing distributions and evaluating skewness, I decided on log and inverse-log transformations. In building the supervised models, I experimented with different techniques (like decision trees and KNN), and I learned how to prevent overfitting. When unexpected errors occurred in coding or model interpretation, I carefully debugged step-by-step or sought help from trusted resources.

From this project, I've gained a much deeper understanding of how real-world datasets behave and how thoughtful preprocessing significantly impacts the success of a machine learning model. I learned the value of visualizing the data early on and how performance metrics like RMSE or accuracy need to be interpreted in the context of the problem, not just as abstract numbers. I also now appreciate the differences between regression and classification models and how each offers unique insights depending on the question being asked. Overall, this experience strengthened my skills in exploratory data analysis, model building, and result interpretation, while reinforcing the importance of flexibility and curiosity when facing messy, imperfect data.