**Introduction:**

This report demonstrates the process of training and testing of a regression model. The dataset consists of both numerical and categorical data, with 2000 samples describing the conditions of different countries and the target for prediction, the expectancies for each sample. The process includes exploratory data analysis, setting up the evaluation framework, baseline model selection and hyperparameter tuning.

**Exploratory data analysis:**

**Data distribution:**

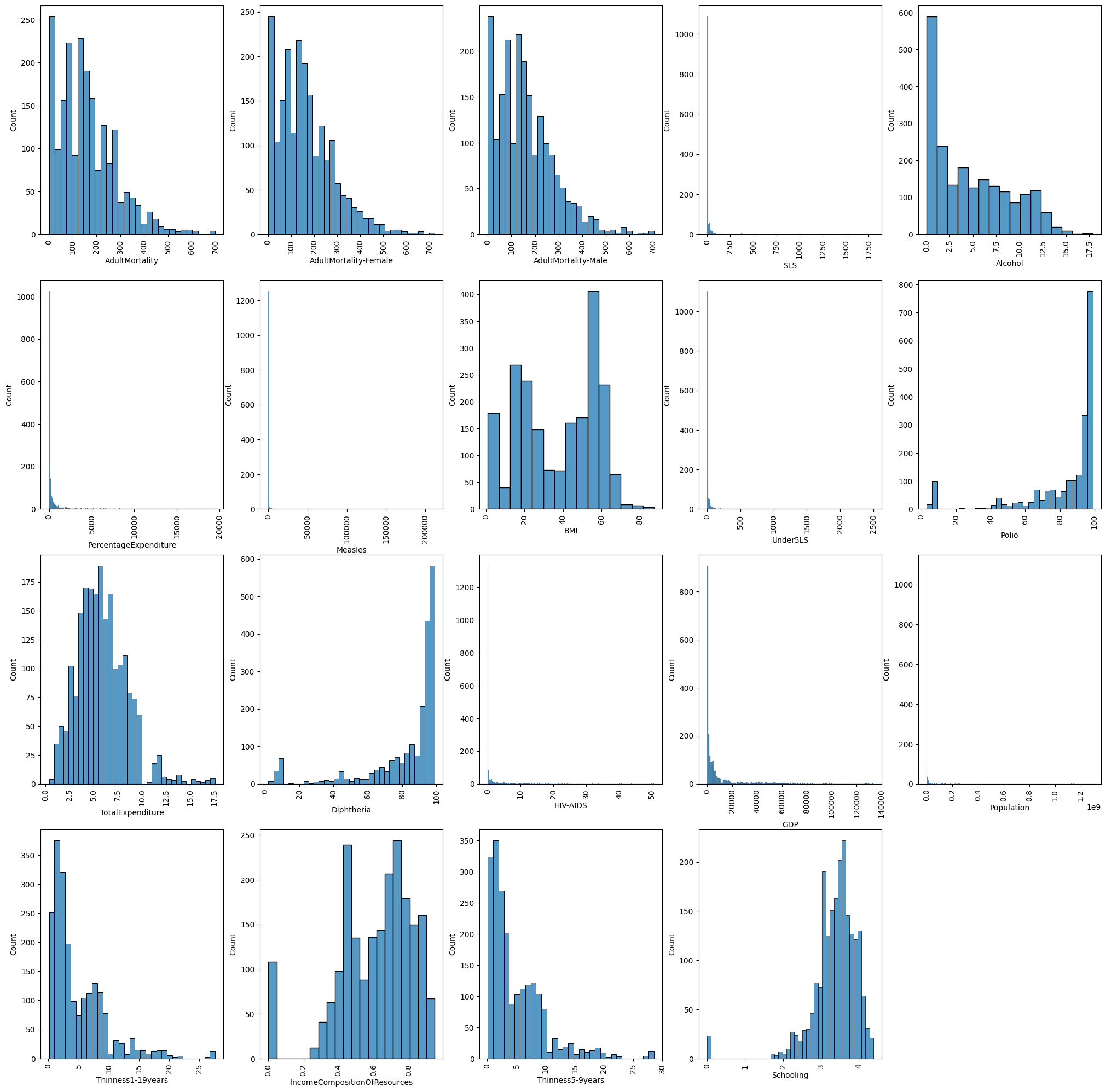


Figure 1. Data distribution of numerical columns

As it can be seen from Figure 1, the overall distribution of the columns is skewed. For instance, adult mortality columns are right skewed, and schooling, polio are left skewed. Some columns are extremely skewed such as percentage expenditures, GDP, and HIV-AIDS.

**Outliers:**

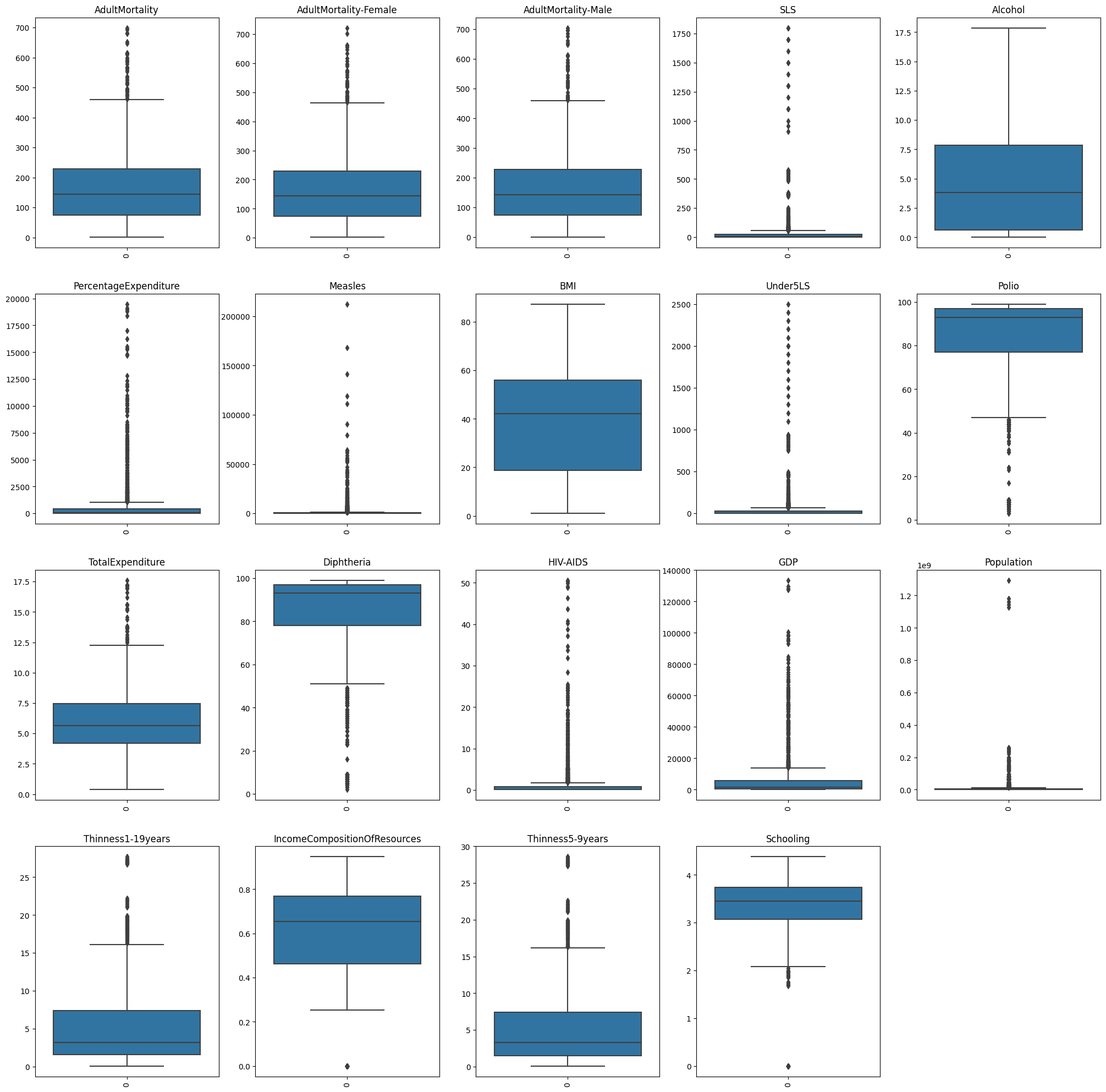


Figure 2. Box plot of numerical features

According to Tukey’s rule, any data point that is outside of the box in the boxplot is considered an outlier. By that rule, almost all features in the datasets have a great number of outliers. However, since the outliers are not largely differentiated from the rest of the data points, they will not be removed in this process. Another reason not to remove the outliers is to retain the amount of data for modeling purposes.

**Relationship with target variable:**

* Correlation:

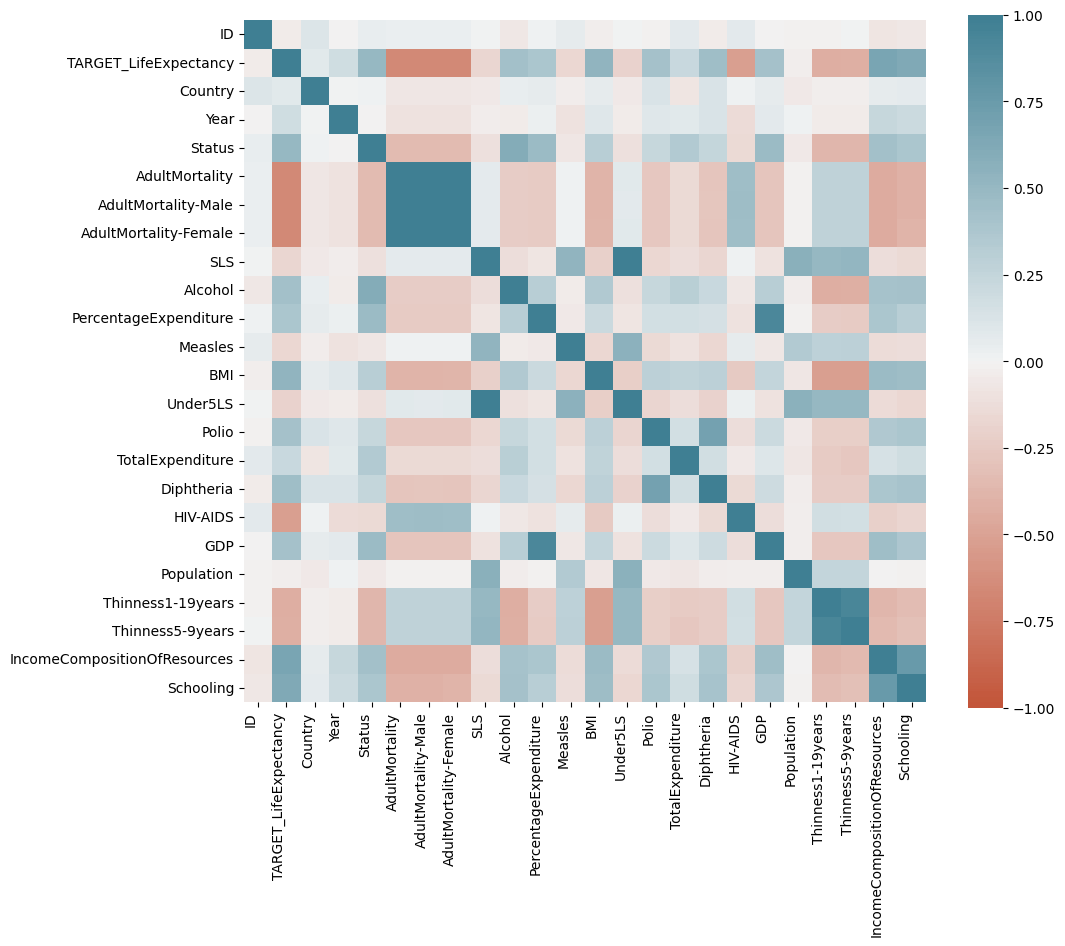


Figure 3. Correlation matrix of the data set

* Numerical variables:

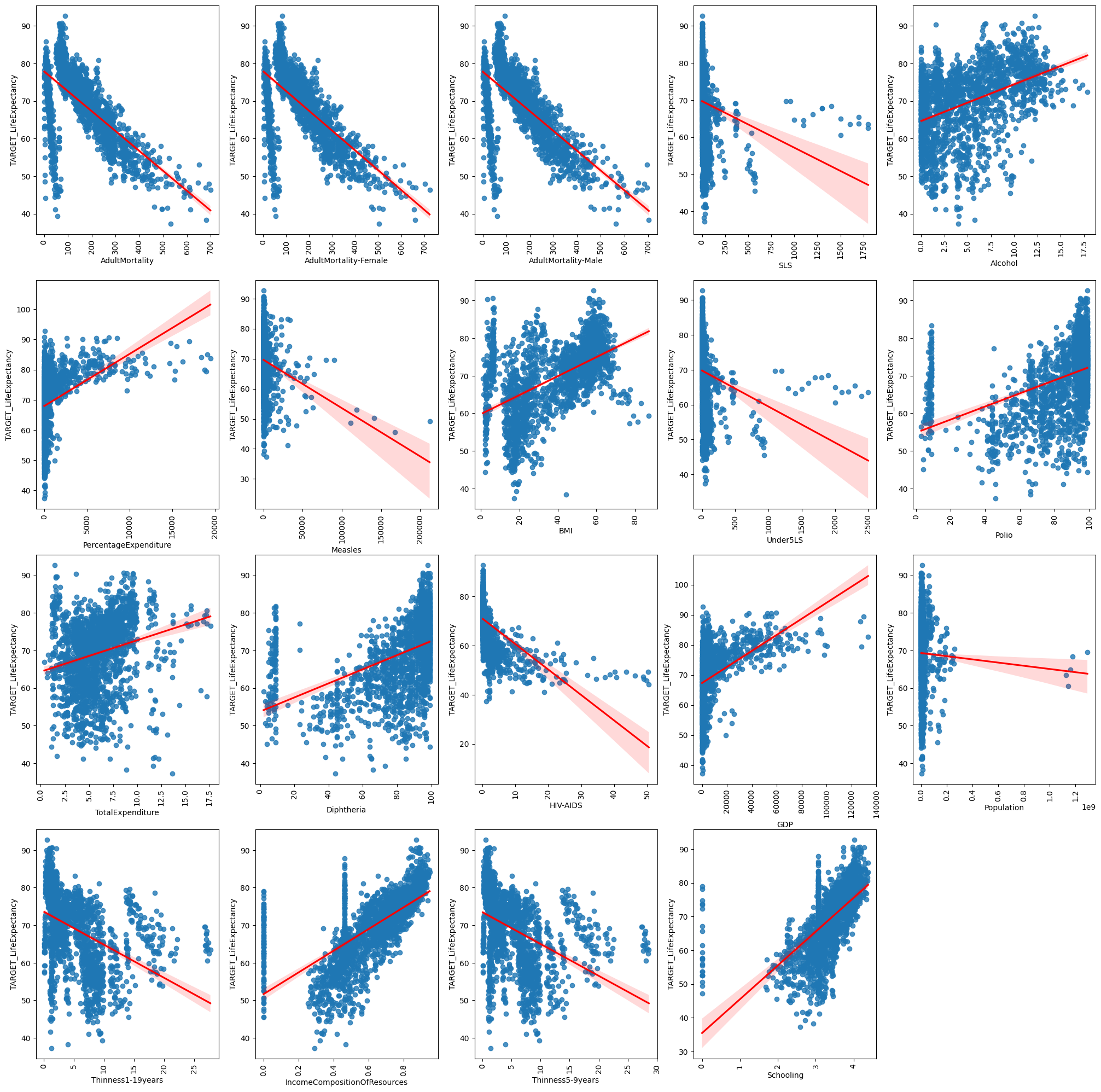


Figure 4. Scatter plot of numerical variables and target

* Categorical variables:

A graph with different colored lines

Description automatically generated

Figure 5. Bar plot of average life expectancy over the years

It can be seen that by plotting the target variable grouped by each year, the average life expectancy grows by each year, suggesting that the variable year has a direct correlation with the target.

A graph of a blue and orange bar

Description automatically generated

Figure 6. Bar plot of average life expectancy for each status

For the Status variable, 1 means that the country is a developing country, and 0 means that the country is a developed country. With the encoding defined, the graph shows that people from developing countries have an average of 10 years more life expectancy than people from developed countries.

**Evaluation framework:**

**Data splitting:**

At first glance, the dataset can be viewed as a time series dataset, due to the fact that the measurements of attributes are marked over intervals of time. However, since the task is to predict the test set which is in the same period as the training set, the training set will not be considered as time series data and therefore be split and validated normally. The dataset will be split into 3 sets, as below.

* Training set (60%): Used to pick the best modeling technique for the dataset.
* Validation set (20%): Used to pick the best version of the training model by tuning hyperparameters.
* Test set (20%): Used to test the final hypothesis.

After splitting, the data from each set must preserve similar distribution and have no intersections with each other.

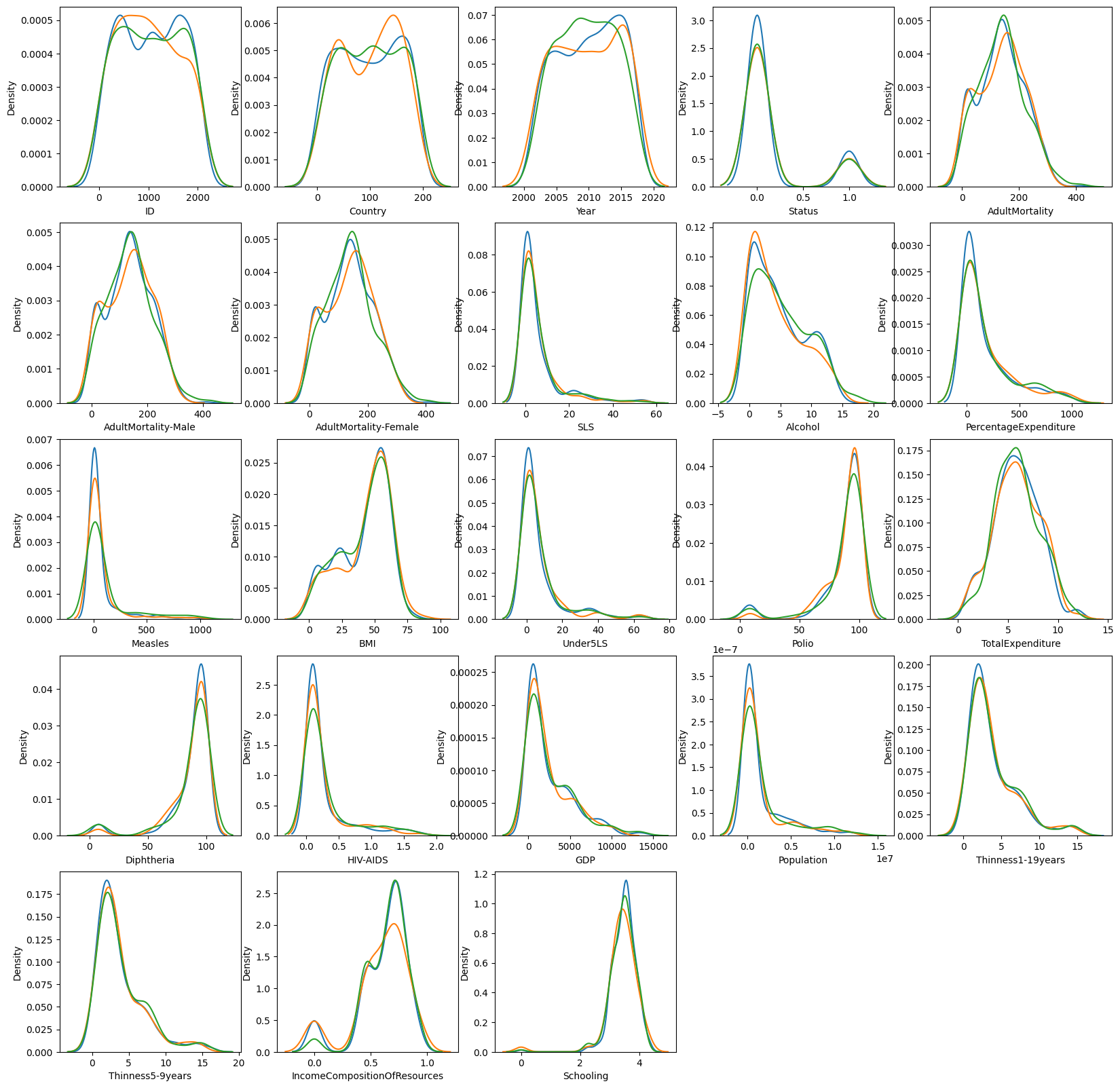
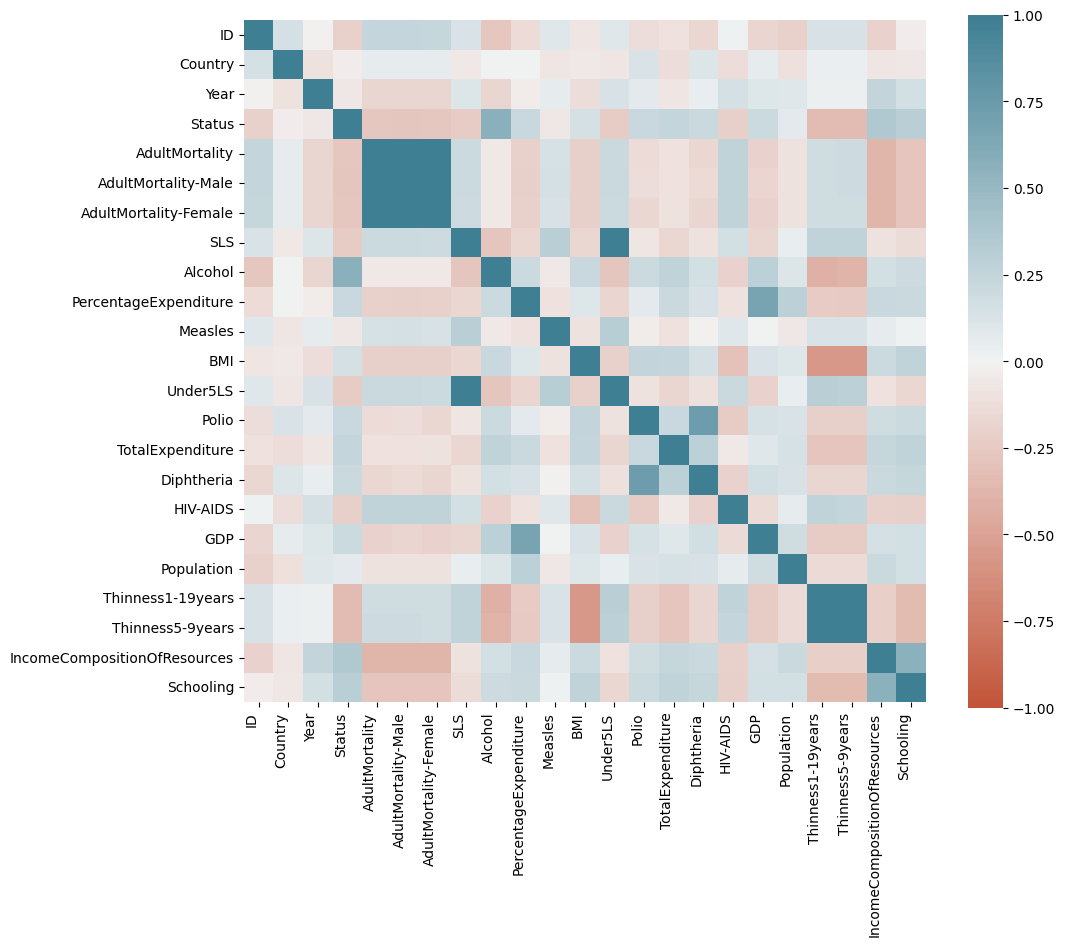
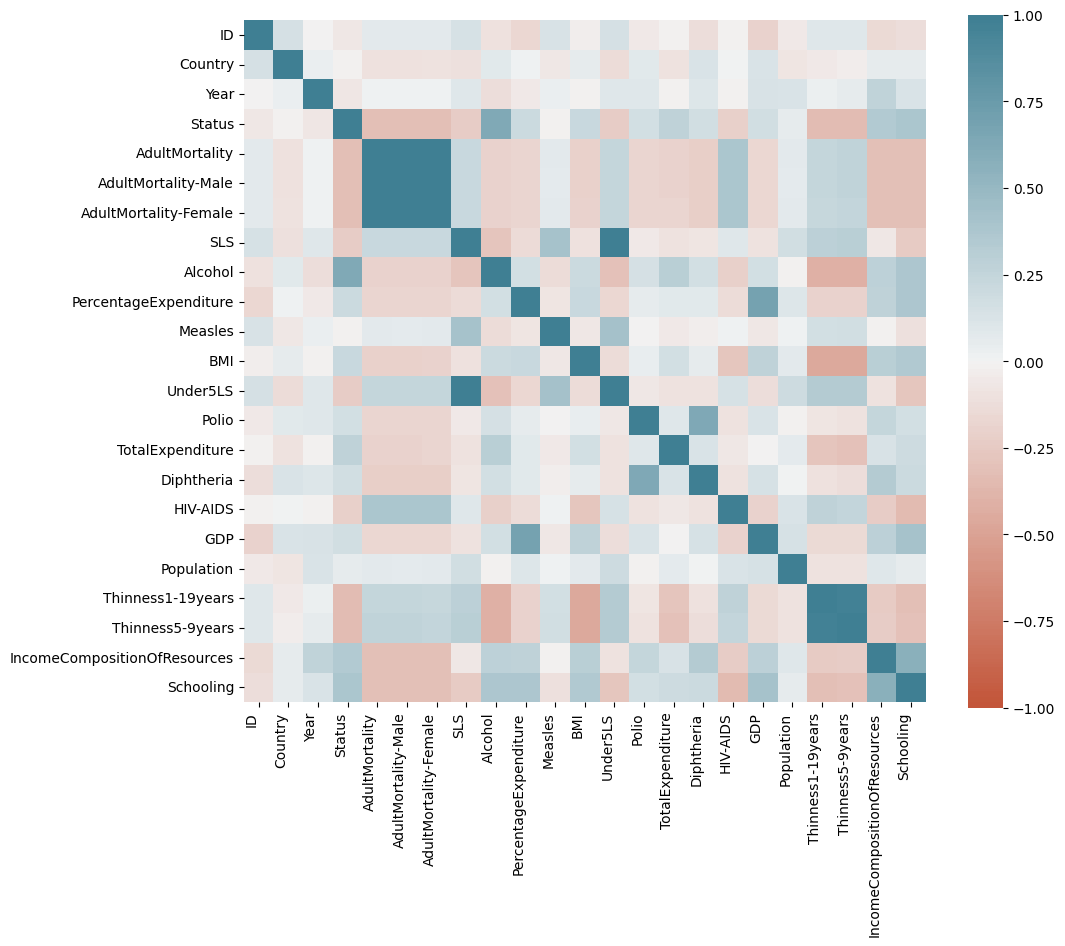


Figure 7. Distribution of numerical variables for 3 sets

A screenshot of a graph

Description automatically generated

Figure 8. Correlation matrix of train, validation, and test set, respectively.

Figure 8 demonstrates the distribution and the correlations for each set of training, validation, and testing. It shows that the distributions of numerical variables are not extremely different.

**Evaluation metrics:**

Since the task is a regression task, there is a possibility which the model might be overfit to the data. Therefore, a metric which measures the fit degree of the model is required, such as R2 , which will be the main evaluation metric.

However, there may be occasions where R2 is low, while substantial relations are in presence. Since the R2 metrics can mislead the development direction, we will use another metric to make sure of the regression results. Mean squared error can help with measuring the overall error of the current model, and give another perspective on the model’s fit.

**Data preprocessing:**

**Numerical data:**

* Scaling:

Since the data consists of plenty of outliers, using the Robust Scaler is useful in this case.

* Logarithm transformation:

Since a lot of the distributions are skewed, they need to be transformed back to Gaussian distribution. To do this, Box-Cox or Yeo-Johnson logarithm transformation can perform the job. However, since Box-Cox covers only positive data, Yeo-Johnson is used for this data set, on variables that have high correlation to the target variable. The results of scaling and transforming is presented in figure 9.

A collage of graphs

Description automatically generated

Figure 9. Results after scaling and transforming

**Categorical data:**

For all categorical data, because they do not have any hidden ranking or meaning within them, using One Hot Encoder to transform the data to binary will be adequate for numerical models.

**Baseline model selection:**

As shown in Figure 3 and Figure 4 , some variables have the ability to represent a linear correlation between themselves and the target variable, such as BMI, AdultMortality, and Schooling. Meanwhile, some variables show non-linear relationships with the target variable, such as GDP, HIV-AIDS, and PercentageExpenditure. Therefore, the candidate models include 2 linear models such as SGD regressor (Stochastic Gradient Descent) and Linear Regression (Ordinary Least Squares), and 1 polynomial regression model.

Since SGD regressor is better suited to datasets with more than 100,000 samples, comparisons will only be performed on linear regression and polynomial regression. To compare the 2 models, each model is set to perform with 5-fold cross validation on the training set, to see which model is better for unseen data.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Linear Regression | Polynomial Regression | Score type |
| 0 | -8.858969e+10 | -1.582043e+13 | r2 score |
| 1 | 7.798550e+13 | 4.400319e+14 | MSE |

Figure 10. Table of linear and polynomial regression scores

Since Linear Regression has a higher R2 score and a lower MSE, Linear Regression is chosen as the baseline model. Large numbers in both models are likely to be caused by multicollinearity, which is a scenario where multiple non-target variables correlate with each other. For example, AdultMortality, AdultMortality-Female, and AdultMortality-Male intercorrelates, etc., as shown in the figure below. To solve this problem, regularization is needed.

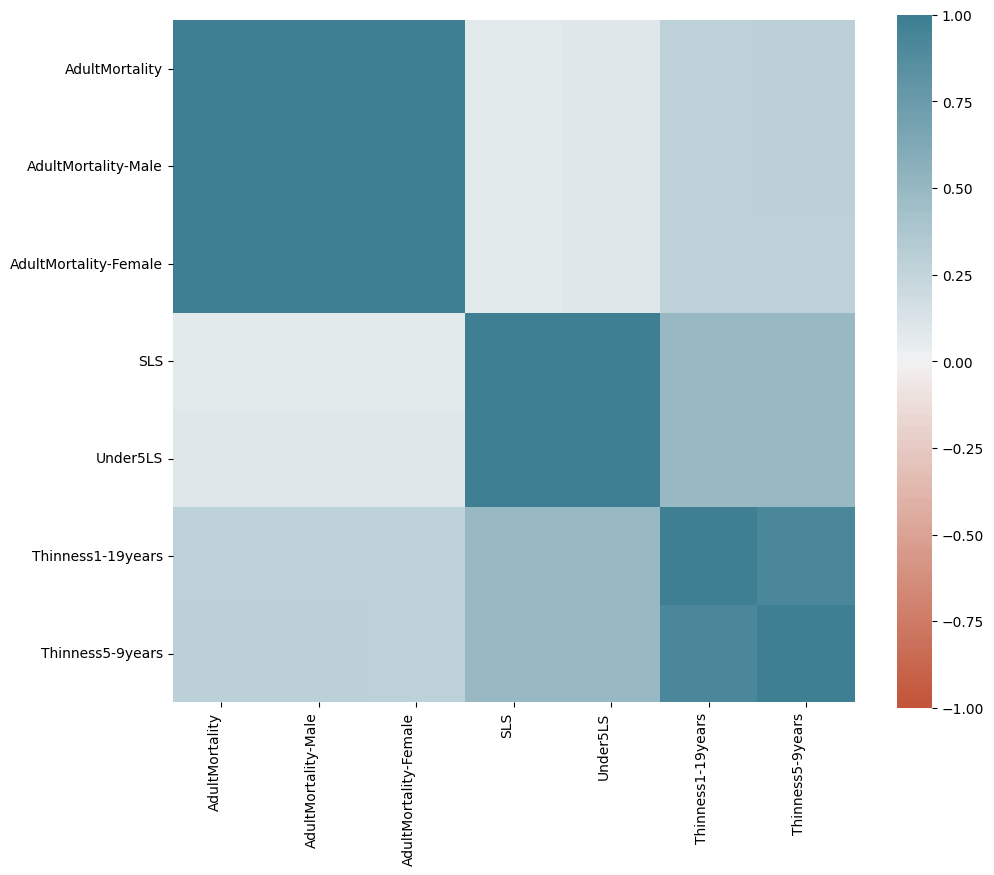


Figure 11. Correlation matrix of intercorrelated variables

Applying regularization:

To experiment with Ridge and Lasso regularization, a series of alpha values are tested on both models to see the trend of both performances.

A graph with a red line and blue line

Description automatically generated

Figure 12. Performance of Ridge and Lasso regression

The chart shows that Ridge regression outperformed Lasso in this dataset. Hence, Ridge regularization is chosen for hyperparameter tuning.

**Hyperparameter tuning:**

In this section, the alpha value of Ridge regression is brought upon for tuning and measured for performance. The goodness of the Ridge regression model is judged based on training data and on validation data to measure the generalization gap. The best value for alpha would mean the best validation score and lowest generalization gap.

A graph of a curve

Description automatically generated

Figure 13. Performance of Ridge validation and training with lambda values

As it is described in the graph, both the score of training and validation set gradually declines from the top and have very little gap between them. Therefore, it is safe to choose the value at the start of the range, alpha = 0.1.

Hypothesis testing:

A graph of performance by set

Description automatically generated

Figure 14. Performance of 3 sets

The bar chart shows that there are little to no differences in the testing set compared to other sets. With that being said, the hypothesis is considered accurate with respect to the dataset.

**References:**

https://wserver.crc.losrios.edu/~larsenl/ExtraMaterials/MisconceptionsR2.pdf