Introduction. In the aftermath of COVID-19, countries are striving to improve their life expectancy rates with limited resources. The challenge is to identify the most influential factors to prioritize government efforts using regression models. While previous studies have examined demographic variables, income composition, and mortality rates, they have overlooked the impact of immunization and the human development index. This study seeks to fill this gap by examining essential immunizations such as Hepatitis B, Polio, and Diphtheria, as well as mortality, economic, social, and health-related factors. By observing 183 countries, this study identifies factors that contribute to lower life expectancy and recommends areas for improvement.

Aim. The objective of this study is to identify the factors that have the greatest impact on life expectancy. By doing so, it can serve as a valuable reference for governments around the world in their efforts to improve the life expectancy of their citizens.

Data Exploration. This research is based on the Kaggle dataset {*Kumar, R. (2018). Life Expectancy (WHO). Retrieved February 23, 2023 from https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who*}. We only focus on the data collected in 2015. Our target response is *life expectancy* in age and 14 other standardised explanatory variables.

The heatmap of the correlation between each pair of explanatory variables is presented at Figure 1a. It can be seen that 'Measles' and 'Infant. deaths, 'Diphtheria' and 'Hepatitis. B', 'thinness.5.9.years' and 'thinness.1.19.years', 'Schooling' and 'Income.composition.of.resources' are highly correlated. We could therefore expect them to have similar parameters in the regression model.

Model Fitting. I employed three distinct regression techniques, namely multiple linear regression, lasso regression, and principal component regression, and tested their efficacy. Initially, I partitioned the complete data set into two halves, i.e., training and test sets. I then trained each model on the training set and evaluated its performance on the test set using mean square error as a metric to assess its ability to predict unseen life expectancy.

Multiple Linear Regression. By performing back deletion, I sequentially find out all the significant variables in the prediction of life expectancy, the relevant information is presented here.

```
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      70.9829     0.2885 246.007  < 2e-16 ***
Adult.Mortality                  -2.7656     0.4444  -6.223 6.58e-08 ***
Hepatitis.B                       2.1426     0.6354   3.372  0.00136 **
Polio                            -0.9520     0.3671  -2.594  0.01210 *
Diphtheria                       -1.4332     0.7262  -1.974  0.05338 .
Population                       -0.6760     0.2342  -2.887  0.00552 **
Income.composition.of.resources   5.8507     0.4339  13.485  < 2e-16 ***
---
F-statistic: 106.3 on 6 and 56 DF,  p-value: < 2.2e-16
```

By checking the F-statistics, the result is significant, which means that the model with fewer variables is preferred. The MSE on the test set is recorded as 7.825.

Lasso Regression. This method could be very suitable in our case because we are having 14 distinct predictors and from Figure 1a, we can see a few features are correlated to others, we might be able to drop some of those. We confirm it by using Lasso regression. 10-fold cross-validation is used to search optimal penalty term λ across 100 values from 1e-2 to 1e10. The optimal λ reported is 0.32, indicating that the model has effectively shrunk some of the coefficients toward zero. The remaining parameters can be seen as the most important predictors of life expectancy. The MSE on the test set is recorded as 6.108.

Principal Component Regression. There is another approach to eliminate unnecessary predictors, which involves calculating the principal components of the feature sets. As shown in Figure 1b, we can observe that the first 9 principal components account for approximately 96% of the variation. To ensure that this percentage of variance explained is sufficient to explain the response, we refer to the validation plot. The plot indicates that at n=9, the root-mean-square error of prediction (RMSEP) is quite low, implying that 9 components are enough to achieve a good response. The MSE is measured at 6.534, which is lower than the MSE obtained by linear regression. This indicates that the last 5 unused components are indeed redundant.

Model selection and validation. After comparing the mean square error (MSE) of the three models, I claim that Lasso regression provides the best fit and captures the most response. However, before proceeding, we must verify that our Lasso model follows the assumptions of linear regression. Figure 2 presents four diagnostic plots of the residuals, which we will examine one by one.

Firstly, figure 2a plots the residuals against the fitted values. The scatters are randomly spread across the plot and show no apparent pattern. The orange LOESS curve is relatively flat and close to the dashed line at zero, indicating that the linear relationship assumption is reasonable, and our residuals have a mean of zero. The plot also indicates the existence of a few potential outliers, which are far from the zero line. Their influence to our model will be examined later.

Figure 2b examines whether our residuals follow a normal distribution. The plot shows that our residuals are mostly in a straight line. Some deviations are near the ends, but they are generally small. This is a solid result suggesting that our residuals are normally distributed.

Figure 2c shows the standardised residuals are scattered randomly against fitted values without any obvious pattern. The orange LOESS curve again shows the trend. We can tell except at the far two ends where a few points pull it down, the line is relatively flat. It shows our residuals are homoscedasticity which means constant variance.

Figure 2d displays the standardised residuals against leverage. The residuals are centred around zero and read 2-3 standard deviations away from zero, and approximately symmetrically about zero. It follows the assumption of normal distribution. Although the LOESS line is a bit curved, and there are a few points with relatively large leverage which means they can be influential, there is no point whose cook's distance is larger than 0.5 being highlighted. It means our regression will not change significantly if those high leverage points are removed. Thus we could leave them there and stick with our current model.

In summary, all assumptions of our Lasso model have been met: residuals are linear, have a mean of zero and constant variance, and the model is barely impacted by the removal of some points. Thus, we can confirm that our Lasso model is valid for use.

Model Interpretation. Hereby I present the parameters obtained by Lasso regression:

```
(Intercept)  Adult.Mortality infant.deaths Hepatitis.B  Measles BMI Polio Diphtheria   HIV.AIDS   GDP  Population
 70.95397       -1.828603    -0.5396649     0.3115052     0      0    0       0       -0.8708945   0  -0.03672227
thinness..1.19.years    thinness.5.9.years      Income.composition.of.resources   Schooling
          0                    -0.1130365                     4.99845476               0.02940733
```

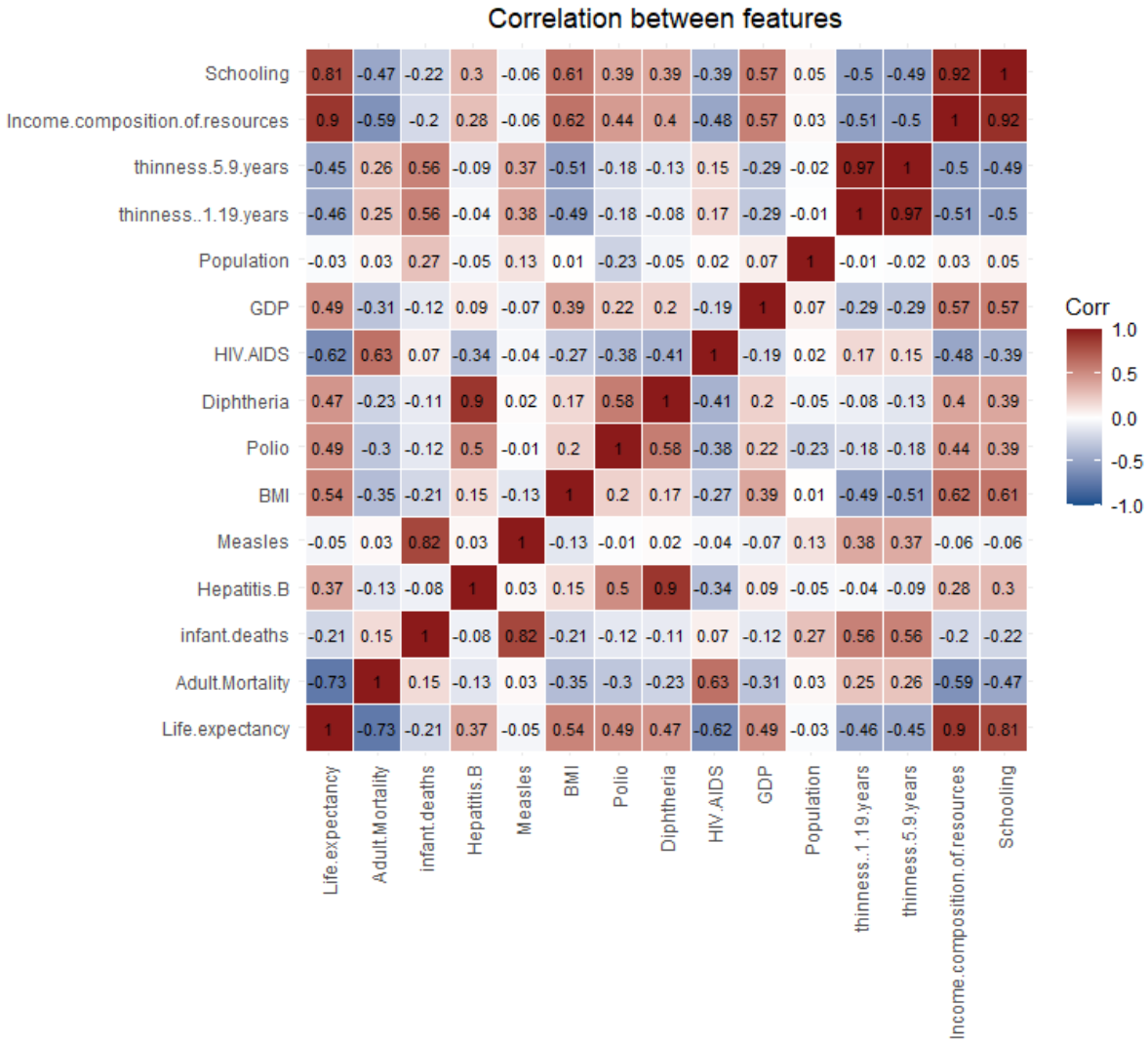
To begin, the intercept of 70.95 is the expected value of life expectancy when all predictors are equal to zero. It represents the *average life expectancy* of the studied 183 countries in 2015.

From the results, we can see that the *Adult Mortality Rate*, *number of infant deaths per 1000*, *death per 1000 HIV.AIDS patients*, *Prevalence of thinness among children for Ages 5 to 9* have significant negative coefficients. This suggests that life expectancy is expected to decrease as these features increase. *Population* also has a negative coefficient but it is rather small, suggesting it is less influential. On the other hand, features like *Hepatitis B immunization*, *Human Index of Income Composition of Resources* and *Schooling* have positive coefficients, indicating that an increase in these predictors would lead to an increase in life expectancy.

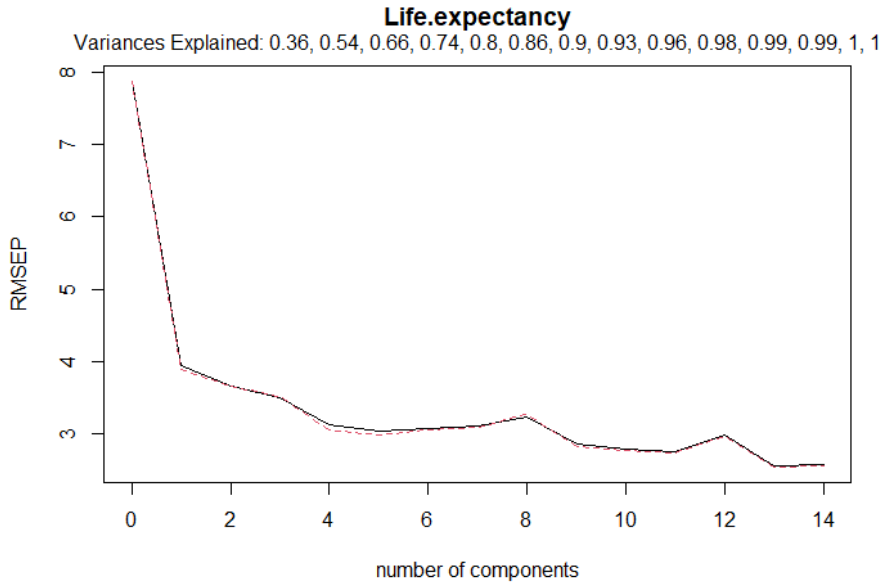
Although some features are set to zero in this model, it does not necessarily mean they are unrelated to life expectancy. This could be due to their high correlation with other features or minimal effect on life expectancy. To explore this further, we can refer to the correlation map in Figure 1a. The correlation between the *number of Measles per 1000 population* and *number of infant deaths per 1000* is 0.82, indicating a strong positive association between the two variables. This suggests that the negative impact of measles on life expectancy is similar to that of infant deaths. Similarly, *Diphtheria tetanus toxoid and pertussis immunization* is 90% correlated to *Hepatitis B immunization*, which means the more widespread the availability of vaccines for these two diseases, the greater the positive impact on life expectancy. Additionally, *Prevalence of thinness among children for Ages 10 to 19* is highly positively correlated with the prevalence for Ages 5 to 9, indicating it also has a negative impact on life expectancy.

Lastly, the features *Polio immunization* and *GDP* are set to zero in the model, indicating that they are less influential to life expectancy. However, they do have some positive impact, as suggested by their correlation with features like *Hepatitis. B* and *Income Composition of Resources.*, etc.

Guidance to government. The results of the analysis can provide the government with a clear direction to prioritize areas that can improve life expectancy. The first priority should be given to increasing the human index of income composition of resources by effectively utilizing available resources. Secondly, investing in healthcare and policing can reduce the adult mortality rate, infant deaths, and the spread of measles, which have significantly positive impact on life expectancy. Furthermore, it is beneficial to promote vaccination programs for Hepatitis B and Diphtheria. Additionally, providing higher levels of education and nutritional supplementation for adolescents can help to reduce thinness and support youth growth. It is worth noting that a higher GDP does not necessarily lead to higher average life expectancy due to factors such as income inequality. The 3D plot shown in Figure 3 illustrates the relationship between life expectancy and the two most influential features: adult mortality, and income composition of resources. The trend obeys our inferred parameters. Countries such as Cyprus, Chile, and Australia with high life expectancy and income composition, and low adult mortality rates, can serve as good examples of policies and management. Meanwhile, more attention and aid should be given to countries like Central African Republic, Sierra Leone, and Lesotho to support their development.

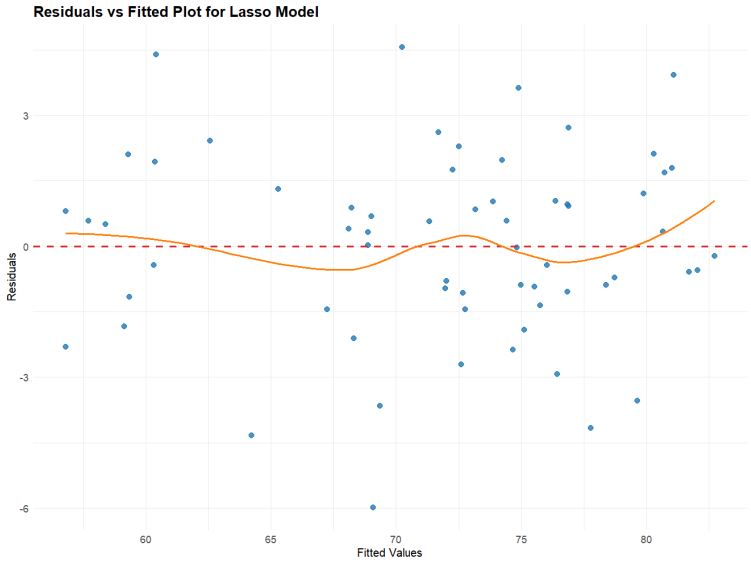


(a) Correlation map between explanatory variables

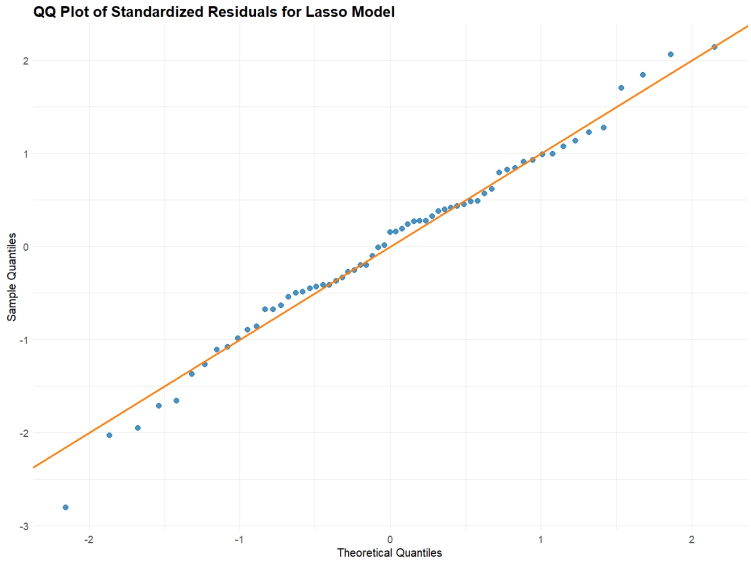


(b) Variance/Response Explained by components

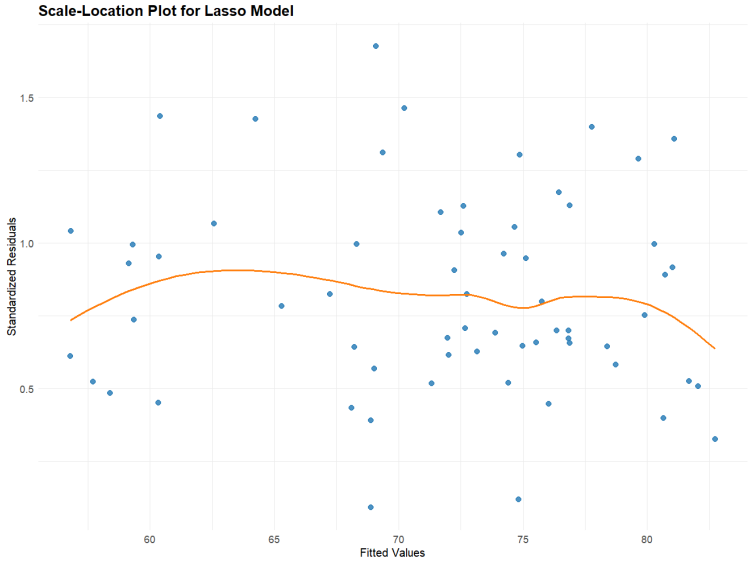
Figure 1: Features/Components Visualisation



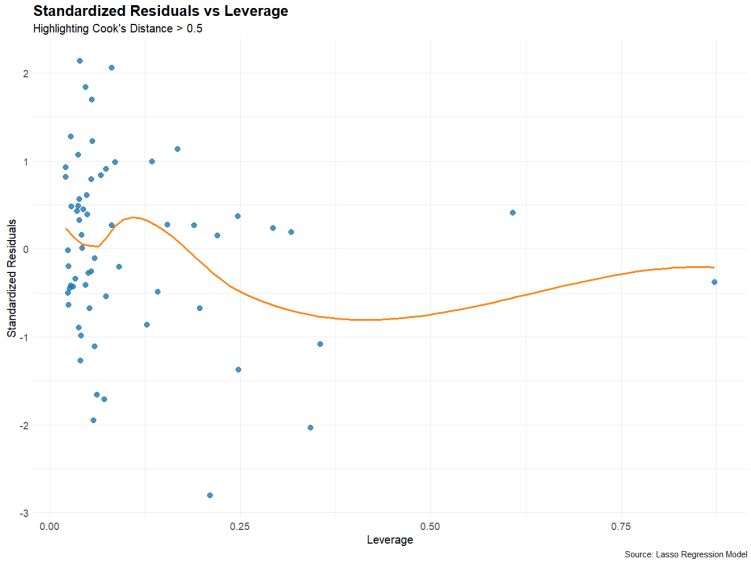
(a)



(b)



(c)



(d)

Figure 2: Diagnostic Plots of Lasso Model

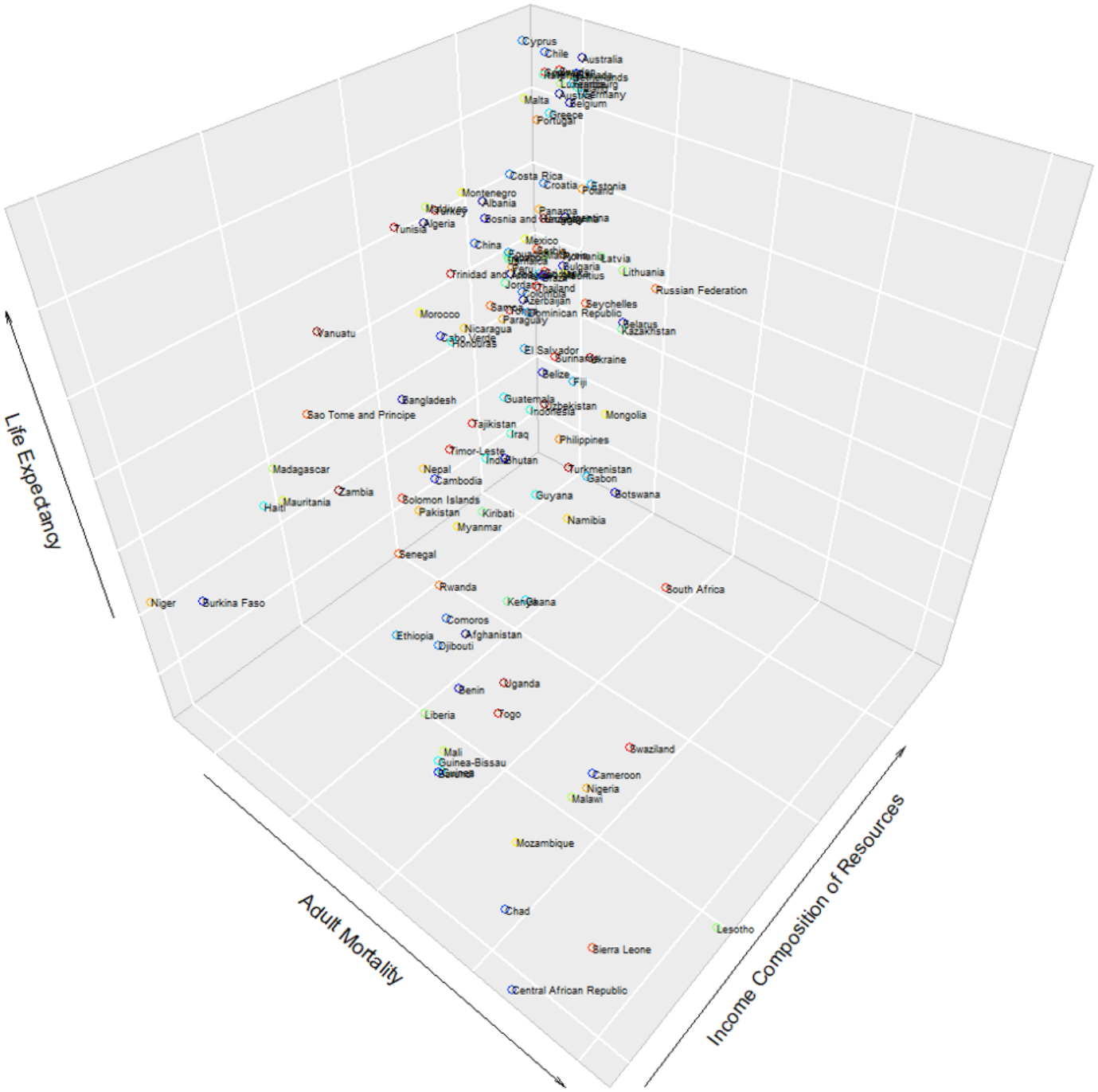


Figure 3: Life Expectancy versus two most important features