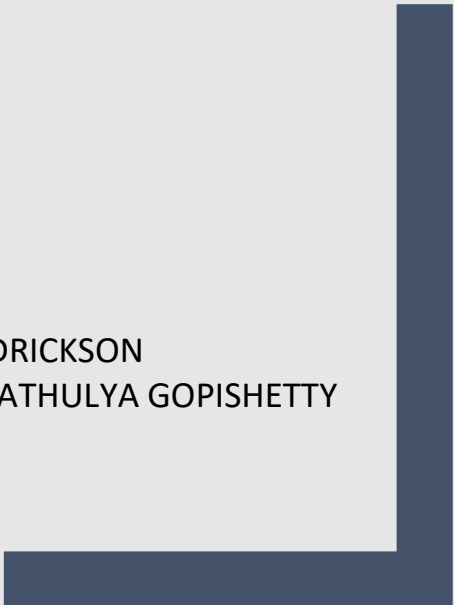




APPLIED STATISTICS AND VISUALIZATION FOR ANALYTICS: FINAL PROJECT

LIFE EXPECTANCY

BY GROUP 2:
AUSTEN HENDRICKSON
VENKATA SRI ATHULYA GOPISHETTY
WEI WANG



Abstract

Reasonable predictions of life expectancy are of great significance to social progress and economic development. There have been many studies undertaken in the past attempting to see which factors seem to affect life expectancy by considering demographic variables such as, income composition, and mortality rates. But this time we will add the effect of immunization and human development index which has not always taken into account in the past. It can not only help the government formulate policies based on population size and macro-control of population, but also improve the health of residents and promote the development of public health. This study empirically analyzes the influencing factors on life expectancy, establishes different models that predict the life expectancy of the population, and studies the influence of those factors on the model. From this, the main factors that affect the life expectancy of a population are analyzed. After data analysis, it was found that the best model for life expectancy prediction was formed from the following factors, Adult Mortality, Infant deaths, Polio, Diphtheria, Income, and Schooling, while Alcohol, Measles, etc. appear to have no significant impact on life expectancy.

The Dataset

Source: The original data for this study came from the life expectancy database in the Kaggle dataset. The data was collected from the World Health Organization (WHO) and the United Nations website with the help of Deeksha Russell and Duan Wang. Within the dataset is data related to health factors that were collected from the WHO database, and the economic data within the dataset was collected from the United Nations website. The accuracy of the study depends on the availability and accuracy of the data. Since the data used in this article is collected from the WHO and UN, the data is presumed to be accurate and reliable.

The data set contains data for 193 countries and regions from 2000 to 2015, consisting of 22 columns and 2938 rows, which means that our data includes 21 predictors (Appendix A). The 21 predictors are split into several categories: immune-related factors, mortality factors, economic factors and social factors. These factors are what were used to analyze life expectancy. Unfortunately, there was some missing data within the dataset. The missing data values were distributed in the fields of population, hepatitis B and GDP, and a few countries such as Vanuatu, Tonga, Togo, and Cape Verde were missing practically every value. The goal of this project was to try and find an appropriate model to predict life expectancy by analyzing the existing data and determine the importance of the predictor variables.

Data Exploration and Preprocessing

The first thing that was done was examining the data for any missing values or values that seemed weird. Right away it was noticed that there was a fair amount of missing values in the dataset, so a closer look was taken at each column. The first column that was looked at was life expectancy because this was the column that was of most interest and because it is what was trying to be predicted based on the other variables. It was found that there were ten missing values in life expectancy, all of which were individual countries with one row of data and missing data from several other columns as well. Due to this it was decided that these ten rows would just be deleted. Additionally, a boxplot was created, as seen in appendix B, that showed a few outliers in life expectancy. These outliers were significantly different from the rest of the data and so were removed from the dataset. Next, the remaining columns were looked at and instead of deleting missing values they were filled in with the mean value of each country for that particular variable.

After the missing values were taken care of, a closer look was taken at some values that didn't make sense. It was discovered that not all the countries were correctly labeled as developed or developing. For instance, Canada was labeled as developing, so each country was examined and then verified whether or not it was correctly classified. After this the column Hepatitis B was looked at since it had a lot of rows with the value zero. At first, it was thought that this was a mistake but upon further research it was learned that many countries didn't start using Hepatitis B vaccines for several years and so these values were left alone. Finally, the status column was converted into binary values with 1 for developed and 0 for developing so that they could be included in the analysis.

After the initial preprocessing a correlation plot was created and can be seen in Appendix C. This plot uses pie charts to show which variables are the most correlated. Even though the pie charts showed which variables were most correlated they did not show the distribution and so another correlation plot was created as seen below in order to figure out which variables appeared to be related to life expectancy and establish a starting point for the analysis. In the correlation chart it can be seen that there are many linear relationships both strong and weak as well as some nonlinear relationships. The strong linear relationships provided the best variables to start exploring with. Summary statistics after initial data pre-processing can be seen in Appendix B.

Life Expectancy: Selected Variables

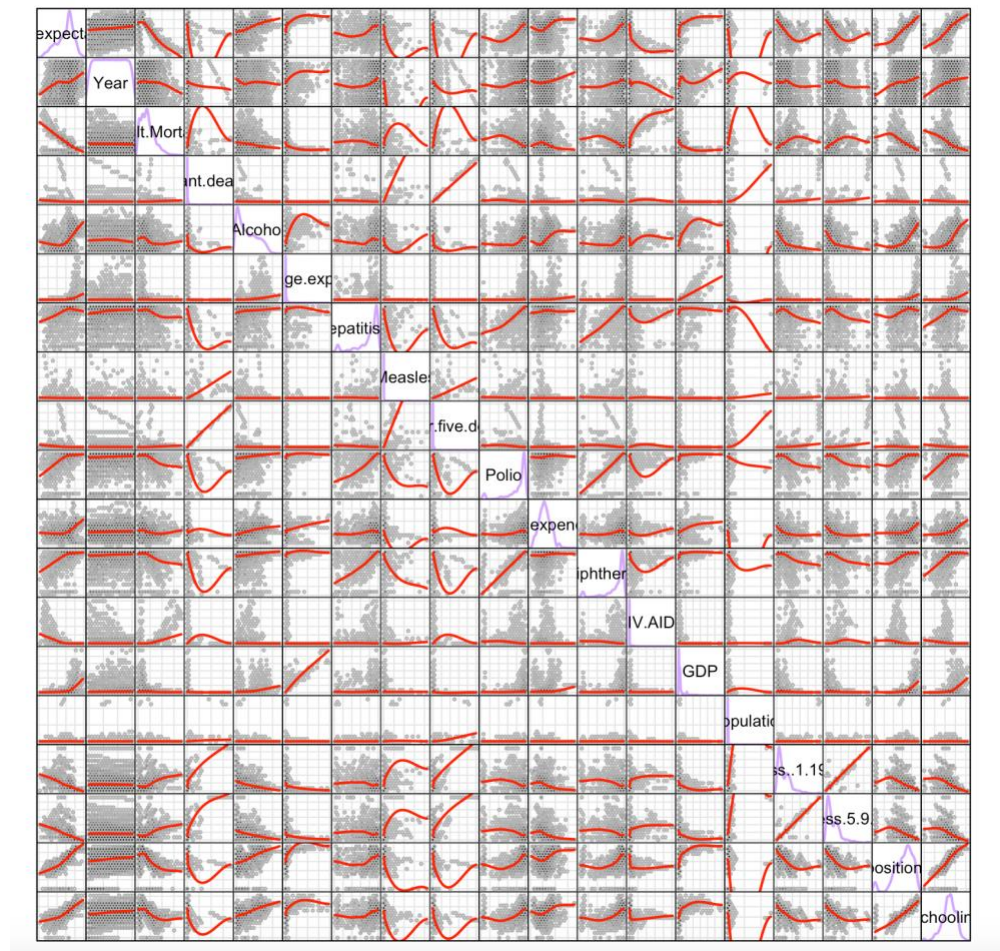


Figure 1. Correlation plot for life expectancy after initial preprocessing

Predictive Models

Linear Regression

Through the correlation graph, it was found that many columns in the dataset had a linear relationship with life expectancy, such as schooling and income composition of resources, however it was also apparent that some of the linear correlations were not very strong. Before fitting them, it was important to figure out which explanatory variables were the most useful for life expectancy prediction. So the decision was made to divide the data between a training dataset and test dataset. Random sampling split 80% of the data into the training set with 20% in the testing set. These techniques are primarily used to avoid overfitting to the data and avoid building overly complicated models. If the data was not split, the model would be fit to all of the data and any

measurement of accuracy would not be useful for comparison with new data. The same split was used for the other prediction models discussed later on.

According to the initial analysis of factors that appeared to have an influence on life expectancy it was assumed that the theoretical predictive model should be a linear model, containing all the predictors in the existing dataset. Using the summary function for the `lm.all` model in Appendix D, the P-value for each explanatory variable was checked since linear regression analysis is a form of inferential statistics. The p-values helped determine whether the relationships that were observed in the sample also existed in the larger population. The p-value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable. If there is no correlation, there is no association between the changes in the independent variable and the shifts in the dependent variable. If the p-value for a variable is less than a significance level of .05, then the sample data provides enough evidence to reject the null hypothesis for the entire population.

From the `lm.all` model the explanatory variables: Status, AdultMortality, infant.deaths, under.five.deaths, Polio, Diphtheria, HIV.AIDS, Income.composition.of.resources, and Schooling all had p-values less than 0.05. This means the null hypothesis can be rejected and that there is a significant linear association between the explanatory variables and the response variable. In addition, the adjusted R-squared for the `lm.all` model is 0.8152, which means that the model explains about 81.52% the variability of the response data around its mean. This seemed like a good starting point and that the model could be improved by comparing linear models based on the significant variables using the Adjusted R-squared.

When creating the second linear model `lm.fit1`, the plan was to reduce the number of explanatory variables by removing the ones that had no significant linear association with life expectancy. But the results generated were about the same as before and can be seen in Appendix D. The adjusted R-squared is 0.8094. Realizing that the most effective variables for predicting life expectancy had been chosen, and wanting to improve the accuracy and fit of the model for the training data, a third model `lm.fit2` was created. The `stat_smooth` function was chosen for fitting the relationship between life expectancy and the important variables. In order to view the smoothing and changes to the model `ggplot` was used and can be seen in Appendix E. Using `ggplot` It can be seen that there are four variables that could be fitted better. When using the smoothing model, there are 10 different smoothing lines to choose from. The goal was to pick a smoothing line that was closer to a linear model and less wiggly. When fitting the model, it can be easy to overfit it and so to avoid this, fitting would be stopped when there were no longer significant changes to the adjusted R-squared value and when the plot was similar to a linear model. After fitting the HIV.AIDS with the smooth function `ns(x,5)`, Adult.Mortality with `ns(x,3)`, and Income.composition.of.resources with `ns(x,4)`, the Adjusted R-squared increased significantly for the final training model in Appendix D. (0.8814) At the same time it was found that the standard error of the model was 3.319 and represents the average distance that the observed values fall from the regression line. This means that the regression model on average deviates from the actual values in the training data set by 3.319. Even though it is lower than the initial model `lm.all`, to further evaluate the model, summary diagnostic plots were run but did not indicate any assumption had been violated.

Diagnostic Plot Analysis for linear regression

Residuals vs Fitted

For the Residuals vs Fitted in order to not violate the linear model assumption, it is expected to see equally spread residuals around a horizontal line without distinct patterns (for example no steep line) and that the horizontal line is around 0. This would mean that there could be a linear relationship between predictor variables and an outcome variable. Compared with the initial output it can be seen that after fitting the model below that the line is fairly flat except for the highest values. In addition, the point distribution is more uniform than before. There is a significant improvement than shown in the unaltered data.

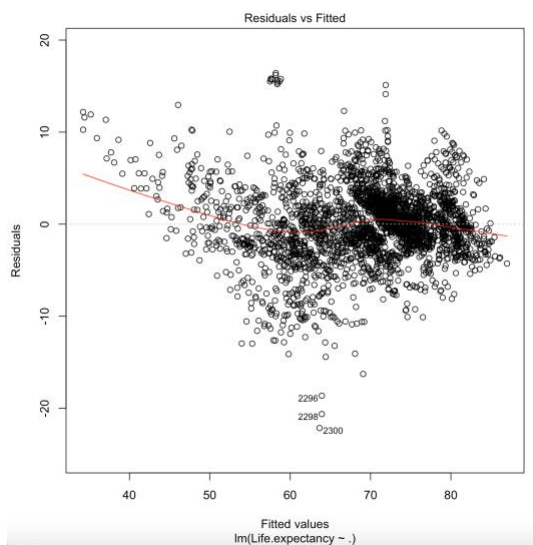


Figure2. Residuals vs Fitted (Initial)

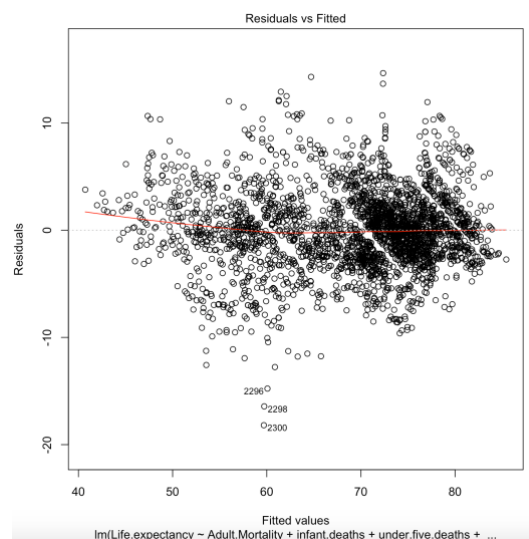


Figure3. Residuals vs Fitted (Final)

Normal Q-Q

A normal Q-Q plot shows if residuals are normally distributed. It is created by plotting two sets of quantiles against one another. It is expected to see the residuals closely fitted to the dashed line. If the residuals are not closely fitted to the line it means that they are not normally distributed and that the model is not a good fit. When comparing the two figures below it is apparent that there was some skewness in the initial data, which has been slightly improved but not entirely eliminated through the fitting of the response variable and removal of outliers.

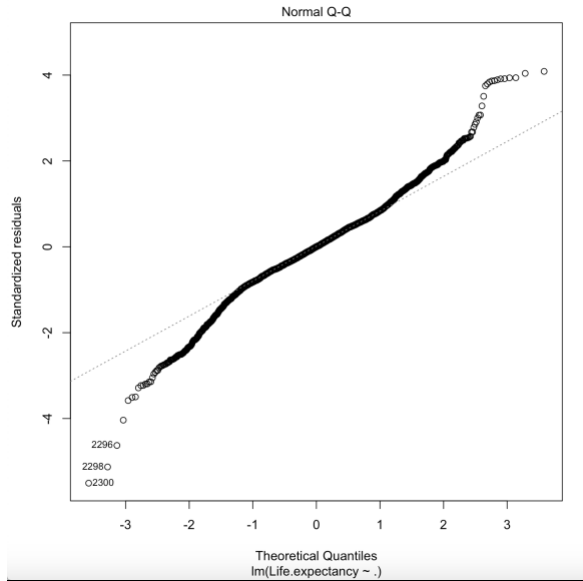


Figure4. Normal Q-Q (Initial)

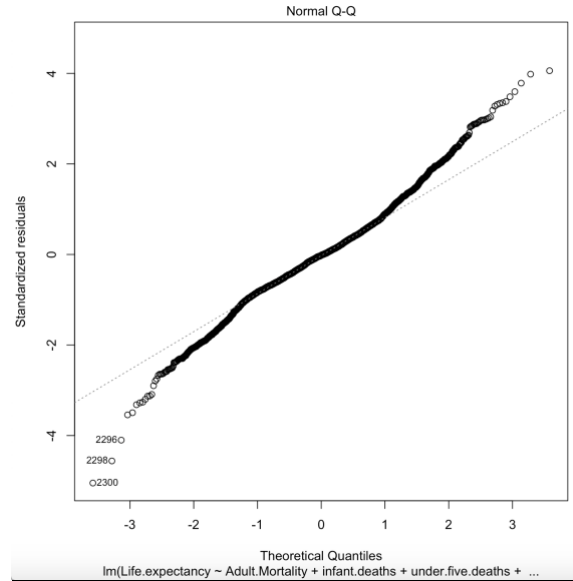


Figure5. Normal Q-Q (Final)

Scale Location

For scale location, it could show if residuals are spread equally along the ranges of predictors. This is how to check the assumption of equal variance (homoscedasticity). It's good if a horizontal line with equally (randomly) spread points can be observed. According to the figures below, it can be seen that after fitting the model, the distribution of residuals is more uniform. But the variance generated by the data is not constant, and most of them are concentrated in a higher range. Although there is significant improvement, it is still not perfect.

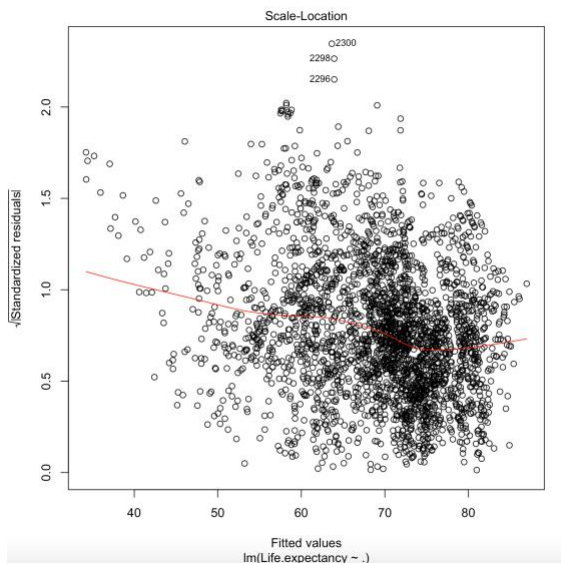


Figure6. Scale Location (Initial)

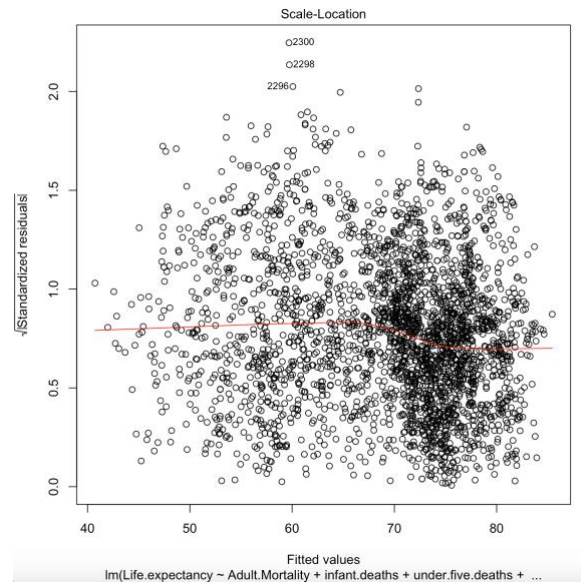


Figure7. Scale Location (Final)

Residuals vs Leverage

Residuals vs Leverage plot is helpful for finding influential cases (i.e., subjects) if any. From the figures below Cook's distance lines (a red dashed line) can barely be seen because all cases are well inside of Cook's distance lines. That means that there are no outliers influential to the regression results. In addition to not wanting points outside of the Cook's distance lines it is expected to see a straight line centered around zero with constant variability around it. This is the case for both models with the fitted model being a slightly better representation.

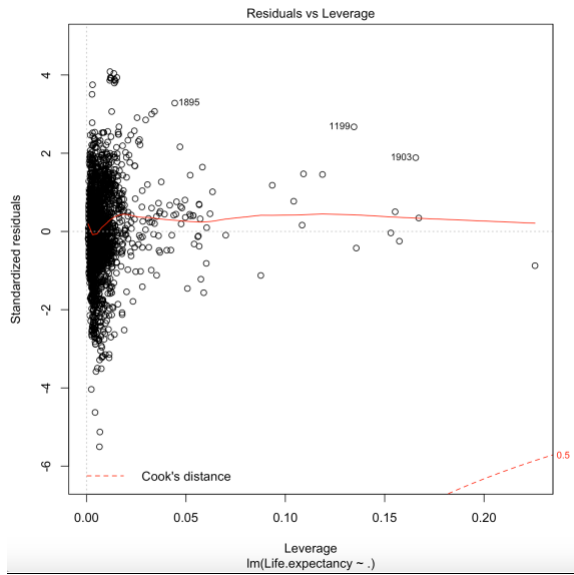


Figure 8. *Residuals vs Leverage (Initial)*

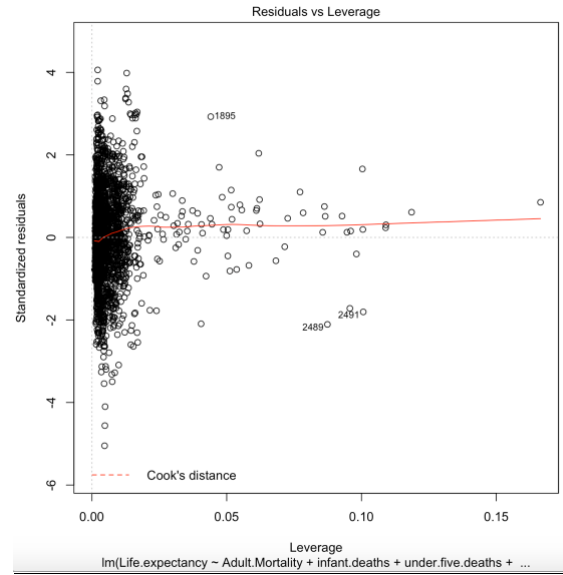


Figure 9. *Residuals vs Leverage (Final)*

Finally, the mean squared error (MSE) was calculated for the test dataset (586 records) to assess the quality of the lm.fit2 model as a predictor using the formula below.

$$MSE = \frac{1}{n} \sum \left(y - \hat{y} \right)^2$$

The square of the difference
between actual and
predicted

The result is 3.172522. This value will be compared with other models for determining the best model for the data.

Decision Tree

In addition to linear modeling a decision tree was also used for predictive modeling because they work for both categorical and numerical data. For the decision tree modeling two methods were chosen, a decision tree analyzing categorical variables and one for analyzing continuous variables as well.

Categorical prediction

For the first part of the analysis, a new column was created named Label. The label column was then split based on quartile range, Q1 comes from the 0-25% quartile range, Q2 is the 25-50% quartile range, Q3 is the 50% - 75% quartile range and Q4 is on 75% - 100% quartile range. Countries with low life expectancy were deemed to be values below 61.92 and were tagged as Q1. Q2 for the life expectancy values between 61.92 and 70.00, Q3 for the life expectancy values between 70.00 and 74.80, and Q4 for life expectancy values greater than 74.80 which are considered to be long life expectancy values.

The decision tree model was prepared using the trained data, later labels were predicted by the model using the test data. The accuracy obtained on categorical target prediction was 79.5%. The root node starts with comparing Income composition of resources and gets down to the terminal nodes Q1, Q2, Q3, Q4 by passing on through HIV.AIDS, Income composition of resources, Adult mortality, Percentage expenditure, Infant deaths, GDP, and Percentage.expenditure. However, it is clear from the decision tree that income composition of resources and HIV.AIDS indeed plays a major role in decision making regarding life expectancy.

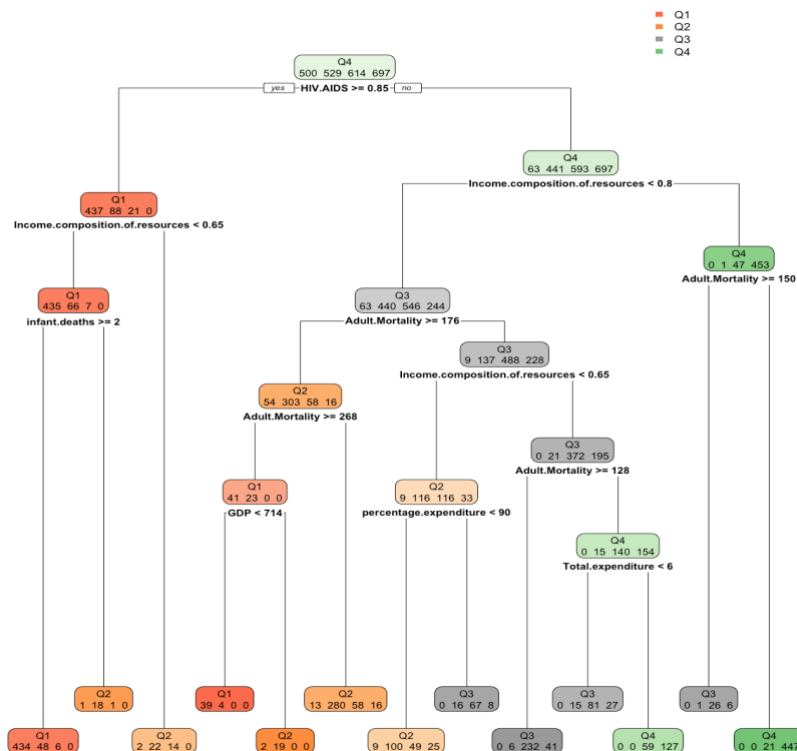


Figure10. Decision tree for categorical prediction

Continuous prediction

For the second part of the decision tree model, prediction of life expectancy is conducted using the continuous target variables. The model is created using the trained data for the estimation of Life expectancy. From figure 10, it is apparent that there are 8 leaf nodes for Life expectancy that are passed down through 6 decision nodes and a root node that involves Under five deaths, Adult mortality, HIV/AIDS and Income composition of resources. It can be seen that there is a similar root node on HIV/AIDS for the categorical decision tree model. Yet again, Income composition of resources plays a crucial role in deciding the life expectancy value under the given conditions. By observing closely, it is evident that the higher the income, the higher the life expectancy as well. MSE value for the test data on this model is 14.988. Likewise, Appendix F shows the plot between predicted life expectancy values VS actual life expectancy values.

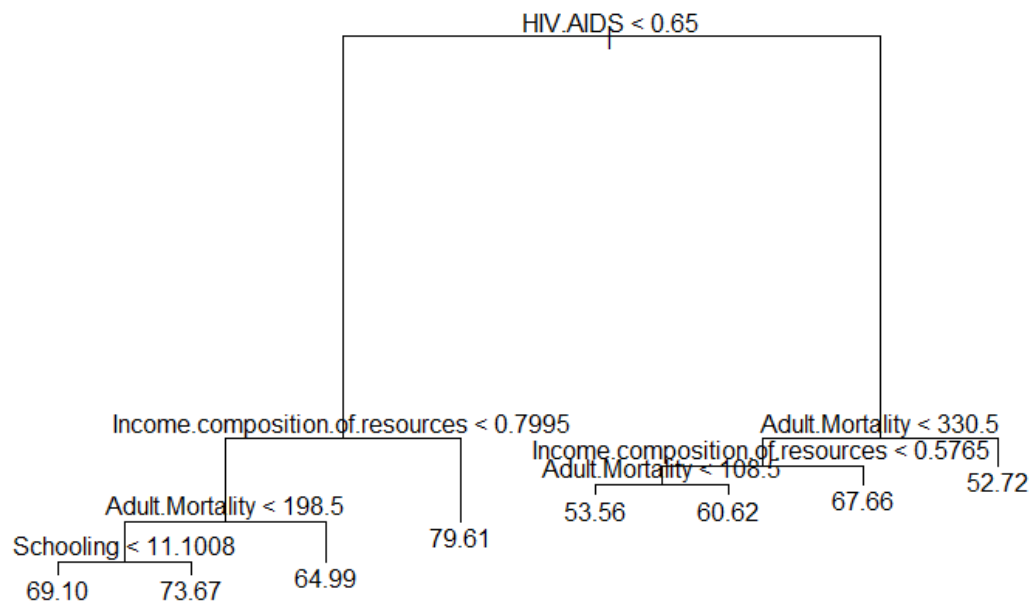


Figure11. Unpruned Decision tree for continuous prediction

Then when using cross validation to check whether pruning would improve the performance of the model, it was discovered that using 8 terminal nodes resulted in having the least deviation as per the plot in figure 11.

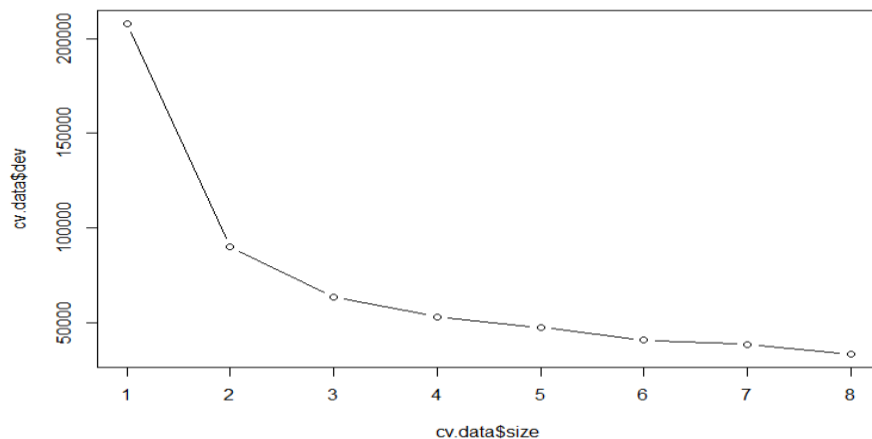


Figure12. Pruning plot (Size of tree VS Deviation)

Wanting to still explore the option of if there was any decrease in the MSE value for a pruned tree, a 5 terminal node tree was used. But it did not decrease the MSE in fact, it actually increased mean squared error (MSE) to 17.99. It was found that the 8-node tree produced the best results and so the unpruned version was used. Below picture displays the pruned version of the former tree. In conclusion, HIV/AIDS turned out to be the most important decision-making root node in both the decision trees. We can safely assume that countries with higher income tend to have higher Life expectancy from the decision tree model. Unpruned version of the decision tree recorded the lowest MSE compared to the pruned decision tree. Reduction of nodes certainly decreased the size of the tree, but it did not decrease the error rate.

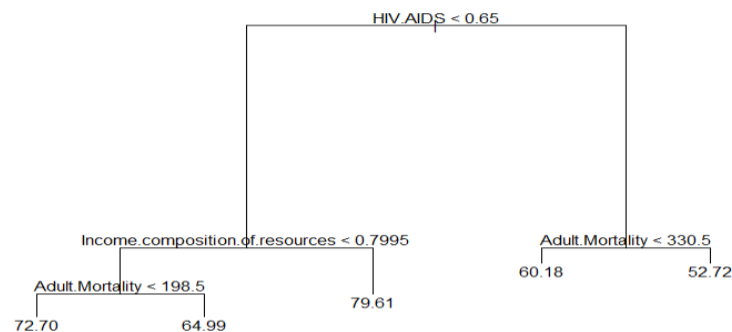


Figure13. Pruned Decision tree for continuous prediction

Random Forest:

In addition to using linear regression and decision trees, a random forest regression was also used to see if a better representation of the data could be created and to verify the importance of the

variables being used. With the random forest regression, a prediction accuracy of 90.27% was produced for categorical data. Initially with the continuous decision tree model an MSE of 14.988 was produced but when a random forest was used to sample a random subset of the predictors, an MSE of 2.643 was obtained which is substantially better. Next, the importance of the variables in the model was looked at. This is given in the figures below.

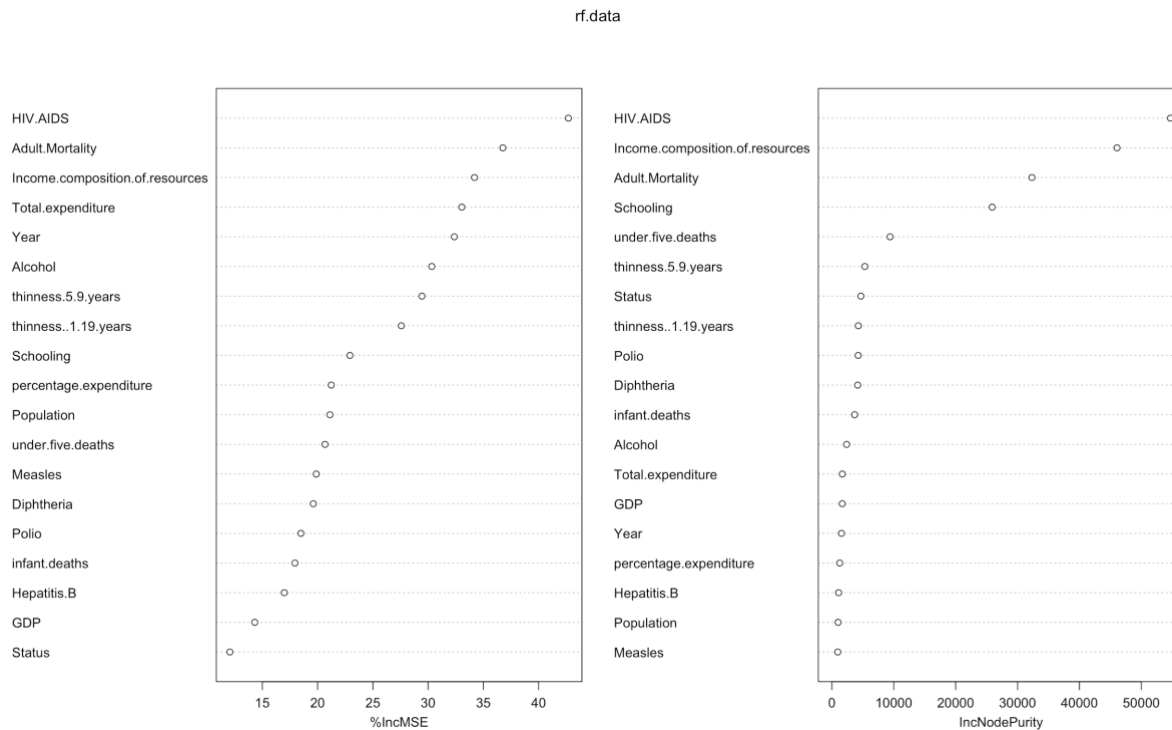


Figure 14. show the importance of the variables in the model by displaying the % Increase in MSE and Increase in Node Purity.

%in MSE indicates how much the MSE would increase if a variable was shuffled around randomly. A higher %MSE indicates a more important variable. Increase in Node Purity indicates how much purer splits involving a particular variable are over all trees. When looking at the IncNodePurity it is clear that the variables chosen for the linear regression are rated as the most important which helps validate that those variables are the best representation of the model.

Model Comparison and Conclusion

According to the research we did before, the data has underlying linear relationships within it. Though the mean squared error value (MSE) is lowest for Random forest, it is in a comparable range with Linear regression MSE value with a slight difference for the continuous target prediction. The random forest model was also able to classify those variables that were

substantially more important than others. It was also able to deal with nonlinear relations without transforming variables. Alternatively, linear regression works on significant variables in the model to produce fruitful results. From the variable importance plot, HIV/AIDS appears to have high importance followed by Adult mortality and Income composition of resources for discerning on Life Expectancy estimation. Even though both decision tree and random forest performed well on categorical and continuous data prediction, the random forest worked perfectly for the categorical target variables yielding the highest accuracy when compared to decision tree model.

Model	MSE for continuous target attribute	Accuracy for categorical target attribute
Linear Regression	3.172	-
Decision Tree	14.988	79.52%
Random Forest	2.643	92.15%

Table1. Comparison of model results

Variables that are highly correlated with the life expectancy like Adult Mortality, HIV AIDS, Income composition resources ended up being the most important factors in deciding the target variable in random forest. Significant variables that resonate with the Life expectancy were able to be found through the use of the random forest. Features like Alcohol, Hepatitis B, Total percentage expenditure are positively correlated with the Life expectancy and are assumed to be important factors in estimating the target variable. However, there is no significant effect derived when they are placed in an initial complete linear regression model. Besides, attributes like Infant deaths that are mildly correlated with Life expectancy proved to be significant in the best performing linear model. In conclusion, the status of a country has a substantial effect on Life expectancy since the developed countries have a higher schooling rate, income composition of resources and lower HIV/AIDS deaths and adult mortality rate than developing countries and therefore higher Life expectancy.

Finally, this project can be considered to be in initial stages of data analysis, and much more needs to be accomplished. Individual country analysis could also be done by grouping the data. This project shows that models like Random forest and Decision trees are reliable for these kinds of data prediction. What's more, linear regression also performs well on this data due to concealed linear relationships among the variables.

Challenges and Additional Efforts

There are several challenges during the data exploration part of this project. The first challenge comes from cleaning the data. Initially, the first method adopted was deleting missing items to organize the data. However, this caused the loss of nearly one-third of the sample size. This was because for each country there are different columns that have null values. The next method used in the study was to fill in the null values with average values for individual countries. This was to ensure the completeness and richness of the data as much as possible. In addition, when constructing the model, not all the weakly correlated variables were fitted, which may reduce the model's explanatory power and lose the predictive power of some models. In order to make the linear model perform better and improve the R-square of the model, nonlinearly fits could be used for all the weakly correlated variables.

Additional challenges encountered included the discovery of incorrect data in the original file. For example, most of the values for the BMI variable were greater than 50, which is obviously unreasonable. Since the BMI values were shocking, information was consulted on how to calculate the BMI value to ensure that the data was accurate. Finally, this part of the data needed to just be deleted because a reasonable value for BMI should be between 18-28. Also as previously mentioned, there were some errors in the value of the status variable. For example, the original document marked some developed countries as developing countries, like the UK and the USA, which forced the manual review and modification of the data, which took a lot of time to correct. Finally, when using decision trees and free trees to build models, the predicted variables in this study are continuous variables, not category or binary, which increases the difficulty of prediction for this study. To resolve this difficulty, two measures were taken. On the one hand, set labels for continuous variables, and then perform category prediction. On the other hand, continuous numerical modeling is used for prediction. The purpose of this is that the model is more convincing, predictable, and compared from two perspectives.

References:

KumarRajarshi. (2018, February 10). Life Expectancy (WHO). Retrieved from <https://www.kaggle.com/kumarajarshi/life-expectancy-who>

Prabhakaran,S.(n.d.).eval(ez_write_tag([[468,60], 'r_statistics_co-box-3', 'ezslot_0', 109, '0', '0']));Linear Regression. Retrieved from <http://r-statistics.co/Linear-Regression.html>

Bommae. (n.d.). University of Virginia Library Research Data Services Sciences. Retrieved from <https://data.library.virginia.edu/diagnostic-plots/>

Stephanie. (2019, January 20). Explanatory Variable & Response Variable: Simple Definition and Uses. Retrieved from <https://www.statisticshowto.com/explanatory-variable/>

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, 63(4), 308–319. doi: 10.1198/tast.2009.08199

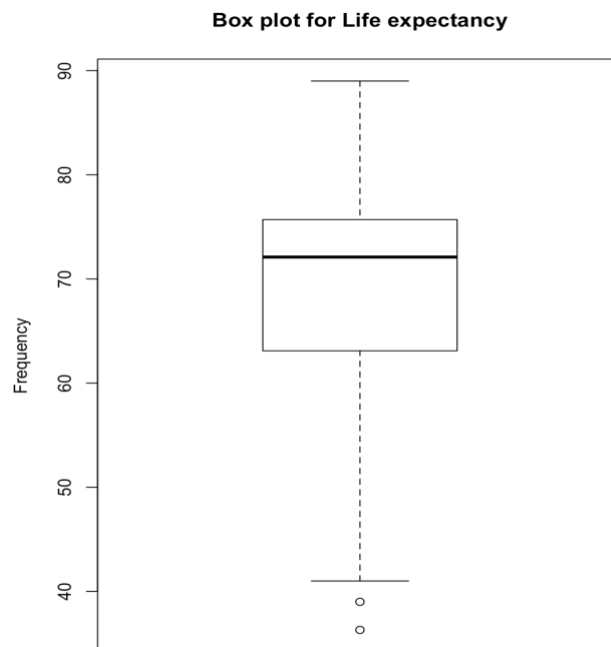
Olive, D. J. (2017). Multiple Linear Regression. *Linear Regression*, 17–83. doi: 10.1007/978-3-319-55252-1_2

Appendix A _Data

The introduction of each variable from the dataset.

# Country	# Year	# Status	# Life expectancy	# Adult Mortality	# infant deaths
Country	Year	Developed or Developing status	Life Expectancy in age	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)	Number of Infant Deaths per 1000 population
# Alcohol	# percentage expendi	# Hepatitis B	# Measles	# BMI	# under-five deaths
Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)	Expenditure on health as a percentage of Gross Domestic Product per capita(%)	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)	Measles - number of reported cases per 1000 population	Average Body Mass Index of entire population	Number of under-five deaths per 1000 population
# Polio	# Total expenditure	# Diphtheria	# HIV/AIDS	# GDP	# Population
Polio (Pol3) immunization coverage among 1-year-olds (%)	General government expenditure on health as a percentage of total government expenditure (%)	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)	Deaths per 1 000 live births HIV/AIDS (0-4 years)	Gross Domestic Product per capita (in USD)	Population of the country
# thinness 1-19 years	# thinness 5-9 years				
Prevalence of thinness among children and adolescents for Age 10 to 19 (%)	Prevalence of thinness among children for Age 5 to 9(%)				

Appendix B_box plot for the outlier of the life.expectancy and Summary statistics for variables after preprocessing

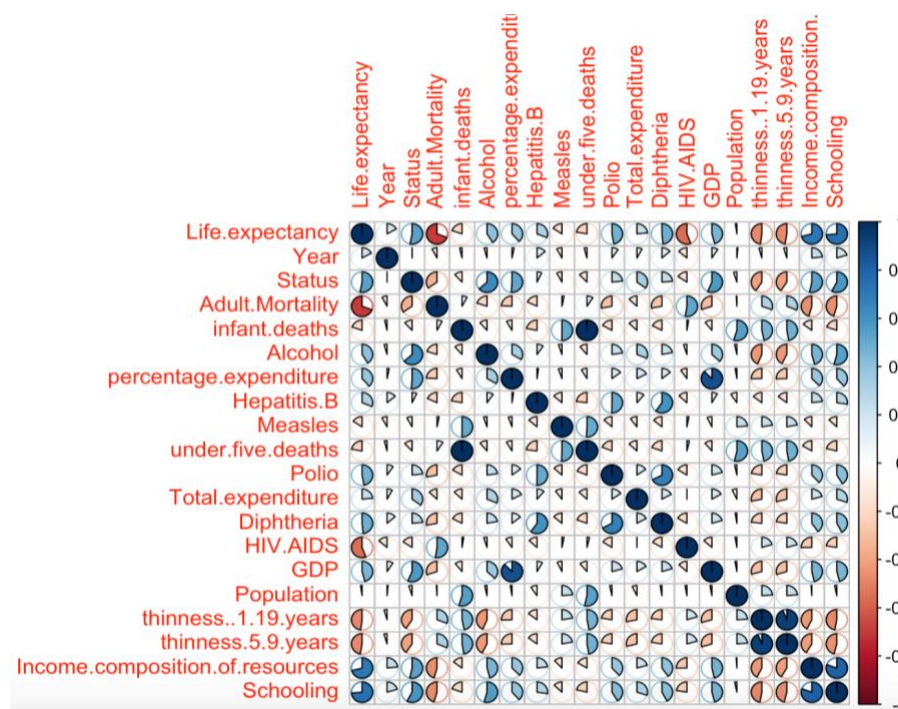


Life expectancy	Year	Status	Adult.Mortality	infant.deaths	Alcohol	percentage.expenditure
Min. :41.00	Min. :2000	Min. :0.0000	Min. : 1.0	Min. : 0.00	Min. : 0.010	Min. : 0.000
1st Qu.:63.12	1st Qu.:2004	1st Qu.:0.0000	1st Qu.: 74.0	1st Qu.: 0.00	1st Qu.: 0.950	1st Qu.: 4.807
Median :72.10	Median :2008	Median :0.0000	Median :144.0	Median : 3.00	Median : 3.695	Median : 65.723
Mean :69.25	Mean :2008	Mean :0.1914	Mean :164.5	Mean : 30.41	Mean : 4.605	Mean : 740.808
3rd Qu.:75.70	3rd Qu.:2012	3rd Qu.:0.0000	3rd Qu.:227.0	3rd Qu.: 22.00	3rd Qu.: 7.680	3rd Qu.: 443.088
Max. :89.00	Max. :2015	Max. :1.0000	Max. :723.0	Max. :1800.00	Max. :17.870	Max. :19479.912

Hepatitis.B	Measles	under.five.deaths	Polio	Total.expenditure	Diphtheria	HIV.AIDS
Min. : 1.00	Min. : 0	Min. : 0.00	Min. : 3.00	Min. : 0.370	Min. : 2.0	Min. : 0.100
1st Qu.:73.62	1st Qu.: 0	1st Qu.: 0.00	1st Qu.:78.00	1st Qu.: 4.280	1st Qu.:78.0	1st Qu.: 0.100
Median :88.00	Median : 17	Median : 4.00	Median :93.00	Median : 5.705	Median :93.0	Median : 0.100
Mean :78.86	Mean : 2428	Mean : 42.17	Mean :82.41	Mean : 5.905	Mean :82.2	Mean : 1.748
3rd Qu.:96.00	3rd Qu.: 362	3rd Qu.: 28.00	3rd Qu.:97.00	3rd Qu.: 7.430	3rd Qu.:97.0	3rd Qu.: 0.800
Max. :99.00	Max. :212183	Max. :2500.00	Max. :99.00	Max. :17.600	Max. :99.0	Max. :50.600

GDP	Population	thinness..1.19.years	thinness.5.9.years	Income.composition.of.resources	Schooling
Min. : 1.68	Min. :3.400e+01	Min. : 0.100	Min. : 0.100	Min. :0.0000	Min. : 0.00
1st Qu.: 561.92	1st Qu.:4.169e+05	1st Qu.: 1.600	1st Qu.: 1.600	1st Qu.:0.5040	1st Qu.:10.30
Median : 2833.37	Median :3.626e+06	Median : 3.400	Median : 3.400	Median :0.6730	Median :12.30
Mean : 7298.16	Mean :1.285e+07	Mean : 4.862	Mean : 4.893	Mean :0.6294	Mean :12.03
3rd Qu.: 5456.51	3rd Qu.:1.415e+07	3rd Qu.: 7.100	3rd Qu.: 7.200	3rd Qu.:0.7808	3rd Qu.:14.30
Max. :119172.74	Max. :1.294e+09	Max. :27.700	Max. :28.600	Max. :0.9480	Max. :20.70

Appendix C_Correlation pie chart for numeric variables



Appendix D _summary for lm.all, lm.fit1, lm.fit2

```
summary(lm.all)
```

```
Call:
lm(formula = Life.expectancy ~ ., data = train)

Residuals:
    Min       1Q   Median       3Q      Max
-21.8287  -2.1564  -0.0081   2.3175  16.7973

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.862e+01  3.942e+01   2.502  0.0124 *
Year        -2.173e-02  1.970e-02  -1.103  0.2701
Status        1.426e+00  3.342e-01   4.267 2.06e-05 ***
Adult.Mortality -1.882e-02  9.175e-04 -20.516 < 2e-16 ***
infant.deaths  9.880e-02  9.223e-03  10.712 < 2e-16 ***
Alcohol       5.852e-02  2.984e-02   1.961  0.0500 *
percentage.expenditure 1.395e-04  8.801e-05   1.585  0.1130
Hepatitis.B   2.661e-03  4.554e-03   0.584  0.5590
Measles      -2.220e-05  8.750e-06  -2.538  0.0112 *
under.five.deaths -7.396e-02  6.772e-03 -10.921 < 2e-16 ***
Polio         2.424e-02  5.032e-03   4.817 1.55e-06 ***
Total.expenditure -2.201e-02  3.842e-02  -0.573  0.5667
Diphtheria    3.420e-02  5.449e-03   6.277 4.11e-10 ***
HIV.AIDS     -4.868e-01  2.099e-02 -23.192 < 2e-16 ***
GDP          2.890e-05  1.361e-05   2.124  0.0338 *
Population    1.320e-09  1.807e-09   0.730  0.4654
thinness..119.years -8.785e-02  5.807e-02  -1.513  0.1305
thinness.5.9.years -4.398e-02  5.683e-02  -0.774  0.4390
Income.composition.of.resources 6.218e+00  7.202e-01   8.634 < 2e-16 ***
Schooling     7.735e-01  4.877e-02  15.861 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.093 on 2320 degrees of freedom
Multiple R-squared:  0.8167,    Adjusted R-squared:  0.8152
F-statistic: 544 on 19 and 2320 DF,  p-value: < 2.2e-16
```

summary(lm.fit1)

```
Call:
lm(formula = Life.expectancy ~ Status + Adult.Mortality + infant.deaths +
    under.five.deaths + Polio + Diphtheria + HIV.AIDS + Income.composition.of.resources +
    Schooling, data = train)

Residuals:
    Min       1Q   Median       3Q      Max
-22.0893  -2.2142   0.0713   2.3219  18.0643

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  53.5925184  0.5473094  97.920 < 2e-16 ***
Status       2.6950254  0.2735311   9.853 < 2e-16 ***
Adult.Mortality -0.0191660  0.0009188 -20.859 < 2e-16 ***
infant.deaths  0.0937834  0.0089823  10.441 < 2e-16 ***
under.five.deaths -0.0722661  0.0066370 -10.888 < 2e-16 ***
Polio         0.0257517  0.0049985   5.152 2.79e-07 ***
Diphtheria    0.0350009  0.0050973   6.867 8.40e-12 ***
HIV.AIDS     -0.4905563  0.0209662 -23.397 < 2e-16 ***
Income.composition.of.resources 6.9845961  0.7113597   9.819 < 2e-16 ***
Schooling     0.8244098  0.0474867  17.361 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.157 on 2330 degrees of freedom
Multiple R-squared:  0.8101,    Adjusted R-squared:  0.8094
F-statistic: 1105 on 9 and 2330 DF,  p-value: < 2.2e-16
```

summary(lm.fit2)

```

Call:
lm(formula = Life.expectancy ~ Status + ns(Adult.Mortality, 3) +
    infant.deaths + under.five.deaths + Polio + Diphtheria +
    ns(HIV.AIDS, 6) + ns(Income.composition.of.resources, 4) +
    Schooling, data = train)

Residuals:
    Min       1Q   Median       3Q      Max
-18.8972  -2.0239  -0.0763   1.7462  15.4724

Coefficients: (3 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    48.642042    1.281303   37.963 < 2e-16 ***
Status          0.743724    0.306402    2.427 0.015288 *
ns(Adult.Mortality, 3)1  -6.415688    0.522658  -12.275 < 2e-16 ***
ns(Adult.Mortality, 3)2  -3.132309    0.718806   -4.358 1.37e-05 ***
ns(Adult.Mortality, 3)3  -8.943355    1.140939   -7.839 6.88e-15 ***
infant.deaths     0.052406    0.007443    7.041 2.50e-12 ***
under.five.deaths -0.041010    0.005490   -7.470 1.13e-13 ***
Polio            0.014302    0.003961    3.611 0.000311 ***
Diphtheria       0.023847    0.004046    5.894 4.31e-09 ***
ns(HIV.AIDS, 6)1    -82.929512   10.195357  -8.134 6.68e-16 ***
ns(HIV.AIDS, 6)2    119.415817   13.514061    8.836 < 2e-16 ***
ns(HIV.AIDS, 6)3         NA         NA         NA      NA
ns(HIV.AIDS, 6)4     0.139654    1.686681    0.083 0.934019
ns(HIV.AIDS, 6)5         NA         NA         NA      NA
ns(HIV.AIDS, 6)6         NA         NA         NA      NA
ns(Income.composition.of.resources, 4)1  4.013615    0.432021    9.290 < 2e-16 ***
ns(Income.composition.of.resources, 4)2  5.761833    0.468952   12.287 < 2e-16 ***
ns(Income.composition.of.resources, 4)3  4.440276    0.849643    5.226 1.89e-07 ***
ns(Income.composition.of.resources, 4)4  16.114986    0.788252   20.444 < 2e-16 ***
Schooling        0.199195    0.043215    4.609 4.26e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

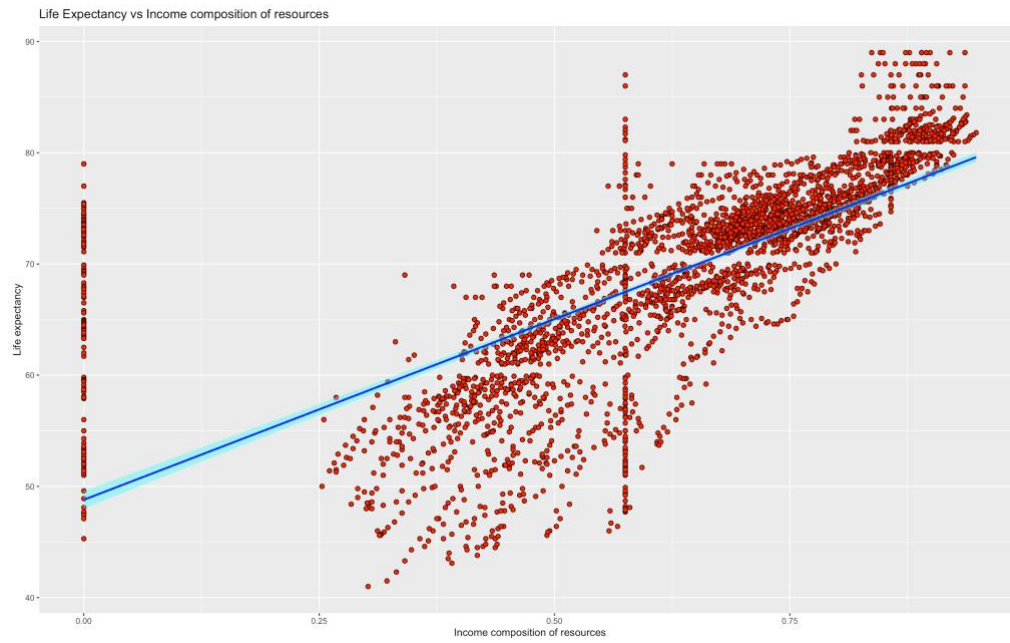
Residual standard error: 3.279 on 2323 degrees of freedom
Multiple R-squared:  0.8822,    Adjusted R-squared:  0.8814
F-statistic: 1087 on 16 and 2323 DF,  p-value: < 2.2e-16

```

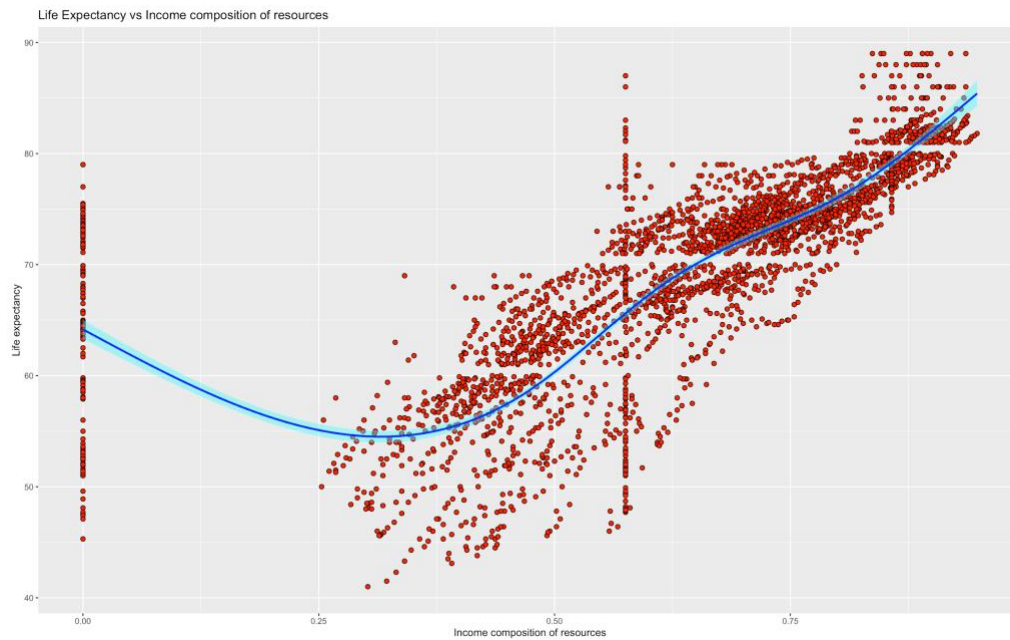
Appendix E_smooth model

Smooth model fitting for income composition of resources & Life.expectancy

Initial

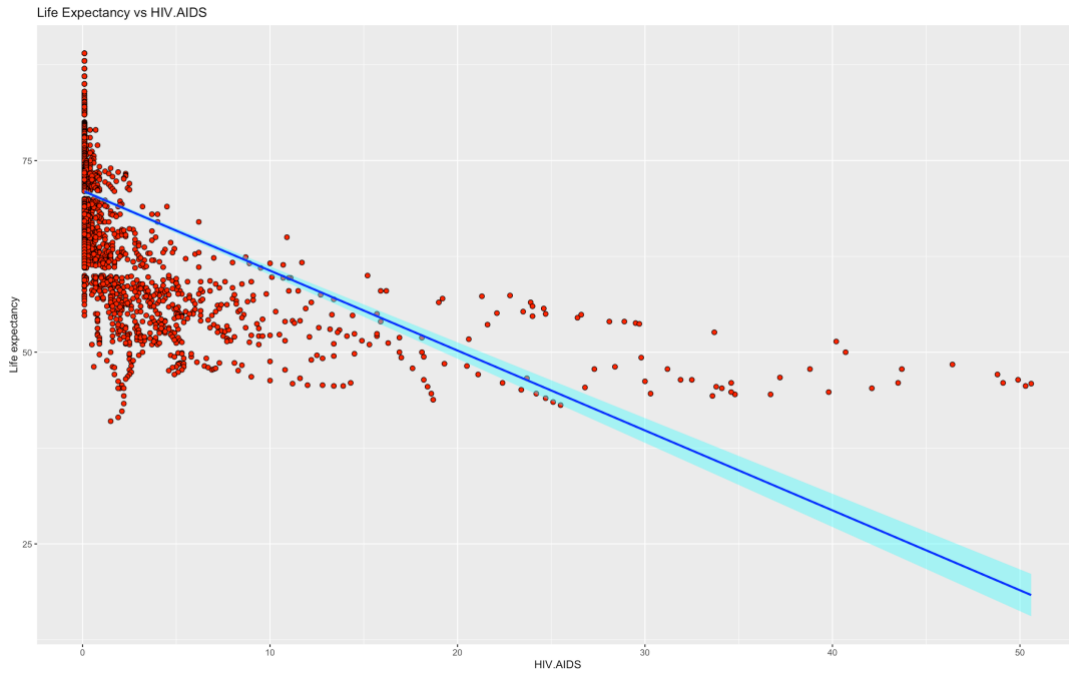


Final

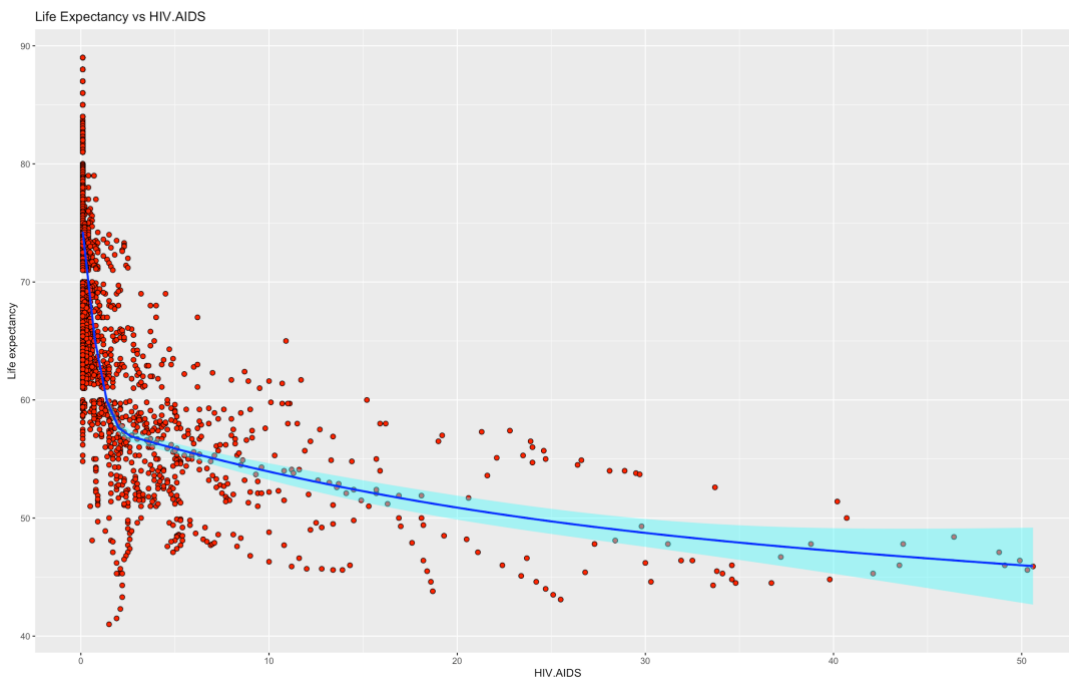


Smooth model fitting for HIV.AIDS& Life.expectancy

Initial

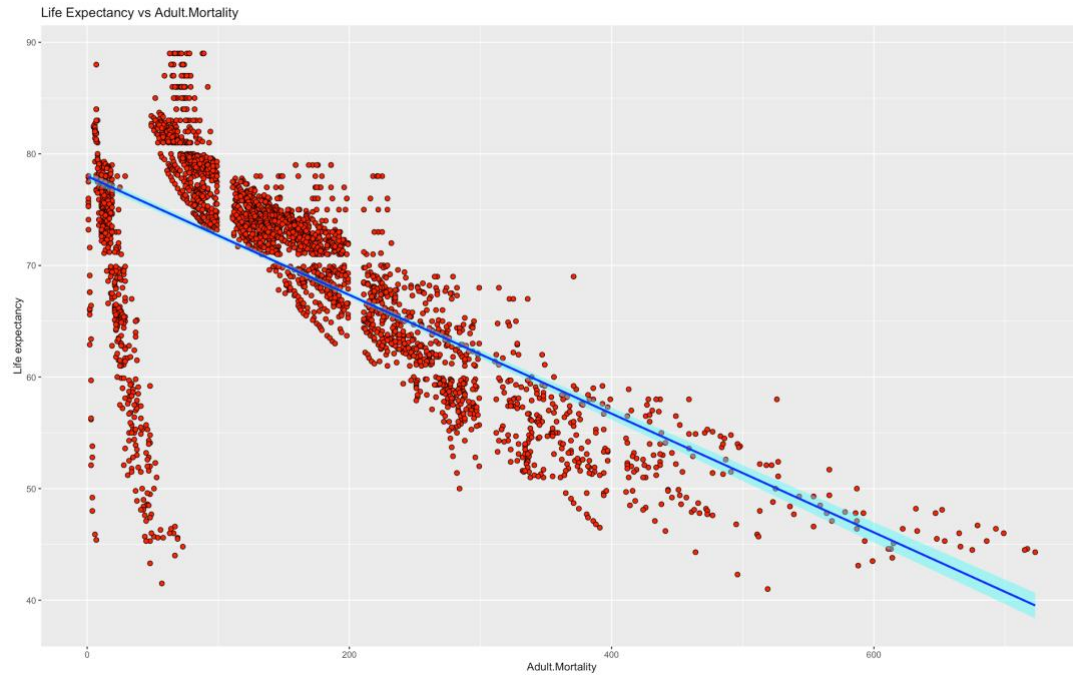


Final

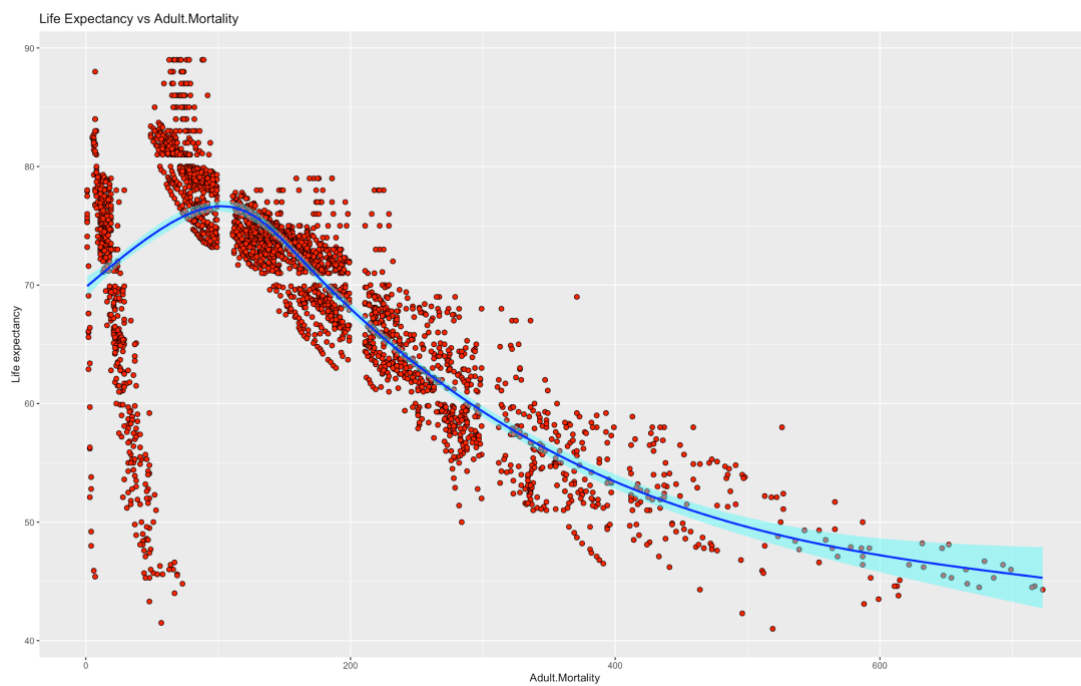


Smooth model fitting for Adult.Mortality & Life.expectancy

Initial



Final



Appendix F_DDecision tree Continuous

