MSDS 6372 Applied Statistics
2021-06-06
Team: Joseph Lazarus, Rick Fontenot, Satvik Ajmera

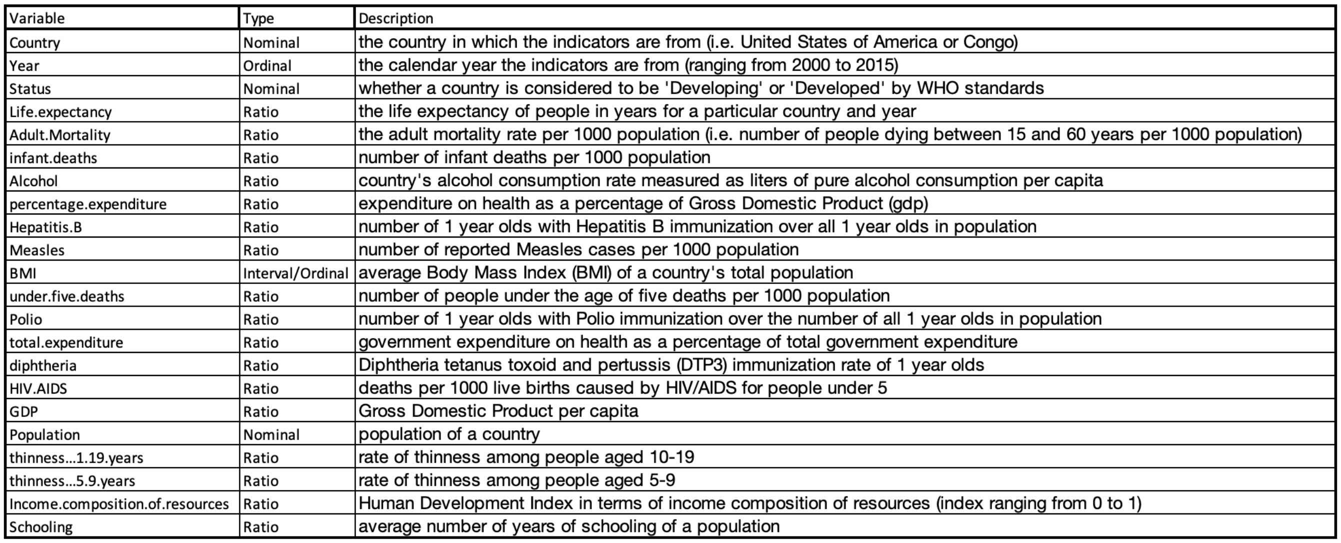
Project 1: Predicting Life Expectancy

I. Introduction

Life expectancy is a key metric for assessing human population health. It tells us the average age of death in a population. Estimates suggest that in pre-modern times, ie before The Age of Enlightenment (17th and 18th century) life expectancy was around 30 years. The world Health Organization’s (WHO) primary goal is to direct and coordinate international health within the United Nations (UN). Life expectancy provides a strong signal to the WHO focus their efforts.
Our goal is to predict life expectancy from a data set collected by the WHO. The results could give officials in the WHO an insight into which metrics have the greatest impact on world health.

II. Data Description

The data set comes from Kaggle, a popular Data Science website which is compiled from the Global Health Observatory under the WHO, who keeps tracks of global health status. There are 22 columns with 2,938 observations from 193 Countries over the years of 2000 to 2015.



Variable	Type	Description
Country	Nominal	the country in which the indicators are from (i.e. United States of America or Congo)
Year	Ordinal	the calendar year the indicators are from (ranging from 2000 to 2015)
Status	Nominal	whether a country is considered to be 'Developing' or 'Developed' by WHO standards
Life.expectancy	Ratio	the life expectancy of people in years for a particular country and year
Adult.Mortality	Ratio	the adult mortality rate per 1000 population (i.e. number of people dying between 15 and 60 years per 1000 population)
infant.deaths	Ratio	number of infant deaths per 1000 population
Alcohol	Ratio	country's alcohol consumption rate measured as liters of pure alcohol consumption per capita
percentage.expenditure	Ratio	expenditure on health as a percentage of Gross Domestic Product (gdp)
Hepatitis.B	Ratio	number of 1 year olds with Hepatitis B immunization over all 1 year olds in population
Measles	Ratio	number of reported Measles cases per 1000 population
BMI	Interval/Ordinal	average Body Mass Index (BMI) of a country's total population
under.five.deaths	Ratio	number of people under the age of five deaths per 1000 population
Polio	Ratio	number of 1 year olds with Polio immunization over the number of all 1 year olds in population
total.expenditure	Ratio	government expenditure on health as a percentage of total government expenditure
diphtheria	Ratio	Diphtheria tetanus toxoid and pertussis (DTP3) immunization rate of 1 year olds
HIV.AIDS	Ratio	deaths per 1000 live births caused by HIV/AIDS for people under 5
GDP	Ratio	Gross Domestic Product per capita
Population	Nominal	population of a country
thinness...1.19.years	Ratio	rate of thinness among people aged 10-19
thinness...5.9.years	Ratio	rate of thinness among people aged 5-9
Income.composition.of.resources	Ratio	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
Schooling	Ratio	average number of years of schooling of a population

Figure 1: Data Description

Through our exploration of the data set, many variables appear to have flawed data including errors in population numbers which then impact other variables that are derived on a per capita basis. There are also a significant amount of missing data points. These issues as well as methods of improvement will be summarized in our exploratory data analysis.

III. Exploratory Data Analysis (EDA)

Based on the goal of predicting Life Expectancy through regression and non-parametric prediction modeling our EDA is focused handling missing values, data issues, covariance between predictor variables and linearity of relationships.

Summary Statistics

Country level life expectancy ranges from 36.3 to 89 with a median of 72.1 See Appendix I for a full summary statistics table.

Missing Values

The Population variable is missing over 20% of it's observations including missing all 16 years of data for 40 of the countries. GDP is missing 15% observations as well. In the next section we will explore further issues on these parameters and our method to improve our set. Hepatitis.B is missing 19% but due to lack of available sources we did not have a reliable way to improve this parameter. For other parameters with less than 7% missing values, we chose to omit NA values rather than impute. See Appendix I for plots and details.

Data Issues

In order to reduce the missing values for population, we merged in an updated data set from the World Bank (outside of kaggle). This process highlighted an issue with unreliable population numbers in the original data set beyond just missing values. Since many of the other variables are per capita ratios calculated with population, this data reliability issue impacted. See the appendix for more information on how we analyzed and improved data reliability. See Appendix I for plots and details.

Collinearity and initial variable reduction

Infant deaths and under 5 deaths are highly correlated with an R-square of 0.99. In single parameter regression models, under five deaths has a higher correlation to Life Expectancy. Measles is also correlated with both of these parameters and has a weaker correlation to life expectancy. We chose to include under five deaths for model analysis. Infant deaths and measles were dropped from our data set.

Beyond collinearity, Hepatitis.B (R-square=0.06) and Total.expenditure (R-square=0.03) displayed weak to no correlation with Life Expectancy and have high percentage of missing values. These columns were dropped from modeling data sets.

Linearity of relationships to Life expectancy

Plots vs. Life expectancy for each potential predictor variable showed non-linear relationships where a log transformation for the following variables improved diagnostic residual plots to meet assumptions of linear regression: Corrected.expenditure, EstGDPpercapita, HIV.AIDS, under.five.deaths and adj.AdultMortality. We created transformed parameters for these variables to feed into linear modeling variable selections and kept the untransformed data for non-parametric models. See Appendix I

IV. Addressing Objective 1

Build an interpretable model using Multiple Linear Regression techniques to predict Life Expectancy from the WHO data set.

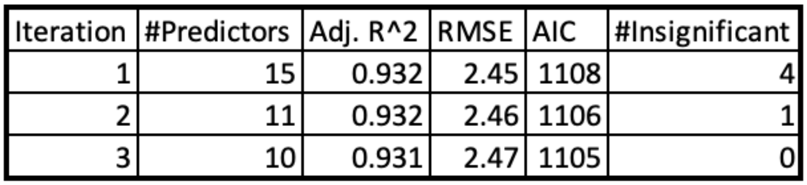
Data Preparation

We choose to split on our data set on year. This strategy gives us two advantages in our analysis. We capture nearly identical composition of countries by region and status, in our training and test sets. In addition, any trends in the data could be converted into a time series analysis. We achieved a 70/30 split by gathering observations from 2000 - 2010 in our training. Following years were collected into our test set.

Approach

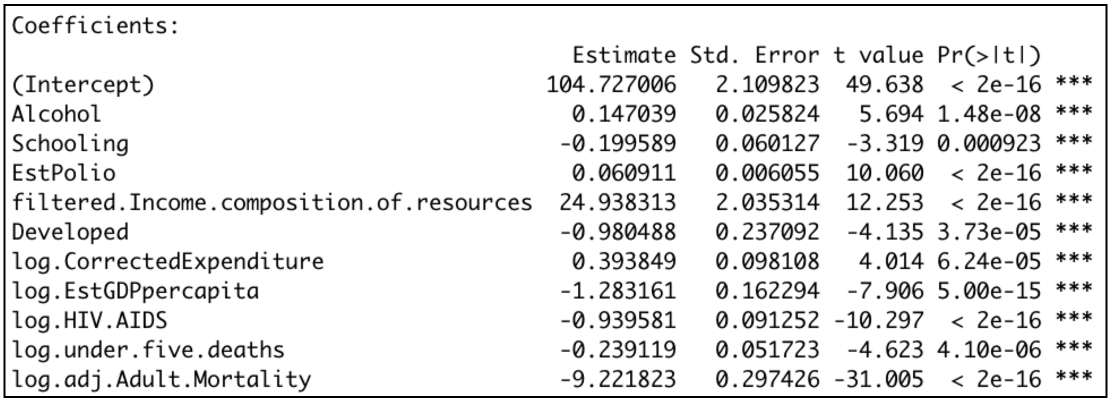
Our initial model is built intuition from our EDA based on the following findings:
1) Cleaning the data issues as described above and in Appendix I
2) Dropping infant.deaths and Measles due to collinear relationships to stronger predictors
3) Dropping the non-correlated Hepatitis.B and Total.expenditure due to high missing values
4) Log transformation of 5 variables as described above and in Appendix I figures 11 - 17
5) Model all 15 remaining cleaned and transformed predictors, then iteratively remove non-significant variables based on p-values to reduce the model

After 3 iterations, the remaining 10 variables are all highly significant



Iteration	#Predictors	Adj. R^2	RMSE	AIC	#Insignificant
1	15	0.932	2.45	1108	4
2	11	0.932	2.46	1106	1
3	10	0.931	2.47	1105	0

Figure 4



```
Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          104.727006   2.109823  49.638  < 2e-16 ***
Alcohol                                0.147039   0.025824   5.694 1.48e-08 ***
Schooling                             -0.199589   0.060127  -3.319 0.000923 ***
EstPolio                               0.060911   0.006055  10.060  < 2e-16 ***
filtered.Income.composition.of.resources  24.938313   2.035314  12.253  < 2e-16 ***
Developed                             -0.980488   0.237092  -4.135 3.73e-05 ***
log.CorrectedExpenditure               0.393849   0.098108   4.014 6.24e-05 ***
log.EstGDPpercapita                   -1.283161   0.162294  -7.906 5.00e-15 ***
log.HIV.AIDS                          -0.939581   0.091252 -10.297  < 2e-16 ***
log.under.five.deaths                 -0.239119   0.051723  -4.623 4.10e-06 ***
log.adj.Adult.Mortality               -9.221823   0.297426 -31.005  < 2e-16 ***
```

Figure 5

Check Assumptions

Residuals from iteration 3 model scatter along the range of predicted values. This meets the constant variance assumption. The high leverage observations all have standardized residuals centered near zero and do not cross the Cook's D threshold therefore do not appear to be highly influential. Departure from normality is observed but the sample size is sufficiently large to meet the requirements of the central limit theorem.

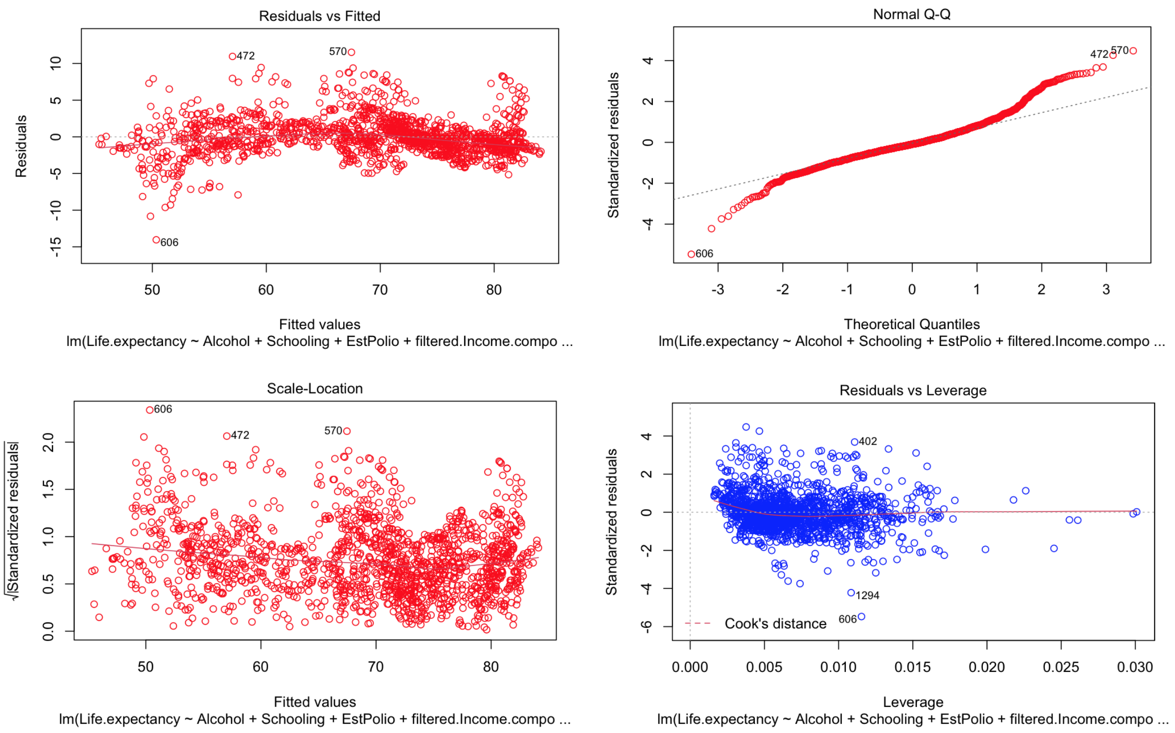


Figure 6

Model Interpretation

Life Expectancy = 104.7 + (0.15 x Alcohol) - (0.20 x Schooling) + (0.06 x Polio) + 24.9 x Income...resources) - (0.98 x Developed) + [0.394 x log(Expenditure)] - [1.28 x log(GDP)] - [0.94 x log(HIV.AIDS)] - [0.24 x log(under.five.deaths)] - [9.22 x log(Adult Mortality)]

Parameter Interpretations

Holding all other variables constant the effects of each individual predictors are:

- For every 1 unit increase in Alcohol consumption rate the mean Life expectancy increases 0.15 years. This positive relationship is not intuitive and may need to be explored further.
- For every 1 year increase in average schooling, mean life expectancy decreases by 0.2 years. This negative relationship is not intuitive and not inline with a positive relationship shown in individual scatterplot. This may be a sign of overfitting to look at further.
- For every 1% increase in the Polio vaccination rate among 1 year olds, the mean life expectancy increases by 0.06 years
- For every 1 unit increase in the Human development index in terms of income composition of resources, the mean life expectancy increases by 24.9 years. However note that the median value of composition is 0.68 and the max is 0.95 and likely on a scale of 0 to 1 so a full 1 unit increase is not likely
- As a country transitions from undeveloped to developed status the mean life expectancy increases by 0.02 years
- For a 100% increase in Expenditure, the difference in the expected mean Life expectancy will be 0.394 * log(1+1) = 0.27 years

- For a 100% increase in GDP, the decrease in the expected mean Life expectancy will be 1.28 * log(1+1) = 0.88 years. This negative relationship is not intuitive and could be interested to study further
- For a 100% increase in ratio of deaths caused HIV.AIDS, the decrease in the expected mean Life expectancy will be 0.94 * log(1+1) = 0.65 years.
- For a 100% increase in rate of deaths for people under 5 years old, the decrease in the expected mean Life expectancy will be 0.24 * log(1+1) = 0.17 years.
- For a 100% increase in Adult mortality, the decrease in the expected mean Life expectancy will be 9.22 * log(1+1) = 6.4 years.

V. Addressing Objective 2

Build a more complex model using MLR to predict Life Expectancy from the WHO data set.

Data Preparation

The same technique described above is used to prepare the data in a train test split.

Approach

We used LASSO to identify key variables in the data, See table 1 for results. By varying what variables LASSO would consider and altering lambda, one of the hyperparameters, we came up with three models. Low Lambda value allowed variables to enter the model. Increasing this value restricted the number of variables selected. To achieve the goal of building a more complex model we altered the set of variables LASSO would consider to include various interaction terms. Then tuned the hyperparameter to come up with the model listed last in table 1.

A Précis of LASSO selection

LASSO applies a cost function to the residual sums of squares (RSS). In simple terms this discourages over fitting. RSS measures the amount of error between the regression function and the data set. The hyperparameters of Lambda and Alpha apply regularization of the coefficients. When the penalty term applied to the coefficients reduces them to zero or a very small number the algorithm deselects those terms.

Results

Table 1: Lasso selection Results

ID	Predictors Selected by Lasso	Test RMSE	AIC	Lambda	Alpha
1	17 Predictors	2.33	984.54	0.01	1
2	5 Predictors	2.45	1016.40	0.5	1
3	6 Predictors including interaction terms	2.41	1002.29	0.0132	1

We used the caret package to rank the coefficients produced by LASSO to rank them in order of importance. The following variables were used to construct the complex model to predict Life Expectancy: **Income Composition of Resources**, **Adult Mortality**, **Thinness 1-19 * Developed**, and **Alcohol * Developed.**

VI. Non Parametric Models

Use non parametric models techniques to build a model that best predicts Life Expectancy from the WHO data set.

A Précis of KNN Regression

K-Nearest Neighbors is a non parametric method that associates the relative distance between independent variables and the continuous outcome variable. One can think of the adage, "birds of a feather.." The Hyperparamter of K determines the nearest points the algorithm will consider in averaging the distance. More in Appendix II, paragraph 2

Data Preparation

Preparing the data for KNN often requires us to scale the data. Since Euclidean distance is used we do not want the algorithm to be affected by differences in magnitudes. To account for this the data is scaled on a range from 0 through 1. This makes prediction with KNN regression much more powerful but at the cost of interpretability. Once scaled the data was divided into train/test sets as described in the previous sections.

Variable Selection

To optimize the KNN regression we used a variable selection method tailored to KNN. Backward elimination feature selection with random KNN, rknnBeg. (see appendix). Our selection process chose the following five scaled variables; **Adult mortality**, **Income composition of resources**, **HIV**, **thinness 5-9**, **BMI**. More details in Appendix I paragraph 2 B.

Comparison of KNN Regression Results

Using the five variables from our selection process. We ran the model through a loop each time increasing K by 1 and stopping at 50. At each step we calculated AIC (see appendix for AIC equation) and RMSE for that K value.(Figure 8). From these 50 models we selected the best performing model. The optimal K value we found, K=8. See Appendix **I**, figure 7.

Table 2 is a comparison of performance between KNN regression. Our top two KNN models were achieved by the model selection process described above. Out performing KNN regression conducted on a model with all variables. (see Appendix **II**, table 2 for full results)

Table 2: Comparison of KNN models

ID	*Model*	*RMSE*	*AIC*
1	*K = 8 (5 variables)*	*2.26*	*929.13*
2	*K = 11 (5 Variables)*	*2.28*	*937.60*

Plotting the observed vs predicted Life Expectancy number produced a graph that generally follows a 45 degree line. This is the type of trend we want. Tells us that the model is doing a good job of predicting Life Expectancy. See Appendix **II**, figure 8.

Insights

Performance metrics between the All Variable and 5 variable groups of KNN models are not far off. Scaling the data shrank our x values so that they were all relatively close together. With 5 variables the trends were still apparent to the algorithm. The AIC score improved by reducing the penalty for the number of predictors in the model. See Appendix II, For AIC formula applied to all the models.

VII. Summary

To recap our findings from objectives 1 and 2, the table below summarizes our final models:

Model	*Number of Parameters*	*RMSE*	*AIC*
Manual MLR	*10*	*2.47*	*1105*
Complex MLR	*6*	*2.41*	*1002*
KNN-regression	*5*	*2.26*	*929*

Using MLR to predict life expectancy from the WHO and combining it with the world bank data we found that most factors with the most predictive power from this data set were **Alcohol, Schooling, Polio, Income.composition.of.resources, Developed, Expenditure, GDP, HIV.AIDS, under.five.deaths, Adult Mortality.**

When LASSO assisted in selection variables and interaction terms this created a more complex model; **Income Composition of Resources**, **Adult Mortality**, **Thinness 1-19 * Developed**, and **Alcohol * Developed**. While this model added complexity in the search for useful interaction terms, it did not improve the AIC or RMSE compared to our prior MLR model based on intuition from EDA.

Running the knn-regression non parametric model allowed us to make predictions based on relationships with curvature and nearest neighbors rather than just linear models. This produced the best model in terms of error rates exhibited in RMSE and AIC. The optimal variables here were **Adult mortality**, **Income composition of resources**, **HIV**, **thinness 5-9**, **BMI**.

Scope of Inference

The data set is from an observational study rather than a designed, controlled experiment. Thus causal inference can not be made and our inference is limited to this data set. With that in mind, the study may indicate areas to search for improvements to life expectancy by using tools such as increasing amount of income composition, taking steps to reduce deaths from HIV & AIDS through prevention measures, increasing polio vaccination rates.

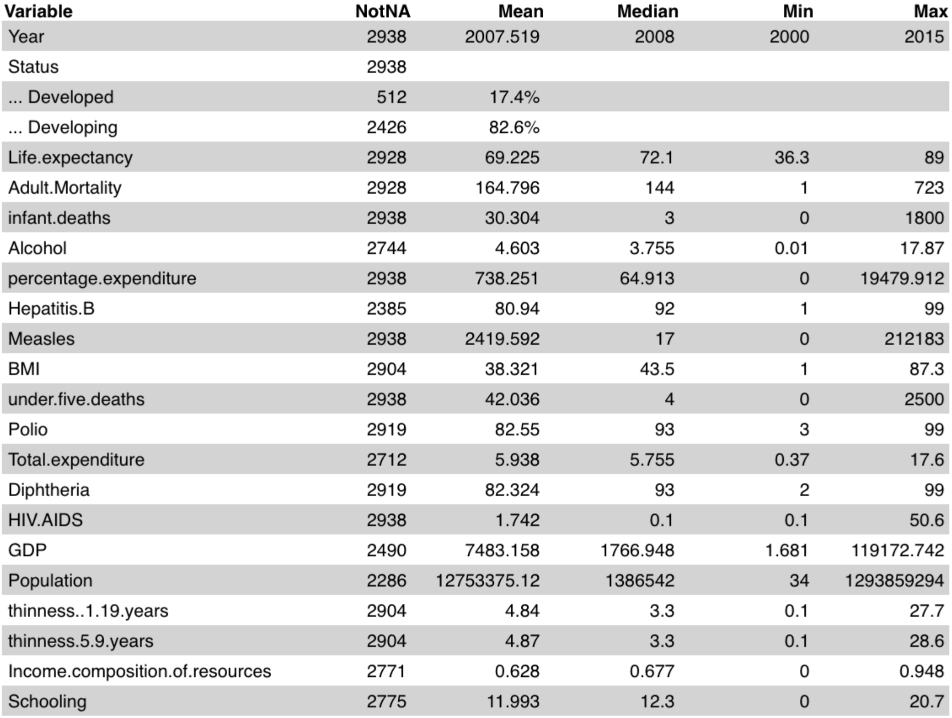
Taking measures to prevent deaths of kids under 5 is also significant but would require further study into the causes of these deaths. Similarly, while Adult mortality rate is a significant predictor of mean life expectancy, further studies into the causes of these deaths on a country level would be needed to assess areas for improvements.

Given more time we'd like to explore more variable selection methods to use with KNN regression. Also, explore decision tree boosting which we only touched upon in my research in reading about trees.

APPENDIX I

Summary Statistics

The following table shows summaries of all variables available in the data set.



Variable	NotNA	Mean	Median	Min	Max
Year	2938	2007.519	2008	2000	2015
Status	2938				
... Developed	512	17.4%			
... Developing	2426	82.6%			
Life.expectancy	2928	69.225	72.1	36.3	89
Adult.Mortality	2928	164.796	144	1	723
infant.deaths	2938	30.304	3	0	1800
Alcohol	2744	4.603	3.755	0.01	17.87
percentage.expenditure	2938	738.251	64.913	0	19479.912
Hepatitis.B	2385	80.94	92	1	99
Measles	2938	2419.592	17	0	212183
BMI	2904	38.321	43.5	1	87.3
under.five.deaths	2938	42.036	4	0	2500
Polio	2919	82.55	93	3	99
Total.expenditure	2712	5.938	5.755	0.37	17.6
Diphtheria	2919	82.324	93	2	99
HIV.AIDS	2938	1.742	0.1	0.1	50.6
GDP	2490	7483.158	1766.948	1.681	119172.742
Population	2286	12753375.12	1386542	34	1293859294
thinness..1.19.years	2904	4.84	3.3	0.1	27.7
thinness.5.9.years	2904	4.87	3.3	0.1	28.6
Income.composition.of.resources	2771	0.628	0.677	0	0.948
Schooling	2775	11.993	12.3	0	20.7

Life Expectancy is increasing over time and there appears to be a split between Developed and Developing countries. This could be an interesting interaction to explore in modeling but may also be described by the differences in individual metrics within these populations

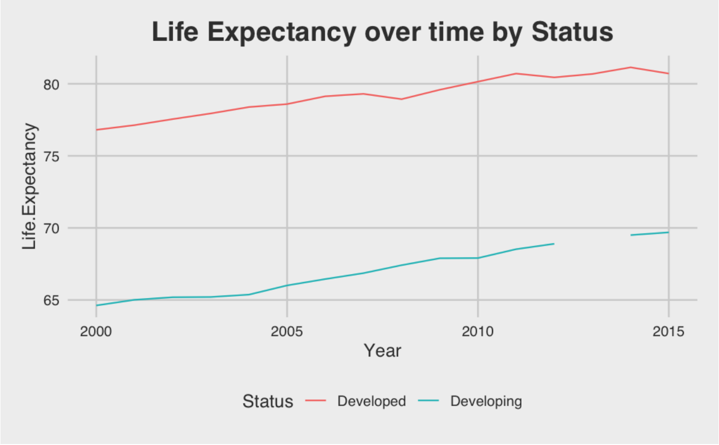


Figure 3 Life Expectancy over time

Data Issues

Details on data reliability concerns and adjustments we made.

Figure 3 shows population by year for the country of Rwanda as an example. Intuitively it does not make sense for the population to drop to near zero one year then back at high levels a year or two later. The replacement data from the current world bank set shows a more reliable trend in population and matches the high points on the trend of original data indicating that many years in the kaggle set are in error by factors of 10 and 100. Figure 4 shows that when compared across all countries there is a high percentage of observations where these factor errors occur. We chose to use the new replacement population data from World Bank (renamed to ("EstPopulation"). In addition to correcting these errors, the missing values were also reduced from 652 to just 4 observations

Figure 3: Rwanda population example

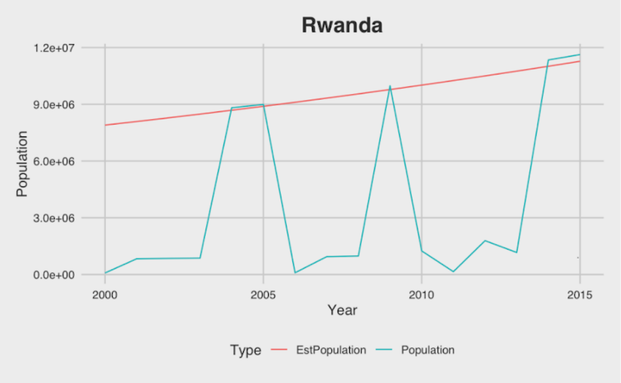


Figure 4: Original Population data vs. new WHO

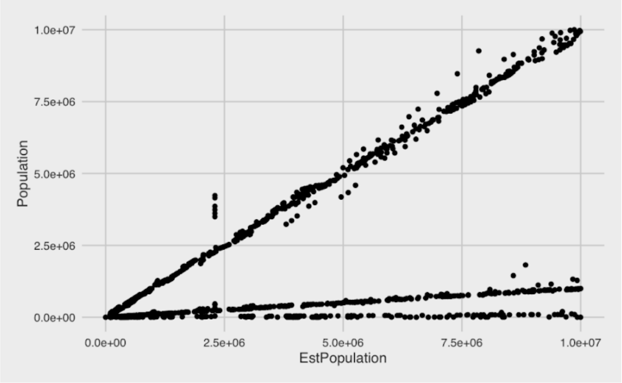


figure 9

The GDP per capita data in the original set shows the same issue as it appears to be calculated from the incorrect population data (see similar plots as above in Figures 5 & 6). We chose to use corrected data from the current World Bank source in place of the kaggle column (renamed to (EstGDPpercapita).

Figure 10 left Rwanda GDP example

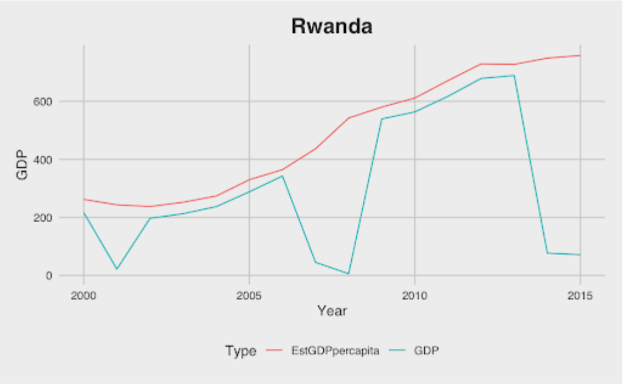
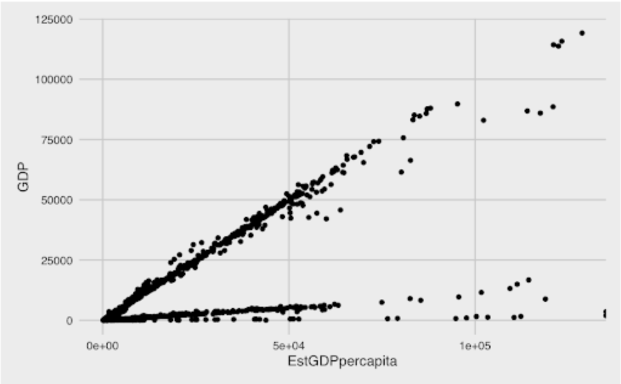


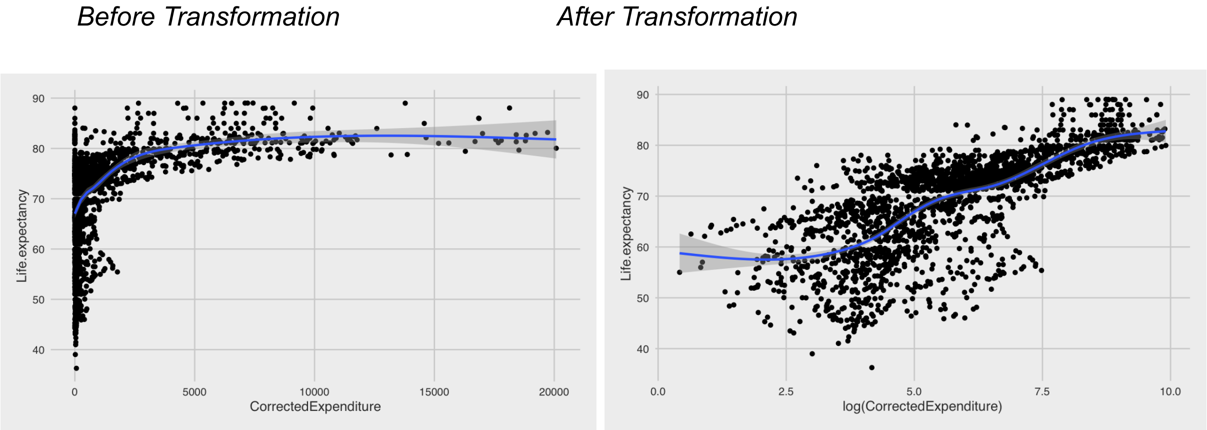
Figure 10 right Original GDP data vs. new WHO



All other potential predictor variables that were per capita based were checked similarly and the percentage.expenditure and Adult.mortality also showed similar issues, all others appeared to be reliable. Adult Mortality was replaced with current data from the WHO (renamed “new.AdultM.ortality”) and percentage expenditure was corrected with an error factor calculated by the ratio of EstGDP/originalGDP (renamed “Corrected.expenditure).

Linearity of relationships to Life expectancy, transformations and improvements

Corrected expenditure exhibited a non-linear relationship to Life expectancy with an R-squared of 0.25 and non-constant variance on residuals. After a log transformation of Expenditure the R-square increased to 0.55 and the residuals appear to have constant variance


Before Transformation

After Transformation

Figure 11

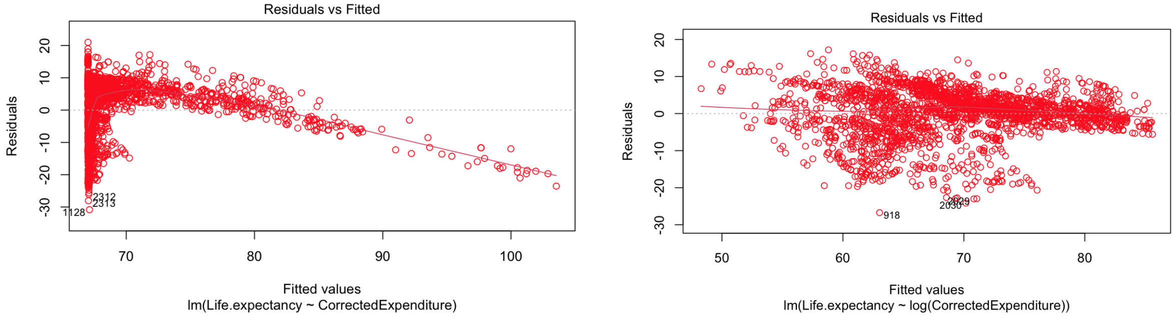
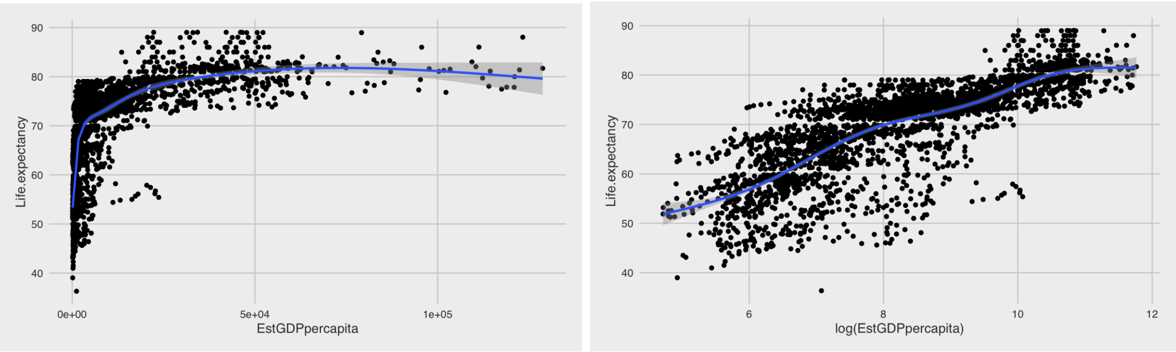


Figure 12

Estimated GDP per capita exhibited a non-linear relationship to Life expectancy with an R-squared of 0.32 and non-constant variance on residuals. After a log transformation of Expenditure the R-square increased to 0.62 and the residuals appear to have constant variance


Before Transformation

After Transformation

Figure 13

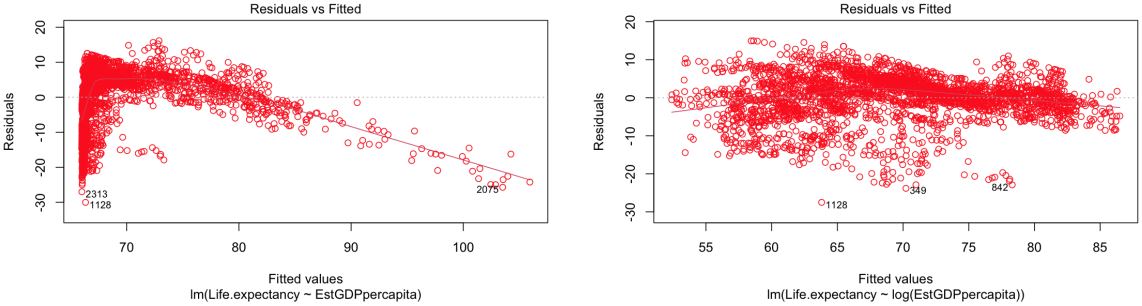


Figure 14

HIV.AIDS exhibited a non-linear relationship to Life expectancy with an R-squared of 0.31 and non-constant variance on residuals. After a log transformation of Expenditure the R-square increased to 0.66 and the residuals appear to have constant variance

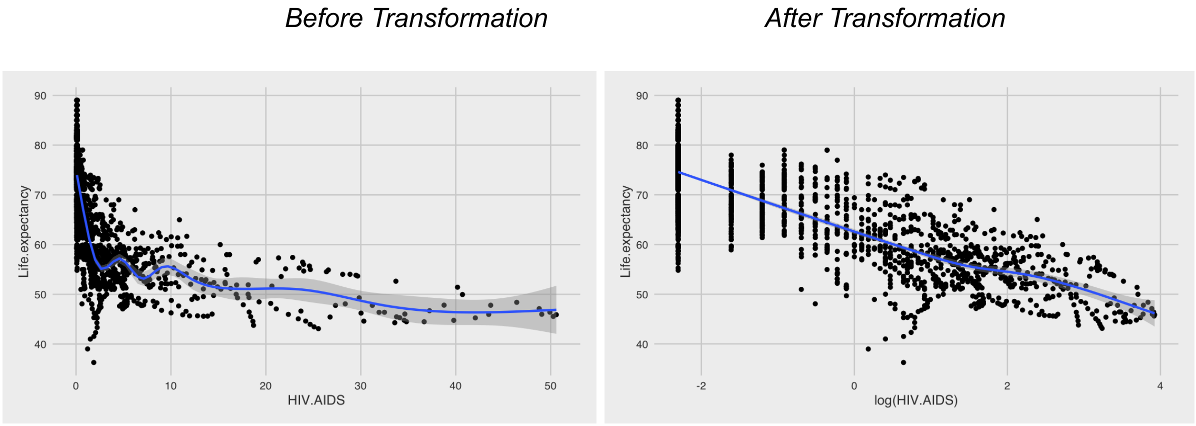


Figure 15

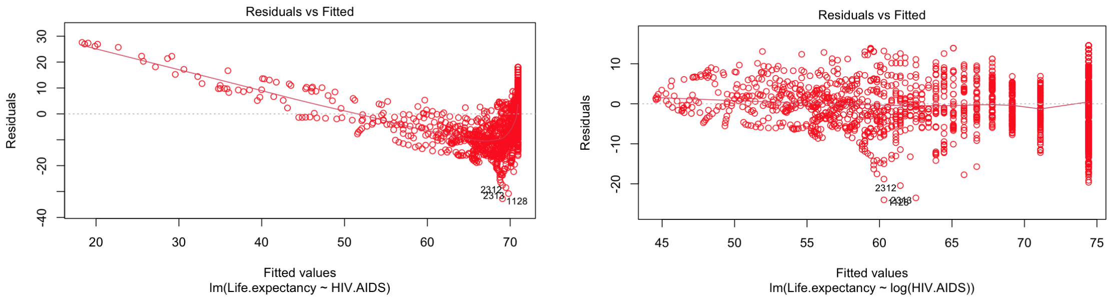
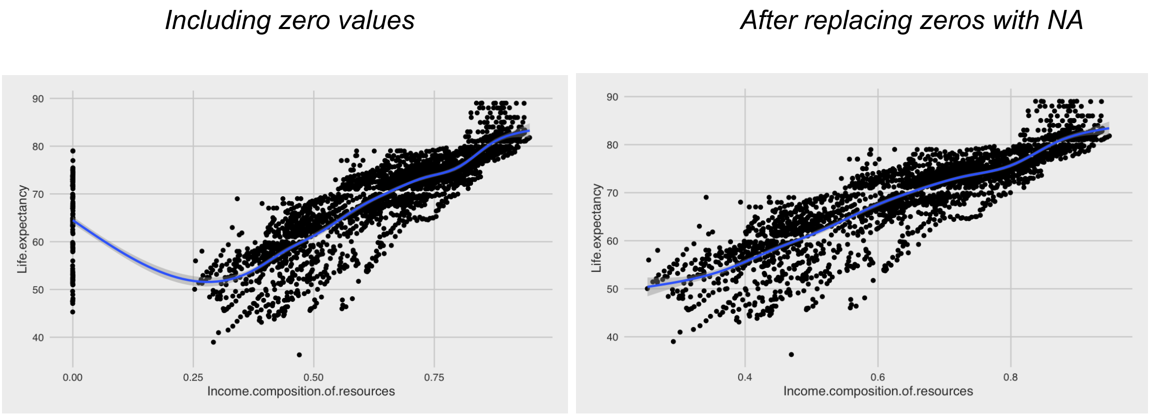


Figure 16

Income.composition.of.resources had an issue with zero values instead of NA(missing values) with an R-squared of 0.52 and issues with high leverage / high residual values. After replacing the zeros with NA, the R-square increased to 0.79 and the residuals diagnostics are improved


Including zero values

After replacing zeros with NA

Figure 17

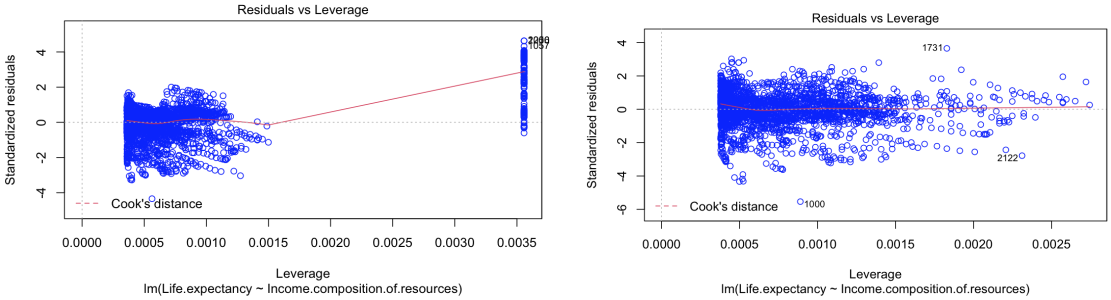


Figure 18

Under five deaths exhibited a non-linear relationship to Life expectancy with an R-squared of 0.05 and non-constant variance on residuals. After a log transformation of Expenditure the R-square increased to 0.38 and the residuals appear to have constant variance

Before Transformation *After Transformation*

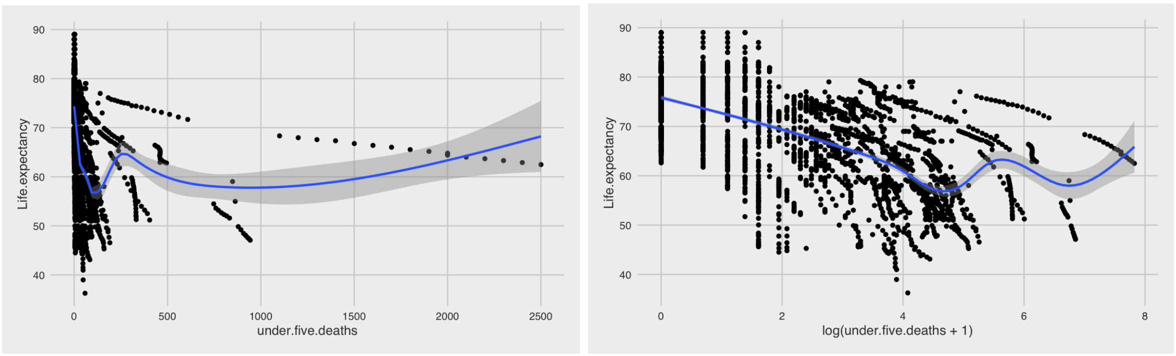


Figure 19

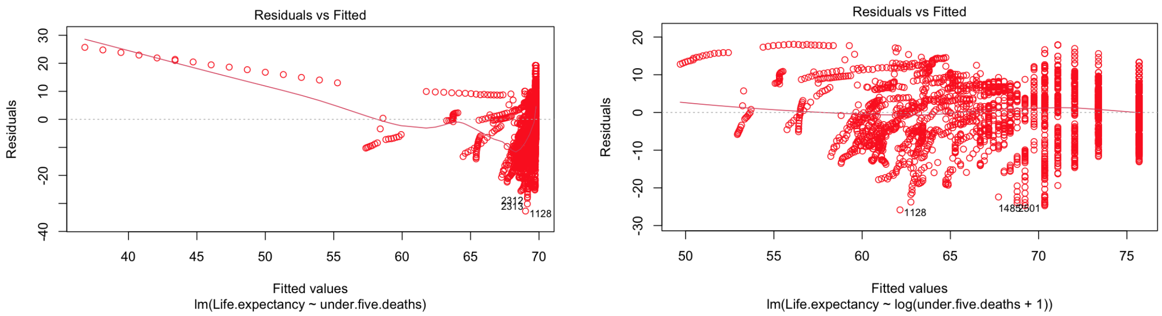


Figure 20

APPENDIX II

I. AIC

A. Function used to calculate AIC on all models

```
calcAIC <- function(actual, predicted, parameters){

  resids = actual - predicted

  n = length(predicted)

  sse = sum(resids^2)

  AIC =  n * log(sse/n) + 2*(parameters + 1)

  print(return(AIC))

}
```

II. KNN Regression

A. Précis of KNN Regression continued

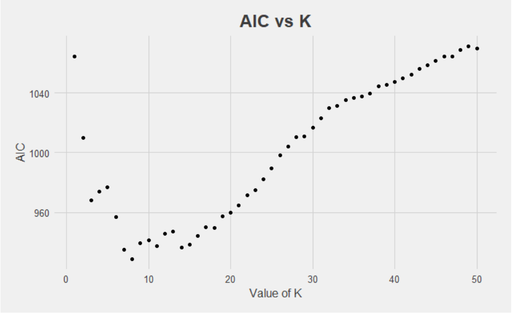
1. The Bias/Variance trade off in KNN regression is relative to the value of K. When K=1 the model will perfectly track the independent variables. As the value of K increases the model will produce a smoother fit. Parametric models will tend to outperform non parametric models like KNN; when the parametric form that has been selected is close to the true form data. Over simplification is non parametric models tend to handle non-linearity well.

   Overfit in KNN can occur when the value of K is too low for the reasons discussed above. Adding predictors can create a spatial dimension problem. When we go from 1 predictor to 2, the spatial calculations leave 1 dimension and enter 2 dimensions. Increasing variables increases dimensions and can result in problems of observations not having near neighbors.

B. Recursive Backward Elimination Feature Selection with Random KNN

1. Backward elimination starts with all features and variables, testing it with the dependent variable (Life Expectancy). Under a selected fitting model of criterion of K nearest neighbor of 5.
2. K = 5 was chosen based on model performance results of AIC and RMSE from a range of K of 1 through 50.

C. Optimizing the Hyperparameter



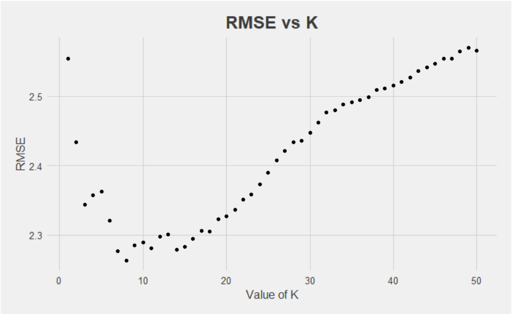


Figure 7

D. KNN Regression Observed vs predicted

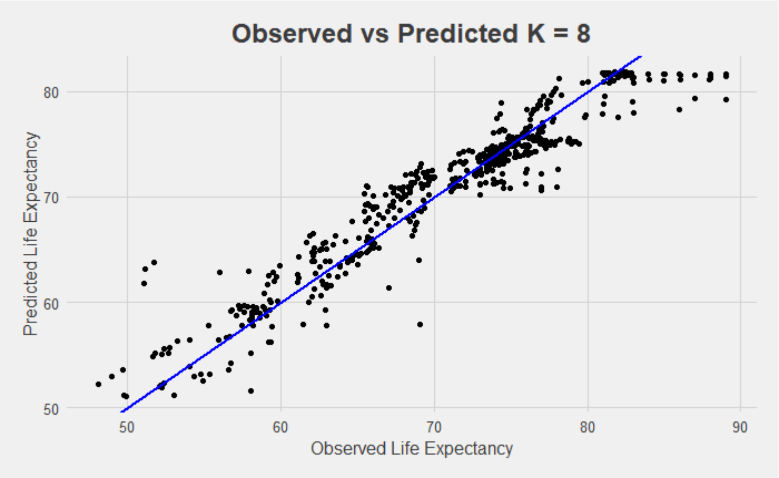


Figure 8

Table 2 continued

ID	*Model*	*RMSE*	*AIC*
1	*K = 8 (5 variables)*	*2.26*	*929.13*
2	*K = 11 (5 Variables)*	*2.28*	*937.60*
3	*K = 6 (all variables)*	*2.57*	*1100.587*
4	*K= 5 (all variables)*	*2.58*	*1102.89*

For all code used in EDA, data prep and modeling, visit our github repository at:

https://github.com/JosephLazarus/Life_Expectancy