**Purpose**

Watching my parents age, in addition to getting older myself, I have been interested in methods to extend life, in a quality manner. I have read articles about blood pressure control, eating habits, effects of anxiety and depression, quality and quantity of sleep, effects of medications and exercise. It seems everyone has the magic bullet, or at least a piece to the puzzle for becoming “younger” and living longer.

Over the years, many have studied, collectively, some of the known and measurable variables on longevity of life. For instance, a study was conducted in the early 2000s by the University of Colorado, to validate and expand upon previous studies1. Rather than look at variables affecting an individual (like the ones I was researching for myself), they looked at variables on the country/world-wide level.

In this study, I would like to answer the following questions:

1. What are top factors affecting life expectancy?
2. With what accuracy can we predict a country's life expectancy based upon these factors?
3. Is it possible to predict if a population will have a life expectancy >= 65?

**Study Outline**

Coding was completed in R using RStudio IDE. The following defines the process used to complete the study:

1. Load and Clean Data
2. Plot and Visualize Data (from 2014 only)
3. Scale/Normalize Data
4. Perform Exploratory Analysis
   1. Scatter Plot Matrix of Scaled Data
   2. Calculate Descriptive Statistics
   3. Corrplot() Analysis
   4. Create Training and Testing Data
5. Unsupervised Modeling: Perform Clustering and Dimension Reduction
   1. Perform Kmeans Iteration
   2. Elbow Plot: Determine Number of Clusters
   3. Perform PCA: Verify Cluster Number
   4. Create Covariance and Correlation Matrices
   5. Select the Number of Factors
      1. Plotting Factors and Loadings
      2. Extract Insights
6. Bi plot of Loadings and Component scores
   1. Supervised Modeling
      1. Perform Multivariate Linear Regression Modeling
      2. Obtain Predictor Correlations
      3. Perform Linear Model (training data, top 3 +/- correlated predictors)
      4. Unrestricted Model: Perform Linear Model (training data, all predictors)
      5. Detecting Multicollinearity: Valuation Inflation Factor
      6. Predictor Selection Using Akaike Information Criteria (AIC)
      7. Run Final Training Model
      8. Run Testing Model
      9. Visualize Linear Model Results
7. Classification Modeling
   1. Load, Clean and Scale Data
   2. Perform Generalized Linear Model
      1. Convert Dependent Variable to Type Factor
      2. Run Model / Obtain Log(odds)
      3. Define Classification Rule / Make Predictions
      4. Create Confusion Matrix and Statistics - Compare Predicted Classes to Actual

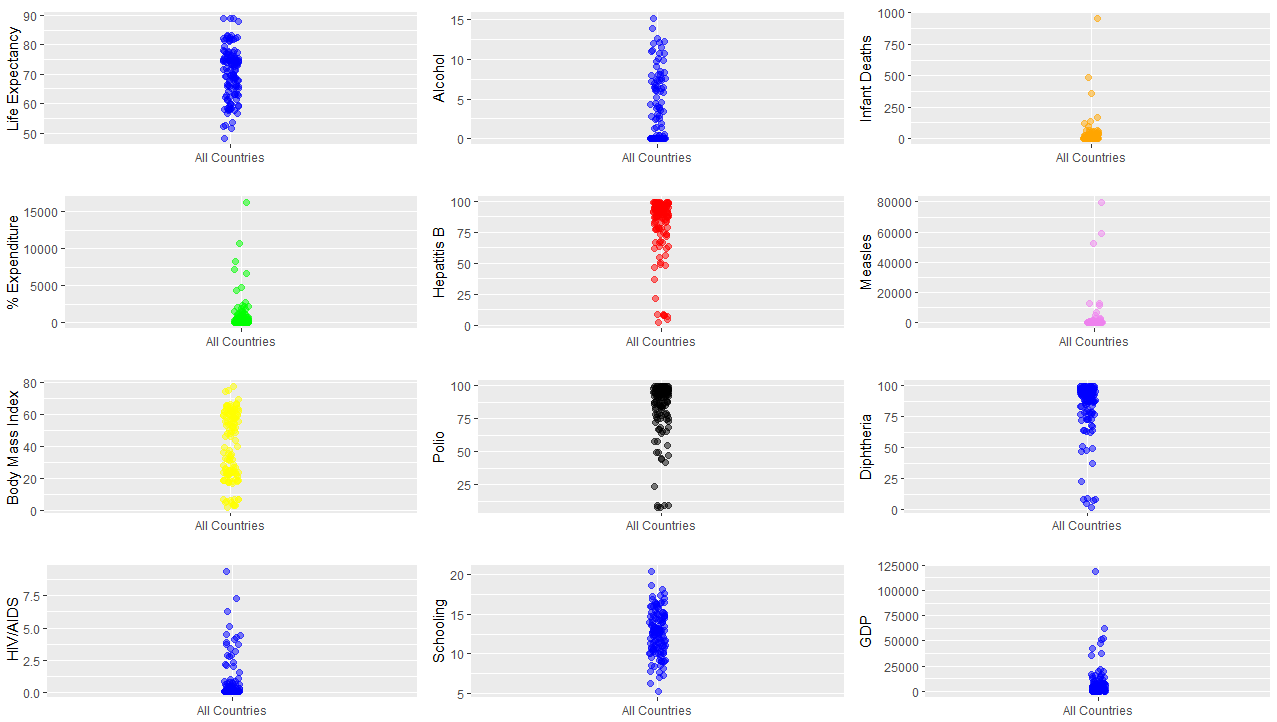
**Dataset Description**

I was unable to find a quality dataset that included personal factors that might affect an individual’s life span. Rather, I obtained population level data compiled by the World Health Organization.

The data follows specific 22 features for 184 countries from 2000-2015. The following table defines the dataset:

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

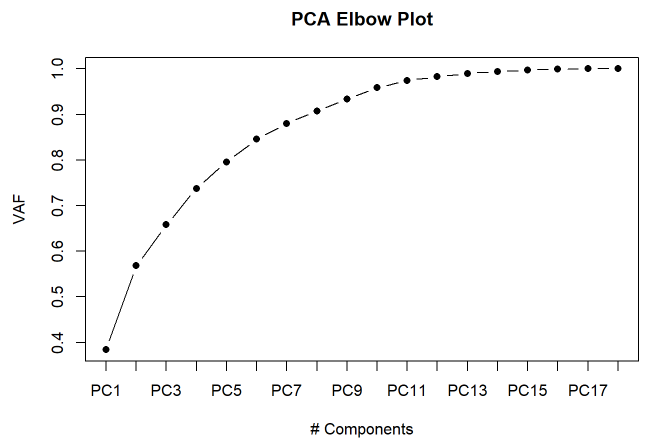
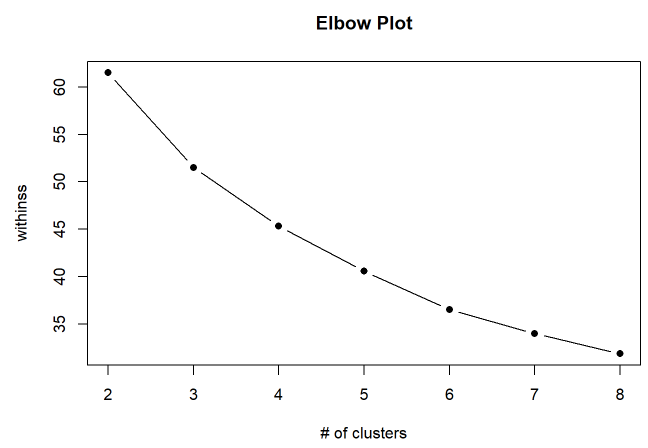
Data was loaded into the model via a csv file and cleaned of empty data. For simplicity, data from 2014 was extracted and modeled. A visualization of a portion of the data is shown below:



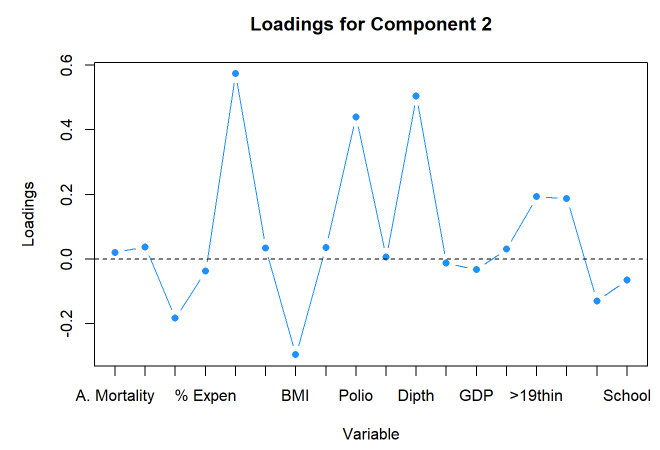
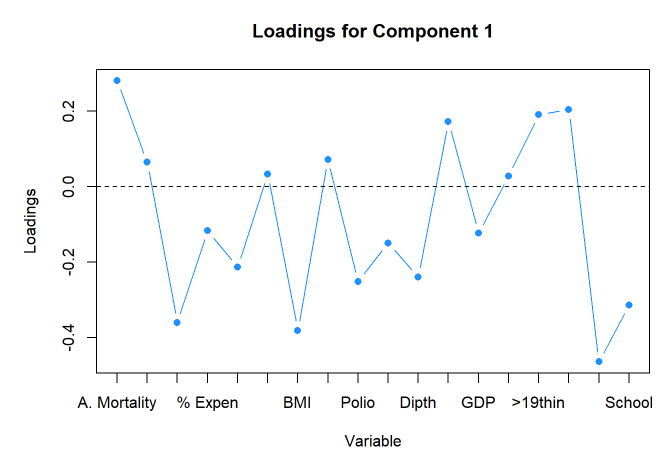
Prior to modeling, the data was normalized to rid of potential size bias. Descriptive statistics were developed for 19 features. In addition, a corrplot() function was created to aid in data visualization. This function is a “visual exploratory tool on correlation matrix that supports automatic variable reordering to help detect hidden patterns among variables”. Data exploratory analysis results can be found within the programming model.

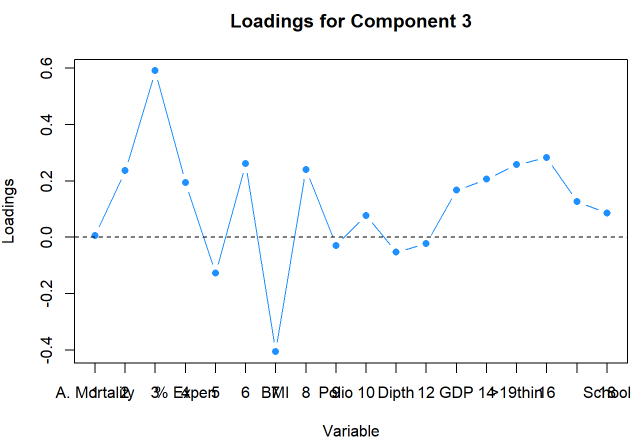
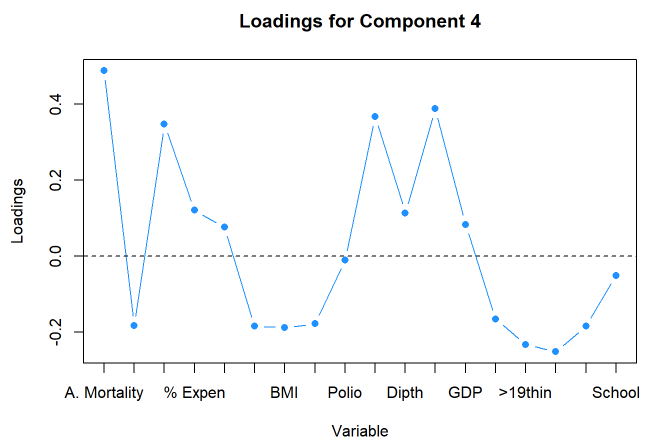
**Unsupervised Analysis**

Both Kmeans and Principal Component Analysis was performed on the full set of normalized 2014 data. It was difficult to make a definitive decision of number of clusters based upon the elbow curves. Arguments could be made to choose anywhere from 2 to 6. In this study, dimension reduction was chosen from 18 features down to 4. With this level of reduction, 84.6% of the variance was still accounted.

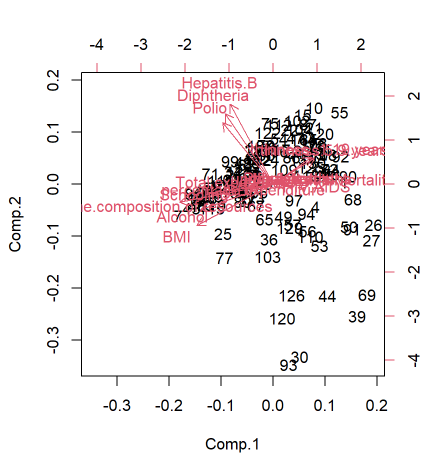


Component loading graphs were created to better understand the relationship between the Loading and Variables.



A biplot was created for the relationship between the first 2 components. Although the map shows 3 distinct areas of life expectancy facors, the countries association with them is indistinguishable. It was difficult to gleen additional information from this tool.



**Supervised Analysis**

Data was split into training and testing sets prior to performing multivariate linear regression. Reduced factor results from PCA were not input into the LR model. Rather, multiple methods were used to determine which factors to include.

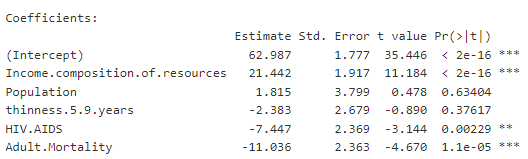
A correlation vector was developed to determine the relative correlations to life expectancy for each predictor. As a baseline model to determine effect on life expectancy, the top 3 positive and negative factors were input into the linear model, using training data.

For comparison, a second training model was performed using all predictors. Coefficients, p-values, r2 vales and sum of squares were evaluated and compared.

Using the Variance Inflation Factor for each predictor, multicollinearity was evaluated. From this analysis, 2 variables (infant.deaths and under.five.deaths) were found to have highly inflated variance and therefore removed from the model.

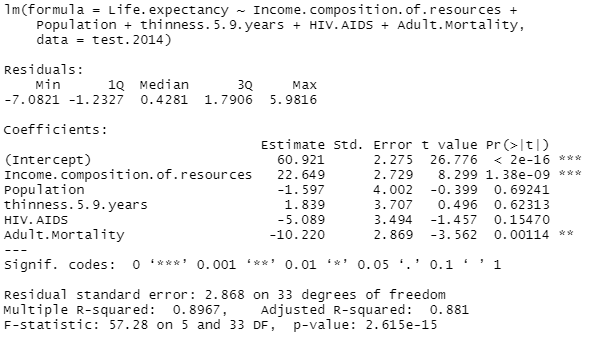
Akaike Information Criteria was incorporated for predictor selection. The lowest AIC iteration resulted in the following model:





The results seemed reasonable. However, future model would perhaps not include Population due to the large P-value. The overall model had and adjusted R-squared of 0.86 with a P-value less than 2.2e-16.

The same model was performed with the holdout testing data and resulted in similar results.



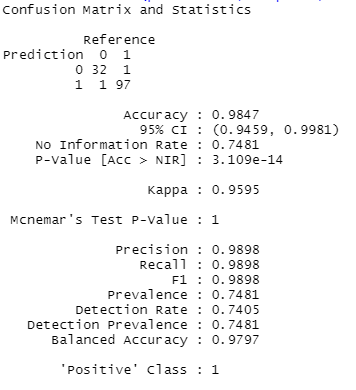
Results were tabulated and are presented in the following graphs. Note the linear model is not precise at describing real world behavior as some of the data points are 5 deviations from the model. Although not perfect, the Density plot comparing the model to actual data aligned fairly well.



**Classification Modeling**

The data set did not include categorical variables. Therefore, I created a column detailing (0 or 1) whether or not life expectancy >= 65 for each country.

Again, the data was loaded, cleaned and scaled. Since the dependent variable was categorical, rather than numerical, it was converted into a type factor. A Generalized Linear Model was run, using the same factors as in the linear model. The model results were used to create the following confusion matrix:



These results demonstrate that the chosen predictors are exceptional at determining if a country would have a life expectancy greater than 65 years.

**Improvements and Future Work**

1. Use larger data set to train the model rather than simply using data from a single year (2014).
2. Perform different types of models to determine whether or not linear model is the most appropriate vehicle for this analysis.
3. Use outputs from PCA analysis as inputs into the linear model. I did not understand how to clearly interpret the results from such a model – with combined and rotated predictors.