

Analysis of Life Expectancy Data Set

STAT 4355.001

May 13, 2021

Team Name : VMC

VMC Team Members: Caleb Captain, Vismaya Joseph, Muhammad Munir

Data Description

Several factors have an effect on our life expectancy, such as income, mortality rates and demographic variables. These factors have been studied before however to increase the spectrum of variables, we have included immunization and human development index both of which extremely important factors were not accounted for before. The variables are distributed across 193 countries across a span of 15 years (2000-2015), yielding in 2,985 data points.

The data is derived from The Global Health Observatory (GHO) data repository under World Health Organization (WHO). Only the most critical factors which are more representative were chosen. As our health and living quality improves there has been a huge improvement in the mortality rates of developing countries, therefore the goal of this data set is to outline those correlations that have helped improve our life expectancy.

The dataset consists of 22 Columns outlined below in the predictor / response variable section and 2,938 rows with 20 predicting variables. All predicting variables are divided into several broad categories including but limited to: Immunization related factors, Mortality factors, Economical factors, and Social factors.

Predictor/Response Variables:

For our data set we have a large number of predictor values and they include the following:

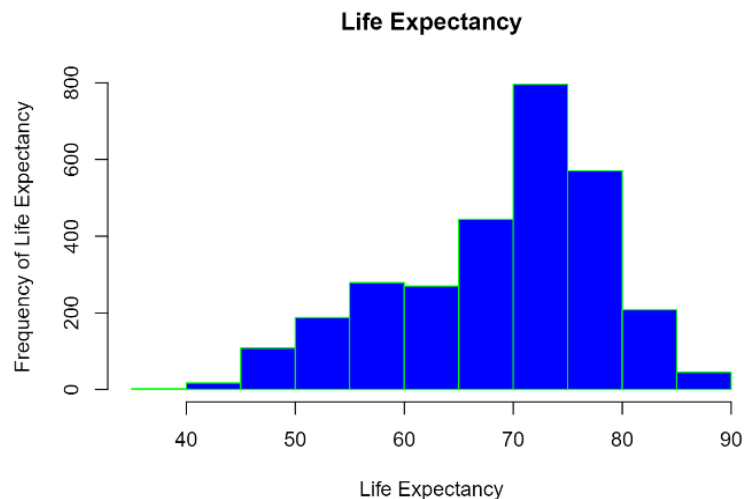
1. Country
2. Year
3. Status
 - a. Developed or developing
4. Life Expectancy
 - a. Life expectancy in age
5. Adult Mortality Rate
 - a. Mortality rate of both genders between 15 and 60 years per 1000
6. Infant death count
 - a. Number of infant deaths per 1000 population
7. Alcohol level
 - a. Per capita recorded alcohol consumption in litres
8. Percentage Expenditure
 - a. Health expenditure percentage of gross domestic product per capita
9. Hepatitis B
 - a. Immunization coverage for Hepatitis B among 1 year olds
10. Measles
 - a. Number of measles cases per 1000 population

11. Body Mass Index average
 - a. Average BMI of population
12. Under 5 deaths
 - a. Number of deaths under age 5 per 1000 population
13. Polio
 - a. Immunization coverage for Polio among 1 year olds
14. Total Expenditure
 - a. Government spending on health as percentage of total spending
15. Diphtheria
 - a. Coverage of diphtheria among 1 year olds
16. HIV/AIDS
 - a. Deaths per 1000 population from HIV/AIDS from 0-4 years
17. GDP
 - a. Gross domestic product per capita
18. Population
19. thinness 10-19 years
 - a. Prevalence of thinness in adolescents
20. thinness 5-9 years
 - a. Prevalence of thinness in children
21. income composition
 - a. Human development index in terms of income composition
22. Schooling
 - a. Average number of years of schooling

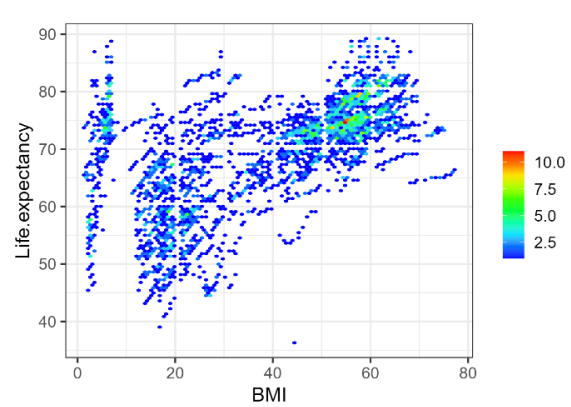
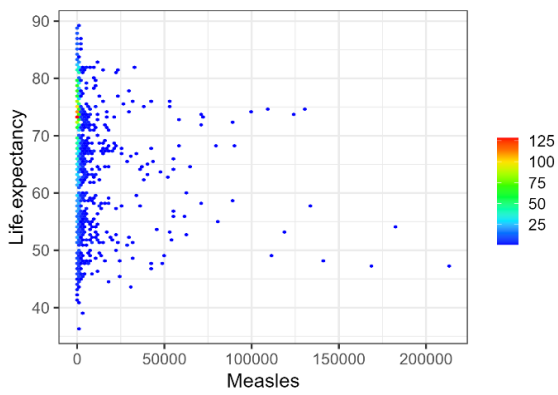
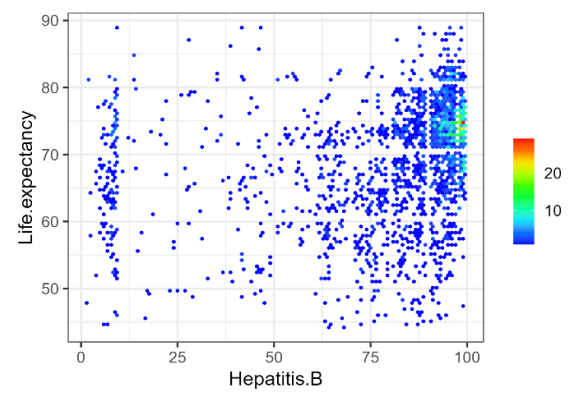
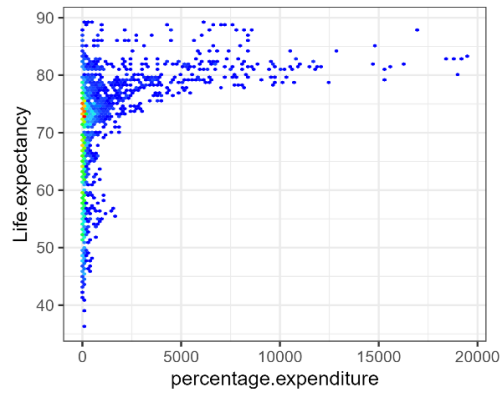
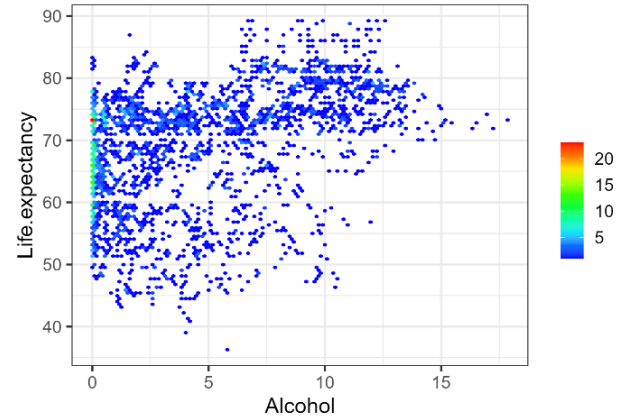
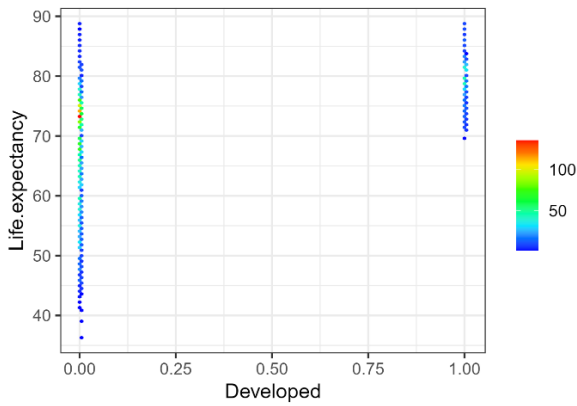
From these values we will determine if there is a correlation or association to the response variable which is: life expectancy rate.

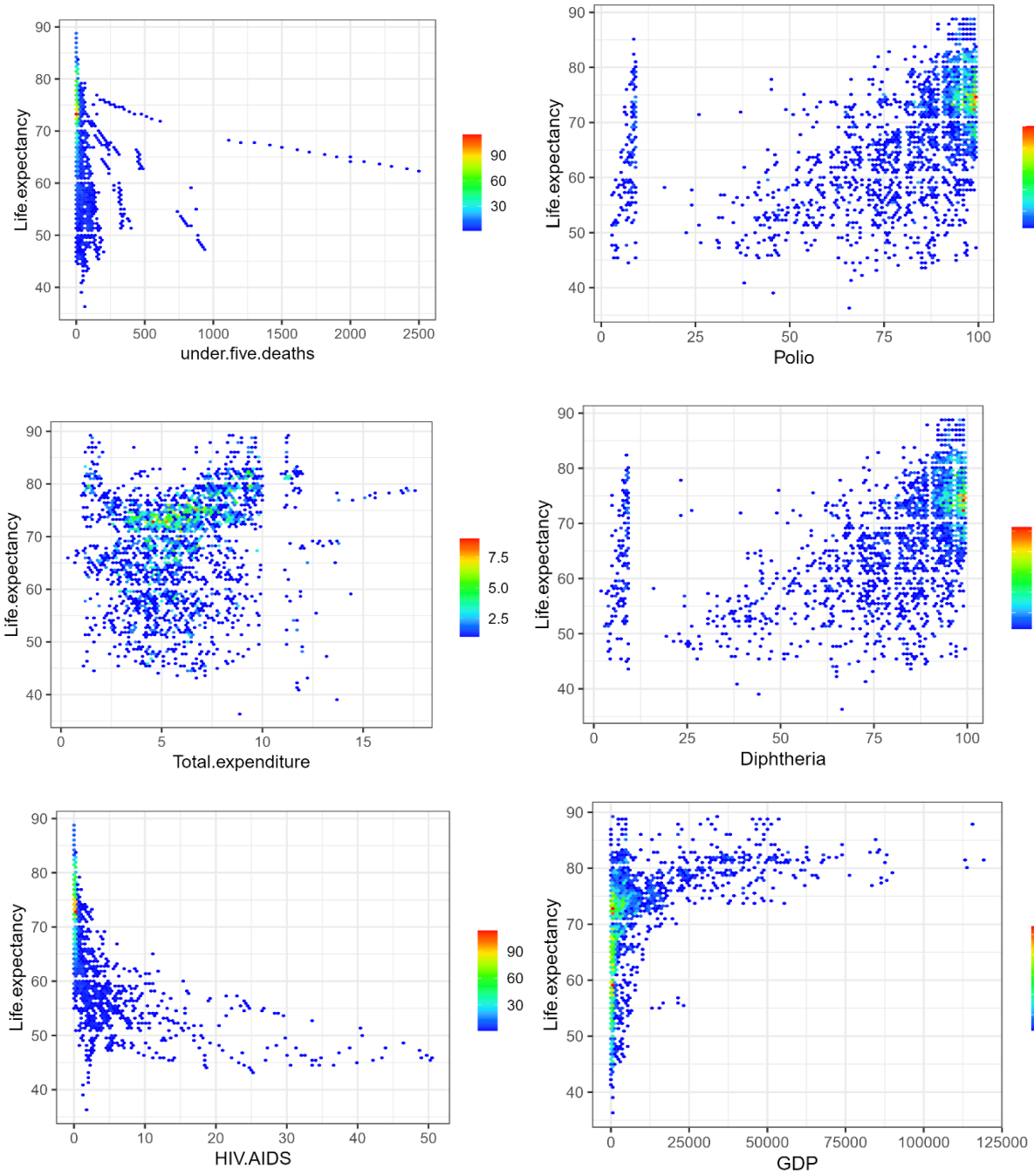
Histogram of Life Expectancy response variable:

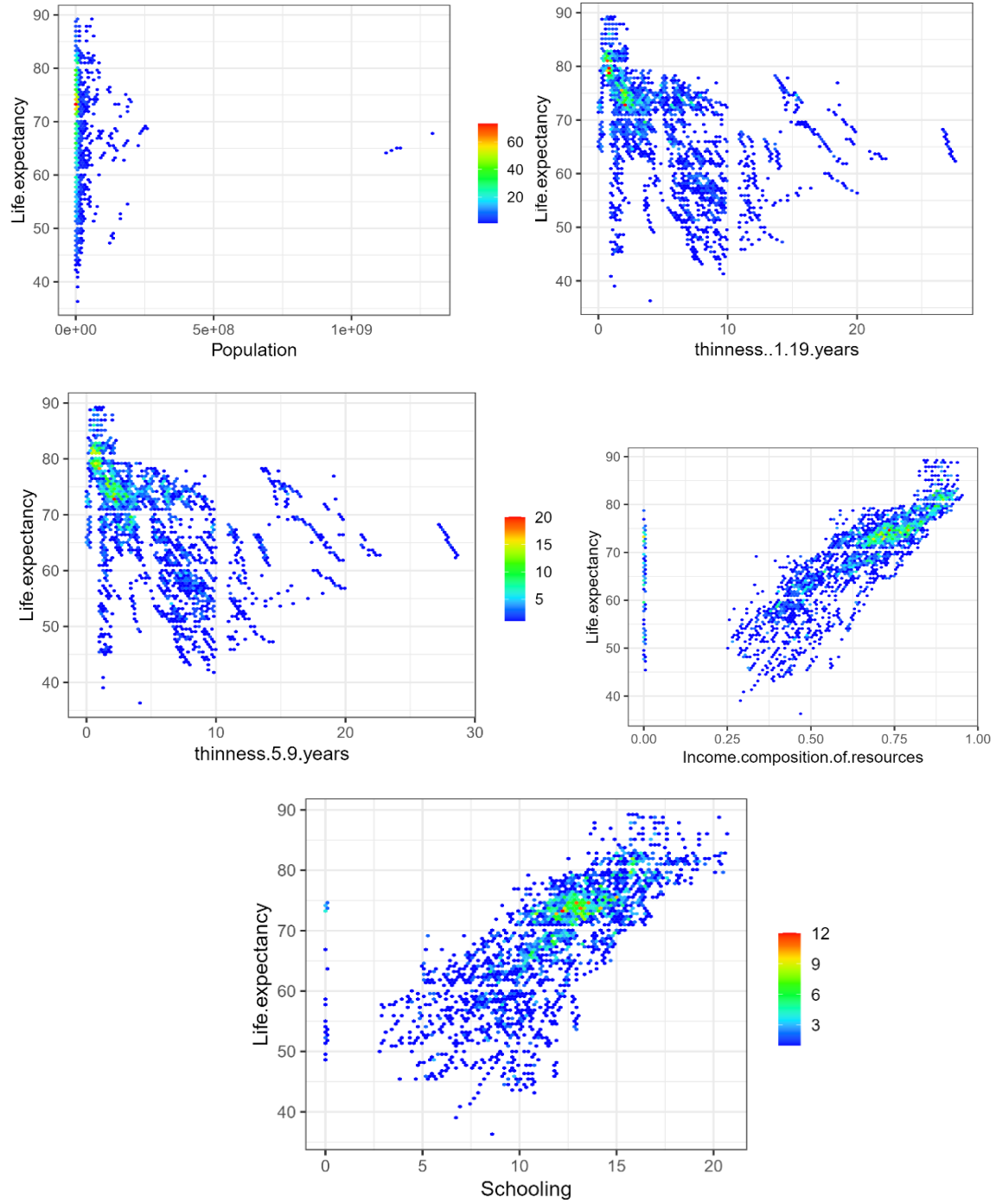
As shown by the histogram, most countries have a life expectancy of around 65 - 80 and the distribution is mostly normal but has a right skew.



Scatterplots of Variables vs. Life Expectancy







Data Analysis

Data Analysis Goal:

The analysis goal is to discover and create the relevant features needed to train a multiple linear regression model for predicting the average life expectancy. To find relevant features, exploratory data analysis is required. Some of the steps in the analysis process include but are not limited to: finding out how many null values are in each column, having an accurate count of how many unique values are in each feature column, graphing and discovering what types of distributions each feature column has, performing hypothesis testing to see which features are most relevant. During this entire process, graphing and display will be required to justify any necessary changes to the data. After the data is properly cleaned, hypothesis tested, and split, we can then use multiple linear regression on the data.

Data Analysis Plan:

The data analysis plan is to first find the relevant statistics of the data as well as displaying the data in the form of graphs. The quantiles and distributions of the data will be useful to see how the data points are spread. Graphing the data will also help with this as well and show us the graphical distribution of the data. Both of these processes will be used to find another useful metric: how many null items exist in the data. If there are null items in feature columns, we can find the mean value and use that to fill in the null items. Another piece of useful information that can be found in the statistical analysis is outliers of the dataset and getting rid of those as well. After the data has been cleaned, we can do hypothesis testing on the data and figure out the correlation values of each feature column with respect to the target column: average life expectancy. Finally, we can train the multiple linear regression algorithm on the newly cleaned data. From there, we can gather model metrics, error rates, R Squared, Adjusted R Squared, etc.

Data Cleaning:

The data we used needed ample amounts of cleaning. There were 1,293 rows with null values. We proceeded to drop all of the rows with null values, leaving us with a data frame of dimensions 1,645 rows by 18 columns.

Adult mortality is a feature, however, that would be highly correlated to life expectancy, and as such could cause over fitting. Infant mortality, Country, and Year were dropped due to not having

uses in the regression. The target is to find a linear relationship to Life Expectancy based on relevant numerical values, thus these three values would be detrimental and/or unnecessary.

The feature column denoting Development Status was binarily factored, as the only two options for the column were Developed and Developing.

Features used in the fitting of Model 1: Status, Alcohol, Percentage Expenditure, Hepatitis B, Measles, BMI, Under Five Deaths, Polio, Total Expenditure, Diphtheria, HIV/AIDS, GDP, Population, Thinness: 10-19 years, Thinness: 5-9 years, Income Composition of Resources, and Schooling.

Features used in the fitting of Model 2: Status, Alcohol, Percentage Expenditure, Hepatitis B, BMI, Under Five Deaths, Polio, Total Expenditure, Diphtheria, HIV/AID, Income Composition of Resources and Schooling.

We also tried to implement a strategy where we replaced in all the null values with in our data with means of the column so that they would simply appear as one point on the model and not disturb the analysis, however we decided that the null column removal and null row removal was a better strategy, as our accuracy of the model had a more significant increase.

Model Fitting

In our initial fit, we regressed across all the variables provided to us. We quickly noticed several variables did not correlate with life expectancy and removed them from our model in the variable selection process. We removed infant and adult mortality from our model as those are used to calculate the life expectancy and would have had a correlation with it regardless.

Initial Model

```

call:
lm(formula = Life.expectancy ~ Status + Alcohol + percentage.expenditure +
  Hepatitis.B + Measles + BMI + under.five.deaths + Polio +
  Total.expenditure + Diphtheria + HIV.AIDS + GDP + Population +
  thinness..1.19.years + thinness.5.9.years + Income.composition.of.resources +
  Schooling, data = data1)

Residuals:
    Min       1Q   Median       3Q      Max
-16.9057  -2.5792   0.1306   2.6794  13.4824

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.662e+01  7.377e-01  63.194 < 2e-16 ***
Status       1.591e+00  3.771e-01   4.220 2.58e-05 ***
Alcohol      -2.211e-01  3.637e-02  -6.080 1.50e-09 ***
percentage.expenditure  4.512e-04  2.018e-04   2.236 0.02550 *
Hepatitis.B  -1.018e-02  4.991e-03  -2.040 0.04149 *
Measles       4.373e-05  1.397e-05   3.130 0.00178 **
BMI           4.631e-02  6.698e-03   6.915 6.71e-12 ***
under.five.deaths -3.049e-03  1.047e-03  -2.912 0.00364 **
Polio         1.455e-02  5.762e-03   2.525 0.01166 *
Total.expenditure  8.881e-02  4.562e-02   1.947 0.05173 .
Diphtheria    2.148e-02  6.616e-03   3.247 0.00119 **
HIV.AIDS     -5.967e-01  1.759e-02 -33.930 < 2e-16 ***
GDP           3.813e-06  3.180e-05   0.120 0.90458
Population    3.119e-09  2.294e-09   1.359 0.17425
thinness..1.19.years  7.317e-03  5.938e-02   0.123 0.90194
thinness.5.9.years -5.557e-02  5.844e-02  -0.951 0.34186
Income.composition.of.resources  1.233e+01  9.225e-01  13.363 < 2e-16 ***
Schooling     1.011e+00  6.615e-02  15.279 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.021 on 1626 degrees of freedom
Multiple R-squared:  0.7926,    Adjusted R-squared:  0.7905
F-statistic: 365.6 on 17 and 1626 DF, p-value: < 2.2e-16

```

There are several variables with p-value greater than the significance value 0.05. These variables were discarded while selecting our reduced model.

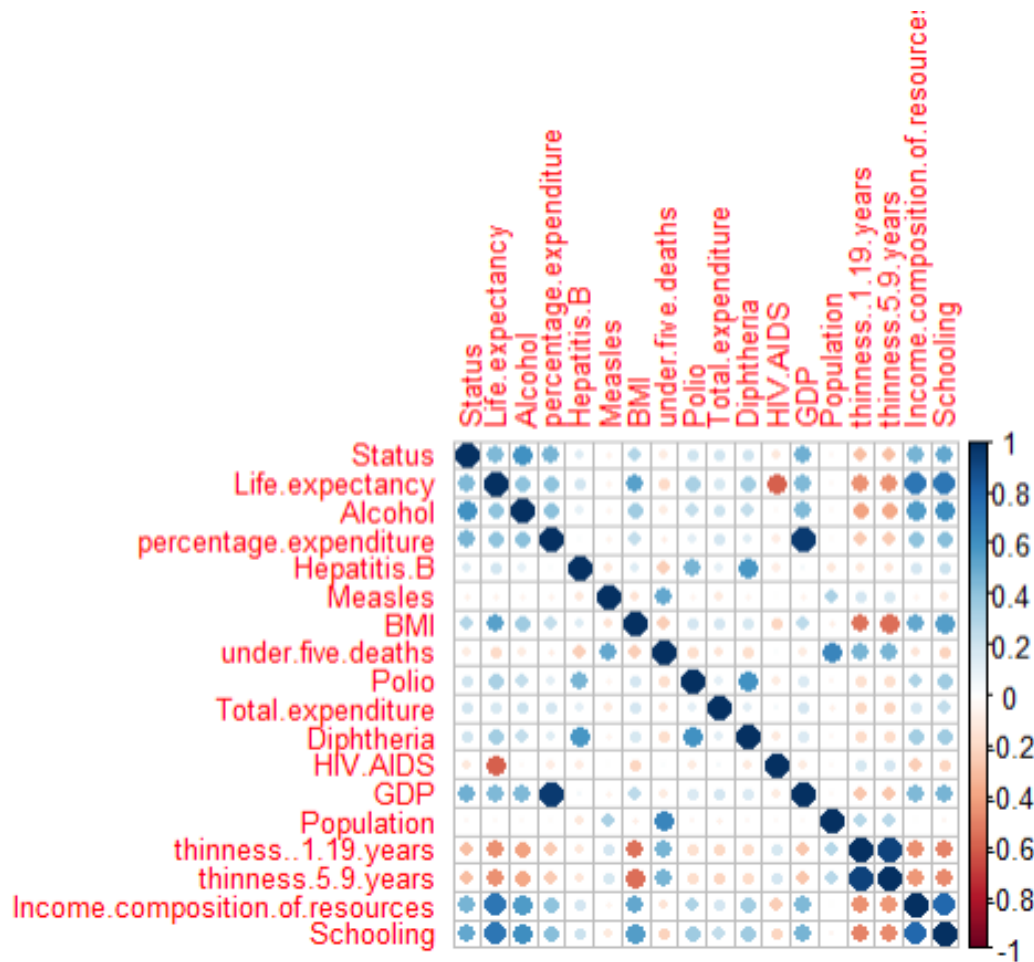
```

> var(mod1)
              Status              Alcohol      percentage.expenditure
              1.815203              2.184843              12.838936
Hepatitis.B              Measles
              1.649422              1.423641              1.778719
under.five.deaths              Polio      Total.expenditure
              2.586541              1.701033              1.117296
Diphtheria              HIV.AIDS      GDP
              2.068821              1.143776              13.565135
Population      thinness..1.19.years      thinness.5.9.years
              1.682323              7.371159              7.304540
Income.composition.of.resources      Schooling
              2.902628              3.479018

```

The multicollinearity test identifies several variables that could be potentially correlated, such as GDP, Percentage-Expenditure, Thinness in 5 - 9 years (children) and Thinness in 10 - 19 years (adolescence). We came up with the conclusion that a higher GDP would result in a higher Percentage Expenditure on Health as there is more room for money to be spent after the bare necessities are taken care of. Also the correlation between Thinness in children and adolescence is correlated majorly because if a kid is thin it is likely to carry over to adolescence.

Variable Selection



As a sanity check we used the above diagram to visualize variables with low correlation and removed them from our reduced model, coincidentally we were able to get rid of all of the variables that seemed to exhibit multicollinearity.

Reduced Model

```

Residuals:
    Min       1Q   Median       3Q      Max
-16.9550  -2.5406   0.1157   2.7308  13.7547

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.604e+01  6.627e-01  69.469  < 2e-16 ***
Status       1.592e+00  3.763e-01   4.231  2.45e-05 ***
Alcohol      -2.143e-01  3.585e-02  -5.978  2.76e-09 ***
percentage.expenditure  4.777e-04  6.659e-05   7.175  1.10e-12 ***
Hepatitis.B  -1.077e-02  4.981e-03  -2.163  0.030712 *
Measles      4.538e-05  1.382e-05   3.283  0.001049 **
BMI          5.052e-02  6.239e-03   8.098  1.08e-15 ***
under.five.deaths -2.813e-03  7.977e-04  -3.526  0.000433 ***
Polio        1.453e-02  5.747e-03   2.528  0.011557 *
Total.expenditure  9.326e-02  4.544e-02   2.053  0.040275 *
Diphtheria   2.194e-02  6.602e-03   3.323  0.000909 ***
HIV.AIDS     -6.000e-01  1.749e-02 -34.299  < 2e-16 ***
Income.composition.of.resources  1.244e+01  9.176e-01  13.561  < 2e-16 ***
Schooling    1.019e+00  6.566e-02  15.513  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.022 on 1630 degrees of freedom
Multiple R-squared:  0.792,    Adjusted R-squared:  0.7904
F-statistic: 477.5 on 13 and 1630 DF,  p-value: < 2.2e-16

```

Our reduced model did not contain any variables that did not correlate with life-expectancy. There was no improvement on residual standard error, multiple R-squared and adjusted R-squared.

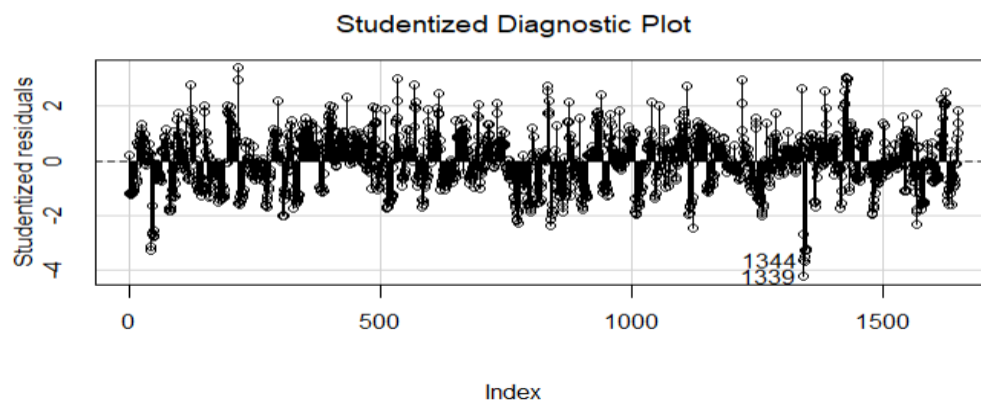
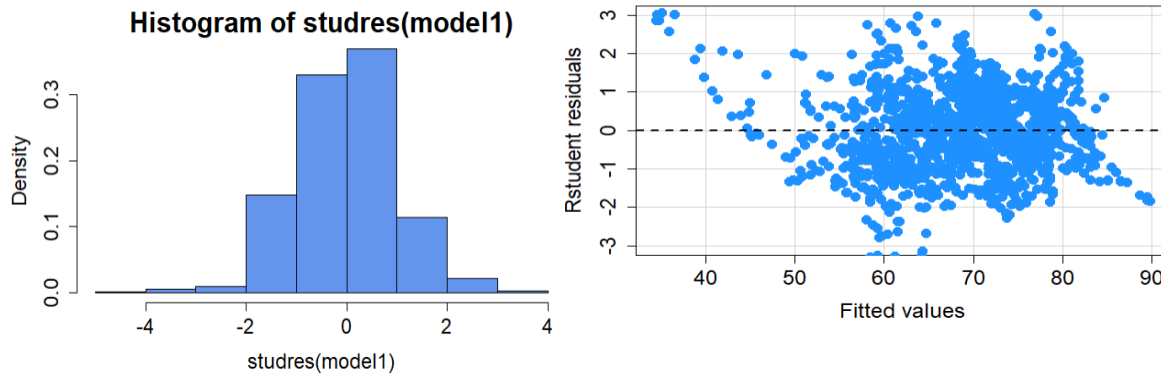
Status	Alcohol	percentage.expenditure
1.806372	2.121359	1.397093
Hepatitis.B	Measles	BMI
1.641936	1.393016	1.542735
under.five.deaths	Polio	Total.expenditure
1.500718	1.691680	1.107988
Diphtheria	HIV.AIDS	Income.composition.of.resources
2.059171	1.131271	2.871253
Schooling		
3.426496		

For the remaining variables the variance influence analysis indicated that there was not any multicollinearity.

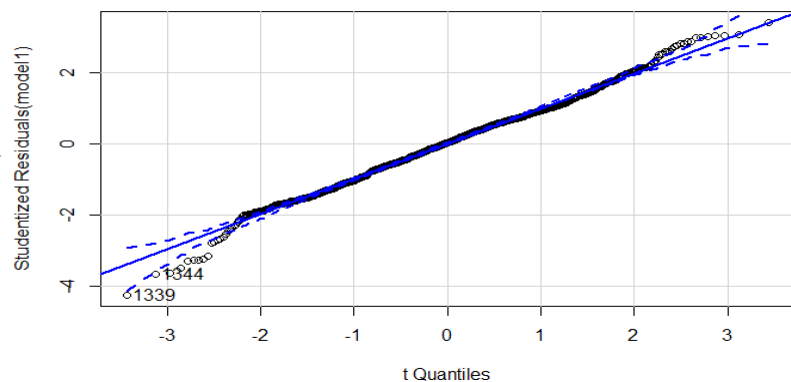
Residual Analysis

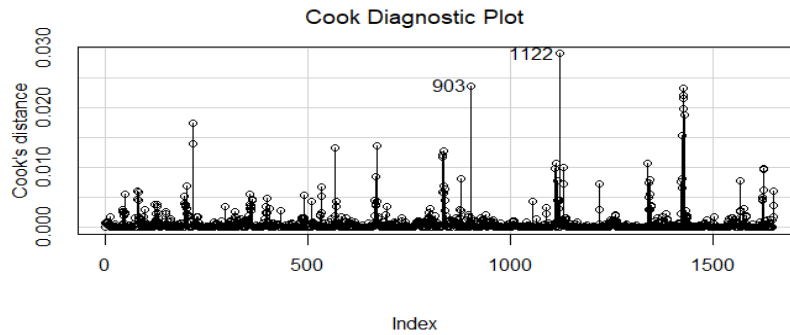
The first model, titled Model 1, Multiple R² value: 79.25%, Adjusted R² value: 79.03%. The second model, titled Model 2, Multiple R² value: 79.20%, Adjusted R² value: 79.04%.

We can see the histogram has a normal distribution and that the Rstudent residual V Fitted values is near 0 and furthermore the Studentized residual has a mean of 0, all these indicate that the residuals are normally distributed and there is not a need to do any transformations.



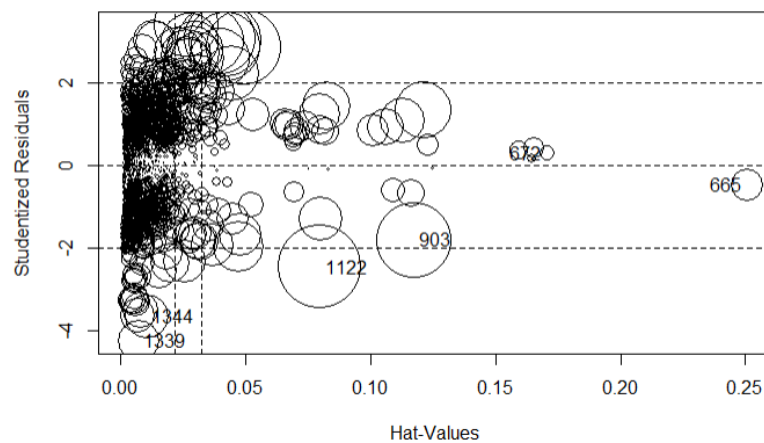
The QQ plot was consistent with our previous analysis and indicated that residuals are normal.





Outlier removal

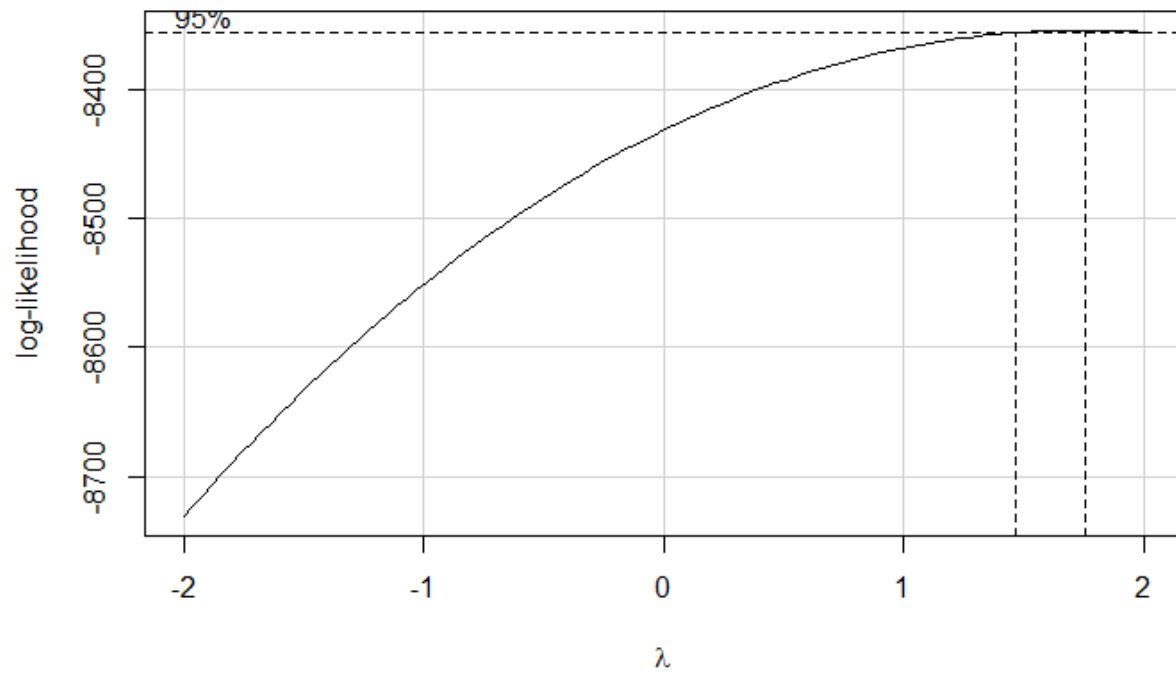
Shown here are outlier graphs for the reduced model. Several outliers were detected and removed from the model to increase the precision. After the removal of the outliers there was an increase in the normality of the residuals.



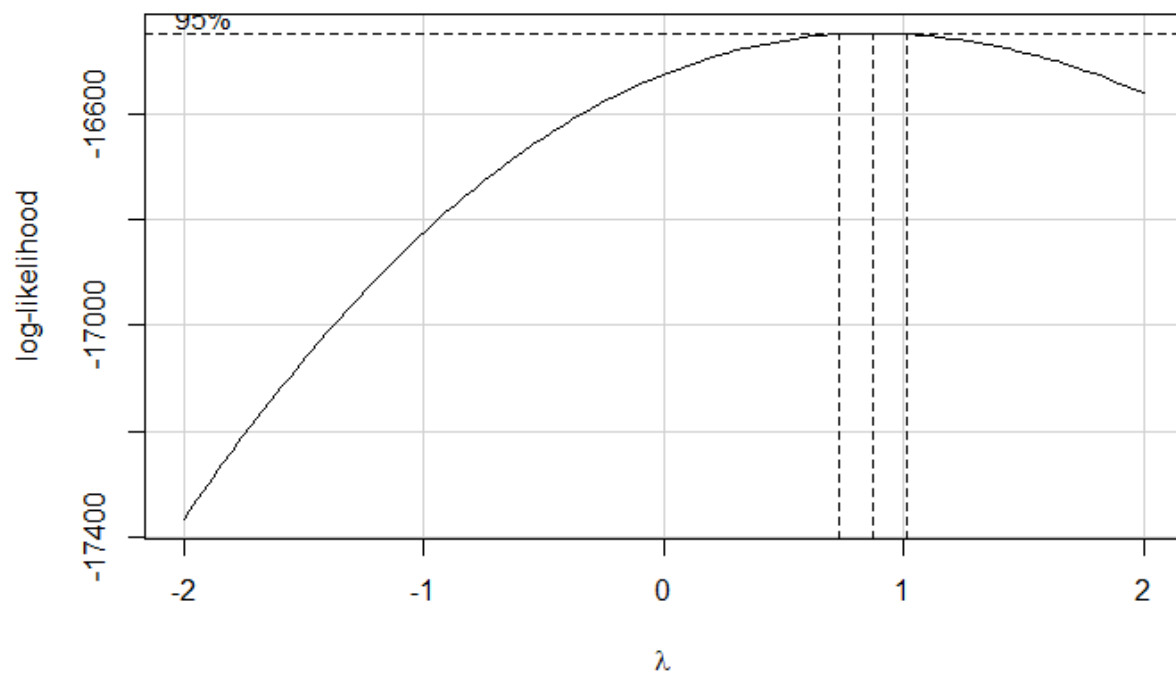
Transformations

As the residuals indicated that the normality assumption was correct, there was no need to perform any transformations. As an experiment we still plotted a boxCox plot which resulted in a lambda of 2, performing the transformation with $\lambda = 2$ we realised that our R-Squared and

Adjusted R-Squared became worse.



After using the suggested transformation our log-likelihood further declined into the negatives indicating a transformation is not the best solution here



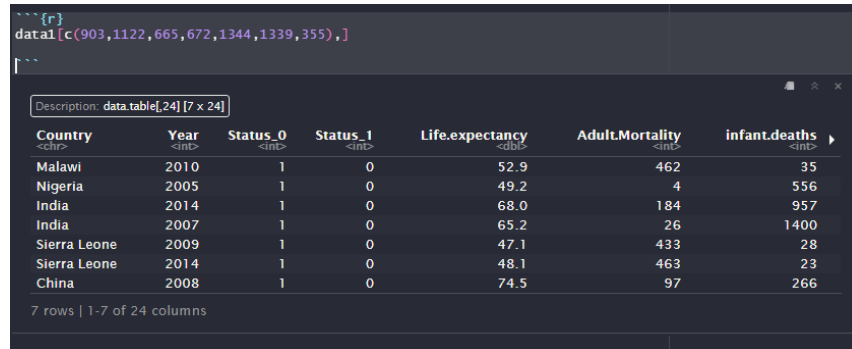
Conclusion

Many countries stood out as outliers after our data analysis and they are the following.

```

***{R}
data1[c(903,1122,665,672,1344,1339,355),]
***

```



Country	Year	Status_0	Status_1	Life expectancy	Adult Mortality	infant deaths
Malawi	2010	1	0	52.9	462	35
Nigeria	2005	1	0	49.2	4	556
India	2014	1	0	68.0	184	957
India	2007	1	0	65.2	26	1400
Sierra Leone	2009	1	0	47.1	433	28
Sierra Leone	2014	1	0	48.1	463	23
China	2008	1	0	74.5	97	266

7 rows | 1-7 of 24 columns

However we were able to come to the conclusion that all of these variables play a very large effect on the response variable of Life Expectancy. Our questions in our analysis goals were as follows:

What relationship, if any, is there between Life Expectancy and monetary features, such as Income Composition of Resources, GDP, Percentage Expenditure?

How does the Education and Development Status of a Country affect the mean Life expectancy?

Does Alcohol Consumption negatively affect Life expectancy?

For our monetary features analysis, there were not very useful results. The only monetary feature that had a high enough correlation to be relevant for Model 2 was Income Composition of Resources. The other two features did not have a revealing correlation.

Education displayed a positive correlation with Life Expectancy. . The number of years of schooling results in an increase in life expectancy as there are more professionals skilled to treat and serve the community. The developed countries as predicted had a higher mean life expectancy than developing countries as they have better resources and more access to better healthcare.

Regarding alcohol consumption, we were given surprising results in which there was a small positive correlation between both variables. The higher the alcohol consumption rate, the higher the life expectancy. This goes against our assumption that higher Alcohol Consumption would be detrimental to overall life expectancy.

Reflections

Regarding our models, we did notice several aspects that could be improved upon. Firstly: missing data. There are 2,563 null values in the data, further broken into 1,289 data entries, which was cause for concern. One method utilized earlier in the project was filling in the data with the mean values, however, this would have several detrimental effects to the data. Population was so sparse that inputting the mean would most likely not help. Several other features were missing similar, if not more data entries.

This led us to the method used in the project: dropping null values. This, too, comes with its benefits and drawbacks. One benefit is that we would be able to use more features for modeling, which led to an overall better fitted model. The drawback was that we'd lose almost half of the data entries because of this, leaving us with 1,649 data entries to work with. A better solution might be to use a mixture of both solutions: dropping columns with too high of null values while keeping some that could be filled in with the column mean. Further exploration into filling in values extrapolated data from other reports could also be used.

An aspect of the project that was beneficial was that, after fitting, the models were of a Normal distribution. This means that transformations were not necessary, however, they could still be used regardless. Outlier detection was fairly straightforward and a simple process with few outliers to remove. Finally, multicollinearity wasn't a factor at all in the data. We were given guidance early on as to which features not to include, such as Adult Mortality and Infant Mortality, which helped greatly in the fitting of both models.

Materials and Appendix

Acknowledgements

- The data was collected from WHO and the United Nations website with the help of Deeksha Russell and Duan Wang.

Members' Roles:

- Caleb Captain: Cleaning, correlation testing the data to find relevant features, and metrics evaluation.
- Vismaya Joseph: Data Description, Understanding the Data and multiple linear regression modeling.
- Muhammad Munir: Graphing, plotting, and transformations on models for improvement.

Reference:

- Kaggle; <https://www.kaggle.com/kumaraajarshi/life-expectancy-who>

