Factors influencing life expectancy

Carson Young

26 November 2022

Outline and thought process

Here are some of my initial thoughts on the steps to analyse this dataset.

- **Data Exploration**- Understanding the variables and correlations between them.
- **Data Cleaning**- Correcting inconsistency or missing data.
- **Feature Selection**- There is a mix of continuous and categorical variables, 22 in total. We need to consider which combination affects life expectancy the most.
- **Models and visualizations**- We could try linear regression and k-means clustering.
- **Mathematical tools**- Could use correlation matrices, model section algorithms, scatter plots and principal component analysis.
- **Data Analysis**- What does the data and models tell us in plain English? What are the key contributors to life expectancy? Do countries of similar life expectancy share similar attributes?

We will utilise R libraries and omit mathematical details for clarity.

It is common to split the data into train and test sets to scrutinise the relevance of models. For brevity, we will only focus on understanding the implications of the data.

Data Exploration

Begin by importing the dataset and constructing basic descriptive statistics.
The dataset contains data from 193 countries, with 22 features. These include economic, education and health factors.

```
life <- read.csv("Life Expectancy Data.csv")
str(life)
```

```
## 'data.frame':    2938 obs. of  22 variables:
##  $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
##  $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
##  $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
##  $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
##  $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
```

```
##  $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
##  $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
##  $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
##  $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.4
##  $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

Year and country are the categorical variables.

```
summary(life)
```

```
##    Country               Year         Status          Life.expectancy
##  Length:2938        Min.   :2000   Length:2938        Min.   :36.30
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:63.10
##  Mode  :character   Median :2008   Mode  :character   Median :72.10
##                     Mean   :2008                      Mean   :69.22
##                     3rd Qu.:2012                      3rd Qu.:75.70
##                     Max.   :2015                      Max.   :89.00
##                                                       NA's   :10
##  Adult.Mortality infant.deaths       Alcohol        percentage.expenditure
##  Min.   :  1.0   Min.   :   0.0   Min.   : 0.0100   Min.   :    0.000
##  1st Qu.: 74.0   1st Qu.:   0.0   1st Qu.: 0.8775   1st Qu.:    4.685
##  Median :144.0   Median :   3.0   Median : 3.7550   Median :   64.913
##  Mean   :164.8   Mean   :  30.3   Mean   : 4.6029   Mean   :  738.251
##  3rd Qu.:228.0   3rd Qu.:  22.0   3rd Qu.: 7.7025   3rd Qu.:  441.534
##  Max.   :723.0   Max.   :1800.0   Max.   :17.8700   Max.   :19479.912
##  NA's   :10                       NA's   :194
##   Hepatitis.B       Measles               BMI        under.five.deaths
##  Min.   : 1.00   Min.   :     0.0   Min.   : 1.00   Min.   :   0.00
##  1st Qu.:77.00   1st Qu.:     0.0   1st Qu.:19.30   1st Qu.:   0.00
##  Median :92.00   Median :    17.0   Median :43.50   Median :   4.00
##  Mean   :80.94   Mean   :  2419.6   Mean   :38.32   Mean   :  42.04
##  3rd Qu.:97.00   3rd Qu.:   360.2   3rd Qu.:56.20   3rd Qu.:  28.00
##  Max.   :99.00   Max.   :212183.0   Max.   :87.30   Max.   :2500.00
##  NA's   :553                        NA's   :34
##      Polio       Total.expenditure   Diphtheria       HIV.AIDS
##  Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100
##  1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100
##  Median :93.00   Median : 5.755    Median :93.00   Median : 0.100
##  Mean   :82.55   Mean   : 5.938    Mean   :82.32   Mean   : 1.742
##  3rd Qu.:97.00   3rd Qu.: 7.492    3rd Qu.:97.00   3rd Qu.: 0.800
##  Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600
##  NA's   :19      NA's   :226       NA's   :19
##       GDP              Population        thinness..1.19.years
##  Min.   :     1.68   Min.   :3.400e+01   Min.   : 0.10
##  1st Qu.:   463.94   1st Qu.:1.958e+05   1st Qu.: 1.60
##  Median :  1766.95   Median :1.387e+06   Median : 3.30
##  Mean   :  7483.16   Mean   :1.275e+07   Mean   : 4.84
##  3rd Qu.:  5910.81   3rd Qu.:7.420e+06   3rd Qu.: 7.20
##  Max.   :119172.74   Max.   :1.294e+09   Max.   :27.70
##  NA's   :448         NA's   :652         NA's   :34
##  thinness.5.9.years Income.composition.of.resources   Schooling
##  Min.   : 0.10      Min.   :0.0000                  Min.   : 0.00
##  1st Qu.: 1.50      1st Qu.:0.4930                  1st Qu.:10.10
##  Median : 3.30      Median :0.6770                  Median :12.30
##  Mean   : 4.87      Mean   :0.6276                  Mean   :11.99
##  3rd Qu.: 7.20      3rd Qu.:0.7790                  3rd Qu.:14.30
```

```
##  Max.   :28.60      Max.   :0.9480                  Max.   :20.70
##  NA's   :34         NA's   :167                     NA's   :163
```

Note down a few key facts

- Mean life expectancy is 69.22
- Mean years of schooling is 12
- Mean adulat mortality rate is 164.8 (per 1000)

Data Cleaning

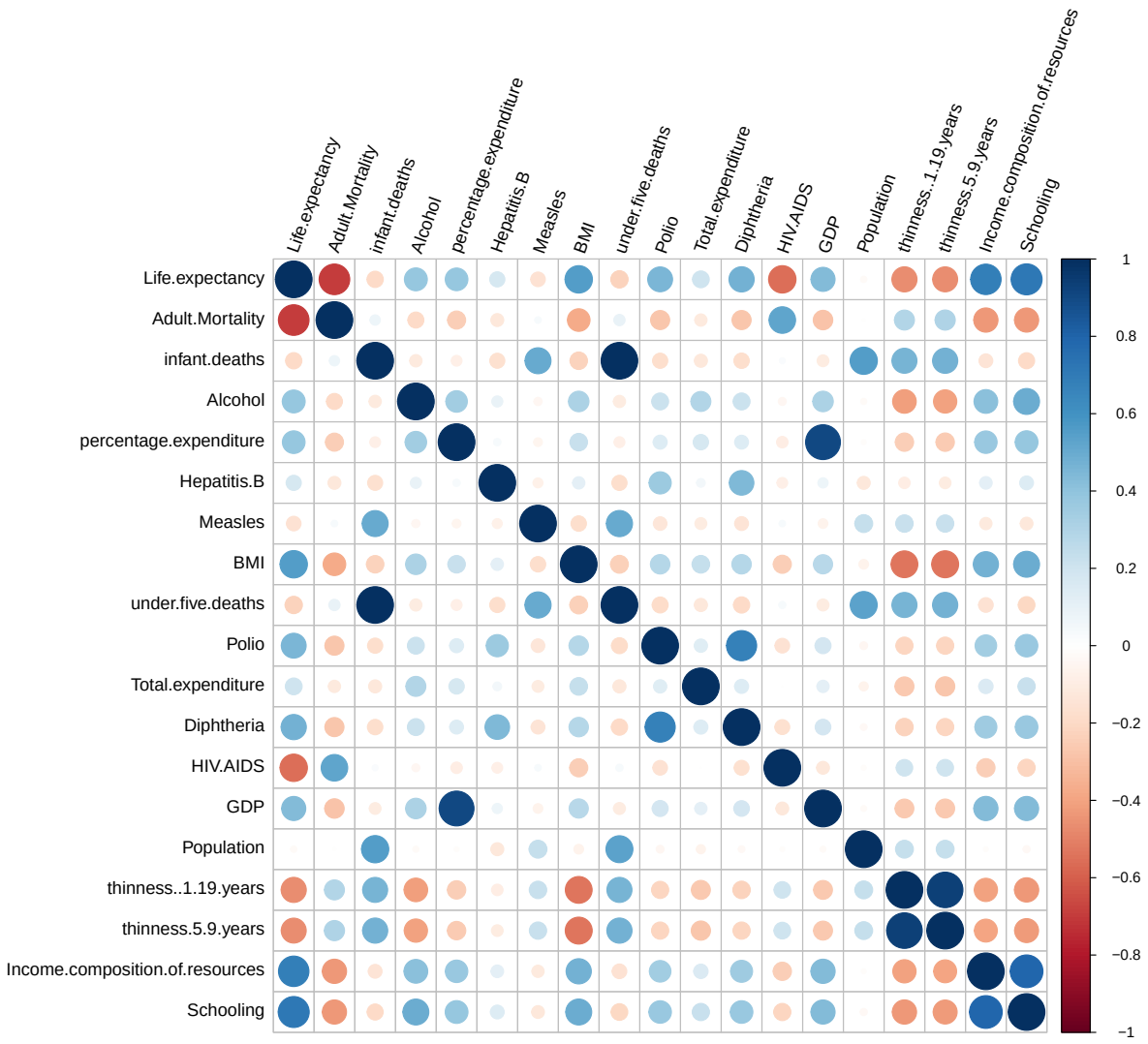
There are missing values in many columns (NA's above). There are many ways to fix this, each with different implications. We shall use a straightforward way- **imputation by median** since the missing values account for less than half the rows.

```
life$Life.expectancy[is.na(life$Life.expectancy)] <-
  +median(life$Life.expectancy,  na.rm = TRUE)
life$Adult.Mortality[is.na(life$Adult.Mortality)] <-
  +median(life$Adult.Mortality,  na.rm = TRUE)
life$Alcohol[is.na(life$Alcohol)] <- median(life$Alcohol,  na.rm = TRUE)
life$Hepatitis.B[is.na(life$Hepatitis.B)] <-
  +median(life$Hepatitis.B,  na.rm = TRUE)
life$BMI[is.na(life$BMI)] <- median(life$BMI,  na.rm = TRUE)
life$Polio[is.na(life$Polio)] <- median(life$Polio,  na.rm = TRUE)
life$Total.expenditure[is.na(life$Total.expenditure)] <-
  +median(life$Total.expenditure,  na.rm = TRUE)
life$Diphtheria[is.na(life$Diphtheria)] <-
  +median(life$Diphtheria,  na.rm = TRUE)
life$GDP[is.na(life$GDP)] <- median(life$GDP,  na.rm = TRUE)
life$Population[is.na(life$Population)] <-
  +median(life$Population,  na.rm = TRUE)
life$thinness..1.19.years[is.na(life$thinness..1.19.years)] <-
  +median(life$thinness..1.19.years,  na.rm = TRUE)
life$thinness.5.9.years[is.na(life$thinness.5.9.years)] <-
  +median(life$thinness.5.9.years,  na.rm = TRUE)
life$Income.composition.of.resources[is.na(life$Income.composition.of.resources)]<-
  +median(life$Income.composition.of.resources,  na.rm = TRUE)
life$Schooling[is.na(life$Schooling)] <- median(life$Schooling,  na.rm = TRUE)
```

Correlation Matrix

The goal here is to identify variables that are correlated.

```
life_numerical = subset(life, select = -c(Country,Status,Year) )
corrmatrix = cor(life_numerical,use = "complete.obs")
corrplot(corrmatrix,tl.srt=70,tl.col="black")
```



The size of the circle corresponds to the level of correlation. Some obvious correlations are expected:

- Infant deaths vs under five deaths
- GDP vs percentage expenditure
- Adult mortality vs life expectancy
- Thinness 1-19 years vs thinness 5-9 years

Strong correlation can cause problems when running regression models. A solution is to drop one variable in each highly correlated pair. The **variance inflation factor** can tell us which variables to drop.

```
head(corrmatrix,1)
```

```
##                 Life.expectancy Adult.Mortality infant.deaths   Alcohol
## Life.expectancy               1      -0.6963901    -0.1967691 0.3889175
##                 percentage.expenditure Hepatitis.B    Measles       BMI
## Life.expectancy              0.3814181   0.1702186 -0.1577666 0.5569012
##                 under.five.deaths     Polio Total.expenditure Diphtheria
## Life.expectancy        -0.2227382 0.4583993         0.2088437  0.4722108
##                   HIV.AIDS       GDP  Population thinness..1.19.years
## Life.expectancy -0.5567034 0.4304613 -0.02901388           -0.4680022
##                 thinness.5.9.years Income.composition.of.resources Schooling
## Life.expectancy         -0.4624732                       0.6886616 0.7130535
```

We expect adult mortality, schooling, income composition of resources, BMI and HIV.AIDS to be the more significant contributors.

Less significant factors include population, measles, hepatitis B and infant deaths.

Infant deaths are worth investigating as we would expect from common sense that a high infant mortality rate lowers life expectancy. A possible explanation is the highly skewed nature of this feature. Around 80% of rows have an infant mortality rate of zero. Hence, it does not contribute much to life expectancy. Reducing child mortality is one of the millennium goals and we expect this figure to decrease further. The same reasoning applies to measles and hepatitis B.

Models and visulisation

Variance inflation factor

Remove one feature from each colinear pair

```
life$Status <- factor(life$Status)
model <- lm(Life.expectancy~.-Country,data=life)
vif(model)
```

```
##                             Year                           Status
##                         1.156924                         1.885164
##                  Adult.Mortality                    infant.deaths
##                         1.744911                       177.439333
##                          Alcohol          percentage.expenditure
##                         1.893620                         5.817572
##                      Hepatitis.B                          Measles
##                         1.312691                         1.382430
##                              BMI                under.five.deaths
##                         1.721053                       176.391427
##                            Polio               Total.expenditure
##                         1.939102                         1.220162
##                         Diphtheria                          HIV.AIDS
##                         2.165262                         1.441125
##                              GDP                       Population
##                         6.035273                         1.491045
##             thinness..1.19.years              thinness.5.9.years
##                         8.775830                         8.874575
## Income.composition.of.resources                         Schooling
##                         3.060504                         3.332427
```

Drop the following features since they have the higher VIF:

- Infant deaths
- GDP
- thinness 5-9 years

Linear regression

Life expectancy is our response variable. The other variables are predictors.

```
model2 <- lm(Life.expectancy~.-Country -infant.deaths -GDP -thinness.5.9.years,data=life)
summary(model2)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ . - Country - infant.deaths -
##     GDP - thinness.5.9.years, data = life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.3308  -2.3225  -0.0477   2.4037  17.2942
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     8.506e+01  3.551e+01   2.395  0.01666 *
## Year                           -1.474e-02  1.775e-02  -0.830  0.40643
## StatusDeveloping               -1.614e+00  2.764e-01  -5.837 5.89e-09 ***
## Adult.Mortality                -2.044e-02  8.128e-04 -25.145  < 2e-16 ***
## Alcohol                         1.486e-02  2.657e-02   0.559  0.57608
## percentage.expenditure          2.656e-04  4.457e-05   5.959 2.85e-09 ***
## Hepatitis.B                    -1.940e-02  3.800e-03  -5.106 3.50e-07 ***
## Measles                        -2.380e-05  7.836e-06  -3.038  0.00240 **
## BMI                             4.648e-02  4.991e-03   9.313  < 2e-16 ***
## under.five.deaths              -1.996e-03  7.027e-04  -2.841  0.00453 **
## Polio                           3.200e-02  4.551e-03   7.032 2.53e-12 ***
## Total.expenditure               6.169e-02  3.473e-02   1.776  0.07583 .
## Diphtheria                      4.801e-02  4.722e-03  10.168  < 2e-16 ***
## HIV.AIDS                       -4.818e-01  1.805e-02 -26.699  < 2e-16 ***
## Population                      3.556e-09  1.703e-09   2.088  0.03685 *
## thinness..1.19.years           -5.520e-02  2.439e-02  -2.263  0.02373 *
## Income.composition.of.resources  6.462e+00  6.470e-01   9.989  < 2e-16 ***
## Schooling                       6.824e-01  4.274e-02  15.967  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.149 on 2920 degrees of freedom
## Multiple R-squared:  0.8108, Adjusted R-squared:  0.8097
## F-statistic:   736 on 17 and 2920 DF,  p-value: < 2.2e-16
```

Select features using stepwise selection

```
model3 <- step(model2, scope=~., direction = "both", trace = FALSE)
summary(model3)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Status + Adult.Mortality + percentage.expenditure +
```

```
##     Hepatitis.B + Measles + BMI + under.five.deaths + Polio +
##     Total.expenditure + Diphtheria + HIV.AIDS + Population +
##     thinness..1.19.years + Income.composition.of.resources +
##     Schooling, data = life)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.2494  -2.3183  -0.0489   2.4169  17.3806
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.564e+01  6.743e-01  82.516  < 2e-16 ***
## StatusDeveloping               -1.699e+00  2.560e-01  -6.636 3.83e-11 ***
## Adult.Mortality                -2.045e-02  8.076e-04 -25.319  < 2e-16 ***
## percentage.expenditure          2.676e-04  4.450e-05   6.014 2.04e-09 ***
## Hepatitis.B                    -1.933e-02  3.798e-03  -5.089 3.82e-07 ***
## Measles                        -2.331e-05  7.821e-06  -2.980  0.00291 **
## BMI                             4.646e-02  4.990e-03   9.312  < 2e-16 ***
## under.five.deaths              -1.968e-03  7.010e-04  -2.807  0.00503 **
## Polio                           3.214e-02  4.548e-03   7.067 1.97e-12 ***
## Total.expenditure               6.147e-02  3.436e-02   1.789  0.07371 .
## Diphtheria                      4.784e-02  4.714e-03  10.148  < 2e-16 ***
## HIV.AIDS                       -4.792e-01  1.789e-02 -26.795  < 2e-16 ***
## Population                      3.545e-09  1.702e-09   2.082  0.03738 *
## thinness..1.19.years           -5.912e-02  2.388e-02  -2.475  0.01337 *
## Income.composition.of.resources 6.380e+00  6.402e-01   9.966  < 2e-16 ***
## Schooling                       6.849e-01  4.179e-02  16.389  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.148 on 2922 degrees of freedom
## Multiple R-squared:  0.8107, Adjusted R-squared:  0.8097
## F-statistic: 834.2 on 15 and 2922 DF,  p-value: < 2.2e-16
```

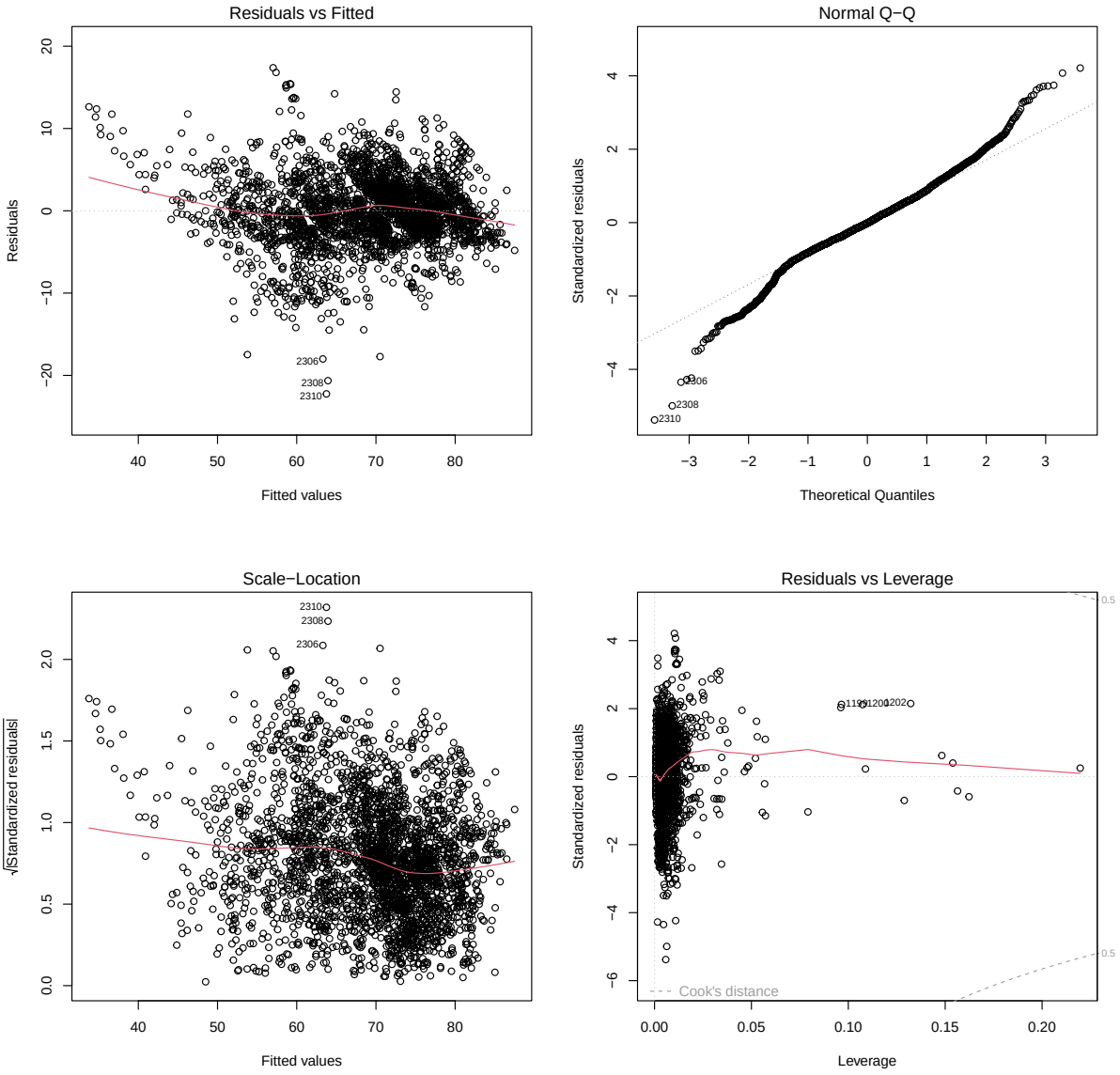
Alcohol is a factor dropped. Consumption of alcohol alone does not reflect the health of its people. In less developed countries alcohol could be a social issue, whereas in developed countries it could be a 'luxury' that is enjoyed. This makes it a poor indicator of life expectancy. Perhaps if we had alcohol abuse data, that may be a better indicator.

The final selected model has **15 features**. These are the key contributors to life expectancy. Features such as population, thinness 1-19 yr and total expenditure have different significant levels and slightly weaker evidence to include them. Other feature selection algorithms may remove these and reduce the model to **12 features**.

Diagnostic plots

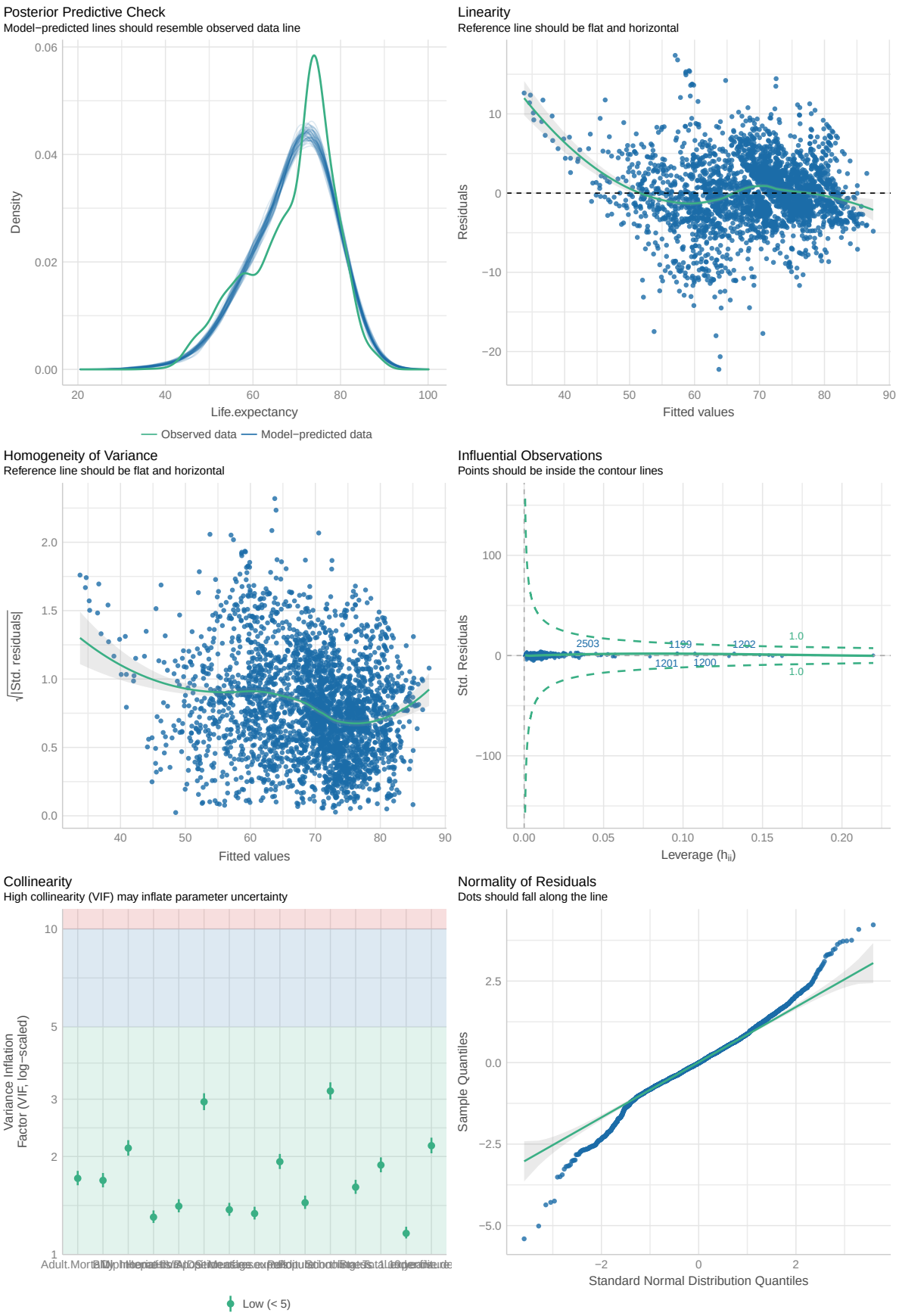
This section is to verify the assumptions of a linear model. Included just for completeness. Look for linear, independent and constant variance in the residual plots. Look for points with high leverage, cook's distance or residuals.

```
par(mfrow=c(2,2))
plot(model3)
```



Linear regression assumptions met.

```
# Using performance package
check_model(model3)
```


Posterior Predictive Check
Model−predicted lines should resemble observed data line
Density
Life.expectancy
Observed data
Model−predicted data
Linearity
Reference line should be flat and horizontal
Residuals
Fitted values
Homogeneity of Variance
Reference line should be flat and horizontal
√|Std. residuals|
Fitted values
Influential Observations
Points should be inside the contour lines
Std. Residuals
Leverage (hii)
2503
1199
1202
1201
1200
1.0
1.0
Collinearity
High collinearity (VIF) may inflate parameter uncertainty
Variance Inflation Factor (VIF, log−scaled)
Low (< 5)
Normality of Residuals
Dots should fall along the line
Sample Quantiles
Standard Normal Distribution Quantiles

Principal component analysis

Since there are 22 variables, it is impractical to plot every single variable against each other and attempt to deduce patterns.

Instead, principle component analysis is a clever transformation that captures most of the information in two new variables. Namely, the first two **principal components**.

```
pca <- prcomp(life_numerical,scale=TRUE)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.474 1.6364 1.30731 1.19036 1.11458 0.93545 0.91672
## Proportion of Variance 0.322 0.1409 0.08995 0.07458 0.06538 0.04606 0.04423
## Cumulative Proportion  0.322 0.4630 0.55292 0.62750 0.69288 0.73894 0.78317
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.86690 0.78394 0.76678 0.70600 0.66287 0.64239 0.56217
## Proportion of Variance 0.03955 0.03235 0.03095 0.02623 0.02313 0.02172 0.01663
## Cumulative Proportion  0.82272 0.85507 0.88601 0.91224 0.93537 0.95709 0.97372
##                           PC15    PC16    PC17    PC18    PC19
## Standard deviation     0.44540 0.38448 0.30018 0.24542 0.05202
## Proportion of Variance 0.01044 0.00778 0.00474 0.00317 0.00014
## Cumulative Proportion  0.98416 0.99195 0.99669 0.99986 1.00000
```

We can see the first two component captures around half the variance (46%) in the entire dataset.

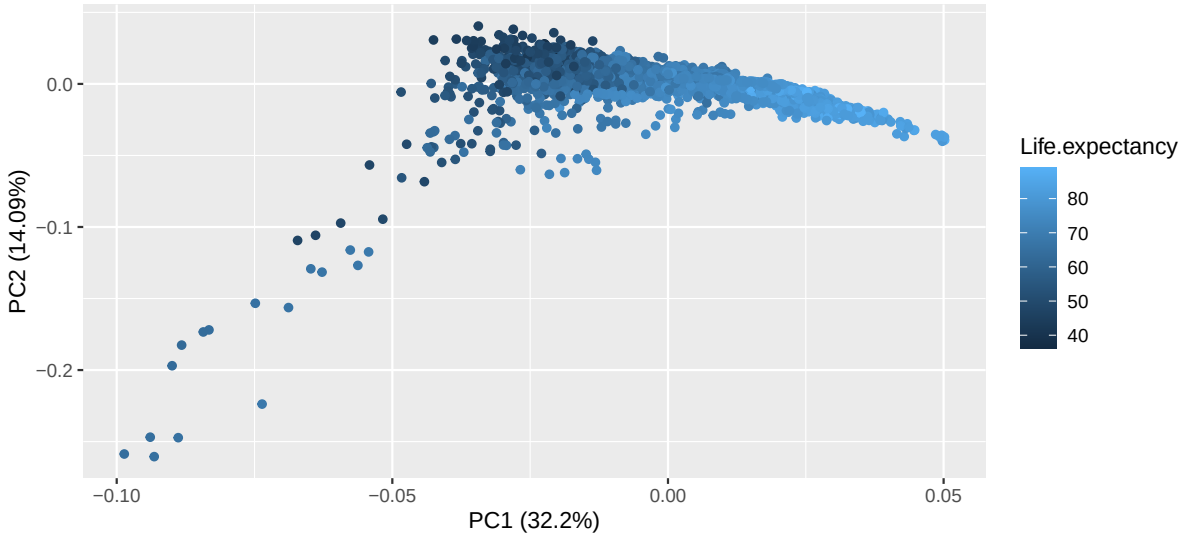
```
library(ggfortify)
```

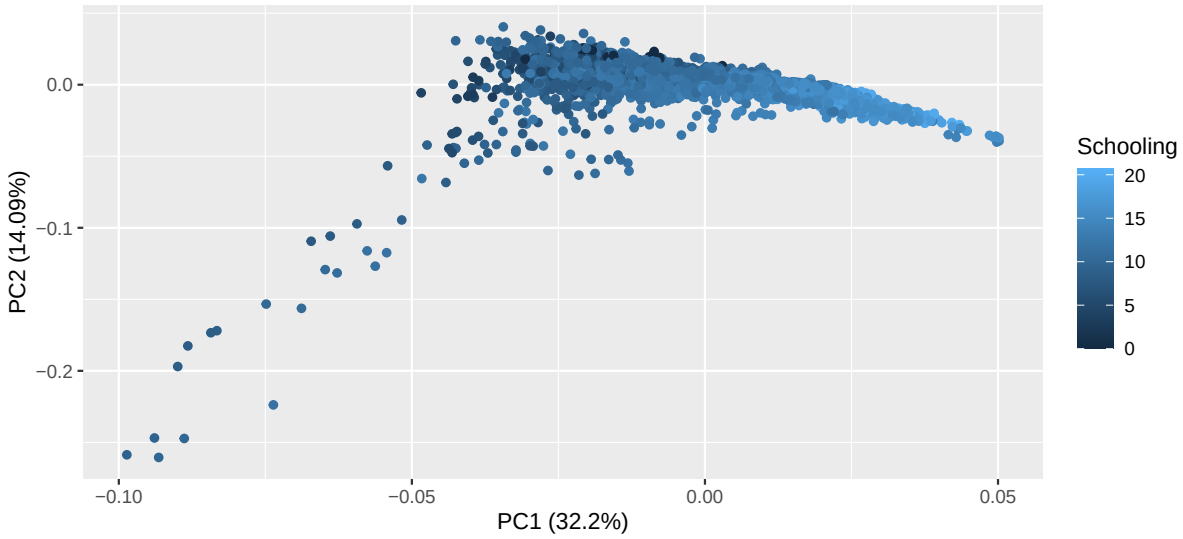
```
## Loading required package: ggplot2
```

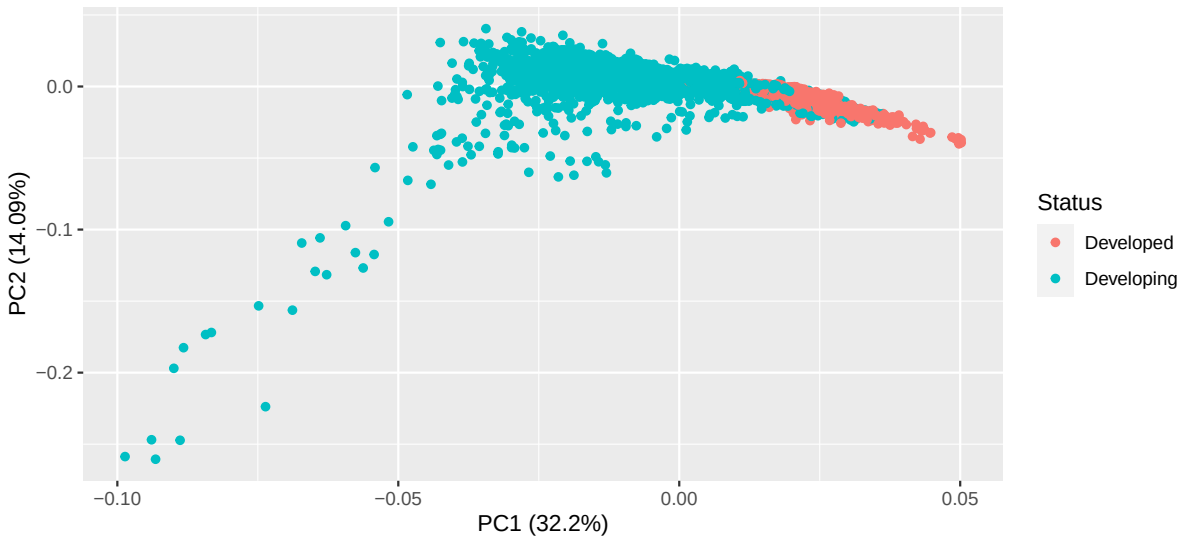
```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
plot1 <- autoplot(pca,data=life,colour="Life.expectancy")
plot2 <- autoplot(pca,data=life,colour="Schooling")
plot3 <- autoplot(pca,data=life,colour="Status")
grid.arrange(plot1, plot2, plot3, nrow=3)
```







We can see a striking connection between life expectancy and the status of a country. The **clustering of country status agrees with the clustering of life expectancy**. Since we are plotting against the two principal components (they are linear combinations of the predictors), we can infer **countries of similar life expectancy share similar attributes**.

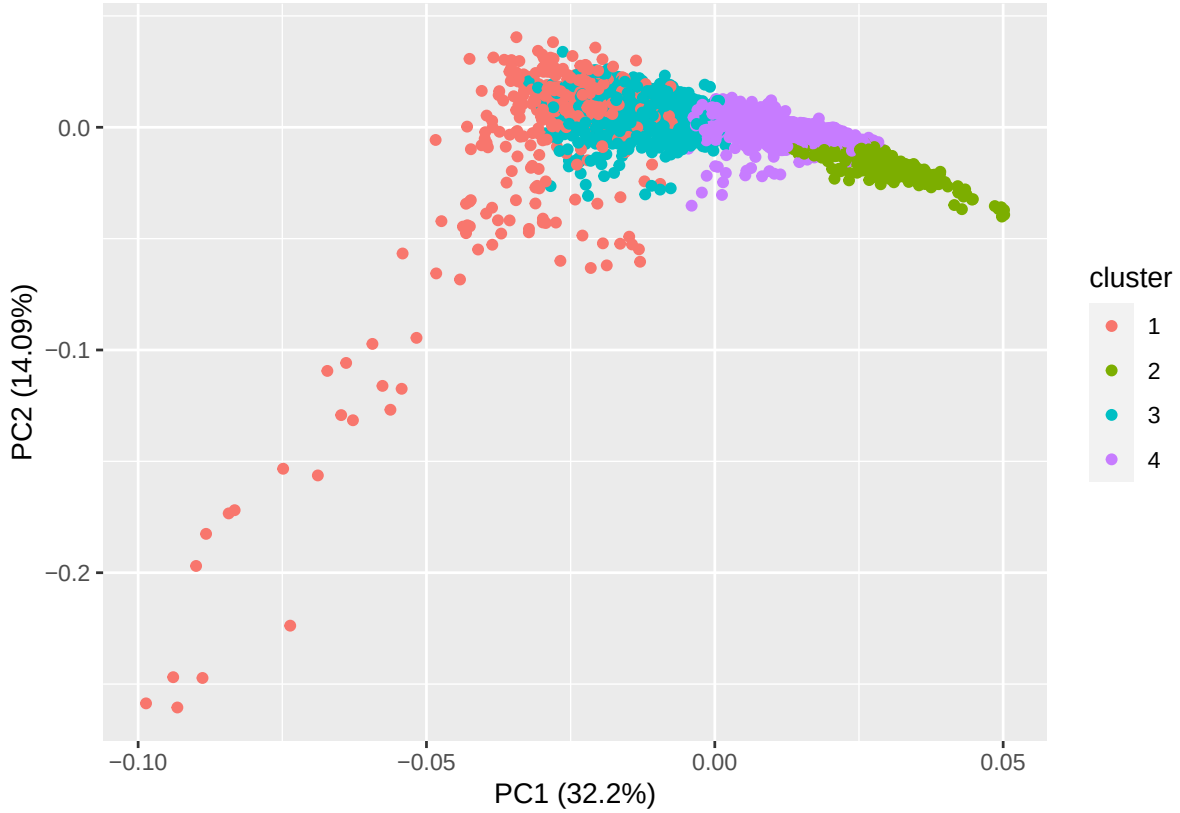
Although schooling does not directly affect life expectancy. It is reasonable to assume countries with poor education systems also have a poor healthcare systems and a lower standard of living. Hence we

also observe a similar clustering pattern.

A few outliers are coming from developing countries. This could be due to missing or incorrect data. Data from some developing countries are incomplete and imputation may have distorted the data.

K-means Clustering

```
library(cluster)
life_numerical_scale <- scale(life_numerical)
lifekmeans <- kmeans(life_numerical_scale, 4)
autoplot(lifekmeans, data = life_numerical_scale)
```



Going through the which cluster each data point was assigned to Reveals:

- Cluster 2- Australia, Denmark, Canada, New Zealand
- Cluster 4- Estonia, Fiji, Mexico
- Cluster 1 & 3- Gambia, Malawai, Equatorial Guinea

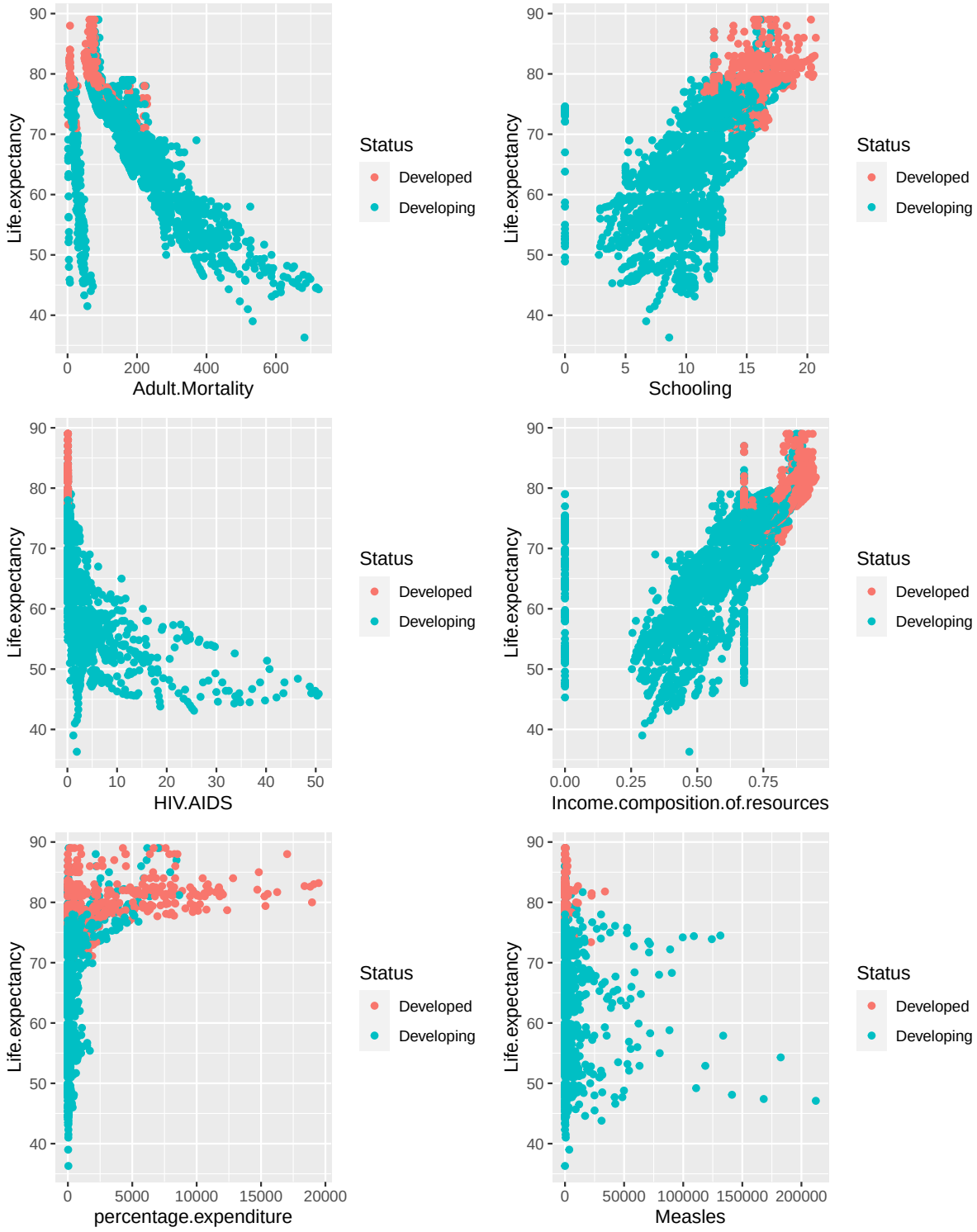
These groupings roughly describes the level of human development and life expectancy.

Cluster plots

```
plot1 <- ggplot() +
  geom_point(data = life,
             mapping = aes(x = Adult.Mortality, y = Life.expectancy,
                                colour = Status))

plot2 <- ggplot() +
  geom_point(data = life,
             mapping = aes(x = Schooling, y = Life.expectancy,
                                colour = Status))
plot3 <- ggplot() +
  geom_point(data = life,
             mapping = aes(x = HIV.AIDS, y = Life.expectancy,
                                colour = Status))
plot4 <- ggplot() +
  geom_point(data = life,
             mapping = aes(x = Income.composition.of.resources, y = Life.expectancy,
                                colour = Status))
plot5 <- ggplot() +
  geom_point(data = life,
             mapping = aes(x = percentage.expenditure, y = Life.expectancy,
                                colour = Status))
plot6 <- ggplot() +
  geom_point(data = life,
             mapping = aes(x = Measles, y = Life.expectancy,
                                colour = Status))

grid.arrange(plot1, plot2, plot3,plot4,plot5, plot6,ncol=2, nrow=3)
```


Life.expectancy
Adult.Mortality
Schooling
HIV.AIDS
Income.composition.of.resources
percentage.expenditure
Measles
Status
Developed
Developing

As expected, a higher adult mortality rate is related to a lower life expectancy. Hence the strong negative trend.

Schooling was discussed in the previous section. Some minor trends look like vertical lines. These are due to the missing data where years of schooling were set to zero or as the median of the entire dataset. However, ignoring these points, we can still observe a clear separation between developed and developing countries.

HIV.AIDS and Measles both have a similar characteristic spread. We see developed nations forming a straight line cluster on the top left of the graph. Indicating high life expectancy and extremely low levels of these diseases. High rates of disease is linked to lower life expectancy.

Income composition of resources in this dataset is a form of the human development index. Hence the strong linear trend. Again we see the vertical trends caused by data issues.

Percentage expenditure (on healthcare) shows again a clear clustering of developing and developed nations. Developing nations spend a small proportion of their GDP on healthcare. Some developed nations seem to spend a small proportion of their GDP as well, but this could be due to a very high GDP.

Conclusion

What does the data and models tell us in plain English?

Life expectancy depends on a few contributors (described below). From the regression models and visulisations, we can see countries of similar life expectancy share similar attributes. Developed countries have a higher life expectancy. They have longer years of schooling, more spending on healthcare and lower disease/mortality rates. The opposite can be said for developing countries.

To improve a country's life expectancy, improvements must be made to these factors. These factors are tied to economic development. The question of how shall be left with policy makers.

What are the key contributors to life expectancy?

Using the regression model and cluster plots, the main factors influencing life expectancy are:

- Status (Developing vs developed)
- Adult mortality
- Percentage expenditure (on health)
- Hepatitis B
- Measles
- BMI
- Under five deaths
- Polio
- Diphtheria
- HIV.AIDS
- Income composition of resources
- Schooling

These are a mix of economic, health and education indicators.They all relate to the welfare of humans.