Stat230a Final Project (Life Expectancy)

Lauren Flemmer

Read in data

```
lifeExpectancy <- read.csv('/Users/laurenflemmer/Desktop/life_expectancy_proj/Life Expectancy Data.csv')

lifeExpectancy <- lifeExpectancy %>% drop_na()

head(lifeExpectancy, 5)
```

```
##       Country Year     Status Life.expectancy Adult.Mortality infant.deaths
## 1 Afghanistan 2015 Developing            65.0             263            62
## 2 Afghanistan 2014 Developing            59.9             271            64
## 3 Afghanistan 2013 Developing            59.9             268            66
## 4 Afghanistan 2012 Developing            59.5             272            69
## 5 Afghanistan 2011 Developing            59.2             275            71
##   Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths
## 1    0.01              71.279624          65    1154 19.1                83
## 2    0.01              73.523582          62     492 18.6                86
## 3    0.01              73.219243          64     430 18.1                89
## 4    0.01              78.184215          67    2787 17.6                93
## 5    0.01               7.097109          68    3013 17.2                97
##   Polio Total.expenditure Diphtheria HIV.AIDS       GDP Population
## 1     6              8.16         65      0.1 584.25921   33736494
## 2    58              8.18         62      0.1 612.69651     327582
## 3    62              8.13         64      0.1 631.74498   31731688
## 4    67              8.52         67      0.1 669.95900    3696958
## 5    68              7.87         68      0.1  63.53723    2978599
##   thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 1                 17.2                17.3                           0.479
## 2                 17.5                17.5                           0.476
## 3                 17.7                17.7                           0.470
## 4                 17.9                18.0                           0.463
## 5                 18.2                18.2                           0.454
##   Schooling
## 1      10.1
## 2      10.0
## 3       9.9
## 4       9.8
## 5       9.5
```

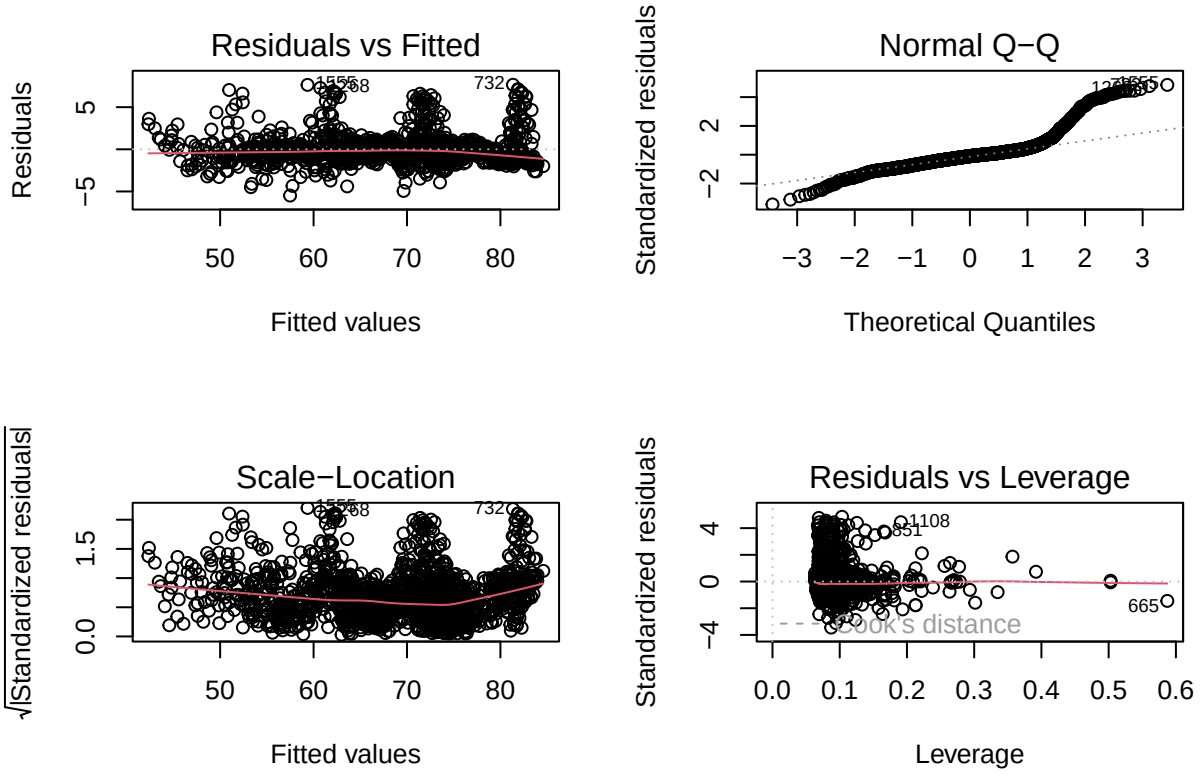
```
colSums(is.na(lifeExpectancy))
```

```
##                         Country                            Year
##                               0                               0
##                          Status                 Life.expectancy
```

```
##                                0                                0
##                  Adult.Mortality                    infant.deaths
##                                0                                0
##                          Alcohol           percentage.expenditure
##                                0                                0
##                      Hepatitis.B                         Measles
##                                0                                0
##                              BMI                  under.five.deaths
##                                0                                0
##                            Polio                Total.expenditure
##                                0                                0
##                       Diphtheria                         HIV.AIDS
##                                0                                0
##                              GDP                       Population
##                                0                                0
##             thinness..1.19.years                thinness.5.9.years
##                                0                                0
## Income.composition.of.resources                         Schooling
##                                0                                0
```

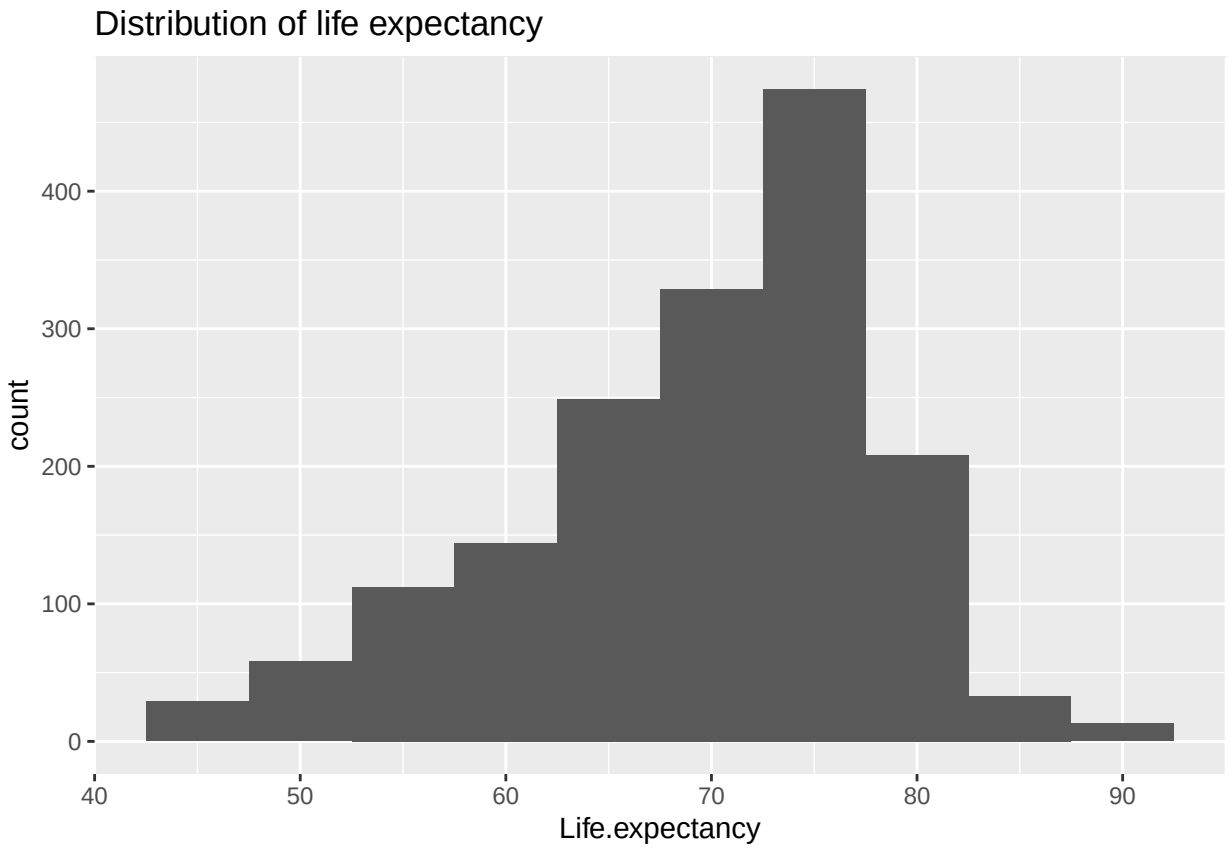
To determine which model is appropriate, plot diagnostics

```
par(mfrow=c(2,2))
# diagnostic plots
basic_model <- lm(Life.expectancy ~ ., data = lifeExpectancy)
plot(basic_model)
```

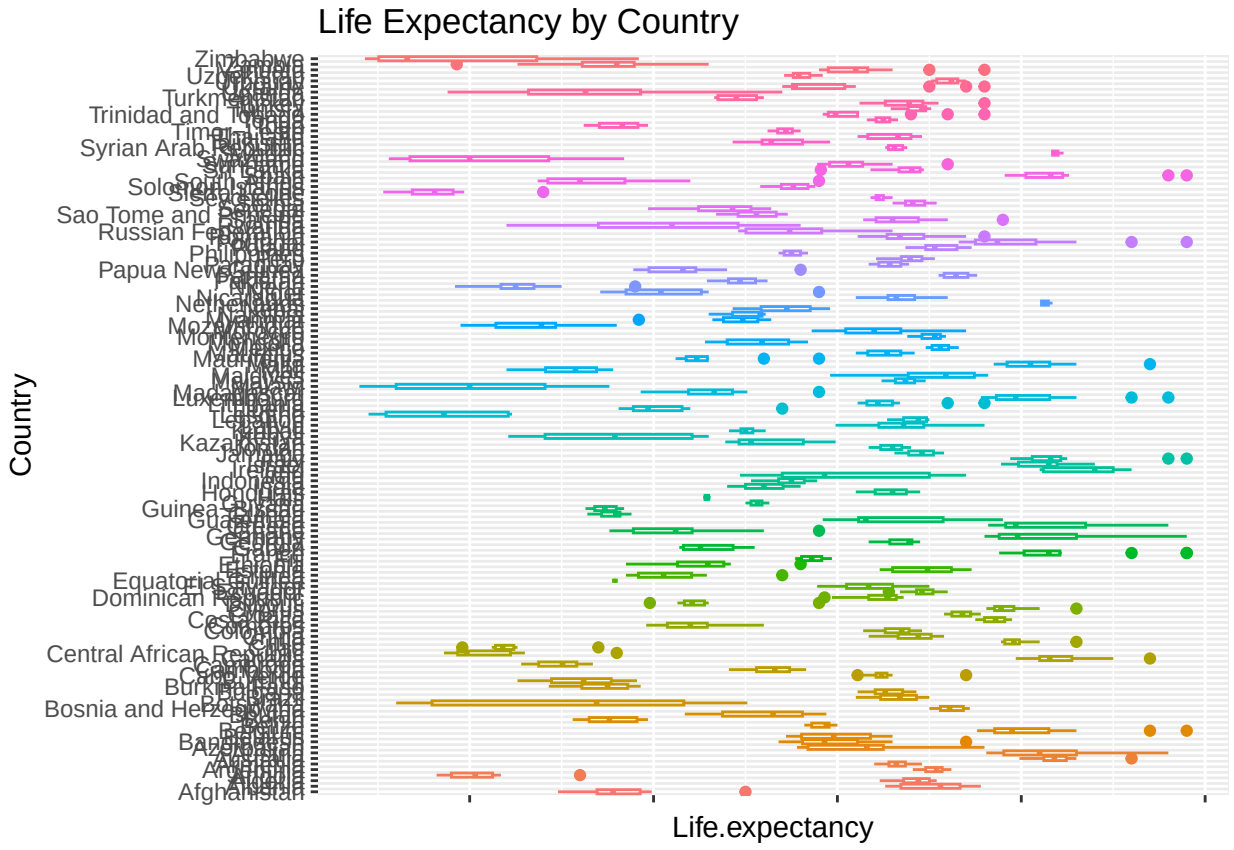



```
# plot response
ggplot(data=lifeExpectancy) +
```

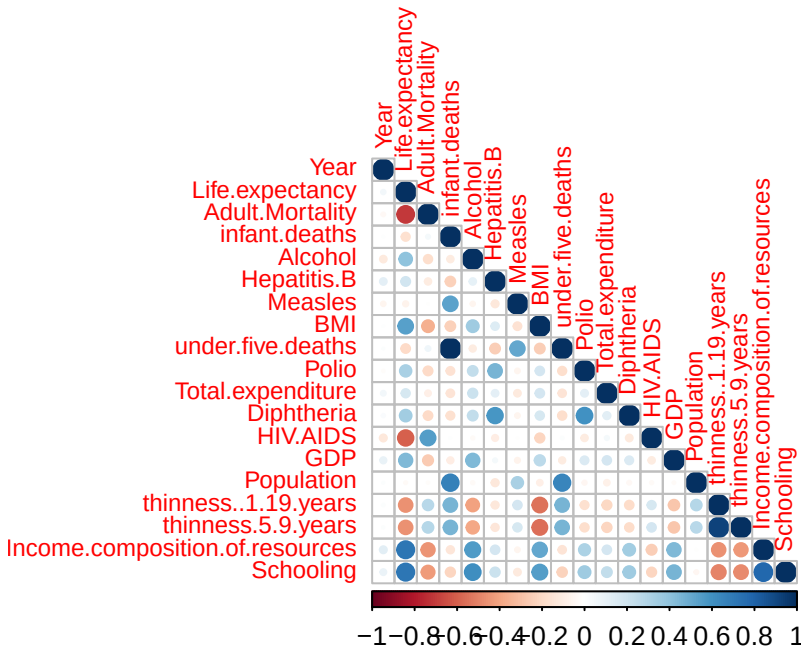
```
geom_histogram(mapping=aes(x=Life.expectancy), bins=10) +
ggtitle("Distribution of life expectancy")
```




```
ggplot(data=lifeExpectancy) +
geom_boxplot(mapping=aes(x=Country, y=Life.expectancy, color=Country), bins=10) +
ggtitle("Life Expectancy by Country") +
theme(legend.position = "none") +
coord_flip() +
theme(axis.text.x = element_text(angle = 80, vjust = 0.5, hjust=1, size=0.01))
```




```
# correlation plot
# include only continuous vars
# removed percentage expenditure bc redundant
continuous_vars <- lifeExpectancy %>% select(Year, Life.expectancy,
                                             Adult.Mortality, infant.deaths,
                                             Alcohol, 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)


corMat <- cor(continuous_vars)
corrplot(corMat, type="lower", method="circle",  tl.cex = 0.75)
```



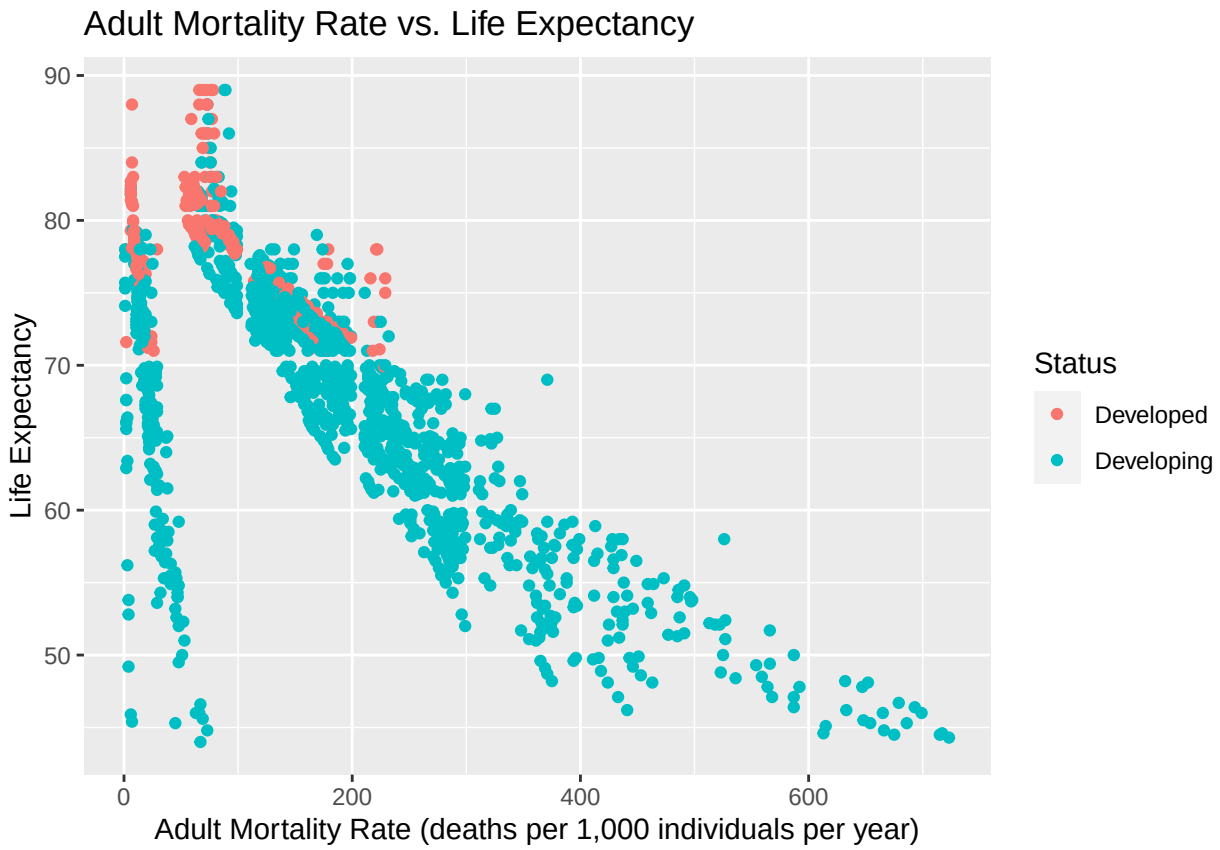

```
#corMat
```

Variables correlated with life expectancy ($|r| > 0.4$):

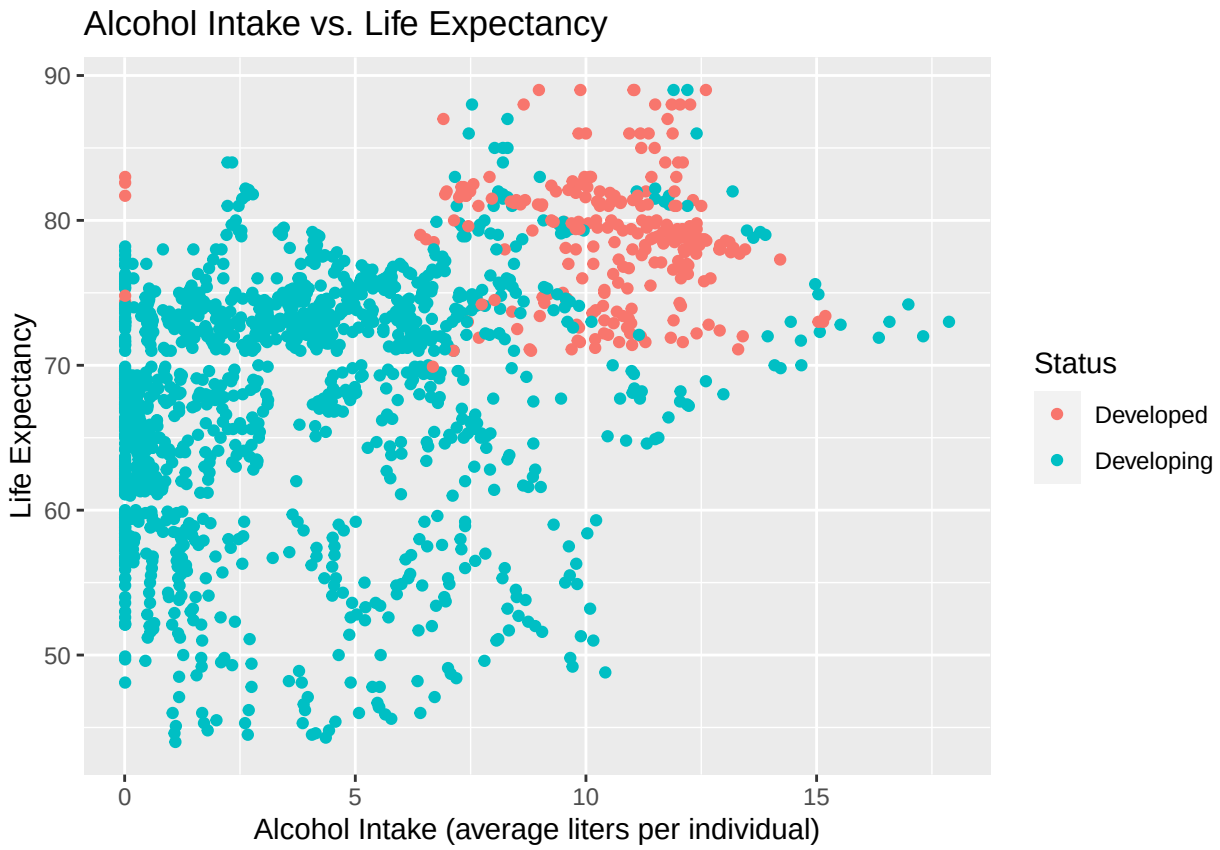
- Adult mortality rate
- Alcohol intake
- BMI
- HIV/AIDS
- GDP
- Prevalence of thinness among ages 10-19
- Prevalence in thinness among ages 5-9
- Human Development Index (in terms of income composition of resources)
- Number of years of schooling

EDA for variables correlated with life expectancy

```
# adult mortality vs. life expectancy
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=Adult.Mortality, y=Life.expectancy, color=Status)) +
  ggtitle("Adult Mortality Rate vs. Life Expectancy") +
  xlab("Adult Mortality Rate (deaths per 1,000 individuals per year)") +
  ylab("Life Expectancy")
```

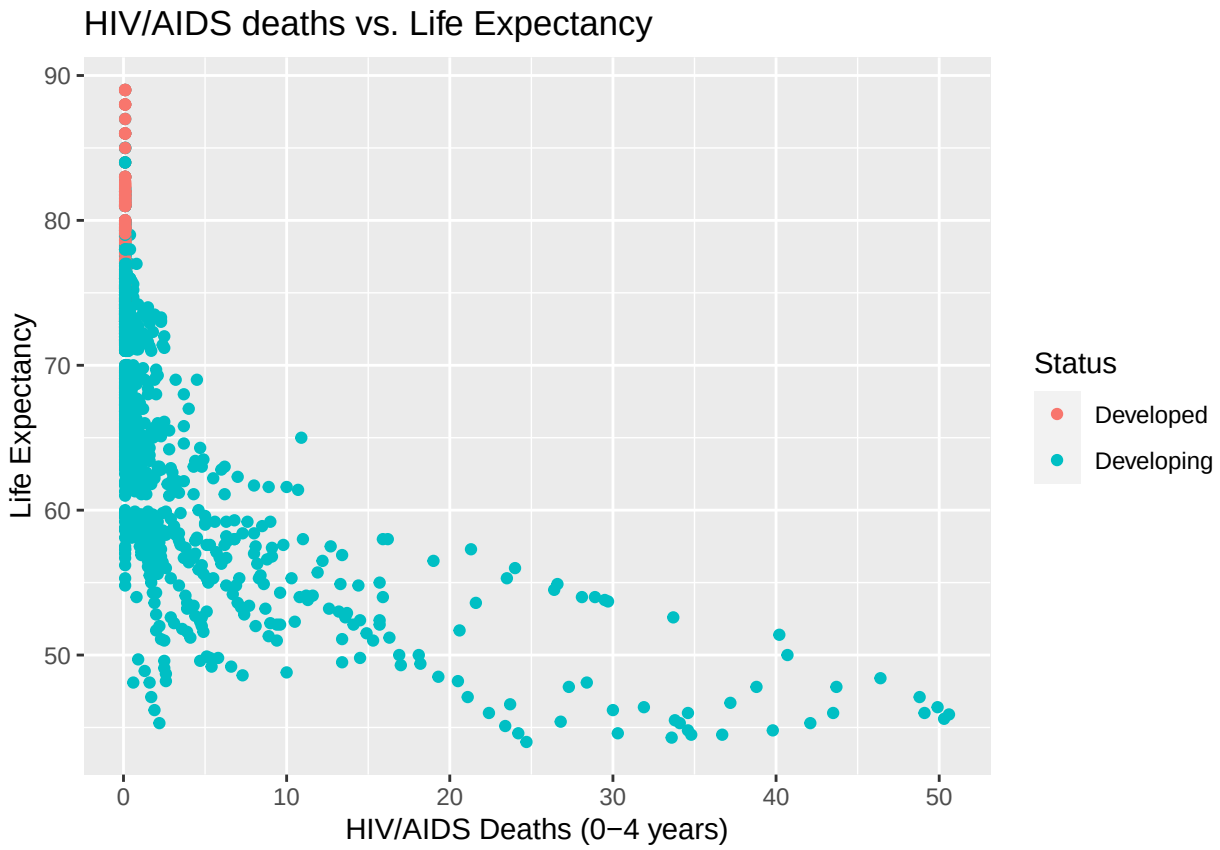



```
# alcohol intake vs life expectancy
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=Alcohol, y=Life.expectancy, color=Status)) +
  ggtitle("Alcohol Intake vs. Life Expectancy") +
  xlab("Alcohol Intake (average liters per individual)") +
  ylab("Life Expectancy")
```

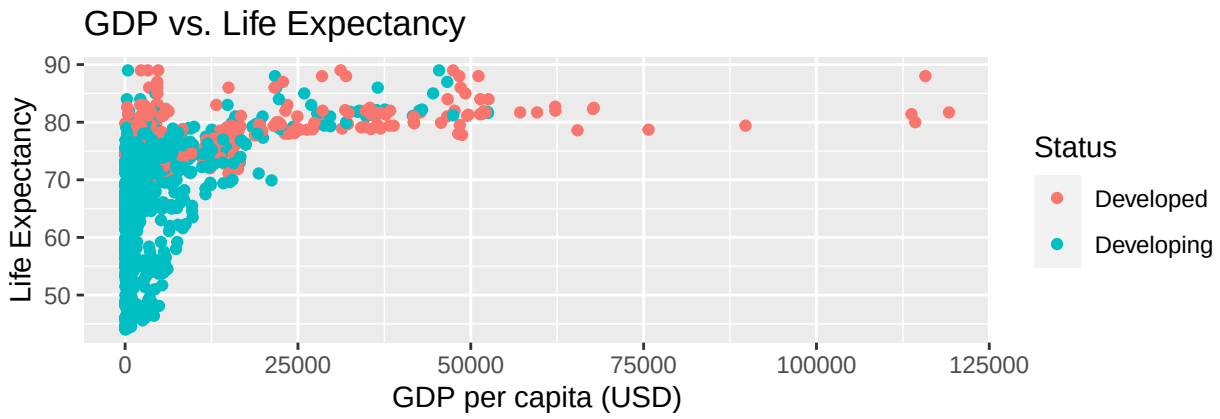


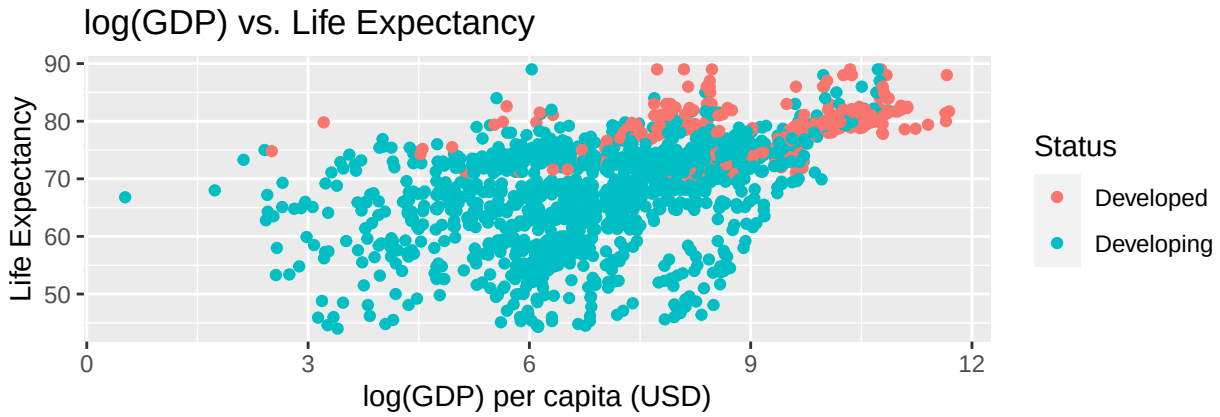

```
# bmi vs. life expectancy
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=BMI, y=Life.expectancy, color=Status)) +
  ggtitle("BMI vs. Life Expectancy") +
  xlab("BMI") +
  ylab("Life Expectancy")
```


```
# hiv/aids vs life expectancy
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=HIV.AIDS, y=Life.expectancy, color=Status)) +
  ggtitle("HIV/AIDS deaths vs. Life Expectancy") +
  xlab("HIV/AIDS Deaths (0-4 years)") +
  ylab("Life Expectancy")
```

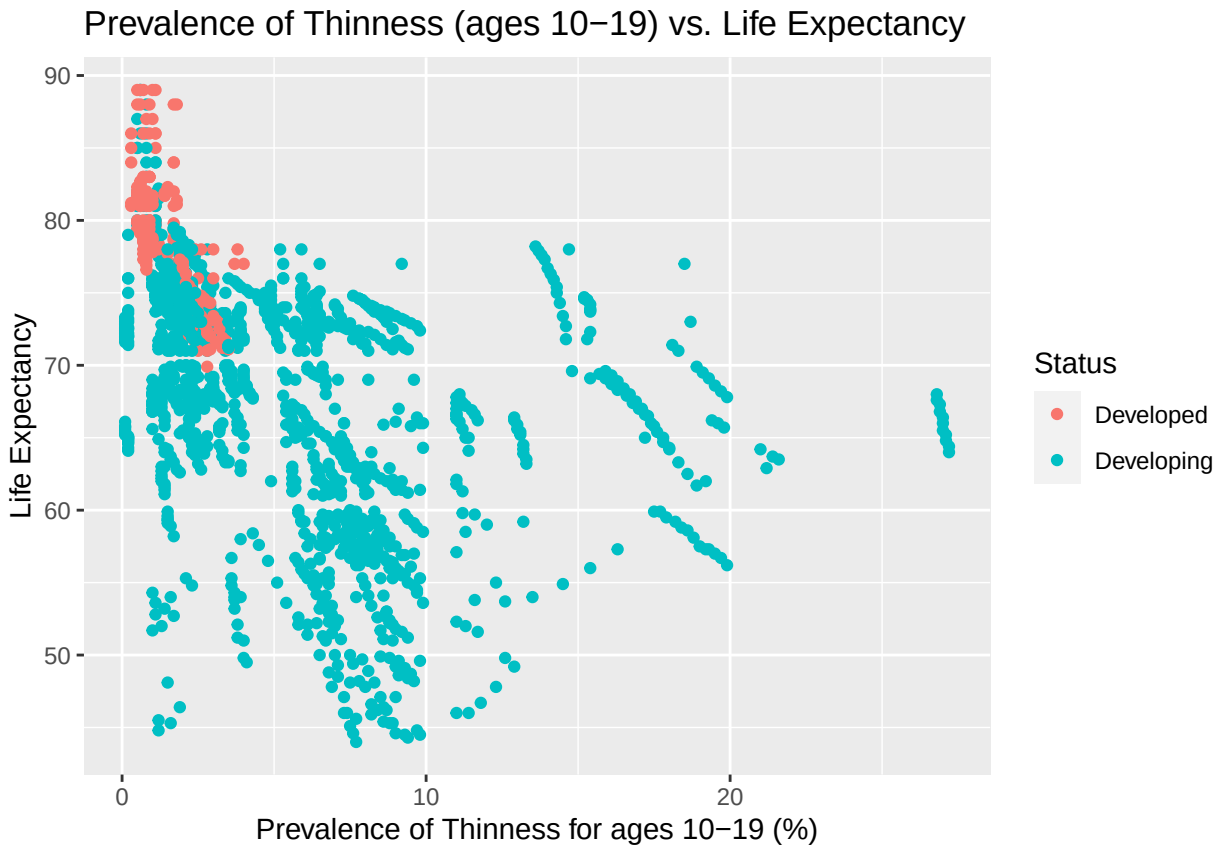



```
# gdp vs life expectancy
gdp_plot <- ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=GDP, y=Life.expectancy, color=Status)) +
  ggtitle("GDP vs. Life Expectancy") +
  xlab("GDP per capita (USD)") +
  ylab("Life Expectancy")

gdp_log_plot <- ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=log(GDP), y=Life.expectancy, color=Status)) +
  ggtitle("log(GDP) vs. Life Expectancy") +
  xlab("log(GDP) per capita (USD)") +
  ylab("Life Expectancy")

grid.arrange(gdp_plot, gdp_log_plot)
```

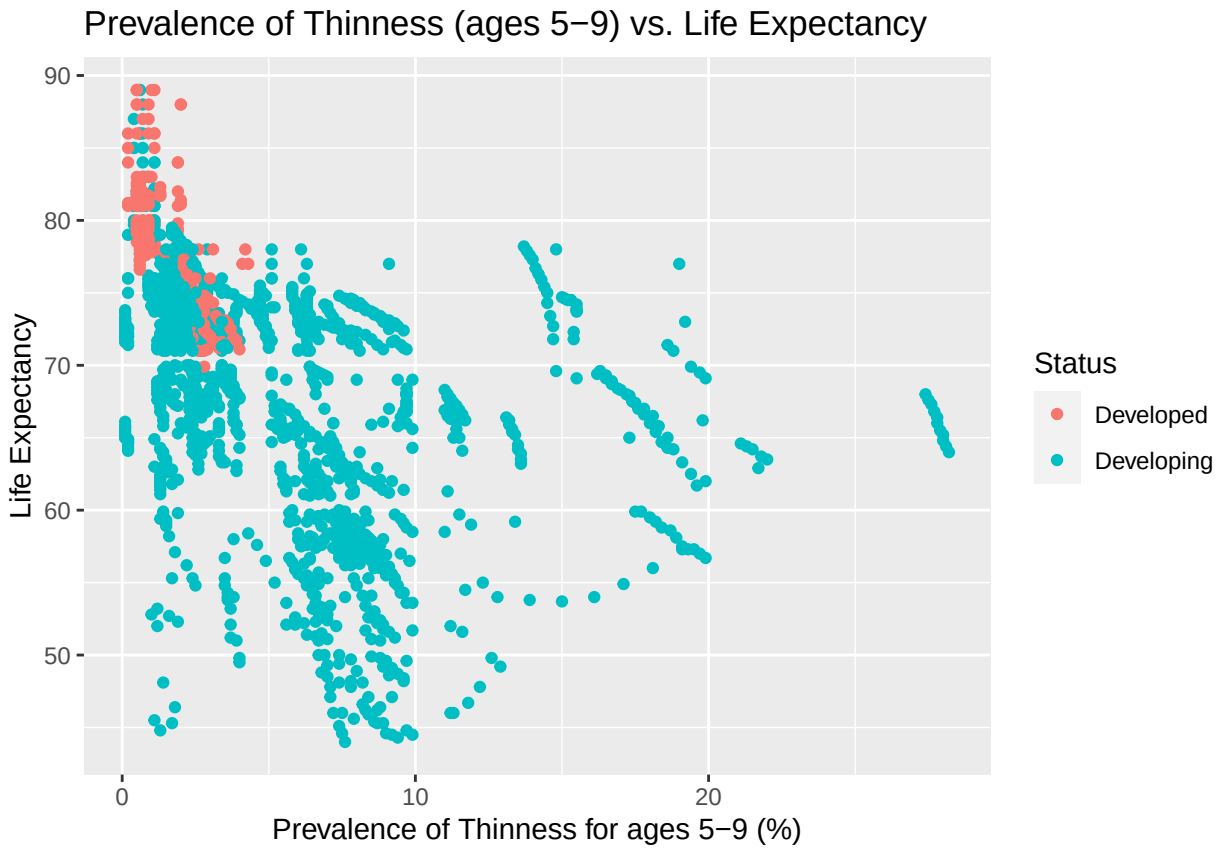




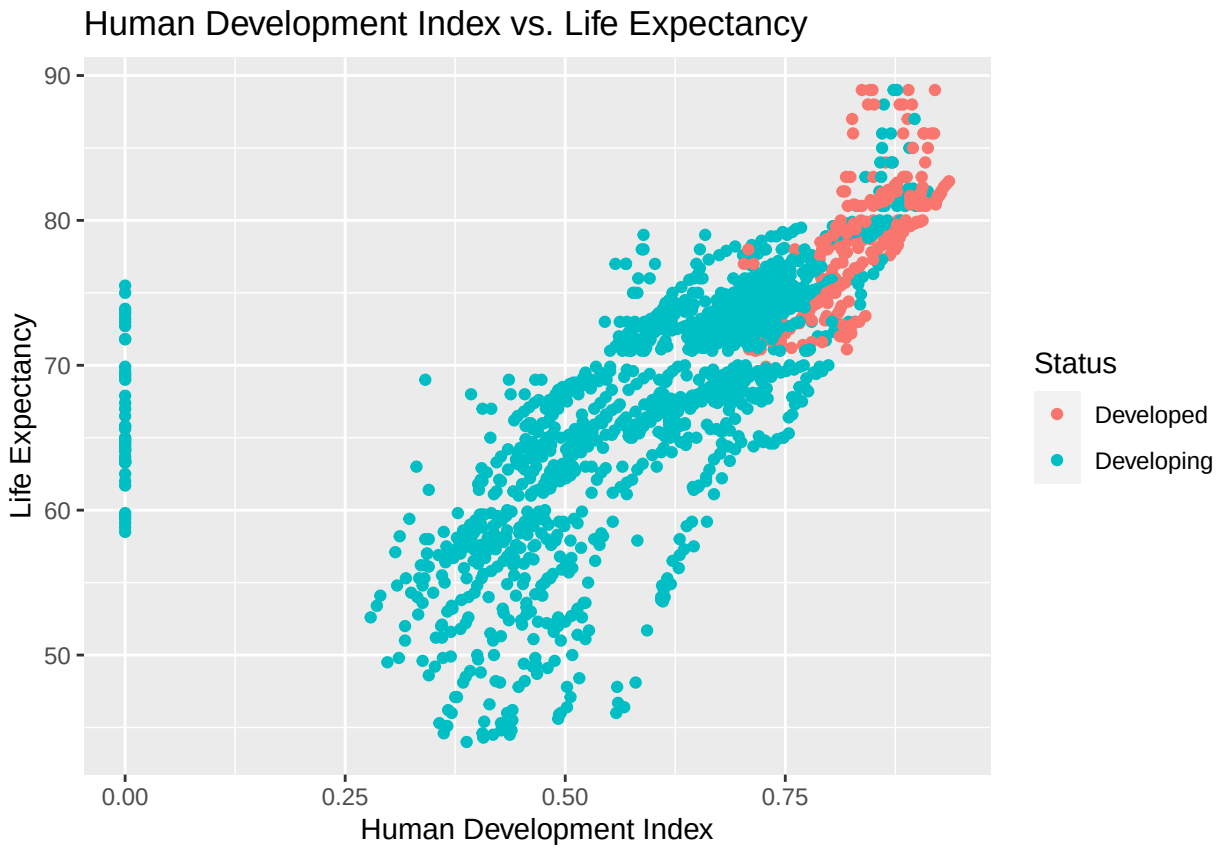

```
# prevalence of thinness (ages 10-19)
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=thinness..1.19.years, y=Life.expectancy, color=Status)) +
  ggtitle("Prevalence of Thinness (ages 10-19) vs. Life Expectancy") +
  xlab("Prevalence of Thinness for ages 10-19 (%)") +
  ylab("Life Expectancy")
```



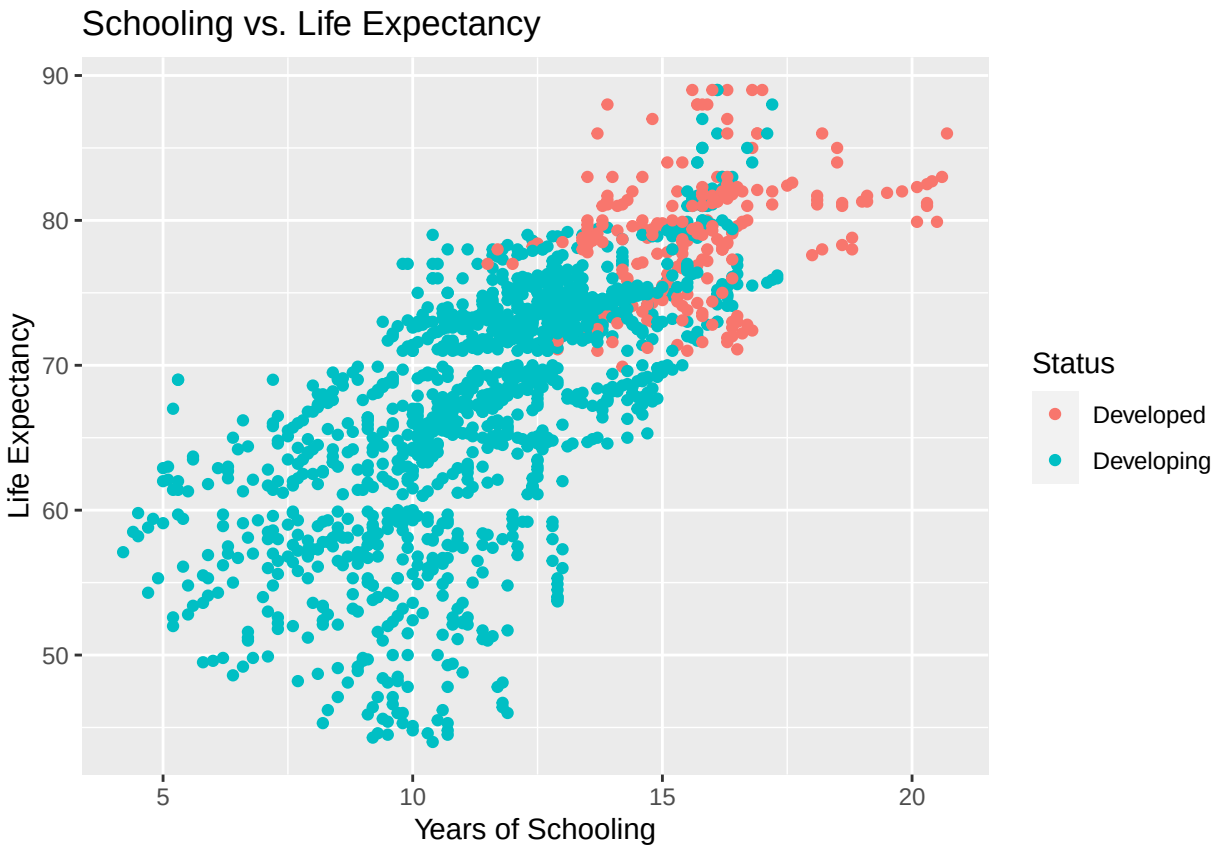

```
# prevalence of thinness (ages 5-9)
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=thinness.5.9.years, y=Life.expectancy, color=Status)) +
  ggtitle("Prevalence of Thinness (ages 5-9) vs. Life Expectancy") +
  xlab("Prevalence of Thinness for ages 5-9 (%)") +
  ylab("Life Expectancy")
```




```
# human development index vs. life expectancy
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=Income.composition.of.resources, y=Life.expectancy, color=Status)) +
  ggtitle("Human Development Index vs. Life Expectancy") +
  xlab("Human Development Index") +
  ylab("Life Expectancy")
```

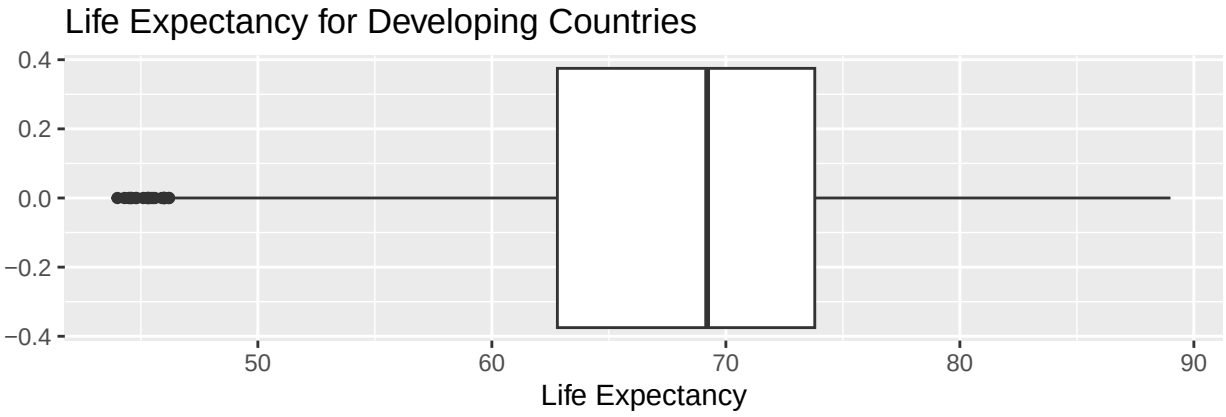


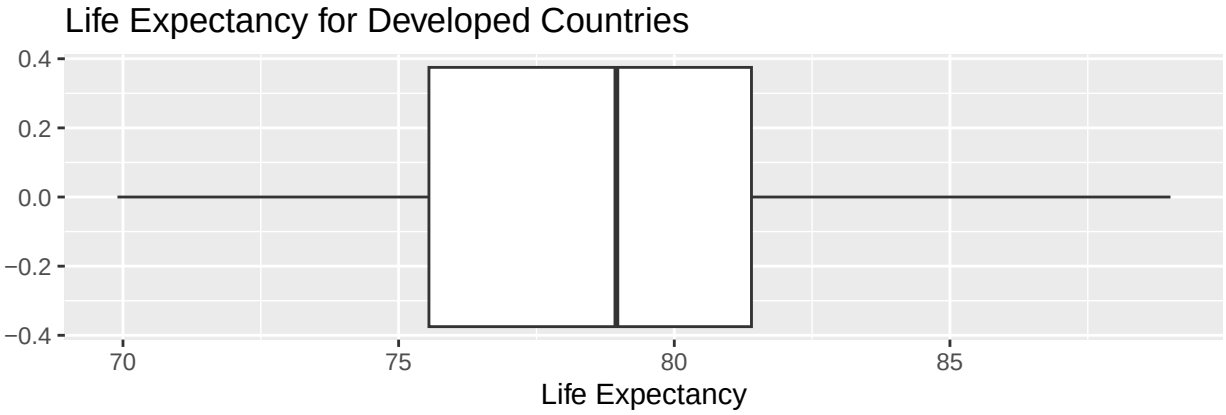

```
# years of schooling vs. life expectancy
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=Schooling, y=Life.expectancy, color=Status)) +
  ggtitle("Schooling vs. Life Expectancy") +
  xlab("Years of Schooling") +
  ylab("Life Expectancy")
```



EDA for non-continuous variables

```
# status vs. life expectancy
developing_expectancy <- lifeExpectancy %>% filter(Status == "Developing") %>% select(Life.expectancy)
developed_expectancy <- lifeExpectancy %>% filter(Status == "Developed") %>% select(Life.expectancy)

boxplot_developing <- ggplot(data=developing_expectancy) +
  geom_boxplot(mapping=aes(x=Life.expectancy)) +
  ggtitle("Life Expectancy for Developing Countries") +
  xlab("Life Expectancy")

boxplot_developed <- ggplot(data=developed_expectancy) +
  geom_boxplot(mapping=aes(x=Life.expectancy)) +
  ggtitle("Life Expectancy for Developed Countries") +
  xlab("Life Expectancy")

grid.arrange(boxplot_developing, boxplot_developed)
```





Initial modeling using 'geeglm' package

Model 1 (Independent correlation structure, gaussian) Predictors:

- Status (Developed/Developing)
- Adult mortality rate
- Alcohol intake
- BMI
- HIV/AIDS
- log(GDP)
- Prevalence of thinness among ages 10-19
- Prevalence in thinness among ages 5-9
- Human Development Index (in terms of income composition of resources)
- Number of years of schooling

```
# convert country to factor
lifeExpectancy$Country <- as.factor(lifeExpectancy$Country)

# convert HDI from 0-1 to 0-100 for better interpretability of model estimates
lifeExpectancy$Income.composition.of.resources <- lifeExpectancy$Income.composition.of.resources * 100
```

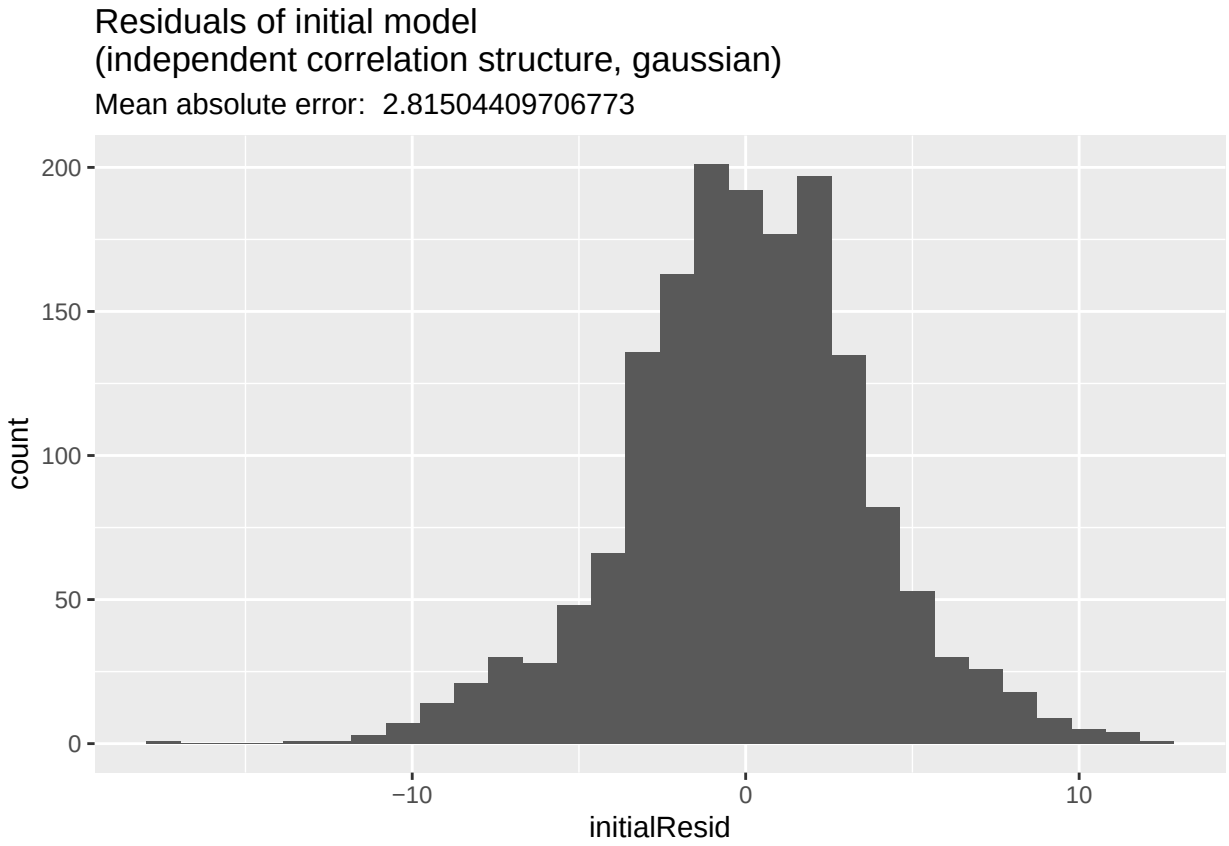
```
# fit initial model
initialModel <- geeglm(formula = Life.expectancy ~ Status + Year + Adult.Mortality + Alcohol + BMI
                               + HIV.AIDS + log(GDP) + thinness..1.19.years + thinness.5.9.years
                               + Income.composition.of.resources + Schooling, id = Country,
                               data = lifeExpectancy, family = "gaussian")
```

```
summary(initialModel)$coefficients
```

```
##                                    Estimate      Std.err       Wald
## (Intercept)                     358.66456721 75.453919934 22.5950515
## StatusDeveloping                 -1.36926240  0.819073000  2.7946546
## Year                             -0.15247061  0.038030626 16.0733104
## Adult.Mortality                  -0.01748929  0.002305808 57.5305255
## Alcohol                          -0.18849973  0.085152151  4.9003850
## BMI                               0.02743665  0.010028907  7.4843649
## HIV.AIDS                         -0.44441915  0.049598995 80.2859892
## log(GDP)                          0.49802463  0.095407183 27.2483054
## thinness..1.19.years             -0.03846752  0.068831247  0.3123324
## thinness.5.9.years               -0.03477710  0.057004658  0.3721912
## Income.composition.of.resources   0.11102774  0.027585366 16.1996464
## Schooling                         0.93039398  0.164142456 32.1286090
##                                      Pr(>|W|)
## (Intercept)                     1.999991e-06
## StatusDeveloping                9.457914e-02
## Year                            6.093684e-05
## Adult.Mortality                 3.330669e-14
## Alcohol                         2.685071e-02
## BMI                             6.223702e-03
## HIV.AIDS                        0.000000e+00
## log(GDP)                        1.789319e-07
## thinness..1.19.years            5.762524e-01
## thinness.5.9.years              5.418123e-01
## Income.composition.of.resources 5.700475e-05
## Schooling                       1.442967e-08
```

```
# plot residuals
initialResid <- as.vector(initialModel$residuals)
mean_abs_resid_initial <- mean(abs(initialResid))

ggplot() +
  geom_histogram(aes(initialResid)) +
  ggtitle(label = "Residuals of initial model \n(independent correlation structure, gaussian)",
      subtitle = paste("Mean absolute error: ", mean_abs_resid_initial))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Model 2 (Independent correlation structure) Predictors:

(Backward selection done using Robust Z)

- Status (Developed/Developing)
- Adult mortality rate
- Alcohol intake
- BMI
- HIV/AIDS
- log(GDP)
- Human Development Index (in terms of income composition of resources)
- Number of years of schooling

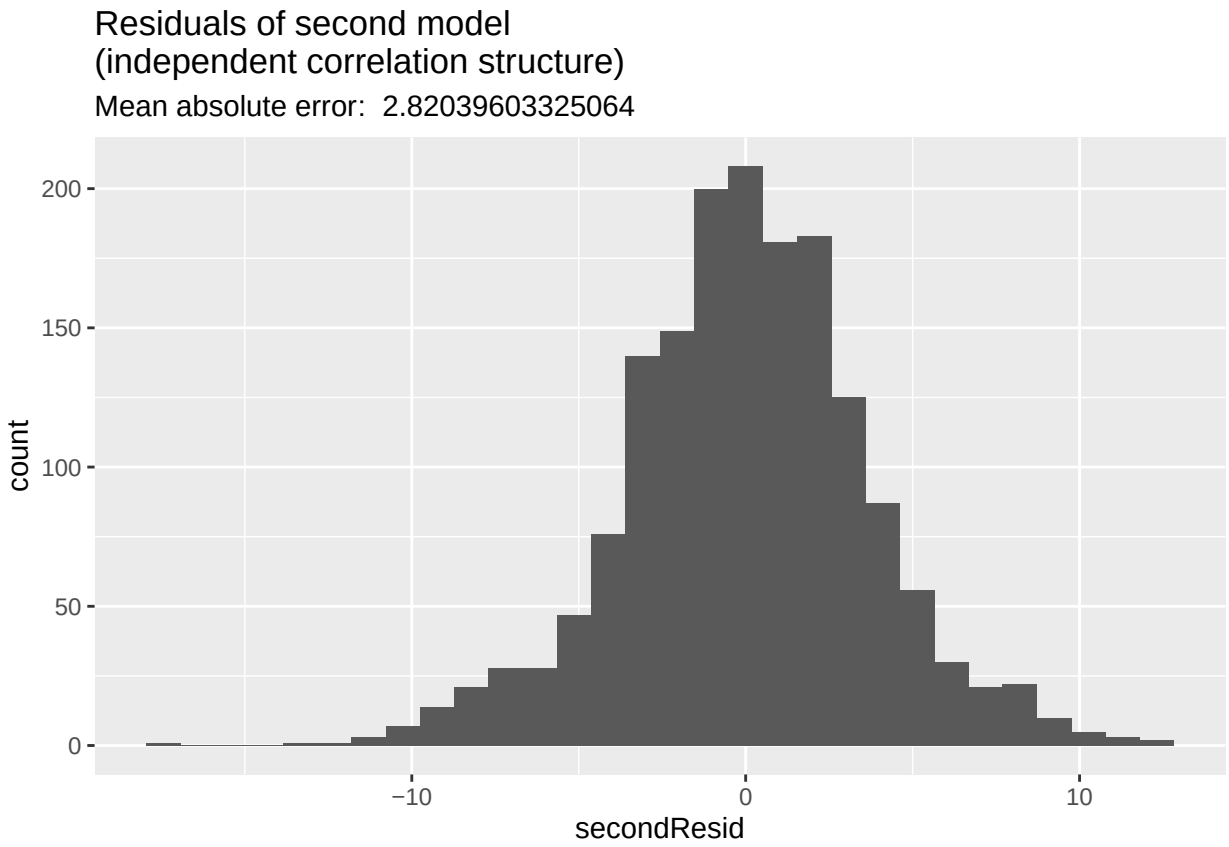
```
# fit second model
secondModel <- geeglm(formula = Life.expectancy ~ Status + Year + Adult.Mortality + Alcohol + BMI +
                              HIV.AIDS + log(GDP) + Income.composition.of.resources + Schooling,
                              id = Country, data = lifeExpectancy, family = "gaussian")
summary(secondModel)$coefficients
```

```
##                                  Estimate     Std.err      Wald     Pr(>|W|)
## (Intercept)                   362.17234816 74.85685334 23.408169 1.310210e-06
## StatusDeveloping               -1.37665316  0.82975532  2.752637 9.709416e-02
## Year                           -0.15465680  0.03767976 16.846968 4.051801e-05
## Adult.Mortality                -0.01753501  0.00229392 58.432646 2.098322e-14
## Alcohol                        -0.17648021  0.08527053  4.283454 3.848506e-02
## BMI                             0.03410873  0.01044020 10.673663 1.086717e-03
## HIV.AIDS                       -0.44699537  0.04902540 83.131154 0.000000e+00
## log(GDP)                        0.49770594  0.09523882 27.309744 1.733357e-07
```

```
## Income.composition.of.resources   0.11214319  0.02707574 17.154775 3.445422e-05
## Schooling                         0.94413800  0.15991610 34.856724 3.548818e-09
```

```r
# plot residuals
secondResid <- as.vector(secondModel$residuals)
mean_abs_resid_second <- mean(abs(secondResid))

ggplot() +
  geom_histogram(aes(secondResid)) +
  ggtitle(label = "Residuals of second model \n(independent correlation structure)",
      subtitle = paste("Mean absolute error: ", mean_abs_resid_second))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



"AR1" correlation

Model 1 (AR1 correlation structure)Predictors:

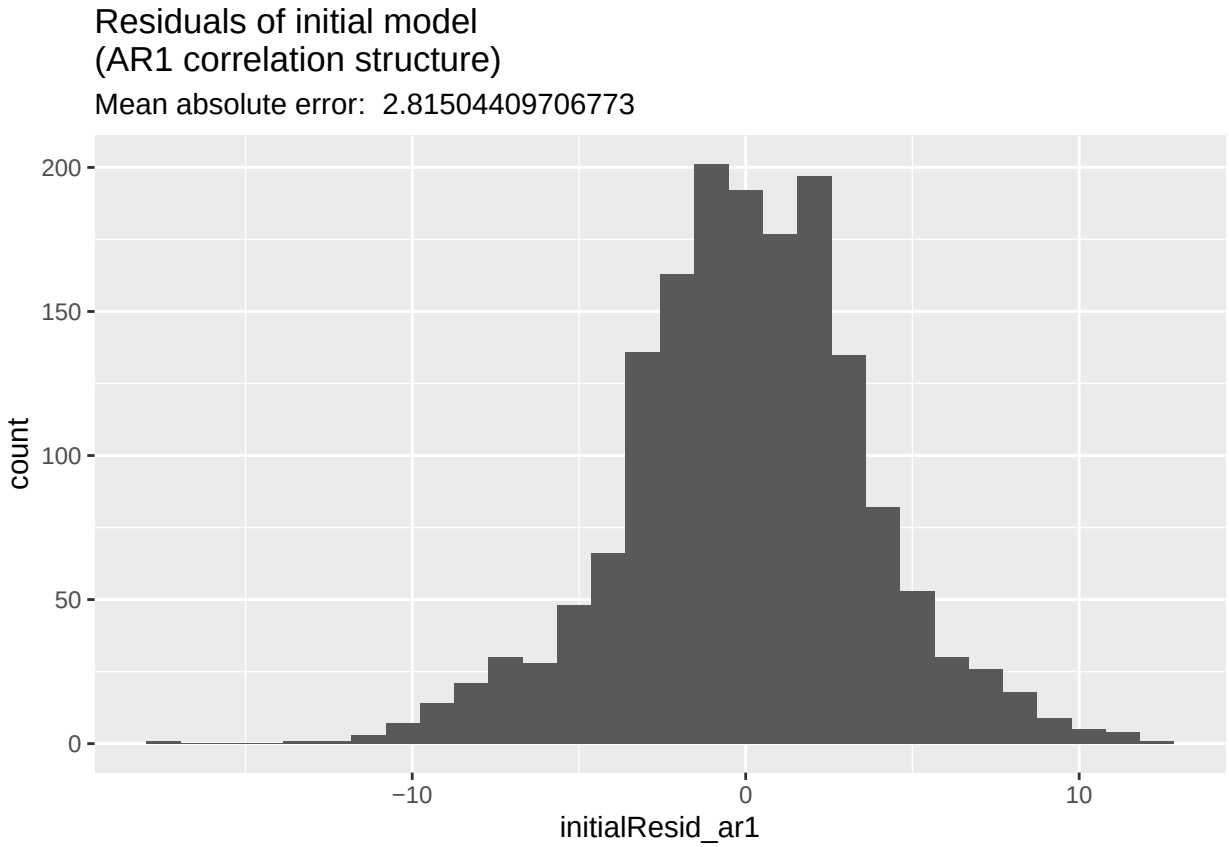
- Status (Developed/Developing)
- Adult mortality rate
- Alcohol intake
- BMI
- HIV/AIDS
- log(GDP)
- Prevalence of thinness among ages 10-19
- Prevalence in thinness among ages 5-9
- Human Development Index (in terms of income composition of resources)
- Number of years of schooling

```r
# fit initial model
initialModel_ar1 <- geeglm(formula = Life.expectancy ~ Status + Year + Adult.Mortality + Alcohol + BMI
                           + HIV.AIDS + log(GDP) + thinness..1.19.years + thinness.5.9.years
                           + Income.composition.of.resources + Schooling, id = Country,
                           data = lifeExpectancy, corstr = "ar1", family = "gaussian")
summary(initialModel_ar1)$coefficients
```

```
##                                     Estimate      Std.err         Wald
## (Intercept)                     -8.615082e+01 5.706060e+01  2.279534421
## StatusDeveloping                -5.824818e+00 1.040977e+00 31.309973394
## Year                             7.304077e-02 2.906636e-02  6.314649336
## Adult.Mortality                 -7.352253e-04 5.368132e-04  1.875834428
## Alcohol                         -5.622125e-03 2.282424e-02  0.060674809
## BMI                              1.463369e-03 2.999144e-03  0.238074626
## HIV.AIDS                        -4.078279e-01 5.870129e-02 48.267915457
## log(GDP)                         1.952508e-03 3.141030e-02  0.003864038
## thinness..1.19.years            -4.008282e-02 3.277668e-02  1.495500857
## thinness.5.9.years               1.521806e-02 2.768839e-02  0.302080819
## Income.composition.of.resources  1.683596e-02 4.707713e-03 12.789572690
## Schooling                        1.103759e+00 1.234923e-01 79.885624828
##                                     Pr(>|W|)
## (Intercept)                     1.310912e-01
## StatusDeveloping                2.199468e-08
## Year                            1.197444e-02
## Adult.Mortality                 1.708083e-01
## Alcohol                         8.054325e-01
## BMI                             6.256002e-01
## HIV.AIDS                        3.717804e-12
## log(GDP)                        9.504343e-01
## thinness..1.19.years            2.213649e-01
## thinness.5.9.years              5.825809e-01
## Income.composition.of.resources 3.485567e-04
## Schooling                       0.000000e+00
```

```r
# plot residuals
initialResid_ar1 <- as.vector(initialModel_ar1$residuals)
mean_abs_resid_initial_ar1 <- mean(abs(initialResid_ar1))

ggplot() +
  geom_histogram(aes(initialResid_ar1)) +
  ggtitle(label = "Residuals of initial model \n(AR1 correlation structure)",
      subtitle = paste("Mean absolute error: ", mean_abs_resid_initial_ar1))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Using AR1 Correlation structure, perform backwards selection of predictors

Model 2 (AR1 Correlation Structure) Predictors:

- Status (Developed/Developing)
- Year
- HIV/AIDS
- Human Development Index (in terms of income composition of resources)
- Number of years of schooling

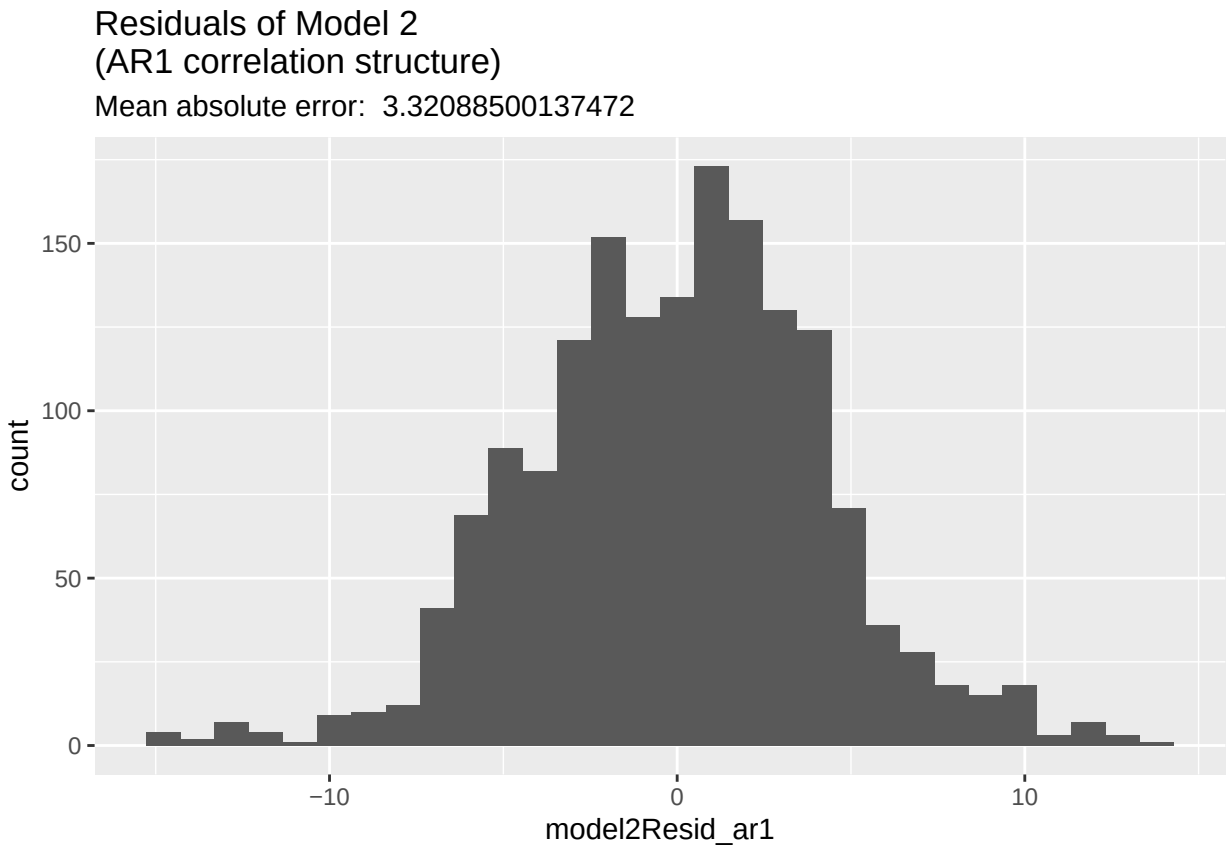
```
# fit initial model
model2_ar1 <- geeglm(formula = Life.expectancy ~ Status + Year + HIV.AIDS
                             + Income.composition.of.resources + Schooling, id = Country,
                             data = lifeExpectancy, corstr = "ar1", family = "gaussian")
summary(model2_ar1)$coefficients
```

```
##                                    Estimate      Std.err      Wald
## (Intercept)                     -92.69412291 56.417504072  2.699459
## StatusDeveloping                 -5.95221369  1.031812649 33.277856
## Year                              0.07619169  0.028783899  7.006739
## HIV.AIDS                         -0.40892008  0.060509651 45.669640
## Income.composition.of.resources   0.01685699  0.004582926 13.529264
## Schooling                         1.11410328  0.122475875 82.746532
##                                      Pr(>|W|)
## (Intercept)                     1.003823e-01
## StatusDeveloping                7.988718e-09
## Year                            8.120345e-03
## HIV.AIDS                        1.399758e-11
```

```
## Income.composition.of.resources 2.348722e-04
## Schooling                       0.000000e+00
```

```
# plot residuals
model2Resid_ar1 <- as.vector(model2_ar1$residuals)
mean_abs_resid_model2_ar1 <- mean(abs(model2Resid_ar1))

ggplot() +
  geom_histogram(aes(model2Resid_ar1)) +
  ggtitle(label = "Residuals of Model 2 \n(AR1 correlation structure)",
      subtitle = paste("Mean absolute error: ", mean_abs_resid_model2_ar1))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Compare Independent vs. AR-1 Standard Error:

(Same predictors in both models)

- Status (Developed/Developing)
- Adult mortality rate
- HIV/AIDS
- Human Development Index (in terms of income composition of resources)
- Number of years of schooling

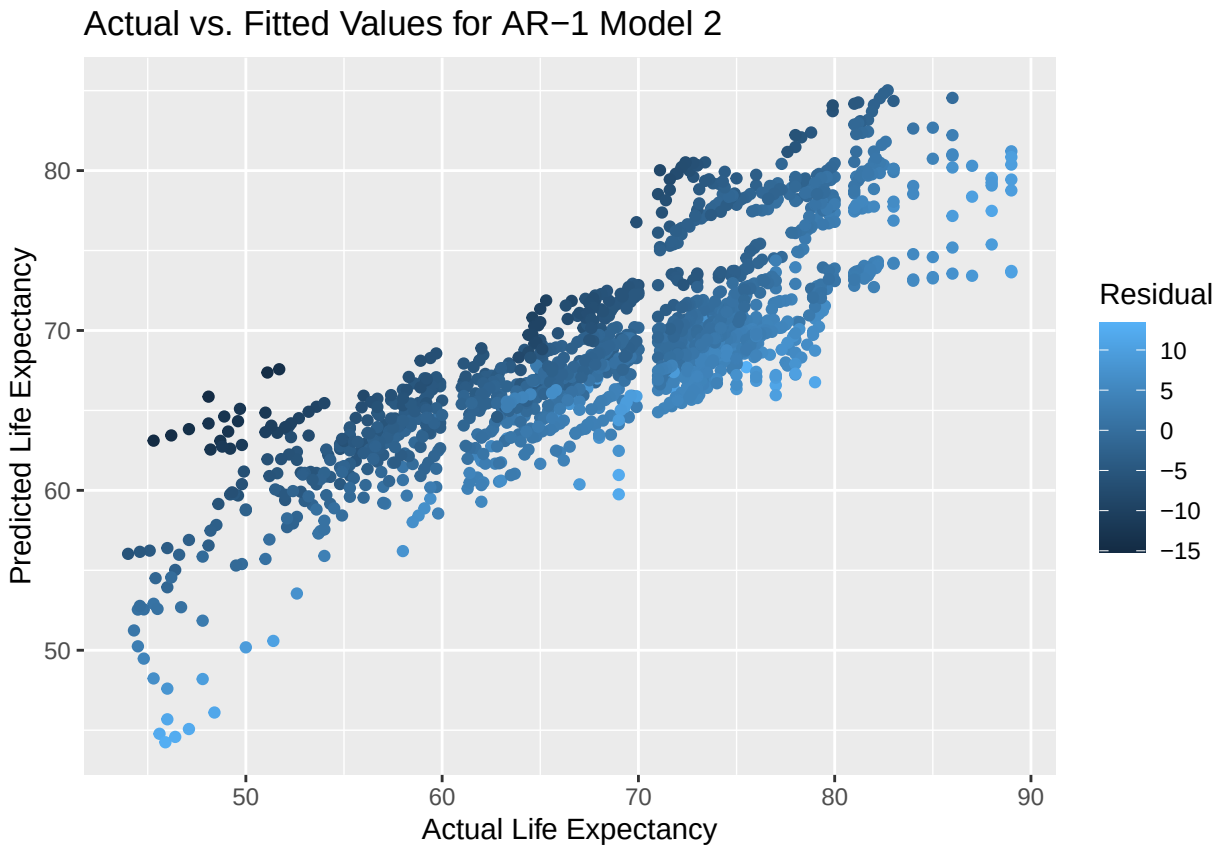
```
# fit independent model w/ same predictors as final AR-1 model to compare standard errors
testIndepModel <- geeglm(formula = Life.expectancy ~ Status + Year + HIV.AIDS
                                 + Income.composition.of.resources + Schooling,
                                 id = Country, data = lifeExpectancy, family = "gaussian")
```

```
summary(testIndepModel)$coefficients
```

```
##                                    Estimate     Std.err      Wald     Pr(>|W|)
## (Intercept)                     360.5058009 84.26742318 18.302275 1.884819e-05
## StatusDeveloping                 -1.2318027  0.97380582  1.600065 2.058940e-01
## Year                             -0.1556290  0.04253417 13.387679 2.532826e-04
## HIV.AIDS                         -0.6413125  0.08679507 54.594540 1.481038e-13
## Income.composition.of.resources   0.1434149  0.03599286 15.876533 6.761189e-05
## Schooling                         1.1997387  0.20442289 34.444057 4.386801e-09
```

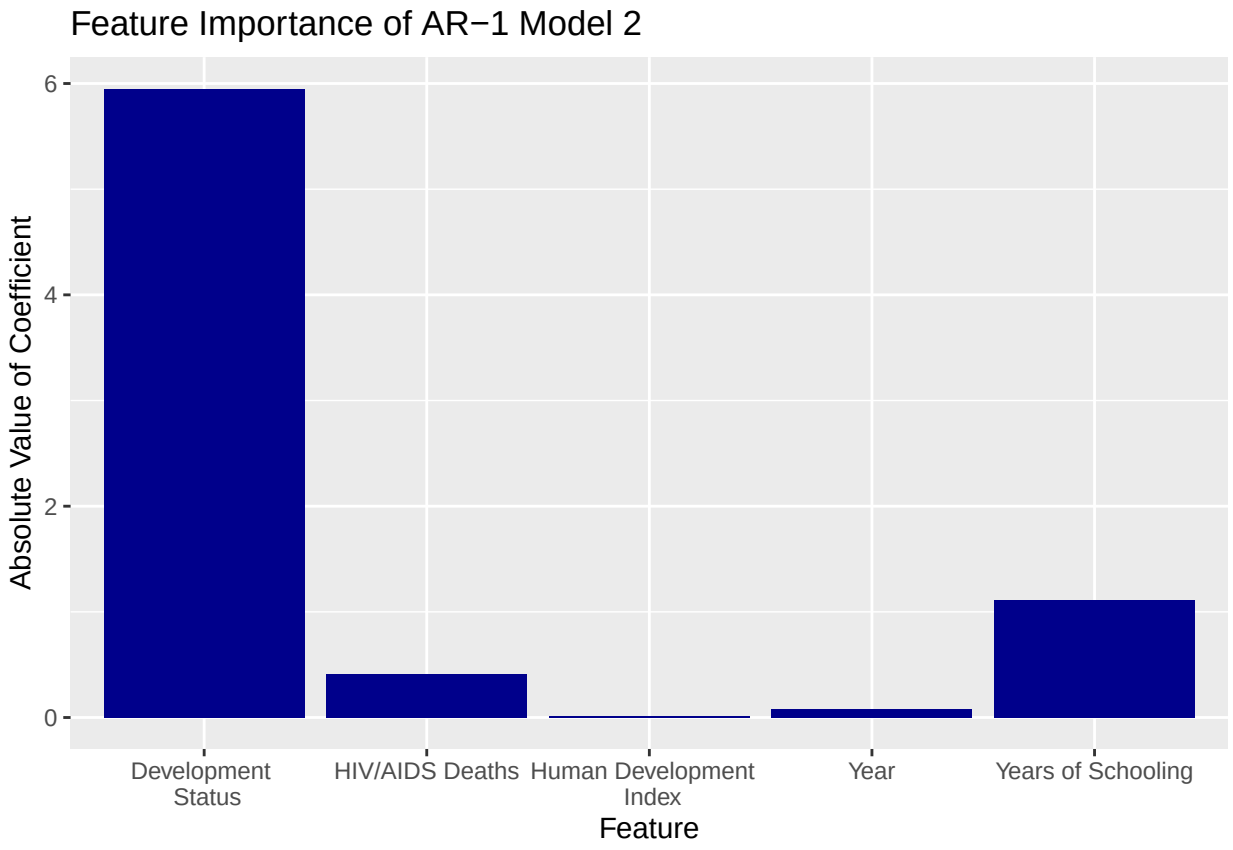
Using AR-1 Model 2, Examine Predicted vs. Actual Values

```
fitted_vals <- model2_ar1$fitted.values
actual_vals <- lifeExpectancy$Life.expectancy
actual_fitted_df <- as.data.frame(cbind(fitted_vals, actual_vals, model2Resid_ar1))
names(actual_fitted_df)[1] <- "fitted_vals"
names(actual_fitted_df)[3] <- "Residual"


ggplot(actual_fitted_df) +
  geom_point(aes(x=actual_vals, fitted_vals, color=Residual)) +
  ggtitle("Actual vs. Fitted Values for AR-1 Model 2") +
  xlab("Actual Life Expectancy") +
  ylab("Predicted Life Expectancy")
```



Feature Importance in AR-1 Model 2

```
feature_names <- c("Development \n Status", "Year", "HIV/AIDS Deaths", "Human Development  \n Index", "Y
abs_coeff <- abs(model2_ar1$coefficients[-1])

ggplot() +
  geom_col(aes(x=feature_names, y=abs_coeff), fill = "dark blue") +
  ggtitle("Feature Importance of AR-1 Model 2") +
  xlab("Feature") +
  ylab("Absolute Value of Coefficient")
```

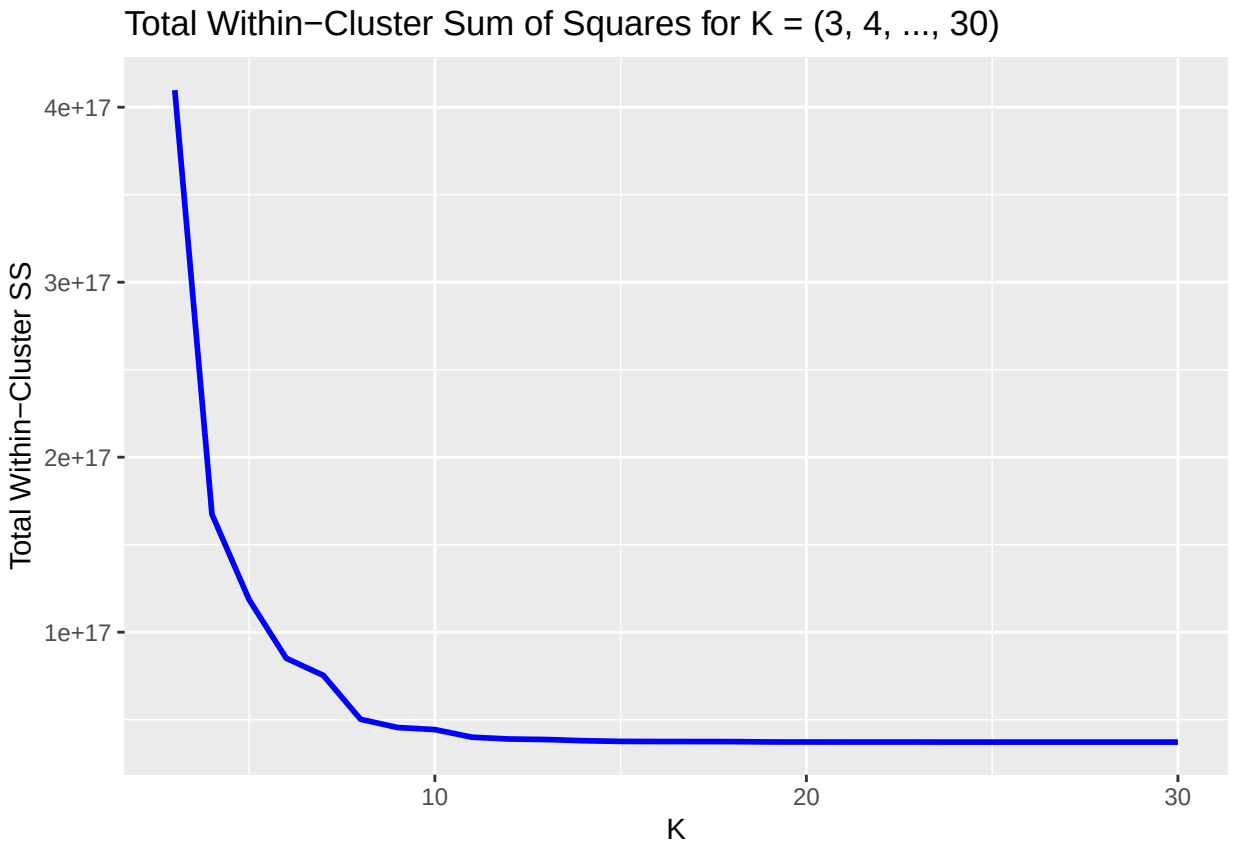


Exploring new clustering based on similarity of countries

K-Means to create new country clusters

```
kmeans_vars <- continuous_vars[, c(-1, -2, -18)]

k_vec <- 3:30
tot_within_SS <- c()

# Cross-Validate to find the best K, which minimizes the total within cluster sum of squares
for (k in k_vec) {

  clustering_k <- kmeans(kmeans_vars, centers=k, nstart=10, iter.max=100)
  tot_within_SS <- c(tot_within_SS, clustering_k$tot.withinss)
```

```
}

# plot K vs. total within-cluster SS
ggplot() +
  geom_line(aes(x=k_vec, y=tot_within_SS), color="blue", lwd = 1) +
  ggtitle("Total Within-Cluster Sum of Squares for K = (3, 4, ..., 30)") +
  xlab("K") +
  ylab("Total Within-Cluster SS")
```



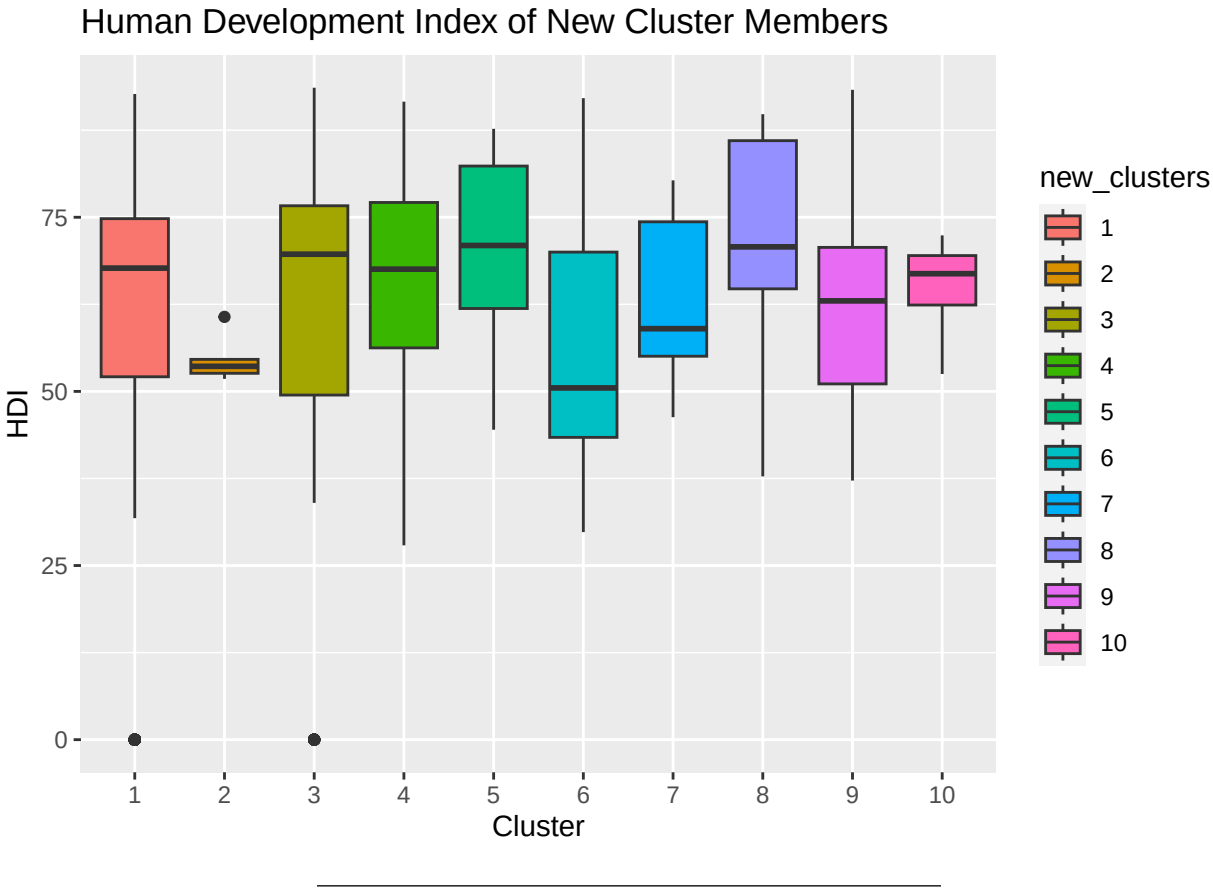
Optimal number of clusters: K = 10

```
# perform final clustering
kmeans10 <- kmeans(kmeans_vars, centers=10, nstart=10)
new_clusters <- factor(kmeans10$cluster)
```

Compare human development index values for cluster members

```
HDI <- lifeExpectancy$Income.composition.of.resources

# get countries belonging to each cluster
lifeExpectancy$cluster <- new_clusters
countries_cluster <- lifeExpectancy %>% group_by(cluster) %>% summarise(countries = str_c(unique(Country

ggplot() +
  geom_boxplot(aes(x=new_clusters, y=HDI, fill=new_clusters)) +
```

```
ggtitle("Human Development Index of New Cluster Members") +
xlab("Cluster")
```



Countries in Cluster 1: Algeria, Argentina, Colombia, Italy, Kenya, Myanmar, Poland, South Africa, Spain, Uganda, Ukraine

Countries in Cluster 2: Afghanistan, Albania, Algeria, Angola, Argentina, Armenia, Australia, Belgium, Bosnia and Herzegovina, Botswana, Brazil, Cameroon, Canada, Central African Republic, Colombia, Costa Rica, Croatia, Eritrea, Ghana, Greece, Indonesia, Iraq, Ireland, Jamaica, Jordan, Kenya, Latvia, Lebanon, Lesotho, Liberia, Lithuania, Madagascar, Malaysia, Mauritania, Mongolia, Morocco, Mozambique, Myanmar, Namibia, Nepal, Nicaragua, Panama, Papua New Guinea, Peru, Poland, Romania, Senegal, South Africa, Spain, Syrian Arab Republic, Tunisia, Turkmenistan, Uganda, Ukraine, Uruguay, Uzbekistan

Countries in Cluster 3: Ethiopia, France, Germany, Philippines, Thailand, Turkey

Countries in Cluster 4: Afghanistan, Algeria, Angola, Australia, Brazil, Canada, Ghana, Indonesia, Iraq, Madagascar, Malaysia, Morocco, Mozambique, Nepal, Peru, Romania, Uganda

Countries in Cluster 5: Brazil, Indonesia, Pakistan

Countries in Cluster 6: Austria, Azerbaijan, Belarus, Belgium, Benin, Bulgaria, Burundi, Chad, Dominican Republic, El Salvador, France, Germany, Greece, Guinea, Honduras, Italy, Jordan,

Mexico, Myanmar, Nicaragua, Papua New Guinea, Paraguay, Philippines, Rwanda, Senegal, Serbia, Sierra Leone, South Africa, Sweden, Tajikistan, Thailand, Togo, Tunisia, Turkey

Countries in Cluster 7: Bangladesh, India, Mexico, Nigeria, Russian Federation

Countries in Cluster 8: Afghanistan, Bangladesh, Brazil, Burkina Faso, Cambodia, Cameroon, Chad, Chile, Ecuador, Ghana, Guatemala, India, Indonesia, Kazakhstan, Madagascar, Malawi, Mali, Mexico, Mozambique, Netherlands, Niger, Nigeria, Pakistan, Romania, Russian Federation, Senegal, Syrian Arab Republic, Zambia, Zimbabwe

Countries in Cluster 9: Afghanistan, Albania, Algeria, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, Bangladesh, Belarus, Belgium, Belize, Benin, Bhutan, Bosnia and Herzegovina, Botswana, Brazil, Bulgaria, Burkina Faso, Burundi, Cabo Verde, Cambodia, Cameroon, Canada, Central African Republic, Chad, Chile, China, Colombia, Comoros, Costa Rica, Croatia, Cyprus, Djibouti, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Eritrea, Estonia, Ethiopia, Fiji, France, Gabon, Georgia, Germany, Ghana, Greece, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Honduras, Iraq, Ireland, Israel, Italy, Jamaica, Jordan, Kazakhstan, Kenya, Kiribati, Latvia, Lebanon, Lesotho, Liberia, Lithuania, Luxembourg, Madagascar, Malawi, Malaysia, Maldives, Mali, Malta, Mauritania, Mauritius, Mexico, Mongolia, Montenegro, Morocco, Mozambique, Myanmar, Namibia, Netherlands, Nicaragua, Niger, Nigeria, Pakistan, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Romania, Russian Federation, Rwanda, Samoa, Sao Tome and Principe, Senegal, Serbia, Seychelles, Sierra Leone, Solomon Islands, South Africa, Spain, Sri Lanka, Suriname, Swaziland, Sweden, Syrian Arab Republic, Tajikistan, Thailand, Timor-Leste, Togo, Tonga, Trinidad and Tobago, Tunisia, Turkey, Turkmenistan, Uganda, Ukraine, Uruguay, Uzbekistan, Vanuatu, Zambia, Zimbabwe

Countries in Cluster 10: India

Fit Model 1 using new clusters

Model 1 (Unstructured correlation structure, clustering by K-Means clusters):

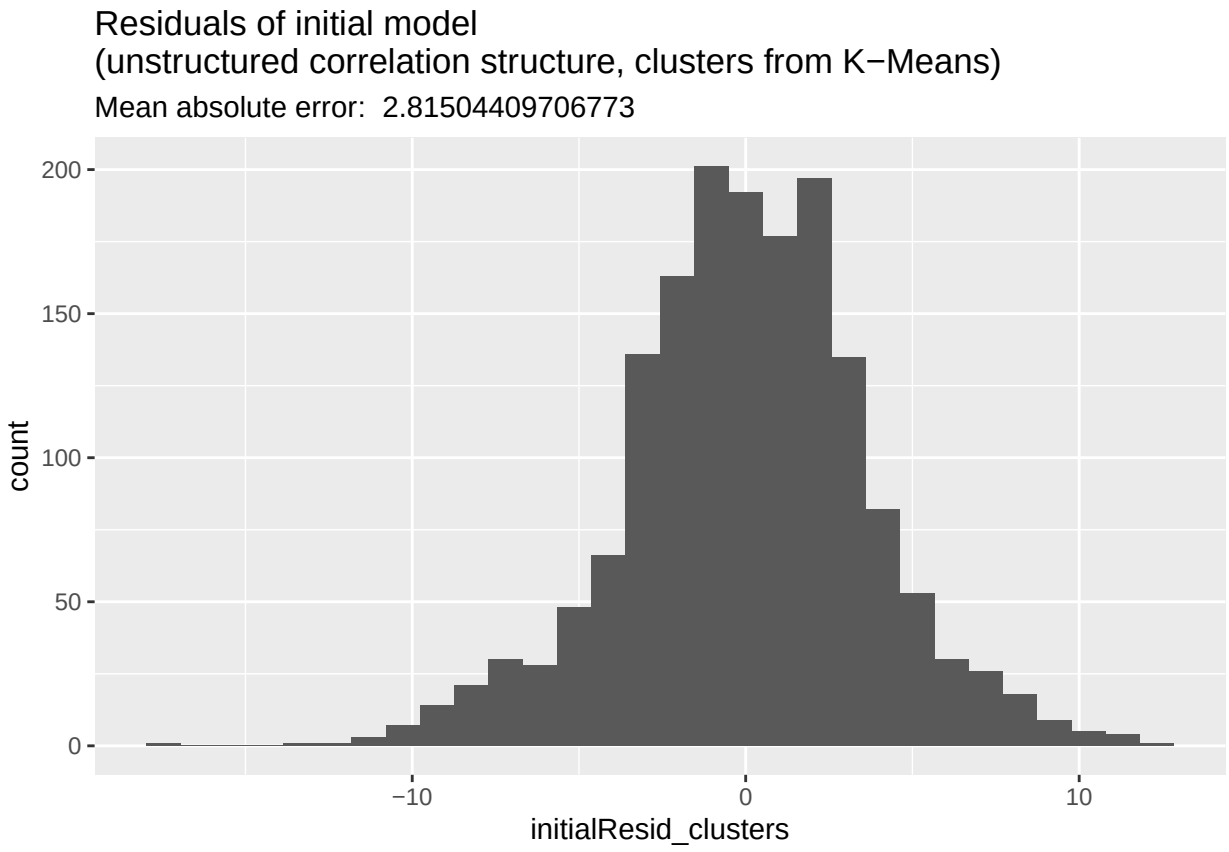
```
# fit initial model
initialModel_clusters <- geeglm(formula = Life.expectancy ~ Status + Year + Adult.Mortality + Alcohol +
                                + HIV.AIDS + log(GDP) + thinness..1.19.years + thinness.5.9.years
                                + Income.composition.of.resources + Schooling, id = cluster,
                                data = lifeExpectancy, corstr = "unstructured", family = "gaussian")
summary(initialModel_clusters)$coefficients
```

```
##                                    Estimate      Std.err       Wald
## (Intercept)                     412.836364926 3.869385e+02  1.1383400
## StatusDeveloping                 -1.473929273 7.888183e-01  3.4913980
## Year                             -0.181148442 1.923979e-01  0.8864789
## Adult.Mortality                  -0.008920334 2.079018e-03 18.4096550
## Alcohol                          -0.236824951 7.681688e-02  9.5047690
## BMI                               0.029508709 2.773239e-02  1.1322070
## HIV.AIDS                         -0.577117408 6.453359e-02 79.9754470
## log(GDP)                          0.316827281 3.834731e-01  0.6826143
## thinness..1.19.years             -0.048453554 6.154738e-02  0.6197723
## thinness.5.9.years               -0.052582088 6.310708e-02  0.6942558
## Income.composition.of.resources   0.123730217 6.308320e-02  3.8470175
```

```
## Schooling                          1.169007353 3.651510e-01 10.2491891
##                                      Pr(>|W|)
## (Intercept)                       0.2860033650
## StatusDeveloping                  0.0616884688
## Year                              0.3464332873
## Adult.Mortality                   0.0000178153
## Alcohol                           0.0020493855
## BMI                               0.2873050751
## HIV.AIDS                          0.0000000000
## log(GDP)                          0.4086879505
## thinness..1.19.years              0.4311318977
## thinness.5.9.years                0.4047205777
## Income.composition.of.resources 0.0498345339
## Schooling                         0.0013674470
```

```r
# plot residuals
initialResid_clusters <- as.vector(initialModel_clusters$residuals)
mean_abs_resid_initial_clusters <- mean(abs(initialResid_clusters))

ggplot() +
  geom_histogram(aes(initialResid_clusters)) +
  ggtitle(label = "Residuals of initial model \n(unstructured correlation structure, clusters from K-Mea
      subtitle = paste("Mean absolute error: ", mean_abs_resid_initial_clusters))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Comparing standard errors from clustering by Country vs. K-Means

```
summary(initialModel_ar1)$coefficients
```

```
##                                     Estimate      Std.err         Wald
## (Intercept)                     -8.615082e+01 5.706060e+01  2.279534421
## StatusDeveloping                -5.824818e+00 1.040977e+00 31.309973394
## Year                             7.304077e-02 2.906636e-02  6.314649336
## Adult.Mortality                 -7.352253e-04 5.368132e-04  1.875834428
## Alcohol                         -5.622125e-03 2.282424e-02  0.060674809
## BMI                              1.463369e-03 2.999144e-03  0.238074626
## HIV.AIDS                        -4.078279e-01 5.870129e-02 48.267915457
## log(GDP)                         1.952508e-03 3.141030e-02  0.003864038
## thinness..1.19.years            -4.008282e-02 3.277668e-02  1.495500857
## thinness.5.9.years               1.521806e-02 2.768839e-02  0.302080819
## Income.composition.of.resources  1.683596e-02 4.707713e-03 12.789572690
## Schooling                        1.103759e+00 1.234923e-01 79.885624828
##                                     Pr(>|W|)
## (Intercept)                     1.310912e-01
## StatusDeveloping                2.199468e-08
## Year                            1.197444e-02
## Adult.Mortality                 1.708083e-01
## Alcohol                         8.054325e-01
## BMI                             6.256002e-01
## HIV.AIDS                        3.717804e-12
## log(GDP)                        9.504343e-01
## thinness..1.19.years            2.213649e-01
## thinness.5.9.years              5.825809e-01
## Income.composition.of.resources 3.485567e-04
## Schooling                       0.000000e+00
```

```
summary(initialModel_clusters)$coefficients
```

```
##                                     Estimate      Std.err       Wald
## (Intercept)                     412.836364926 3.869385e+02  1.1383400
## StatusDeveloping                 -1.473929273 7.888183e-01  3.4913980
## Year                             -0.181148442 1.923979e-01  0.8864789
## Adult.Mortality                  -0.008920334 2.079018e-03 18.4096550
## Alcohol                          -0.236824951 7.681688e-02  9.5047690
## BMI                               0.029508709 2.773239e-02  1.1322070
## HIV.AIDS                         -0.577117408 6.453359e-02 79.9754470
## log(GDP)                          0.316827281 3.834731e-01  0.6826143
## thinness..1.19.years             -0.048453554 6.154738e-02  0.6197723
## thinness.5.9.years               -0.052582088 6.310708e-02  0.6942558
## Income.composition.of.resources   0.123730217 6.308320e-02  3.8470175
## Schooling                         1.169007353 3.651510e-01 10.2491891
##                                     Pr(>|W|)
## (Intercept)                     0.2860033650
## StatusDeveloping                0.0616884688
## Year                            0.3464332873
## Adult.Mortality                 0.0000178153
## Alcohol                         0.0020493855
## BMI                             0.2873050751
## HIV.AIDS                        0.0000000000
## log(GDP)                        0.4086879505
```

```
## thinness..1.19.years            0.4311318977
## thinness.5.9.years              0.4047205777
## Income.composition.of.resources 0.0498345339
## Schooling                       0.0013674470
```