Final Project - WHO Data Set

Douglas Barley, Ethan Haley, Isabel Magnus, John Mazon, Vinayak Kamath, Arushi Arora

11/22/2021

## Loading Data Set - WHO Life Expectancy Data

EDA and modeling

```
#WHO_Final <- WHO_with_Region
WHO_URL <- "https://raw.githubusercontent.com/ebhtra/msds-621/main/FinalProject/fewerNAs.csv"
#https://raw.githubusercontent.com/ebhtra/msds-621/main/FinalProject/finalProjDF.csv"
WHO_Final <- read_csv(WHO_URL )
```

```
## Rows: 672 Columns: 30
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): CountryName, Status, CountryCode, Region, IncomeGroup
## dbl (25): Year, InfantDeaths, Alcohol, Measles, under-five deaths, Polio, To...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

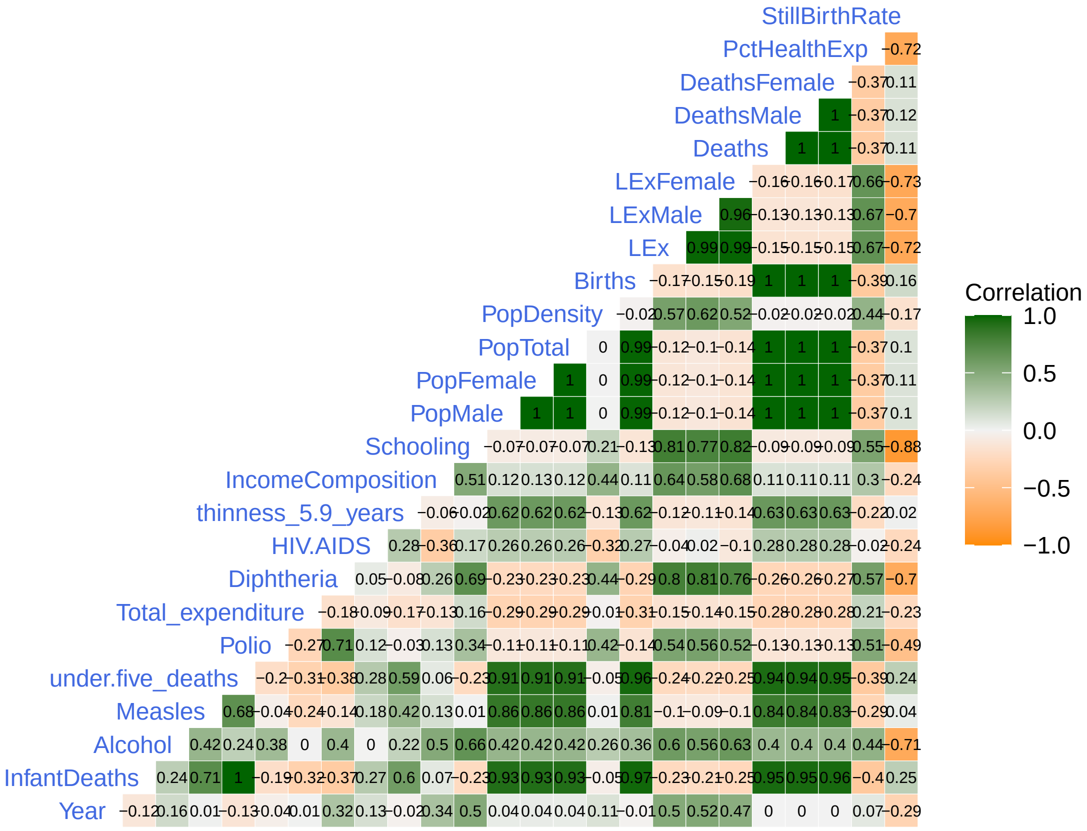
```
# Dropping NA
WHO_Final <- WHO_Final %>%
  drop_na()

# renaming columns
#WHO_Final <- WHO_Final %>%
#  rename(LifeExpectancy = "Life expectancy", AdultMortality = "Adult Mortality", InfantDeaths = "infan
#         , IncomeComposition ="Income composition of resources")

summary(WHO_Final)
```

```
##  CountryName             Year         Status            InfantDeaths
##  Length:636         Min.   :2000   Length:636         Min.   :   0.00
##  Class :character   1st Qu.:2004   Class :character   1st Qu.:   0.00
##  Mode  :character   Median :2008   Mode  :character   Median :   3.00
##                     Mean   :2008                      Mean   :  30.39
##                      3rd Qu.:2011                      3rd Qu.:  18.00  
##                      Max.   :2015                      Max.   :1800.00  
##     Alcohol          Measles          under-five deaths     Polio      
##  Min.   : 0.010   Min.   :     0.0   Min.   :   0.00   Min.   : 3.00  
##  1st Qu.: 1.340   1st Qu.:     0.0   1st Qu.:   0.00   1st Qu.:75.75  
##  Median : 4.285   Median :    15.0   Median :   3.00   Median :93.00  
##  Mean   : 4.958   Mean   :  2732.2   Mean   :  42.08   Mean   :81.22  
##  3rd Qu.: 7.920   3rd Qu.:   337.5   3rd Qu.:  23.25   3rd Qu.:97.00  
##  Max.   :15.520   Max.   :212183.0   Max.   :2500.00   Max.   :99.00  
##  Total expenditure   Diphtheria      HIV/AIDS     thinness 5-9 years
##  Min.   : 0.920    Min.   : 3.0   Min.   : 0.10   Min.   : 0.100    
##  1st Qu.: 4.338    1st Qu.:77.0   1st Qu.: 0.10   1st Qu.: 1.500    
##  Median : 5.825    Median :93.0   Median : 0.10   Median : 3.200    
##  Mean   : 5.990    Mean   :81.2   Mean   : 1.51   Mean   : 4.849    
##  3rd Qu.: 7.803    3rd Qu.:97.0   3rd Qu.: 0.60   3rd Qu.: 7.200    
##  Max.   :13.830    Max.   :99.0   Max.   :49.10   Max.   :28.600    
##  IncomeComposition   Schooling     CountryCode           Region         
##  Min.   :0.0000    Min.   : 0.00   Length:636         Length:636        
##  1st Qu.:0.4918    1st Qu.:10.10   Class :character   Class :character  
##  Median :0.6760    Median :12.40   Mode  :character   Mode  :character  
##  Mean   :0.6178    Mean   :12.02                                        
##  3rd Qu.:0.7792    3rd Qu.:14.32                                        
##  Max.   :0.9480    Max.   :20.40                                        
##  IncomeGroup           PopMale           PopFemale           PopTotal      
##  Length:636         Min.   :    35.7   Min.   :    40.3   Min.   :     76  
##  Class :character   1st Qu.:   987.7   1st Qu.:  1015.6   1st Qu.:   2026  
##  Mode  :character   Median :  3622.1   Median :  3691.3   Median :   7317  
##                     Mean   : 18906.2   Mean   : 18574.6   Mean   :  37481  
##                     3rd Qu.: 11826.6   3rd Qu.: 12007.2   3rd Qu.:  23561  
##                     Max.   :722508.0   Max.   :684339.9   Max.   :1406848  
##    PopDensity           Births               LEx           LExMale     
##  Min.   :   1.543   Min.   :     7.22   Min.   :37.61   Min.   :37.14  
##  1st Qu.:  29.482   1st Qu.:   187.11   1st Qu.:62.32   1st Qu.:60.06  
##  Median :  75.680   Median :   680.70   Median :70.69   Median :67.81  
##  Mean   : 181.127   Mean   :  3734.26   Mean   :68.41   Mean   :65.96  
##  3rd Qu.: 150.295   3rd Qu.:  2854.69   3rd Qu.:75.06   3rd Qu.:72.58  
##  Max.   :7988.776   Max.   :139249.38   Max.   :83.32   Max.   :80.58  
##    LExFemale         Deaths           DeathsMale          DeathsFemale     
##  Min.   :38.08   Min.   :    2.35   Min.   :    1.151   Min.   :    1.191  
##  1st Qu.:64.08   1st Qu.:   88.96   1st Qu.:   46.786   1st Qu.:   39.626  
##  Median :73.83   Median :  301.44   Median :  157.033   Median :  141.575  
##  Mean   :70.90   Mean   : 1476.66   Mean   :  790.999   Mean   :  685.660  
##  3rd Qu.:78.05   3rd Qu.:  902.25   3rd Qu.:  476.014   3rd Qu.:  428.293  
##  Max.   :86.47   Max.   :48592.46   Max.   :27361.746   Max.   :21894.284  
##   PctHealthExp    StillBirthRate  
##  Min.   : 1.150   Min.   : 1.790  
##  1st Qu.: 6.287   1st Qu.: 4.925  
##  Median : 9.200   Median :10.875  
##  Mean   : 9.697   Mean   :12.877  
##  3rd Qu.:12.675   3rd Qu.:17.698  
##  Max.   :31.960   Max.   :44.800
```

```
#calculate mean of each death column
WHO_Final %>%
    group_by(Status) %>%
    summarize(count = n(),
             LifExpMean = mean(LEx, na.rm=TRUE),
             DeathsMean = mean(Deaths, na.rm=TRUE),
             InfdeaMean = mean(InfantDeaths, na.rm=TRUE)
             )
```

```
## # A tibble: 2 x 5
##   Status     count LifExpMean DeathsMean InfdeaMean
##   <chr>      <int>      <dbl>      <dbl>      <dbl>
## 1 Developed    116       77.9       844.      0.612
## 2 Developing   520       66.3      1618.     37.0
```

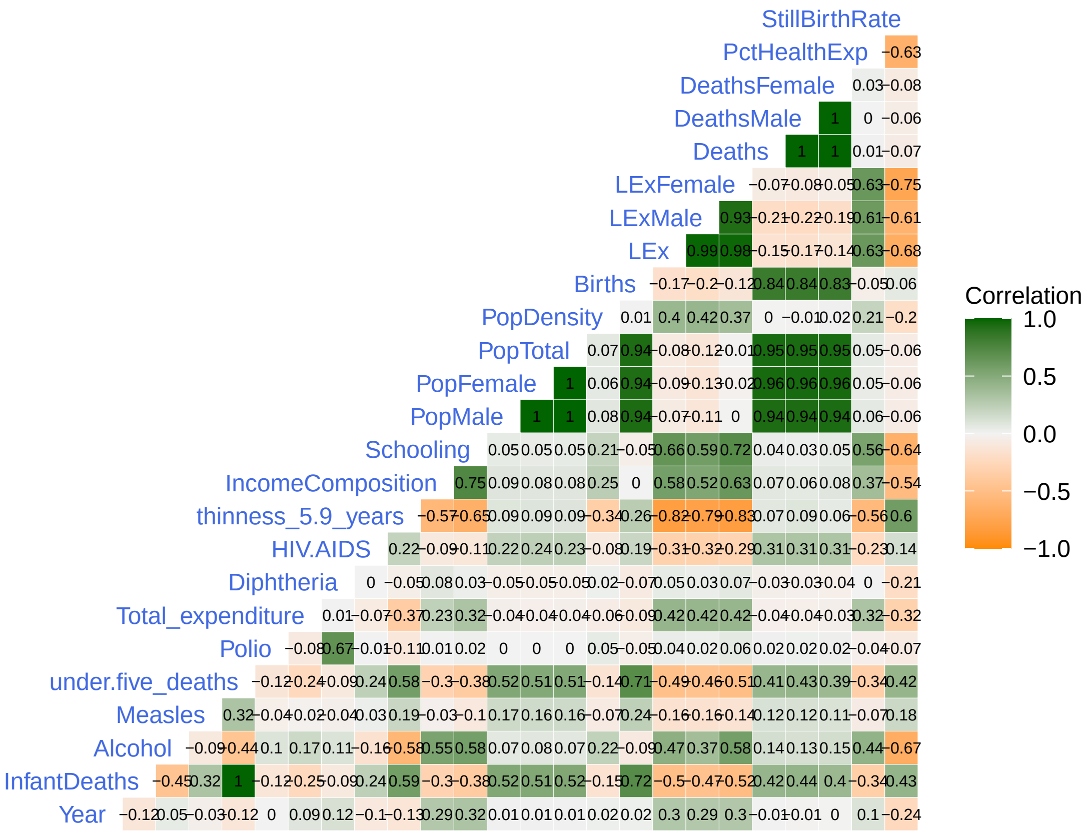
```
#Correlation between variables

#Region=='South Asia'
WHO_Final_Numeric <- WHO_Final %>%
  filter( Region=='South Asia') %>%
  select_if(is.numeric)

ggcorr(WHO_Final_Numeric,
       label = T,
       label_size = 2,
       label_round = 2,
       hjust = 1,
       size = 3,
       color = "royalblue",
       layout.exp = 5,
       low = "darkorange",
       mid = "gray95",
       high = "darkgreen",
       name = "Correlation")
```



```
lm.region <- lm(formula = LEx ~ . , data = WHO_Final_Numeric)
summary(lm.region)
```

```
##
## Call:
## lm(formula = LEx ~ ., data = WHO_Final_Numeric)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0231830 -0.0088617  0.0003408  0.0073353  0.0277118
##
## Coefficients: (2 not defined because of singularities)
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.676e+01  4.758e+00  -3.523  0.00648 **
## Year                  8.618e-03  2.484e-03   3.469  0.00706 **
## InfantDeaths          1.890e-03  1.339e-03   1.412  0.19162
## Alcohol               1.011e-02  2.800e-02   0.361  0.72636
## Measles               1.116e-06  1.264e-06   0.883  0.40004
## `under-five deaths`  -1.205e-03  6.995e-04  -1.722  0.11916
## Polio                 8.925e-04  4.026e-04   2.217  0.05383 .
## `Total expenditure`  -2.267e-02  8.843e-03  -2.564  0.03049 *
## Diphtheria           -5.571e-04  1.371e-03  -0.406  0.69406
## `HIV/AIDS`           -1.010e-01  2.108e-01  -0.479  0.64341
## `thinness 5-9 years`  1.725e-03  1.289e-03   1.338  0.21365
## IncomeComposition     5.858e-02  7.141e-02   0.820  0.43324
## Schooling           -1.793e-02  2.046e-02  -0.876  0.40373
## PopMale             -3.356e-06  1.395e-05  -0.241  0.81526
## PopFemale            1.239e-06  1.530e-05   0.081  0.93723
## PopTotal                    NA         NA      NA       NA
## PopDensity          -5.438e-06  4.590e-05  -0.118  0.90829
## Births              -2.506e-05  1.448e-05  -1.730  0.11772
## LExMale              5.112e-01  1.664e-02  30.720 2.01e-10 ***
## LExFemale            4.868e-01  1.479e-02  32.909 1.09e-10 ***
## Deaths               6.617e-04  2.360e-04   2.804  0.02057 *
## DeathsMale          -1.077e-03  4.636e-04  -2.323  0.04527 *
## DeathsFemale                NA         NA      NA       NA
## PctHealthExp        -1.375e-02  4.626e-03  -2.972  0.01565 *
## StillBirthRate      -1.667e-03  2.551e-03  -0.653  0.52993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02245 on 9 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 8.587e+04 on 22 and 9 DF,  p-value: < 2.2e-16
```

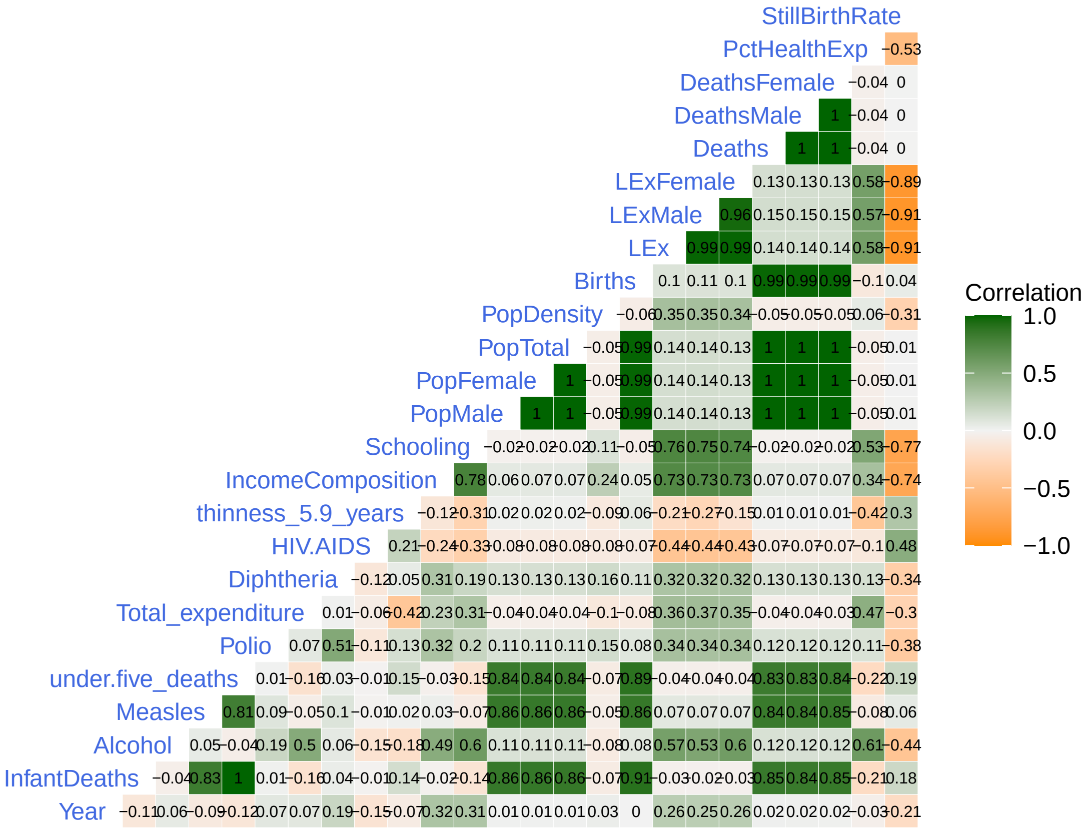
```
#Region=='Europe & Central Asia'
WHO_Final_Numeric <- WHO_Final %>%
  filter( Region=='Europe & Central Asia') %>%
  select_if(is.numeric)

ggcorr(WHO_Final_Numeric,
       label = T,
       label_size = 2,
       label_round = 2,
       hjust = 1,
       size = 3,
       color = "royalblue",
       layout.exp = 5,
       low = "darkorange",
       mid = "gray95",
       high = "darkgreen",
       name = "Correlation")
```



```
lm.region <- lm(formula = LEx ~ . , data = WHO_Final_Numeric)
summary(lm.region)
```

```
##
## Call:
## lm(formula = LEx ~ ., data = WHO_Final_Numeric)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.198829 -0.023877  0.001607  0.025271  0.164385
##
## Coefficients: (2 not defined because of singularities)
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.203e+00  1.672e+00  -1.317 0.189821
## Year                  4.293e-04  8.523e-04   0.504 0.615201
## InfantDeaths          7.745e-03  1.252e-02   0.619 0.537116
## Alcohol               4.445e-03  1.895e-03   2.346 0.020236 *
## Measles              -3.174e-06  1.492e-06  -2.127 0.035002 *
## `under-five deaths`  -4.982e-03  1.035e-02  -0.481 0.631071
## Polio                -1.246e-04  4.028e-04  -0.309 0.757451
## `Total expenditure`  -1.039e-04  1.846e-03  -0.056 0.955170
## Diphtheria           -2.397e-04  3.298e-04  -0.727 0.468397
## `HIV/AIDS`           -6.070e-02  5.757e-02  -1.054 0.293424
## `thinness 5-9 years`  2.290e-02  8.187e-03   2.797 0.005822 **
## IncomeComposition    -2.651e-02  3.343e-02  -0.793 0.428949
## Schooling           -6.297e-05  2.937e-03   -0.021 0.982923
## PopMale             -6.455e-05  1.809e-05   -3.567 0.000482 ***
## PopFemale            7.015e-05  2.182e-05    3.215 0.001592 **
## PopTotal                    NA         NA       NA       NA
## PopDensity           7.356e-05  5.522e-05    1.332 0.184806
## Births              -1.127e-05  2.506e-05   -0.450 0.653569
## LExMale              4.972e-01  3.091e-03 160.833  < 2e-16 ***
## LExFemale            5.192e-01  4.715e-03 110.108  < 2e-16 ***
## Deaths               3.687e-04  9.513e-05    3.876 0.000157 ***
## DeathsMale          -9.196e-04  1.550e-04   -5.933 1.91e-08 ***
## DeathsFemale                NA         NA       NA       NA
## PctHealthExp        -8.390e-04  1.842e-03   -0.455 0.649455
## StillBirthRate       8.925e-03  2.236e-03    3.991 0.000102 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05211 on 153 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 7.664e+04 on 22 and 153 DF,  p-value: < 2.2e-16
```

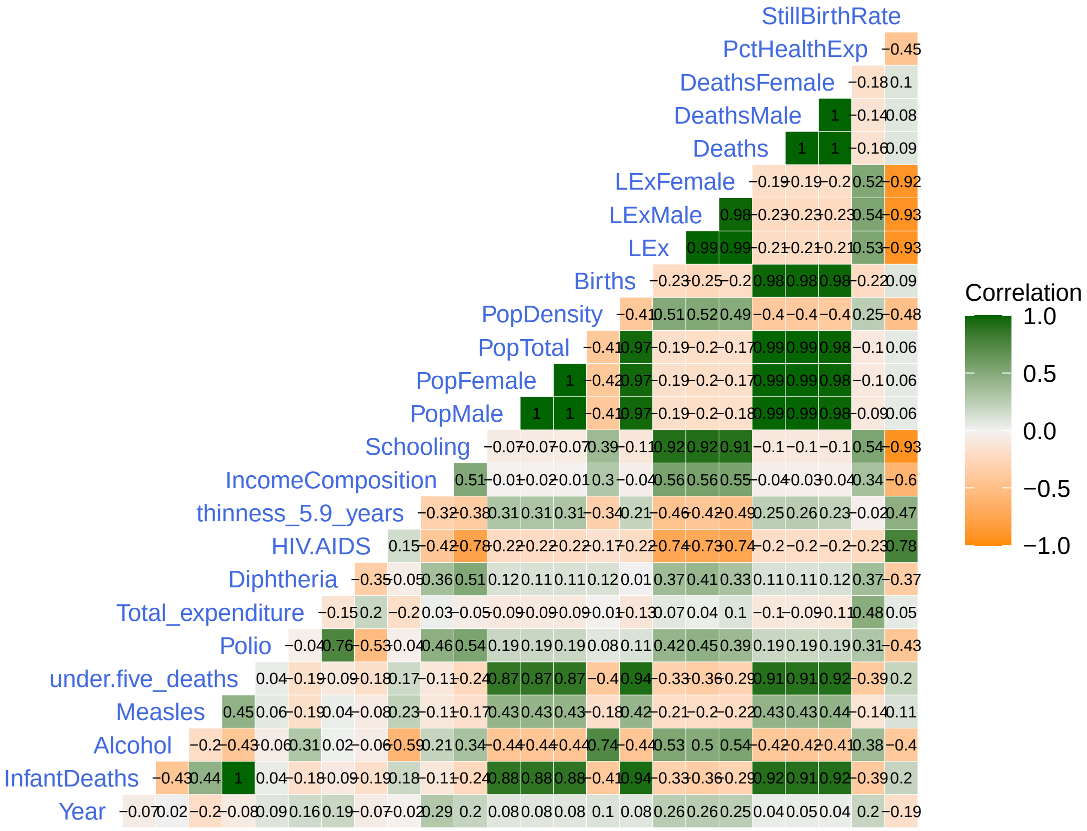
```
#Region=='East Asia & Pacific'
WHO_Final_Numeric <- WHO_Final %>%
  filter( Region=='East Asia & Pacific') %>%
  select_if(is.numeric)

ggcorr(WHO_Final_Numeric,
       label = T,
       label_size = 2,
       label_round = 2,
       hjust = 1,
       size = 3,
       color = "royalblue",
       layout.exp = 5,
       low = "darkorange",
       mid = "gray95",
       high = "darkgreen",
       name = "Correlation")
```



```r
lm.region <- lm(formula = LEx ~ . , data = WHO_Final_Numeric)
summary(lm.region)
```

```
##
## Call:
## lm(formula = LEx ~ ., data = WHO_Final_Numeric)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.08368 -0.02528 -0.00445  0.02203  0.11650
##
## Coefficients: (2 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.690e+00  2.100e+00   0.805  0.42343
## Year                   -1.158e-03  1.086e-03  -1.066  0.29014
## InfantDeaths            2.219e-03  6.242e-03   0.355  0.72326
## Alcohol                 2.971e-04  2.921e-03   0.102  0.91927
## Measles                -2.487e-07  8.512e-07  -0.292  0.77101
## `under-five deaths`    -1.549e-03  4.703e-03  -0.329  0.74287
## Polio                  -2.229e-04  2.245e-04  -0.993  0.32395
## `Total expenditure`     3.858e-03  2.327e-03   1.658  0.10170
## Diphtheria              3.794e-04  2.376e-04   1.597  0.11466
## `HIV/AIDS`              2.015e-03  1.841e-02   0.109  0.91314
## `thinness 5-9 years`    1.640e-03  1.609e-03   1.019  0.31158
## IncomeComposition      -3.592e-02  4.277e-02  -0.840  0.40373
## Schooling             1.331e-02  3.064e-03   4.346 4.41e-05 ***
## PopMale              -2.301e-05  5.582e-06  -4.121 9.83e-05 ***
## PopFemale             2.575e-05  7.636e-06   3.372  0.00120 **
## PopTotal                     NA         NA      NA       NA
## PopDensity            1.364e-05  4.701e-06   2.902  0.00490 **
## Births               -1.430e-05  5.258e-06  -2.720  0.00815 **
## LExMale               5.128e-01  3.744e-03 136.951  < 2e-16 ***
## LExFemale             4.920e-01  4.601e-03 106.931  < 2e-16 ***
## Deaths                1.703e-05  5.509e-05   0.309  0.75809
## DeathsMale           -2.540e-05  1.140e-04  -0.223  0.82424
## DeathsFemale                 NA         NA      NA       NA
## PctHealthExp         -2.144e-03  1.707e-03  -1.257  0.21294
## StillBirthRate        1.183e-02  2.806e-03   4.217 7.00e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04437 on 73 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 9.525e+04 on 22 and 73 DF,  p-value: < 2.2e-16
```

```
#Region=='Middle East & North Africa'
WHO_Final_Numeric <- WHO_Final %>%
  filter( Region=='Middle East & North Africa') %>%
  select_if(is.numeric)

ggcorr(WHO_Final_Numeric,
       label = T,
       label_size = 2,
       label_round = 2,
       hjust = 1,
       size = 3,
       color = "royalblue",
       layout.exp = 5,
       low = "darkorange",
       mid = "gray95",
       high = "darkgreen",
       name = "Correlation")
```



```
lm.region <- lm(formula = LEx ~ . , data = WHO_Final_Numeric)
summary(lm.region)
```

```
##
## Call:
## lm(formula = LEx ~ ., data = WHO_Final_Numeric)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.055629 -0.015828 -0.004236  0.023037  0.068263
##
## Coefficients: (2 not defined because of singularities)
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.664e-02  2.730e+00   0.028 0.977797
## Year                  2.289e-04  1.514e-03   0.151 0.880839
## InfantDeaths          6.054e-03  1.455e-02   0.416 0.680415
## Alcohol               1.138e-02  9.636e-03   1.181 0.247071
## Measles               4.443e-06  3.838e-06   1.158 0.256391
## `under-five deaths`  -2.420e-03  1.086e-02  -0.223 0.825224
## Polio                 5.193e-04  1.043e-03   0.498 0.622323
## `Total expenditure`   5.029e-03  4.871e-03   1.032 0.310435
## Diphtheria           -5.141e-04  8.241e-04  -0.624 0.537596
## `HIV/AIDS`            4.286e-02  2.532e-02   1.693 0.101197
## `thinness 5-9 years` -1.415e-02  9.834e-03  -1.439 0.160933
## IncomeComposition    -1.151e-01  4.926e-02  -2.336 0.026612 *
## Schooling             5.579e-03  1.083e-02   0.515 0.610258
## PopMale              -1.318e-04  3.807e-05  -3.463 0.001681 **
## PopFemale             1.678e-04  4.131e-05   4.061 0.000339 ***
## PopTotal                     NA         NA      NA       NA
## PopDensity            2.743e-05  3.412e-05   0.804 0.428104
## Births               -4.987e-05  2.155e-05  -2.315 0.027913 *
## LExMale               4.824e-01  1.476e-02  32.688  < 2e-16 ***
## LExFemale             5.097e-01  1.173e-02  43.442  < 2e-16 ***
## Deaths                8.196e-04  5.185e-04   1.581 0.124779
## DeathsMale           -2.277e-03  1.002e-03  -2.273 0.030597 *
## DeathsFemale                 NA         NA      NA       NA
## PctHealthExp          4.057e-05  3.142e-03   0.013 0.989786
## StillBirthRate       -7.422e-03  6.025e-03  -1.232 0.227922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03702 on 29 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 5.15e+04 on 22 and 29 DF,  p-value: < 2.2e-16
```

Time Series Analysis

```
# subset the data
WHO_Final_RegionSA <- subset(WHO_Final, Region=='South Asia')

WHO_Final_RegionSA2000 <- subset(WHO_Final_RegionSA, Year == "2000")
WHO_Final_RegionSA2015 <- subset(WHO_Final_RegionSA, Year == "2015")

# estimate simple regression models using 1982 and 1988 data
who2000_mod <- lm(LEx ~ Schooling, data = WHO_Final_RegionSA2000)
who2015_mod <- lm(LEx ~ Schooling, data = WHO_Final_RegionSA2015)


coeftest(who2000_mod, vcov. = vcovHC, type = "HC1")
```

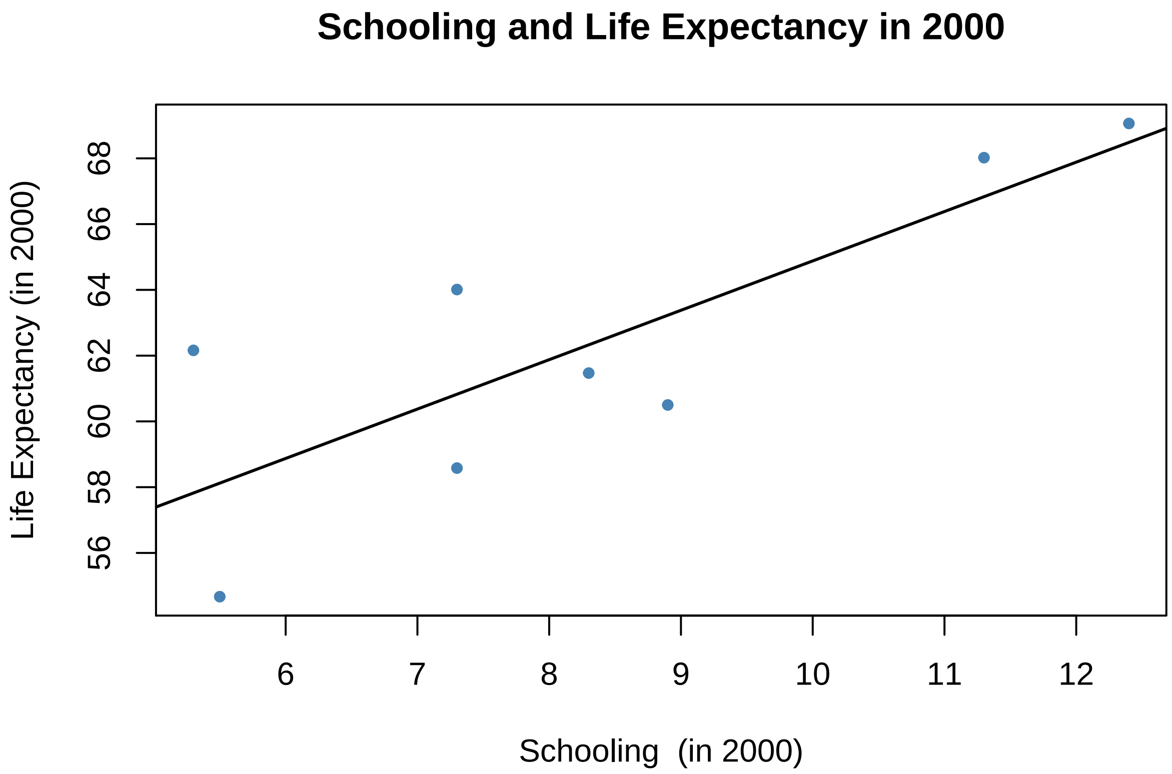
```
##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 49.86318    4.51156 11.0523 3.265e-05 ***
## Schooling    1.50173    0.44179  3.3992   0.01451 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(who2015_mod, vcov. = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
```

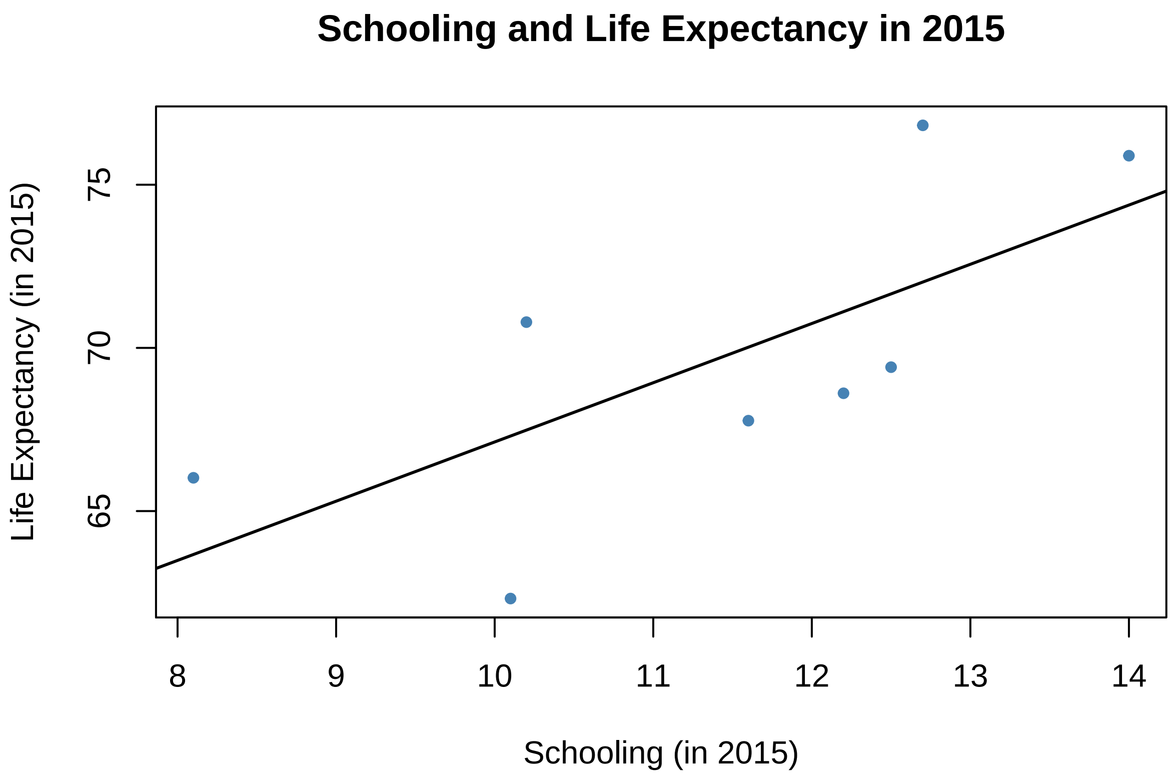
```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 48.96835    7.64208  6.4077 0.0006813 ***
## Schooling    1.81491    0.64216  2.8262 0.0301059 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# plot the observations and add the estimated regression line for 2000 data
plot(x = WHO_Final_RegionSA2000$Schooling,
     y = WHO_Final_RegionSA2000$LEx,
     xlab = "Schooling  (in 2000)",
     ylab = "Life Expectancy (in 2000)",
     main = "Schooling and Life Expectancy in 2000",
     #ylim = c(0, 4.5),
     pch = 20,
     col = "steelblue")

abline(who2000_mod, lwd = 1.5)
```



```
# plot the observations and add the estimated regression line for 2015 data
plot(x = WHO_Final_RegionSA2015$Schooling,
     y = WHO_Final_RegionSA2015$LEx,
     xlab = "Schooling (in 2015)",
     ylab = "Life Expectancy (in 2015)",
     main = "Schooling and Life Expectancy in 2015",
     #ylim = c(0, 4.5),
```

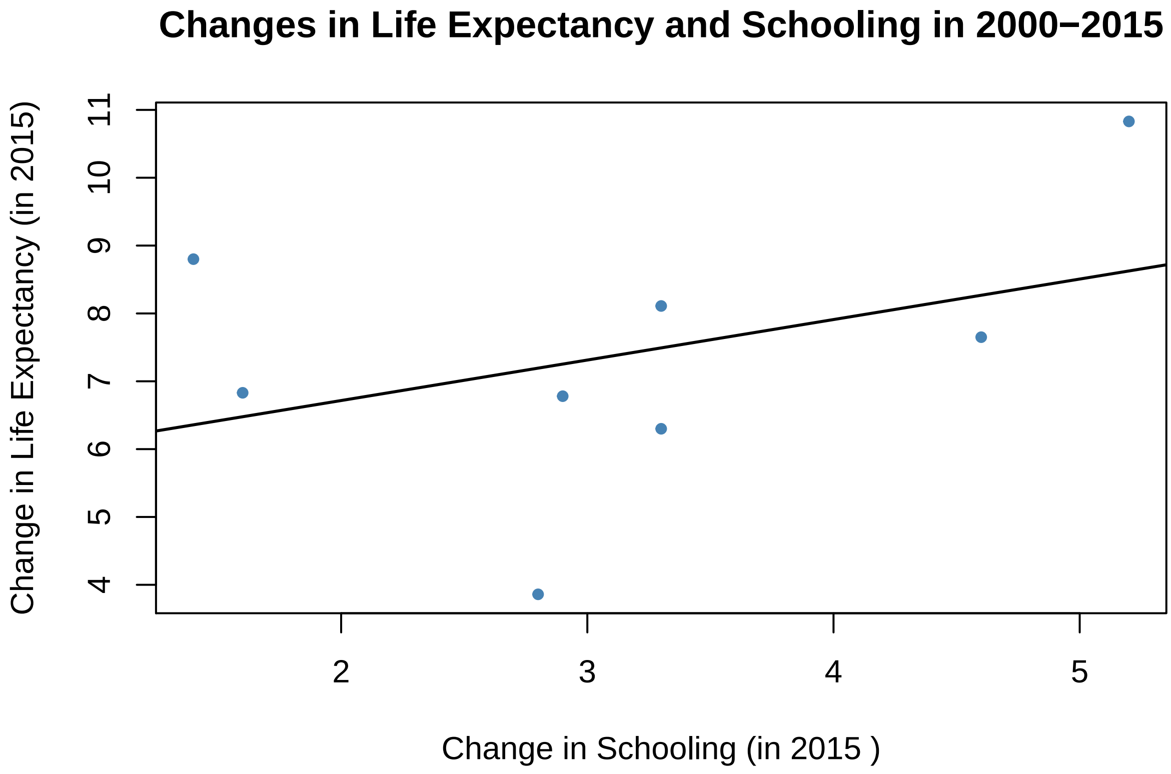
```r
     pch = 20,
     col = "steelblue")

abline(who2015_mod, lwd = 1.5)
```



```r
# compute the differences
diff_LEx <- WHO_Final_RegionSA2015$LEx - WHO_Final_RegionSA2000$LEx
diff_Schooling <- WHO_Final_RegionSA2015$Schooling - WHO_Final_RegionSA2000$Schooling

# estimate a regression using differenced data
who_diff_mod <- lm(diff_LEx ~ diff_Schooling)

coeftest(who_diff_mod, vcov = vcovHC, type = "HC1")
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.52211    2.13566  2.5857  0.04145 *
## diff_Schooling  0.59694    0.61698  0.9675  0.37065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# plot the differenced data
plot(x = diff_Schooling,
     y = diff_LEx,
     xlab = "Change in Schooling (in 2015 )",
     ylab = "Change in Life Expectancy (in 2015)",
     main = "Changes in Life Expectancy and Schooling in 2000-2015",
     pch = 20,
     col = "steelblue")

# add the regression line to plot
abline(who_diff_mod, lwd = 1.5)
```


Changes in Life Expectancy and Schooling in 2000−2015
Change in Life Expectancy (in 2015)
Change in Schooling (in 2015)