data_exploration

```
# install.packages("rmarkdown")
# install.packages("readr")
# install.packages("leaps")
# install.packages("corrplot")
library(readr)
library(leaps)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

Loading and cleaning dataset

```
# Loading and cleaning data from source

col_names <- c('country', 'year', 'status', 'life_expectancy', 'adult_mortality',
               'infant_deaths', 'alcohol', 'percent_expend', 'hep_b', 'measles',
               'bmi', 'deaths_under5', 'polio', 'total_expend', 'diptheria', 'hiv_aids',
               'gdp', 'population', 'thin_1_19', 'thin_5_9', 'income_comp', 'schooling')

data <- read_csv('data/life_expectancy_raw.csv',
                 col_names = TRUE,
                 col_types = cols(population = col_double()),
                 trim_ws = TRUE)

colnames(data) <- col_names

head(data, 10) #view first 10 rows
```

```
## # A tibble: 10 x 22
##    country     year status   life_expectancy adult_mortality infant_deaths
##    <chr>      <int> <chr>              <dbl>           <int>         <int>
##  1 Afghanist~  2015 Develop~            65.0             263            62
##  2 Afghanist~  2014 Develop~            59.9             271            64
##  3 Afghanist~  2013 Develop~            59.9             268            66
##  4 Afghanist~  2012 Develop~            59.5             272            69
##  5 Afghanist~  2011 Develop~            59.2             275            71
##  6 Afghanist~  2010 Develop~            58.8             279            74
##  7 Afghanist~  2009 Develop~            58.6             281            77
##  8 Afghanist~  2008 Develop~            58.1             287            80
##  9 Afghanist~  2007 Develop~            57.5             295            82
## 10 Afghanist~  2006 Develop~            57.3             295            84
## # ... with 16 more variables: alcohol <dbl>, percent_expend <dbl>,
## #   hep_b <int>, measles <int>, bmi <dbl>, deaths_under5 <int>,
## #   polio <int>, total_expend <dbl>, diptheria <int>, hiv_aids <dbl>,
## #   gdp <dbl>, population <int>, thin_1_19 <dbl>, thin_5_9 <dbl>,
## #   income_comp <dbl>, schooling <dbl>
```

Data exploration and scoping

```
# Picking data from one year- try most recent year

data_2015 = data[data$year == 2015, ] # Most recent year available

missing_2015 <- colSums(is.na(data_2015))
print("Columns with more than 10% missing data for 2015:")
```

```
## [1] "Columns with more than 10% missing data for 2015:"
```

```
print(missing_2015[missing_2015 > 18])
```

```
##      alcohol total_expend          gdp   population
##          177          181           29           42
```

```
# Two predictors with data mostly incomplete- try the next most recent year

data_2014 = data[data$year == 2014, ] # Next most recent year
missing_2014 <- colSums(is.na(data_2014))
print("Columns with more than 10% missing data for 2014:")
```

```
## [1] "Columns with more than 10% missing data for 2014:"
```

```
print(missing_2014[missing_2014 > 18])
```

```
##        gdp population
##         28         42
```

```
# clean out 2014 data with complete columns to use as input for analysis

input <- data_2014[complete.cases(data_2014), ]


# convert country status to numerical dummy variable
input["status_code"] <- NA
input$status_code[input$status == 'Developed'] <- 1
input$status_code[input$status == 'Developing'] <- 0

# drop unused columns for regression model
input <- input[, !(colnames(input) %in% c('country', 'year', 'status'))]

head(input, 10) #view first 10 rows
```

```
## # A tibble: 10 x 20
##    life_expectancy adult_mortality infant_deaths alcohol percent_expend
##              <dbl>           <int>         <int>   <dbl>          <dbl>
##  1            59.9             271            64  0.0100           73.5
##  2            77.5               8             0  4.51            429.
##  3            75.4              11            21  0.0100           54.2
##  4            51.7             348            67  8.33             24.0
##  5            76.2             118             8  7.93            847.
##  6            74.6              12             1  3.91            296.
##  7            82.7               6             1  9.71          10769.
##  8            81.4              66             0 12.3            8350.
##  9            72.5             119             5  0.0100          306.
## 10            71.4             132            98  0.0100           10.4
```

```
## # ... with 15 more variables: hep_b <int>, measles <int>, bmi <dbl>,
## #   deaths_under5 <int>, polio <int>, total_expend <dbl>, diptheria <int>,
## #   hiv_aids <dbl>, gdp <dbl>, population <int>, thin_1_19 <dbl>,
## #   thin_5_9 <dbl>, income_comp <dbl>, schooling <dbl>, status_code <dbl>
```

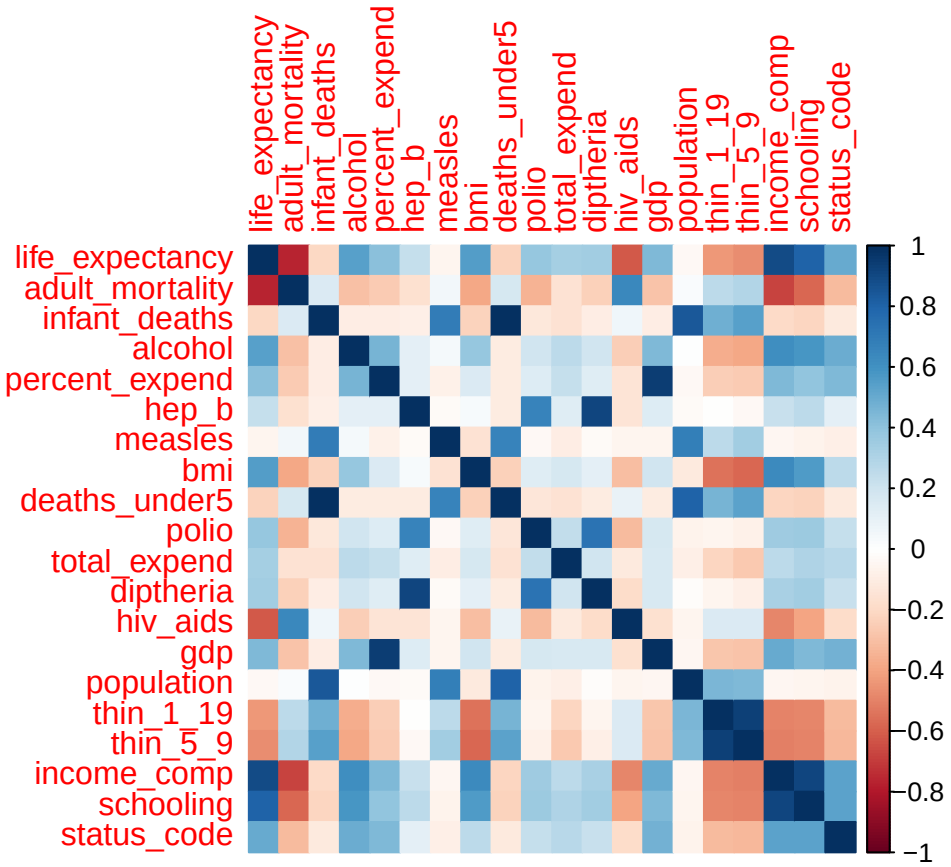
```
write.csv(input, file='data/life_expectancy_input.csv')
```

Check correlation of predictors

```
str(input)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    130 obs. of  20 variables:
##  $ life_expectancy: num  59.9 77.5 75.4 51.7 76.2 74.6 82.7 81.4 72.5 71.4 ...
##  $ adult_mortality: int  271 8 11 348 118 12 6 66 119 132 ...
##  $ infant_deaths  : int  64 0 21 67 8 1 1 0 5 98 ...
##  $ alcohol        : num  0.01 4.51 0.01 8.33 7.93 ...
##  $ percent_expend : num  73.5 428.7 54.2 24 847.4 ...
##  $ hep_b          : int  62 98 95 64 94 93 91 98 94 97 ...
##  $ measles        : int  492 0 0 11699 1 13 340 117 0 289 ...
##  $ bmi            : num  18.6 57.2 58.4 22.7 62.2 54.1 66.1 57.1 51.5 17.7 ...
##  $ deaths_under5  : int  86 1 24 101 9 1 1 0 6 121 ...
##  $ polio          : int  58 98 95 68 92 95 92 98 97 97 ...
##  $ total_expend   : num  8.18 5.88 7.21 3.31 4.79 ...
##  $ diptheria      : int  62 98 95 64 94 93 92 98 94 97 ...
##  $ hiv_aids       : num  0.1 0.1 0.1 2 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ gdp            : num  613 4576 548 479 12245 ...
##  $ population     : int  327582 288914 39113313 2692466 42981515 29622 2346694 8541575 953579 159452
##  $ thin_1_19      : num  17.5 1.2 6 8.5 1 2.1 0.6 1.8 2.8 18.1 ...
##  $ thin_5_9       : num  17.5 1.3 5.8 8.3 0.9 2.1 0.6 2 2.9 18.6 ...
##  $ income_comp    : num  0.476 0.761 0.741 0.527 0.825 0.739 0.936 0.892 0.752 0.57 ...
##  $ schooling      : num  10 14.2 14.4 11.4 17.3 12.7 20.4 15.9 12.2 10 ...
##  $ status_code    : num  0 0 0 0 0 0 1 1 0 0 ...
```

```
correlation <- cor(input)
corrplot(correlation, method = 'color')
```



Predictors with strong positive correlation with life_expectancy: - income_comp - schooling

Predictors with strong negative correlation with life_expectancy: - adult_mortality - hiv_aids

Stepwise selection using AIC

```
attach(input)

nullmodel <- lm(life_expectancy~1, data=input)
fullmodel <- lm(life_expectancy~., data=input)

step(nullmodel, data=input, scope=list(upper=fullmodel, lower=nullmodel, direction='both', k=2, test='F
```

```
## Start:  AIC=560.62
## life_expectancy ~ 1
##
##                   Df Sum of Sq    RSS    AIC
## + income_comp      1    7651.5 1901.8 352.79
## + schooling        1    6203.3 3350.1 426.40
## + adult_mortality  1    5643.0 3910.4 446.50
## + hiv_aids         1    3594.6 5958.8 501.26
## + bmi              1    2894.7 6658.7 515.70
## + alcohol          1    2806.6 6746.8 517.41
## + status_code      1    2425.8 7127.6 524.54
## + thin_5_9         1    2031.1 7522.3 531.55
## + gdp              1    1892.4 7661.0 533.93
```

```
## + thin_1_19        1    1834.3 7719.1 534.91
## + percent_expend   1    1640.9 7912.5 538.13
## + polio            1    1411.5 8141.8 541.84
## + diptheria        1    1114.0 8439.4 546.51
## + total_expend     1    1041.4 8511.9 547.62
## + hep_b            1     542.1 9011.2 555.03
## + deaths_under5    1     499.4 9053.9 555.64
## + infant_deaths    1     384.8 9168.6 557.28
## <none>                         9553.3 560.62
## + measles          1      24.4 9529.0 562.29
## + population       1      11.1 9542.3 562.47
##
## Step:  AIC=352.79
## life_expectancy ~ income_comp
##
##                   Df Sum of Sq    RSS    AIC
## + adult_mortality  1     477.6 1424.2 317.20
## + hiv_aids         1     411.7 1490.1 323.08
## + total_expend     1      89.1 1812.7 348.55
## + polio            1      46.9 1854.9 351.55
## <none>                         1901.8 352.79
## + diptheria        1      25.6 1876.2 353.03
## + hep_b            1      14.8 1887.0 353.77
## + deaths_under5    1      12.7 1889.1 353.92
## + schooling        1      11.7 1890.1 353.99
## + infant_deaths    1       7.0 1894.8 354.31
## + status_code      1       6.0 1895.8 354.38
## + bmi              1       5.1 1896.6 354.44
## + percent_expend   1       3.0 1898.8 354.59
## + measles          1       1.7 1900.1 354.67
## + alcohol          1       1.0 1900.8 354.73
## + thin_5_9         1       0.6 1901.2 354.75
## + population       1       0.6 1901.2 354.75
## + thin_1_19        1       0.4 1901.3 354.76
## + gdp              1       0.3 1901.5 354.77
## - income_comp      1    7651.5 9553.3 560.62
##
## Step:  AIC=317.2
## life_expectancy ~ income_comp + adult_mortality
##
##                   Df Sum of Sq    RSS    AIC
## + hiv_aids         1    120.75 1303.4 307.68
## + total_expend     1    102.84 1321.3 309.45
## + diptheria        1     22.36 1401.8 317.14
## <none>                         1424.2 317.20
## + status_code      1     18.90 1405.3 317.46
## + polio            1     13.77 1410.4 317.93
## + alcohol          1     13.39 1410.8 317.97
## + hep_b            1     12.39 1411.8 318.06
## + percent_expend   1     10.38 1413.8 318.24
## + thin_5_9         1      6.70 1417.5 318.58
## + deaths_under5    1      6.37 1417.8 318.61
## + infant_deaths    1      3.81 1420.4 318.85
## + thin_1_19        1      2.75 1421.4 318.94
```

```
## + gdp              1      1.48 1422.7 319.06
## + bmi              1      0.41 1423.8 319.16
## + measles          1      0.34 1423.8 319.16
## + schooling        1      0.32 1423.9 319.17
## + population       1      0.25 1423.9 319.17
## - adult_mortality  1    477.62 1901.8 352.79
## - income_comp      1   2486.19 3910.4 446.50
##
## Step:  AIC=307.68
## life_expectancy ~ income_comp + adult_mortality + hiv_aids
##
##                   Df Sum of Sq    RSS    AIC
## + total_expend     1    103.09 1200.3 298.97
## + status_code      1     26.23 1277.2 307.03
## <none>                         1303.4 307.68
## + diptheria        1     19.34 1284.1 307.73
## + percent_expend   1     15.72 1287.7 308.10
## + thin_5_9         1     13.39 1290.0 308.34
## + alcohol          1     10.52 1292.9 308.62
## + hep_b            1      9.33 1294.1 308.74
## + deaths_under5    1      8.80 1294.6 308.80
## + infant_deaths    1      6.39 1297.0 309.04
## + thin_1_19        1      5.97 1297.5 309.08
## + polio            1      5.88 1297.5 309.09
## + gdp              1      3.87 1299.5 309.29
## + measles          1      3.42 1300.0 309.34
## + bmi              1      1.15 1302.3 309.56
## + population       1      0.23 1303.2 309.65
## + schooling        1      0.16 1303.3 309.66
## - hiv_aids         1    120.75 1424.2 317.20
## - adult_mortality  1    186.67 1490.1 323.08
## - income_comp      1   2373.58 3677.0 440.50
##
## Step:  AIC=298.97
## life_expectancy ~ income_comp + adult_mortality + hiv_aids +
##     total_expend
##
##                   Df Sum of Sq    RSS    AIC
## <none>                         1200.3 298.97
## + status_code      1     12.24 1188.1 299.63
## + diptheria        1     10.17 1190.2 299.86
## + percent_expend   1      6.61 1193.7 300.25
## + hep_b            1      5.08 1195.2 300.41
## + thin_5_9         1      4.55 1195.8 300.47
## + alcohol          1      3.77 1196.6 300.56
## + deaths_under5    1      3.50 1196.8 300.59
## + gdp              1      2.48 1197.8 300.70
## + thin_1_19        1      2.17 1198.2 300.73
## + infant_deaths    1      2.01 1198.3 300.75
## + bmi              1      1.38 1199.0 300.82
## + schooling        1      1.35 1199.0 300.82
## + measles          1      0.84 1199.5 300.87
## + polio            1      0.49 1199.8 300.91
## + population       1      0.07 1200.3 300.96
```

```
## - total_expend     1    103.09 1303.4 307.68
## - hiv_aids         1    121.01 1321.3 309.45
## - adult_mortality  1    194.15 1394.5 316.46
## - income_comp      1   2054.61 3254.9 426.65
```

```
##
## Call:
## lm(formula = life_expectancy ~ income_comp + adult_mortality +
##     hiv_aids + total_expend, data = input)
##
## Coefficients:
##     (Intercept)      income_comp  adult_mortality         hiv_aids
##        47.26321         36.59524         -0.01744         -0.81982
##    total_expend
##         0.36635
```

```
model1 <- lm(life_expectancy ~ income_comp + adult_mortality + hiv_aids + total_expend)

anova(model1)
```

```
## Analysis of Variance Table
##
## Response: life_expectancy
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## income_comp       1 7651.5  7651.5 796.820 < 2.2e-16 ***
## adult_mortality   1  477.6   477.6  49.739 1.059e-10 ***
## hiv_aids          1  120.8   120.8  12.575  0.000551 ***
## total_expend      1  103.1   103.1  10.736  0.001360 **
## Residuals       125 1200.3     9.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = life_expectancy ~ income_comp + adult_mortality +
##     hiv_aids + total_expend)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5461  -1.6357  -0.0831   1.6409   9.9919
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     47.263214   2.070032  22.832  < 2e-16 ***
## income_comp     36.595239   2.501808  14.628  < 2e-16 ***
## adult_mortality -0.017439   0.003878  -4.497 1.56e-05 ***
## hiv_aids        -0.819817   0.230941  -3.550 0.000544 ***
## total_expend     0.366347   0.111809   3.277 0.001360 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.099 on 125 degrees of freedom
## Multiple R-squared:  0.8744, Adjusted R-squared:  0.8703
## F-statistic: 217.5 on 4 and 125 DF,  p-value: < 2.2e-16
```