#the package used to create a formal summary table   
library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

#reading in the data  
lifeExpect1 = read.csv("../Life\_Expectancy\_Data1.csv")  
  
#First a few quick summaries of the data  
dim(lifeExpect1) # this tells us there are 217 rows and 29 columns

## [1] 217 29

attributes(lifeExpect1)

## $names  
## [1] "Country.Name" "Country.Code" "Continent"   
## [4] "SP.DYN.LE00.IN" "EG.ELC.ACCS.ZS" "NY.ADJ.NNTY.KD.ZG"   
## [7] "NY.ADJ.NNTY.PC.KD.ZG" "SH.HIV.INCD.14" "SE.PRM.UNER"   
## [10] "SE.PRM.CUAT.ZS" "SE.TER.CUAT.BA.ZS" "SP.DYN.IMRT.IN"   
## [13] "SE.PRM.CMPT.ZS" "SE.ADT.LITR.ZS" "FR.INR.RINR"   
## [16] "SP.POP.GROW" "EN.POP.DNST" "SP.POP.TOTL"   
## [19] "SH.XPD.CHEX.PC.CD" "SH.XPD.CHEX.GD.ZS" "SL.UEM.TOTL.NE.ZS"   
## [22] "NY.GDP.MKTP.KD.ZG" "NY.GDP.PCAP.CD" "SP.DYN.CBRT.IN"   
## [25] "EG.FEC.RNEW.ZS" "SH.HIV.INCD" "SH.H2O.SMDW.ZS"   
## [28] "SI.POV.LMIC" "SE.COM.DURS"   
##   
## $class  
## [1] "data.frame"  
##   
## $row.names  
## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18  
## [19] 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36  
## [37] 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54  
## [55] 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72  
## [73] 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90  
## [91] 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108  
## [109] 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126  
## [127] 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144  
## [145] 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162  
## [163] 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180  
## [181] 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198  
## [199] 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216  
## [217] 217

#next looking at the variables more specifically  
str(lifeExpect1)

## 'data.frame': 217 obs. of 29 variables:  
## $ Country.Name : chr "Afghanistan" "Albania" "Algeria" "American Samoa" ...  
## $ Country.Code : chr "AFG" "ALB" "DZA" "ASM" ...  
## $ Continent : chr "Asia" "Europe" "Africa" "Australia/Oceania" ...  
## $ SP.DYN.LE00.IN : num 64.8 78.6 76.9 NA NA ...  
## $ EG.ELC.ACCS.ZS : num 97.7 100 99.5 NA 100 ...  
## $ NY.ADJ.NNTY.KD.ZG : num NA 0.146 2.938 NA NA ...  
## $ NY.ADJ.NNTY.PC.KD.ZG: num NA 0.574 0.966 NA NA ...  
## $ SH.HIV.INCD.14 : int 200 NA 200 NA NA 6200 NA 100 100 NA ...  
## $ SE.PRM.UNER : int NA 3359 12511 NA NA NA NA 18518 14928 NA ...  
## $ SE.PRM.CUAT.ZS : num NA NA NA NA NA NA NA NA NA NA ...  
## $ SE.TER.CUAT.BA.ZS : num NA NA NA NA NA NA NA NA NA NA ...  
## $ SP.DYN.IMRT.IN : num 46.4 8.6 20 NA 2.5 49.9 5.6 8.1 10.2 NA ...  
## $ SE.PRM.CMPT.ZS : num 84.3 103.3 101.4 NA NA ...  
## $ SE.ADT.LITR.ZS : num NA NA NA NA NA NA NA NA NA NA ...  
## $ FR.INR.RINR : num NA 4.99 8.51 NA NA ...  
## $ SP.POP.GROW : num 2.313 -0.426 1.934 -0.269 0.179 ...  
## $ EN.POP.DNST : num 58.3 104.2 18.1 276.6 164.1 ...  
## $ SP.POP.TOTL : int 38041757 2854191 43053054 55312 77146 31825299 97115 44938712 2957728 106310 ...  
## $ SH.XPD.CHEX.PC.CD : num 65.8 NA 248.2 NA 2744.2 ...  
## $ SH.XPD.CHEX.GD.ZS : num 13.24 NA 6.24 NA 6.71 ...  
## $ SL.UEM.TOTL.NE.ZS : num NA 11.5 NA NA NA ...  
## $ NY.GDP.MKTP.KD.ZG : num 3.912 2.113 1 -0.488 2.016 ...  
## $ NY.GDP.PCAP.CD : num 494 5396 3990 11715 40897 ...  
## $ SP.DYN.CBRT.IN : num 31.8 11.6 23.6 NA 7 ...  
## $ EG.FEC.RNEW.ZS : logi NA NA NA NA NA NA ...  
## $ SH.HIV.INCD : int 1300 100 1400 NA NA 16000 NA 5300 500 NA ...  
## $ SH.H2O.SMDW.ZS : num 26.8 70.7 73.1 98.4 90.6 ...  
## $ SI.POV.LMIC : num NA NA NA NA NA NA NA 4.9 9.8 NA ...  
## $ SE.COM.DURS : int 9 9 10 NA 10 6 11 14 12 13 ...

summary(lifeExpect1)

## Country.Name Country.Code Continent SP.DYN.LE00.IN   
## Length:217 Length:217 Length:217 Min. :53.28   
## Class :character Class :character Class :character 1st Qu.:67.89   
## Mode :character Mode :character Mode :character Median :74.23   
## Mean :72.93   
## 3rd Qu.:78.48   
## Max. :85.08   
## NA's :19   
## EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SH.HIV.INCD.14   
## Min. : 6.721 Min. :-30.792 Min. :-32.5432 Min. : 100   
## 1st Qu.: 84.762 1st Qu.: 1.225 1st Qu.: 0.5222 1st Qu.: 100   
## Median :100.000 Median : 3.660 Median : 2.7583 Median : 500   
## Mean : 86.470 Mean : 4.030 Mean : 2.6585 Mean : 1650   
## 3rd Qu.:100.000 3rd Qu.: 6.242 3rd Qu.: 5.0702 3rd Qu.: 1100   
## Max. :100.000 Max. : 50.172 Max. : 47.2518 Max. :20000   
## NA's :1 NA's :79 NA's :79 NA's :127   
## SE.PRM.UNER SE.PRM.CUAT.ZS SE.TER.CUAT.BA.ZS SP.DYN.IMRT.IN   
## Min. : 0 Min. : 49.55 Min. : 4.322 Min. : 1.60   
## 1st Qu.: 1262 1st Qu.: 81.77 1st Qu.:11.898 1st Qu.: 5.70   
## Median : 7359 Median : 93.69 Median :19.665 Median :14.30   
## Mean : 98650 Mean : 87.74 Mean :19.864 Mean :20.97   
## 3rd Qu.: 78956 3rd Qu.: 99.24 3rd Qu.:25.721 3rd Qu.:31.50   
## Max. :1712650 Max. :100.00 Max. :46.631 Max. :82.40   
## NA's :99 NA's :181 NA's :179 NA's :24   
## SE.PRM.CMPT.ZS SE.ADT.LITR.ZS FR.INR.RINR SP.POP.GROW   
## Min. : 54.73 Min. : 58.00 Min. :-78.518 Min. :-1.6095   
## 1st Qu.: 85.82 1st Qu.: 89.89 1st Qu.: 3.176 1st Qu.: 0.3882   
## Median : 97.40 Median : 95.74 Median : 6.354 Median : 1.0946   
## Mean : 93.05 Mean : 92.04 Mean : 6.220 Mean : 1.1917   
## 3rd Qu.:101.45 3rd Qu.: 97.56 3rd Qu.: 9.214 3rd Qu.: 1.9556   
## Max. :120.45 Max. :100.00 Max. : 39.877 Max. : 4.4687   
## NA's :89 NA's :192 NA's :104 NA's :1   
## EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS  
## Min. : 0.137 Min. :1.076e+04 Min. : 19.85 Min. : 1.525   
## 1st Qu.: 38.177 1st Qu.:7.779e+05 1st Qu.: 85.73 1st Qu.: 4.444   
## Median : 92.842 Median :6.661e+06 Median : 392.43 Median : 6.272   
## Mean : 446.043 Mean :3.545e+07 Mean : 1143.71 Mean : 6.595   
## 3rd Qu.: 233.011 3rd Qu.:2.544e+07 3rd Qu.: 1160.93 3rd Qu.: 8.202   
## Max. :19466.444 Max. :1.408e+09 Max. :10921.01 Max. :23.962   
## NA's :1 NA's :1 NA's :31 NA's :31   
## SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN   
## Min. : 0.100 Min. :-11.143 Min. : 228.2 Min. : 5.90   
## 1st Qu.: 3.810 1st Qu.: 1.183 1st Qu.: 2369.7 1st Qu.:10.62   
## Median : 5.660 Median : 2.605 Median : 7027.6 Median :17.19   
## Mean : 7.674 Mean : 2.811 Mean : 18605.5 Mean :19.37   
## 3rd Qu.: 9.960 3rd Qu.: 4.778 3rd Qu.: 23330.8 3rd Qu.:27.04   
## Max. :28.470 Max. : 19.536 Max. :189487.1 Max. :45.64   
## NA's :96 NA's :14 NA's :12 NA's :13   
## EG.FEC.RNEW.ZS SH.HIV.INCD SH.H2O.SMDW.ZS SI.POV.LMIC   
## Mode:logical Min. : 100 Min. : 5.581 Min. : 0.000   
## NA's:217 1st Qu.: 500 1st Qu.: 54.157 1st Qu.: 2.825   
## Median : 1100 Median : 88.908 Median : 6.600   
## Mean : 7574 Mean : 73.702 Mean :10.127   
## 3rd Qu.: 4900 3rd Qu.: 98.604 3rd Qu.: 9.800   
## Max. :210000 Max. :100.000 Max. :63.800   
## NA's :88 NA's :89 NA's :195   
## SE.COM.DURS   
## Min. : 0.000   
## 1st Qu.: 9.000   
## Median :10.000   
## Mean : 9.919   
## 3rd Qu.:12.000   
## Max. :17.000   
## NA's :19

#we can create a nice output table for each variable using the stargazer package  
stargazer(lifeExpect1, type = "html", out = "../lifeExpectSummary.html")

##   
## <table style="text-align:center"><tr><td colspan="6" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">Statistic</td><td>N</td><td>Mean</td><td>St. Dev.</td><td>Min</td><td>Max</td></tr>  
## <tr><td colspan="6" style="border-bottom: 1px solid black"></td></tr><tr><td style="text-align:left">SP.DYN.LE00.IN</td><td>198</td><td>72.927</td><td>7.471</td><td>53.283</td><td>85.078</td></tr>  
## <tr><td style="text-align:left">EG.ELC.ACCS.ZS</td><td>216</td><td>86.470</td><td>24.424</td><td>6.721</td><td>100.000</td></tr>  
## <tr><td style="text-align:left">NY.ADJ.NNTY.KD.ZG</td><td>138</td><td>4.030</td><td>6.713</td><td>-30.792</td><td>50.172</td></tr>  
## <tr><td style="text-align:left">NY.ADJ.NNTY.PC.KD.ZG</td><td>138</td><td>2.659</td><td>6.534</td><td>-32.543</td><td>47.252</td></tr>  
## <tr><td style="text-align:left">SH.HIV.INCD.14</td><td>90</td><td>1,650.444</td><td>3,323.894</td><td>100</td><td>20,000</td></tr>  
## <tr><td style="text-align:left">SE.PRM.UNER</td><td>118</td><td>98,649.650</td><td>258,179.500</td><td>0</td><td>1,712,650</td></tr>  
## <tr><td style="text-align:left">SE.PRM.CUAT.ZS</td><td>36</td><td>87.740</td><td>15.082</td><td>49.555</td><td>100.000</td></tr>  
## <tr><td style="text-align:left">SE.TER.CUAT.BA.ZS</td><td>38</td><td>19.864</td><td>9.996</td><td>4.322</td><td>46.631</td></tr>  
## <tr><td style="text-align:left">SP.DYN.IMRT.IN</td><td>193</td><td>20.971</td><td>19.228</td><td>1.600</td><td>82.400</td></tr>  
## <tr><td style="text-align:left">SE.PRM.CMPT.ZS</td><td>128</td><td>93.046</td><td>12.938</td><td>54.729</td><td>120.447</td></tr>  
## <tr><td style="text-align:left">SE.ADT.LITR.ZS</td><td>25</td><td>92.038</td><td>10.552</td><td>57.999</td><td>100.000</td></tr>  
## <tr><td style="text-align:left">FR.INR.RINR</td><td>113</td><td>6.220</td><td>10.584</td><td>-78.518</td><td>39.877</td></tr>  
## <tr><td style="text-align:left">SP.POP.GROW</td><td>216</td><td>1.192</td><td>1.092</td><td>-1.610</td><td>4.469</td></tr>  
## <tr><td style="text-align:left">EN.POP.DNST</td><td>216</td><td>446.043</td><td>1,996.603</td><td>0.137</td><td>19,466.440</td></tr>  
## <tr><td style="text-align:left">SP.POP.TOTL</td><td>216</td><td>35,445,760.000</td><td>138,416,498.000</td><td>10,764</td><td>1,407,745,000</td></tr>  
## <tr><td style="text-align:left">SH.XPD.CHEX.PC.CD</td><td>186</td><td>1,143.706</td><td>1,838.133</td><td>19.850</td><td>10,921.010</td></tr>  
## <tr><td style="text-align:left">SH.XPD.CHEX.GD.ZS</td><td>186</td><td>6.595</td><td>3.026</td><td>1.525</td><td>23.962</td></tr>  
## <tr><td style="text-align:left">SL.UEM.TOTL.NE.ZS</td><td>121</td><td>7.674</td><td>5.631</td><td>0.100</td><td>28.470</td></tr>  
## <tr><td style="text-align:left">NY.GDP.MKTP.KD.ZG</td><td>203</td><td>2.811</td><td>3.230</td><td>-11.143</td><td>19.536</td></tr>  
## <tr><td style="text-align:left">NY.GDP.PCAP.CD</td><td>205</td><td>18,605.510</td><td>27,774.070</td><td>228.214</td><td>189,487.100</td></tr>  
## <tr><td style="text-align:left">SP.DYN.CBRT.IN</td><td>204</td><td>19.373</td><td>9.733</td><td>5.900</td><td>45.637</td></tr>  
## <tr><td style="text-align:left">SH.HIV.INCD</td><td>129</td><td>7,574.419</td><td>22,210.870</td><td>100</td><td>210,000</td></tr>  
## <tr><td style="text-align:left">SH.H2O.SMDW.ZS</td><td>128</td><td>73.702</td><td>29.788</td><td>5.581</td><td>100.000</td></tr>  
## <tr><td style="text-align:left">SI.POV.LMIC</td><td>22</td><td>10.127</td><td>13.872</td><td>0.000</td><td>63.800</td></tr>  
## <tr><td style="text-align:left">SE.COM.DURS</td><td>198</td><td>9.919</td><td>2.589</td><td>0</td><td>17</td></tr>  
## <tr><td colspan="6" style="border-bottom: 1px solid black"></td></tr></table>

#to confirm, all of our predictor variables (besides country name and country code) seem to be continious  
#lets to a check on how many levels there are per variables  
sapply(lifeExpect1,function(x) length(unique(x)))

## Country.Name Country.Code Continent   
## 217 217 7   
## SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG   
## 199 93 139   
## NY.ADJ.NNTY.PC.KD.ZG SH.HIV.INCD.14 SE.PRM.UNER   
## 139 31 118   
## SE.PRM.CUAT.ZS SE.TER.CUAT.BA.ZS SP.DYN.IMRT.IN   
## 35 39 150   
## SE.PRM.CMPT.ZS SE.ADT.LITR.ZS FR.INR.RINR   
## 129 26 114   
## SP.POP.GROW EN.POP.DNST SP.POP.TOTL   
## 217 217 217   
## SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS   
## 187 187 117   
## NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN   
## 204 206 182   
## EG.FEC.RNEW.ZS SH.HIV.INCD SH.H2O.SMDW.ZS   
## 1 57 122   
## SI.POV.LMIC SE.COM.DURS   
## 22 15

#excluding re\_energy/EG.FEC.RNEW.ZS (which is column with all missing), we can see that continent would be worth turing into  
#a factor within R  
table(lifeExpect1$Continent)

##   
## Africa Asia Australia/Oceania   
## 1 54 50 18   
## Europe North America South America   
## 48 34 12

#we can see that one value is left blank, lets look at the country name  
lifeExpect1[lifeExpect1$Continent == "",]$Country.Name

## [1] "Solomon Islands"

#we can see that it's the Solomon Islands, therefore it's continent should be 'Australia/Oceania'  
lifeExpect1[lifeExpect1$Continent == "",]$Continent = "Australia/Oceania"  
  
#now lets look at the table again  
table(lifeExpect1$Continent)

##   
## Africa Asia Australia/Oceania Europe   
## 54 50 19 48   
## North America South America   
## 34 12

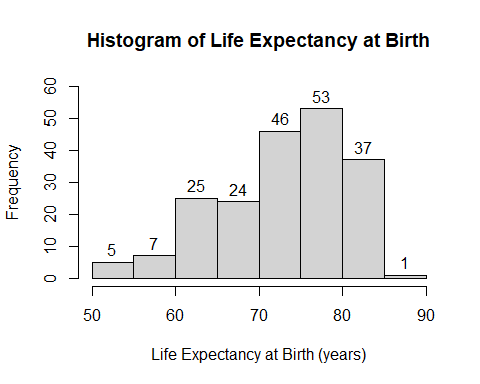
#good amount of observations per category, we shall keep all, time to convert  
lifeExpect1$Continent = as.factor(lifeExpect1$Continent)  
  
#the variable SP.POP.TOTL is in scientific notation format, we should convert this before going on so all continuous variables are  
#in the same format, this is done by changing the option of how R represents these variables  
options(scipen = 999)  
  
#The codebook/legend in the coursework tells us that SP.DYN.LE00.IN (SP.DYN.LE00.IN) is Life expectancy at birth, total (years)  
#Therefore SP.DYN.LE00.IN (SP.DYN.LE00.IN) is the variable we are interested in predicting/ the dependent variable  
summary(lifeExpect1$SP.DYN.LE00.IN)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 53.28 67.89 74.23 72.93 78.48 85.08 19

var(complete.cases(lifeExpect1$SP.DYN.LE00.IN))

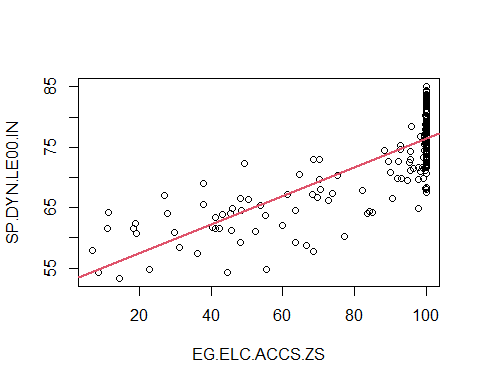
## [1] 0.08026114

#creating a histogram of life expectancy (SP.DYN.LE00.IN) with bins of size 5 each.  
hist(lifeExpect1$SP.DYN.LE00.IN,  
 xlab = "Life Expectancy at Birth (years)",  
 main = "Histogram of Life Expectancy at Birth",  
 ylim = c(0,60),  
 labels = TRUE)

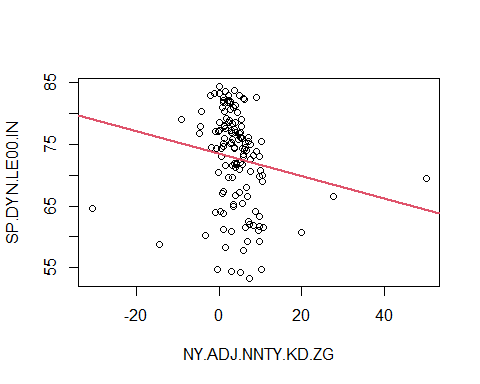


#From the above histogram we can see that the mode bin is 75 - 80.  
  
#going forward, we should remove country name and country code from our dataset of variables to analyze since they  
#provide little to no usage in predicting future countries (a factor/categorical variable which has 1 unique value per  
#observation provides little/no insight for predictions)  
lifeExpect1.cnames = subset(lifeExpect1, select = c(Country.Name,Country.Code))  
lifeExpect1 = subset(lifeExpect1, select = -c(Country.Name,Country.Code))  
  
#after removing those vars, lets investigate the relationship between the dependent variable, SP.DYN.LE00.IN, and all predictor variables  
#to do this, we will write a function which we will loop all the variables though  
relationCheck = function(data){  
   
 #removing the NA' values from the data/getting only the complete cases  
 data = na.omit(data)  
   
 #create model and summary  
 model = lm(SP.DYN.LE00.IN ~ ., data = data)  
 print(summary(model))  
   
 #we then get the relevant plots for the model to evaluate  
 plot(data[[names(data)[2]]],data[[names(data)[1]]],  
 xlab = names(data)[2],  
 ylab = names(data)[1])  
 abline(model,col=2,lwd=2)  
   
 #Get plots for each linear model, uncomment to see all (produces alot of plots)  
 #plot(model)  
   
}  
  
#Here we set a graphical setting to ask for user input before  
#each graph.   
####Uncomment this if running this section to see all graphs!####  
#par(ask = T)  
  
#We then make a temp dataframe that doesn't include Continent and   
#EG.FEC.RNEW.ZS since it has no values  
loopLifeExpect = subset(lifeExpect1,select = -c(Continent, EG.FEC.RNEW.ZS))  
  
#preforming a loop on lifeExpect so that we can view each continuous variables  
for (i in 1:ncol(loopLifeExpect)){  
 if (i != 1) {  
 relationCheck(subset(lifeExpect1, select = c(names(loopLifeExpect)[1],names(loopLifeExpect)[i])))  
 }  
}

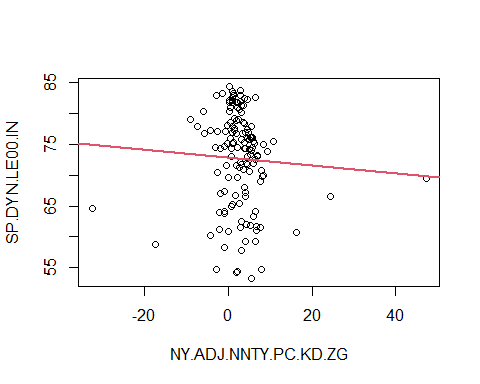
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.2018 -2.9662 0.0778 3.5547 8.8133   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 52.80646 1.14350 46.18 <0.0000000000000002 \*\*\*  
## EG.ELC.ACCS.ZS 0.23601 0.01287 18.34 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.544 on 196 degrees of freedom  
## Multiple R-squared: 0.6319, Adjusted R-squared: 0.63   
## F-statistic: 336.4 on 1 and 196 DF, p-value: < 0.00000000000000022



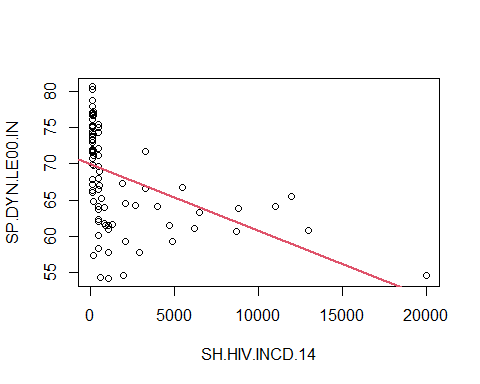
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.822 -5.126 1.861 5.180 10.966   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 73.42071 0.76254 96.285 <0.0000000000000002 \*\*\*  
## NY.ADJ.NNTY.KD.ZG -0.18150 0.09764 -1.859 0.0652 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.673 on 136 degrees of freedom  
## Multiple R-squared: 0.02478, Adjusted R-squared: 0.01761   
## F-statistic: 3.455 on 1 and 136 DF, p-value: 0.06522



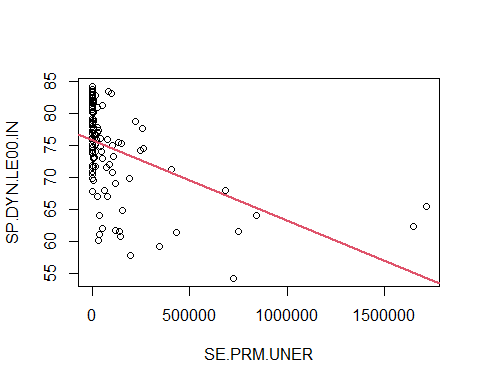
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.225 -5.266 1.703 5.456 11.520   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.8604 0.7134 102.138 <0.0000000000000002 \*\*\*  
## NY.ADJ.NNTY.PC.KD.ZG -0.0644 0.1014 -0.635 0.527   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.758 on 136 degrees of freedom  
## Multiple R-squared: 0.002954, Adjusted R-squared: -0.004377   
## F-statistic: 0.403 on 1 and 136 DF, p-value: 0.5266



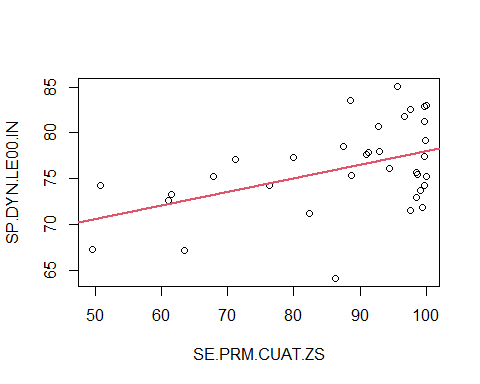
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.036 -3.925 1.664 4.659 10.859   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.9151996 0.7188299 97.263 < 0.0000000000000002 \*\*\*  
## SH.HIV.INCD.14 -0.0009143 0.0001946 -4.699 0.00000959 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.101 on 88 degrees of freedom  
## Multiple R-squared: 0.2006, Adjusted R-squared: 0.1915   
## F-statistic: 22.08 on 1 and 88 DF, p-value: 0.000009586



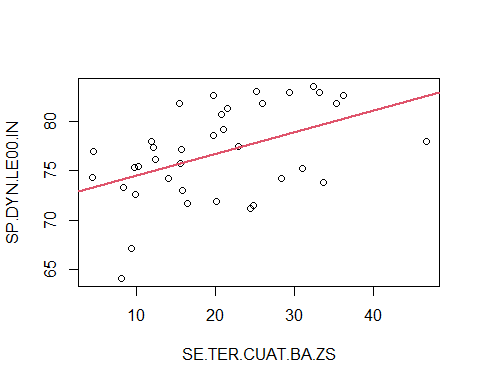
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.6380 -3.3273 0.9585 5.2488 11.0743   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 75.929200372 0.614571070 123.548 < 0.0000000000000002 \*\*\*  
## SE.PRM.UNER -0.000012581 0.000002194 -5.735 0.0000000839 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.111 on 112 degrees of freedom  
## Multiple R-squared: 0.227, Adjusted R-squared: 0.2201   
## F-statistic: 32.89 on 1 and 112 DF, p-value: 0.00000008392



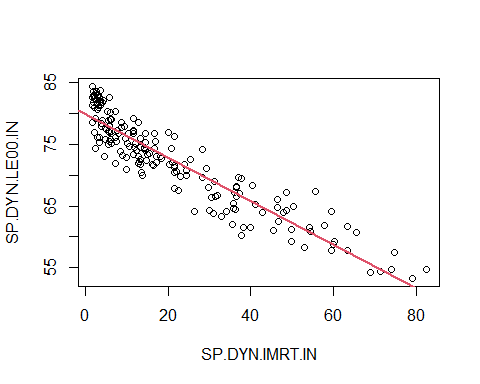
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.8666 -3.0304 0.9749 3.3324 7.6764   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 63.16087 4.34142 14.548 0.00000000000000065 \*\*\*  
## SE.PRM.CUAT.ZS 0.14878 0.04897 3.038 0.00463 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.327 on 33 degrees of freedom  
## Multiple R-squared: 0.2186, Adjusted R-squared: 0.1949   
## F-statistic: 9.231 on 1 and 33 DF, p-value: 0.004627



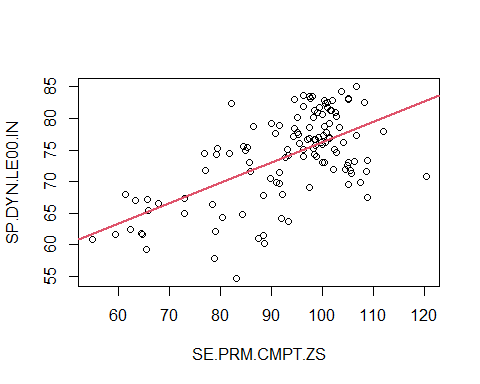
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.923 -3.828 1.056 3.360 6.132   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.30036 1.56573 46.177 < 0.0000000000000002 \*\*\*  
## SE.TER.CUAT.BA.ZS 0.21874 0.06987 3.131 0.00351 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.195 on 35 degrees of freedom  
## Multiple R-squared: 0.2188, Adjusted R-squared: 0.1965   
## F-statistic: 9.802 on 1 and 35 DF, p-value: 0.00351



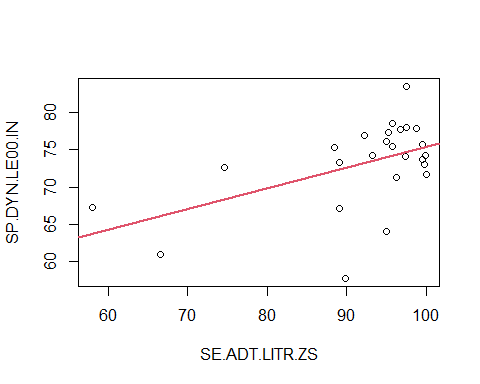
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.4624 -1.9011 -0.2795 2.2331 7.0154   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 79.83931 0.30454 262.17 <0.0000000000000002 \*\*\*  
## SP.DYN.IMRT.IN -0.35156 0.01057 -33.26 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.79 on 182 degrees of freedom  
## Multiple R-squared: 0.8588, Adjusted R-squared: 0.858   
## F-statistic: 1107 on 1 and 182 DF, p-value: < 0.00000000000000022



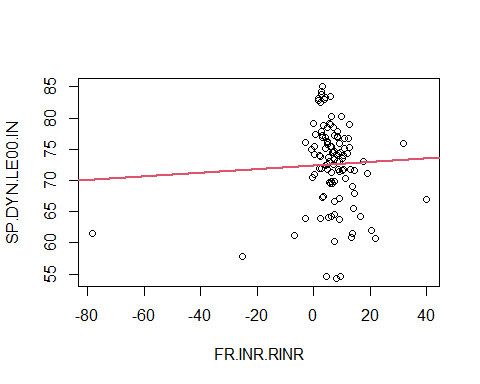
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.1265 -3.2474 0.5287 4.3403 11.9424   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 44.1031 3.6448 12.100 < 0.0000000000000002 \*\*\*  
## SE.PRM.CMPT.ZS 0.3211 0.0388 8.276 0.000000000000204 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.474 on 120 degrees of freedom  
## Multiple R-squared: 0.3634, Adjusted R-squared: 0.3581   
## F-statistic: 68.49 on 1 and 120 DF, p-value: 0.0000000000002037



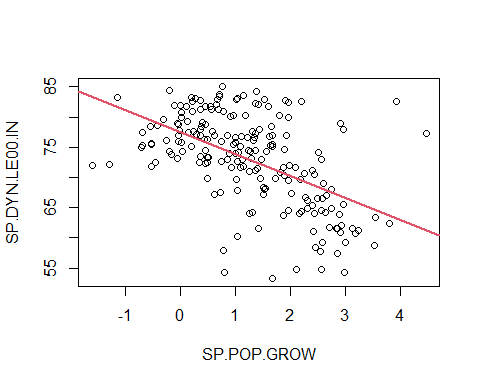
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.739 -2.257 1.004 3.283 8.876   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 47.65018 9.14325 5.212 0.0000276 \*\*\*  
## SE.ADT.LITR.ZS 0.27668 0.09872 2.803 0.0101 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.103 on 23 degrees of freedom  
## Multiple R-squared: 0.2546, Adjusted R-squared: 0.2222   
## F-statistic: 7.855 on 1 and 23 DF, p-value: 0.01011



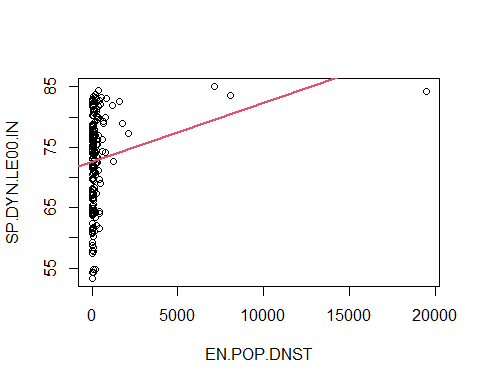
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.324 -2.987 1.202 4.287 12.565   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.42717 0.74392 97.359 <0.0000000000000002 \*\*\*  
## FR.INR.RINR 0.02868 0.06007 0.477 0.634   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.721 on 108 degrees of freedom  
## Multiple R-squared: 0.002106, Adjusted R-squared: -0.007133   
## F-statistic: 0.228 on 1 and 108 DF, p-value: 0.634



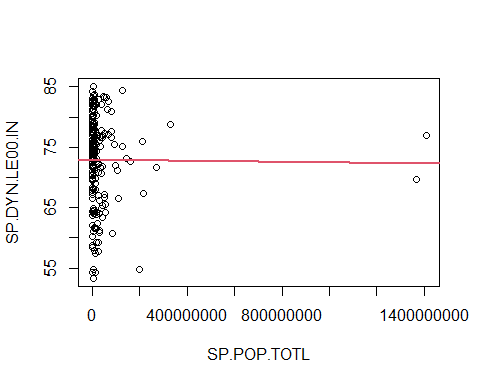
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.3005 -3.7426 -0.0117 4.0213 19.3305   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 77.5354 0.6848 113.218 < 0.0000000000000002 \*\*\*  
## SP.POP.GROW -3.6287 0.4090 -8.872 0.000000000000000461 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.325 on 195 degrees of freedom  
## Multiple R-squared: 0.2876, Adjusted R-squared: 0.2839   
## F-statistic: 78.71 on 1 and 195 DF, p-value: 0.0000000000000004611



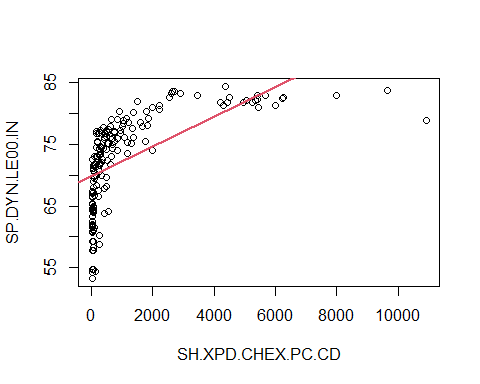
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.313 -5.192 1.291 5.256 11.430   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.5888395 0.5361179 135.397 < 0.0000000000000002 \*\*\*  
## EN.POP.DNST 0.0009740 0.0003311 2.942 0.00366 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.348 on 195 degrees of freedom  
## Multiple R-squared: 0.04249, Adjusted R-squared: 0.03758   
## F-statistic: 8.653 on 1 and 195 DF, p-value: 0.003662



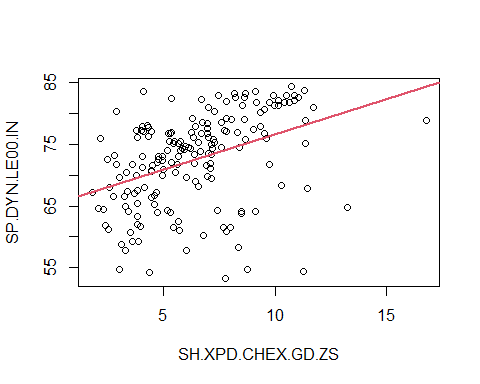
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.692 -5.051 1.259 5.521 12.104   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.9775472143399 0.5529563405929 131.977 <0.0000000000000002 \*\*\*  
## SP.POP.TOTL -0.0000000004403 0.0000000037039 -0.119 0.905   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.494 on 195 degrees of freedom  
## Multiple R-squared: 7.247e-05, Adjusted R-squared: -0.005055   
## F-statistic: 0.01413 on 1 and 195 DF, p-value: 0.9055



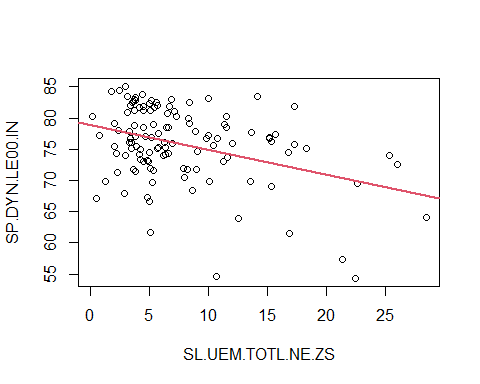
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.351 -3.210 1.485 4.494 8.569   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.7426973 0.5190206 134.37 <0.0000000000000002 \*\*\*  
## SH.XPD.CHEX.PC.CD 0.0024170 0.0002392 10.11 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.923 on 175 degrees of freedom  
## Multiple R-squared: 0.3685, Adjusted R-squared: 0.3649   
## F-statistic: 102.1 on 1 and 175 DF, p-value: < 0.00000000000000022



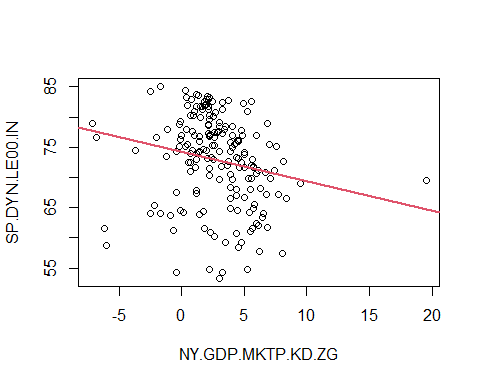
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -23.658 -3.407 1.620 4.716 13.732   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65.097 1.377 47.289 < 0.0000000000000002 \*\*\*  
## SH.XPD.CHEX.GD.ZS 1.144 0.199 5.749 0.0000000391 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.836 on 175 degrees of freedom  
## Multiple R-squared: 0.1589, Adjusted R-squared: 0.1541   
## F-statistic: 33.06 on 1 and 175 DF, p-value: 0.0000000391



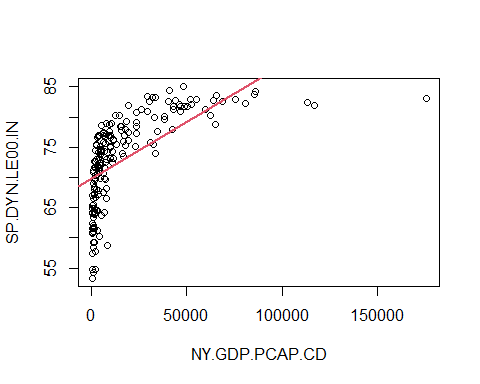
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.9228 -3.5769 0.9043 4.3836 10.2194   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 78.83915 0.87308 90.300 < 0.0000000000000002 \*\*\*  
## SL.UEM.TOTL.NE.ZS -0.39526 0.09141 -4.324 0.0000324 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.635 on 117 degrees of freedom  
## Multiple R-squared: 0.1378, Adjusted R-squared: 0.1304   
## F-statistic: 18.7 on 1 and 117 DF, p-value: 0.00003241



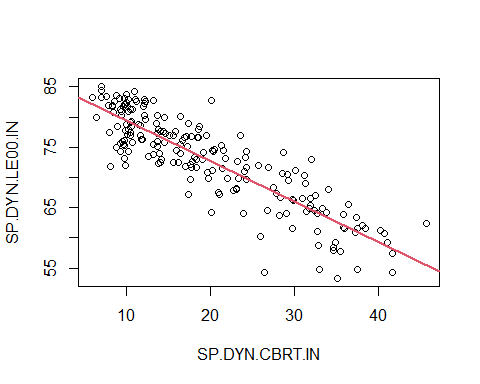
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.108 -3.732 1.134 5.252 11.063   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 74.2516 0.7365 100.813 <0.0000000000000002 \*\*\*  
## NY.GDP.MKTP.KD.ZG -0.4901 0.1767 -2.774 0.0061 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.307 on 185 degrees of freedom  
## Multiple R-squared: 0.03993, Adjusted R-squared: 0.03474   
## F-statistic: 7.695 on 1 and 185 DF, p-value: 0.006105



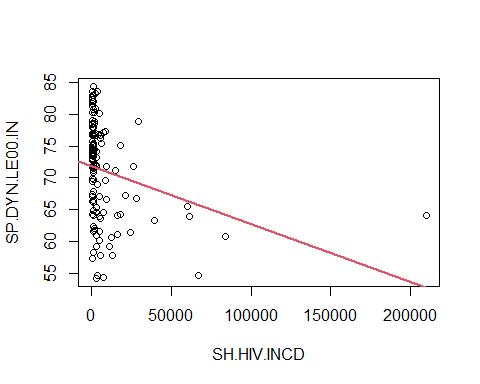
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -19.860 -3.050 1.344 4.213 8.530   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.80790847 0.51615970 135.25 <0.0000000000000002 \*\*\*  
## NY.GDP.PCAP.CD 0.00018823 0.00001737 10.84 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.87 on 187 degrees of freedom  
## Multiple R-squared: 0.3858, Adjusted R-squared: 0.3825   
## F-statistic: 117.5 on 1 and 187 DF, p-value: < 0.00000000000000022



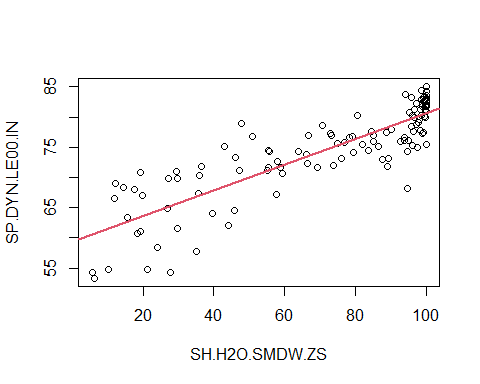
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.0568 -2.2517 0.4703 2.3831 10.1750   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 86.1321 0.5831 147.71 <0.0000000000000002 \*\*\*  
## SP.DYN.CBRT.IN -0.6717 0.0266 -25.25 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.631 on 196 degrees of freedom  
## Multiple R-squared: 0.7649, Adjusted R-squared: 0.7637   
## F-statistic: 637.7 on 1 and 196 DF, p-value: < 0.00000000000000022



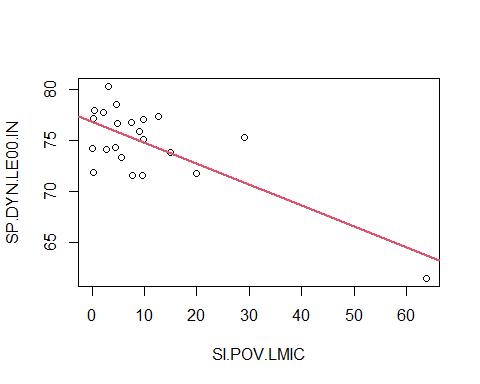
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.341 -5.485 0.882 5.292 12.623   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 71.82405378 0.68434179 104.953 < 0.0000000000000002 \*\*\*  
## SH.HIV.INCD -0.00009051 0.00002926 -3.093 0.00244 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.354 on 127 degrees of freedom  
## Multiple R-squared: 0.07005, Adjusted R-squared: 0.06273   
## F-statistic: 9.567 on 1 and 127 DF, p-value: 0.002436



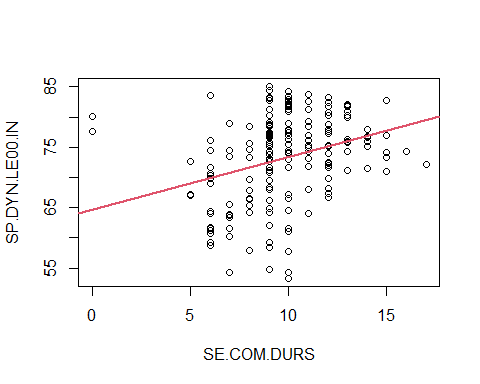
##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.2567 -2.4605 0.4529 2.4906 9.5086   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 59.27003 0.92651 63.97 <0.0000000000000002 \*\*\*  
## SH.H2O.SMDW.ZS 0.21282 0.01187 17.93 <0.0000000000000002 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.896 on 117 degrees of freedom  
## Multiple R-squared: 0.7331, Adjusted R-squared: 0.7308   
## F-statistic: 321.3 on 1 and 117 DF, p-value: < 0.00000000000000022



##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.8959 -2.2073 0.3867 1.4729 4.4338   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 76.76408 0.70150 109.43 < 0.0000000000000002 \*\*\*  
## SI.POV.LMIC -0.20441 0.04146 -4.93 0.0000807 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.636 on 20 degrees of freedom  
## Multiple R-squared: 0.5486, Adjusted R-squared: 0.526   
## F-statistic: 24.3 on 1 and 20 DF, p-value: 0.00008074



##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -20.0991 -5.2083 0.6082 4.9184 15.4715   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.6025 2.0518 31.486 < 0.0000000000000002 \*\*\*  
## SE.COM.DURS 0.8780 0.2031 4.322 0.0000255 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.097 on 180 degrees of freedom  
## Multiple R-squared: 0.09403, Adjusted R-squared: 0.089   
## F-statistic: 18.68 on 1 and 180 DF, p-value: 0.0000255



#remove the variable to avoid confusion in later parts  
rm(loopLifeExpect)  
  
#From these graphs we can see that the following variables have  
#a good linear relationship with Life Expectancy  
#We should look to keep these in a linear model  
#EG.ELC.ACCS.ZS  
#SP.DYN.IMRT.IN  
#SP.DYN.CBRT.IN  
#SE.PRM.CMPT.ZS  
#SP.POP.GROW  
  
#before modelling , we need to scale the continuous variables (excluding the dependant variable), can achieve this using a for loop  
for (i in names(lifeExpect1)) {  
 if (is.numeric(lifeExpect1[[i]]) & i != "SP.DYN.LE00.IN") {  
 lifeExpect1[[i]] = as.vector(scale(lifeExpect1[[i]]))  
 }  
}  
  
#now lets do a boxplot to check for outliers (excluding the dependent variable)  
boxplot(subset(lifeExpect1, select = -c(Continent, SP.DYN.LE00.IN)))  
  
#Using these boxplots and the previous linear graphs, we have  
#Some outliers we should remove to get better fits  
#SH.HIV.INCD.14, remove row (20000+)  
#SE.PRM.UNER, remove rows (1500000+)  
#EN.POP.DNST, remove rows (5000+)  
#SP.POP.TOTL, remove rows (1200000000+)  
#SH.XPD.CHEX.PC.CD, remove rows (~8000+)  
#NY.GDP.MKTP.KD.ZG, remove row (19+)  
#NY.GDP.PCAP.CD, remove rows (100000+)  
#SH.HIV.INCD, remove row (150000+)  
#SI.POV.LMIC, remove row (60+)  
  
summary(lifeExpect1)

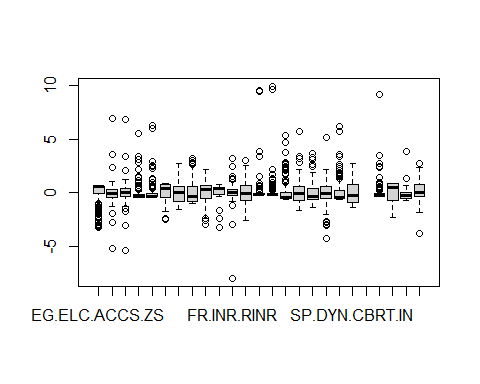
## Continent SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG   
## Africa :54 Min. :53.28 Min. :-3.26523 Min. :-5.18696   
## Asia :50 1st Qu.:67.89 1st Qu.:-0.06994 1st Qu.:-0.41783   
## Australia/Oceania:19 Median :74.23 Median : 0.55395 Median :-0.05516   
## Europe :48 Mean :72.93 Mean : 0.00000 Mean : 0.00000   
## North America :34 3rd Qu.:78.48 3rd Qu.: 0.55395 3rd Qu.: 0.32944   
## South America :12 Max. :85.08 Max. : 0.55395 Max. : 6.87298   
## NA's :19 NA's :1 NA's :79   
## NY.ADJ.NNTY.PC.KD.ZG SH.HIV.INCD.14 SE.PRM.UNER SE.PRM.CUAT.ZS   
## Min. :-5.38773 Min. :-0.4664 Min. :-0.38210 Min. :-2.5318   
## 1st Qu.:-0.32697 1st Qu.:-0.4664 1st Qu.:-0.37721 1st Qu.:-0.3955   
## Median : 0.01527 Median :-0.3461 Median :-0.35359 Median : 0.3942   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.: 0.36912 3rd Qu.:-0.1656 3rd Qu.:-0.07628 3rd Qu.: 0.7628   
## Max. : 6.82514 Max. : 5.5205 Max. : 6.25147 Max. : 0.8129   
## NA's :79 NA's :127 NA's :99 NA's :181   
## SE.TER.CUAT.BA.ZS SP.DYN.IMRT.IN SE.PRM.CMPT.ZS SE.ADT.LITR.ZS   
## Min. :-1.55488 Min. :-1.0074 Min. :-2.9617 Min. :-3.2259   
## 1st Qu.:-0.79695 1st Qu.:-0.7942 1st Qu.:-0.5584 1st Qu.:-0.2032   
## Median :-0.01989 Median :-0.3469 Median : 0.3368 Median : 0.3505   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.58598 3rd Qu.: 0.5476 3rd Qu.: 0.6498 3rd Qu.: 0.5231   
## Max. : 2.67793 Max. : 3.1947 Max. : 2.1179 Max. : 0.7546   
## NA's :179 NA's :24 NA's :89 NA's :192   
## FR.INR.RINR SP.POP.GROW EN.POP.DNST SP.POP.TOTL   
## Min. :-8.00618 Min. :-2.56617 Min. :-0.2233 Min. :-0.25600   
## 1st Qu.:-0.28760 1st Qu.:-0.73604 1st Qu.:-0.2043 1st Qu.:-0.25046   
## Median : 0.01275 Median :-0.08888 Median :-0.1769 Median :-0.20795   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.: 0.28295 3rd Qu.: 0.69979 3rd Qu.:-0.1067 3rd Qu.:-0.07228   
## Max. : 3.18004 Max. : 3.00209 Max. : 9.5264 Max. : 9.91427   
## NA's :104 NA's :1 NA's :1 NA's :1   
## SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG  
## Min. :-0.61141 Min. :-1.6754 Min. :-1.3450 Min. :-4.3196   
## 1st Qu.:-0.57557 1st Qu.:-0.7110 1st Qu.:-0.6862 1st Qu.:-0.5041   
## Median :-0.40872 Median :-0.1069 Median :-0.3576 Median :-0.0638   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.00937 3rd Qu.: 0.5309 3rd Qu.: 0.4060 3rd Qu.: 0.6089   
## Max. : 5.31915 Max. : 5.7387 Max. : 3.6932 Max. : 5.1771   
## NA's :31 NA's :31 NA's :96 NA's :14   
## NY.GDP.PCAP.CD SP.DYN.CBRT.IN EG.FEC.RNEW.ZS SH.HIV.INCD   
## Min. :-0.6617 Min. :-1.3843 Mode:logical Min. :-0.3365   
## 1st Qu.:-0.5846 1st Qu.:-0.8998 NA's:217 1st Qu.:-0.3185   
## Median :-0.4169 Median :-0.2240 Median :-0.2915   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.1701 3rd Qu.: 0.7875 3rd Qu.:-0.1204   
## Max. : 6.1526 Max. : 2.6985 Max. : 9.1138   
## NA's :12 NA's :13 NA's :88   
## SH.H2O.SMDW.ZS SI.POV.LMIC SE.COM.DURS   
## Min. :-2.2868 Min. :-0.73005 Min. :-3.83095   
## 1st Qu.:-0.6561 1st Qu.:-0.52640 1st Qu.:-0.35501   
## Median : 0.5105 Median :-0.25427 Median : 0.03121   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.: 0.8359 3rd Qu.:-0.02359 3rd Qu.: 0.80364   
## Max. : 0.8828 Max. : 3.86912 Max. : 2.73472   
## NA's :89 NA's :195 NA's :19

#From having a look at the summary, we can see that the some of the variables have a high amount of missing data  
#SH.HIV.INCD.14; c\_hiv ( Children (ages 0-14) newly infected with HIV): NA = 127  
#SE.PRM.CUAT.ZS; pri\_edu (Educational attainment, at least completed primary): NA = 181  
#SE.TER.CUAT.BA.ZS; bcs\_edu (Educational attainment, at least Bachelor's or equivalent): NA = 179  
#SE.ADT.LITR.ZS; lit\_rate (Literacy rate, adult total): NA = 192  
#EG.FEC.RNEW.ZS; re\_energy (Renewable energy consumption): NA = 217 (All is missing!)  
#SI.POV.LMIC; pov\_head ( Poverty headcount ratio at \$3.20 a day): NA = 195  
  
#in part 2, we will tackle these missing values accordingly  
  
# library imports  
library(mice)

##   
## Attaching package: 'mice'

## The following object is masked from 'package:stats':  
##   
## filter

## The following objects are masked from 'package:base':  
##   
## cbind, rbind



# summarising the analysis  
attributes(lifeExpect1)

## $names  
## [1] "Continent" "SP.DYN.LE00.IN" "EG.ELC.ACCS.ZS"   
## [4] "NY.ADJ.NNTY.KD.ZG" "NY.ADJ.NNTY.PC.KD.ZG" "SH.HIV.INCD.14"   
## [7] "SE.PRM.UNER" "SE.PRM.CUAT.ZS" "SE.TER.CUAT.BA.ZS"   
## [10] "SP.DYN.IMRT.IN" "SE.PRM.CMPT.ZS" "SE.ADT.LITR.ZS"   
## [13] "FR.INR.RINR" "SP.POP.GROW" "EN.POP.DNST"   
## [16] "SP.POP.TOTL" "SH.XPD.CHEX.PC.CD" "SH.XPD.CHEX.GD.ZS"   
## [19] "SL.UEM.TOTL.NE.ZS" "NY.GDP.MKTP.KD.ZG" "NY.GDP.PCAP.CD"   
## [22] "SP.DYN.CBRT.IN" "EG.FEC.RNEW.ZS" "SH.HIV.INCD"   
## [25] "SH.H2O.SMDW.ZS" "SI.POV.LMIC" "SE.COM.DURS"   
##   
## $row.names  
## [1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18  
## [19] 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36  
## [37] 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54  
## [55] 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72  
## [73] 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90  
## [91] 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108  
## [109] 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126  
## [127] 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144  
## [145] 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162  
## [163] 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180  
## [181] 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198  
## [199] 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216  
## [217] 217  
##   
## $class  
## [1] "data.frame"

str(lifeExpect1)

## 'data.frame': 217 obs. of 27 variables:  
## $ Continent : Factor w/ 6 levels "Africa","Asia",..: 2 4 1 3 4 1 5 6 2 5 ...  
## $ SP.DYN.LE00.IN : num 64.8 78.6 76.9 NA NA ...  
## $ EG.ELC.ACCS.ZS : num 0.46 0.554 0.533 NA 0.554 ...  
## $ NY.ADJ.NNTY.KD.ZG : num NA -0.579 -0.163 NA NA ...  
## $ NY.ADJ.NNTY.PC.KD.ZG: num NA -0.319 -0.259 NA NA ...  
## $ SH.HIV.INCD.14 : num -0.436 NA -0.436 NA NA ...  
## $ SE.PRM.UNER : num NA -0.369 -0.334 NA NA ...  
## $ SE.PRM.CUAT.ZS : num NA NA NA NA NA NA NA NA NA NA ...  
## $ SE.TER.CUAT.BA.ZS : num NA NA NA NA NA NA NA NA NA NA ...  
## $ SP.DYN.IMRT.IN : num 1.3225 -0.6434 -0.0505 NA -0.9606 ...  
## $ SE.PRM.CMPT.ZS : num -0.674 0.794 0.643 NA NA ...  
## $ SE.ADT.LITR.ZS : num NA NA NA NA NA NA NA NA NA NA ...  
## $ FR.INR.RINR : num NA -0.116 0.216 NA NA ...  
## $ SP.POP.GROW : num 1.027 -1.482 0.68 -1.338 -0.928 ...  
## $ EN.POP.DNST : num -0.1942 -0.1712 -0.2143 -0.0849 -0.1412 ...  
## $ SP.POP.TOTL : num 0.0188 -0.2355 0.055 -0.2557 -0.2555 ...  
## $ SH.XPD.CHEX.PC.CD : num -0.586 NA -0.487 NA 0.871 ...  
## $ SH.XPD.CHEX.GD.ZS : num 2.1965 NA -0.1163 NA 0.0384 ...  
## $ SL.UEM.TOTL.NE.ZS : num NA 0.674 NA NA NA ...  
## $ NY.GDP.MKTP.KD.ZG : num 0.341 -0.216 -0.561 -1.021 -0.246 ...  
## $ NY.GDP.PCAP.CD : num -0.652 -0.476 -0.526 -0.248 0.803 ...  
## $ SP.DYN.CBRT.IN : num 1.277 -0.797 0.433 NA -1.271 ...  
## $ EG.FEC.RNEW.ZS : logi NA NA NA NA NA NA ...  
## $ SH.HIV.INCD : num -0.282 -0.337 -0.278 NA NA ...  
## $ SH.H2O.SMDW.ZS : num -1.5745 -0.1019 -0.0189 0.8278 0.5686 ...  
## $ SI.POV.LMIC : num NA NA NA NA NA ...  
## $ SE.COM.DURS : num -0.355 -0.355 0.0312 NA 0.0312 ...

head(lifeExpect1)

## Continent SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG  
## 1 Asia 64.833 0.4597853 NA  
## 2 Europe 78.573 0.5539554 -0.5785355  
## 3 Africa 76.880 0.5334836 -0.1627032  
## 4 Australia/Oceania NA NA NA  
## 5 Europe NA 0.5539554 NA  
## 6 Africa 61.147 -1.6704932 -0.4331419  
## NY.ADJ.NNTY.PC.KD.ZG SH.HIV.INCD.14 SE.PRM.UNER SE.PRM.CUAT.ZS  
## 1 NA -0.4363691 NA NA  
## 2 -0.3190389 NA -0.3690868 NA  
## 3 -0.2589729 -0.4363691 -0.3336386 NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 -0.7289401 1.3687428 NA NA  
## SE.TER.CUAT.BA.ZS SP.DYN.IMRT.IN SE.PRM.CMPT.ZS SE.ADT.LITR.ZS FR.INR.RINR  
## 1 NA 1.32246862 -0.6736735 NA NA  
## 2 NA -0.64336894 0.7942993 NA -0.115807  
## 3 NA -0.05049729 0.6429495 NA 0.216203  
## 4 NA NA NA NA NA  
## 5 NA -0.96060727 NA NA NA  
## 6 NA 1.50449061 NA NA -1.236396  
## SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS  
## 1 1.0273214 -0.19421671 0.01875497 -0.5864102 2.19645792  
## 2 -1.4819621 -0.17122858 -0.23546015 NA NA  
## 3 0.6800355 -0.21434746 0.05495945 -0.4871792 -0.11627315  
## 4 -1.3381442 -0.08488569 -0.25568085 NA NA  
## 5 -0.9276740 -0.14119112 -0.25552311 0.8707142 0.03844599  
## 6 1.8791524 -0.21061545 -0.02615628 -0.5834072 -1.34222918  
## SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN  
## 1 NA 0.3405489 -0.6520948 1.2770042  
## 2 0.6741893 -0.2160889 -0.4756180 -0.7966202  
## 3 NA -0.5607544 -0.5262404 0.4325329  
## 4 NA -1.0213127 -0.2480784 NA  
## 5 NA -0.2463859 0.8026128 -1.2713078  
## 6 NA -1.0636722 -0.5687276 2.1431549  
## EG.FEC.RNEW.ZS SH.HIV.INCD SH.H2O.SMDW.ZS SI.POV.LMIC SE.COM.DURS  
## 1 NA -0.2824932 -1.57453741 NA -0.35500630  
## 2 NA -0.3365208 -0.10192266 NA -0.35500630  
## 3 NA -0.2779909 -0.01887393 NA 0.03120934  
## 4 NA NA 0.82783794 NA NA  
## 5 NA NA 0.56861090 NA 0.03120934  
## 6 NA 0.3793450 NA NA -1.51365323

summary(lifeExpect1)

## Continent SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG   
## Africa :54 Min. :53.28 Min. :-3.26523 Min. :-5.18696   
## Asia :50 1st Qu.:67.89 1st Qu.:-0.06994 1st Qu.:-0.41783   
## Australia/Oceania:19 Median :74.23 Median : 0.55395 Median :-0.05516   
## Europe :48 Mean :72.93 Mean : 0.00000 Mean : 0.00000   
## North America :34 3rd Qu.:78.48 3rd Qu.: 0.55395 3rd Qu.: 0.32944   
## South America :12 Max. :85.08 Max. : 0.55395 Max. : 6.87298   
## NA's :19 NA's :1 NA's :79   
## NY.ADJ.NNTY.PC.KD.ZG SH.HIV.INCD.14 SE.PRM.UNER SE.PRM.CUAT.ZS   
## Min. :-5.38773 Min. :-0.4664 Min. :-0.38210 Min. :-2.5318   
## 1st Qu.:-0.32697 1st Qu.:-0.4664 1st Qu.:-0.37721 1st Qu.:-0.3955   
## Median : 0.01527 Median :-0.3461 Median :-0.35359 Median : 0.3942   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.: 0.36912 3rd Qu.:-0.1656 3rd Qu.:-0.07628 3rd Qu.: 0.7628   
## Max. : 6.82514 Max. : 5.5205 Max. : 6.25147 Max. : 0.8129   
## NA's :79 NA's :127 NA's :99 NA's :181   
## SE.TER.CUAT.BA.ZS SP.DYN.IMRT.IN SE.PRM.CMPT.ZS SE.ADT.LITR.ZS   
## Min. :-1.55488 Min. :-1.0074 Min. :-2.9617 Min. :-3.2259   
## 1st Qu.:-0.79695 1st Qu.:-0.7942 1st Qu.:-0.5584 1st Qu.:-0.2032   
## Median :-0.01989 Median :-0.3469 Median : 0.3368 Median : 0.3505   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.58598 3rd Qu.: 0.5476 3rd Qu.: 0.6498 3rd Qu.: 0.5231   
## Max. : 2.67793 Max. : 3.1947 Max. : 2.1179 Max. : 0.7546   
## NA's :179 NA's :24 NA's :89 NA's :192   
## FR.INR.RINR SP.POP.GROW EN.POP.DNST SP.POP.TOTL   
## Min. :-8.00618 Min. :-2.56617 Min. :-0.2233 Min. :-0.25600   
## 1st Qu.:-0.28760 1st Qu.:-0.73604 1st Qu.:-0.2043 1st Qu.:-0.25046   
## Median : 0.01275 Median :-0.08888 Median :-0.1769 Median :-0.20795   
## Mean : 0.00000 Mean : 0.00000 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.: 0.28295 3rd Qu.: 0.69979 3rd Qu.:-0.1067 3rd Qu.:-0.07228   
## Max. : 3.18004 Max. : 3.00209 Max. : 9.5264 Max. : 9.91427   
## NA's :104 NA's :1 NA's :1 NA's :1   
## SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG  
## Min. :-0.61141 Min. :-1.6754 Min. :-1.3450 Min. :-4.3196   
## 1st Qu.:-0.57557 1st Qu.:-0.7110 1st Qu.:-0.6862 1st Qu.:-0.5041   
## Median :-0.40872 Median :-0.1069 Median :-0.3576 Median :-0.0638   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.00937 3rd Qu.: 0.5309 3rd Qu.: 0.4060 3rd Qu.: 0.6089   
## Max. : 5.31915 Max. : 5.7387 Max. : 3.6932 Max. : 5.1771   
## NA's :31 NA's :31 NA's :96 NA's :14   
## NY.GDP.PCAP.CD SP.DYN.CBRT.IN EG.FEC.RNEW.ZS SH.HIV.INCD   
## Min. :-0.6617 Min. :-1.3843 Mode:logical Min. :-0.3365   
## 1st Qu.:-0.5846 1st Qu.:-0.8998 NA's:217 1st Qu.:-0.3185   
## Median :-0.4169 Median :-0.2240 Median :-0.2915   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.1701 3rd Qu.: 0.7875 3rd Qu.:-0.1204   
## Max. : 6.1526 Max. : 2.6985 Max. : 9.1138   
## NA's :12 NA's :13 NA's :88   
## SH.H2O.SMDW.ZS SI.POV.LMIC SE.COM.DURS   
## Min. :-2.2868 Min. :-0.73005 Min. :-3.83095   
## 1st Qu.:-0.6561 1st Qu.:-0.52640 1st Qu.:-0.35501   
## Median : 0.5105 Median :-0.25427 Median : 0.03121   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.: 0.8359 3rd Qu.:-0.02359 3rd Qu.: 0.80364   
## Max. : 0.8828 Max. : 3.86912 Max. : 2.73472   
## NA's :89 NA's :195 NA's :19

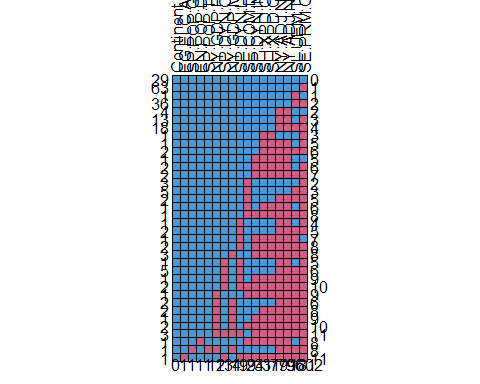
# removing these as all/most of the values are null(cut-ff threshold - >50%, ideally should be >30%)  
lifeExpect1$SH.HIV.INCD.14 <- NULL  
lifeExpect1$SE.PRM.UNER <- NULL # might be important  
# lifeExpect1$SE.PRM.CUAT.ZS <- NULL   
lifeExpect1$SE.TER.CUAT.BA.ZS <- NULL  
lifeExpect1$SE.PRM.CMPT.ZS <- NULL # since we removed $SE.PRM.UNER  
lifeExpect1$SE.ADT.LITR.ZS <- NULL  
lifeExpect1$FR.INR.RINR <- NULL  
# lifeExpect1$SL.UEM.TOTL.NE.ZS <- NULL might be imp  
lifeExpect1$EG.FEC.RNEW.ZS <- NULL  
lifeExpect1$SH.H2O.SMDW.ZS <- NULL  
lifeExpect1$SH.HIV.INCD <- NULL  
lifeExpect1$SI.POV.LMIC <- NULL  
summary(lifeExpect1)

## Continent SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG   
## Africa :54 Min. :53.28 Min. :-3.26523 Min. :-5.18696   
## Asia :50 1st Qu.:67.89 1st Qu.:-0.06994 1st Qu.:-0.41783   
## Australia/Oceania:19 Median :74.23 Median : 0.55395 Median :-0.05516   
## Europe :48 Mean :72.93 Mean : 0.00000 Mean : 0.00000   
## North America :34 3rd Qu.:78.48 3rd Qu.: 0.55395 3rd Qu.: 0.32944   
## South America :12 Max. :85.08 Max. : 0.55395 Max. : 6.87298   
## NA's :19 NA's :1 NA's :79   
## NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW   
## Min. :-5.38773 Min. :-2.5318 Min. :-1.0074 Min. :-2.56617   
## 1st Qu.:-0.32697 1st Qu.:-0.3955 1st Qu.:-0.7942 1st Qu.:-0.73604   
## Median : 0.01527 Median : 0.3942 Median :-0.3469 Median :-0.08888   
## Mean : 0.00000 Mean : 0.0000 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.: 0.36912 3rd Qu.: 0.7628 3rd Qu.: 0.5476 3rd Qu.: 0.69979   
## Max. : 6.82514 Max. : 0.8129 Max. : 3.1947 Max. : 3.00209   
## NA's :79 NA's :181 NA's :24 NA's :1   
## EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS  
## Min. :-0.2233 Min. :-0.25600 Min. :-0.61141 Min. :-1.6754   
## 1st Qu.:-0.2043 1st Qu.:-0.25046 1st Qu.:-0.57557 1st Qu.:-0.7110   
## Median :-0.1769 Median :-0.20795 Median :-0.40872 Median :-0.1069   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.:-0.1067 3rd Qu.:-0.07228 3rd Qu.: 0.00937 3rd Qu.: 0.5309   
## Max. : 9.5264 Max. : 9.91427 Max. : 5.31915 Max. : 5.7387   
## NA's :1 NA's :1 NA's :31 NA's :31   
## SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN   
## Min. :-1.3450 Min. :-4.3196 Min. :-0.6617 Min. :-1.3843   
## 1st Qu.:-0.6862 1st Qu.:-0.5041 1st Qu.:-0.5846 1st Qu.:-0.8998   
## Median :-0.3576 Median :-0.0638 Median :-0.4169 Median :-0.2240   
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.4060 3rd Qu.: 0.6089 3rd Qu.: 0.1701 3rd Qu.: 0.7875   
## Max. : 3.6932 Max. : 5.1771 Max. : 6.1526 Max. : 2.6985   
## NA's :96 NA's :14 NA's :12 NA's :13   
## SE.COM.DURS   
## Min. :-3.83095   
## 1st Qu.:-0.35501   
## Median : 0.03121   
## Mean : 0.00000   
## 3rd Qu.: 0.80364   
## Max. : 2.73472   
## NA's :19

# Little EDA  
# creating a table format for the missing values, md.pattern visualises the table giving insight  
# into the number of missing values for every column  
dim(lifeExpect1)

## [1] 217 17

md.pattern(lifeExpect1, rotate.names = TRUE)



## Continent EG.ELC.ACCS.ZS SP.POP.GROW EN.POP.DNST SP.POP.TOTL NY.GDP.PCAP.CD  
## 29 1 1 1 1 1 1  
## 63 1 1 1 1 1 1  
## 1 1 1 1 1 1 1  
## 36 1 1 1 1 1 1  
## 4 1 1 1 1 1 1  
## 13 1 1 1 1 1 1  
## 18 1 1 1 1 1 1  
## 1 1 1 1 1 1 1  
## 1 1 1 1 1 1 1  
## 2 1 1 1 1 1 1  
## 1 1 1 1 1 1 1  
## 2 1 1 1 1 1 1  
## 2 1 1 1 1 1 1  
## 3 1 1 1 1 1 1  
## 5 1 1 1 1 1 1  
## 2 1 1 1 1 1 1  
## 1 1 1 1 1 1 1  
## 1 1 1 1 1 1 1  
## 1 1 1 1 1 1 1  
## 2 1 1 1 1 1 1  
## 1 1 1 1 1 1 1  
## 2 1 1 1 1 1 1  
## 3 1 1 1 1 1 1  
## 1 1 1 1 1 1 1  
## 5 1 1 1 1 1 1  
## 1 1 1 1 1 1 1  
## 2 1 1 1 1 1 1  
## 1 1 1 1 1 1 0  
## 2 1 1 1 1 1 0  
## 2 1 1 1 1 1 0  
## 1 1 1 1 1 1 0  
## 2 1 1 1 1 1 0  
## 3 1 1 1 1 1 0  
## 1 1 1 1 0 1 1  
## 1 1 1 0 1 0 0  
## 1 1 0 1 1 1 1  
## 0 1 1 1 1 12  
## SP.DYN.CBRT.IN NY.GDP.MKTP.KD.ZG SP.DYN.LE00.IN SE.COM.DURS SP.DYN.IMRT.IN  
## 29 1 1 1 1 1  
## 63 1 1 1 1 1  
## 1 1 1 1 1 1  
## 36 1 1 1 1 1  
## 4 1 1 1 1 1  
## 13 1 1 1 1 1  
## 18 1 1 1 1 1  
## 1 1 1 1 1 1  
## 1 1 1 1 1 1  
## 2 1 1 1 1 1  
## 1 1 1 1 1 0  
## 2 1 1 1 1 0  
## 2 1 1 1 1 0  
## 3 1 1 1 0 1  
## 5 1 1 1 0 1  
## 2 1 1 1 0 1  
## 1 1 1 1 0 1  
## 1 1 1 1 0 0  
## 1 1 1 0 1 1  
## 2 1 1 0 1 1  
## 1 1 1 0 1 0  
## 2 1 1 0 1 0  
## 3 1 0 1 1 0  
## 1 0 1 0 1 1  
## 5 0 1 0 1 1  
## 1 0 1 0 1 0  
## 2 0 1 0 0 0  
## 1 1 1 1 0 0  
## 2 1 0 1 1 1  
## 2 1 0 1 1 1  
## 1 1 0 1 1 0  
## 2 1 0 1 0 0  
## 3 0 0 0 1 0  
## 1 1 1 1 0 0  
## 1 1 0 1 1 1  
## 1 0 1 0 0 0  
## 13 14 19 19 24  
## SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG  
## 29 1 1 1 1  
## 63 1 1 1 1  
## 1 1 1 1 1  
## 36 1 1 1 1  
## 4 1 1 0 0  
## 13 1 1 0 0  
## 18 1 1 0 0  
## 1 0 0 1 1  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 2 0 0 0 0  
## 3 1 1 1 1  
## 5 1 1 1 1  
## 2 1 1 0 0  
## 1 0 0 0 0  
## 1 0 0 0 0  
## 1 1 1 0 0  
## 2 1 1 0 0  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 3 0 0 0 0  
## 1 1 1 0 0  
## 5 1 1 0 0  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 1 0 0 0 0  
## 2 1 1 0 0  
## 2 0 0 0 0  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 3 0 0 0 0  
## 1 0 0 0 0  
## 1 1 1 0 0  
## 1 0 0 0 0  
## 31 31 79 79  
## SL.UEM.TOTL.NE.ZS SE.PRM.CUAT.ZS   
## 29 1 1 0  
## 63 1 0 1  
## 1 0 1 1  
## 36 0 0 2  
## 4 1 1 2  
## 13 1 0 3  
## 18 0 0 4  
## 1 1 0 3  
## 1 1 0 5  
## 2 0 0 6  
## 1 1 1 5  
## 2 1 0 6  
## 2 0 0 7  
## 3 1 0 2  
## 5 0 0 3  
## 2 0 0 5  
## 1 1 0 6  
## 1 0 0 8  
## 1 1 0 4  
## 2 0 0 5  
## 1 0 1 7  
## 2 0 0 8  
## 3 0 0 8  
## 1 1 0 5  
## 5 0 0 6  
## 1 0 0 9  
## 2 0 0 10  
## 1 0 0 9  
## 2 0 0 6  
## 2 0 0 8  
## 1 0 0 9  
## 2 0 0 10  
## 3 0 0 11  
## 1 1 0 8  
## 1 0 0 8  
## 1 0 0 11  
## 96 181 602

# summing up the number of null values for all the columns  
colSums(is.na(lifeExpect1))

## Continent SP.DYN.LE00.IN EG.ELC.ACCS.ZS   
## 0 19 1   
## NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS   
## 79 79 181   
## SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST   
## 24 1 1   
## SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS   
## 1 31 31   
## SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD   
## 96 14 12   
## SP.DYN.CBRT.IN SE.COM.DURS   
## 13 19

# mice imputation  
my\_imp = mice(lifeExpect1, m=5,   
 method = c("", "cart","cart","cart","cart","cart",  
 "cart","cart","cart","cart","cart","cart","cart",  
 "cart","cart","cart","cart"),  
 maxit=10)

##   
## iter imp variable  
## 1 1 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 1 2 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 1 3 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 1 4 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 1 5 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 2 1 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 2 2 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 2 3 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 2 4 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 2 5 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 3 1 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 3 2 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 3 3 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 3 4 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 3 5 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 4 1 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 4 2 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 4 3 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 4 4 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 4 5 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 5 1 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 5 2 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 5 3 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 5 4 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 5 5 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 6 1 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 6 2 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 6 3 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 6 4 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 6 5 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 7 1 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 7 2 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 7 3 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 7 4 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 7 5 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 8 1 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 8 2 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 8 3 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 8 4 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 8 5 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 9 1 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 9 2 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 9 3 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 9 4 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 9 5 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 10 1 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 10 2 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 10 3 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 10 4 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 10 5 SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG NY.ADJ.NNTY.PC.KD.ZG SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS

## Warning: Number of logged events: 50

# summary to inspect and assign the mean value for further inspection  
summary(lifeExpect1$SP.DYN.LE00.IN)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 53.28 67.89 74.23 72.93 78.48 85.08 19

meann <- as.numeric(format(round(summary(lifeExpect1$SP.DYN.LE00.IN)[4], 2)))  
  
# empty vector to hold difference values later  
a <- c()  
  
my\_imp$imp$SP.DYN.LE00.IN

## 1 2 3 4 5  
## 4 75.79700 68.19100 73.94390 71.68200 74.23585  
## 5 83.19756 83.19756 82.90732 82.74634 82.01220  
## 28 72.27400 66.46700 79.97317 74.91463 80.07400  
## 37 82.30244 81.70732 80.22700 78.49756 82.90000  
## 51 73.32100 75.79700 76.68000 77.65700 78.50600  
## 56 68.36900 69.49500 64.83300 68.19100 67.92300  
## 76 71.77700 78.93000 75.79700 78.49756 72.53400  
## 78 80.94146 82.80488 80.07400 79.97317 79.66829  
## 95 82.01220 73.91800 81.78537 81.28049 81.74634  
## 125 71.09700 71.09700 69.90600 67.44400 70.60400  
## 131 82.95854 82.30244 82.44634 82.44634 82.56098  
## 138 69.65600 72.59100 70.60400 72.59100 69.90600  
## 147 80.68293 71.68200 79.66829 78.80200 81.28049  
## 151 77.63415 80.68293 79.66829 77.97200 80.07400  
## 165 80.94146 80.94146 81.79268 81.78537 82.01220  
## 173 71.51300 70.90700 73.31700 72.53400 73.00500  
## 182 72.53400 75.88100 71.90100 73.32100 77.01600  
## 201 73.91800 67.92300 77.01600 76.91200 77.01600  
## 202 74.52600 76.20300 71.72500 75.27000 76.29300

# shows the mean values of the different variants of the imputed columns and   
# assigns difference for further evaluation  
for(i in 1:5){  
 diff <- as.numeric(mean(my\_imp$imp$SP.DYN.LE00.IN[,i])- meann)  
 print(  
 sprintf("%d column, mean:%.3f, diff:%.3f", i, mean(my\_imp$imp$SP.DYN.LE00.IN[,i]), diff)  
 )  
 a[i] = diff  
}

## [1] "1 column, mean:76.076, diff:3.146"  
## [1] "2 column, mean:75.301, diff:2.371"  
## [1] "3 column, mean:76.540, diff:3.610"  
## [1] "4 column, mean:76.448, diff:3.518"  
## [1] "5 column, mean:77.335, diff:4.405"

min = which(a == min(a))  
cat("The least differnce between the mean the the imputed values is shown in column:", min, '\n')

## The least differnce between the mean the the imputed values is shown in column: 2

final\_clean\_ds = complete(my\_imp, min)  
  
# library imports  
library(corrplot)

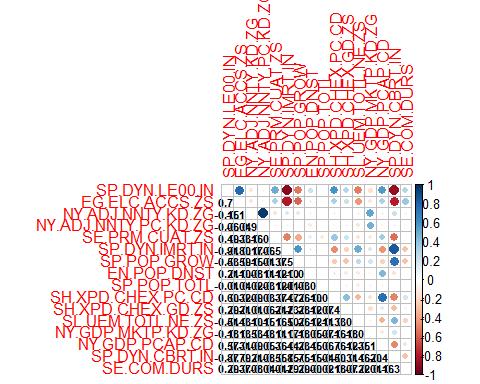
## corrplot 0.92 loaded

library(olsrr)

##   
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':  
##   
## rivers

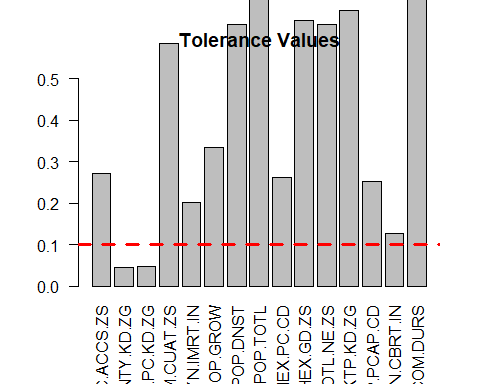
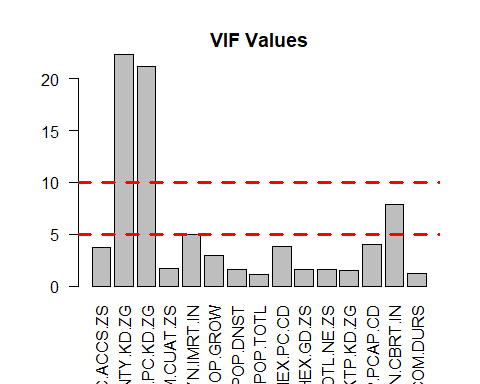
tol\_vif\_plot <- function(pred\_vars) {  
 #linear model between SP.DYN.LE00.IN (life expectancy) and all other predictive variables in the dataset  
 model = lm(pred\_vars$SP.DYN.LE00.IN ~ ., data=pred\_vars)  
   
 #gets the tolerance and TIF values of the model  
 tol\_vif\_values <- ols\_vif\_tol(model)  
   
 #displays the TIF and Tolerance values  
 tol\_vif\_values  
   
 #creates a new barplot for the TIF values  
 barplot(tol\_vif\_values$VIF, names.arg=tol\_vif\_values$Variables, las=2, main = "VIF Values", ylim = c(0, 20))  
   
 #adds dashed lines to bar chart  
 abline(h = 5, lwd = 3, lty = 2,col = 'red')  
 abline(h = 10, lwd = 3, lty = 2,col = 'red')  
   
 #creates a new barplot for the Tolerance values  
 barplot(tol\_vif\_values$Tolerance, names.arg=tol\_vif\_values$Variables, las=2, main = "Tolerance Values", ylim = c(0, 0.5))  
   
 #adds dashed line to bar chart  
 abline(h = 0.1, lwd = 3, lty = 2,col = 'red')  
}  
  
par(mfrow=c(1,1))  
  
#create subsets, one of the predictive variables, one without  
pred\_vars <- subset (final\_clean\_ds, select = -c(Continent))  
other\_vars <- subset (final\_clean\_ds, select = c(Continent))  
  
#set all missing values to 0  
pred\_vars[is.na(pred\_vars)] <- 0  
  
#get correlation matrix for all predictive variables  
cor\_matrix <-cor(pred\_vars)  
  
#get a plot which shows the correlation matrix (titles on left and top, text color black, text size = 0.7)  
cor\_plot <- corrplot.mixed(cor\_matrix, tl.pos = "lt", lower.col = "black", number.cex = 0.7)



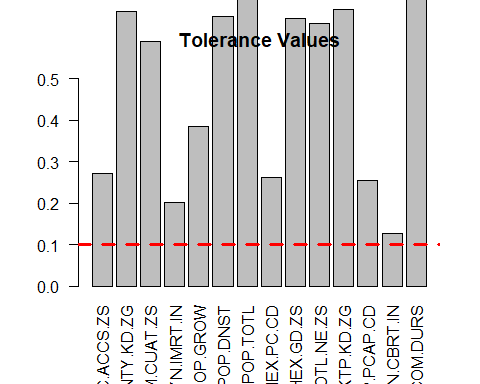
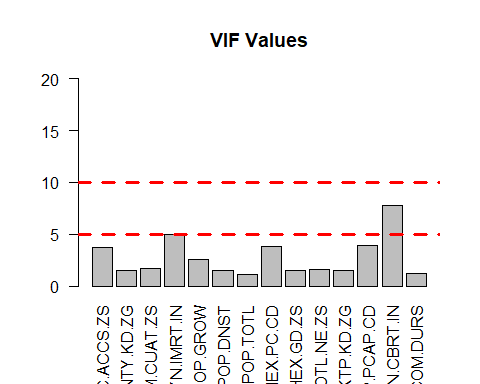
#the variables with the highest correlation are NY.ADJ.NNTY.KD.ZG and NY.ADJ.NNTY.PC.KD.ZG  
cor(lifeExpect1$NY.ADJ.NNTY.KD.ZG, lifeExpect1$NY.ADJ.NNTY.PC.KD.ZG, method = "pearson", use = "complete.obs")

## [1] 0.9836795

#The very high correlation between some variables suggests there may be some collinearity in our data, especially when looking at NY.ADJ.NNTY.KD.ZG and NY.ADJ.NNTY.PC.KD.ZG  
  
tol\_vif\_plot(pred\_vars)



#when looking at our VIF values, we have two values above 10 which is not within tolerance, NY.ADJ.NNTY.KD.ZG and NY.ADJ.NNTY.PC.KD.ZG  
#when we look at the corresponding Tolerance stats, they are both below 0.1, which is also not within tolerance.  
  
#we choose to remove NY.ADJ.NNTY.PC.KD.ZG as it is the average value per capita, compared to the total amount.  
pred\_vars <- subset (pred\_vars, select = -c(NY.ADJ.NNTY.PC.KD.ZG))  
  
tol\_vif\_plot(pred\_vars)



#combining our variables  
final\_clean\_ds <- cbind(other\_vars, pred\_vars)  
  
reg\_data <- final\_clean\_ds  
  
summary(final\_clean\_ds)

## Continent SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG   
## Africa :54 Min. :53.28 Min. :-3.265234 Min. :-5.18696   
## Asia :50 1st Qu.:68.19 1st Qu.:-0.060197 1st Qu.:-0.39709   
## Australia/Oceania:19 Median :74.24 Median : 0.553955 Median :-0.08217   
## Europe :48 Mean :73.13 Mean : 0.001855 Mean :-0.04281   
## North America :34 3rd Qu.:78.57 3rd Qu.: 0.553955 3rd Qu.: 0.29676   
## South America :12 Max. :85.08 Max. : 0.553955 Max. : 6.87298   
## SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST   
## Min. :-2.53180 Min. :-1.00741 Min. :-2.566172 Min. :-0.223332   
## 1st Qu.:-1.32050 1st Qu.:-0.79939 1st Qu.:-0.730943 1st Qu.:-0.204227   
## Median :-0.01614 Median :-0.39894 Median :-0.085880 Median :-0.176960   
## Mean :-0.32121 Mean :-0.05608 Mean : 0.004099 Mean :-0.000815   
## 3rd Qu.: 0.71415 3rd Qu.: 0.47477 3rd Qu.: 0.704884 3rd Qu.:-0.107211   
## Max. : 0.81288 Max. : 3.19469 Max. : 3.002089 Max. : 9.526383   
## SP.POP.TOTL SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS  
## Min. :-0.256003 Min. :-0.61141 Min. :-1.675398 Min. :-1.3450   
## 1st Qu.:-0.250425 1st Qu.:-0.56550 1st Qu.:-0.740729 1st Qu.:-0.5636   
## Median :-0.207116 Median :-0.38977 Median :-0.076711 Median :-0.1072   
## Mean :-0.000888 Mean : 0.06338 Mean :-0.003524 Mean : 0.3726   
## 3rd Qu.:-0.072824 3rd Qu.: 0.12310 3rd Qu.: 0.477418 3rd Qu.: 1.1413   
## Max. : 9.914275 Max. : 5.31915 Max. : 5.738704 Max. : 3.6933   
## NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS   
## Min. :-4.31956 Min. :-0.66167 Min. :-1.384329 Min. :-3.83095   
## 1st Qu.:-0.49759 1st Qu.:-0.58457 1st Qu.:-0.895976 1st Qu.:-0.35501   
## Median :-0.06143 Median :-0.42312 Median :-0.213330 Median : 0.03121   
## Mean : 0.02438 Mean :-0.01573 Mean :-0.009062 Mean :-0.04176   
## 3rd Qu.: 0.63482 3rd Qu.: 0.12390 3rd Qu.: 0.723510 3rd Qu.: 0.80364   
## Max. : 5.17711 Max. : 6.15256 Max. : 2.698498 Max. : 2.73472

# reg\_data <- final\_clean\_ds[!names(final\_clean\_ds) %in% c("Country.Code", "Country.Name", "Continent")]  
reg\_data <- final\_clean\_ds[!names(final\_clean\_ds) %in% c("Country.Code", "Country.Name")]  
  
  
library('MASS')

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:olsrr':  
##   
## cement

library('leaps')  
library('caret')

## Loading required package: ggplot2

## Loading required package: lattice

attach(reg\_data, warn.conflicts = F)  
  
index <- createDataPartition(SP.DYN.LE00.IN , p=0.8, times = 1, list = F)  
  
reg\_data\_train <- reg\_data[index,]  
reg\_data\_test <- reg\_data[-index,]  
  
nrow(reg\_data\_train)

## [1] 176

nrow(reg\_data\_test)

## [1] 41

full\_model <- lm(SP.DYN.LE00.IN ~ ., data = reg\_data\_train, x = T)  
  
null\_model <- lm(SP.DYN.LE00.IN ~ 1, data = reg\_data\_train)  
  
# looks like   
# Continent  
# SP.DYN.IMRT.IN  
# SP.POP.GROW  
# EN.POP.DNST  
# SH.XPD.CHEX.PC.CD  
# SP.DYN.CBRT.IN   
  
# are the significant variables, taking a first look at the full model  
# interpreting lm summary https://stats.stackexchange.com/questions/5135/interpretation-of-rs-lm-output  
summary(full\_model)

##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = reg\_data\_train, x = T)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.7218 -1.2916 0.0493 1.1147 4.6302   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 72.444579 0.531540 136.292 < 0.0000000000000002  
## ContinentAsia 0.550566 0.690616 0.797 0.42654  
## ContinentAustralia/Oceania 0.354375 0.804502 0.440 0.66019  
## ContinentEurope 0.501314 0.818902 0.612 0.54131  
## ContinentNorth America 0.527580 0.740400 0.713 0.47718  
## ContinentSouth America 0.493159 0.929787 0.530 0.59659  
## EG.ELC.ACCS.ZS 0.523753 0.349958 1.497 0.13651  
## NY.ADJ.NNTY.KD.ZG 0.218981 0.217582 1.006 0.31577  
## SE.PRM.CUAT.ZS -0.199602 0.192643 -1.036 0.30175  
## SP.DYN.IMRT.IN -4.013476 0.413056 -9.717 < 0.0000000000000002  
## SP.POP.GROW 1.704596 0.305468 5.580 0.000000103780  
## EN.POP.DNST 0.563444 0.188614 2.987 0.00327  
## SP.POP.TOTL -0.171877 0.157211 -1.093 0.27595  
## SH.XPD.CHEX.PC.CD 0.940615 0.283948 3.313 0.00115  
## SH.XPD.CHEX.GD.ZS 0.072441 0.216391 0.335 0.73825  
## SL.UEM.TOTL.NE.ZS 0.002674 0.182955 0.015 0.98836  
## NY.GDP.MKTP.KD.ZG -0.161567 0.199079 -0.812 0.41827  
## NY.GDP.PCAP.CD -0.012298 0.308248 -0.040 0.96823  
## SP.DYN.CBRT.IN -3.311894 0.488884 -6.774 0.000000000241  
## SE.COM.DURS 0.133027 0.193127 0.689 0.49197  
##   
## (Intercept) \*\*\*  
## ContinentAsia   
## ContinentAustralia/Oceania   
## ContinentEurope   
## ContinentNorth America   
## ContinentSouth America   
## EG.ELC.ACCS.ZS   
## NY.ADJ.NNTY.KD.ZG   
## SE.PRM.CUAT.ZS   
## SP.DYN.IMRT.IN \*\*\*  
## SP.POP.GROW \*\*\*  
## EN.POP.DNST \*\*   
## SP.POP.TOTL   
## SH.XPD.CHEX.PC.CD \*\*   
## SH.XPD.CHEX.GD.ZS   
## SL.UEM.TOTL.NE.ZS   
## NY.GDP.MKTP.KD.ZG   
## NY.GDP.PCAP.CD   
## SP.DYN.CBRT.IN \*\*\*  
## SE.COM.DURS   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.144 on 156 degrees of freedom  
## Multiple R-squared: 0.9224, Adjusted R-squared: 0.9129   
## F-statistic: 97.57 on 19 and 156 DF, p-value: < 0.00000000000000022

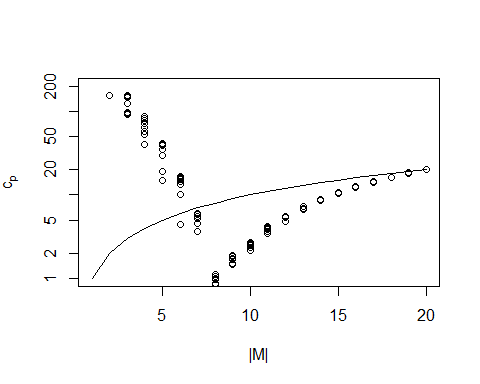
# interpreting anova https://stats.stackexchange.com/questions/115304/interpreting-output-from-anova-when-using-lm-as-input  
anova(full\_model)

## Analysis of Variance Table  
##   
## Response: SP.DYN.LE00.IN  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Continent 5 4904.7 980.93 213.3585 < 0.00000000000000022 \*\*\*  
## EG.ELC.ACCS.ZS 1 1375.2 1375.21 299.1174 < 0.00000000000000022 \*\*\*  
## NY.ADJ.NNTY.KD.ZG 1 44.0 44.01 9.5729 0.0023408 \*\*   
## SE.PRM.CUAT.ZS 1 270.9 270.92 58.9258 0.000000000001675 \*\*\*  
## SP.DYN.IMRT.IN 1 1253.5 1253.45 272.6336 < 0.00000000000000022 \*\*\*  
## SP.POP.GROW 1 61.3 61.30 13.3321 0.0003556 \*\*\*  
## EN.POP.DNST 1 133.0 133.01 28.9307 0.000000269639920 \*\*\*  
## SP.POP.TOTL 1 0.1 0.13 0.0292 0.8645139   
## SH.XPD.CHEX.PC.CD 1 256.6 256.60 55.8123 0.000000000005326 \*\*\*  
## SH.XPD.CHEX.GD.ZS 1 7.6 7.60 1.6532 0.2004290   
## SL.UEM.TOTL.NE.ZS 1 1.1 1.07 0.2330 0.6300048   
## NY.GDP.MKTP.KD.ZG 1 2.9 2.93 0.6381 0.4256018   
## NY.GDP.PCAP.CD 1 0.7 0.73 0.1580 0.6915632   
## SP.DYN.CBRT.IN 1 209.1 209.07 45.4745 0.000000000284114 \*\*\*  
## SE.COM.DURS 1 2.2 2.18 0.4745 0.4919682   
## Residuals 156 717.2 4.60   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

library(leaps) # step  
  
# our data is not that large and the number of predictors is also manageable, we  
# can use exhaustive search to find the best model given a metric  
  
# exhaustive sear via Mallow's Cp ==============================================  
X <- full\_model$x  
y <- reg\_data\_train$SP.DYN.LE00.IN  
  
all\_models <- leaps(X, y, int = F, strictly.compatible = F, method="Cp")  
  
# source: lab 3  
plot(all\_models$size, all\_models$Cp, log = "y", xlab = "|M|",  
 ylab = expression(c[p]), ylim = c(1,200))

## Warning in xy.coords(x, y, xlabel, ylabel, log): 1 y value <= 0 omitted from  
## logarithmic plot

lines(all\_models$size, all\_models$size)



min\_cp\_val <- all\_models$Cp == min(all\_models$Cp)  
min(all\_models$Cp)

## [1] -0.8554789

min\_cp <- all\_models$which[min\_cp\_val, ]  
min\_cp

## (Intercept) ContinentAsia   
## TRUE FALSE   
## ContinentAustralia/Oceania ContinentEurope   
## FALSE FALSE   
## ContinentNorth America ContinentSouth America   
## FALSE FALSE   
## EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG   
## TRUE FALSE   
## SE.PRM.CUAT.ZS SP.DYN.IMRT.IN   
## FALSE TRUE   
## SP.POP.GROW EN.POP.DNST   
## TRUE TRUE   
## SP.POP.TOTL SH.XPD.CHEX.PC.CD   
## FALSE TRUE   
## SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS   
## FALSE FALSE   
## NY.GDP.MKTP.KD.ZG NY.GDP.PCAP.CD   
## FALSE FALSE   
## SP.DYN.CBRT.IN SE.COM.DURS   
## TRUE FALSE

model\_full\_cp <- lm(SP.DYN.LE00.IN ~ Continent + EG.ELC.ACCS.ZS +   
 SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST +   
 SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN,  
 data = reg\_data\_train)  
summary(model\_full\_cp)

##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ Continent + EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN +   
## SP.POP.GROW + EN.POP.DNST + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN,   
## data = reg\_data\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.7288 -1.3028 -0.0025 1.2159 4.8613   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.4861 0.4938 146.783 < 0.0000000000000002 \*\*\*  
## ContinentAsia 0.4524 0.6546 0.691 0.490458   
## ContinentAustralia/Oceania 0.5026 0.7354 0.683 0.495286   
## ContinentEurope 0.4937 0.7826 0.631 0.529046   
## ContinentNorth America 0.6336 0.7149 0.886 0.376753   
## ContinentSouth America 0.7147 0.8895 0.803 0.422900   
## EG.ELC.ACCS.ZS 0.6640 0.3240 2.049 0.042023 \*   
## SP.DYN.IMRT.IN -3.9239 0.3681 -10.659 < 0.0000000000000002 \*\*\*  
## SP.POP.GROW 1.6760 0.2942 5.697 0.000000055155 \*\*\*  
## EN.POP.DNST 0.5520 0.1514 3.646 0.000357 \*\*\*  
## SH.XPD.CHEX.PC.CD 0.9810 0.1820 5.389 0.000000242716 \*\*\*  
## SP.DYN.CBRT.IN -3.1207 0.4593 -6.795 0.000000000191 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.119 on 164 degrees of freedom  
## Multiple R-squared: 0.9203, Adjusted R-squared: 0.9149   
## F-statistic: 172.1 on 11 and 164 DF, p-value: < 0.00000000000000022

anova(model\_full\_cp)

## Analysis of Variance Table  
##   
## Response: SP.DYN.LE00.IN  
## Df Sum Sq Mean Sq F value Pr(>F)   
## Continent 5 4904.7 980.93 218.430 < 0.00000000000000022 \*\*\*  
## EG.ELC.ACCS.ZS 1 1375.2 1375.21 306.228 < 0.00000000000000022 \*\*\*  
## SP.DYN.IMRT.IN 1 1557.6 1557.63 346.847 < 0.00000000000000022 \*\*\*  
## SP.POP.GROW 1 51.9 51.86 11.547 0.0008518 \*\*\*  
## EN.POP.DNST 1 141.1 141.12 31.423 0.000000086128226 \*\*\*  
## SH.XPD.CHEX.PC.CD 1 265.8 265.79 59.186 0.000000000001264 \*\*\*  
## SP.DYN.CBRT.IN 1 207.3 207.34 46.170 0.000000000190761 \*\*\*  
## Residuals 164 736.5 4.49   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# testing wrapper methods ======================================================  
  
model\_step\_backward <- step(full\_model, method = "backward")

## Start: AIC=287.26  
## SP.DYN.LE00.IN ~ Continent + EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG +   
## SE.PRM.CUAT.ZS + SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST +   
## SP.POP.TOTL + SH.XPD.CHEX.PC.CD + SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS +   
## NY.GDP.MKTP.KD.ZG + NY.GDP.PCAP.CD + SP.DYN.CBRT.IN + SE.COM.DURS  
##   
## Df Sum of Sq RSS AIC  
## - Continent 5 3.25 720.47 278.06  
## - SL.UEM.TOTL.NE.ZS 1 0.00 717.22 285.26  
## - NY.GDP.PCAP.CD 1 0.01 717.23 285.26  
## - SH.XPD.CHEX.GD.ZS 1 0.52 717.74 285.39  
## - SE.COM.DURS 1 2.18 719.40 285.80  
## - NY.GDP.MKTP.KD.ZG 1 3.03 720.25 286.00  
## - NY.ADJ.NNTY.KD.ZG 1 4.66 721.88 286.40  
## - SE.PRM.CUAT.ZS 1 4.94 722.16 286.47  
## - SP.POP.TOTL 1 5.50 722.72 286.61  
## <none> 717.22 287.26  
## - EG.ELC.ACCS.ZS 1 10.30 727.52 287.77  
## - EN.POP.DNST 1 41.03 758.25 295.05  
## - SH.XPD.CHEX.PC.CD 1 50.45 767.67 297.23  
## - SP.POP.GROW 1 143.17 860.39 317.29  
## - SP.DYN.CBRT.IN 1 210.99 928.22 330.65  
## - SP.DYN.IMRT.IN 1 434.06 1151.28 368.55  
##   
## Step: AIC=278.06  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + SE.PRM.CUAT.ZS +   
## SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +   
## SH.XPD.CHEX.PC.CD + SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS +   
## NY.GDP.MKTP.KD.ZG + NY.GDP.PCAP.CD + SP.DYN.CBRT.IN + SE.COM.DURS  
##   
## Df Sum of Sq RSS AIC  
## - NY.GDP.PCAP.CD 1 0.00 720.47 276.06  
## - SL.UEM.TOTL.NE.ZS 1 0.11 720.58 276.09  
## - SH.XPD.CHEX.GD.ZS 1 0.56 721.03 276.19  
## - SE.COM.DURS 1 2.90 723.37 276.77  
## - NY.GDP.MKTP.KD.ZG 1 3.28 723.75 276.86  
## - SE.PRM.CUAT.ZS 1 4.56 725.04 277.17  
## - NY.ADJ.NNTY.KD.ZG 1 5.05 725.52 277.29  
## - SP.POP.TOTL 1 5.21 725.69 277.33  
## <none> 720.47 278.06  
## - EG.ELC.ACCS.ZS 1 18.57 739.04 280.54  
## - EN.POP.DNST 1 42.11 762.59 286.06  
## - SH.XPD.CHEX.PC.CD 1 51.93 772.40 288.31  
## - SP.POP.GROW 1 150.84 871.31 309.51  
## - SP.DYN.CBRT.IN 1 226.20 946.67 324.11  
## - SP.DYN.IMRT.IN 1 446.77 1167.25 360.98  
##   
## Step: AIC=276.06  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + SE.PRM.CUAT.ZS +   
## SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +   
## SH.XPD.CHEX.PC.CD + SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS +   
## NY.GDP.MKTP.KD.ZG + SP.DYN.CBRT.IN + SE.COM.DURS  
##   
## Df Sum of Sq RSS AIC  
## - SL.UEM.TOTL.NE.ZS 1 0.11 720.58 274.09  
## - SH.XPD.CHEX.GD.ZS 1 0.61 721.09 274.21  
## - SE.COM.DURS 1 2.90 723.37 274.77  
## - NY.GDP.MKTP.KD.ZG 1 3.28 723.76 274.86  
## - SE.PRM.CUAT.ZS 1 4.57 725.04 275.17  
## - NY.ADJ.NNTY.KD.ZG 1 5.05 725.52 275.29  
## - SP.POP.TOTL 1 5.26 725.73 275.34  
## <none> 720.47 276.06  
## - EG.ELC.ACCS.ZS 1 18.65 739.13 278.56  
## - EN.POP.DNST 1 59.32 779.80 287.98  
## - SH.XPD.CHEX.PC.CD 1 108.33 828.80 298.71  
## - SP.POP.GROW 1 150.85 871.32 307.52  
## - SP.DYN.CBRT.IN 1 227.70 948.18 322.39  
## - SP.DYN.IMRT.IN 1 447.58 1168.05 359.10  
##   
## Step: AIC=274.09  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + SE.PRM.CUAT.ZS +   
## SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +   
## SH.XPD.CHEX.PC.CD + SH.XPD.CHEX.GD.ZS + NY.GDP.MKTP.KD.ZG +   
## SP.DYN.CBRT.IN + SE.COM.DURS  
##   
## Df Sum of Sq RSS AIC  
## - SH.XPD.CHEX.GD.ZS 1 0.56 721.14 272.22  
## - SE.COM.DURS 1 3.03 723.61 272.82  
## - NY.GDP.MKTP.KD.ZG 1 3.20 723.78 272.86  
## - SE.PRM.CUAT.ZS 1 4.80 725.39 273.26  
## - SP.POP.TOTL 1 5.16 725.75 273.34  
## - NY.ADJ.NNTY.KD.ZG 1 5.66 726.25 273.46  
## <none> 720.58 274.09  
## - EG.ELC.ACCS.ZS 1 18.59 739.18 276.57  
## - EN.POP.DNST 1 59.92 780.51 286.14  
## - SH.XPD.CHEX.PC.CD 1 110.55 831.14 297.21  
## - SP.POP.GROW 1 155.01 875.60 306.38  
## - SP.DYN.CBRT.IN 1 231.68 952.27 321.15  
## - SP.DYN.IMRT.IN 1 494.10 1214.68 363.99  
##   
## Step: AIC=272.22  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + SE.PRM.CUAT.ZS +   
## SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +   
## SH.XPD.CHEX.PC.CD + NY.GDP.MKTP.KD.ZG + SP.DYN.CBRT.IN +   
## SE.COM.DURS  
##   
## Df Sum of Sq RSS AIC  
## - SE.COM.DURS 1 2.96 724.10 270.94  
## - NY.GDP.MKTP.KD.ZG 1 3.76 724.90 271.14  
## - SE.PRM.CUAT.ZS 1 4.65 725.79 271.35  
## - SP.POP.TOTL 1 5.51 726.65 271.56  
## - NY.ADJ.NNTY.KD.ZG 1 6.05 727.19 271.69  
## <none> 721.14 272.22  
## - EG.ELC.ACCS.ZS 1 19.62 740.76 274.95  
## - EN.POP.DNST 1 60.40 781.55 284.38  
## - SH.XPD.CHEX.PC.CD 1 131.25 852.39 299.65  
## - SP.POP.GROW 1 154.62 875.76 304.41  
## - SP.DYN.CBRT.IN 1 234.23 955.37 319.72  
## - SP.DYN.IMRT.IN 1 507.44 1228.58 363.99  
##   
## Step: AIC=270.94  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + SE.PRM.CUAT.ZS +   
## SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +   
## SH.XPD.CHEX.PC.CD + NY.GDP.MKTP.KD.ZG + SP.DYN.CBRT.IN  
##   
## Df Sum of Sq RSS AIC  
## - NY.GDP.MKTP.KD.ZG 1 3.13 727.23 269.70  
## - NY.ADJ.NNTY.KD.ZG 1 4.99 729.10 270.15  
## - SE.PRM.CUAT.ZS 1 5.41 729.52 270.25  
## - SP.POP.TOTL 1 6.10 730.20 270.42  
## <none> 724.10 270.94  
## - EG.ELC.ACCS.ZS 1 23.77 747.87 274.63  
## - EN.POP.DNST 1 59.59 783.69 282.86  
## - SH.XPD.CHEX.PC.CD 1 139.62 863.72 299.97  
## - SP.POP.GROW 1 152.84 876.95 302.65  
## - SP.DYN.CBRT.IN 1 232.91 957.01 318.03  
## - SP.DYN.IMRT.IN 1 506.48 1230.58 362.28  
##   
## Step: AIC=269.7  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + SE.PRM.CUAT.ZS +   
## SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +   
## SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN  
##   
## Df Sum of Sq RSS AIC  
## - NY.ADJ.NNTY.KD.ZG 1 2.59 729.82 268.33  
## - SE.PRM.CUAT.ZS 1 5.84 733.08 269.11  
## - SP.POP.TOTL 1 6.95 734.18 269.38  
## <none> 727.23 269.70  
## - EG.ELC.ACCS.ZS 1 22.63 749.86 273.10  
## - EN.POP.DNST 1 60.09 787.33 281.68  
## - SH.XPD.CHEX.PC.CD 1 143.92 871.15 299.48  
## - SP.POP.GROW 1 149.72 876.95 300.65  
## - SP.DYN.CBRT.IN 1 231.18 958.41 316.28  
## - SP.DYN.IMRT.IN 1 519.92 1247.16 362.63  
##   
## Step: AIC=268.33  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SE.PRM.CUAT.ZS + SP.DYN.IMRT.IN +   
## SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL + SH.XPD.CHEX.PC.CD +   
## SP.DYN.CBRT.IN  
##   
## Df Sum of Sq RSS AIC  
## - SE.PRM.CUAT.ZS 1 5.95 735.78 267.76  
## - SP.POP.TOTL 1 6.70 736.53 267.94  
## <none> 729.82 268.33  
## - EG.ELC.ACCS.ZS 1 24.75 754.57 272.20  
## - EN.POP.DNST 1 59.02 788.85 280.02  
## - SH.XPD.CHEX.PC.CD 1 146.83 876.66 298.59  
## - SP.POP.GROW 1 151.03 880.86 299.43  
## - SP.DYN.CBRT.IN 1 228.68 958.51 314.30  
## - SP.DYN.IMRT.IN 1 517.78 1247.60 360.70  
##   
## Step: AIC=267.76  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN + SP.POP.GROW +   
## EN.POP.DNST + SP.POP.TOTL + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN  
##   
## Df Sum of Sq RSS AIC  
## - SP.POP.TOTL 1 5.10 740.87 266.97  
## <none> 735.78 267.76  
## - EG.ELC.ACCS.ZS 1 34.15 769.93 273.74  
## - EN.POP.DNST 1 57.96 793.74 279.10  
## - SH.XPD.CHEX.PC.CD 1 143.14 878.92 297.04  
## - SP.POP.GROW 1 153.00 888.78 299.01  
## - SP.DYN.CBRT.IN 1 224.98 960.75 312.71  
## - SP.DYN.IMRT.IN 1 514.26 1250.03 359.04  
##   
## Step: AIC=266.97  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN + SP.POP.GROW +   
## EN.POP.DNST + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN  
##   
## Df Sum of Sq RSS AIC  
## <none> 740.87 266.97  
## - EG.ELC.ACCS.ZS 1 33.29 774.17 272.71  
## - EN.POP.DNST 1 59.62 800.49 278.59  
## - SH.XPD.CHEX.PC.CD 1 146.17 887.05 296.66  
## - SP.POP.GROW 1 151.55 892.43 297.73  
## - SP.DYN.CBRT.IN 1 221.23 962.11 310.96  
## - SP.DYN.IMRT.IN 1 522.27 1263.15 358.87

summary(model\_step\_backward)

##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN +   
## SP.POP.GROW + EN.POP.DNST + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN,   
## data = reg\_data\_train, x = T)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0534 -1.4002 0.0123 1.2206 4.7652   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.8858 0.1593 457.444 < 0.0000000000000002 \*\*\*  
## EG.ELC.ACCS.ZS 0.7861 0.2853 2.756 0.006498 \*\*   
## SP.DYN.IMRT.IN -3.9405 0.3610 -10.915 < 0.0000000000000002 \*\*\*  
## SP.POP.GROW 1.6445 0.2797 5.880 0.0000000214153 \*\*\*  
## EN.POP.DNST 0.5463 0.1481 3.688 0.000305 \*\*\*  
## SH.XPD.CHEX.PC.CD 0.9926 0.1719 5.774 0.0000000362023 \*\*\*  
## SP.DYN.CBRT.IN -3.1420 0.4423 -7.104 0.0000000000323 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.094 on 169 degrees of freedom  
## Multiple R-squared: 0.9198, Adjusted R-squared: 0.917   
## F-statistic: 323.1 on 6 and 169 DF, p-value: < 0.00000000000000022

model\_step\_forward <- step(null\_model, method = "forward",  
 scope =   
 # the predictors signified as significant by the lm full\_model   
 # summary  
 ~ Continent + SP.DYN.IMRT.IN + SP.POP.GROW +   
 EN.POP.DNST + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN,  
 data = reg\_data\_train)

## Start: AIC=699.1  
## SP.DYN.LE00.IN ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + SP.DYN.IMRT.IN 1 7731.0 1509.1 382.19  
## + SP.DYN.CBRT.IN 1 6923.6 2316.5 457.61  
## + Continent 5 4904.7 4335.4 575.92  
## + SH.XPD.CHEX.PC.CD 1 3295.4 5944.7 623.48  
## + SP.POP.GROW 1 2495.4 6744.7 645.70  
## + EN.POP.DNST 1 482.3 8757.8 691.67  
## <none> 9240.1 699.10  
##   
## Step: AIC=382.19  
## SP.DYN.LE00.IN ~ SP.DYN.IMRT.IN  
##   
## Df Sum of Sq RSS AIC  
## + SH.XPD.CHEX.PC.CD 1 309.6 1199.5 343.77  
## + SP.DYN.CBRT.IN 1 275.0 1234.1 348.78  
## + EN.POP.DNST 1 157.9 1351.2 364.73  
## <none> 1509.1 382.19  
## + Continent 5 80.7 1428.4 382.51  
## + SP.POP.GROW 1 5.8 1503.3 383.51  
## - SP.DYN.IMRT.IN 1 7731.0 9240.1 699.10  
##   
## Step: AIC=343.77  
## SP.DYN.LE00.IN ~ SP.DYN.IMRT.IN + SH.XPD.CHEX.PC.CD  
##   
## Df Sum of Sq RSS AIC  
## + SP.DYN.CBRT.IN 1 180.7 1018.8 317.04  
## + EN.POP.DNST 1 131.7 1067.8 325.30  
## + Continent 5 68.7 1130.8 343.39  
## <none> 1199.5 343.77  
## + SP.POP.GROW 1 2.2 1197.3 345.45  
## - SH.XPD.CHEX.PC.CD 1 309.6 1509.1 382.19  
## - SP.DYN.IMRT.IN 1 4745.2 5944.7 623.48  
##   
## Step: AIC=317.04  
## SP.DYN.LE00.IN ~ SP.DYN.IMRT.IN + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN  
##   
## Df Sum of Sq RSS AIC  
## + SP.POP.GROW 1 187.12 831.70 283.33  
## + EN.POP.DNST 1 93.60 925.23 302.08  
## <none> 1018.82 317.04  
## + Continent 5 32.88 985.95 321.27  
## - SP.DYN.CBRT.IN 1 180.67 1199.50 343.77  
## - SH.XPD.CHEX.PC.CD 1 215.30 1234.12 348.78  
## - SP.DYN.IMRT.IN 1 945.77 1964.59 430.61  
##   
## Step: AIC=283.33  
## SP.DYN.LE00.IN ~ SP.DYN.IMRT.IN + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN +   
## SP.POP.GROW  
##   
## Df Sum of Sq RSS AIC  
## + EN.POP.DNST 1 57.54 774.17 272.71  
## <none> 831.70 283.33  
## + Continent 5 18.78 812.93 289.31  
## - SH.XPD.CHEX.PC.CD 1 124.24 955.94 305.83  
## - SP.POP.GROW 1 187.12 1018.82 317.04  
## - SP.DYN.CBRT.IN 1 365.57 1197.27 345.45  
## - SP.DYN.IMRT.IN 1 794.13 1625.84 399.30  
##   
## Step: AIC=272.71  
## SP.DYN.LE00.IN ~ SP.DYN.IMRT.IN + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN +   
## SP.POP.GROW + EN.POP.DNST  
##   
## Df Sum of Sq RSS AIC  
## <none> 774.17 272.71  
## + Continent 5 18.81 755.35 278.38  
## - EN.POP.DNST 1 57.54 831.70 283.33  
## - SH.XPD.CHEX.PC.CD 1 124.64 898.80 296.98  
## - SP.POP.GROW 1 151.06 925.23 302.08  
## - SP.DYN.CBRT.IN 1 292.21 1066.37 327.07  
## - SP.DYN.IMRT.IN 1 829.20 1603.37 398.85

summary(model\_step\_forward)

##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ SP.DYN.IMRT.IN + SH.XPD.CHEX.PC.CD +   
## SP.DYN.CBRT.IN + SP.POP.GROW + EN.POP.DNST, data = reg\_data\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.746 -1.463 -0.008 1.151 4.750   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.8545 0.1620 449.771 < 0.0000000000000002 \*\*\*  
## SP.DYN.IMRT.IN -4.4010 0.3261 -13.494 < 0.0000000000000002 \*\*\*  
## SH.XPD.CHEX.PC.CD 0.8982 0.1717 5.232 0.000000489863798 \*\*\*  
## SP.DYN.CBRT.IN -3.4742 0.4337 -8.010 0.000000000000174 \*\*\*  
## SP.POP.GROW 1.6418 0.2851 5.759 0.000000038667821 \*\*\*  
## EN.POP.DNST 0.5365 0.1509 3.555 0.00049 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.134 on 170 degrees of freedom  
## Multiple R-squared: 0.9162, Adjusted R-squared: 0.9138   
## F-statistic: 371.8 on 5 and 170 DF, p-value: < 0.00000000000000022

# Here I am testing out the forwards/backwards step-wise algorithm introduced  
# in the lecture  
motel\_step\_both <- stepAIC(full\_model, directoin = "both") # worse than step forward

## Start: AIC=287.26  
## SP.DYN.LE00.IN ~ Continent + EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG +   
## SE.PRM.CUAT.ZS + SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST +   
## SP.POP.TOTL + SH.XPD.CHEX.PC.CD + SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS +   
## NY.GDP.MKTP.KD.ZG + NY.GDP.PCAP.CD + SP.DYN.CBRT.IN + SE.COM.DURS  
##   
## Df Sum of Sq RSS AIC  
## - Continent 5 3.25 720.47 278.06  
## - SL.UEM.TOTL.NE.ZS 1 0.00 717.22 285.26  
## - NY.GDP.PCAP.CD 1 0.01 717.23 285.26  
## - SH.XPD.CHEX.GD.ZS 1 0.52 717.74 285.39  
## - SE.COM.DURS 1 2.18 719.40 285.80  
## - NY.GDP.MKTP.KD.ZG 1 3.03 720.25 286.00  
## - NY.ADJ.NNTY.KD.ZG 1 4.66 721.88 286.40  
## - SE.PRM.CUAT.ZS 1 4.94 722.16 286.47  
## - SP.POP.TOTL 1 5.50 722.72 286.61  
## <none> 717.22 287.26  
## - EG.ELC.ACCS.ZS 1 10.30 727.52 287.77  
## - EN.POP.DNST 1 41.03 758.25 295.05  
## - SH.XPD.CHEX.PC.CD 1 50.45 767.67 297.23  
## - SP.POP.GROW 1 143.17 860.39 317.29  
## - SP.DYN.CBRT.IN 1 210.99 928.22 330.65  
## - SP.DYN.IMRT.IN 1 434.06 1151.28 368.55  
##   
## Step: AIC=278.06  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + SE.PRM.CUAT.ZS +   
## SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +   
## SH.XPD.CHEX.PC.CD + SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS +   
## NY.GDP.MKTP.KD.ZG + NY.GDP.PCAP.CD + SP.DYN.CBRT.IN + SE.COM.DURS  
##   
## Df Sum of Sq RSS AIC  
## - NY.GDP.PCAP.CD 1 0.00 720.47 276.06  
## - SL.UEM.TOTL.NE.ZS 1 0.11 720.58 276.09  
## - SH.XPD.CHEX.GD.ZS 1 0.56 721.03 276.19  
## - SE.COM.DURS 1 2.90 723.37 276.77  
## - NY.GDP.MKTP.KD.ZG 1 3.28 723.75 276.86  
## - SE.PRM.CUAT.ZS 1 4.56 725.04 277.17  
## - NY.ADJ.NNTY.KD.ZG 1 5.05 725.52 277.29  
## - SP.POP.TOTL 1 5.21 725.69 277.33  
## <none> 720.47 278.06  
## - EG.ELC.ACCS.ZS 1 18.57 739.04 280.54  
## - EN.POP.DNST 1 42.11 762.59 286.06  
## - SH.XPD.CHEX.PC.CD 1 51.93 772.40 288.31  
## - SP.POP.GROW 1 150.84 871.31 309.51  
## - SP.DYN.CBRT.IN 1 226.20 946.67 324.11  
## - SP.DYN.IMRT.IN 1 446.77 1167.25 360.98  
##   
## Step: AIC=276.06  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + SE.PRM.CUAT.ZS +   
## SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +   
## SH.XPD.CHEX.PC.CD + SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS +   
## NY.GDP.MKTP.KD.ZG + SP.DYN.CBRT.IN + SE.COM.DURS  
##   
## Df Sum of Sq RSS AIC  
## - SL.UEM.TOTL.NE.ZS 1 0.11 720.58 274.09  
## - SH.XPD.CHEX.GD.ZS 1 0.61 721.09 274.21  
## - SE.COM.DURS 1 2.90 723.37 274.77  
## - NY.GDP.MKTP.KD.ZG 1 3.28 723.76 274.86  
## - SE.PRM.CUAT.ZS 1 4.57 725.04 275.17  
## - NY.ADJ.NNTY.KD.ZG 1 5.05 725.52 275.29  
## - SP.POP.TOTL 1 5.26 725.73 275.34  
## <none> 720.47 276.06  
## - EG.ELC.ACCS.ZS 1 18.65 739.13 278.56  
## - EN.POP.DNST 1 59.32 779.80 287.98  
## - SH.XPD.CHEX.PC.CD 1 108.33 828.80 298.71  
## - SP.POP.GROW 1 150.85 871.32 307.52  
## - SP.DYN.CBRT.IN 1 227.70 948.18 322.39  
## - SP.DYN.IMRT.IN 1 447.58 1168.05 359.10  
##   
## Step: AIC=274.09  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + SE.PRM.CUAT.ZS +   
## SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +   
## SH.XPD.CHEX.PC.CD + SH.XPD.CHEX.GD.ZS + NY.GDP.MKTP.KD.ZG +   
## SP.DYN.CBRT.IN + SE.COM.DURS  
##   
## Df Sum of Sq RSS AIC  
## - SH.XPD.CHEX.GD.ZS 1 0.56 721.14 272.22  
## - SE.COM.DURS 1 3.03 723.61 272.82  
## - NY.GDP.MKTP.KD.ZG 1 3.20 723.78 272.86  
## - SE.PRM.CUAT.ZS 1 4.80 725.39 273.26  
## - SP.POP.TOTL 1 5.16 725.75 273.34  
## - NY.ADJ.NNTY.KD.ZG 1 5.66 726.25 273.46  
## <none> 720.58 274.09  
## - EG.ELC.ACCS.ZS 1 18.59 739.18 276.57  
## - EN.POP.DNST 1 59.92 780.51 286.14  
## - SH.XPD.CHEX.PC.CD 1 110.55 831.14 297.21  
## - SP.POP.GROW 1 155.01 875.60 306.38  
## - SP.DYN.CBRT.IN 1 231.68 952.27 321.15  
## - SP.DYN.IMRT.IN 1 494.10 1214.68 363.99  
##   
## Step: AIC=272.22  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + SE.PRM.CUAT.ZS +   
## SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +   
## SH.XPD.CHEX.PC.CD + NY.GDP.MKTP.KD.ZG + SP.DYN.CBRT.IN +   
## SE.COM.DURS  
##   
## Df Sum of Sq RSS AIC  
## - SE.COM.DURS 1 2.96 724.10 270.94  
## - NY.GDP.MKTP.KD.ZG 1 3.76 724.90 271.14  
## - SE.PRM.CUAT.ZS 1 4.65 725.79 271.35  
## - SP.POP.TOTL 1 5.51 726.65 271.56  
## - NY.ADJ.NNTY.KD.ZG 1 6.05 727.19 271.69  
## <none> 721.14 272.22  
## - EG.ELC.ACCS.ZS 1 19.62 740.76 274.95  
## - EN.POP.DNST 1 60.40 781.55 284.38  
## - SH.XPD.CHEX.PC.CD 1 131.25 852.39 299.65  
## - SP.POP.GROW 1 154.62 875.76 304.41  
## - SP.DYN.CBRT.IN 1 234.23 955.37 319.72  
## - SP.DYN.IMRT.IN 1 507.44 1228.58 363.99  
##   
## Step: AIC=270.94  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + SE.PRM.CUAT.ZS +   
## SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +   
## SH.XPD.CHEX.PC.CD + NY.GDP.MKTP.KD.ZG + SP.DYN.CBRT.IN  
##   
## Df Sum of Sq RSS AIC  
## - NY.GDP.MKTP.KD.ZG 1 3.13 727.23 269.70  
## - NY.ADJ.NNTY.KD.ZG 1 4.99 729.10 270.15  
## - SE.PRM.CUAT.ZS 1 5.41 729.52 270.25  
## - SP.POP.TOTL 1 6.10 730.20 270.42  
## <none> 724.10 270.94  
## - EG.ELC.ACCS.ZS 1 23.77 747.87 274.63  
## - EN.POP.DNST 1 59.59 783.69 282.86  
## - SH.XPD.CHEX.PC.CD 1 139.62 863.72 299.97  
## - SP.POP.GROW 1 152.84 876.95 302.65  
## - SP.DYN.CBRT.IN 1 232.91 957.01 318.03  
## - SP.DYN.IMRT.IN 1 506.48 1230.58 362.28  
##   
## Step: AIC=269.7  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG + SE.PRM.CUAT.ZS +   
## SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL +   
## SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN  
##   
## Df Sum of Sq RSS AIC  
## - NY.ADJ.NNTY.KD.ZG 1 2.59 729.82 268.33  
## - SE.PRM.CUAT.ZS 1 5.84 733.08 269.11  
## - SP.POP.TOTL 1 6.95 734.18 269.38  
## <none> 727.23 269.70  
## - EG.ELC.ACCS.ZS 1 22.63 749.86 273.10  
## - EN.POP.DNST 1 60.09 787.33 281.68  
## - SH.XPD.CHEX.PC.CD 1 143.92 871.15 299.48  
## - SP.POP.GROW 1 149.72 876.95 300.65  
## - SP.DYN.CBRT.IN 1 231.18 958.41 316.28  
## - SP.DYN.IMRT.IN 1 519.92 1247.16 362.63  
##   
## Step: AIC=268.33  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SE.PRM.CUAT.ZS + SP.DYN.IMRT.IN +   
## SP.POP.GROW + EN.POP.DNST + SP.POP.TOTL + SH.XPD.CHEX.PC.CD +   
## SP.DYN.CBRT.IN  
##   
## Df Sum of Sq RSS AIC  
## - SE.PRM.CUAT.ZS 1 5.95 735.78 267.76  
## - SP.POP.TOTL 1 6.70 736.53 267.94  
## <none> 729.82 268.33  
## - EG.ELC.ACCS.ZS 1 24.75 754.57 272.20  
## - EN.POP.DNST 1 59.02 788.85 280.02  
## - SH.XPD.CHEX.PC.CD 1 146.83 876.66 298.59  
## - SP.POP.GROW 1 151.03 880.86 299.43  
## - SP.DYN.CBRT.IN 1 228.68 958.51 314.30  
## - SP.DYN.IMRT.IN 1 517.78 1247.60 360.70  
##   
## Step: AIC=267.76  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN + SP.POP.GROW +   
## EN.POP.DNST + SP.POP.TOTL + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN  
##   
## Df Sum of Sq RSS AIC  
## - SP.POP.TOTL 1 5.10 740.87 266.97  
## <none> 735.78 267.76  
## - EG.ELC.ACCS.ZS 1 34.15 769.93 273.74  
## - EN.POP.DNST 1 57.96 793.74 279.10  
## - SH.XPD.CHEX.PC.CD 1 143.14 878.92 297.04  
## - SP.POP.GROW 1 153.00 888.78 299.01  
## - SP.DYN.CBRT.IN 1 224.98 960.75 312.71  
## - SP.DYN.IMRT.IN 1 514.26 1250.03 359.04  
##   
## Step: AIC=266.97  
## SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN + SP.POP.GROW +   
## EN.POP.DNST + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN  
##   
## Df Sum of Sq RSS AIC  
## <none> 740.87 266.97  
## - EG.ELC.ACCS.ZS 1 33.29 774.17 272.71  
## - EN.POP.DNST 1 59.62 800.49 278.59  
## - SH.XPD.CHEX.PC.CD 1 146.17 887.05 296.66  
## - SP.POP.GROW 1 151.55 892.43 297.73  
## - SP.DYN.CBRT.IN 1 221.23 962.11 310.96  
## - SP.DYN.IMRT.IN 1 522.27 1263.15 358.87

summary(motel\_step\_both)

##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN +   
## SP.POP.GROW + EN.POP.DNST + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN,   
## data = reg\_data\_train, x = T)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0534 -1.4002 0.0123 1.2206 4.7652   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.8858 0.1593 457.444 < 0.0000000000000002 \*\*\*  
## EG.ELC.ACCS.ZS 0.7861 0.2853 2.756 0.006498 \*\*   
## SP.DYN.IMRT.IN -3.9405 0.3610 -10.915 < 0.0000000000000002 \*\*\*  
## SP.POP.GROW 1.6445 0.2797 5.880 0.0000000214153 \*\*\*  
## EN.POP.DNST 0.5463 0.1481 3.688 0.000305 \*\*\*  
## SH.XPD.CHEX.PC.CD 0.9926 0.1719 5.774 0.0000000362023 \*\*\*  
## SP.DYN.CBRT.IN -3.1420 0.4423 -7.104 0.0000000000323 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.094 on 169 degrees of freedom  
## Multiple R-squared: 0.9198, Adjusted R-squared: 0.917   
## F-statistic: 323.1 on 6 and 169 DF, p-value: < 0.00000000000000022

# evaluation ===================================================================  
  
summary(full\_model)

##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ ., data = reg\_data\_train, x = T)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.7218 -1.2916 0.0493 1.1147 4.6302   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 72.444579 0.531540 136.292 < 0.0000000000000002  
## ContinentAsia 0.550566 0.690616 0.797 0.42654  
## ContinentAustralia/Oceania 0.354375 0.804502 0.440 0.66019  
## ContinentEurope 0.501314 0.818902 0.612 0.54131  
## ContinentNorth America 0.527580 0.740400 0.713 0.47718  
## ContinentSouth America 0.493159 0.929787 0.530 0.59659  
## EG.ELC.ACCS.ZS 0.523753 0.349958 1.497 0.13651  
## NY.ADJ.NNTY.KD.ZG 0.218981 0.217582 1.006 0.31577  
## SE.PRM.CUAT.ZS -0.199602 0.192643 -1.036 0.30175  
## SP.DYN.IMRT.IN -4.013476 0.413056 -9.717 < 0.0000000000000002  
## SP.POP.GROW 1.704596 0.305468 5.580 0.000000103780  
## EN.POP.DNST 0.563444 0.188614 2.987 0.00327  
## SP.POP.TOTL -0.171877 0.157211 -1.093 0.27595  
## SH.XPD.CHEX.PC.CD 0.940615 0.283948 3.313 0.00115  
## SH.XPD.CHEX.GD.ZS 0.072441 0.216391 0.335 0.73825  
## SL.UEM.TOTL.NE.ZS 0.002674 0.182955 0.015 0.98836  
## NY.GDP.MKTP.KD.ZG -0.161567 0.199079 -0.812 0.41827  
## NY.GDP.PCAP.CD -0.012298 0.308248 -0.040 0.96823  
## SP.DYN.CBRT.IN -3.311894 0.488884 -6.774 0.000000000241  
## SE.COM.DURS 0.133027 0.193127 0.689 0.49197  
##   
## (Intercept) \*\*\*  
## ContinentAsia   
## ContinentAustralia/Oceania   
## ContinentEurope   
## ContinentNorth America   
## ContinentSouth America   
## EG.ELC.ACCS.ZS   
## NY.ADJ.NNTY.KD.ZG   
## SE.PRM.CUAT.ZS   
## SP.DYN.IMRT.IN \*\*\*  
## SP.POP.GROW \*\*\*  
## EN.POP.DNST \*\*   
## SP.POP.TOTL   
## SH.XPD.CHEX.PC.CD \*\*   
## SH.XPD.CHEX.GD.ZS   
## SL.UEM.TOTL.NE.ZS   
## NY.GDP.MKTP.KD.ZG   
## NY.GDP.PCAP.CD   
## SP.DYN.CBRT.IN \*\*\*  
## SE.COM.DURS   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.144 on 156 degrees of freedom  
## Multiple R-squared: 0.9224, Adjusted R-squared: 0.9129   
## F-statistic: 97.57 on 19 and 156 DF, p-value: < 0.00000000000000022

summary(model\_full\_cp)

##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ Continent + EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN +   
## SP.POP.GROW + EN.POP.DNST + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN,   
## data = reg\_data\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.7288 -1.3028 -0.0025 1.2159 4.8613   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.4861 0.4938 146.783 < 0.0000000000000002 \*\*\*  
## ContinentAsia 0.4524 0.6546 0.691 0.490458   
## ContinentAustralia/Oceania 0.5026 0.7354 0.683 0.495286   
## ContinentEurope 0.4937 0.7826 0.631 0.529046   
## ContinentNorth America 0.6336 0.7149 0.886 0.376753   
## ContinentSouth America 0.7147 0.8895 0.803 0.422900   
## EG.ELC.ACCS.ZS 0.6640 0.3240 2.049 0.042023 \*   
## SP.DYN.IMRT.IN -3.9239 0.3681 -10.659 < 0.0000000000000002 \*\*\*  
## SP.POP.GROW 1.6760 0.2942 5.697 0.000000055155 \*\*\*  
## EN.POP.DNST 0.5520 0.1514 3.646 0.000357 \*\*\*  
## SH.XPD.CHEX.PC.CD 0.9810 0.1820 5.389 0.000000242716 \*\*\*  
## SP.DYN.CBRT.IN -3.1207 0.4593 -6.795 0.000000000191 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.119 on 164 degrees of freedom  
## Multiple R-squared: 0.9203, Adjusted R-squared: 0.9149   
## F-statistic: 172.1 on 11 and 164 DF, p-value: < 0.00000000000000022

summary(model\_step\_backward)

##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN +   
## SP.POP.GROW + EN.POP.DNST + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN,   
## data = reg\_data\_train, x = T)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0534 -1.4002 0.0123 1.2206 4.7652   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.8858 0.1593 457.444 < 0.0000000000000002 \*\*\*  
## EG.ELC.ACCS.ZS 0.7861 0.2853 2.756 0.006498 \*\*   
## SP.DYN.IMRT.IN -3.9405 0.3610 -10.915 < 0.0000000000000002 \*\*\*  
## SP.POP.GROW 1.6445 0.2797 5.880 0.0000000214153 \*\*\*  
## EN.POP.DNST 0.5463 0.1481 3.688 0.000305 \*\*\*  
## SH.XPD.CHEX.PC.CD 0.9926 0.1719 5.774 0.0000000362023 \*\*\*  
## SP.DYN.CBRT.IN -3.1420 0.4423 -7.104 0.0000000000323 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.094 on 169 degrees of freedom  
## Multiple R-squared: 0.9198, Adjusted R-squared: 0.917   
## F-statistic: 323.1 on 6 and 169 DF, p-value: < 0.00000000000000022

summary(model\_step\_forward)

##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ SP.DYN.IMRT.IN + SH.XPD.CHEX.PC.CD +   
## SP.DYN.CBRT.IN + SP.POP.GROW + EN.POP.DNST, data = reg\_data\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.746 -1.463 -0.008 1.151 4.750   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.8545 0.1620 449.771 < 0.0000000000000002 \*\*\*  
## SP.DYN.IMRT.IN -4.4010 0.3261 -13.494 < 0.0000000000000002 \*\*\*  
## SH.XPD.CHEX.PC.CD 0.8982 0.1717 5.232 0.000000489863798 \*\*\*  
## SP.DYN.CBRT.IN -3.4742 0.4337 -8.010 0.000000000000174 \*\*\*  
## SP.POP.GROW 1.6418 0.2851 5.759 0.000000038667821 \*\*\*  
## EN.POP.DNST 0.5365 0.1509 3.555 0.00049 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.134 on 170 degrees of freedom  
## Multiple R-squared: 0.9162, Adjusted R-squared: 0.9138   
## F-statistic: 371.8 on 5 and 170 DF, p-value: < 0.00000000000000022

summary(motel\_step\_both)

##   
## Call:  
## lm(formula = SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN +   
## SP.POP.GROW + EN.POP.DNST + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN,   
## data = reg\_data\_train, x = T)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.0534 -1.4002 0.0123 1.2206 4.7652   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 72.8858 0.1593 457.444 < 0.0000000000000002 \*\*\*  
## EG.ELC.ACCS.ZS 0.7861 0.2853 2.756 0.006498 \*\*   
## SP.DYN.IMRT.IN -3.9405 0.3610 -10.915 < 0.0000000000000002 \*\*\*  
## SP.POP.GROW 1.6445 0.2797 5.880 0.0000000214153 \*\*\*  
## EN.POP.DNST 0.5463 0.1481 3.688 0.000305 \*\*\*  
## SH.XPD.CHEX.PC.CD 0.9926 0.1719 5.774 0.0000000362023 \*\*\*  
## SP.DYN.CBRT.IN -3.1420 0.4423 -7.104 0.0000000000323 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.094 on 169 degrees of freedom  
## Multiple R-squared: 0.9198, Adjusted R-squared: 0.917   
## F-statistic: 323.1 on 6 and 169 DF, p-value: < 0.00000000000000022

anova(model\_full\_cp, full\_model)

## Analysis of Variance Table  
##   
## Model 1: SP.DYN.LE00.IN ~ Continent + EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN +   
## SP.POP.GROW + EN.POP.DNST + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN  
## Model 2: SP.DYN.LE00.IN ~ Continent + EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG +   
## SE.PRM.CUAT.ZS + SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST +   
## SP.POP.TOTL + SH.XPD.CHEX.PC.CD + SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS +   
## NY.GDP.MKTP.KD.ZG + NY.GDP.PCAP.CD + SP.DYN.CBRT.IN + SE.COM.DURS  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 164 736.49   
## 2 156 717.22 8 19.273 0.524 0.8372

anova(model\_step\_backward, full\_model)

## Analysis of Variance Table  
##   
## Model 1: SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN + SP.POP.GROW +   
## EN.POP.DNST + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN  
## Model 2: SP.DYN.LE00.IN ~ Continent + EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG +   
## SE.PRM.CUAT.ZS + SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST +   
## SP.POP.TOTL + SH.XPD.CHEX.PC.CD + SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS +   
## NY.GDP.MKTP.KD.ZG + NY.GDP.PCAP.CD + SP.DYN.CBRT.IN + SE.COM.DURS  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 169 740.87   
## 2 156 717.22 13 23.652 0.3957 0.9696

anova(model\_step\_forward, full\_model)

## Analysis of Variance Table  
##   
## Model 1: SP.DYN.LE00.IN ~ SP.DYN.IMRT.IN + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN +   
## SP.POP.GROW + EN.POP.DNST  
## Model 2: SP.DYN.LE00.IN ~ Continent + EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG +   
## SE.PRM.CUAT.ZS + SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST +   
## SP.POP.TOTL + SH.XPD.CHEX.PC.CD + SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS +   
## NY.GDP.MKTP.KD.ZG + NY.GDP.PCAP.CD + SP.DYN.CBRT.IN + SE.COM.DURS  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 170 774.17   
## 2 156 717.22 14 56.945 0.8847 0.5766

anova(motel\_step\_both, full\_model)

## Analysis of Variance Table  
##   
## Model 1: SP.DYN.LE00.IN ~ EG.ELC.ACCS.ZS + SP.DYN.IMRT.IN + SP.POP.GROW +   
## EN.POP.DNST + SH.XPD.CHEX.PC.CD + SP.DYN.CBRT.IN  
## Model 2: SP.DYN.LE00.IN ~ Continent + EG.ELC.ACCS.ZS + NY.ADJ.NNTY.KD.ZG +   
## SE.PRM.CUAT.ZS + SP.DYN.IMRT.IN + SP.POP.GROW + EN.POP.DNST +   
## SP.POP.TOTL + SH.XPD.CHEX.PC.CD + SH.XPD.CHEX.GD.ZS + SL.UEM.TOTL.NE.ZS +   
## NY.GDP.MKTP.KD.ZG + NY.GDP.PCAP.CD + SP.DYN.CBRT.IN + SE.COM.DURS  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 169 740.87   
## 2 156 717.22 13 23.652 0.3957 0.9696

reg\_data\_test$pred\_red\_SP.DYN.LE00.IN <- predict(model\_full\_cp, reg\_data\_test)  
reg\_data\_test$pred\_ful\_SP.DYN.LE00.IN <- predict(full\_model, reg\_data\_test)  
  
RMSE(reg\_data\_test$pred\_red\_SP.DYN.LE00.IN, reg\_data\_test$SP.DYN.LE00.IN)

## [1] 2.629916

# 1.472737  
R2(reg\_data\_test$pred\_red\_SP.DYN.LE00.IN, reg\_data\_test$SP.DYN.LE00.IN)

## [1] 0.8845503

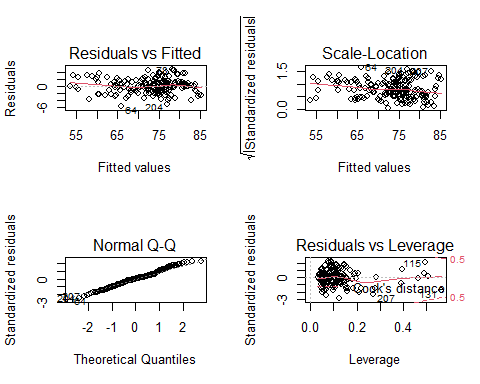
# 0.9570669  
  
RMSE(reg\_data\_test$pred\_ful\_SP.DYN.LE00.IN, reg\_data\_test$SP.DYN.LE00.IN)

## [1] 2.600094

# 1.482689  
R2(reg\_data\_test$pred\_ful\_SP.DYN.LE00.IN, reg\_data\_test$SP.DYN.LE00.IN)

## [1] 0.886316

# 0.9569355  
  
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page  
plot(full\_model)



# load libraries  
library(car)

## Loading required package: carData

library(stats)  
  
par(mfrow=c(1,1))  
  
# checking the data and dimensions   
head(final\_clean\_ds)

## Continent SP.DYN.LE00.IN EG.ELC.ACCS.ZS NY.ADJ.NNTY.KD.ZG  
## 1 Asia 64.83300 0.4597853 -0.087145738  
## 2 Europe 78.57300 0.5539554 -0.578535460  
## 3 Africa 76.88000 0.5334836 -0.162703224  
## 4 Australia/Oceania 68.19100 0.4025913 -0.745664111  
## 5 Europe 83.19756 0.5539554 -0.006125398  
## 6 Africa 61.14700 -1.6704932 -0.433141893  
## SE.PRM.CUAT.ZS SP.DYN.IMRT.IN SP.POP.GROW EN.POP.DNST SP.POP.TOTL  
## 1 -1.3204967 1.32246862 1.0273214 -0.19421671 0.01875497  
## 2 0.7939374 -0.64336894 -1.4819621 -0.17122858 -0.23546015  
## 3 -1.3204967 -0.05049729 0.6800355 -0.21434746 0.05495945  
## 4 0.7988553 0.02751213 -1.3381442 -0.08488569 -0.25568085  
## 5 0.7988553 -0.96060727 -0.9276740 -0.14119112 -0.25552311  
## 6 -1.7367281 1.50449061 1.8791524 -0.21061545 -0.02615628  
## SH.XPD.CHEX.PC.CD SH.XPD.CHEX.GD.ZS SL.UEM.TOTL.NE.ZS NY.GDP.MKTP.KD.ZG  
## 1 -0.5864102 2.19645792 -0.4712794 0.3405489  
## 2 -0.4675390 -0.91836345 0.6741893 -0.2160889  
## 3 -0.4871792 -0.11627315 -0.4268814 -0.5607544  
## 4 -0.1579418 1.03956043 1.3579186 -1.0213127  
## 5 0.8707142 0.03844599 -0.7394434 -0.2463859  
## 6 -0.5834072 -1.34222918 3.6932539 -1.0636722  
## NY.GDP.PCAP.CD SP.DYN.CBRT.IN SE.COM.DURS  
## 1 -0.6520948 1.2770042 -0.35500630  
## 2 -0.4756180 -0.7966202 -0.35500630  
## 3 -0.5262404 0.4325329 0.03120934  
## 4 -0.2480784 0.1669339 -1.51365323  
## 5 0.8026128 -1.2713078 0.03120934  
## 6 -0.5687276 2.1431549 -1.51365323

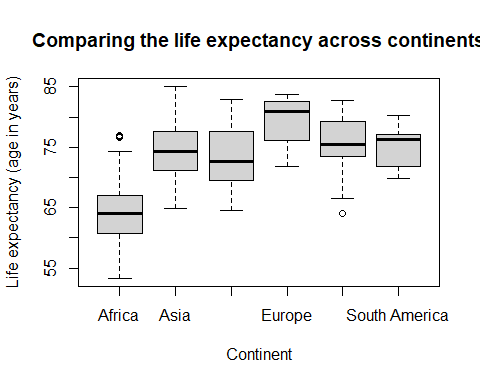
dim(final\_clean\_ds)

## [1] 217 16

# one-way ANOVA ( differences in average life expectancy across the continents)  
life\_expectancy <- final\_clean\_ds$SP.DYN.LE00.IN  
continent <- final\_clean\_ds$Continent  
  
# shows the average life expectancy across each continent  
group.means<-tapply(life\_expectancy, continent, mean)  
group.means

## Africa Asia Australia/Oceania Europe   
## 64.11014 74.61739 73.49023 79.34403   
## North America South America   
## 75.63260 75.09100

# boxplot to show the comparison of life expectancy within each continent  
boxplot(life\_expectancy~continent,main='Comparing the life expectancy across continents',  
 xlab='Continent', col="light gray", ylab = "Life expectancy (age in years)",)



# factor variables  
# final\_clean\_ds$Continent <- factor(final\_clean\_ds$Continent, levels=c("Africa","Asia","Australia/Oceania", "Europe", "North America", "South America"))  
  
anova1way<-aov(life\_expectancy~as.factor(continent),data=final\_clean\_ds)  
summary(anova1way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## as.factor(continent) 5 6619 1323.8 55.49 <0.0000000000000002 \*\*\*  
## Residuals 211 5034 23.9   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# reject the null hypothesis because our P value in our ANOVA table is significant at the 0.001 level.  
  
# post-hoc tests  
cat("Bonferroni post-hoc test","\n")

## Bonferroni post-hoc test

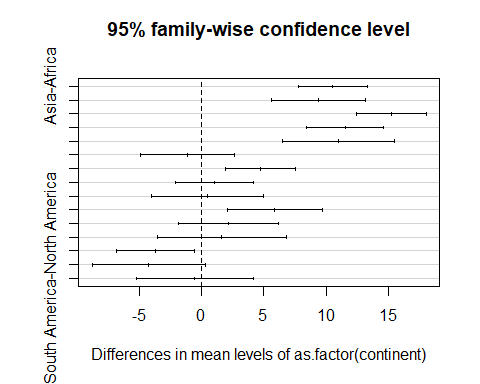
# Bonferroni post-hoc test  
pairwise.t.test(final\_clean\_ds$SP.DYN.LE00.IN, final\_clean\_ds$Continent, p.adj = "bonferroni")

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: final\_clean\_ds$SP.DYN.LE00.IN and final\_clean\_ds$Continent   
##   
## Africa Asia Australia/Oceania Europe   
## Asia < 0.0000000000000002 - - -   
## Australia/Oceania 0.00000000016 1.00000 - -   
## Europe < 0.0000000000000002 0.00004729625 0.00023 -   
## North America < 0.0000000000000002 1.00000 1.00000 0.01251  
## South America 0.00000000039 1.00000 1.00000 0.11311  
## North America  
## Asia -   
## Australia/Oceania -   
## Europe -   
## North America -   
## South America 1.00000   
##   
## P value adjustment method: bonferroni

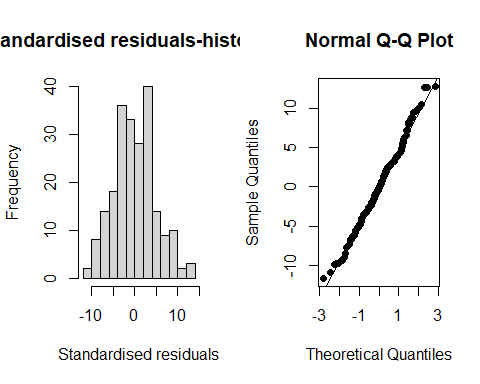
cat("\n","Tukey post-hoc test","\n")

##   
## Tukey post-hoc test

# Tukey post-hoc test  
tukey.continent<-TukeyHSD(anova1way)  
  
plot(tukey.continent)



# As seen, the North-America - Asia is outside of the mean  
  
# one-way ANOVA assumptions  
final\_clean\_ds$residuals1<-anova1way$residuals  
par(mfrow=c(1,2))  
  
# standardised residuals histogram  
hist(final\_clean\_ds$residuals1, main="Standardised residuals-histogram",xlab="Standardised residuals")  
  
# qq plot  
qqnorm(final\_clean\_ds$residuals1,pch=19)  
qqline(final\_clean\_ds$residuals1)



# Shapiro test  
shapiro.test(final\_clean\_ds$residuals1)

##   
## Shapiro-Wilk normality test  
##   
## data: final\_clean\_ds$residuals1  
## W = 0.99242, p-value = 0.3266

# Shapiro test shows the data is/n't normally distributed  
  
# Bartlett test  
bartlett.test(final\_clean\_ds$residuals1~as.numeric(final\_clean\_ds$Continent))

##   
## Bartlett test of homogeneity of variances  
##   
## data: final\_clean\_ds$residuals1 by as.numeric(final\_clean\_ds$Continent)  
## Bartlett's K-squared = 15.715, df = 5, p-value = 0.007706

# levene Test  
leveneTest(life\_expectancy~factor(final\_clean\_ds$Continent))

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 5 2.0757 0.06969 .  
## 211   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# welche one-way ANOVA  
continent.welchtest<-oneway.test(life\_expectancy~Continent,data=final\_clean\_ds)  
continent.welchtest

##   
## One-way analysis of means (not assuming equal variances)  
##   
## data: life\_expectancy and Continent  
## F = 48.824, num df = 5.000, denom df = 65.258, p-value <  
## 0.00000000000000022