DS6050.FinalProject.Bender

Alex Bender

March 24, 2019

Data Description: Countries of the World - This data comes from the US CIA. It is general characteristsics on different nations of the world. Some columns included in the data are Region, Population, Area (sq. mi.), Pop. Density (per sq. mi.), Coastline (coast/area ratio), Net migration, Arable (%), Crops (%), Other (%), Climate.

UN Human Development Data - This comes from the UN's 2015 Human Development report, which was used to calculate the Human Development Index. The datasets measures status of different nations in different metrics of human development. Some columns included in the data are Life Expectancy at Birth, Expected Years of Education, Mean Years of Education, Gross National Income (GNI) per Capita, GNI per Capita Rank Minus HDI Rank.

World Happiness - The World Happiness Report was released at the United Nations at an event celebrating International Day of Happiness on March 20th. The report continues to gain global recognition. Happiness Score is based on the World Happiness Report which includes GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust Government Corruption, etc.

I am going to explore world happiness(WH), human development(HDI), and country characteristics(CC) data in order to derive interesting insights. I will be operating on the assumption of using the HDI data from 2014, WH data from 2016, and CC data from as recent as 2017. Though the data is not all from the same time period, I will be assuming that the variation between the few years won't be significant enough to skew results significantly.

```
#load country characteristics data
cc_data <- read.csv("countries of the world.csv", dec=",")

#load 2016 world happiness data
wh2016 <- read.csv("happiness2016.csv")

#load 2014 HDI data
hdi2014 <- read.csv("human_development.csv", stringsAsFactors = FALSE)

#gather data understanding
str(cc_data)
```

```
## 'data.frame':    227 obs. of  20 variables:
##  $ Country                           : Factor w/ 227 levels "Afghanistan
",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ Region                            : Factor w/ 11 levels "ASIA (EX. NEAR
```

```
EAST)         ",..: 1 4 7 9 11 10 5 5 5 3 ...
##  $ Population                        : int  31056997 3581655 32930091
57794 71201 12127071 13477 69108 39921833 2976372 ...
##  $ Area..sq..mi..                    : int  647500 28748 2381740 199 468
1246700 102 443 2766890 29800 ...
##  $ Pop..Density..per.sq..mi..        : num  48 124.6 13.8 290.4 152.1 ...
##  $ Coastline..coast.area.ratio.      : num  0 1.26 0.04 58.29 0 ...
##  $ Net.migration                     : num  23.06 -4.93 -0.39 -20.71 6.6
...
##  $ Infant.mortality..per.1000.births.: num  163.07 21.52 31 9.27 4.05 ...
##  $ GDP....per.capita.                : int  700 4500 6000 8000 19000 1900
8600 11000 11200 3500 ...
##  $ Literacy....                      : num  36 86.5 70 97 100 42 95 89
97.1 98.6 ...
##  $ Phones..per.1000.                 : num  3.2 71.2 78.1 259.5 497.2 ...
##  $ Arable....                        : num  12.13 21.09 3.22 10 2.22 ...
##  $ Crops....                         : num  0.22 4.42 0.25 15 0 0.24 0
4.55 0.48 2.3 ...
##  $ Other....                         : num  87.7 74.5 96.5 75 97.8 ...
##  $ Climate                           : num  1 3 1 2 3 NA 2 2 3 4 ...
##  $ Birthrate                         : num  46.6 15.11 17.14 22.46 8.71
...
##  $ Deathrate                         : num  20.34 5.22 4.61 3.27 6.25 ...
##  $ Agriculture                       : num  0.38 0.232 0.101 NA NA 0.096
0.04 0.038 0.095 0.239 ...
##  $ Industry                          : num  0.24 0.188 0.6 NA NA 0.658
0.18 0.22 0.358 0.343 ...
##  $ Service                           : num  0.38 0.579 0.298 NA NA 0.246
0.78 0.743 0.547 0.418 ...
```

```r
summary(cc_data)
```

```
##              Country                                   Region
##  Afghanistan    :  1   SUB-SAHARAN AFRICA                 :51
##  Albania        :  1   LATIN AMER. & CARIB                :45
##  Algeria        :  1   ASIA (EX. NEAR EAST)               :28
##  American Samoa :  1   WESTERN EUROPE                     :28
##  Andorra        :  1   OCEANIA                            :21
##  Angola         :  1   NEAR EAST                          :16
##  (Other)        :221   (Other)                            :38
##    Population        Area..sq..mi..     Pop..Density..per.sq..mi..
##  Min.   :7.026e+03   Min.   :       2   Min.   :    0.00
##  1st Qu.:4.376e+05   1st Qu.:    4648   1st Qu.:   29.15
##  Median :4.787e+06   Median :   86600   Median :   78.80
##  Mean   :2.874e+07   Mean   :  598227   Mean   :  379.05
##  3rd Qu.:1.750e+07   3rd Qu.:  441811   3rd Qu.:  190.15
##  Max.   :1.314e+09   Max.   :17075200   Max.   :16271.50
##
##  Coastline..coast.area.ratio. Net.migration
##  Min.   :  0.00               Min.   :-20.99000
```

```
##  1st Qu.:  0.10              1st Qu.: -0.92750
##  Median :  0.73              Median :  0.00000
##  Mean   : 21.17              Mean   :  0.03812
##  3rd Qu.: 10.35              3rd Qu.:  0.99750
##  Max.   :870.66              Max.   : 23.06000
##                              NA's   :3
##  Infant.mortality..per.1000.births. GDP....per.capita.  Literacy....
##  Min.   :  2.29                     Min.   :  500      Min.   : 17.60
##  1st Qu.:  8.15                     1st Qu.: 1900      1st Qu.: 70.60
##  Median : 21.00                     Median : 5550      Median : 92.50
##  Mean   : 35.51                     Mean   : 9690      Mean   : 82.84
##  3rd Qu.: 55.70                     3rd Qu.:15700      3rd Qu.: 98.00
##  Max.   :191.19                     Max.   :55100      Max.   :100.00
##  NA's   :3                          NA's   :1          NA's   :18
##  Phones..per.1000.   Arable....      Crops....        Other....
##  Min.   :   0.2    Min.   : 0.00   Min.   : 0.000   Min.   : 33.33
##  1st Qu.:  37.8    1st Qu.: 3.22   1st Qu.: 0.190   1st Qu.: 71.65
##  Median : 176.2    Median :10.42   Median : 1.030   Median : 85.70
##  Mean   : 236.1    Mean   :13.80   Mean   : 4.564   Mean   : 81.64
##  3rd Qu.: 389.6    3rd Qu.:20.00   3rd Qu.: 4.440   3rd Qu.: 95.44
##  Max.   :1035.6    Max.   :62.11   Max.   :50.680   Max.   :100.00
##  NA's   :4         NA's   :2       NA's   :2        NA's   :2
##     Climate        Birthrate       Deathrate       Agriculture
##  Min.   :1.000   Min.   : 7.29   Min.   : 2.290   Min.   :0.00000
##  1st Qu.:2.000   1st Qu.:12.67   1st Qu.: 5.910   1st Qu.:0.03775
##  Median :2.000   Median :18.79   Median : 7.840   Median :0.09900
##  Mean   :2.139   Mean   :22.11   Mean   : 9.241   Mean   :0.15084
##  3rd Qu.:3.000   3rd Qu.:29.82   3rd Qu.:10.605   3rd Qu.:0.22100
##  Max.   :4.000   Max.   :50.73   Max.   :29.740   Max.   :0.76900
##  NA's   :22      NA's   :3       NA's   :4        NA's   :15
##     Industry         Service
##  Min.   :0.0200   Min.   :0.0620
##  1st Qu.:0.1930   1st Qu.:0.4293
##  Median :0.2720   Median :0.5710
##  Mean   :0.2827   Mean   :0.5653
##  3rd Qu.:0.3410   3rd Qu.:0.6785
##  Max.   :0.9060   Max.   :0.9540
##  NA's   :16       NA's   :15
```

```r
head(cc_data)
```

```
##            Country                               Region Population
## 1     Afghanistan         ASIA (EX. NEAR EAST)             31056997
## 2         Albania  EASTERN EUROPE                           3581655
## 3         Algeria  NORTHERN AFRICA                         32930091
## 4 American Samoa  OCEANIA                                     57794
## 5         Andorra  WESTERN EUROPE                             71201
## 6          Angola  SUB-SAHARAN AFRICA                      12127071
##   Area..sq..mi.. Pop..Density..per.sq..mi.. Coastline..coast.area.ratio.
## 1         647500                       48.0                         0.00
```

```
## 2          28748                      124.6                         1.26
## 3        2381740                       13.8                         0.04
## 4            199                      290.4                        58.29
## 5            468                      152.1                         0.00
## 6        1246700                        9.7                         0.13
##   Net.migration Infant.mortality..per.1000.births. GDP....per.capita.
## 1         23.06                              163.07                700
## 2         -4.93                               21.52               4500
## 3         -0.39                               31.00               6000
## 4        -20.71                                9.27               8000
## 5          6.60                                4.05              19000
## 6          0.00                              191.19               1900
##   Literacy.... Phones..per.1000. Arable.... Crops.... Other.... Climate
## 1         36.0               3.2      12.13      0.22     87.65       1
## 2         86.5              71.2      21.09      4.42     74.49       3
## 3         70.0              78.1       3.22      0.25     96.53       1
## 4         97.0             259.5      10.00     15.00     75.00       2
## 5        100.0             497.2       2.22      0.00     97.78       3
## 6         42.0               7.8       2.41      0.24     97.35      NA
##   Birthrate Deathrate Agriculture Industry Service
## 1     46.60     20.34       0.380    0.240   0.380
## 2     15.11      5.22       0.232    0.188   0.579
## 3     17.14      4.61       0.101    0.600   0.298
## 4     22.46      3.27          NA       NA      NA
## 5      8.71      6.25          NA       NA      NA
## 6     45.11     24.20       0.096    0.658   0.246
```

```
#gather data understanding
str(hdi2014)
```

```
## 'data.frame':    195 obs. of  8 variables:
##  $ HDI.Rank                             : int  1 2 3 4 5 6 6 8 9 9 ...
##  $ Country                              : chr  "Norway" "Australia"
"Switzerland" "Denmark" ...
##  $ Human.Development.Index..HDI.        : num  0.944 0.935 0.93 0.923
0.922 0.916 0.916 0.915 0.913 0.913 ...
##  $ Life.Expectancy.at.Birth             : num  81.6 82.4 83 80.2 81.6
80.9 80.9 79.1 82 81.8 ...
##  $ Expected.Years.of.Education          : num  17.5 20.2 15.8 18.7 17.9
16.5 18.6 16.5 15.9 19.2 ...
##  $ Mean.Years.of.Education              : num  12.6 13 12.8 12.7 11.9
13.1 12.2 12.9 13 12.5 ...
##  $ Gross.National.Income..GNI..per.Capita: chr  "64,992" "42,261" "56,431"
"44,025" ...
##  $ GNI.per.Capita.Rank.Minus.HDI.Rank    : int  5 17 6 11 9 11 16 3 11 23
...
```

```
summary(hdi2014)
```

```
##     HDI.Rank        Country           Human.Development.Index..HDI.
##  Min.   :  1.00   Length:195         Min.   :0.3480
```

```
##  1st Qu.: 47.75   Class :character   1st Qu.:0.5770
##  Median : 94.00   Mode  :character   Median :0.7210
##  Mean   : 94.31                      Mean   :0.6918
##  3rd Qu.:141.25                      3rd Qu.:0.8000
##  Max.   :188.00                      Max.   :0.9440
##  NA's   :7
##  Life.Expectancy.at.Birth Expected.Years.of.Education
##  Min.   :49.00            Min.   : 4.10
##  1st Qu.:65.75            1st Qu.:11.10
##  Median :73.10            Median :13.10
##  Mean   :71.07            Mean   :12.86
##  3rd Qu.:76.80            3rd Qu.:14.90
##  Max.   :84.00            Max.   :20.20
##
##  Mean.Years.of.Education Gross.National.Income..GNI..per.Capita
##  Min.   : 1.400          Length:195
##  1st Qu.: 5.550          Class :character
##  Median : 8.400          Mode  :character
##  Mean   : 8.079
##  3rd Qu.:10.600
##  Max.   :13.100
##
##  GNI.per.Capita.Rank.Minus.HDI.Rank
##  Min.   :-84.0000
##  1st Qu.: -9.0000
##  Median :  1.5000
##  Mean   :  0.1862
##  3rd Qu.: 11.0000
##  Max.   : 47.0000
##  NA's   :7
```

```r
head(hdi2014)
```

```
##   HDI.Rank     Country Human.Development.Index..HDI.
## 1        1      Norway                         0.944
## 2        2   Australia                         0.935
## 3        3 Switzerland                         0.930
## 4        4     Denmark                         0.923
## 5        5 Netherlands                         0.922
## 6        6     Germany                         0.916
##   Life.Expectancy.at.Birth Expected.Years.of.Education
## 1                     81.6                        17.5
## 2                     82.4                        20.2
## 3                     83.0                        15.8
## 4                     80.2                        18.7
## 5                     81.6                        17.9
## 6                     80.9                        16.5
##   Mean.Years.of.Education Gross.National.Income..GNI..per.Capita
## 1                    12.6                                 64,992
## 2                    13.0                                 42,261
```

```
## 3                    12.8                                    56,431
## 4                    12.7                                    44,025
## 5                    11.9                                    45,435
## 6                    13.1                                    43,919
##   GNI.per.Capita.Rank.Minus.HDI.Rank
## 1                                  5
## 2                                 17
## 3                                  6
## 4                                 11
## 5                                  9
## 6                                 11
```

```
#gather data understanding
str(wh2016)
```

```
## 'data.frame':    157 obs. of  13 variables:
##  $ Country                      : Factor w/ 157 levels "Afghanistan",..: 
38 135 58 104 45 26 98 99 7 134 ...
##  $ Region                       : Factor w/ 10 levels "Australia and New 
Zealand",..: 10 10 10 10 10 6 10 1 1 10 ...
##  $ Happiness.Rank               : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Happiness.Score              : num  7.53 7.51 7.5 7.5 7.41 ...
##  $ Lower.Confidence.Interval    : num  7.46 7.43 7.33 7.42 7.35 ...
##  $ Upper.Confidence.Interval    : num  7.59 7.59 7.67 7.58 7.47 ...
##  $ Economy..GDP.per.Capita.     : num  1.44 1.53 1.43 1.58 1.41 ...
##  $ Family                       : num  1.16 1.15 1.18 1.13 1.13 ...
##  $ Health..Life.Expectancy.     : num  0.795 0.863 0.867 0.796 0.811 ...
##  $ Freedom                      : num  0.579 0.586 0.566 0.596 0.571 ...
##  $ Trust..Government.Corruption.: num  0.445 0.412 0.15 0.358 0.41 ...
##  $ Generosity                   : num  0.362 0.281 0.477 0.379 0.255 ...
##  $ Dystopia.Residual            : num  2.74 2.69 2.83 2.66 2.83 ...
```

```
summary(wh2016)
```

```
##         Country                                Region   Happiness.Rank
##  Afghanistan:  1   Sub-Saharan Africa             :38   Min.   :  1.00
##  Albania    :  1   Central and Eastern Europe     :29   1st Qu.: 40.00
##  Algeria    :  1   Latin America and Caribbean    :24   Median : 79.00
##  Angola     :  1   Western Europe                 :21   Mean   : 78.98
##  Argentina  :  1   Middle East and Northern Africa:19   3rd Qu.:118.00
##  Armenia    :  1   Southeastern Asia              : 9   Max.   :157.00
##  (Other)    :151   (Other)                        :17
##  Happiness.Score Lower.Confidence.Interval Upper.Confidence.Interval
##  Min.   :2.905   Min.   :2.732             Min.   :3.078
##  1st Qu.:4.404   1st Qu.:4.327             1st Qu.:4.465
##  Median :5.314   Median :5.237             Median :5.419
##  Mean   :5.382   Mean   :5.282             Mean   :5.482
##  3rd Qu.:6.269   3rd Qu.:6.154             3rd Qu.:6.434
##  Max.   :7.526   Max.   :7.460             Max.   :7.669
##
##  Economy..GDP.per.Capita.     Family       Health..Life.Expectancy.
```

```
##  Min.   :0.0000          Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.6702          1st Qu.:0.6418   1st Qu.:0.3829
##  Median :1.0278          Median :0.8414   Median :0.5966
##  Mean   :0.9539          Mean   :0.7936   Mean   :0.5576
##  3rd Qu.:1.2796          3rd Qu.:1.0215   3rd Qu.:0.7299
##  Max.   :1.8243          Max.   :1.1833   Max.   :0.9528
##
##     Freedom       Trust..Government.Corruption.   Generosity
##  Min.   :0.0000   Min.   :0.00000              Min.   :0.0000
##  1st Qu.:0.2575   1st Qu.:0.06126              1st Qu.:0.1546
##  Median :0.3975   Median :0.10547              Median :0.2225
##  Mean   :0.3710   Mean   :0.13762              Mean   :0.2426
##  3rd Qu.:0.4845   3rd Qu.:0.17554              3rd Qu.:0.3119
##  Max.   :0.6085   Max.   :0.50521              Max.   :0.8197
##
##  Dystopia.Residual
##  Min.   :0.8179
##  1st Qu.:2.0317
##  Median :2.2907
##  Mean   :2.3258
##  3rd Qu.:2.6646
##  Max.   :3.8377
##
```

```
head(wh2016)
```

```
##       Country          Region Happiness.Rank Happiness.Score
## 1     Denmark Western Europe              1           7.526
## 2 Switzerland Western Europe              2           7.509
## 3     Iceland Western Europe              3           7.501
## 4      Norway Western Europe              4           7.498
## 5     Finland Western Europe              5           7.413
## 6      Canada  North America              6           7.404
##   Lower.Confidence.Interval Upper.Confidence.Interval
## 1                     7.460                     7.592
## 2                     7.428                     7.590
## 3                     7.333                     7.669
## 4                     7.421                     7.575
## 5                     7.351                     7.475
## 6                     7.335                     7.473
##   Economy..GDP.per.Capita.  Family Health..Life.Expectancy. Freedom
## 1                  1.44178 1.16374                  0.79504 0.57941
## 2                  1.52733 1.14524                  0.86303 0.58557
## 3                  1.42666 1.18326                  0.86733 0.56624
## 4                  1.57744 1.12690                  0.79579 0.59609
## 5                  1.40598 1.13464                  0.81091 0.57104
## 6                  1.44015 1.09610                  0.82760 0.57370
##   Trust..Government.Corruption. Generosity Dystopia.Residual
## 1                       0.44453    0.36171           2.73939
## 2                       0.41203    0.28083           2.69463
```

```
## 3                           0.14975    0.47678            2.83137
## 4                           0.35776    0.37895            2.66465
## 5                           0.41004    0.25492            2.82596
## 6                           0.31329    0.44834            2.70485
```

As you can see all of the loaded data has different numbers of rows, meaning that different countries are included in the different datasets. This will be reconcilied by only including the countries located in the dataset with the least amount of countries (world happiness 2016). Since my analysis relies on merging the unqiue nation metrics from the various datasets, it would be wise to only included the countries with full data. Still, this leaves us with over 150 countries, which is enough to run both numeric prediction (regression) and classification data mining tasks.The data set is a similar size to the built in Iris dataset, with more dimensions, so it should still be fine. I will combat this small amount of data with the use of k-fold Cross-validation.

My plan of attack: - I am going to use the 2016 World Happiness data as my basis for dependent variables. - I plan to do numeric prediction/regression by predicting the happiness score of a nation using the happiness score column from the wh2016 data. I will be exploring multiple linear regression, regression tree, neural network, and kNN models. - As my independent variables I will be using the Human Development Index and Country Characteristics data in order to attempt to predict the target/response variables from the world happiness data. - I won’t be using the additional features in the World Happiness data as predictors for two reasons. 1) The world happiness score is a direct calculation from these features, so I don’t want to risk overfitting by the model just memorizing the calculation essentially. 2) I want to derive novel insights from a various set of predictors from the other two datasets.

If I have time: - Classification for region of the world based on the features - Derive an attribute that is a binary indicator of happy or not in order to do binary classification using SVM and/or Logistic regression.

Now we have to merge the datasets properly. First we have to match up the rows based on Country name. If different datasets name countries differently, this will pose a challenge.

```
library(tidyverse)

## -- Attaching packages ---------------------------------------------------
----------------- tidyverse 1.2.1 --

## v ggplot2 3.1.0       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------
----------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

To reduce chances for errors down the line, let's trim any additional whitespaces from the data.

```
#Get rid of unnecessary whitespace
cc_data$Country <- str_trim(cc_data$Country) %>% as.factor()
cc_data$Region <- str_trim(cc_data$Region) %>% as.factor()
hdi2014$Country <- str_trim(hdi2014$Country) %>% as.factor()
wh2016$Country <- str_trim(wh2016$Country) %>% as.factor()
wh2016$Region <- str_trim(wh2016$Region) %>% as.factor()
```

Let's only select the two target/response variables and the country from the world happiness data.

```
#select only the needed columns
wh2016 <- select(wh2016, Country, Region, Happiness.Score)
```

We know from this dataset that we will be looking at 157 countries.

I am going to standardize all of the country names using a convenient function in the standardize text package. The function recognizes common variations of country names and essentially "Autocorrects" them into a standard format. This will help a lot when joining datasets together.

```
#install.packages("StandardizeText")
library(StandardizeText)

#Standardize column using default country names
hdi2014$Country <- standardize.countrynames(hdi2014$Country,suggest="auto",
verbose = T)

##
## The following names were not recoginized and left unchanged:
## [1] "Arab States"                      "CÃ´te d'Ivoire"
## [3] "Cabo Verde"                       "East Asia and the Pacific"
## [5] "Europe and Central Asia"          "Latin America and the Caribbean"
## [7] "South Asia"                       "Sub-Saharan Africa"
##
## The following names were changed:
##                                    Original
## 1           Bolivia (Plurinational State of)
## 2                                      Congo
## 3         Congo (Democratic Republic of the)
## 4                  Iran (Islamic Republic of)
## 5                         Korea (Republic of)
## 6           Lao People's Democratic Republic
## 7           Micronesia (Federated States of)
## 8                       Moldova (Republic of)
## 9                         Palestine, State of
## 10                      Saint Kitts and Nevis
## 11                                Saint Lucia
## 12          Saint Vincent and the Grenadines
```

```
## 13                                  Slovakia
## 14             Tanzania (United Republic of)
## 15 The former Yugoslav Republic of Macedonia
## 16        Venezuela (Bolivarian Republic of)
## 17                                  Viet Nam
##                         Modified
## 1                        Bolivia
## 2                 Congo Republic
## 3       Congo Democratic Republic
## 4                           Iran
## 5                 Korea Republic
## 6                           Laos
## 7                     Micronesia
## 8                        Moldova
## 9           Palestinian Territory
## 10            St. Kitts and Nevis
## 11                     St. Lucia
## 12 St. Vincent and the Grenadines
## 13               Slovak Republic
## 14                      Tanzania
## 15                     Macedonia
## 16                     Venezuela
## 17                       Vietnam
##
## The following suggested changes were applied:
##                 Original Suggested
## 1 Hong Kong, China (SAR) Hong Kong
```

I can see here that some useful changes were made! Also it seems there was an encoding error for "CÃ´te d'Ivoire", so I will manually change this to "Cote d'Ivoire" in the hdi2014 dataset.

```
#Replace mistake manually
hdi2014$Country[which(hdi2014$Country=="CÃ´te d'Ivoire")] <- "Cote d'Ivoire"

#Make sure it worked
hdi2014$Country[which(hdi2014$Country=="CÃ´te d'Ivoire")]

## character(0)

hdi2014$Country[which(hdi2014$Country=="Cote d'Ivoire")]

## [1] "Cote d'Ivoire"

#Standardize column using default country names
wh2016$Country <- standardize.countrynames(wh2016$Country,suggest="auto",
verbose = T)

##
## The following names were not recoginized and left unchanged:
## [1] "North Cyprus"      "Somaliland Region"
```

```
##
## The following names were changed:
##                  Original                  Modified
## 1     Congo (Brazzaville)           Congo Republic
## 2        Congo (Kinshasa) Congo Democratic Republic
## 3             Ivory Coast            Cote d'Ivoire
## 4 Palestinian Territories     Palestinian Territory
## 5                  Russia        Russian Federation
## 6                Slovakia           Slovak Republic
## 7             South Korea            Korea Republic
## 8                   Syria      Syrian Arab Republic
```

This data had 8 names that required standardization! Also, some names weren't recognized, but we will see if we need those.

```
#Standardize column using default country names
cc_data$Country <- standardize.countrynames(cc_data$Country,suggest="auto")

##
## Note: 3 names were not recoginized and left unchanged.
##
## The following names were changed:
##                            Original                          Modified
## 1                 Antigua & Barbuda               Antigua and Barbuda
## 2                      Bahamas, The                           Bahamas
## 3                Bosnia & Herzegovina            Bosnia and Herzegovina
## 4               British Virgin Is.                 Virgin Islands BR
## 5                            Brunei                 Brunei Darussalam
## 6              Central African Rep.          Central African Republic
## 7                   Congo, Dem. Rep.         Congo Democratic Republic
## 8                       East Timor                       Timor-Leste
## 9                       Gambia, The                            Gambia
## 10                     Korea, North         Korea Democratic Republic
## 11                     Korea, South                    Korea Republic
## 12                            Macau                             Macao
## 13                           Russia                Russian Federation
## 14                     Saint Helena                         St. Helena
## 15               Saint Kitts & Nevis               St. Kitts and Nevis
## 16                      Saint Lucia                          St. Lucia
## 17 Saint Vincent and the Grenadines St. Vincent and the Grenadines
## 18               Sao Tome & Principe              Sao Tome and Principe
## 19                         Slovakia                   Slovak Republic
## 20             St Pierre & Miquelon           St. Pierre and Miquelon
## 21                            Syria              Syrian Arab Republic
## 22                Trinidad & Tobago               Trinidad and Tobago
## 23                Turks & Caicos Is          Turks and Caicos Islands
##
## The following suggested changes were applied:
##               Original      Suggested
```

```
## 1 Congo, Repub. of the Congo Republic
## 2 Micronesia, Fed. St.     Micronesia
```

Perfect! Now all country names should be standardized, so we can do a join.Let's see which countries from the world happiness data do not have a match in the human development data using an anti-join.

```
#make sure the country name standardization worked
non_matches <- anti_join(wh2016[1], hdi2014[2], by = "Country")
non_matches2 <- anti_join(wh2016[1], cc_data[1] , by = "Country")

#print out distinct non-matches
distinct(rbind(non_matches,non_matches2))

##                     Country
## 1             Puerto Rico
## 2                  Taiwan
## 3            North Cyprus
## 4                 Somalia
## 5                  Kosovo
## 6        Somaliland Region
## 7              Montenegro
## 8   Palestinian Territory
## 9                 Myanmar
## 10            South Sudan
```

The countries that have a happiness score but do not have any data in either the HDI data or the CC data are Puerto Rico, Taiwan, North Cyprus, Somalia, Kosovo, Somaliland Region, Montenegro, Palestinian Territory, Myanmar, and South Sudan. Because these countries don't have any data from the other datasets for the dependent/response/target variable, they would essentially be complete empty rows. From the eventual merged data, I will be removing these countries for the analysis. This will leave us with 147 countries to analyze, which is still enough to draw meaningful insights.

In order to merge the datasets, I am going to do an inner join, which will essentially pull all the data that both datasets have based on matching the "Country" column. For example, a sample row will contain the data columns located in the all three datasets for a single country name, such as France.

```
#merge datasets to make final dataset
world_df <- inner_join(wh2016, hdi2014, by = "Country")

world_df <- inner_join(world_df, cc_data, by = "Country")

nrow(world_df)

## [1] 147
```

147 countries remaining, perfect!

Let's make sure the join worked correctly by spot checking a country

```
cc_data[which(cc_data$Country=="France"),"Population"]

## [1] 60876136

hdi2014[which(hdi2014$Country=="France"),"Mean.Years.of.Education"]

## [1] 11.1

wh2016[which(wh2016$Country=="France"),"Happiness.Score"]

## [1] 6.478

world_df[which(world_df$Country=="France"),c("Population",
"Mean.Years.of.Education", "Happiness.Score")]

##    Population Mean.Years.of.Education Happiness.Score
## 31   60876136                    11.1           6.478
```

Perfect, they match!

Now let's explore the data!

```
str(world_df)

## 'data.frame':    147 obs. of  29 variables:
##  $ Country                             : chr  "Denmark" "Switzerland"
"Iceland" "Norway" ...
##  $ Region.x                            : Factor w/ 10 levels "Australia
and New Zealand",..: 10 10 10 10 10 6 10 1 1 10 ...
##  $ Happiness.Score                     : num  7.53 7.51 7.5 7.5 7.41 ...
##  $ HDI.Rank                            : int  4 3 16 1 24 9 5 9 2 14 ...
##  $ Human.Development.Index..HDI.       : num  0.923 0.93 0.899 0.944
0.883 0.913 0.922 0.913 0.935 0.907 ...
##  $ Life.Expectancy.at.Birth            : num  80.2 83 82.6 81.6 80.8 82
81.6 81.8 82.4 82.2 ...
##  $ Expected.Years.of.Education         : num  18.7 15.8 19 17.5 17.1
15.9 17.9 19.2 20.2 15.8 ...
##  $ Mean.Years.of.Education             : num  12.7 12.8 10.6 12.6 10.3
13 11.9 12.5 13 12.1 ...
##  $ Gross.National.Income..GNI..per.Capita: chr  "44,025" "56,431" "35,182"
"64,992" ...
##  $ GNI.per.Capita.Rank.Minus.HDI.Rank  : int  11 6 12 5 0 11 9 23 17 -1
...
##  $ Region.y                            : Factor w/ 11 levels "ASIA (EX.
NEAR EAST)",..: 11 11 11 11 11 8 11 9 9 11 ...
##  $ Population                          : int  5450661 7523934 299388
4610820 5231372 33098932 16491461 4076140 20264082 9016596 ...
##  $ Area..sq..mi..                      : int  43094 41290 103000 323802
338145 9984670 41526 268680 7686850 449964 ...
##  $ Pop..Density..per.sq..mi..          : num  126.5 182.2 2.9 14.2 15.5
```

```
...
##  $ Coastline..coast.area.ratio.          : num  16.97 0 4.83 7.77 0.37 ...
##  $ Net.migration                         : num  2.48 4.05 2.38 1.74 0.95
5.96 2.91 4.05 3.98 1.67 ...
##  $ Infant.mortality..per.1000.births.    : num  4.56 4.39 3.31 3.7 3.57
4.75 5.04 5.85 4.69 2.77 ...
##  $ GDP....per.capita.                    : int  31100 32700 30900 37800
27400 29800 28600 21600 29000 26800 ...
##  $ Literacy....                          : num  100 99 99.9 100 100 97 99
99 100 99 ...
##  $ Phones..per.1000.                     : num  615 681 648 462 405 ...
##  $ Arable....                            : num  54.02 10.42 0.07 2.87 7.19
...
##  $ Crops....                             : num  0.19 0.61 0 0 0.03 0.02
0.97 6.99 0.04 0.01 ...
##  $ Other....                             : num  45.8 89 99.9 97.1 92.8 ...
##  $ Climate                               : num  3 3 3 3 3 NA 3 3 1 3 ...
##  $ Birthrate                             : num  11.13 9.71 13.64 11.46
10.45 ...
##  $ Deathrate                             : num  10.36 8.49 6.72 9.4 9.86
...
##  $ Agriculture                           : num  0.018 0.015 0.086 0.021
0.028 0.022 0.021 0.043 0.038 0.011 ...
##  $ Industry                              : num  0.246 0.34 0.15 0.415
0.295 0.294 0.244 0.273 0.262 0.282 ...
##  $ Service                               : num  0.735 0.645 0.765 0.564
0.676 0.684 0.736 0.684 0.7 0.707 ...
```

Here we see some features that will require attention. 1) there are two region columns from 2 different datasets. We will only use one of these. I am going to use the regions from the World Happiness data and remove the other column. 2) HDI rank acts as a row number in the hdi data and doesn't add information beyond the HDI score, so we will remove the rank column. 3) Since the rank column won't be used I will also remove the "GNI.per.Capita.Rank.Minus.HDI.Rank" columns since it relies on rank and I could not find an explanation of what this column means. 4) The punctuation located within the feature names got coerced into "." periods, so I will be renaming some columns to make them more readable. 5) The gross national income columns is currently a character instead of a numeric.

```
#get rid of duplicate region column, HDI rank column,
GNI.per.Capita.Rank.Minus.HDI.Rank column
world_df <-select(world_df, -Region.y)
world_df <- select(world_df, -HDI.Rank)
world_df <- select(world_df, -GNI.per.Capita.Rank.Minus.HDI.Rank)
```

Now rename columns for readability.

```
world_df <- rename(world_df, Region = "Region.x", HDI.Score =
"Human.Development.Index..HDI.", Gross.National.Income.per.Capita
="Gross.National.Income..GNI..per.Capita", Area.sq.mi =
```

```
"Area..sq..mi..",Pop.Density.per.sq.mi =
"Pop..Density..per.sq..mi..",Coast.Area.Ratio=
"Coastline..coast.area.ratio.", Infant.Mortality.per.1000.births=
"Infant.mortality..per.1000.births.", GDP.per.capita = "GDP....per.capita.",
Literacy.percent ="Literacy....", Phones.per.1000.people
="Phones..per.1000.", Arable.percent="Arable....", Crops.percent =
"Crops....", Other.Land.Use.percent= "Other....")
```

Now change Gross National Income (GNI) into a numeric using regex to recognize the

```
world_df$Gross.National.Income.per.Capita <- as.numeric(gsub(",", "",
world_df$Gross.National.Income.per.Capita))
```

Now let's explore the data some more.

```
anyNA(world_df)

## [1] TRUE
```

There are NA values, so let's try to handle these.

```
summary(world_df)

##    Country                                           Region    Happiness.Score
##  Length:147         Sub-Saharan Africa                :35   Min.   :2.905
##  Class :character   Central and Eastern Europe        :27   1st Qu.:4.383
##  Mode  :character   Latin America and Caribbean       :23   Median :5.314
##                     Western Europe                    :20   Mean   :5.386
##                     Middle East and Northern Africa:18   3rd Qu.:6.296
##                     Southeastern Asia                 : 8   Max.   :7.526
##                     (Other)                           :16
##    HDI.Score      Life.Expectancy.at.Birth Expected.Years.of.Education
##  Min.   :0.3480   Min.   :50.90            Min.   : 5.40
##  1st Qu.:0.5885   1st Qu.:65.90            1st Qu.:11.25
##  Median :0.7330   Median :74.00            Median :13.50
##  Mean   :0.7056   Mean   :71.75            Mean   :13.19
##  3rd Qu.:0.8355   3rd Qu.:77.50            3rd Qu.:15.25
##  Max.   :0.9440   Max.   :84.00            Max.   :20.20
##
##  Mean.Years.of.Education Gross.National.Income.per.Capita
##  Min.   : 1.400          Min.   :   680
##  1st Qu.: 6.000          1st Qu.:  4198
##  Median : 8.500          Median : 12122
##  Mean   : 8.291          Mean   : 18226
##  3rd Qu.:10.900          3rd Qu.: 25486
##  Max.   :13.100          Max.   :123124
##
##    Population          Area.sq.mi        Pop.Density.per.sq.mi
##  Min.   :2.877e+05   Min.   :     316   Min.   :   1.8
##  1st Qu.:4.493e+06   1st Qu.:   64894   1st Qu.:  27.1
##  Median :1.018e+07   Median :  236800   Median :  66.9
##  Mean   :4.317e+07   Mean   :  877074   Mean   : 206.3
```

```
##  3rd Qu.:2.968e+07   3rd Qu.:  700057   3rd Qu.: 127.2
##  Max.   :1.314e+09   Max.   :17075200   Max.   :6482.2
##
##  Coast.Area.Ratio Net.migration      Infant.Mortality.per.1000.births
##  Min.   : 0.000   Min.   :-10.8300   Min.   :  2.290
##  1st Qu.: 0.005   1st Qu.: -0.7750   1st Qu.:  8.685
##  Median : 0.240   Median :  0.0000   Median : 24.600
##  Mean   : 2.665   Mean   :  0.2145   Mean   : 39.395
##  3rd Qu.: 1.365   3rd Qu.:  0.6700   3rd Qu.: 64.605
##  Max.   :67.120   Max.   : 23.0600   Max.   :191.190
##
##  GDP.per.capita  Literacy.percent Phones.per.1000.people Arable.percent
##  Min.   :  500   Min.   : 17.60   Min.   :  0.20         Min.   : 0.070
##  1st Qu.: 1850   1st Qu.: 69.78   1st Qu.: 26.95         1st Qu.: 4.465
##  Median : 5400   Median : 90.80   Median :139.00         Median :12.310
##  Mean   : 9635   Mean   : 81.51   Mean   :201.71         Mean   :16.120
##  3rd Qu.:13050   3rd Qu.: 98.00   3rd Qu.:317.90         3rd Qu.:23.330
##  Max.   :55100   Max.   :100.00   Max.   :898.00         Max.   :62.110
##                  NA's   :3        NA's   :1
##  Crops.percent    Other.Land.Use.percent    Climate         Birthrate
##  Min.   : 0.000   Min.   :33.91            Min.   :1.000   Min.   : 7.29
##  1st Qu.: 0.260   1st Qu.:70.39            1st Qu.:2.000   1st Qu.:11.95
##  Median : 1.080   Median :85.38            Median :2.000   Median :20.41
##  Mean   : 3.045   Mean   :80.84            Mean   :2.172   Mean   :22.64
##  3rd Qu.: 3.165   3rd Qu.:93.85            3rd Qu.:3.000   3rd Qu.:30.86
##  Max.   :23.320   Max.   :99.93            Max.   :4.000   Max.   :50.73
##                                            NA's   :16      NA's   :1
##    Deathrate       Agriculture        Industry          Service
##  Min.   : 2.410   Min.   :0.0000   Min.   :0.0400   Min.   :0.1770
##  1st Qu.: 6.213   1st Qu.:0.0400   1st Qu.:0.2210   1st Qu.:0.4255
##  Median : 8.870   Median :0.1010   Median :0.2940   Median :0.5500
##  Mean   : 9.827   Mean   :0.1534   Mean   :0.3070   Mean   :0.5390
##  3rd Qu.:12.207   3rd Qu.:0.2255   3rd Qu.:0.3575   3rd Qu.:0.6475
##  Max.   :29.500   Max.   :0.7690   Max.   :0.8010   Max.   :0.9060
##  NA's   :1
```

The aren't very many NA values since the datasets were pretty full, but there are 3 in Literacy.percent, 1 in Phones.per.1000, 16 in Climate, 1 in Birthrate, and 1 in Deathrate. Since there aren't a lot, I am going to impute these values. This can be done in a variety of ways. Since they are all numerical features, I could impute them with the mean or median. Also, I could impute using similar countries based on a model such as kNN. Since the dimensionality is quite high for kNN which works best on low dimensions 5-15 and there are 25, I won't use this method. Additionaly, due to the small amount of NAs, a central tendency imputation will likely be quite accurate and/or not skew the model much.

I will replace all with median, since it is less sensitive to outliers.

```r
#replace Literacy.percent with the median
world_df$Literacy.percent[is.na(world_df$Literacy.percent)] <-
median(world_df$Literacy.percent, na.rm = T)
```

```r
#replace Phones.per.1000 with the median
world_df$Phones.per.1000.people[is.na(world_df$Phones.per.1000.people)] <-
median(world_df$Phones.per.1000.people, na.rm = T)

#replace Climate with the median
world_df$Climate[is.na(world_df$Climate)] <- median(world_df$Climate, na.rm =
T)

#replace Birthrate with the median
world_df$Birthrate[is.na(world_df$Birthrate)] <- median(world_df$Birthrate,
na.rm = T)

#replace Deathrate with the median
world_df$Deathrate[is.na(world_df$Deathrate)] <- median(world_df$Deathrate,
na.rm = T)

#make sure it worked
anyNA(world_df)

## [1] FALSE
```

Awesome! All NA values have been imputed with the median.

Now are there any outliers?!?

There are multiple ways to detect outliers, such as using the IQR and the z-score, then we will make a determination of what to do with these values if any. Outliers can also be detected by plotting a linear regression model and with principal component analysis.

I am going to detect outliers using the Z/score.

Outlier = +/- 3 standard deviations from the mean. (A.k.a +/- z-score of 3). 3 is a rule of thumb, it does not work in every instance. We are going to use 3 in this instance since the variance is relatively standard and there are no observable clusters in the data. We could explore other values for the cutoff based on variance in the features because sometimes if there is pretty low variance and values tend to stick around a certain point, then a lower threshold may be better than the 3 heuristic. However, we are making the decision to explore IQR as another outlier detection method rather than other z-score cutoffs.

```r
#create a function that calculates outliers based on zscore and formula above
#function takes in a continuous variable and spits out the z scores
outlier.z <- function(cvar) {
a <- sd(cvar)
b <- mean(cvar)
c <- ((b-cvar)/(a))
c
}
```

First let’s output all observation numbers of the rows that have a abs(z-score) >3

```r
#detect which countries have large z-scores
n <- ncol(world_df)
for (i in 3:n){
p <- world_df[abs(outlier.z(world_df[,i]))>3,"Country"]
print(p)
}
```

```
## character(0)
## character(0)
## character(0)
## character(0)
## character(0)
## [1] "Singapore" "Qatar"     "Kuwait"
## [1] "China" "India"
## [1] "Canada"             "Australia"          "United States"
## [4] "Brazil"             "Russian Federation" "China"
## [1] "Singapore" "Hong Kong"
## [1] "Malta"     "Hong Kong"
## [1] "Singapore"   "Qatar"       "Kuwait"      "Afghanistan"
## [1] "Angola"      "Afghanistan"
## [1] "Luxembourg"
## [1] "Niger"
## [1] "United States"
## [1] "Bangladesh"
## [1] "Malaysia"    "Philippines" "Comoros"
## character(0)
## character(0)
## character(0)
## [1] "Botswana"
## [1] "Liberia"
## [1] "Qatar"
## character(0)
```

Based on the z-score method we can see that many countries have values that are larger than 3 z-scores from the mean in certain features. The countries fall on both high and low ends of the spectrum. These listed countries would be considered outliers by this method.

For example, you can see that China and India are outliers in terms of population, and the 6 largest countries are outliers in terms of land area. Singapore and Hong Kong are outliers in terms of Population Density. Luxembourg is an outlier in terms of GDP per capita. All of these findings match with what we would expect logically, which is a good sign.

Now let’s explore some of these outliers visually.

```r
library(ggplot2)
#install.packages("GridExtra")
library(gridExtra)
```

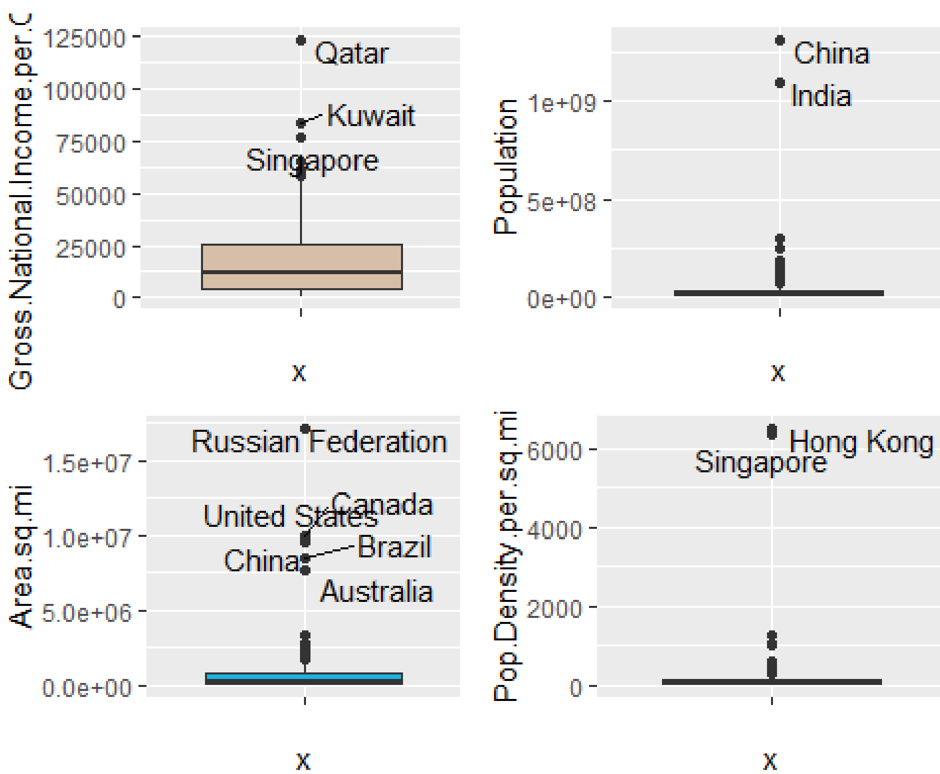
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

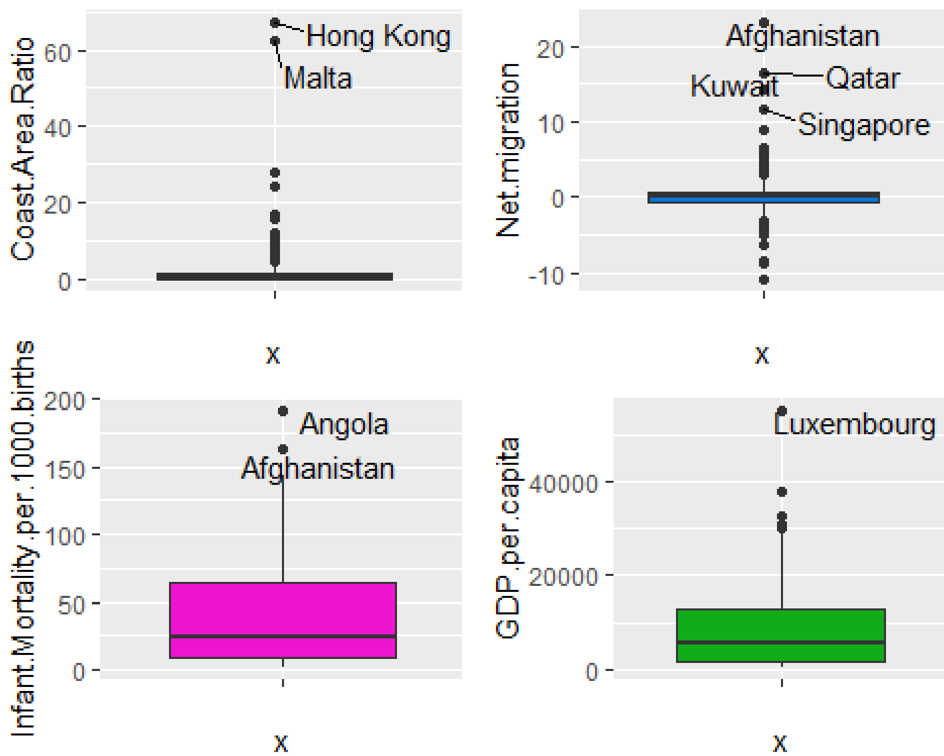
```r
#install.packages("ggrepel")
library(ggrepel)
attach(world_df)

#make boxplot for GNI per capita that labels outliers
g1 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Gross.National.Income.per.Capita))>3,
Country, "")) %>%
  ggplot(., aes(x = "", y = Gross.National.Income.per.Capita)) +
    geom_boxplot(fill = "#d6bea9") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)

#make boxplot for population that labels outliers
g2 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Population))>3, Country, "")) %>%
  ggplot(., aes(x = "", y = Population)) +
    geom_boxplot(fill = "#0066cc") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)


#make boxplot for land area that labels outliers
g3 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Area.sq.mi))>3, Country, "")) %>%
  ggplot(., aes(x = "", y = Area.sq.mi)) +
    geom_boxplot(fill = "#1cb2e3") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)
#make boxplot for population density that labels outliers
g4 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Pop.Density.per.sq.mi))>3, Country,
"")) %>%
  ggplot(., aes(x = "", y = Pop.Density.per.sq.mi)) +
    geom_boxplot(fill = "#54acbe") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)

#make boxplot for coast area ratio that labels outliers
g5 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Coast.Area.Ratio))>3, Country, ""))
%>%
  ggplot(., aes(x = "", y = Coast.Area.Ratio)) +
    geom_boxplot(fill = "#cbe123") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)

#make boxplot for net migration that labels outliers
g6 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Net.migration))>3, Country, "")) %>%
  ggplot(., aes(x = "", y = Net.migration)) +
```

```
    geom_boxplot(fill = "#0077dd") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)

#make boxplot for infant mortality that labels outliers
g7 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Infant.Mortality.per.1000.births))>3,
Country, "")) %>%
  ggplot(., aes(x = "", y = Infant.Mortality.per.1000.births)) +
    geom_boxplot(fill = "#ef14d1") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)

#make boxplot for GDP Per capita that labels outliers
g8 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(GDP.per.capita))>3, Country, "")) %>%
  ggplot(., aes(x = "", y = GDP.per.capita)) +
    geom_boxplot(fill = "#12ab1a") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)

#make boxplot for literacy that labels outliers
g9 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Literacy.percent))>3, Country, ""))
%>%
  ggplot(., aes(x = "", y = Literacy.percent)) +
    geom_boxplot(fill = "#bcde28") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)

#make boxplot for phones that labels outliers
g10 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Phones.per.1000.people))>3, Country,
"")) %>%
  ggplot(., aes(x = "", y = Phones.per.1000.people)) +
    geom_boxplot(fill = "#9aad3f") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)

#make boxplot for arable that labels outliers
g11 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Arable.percent))>3, Country, "")) %>%
  ggplot(., aes(x = "", y = Arable.percent)) +
    geom_boxplot(fill = "#1834af") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)

#make boxplot for crops that labels outliers
g12 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Crops.percent))>3, Country, "")) %>%
  ggplot(., aes(x = "", y = Crops.percent)) +
    geom_boxplot(fill = "#fe1568") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)

#make boxplot for deathrate that labels outliers
```

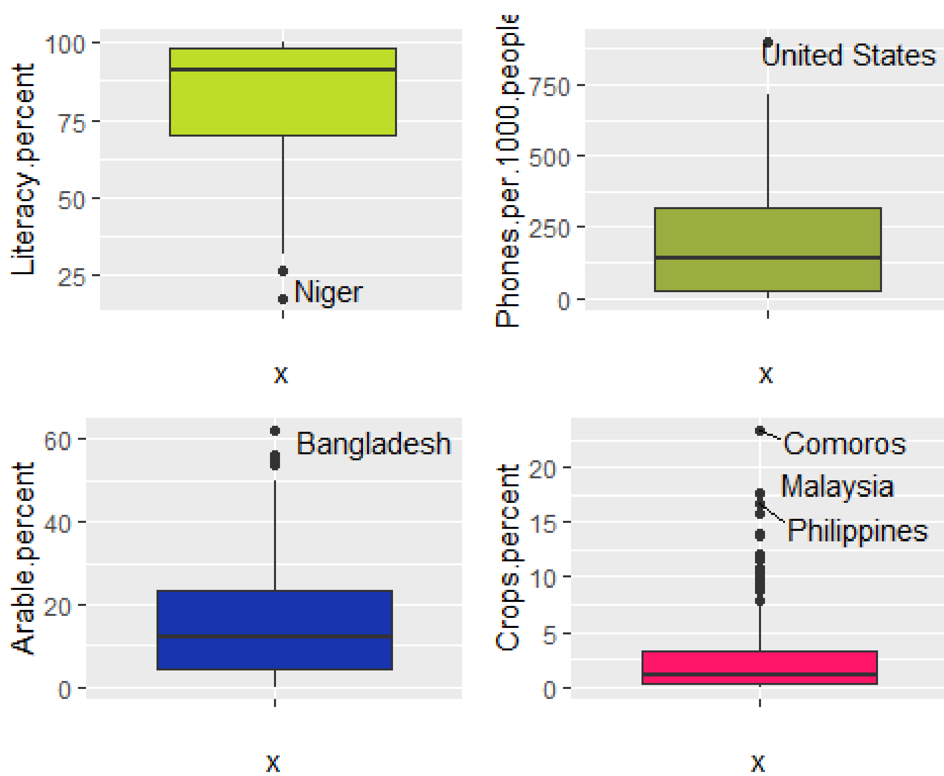
```
g13 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Deathrate))>3, Country, "")) %>%
  ggplot(., aes(x = "", y = Deathrate)) +
    geom_boxplot(fill = "#b15c16") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)


#make boxplot for agriculture that labels outliers
g14 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Agriculture))>3, Country, "")) %>%
  ggplot(., aes(x = "", y = Agriculture)) +
    geom_boxplot(fill = "#ed1478") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)

#make boxplot for industry that labels outliers
g15 <- world_df %>%
  mutate(outlier = ifelse(abs(outlier.z(Industry))>3, Country, "")) %>%
  ggplot(., aes(x = "", y = Industry)) +
    geom_boxplot(fill = "#bb0000") +
    geom_text_repel(aes(label = outlier), hjust = -0.2)
detach(world_df)

grid.arrange(g1, g2, g3, g4, nrow = 2)
```



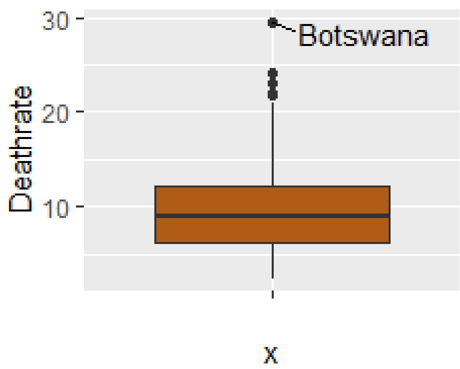

```
grid.arrange(g5, g6, g7, g8, nrow = 2)
```

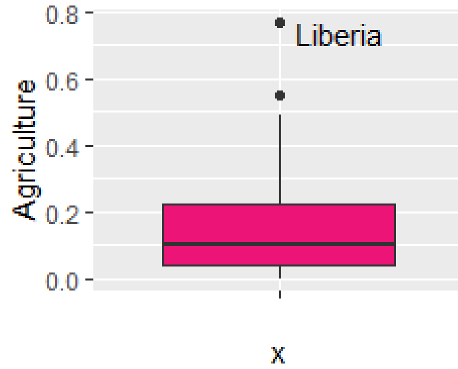


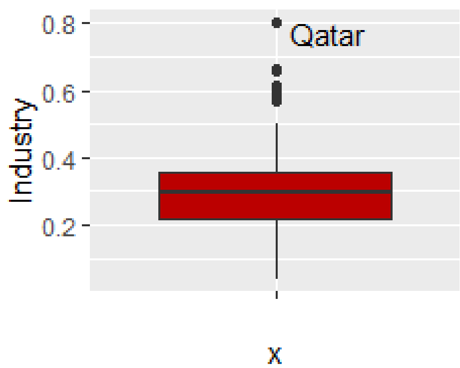

```r
grid.arrange(g9, g10, g11, g12, nrow = 2)
```




```r
grid.arrange(g13, g14, g15, nrow = 2)
```







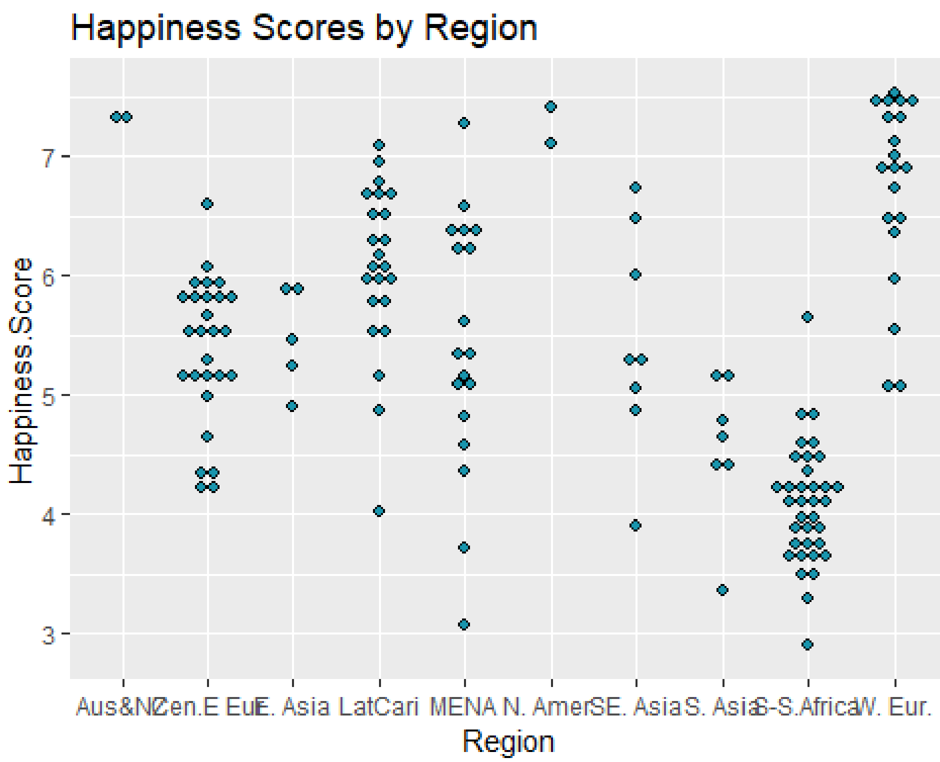
These boxplots (representing all of the features that contain outliers based on z-score) clearly show a good story into the data. All of the findings make sense with what we would expect and accessible layout the distributions of the features as well as label the outlier countries. The removal of outliers is something to take seriously. It could have big negative ramifications if you remove outliers without justification. Since all of the outliers here represent actual conditions out in the world and are not based on data input error, I am going to make the careful decision to leave them all in the dataset. This is an assumption that will be marked. I will pay particular attention to output result with these in mind. However, I fully expect that the presence of these data points won't affect our analysis to a grand extent.

Also, since we don't have very many observations, once we scale our data, outliers now may not remain outliers. Lastly, once more data is collected (such as with the other countries in the world), it's possible outliers won't be outliers anymore.

Let's do some more exploratory visualizations to get a sense of relationships between variables.

```
r1 <- ggplot(world_df, aes(Region, Happiness.Score))

#show region names
levels(world_df$Region)

##  [1] "Australia and New Zealand"       "Central and Eastern Europe"
##  [3] "Eastern Asia"                    "Latin America and Caribbean"
##  [5] "Middle East and Northern Africa" "North America"
```

```
##  [7] "Southeastern Asia"                "Southern Asia"
##  [9] "Sub-Saharan Africa"               "Western Europe"

#abbreviate region names to make graph more legible
reg_abbr <- c("Australia and New Zealand" = "Aus&NZ", "Central and Eastern
Europe" = "Cen.E Eur","Eastern Asia" = "E. Asia", "Latin America and
Caribbean" = "LatCari", "Middle East and Northern Africa" = "MENA", "North
America" = "N. Amer.", "Southeastern Asia" = "SE. Asia", "Southern Asia" =
"S. Asia", "Sub-Saharan Africa" = "S-S.Africa", "Western Europe" = "W. Eur.")

#plot dotplot to show distribution
r1 + geom_dotplot(binaxis = "y", stackdir = "center", binwidth = 0.1, fill =
"#1995AD") + scale_x_discrete(labels = reg_abbr) + ggtitle("Happiness Scores
by Region")
```



This dot plot nicely shows some distribution of happiness scores by region. You can see some regions such as the Middle East and North Africa have a large varaince. For example, there is a country with a happiness score of just above 3, but there is also a country with a happiness score of over 7. Also, you can see that the regions North America and Australia & NZ only have two members each.

Because the variance within regions is high and some regions only have a few members, I anticipate my eventual classification of region may be difficult. I can try to combat this with a few methods such as stratified sampling, sampling with replacement, and k-fold cross validation. We will see how it turns out!

My first model I am going to explore is Multiple Linear Regression in hopes of predicting the Happiness score of a nation.

Multiple Linear Regression

Let's do some correlation/collinearity analysis, since multicollinearity can doom a regression model.

Let's explore correlations to the response variable Happiness Score since the full correlation table would be hard to digest.

```
#correlations just to the response variable Sale Price
cormatx.response <- round(cor(world_df[c(4:26)], world_df[3]),2)
cormatx.response

##                                   Happiness.Score
## HDI.Score                                    0.83
## Life.Expectancy.at.Birth                     0.78
## Expected.Years.of.Education                  0.74
## Mean.Years.of.Education                      0.71
## Gross.National.Income.per.Capita             0.68
## Population                                  -0.02
## Area.sq.mi                                   0.16
## Pop.Density.per.sq.mi                        0.09
## Coast.Area.Ratio                             0.16
## Net.migration                                0.17
## Infant.Mortality.per.1000.births            -0.71
## GDP.per.capita                               0.73
## Literacy.percent                             0.66
## Phones.per.1000.people                       0.73
## Arable.percent                              -0.06
## Crops.percent                               -0.17
## Other.Land.Use.percent                       0.09
## Climate                                      0.29
## Birthrate                                   -0.70
## Deathrate                                   -0.49
## Agriculture                                 -0.69
## Industry                                     0.19
## Service                                      0.53
```

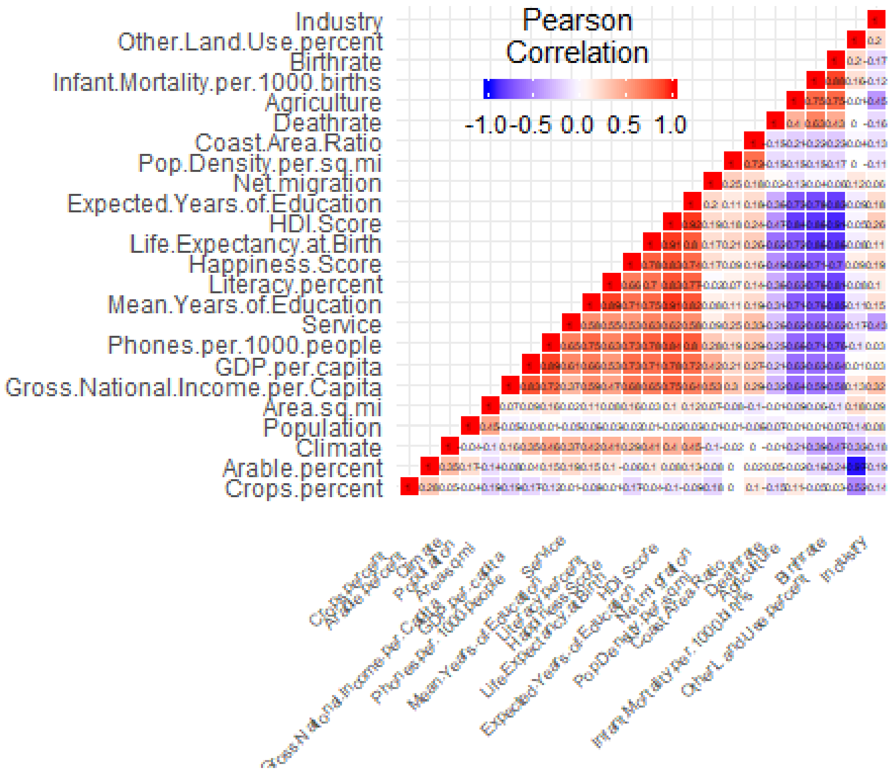
High values indicate high correlations, and when there are multiple features correlated with one another (which is not visualized here…yet), that indicates collinearity, which is not ideal for a regression analysis. Essentially, the same information is conveyed by multiple variables. Right off the bat I can see some high correlations such as between HDI Score and Happiness Score.

This is a little difficult to visualize, though. Let's see if we can visualize it better.

Using starter code from STHDA [____], we are going to create a correlation matrix that is shaded by intensity of correlation.

```
#create full correlation matrix
cormatx <- round(cor(world_df[3:26]), 2)

reorder_cormatx <- function(cormat){
# Use correlation between variables as distance
dd <- as.dist((1-cormatx)/2)
hc <- hclust(dd)
cormat <-cormat[hc$order, hc$order]
}

# Get upper triangle of the correlation matrix
  get_upper_tri <- function(cormat){
    cormat[lower.tri(cormat)]<- NA
    return(cormat)
  }

#install.packages("reshape2")
library(reshape2)

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths

# Reorder the correlation matrix
cormatx <- reorder_cormatx(cormatx)
upper_tri <- get_upper_tri(cormatx)
# Melt the correlation matrix
melted_cormatx <- melt(upper_tri, na.rm = TRUE)
# Create a ggheatmap
ggheatmap <- ggplot(melted_cormatx, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
    name="Pearson\nCorrelation") +
  theme_minimal()+ # minimal theme
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 6, hjust = 1.5))+
 coord_fixed()

#format heatmap
ggheatmap +
geom_text(aes(Var2, Var1, label = value), color = "black", size = 1.2) +
theme(
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.background = element_blank(),
  axis.ticks = element_blank(),
  legend.justification = c(1, 0),
```

```
legend.position = c(0.6, 0.7),
legend.direction = "horizontal")+
guides(fill = guide_colorbar(barwidth = 5, barheight = .5,
              title.position = "top", title.hjust = 0.5))
```



Ahh much better.

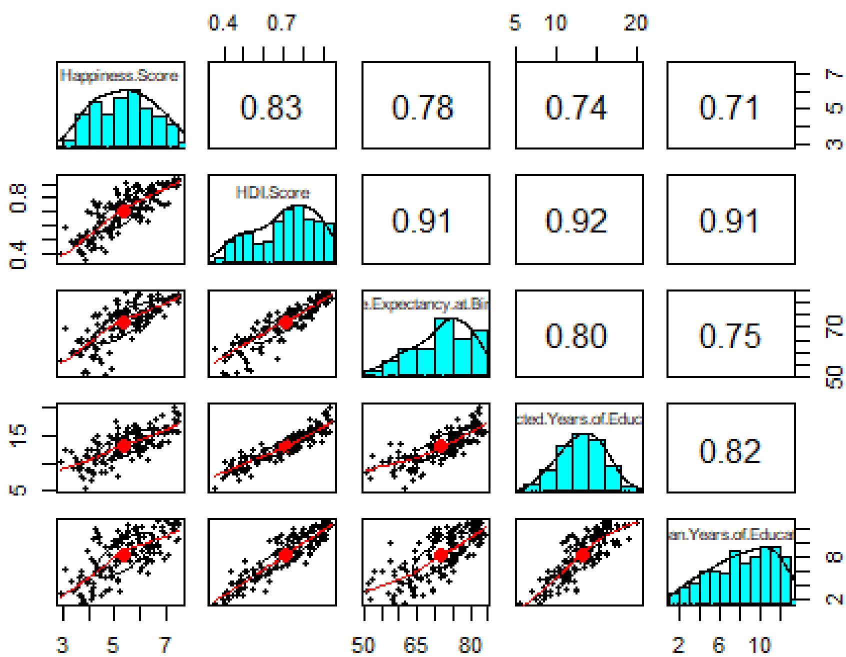
Now we can visualize our correlations. As we can see, some overall correlations between the variables are pretty high which means there is likely collinearity. The strongest correlations (which we are going to count as ones with an absolute value >=.75) are between HDI.Score and other variables. Since HDI Score is essentially another dependent variable that was calculated based on a variety of factors, I am going to drop it and keep the other features and see how this improves collinearity. Collinearity exists when too many features explain eachother, and it seems that the correlations between the variables are high enough to explain eachother.

Since feature removal is a big deal and can have large adverse impacts to a model if done incorrectly, I won't make any additional removals before creating a model and finding the Variance Inflation Factor (VIF) for each predictor. This VIF value will help me understand when it's safe to remove features.

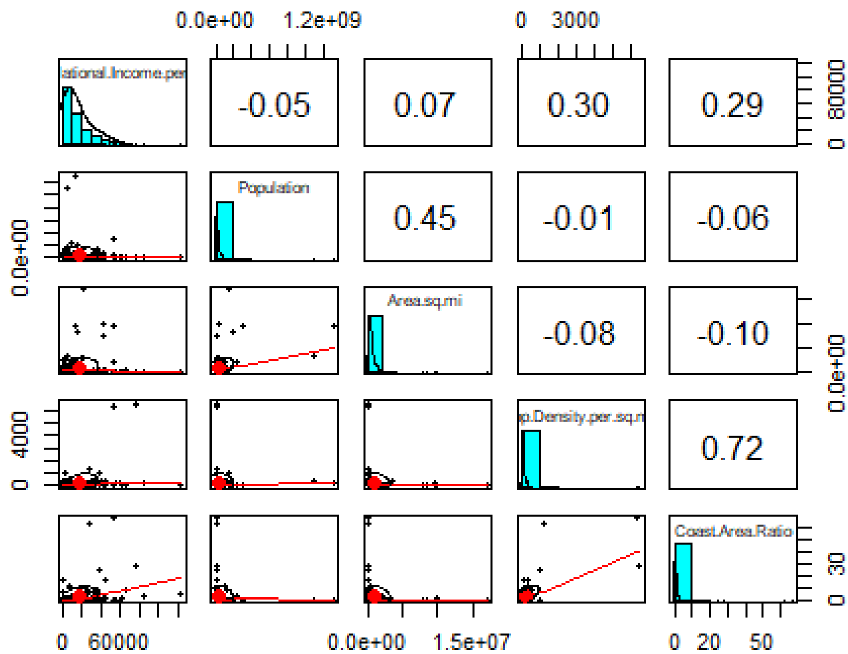
I will be building this regression model shortly.

First, I want to do some more exploratory data visualization with pairs.panels() and inspect distribution of features to see if I need to apply transforms in order to make a normally distributed dataset. Linear regression is parametric and assuming features are normally distributed.

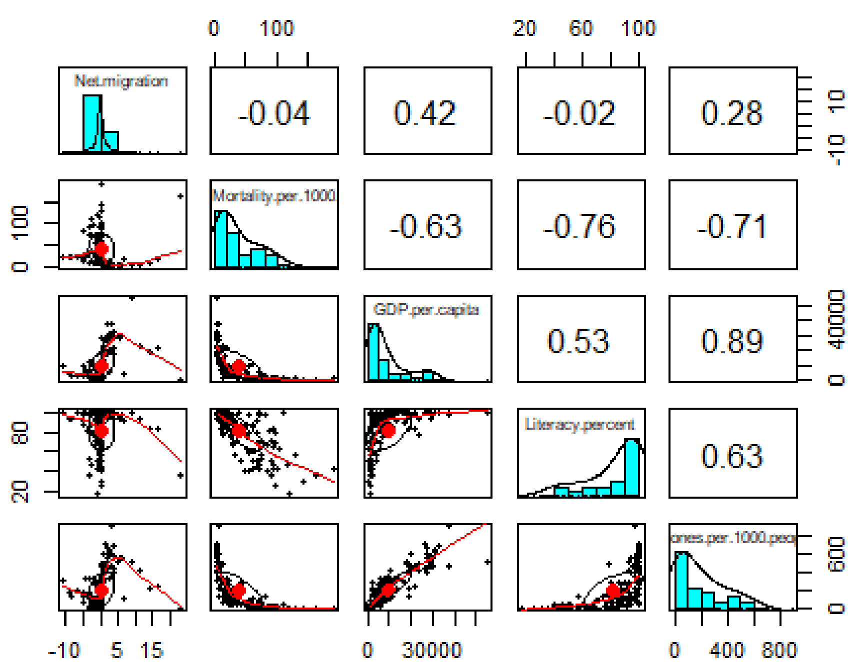
```r
#install.packages("psych")
library(psych)

##
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha

pairs.panels(world_df[3:7])
```



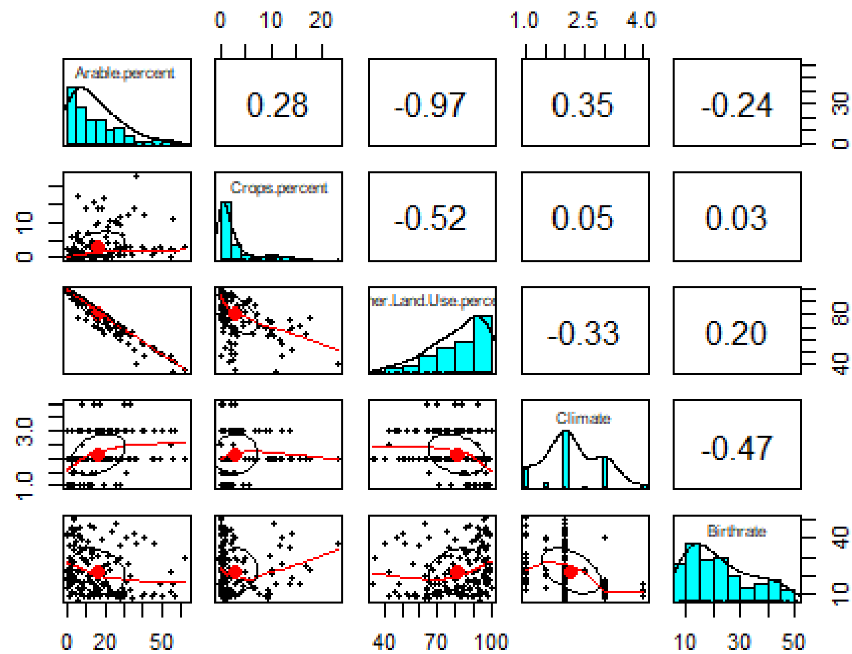

```r
pairs.panels(world_df[8:12])
```



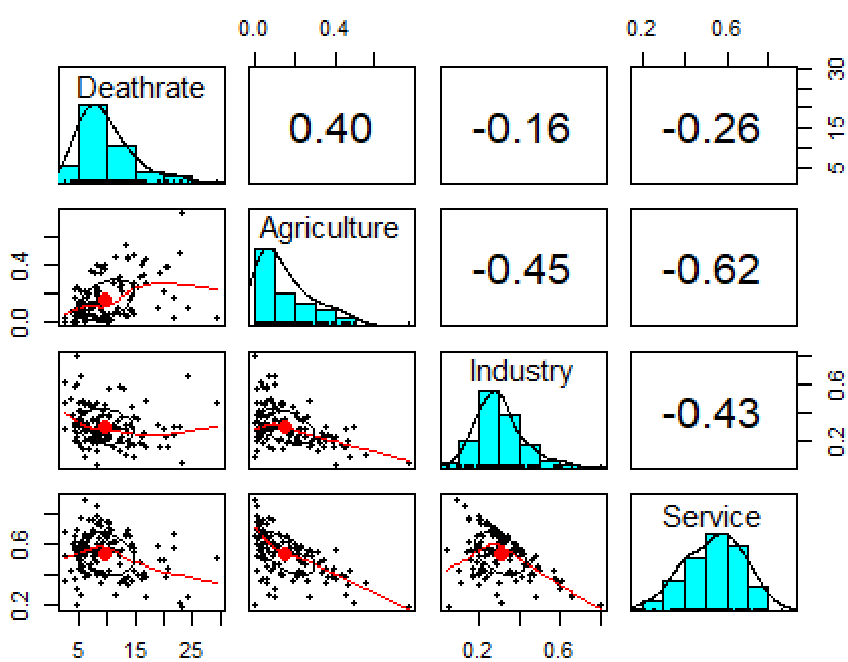

```
pairs.panels(world_df[13:17])
```




```
pairs.panels(world_df[18:22])
```




```
pairs.panels(world_df[23:26])
```



When the oval (correlation elipse) is stretched, it means a strong correlation. We can see that Expected years of education and mean years of education have strong correlations and likely explain eachother, for example. There could be collinearity between these.

Regression assumes normality, so let's see if any transforms help the data look more normally distributed. Age and absences in particular look off.

Let's explore a few of these closer. In particular we are concerned about transforming the variables such as HDI.Score, Life Expectancy, Mean years of education, Gross National Income, Population, Area sq mi, population density, coast area ratio, net migration, infant mortality, GDP per capita, literacy, phones per 1000, arable percent, crops percent, other land use percent, climate, birthrate, deathrate, agriculture to make them resemble normal distributions more closely. I am deeming the other features to be fairly normally distributed.

Disclaimer: it is possible that some of these features may not even make it into the final model due to feature selection and backfitting, but I am going to normalize them for good measure.

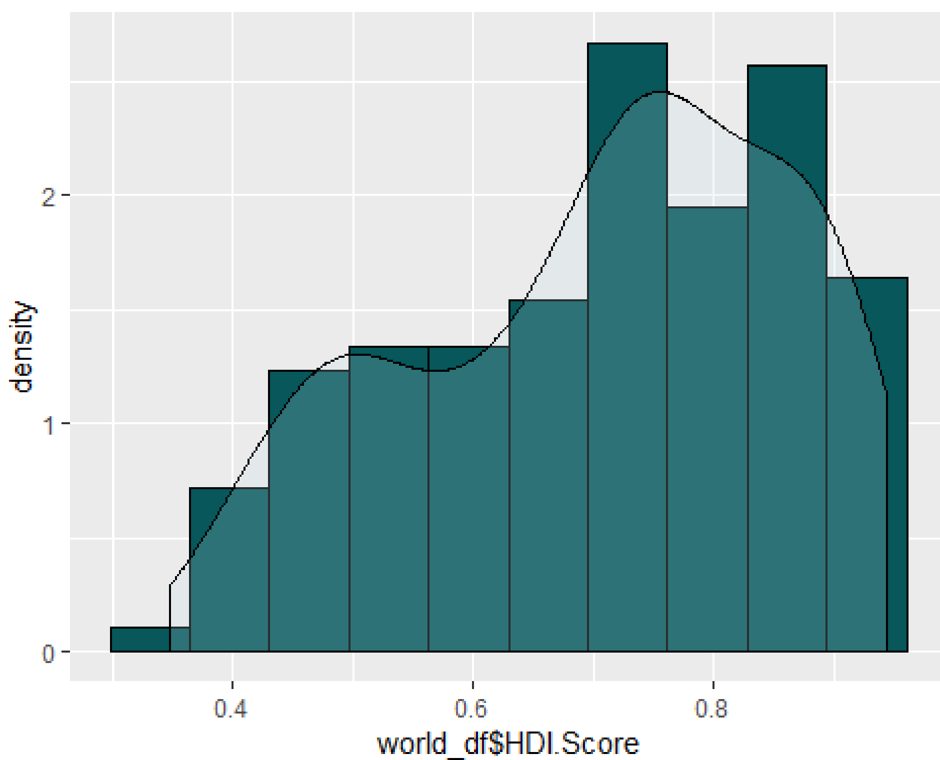
To make this faster I will make function to min/max and z-score transform features

```
#normalize columns with min-max normalization by creating a function that
takes in an argument "x" and normalizes between 0-1 using the min and max
method
normalize <- function(x) {
  return( (x-min(x))/ diff(range(x)))
}

#standardize columns with z-score standardization by creating a function that
takes in an argument "y" and standardizes between +/- z-scores using z-score
standardization method
zstandardize <- function(y) {
  return( (y-mean(y))/ (sd(y)))
}
```

First let's start with HDI.Score.

```
# Histogram with density instead of count on y-axis
# Overlay with transparent density plot

#ORIGINAL
a <- ggplot(world_df, aes(x=world_df$HDI.Score)) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#LOG TRANSFORM
a1 <- ggplot(world_df, aes(x=log(world_df$HDI.Score))) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")
```

```r
#INVERSE TRANSFORM
a2 <- ggplot(world_df, aes(x=1/((world_df$HDI.Score)))) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#SQRT TRANSFORM
a3 <- ggplot(world_df, aes(x=sqrt(world_df$HDI.Score))) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#SQUARE TRANSFORM
a4 <- ggplot(world_df, aes(x=(world_df$HDI.Score)^2)) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#MIN/MAX TRANSFORM
a5 <- ggplot(world_df, aes(x=normalize(world_df$HDI.Score))) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#Z-SCORE TRANSFORM
a6 <- ggplot(world_df, aes(x= zstandardize(world_df$HDI.Score))) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")
#print original
a
```



```r
#print options
grid.arrange(a1,a2,a3,a4,a5,a6, nrow=3)
```

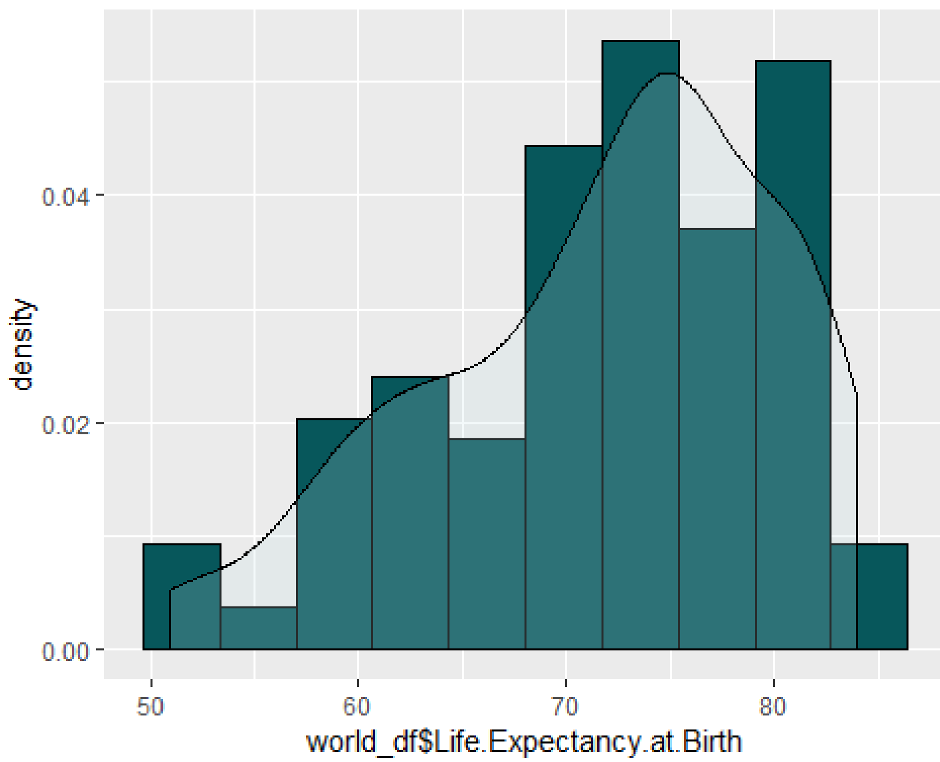
It looks like the Square ^2 transform makes it most resemble a normal distribution, so let's replace it with its square.

```r
#make new data frame that is more normally distributed
world_norm_dist <- world_df

#replace feature
world_norm_dist$HDI.Score <- (world_df$HDI.Score)^2
```

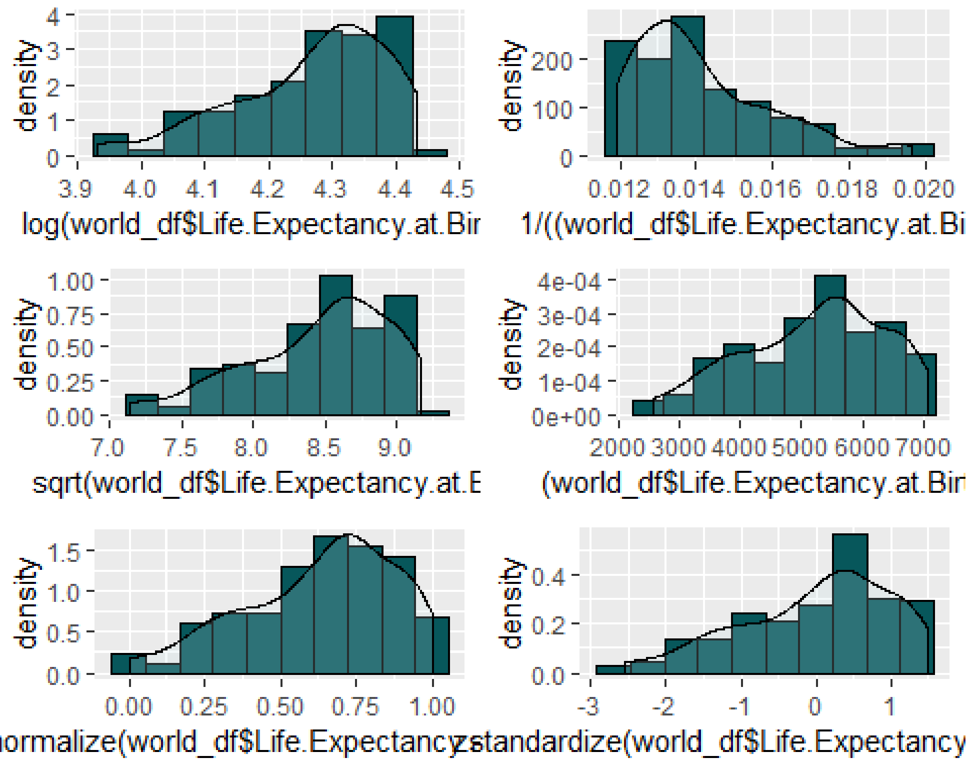
Now Life Expectancy.

```r
# Histogram with density instead of count on y-axis
# Overlay with transparent density plot

#ORIGINAL
a <- ggplot(world_df, aes(x=world_df$Life.Expectancy.at.Birth)) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#LOG TRANSFORM
a1 <- ggplot(world_df, aes(x=log(world_df$Life.Expectancy.at.Birth))) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")
```

```r
#INVERSE TRANSFORM
a2 <- ggplot(world_df, aes(x=1/((world_df$Life.Expectancy.at.Birth)))) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#SQRT TRANSFORM
a3 <- ggplot(world_df, aes(x=sqrt(world_df$Life.Expectancy.at.Birth))) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#SQUARE TRANSFORM
a4 <- ggplot(world_df, aes(x=(world_df$Life.Expectancy.at.Birth)^2)) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#MIN/MAX TRANSFORM
a5 <- ggplot(world_df, aes(x=normalize(world_df$Life.Expectancy.at.Birth))) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#Z-SCORE TRANSFORM
a6 <- ggplot(world_df, aes(x=
zstandardize(world_df$Life.Expectancy.at.Birth))) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")
#print original
a
```




```r
#print options
grid.arrange(a1,a2,a3,a4,a5,a6, nrow=3)
```



It looks like the min/max transform makes it most resemble a normal distribution since it slightly reduces the left skew, so let's replace it with it.

```
#replace feature
world_norm_dist$Life.Expectancy.at.Birth <-
normalize(world_df$Life.Expectancy.at.Birth)
```

The above iterations of testing and transforming features shows the process visually. Now, in orer to speed things up for the remaining 18 features I want to transform, I am going to use the bestNormalize package. This package checks all of the transforms (plus more complicated ones) similar to how I have been doing, then transforms the data based on the best transform. The best transform is determined by the Estimated Normality Statistics (Pearson P / df). The lower the value ==> the more normal it is. The function is doing repeated CV in order to find the best transform.

The orderNorm method guarantees normality, so I will set this to false since it is not as natural of a transform.

Use bestNormalize for remaining features.

```
#install.packages("bestNormalize")
library(bestNormalize)
set.seed(300)

# Pick the best one automatically for the remaining features
#k = number of folds and r = number of repeats for the CV. Helps with
run=time performance
```

```r
mean.edu.t <- bestNormalize(world_df$Mean.Years.of.Education, allow_orderNorm
= F, k = 5, r = 3)

gni.t <- bestNormalize(world_df$Gross.National.Income.per.Capita,
allow_orderNorm = F, k = 5, r = 3)

## Warning in bestNormalize(world_df$Gross.National.Income.per.Capita,
allow_orderNorm = F, : exp_x  did not work;  Error in exp_x(standardize =
TRUE, warn = TRUE, x = c(44025, 56431, 35182,  :
##   Transformation finite for less than 3 x values

pop.t <- bestNormalize(world_df$Population, allow_orderNorm = F, k = 5, r =
3)

## Warning in bestNormalize(world_df$Population, allow_orderNorm = F, k = 5,
: exp_x  did not work;  Error in exp_x(standardize = TRUE, warn = TRUE, x =
c(5450661L, 7523934L,  :
##   Transformation finite for less than 3 x values

area.t <- bestNormalize(world_df$Area.sq.mi, allow_orderNorm = F, k = 5, r =
3)

## Warning in bestNormalize(world_df$Area.sq.mi, allow_orderNorm = F, k = 5,
: exp_x  did not work;  Error in exp_x(standardize = TRUE, warn = TRUE, x =
c(43094L, 41290L,  :
##   Transformation finite for less than 3 x values

pop.den.t <- bestNormalize(world_df$Pop.Density.per.sq.mi, allow_orderNorm =
F, k = 5, r = 3)

## Warning in bestNormalize(world_df$Pop.Density.per.sq.mi, allow_orderNorm =
F, : exp_x  did not work;  Error in exp_x(standardize = TRUE, warn = TRUE, x
= c(126.5, 182.2, 2.9,  :
##   Transformation finite for less than 3 x values

coast.t <- bestNormalize(world_df$Coast.Area.Ratio, allow_orderNorm = F, k =
5, r = 3)

## Warning in bestNormalize(world_df$Coast.Area.Ratio, allow_orderNorm = F, :
boxcox  did not work;  Error in estimate_boxcox_lambda(x, ...) : x must be
positive

migrate.t <- bestNormalize(world_df$Net.migration, allow_orderNorm = F, k =
5, r = 3)

## Warning in bestNormalize(world_df$Net.migration, allow_orderNorm = F, k =
5, : boxcox  did not work;  Error in estimate_boxcox_lambda(x, ...) : x must
be positive

infant.t <- bestNormalize(world_df$Infant.Mortality.per.1000.births,
allow_orderNorm = F, k = 5, r = 3)
```

```
gdp.t <- bestNormalize(world_df$GDP.per.capita, allow_orderNorm = F, k = 5, r
= 3)

## Warning in bestNormalize(world_df$GDP.per.capita, allow_orderNorm = F, k =
5, : exp_x  did not work;  Error in exp_x(standardize = TRUE, warn = TRUE, x
= c(31100L, 32700L,  :
##   Transformation finite for less than 3 x values

literacy.t <- bestNormalize(world_df$Literacy.percent, allow_orderNorm = F, k
= 5, r = 3)

phone.t <- bestNormalize(world_df$Phones.per.1000.people, allow_orderNorm =
F, k = 5, r = 3)

## Warning in bestNormalize(world_df$Phones.per.1000.people, allow_orderNorm
= F, : exp_x  did not work;  Error in exp_x(standardize = TRUE, warn = TRUE,
x = c(614.6, 680.9, 647.7,  :
##   Transformation finite for less than 3 x values

arable.t <- bestNormalize(world_df$Arable.percent, allow_orderNorm = F, k =
5, r = 3)

crop.t <- bestNormalize(world_df$Crops.percent, allow_orderNorm = F, k = 5, r
= 3)

## Warning in bestNormalize(world_df$Crops.percent, allow_orderNorm = F, k =
5, : boxcox  did not work;  Error in estimate_boxcox_lambda(x, ...) : x must
be positive

other.t <- bestNormalize(world_df$Other.Land.Use.percent, allow_orderNorm =
F, k = 5, r = 3)

climate.t <- bestNormalize(world_df$Climate, allow_orderNorm = F, k = 5, r =
3)

birth.t <- bestNormalize(world_df$Birthrate, allow_orderNorm = F, k = 5, r =
3)

death.t <- bestNormalize(world_df$Deathrate, allow_orderNorm = F, k = 5, r =
3)

agricul.t <- bestNormalize(world_df$Agriculture, allow_orderNorm = F, k = 5,
r = 3)

## Warning in bestNormalize(world_df$Agriculture, allow_orderNorm = F, k = 5,
: boxcox  did not work;  Error in estimate_boxcox_lambda(x, ...) : x must be
positive
```

As you can see, not every transform works for every feature. Yet, the best one is still chosen. This takes quite a while to run because it is doing 5 fold CV with 3 repeats for every feature to ensure the best transform is chosen.

Now that we have found the best transform for all of the remaining features, I am going to show an example output and then replace the values in our dataframe with the transformed values. The transformed values from the bestNormalize function can be accessed in the $x.t call.

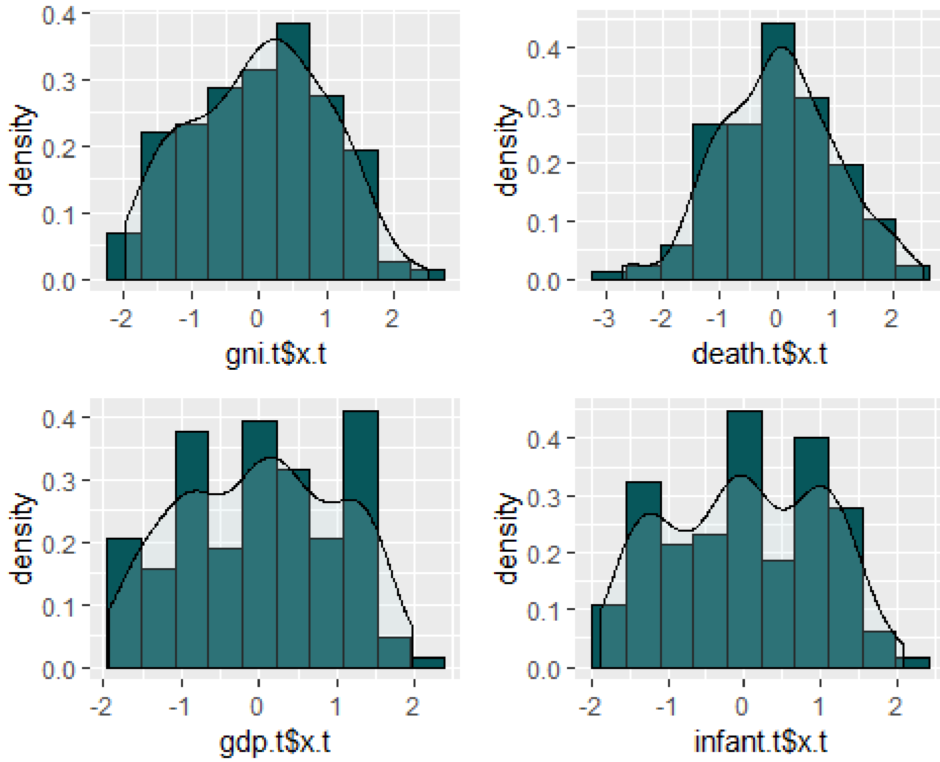
An example output for the Infant Mortality feature is:

```
#Show chosen transform and statistics
infant.t

## Best Normalizing transformation with 147 Observations
##  Estimated Normality Statistics (Pearson P / df, lower => more normal):
##  - No transform: 4.068
##  - Box-Cox: 1.5484
##  - Log_b(x+a): 1.7473
##  - sqrt(x+a): 2.2009
##  - exp(x): 18.7236
##  - arcsinh(x): 1.7036
##  - Yeo-Johnson: 1.5773
## Estimation method: Out-of-sample via CV with 5 folds and 3 repeats
##
## Based off these, bestNormalize chose:
## Standardized Box Cox Transformation with 147 nonmissing obs.:
##  Estimated statistics:
##  - lambda = 0.1229329
##  - mean (before standardization) = 3.953822
##  - sd (before standardization) = 1.638873

#Show transformed values
head(infant.t$x.t)

## [1] -1.394716 -1.422587 -1.625708 -1.546428 -1.572004 -1.364624
```

Let’s see if this actually works visually.

```
#spot check for gni
a1 <- ggplot(world_df, aes(x= gni.t$x.t)) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#spot check for deathrate
a2 <- ggplot(world_df, aes(x= death.t$x.t)) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#spot check for gdp per capita
a3 <- ggplot(world_df, aes(x= gdp.t$x.t)) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

#Box Cox transform for infant mortality
```

```r
a4 <- ggplot(world_df, aes(x= infant.t$x.t)) +
geom_histogram(aes(y=..density..),bins=10, colour="black",
fill="#07575b")+geom_density(alpha=.2, fill="#c4dfe6")

grid.arrange(a1,a2,a3,a4,nrow=2)
```



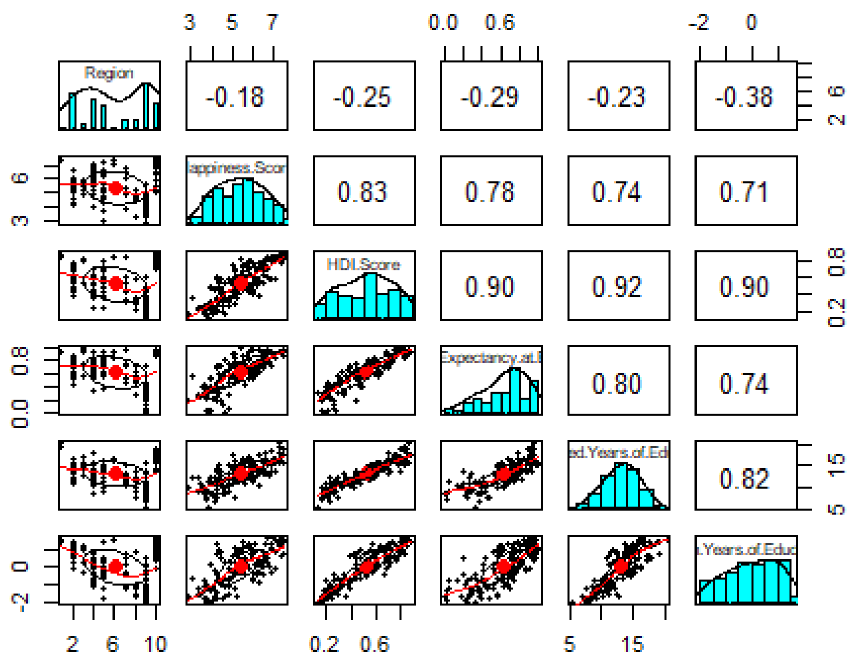
Looks a lot better than before! Now let's apply all of these to the normally distributed data frame. If at the end of our regression analysis we have to reverse any transform, we can easily access which transform was applied using the $chosen_transform call.

```r
#replace features with transforms
world_norm_dist$Mean.Years.of.Education <- mean.edu.t$x.t
world_norm_dist$Gross.National.Income.per.Capita <- gni.t$x.t
world_norm_dist$Population <- pop.t$x.t
world_norm_dist$Area.sq.mi <- area.t$x.t
world_norm_dist$Pop.Density.per.sq.mi <- pop.den.t$x.t
world_norm_dist$Coast.Area.Ratio <- coast.t$x.t
world_norm_dist$Net.migration <- migrate.t$x.t
world_norm_dist$Infant.Mortality.per.1000.births <- infant.t$x.t
world_norm_dist$GDP.per.capita <- gdp.t$x.t
world_norm_dist$Literacy.percent <- literacy.t$x.t
world_norm_dist$Phones.per.1000.people <- phone.t$x.t
world_norm_dist$Arable.percent <- arable.t$x.t
world_norm_dist$Crops.percent <- crop.t$x.t
world_norm_dist$Other.Land.Use.percent <- other.t$x.t
world_norm_dist$Climate <- climate.t$x.t
world_norm_dist$Birthrate <- birth.t$x.t
```

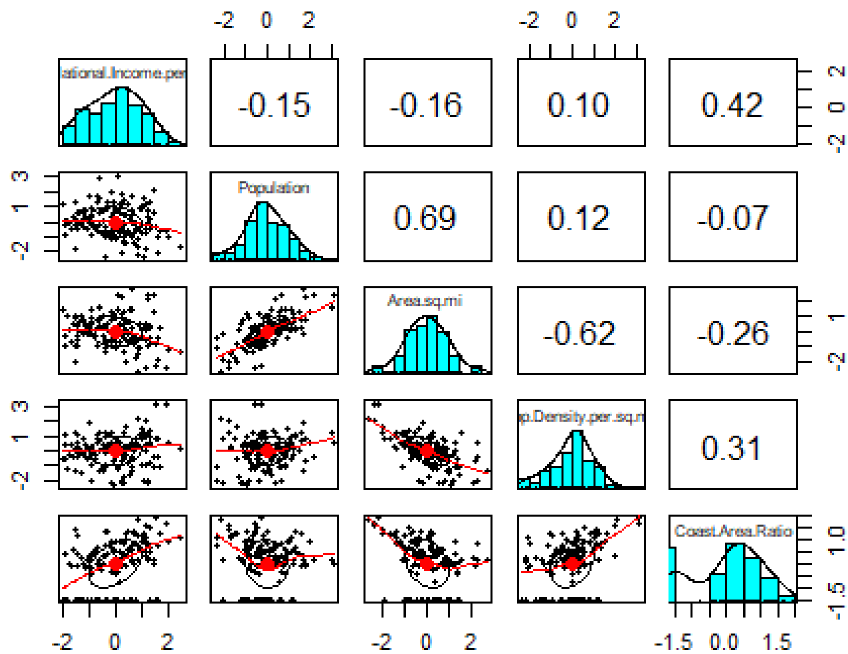
```
world_norm_dist$Deathrate <- death.t$x.t
world_norm_dist$Agriculture <- agricul.t$x.t
```

Look at the pairs.panels again.

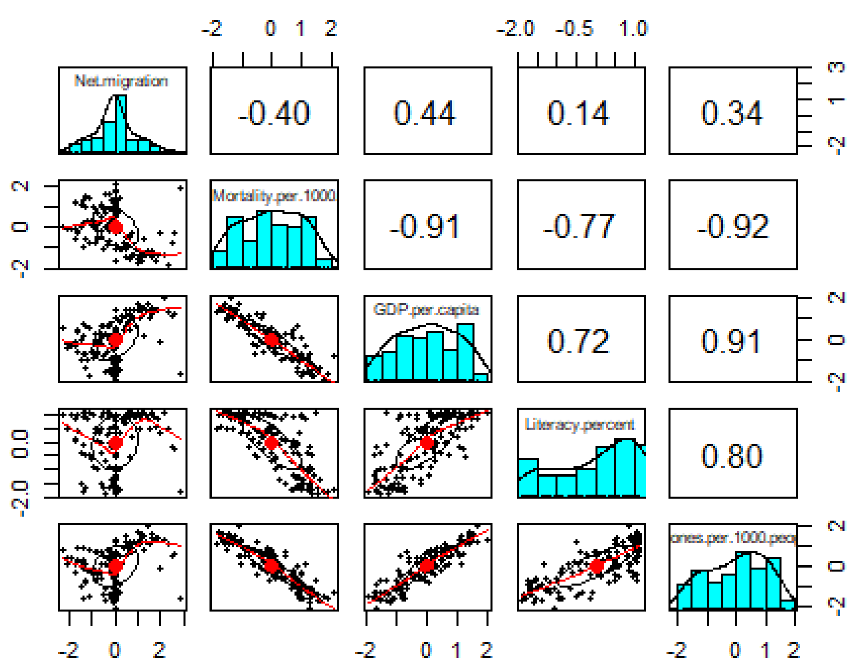
```
pairs.panels(world_norm_dist[2:7])
```



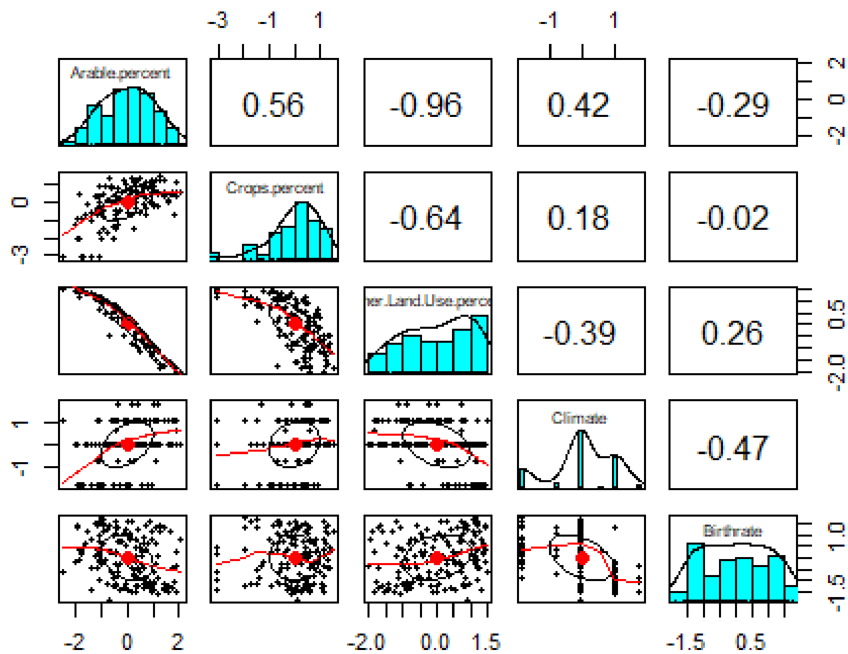

```
pairs.panels(world_norm_dist[8:12])
```



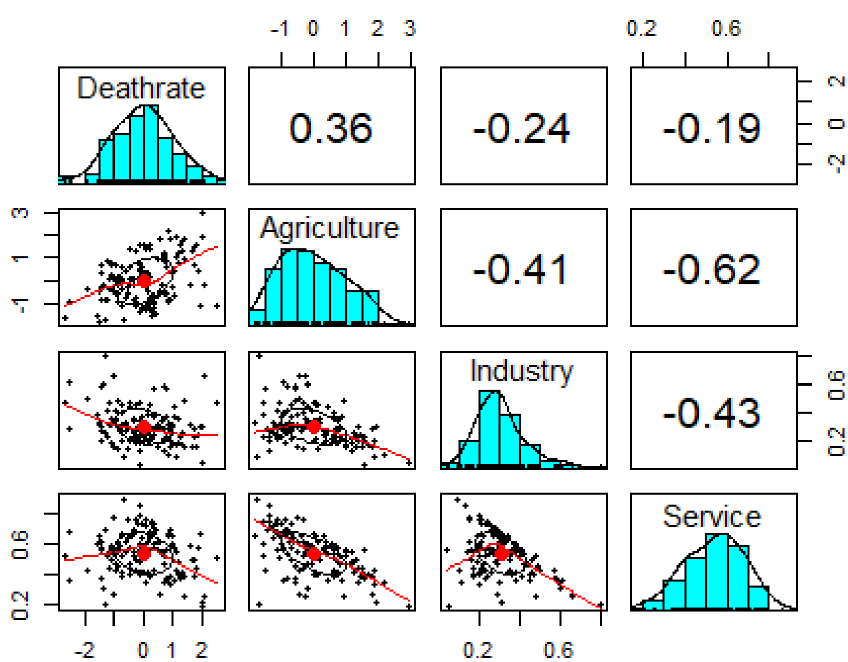

```
pairs.panels(world_norm_dist[13:17])
```




```
pairs.panels(world_norm_dist[18:22])
```




```r
pairs.panels(world_norm_dist[23:26])
```



It looks much better!

Now dummy code the “Region” category for the normalized dataset since in this instance it is a predictor and not a response variable.

```
region_vars <- model.matrix( ~ Region - 1, data=world_norm_dist )
head(region_vars[,-10])

##   RegionAustralia and New Zealand RegionCentral and Eastern Europe
## 1                               0                                0
## 2                               0                                0
## 3                               0                                0
## 4                               0                                0
## 5                               0                                0
## 6                               0                                0
##   RegionEastern Asia RegionLatin America and Caribbean
## 1                  0                                 0
## 2                  0                                 0
## 3                  0                                 0
## 4                  0                                 0
## 5                  0                                 0
## 6                  0                                 0
##   RegionMiddle East and Northern Africa RegionNorth America
## 1                                     0                   0
## 2                                     0                   0
## 3                                     0                   0
## 4                                     0                   0
## 5                                     0                   0
## 6                                     0                   1
##   RegionSoutheastern Asia RegionSouthern Asia RegionSub-Saharan Africa
## 1                       0                   0                        0
## 2                       0                   0                        0
## 3                       0                   0                        0
## 4                       0                   0                        0
## 5                       0                   0                        0
## 6                       0                   0                        0

#add dummy columns -1 to the data. There is always one less columns than
there are levels
world_norm_dist <- cbind(world_norm_dist, region_vars[,-10])

#do a quick spot check
head(world_df$Region)

## [1] Western Europe Western Europe Western Europe Western Europe
## [5] Western Europe North America
## 10 Levels: Australia and New Zealand ... Western Europe
```

The binary dummy variables match with the actual values! Sweet. If all values are 0, this means that the region is Western Europe. This will be represented by the intercept in the regression model.

Now I am going to remove the original region column.

```r
world_norm_dist <- world_norm_dist[-2]
```

Remove spaces and special characters from new variable names.

```r
world_norm_dist <- rename(world_norm_dist, Region.AusNZ = "RegionAustralia
and New Zealand", Region.Cen.E.Eur = "RegionCentral and Eastern Europe",
Region.E.Asia = "RegionEastern Asia", Region.LatCari = "RegionLatin America
and Caribbean", Region.MENA = "RegionMiddle East and Northern Africa",
Region.N.Amer = "RegionNorth America", Region.SE.Asia = "RegionSoutheastern
Asia", Region.S.Asia = "RegionSouthern Asia", Region.SS.Africa = "RegionSub-
Saharan Africa")
```

Create an easy list of predictors to pull from for the regression model.

```r
#prepare the list of predictor names for multiple regression
var_names <- names(world_norm_dist[3:34])
formula <- as.formula(paste('Happiness.Score ~ '
,paste(var_names,collapse='+')))

#make sure it worked
formula

## Happiness.Score ~ HDI.Score + Life.Expectancy.at.Birth +
Expected.Years.of.Education +
##     Mean.Years.of.Education + Gross.National.Income.per.Capita +
##     Population + Area.sq.mi + Pop.Density.per.sq.mi + Coast.Area.Ratio +
##     Net.migration + Infant.Mortality.per.1000.births + GDP.per.capita +
##     Literacy.percent + Phones.per.1000.people + Arable.percent +
##     Crops.percent + Other.Land.Use.percent + Climate + Birthrate +
##     Deathrate + Agriculture + Industry + Service + Region.AusNZ +
##     Region.Cen.E.Eur + Region.E.Asia + Region.LatCari + Region.MENA +
##     Region.N.Amer + Region.SE.Asia + Region.S.Asia + Region.SS.Africa
```

I am going to build a multiple regression model with the aim of using the VIF to help with feature selection. If there are variables that explain eachother too much, I will know to remove them. Any VIF above 20 or so is considered high

```r
#make model for all features
m1 <- lm(Happiness.Score ~ HDI.Score + Life.Expectancy.at.Birth +
Expected.Years.of.Education +
    Mean.Years.of.Education + Gross.National.Income.per.Capita +
    Population + Area.sq.mi + Pop.Density.per.sq.mi + Coast.Area.Ratio +
    Net.migration + Infant.Mortality.per.1000.births + GDP.per.capita +
    Literacy.percent + Phones.per.1000.people + Arable.percent +
    Crops.percent + Other.Land.Use.percent + Climate + Birthrate +
    Deathrate + Agriculture + Industry + Service + Region.AusNZ +
    Region.Cen.E.Eur + Region.E.Asia + Region.LatCari + Region.MENA +
    Region.N.Amer + Region.SE.Asia + Region.S.Asia + Region.SS.Africa, data =
world_norm_dist)
```

Now lets looks at the Varaince Inflation Factor numbers to get a sense of multicollinearity.

```r
#install.packages("car")
library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:psych':
##
##     logit

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

round(vif(m1),2)

##                         HDI.Score         Life.Expectancy.at.Birth
##                            467.72                            32.33
##      Expected.Years.of.Education         Mean.Years.of.Education
##                             23.64                            43.46
## Gross.National.Income.per.Capita                       Population
##                             96.31                            72.51
##                        Area.sq.mi          Pop.Density.per.sq.mi
##                            117.14                            61.54
##                  Coast.Area.Ratio                    Net.migration
##                              2.08                             2.51
## Infant.Mortality.per.1000.births                   GDP.per.capita
##                             20.08                            17.41
##                  Literacy.percent           Phones.per.1000.people
##                             11.75                            17.02
##                    Arable.percent                    Crops.percent
##                             23.39                             3.50
##            Other.Land.Use.percent                          Climate
##                             21.80                             2.56
##                         Birthrate                        Deathrate
##                             15.70                             6.35
##                       Agriculture                         Industry
##                             60.83                            28.34
##                           Service                     Region.AusNZ
##                             36.70                             1.52
##                  Region.Cen.E.Eur                    Region.E.Asia
##                              5.20                             1.75
##                    Region.LatCari                      Region.MENA
##                              4.72                             5.33
##                     Region.N.Amer                   Region.SE.Asia
```

```
##                              1.68                             2.80
##                      Region.S.Asia                 Region.SS.Africa
##                              3.21                             9.82
```

Here we can see that several features have very high VIFs. This signals that features explain eachother and there is multicollinearity. However, it is important to note that multicollinearity can sometimes be ignored, if the collinearity does not affect statistical significance. For example, "If your model has x, z, and xz, both x and z are likely to be highly correlated with their product. This is not something to be concerned about, however, because the p-value for xz is not affected by the multicollinearity."[___]. It is not always reason for alarm when features are derived from eachother. It makes sense that they would explain eachother, yet they don't affect p-values.

Yet, the HDI.Score VIF is extremely high. We saw high correlations earlier in the correlation matrix too. Because HDI.Score is a direct calculation from every other feature in the HDI datset, it is explain by all the other features. I am going to remove HDI.Score.

```
#make model for features
m2 <- lm(Happiness.Score ~ Life.Expectancy.at.Birth +
Expected.Years.of.Education +
    Mean.Years.of.Education + Gross.National.Income.per.Capita +
    Population + Area.sq.mi + Pop.Density.per.sq.mi + Coast.Area.Ratio +
    Net.migration + Infant.Mortality.per.1000.births + GDP.per.capita +
    Literacy.percent + Phones.per.1000.people + Arable.percent +
    Crops.percent + Other.Land.Use.percent + Climate + Birthrate +
    Deathrate + Agriculture + Industry + Service + Region.AusNZ +
    Region.Cen.E.Eur + Region.E.Asia + Region.LatCari + Region.MENA +
    Region.N.Amer + Region.SE.Asia + Region.S.Asia + Region.SS.Africa, data =
world_norm_dist)

round(vif(m2),2)
```

```
##         Life.Expectancy.at.Birth      Expected.Years.of.Education
##                            11.91                             7.45
##          Mean.Years.of.Education Gross.National.Income.per.Capita
##                            14.30                            20.98
##                       Population                       Area.sq.mi
##                            72.31                           117.09
##            Pop.Density.per.sq.mi                 Coast.Area.Ratio
##                            61.47                             2.08
##                    Net.migration Infant.Mortality.per.1000.births
##                             2.51                            17.37
##                   GDP.per.capita                 Literacy.percent
##                            17.34                            11.56
##           Phones.per.1000.people                   Arable.percent
##                            17.02                            23.36
##                    Crops.percent           Other.Land.Use.percent
##                             3.49                            21.75
##                          Climate                        Birthrate
##                             2.47                            15.44
```

```
##                          Deathrate                         Agriculture
##                               5.71                               59.94
##                           Industry                             Service
##                              28.27                               36.40
##                       Region.AusNZ                     Region.Cen.E.Eur
##                               1.52                                4.90
##                      Region.E.Asia                      Region.LatCari
##                               1.75                                4.59
##                        Region.MENA                      Region.N.Amer
##                               5.24                                1.67
##                     Region.SE.Asia                      Region.S.Asia
##                               2.68                                3.02
##                   Region.SS.Africa
##                               9.45
```

Now I am going to remove Area, as it's information is explained by other features such as the land usage % stats.

```
#make model for features
m3 <- lm(Happiness.Score ~ Life.Expectancy.at.Birth +
Expected.Years.of.Education +
    Mean.Years.of.Education + Gross.National.Income.per.Capita +
    Population + Pop.Density.per.sq.mi + Coast.Area.Ratio +
    Net.migration + Infant.Mortality.per.1000.births + GDP.per.capita +
    Literacy.percent + Phones.per.1000.people + Arable.percent +
    Crops.percent + Other.Land.Use.percent + Climate + Birthrate +
    Deathrate + Agriculture + Industry + Service + Region.AusNZ +
    Region.Cen.E.Eur + Region.E.Asia + Region.LatCari + Region.MENA +
    Region.N.Amer + Region.SE.Asia + Region.S.Asia + Region.SS.Africa, data =
world_norm_dist)

round(vif(m3),2)
```

```
##           Life.Expectancy.at.Birth      Expected.Years.of.Education
##                              11.73                               7.44
##            Mean.Years.of.Education Gross.National.Income.per.Capita
##                              14.26                              20.96
##                         Population            Pop.Density.per.sq.mi
##                               1.81                               5.39
##                   Coast.Area.Ratio                      Net.migration
##                               2.07                               2.50
## Infant.Mortality.per.1000.births                      GDP.per.capita
##                              16.93                              17.29
##                   Literacy.percent           Phones.per.1000.people
##                              11.56                              16.83
##                     Arable.percent                      Crops.percent
##                              23.36                               3.40
##             Other.Land.Use.percent                            Climate
##                              21.74                               2.38
##                          Birthrate                          Deathrate
##                              14.49                               5.46
```

```
##                         Agriculture                              Industry
##                               59.55                                 28.00
##                             Service                          Region.AusNZ
##                               36.18                                  1.48
##                      Region.Cen.E.Eur                       Region.E.Asia
##                                4.90                                  1.75
##                      Region.LatCari                           Region.MENA
##                                4.58                                  5.22
##                       Region.N.Amer                        Region.SE.Asia
##                                1.49                                  2.68
##                       Region.S.Asia                      Region.SS.Africa
##                                3.02                                  9.35
```

Agriculture, service, and indsutry are all dependent on one another, so there is no cause for alarm that their VIFs are now the highest. We are going to leave the remaining feature selection to PCA.

Principal component analysis - works best when there is high correlation between variables.. perfect!

```
wdata <-world_norm_dist[,-c(1,2)]
pcal <- princomp(wdata, scores = TRUE, cor = TRUE)
summary(pcal)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4     Comp.5
## Standard deviation     3.363189 1.9512695 1.58377157 1.39397981 1.33750782
## Proportion of Variance 0.353470 0.1189829 0.07838539 0.06072437 0.05590397
## Cumulative Proportion  0.353470 0.4724529 0.55083831 0.61156267 0.66746665
##                            Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     1.27432844 1.08905268 1.06494212 1.02606244
## Proportion of Variance 0.05074728 0.03706362 0.03544068 0.03290013
## Cumulative Proportion  0.71821393 0.75527755 0.79071822 0.82361835
##                           Comp.10    Comp.11    Comp.12    Comp.13
## Standard deviation     1.00227067 0.92221033 0.80675731 0.75109153
## Proportion of Variance 0.03139208 0.02657725 0.02033929 0.01762933
## Cumulative Proportion  0.85501043 0.88158768 0.90192697 0.91955630
##                           Comp.14    Comp.15     Comp.16     Comp.17
## Standard deviation     0.72290785 0.58508640 0.552073197 0.520098268
## Proportion of Variance 0.01633112 0.01069769 0.009524525 0.008453194
## Cumulative Proportion  0.93588742 0.94658511 0.956109632 0.964562826
##                            Comp.18     Comp.19     Comp.20     Comp.21
## Standard deviation     0.463662392 0.434279806 0.419441117 0.329772123
## Proportion of Variance 0.006718213 0.005893717 0.005497839 0.003398427
## Cumulative Proportion  0.971281038 0.977174756 0.982672595 0.986071021
##                            Comp.22     Comp.23     Comp.24     Comp.25
## Standard deviation     0.313791327 0.274180777 0.249352117 0.236167021
## Proportion of Variance 0.003077031 0.002349222 0.001943015 0.001742964
## Cumulative Proportion  0.989148053 0.991497274 0.993440289 0.995183254
##                            Comp.26    Comp.27     Comp.28     Comp.29
```

```
## Standard deviation     0.209379820 0.19678861 0.191040325 0.1456040729
## Proportion of Variance 0.001369997 0.00121018 0.001140513 0.0006625171
## Cumulative Proportion  0.996553251 0.99776343 0.998903943 0.9995664605
##                             Comp.30      Comp.31      Comp.32
## Standard deviation     0.0910302638 0.0630503432 4.014237e-02
## Proportion of Variance 0.0002589534 0.0001242296 5.035655e-05
## Cumulative Proportion  0.9998254139 0.9999496435 1.000000e+00
```

1st component explains 35% of variance in data, 2nd component explains 11% (cumulative 46%).

Eigen values = standard deviation of PCs squared. We could use the first 5, but I'm only going to do 3

All 32 components explain the full variation in the data.

Now let's calculate loadings. These tell us the correlations between each feature and the components. Theoretically, the features most highly correlated to the first few components are the best to use. And the features most correlated with the last components are the ones that are explained by other features.

```
#same thing
pcal$loadings
```

```
##
## Loadings:
##                                  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
## HDI.Score                         0.292
## Life.Expectancy.at.Birth          0.273
## Expected.Years.of.Education       0.268
## Mean.Years.of.Education           0.264         0.154         0.139
## Gross.National.Income.per.Capita  0.272  0.132                       0.112
## Population                                             0.634        -0.194
## Area.sq.mi                               0.229  0.187  0.480  0.160 -0.234
## Pop.Density.per.sq.mi                   -0.358 -0.300        -0.180  0.107
## Coast.Area.Ratio                  0.136        -0.282        -0.223 -0.168
## Net.migration                     0.107  0.117         0.147 -0.470  0.273
## Infant.Mortality.per.1000.births -0.284
## GDP.per.capita                    0.278
## Literacy.percent                  0.250         0.180         0.186
## Phones.per.1000.people            0.286
## Arable.percent                          -0.453         0.185
## Crops.percent                           -0.355 -0.272         0.118
## Other.Land.Use.percent                   0.456        -0.135
## Climate                           0.119 -0.246  0.225               -0.106
## Birthrate                        -0.274
## Deathrate                        -0.123 -0.110  0.429        -0.132  0.189
## Agriculture                      -0.258 -0.115
## Industry                                 0.226 -0.217  0.136  0.340  0.302
## Service                           0.207         0.119 -0.102 -0.247 -0.256
## Region.AusNZ                                                 -0.111
```

```
## Region.Cen.E.Eur                        -0.164  0.263         0.467  0.243
## Region.E.Asia                                          0.126 -0.123 -0.101
## Region.LatCari                                 -0.144 -0.242  0.194 -0.594
## Region.MENA                              0.176 -0.390                0.279
## Region.N.Amer                                   0.136  0.265 -0.119 -0.157
## Region.SE.Asia                                 -0.168  0.162              
## Region.S.Asia                                          0.238 -0.102       
## Region.SS.Africa                 -0.213         0.179        -0.267  0.109
##                                  Comp.7 Comp.8 Comp.9 Comp.10 Comp.11
## HDI.Score                                                            
## Life.Expectancy.at.Birth                                              
## Expected.Years.of.Education                                           
## Mean.Years.of.Education                                               
## Gross.National.Income.per.Capita                              0.127  
## Population                                           -0.140          
## Area.sq.mi                                           -0.125          
## Pop.Density.per.sq.mi                    0.172                        
## Coast.Area.Ratio                 -0.213 -0.155        -0.133   0.185  
## Net.migration                                         0.115          
## Infant.Mortality.per.1000.births                                      
## GDP.per.capita                                                 0.112  
## Literacy.percent                                                      
## Phones.per.1000.people                                                
## Arable.percent                                                        
## Crops.percent                           -0.198        -0.244          
## Other.Land.Use.percent                   0.142                        
## Climate                          -0.193                        0.147  
## Birthrate                                                             
## Deathrate                        -0.103                        0.339  
## Agriculture                                     0.118         -0.246  
## Industry                         -0.153                        0.431  
## Service                           0.155                       -0.120  
## Region.AusNZ                      0.186 -0.407  0.686 -0.283          
## Region.Cen.E.Eur                                              -0.216  
## Region.E.Asia                    -0.180  0.717  0.127 -0.394          
## Region.LatCari                                                 0.238  
## Region.MENA                       0.294        -0.181 -0.227  -0.234  
## Region.N.Amer                           -0.214 -0.523  0.149  -0.354  
## Region.SE.Asia                   -0.665 -0.140  0.257  0.379  -0.287  
## Region.S.Asia                     0.430  0.219  0.233  0.548   0.290  
## Region.SS.Africa                 -0.152 -0.114 -0.142 -0.271   0.163  
##                                  Comp.12 Comp.13 Comp.14 Comp.15 Comp.16
## HDI.Score                                                               
## Life.Expectancy.at.Birth         -0.101   0.137                         
## Expected.Years.of.Education               0.156                  -0.105 
## Mean.Years.of.Education           0.156                   0.201         
## Gross.National.Income.per.Capita                                        
## Population                       -0.209  -0.140           0.255         
## Area.sq.mi                       -0.169          -0.143                 
## Pop.Density.per.sq.mi                    -0.119   0.137   0.218   0.151 
```

```
## Coast.Area.Ratio                  0.310   0.140  -0.585   0.157  -0.414
## Net.migration                    -0.210           0.415   0.294  -0.446
## Infant.Mortality.per.1000.births  0.103
## GDP.per.capita
## Literacy.percent                         -0.130           0.215
## Phones.per.1000.people
## Arable.percent                           -0.121          -0.370  -0.180
## Crops.percent                     0.114           0.199   0.475   0.170
## Other.Land.Use.percent                                    0.396   0.156
## Climate                          -0.317   0.686                   0.257
## Birthrate                                         0.101           0.102
## Deathrate                         0.119
## Agriculture                               0.261                  -0.331
## Industry                          0.200           0.198
## Service                          -0.101  -0.341  -0.208           0.288
## Region.AusNZ                      0.250           0.106  -0.102   0.142
## Region.Cen.E.Eur                  0.155          -0.153   0.217  -0.180
## Region.E.Asia                     0.224   0.210   0.166  -0.187
## Region.LatCari                           -0.129   0.376          -0.245
## Region.MENA                      -0.255   0.122  -0.186
## Region.N.Amer                     0.538   0.182   0.182
## Region.SE.Asia                           -0.148          -0.144
## Region.S.Asia                     0.217   0.123  -0.147
## Region.SS.Africa                                                  0.232
##                                  Comp.17 Comp.18 Comp.19 Comp.20 Comp.21
## HDI.Score                                         0.108
## Life.Expectancy.at.Birth         -0.118           0.277  -0.164  -0.125
## Expected.Years.of.Education       0.363   0.203           0.179  -0.684
## Mean.Years.of.Education           0.249   0.239           0.226   0.282
## Gross.National.Income.per.Capita                         -0.117   0.250
## Population                       -0.191   0.132   0.101   0.102
## Area.sq.mi                                               -0.131
## Pop.Density.per.sq.mi            -0.324   0.243   0.231   0.292  -0.123
## Coast.Area.Ratio                                 -0.115
## Net.migration                                    -0.316
## Infant.Mortality.per.1000.births  0.120                   0.255
## GDP.per.capita                           -0.110   0.185  -0.308   0.200
## Literacy.percent                  0.347   0.187           0.206   0.207
## Phones.per.1000.people                   -0.132   0.252
## Arable.percent
## Crops.percent                     0.328  -0.358          -0.315
## Other.Land.Use.percent           -0.162           0.157          -0.121
## Climate                          -0.140          -0.196   0.180   0.187
## Birthrate                         0.128   0.244  -0.153   0.132   0.260
## Deathrate                                -0.489   0.329   0.334
## Agriculture                       0.126   0.130   0.337  -0.119
## Industry                                         -0.188
## Service                                  -0.213  -0.382
## Region.AusNZ                     -0.250                   0.118
## Region.Cen.E.Eur                 -0.389          -0.293  -0.102  -0.160
```

```
## Region.E.Asia                            -0.135  -0.152
## Region.LatCari                   -0.120                   0.182
## Region.MENA                              -0.217           0.374
## Region.N.Amer
## Region.SE.Asia                    0.119  -0.181                  -0.103
## Region.S.Asia                     0.191
## Region.SS.Africa                          0.358          -0.242  -0.237
##                                  Comp.22 Comp.23 Comp.24 Comp.25 Comp.26
## HDI.Score                         0.105   0.223
## Life.Expectancy.at.Birth         -0.503   0.427           0.356   0.304
## Expected.Years.of.Education       0.255   0.183                  -0.171
## Mean.Years.of.Education                           0.449           0.249
## Gross.National.Income.per.Capita  0.396   0.166  -0.207   0.194
## Population
## Area.sq.mi
## Pop.Density.per.sq.mi             0.103
## Coast.Area.Ratio
## Net.migration
## Infant.Mortality.per.1000.births  0.157   0.226  -0.499   0.480
## GDP.per.capita                    0.405                   0.176  -0.158
## Literacy.percent                 -0.306  -0.298  -0.331   0.165  -0.143
## Phones.per.1000.people                   -0.145  -0.549  -0.516   0.340
## Arable.percent                                            0.160
## Crops.percent                             0.101
## Other.Land.Use.percent                                           -0.136
## Climate
## Birthrate                         0.112   0.514          -0.336   0.257
## Deathrate                                 0.152   0.115           0.116
## Agriculture                       0.120
## Industry                         -0.268
## Service
## Region.AusNZ
## Region.Cen.E.Eur                  0.223                           0.222
## Region.E.Asia                                                     0.115
## Region.LatCari                    0.169  -0.164   0.106   0.141   0.291
## Region.MENA                              -0.234   0.107   0.148   0.257
## Region.N.Amer
## Region.SE.Asia                    0.101                           0.211
## Region.S.Asia                            -0.205                   0.202
## Region.SS.Africa                         -0.262           0.253   0.467
##                                  Comp.27 Comp.28 Comp.29 Comp.30 Comp.31
## HDI.Score                                 0.185           0.122
## Life.Expectancy.at.Birth                 -0.108
## Expected.Years.of.Education
## Mean.Years.of.Education           0.446  -0.151
## Gross.National.Income.per.Capita          0.588   0.150
## Population                                                       -0.526
## Area.sq.mi                                               -0.144   0.673
## Pop.Density.per.sq.mi                            -0.105           0.482
## Coast.Area.Ratio
```

```
## Net.migration
## Infant.Mortality.per.1000.births  0.396  -0.231  -0.156
## GDP.per.capita                   -0.255  -0.603  -0.163
## Literacy.percent                 -0.427
## Phones.per.1000.people            0.250  -0.166   0.116
## Arable.percent                           -0.156   0.665
## Crops.percent                                     0.113
## Other.Land.Use.percent                   -0.193   0.615
## Climate
## Birthrate                        -0.414  -0.197
## Deathrate                        -0.215
## Agriculture                                               0.668   0.109
## Industry                                 -0.126           0.446
## Service                                                   0.515
## Region.AusNZ
## Region.Cen.E.Eur                 -0.198
## Region.E.Asia
## Region.LatCari
## Region.MENA                      -0.154
## Region.N.Amer
## Region.SE.Asia
## Region.S.Asia
## Region.SS.Africa
##                                  Comp.32
## HDI.Score                         0.864
## Life.Expectancy.at.Birth         -0.183
## Expected.Years.of.Education      -0.161
## Mean.Years.of.Education          -0.221
## Gross.National.Income.per.Capita -0.360
## Population
## Area.sq.mi
## Pop.Density.per.sq.mi
## Coast.Area.Ratio
## Net.migration
## Infant.Mortality.per.1000.births
## GDP.per.capita
## Literacy.percent
## Phones.per.1000.people
## Arable.percent
## Crops.percent
## Other.Land.Use.percent
## Climate
## Birthrate
## Deathrate
## Agriculture
## Industry
## Service
## Region.AusNZ
## Region.Cen.E.Eur
## Region.E.Asia
```

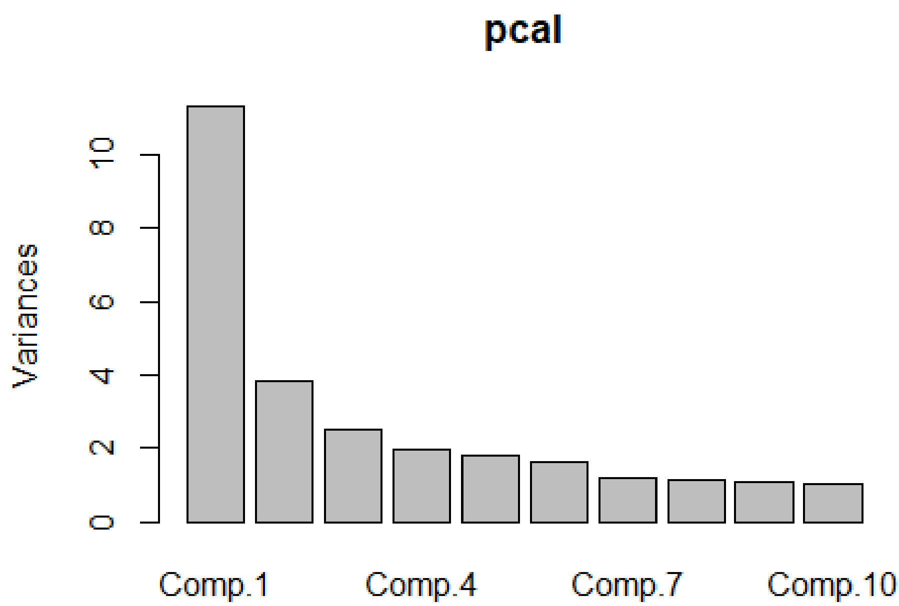
```
## Region.LatCari
## Region.MENA
## Region.N.Amer
## Region.SE.Asia
## Region.S.Asia
## Region.SS.Africa
##
##                 Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings      1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.031  0.031  0.031  0.031  0.031  0.031  0.031  0.031
## Cumulative Var  0.031  0.062  0.094  0.125  0.156  0.187  0.219  0.250
##                 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13 Comp.14 Comp.15
## SS loadings      1.000   1.000   1.000   1.000   1.000   1.000   1.000
## Proportion Var  0.031   0.031   0.031   0.031   0.031   0.031   0.031
## Cumulative Var  0.281   0.312   0.344   0.375   0.406   0.437   0.469
##                 Comp.16 Comp.17 Comp.18 Comp.19 Comp.20 Comp.21 Comp.22
## SS loadings       1.000   1.000   1.000   1.000   1.000   1.000   1.000
## Proportion Var    0.031   0.031   0.031   0.031   0.031   0.031   0.031
## Cumulative Var    0.500   0.531   0.562   0.594   0.625   0.656   0.687
##                 Comp.23 Comp.24 Comp.25 Comp.26 Comp.27 Comp.28 Comp.29
## SS loadings       1.000   1.000   1.000   1.000   1.000   1.000   1.000
## Proportion Var    0.031   0.031   0.031   0.031   0.031   0.031   0.031
## Cumulative Var    0.719   0.750   0.781   0.812   0.844   0.875   0.906
##                 Comp.30 Comp.31 Comp.32
## SS loadings       1.000   1.000   1.000
## Proportion Var    0.031   0.031   0.031
## Cumulative Var    0.938   0.969   1.000
```

This tells a similar story as before because HDI score is HIGHLY correlated with Component 32 (.864). Area sq mi is highly correlated with comp 31 (.673). Then agriculture would be next since it has a high correlation with Component 30. Other land use and arable percent are also multicollinear. This all makes sense logically because these features are mutually exclusive and are dependent on each others' calculated value.

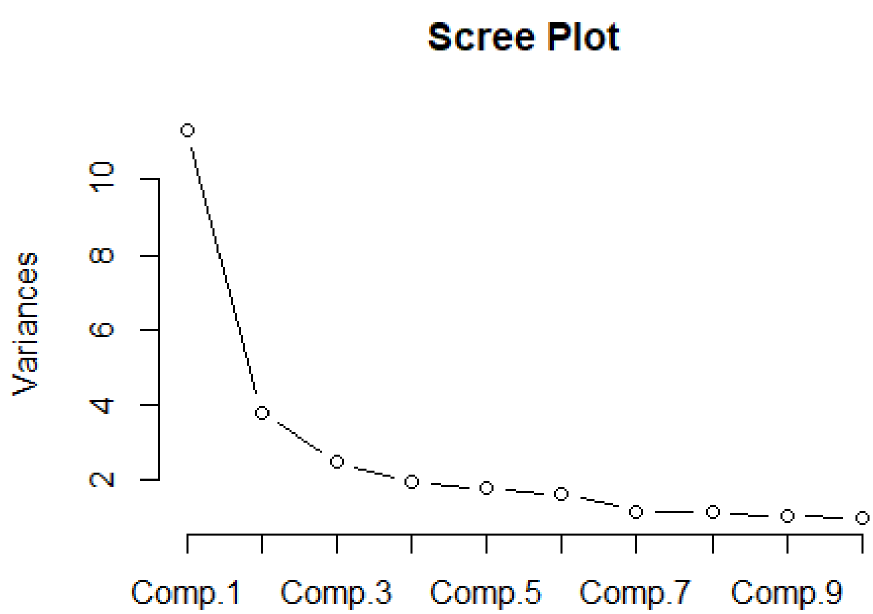
DISCLAIMER: It is also important to note that Region.Western.Europe is excluded from the model since it is explained by the intercept (0 values in all other region dummies).

Let's now look at the scree plot of the eigenvalues

```
plot(pcal)
```



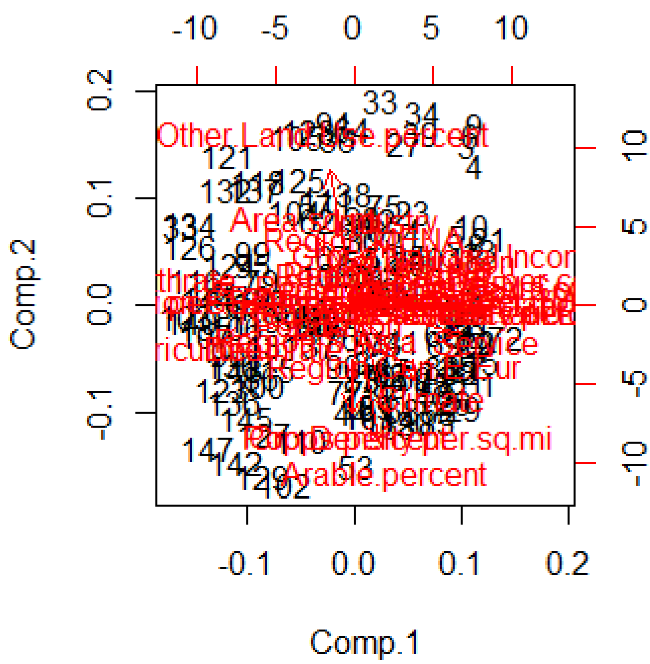

```
screeplot(pcal, type = "line", main = "Scree Plot")
```



This shows us the importance of the first few components.

Now a biplot of score variables

```r
biplot(pcal)
```



This is too messy to read.

Scores of the components.

```r
pcal$scores[1:5,]
```

```
##     Comp.1     Comp.2    Comp.3      Comp.4     Comp.5      Comp.6
## 1 5.175588 -1.7033081 0.6554601 -0.02074137 -1.6199347  0.46277862
## 2 4.604376  0.2358656 0.8396040 -0.15183043 -0.7211958  1.07756950
## 3 4.299446  3.5512105 2.1296445 -2.52313167 -1.8961323 -0.44785387
## 4 4.577120  3.1169056 1.5314561 -0.34496318 -0.7133152  0.53488547
## 5 3.883556  1.5548817 1.5785306 -0.32933855 -0.7237401 -0.05012623
##         Comp.7      Comp.8      Comp.9      Comp.10    Comp.11    Comp.12
## 1 -0.047744728 -0.57281011 -0.33778276 -0.007054186  0.8779250 -0.4306642
## 2  0.079534176  0.35338573 -0.07631097  0.730619414  0.2826630 -1.1179809
## 3 -0.000828971  0.32857429  0.32623112  1.581041944 -0.5592776 -1.0686336
## 4 -0.772130933  0.18514636 -0.05790453  1.024657218  1.3047392 -0.5150073
## 5 -0.218092300  0.07668175 -0.05243349  0.638292556  0.6107551 -0.9818318
##       Comp.13    Comp.14    Comp.15    Comp.16      Comp.17     Comp.18
## 1 -0.3089150 -0.4740435 -1.0677930 -0.9450420  0.631270380  0.55532814
## 2 -0.1020641  1.6584938  0.4833660  1.0431751 -0.004716887  0.05374748
## 3  1.1396300 -1.5798728 -0.5794377 -0.6987329  0.635245393  0.37896393
## 4  0.7191781 -0.9934749 -0.5096451 -0.5740363 -0.233521139  0.95610525
## 5  0.3747044 -0.6379355 -0.3903499  0.0555760  0.191661105 -0.03838195
```

```
##        Comp.19    Comp.20     Comp.21     Comp.22      Comp.23      Comp.24
## 1 -0.16705639  0.4728736  0.28404524  0.06083372  0.355299310  0.008218471
## 2  0.79858255 -0.1922257  0.22269628 -0.05500484 -0.007848431 -0.010451781
## 3 -0.09412429 -0.2756977 -0.16548488  0.06679950  0.062321739 -0.421059191
## 4  0.41156465  0.3899458  0.48388236 -0.08472126  0.148797565  0.173668314
## 5  0.27562548 -0.1482611 -0.01099987 -0.25292171 -0.010647661  0.049841697
##        Comp.25     Comp.26     Comp.27     Comp.28     Comp.29     Comp.30
## 1 -0.08985138 -0.02409604  0.11476269 -0.06856710  0.05962942 -0.05324456
## 2 -0.15617129 -0.01171687  0.45463180 -0.01594777  0.04496703  0.08564744
## 3 -0.53697215 -0.01674172 -0.09190221 -0.09469452 -0.55018409  0.07806191
## 4 -0.07986937 -0.10266081 -0.01187825 -0.04270207 -0.11773428  0.08465605
## 5 -0.09301560 -0.40645393 -0.21577262  0.06626700  0.21934331 -0.01414491
##        Comp.31      Comp.32
## 1 -0.03809410  0.002925537
## 2 -0.02789215  0.035079381
## 3 -0.00399339 -0.031871110
## 4 -0.04024532  0.033203837
## 5 -0.03195102 -0.012052959
```

This also shows the relative importance of the different components.

I am going to start back-fitting my model using statistical significance (p-value) in order to find our final regression model to predict happiness score. The VIF analysis and PCA has led me to exclude HDI.Score and Area.sq.mi from my regression model equation. Feature removal is no small decision, but the justification behind this one is multicollinearity. The rest of the feature selection will be done by backfitting by p-value. Any feature that has a p-value of greater than 0.05 will be considered not statistically significant.

```
#remove HDI score and Area Sq mi
world_norm_dist2 <- world_norm_dist[-c(3,9)]
```