What makes people happy?

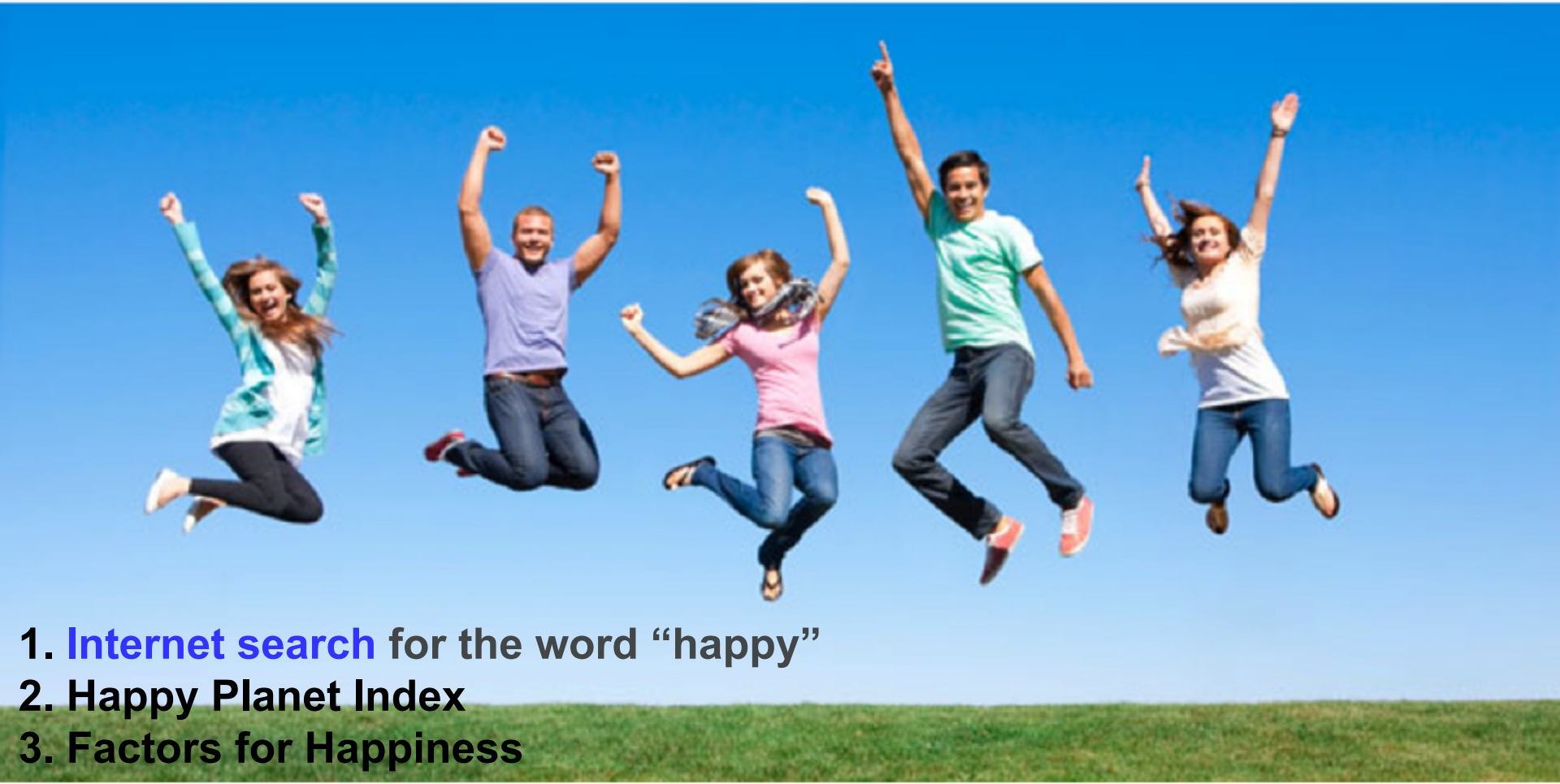
Life expectancy, Wellbeing, Travel, Family, Health, Love, … → **Happiness**

https://www.psychologies.co.uk/tests/happy.html

TEST: Are you ready to be …

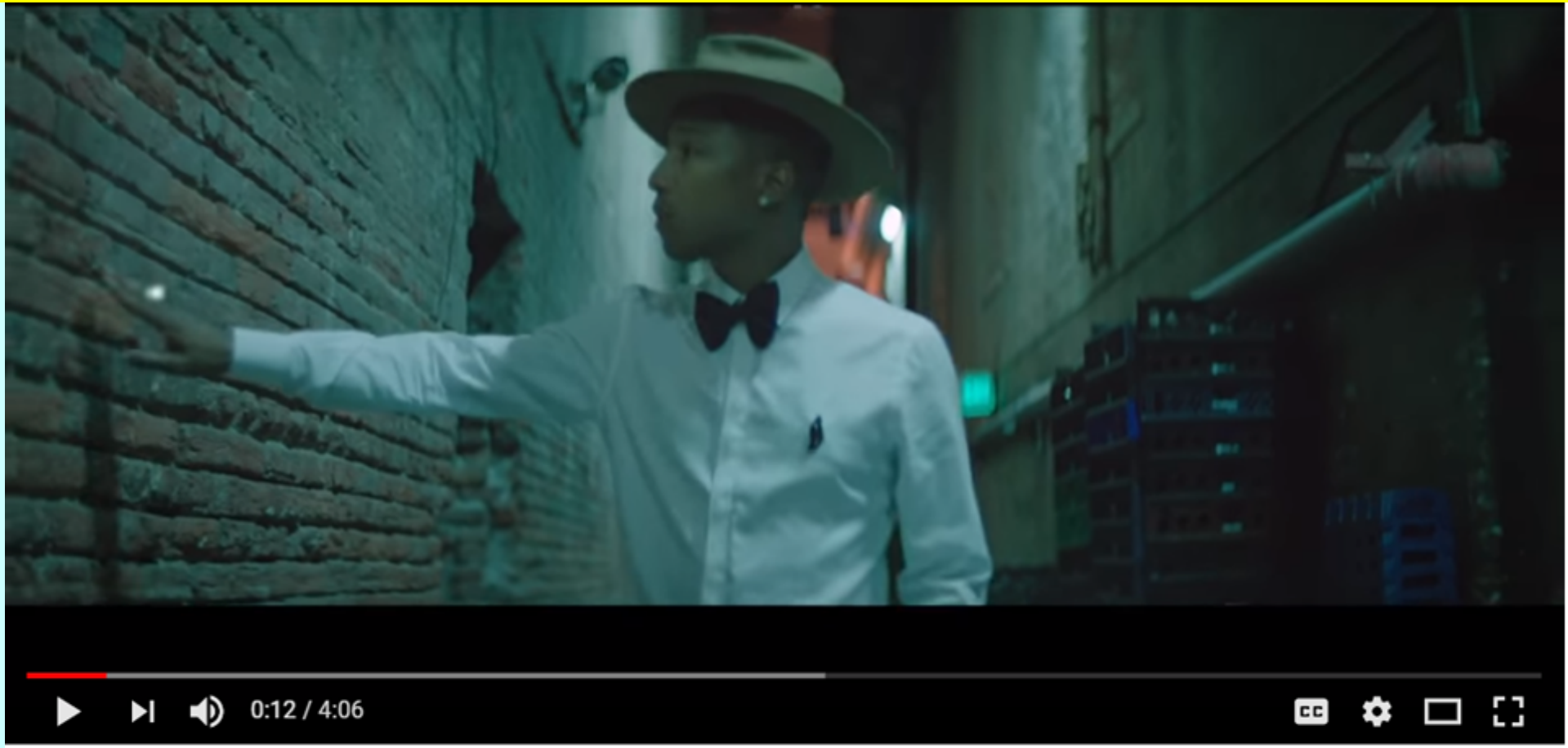
TEST: Are you ready to be happy?

Is happiness there for you to grasp with both hands? Or are you so preoccupied



1. Internet search for the word "happy"
2. Happy Planet Index
3. Factors for Happiness

1. Internet searches for the word "happy" (Pharrell Williams song)



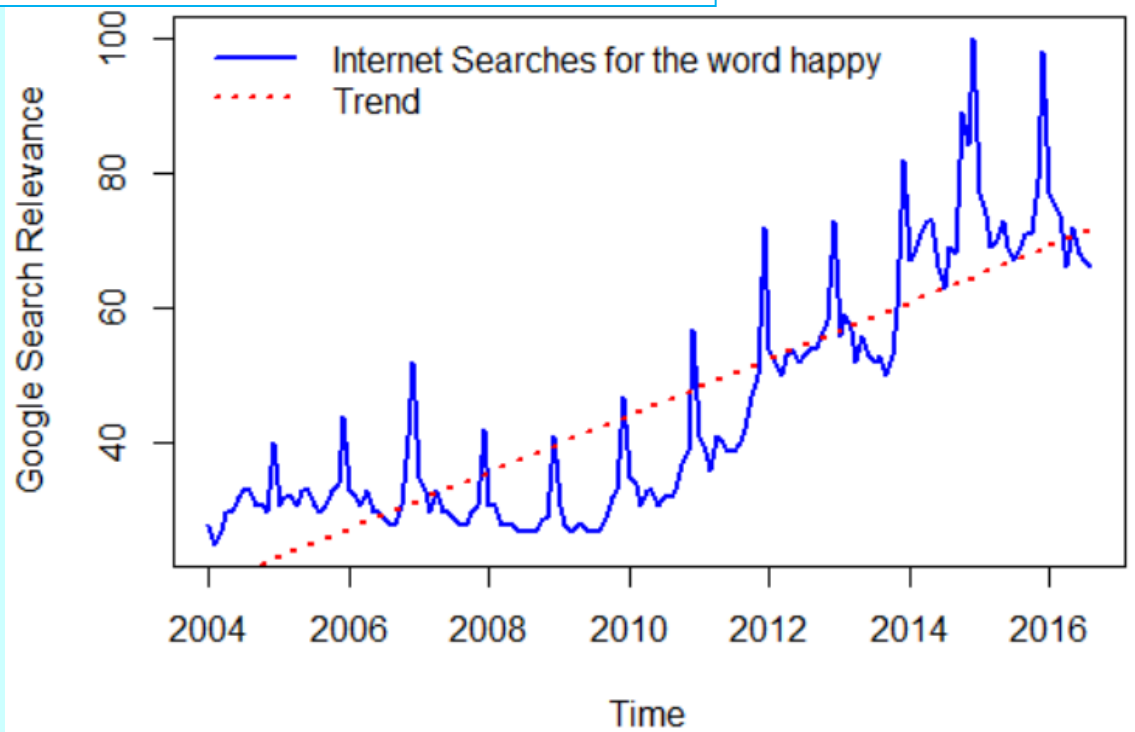
Folks are doing **internet searches** for the word "happy" more than ever before. And no, the **Pharrell Williams** song released late in 2013 isn't single handedly driving the interest. The upward trend clearly started before the song was released!

Ref: https://blog.plot.ly/post/148975591782/the-data-behind-happiness

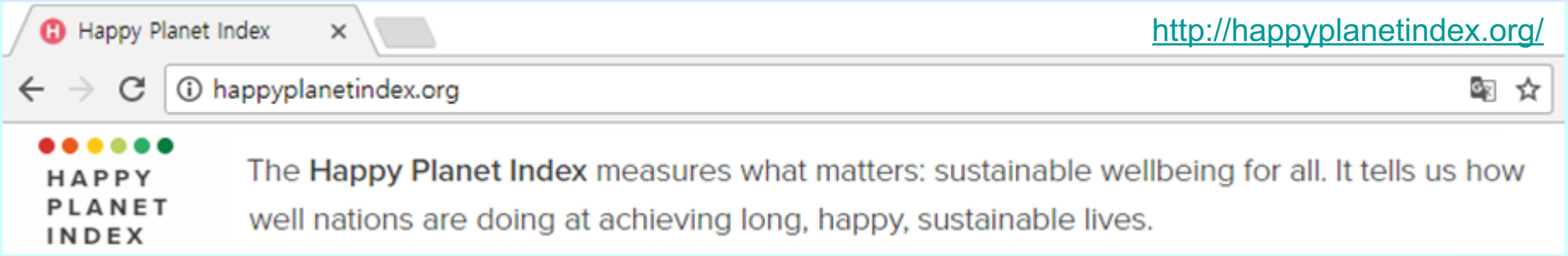
►Internet Searches Trend for the word "happy" (Pharrell Williams song)

```
> #[Data Source] https://plot.ly/~Dreamshot/8235/happy/
> load('HappyTrendData.RData')
> str(happy_trend)
'data.frame':   152 obs. of  2 variables:
 $ Time : Date, format: "2004-01-01" "2004-02-01" ...
 $ Happy: num  28 25 27 30 30 32 33 33 31 31 ...
```

```
plot(happy_trend, type='l', col='blue', lwd=2,
     ylab='Google Search Relevance')
abline(lm(Happy~Time, data=happy_trend), col='red', lty=3, lwd=2)
legend('topleft', col=c('blue','red'),
 lty=c(1,3), lwd=c(2,2), bty="n",
 legend=c('Internet Searches for the word happy','Trend'))
```

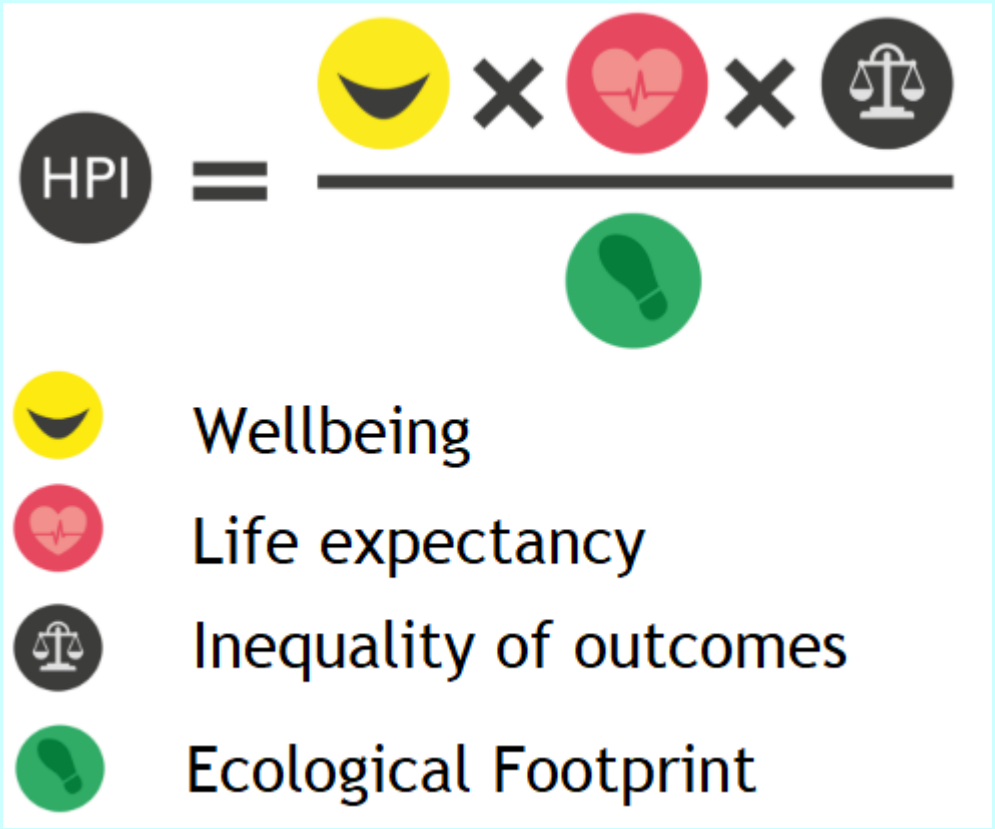


2. Happy Planet Index



http://happyplanetindex.org/

How is the Happy Planet Index calculated?



http://happyplanetindex.org/about

(1) Happy Planet Index (HPI) Data

```
> #[Data Source] http://happyplanetindex.org/about
> HP <- read.csv('HappyIndex.csv', header=TRUE)
> str(HP)
'data.frame':   140 obs. of  14 variables:
 $ HPI.Rank                          : int  110 13 30 19 73 105 43 8 102 87 ...
 $ Country                           : Factor w/ 140 levels "Afghanistan",..: 1 2 3 4
 $ Region                            : Factor w/ 6 levels "Americas","Asia Pacific",..
 $ Average.Life..Expectancy          : num  59.7 77.3 74.3 75.9 74.4 82.1 81 70.8 70.
 $ Average.Wellbeing..0.10.          : num  3.8 5.5 5.6 6.5 4.3 7.2 7.4 4.7 5.7 6.9 .
 $ Happy.Life.Years                  : num  12.4 34.4 30.5 40.2 24 53.1 54.4 23.3 34
 $ Footprint..gha.capita.            : num  0.8 2.2 2.1 3.1 2.2 9.3 6.1 0.7 5.1 7.4 .
 $ Inequality.of.Outcomes            : Factor w/ 44 levels "10%","11%","12%",..: 33 8
 $ Inequality.adjusted.Life.Expectancy: num  38.3 69.7 60.5 68.3 66.9 78.6 78 56.6 66.
 $ Inequality.adjusted.Wellbeing     : num  3.4 5.1 5.2 6 3.7 6.9 7.1 4.3 5.3 6.6 ...
 $ Happy.Planet.Index                : num  20.2 36.8 33.3 35.2 25.7 21.2 30.5 38.4 2
 $ X.GDP.capita...PPP..              : Factor w/ 140 levels " $1,019 "," $1,159 ",..:
 $ Population                        : Factor w/ 140 levels " 1,129,303 ",..: 65 48 8
 $ GINI.index                        : Factor w/ 58 levels "24.7","25.6",..: 58 12 58
```

Creating a data frame with Rank, … , HPI columns

```
> HD <- data.frame(Rank=HP[,1],Country=HP[,2],LifeExpectancy=HP[,4],Wellbeing=HP[,5],
+                  Footprint=HP[,7],InequalityOutcome=HP[,8],HPI=HP[,11])
> head(HD)
  Rank     Country LifeExpectancy Wellbeing Footprint InequalityOutcome  HPI
1  110 Afghanistan           59.7       3.8       0.8               43% 20.2
2   13     Albania           77.3       5.5       2.2               17% 36.8
3   30     Algeria           74.3       5.6       2.1               24% 33.3
4   19   Argentina           75.9       6.5       3.1               16% 35.2
5   73     Armenia           74.4       4.3       2.2               22% 25.7
6  105   Australia           82.1       7.2       9.3                8% 21.2
```

Removing "%" character from InequalityOutcome variable

```
#InequalityOutcome: remove "%"
HD$InequalityOutcome <- sapply(HD$InequalityOutcome,FUN=function(x)
  as.character(gsub("%","",as.character(x),fixed=TRUE)))
HD$InequalityOutcome <- as.numeric(HD$InequalityOutcome)
```

#sub(pattern, replacement, x): pattern -> replacement in x

Sorting data by Rank

```
> HRank <- HD[order(HD$Rank,decreasing=FALSE),]
> dim(HRank)
[1] 140   7
> head(HRank)
    Rank    Country LifeExpectancy Wellbeing Footprint InequalityOutcome  HPI
29     1 Costa Rica           79.1       7.3       2.8                15 44.7
80     2     Mexico           76.4       7.3       2.9                19 40.7
27     3   Colombia           73.7       6.4       1.9                24 40.7
135    4    Vanuatu           71.3       6.5       1.9                22 40.6
137    5    Vietnam           75.5       5.5       1.7                19 40.3
97     6     Panama           77.2       6.9       2.8                19 39.5
> tail(HRank)
    Rank       Country LifeExpectancy Wellbeing Footprint InequalityOutcome  HPI
30   135 Cote d'Ivoire           50.8       3.8       1.3                45 14.4
81   136      Mongolia           68.6       4.9       6.1                22 14.3
12   137         Benin           59.2       3.2       1.4                44 13.4
124  138          Togo           58.6       2.9       1.1                43 13.2
73   139    Luxembourg           81.1       7.0      15.8                 7 13.2
24   140          Chad           50.8       4.0       1.5                51 12.8
```

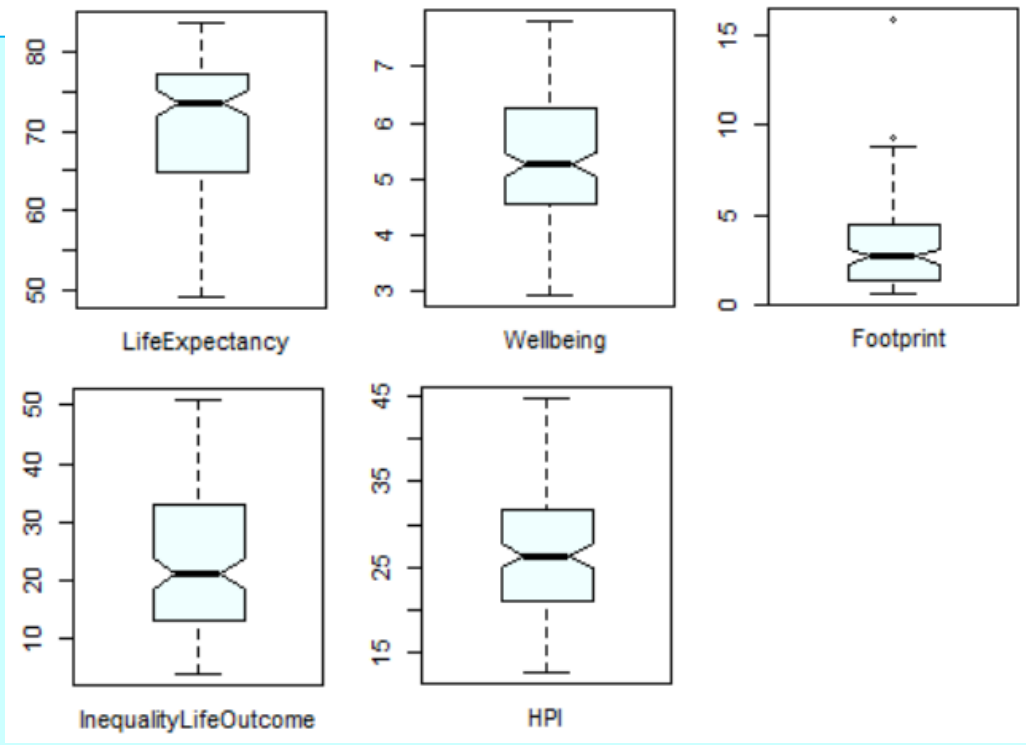
https://www.r-bloggers.com/apply-lapply-rapply-sapply-functions-in-r/

(2) Descriptive Statistics

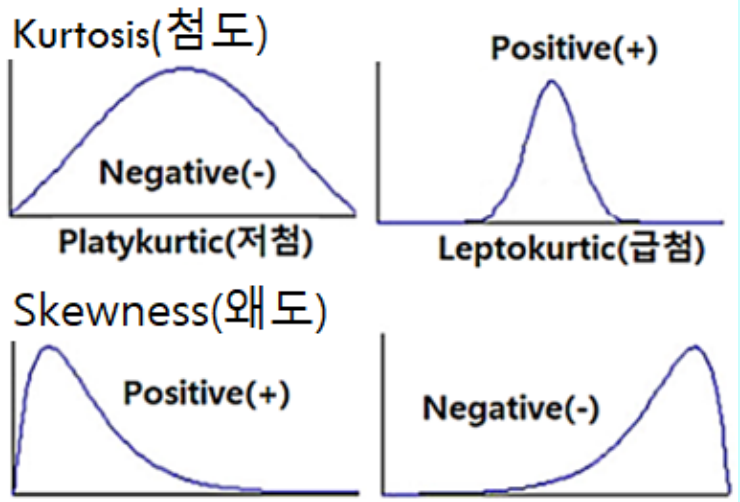
```
> summary(HD[,3:7])
 LifeExpectancy    Wellbeing       Footprint      InequalityOutcome      HPI       
 Min.   :48.90   Min.   :2.900   Min.   : 0.60   Min.   : 4.0      Min.   :12.80  
 1st Qu.:65.03   1st Qu.:4.575   1st Qu.: 1.40   1st Qu.:13.0      1st Qu.:21.18  
 Median :73.50   Median :5.250   Median : 2.70   Median :21.0      Median :26.30  
 Mean   :70.92   Mean   :5.408   Mean   : 3.26   Mean   :23.3      Mean   :26.41  
 3rd Qu.:77.05   3rd Qu.:6.225   3rd Qu.: 4.45   3rd Qu.:33.0      3rd Qu.:31.55  
 Max.   :83.60   Max.   :7.800   Max.   :15.80   Max.   :51.0      Max.   :44.70  
```

```
##Boxplot
xa = c("LifeExpectancy","Wellbeing","Footprint",
       "InequalityLifeOutcome","HPI")
par(mfrow=c(2,3))
for(i in 3:7) boxplot(HD[,i],notch=TRUE,col='azure',xlab=xa[i-2])
par(mfrow=c(1,1))
```

#notch=TRUE will show the confidence interval around the median



Quantity	Definition	R function
Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	mean(x)
Standard Deviation	$s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2}$	sd(x)
Standard Error	$se = \frac{s}{\sqrt{n}}$	sd(x)/sqrt(n)



median absolute deviation (MAD) $= \text{median}(|X_i - \tilde{X}|)$

$$\tilde{X} = \text{median}(X)$$

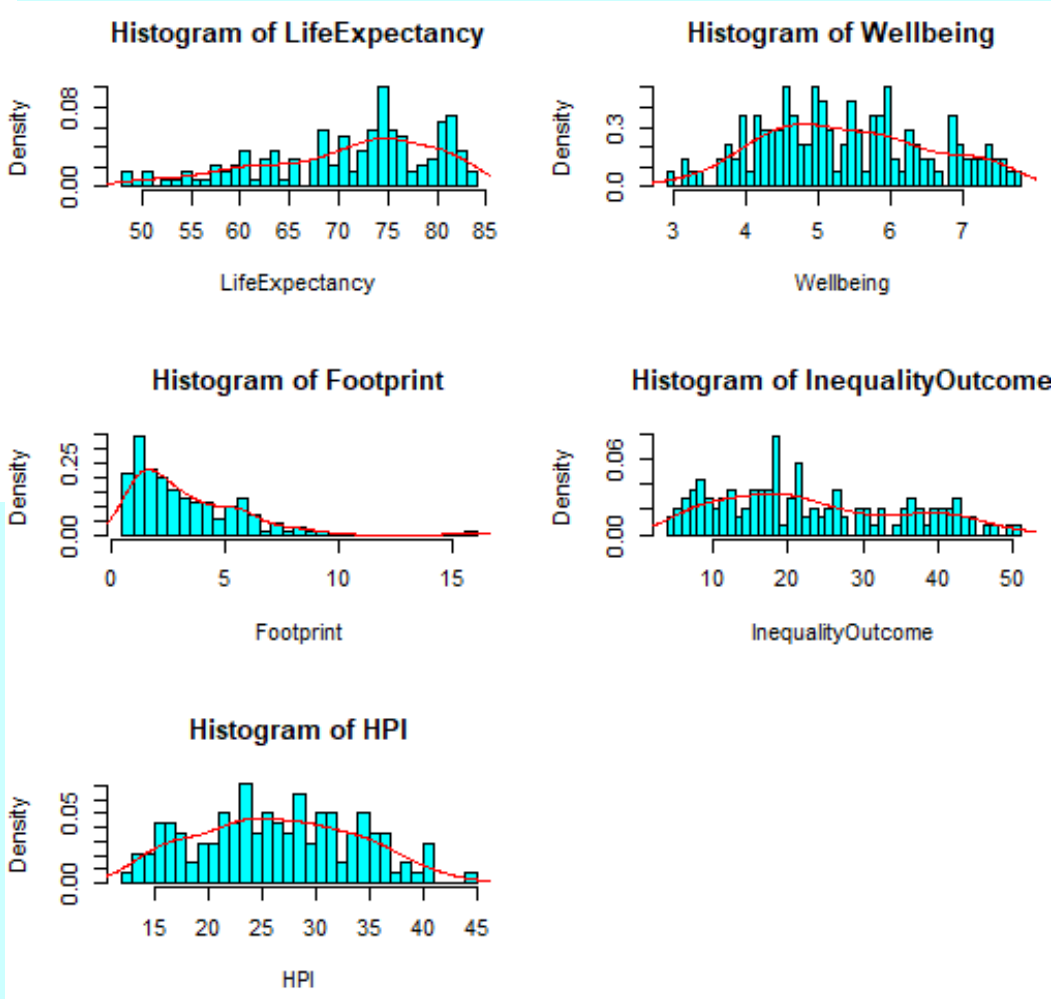
After removing the specified outlier observations, the **trimmed** mean is found using a standard arithmetic averaging formula.

```
> mad(HD[,3])
[1] 8.8956
> mean(DescTools::Trim(HD[,3]))
[1] 71.70357
```

```
> ##Describe
> library(psych)
> describe(HD[,3:7])
                   vars   n  mean    sd median trimmed   mad  min  max range  skew kurtosis   se
LifeExpectancy        1 140 70.92  8.75  73.50   71.70  8.90 48.9 83.6  34.7 -0.68    -0.39 0.74
Wellbeing             2 140  5.41  1.15   5.25    5.38  1.11  2.9  7.8   4.9  0.19    -0.79 0.10
Footprint             3 140  3.26  2.30   2.70    2.96  2.08  0.6 15.8  15.2  1.67     5.01 0.19
InequalityOutcome     4 140 23.30 12.12  21.00   22.65 13.34  4.0 51.0  47.0  0.44    -0.92 1.02
HPI                   5 140 26.41  7.32  26.30   26.31  7.71 12.8 44.7  31.9  0.11    -0.80 0.62
```

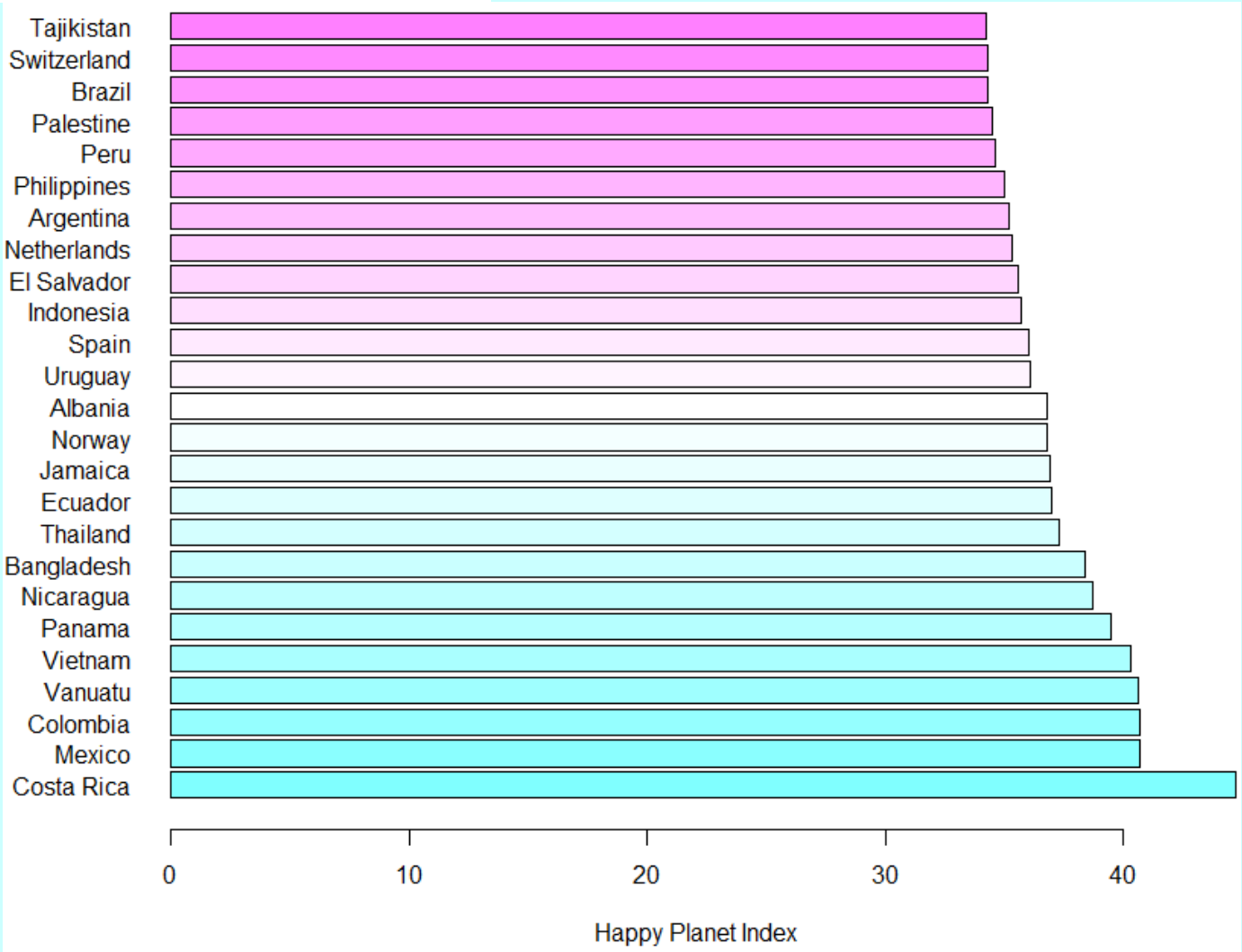
Density Distributions

```
attach(HD); par(mfrow=c(3,2))
hist(LifeExpectancy,breaks=40,
     freq=FALSE,col='cyan')
lines(density(LifeExpectancy),col=2)
hist(Wellbeing,breaks=40,
     freq=FALSE,col='cyan')
lines(density(Wellbeing),col=2)
hist(Footprint,breaks=40,
     freq=FALSE,col='cyan')
lines(density(Footprint),col=2)
hist(InequalityOutcome,breaks=40,
     freq=FALSE,col='cyan')
lines(density(InequalityOutcome),col=2)
hist(HPI,breaks=40,
     freq=FALSE,col='cyan')
lines(density(HPI),col=2)
par(mfrow=c(1,1)); detach(HD)
```



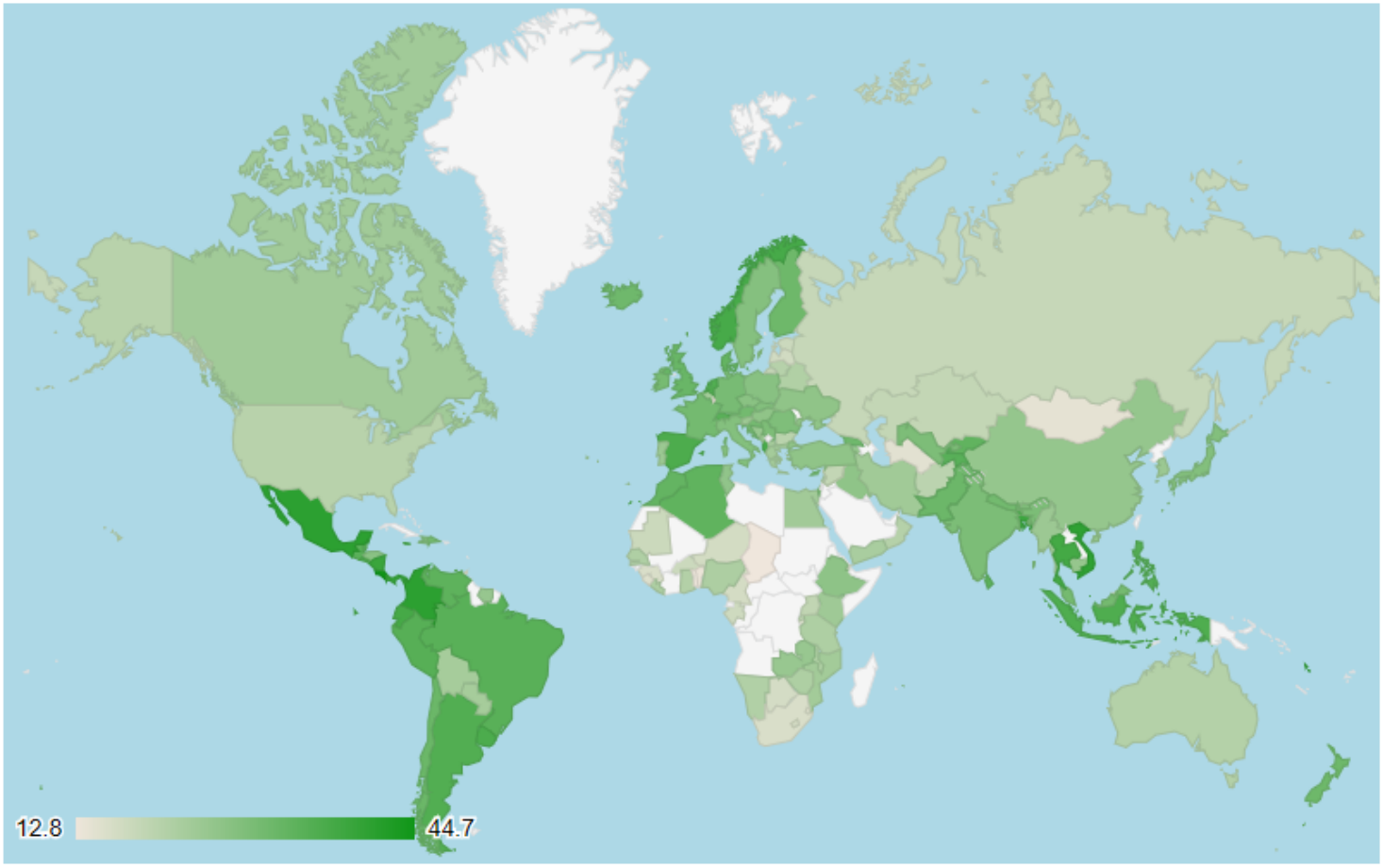
Bar Chart of 25 Higher HPI Countries

```
HRank25 <- subset(HRank, Rank<26)
par(mar=c(4.5,5.5,0.2,0.5))
barplot(HRank25$HPI, names.arg=HRank25$Country,
      horiz=TRUE, col=cm.colors(25), las=1,
      cex.names=1, xlab="Happy Planet Index")
```



(3) HPI Map

```
library(googleVis)
GC <- gvisGeoChart(HD, locationvar='Country', colorvar='HPI',
                   options=list(width=800,height=500,
                   backgroundColor='lightblue'))
GT <- gvisTable(HRank[,c(1,2,7)],options=list(width=250,height=500))
plot(gvisMerge(GC,GT,horizontal=TRUE))
```

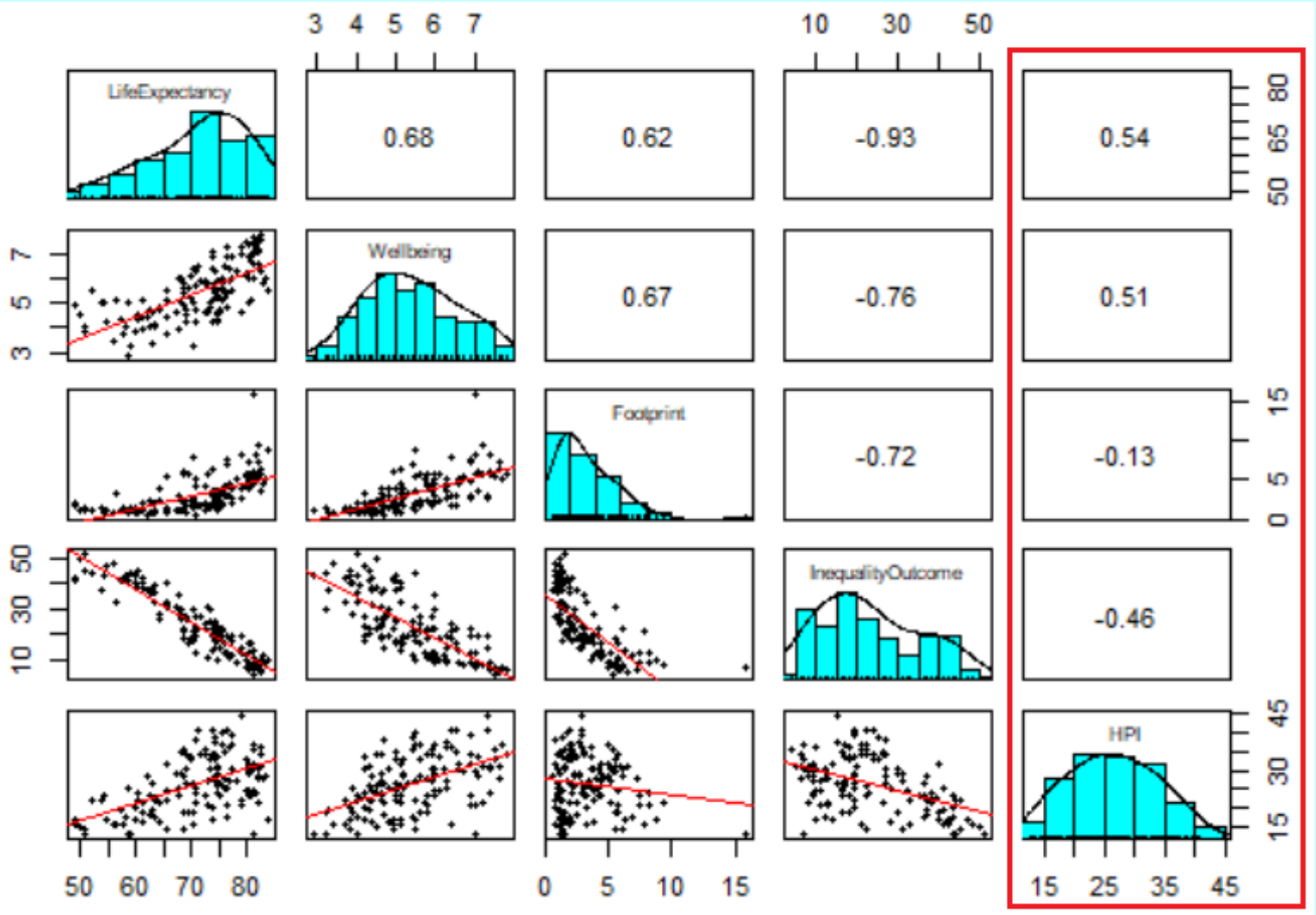


Rank	Country	HPI
1	Costa Rica	44.7
2	Mexico	40.7
3	Colombia	40.7
4	Vanuatu	40.6
5	Vietnam	40.3
6	Panama	39.5
7	Nicaragua	38.7
8	Bangladesh	38.4
9	Thailand	37.3
10	Ecuador	37
11	Jamaica	36.9
12	Norway	36.8
13	Albania	36.8
14	Uruguay	36.1
15	Spain	36
16	Indonesia	35.7
17	El Salvador	35.6
18	Netherlands	35.3
19	Argentina	35.2
20	Philippines	35
21	Peru	34.6
22	Palestine	34.5
23	Brazil	34.3

(1) Correlations

3. Factors for Happiness

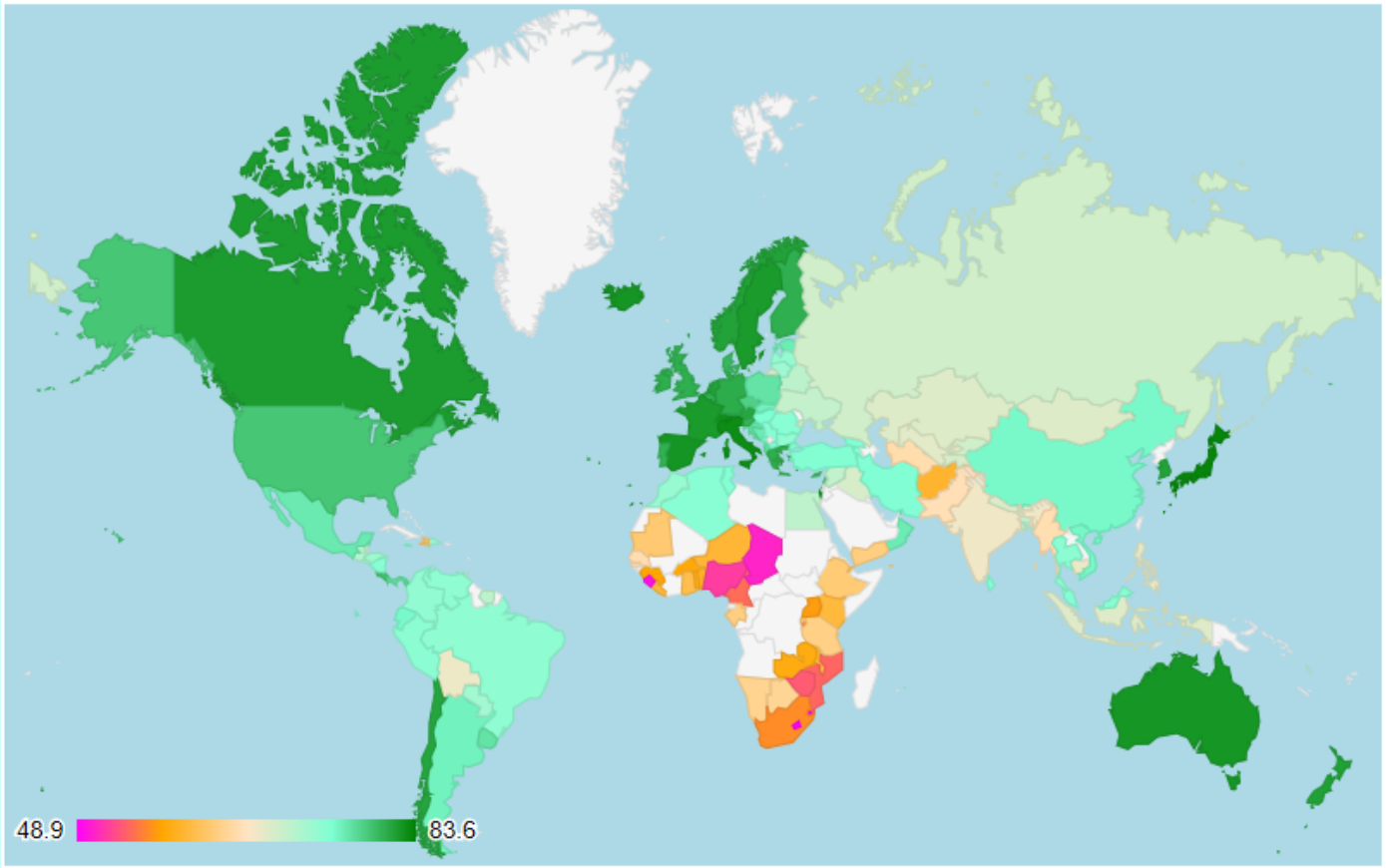
```
##Visualization of Correlations
psych::pairs.panels(HD[,c(3:7)],lm=TRUE,ellipse=FALSE)
```



Moderate positive correlations: **LifeExpectancy**, **Wellbeing**
Moderate negative correlation: InequalityOutcome

(2) LifeExpectancy Map

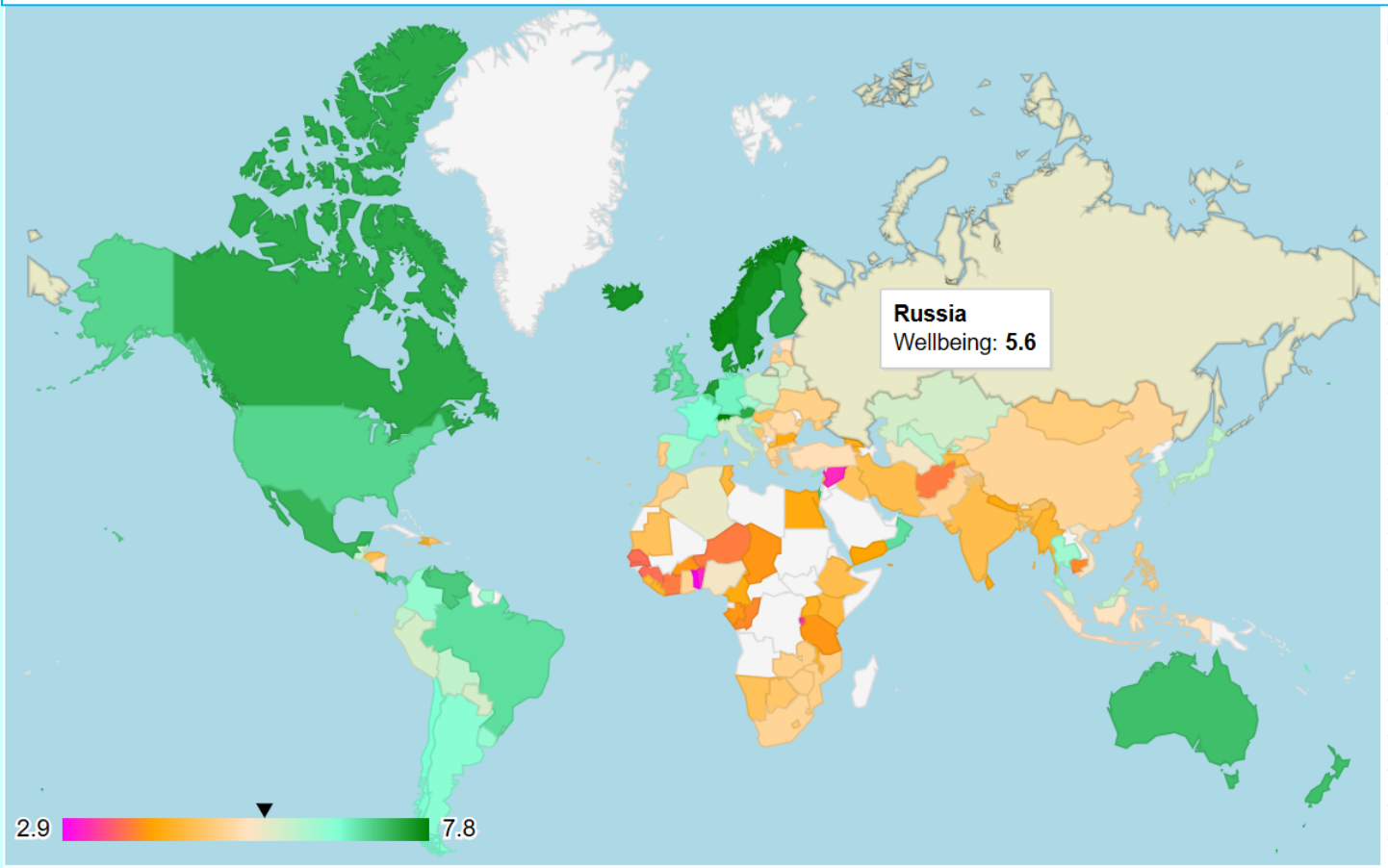
```
library(googleVis)
GC <- gvisGeoChart(HD, locationvar='Country', colorvar='LifeExpectancy',
  options=list(width=800,height=500,
               backgroundColor='lightblue',
  colorAxis="{values:[48.9,57.6,66.3,75.0,83.6],
  colors:[ \'magenta',\'orange',\'bisque',\'aquamarine',\'green']}"))
GT <- gvisTable(HRank[,c(1:3)],options=list(width=300,height=500))
plot(gvisMerge(GC,GT,horizontal=TRUE))
```



Rank	Country	LifeExpectancy
1	Costa Rica	79.1
2	Mexico	76.4
3	Colombia	73.7
4	Vanuatu	71.3
5	Vietnam	75.5
6	Panama	77.2
7	Nicaragua	74.3
8	Bangladesh	70.8
9	Thailand	74.1
10	Ecuador	75.4
11	Jamaica	75.3
12	Norway	81.3
13	Albania	77.3
14	Uruguay	76.9
15	Spain	82.2
16	Indonesia	68.5
17	El Salvador	72.5
18	Netherlands	81.2
19	Argentina	75.9
20	Philippines	67.9
21	Peru	74.1
22	Palestine	72.6
23	Brazil	73.9

(3) Wellbeing Map

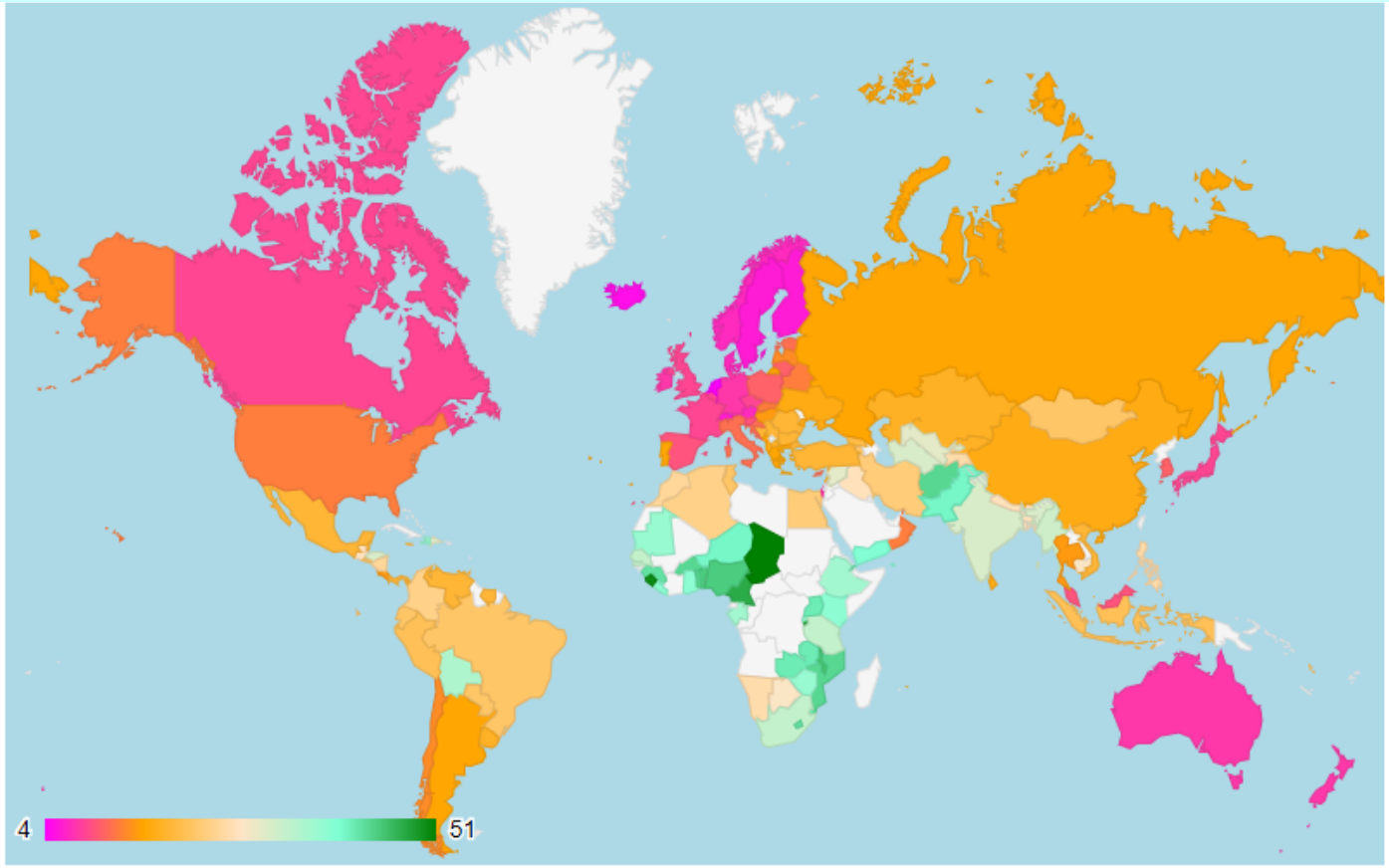
```
library(googleVis)
GC <- gvisGeoChart(HD, locationvar='Country', colorvar='Wellbeing',
                   options=list(width=800,height=500,
                                backgroundColor='lightblue',
                                colorAxis="{values:[2.9,4.1,5.4,6.6,7.8],
  colors:[ \'magenta',\'orange',\'bisque',\'aquamarine',\'green']}"))
GT <- gvisTable(HRank[,c(1,2,4)],options=list(width=300,height=500))
plot(gvisMerge(GC,GT,horizontal=TRUE))
```



Rank	Country	Wellbeing
1	Switzerland	7.8
2	Norway	7.7
3	Iceland	7.6
4	Sweden	7.6
5	Denmark	7.5
6	Netherlands	7.5
7	Austria	7.4
8	Canada	7.4
9	Finland	7.4
10	Costa Rica	7.3
11	Mexico	7.3
12	Australia	7.2
13	New Zealand	7.2
14	Israel	7.1
15	Venezuela	7.1
16	Ireland	7
17	Luxembourg	7
18	United States of America	7
19	Belgium	6.9
20	Brazil	6.9
21	Oman	6.9
22	Panama	6.9
23	United Kingdom	6.9

(4) InequalityOutcome Map

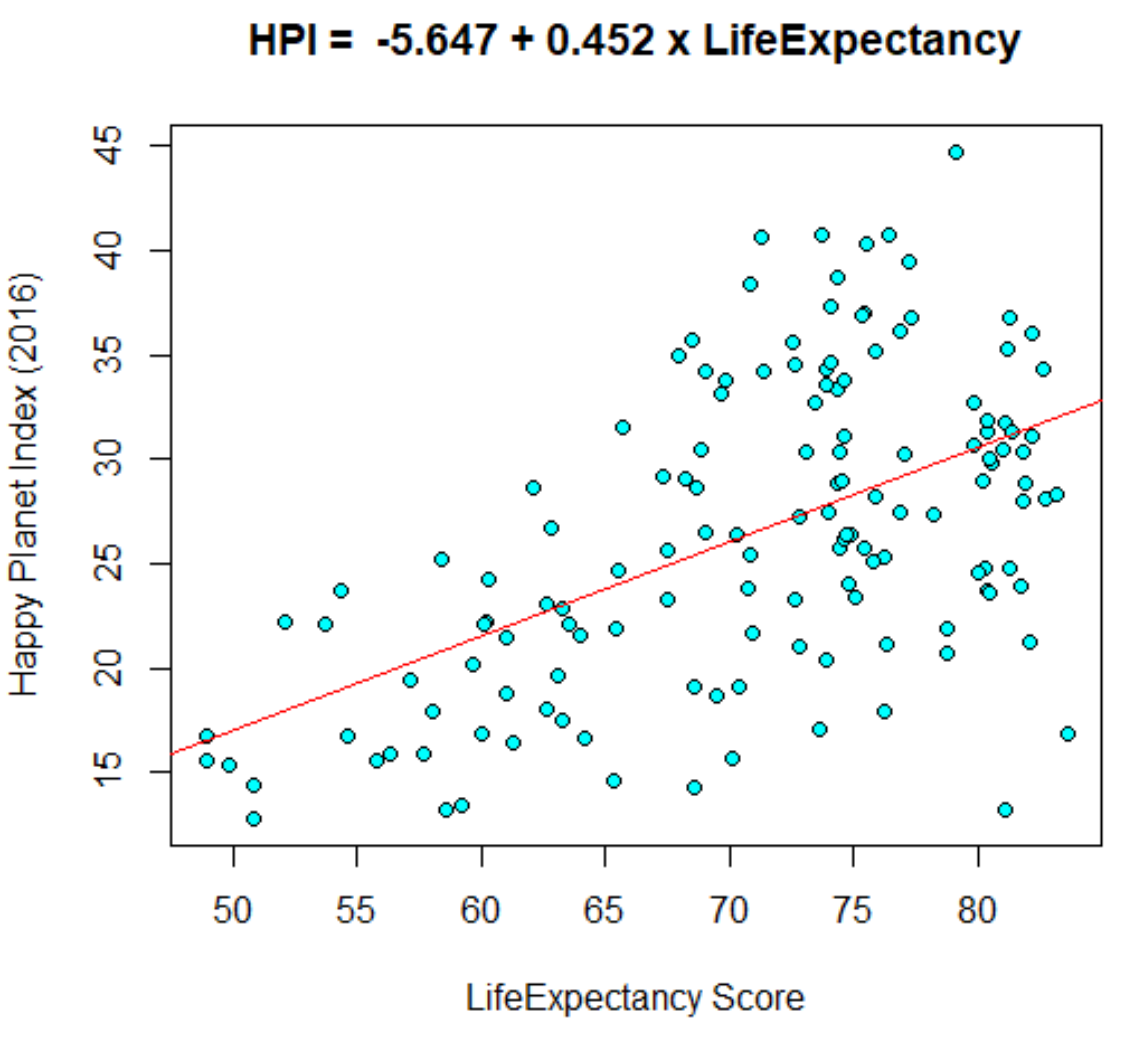
```
library(googleVis)
GC <- gvisGeoChart(HD, locationvar='Country', colorvar='InequalityOutcome',
                   options=list(width=800,height=500,
                                backgroundColor='lightblue',
                                colorAxis="{values:[4.0,15.8,27.5,39.3,51.0],
                                colors:[ \'magenta',\'orange',\'bisque',\'aquamarine',\'green']}"))
GT <- gvisTable(HRank[,c(1,2,6)],options=list(width=300,height=500))
plot(gvisMerge(GC,GT,horizontal=TRUE))
```



Rank	Country	InequalityOutcome
1	Costa Rica	15
2	Mexico	19
3	Colombia	24
4	Vanuatu	22
5	Vietnam	19
6	Panama	19
7	Nicaragua	25
8	Bangladesh	27
9	Thailand	15
10	Ecuador	22
11	Jamaica	21
12	Norway	7
13	Albania	17
14	Uruguay	18
15	Spain	10
16	Indonesia	21
17	El Salvador	22
18	Netherlands	4
19	Argentina	16
20	Philippines	26
21	Peru	21
22	Palestine	24
23	Brazil	22

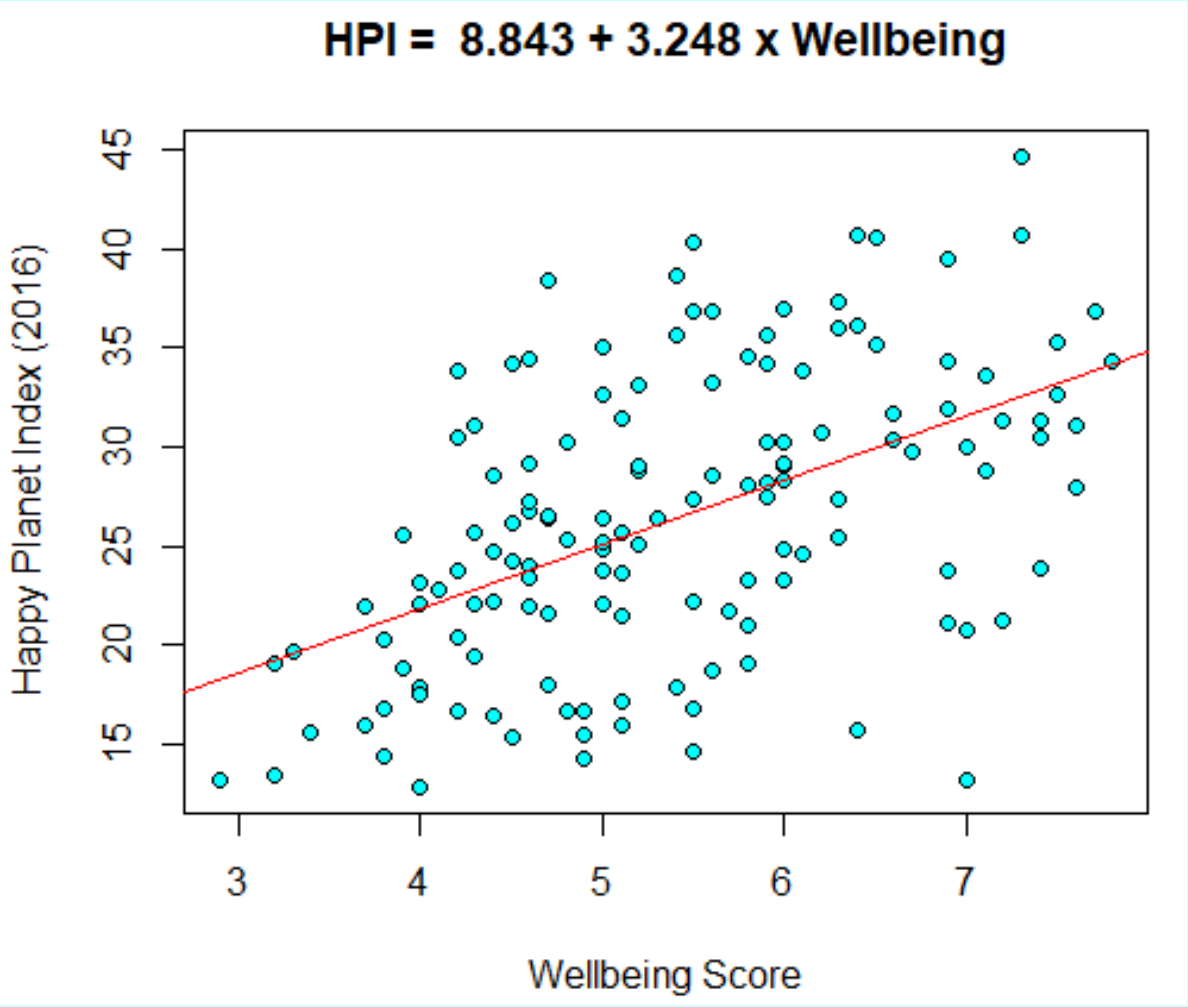
(5) Linear Regression for Happiness vs. LifeExpectancy

```
fitL <- lm(HPI ~ LifeExpectancy, data=HD)
title = paste('HPI = ',round(fitL$coefficients[1],3),
              '+', round(fitL$coefficients[2],3), 'x LifeExpectancy')
plot(HD[,3],HD[,7], pch=21, bg='cyan', xlab='LifeExpectancy Score',
     ylab='Happy Planet Index (2016)', main=title)
abline(fitL, col=2)
```



(6) Linear Regression for Happiness vs. Wellbeing

```
fitW <- lm(HPI ~ Wellbeing, data=HD)
title = paste('HPI = ',round(fitW$coefficients[1],3),
              '+', round(fitW$coefficients[2],3), 'x Wellbeing')
plot(HD[,4], HD[,7], pch=21, bg='cyan', xlab='Wellbeing Score',
     ylab='Happy Planet Index (2016)', main=title)
abline(fitW, col=2)
```



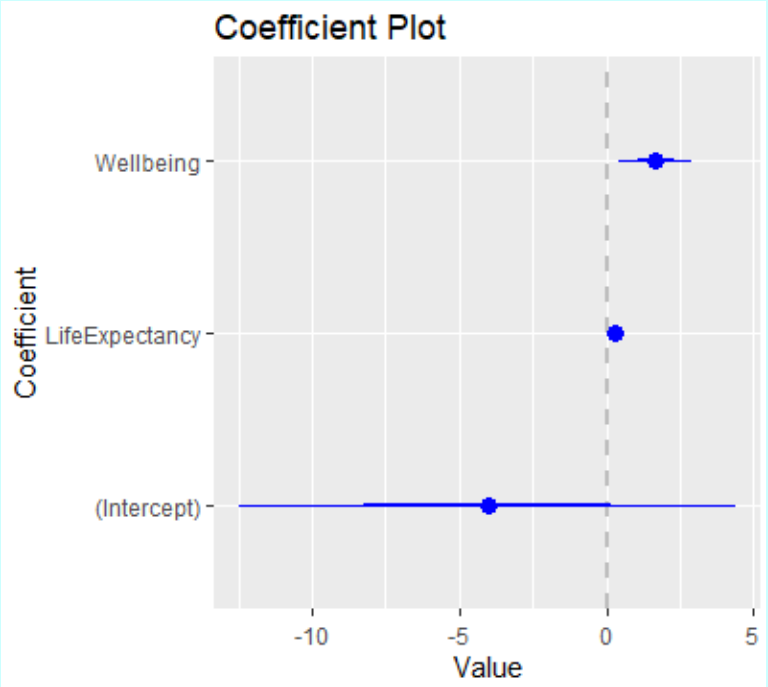
Conduct also one way ANOVA test for HPI and Wellbeing variables.
→ Exercise 1.

(7) Linear Regression for Happiness vs. (LifeExpectancy, Wellbeing)

```
> fitA <- lm(HPI ~ LifeExpectancy + Wellbeing, data=HD)
> summary(fitA)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.03547    4.22310  -0.956 0.340973
LifeExpectancy  0.30157    0.08028   3.756 0.000254 ***
Wellbeing       1.67442    0.61195   2.736 0.007040 **
---
Residual standard error: 6.04 on 137 degrees of freedom
Multiple R-squared:  0.3289,    Adjusted R-squared:  0.3191
F-statistic: 33.56 on 2 and 137 DF,  p-value: 1.37e-12
```

HPI = -4.035 + 0.302 LifeExpectancy + 1.674 Wellbeing

```
library(coefplot)
coefplot(fitA)
```



(8) Clustering

kmeans(x, centers, ...) {stats}

Perform k-means clustering on a data matrix.

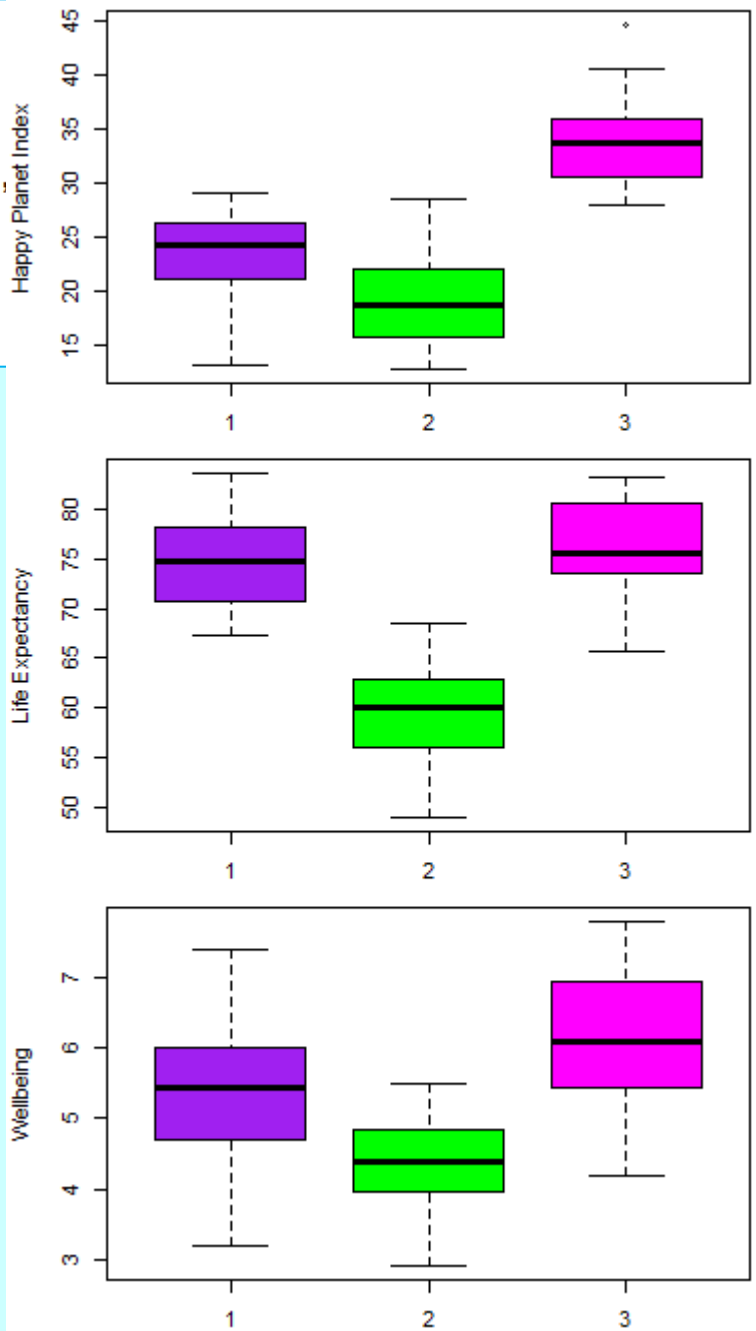
```
> kc <- kmeans(HD[,c(3,4,7)], centers=3, nstart=10)
> HDC <- data.frame(HD,Cluster=kc$cluster)
> head(HDC,3)
  Rank     Country LifeExpectancy Wellbeing Footprint InequalityOutcome  HPI Cluster
1  110 Afghanistan           59.7       3.8       0.8                43 20.2       2
2   13     Albania           77.3       5.5       2.2                17 36.8       3
3   30     Algeria           74.3       5.6       2.1                24 33.3       3
> table(HDC$Cluster)

 1  2  3
46 39 55
```

Mean values of Happiness, Life expectancy, and Wellbeing by Cluster

```
> library(dplyr)
> Mean_by_Cluster <- HDC %>%
+   select(HPI, LifeExpectancy, Wellbeing, Cluster) %>%
+   group_by(Cluster) %>%
+   summarise(mean_HPI=mean(HPI), mean_LifeExpectancy=mean(LifeExpectancy),
+             mean_Wellbeing=mean(Wellbeing))
> Mean_by_Cluster
# A tibble: 3 x 4
  Cluster mean_HPI mean_LifeExpectancy mean_Wellbeing
    <int>    <dbl>               <dbl>          <dbl>
1       1     23.6                74.6           5.45
2       2     19.2                59.1           4.33
3       3     33.8                76.2           6.14
```

Boxplots of HPI, Life expectancy, and Wellbeing by Cluster

```
cols <- c("purple","green","magenta")
par(mfrow=c(3,1),mar=c(2,4,1,1))
boxplot(HPI~Cluster, boxwex=0.75,xlab="Cluster",
        ylab="Happy Planet Index",col=cols, data=HDC)
boxplot(LifeExpectancy~Cluster, boxwex=0.75, xlab="Cluster",
        ylab="Life Expectancy",col=cols, data=HDC)
boxplot(Wellbeing~Cluster, boxwex=0.75,xlab="Cluster",
        ylab="Wellbeing", col=cols, data=HDC)
par(mfrow=c(1,1))
```



Interpretation by Cluster

1: Middle scores of HPI, LifeExpectancy, and Wellbeing

2: Lowest scores of HPI, LifeExpectancy, and Wellbeing

3: Highest scores of HPI, LifeExpectancy, and Wellbeing