GapMinder Country Visualizations

Will Koehrsen

April 7, 2017

Introduction

I decided to explore six different measures for countries in 2006. The six statistics I choose were: inequality, poverty, billionaires, energy use, literacy, and life expectancy.

1. Inquality: measured by the GINI index, which describes how much the wealth distribution in a country varies from perfectly equal. A figure of 0 represents total equality, while 100 represents total inequality.
2. Poverty: measured in the percentage of the total population subsisting on <$2 per day.
3. Billionaires: number of billionairies per one million inhabitants
4. Energy Use Per Person: measured in tonnes of oil equivalent (TOE)
5. Literacy: measured as percentage of adults (15+) who demonstrate literacy
6. Life Expectancy: life expectancy in years at birth.

** All data is sourced from GapMinder **

Data Wrangling

First, I need to load in the data from separate csv files and combine into a single dataframe. I will use the merge function to merge on country names. To merge multiple data frames at once, I can use the reduce function combined with merge and the list of dataframes to merge.

Visualization and Correlation Analysis

The data wrangling phase was relatively simple because GapMinder provides the data in a tidy format. I now had all the statistics for 2006 in a single dataframe. I saw that I had a lot of NAs, which might be a problem. The billionaires per million inhabitants had many zeros instead of na. I decided to leave that as is, because it is not missing data. Most countries really do not have any billionaires. I may subset this data when plotting or generating correlation statistics to exclude the countries with 0 billionaires.

```
# First I should take a look at the overall summary

summary(df[,2:7])
```

```
##    inequality     poverty_rate       energy          literacy_rate
##  Min.   :27.66   Min.   : 0.14   Min.   : 0.00911   Min.   :26.18
##  1st Qu.:34.34   1st Qu.: 4.62   1st Qu.: 0.51512   1st Qu.:62.75
##  Median :42.77   Median :14.42   Median : 1.14679   Median :88.12
##  Mean   :43.54   Mean   :27.30   Mean   : 2.26415   Mean   :76.17
##  3rd Qu.:52.15   3rd Qu.:45.04   3rd Qu.: 2.88191   3rd Qu.:92.30
##  Max.   :67.40   Max.   :95.15   Max.   :18.74956   Max.   :99.02
##  NA's   :195     NA's   :194     NA's   :80         NA's   :214
##  Life.expectancy billionaire_pm
##  Min.   :43.10   Min.   : 0.0000
##  1st Qu.:64.75   1st Qu.: 0.0000
##  Median :72.70   Median : 0.0000
##  Mean   :70.01   Mean   : 0.2298
##  3rd Qu.:76.80   3rd Qu.: 0.0000
##  Max.   :84.40   Max.   :30.7286
##  NA's   :40
```

Maybe I should subset the billionaires. Here is the summary for countries with at least one billionaire.

```
summary(subset(df, billionaire_pm != 0, billionaire_pm))
```

```
##  billionaire_pm
##  Min.   : 0.006103
##  1st Qu.: 0.085071
##  Median : 0.242550
##  Mean   : 1.120716
##  3rd Qu.: 0.774383
##  Max.   :30.728575
```

```
# Which country can the maximum be?

subset(df, df$billionaire_pm==max(df$billionaire_pm))
```

```
##     Country inequality poverty_rate energy literacy_rate Life.expectancy
## 139  Monaco         NA           NA     NA            NA              NA
##     billionaire_pm
## 139       30.72857
```

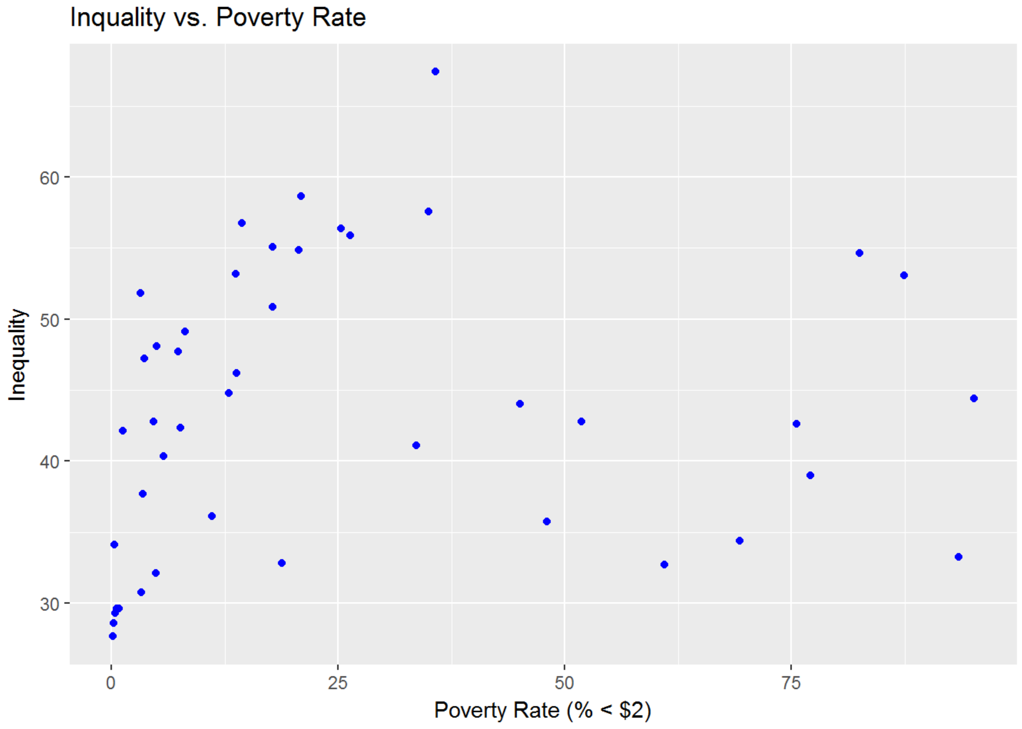
```
# Which country has the highest literacy rate?
subset(df, df$literacy_rate == max(df$literacy_rate, na.rm=TRUE))
```

```
##     Country inequality poverty_rate   energy literacy_rate Life.expectancy
## 213   Tonga         NA           NA 0.561112      99.01846            70.1
##     billionaire_pm
## 213              0
```

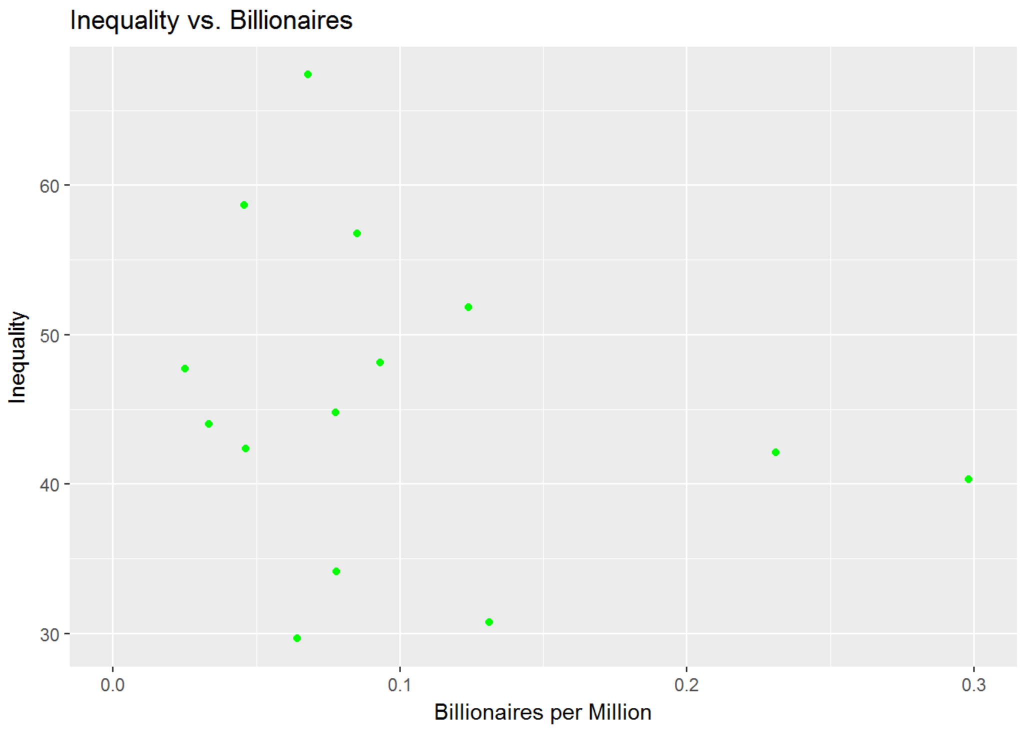
```
# Finally, who has the lowest inequality?
subset(df, df$inequality == min(df$inequality, na.rm=TRUE))
```

```
##             Country inequality poverty_rate  energy literacy_rate
## 190 Slovak Republic      27.66         0.14 3.45743            NA
##     Life.expectancy billionaire_pm
## 190            74.5              0
```

```
## Warning: Removed 195 rows containing missing values (geom_point).
```



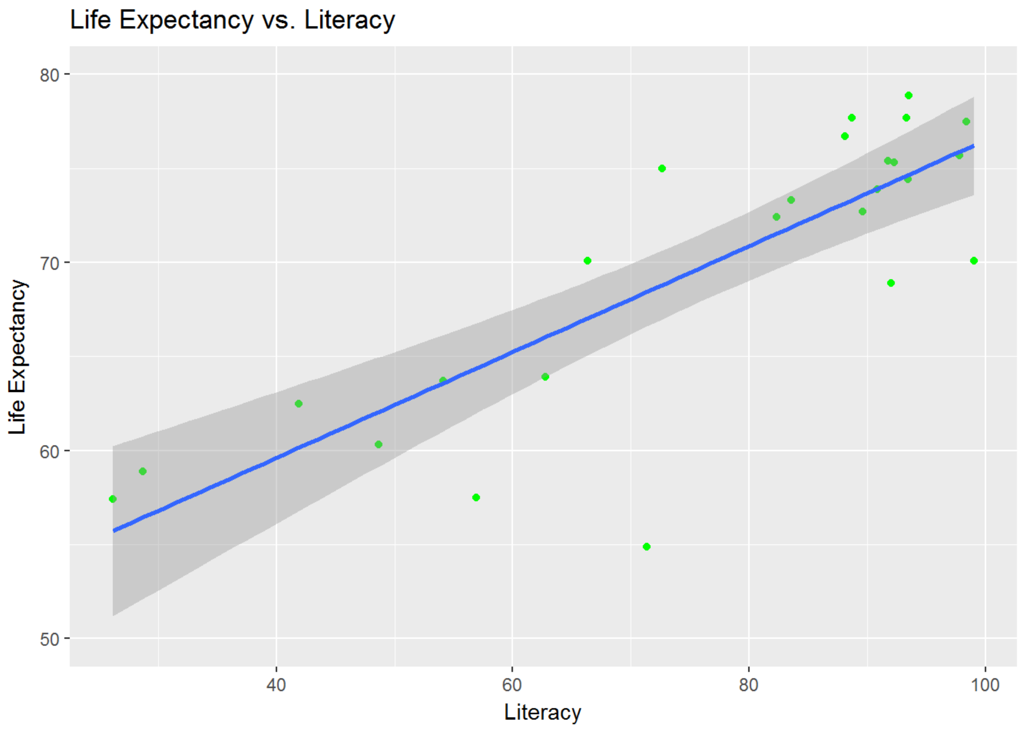
```
## Warning: Removed 35 rows containing missing values (geom_point).
```



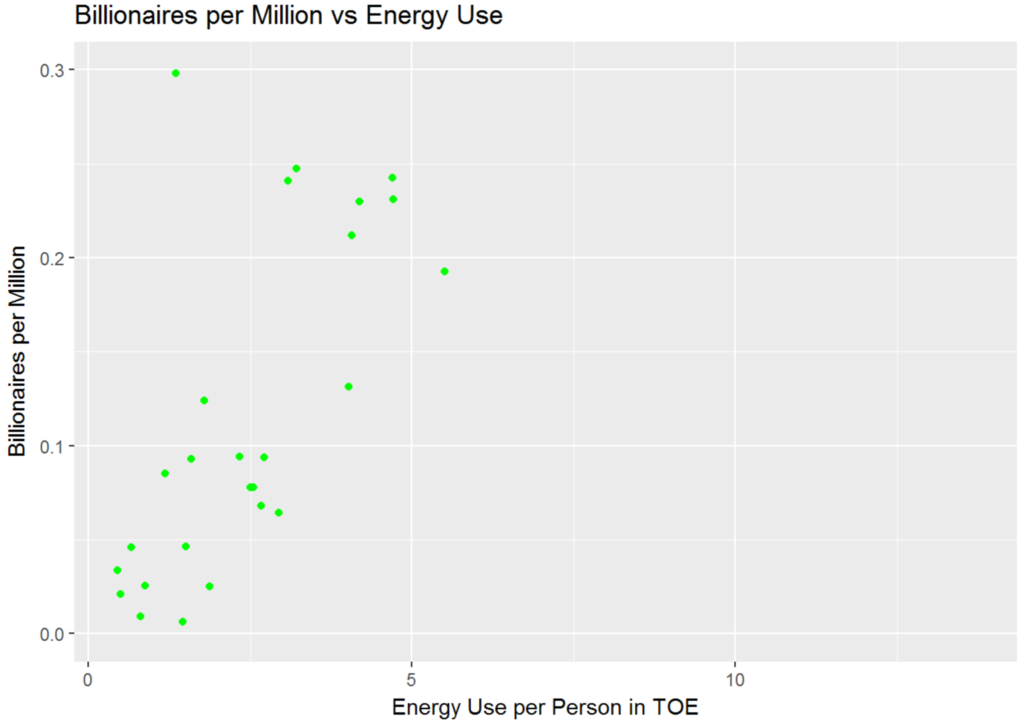
```
## Warning: Removed 214 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 214 rows containing missing values (geom_point).
```

These are fairly intriguing. It would be better if data existed for every country, but that will not always be the case.

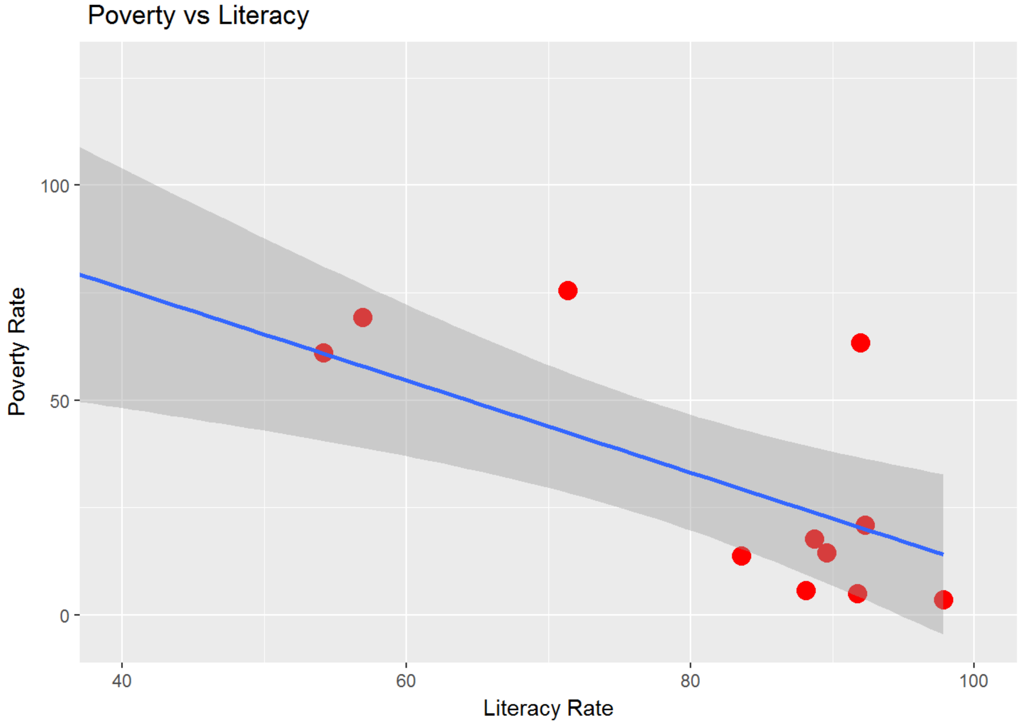


```
## Warning: Removed 2 rows containing missing values (geom_point).
```



```
## Warning: Removed 227 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 227 rows containing missing values (geom_point).
```

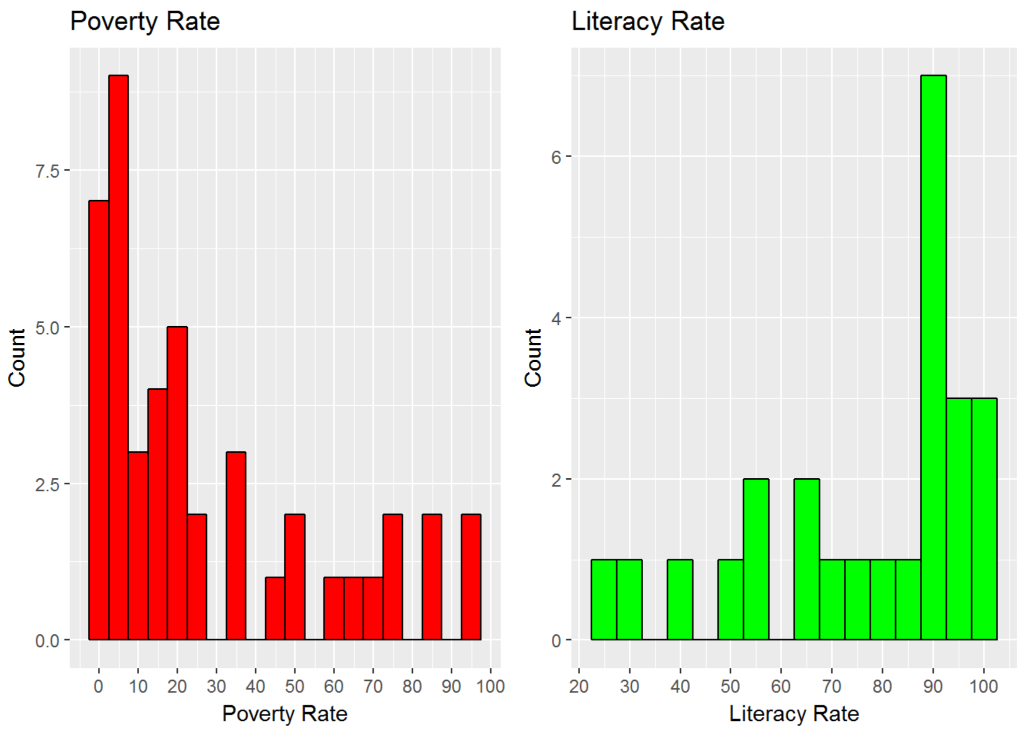


That is definitely what I expected. As a country increases its literacy rate, that is a good indicator of the wealth of a country, and hence the poverty rate will decrease. Or maybe the poverty rate decreasing drives up the literacy rate. At this point, it is clear this is a correlation, but the causation direction cannot be determined without looking at the changes over time.

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

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

```
## Warning: Removed 194 rows containing non-finite values (stat_bin).
## Warning: Removed 214 rows containing non-finite values (stat_bin).
```



Finally, I want to see the correlations between each row. I do not know exactly what to expect, so I will calculate the correlation between each pair of rows to see the highest and then I can graph those in a scatterplot to check the statistics.

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## 
## Attaching package: 'Hmisc'
```

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

```
## The following objects are masked from 'package:dplyr':
## 
##     combine, src, summarize
```

```
## The following objects are masked from 'package:base':
## 
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

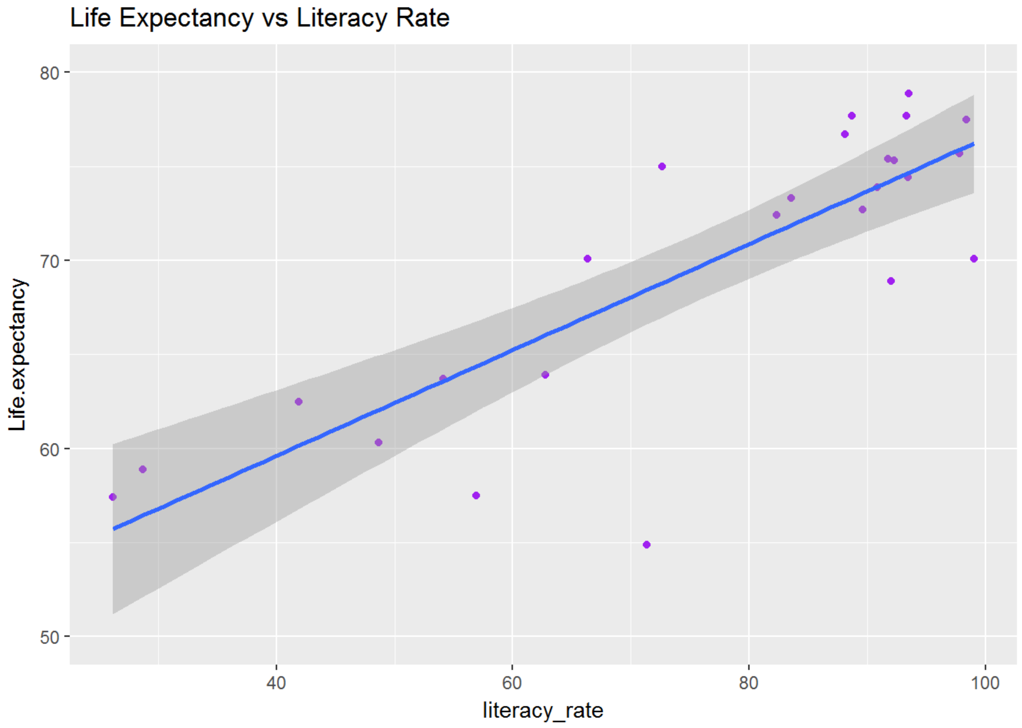
```
##                  inequality poverty_rate energy literacy_rate
## inequality             1.00         0.13  -0.35          0.70
## poverty_rate           0.13         1.00  -0.55         -0.76
## energy                -0.35        -0.55   1.00          0.27
## literacy_rate          0.70        -0.76   0.27          1.00
## Life.expectancy       -0.02        -0.79   0.44          0.83
## billionaire_pm         0.02        -0.33   0.48          0.21
##                  Life.expectancy billionaire_pm
## inequality                 -0.02           0.02
## poverty_rate               -0.79          -0.33
## energy                      0.44           0.48
## literacy_rate               0.83           0.21
## Life.expectancy             1.00           0.32
## billionaire_pm              0.32           1.00
## 
## n
##                  inequality poverty_rate energy literacy_rate
## inequality              44           44     40            11
## poverty_rate            44           45     41            12
## energy                  40           41    159            21
## literacy_rate           11           12     21            25
## Life.expectancy         44           45    158            25
## billionaire_pm          44           45    159            25
##                 Life.expectancy billionaire_pm
## inequality                   44             44
## poverty_rate                 45             45
## energy                      158            159
## literacy_rate                25             25
## Life.expectancy             199            199
## billionaire_pm              199            239
##
## P
##                 inequality poverty_rate energy literacy_rate
## inequality                 0.4062       0.0249 0.0175
## poverty_rate    0.4062                  0.0002 0.0043
## energy          0.0249     0.0002              0.2301
## literacy_rate   0.0175     0.0043       0.2301
## Life.expectancy 0.9087     0.0000       0.0000 0.0000
## billionaire_pm  0.9107     0.0261       0.0000 0.3189
##                 Life.expectancy billionaire_pm
## inequality      0.9087          0.9107
## poverty_rate    0.0000          0.0261
## energy          0.0000          0.0000
## literacy_rate   0.0000          0.3189
## Life.expectancy                 0.0000
## billionaire_pm  0.0000
```

It appears as those life expectancty and literacy rate are the most highly correlated country statistics. We should take a look at that plot to confirm.

```
ggplot(aes(x=literacy_rate, y=Life.expectancy), data=df) + geom_point(color='purple') + coord_cartesian(ylim=c
(50,80)) +
  geom_smooth(method='lm') + labs(title='Life Expectancy vs Literacy Rate')
```

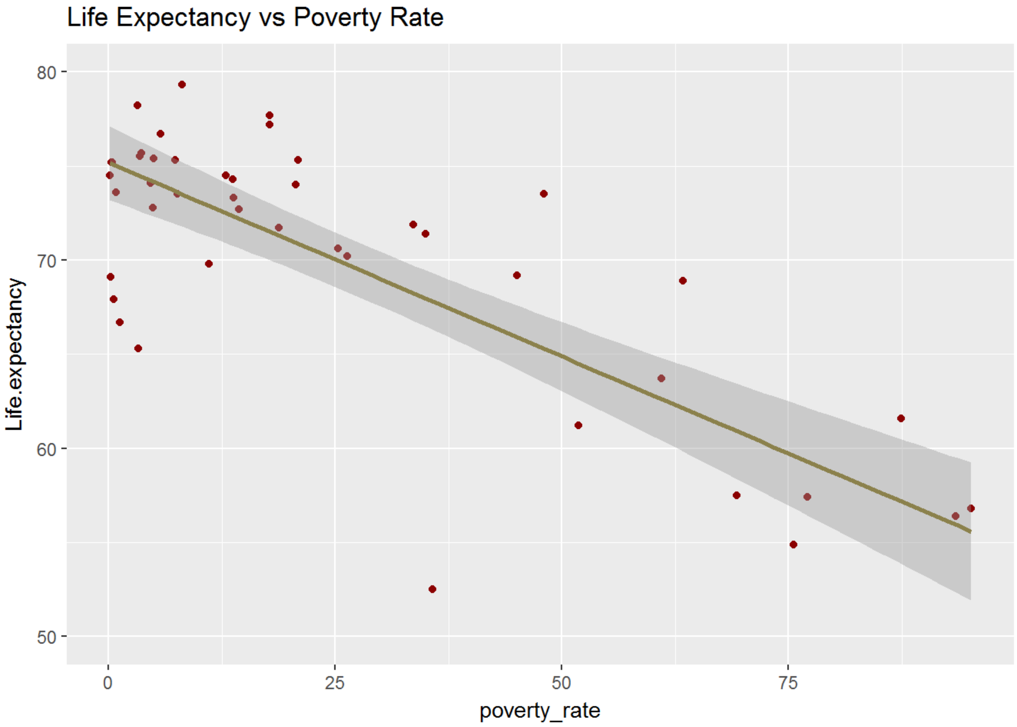
```
## Warning: Removed 214 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 214 rows containing missing values (geom_point).
```



Looking at the correlations again, the most negative correlation is between poverty and life expectancy. One more plot to make sure that can be confirmed visually (a plot is worth a thousand statistics)

```
## Warning: Removed 194 rows containing non-finite values (stat_smooth).
## Warning: Removed 194 rows containing missing values (geom_point).
```



To conclude, the most highly correlated country statistics for 2006 were life expectancy and literacy rate. The most negatively correlated statistics for countries in 2006 were life expectancy and poverty rate. Inquality, which was what most started my investigation, was most negatively correlated with energy per person. I also discovered that Monaco is where all the rich billionaires like to live, with an astounding 30 billionaires per one million residents. Moreover, it appears that Tonga has the highest literacy rate and the Slovak Republic has the lowest inequality among the countries with data on record.