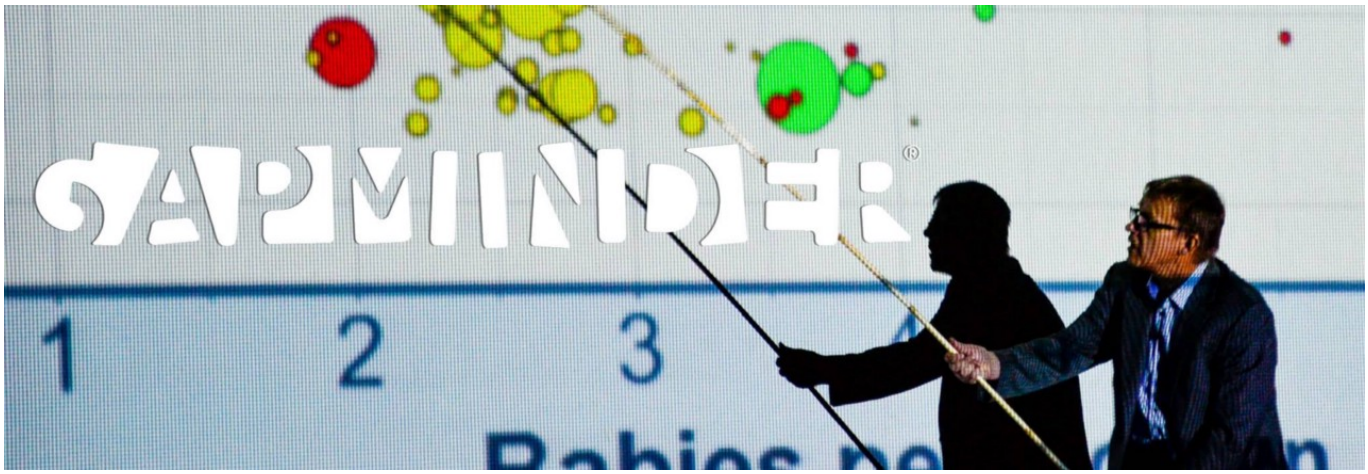
Exploratory Data Analysis in R of Global Data from GapMinder

Hamza Bendemra, Ph.D. Follow
Feb 3, 2018 · 7 min read



Hans Rosling/GapMinder

In this post, I perform an Exploratory Data Analysis (EDA) on two data sets from GapMinder. This post includes the R code used (also found in this GitHub repo). In summary:

- Method: Exploratory Data Analysis (EDA), Correlation, Linear Regression
- Program/Platform: R/RStudio
- Sources: World Health Organization, World Bank

The Data

In this data analysis, I use data available on GapMinder's data webpage. Specifically, I focused on:

- *GDP/capita (US$, inflation-adjusted)* from the World Bank (WB) and
- *Prevalence of HIV among adults aged 15–49 (%)* from the World Health Organisation (WHO).

The Question

The question I am asking in this analysis:

Is there a correlation between GDP per Capita and prevalence of HIV in the 15–49 age bracket? And if yes, how strong is that correlation?

My expectation is that there is a negative correlation between GDP per capita and HIV Prevalence; meaning that poorer countries have higher prevalence of HIV.

Data Wrangling

Let's do some initial data wrangling in R on the CSV files downloaded from GapMinder to prep our data for analysis.

GDP per capita data

The data from GapMinder was in the form of CSV files that needed to be reorganised according to key-value pairs in the original CSV tables. I used the function *gather()* from the amazing 'tidyr' library.

Let's have a look at the structure of the resulting data frame structure and determine what time frame this data set covers:

```
## 'data.frame': 14300 obs. of 3 variables:
## $ Income per person (fixed 2000 US$):
Factor w/ 275 levels "Abkhazia","Afghanistan",..: 1 2 3 5 6 7 8 9 10
12 ...
## $ Year : int 1960 1960 1960 1960 1960 1960 1960 1960 1960 1960 ...
## $ GDP : num NA NA NA NA 1280 ...
```

```
## [1] 1960 2011
```

The resulting dataframe features 14,300 observations of 3 variables (Country, Year, GDP). The column 'Country' lists 275 countries. GDP per Capita is provided for the 275 countries from 1960–2011.

HIV prevalence data

Let's perform similar data tidying on the HIV prevalence data. Again, I'll be using the function *gather()* from the 'tidyr' library.

Again, let's have a look at the structure of the resulting data frame structure:

```
## 'data.frame': 9075 obs. of 3 variables:
## $ Estimated HIV Prevalence% - (Ages 15-49):
Factor w/ 275 levels "Abkhazia","Afghanistan",..: 1 2 3 5 6 7 8 9 10
12 ...
## $ Year : int 1979 1979 1979 1979 1979 1979 1979 1979 1979 1979 ...
## $ HIV_prev : num NA NA NA NA NA ...
```

The resulting dataframe features 9075 observations of 3 variables (Country, Year, Estimated HIV Prevalence). The column 'Country' lists 275 countries. GDP per Capita is provided for the 275 countries from 1979–2011.

Combining dataframes

Let's combine the two dataframes to allow us to compare GDP per capita and HIV prevalence. I'm using the *merge()* function. Let's then have a look at the structure of the resulting dataframe.

```
## 'data.frame': 9075 obs. of 4 variables:
## $ Country : Factor w/ 275 levels "Abkhazia","Afghanistan",..: 1 1
1 1 1 1 1 1 1 1 ...
## $ Year : int 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 ...
## $ GDP : num NA NA NA NA NA NA NA NA NA NA ...
## $ HIV_prev: num NA NA NA NA NA NA NA NA NA NA ...
```

We now have a data frame for our subsequent data exploration. This dataframe is organised into four columns: (i) country, (ii) year, (iii) GDP/Capita for that country in that year, and (iv) HIV Prevalence for that country in that year.

Exploring the data

Datasets often feature missing data. I suspect it would be the case with this dataset also even if it was sourced from an official organisation. Let's have a look at the percentage of missing GDP data in the combined dataframe.

```
## [1] 35.30579
```

About 35.3% of the GDP per Capita column in the combined dataframe have missing data. This is quite substantial and is most likely due the fact that consistent measurements of GDP are costly and have only started in the last few decades (N.B. the World Bank itself was founded in 1945 after WWII).

Rather than replacing the missing data with an average or an estimate, missing data will be dismissed (i.e. not plotted) in the subsequent analysis.

Let's have a look at the percentage of missing data on the HIV Prevalence front.

```
## [1] 63.62534
```

We have an even higher rate at 63.6% for HIV Prevalence missing data in the combined dataframe. This is of course due to the fact that the initial dataframe from the World Health Organisation included measurement for a timeframe starting 1979, whereas our combined dataframe timeframe starts in 1960.

The lag in consistent measurements of HIV associated metrics have only really been performed on a large scale from the early-1980s when HIV/Aids became a recognised major health crisis.

Let's get an initial sense of the type of distribution we may getting from both data sets. For the GDP per capita:

```
## Min. - 1st Qu. - Median - Mean - 3rd Qu. - Max. - NA's
## 54.51 - 590.26 - 2038.88 - 7315.07 - 9239.73 - 108111.21 - 3204
```

Regarding the HIV prevalence:

```
## Min. - 1st Qu. - Median - Mean - 3rd Qu. - Max. - NA's
## 0.010 - 0.100 - 0.300 - 1.743 - 1.200 - 26.500 - 5774
```

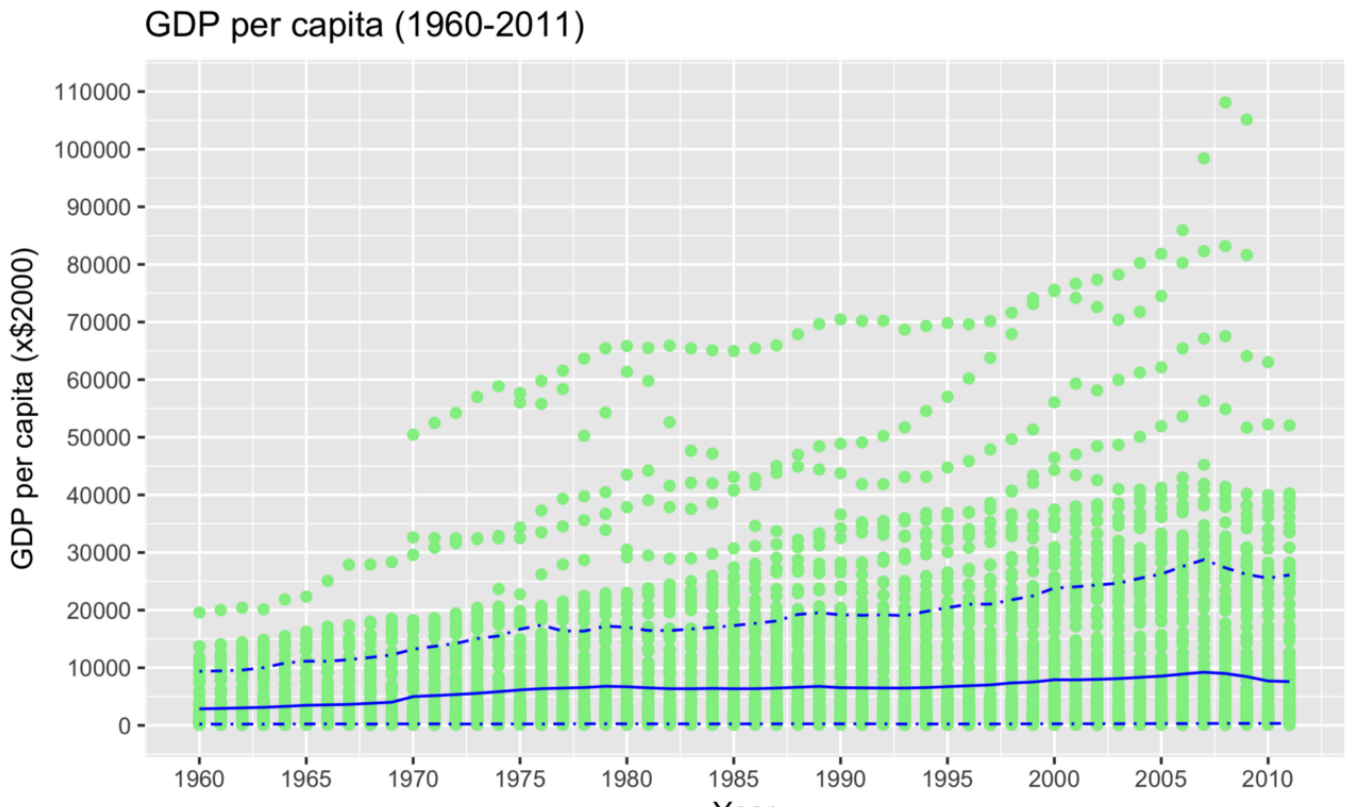
This quick glance clearly shows that some countries in our dataframe are considerable pulling the distribution's mean up (compared to the median). This is particularly the case with the HIV Prevalence data.

The Plots

In this section, I generate various plots (using ggplot) to get an overview of the distribution and attempt to identify trends and patterns.

First, we look at the overall data set and generate a scatter plot of GDP per Capita for the 275 countries listed in our dataset from 1960 to 2011.

We also overlay the mean and the upper-lower limits (5%; 95%) of GDPs on our plot. This will give us a better understanding of where the bulk of our distribution lies.

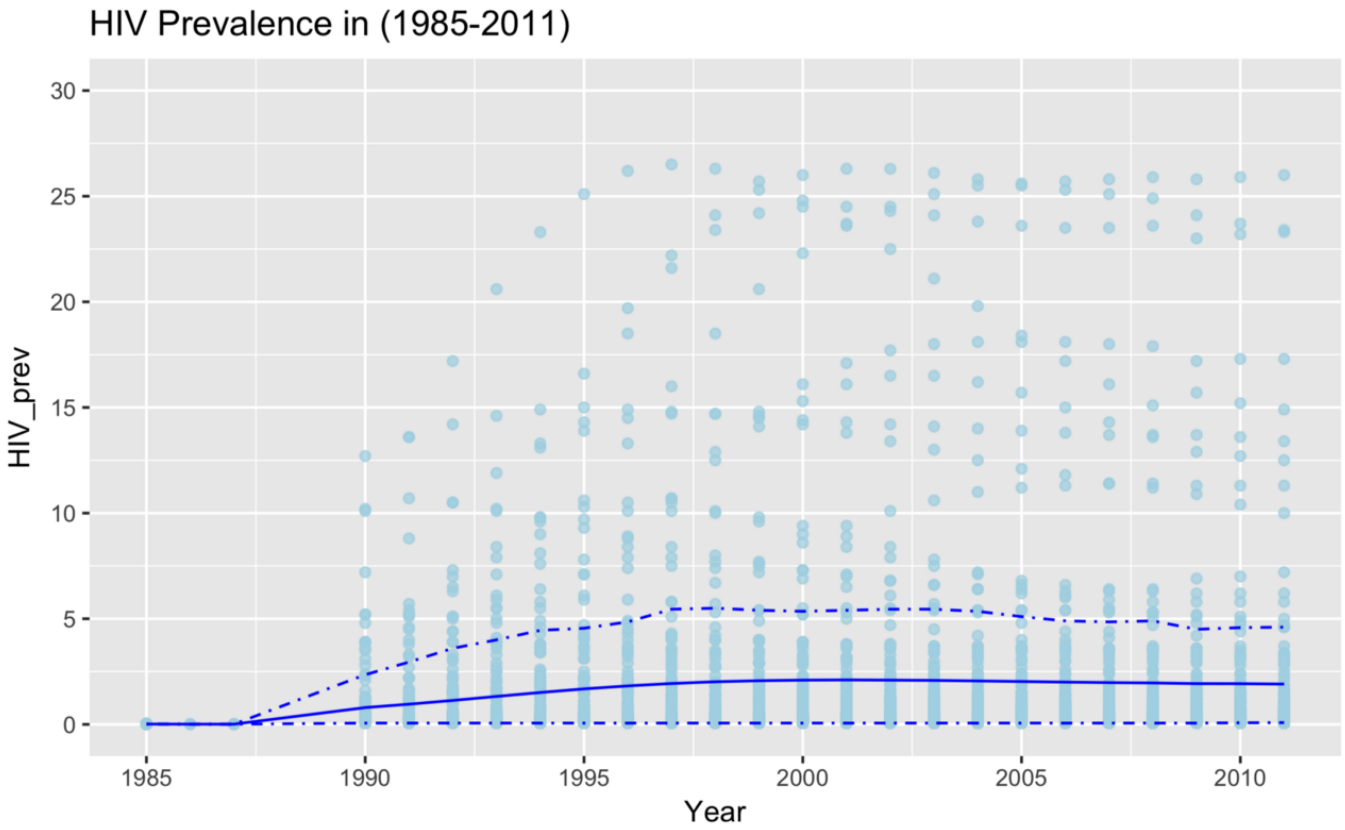


Year

The plot shows an overall increasing trend in Global GDP between 1960 and 2011. The bulk of the data falls at a maximum GDP per Capita of USD 30,000 (xUSD 2,000).

Let's look at the data set on HIV Prevalence by generating a scatter plot of HIV Prevalence for the 275 countries listed in our dataset from 1960 to 2011.

Similarly to GPD Per Capita, we also overlay the mean and the upper-lower limits (5%; 95%) of HIV prevalence on our plot. We'll focus on data from 1985 to 2011.

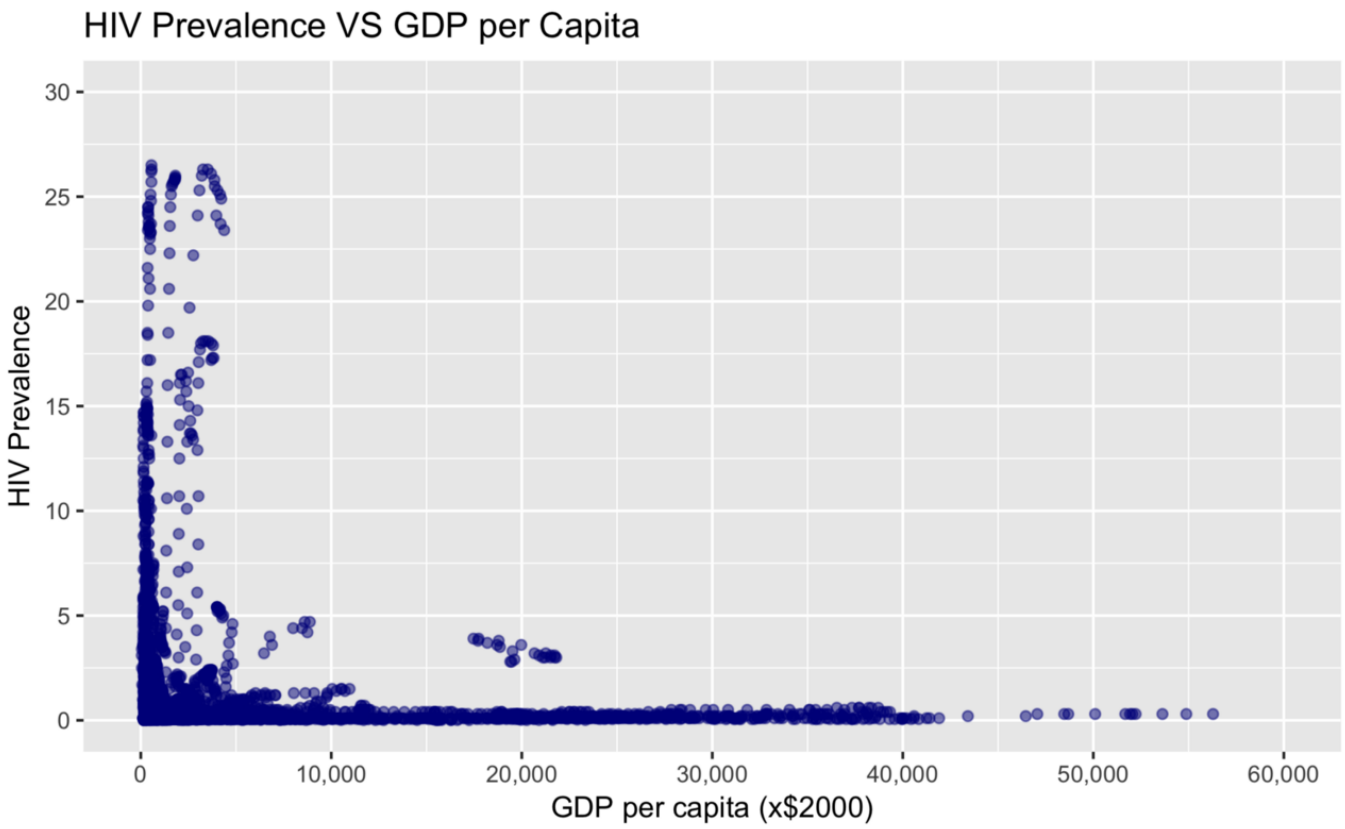


According to our data, the rate of HIV prevalence has increased between 1985 and 2011 with a stagnation in the mean from the early 2000s and slight decline since 2005. This would correspond to advances in preventative measures to reduce the incidence and likelihood of contracting HIV.

Correlation and Linear Regression

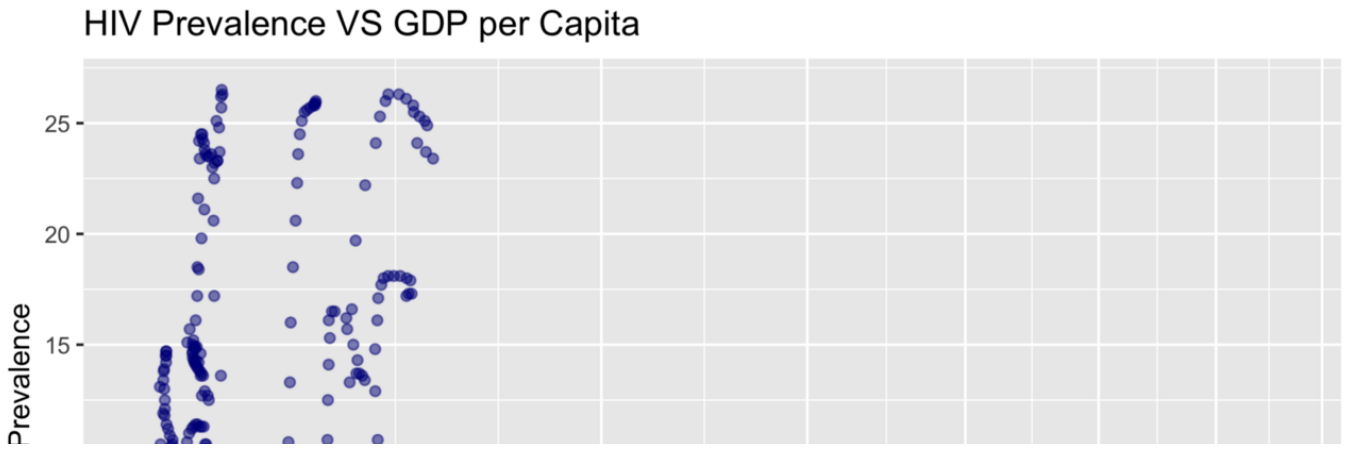
Now, let's focus on the combined dataframe we created to earlier to investigate the correlation between the two variables of interest. To have a first look at correlation, let's

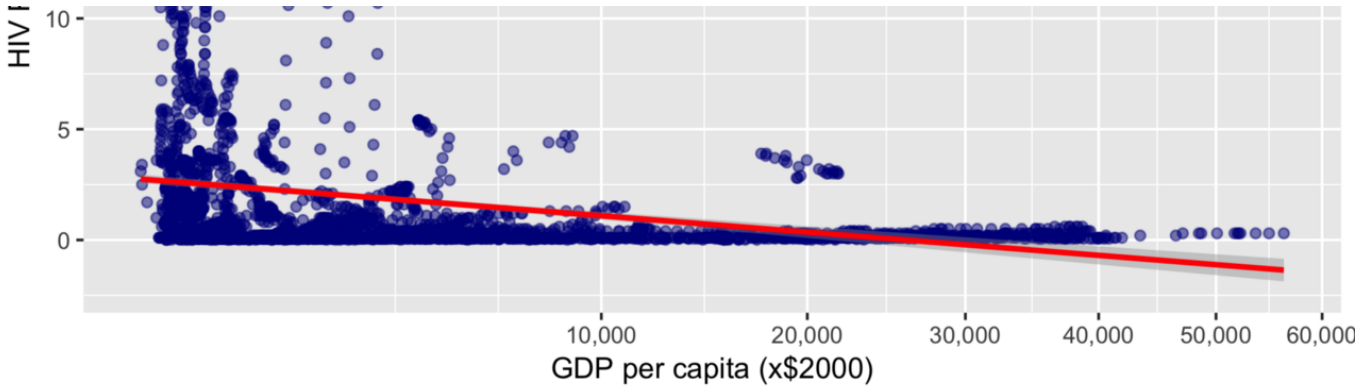
plot HIV Prevalance v. GDP per Capita.



The scatter plot clearly indicates that the lower GDP per Capita data points (i.e. countries) have a much higher HIV prevalence compared countries with higher GDP per Capita.

Let's take a closer by creating a plot with a square root scale applied to the x-axis to further emphasise countries with lower GDP per capita. We'll also use the R function geom_smooth() to perform a simple linear regression to better visualise the relationship between the two variables.





The scatter plot above further indicates that countries which shower GDP per capita have on average higher HIV prevalence.

Let's calculate the correlation factor between both variables using Pearson's method.

```
## ## Pearson's product-moment correlation
## ## data: gdp.HIV$HIV_prev and gdp.HIV$GDP
## t = -10.938, df = 3183, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.2235800 -0.1566303
## sample estimates:
## cor
## -0.1903264
```

The resulting correlation factor is -0.19 which is a negative but weak correlation. This matches with the linear regression plotted above.

This negative but weak correlation negative correlation matches other published results which found that an individual's wealth (rather that the GDP per Capita of the country in which that individual resides) is a stronger indicator of HIV prevalence within an individual's particular community.

Concluding Remarks

In this project, we collected data from public sources (GapMinder, WHO, WB). We performed data wrangling and an initial exploratory data analysis. Then, we derived a correlation factor and applied linear regression to assess the linear relationship between two variables of interest (GDP per capita, HIV prevalence).

Disclosure: this study was completed as part of a lesson in UDacity's Data Analyst Nanodegree.

Data Science R Programming Exploratory Data Analysis Udacity Data Analyst