Gapminder exploration

Zoe Vernon

August 29, 2018

1 Introduction

Gapminder is an excellent organization aimed at increasing the use and understanding of statistics on a number of global topics. They collect a variety of data from many sources and aim to produce fact-based statistics reflecting the current state of our world. The data we are exploring throughout this analysis consists of population, life expectency and GDP information for many countries through time.

The data can be found from `https://raw.githubusercontent.com/resbaz/r-novice-gapminder-files/master/data/gapminder-FiveYearData.csv` if you would like to download it yourself.

Fortunately, the data was already very clean, so we did not conduct any major modifications to the data.

2 Visualizing the gapminder data (ggplot2)

We are interested in exploring life expectancy as a function of GDP. Figure 1 shows a scatterplot of life expectancy versus GDP.

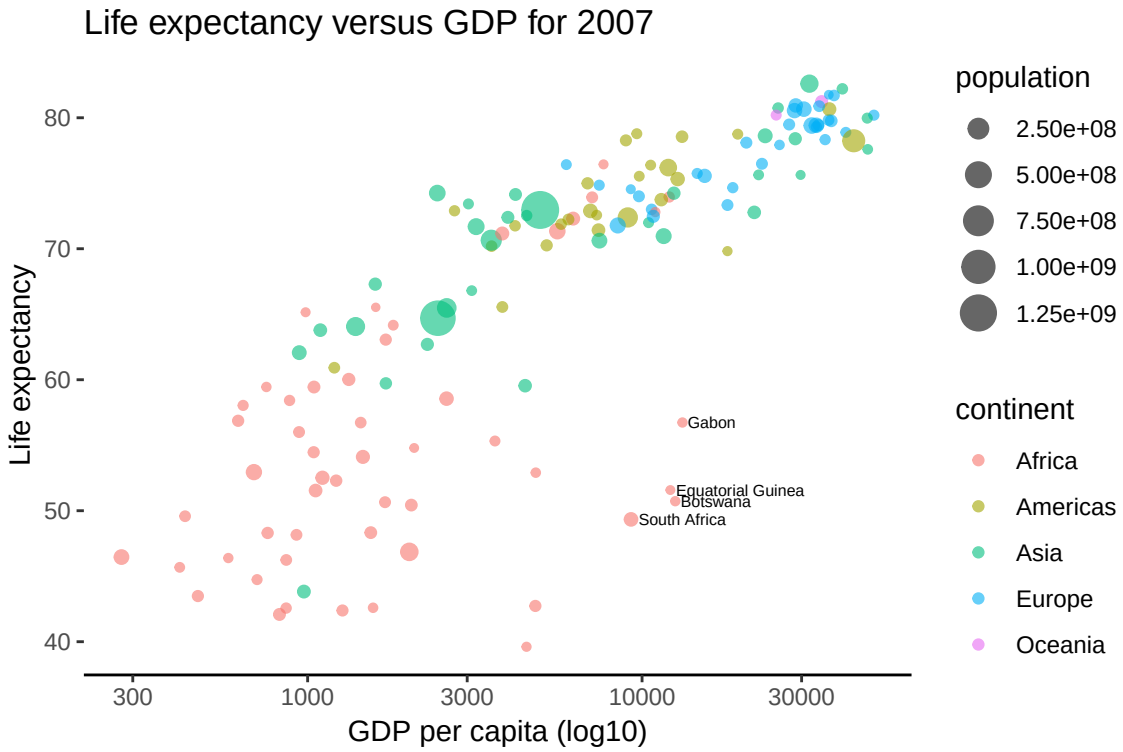

Figure 1: Life expectancy versus GDP for all countries in the year 2007

It certainly appears as though there is some kind of rapid increase in the low GDP range, which slows to a gradual increase in the high GDP range. Several African countries have surprisingly low life expectency for their GDP.

Next, we explore change in life expectancy over time. Figure 2 shows a series of boxplots, one for each year-continent combination. Each data point corresponds to the life expectency of a country for the given year in the given continent.

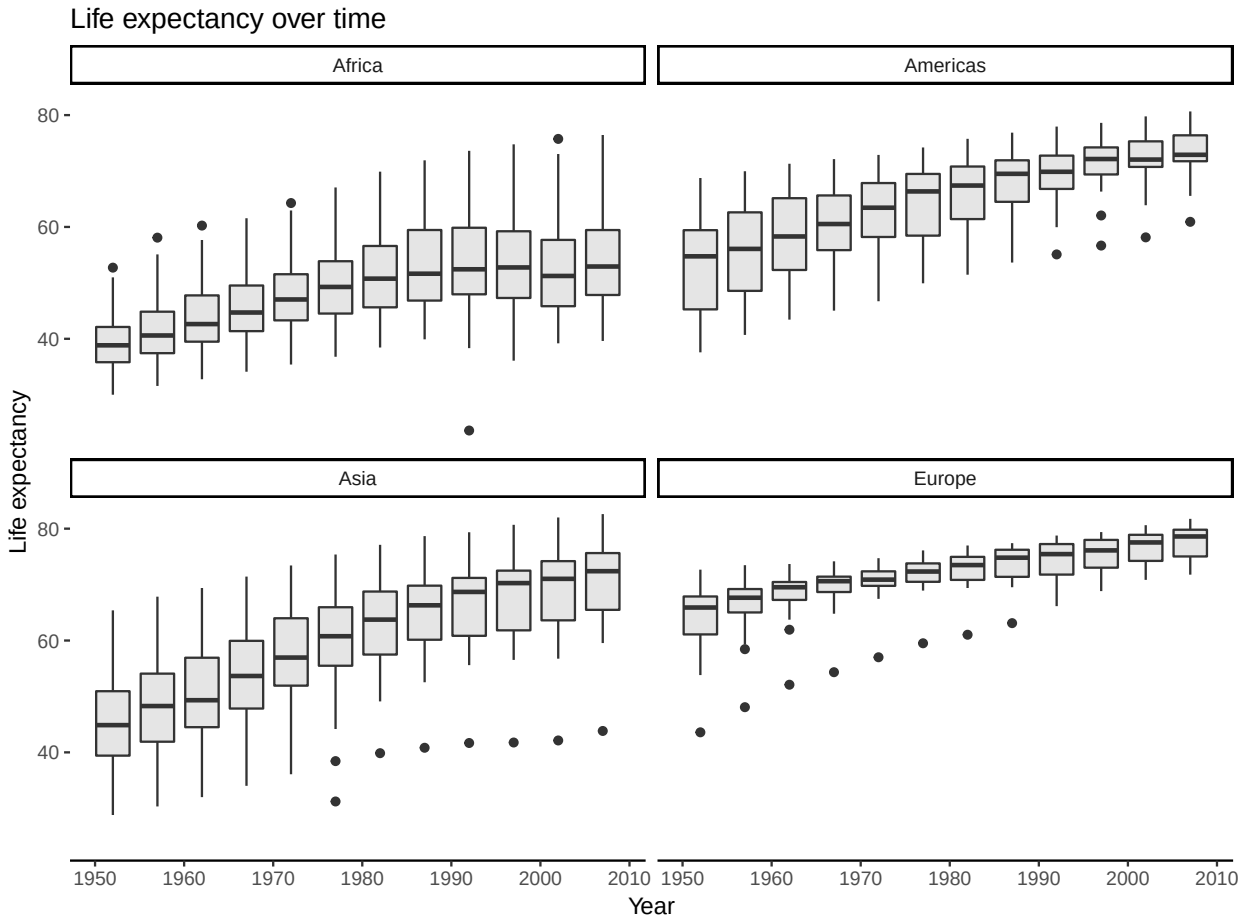

Figure 2: Life expectancy over time

We see that the life expectancy increased in Africa from 1950 up until the 1990s but has stayed fairly constant with a median of around 50 years since the 1990s. The Americas, Asia, and Europe on the other hand, have experienced continued growth.

2.1 Comparing GDP across continents (dplyr)

Table 1 compares GDP per capita across continents.

continent	countries	mean	SD
Oceania	2	29810.19	6540.99
Europe	30	25054.48	11800.34
Asia	33	12473.03	14154.94
Americas	25	11003.03	9713.21
Africa	52	3089.03	3618.16

Table 1: A table displaying the mean and standard deviation of GDP per capita in 2007 for each continent

Clearly Oceania is leading the way in terms of GDP per cap.

Next, we want to ask about raw GDP (i.e. overall GDP for each country, rather than standardized by per capita). Table 2 shows the average total GDP for each continent for 2007.

continent	countries	mean	SD
Americas	25	777	2573
Asia	33	628	1344
Europe	30	493	678
Oceania	2	404	424
Africa	52	46	92

Table 2: A table displaying the mean and standard deviation of GDP (in billions) in 2007 for each continent

3 tidyr()

See http://swcarpentry.github.io/r-novice-gapminder/14-tidyr/ for more ways to use “‘tidyr()“ on this data.

The gapminder data that we used for visualization was already in a clean usuable format. Here we are given a dataset that requires some processing to get in a more useful form.

```
# load the wide format of the data
gapminder_wide <- loadGapminderData(path = "data/",
                                    filename = "gapminder_wide.csv")

# look at wide_data comapred to original
dim(gapminder_wide)

## [1] 142  38

dim(gapminder)

## [1] 1704    6

head(gapminder_wide)[, c(1:4, 37:38)]

##   continent      country gdpPercap_1952 gdpPercap_1957 pop_2002 pop_2007
## 1    Africa      Algeria      2449.0082      3013.9760 31287142 33333216
## 2    Africa       Angola      3520.6103      3827.9405 10866106 12420476
## 3    Africa        Benin      1062.7522       959.6011  7026113  8078314
## 4    Africa     Botswana       851.2411       918.2325  1630347  1639131
## 5    Africa Burkina Faso       543.2552       617.1835 12251209 14326203
## 6    Africa      Burundi       339.2965       379.5646  7021078  8390505

head(gapminder)
```

```
##       country year population continent life_exp gdp_per_cap
## 1 Afghanistan 1952    8425333      Asia   28.801    779.4453
## 2 Afghanistan 1957    9240934      Asia   30.332    820.8530
## 3 Afghanistan 1962   10267083      Asia   31.997    853.1007
## 4 Afghanistan 1967   11537966      Asia   34.020    836.1971
## 5 Afghanistan 1972   13079460      Asia   36.088    739.9811
## 6 Afghanistan 1977   14880372      Asia   38.438    786.1134
```

We can that the wide version now has a separate column for each year of GDP, life expectancy, and population. This data becomes much easier to work with and understand if we can make year into a column. First, we use the "'gather()"' function to create a long version of the data where we only have 4 columns, one each for continent and country, one (obstype_year) that gives the key for the values (obs_values) in the fourth column.

```
gapminder_long <- gapminder_wide %>%
  gather(key   = obstype_year, # name new column for ID variable
         value = obs_values,   # name column that will contain the observations
         starts_with('pop'),             # only include columns that start
         starts_with('lifeExp'),         # pop, lifeExp, or gdpPercap
         starts_with('gdpPercap'))

head(gapminder_long)
```

```
##   continent      country obstype_year obs_values
## 1    Africa      Algeria     pop_1952    9279525
## 2    Africa       Angola     pop_1952    4232095
## 3    Africa        Benin     pop_1952    1738315
## 4    Africa     Botswana     pop_1952     442308
## 5    Africa Burkina Faso     pop_1952    4469979
## 6    Africa      Burundi     pop_1952    2445618
```

```
tail(gapminder_long)
```

```
##      continent        country   obstype_year obs_values
## 5107    Europe         Sweden gdpPercap_2007  33859.748
## 5108    Europe    Switzerland gdpPercap_2007  37506.419
## 5109    Europe         Turkey gdpPercap_2007   8458.276
## 5110    Europe United Kingdom gdpPercap_2007  33203.261
## 5111   Oceania      Australia gdpPercap_2007  34435.367
## 5112   Oceania    New Zealand gdpPercap_2007  25185.009
```

Next, we separate the obstype_year variable into two separate columns, one that contains the year of the observation and another that tells whether the value in obs_values is gdp, life expectancy, or population.

```
# separate the obstype_year into two
gapminder_long <- gapminder_long %>%
  separate(obstype_year,
           into = c('obs_type','year'),
           sep="_")

head(gapminder_long)
```

```
##   continent      country obs_type year obs_values
```

```
## 1    Africa      Algeria      pop 1952    9279525
## 2    Africa       Angola      pop 1952    4232095
## 3    Africa        Benin      pop 1952    1738315
## 4    Africa     Botswana      pop 1952     442308
## 5    Africa Burkina Faso      pop 1952    4469979
## 6    Africa      Burundi      pop 1952    2445618

tail(gapminder_long)

##      continent        country  obs_type year obs_values
## 5107    Europe         Sweden gdpPercap 2007  33859.748
## 5108    Europe    Switzerland gdpPercap 2007  37506.419
## 5109    Europe         Turkey gdpPercap 2007   8458.276
## 5110    Europe United Kingdom gdpPercap 2007  33203.261
## 5111   Oceania      Australia gdpPercap 2007  34435.367
## 5112   Oceania    New Zealand gdpPercap 2007  25185.009
```

Finally, we spread the long version of the data to get the original intermediate version.

```
gapminder_normal <- gapminder_long %>%
  spread(obs_type, obs_values)

# check that is looks like the original data
head(gapminder_normal)

##   continent country year gdpPercap lifeExp      pop
## 1    Africa Algeria 1952  2449.008  43.077  9279525
## 2    Africa Algeria 1957  3013.976  45.685 10270856
## 3    Africa Algeria 1962  2550.817  48.303 11000948
## 4    Africa Algeria 1967  3246.992  51.407 12760499
## 5    Africa Algeria 1972  4182.664  54.518 14760787
## 6    Africa Algeria 1977  4910.417  58.014 17152804

dim(gapminder_normal)

## [1] 1704    6

dim(gapminder_long)

## [1] 5112    5
```