Gapminder exploration

Zoe Vernon

8/31/2018

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)

1. We are interested in exploring life expectancy as a function of GDP. Create scatterplot of life expectancy versus GDP using `ggplot()` where the size of points are based on the population of the country and they are colored by the continent the country resides in.
2. Next, we explore change in life expectancy over time. Create a series of boxplots using `ggplot()`, one for each year-continent combination, where each data point corresponds to the life expectency of a country for the given year in the given continent.

2.1 Comparing GDP across continents (dplyr)

1. Compute the mean and variance of the GDP for each continent without using `dplyr()`
2. Perform the same computation using `group_by()` and `summarize()` and print the table using the `kable()` function
3. Next, we want to ask about raw GDP (i.e. overall GDP for each country, rather than standardized by per capita). Create a table that shows the average total GDP for each continent in 2007.

2.2 Using `tidyr()` with the gapminder 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. Transform the `gapminder_wide` dataset so that it is in the same form as the original gapminder dataset.

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

# look at wide_data compared 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
```

1. Use the `gather()` and `separate()` functions to create a long version of the data where we only have five columns: continent, country, the value of an observation, the type of observation (i.e. GDP, life expectancy, or population size), and the year of the observation.
2. Finally, use `spread()` to convert the long version of the data to get the original intermediate version.