HDS 5230 High performance computing

Homework Week 1

*Miao Cai**

2019-01-17

Questions

1) Load the data.table package, then convert this dataframe into a data.table. Save the resulting data.table as `gapminder_dt`.

```
library(gapminder)
library(data.table)
data(gapminder)
gapminder_dt = as.data.table(gapminder)
```

2) Practicing using the i and j arguments to subset the data.table by writing the code to get the following subsets:

a. Only the first 30 rows

```
gapminder_dt[1:30]
```

```
##         country continent year lifeExp      pop gdpPercap
##  1: Afghanistan      Asia 1952  28.801  8425333  779.4453
##  2: Afghanistan      Asia 1957  30.332  9240934  820.8530
##  3: Afghanistan      Asia 1962  31.997 10267083  853.1007
##  4: Afghanistan      Asia 1967  34.020 11537966  836.1971
##  5: Afghanistan      Asia 1972  36.088 13079460  739.9811
##  6: Afghanistan      Asia 1977  38.438 14880372  786.1134
##  7: Afghanistan      Asia 1982  39.854 12881816  978.0114
##  8: Afghanistan      Asia 1987  40.822 13867957  852.3959
##  9: Afghanistan      Asia 1992  41.674 16317921  649.3414
## 10: Afghanistan      Asia 1997  41.763 22227415  635.3414
## 11: Afghanistan      Asia 2002  42.129 25268405  726.7341
## 12: Afghanistan      Asia 2007  43.828 31889923  974.5803
## 13:     Albania    Europe 1952  55.230  1282697 1601.0561
## 14:     Albania    Europe 1957  59.280  1476505 1942.2842
## 15:     Albania    Europe 1962  64.820  1728137 2312.8890
## 16:     Albania    Europe 1967  66.220  1984060 2760.1969
## 17:     Albania    Europe 1972  67.690  2263554 3313.4222
## 18:     Albania    Europe 1977  68.930  2509048 3533.0039
## 19:     Albania    Europe 1982  70.420  2780097 3630.8807
## 20:     Albania    Europe 1987  72.000  3075321 3738.9327
## 21:     Albania    Europe 1992  71.581  3326498 2497.4379
## 22:     Albania    Europe 1997  72.950  3428038 3193.0546
## 23:     Albania    Europe 2002  75.651  3508512 4604.2117
## 24:     Albania    Europe 2007  76.423  3600523 5937.0295
## 25:     Algeria    Africa 1952  43.077  9279525 2449.0082
## 26:     Algeria    Africa 1957  45.685 10270856 3013.9760
## 27:     Algeria    Africa 1962  48.303 11000948 2550.8169
## 28:     Algeria    Africa 1967  51.407 12760499 3246.9918
```

*Department of Epidemiology and Biostatistics, Saint Louis University. Email address miao.cai@slu.edu

```
## 29:     Algeria    Africa 1972  54.518 14760787 4182.6638
## 30:     Algeria    Africa 1977  58.014 17152804 4910.4168
##         country continent year lifeExp      pop gdpPercap
```

b. Only rows where year is 1952

```
gapminder_dt[year == 1952]
```

```
##                 country continent year lifeExp      pop gdpPercap
##   1:        Afghanistan      Asia 1952  28.801  8425333  779.4453
##   2:            Albania    Europe 1952  55.230  1282697 1601.0561
##   3:            Algeria    Africa 1952  43.077  9279525 2449.0082
##   4:             Angola    Africa 1952  30.015  4232095 3520.6103
##   5:          Argentina  Americas 1952  62.485 17876956 5911.3151
##  ---
## 138:            Vietnam      Asia 1952  40.412 26246839  605.0665
## 139: West Bank and Gaza      Asia 1952  43.160  1030585 1515.5923
## 140:        Yemen, Rep.      Asia 1952  32.548  4963829  781.7176
## 141:             Zambia    Africa 1952  42.038  2672000 1147.3888
## 142:           Zimbabwe    Africa 1952  48.451  3080907  406.8841
```

c. Only the rows where continent is Africa

```
gapminder_dt[continent == "Africa"]
```

```
##       country continent year lifeExp      pop gdpPercap
##   1:  Algeria    Africa 1952  43.077  9279525 2449.0082
##   2:  Algeria    Africa 1957  45.685 10270856 3013.9760
##   3:  Algeria    Africa 1962  48.303 11000948 2550.8169
##   4:  Algeria    Africa 1967  51.407 12760499 3246.9918
##   5:  Algeria    Africa 1972  54.518 14760787 4182.6638
##  ---
## 620: Zimbabwe    Africa 1987  62.351  9216418  706.1573
## 621: Zimbabwe    Africa 1992  60.377 10704340  693.4208
## 622: Zimbabwe    Africa 1997  46.809 11404948  792.4500
## 623: Zimbabwe    Africa 2002  39.989 11926563  672.0386
## 624: Zimbabwe    Africa 2007  43.487 12311143  469.7093
```

d. Only rows where the year is 2007, with only the country column and the lifeExp column.

```
gapminder_dt[year == 2007, .(country, lifeExp)]
```

```
##                 country lifeExp
##   1:        Afghanistan  43.828
##   2:            Albania  76.423
##   3:            Algeria  72.301
##   4:             Angola  42.731
##   5:          Argentina  75.320
##  ---
## 138:            Vietnam  74.249
## 139: West Bank and Gaza  73.422
## 140:        Yemen, Rep.  62.698
## 141:             Zambia  42.384
## 142:           Zimbabwe  43.487
```

3) Now you will need to figure out what code to write to answer the following questions:

a. Which 5 countries have the highest population in 1952? What about 1987? What about 2007?

```
(setorder(gapminder_dt[year == 1952, .(country, pop)], -pop)[1:5])
```

```
##          country       pop
## 1:         China 556263527
## 2:         India 372000000
## 3: United States 157553000
## 4:         Japan  86459025
## 5:     Indonesia  82052000
```

```
(setorder(gapminder_dt[year == 1987, .(country, pop)], -pop)[1:5])
```

```
##          country        pop
## 1:         China 1084035000
## 2:         India  788000000
## 3: United States  242803533
## 4:     Indonesia  169276000
## 5:        Brazil  142938076
```

```
(setorder(gapminder_dt[year == 2007, .(country, pop)], -pop)[1:5])
```

```
##          country        pop
## 1:         China 1318683096
## 2:         India 1110396331
## 3: United States  301139947
## 4:     Indonesia  223547000
## 5:        Brazil  190010647
```

b. Which 5 countries have the lowest population in 1952? What about 1987? What about 2007?

```
(setorder(gapminder_dt[year == 1952, .(country, pop)], pop)[1:5])
```

```
##                  country    pop
## 1: Sao Tome and Principe  60011
## 2:              Djibouti  63149
## 3:               Bahrain 120447
## 4:               Iceland 147962
## 5:               Comoros 153936
```

```
(setorder(gapminder_dt[year == 1987, .(country, pop)], pop)[1:5])
```

```
##                  country    pop
## 1: Sao Tome and Principe 110812
## 2:               Iceland 244676
## 3:              Djibouti 311025
## 4:     Equatorial Guinea 341244
## 5:               Comoros 395114
```

```
(setorder(gapminder_dt[year == 2007, .(country, pop)], pop)[1:5])
```

```
##                  country    pop
## 1: Sao Tome and Principe 199579
## 2:               Iceland 301931
## 3:              Djibouti 496374
## 4:     Equatorial Guinea 551201
## 5:            Montenegro 684736
```

c. Which 5 countries have the highest lifeExp in 1952? What about 1987? What about 2007?

```
(setorder(gapminder_dt[year == 1952, .(country, lifeExp)], -lifeExp)[1:5])
```

```
##        country lifeExp
## 1:      Norway   72.67
## 2:     Iceland   72.49
## 3: Netherlands   72.13
## 4:      Sweden   71.86
## 5:     Denmark   70.78
```

```
(setorder(gapminder_dt[year == 1987, .(country, lifeExp)], -lifeExp)[1:5])
```

```
##        country lifeExp
## 1:       Japan   78.67
## 2: Switzerland   77.41
## 3:     Iceland   77.23
## 4:      Sweden   77.19
## 5:       Spain   76.90
```

```
(setorder(gapminder_dt[year == 2007, .(country, lifeExp)], -lifeExp)[1:5])
```

```
##             country lifeExp
## 1:            Japan  82.603
## 2: Hong Kong, China  82.208
## 3:          Iceland  81.757
## 4:      Switzerland  81.701
## 5:        Australia  81.235
```

d. Which 5 countries have the lowest lifeExp in 1952? What about 1987? What about 2007?

```
(setorder(gapminder_dt[year == 1952, .(country, lifeExp)], lifeExp)[1:5])
```

```
##         country lifeExp
## 1:  Afghanistan  28.801
## 2:       Gambia  30.000
## 3:       Angola  30.015
## 4: Sierra Leone  30.331
## 5:   Mozambique  31.286
```

```
(setorder(gapminder_dt[year == 1987, .(country, lifeExp)], lifeExp)[1:5])
```

```
##          country lifeExp
## 1:        Angola  39.906
## 2:  Sierra Leone  40.006
## 3:   Afghanistan  40.822
## 4: Guinea-Bissau  41.245
## 5:    Mozambique  42.861
```

```
(setorder(gapminder_dt[year == 2007, .(country, lifeExp)], lifeExp)[1:5])
```

```
##         country lifeExp
## 1:    Swaziland  39.613
## 2:   Mozambique  42.082
## 3:       Zambia  42.384
## 4: Sierra Leone  42.568
## 5:      Lesotho  42.592
```

e. Calculate the average life expectancy by country across all years for only countries in Asia. Which Country in Asia has the highest and lowest average life expectancy?

```
# The highest average life expectancy country in Asia
(setorder(gapminder_dt[continent == "Asia",
                       .(mean_lifeExp = mean(lifeExp)),
                       by = country],
          -mean_lifeExp))[1]
```

```
##    country mean_lifeExp
## 1:   Japan     74.82692
```

```
# The lowest average life expectancy country in Asia
(setorder(gapminder_dt[continent == "Asia",
                       .(mean_lifeExp = mean(lifeExp)),
                       by = country],
          mean_lifeExp))[1]
```

```
##        country mean_lifeExp
## 1: Afghanistan     37.47883
```

f. Create a new column that is pop*gdpPercap. Which countries have the highest value for this column?

```
options(scipen = 999)
# Top ten countries
(setorder(gapminder_dt[, .(country, pop_gdp = pop*gdpPercap)], -pop_gdp)[1:10])
```

```
##           country        pop_gdp
##  1: United States 12934458535085
##  2: United States 11247278678121
##  3: United States  9761353098899
##  4: United States  8221624217606
##  5: United States  7256025860958
##  6:         China  6539500929092
##  7: United States  5806915391021
##  8: United States  5301732427679
##  9: United States  4576999719662
## 10:         Japan  4035134797102
```

4) What is the correlation between gdpPercap and Life expectancy? Using your data.table object, call the correlation function from the j argument (instead of doing it the 'base' R way).

```
gapminder_dt[, cor(gdpPercap, lifeExp)]
```

```
## [1] 0.5837062
```

5) HARDER QUESTION: I want you to create a linear model estimating the slope between year and lifeExp for each country individually, then extract the slopes/country names and save the resulting table. This resulting table should have one row for every country, with two variables (country and slope). It may be helpful to start by just fitting a model for the whole dataset and figuring out how to extract the slope from that model.

a. Which 5 countries have the highest average increase in life expectancy over time (biggest slope)?

```
q5 = gapminder_dt[, .(increase = lm(lifeExp ~ year)$coefficients),
                  by = country]

setorder(q5, -increase)[1:5]
```

```
##     country    increase
## 1: Zimbabwe 236.7981946
## 2:   Zambia 165.6079669
```

```
## 3:   Rwanda 132.2049753
## 4:     Oman   0.7721790
## 5:  Vietnam   0.6716154
```

b. Which 5 countries have the lowest average increase in life expectancy over time (biggest slope)?

```
setorder(q5, increase)[1:5]
```

```
##         country  increase
## 1:         Oman -1470.086
## 2:      Vietnam -1271.983
## 3: Saudi Arabia -1227.250
## 4:    Indonesia -1201.937
## 5:        Libya -1178.944
```