R: Parallel Code, bootstrapping, and Profiling

Introduction

This assignment will cover bootstrapping, parallel code, and code profiling.

Create an rmarkdown file documenting your code and answers. Knit the markdown document to either Word or PDF format, then submit the resulting document as your homework solution. Be sure to add commentary through the assignment as appropriate.

We will use the 'gapminder' dataset, which is available in the gapminder R package. Install the gapminder package then load the data using the following:

```
install.packages('gapminder')

library(gapminder)

data(gapminder)
```

Examine the help for the dataset to understand the context.

Questions

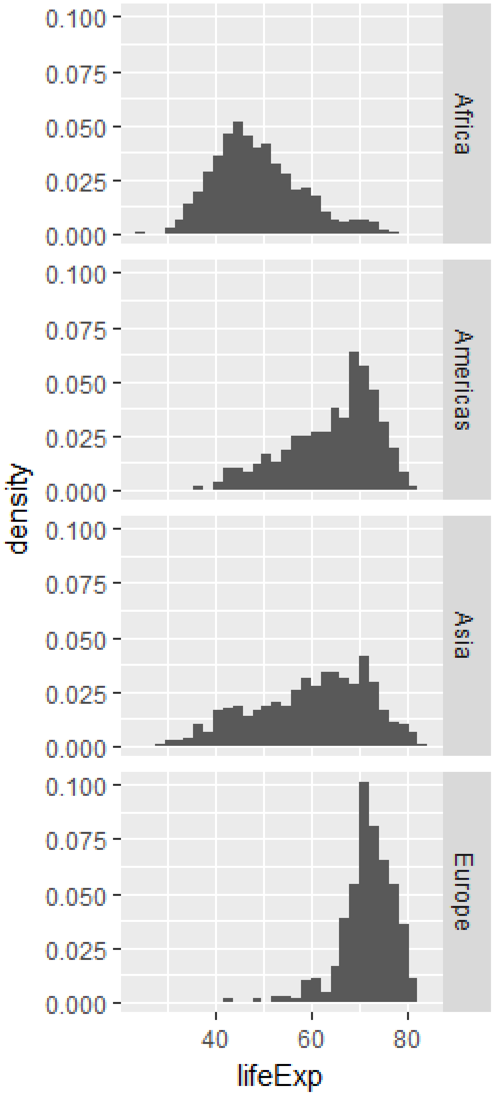
1) Load up the dataset and convert to a data.table
2) We will focus on answering the following question: 'Is the median life expectancy in 2007 for Asia, Americas, Europe, and Africa significantly different from each other? Estimate the median life expectancy in 2007 for Asia, Americas, Europe, and Africa, and bootstrap confidence intervals for each continent.'
    a. First calculate the median life expectancy for Asia, Americas, Europe, and Africa for 2007. This is your 'point estimate' of the median for each continent.
    b. For the following bootstrapped approach, implement both a parallel version using foreach(), and a non-parallel version using foreach(). Time the difference in the approaches. You can simply use Sys.time() as I did in the lecture.

c. Now bootstrap a distribution of 100,000 possible medians for each continent. You should end up with 4 separate distributions of medians (you must do this separately for each continent – in class, we only did a single bootstrapped distribution).

d. Summarize the 95% of these 4 distributions by calculating the 2.5% quantile and the 97.5% quantile. This is done separately for each continent (each distribution). You can just directly calculate this using the quantile function in R like so: `quantile(sample, c(.975,.025))`

e. Create a separate histogram for each of the 4 distributions. Stack these distributions on top of each other. This is easily done with ggplot2 and facet_grid if the data is all in one dataframe in a long format with a group indicator. Here is sample code (without the bootstrap part!):

```
## sample histogram
library(ggplot2)
library(data.table)
gapminder_dt <- data.table(gapminder)
ggplot(gapminder_dt[continent != 'Oceania'],
       aes(x=lifeExp)) +
  geom_histogram(aes(y=..density..)) +
  facet_grid(continent~.)
```


density
Africa
Americas
Asia
Europe
lifeExp