Trabalhando com dados do censo 2010

Elias Teixeira Krainski

Curso de Especialização em
Data Science & Big Data
Universidade Federal do Paraná

28 Abril 2018

Introdução

Será?

http://g1.globo.com/globo-news/noticia/2013/01/profissao-de-estatistico-tem-segunda-maior-media-salarial-do-brasil.html

Edição do dia 23/01/2013

23/01/2013 10h56 - Atualizado em 23/01/2013 10h56

Profissão de estatístico tem segunda maior média salarial do Brasil

Segundo diretora do ENCE-IBGE, Denise Britz, mercado está em expansão porque a sociedade não toma decisões sem se basear em informações.

No jornal OGlobo: *média no país: R$ 5.416 por mês. Só perde para os médicos, com ganho médio mensal de R$ 6.940.*

Vamos ver…

Microdados do CENSO 2010

Disponíveis em um arquivo compactados (.zip) por estado (exceto SP que são dois arquivos) em:

https://ww2.ibge.gov.br/home/estatistica/populacao/censo2010/resultados_gerais_amostra/resultados_gerais_amostra_tab_uf_microdados.shtm

Prepara para leitura dos dados

```
setwd('~diretoriodados')
(z0 <- dir())
(zz <- z0[grep('zip', z0)][18:19]) ### PR e RJ
names(zz) <- gsub('.zip', '', zz, fixed=TRUE)
library(readr) ## para usar read_fwf() (eficiente)

### define posicoes das colunas (ver Documentacao)
ww <- fwf_positions(
  c(1, 29, 53, 159, 247, 322),
  c(7, 44, 53, 161, 253, 327),
  c('mun', 'peso', 'urb', 'grad', 'rend', 'rendt'))

### define classes (opcional, facilita)
colcl <- do.call('cols', list('i', 'd', 'i', 'i', 'd', 'd'))
```

Leitura dos dados

```
res <- lapply(zz[c(18:19)], function(z) {
  system(paste('unzip', z))
  uf <- gsub('.zip', '', z, fixed=TRUE)
  fl <- dir(paste0(uf, '/Pessoas'))
  r <- read_fwf(paste0(uf, '/Pessoas/', fl), ww, colcl)
  system(paste('rm -r', uf))
  r$peso <- r$peso <- 1e-13
  return(r)
})
setwd('..')
save('res', file='pesourbgradrend.RData', compress='xz')

load('../data/pesourbgradrend.RData')
```

Verificação inicial

```
sapply(res, dim)
```

```
##             PR      RJ
## [1,] 1293034 1143650
## [2,]       6       6
```

```
head(res[[1]],2)
```

```
##       mun     peso urb grad rend rendt
## 1 4100103 3.070993   1   NA  510     0
## 2 4100103 3.070993   1   NA   NA   510
```

```
### soma do peso amostral é a populacao total
sum(res$PR$peso) ## Populacao no PR em 2010
```

```
## [1] 10444526
```

```
tapply(res$PR$peso, res$PR$urb, sum) ## Urbana e rural
```

```
##       1       2
## 8913240 1531286
```

Salário dos estatísticos

```
### Estatisticos
sapply(res, function(x) table(x$grad==462)) ## na amostra
```

```
##           PR    RJ
## FALSE 78478 89612
## TRUE     52   163
```

```
(etot <- sapply(res, function(x)
    sum(x$peso[which(x$grad==462)]))) ## total
```

```
##        PR        RJ
##  899.6029 2874.5831
```

```
stot <- sapply(res, function(x) {
    ii <- which(x$grad==462)
    sum(x$peso[ii]*x$rendt[ii], na.rm=TRUE)
})
stot/etot ### rendimento total medio
```

```
##       PR       RJ
## 5802.619 5157.721
```