data.Table/dplyr/plyr

Montse Figueiro

7 de junio de 2016

## **Ejercicio tasa.paro**

En cada periodo, calcula el porcentaje de parados que aporta cada provincia (de manera que la suma de los porcentajes de todas las provincias sumen el 100% en cada periodo. Hazlo para el total por sexos (hombres + mujeres). Usa los tres paquetes: plyr, data.table y dplyr.

### **Con data.table:**

library(data.table)  
paro <- read.table("paro.csv", header = T, sep = "\t")  
paro$Periodo <- gsub("IV", "4", paro$Periodo)  
paro$Periodo <- gsub("III", "3", paro$Periodo)  
paro$Periodo <- gsub("II", "2", paro$Periodo)  
paro$Periodo <- gsub("I", "1", paro$Periodo)  
  
paro$Situation <- as.character(paro$Situation)  
  
paro$Situation[paro$Situation == "Active population"] <- "active"  
paro$Situation[paro$Situation == "Inactive persons"] <- "inactive"  
  
  
paro$Situation[paro$Situation == "Unemployed persons"] <- "unemployed"  
paro$Situation[paro$Situation == "Employed persons"] <- "employed"  
paro$Situation[paro$Situation == "Parados que buscan primer empleo"] <- "never\_employed"  
  
paro$Situation <- factor(paro$Situation)

parados <- paro[paro$Situation=="unemployed",]  
parados.dt <- data.table(parados)  
res <- parados.dt[, list(total = sum(value)), by = c("Periodo", "Provinces")]  
res2 <- res[, pct := 100\* total / sum(total), by = c("Periodo")]  
head(res2)

## Periodo Provinces total pct  
## 1: 2014Q4 02 Albacete 51.1 0.9363090  
## 2: 2014Q4 03 Alicante/Alacant 221.5 4.0585605  
## 3: 2014Q4 04 AlmerÃ­a 126.5 2.3178687  
## 4: 2014Q4 01 Araba/Ãlava 27.0 0.4947230  
## 5: 2014Q4 33 Asturias 98.4 1.8029903  
## 6: 2014Q4 05 Ãvila 18.9 0.3463061

#### *comprobación:*

a <- res2[res2$Periodo=="2014Q4",]  
sum(a$pct)

## [1] 100

sum(res2$pct)#igual a 16, cada periodo suma le 100%

## [1] 1600

unique(res2$Periodo) # hay 16 periodos diferentes

## [1] "2014Q4" "2014Q3" "2014Q2" "2014Q1" "2013Q4" "2013Q3" "2013Q2"  
## [8] "2013Q1" "2012Q4" "2012Q3" "2012Q2" "2012Q1" "2011Q4" "2011Q3"  
## [15] "2011Q2" "2011Q1"

### **Con plyr:**

library(plyr)  
parados2 <- parados  
sum(parados2$value)

## [1] 89941.6

parados2$Gender <- NULL  
 prueba2 <- ddply(parados2, .(Periodo,Provinces), plyr::summarize, total= sum(value))  
 prueba2 <- ddply(prueba2,.(Periodo),transform,pct=100\*total/sum(total))  
 head(prueba2)

## Periodo Provinces total pct  
## 1 2011Q1 01 Araba/Ãlava 20.0 0.4063884  
## 2 2011Q1 02 Albacete 45.3 0.9204698  
## 3 2011Q1 03 Alicante/Alacant 205.7 4.1797050  
## 4 2011Q1 04 AlmerÃ­a 114.9 2.3347015  
## 5 2011Q1 05 Ãvila 20.8 0.4226440  
## 6 2011Q1 06 Badajoz 88.8 1.8043646

sum(prueba2$total)

## [1] 89941.6

#### *comprobación:*

sum(prueba2$pct)

## [1] 1600

### **Con dplyr:**

library(dplyr)  
parados <- paro[paro$Situation=="unemployed",]  
prueba3 <- parados%>%dplyr::select(Periodo,Provinces,value)%>%group\_by(Provinces,Periodo)%>%  
 mutate(tottal=sum(value))%>%dplyr::select(Periodo,Provinces,tottal)%>%filter(!duplicated(Provinces))%>%  
 group\_by(Periodo)%>%mutate(pct=100\*tottal/sum(tottal))  
prueba3

## Source: local data frame [832 x 4]  
## Groups: Periodo [16]  
##   
## Periodo Provinces tottal pct  
## (chr) (fctr) (dbl) (dbl)  
## 1 2014Q4 02 Albacete 51.1 0.9363090  
## 2 2014Q4 03 Alicante/Alacant 221.5 4.0585605  
## 3 2014Q4 04 AlmerÃ­a 126.5 2.3178687  
## 4 2014Q4 01 Araba/Ãlava 27.0 0.4947230  
## 5 2014Q4 33 Asturias 98.4 1.8029903  
## 6 2014Q4 05 Ãvila 18.9 0.3463061  
## 7 2014Q4 06 Badajoz 100.4 1.8396365  
## 8 2014Q4 07 Balears, Illes 111.4 2.0411903  
## 9 2014Q4 08 Barcelona 545.2 9.9897391  
## 10 2014Q4 48 Bizkaia 100.5 1.8414688  
## .. ... ... ... ...

#### *comprobación:*

sum(prueba3$pct) #16 esto si que representa 100% cada periodo, porcentaje que cada provincia representa en cada periodo.

## [1] 1600