Assignment 8

edelsonc

10/2/2016

Florida publishes the salaries of all state employees, as well as csv files containing the same information. We will use this imformation to learn about the people who work for the Florida State School System.

Importing the Data

Data for employees of the Florida State University System was downloaded from https://prod.flbog.net:4445/pls/apex/f?p=140:1:0::::: and saved locally as a single csv. This file was then read into memory

```
# reading the FLorida University System (FUS) data into R
FUS <- read.csv("emp.csv")
str(FUS)
```

```
## 'data.frame':    86021 obs. of  12 variables:
##  $ University     : Factor w/ 12 levels "FAMU","FAU","FGCU",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Budget.Entity  : Factor w/ 13 levels "","Auxiliaries",..: 6 4 6 6 4 6 4 7 6 6 ...
##  $ Position.Number: Factor w/ 52730 levels "0","00000","00000000",..: 26326 26383 26383 26334 26186 1
##  $ Last.Name      : Factor w/ 24693 levels "AAGARD","AALO",..: 11 13 13 29 61 61 61 79 79 80 ...
##  $ First.Name     : Factor w/ 10452 levels "A","A COSKUN",..: 7587 6419 6419 10111 8285 8285 8285 200
##  $ MI             : Factor w/ 28 levels " ",".","A","B",..: 21 6 6 2 27 27 27 2 2 20 ...
##  $ Employee.Type  : Factor w/ 2 levels "OPS","SALARIED": 2 2 2 2 2 2 2 2 2 2 ...
##  $ FTE            : num  0.77 0.3 0.7 0.75 0.2 0.72 0.08 0.41 0.59 0.75 ...
##  $ Class.Code     : Factor w/ 2446 levels "0001","0002",..: 1840 1840 1840 1841 1840 1840 1840 2144 1
##  $ Class.Title    : Factor w/ 4289 levels "","ACAD ADVSR/RETENTION SPCLST",..: 3377 3377 3377 583 337
##  $ Annual.Salary  : int  160000 34023 79387 92195 26911 97457 10186 22022 31691 45345 ...
##  $ OPS.Term.Amount: int  NA NA NA NA NA NA NA NA NA NA ...
```

As seen, it is already a dataframe, and primarily contains factor data.

Number of Employees

Since each row constitues a single employee, we can get a first glance at how many employees there are just by looking at the number of rows in our data frame

```
nrow(FUS)
```

```
## [1] 86021
```

So at most 86021 people. However, this doesn't account for the fact that some people may work multiple jobs in the Florida State. University System. This can be fixed by instead counting the number of unique names (assuming nobody is adobting multiple names between jobs)

```
length(unique(paste(FUS$Last.Name, FUS$First.Name)))
```

```
## [1] 52236
```

This gives us a total of 52236 employees. There is also a problem with this method, in that it counts people with the same name as the same person. So this number may underestimate the total number of state university system employees. So it is probably safe to assume the real number of employees lies in the range 52236 - 86021.

Number of Professors

We can easily find the number of full professors by simply checking if a rows class title is 'Professor'

```
sum(FUS$Class.Title == "PROFESSOR")
```

```
## [1] 6978
```

Alternativly, we could have checked who had a class code of 9001

```
sum(FUS$Class.Code == 9001)
```

```
## [1] 6980
```

Surprisingly, there is are two rows that are different for the two searches. We can easily find these by combining the previous two expressions

```
not_prof <- FUS$Class.Title != "PROFESSOR" & FUS$Class.Code == 9001
FUS[not_prof,c(1,4,5,9,10)]
```

```
##       University Last.Name First.Name Class.Code Class.Title
## 18296        FPU     AVENT      RANDY       9001   PRESIDENT
## 18297        FPU     AVENT      RANDY       9001   PRESIDENT
```

Here it appears that either one of the two pieces of information were entered incorrectly, and the the row, unfortunatly, got duplicated.

Median Salary

There are a number of options for getting this infomration. The simplest is to subset based on our earlier logic and use the `summary` function for the five number summary

```
prof <- FUS$Class.Title == "PROFESSOR"
summary(FUS[prof, 11])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       3   24530   80500   87650  123400  984800     404
```

showing that the median salary is 80500 (with some poor guy getting paid pennies because someone mistyped).

Number of Female Professors

```
# download html table with names
wget http://deron.meranda.us/data/census-dist-female-first.txt
awk '{print $1}' census-dist-female-first.txt > names_female.csv
```

This created a new file that was a single column of female names

```
head names_female.csv
```

```
## MARY
## PATRICIA
## LINDA
## BARBARA
## ELIZABETH
## JENNIFER
## MARIA
## SUSAN
## MARGARET
## DOROTHY
```

Now all we have to do is merge our data based on this `names_female.csv` on the `First.Name` column (an inner join) and well have a dataframe of female professors

```
# create names dataframe
fnames <- read.csv("names_female.csv", header=FALSE)

# merge without sort on first names
female_prof <- merge(FUS[prof,], fnames, by.x = "First.Name",
                     by.y = "V1", sort = FALSE)

# view structure of new dataframe
str(female_prof)
```

```
## 'data.frame':    4668 obs. of  12 variables:
##  $ First.Name     : Factor w/ 10452 levels "A","A COSKUN",..: 6419 6419 6419 6419 6419 6419 6419 6419
##  $ University     : Factor w/ 12 levels "FAMU","FAU","FGCU",..: 9 8 1 9 9 9 11 11 8 6 ...
##  $ Budget.Entity  : Factor w/ 13 levels "","Auxiliaries",..: 6 6 6 6 6 4 4 6 6 6 ...
##  $ Position.Number: Factor w/ 52730 levels "0","00000","00000000",..: 7817 31562 25680 7817 6771 677
##  $ Last.Name      : Factor w/ 24693 levels "AAGARD","AALO",..: 11208 4886 22167 11208 11323 11323 70
##  $ MI             : Factor w/ 28 levels " ",".","A","B",..: 7 3 2 7 12 12 7 12 5 18 ...
##  $ Employee.Type  : Factor w/ 2 levels "OPS","SALARIED": 2 2 2 2 2 2 2 2 2 2 ...
##  $ FTE            : num  0.7 0.75 1 0.3 0.6 0.4 0.18 1 1 0.75 ...
##  $ Class.Code     : Factor w/ 2446 levels "0001","0002",..: 1840 1840 1840 1840 1840 1840 1840 1840
##  $ Class.Title    : Factor w/ 4289 levels "","ACAD ADVSR/RETENTION SPCLST",..: 3377 3377 3377 3377 33
##  $ Annual.Salary  : int  101666 185000 196445 43571 96665 64443 28110 166338 176636 102541 ...
##  $ OPS.Term.Amount: int  NA NA NA NA NA NA NA NA NA NA ...
```

```
# count number of rows (number of female professors)
nrow(female_prof)
```

```
## [1] 4668
```