Homework8data.tablekey

James Dickens

3/28/2021

library(tidyverse)

## -- Attaching packages --------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.5  
## v tidyr 1.1.1 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(dplyr)  
library(data.table)

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

## The following object is masked from 'package:purrr':  
##   
## transpose

# 1.a) Use `fread()` and relative paths to load in the   
# `data.table` from college\_score.csv.  
  
collegedata <- fread("college\_score.csv")  
collegedata

## UNITID OPEID MN\_EARN\_WNE\_P6  
## 1: 100654 100200 27800  
## 2: 100663 105200 37600  
## 3: 100690 2503400 39400  
## 4: 100706 105500 41300  
## 5: 100724 100500 23500  
## ---   
## 7171: 48305401 4224601 <NA>  
## 7172: 48387801 4223701 <NA>  
## 7173: 48400201 4228101 <NA>  
## 7174: 48414501 4232401 <NA>  
## 7175: 48778201 4221501 <NA>  
## INSTNM SAT\_AVG ADM\_RATE  
## 1: Alabama A & M University 849 0.8738  
## 2: University of Alabama at Birmingham 1125 0.5814  
## 3: Amridge University NA NA  
## 4: University of Alabama in Huntsville 1257 0.7628  
## 5: Alabama State University 825 0.4590  
## ---   
## 7171: Barber School of Pittsburgh - Ambridge NA NA  
## 7172: Bay Area Medical Academy - San Jose Satellite Location NA NA  
## 7173: High Desert Medical College NA NA  
## 7174: Vantage College-San Antonio NA NA  
## 7175: BCI - Malden NA NA  
## UGDS COSTT4\_A AVGFACSAL GRAD\_DEBT\_MDN AGE\_ENTRY ICLEVEL  
## 1: 4616 22667 7028 32750 20.28374137 1  
## 2: 12047 22684 10517 21833 23.60797466 1  
## 3: 293 13380 3857 22890 33.6722973 1  
## 4: 6346 22059 9463 22647 22.72791963 1  
## 5: 4704 19242 7952 31500 20.13099042 1  
## ---   
## 7171: NA NA NA 13583 28.84732824 3  
## 7172: NA NA NA 8850 26.50239234 3  
## 7173: NA NA NA 7521 24.96246649 3  
## 7174: NA NA NA 9891.5 28.38990826 3  
## 7175: NA NA NA 4229 30.18333333 3

# b) Use and show data.table R coding to determine how  
# to determine how many rows the college\_score data  
# has.  
collegedata[, .(.N)]

## N  
## 1: 7175

# c) Use read\_csv and relative paths to load in the   
# `data.table` from college\_score.csv.  
collegedata1 <- read\_csv("college\_score.csv")

## Parsed with column specification:  
## cols(  
## UNITID = col\_double(),  
## OPEID = col\_double(),  
## MN\_EARN\_WNE\_P6 = col\_character(),  
## INSTNM = col\_character(),  
## SAT\_AVG = col\_double(),  
## ADM\_RATE = col\_double(),  
## UGDS = col\_double(),  
## COSTT4\_A = col\_double(),  
## AVGFACSAL = col\_double(),  
## GRAD\_DEBT\_MDN = col\_character(),  
## AGE\_ENTRY = col\_character(),  
## ICLEVEL = col\_double()  
## )

collegedata1

## # A tibble: 7,175 x 12  
## UNITID OPEID MN\_EARN\_WNE\_P6 INSTNM SAT\_AVG ADM\_RATE UGDS COSTT4\_A AVGFACSAL  
## <dbl> <dbl> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 100654 1.00e5 27800 Alaba~ 849 0.874 4616 22667 7028  
## 2 100663 1.05e5 37600 Unive~ 1125 0.581 12047 22684 10517  
## 3 100690 2.50e6 39400 Amrid~ NA NA 293 13380 3857  
## 4 100706 1.06e5 41300 Unive~ 1257 0.763 6346 22059 9463  
## 5 100724 1.00e5 23500 Alaba~ 825 0.459 4704 19242 7952  
## 6 100751 1.05e5 38500 The U~ 1202 0.526 31663 28422 9802  
## 7 100760 1.01e5 26000 Centr~ NA NA 1492 13868 5960  
## 8 100812 1.01e5 37400 Athen~ NA NA 2888 NA 8367  
## 9 100830 8.31e5 33200 Aubur~ 1009 0.766 4171 19255 7251  
## 10 100858 1.01e5 41700 Aubur~ 1217 0.805 22095 29794 9945  
## # ... with 7,165 more rows, and 3 more variables: GRAD\_DEBT\_MDN <chr>,  
## # AGE\_ENTRY <chr>, ICLEVEL <dbl>

# d) Use tidyverse coding to determine how many rows the  
# college\_score data has.  
collegedata1 %>%  
 count()

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 7175

#2 Use a data.table command to display the structure of  
# the college\_score data  
str(collegedata)

## Classes 'data.table' and 'data.frame': 7175 obs. of 12 variables:  
## $ UNITID : int 100654 100663 100690 100706 100724 100751 100760 100812 100830 100858 ...  
## $ OPEID : int 100200 105200 2503400 105500 100500 105100 100700 100800 831000 100900 ...  
## $ MN\_EARN\_WNE\_P6: chr "27800" "37600" "39400" "41300" ...  
## $ INSTNM : chr "Alabama A & M University" "University of Alabama at Birmingham" "Amridge University" "University of Alabama in Huntsville" ...  
## $ SAT\_AVG : int 849 1125 NA 1257 825 1202 NA NA 1009 1217 ...  
## $ ADM\_RATE : num 0.874 0.581 NA 0.763 0.459 ...  
## $ UGDS : int 4616 12047 293 6346 4704 31663 1492 2888 4171 22095 ...  
## $ COSTT4\_A : int 22667 22684 13380 22059 19242 28422 13868 NA 19255 29794 ...  
## $ AVGFACSAL : int 7028 10517 3857 9463 7952 9802 5960 8367 7251 9945 ...  
## $ GRAD\_DEBT\_MDN : chr "32750" "21833" "22890" "22647" ...  
## $ AGE\_ENTRY : chr "20.28374137" "23.60797466" "33.6722973" "22.72791963" ...  
## $ ICLEVEL : int 1 1 1 1 1 1 2 1 1 1 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

#3. Use data.table to calculate the median for average  
# faculty salaries.  
collegedata[, .(AVEFALSAL = median(AVGFACSAL, na = T))]

## AVEFALSAL  
## 1: 6217

# 4 Use Tidyverse coding to change the levels of the variable  
# ICLEVEL. In particular change 1 to OneYear, 2 to TwoYear,  
# and 3 to ThreeYear. Select only variables INSTNM, SAT\_AVG,  
# ADM\_RATE, and ICLEVEL  
collegedata1%>%  
 select(INSTNM, SAT\_AVG, ADM\_RATE, ICLEVEL)%>%  
 mutate(ICLEVEL = recode(ICLEVEL, "1" = "OneYear",  
 "2" = "TwoYear", "3" = "ThreeYear")) -> coldata  
coldata

## # A tibble: 7,175 x 4  
## INSTNM SAT\_AVG ADM\_RATE ICLEVEL  
## <chr> <dbl> <dbl> <chr>   
## 1 Alabama A & M University 849 0.874 OneYear  
## 2 University of Alabama at Birmingham 1125 0.581 OneYear  
## 3 Amridge University NA NA OneYear  
## 4 University of Alabama in Huntsville 1257 0.763 OneYear  
## 5 Alabama State University 825 0.459 OneYear  
## 6 The University of Alabama 1202 0.526 OneYear  
## 7 Central Alabama Community College NA NA TwoYear  
## 8 Athens State University NA NA OneYear  
## 9 Auburn University at Montgomery 1009 0.766 OneYear  
## 10 Auburn University 1217 0.805 OneYear  
## # ... with 7,165 more rows

# 5 The data.table representation of your data produced  
# by the coding of #4 will show the first five rows and the  
# last five rows of the data set. Use R coding to change  
# the output of #4 to data.table output.  
  
as.data.table(coldata)

## INSTNM SAT\_AVG ADM\_RATE  
## 1: Alabama A & M University 849 0.8738  
## 2: University of Alabama at Birmingham 1125 0.5814  
## 3: Amridge University NA NA  
## 4: University of Alabama in Huntsville 1257 0.7628  
## 5: Alabama State University 825 0.4590  
## ---   
## 7171: Barber School of Pittsburgh - Ambridge NA NA  
## 7172: Bay Area Medical Academy - San Jose Satellite Location NA NA  
## 7173: High Desert Medical College NA NA  
## 7174: Vantage College-San Antonio NA NA  
## 7175: BCI - Malden NA NA  
## ICLEVEL  
## 1: OneYear  
## 2: OneYear  
## 3: OneYear  
## 4: OneYear  
## 5: OneYear  
## ---   
## 7171: ThreeYear  
## 7172: ThreeYear  
## 7173: ThreeYear  
## 7174: ThreeYear  
## 7175: ThreeYear

# 6 Now use a data.table method to output observational rows  
# of the original data set (550 - 555) inclusive.  
collegedata[c(550,551,552,553,554,555)]

## UNITID OPEID MN\_EARN\_WNE\_P6 INSTNM SAT\_AVG  
## 1: 126942 954200 29100 Community College of Denver NA  
## 2: 126951 793806 <NA> Lincoln College of Technology-Denver NA  
## 3: 126979 135200 PrivacySuppressed Denver Seminary NA  
## 4: 127060 137100 48400 University of Denver 1262  
## 5: 127158 559600 24900 Emily Griffith Technical College NA  
## 6: 127185 135300 30200 Fort Lewis College 1014  
## ADM\_RATE UGDS COSTT4\_A AVGFACSAL GRAD\_DEBT\_MDN AGE\_ENTRY ICLEVEL  
## 1: NA 7622 14516 5176 10460 25.99357172 2  
## 2: NA 1033 NA 3924 12000 23.9295393 2  
## 3: NA NA NA 5956 PrivacySuppressed <NA> 1  
## 4: 0.5347 5738 59656 10680 23804 20.26430518 1  
## 5: NA 4693 NA NA PrivacySuppressed 29.43791946 3  
## 6: 0.9212 3443 24298 7152 18866 21.81433824 1