census.R

manwthglasses

Sun Sep 30 22:05:45 2018

#Set working directory   
setwd("/home/manwthglasses/Documents/IntroToDataScience/")  
#Set URl and load dataframe  
urlToRead <- "http://www2.census.gov/programs-surveys/popest/tables/2010-2011/state/totals/nst-est2011-01.csv"  
testFrame <- read.csv(url(urlToRead))  
str(testFrame)

## 'data.frame': 66 obs. of 10 variables:  
## $ table.with.row.headers.in.column.A.and.column.headers.in.rows.3.through.4...leading.dots.indicate.sub.parts.: Factor w/ 65 levels "",".Alabama",..: 62 53 1 64 55 54 60 65 2 3 ...  
## $ X : Factor w/ 60 levels "","1,052,567",..: 1 59 60 27 38 47 10 49 32 50 ...  
## $ X.1 : Factor w/ 59 levels "","1,052,567",..: 1 1 59 27 38 47 10 49 32 50 ...  
## $ X.2 : Factor w/ 60 levels "","1,052,528",..: 1 60 21 28 39 48 10 51 33 50 ...  
## $ X.3 : Factor w/ 59 levels "","1,051,302",..: 1 1 21 28 38 48 10 50 33 51 ...  
## $ X.4 : logi NA NA NA NA NA NA ...  
## $ X.5 : logi NA NA NA NA NA NA ...  
## $ X.6 : logi NA NA NA NA NA NA ...  
## $ X.7 : logi NA NA NA NA NA NA ...  
## $ X.8 : logi NA NA NA NA NA NA ...

#Removing rows and columns  
testFrame <- testFrame[-1:-8,]  
summary(testFrame[,6:10])

## X.4 X.5 X.6 X.7   
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:58 NA's:58 NA's:58 NA's:58   
## X.8   
## Mode:logical   
## NA's:58

testFrame <- testFrame[,1:5]  
tail(testFrame,5)

## table.with.row.headers.in.column.A.and.column.headers.in.rows.3.through.4...leading.dots.indicate.sub.parts.  
## 62 Note: The April 1, 2010 Population Estimates base reflects changes to the Census 2010 population from geographic program revisions. It does not reflect changes from the Count Question Resolution program. See Geographic Terms and Definitions at http://www.census.gov/popest/about/geo/terms.html for a list of the states that are included in each region.  
## 63 Suggested Citation:  
## 64 Table 1. Annual Estimates of the Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2011 (NST-EST2011-01)  
## 65 Source: U.S. Census Bureau, Population Division  
## 66 Release Date: December 2011  
## X X.1 X.2 X.3  
## 62   
## 63   
## 64   
## 65   
## 66

testFrame <- testFrame[-52:-58,]  
str(testFrame)

## 'data.frame': 51 obs. of 5 variables:  
## $ table.with.row.headers.in.column.A.and.column.headers.in.rows.3.through.4...leading.dots.indicate.sub.parts.: Factor w/ 65 levels "",".Alabama",..: 2 3 4 5 6 7 8 9 10 11 ...  
## $ X : Factor w/ 60 levels "","1,052,567",..: 32 50 41 19 28 33 23 54 45 13 ...  
## $ X.1 : Factor w/ 59 levels "","1,052,567",..: 32 50 41 19 28 33 23 54 45 13 ...  
## $ X.2 : Factor w/ 60 levels "","1,052,528",..: 33 50 42 19 29 34 24 55 46 13 ...  
## $ X.3 : Factor w/ 59 levels "","1,051,302",..: 33 51 42 19 29 34 24 58 46 13 ...

#Renaming rows and columns  
testFrame$stateName <- testFrame[,1]  
str(testFrame)

## 'data.frame': 51 obs. of 6 variables:  
## $ table.with.row.headers.in.column.A.and.column.headers.in.rows.3.through.4...leading.dots.indicate.sub.parts.: Factor w/ 65 levels "",".Alabama",..: 2 3 4 5 6 7 8 9 10 11 ...  
## $ X : Factor w/ 60 levels "","1,052,567",..: 32 50 41 19 28 33 23 54 45 13 ...  
## $ X.1 : Factor w/ 59 levels "","1,052,567",..: 32 50 41 19 28 33 23 54 45 13 ...  
## $ X.2 : Factor w/ 60 levels "","1,052,528",..: 33 50 42 19 29 34 24 55 46 13 ...  
## $ X.3 : Factor w/ 59 levels "","1,051,302",..: 33 51 42 19 29 34 24 58 46 13 ...  
## $ stateName : Factor w/ 65 levels "",".Alabama",..: 2 3 4 5 6 7 8 9 10 11 ...

colnames(testFrame)

## [1] "table.with.row.headers.in.column.A.and.column.headers.in.rows.3.through.4...leading.dots.indicate.sub.parts."  
## [2] "X"   
## [3] "X.1"   
## [4] "X.2"   
## [5] "X.3"   
## [6] "stateName"

testFrame <- testFrame[,-1]  
colnames(testFrame)

## [1] "X" "X.1" "X.2" "X.3" "stateName"

testFrame$stateName

## [1] .Alabama .Alaska .Arizona   
## [4] .Arkansas .California .Colorado   
## [7] .Connecticut .Delaware .District of Columbia  
## [10] .Florida .Georgia .Hawaii   
## [13] .Idaho .Illinois .Indiana   
## [16] .Iowa .Kansas .Kentucky   
## [19] .Louisiana .Maine .Maryland   
## [22] .Massachusetts .Michigan .Minnesota   
## [25] .Mississippi .Missouri .Montana   
## [28] .Nebraska .Nevada .New Hampshire   
## [31] .New Jersey .New Mexico .New York   
## [34] .North Carolina .North Dakota .Ohio   
## [37] .Oklahoma .Oregon .Pennsylvania   
## [40] .Rhode Island .South Carolina .South Dakota   
## [43] .Tennessee .Texas .Utah   
## [46] .Vermont .Virginia .Washington   
## [49] .West Virginia .Wisconsin .Wyoming   
## 65 Levels: .Alabama .Alaska .Arizona .Arkansas .California ... West

#cleaning up elements  
testFrame$stateName <- gsub("\\.","", testFrame$stateName)   
testFrame$stateName

## [1] "Alabama" "Alaska" "Arizona"   
## [4] "Arkansas" "California" "Colorado"   
## [7] "Connecticut" "Delaware" "District of Columbia"  
## [10] "Florida" "Georgia" "Hawaii"   
## [13] "Idaho" "Illinois" "Indiana"   
## [16] "Iowa" "Kansas" "Kentucky"   
## [19] "Louisiana" "Maine" "Maryland"   
## [22] "Massachusetts" "Michigan" "Minnesota"   
## [25] "Mississippi" "Missouri" "Montana"   
## [28] "Nebraska" "Nevada" "New Hampshire"   
## [31] "New Jersey" "New Mexico" "New York"   
## [34] "North Carolina" "North Dakota" "Ohio"   
## [37] "Oklahoma" "Oregon" "Pennsylvania"   
## [40] "Rhode Island" "South Carolina" "South Dakota"   
## [43] "Tennessee" "Texas" "Utah"   
## [46] "Vermont" "Virginia" "Washington"   
## [49] "West Virginia" "Wisconsin" "Wyoming"

str(testFrame)

## 'data.frame': 51 obs. of 5 variables:  
## $ X : Factor w/ 60 levels "","1,052,567",..: 32 50 41 19 28 33 23 54 45 13 ...  
## $ X.1 : Factor w/ 59 levels "","1,052,567",..: 32 50 41 19 28 33 23 54 45 13 ...  
## $ X.2 : Factor w/ 60 levels "","1,052,528",..: 33 50 42 19 29 34 24 55 46 13 ...  
## $ X.3 : Factor w/ 59 levels "","1,051,302",..: 33 51 42 19 29 34 24 58 46 13 ...  
## $ stateName: chr "Alabama" "Alaska" "Arizona" "Arkansas" ...

testFrame$april10census

## NULL

testFrame$april10census <- gsub(",", "", testFrame$X)   
testFrame$april10base <- gsub(",","",testFrame$X.1)  
testFrame$july10pop <- gsub(",","",testFrame$X.2)  
testFrame$july11pop <- gsub(",","",testFrame$X.3)  
testFrame$april10census <- as.numeric(gsub(" ","", testFrame$april10census))  
testFrame$april10base <- as.numeric(gsub(" ","", testFrame$april10base))  
testFrame$july10pop <- as.numeric(gsub(" ","", testFrame$july10pop))  
testFrame$july11pop <- as.numeric(gsub(" ","", testFrame$july11pop))  
testFrame <- testFrame[,-1:-4]  
head(testFrame,5)

## stateName april10census april10base july10pop july11pop  
## 9 Alabama 4779736 4779735 4785401 4802740  
## 10 Alaska 710231 710231 714146 722718  
## 11 Arizona 6392017 6392013 6413158 6482505  
## 12 Arkansas 2915918 2915921 2921588 2937979  
## 13 California 37253956 37253956 37338198 37691912

rownames(testFrame) <- NULL  
head(testFrame,5)

## stateName april10census april10base july10pop july11pop  
## 1 Alabama 4779736 4779735 4785401 4802740  
## 2 Alaska 710231 710231 714146 722718  
## 3 Arizona 6392017 6392013 6413158 6482505  
## 4 Arkansas 2915918 2915921 2921588 2937979  
## 5 California 37253956 37253956 37338198 37691912

#Sorting data frames  
sortedStates <- testFrame[order(-testFrame$july11pop),]  
head(sortedStates,5)

## stateName april10census april10base july10pop july11pop  
## 5 California 37253956 37253956 37338198 37691912  
## 44 Texas 25145561 25145561 25253466 25674681  
## 33 New York 19378102 19378104 19395206 19465197  
## 10 Florida 18801310 18801311 18838613 19057542  
## 14 Illinois 12830632 12830632 12841980 12869257

write.csv(testFrame, "testFrame.WM.csv")