Bing-Je_Wu_HW3

Bing-Je Wu

4/19/2019

#Step 1. Create a function (readStates) to read a CSV file into R

```
urlRemote <- "https://www2.census.gov/"
path <- "programs-surveys/popest/tables/2010-2011/state/totals/"
fileName <- "nst-est2011-01.csv"
urlToRead <- paste0(urlRemote, path, fileName)
```

```
readStates <- function(inputURL) {
    library(RCurl)
    Temp <- getURL(inputURL)
    return(read.csv(text = Temp))
}
```

```
mytable <- readStates(urlToRead)
```

#Step 2. Clean the dataframe

Remove empty columns:

```
mytable <- mytable[, 1:5]
```

Remove top 8 rows:

```
mytable <- mytable[-1:-8, ]
rownames(mytable) <- NULL
```

Remove bottom 6 rows:

```
mytable <- mytable[-52:-58, ]
```

Rename column, remove the old column, and normoalize it:

```
mytable$stateName <- mytable[, 1]
mytable <- mytable[, -1]
mytable$stateName <- gsub("\\.", "", mytable$stateName)
```

Normalize X,X.1,X.2,X.3 variables:

```
mytable$base2010 <- gsub("\\,", "", mytable$X)
mytable$base2010 <- as.numeric(mytable$base2010)

mytable$base2011 <- gsub("\\,", "", mytable$X.1)
mytable$base2011 <- as.numeric(mytable$base2011)

mytable$Jul2010 <- gsub("\\,", "", mytable$X.2)
mytable$Jul2010 <- as.numeric(mytable$Jul2010)

mytable$Jul2011 <- gsub("\\,", "", mytable$X.3)
mytable$Jul2011 <- as.numeric(mytable$Jul2011)
```

Remove old X, X.1, X.2, X.3 columns:

```
mytable <- mytable[, -1:-4]
```

Analyze the dataset:

```
summary(mytable)
```

```
##   stateName            base2010           base2011       
##  Length:51          Min.   :  563626   Min.   :  563626  
##  Class :character   1st Qu.: 1696962   1st Qu.: 1696962  
##  Mode  :character   Median : 4339367   Median : 4339362  
##                     Mean   : 6053834   Mean   : 6053834  
##                     3rd Qu.: 6636084   3rd Qu.: 6636084  
##                     Max.   :37253956   Max.   :37253956  
##     Jul2010            Jul2011        
##  Min.   :  564554   Min.   :  568158  
##  1st Qu.: 1700622   1st Qu.: 1713813  
##  Median : 4347223   Median : 4369356  
##  Mean   : 6065298   Mean   : 6109645  
##  3rd Qu.: 6649208   3rd Qu.: 6708787  
##  Max.   :37338198   Max.   :37691912
```

```
str(mytable)
```

```
## 'data.frame':    51 obs. of  5 variables:
##  $ stateName: chr  "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ base2010 : num  4779736 710231 6392017 2915918 37253956 ...
##  $ base2011 : num  4779735 710231 6392013 2915921 37253956 ...
##  $ Jul2010  : num  4785401 714146 6413158 2921588 37338198 ...
##  $ Jul2011  : num  4802740 722718 6482505 2937979 37691912 ...
```

#Step 3. Store and explore the dataset

Store the dataset as dfStates:

```
dfStates <- mytable
```

Calculate the mean for the July2011 data:

```
mean(dfStates$Jul2011)
```

```
## [1] 6109645
```

#Step 4: Find the state with the Highest Population

```
dfStates[which.max(dfStates$Jul2011), ]
```

```
##    stateName base2010 base2011  Jul2010  Jul2011
## 5 California 37253956 37253956 37338198 37691912
```

Based on the July2011 data, California is the state that has the highest population.

```
dfStates[order(dfStates$Jul2011, decreasing = FALSE), ]
```

```
##                stateName base2010 base2011  Jul2010  Jul2011
## 51               Wyoming   563626   563626   564554   568158
## 9  District of Columbia   601723   601723   604912   617996
## 46              Vermont   625741   625741   625909   626431
## 35         North Dakota   672591   672591   674629   683932
## 2                Alaska   710231   710231   714146   722718
## 42         South Dakota   814180   814180   816598   824082
## 8              Delaware   897934   897934   899792   907135
## 27              Montana   989415   989415   990958   998199
## 40         Rhode Island  1052567  1052567  1052528  1051302
## 30        New Hampshire  1316470  1316472  1316807  1318194
## 20                Maine  1328361  1328361  1327379  1328188
## 12               Hawaii  1360301  1360301  1363359  1374810
## 13                Idaho  1567582  1567582  1571102  1584985
## 28             Nebraska  1826341  1826341  1830141  1842641
## 49        West Virginia  1852994  1852996  1854368  1855364
## 32           New Mexico  2059179  2059180  2065913  2082224
## 29               Nevada  2700551  2700551  2704283  2723322
## 45                 Utah  2763885  2763885  2775479  2817222
## 17               Kansas  2853118  2853118  2859143  2871238
## 4              Arkansas  2915918  2915921  2921588  2937979
## 25          Mississippi  2967297  2967297  2970072  2978512
## 16                 Iowa  3046355  3046350  3050202  3062309
## 7           Connecticut  3574097  3574097  3575498  3580709
## 37             Oklahoma  3751351  3751354  3760184  3791508
## 38               Oregon  3831074  3831074  3838332  3871859
## 18             Kentucky  4339367  4339362  4347223  4369356
## 19            Louisiana  4533372  4533372  4545343  4574836
## 41       South Carolina  4625364  4625364  4637106  4679230
## 1               Alabama  4779736  4779735  4785401  4802740
## 6              Colorado  5029196  5029196  5047692  5116796
## 24            Minnesota  5303925  5303925  5310658  5344861
## 50            Wisconsin  5686986  5686986  5691659  5711767
## 21             Maryland  5773552  5773552  5785681  5828289
## 26             Missouri  5988927  5988927  5995715  6010688
## 43            Tennessee  6346105  6346110  6357436  6403353
## 3               Arizona  6392017  6392013  6413158  6482505
## 15              Indiana  6483802  6483800  6490622  6516922
## 22        Massachusetts  6547629  6547629  6555466  6587536
## 48           Washington  6724540  6724540  6742950  6830038
## 47             Virginia  8001024  8001030  8023953  8096604
## 31           New Jersey  8791894  8791894  8799593  8821155
## 34       North Carolina  9535483  9535475  9560234  9656401
## 11              Georgia  9687653  9687660  9712157  9815210
## 23             Michigan  9883640  9883635  9877143  9876187
## 36                 Ohio 11536504 11536502 11537968 11544951
## 39         Pennsylvania 12702379 12702379 12717722 12742886
## 14             Illinois 12830632 12830632 12841980 12869257
## 10              Florida 18801310 18801311 18838613 19057542
## 33             New York 19378102 19378104 19395206 19465197
## 44                Texas 25145561 25145561 25253466 25674681
## 5            California 37253956 37253956 37338198 37691912
```

#Step 5: Explore the distribution of the states

Write a function:

```
below_percentage <- function(inputvector, inputnumber) {
    Total_number <- length(inputvector)
    Number_below <- length(inputvector[inputvector < inputnumber])
    return(Number_below/Total_number)
}
```

Test the function:

```
A = c(1, 2, 3, 4, 5)
a = 2
below_percentage(A, a)
```

```
## [1] 0.2
```

Test vector ‘dfStates$Jul2011Num'andthemeanofdfStates$Jul2011Num’:

```
below_percentage(dfStates$Jul2011, mean(dfStates$Jul2011))
```

```
## [1] 0.6666667
```