stats-380-assignment-3

Sooyong Choi

01 October 2020

```
# Question 1a

setwd("C:/Users/rick9/Desktop/Stats 380/A3")
text = readLines("sales.txt")
```

```
> text
[1] "<|Boston->{<|Date->Wed 1 Jan 2014,Sales->198|>,<|Date->Thu 2 Jan 2014,Sales->189|>,<|Date->Fri 3 Jan 2014
,Sales->196|>,<|Date->Sat 4 Jan 2014,Sales->194|>,<|Date->Sun 5 Jan 2014,Sales->193|>,<|Date->Mon 6 Jan 2014,S
```

```
# Question 1b

# finding the index for countries in the data
match = unlist(gregexpr('[a-zA-Z]+[[:space:]]?[a-zA-Z]+[[:space:]]?[a-zA-Z]+-
>\\{', text))

# finding the ending index for the end of city data
endmatch = c(unlist(gregexpr('\\},', text)), nchar(text))

splitData = substring(text, match, endmatch)
```

```
> splitData
[1] "Boston->{<|Date->Wed 1 Jan 2014,Sales->198|>,<|Date->Thu 2 Jan
2014,Sales->189|>,<|Date->Fri 3 Jan 2014,Sales->196|>,<|Date->Sat 4 Jan
2014,Sales->194|>,<|Date->Sun 5 Jan 2014,Sales->193|>,<|Date->Mon 6 Jan
2014,Sales->184|>,<|Date->Tue 7 Jan 2014,Sales->190|>,<|Date->Wed 8 Jan
2014,Sales->196|>,<|Date->Thu 9 Jan 2014,Sales->189|>,<|Date->Fri 10 Jan
2014,Sales->194|>,<|Date->Sat 11 Jan 2014,Sales->206|>,<|Date->Sun 12 Jan
2014,Sales->191|>,<|Date->Mon 13 Jan 2014,Sales->187|>,<|Date->Tue 14 Jan
2014,Sales->199|>,<|Date->Wed 15 Jan 2014,Sales->193|>,<|Date->Thu 16 Jan
2014,Sales->184|>,<|Date->Fri 17 Jan 2014,Sales->190|>,<|Date->Sat 18 Jan
2014,Sales->196|>,<|Date->Sun 19 Jan 2014,Sales->196|>,<|Date->Mon 20 Jan
2014,Sales->191|>,<|Date->Tue 21 Jan 2014,Sales->192|>,<|Date->Wed 22 Jan
2014,Sales->188|>,<|Date->Thu 23 Jan 2014,Sales->188|>,<|Date->Fri 24 Jan
2014,Sales->195|>,<|Date->Sat 25 Jan 2014,Sales->181|>,<|Date->Sun 26 Jan
2014,Sales->199|>,<|Date->Mon 27 Jan 2014,Sales->186|>,<|Date->Tue 28 Jan
2014,Sales->188|>,<|Date->Wed 29 Jan 2014,Sales->185|>,<|Date->Thu 30 Jan
2014,Sales->197|>,<|Date->Fri 31 Jan 2014,Sales->188|>,<|Date->Sat 1 Feb
2014,Sales->193|>,<|Date->Sun 2 Feb 2014,Sales->195|>,<|Date->Mon 3 Feb
2014,Sales->193|>,<|Date->Tue 4 Feb 2014,Sales->182|>,<|Date->Wed 5 Feb
2014,Sales->203|>,<|Date->Thu 6 Feb 2014,Sales->188|>,<|Date->Fri 7 Feb
2014,Sales->192|>,<|Date->Sat 8 Feb 2014,Sales->195|>,<|Date->Sun 9 Feb
2014,Sales->179|>,<|Date->Mon 10 Feb 2014,Sales->187|>,<|Date->Tue 11 Feb
2014,Sales->195|>,<|Date->Wed 12 Feb 2014,Sales->193|>,<|Date->Thu 13 Feb
2014,Sales->201|>,<|Date->Fri 14 Feb 2014,Sales->197|>,<|Date->Sat 15 Feb
2014,Sales->188|>,<|Date->Sun 16 Feb 2014,Sales->187|>,<|Date->Mon 17 Feb
2014,Sales->198|>,<|Date->Tue 18 Feb 2014,Sales->187|>,<|Date->Wed 19 Feb
2014,Sales->194|>,<|Date->Thu 20 Feb 2014,Sales->188|>,<|Date->Fri 21 Feb
```

```
2014,Sales->185|>,<|Date->Sat 22 Feb 2014,Sales->193|>,<|Date->Sun 23 Feb
2014,Sales->196|>,<|Date->Mon 24 Feb 2014,Sales->196|>,<|Date->Tue 25 Feb
2014,Sales->190|>,<|Date->Wed 26 Feb 2014,Sales->188|>,<|Date->Thu 27 Feb
2014,Sales->186|>,<|Date->Fri 28 Feb 2014,Sales->191|>,<|Date->Sat 1 Mar
2014,Sales->185|>,<|Date->Sun 2 Mar 2014,Sales->193|>,<|Date->Mon 3 Mar
2014,Sales->189|>,<|Date->Tue 4 Mar 2014,Sales->194|>,<|Date->Wed 5 Mar
2014,Sales->183|>,<|Date->Thu 6 Mar 2014,Sales->187|>,<|Date->Fri 7 Mar
2014,Sales->193|>,<|Date->Sat 8 Mar 2014,Sales->193|>,<|Date->Sun 9 Mar
2014,Sales->197|>,<|Date->Mon 10 Mar 2014,Sales->191|>,<|Date->Tue 11 Mar
2014,Sales->191|>,<|Date->Wed 12 Mar 2014,Sales->186|>,<|Date->Thu 13 Mar
2014,Sales->185|>,<|Date->Fri 14 Mar 2014,Sales->191|>,<|Date->Sat 15 Mar
2014,Sales->190|>,<|Date->Sun 16 Mar 2014,Sales->196|>,<|Date->Mon 17 Mar
2014,Sales->181|>,<|Date->Tue 18 Mar 2014,Sales->190|>,<|Date->Wed 19 Mar
2014,Sales->186|>,<|Date->Thu 20 Mar 2014,Sales->194|>,<|Date->Fri 21 Mar
2014,Sales->187|>,<|Date->Sat 22 Mar 2014,Sales->194|>,<|Date->Sun 23 Mar
2014,Sales->192|>,<|Date->Mon 24 Mar 2014,Sales->189|>,<|Date->Tue 25 Mar
2014,Sales->199|>,<|Date->Wed 26 Mar 2014,Sales->195|>,<|Date->Thu 27 Mar
2014,Sales->195|>,<|Date->Fri 28 Mar 2014,Sales->197|>,<|Date->Sat 29 Mar
2014,Sales->189|>,<|Date->Sun 30 Mar 2014,Sales->198|>,<|Date->Mon 31 Mar
2014,Sales->193|>,<|Date->Tue 1 Apr 2014,Sales->193|>,<|Date->Wed 2 Apr
2014,Sales->183|>,<|Date->Thu 3 Apr 2014,Sales->184|>,<|Date->Fri 4 Apr
2014,Sales->194|>,<|Date->Sat 5 Apr 2014,Sales->194|>,<|Date->Sun 6 Apr
2014,Sales->193|>,<|Date->Mon 7 Apr 2014,Sales->193|>,<|Date->Tue 8 Apr
2014,Sales->194|>,<|Date->Wed 9 Apr 2014,Sales->195|>,<|Date->Thu 10 Apr
2014,Sales->191|>}"
[2] "New York City->{<|Date->Wed 1 Jan 2014,Sales->220|>,<|Date->Thu 2 Jan
2014,Sales->232|>,<|Date->Fri 3 Jan…
```

```
# Question 1c

# finding the index for data and sales
dateindex = gregexpr("[a-zA-Z]+ [0-9]+ [a-zA-Z]+ [0-9]+,Sales->[0-9]+", text)

# extracting the data using the index
datedata = regmatches(text, dateindex)

# remove 'Sales->' from the text file
data = lapply(data,function(datedata)(gsub("Sales->","",datedata)))

# separate the string by ,
unlistedextractedData = unlist(strsplit(data[[1]], ","))
extractedData = list(unlistedextractedData)

head(extractedData[[1]])

tail(extractedData[[1]])
```

```
> head(extractedData[[1]])
[1] "Wed 1 Jan 2014" "198"           "Thu 2 Jan 2014" "189"           "Fri 3 Jan 2014" "196"
> tail(extractedData[[1]])
[1] "Tue 8 Apr 2014"  "228"            "Wed 9 Apr 2014"  "228"            "Thu 10 Apr 2014"
[6] "231"
```

```r
# Question 1d

salesMatrix = matrix(unlistedextractedData, ncol = 2, byrow = TRUE)

# finding the comma index to count how many datas belong to each country
commaIndex = unlist(gregexpr(",", text))

# checking how many data belongs to each country
# all the cities have exactly the same data length, 200
cityDataLen = length(commaIndex[commaIndex < match[2]])
length(commaIndex[(match[2] < commaIndex) & (commaIndex < match[3])])
```

```
## [1] 200
```

```r
length(commaIndex[(match[3] < commaIndex) & (commaIndex < match[4])])
```

```
## [1] 200
```

```r
length(commaIndex[(match[4] < commaIndex) & (commaIndex < match[5])])
```

```
## [1] 200
```

```r
length(commaIndex[(match[5] < commaIndex) & (commaIndex < match[6])])
```

```
## [1] 200
```

```r
cityNames = c("Boston", "New York City", "Paris", "London", "Shanghai", "Toky
o")

# cityDataLen is divided by 2 as the date and the sales are the in the same r
ow
salesData = list(cbind(rep(cityNames, each = cityDataLen/2), salesMatrix))
```

```
## Warning in cbind(rep(cityNames, each = cityDataLen/2), salesMatrix): numbe
r
## of rows of result is not a multiple of vector length (arg 1)
```

```r
head(salesData[[1]])
```

```
> head(salesData[[1]])
     [,1]     [,2]              [,3]
[1,] "Boston" "Wed 1 Jan 2014" "198"
[2,] "Boston" "Thu 2 Jan 2014" "189"
[3,] "Boston" "Fri 3 Jan 2014" "196"
[4,] "Boston" "Sat 4 Jan 2014" "194"
[5,] "Boston" "Sun 5 Jan 2014" "193"
[6,] "Boston" "Mon 6 Jan 2014" "184"
>
```