L2 prep M0, M1, M3

Learning Spoons

2018년 4월 17일

install.packages("package\_name")  
library(package\_name)

# csv  
dataset <-   
 read.csv("filename.csv", header = TRUE, stringsAsFactors = FALSE)  
 # if the first line is "header" (default)  
dataset <-   
 read.csv("filename.csv", header = FALSE, stringsAsFactors = FALSE)  
 # if no "header" in the data file  
write.csv(dataset, "filename.csv")  
  
# txt  
# read.csv 대신에 read.table  
# write.csv 대신에 write.table  
dataset <- read.table(나머지 문법은 read.csv와 똑같고 sep 옵션 추가)  
# ex) sep = ","  
# sep = " "  
# sep = "-"  
  
# xls, xlsx  
install.packages("readxl") # 최초 사용시에만 설치 필요   
library(readxl)  
dataset <- read\_excel("filename.xlsx")   
 # if first line is "header"  
dataset <- read\_excel("filename.xlsx", col\_names=FALSE)  
 # if no "header" in the data file  
  
# rda - 데이터셋 한 개 저장   
load("filename.rda")  
save(dataset, "filename.rda")  
  
# rdata - 현재 메모리에 있는 모든 변수 저장  
load("2018-04-17.rdata")  
save(dataset, "2018-04-17.rdata")

## Module 1

install.packages("ISLR")  
library(ISLR)  
class(Carseats)  
head(Carseats) # 처음 6개 관찰값  
tail(Carseats,5) # 처음 6개 관찰값  
View(Carseats) # "Viewer"에 보여줌  
dim(Carseats) # 차원 (dimension)  
str(Carseats) # 구조 (structure)  
summary(Carseats) # 기초 통계량 (descriptive stat)

# install.packages("ISLR")  
library(ISLR)  
class(Carseats)

## [1] "data.frame"

head(Carseats) # 처음 6개 관찰값

## Sales CompPrice Income Advertising Population Price ShelveLoc Age  
## 1 9.50 138 73 11 276 120 Bad 42  
## 2 11.22 111 48 16 260 83 Good 65  
## 3 10.06 113 35 10 269 80 Medium 59  
## 4 7.40 117 100 4 466 97 Medium 55  
## 5 4.15 141 64 3 340 128 Bad 38  
## 6 10.81 124 113 13 501 72 Bad 78  
## Education Urban US  
## 1 17 Yes Yes  
## 2 10 Yes Yes  
## 3 12 Yes Yes  
## 4 14 Yes Yes  
## 5 13 Yes No  
## 6 16 No Yes

tail(Carseats,5) # 처음 6개 관찰값

## Sales CompPrice Income Advertising Population Price ShelveLoc Age  
## 396 12.57 138 108 17 203 128 Good 33  
## 397 6.14 139 23 3 37 120 Medium 55  
## 398 7.41 162 26 12 368 159 Medium 40  
## 399 5.94 100 79 7 284 95 Bad 50  
## 400 9.71 134 37 0 27 120 Good 49  
## Education Urban US  
## 396 14 Yes Yes  
## 397 11 No Yes  
## 398 18 Yes Yes  
## 399 12 Yes Yes  
## 400 16 Yes Yes

View(Carseats) # "Viewer"에 보여줌  
dim(Carseats) # 차원 (dimension)

## [1] 400 11

str(Carseats) # 구조 (structure)

## 'data.frame': 400 obs. of 11 variables:  
## $ Sales : num 9.5 11.22 10.06 7.4 4.15 ...  
## $ CompPrice : num 138 111 113 117 141 124 115 136 132 132 ...  
## $ Income : num 73 48 35 100 64 113 105 81 110 113 ...  
## $ Advertising: num 11 16 10 4 3 13 0 15 0 0 ...  
## $ Population : num 276 260 269 466 340 501 45 425 108 131 ...  
## $ Price : num 120 83 80 97 128 72 108 120 124 124 ...  
## $ ShelveLoc : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...  
## $ Age : num 42 65 59 55 38 78 71 67 76 76 ...  
## $ Education : num 17 10 12 14 13 16 15 10 10 17 ...  
## $ Urban : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...  
## $ US : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...

summary(Carseats) # 기초 통계량 (descriptive stat)

## Sales CompPrice Income Advertising   
## Min. : 0.000 Min. : 77 Min. : 21.00 Min. : 0.000   
## 1st Qu.: 5.390 1st Qu.:115 1st Qu.: 42.75 1st Qu.: 0.000   
## Median : 7.490 Median :125 Median : 69.00 Median : 5.000   
## Mean : 7.496 Mean :125 Mean : 68.66 Mean : 6.635   
## 3rd Qu.: 9.320 3rd Qu.:135 3rd Qu.: 91.00 3rd Qu.:12.000   
## Max. :16.270 Max. :175 Max. :120.00 Max. :29.000   
## Population Price ShelveLoc Age   
## Min. : 10.0 Min. : 24.0 Bad : 96 Min. :25.00   
## 1st Qu.:139.0 1st Qu.:100.0 Good : 85 1st Qu.:39.75   
## Median :272.0 Median :117.0 Medium:219 Median :54.50   
## Mean :264.8 Mean :115.8 Mean :53.32   
## 3rd Qu.:398.5 3rd Qu.:131.0 3rd Qu.:66.00   
## Max. :509.0 Max. :191.0 Max. :80.00   
## Education Urban US   
## Min. :10.0 No :118 No :142   
## 1st Qu.:12.0 Yes:282 Yes:258   
## Median :14.0   
## Mean :13.9   
## 3rd Qu.:16.0   
## Max. :18.0

# A data frame with 400 observations on the following 11 variables.  
# `Sales` Unit sales (in thousands) at each location  
# `CompPrice` Price charged by competitor at each location  
# `Income` Community income level (in thousands of dollars)  
# `Advertising` Local advertising budget for company at each location   
# (in thousands of dollars)  
# `Population` Population size in region (in thousands)  
# `Price` Price company charges for car seats at each site  
# `ShelveLoc` A factor with levels Bad, Good and Medium indicating   
# the quality of the shelving location for the car seats at each site  
# `Age` Average age of the local population  
# `Education` Education level at each location  
# `Urban` A factor with levels No and Yes to indicate whether the store is   
# in an urban or rural location  
# `US`` A factor with levels No and Yes to indicate whether the store is in the US or not  
# James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013)   
# An Introduction to Statistical Learning with applications in R,   
# www.StatLearning.com, Springer-Verlag, New York

dplyr

head(Carseats)

## Sales CompPrice Income Advertising Population Price ShelveLoc Age  
## 1 9.50 138 73 11 276 120 Bad 42  
## 2 11.22 111 48 16 260 83 Good 65  
## 3 10.06 113 35 10 269 80 Medium 59  
## 4 7.40 117 100 4 466 97 Medium 55  
## 5 4.15 141 64 3 340 128 Bad 38  
## 6 10.81 124 113 13 501 72 Bad 78  
## Education Urban US  
## 1 17 Yes Yes  
## 2 10 Yes Yes  
## 3 12 Yes Yes  
## 4 14 Yes Yes  
## 5 13 Yes No  
## 6 16 No Yes

# install.packages("dplyr")  
library(dplyr)

Carseats$Revenue <- Carseats$Sales

Basics

# rename (이름 바꾸기)   
Carseats <- rename(Carseats, Sales = Revenue)  
names(Carseats)[names(Carseats)=="Sales"] <- "Revenue"  
  
# filter (관찰값 추출, Row 추출)  
temp <- filter(Carseats, Income > 100)  
temp <- Carseats %>% filter(Income > 100)  
temp <- Carseats[Carseats$Income > 100,]  
  
temp <- filter(Carseats, Age >= 30 & Age < 40)  
temp <- Carseats %>% filter(Age >= 30 & Age < 40)  
temp <- Carseats[((Carseats$Age >= 30) & (Carseats$Age < 40)),]  
  
# select (변수 추출, Column 선택)  
temp <- select(Carseats, Income, Population)  
temp <- Carseats %>% select(Income, Population)  
temp <- Carseats[,c("Income", "Population")]  
  
# arrange (정렬)  
Carseats <- arrange(Carseats, Price)  
Carseats <- Carseats %>% arrange(Price)  
Carseats <- Carseats[order(Carseats$Price),]  
  
Carseats <- arrange(Carseats, desc(Price))  
Carseats <- Carseats %>% arrange(desc(Price))  
Carseats <- Carseats[order(Carseats$Price, decreasing = TRUE),]  
  
# mutate (새로운 변수)  
Carseats <- mutate(Carseats,   
 AdvPerCapita = Advertising/Population,  
 RevPerCapita = Revenue/Population)  
Carseats <- Carseats %>%   
 mutate(AdvPerCapita = Advertising/Population,  
 RevPerCapita = Revenue/Population)  
Carseats$AdvPerCapita <- Carseats$Advertising/Carseats$Population  
Carseats$RevPerCapita <- Carseats$Revenue/Carseats$Population  
  
Carseats <- mutate(Carseats,   
 AgeClass = ifelse(Age>=60, "Silver", "non-Silver"))  
Carseats <- Carseats %>%   
 mutate(AgeClass = ifelse(Age>=60, "Silver", "non-Silver"))  
Carseats$AgeClass <-   
 ifelse(Carseats$Age >= 60, "Silver", "non-Silver")

Advanced

# successive treatments  
focusCity <- Carseats %>%   
 filter(Income > 100) %>%  
 filter(Age >= 30 & Age < 40) %>%  
 mutate(AdvPerCapita = Advertising/Population) %>%  
 select(Revenue, Income, Age, Population, Education, AdvPerCapita) %>%  
 arrange(Revenue)  
print(focusCity)

## Revenue Income Age Population Education AdvPerCapita  
## 1 5.04 114 34 298 16 0.00000000  
## 2 5.32 116 39 170 16 0.00000000  
## 3 6.80 117 38 337 10 0.01483680  
## 4 7.49 119 35 178 13 0.03370787  
## 5 7.67 117 36 400 10 0.02000000  
## 6 8.55 111 36 480 16 0.04791667  
## 7 8.97 107 33 144 13 0.00000000  
## 8 9.03 102 35 123 16 0.10569106  
## 9 9.39 118 32 445 15 0.03146067  
## 10 9.58 104 37 353 17 0.06515581  
## 11 10.36 105 34 428 12 0.04205607  
## 12 10.59 120 30 262 10 0.05725191  
## 13 12.57 108 33 203 14 0.08374384

Carseats %>%  
 mutate(AgeClass =   
 ifelse(Age < 30, "Twenties",   
 ifelse(Age < 40, "Thirties", "FourtyAbove"))) %>%  
 group\_by(AgeClass) %>%  
 summarise(avgRevenue = mean(Sales))

## # A tibble: 3 x 2  
## AgeClass avgRevenue  
## <chr> <dbl>  
## 1 FourtyAbove 7.30  
## 2 Thirties 8.26  
## 3 Twenties 7.76

appendix

# C, C++  
for(int i=1; i <=10, i=i+1) {  
 printf("value of i: %d\n", i);  
}  
  
# MATLAB  
for i=1:10 {  
 disp("value of i: %d", i)  
}  
  
# Python  
for i in range(1, 11):  
 print "value of i: %d" % (i)  
  
## R  
for (i in 1:10) {  
 print(paste("value of i:", i))  
}