LAB1 STAT515

Mykola Signayevskyy

2023-02-03

```
carseats = read.csv('carseats.csv')
head(carseats)
```

a) Read in the "carseats" dataset, look at the first few rows and inspect the data types of the variables in the dataframe.

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

```
str(carseats)
```

```
## 'data.frame':    400 obs. of  11 variables:
##  $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
##  $ CompPrice  : int  138 111 113 117 141 124 115 136 132 132 ...
##  $ Income     : int  73 48 35 100 64 113 105 81 110 113 ...
##  $ Advertising: int  11 16 10 4 3 13 0 15 0 0 ...
##  $ Population : int  276 260 269 466 340 501 45 425 108 131 ...
##  $ Price      : int  120 83 80 97 128 72 108 120 124 124 ...
##  $ ShelveLoc  : chr  "Bad" "Good" "Medium" "Medium" ...
##  $ Age        : int  42 65 59 55 38 78 71 67 76 76 ...
##  $ Education  : int  17 10 12 14 13 16 15 10 10 17 ...
##  $ Urban      : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ US         : chr  "Yes" "Yes" "Yes" "Yes" ...
```

```
carseats$ShelveLoc=as.factor(carseats$ShelveLoc)
carseats$Urban=as.factor(carseats$Urban)
carseats$US=as.factor(carseats$US)
str(carseats) #checking changes
```

b) Change the variables "ShelveLoc, urban, US" into a factor variables.

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

```
carseats$Profit = carseats$Income - carseats$Advertising
head(carseats$Profit) #checking result
```

c) create a new variable called "profit" which stands for "Income - Advertising"

```
## [1]  62  32  25  96  61 100
```

```
table(is.na(carseats))
```

d) Check for missing data. If you have missing data remove the corresponding rows from the dataset.

```
##
## FALSE  TRUE
##  4797     3
```

```
carseats = na.omit(carseats)
table(is.na(carseats)) #checking result
```

```
##
## FALSE
##  4776
```

```
length(which(carseats$ShelveLoc=='Good'))
```

e) How many "Good" shelving locations are there in the dataset?

```
## [1] 85
```

```
length(which(carseats$US=="Yes"))
```

f) How many stores are inside the USA? create a separate data frame containing all stores from USA. #name the data set as "stores_USA"

```
## [1] 256
```

```
stores_USA = carseats[carseats$US=='Yes',]
head(stores_USA)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
```

```
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 6 10.81       124    113          13        501    72       Bad  78        16
## 8 11.85       136     81          15        425   120      Good  67        10
##   Urban  US Profit
## 1   Yes Yes     62
## 2   Yes Yes     32
## 3   Yes Yes     25
## 4   Yes Yes     96
## 6    No Yes    100
## 8   Yes Yes     66
```

```
HighUrban_USSales = subset(stores_USA, Sales > 7 & Urban == "Yes")
head(HighUrban_USSales)
```

g) create another data set called "HighUrban_USSales" using 'stores_USA' data set. #where, sales are greater than 7 thousand and stores are located in Urban areas.

```
##    Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1   9.50       138     73          11        276   120       Bad  42        17
## 2  11.22       111     48          16        260    83      Good  65        10
## 3  10.06       113     35          10        269    80    Medium  59        12
## 4   7.40       117    100           4        466    97    Medium  55        14
## 8  11.85       136     81          15        425   120      Good  67        10
## 12 11.96       117     94           4        503    94      Good  50        13
##    Urban  US Profit
## 1    Yes Yes     62
## 2    Yes Yes     32
## 3    Yes Yes     25
## 4    Yes Yes     96
## 8    Yes Yes     66
## 12   Yes Yes     90
```

```
HighUrban_USSales = subset(HighUrban_USSales, select = -c(Urban , US))
head(HighUrban_USSales)
```

h) Remove "US" and "Urban" columns from the "HighUrban_USSales" dataset.

```
##    Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1   9.50       138     73          11        276   120       Bad  42        17
## 2  11.22       111     48          16        260    83      Good  65        10
## 3  10.06       113     35          10        269    80    Medium  59        12
## 4   7.40       117    100           4        466    97    Medium  55        14
## 8  11.85       136     81          15        425   120      Good  67        10
## 12 11.96       117     94           4        503    94      Good  50        13
##    Profit
## 1      62
## 2      32
## 3      25
## 4      96
## 8      66
## 12     90
```

```
write.csv(HighUrban_USSales, "HighUrban_USSales.csv")
```

i) For one the above subset, write to a new CSV file