Lab-assignment01 | STAT 515 | 002

saivarun tanjore raghavendra (G number: G01475545)

2024-01-31

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

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

```
data = read.csv("/Users/trsaivarun/Desktop/R programs/lab assignments/carseats(1).csv")
```

```
head(data)
```

```
##   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(data)
```

```
## '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" ...
```

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

```
data$ShelveLoc = factor(data$ShelveLoc)
str(data)
```

```
## '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      : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ US         : chr  "Yes" "Yes" "Yes" "Yes" ...
```

```
data$US = factor(data$US)
str(data)
```

```
## '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      : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ US         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

```
data$Urban = factor(data$Urban)
str(data)
```

```
## '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 ...
```

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

```
data$profit = data$Income - data$Advertising
data$profit
```

```
##   [1]  62  32  25  96  61 100 105  66 110 113  69  90  33  17 106  90  32  61
##  [19] 110  60  88  17  40  31 103  32 104 118  74  84  94  42  20  25  54  73
##  [37]  76  36  73  60  98  53  69  31  73  63  76  98  52  93  14  90  37  51
##  [55]  90  76  82  91  78  67  83  32  45  78  55  26  92  47  49  59  66  35
##  [73]  45  80  63  88  77  59  47  67  84  72  79  29  25 103  75  60  35  63
##  [91]  22  35 113  30  92  15  32  77  53  44  58  93  22  91  96  92  33 107
## [109]  77  65  55 106  94  18  78  35  75  53  86  86  94  79  95 103 113  78
## [127]  66  45  97 113  71  66  78  96  31  80  75  42  91  52  50  42  84  81
## [145]  68  52  83  45 119 107  76  41  78  29  59  72  34  50  89  60  28  16
## [163]  74  64  64  51  50  73  89  26  27  94  89  86  24  89  98  72  57  22
## [181]  97  83  56  68  26  89  51  32  37  99  24  29  26  63  80  89  22  61
## [199]  75  83  92  83  74  82  80  21  67 105  54  10  39 104  50  79 112  68
## [217]  33  44  49  60 105  44 113  36  82  25  33  54  60 104  60  69  70  58
## [235]  51  24  18  20  24 105  80  63  46  12  30  43  36 114  52  67  95 106
## [253]  97  19  81  73  40  48  38  26 109  38  62  20  24  25  81  75  57  69
## [271]  26  56  33  98  91 108  55  36 111  44  76  62  96 110  35  15 107  40
## [289]  40  52  97  70  50  84  73  21  31  70  63  23  77  93  64  36  86   3
## [307]  31  92  61  98  36  56 112  78  23  13  31  30  62  26  58  34  40  87
## [325]  61  58  30  21  65  45  59  48  13  53 108  55  29  38  24  40  29 120
## [343]  89  32  80  68 107  39  82   9  84  99  89  55  30 100 109  70  86  51
## [361]  79  15  55  74   5  30  45 106  12  78  19  81  50  71  40  42  41  61
## [379]  85 111  NA  44   9 117  22  60 116  59  78  34  66  63  29  41  39  91
## [397]  20  14  72  37
```

```
head(data)
```

```
##   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 profit
## 1   Yes Yes     62
## 2   Yes Yes     32
## 3   Yes Yes     25
## 4   Yes Yes     96
## 5   Yes  No     61
## 6    No Yes    100
```

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

#Here removing missing values

```
table(is.na(data))
```

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

```
data=na.omit(data) #deleted the null values here
head(data)
```

```
##   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 profit
## 1   Yes Yes     62
## 2   Yes Yes     32
## 3   Yes Yes     25
## 4   Yes Yes     96
## 5   Yes  No     61
## 6    No Yes    100
```

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

```
data1=subset(data,data$ShelveLoc == "Good")
head(data1)
```

```
##    Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 2  11.22       111     48          16        260    83      Good  65        10
## 8  11.85       136     81          15        425   120      Good  67        10
## 12 11.96       117     94           4        503    94      Good  50        13
## 14 10.96       115     28          11         29    86      Good  53        18
## 15 11.17       107    117          11        148   118      Good  52        18
## 17  7.58       118     32           0        284   110      Good  63        13
##    Urban  US profit
## 2    Yes Yes     32
## 8    Yes Yes     66
## 12   Yes Yes     90
## 14   Yes Yes     17
## 15   Yes Yes    106
## 17   Yes  No     32
```

```
nrow(data1)
```

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

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"

```
stores_USA = subset(data, data$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
```

```
nrow(stores_USA)
```

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

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.

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

```
##    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
```

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

```
HighUrban_USSales2 = HighUrban_USSales[,-c(10,11)]
head(HighUrban_USSales2)
```

```
##    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
```

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

```
write.csv(data,'/Users/trsaivarun/Desktop/R programs/lab assignments/carseats_pure.csv', row.names = F)
```

Q2) See the following code of a function and explain what it does. Suggest a suitable name for the function and rename. Demonstrate how the function works when you have numerical data and character data. function1 <- function(x) { if (length(x) <= 1) return(NULL) x[-length(x)] }

A) The following function takes "x" as parameter and then it checks for its length, and if the length is less than or equal to 1 then it returns NULL value. Or else, it removes that last character from the variable and returns the remaining part of it.

```
cutter <- function(x) {
if (length(x) <= 1) return(NULL)
x[-length(x)]
}

digits = c(1,2,3,4,5)
letters = c("Benz","Toyota","BMW")

res1 = cutter(digits)
cat(res1)
```

```
## 1 2 3 4
```

```
res2 = cutter(letters)
cat(res2)
```

```
## Benz Toyota
```

Q3) Write a function to compute the sample variance of a numerical vector. Use the equation of the variance to write the function.

```
sample_v <- function(var) {

  square = sum((var-mean(var))^2)
  s_varience = square/(length(var)-1)
  return (s_varience)

}


data = c(6,6,6,7,8,9,2,2)

res = sample_v(data)
cat(res)
```

```
## 6.5
```