Análise de Dados - Aula 1

Ítalo e José Antonio

2024-10-05

Análise de Dados - Aula 1

Questão 1 - ok

Descrição:
O domínio do problema é a Ciência da Computação, focando no desempenho de hardware de computadores. O conhecimento esperado envolve dados de desempenho relativo da CPU, descritos em termos de tempo de ciclo, tamanho de memória, etc.

```
# C:/Users/2019101100910126/Documents/Github/EstudoDirigindo-MinecaoDeDados/Computer Hardware/computer+
computer_data <- read.table("C:/Users/josej/OneDrive/Documentos/GitHub/EstudoDirigindo-MinecaoDeDados/C
colnames(computer_data) <- c("VendorName", "ModelName", "MYCT", "MMIN", "MMAX", "CACH", "CHMIN", "CHMAX
View(computer_data)
```

Questão 2 - ok

Descrição:
Aqui, utilizamos as funções `dim()`, `nrow()` e `ncol()` temos o numero de amostra de 209 e o numero de variaveis de 10

```
print(paste('Número de linhas(amostras):', nrow(computer_data)))
```

```
## [1] "Número de linhas(amostras): 209"
```

```
print(paste('Número de colunas(variaveis):', ncol(computer_data)))
```

```
## [1] "Número de colunas(variaveis): 10"
```

Questão 3 - ok

Descrição:
Utilizamos a função `str()` para exibir a estrutura do conjunto de dados.

```
# Exibir a estrutura do dataset
str(computer_data)
```

```
## 'data.frame':    209 obs. of  10 variables:
##  $ VendorName: chr  "adviser" "amdahl" "amdahl" "amdahl" ...
##  $ ModelName : chr  "32/60" "470v/7" "470v/7a" "470v/7b" ...
##  $ MYCT      : int  125 29 29 29 29 26 23 23 23 23 ...
##  $ MMIN      : int  256 8000 8000 8000 8000 8000 16000 16000 16000 32000 ...
##  $ MMAX      : int  6000 32000 32000 32000 16000 32000 32000 32000 64000 64000 ...
##  $ CACH      : int  256 32 32 32 32 64 64 64 64 128 ...
##  $ CHMIN     : int  16 8 8 8 8 8 16 16 16 32 ...
##  $ CHMAX     : int  128 32 32 32 16 32 32 32 32 64 ...
##  $ PRP       : int  198 269 220 172 132 318 367 489 636 1144 ...
##  $ ERP       : int  199 253 253 253 132 290 381 381 749 1238 ...
```

A maioria dos campos vieram com seus datatype certos, menos os dois primeiros, que vieram como chr entao vou fazer a tranformação dele para factor(Variável categórica)

```
# Transformando colunas de character para factor
computer_data$VendorName <- as.factor(computer_data$VendorName)
computer_data$ModelName <- as.factor(computer_data$ModelName)

# Exibir a estrutura do dataset com as alterações feitas
str(computer_data)
```

```
## 'data.frame':    209 obs. of  10 variables:
##  $ VendorName: Factor w/ 30 levels "adviser","amdahl",..: 1 2 2 2 2 2 2 2 2 2 ...
##  $ ModelName : Factor w/ 209 levels "100","1100/61-h1",..: 30 63 64 65 66 67 75 76 77 78 ...
##  $ MYCT      : int  125 29 29 29 29 26 23 23 23 23 ...
##  $ MMIN      : int  256 8000 8000 8000 8000 8000 16000 16000 16000 32000 ...
##  $ MMAX      : int  6000 32000 32000 32000 16000 32000 32000 32000 64000 64000 ...
##  $ CACH      : int  256 32 32 32 32 64 64 64 64 128 ...
##  $ CHMIN     : int  16 8 8 8 8 8 16 16 16 32 ...
##  $ CHMAX     : int  128 32 32 32 16 32 32 32 32 64 ...
##  $ PRP       : int  198 269 220 172 132 318 367 489 636 1144 ...
##  $ ERP       : int  199 253 253 253 132 290 381 381 749 1238 ...
```

Questão 4 - ok

Descrição:
Usado a função `summary()` temos resumo estatístico do dataframe, mas para garantir que nao exista valore NA, uma funcao de soma(sum) que vai contar a quantidade de valores NA no dataframe.

```
# Resumo do dataset
summary(computer_data)
```

```
##      VendorName       ModelName        MYCT             MMIN
##  ibm      : 32   100       :  1   Min.   :  17.0   Min.   :   64
##  nas      : 19   1100/61-h1:  1   1st Qu.:  50.0   1st Qu.:  768
##  honeywell: 13   1100/81   :  1   Median : 110.0   Median : 2000
##  ncr      : 13   1100/82   :  1   Mean   : 203.8   Mean   : 2868
##  sperry   : 13   1100/83   :  1   3rd Qu.: 225.0   3rd Qu.: 4000
##  siemens  : 12   1100/84   :  1   Max.   :1500.0   Max.   :32000
##  (Other)  :107   (Other)   :203
##       MMAX            CACH            CHMIN            CHMAX
```

```
##  Min.   :   64   Min.   :  0.00   Min.   : 0.000   Min.   :  0.00
##  1st Qu.: 4000   1st Qu.:  0.00   1st Qu.: 1.000   1st Qu.:  5.00
##  Median : 8000   Median :  8.00   Median : 2.000   Median :  8.00
##  Mean   :11796   Mean   : 25.21   Mean   : 4.699   Mean   : 18.27
##  3rd Qu.:16000   3rd Qu.: 32.00   3rd Qu.: 6.000   3rd Qu.: 24.00
##  Max.   :64000   Max.   :256.00   Max.   :52.000   Max.   :176.00
##
##       PRP              ERP
##  Min.   :   6.0   Min.   :  15.00
##  1st Qu.:  27.0   1st Qu.:  28.00
##  Median :  50.0   Median :  45.00
##  Mean   : 105.6   Mean   :  99.33
##  3rd Qu.: 113.0   3rd Qu.: 101.00
##  Max.   :1150.0   Max.   :1238.00
##
```

```r
# Contar o número total de NAs
quant_na <- sum(is.na(computer_data))

# Exibir o total de NAs
print(paste("Quantidade de dados ausentes(NA):", quant_na))
```

```
## [1] "Quantidade de dados ausentes(NA): 0"
```

Questão 5 - Parei aqui, comentar sobre o segundo caso abaixo!!!

Descrição:
Aqui, fiz uma contagem de amostra para cada classe, primeiro de VendorName e depos de ModelName

VendorName Nessa coluna/variavel foi possivel fazer a contagem de amostras sem grandes questoes e foi possivel demostrar as ammostras mais repetivas e seus valores.

```r
# Contar amostras em cada classe
count_VendorName <- table(computer_data$VendorName)

# Criar um data frame organizado
df_porcentagem_VendorName <- data.frame(
  porcentagem = round(((count_VendorName / sum(count_VendorName)) * 100), 2)
)

kable(df_porcentagem_VendorName, caption = "Porcentagem da Classe: Vendor_Name")
```

Table 1: Porcentagem da Classe: Vendor_Name

porcentagem.Var1	porcentagem.Freq
adviser	0.48
amdahl	4.31
apollo	0.96
basf	0.96
bti	0.96

porcentagem.Var1	porcentagem.Freq
burroughs	3.83
c.r.d	2.87
cambex	2.39
cdc	4.31
dec	2.87
dg	3.35
formation	2.39
four-phase	0.48
gould	1.44
harris	3.35
honeywell	6.22
hp	3.35
ibm	15.31
ipl	2.87
magnuson	2.87
microdata	0.48
nas	9.09
ncr	6.22
nixdorf	1.44
perkin-elmer	1.44
prime	2.39
siemens	5.74
sperry	6.22
sratus	0.48
wang	0.96

```r
# Contar amostras em cada classe
count_ModelName <- table(computer_data$ModelName)

# Criar um data frame organizado
df_porcentagem_ModelName <- data.frame(
  porcentagem = round(((count_ModelName / sum(count_ModelName)) * 100), 2)
)

kable(head(df_porcentagem_ModelName, 10), caption = "Porcentagem da Classe: Model_Name")
```

Table 2: Porcentagem da Classe: Model_Name

porcentagem.Var1	porcentagem.Freq
100	0.48
1100/61-h1	0.48
1100/81	0.48
1100/82	0.48
1100/83	0.48
1100/84	0.48
1100/93	0.48
1100/94	0.48
1636-1	0.48
1636-10	0.48

Questão 6

Descrição:
(Inclua a descrição e o código para a Questão 6)

Questão 7

Descrição:
(Inclua a descrição e o código para a Questão 7)

Questão 8

Descrição:
Calculamos as médias, os valores mínimos e máximos das variáveis `Sepal.Length` e `Petal.Length` para as espécies `setosa`, `versicolor` e `virginica`.

```
# Setosa
mean(iris$Sepal.Length[1:50])
```

```
## [1] 5.006
```

```
mean(iris$Petal.Length[1:50])
```

```
## [1] 1.462
```

```
min(iris$Sepal.Length[1:50])
```

```
## [1] 4.3
```

```
min(iris$Petal.Length[1:50])
```

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

```
max(iris$Sepal.Length[1:50])
```

```
## [1] 5.8
```

```
max(iris$Petal.Length[1:50])
```

```
## [1] 1.9
```

```
# Versicolor
mean(iris$Sepal.Length[51:100])
```

```
## [1] 5.936
```

```r
mean(iris$Petal.Length[51:100])
```

```
## [1] 4.26
```

```r
# Virginica
mean(iris$Sepal.Length[101:150])
```

```
## [1] 6.588
```

```r
mean(iris$Petal.Length[101:150])
```

```
## [1] 5.552
```

Questão 9

Descrição:
(Inclua a descrição e o código para a Questão 9)

Questão 10

Descrição:
(Inclua a descrição e o código para a Questão 10)

Questão 11

Descrição:
(Inclua a descrição e o código para a Questão 11)

Questão 12

Descrição:
(Inclua a descrição e o código para a Questão 12)

Questão 13

Descrição:
(Inclua a descrição e o código para a Questão 13)

Questão 14

Descrição:
(Inclua a descrição e o código para a Questão 14)

Questão 15

Descrição:
(Inclua a descrição e o código para a Questão 15)