Aquisição de Arquivos e Datasets

Editando Variáveis Texto

Delermando Branquinho Filho

Fixação de vetores de caracteres - tolower(), toupper()

```
cameraData <- read.csv("./data/cameras.csv")
names(cameraData)
```

```
## [1] "address"      "direction"    "street"       "crossStreet"
## [5] "intersection" "Location.1"
```

```
tolower(names(cameraData))
```

```
## [1] "address"      "direction"    "street"       "crossstreet"
## [5] "intersection" "location.1"
```

Fixação de vetores de caracteres - strsplit()

- Good for automatically splitting variable names
- Important parameters: *x*, *split*

```
splitNames = strsplit(names(cameraData),"\\.")
splitNames[[5]]
```

```
## [1] "intersection"
```

```
splitNames[[6]]
```

```
## [1] "Location" "1"
```

Rapidamente à parte - listas

```
mylist <- list(letters = c("A", "b", "c"), numbers = 1:3, matrix(1:25, ncol = 5))
head(mylist)
```

```
## $letters
## [1] "A" "b" "c"
##
## $numbers
## [1] 1 2 3
##
## [[3]]
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    6   11   16   21
## [2,]    2    7   12   17   22
## [3,]    3    8   13   18   23
## [4,]    4    9   14   19   24
```

```
## [5,]    5   10   15   20   25
```

```
mylist[1]
```

```
## $letters
## [1] "A" "b" "c"
```

```
mylist$letters
```

```
## [1] "A" "b" "c"
```

```
mylist[[1]]
```

```
## [1] "A" "b" "c"
```

Fixação de vetores de caracteres - sapply()

- Aplica uma função a cada elemento em um vetor ou lista
- Parâmetros importantes: X , _ FUN_

```
splitNames[[6]][1]
```

```
## [1] "Location"
```

```
firstElement <- function(x){x[1]}
sapply(splitNames,firstElement)
```

```
##  [1] "address"      "direction"    "street"       "crossStreet"
##  [5] "intersection" "Location"
```

Experimento para os alunos

```
reviews <- read.csv("./data/reviews.csv"); solutions <- read.csv("./data/solutions.csv")
head(reviews,2)
```

```
##   id solution_id reviewer_id      start       stop time_left accept
## 1  1           3          27 1304095698 1304095758      1754      1
## 2  2           4          22 1304095188 1304095206      2306      1
```

```
head(solutions,2)
```

```
##   id problem_id subject_id      start       stop time_left answer
## 1  1        156         29 1304095119 1304095169      2343      B
## 2  2        269         25 1304095119 1304095183      2329      C
```

Fixação de vetores de caracteres - sub()

- Important parameters: *pattern*, *replacement*, *x*

```
names(reviews)
```

```
## [1] "id"          "solution_id" "reviewer_id" "start"       "stop"
## [6] "time_left"   "accept"
```

```
sub("_","",names(reviews),)
```

```
## [1] "id"         "solutionid" "reviewerid" "start"      "stop"
## [6] "timeleft"   "accept"
```

Fixação de vetores de caracteres - gsub()

```
testName <- "this_is_a_test"
sub("_","",testName)
```

```
## [1] "thisis_a_test"
```

```
gsub("_","",testName)
```

```
## [1] "thisisatest"
```

Procurando valores - grep(),grepl()

```
grep("Alameda",cameraData$intersection)
```

```
## [1]  4  5 36
```

```
table(grepl("Alameda",cameraData$intersection))
```

```
##
## FALSE  TRUE
##    77     3
```

```
cameraData2 <- cameraData[!grepl("Alameda",cameraData$intersection),]
```

Mais de on grep()

```
grep("Alameda",cameraData$intersection,value=TRUE)
```

```
## [1] "The Alameda  & 33rd St"   "E 33rd  & The Alameda"
## [3] "Harford \n & The Alameda"
```

```
grep("JeffStreet",cameraData$intersection)
```

```
## integer(0)
```

```
length(grep("JeffStreet",cameraData$intersection))
```

```
## [1] 0
```

Funções de seqüência mais úteis

```
library(stringr)
nchar("Delermando Branquinho Filho")
```

```
## [1] 27
```

```
substr("Delermando Branquinho Filho",1,7)
```

```
## [1] "Delerma"
```

```
paste("Delermando","Branquinho","Filho")
```

```
## [1] "Delermando Branquinho Filho"
```

```
paste0("Delermando","Branquinho","Filho")
```

```
## [1] "DelermandoBranquinhoFilho"
```

```
str_trim("Delermando      ")
```

```
## [1] "Delermando"
```

Pontos importantes sobre o texto em conjuntos de dados

- Os nomes das variáveis devem ser
- Minúsculas sempre que possível
- Descritivo (Diagnóstico versus Dx)
- Não duplicado
- Não tem sublinhados ou pontos ou espaços em branco
- Variáveis com valores de caractere
- Deve ser feito geralmente em variáveis fatoras (depende da aplicação)
- Deve ser descritivo (use TRUE / FALSE em vez de 0/1 e Masculino / Feminino versus 0/1 ou M / F)