Aula 3 - Manipular dados no R

Marcelo Prudente

9 de janeiro de 2020

Sumário

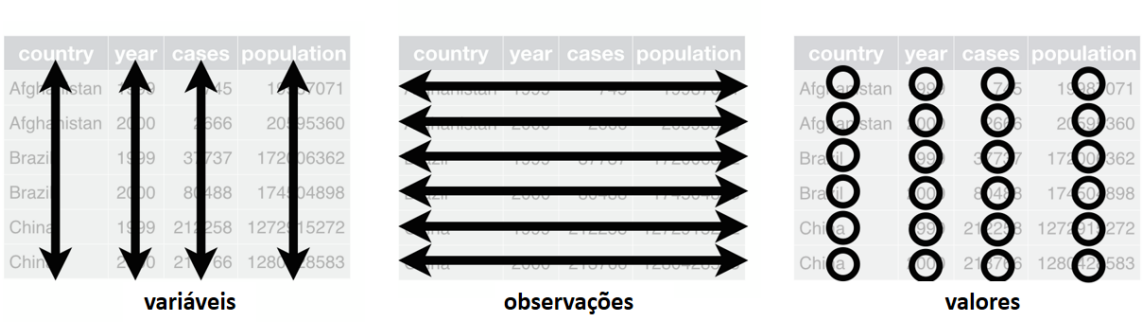
1 Tidy data

O tidyverse baseia-se na ideia de que os dados devem estar ordenados ou arrumados (tidy) para a manpulação (ver capítulo 12 de R for Data Science).

O conceito se baseia em três regras:

1. Cada variável deve ter a sua própria coluna.
2. Cada observação tem sua própria linha.
3. Cada valor tem sua própria célula

Essas regras estão interrelacionadas e devem ser plenamente satisfeitas para que tenhamos uma base de dados arrumada (tidy data).



Fonte: R for data Science. Cap. 12

Figura 1: Tidy Data

2 Manipulação de dados com *dplyr*

A manipulação de dados demanda um bom tempo de qualquer analista de dados. Remover colunas, criar colunas, fundir tabelas, renomear variáveis, sumarizar variáveis, entre outros, são tarefas comuns bastante facilitadas pelo programa.

No **R** há diversas formas de manipular dados por meio de diversos pacotes. Podemos elencar, por exemplo, o pacote **base**, já instalado no sistema, o pacote **dplyr** do **tidyverse** ou o pacote **data.table**. Na prática, todos tem o mesmo objetivo, mas a sintaxe a performance são distintas.

Para esse curso, daremos grande ênfase à abordagem do **dplyr**, cuja linguagem é próxima a do SQL, por sua facilidade, disseminação e performance. Se o nosso objetivo fosse trabalhar intensivamente com bases grandes, daríamos ênfase ao **data.table**. Caso você queira entender melhor essa forma de manipular dados recomendo acessar **este link**.

No entanto, o objetivo deste tópico também é apresentar algumas estratégias importantes para a manipulação dos dados com o pacote **base** - que já vem instalado como padrão no R. Afinal, é muito comum observar nos fóruns o uso da sintaxe desse pacote e, assim, é necessário conhecê-la um pouco. Ainda, essa pequena introdução à manipulação dos dados tem por objetivo explicitar a lógica por trás da linguagem do **R**.

Portanto, os comandos aprendidos nesta aula trazem “massa crítica” para as aulas seguintes e para os exercícios.

Iniciaremos nosso estudo com banco de dados bem comportado que está na nossa pasta do Github. Baixe o arquivo **dados_sociais.csv**. Com ele, vamos acessar questões recorrentes na análise de dados.

Qual o primeiro passo? Ler os dados!

```
# limpar ambiente global
rm(list = ls())
# carregar pacotes
library(data.table); library(readr)
# com fread
dados_sociais <- fread("C:/curso_r/dados/dados_sociais.csv", dec = ",")

# com read_csv2 - sep = ";" e dec = ","
dados_sociais <- read_csv2("C:/curso_r/dados/dados_sociais.csv",
                           locale = locale(encoding = "latin1"))
```

2.1 Como estão os meus dados

Com os dados carregados, devemos gastar nossa energia em conhecer os dados para promover as mudanças necessárias (desejadas).

O primeiro passo é olhar a nossa base de dados sociais e verificar se as estruturas de dados foram corretamente importadas. Com isso, queremos dizer que dados de texto são importados como texto (*character*) e dados numéricos como números (*integer, numeric, double*).

```
library(dplyr)
glimpse(dados_sociais)
```

```
## Observations: 16,695
## Variables: 8
## $ ano          <dbl> 1991, 1991, 1991, 1991, 1991, 1991, 1991, 1991, 1991, ..
## $ uf           <dbl> 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11, 11..
## $ cod_ibge     <dbl> 1100015, 1100023, 1100031, 1100049, 1100056, 1100064, ..
## $ municipio    <chr> "ALTA FLORESTA D'OESTE", "ARIQUEMES", "CABIXI", "CACOA..
## $ esp_vida     <dbl> 62.01, 66.02, 63.16, 65.03, 62.73, 64.46, 59.32, 62.76..
## $ tx_analf_15m <dbl> 23.55, 17.18, 24.57, 21.41, 20.26, 25.44, 30.49, 19.24..
## $ pop          <dbl> 23417, 56061, 7601, 69173, 19451, 25441, 11968, 7418, ..
## $ rdpc         <dbl> 198.46, 319.47, 116.38, 320.24, 240.10, 224.82, 81.38,..
```

Lembre-se que a manipulação permite melhorar a precisão dos dados analisados. Com o **R** é possível fazer isso de forma rápida e transparente, sobretudo em comparação ao uso de planilhas do Excel.

2.2 Manipulação de dados com o R

Entre os 15176 pacotes do **R** no CRAN, alguns foram especificamente desenhados para manipular dados e explorar os dados.

Para as aulas, utilizaremos os pacotes do *tidyverse*. O *tidyverse* é uma coleção de pacotes desenhados para ciência de dados (**ver informações neste link**):

- *dplyr* - manipulação dos dados
- *lubridate* - manipulação de datas
- *tidyr* - pivotar dados
- *ggplot2* - gráficos
- *purrr* - programação funcional
- *readr* - leitura de dados
- *tibble* - nova versão do data.frame
- *forcats* - manipulação de *factors*
- *stringr* - manipulação de *strings*

Embora nosso foco seja analisar os dados com o tidyvers, fazemos menção honrosa ao pacote *data.table()*. Atualmente, Hadley Wickham trabalha no pacote dtplyr (*neste link*) cuja função é traduzir as expressões do dplyr para o data.table. Com certeza, a manipulação de bases de dados grandes (maiores que 1 Gb, por exemplo) ganha muito mais eficiência com as soluções propostas por *Matt Dowle*.

3 Manipulação de dados com dplyr

O pacote *dplyr* é hoje um dos mais utilizados para a manipulação de dados por algumas razões:

- sintaxe amigável e simples (próxima à do SQL)
- velocidade em comparação ao pacote de base
- possibilidade de encadear funções - algo muito útil!
- diversos tutoriais disponíveis (livros, artigos no Rpubs, discussões no stackoverflow)

Recomendo fortemente ler o *R for data Science* onde os tópicos de manipulação de dados são explicados com maestria.

Comandos básicos do dplyr

Os seis mais importantes comandos (ou verbos) de manipulação de dados do *dplyr* são:

1. filter() - seleciona linhas
2. select() - seleciona colunas
3. arrange() - ordena os dados
4. mutate() - cria novas variáveis e renomeia
5. group_by() - agrupa os dados
6. summarise()- sumariza os dados

Lembre desses comandos! Eles serão seus grandes amigos!!

3.1 filter (): filtrando as linhas

Um dos problemas mais básicos da manipulação de dados é a necessidade de extrair subconjuntos dessas informações. Por exemplo, em dados segregados por Estado, você pode se interessar por analisar apenas essa unidade da federação específica, uma região (Nordeste ou Sul) ou um ano. Mas como implementar isso? Com o *dplyr*, essa é uma tarefa intuitiva, conforme o comando a seguir. Observe que para alcançarmos nosso objetivo utilizamos o operador lógico da igualdade.

```
# filtrar apenas informações do ano de 2000
dados_2010 <-  filter(dados_sociais, ano == 2000)
```

A aborgagem do dplyr é simples, mas a sua contraparte da base do sistema também:

```
dados_2010 <- subset(dados_sociais, ano == 2010)
```

São soluções muito parecidas e fáceis, não?

Exercício

Tente filtrar as linhas dos dados de tal forma que:

- o ano seja 2010 - a taxa de analfabetismo seja maior que a média - o Estado seja do Nordeste

Dica: utilize os comandos lógicos e de medida aprendidos no capítulo 1.

3.1.1 Filtrar variáveis de texto

É possível filtrar bancos de dados por meio de variáveis que contenham observações do tipo de textos. Esse tipo de filtragem é comum quando buscamos subconjunto de dados com algum padrão. Se quisermos filtrar as cidades com nomes que contenham "ara", utilizamos o código a seguir. Nesse caso, utilizamos a função *grepl()*. Para maior informação, lembre-se sempre de olhar a documentação do comando ?grepl.

```
# sensível a maiúsculas e minúsculas
filter(dados_sociais, grepl("Ara", municipio))
```

```
## # A tibble: 30 x 8
##      ano    uf cod_ibge municipio esp_vida tx_analf_15m   pop  rdpc
##    <dbl> <dbl>    <dbl> <chr>        <dbl>        <dbl> <dbl> <dbl>
##  1  1991    31  3103207 Araçaí        65.8         14.4  1947  310.
```

```
##  2  1991    31  3103306 Aracitaba     64.9         25.4  2386  215.
##  3  1991    31  3103405 Araçuaí       64.1         34.6 33357  169.
##  4  1991    31  3103504 Araguari      71.0         11.1 90294  434.
##  5  1991    31  3103603 Arantina      66.3         24.4  2638  181.
##  6  1991    31  3103702 Araponga      60.2         42.4  7847  125.
##  7  1991    31  3103751 Araporã       69.1         18.1  4288  281.
##  8  1991    31  3103801 Arapuá        69.6         18.3  3093  244.
##  9  1991    31  3103900 Araújos       67.6         17.7  5500  258.
## 10  1991    31  3104007 Araxá         67.4         11.0 69217  439.
## # ... with 20 more rows
```

```
# ignoramos maiúsculas e minúsculas
filter(dados_sociais, grepl("Ara", municipio, ignore.case = TRUE ))
```

```
## # A tibble: 1,377 x 8
##      ano    uf cod_ibge municipio       esp_vida tx_analf_15m   pop  rdpc
##    <dbl> <dbl>    <dbl> <chr>              <dbl>        <dbl> <dbl> <dbl>
##  1  1991    11  1100072 CORUMBIARA          59.3         30.5 11968  81.4
##  2  1991    11  1100122 JI-PARANÁ           64.5         18.1 99247 312.
##  3  1991    11  1100403 ALTO PARAÍSO        59.7         22.1  9261 135.
##  4  1991    11  1101807 VALE DO PARAÍSO     63.8         27.0  8801 125.
##  5  1991    12  1200609 TARAUACÁ            61.7         61.9 24741 131.
##  6  1991    13  1301001 CARAUARI            60           60.9 19073 145.
##  7  1991    13  1301902 ITACOATIARA         65.0         19.8 58040 219.
##  8  1991    13  1301951 ITAMARATI           61.6         82.7  9085 166.
##  9  1991    13  1302801 MARAÃ               61.0         63.1 11766 127.
## 10  1991    14  1400209 CARACARAÍ           63.7         32.8  6810 310.
## # ... with 1,367 more rows
```

A manipulação de *strings* é um capítulo a parte na análise de dados. Há uma série de expressões regulares na ciência da computação e *este livro, Handling Strings with R* pode ser muito útil quando temos dúvidas nesse tema.

Vamos a um exemplo. Para filtrar todas as cidades com nomes iniciados em "São José de" utilizamos o *regex* ^ .

```
filter(dados_sociais, grepl("^São José de", municipio, ignore.case = TRUE
↪  ))
```

```
## # A tibble: 21 x 8
```

```
##       ano    uf cod_ibge municipio             esp_vida tx_analf_15m    pop   r
##     <dbl> <dbl>    <dbl> <chr>                    <dbl>        <dbl>  <dbl> <d
##  1   1991    21  2111201 SÃO JOSÉ DE RIBAMAR       58.2         20.5  74681 17
##  2   1991    24  2412203 SÃO JOSÉ DE MIPIBU        60.5         45.7  27874 15
##  3   1991    25  2514305 SÃO JOSÉ DE CAIANA        59.2         59.2   4724  4
##  4   1991    25  2514404 SÃO JOSÉ DE ESPINHAR~     58.6         53.5   6151  6
##  5   1991    25  2514503 SÃO JOSÉ DE PIRANHAS      60.9         47.4  17511  9
##  6   1991    25  2514552 SÃO JOSÉ DE PRINCESA      56.2         51.2   5816  8
##  7   1991    33  3305133 SÃO JOSÉ DE UBÁ           66.2         31.4   6004 33
##  8   2000    21  2111201 SÃO JOSÉ DE RIBAMAR       65.7         12.1 117856 27
##  9   2000    24  2412203 SÃO JOSÉ DE MIPIBU        68.6         32    34637 20
## 10   2000    25  2514305 SÃO JOSÉ DE CAIANA        63.2         42.6   5709 11
## # ... with 11 more rows
```

3.2 select(): selecionando as colunas

Às vezes, não queremos filtrar as linhas, mas apenas selecionar algumas colunas, sobretudo quando estamos diante de bancos com muitas variáveis (+ 50). Mais uma vez, o *dplyr* oferece um comando intuitivo para esse tipo de esforço. Claro, para selecionar as colunas é necessário conhecer seus nomes.

```
# quais os nomes das colunas
colnames(dados_sociais)

# selecionar ano, uf e taxa de analfabetismo
dados_select <- select(dados_sociais, ano, uf, tx_analf_15m )

# olhar a nova base
head(dados_select)
```

Note que o comando **subset()**, utilizado para filtrar linhas, também conta com uma opção *select*, conforme o código abaixo:

```
dados_select <- subset(dados_sociais, select = c("ano", "uf",
↪  "tx_analf_15m"))
head(dados_select)
```

```
## # A tibble: 6 x 3
##     ano    uf tx_analf_15m
##   <dbl> <dbl>        <dbl>
```

```
## 1  1991    11         23.6
## 2  1991    11         17.2
## 3  1991    11         24.6
## 4  1991    11         21.4
## 5  1991    11         20.3
## 6  1991    11         25.4
```

Você deve estar se perguntando: qual a vantagem dos comandos do *dplyr* sobre o *subset*? Calma! Chegaremos lá. Antes, vamos a mais alguns exemplos.

Com o comando select podemos selecionar um intervalo de variáveis de acordo com seus nomes ou mesmo sua posição no banco.

```
# Selecionar as variáveis de nome até municípios
select(dados_sociais, ano:municipio)
```

```
## # A tibble: 16,695 x 4
##      ano    uf cod_ibge municipio
##    <dbl> <dbl>    <dbl> <chr>
##  1  1991    11  1100015 ALTA FLORESTA D'OESTE
##  2  1991    11  1100023 ARIQUEMES
##  3  1991    11  1100031 CABIXI
##  4  1991    11  1100049 CACOAL
##  5  1991    11  1100056 CEREJEIRAS
##  6  1991    11  1100064 COLORADO DO OESTE
##  7  1991    11  1100072 CORUMBIARA
##  8  1991    11  1100080 COSTA MARQUES
##  9  1991    11  1100098 ESPIGÃO D'OESTE
## 10  1991    11  1100106 GUAJARÁ-MIRIM
## # ... with 16,685 more rows
```

```
# Selecionar as 4 primeiras variáveis pela posição
select(dados_sociais, 1:4)
```

```
## # A tibble: 16,695 x 4
##      ano    uf cod_ibge municipio
##    <dbl> <dbl>    <dbl> <chr>
##  1  1991    11  1100015 ALTA FLORESTA D'OESTE
##  2  1991    11  1100023 ARIQUEMES
##  3  1991    11  1100031 CABIXI
```

```
##  4  1991    11  1100049 CACOAL
##  5  1991    11  1100056 CEREJEIRAS
##  6  1991    11  1100064 COLORADO DO OESTE
##  7  1991    11  1100072 CORUMBIARA
##  8  1991    11  1100080 COSTA MARQUES
##  9  1991    11  1100098 ESPIGÃO D'OESTE
## 10  1991    11  1100106 GUAJARÁ-MIRIM
## # ... with 16,685 more rows
```

Da mesma forma, podemos excluir variáveis com muita facilidade

```
# Todas variáveis, exceto município
dados_select <- select(dados_sociais, -municipio)
```

Ainda, como em qualquer operação do **R**, é possível criar um vetor com as variáveis que se quer selecionar.

```
# Selecionar as variáveis uf e tx_analf_15m
minha_selecao <- c("uf", "tx_analf_15m")
dados_select <- dados_sociais %>% select(minha_selecao)
# mostrar o resultado
head(dados_select)
```

Ainda, quem pode selecionar tem o poder de reordenar as colunas. Use isso a seu favor.

```
# Você pode reordenar grupos
dados_select <- select(dados_sociais,  cod_ibge:rdpc, ano:uf)

# Você pode colocar a variável como a primeira do banco
# note o everything
dados_select <- select(dados_sociais, cod_ibge, everything())
```

Atente: o **everything()** retorna todas outras variáveis do banco.

3.2.1 select(): select helpers

A função *select()* apresenta outras funções que permitem selecionar as variáveis de acordo com os padrões de seus nomes. Para esse exemplo, utilizamos a base de dados de atendimento hospitalar do DATASUS que contém 113 colunas. Numa situação dessas, é importante identificar o padrão de nomenclatura do banco e com os select helpers fica muito fácil selecionar as variáveis desejadas.

```
# ler dados datasus
library(read.dbc)
```

```
sihsus <- read.dbc("C:/curso_r/dados/RDSE1701.dbc")

# nome das colunas
colnames(sihsus)
```

```
##   [1] "UF_ZI"      "ANO_CMPT"   "MES_CMPT"   "ESPEC"      "CGC_HOSP"
##   [6] "N_AIH"      "IDENT"      "CEP"        "MUNIC_RES"  "NASC"
##  [11] "SEXO"       "UTI_MES_IN" "UTI_MES_AN" "UTI_MES_AL" "UTI_MES_TO"
##  [16] "MARCA_UTI"  "UTI_INT_IN" "UTI_INT_AN" "UTI_INT_AL" "UTI_INT_TO"
##  [21] "DIAR_ACOM"  "QT_DIARIAS" "PROC_SOLIC" "PROC_REA"   "VAL_SH"
##  [26] "VAL_SP"     "VAL_SADT"   "VAL_RN"     "VAL_ACOMP"  "VAL_ORTP"
##  [31] "VAL_SANGUE" "VAL_SADTSR" "VAL_TRANSP" "VAL_OBSANG" "VAL_PED1AC"
##  [36] "VAL_TOT"    "VAL_UTI"    "US_TOT"     "DT_INTER"   "DT_SAIDA"
##  [41] "DIAG_PRINC" "DIAG_SECUN" "COBRANCA"   "NATUREZA"   "NAT_JUR"
##  [46] "GESTAO"     "RUBRICA"    "IND_VDRL"   "MUNIC_MOV"  "COD_IDADE"
##  [51] "IDADE"      "DIAS_PERM"  "MORTE"      "NACIONAL"   "NUM_PROC"
##  [56] "CAR_INT"    "TOT_PT_SP"  "CPF_AUT"    "HOMONIMO"   "NUM_FILHOS"
##  [61] "INSTRU"     "CID_NOTIF"  "CONTRACEP1" "CONTRACEP2" "GESTRISCO"
##  [66] "INSC_PN"    "SEQ_AIH5"   "CBOR"       "CNAER"      "VINCPREV"
##  [71] "GESTOR_COD" "GESTOR_TP"  "GESTOR_CPF" "GESTOR_DT"  "CNES"
##  [76] "CNPJ_MANT"  "INFEHOSP"   "CID_ASSO"   "CID_MORTE"  "COMPLEX"
##  [81] "FINANC"     "FAEC_TP"    "REGCT"      "RACA_COR"   "ETNIA"
##  [86] "SEQUENCIA"  "REMESSA"    "AUD_JUST"   "SIS_JUST"   "VAL_SH_FED"
##  [91] "VAL_SP_FED" "VAL_SH_GES" "VAL_SP_GES" "VAL_UCI"    "MARCA_UCI"
##  [96] "DIAGSEC1"   "DIAGSEC2"   "DIAGSEC3"   "DIAGSEC4"   "DIAGSEC5"
## [101] "DIAGSEC6"   "DIAGSEC7"   "DIAGSEC8"   "DIAGSEC9"   "TPDISEC1"
## [106] "TPDISEC2"   "TPDISEC3"   "TPDISEC4"   "TPDISEC5"   "TPDISEC6"
## [111] "TPDISEC7"   "TPDISEC8"   "TPDISEC9"
```

```
# selecionar colunas que começam com VAL
sihsus1 <- select(sihsus, starts_with("val"))
head(sihsus1)
```

```
##    VAL_SH  VAL_SP VAL_SADT VAL_RN VAL_ACOMP VAL_ORTP VAL_SANGUE VAL_SADTSR
## 1 6990.61 1476.66        0      0         0        0          0          0
## 2  745.80  209.98        0      0         0        0          0          0
## 3 7077.69 1699.08        0      0         0        0          0          0
## 4 2142.77  692.39        0      0         0        0          0          0
## 5  416.38  241.54        0      0         0        0          0          0
```

```
## 6  997.50  341.61        0      0         0        0           0           0
##   VAL_TRANSP VAL_OBSANG VAL_PED1AC VAL_TOT VAL_UTI VAL_SH_FED VAL_SP_FED
## 1          0          0          0 8467.27 7659.52          0          0
## 2          0          0          0  955.78    0.00          0          0
## 3          0          0          0 8776.77 7180.80          0          0
## 4          0          0          0 2835.16 1436.16          0          0
## 5          0          0          0  657.92    0.00          0          0
## 6          0          0          0 1339.11    0.00          0          0
##   VAL_SH_GES VAL_SP_GES VAL_UCI
## 1          0          0       0
## 2          0          0       0
## 3          0          0       0
## 4          0          0       0
## 5          0          0       0
## 6          0          0       0
```

```
# selecionar colunas que contém UTI
sihsus2<- select(sihsus, contains("UTI"))
head(sihsus2)
```

```
##   UTI_MES_IN UTI_MES_AN UTI_MES_AL UTI_MES_TO MARCA_UTI UTI_INT_IN UTI_INT_AN
## 1          0          0          0         16        75          0          0
## 2          0          0          0          0        00          0          0
## 3          0          0          0         15        75          0          0
## 4          0          0          0          3        75          0          0
## 5          0          0          0          0        00          0          0
## 6          0          0          0          0        00          0          0
##   UTI_INT_AL UTI_INT_TO VAL_UTI
## 1          0          0 7659.52
## 2          0          0    0.00
## 3          0          0 7180.80
## 4          0          0 1436.16
## 5          0          0    0.00
## 6          0          0    0.00
```

3.3 Encadeando funções com o pipe

Uma das vantagens da aborgagem de manipulação de dados do *dplyr* é o pipe. Introduzido pelo pacote *magrittr*, o *pipe* (expresso pelo comando **%>%**) é uma ferramenta para expressar uma sequência

de múltiplas operações com clareza. Em outras palavras, podemos encadear operações (e essa é uma vantagem relevante) de manipulação de dados. Por exemplo, podemos utilizar o *select* e o *filter* em uma mesma cadeia de comandos.

Mas como isso funciona? Basicamente, o pipe transforma **f(x)** em **x %>% y**.

```
library(magrittr)

x = c(1.555, 2.555, 3.555, 4.555)
# tirar o log de x
log(x)

# mesma coisa com o pipe
x %>% log()

# vc pode ir além!
x %>% log() %>% round(2) # UAU!
```

Como o **R** é uma linguagem funcional, o uso dos *pipes* ajuda a reduzir o número de parênteses nas funções e a deixar o código organizado. Além disso, auxilia a leitura dos códigos da direita para a esquerda. Por fim, para a manipulação de dados a ser feita com o *dplyr*, o uso do *pipe* permite um acesso mais fácil às variáveis

Se aplicarmos filtros para as linhas e selecionarmos colunas com o *dplyr*, o código ficará assim:

```
# esperança de vida nos municipios do Estado de Sergipe em 2010
esp_vida_se_2010 <- dados_sociais %>%
  filter(ano == 2010 & uf == 28) %>%
  select(municipio, esp_vida)
head(esp_vida_se_2010, 4)
```

```
## # A tibble: 4 x 2
##   municipio               esp_vida
##   <chr>                      <dbl>
## 1 AMPARO DE SÃO FRANCISCO     68.7
## 2 AQUIDABÃ                    69.8
## 3 ARACAJU                     74.4
## 4 ARAUÁ                       72
```

Assim, no código acima, o pipe indica a seleção da base dados_sociais, a qual se aplicará um filtro de linhas e serão selecionadas algumas colunas. Muito simples, não? A partir de agora, vamos utilizar sempre o pipe para a manipulação.

3.4 mutate(): criar novas variáveis

A criação de novas variáveis no *R* é bastante facilitada com o *dplyr*. Mostraremos a maneira dessa aborgagem em comparação à forma usual do sistema.

Com o banco de dados sociais, vamos calcular a renda total do município e o logarítimo da população.

```
# encontrar a renda total e logaritmo da populacao
dados_sociais <- dados_sociais %>%
  mutate(renda_total = rdpc * pop,
         log_pop = log(pop))
# veja as novas colunas criadas
head(dados_sociais, 2)
```

```
## # A tibble: 2 x 10
##     ano    uf cod_ibge municipio esp_vida tx_analf_15m   pop  rdpc renda_tota
##   <dbl> <dbl>    <dbl> <chr>        <dbl>        <dbl> <dbl> <dbl>       <dbl
## 1  1991    11  1100015 ALTA FLO~     62.0         23.6 23417  198.    4647338
## 2  1991    11  1100023 ARIQUEMES     66.0         17.2 56061  319.   17909808
## # ... with 1 more variable: log_pop <dbl>
```

```
colnames(dados_sociais)
```

```
##  [1] "ano"          "uf"           "cod_ibge"     "municipio"    "esp_vida"
##  [6] "tx_analf_15m" "pop"          "rdpc"         "renda_total"  "log_pop"
```

Essa é uma abordagem muito mais concisa em relação à base do sistema, observe:

```
# sintaxe verborrágica
dados_sociais$renda_total <- dados_sociais$rdpc *dados_sociais$pop
dados_sociais$log_pop <- log(dados_sociais$pop)
```

O mutate é um comando simples e é por meio dele que faremos a maior parte das nossas construções de variáveis.

3.5 rename()

Renomear variáveis pode ser um pequeno pesadela no R. O *dplyr* apresenta uma forma bastente simplificada para renomear variáveis do banco.

```
# Renomear ano e uf
dados_sociais %>%
  rename(ANO = ano,
         UF = uf)
```

3.6 summarise() e group_by(): agregar os dados

Agora, suponha o seu interesse em identificar a média e mediana da esperança de vida dos municípios brasileiros para cada ano. Como fazer isso?

Esse tipo de operação, em que agrupamos uma estatística por grupo, é muito comum e de fácil solução aqui. Devemos sempre perguntar: quero observar esse fenômeno sob qual nível de agregação? Pensando assim, fica fácil entender o comando abaixo.

```
# agrupar por ano
dados_sociais %>%   # define o banco a ser analisado
  group_by(ano) %>% # agrupa pela variável ano
  summarise(media = mean(esp_vida), # extrai a média
            mediana = median(esp_vida)) # extrai a mediana
```

```
## # A tibble: 3 x 3
##     ano media mediana
##   <dbl> <dbl>   <dbl>
## 1  1991  63.7    64.5
## 2  2000  68.4    69.0
## 3  2010  73.1    73.5
```

```
# agrupar por ano e uf
dados_sociais %>% # banco
  group_by(ano, uf) %>% # agrupa pelas vars ano e uf
  summarise(media = mean(esp_vida),
            mediana = median(esp_vida)) %>%
  head(10)
```

```
## # A tibble: 10 x 4
## # Groups:   ano [1]
##      ano    uf media mediana
##    <dbl> <dbl> <dbl>   <dbl>
##  1  1991    11  62.1    62.7
##  2  1991    12  63.0    63.6
##  3  1991    13  61.7    61.9
```

```
##  4  1991    14  61.4    60.4
##  5  1991    15  62.5    62.9
##  6  1991    16  63.2    63.9
##  7  1991    17  59.9    60.2
##  8  1991    21  57.0    57.0
##  9  1991    22  58.6    58.8
## 10  1991    23  59.9    59.8
```

- Note que executamos duas operações com o *pipe*. Primeiro, agrupamos pela(s) variável(eis) de interesse. Depois, pedimos a média e a mediana para cada um desses anos.

3.7 Exercício

Filtre os dados Extraia a média da esperança de vida por Estado a cada ano. Armazene os dados em um objeto chamado: “med_esp_ano_uf”

4 Mesclar dados no R

4.1 JOIN (merge): melhor que o PROCV

Em geral, as bases de dados disponíveis não estão completas: precisam ser cruzadas para obter maiores informações. Por exemplo, no caso da base **dados_sociais**, não foi especificada a região a qual os municípios e estados pertencem. Como proceder?

No excel, certamente você iria se atrapalhar um pouco até aprender a utilizar o PROCV. Felizmente, o **R** apresenta soluções rápidas e intuitivas para essa atividade.

Examinando os dados, nota-se que as regiões representam o primeiro número da variável *uf*.

Código	Região
1	Norte
2	Nordeste
3	Sudeste
4	Sul
5	Centro Oeste

Portanto, essa é uma chave para mesclar as tabelas.

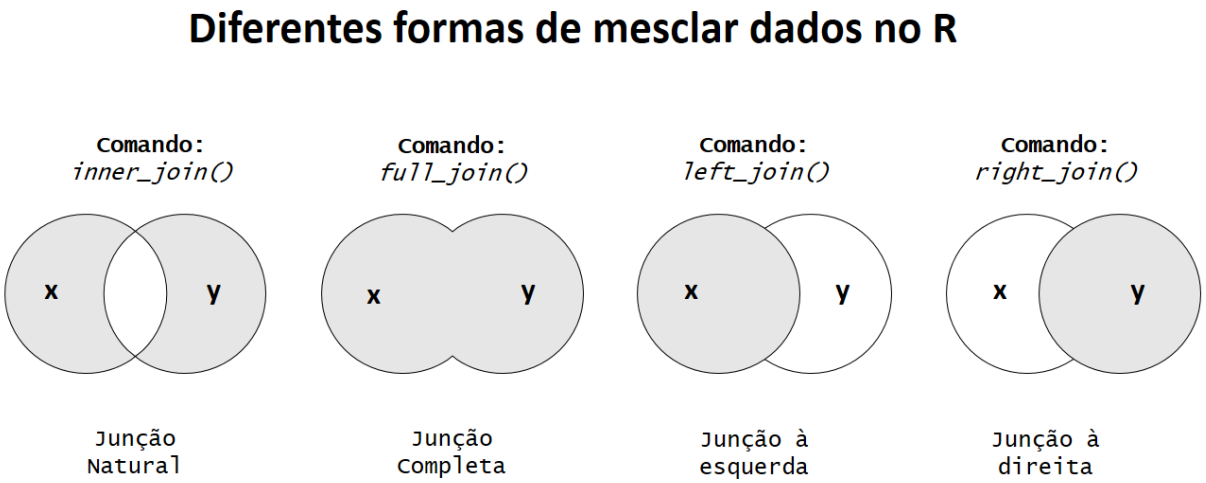


Figura 2: Ilustração das formas de merge no R

Se os nomes das variáveis chaves são iguais, o *dplyr* as identifica

```
inner_join( x , y )     é igual a    inner_join( x , y , by = "chave" )
```

Se os nomes das variáveis chaves são distintos, devem ser identificados

```
inner_join( x , y , by = c( "chave_de_x" = "chave_de_y"  ) )


inner_join( x , y , by = c( "chave1_de_x" = "chave1_de_y" ,
                            "chave2_de_x" = "chave2_de_y" ) )
```

Figura 3: Ilustração das formas de merge no R

4.2 JOIN (merge): junção natural

- Para ilustrar as formas de mesclar dados no **R**, vamos criar dois pequenos dataframes.

```
df1 <- tibble(letras = letters[1:8], X = 1:8)
df2 <- tibble(letras = letters[5:12], Y = 1:8)
```

- Assim, o comando geral é:

```
# Apenas os dados em comum
inner_join(df1, df2)
# Idêntico, mas preferível!
inner_join(df1, df2, by = "letras")
```

4.3 JOIN (merge): outros casos

```
# Junção total
full_join(df1, df2, by = "letras")

# Junção à esquerda
left_join(df1, df2, by = "letras")

# Junção à direita
right_join(df1, df2, by = "letras")
```

- Ainda, pode-se mesclar apenas os dados não coincidentes.

```
anti_join(df1, df2, by ="letras")
anti_join(df2, df1, by ="letras")
```

5 Exercícios

5.1 Fixação do comandos básicos

- Aplicaremos os comandos aprendidos para efetura análises dos dados sobre o programa Financiamento Estudantil (FIES) e do banco *dados_sociais*.
- Essa base administrativa retrata a população dos alunos matriculados no programa FIES.
- Para esse execício, será utilizada uma amostra de 10% das observações.
- Acesse o arquivo: **exercicio_fixacao_dplyr.pdf**

5.2 Primeiro passo: quem são as variáveis?

- Vamos seguir os seguintes passos:

1. No diretório atual encontram-se os arquivos? Tente utilizar **list.files()**.
2. Baixe o arquivo **fies_sample.csv**.

3. Cheque a estrutura do banco. **str()**
4. Identifique o nome das variáveis **colnames()**. Se quiser, salve em um objeto.

5.3 select(): selecionando variáveis relevantes

- O banco tem 50 variáveis. Nem todas são relevantes.
- Selecione algumas variáveis relevantes e salve no objeto fies_sub:
  - UF, código do contrato, raça, sexo, valor da mensalidade, nome da mantenedora, a data de nascimento, quantidade de semestres financiados, descrição e código do curso e situação de ensino médio escola pública .
- **Pergunta:** é possível selecionar as variáveis apenas por alguns atributos dos nomes (ex: DS, CO, NO, ST ou QT)?

5.4 Dicas

- Em caso de dúvida, use os mecanismos de ajuda:
  - help(comando): help(mutate)
  - ?comando: ?mutate
  - também use os *cheetsheets* do *dplyr* **aqui** ou na pasta *cheet_sheets*
  - ou acesse a página oficial do *dplyr* **aqui**

6 Mais funções para manipulação de dados

6.1 funções auxiliares

- Na sumarização dos dados, algumas funções são muito úteis para

Função	Decrição
n()	Número de observações no grupo
n_distinct()	Valores únicos de um vetor
cumsum()	Soma cumulativa
rank()	Ranqueia Variáveis
any()	Alguns valores são verdadeiros?
all()	Todos valores são verdadeiros?
quantile()	Quantis
ifelse()	Criar booleano

Acesse a pasta cheat_sheet para mais dicas

6.2 arrange(): classificando os dados

- No *Excel* é comum ordenar os dados. O comando **arrange()** permite fazer isso com muita facilidade.

```
# Exibir os dados de acordo com a menor população
dados_sociais %>% arrange(pop)
```

- **Exercício**: de acordo com a base, quais os três municípios com a menor população no ano de 2010?

6.3 n(): contando informações

Quando observamos uma base de dados, muitas vezes queremos contar quantos valores uma determinada observação carrega. Essa informação é facilmente adquirida com a combinação do comando group_by() e summarise.

```
# Utilizar o exemplo do datasus
library(read.dbc)
sihsus <- read.dbc("C:/curso_r/dados/RDSE1701.dbc")

# contar quantos casos por municipio
sihsus %>%
  group_by(MUNIC_RES) %>%
```

```
  summarise(n = n()) %>%
  arrange(-n)
```

```
## # A tibble: 136 x 2
##    MUNIC_RES     n
##    <fct>     <int>
##  1 280030     1732
##  2 280480      671
##  3 280350      393
##  4 280670      326
##  5 280210      296
##  6 280290      214
##  7 280130      213
##  8 280740      170
##  9 280020      141
## 10 280300      126
## # ... with 126 more rows
```

6.4 top_n(): selecionando as maiores ou menores linhas por valor

Essa função é uma forma conveniente de filtrar os dados de acordo com o ranking da variável a ser observada.

```
# filtrar as maiores populacoes do banco
dados_sociais %>%
  filter(ano == 2010) %>%
  top_n( 10, pop) %>%
  arrange(-pop)
```

```
## # A tibble: 10 x 10
##      ano    uf cod_ibge municipio esp_vida tx_analf_15m    pop  rdpc renda_to
##    <dbl> <dbl>    <dbl> <chr>        <dbl>        <dbl>  <dbl> <dbl>       <d
##  1  2010    35  3550308 SÃO PAULO     76.3         3.18 1.12e7 1516.     1.69
##  2  2010    33  3304557 RIO DE J~     75.7         2.88 6.27e6 1493.     9.35
##  3  2010    29  2927408 SALVADOR      75.1         3.97 2.65e6  973      2.58
##  4  2010    53  5300108 BRASÍLIA      77.4         3.47 2.54e6 1715.     4.36
##  5  2010    23  2304400 FORTALEZA     74.4         6.94 2.43e6  846.     2.05
##  6  2010    31  3106200 Belo Hor~     76.4         2.87 2.36e6 1497.     3.53
##  7  2010    13  1302603 MANAUS        74.5         3.78 1.79e6  790.     1.42
```

```
##  8  2010    41  4106902 CURITIBA      76.3         2.13 1.74e6 1581.     2.75
##  9  2010    26  2611606 RECIFE        74.5         7.13 1.52e6 1144.     1.74
## 10  2010    43  4314902 PORTO AL~     76.4         2.27 1.39e6 1758.     2.45
## # ... with 1 more variable: log_pop <dbl>
```

6.5 distinct(): extirpando linhas repetidas

Um problema comum em dados administrativos é a repetição de registros - por exemplo, uma mesma observação é repetida em diversos meses seguidos. As estimativas que fizemos dos contratos do FIES não são tão precisas pois há grande números de contratos repetidos ao longo dos meses. No entanto, buscamos apenas informações únicas sobre os contratos - por exemplo, quantas pessoas inscritas pelo FIES são do sexo feminino e cursam direito. Para isso, devemos identificar a existência de casos repetidos e a de casos únicos.

```
# há casos duplicados?
fies_sample %>%
  select(CO_CONTRATO_FIES) %>%
  duplicated() %>%
  sum()
```

```
## [1] 200351
```

```
# quantos casos únicos?
fies_sample %>%
  summarise(unicos = n_distinct(CO_CONTRATO_FIES))
```

```
##   unicos
## 1 606856
```

- Você percebeu que a soma do resultado do comando duplicated() é uma soma de um vetor lógico?*

Como a base administrativa do FIES é semestral, cada contrato tem diversas observações para cada mês. Então, é necessário encontrar as observações únicas:

```
fies_sub_dist <- fies_sample %>%
  distinct(CO_CONTRATO_FIES, .keep_all = TRUE)
```

O argumento *.keep_all = TRUE* mantém todas as variáveis no banco.

6.6 Exemplo

- Veja um exemplo de como encontrar o valor médio da mensalidade

```
mens_uf <- fies_sample %>%
  group_by(SG_UF) %>%
  select(SG_UF, VL_MENSALIDADE)%>%
  summarise(media_mens = mean(VL_MENSALIDADE, na.rm = TRUE))
mens_uf
```

```
## # A tibble: 27 x 2
##    SG_UF media_mens
##    <chr>      <dbl>
##  1 AC         1161.
##  2 AL          841.
##  3 AM          831.
##  4 AP         1300.
##  5 BA         1060.
##  6 CE         1121.
##  7 DF         1136.
##  8 ES         1155.
##  9 GO         1060.
## 10 MA          959.
## # ... with 17 more rows
```

Veja como o pipe **%>%** permite encadear uma grande quantidade de comandos.

6.7 ifelse(): criar booleno

Outra função comum para organizar base de dados é a criação de booleanos. Imagine que você quer categorizar as mensalidades médias entre os estados de acordo com um determinado valor. Se a mensalidade for superior a R$1.000,00, você classificará como **1**, nos outros casos como **0**. Essa operação é muito simples. Como você está criando uma nova coluna, há necessidade de utilizar o *mutate()* em conjunto com um operador lógico.

Primeiro, verifique quais os argumentos do *ifelse()*. ?ifelse

```
mens_uf %>%
  mutate(acima_mil = ifelse(media_mens >= 1000, 1, 0),
         acima_mil2 = ifelse(media_mens >= 1000, "Maior que R$1.000",
                             "Menor que R$1.000"))
```

```
## # A tibble: 27 x 4
##    SG_UF media_mens acima_mil acima_mil2
##    <chr>      <dbl>     <dbl> <chr>
##  1 AC         1161.         1 Maior que R$1.000
##  2 AL          841.         0 Menor que R$1.000
##  3 AM          831.         0 Menor que R$1.000
##  4 AP         1300.         1 Maior que R$1.000
##  5 BA         1060.         1 Maior que R$1.000
##  6 CE         1121.         1 Maior que R$1.000
##  7 DF         1136.         1 Maior que R$1.000
##  8 ES         1155.         1 Maior que R$1.000
##  9 GO         1060.         1 Maior que R$1.000
## 10 MA          959.         0 Menor que R$1.000
## # ... with 17 more rows
```

O comando não limita a criação de booleanos numéricos, pois permite criar o mesmo um classificador de texto. Ainda, é possível fazer múltiplas condições encadeadas

```
mens_uf %>%
  mutate(tres_classe = ifelse(media_mens <900, 0,
                              ifelse(media_mens >= 900 & media_mens <= 1100,
                              ↪  1, 2)))
```

```
## # A tibble: 27 x 3
##    SG_UF media_mens tres_classe
##    <chr>      <dbl>       <dbl>
##  1 AC         1161.           2
##  2 AL          841.           0
##  3 AM          831.           0
##  4 AP         1300.           2
##  5 BA         1060.           1
##  6 CE         1121.           2
##  7 DF         1136.           2
##  8 ES         1155.           2
##  9 GO         1060.           1
## 10 MA          959.           1
## # ... with 17 more rows
```

6.8 round() - arredondar

No *Excel* arredondar os números exige pouco esforço. Assim também ocorre no **R**.

```
# Gerar uma distribuição normal aleatória
x <- rnorm(10, 5, 1)
# Arredondar
round(x)
# Arredondar com duas casas decimais
round(x, digits = 2)
# Ou ainda...
round(x, 2)
```

Claro, você pode aplicar isso à lógica de manipulação do *dplyr*. Nesse caso, vamos aplicar não só o arredondamento comum, mas o *ceiling()* e o *floor()*.

```
#
mens_uf %>%
  mutate(media_mens1 = round(media_mens, 1),
         media_mens2 = ceiling(media_mens),
         media_mens3 = floor(media_mens))
```

```
## # A tibble: 27 x 5
##    SG_UF media_mens media_mens1 media_mens2 media_mens3
##    <chr>      <dbl>       <dbl>       <dbl>       <dbl>
##  1 AC         1161.       1161.        1162        1161
##  2 AL          841.        841          841         840
##  3 AM          831.        831.         831         830
##  4 AP         1300.       1300.        1301        1300
##  5 BA         1060.       1060         1061        1060
##  6 CE         1121.       1121.        1122        1121
##  7 DF         1136.       1136         1137        1136
##  8 ES         1155.       1155.        1155        1154
##  9 GO         1060.       1060.        1060        1059
## 10 MA          959.        959          960         959
## # ... with 17 more rows
```

6.9 any() - algum valor é verdadeiro?

- Em um banco grande não é possível inspecionar visualmente elementos como os *NAs* ou outras informações.

- A função *any()* permite identificar facilmente se algum elemento possui determinada característica especificada.

```
# Algum elemento do banco "df_na" é NA?
any(is.na(df_na))
# A coluna letras do banco "df_na" é NA?
any(is.na(df_na$letras))
# A coluna letras do banco "df_na" contém a letra E?
any(df_na == "E")
any(df_na$idade >10)
```

Outra utilidade é a criação de variáveis de acordo com grupos. Para explicitar a situação, Para o exemplo, vamos utilizar dados de uma pesquisa de domicílios.

```
# baixar os dados
dom <- read_csv2("C:/curso_r/dados/dados_domiciliares.csv",
          locale = locale(encoding = "Latin1"))
```

```
## Using ',' as decimal and '.' as grouping mark. Use read_delim() for more cont
```

```
## Parsed with column specification:
## cols(
##   domicilio = col_double(),
##   condicao_domicilio = col_character(),
##   sexo = col_character(),
##   idade = col_double(),
##   cor = col_character(),
##   sabe_ler = col_character(),
##   escolaridade = col_character()
## )
```

```
# olhar os dados
glimpse(dom)
```

```
## Observations: 4,000
## Variables: 7
## $ domicilio          <dbl> 1.1e+10, 1.1e+10, 1.1e+10, 1.1e+10, 1.1e+10, 1.1..
## $ condicao_domicilio <chr> "Pessoa responsável pelo domicílio", "Filho(a) s..
## $ sexo               <chr> "Mulher", "Homem", "Mulher", "Homem", "Homem", "..
## $ idade              <dbl> 51, 22, 43, 33, 6, 24, 22, 36, 50, 18, 17, 34, 1..
## $ cor                <chr> "Parda", "Parda", "Parda", "Branca", "Parda", "P..
```

```
## $ sabe_ler           <chr> "Não", "Sim", "Não", "Não", "Sim", "Sim", "Não",..
## $ escolaridade       <chr> "Regular do ensino fundamental ou do 1° grau", N..
```

```
# nome das colunas
colnames(dom)
```

```
## [1] "domicilio"          "condicao_domicilio" "sexo"
## [4] "idade"              "cor"                "sabe_ler"
## [7] "escolaridade"
```

Agora, como saber quantos domicílio tem idosos (pessoas acima de 60 anos)? Você pode agrupar os dados por domicílios (*group_by()*) e utilizar o comando *any()* em conjunto com o *ifelse()*.

```
dom <- dom %>%
  group_by(domicilio) %>%
  mutate(dom_idoso = ifelse(any(idade >= 60),
                            "Dom. tem idoso", "Não tem idoso"))

# quantos domicílios tem idosos
dom %>%
  distinct(dom_idoso , .keep_all = TRUE) %>%
  group_by(dom_idoso) %>%
  count()
```

```
## # A tibble: 2 x 2
## # Groups:   dom_idoso [2]
##   dom_idoso          n
##   <chr>          <int>
## 1 Dom. tem idoso   340
## 2 Não tem idoso    975
```

6.10 cut(): transformar dados numéricos em categóricos

Imagine que você tem um vetor com a idade de diversos indivíduo e há a necessidade de reclassifica-la faixas etárias. Por exemplo, a cada 5 anos.

```
# cortar as idades em intervalos de 5 anos
dom <- dom %>%
  mutate( idade_cut = cut(idade, seq(0,100, 5)))
```

É claro que essa solução, embora prática, não gera um resultado legível. Nesse caso, você pode utilizar uma função pronta, como a *age_cat*, que está na nossa pasta de funções.

```
# ler uma função criada
source("C:/curso_r/funcoes/age_cat.R")
# ler os argumentos da funcao
args(age.cat)
```

```
## function (x, lower = 0, upper, by = 10, sep = "-", above.char = "+")
## NULL
```

```
# criar categoria de idades
dom <- dom %>%
  mutate( idade_cat = age.cat(idade, lower = 0,
                              upper = 90))
```

6.11 paste(): concatenar strings

-Muitas vezes faz-se necessário editar elementos de texto no **R**. Por exemplo, os nomes de um banco.
- A função paste permite fazer isso de forma direta.

```
# Irmãos Peixoto
irmaos <- c("Edgar", "Edclésia", "Edmar", "Edésia", "Edésio")

# Como colocar os sobrenomes?
paste(irmaos, "Peixoto")
```

No *dplyr*, você poder criar variáveis juntando colunas. Suponha que você quer uma categoria para especificar o sexo e cor das pessoas pesquisadas.

```
dom <- dom %>%
  mutate(sexo_cor = paste(sexo, "-", cor))

dom %>%
  group_by(sexo_cor) %>%
  count()
```

```
## # A tibble: 10 x 2
## # Groups:   sexo_cor [10]
##    sexo_cor              n
##    <chr>             <int>
##  1 Homem - Amarela       8
##  2 Homem - Branca      592
##  3 Homem - Indígena      6
```

```
##  4 Homem - Parda      1240
##  5 Homem - Preta       161
##  6 Mulher - Amarela      4
##  7 Mulher - Branca     606
##  8 Mulher - Indígena     7
##  9 Mulher - Parda     1243
## 10 Mulher - Preta      133
```

6.12 gsub() ou str_replace(): padrões e substituição

Há uma série de comandos que facilitam a identificação de padrões e substituição desses valores. Por exemplo, suponha a necessidade de substituir um **-** por uma **@**.

```
# funcionamento do gsub
gsub("-", "@", "curso-hotmail.com")
```

```
## [1] "curso@hotmail.com"
```

```
# funcionamento do str_replace
library(stringr)
str_replace("curso-hotmail.com","-", "@")
```

```
## [1] "curso@hotmail.com"
```

Essas operações podem ser utilizadas em conjunto com *mutate*, veja:

```
# aplicando a um data.frame
dom %>%
  mutate(sexo_cor2 = gsub("-", "@", sexo_cor)) %>%
  select(sexo_cor2) %>%  #selecionar variável de interesse
  head(10) # mostrar 10 primeiros
```

```
## Adding missing grouping variables: domicilio
```

```
## # A tibble: 10 x 2
## # Groups:   domicilio [4]
##       domicilio sexo_cor2
##           <dbl> <chr>
##  1 11000001601 Mulher @ Parda
##  2 11000001601 Homem @ Parda
```

```
##  3 11000001603 Mulher @ Parda
##  4 11000001603 Homem @ Branca
##  5 11000001603 Homem @ Parda
##  6 11000001604 Homem @ Parda
##  7 11000001604 Mulher @ Branca
##  8 11000001605 Homem @ Parda
##  9 11000001605 Mulher @ Parda
## 10 11000001605 Homem @ Branca
```

Exercício

Repita o último exemplo aplicando desta vez a função *str_replace*.

6.13 substr(): extrair subconjunto de um vetor character

Algumas vezes precisamos extrair subconjunto de um vetor *character*. Por exemplo, temos a ação orçamentária com sua descrição, mas queremos apenas o código da ação. Nesse caso, como o código da ação é fixo - apena 4 dígitos, é fácil recuperar essa informação.

```
acoes <- c("0005 - Sentenças Judiciais Transitadas em Julgado
↪  (Precatórios)",
           "00IN - Benefícios de Prestação Continuada (BPC) à Pessoa com
            ↪  Deficiência e da Renda Mensal Vitalícia (RMV) por
            ↪  Invalidez",
           "20JQ - Realização e Apoio a Eventos de Esporte, Lazer e
            ↪  Inclusão Social")

# substring do vetor acoes
substr(acoes, 1, 4)
```

```
## [1] "0005" "00IN" "20JQ"
```

Outra possibilidade poderia ser extrair apenas as descrições das ações. Nesse caso, combinamos o *substr* com o comando *nchar*.

```
# começamos na 8ª posicao
# e terminamos no tamanho da observacao
substr(acoes, 8, nchar(acoes))
```

```
## [1] "Sentenças Judiciais Transitadas em Julgado (Precatórios)"
```

```
## [2] "Benefícios de Prestação Continuada (BPC) à Pessoa com Deficiência e da R
## [3] "Realização e Apoio a Eventos de Esporte, Lazer e Inclusão Social"
```

A mesma operação pode ser feita por meio do comando *str_sub()* do pacote **stringr**.

```
library(stringr)
str_sub(acoes, 1, 4)
```

```
## [1] "0005" "00IN" "20JQ"
```

```
str_sub(acoes, 8, nchar(acoes))
```

```
## [1] "Sentenças Judiciais Transitadas em Julgado (Precatórios)"
## [2] "Benefícios de Prestação Continuada (BPC) à Pessoa com Deficiência e da R
## [3] "Realização e Apoio a Eventos de Esporte, Lazer e Inclusão Social"
```

Exercício Você pode tentar extrair as três primeiras letras de todas cidades do banco *dados_sociais*.

6.14 grepl():padrões e substituição

A função *grepl* é bastante utilizada para encontrar padrões de palavras nas *strings*.

```
dom %>% filter(grepl("Pret", cor, ignore.case = TRUE))
```

```
## # A tibble: 294 x 11
## # Groups:   domicilio [200]
##    domicilio condicao_domici~ sexo  idade cor   sabe_ler escolaridade dom_ido
##        <dbl> <chr>            <chr> <dbl> <chr> <chr>    <chr>        <chr>
##  1   1.10e10 Pessoa responsá~ Mulh~    60 Preta Não      Antigo prim~ Dom. te
##  2   1.10e10 Pessoa responsá~ Mulh~    40 Preta Não      Superior -
~ Dom. tem~
##  3   1.10e10 Pessoa responsá~ Homem    83 Preta Não      <NA>         Dom. te
##  4   1.10e10 Pai, mãe, padra~ Homem   103 Preta Não      Classe de a~ Dom. te
##  5   1.10e10 Neto(a)          Mulh~    22 Preta Não      Regular do ~ Dom. te
##  6   1.10e10 Cônjuge ou comp~ Mulh~    77 Preta Sim      <NA>         Dom. te
##  7   1.10e10 Pessoa responsá~ Homem    34 Preta Não      Regular do ~ Não tem
##  8   1.10e10 Filho(a) do res~ Mulh~     9 Preta Sim      <NA>         Não tem
```

```
##  9   1.10e10 Pessoa responsá~ Mulh~    24 Preta Não      Regular do ~ Não tem
## 10   1.10e10 Cônjuge ou comp~ Mulh~    35 Preta Não      Educação de~ Não tem
## # ... with 284 more rows, and 3 more variables: idade_cut <fct>,
## #   idade_cat <fct>, sexo_cor <chr>
```

6.15 lead() e lag(): achar valores anteriores e posteriores em um vetor

Outra tarefa comum em uma base de dados é tentar observar uma variável defasada ou adiantada.

```
x = 1:10
# uma defasagem
lag(x)
```

```
##  [1] NA  1  2  3  4  5  6  7  8  9
```

```
# duas defasagens
lag(x, 2)
```

```
##  [1] NA NA  1  2  3  4  5  6  7  8
```

```
# adiantar uma vez
lead(x)
```

```
##  [1]  2  3  4  5  6  7  8  9 10 NA
```

```
# adiantar duas vezes
lead(x, 2)
```

```
##  [1]  3  4  5  6  7  8  9 10 NA NA
```

Claro isso, pode ser aplicado à fórmula básica do *dplyr*.

```
# criar uma série
df = tibble(a = rep(1:5, 3),
            b = rnorm(15, 15, 10))

# lead e lag
df %>%
  mutate(c = lead(b, 1),
         d = lag(b, 2))
```

```
## # A tibble: 15 x 4
##        a      b      c      d
##    <int>  <dbl>  <dbl>  <dbl>
##  1     1 32.8   20.5   NA
##  2     2 20.5    3.08  NA
##  3     3  3.08   9.10  32.8
##  4     4  9.10   7.34  20.5
##  5     5  7.34  21.0    3.08
##  6     1 21.0    0.910  9.10
##  7     2  0.910 21.5    7.34
##  8     3 21.5    8.86  21.0
##  9     4  8.86  12.6    0.910
## 10     5 12.6    6.88  21.5
## 11     1  6.88  28.3    8.86
## 12     2 28.3    0.154 12.6
## 13     3  0.154 16.8    6.88
## 14     4 16.8   19.2   28.3
## 15     5 19.2   NA      0.154
```

No exemplo abaixo, temos um caso mais concreto. Suponha que você quer saber a variação da renda per capita nos municípios do Estado de Rondônia.

```
TO <- dados_sociais %>%
  filter(uf == "11") %>%
  select(ano:municipio, rdpc)

# exibir primeiras linhas
head(TO)
```

```
## # A tibble: 6 x 5
##     ano    uf cod_ibge municipio              rdpc
##   <dbl> <dbl>    <dbl> <chr>                 <dbl>
## 1  1991    11  1100015 ALTA FLORESTA D'OESTE  198.
## 2  1991    11  1100023 ARIQUEMES              319.
## 3  1991    11  1100031 CABIXI                 116.
## 4  1991    11  1100049 CACOAL                 320.
## 5  1991    11  1100056 CEREJEIRAS             240.
## 6  1991    11  1100064 COLORADO DO OESTE      225.
```

```
# lead
TO <- TO %>%
  group_by(cod_ibge) %>%
  mutate(lag_rdpc = lag(rdpc),
         var_rdpc = (rdpc - lag_rdpc)/ lag_rdpc,
         var_rdpc_pec = paste(round(var_rdpc*100, 2), "%"))
```

Exercício

- Faça o uso do o lead e lag para todo o banco dados sociais. - Observe quais municípios tiveram o maior aumento da renda per capita no período. - Quais tiveram redução. - Crie uma variável booleana para isso.

6.16 bind_cols e bind_rows

- Os comandos **bind_** permitem ligar colunas e linhas de bancos com dimensões iguais.

```
# extrair linhas específicas
um  <-  dados_sociais[1:4, ]
dois <- dados_sociais[7011:7014, ]

# ligar em um novo objeto
meu_bind <- bind_rows(um, dois)
```

- Os comandos **bind_** permitem ligar colunas e linhas de bancos com dimensões iguais.

```
# extrair colunas
um   <- dados_sociais[ , 3 ]
dois <- dados_sociais[ , 8 ]

# ligar em um novo objeto
meu_bind2 <- bind_cols(um, dois)
```

6.17 NAs: valores não especificados ou perdidos

- Ao realizar o **full_join**, o resultado apresenta algumas observações como *NA* (ver exemplo acima).
- Os *NAs* podem representar tanto informações indeterminadas quanto valores propositadamente omitidos.
- De qualquer forma, lidar com os NAs é muito fácil:

```
# Data frame com NAs
df_na <- tibble( letras = LETTERS[1:10],
                 idade =  c( seq(25L, 40L, 5L), NA,
                             NA, 32L, rep(NA, 3)))

# é possível saber se um vetor é NA
is.na(df_na)
# Também é possível remover esses valores
na.omit(df_na)
```

6.18 replace_na() : substituir NAs

- Porém, não é recomendado retirar os NAs sem alguma reflexão. Afinal, eles podem dizer alguma coisa sobre os dados. Ou, ainda, serem apenas campos numúricos não preenchidos.
- Por isso, é melhor substituir os NAs.

```
# Substituindo NAs por números
replace_na(df_na, list(idade = 0))

# Substituindo NAs por textos
replace_na(df_na, list(idade = "Idade não informada"))
```

6.19 quantile() e ntile()

Para calcular os intervalos de uma distribuição acumulada de uma variável (quantis), há funções bastante diretas.

```
# calcular quintis
quantile(dom$idade)
```

```
##   0%  25%  50%  75% 100%
##    0   16   32   48  103
```

```
# calcular decis
quantile(dom$idade, seq(0.1, 1, 0.1))
```

```
##  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    7   14   19   26   32   38   44   51   61  103
```

Exercício

Encontre os percentis da variável idade no banco *domicílio*.

No dplyr, a função *ntile()* divide o vetor em *n* grupos de mesmo tamanho.

```
dom <- dom %>%
  ungroup() %>% # desagrupa os dados
  mutate(quintil = ntile(idade, 5),
         decil = ntile(idade, 10))

# todos os grupos tem mesmo tamanho
dom %>%
  group_by(decil) %>%
  count()
```

```
## # A tibble: 10 x 2
## # Groups:   decil [10]
##    decil     n
##    <int> <int>
##  1     1   400
##  2     2   400
##  3     3   400
##  4     4   400
##  5     5   400
##  6     6   400
##  7     7   400
##  8     8   400
##  9     9   400
## 10    10   400
```

7 Trabalhar com datas

Lidar com datas pode ser um pequeno problema na análise de dados. Em geral, quando importamos planilhas as datas são importadas como texto e, dessa forma, não nos permite extrair informações como o ano, mês, dia da semana, entre outras, que podem ser úteis para nossas análises. Vamos utilizar o exemplo do seguro defeso.

- Primeiro, instale e ative o pacote *lubridate*:
  - você lembra como instalar e ativar pacotes?
  - dica: *inst.…*

7.1 lubridate: transformando as datas em datas (!)

- O lubridate opera datas levando em conta que:
  - *y* = ano
  - *m* = mês
  - *d* = dia
- Assim, você pode transformar *characters* em datas caso tenham dia, mês e ano, da seguinte forma:

```
# ano, mês e dia sem separador
ymd("20180131")
```

```
## [1] "2018-01-31"
```

```
# mês, dia e ano com separador "-"
mdy("01-31-2018")
```

```
## [1] "2018-01-31"
```

```
# dia, mês e ano com separador "/"
dmy("31/01/2018")
```

```
## [1] "2018-01-31"
```

Observe também que quando as datas são apresentadas no formato mês/ano, não há como transformar em datas. Por isso, você pode fazer uso do *paste()* para completar as suas datas.

```
# criamos um vetor com datas mes (m) e ano (y)
datas <- c("01/2014", "03/2016")
# adicionamos o dia (d) com o paste0
dmy(paste0("01", datas))
```

```
## [1] "2014-01-01" "2016-03-01"
```

7.2 lubridate: acessando informações das datas

- Estruturar os dados para o formato de datas permite extrair informações básicas para realizar operações lógicas.

```r
# criar um vetor de datas
datas<- ymd(c("20180131", "20170225", "20160512"))
# quais os anos
year(datas)
```

```
## [1] 2018 2017 2016
```

```r
# quais os meses
months(datas) # formato nome
```

```
## [1] "janeiro"   "fevereiro" "maio"
```

```r
month(datas) # fomato numero
```

```
## [1] 1 2 5
```

```r
# dias
day(datas)
```

```
## [1] 31 25 12
```

```r
# dia da semana
wday(datas)
```

```
## [1] 4 7 5
```

7.3 Exercício

- Leia os dados do seguro defeso na pasta dados. Lembre-se de efetuar uma inspeção visual antes.
- Peça o **head()** do banco. Quantas datas você identifica?
- A partir do banco seguro defeso, verifique:
  - as datas presentes no banco tem a classe de data? Utilize **is.Date()**
  - a data de início do defeso ocorre no período abrangido pelo banco de dados?
  - a data de início do defeso ocorre no período abrangido pelo banco de dados?
  - *melhor*: o defeso ocorre no período do banco de dados?
  - como retirar do banco as linhas em que o defeso não corresponda ao período abrangido pelo banco de dados?
  - em que dias os saques das parcelas ocorreram? Qual dia concentra mais saques?