Tratamento de Dados

Prof. Ricardo B. Sampaio

Introdução ao pacote dplyr

Aula do dia 20 de abril de 2018 - Ciência de Dados para Todos e Ciência de Dados Aplicada

```
library(jsonlite); library(readxl)
library(tidyverse) #pacotes dplyr, ggplot2, tidyr, purr, readr, tibble
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

Funções básicas do dplyr também conhecidos como vebos de ação:

Escolher as observações: **filter()** Reorganziar as observações: **arrange()** Escolher as variáveis: **select()** Criar novas variáveis com base nas variáveis existentes: **mutate()** Juntar valores múltiplos em um único valor: **summary()**

Além das funções filter e arrange as observações podem ser divididas em pedaços menores utilizando **group_by()**

Todos os verbos funcionam de forma similar, o primeiro argumento é o Data Frame e os argumentos subsequentes descrevem o que fazer com os dados usando os nomes das variáveis SEM as aspas.

Para que sejam aplicados os conceitos acima descritos é primeiro feito a leitura dos dados para o ambiente de trabalho e estes são inicialmente analisados e transformados para o formato Data Frame(DF). Lendo os arquivos para Análise de Ciência do lattes (perfis, publicação e orientação), a BDTD, o OASIS e o DGP.

```
setwd("~/Desktop/CD4A") #Pasta onde estão os arquivos para análise CD4A data
#unb.perf <- fromJSON("data/unb.perfis.json")
unb.ori <- fromJSON("data/unb.relatorioOrientacao.json")
unb.pub <- fromJSON("data/unb.relatorioProducaoBibiografica.json")
#unb.oasis <- fromJSON("data/oasisbr_unb.json")
#unb.bdtd <- fromJSON("data/bdtd_unb.json")
#unb.dgp <- list(); for (i in 1:5) (unb.dgp[[i]] <- read_xls("data/UnBCD-01-2018.xls", sheet = i))
```

Depois de feita a importação dos dados é imporatante analisar os dados e transformá-los do formato lista para o formato DF. Isso é feito como exemplo para os arquivos de publicação e orientação

```
#análise da estrutura e importação
#Produção bibliográfica
names(unb.pub);
```

```
## [1] "PERIODICO"         "LIVRO"             "CAPITULO_DE_LIVRO"
## [4] "TEXTO_EM_JORNAIS"  "ARTIGO_ACEITO"
```

```r
names(unb.pub$PERIODICO)
```

```
## [1] "2012" "2013" "2014" "2015" "2016" "2017"
```

```r
unb.pub.df <- data.frame()
for (i in 1:length(unb.pub$PERIODICO))
  unb.pub.df <- rbind(unb.pub.df, unb.pub$PERIODICO[[i]])
names(unb.pub.df)
```

```
##  [1] "natureza"         "titulo"           "periodico"
##  [4] "ano"              "volume"           "issn"
##  [7] "paginas"          "doi"              "autores"
## [10] "autores-endogeno"
```

```r
dim(unb.pub.df)
```

```
## [1] 9806   10
```

```r
#str(unb.pub.df); summary(unb.pub.df)
glimpse(unb.pub.df)
```

```
## Observations: 9,806
## Variables: 10
## $ natureza           <chr> "COMPLETO", "COMPLETO", "COMPLETO", "COMPLE...
## $ titulo             <chr> "A survey of non-abelian tensor products of...
## $ periodico          <chr> "Boletim da Sociedade Paranaense de Matem\u...
## $ ano                <chr> "2012", "2012", "2012", "2012", "2012", "20...
## $ volume             <chr> "30", "16", "26", "6", "18", "14", "15", "6...
## $ issn               <chr> "21751188", "01048740", "01023306", "187090...
## $ paginas            <chr> "77 - 89", "297 - 306", "607 - 618", "359 -...
## $ doi                <chr> "10.5269/bspm.v30i1.13350", "", "10.1590/S0...
## $ autores            <list> [<"Nakaoka, Irene N.", "Rocco, Nora\u00ed ...
## $ `autores-endogeno` <list> ["0000507838194708", "0002528252697017", "...
```

```r
#Orientação
names(unb.ori)
```

```
## [1] "ORIENTACAO_EM_ANDAMENTO_DE_POS_DOUTORADO"
## [2] "ORIENTACAO_EM_ANDAMENTO_DOUTORADO"
## [3] "ORIENTACAO_EM_ANDAMENTO_MESTRADO"
## [4] "ORIENTACAO_EM_ANDAMENTO_GRADUACAO"
## [5] "ORIENTACAO_EM_ANDAMENTO_INICIACAO_CIENTIFICA"
## [6] "ORIENTACAO_CONCLUIDA_POS_DOUTORADO"
## [7] "ORIENTACAO_CONCLUIDA_DOUTORADO"
## [8] "ORIENTACAO_CONCLUIDA_MESTRADO"
```

```r
(print(names(unb.ori$ORIENTACAO_CONCLUIDA_DOUTORADO))) ==
  (print(names(unb.ori$ORIENTACAO_CONCLUIDA_MESTRADO)))
```

```
## [1] "2012" "2013" "2014" "2015" "2016" "2017"
## [1] "2012" "2013" "2014" "2015" "2016" "2017"
```

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

```r
print(names(unb.ori$ORIENTACAO_CONCLUIDA_DOUTORADO$`2012`)) ==
  print(names(unb.ori$ORIENTACAO_CONCLUIDA_MESTRADO$`2012`))
```

```
##  [1] "natureza"                    "titulo"
##  [3] "ano"                         "id_lattes_aluno"
```

```
##  [5] "nome_aluno"                  "instituicao"
##  [7] "curso"                       "codigo_do_curso"
##  [9] "bolsa"                       "agencia_financiadora"
## [11] "codigo_agencia_financiadora" "nome_orientadores"
## [13] "id_lattes_orientadores"
##  [1] "natureza"                    "titulo"
##  [3] "ano"                         "id_lattes_aluno"
##  [5] "nome_aluno"                  "instituicao"
##  [7] "curso"                       "codigo_do_curso"
##  [9] "bolsa"                       "agencia_financiadora"
## [11] "codigo_agencia_financiadora" "nome_orientadores"
## [13] "id_lattes_orientadores"
```

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

```
unb.ori.tipo.df <- data.frame(); unb.ori.df <- data.frame()
for (i in 1:length(unb.ori[[1]]))
  unb.ori.tipo.df <- rbind(unb.ori.tipo.df, unb.ori$ORIENTACAO_CONCLUIDA_POS_DOUTORADO[[i]])
unb.ori.df <- rbind(unb.ori.df, unb.ori.tipo.df); unb.ori.tipo.df <- data.frame()
for (i in 1:length(unb.ori[[1]]))
  unb.ori.tipo.df <- rbind(unb.ori.tipo.df, unb.ori$ORIENTACAO_CONCLUIDA_DOUTORADO[[i]])
unb.ori.df <- rbind(unb.ori.df, unb.ori.tipo.df); unb.ori.tipo.df <- data.frame()
for (i in 1:length(unb.ori[[1]]))
  unb.ori.tipo.df <- rbind(unb.ori.tipo.df, unb.ori$ORIENTACAO_CONCLUIDA_MESTRADO[[i]])
unb.ori.df <- rbind(unb.ori.df, unb.ori.tipo.df); unb.ori.tipo.df <- data.frame()

#str(unb.ori.df); summary(unb.ori.df)
glimpse(unb.ori.df)
```

```
## Observations: 6,464
## Variables: 13
## $ natureza                    <chr> "Supervis\u00e3o de p\u00f3s-douto...
## $ titulo                      <chr> "", "Organiza\u00e7\u00e3o das ass...
## $ ano                         <chr> "2012", "2012", "2012", "2012", "2...
## $ id_lattes_aluno             <chr> "", "", "", "", "", "", "", "", ""...
## $ nome_aluno                  <chr> "Jansen Rodrigo Pereira Santos", "...
## $ instituicao                 <chr> "Universidade de Bras\u00edlia", "...
## $ curso                       <chr> "", "", "", "", "", "", "", "", ""...
## $ codigo_do_curso             <chr> "", "", "", "", "", "", "", "", ""...
## $ bolsa                       <chr> "SIM", "SIM", "SIM", "SIM", "NAO",...
## $ agencia_financiadora        <chr> "Embrapa", "Conselho Nacional de D...
## $ codigo_agencia_financiadora <chr> "002600000997", "002200000000", "0...
## $ nome_orientadores           <list> ["Robert Neil Gerard Miller", "Ro...
## $ id_lattes_orientadores      <list> ["0960398662960668", "14566574218...
```

Os dois conjunto de dados que serão utilizados como referência para os próximos exemplos são unb.pub.df e unb.ori.df

Use filter() para filtrar observações

O comando filter permite que se faça subconjunto dos dados com base nas informações contidas nas observações.

```
glimpse(unb.pub.df)
```

```
## Observations: 9,806
## Variables: 10
```

```
## $ natureza           <chr> "COMPLETO", "COMPLETO", "COMPLETO", "COMPLE...
## $ titulo             <chr> "A survey of non-abelian tensor products of...
## $ periodico          <chr> "Boletim da Sociedade Paranaense de Matem\u...
## $ ano                <chr> "2012", "2012", "2012", "2012", "2012", "20...
## $ volume             <chr> "30", "16", "26", "6", "18", "14", "15", "6...
## $ issn               <chr> "21751188", "01048740", "01023306", "187090...
## $ paginas            <chr> "77 - 89", "297 - 306", "607 - 618", "359 -...
## $ doi                <chr> "10.5269/bspm.v30i1.13350", "", "10.1590/S0...
## $ autores            <list> [<"Nakaoka, Irene N.", "Rocco, Nora\u00ed ...
## $ `autores-endogeno` <list> ["0000507838194708", "0002528252697017", "...
```

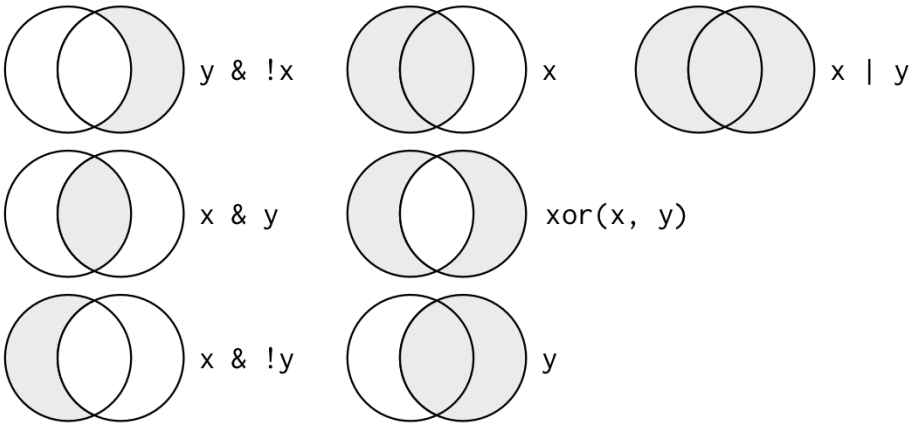
```
#View(unb.pub.df) Usar no RStudio
dim(unb.pub.df)
```

```
## [1] 9806   10
```

```
df <- filter(unb.pub.df, ano == 2017)
glimpse(df)
```

```
## Observations: 1,028
## Variables: 10
## $ natureza           <chr> "COMPLETO", "COMPLETO", "COMPLETO", "COMPLE...
## $ titulo             <chr> "Imers\u00e3o internacional: uma avalia\u00...
## $ periodico          <chr> "Revista Gestao Universitaria na America La...
## $ ano                <chr> "2017", "2017", "2017", "2017", "2017", "20...
## $ volume             <chr> "10", "23", "12", "1", "114", "102", "292",...
## $ issn               <chr> "19834545", "15167313", "2238975X", "030486...
## $ paginas            <chr> "280 - 301", "303 - 320", "183 - 205", "1 -...
## $ doi                <chr> "", "10.1590/1516-731320170020002", "", "10...
## $ autores            <list> [<"NASCIMENTO, A. A.", "Nunes, A.", "GON\u...
## $ `autores-endogeno` <list> ["0002528252697017", "0022322386442215", "...
```

Você pode usar também outrs operadores de comparação como >, >=, <, <=, != (não é igual), e == (igual). Veja que existe uma grande diferença entre = e == O primeiro é utilizado para atribuir valores como <- e o segundo é para comparar dois lados.



Além desses tem também os operadores lógicos

```
df <- filter(unb.pub.df, ano == 2015 | ano == 2016)
table(df$ano)
```

```
##
## 2015 2016
## 1826 1953
```

Caso queira incluir vários sem ter que repetir a variável pode ser usado o comando %in%

```
df <- filter(unb.pub.df, ano %in% c(2015, 2016, 2017))
table(df$ano)
```

```
##
## 2015 2016 2017
## 1826 1953 1028
```

Reorganize as observações usando arrange()

arrange() funciona de maneira similar que o filter mas ao invés de selecionar as observações ele reorganiza em ordem crescente ou decresente ou outra pré-definida

```
glimpse(unb.ori.df)
```

```
## Observations: 6,464
## Variables: 13
## $ natureza                    <chr> "Supervis\u00e3o de p\u00f3s-douto...
## $ titulo                      <chr> "", "Organiza\u00e7\u00e3o das ass...
## $ ano                         <chr> "2012", "2012", "2012", "2012", "2...
## $ id_lattes_aluno             <chr> "", "", "", "", "", "", "", "", ""...
## $ nome_aluno                  <chr> "Jansen Rodrigo Pereira Santos", "...
## $ instituicao                 <chr> "Universidade de Bras\u00edlia", "...
## $ curso                       <chr> "", "", "", "", "", "", "", "", ""...
## $ codigo_do_curso             <chr> "", "", "", "", "", "", "", "", ""...
## $ bolsa                       <chr> "SIM", "SIM", "SIM", "SIM", "NAO",...
## $ agencia_financiadora        <chr> "Embrapa", "Conselho Nacional de D...
## $ codigo_agencia_financiadora <chr> "002600000997", "002200000000", "0...
## $ nome_orientadores           <list> ["Robert Neil Gerard Miller", "Ro...
## $ id_lattes_orientadores      <list> ["0960398662960668", "14566574218...
```

```
df <- arrange(unb.ori.df, desc(curso), nome_aluno)
glimpse(df)
```

```
## Observations: 6,464
## Variables: 13
## $ natureza                    <chr> "Disserta\u00e7\u00e3o de mestrado...
## $ titulo                      <chr> "Narrativas sobre a prostitui\u00e...
## $ ano                         <chr> "2016", "2013", "2013", "2013", "2...
## $ id_lattes_aluno             <chr> "6513342593381137", "5746071830265...
## $ nome_aluno                  <chr> "Cyntia Cristina de Carvalho e Sil...
## $ instituicao                 <chr> "Universidade de Bras\u00edlia", "...
## $ curso                       <chr> "sociologia", "psicologia social, ...
## $ codigo_do_curso             <chr> "90000010", "90000021", "90000021"...
## $ bolsa                       <chr> "NAO", "NAO", "SIM", "NAO", "NAO",...
## $ agencia_financiadora        <chr> "", "", "Coordena\u00e7\u00e3o de ...
## $ codigo_agencia_financiadora <chr> "", "", "045000000000", "", "", ""...
## $ nome_orientadores           <list> ["Hayd\u00e9e Gl\u00f3ria Cruz Ca...
## $ id_lattes_orientadores      <list> ["6889569648252727", "27632416142...
```

Selecione as variáveis com a função select()

As vezes para facilitar a manipulação dos dados você precisa apenas de algumas poucas colunas. Para isso você pode usar o select() para definir quais colunas serão utilizadas.

```
df <- select(unb.ori.df, natureza, ano, curso)
glimpse(df)
```

```
## Observations: 6,464
## Variables: 3
## $ natureza <chr> "Supervis\u00e3o de p\u00f3s-doutorado", "Supervis\u0...
## $ ano      <chr> "2012", "2012", "2012", "2012", "2012", "2012", "2012...
## $ curso    <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", ",...
```

```
table(df[,c(1:2)])
```

```
##                                      ano
## natureza                              2012 2013 2014 2015 2016 2017
##   Disserta<U+00E7><U+00E3>o de mestrado      760  933  877  817  777  463
##   Supervis<U+00E3>o de p<U+00F3>s-doutorado   37   44   72   48   39   13
##   Tese de doutorado                    221  310  272  288  300  193
```

```
# outra opção de selecionar as variáveis
df <- select(unb.ori.df, -(curso:nome_orientadores))
glimpse(df)
```

```
## Observations: 6,464
## Variables: 7
## $ natureza              <chr> "Supervis\u00e3o de p\u00f3s-doutorado"...
## $ titulo                <chr> "", "Organiza\u00e7\u00e3o das assemble...
## $ ano                   <chr> "2012", "2012", "2012", "2012", "2012",...
## $ id_lattes_aluno       <chr> "", "", "", "", "", "", "", "", "", "",...
## $ nome_aluno            <chr> "Jansen Rodrigo Pereira Santos", "Renat...
## $ instituicao           <chr> "Universidade de Bras\u00edlia", "Unive...
## $ id_lattes_orientadores <list> ["0960398662960668", "1456657421809884...
```

algumas funções de busca mais aprimorada podem ser utilizadas no select() starts_with("abc"): combina nomes que começam com "abc". ends_with("xyz"): combina nomes que terminam com "xyz". contains("ijk"): corresponde a nomes que contêm "ijk". matches("(.)\1"): seleciona as variáveis que correspondem a uma expressão regular. Este corresponde a quaisquer variáveis que contenham caracteres repetidos. num_range("x", 1:3) corresponde a x1, x2 e x3.

Para renomear as variáveis a melhor opção é utilizar rename() apesar de poder ser feito isso com select() também

```
df <- rename(unb.ori.df, id.lattes.aluno = id_lattes_aluno)
glimpse(df)
```

```
## Observations: 6,464
## Variables: 13
## $ natureza                   <chr> "Supervis\u00e3o de p\u00f3s-douto...
## $ titulo                     <chr> "", "Organiza\u00e7\u00e3o das ass...
## $ ano                        <chr> "2012", "2012", "2012", "2012", "2...
## $ id.lattes.aluno            <chr> "", "", "", "", "", "", "", "", ""...
## $ nome_aluno                 <chr> "Jansen Rodrigo Pereira Santos", "...
## $ instituicao                <chr> "Universidade de Bras\u00edlia", "...
## $ curso                      <chr> "", "", "", "", "", "", "", "", ""...
## $ codigo_do_curso            <chr> "", "", "", "", "", "", "", "", ""...
## $ bolsa                      <chr> "SIM", "SIM", "SIM", "SIM", "NAO",...
## $ agencia_financiadora       <chr> "Embrapa", "Conselho Nacional de D...
## $ codigo_agencia_financiadora <chr> "002600000997", "002200000000", "0...
## $ nome_orientadores          <list> ["Robert Neil Gerard Miller", "Ro...
```

```
## $ id_lattes_orientadores      <list> ["0960398662960668", "14566574218...
```

Para reorganizar a ordem das variáveis utilizar o select() e pode-se valer também do everything() caso queira colocar algumas em primeiro e as outras seguindo

```
df <- select(unb.ori.df, id_lattes_aluno, nome_aluno, everything())
glimpse(df)
```

```
## Observations: 6,464
## Variables: 13
## $ id_lattes_aluno             <chr> "", "", "", "", "", "", "", "", ""...
## $ nome_aluno                  <chr> "Jansen Rodrigo Pereira Santos", "...
## $ natureza                    <chr> "Supervis\u00e3o de p\u00f3s-douto...
## $ titulo                      <chr> "", "Organiza\u00e7\u00e3o das ass...
## $ ano                         <chr> "2012", "2012", "2012", "2012", "2...
## $ instituicao                 <chr> "Universidade de Bras\u00edlia", "...
## $ curso                       <chr> "", "", "", "", "", "", "", "", ""...
## $ codigo_do_curso             <chr> "", "", "", "", "", "", "", "", ""...
## $ bolsa                       <chr> "SIM", "SIM", "SIM", "SIM", "NAO",...
## $ agencia_financiadora        <chr> "Embrapa", "Conselho Nacional de D...
## $ codigo_agencia_financiadora <chr> "002600000997", "002200000000", "0...
## $ nome_orientadores           <list> ["Robert Neil Gerard Miller", "Ro...
## $ id_lattes_orientadores      <list> ["0960398662960668", "14566574218...
```

Criar novas variáveis utilizanto o mutate()

Além de selecionar, reorganizar e suprimir variáveis é possível criar novas colunas com mutate() tomando como base as variáveis existentes.

```
glimpse(unb.pub.df)
```

```
## Observations: 9,806
## Variables: 10
## $ natureza           <chr> "COMPLETO", "COMPLETO", "COMPLETO", "COMPLE...
## $ titulo             <chr> "A survey of non-abelian tensor products of...
## $ periodico          <chr> "Boletim da Sociedade Paranaense de Matem\u...
## $ ano                <chr> "2012", "2012", "2012", "2012", "2012", "20...
## $ volume             <chr> "30", "16", "26", "6", "18", "14", "15", "6...
## $ issn               <chr> "21751188", "01048740", "01023306", "187090...
## $ paginas            <chr> "77 - 89", "297 - 306", "607 - 618", "359 -...
## $ doi                <chr> "10.5269/bspm.v30i1.13350", "", "10.1590/S0...
## $ autores            <list> [<"Nakaoka, Irene N.", "Rocco, Nora\u00ed ...
## $ `autores-endogeno` <list> ["0000507838194708", "0002528252697017", "...
```

```
df <- select(unb.pub.df, periodico, ano, volume, issn)
glimpse(df)
```

```
## Observations: 9,806
## Variables: 4
## $ periodico <chr> "Boletim da Sociedade Paranaense de Matem\u00e1tica"...
## $ ano       <chr> "2012", "2012", "2012", "2012", "2012", "2012", "201...
## $ volume    <chr> "30", "16", "26", "6", "18", "14", "15", "6", "6", "...
## $ issn      <chr> "21751188", "01048740", "01023306", "18709095", "010...
```

```
df <- mutate(df, per.issn = paste(df$issn, df$periodico, sep = " - "))
glimpse(df)
```

```
## Observations: 9,806
## Variables: 5
## $ periodico <chr> "Boletim da Sociedade Paranaense de Matem\u00e1tica"...
## $ ano       <chr> "2012", "2012", "2012", "2012", "2012", "2012", "201...
## $ volume    <chr> "30", "16", "26", "6", "18", "14", "15", "6", "6", "...
## $ issn      <chr> "21751188", "01048740", "01023306", "18709095", "010...
## $ per.issn  <chr> "21751188 - Boletim da Sociedade Paranaense de Matem...
```

Sumarize grupos utilizando summarise()

Utilizando summarise() com group_by() pode ser feito resumos de partes do conjunto de dados.

```
df <- group_by(unb.pub.df, periodico)
df <- summarise(df, count = n())
head(arrange(df, desc(count)), 10)
```

```
## # A tibble: 10 x 2
##                                                 periodico count
##                                                     <chr> <int>
##  1                                               Plos One    81
##  2                  "Tempus: Actas de Sa\u00fade Coletiva"    61
##  3      "Revista Eletr\u00f4nica Gest\u00e3o & Sa\u00fade"    48
##  4                      Revista de Enfermagem UFPE On Line    44
##  5         "Ci\u00eancia e Sa\u00fade Coletiva (Impresso)"    32
##  6                       "Anu\u00e1rio Antropol\u00f3gico"    31
##  7                              Journal of Algebra (Print)    29
##  8                      Sociedade e Estado (UnB. Impresso)    28
##  9 "Revista Psicologia: Organiza\u00e7\u00f5es e Trabalho"    24
## 10                        Business Management Review (BMR)    22
```

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).