# Manipulação de Dados - Parte I

*Paulo Henrique S. Guimarães*

## 1 Conteúdo do curso

- Importação de dados
- Transformação de dados e comandos básicos (*tibbles*, operador *pipe*)
- Análise exploratória de dados
- Trabalhando com fatores e *strings*.

### 1.1 Introdução

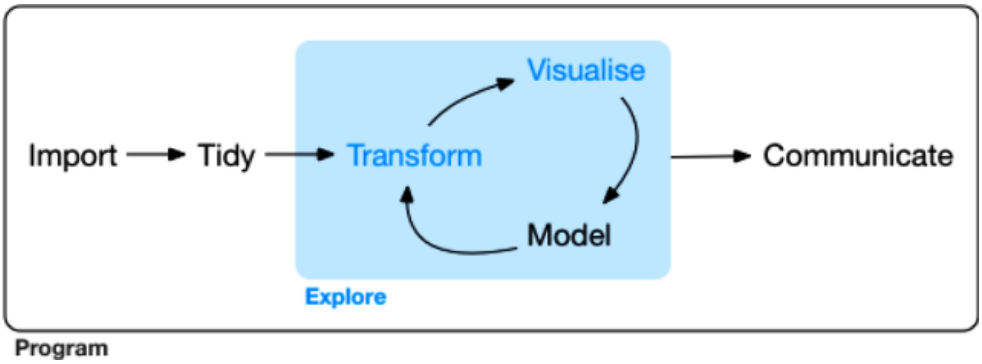


Figure 1: Exploração de dados.

### 1.2 Pacote Tidyverse

O *tidyverse* é um pacote do R, cuja única função é carregar outros pacotes do R. O conjunto desses pacotes forma o *tidyverse*.

**Principais pacotes**

- readr
- tidyr
- dplyr
- ggplot2.

## 2 Importação e Exportação de dados

- Dados delimitados em texto (csv)
- Microsoft Excel (xls, xlsx)


Figure 2: Universo tidyverse.

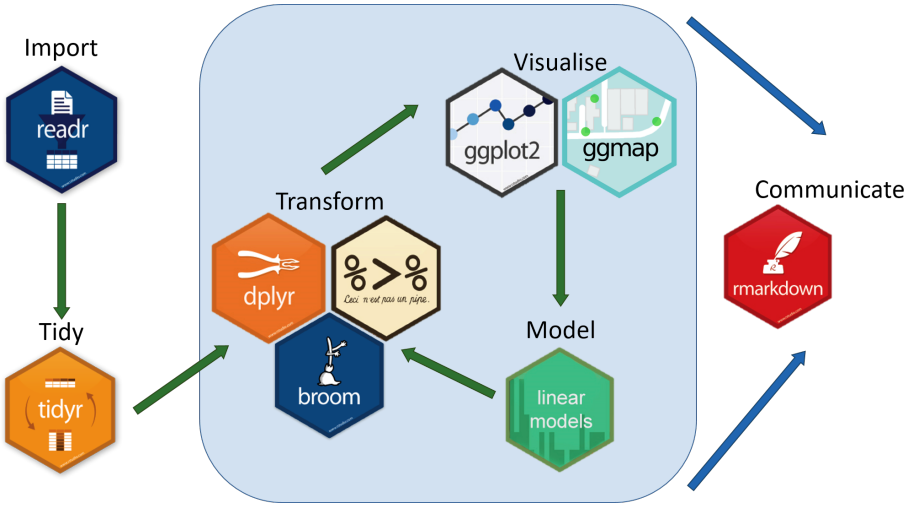


Figure 3: Fluxo de trabalho.

- Arquivos nativos do R (RData, rds)
- Formato fst
- SQLite (SQLITE)
- Texto não estruturado (txt)
- Outros formatos...

```
getwd() # ver meu diretório
```

```
## [1] "H:/Manipulaçao_dados"
```

## 2.1 Importação de dados

O R possui funções para importar dados de arquivos *.csv* - **read.csv** e o **read.csv2**.

- No *tidyverse* temos a opção de utilizar o pacote *readr* que oferece maior velocidade de leitura e melhor

tratamento das classes de colunas.

- Comando **read_csv**.

```
# install.packages('tidyverse')

library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.1
```

```
## -- Attaching packages ---------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.6.1
```

```
## Warning: package 'tibble' was built under R version 3.6.1
```

```
## Warning: package 'tidyr' was built under R version 3.6.1
```

```
## Warning: package 'dplyr' was built under R version 3.6.1
```

```
## Warning: package 'stringr' was built under R version 3.6.1
```

```
## -- Conflicts ------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)

link <- "http://transparencia.al.gov.br/media/arquivo/comparativo_despesas-2017.zip"

# download.file(link, destfile = "E:/DS/comparativo_despesas-2017.zip") # baixar o arquivo

 # unzip("H:/Manipulaçao_dados/comparativo_despesas-2017.zip")

df_despesas <- read_delim("comparativo_despesas-2017.txt",
                          delim = "|",locale = locale(encoding = "ISO-8859-1"),
                          # progress = FALSE, para nao mostrar a barra de progresso
                          progress = FALSE)
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   ANO = col_double(),
##   MES = col_double(),
##   NATUREZA1 = col_double(),
##   NATUREZA2 = col_double(),
##   NATUREZA3 = col_double(),
##   NATUREZA4 = col_double(),
##   NATUREZA5 = col_double(),
##   NATUREZA6 = col_double(),
##   NATUREZA = col_double(),
##   DATA_REGISTRO = col_date(format = ""),
##   PROJETO_ATIVIDADE_ID = col_double(),
##   PROGRAMA_ID = col_double(),
##   SUB_FUNCAO_ID = col_double(),
```

```
##   PT_FUNCAO_ID = col_double(),
##   FONTE_MAE_ID = col_double(),
##   FONTE_ID = col_double(),
##   FL_DIARIA = col_double(),
##   FL_FAVORECIDO = col_double(),
##   SUBTITULO = col_double(),
##   VALOR_EMPENHADO = col_double()
##   # ... with 2 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning: 22 parsing failures.
##    row       col expected    actual                            file
## 261996 ANO       a double ANO       'comparativo_despesas-2017.txt'
## 261996 MES       a double MES       'comparativo_despesas-2017.txt'
## 261996 NATUREZA1 a double NATUREZA1 'comparativo_despesas-2017.txt'
## 261996 NATUREZA2 a double NATUREZA2 'comparativo_despesas-2017.txt'
## 261996 NATUREZA3 a double NATUREZA3 'comparativo_despesas-2017.txt'
## ...... ......... ........ ......... ..............................
## See problems(...) for more details.
```

```
# names(df_despesas)
# head(df_despesas)
# tail(df_despesas)
# class(df_despesas)
# str(df_despesas)
#summary(df_despesas)

glimpse(df_despesas)
```

```
## Observations: 261,996
## Variables: 44
## $ ANO                  <dbl> 2017, 2017, 2017, 2017, 2017, 2017, 2017,...
## $ MES                  <dbl> 11, 12, 11, 3, 7, 9, 12, 6, 3, 3, 10, 7, ...
## $ UG                   <chr> "410548", "510023", "540035", "210013", "...
## $ DESCRICAO_UG         <chr> "AG DE MODERNIZACAO DA GESTAO DE PROCESSO...
## $ GESTAO               <chr> "41548", "00001", "00001", "00001", "0256...
## $ PT                   <chr> "04122000420010000", "10122000420930000",...
## $ PT_DESCRICAO         <chr> "MANUTENCAO DAS ATIVIDADES DO ORGAO", "MA...
## $ FONTE_MAE            <chr> "0100", "0100", "0100", "0100", "0291", "...
## $ DESCRICAO_FONTE_MAE  <chr> "RECURSOS ORDINARIOS", "RECURSOS ORDINARI...
## $ FONTE                <chr> "0100000000", "0100000000", "0100000000",...
## $ DESCRICAO_FONTE      <chr> "RECURSOS ORDINARIOS", "RECURSOS ORDINARI...
## $ PI                   <chr> "000344", "003442", "001829", "002247", "...
## $ CODIGO_FAVORECIDO    <chr> "PF0000001", "PF0000001", "16493672449", ...
## $ NOME_FAVORECIDO      <chr> "FOLHA PAGTOPESSOAL", "FOLHA PAGTOPESSOAL...
## $ NATUREZA1            <dbl> 3e+08, 3e+08, 3e+08, 3e+08, 3e+08, 3e+08,...
## $ DESCRICAO_NATUREZA1  <chr> "DESPESA", "DESPESA", "DESPESA", "DESPESA...
## $ NATUREZA2            <dbl> 3.3e+08, 3.3e+08, 3.3e+08, 3.3e+08, 3.3e+...
## $ DESCRICAO_NATUREZA2  <chr> "DESPESAS CORRENTES", "DESPESAS CORRENTES...
## $ NATUREZA3            <dbl> 3.31e+08, 3.31e+08, 3.33e+08, 3.33e+08, 3...
## $ DESCRICAO_NATUREZA3  <chr> "PESSOAL E ENCARGOS SOCIAIS", "PESSOAL E ...
## $ NATUREZA4            <dbl> 331900000, 331900000, 333900000, 33390000...
## $ DESCRICAO_NATUREZA4  <chr> "APLICACOES DIRETAS", "APLICACOES DIRETAS...
## $ NATUREZA5            <dbl> 331900000, 331900000, 333900000, 33390000...
```

```
## $ DESCRICAO_NATUREZA5  <chr> "APLICACOES DIRETAS", "APLICACOES DIRETAS...
## $ NATUREZA6            <dbl> 331901100, 331901100, 333903600, 33390390...
## $ DESCRICAO_NATUREZA6  <chr> "VENC.E VANTAGENS FIXAS - PESSOAL CIVIL",...
## $ NATUREZA             <dbl> 331901129, 331901104, 333903615, 33390395...
## $ DESCRICAO_NATUREZA   <chr> "COMPLEMENTACAO SALARIAL- PESSOAL CIVIL (...
## $ DATA_REGISTRO        <date> 2017-11-30, 2017-12-15, 2017-11-14, 2017...
## $ PROJETO_ATIVIDADE_ID <dbl> 201720010000, 201720930000, 201723700000,...
## $ PROGRAMA_ID          <dbl> 20170004, 20170004, 20170004, 20170004, 2...
## $ SUB_FUNCAO_ID        <dbl> 2017122, 2017122, 2017181, 2017122, 20170...
## $ PT_FUNCAO_ID         <dbl> 201704, 201710, 201706, 201704, 201702, 2...
## $ FONTE_MAE_ID         <dbl> 20170100, 20170100, 20170100, 20170100, 2...
## $ FONTE_ID             <dbl> 2.017010e+13, 2.017010e+13, 2.017010e+13,...
## $ FL_DIARIA            <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1,...
## $ FL_FAVORECIDO        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1,...
## $ SUBTITULO            <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ DESCRICAO_SUBTITULO  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ ORGAO                <chr> "13548", "27000", "19000", "11000", "0200...
## $ ORGAO_DESCRICAO      <chr> "AGENCIA DE MODERNIZACAO DA GEST.DE PROCE...
## $ VALOR_EMPENHADO      <dbl> 6058.93, 5379.41, 13582.22, 996.47, 3840....
## $ VALOR_LIQUIDADO      <dbl> 6058.93, 5379.41, 6791.11, 996.47, 0.00, ...
## $ VALOR_PAGO           <dbl> 0, 0, 0, 0, 0, 1091, 0, 0, 0, 0, 0, 0, 0,...
```

## 2.2 Exportação de dados

Para criarmos um arquivo **.csv** basta utilizar a função **readr::write_csv**.

```
library(readr)

x<-1000
dados<-data.frame(y = runif(x),
                  z = rep('a',x))
head(dados)
```

```
##            y z
## 1 0.49151700 a
## 2 0.65436393 a
## 3 0.55313227 a
## 4 0.42937510 a
## 5 0.73798233 a
## 6 0.01432261 a
```

```
dir<-'H:/Manipulaçao_dados/dados.csv'
write_csv(dados,path = dir)

dados2<-read_csv(dir,col_types = cols(y = col_double(),z = col_character()))
print(head(dados2))
```

```
## # A tibble: 6 x 2
##        y z
##    <dbl> <chr>
## 1 0.492  a
## 2 0.654  a
## 3 0.553  a
## 4 0.429  a
## 5 0.738  a
```

```
## 6 0.0143 a
```

### 2.3 Arquivos Excel (xls e xlsx)

```
# install.packages('writexl')

library(writexl)
```

```
## Warning: package 'writexl' was built under R version 3.6.2
```

```
N<-500
df<-data.frame(y = seq(1,N),
               z = rep('b',N))
write_xlsx(df,path = dir)
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.6.1
```

```
# endereço do exemplo: http://www.uel.br/pessoal/silvano/Dados/Tilapia.xls

tilapias<-read_excel('Tilapia.xls')
# print(tilapias)
head(tilapias)
```

```
## # A tibble: 6 x 8
##   Turma Equipe Inducao  Peso  Comp   Alt Comp_cabeca Recup
##   <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl>       <dbl> <dbl>
## 1  2000   2021     165  408.  29     7.3         7.8  17
## 2  2000   2021     183  400   29.5   9           7.3   8.9
## 3  2000   2021     161  397.  29.3   8.7         8    28.7
## 4  2000   2021     108  432.  29.5   9.1         7.6 115
## 5  2000   2021     146  336.  26.2   8.5         6.9   8.9
## 6  2000   2021     147  309.  25.8   8.1         6.7  20.5
```

# 3 Referência


1) Livro: R for Data Science - Hadley Wickham & Garrett Grolemund. Alta Books, 2019.