Formação Cientista de Dados

Dia 05 - Laboratório

Vítor Wilher

Cientista de Dados | Mestre em Economia

Plano de Voo

Pacotes para Data Science

Pacotes para Banco de Dados

Exemplo SQL

Exemplo de BI

Produção de Relatórios e Apresentações

Exemplo de Machine Learning

Pacotes para Data Science

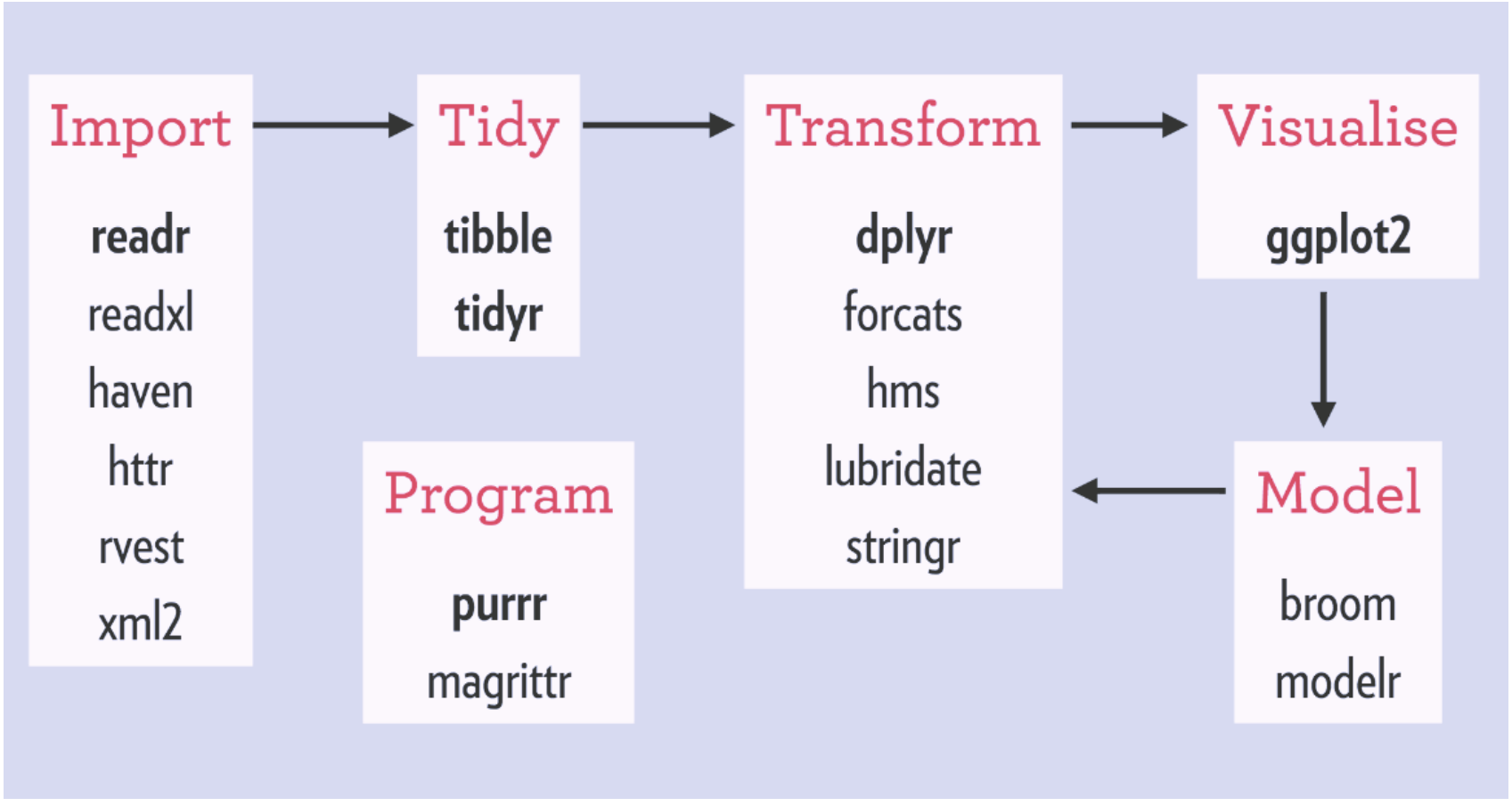

Figure 1: Pacotes do `tidyverse`

Pacotes para Banco de Dados

- **DBI** - Pacote para interação com *sistemas de gerenciamento de bases de dados*. Ver aqui;
- **RODBC** - Pacote para bases de dados ODBC. Ver aqui;
- **RJDBC** - Pacote para bases de dados JDBC. Ver aqui;
- **data.table** - Pacote para tornar mais rápido a leitura e tratamento de dados em `data frames`. Ver aqui;
- Diversos pacotes para leitura de extensões externas, como xlsx, xml, json, sas, spss, stata, etc.
- **sqldf** - Manipulação de R data frames com SQL;

Exemplo SQL

O pacote RSQLite permite fazer a conexão com *databases* SQL, totalmente compatível com interfaces DBI. Assim, vamos instalar os pacotes abaixo para começar nossa integração com o mundo SQL.

```
install.packages('DBI')
install.packages('RSQLite')
```

Exemplo SQL

```
library(RSQLite)
library(DBI)
```

Exemplo SQL

Agora, podemos criar uma conexão com uma database. No caso abaixo, estamos criando uma conexão vazia apenas como exercício. Vamos adicionar coisas a ela daqui a pouco.

```
con <- dbConnect(RSQLite::SQLite(), ":memory:") # Substituir por sua database
dbListTables(con)
```

```
## character(0)
```

Exemplo SQL

Uma vez criada a conexão podemos adicionar um *dataset* apenas para ver algumas funções.

```
dbWriteTable(con, "mtcars", mtcars)
```

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

```
dbListTables(con)
```

```
## [1] "mtcars"
```

```
dbListFields(con, "mtcars")
```

```
##  [1] "row_names" "mpg"       "cyl"       "disp"      "hp"
##  [6] "drat"      "wt"        "qsec"      "vs"        "am"
## [11] "gear"      "carb"
```

Exemplo SQL

```
dbReadTable(con, "mtcars")
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
## Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
## Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
## Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
## Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
## Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
## Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
## Fiat 128          32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic       30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla    33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona     21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Dodge Challenger  15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
## AMC Javelin       15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
## Camaro Z28        13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
## Pontiac Firebird  19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
## Fiat X1-9         27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2     26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa      30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Ford Pantera L    15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
## Ferrari Dino      19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
```

Exemplo SQL

Vamos agora usar um *SQL query* para pegar todos os dados quando `cyl=4`, como abaixo.

```
res <- dbSendQuery(con, "SELECT * FROM mtcars WHERE cyl = 4")
dbFetch(res)
```

```
##                 mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Datsun 710     22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Merc 240D      24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230       22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Fiat 128       32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
## Honda Civic    30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
## Toyota Corolla 33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
## Toyota Corona  21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
## Fiat X1-9      27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
## Porsche 914-2  26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
## Lotus Europa   30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
## Volvo 142E     21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

Exemplo SQL

```
res <- dbSendQuery(con, "SELECT * FROM mtcars WHERE cyl = 4")
while(!dbHasCompleted(res)){
  chunk <- dbFetch(res, n = 5)
  print(nrow(chunk))
}
```

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

Exemplo SQL

Limpe os resultados e feche a conexão com a *database* com as funções abaixo.

```
dbClearResult(res)
dbDisconnect(con)
```

Exemplo SQL

É possível executar funções SQL com o pacote **sqldf**. Instale-o e carregue-o no RStudio. Carregue também o pacote **PASWR**. Vamos utilizar o *dataset titanic3* desse pacote para exemplificar algumas funções SQL.

```
library(sqldf)
library(PASWR)
data("titanic3")
```

Exemplo SQL

Uma vez que tenha carregado esses pacotes e *dataset*, utilize a função `sqldf` para contar o número de linhas do dataset `titanic3`. Utilizando funções do R, como seria?

```
sqldf("select count(*) from titanic3")
```

```
##   count(*)
## 1     1309
```

```
nrow(titanic3)
```

```
## [1] 1309
```

Exemplo SQL

Selecione agora todas as linhas e colunas do objeto titanic3 e coloque em um novo objeto, chamado `TitanicData` com a função sqldf. Utilizando funções do R, como seria?

```
TitanicData <- sqldf("select * from titanic3")
TitanicData <- titanic3[ , ]
```

Exemplo SQL

Selecione as duas primeiras colunas do objeto titanic3 e coloque em um objeto chamado `TitanicSubset2Cols` com a função sqldf. Utilizando funções do R, como seria?

```
colnames(titanic3)
```

```
##  [1] "pclass"    "survived"  "name"      "sex"       "age"
##  [6] "sibsp"     "parch"     "ticket"    "fare"      "cabin"
## [11] "embarked"  "boat"      "body"      "home.dest"
```

```
TitanicSubset2Cols <- sqldf("select pclass,survived
                            from titanic3")
```

Exemplo SQL

Faça o *print* das seis primeiras linhas do objeto titanic3 com a função sqldf. Utilizando funções do R, como seria?

```
sqldf("select * from titanic3 limit 6")
```

```
##   pclass survived                                 name    sex     age sibsp
## 1    1st        1    Allen, Miss. Elisabeth Walton female 29.0000     0
## 2    1st        1   Allison, Master. Hudson Trevor   male  0.9167     1
## 3    1st        0     Allison, Miss. Helen Loraine female  2.0000     1
## 4    1st        0 Allison, Mr. Hudson Joshua Crei   male 30.0000     1
## 5    1st        0 Allison, Mrs. Hudson J C (Bessi female 25.0000     1
## 6    1st        1              Anderson, Mr. Harry   male 48.0000     0
##   parch ticket     fare   cabin    embarked boat body
## 1     0  24160 211.3375      B5 Southampton    2   NA
## 2     2 113781 151.5500 C22 C26 Southampton   11   NA
## 3     2 113781 151.5500 C22 C26 Southampton        NA
## 4     2 113781 151.5500 C22 C26 Southampton       135
## 5     2 113781 151.5500 C22 C26 Southampton        NA
## 6     0  19952  26.5500     E12 Southampton    3   NA
##                         home.dest
## 1                     St Louis, MO
## 2 Montreal, PQ / Chesterville, ON
## 3 Montreal, PQ / Chesterville, ON
## 4 Montreal, PQ / Chesterville, ON
## 5 Montreal, PQ / Chesterville, ON
## 6                     New York, NY
```

Exemplo de BI

O script `purchases.R` traz um exemplo de clusterização em vendas de telefone celular.

Produção de Relatórios e Apresentações

Vamos mostrar agora como é possível integrar todas as etapas de data science com o R, comunicando os resultados encontrados através de relatórios e apresentações.

Exemplo de Machine Learning

O arquivo `ml.Rmd` traz um exemplo de aplicação de algoritmos de ML a um problema de retenção de clientes em uma empresa de telefonia.