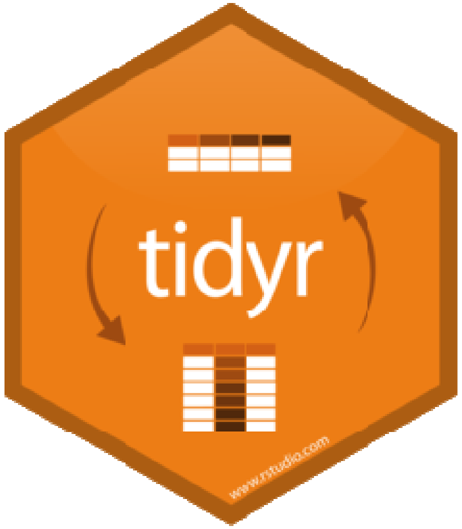
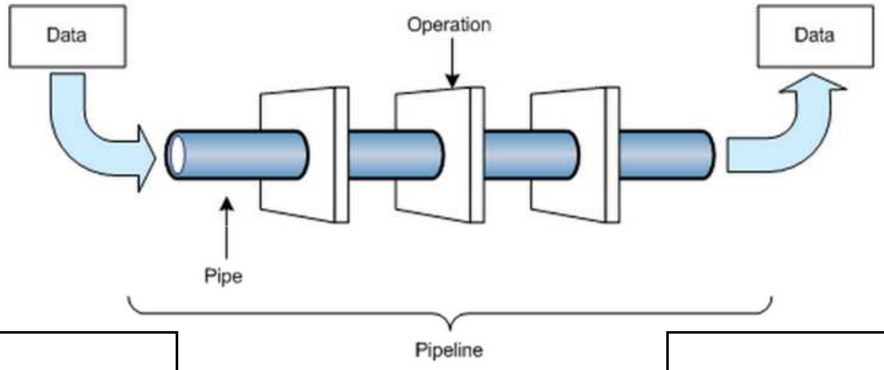
III. *Wrangling*

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL


tidyr
www.rstudio.com

III. *Wrangling*

- Utilização de ***pipes***



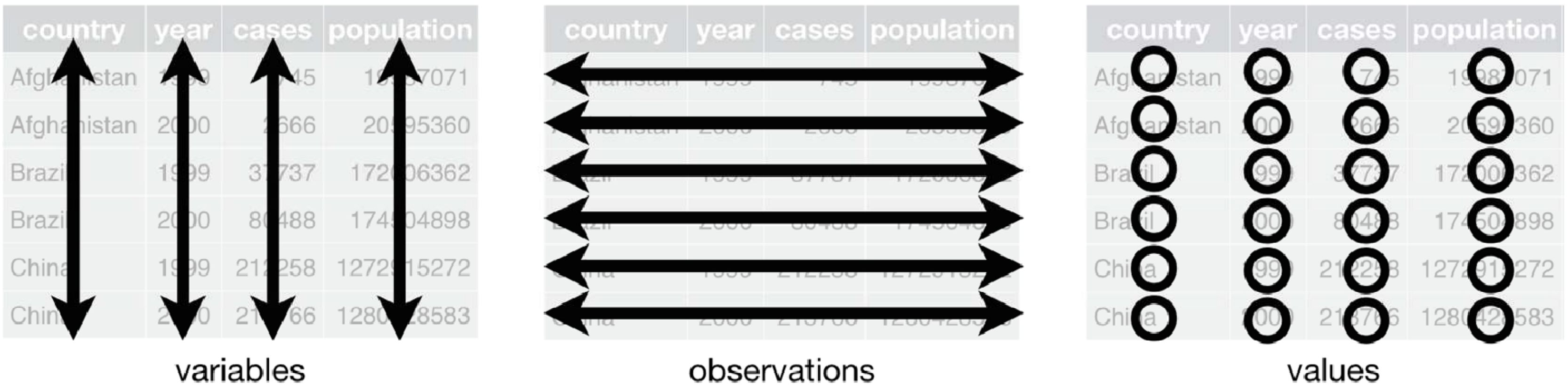
```
> temp[order(temp$Dia_Temp),]
    Dias Dia_Temp
4  sexta       12
1  terca       14
2 quarta       15
3 quinta       20
```

```
> temp %>% arrange(Dia_Temp)
    Dias Dia_Temp
1  sexta       12
2  terca       14
3 quarta       15
4 quinta       20
```

Ctrl + ⇧ Shift + M

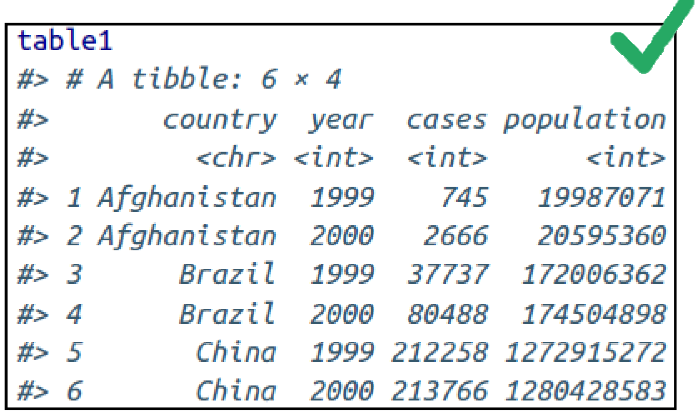
III. *Wrangling*

- Dados em formato ***tidy*** (reshape)

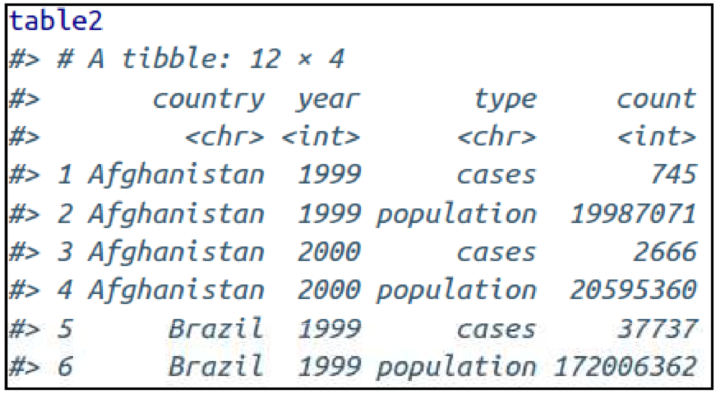


III. *Wrangling*

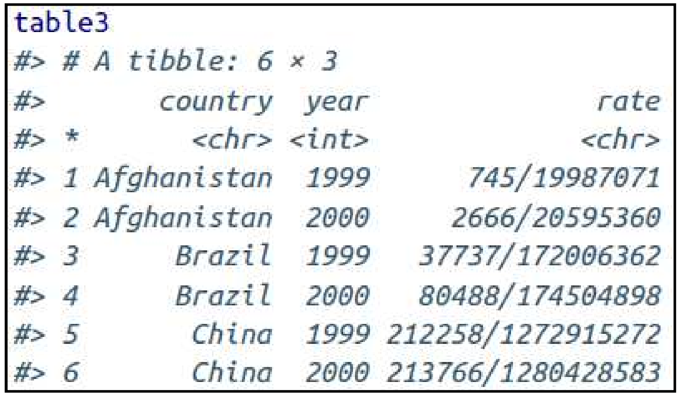
- Informação em formato *tidy*



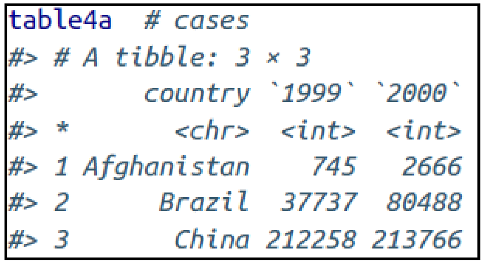
```
table1
#> # A tibble: 6 × 4
#>       country  year  cases population
#>         <chr> <int>  <int>      <int>
#> 1 Afghanistan  1999    745   19987071
#> 2 Afghanistan  2000   2666   20595360
#> 3      Brazil  1999  37737  172006362
#> 4      Brazil  2000  80488  174504898
#> 5       China  1999 212258 1272915272
#> 6       China  2000 213766 1280428583
```



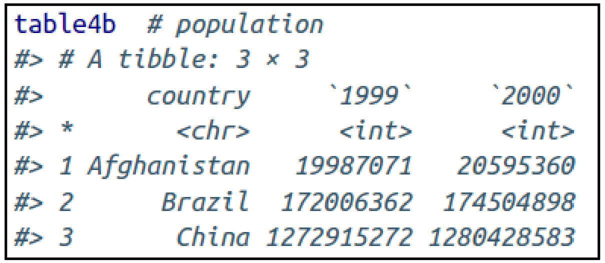
```
table2
#> # A tibble: 12 × 4
#>       country  year       type      count
#>         <chr> <int>      <chr>      <int>
#> 1 Afghanistan  1999      cases        745
#> 2 Afghanistan  1999 population   19987071
#> 3 Afghanistan  2000      cases       2666
#> 4 Afghanistan  2000 population   20595360
#> 5      Brazil  1999      cases      37737
#> 6      Brazil  1999 population  172006362
```



```
table3
#> # A tibble: 6 × 3
#>       country  year              rate
#> *       <chr> <int>             <chr>
#> 1 Afghanistan  1999      745/19987071
#> 2 Afghanistan  2000     2666/20595360
#> 3      Brazil  1999   37737/172006362
#> 4      Brazil  2000   80488/174504898
#> 5       China  1999 212258/1272915272
#> 6       China  2000 213766/1280428583
```



```
table4a  # cases
#> # A tibble: 3 × 3
#>       country `1999` `2000`
#> *       <chr>  <int>  <int>
#> 1 Afghanistan    745   2666
#> 2      Brazil  37737  80488
#> 3       China 212258 213766
```



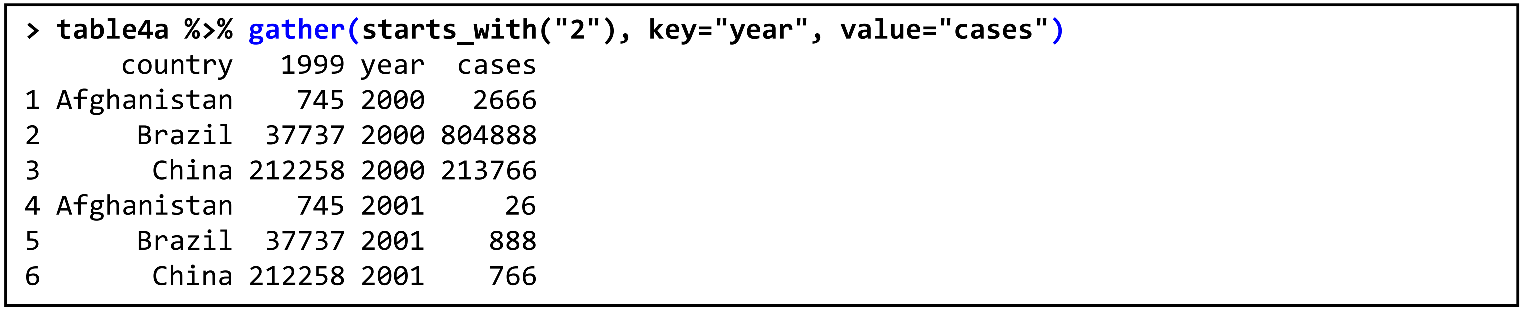
```
table4b  # population
#> # A tibble: 3 × 3
#>       country     `1999`     `2000`
#> *       <chr>      <int>      <int>
#> 1 Afghanistan   19987071   20595360
#> 2      Brazil  172006362  174504898
#> 3       China 1272915272 1280428583
```

III. *Wrangling*

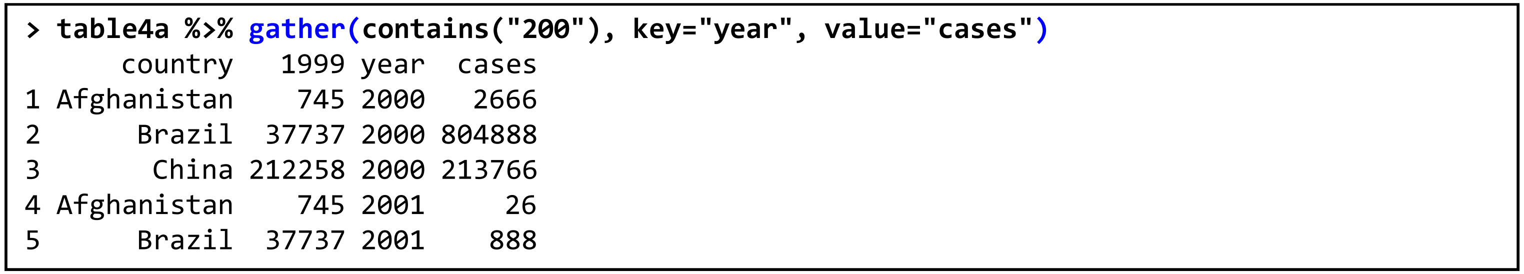
- Recolher numa variável informação dispersa por várias variáveis `gather`

```
> library(tidyr)
> table4a<-data.frame(country=c("Afghanistan","Brazil","China"), '1999'=c(745,37737,212258),
'2000'=c(2666,804888,213766), '2001'=c(26,888,766))
> table4a
      country   1999   2000 2001
1 Afghanistan    745   2666   26
2      Brazil  37737 804888  888
3       China 212258 213766  766

> table4a %>% gather(`1999`,`2000`,"2001", key="year", value="cases")
      country year  cases
1 Afghanistan 1999    745
2      Brazil 1999  37737
3       China 1999 212258
4 Afghanistan 2000   2666
5      Brazil 2000 804888
6       China 2000 213766
7 Afghanistan 2001     26
8      Brazil 2001    888
9       China 2001    766
```

III. *Wrangling*



```
> table4a %>% gather(starts_with("2"), key="year", value="cases")
      country   1999 year   cases
1 Afghanistan    745 2000    2666
2      Brazil  37737 2000 804888
3       China 212258 2000 213766
4 Afghanistan    745 2001      26
5      Brazil  37737 2001     888
6       China 212258 2001     766
```



```
> table4a %>% gather(contains("200"), key="year", value="cases")
      country   1999 year   cases
1 Afghanistan    745 2000    2666
2      Brazil  37737 2000 804888
3       China 212258 2000 213766
4 Afghanistan    745 2001      26
5      Brazil  37737 2001     888
```

III. *Wrangling*

- Distribuir valores de uma variável por várias colunas **spread**

```
> library(tidyr)
> table2<-
data.frame(country=c("Afghanistan","Afghanistan","Afghanistan","Afghanistan","Brazil"),
year=c(1999,1999,2000,2000,1999), type=c("cases","population","cases","population","cases"),
count=c(745,19987071,2666,20595360,377737))

> table2
      country year       type    count
1 Afghanistan 1999      cases      745
2 Afghanistan 1999 population 19987071
3 Afghanistan 2000      cases     2666
4 Afghanistan 2000 population 20595360
5      Brazil 1999      cases   377737

> table2 %>% spread(key = type, value = count)
      country year  cases population
1 Afghanistan 1999    745   19987071
2 Afghanistan 2000   2666   20595360
3      Brazil 1999 377737         NA
```

III. *Wrangling*

- Separar dados de uma coluna em múltiplas colunas `separate`

```
>table3<-data.frame(country=c("Afghanistan","Afghanistan","Brazil","Brazil","China","China"),
year=c(1999,2000,1999,2000,1999,2000),rate=c("45/19987071","2666/20595360","37737/172006362","8
0488/174504898","212258/1272915272","213766/1280428583"))
> table3
      country year              rate
1 Afghanistan 1999       45/19987071
2 Afghanistan 2000     2666/20595360
3      Brazil 1999   37737/172006362
4      Brazil 2000   80488/174504898
5       China 1999 212258/1272915272
6       China 2000 213766/1280428583

> table3 %>% separate(rate, into = c("cases","population"), sep = "/", convert = T)
      country year  cases population
1 Afghanistan 1999     45   19987071
2 Afghanistan 2000   2666   20595360
3      Brazil 1999  37737  172006362
4      Brazil 2000  80488  174504898
5       China 1999 212258 1272915272
6       China 2000 213766 1280428583
```

III. *Wrangling*

- Separar dados de uma coluna em múltiplas colunas `unite`

```
> table3<- data.frame( country=c("Afghanistan","Afghanistan","Brazil","Brazil","China","China"),
                       century=c(19,20,19,20,19,20),year=c("99","00","99","00","99","00"),
rate=c("45/19987071","2666/20595360","37737/172006362","80488/174504898","212258/1272915272","213766/12804
28583"))
> table3
      country century year              rate
1 Afghanistan      19   99       45/19987071
2 Afghanistan      20   00     2666/20595360
3      Brazil      19   99   37737/172006362
4      Brazil      20   00   80488/174504898
5       China      19   99 212258/1272915272
6       China      20   00 213766/1280428583

> table3 %>% unite(new, century, year, sep = "")
      country  new              rate
1 Afghanistan 1999       45/19987071
2 Afghanistan 2000     2666/20595360
3      Brazil 1999   37737/172006362
4      Brazil 2000   80488/174504898
5       China 1999 212258/1272915272
6       China 2000 213766/1280428583
```

Instituto Nacional de Estatística
Statistics Portugal

Exercícios2.pdf

Questão 1

dplyr
www.rstudio.com

III. Wrangling

- Data frame *mtcars*

```
> ? mtcars
[, 1]      mpg       Miles/(US) gallon
[, 2]      cyl       Number of cylinders
[, 3]      disp      Displacement (cu.in.)
[, 4]      hp        Gross horsepower
[, 5]      drat      Rear axle ratio
[, 6]      wt        Weight (1000 lbs)
[, 7]      qsec      1/4 mile time
[, 8]      vs        Engine (0 = V-shaped, 1 = straight)
[, 9]      am        Transmission (0 = automatic, 1 = manual)
[,10]      gear      Number of forward gears
[,11]      carb      Number of carburetors
```

```
> carros <- mtcars
> carros
                   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
```

III. *Wrangling*

- Selecionar variáveis `select()`

Quando pretendemos restringir a um conjunto de variáveis de interesse

```
> carros %>% select(mpg,disp:wt)
                     mpg  disp  hp drat    wt
Mazda RX4           21.0 160.0 110 3.90 2.620
Mazda RX4 Wag       21.0 160.0 110 3.90 2.875
Datsun 710          22.8 108.0  93 3.85 2.320…
```

```
> carros %>% select(starts_with("d"))
                     disp drat
Mazda RX4           160.0 3.90
Mazda RX4 Wag       160.0 3.90
…
```

```
> carros %>% select(hp, everything())
                     hp  mpg cyl  disp drat    wt  qsec vs am gear carb
Mazda RX4           110 21.0   6 160.0 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag       110 21.0   6 160.0 3.90 2.875 17.02  0  1    4    4…
```

III. *Wrangling*

- Alterar os nomes das variáveis `rename(), select()`

```
> carros %>% rename(consume_mpg = mpg)
                  consume_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
…
```

```
> carros %>% rename(consumo_mpg = mpg, cilindros = cyl)
                  consumo_mpg cilindros  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
…
```

```
> carros %>% select(consume_mpg = mpg, cyl)
                  consume_mpg cyl
Mazda RX4                21.0   6
Mazda RX4 Wag            21.0   6
…
```

III. *Wrangling*

- Filtrar os dados por condições `filter()`

Quando pretendemos formar subconjuntos baseados nos valores das variáveis

```
> carros %>% filter(mpg>21 , wt<2)
   mpg cyl disp  hp drat    wt  qsec vs  am gear carb
1 30.4   4 75.7  52 4.93 1.615 18.52  1  1    4    2
2 33.9   4 71.1  65 4.22 1.835 19.90  1  1    4    1
3 27.3   4 79.0  66 4.08 1.935 18.90  1  1    4    1 …
```

```
> carros %>% filter(mpg>21 & wt<2)
     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
  mpg cyl disp  hp drat    wt  qsec vs am gear carb
1 30.4   4 75.7  52 4.93 1.615 18.52  1  1    4    2
2 33.9   4 71.1  65 4.22 1.835 19.90  1  1    4    1
3 27.3   4 79.0  66 4.08 1.935 18.90  1  1    4    1 …
```

```
> carros %>% filter(mpg>21 | wt<2)
     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
1   22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
2   21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
3   24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
```

III. *Wrangling*

- Filtrar os dados com base na posição `slice() / filter ()`

```
> carros %>% slice(1)
    mpg cyl disp  hp drat   wt  qsec vs am gear carb
 1  21   6  160 110  3.9 2.62 16.46  0  1    4    4
```

```
> carros %>% slice(10:n())
 mpg cyl  disp  hp drat    wt  qsec vs am gear carb
1  19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
2  17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
3  16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
4  17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
…
```

```
> carros %>% filter(between(row_number(),10,n()))
    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
1  19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
2  17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
3  16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
4  17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
…
```

III. *Wrangling*

- Filtrar os dados duplicados `distinct()`

```
> carros %>% distinct()
    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
1  21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
2  21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
3  22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
4  21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
…
```

```
> carros %>% distinct(cyl)
  cyl
1   6
2   4
3   8
…
```

```
> carros %>% distinct(cyl, .keep_all = T)
   mpg cyl disp  hp drat   wt  qsec vs am gear carb
1 21.0   6  160 110 3.90 2.62 16.46  0  1    4    4
2 22.8   4  108  93 3.85 2.32 18.61  1  1    4    1
3 18.7   8  360 175 3.15 3.44 17.02  0  0    3    2
…
```

III. *Wrangling*

- Retirar uma amostra dos dados `sample_n/sample_frac/top_n`

```
>  carros %>% sample_n(5, replace = T)
   mpg cyl  disp  hp drat    wt  qsec vs am gear carb           modelo
1 17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4        Merc 280C
2 19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2 Pontiac Firebird
3 18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1          Valiant
4 21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4        Mazda RX4
5 21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1   Hornet 4 Drive
```

```
> carros %>% sample_frac(0.2, replace = T)
   mpg cyl  disp  hp drat    wt  qsec vs am gear carb             modelo
1 15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2        AMC Javelin
2 10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4 Cadillac Fleetwood
3 16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3         Merc 450SE
4 15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8      Maserati Bora
5 19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4           Merc 280
6 19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6       Ferrari Dino
```

```
>   carros %>% top_n(5, disp)
   mpg cyl disp  hp drat    wt  qsec vs am gear carb              modelo
1 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2   Hornet Sportabout
2 14.3   8  360 245 3.21 3.570 15.84  0  0    3    4          Duster 360
3 10.4   8  472 205 2.93 5.250 17.98  0  0    3    4  Cadillac Fleetwood
4 10.4   8  460 215 3.00 5.424 17.82  0  0    3    4 Lincoln Continental
5 14.7   8  440 230 3.23 5.345 17.42  0  0    3    4   Chrysler Imperial
6 19.2   8  400 175 3.08 3.845 17.05  0  0    3    2    Pontiac Firebird
```

III. *Wrangling*

- Organizar informação `arrange()`

Reorganizar os dados por uma ou mais variáveis

```
>  carros %>% arrange(mpg)
    mpg cyl  disp  hp drat    wt  qsec vs am gear carb              modelo
1  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4  Cadillac Fleetwood
2  10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4 Lincoln Continental
3  13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4          Camaro Z28
4  14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4          Duster 360
5  14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4   Chrysler Imperial
6  15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8       Maserati Bora
7  15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3          Merc 450SLC…
```

```
> carros %>% arrange(cyl, desc(mpg))
    mpg cyl  disp  hp drat    wt  qsec vs am gear carb              modelo
1  33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1      Toyota Corolla
2  32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1            Fiat 128
3  30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2         Honda Civic
4  30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2        Lotus Europa
5  27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1           Fiat X1-9
6  26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2       Porsche 914-2
…
```

III. *Wrangling*

- Criar ou atualizar variáveis com informação de variáveis existentes `mutate/transmute`

```
> carros %>% mutate(l100 = (100*3.785411784)/(1.609344*mpg))
   mpg cyl  disp  hp drat    wt  qsec vs am gear carb              modelo      l100
1  21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4           Mazda RX4 11.200694
2  21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4       Mazda RX4 Wag 11.200694
3  22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1          Datsun 710 10.316429
4  21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1      Hornet 4 Drive 10.991336
5  18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2   Hornet Sportabout 12.578320
…
```

```
> carros %>% transmute(l100 = (100*3.785411784)/(1.609344*mpg))
       l100
1  11.200694
2  11.200694
…
```

```
> carros %>% mutate(consumo=cut(l100, breaks = c(0,10,15,Inf), labels=c("baixo","médio","alto")))

   mpg cyl  disp  hp drat    wt  qsec vs am gear carb              modelo      l100 consumo
1  21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4           Mazda RX4 11.200694   médio
2  21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4       Mazda RX4 Wag 11.200694   médio
…
```

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

III. *Wrangling*

- Sintetizar informação de forma agregada `summarise()`

  Centrais: mean(), median()

  Distribuição: sd(), IQR(), mad()

  Intervalos: min(), max(), quantile()

  Posições: first(), last(), nth(),

  Contagens: n(), n_distinct()

  Lógica: any(), all()

```
> carros %>% summarise(média = mean(mpg))
     média
1 20.09062
```

```
> carros %>% summarise(num_carros = n())
  num_carros
1         32
```

```
> carros %>% summarise(desviopadrao = sd(mpg, na.rm = T))
  desviopadrao
1     6.026948
```

III. *Wrangling*

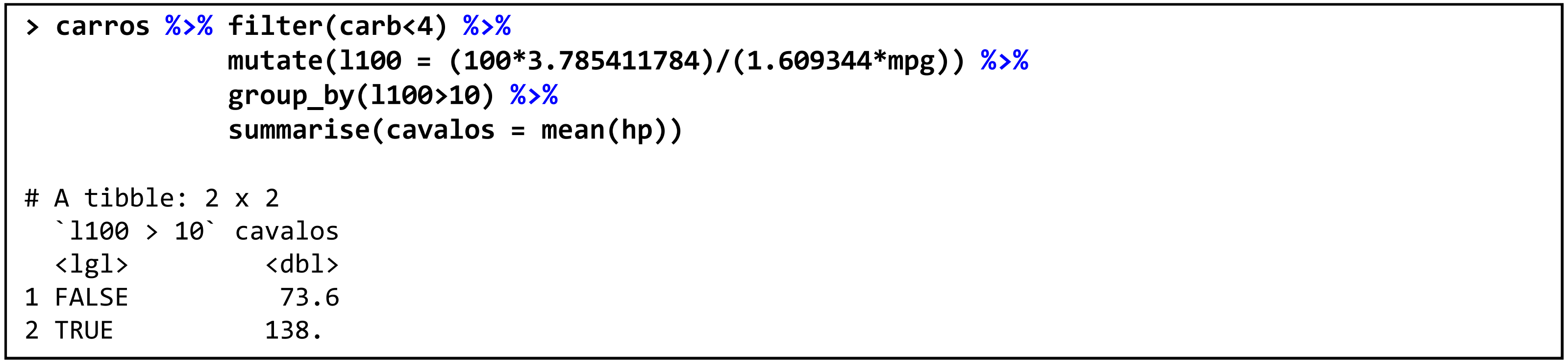
- Fazer cálculos agrupados por determinados critérios `group_by()`

```
> carros %>% group_by(gear) %>% summarise(média = mean(mpg))
# A tibble: 3 x 2
   gear média
  <dbl> <dbl>
1     3  16.1
2     4  24.5
3     5  21.4
```

```
> carros %>% group_by(gear,cyl) %>% summarise(média = mean(mpg))
# A tibble: 8 x 3
# Groups:   gear [3]
   gear   cyl média
  <dbl> <dbl> <dbl>
1     3     4  21.5
2     3     6  19.8
3     3     8  15.0
4     4     4  26.9
5     4     6  19.8
6     5     4  28.2
7     5     6  19.7
8     5     8  15.4
```

III. *Wrangling*

- Combinando múltiplas operações com *pipes*



```
> carros %>% filter(carb<4) %>%
             mutate(l100 = (100*3.785411784)/(1.609344*mpg)) %>%
             group_by(l100>10) %>%
             summarise(cavalos = mean(hp))

# A tibble: 2 x 2
  `l100 > 10` cavalos
  <lgl>         <dbl>
1 FALSE          73.6
2 TRUE          138.
```

```
Quais os carros mais rápidos com velocidades manuais para os diferentes numero de cilindros?
```

```
carros %>% filter(am==1) %>%
           group_by(cyl) %>%
           top_n(1, qsec)
```

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

Exercícios2.pdf

Questão 2

III. *Wrangling*

- Combinar informação de dois *data sets.* *left_join, right_join, inner_join…*

```
left_join(A,B)
A %>% left_join(B)
A %>% left_join(B , by="chave")
A %>% left_join(B , by=c("chave1" = "chave2")
```

Função	descrição
`left_join(A,B)`	Mantém A e correspondentes B se não existir em B fica com ***NA***
`right_join(A,B)`	Mantém B e correspondentes A se não existir em A fica com ***NA***
`inner_join(A,B)`	Mantém tudo que existe **simultaneamente** em A e B. O resto é eliminado
`full_join(A,B)`	Mantém tudo de A e B. Caso não exista correspondencia fica *NA*
`semi_join(A,B)`	Mantém A que existam em B. As restantes A são eliminadas.
`anti_join(A,B)`	Mantém A que **NÃO** existam em B.
`nested_join(A,B)`	Associa a cada A as observações correspondentes B (subtabela)

III. *Wrangling*

```
> carros
   mpg cyl  disp  hp drat    wt  qsec vs am gear carb            modelo
1 21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4         Mazda RX4
2 21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4     Mazda RX4 Wag
3 22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1        Datsun 710
4 21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1    Hornet 4 Drive
5 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2 Hornet Sportabout
6 18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1           Valiant…
> carros2
               modelo      l100
1           Mazda RX4 11.200694
2       Mazda RX4 Wag 11.200694
3          Datsun 710 10.316429
4      Hornet 4 Drive 10.991336
5   Hornet Sportabout 12.578320
6             Valiant 12.995281
…
> carros %>% left_join(carros2)

Joining, by = "modelo"
   mpg cyl  disp  hp drat    wt  qsec vs am gear carb            modelo      l100
1 21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4         Mazda RX4 11.200694
2 21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4     Mazda RX4 Wag 11.200694
3 22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1        Datsun 710 10.316429
4 21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1    Hornet 4 Drive 10.991336
5 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2 Hornet Sportabout 12.578320
…
```

III. *Wrangling*

III. *Wrangling*

- Criar ou alterar variáveis de texto

```
> carros %>% mutate(modelo2=str_sub(modelo,1,3))
    mpg cyl  disp  hp drat    wt  qsec vs     am gear carb             modelo modelo2
1  21.0   6 160.0 110 3.90 2.620 16.46  0 Manual    4    4          Mazda RX4     Maz
2  21.0   6 160.0 110 3.90 2.875 17.02  0 Manual    4    4      Mazda RX4 Wag     Maz
3  22.8   4 108.0  93 3.85 2.320 18.61  1 Manual    4    1         Datsun 710     Dat
…
```

```
> carros %>% mutate(modelo=str_to_upper(modelo))
    mpg cyl  disp  hp drat    wt  qsec vs     am gear carb             modelo
1  21.0   6 160.0 110 3.90 2.620 16.46  0 Manual    4    4          MAZDA RX4
2  21.0   6 160.0 110 3.90 2.875 17.02  0 Manual    4    4      MAZDA RX4 WAG
…
```

```
> carros %>% mutate(modelo=str_replace(modelo,"Merc", "MERCEDES"))
    mpg cyl  disp  hp drat    wt  qsec vs     am gear carb             modelo
…
7  14.3   8 360.0 245 3.21 3.570 15.84  0   AUTO    3    4         Duster 360
8  24.4   4 146.7  62 3.69 3.190 20.00  1   AUTO    4    2      MERCEDES 240D
9  22.8   4 140.8  95 3.92 3.150 22.90  1   AUTO    4    2       MERCEDES 230
…
```

III. *Wrangling*

- Criar ou alterar variáveis de texto

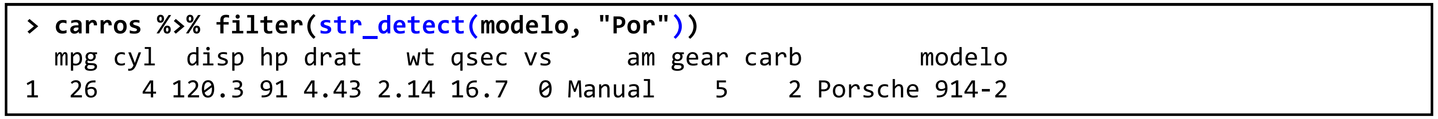
```
> carros %>% mutate(marca=word(modelo,1))
    mpg cyl  disp  hp drat    wt  qsec vs     am gear carb    marca
1  21.0   6 160.0 110 3.90 2.620 16.46  0 Manual    4    4    Mazda
2  21.0   6 160.0 110 3.90 2.875 17.02  0 Manual    4    4    Mazda
3  22.8   4 108.0  93 3.85 2.320 18.61  1 Manual    4    1   Datsun
…
```

```
> carros %>% mutate(comp=str_length(modelo))
 mpg cyl  disp  hp drat    wt  qsec vs am gear carb              modelo    comp
1  21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4           Mazda RX4     9
2  21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4       Mazda RX4 Wag    13
3  22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1          Datsun 710    10
```

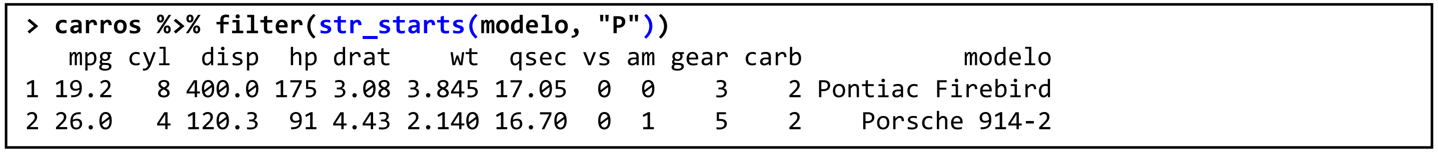
```
carros %>% mutate(Numeros=str_count(modelo, pattern = "\\d"))
 mpg cyl  disp  hp drat    wt  qsec vs am gear carb              modelo Numeros
1  21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4           Mazda RX4       1
2  21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4       Mazda RX4 Wag       1
3  22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1          Datsun 710       3
```

III. *Wrangling*

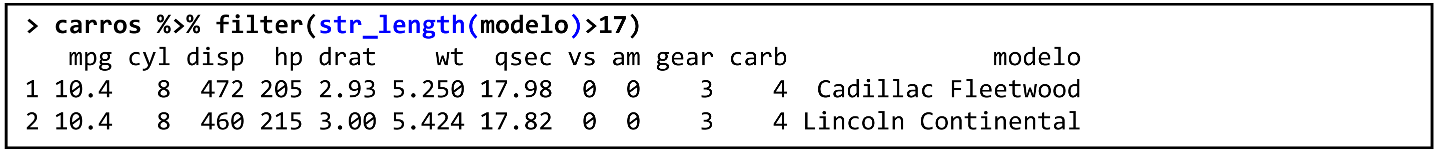
- Filtrar dados por uma variável de texto



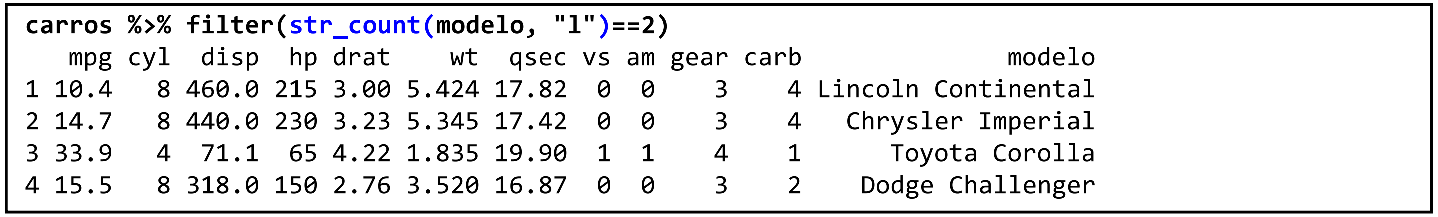
```
> carros %>% filter(str_detect(modelo, "Por"))
  mpg cyl  disp hp drat   wt qsec vs     am gear carb       modelo
1  26   4 120.3 91 4.43 2.14 16.7  0 Manual    5    2 Porsche 914-2
```



```
> carros %>% filter(str_starts(modelo, "P"))
   mpg cyl  disp  hp drat    wt  qsec vs am gear carb           modelo
1 19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2 Pontiac Firebird
2 26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2    Porsche 914-2
```



```
> carros %>% filter(str_length(modelo)>17)
   mpg cyl disp  hp drat    wt  qsec vs am gear carb              modelo
1 10.4   8  472 205 2.93 5.250 17.98  0  0    3    4  Cadillac Fleetwood
2 10.4   8  460 215 3.00 5.424 17.82  0  0    3    4 Lincoln Continental
```



```
carros %>% filter(str_count(modelo, "l")==2)
   mpg cyl  disp  hp drat    wt  qsec vs am gear carb              modelo
1 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4 Lincoln Continental
2 14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4   Chrysler Imperial
3 33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1      Toyota Corolla
4 15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2    Dodge Challenger
```

Instituto Nacional de Estatística
Statistics Portugal

III. *Wrangling*

- Filtrar dados por uma variável de texto (Expressões Regulares)

```
> carros %>% filter(str_detect(modelo, pattern = "^P"))
  mpg cyl  disp  hp drat    wt  qsec vs     am gear carb           modelo
1 19.2   8 400.0 175 3.08 3.845 17.05  0   AUTO    3    2 Pontiac Firebird
2 26.0   4 120.3  91 4.43 2.140 16.70  0 Manual    5    2    Porsche 914-2
```

```
> carros %>% filter(str_detect(modelo, pattern = "c$"))
 mpg cyl disp hp drat    wt  qsec vs     am gear carb      modelo
1 30.4   4 75.7 52 4.93 1.615 18.52  1 Manual    4    2 Honda Civic
```

```
> carros %>% filter(str_detect(modelo, pattern = "\\d"))
  mpg cyl  disp  hp drat    wt  qsec vs     am gear carb         modelo
1   21.0   6 160.0 110 3.90 2.620 16.46  0 Manual    4    4      Mazda RX4
2   21.0   6 160.0 110 3.90 2.875 17.02  0 Manual    4    4  Mazda RX4 Wag
3   22.8   4 108.0  93 3.85 2.320 18.61  1 Manual    4    1     Datsun 710
```

```
> carros %>% filter(str_detect(modelo, pattern= "^[A-Za-z]+[[:space:]]+\\d{3}$"))
```

```
 mpg cyl  disp  hp drat   wt  qsec vs am gear carb     modelo
1 22.8   4 108.0  93 3.85 2.32 18.61  1  1    4    1 Datsun 710
2 14.3   8 360.0 245 3.21 3.57 15.84  0  0    3    4 Duster 360
3 22.8   4 140.8  95 3.92 3.15 22.90  1  0    4    2   Merc 230
4 19.2   6 167.6 123 3.92 3.44 18.30  1  0    4    4   Merc 280
5 32.4   4  78.7  66 4.08 2.20 19.47  1  1    4    1   Fiat 128
```

Exercícios2.pdf

Questões 3 e 4

III. *Wrangling*

III. *Wrangling*

- Extrair informação de variáveis do tipo ***date*** com pacote *lubridate*

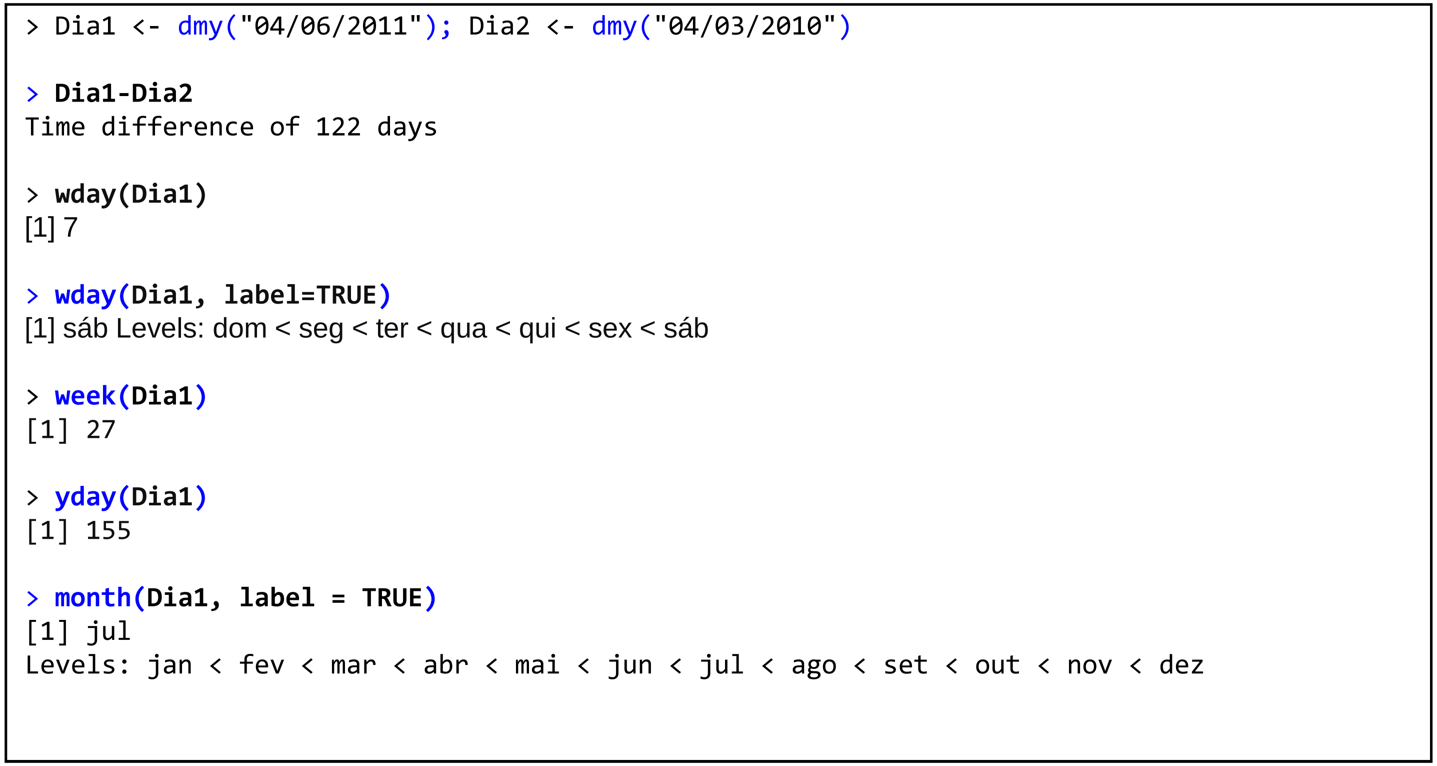
```
> library(lubridate)
> mtr
# Source:   lazy query [?? x 3]
# Database: OraConnection
   ID     DATA_COMPLETA    CONSUMO
   <chr> <chr>              <dbl>
 1 3204  2016-02-13 05:30     676
 2 3204  2016-02-13 05:45     896
 3 3204  2016-02-13 06:00     676
 4 3204  2016-02-13 06:15     360
# ... with more rows
```

Converter variável *text* em tipo *date*

```
> ymd("20110604")

> mdy("06-04-2011")

> dmy("04/06/2011")

> ymd_hms("2011-06-04 12:00:00")

> ymd_hm("2011-08-10 14:00", tz = "Pacific/Auckland")
[1] "2011-08-10 14:00:00 NZST"
```

III. *Wrangling*

- Operações possíveis com variáveis do tipo *date*



```
> Dia1 <- dmy("04/06/2011"); Dia2 <- dmy("04/03/2010")

> Dia1-Dia2
Time difference of 122 days

> wday(Dia1)
[1] 7

> wday(Dia1, label=TRUE)
[1] sáb Levels: dom < seg < ter < qua < qui < sex < sáb

> week(Dia1)
[1] 27

> yday(Dia1)
[1] 155

> month(Dia1, label = TRUE)
[1] jul
Levels: jan < fev < mar < abr < mai < jun < jul < ago < set < out < nov < dez
```

III. *Wrangling*

```
> library(lubridate)
> mtr
# Source:   lazy query [?? x 3]
# Database: OraConnection
   ID    DATA_COMPLETA    CONSUMO
   <chr> <chr>              <dbl>
 1 3204  2016-02-13 05:30     676
 2 3204  2016-02-13 05:45     896
 3 3204  2016-02-13 06:00     676
 4 3204  2016-02-13 06:15     360
 5 3204  2016-02-13 06:30     576
 6 3204  2016-02-13 06:45    2536
 7 3204  2016-02-13 07:00    6612
 8 3204  2016-02-13 07:15    3024
 9 3204  2016-02-13 07:30    3108
10 3204  2016-02-13 07:45    3428
# ... with more rows

> mtr<-mtr %>% mutate(DATA = ymd_hm(DATA_COMPLETA)) %>%
             mutate(Dia_da_semana = wday(DATA)) %>%
             mutate(FDS = (Dia_da_semana==1 | Dia_da_semana==7))
```