Explorando dados no R

Disciplina de Ciência de Dados 2025.2 - UESC

Gabriel Rodrigues

2025-09-04

Importando a base de dados

```
# Importar dataset
df <- read.csv("gapminder_error.csv", header=TRUE)

# Checar coisas básicas
## dimensões da tabela
dim(df)
```

```
[1] 201   5
```

```
## colunas
colnames(df)
```

```
[1] "Country"        "LifeExpectancy" "FertilityRate"  "Population"
[5] "Region"
```

```
## começo da tabela
head(df)
```

```
              Country LifeExpectancy FertilityRate Population
1         Afghanistan             51          7.81   19701940
2             Albania           74.2          2.47    3121965
3             Algeria           73.2          2.63   31183658
4              Angola           52.6          6.88   15058638
5 Antigua and Barbuda           73.9          2.32      77648
```

```
6           Argentina           74.3          2.54   37057453
                     Region
1                 South Asia
2      Europe & Central Asia
3 Middle East & North Africa
4         Sub-Saharan Africa
5                    America
6                    America
```

Resumo da tabela

```
summary(df)
```

```
   Country          LifeExpectancy     FertilityRate    Population
 Length:201         Length:201         Min.   :0.880   Length:201
 Class :character   Class :character   1st Qu.:1.770   Class :character
 Mode  :character   Mode  :character   Median :2.800   Mode  :character
                                       Mean   :3.298
                                       3rd Qu.:4.540
                                       Max.   :7.810

    Region
 Length:201
 Class :character
 Mode  :character
```

Esses números estão corretos? Classes

```
class(df)
```

```
[1] "data.frame"
```

```
class(df$LifeExpectancy)
```

```
[1] "character"
```

```
class(df$Population)
```

```
[1] "character"
```

```
## Colunas que deveriam ser númericas não estão numéricas

# transformar em numérico - remover variáveis não-numéricas
df$LifeExpectancy <- as.numeric(df$LifeExpectancy)
```

```
Warning: NAs introduced by coercion
```

```
## resumir
summary(df)
```

```
   Country           LifeExpectancy   FertilityRate    Population
 Length:201         Min.   : 45.70   Min.   :0.880   Length:201
 Class :character   1st Qu.: 61.40   1st Qu.:1.770   Class :character
 Mode  :character   Median : 71.45   Median :2.800   Mode  :character
                    Mean   : 68.91   Mean   :3.298
                    3rd Qu.: 75.92   3rd Qu.:4.540
                    Max.   :142.10   Max.   :7.810
                    NA's   :1
    Region
 Length:201
 Class :character
 Mode  :character
```

Transformar novamente

```
df$Population <- as.numeric(df$Population)
```

```
Warning: NAs introduced by coercion
```

```r
summary(df)
```

```
   Country          LifeExpectancy   FertilityRate     Population
 Length:201         Min.   : 45.70   Min.   :0.880   Min.   :1.840e+02
 Class :character   1st Qu.: 61.40   1st Qu.:1.770   1st Qu.:1.155e+06
 Mode  :character   Median : 71.45   Median :2.800   Median :5.380e+06
                    Mean   : 68.91   Mean   :3.298   Mean   :3.026e+07
                    3rd Qu.: 75.92   3rd Qu.:4.540   3rd Qu.:1.879e+07
                    Max.   :142.10   Max.   :7.810   Max.   :1.270e+09
                    NA's   :1                        NA's   :1
    Region
 Length:201
 Class :character
 Mode  :character
```

Checar itens repetidos

```r
# Checar itens repetidos
duplicated(df)
```

```
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
[169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
[181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
[193]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```r
## Quantidade de duplicados
table(duplicated(df))
```

```
FALSE  TRUE
  191    10
```

```r
## Remover itens duplicados
df2 <- unique(df)
table(duplicated(df2))
```

```
FALSE
  191
```

Remover valores incorretos - Expectativa de vida

```r
summary(df2)
```

```
   Country           LifeExpectancy   FertilityRate     Population
 Length:191         Min.   : 45.70   Min.   :0.880   Min.   :1.840e+02
 Class :character   1st Qu.: 61.40   1st Qu.:1.790   1st Qu.:1.196e+06
 Mode  :character   Median : 71.45   Median :2.800   Median :5.599e+06
                    Mean   : 68.87   Mean   :3.298   Mean   :3.155e+07
                    3rd Qu.: 75.97   3rd Qu.:4.500   3rd Qu.:1.904e+07
                    Max.   :142.10   Max.   :7.810   Max.   :1.270e+09
                    NA's   :1                        NA's   :1
    Region
 Length:191
 Class :character
 Mode  :character
```

```r
## Expectativa de vida muito acima do normal
df2[df2$LifeExpectancy == 142.10, ]
```

```
     Country LifeExpectancy FertilityRate Population     Region
NA      <NA>             NA            NA         NA       <NA>
131 Pakistan          142.1          4.67  138250487 South Asia
```

```r
## Checar se existem outras
df2[df2$LifeExpectancy >= 100, ]
```

```
     Country LifeExpectancy FertilityRate Population     Region
NA      <NA>             NA            NA         NA       <NA>
131 Pakistan          142.1          4.67  138250487 South Asia
```

```r
## Remover outlier
df2 <- df2[-131,]
## Checar com summary()
summary(df2)
```

```
   Country           LifeExpectancy  FertilityRate     Population
 Length:190         Min.   :45.70   Min.   :0.880   Min.   :1.840e+02
 Class :character   1st Qu.:61.40   1st Qu.:1.785   1st Qu.:1.185e+06
 Mode  :character   Median :71.40   Median :2.785   Median :5.386e+06
                    Mean   :68.48   Mean   :3.291   Mean   :3.099e+07
                    3rd Qu.:75.90   3rd Qu.:4.452   3rd Qu.:1.882e+07
                    Max.   :82.66   Max.   :7.810   Max.   :1.270e+09
                    NA's   :1                       NA's   :1
    Region
 Length:190
 Class :character
 Mode  :character
```

Remover valores incorretos - População

```r
df2[df2$Population == 1.840e+02, ]
```

```
    Country LifeExpectancy FertilityRate Population  Region
NA      <NA>             NA            NA         NA    <NA>
182 Uruguay           74.6          2.27        184 America
```

```r
df2[df2$Country == "Uruguay", ]
```

```
    Country LifeExpectancy FertilityRate Population  Region
182 Uruguay           74.6          2.27        184 America
```

```r
df2[df2$Population <= 1e+03, ]
```

```
    Country LifeExpectancy FertilityRate Population  Region
NA      <NA>             NA            NA         NA    <NA>
182 Uruguay           74.6          2.27        184 America
```

```r
## Remover outlier - removendo com filtro (RECOMENDADO: remover múltiplos itens)
df2 <- df2[df2$Population >= 1000, ]
## Checar remoção
df2[df2$Country == "Uruguay", ]
```

```
   Country LifeExpectancy FertilityRate Population Region
NA    <NA>             NA            NA         NA   <NA>
```

```r
## Checar tabela
summary(df2)
```

```
   Country          LifeExpectancy  FertilityRate     Population
 Length:189         Min.   :45.70   Min.   :0.880   Min.   :5.617e+04
 Class :character   1st Qu.:61.25   1st Qu.:1.778   1st Qu.:1.218e+06
 Mode  :character   Median :71.30   Median :2.785   Median :5.599e+06
                    Mean   :68.42   Mean   :3.299   Mean   :3.115e+07
                    3rd Qu.:75.95   3rd Qu.:4.480   3rd Qu.:1.890e+07
                    Max.   :82.66   Max.   :7.810   Max.   :1.270e+09
                    NA's   :2       NA's   :1       NA's   :1
    Region
 Length:189
```

```
Class :character
Mode  :character
```