Neste notebook trarei dados a respeito da violência policial em alguns estados dos EUA, descobriremos se algumas variáveis estão correlacionadas e, se possível, criaremos um modelo de regressão linear. Serão mostrados também alguns erros comuns que ocorrem durante o processo e como resolvê-los.

```
dados = read.csv("deaths_arrests.csv")
summary(dados)
```

```
##    ï..State             City                PD
##  Length:1006        Length:1006        Length:1006
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Black.People.Killed.by.Police..1.1.2013.12.31.2019.
##  Min.   :   1.00
##  1st Qu.:   2.00
##  Median :   5.00
##  Mean   :  37.37
##  3rd Qu.:  10.00
##  Max.   :1957.00
##  NA's   :913
##  Hispanic.People.Killed.by.Police..1.1.2013.12.31.2019.
##  Min.   :   1.00
##  1st Qu.:   1.00
##  Median :   3.00
##  Mean   :  31.92
##  3rd Qu.:   8.50
##  Max.   :1338.00
##  NA's   :935
##  Native.American.People.Killed.by.Police..1.1.2013.12.31.2019.
##  Min.   :  1.00
##  1st Qu.:  1.00
##  Median :  2.00
##  Mean   : 12.42
##  3rd Qu.:  3.25
##  Max.   :111.00
##  NA's   :994
##  Asian.People.Killed.by.Police..1.1.2013.12.31.2019.
##  Min.   :  1.000
##  1st Qu.:  1.000
##  Median :  1.000
##  Mean   :  6.424
##  3rd Qu.:  2.000
##  Max.   :120.000
##  NA's   :973
##  Pacific.Islanders.Killed.by.Police..1.1.2013.12.31.2019.
##  Min.   : 1.000
##  1st Qu.: 1.000
##  Median : 1.000
##  Mean   : 8.222
##  3rd Qu.:10.000
##  Max.   :42.000
##  NA's   :997
##  White.People.Killed.by.Police..1.1.2013.12.31.2019.
##  Min.   :   1.00
##  1st Qu.:   2.00
##  Median :   4.00
##  Mean   :  49.15
##  3rd Qu.:   9.00
##  Max.   :3417.00
##  NA's   :915
##  Unknown.Race.People.Killed.by.Police..1.1.2013.12.31.2019.
##  Min.   :  1.0
##  1st Qu.:  1.0
##  Median :  2.0
##  Mean   : 18.1
##  3rd Qu.:  4.0
##  Max.   :641.0
##  NA's   :956
##  All.People.Killed.by.Police..1.1.2013.12.31.2019.    Total
##  Min.   :   0.00                                   Length:1006
##  1st Qu.:   8.00                                   Class :character
##  Median :  13.00                                   Mode  :character
##  Mean   : 113.28
##  3rd Qu.:  24.75
##  Max.   :7626.00
##  NA's   :904
##     Black              White           Amer..Indian          Asian
##  Length:1006        Length:1006        Length:1006        Length:1006
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    Hawaiian         Asian.Pacific.Islander    Other
##  Length:1006        Length:1006            Length:1006
##  Class :character   Class :character       Class :character
##  Mode  :character   Mode  :character       Mode  :character
##
##
##
##
##  Two.or.more.races    Hispanic         Black.White.Dissimilarity.Index..2010.
##  Length:1006        Length:1006        Min.   :17.24
##  Class :character   Class :character   1st Qu.:34.62
##  Mode  :character   Mode  :character   Median :49.84
##                                        Mean   :47.77
##                                        3rd Qu.:58.69
##                                        Max.   :82.48
##                                        NA's   :906
##  Murder.and.nonnegligent.manslaughter  Murder.Rate
##  Length:1006                          Min.   :  0.00
##  Class :character                     1st Qu.:  3.40
##  Mode  :character                     Median :  6.60
##                                       Mean   : 13.64
##                                       3rd Qu.: 15.72
##                                       Max.   :104.10
##                                       NA's   :904
##  Avg.Annual.Police.Homicide.Rate
##  Min.   : 0.000
##  1st Qu.: 3.300
##  Median : 4.650
##  Mean   : 5.021
##  3rd Qu.: 6.275
##  Max.   :17.900
##  NA's   :904
##  Avg.Annual.Police.Homicide.Rate.for.Black.People
##  Min.   : 0.00
##  1st Qu.: 5.20
##  Median : 9.55
##  Mean   :12.59
##  3rd Qu.:16.25
##  Max.   :71.50
##  NA's   :904
##  Avg.Annual.Police.Homicide.Rate.for.White.People
##  Min.   : 0.000
##  1st Qu.: 1.300
##  Median : 3.050
##  Mean   : 3.684
##  3rd Qu.: 4.950
##  Max.   :17.700
##  NA's   :904
##  Avg.Annual.Police.Homicide.Rate.for.Hispanic.People Black.White.Disparity
##  Min.   : 0.000                                      Length:1006
##  1st Qu.: 0.000                                      Class :character
##  Median : 3.300                                      Mode  :character
##  Mean   : 3.802
##  3rd Qu.: 5.600
##  Max.   :17.200
##  NA's   :904
##  Hispanic.White.Disparity Violent.crimes.2013..if.reported.by.agency.
##  Length:1006              Min.   :   113
##  Class :character         1st Qu.:  1431
##  Mode  :character         Median :  2552
##                           Mean   :  8997
##                           3rd Qu.:  5085
##                           Max.   :445359
##                           NA's   :907
##  Violent.crimes.2014..if.reported.by.agency.
##  Min.   :   120
##  1st Qu.:  1383
##  Median :  2683
##  Mean   :  9196
##  3rd Qu.:  5193
##  Max.   :441400
##  NA's   :910
##  Violent.crimes.2015..if.reported.by.agency.
##  Min.   :   144
##  1st Qu.:  1441
##  Median :  2766
##  Mean   :  9476
##  3rd Qu.:  5406
##  Max.   :459609
##  NA's   :909
##  Violent.crimes.2016..if.reported.by.agency.
##  Min.   :   152
##  1st Qu.:  1726
##  Median :  3206
##  Mean   : 10316
##  3rd Qu.:  5760
##  Max.   :490035
##  NA's   :911
##  Violent.crimes.2017..if.reported.by.agency.
##  Min.   :   169
##  1st Qu.:  1673
##  Median :  3220
##  Mean   : 10506
##  3rd Qu.:  5736
##  Max.   :483253
##  NA's   :914
##  Violent.crimes.2018..if.reported.by.agency.
##  Min.   :   160
##  1st Qu.:  1636
##  Median :  2953
##  Mean   :  9898
##  3rd Qu.:  5480
##  Max.   :480067
##  NA's   :909
##  Average.Violent.Crimes.Reported..2013.17. Violent.Crime.Rate
##  Min.   :   143                            Min.   : 0.700
##  1st Qu.:  1539                            1st Qu.: 4.500
##  Median :  2755                            Median : 7.100
##  Mean   :  9348                            Mean   : 7.635
##  3rd Qu.:  5335                            3rd Qu.:10.100
##  Max.   :466621                            Max.   :18.900
##  NA's   :905                               NA's   :905
##  X2013.Total.Arrests..UCR.Data. X2014.Total.Arrests X2015.Total.Arrests
##  Min.   :  2742                Min.   :  2500      Min.   :  2566
##  1st Qu.:  9634                1st Qu.:  9020      1st Qu.:  8954
##  Median : 17526                Median : 16248      Median : 15019
##  Mean   : 27958                Mean   : 25761      Mean   : 23460
##  3rd Qu.: 30561                3rd Qu.: 27521      3rd Qu.: 25723
##  Max.   :393809                Max.   :387727      Max.   :339470
##  NA's   :916                   NA's   :914         NA's   :917
##  X2016.Total.Arrests X2017.Total.Arrests X2018.Total.Arrests
##  Min.   :  2422      Min.   :  2405      Min.   :  2064
##  1st Qu.:  9106      1st Qu.:  7925      1st Qu.:  7836
##  Median : 13816      Median : 13247      Median : 13303
##  Mean   : 22174      Mean   : 21038      Mean   : 20182
##  3rd Qu.: 24121      3rd Qu.: 23632      3rd Qu.: 21462
##  Max.   :314864      Max.   :286225      Max.   :246773
##  NA's   :914         NA's   :913         NA's   :911
##  Estimated.Average.Arrests.per.Year Killings.by.Police.per.10k.Arrests
##  Length:1006                        Min.   : 0.00
##  Class :character                   1st Qu.: 6.70
##  Mode  :character                   Median : 9.50
##                                     Mean   :10.83
##                                     3rd Qu.:13.00
##                                     Max.   :39.70
##                                     NA's   :909
```

Podemos observar que os dados possuem muitas linhas completamente em branco. Além disso temos duas linhas com os totais, que não utilizaremos. Os dados relevantes para a análise vão somente até a linha 100. Vamos remover todas as outras. No resultado mostraremos apenas das linhas 85 até 100 e colunas 1 a 8, simplesmente para não exibirmos a tabela inteira, que é muito grande.

```
dados = dados[-c(101:1006), ]
dados[85:100, 1:8]
```

```
##           ï..State           City                                    PD
## 85      California       San Jose               San Jose Police Department
## 86      California      Santa Ana              Santa Ana Police Department
## 87         Arizona     Scottsdale             Scottsdale Police Department
## 88      Washington        Seattle                Seattle Police Department
## 89      Washington        Spokane                Spokane Police Department
## 90        Missouri      St. Louis St. Louis Metropolitan Police Department
## 91       Minnesota       St. Paul               St. Paul Police Department
## 92         Florida St. Petersburg         St. Petersburg Police Department
## 93      California       Stockton               Stockton Police Department
## 94         Florida          Tampa                  Tampa Police Department
## 95            Ohio         Toledo                 Toledo Police Department
## 96         Arizona         Tucson                 Tucson Police Department
## 97        Oklahoma          Tulsa                  Tulsa Police Department
## 98        Virginia Virginia Beach         Virginia Beach Police Department
## 99          Kansas        Wichita                Wichita Police Department
## 100 North Carolina  Winston-Salem          Winston-Salem Police Department
##     Black.People.Killed.by.Police..1.1.2013.12.31.2019.
## 85                                                    1
## 86                                                    1
## 87                                                    1
## 88                                                    4
## 89                                                    1
## 90                                                   36
## 91                                                    5
## 92                                                    2
## 93                                                    5
## 94                                                    4
## 95                                                    3
## 96                                                    2
## 97                                                    7
## 98                                                    3
## 99                                                    1
## 100                                                   3
##     Hispanic.People.Killed.by.Police..1.1.2013.12.31.2019.
## 85                                                      10
## 86                                                      10
## 87                                                       1
## 88                                                       2
## 89                                                      NA
## 90                                                      NA
## 91                                                       1
## 92                                                      NA
## 93                                                       5
## 94                                                      NA
## 95                                                      NA
## 96                                                      10
## 97                                                       4
## 98                                                      NA
## 99                                                       2
## 100                                                     NA
##     Native.American.People.Killed.by.Police..1.1.2013.12.31.2019.
## 85                                                             NA
## 86                                                             NA
## 87                                                             NA
## 88                                                              1
## 89                                                              2
## 90                                                             NA
## 91                                                              2
## 92                                                             NA
## 93                                                             NA
## 94                                                             NA
## 95                                                             NA
## 96                                                             NA
## 97                                                             NA
## 98                                                             NA
## 99                                                             NA
## 100                                                            NA
##     Asian.People.Killed.by.Police..1.1.2013.12.31.2019.
## 85                                                    1
## 86                                                    1
## 87                                                   NA
## 88                                                    2
## 89                                                   NA
## 90                                                   NA
## 91                                                    2
## 92                                                   NA
## 93                                                    2
## 94                                                    2
## 95                                                   NA
## 96                                                   NA
## 97                                                    1
## 98                                                   NA
## 99                                                   NA
## 100                                                  NA
##     Pacific.Islanders.Killed.by.Police..1.1.2013.12.31.2019.
## 85                                                         1
## 86                                                        NA
## 87                                                        NA
## 88                                                         1
## 89                                                        NA
## 90                                                        NA
## 91                                                        NA
## 92                                                        NA
## 93                                                         1
## 94                                                        NA
## 95                                                        NA
## 96                                                        NA
## 97                                                        NA
## 98                                                        NA
## 99                                                        NA
## 100                                                       NA
```

Agora iremos procurar dados faltantes nas coluna que utilizaremos no nosso modelo de regressão linear. Utilizaremos as colunas com o número de habitantes negros de uma cidade e o número de habitantes negros mortos pela polícia na mesma cidade.

```
summary(dados$Black.People.Killed.by.Police..1.1.2013.12.31.2019.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   2.000   5.000   8.341   9.500  56.000       9
```

Temos 9 dados faltantes na coluna escolhida. Podemos verificar em quais linhas estão esses dados, se quisermos.

```
dados[is.na(dados$Black.People.Killed.by.Police..1.1.2013.12.31.2019.), 1:4]
```

```
##      ï..State           City                             PD
## 14    Arizona       Chandler       Chandler Police Department
## 18 California    Chula Vista    Chula Vista Police Department
## 39     Hawaii       Honolulu       Honolulu Police Department
## 42 California         Irvine         Irvine Police Department
## 47      Texas         Laredo         Laredo Police Department
## 54      Texas        Lubbock        Lubbock Police Department
## 57    Arizona           Mesa           Mesa Police Department
## 74      Texas          Plano          Plano Police Department
## 82 California San Bernardino San Bernardino Police Department
##    Black.People.Killed.by.Police..1.1.2013.12.31.2019.
## 14                                                  NA
## 18                                                  NA
## 39                                                  NA
## 42                                                  NA
## 47                                                  NA
## 54                                                  NA
## 57                                                  NA
## 74                                                  NA
## 82                                                  NA
```

Uma opção comum nesse caso é substituir os dados faltantes pela mediana. Entretanto nesse caso em particular o que ocorreu é que, ao invés do valor zero ser colocado nas cidades que não registraram habitantes negros mortos pela polícia no período, deixaram o espaço em branco. Portanto vamos substituir os valores faltantes por zero.

```
dados[is.na(dados$Black.People.Killed.by.Police..1.1.2013.12.31.2019.), ]$Black.People.Killed.by.Police..1.1.201
3.12.31.2019. = 0
```

E agora verificamos novamente. Se não houverem valores em branco devemos receber como resposta uma tabela sem nenhuma linha.

```
dados[is.na(dados$Black.People.Killed.by.Police..1.1.2013.12.31.2019.), 1:4 ]
```

```
## [1] ï..State
## [2] City
## [3] PD
## [4] Black.People.Killed.by.Police..1.1.2013.12.31.2019.
## <0 linhas> (ou row.names de comprimento 0)
```

Repetiremos o mesmo processo com o número de habitantes negros.

```
dados[is.na(dados$Black), 1:4 ]
```

```
## [1] ï..State
## [2] City
## [3] PD
## [4] Black.People.Killed.by.Police..1.1.2013.12.31.2019.
## <0 linhas> (ou row.names de comprimento 0)
```

Não encontramos dados faltantes para o número de habitantes negros. Agora vamos visualizar os dados de outra maneira. Primeiramente precisaremos mudar o tipo de dado presente na coluna com o total de habitantes negros de cada cidade. Os dados são do tipo texto e precisamos passá-los para número.

```
strtoi(dados$Black)
```

```
##   [1]  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
##  [19]  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
##  [37]  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA 478  NA  NA  NA  NA  NA  NA  NA
##  [55]  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
##  [73]  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
##  [91]  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
```

O resultado acima não era esperado. Vemos aqui que estamos com vários valores vazios. Isso acontece pois valores que não são interpretados pelo R são retornados como um valor ausente. Especificamente nesse caso isso ocorreu pois os valores foram escritos com vírgulas. Vamos remover as vírgulas e tentar novamente.

```
library(stringr)
dados$Black = str_replace(dados$Black, ",", "")
```

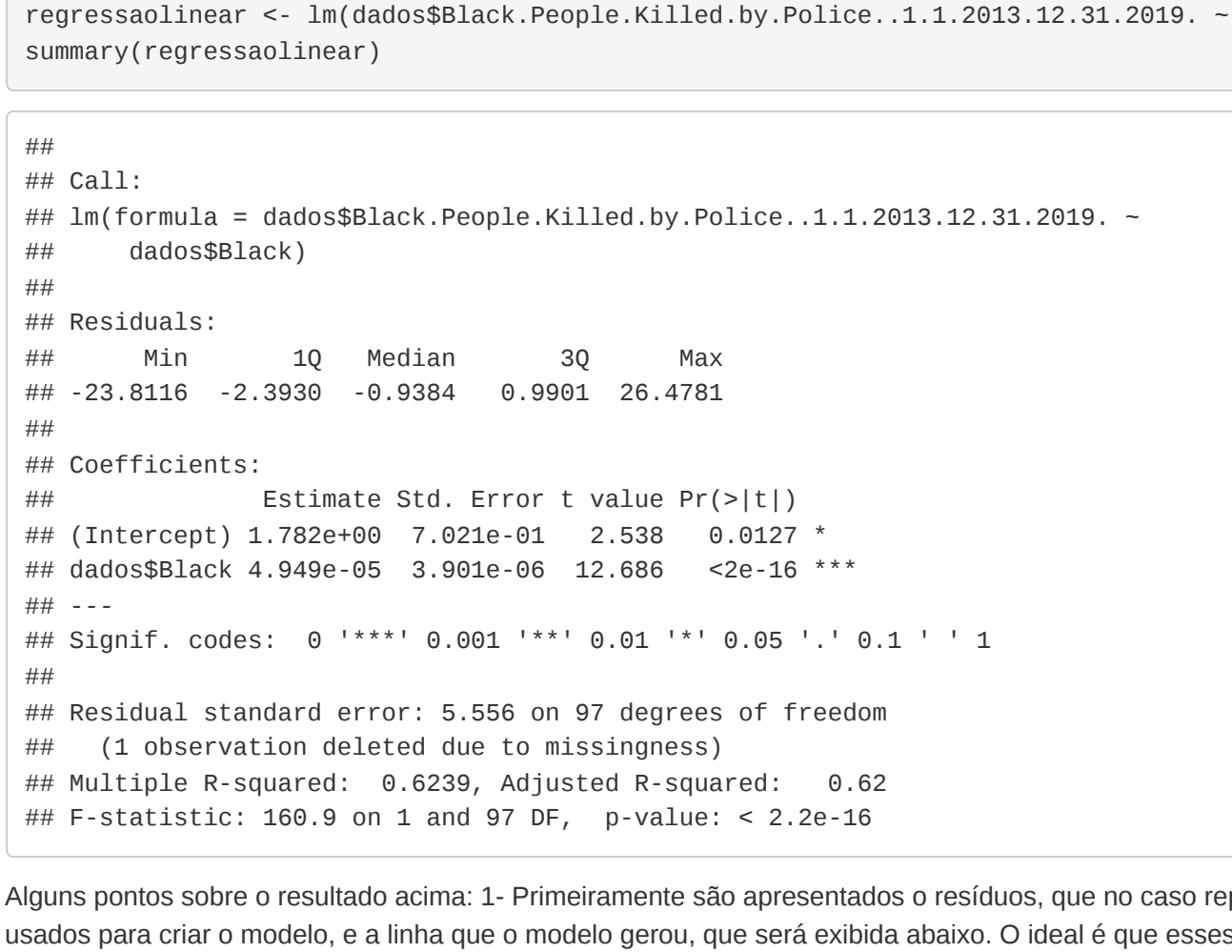
Agora sem as vírgulas podemos transformar os valores em inteiros.

```
dados$Black = strtoi(dados$Black)
dados$Black
```

```
##   [1]  14878   8209  15308  67087 224316  49003  60760  26677 392938 124542
##  [11] 155258 138073  97637  10580 252007  65204 872286   9972 132307 208208
##  [21]  24391 217694  11912 294159 301053  58388 586573  92285  18155  38514
##  [31] 136941   6743  37885  32164  12766 108233  12471   6066  26690 485956
##  [41] 223053   3494  25550 247516  59060 135916    478  62008  42336   9541
##  [51]  59925 347380 135138  18744  16507 408075  14101 169272  64993 233325
##  [61]  69971 204866     NA 138074 102452  41561 106637  85744  55128  63584
##  [71] 644287  86788  78847  19199  35462 115976   5990  19917  83346  64967
##  [81]  83365  29897  82497  46781  27508   3177   3484  47113   4643 156389
##  [91]  43620  57489  33507  83032  76820  23362  61230  83210  42676  78065
```

Agora finalmente podemos visualizar os dados em um gráfico.

```
plot(dados$Black, dados$Black.People.Killed.by.Police..1.1.2013.12.31.2019.)
```



E então criar o modelo de regressão linear.

```
regressaolinear <- lm(dados$Black.People.Killed.by.Police..1.1.2013.12.31.2019. ~ dados$Black)
summary(regressaolinear)
```

```
##
## Call:
## lm(formula = dados$Black.People.Killed.by.Police..1.1.2013.12.31.2019. ~
##     dados$Black)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.8116  -2.3930  -0.9384   0.9901  26.4781
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.782e+00  7.021e-01   2.538   0.0127 *
## dados$Black 4.949e-05  3.901e-06  12.686   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.556 on 97 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6239, Adjusted R-squared:   0.62
## F-statistic: 160.9 on 1 and 97 DF,  p-value: < 2.2e-16
```

Alguns pontos sobre o resultado acima: 1- Primeiramente são apresentados o resíduos, que no caso representam a distância entre os dados usados para criar o modelo, e a linha que o modelo gerou, que será exibida abaixo. O ideal é que esses resíduos tenham uma distribuição normal, logo os valores mínimo e máximo idealmente estão a uma mesma distancia de zero, assim como o primeiro e o terceiro quartis. Também é interessante que a mediana seja próxima de zero.

2- O coeficiente de determinação (r^2), nos diz o quanto a inclusão de uma variável independente reduz a variância observada na variável dependente. Em outras palavras, o valor encontrado de 0.60 indica que o número de habitantes negros de uma cidade “explica” 60% da variância encontrada no número de habitantes negros mortos pela polícia.

3- O valor de p encontrado de 0,00000000000000022 (2.2e-16) é um indicativo de que o resultado não se deve a aleatoriedade. É interessante que esse valor esteja abaixo de 0,05.

```
plot(dados$Black, dados$Black.People.Killed.by.Police..1.1.2013.12.31.2019., main = "Regressão Linear", xlab = "N
úmero de habitantes negros", ylab = "Número de habitantes negros mortos pela polícia")
abline(regressaolinear, col = "blue")
```


O objetivo aqui foi demonstrar o processo de criação de um modelo simples de regressão linear. Os dados, que são reais, são utilizados apenas com um fim didático. Não seria prudente elaborar grandes conclusões ou previsões sobre esses dados, o intuito é mostrar um pouco do código usado e possíveis problemas enfrentados durante o tratamento dos dados.