IC

Beatriz Vianna

June 24, 2018

1. Introdução

2. Base de dados

```
library(kernlab)
data (spam)
```

Spam é um conjunto de dados coletado no Hewlett-Packard Labs, que classifica 4601 e-mails como spam ou nonspam. Além desta variável classificatória há 57 variáveis numéricas indicando a frequência de certas palavras e caracteres em cada um destes e-mails.

Trabalharemos com esse data frame, com 4601 observações de e-mails e 58 variáveis.
As primeiras 48 variáveis contêm a frequência do nome da variável no e-mail. Se o nome da variável começa com "num" ela indica a frequência do número correspondente. As variáveis 49-54 indicam a frequência dos caracteres ';', '(', '[', '!', '$', e '#'. As variáveis 55 - 57 contêm a média, a mais longa e e o total de sequências de letras em caixa alta. A variável 58 indica o tipo de e-mail, sendo 2788 classificados como "nonspam" e 1813 classificados como "spam".

Abaixo as primeiras 5 observações da base:

```
##   make address  all num3d  our over remove internet order mail receive
## 1 0.00    0.64 0.64     0 0.32 0.00   0.00     0.00  0.00 0.00    0.00
## 2 0.21    0.28 0.50     0 0.14 0.28   0.21     0.07  0.00 0.94    0.21
## 3 0.06    0.00 0.71     0 1.23 0.19   0.19     0.12  0.64 0.25    0.38
## 4 0.00    0.00 0.00     0 0.63 0.00   0.31     0.63  0.31 0.63    0.31
## 5 0.00    0.00 0.00     0 0.63 0.00   0.31     0.63  0.31 0.63    0.31
##   will people report addresses free business email  you credit your font
## 1 0.64   0.00   0.00      0.00 0.32     0.00  1.29 1.93   0.00 0.96    0
## 2 0.79   0.65   0.21      0.14 0.14     0.07  0.28 3.47   0.00 1.59    0
## 3 0.45   0.12   0.00      1.75 0.06     0.06  1.03 1.36   0.32 0.51    0
## 4 0.31   0.31   0.00      0.00 0.31     0.00  0.00 3.18   0.00 0.31    0
## 5 0.31   0.31   0.00      0.00 0.31     0.00  0.00 3.18   0.00 0.31    0
##   num000 money hp hpl george num650 lab labs telnet num857 data num415
## 1   0.00  0.00  0   0      0      0   0    0      0      0    0      0
## 2   0.43  0.43  0   0      0      0   0    0      0      0    0      0
## 3   1.16  0.06  0   0      0      0   0    0      0      0    0      0
## 4   0.00  0.00  0   0      0      0   0    0      0      0    0      0
## 5   0.00  0.00  0   0      0      0   0    0      0      0    0      0
##   num85 technology num1999 parts pm direct cs meeting original project
## 1     0          0    0.00     0  0   0.00  0       0     0.00       0
## 2     0          0    0.07     0  0   0.00  0       0     0.00       0
## 3     0          0    0.00     0  0   0.06  0       0     0.12       0
## 4     0          0    0.00     0  0   0.00  0       0     0.00       0
## 5     0          0    0.00     0  0   0.00  0       0     0.00       0
##     re  edu table conference charSemicolon charRoundbracket
## 1 0.00 0.00     0          0          0.00            0.000
## 2 0.00 0.00     0          0          0.00            0.132
## 3 0.06 0.06     0          0          0.01            0.143
## 4 0.00 0.00     0          0          0.00            0.137
## 5 0.00 0.00     0          0          0.00            0.135
##   charSquarebracket charExclamation charDollar charHash capitalAve
## 1                 0           0.778      0.000    0.000      3.756
## 2                 0           0.372      0.180    0.048      5.114
## 3                 0           0.276      0.184    0.010      9.821
## 4                 0           0.137      0.000    0.000      3.537
## 5                 0           0.135      0.000    0.000      3.537
##   capitalLong capitalTotal type
## 1          61          278 spam
## 2         101         1028 spam
## 3         485         2259 spam
## 4          40          191 spam
## 5          40          191 spam
```

Um resumo das variáveis da base:

```
summary(spam)
```

```
##       make           address            all             num3d
##  Min.   :0.0000   Min.   : 0.000   Min.   :0.0000   Min.   : 0.00000
##  1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 0.00000
##  Median :0.0000   Median : 0.000   Median :0.0000   Median : 0.00000
##  Mean   :0.1046   Mean   : 0.213   Mean   :0.2807   Mean   : 0.06542
##  3rd Qu.:0.0000   3rd Qu.: 0.000   3rd Qu.:0.4200   3rd Qu.: 0.00000
##  Max.   :4.5400   Max.   :14.280   Max.   :5.1000   Max.   :42.81000
##       our               over            remove          internet
##  Min.   : 0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median :0.0000   Median :0.0000   Median : 0.0000
##  Mean   : 0.3122   Mean   :0.0959   Mean   :0.1142   Mean   : 0.1053
##  3rd Qu.: 0.3800   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.: 0.0000
##  Max.   :10.0000   Max.   :5.8800   Max.   :7.2700   Max.   :11.1100
##      order              mail            receive             will
##  Min.   :0.00000   Min.   : 0.0000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.00000   Median : 0.0000   Median :0.00000   Median :0.1000
##  Mean   :0.09007   Mean   : 0.2394   Mean   :0.05982   Mean   :0.5417
##  3rd Qu.:0.00000   3rd Qu.: 0.1600   3rd Qu.:0.00000   3rd Qu.:0.8000
##  Max.   :5.26000   Max.   :18.1800   Max.   :2.61000   Max.   :9.6700
##      people            report           addresses           free
##  Min.   :0.00000   Min.   : 0.00000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.:0.00000   1st Qu.: 0.00000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median :0.00000   Median : 0.00000   Median :0.0000   Median : 0.0000
##  Mean   :0.09393   Mean   : 0.05863   Mean   :0.0492   Mean   : 0.2488
##  3rd Qu.:0.00000   3rd Qu.: 0.00000   3rd Qu.:0.0000   3rd Qu.: 0.1000
##  Max.   :5.55000   Max.   :10.00000   Max.   :4.4100   Max.   :20.0000
##     business          email             you             credit
##  Min.   :0.0000   Min.   :0.0000   Min.   : 0.000   Min.   : 0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.000   1st Qu.: 0.00000
##  Median :0.0000   Median :0.0000   Median : 1.310   Median : 0.00000
##  Mean   :0.1426   Mean   :0.1847   Mean   : 1.662   Mean   : 0.08558
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.: 2.640   3rd Qu.: 0.00000
##  Max.   :7.1400   Max.   :9.0900   Max.   :18.750   Max.   :18.18000
##       your              font             num000           money
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.00000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.00000
##  Median : 0.2200   Median : 0.0000   Median :0.0000   Median : 0.00000
##  Mean   : 0.8098   Mean   : 0.1212   Mean   :0.1016   Mean   : 0.09427
##  3rd Qu.: 1.2700   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 0.00000
##  Max.   :11.1100   Max.   :17.1000   Max.   :5.4500   Max.   :12.50000
##        hp               hpl              george            num650
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median :0.0000
##  Mean   : 0.5495   Mean   : 0.2654   Mean   : 0.7673   Mean   :0.1248
##  3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.: 0.0000   3rd Qu.:0.0000
##  Max.   :20.8300   Max.   :16.6600   Max.   :33.3300   Max.   :9.0900
##       lab                labs            telnet             num857
##  Min.   : 0.00000   Min.   :0.0000   Min.   : 0.00000   Min.   :0.00000
##  1st Qu.: 0.00000   1st Qu.:0.0000   1st Qu.: 0.00000   1st Qu.:0.00000
##  Median : 0.00000   Median :0.0000   Median : 0.00000   Median :0.00000
##  Mean   : 0.09892   Mean   :0.1029   Mean   : 0.06475   Mean   :0.04705
##  3rd Qu.: 0.00000   3rd Qu.:0.0000   3rd Qu.: 0.00000   3rd Qu.:0.00000
##  Max.   :14.28000   Max.   :5.8800   Max.   :12.50000   Max.   :4.76000
##       data              num415            num85            technology
##  Min.   : 0.00000   Min.   :0.00000   Min.   : 0.0000   Min.   :0.00000
##  1st Qu.: 0.00000   1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:0.00000
##  Median : 0.00000   Median :0.00000   Median : 0.0000   Median :0.00000
##  Mean   : 0.09723   Mean   :0.04784   Mean   : 0.1054   Mean   :0.09748
##  3rd Qu.: 0.00000   3rd Qu.:0.00000   3rd Qu.: 0.0000   3rd Qu.:0.00000
##  Max.   :18.18000   Max.   :4.76000   Max.   :20.0000   Max.   :7.69000
##     num1999          parts              pm               direct
##  Min.   :0.000   Min.   :0.0000   Min.   : 0.00000   Min.   :0.00000
##  1st Qu.:0.000   1st Qu.:0.0000   1st Qu.: 0.00000   1st Qu.:0.00000
##  Median :0.000   Median :0.0000   Median : 0.00000   Median :0.00000
##  Mean   :0.137   Mean   :0.0132   Mean   : 0.07863   Mean   :0.06483
##  3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.: 0.00000   3rd Qu.:0.00000
##  Max.   :6.890   Max.   :8.3300   Max.   :11.11000   Max.   :4.76000
##        cs             meeting           original         project
##  Min.   :0.00000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median :0.00000   Median : 0.0000   Median :0.0000   Median : 0.0000
##  Mean   :0.04367   Mean   : 0.1323   Mean   :0.0461   Mean   : 0.0792
##  3rd Qu.:0.00000   3rd Qu.: 0.0000   3rd Qu.:0.0000   3rd Qu.: 0.0000
##  Max.   :7.14000   Max.   :14.2800   Max.   :3.5700   Max.   :20.0000
##        re               edu              table             conference
##  Min.   : 0.0000   Min.   : 0.0000   Min.   :0.000000   Min.   : 0.00000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.:0.000000   1st Qu.: 0.00000
##  Median : 0.0000   Median : 0.0000   Median :0.000000   Median : 0.00000
##  Mean   : 0.3012   Mean   : 0.1798   Mean   :0.005444   Mean   : 0.03187
##  3rd Qu.: 0.1100   3rd Qu.: 0.0000   3rd Qu.:0.000000   3rd Qu.: 0.00000
##  Max.   :21.4200   Max.   :22.0500   Max.   :2.170000   Max.   :10.00000
##  charSemicolon     charRoundbracket charSquarebracket charExclamation
##  Min.   :0.00000   Min.   :0.000    Min.   :0.00000   Min.   : 0.0000
##  1st Qu.:0.00000   1st Qu.:0.000    1st Qu.:0.00000   1st Qu.: 0.0000
##  Median :0.00000   Median :0.065    Median :0.00000   Median : 0.0000
##  Mean   :0.03857   Mean   :0.139    Mean   :0.01698   Mean   : 0.2691
##  3rd Qu.:0.00000   3rd Qu.:0.188    3rd Qu.:0.00000   3rd Qu.: 0.3150
##  Max.   :4.38500   Max.   :9.752    Max.   :4.08100   Max.   :32.4780
##    charDollar         charHash          capitalAve        capitalLong
##  Min.   :0.00000   Min.   : 0.00000   Min.   :   1.000   Min.   :   1.00
##  1st Qu.:0.00000   1st Qu.: 0.00000   1st Qu.:   1.588   1st Qu.:   6.00
##  Median :0.00000   Median : 0.00000   Median :   2.276   Median :  15.00
##  Mean   :0.07581   Mean   : 0.04424   Mean   :   5.191   Mean   :  52.17
##  3rd Qu.:0.05200   3rd Qu.: 0.00000   3rd Qu.:   3.706   3rd Qu.:  43.00
##  Max.   :6.00300   Max.   :19.82900   Max.   :1102.500   Max.   :9989.00
##   capitalTotal          type
##  Min.   :    1.0   nonspam:2788
##  1st Qu.:   35.0   spam   :1813
##  Median :   95.0
##  Mean   :  283.3
##  3rd Qu.:  266.0
##  Max.   :15841.0
```

Um histograma com a média de distribuição de frequência de cada uma das variáveis da base, divididas entre spam e nonspam:

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:kernlab':
##
##     alpha
```

```
##
## Attaching package: 'magrittr'
```

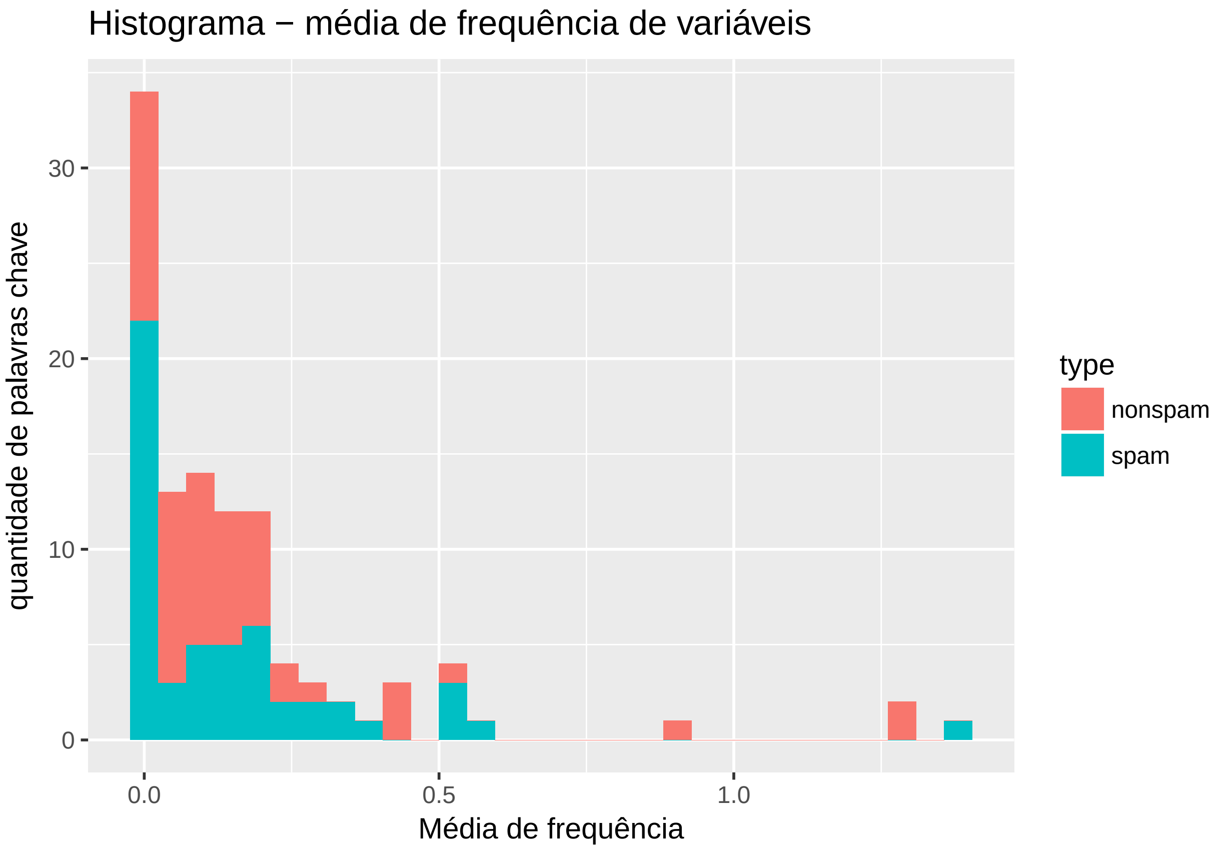
```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

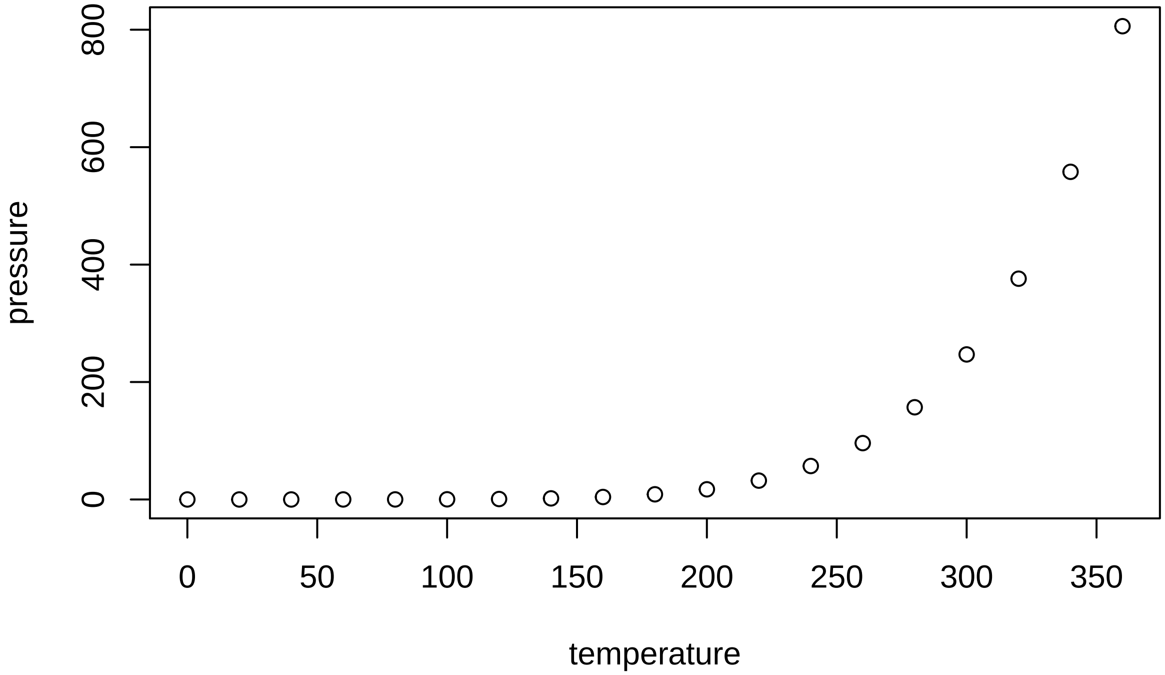
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```r
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.