R Notebook

Code

s…	to_multiple	fr…	cc	sent_email	time	im…	attach	dollar	win…
<fctr>	<fctr>	<fctr>	<int>	<fctr>	<S3: POSIXct>	<dbl>	<dbl>	<dbl>	<fctr>
0	0	1	0	0	2012-01-01 07:16:41	0	0	0	no
0	0	1	0	0	2012-01-01 08:03:59	0	0	0	no
0	0	1	0	0	2012-01-01 17:00:32	0	0	4	no
0	0	1	0	0	2012-01-01 10:09:49	0	0	0	no
0	0	1	0	0	2012-01-01 11:00:01	0	0	0	no
0	0	1	0	0	2012-01-01 11:04:46	0	0	0	no
0	1	1	0	1	2012-01-01 18:55:06	0	0	0	no
0	1	1	1	1	2012-01-01 19:45:21	1	1	0	no
0	0	1	0	0	2012-01-01 22:08:59	0	0	0	no
0	0	1	0	0	2012-01-01 19:12:00	0	0	0	no

1-10 of 3,921 rows | 1-10 of 21 columns Previous **1** 2 3 4 5 6 ... 100 Next

Hide

```
dim(email)
```

```
[1] 3921   21
```

1 - Variables categoricas binomiales: to_multiple, from, sent_email, image, winner, format, re_subj, exclaim_subj, urgent_subj. Variables categoricas ordinales: number.

2 - variables cuantitativas: cc, time, attach, dollar, password, num_char, line_breaks, exclaim_mess

Hide

```
write.csv(x = email, file = "email.csv", row.names = FALSE, col.names = TRUE)
```

```
attempt to set 'col.names' ignored
```

Hide

```


sum(is.na(email))
```

```
[1] 0
```

La variable dependiente a predecir es spam, en el que se recoge si un mail va a la bandeja de spam del correo del destinatario, o por el contrario va a la bandeja de recibidos.

Se trata de una variable categorica bonimial: (0 - NO / 1 - SI)

Hide

```
library(summarytools)

table(email$spam)
```

```

   0    1 
3554  367 
```

Hide

```
freq(email$spam, style = "rmarkdown")
```

```
### Frequencies
#### email$spam
**Type:** Factor

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|      **0** | 3554 |   90.64 |        90.64 |   90.64 |        90.64 |
|      **1** |  367 |    9.36 |       100.00 |    9.36 |       100.00 |
| **\<NA\>** |    0 |         |              |    0.00 |       100.00 |
|  **Total** | 3921 |  100.00 |       100.00 |  100.00 |       100.00 |
```

Podemos ver en esta tabla de frecuencias de los 3921 mails analizados, 3554(90.64%) son catalogados como NO spam(en caso de que 0 sea que no) y 367 (9.36%) son catalogados como si spam. No tenemos valores nulos en esta variable.

Vamos a comenzar analizando las variables cuantitativas

CC

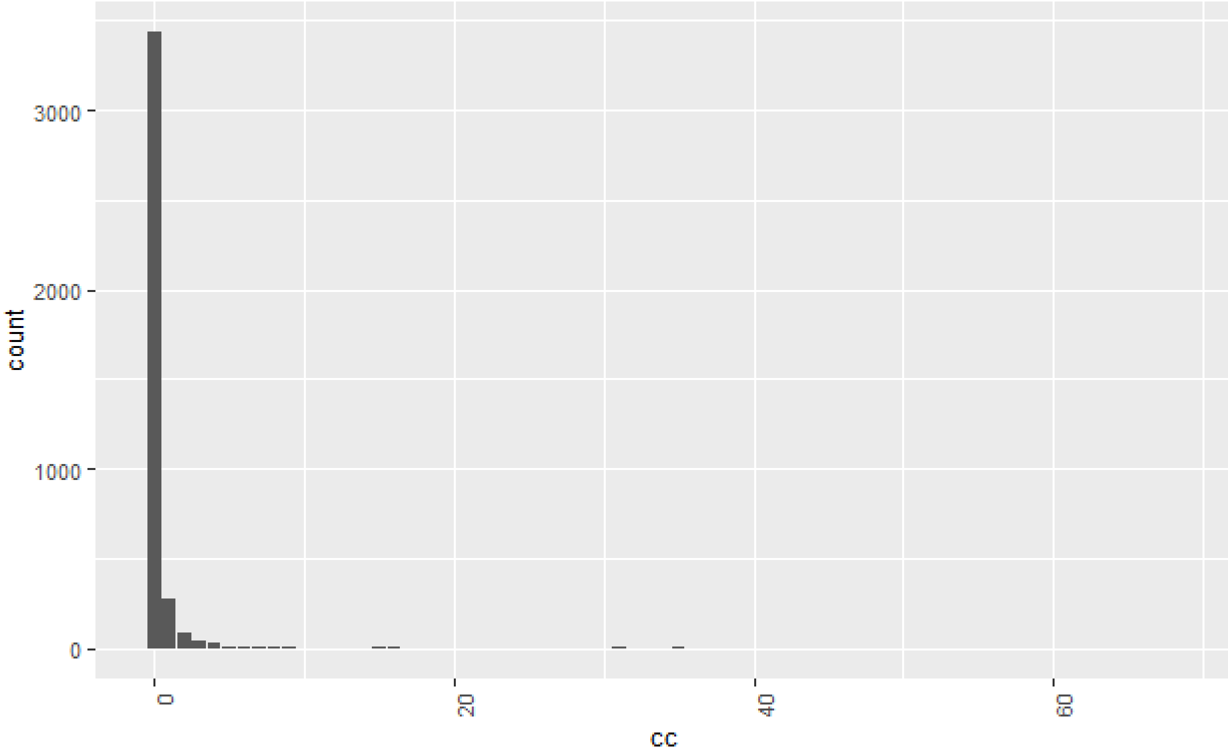
Hide

```
summary(email$cc)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.4045  0.0000 68.0000 
```

¿ Cuanta gente suele ir en copia en estos correos?

Vamos a realizar un histograma y un grafico de bigotes



Hide

```
ctable(email$cc, email$spam)
```

```
Cross-Tabulation, Row Proportions
cc * spam
Data Frame: email

------- ------ -------------- ------------- --------------
          spam              0             1          Total
     cc
      0         3087 ( 89.8%)   349 ( 10.2%)   3436 (100.0%)
      1          278 (100.0%)     0 (  0.0%)    278 (100.0%)
      2           80 (100.0%)     0 (  0.0%)     80 (100.0%)
      3           39 ( 95.1%)     2 (  4.9%)     41 (100.0%)
      4           21 ( 63.6%)    12 ( 36.4%)     33 (100.0%)
      5            7 (100.0%)     0 (  0.0%)      7 (100.0%)
      6            9 (100.0%)     0 (  0.0%)      9 (100.0%)
      7            8 (100.0%)     0 (  0.0%)      8 (100.0%)
      8            2 (100.0%)     0 (  0.0%)      2 (100.0%)
      9            2 (100.0%)     0 (  0.0%)      2 (100.0%)
     12            0 (  0.0%)     1 (100.0%)      1 (100.0%)
     13            1 (100.0%)     0 (  0.0%)      1 (100.0%)
     15            3 (100.0%)     0 (  0.0%)      3 (100.0%)
     16            3 (100.0%)     0 (  0.0%)      3 (100.0%)
     18            1 (100.0%)     0 (  0.0%)      1 (100.0%)
     19            0 (  0.0%)     1 (100.0%)      1 (100.0%)
     21            1 (100.0%)     0 (  0.0%)      1 (100.0%)
     23            0 (  0.0%)     1 (100.0%)      1 (100.0%)
     25            1 (100.0%)     0 (  0.0%)      1 (100.0%)
     31            2 (100.0%)     0 (  0.0%)      2 (100.0%)
     33            1 (100.0%)     0 (  0.0%)      1 (100.0%)
     35            5 (100.0%)     0 (  0.0%)      5 (100.0%)
     38            1 (100.0%)     0 (  0.0%)      1 (100.0%)
     50            0 (  0.0%)     1 (100.0%)      1 (100.0%)
     64            1 (100.0%)     0 (  0.0%)      1 (100.0%)
     68            1 (100.0%)     0 (  0.0%)      1 (100.0%)
  Total         3554 ( 90.6%)   367 (  9.4%)   3921 (100.0%)
------- ------ -------------- ------------- --------------
```

Vamos a conbvertir esta variable en binaria, entre los que si van en copia y los que no.

Hide

```
ctable(email$cc_binary, email$spam)
```

```
Cross-Tabulation, Row Proportions
cc_binary * spam
Data Frame: email

----------- ------ ------------- ------------ --------------
              spam             0            1          Total
  cc_binary
          0         3087 (89.8%)   349 (10.2%)   3436 (100.0%)
          1          467 (96.3%)    18 ( 3.7%)    485 (100.0%)
      Total         3554 (90.6%)   367 ( 9.4%)   3921 (100.0%)
----------- ------ ------------- ------------ --------------
```

attach

Cuantos documentos adjuntos suelen llevar estos mails ??

Hide

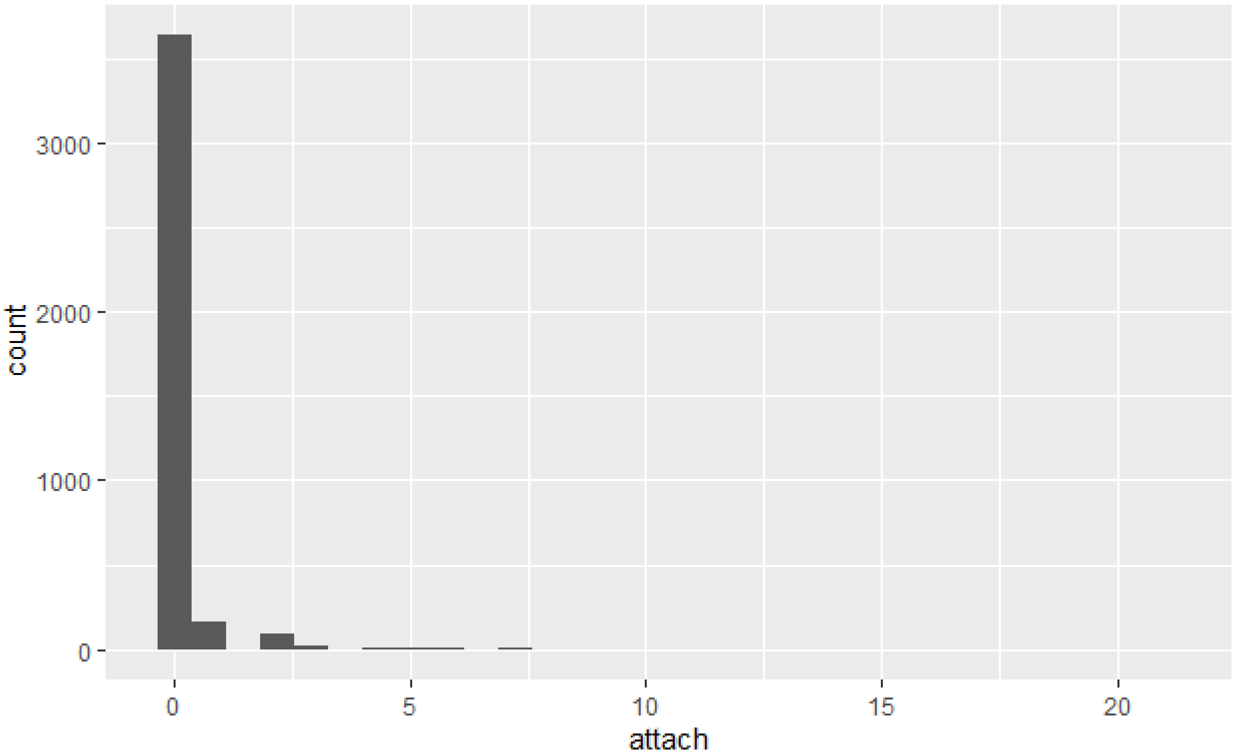
```
table(email$attach)
```

```

   0    1    2    3    4    5    6    7    8    9   10   20   21 
3638  158   90   19    3    4    2    2    1    1    1    1    1 
```

Hide

```
summary(email$attach)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.1329  0.0000 21.0000 
```



Hide

```
ctable(email$attach, email$spam)
```

```
Cross-Tabulation, Row Proportions
attach * spam
Data Frame: email

-------- ------ -------------- ------------ --------------
           spam              0            1          Total
  attach
       0          3315 ( 91.1%)   323 ( 8.9%)   3638 (100.0%)
       1           150 ( 94.9%)     8 ( 5.1%)    158 (100.0%)
       2            54 ( 60.0%)    36 (40.0%)     90 (100.0%)
       3            19 (100.0%)     0 ( 0.0%)     19 (100.0%)
       4             3 (100.0%)     0 ( 0.0%)      3 (100.0%)
       5             4 (100.0%)     0 ( 0.0%)      4 (100.0%)
       6             2 (100.0%)     0 ( 0.0%)      2 (100.0%)
       7             2 (100.0%)     0 ( 0.0%)      2 (100.0%)
       8             1 (100.0%)     0 ( 0.0%)      1 (100.0%)
       9             1 (100.0%)     0 ( 0.0%)      1 (100.0%)
      10             1 (100.0%)     0 ( 0.0%)      1 (100.0%)
      20             1 (100.0%)     0 ( 0.0%)      1 (100.0%)
      21             1 (100.0%)     0 ( 0.0%)      1 (100.0%)
   Total          3554 ( 90.6%)   367 ( 9.4%)   3921 (100.0%)
-------- ------ -------------- ------------ --------------
```

Vamos a convertir esta variable en binaria, de tal forma que los correos que NO lleven adjunto será un 0 y los que si un 1

Hide

```
ctable(email$attach_binary, email$spam)
```

```
Cross-Tabulation, Row Proportions
attach_binary * spam
Data Frame: email

--------------- ------ ------------- ------------ --------------
                  spam             0            1          Total
  attach_binary
              0          3315 (91.1%)   323 ( 8.9%)   3638 (100.0%)
              1           239 (84.5%)    44 (15.5%)    283 (100.0%)
          Total          3554 (90.6%)   367 ( 9.4%)   3921 (100.0%)
--------------- ------ ------------- ------------ --------------
```

Podemos ver como el 15.5% de los correos que contiene algun adjunto es catalogado como spam.

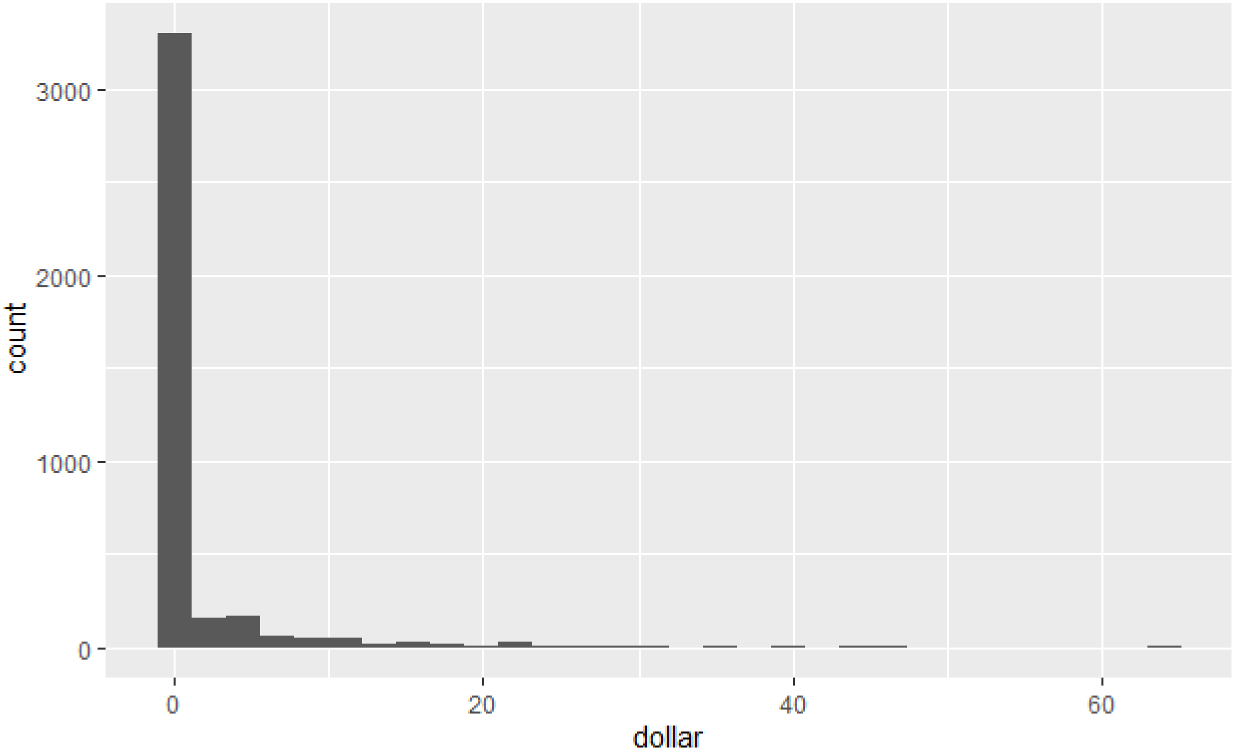
dollar

Variable que recoge las veces que aparece el simbolo dolar.

Hide

```
summary(email$dollar)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.000   0.000   1.467   0.000  64.000 
```



Hide

```
ctable(email$dollar, email$spam)
```

```
Cross-Tabulation, Row Proportions
dollar * spam
Data Frame: email

-------- ------ -------------- ------------- --------------
          spam              0             1          Total
  dollar
       0         2886 ( 90.9%)  289 (  9.1%)  3175 (100.0%)
       1           97 ( 80.8%)   23 ( 19.2%)   120 (100.0%)
       2          122 ( 80.8%)   29 ( 19.2%)   151 (100.0%)
       3            4 ( 40.0%)    6 ( 60.0%)    10 (100.0%)
       4          139 ( 95.2%)    7 (  4.8%)   146 (100.0%)
       5           15 ( 75.0%)    5 ( 25.0%)    20 (100.0%)
       6           43 ( 97.7%)    1 (  2.3%)    44 (100.0%)
       7            9 ( 75.0%)    3 ( 25.0%)    12 (100.0%)
       8           34 ( 97.1%)    1 (  2.9%)    35 (100.0%)
       9           10 (100.0%)    0 (  0.0%)    10 (100.0%)
      10           22 (100.0%)    0 (  0.0%)    22 (100.0%)
      11           10 (100.0%)    0 (  0.0%)    10 (100.0%)
      12           20 (100.0%)    0 (  0.0%)    20 (100.0%)
      13            7 (100.0%)    0 (  0.0%)     7 (100.0%)
      14           14 (100.0%)    0 (  0.0%)    14 (100.0%)
      15            5 (100.0%)    0 (  0.0%)     5 (100.0%)
      16           23 (100.0%)    0 (  0.0%)    23 (100.0%)
      17            2 (100.0%)    0 (  0.0%)     2 (100.0%)
      18           14 (100.0%)    0 (  0.0%)    14 (100.0%)
      19            0 (  0.0%)    1 (100.0%)     1 (100.0%)
      20           10 (100.0%)    0 (  0.0%)    10 (100.0%)
      21            7 (100.0%)    0 (  0.0%)     7 (100.0%)
      22           12 (100.0%)    0 (  0.0%)    12 (100.0%)
      23            7 (100.0%)    0 (  0.0%)     7 (100.0%)
      24            7 (100.0%)    0 (  0.0%)     7 (100.0%)
      25            3 (100.0%)    0 (  0.0%)     3 (100.0%)
      26            6 ( 85.7%)    1 ( 14.3%)     7 (100.0%)
      27            1 (100.0%)    0 (  0.0%)     1 (100.0%)
      28            5 (100.0%)    0 (  0.0%)     5 (100.0%)
      29            1 (100.0%)    0 (  0.0%)     1 (100.0%)
      30            1 (100.0%)    0 (  0.0%)     1 (100.0%)
      32            2 (100.0%)    0 (  0.0%)     2 (100.0%)
      34            1 (100.0%)    0 (  0.0%)     1 (100.0%)
      36            1 ( 50.0%)    1 ( 50.0%)     2 (100.0%)
      40            3 (100.0%)    0 (  0.0%)     3 (100.0%)
      44            3 (100.0%)    0 (  0.0%)     3 (100.0%)
      46            2 (100.0%)    0 (  0.0%)     2 (100.0%)
      48            1 (100.0%)    0 (  0.0%)     1 (100.0%)
      54            1 (100.0%)    0 (  0.0%)     1 (100.0%)
      63            1 (100.0%)    0 (  0.0%)     1 (100.0%)
      64            3 (100.0%)    0 (  0.0%)     3 (100.0%)
   Total         3554 ( 90.6%)  367 (  9.4%)  3921 (100.0%)
-------- ------ -------------- ------------- --------------
```

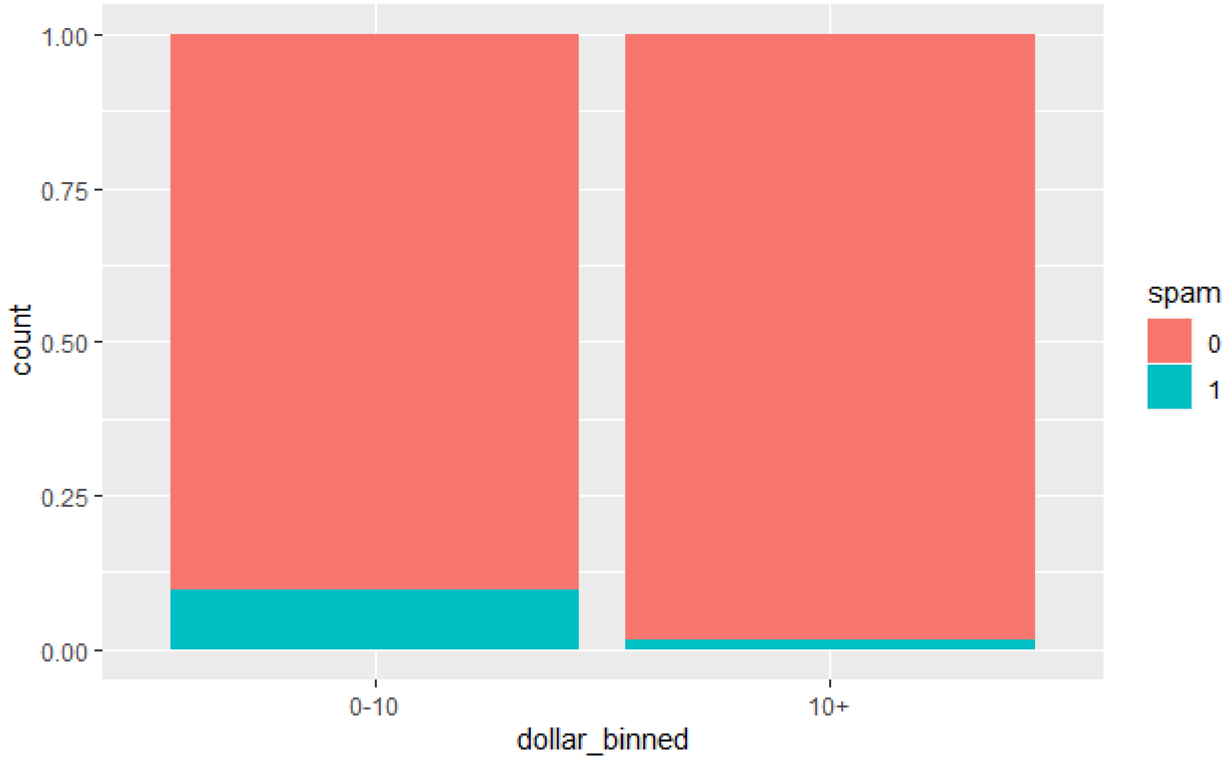
Hide

```
ctable(email$dollar_binned, email$spam)
```

```
Cross-Tabulation, Row Proportions
dollar_binned * spam
Data Frame: email

--------------- ------ ------------- ------------ ---------------
                  spam             0            1           Total
  dollar_binned
           0-10          3359 (90.2%)  364 ( 9.8%)   3723 (100.0%)
            10+           195 (98.5%)    3 ( 1.5%)    198 (100.0%)
          Total          3554 (90.6%)  367 ( 9.4%)   3921 (100.0%)
--------------- ------ ------------- ------------ ---------------
```

Hide

```
ggplot(email, aes(dollar_binned, fill = spam)) + geom_bar(position = "fill")
```



password

Cuantas veces se repite la palabra password dentro de una correo electronico

Hide

```
table(email$password)
```

```
   0    1    2    3    4    5    6    8   11   13   18   22   28
3809   22   39    8   23    5    3    5    2    1    1    2    1
```

Hide

```
ctable(email$password, email$spam)
```

```
Cross-Tabulation, Row Proportions
password * spam
Data Frame: email

---------- ------ --------------- ------------ ---------------
             spam               0            1           Total
  password
         0           3446 ( 90.5%)   363 ( 9.5%)   3809 (100.0%)
         1             20 ( 90.9%)     2 ( 9.1%)     22 (100.0%)
         2             37 ( 94.9%)     2 ( 5.1%)     39 (100.0%)
         3              8 (100.0%)     0 ( 0.0%)      8 (100.0%)
         4             23 (100.0%)     0 ( 0.0%)     23 (100.0%)
         5              5 (100.0%)     0 ( 0.0%)      5 (100.0%)
         6              3 (100.0%)     0 ( 0.0%)      3 (100.0%)
         8              5 (100.0%)     0 ( 0.0%)      5 (100.0%)
        11              2 (100.0%)     0 ( 0.0%)      2 (100.0%)
        13              1 (100.0%)     0 ( 0.0%)      1 (100.0%)
        18              1 (100.0%)     0 ( 0.0%)      1 (100.0%)
        22              2 (100.0%)     0 ( 0.0%)      2 (100.0%)
        28              1 (100.0%)     0 ( 0.0%)      1 (100.0%)
     Total           3554 ( 90.6%)   367 ( 9.4%)   3921 (100.0%)
---------- ------ --------------- ------------ ---------------
```

Vamos a convertir la variable password en binaria, 0 - ninguna / 1 - alguna vez

Hide

```
email <- email %>%
  mutate(password_binary = ifelse(password == 0,0,1))

ctable(email$password_binary, email$spam)
```

```
Cross-Tabulation, Row Proportions
password_binary * spam
Data Frame: email

---------------- ------ ------------- ------------ ---------------
                   spam             0            1           Total
  password_binary
                0          3446 (90.5%)   363 ( 9.5%)   3809 (100.0%)
                1           108 (96.4%)     4 ( 3.6%)    112 (100.0%)
            Total          3554 (90.6%)   367 ( 9.4%)   3921 (100.0%)
---------------- ------ ------------- ------------ ---------------
```

En un 9.5% de los casos en los que no aparece la palabra password se declara como spam y solo un 3.6% cuando aparece alguna vez la palabra password.

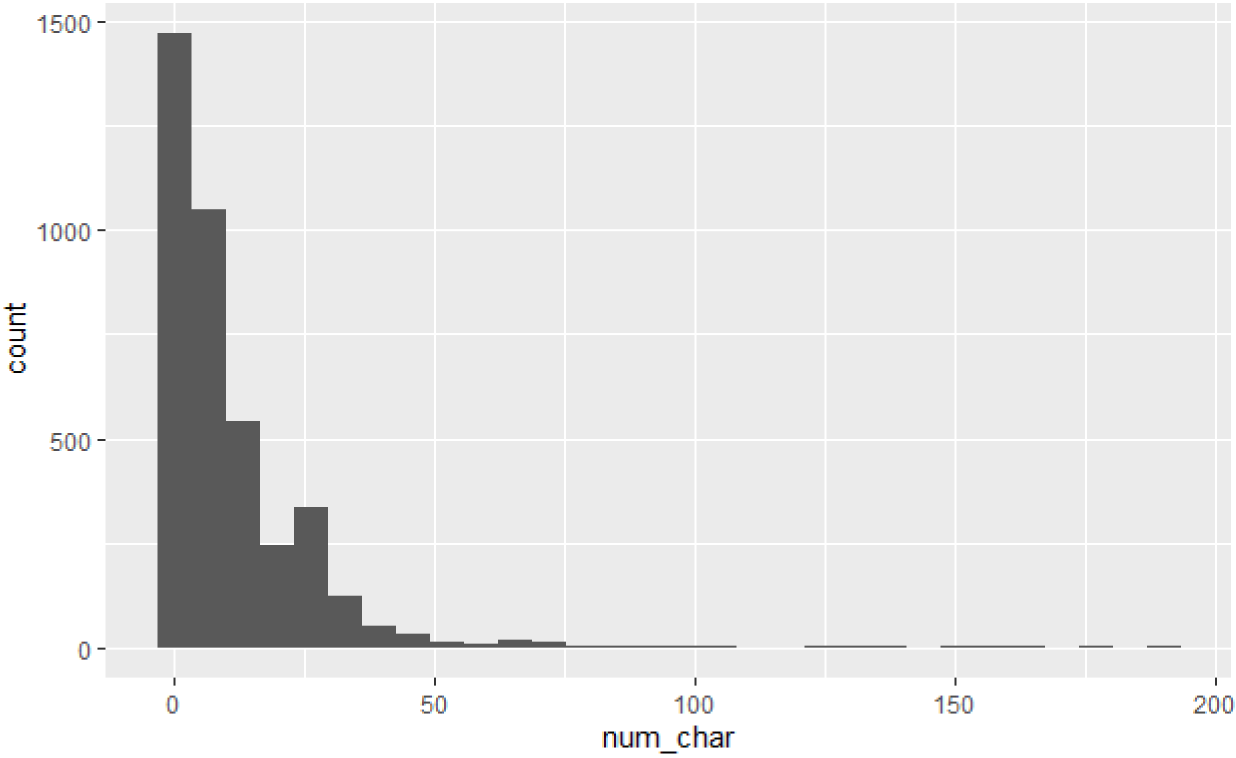
num_char

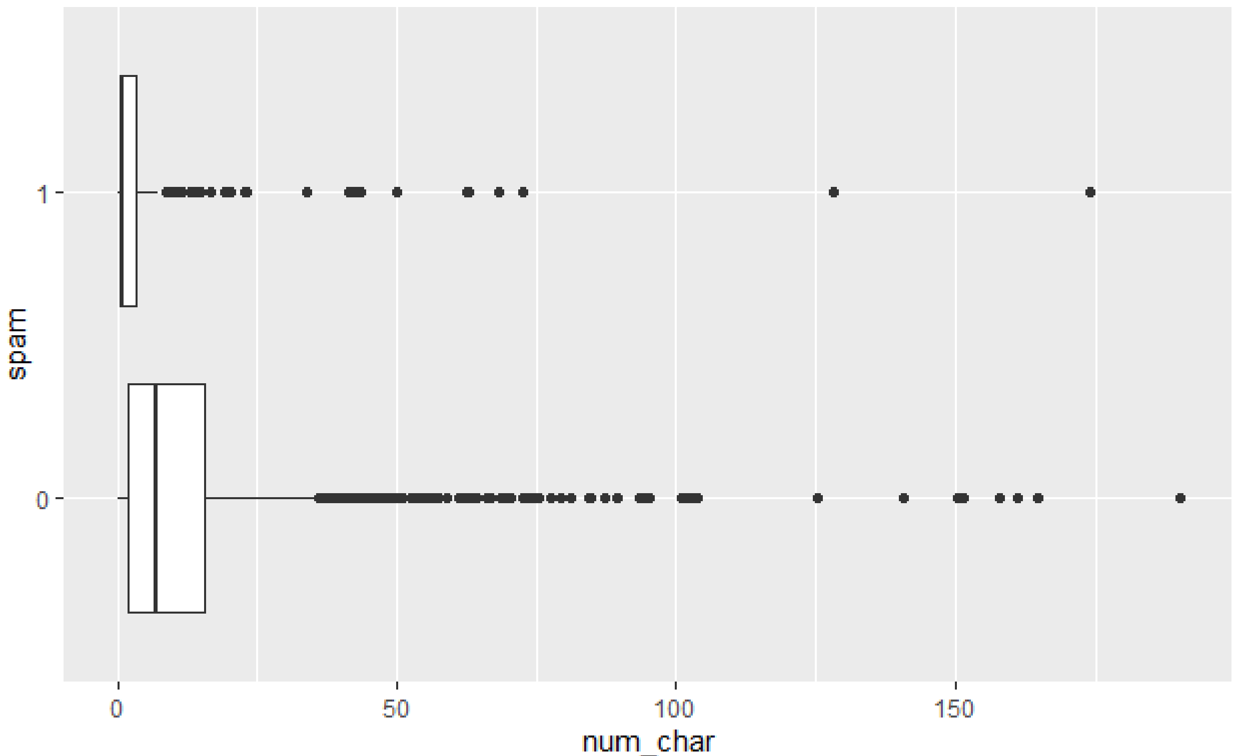
¿Son determinantes los caracteres a la hora de establecer un correo como spam?

Hide

```
summary(email$num_char)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.001   1.459   5.856  10.707  14.084 190.087 
```





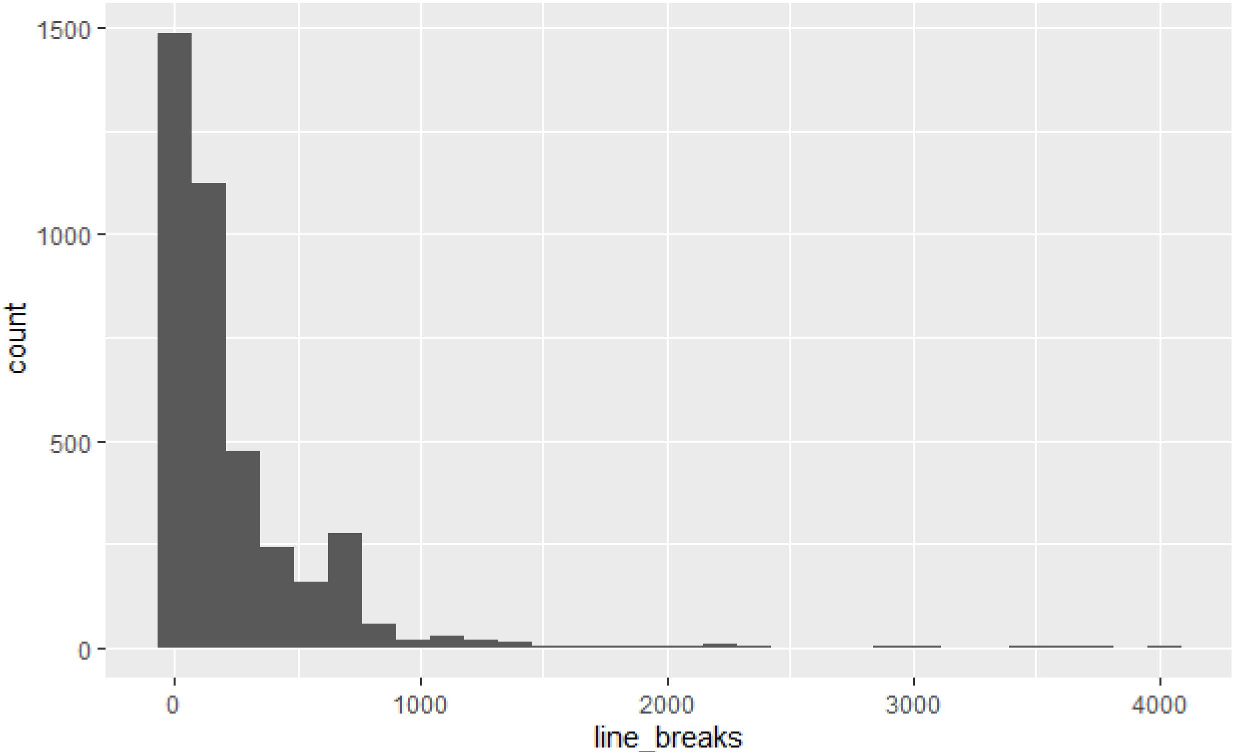
line_breaks - saltos de linea

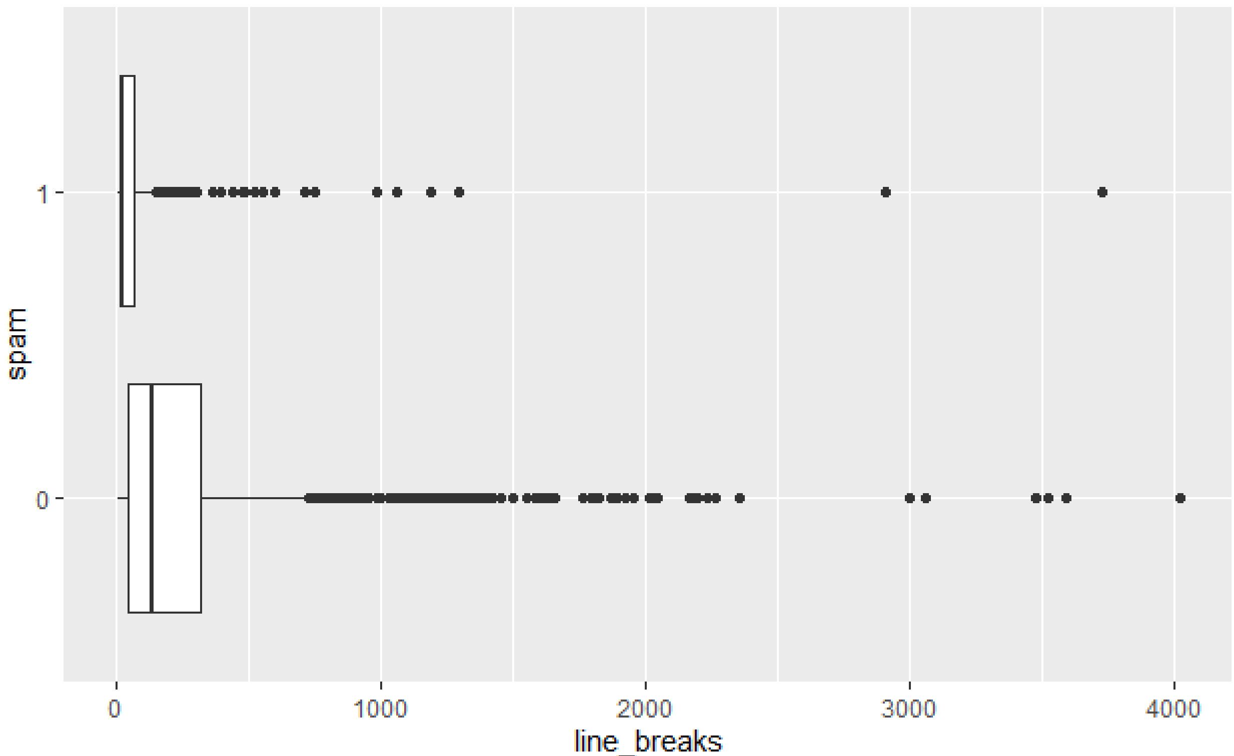
¿ Los saltos de liena son determinantes a la hora de declarar un correo como spam?

Hide

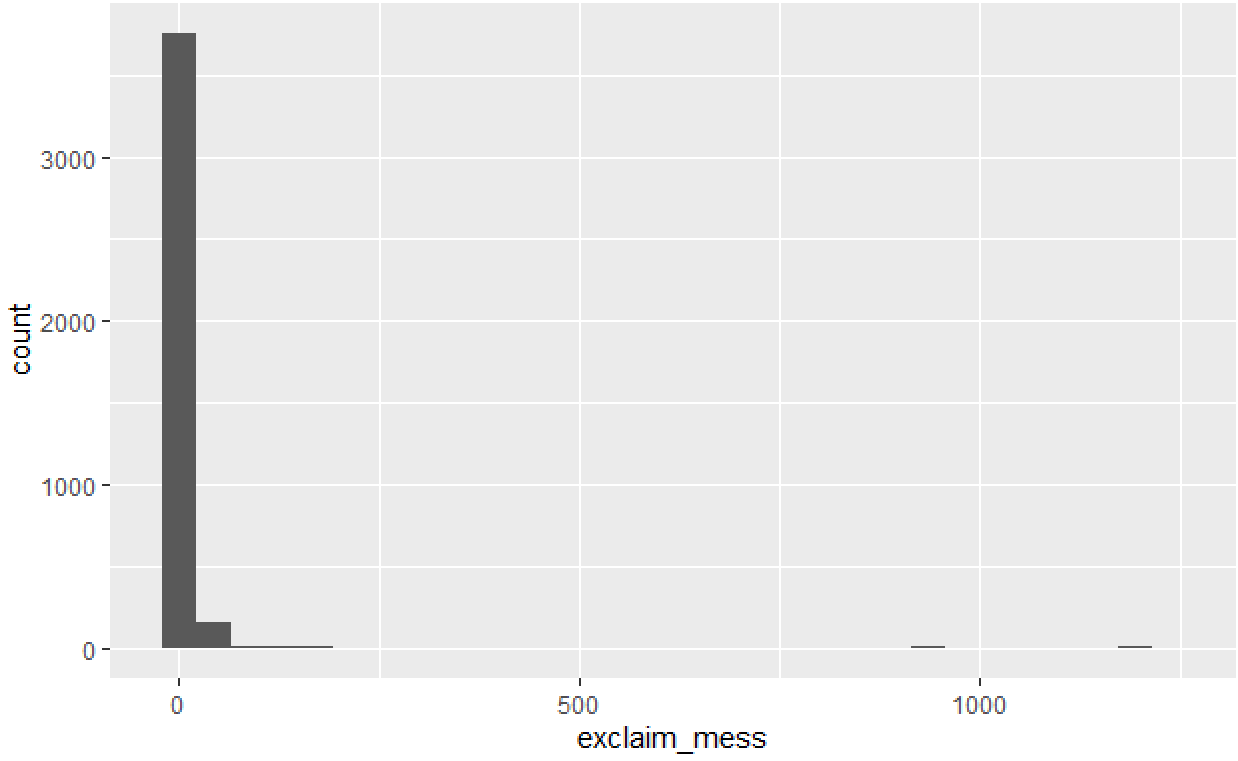
```
summary(email$line_breaks)
```

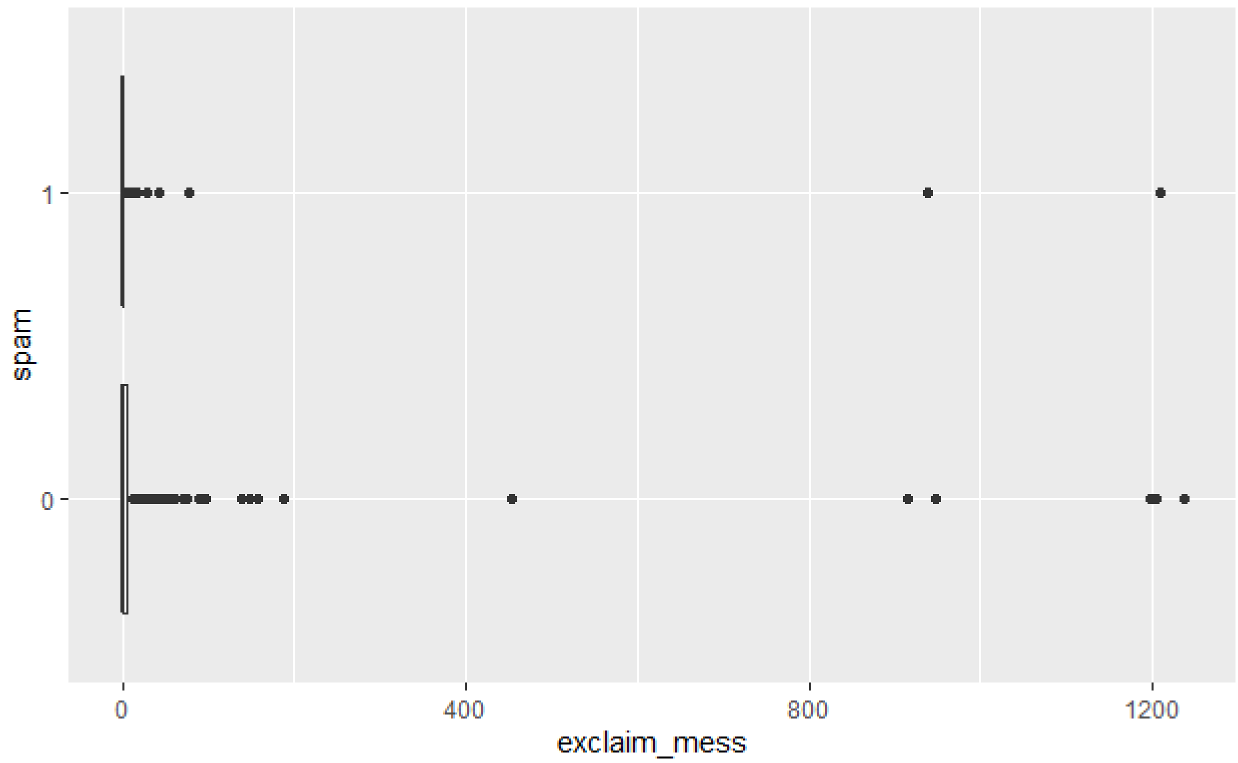
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    1.0    34.0   119.0   230.7   298.0  4022.0 
```





exclaim_mess





Hide

```
ctable(email$exclaim_mess, email$spam)
```

```
Cross-Tabulation, Row Proportions
exclaim_mess * spam
Data Frame: email

-------------- ------ -------------- ------------- --------------
                 spam              0             1          Total
  exclaim_mess
             0          1219 ( 84.9%)  216 ( 15.1%)  1435 (100.0%)
             1           650 ( 88.7%)   83 ( 11.3%)   733 (100.0%)
             2           482 ( 95.1%)   25 (  4.9%)   507 (100.0%)
             3           116 ( 90.6%)   12 (  9.4%)   128 (100.0%)
             4           185 ( 97.4%)    5 (  2.6%)   190 (100.0%)
             5           112 ( 99.1%)    1 (  0.9%)   113 (100.0%)
             6           112 ( 97.4%)    3 (  2.6%)   115 (100.0%)
             7            49 ( 96.1%)    2 (  3.9%)    51 (100.0%)
             8            91 ( 97.8%)    2 (  2.2%)    93 (100.0%)
             9            40 ( 88.9%)    5 ( 11.1%)    45 (100.0%)
            10            83 ( 97.6%)    2 (  2.4%)    85 (100.0%)
            11            17 (100.0%)    0 (  0.0%)    17 (100.0%)
            12            53 ( 94.6%)    3 (  5.4%)    56 (100.0%)
            13            20 (100.0%)    0 (  0.0%)    20 (100.0%)
            14            42 ( 97.7%)    1 (  2.3%)    43 (100.0%)
            15            11 (100.0%)    0 (  0.0%)    11 (100.0%)
            16            28 ( 96.6%)    1 (  3.4%)    29 (100.0%)
            17            11 ( 91.7%)    1 (  8.3%)    12 (100.0%)
            18            26 (100.0%)    0 (  0.0%)    26 (100.0%)
            19             5 (100.0%)    0 (  0.0%)     5 (100.0%)
            20            29 (100.0%)    0 (  0.0%)    29 (100.0%)
            21             9 (100.0%)    0 (  0.0%)     9 (100.0%)
            22            15 (100.0%)    0 (  0.0%)    15 (100.0%)
            23             3 (100.0%)    0 (  0.0%)     3 (100.0%)
            24            11 (100.0%)    0 (  0.0%)    11 (100.0%)
            25             6 (100.0%)    0 (  0.0%)     6 (100.0%)
            26            11 (100.0%)    0 (  0.0%)    11 (100.0%)
            27             1 (100.0%)    0 (  0.0%)     1 (100.0%)
            28             5 ( 83.3%)    1 ( 16.7%)     6 (100.0%)
            29             8 (100.0%)    0 (  0.0%)     8 (100.0%)
            30            13 (100.0%)    0 (  0.0%)    13 (100.0%)
            31            12 (100.0%)    0 (  0.0%)    12 (100.0%)
            32            13 (100.0%)    0 (  0.0%)    13 (100.0%)
            33             3 (100.0%)    0 (  0.0%)     3 (100.0%)
            34             3 (100.0%)    0 (  0.0%)     3 (100.0%)
            35             2 (100.0%)    0 (  0.0%)     2 (100.0%)
            36             3 (100.0%)    0 (  0.0%)     3 (100.0%)
            38             3 (100.0%)    0 (  0.0%)     3 (100.0%)
            39             1 (100.0%)    0 (  0.0%)     1 (100.0%)
            40             2 (100.0%)    0 (  0.0%)     2 (100.0%)
            41             1 (100.0%)    0 (  0.0%)     1 (100.0%)
            42             1 (100.0%)    0 (  0.0%)     1 (100.0%)
            43             2 ( 66.7%)    1 ( 33.3%)     3 (100.0%)
            44             3 (100.0%)    0 (  0.0%)     3 (100.0%)
            45             5 (100.0%)    0 (  0.0%)     5 (100.0%)
            46             3 (100.0%)    0 (  0.0%)     3 (100.0%)
            47             2 (100.0%)    0 (  0.0%)     2 (100.0%)
            48             1 (100.0%)    0 (  0.0%)     1 (100.0%)
```

```
            49            3 (100.0%)     0 (  0.0%)      3 (100.0%)
            52            1 (100.0%)     0 (  0.0%)      1 (100.0%)
            54            1 (100.0%)     0 (  0.0%)      1 (100.0%)
            55            4 (100.0%)     0 (  0.0%)      4 (100.0%)
            57            2 (100.0%)     0 (  0.0%)      2 (100.0%)
            58            2 (100.0%)     0 (  0.0%)      2 (100.0%)
            62            2 (100.0%)     0 (  0.0%)      2 (100.0%)
            71            1 (100.0%)     0 (  0.0%)      1 (100.0%)
            75            1 (100.0%)     0 (  0.0%)      1 (100.0%)
            78            0 (  0.0%)     1 (100.0%)      1 (100.0%)
            89            1 (100.0%)     0 (  0.0%)      1 (100.0%)
            94            1 (100.0%)     0 (  0.0%)      1 (100.0%)
            96            1 (100.0%)     0 (  0.0%)      1 (100.0%)
           139            1 (100.0%)     0 (  0.0%)      1 (100.0%)
           148            1 (100.0%)     0 (  0.0%)      1 (100.0%)
           157            1 (100.0%)     0 (  0.0%)      1 (100.0%)
           187            1 (100.0%)     0 (  0.0%)      1 (100.0%)
           454            1 (100.0%)     0 (  0.0%)      1 (100.0%)
           915            1 (100.0%)     0 (  0.0%)      1 (100.0%)
           939            0 (  0.0%)     1 (100.0%)      1 (100.0%)
           947            1 (100.0%)     0 (  0.0%)      1 (100.0%)
          1197            1 (100.0%)     0 (  0.0%)      1 (100.0%)
          1203            2 (100.0%)     0 (  0.0%)      2 (100.0%)
          1209            0 (  0.0%)     1 (100.0%)      1 (100.0%)
          1236            1 (100.0%)     0 (  0.0%)      1 (100.0%)
         Total         3554 ( 90.6%)   367 (  9.4%)   3921 (100.0%)
-------------- ------ --------------- -------------- ---------------
```

Vamos ahora a analizar las variables categoricas.

to_multiple

Vamos a ver primero, del total de correos analizados, cuantos se envian a multiples personas y cuantos no

Hide

```
freq(email$to_multiple, style = "rmarkdown")
```

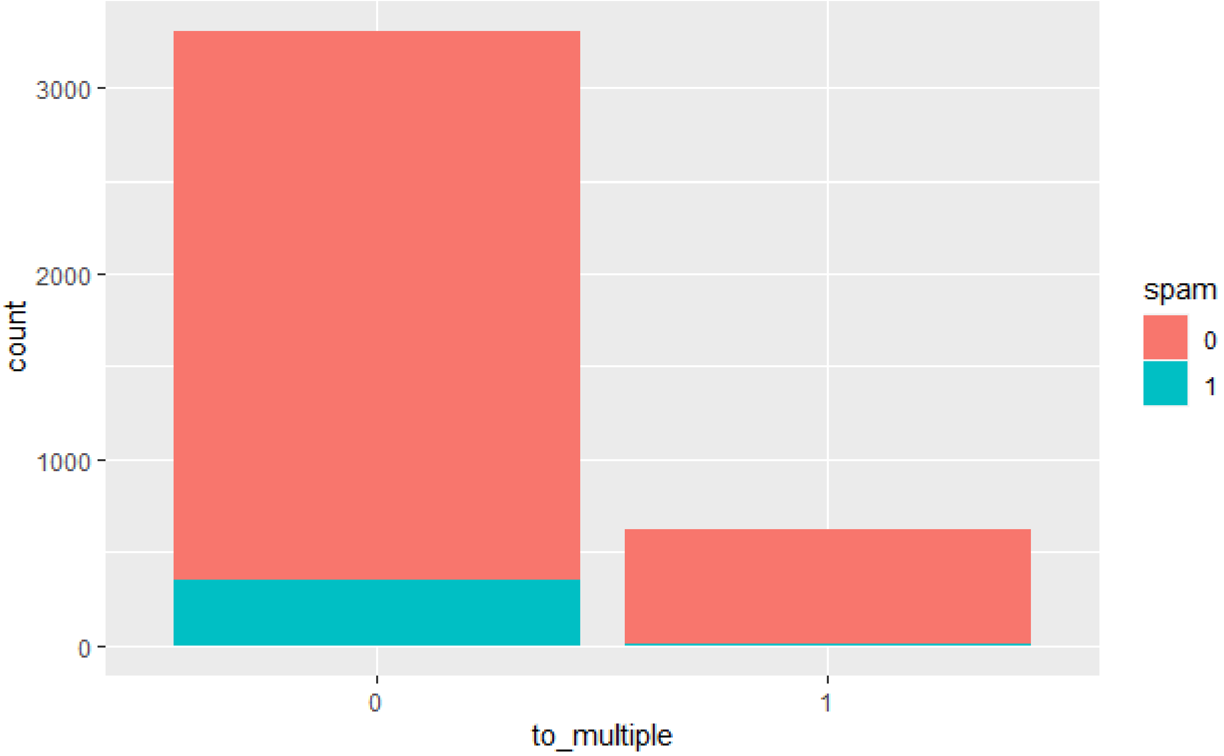
```
### Frequencies
#### email$to_multiple
**Type:** Factor

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|      **0** | 3301 |   84.19 |        84.19 |   84.19 |        84.19 |
|      **1** |  620 |   15.81 |       100.00 |   15.81 |       100.00 |
| **\<NA\>** |    0 |         |              |    0.00 |       100.00 |
|  **Total** | 3921 |  100.00 |       100.00 |  100.00 |       100.00 |
```

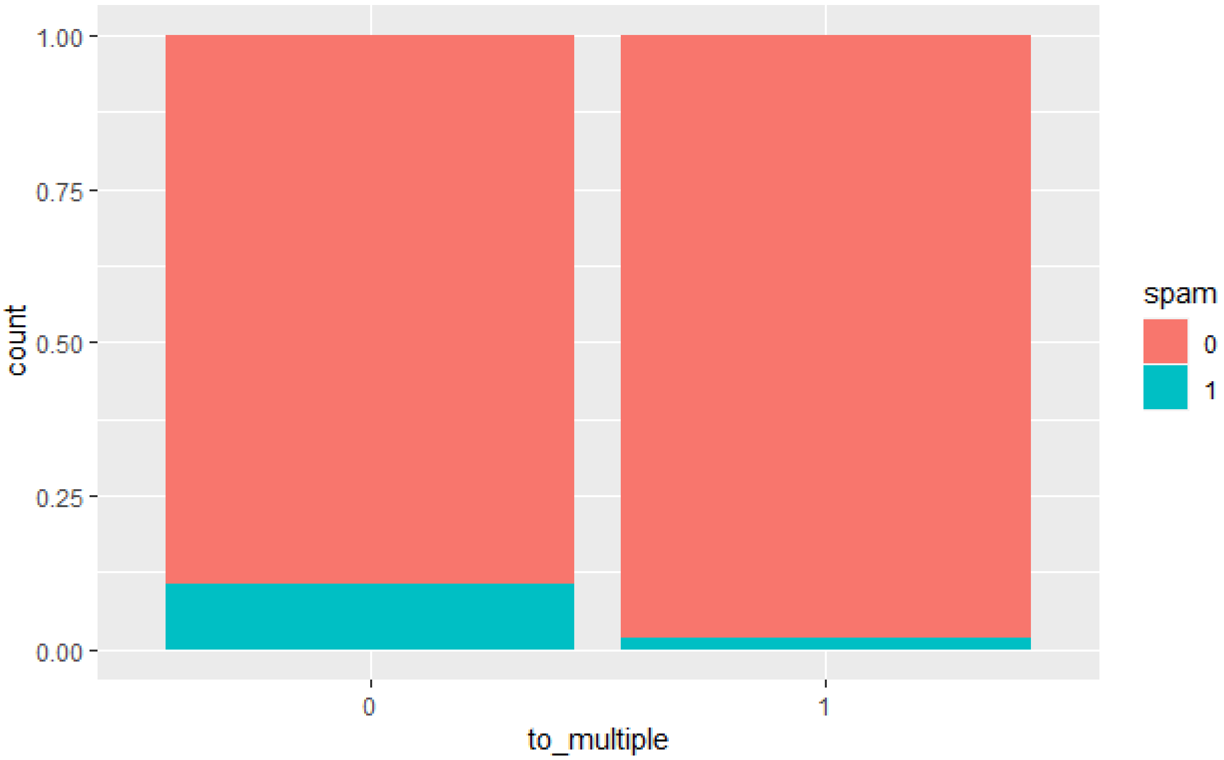
influye que el correo se envie a multiples personas a la vvez, para declararlo como spam ??

Hide

```
ctable(email$to_multiple, email$spam)
```

```
Cross-Tabulation, Row Proportions
to_multiple * spam
Data Frame: email

------------- ------ ------------- ------------ ---------------
                spam             0            1           Total
  to_multiple
            0          2946 (89.2%)  355 (10.8%)   3301 (100.0%)
            1           608 (98.1%)   12 ( 1.9%)    620 (100.0%)
        Total          3554 (90.6%)  367 ( 9.4%)   3921 (100.0%)
------------- ------ ------------- ------------ ---------------
```





from

¿DE donde proviene el correo, cuantos correos tienen un origen y cuantos no??

Hide

```
freq(email$from, style = "rmarkdown")
```

```
### Frequencies
#### email$from
**Type:** Factor

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|      **0** |    3 |   0.077 |        0.077 |   0.077 |        0.077 |
|      **1** | 3918 |  99.923 |      100.000 |  99.923 |      100.000 |
| **\<NA\>** |    0 |         |              |   0.000 |      100.000 |
|  **Total** | 3921 | 100.000 |      100.000 | 100.000 |      100.000 |
```

Hide

```
ctable(email$from, email$spam)
```

```
Cross-Tabulation, Row Proportions
from * spam
Data Frame: email

------- ------ ------------- ------------- ---------------
          spam             0             1           Total
   from
      0            0 ( 0.0%)     3 (100.0%)      3 (100.0%)
      1         3554 (90.7%)   364 (  9.3%)   3918 (100.0%)
  Total         3554 (90.6%)   367 (  9.4%)   3921 (100.0%)
------- ------ ------------- ------------- ---------------
```

sent_email

¿ del total de mails, cuantos son enviados y cuantos no?

Hide

```
freq(email$sent_email, style = "rmarkdown")
```

```
### Frequencies
#### email$sent_email
**Type:** Factor

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|      **0** | 2831 |   72.20 |        72.20 |   72.20 |        72.20 |
|      **1** | 1090 |   27.80 |       100.00 |   27.80 |       100.00 |
| **\<NA\>** |    0 |         |              |    0.00 |       100.00 |
|  **Total** | 3921 |  100.00 |       100.00 |  100.00 |       100.00 |
```

Hide

```
ctable(email$sent_email, email$spam)
```

```
Cross-Tabulation, Row Proportions
sent_email * spam
Data Frame: email

------------ ------ --------------- ------------ ---------------
               spam               0            1           Total
  sent_email
           0          2464 ( 87.0%)   367 (13.0%)   2831 (100.0%)
           1          1090 (100.0%)     0 ( 0.0%)   1090 (100.0%)
       Total          3554 ( 90.6%)   367 ( 9.4%)   3921 (100.0%)
------------ ------ --------------- ------------ ---------------
```

image - imagen

Cuantos correos del total analizados tienen imagen y cuantos no

Hide

```
freq(email$image, style = "rmarkdown")
```

```
### Frequencies
#### email$image
**Type:** Numeric

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|      **0** | 3811 |  97.195 |       97.195 |  97.195 |       97.195 |
|      **1** |   76 |   1.938 |       99.133 |   1.938 |       99.133 |
|      **2** |   17 |   0.434 |       99.566 |   0.434 |       99.566 |
|      **3** |   11 |   0.281 |       99.847 |   0.281 |       99.847 |
|      **4** |    2 |   0.051 |       99.898 |   0.051 |       99.898 |
|      **5** |    2 |   0.051 |       99.949 |   0.051 |       99.949 |
|      **9** |    1 |   0.026 |       99.974 |   0.026 |       99.974 |
|     **20** |    1 |   0.026 |      100.000 |   0.026 |      100.000 |
| **\<NA\>** |    0 |         |              |   0.000 |      100.000 |
|  **Total** | 3921 | 100.000 |      100.000 | 100.000 |      100.000 |
```

Influye el numero de imagenes a la hora de declarar un correo como spam??

Hide

```
ctable(email$image, email$spam)
```

```
Cross-Tabulation, Row Proportions
image * spam
Data Frame: email

------- ------ -------------- ------------ --------------
         spam               0            1          Total
  image
      0          3446 ( 90.4%)   365 ( 9.6%)   3811 (100.0%)
      1            74 ( 97.4%)     2 ( 2.6%)     76 (100.0%)
      2            17 (100.0%)     0 ( 0.0%)     17 (100.0%)
      3            11 (100.0%)     0 ( 0.0%)     11 (100.0%)
      4             2 (100.0%)     0 ( 0.0%)      2 (100.0%)
      5             2 (100.0%)     0 ( 0.0%)      2 (100.0%)
      9             1 (100.0%)     0 ( 0.0%)      1 (100.0%)
     20             1 (100.0%)     0 ( 0.0%)      1 (100.0%)
  Total          3554 ( 90.6%)   367 ( 9.4%)   3921 (100.0%)
------- ------ -------------- ------------ --------------
```

WINNER

Cuantas veces aparece la palabra ganador y cuantas no ¿¿?

Hide

```
summary(email$winner)
```

```
  no  yes
3857   64
```

Hide

```
freq(email$winner, style = "rmarkdown")
```

```
### Frequencies
#### email$winner
**Type:** Factor

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|     **no** | 3857 |   98.37 |        98.37 |   98.37 |        98.37 |
|    **yes** |   64 |    1.63 |       100.00 |    1.63 |       100.00 |
| **\<NA\>** |    0 |         |              |    0.00 |       100.00 |
|  **Total** | 3921 |  100.00 |       100.00 |  100.00 |       100.00 |
```

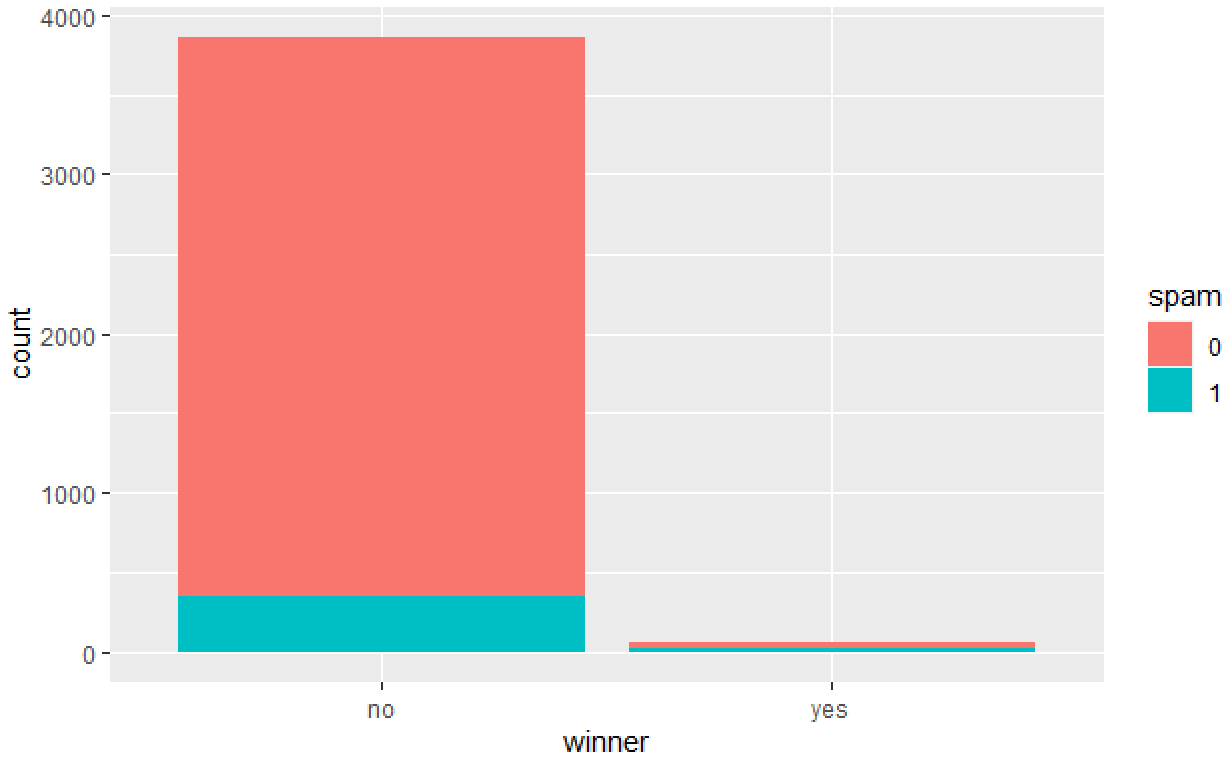
Influye la palabra winner a la hora de declarar un correo como spam

Hide

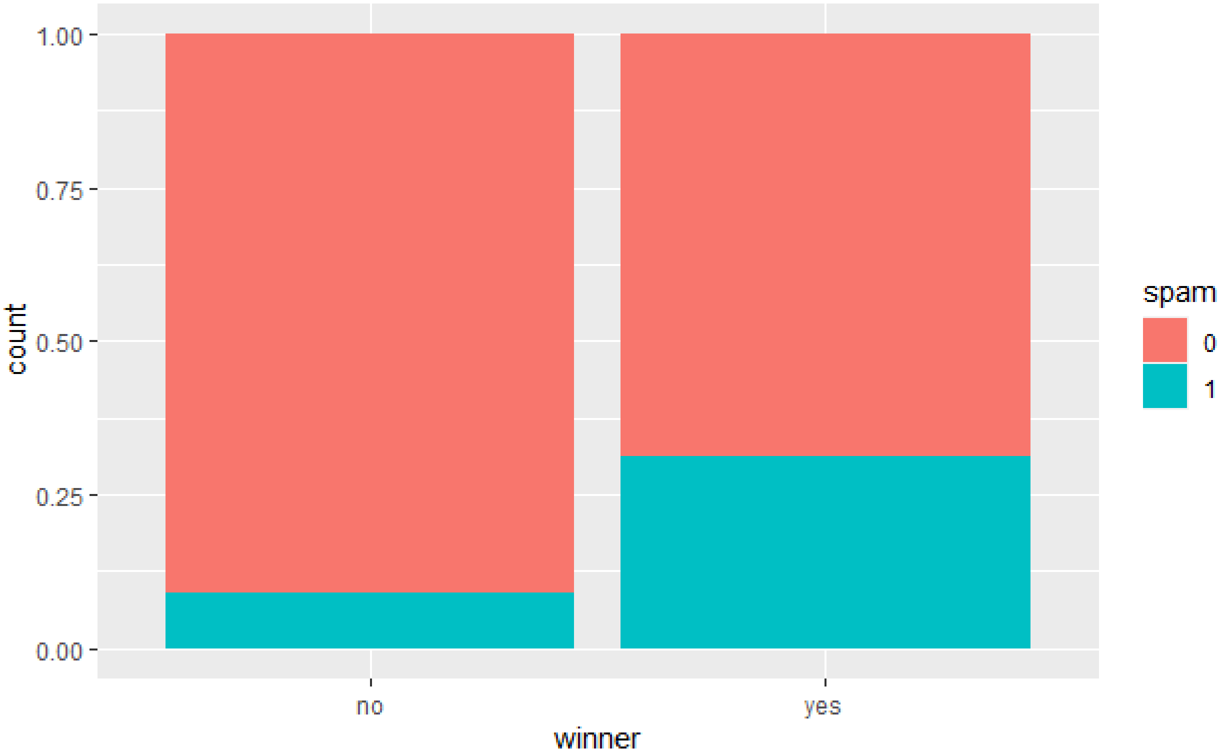
```
ctable(email$winner, email$spam)
```

```
Cross-Tabulation, Row Proportions
winner * spam
Data Frame: email

-------- ------ ------------- ------------ --------------
           spam             0            1          Total
  winner
      no          3510 (91.0%)   347 ( 9.0%)  3857 (100.0%)
     yes            44 (68.8%)    20 (31.2%)    64 (100.0%)
   Total          3554 (90.6%)   367 ( 9.4%)  3921 (100.0%)
-------- ------ ------------- ------------ --------------
```

vamos a ver un grafico de esto en valores absolutos



vamos a ver un grafico en terminos relativos



inherit - heredar

Hide

```
summary(email$inherit)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.000   0.000   0.038   0.000   9.000 
```

Hide

```
table(email$inherit)
```

```

   0    1    2    6    9 
3793  122    3    2    1 
```

Vamos a ver el numero de veces que se repite la palabra heredar(inherit)

Hide

```
freq(email$inherit, style = "rmarkdown")
```

```
### Frequencies
#### email$inherit
**Type:** Numeric

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|      **0** | 3793 |  96.736 |       96.736 |  96.736 |       96.736 |
|      **1** |  122 |   3.111 |       99.847 |   3.111 |       99.847 |
|      **2** |    3 |   0.077 |       99.923 |   0.077 |       99.923 |
|      **6** |    2 |   0.051 |       99.974 |   0.051 |       99.974 |
|      **9** |    1 |   0.026 |      100.000 |   0.026 |      100.000 |
| **\<NA\>** |    0 |         |              |   0.000 |      100.000 |
|  **Total** | 3921 | 100.000 |      100.000 | 100.000 |      100.000 |
```

Hide

```
ctable(email$inherit, email$spam)
```

```
Cross-Tabulation, Row Proportions
inherit * spam
Data Frame: email

--------- ------ -------------- ------------- --------------
            spam              0             1          Total
  inherit
        0          3440 ( 90.7%)   353 (  9.3%)  3793 (100.0%)
        1           109 ( 89.3%)    13 ( 10.7%)   122 (100.0%)
        2             3 (100.0%)     0 (  0.0%)     3 (100.0%)
        6             2 (100.0%)     0 (  0.0%)     2 (100.0%)
        9             0 (  0.0%)     1 (100.0%)     1 (100.0%)
    Total          3554 ( 90.6%)   367 (  9.4%)  3921 (100.0%)
--------- ------ -------------- ------------- --------------
```

viagra

Hide

```
unique(email$viagra)
```

```
[1] 0 8
```

Del total de correos analizados, cuantos contienen la palabra viagra y cuantos no

Hide

```
freq(email$viagra, style = "rmarkdown")
```

```
### Frequencies
#### email$viagra
**Type:** Numeric

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|      **0** | 3920 |  99.974 |       99.974 |  99.974 |       99.974 |
|      **8** |    1 |   0.026 |      100.000 |   0.026 |      100.000 |
| **\<NA\>** |    0 |         |              |   0.000 |      100.000 |
|  **Total** | 3921 | 100.000 |      100.000 | 100.000 |      100.000 |
```

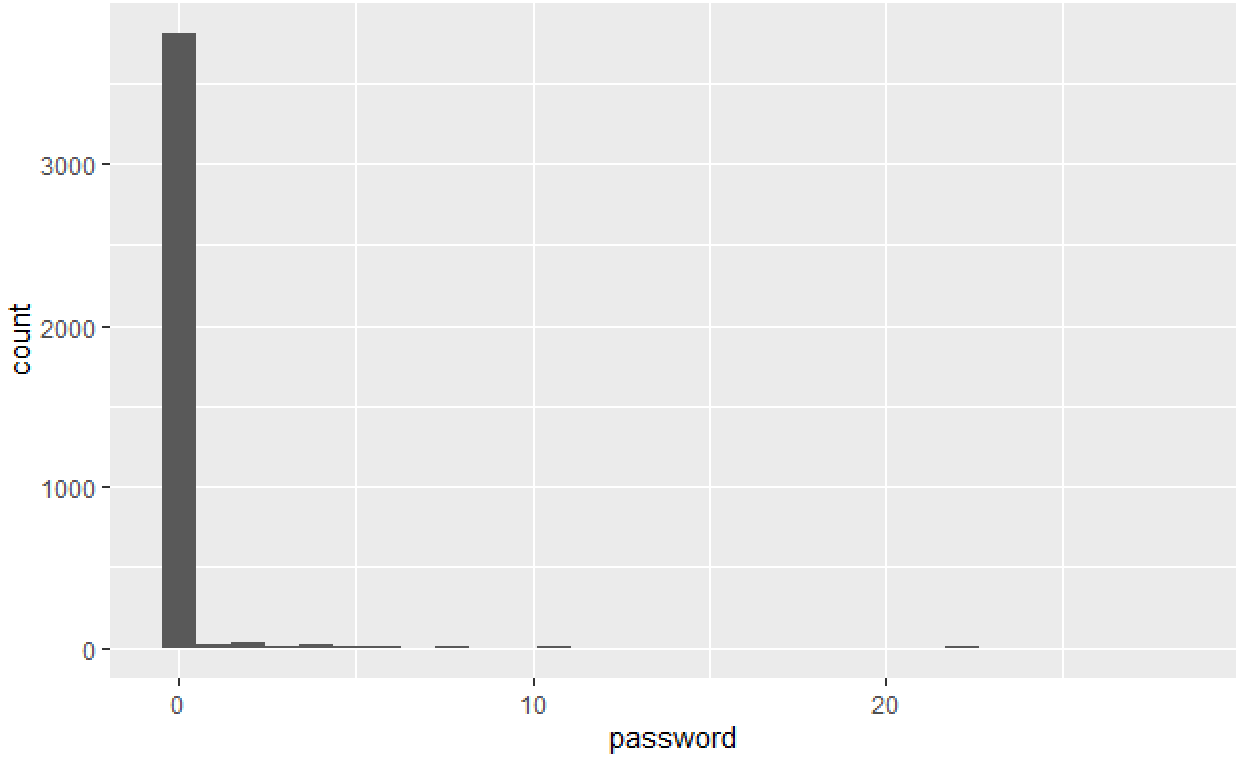
password

Cuantas veces aparece la palabra password repetida en estos correos

Hide

```
freq(email$password, style = "rmarkdown")
```

```
### Frequencies
#### email$password
**Type:** Numeric

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|      **0** | 3809 |  97.144 |       97.144 |  97.144 |       97.144 |
|      **1** |   22 |   0.561 |       97.705 |   0.561 |       97.705 |
|      **2** |   39 |   0.995 |       98.699 |   0.995 |       98.699 |
|      **3** |    8 |   0.204 |       98.903 |   0.204 |       98.903 |
|      **4** |   23 |   0.587 |       99.490 |   0.587 |       99.490 |
|      **5** |    5 |   0.128 |       99.617 |   0.128 |       99.617 |
|      **6** |    3 |   0.077 |       99.694 |   0.077 |       99.694 |
|      **8** |    5 |   0.128 |       99.821 |   0.128 |       99.821 |
|     **11** |    2 |   0.051 |       99.872 |   0.051 |       99.872 |
|     **13** |    1 |   0.026 |       99.898 |   0.026 |       99.898 |
|     **18** |    1 |   0.026 |       99.923 |   0.026 |       99.923 |
|     **22** |    2 |   0.051 |       99.974 |   0.051 |       99.974 |
|     **28** |    1 |   0.026 |      100.000 |   0.026 |      100.000 |
| **\<NA\>** |    0 |         |              |   0.000 |      100.000 |
|  **Total** | 3921 | 100.000 |      100.000 | 100.000 |      100.000 |
```



Influye la aparicion de la palabra password a la hora de determinar un correo como spam

Hide

```
ctable(email$password, email$spam)
```

```
Cross-Tabulation, Row Proportions
password * spam
Data Frame: email

---------- ------ -------------- ------------ --------------
             spam              0            1          Total
  password
         0          3446 ( 90.5%)  363 ( 9.5%)  3809 (100.0%)
         1            20 ( 90.9%)    2 ( 9.1%)    22 (100.0%)
         2            37 ( 94.9%)    2 ( 5.1%)    39 (100.0%)
         3             8 (100.0%)    0 ( 0.0%)     8 (100.0%)
         4            23 (100.0%)    0 ( 0.0%)    23 (100.0%)
         5             5 (100.0%)    0 ( 0.0%)     5 (100.0%)
         6             3 (100.0%)    0 ( 0.0%)     3 (100.0%)
         8             5 (100.0%)    0 ( 0.0%)     5 (100.0%)
        11             2 (100.0%)    0 ( 0.0%)     2 (100.0%)
        13             1 (100.0%)    0 ( 0.0%)     1 (100.0%)
        18             1 (100.0%)    0 ( 0.0%)     1 (100.0%)
        22             2 (100.0%)    0 ( 0.0%)     2 (100.0%)
        28             1 (100.0%)    0 ( 0.0%)     1 (100.0%)
     Total          3554 ( 90.6%)  367 ( 9.4%)  3921 (100.0%)
---------- ------ -------------- ------------ --------------
```

Creamos una variable dependiente binaria

s...	to_multiple	fr...	cc	sent_email	time	im...	attach	dollar	win...
<fctr>	<fctr>	<fctr>	<int>	<fctr>	<S3: POSIXct>	<dbl>	<dbl>	<dbl>	<fctr>

s…	to_multiple	fr…	cc	sent_email	time	im…	attach	dollar	win…
<fctr>	<fctr>	<fctr>	<int>	<fctr>	<S3: POSIXct>	<dbl>	<dbl>	<dbl>	<fctr>
0	0	1	0	0	2012-01-01 07:16:41	0	0	0	no
0	0	1	0	0	2012-01-01 08:03:59	0	0	0	no
0	0	1	0	0	2012-01-01 17:00:32	0	0	4	no
0	0	1	0	0	2012-01-01 10:09:49	0	0	0	no
0	0	1	0	0	2012-01-01 11:00:01	0	0	0	no
0	0	1	0	0	2012-01-01 11:04:46	0	0	0	no

6 rows | 1-10 of 22 columns

Hide

```
ctable(email$password_binary, email$spam)
```

```
Cross-Tabulation, Row Proportions
password_binary * spam
Data Frame: email

----------------- ------ ------------- ------------ ---------------
                    spam             0            1           Total
  password_binary
                0          3446 (90.5%)   363 ( 9.5%)   3809 (100.0%)
                1           108 (96.4%)     4 ( 3.6%)    112 (100.0%)
            Total          3554 (90.6%)   367 ( 9.4%)   3921 (100.0%)
----------------- ------ ------------- ------------ ---------------
```

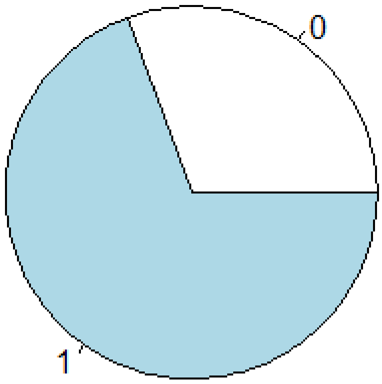
format

Cuantos correos tienen formato y cuantos no lo tienen

Hide

```
freq(email$format, style = "rmarkdown")
```

```
### Frequencies
#### email$format
**Type:** Factor

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|      **0** | 1195 |   30.48 |        30.48 |   30.48 |        30.48 |
|      **1** | 2726 |   69.52 |       100.00 |   69.52 |       100.00 |
| **\<NA\>** |    0 |         |              |    0.00 |       100.00 |
|  **Total** | 3921 |  100.00 |       100.00 |  100.00 |       100.00 |
```

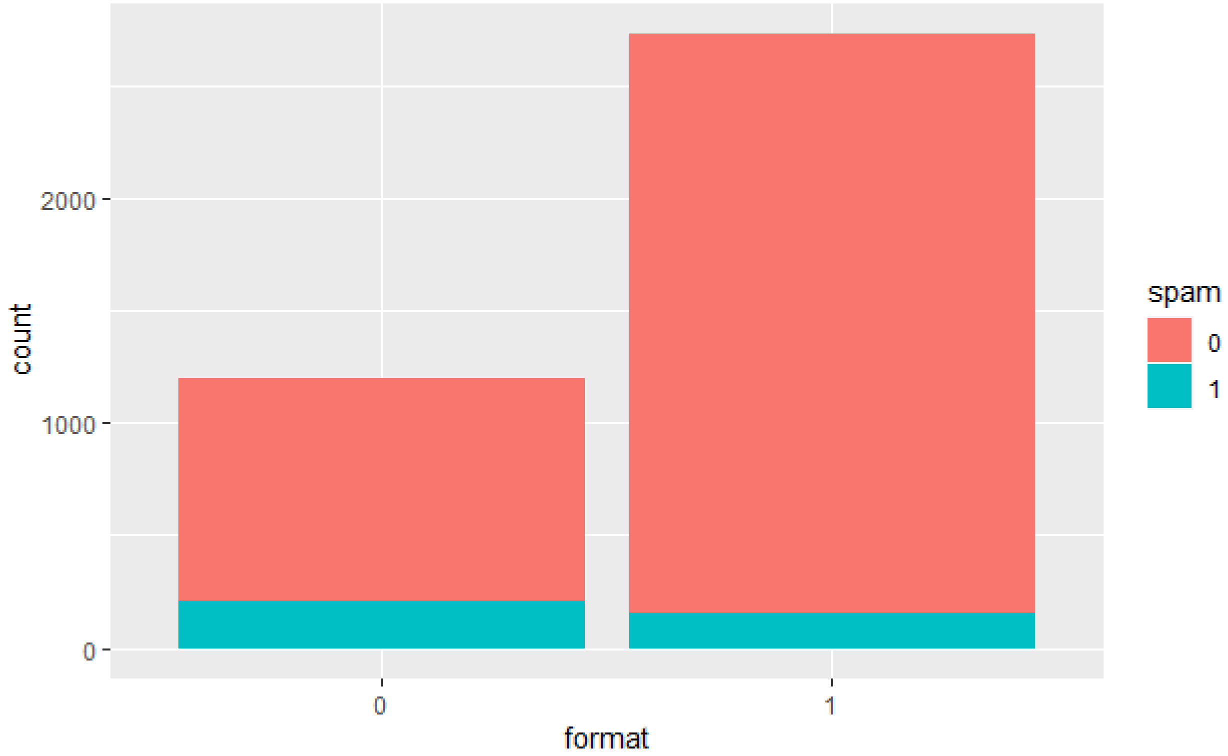


Influye que los correos tengan formato a la hora de declararlos como spam??

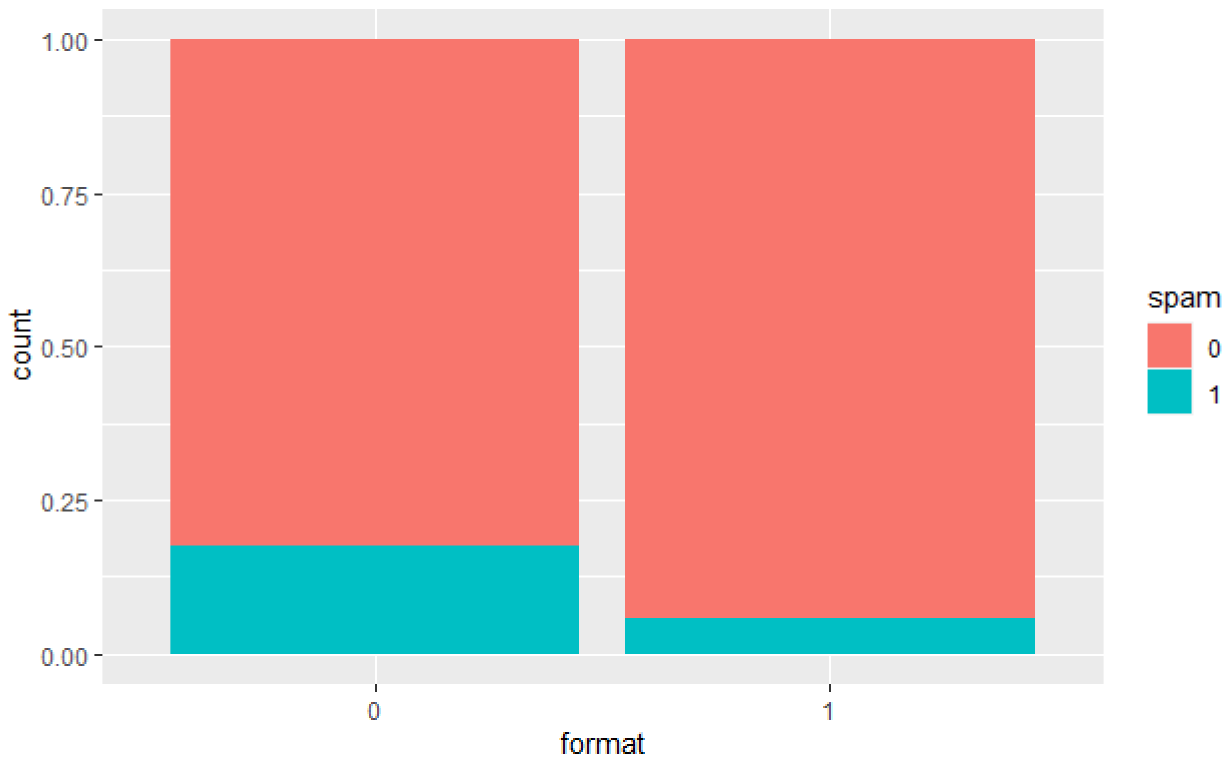
Hide

```
ctable(email$format, email$spam)
```

```
Cross-Tabulation, Row Proportions
format * spam
Data Frame: email

-------- ------ -------------- ------------- ---------------
           spam              0             1           Total
  format                                                    
       0           986 (82.5%)   209 (17.5%)   1195 (100.0%)
       1          2568 (94.2%)   158 ( 5.8%)   2726 (100.0%)
   Total          3554 (90.6%)   367 ( 9.4%)   3921 (100.0%)
-------- ------ -------------- ------------- ---------------
```



Vamos a ver el grafico en terminos relativos



re_subj

Hide

```
summary(email$re_subj)
```

```
   0    1 
2896 1025 
```

Hide

```
unique(email$re_subj)
```

```
[1] 0 1
Levels: 0 1
```

Cuantas veces tenemos re_subj y cuantos no en estos correos

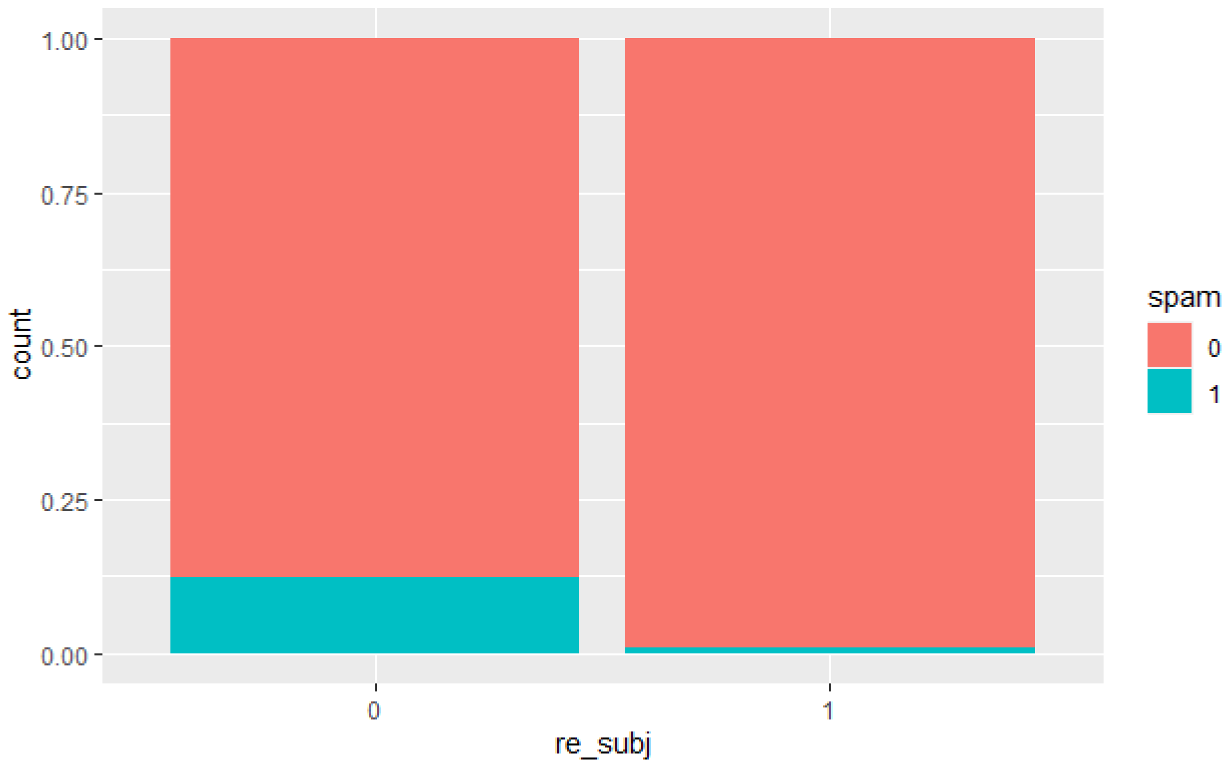
Hide

```
ctable(email$re_subj, email$spam)
```

```
Cross-Tabulation, Row Proportions
re_subj * spam
Data Frame: email

--------- ------ ------------- ------------ --------------
            spam             0            1          Total
  re_subj
        0          2537 (87.6%)   359 (12.4%)   2896 (100.0%)
        1          1017 (99.2%)     8 ( 0.8%)   1025 (100.0%)
    Total          3554 (90.6%)   367 ( 9.4%)   3921 (100.0%)
--------- ------ ------------- ------------ --------------
```

Hide

```
ggplot(email, aes(re_subj, fill = spam)) + geom_bar(position = "fill")
```



Son declarados como spam mayoritariamente los correos que NO tiene r_subj, en concreto un 12%.

exclaim_subj

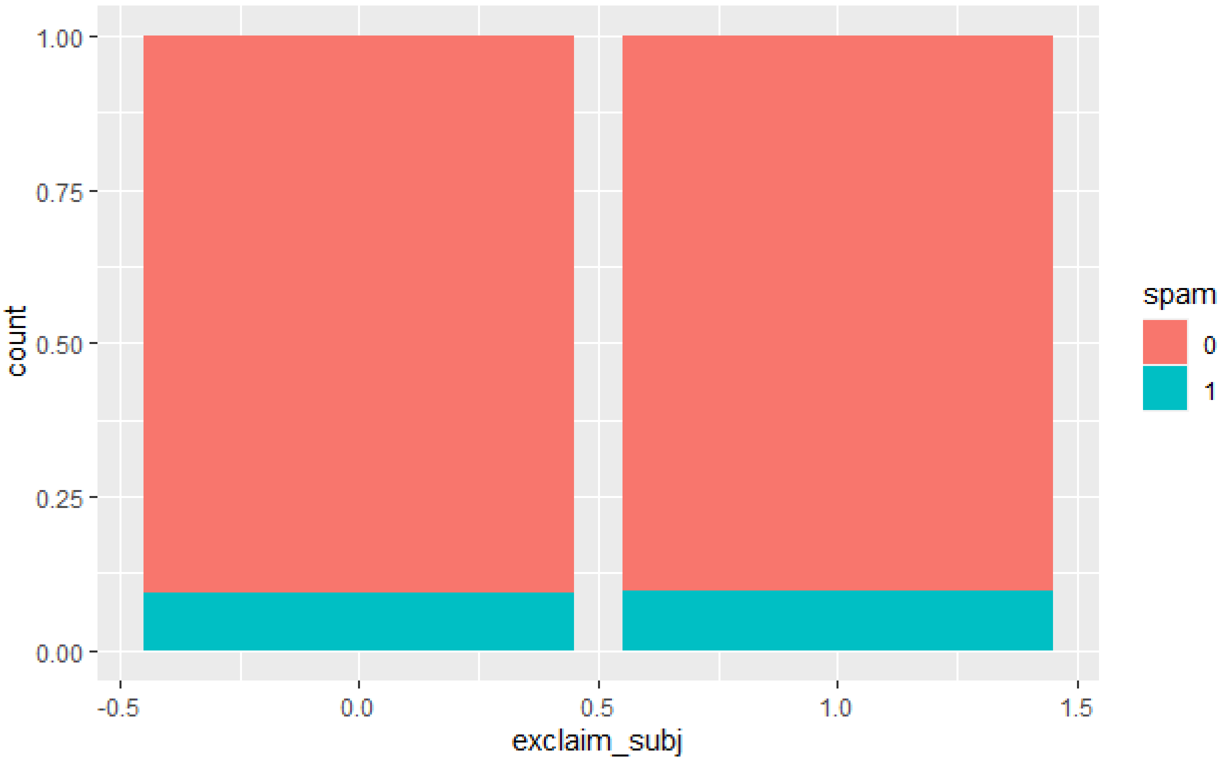
Hide

```
ctable(email$exclaim_subj, email$spam)
```

```
Cross-Tabulation, Row Proportions
exclaim_subj * spam
Data Frame: email

-------------- ------ ------------- ------------ ---------------
                 spam             0            1           Total
  exclaim_subj
             0          3269 (90.7%)   337 ( 9.3%)   3606 (100.0%)
             1           285 (90.5%)    30 ( 9.5%)    315 (100.0%)
         Total          3554 (90.6%)   367 ( 9.4%)   3921 (100.0%)
-------------- ------ ------------- ------------ ---------------
```

Hide

```
ggplot(email, aes(exclaim_subj, fill = spam)) + geom_bar(position = "fill")
```



Parece que los correos exclaim y los que no lo tienen, son declarados como spam en la misma proporcion (9.3 - 9.5), por lo que no parece que sea una variable determinante a la hora de diferenciar entre spam y no spam

urgent_subj

Hide

```
freq(email$urgent_subj, style = "rmarkdown")
```

```
### Frequencies
#### email$urgent_subj
**Type:** Factor

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|      **0** | 3914 |   99.82 |        99.82 |   99.82 |        99.82 |
|      **1** |    7 |    0.18 |       100.00 |    0.18 |       100.00 |
| **\<NA\>** |    0 |         |              |    0.00 |       100.00 |
|  **Total** | 3921 |  100.00 |       100.00 |  100.00 |       100.00 |
```

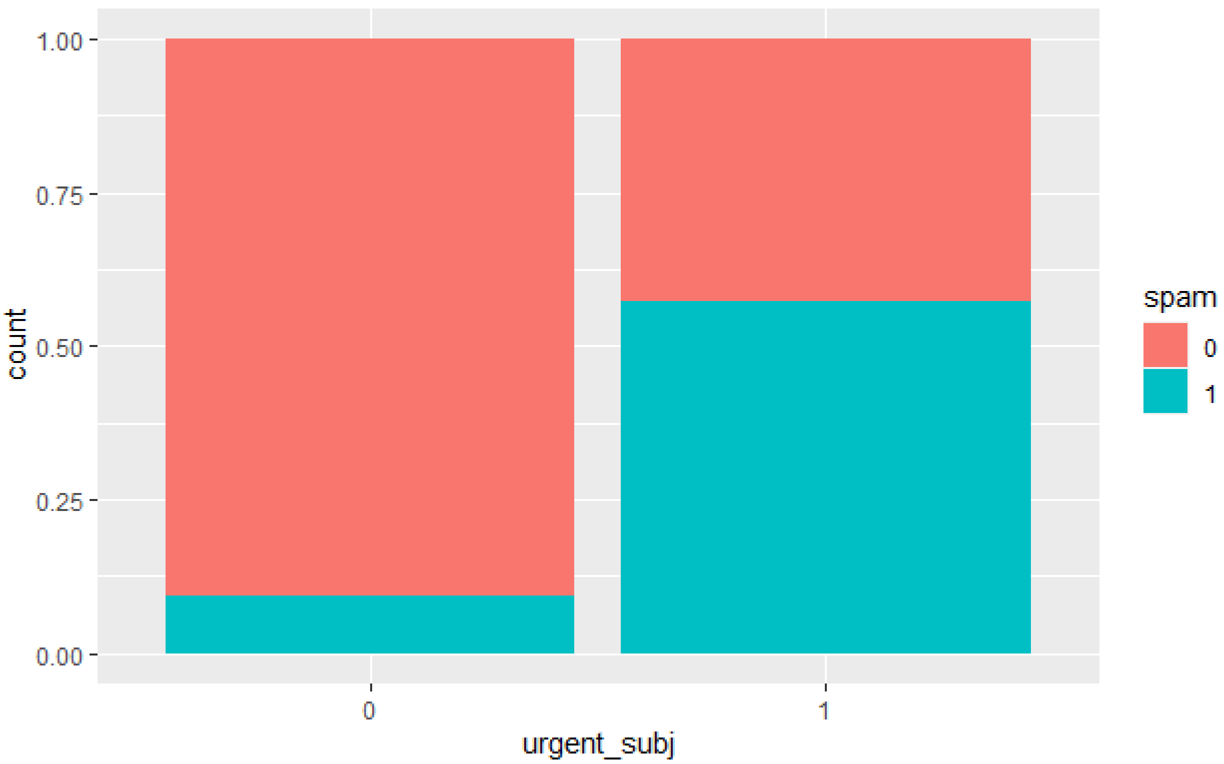
Hide

```
ctable(email$urgent_subj, email$spam)
```

```
Cross-Tabulation, Row Proportions
urgent_subj * spam
Data Frame: email

------------- ------ ------------- ------------ --------------
                spam             0            1          Total
  urgent_subj
            0          3551 (90.7%)  363 ( 9.3%)  3914 (100.0%)
            1             3 (42.9%)    4 (57.1%)     7 (100.0%)
        Total          3554 (90.6%)  367 ( 9.4%)  3921 (100.0%)
------------- ------ ------------- ------------ --------------
```

Hide

```
ggplot(email, aes(urgent_subj, fill = spam)) + geom_bar(position = "fill")
```



Solo tenemos 7 correos que son urgent_subj, lo que supone un 0.18%. Eso si el 57% de estos(4) son declarados spam, frente al 43%(3) que no lo son. El porcentaje de casos no urgent_subj, no es muy representativo, pero vamos a mantener esta vaariable.

time

De la variable time vamos a obtener los meses, por si hubiera algun mes en el que se declaren mas correos como spam.

Hide

```
class(email$time)
```

```
[1] "POSIXct" "POSIXt"
```

Hide

```
class(email$time)
```

```
[1] "Date"
```

Hide

```

email$mes <-  format(as.Date(email$time), "%m")

# ya tenemos una nueva colunmna con los meses en los que se envian los correos
```

Vamos a ver en que meses se envian mas correos y si estos influyen a la hora de establecer un correo como spam

Hide

```
freq(email$mes, style = "rmarkdown")
```

```
### Frequencies
#### email$mes
**Type:** Character

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|     **01** | 1300 |   33.15 |        33.15 |   33.15 |        33.15 |
|     **02** | 1326 |   33.82 |        66.97 |   33.82 |        66.97 |
|     **03** | 1291 |   32.93 |        99.90 |   32.93 |        99.90 |
|     **04** |    4 |    0.10 |       100.00 |    0.10 |       100.00 |
| **\<NA\>** |    0 |         |              |    0.00 |       100.00 |
|  **Total** | 3921 |  100.00 |       100.00 |  100.00 |       100.00 |
```

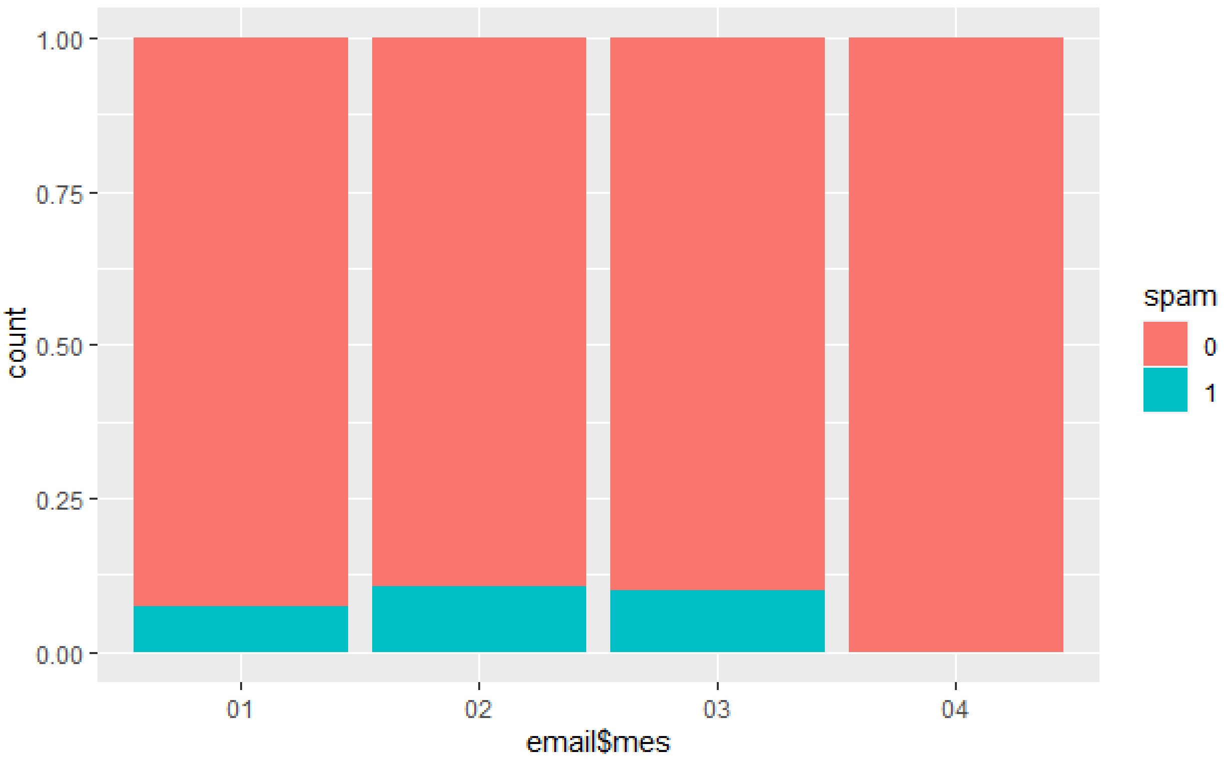
Hide

```

ctable(email$mes, email$spam)
```

```
Cross-Tabulation, Row Proportions
mes * spam
Data Frame: email

------- ------ -------------- ------------ --------------
          spam              0            1          Total
    mes
     01          1206 ( 92.8%)    94 ( 7.2%)   1300 (100.0%)
     02          1183 ( 89.2%)   143 (10.8%)   1326 (100.0%)
     03          1161 ( 89.9%)   130 (10.1%)   1291 (100.0%)
     04             4 (100.0%)     0 ( 0.0%)      4 (100.0%)
  Total          3554 ( 90.6%)   367 ( 9.4%)   3921 (100.0%)
------- ------ -------------- ------------ --------------
```

Hide

```
ggplot(email, aes(email$mes, fill = spam)) + geom_bar(position = "fill")
```



En este conjunto de datos solo tenemos informacion de correos enviados de Enero - Abril en proporcion de un 33% los tres primeros meses y de un 1% el mes de Abril.

Los mails declarados como spam se reparten entre los 3 primeros meses, de forma bastante homogenea, no parece muy significativa esta variable.

number

¿Con que frecuencia aparecen los numeros grandes, pequeños o no hay numeros, en estos correos?

Hide

```
library(summarytools)
```

```
Registered S3 method overwritten by 'pryr':
  method      from
  print.bytes Rcpp
```

Hide

```
summary(email$number)
```

```
 none small   big 
  549  2827   545 
```

Hide

```
freq(email$number, style= "rmarkdown")
```

```
### Frequencies
#### email$number
**Type:** Factor

|       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
|-----------:|-----:|--------:|-------------:|--------:|-------------:|
|   **none** |  549 |   14.00 |        14.00 |   14.00 |        14.00 |
|  **small** | 2827 |   72.10 |        86.10 |   72.10 |        86.10 |
|    **big** |  545 |   13.90 |       100.00 |   13.90 |       100.00 |
| **\<NA\>** |    0 |         |              |    0.00 |       100.00 |
|  **Total** | 3921 |  100.00 |       100.00 |  100.00 |       100.00 |
```

Influye que haya numeros o su tamaño a la hora de catalogar el correo como spam??

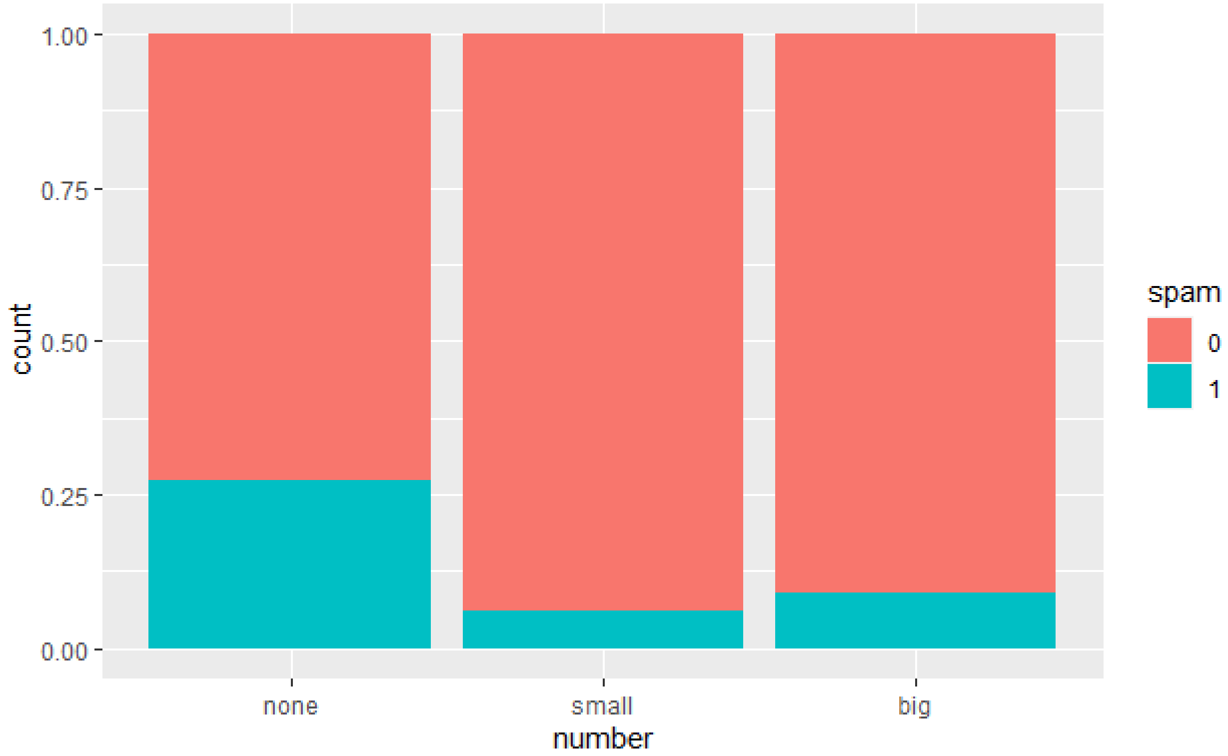
Hide

```
library(ggplot2)

ctable(email$number, email$spam)
```

```
Cross-Tabulation, Row Proportions  
number * spam  
Data Frame: email  

-------- ------ ------------- ------------ --------------
           spam             0            1          Total
  number                                                 
    none           400 (72.9%)   149 (27.1%)    549 (100.0%)
   small          2659 (94.1%)   168 ( 5.9%)   2827 (100.0%)
     big           495 (90.8%)    50 ( 9.2%)    545 (100.0%)
   Total          3554 (90.6%)   367 ( 9.4%)   3921 (100.0%)
-------- ------ ------------- ------------ --------------
```

Hide

```
ggplot(email, aes(x = number, fill = spam)) + geom_bar(position = "fill")
```

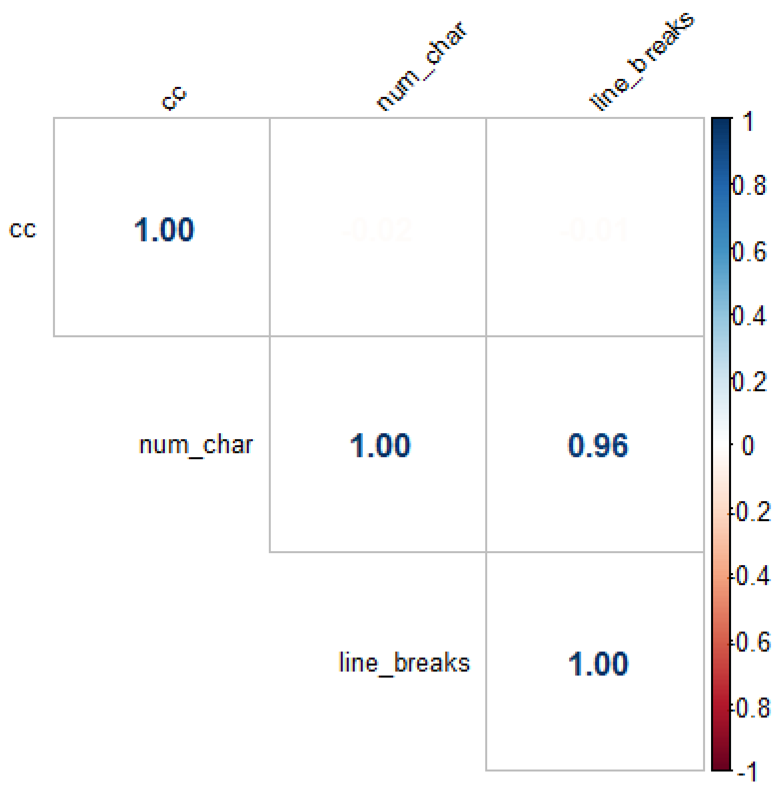


Podemos observar que cuando no hay numeros en los correos, la posibilidad de catalogar el correo como spam se dispara a un 27%, mientras que cuando los hay pequeños es de un 6% y grandes de un 9%

Analisis bidimensional

Vamos a analizar si las variables cuantitativas estan correlacionadas entre ellas, de ser asi, habria que eliminarlas:

cc, num_chart, line_breaks



El numero de caracteres del correo y el de saltos de linea estan altamente correlacionados (0.96), vamos a mantenerlos de momento

Conclusion.

1 - Variables a eliminar: time, exclaim_subj, viagra,

2 - variables que nos quedamos: password_binary, dollar_binnes, attach_binary, to_multiple, from, cc, sent_email, image, winner, inherit, num_chart, line_breaks, format, re_subj urgent_subj, exclaim_mess, number

3 - Se han modificado 3 variables (password_binary, dollar_binnes, attach_binary), dejandolas como binarias, aunque deberian haber sido mas…

vamos a eliminar las variables que no correspondan:

Hide

```
email <- subset( email, select = -c(time, exclaim_subj, viagra,password, dollar, attach ) )


head(email)
```

Vamos a convertir a las variables categoricas en variables factores ordenadas

Hide

```
email$password_binary = factor(email$password_binary= order = TRUE, levels = c(0, 1))
```

```
Error: inesperado '=' in "email$password_binary = factor(email$password_binary="
```

Hide

```
NA
```

Aqui terminamos el analisis descriptivo de este dataset y tenemos que comenzar con el analisis predictivo.

Hide

```
modelo = glm(spam~ password_binary, dollar_binnes, attach_binary, data = email)
```

```
Error in glm(spam ~ password_binary, dollar_binnes, attach_binary, data = email) :
  objeto 'dollar_binnes' no encontrado
```