Práctica Evaluable

Javier de Castro Rodríguez

2023-02-26

En la detección de Spam se utilizan con frecuencia técnicas de machine learning para mejorar los índices de detección de correos no deseados. En el dataset adjunto, se han seleccionado para cada mensaje una serie de términos clave que suelen aparecer con frecuencia en los mensajes spam.

Posteriormente, se ha realizado una codificación vectorial de los correos electrónicos considerando esos términos clave. Para cada correo disponemos de la clasificación por parte de los expertos humanos. Se pide realizar las siguientes tareas:

Accedemos al directorio de trabajo:

```
currentDir <- getwd()
parentPath <- dirname(currentDir)
```

Cargamos librerías e instalamos paquetes en caso de no tenerlos ya instalados:

```
libs <- c("tidyverse","skimr","caret", "ROCR", "dplyr", "knitr","ggplot2",
"DataExplorer", "corrplot", "Hmisc", "Matrix","tm", "e1071",
"gridExtra","reshape2", "factoextra", "cluster", "caTools",
"tinytex","FactoMineR", "mice", "kohonen", "RColorBrewer", "RCurl", "plot3D")

for (i in libs){
    print(i)
    if(!require(i, character.only = TRUE))
      { install.packages(i, dependencies=TRUE); library(i) }
  }

## [1] "tidyverse"

## Loading required package: tidyverse

## Warning: package 'tidyverse' was built under R version 4.2.2

## Warning: package 'ggplot2' was built under R version 4.2.2

## Warning: package 'tidyr' was built under R version 4.2.2

## Warning: package 'readr' was built under R version 4.2.2

## Warning: package 'purrr' was built under R version 4.2.2

## Warning: package 'dplyr' was built under R version 4.2.2
```

```
## Warning: package 'stringr' was built under R version 4.2.2

## Warning: package 'forcats' was built under R version 4.2.2

## Warning: package 'lubridate' was built under R version 4.2.2

## ── Attaching core tidyverse packages ──────────────────────────────────────
tidyverse 2.0.0 ──
## ✔ dplyr     1.1.0     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.1     ✔ tibble    3.1.8
## ✔ lubridate 1.9.2     ✔ tidyr     1.3.0
## ✔ purrr     1.0.1

## ── Conflicts ────────────────────────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force
all conflicts to become errors

## [1] "skimr"

## Loading required package: skimr

## Warning: package 'skimr' was built under R version 4.2.2

## [1] "caret"

## Loading required package: caret

## Warning: package 'caret' was built under R version 4.2.2

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 4.2.2

##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift

## [1] "ROCR"

## Loading required package: ROCR

## Warning: package 'ROCR' was built under R version 4.2.2

## [1] "dplyr"
## [1] "knitr"

## Loading required package: knitr
```

```
## Warning: package 'knitr' was built under R version 4.2.2

## [1] "ggplot2"
## [1] "DataExplorer"

## Loading required package: DataExplorer

## Warning: package 'DataExplorer' was built under R version 4.2.2

## [1] "corrplot"

## Loading required package: corrplot

## Warning: package 'corrplot' was built under R version 4.2.2

## corrplot 0.92 loaded

## [1] "Hmisc"

## Loading required package: Hmisc

## Warning: package 'Hmisc' was built under R version 4.2.2

## Loading required package: survival
##
## Attaching package: 'survival'
##
## The following object is masked from 'package:caret':
##
##     cluster
##
## Loading required package: Formula
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
##
## The following objects are masked from 'package:base':
##
##     format.pval, units

## [1] "Matrix"

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 4.2.2

##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
```

```
##
##     expand, pack, unpack

## [1] "tm"

## Loading required package: tm

## Warning: package 'tm' was built under R version 4.2.2

## Loading required package: NLP
##
## Attaching package: 'NLP'
##
## The following object is masked from 'package:ggplot2':
##
##     annotate

## [1] "e1071"

## Loading required package: e1071
##
## Attaching package: 'e1071'
##
## The following object is masked from 'package:Hmisc':
##
##     impute

## [1] "gridExtra"

## Loading required package: gridExtra

## Warning: package 'gridExtra' was built under R version 4.2.2

##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine

## [1] "reshape2"

## Loading required package: reshape2

## Warning: package 'reshape2' was built under R version 4.2.2

##
## Attaching package: 'reshape2'
##
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
## [1] "factoextra"

## Loading required package: factoextra

## Warning: package 'factoextra' was built under R version 4.2.2

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

## [1] "cluster"

## Loading required package: cluster

## Warning: package 'cluster' was built under R version 4.2.2

## [1] "caTools"

## Loading required package: caTools

## Warning: package 'caTools' was built under R version 4.2.2

## [1] "tinytex"

## Loading required package: tinytex

## Warning: package 'tinytex' was built under R version 4.2.2

## [1] "FactoMineR"

## Loading required package: FactoMineR

## Warning: package 'FactoMineR' was built under R version 4.2.2

## [1] "mice"

## Loading required package: mice

## Warning: package 'mice' was built under R version 4.2.2

##
## Attaching package: 'mice'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     cbind, rbind

## [1] "kohonen"

## Loading required package: kohonen

## Warning: package 'kohonen' was built under R version 4.2.2
```

```
##
## Attaching package: 'kohonen'
##
## The following object is masked from 'package:purrr':
##
##     map

## [1] "RColorBrewer"

## Loading required package: RColorBrewer

## [1] "RCurl"

## Loading required package: RCurl

## Warning: package 'RCurl' was built under R version 4.2.2

##
## Attaching package: 'RCurl'
##
## The following object is masked from 'package:mice':
##
##     complete
##
## The following object is masked from 'package:tidyr':
##
##     complete

## [1] "plot3D"

## Loading required package: plot3D

## Warning: package 'plot3D' was built under R version 4.2.2
```

En el análisis previo del archivo *spam.csv*, comprobamos que las variables no tienen nombre y que corresponden con el archivo *nombres*.

Cargamos dataset y archivo "nombres" para modificar columnas

```
nombre_variables <- read.table("./nombre_variables", header = FALSE, sep =
"")
nombres <- c("make",
"address","all","3d","our","over","remove","internet","order","mail","receive
","will","people","report","addresses","free","business","email", "you",
"credit", "your","font", "000", "money", "hp", "hpl", "george", "650", "lab",
"labs", "telnet", "857","data","415", "85", "technology", "1999", "parts",
"pm", "direct", "cs", "meeting","original", "project", "re", "edu", "table",
"conference", ";", "(", "[","!", "$",
"#", "cap_run_length_average",
"cap_run_length_longest","cap_run_length_total", "clase")
```

Cargamos dataset:

```
spam <-read.table("./spam.data", col.names = nombres)
head(spam)

##   make address  all X3d  our over remove internet order mail receive will
## 1 0.00    0.64 0.64   0 0.32 0.00   0.00     0.00  0.00 0.00    0.00 0.64
## 2 0.21    0.28 0.50   0 0.14 0.28   0.21     0.07  0.00 0.94    0.21 0.79
## 3 0.06    0.00 0.71   0 1.23 0.19   0.19     0.12  0.64 0.25    0.38 0.45
## 4 0.00    0.00 0.00   0 0.63 0.00   0.31     0.63  0.31 0.63    0.31 0.31
## 5 0.00    0.00 0.00   0 0.63 0.00   0.31     0.63  0.31 0.63    0.31 0.31
## 6 0.00    0.00 0.00   0 1.85 0.00   0.00     1.85  0.00 0.00    0.00 0.00
##   people report addresses free business email  you credit your font X000
money
## 1   0.00   0.00      0.00 0.32     0.00  1.29 1.93   0.00 0.96    0 0.00
0.00
## 2   0.65   0.21      0.14 0.14     0.07  0.28 3.47   0.00 1.59    0 0.43
0.43
## 3   0.12   0.00      1.75 0.06     0.06  1.03 1.36   0.32 0.51    0 1.16
0.06
## 4   0.31   0.00      0.00 0.31     0.00  0.00 3.18   0.00 0.31    0 0.00
0.00
## 5   0.31   0.00      0.00 0.31     0.00  0.00 3.18   0.00 0.31    0 0.00
0.00
## 6   0.00   0.00      0.00 0.00     0.00  0.00 0.00   0.00 0.00    0 0.00
0.00
##   hp hpl george X650 lab labs telnet X857 data X415 X85 technology X1999
parts
## 1  0   0      0    0   0    0      0    0    0    0   0          0  0.00
0
## 2  0   0      0    0   0    0      0    0    0    0   0          0  0.07
0
## 3  0   0      0    0   0    0      0    0    0    0   0          0  0.00
0
## 4  0   0      0    0   0    0      0    0    0    0   0          0  0.00
0
## 5  0   0      0    0   0    0      0    0    0    0   0          0  0.00
0
## 6  0   0      0    0   0    0      0    0    0    0   0          0  0.00
0
##   pm direct cs meeting original project   re  edu table conference   X.
X..1
## 1  0   0.00  0       0     0.00       0 0.00 0.00     0          0 0.00
0.000
## 2  0   0.00  0       0     0.00       0 0.00 0.00     0          0 0.00
0.132
## 3  0   0.06  0       0     0.12       0 0.06 0.06     0          0 0.01
0.143
```

```
## 4  0   0.00  0       0     0.00       0 0.00 0.00     0          0 0.00
0.137
## 5  0   0.00  0       0     0.00       0 0.00 0.00     0          0 0.00
0.135
## 6  0   0.00  0       0     0.00       0 0.00 0.00     0          0 0.00
0.223
##   X..2  X..3  X..4  X..5 cap_run_length_average cap_run_length_longest
## 1    0 0.778 0.000 0.000                  3.756                     61
## 2    0 0.372 0.180 0.048                  5.114                    101
## 3    0 0.276 0.184 0.010                  9.821                    485
## 4    0 0.137 0.000 0.000                  3.537                     40
## 5    0 0.135 0.000 0.000                  3.537                     40
## 6    0 0.000 0.000 0.000                  3.000                     15
##   cap_run_length_total clase
## 1                  278     1
## 2                 1028     1
## 3                 2259     1
## 4                  191     1
## 5                  191     1
## 6                   54     1
```

En el **primer apartado** de la práctica se nos pide sustituir un 2% de valores por NAs de manera aleatoria, imputando a continuación los valores faltantes.

```
random_NA <- as.data.frame(lapply(spam, \(x) replace(x, sample(length(x),
.02*length(x)), NA)))

sum(is.na(random_NA)) #total datos faltantes

## [1] 5336
```

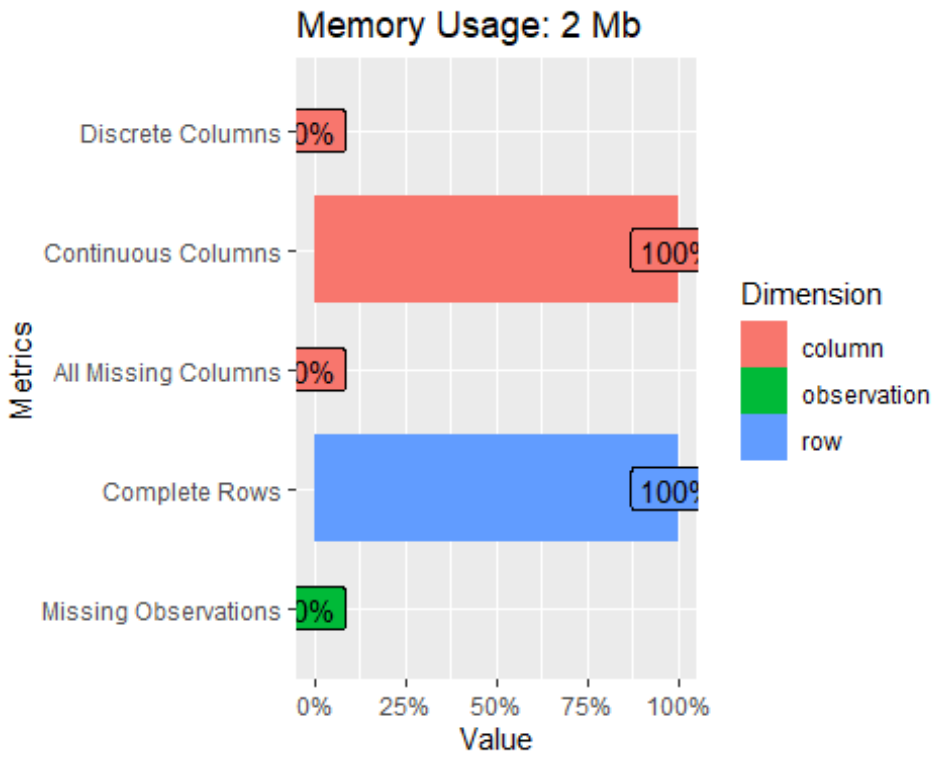
Para la imputación de valores faltantes he optado por el método "pmm" *(Predictive Mean Matching)* del paquete "mice". Nuestros datos son completamente al azar *MCAR* ya que los hemos generado de una manera aleatoria. El método funciona para variables numéricas y continuas.

Calcula el valor predicho utilizando un modelo de regresión y elige los 5 elementos más cercanos al valor predicho (por Distancia euclidiana).

```
set.seed(1234)
imputed_data <- mice(random_NA, m= 3, maxit = 10, method = "pmm", print =
FALSE)

## Warning: Number of logged events: 58

spam <- mice::complete(imputed_data)
```

En el **segundo apartado** se nos pide una análisis exploratorio.

```
plot_intro(spam) #Comprobamos que ya no tenemos faltantes
```



Vemos dimensiones y características de nuestro dataset:

```
dim(spam)

## [1] 4601   58

glimpse(spam)

## Rows: 4,601
## Columns: 58
## $ make                    <dbl> 0.00, 0.21, 0.06, 0.00, 0.00, 0.00, 0.00,
0.00,...
## $ address                 <dbl> 0.64, 0.28, 0.00, 0.00, 0.00, 0.00, 0.00,
0.00,...
## $ all                     <dbl> 0.64, 0.50, 0.71, 0.00, 0.00, 0.00, 0.00,
0.00,...
## $ X3d                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,...
## $ our                     <dbl> 0.32, 0.14, 1.23, 0.63, 0.63, 1.85, 1.92,
1.88,...
## $ over                    <dbl> 0.00, 0.28, 0.19, 0.00, 0.00, 0.00, 0.00,
0.00,...
## $ remove                  <dbl> 0.00, 0.21, 0.19, 0.31, 0.31, 0.00, 0.00,
0.00,...
```

```
## $ internet                <dbl> 0.00, 0.07, 0.12, 0.63, 0.63, 1.85, 0.00,
1.88,…
## $ order                   <dbl> 0.00, 0.00, 0.64, 0.31, 0.31, 0.00, 0.00,
0.00,…
## $ mail                    <dbl> 0.00, 0.94, 0.25, 0.63, 0.63, 0.00, 0.64,
0.00,…
## $ receive                 <dbl> 0.00, 0.21, 0.38, 0.31, 0.31, 0.00, 0.96,
0.00,…
## $ will                    <dbl> 0.64, 0.79, 0.45, 0.31, 0.31, 0.00, 1.28,
0.00,…
## $ people                  <dbl> 0.00, 0.65, 0.12, 0.31, 0.31, 0.00, 0.00,
0.00,…
## $ report                  <dbl> 0.00, 0.21, 0.70, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ addresses               <dbl> 0.00, 0.14, 1.75, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ free                    <dbl> 0.32, 0.14, 0.06, 0.31, 0.31, 0.00, 0.96,
0.00,…
## $ business                <dbl> 0.00, 0.07, 0.06, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ email                   <dbl> 1.29, 0.28, 1.03, 0.00, 0.00, 0.00, 0.32,
0.00,…
## $ you                     <dbl> 1.93, 3.47, 1.36, 3.18, 3.18, 0.00, 3.85,
0.00,…
## $ credit                  <dbl> 0.00, 0.00, 0.32, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ your                    <dbl> 0.96, 1.59, 0.51, 0.31, 0.31, 0.00, 0.64,
0.00,…
## $ font                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ X000                    <dbl> 0.00, 0.43, 1.16, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ money                   <dbl> 0.00, 0.43, 0.06, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ hp                      <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 1.99, 0.00,
0.00,…
## $ hpl                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ george                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ X650                    <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ lab                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ labs                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ telnet                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ X857                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
```

```
## $ data                    <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ X415                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ X85                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ technology              <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ X1999                   <dbl> 0.00, 0.07, 0.00, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ parts                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ pm                      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ direct                  <dbl> 0.00, 0.00, 0.06, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ cs                      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ meeting                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ original                <dbl> 0.00, 0.00, 0.12, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ project                 <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ re                      <dbl> 0.00, 0.00, 0.06, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ edu                     <dbl> 0.00, 0.00, 0.06, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ table                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ conference              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ X.                      <dbl> 0.000, 0.000, 0.010, 0.000, 0.000, 0.000,
0.000…
## $ X..1                    <dbl> 0.000, 0.132, 0.143, 0.137, 0.135, 0.223,
0.054…
## $ X..2                    <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 0.000,
0.000…
## $ X..3                    <dbl> 0.778, 0.372, 0.276, 0.137, 0.135, 0.000,
0.164…
## $ X..4                    <dbl> 0.000, 0.180, 0.184, 0.000, 0.000, 0.000,
0.054…
## $ X..5                    <dbl> 0.000, 0.048, 0.010, 0.000, 0.000, 0.000,
0.000…
## $ cap_run_length_average <dbl> 3.756, 5.114, 9.821, 3.537, 3.537, 3.000,
1.671…
## $ cap_run_length_longest <int> 61, 101, 485, 40, 40, 15, 4, 11, 445, 43,
6, 11…
## $ cap_run_length_total   <int> 278, 1028, 2259, 191, 418, 54, 112, 49,
1257, 7…
```

```
## $ clase                  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1,…
```

El dataset consta de 4601 filas y 58 columnas. Después de hacer *glimpse()* vemos que la mayor parte de las variables son de tipo *doble*. Variable *clase,* para clasificar como spam o no.

Para saber más de la distribución de variables hacemos una descripción estadística

Descripción estadística

```
options(scipen = 999, digits=3) # notación científica
skim(spam)
```

Data summary

Name	spam
Number of rows	4601
Number of columns	58
_____	
Column type frequency:	
numeric	58
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
make	0	1	0.10	0.30	0	0.00	0.00	0.00	4.54	▇_____
address	0	1	0.21	1.29	0	0.00	0.00	0.00	14.28	▇_____
all	0	1	0.28	0.51	0	0.00	0.00	0.42	5.10	▇_____

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
X3d	0	1	0.07	1.40	0	0.00	0.00	0.00	42.81	▇▁▁▁▁
our	0	1	0.31	0.67	0	0.00	0.00	0.39	10.00	▇▁▁▁▁
over	0	1	0.10	0.27	0	0.00	0.00	0.00	5.88	▇▁▁▁▁
remove	0	1	0.11	0.39	0	0.00	0.00	0.00	7.27	▇▁▁▁▁
internet	0	1	0.10	0.40	0	0.00	0.00	0.00	11.11	▇▁▁▁▁
order	0	1	0.09	0.28	0	0.00	0.00	0.00	5.26	▇▁▁▁▁
mail	0	1	0.24	0.65	0	0.00	0.00	0.16	18.18	▇▁▁▁▁
receive	0	1	0.06	0.20	0	0.00	0.00	0.00	2.61	▇▁▁▁▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
will	0	1	0.54	0.87	0	0.00	0.11	0.80	9.67	▇▁▁▁▁
people	0	1	0.09	0.30	0	0.00	0.00	0.00	5.55	▇▁▁▁▁
report	0	1	0.06	0.34	0	0.00	0.00	0.00	10.00	▇▁▁▁▁
addresses	0	1	0.05	0.26	0	0.00	0.00	0.00	4.41	▇▁▁▁▁
free	0	1	0.25	0.83	0	0.00	0.00	0.11	20.00	▇▁▁▁▁
business	0	1	0.14	0.44	0	0.00	0.00	0.00	7.14	▇▁▁▁▁
email	0	1	0.18	0.53	0	0.00	0.00	0.00	9.09	▇▁▁▁▁
you	0	1	1.66	1.77	0	0.00	1.31	2.63	18.75	▇▁▁▁▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
credit	0	1	0.08	0.51	0	0.00	0.00	0.00	18.18	▇▁▁▁▁
your	0	1	0.81	1.21	0	0.00	0.22	1.27	11.11	▇▁▁▁▁
font	0	1	0.12	1.02	0	0.00	0.00	0.00	17.10	▇▁▁▁▁
X000	0	1	0.10	0.35	0	0.00	0.00	0.00	5.45	▇▁▁▁▁
money	0	1	0.09	0.44	0	0.00	0.00	0.00	12.50	▇▁▁▁▁
hp	0	1	0.55	1.67	0	0.00	0.00	0.00	20.83	▇▁▁▁▁
hpl	0	1	0.27	0.89	0	0.00	0.00	0.00	16.66	▇▁▁▁▁
george	0	1	0.78	3.41	0	0.00	0.00	0.00	33.33	▇▁▁▁▁

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
X650	0	1	0.12	0.54	0	0.00	0.00	0.00	9.09	▇_____
lab	0	1	0.10	0.60	0	0.00	0.00	0.00	14.28	▇_____
labs	0	1	0.10	0.46	0	0.00	0.00	0.00	5.88	▇_____
telnet	0	1	0.06	0.40	0	0.00	0.00	0.00	12.50	▇_____
X857	0	1	0.05	0.33	0	0.00	0.00	0.00	4.76	▇_____
data	0	1	0.10	0.56	0	0.00	0.00	0.00	18.18	▇_____
X415	0	1	0.05	0.33	0	0.00	0.00	0.00	4.76	▇_____
X85	0	1	0.11	0.53	0	0.00	0.00	0.00	20.00	▇_____

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
technology	0	1	0.10	0.40	0	0.00	0.00	0.00	7.69	▇_____
X1999	0	1	0.13	0.42	0	0.00	0.00	0.00	6.89	▇_____
parts	0	1	0.01	0.22	0	0.00	0.00	0.00	8.33	▇_____
pm	0	1	0.08	0.43	0	0.00	0.00	0.00	11.11	▇_____
direct	0	1	0.06	0.35	0	0.00	0.00	0.00	4.76	▇_____
cs	0	1	0.04	0.36	0	0.00	0.00	0.00	7.14	▇_____
meeting	0	1	0.13	0.76	0	0.00	0.00	0.00	14.28	▇_____
original	0	1	0.05	0.22	0	0.00	0.00	0.00	3.57	▇_____

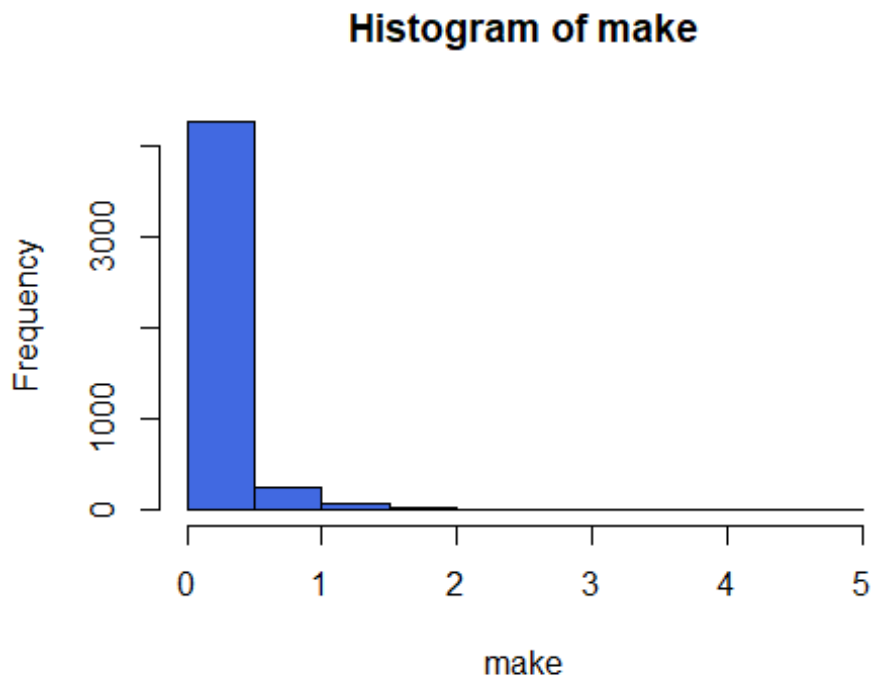
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
project	0	1	0.08	0.63	0	0.00	0.00	0.00	20.00	
re	0	1	0.30	1.03	0	0.00	0.00	0.12	21.42	
edu	0	1	0.18	0.91	0	0.00	0.00	0.00	22.05	
table	0	1	0.01	0.08	0	0.00	0.00	0.00	2.17	
conference	0	1	0.03	0.29	0	0.00	0.00	0.00	10.00	
X.	0	1	0.04	0.23	0	0.00	0.00	0.00	4.37	
X..1	0	1	0.14	0.27	0	0.00	0.06	0.19	9.75	
X..2	0	1	0.02	0.11	0	0.00	0.00	0.00	4.08	

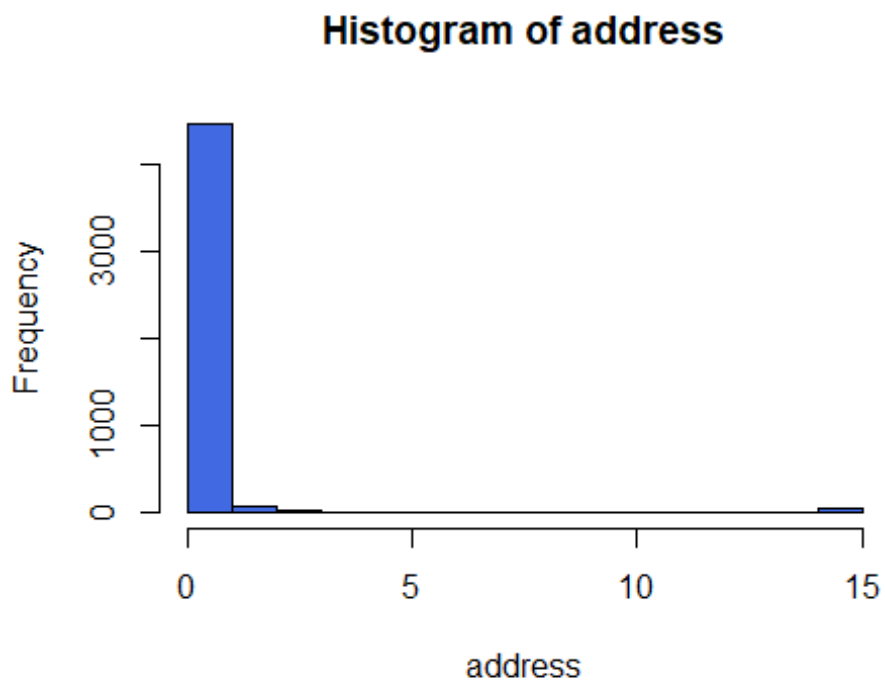
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
X..3	0	1	0.27	0.81	0	0.00	0.00	0.31	32.48	▇▁▁▁▁
X..4	0	1	0.07	0.24	0	0.00	0.00	0.05	6.00	▇▁▁▁▁
X..5	0	1	0.04	0.43	0	0.00	0.00	0.00	19.83	▇▁▁▁▁
cap_run_length_average	0	1	5.19	31.80	1	1.59	2.28	3.70	1102.50	▇▁▁▁▁
cap_run_length_longest	0	1	52.14	194.82	1	6.00	15.00	43.00	9989.00	▇▁▁▁▁
cap_run_length_total	0	1	282.98	605.38	1	35.00	95.00	266.00	15841.00	▇▁▁▁▁
clase	0	1	0.39	0.49	0	0.00	0.00	1.00	1.00	▇▁▁▁▇

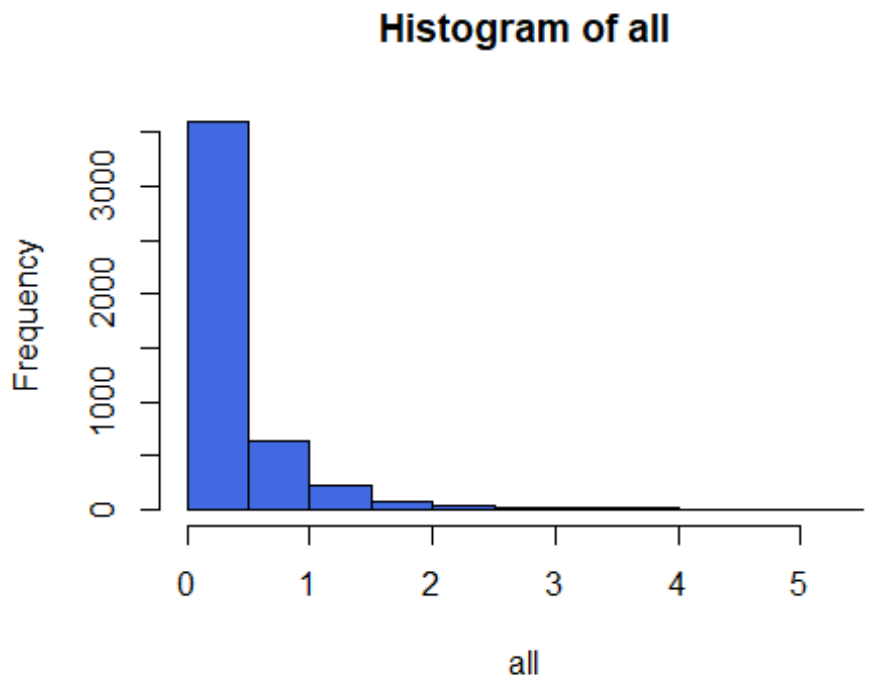
- Se observa que un gran número de *means* están muy próximas a cero. No se espera que tengamos valores muy altos.
- Hay muchas variables con valores igual a cero.
- Las *sd* son altas. Esto nos indica que los en general los valores no están agrupados en torno a la media por lo que tendremos un número alto de *outliers*

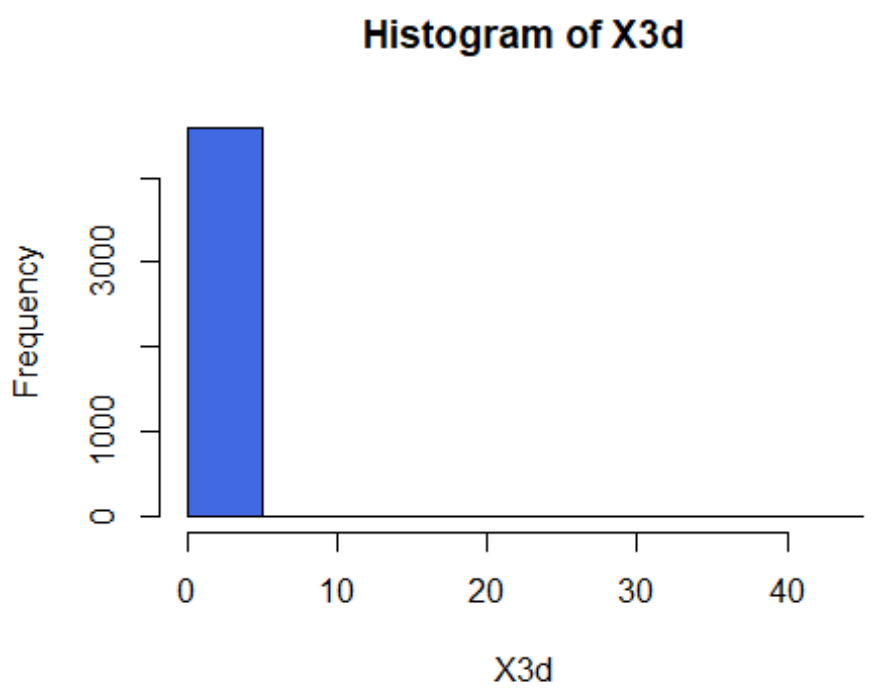
- Podemos visualizar la distribución de las variables mediante histogramas.

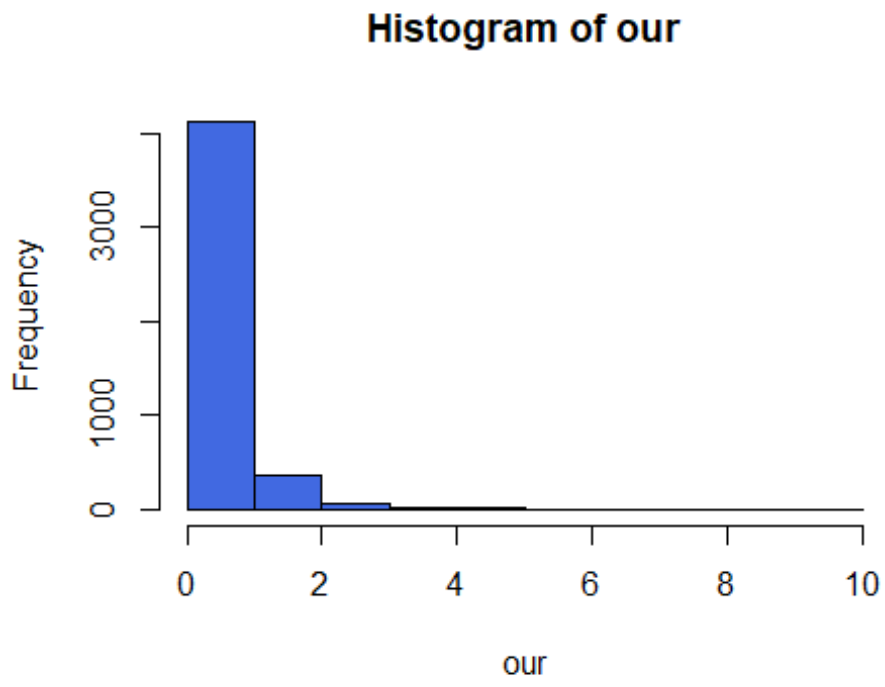
```
for (var in colnames(spam)[1:57]) {
  hist(unlist(spam[,var]), col='royalblue',
       main= paste('Histogram of', var),
       xlab=var)
}
```

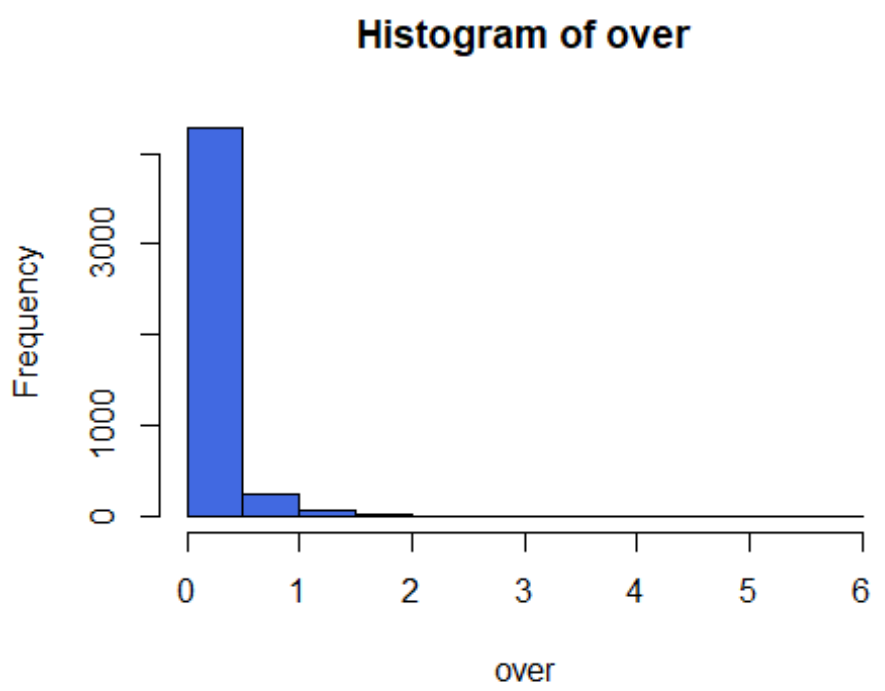

Histogram of make
Frequency
0
1000
3000
0
1
2
3
4
5
make


Histogram of address
Frequency
0
1000
3000
0
5
10
15
address

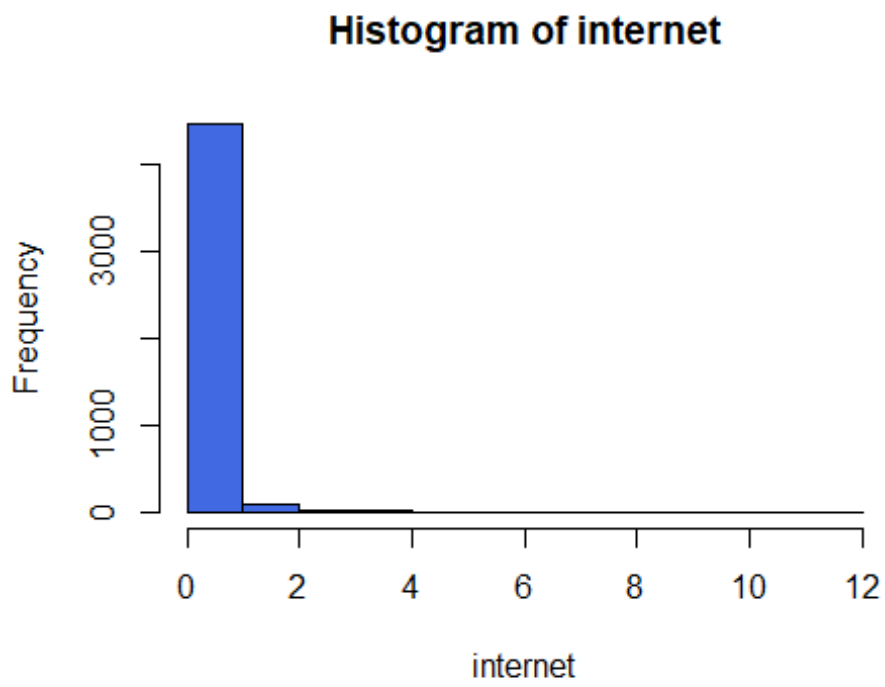

Histogram of all
Frequency
0
1000
2000
3000
0
1
2
3
4
5
all

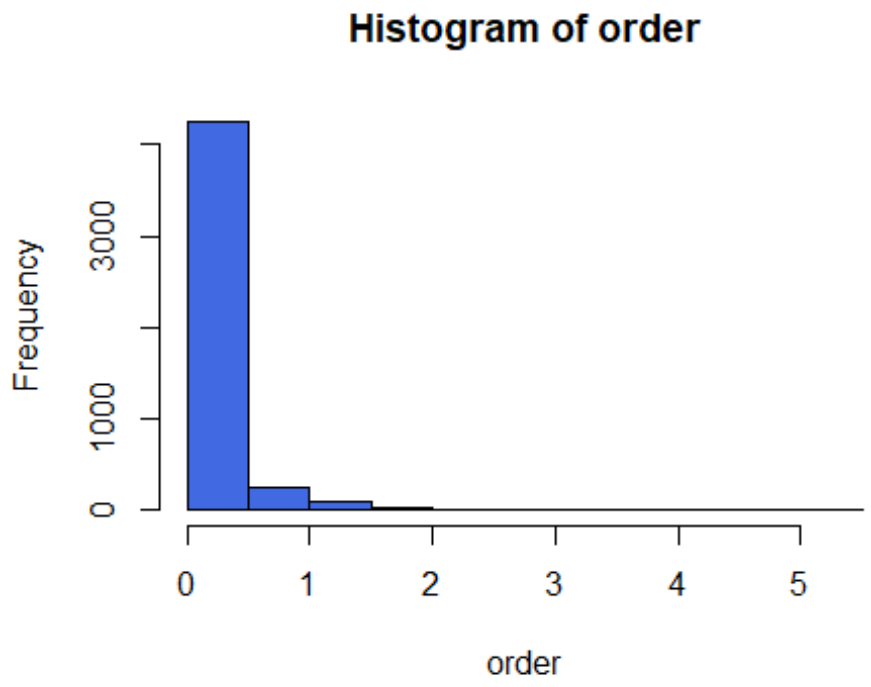

Histogram of X3d
Frequency
0
1000
3000
0
10
20
30
40
X3d

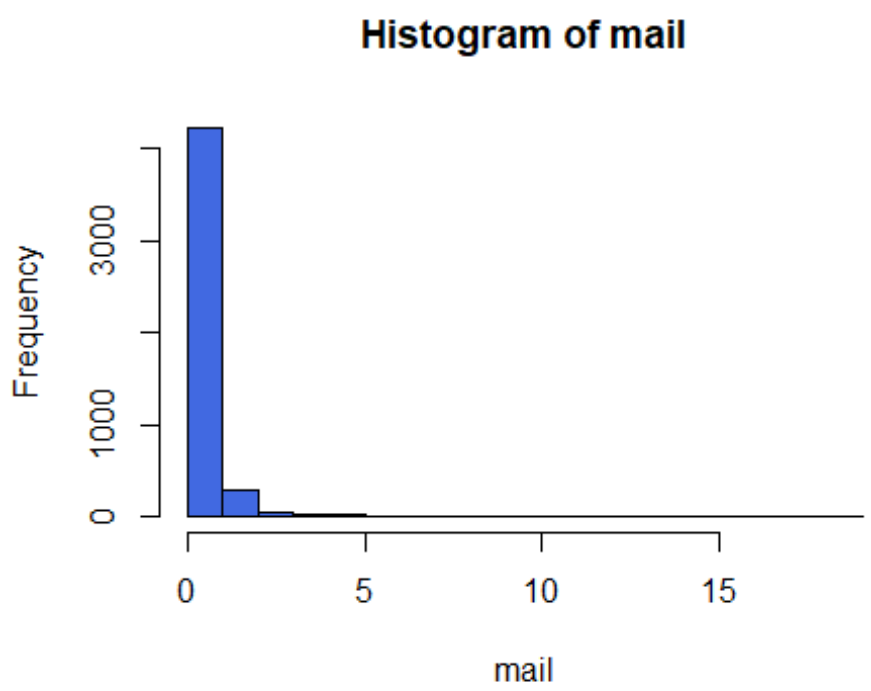

Histogram of our
Frequency
0
1000
3000
0
2
4
6
8
10
our

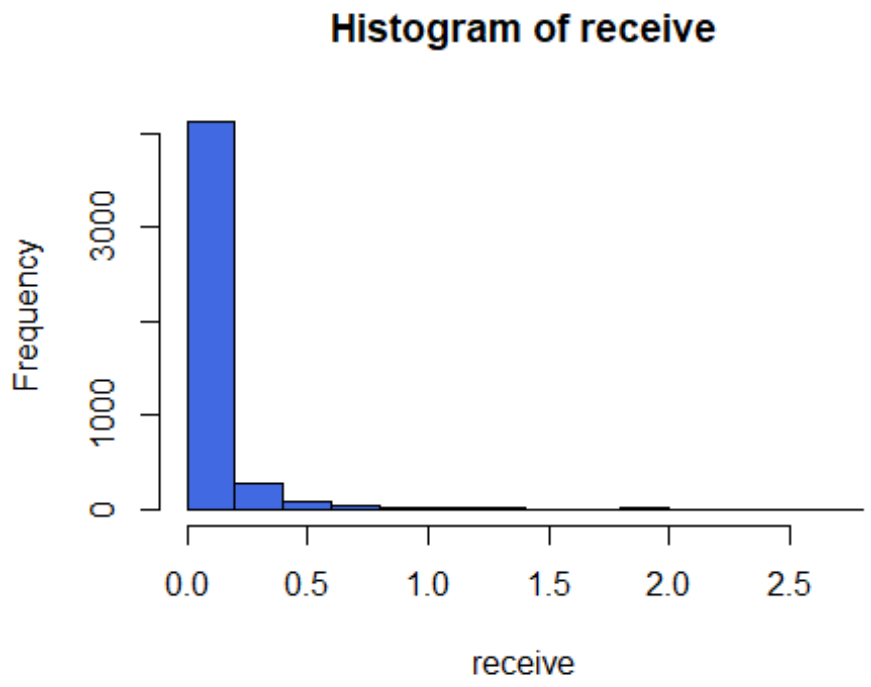

Histogram of over
Frequency
0
1000
3000
0
1
2
3
4
5
6
over

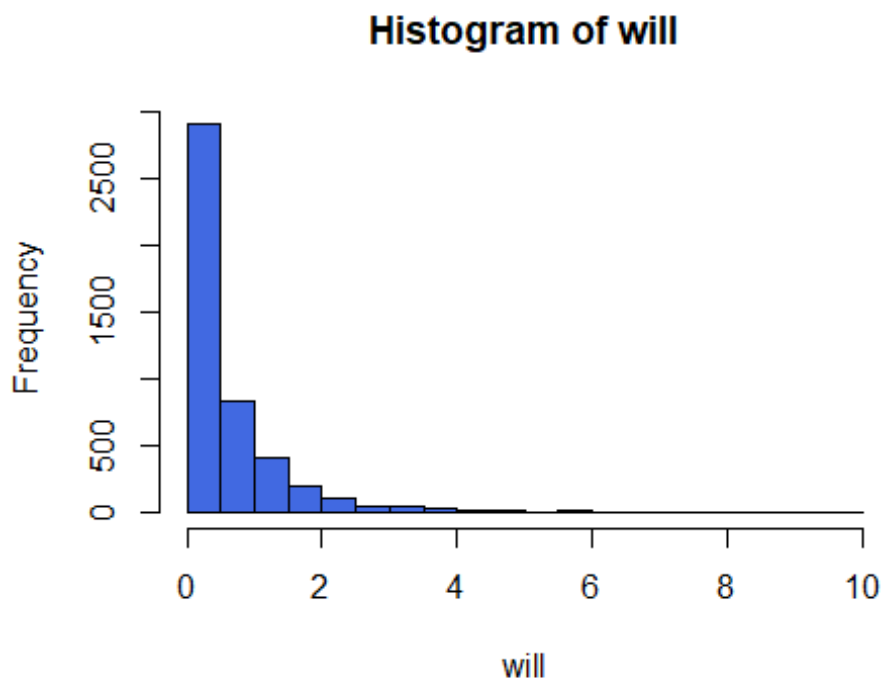

Histogram of remove
Frequency
0
1000
3000
0
2
4
6
remove

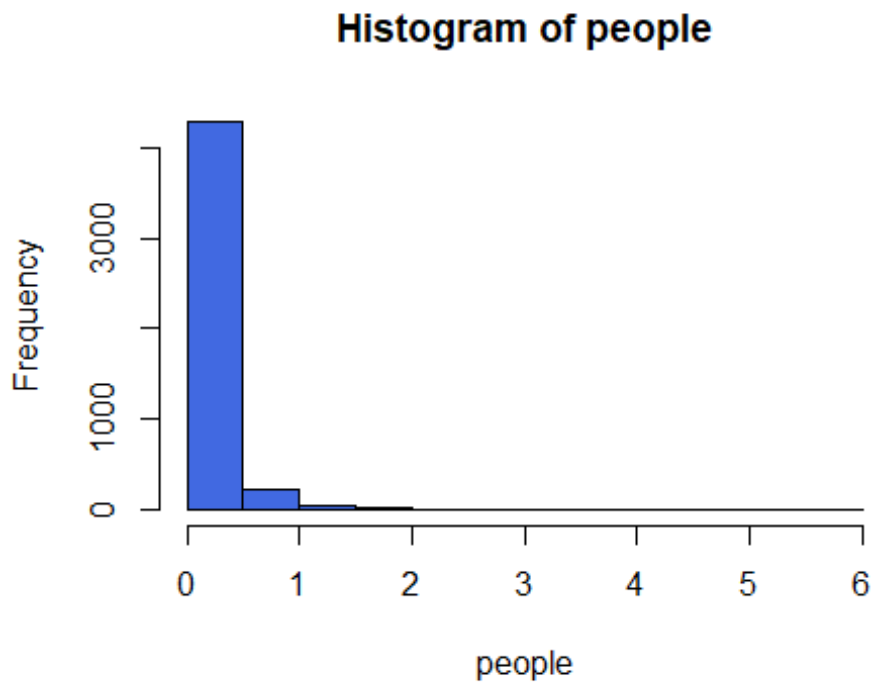

Histogram of internet
Frequency
0
1000
3000
0
2
4
6
8
10
12
internet

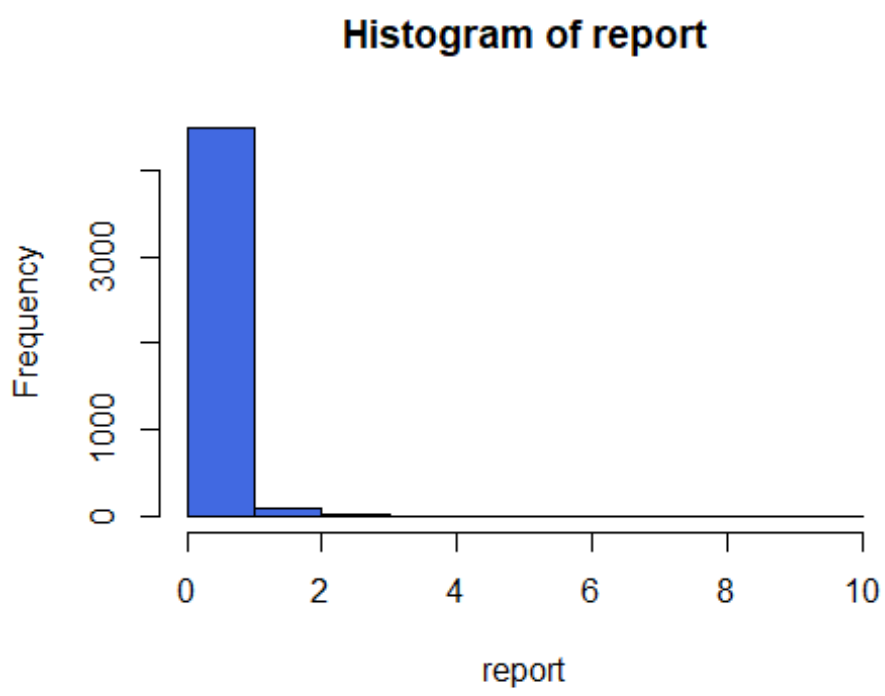

Histogram of order
Frequency
0
1000
3000
0
1
2
3
4
5
order

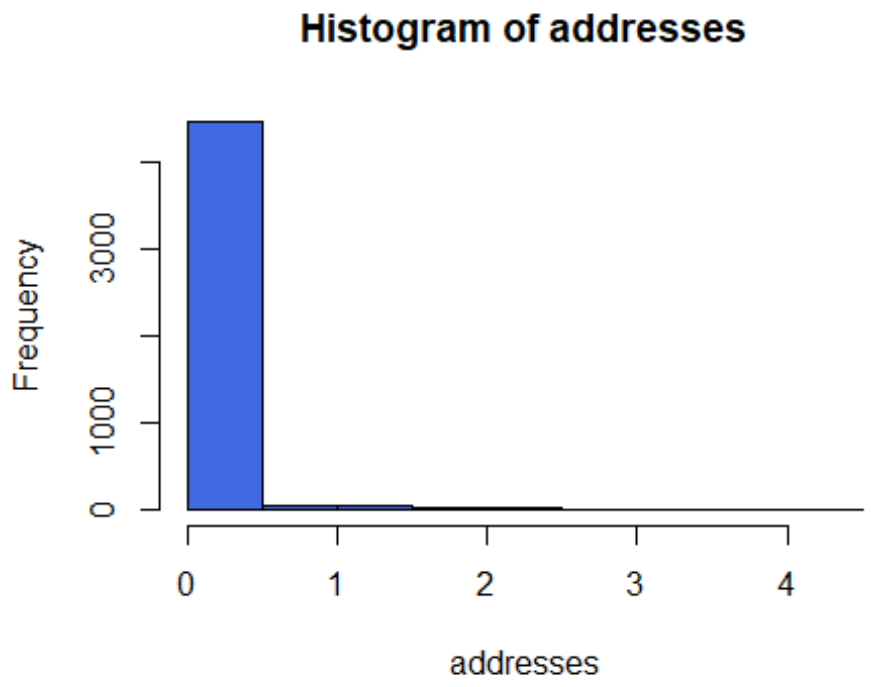

Histogram of mail
Frequency
0
1000
3000
0
5
10
15
mail

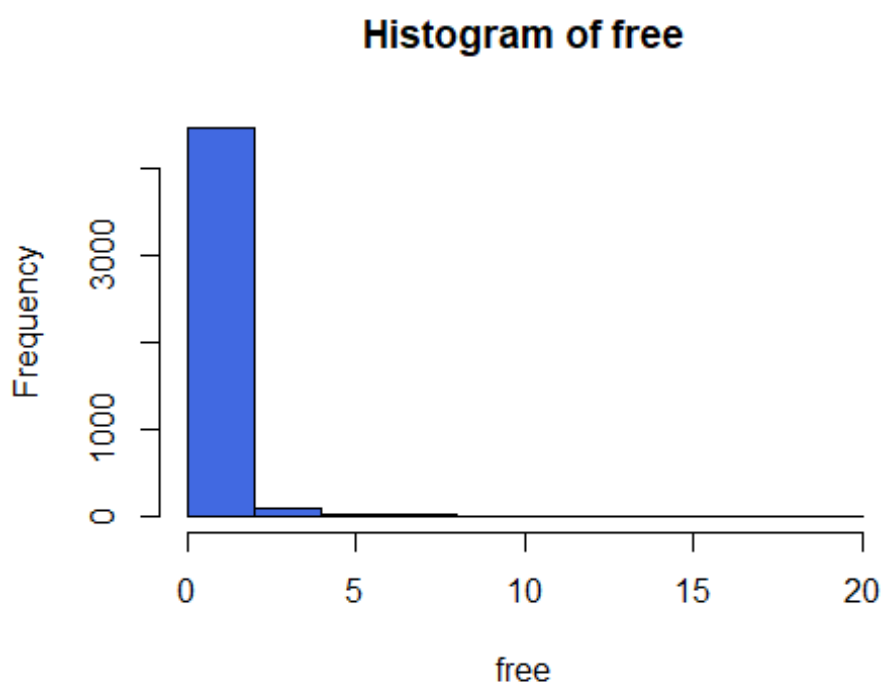

Histogram of receive
Frequency
0
1000
3000
0.0
0.5
1.0
1.5
2.0
2.5
receive

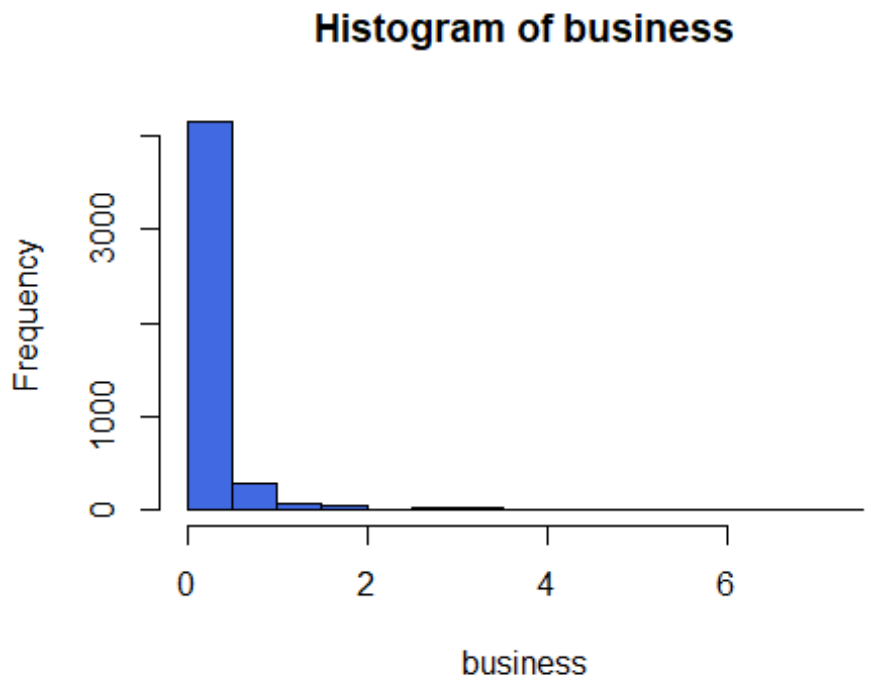

Histogram of will
Frequency
0
500
1500
2500
0
2
4
6
8
10
will

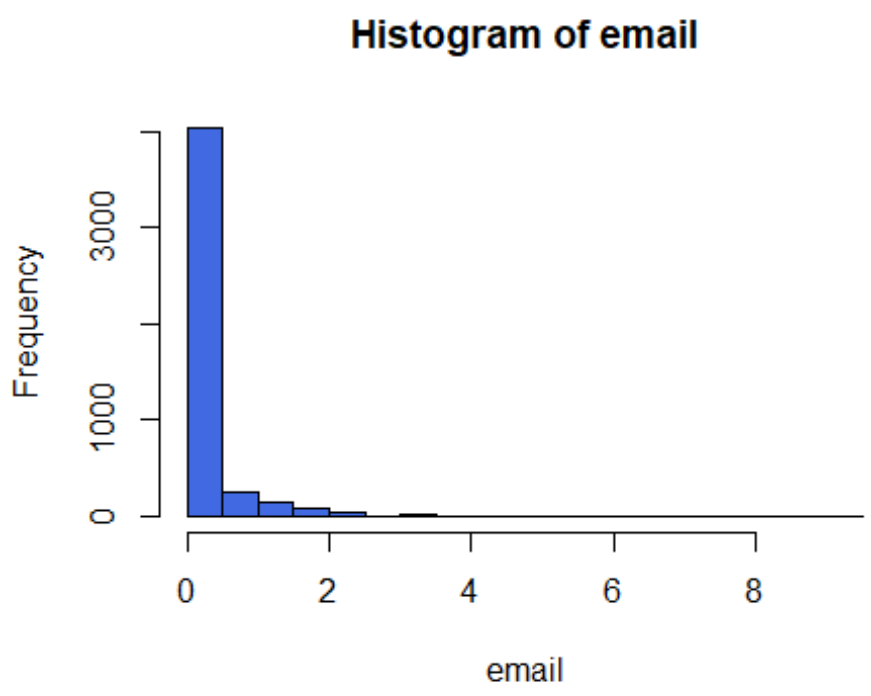

Histogram of people
Frequency
0
1000
3000
0
1
2
3
4
5
6
people

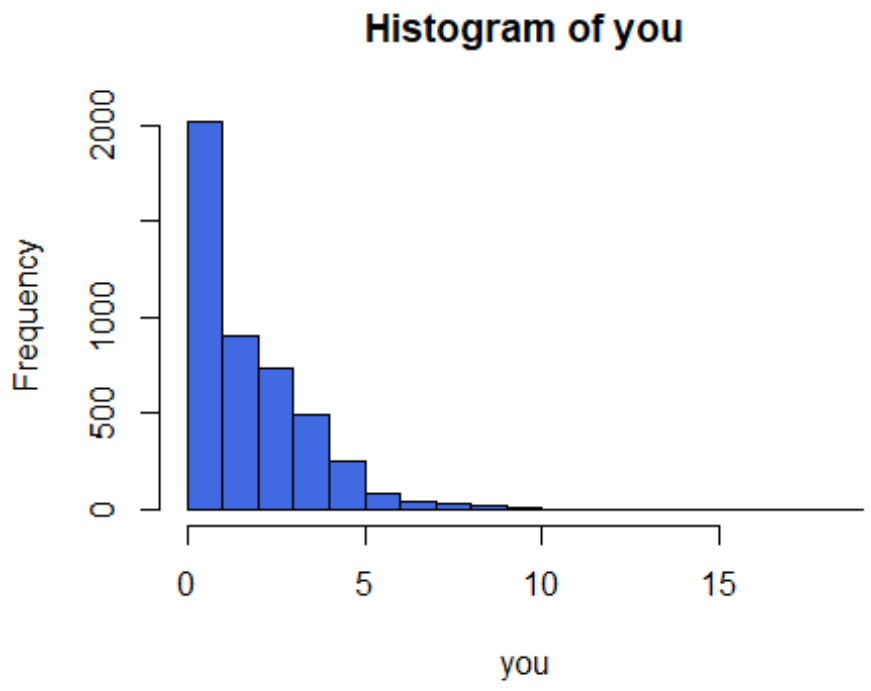

Histogram of report
Frequency
0
1000
3000
0
2
4
6
8
10
report

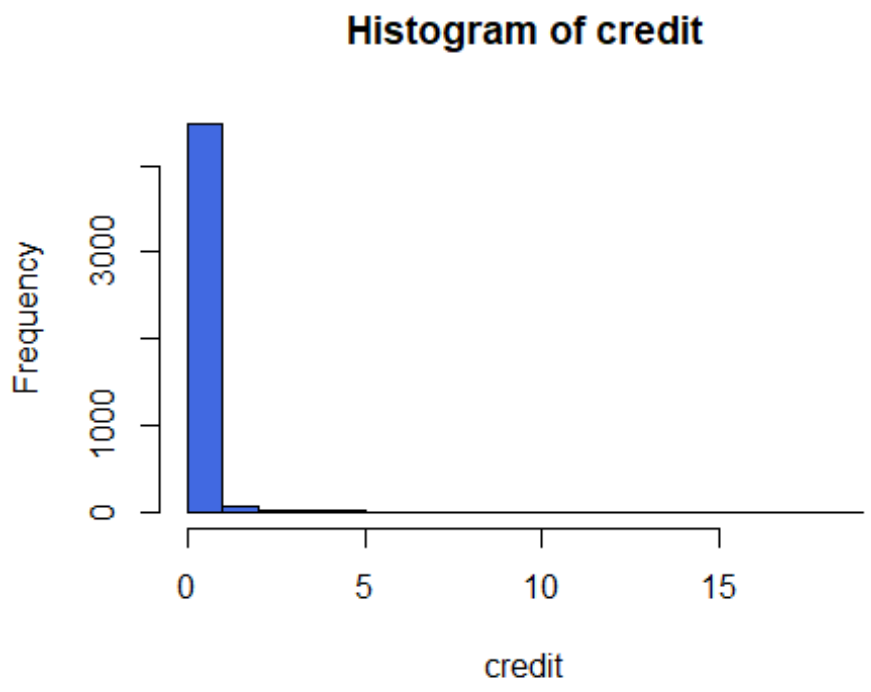

Histogram of addresses
Frequency
0
1000
3000
0
1
2
3
4
addresses

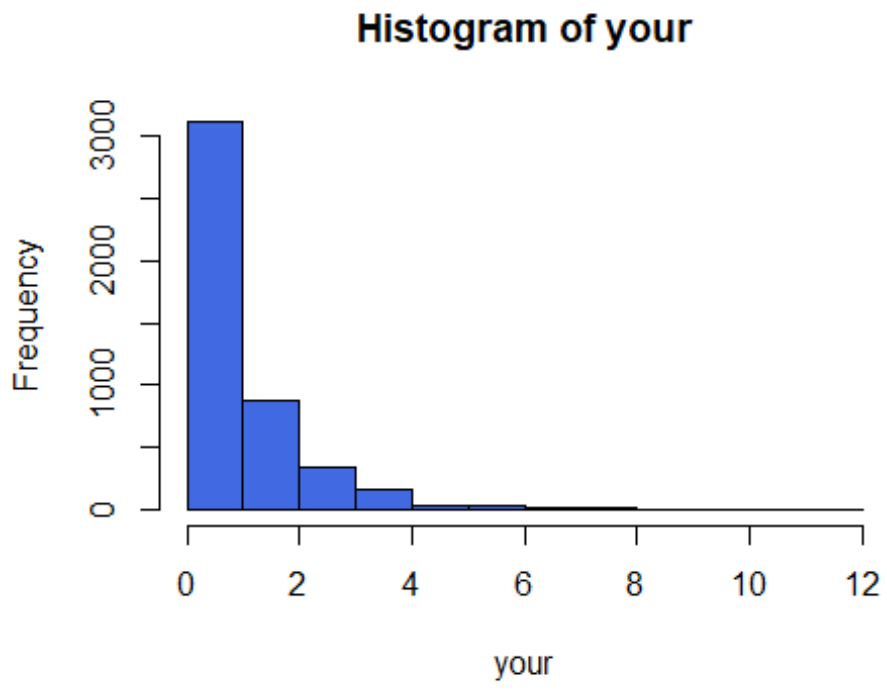

Histogram of free
Frequency
0
1000
3000
0
5
10
15
20
free

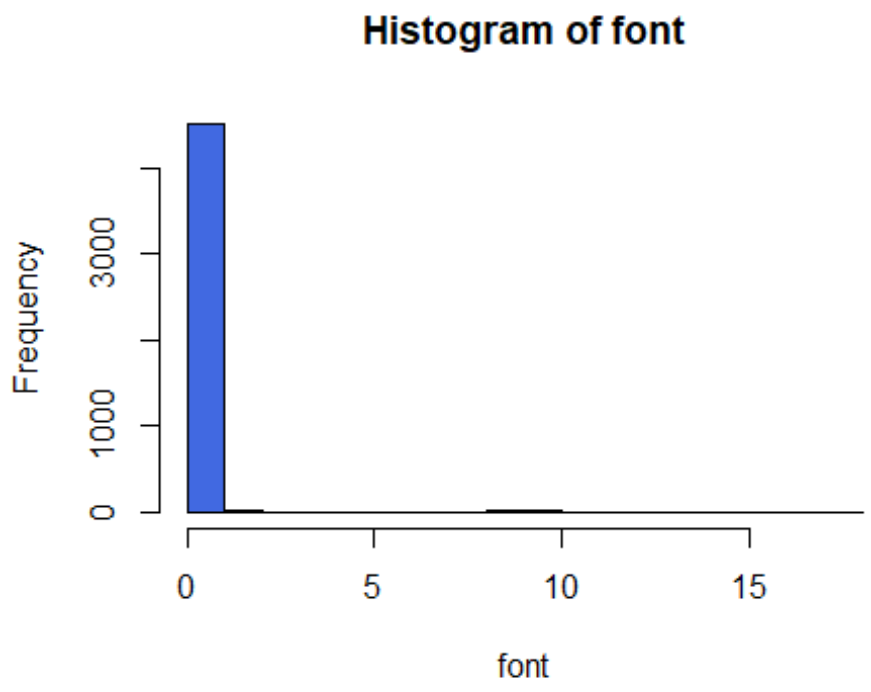

Histogram of business
Frequency
0
1000
3000
0
2
4
6
business

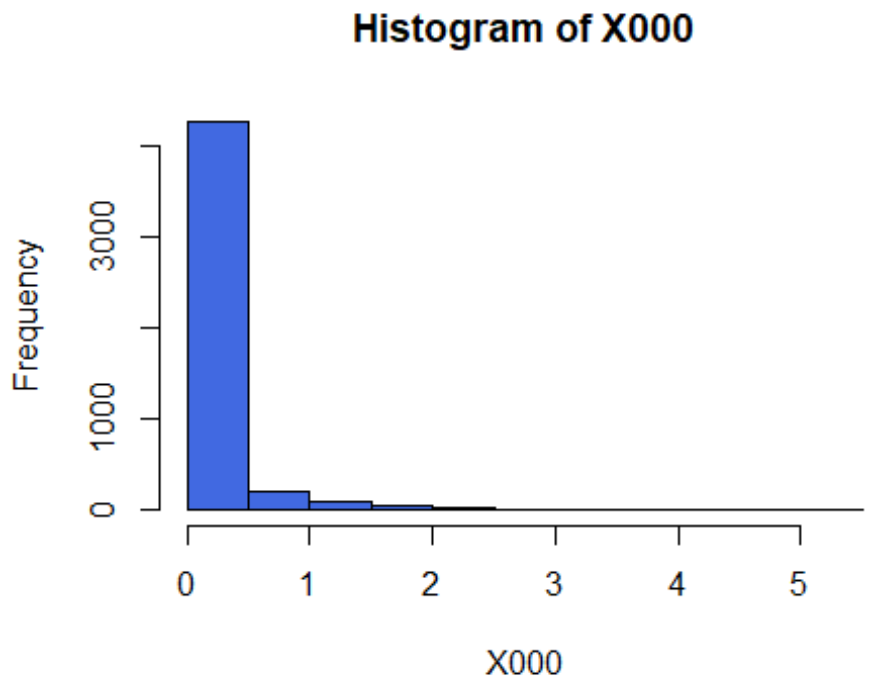

Histogram of email
Frequency
0
1000
3000
0
2
4
6
8
email

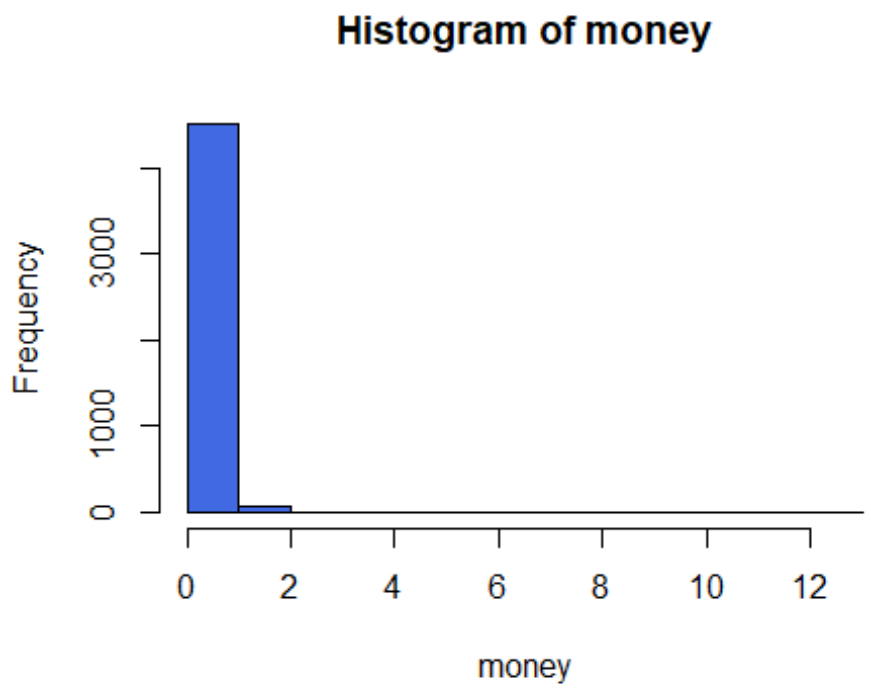

Histogram of you
Frequency
0
500
1000
2000
0
5
10
15
you

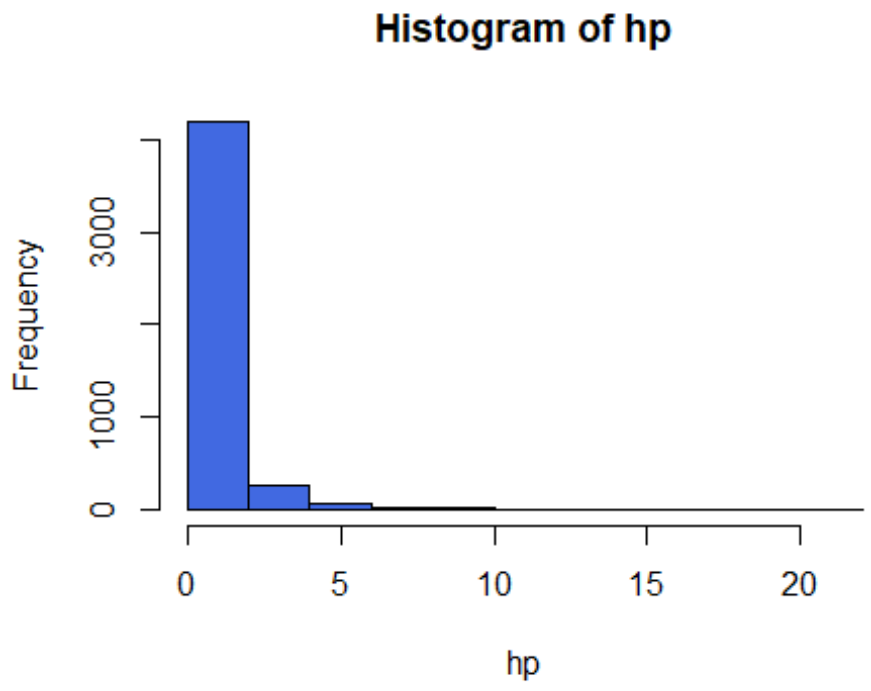

Histogram of credit
Frequency
0
1000
3000
0
5
10
15
credit

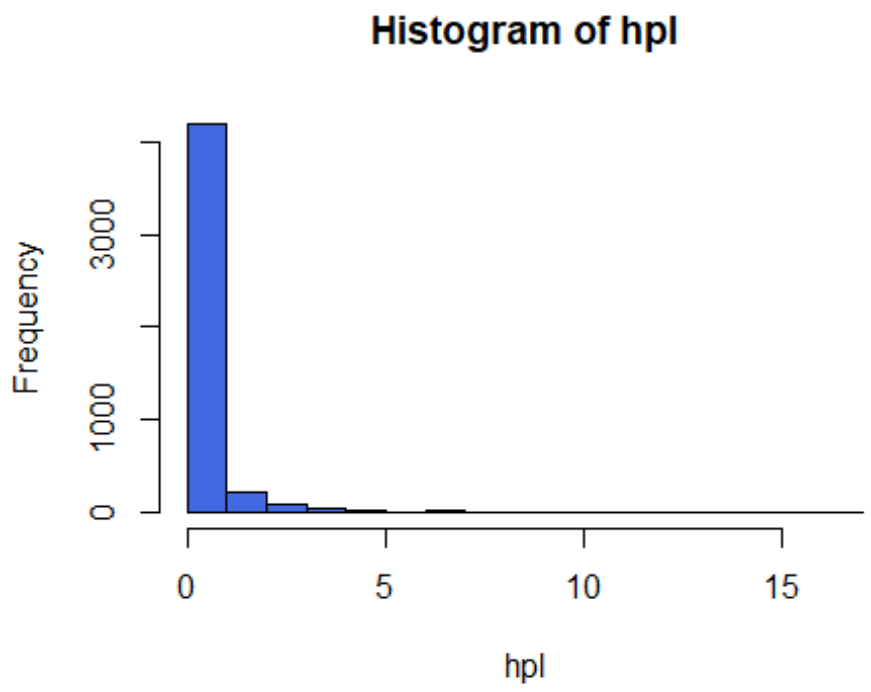

Histogram of your
Frequency
0
1000
2000
3000
0
2
4
6
8
10
12
your

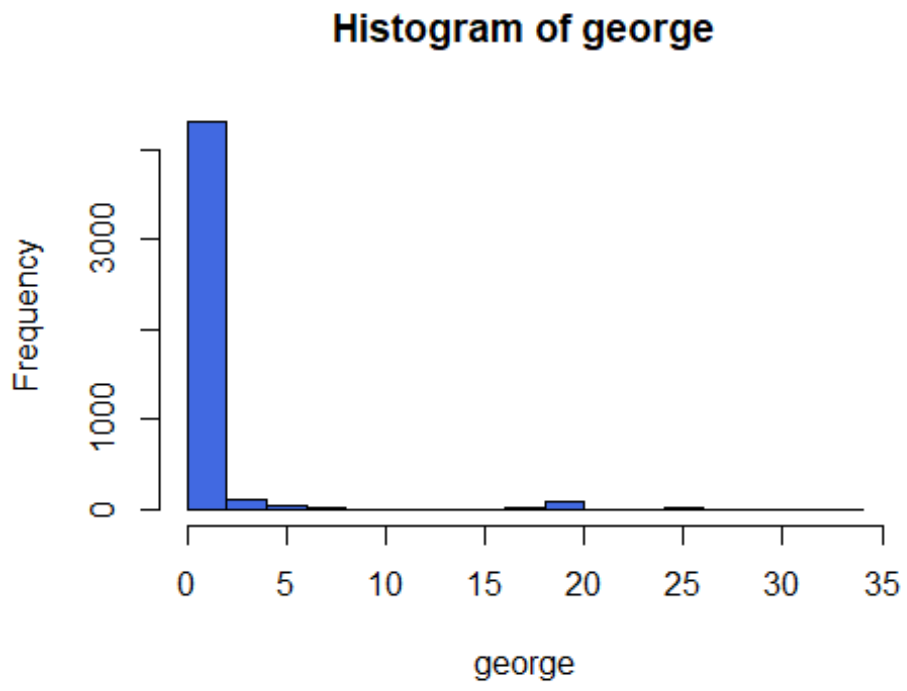

Histogram of font
Frequency
0
1000
3000
0
5
10
15
font

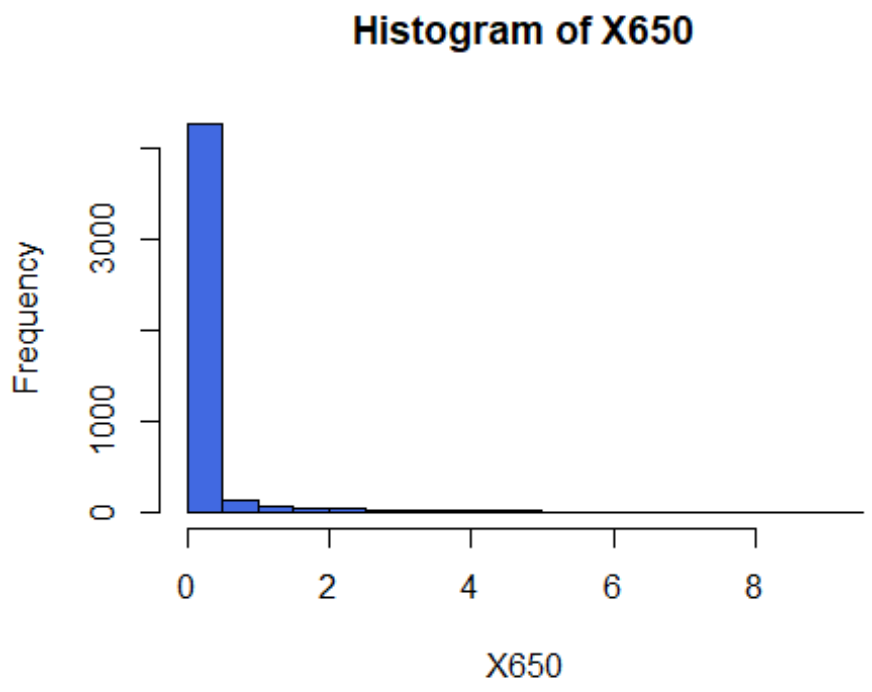

Histogram of X000
Frequency
0
1000
3000
0
1
2
3
4
5
X000

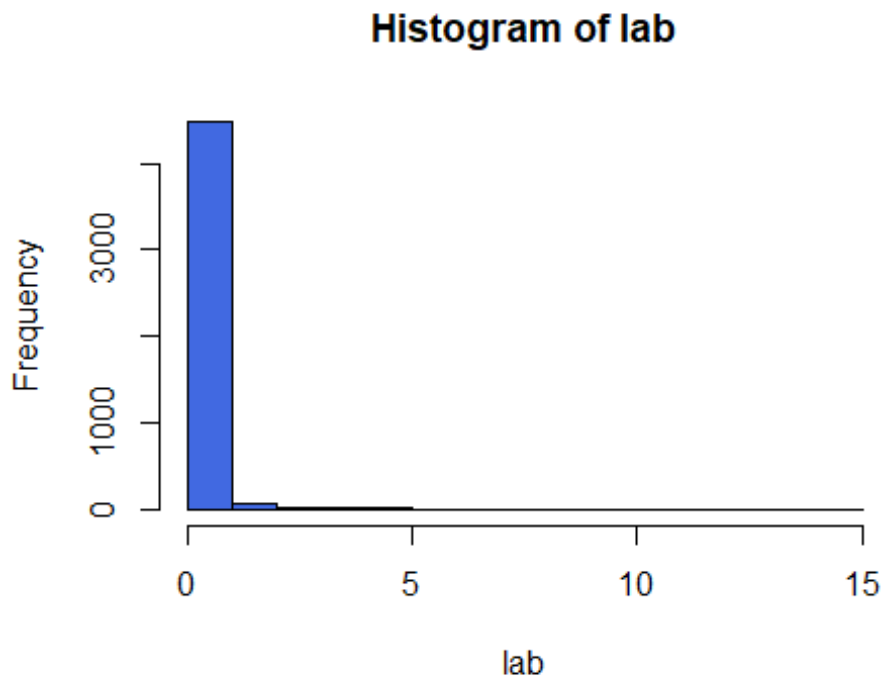

Histogram of money
Frequency
0
1000
3000
0
2
4
6
8
10
12
money

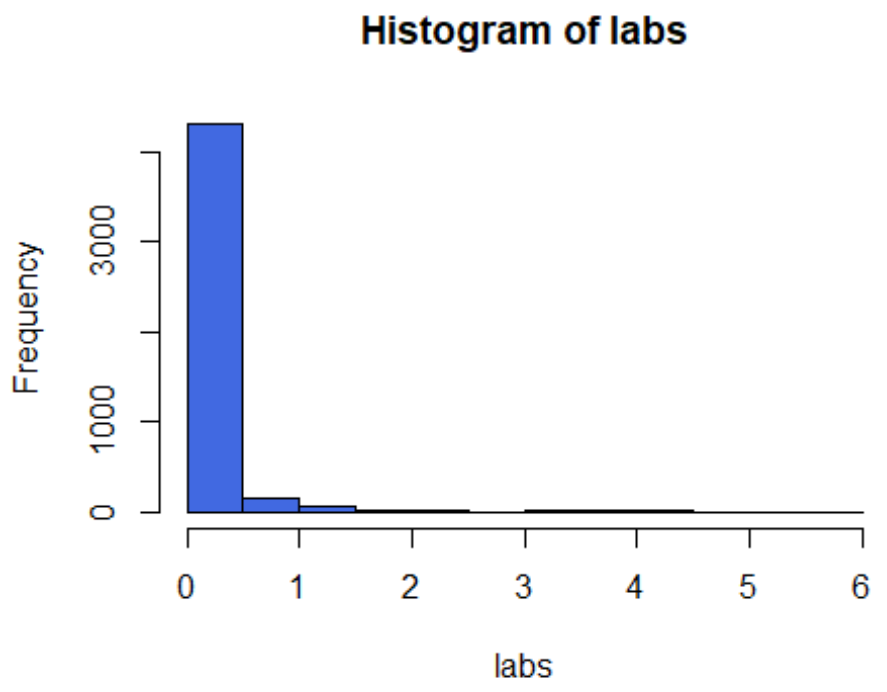

Histogram of hp
Frequency
0
1000
3000
0
5
10
15
20
hp

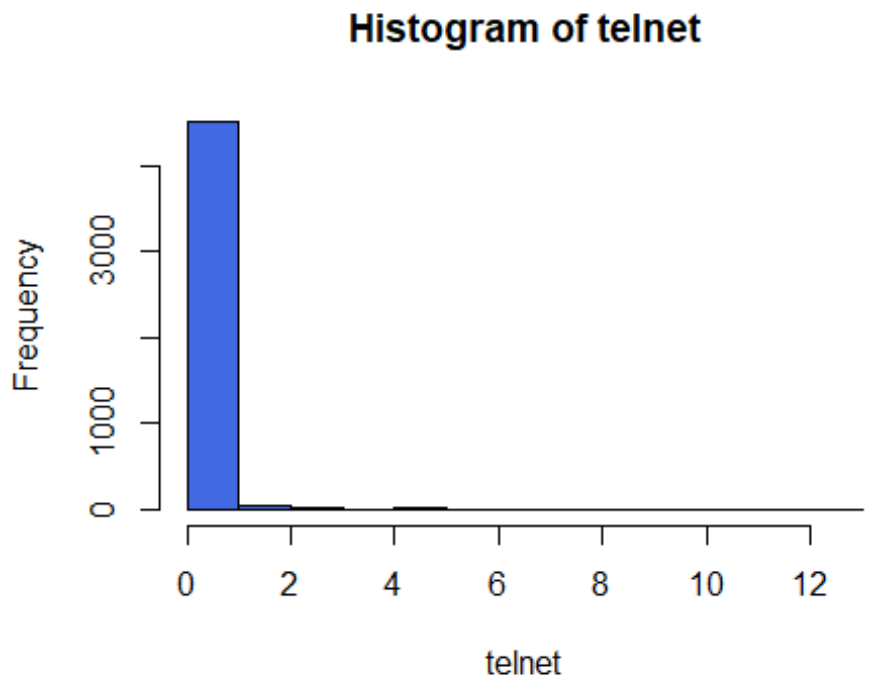

Histogram of hpl
Frequency
0
1000
3000
0
5
10
15
hpl

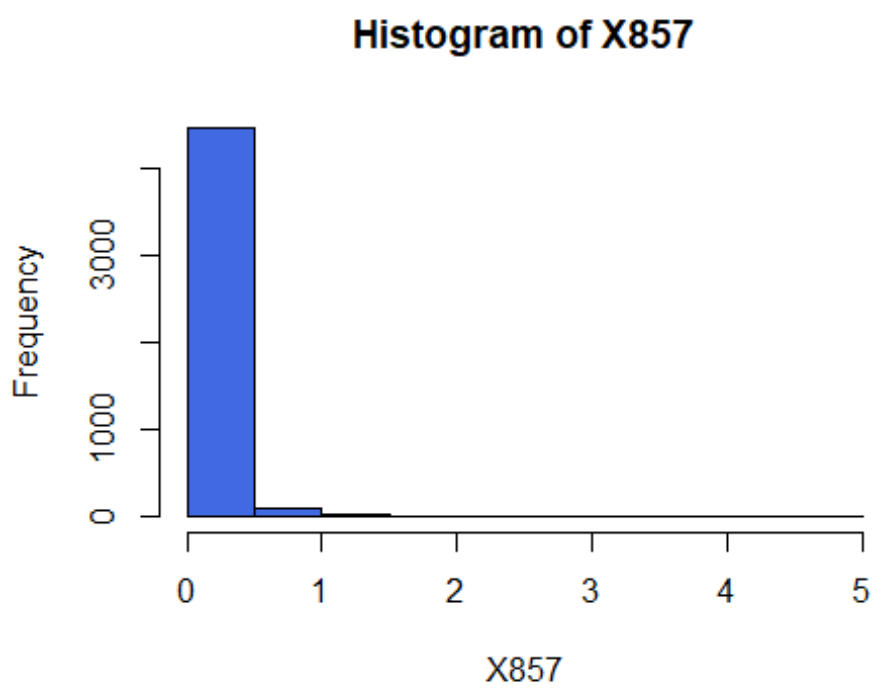

Histogram of george
Frequency
0
1000
3000
0
5
10
15
20
25
30
35
george

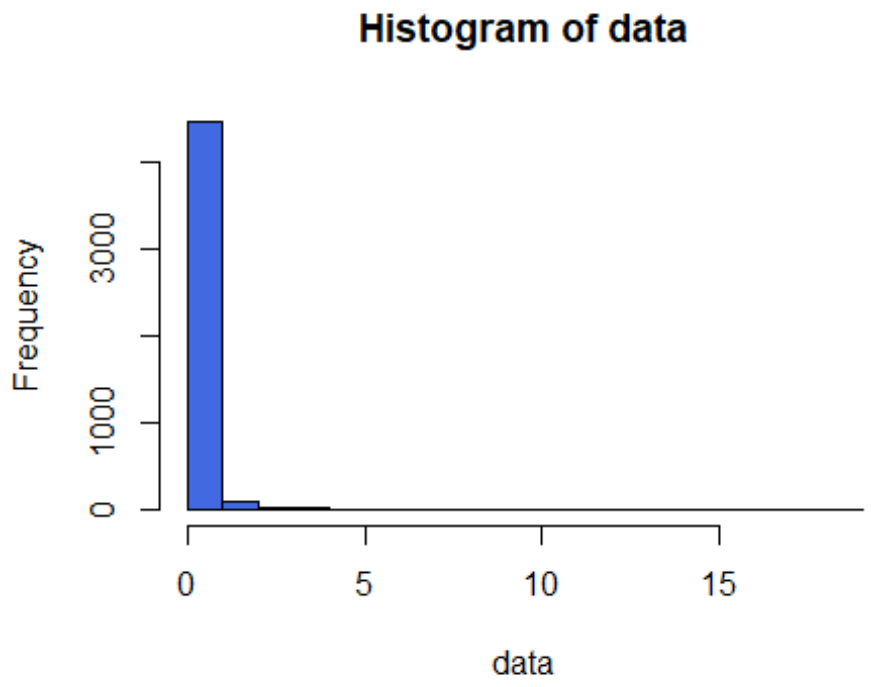

Histogram of X650
Frequency
0
1000
3000
0
2
4
6
8
X650

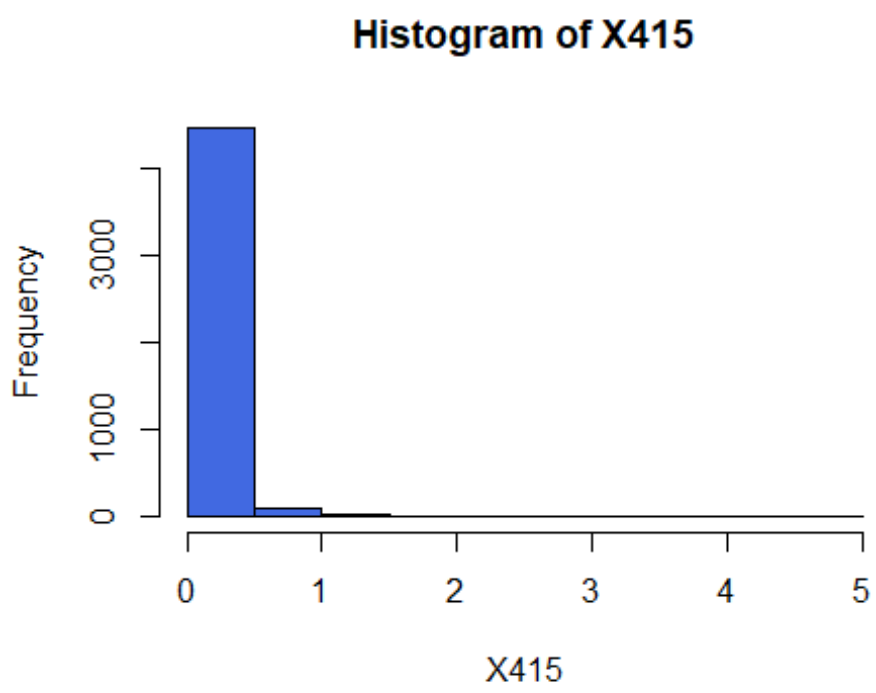

Histogram of lab
Frequency
0
1000
3000
0
5
10
15
lab

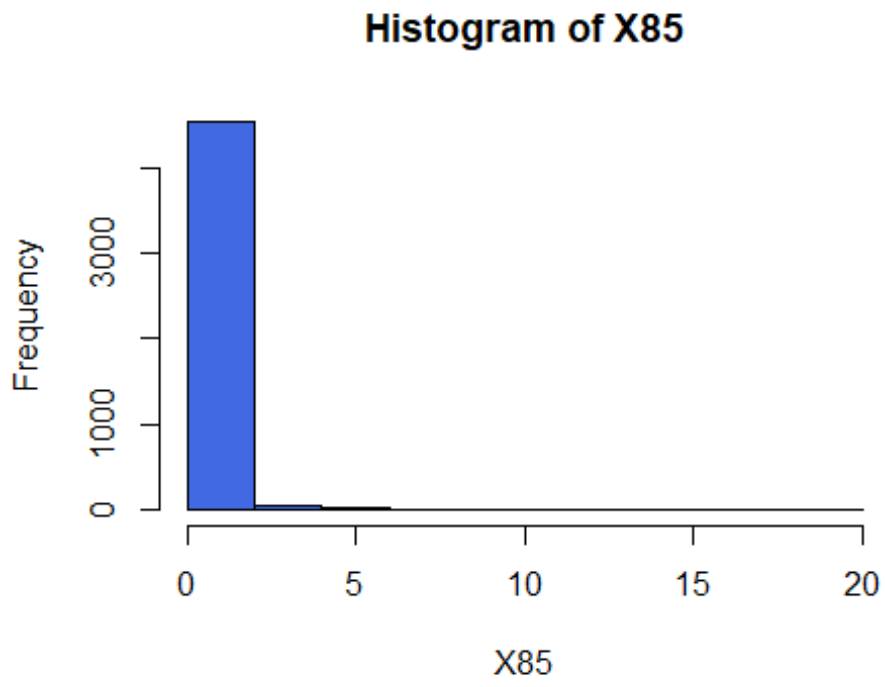

Histogram of labs
Frequency
0
1000
3000
0
1
2
3
4
5
6
labs

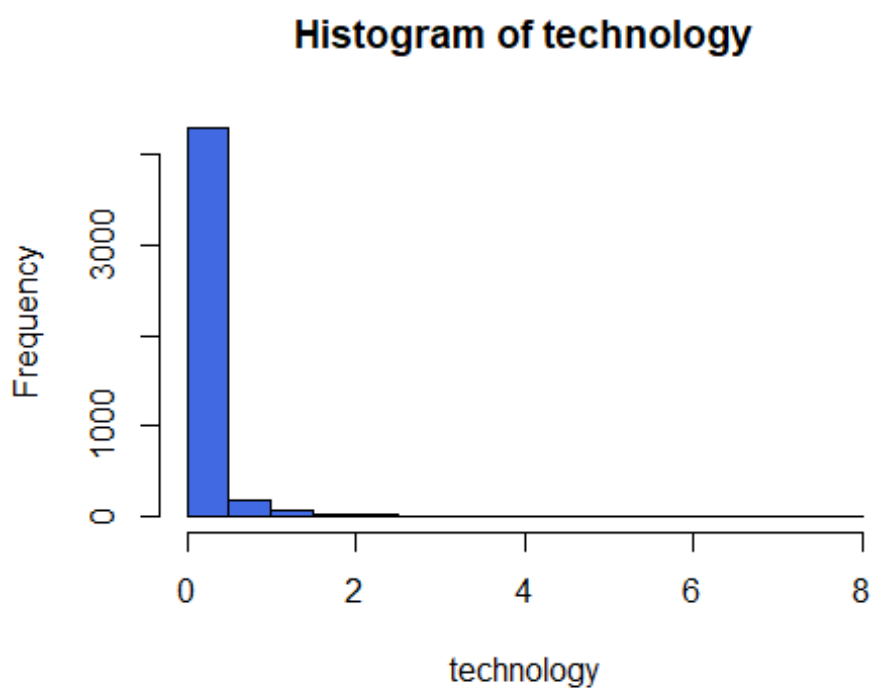

Histogram of telnet
Frequency
0
1000
3000
0
2
4
6
8
10
12
telnet

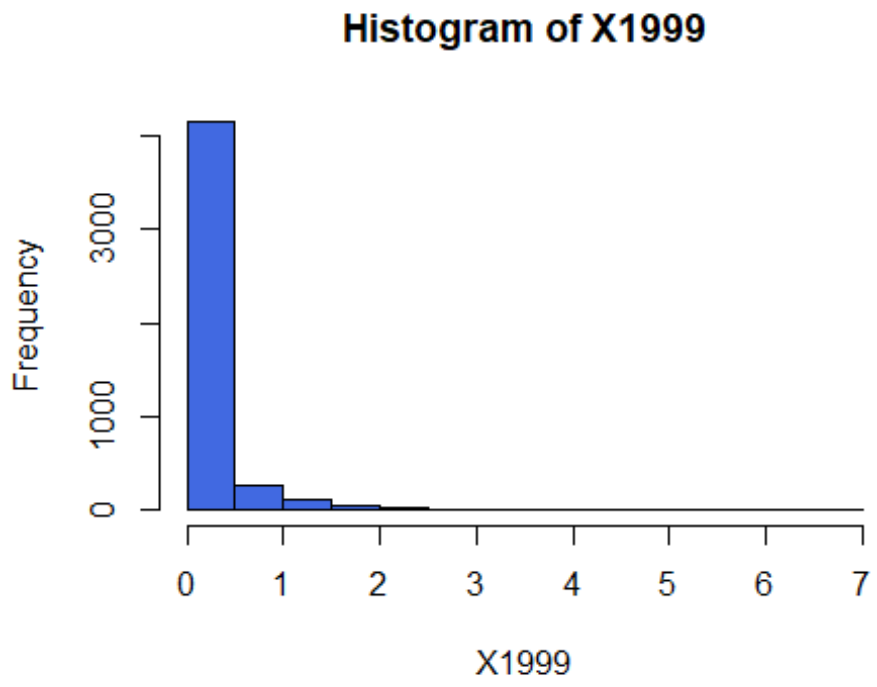

Histogram of X857
Frequency
0
1000
3000
0
1
2
3
4
5
X857

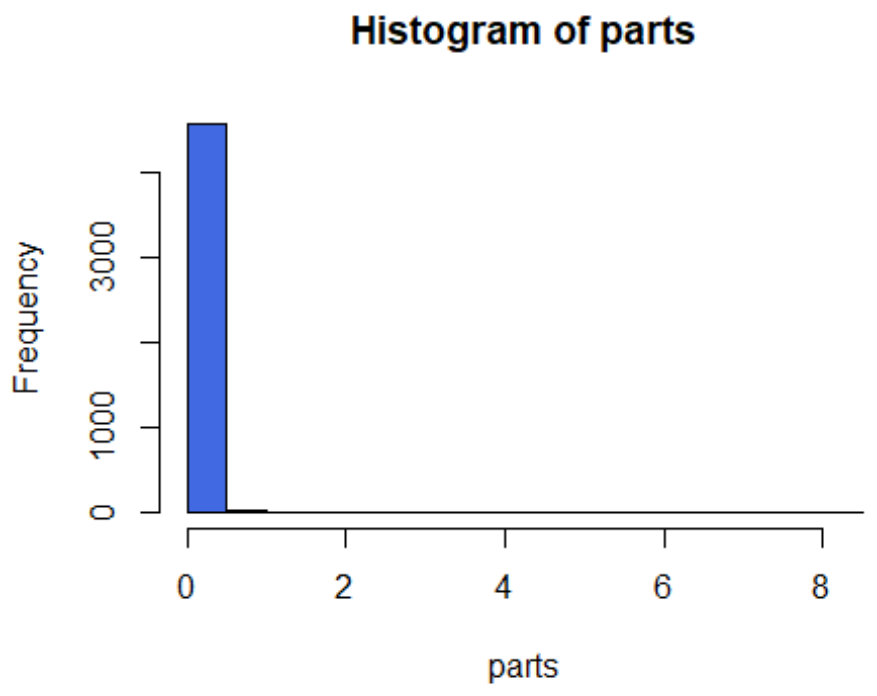

Histogram of data
Frequency
0
1000
3000
0
5
10
15
data

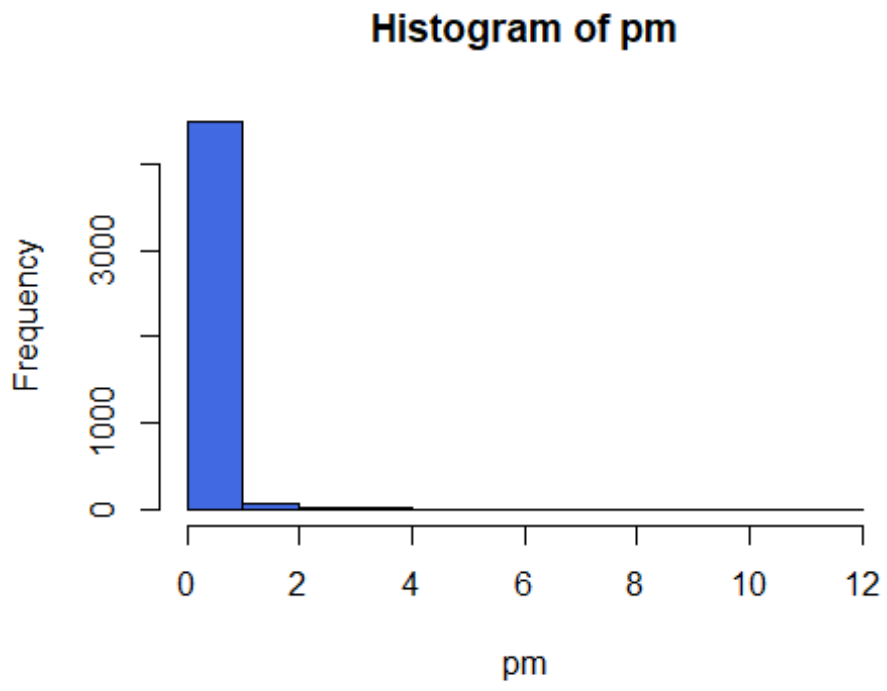

Histogram of X415
Frequency
0
1000
3000
0
1
2
3
4
5
X415

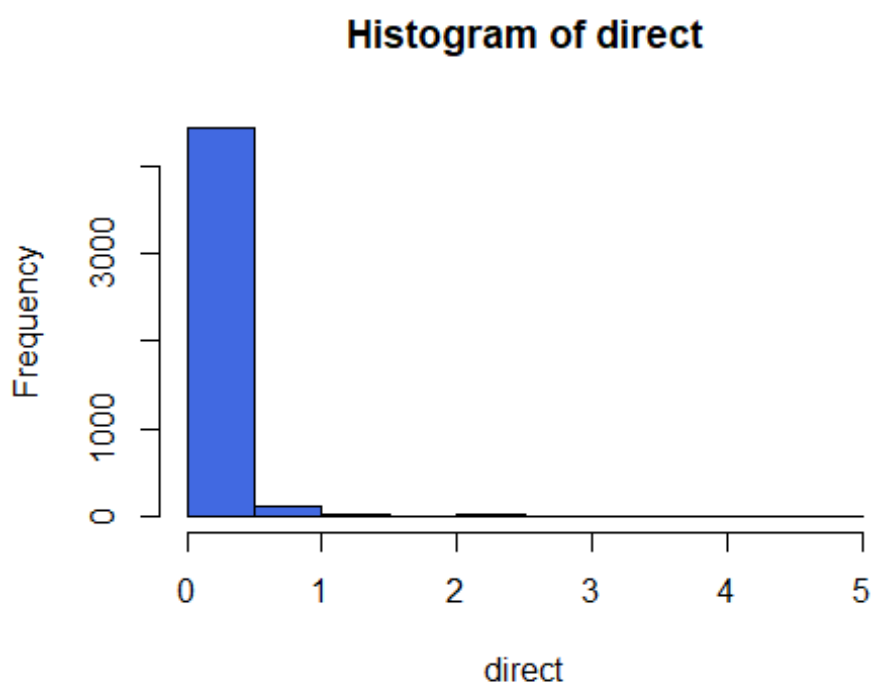

Histogram of X85
Frequency
0
1000
3000
0
5
10
15
20
X85

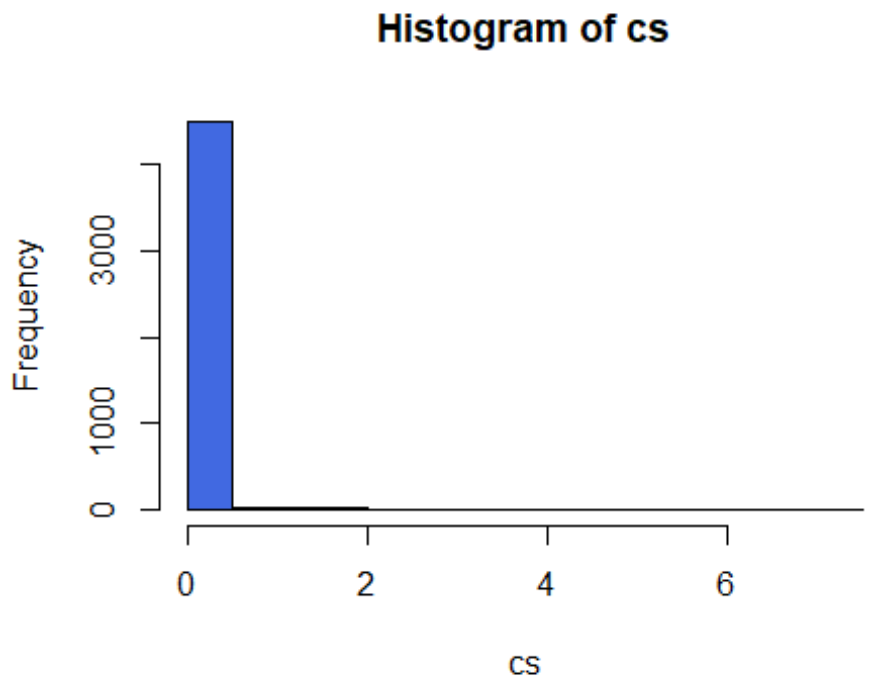

Histogram of technology
Frequency
0
1000
3000
0
2
4
6
8
technology

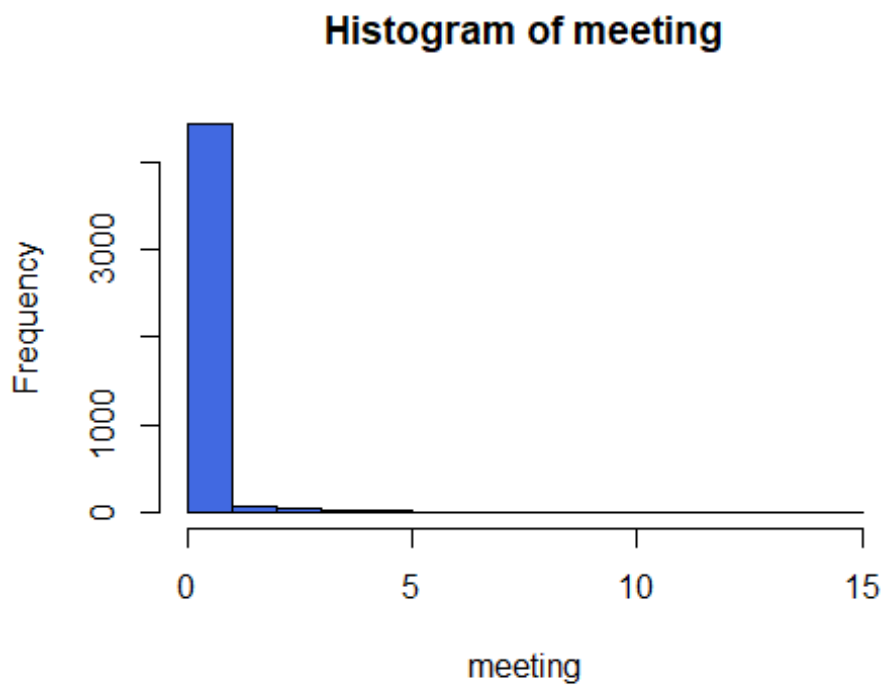

Histogram of X1999
Frequency
0
1000
3000
0
1
2
3
4
5
6
7
X1999

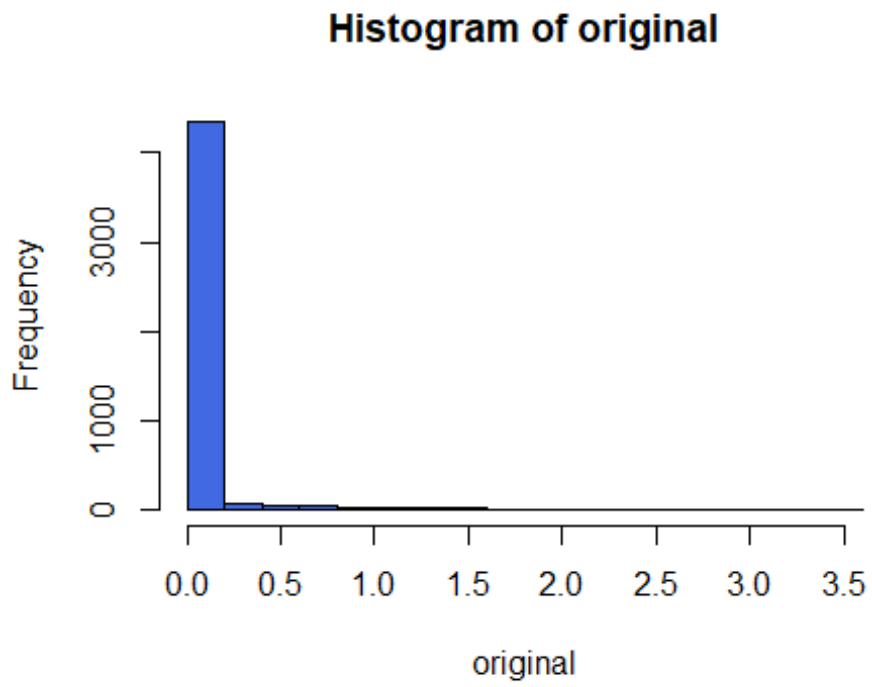

Histogram of parts
Frequency
0
1000
3000
0
2
4
6
8
parts

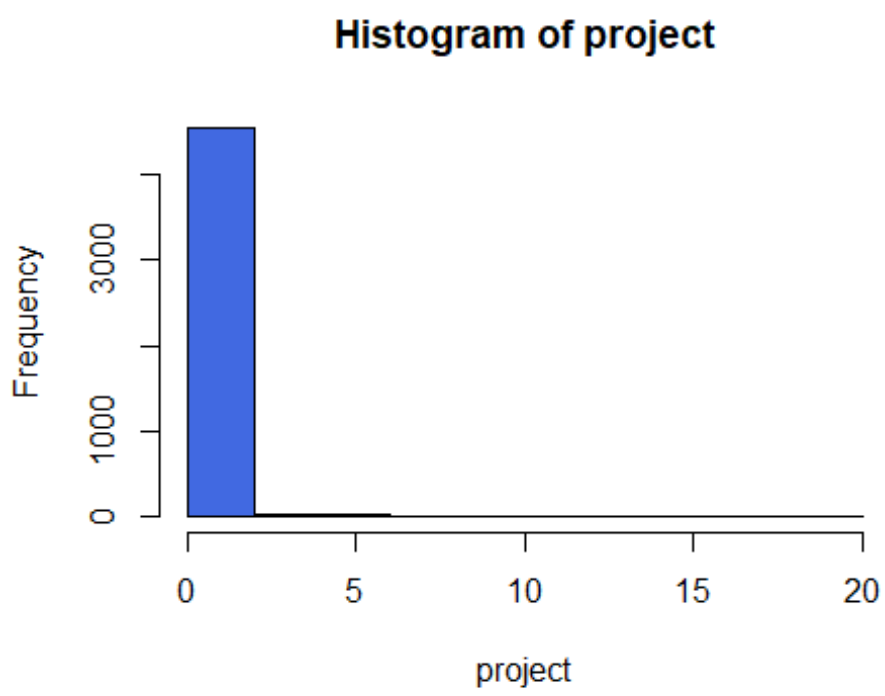

Histogram of pm
Frequency
0
1000
3000
0
2
4
6
8
10
12
pm

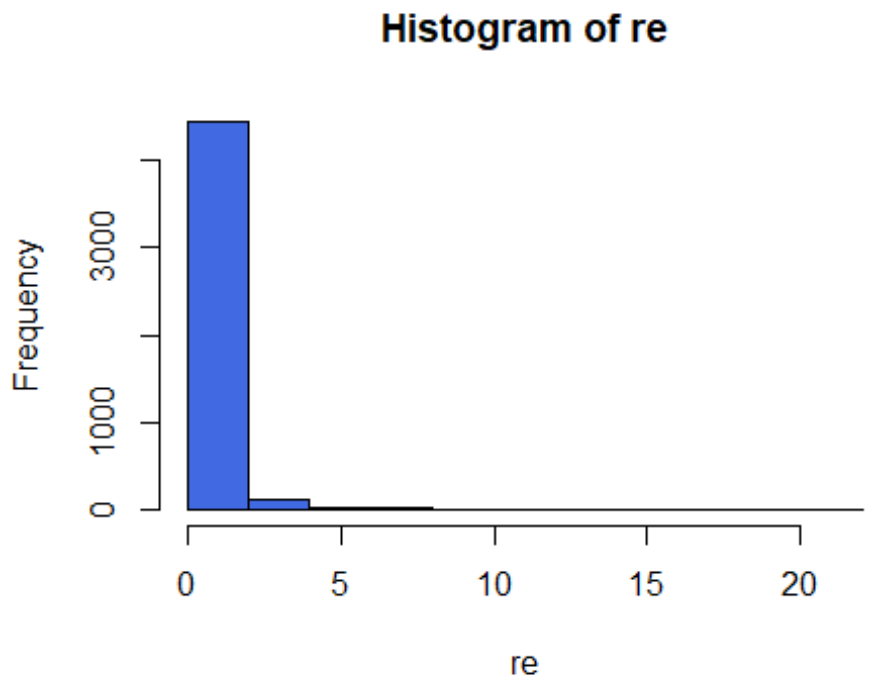

Histogram of direct
Frequency
0
1000
3000
0
1
2
3
4
5
direct

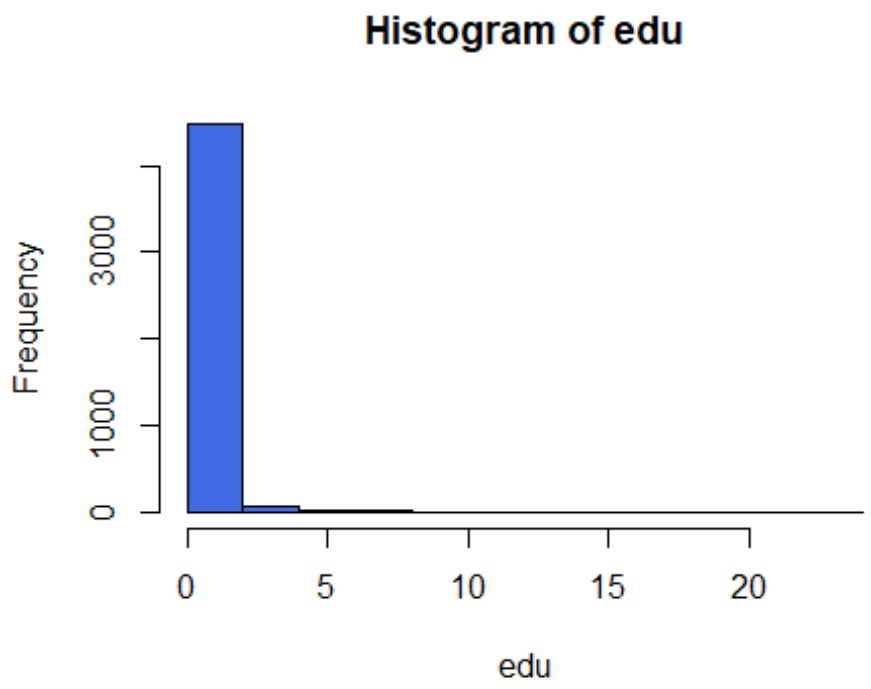

Histogram of cs
Frequency
0
1000
3000
0
2
4
6
cs

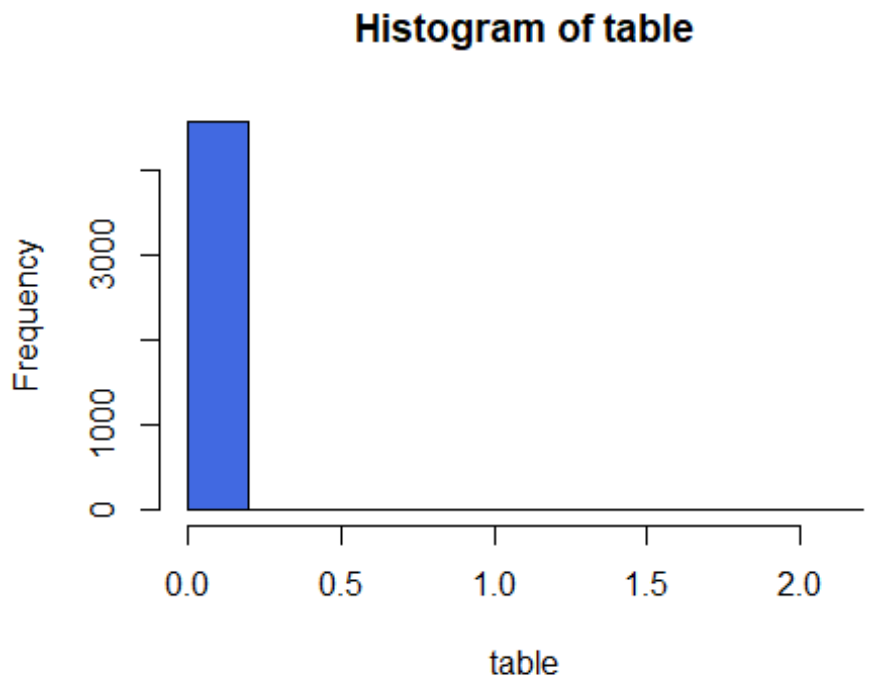

Histogram of meeting
Frequency
0
1000
3000
0
5
10
15
meeting

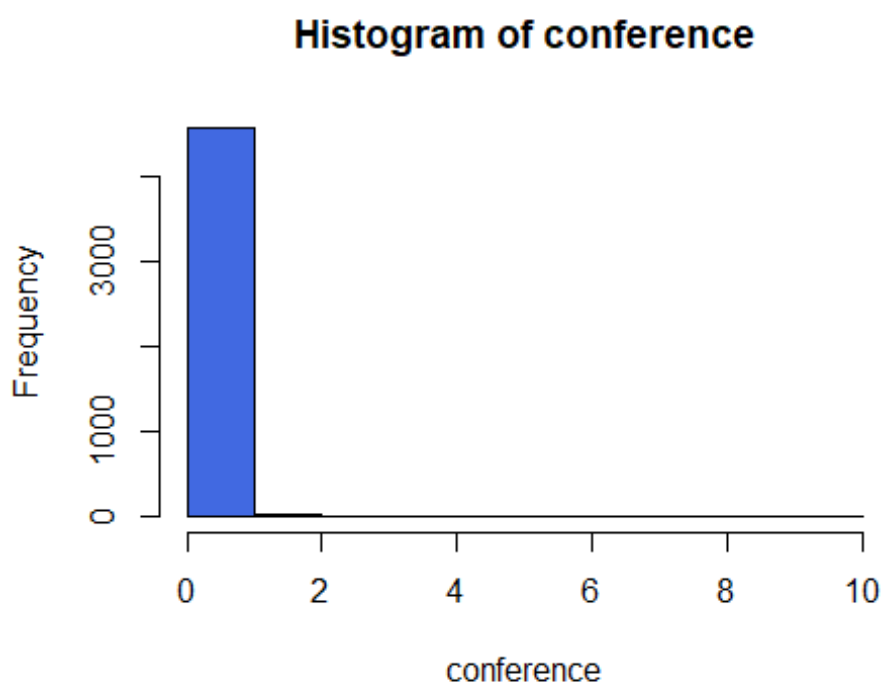

Histogram of original
Frequency
0
1000
3000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
original

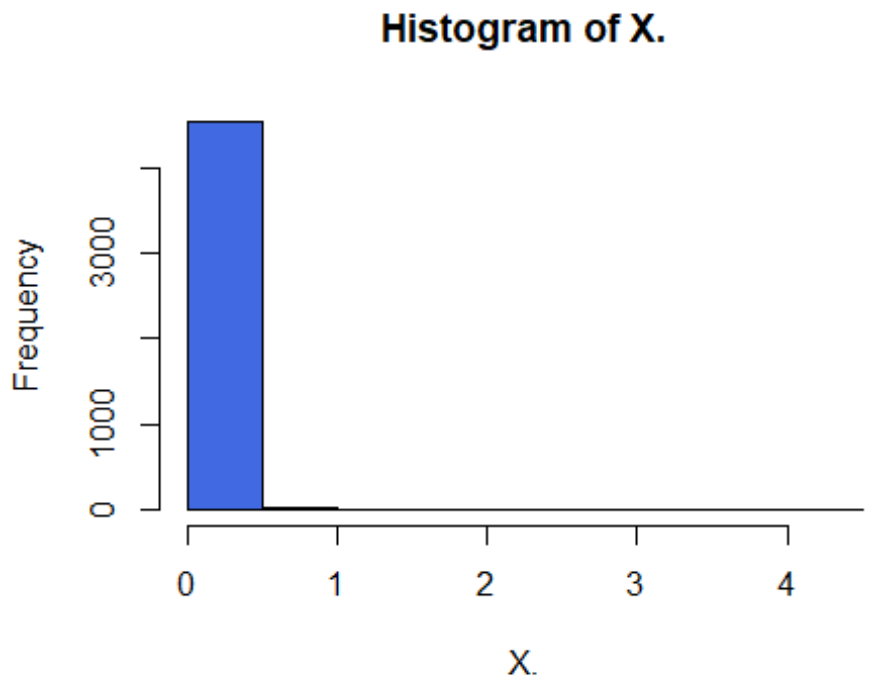

Histogram of project
Frequency
0
1000
3000
0
5
10
15
20
project

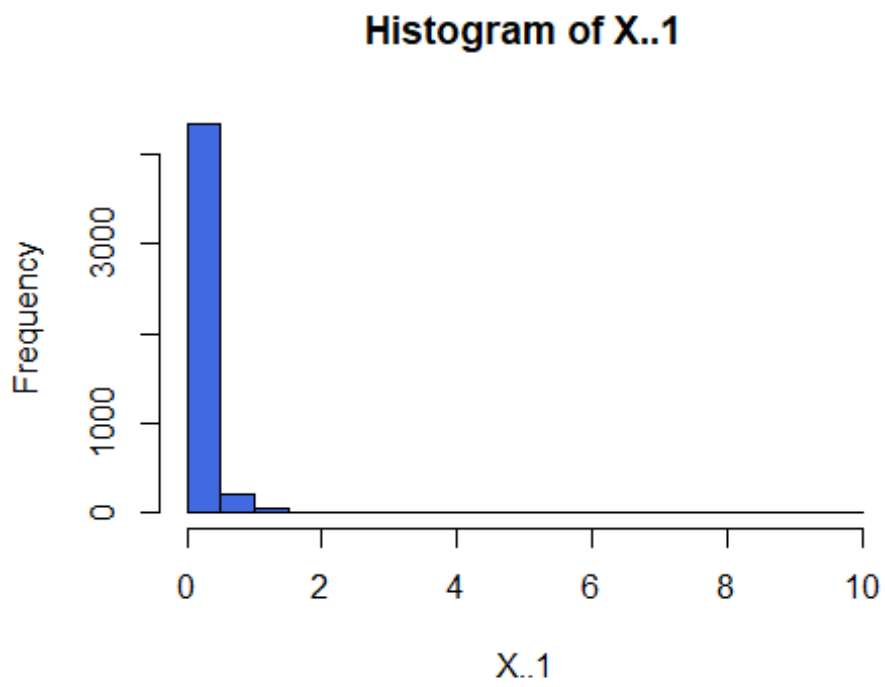

Histogram of re
Frequency
0
1000
3000
0
5
10
15
20
re

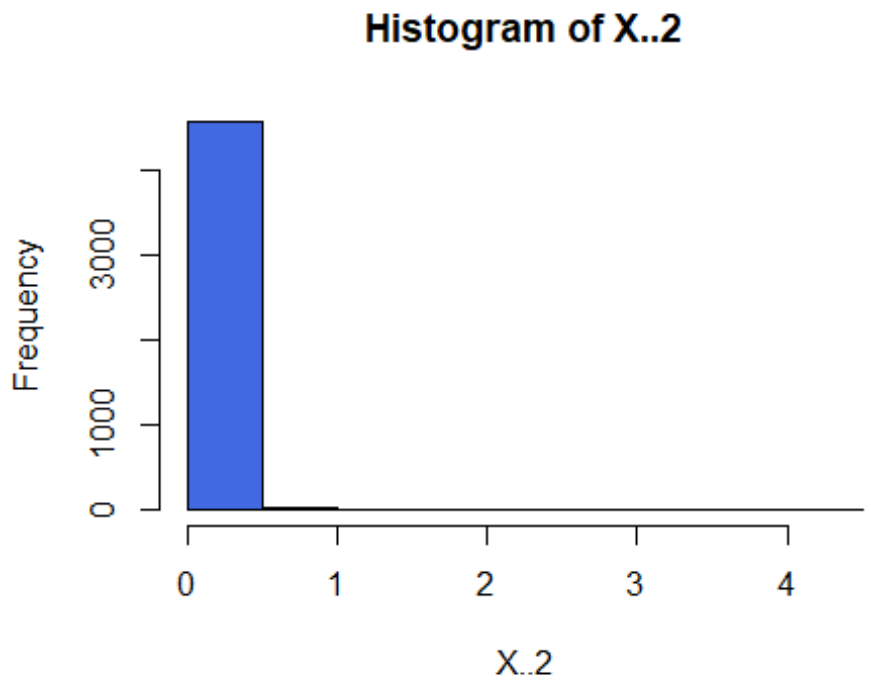

Histogram of edu
Frequency
0
1000
3000
0
5
10
15
20
edu

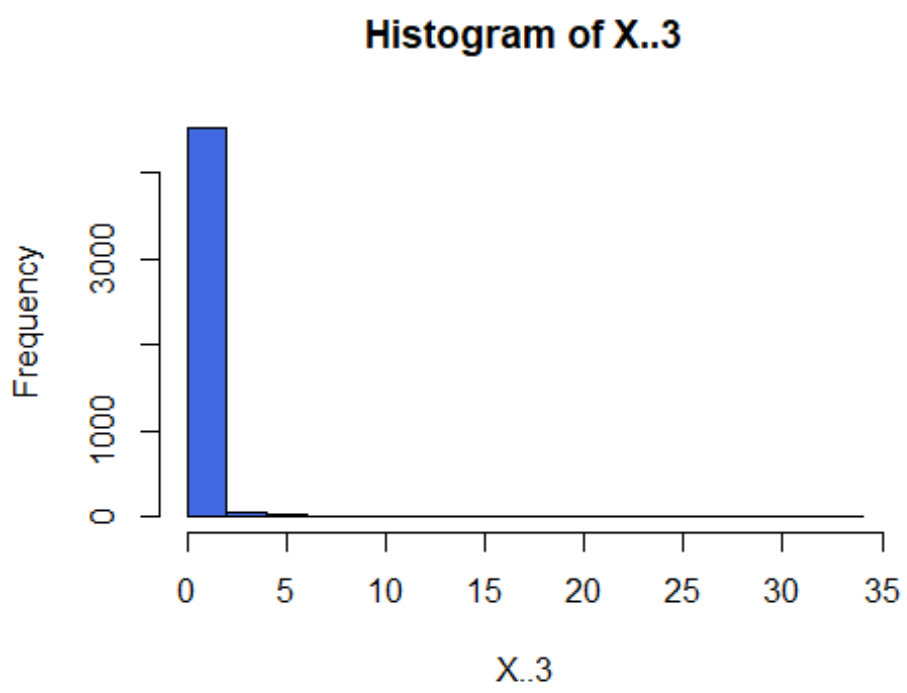

Histogram of table
Frequency
0
1000
3000
0.0
0.5
1.0
1.5
2.0
table

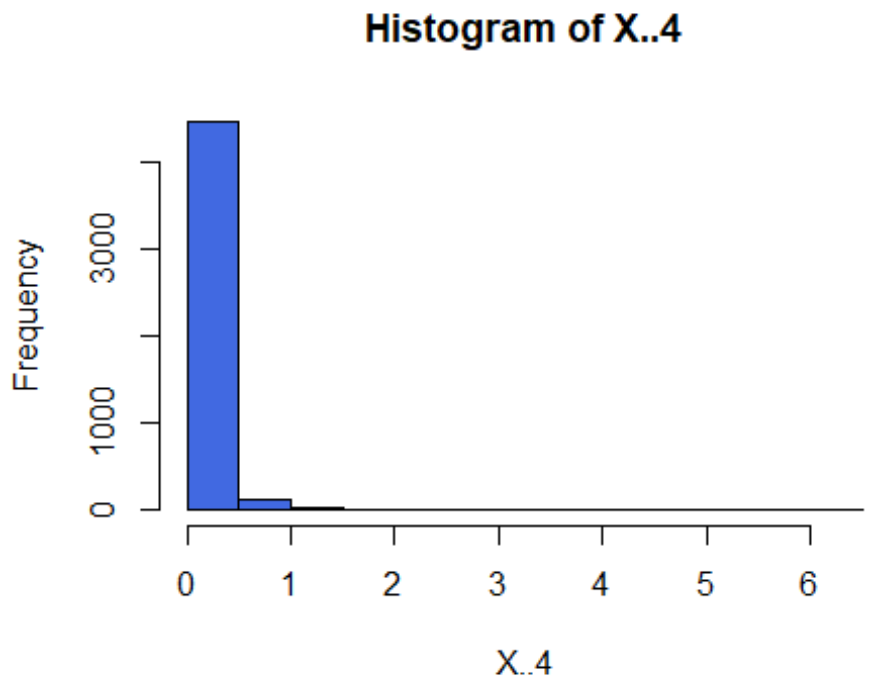

Histogram of conference
Frequency
0
1000
3000
0
2
4
6
8
10
conference

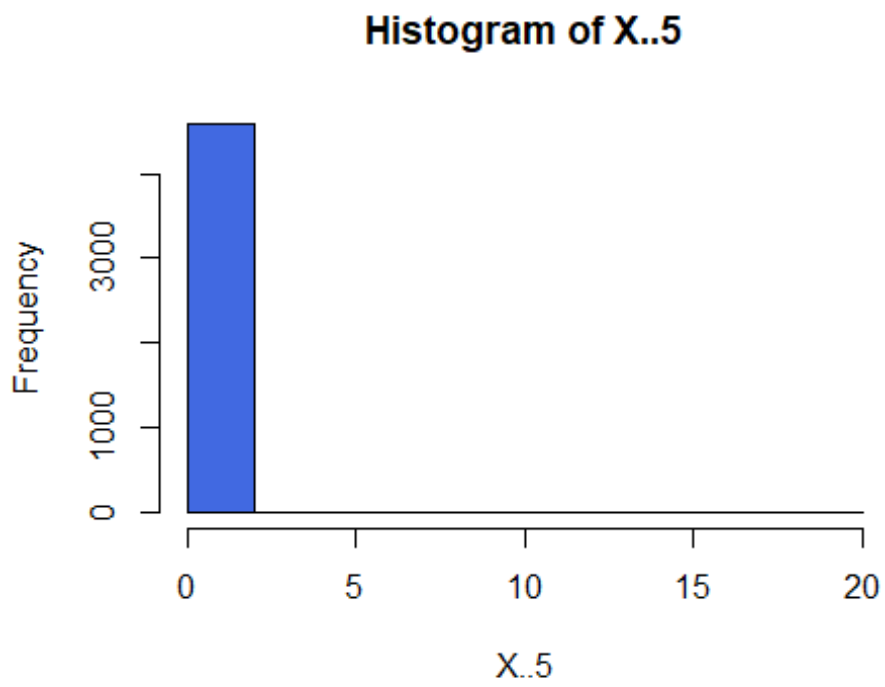

Histogram of X.
Frequency
0
1000
3000
0
1
2
3
4
X.

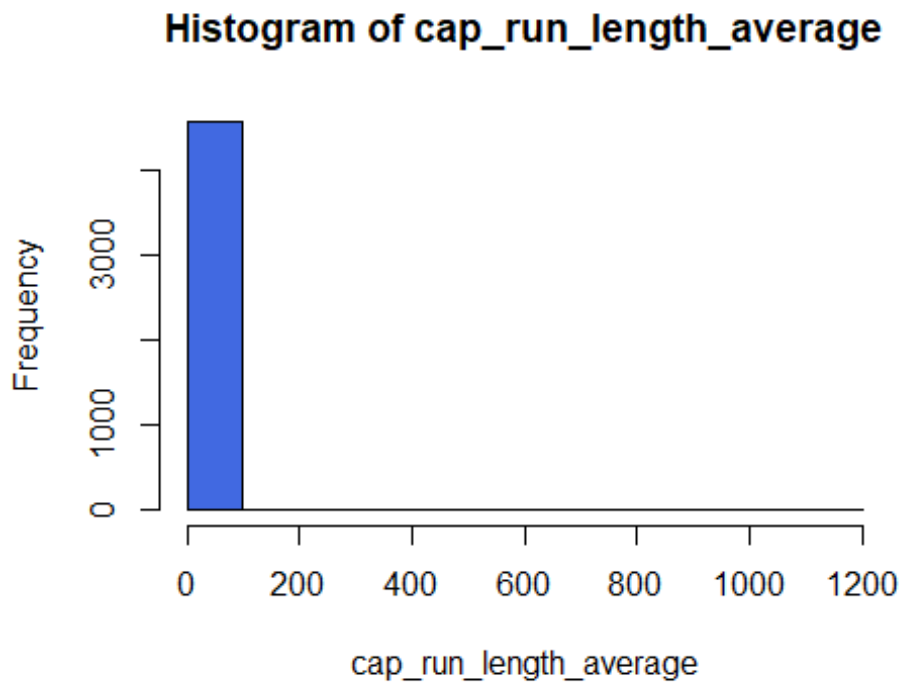

Histogram of X..1
Frequency
0
1000
3000
0
2
4
6
8
10
X..1

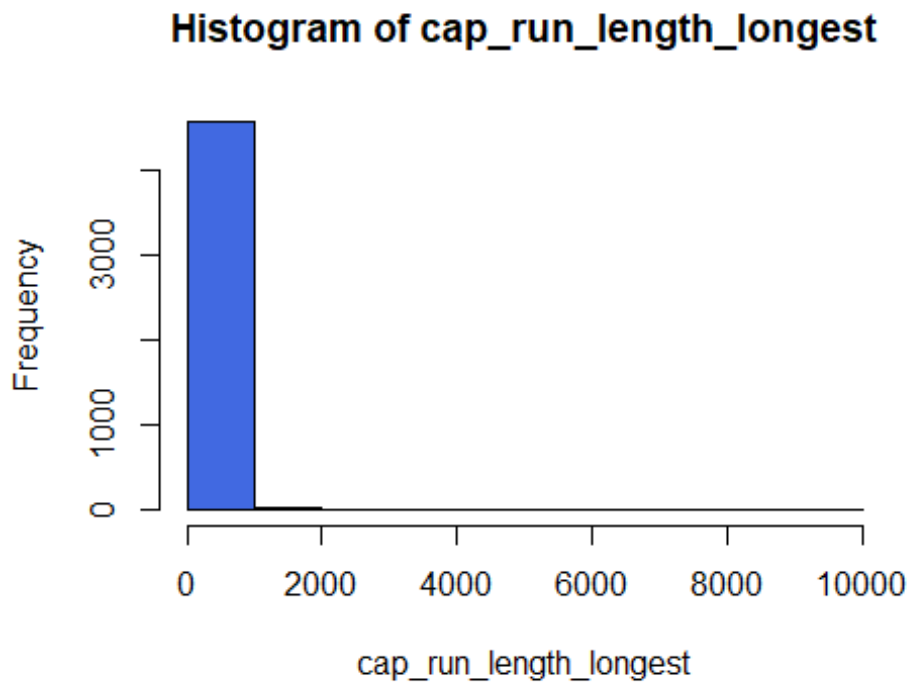

Histogram of X..2
Frequency
0
1000
3000
0
1
2
3
4
X..2

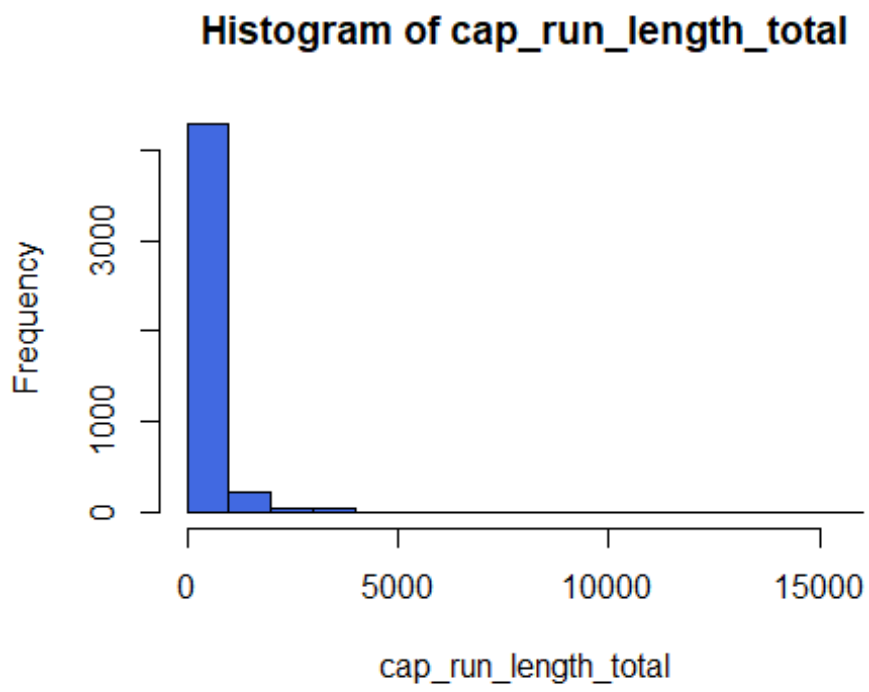

Histogram of X..3
Frequency
0
1000
3000
0
5
10
15
20
25
30
35
X..3


Histogram of X..4
Frequency
0
1000
3000
0
1
2
3
4
5
6
X..4


Histogram of X..5
Frequency
0
1000
3000
0
5
10
15
20
X..5


Histogram of cap_run_length_average
Frequency
0
1000
3000
0
200
400
600
800
1000
1200
cap_run_length_average





Revisamos la presencia de *outliers*

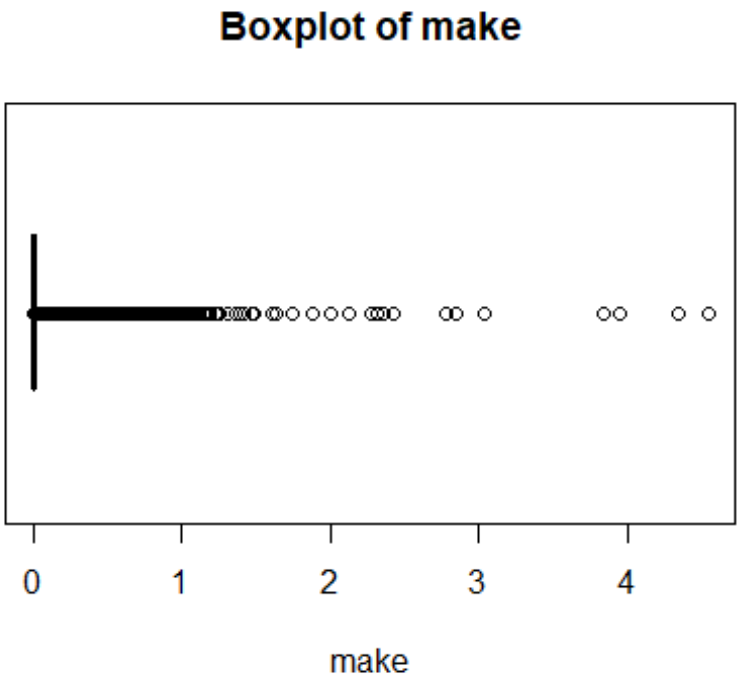
```r
detect_outliers <- function(inp, na.rm=TRUE) {
  i.qnt <- quantile(inp, probs=c(.25, .75), na.rm=na.rm)
  i.max <- 1.5 * IQR(inp, na.rm=na.rm)
  otp <- inp
  otp[inp < (i.qnt[1] - i.max)] <- NA
  otp[inp > (i.qnt[2] + i.max)] <- NA

  sum(is.na(otp))
}
```

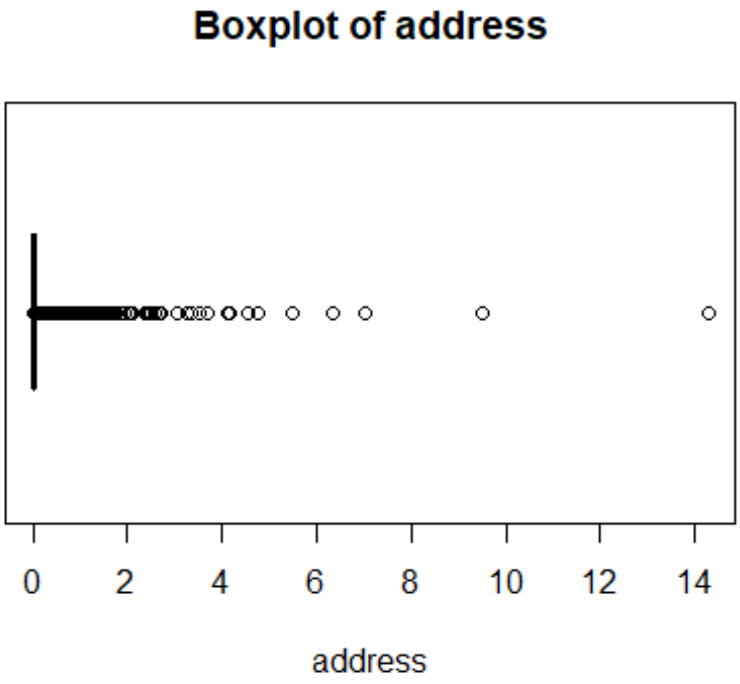
Podemos ver el número en cada variable. Tomamos *make* y *address* como ejemplo:

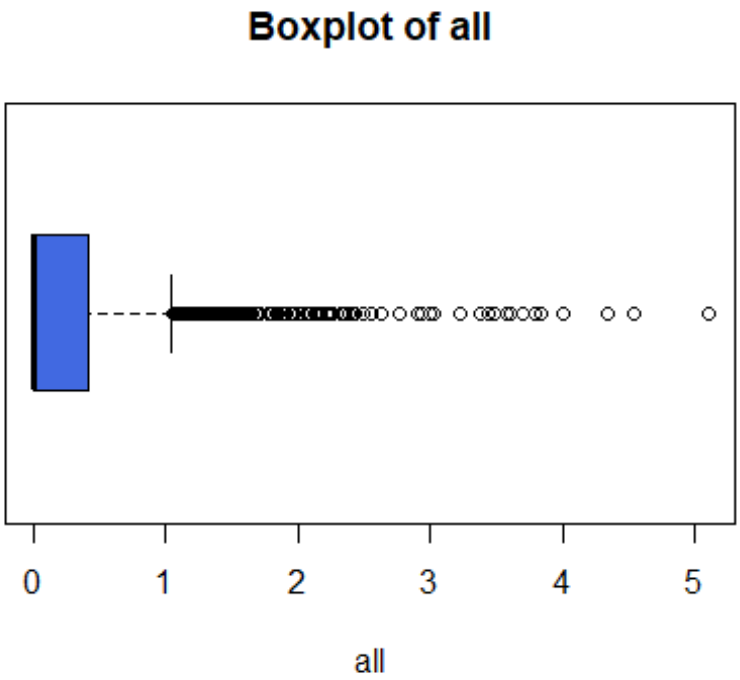
```r
detect_outliers(spam$make)

## [1] 1048

detect_outliers(spam$address)

## [1] 904
```

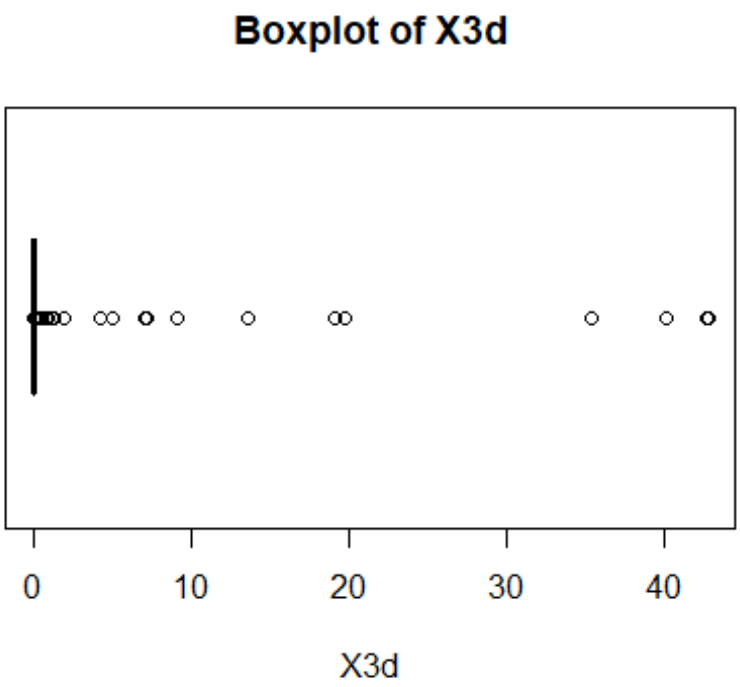
Para visualizar esta información usamos boxplots. Con los gráficos podremos confirmar lo que el análisis estadístico sugería. El número elevado de *outliers*. No haremos ningún tratamiento de ellos porque la mayor parte de las columnas tienen valores cero *media* si queremos reemplazarlos con este valor.

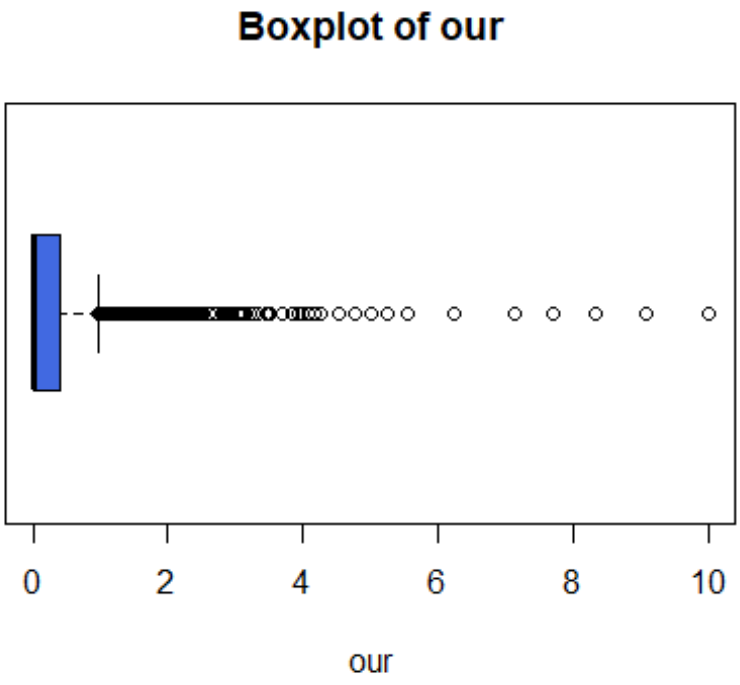
```r
for (var in colnames(spam)[1:57]) {
  boxplot(unlist(spam[,var]),
          horizontal = TRUE,
          col='royalblue',
          main= paste('Boxplot of', var),
          xlab=var)
}
```

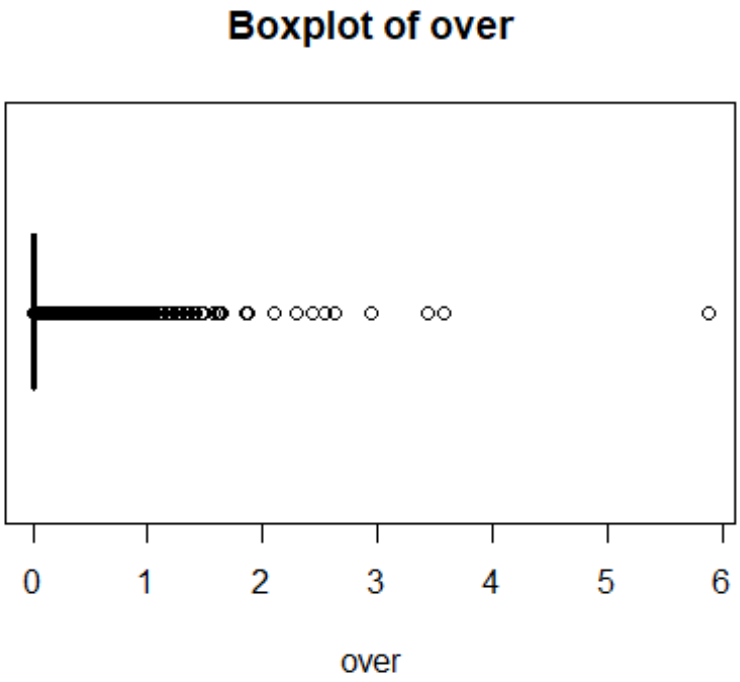

Boxplot of make
0
1
2
3
4
make

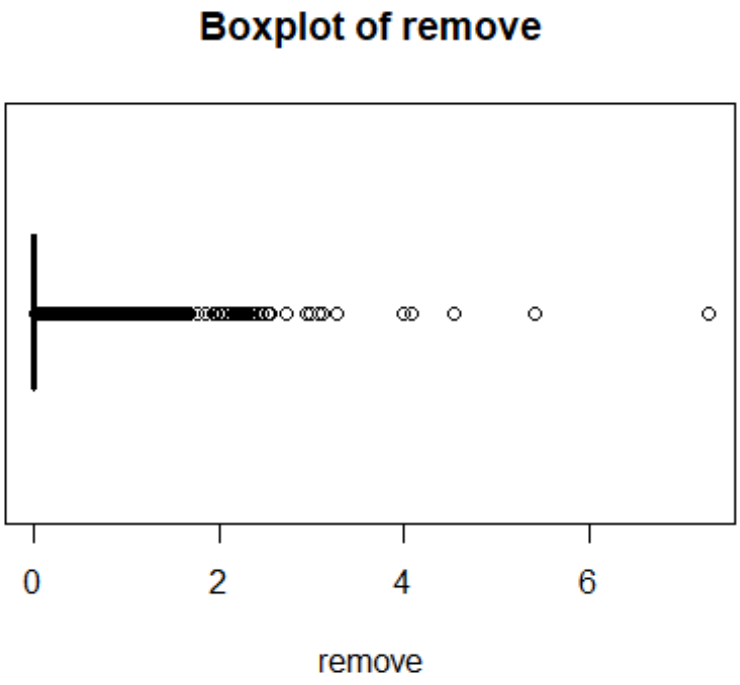

Boxplot of address
0
2
4
6
8
10
12
14
address

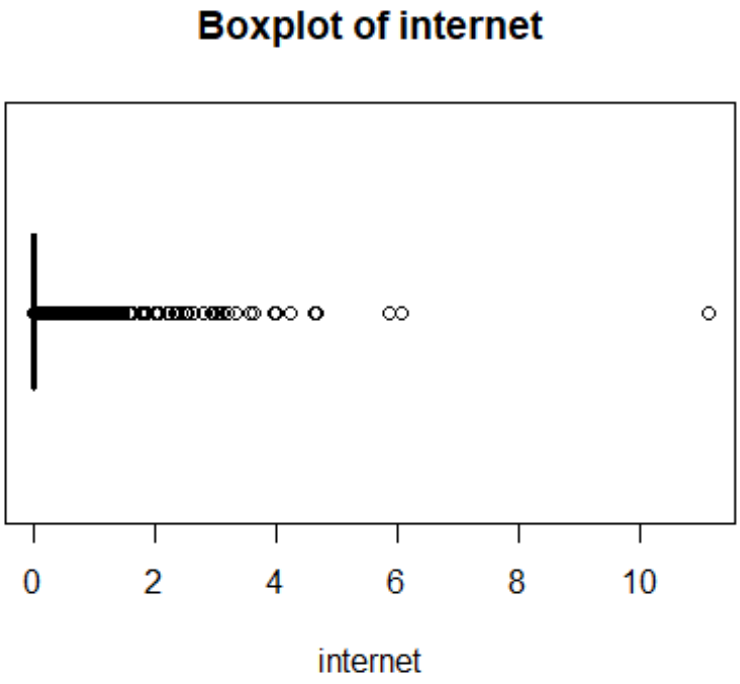

Boxplot of all
0
1
2
3
4
5
all

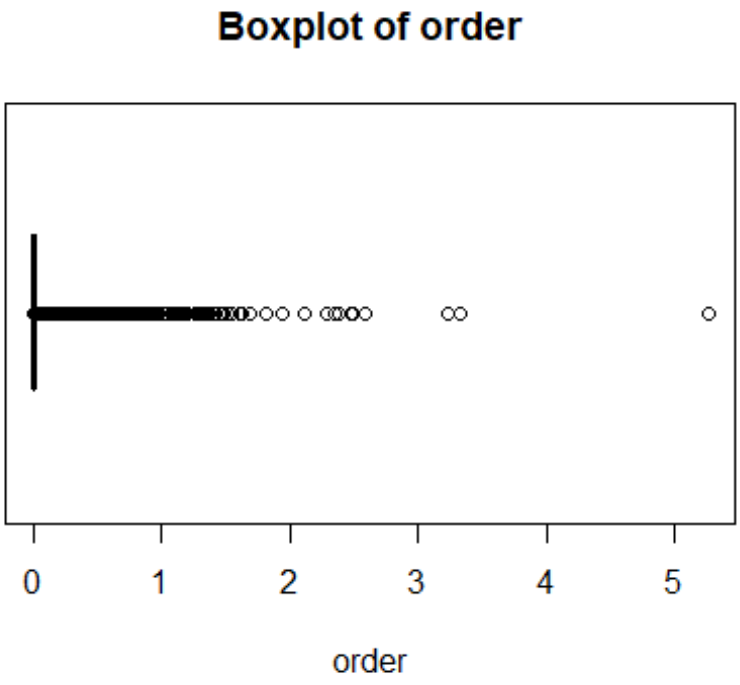

Boxplot of X3d
0
10
20
30
40
X3d

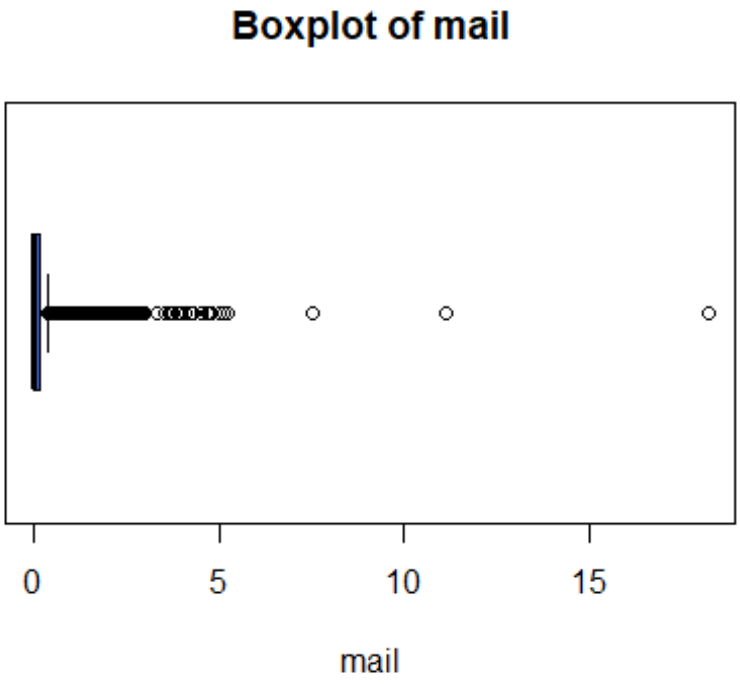

Boxplot of our
0
2
4
6
8
10
our

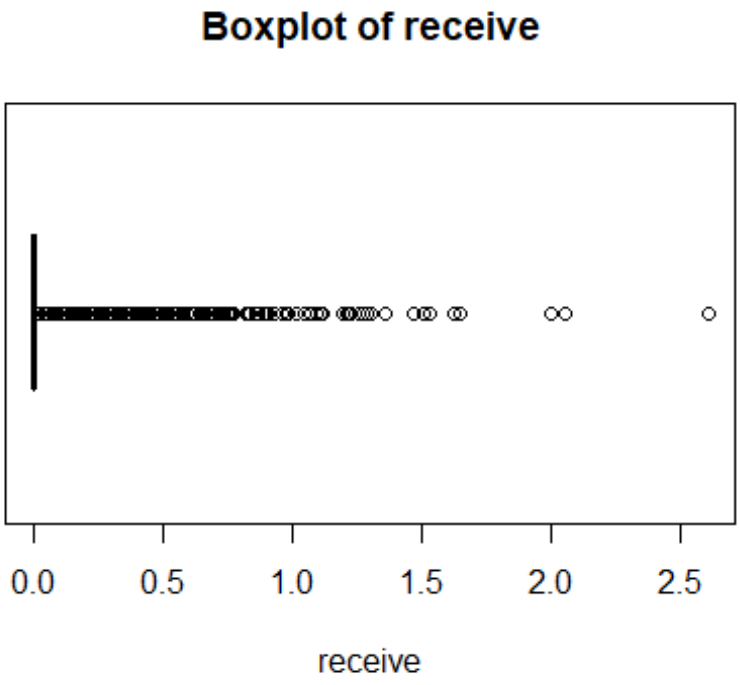

Boxplot of over
0
1
2
3
4
5
6
over

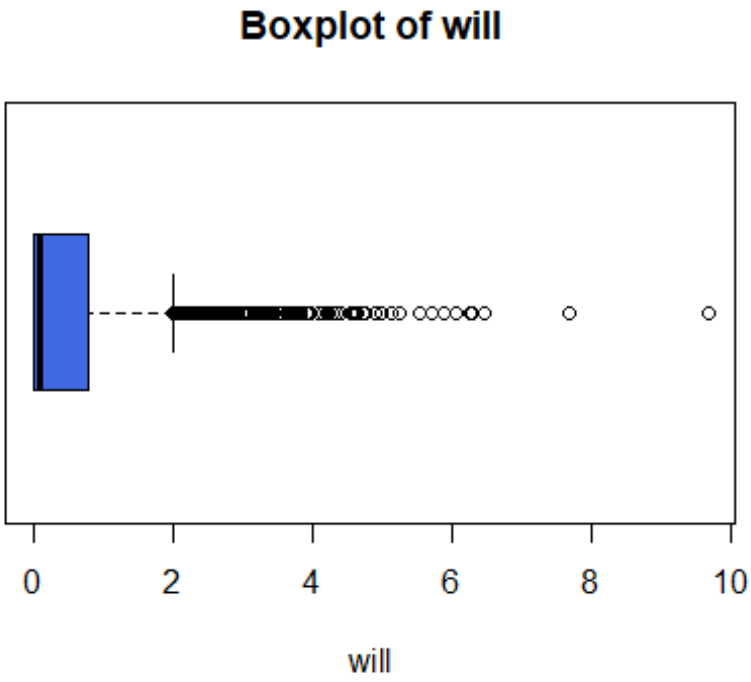

Boxplot of remove
0
2
4
6
remove

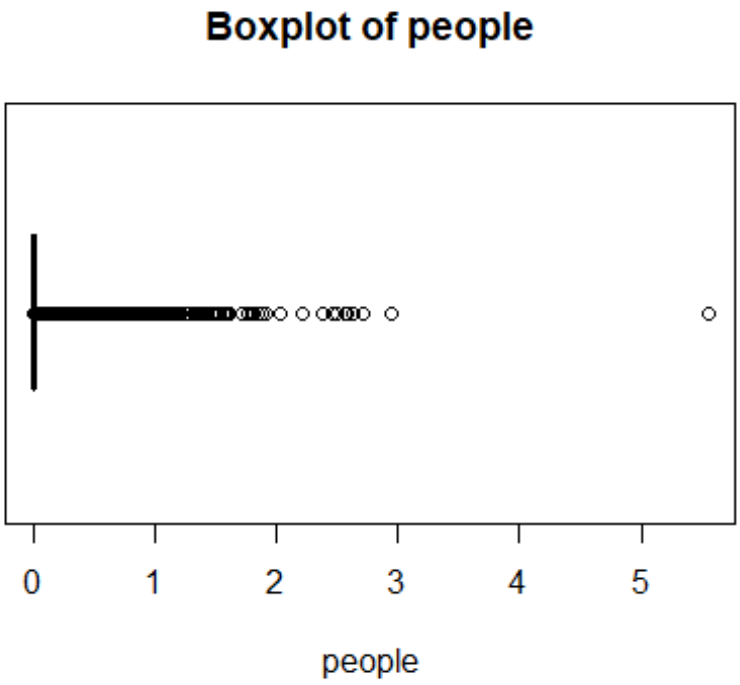

Boxplot of internet
0
2
4
6
8
10
internet

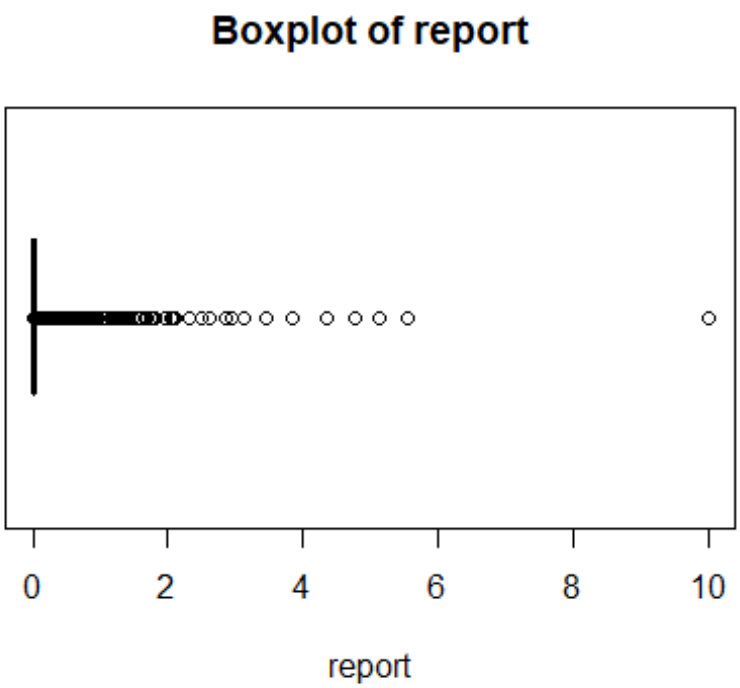

Boxplot of order
0
1
2
3
4
5
order

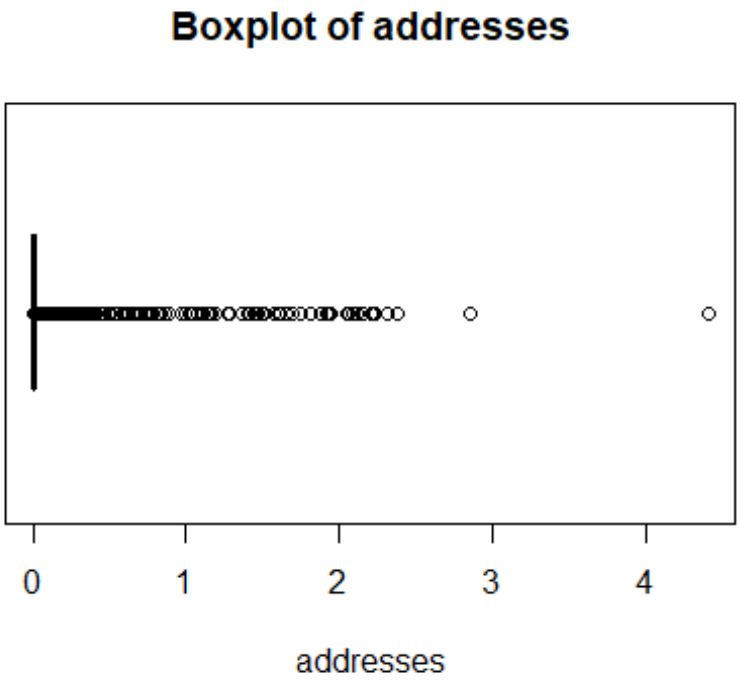

Boxplot of mail
0
5
10
15
mail

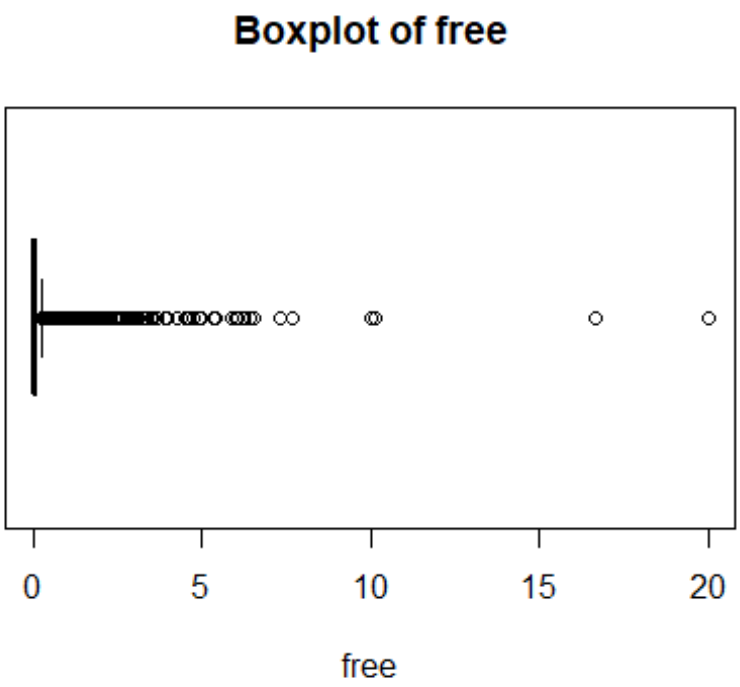

Boxplot of receive
0.0
0.5
1.0
1.5
2.0
2.5
receive

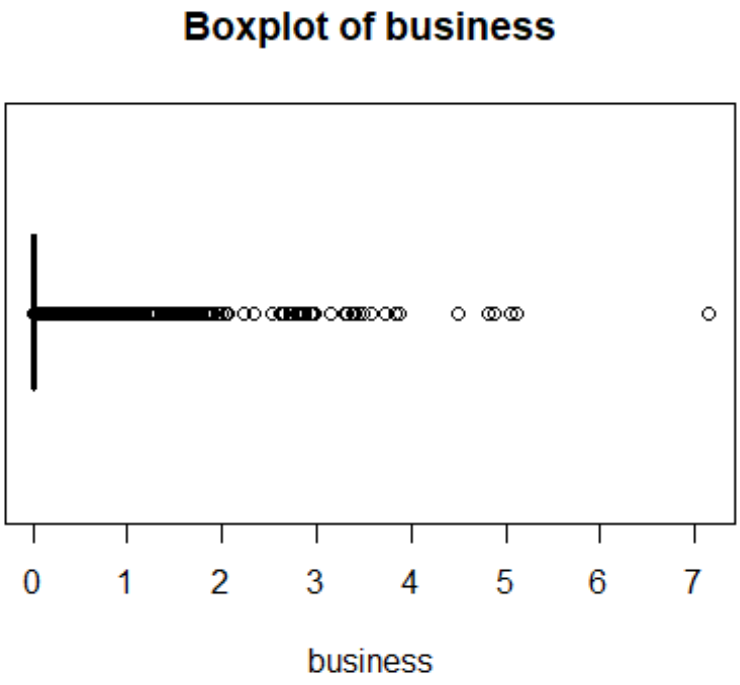

Boxplot of will
0
2
4
6
8
10
will

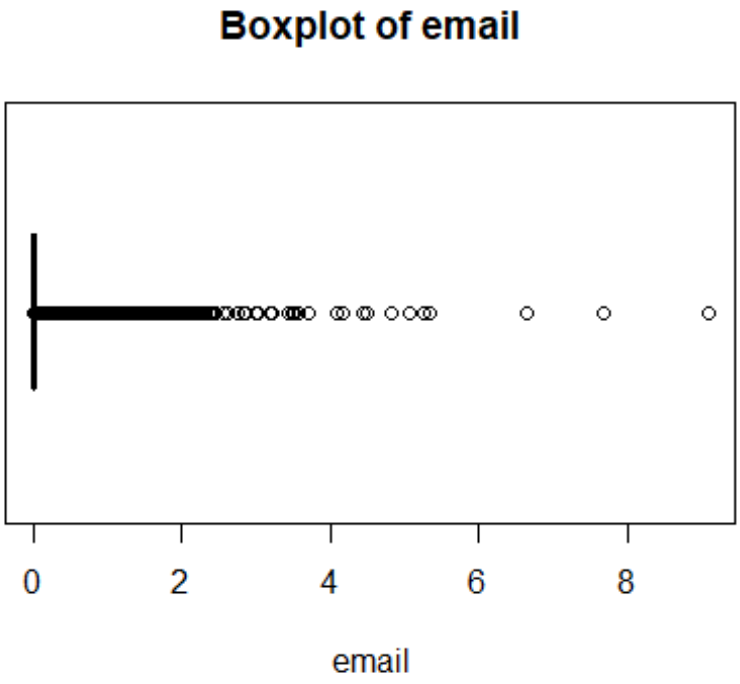

Boxplot of people
0
1
2
3
4
5
people

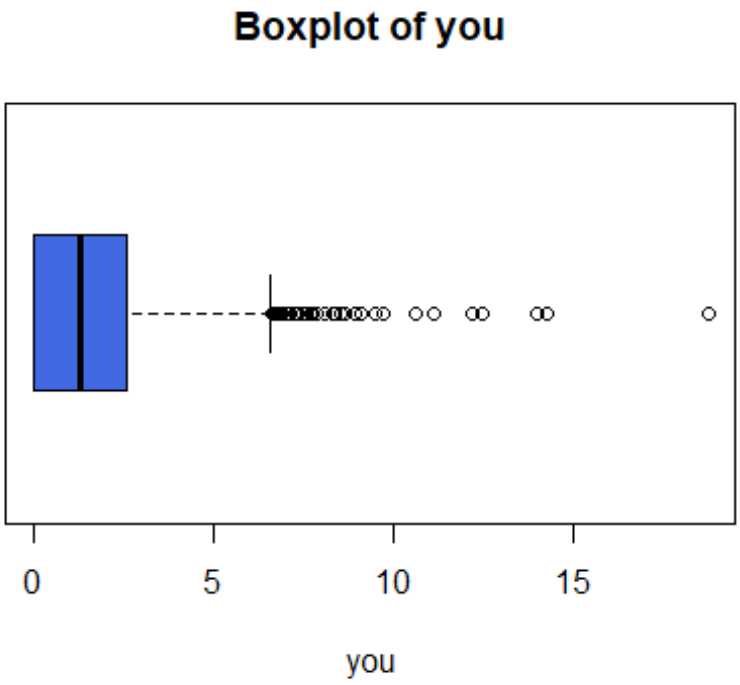

Boxplot of report
0
2
4
6
8
10
report

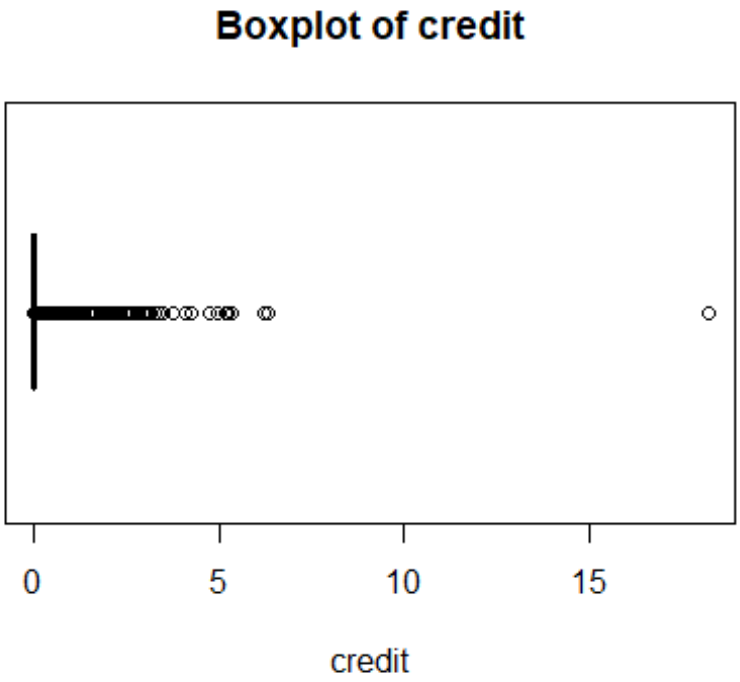

Boxplot of addresses
0
1
2
3
4
addresses

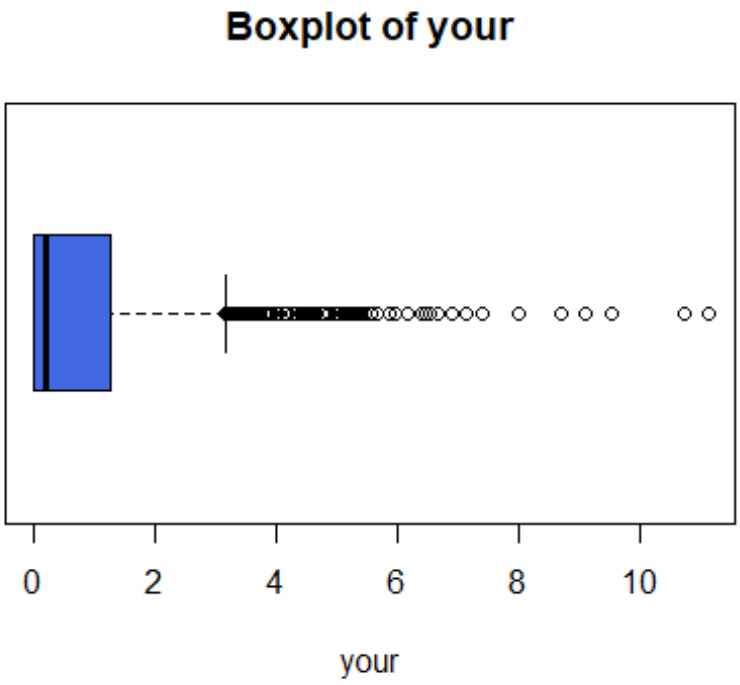

Boxplot of free
0
5
10
15
20
free

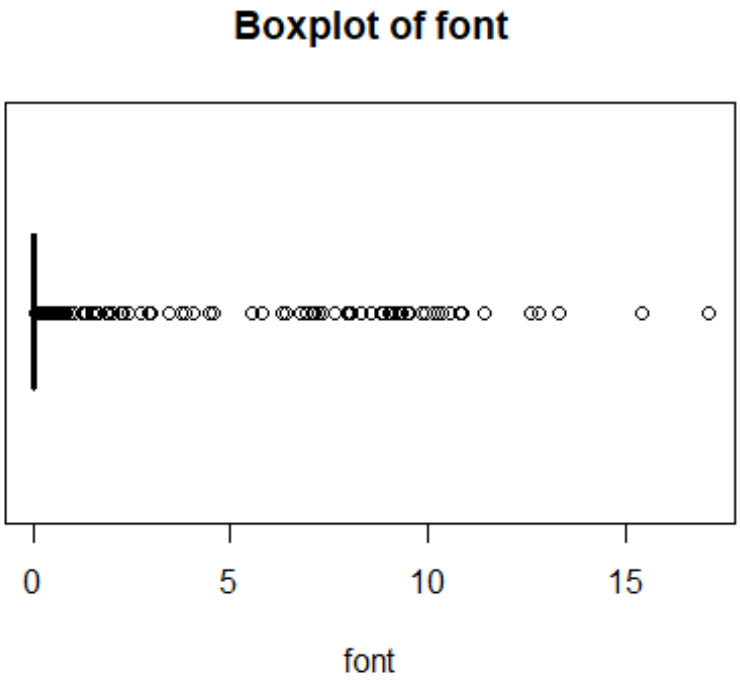

Boxplot of business
0
1
2
3
4
5
6
7
business

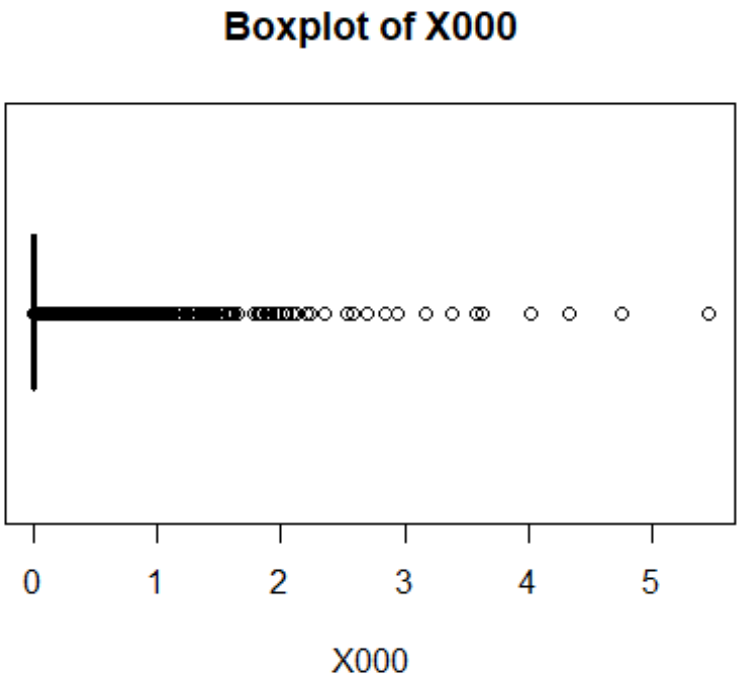

Boxplot of email
0
2
4
6
8
email

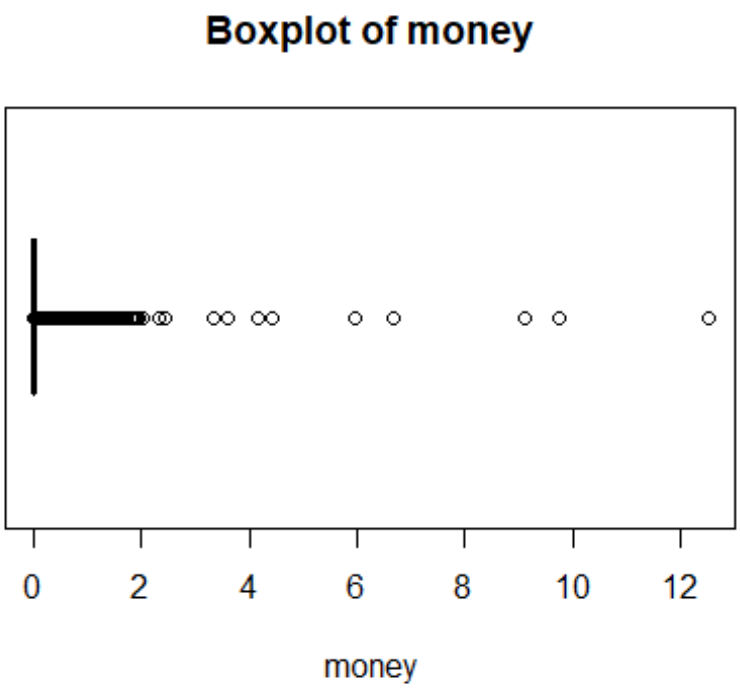

Boxplot of you
0
5
10
15
you

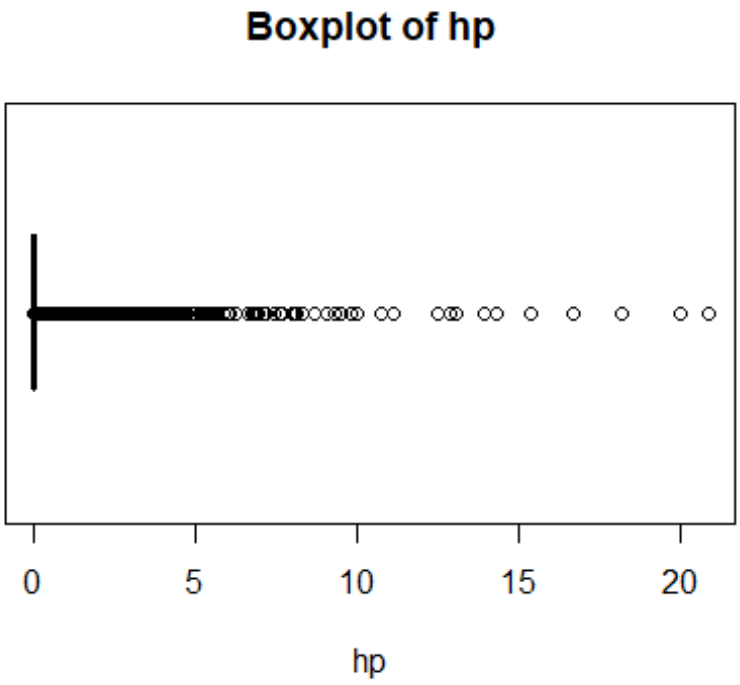

Boxplot of credit
0
5
10
15
credit

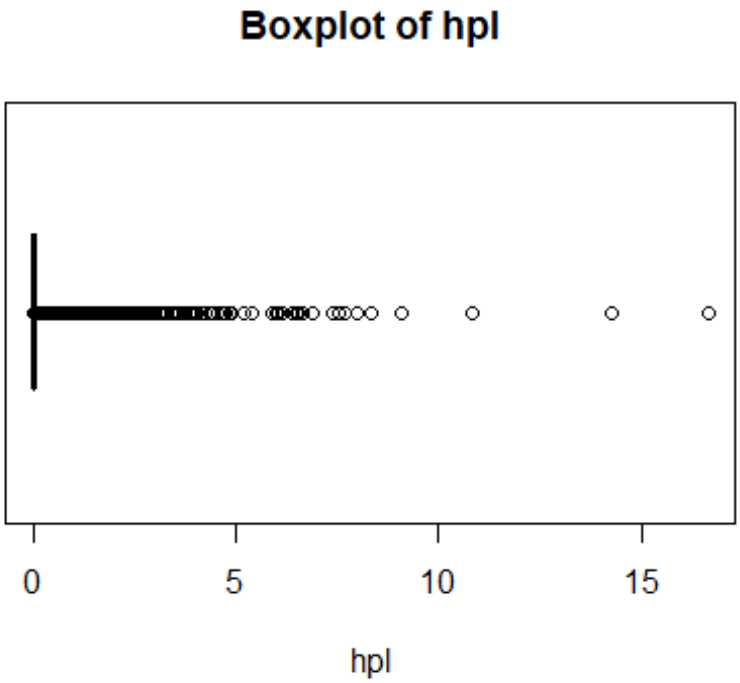

Boxplot of your
0
2
4
6
8
10
your

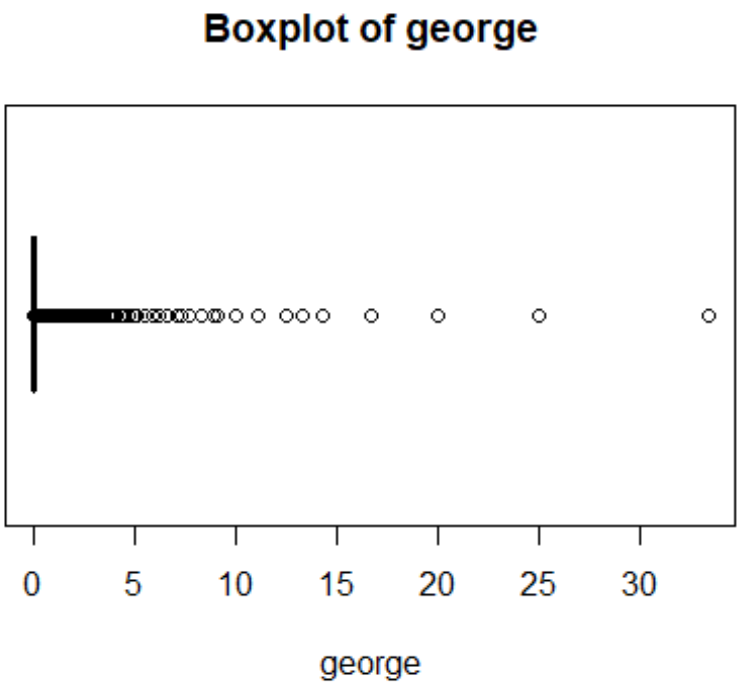

Boxplot of font
0
5
10
15
font

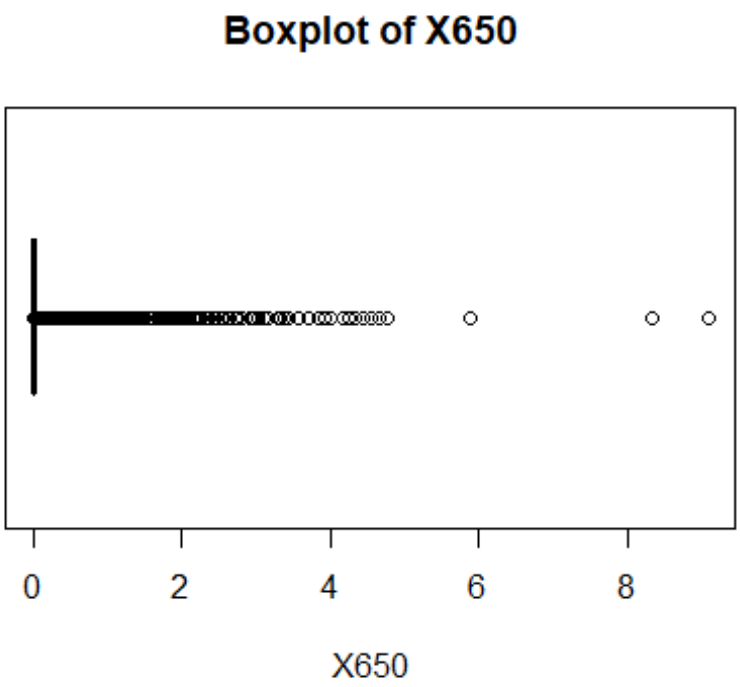

Boxplot of X000
0
1
2
3
4
5
X000

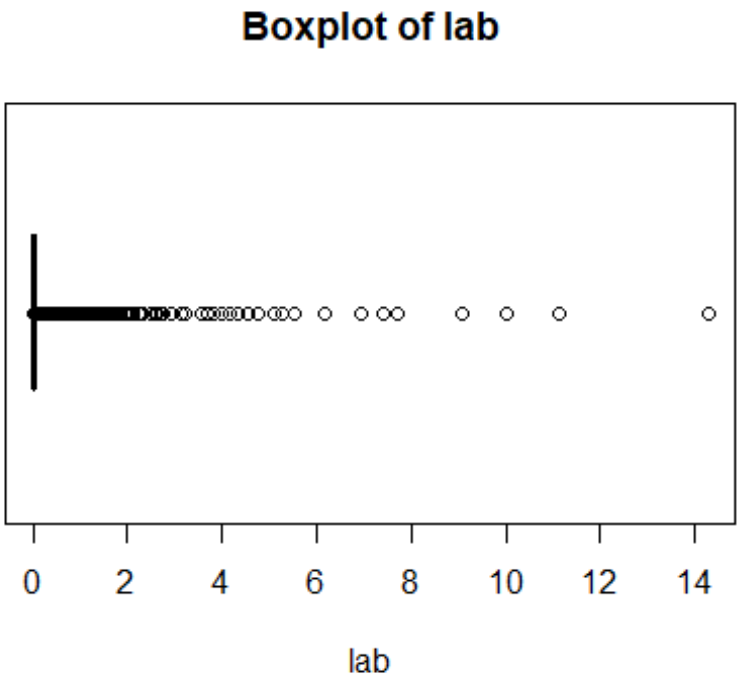

Boxplot of money
0
2
4
6
8
10
12
money

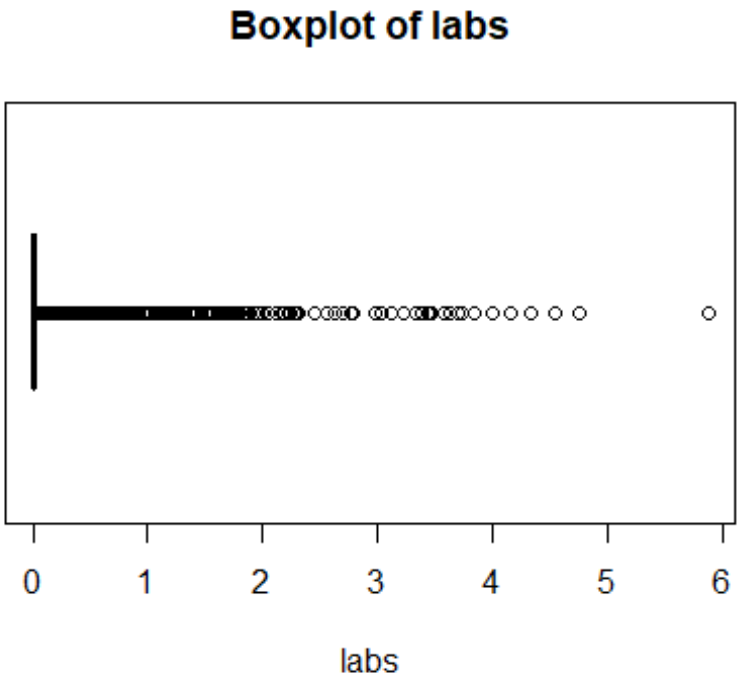

Boxplot of hp
0
5
10
15
20
hp

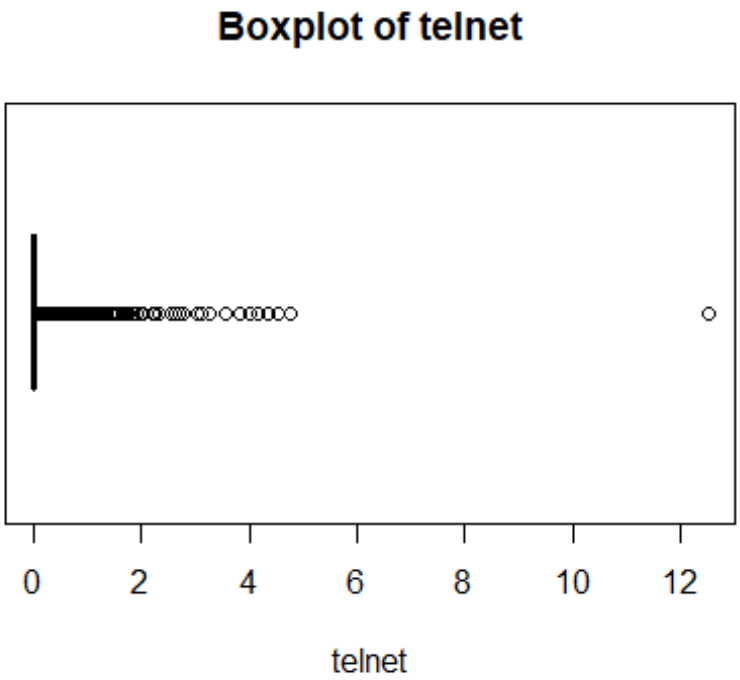

Boxplot of hpl
0
5
10
15
hpl

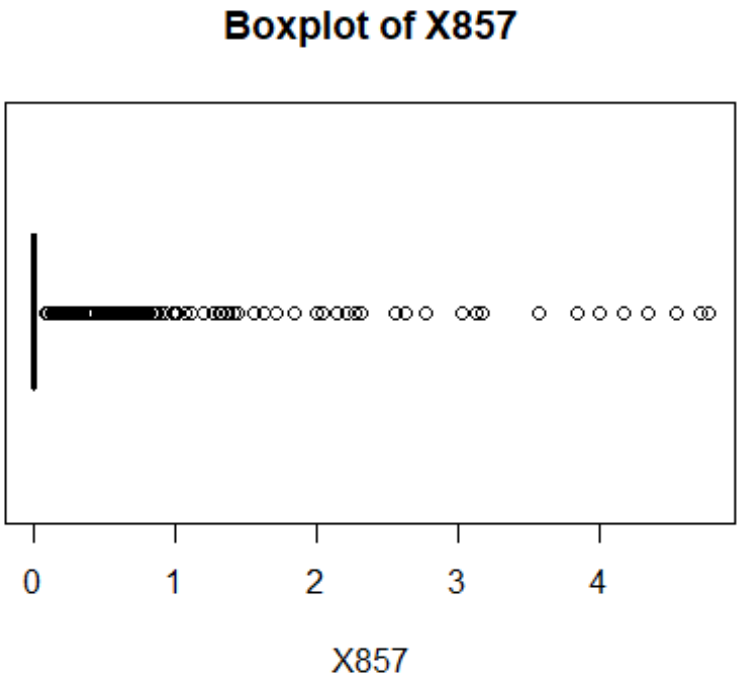

Boxplot of george
0
5
10
15
20
25
30
george

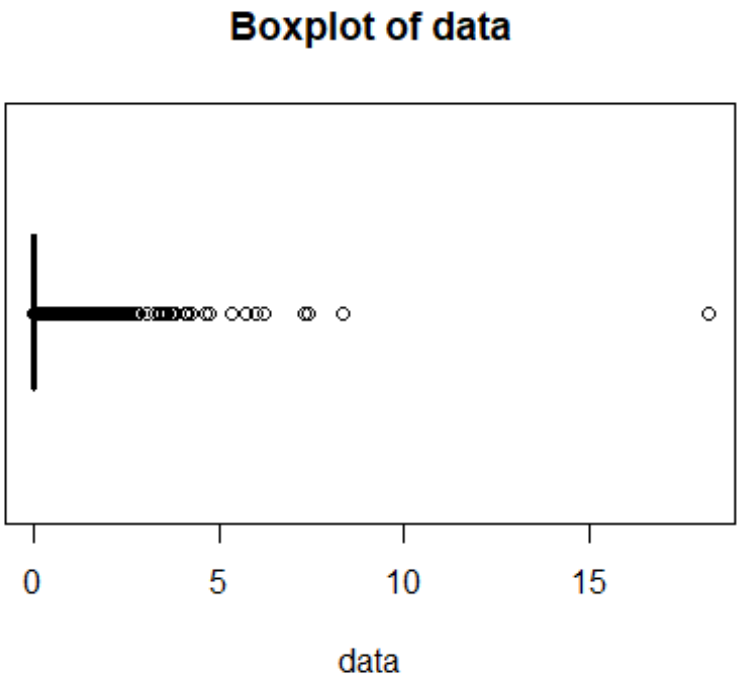

Boxplot of X650
0
2
4
6
8
X650

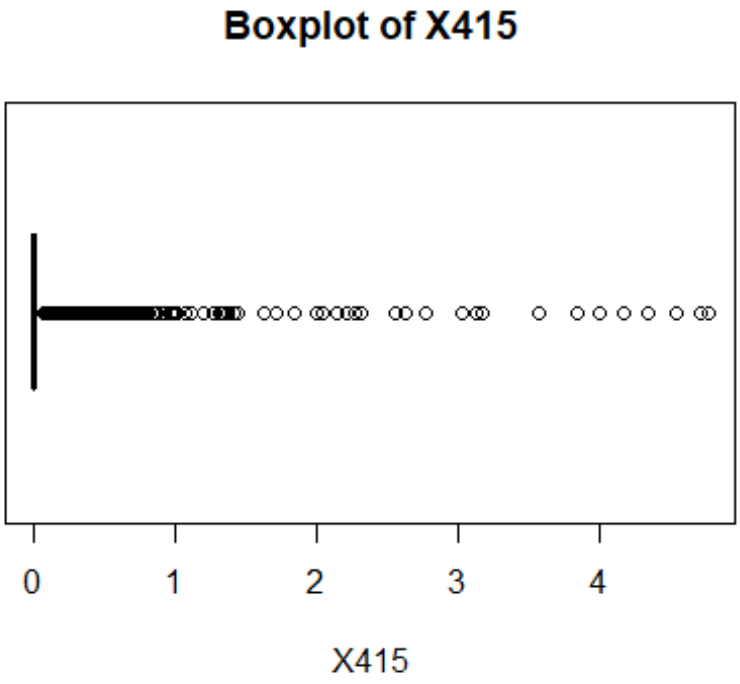

Boxplot of lab
0
2
4
6
8
10
12
14
lab

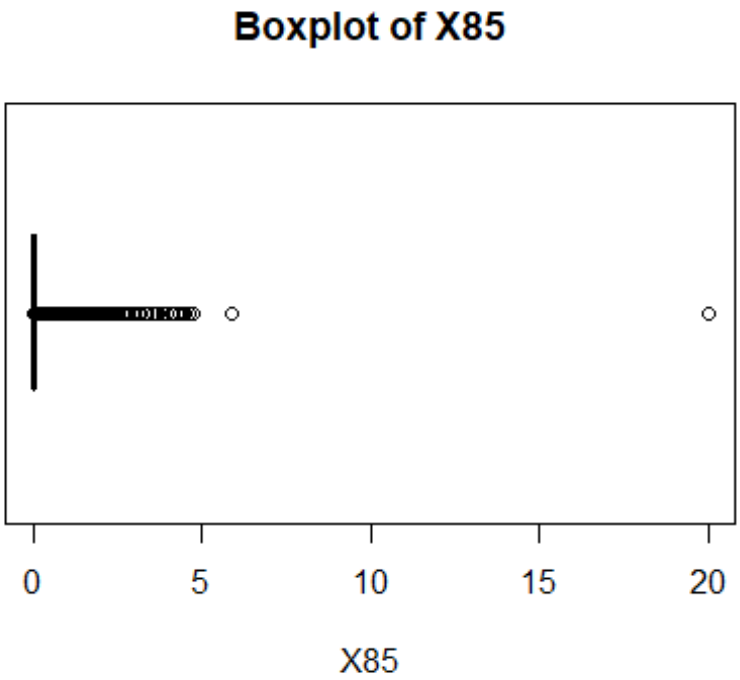

Boxplot of labs
0
1
2
3
4
5
6
labs

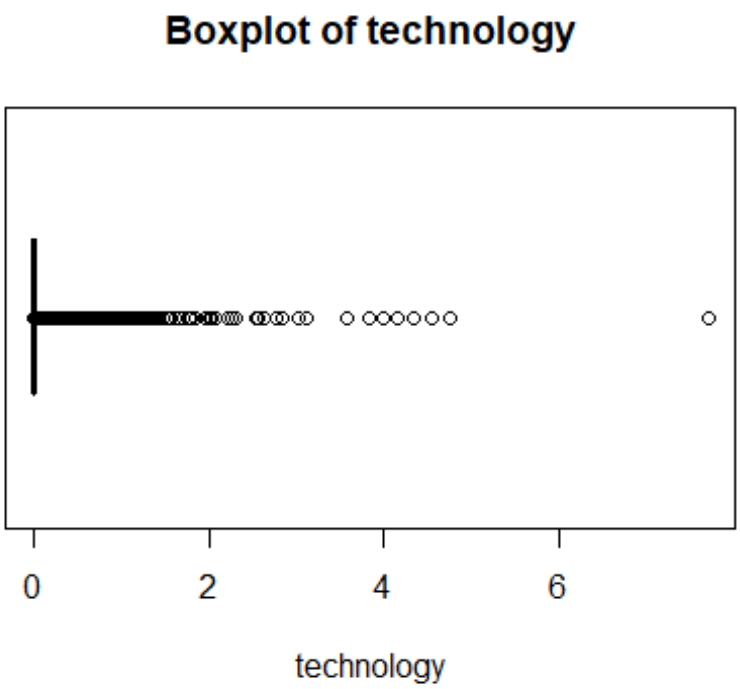

Boxplot of telnet
0
2
4
6
8
10
12
telnet

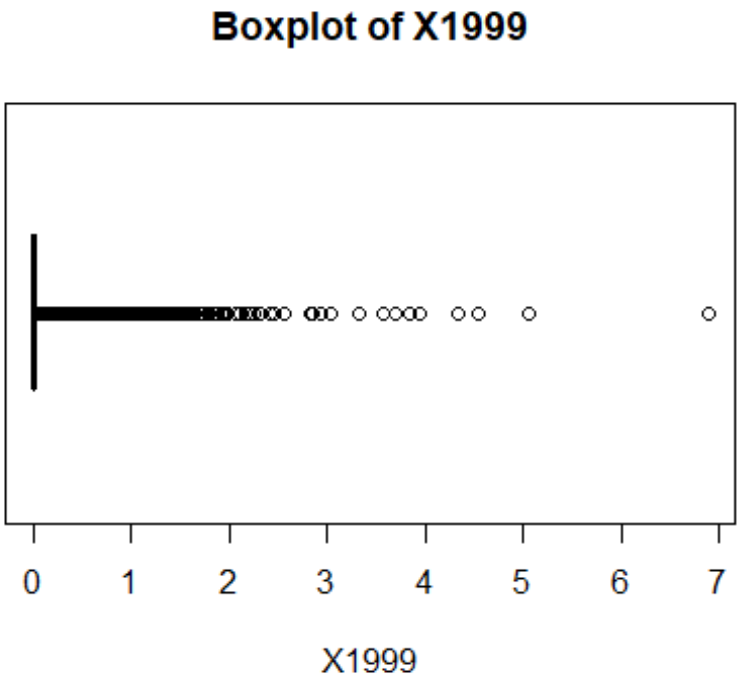

Boxplot of X857
0
1
2
3
4
X857

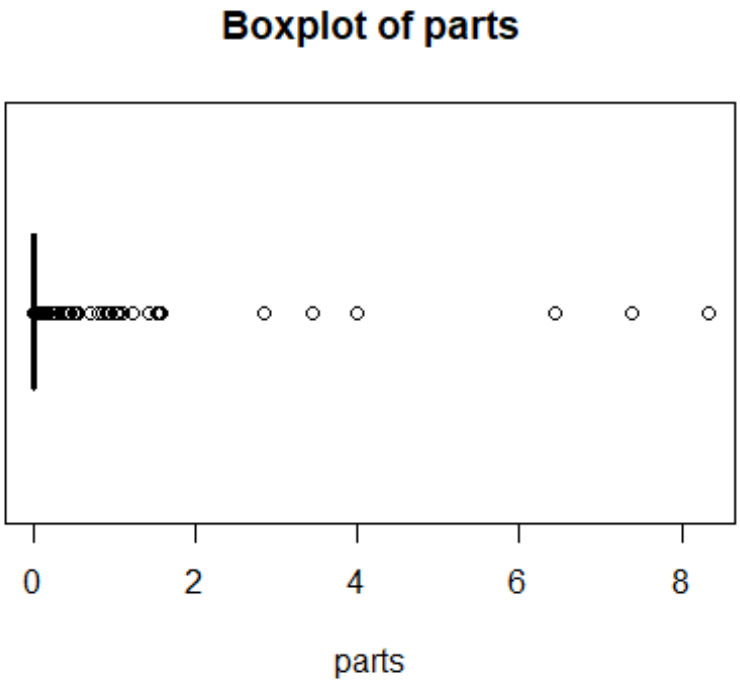

Boxplot of data
0
5
10
15
data

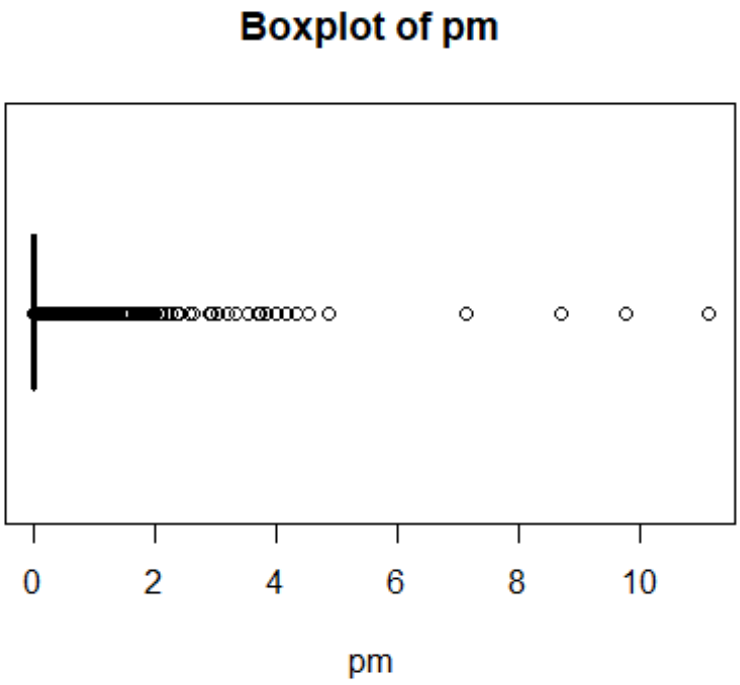

Boxplot of X415
0
1
2
3
4
X415

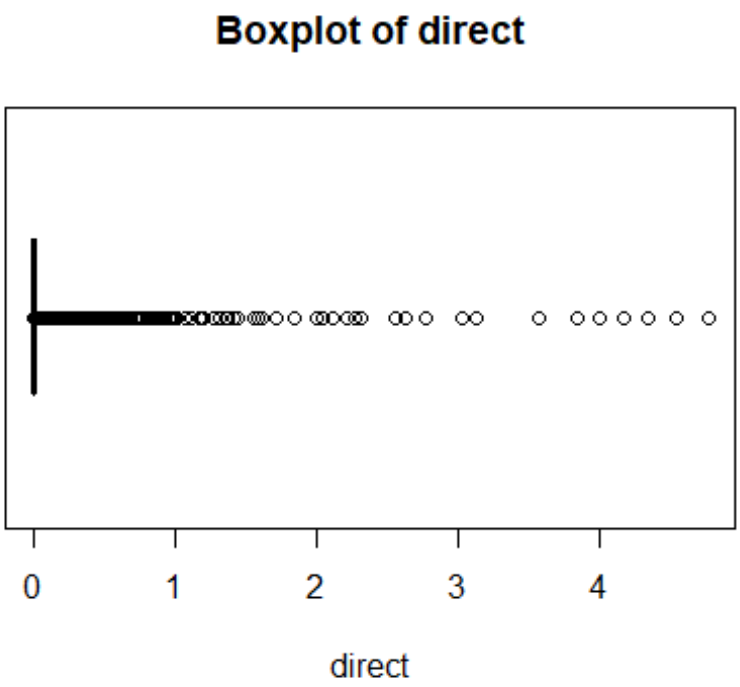

Boxplot of X85
0
5
10
15
20
X85

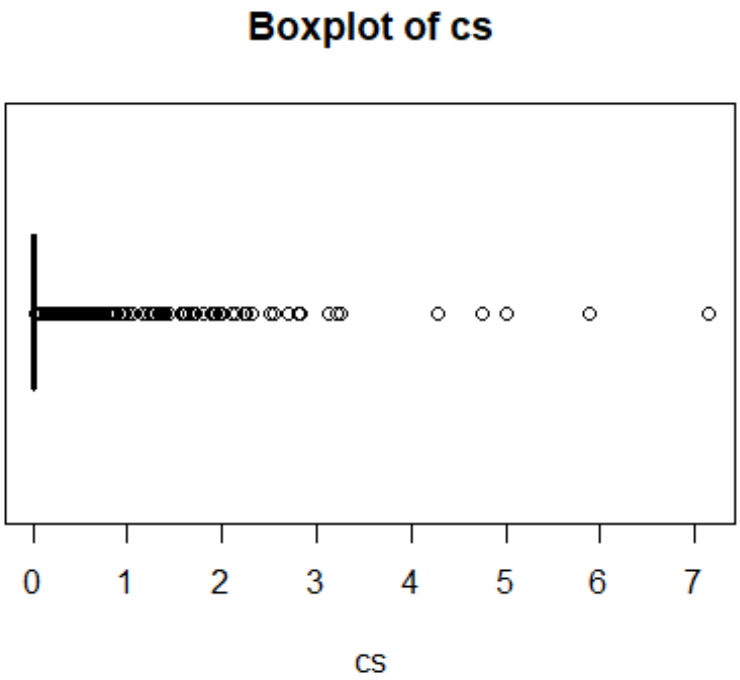

Boxplot of technology
0
2
4
6
technology

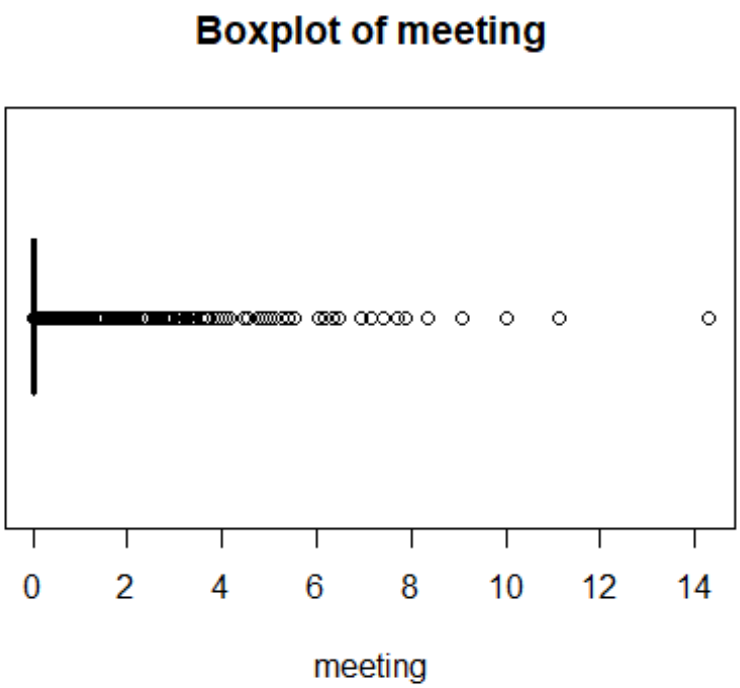

Boxplot of X1999
0
1
2
3
4
5
6
7
X1999

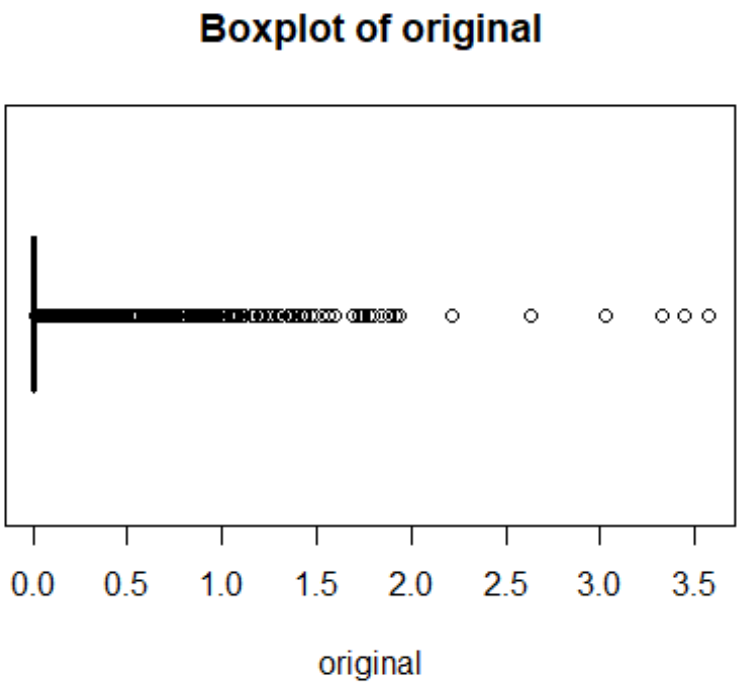

Boxplot of parts
0
2
4
6
8
parts

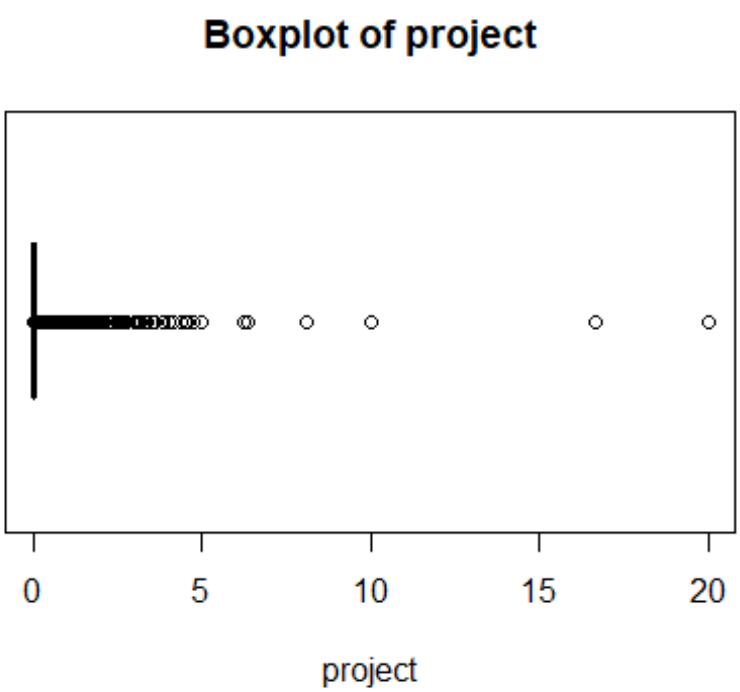

Boxplot of pm
0
2
4
6
8
10
pm

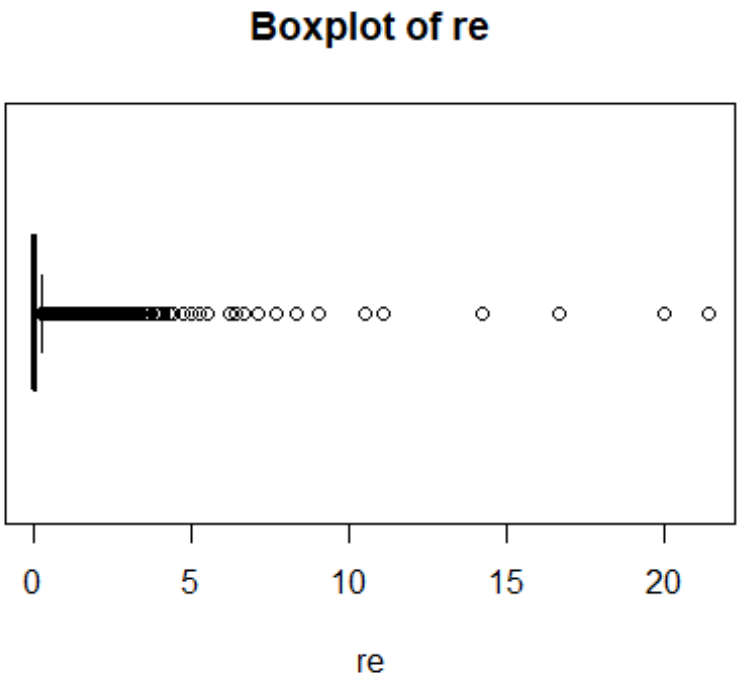

Boxplot of direct
0
1
2
3
4
direct

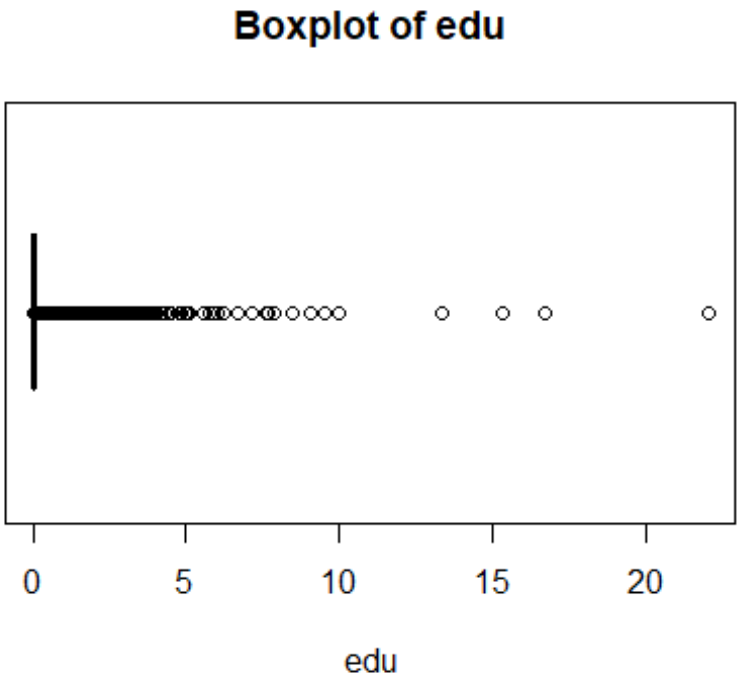

Boxplot of cs
0
1
2
3
4
5
6
7
cs

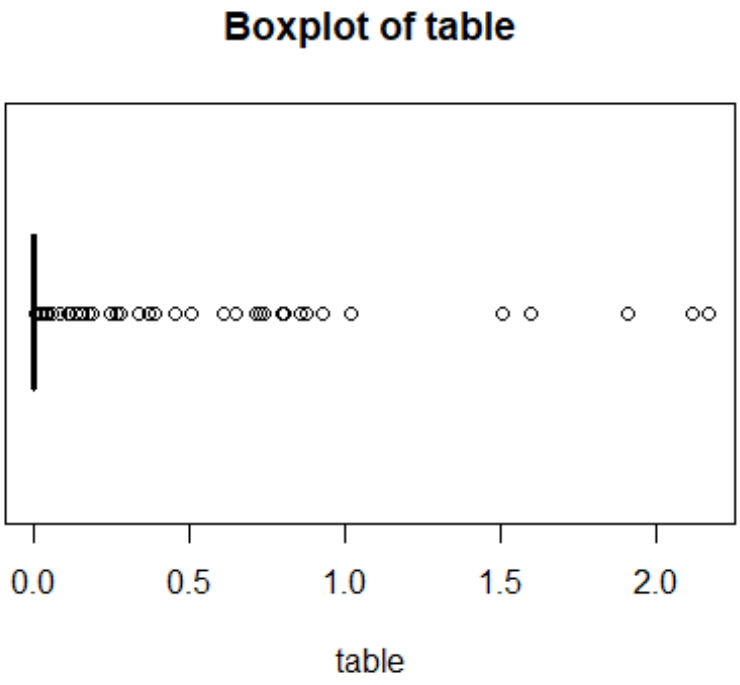

Boxplot of meeting
0
2
4
6
8
10
12
14
meeting

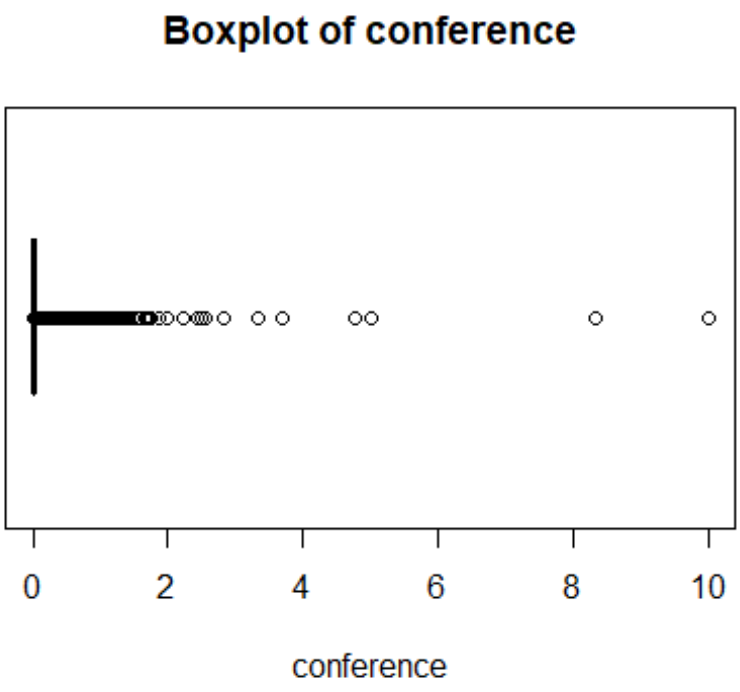

Boxplot of original
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
original

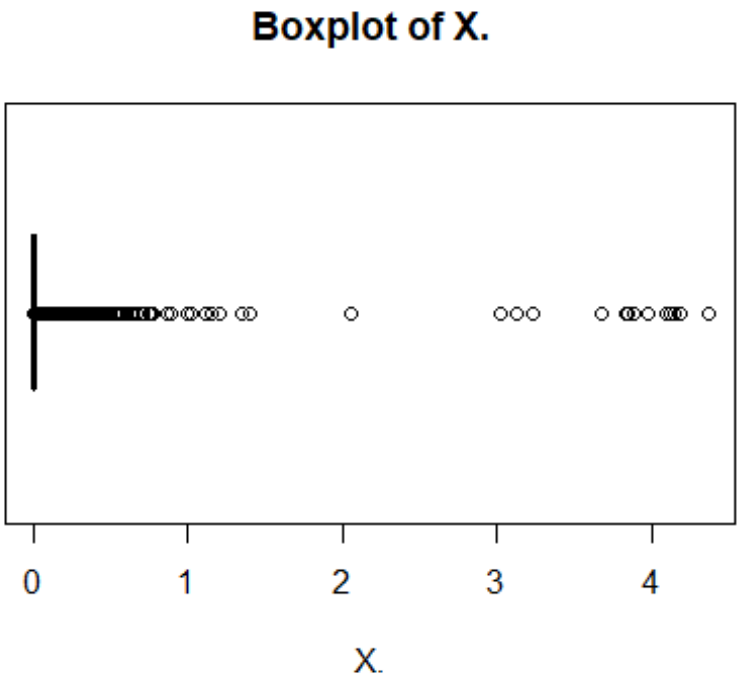

Boxplot of project
0
5
10
15
20
project

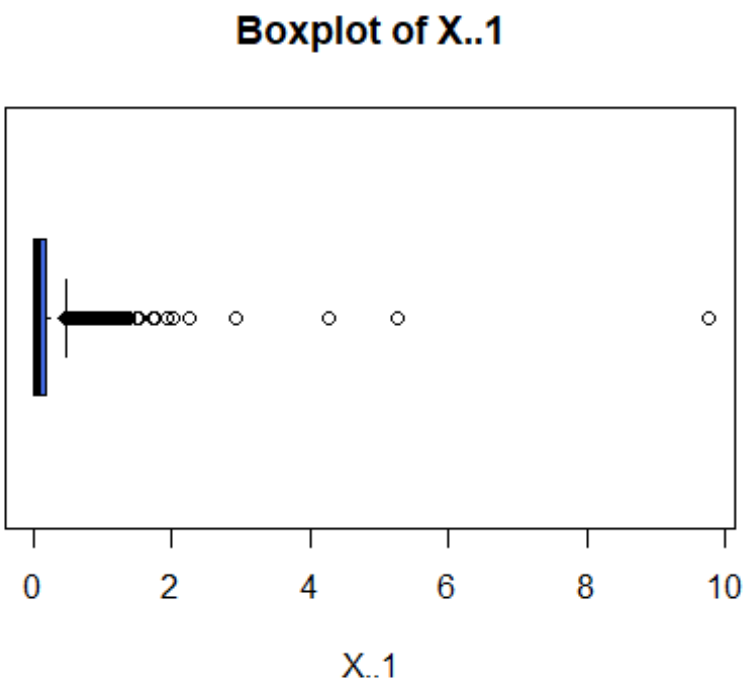

Boxplot of re
0
5
10
15
20
re

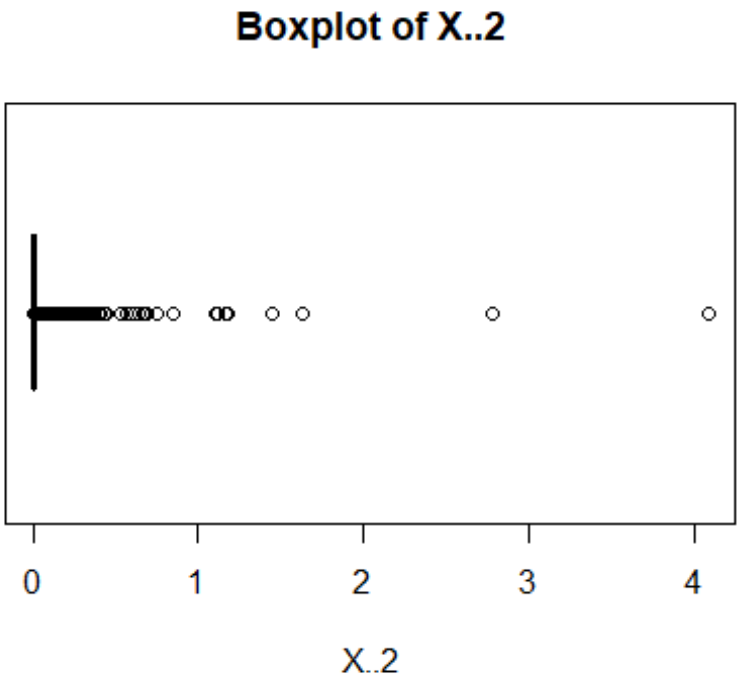

Boxplot of edu
0
5
10
15
20
edu

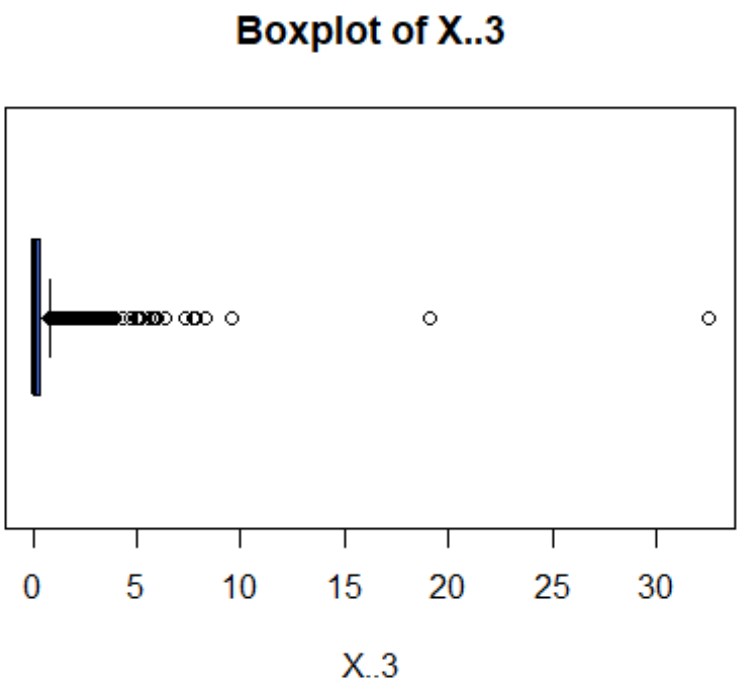

Boxplot of table
0.0
0.5
1.0
1.5
2.0
table

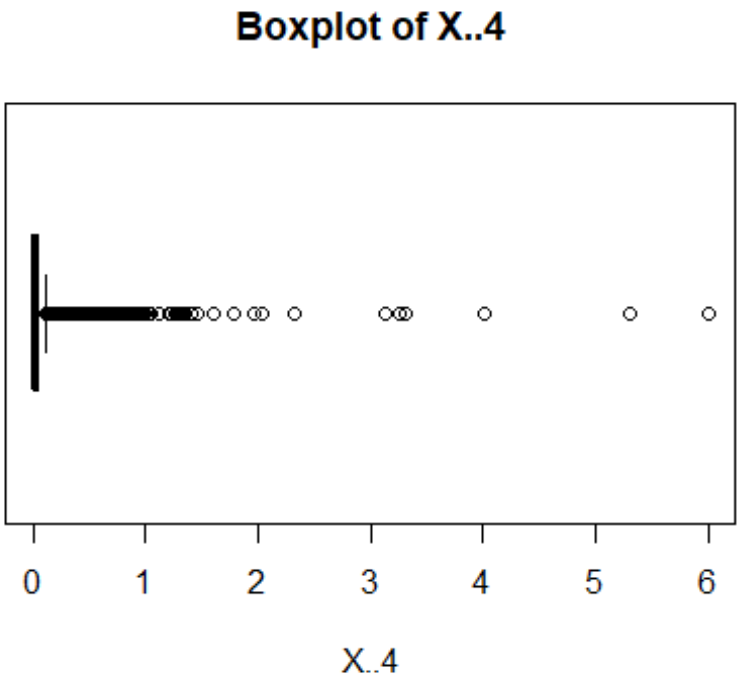

Boxplot of conference
0
2
4
6
8
10
conference

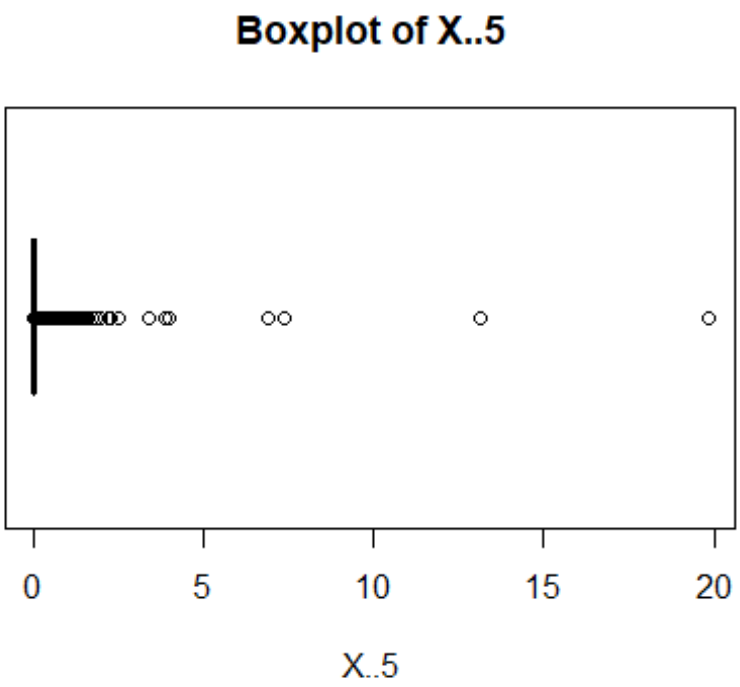

Boxplot of X.
0
1
2
3
4
X.

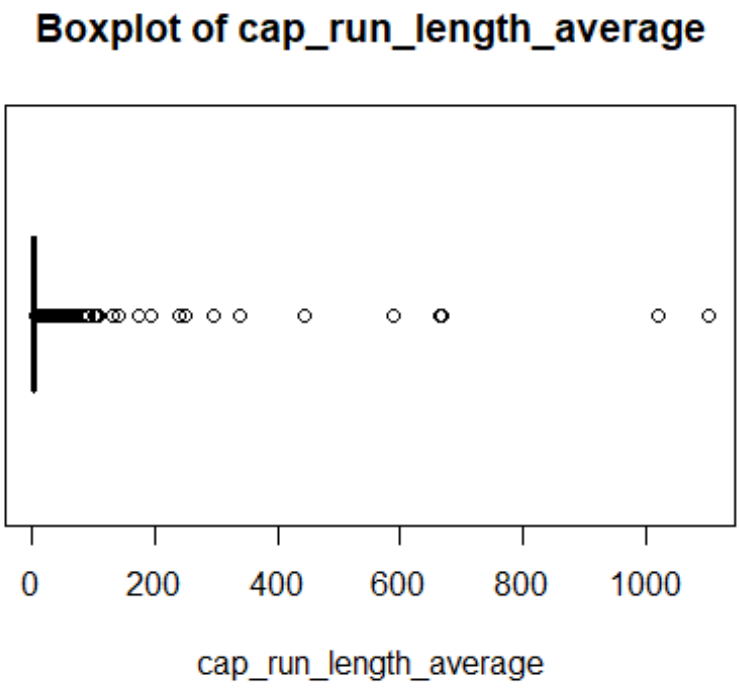

Boxplot of X..1
0
2
4
6
8
10
X..1

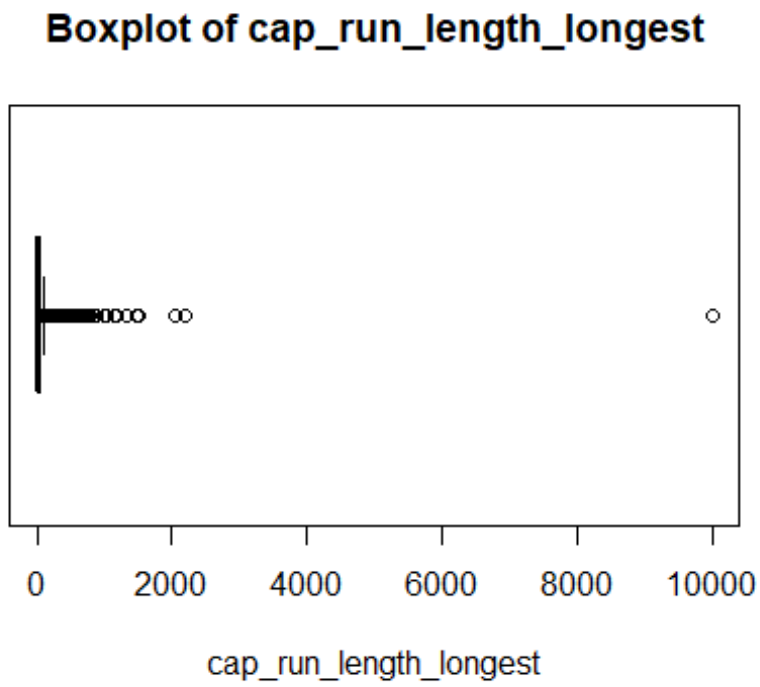

Boxplot of X..2
0
1
2
3
4
X..2

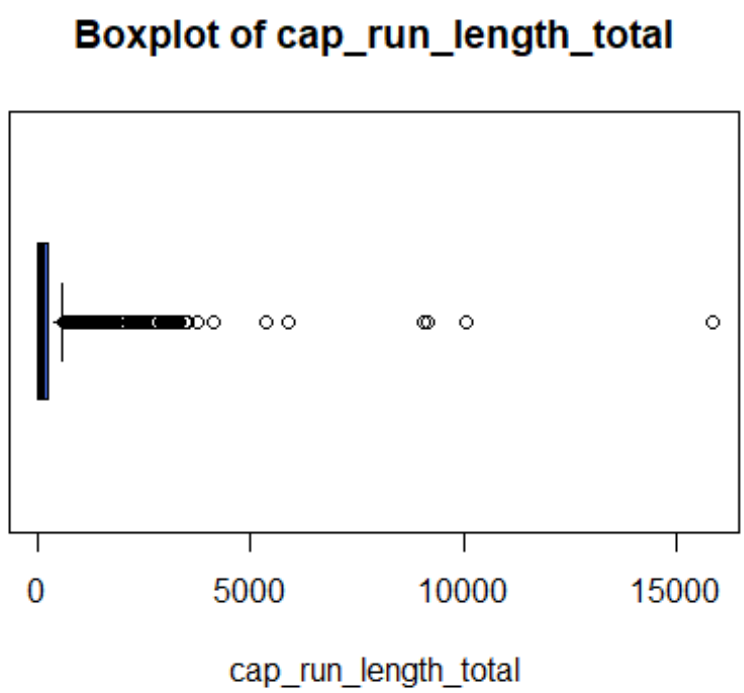

Boxplot of X..3
0
5
10
15
20
25
30
X..3


Boxplot of X..4
0
1
2
3
4
5
6
X..4


Boxplot of X..5
0
5
10
15
20
X..5


Boxplot of cap_run_length_average
0
200
400
600
800
1000
cap_run_length_average


Boxplot of cap_run_length_longest
0
2000
4000
6000
8000
10000
cap_run_length_longest


Boxplot of cap_run_length_total
0
5000
10000
15000
cap_run_length_total

El **apartado tercero** señala que debemos eliminar las palabras que tengan una correlación elevada con otras, ver la frecuencia con la que aparecen y las de menor aparición eliminarlas.

Para cumplir con el objetivo de este punto, realizamos una serie de modificaciones previas consistentes en cambiar la variable *clase* a factor y convertir la variable *capital_run_length* en entero.

```
cols_to_int <- c( 'cap_run_length_average')
cols_to_factor <- c('clase')

spam <- spam %>%
  mutate_at(cols_to_int, as.integer) %>%
  mutate_at(cols_to_factor, factor)
```

Para revisar que los cambios se han realizado correctamente:

```
glimpse(spam)

## Rows: 4,601
## Columns: 58
## $ make                    <dbl> 0.00, 0.21, 0.06, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ address                 <dbl> 0.64, 0.28, 0.00, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ all                     <dbl> 0.64, 0.50, 0.71, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ X3d                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ our                     <dbl> 0.32, 0.14, 1.23, 0.63, 0.63, 1.85, 1.92,
1.88,…
## $ over                    <dbl> 0.00, 0.28, 0.19, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ remove                  <dbl> 0.00, 0.21, 0.19, 0.31, 0.31, 0.00, 0.00,
0.00,…
## $ internet                <dbl> 0.00, 0.07, 0.12, 0.63, 0.63, 1.85, 0.00,
1.88,…
## $ order                   <dbl> 0.00, 0.00, 0.64, 0.31, 0.31, 0.00, 0.00,
0.00,…
## $ mail                    <dbl> 0.00, 0.94, 0.25, 0.63, 0.63, 0.00, 0.64,
0.00,…
## $ receive                 <dbl> 0.00, 0.21, 0.38, 0.31, 0.31, 0.00, 0.96,
0.00,…
## $ will                    <dbl> 0.64, 0.79, 0.45, 0.31, 0.31, 0.00, 1.28,
0.00,…
## $ people                  <dbl> 0.00, 0.65, 0.12, 0.31, 0.31, 0.00, 0.00,
0.00,…
## $ report                  <dbl> 0.00, 0.21, 0.70, 0.00, 0.00, 0.00, 0.00,
0.00,…
```

```
## $ addresses               <dbl> 0.00, 0.14, 1.75, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ free                    <dbl> 0.32, 0.14, 0.06, 0.31, 0.31, 0.00, 0.96,
0.00,…
## $ business                <dbl> 0.00, 0.07, 0.06, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ email                   <dbl> 1.29, 0.28, 1.03, 0.00, 0.00, 0.00, 0.32,
0.00,…
## $ you                     <dbl> 1.93, 3.47, 1.36, 3.18, 3.18, 0.00, 3.85,
0.00,…
## $ credit                  <dbl> 0.00, 0.00, 0.32, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ your                    <dbl> 0.96, 1.59, 0.51, 0.31, 0.31, 0.00, 0.64,
0.00,…
## $ font                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ X000                    <dbl> 0.00, 0.43, 1.16, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ money                   <dbl> 0.00, 0.43, 0.06, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ hp                      <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 1.99, 0.00,
0.00,…
## $ hpl                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ george                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ X650                    <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ lab                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ labs                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ telnet                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ X857                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ data                    <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ X415                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ X85                     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ technology              <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ X1999                   <dbl> 0.00, 0.07, 0.00, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ parts                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ pm                      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
```

```
## $ direct                  <dbl> 0.00, 0.00, 0.06, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ cs                      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ meeting                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ original                <dbl> 0.00, 0.00, 0.12, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ project                 <dbl> 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ re                      <dbl> 0.00, 0.00, 0.06, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ edu                     <dbl> 0.00, 0.00, 0.06, 0.00, 0.00, 0.00, 0.00,
0.00,…
## $ table                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ conference              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0,…
## $ X.                      <dbl> 0.000, 0.000, 0.010, 0.000, 0.000, 0.000,
0.000…
## $ X..1                    <dbl> 0.000, 0.132, 0.143, 0.137, 0.135, 0.223,
0.054…
## $ X..2                    <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 0.000,
0.000…
## $ X..3                    <dbl> 0.778, 0.372, 0.276, 0.137, 0.135, 0.000,
0.164…
## $ X..4                    <dbl> 0.000, 0.180, 0.184, 0.000, 0.000, 0.000,
0.054…
## $ X..5                    <dbl> 0.000, 0.048, 0.010, 0.000, 0.000, 0.000,
0.000…
## $ cap_run_length_average <int> 3, 5, 9, 3, 3, 3, 1, 2, 9, 1, 1, 1, 3, 2,
1, 5,…
## $ cap_run_length_longest <int> 61, 101, 485, 40, 40, 15, 4, 11, 445, 43,
6, 11…
## $ cap_run_length_total   <int> 278, 1028, 2259, 191, 418, 54, 112, 49,
1257, 7…
## $ clase                   <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1,…
```

Correlación existente entre palabras:

```
cor(spam[, 1:54])
```

```
##                 make  address      all       X3d      our     over   remove
## make         1.00000 -0.01659  0.06007  0.013642  0.02499  0.06390  0.01043
## address     -0.01659  1.00000 -0.03360 -0.006920 -0.01928 -0.02532  0.00334
## all          0.06007 -0.03360  1.00000 -0.020126  0.08199  0.08261  0.03596
## X3d          0.01364 -0.00692 -0.02013  1.000000  0.00336 -0.00995  0.01966
## our          0.02499 -0.01928  0.08199  0.003358  1.00000  0.06224  0.14910
## over         0.06390 -0.02532  0.08261 -0.009947  0.06224  1.00000  0.05736
## remove       0.01043  0.00334  0.03596  0.019663  0.14910  0.05736  1.00000
## internet    -0.00489 -0.01131  0.01230  0.010358  0.02760  0.08014  0.03875
## order        0.10873 -0.00463  0.08943 -0.002500  0.02250  0.11300  0.05621
## mail         0.04131  0.03394  0.03239 -0.004945  0.03628  0.01461  0.05435
## receive      0.19651 -0.00672  0.04945 -0.012909  0.06924  0.05569  0.15125
## will         0.09534 -0.04193  0.07740 -0.019205  0.07085  0.00509 -0.00368
## people       0.06502 -0.01944  0.05025 -0.013184  0.04242  0.07893  0.01391
## report       0.03775 -0.00223  0.01904  0.002287  0.00346  0.00966 -0.01785
## addresses    0.02867  0.00491  0.11745  0.002591  0.05586  0.16693  0.04002
## free         0.06714 -0.00959  0.06219  0.007239  0.08494  0.01841  0.14458
## business     0.08632 -0.01911  0.03580  0.003485  0.14377  0.06801  0.19951
## email        0.05686  0.03393  0.11041  0.019530  0.06508  0.07843  0.11871
## you          0.13074 -0.05624  0.14091 -0.010884  0.09466  0.09856  0.11052
## credit       0.01854 -0.01347  0.03539 -0.005220  0.02960  0.05577  0.04925
## your         0.20551 -0.01688  0.15454  0.008200  0.14777  0.10475  0.12788
## font        -0.02423 -0.00885 -0.03529  0.028159 -0.01968  0.00808 -0.00198
## X000         0.13662 -0.02090  0.11952  0.011775  0.06961  0.21735  0.07020
## money        0.18934  0.00210  0.04109  0.027531  0.00251  0.06009  0.02759
## hp          -0.07230 -0.04361 -0.08699 -0.015195 -0.07240 -0.08483 -0.08982
## hpl         -0.05956 -0.03829 -0.06348 -0.013692 -0.07745 -0.08512 -0.08050
## george      -0.06782 -0.02963 -0.11028 -0.010756 -0.09011 -0.06973 -0.06597
## X650        -0.04793 -0.02887 -0.04948 -0.010286 -0.06149 -0.06571 -0.06498
## lab         -0.04218 -0.02191 -0.05792 -0.007747  0.03063 -0.04919 -0.04868
## labs        -0.05481 -0.02785 -0.03255 -0.010547 -0.05598 -0.04988 -0.05889
## telnet      -0.03681 -0.01762 -0.03217 -0.007489 -0.04375 -0.04699 -0.04493
## X857        -0.03165 -0.00322 -0.06157 -0.006697 -0.02555 -0.03618 -0.04058
## data        -0.03193 -0.02499 -0.05361 -0.008097 -0.03325 -0.03311 -0.04177
## X415        -0.02844 -0.00430 -0.06221 -0.006723 -0.02607 -0.03776 -0.04105
## X85         -0.04571 -0.02432 -0.04789 -0.006180 -0.05080 -0.05412 -0.05277
## technology  -0.05500 -0.02820 -0.04360 -0.006532 -0.05125 -0.05228 -0.05337
## X1999       -0.05856 -0.02278 -0.06886 -0.007580 -0.07608 -0.05954 -0.04138
## parts       -0.00687 -0.00893  0.03224 -0.002665  0.13176 -0.01779 -0.01483
## pm          -0.01072 -0.01875 -0.01224 -0.004474 -0.04006 -0.04866 -0.04590
## direct      -0.03661 -0.01487 -0.04785 -0.007640 -0.02174 -0.03047 -0.02276
## cs          -0.00828 -0.01510 -0.03140 -0.005678 -0.04776 -0.02927 -0.03330
## meeting     -0.02524 -0.02512 -0.00657 -0.008074  0.11198 -0.05416 -0.04731
## original    -0.02062 -0.00259 -0.04822 -0.009222 -0.05020 -0.02977 -0.04693
## project     -0.02303 -0.02038 -0.03877 -0.006053  0.02073 -0.02947 -0.03540
## re          -0.03587 -0.01486 -0.04856 -0.012810 -0.04294 -0.05311 -0.05043
```

```
## edu        -0.03410 -0.02506 -0.05211 -0.009137 -0.07714 -0.03281 -0.05332
## table      -0.00101 -0.00866  0.02686 -0.003424 -0.02730 -0.01532 -0.01836
## conference -0.01936 -0.01590 -0.02533 -0.001959 -0.03270 -0.03153 -0.03174
## X.         -0.02656 -0.00827 -0.03510 -0.001174 -0.03250 -0.01832 -0.03288
## X..1       -0.02229 -0.04956 -0.01628 -0.012274 -0.04765 -0.00865 -0.05356
## X..2       -0.03303 -0.01855 -0.03353 -0.007194 -0.02714 -0.01683 -0.02664
## X..3        0.05863 -0.01534  0.10627 -0.002882  0.03131  0.06535  0.04994
## X..4        0.11816 -0.00963  0.08607  0.011197  0.04485  0.10744  0.06217
## X..5       -0.00863  0.00200 -0.00144 -0.000309  0.00145  0.02011  0.04666
##             internet     order     mail  receive      will   people
report
## make       -0.004892  0.108732  0.04131  0.19651  0.095340  0.06502
0.03775
## address    -0.011313 -0.004634  0.03394 -0.00672 -0.041926 -0.01944
-0.00223
## all         0.012297  0.089431  0.03239  0.04945  0.077404  0.05025
0.01904
## X3d         0.010358 -0.002500 -0.00495 -0.01291 -0.019205 -0.01318
0.00229
## our         0.027605  0.022498  0.03628  0.06924  0.070847  0.04242
0.00346
## over        0.080138  0.112998  0.01461  0.05569  0.005089  0.07893
0.00966
## remove      0.038754  0.056206  0.05435  0.15125 -0.003679  0.01391
-0.01785
## internet    1.000000  0.102891  0.08579  0.12411 -0.003589  0.02381
0.01299
## order       0.102891  1.000000  0.13201  0.13206  0.029898  0.03214
0.06140
## mail        0.085792  0.132014  1.00000  0.11988  0.083164  0.04257
0.01648
## receive     0.124111  0.132059  0.11988  1.00000  0.119363  0.04793
0.04706
## will       -0.003589  0.029898  0.08316  0.11936  1.000000  0.00270
0.00580
## people      0.023811  0.032142  0.04257  0.04793  0.002701  1.00000
0.06830
## report      0.012994  0.061402  0.01648  0.04706  0.005797  0.06830
1.00000
## addresses   0.072240  0.231336  0.20704  0.06462  0.023955  0.07682
-0.01600
## free        0.049257  0.005805  0.02603  0.09667 -0.027056  0.00439
0.00328
## business    0.218452  0.162180  0.08438  0.17240  0.066436  0.05084
0.01528
## email       0.030704  0.098781  0.03914  0.08890  0.011428  0.07203
-0.02885
## you         0.022434  0.037978  0.08845  0.14701  0.085806  0.11718
0.01364
## credit      0.106604  0.123404  0.03479  0.14661  0.011287 -0.01697
```

```
0.03253
## your        0.152422  0.153706  0.09713  0.29818  0.103979  0.05647
0.05061
## font       -0.019103 -0.022467  0.01110 -0.01100 -0.046024 -0.02999
-0.01967
## X000        0.086430  0.122141  0.09310  0.10460  0.012519  0.11691
0.04498
## money       0.033476  0.101961  0.05144  0.05619  0.017180  0.08049
0.04089
## hp         -0.050650 -0.070223 -0.03585 -0.07659 -0.024421 -0.05833
-0.03867
## hpl        -0.038978 -0.047806 -0.01353 -0.07673  0.011892 -0.06730
-0.04412
## george     -0.056292 -0.064691 -0.06881 -0.06505 -0.120687 -0.05514
-0.02943
## X650       -0.050887 -0.054301  0.02032 -0.06004 -0.037907 -0.05999
-0.02195
## lab        -0.036524 -0.045680 -0.02689 -0.04712  0.118500 -0.02729
-0.01432
## labs       -0.042894 -0.045834  0.00242 -0.05537 -0.014966 -0.05246
-0.03219
## telnet     -0.035639 -0.039980 -0.02334 -0.03151 -0.038149 -0.03609
-0.02095
## X857       -0.033859 -0.033400 -0.01512 -0.03895 -0.055381 -0.03335
-0.02206
## data       -0.037144 -0.015479 -0.03198 -0.04150 -0.017409 -0.04515
-0.01251
## X415       -0.034744 -0.031646 -0.01420 -0.03869 -0.056598 -0.03164
-0.02235
## X85        -0.031899 -0.038822 -0.01972 -0.04843 -0.049139 -0.04908
-0.02729
## technology -0.032775 -0.055115 -0.01665 -0.05244 -0.013858 -0.04612
0.00541
## X1999      -0.014333 -0.033387 -0.00207 -0.03369 -0.029871 -0.04499
-0.03051
## parts      -0.012059 -0.002476 -0.01787 -0.00472 -0.025234 -0.01002
-0.00189
## pm         -0.028973 -0.040109 -0.01424 -0.04336  0.016851 -0.01631
-0.02656
## direct     -0.005543 -0.010254  0.00495 -0.02680 -0.037190 -0.01329
-0.02820
## cs         -0.003664 -0.035309 -0.02449 -0.03455 -0.022056 -0.01936
-0.02131
## meeting    -0.043482 -0.048773 -0.05389 -0.04270  0.119774 -0.03804
0.00487
## original   -0.000622 -0.035678  0.02320 -0.03971 -0.023725 -0.02131
-0.00867
## project    -0.030573 -0.035572 -0.02982 -0.03716  0.020094 -0.02501
-0.01339
## re         -0.002283 -0.074810 -0.03549 -0.06556 -0.088980 -0.04214
```

```
0.00046
## edu        -0.037672 -0.056935 -0.03145 -0.05042 -0.070711 -0.02167
-0.02072
## table      -0.007304  0.001649 -0.01383 -0.01970  0.000301 -0.01387
0.04584
## conference -0.021339 -0.026100 -0.01700 -0.02273  0.032606 -0.02031
-0.01654
## X.         -0.027923 -0.014911  0.01749 -0.03264 -0.027934 -0.02289
-0.01944
## X..1       -0.035222 -0.029811  0.00294 -0.05734 -0.027382 -0.05115
-0.00915
## X..2       -0.018751  0.014616  0.00691 -0.02458 -0.044731 -0.02760
-0.01338
## X..3        0.031852  0.042780  0.03612  0.02273  0.011695  0.04085
-0.00768
## X..4        0.054516  0.150734  0.06924  0.07023  0.013341  0.20454
0.08983
## X..5       -0.006532 -0.000741  0.04573  0.00140 -0.030785 -0.01432
0.00630
##            addresses      free  business    email       you   credit
your
## make        0.028672  0.067140  0.086322  0.05686  0.130742  0.01854
0.20551
## address     0.004908 -0.009591 -0.019112  0.03393 -0.056237 -0.01347
-0.01688
## all         0.117451  0.062193  0.035800  0.11041  0.140908  0.03539
0.15454
## X3d         0.002591  0.007239  0.003485  0.01953 -0.010884 -0.00522
0.00820
## our         0.055863  0.084939  0.143774  0.06508  0.094660  0.02960
0.14777
## over        0.166933  0.018408  0.068006  0.07843  0.098562  0.05577
0.10475
## remove      0.040023  0.144580  0.199513  0.11871  0.110524  0.04925
0.12788
## internet    0.072240  0.049257  0.218452  0.03070  0.022434  0.10660
0.15242
## order       0.231336  0.005805  0.162180  0.09878  0.037978  0.12340
0.15371
## mail        0.207044  0.026030  0.084383  0.03914  0.088449  0.03479
0.09713
## receive     0.064617  0.096674  0.172396  0.08890  0.147012  0.14661
0.29818
## will        0.023955 -0.027056  0.066436  0.01143  0.085806  0.01129
0.10398
## people      0.076821  0.004385  0.050836  0.07203  0.117175 -0.01697
0.05647
## report     -0.016000  0.003281  0.015279 -0.02885  0.013640  0.03253
0.05061
## addresses   1.000000 -0.000634  0.018639  0.25150  0.043228  0.01832
```

```
0.07181
## free       -0.000634  1.000000  0.048411  0.06383  0.085489  0.02762
0.09874
## business    0.018639  0.048411  1.000000  0.04987  0.085695  0.18708
0.20847
## email       0.251498  0.063830  0.049868  1.00000  0.091223  0.01124
0.13664
## you         0.043228  0.085489  0.085695  0.09122  1.000000  0.03361
0.30745
## credit      0.018316  0.027624  0.187076  0.01124  0.033607  1.00000
0.12158
## your        0.071806  0.098742  0.208465  0.13664  0.307453  0.12158
1.00000
## font       -0.004387 -0.008502 -0.020904 -0.02673 -0.023431  0.02901
-0.02029
## X000        0.367325  0.052901  0.097845  0.09620  0.112158  0.04034
0.12459
## money       0.033713  0.101782  0.045065  0.09348  0.181780  0.05970
0.16445
## hp         -0.049319 -0.089980 -0.054505 -0.03597 -0.198088 -0.04715
-0.15343
## hpl        -0.030306 -0.079232 -0.075840 -0.02648 -0.160997 -0.04842
-0.13278
## george     -0.041637 -0.003323 -0.069837 -0.06934 -0.158073 -0.03677
-0.12652
## X650       -0.033186 -0.063474 -0.063510  0.04076 -0.118919 -0.03692
-0.09410
## lab        -0.028788 -0.046483 -0.045259 -0.05046 -0.088071 -0.02713
-0.06404
## labs       -0.033890 -0.058798 -0.039665  0.02818 -0.108703 -0.03669
-0.09005
## telnet     -0.026464 -0.043011 -0.040189  0.03150 -0.087842 -0.02592
-0.06764
## X857       -0.020939 -0.039140 -0.040926 -0.03853 -0.076529 -0.02293
-0.04769
## data       -0.015579 -0.044048 -0.040157 -0.02666 -0.095411 -0.02738
-0.08069
## X415       -0.023638 -0.039411 -0.042802 -0.04180 -0.075967 -0.02370
-0.05075
## X85        -0.021964 -0.056292 -0.056457  0.02291 -0.092092 -0.03095
-0.08160
## technology -0.034563 -0.060008 -0.035748  0.03028 -0.136223 -0.03712
-0.10441
## X1999      -0.035888 -0.060493 -0.048470 -0.03470 -0.134498 -0.03389
-0.11319
## parts      -0.011361  0.017566 -0.015506  0.00654 -0.032606  0.00989
-0.00700
## pm         -0.016171 -0.034055 -0.042688 -0.02382 -0.038104 -0.02753
-0.06120
## direct      0.046645 -0.029958  0.008928  0.00796 -0.060732 -0.00500
```

```
-0.01677
## cs         -0.013423 -0.023174 -0.036350 -0.02971 -0.049359 -0.01975
-0.05855
## meeting    -0.030742 -0.042440 -0.040692 -0.04585 -0.087335 -0.02810
-0.07905
## original    0.047055 -0.044064 -0.050136 -0.01680 -0.050955 -0.02972
-0.04905
## project    -0.018321 -0.032580 -0.021401 -0.03423 -0.068312 -0.01359
-0.06281
## re         -0.039109 -0.045388 -0.058429 -0.04789  0.113371 -0.04246
-0.03354
## edu        -0.025447 -0.045340 -0.057741 -0.04025 -0.000953 -0.03011
-0.07779
## table      -0.011878 -0.018625 -0.011410  0.01883 -0.003724 -0.00776
0.00947
## conference -0.021064 -0.028389 -0.030238 -0.01486 -0.040016 -0.01795
-0.04976
## X.         -0.018543 -0.026267 -0.030590 -0.04028 -0.042528 -0.01981
-0.05864
## X..1       -0.002237 -0.046647 -0.036387 -0.03207 -0.128232 -0.01860
-0.08422
## X..2       -0.002366 -0.030073 -0.036148 -0.01708 -0.061791 -0.01231
-0.04345
## X..3        0.017092  0.103662  0.077016  0.03666  0.152823  0.04595
0.07724
## X..4        0.117509  0.049052  0.101244  0.06433  0.091416  0.03447
0.13485
## X..5       -0.005635  0.035092 -0.000728  0.02140 -0.001578  0.00771
-0.00364
##                font      X000     money        hp      hpl   george
X650
## make       -0.02423  0.136618  0.189340 -0.072302 -0.05956 -0.06782
-0.047934
## address    -0.00885 -0.020904  0.002095 -0.043613 -0.03829 -0.02963
-0.028874
## all        -0.03529  0.119519  0.041087 -0.086986 -0.06348 -0.11028
-0.049477
## X3d         0.02816  0.011775  0.027531 -0.015195 -0.01369 -0.01076
-0.010286
## our        -0.01968  0.069610  0.002509 -0.072395 -0.07745 -0.09011
-0.061488
## over        0.00808  0.217346  0.060086 -0.084833 -0.08512 -0.06973
-0.065710
## remove     -0.00198  0.070198  0.027588 -0.089816 -0.08050 -0.06597
-0.064985
## internet   -0.01910  0.086430  0.033476 -0.050650 -0.03898 -0.05629
-0.050887
## order      -0.02247  0.122141  0.101961 -0.070223 -0.04781 -0.06469
-0.054301
## mail        0.01110  0.093103  0.051442 -0.035850 -0.01353 -0.06881
```

```
0.020322
## receive    -0.01100  0.104600  0.056193 -0.076590 -0.07673 -0.06505
-0.060045
## will       -0.04602  0.012519  0.017180 -0.024421  0.01189 -0.12069
-0.037907
## people     -0.02999  0.116912  0.080491 -0.058334 -0.06730 -0.05514
-0.059991
## report     -0.01967  0.044981  0.040886 -0.038674 -0.04412 -0.02943
-0.021952
## addresses  -0.00439  0.367325  0.033713 -0.049319 -0.03031 -0.04164
-0.033186
## free       -0.00850  0.052901  0.101782 -0.089980 -0.07923 -0.00332
-0.063474
## business   -0.02090  0.097845  0.045065 -0.054505 -0.07584 -0.06984
-0.063510
## email      -0.02673  0.096204  0.093485 -0.035966 -0.02648 -0.06934
0.040758
## you        -0.02343  0.112158  0.181780 -0.198088 -0.16100 -0.15807
-0.118919
## credit      0.02901  0.040342  0.059695 -0.047149 -0.04842 -0.03677
-0.036920
## your       -0.02029  0.124588  0.164452 -0.153432 -0.13278 -0.12652
-0.094102
## font        1.00000  0.022551 -0.011894 -0.038442 -0.03480 -0.02698
-0.027057
## X000        0.02255  1.000000  0.050429 -0.086648 -0.08167 -0.06603
-0.066676
## money      -0.01189  0.050429  1.000000 -0.066634 -0.06141 -0.04767
-0.046777
## hp         -0.03844 -0.086648 -0.066634  1.000000  0.50857 -0.01102
0.337025
## hpl        -0.03480 -0.081670 -0.061412  0.508572  1.00000  0.00879
0.376836
## george     -0.02698 -0.066027 -0.047669 -0.011024  0.00879  1.00000
0.031707
## X650       -0.02706 -0.066676 -0.046777  0.337025  0.37684  0.03171
1.000000
## lab        -0.01970 -0.047258 -0.035038  0.216671  0.21871  0.02867
0.340282
## labs       -0.02668 -0.063113 -0.046848  0.435298  0.39782  0.04221
0.563335
## telnet     -0.01878 -0.046096 -0.031616  0.347421  0.34126  0.05107
0.521258
## X857       -0.01679 -0.041311 -0.028242  0.360868  0.35339  0.07908
0.558345
## data       -0.01927 -0.047867 -0.035425  0.000399  0.00717 -0.02234
0.010279
## X415       -0.01705 -0.042055 -0.029591  0.363810  0.35991  0.07715
0.566679
## X85        -0.02333 -0.050716 -0.038084  0.313733  0.32770  0.03104
```

```
0.578761
## technology -0.02698 -0.060146 -0.045271  0.387785  0.35122  0.03691
0.566193
## X1999      -0.03252 -0.070530 -0.055240  0.124059  0.15606 -0.01712
0.018894
## parts      -0.00490 -0.017043 -0.007876  0.008132  0.02167 -0.01031
-0.012843
## pm         -0.01796 -0.037673 -0.032983  0.048518  0.05949  0.00178
0.037051
## direct     -0.02150 -0.004115 -0.022111  0.326581  0.32135  0.05765
0.512885
## cs         -0.01420 -0.031338 -0.025700  0.001990  0.01837 -0.01773
0.021216
## meeting    -0.02023 -0.050028 -0.033989  0.019556  0.04692 -0.00735
0.004751
## original   -0.02131 -0.040093 -0.037067  0.107093  0.13177  0.00876
0.076317
## project    -0.01220 -0.035363 -0.024075 -0.002204  0.01417 -0.01255
-0.000683
## re         -0.03061 -0.053903 -0.045531  0.048808 -0.00779 -0.00829
-0.006670
## edu        -0.01964 -0.053022 -0.032070 -0.045508 -0.03600 -0.03826
-0.008321
## table       0.02665 -0.018025 -0.011132  0.001083  0.03241 -0.01092
0.002878
## conference -0.01318 -0.030575 -0.015888 -0.003065  0.03225  0.00615
0.003535
## X.          0.40042 -0.027564 -0.018731  0.022590  0.01207 -0.02023
-0.025257
## X..1       -0.04518 -0.037888 -0.032199  0.134436  0.14247 -0.02995
0.316406
## X..2       -0.00146  0.000471 -0.020148  0.038424  0.06547 -0.01969
0.032667
## X..3       -0.00466  0.072444  0.051143 -0.090146 -0.07795 -0.06639
-0.063243
## X..4       -0.01057  0.309988  0.103619 -0.085560 -0.07995 -0.06843
-0.060547
## X..5        0.18851  0.020064  0.000811  0.059285 -0.02062 -0.02057
-0.011335
##                 lab     labs   telnet      X857      data      X415
X85
## make       -0.04218 -0.05481 -0.03681 -0.031654 -0.031930 -0.028438
-0.04571
## address    -0.02191 -0.02785 -0.01762 -0.003219 -0.024991 -0.004305
-0.02432
## all        -0.05792 -0.03255 -0.03217 -0.061568 -0.053612 -0.062208
-0.04789
## X3d        -0.00775 -0.01055 -0.00749 -0.006697 -0.008097 -0.006723
-0.00618
## our         0.03063 -0.05598 -0.04375 -0.025549 -0.033255 -0.026074
```

```
-0.05080
## over       -0.04919 -0.04988 -0.04699 -0.036175 -0.033107 -0.037758
-0.05412
## remove     -0.04868 -0.05889 -0.04493 -0.040579 -0.041770 -0.041052
-0.05277
## internet   -0.03652 -0.04289 -0.03564 -0.033859 -0.037144 -0.034744
-0.03190
## order      -0.04568 -0.04583 -0.03998 -0.033400 -0.015479 -0.031646
-0.03882
## mail       -0.02689  0.00242 -0.02334 -0.015118 -0.031984 -0.014195
-0.01972
## receive    -0.04712 -0.05537 -0.03151 -0.038950 -0.041498 -0.038693
-0.04843
## will        0.11850 -0.01497 -0.03815 -0.055381 -0.017409 -0.056598
-0.04914
## people     -0.02729 -0.05246 -0.03609 -0.033349 -0.045154 -0.031637
-0.04908
## report     -0.01432 -0.03219 -0.02095 -0.022061 -0.012510 -0.022350
-0.02729
## addresses  -0.02879 -0.03389 -0.02646 -0.020939 -0.015579 -0.023638
-0.02196
## free       -0.04648 -0.05880 -0.04301 -0.039140 -0.044048 -0.039411
-0.05629
## business   -0.04526 -0.03967 -0.04019 -0.040926 -0.040157 -0.042802
-0.05646
## email      -0.05046  0.02818  0.03150 -0.038529 -0.026663 -0.041805
0.02291
## you        -0.08807 -0.10870 -0.08784 -0.076529 -0.095411 -0.075967
-0.09209
## credit     -0.02713 -0.03669 -0.02592 -0.022932 -0.027376 -0.023699
-0.03095
## your       -0.06404 -0.09005 -0.06764 -0.047695 -0.080695 -0.050749
-0.08160
## font       -0.01970 -0.02668 -0.01878 -0.016790 -0.019274 -0.017050
-0.02333
## X000       -0.04726 -0.06311 -0.04610 -0.041311 -0.047867 -0.042055
-0.05072
## money      -0.03504 -0.04685 -0.03162 -0.028242 -0.035425 -0.029591
-0.03808
## hp          0.21667  0.43530  0.34742  0.360868  0.000399  0.363810
0.31373
## hpl         0.21871  0.39782  0.34126  0.353392  0.007173  0.359908
0.32770
## george      0.02867  0.04221  0.05107  0.079077 -0.022341  0.077152
0.03104
## X650        0.34028  0.56333  0.52126  0.558345  0.010279  0.566679
0.57876
## lab         1.00000  0.37876  0.41183  0.507947 -0.001629  0.506612
0.32555
## labs        0.37876  1.00000  0.60784  0.671341 -0.011684  0.681696
```

```
0.54285
## telnet      0.41183  0.60784  1.00000  0.734945 -0.013985  0.732029
0.52020
## X857        0.50795  0.67134  0.73495  1.000000 -0.016906  0.986706
0.55172
## data       -0.00163 -0.01168 -0.01399 -0.016906  1.000000 -0.017518
-0.00430
## X415        0.50661  0.68170  0.73203  0.986706 -0.017518  1.000000
0.55933
## X85         0.32555  0.54285  0.52020  0.551723 -0.004297  0.559334
1.00000
## technology  0.41030  0.62664  0.67720  0.726499 -0.005999  0.724857
0.51785
## X1999       0.01400  0.06023  0.03514  0.030817  0.048523  0.032428
0.02218
## parts      -0.00699 -0.00269 -0.00860 -0.007593  0.022899 -0.007741
-0.01120
## pm          0.00976  0.03823  0.03845  0.041319  0.067760  0.041718
0.02186
## direct      0.46999  0.60127  0.69970  0.850475 -0.019525  0.841996
0.50834
## cs         -0.01538  0.03826 -0.01280 -0.010118 -0.001167  0.002187
0.02331
## meeting     0.43157 -0.01653 -0.01033 -0.007724  0.010331 -0.008141
0.03142
## original    0.05917  0.07523  0.08820  0.113625  0.006800  0.113919
0.06271
## project    -0.00420 -0.00607  0.06368  0.003908  0.023340  0.005943
0.00101
## re         -0.01099 -0.00269 -0.00830  0.011338  0.000233  0.013131
-0.01556
## edu        -0.02319 -0.01198 -0.02265 -0.019207 -0.019220 -0.014708
-0.01468
## table      -0.00314 -0.00490 -0.00291  0.000185 -0.000709 -0.000105
0.01484
## conference -0.00361 -0.00489 -0.01371 -0.011788  0.004328 -0.010872
0.00178
## X.         -0.01948 -0.01990 -0.01672 -0.009574 -0.005079 -0.010298
-0.02247
## X..1        0.16052  0.23261  0.23504  0.305435  0.030135  0.305958
0.20228
## X..2        0.00629  0.00426  0.01107  0.012043  0.111004  0.013185
0.03630
## X..3       -0.04209 -0.06215 -0.04469 -0.041272 -0.046161 -0.039443
-0.04925
## X..4       -0.04992 -0.06514 -0.04705 -0.041975 -0.047808 -0.039923
-0.04841
## X..5        0.00202  0.05733 -0.00139 -0.011364 -0.010001 -0.010693
-0.00973
##            technology    X1999    parts       pm    direct       cs
```

```
meeting
## make         -0.05500 -0.05856 -0.00687 -0.01072 -0.036612 -0.00828
-0.025237
## address      -0.02820 -0.02278 -0.00893 -0.01875 -0.014869 -0.01510
-0.025121
## all          -0.04360 -0.06886  0.03224 -0.01224 -0.047850 -0.03140
-0.006569
## X3d          -0.00653 -0.00758 -0.00267 -0.00447 -0.007640 -0.00568
-0.008074
## our          -0.05125 -0.07608  0.13176 -0.04006 -0.021736 -0.04776
0.111979
## over         -0.05228 -0.05954 -0.01779 -0.04866 -0.030470 -0.02927
-0.054159
## remove       -0.05337 -0.04138 -0.01483 -0.04590 -0.022756 -0.03330
-0.047307
## internet     -0.03278 -0.01433 -0.01206 -0.02897 -0.005543 -0.00366
-0.043482
## order        -0.05512 -0.03339 -0.00248 -0.04011 -0.010254 -0.03531
-0.048773
## mail         -0.01665 -0.00207 -0.01787 -0.01424  0.004946 -0.02449
-0.053889
## receive      -0.05244 -0.03369 -0.00472 -0.04336 -0.026801 -0.03455
-0.042701
## will         -0.01386 -0.02987 -0.02523  0.01685 -0.037190 -0.02206
0.119774
## people       -0.04612 -0.04499 -0.01002 -0.01631 -0.013293 -0.01936
-0.038040
## report        0.00541 -0.03051 -0.00189 -0.02656 -0.028196 -0.02131
0.004871
## addresses    -0.03456 -0.03589 -0.01136 -0.01617  0.046645 -0.01342
-0.030742
## free         -0.06001 -0.06049  0.01757 -0.03406 -0.029958 -0.02317
-0.042440
## business     -0.03575 -0.04847 -0.01551 -0.04269  0.008928 -0.03635
-0.040692
## email         0.03028 -0.03470  0.00654 -0.02382  0.007962 -0.02971
-0.045852
## you          -0.13622 -0.13450 -0.03261 -0.03810 -0.060732 -0.04936
-0.087335
## credit       -0.03712 -0.03389  0.00989 -0.02753 -0.004999 -0.01975
-0.028099
## your         -0.10441 -0.11319 -0.00700 -0.06120 -0.016766 -0.05855
-0.079051
## font         -0.02698 -0.03252 -0.00490 -0.01796 -0.021498 -0.01420
-0.020235
## X000         -0.06015 -0.07053 -0.01704 -0.03767 -0.004115 -0.03134
-0.050028
## money        -0.04527 -0.05524 -0.00788 -0.03298 -0.022111 -0.02570
-0.033989
## hp            0.38779  0.12406  0.00813  0.04852  0.326581  0.00199
```

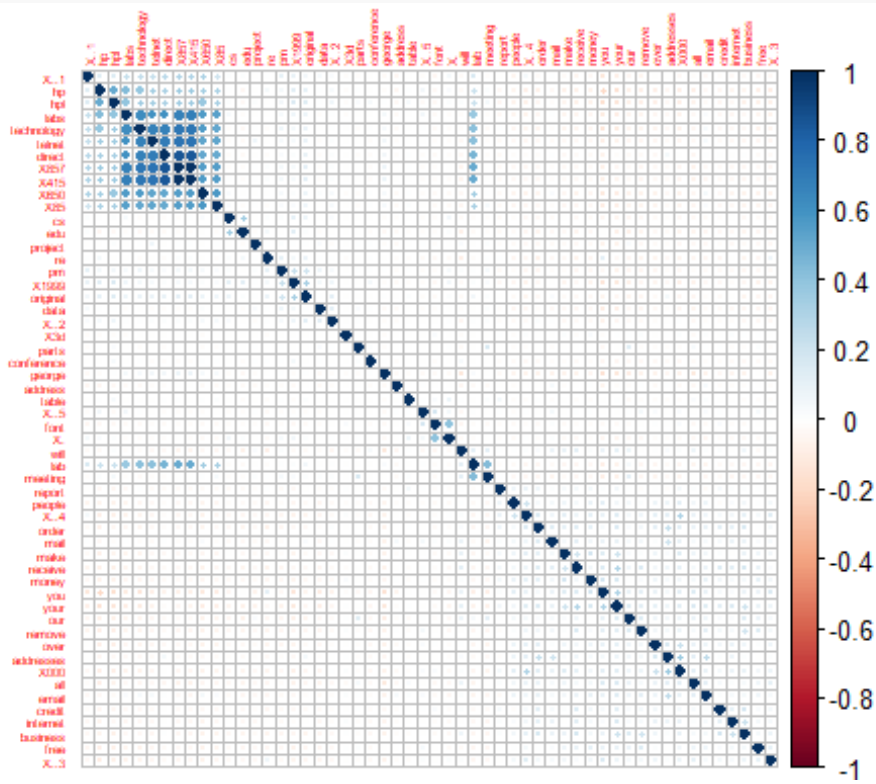
```
0.019556
## hpl           0.35122  0.15606  0.02167  0.05949  0.321353  0.01837
0.046916
## george        0.03691 -0.01712 -0.01031  0.00178  0.057651 -0.01773
-0.007350
## X650          0.56619  0.01889 -0.01284  0.03705  0.512885  0.02122
0.004751
## lab           0.41030  0.01400 -0.00699  0.00976  0.469990 -0.01538
0.431575
## labs          0.62664  0.06023 -0.00269  0.03823  0.601274  0.03826
-0.016528
## telnet        0.67720  0.03514 -0.00860  0.03845  0.699695 -0.01280
-0.010328
## X857          0.72650  0.03082 -0.00759  0.04132  0.850475 -0.01012
-0.007724
## data         -0.00600  0.04852  0.02290  0.06776 -0.019525 -0.00117
0.010331
## X415          0.72486  0.03243 -0.00774  0.04172  0.841996  0.00219
-0.008141
## X85           0.51785  0.02218 -0.01120  0.02186  0.508339  0.02331
0.031422
## technology    1.00000  0.04692 -0.01114  0.06995  0.675571 -0.01958
-0.022601
## X1999         0.04692  1.00000 -0.00321  0.22716  0.018682  0.10111
0.023364
## parts        -0.01114 -0.00321  1.00000  0.00263 -0.010175 -0.00564
0.168020
## pm            0.06995  0.22716  0.00263  1.00000  0.023306  0.01883
0.036993
## direct        0.67557  0.01868 -0.01017  0.02331  1.000000 -0.01560
-0.016717
## cs           -0.01958  0.10111 -0.00564  0.01883 -0.015599  1.00000
-0.012309
## meeting      -0.02260  0.02336  0.16802  0.03699 -0.016717 -0.01231
1.000000
## original      0.08319  0.32301 -0.00686  0.19361  0.097512  0.07858
0.017484
## project       0.00632  0.00347  0.00149  0.01387 -0.002271 -0.00780
0.014446
## re           -0.01130  0.04833 -0.00187  0.10355 -0.010240  0.01207
0.002506
## edu          -0.03446  0.12721 -0.00988 -0.00271 -0.027652  0.34661
-0.019544
## table        -0.00214 -0.01902 -0.00409  0.00363  0.000478 -0.00728
0.011813
## conference    0.00978  0.04850  0.01761  0.00977 -0.015779 -0.00136
-0.000562
## X.           -0.01965  0.05130  0.00908  0.03630 -0.019812  0.05584
-0.007803
## X..1          0.24703  0.09227 -0.01036  0.10078  0.270361  0.01387
```

```
-0.011581
## X..2          0.00126  0.06916  0.00171  0.03698  0.014240  0.02947
0.011000
## X..3         -0.05978 -0.05286 -0.01505 -0.02241 -0.029674 -0.02578
-0.036394
## X..4         -0.05722 -0.06250 -0.01267 -0.04296 -0.016871 -0.03622
-0.041838
## X..5          0.00897 -0.01690 -0.00360 -0.01072 -0.010607 -0.01156
-0.003657
##             original   project        re       edu     table conference
## make       -0.020621 -0.023035 -0.035873 -0.034105 -0.001011  -0.019356
## address    -0.002594 -0.020381 -0.014864 -0.025062 -0.008658  -0.015900
## all        -0.048223 -0.038766 -0.048558 -0.052115  0.026861  -0.025327
## X3d        -0.009222 -0.006053 -0.012810 -0.009137 -0.003424  -0.001959
## our        -0.050202  0.020726 -0.042937 -0.077141 -0.027299  -0.032703
## over       -0.029774 -0.029474 -0.053106 -0.032807 -0.015323  -0.031528
## remove     -0.046928 -0.035397 -0.050432 -0.053318 -0.018359  -0.031744
## internet   -0.000622 -0.030573 -0.002283 -0.037672 -0.007304  -0.021339
## order      -0.035678 -0.035572 -0.074810 -0.056935  0.001649  -0.026100
## mail        0.023199 -0.029820 -0.035488 -0.031454 -0.013833  -0.017003
## receive    -0.039710 -0.037160 -0.065565 -0.050417 -0.019697  -0.022731
## will       -0.023725  0.020094 -0.088980 -0.070711  0.000301   0.032606
## people     -0.021314 -0.025009 -0.042143 -0.021665 -0.013872  -0.020310
## report     -0.008669 -0.013395  0.000460 -0.020718  0.045841  -0.016535
## addresses   0.047055 -0.018321 -0.039109 -0.025447 -0.011878  -0.021064
## free       -0.044064 -0.032580 -0.045388 -0.045340 -0.018625  -0.028389
## business   -0.050136 -0.021401 -0.058429 -0.057741 -0.011410  -0.030238
## email      -0.016804 -0.034225 -0.047886 -0.040245  0.018831  -0.014864
## you        -0.050955 -0.068312  0.113371 -0.000953 -0.003724  -0.040016
## credit     -0.029719 -0.013586 -0.042456 -0.030114 -0.007759  -0.017949
## your       -0.049048 -0.062806 -0.033541 -0.077785  0.009475  -0.049762
## font       -0.021310 -0.012198 -0.030609 -0.019640  0.026652  -0.013185
## X000       -0.040093 -0.035363 -0.053903 -0.053022 -0.018025  -0.030575
## money      -0.037067 -0.024075 -0.045531 -0.032070 -0.011132  -0.015888
## hp          0.107093 -0.002204  0.048808 -0.045508  0.001083  -0.003065
## hpl         0.131770  0.014170 -0.007791 -0.036000  0.032413   0.032248
## george      0.008761 -0.012548 -0.008292 -0.038259 -0.010922   0.006146
## X650        0.076317 -0.000683 -0.006670 -0.008321  0.002878   0.003535
## lab         0.059172 -0.004201 -0.010995 -0.023194 -0.003140  -0.003606
## labs        0.075226 -0.006072 -0.002694 -0.011980 -0.004895  -0.004893
## telnet      0.088201  0.063679 -0.008298 -0.022649 -0.002910  -0.013712
## X857        0.113625  0.003908  0.011338 -0.019207  0.000185  -0.011788
## data        0.006800  0.023340  0.000233 -0.019220 -0.000709   0.004328
## X415        0.113919  0.005943  0.013131 -0.014708 -0.000105  -0.010872
## X85         0.062712  0.001007 -0.015558 -0.014682  0.014842   0.001784
## technology  0.083188  0.006323 -0.011302 -0.034456 -0.002141   0.009778
## X1999       0.323007  0.003469  0.048330  0.127211 -0.019020   0.048497
## parts      -0.006859  0.001491 -0.001870 -0.009884 -0.004085   0.017611
## pm          0.193614  0.013866  0.103555 -0.002706  0.003633   0.009767
## direct      0.097512 -0.002271 -0.010240 -0.027652  0.000478  -0.015779
```

```
## cs          0.078581 -0.007805  0.012074  0.346614 -0.007277  -0.001362
## meeting     0.017484  0.014446  0.002506 -0.019544  0.011813  -0.000562
## original    1.000000  0.008132  0.081701  0.013216  0.016567  -0.004875
## project     0.008132  1.000000  0.004166 -0.015281 -0.004321   0.001037
## re          0.081701  0.004166  1.000000  0.041279 -0.013015   0.004263
## edu         0.013216 -0.015281  0.041279  1.000000 -0.010794  -0.015767
## table       0.016567 -0.004321 -0.013015 -0.010794  1.000000  -0.006656
## conference -0.004875  0.001037  0.004263 -0.015767 -0.006656   1.000000
## X.          0.010448 -0.007098 -0.024686  0.017230  0.076991  -0.002018
## X..1        0.058764 -0.003649  0.002402  0.013818 -0.005963  -0.007849
## X..2        0.113004 -0.011405  0.007446 -0.002190 -0.005032  -0.006544
## X..3       -0.048649 -0.032571  0.067387 -0.030952 -0.018766  -0.026425
## X..4       -0.053125 -0.036682 -0.047511 -0.045632 -0.019292  -0.030556
## X..5       -0.013234  0.000730 -0.023269 -0.015130  0.007683  -0.008616
##                  X.     X..1      X..2     X..3     X..4      X..5
## make       -0.02656 -0.02229 -0.033026  0.05863  0.11816 -0.008628
## address    -0.00827 -0.04956 -0.018548 -0.01534 -0.00963  0.001997
## all        -0.03510 -0.01628 -0.033534  0.10627  0.08607 -0.001445
## X3d        -0.00117 -0.01227 -0.007194 -0.00288  0.01120 -0.000309
## our        -0.03250 -0.04765 -0.027139  0.03131  0.04485  0.001447
## over       -0.01832 -0.00865 -0.016825  0.06535  0.10744  0.020111
## remove     -0.03288 -0.05356 -0.026639  0.04994  0.06217  0.046662
## internet   -0.02792 -0.03522 -0.018751  0.03185  0.05452 -0.006532
## order      -0.01491 -0.02981  0.014616  0.04278  0.15073 -0.000741
## mail        0.01749  0.00294  0.006915  0.03612  0.06924  0.045729
## receive    -0.03264 -0.05734 -0.024577  0.02273  0.07023  0.001397
## will       -0.02793 -0.02738 -0.044731  0.01170  0.01334 -0.030785
## people     -0.02289 -0.05115 -0.027604  0.04085  0.20454 -0.014317
## report     -0.01944 -0.00915 -0.013384 -0.00768  0.08983  0.006303
## addresses  -0.01854 -0.00224 -0.002366  0.01709  0.11751 -0.005635
## free       -0.02627 -0.04665 -0.030073  0.10366  0.04905  0.035092
## business   -0.03059 -0.03639 -0.036148  0.07702  0.10124 -0.000728
## email      -0.04028 -0.03207 -0.017080  0.03666  0.06433  0.021397
## you        -0.04253 -0.12823 -0.061791  0.15282  0.09142 -0.001578
## credit     -0.01981 -0.01860 -0.012311  0.04595  0.03447  0.007708
## your       -0.05864 -0.08422 -0.043454  0.07724  0.13485 -0.003637
## font        0.40042 -0.04518 -0.001455 -0.00466 -0.01057  0.188511
## X000       -0.02756 -0.03789  0.000471  0.07244  0.30999  0.020064
## money      -0.01873 -0.03220 -0.020148  0.05114  0.10362  0.000811
## hp          0.02259  0.13444  0.038424 -0.09015 -0.08556  0.059285
## hpl         0.01207  0.14247  0.065471 -0.07795 -0.07995 -0.020620
## george     -0.02023 -0.02995 -0.019688 -0.06639 -0.06843 -0.020566
## X650       -0.02526  0.31641  0.032667 -0.06324 -0.06055 -0.011335
## lab        -0.01948  0.16052  0.006289 -0.04209 -0.04992  0.002016
## labs       -0.01990  0.23261  0.004264 -0.06215 -0.06514  0.057333
## telnet     -0.01672  0.23504  0.011073 -0.04469 -0.04705 -0.001390
## X857       -0.00957  0.30543  0.012043 -0.04127 -0.04197 -0.011364
## data       -0.00508  0.03014  0.111004 -0.04616 -0.04781 -0.010001
## X415       -0.01030  0.30596  0.013185 -0.03944 -0.03992 -0.010693
## X85        -0.02247  0.20228  0.036304 -0.04925 -0.04841 -0.009729
```

```
## technology -0.01965  0.24703  0.001263 -0.05978 -0.05722  0.008967
## X1999       0.05130  0.09227  0.069162 -0.05286 -0.06250 -0.016904
## parts       0.00908 -0.01036  0.001713 -0.01505 -0.01267 -0.003602
## pm          0.03630  0.10078  0.036978 -0.02241 -0.04296 -0.010718
## direct     -0.01981  0.27036  0.014240 -0.02967 -0.01687 -0.010607
## cs          0.05584  0.01387  0.029467 -0.02578 -0.03622 -0.011557
## meeting    -0.00780 -0.01158  0.011000 -0.03639 -0.04184 -0.003657
## original    0.01045  0.05876  0.113004 -0.04865 -0.05313 -0.013234
## project    -0.00710 -0.00365 -0.011405 -0.03257 -0.03668  0.000730
## re         -0.02469  0.00240  0.007446  0.06739 -0.04751 -0.023269
## edu         0.01723  0.01382 -0.002190 -0.03095 -0.04563 -0.015130
## table       0.07699 -0.00596 -0.005032 -0.01877 -0.01929  0.007683
## conference -0.00202 -0.00785 -0.006544 -0.02643 -0.03056 -0.008616
## X.          1.00000  0.04983  0.009310  0.01965  0.00580  0.048584
## X..1        0.04983  1.00000  0.020923 -0.03010  0.04733 -0.012516
## X..2        0.00931  0.02092  1.000000 -0.03079 -0.02370 -0.006766
## X..3        0.01965 -0.03010 -0.030793  1.00000  0.13775  0.019879
## X..4        0.00580  0.04733 -0.023697  0.13775  1.00000  0.005716
## X..5        0.04858 -0.01252 -0.006766  0.01988  0.00572  1.000000
```

```
correlations <- cor(spam[, 1:54])
corrplot(correlations, order = 'hclust',tl.cex = .35)
```



Con *findcorrelation()* identificamos las columnas que se recomienda eliminar:

```
highCorr <- findCorrelation(correlations,verbose = T, names = T, cutoff =
.80)

## Compare row 34  and column  32 with corr  0.987
##   Means:  0.149 vs 0.06 so flagging column 34
## Compare row 32  and column  40 with corr  0.85
##   Means:  0.133 vs 0.057 so flagging column 32
## All correlations <= 0.8

print(highCorr)

## [1] "X415" "X857"
```

Eliminamos las variables que presentan una correlación alta

```
filteredCorrData <- spam[, -c(32, 34)]
dim(filteredCorrData)

## [1] 4601   56
```

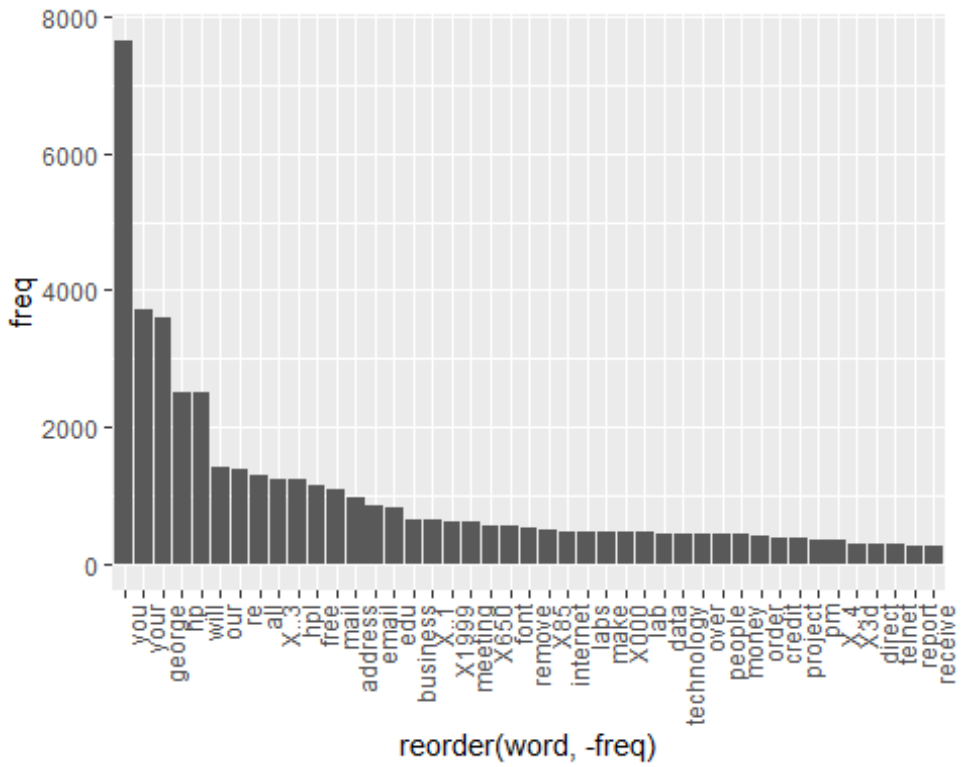
Frecuencia de palabras:

```
numeric_cols <- sapply(filteredCorrData, is.numeric)
freq<- sort(colSums(as.matrix(filteredCorrData[,1:52])), decreasing=TRUE)
wf<- data.frame(word=names(freq), freq=freq)
(wf)

##                  word   freq
## you               you 7650.2
## your             your 3718.3
## george         george 3603.8
## hp                 hp 2517.9
## will             will 2500.3
## our               our 1430.0
## re                 re 1398.5
## all               all 1288.1
## X..3             X..3 1232.6
## hpl               hpl 1227.0
## free             free 1157.0
## mail             mail 1099.2
## address       address  981.4
## email           email  848.6
## edu               edu  824.2
## business     business  655.8
## X..1             X..1  635.4
## X1999           X1999  620.1
```

```
## meeting       meeting  605.6
## X650             X650  567.5
## font             font  554.4
## remove         remove  528.6
## X85               X85  489.4
## internet     internet  481.5
## labs             labs  479.0
## make             make  477.2
## X000             X000  472.7
## lab               lab  460.7
## data             data  453.6
## technology technology  447.2
## over             over  438.8
## people         people  433.5
## money           money  432.8
## order           order  415.8
## credit         credit  385.1
## project       project  377.2
## pm                 pm  355.8
## X..4             X..4  342.2
## X3d               X3d  301.0
## direct         direct  298.2
## telnet         telnet  294.9
## report         report  274.7
## receive       receive  271.9
## addresses   addresses  228.9
## original     original  208.9
## X..5             X..5  202.8
## cs                 cs  200.6
## X.                 X.  172.7
## conference conference  147.7
## X..2             X..2   78.6
## parts           parts   60.7
## table           table   26.8
```

Eliminaremos las palabras con frecuencia < 230. *"table", "parts", "X..2", "conference", "X.", "cs", "X..5", "original", "addresses"*.

```
filteredlowfreq <- filteredCorrData[, -c(45,36,49,46,47,39,52,41)]

frecuencia_palabras <- ggplot(subset(wf, freq>230), aes(x=reorder(word,
-freq), y =freq)) +
  geom_bar(stat = "identity") +
  theme(axis.text.x=element_text(angle=90, hjust=1))
frecuencia_palabras
```



En el **cuarto punto**, proyectaremos los datos sobre un subespacio de dimensión menor utilizando PCA.

El PCA nos dará un número de componentes que nos servirán para condensar las variables del dataset.

```
pcaspam <- prcomp(spam[,1:57], scale = TRUE) # centrar variables para que
tengan media cero y al indicar
TRUE la desviación estándar de uno

summary(pcaspam)

## Importance of components:
##                           PC1    PC2    PC3   PC4    PC5    PC6    PC7
PC8
## Standard deviation     2.568 1.8049 1.4105 1.264 1.2459 1.2056 1.1921
1.1767
## Proportion of Variance 0.116 0.0571 0.0349 0.028 0.0272 0.0255 0.0249
0.0243
## Cumulative Proportion  0.116 0.1729 0.2078 0.236 0.2631 0.2885 0.3135
0.3378
##                            PC9   PC10   PC11  PC12   PC13   PC14  PC15
PC16
## Standard deviation     1.1378 1.1310 1.1002 1.067 1.0596 1.0460 1.040
1.0288
```

```
## Proportion of Variance 0.0227 0.0224 0.0212 0.020 0.0197 0.0192 0.019
0.0186
## Cumulative Proportion  0.3605 0.3829 0.4042 0.424 0.4438 0.4630 0.482
0.5006
##                          PC17   PC18   PC19   PC20   PC21   PC22  PC23
PC24
## Standard deviation     1.0251 1.0109 1.0059 1.0023 0.9961 0.9896 0.984
0.9734
## Proportion of Variance 0.0184 0.0179 0.0177 0.0176 0.0174 0.0172 0.017
0.0166
## Cumulative Proportion  0.5190 0.5370 0.5547 0.5723 0.5897 0.6069 0.624
0.6405
##                          PC25   PC26   PC27   PC28   PC29   PC30   PC31
PC32
## Standard deviation     0.9684 0.9642 0.9586 0.9492 0.9349 0.9302 0.9143
0.9111
## Proportion of Variance 0.0164 0.0163 0.0161 0.0158 0.0153 0.0152 0.0147
0.0146
## Cumulative Proportion  0.6570 0.6733 0.6894 0.7052 0.7205 0.7357 0.7504
0.7650
##                          PC33   PC34   PC35   PC36   PC37   PC38   PC39
PC40
## Standard deviation     0.8956 0.8846 0.8779 0.8658 0.8579 0.8522 0.8415
0.828
## Proportion of Variance 0.0141 0.0137 0.0135 0.0132 0.0129 0.0127 0.0124
0.012
## Cumulative Proportion  0.7790 0.7927 0.8063 0.8194 0.8323 0.8451 0.8575
0.870
##                          PC41   PC42   PC43   PC44   PC45   PC46    PC47
## Standard deviation     0.8145 0.8122 0.7835 0.7772 0.7656 0.7626 0.72651
## Proportion of Variance 0.0116 0.0116 0.0108 0.0106 0.0103 0.0102 0.00926
## Cumulative Proportion  0.8812 0.8927 0.9035 0.9141 0.9244 0.9346 0.94386
##                           PC48    PC49    PC50    PC51    PC52    PC53
PC54
## Standard deviation     0.69947 0.67264 0.63778 0.61397 0.60977 0.58222
0.55378
## Proportion of Variance 0.00858 0.00794 0.00714 0.00661 0.00652 0.00595
0.00538
## Cumulative Proportion  0.95244 0.96038 0.96751 0.97413 0.98065 0.98660
0.99198
##                           PC55    PC56    PC57
## Standard deviation     0.50861 0.43095 0.11329
## Proportion of Variance 0.00454 0.00326 0.00023
## Cumulative Proportion  0.99652 0.99977 1.00000
```

```
eig_val <- get_eigenvalue(pcaspam)
eig_val
```

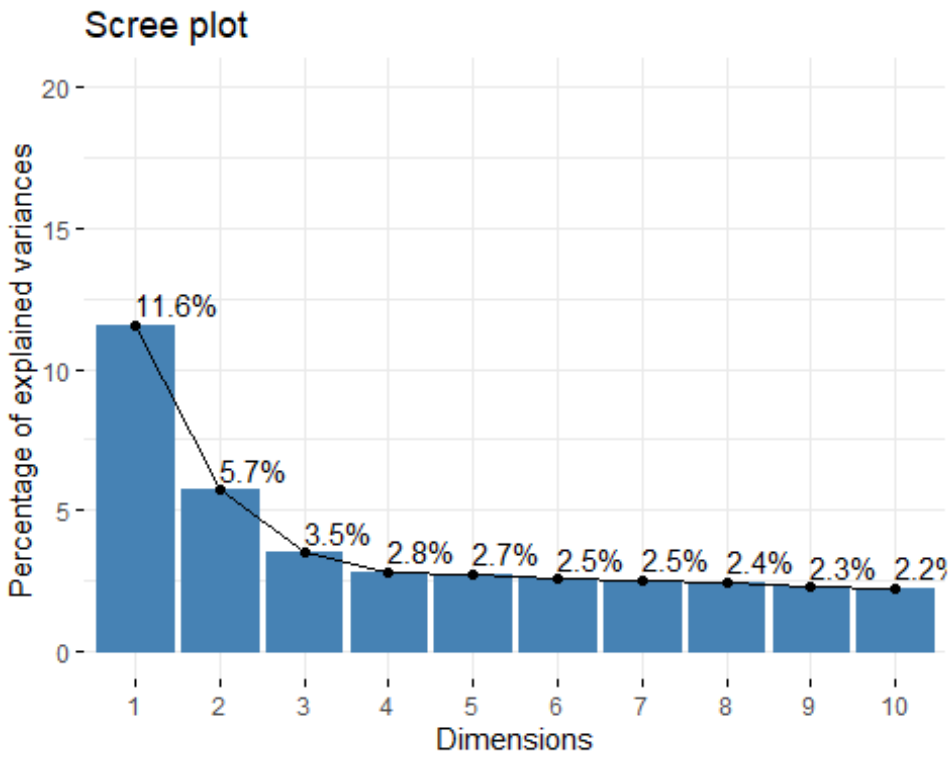
```
##        eigenvalue variance.percent cumulative.variance.percent
## Dim.1      6.5962          11.5723                        11.6
```

```
## Dim.2      3.2576          5.7151                          17.3
## Dim.3      1.9896          3.4906                          20.8
## Dim.4      1.5980          2.8035                          23.6
## Dim.5      1.5522          2.7231                          26.3
## Dim.6      1.4534          2.5499                          28.9
## Dim.7      1.4210          2.4930                          31.3
## Dim.8      1.3847          2.4292                          33.8
## Dim.9      1.2945          2.2710                          36.0
## Dim.10     1.2792          2.2442                          38.3
## Dim.11     1.2103          2.1234                          40.4
## Dim.12     1.1389          1.9981                          42.4
## Dim.13     1.1228          1.9699                          44.4
## Dim.14     1.0941          1.9195                          46.3
## Dim.15     1.0823          1.8988                          48.2
## Dim.16     1.0584          1.8568                          50.1
## Dim.17     1.0508          1.8435                          51.9
## Dim.18     1.0219          1.7929                          53.7
## Dim.19     1.0119          1.7752                          55.5
## Dim.20     1.0046          1.7624                          57.2
## Dim.21     0.9921          1.7406                          59.0
## Dim.22     0.9793          1.7181                          60.7
## Dim.23     0.9676          1.6976                          62.4
## Dim.24     0.9475          1.6623                          64.1
## Dim.25     0.9377          1.6451                          65.7
## Dim.26     0.9298          1.6312                          67.3
## Dim.27     0.9190          1.6123                          68.9
## Dim.28     0.9011          1.5808                          70.5
## Dim.29     0.8740          1.5333                          72.1
## Dim.30     0.8654          1.5182                          73.6
## Dim.31     0.8360          1.4666                          75.0
## Dim.32     0.8301          1.4563                          76.5
## Dim.33     0.8020          1.4070                          77.9
## Dim.34     0.7825          1.3728                          79.3
## Dim.35     0.7707          1.3521                          80.6
## Dim.36     0.7495          1.3150                          81.9
## Dim.37     0.7360          1.2912                          83.2
## Dim.38     0.7263          1.2743                          84.5
## Dim.39     0.7081          1.2422                          85.7
## Dim.40     0.6863          1.2041                          87.0
## Dim.41     0.6633          1.1638                          88.1
## Dim.42     0.6596          1.1572                          89.3
## Dim.43     0.6139          1.0770                          90.4
## Dim.44     0.6041          1.0598                          91.4
## Dim.45     0.5862          1.0284                          92.4
## Dim.46     0.5816          1.0203                          93.5
## Dim.47     0.5278          0.9260                          94.4
## Dim.48     0.4893          0.8583                          95.2
## Dim.49     0.4524          0.7938                          96.0
## Dim.50     0.4068          0.7136                          96.8
## Dim.51     0.3770          0.6613                          97.4
```

```
## Dim.52     0.3718          0.6523                        98.1
## Dim.53     0.3390          0.5947                        98.7
## Dim.54     0.3067          0.5380                        99.2
## Dim.55     0.2587          0.4538                        99.7
## Dim.56     0.1857          0.3258                       100.0
## Dim.57     0.0128          0.0225                       100.0
```

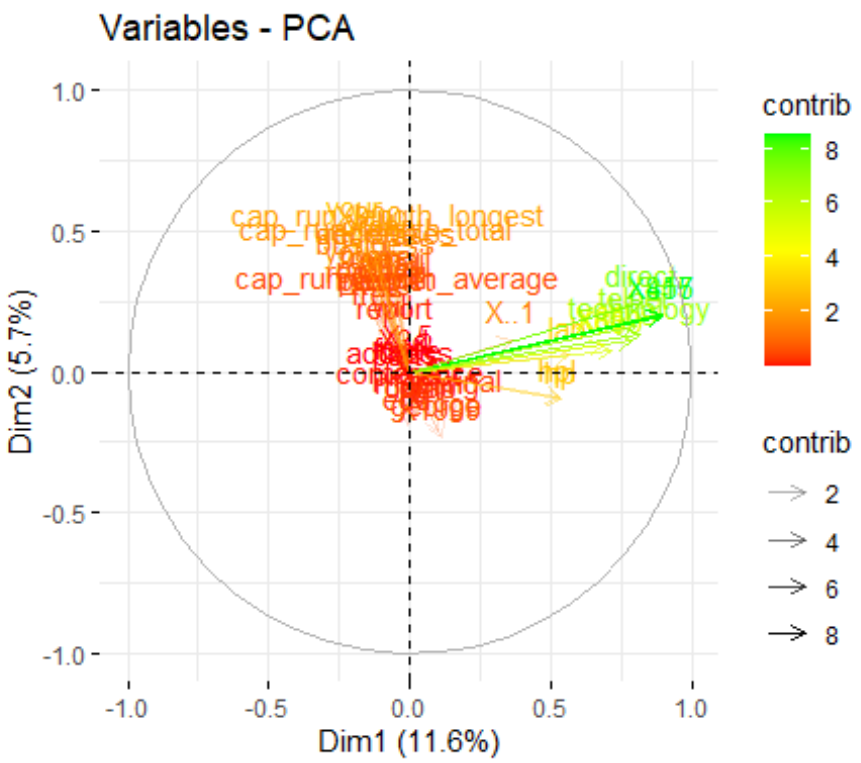
Como podemos ver, hay **20 componentes** que nos dan un *Eigenvalor* (denotado por la desviación estándar) mayor que 1. Este límite será nuestro criterio para seleccionar los componentes que mantendremos

```
fviz_screeplot(pcaspam, addlabels = TRUE, ylim = c(0, 20))
```



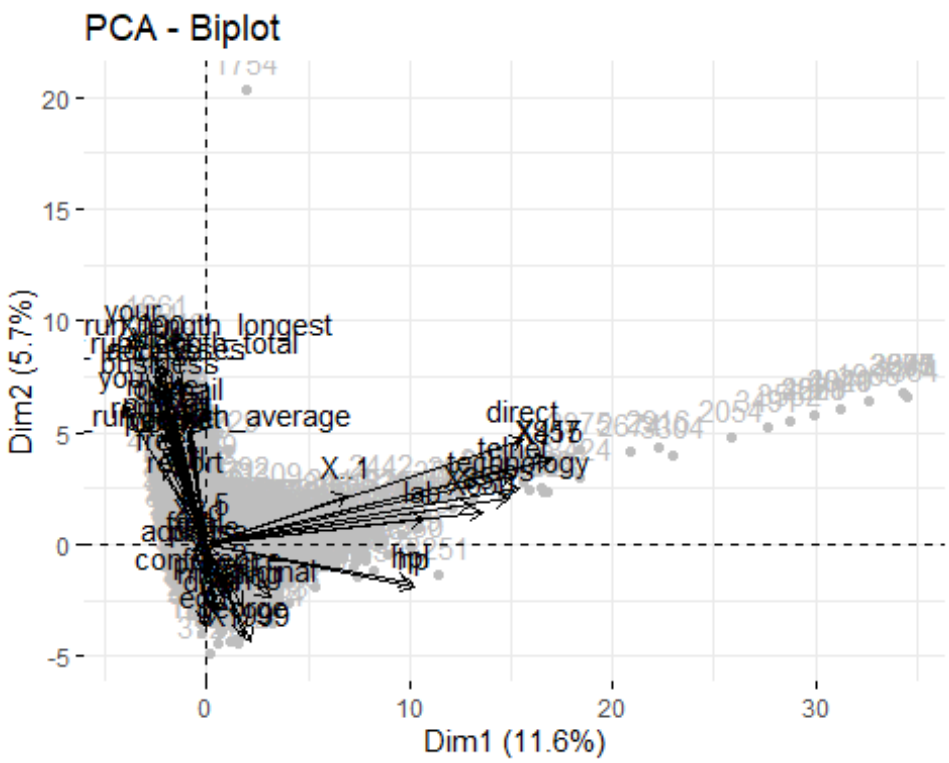
La componente 1 representa un 11,6% de la varianza total de las variables seleccionadas. La varianza de las otras componentes es 5,7% y 3,6% y luego se homogeniza en torno al 2,5%

```
fviz_pca_var(pcaspam,repel = F, colvar="cos2", col.var = "contrib", alpha.var
= "contrib", gradient.cols=c("#FF0000","#FFFF00","#00FF00"))
```



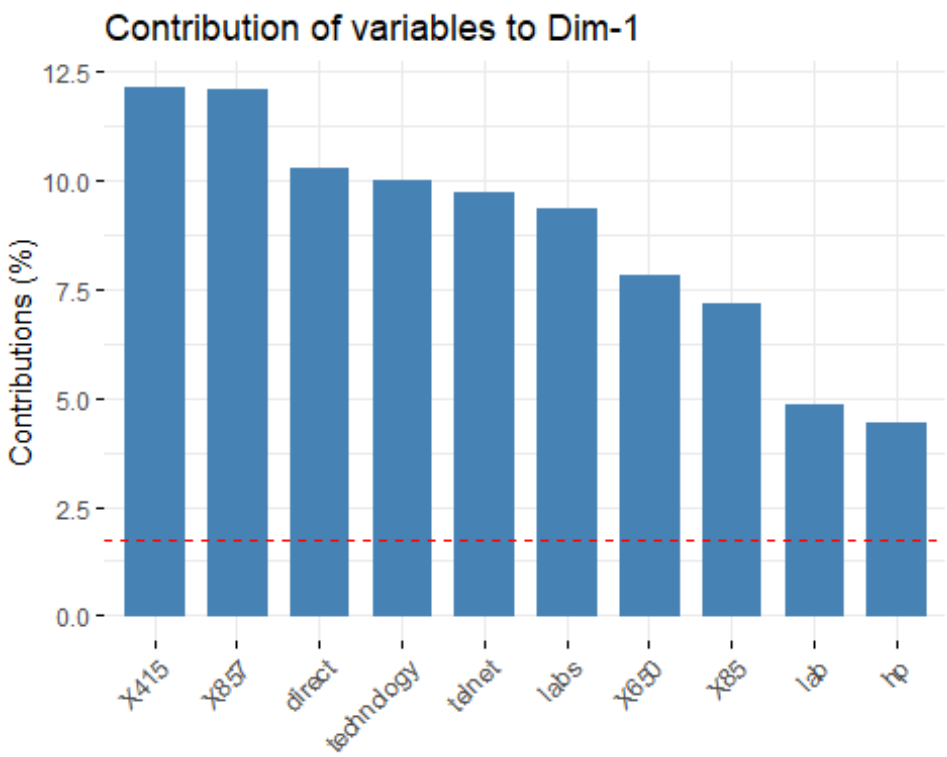
Para saber cómo está compuesto cada uno de estos componentes, podemos generar un Biplot. Este tipo de gráfico nos aparecerá como vectores en dos dimensiones (que serán los dos primeros componentes del análisis).

```
fviz_pca_biplot(pcaspam, repel = F, col.var = "black", col.ind = "gray")
```

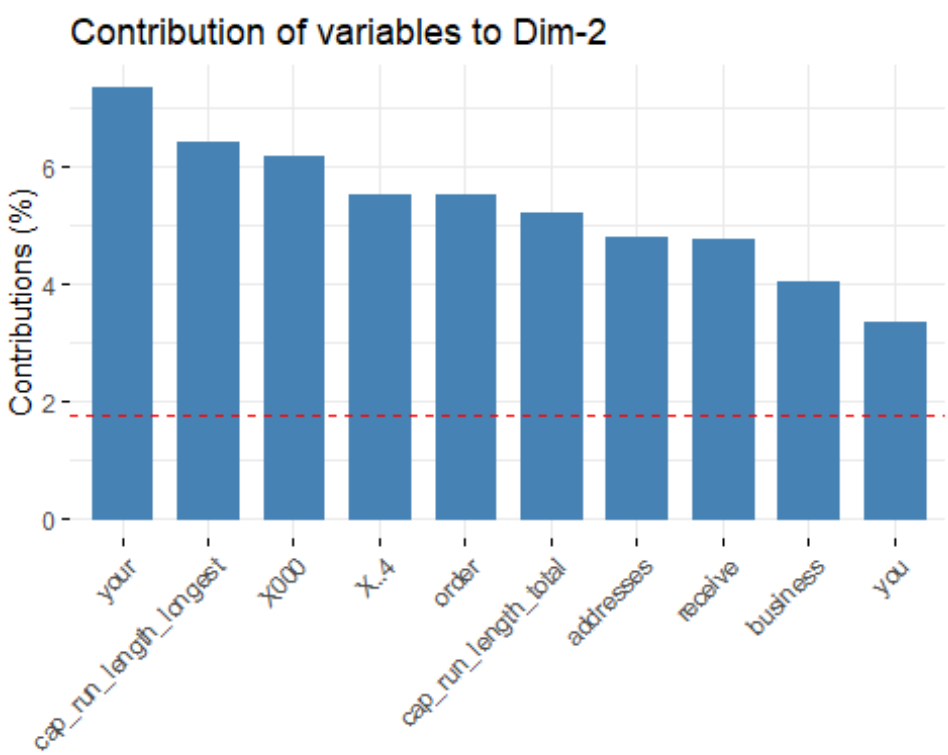


Con estos gráficos anteriores si podemos intuir una estructura de grupo. Estos conjuntos nos permiten representar las tres contenidos principales de los mensajes de spam.

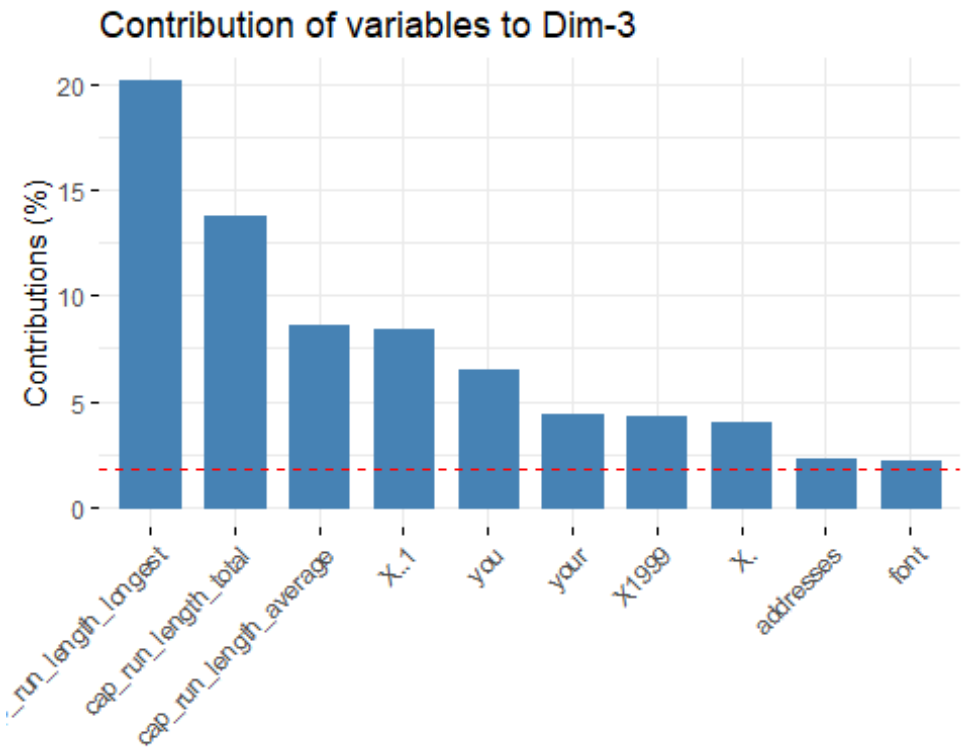
```
fviz_contrib(pcaspam, choice = "var", axes = 1, top = 10)
```




```
fviz_contrib(pcaspam, choice = "var", axes = 2, top = 10)
```



```
fviz_contrib(pcaspam, choice = "var", axes = 3, top = 10)
```



La primera dimensión es la que más grado de variación tiene. Representa el contenido relacionado con la tecnología. La segunda es la que denota un contenido más personal y de negocios. La tercera estaría relacionada con esas palabras que tienen mayúsculas y signos ortográficos.

El **apartado cinco** nos indica hacer un clustering atendiendo al contenido semántico

```
spam_escalado <- scale(spam[,1:57])
head(spam_escalado)

##        make address    all     X3d     our   over remove internet  order
## [1,] -0.346  0.3307  0.712 -0.0469  0.0138 -0.349 -0.293  -0.2615 -0.324
## [2,]  0.354  0.0517  0.435 -0.0469 -0.2558  0.675  0.242  -0.0866 -0.324
## [3,] -0.146 -0.1653  0.850 -0.0469  1.3768  0.346  0.191   0.0383  1.970
## [4,] -0.346 -0.1653 -0.554 -0.0469  0.4781 -0.349  0.497   1.3126  0.787
## [5,] -0.346 -0.1653 -0.554 -0.0469  0.4781 -0.349  0.497   1.3126  0.787
## [6,] -0.346 -0.1653 -0.554 -0.0469  2.3055 -0.349 -0.293   4.3608 -0.324
##         mail receive   will  people report addresses    free business
email
## [1,] -0.3699  -0.296  0.111 -0.3116 -0.178    -0.191  0.0825   -0.322
2.093
```

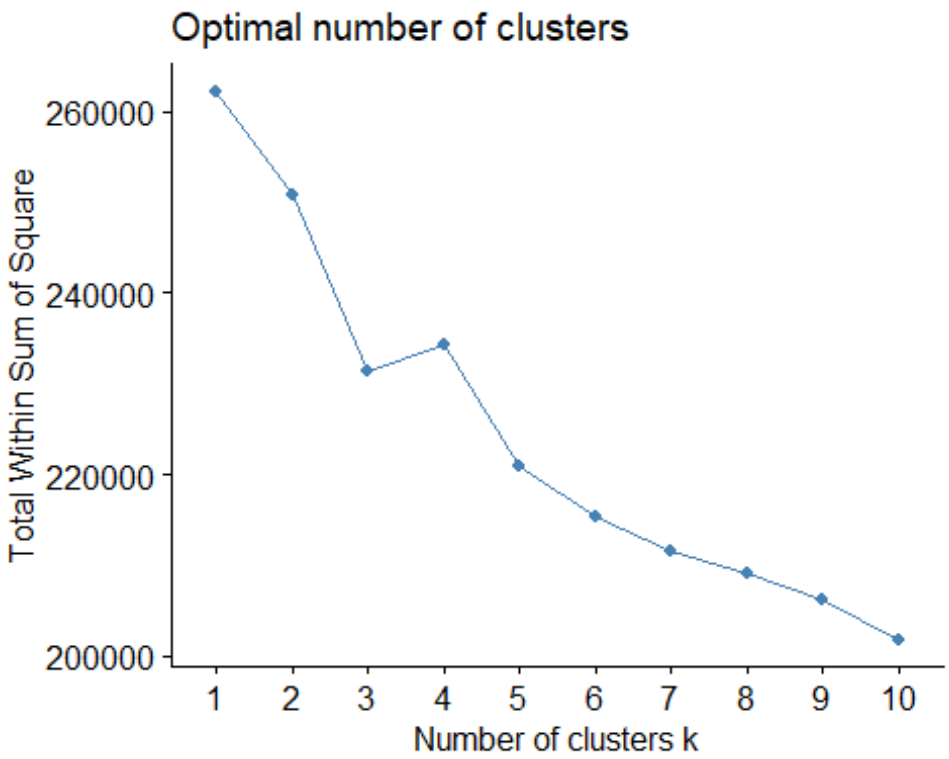
```
## [2,]  1.0854   0.755  0.285  1.8382  0.447     0.346 -0.1341   -0.164
0.181
## [3,]  0.0172   1.605 -0.108  0.0853  1.904     6.528 -0.2304   -0.186
1.601
## [4,]  0.6055   1.255 -0.269  0.7137 -0.178    -0.191  0.0704   -0.322
-0.349
## [5,]  0.6055   1.255 -0.269  0.7137 -0.178    -0.191  0.0704   -0.322
-0.349
## [6,] -0.3699  -0.296 -0.627 -0.3116 -0.178    -0.191 -0.3026   -0.322
-0.349
##         you credit   your   font   X000  money     hp    hpl george  X650
## [1,]  0.151 -0.165  0.126 -0.118 -0.292 -0.213 -0.328 -0.299 -0.229 -0.23
## [2,]  1.018 -0.165  0.648 -0.118  0.930  0.759 -0.328 -0.299 -0.229 -0.23
## [3,] -0.171  0.465 -0.247 -0.118  3.004 -0.077 -0.328 -0.299 -0.229 -0.23
## [4,]  0.855 -0.165 -0.413 -0.118 -0.292 -0.213 -0.328 -0.299 -0.229 -0.23
## [5,]  0.855 -0.165 -0.413 -0.118 -0.292 -0.213 -0.328 -0.299 -0.229 -0.23
## [6,] -0.937 -0.165 -0.670 -0.118 -0.292 -0.213  0.866 -0.299 -0.229 -0.23
##         lab   labs telnet   X857   data   X415    X85 technology  X1999
parts
## [1,] -0.168 -0.227  -0.16 -0.143 -0.175 -0.145 -0.199     -0.241 -0.321
-0.0598
## [2,] -0.168 -0.227  -0.16 -0.143 -0.175 -0.145 -0.199     -0.241 -0.154
-0.0598
## [3,] -0.168 -0.227  -0.16 -0.143 -0.175 -0.145 -0.199     -0.241 -0.321
-0.0598
## [4,] -0.168 -0.227  -0.16 -0.143 -0.175 -0.145 -0.199     -0.241 -0.321
-0.0598
## [5,] -0.168 -0.227  -0.16 -0.143 -0.175 -0.145 -0.199     -0.241 -0.321
-0.0598
## [6,] -0.168 -0.227  -0.16 -0.143 -0.175 -0.145 -0.199     -0.241 -0.321
-0.0598
##          pm  direct     cs meeting original project     re    edu  table
## [1,] -0.178 -0.1852 -0.121  -0.172   -0.205   -0.13 -0.294 -0.196 -0.073
## [2,] -0.178 -0.1852 -0.121  -0.172   -0.205   -0.13 -0.294 -0.196 -0.073
## [3,] -0.178 -0.0138 -0.121  -0.172    0.337   -0.13 -0.236 -0.131 -0.073
## [4,] -0.178 -0.1852 -0.121  -0.172   -0.205   -0.13 -0.294 -0.196 -0.073
## [5,] -0.178 -0.1852 -0.121  -0.172   -0.205   -0.13 -0.294 -0.196 -0.073
## [6,] -0.178 -0.1852 -0.121  -0.172   -0.205   -0.13 -0.294 -0.196 -0.073
##      conference     X.     X..1   X..2     X..3   X..4     X..5
## [1,]     -0.112 -0.160 -0.51416 -0.156  0.62676 -0.305 -0.10280
## [2,]     -0.112 -0.160 -0.02270 -0.156  0.12791  0.434  0.00916
## [3,]     -0.112 -0.118  0.01825 -0.156  0.00995  0.450 -0.07948
## [4,]     -0.112 -0.160 -0.00409 -0.156 -0.16084 -0.305 -0.10280
## [5,]     -0.112 -0.160 -0.01153 -0.156 -0.16330 -0.305 -0.10280
## [6,]     -0.112 -0.160  0.31611 -0.156 -0.32917 -0.305 -0.10280
##      cap_run_length_average cap_run_length_longest cap_run_length_total
## [1,]               -0.05549                 0.0455             -0.00823
## [2,]                0.00738                 0.2508              1.23067
## [3,]                0.13313                 2.2219              3.26412
## [4,]               -0.05549                -0.0623             -0.15194
```

```
## [5,]                -0.05549                 -0.0623              0.22303
## [6,]                -0.05549                 -0.1906             -0.37825
```

Revisamos cuál sería el número óptimo:

Usamos el método *Elbow* que calcula la varianza total en función del número de *clusters* y escoge como óptimo el valor en el cual añadir más *clusters* apenas consigue mejoría.

```
fviz_nbclust(spam_escalado, kmeans, method = "wss")
```



El primer método que usaremos es el *k-means clustering*. Un método no jerárquico para agrupar objetos. Este algoritmo trabaja de la siguiente manera:

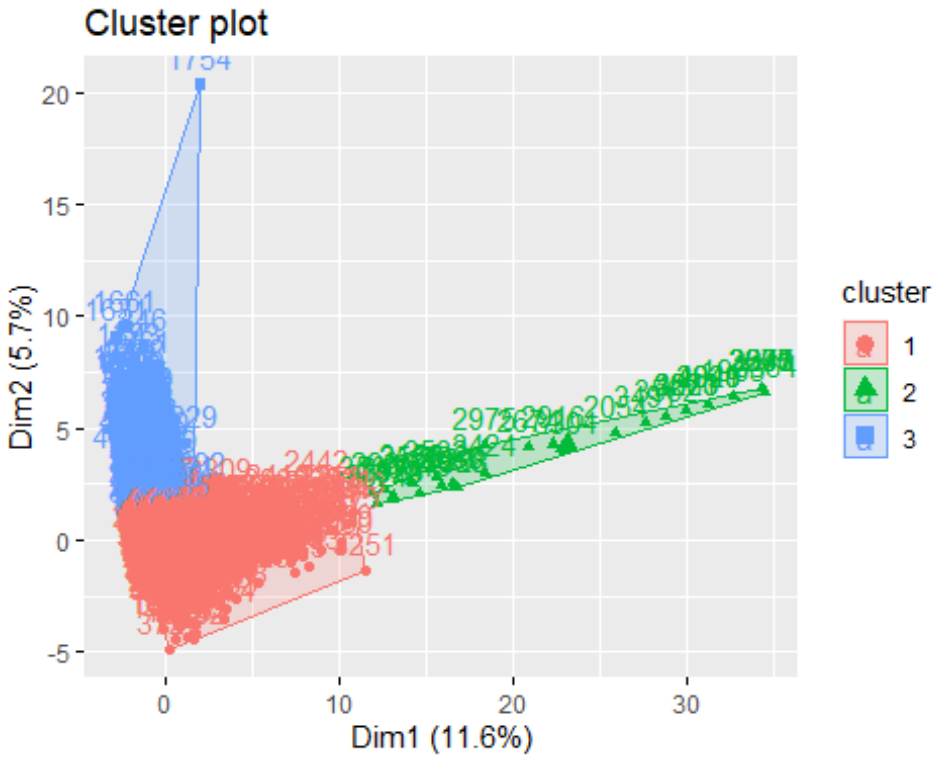
- Asigna un clúster inicial (de 1 a K) de manera aleatoria a cada observación.
- Itera hasta que la asignación de cada clúster deje de cambiar

```
set.seed(1234)
kmcluster <- kmeans(spam_escalado, centers = 3, nstart = 25) # centers = 3
porque ya sabemos el número
de grupos
str(kmcluster)

## List of 9
##  $ cluster     : int [1:4601] 1 3 3 1 3 1 3 1 3 1 ...
##  $ centers     : num [1:3, 1:57] -0.16351 -0.34564 0.41964 0.00551
```

```
-0.02493 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:3] "1" "2" "3"
##   .. ..$ : chr [1:57] "make" "address" "all" "X3d" ...
##  $ totss       : num 262200
##  $ withinss    : num [1:3] 141091 3607 86778
##  $ tot.withinss: num 231477
##  $ betweenss   : num 30723
##  $ size        : int [1:3] 3265 35 1301
##  $ iter        : int 2
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

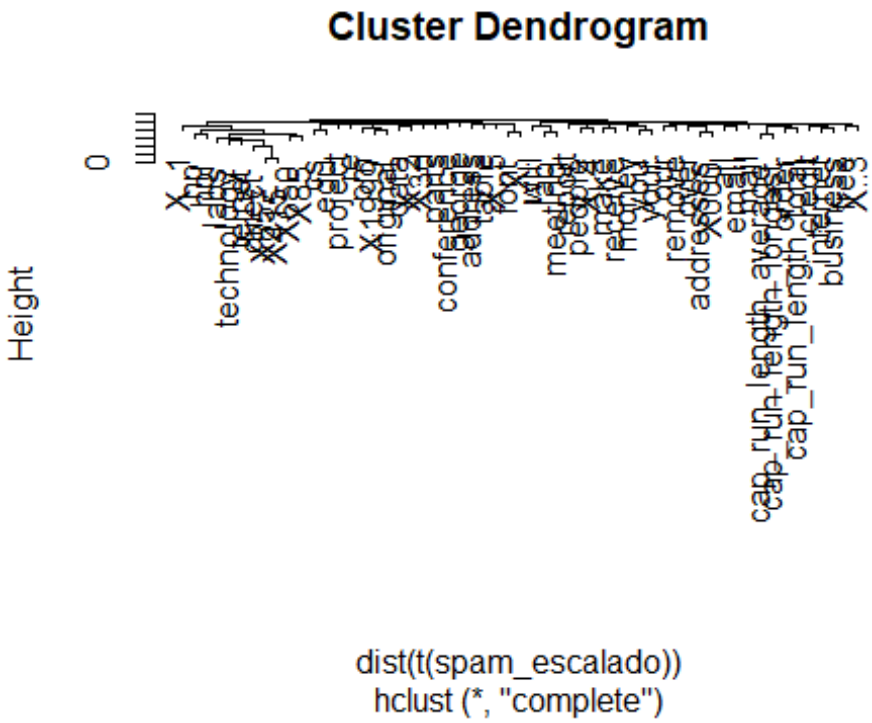
Al plotear podemos confirmar los tres grupos

```
fviz_cluster(kmcluster, data = spam_escalado)
```

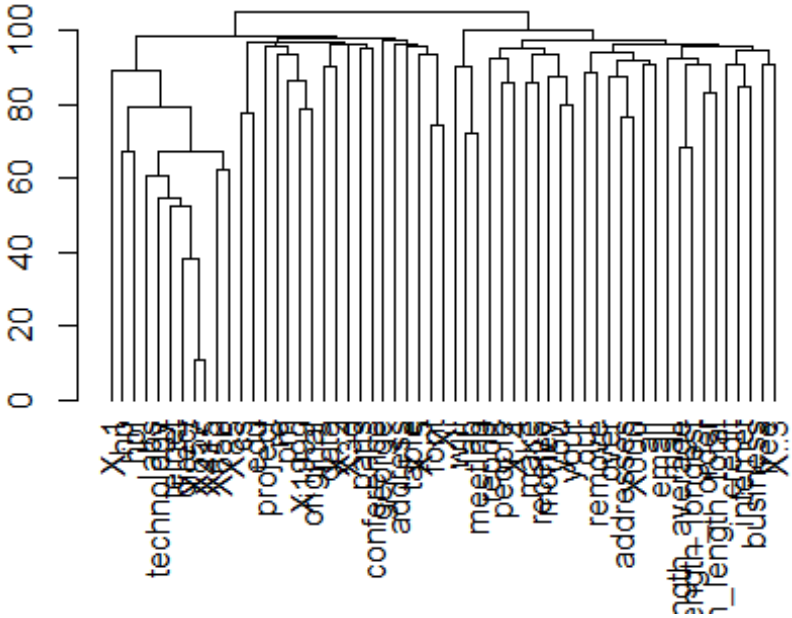


El segundo método será el clustering jerárquico. Cada hoja del dendrograma representa un elemento u observación. Conforme ascendemos por el árbol, algunas de las hojas se fusionan en ramas. Estas corresponden a observaciones que son similares unas a otras. Si ascendemos más en el árbol, las ramas se fusionan con hojas o con otras ramas. Las uniones más tempranas (más abajo en el árbol) corresponden con grupos de observaciones más similares entre sí. Por el contrario, las observaciones que se unen más arriba del árbol (cerca del final del árbol, más tardías) tienden a ser bastante diferentes.

```r
d3 <- dist(spam_escalado)

hc <- hclust(dist(t(spam_escalado)))
plot(hc)
```

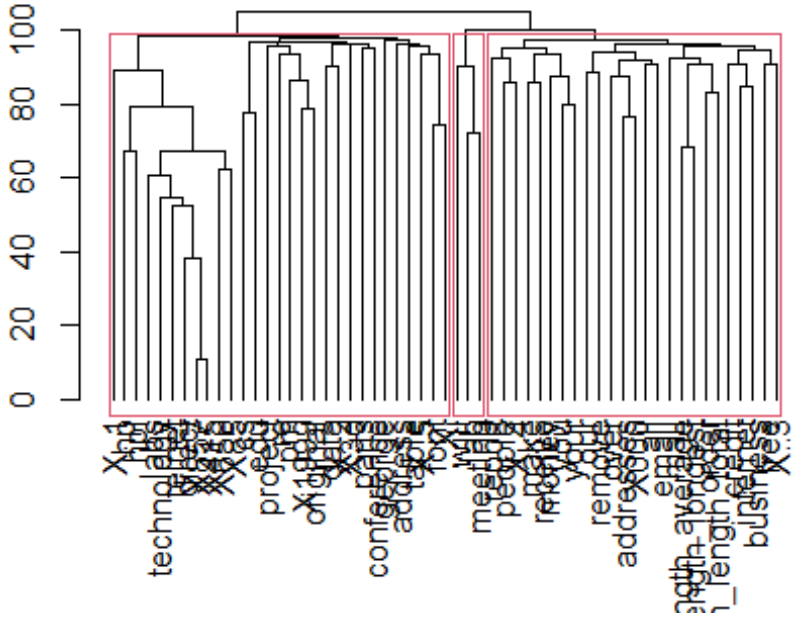


```
plot(as.dendrogram(hc))
```



En este último plot podemos distinguir los tres grupos

```
plot(as.dendrogram(hc))
rect.hclust(hc, k = 3)
```



Después de utilizar varios métodos de clustering vemos que los resultados son similares pero que en el caso de usar un clustering jerárquico debemos seleccionar bien cual será nuestro tipo de *linkage* y las altura de corte para obtener los clústeres.

Ambos métodos asignan forzosamente cada observación a un cluster. En nuestro ejemplo se ve la obs.1754 arriba del diagrama en azul. Esto a veces podría no ser apropiado.

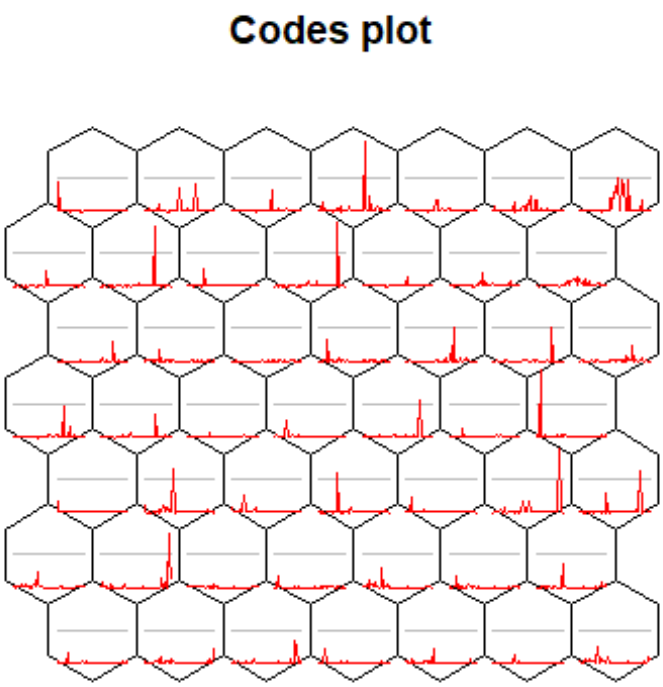
La **parte final** de la práctica nos habla de visualizar la estructura semántica de los documentos con mpas autoorganizativos.

Comenzamos normalizando nuestro conjunto de datos:

```
set.seed(100)
scalado_spam <- scale((spam[,1:57]), center=T,scale=T)
```

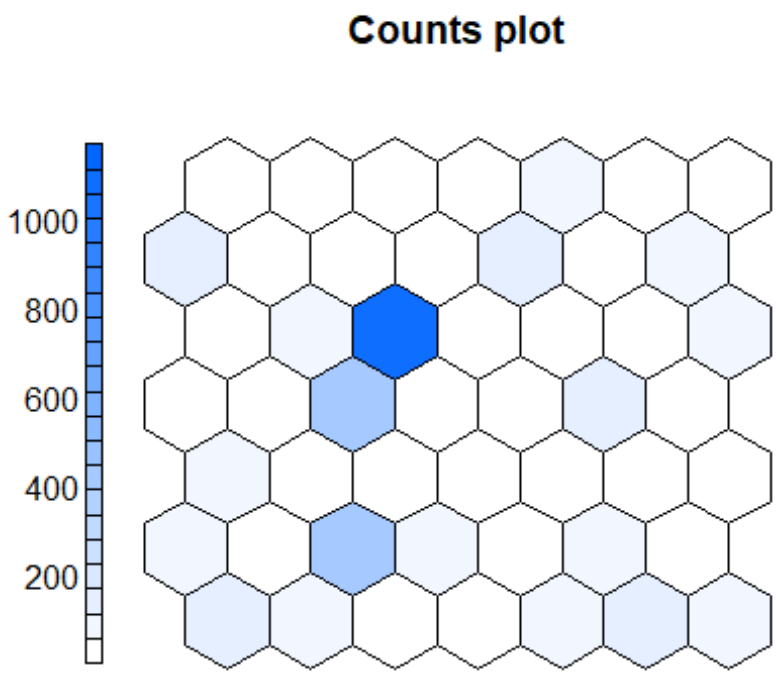
y comenzamos creando un *mapa de kohonen,* de 7x7

```
som <- som(scalado_spam, grid = somgrid(7,7,"hexagonal"), rlen = 1000)
plot(som, shape = "straight")

## Warning in par(opar): argument 1 does not name a graphical parameter
```



Ahora haremos una gráfica de densidad. El número de instancias en cada celda nos servirá para identificar áreas de alta densidad.

```
degrade.bleu <- function(n){
   return(rgb(0,0.4,1,alpha=seq(0,1,1/n)))}

plot(som, type="count", shape = "straight", palette.name = degrade.bleu)
```



Podemos ver el número de elementos en cada nodo de la siguiente forma:

```
nb <- table(som$unit.classif)
print(nb)

##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
16
##  177   69   21   58   65  139   83   74    8  444  109   37  110   38  116
11
##   17   18   19   20   21   22   23   24   25   26   27   28   29   30   31
32
##   57    7   60    5   14   21   27  469   45   17  163    6   39  110 1167
32
##   33   34   35   36   37   38   39   40   41   42   43   44   45   46   47
48
##   15   16   76  121    5   51    4  124   60  102   35   21   39    5   87
21
##   49
##   21
```

Comprobamos a qué nodo se ha asignado a cada observación de nuestro dataframe:

```
print(som$unit.classif)
```

```
##    [1]  6 10  7 24 24 31 17 31  8 10 17 31  6 17 10 40 10 10 30  7 31  1
19  6
##   [25] 31  1 31 13 31  8 15 11 10  6 17 11 19 10  6 11 32 19 19 30  8  1
10 17
##   [49]  2 10 10  5 31  4  5 24 10 27  6 11  6 24 13  6 10 10 40 13 10 40
12  6
##   [73]  6 10 10  5 11  1 10  6 24 10  6  4 31 32 10  6  1 41 12  2  2  2
7  6
##   [97] 24 31 31 15 19 31 40 31 40 31 40  4 40 13 13 31 31 31 25 24 31 31
12 12
##  [121] 31 31  2 31 31  8 10 24 10 10 10 10 24 30  5 31 31  5  6  1  5 30
2 30
##  [145]  6  5  5 31 30 32 10 12 19 31 11 38  1 10 10 14 14  7 24  6  7 10
13  6
##  [169] 31 24 11 10 24 24 32 17  8 10  6 17 30 30 10 10 24 31 31 19 15 10
13  8
##  [193] 10 31 10 31  5 10  5  6 12 10 10 10 30 24 24  8  6  7 30 13  5  5
19  6
##  [217] 31  6  7 10  7 15  1 10 10 10  6 30 31 12  6 24 31 30 18 31 31 10
13 14
##  [241] 31  1 13 32 10  6 13 13 10  1 31  4 10 10 10 24  6 10 10 10 10  1
10  6
##  [265]  6  1 10 10 13  1  1 31 38 19 13 31 10 13 31 13 13  6  6  8  1  6
31 24
##  [289] 10 10 13  8 10 10 31  5 10 10  1  4  8 17 27 10 10 10 31  1 31 10
24  6
##  [313]  1 10 31 17 31 10  4  4 10 40  6 31 40 31  1  6 10  1 17 10 24 31
10 10
##  [337] 10  6 10  6 10 10 10 10 10  9  6 31 10  6 10 31 24 10 10 10 31 13
13  2
##  [361] 31 11 27 10 19 10 24 10 10 11  2 11  9 11 13 31  2 10 24 10 17 24
7 24
##  [385]  1 10 10  1 10  7 10 31  7 40  1 38  6 32  7 24 17 15 24 12  1 10
19  2
##  [409] 31 24  7 31 10  1 14 10  8  8  8  2  8 31 10 24  2 38 13 13 11 10
30 27
##  [433] 30  6 10 17 31 31 10  4 31 12  2 30 31  6 14 10 10 10 10 31  7 10
10 31
##  [457] 31 32 14 14 24 24 11 31 30  3 17  1 24 24 24 24 24  1 10 31 24 31
31 31
##  [481]  8 31 31 10  8 12  8 32  7  5  4 10  8 14 10 31 24 32 10  2 31 31
7  7
##  [505] 31  5 31 30 17 32 10 24  7  2  1 31  1 31  4 24  4  4  3 10 10  7
1 24
##  [529]  1  2 10  5  5  4  6 31 31 31 15  6 10 10 10 10 10 16 31  6  6  6
24  8
```

```
##  [553] 32  6  2  8 10 10 30 10 31 10 10 13 13 13 15 13 31  2  8 32  8 10
15  1
##  [577]  1  1 13 38  1  2 10  1  1 31 24  1 15 24 13  6  8 31 31  5  4 13
4 15
##  [601]  2  5 25  8 10 10 32 17 24 10  7  5  7  7  2 10 31  4  4  5 14 10
24  1
##  [625] 10 15 13 31  5 10  1 24  2 24  6  6  8  6  8  8  5 15 12 24 17 13
12 10
##  [649] 31 12  1 30  6 10 10 10 10 10 31 30 10 24 31  2  1  5 31  6 10 14
5 10
##  [673]  9  6 19  2 10 24  1  2  2 31  4 15 31 10  2 10 13  6  6 31  5 30
31  7
##  [697] 10  7  7 10  1 14  6 10  1 13  5 24  6 10 10  7 10 13 10 13 13 10
10  1
##  [721]  1  7  7  7  7 10  5 10 31 13 31 10 30  7 14 10 19 31  5 24 10 31
13 14
##  [745] 31 30  7 27 31 24 10 13 30  8 10 10 10  5 30 31 10 10 10 12 10  1
10  8
##  [769] 31 17 14 10 10 31  4  5 19 10  7 13 10 27 10 31 24 31 31 10 10 13
10  1
##  [793] 31  1  1 15 10 10 10 10 13  1  7  1 10 24  9 31 13 10  9 10  7 24
1  6
##  [817]  2  1  8 10 10 10 31 31 24  5 20 15  2 10 12 10  7  7  7 10 17  8
6  8
##  [841] 10 13  8  2 31 12 13  1 15  7  7  7 24 31  8 31 31 10 31  8 31 31
25  8
##  [865] 13 10 10 24 31 10  6 10  8 10 31  3 24 17 10 10 15  1 10 24 10 10
10 10
##  [889]  9 31  1 10 30 31  2 24 10 31 10 10 10 13 11  6  2 30 32 31 10 13
2  1
##  [913] 38 10 12  8  5 10 14 24 10 10 10 14 10 17 13 14 13 11  2  5 10 13
1 28
##  [937] 31  1  1  1  6  6 11 11  9 31 31 13 31 31  8 10  7 11 19 19 19 19
30 19
##  [961] 14 12 10  2 11 11  6 14  6 15 24 15 24  8 12 12  1 31 14 10 12 10
4 14
##  [985] 10 14  5 13  8 10 24  7 24 15  6 13 13 13 24 10 10 10 17  3 10 10
1  5
## [1009]  1 16 16  6  6  1  3  1  1 31 10 30  1 24 38 10 14  7 10  1 15  5
10 10
## [1033]  7 10 24  1 19  5 10  1 15 13 10 31 13 40 13  1 10 10 13  1 31 10
10 40
## [1057] 15  2  1 14 10 31 24 13  1 13  4 15  8  3 31 10 10 15 13 13 10 15
8 13
## [1081] 10  1  1 10 10 10 10  5 15 30 14 13 10 10 10 10 14  8 13  7  6  1
28 10
## [1105]  6 10  5  5 10 17 24 14  1 10  2 10 28 10 10 10 15  2 28 10 10 10
11 24
## [1129] 10 10 30 10 10  8 24 10 31 10 11  2 10 10  6 24 10 30 24  3 10  5
11 10
```

```
## [1153] 31 17 10  3  1  3 31 12 24  2  4  5 15 12 12 10  8  6 13  2 10 13
13 13
## [1177] 31 15 24 40  6 12  8 10 40  1 12 40  6 24 12  8 30 10 30 10 10 15
6  6
## [1201]  2 11  7 12 12 19 30 31  2 24 10 28 30 24 10 10 10 13  2 15  8  8
8 31
## [1225] 10 19  2 30  7 12 30 10 10 31 12 31  1  8 10 31 38 10 10 25 10 13
31 10
## [1249] 12 10 30 10 30  8 19 10 13 10  1 10  5  1  5  5  5 31  7 13  7 10
4 10
## [1273] 10 10 24  7  8  6  6  7  7 31 13 10 31  7 31  4  3  5  4 10 31 20
25 10
## [1297] 13 13 25 32  6  6 10  7  1 41  7 24 24 10 15 32 31  2  2  8 38  2
30  1
## [1321]  1 19 30 32 10 10 24 31  8  1 25 16 25 16 25 16 31  5  6 10 14 31
31  4
## [1345] 10 17 17 38  1 10 10 40  7 10  5 14  8  2  8  8  7  3  6  8 31 31
10 10
## [1369]  4 31  8 31 10 10 10 10 24 11 10  1 31 10 24 10  1  2 13  2 24  3
24 17
## [1393] 10 31  8 10 31  4 30 19 11 14 30 31  5  8 10 30  1 32  2 13 11 38
24 24
## [1417]  2  7 31 24  4 30  1  7 31 19  2  8 10 24 13 12  1 38  5 40  1  1
5 10
## [1441] 40  6 13 10 17 10 27  2 10 10  1 24  1  4 10 13 17 18 24  8 17 10
31 31
## [1465] 10 17 30 30 31 10  5 10 30 30 10  7 10 24 31 10  7 10 24 16 31 31
5 38
## [1489]  2 10 10  6 15 10 10 10 13 17  5 10  5  5  1 24 10  8 31 31  4 13
10 31
## [1513] 14 13  4  5 10 31  1 31  6 31 31 10  6 11 31 24 10 13  6  1  3 31
10  6
## [1537]  6  6 13  2  4 10 31  5 13 24  6  7 10 30 24 10 10 10 32 18 31 17
11 10
## [1561] 10 10 24 32 25 13  6 11  1 10 19 24 13 17 24  4  7 10  7 31 10 31
8 10
## [1585] 31 31  8 15  5 24  8 16  7 11  6 10  6 31 10 10 10 10 10 10  7 10
31  7
## [1609] 10 10  4  7  1  1 31 31 31 31 32  7 25 31 15 31  5 24 14  2  4 32
16 31
## [1633] 19 31 30 31 31 31  1  2 31 27 10  6  1 24 27  1 31 31 12 25  8  2
31  7
## [1657] 10  1 31 30  7  2 10 31 10 10 17 10 28 30  1 31  6  2  4 31 24 31
15 25
## [1681] 31 10 10 24  1 14 11 38 31  4 31  3  1 30 31 41 24 17  1 24 31 15
1 15
## [1705] 30 25 10 18  1 31 31 31  3  2 31 24 10 38 10 17 31 38 31  1 17 17
19 10
## [1729] 30 10 13  1  2 25 24 25 24  1  6 25 24 13  1 30 11 13 13 13  4  6
4 25
```

```
## [1753] 31  9  1 10  6 13 10  1  7  6 32  7  7 13  3 13 31  7 24  7  7 10
24 31
## [1777]  1  6 13  8  2 11 15 10 14 17 12 24  8 14 13  7 24  5  6 13  1 14
10 32
## [1801]  4 32 11 13 35 10  7  7 24  5 10 10 10  2  7 36 40 31 42 31 36 33
24 47
## [1825] 15 40 31 31 11 27 31 27 22 24 41 31 11 11 47 35 31 31 48 19 24 24
24 31
## [1849] 47 31 11 38 27  4 31 31 31 27 40 22 22 11 10 40 10 31  1 15 31 35
31 31
## [1873] 38 12 24 31 31 47 48 31 48 27 47 31 31  6 31 31 31 31 31  1 33 40
31 31
## [1897] 38 40 31 22  6 31 22 31 31 40 44 41 31 10 31 31 27 22 31 31 24 40
40 11
## [1921] 35 27 48 48 10 35 40 27 27 35 40  1 27 47 35 31 31 31  6 38 35 36
31 27
## [1945] 31 40 24 31 31 33 31 31 31 31 15 31 41 31 48 31 35 31 44 31 49 36
19 45
## [1969] 31 38 31 31  1 31 24  1 15 19 31 41 40 15 47 31 11 30 30 47 47 45
15 31
## [1993] 35 40 26 31 31 42 19 31 35 31 31 31 27 42 31 41 15 42 42 27 31 10
31 31
## [2017] 38 27 35 31 31 31 31 31 24 31 31 27 31 31 35 36 31 40 31 31 31 48
35  1
## [2041]  6 41 31 31 41 15 45 45 31 41 24 24 42 49 24 36 42 40 24 42  1  1
41 31
## [2065] 19 44 42 31 24 31 31 47 35 42 31 40 41 15 31 31 31 15 31 31 31 10
35 47
## [2089] 27  4 24 31 24 31  1 31 24 31 31 31 36 31 31 42 27 23 35 35 31 25
38 36
## [2113] 36 31 31  1 31 31 38 38 42 15 10 47 31 36 31 31 41 41 31 15 22 31
41 31
## [2137] 31 13 15 31 27 31 31 41 41 31 31  1 36 31 42 44 47 41 27 27 47 47
35 27
## [2161] 44 31 31 31 31 24 10 17 37 31 31 42 47 27 31 38 37 31 38 27 11 10
40 31
## [2185] 31 40 31 36 24 40 31 31 31 31 31 27 24 31 47 42 42 42 30 30 15 40
11 31
## [2209] 31 36 36 47 42 24 31 15 31 31 42 26 47 42 27 41 26 24 30 31 24 47
27 42
## [2233] 41 27 31 24 19 40 31 31 31 31 31 48 31 31 31 31 24 42 25 31 31 31
31 31
## [2257] 31 24 15 17 31 31 31 45 47 31 24 40 31 45 49 31 31 31 31 36  1 42
17 31
## [2281] 24 31 31 31 42 31 35  1 31 31 42 31 31 10 40 40 36 24 31 48 31 22
27 34
## [2305] 10 10 13 24 31 30 31 42 11 31 11 31 31 27 27 31 31 31 31 44 24 31
31 42
## [2329] 34  1  1 31 31 40 40 31 31 31 31 15 27 35 31 34 42 40 31 31 31  1
27 31
```

```
## [2353] 35 27 40 40 36 27 41 31 31 11 40 45 24 30 45 31 31 44  1 40 47 31
31 31
## [2377] 15 45 31 31 11 30 35 24 24 31 23 31 31 24 24 35 31 31 41 31 33 40
47 27
## [2401] 42 31 15 15 42 40 31 45 31 42 24 15 35 15 31 19 47 10 35 31 42 24
1 31
## [2425] 31 35 42 35 24 47 31 10 26 27 31 31 11 31 24 24 36 42 41 31 24  1
1 24
## [2449] 31 42 31 35 35 31 31 27 31 31 47 35 35 13 31 31 10 31 40 40 30 27
40 24
## [2473] 24 30 24 35 24 41 31 24 31 47 42 17 42 17 11 24 31 31 10 41 19 44
42 24
## [2497] 27 31 33 48 24 24 24 40 31 31 31 24 38 40 10 17 31 11 24 11 42 47
31 47
## [2521] 31 36 31 31 31 10 10 36 31 31 24 24 31 31 19 35 10 31 31 24 24 23
31 41
## [2545] 27 24 24 31 24 11 31 47 15 31 42 31 30 36 27 30 27 47 47 35 11 31
24 41
## [2569] 36 31 36 33 41 24 49 31 24 27 31 31 17 45 31 31 31 41 31 25 42 27
31 36
## [2593] 36 31 35 35 31 35 42 27 31 31 45  6 19 27 41 31 27 31 31 31 31 19
44 31
## [2617] 47 31 31 31 45 23 31  1 31  4 31 35 27 45 49 31 31 31 47 31 36 31
31 27
## [2641] 22 22 22 10 24 24 27 24 26 47 31 24 35 31 31 47 27 42 42 31 42 31
40 15
## [2665]  1 47 31  1 31 31 31 11 31 49 42 31 31 27 40 25 27 31 15 41 31 31
24 47
## [2689] 11 11 47 27 27 10 31 23 49 31 19 11 40 27 31 27 23 31 40 24 10 31
31 31
## [2713] 11 31 31 31 27 24 42 45 23 35 44 27 31  6 31 24 24 24 24 31 24 31
31 40
## [2737] 31 31 31 11 31 36 27 31 27 40 19 34 27 31 48 31 31 38 34  6 41 47
31 31
## [2761] 48 42 41 34 31 45 31 31 43 41 31 31 45 11 31 35 42 40 23 31 31 31
15 31
## [2785] 47 31 27 15  6 41 23 41 31 24 42 31 31 33 11 31 31 42 17 31 36 11
38 31
## [2809] 31 31 24 15 36 36 36 41 31 40 11 31 31 42 31 31 40 15 31 15 31 24
31 24
## [2833] 38 40 31 27 38  1 17 31 24 27 31 27 31 27 41 23 49 31 11 31 31 31
31 31
## [2857]  6 30 31 30 35 31 41 31 45 42 38 31 45  1 31 31 40 40 35 27 31 40
42 25
## [2881] 44 41 27 27 24 47 44 44 44 31 41 48 35 24 47 30 27 24 19 42 31 27
31 35
## [2905] 23 25 41 33 24 31 24 31 31 27 31 49 40  8 30 35 27 31  1 31 33 31
47 24
## [2929] 27 27 34 30 30 31 31 31 27  1  1  1 40 41 19 31 31 31 31 11 24 35
31 20
```

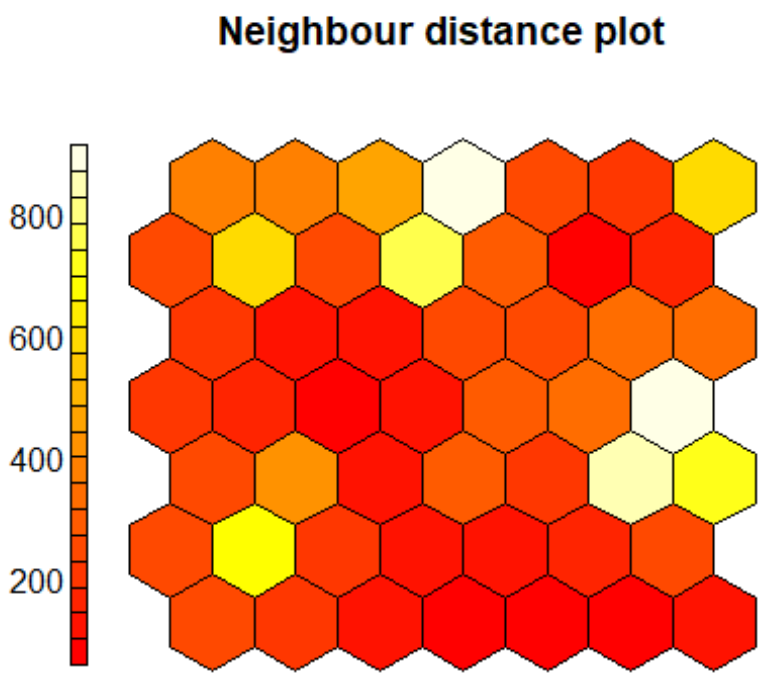
```
## [2953] 31 20 31 31 31 25 27 42 31 31 47 31 41 31 48 15 31 31 44  6 40 31
49 44
## [2977] 31 31 11 31 10 11 31 38 27 48 31 31 35 31 31 47 31 10 27 35 47 42
27 31
## [3001] 27 19 31 31 24 17 31 31 31  1 31  2 45 27 31 31 47 47 11 42 42 31
42 31
## [3025]  7 24 35 26 31 31 31 40 26 38 33 31  6 24 31 45 23  6 31 31 31 31
31 38
## [3049] 31 31 31 31 31 31 31  6 31 41 11 42 31 44 31 31 31 35 30  4 24 31
27 35
## [3073] 31 40 24 31 35 38 43 43 43 42 43 43 31 31 11 41 24 15 31 24 47 43
31 10
## [3097] 35 24 23 31 43 43 15 43 31 30 31 43  6 35 22  6 24 31 41 40 45 43
27 43
## [3121] 43 43 43 43 31 31 43 43 43 43 43 43 43 43 43 43 36 43 43 43 43 43
43 43
## [3145] 43 31 31 31 38 36 38 31 31 34 31 22 31 42 47 31 31 31 41 30 26 19
23 36
## [3169] 36 47 26 29 29 47 29 36 19 47 19 26 31 23 31 26 31  2 15 41 40 44
24 31
## [3193] 22 31 13 31 31 31 31 31 31 31 31 42 31 24 31 31 10  1 31 31 31 31
10 24
## [3217] 40 31 24  4 31 31 31 26 31 17 40 47 36 44 36 31 31 31 36 31 31 24
39 15
## [3241] 41 31  5 45 31 49 18 31 42  1 22 42 31 31 42 31 31 31 31 26 15 38
31 31
## [3265] 31 30 40 31 42 40 31 27 49 31 47 31 13 25 31 31 42 47 31 31 31 31
42 31
## [3289] 29 40 31  1  1 31 31 48 31 11 31 31 23 38 19 49 31 31 31 11  1 10
25 31
## [3313]  1 10 31 24 47 31 31 42 35 45 27 30 31 31 31 31 31 31 47 47 24 40
31 25
## [3337] 24 31 31 31 40 31 35 31 35 35 31 31 31 39 39 24 31 19 27 24 31 31
40 31
## [3361] 31 45 24 49 40 47 35 24 31 42 27 31 10 31  1  1 36 10 46 15  1 31
1 27
## [3385] 31 27 33 42 38 15 41 31 31 31 15 47 48 31 13 31 31 40 15 31 31 31
31 15
## [3409] 20 40 31 31 31 47 27 31 31 31 11 27 31 31 38 49 15 47 31 31 31 15
31 27
## [3433] 31 31 27 24 31 17 31 31 10 31 31 45 31 27 27 35 24 11 30 31  6 31
45 10
## [3457] 40 27 27 19 23 32 42 31 31 31 31 31 40 27  6 34 35 45 33 36 40 23
24 15
## [3481] 31 31 42 41  1 31 31 38 31 42 49 36 36 31 27 45 31 42 31  6  6 42
27 31
## [3505] 40 27 31 31 27 31 31 40 31 31 27 31 31 31 31 31 24 27 36 36 40 24
31 31
## [3529] 35 24 31 42 31 15 36 42 31 39 31 42 31 31 31  1 40 40 40 24 31 31
31 27
```

```
## [3553] 27 25 34 31 40 24 31 11 31 49 49 47 31 31 11 17 27 27 42 31 40 31
27 11
## [3577] 31 31 36 24 36 30 36 36 36 31 36 31 36 36 36 36 36 31 31 36 36 36
36 36
## [3601] 31 31 36 36 31 31 31 31 36 36 31 36 36 47 31 36 48 31 36 31 36 36
36 36
## [3625] 31 36 36 36 36 36 36 36 31 36 37 31 37 36 36 31 36 36 36 36 36 36
36 36
## [3649] 36 36 36 36 36 36 36 36 36 36 36 47 36 36 36 36 36 31 31 31 36 36
36 36
## [3673] 47 47 48 36 31 31 31 31 31 31 31 31 31 31 31 31 31 31 44 40 31 47
30 31
## [3697] 24 26 48 31 31 31 33 34 27 31  1  1  1 41  1 15 31 45 31 35 31 35
31  2
## [3721] 31 24 31 24 31 45 45 23 45 31 15  1 31 40 31 31 42 31 31 42 31 31
31 31
## [3745] 31 36 31  1 36 31 47 45 31 31 45 31 40 10 38 27  4 27 23 38 27 17
31 31
## [3769] 42 24 11 47 31 31 24 31 31 17 34 31 27 33 15 19 31 31  1 30 46 31
40 30
## [3793] 31 42 31 45 45 31 15 31 31 24 10 38 24 11 31 31 47 32 32 32 31 31
31 31
## [3817] 31 42 27 31 46 31 46 27 41 47 30 19 19 27 42 24 31 31 19 12 31 34
19 31
## [3841] 31 31 31 31 47 34 31 31 31 47 35  4 31 31 47 42 27 11 31 40 15 31
42 32
## [3865] 31 47 31 17 31 27 31 31 10 31 47 42 31 40 40 42 47 42  6 34 15 24
47 31
## [3889] 25 24 42 45 31 21 31 31  8 31 31 31 31 31 31 27 31 31 17 31 31 31
31 31
## [3913] 18 36 31 36 31 24 31 42 31 35 27  1  1 31  1 31 26 31 35 31 15 27
42 24
## [3937] 40 24 31 42  4 31 40 25 31 40 35 22 11 27 35 38 31 31 42 31 22 19
27 47
## [3961] 31 31 45 23 31 31 31 31 24 40 27 24 31 31 31 31 24 31 31 11 31 31
48 31
## [3985] 31 31 31 31 31 38 31 24 34 41 31 31 27 32 32 31 47 36 47 36 33 31
46 31
## [4009] 31 47 11 35 31 41 31 31 41 19 31 49 49 30 36 42 42 40 27 31 15 47
24 40
## [4033] 42 47 42 31 35 31 35 49 31 11  6 24 31 29 38 22 31 30 31 24 24 24
24 24
## [4057] 24 24 24 24 31  6 29 27 24 15 29 40 29 24 27 24 24 24 24 27 24 24
27 11
## [4081] 24  4 40 40 24 31 24 31  4 27 27 27 29 31 24 24 23 15 24 24 24 24
18 24
## [4105] 24 24 27 24 31 24 24 11 29 24 24 24 31 24 10 40 29 24 29 24 24 11
25 24
## [4129] 24 10  4 30 24  6 24 23 31 30 30 31 10 31 31 15 24 29 31 31 24 15
24 40
```

```
## [4153]  6  4 31 30 24 24 30 30 24 31 31 27 30 30 24 30 30 31 24 31 30 31
24 11
## [4177] 31 31 30 24 24 31 24 24 24 11 31 24 24 24 24 15 24  4 24 11 24 24
24 31
## [4201] 15 24 24 30 24 30 30 41 27 27 30 31 27 31  6 31  6 11 24 31 11 24
31 30
## [4225] 31 24 24 31 24 31 24 31 24 24 31 24 44 31 31 24  4 24 27 31  3 24
3 24
## [4249] 31 24 24 31 24 15 15  6 24 15 24 25 27 25 24 38 24 21 14 21 21 21
11 24
## [4273] 24 21 31 21 24 27 24 15 41 24 24 35 26 24 24 19  6 10 24 15 31 31
24 11
## [4297] 24 40 24 24 24 11 21 25 25 21 24 31 21 21 21 21 21 24 24 24 25 11
31 16
## [4321] 35 10 31 40 24 27  6 31 30 15 24 11 29 24 24 24 25 31 11 24 24 31
24 25
## [4345] 25  6 23 24 24 24 31 31  6 10 29 24 29 24 29 29 14  4 31 31 23 31
31 31
## [4369] 27 27 11 24 27 15 31 27 27 11 15 11 31 24 31 40 29 24 16  6 31 23
11 24
## [4393] 31 31 24 31 24 24 24 24 24 24 31 31 25 31 31  6 24 31 15 29 15 29
4 29
## [4417] 31 25 31 24 24  6 23 24 31 24 31  6 31 11 24 31 31 31 24 31 24 24
24 24
## [4441] 24 24 24 31 31 31 31 24 24 24 24 24 31  1 24 24 24 24 29  3 24 15
4  3
## [4465]  3 29 11 31 29 24 24 22 22 24 24 29 11 24 24 24 24 15 31 31 31 24
40 24
## [4489] 31 24 24 24 24 29 24 11 31 24 29 30 15 40 27 29 24 27 29 24 29 24
24 29
## [4513] 24 29 15 30 29 29 24 25 24 29 31 31 24 24 31 31 31 24 25 24 24 31
31 31
## [4537] 31 31 40 25 24 24 31 31 40 31 36 31 24 26 31 31 29 37 31 24 24 15
31 24
## [4561] 31 24 24 24 24 24 31 24 15 24 30 31 31 24 24 31 31 24 24 31 22 24
31 11
## [4585] 24 29 27 31 11 31 24 24 19 29 24 24 27 24 31 15 24
```

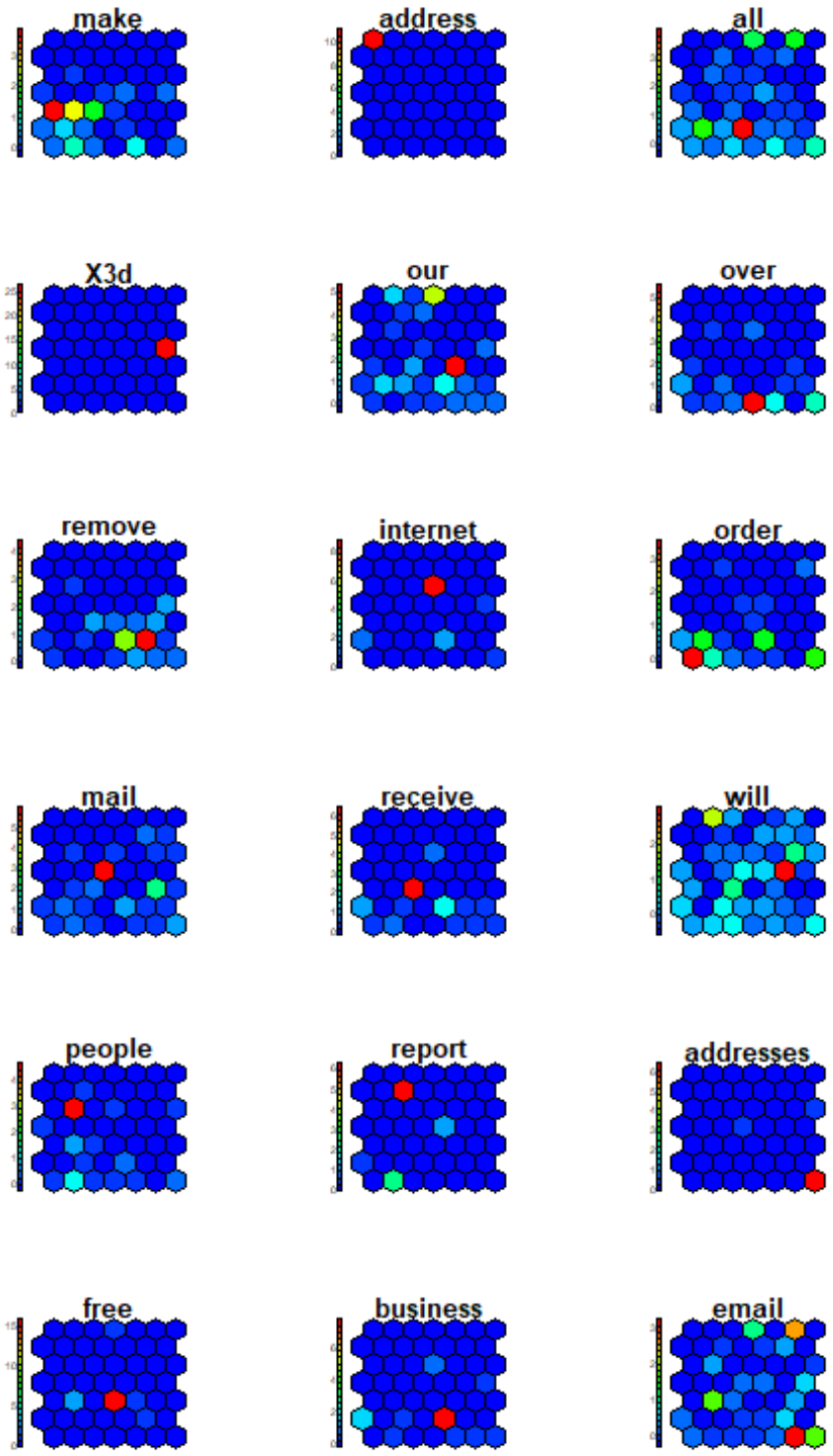
Gráfica de distancias.Es un mapa con la distancia euclidiana entre los vectores de cada neurona con su vecina representada con una degradación de colores.

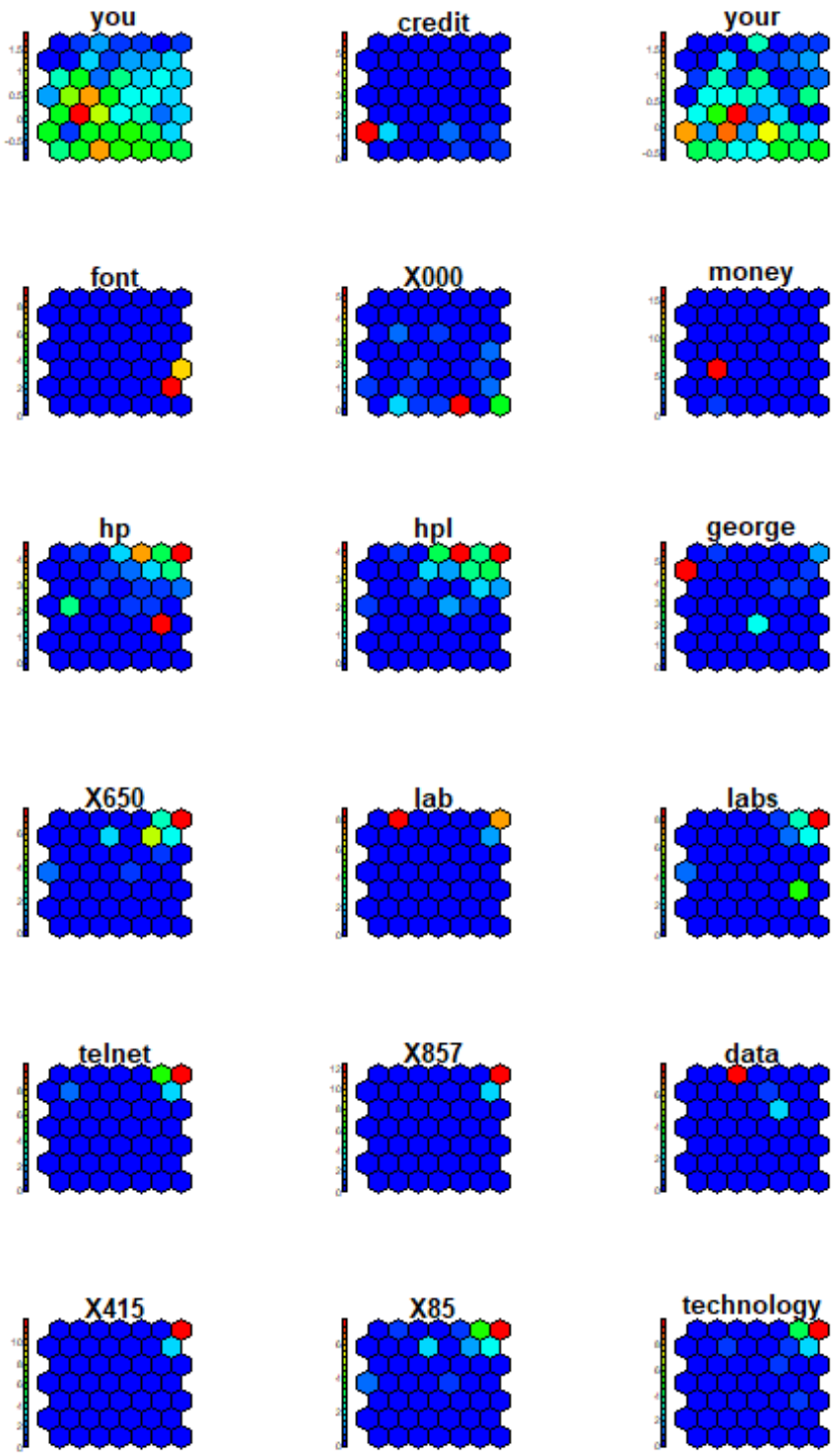
```
plot(som, type="dist.neighbours", shape = "straight")
```

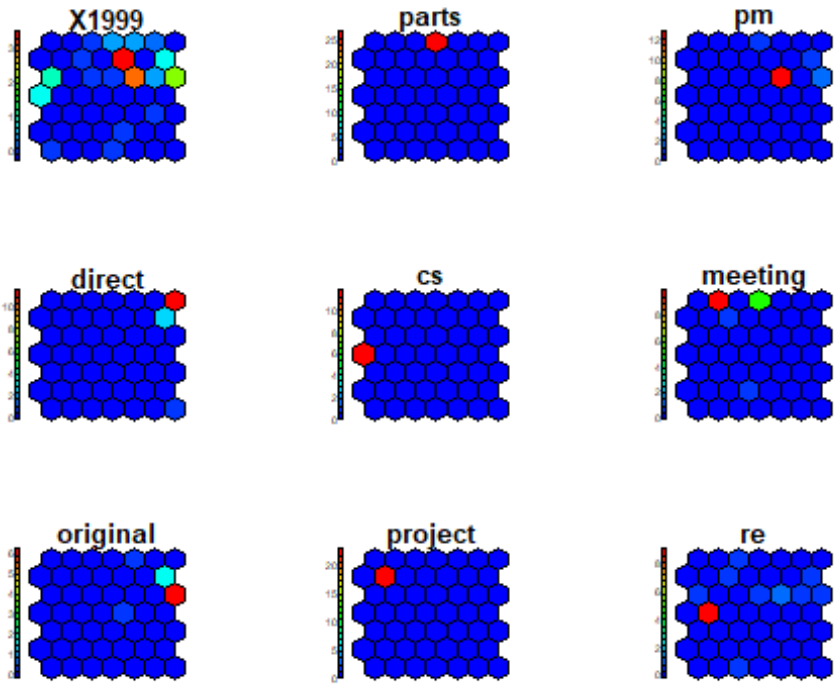


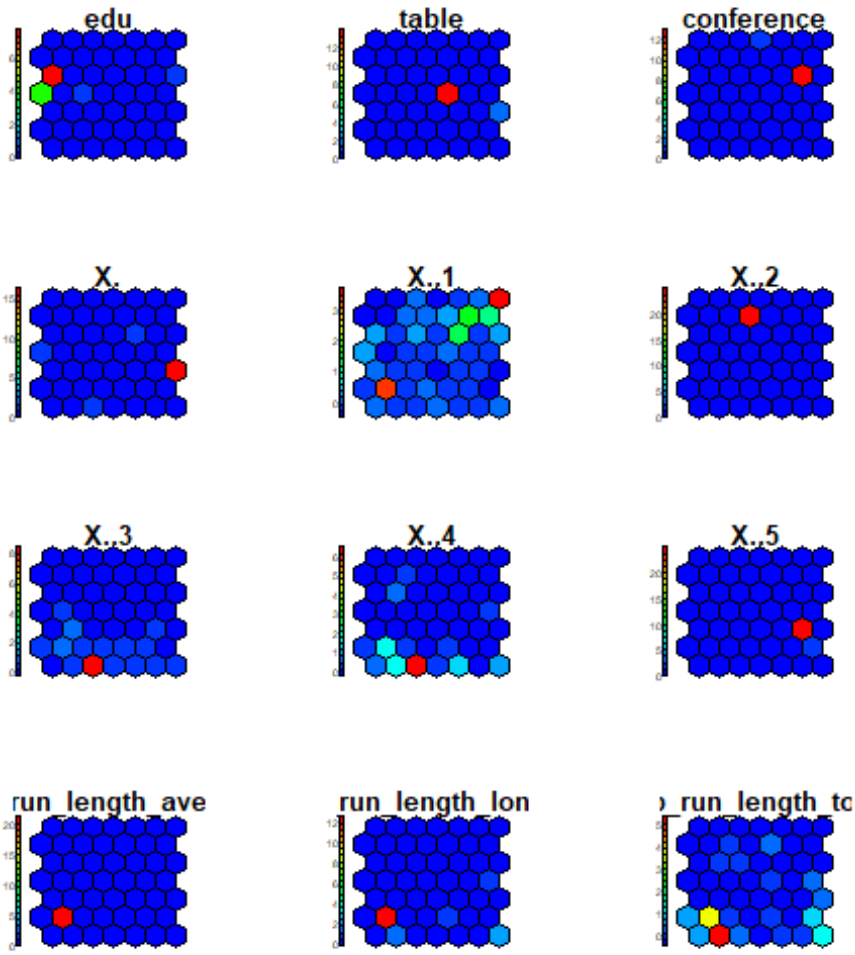
Hacemos un gráfico para cada variable resaltando los contrastes entre las áreas de alto y bajo valor.

```
coolBlueHotRed <- function(n, alpha = 1) {
  rainbow(n, end=4/6, alpha=alpha)[n:1]}

par(mfrow=c(3,3))
for (j in 1:ncol(spam[1:57])){
  plot(som,type="property",property=getCodes(som,1)[,j],
palette.name=coolBlueHotRed,main=colnames(spam[1:57])[j],cex=0.5, shape =
"straight")}
```


make
address
all
X3d
our
over
remove
internet
order
mail
receive
will
people
report
addresses
free
business
email


you
credit
your
font
X000
money
hp
hpl
george
X650
lab
labs
telnet
X857
data
X415
X85
technology


X1999
parts
pm
direct
cs
meeting
original
project
re


edu
table
conference
X.
X..1
X..2
X..3
X..4
X..5
run_length_ave
run_length_lon
run_length_to

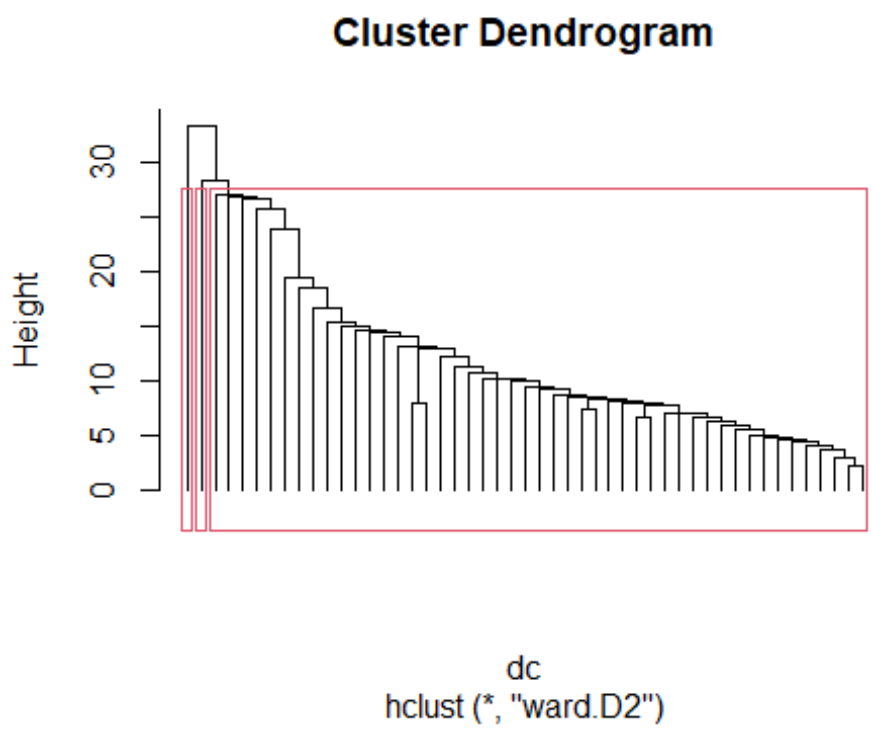
```r
# Importancia de cada variable - varianza ponderada por el tamaño de la celda
sigma2 <- sqrt(apply(getCodes(som,1),2,function(x,effectif)
   {m<-sum(effectif*(x-weighted.mean(x,effectif))^2)/(sum(effectif)-1)},
   effectif=nb))
 print(sort(sigma2,decreasing=T))
```

```
##                 george                address                   font
##                  0.960                  0.959                  0.935
##                    X3d                   X857                   X415
##                  0.930                  0.926                  0.924
##                     X. cap_run_length_average                  parts
##                  0.902                  0.890                  0.869
##              addresses                 direct                  table
##                  0.858                  0.857                  0.843
##                   X..5                   X650                 telnet
##                  0.842                  0.836                  0.833
##               original                  money             technology
##                  0.827                  0.826                  0.801
##                     cs                    lab                  order
##                  0.795                  0.787                  0.772
##             conference                receive                    edu
##                  0.767                  0.766                  0.766
##                     pm                 credit               internet
##                  0.759                  0.756                  0.755
##                project                   X000                   X..2
##                  0.753                  0.751                  0.746
##                 people               business                 remove
##                  0.745                  0.743                  0.740
##                   data   cap_run_length_total                   labs
##                  0.735                  0.734                  0.733
##                    hpl                 report                     re
##                  0.732                  0.720                  0.702
##                     hp                   over                meeting
##                  0.701                  0.701                  0.696
##                    all                  X1999                   make
##                  0.689                  0.689                  0.687
##                    our                  email                   your
##                  0.685                  0.668                  0.662
##                    X85                   free                    you
##                  0.659                  0.658                  0.644
##                   mail                   will cap_run_length_longest
##                  0.639                  0.638                  0.615
##                   X..3                   X..4                   X..1
##                  0.603                  0.571                  0.415
```

Clustering SOM:

Probamos con tres clústeres que es número con el que venimos trabajando

```
# Matriz de distancia entre nodos
dc <- dist(getCodes(som,1))
# HAC
cah <- hclust(dc,method="ward.D2",members=nb)
plot(cah,hang=-1,labels=F)


rect.hclust(cah,k=3)
```



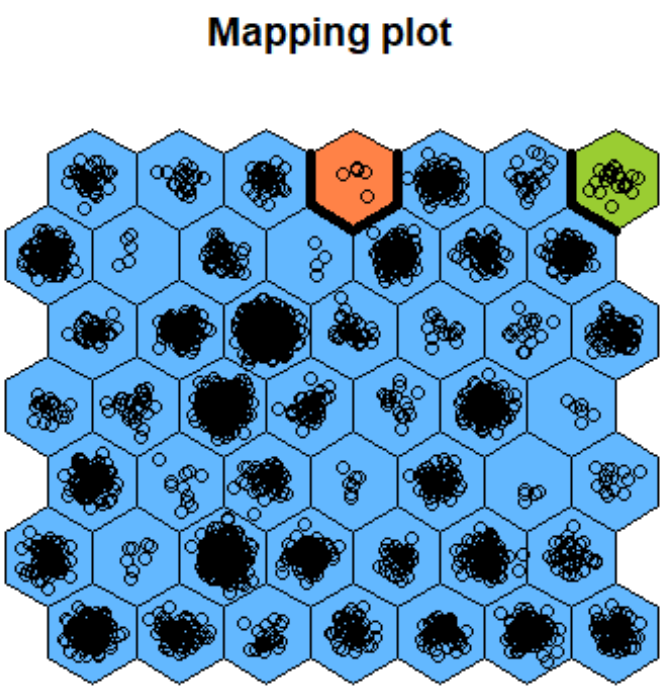
Clúster al que pertenece cada nodo del mapa.

```
groupes <- cutree(cah, k=3)
print(groupes)

##  V1  V2  V3  V4  V5  V6  V7  V8  V9 V10 V11 V12 V13 V14 V15 V16 V17 V18
V19 V20
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1
## V21 V22 V23 V24 V25 V26 V27 V28 V29 V30 V31 V32 V33 V34 V35 V36 V37 V38
V39 V40
##   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
1   1
```

```
## V41 V42 V43 V44 V45 V46 V47 V48 V49
##   1   1   1   1   1   2   1   1   3
```

agrupaciones en el mapa: clusterización de nodos

```
plot(som,type="mapping",bgcol=c("steelblue1","sienna1","yellowgreen")[groupes
],
      shape = "straight")
add.cluster.boundaries(som,clustering=groupes)
```

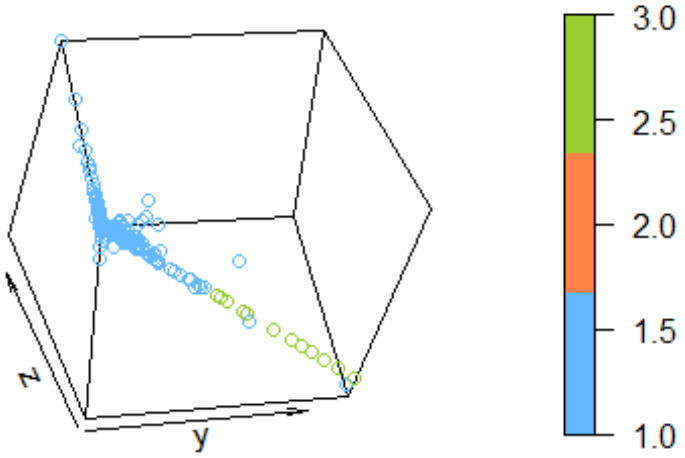


Asignamos a cada registro su clúster

```
ind.groupe <- groupes[som$unit.classif]
spam$groupes <- ind.groupe
```

Representación 3D

```
points3D(spam$X857, spam$X415 , spam$george   ,colvar=ind.groupe,
          col=c("steelblue1","sienna1","yellowgreen"),
          phi=45,theta=85)
```



Las técnicas de clustering que usamos *k-means* y *jerarquización* arrojan resultados parecidos a los mapas (SOM). La ventaja de los mapas es que los resultados son visuales y el análisis resulta más intuitivo. Además, los SOM conservan tu topología inicial, esto es ; que aquellos datos que son próximos en el espacio multidimensional, se mantendrán próximps en el bidimensional, por lo que hace a los SOM más precisos.