Memoria de la práctica 1

Adrián de la Torre Rodríguez

Índice

Lectura de datos

Para leer los datos vamos a usar la librería `tidyverse` y para observar algunas características de los datos vamos a usar la librería `funModeling`. Podemos observar que tenemos un conjunto de datos con 200000 filas y 202 variables todas de tipo numérico salvo el ID de la fila.

En cuanto a la calidad del conjunto de datos; el porcentaje de ceros de cada variable es nulo o casi nulo al igual que los valores perdidos. Por lo tanto no habrá que realizar mucha limpieza de datos.

```
library(tidyverse)
```

```
## -- Attaching packages ------------------------------------------------ tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0       v purrr   0.3.1
## v tibble  2.0.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
```

```
## -- Conflicts --------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(funModeling)
```

```
## Loading required package: Hmisc
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## 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
```

```
## funModeling v.1.7 :)
## Examples and tutorials at livebook.datascienceheroes.com
```

```
data_raw <- read_csv('train_ok.csv')
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   ID_code = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
dim(data_raw)
```

```
## [1] 200000    202
```

```
df_status(data_raw)
```

```
##     variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
## 1    ID_code       0    0.00    0 0.00     0     0 character 200000
## 2     target  181989   90.99    0 0.00     0     0   numeric      2
## 3      var_0       0    0.00   17 0.01     0     0   numeric  94670
## 4      var_1       1    0.00   11 0.01     0     0   numeric 108931
## 5      var_2       0    0.00   19 0.01     0     0   numeric  86554
## 6      var_3       0    0.00   16 0.01     0     0   numeric  74593
## 7      var_4       0    0.00   22 0.01     0     0   numeric  63513
## 8      var_5       3    0.00   21 0.01     0     0   numeric 141017
## 9      var_6       0    0.00   21 0.01     0     0   numeric  38599
## 10     var_7       0    0.00   20 0.01     0     0   numeric 103059
## 11     var_8       1    0.00   17 0.01     0     0   numeric  98615
## 12     var_9       0    0.00   19 0.01     0     0   numeric  49417
## 13    var_10       1    0.00   12 0.01     0     0   numeric 128760
## 14    var_11       2    0.00   24 0.01     0     0   numeric 130183
## 15    var_12       0    0.00   26 0.01     0     0   numeric   9560
## 16    var_13       0    0.00   20 0.01     0     0   numeric 115175
## 17    var_14       0    0.00   20 0.01     0     0   numeric  79119
## 18    var_15       0    0.00   20 0.01     0     0   numeric  19810
## 19    var_16       0    0.00   21 0.01     0     0   numeric  86914
## 20    var_17       2    0.00   18 0.01     0     0   numeric 137819
## 21    var_18       0    0.00   18 0.01     0     0   numeric 139507
## 22    var_19       0    0.00   24 0.01     0     0   numeric 144171
## 23    var_20       0    0.00   17 0.01     0     0   numeric 127758
## 24    var_21       0    0.00   21 0.01     0     0   numeric 140057
## 25    var_22       2    0.00   28 0.01     0     0   numeric  90657
## 26    var_23       0    0.00   22 0.01     0     0   numeric  24913
## 27    var_24       0    0.00   14 0.01     0     0   numeric 105096
## 28    var_25       0    0.00   27 0.01     0     0   numeric  14852
## 29    var_26       1    0.00   24 0.01     0     0   numeric 127082
## 30    var_27       5    0.00   13 0.01     0     0   numeric  60184
## 31    var_28       0    0.00   21 0.01     0     0   numeric  35858
## 32    var_29       0    0.00   25 0.01     0     0   numeric  88335
## 33    var_30       2    0.00   21 0.01     0     0   numeric 145968
## 34    var_31       0    0.00   16 0.01     0     0   numeric  77387
## 35    var_32       2    0.00   19 0.01     0     0   numeric  85964
```

```
## 36    var_33       0    0.00   31 0.02     0     0   numeric 112234
## 37    var_34       0    0.00   24 0.01     0     0   numeric  25163
## 38    var_35       2    0.00   20 0.01     0     0   numeric 122372
## 39    var_36       1    0.00   17 0.01     0     0   numeric  96400
## 40    var_37       0    0.00   17 0.01     0     0   numeric  79037
## 41    var_38       0    0.00   27 0.01     0     0   numeric 115355
## 42    var_39       2    0.00   17 0.01     0     0   numeric 112672
## 43    var_40       0    0.00   23 0.01     0     0   numeric 141868
## 44    var_41       0    0.00   20 0.01     0     0   numeric 131889
## 45    var_42       0    0.00   12 0.01     0     0   numeric  31592
## 46    var_43       0    0.00   21 0.01     0     0   numeric  15188
## 47    var_44       0    0.00   16 0.01     0     0   numeric 127695
## 48    var_45       0    0.00   17 0.01     0     0   numeric 169958
## 49    var_46       0    0.00   24 0.01     0     0   numeric  93444
## 50    var_47       0    0.00   28 0.01     0     0   numeric 154763
## 51    var_48       0    0.00   23 0.01     0     0   numeric 152027
## 52    var_49       0    0.00   14 0.01     0     0   numeric 140632
## 53    var_50       0    0.00   21 0.01     0     0   numeric  32307
## 54    var_51       0    0.00   27 0.01     0     0   numeric 143437
## 55    var_52       2    0.00   13 0.01     0     0   numeric 121309
## 56    var_53       0    0.00   27 0.01     0     0   numeric  33459
## 57    var_54       0    0.00   23 0.01     0     0   numeric 144764
## 58    var_55       0    0.00   23 0.01     0     0   numeric 128070
## 59    var_56       0    0.00   33 0.02     0     0   numeric 103035
## 60    var_57       0    0.00   19 0.01     0     0   numeric  35543
## 61    var_58       4    0.00   30 0.02     0     0   numeric 113901
## 62    var_59       0    0.00   17 0.01     0     0   numeric  37742
## 63    var_60       0    0.00   20 0.01     0     0   numeric 113755
## 64    var_61       1    0.00   22 0.01     0     0   numeric 159354
## 65    var_62       3    0.00   21 0.01     0     0   numeric  74776
## 66    var_63       1    0.00   23 0.01     0     0   numeric  97092
## 67    var_64       0    0.00   23 0.01     0     0   numeric  59378
## 68    var_65       0    0.00   18 0.01     0     0   numeric 108345
## 69    var_66       0    0.00   22 0.01     0     0   numeric  47720
## 70    var_67       0    0.00   14 0.01     0     0   numeric 137244
## 71    var_68       0    0.00   14 0.01     0     0   numeric    451
## 72    var_69       2    0.00   18 0.01     0     0   numeric 110338
## 73    var_70       0    0.00   26 0.01     0     0   numeric 153176
## 74    var_71       1    0.00   22 0.01     0     0   numeric  13527
## 75    var_72       2    0.00   15 0.01     0     0   numeric 110110
## 76    var_73       0    0.00   13 0.01     0     0   numeric 142575
## 77    var_74       0    0.00   16 0.01     0     0   numeric 161047
## 78    var_75       0    0.00   27 0.01     0     0   numeric 129372
## 79    var_76       2    0.00   24 0.01     0     0   numeric 139307
## 80    var_77       0    0.00   16 0.01     0     0   numeric 106806
## 81    var_78       0    0.00   19 0.01     0     0   numeric  72254
## 82    var_79       0    0.00   22 0.01     0     0   numeric  53212
## 83    var_80       0    0.00   17 0.01     0     0   numeric 136425
## 84    var_81       0    0.00   24 0.01     0     0   numeric  79063
## 85    var_82       0    0.00   14 0.01     0     0   numeric 144819
## 86    var_83       1    0.00   24 0.01     0     0   numeric 144271
## 87    var_84       1    0.00    9 0.00     0     0   numeric 133761
## 88    var_85       0    0.00   30 0.02     0     0   numeric 108424
## 89    var_86       1    0.00   21 0.01     0     0   numeric 140584
```

```
## 90    var_87       0    0.00   26 0.01     0     0   numeric 125285
## 91    var_88       0    0.00   23 0.01     0     0   numeric  84914
## 92    var_89       0    0.00   34 0.02     0     0   numeric 103516
## 93    var_90       0    0.00   11 0.01     0     0   numeric 157205
## 94    var_91       0    0.00   26 0.01     0     0   numeric   7961
## 95    var_92       0    0.00   28 0.01     0     0   numeric 110734
## 96    var_93       0    0.00   17 0.01     0     0   numeric  26706
## 97    var_94       0    0.00   24 0.01     0     0   numeric  89143
## 98    var_95      12    0.01   17 0.01     0     0   numeric  29387
## 99    var_96       0    0.00   18 0.01     0     0   numeric 148092
## 100   var_97       0    0.00   20 0.01     0     0   numeric 158726
## 101   var_98       1    0.00   17 0.01     0     0   numeric  33265
## 102   var_99       5    0.00   21 0.01     0     0   numeric  69298
## 103  var_100       1    0.00   15 0.01     0     0   numeric 150719
## 104  var_101       0    0.00   14 0.01     0     0   numeric 122290
## 105  var_102       0    0.00   29 0.01     0     0   numeric 146224
## 106  var_103       0    0.00   18 0.01     0     0   numeric   9376
## 107  var_104       0    0.00   30 0.02     0     0   numeric  72624
## 108  var_105       0    0.00   16 0.01     0     0   numeric  39115
## 109  var_106       0    0.00   16 0.01     0     0   numeric  71061
## 110  var_107       0    0.00   16 0.01     0     0   numeric 137818
## 111  var_108       0    0.00   21 0.01     0     0   numeric   8525
## 112  var_109       0    0.00   20 0.01     0     0   numeric 112169
## 113  var_110       0    0.00   22 0.01     0     0   numeric 106113
## 114  var_111       0    0.00   19 0.01     0     0   numeric  46464
## 115  var_112       0    0.00   21 0.01     0     0   numeric  60480
## 116  var_113       0    0.00   19 0.01     0     0   numeric 116489
## 117  var_114       0    0.00   18 0.01     0     0   numeric  43083
## 118  var_115       1    0.00   22 0.01     0     0   numeric  86726
## 119  var_116       2    0.00   17 0.01     0     0   numeric  63466
## 120  var_117       0    0.00   24 0.01     0     0   numeric 164449
## 121  var_118       0    0.00   19 0.01     0     0   numeric 143658
## 122  var_119       1    0.00   22 0.01     0     0   numeric 112395
## 123  var_120       0    0.00   19 0.01     0     0   numeric 158256
## 124  var_121       0    0.00   24 0.01     0     0   numeric  64694
## 125  var_122       2    0.00   18 0.01     0     0   numeric 121764
## 126  var_123       0    0.00   22 0.01     0     0   numeric 129883
## 127  var_124       1    0.00   19 0.01     0     0   numeric  91018
## 128  var_125       0    0.00   14 0.01     0     0   numeric  16059
## 129  var_126       0    0.00   16 0.01     0     0   numeric  32410
## 130  var_127       2    0.00   17 0.01     0     0   numeric  95706
## 131  var_128       1    0.00   17 0.01     0     0   numeric  98197
## 132  var_129       0    0.00   21 0.01     0     0   numeric 113418
## 133  var_130       0    0.00   22 0.01     0     0   numeric  36638
## 134  var_131      15    0.01   22 0.01     0     0   numeric  21464
## 135  var_132       0    0.00   18 0.01     0     0   numeric  57922
## 136  var_133       0    0.00   24 0.01     0     0   numeric  19235
## 137  var_134       3    0.00   22 0.01     0     0   numeric 131610
## 138  var_135       0    0.00   14 0.01     0     0   numeric 140771
## 139  var_136       0    0.00   23 0.01     0     0   numeric 156598
## 140  var_137       0    0.00   16 0.01     0     0   numeric 144389
## 141  var_138       4    0.00   18 0.01     0     0   numeric 117417
## 142  var_139       0    0.00   24 0.01     0     0   numeric 137287
## 143  var_140       0    0.00   24 0.01     0     0   numeric 121377
```

```
## 144  var_141       2    0.00   18 0.01     0     0   numeric 134437
## 145  var_142       0    0.00   19 0.01     0     0   numeric 128601
## 146  var_143       0    0.00   20 0.01     0     0   numeric  94370
## 147  var_144       0    0.00   24 0.01     0     0   numeric  40595
## 148  var_145       1    0.00   18 0.01     0     0   numeric 108520
## 149  var_146       0    0.00   14 0.01     0     0   numeric  84312
## 150  var_147       3    0.00   15 0.01     0     0   numeric 137551
## 151  var_148       0    0.00   21 0.01     0     0   numeric  10608
## 152  var_149       0    0.00   15 0.01     0     0   numeric 148495
## 153  var_150       0    0.00   15 0.01     0     0   numeric  83658
## 154  var_151       0    0.00   23 0.01     0     0   numeric 109660
## 155  var_152       0    0.00   24 0.01     0     0   numeric  95821
## 156  var_153       0    0.00   14 0.01     0     0   numeric  73727
## 157  var_154       1    0.00   13 0.01     0     0   numeric 119338
## 158  var_155       2    0.00   26 0.01     0     0   numeric 127451
## 159  var_156       0    0.00   11 0.01     0     0   numeric  40634
## 160  var_157       2    0.00   14 0.01     0     0   numeric 126529
## 161  var_158       0    0.00   23 0.01     0     0   numeric 144545
## 162  var_159       0    0.00   20 0.01     0     0   numeric 112826
## 163  var_160       0    0.00   21 0.01     0     0   numeric 156262
## 164  var_161       0    0.00   16 0.01     0     0   numeric  11071
## 165  var_162       0    0.00   18 0.01     0     0   numeric  57395
## 166  var_163       0    0.00   22 0.01     0     0   numeric 123160
## 167  var_164       0    0.00   18 0.01     0     0   numeric 122737
## 168  var_165       0    0.00   20 0.01     0     0   numeric 119396
## 169  var_166       0    0.00   24 0.01     0     0   numeric  17902
## 170  var_167       2    0.00   24 0.01     0     0   numeric 140940
## 171  var_168       0    0.00   17 0.01     0     0   numeric  97225
## 172  var_169       0    0.00   18 0.01     0     0   numeric  18242
## 173  var_170       5    0.00   23 0.01     0     0   numeric 113712
## 174  var_171       1    0.00   16 0.01     0     0   numeric 125911
## 175  var_172       0    0.00   20 0.01     0     0   numeric 143355
## 176  var_173       1    0.00   21 0.01     0     0   numeric 128108
## 177  var_174       0    0.00   19 0.01     0     0   numeric 134936
## 178  var_175       0    0.00   22 0.01     0     0   numeric  92654
## 179  var_176       0    0.00   18 0.01     0     0   numeric 142515
## 180  var_177       0    0.00   11 0.01     0     0   numeric  85718
## 181  var_178       2    0.00   21 0.01     0     0   numeric 145221
## 182  var_179       5    0.00   26 0.01     0     0   numeric  90086
## 183  var_180       1    0.00   14 0.01     0     0   numeric 123473
## 184  var_181       0    0.00   25 0.01     0     0   numeric  56162
## 185  var_182       2    0.00   22 0.01     0     0   numeric 149181
## 186  var_183       0    0.00   21 0.01     0     0   numeric 117523
## 187  var_184       2    0.00   22 0.01     0     0   numeric 145173
## 188  var_185       0    0.00   15 0.01     0     0   numeric 120744
## 189  var_186       0    0.00   19 0.01     0     0   numeric  98058
## 190  var_187       1    0.00   15 0.01     0     0   numeric 157022
## 191  var_188       0    0.00   22 0.01     0     0   numeric 108803
## 192  var_189       3    0.00   23 0.01     0     0   numeric  41762
## 193  var_190       1    0.00   19 0.01     0     0   numeric 114954
## 194  var_191       0    0.00   24 0.01     0     0   numeric  94261
## 195  var_192       3    0.00   21 0.01     0     0   numeric  59064
## 196  var_193       0    0.00   23 0.01     0     0   numeric 110554
## 197  var_194       0    0.00   16 0.01     0     0   numeric  97065
```

```
## 198  var_195       2    0.00   19 0.01     0     0   numeric  57867
## 199  var_196       1    0.00   21 0.01     0     0   numeric 125555
## 200  var_197       0    0.00   14 0.01     0     0   numeric  40536
## 201  var_198       0    0.00   16 0.01     0     0   numeric  94150
## 202  var_199       3    0.00   21 0.01     0     0   numeric 149422
```

Limpieza de datos

Como los datos perdidos representan un porcentaje mínimo, se va a proceder a eliminar las filas que los contienen. Se ha optado por este método ya que reemplazarlos por la media usando el método *mice* de la librería `mice` tardaba demasiado tiempo.

Al eliminar las filas con valores perdidos observamos que tanto en el conjunto de datos de filas con valores perdidos y el conjunto de datos limpio, la proporción de la variable *target* es similiar. Al tener la misma proporción, no surge ningún problema de desbalanceo de clases.

Observamos que ahora ninguna variable contiene datos perdidos y que se han eliminado en total 3959 filas lo que representa el 1.9795% de los datos originales.

```
# el método na.omit elimina las filas con valores perdidos
data_nona <- na.omit(data_raw)
# Guardamos las filas con valores perdidos para compararlas con el dataset limpio
data_dirty <- data_raw[rowSums(is.na(data_raw)) > 0,]
# Vemos la proporicion de la variable target
prop.table(table(data_dirty$target))
```

```
##
##         0         1
## 0.9123516 0.0876484
```

```
prop.table(table(data_nona$target))
```

```
##
##         0         1
## 0.9098964 0.0901036
```

```
# Estadísticas del conjunto de datos nuevo
df_status(data_nona)
```

```
##     variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
## 1    ID_code       0    0.00    0    0     0     0 character 196041
## 2     target  178377   90.99    0    0     0     0   numeric      2
## 3      var_0       0    0.00    0    0     0     0   numeric  93931
## 4      var_1       1    0.00    0    0     0     0   numeric 107855
## 5      var_2       0    0.00    0    0     0     0   numeric  85868
## 6      var_3       0    0.00    0    0     0     0   numeric  74153
## 7      var_4       0    0.00    0    0     0     0   numeric  63220
## 8      var_5       3    0.00    0    0     0     0   numeric 139057
## 9      var_6       0    0.00    0    0     0     0   numeric  38468
## 10     var_7       0    0.00    0    0     0     0   numeric 102087
## 11     var_8       1    0.00    0    0     0     0   numeric  97745
## 12     var_9       0    0.00    0    0     0     0   numeric  49260
## 13    var_10       1    0.00    0    0     0     0   numeric 127204
## 14    var_11       2    0.00    0    0     0     0   numeric 128572
## 15    var_12       0    0.00    0    0     0     0   numeric   9553
## 16    var_13       0    0.00    0    0     0     0   numeric 113943
```

```
## 17    var_14       0    0.00    0    0     0     0   numeric  78668
## 18    var_15       0    0.00    0    0     0     0   numeric  19770
## 19    var_16       0    0.00    0    0     0     0   numeric  86250
## 20    var_17       2    0.00    0    0     0     0   numeric 135996
## 21    var_18       0    0.00    0    0     0     0   numeric 137634
## 22    var_19       0    0.00    0    0     0     0   numeric 142204
## 23    var_20       0    0.00    0    0     0     0   numeric 126226
## 24    var_21       0    0.00    0    0     0     0   numeric 138190
## 25    var_22       2    0.00    0    0     0     0   numeric  89936
## 26    var_23       0    0.00    0    0     0     0   numeric  24864
## 27    var_24       0    0.00    0    0     0     0   numeric 104098
## 28    var_25       0    0.00    0    0     0     0   numeric  14830
## 29    var_26       1    0.00    0    0     0     0   numeric 125566
## 30    var_27       5    0.00    0    0     0     0   numeric  59926
## 31    var_28       0    0.00    0    0     0     0   numeric  35746
## 32    var_29       0    0.00    0    0     0     0   numeric  87643
## 33    var_30       2    0.00    0    0     0     0   numeric 143915
## 34    var_31       0    0.00    0    0     0     0   numeric  76871
## 35    var_32       1    0.00    0    0     0     0   numeric  85350
## 36    var_33       0    0.00    0    0     0     0   numeric 111118
## 37    var_34       0    0.00    0    0     0     0   numeric  25109
## 38    var_35       2    0.00    0    0     0     0   numeric 121023
## 39    var_36       1    0.00    0    0     0     0   numeric  95570
## 40    var_37       0    0.00    0    0     0     0   numeric  78499
## 41    var_38       0    0.00    0    0     0     0   numeric 114136
## 42    var_39       2    0.00    0    0     0     0   numeric 111451
## 43    var_40       0    0.00    0    0     0     0   numeric 139944
## 44    var_41       0    0.00    0    0     0     0   numeric 130245
## 45    var_42       0    0.00    0    0     0     0   numeric  31503
## 46    var_43       0    0.00    0    0     0     0   numeric  15166
## 47    var_44       0    0.00    0    0     0     0   numeric 126169
## 48    var_45       0    0.00    0    0     0     0   numeric 167139
## 49    var_46       0    0.00    0    0     0     0   numeric  92669
## 50    var_47       0    0.00    0    0     0     0   numeric 152416
## 51    var_48       0    0.00    0    0     0     0   numeric 149797
## 52    var_49       0    0.00    0    0     0     0   numeric 138803
## 53    var_50       0    0.00    0    0     0     0   numeric  32222
## 54    var_51       0    0.00    0    0     0     0   numeric 141388
## 55    var_52       2    0.00    0    0     0     0   numeric 119961
## 56    var_53       0    0.00    0    0     0     0   numeric  33370
## 57    var_54       0    0.00    0    0     0     0   numeric 142725
## 58    var_55       0    0.00    0    0     0     0   numeric 126499
## 59    var_56       0    0.00    0    0     0     0   numeric 102098
## 60    var_57       0    0.00    0    0     0     0   numeric  35431
## 61    var_58       4    0.00    0    0     0     0   numeric 112652
## 62    var_59       0    0.00    0    0     0     0   numeric  37638
## 63    var_60       0    0.00    0    0     0     0   numeric 112577
## 64    var_61       1    0.00    0    0     0     0   numeric 156898
## 65    var_62       3    0.00    0    0     0     0   numeric  74285
## 66    var_63       1    0.00    0    0     0     0   numeric  96269
## 67    var_64       0    0.00    0    0     0     0   numeric  59118
## 68    var_65       0    0.00    0    0     0     0   numeric 107312
## 69    var_66       0    0.00    0    0     0     0   numeric  47539
## 70    var_67       0    0.00    0    0     0     0   numeric 135439
```

```
## 71    var_68       0    0.00    0    0     0     0   numeric    451
## 72    var_69       2    0.00    0    0     0     0   numeric 109274
## 73    var_70       0    0.00    0    0     0     0   numeric 150880
## 74    var_71       1    0.00    0    0     0     0   numeric  13499
## 75    var_72       2    0.00    0    0     0     0   numeric 108986
## 76    var_73       0    0.00    0    0     0     0   numeric 140625
## 77    var_74       0    0.00    0    0     0     0   numeric 158494
## 78    var_75       0    0.00    0    0     0     0   numeric 127793
## 79    var_76       2    0.00    0    0     0     0   numeric 137476
## 80    var_77       0    0.00    0    0     0     0   numeric 105775
## 81    var_78       0    0.00    0    0     0     0   numeric  71787
## 82    var_79       0    0.00    0    0     0     0   numeric  52993
## 83    var_80       0    0.00    0    0     0     0   numeric 134635
## 84    var_81       0    0.00    0    0     0     0   numeric  78528
## 85    var_82       0    0.00    0    0     0     0   numeric 142796
## 86    var_83       0    0.00    0    0     0     0   numeric 142259
## 87    var_84       1    0.00    0    0     0     0   numeric 132056
## 88    var_85       0    0.00    0    0     0     0   numeric 107347
## 89    var_86       1    0.00    0    0     0     0   numeric 138698
## 90    var_87       0    0.00    0    0     0     0   numeric 123831
## 91    var_88       0    0.00    0    0     0     0   numeric  84255
## 92    var_89       0    0.00    0    0     0     0   numeric 102550
## 93    var_90       0    0.00    0    0     0     0   numeric 154833
## 94    var_91       0    0.00    0    0     0     0   numeric   7945
## 95    var_92       0    0.00    0    0     0     0   numeric 109613
## 96    var_93       0    0.00    0    0     0     0   numeric  26620
## 97    var_94       0    0.00    0    0     0     0   numeric  88481
## 98    var_95      12    0.01    0    0     0     0   numeric  29296
## 99    var_96       0    0.00    0    0     0     0   numeric 145976
## 100   var_97       0    0.00    0    0     0     0   numeric 156285
## 101   var_98       1    0.00    0    0     0     0   numeric  33158
## 102   var_99       5    0.00    0    0     0     0   numeric  68904
## 103  var_100       1    0.00    0    0     0     0   numeric 148493
## 104  var_101       0    0.00    0    0     0     0   numeric 120950
## 105  var_102       0    0.00    0    0     0     0   numeric 144143
## 106  var_103       0    0.00    0    0     0     0   numeric   9358
## 107  var_104       0    0.00    0    0     0     0   numeric  72220
## 108  var_105       0    0.00    0    0     0     0   numeric  38977
## 109  var_106       0    0.00    0    0     0     0   numeric  70612
## 110  var_107       0    0.00    0    0     0     0   numeric 135985
## 111  var_108       0    0.00    0    0     0     0   numeric   8522
## 112  var_109       0    0.00    0    0     0     0   numeric 111029
## 113  var_110       0    0.00    0    0     0     0   numeric 105050
## 114  var_111       0    0.00    0    0     0     0   numeric  46289
## 115  var_112       0    0.00    0    0     0     0   numeric  60238
## 116  var_113       0    0.00    0    0     0     0   numeric 115241
## 117  var_114       0    0.00    0    0     0     0   numeric  42929
## 118  var_115       1    0.00    0    0     0     0   numeric  86066
## 119  var_116       2    0.00    0    0     0     0   numeric  63182
## 120  var_117       0    0.00    0    0     0     0   numeric 161756
## 121  var_118       0    0.00    0    0     0     0   numeric 141665
## 122  var_119       1    0.00    0    0     0     0   numeric 111198
## 123  var_120       0    0.00    0    0     0     0   numeric 155808
## 124  var_121       0    0.00    0    0     0     0   numeric  64365
```

```
## 125  var_122       2    0.00    0    0     0     0   numeric 120417
## 126  var_123       0    0.00    0    0     0     0   numeric 128221
## 127  var_124       1    0.00    0    0     0     0   numeric  90254
## 128  var_125       0    0.00    0    0     0     0   numeric  16026
## 129  var_126       0    0.00    0    0     0     0   numeric  32319
## 130  var_127       2    0.00    0    0     0     0   numeric  94892
## 131  var_128       1    0.00    0    0     0     0   numeric  97367
## 132  var_129       0    0.00    0    0     0     0   numeric 112245
## 133  var_130       0    0.00    0    0     0     0   numeric  36529
## 134  var_131      15    0.01    0    0     0     0   numeric  21423
## 135  var_132       0    0.00    0    0     0     0   numeric  57674
## 136  var_133       0    0.00    0    0     0     0   numeric  19190
## 137  var_134       3    0.00    0    0     0     0   numeric 129989
## 138  var_135       0    0.00    0    0     0     0   numeric 138916
## 139  var_136       0    0.00    0    0     0     0   numeric 154242
## 140  var_137       0    0.00    0    0     0     0   numeric 142393
## 141  var_138       4    0.00    0    0     0     0   numeric 116133
## 142  var_139       0    0.00    0    0     0     0   numeric 135467
## 143  var_140       0    0.00    0    0     0     0   numeric 119951
## 144  var_141       2    0.00    0    0     0     0   numeric 132719
## 145  var_142       0    0.00    0    0     0     0   numeric 126969
## 146  var_143       0    0.00    0    0     0     0   numeric  93562
## 147  var_144       0    0.00    0    0     0     0   numeric  40469
## 148  var_145       1    0.00    0    0     0     0   numeric 107410
## 149  var_146       0    0.00    0    0     0     0   numeric  83698
## 150  var_147       3    0.00    0    0     0     0   numeric 135731
## 151  var_148       0    0.00    0    0     0     0   numeric  10586
## 152  var_149       0    0.00    0    0     0     0   numeric 146339
## 153  var_150       0    0.00    0    0     0     0   numeric  83066
## 154  var_151       0    0.00    0    0     0     0   numeric 108631
## 155  var_152       0    0.00    0    0     0     0   numeric  94991
## 156  var_153       0    0.00    0    0     0     0   numeric  73278
## 157  var_154       1    0.00    0    0     0     0   numeric 118016
## 158  var_155       2    0.00    0    0     0     0   numeric 125902
## 159  var_156       0    0.00    0    0     0     0   numeric  40514
## 160  var_157       2    0.00    0    0     0     0   numeric 125005
## 161  var_158       0    0.00    0    0     0     0   numeric 142566
## 162  var_159       0    0.00    0    0     0     0   numeric 111587
## 163  var_160       0    0.00    0    0     0     0   numeric 153879
## 164  var_161       0    0.00    0    0     0     0   numeric  11051
## 165  var_162       0    0.00    0    0     0     0   numeric  57104
## 166  var_163       0    0.00    0    0     0     0   numeric 121683
## 167  var_164       0    0.00    0    0     0     0   numeric 121297
## 168  var_165       0    0.00    0    0     0     0   numeric 118080
## 169  var_166       0    0.00    0    0     0     0   numeric  17873
## 170  var_167       2    0.00    0    0     0     0   numeric 139065
## 171  var_168       0    0.00    0    0     0     0   numeric  96372
## 172  var_169       0    0.00    0    0     0     0   numeric  18203
## 173  var_170       4    0.00    0    0     0     0   numeric 112515
## 174  var_171       1    0.00    0    0     0     0   numeric 124380
## 175  var_172       0    0.00    0    0     0     0   numeric 141389
## 176  var_173       1    0.00    0    0     0     0   numeric 126541
## 177  var_174       0    0.00    0    0     0     0   numeric 133205
## 178  var_175       0    0.00    0    0     0     0   numeric  91911
```

```
## 179  var_176       0    0.00    0    0     0     0   numeric 140591
## 180  var_177       0    0.00    0    0     0     0   numeric  85117
## 181  var_178       2    0.00    0    0     0     0   numeric 143183
## 182  var_179       5    0.00    0    0     0     0   numeric  89350
## 183  var_180       1    0.00    0    0     0     0   numeric 122019
## 184  var_181       0    0.00    0    0     0     0   numeric  55916
## 185  var_182       2    0.00    0    0     0     0   numeric 146953
## 186  var_183       0    0.00    0    0     0     0   numeric 116214
## 187  var_184       2    0.00    0    0     0     0   numeric 143148
## 188  var_185       0    0.00    0    0     0     0   numeric 119400
## 189  var_186       0    0.00    0    0     0     0   numeric  97209
## 190  var_187       1    0.00    0    0     0     0   numeric 154579
## 191  var_188       0    0.00    0    0     0     0   numeric 107685
## 192  var_189       3    0.00    0    0     0     0   numeric  41648
## 193  var_190       1    0.00    0    0     0     0   numeric 113668
## 194  var_191       0    0.00    0    0     0     0   numeric  93472
## 195  var_192       3    0.00    0    0     0     0   numeric  58794
## 196  var_193       0    0.00    0    0     0     0   numeric 109459
## 197  var_194       0    0.00    0    0     0     0   numeric  96227
## 198  var_195       2    0.00    0    0     0     0   numeric  57585
## 199  var_196       0    0.00    0    0     0     0   numeric 124068
## 200  var_197       0    0.00    0    0     0     0   numeric  40403
## 201  var_198       0    0.00    0    0     0     0   numeric  93383
## 202  var_199       3    0.00    0    0     0     0   numeric 147249
```

```
# Calculamos el número de filas eliminadas
deleted_rows <- dim(data_raw) - dim(data_nona)
deleted_rows
```

```
## [1] 3959    0
```

```
# Calculamos el porcentaje
deleted_rows * 100 / dim(data_raw)
```

```
## [1] 1.9795 0.0000
```

Una vez tratadas las filas con valores perdidos vamos a eliminar las columnas cuya diversidad en sus valores sea muy alta o muy baja. Finalmente de 202 variables nos quedamos con 171.

```
data_clean <- data_nona

status <- df_status(data_clean)
```

```
##     variable q_zeros p_zeros q_na p_na q_inf p_inf      type unique
## 1    ID_code       0    0.00    0    0     0     0 character 196041
## 2     target  178377   90.99    0    0     0     0   numeric      2
## 3      var_0       0    0.00    0    0     0     0   numeric  93931
## 4      var_1       1    0.00    0    0     0     0   numeric 107855
## 5      var_2       0    0.00    0    0     0     0   numeric  85868
## 6      var_3       0    0.00    0    0     0     0   numeric  74153
## 7      var_4       0    0.00    0    0     0     0   numeric  63220
## 8      var_5       3    0.00    0    0     0     0   numeric 139057
## 9      var_6       0    0.00    0    0     0     0   numeric  38468
## 10     var_7       0    0.00    0    0     0     0   numeric 102087
## 11     var_8       1    0.00    0    0     0     0   numeric  97745
## 12     var_9       0    0.00    0    0     0     0   numeric  49260
## 13    var_10       1    0.00    0    0     0     0   numeric 127204
```

```
## 14    var_11       2    0.00    0    0     0     0   numeric 128572
## 15    var_12       0    0.00    0    0     0     0   numeric   9553
## 16    var_13       0    0.00    0    0     0     0   numeric 113943
## 17    var_14       0    0.00    0    0     0     0   numeric  78668
## 18    var_15       0    0.00    0    0     0     0   numeric  19770
## 19    var_16       0    0.00    0    0     0     0   numeric  86250
## 20    var_17       2    0.00    0    0     0     0   numeric 135996
## 21    var_18       0    0.00    0    0     0     0   numeric 137634
## 22    var_19       0    0.00    0    0     0     0   numeric 142204
## 23    var_20       0    0.00    0    0     0     0   numeric 126226
## 24    var_21       0    0.00    0    0     0     0   numeric 138190
## 25    var_22       2    0.00    0    0     0     0   numeric  89936
## 26    var_23       0    0.00    0    0     0     0   numeric  24864
## 27    var_24       0    0.00    0    0     0     0   numeric 104098
## 28    var_25       0    0.00    0    0     0     0   numeric  14830
## 29    var_26       1    0.00    0    0     0     0   numeric 125566
## 30    var_27       5    0.00    0    0     0     0   numeric  59926
## 31    var_28       0    0.00    0    0     0     0   numeric  35746
## 32    var_29       0    0.00    0    0     0     0   numeric  87643
## 33    var_30       2    0.00    0    0     0     0   numeric 143915
## 34    var_31       0    0.00    0    0     0     0   numeric  76871
## 35    var_32       1    0.00    0    0     0     0   numeric  85350
## 36    var_33       0    0.00    0    0     0     0   numeric 111118
## 37    var_34       0    0.00    0    0     0     0   numeric  25109
## 38    var_35       2    0.00    0    0     0     0   numeric 121023
## 39    var_36       1    0.00    0    0     0     0   numeric  95570
## 40    var_37       0    0.00    0    0     0     0   numeric  78499
## 41    var_38       0    0.00    0    0     0     0   numeric 114136
## 42    var_39       2    0.00    0    0     0     0   numeric 111451
## 43    var_40       0    0.00    0    0     0     0   numeric 139944
## 44    var_41       0    0.00    0    0     0     0   numeric 130245
## 45    var_42       0    0.00    0    0     0     0   numeric  31503
## 46    var_43       0    0.00    0    0     0     0   numeric  15166
## 47    var_44       0    0.00    0    0     0     0   numeric 126169
## 48    var_45       0    0.00    0    0     0     0   numeric 167139
## 49    var_46       0    0.00    0    0     0     0   numeric  92669
## 50    var_47       0    0.00    0    0     0     0   numeric 152416
## 51    var_48       0    0.00    0    0     0     0   numeric 149797
## 52    var_49       0    0.00    0    0     0     0   numeric 138803
## 53    var_50       0    0.00    0    0     0     0   numeric  32222
## 54    var_51       0    0.00    0    0     0     0   numeric 141388
## 55    var_52       2    0.00    0    0     0     0   numeric 119961
## 56    var_53       0    0.00    0    0     0     0   numeric  33370
## 57    var_54       0    0.00    0    0     0     0   numeric 142725
## 58    var_55       0    0.00    0    0     0     0   numeric 126499
## 59    var_56       0    0.00    0    0     0     0   numeric 102098
## 60    var_57       0    0.00    0    0     0     0   numeric  35431
## 61    var_58       4    0.00    0    0     0     0   numeric 112652
## 62    var_59       0    0.00    0    0     0     0   numeric  37638
## 63    var_60       0    0.00    0    0     0     0   numeric 112577
## 64    var_61       1    0.00    0    0     0     0   numeric 156898
## 65    var_62       3    0.00    0    0     0     0   numeric  74285
## 66    var_63       1    0.00    0    0     0     0   numeric  96269
## 67    var_64       0    0.00    0    0     0     0   numeric  59118
```

```
## 68    var_65       0    0.00    0    0     0     0   numeric 107312
## 69    var_66       0    0.00    0    0     0     0   numeric  47539
## 70    var_67       0    0.00    0    0     0     0   numeric 135439
## 71    var_68       0    0.00    0    0     0     0   numeric    451
## 72    var_69       2    0.00    0    0     0     0   numeric 109274
## 73    var_70       0    0.00    0    0     0     0   numeric 150880
## 74    var_71       1    0.00    0    0     0     0   numeric  13499
## 75    var_72       2    0.00    0    0     0     0   numeric 108986
## 76    var_73       0    0.00    0    0     0     0   numeric 140625
## 77    var_74       0    0.00    0    0     0     0   numeric 158494
## 78    var_75       0    0.00    0    0     0     0   numeric 127793
## 79    var_76       2    0.00    0    0     0     0   numeric 137476
## 80    var_77       0    0.00    0    0     0     0   numeric 105775
## 81    var_78       0    0.00    0    0     0     0   numeric  71787
## 82    var_79       0    0.00    0    0     0     0   numeric  52993
## 83    var_80       0    0.00    0    0     0     0   numeric 134635
## 84    var_81       0    0.00    0    0     0     0   numeric  78528
## 85    var_82       0    0.00    0    0     0     0   numeric 142796
## 86    var_83       0    0.00    0    0     0     0   numeric 142259
## 87    var_84       1    0.00    0    0     0     0   numeric 132056
## 88    var_85       0    0.00    0    0     0     0   numeric 107347
## 89    var_86       1    0.00    0    0     0     0   numeric 138698
## 90    var_87       0    0.00    0    0     0     0   numeric 123831
## 91    var_88       0    0.00    0    0     0     0   numeric  84255
## 92    var_89       0    0.00    0    0     0     0   numeric 102550
## 93    var_90       0    0.00    0    0     0     0   numeric 154833
## 94    var_91       0    0.00    0    0     0     0   numeric   7945
## 95    var_92       0    0.00    0    0     0     0   numeric 109613
## 96    var_93       0    0.00    0    0     0     0   numeric  26620
## 97    var_94       0    0.00    0    0     0     0   numeric  88481
## 98    var_95      12    0.01    0    0     0     0   numeric  29296
## 99    var_96       0    0.00    0    0     0     0   numeric 145976
## 100   var_97       0    0.00    0    0     0     0   numeric 156285
## 101   var_98       1    0.00    0    0     0     0   numeric  33158
## 102   var_99       5    0.00    0    0     0     0   numeric  68904
## 103  var_100       1    0.00    0    0     0     0   numeric 148493
## 104  var_101       0    0.00    0    0     0     0   numeric 120950
## 105  var_102       0    0.00    0    0     0     0   numeric 144143
## 106  var_103       0    0.00    0    0     0     0   numeric   9358
## 107  var_104       0    0.00    0    0     0     0   numeric  72220
## 108  var_105       0    0.00    0    0     0     0   numeric  38977
## 109  var_106       0    0.00    0    0     0     0   numeric  70612
## 110  var_107       0    0.00    0    0     0     0   numeric 135985
## 111  var_108       0    0.00    0    0     0     0   numeric   8522
## 112  var_109       0    0.00    0    0     0     0   numeric 111029
## 113  var_110       0    0.00    0    0     0     0   numeric 105050
## 114  var_111       0    0.00    0    0     0     0   numeric  46289
## 115  var_112       0    0.00    0    0     0     0   numeric  60238
## 116  var_113       0    0.00    0    0     0     0   numeric 115241
## 117  var_114       0    0.00    0    0     0     0   numeric  42929
## 118  var_115       1    0.00    0    0     0     0   numeric  86066
## 119  var_116       2    0.00    0    0     0     0   numeric  63182
## 120  var_117       0    0.00    0    0     0     0   numeric 161756
## 121  var_118       0    0.00    0    0     0     0   numeric 141665
```

```
## 122  var_119       1    0.00    0    0     0     0   numeric 111198
## 123  var_120       0    0.00    0    0     0     0   numeric 155808
## 124  var_121       0    0.00    0    0     0     0   numeric  64365
## 125  var_122       2    0.00    0    0     0     0   numeric 120417
## 126  var_123       0    0.00    0    0     0     0   numeric 128221
## 127  var_124       1    0.00    0    0     0     0   numeric  90254
## 128  var_125       0    0.00    0    0     0     0   numeric  16026
## 129  var_126       0    0.00    0    0     0     0   numeric  32319
## 130  var_127       2    0.00    0    0     0     0   numeric  94892
## 131  var_128       1    0.00    0    0     0     0   numeric  97367
## 132  var_129       0    0.00    0    0     0     0   numeric 112245
## 133  var_130       0    0.00    0    0     0     0   numeric  36529
## 134  var_131      15    0.01    0    0     0     0   numeric  21423
## 135  var_132       0    0.00    0    0     0     0   numeric  57674
## 136  var_133       0    0.00    0    0     0     0   numeric  19190
## 137  var_134       3    0.00    0    0     0     0   numeric 129989
## 138  var_135       0    0.00    0    0     0     0   numeric 138916
## 139  var_136       0    0.00    0    0     0     0   numeric 154242
## 140  var_137       0    0.00    0    0     0     0   numeric 142393
## 141  var_138       4    0.00    0    0     0     0   numeric 116133
## 142  var_139       0    0.00    0    0     0     0   numeric 135467
## 143  var_140       0    0.00    0    0     0     0   numeric 119951
## 144  var_141       2    0.00    0    0     0     0   numeric 132719
## 145  var_142       0    0.00    0    0     0     0   numeric 126969
## 146  var_143       0    0.00    0    0     0     0   numeric  93562
## 147  var_144       0    0.00    0    0     0     0   numeric  40469
## 148  var_145       1    0.00    0    0     0     0   numeric 107410
## 149  var_146       0    0.00    0    0     0     0   numeric  83698
## 150  var_147       3    0.00    0    0     0     0   numeric 135731
## 151  var_148       0    0.00    0    0     0     0   numeric  10586
## 152  var_149       0    0.00    0    0     0     0   numeric 146339
## 153  var_150       0    0.00    0    0     0     0   numeric  83066
## 154  var_151       0    0.00    0    0     0     0   numeric 108631
## 155  var_152       0    0.00    0    0     0     0   numeric  94991
## 156  var_153       0    0.00    0    0     0     0   numeric  73278
## 157  var_154       1    0.00    0    0     0     0   numeric 118016
## 158  var_155       2    0.00    0    0     0     0   numeric 125902
## 159  var_156       0    0.00    0    0     0     0   numeric  40514
## 160  var_157       2    0.00    0    0     0     0   numeric 125005
## 161  var_158       0    0.00    0    0     0     0   numeric 142566
## 162  var_159       0    0.00    0    0     0     0   numeric 111587
## 163  var_160       0    0.00    0    0     0     0   numeric 153879
## 164  var_161       0    0.00    0    0     0     0   numeric  11051
## 165  var_162       0    0.00    0    0     0     0   numeric  57104
## 166  var_163       0    0.00    0    0     0     0   numeric 121683
## 167  var_164       0    0.00    0    0     0     0   numeric 121297
## 168  var_165       0    0.00    0    0     0     0   numeric 118080
## 169  var_166       0    0.00    0    0     0     0   numeric  17873
## 170  var_167       2    0.00    0    0     0     0   numeric 139065
## 171  var_168       0    0.00    0    0     0     0   numeric  96372
## 172  var_169       0    0.00    0    0     0     0   numeric  18203
## 173  var_170       4    0.00    0    0     0     0   numeric 112515
## 174  var_171       1    0.00    0    0     0     0   numeric 124380
## 175  var_172       0    0.00    0    0     0     0   numeric 141389
```

```
## 176  var_173       1    0.00    0    0     0     0   numeric 126541
## 177  var_174       0    0.00    0    0     0     0   numeric 133205
## 178  var_175       0    0.00    0    0     0     0   numeric  91911
## 179  var_176       0    0.00    0    0     0     0   numeric 140591
## 180  var_177       0    0.00    0    0     0     0   numeric  85117
## 181  var_178       2    0.00    0    0     0     0   numeric 143183
## 182  var_179       5    0.00    0    0     0     0   numeric  89350
## 183  var_180       1    0.00    0    0     0     0   numeric 122019
## 184  var_181       0    0.00    0    0     0     0   numeric  55916
## 185  var_182       2    0.00    0    0     0     0   numeric 146953
## 186  var_183       0    0.00    0    0     0     0   numeric 116214
## 187  var_184       2    0.00    0    0     0     0   numeric 143148
## 188  var_185       0    0.00    0    0     0     0   numeric 119400
## 189  var_186       0    0.00    0    0     0     0   numeric  97209
## 190  var_187       1    0.00    0    0     0     0   numeric 154579
## 191  var_188       0    0.00    0    0     0     0   numeric 107685
## 192  var_189       3    0.00    0    0     0     0   numeric  41648
## 193  var_190       1    0.00    0    0     0     0   numeric 113668
## 194  var_191       0    0.00    0    0     0     0   numeric  93472
## 195  var_192       3    0.00    0    0     0     0   numeric  58794
## 196  var_193       0    0.00    0    0     0     0   numeric 109459
## 197  var_194       0    0.00    0    0     0     0   numeric  96227
## 198  var_195       2    0.00    0    0     0     0   numeric  57585
## 199  var_196       0    0.00    0    0     0     0   numeric 124068
## 200  var_197       0    0.00    0    0     0     0   numeric  40403
## 201  var_198       0    0.00    0    0     0     0   numeric  93383
## 202  var_199       3    0.00    0    0     0     0   numeric 147249
```

```r
# Calculamos las columnas con valores diferentes
dif_cols <- status %>%
  filter(unique > 0.9 * nrow(data_clean)) %>%
  select(variable)

# Calculamos las columnas con valores iguales
eq_cols <- status %>%
  filter(unique < 0.2 * nrow(data_clean)) %>%
  select(variable)

# Eliminamos las columnas
remove_cols <- bind_rows(
  list(eq_cols, dif_cols)
)
data_clean <- data_clean %>%
  select(-one_of(remove_cols$variable))

data_clean$ID_code = data_nona$ID_code
data_clean$target = data_nona $ target
```

Tratamiento de Outliers

En este apartado vamos a tratar con los Outliers, los outliers son elementos que se consideran muy separados de la mayoría de valores de su variable. Se considera outlier si el valor excede 1.5*IQR de la variable.

Para tratar con estos datos se han calculado los cuantiles 25 y 75 para ver los datos que están por debajo

y por encima respectivamente. Y los cuantiles 5 y 95 para calcular el valor para asignar a los outliers. A continuación se calcula el rango para considerar si un valor es outlier y se le aplica a todos los outliers los valores nuevos calculados.

Se puede observar que en algunas variables se ha reducido el número de valores únicos lo que significa que los outliers han sido reemplazados por otros valores que ya aparecían en el data set, concretamente con los de los percentiles 5 y 95 de las variables.

Para comprobar el funcionamiento de este método vamos a usar graficos *boxplot* que muestran los outliers. Observamos que para *var_0* se han eliminado todos los outliers

Fuente

```
# Guardamos los datos en otra variable
data_capped <- data_clean[c(-170,-171)]
# Mostramos el estado del data set antes del proceso
df_status(data_capped)
```

```
##     variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
## 1      var_0       0       0    0    0     0     0 numeric  93931
## 2      var_1       1       0    0    0     0     0 numeric 107855
## 3      var_2       0       0    0    0     0     0 numeric  85868
## 4      var_3       0       0    0    0     0     0 numeric  74153
## 5      var_4       0       0    0    0     0     0 numeric  63220
## 6      var_5       3       0    0    0     0     0 numeric 139057
## 7      var_7       0       0    0    0     0     0 numeric 102087
## 8      var_8       1       0    0    0     0     0 numeric  97745
## 9      var_9       0       0    0    0     0     0 numeric  49260
## 10    var_10       1       0    0    0     0     0 numeric 127204
## 11    var_11       2       0    0    0     0     0 numeric 128572
## 12    var_13       0       0    0    0     0     0 numeric 113943
## 13    var_14       0       0    0    0     0     0 numeric  78668
## 14    var_16       0       0    0    0     0     0 numeric  86250
## 15    var_17       2       0    0    0     0     0 numeric 135996
## 16    var_18       0       0    0    0     0     0 numeric 137634
## 17    var_19       0       0    0    0     0     0 numeric 142204
## 18    var_20       0       0    0    0     0     0 numeric 126226
## 19    var_21       0       0    0    0     0     0 numeric 138190
## 20    var_22       2       0    0    0     0     0 numeric  89936
## 21    var_24       0       0    0    0     0     0 numeric 104098
## 22    var_26       1       0    0    0     0     0 numeric 125566
## 23    var_27       5       0    0    0     0     0 numeric  59926
## 24    var_29       0       0    0    0     0     0 numeric  87643
## 25    var_30       2       0    0    0     0     0 numeric 143915
## 26    var_31       0       0    0    0     0     0 numeric  76871
## 27    var_32       1       0    0    0     0     0 numeric  85350
## 28    var_33       0       0    0    0     0     0 numeric 111118
## 29    var_35       2       0    0    0     0     0 numeric 121023
## 30    var_36       1       0    0    0     0     0 numeric  95570
## 31    var_37       0       0    0    0     0     0 numeric  78499
## 32    var_38       0       0    0    0     0     0 numeric 114136
## 33    var_39       2       0    0    0     0     0 numeric 111451
## 34    var_40       0       0    0    0     0     0 numeric 139944
## 35    var_41       0       0    0    0     0     0 numeric 130245
## 36    var_44       0       0    0    0     0     0 numeric 126169
## 37    var_45       0       0    0    0     0     0 numeric 167139
```

```
## 38    var_46       0       0    0    0     0     0 numeric  92669
## 39    var_47       0       0    0    0     0     0 numeric 152416
## 40    var_48       0       0    0    0     0     0 numeric 149797
## 41    var_49       0       0    0    0     0     0 numeric 138803
## 42    var_51       0       0    0    0     0     0 numeric 141388
## 43    var_52       2       0    0    0     0     0 numeric 119961
## 44    var_54       0       0    0    0     0     0 numeric 142725
## 45    var_55       0       0    0    0     0     0 numeric 126499
## 46    var_56       0       0    0    0     0     0 numeric 102098
## 47    var_58       4       0    0    0     0     0 numeric 112652
## 48    var_60       0       0    0    0     0     0 numeric 112577
## 49    var_61       1       0    0    0     0     0 numeric 156898
## 50    var_62       3       0    0    0     0     0 numeric  74285
## 51    var_63       1       0    0    0     0     0 numeric  96269
## 52    var_64       0       0    0    0     0     0 numeric  59118
## 53    var_65       0       0    0    0     0     0 numeric 107312
## 54    var_66       0       0    0    0     0     0 numeric  47539
## 55    var_67       0       0    0    0     0     0 numeric 135439
## 56    var_69       2       0    0    0     0     0 numeric 109274
## 57    var_70       0       0    0    0     0     0 numeric 150880
## 58    var_72       2       0    0    0     0     0 numeric 108986
## 59    var_73       0       0    0    0     0     0 numeric 140625
## 60    var_74       0       0    0    0     0     0 numeric 158494
## 61    var_75       0       0    0    0     0     0 numeric 127793
## 62    var_76       2       0    0    0     0     0 numeric 137476
## 63    var_77       0       0    0    0     0     0 numeric 105775
## 64    var_78       0       0    0    0     0     0 numeric  71787
## 65    var_79       0       0    0    0     0     0 numeric  52993
## 66    var_80       0       0    0    0     0     0 numeric 134635
## 67    var_81       0       0    0    0     0     0 numeric  78528
## 68    var_82       0       0    0    0     0     0 numeric 142796
## 69    var_83       0       0    0    0     0     0 numeric 142259
## 70    var_84       1       0    0    0     0     0 numeric 132056
## 71    var_85       0       0    0    0     0     0 numeric 107347
## 72    var_86       1       0    0    0     0     0 numeric 138698
## 73    var_87       0       0    0    0     0     0 numeric 123831
## 74    var_88       0       0    0    0     0     0 numeric  84255
## 75    var_89       0       0    0    0     0     0 numeric 102550
## 76    var_90       0       0    0    0     0     0 numeric 154833
## 77    var_92       0       0    0    0     0     0 numeric 109613
## 78    var_94       0       0    0    0     0     0 numeric  88481
## 79    var_96       0       0    0    0     0     0 numeric 145976
## 80    var_97       0       0    0    0     0     0 numeric 156285
## 81    var_99       5       0    0    0     0     0 numeric  68904
## 82   var_100       1       0    0    0     0     0 numeric 148493
## 83   var_101       0       0    0    0     0     0 numeric 120950
## 84   var_102       0       0    0    0     0     0 numeric 144143
## 85   var_104       0       0    0    0     0     0 numeric  72220
## 86   var_106       0       0    0    0     0     0 numeric  70612
## 87   var_107       0       0    0    0     0     0 numeric 135985
## 88   var_109       0       0    0    0     0     0 numeric 111029
## 89   var_110       0       0    0    0     0     0 numeric 105050
## 90   var_111       0       0    0    0     0     0 numeric  46289
## 91   var_112       0       0    0    0     0     0 numeric  60238
```

```
## 92   var_113       0       0    0    0     0     0 numeric 115241
## 93   var_114       0       0    0    0     0     0 numeric  42929
## 94   var_115       1       0    0    0     0     0 numeric  86066
## 95   var_116       2       0    0    0     0     0 numeric  63182
## 96   var_117       0       0    0    0     0     0 numeric 161756
## 97   var_118       0       0    0    0     0     0 numeric 141665
## 98   var_119       1       0    0    0     0     0 numeric 111198
## 99   var_120       0       0    0    0     0     0 numeric 155808
## 100  var_121       0       0    0    0     0     0 numeric  64365
## 101  var_122       2       0    0    0     0     0 numeric 120417
## 102  var_123       0       0    0    0     0     0 numeric 128221
## 103  var_124       1       0    0    0     0     0 numeric  90254
## 104  var_127       2       0    0    0     0     0 numeric  94892
## 105  var_128       1       0    0    0     0     0 numeric  97367
## 106  var_129       0       0    0    0     0     0 numeric 112245
## 107  var_132       0       0    0    0     0     0 numeric  57674
## 108  var_134       3       0    0    0     0     0 numeric 129989
## 109  var_135       0       0    0    0     0     0 numeric 138916
## 110  var_136       0       0    0    0     0     0 numeric 154242
## 111  var_137       0       0    0    0     0     0 numeric 142393
## 112  var_138       4       0    0    0     0     0 numeric 116133
## 113  var_139       0       0    0    0     0     0 numeric 135467
## 114  var_140       0       0    0    0     0     0 numeric 119951
## 115  var_141       2       0    0    0     0     0 numeric 132719
## 116  var_142       0       0    0    0     0     0 numeric 126969
## 117  var_143       0       0    0    0     0     0 numeric  93562
## 118  var_144       0       0    0    0     0     0 numeric  40469
## 119  var_145       1       0    0    0     0     0 numeric 107410
## 120  var_146       0       0    0    0     0     0 numeric  83698
## 121  var_147       3       0    0    0     0     0 numeric 135731
## 122  var_149       0       0    0    0     0     0 numeric 146339
## 123  var_150       0       0    0    0     0     0 numeric  83066
## 124  var_151       0       0    0    0     0     0 numeric 108631
## 125  var_152       0       0    0    0     0     0 numeric  94991
## 126  var_153       0       0    0    0     0     0 numeric  73278
## 127  var_154       1       0    0    0     0     0 numeric 118016
## 128  var_155       2       0    0    0     0     0 numeric 125902
## 129  var_156       0       0    0    0     0     0 numeric  40514
## 130  var_157       2       0    0    0     0     0 numeric 125005
## 131  var_158       0       0    0    0     0     0 numeric 142566
## 132  var_159       0       0    0    0     0     0 numeric 111587
## 133  var_160       0       0    0    0     0     0 numeric 153879
## 134  var_162       0       0    0    0     0     0 numeric  57104
## 135  var_163       0       0    0    0     0     0 numeric 121683
## 136  var_164       0       0    0    0     0     0 numeric 121297
## 137  var_165       0       0    0    0     0     0 numeric 118080
## 138  var_167       2       0    0    0     0     0 numeric 139065
## 139  var_168       0       0    0    0     0     0 numeric  96372
## 140  var_170       4       0    0    0     0     0 numeric 112515
## 141  var_171       1       0    0    0     0     0 numeric 124380
## 142  var_172       0       0    0    0     0     0 numeric 141389
## 143  var_173       1       0    0    0     0     0 numeric 126541
## 144  var_174       0       0    0    0     0     0 numeric 133205
## 145  var_175       0       0    0    0     0     0 numeric  91911
```

```
## 146  var_176       0       0    0    0     0     0 numeric 140591
## 147  var_177       0       0    0    0     0     0 numeric  85117
## 148  var_178       2       0    0    0     0     0 numeric 143183
## 149  var_179       5       0    0    0     0     0 numeric  89350
## 150  var_180       1       0    0    0     0     0 numeric 122019
## 151  var_181       0       0    0    0     0     0 numeric  55916
## 152  var_182       2       0    0    0     0     0 numeric 146953
## 153  var_183       0       0    0    0     0     0 numeric 116214
## 154  var_184       2       0    0    0     0     0 numeric 143148
## 155  var_185       0       0    0    0     0     0 numeric 119400
## 156  var_186       0       0    0    0     0     0 numeric  97209
## 157  var_187       1       0    0    0     0     0 numeric 154579
## 158  var_188       0       0    0    0     0     0 numeric 107685
## 159  var_189       3       0    0    0     0     0 numeric  41648
## 160  var_190       1       0    0    0     0     0 numeric 113668
## 161  var_191       0       0    0    0     0     0 numeric  93472
## 162  var_192       3       0    0    0     0     0 numeric  58794
## 163  var_193       0       0    0    0     0     0 numeric 109459
## 164  var_194       0       0    0    0     0     0 numeric  96227
## 165  var_195       2       0    0    0     0     0 numeric  57585
## 166  var_196       0       0    0    0     0     0 numeric 124068
## 167  var_197       0       0    0    0     0     0 numeric  40403
## 168  var_198       0       0    0    0     0     0 numeric  93383
## 169  var_199       3       0    0    0     0     0 numeric 147249
```

```r
# Variable para el índe de columna
i <- 1
# Para cada columna del data set...
for (col in data_capped){
  # Calculamos los cuantiles 25 y 75
  qnt <- quantile(col, probs=c(.25, .75), na.rm = T)
  # Calculamos los cuantiles 5 y 95 que serán los nuevos valores de los outliers
  caps <- quantile(col, probs=c(.05, .95), na.rm = T)
  # H es lo máximo que puede variar un valor del percientil 25 por debajo y 75 por encima. Se calcula u
  H <- 1.5 * IQR(col, na.rm = T)
  # Todos los valores de la columna que superen ese rango H se actualizan por los valores nuevos
  col[col < (qnt[1] - H)] <- caps[1]
  col[col > (qnt[2] + H)] <- caps[2]
  # Actualizamos la columna del data set
  data_capped[i] <- col
  # Aumentamos el índice de la columna
  i <- i + 1
}
# Mostramos de nuevo el conjunto de datos
df_status(data_capped)
```

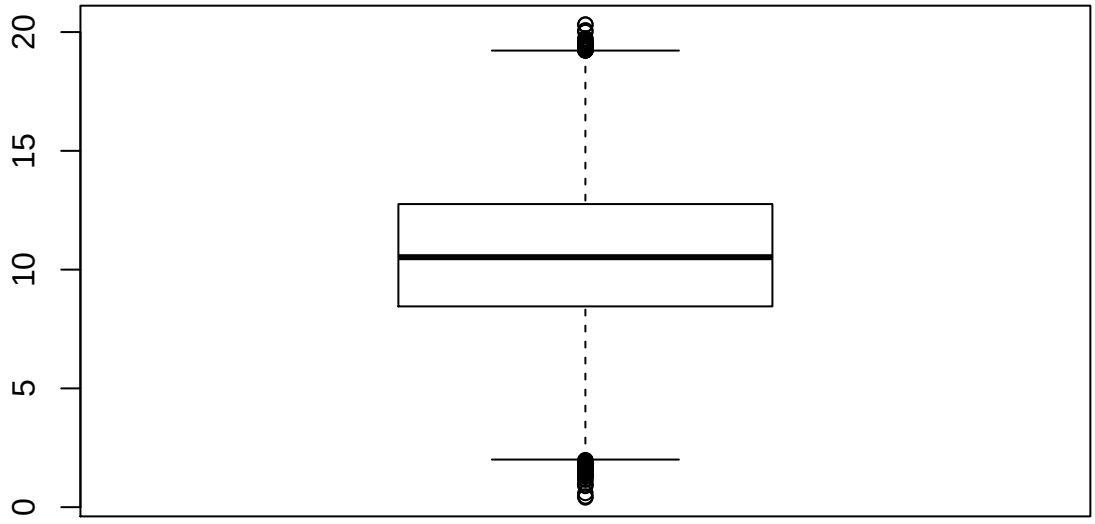
```
##     variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
## 1      var_0       0       0    0    0     0     0 numeric  93839
## 2      var_1       1       0    0    0     0     0 numeric 107849
## 3      var_2       0       0    0    0     0     0 numeric  85824
## 4      var_3       0       0    0    0     0     0 numeric  74131
## 5      var_4       0       0    0    0     0     0 numeric  63149
## 6      var_5       3       0    0    0     0     0 numeric 139053
## 7      var_7       0       0    0    0     0     0 numeric 102069
## 8      var_8       1       0    0    0     0     0 numeric  97744
```

```
## 9      var_9       0       0    0    0     0     0 numeric  49260
## 10    var_10       1       0    0    0     0     0 numeric 127078
## 11    var_11       2       0    0    0     0     0 numeric 128276
## 12    var_13       0       0    0    0     0     0 numeric 113942
## 13    var_14       0       0    0    0     0     0 numeric  78666
## 14    var_16       0       0    0    0     0     0 numeric  86183
## 15    var_17       2       0    0    0     0     0 numeric 135777
## 16    var_18       0       0    0    0     0     0 numeric 137625
## 17    var_19       0       0    0    0     0     0 numeric 142203
## 18    var_20       0       0    0    0     0     0 numeric 126225
## 19    var_21       0       0    0    0     0     0 numeric 137940
## 20    var_22       2       0    0    0     0     0 numeric  89678
## 21    var_24       0       0    0    0     0     0 numeric 104095
## 22    var_26       1       0    0    0     0     0 numeric 125392
## 23    var_27       5       0    0    0     0     0 numeric  59923
## 24    var_29       0       0    0    0     0     0 numeric  87504
## 25    var_30       2       0    0    0     0     0 numeric 143903
## 26    var_31       0       0    0    0     0     0 numeric  76290
## 27    var_32       1       0    0    0     0     0 numeric  85348
## 28    var_33       0       0    0    0     0     0 numeric 111114
## 29    var_35       2       0    0    0     0     0 numeric 120839
## 30    var_36       1       0    0    0     0     0 numeric  95402
## 31    var_37       0       0    0    0     0     0 numeric  78498
## 32    var_38       0       0    0    0     0     0 numeric 113650
## 33    var_39       2       0    0    0     0     0 numeric 111300
## 34    var_40       0       0    0    0     0     0 numeric 139899
## 35    var_41       0       0    0    0     0     0 numeric 130245
## 36    var_44       0       0    0    0     0     0 numeric 125068
## 37    var_45       0       0    0    0     0     0 numeric 167097
## 38    var_46       0       0    0    0     0     0 numeric  92413
## 39    var_47       0       0    0    0     0     0 numeric 152414
## 40    var_48       0       0    0    0     0     0 numeric 149706
## 41    var_49       0       0    0    0     0     0 numeric 138796
## 42    var_51       0       0    0    0     0     0 numeric 141349
## 43    var_52       2       0    0    0     0     0 numeric 119625
## 44    var_54       0       0    0    0     0     0 numeric 142690
## 45    var_55       0       0    0    0     0     0 numeric 126499
## 46    var_56       0       0    0    0     0     0 numeric 101875
## 47    var_58       4       0    0    0     0     0 numeric 112275
## 48    var_60       0       0    0    0     0     0 numeric 112572
## 49    var_61       1       0    0    0     0     0 numeric 156845
## 50    var_62       3       0    0    0     0     0 numeric  73862
## 51    var_63       1       0    0    0     0     0 numeric  96092
## 52    var_64       0       0    0    0     0     0 numeric  59093
## 53    var_65       0       0    0    0     0     0 numeric 107070
## 54    var_66       0       0    0    0     0     0 numeric  47398
## 55    var_67       0       0    0    0     0     0 numeric 135002
## 56    var_69       2       0    0    0     0     0 numeric 109260
## 57    var_70       0       0    0    0     0     0 numeric 150874
## 58    var_72       2       0    0    0     0     0 numeric 108863
## 59    var_73       0       0    0    0     0     0 numeric 140603
## 60    var_74       0       0    0    0     0     0 numeric 158202
## 61    var_75       0       0    0    0     0     0 numeric 127782
## 62    var_76       2       0    0    0     0     0 numeric 137267
```

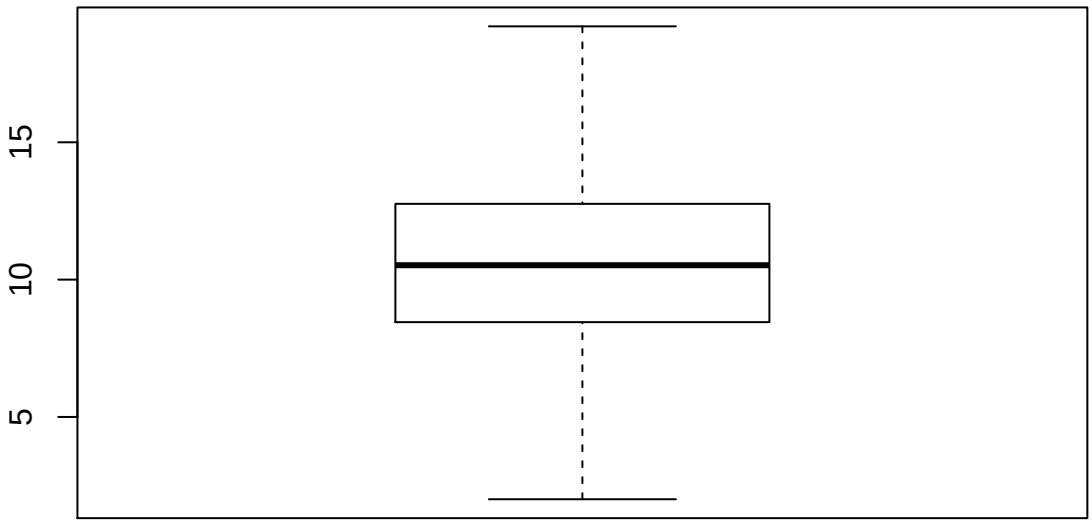
```
## 63    var_77       0       0    0    0     0     0 numeric 105775
## 64    var_78       0       0    0    0     0     0 numeric  71785
## 65    var_79       0       0    0    0     0     0 numeric  52993
## 66    var_80       0       0    0    0     0     0 numeric 134613
## 67    var_81       0       0    0    0     0     0 numeric  78376
## 68    var_82       0       0    0    0     0     0 numeric 142794
## 69    var_83       0       0    0    0     0     0 numeric 142254
## 70    var_84       1       0    0    0     0     0 numeric 132018
## 71    var_85       0       0    0    0     0     0 numeric 107343
## 72    var_86       1       0    0    0     0     0 numeric 138551
## 73    var_87       0       0    0    0     0     0 numeric 123707
## 74    var_88       0       0    0    0     0     0 numeric  84226
## 75    var_89       0       0    0    0     0     0 numeric 102472
## 76    var_90       0       0    0    0     0     0 numeric 154824
## 77    var_92       0       0    0    0     0     0 numeric 109578
## 78    var_94       0       0    0    0     0     0 numeric  88481
## 79    var_96       0       0    0    0     0     0 numeric 145970
## 80    var_97       0       0    0    0     0     0 numeric 156278
## 81    var_99       5       0    0    0     0     0 numeric  68823
## 82   var_100       1       0    0    0     0     0 numeric 148321
## 83   var_101       0       0    0    0     0     0 numeric 120950
## 84   var_102       0       0    0    0     0     0 numeric 144058
## 85   var_104       0       0    0    0     0     0 numeric  71921
## 86   var_106       0       0    0    0     0     0 numeric  70233
## 87   var_107       0       0    0    0     0     0 numeric 135903
## 88   var_109       0       0    0    0     0     0 numeric 111028
## 89   var_110       0       0    0    0     0     0 numeric 104618
## 90   var_111       0       0    0    0     0     0 numeric  46212
## 91   var_112       0       0    0    0     0     0 numeric  60225
## 92   var_113       0       0    0    0     0     0 numeric 115240
## 93   var_114       0       0    0    0     0     0 numeric  42883
## 94   var_115       1       0    0    0     0     0 numeric  85904
## 95   var_116       2       0    0    0     0     0 numeric  63168
## 96   var_117       0       0    0    0     0     0 numeric 161653
## 97   var_118       0       0    0    0     0     0 numeric 141448
## 98   var_119       1       0    0    0     0     0 numeric 111194
## 99   var_120       0       0    0    0     0     0 numeric 155571
## 100  var_121       0       0    0    0     0     0 numeric  64353
## 101  var_122       2       0    0    0     0     0 numeric 120394
## 102  var_123       0       0    0    0     0     0 numeric 128199
## 103  var_124       1       0    0    0     0     0 numeric  89777
## 104  var_127       2       0    0    0     0     0 numeric  94890
## 105  var_128       1       0    0    0     0     0 numeric  97352
## 106  var_129       0       0    0    0     0     0 numeric 112154
## 107  var_132       0       0    0    0     0     0 numeric  57636
## 108  var_134       3       0    0    0     0     0 numeric 129988
## 109  var_135       0       0    0    0     0     0 numeric 138890
## 110  var_136       0       0    0    0     0     0 numeric 154240
## 111  var_137       0       0    0    0     0     0 numeric 142393
## 112  var_138       4       0    0    0     0     0 numeric 115825
## 113  var_139       0       0    0    0     0     0 numeric 135296
## 114  var_140       0       0    0    0     0     0 numeric 119826
## 115  var_141       2       0    0    0     0     0 numeric 132718
## 116  var_142       0       0    0    0     0     0 numeric 126827
```

```
## 117  var_143       0       0    0    0     0     0 numeric  93459
## 118  var_144       0       0    0    0     0     0 numeric  40413
## 119  var_145       1       0    0    0     0     0 numeric 107397
## 120  var_146       0       0    0    0     0     0 numeric  82974
## 121  var_147       3       0    0    0     0     0 numeric 135634
## 122  var_149       0       0    0    0     0     0 numeric 145985
## 123  var_150       0       0    0    0     0     0 numeric  83051
## 124  var_151       0       0    0    0     0     0 numeric 108630
## 125  var_152       0       0    0    0     0     0 numeric  94674
## 126  var_153       0       0    0    0     0     0 numeric  73274
## 127  var_154       1       0    0    0     0     0 numeric 118011
## 128  var_155       2       0    0    0     0     0 numeric 125678
## 129  var_156       0       0    0    0     0     0 numeric  40488
## 130  var_157       2       0    0    0     0     0 numeric 124969
## 131  var_158       0       0    0    0     0     0 numeric 142549
## 132  var_159       0       0    0    0     0     0 numeric 111417
## 133  var_160       0       0    0    0     0     0 numeric 153876
## 134  var_162       0       0    0    0     0     0 numeric  56957
## 135  var_163       0       0    0    0     0     0 numeric 121547
## 136  var_164       0       0    0    0     0     0 numeric 121296
## 137  var_165       0       0    0    0     0     0 numeric 118014
## 138  var_167       2       0    0    0     0     0 numeric 138798
## 139  var_168       0       0    0    0     0     0 numeric  96103
## 140  var_170       4       0    0    0     0     0 numeric 112134
## 141  var_171       1       0    0    0     0     0 numeric 124234
## 142  var_172       0       0    0    0     0     0 numeric 141026
## 143  var_173       1       0    0    0     0     0 numeric 126422
## 144  var_174       0       0    0    0     0     0 numeric 133173
## 145  var_175       0       0    0    0     0     0 numeric  91909
## 146  var_176       0       0    0    0     0     0 numeric 140552
## 147  var_177       0       0    0    0     0     0 numeric  85114
## 148  var_178       2       0    0    0     0     0 numeric 143140
## 149  var_179       5       0    0    0     0     0 numeric  87980
## 150  var_180       1       0    0    0     0     0 numeric 121989
## 151  var_181       0       0    0    0     0     0 numeric  55852
## 152  var_182       2       0    0    0     0     0 numeric 146948
## 153  var_183       0       0    0    0     0     0 numeric 116193
## 154  var_184       2       0    0    0     0     0 numeric 143107
## 155  var_185       0       0    0    0     0     0 numeric 119176
## 156  var_186       0       0    0    0     0     0 numeric  96798
## 157  var_187       1       0    0    0     0     0 numeric 154576
## 158  var_188       0       0    0    0     0     0 numeric 107664
## 159  var_189       3       0    0    0     0     0 numeric  41634
## 160  var_190       1       0    0    0     0     0 numeric 113514
## 161  var_191       0       0    0    0     0     0 numeric  93394
## 162  var_192       3       0    0    0     0     0 numeric  58138
## 163  var_193       0       0    0    0     0     0 numeric 109022
## 164  var_194       0       0    0    0     0     0 numeric  96225
## 165  var_195       2       0    0    0     0     0 numeric  57460
## 166  var_196       0       0    0    0     0     0 numeric 124068
## 167  var_197       0       0    0    0     0     0 numeric  40356
## 168  var_198       0       0    0    0     0     0 numeric  93294
## 169  var_199       3       0    0    0     0     0 numeric 147233
```

```r
# Añadimos de nuevo el id y el target
data_capped$ID_code = data_clean$ID_code
data_capped$target = data_clean$target

# Mostramos los boxplot de var_0 antes y después del método
boxplot(data_clean$var_0)
```




```r
boxplot(data_capped$var_0)
```



Normalizar datos

En este apartado vamos a normalizar los datos, es decir transformar los rangos de sus variables al intervalo [0,1] donde cero sería el mínimo valor que toma la variable antes de ser normalizada y 1 el máximo. Hacemos esta normalización para igualar los rangos en que se mueven las variables para facilitar visualizar la relación entre ellas.

Vamos a usar la librería `BBmisc` y la función *normalize* con el método range que por defecto normaliza al rango [0,1]

```
library(BBmisc)
```

```
##
## Attaching package: 'BBmisc'
```

```
## The following object is masked from 'package:Hmisc':
##
##     %nin%
```

```
## The following objects are masked from 'package:dplyr':
##
##     coalesce, collapse
```

```
data_normalized <- normalize(data_capped, method = "range")
glimpse(data_normalized)
```

```
## Observations: 196,041
## Variables: 171
```

```
## $ var_0   <dbl> 0.4020016, 0.5515723, 0.3836356, 0.5260040, 0.4549388,...
## $ var_1   <dbl> 0.2920740, 0.4009248, 0.4587365, 0.4832331, 0.5108026,...
## $ var_2   <dbl> 0.5848318, 0.7134109, 0.5961954, 0.3899955, 0.6485380,...
## $ var_3   <dbl> 0.3607499, 0.3849098, 0.5892733, 0.5323751, 0.4868141,...
## $ var_4   <dbl> 0.5419442, 0.6369509, 0.4493930, 0.6603891, 0.6279930,...
## $ var_5   <dbl> 0.4286163, 0.7801826, 0.4329165, 0.5889806, 0.6812439,...
## $ var_7   <dbl> 0.5992294, 0.4961099, 0.4015886, 0.4168387, 0.6300154,...
## $ var_8   <dbl> 0.2517625, 0.6522510, 0.2517972, 0.2050500, 0.8070786,...
## $ var_9   <dbl> 0.2474200, 0.5730561, 0.2760407, 0.5953260, 0.5164134,...
## $ var_10  <dbl> 0.5825528, 0.4775109, 0.4799819, 0.5630144, 0.1921329,...
## $ var_11  <dbl> 0.6947765, 0.8406372, 0.2626474, 0.6836524, 0.2364119,...
## $ var_13  <dbl> 0.20359165, 0.49053586, 0.45023043, 0.62073510, 0.4713...
## $ var_14  <dbl> 0.5817530, 0.3303594, 0.4865202, 0.4067548, 0.5063215,...
## $ var_16  <dbl> 0.2570336, 0.8175593, 0.3909068, 0.6035496, 0.4470301,...
## $ var_17  <dbl> 0.4587246, 0.2427775, 0.6010350, 0.6349217, 0.5666593,...
## $ var_18  <dbl> 0.2700105, 0.3445570, 0.2788539, 0.5171970, 0.3175519,...
## $ var_19  <dbl> 0.9043722, 0.8585341, 0.6853806, 0.5409280, 0.4893561,...
## $ var_20  <dbl> 0.4012494, 0.1962028, 0.6096027, 0.6830395, 0.7154601,...
## $ var_21  <dbl> 0.4760946, 0.1838432, 0.5176955, 0.3967090, 0.5355427,...
## $ var_22  <dbl> 0.3982664, 0.7621281, 0.3151422, 0.6566117, 0.8570210,...
## $ var_24  <dbl> 0.6580954, 0.3345335, 0.2777227, 0.5598846, 0.6504446,...
## $ var_26  <dbl> 0.4714392, 0.2703749, 0.3850189, 0.6525387, 0.7242245,...
## $ var_27  <dbl> 0.5924255, 0.3522286, 0.3137868, 0.5367558, 0.2647269,...
## $ var_29  <dbl> 0.5697702, 0.1941790, 0.5799313, 0.3522218, 0.7034165,...
## $ var_30  <dbl> 0.6647587, 0.4759187, 0.4380297, 0.3511253, 0.6049005,...
## $ var_31  <dbl> 0.70951402, 0.29262489, 0.67483544, 0.50528925, 0.4245...
## $ var_32  <dbl> 0.2974424, 0.6375639, 0.4021636, 0.3833210, 0.5522980,...
## $ var_33  <dbl> 0.5811531, 0.6748435, 0.6624215, 0.1764737, 0.6730641,...
## $ var_35  <dbl> 0.7320414, 0.3972642, 0.5118941, 0.1707913, 0.3936234,...
## $ var_36  <dbl> 0.4134441, 0.5323447, 0.6386333, 0.8332391, 0.5319359,...
## $ var_37  <dbl> 0.6881411, 0.5278710, 0.6134446, 0.2932887, 0.5259105,...
## $ var_38  <dbl> 0.2530554, 0.5089447, 0.5120947, 0.4406082, 0.8621453,...
## $ var_39  <dbl> 0.6350551, 0.6266513, 0.5075399, 0.4975153, 0.5308948,...
## $ var_40  <dbl> 0.5365064, 0.4129255, 0.3623932, 0.5661461, 0.5289856,...
## $ var_41  <dbl> 0.4584934, 0.2999536, 0.3106418, 0.2160960, 0.6513917,...
## $ var_44  <dbl> 0.5786170, 0.2442288, 0.3360863, 0.4630537, 0.4263969,...
## $ var_45  <dbl> 0.54529782, 0.22857286, 0.53830774, 0.40666537, 0.0864...
## $ var_46  <dbl> 0.1703606, 0.2594829, 0.4912489, 0.5896389, 0.6588421,...
## $ var_47  <dbl> 0.4706975, 0.6959502, 0.6024435, 0.1431412, 0.6935989,...
## $ var_48  <dbl> 0.5202546, 0.7958625, 0.5910485, 0.2912401, 0.2778010,...
## $ var_49  <dbl> 0.2603832, 0.6916757, 0.2934438, 0.3251139, 0.4530091,...
## $ var_51  <dbl> 0.8243185, 0.4975628, 0.6289227, 0.4216529, 0.4409727,...
## $ var_52  <dbl> 0.5672428, 0.4453166, 0.6179802, 0.6559297, 0.5407206,...
## $ var_54  <dbl> 0.4605964, 0.3136127, 0.6618467, 0.6374181, 0.3046743,...
## $ var_55  <dbl> 0.5662348, 0.6808305, 0.3062294, 0.4921657, 0.5454286,...
## $ var_56  <dbl> 0.2387976, 0.6265373, 0.4491709, 0.3929770, 0.5140003,...
## $ var_58  <dbl> 0.4505161, 0.4464713, 0.4123416, 0.5044793, 0.4478764,...
## $ var_60  <dbl> 0.6671101, 0.3366546, 0.5220792, 0.6509026, 0.3628666,...
## $ var_61  <dbl> 0.7406356, 0.5677415, 0.5233627, 0.3157771, 0.2772083,...
## $ var_62  <dbl> 0.6115468, 0.5862931, 0.5041449, 0.7334560, 0.3817623,...
## $ var_63  <dbl> 0.4514074, 0.3914882, 0.7795153, 0.2237387, 0.3659137,...
## $ var_64  <dbl> 0.7600082, 0.6949663, 0.1947953, 0.7567709, 0.3715575,...
## $ var_65  <dbl> 0.89171160, 0.52387700, 0.40250916, 0.74335677, 0.2544...
## $ var_66  <dbl> 0.5039609, 0.4411906, 0.7459732, 0.4755135, 0.6869545,...
```

```
## $ var_67  <dbl> 0.75606518, 0.40360009, 0.45032979, 0.78480659, 0.2500...
## $ var_69  <dbl> 0.4438580, 0.7412795, 0.3974374, 0.4739855, 0.5064173,...
## $ var_70  <dbl> 0.4567849, 0.7172814, 0.1914654, 0.2146579, 0.7112307,...
## $ var_72  <dbl> 0.6990108, 0.6100835, 0.7497349, 0.5260324, 0.4393746,...
## $ var_73  <dbl> 0.24192848, 0.50042866, 0.31805029, 0.59953480, 0.1437...
## $ var_74  <dbl> 0.7941418, 0.3526320, 0.5019377, 0.3119522, 0.1986823,...
## $ var_75  <dbl> 0.53651449, 0.21783515, 0.56647041, 0.45330235, 0.6763...
## $ var_76  <dbl> 0.3095832, 0.4346266, 0.4621055, 0.6220637, 0.3846618,...
## $ var_77  <dbl> 0.6792981, 0.6823494, 0.6719632, 0.6885157, 0.4986699,...
## $ var_78  <dbl> 0.6007273, 0.5170510, 0.5841309, 0.3991156, 0.6547401,...
## $ var_79  <dbl> 0.2253027, 0.4287808, 0.3032313, 0.6205367, 0.5074805,...
## $ var_80  <dbl> 0.6760273, 0.4164498, 0.5316562, 0.3752400, 0.4261214,...
## $ var_81  <dbl> 0.42297078, 0.77121472, 0.57891444, 0.15849471, 0.3824...
## $ var_82  <dbl> 0.5977223, 0.5781394, 0.8326186, 0.6026944, 0.3894642,...
## $ var_83  <dbl> 0.5257655, 0.3665602, 0.6742989, 0.3569504, 0.3278729,...
## $ var_84  <dbl> 0.4494263, 0.4231610, 0.7214739, 0.6163337, 0.7879207,...
## $ var_85  <dbl> 0.6283043, 0.2942185, 0.5243576, 0.4764602, 0.3263995,...
## $ var_86  <dbl> 0.5824064, 0.0492279, 0.2617398, 0.3115448, 0.5302209,...
## $ var_87  <dbl> 0.6562390, 0.4925122, 0.4449250, 0.6308907, 0.7694739,...
## $ var_88  <dbl> 0.4443736, 0.4377062, 0.7151451, 0.2051925, 0.4448518,...
## $ var_89  <dbl> 0.23316187, 0.93055583, 0.42589220, 0.69641320, 0.5084...
## $ var_90  <dbl> 0.4409832, 0.7182759, 0.4289132, 0.6163010, 0.2089136,...
## $ var_92  <dbl> 0.4291235, 0.3066304, 0.5543504, 0.5642008, 0.3585428,...
## $ var_94  <dbl> 0.7441879, 0.5532248, 0.7082879, 0.2766588, 0.6954877,...
## $ var_96  <dbl> 0.3903504, 0.6985536, 0.6131537, 0.4370072, 0.5025496,...
## $ var_97  <dbl> 0.4725649, 0.6098443, 0.4828146, 0.4974573, 0.7231683,...
## $ var_99  <dbl> 0.2527037, 0.6454383, 0.5765130, 0.7653595, 0.6432963,...
## $ var_100 <dbl> 0.8003294, 0.3651362, 0.6149516, 0.2493628, 0.1917560,...
## $ var_101 <dbl> 0.4662711, 0.2935171, 0.4423666, 0.7911548, 0.4326603,...
## $ var_102 <dbl> 0.5827672, 0.7598108, 0.3519217, 0.4257563, 0.4005563,...
## $ var_104 <dbl> 0.7859094, 0.8284615, 0.8143370, 0.3054891, 0.1080676,...
## $ var_106 <dbl> 0.58857208, 0.38059067, 0.19419723, 0.66869792, 0.5205...
## $ var_107 <dbl> 0.4822430, 0.8197836, 0.2894819, 0.2546754, 0.3354656,...
## $ var_109 <dbl> 0.7131761, 0.2806254, 0.5607253, 0.3315523, 0.8632373,...
## $ var_110 <dbl> 0.3374876, 0.5545683, 0.3941755, 0.4439650, 0.2321647,...
## $ var_111 <dbl> 0.5738186, 0.6288257, 0.4052567, 0.6080764, 0.4470583,...
## $ var_112 <dbl> 0.5684221, 0.7162102, 0.6312490, 0.3355598, 0.2763603,...
## $ var_113 <dbl> 0.1776538, 0.5146869, 0.2776322, 0.6321745, 0.4941424,...
## $ var_114 <dbl> 0.3935455, 0.5871620, 0.5008245, 0.4028017, 0.4626651,...
## $ var_115 <dbl> 0.44968600, 0.62770464, 0.54348937, 0.17702352, 0.1052...
## $ var_116 <dbl> 0.5123714, 0.5621391, 0.3730863, 0.7938642, 0.4032009,...
## $ var_117 <dbl> 0.49382426, 0.39427259, 0.66382135, 0.78419471, 0.6240...
## $ var_118 <dbl> 0.3386917, 0.3768175, 0.2618918, 0.6775648, 0.8819259,...
## $ var_119 <dbl> 0.6707135, 0.6802221, 0.4633578, 0.5834470, 0.5693646,...
## $ var_120 <dbl> 0.3795074, 0.6033407, 0.4535209, 0.5886081, 0.7400183,...
## $ var_121 <dbl> 0.3342973, 0.3123144, 0.2591156, 0.4882464, 0.6261529,...
## $ var_122 <dbl> 0.4860265, 0.7281740, 0.5479120, 0.5867729, 0.6305397,...
## $ var_123 <dbl> 0.31843374, 0.58294943, 0.23531384, 0.06423337, 0.5707...
## $ var_124 <dbl> 0.4967578, 0.2296102, 0.4623095, 0.3978312, 0.4781977,...
## $ var_127 <dbl> 0.4162841, 0.5841131, 0.3200466, 0.3826833, 0.8292372,...
## $ var_128 <dbl> 0.4627056, 0.6032620, 0.5127320, 0.7455652, 0.8508019,...
## $ var_129 <dbl> 0.4968766, 0.5426311, 0.5667863, 0.4350760, 0.3072163,...
## $ var_132 <dbl> 0.7113331, 0.2648675, 0.6158831, 0.6309542, 0.4936860,...
## $ var_134 <dbl> 0.7927690, 0.7045140, 0.7588260, 0.6179221, 0.4633565,...
```

```
## $ var_135 <dbl> 0.6891759, 0.4076667, 0.3442989, 0.6618784, 0.7666765,...
## $ var_136 <dbl> 0.6028356, 0.2918186, 0.5252038, 0.5728989, 0.4086867,...
## $ var_137 <dbl> 0.7561970, 0.5083964, 0.4602273, 0.6290305, 0.6515961,...
## $ var_138 <dbl> 0.5638754, 0.3763168, 0.6026883, 0.3430564, 0.5580611,...
## $ var_139 <dbl> 0.6809085, 0.6924940, 0.5196271, 0.5268856, 0.6415782,...
## $ var_140 <dbl> 0.6688493, 0.5076000, 0.6359075, 0.5143323, 0.6096701,...
## $ var_141 <dbl> 0.2759676, 0.5815108, 0.3545166, 0.8124706, 0.3059363,...
## $ var_142 <dbl> 0.71533107, 0.11980450, 0.44359521, 0.43740208, 0.4896...
## $ var_143 <dbl> 0.4321789, 0.4933373, 0.2041448, 0.2255137, 0.7241149,...
## $ var_144 <dbl> 0.5114095, 0.4946161, 0.5879377, 0.7481356, 0.3232969,...
## $ var_145 <dbl> 0.6649061, 0.7835394, 0.4892417, 0.4260911, 0.6994556,...
## $ var_146 <dbl> 0.5855143, 0.3949763, 0.5802872, 0.5086228, 0.5855936,...
## $ var_147 <dbl> 0.19558998, 0.85091409, 0.61870465, 0.67438795, 0.2830...
## $ var_149 <dbl> 0.7150415, 0.3319556, 0.3860276, 0.4483966, 0.5776691,...
## $ var_150 <dbl> 0.6063123, 0.7513895, 0.3188167, 0.5775057, 0.7181860,...
## $ var_151 <dbl> 0.5504912, 0.7284045, 0.4300718, 0.8284440, 0.1385267,...
## $ var_152 <dbl> 0.57726946, 0.60235099, 0.30625831, 0.49752187, 0.3914...
## $ var_153 <dbl> 0.4814985, 0.4838611, 0.2649108, 0.6131253, 0.7472507,...
## $ var_154 <dbl> 0.61047188, 0.23366369, 0.07663329, 0.83377142, 0.4133...
## $ var_155 <dbl> 0.6179889, 0.5729784, 0.5848789, 0.2317021, 0.8925852,...
## $ var_156 <dbl> 0.3184686, 0.3300808, 0.7277202, 0.2968214, 0.4618358,...
## $ var_157 <dbl> 0.24374314, 0.35245261, 0.60335639, 0.62065388, 0.2603...
## $ var_158 <dbl> 0.4816029, 0.5856939, 0.8353451, 0.5331018, 0.6876160,...
## $ var_159 <dbl> 0.5589288, 0.5196858, 0.5615759, 0.5952036, 0.4111142,...
## $ var_160 <dbl> 0.3516103, 0.5668909, 0.3170690, 0.6037052, 0.6286819,...
## $ var_162 <dbl> 0.2878031, 0.3128747, 0.4215060, 0.7273817, 0.5142182,...
## $ var_163 <dbl> 0.3079106, 0.6641240, 0.3302044, 0.2700173, 0.7428239,...
## $ var_164 <dbl> 0.8281287, 0.4590790, 0.5844658, 0.2672272, 0.8249997,...
## $ var_165 <dbl> 0.2681554, 0.3531586, 0.5381174, 0.4259659, 0.5639787,...
## $ var_167 <dbl> 0.5794264, 0.8079820, 0.4676734, 0.5336934, 0.5548201,...
## $ var_168 <dbl> 0.7742836, 0.4432437, 0.6422093, 0.2510830, 0.3995876,...
## $ var_170 <dbl> 0.3156475, 0.7212068, 0.2239959, 0.2215788, 0.5602593,...
## $ var_171 <dbl> 0.2544151, 0.6947592, 0.4013831, 0.3271340, 0.4420470,...
## $ var_172 <dbl> 0.5217829, 0.5455034, 0.2422273, 0.4675832, 0.4997865,...
## $ var_173 <dbl> 0.61224170, 0.68074356, 0.35172104, 0.86068319, 0.5895...
## $ var_174 <dbl> 0.4590940, 0.7367684, 0.6280377, 0.3686096, 0.4568802,...
## $ var_175 <dbl> 0.2679462, 0.4319299, 0.4571362, 0.2893186, 0.7792516,...
## $ var_176 <dbl> 0.3437638, 0.4108585, 0.5716648, 0.8546804, 0.5011820,...
## $ var_177 <dbl> 0.5819214, 0.7708535, 0.2581696, 0.5972523, 0.5687864,...
## $ var_178 <dbl> 0.4482705, 0.3726469, 0.3261682, 0.3109756, 0.5517236,...
## $ var_179 <dbl> 0.4477489, 0.8309945, 0.4930035, 0.7119213, 0.6772511,...
## $ var_180 <dbl> 0.3201393, 0.4122140, 0.4547961, 0.4724454, 0.3434347,...
## $ var_181 <dbl> 0.3657285, 0.4269517, 0.4143538, 0.2871930, 0.4634077,...
## $ var_182 <dbl> 0.5418523, 0.2067561, 0.2374253, 0.5200957, 0.5758789,...
## $ var_183 <dbl> 0.7079092, 0.5226606, 0.2877561, 0.4415717, 0.6273860,...
## $ var_184 <dbl> 0.7408266, 0.6809836, 0.6908342, 0.2632063, 0.4868065,...
## $ var_185 <dbl> 0.8638058, 0.4671860, 0.4510162, 0.6449273, 0.3552426,...
## $ var_186 <dbl> 0.6609015, 0.4271614, 0.2998850, 0.4427541, 0.3583033,...
## $ var_187 <dbl> 0.3801626, 0.4353474, 0.5763122, 0.4806153, 0.7541316,...
## $ var_188 <dbl> 0.5806859, 0.3928828, 0.2550527, 0.3529513, 0.5121171,...
## $ var_189 <dbl> 0.4724739, 0.3113695, 0.2284322, 0.6884425, 0.3239699,...
## $ var_190 <dbl> 0.5487376, 0.6727947, 0.4895584, 0.5499447, 0.3194835,...
## $ var_191 <dbl> 0.3059213, 0.5218536, 0.6407681, 0.3506974, 0.6253024,...
## $ var_192 <dbl> 0.64773238, 0.58054870, 0.46953213, 0.35373844, 0.2481...
```

```
## $ var_193 <dbl> 0.4241271, 0.8359967, 0.4238958, 0.4121365, 0.7578355,...
## $ var_194 <dbl> 0.5218382, 0.3576564, 0.6854518, 0.7614049, 0.2438781,...
## $ var_195 <dbl> 0.2218936, 0.7772167, 0.9160318, 0.3631397, 0.3328781,...
## $ var_196 <dbl> 0.6789810, 0.6866137, 0.2363369, 0.3468097, 0.5575065,...
## $ var_197 <dbl> 0.4329306, 0.4750229, 0.3776541, 0.7557564, 0.6083960,...
## $ var_198 <dbl> 0.31291822, 0.64221972, 0.42760706, 0.61940479, 0.6210...
## $ var_199 <dbl> 0.5358626, 0.5835937, 0.5591996, 0.4118264, 0.4147939,...
## $ ID_code <chr> "train_0", "train_1", "train_2", "train_3", "train_4",...
## $ target  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ...
```

Estudio de correlaciones

En esta sección vamos estudiar la correlación entre las variables, es decir, si existe alguna dependencia directa o inversa entre algún par de variables. El objetivo es eliminar variables que tengan una alta dependencia de otra ya que no aportan información al conjunto de datos.

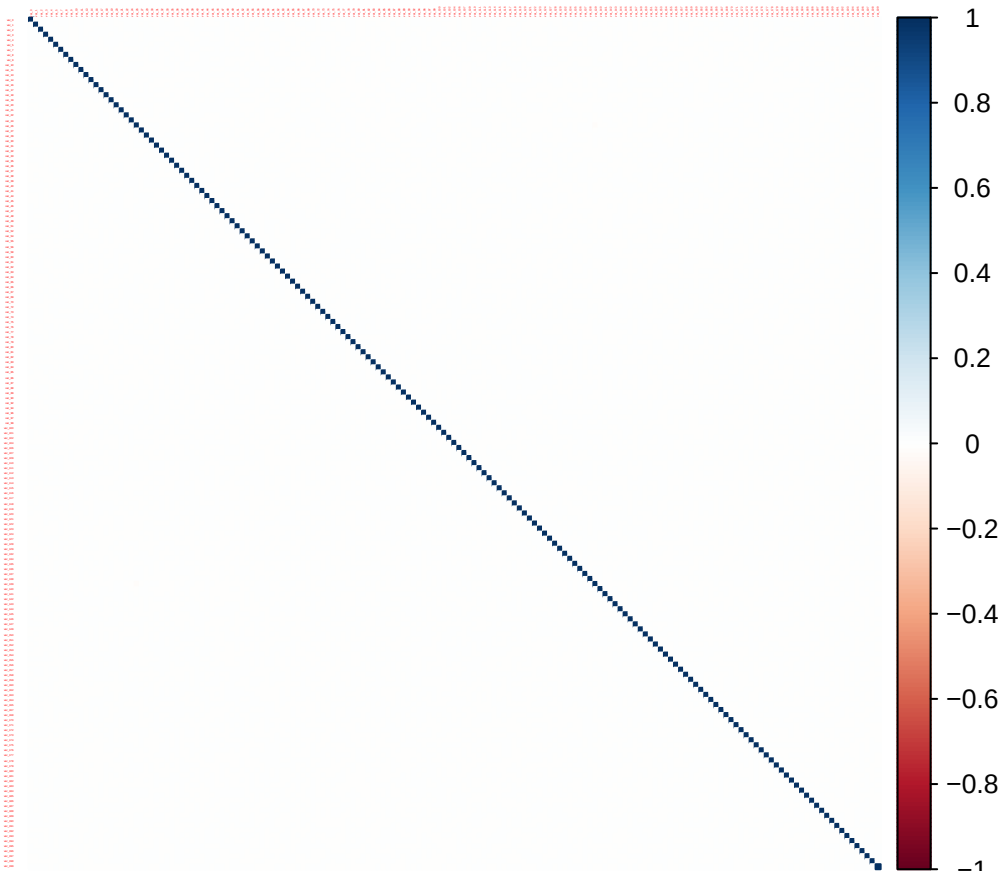
Para estudiar la correlación vamos a usar la librería `corrplot` la cual nos permite visualizarla gráficamente. Tras ver el diagrama siguiente podemos observar que no hay ninguna correlación entre ningún par de variables. Por otro lado se puede observar que hay una correlación máxima entre una variable y sí misma lo que es trivial pero nos demuestra que el método funciona.

Aun así, como tenemos demasiadas variables vamos a eliminar las que tengan una correlacion con la variable *target* menor que el 0.02. Despues de esto nos quedamos con 116 variables.

```
# Librería corrplot para dibujar la matrix de correlaciones
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
# Eliminamos las variables que no son numéricas (id)
data_correlation <- data_normalized[c(-170,-171)]
# Creamos la matriz de correlación de los datos
correlation_table <- cor(data_correlation)
# Dibujamos la matriz con el método de color y con el tamaño de las etiquetas a 0.1 ya que son demasiad
corrplot(correlation_table, method = "color",tl.cex = 0.1)
```




```
# Vemos la correlación de las variables con target
cor_target <- correlation_table(data_normalized, target='target')
# Nos quedamos con las variables con una correlación con target mayor que el 0.02
important_vars <- cor_target %>%
  filter(abs(target) >= 0.02)
# Creamos el data set listo para usarlo en la predicción con las variables más importantes
data_important <- data_normalized %>%
  select(one_of(important_vars$Variable))

# Vemos el nuevo tamaño de los datos
dim(data_important)
```

```
## [1] 196041    116
```

Downsampling

Debido a la cantidad enorme de datos y a disponer de una máquina no muy potente se ha optado por reducir las muestras de la clase mayoritaria para poder crear un modelo de predicción en un tiempo razonable aunque se pierda precisión. Además se balancearán las clases lo que será una ventaja a la hora del entrenamiento del modelo de predicción.

Se va a usar la técnica de *downsampling* que hace justo lo que he comentado. Se establece una semilla para la selección aleatoria de R y despues se eliminan aleatoriamente filas de la clase mayoritaria hasta tener el mismo número de clases.

Vemos que en el nuevo conjunto de datos hay el mismo número de filas en cada clase.

```r
library(caret)
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:survival':
##
##     cluster
```

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

```r
# Establecemos la semilla
set.seed(9560)
# Cambiamos la variable target a tipo factor
data_downsampled <- data_important %>%
  mutate(target = as.factor(ifelse(target == 1, 'Yes', 'No')))

# Hacemos doumsamplin para igualar el número de filas pertenecientes a cada clase
predictors <- select(data_downsampled, -target)
data_downsampled <- downSample(x = predictors, y = data_downsampled$target, yname = 'target')
# Vemos la cantidad de elementos en cada clase
count(data_downsampled, target)
```

```
## # A tibble: 2 x 2
##   target     n
##   <fct>  <int>
## 1 No     17664
## 2 Yes    17664
```

```r
# Nueva variable para la predicción
data_ready <- data_downsampled
```

Predicción

Una vez tenemos los datos preparados se va a proceder a usar distintos modelos de predicción para, en base a las variables de cada fila, intentar predecir la variable *target*. En primer lugar vamos a usar **random forest** que es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. Fuente.

Para ello vamos a usar una función proporcionada en el siguiente GitHub que envuelve la función de entrenamiento del paquete *caret*. El primer paso es barajar aleatoriamente el conjunto de datos y despues dividir los datos en un conjunto de entrenamiento y otro de validación en una proporción de 70% y 30% respectivamente. A continuación entrenamos el *random forest* con el conjunto de entrenamiento y al terminar probamos el modelo prediciendo en el conjunto de datos de validación. Para comprobar su precisión se va a usar la curva ROC que representa gráficamente la cantidad de aciertos al variar un umbral de discriminación. La precisión se mide con el área que se encuentra debajo de esta curva ROC, cuanto más cercana sea a 1 mejor será la precisión.

En este caso hemos obtenido una precisión del 0.98.

```r
library(caret)
library(spatstat)
```

```
## Loading required package: spatstat.data
```

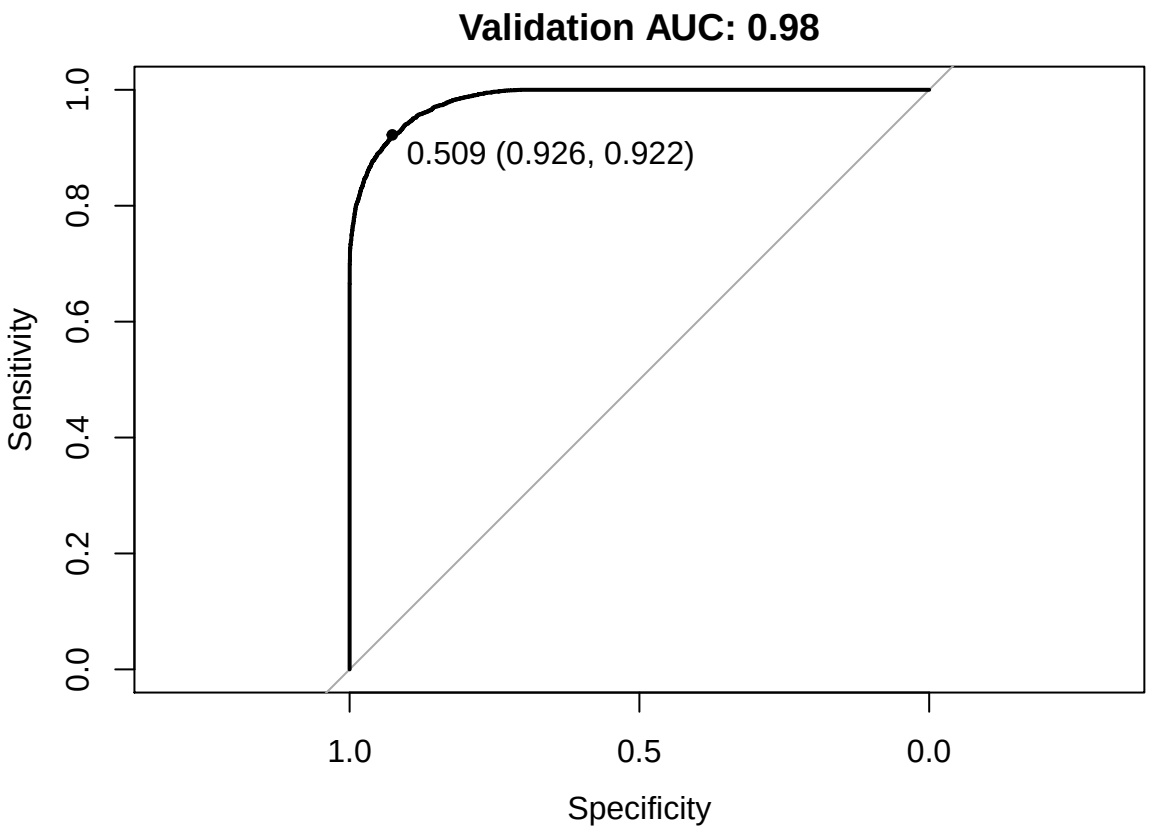
```
## Loading required package: nlme
##
## Attaching package: 'nlme'
## The following object is masked from 'package:BBmisc':
##
##     collapse
## The following object is masked from 'package:dplyr':
##
##     collapse
## Loading required package: rpart
##
## spatstat 1.59-0       (nickname: 'J'ai omis les oeufs de caille')
## For an introduction to spatstat, type 'beginner'
##
## Note: R version 3.4.4 (2018-03-15) is more than 9 months old; we strongly recommend upgrading to the
##
## Attaching package: 'spatstat'
## The following object is masked from 'package:lattice':
##
##     panel.histogram
```

```
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
## The following objects are masked from 'package:spatstat':
##
##     auc, coords, roc
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
# Función para entrenar sobre un conjunto de datos con el método de random forest
trainRF <- function(train_data, rfCtrl = NULL, rfParametersGrid = NULL) {
  if(is.null(rfCtrl)) {
    rfCtrl <- trainControl(
      verboseIter = T,
      classProbs = TRUE,
      method = "repeatedcv",
      number = 10,
      repeats = 1,
      summaryFunction = twoClassSummary)
  }
  if(is.null(rfParametersGrid)) {
    rfParametersGrid <- expand.grid(
      .mtry = c(sqrt(ncol(train_data))))
  }

  rfModel <- train(
```

```
    target ~ .,
    data = train_data,
    method = "rf",
    metric = "ROC",
    trControl = rfCtrl,
    tuneGrid = rfParametersGrid)

  return(rfModel)
}

# Barajamos aleatoriamente los datos
data_ready <- data_ready[sample(1:nrow(data_ready)), ]
# Creamos un conjunto de índices aleatorios cuyo cardinal es el 70% del total del conjunto de datos
trainIndex <- createDataPartition(data_ready$target, p = .7, list = FALSE, times = 1)
# Usando esos índices creamos el conjunto de entrenamiento
data_train <- data_ready[trainIndex,]
# Creamos el conjunto de validación usando los índices que no están en trainIndex
data_val <- data_ready[-trainIndex,]


# Entrenamos el modelo con la función anterior
# rfModel <- trainRF(data_train)
# Guardamos el modelo para poder usarlo posteriormente
# saveRDS(rfModel, file = "model1.rds")
# Leemos el modelo
rfModel <- readRDS("model1.rds")

# Realizamos una predicción sobre el conjunto de validación
predictionValidationProb <- predict(rfModel, data_val, type = "prob")

# Medimos la precisión del modelo usando la curva ROC
auc <- roc(data_val$target, predictionValidationProb[["Yes"]], levels = unique(data_val[["target"]]))
# Dibujamos la curva ROC
roc_validation <- plot.roc(auc, ylim=c(0,1), type = "S" , print.thres = T, main=paste('Validation AUC:'
```

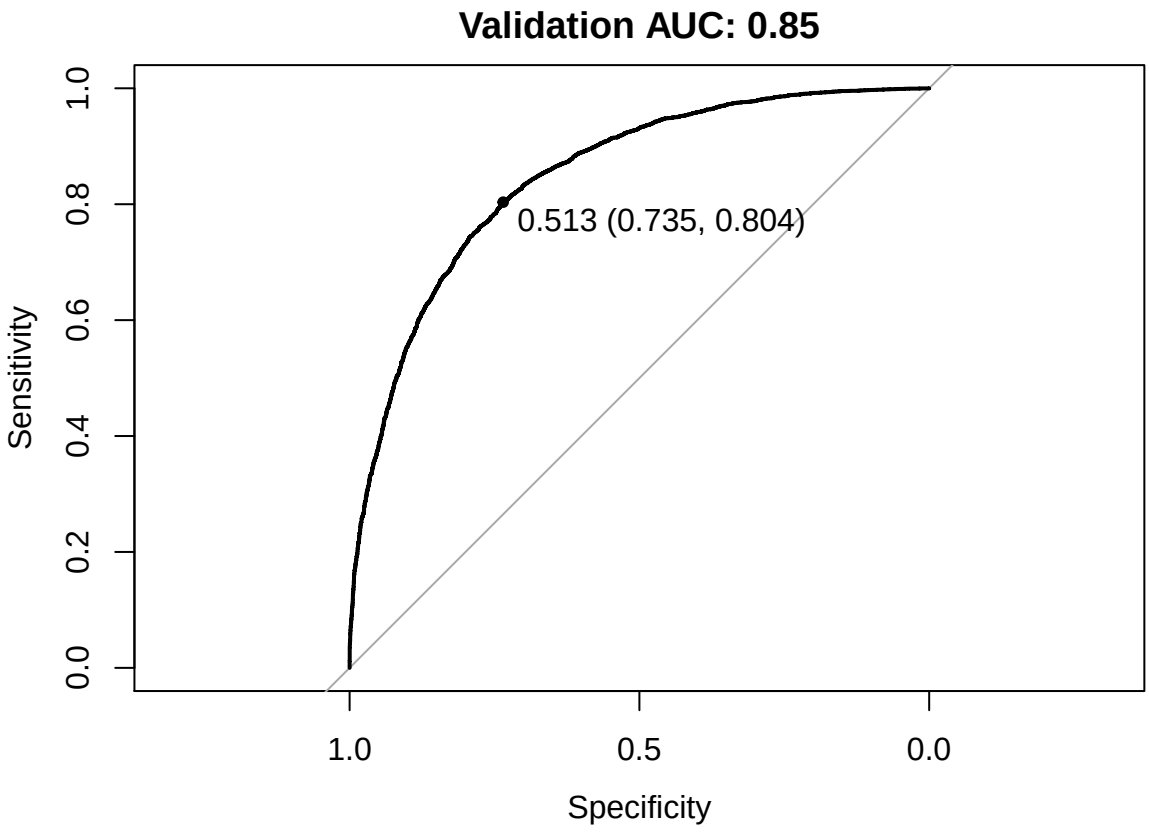


Ahora vamos a usar el método **boosted tree**, que resumidamente consiste en computar una secuencia de árboles muy sumples, donde cada árbol sucesivo es construido con los residuos de la predicción del árbol predecesor(Fuente)

En este caso podemos observar que hemos obtenido peores resultados con un 0.82 de precisión.

```
# Función para entrenar sobre un conjunto de datos con el método de boosted tree
trainBT <- function(train_data, btCtrl = NULL, btParametersGrid = NULL) {
  if(is.null(btCtrl)) {
    btCtrl <- trainControl(
      verboseIter = T,
      method = "repeatedcv",
      number = 10,
      repeats = 1)
  }
  if(is.null(btParametersGrid)) {
    btParametersGrid <- expand.grid(
      .mtry = c(sqrt(ncol(train_data))))
  }

  btModel <- train(
    target ~ .,
    data = train_data,
    method = "gbm",
    trControl = btCtrl,
    verbose = FALSE)
```

```
  return(btModel)
}


# Entrenamos el modelo con la función anterior
# btModel <- trainBT(data_train)
# Guardamos el modelo para poder usarlo posteriormente
# saveRDS(btModel, file = "model2.rds")
# Leemos el modelo
btModel <- readRDS("model2.rds")

# Realizamos una predicción sobre el conjunto de validación
predictionValidationProb <- predict(btModel, data_val, type = "prob")

# Medimos la precisión del modelo usando la curva ROC
auc <- roc(data_val$target, predictionValidationProb[["Yes"]], levels = unique(data_val[["target"]]))
# Dibujamos la curva ROC
roc_validation <- plot.roc(auc, ylim=c(0,1), type = "S" , print.thres = T, main=paste('Validation AUC:'
```



Discusión de resultados

Tras usar dos métodos de predicción (*random forest* y *boosted tree*) observamos que el mejor resultado lo obtiene el random forest con un 0.98 de precisión. Esto es un muy buen resultado ya que ha acertado casi el 100% de los casos. Si ademas tenemos en cuenta de que se trataba de un conjunto de datos balanceados 50/50 el resultado es aún mas prometedor.

Si recordamos, en los datos originales, en torno al 90% de muestras pertenecían a la clase 0. Por lo tanto hemos obtenido un modelo más preciso que si hubiéramos creado uno en el cuál siempre predijera 0 (obtendría un 0.9 de precisión)

Conclusiones

En esta práctica hemos tenido que trabajar con un conjunto de datos relativamente grande de los cuales hemos tenido que preprocesarlos a ciegas, es decir, no teníamos ninguna información de qué significaba cada variable ni qué representaba el conjunto de datos en general. Esta situación se asemeja a muchos casos de la vida real donde tenemos gran cantidad de variables y el analista de datos no tiene mucha información de qué significa cada una.

Como el conjunto de datos era demasiado grande se han tenido que eliminar tanto filas como columnas(variables). Esto ha permitido aprender a identificar datos que no aportan mucha información pero por otro lado ha interferido en la precisión del modelo de predicción. Sin embargo esta interferencia ha sido casi nula ya que hemos obtenido una precisión del 0.98.

Bibliografía

- [1] (https://github.com/jgromero/sige2019)
- [2] (https://es.wikipedia.org/wiki/Curva_ROC)
- [3] (https://es.wikipedia.org/wiki/Random_forest)
- [4] (http://r-statistics.co/Outlier-Treatment-With-R.html)
- [5] (https://www.r-statistics.com/)
- [6] (https://topepo.github.io/caret/model-training-and-tuning.html)