Análise Exploratória de Dados Porto Seguro

Por Igor Magalhães Rodrigues

0. Tabela de Conteúdos

1. Relatório

Como podemos ler na página sobre os dados, o objetivo do *dataset* é analisar a probabilidade de um cliente contratar um novo produto. Para isso teremos algumas variáveis. Trabalharemos apenas com os dados *train.csv* e *metadata.csv*, a descrição indica que os dados de *train.csv* são divididos da seguinte maneira: 1 variável de identificação, 68 variáveis explicativas, 1 variável resposta onde cada linha da tabela representa um cliente. Os dados de metadata representam os tipos de cada variável, sendo esses Qualitativo Nominal, Qualitativo Ordinal, Quantitativo Discreto, Quantitativo Contínua.

A próxima etapa é identificar o que queremos descobrir a partir desses dados. Como dito, o objetivo do *dataset* é analisar a probabilidade de um cliente contratar um novo produto, então nossas perguntas para a análise exploratória estarão relacionadas a identificar quais variáveis podem ser úteis para isso, além de suas descrições estatísticas.

O próximo passo é ler os dados para que possamos de fato analisar usando o R. Para isso, é importante notar que dados com "-999" em train.csv devem ser considerados como os dados faltantes, portanto leremos esses como *NA*. Isso pôde ser identificado criando uma *view()* da tabela criada a partir de *train.csv*, mas uma outra forma de notar seria que "-999" seriam observações discrepantes das outras em vários sentidos.

No nosso *dataset*, não sabemos qual característica de um cliente cada variável representa, no entanto, demos a sorte de ter esses tipos indicados no arquivo *metadata.csv*.

Note que isso ainda é uma desvantagem mesmo tendo os tipos especificados, já que não podemos pensar nas variáveis como tipos diferentes dependendo da análise que queremos fazer. Como é o que temos, seguiremos assim.

Começaremos olhando para os nossos dados, vendo *head*, *tail*, *str*. Notamos a existência de dados faltantes, algumas variáveis que aparentam ser binárias, algumas com muitas classes. Para investigar melhor, podemos ver um sumário da estatística descritiva de cada variável.

Com estatística descritiva queremos ver coisas como o menor valor, primeiro quadrante, mediana, média, terceiro quadrante e valor máximo, isto é, o *summary()* dos dados. Note que isso faz sentido apenas para dados quantitativos, para dados qualitativos apenas veremos as frequências das ocorrências de cada classe. Sobre os dados qualitativos, podemos ver na figura 1 algumas variáveis com maior parte das ocorrências em apenas algumas classes, já outras com poucas ocorrências por classe. Nesse segundo caso, poderíamos agrupar as classes em intervalo, mas como não temos mais informações sobre o que cada variável significa, talvez não seja algo sábio. Outra coisa que podemos ver com o sumário dos dados é a quantidade de dados faltantes, que é substancial em algumas variáveis. Isso pode atrapalhar no caso de uma modelagem preditiva, já que reduz a quantidade de dados possível para fazer o modelo, mas pode ser amenizado descartando variáveis que não influenciam tanto no valor alvo. Veremos mais sobre isso posteriormente.

Analisando os dados das variáveis numéricas, notamos pela figura 2 que várias dessas realmente devem ser variáveis booleanas, isto é, apenas 0 ou 1. Grande parte das variáveis numéricas tem 0 como seu valor mínimo. Para ter uma análise mais visual, vamos fazer *box plots* das variáveis, notando que algumas dessas devem ser separadas, posto que seus valores máximos fariam com que a análise das variáveis de menor magnitude ficasse ofuscada. Separaremos em 3 grupos: **grupo 1** com *var40*, *var45*, *var46*, *var48*; **grupo 2** com apenas a *var52* e o **grupo 3** com todo o resto.

Vendo os *box plots* figura 3, figura 4, figura 5, pode-se notar que várias das observações possuem vários dados discrepantes, enquanto algumas outras não têm. Em uma possível modelagem preditiva seria interessante analisar a acurácia do modelo descartando e levando em conta observações discrepantes.

Voltando ao nosso objetivo, vejamos as correlações entre as variáveis e a variável alvo *y*. Para isso, faremos uma matriz de correlação com os dados numéricos, descartando as observações com dados faltantes. Como pode ser visto na figura 6, existe alguma correlação positiva entre *y* e as variáveis *var54*, *var53*, *var50* e *var57* em ordem decrescente, porém ainda não é uma grande correlação. Também existe alguma correlação negativa entre *y* e as variáveis *var45*, *var46*, *var58*, porém pouco relevante.

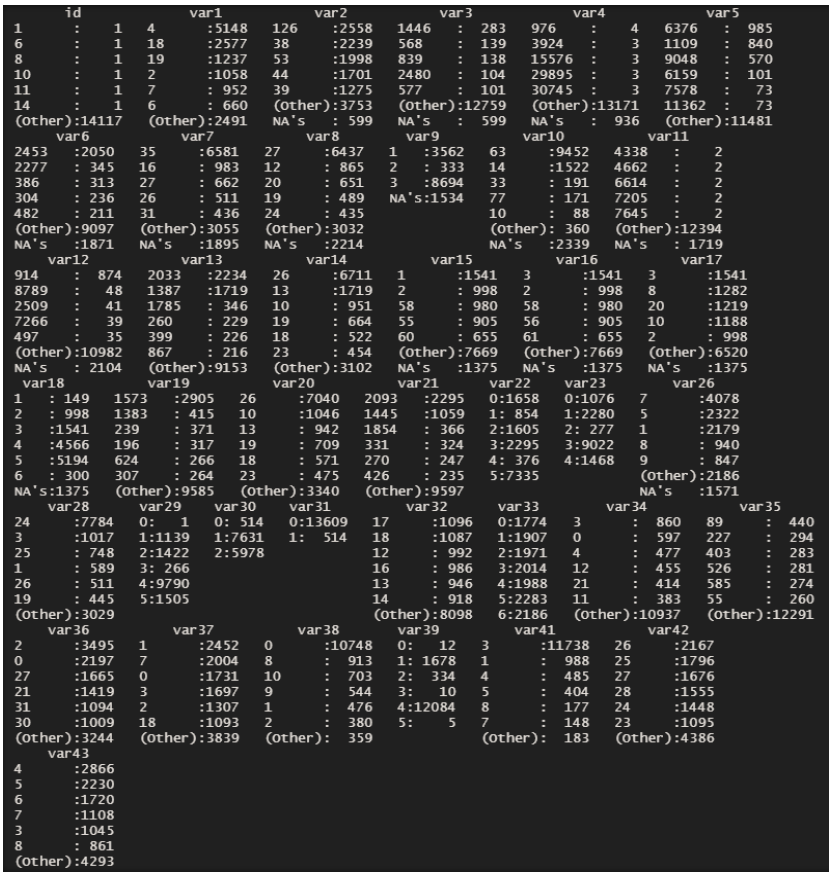
Façamos agora um *PCA (Principal Component Analysis)* para ver as direções dos autovetores pensando em uma redução de dimensionalidade, o resultado pode ser visto na figura 7. Como essa análise não vai se aprofundar tanto, vamos passar para os dados categóricos.

Fazendo o teste do Qui quadrado descobrimos vários candidatos a variáveis que podem influenciar no valor de y. Outra redução de dimensionalidade poderia ser feita aqui, analisando correlação entre as variáveis e usando um número menor em um possível modelo.

Concluiremos nossa análise pensando nos próximos passos. Uma primeira análise dos nossos dados está feita, o que significa que estamos familiarizados com o que trabalharemos. Outras coisas a se fazer com os dados, como analisar melhor variáveis específicas, fazer outras visualizações e outros testes surgem com as tentativas de fazer um modelo preditivo. Neste *dataset* poderíamos usar algum modelo de aprendizado supervisionado, posto que temos a resposta y para um vetor de dados x. Nesse caso queremos fazer uma regressão, já que a saída do modelo é um valor contínuo representando a probabilidade de o cliente comprar o produto. Uma boa forma de começar - embora talvez seja simples - é testar uma regressão linear para ver o quão complexa realmente é essa tarefa. Posteriormente modelos mais sofisticados como gradient boosting (mas não redes neurais, já que a quantidade de dados é pequena) podem ser testados, respeitando a codificação das variáveis categóricas.

2. Apêndice

2.1 Imagens



```
      id              var1            var2            var3            var4            var5
1      :    1   4      :5148   126    :2558   1446   :  283   976    :    4   6376   :  985
6      :    1   18     :2577   38     :2239   568    :  139   3924   :    3   1109   :  840
8      :    1   19     :1237   53     :1998   839    :  138   15576  :    3   9048   :  570
10     :    1   2      :1058   44     :1701   2480   :  104   29895  :    3   6159   :  101
11     :    1   7      : 952   39     :1275   577    :  101   30745  :    3   7578   :   73
14     :    1   6      : 660   (Other):3753   (Other):12759   (Other):13171   11362  :   73
(Other):14117   (Other):2491   NA's   : 599   NA's   :  599   NA's   :  936   (Other):11481
     var6            var7            var8            var9            var10           var11
2453   :2050   35     :6581   27     :6437   1    :3562   63     :9452   4338   :    2
2277   : 345   16     : 983   12     : 865   2    : 333   14     :1522   4662   :    2
386    : 313   27     : 662   20     : 651   3    :8694   33     : 191   6614   :    2
304    : 236   26     : 511   19     : 489   NA's:1534   77     : 171   7205   :    2
482    : 211   31     : 436   24     : 435               10     :  88   7645   :    2
(Other):9097   (Other):3055   (Other):3032               (Other): 360   (Other):12394
NA's   :1871   NA's   :1895   NA's   :2214               NA's   :2339   NA's   : 1719
    var12           var13           var14           var15           var16           var17
914    :  874   2033   :2234   26     :6711   1      :1541   3      :1541   3      :1541
8789   :   48   1387   :1719   13     :1719   2      : 998   2      : 998   8      :1282
2509   :   41   1785   : 346   10     : 951   58     : 980   58     : 980   20     :1219
7266   :   39   260    : 229   19     : 664   55     : 905   56     : 905   10     :1188
497    :   35   399    : 226   18     : 522   60     : 655   61     : 655   2      : 998
(Other):10982   867    : 216   23     : 454   (Other):7669   (Other):7669   (Other):6520
NA's   : 2104   (Other):9153   (Other):3102   NA's   :1375   NA's   :1375   NA's   :1375
 var18           var19           var20           var21      var22   var23          var26
1    : 149   1573   :2905   26     :7040   2093   :2295   0:1658   0:1076   7      :4078
2    : 998   1383   : 415   10     :1046   1445   :1059   1: 854   1:2280   5      :2322
3    :1541   239    : 371   13     : 942   1854   : 366   2:1605   2: 277   1      :2179
4    :4566   196    : 317   19     : 709   331    : 324   3:2295   3:9022   8      : 940
5    :5194   624    : 266   18     : 571   270    : 247   4: 376   4:1468   9      : 847
6    : 300   307    : 264   23     : 475   426    : 235   5:7335            (Other):2186
NA's:1375    (Other):9585   (Other):3340   (Other):9597                     NA's   :1571
    var28           var29      var30      var31          var32          var33          var34          var35
24     :7784   0:   1   0: 514   0:13609   17     :1096   0:1774   3      :  860   89     :  440
3      :1017   1:1139   1:7631   1:  514   18     :1087   1:1907   0      :  597   227    :  294
25     : 748   2:1422   2:5978             12     : 992   2:1971   4      :  477   403    :  283
1      : 589   3: 266                      16     : 986   3:2014   12     :  455   526    :  281
26     : 511   4:9790                      13     : 946   4:1988   21     :  414   585    :  274
19     : 445   5:1505                      14     : 918   5:2283   11     :  383   55     :  260
(Other):3029                               (Other):8098   6:2186   (Other):10937   (Other):12291
    var36           var37          var38          var39          var41          var42
2      :3495   1      :2452   0      :10748   0:   12   3      :11738   26     :2167
0      :2197   7      :2004   8      :  913   1: 1678   1      :  988   25     :1796
27     :1665   0      :1731   10     :  703   2:  334   4      :  485   27     :1676
21     :1419   3      :1697   9      :  544   3:   10   5      :  404   28     :1555
31     :1094   2      :1307   1      :  476   4:12084   8      :  177   24     :1448
30     :1009   18     :1093   2      :  380   5:    5   7      :  148   23     :1095
(Other):3244   (Other):3839   (Other):  359              (Other):  183   (Other):4386
    var43
4      :2866
5      :2230
6      :1720
7      :1108
3      :1045
8      : 861
(Other):4293
```

Figura 1

```
> summary(train %>% select_if(is.numeric))
     var24           var25           var27           var40           var44           var45
 Min.   :0.0000   Min.   :0.000   Min.   :0.00   Min.   : 0.00   Min.   :0.0000   Min.   : 0.0000
 1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:0.00   1st Qu.: 2.00   1st Qu.:1.0000   1st Qu.: 0.0000
 Median :1.0000   Median :2.000   Median :0.00   Median : 5.00   Median :1.0000   Median : 0.0000
 Mean   :0.9586   Mean   :2.093   Mean   :0.37   Mean   : 6.01   Mean   :0.8658   Mean   : 0.2752
 3rd Qu.:2.0000   3rd Qu.:4.000   3rd Qu.:1.00   3rd Qu.:10.00   3rd Qu.:1.0000   3rd Qu.: 0.0000
 Max.   :2.0000   Max.   :4.000   Max.   :1.00   Max.   :20.00   Max.   :1.0000   Max.   :15.0000

     var46           var47           var48           var49           var50
 Min.   : 0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.: 0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median : 0.0000   Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   : 0.2216   Mean   :0.04985   Mean   :0.1061   Mean   :0.3437   Mean   :0.1483
 3rd Qu.: 0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :15.0000   Max.   :5.00000   Max.   :9.0000   Max.   :1.0000   Max.   :1.0000

     var51           var52           var53           var54           var55           var56
 Min.   :0.0000   Min.   : 1.0   Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:15.0   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.2052   1st Qu.:0.1860
 Median :0.0000   Median :24.0   Median :2.000   Median :2.000   Median :0.2095   Median :0.4530
 Mean   :0.0553   Mean   :26.1   Mean   :1.635   Mean   :1.519   Mean   :0.2160   Mean   :0.4901
 3rd Qu.:0.0000   3rd Qu.:36.0   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.2183   3rd Qu.:0.8200
 Max.   :1.0000   Max.   :64.0   Max.   :3.000   Max.   :3.000   Max.   :0.7509   Max.   :1.0000
                  NA's   :1534                                                    NA's   :1587
     var57           var58           var59           var60           var61
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
 1st Qu.:0.0785   1st Qu.:0.0234   1st Qu.:0.1372   1st Qu.:0.052   1st Qu.:0.1760
 Median :0.1226   Median :0.0622   Median :0.2020   Median :0.135   Median :0.2431
 Mean   :0.3458   Mean   :0.1166   Mean   :0.2216   Mean   :0.203   Mean   :0.2709
 3rd Qu.:0.7312   3rd Qu.:0.1594   3rd Qu.:0.2753   3rd Qu.:0.286   3rd Qu.:0.3476
 Max.   :1.0000   Max.   :1.0000   Max.   :0.9138   Max.   :1.000   Max.   :0.9350
 NA's   :1589     NA's   :1571     NA's   :2182     NA's   :6484    NA's   :346
     var62           var63           var64           var65           var66
 Min.   :0.00000   Min.   :0.0000   Min.   :0.004267   Min.   :0.000   Min.   :0.000
 1st Qu.:0.04039   1st Qu.:0.7064   1st Qu.:0.018357   1st Qu.:0.121   1st Qu.:0.002
 Median :0.06026   Median :0.8779   Median :0.023480   Median :0.251   Median :0.003
 Mean   :0.07309   Mean   :0.7792   Mean   :0.028778   Mean   :0.301   Mean   :0.007
 3rd Qu.:0.08883   3rd Qu.:0.9304   3rd Qu.:0.034581   3rd Qu.:0.425   3rd Qu.:0.006
 Max.   :0.81485   Max.   :1.0000   Max.   :0.387387   Max.   :1.000   Max.   :1.000
                                                       NA's   :12197   NA's   :12197
     var67           var68              y
 Min.   :0.0294   Min.   :0.00368   Min.   :0.0000
 1st Qu.:0.1176   1st Qu.:0.13603   1st Qu.:0.0000
 Median :0.1765   Median :0.17647   Median :0.0000
 Mean   :0.2069   Mean   :0.17987   Mean   :0.2018
 3rd Qu.:0.2647   3rd Qu.:0.22059   3rd Qu.:0.0000
 Max.   :0.9118   Max.   :1.00000   Max.   :1.0000
 NA's   :586      NA's   :33
```

Figura 2

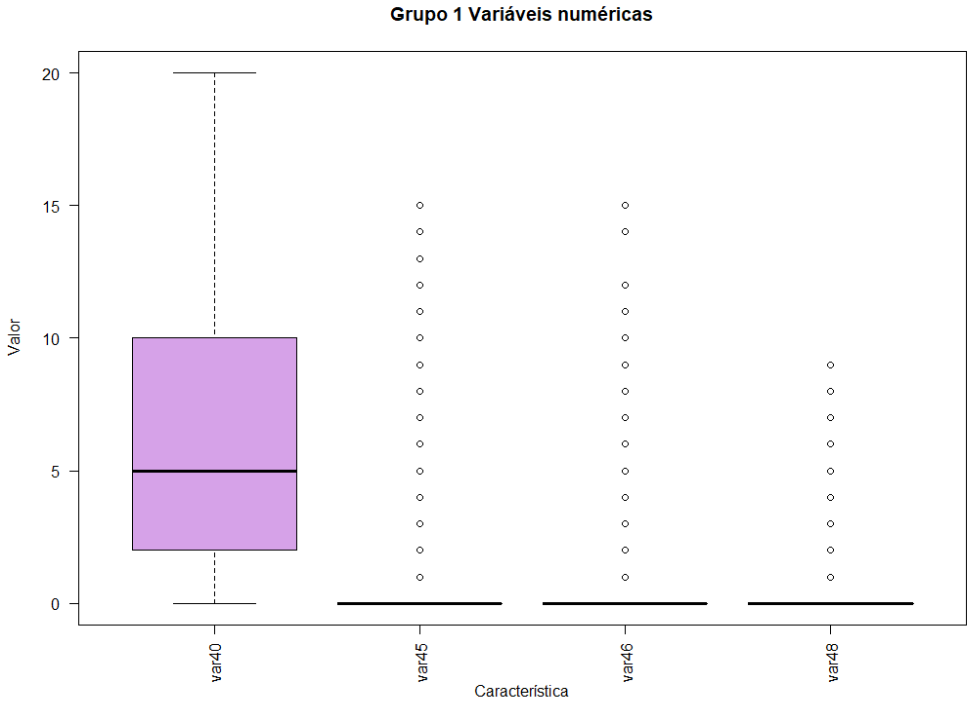


Figura 3

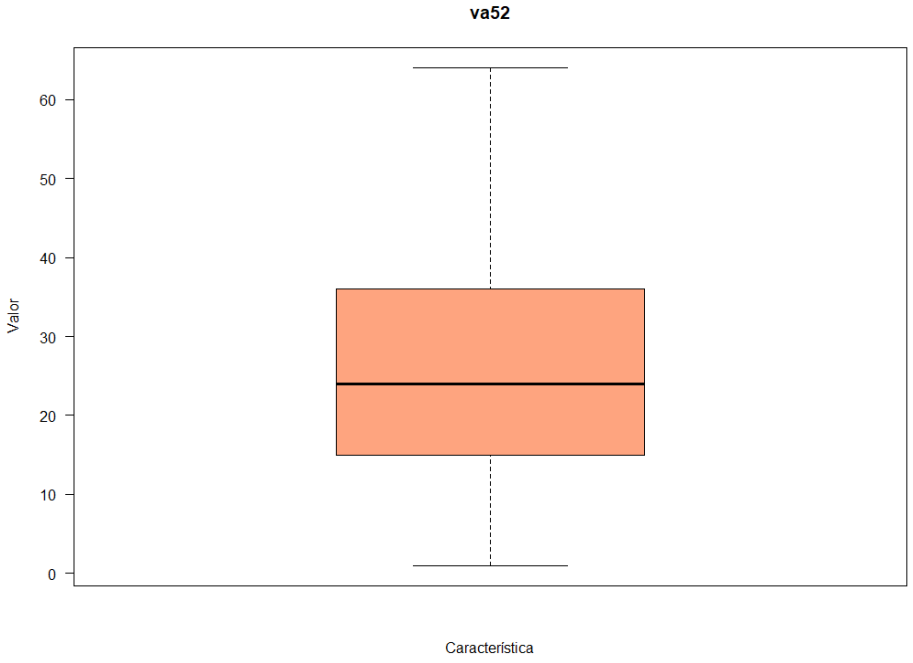


Figura 4

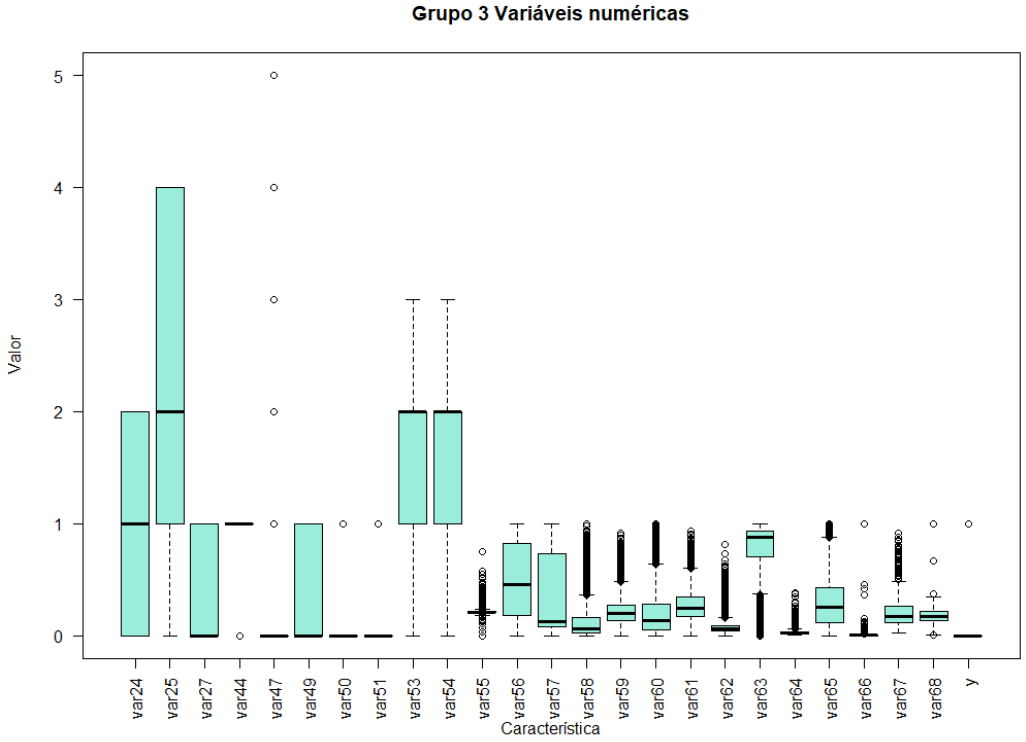


Figura 5

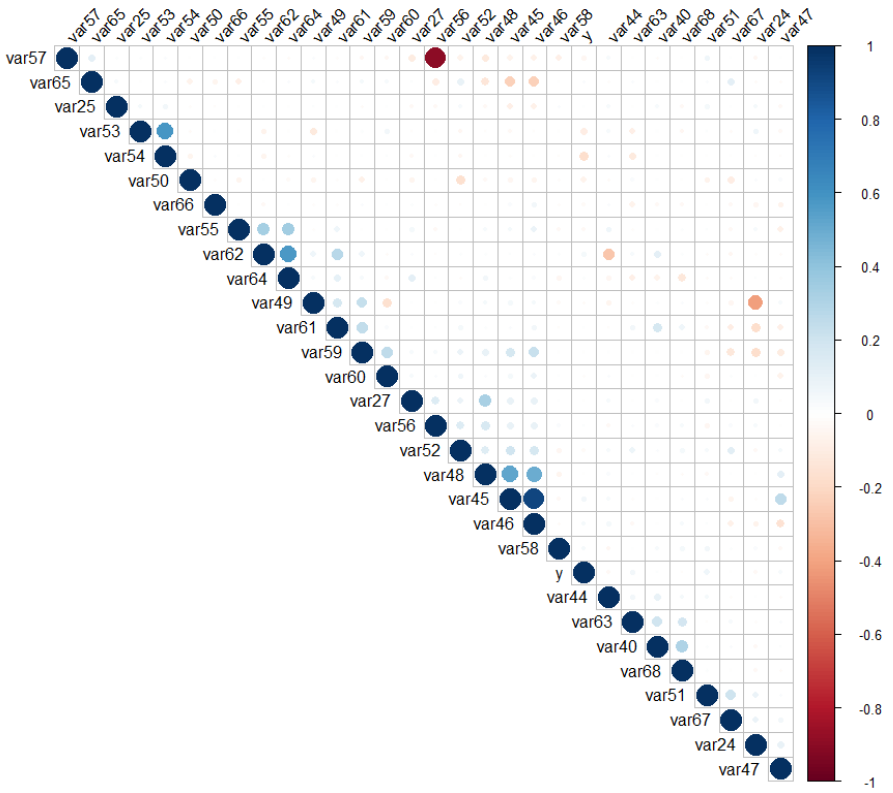


Figura 6

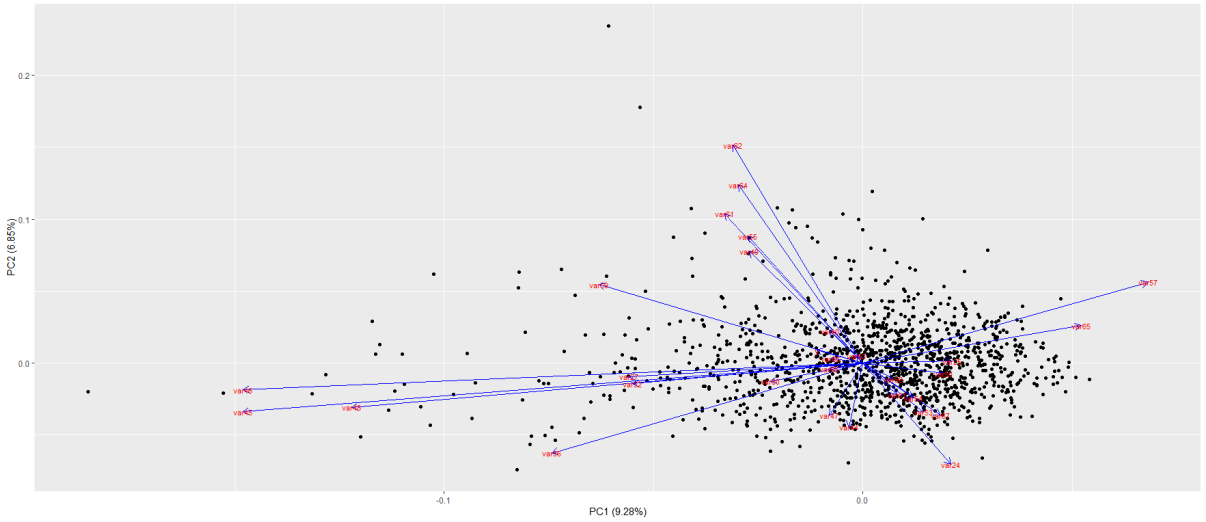


Figura 7

2.2 Código

```
# usando pacotes úteis
library(tidyverse)
library(corrplot)
library(ggcorrplot)
library(ggfortify)
library(ggplot2)

# mudando a pasta
setwd("~/porto-seguro-data-challenge")

# lendo os metadados com tipos das variáveis
metadata <- read_csv(file='metadata.csv')
view(metadata)


#---------------------------------------------------------------------------
# agora importaremos train,csv

# vendo o dataframe, percebe-se que na é representado como -999
# lendo train.csv com "-999" sendo trocado por na
train <- read.csv('train.csv', na.strings="-999")

view(train)

# vamos explorar um pouco melhor os dados de train
head(train)
tail(train)

# e agora ver a quantidade de valores únicos de cada coluna
```

```
sapply(train, function(x) length(unique(x)))


#------------------------------------------------------------------------
# vamos trocar as variáveis para os tipos apropriados

# começamos vendo quais são os tipos possíveis pelos metadados
unique(metadata$`Variavel tipo`)

# então criamos um vetor com os tipos na ordem correspondente às variáveis
rtypes = character(0)
for (i in 1:length(metadata$`Variavel tipo`)) rtypes[i] <- metadata$`Variavel tipo`[i]


# e trocamos as variáveis no vetor pelo atalho com c sendo char, n sendo numeric
counter <- 1
for (var in rtypes) {
  rtypes[counter] <- switch(var, "Qualitativo nominal" = "n",
                                 "Qualitativo ordinal" = "o",
                                 "Quantitativo discreto" = "d",
                                 "Quantitativo continua" = "c")

  counter <- counter + 1
}

print(rtypes)

# e convertemos cada coluna para seu tipo adequado
names <- colnames(train)

for (i in 1:length(rtypes)) {
  type <- rtypes[i]
  name <- names[i]

  if (type == "n") {
          train[[name]] <- as.factor(train[[name]])

  } else if (type == "o") {
          train[[name]] <- factor(train[[name]], order=TRUE)

  } #else if (type == "d") {
          #train[[name]] <- as.integer(train[[name]])}
          else {
          train[[name]] <- as.numeric(train[[name]])

  }
}
```

```
#------------------------------------------------------------------------
# Agora podemos fazer uma análise exploratória mais profunda

head(train)
tail(train)
str(train)

# Vejamos a estatistica descritiva, colocando em um dataframe para facilitar
# Aqui dividimos qualitativos de quantitativos, veremos sumario dos dois
qualitativeSummary <- data.frame(summary(train %>% select_if(is.factor)))
quantitativeSummary <- data.frame(summary(train %>% select_if(is.numeric)))


# de fato analisando
summary(train %>% select_if(is.factor))
summary(train %>% select_if(is.numeric))

# agora uma análise gráfica dos dados numéricos
boxplot(train %>% select_if(is.numeric))

# façamos um separado para var52, outro para as vars 40, 45, 46, 48 e um
# terceiro para todo o resto
grupo1 <- subset(train, select=c(var40, var45, var46, var48)) %>% select_if(is.numeric)
grupo2 <- subset(train, select=c(var52)) %>% select_if(is.numeric)
grupo3 <- subset(train, select=-c(var40, var45, var46, var48, var52)) %>%
select_if(is.numeric)

boxplot(grupo1, xlab='Característica', ylab='Valor', las=2,
        main='Grupo 1 Variáveis numéricas', col=rgb(214, 162, 232, maxColorValue = 255))

boxplot(grupo2, xlab='Característica', ylab='Valor', las=2,
        main='va52', col=rgb(254, 164, 127, maxColorValue = 255))

boxplot(grupo3, xlab='Característica', ylab='Valor', las=2,
        main='Grupo 3 Variáveis numéricas', col=rgb(154, 236, 219, maxColorValue = 255))


#------------------------------------------------------------------------
# para ver a correlação, precisamos excluir os valores faltantes
# checando a correlação entre variáveis numéricas
numcorr <- cor(train %>% select_if(is.numeric), use="complete.obs")

# agora plotando
corrplot(numcorr, type = "upper", order = "hclust",
        tl.col = "black", tl.srt = 45)

# vamos calcular o PCA para essas variáveis, para isso precisamos ignorar os NA
```

```
pca <- prcomp(na.omit(train %>% select_if(is.numeric)), center = TRUE,scale. = TRUE)
autoplot(pca, loadings = TRUE, loadings.colour = 'blue',
         loadings.label = TRUE, loadings.label.size = 3)

# vamos procurar correlação entre as variáveis categóricas, considerando y como
categórica
train2 <- train
train2$y <- as.factor(train2$y)


for (name in colnames(train2)) {
  print(chisq.test(table(train2[c(name, "y")])))
}
```