

# Desafio Data Science

*Carlos Tonhatti*

*September 27, 2018*

## Carregamento dos dados e verificação

```
# Carregar dados

Dados_brutos<-read.csv2("bank-full.csv")
head(Dados_brutos)

##   age         job marital education default balance housing loan contact
## 1 58 management married tertiary     no    2143    yes   no unknown
## 2 44 technician single secondary    no      29    yes   no unknown
## 3 33 entrepreneur married secondary    no       2    yes yes unknown
## 4 47 blue-collar married unknown    no    1506    yes   no unknown
## 5 33          unknown single unknown    no       1    no  no unknown
## 6 35 management married tertiary     no    231    yes   no unknown
##   day month duration campaign pdays previous poutcome  y
## 1   5   may      261        1     -1       0  unknown no
## 2   5   may      151        1     -1       0  unknown no
## 3   5   may       76        1     -1       0  unknown no
## 4   5   may       92        1     -1       0  unknown no
## 5   5   may      198        1     -1       0  unknown no
## 6   5   may      139        1     -1       0  unknown no

dim(Dados_brutos)

## [1] 45211     17

summary(Dados_brutos)

##      age                job            marital           education
##  Min.   :18.00   blue-collar:9732   divorced: 5207   primary   :6851
##  1st Qu.:33.00   management :9458    married :27214   secondary:23202
##  Median :39.00   technician :7597   single   :12790   tertiary  :13301
##  Mean   :40.94   admin.     :5171           unknown  :1857
##  3rd Qu.:48.00   services    :4154
##  Max.   :95.00   retired    :2264
##                  (Other)    :6835
##      default              balance            housing           loan           contact
##  no :44396   Min.   :-8019   no :20081   no :37967   cellular :29285
##  yes: 815   1st Qu.:    72   yes:25130   yes: 7244   telephone: 2906
##                  Median : 448
##                  Mean   : 1362
##                  3rd Qu.: 1428
##                  Max.   :102127
##
##      day                 month            duration           campaign
##  Min.   : 1.00   may     :13766   Min.   : 0.0   Min.   : 1.000
##  1st Qu.: 8.00   jul     : 6895   1st Qu.:103.0   1st Qu.: 1.000
##  Median :16.00   aug     : 6247   Median :180.0   Median : 2.000
```

```

##   Mean    :15.81   jun    : 5341   Mean    : 258.2   Mean    : 2.764
##  3rd Qu.:21.00   nov    : 3970   3rd Qu.: 319.0   3rd Qu.: 3.000
##  Max.    :31.00   apr    : 2932   Max.    :4918.0   Max.    :63.000
##                (Other): 6060
##      pdays      previous      poutcome      y
##  Min.   :-1.0    Min.   : 0.0000  failure: 4901  no  :39922
##  1st Qu.:-1.0    1st Qu.: 0.0000  other   : 1840  yes : 5289
##  Median :-1.0    Median : 0.0000  success: 1511
##  Mean   :40.2    Mean   : 0.5803  unknown:36959
##  3rd Qu.:-1.0    3rd Qu.: 0.0000
##  Max.   :871.0   Max.   :275.0000
##

```

## Questão 1 - Qual profissão tem mais tendência a fazer um empréstimo? De qual tipo?

Primeiro passo foi contar quantos empréstimos por profissão tomando cuidado para descontar os casos que possuem ambos tipos de empréstimos

```

# profissoes
job_per_category<-table(Dados_brutos$job)
personal_loan_job<-table(Dados_brutos$job,Dados_brutos$loan)
housing_loan_job<-table(Dados_brutos$job,Dados_brutos$housing)
both<-table(Dados_brutos$job, Dados_brutos$housing=="yes" & Dados_brutos$loan=="yes")

## Número de empréstimos por categoria excluindo aqueles que fazem ambos empréstimos
grouping_positive_loan_job<-data.frame(personal=personal_loan_job[,2], housing=housing_loan_job[,2])
Total_emprestimos<-rowSums(grouping_positive_loan_job)-both[,2]
Total_emprestimos_percent<-Total_emprestimos/job_per_category
Total_emprestimos_percent<-as.data.frame(Total_emprestimos_percent)

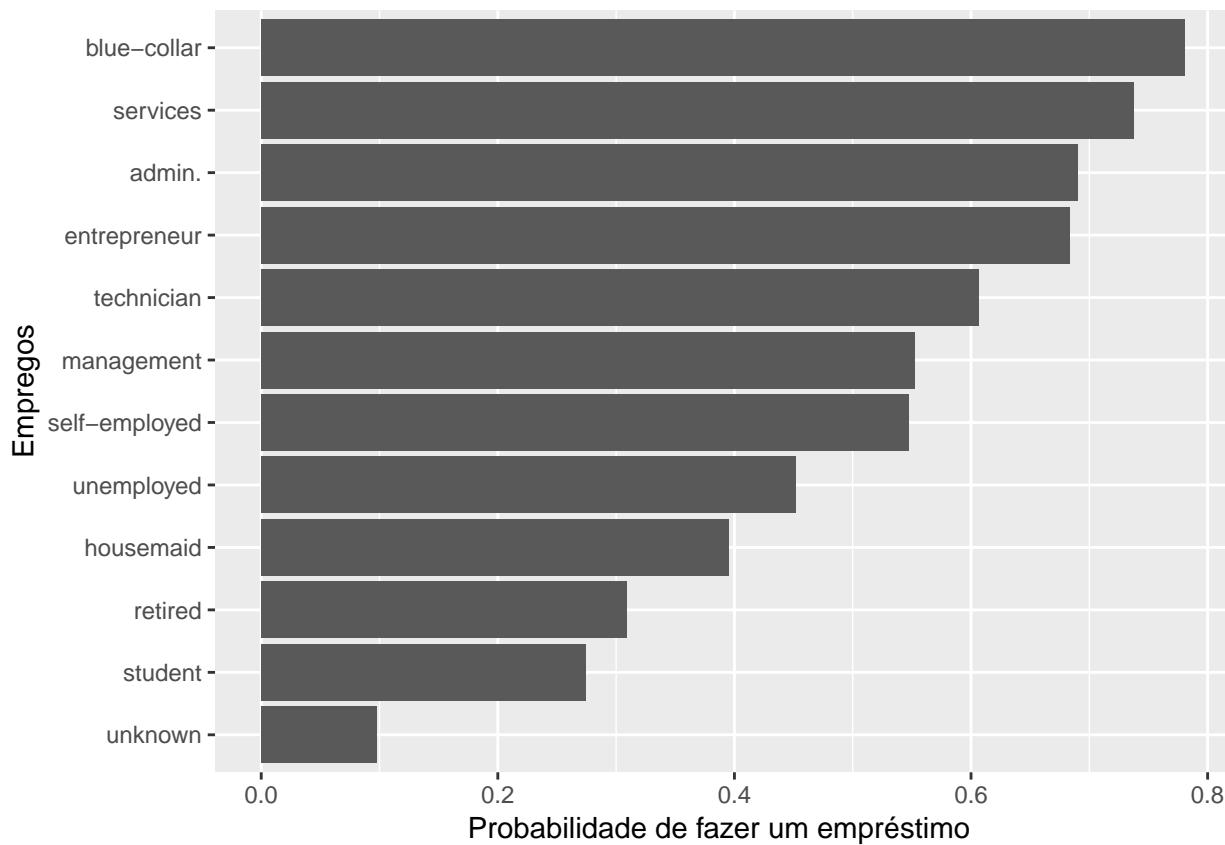
# Número de ambos empréstimos
(Both_category<-both[,2]/job_per_category)

##          admin. blue-collar entrepreneur housemaid management
## 0.116805260 0.116420058 0.140551446 0.048387097 0.074645802
##      retired self-employed        services     student technician
## 0.044611307 0.082330589 0.129754454 0.004264392 0.107410820
##      unemployed       unknown
## 0.048349962 0.006944444

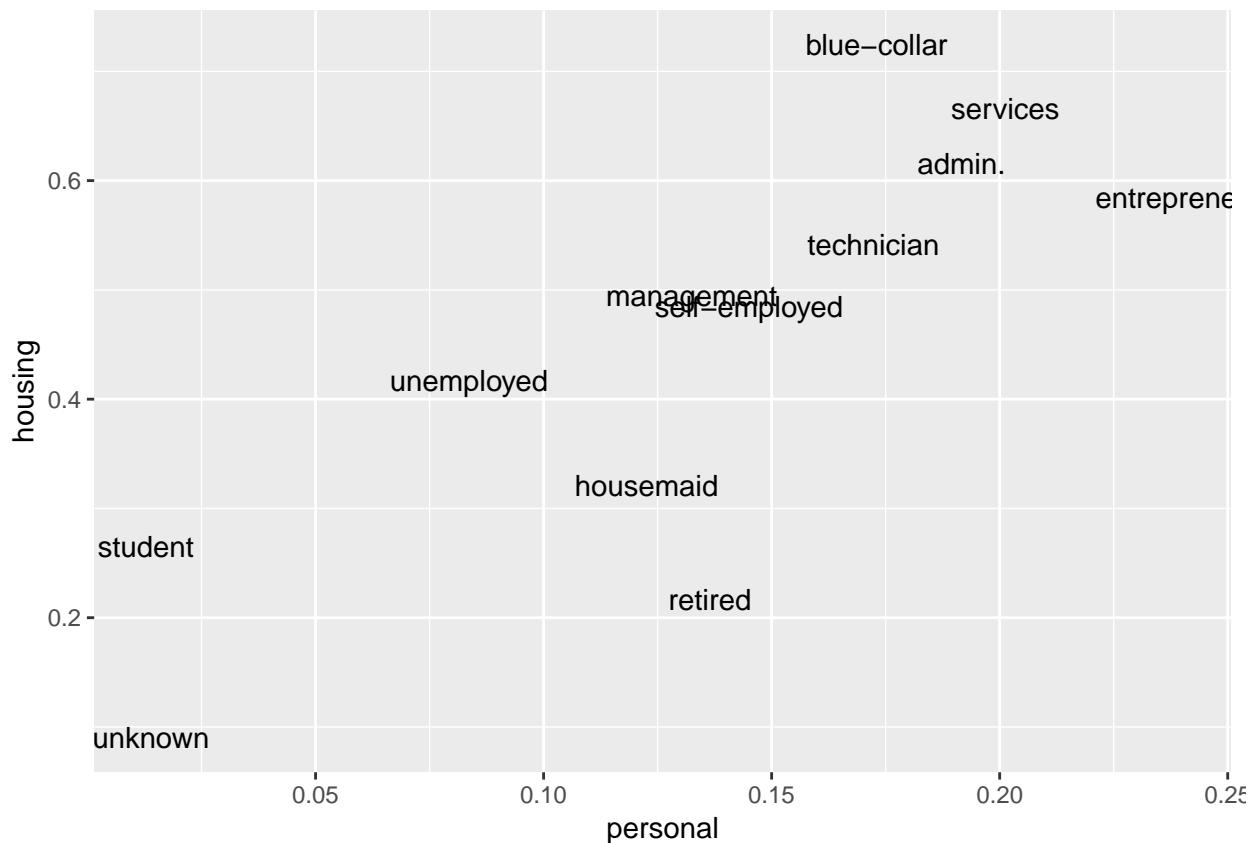
library(ggplot2)

ggplot(Total_emprestimos_percent,aes(reorder(Var1,Freq),y=Freq))+geom_col()+ coord_flip()+
  labs(x="Empregos", y="Probabilidade de fazer um empréstimo ")

```



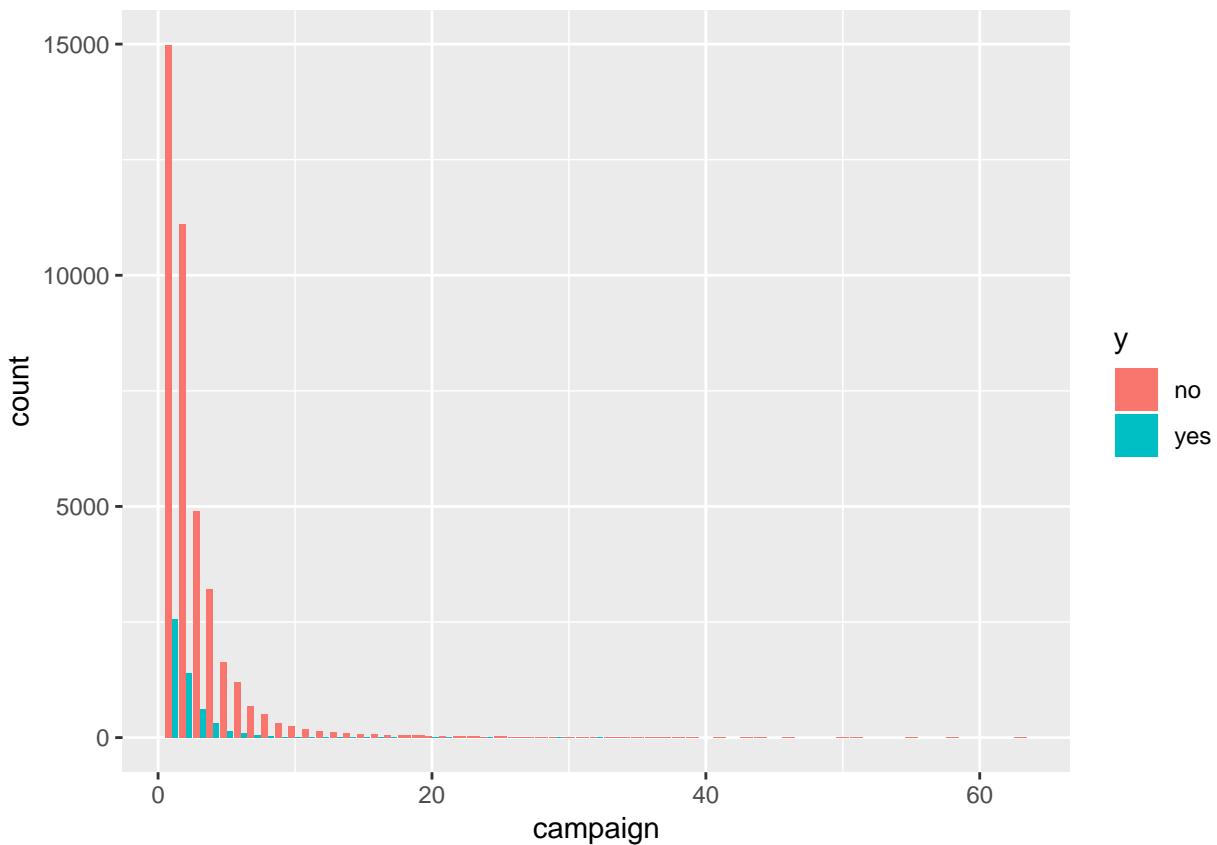
```
# Dividindo por tipo de empréstimo
grouping_percent<-grouping_positive_loan_job/job_per_category
ggplot(grouping_percent, aes(personal, housing)) + geom_text(aes(label=rownames(grouping_percent)))
```



A profissão que tem maior tendência de fazer um empréstimo é a de operário (blue-collar). No banco de dados analisado 78% dos operários fazem algum tipo de empréstimo. Sendo 72% dos operários tem empréstimo imobiliário, 13% tem empréstimo pessoal e 11% possui ambos empréstimos.

**Questão 2 - Fazendo uma relação entre número de contatos e sucesso da campanha quais são os pontos relevantes a serem observados?**

```
### Contatos
ggplot(Dados_brutos,aes(x=campaign, fill=y))+ geom_bar(position = "dodge")
```



Um grande número de contatos não aumenta a chance de sucesso do contato.

### Questão 3 - Baseando-se nos resultados de adesão desta campanha qual o número médio e o máximo de ligações que você indica para otimizar a adesão?

Conhecendo as distribuições do número de chamadas é possível indicar o número de ligações que deve ser feito.

```
Sucess<-Dados_brutos[Dados_brutos$y=="yes",]
Fail<-Dados_brutos[Dados_brutos$y=="no",]
summary(Sucess)
```

```
##      age              job          marital      education
##  Min.   :18.00   management :1301   divorced: 622   primary   : 591
##  1st Qu.:31.00   technician : 840    married :2755  secondary:2450
##  Median :38.00   blue-collar: 708    single  :1912  tertiary  :1996
##  Mean   :41.67   admin.     : 631           unknown  : 252
##  3rd Qu.:50.00   retired    : 516
##  Max.   :95.00   services   : 369
##                  (Other)    : 924
##      default        balance      housing      loan          contact
##  no :5237   Min.   :-3058   no :3354   no :4805   cellular :4369
##  yes: 52   1st Qu.: 210    yes:1935  yes: 484  telephone: 390
##                  Median : 733           unknown  : 530
```

```

##          Mean   : 1804
##          3rd Qu.: 2159
##          Max.   :81204
##
##      day       month      duration      campaign
##  Min.   : 1.00   may     : 925   Min.   : 8.0   Min.   : 1.000
##  1st Qu.: 8.00   aug     : 688   1st Qu.: 244.0  1st Qu.: 1.000
##  Median :15.00   jul     : 627   Median : 426.0  Median : 2.000
##  Mean   :15.16   apr     : 577   Mean   : 537.3  Mean   : 2.141
##  3rd Qu.:22.00   jun     : 546   3rd Qu.: 725.0  3rd Qu.: 3.000
##  Max.   :31.00   feb     : 441   Max.   :3881.0  Max.   :32.000
##          (Other):1485
##      pdays      previous      poutcome      y
##  Min.   :-1.00   Min.   : 0.00  failure: 618   no :    0
##  1st Qu.:-1.00   1st Qu.: 0.00  other   : 307   yes:5289
##  Median :-1.00   Median : 0.00  success: 978
##  Mean   :68.7   Mean   : 1.17  unknown:3386
##  3rd Qu.:98.0   3rd Qu.: 1.00
##  Max.   :854.0   Max.   :58.00
##
##  

summary(Fail)  

##
##      age           job      marital      education
##  Min.   :18.00  blue-collar:9024  divorced: 4585  primary   : 6260
##  1st Qu.:33.00 management: 8157  married  :24459   secondary:20752
##  Median :39.00 technician: 6757  single   :10878   tertiary :11305
##  Mean   :40.84  admin.     :4540
##  3rd Qu.:48.00  services    :3785
##  Max.   :95.00  retired    :1748
##          (Other)   :5911
##      default      balance      housing      loan      contact
##  no :39159   Min.   :-8019   no :16727   no :33162  cellular :24916
##  yes: 763   1st Qu.: 58     yes:23195  yes: 6760  telephone: 2516
##          Median : 417
##          Mean   : 1304
##          3rd Qu.: 1345
##          Max.   :102127
##
##      day       month      duration      campaign
##  Min.   : 1.00   may     :12841   Min.   : 0.0   Min.   : 1.000
##  1st Qu.: 8.00   jul     : 6268   1st Qu.: 95.0  1st Qu.: 1.000
##  Median :16.00   aug     : 5559   Median : 164.0  Median : 2.000
##  Mean   :15.89   jun     : 4795   Mean   : 221.2  Mean   : 2.846
##  3rd Qu.:21.00   nov     : 3567   3rd Qu.: 279.0  3rd Qu.: 3.000
##  Max.   :31.00   apr     : 2355   Max.   :4918.0  Max.   :63.000
##          (Other): 4537
##      pdays      previous      poutcome      y
##  Min.   :-1.00   Min.   : 0.0000  failure: 4283  no :39922
##  1st Qu.:-1.00   1st Qu.: 0.0000  other   : 1533  yes:    0
##  Median :-1.00   Median : 0.0000  success: 533
##  Mean   :36.42   Mean   : 0.5021  unknown:33573
##  3rd Qu.:-1.00   3rd Qu.: 0.0000
##  Max.   :871.00  Max.   :275.0000
##
```

```

# Achar os quantis
quantile(Sucess$campaign, probs = seq(0,1, 0.05))

##    0%    5%   10%   15%   20%   25%   30%   35%   40%   45%   50%   55%   60%   65%   70%
##    1     1     1     1     1     1     1     1     1     1     2     2     2     2     2
## 75% 80% 85% 90% 95% 100%
## 3     3     3     4     5    32

```

Dada as distribuições apresentadas indico que o número médio de ligações seja 2 pois é a média das ligações que se converteram em adesão. O número máximo que eu indico para obter 95% de aproveitamento é de 5 ligações.

## Questão 4 O resultado da campanha anterior tem relevância na campanha atual?

Nesta questão foi necessário limpar os dados pois a campanha previa apresenta duas respostas que não podem ser comparadas (“other”, “unknown”). A relação entre as duas variáveis foi realizada através do teste de contigência usando chi-quadrado.

```

# Selecionar apenas os dados de desempenho da campanha previa e atual
previa_atual<-data.frame(previa=Dados_brutos$poutcome, atual= Dados_brutos$y)

# Limpar os dados. retirar classes "outros" e "desconhecidos" pois não tem como comparar
ss<-ifelse(previa_atual$previa=="other" | previa_atual$previa=="unknown",F,T)
JustCamp<- previa_atual[ss,]
JustCamp<-data.frame(previa=factor(JustCamp$previa),atual=factor(JustCamp$atual))

# Constando as ocorrencias
table(JustCamp)

##          atual
## previa      no yes
##   failure  4283  618
##   success   533  978

# Teste do chi quadrado
X2test<-chisq.test(table(JustCamp))
X2test

##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: table(JustCamp)
## X-squared = 1675.1, df = 1, p-value < 2.2e-16

```

O teste de chi-quadrado mostrou significância estatística. Portanto, há relação entre as duas variáveis.

Sim, o resultado da campanha anterior tem relevância para o resultado da campanha atual. Uma falha na campanha passada pode resultar em uma falha na campanha atual.

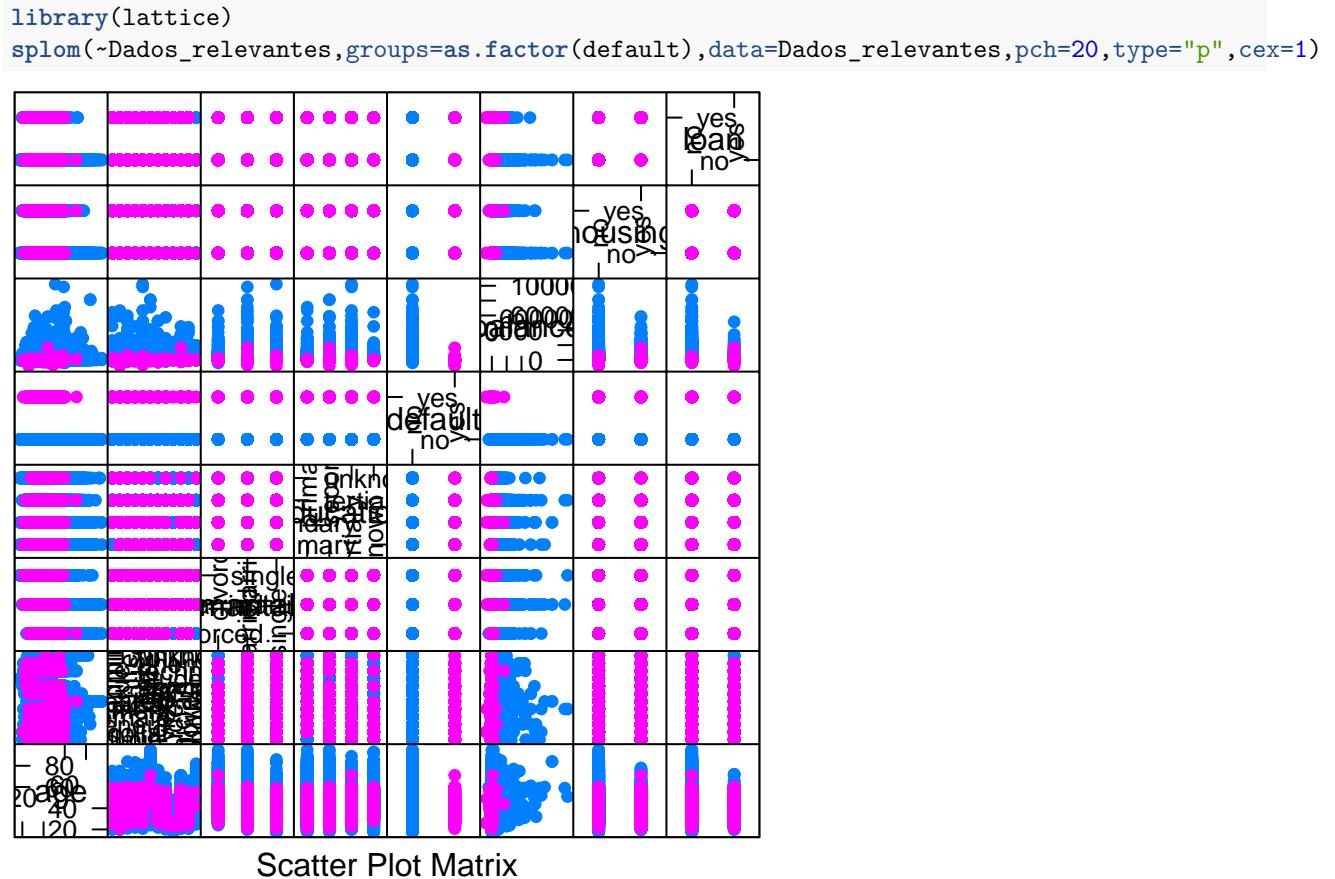
## Questão 5 - Qual o fator determinante para que o banco exija um seguro de crédito?

Para esta análise e a proxima eu exclui as variáveis sobre as campanhas de marketing pois creio que estas não tem relação á exigência de seguro ou o perfil de quem tem empréstimo imobiliário.

```
## Caracteristicas do emprestimo imobiliario  
# apenas dados que interessam. Excluir os dados sobre as campanhas de marketing  
  
Dados_relevantes<-Dados_brutos[,1:8]  
  
head(Dados_relevantes)
```

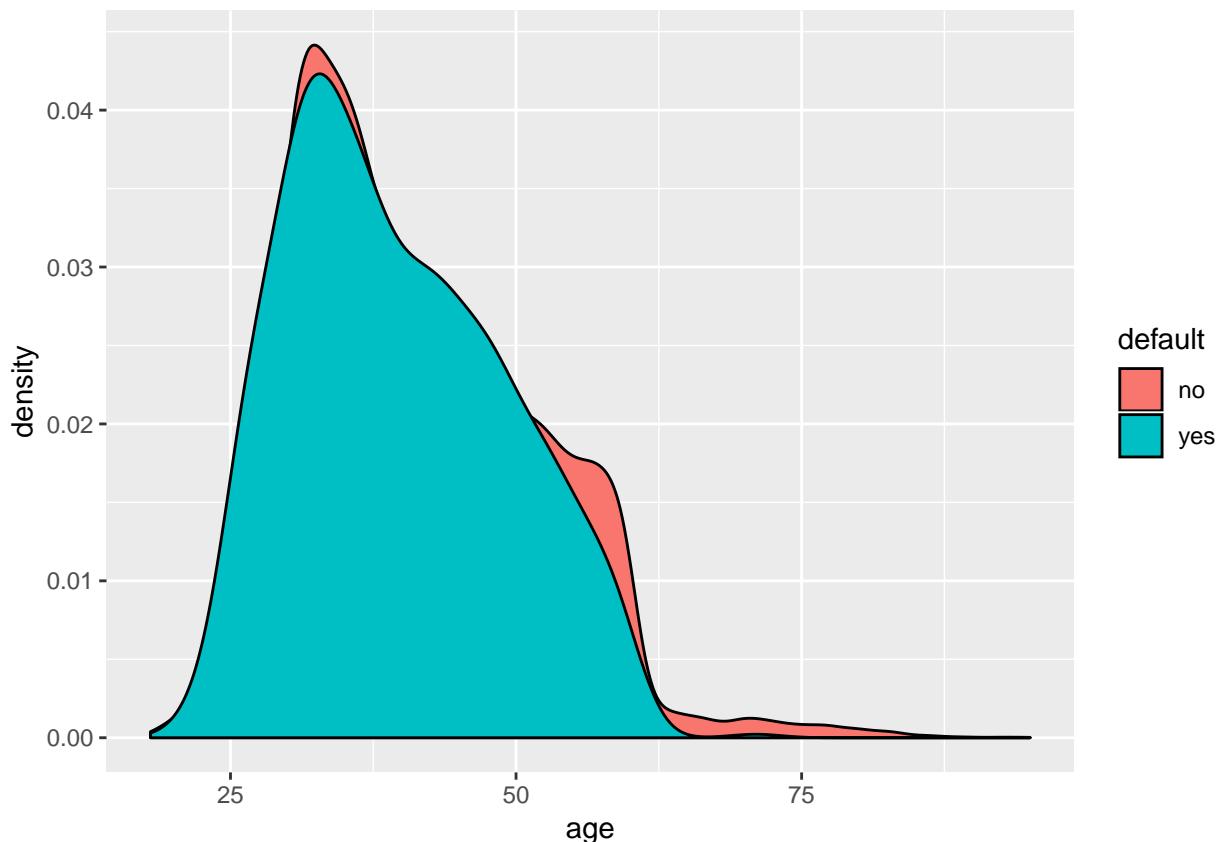
```
##   age      job marital education default balance housing loan  
## 1 58 management married tertiary no 2143 yes no  
## 2 44 technician single secondary no 29 yes no  
## 3 33 entrepreneur married secondary no 2 yes yes  
## 4 47 blue-collar married unknown no 1506 yes no  
## 5 33 unknown single unknown no 1 no no  
## 6 35 management married tertiary no 231 yes no
```

Para encontrar as tendências construi dois graficos um usando todos os dados que considero relevantes e outro para confirmar a tendência observada.



```
# Este grafico está salvo em arquivo separado TodasVar.pdf
```

```
ggplot(Dados_relevantes, aes(x=age, fill=default)) + geom_density()
```



O fator é a idade. O banco oferece o credito default sem necessitar de seguro do credito para quem esta em idade economicamente ativa (<60 anos). Embora a variável saldo (balance) seja melhor para classificar quem tem credito sem seguro não faz sentido o banco usar o saldo baixo como condição para não exigir seguro.

## Questão 6 - Quais são as características mais proeminentes de um cliente que possua empréstimo imobiliário?

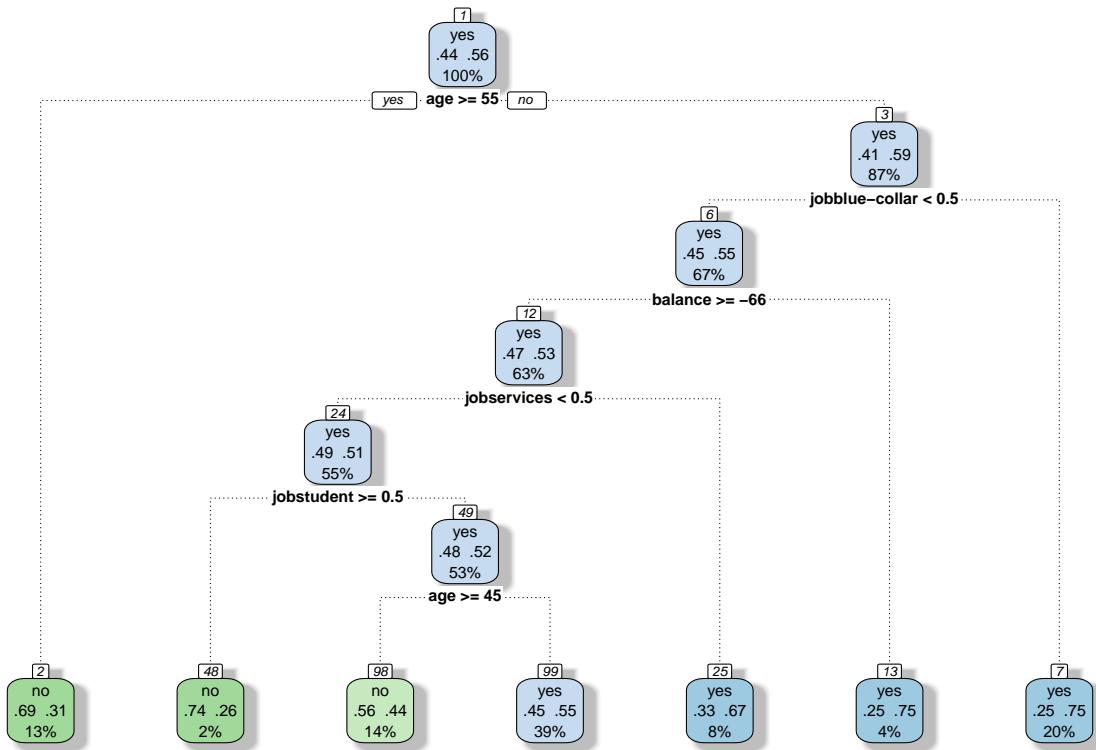
Para encontrar as principais características usei um algoritmo de classificação (árvore de decisão).

```
library(caret)

model<- train(housing ~ ., data = Dados_relevantes,
              method = "rpart")
library(rattle)

## Rattle: A free graphical interface for data science with R.
## Version 5.2.0 Copyright (c) 2006-2018 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

fancyRpartPlot(model$finalModel)
```



Rattle 2018–Sep–27 15:58:23 carlos

O perfil é: uma pessoa que tenha um empréstimo imobiliário tem menos que 55 anos, é operário e tem um saldo bancário maior que -66 euros