

Appus - Questão 1

Lucas Murtinho

April 12, 2016

Preparação

Pasta

```
pasta = 'E:/Drive/appus/'
```

Carregar módulos

```
library(lubridate) # datas
```

Ler dados

```
dados = read.csv2(paste0(pasta, "questao1.csv"),
                  colClasses=c(rep("factor", 2), "character", "factor",
                                "character", rep("factor", 2), "character",
                                rep("factor", 2), "factor", rep("numeric", 3)))
```

Limpeza de dados

Primeira vista

```
str(dados)
```

```
## 'data.frame': 2500 obs. of 15 variables:
## $ Desligamento : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 1 ...
## $ Ex.trainee : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
## $ Data.de.nascimento : chr "16/10/1963" "27/04/1963" "29/01/1960" "14/07/1941" ...
## $ Sexo : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
## $ Data.de.admissão : chr "1/27/03" "10/05/1992" "01/03/1989" "06/05/1991" ...
## $ Cargo : Factor w/ 885 levels "1","10","100",...: 1 112 223 334 444 555
## $ Área : Factor w/ 8 levels "Administrativo",...: 7 1 1 7 3 7 7 3 7 3 .
## $ Salário.mensal.médio : chr "R$ 5.080,00" "R$ 12.498,33" "R$ 9.966,67" "R$ 3.795,00"
## $ Posição.critica : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 2 ...
## $ Gestor : Factor w/ 820 levels "1","10","100",...: 1 112 223 334 445 334
## $ Avaliação.de.desempenho : Factor w/ 4 levels "BOM","EXCELENTE",...: 1 1 1 4 4 1 4 1 1 1
## $ Distância.residência.trabalho..Km.: num 25 29 10 25 14 13 10 19 24 2 ...
## $ Tempo.deslocamento..min. : num 17 20 6 17 9 9 6 13 16 1 ...
## $ Turnover.mercado : num 0.03 0.17 0.23 0.11 0.02 0.16 0.04 0.13 0.09 0.12 ...
## $ X : Factor w/ 1 level "": 1 1 1 1 1 1 1 1 1 1 ...
```

Apagar última coluna

```
dados = dados[,-ncol(dados)]
```

Data de nascimento

```
nasc_1 = as.Date(dados$Data.de.nascimento, "%d/%m/%Y")
nasc_2 = as.Date(dados$Data.de.nascimento, "%m/%d/%y")

dados$Data.de.nascimento = as.Date(nasc_1)

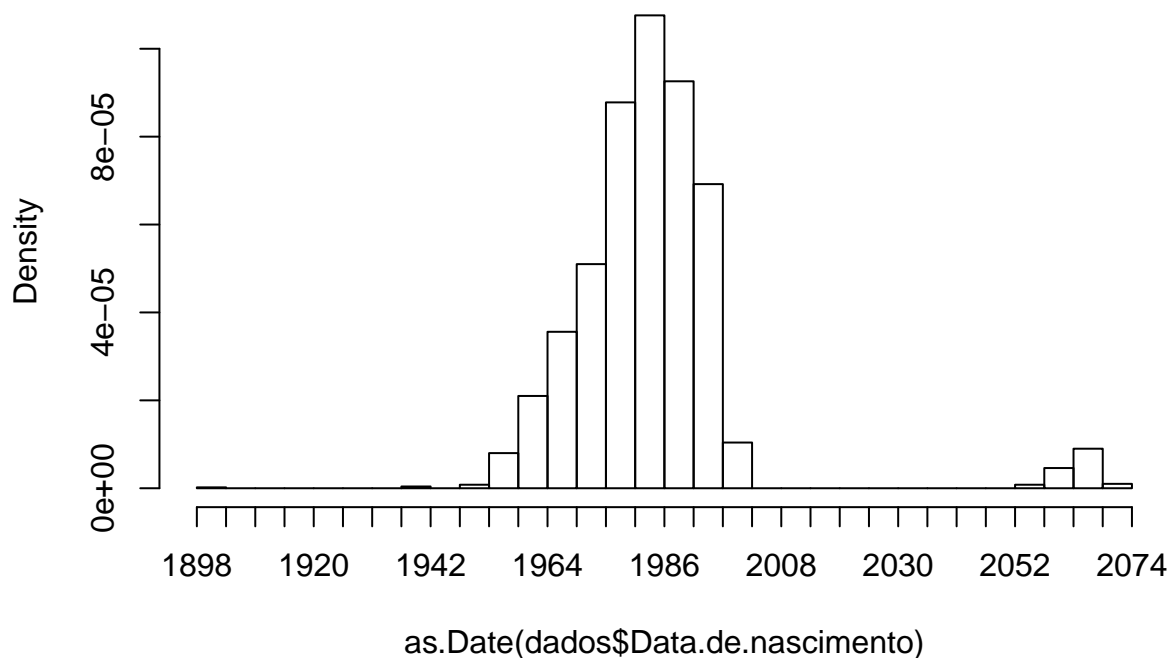
dados$Data.de.nascimento[is.na(dados$Data.de.nascimento)] =
  as.Date(nasc_2[is.na(dados$Data.de.nascimento)])

summary(dados$Data.de.nascimento)
```

```
##           Min.         1st Qu.         Median         Mean         3rd Qu.
## "1900-10-25" "1976-02-22" "1983-02-20" "1984-04-25" "1990-02-13"
##           Max.
## "2068-12-15"
```

```
hist(as.Date(dados$Data.de.nascimento), breaks=30)
```

Histogram of as.Date(dados\$Data.de.nascimento)



```

year(dados$Data.de.nascimento) =
  year(dados$Data.de.nascimento) -
  ifelse(year(dados$Data.de.nascimento)>2000,
        100, 0)

summary(dados$Data.de.nascimento)

```

```

##           Min.         1st Qu.         Median         Mean         3rd Qu.
## "1900-10-25" "1974-09-14" "1982-05-26" "1981-03-27" "1989-01-19"
##           Max.
## "2000-10-25"

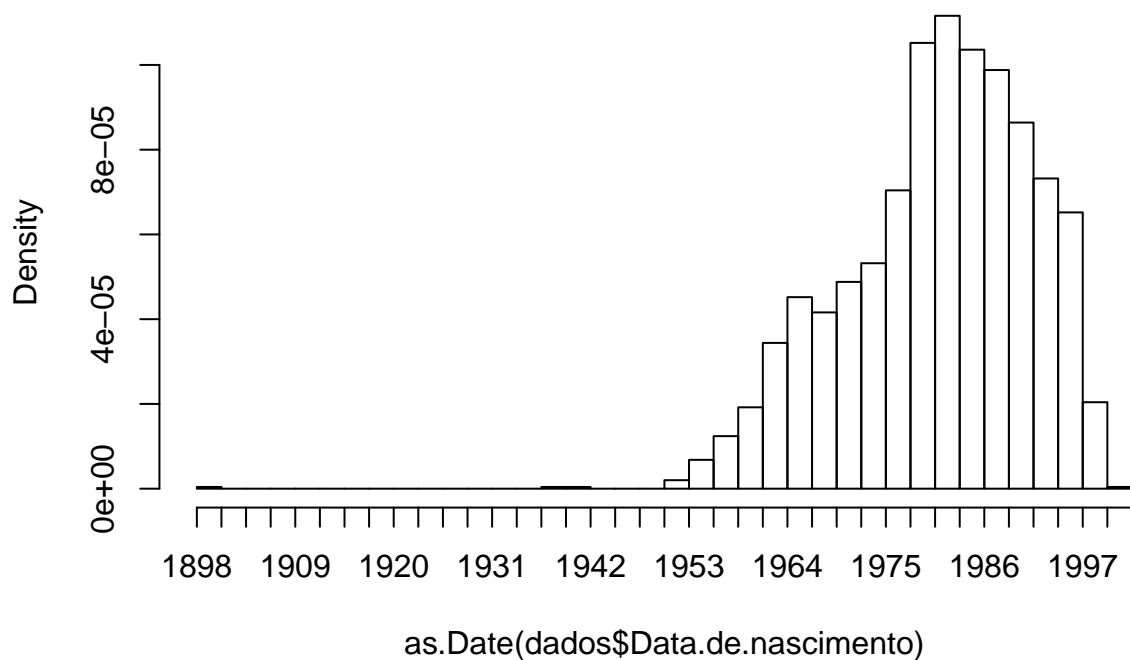
```

```

hist(as.Date(dados$Data.de.nascimento), breaks=30)

```

Histogram of as.Date(dados\$Data.de.nascimento)



```

sum(year(dados$Data.de.nascimento)<1930)

```

```

## [1] 1

```

```

nascimento_errado = dados[which(year(dados$Data.de.nascimento)<1930),]

dados = dados[-which(year(dados$Data.de.nascimento)<1930),]
dim(dados)

```

```

## [1] 2499  14

```

Data de admissão

```
admis_1 = as.Date(dados$Data.de.admissão, "%d/%m/%Y")
admis_2 = as.Date(dados$Data.de.admissão, "%m/%d/%y")

dados$Data.de.admissão = as.Date(admis_1)

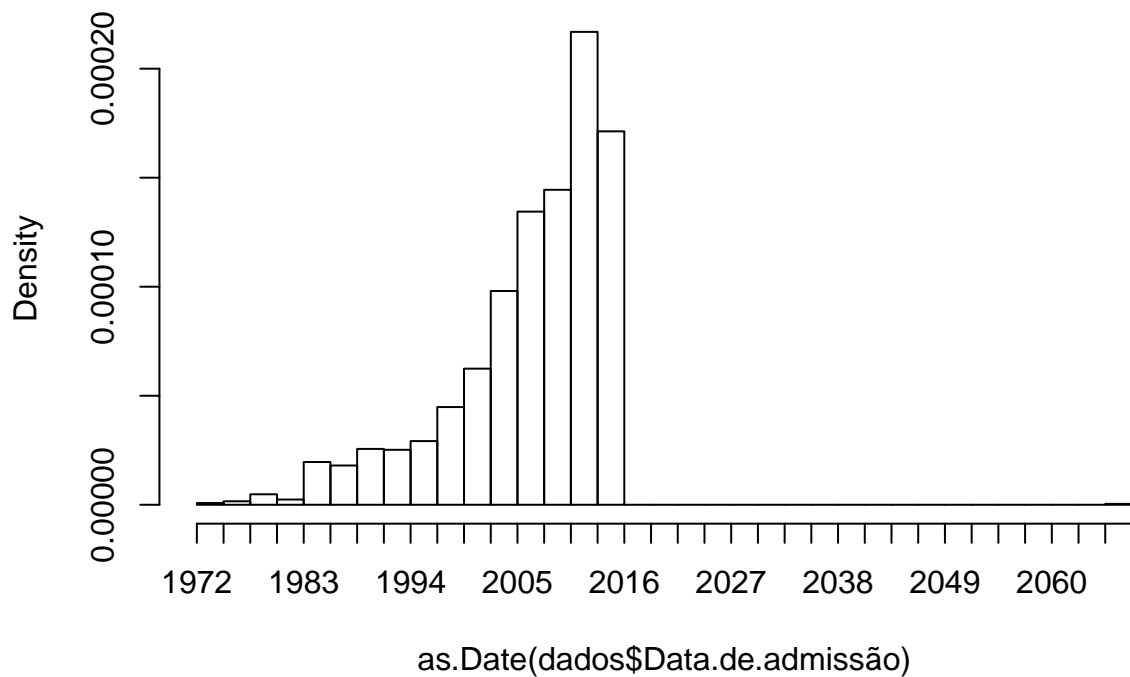
dados$Data.de.admissão[is.na(dados$Data.de.admissão)] =
  as.Date(admis_2[is.na(dados$Data.de.admissão)])

summary(dados$Data.de.admissão)
```

```
##           Min.       1st Qu.         Median         Mean       3rd Qu.
## "1973-01-08" "2003-02-18" "2008-11-08" "2006-11-27" "2013-01-04"
##           Max.
## "2068-05-15"
```

```
hist(as.Date(dados$Data.de.admissão), breaks=30)
```

Histogram of as.Date(dados\$Data.de.admissão)



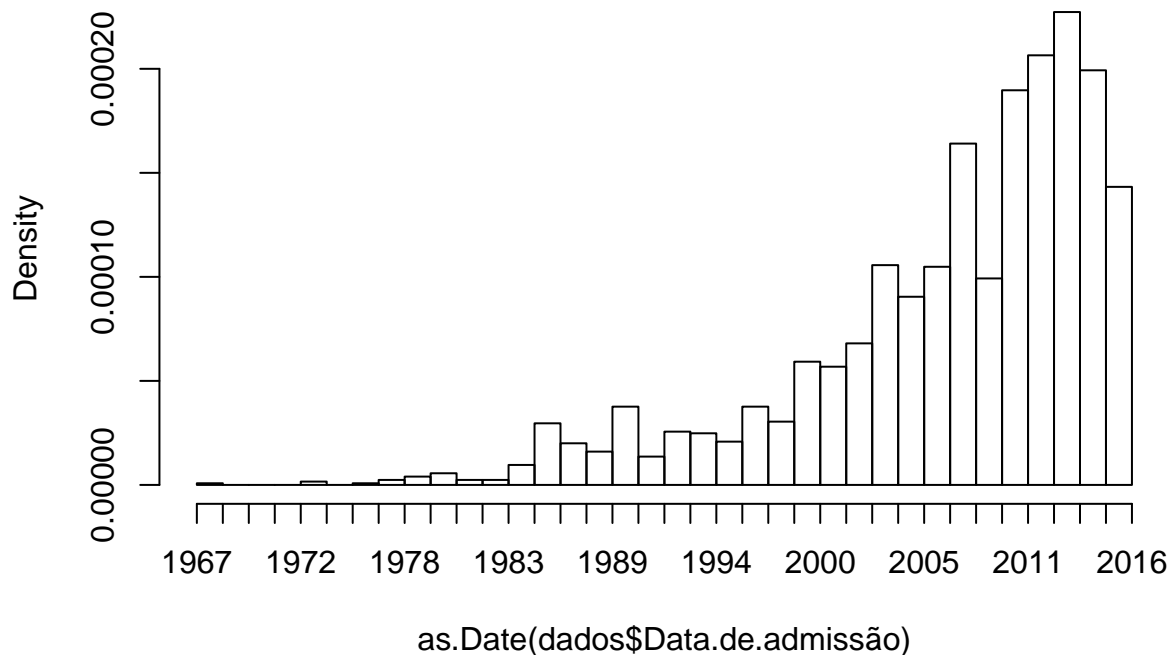
```
year(dados$Data.de.admissão) =
  year(dados$Data.de.admissão) -
  ifelse(year(dados$Data.de.admissão) >= 2068,
    100, 0)
```

```
summary(dados$Data.de.admissão)
```

```
##           Min.          1st Qu.          Median          Mean          3rd Qu.
## "1968-05-15" "2003-02-15" "2008-11-08" "2006-11-12" "2013-01-04"
##           Max.
## "2015-12-21"
```

```
hist(as.Date(dados$Data.de.admissão), breaks=30)
```

Histogram of as.Date(dados\$Data.de.admissão)



```
table(dados$Área[(dados$Data.de.nascimento+years(10))>dados$Data.de.admissão])
```

```
##
## Administrativo      Aprendiz      Coordenação      Estagiário      Executivo
##           15           0           4           0           1
##      Gerência      Operacional      Trainee
##           1           30           0
```

```
admissao_errada = dados[which(dados$Data.de.nascimento+years(10)>dados$Data.de.admissão),]
```

```
dados = dados[-which(dados$Data.de.nascimento+years(10)>dados$Data.de.admissão),]
```

Salário mensal médio

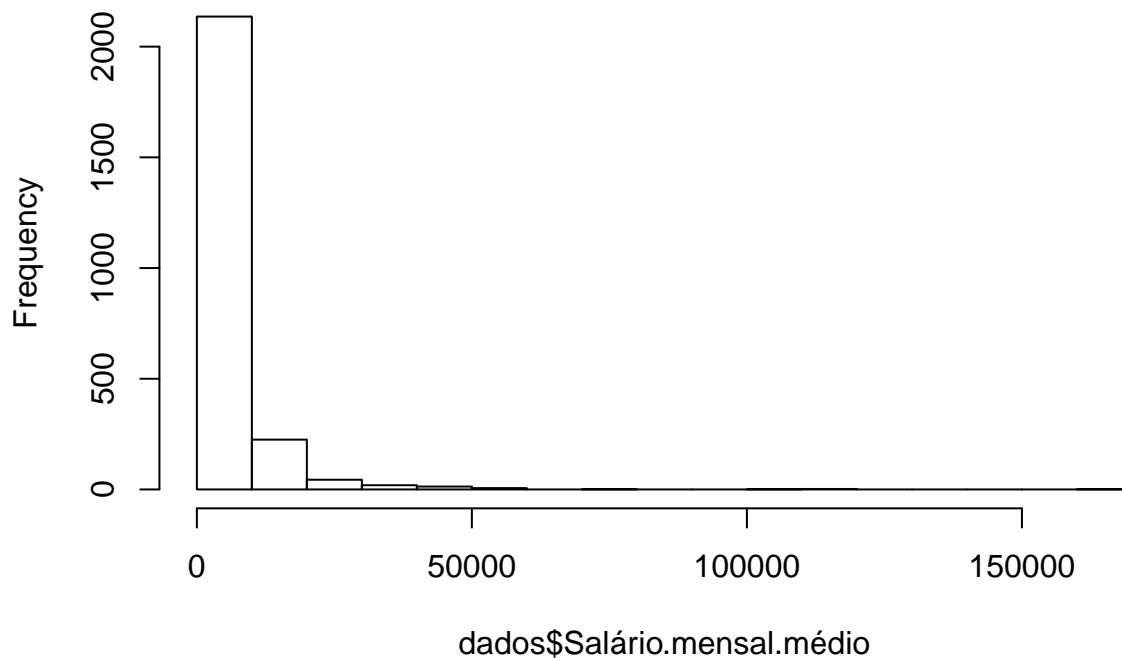
```
dados$Salário.mensal.médio = gsub("[^0-9,]", "", dados$Salário.mensal.médio)
dados$Salário.mensal.médio = gsub(",", ".", dados$Salário.mensal.médio)
dados$Salário.mensal.médio = as.numeric(dados$Salário.mensal.médio)

summary(dados$Salário.mensal.médio)
```

```
##      Min.   1st Qu.   Median     Mean   3rd Qu.     Max.
##    616.7   3236.0   4422.0   6456.0   6625.0  166800.0
```

```
hist(dados$Salário.mensal.médio)
```

Histogram of dados\$Salário.mensal.médio



```
sum(dados$Salário.mensal.médio>50000)
```

```
## [1] 11
```

```
dados$Área[dados$Salário.mensal.médio>50000]
```

```
## [1] Gerência      Gerência      Coordenação  Executivo    Coordenação
## [6] Executivo     Executivo     Executivo     Executivo     Gerência
## [11] Gerência
## 8 Levels: Administrativo Aprendiz Coordenação Estagiário ... Trainee
```

AValiação de desempenho

```
head(dados$Avaliação.de.desempenho)
```

```
## [1] BOM    BOM    BOM    OTIMO OTIMO BOM  
## Levels: BOM EXCELENTE INSATISFATORIO OTIMO
```

```
table(levels(dados$Avaliação.de.desempenho))
```

```
##  
##          BOM          EXCELENTE INSATISFATORIO          OTIMO  
##          1            1            1            1
```

```
dados$Avaliação.de.desempenho =  
  factor(dados$Avaliação.de.desempenho,  
        levels=c("INSATISFATORIO",  
                  "BOM", "OTIMO",  
                  "EXCELENTE"),  
        ordered=TRUE)
```

```
head(dados$Avaliação.de.desempenho)
```

```
## [1] BOM    BOM    BOM    OTIMO OTIMO BOM  
## Levels: INSATISFATORIO < BOM < OTIMO < EXCELENTE
```