R Notebook

Pŕe-processamento em R com a base de dados *bancoufba.csv*

Agenda

- Introdução
- Limpesa
- Exploração
- Transformação
- Redução de Dimensionalidade

```
# lendo a base de dados
banco = read.csv('../data/bancoufba.csv')

# listando o nome dos atributos
names(banco)
```

```
## [1] "cpf"              "sexo"             "salario"
## [4] "estado"           "poupanca"         "altura"
## [7] "peso"             "total.emprestimo" "financiamento"
```

```
# visualizando alguns atributos
head(banco)
```

```
##         cpf sexo salario estado poupanca altura   peso total.emprestimo
## 1 741132012    M 5124.00     MA -1000.00   1.93  85.62                0
## 2 246313940    F 4772.45     SP 15012.90   1.84  99.12               NA
## 3 431872706    F 5001.80     BA     0.00   1.66 104.39                0
## 4 127070574    F 5279.32     DF     0.00   1.73  72.57                0
## 5 620680271    M 5327.52     RS  8172.68   2.05  97.89                0
## 6 939260408    M 5229.73     SE     0.00   1.69  74.25                0
##   financiamento
## 1             S
## 2             S
## 3             N
## 4             S
## 5             S
## 6             S
```

```
# verificando os níveis de um determinado atributo
levels(banco$sexo)
```

```
## [1] "F" "M"
```

```
levels(banco$estado)
```

```
##  [1] "AC" "AL" "AM" "AP" "BA" "CE" "DF" "ES" "GO" "MA" "MG" "MS" "MT" "PA"
## [15] "PB" "PE" "PI" "PR" "RJ" "RN" "RO" "RR" "RS" "SC" "SE" "SP" "TO"
```

```
# média -  no exemplo todos os pesos das instâncias
mean(banco$peso)
```

```
## [1] 92.8259
```

```
# mediana
median(banco$peso)
```

```
## [1] 95.18
```

```
# mínino e maximo
min(banco$peso)
```

```
## [1] 55.55
```

```
max(banco$peso)
```

```
## [1] 129.89
```

```
# Moda - o valor mais frequente nas minhas instâncias - rever isso.
sort(table(banco$estado), decreasing = T)[1]
```

```
## AC
## 13
```

```
# Resumo dos Dados no atributo Salário
summary(banco$salario)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    4000    4581    5094    5799    5489   25720       5
```

```
# Resumo dos dados em toda a base
summary(banco)
```

```
##       cpf            sexo       salario          estado       poupanca
##  Min.   : 19593780   F: 92   Min.   : 4000   AC     : 13   Min.   :-1000
##  1st Qu.:301568359   M:118   1st Qu.: 4581   MS     : 13   1st Qu.:    0
##  Median :571129336           Median : 5094   SC     : 12   Median :    0
##  Mean   :531321108           Mean   : 5799   MA     : 11   Mean   : 3908
##  3rd Qu.:766159015           3rd Qu.: 5489   MG     : 11   3rd Qu.: 7306
##  Max.   :999966843           Max.   :25716   SP     : 11   Max.   :19529
##                              NA's   :5       (Other):139   NA's   :5
##      altura            peso        total.emprestimo financiamento
##  Min.   : 1.400   Min.   : 55.55   Min.   :  0.0    N   :121
##  1st Qu.: 1.620   1st Qu.: 77.97   1st Qu.:  0.0    S   : 84
##  Median : 1.785   Median : 95.18   Median :  0.0    NA's:  5
##  Mean   : 2.240   Mean   : 92.83   Mean   :154.0
##  3rd Qu.: 1.978   3rd Qu.:106.23   3rd Qu.:443.3
##  Max.   :11.980   Max.   :129.89   Max.   :600.0
##                                    NA's   :5
```

```
# calculando a variãnça - dado peso é atributo Racional
var(banco$peso)
```

```
## [1] 393.9874
```

```
# calculando o desvio padrão - ate agora todas essas métricas estatísticas o R ja tem as funções
sd(banco$peso)
```

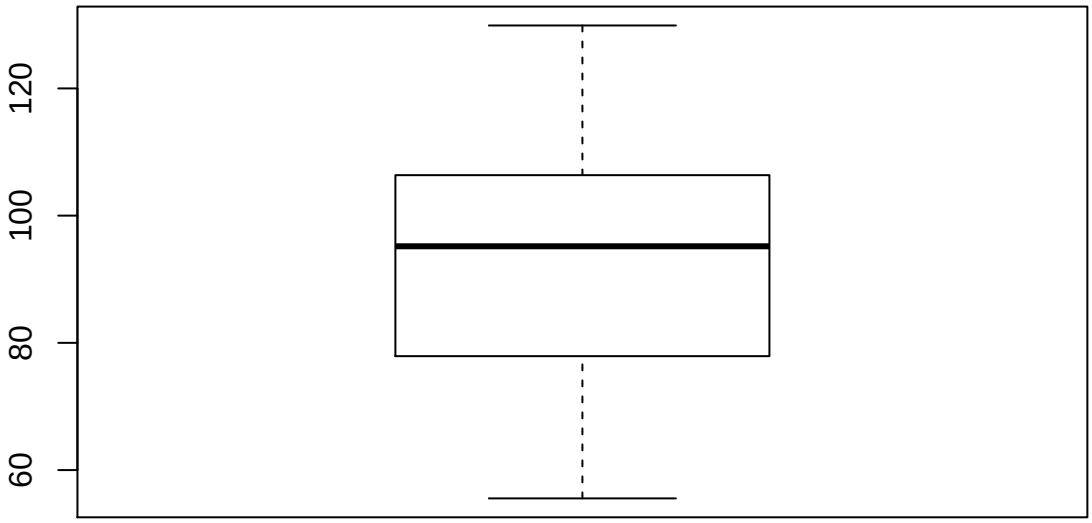
```
## [1] 19.84912
```

Visualização

Boxplot

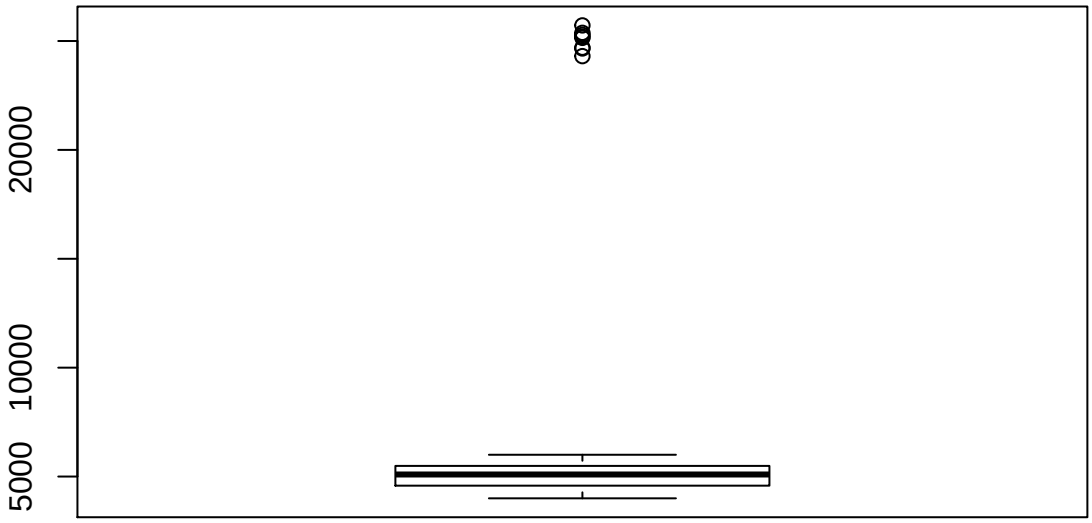
Para o pré processamento dos dados podemos visualiza-los os dados para compreender uma maior abstração e fazer melhores induções sobre os mesmo para então conferir se uma determinada hipótesi aplica aos dados. Nessa etapa de pré-processamento podemos utilizar o **Boxplot** para avaliar a distribuição empírica dos dados. O **Boxplot** é formado pelo primeiro, terceiro quartil e pela mediana.

```
# boxplot so aceita valores numericos ? por usar quartis e medianas, pode ser que sim
boxplot(banco$peso)
```



As hastes inferiores e superiores se estendem, respectivamente, do quartil inferior até o menor valor não inferior ao limite inferior e do quartil superior até o maior valor não superior ao limite superior. Pontos foras desses limites são considerados valores discrepantes (**outliers**). Como exemplos temos:

```
boxplot(banco$salario)
```

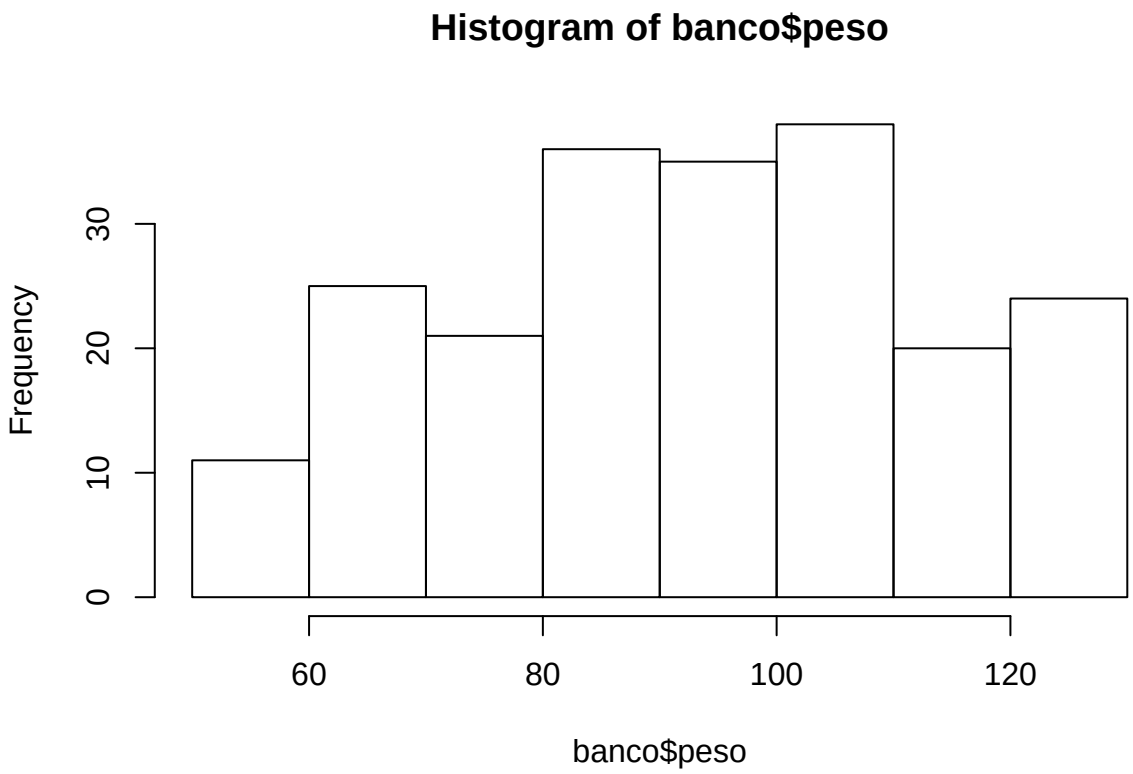


Histograma

Distribuição de frequências é um agrupamento de dados em classes contabilizando o número de ocorrências em cada clases. O número de ocorrências de uma determinada classe recebe o nome de frequência absoluta.

O *Objetivo* é apresentar os dados de uma maneira mais concisa e que nos permita extrair informações sobre seu comportamento.

```
hist(banco$peso)
```



Transformando dados e Reduzindo a dimensionalidade

Transformação

Várias técnicas de AM são limitadas ao tipo dos atributos: apenas valores numéricos ou apenas valores simbólicos. RNA e SVM são exemplos de técnicas que lidam apenas com dados numéricos. Pode-se fazer a conversão de valores como solução.

Conversão Simbólico-Numérico

Atributo nominal com dois valores que representam presença ou ausência de uma característica. Pode-se substituir por um dígito binário.

Atributo nominal com dois valores que representam relação de ordem. Pode-se substituir por um dígito binário

```
# como exemplo o atributo sexo, que é nominal(simbólico), e queremos trabalhar com dados numéricos.
levels(banco$sexo)
```

```
## [1] "F" "M"
```

```
# procura as instância onde encontrar o atributo 'F'
which(banco$sexo=="F")
```

```
##  [1]   2   3   4   7   8  10  12  17  18  19  20  22  23  24  26  28  34
## [18]  35  38  43  45  48  51  53  58  62  66  68  69  72  73  74  76  77
## [35]  81  82  84  90  91  93  94  96 100 104 109 114 116 118 119 120 121
## [52] 122 123 124 125 129 132 133 135 136 137 138 142 148 149 150 151 152
## [69] 153 155 157 159 160 161 163 164 168 171 172 173 175 176 180 187 188
```

```
## [86] 192 193 196 202 203 204 210
```

```r
# vetor que armazena os atributos do sexo
temp<-as.vector(banco$sexo)

# faz a tranformação de valores - 0 para M, 1 para F
temp[which(temp=="M")] = "0"
temp[which(temp=="F")] = "1"

#atribui na minha base os atributos tranformados como inteiros
banco$sexo <- as.integer(temp)

# lista os novos valores dos atributos
head(banco$sexo)
```

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

Conversão Numérico-Simbólico

- Uma parcela dos algortimos de classificação e de associaçãi foram desenvolvidos para trabalhar com valores quantitativos. Atributo quantitativo do tipo discreto ou binário, com apenas dois valores.
  - Conversão Trivial: associar um nomea cada valor.
- Atributos quantitativos numéricos
  - Discretização : transformação de valores numéricos em intervalos.
  - Existem vários métodos de discretização, o mais simples é a média.

Tranformação dos atributos numéricos

- Transformar um valor numérico em outro valor numérico.
- Isso ocorre quando os limites inferior e superior de valores dos atributos são muito diferentes.
- Ou quando vários atributos estão em escalas diferentes.
- A transformação é necessária para evitar que um atributo predomine sobre outro. Pode se usar por exemplo a *normalização* dos dados.

DATASETS

- Machine Learning Repository
- UCI KDD Archive
- Delve