UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
IMD – INSTITUTO METRÓPOLE DIGITAL
Disciplina: IMD0033 – PROBABILIDADE – 24N12
Professora: Ismenia

LISTA DE EXERCÍCIOS 1

Atividade para pontuação na Unidade 1 – Exercícios
(Valor 3,0 pontos)

1. Objetivo

Trabalhar conceitos variados de Estatística Descritiva, por meio de análise de uma base de dados reais multivariada.

2. Conhecimentos necessários

Tipos de variáveis (Qualitativa, Quantitativa e suas classificações), Medidas de Tendência Central e de Variabilidade, Gráficos e Tabelas.

3. Informações preliminares:

Esta base de dados, denominada de **adult** é tradicional na literatura de Machine Learning. Estes dados foram extraídos de uma base de dados do Bureau do Censo Americano (http://www.census.gov/ftp/pub/DES/www/welcome.html) do ano de 1994 e tem como doadores Ronny Kohavi e Barry Becker,(Data Mining and Visualization Silicon Graphics). Eles disponibilizam o e-mail: ronnyk@sgi.com para questionamentos.

São 32.561 observações e 15 variáveis.

O objetivo dos pesquisadores inicialmente era executar uma tarefa de predição para determinar se uma pessoa recebia acima de 50k de renda por ano.

Variáveis:

age: continuous.
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt: continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num: continuous.
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex: Female, Male.
capital-gain: continuous.
capital-loss: continuous.

```
hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany,
Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran,
Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal,
Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia,
Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador,
Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

class: >50K, <=50K.
```

Roteiro

Execute o comando de importação do arquivo adult.csv no R-Studio e logo após os comandos abaixo para nomear suas colunas (associar os nomes às variáveis), verificar a dimensão da base de dados e visualizar suas seis primeiras linhas.

```
names(adult)<-c("age","workclass","fnlwgt","education","education-num","marital-
status","occupation","relationship","race","sex","capital-gain","capital-loss","hours-per-
week","native-country","class" )
dim(adult)
head(adult)
```

Você deve executar comandos que permitam obter informações acerda das variáveis e de seu comportamento, como por exemplo:

```
summary(adult)
table(adult$class)
table(adult$sex)
table(adult$race)
table(adult$relationship)
table(adult$class,adult$sex)
boxplot(adult$age~adult$class,col=c(3,4),main="Idade vs class",sub="class")
boxplot(adult$`hours-per-week`~adult$class,col=c(3,4),main="Horas por semana vs
   class",sub="class")
boxplot(adult$age~adult$sex,col=c(3,4),main="Age vs class",sub="class")
hist(adult$`hours-per-week`)
hist(adult$age)
plot(table(adult$education))
plot(table(adult$race))
plot(table(adult$relationship))
pie(table(adult$class))
pie(table(adult$sex))
hist(adult$`capital-gain`-adult$`capital-loss`)
summary(adult$`capital-gain`-adult$`capital-loss`)
```

e outros que você julgar conveniente (substitua o nome das variáveis por aquelas que você tiver interesse em analisar)

<u>Responda</u>

1) (Valor: 1,5 ponto) O que você consegue dizer a respeito dessa base de dados? Como são as pessoas que estão inseridas nessa base? Escreva um texto com os resultados que você julgou mais importantes.

2) (Valor: 1,5 ponto) Os pesquisadores de começaram a trabalhar com essa base de dados tinham por objetivo identificar os fatores que poderiam 'prever' quem recebia mais de US$ 50.000 dólares por ano (variável class). Que suposições você faria a respeito? Qual o perfil das pessoas que recebem mais de US$ 50.000 dólares por ano de acordo com os dados dessa base?