Resumo da atividade da disciplina exploração e mineração de dados

Randerson Douglas R. Santos

Universidade Federal de Alagoas (UFAL)
Instituto de Computação (IC)

rdrs@ic.ufal.br

1. Atividade:

1. Escolher um dataset de sua preferência.
2. Elaborar algumas perguntas a serem feitas ao dataset.
3. Explorar os dados, visando responder as perguntas elaboradas.
4. Escolher um dataset de sua preferência.

2. Dataset:

O dataset escolhido é O Framingham Heart Study, que é um estudo prospectivo de longo prazo da etiologia da doença cardiovascular entre uma população de indivíduos vivos livres na comunidade de Framingham, Massachusetts. Contendo algumas variáveis que supostamente possuem uma ligação com doença cardiovascular.

Data Analisys	
Dataset:	https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset

3. Questões de Análise

Questões	Variáveis
3.1 – Qual a relação de pacientes que hipertensos com tendência a ter um AVC?	sysbp3 diabp3
3.2 – Qual a relação (quantidade) de pacientes por sexo que possuem/não possuem diabetes?	sex3 diabetes3
3.3 – Qual risco cardiovascular mediante a analise de níveis de colesterol (pessoas mais sujeitas – por sexo)?	sex3 ldlc3 hdlc3

4. Análise

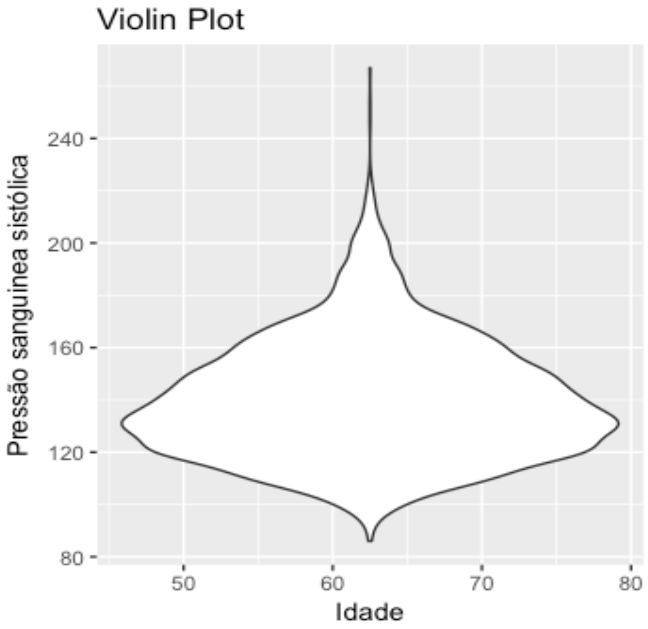
Verificando o índice de correlação entre as variáveis conforme pode ser visto no tabela abaixo um relacionamento negativo entre as seguintes variáveis: sex3 e diabp3; sex3 e diabetes3; diabp3 e age3;

	Sex3	Sysbp3	Diabp3	Age3	Diabetes3
Sex3	1.00000000	0.03593315	-0.05798673	0.03101508	-0.02784132
Sysbp3	0.03593315	1.00000000	0.65612871	0.34204125	0.16175283
Diabp3	- 0.05798673	0.65612871	1.00000000	- 0.03599290	0.03873005
Age3	0.03101508	0.34204125	-0.03599290	1.00000000	0.10815509
Diabetes3	- 0.02784132	0.16175283	0.03873005	0.10815509	1.00000000

Como análise do primeiro questionamento.
1 – Removemos todas as linhas que possuíam valores nulos.

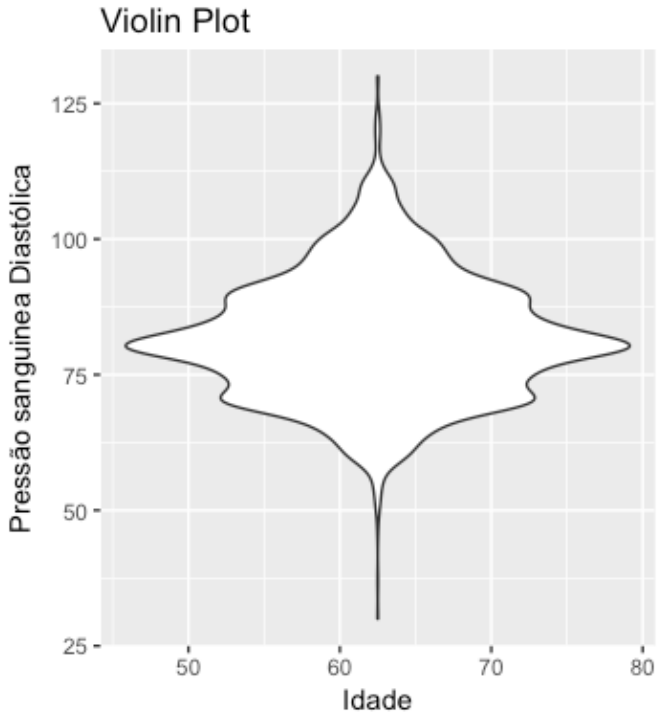
2 – Foi feita uma analise quanto a distribuição dos dados referente a variância da idade e pressão sistólica e apresentando os dados utilizando o gráfico violin plot e verificamos o aumento da pressão sanguínea sistólica após os 60 anos de idade.

```
library(ggplot2)
ggplot(remove_nulos,aes(remove_nulos$age3,remove_nulos$sysbp3))+geom
_violin()+ ggtitle("Violin Plot") +xlab("Idade") + ylab("Pressão sanguinea
sistólica")
```



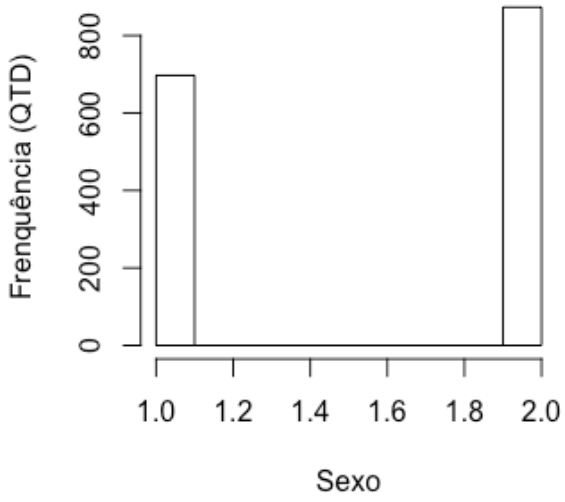
Da mesma forma foi feita uma analise quanto a distribuição dos dados referente a variância da idade e pressão diastólica e apresentando os dados utilizando o gráfico violin plot e verificamos o aumento da pressão sanguínea diastólica após os 60 anos de idade de forma menos acentuada comparando com a análise da pressão sistólica (que aumenta significativamente).

```
ggplot(remove_nulos,aes(remove_nulos$age3,remove_nulos$diabp3))+geom_violin()+ ggtitle("Violin Plot") +xlab("Idade") + ylab("Pressão sanguínea Diastólica")
```



3 - Referente a segundo questionamento realizamos a análise de quantidade de pessoas que são hipertensas (por sexo, sendo 1 = Homens e 2 = Mulheres). Sendo observado um maior número de mulheres hipertensas do que homens.

```
dataset1 <- subset(handouts_fhs, select = c("sex3","sysbp3","diabp3"))
remove_nulos <- na.omit(dataset1)
pessoas_hyp <- subset(remove_nulos, sysbp3 > 120 & diabp3 > 80)
hist(pessoas_hyp$sex3, xlab = "Sexo", ylab = "Frequência (QTD)")
```



Conforme a terceira pergunta podemos ver que 8 pessoas do sexo feminino e 3 do sexo masculino estão com alto risco cardiovascular mediante aos índices de colesterol.

Comando utilizado:
pessoas_cardio <- subset(remove_nulos, ldlc3 < 70 & hdlc3 > 40)
a <- table(pessoas_cardio $sex3)
View(a)
pie(a, main="Total por sexo (Risco cardiovascular alto)")