LABORATÓRIO 1: Relação entre Variáveis

Fernando Bispo

Sumário

Questão 1

Introdução

Os dados disponibilizados pelo *California Standardized Testing and Reporting* (STAR) contêm informações sobre performance de exames, características da escola e informações demográficas dos alunos. Os dados são provenientes de 428 escolas de ensino fundamental de distritos da Califórnia em 1998 e 1999. Os escores dos testes são médias dos escores de leitura e matemática em testes de *Stanford* padronizados que foram administrados em estudantes da 5ª série. Para as análises a serem feitas serão consideradas apenas as variáveis **MATH-SCR** (média do escore de matemática) e **STR** (razão do número de estudantes por professores:(ESTUD/PROF)

A Tabela 1 traz uma breve visão do conjunto de dados em análise.

Item a

Com base na Tabela 2, que traz as principais medidas de tendência central e de dispersão, se constata que os dados aparentam ter uma distribuição simétrica, tendo em vista o valor da Mediana ser próximo ao da Média, contudo o Coeficiente de Variação (CV), que representa a variabilidade dos dados em relação à média, traz que a variável MATH-SCR possui os dados mais agrupados (ou mais homogêneos) em relação a média, já a variável STR possui uma maior variabilidade.

Tabela 1: Apresentação parcial do conjunto de dados.

observation_number	str	math_scr
1	17,88991	690,0
2	21,52466	661,9
3	18,69723	650,9
4	17,35714	643,5
5	18,67133	639,9
6	21,40625	605,4

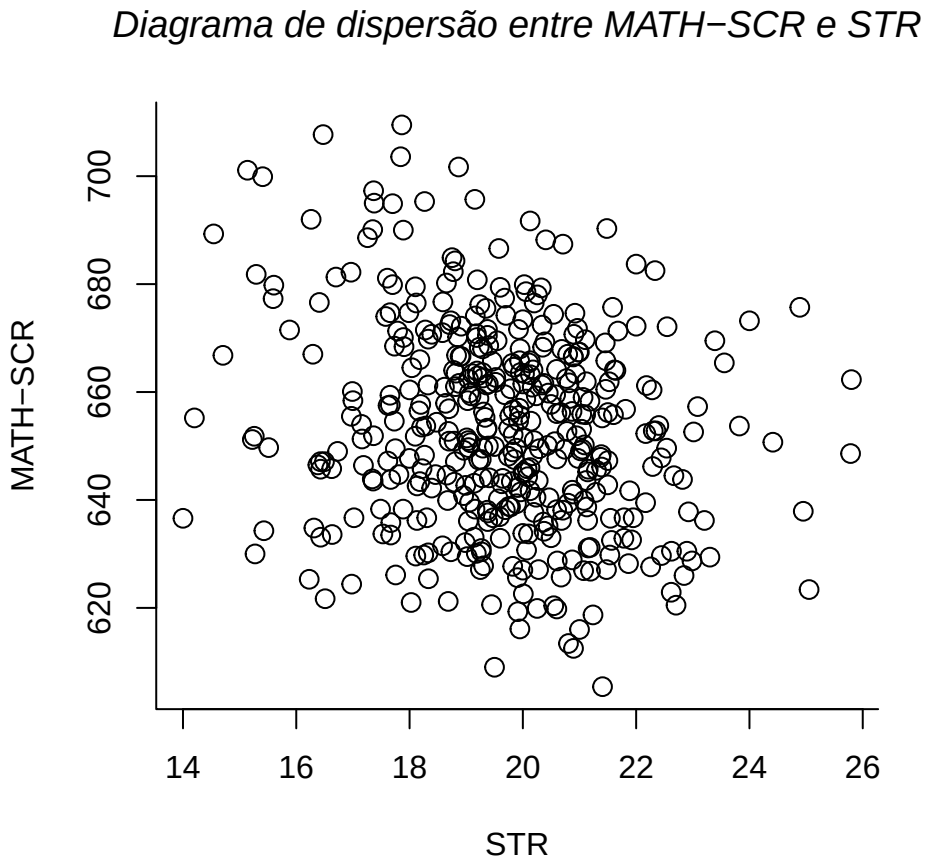
Tabela 2: Medidas tendência central e dispersão.

	Min	Q1	Med	Média	Q3	Max	Desvio Padrão	CV
math_scr	605,4	639,35	652,45	653,34	665,90	709,5	18,75	0,03
str	14,0	18,58	19,72	19,64	20,87	25,8	1,89	0,10

Note:

Fonte: California Standardized Testing and Reporting (STAR)

Item b



A escolha da variável a ser avaliada é de suma importância para obtenção de resultados coerentes, tendo em vista a não indicação da variável a ser observada, uma análise cuidadosa é necessária, para tanto, após essa análise se constatou que faz mais sentido a avaliação da variável MATH-SCR (média do escore de matemática) em relação a variável STR (razão do número de estudantes por professores) e com isso em mente se constata uma leve e aparente relação negativa entre as variáveis, ou seja, a medida que o número de alunos por professor aumenta, menor a média do escore, contudo essa é uma interpretação preliminar.

Item c

Sendo muitos os momentos necessário em que a implementação/criação de funções próprias personalizadas para o estudo de determinados fenômenos, que não estão disponíveis em pacotes conhecidos, se faz necessário, e o aprimoramento dessa habilidade deve ser incentivado, para tanto, segue a construção de uma função responsável por calcular o Desvio Padrão dos dados inseridos

```
# Criação da função:
dp = function(x){
  n = length(x)
  m = sum(x)/n
  desvio = (x - m)^2
  var = sum(desvio)/(n-1)
  dp = sqrt(var)

  return(dp)
}

# Desvio Padrão da variável MATH-SCR:
dp(dados1$math_scr)
```

```
[1] 18.7542
```

```
# Desvio Padrão da variável STR:
dp(dados1$str)
```

```
[1] 1.891812
```

Para avaliar a eficácia dessa função, um comparativo dos valores de Desvio Padrão obtidos pela função criada e os resultados dos Desvios Padrão obtidos na Tabela 2 mostra que a função cumpriu o seu papel de forma satisfatória.

Item d

Para esta etapa da questão é requisitada a criação de uma função que calcule o Coeficiente de Correlação Linear de Pearson, sendo implementada a seguir.

```
# Criação da função:
corr_pearson = function(x, y){
  n = length(x)
  soma_x = sum(x)
  soma_x2 = sum(x^2)
  s_xx = soma_x2-(((soma_x)^2)/n)
```

```
  soma_y = sum(y)
  soma_y2 = sum(y^2)
  s_yy = soma_y2-(((soma_y)^2)/n)

  soma_xy = sum(x*y)
  s_xy = soma_xy-((soma_x*soma_y)/n)

  r = s_xy/(sqrt(s_xx*s_yy))

  return(r)
}

# Função criada
corr_pearson(dados1$math_scr, dados1$str)
```

```
[1] -0.1955534
```

```
# Função já implementada
stats::cor(dados1$math_scr, dados1$str)
```

```
[1] -0.1955534
```

Através do comparativo do resultado da função criada com a função já existente, a função criada mostra que cumpriu o seu papel de forma satisfatória.

Questão 2

Introdução

O Instituto Nacional de Diabetes e de Doenças Digestivas e Renais *(National Institute of Diabetes and Digestive and Kidney Diseases)* dos EUA conduziram um estudo com 768 mulheres da tribo Pima, que residem próximo a Phoenix. As seguintes variáveis foram coletadas: número de gestações (**pregnant**), concentração de glicose no plasma (obtido 2 horas depois da realização de um teste de tolerância a glicose) (**glucose**), pressão sanguínea diastólica (mmHg)(**diastolic**), largura do tríceps (mm)(**triceps**), nível de insulina (µU/ml)(**insulin**), índice de massa corpórea (kg/m^2)(**bmi**), nível de função diabética (**diabetes**), idade (em anos) (**age**) e um teste para avaliação de sinais de diabetes (0 = negativo e 1 = positivo) (**test**). No banco de dados as variáveis encontram-se apresentadas nessa mesma ordem, mas com seus nomes em inglês.

A fim de conhecer o conjunto de dados, segue a Tabela 3 contendo as principais medidas resumo do conjunto de dados disponibilizado.

Tabela 3: Apresentação parcial do conjunto de dados.

pregnant	glucose	diastolic	triceps	insulin	bmi	diabetes	age	test
6	148	72	35	0	33,6	0,627	50	1
1	85	66	29	0	26,6	0,351	31	0
8	183	64	0	0	23,3	0,672	32	1
1	89	66	23	94	28,1	0,167	21	0
0	137	40	35	168	43,1	2,288	33	1
5	116	74	0	0	25,6	0,201	30	0

Tabela 4: Medidas tendência central e dispersão para os dados originais.

	Min	Q1	Med	Média	Q3	Max	Desvio Padrão	CV
age	21,00	24,00	29,00	33,24	41,00	81,00	11,76	0,35
bmi	0,00	27,30	32,00	31,99	36,60	67,10	7,88	0,25
diabetes	0,08	0,24	0,37	0,47	0,63	2,42	0,33	0,70
diastolic	0,00	62,00	72,00	69,11	80,00	122,00	19,36	0,28
glucose	0,00	99,00	117,00	120,89	140,50	199,00	31,97	0,26
insulin	0,00	0,00	30,50	79,80	127,50	846,00	115,24	1,44
pregnant	0,00	1,00	3,00	3,85	6,00	17,00	3,37	0,88
test	0,00	0,00	0,00	0,35	1,00	1,00	0,48	1,37
triceps	0,00	0,00	23,00	20,54	32,00	99,00	15,95	0,78

Note:

Fonte: Instituto Nacional de Diabetes e de Doenças Digestivas e Renais - EUA

Item a.1

Item a.2

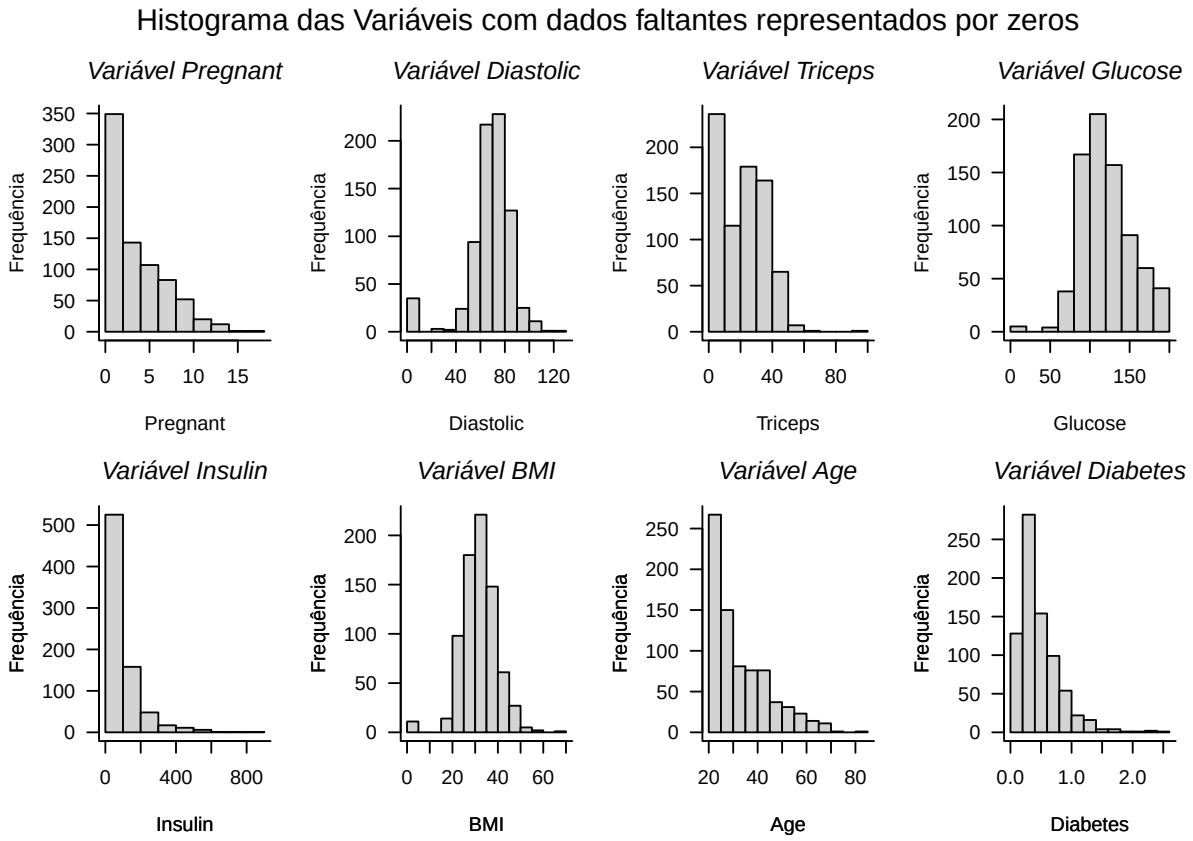


Tabela 5: Quantitativo de dados faltante.

pregnant	glucose	diastolic	triceps	insulin	bmi	diabetes	age	test
0	5	35	227	374	11	0	0	0

Tabela 6: Medidas tendência central e dispersão para os dados sem informações faltantes.

	Min	Q1	Med	Média	Q3	Max	Desvio Padrão	CV
age	21,00	24,00	29,00	33,24	41,00	81,00	11,76	0,35
bmi	18,20	27,50	32,30	32,46	36,60	67,10	6,92	0,21
diabetes	0,08	0,24	0,37	0,47	0,63	2,42	0,33	0,70
diastolic	24,00	64,00	72,00	72,41	80,00	122,00	12,38	0,17
glucose	44,00	99,00	117,00	121,69	141,00	199,00	30,54	0,25
insulin	14,00	76,00	125,00	155,55	190,00	846,00	118,78	0,76
pregnant	0,00	1,00	3,00	3,85	6,00	17,00	3,37	0,88
test	0,00	0,00	0,00	0,35	1,00	1,00	0,48	1,37
triceps	7,00	22,00	29,00	29,15	36,00	99,00	10,48	0,36

Note:

Fonte: Instituto Nacional de Diabetes e de Doenças Digestivas e Renais - EUA

Levando em consideração o fato de haverem muitos dados faltantes constatou-se que esses dados foram substituídos por valores zero, o que influencia negativamente na interpretação dos resultados das principais medidas resumo calculadas, logo essas observações serão removidas e novos cálculos serão realizados a fim de se obter medidas mais fidedignas.

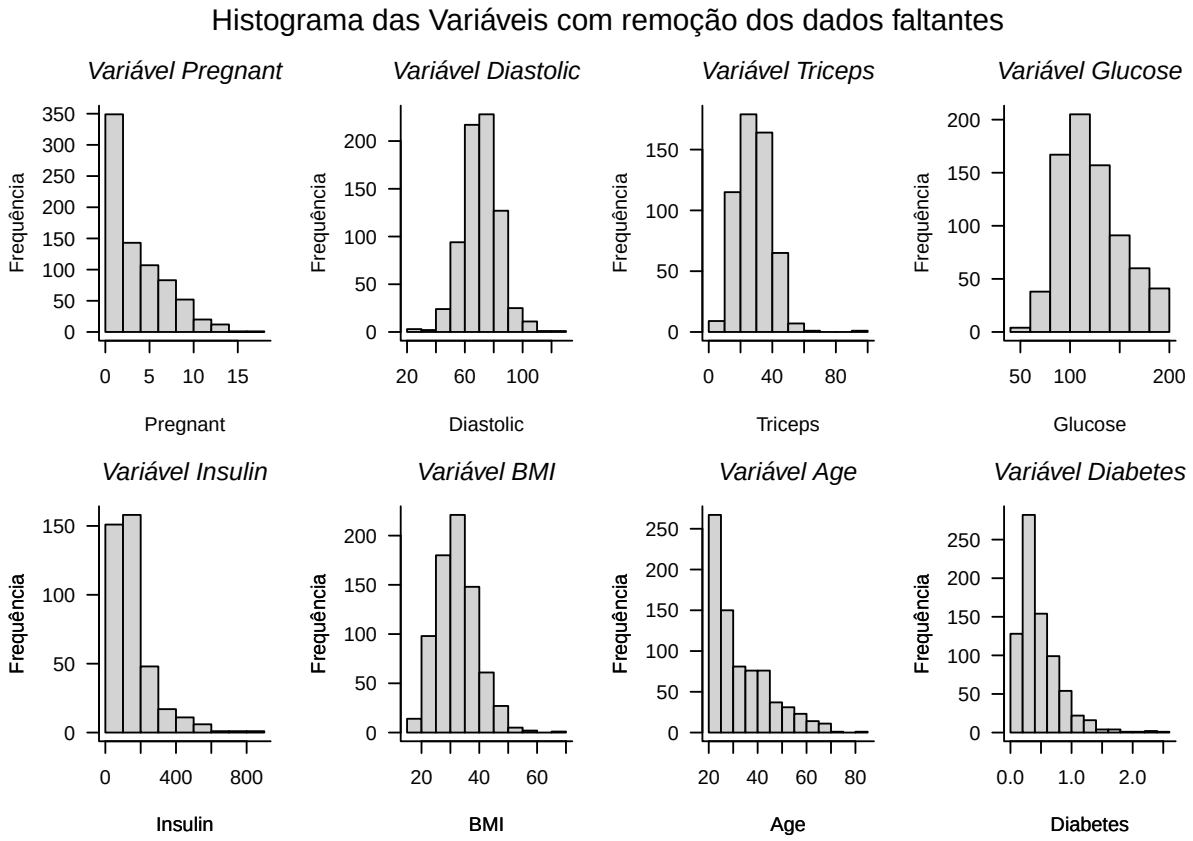
É importante ressaltar que a variável *test* é uma variável categórica, logo não faz sentido haver interpretações acerca das medidas resumo dessa variável.

Item b.1

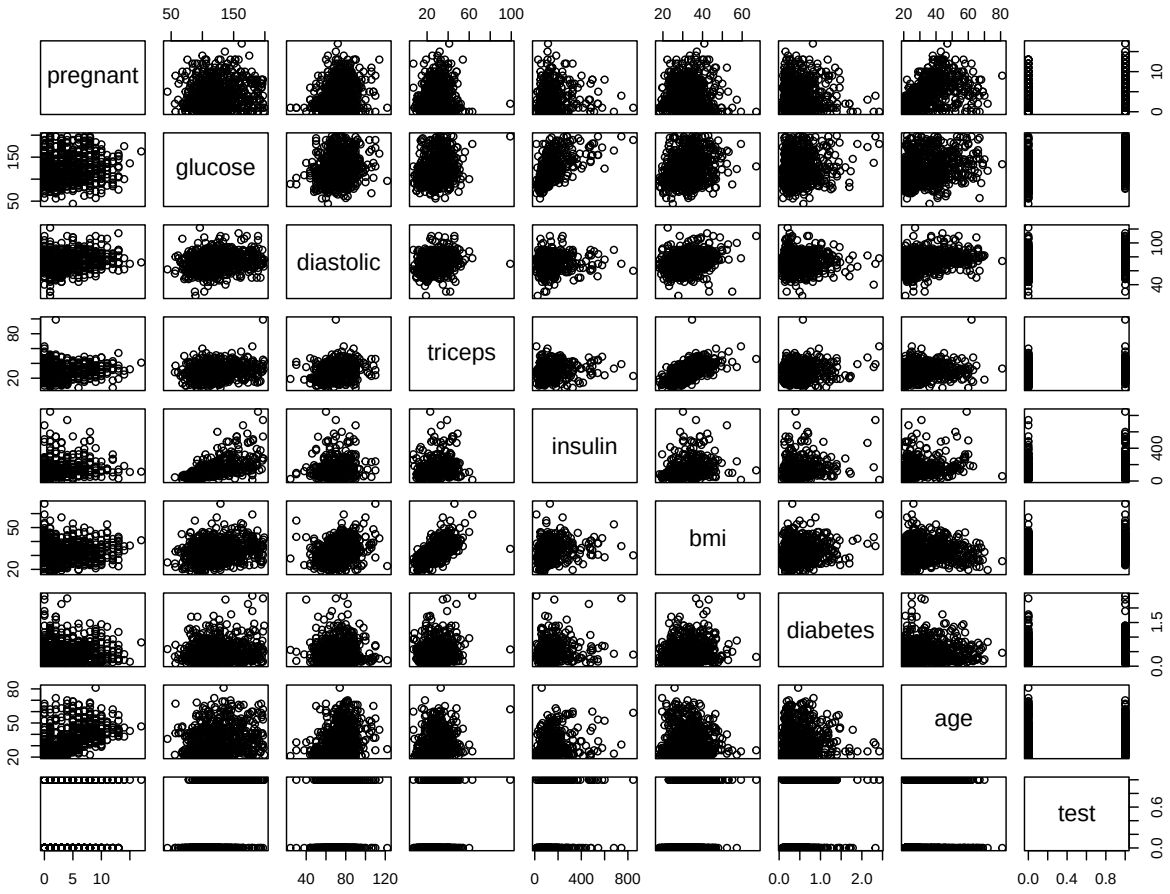
Após uma análise dos dados se constata que há variáveis com valores zerados, não fazendo sentido esse valor para a variável em questão, como por exemplo a variável **triceps** (largura do tríceps (mm)). Com base nesta anbálise, esses valores descrepantes foram removidos da análise sendo exposto na Tabela X o quantitativo de dados faltantes que cada variável possui e foram removidos, gerando-se em seguida a Tabela XX com as medidas resumo sem essas observações.

Item b.2

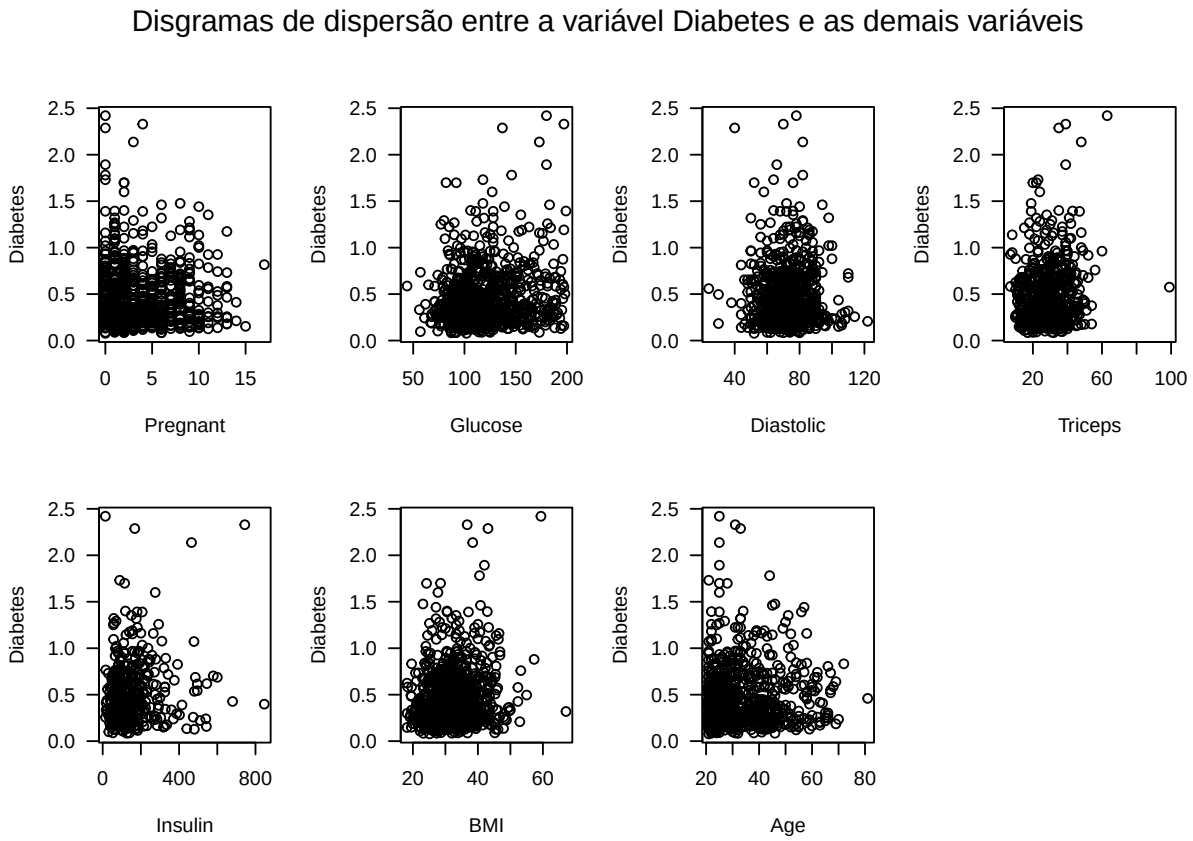
Item b.3



Item c.1


pregnant
glucose
diastolic
triceps
insulin
bmi
diabetes
age
test

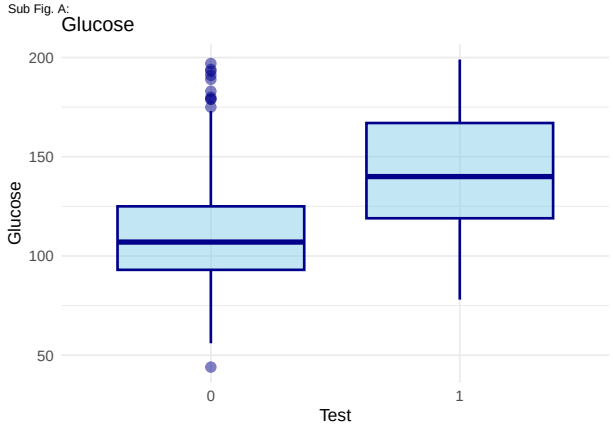
Item c.2


Disgramas de dispersão entre a variável Diabetes e as demais variáveis
Diabetes
Pregnant
Glucose
Diastolic
Triceps
Insulin
BMI
Age

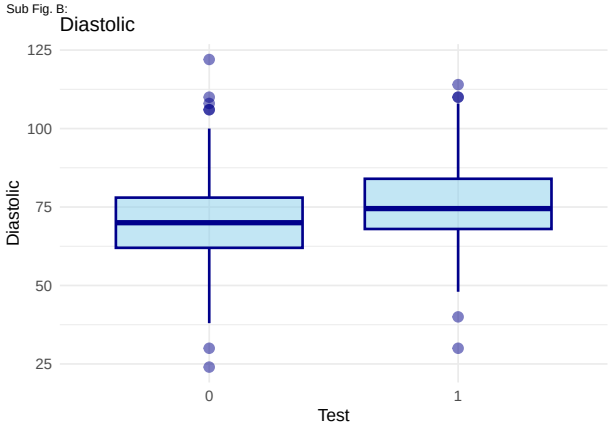
Item d

Figura 6: BoxPlot das variáveis em análise.

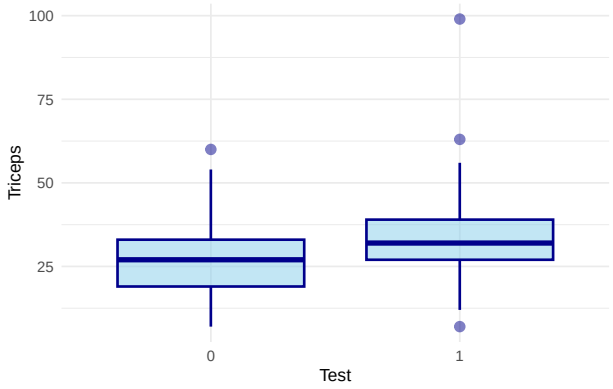
Sub Fig. A:



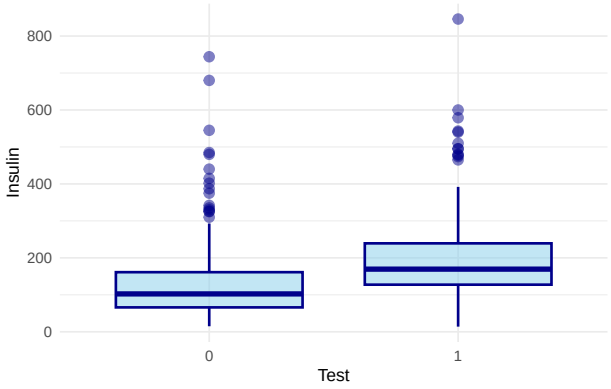
Sub Fig. B:

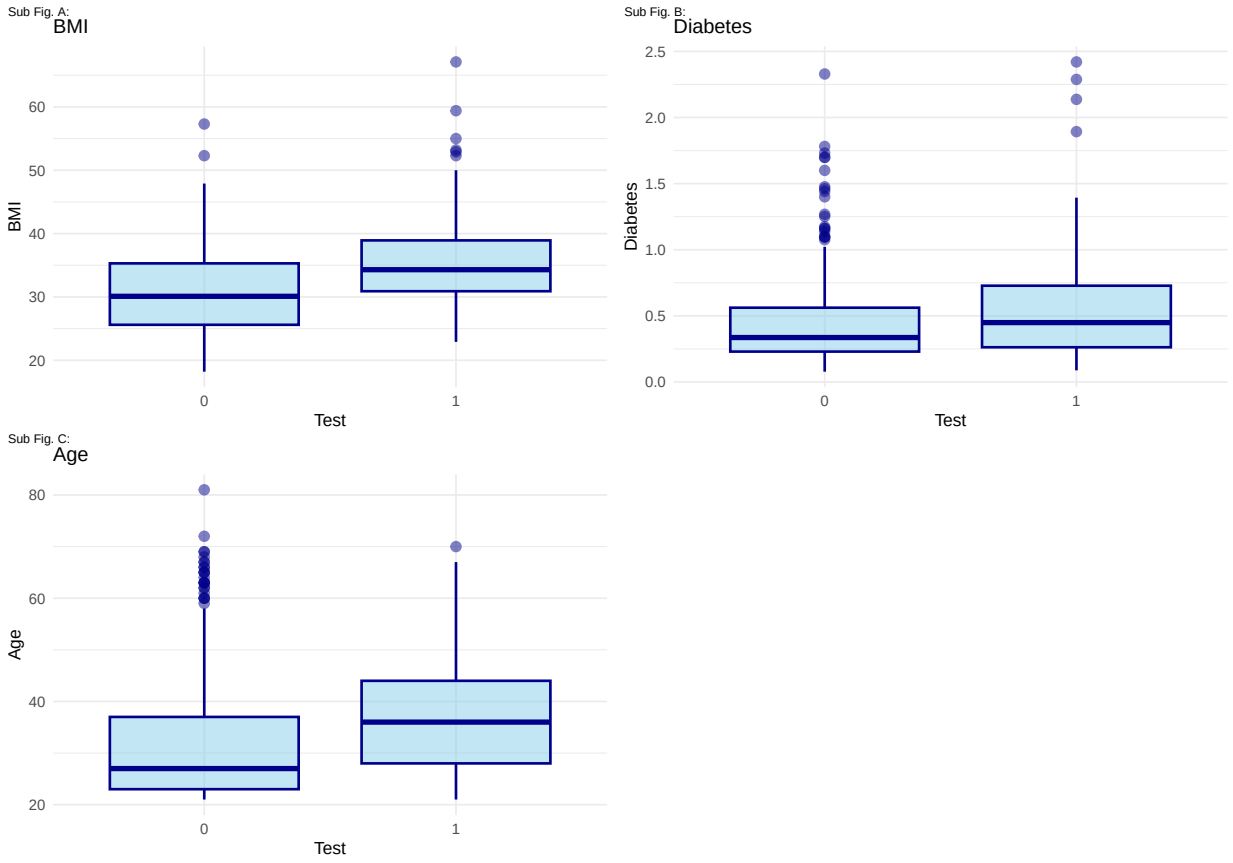


Sub Fig. C:



Sub Fig. D:





Fonte: Instituto Nacional de Diabetes e de Doenças Digestivas e Renais

Com base na Figura 6 se constata que há uma maior variabilidade entre os indivíduos que apresentaram resultado positivo para a avaliação de sinais de diabetes, além do fato da mediana dos dados serem superiores em indivíduos com esse resultado, ou seja, os indicadores para esses indivíduos são levemente alterados em comparação com os que tiveram resultado negativo.