Aula 2

Renato Rodrigues Silva

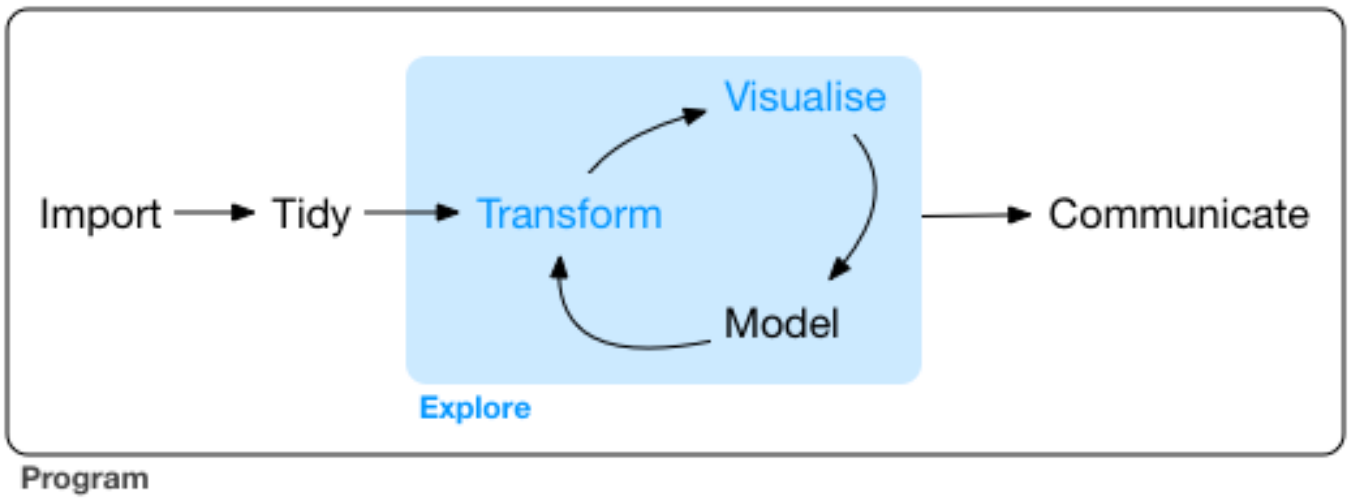
Universidade Federal de Goiás.

(2020-12-15)

Introdução

- Na aula de hoje será visto basicamente alguns fundamentos de manipulação e visualização de dados no software R.
- Manipulação consiste em fazer a limpeza e formatação dos dados.
- Visualização consiste na elaboração de gráficos.

Ciclo da análise de dados



Nessa aula, o objetivo é aprender a arrumar (tidy), transformar e visualizar os dados.

Fonte:

Conceito de dados organizados no R

- **Dado:** é a informação coletada e registrada, referente a uma variável (VIEIRA, 2018).
- **Variável:** é uma condição ou característica que descreve uma pessoa, um animal, um lugar, um objeto, uma ideia. A variável pode assumir valores diferentes em diferentes unidades (VIEIRA, 2018).
- Existem três regras inter-relacionadas que tornam um conjunto de dados organizado:
- Cada variável deve ter sua própria coluna.
- Cada observação deve ter sua própria linha.
- Cada valor deve ter sua própria célula.

Dados e Variáveis

A Figura 1.1 resume a classificação das variáveis.

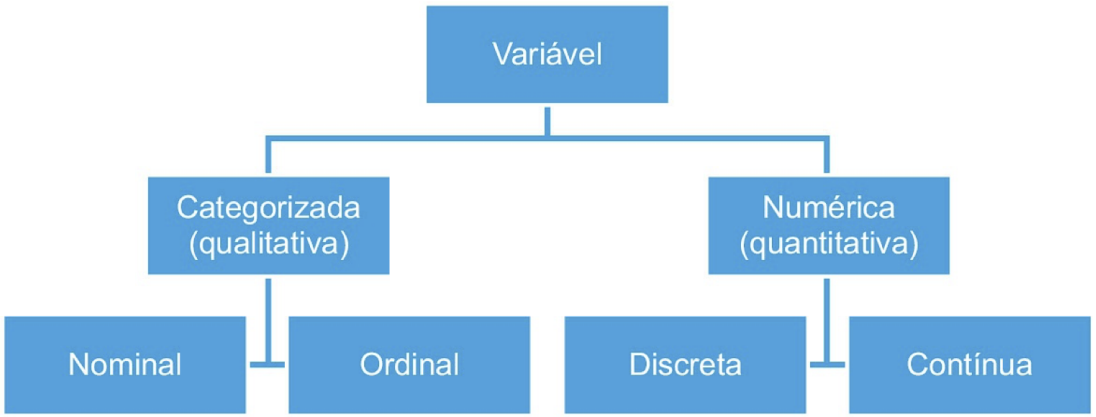


Figura 1.1 – Tipos de variáveis.

Fonte: VIEIRA, 2018.

Exemplo: Associações entre variáveis relacionadas ao diabetes em mulheres indígenas do povo Pima

- Por muitos anos, os cientistas questionaram por que tantas mulheres indígenas dos povos Pima sofrem de diabetes em relação a outras etnias.

Hipóteses Principais do Estudo

- Existe uma diferença nas médias para o índice de massa corporal (IMC) e número de gestações para aqueles que testaram positivo e aqueles que testaram negativo para diabetes ?
- Existe uma relação entre os resultados do teste para diabetes e o pedigree das nativas?

Variáveis do banco de dados

- `Pregnancies`: (Número de Gestações)
- `Glucose`: (Medição de Glicose)
- `BloodPressure`: (Pressão Sanguínea)
- `SkinThickness`: (Espessura da Pele)
- `Insulin`: (Insulina)
- `BMI`: (Índice de Massa Corporal)
- `DiabetesPedigreeFunction`: (Função definida como uma síntese da história de diabetes mellitus em parentes e a relação genética desses parentes com o sujeito.)
- `Age`: (Idade)

Tidyverse

- O pacote `tidyverse` do software estatístico `R` é utilizado para importar, manipular e visualizar dados no R.
- Para instalar o pacote, podemos usar o seguinte comando

```
install.packages("tidyverse")
```

Fonte:

Importar dados

```
dat = read.csv("diabetes.csv", header = TRUE)
```

Visão geral do data.frame()

- A função `glimpse()` fornece uma visão geral do `data.frame()`

```
glimpse(dat)
```

```
## Rows: 768
## Columns: 9
## $ Pregnancies              <int> 6, 1, 8, 1, 0, 5, 3, 10, 2, 8, 4, 10, 10,
## $ Glucose                  <int> 148, 85, 183, 89, 137, 116, 78, 115, 197,
## $ BloodPressure            <int> 72, 66, 64, 66, 40, 74, 50, 0, 70, 96, 92
## $ SkinThickness            <int> 35, 29, 0, 23, 35, 0, 32, 0, 45, 0, 0, 0,
## $ Insulin                  <int> 0, 0, 0, 94, 168, 0, 88, 0, 543, 0, 0, 0,
## $ BMI                      <dbl> 33.6, 26.6, 23.3, 28.1, 43.1, 25.6, 31.0,
## $ DiabetesPedigreeFunction <dbl> 0.627, 0.351, 0.672, 0.167, 2.288, 0.201,
## $ Age                      <int> 50, 31, 32, 21, 33, 30, 26, 29, 53, 54, 3
## $ Outcome                  <int> 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1,
```

Introdução a manipulação de dados - Uso da biblioteca `dplyr`

- A biblioteca `dplyr` faz parte do conjunto de bibliotecas `tidyverse`. Na prática, quando o usuário instala o `tidyverse` a biblioteca `dplyr` já estará instalada.
- O principal objetivo da biblioteca `dplyr` é fazer manipulação de dados.

dplyr - Principais funções

- `filter()` - filtra linhas
- `select()` - seleciona colunas
- `mutate()` - cria/modifica colunas
- `arrange()` - ordena a planilha
- `summarise()` - calcula algumas medidas resumo no conjunto de dados.

Usando a função filter()

- Selecione apenas as linhas em que o número de gestações seja maior do que 2

```
filter(dat, Pregnancies > 2)
```

- Selecione apenas as linhas em que o número de gestações seja maior do que 2 e glicose menor do que 126

```
filter(dat, Pregnancies > 2 & Glucose < 126)
```

- Selecione apenas as linhas em que o nível de insulina seja igual a zero

```
filter(dat, Insulin == 0)
```

Usando a função filter()

- Selecione apenas as linhas em que o número de gestações seja maior do que 2

```
##     Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1             6     148            72            35       0 33.6
## 2             8     183            64             0       0 23.3
## 3             5     116            74             0       0 25.6
## 4             3      78            50            32      88 31.0
## 5            10     115             0             0       0 35.3
## 6             8     125            96             0       0  0.0
## 7             4     110            92             0       0 37.6
## 8            10     168            74             0       0 38.0
## 9            10     139            80             0       0 27.1
## 10            5     166            72            19     175 25.8
## 11            7     100             0             0       0 30.0
## 12            7     107            74             0       0 29.6
## 13            3     126            88            41     235 39.3
## 14            8      99            84             0       0 35.4
## 15            7     196            90             0       0 39.8
## 16            9     119            80            35       0 29.0
## 17           11     143            94            33     146 36.6
## 18           10     125            70            26     115 31.1
## 19            7     147            76             0       0 39.4
```

Usando a função filter()

- Selecione apenas as linhas em que o número de gestações seja maior do que 2 e glicose menor do que 126

```
##     Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1             5     116            74             0       0 25.6
## 2             3      78            50            32      88 31.0
## 3            10     115             0             0       0 35.3
## 4             8     125            96             0       0  0.0
## 5             4     110            92             0       0 37.6
## 6             7     100             0             0       0 30.0
## 7             7     107            74             0       0 29.6
## 8             8      99            84             0       0 35.4
## 9             9     119            80            35       0 29.0
## 10           10     125            70            26     115 31.1
## 11            5     117            92             0       0 34.1
## 12            5     109            75            26       0 36.0
## 13            3      88            58            11      54 24.8
## 14            6      92            92             0       0 19.9
## 15           10     122            78            31       0 27.6
## 16            4     103            60            33     192 24.0
## 17            9     102            76            37       0 32.9
## 18            4     111            72            47     207 37.1
## 19            7     106            92            18       0 22.7
```

Usando a função filter()

- Selecione apenas as linhas em que o nível de insulina seja igual a zero

```
##     Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1             6     148            72            35       0 33.6
## 2             1      85            66            29       0 26.6
## 3             8     183            64             0       0 23.3
## 4             5     116            74             0       0 25.6
## 5            10     115             0             0       0 35.3
## 6             8     125            96             0       0  0.0
## 7             4     110            92             0       0 37.6
## 8            10     168            74             0       0 38.0
## 9            10     139            80             0       0 27.1
## 10            7     100             0             0       0 30.0
## 11            7     107            74             0       0 29.6
## 12            8      99            84             0       0 35.4
## 13            7     196            90             0       0 39.8
## 14            9     119            80            35       0 29.0
## 15            7     147            76             0       0 39.4
## 16            5     117            92             0       0 34.1
## 17            5     109            75            26       0 36.0
## 18            6      92            92             0       0 19.9
## 19           10     122            78            31       0 27.6
## 20           11     138            76             0       0 33.2
```

Usando a função select()

- Selecione apenas as colunas referentes a glicose

```
select(dat, Glucose)
```

- Selecione todas as variáveis que inicia com B

```
select(dat, starts_with("B"))
```

- Selecione apenas as colunas referentes a glicose, e pressão sanguínea

```
vars = c("Glucose", "Pregnancies", "Insulin")

select(dat, one_of(vars))
```

Usando a função select()

- Selecione apenas as colunas referentes a glicose

```
##     Glucose
## 1       148
## 2        85
## 3       183
## 4        89
## 5       137
## 6       116
## 7        78
## 8       115
## 9       197
## 10      125
## 11      110
## 12      168
## 13      139
## 14      189
## 15      166
## 16      100
## 17      118
## 18      107
## 19      103
## 20      115
```

Usando a função select()

- Selecione todas as variáveis que inicia com B

```
##     BloodPressure  BMI
## 1              72 33.6
## 2              66 26.6
## 3              64 23.3
## 4              66 28.1
## 5              40 43.1
## 6              74 25.6
## 7              50 31.0
## 8               0 35.3
## 9              70 30.5
## 10             96  0.0
## 11             92 37.6
## 12             74 38.0
## 13             80 27.1
## 14             60 30.1
## 15             72 25.8
## 16              0 30.0
## 17             84 45.8
## 18             74 29.6
## 19             30 43.3
## 20             70 34.6
```

Usando a função select()

- Selecione apenas as colunas referentes a glicose, e pressão sanguínea

```
##     Glucose Pregnancies Insulin
## 1       148           6       0
## 2        85           1       0
## 3       183           8       0
## 4        89           1      94
## 5       137           0     168
## 6       116           5       0
## 7        78           3      88
## 8       115          10       0
## 9       197           2     543
## 10      125           8       0
## 11      110           4       0
## 12      168          10       0
## 13      139          10       0
## 14      189           1     846
## 15      166           5     175
## 16      100           7       0
## 17      118           0     230
## 18      107           7       0
## 19      103           1      83
## 20      115           1      96
```

Usando a função select()

- Selecione apenas as colunas referentes a glicose, e pressão sanguínea

```
##     Glucose Pregnancies Insulin
## 1       148           6       0
## 2        85           1       0
## 3       183           8       0
## 4        89           1      94
## 5       137           0     168
## 6       116           5       0
## 7        78           3      88
## 8       115          10       0
## 9       197           2     543
## 10      125           8       0
## 11      110           4       0
## 12      168          10       0
## 13      139          10       0
## 14      189           1     846
## 15      166           5     175
## 16      100           7       0
## 17      118           0     230
## 18      107           7       0
## 19      103           1      83
## 20      115           1      96
```

Usando a função mutate()

- Transformar a variável resultados (Outcome) em uma variável categorica

```
mutate(dat, Outcome = factor(Outcome))  %>% glimpse()
```

- Calcule transformar as medidas de glicose em escala logaritmica

```
mutate(dat, lGlucose = log(Glucose))  %>% select(Glucose, lGlucose)
```

- Vamos criar uma nova variável denominada $QUICKI = \frac{1}{\log \text{insulina} + \log \text{glicose}}$.

```
mutate(dat, QUICKI =  1 / (log10(Glucose) + log10(Insulin)))  %>% sel
```

Fonte:

- O comando %>% encaminhará um valor, ou o resultado de uma expressão, para a próxima de função

Usando a função mutate()

- Transformar a variável resultados (Outcome) em uma variável categorica

```
## Rows: 768
## Columns: 9
## $ Pregnancies              <int> 6, 1, 8, 1, 0, 5, 3, 10, 2, 8, 4, 10, 10,
## $ Glucose                  <int> 148, 85, 183, 89, 137, 116, 78, 115, 197,
## $ BloodPressure            <int> 72, 66, 64, 66, 40, 74, 50, 0, 70, 96, 92
## $ SkinThickness            <int> 35, 29, 0, 23, 35, 0, 32, 0, 45, 0, 0, 0,
## $ Insulin                  <int> 0, 0, 0, 94, 168, 0, 88, 0, 543, 0, 0, 0,
## $ BMI                      <dbl> 33.6, 26.6, 23.3, 28.1, 43.1, 25.6, 31.0,
## $ DiabetesPedigreeFunction <dbl> 0.627, 0.351, 0.672, 0.167, 2.288, 0.201,
## $ Age                      <int> 50, 31, 32, 21, 33, 30, 26, 29, 53, 54, 3
## $ Outcome                  <fct> 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1,
```

Usando a função mutate()

- Calcule transformar as medidas de glicose em escala logaritmica

```
##     Glucose lGlucose
## 1       148 4.997212
## 2        85 4.442651
## 3       183 5.209486
## 4        89 4.488636
## 5       137 4.919981
## 6       116 4.753590
## 7        78 4.356709
## 8       115 4.744932
## 9       197 5.283204
## 10      125 4.828314
## 11      110 4.700480
## 12      168 5.123964
## 13      139 4.934474
## 14      189 5.241747
## 15      166 5.111988
## 16      100 4.605170
## 17      118 4.770685
## 18      107 4.672829
## 19      103 4.634729
## 20      115 4.744932
```

Usando a função mutate()

- Vamos criar uma nova variável denominada $QUICKI = \frac{1}{\log \text{insulina} + \log \text{glicose}}$.

```
##     Glucose Insulin    QUICKI
## 1       148       0 0.0000000
## 2        85       0 0.0000000
## 3       183       0 0.0000000
## 4        89      94 0.2549383
## 5       137     168 0.2292511
## 6       116       0 0.0000000
## 7        78      88 0.2606490
## 8       115       0 0.0000000
## 9       197     543 0.1988362
## 10      125       0 0.0000000
## 11      110       0 0.0000000
## 12      168       0 0.0000000
## 13      139       0 0.0000000
## 14      189     846 0.1921661
## 15      166     175 0.2240572
## 16      100       0 0.0000000
## 17      118     230 0.2255498
## 18      107       0 0.0000000
```

Usando a função arrange()

- Ordenar o conjunto de dados de forma crescente pela coluna número de gestações

```
arrange(dat, Pregnancies)
```

- Ordenar o conjunto de dados de forma decrescente pela coluna número de gestações

```
arrange(dat, desc(Pregnancies))
```

- Ordenar o conjunto de dados de forma crescente pela colunas número de gestações, glicose e pressão sanguínea

```
arrange(dat, Pregnancies, Glucose, BloodPressure )
```

Usando a função arrange()

- Ordenar o conjunto de dados de forma crescente pela coluna número de gestações

```
##     Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1             0     137            40            35     168 43.1
## 2             0     118            84            47     230 45.8
## 3             0     180            66            39       0 42.0
## 4             0     100            88            60     110 46.8
## 5             0     146            82             0       0 40.5
## 6             0     105            64            41     142 41.5
## 7             0     109            88            30       0 32.5
## 8             0     131             0             0       0 43.2
## 9             0     101            65            28       0 24.6
## 10            0     125            96             0       0 22.5
## 11            0      95            85            25      36 37.4
## 12            0     162            76            56     100 53.2
## 13            0     113            76             0       0 33.3
## 14            0     105            84             0       0 27.9
## 15            0     100            70            26      50 30.8
## 16            0      93            60            25      92 28.7
## 17            0     129            80             0       0 31.2
## 18            0     102            75            23       0  0.0
## 19            0     114            80            34     285 44.2
```

Usando a função arrange()

- Ordenar o conjunto de dados de forma decrescente pela coluna número de gestações

```
##     Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1            17     163            72            41     114 40.9
## 2            15     136            70            32     110 37.1
## 3            14     100            78            25     184 36.6
## 4            14     175            62            30       0 33.6
## 5            13     145            82            19     110 22.2
## 6            13     126            90             0       0 43.4
## 7            13     106            72            54       0 36.6
## 8            13     106            70             0       0 34.2
## 9            13     152            90            33      29 26.8
## 10           13     129             0            30       0 39.9
## 11           13      76            60             0       0 32.8
## 12           13     104            72             0       0 31.2
## 13           13     158           114             0       0 42.3
## 14           13     153            88            37     140 40.6
## 15           12     151            70            40     271 41.8
## 16           12      92            62             7     258 27.6
## 17           12     106            80             0       0 23.6
## 18           12      88            74            40      54 35.3
## 19           12     140            82            43     325 39.2
```

Usando a função arrange()

- Ordenar o conjunto de dados de forma crescente pela colunas número de gestações, glicose e pressão sanguínea

```
##     Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 1             0      57            60             0       0 21.7
## 2             0      67            76             0       0 45.3
## 3             0      73             0             0       0 21.1
## 4             0      74            52            10      36 27.8
## 5             0      78            88            29      40 36.9
## 6             0      84            64            22      66 35.8
## 7             0      84            82            31     125 38.2
## 8             0      86            68            32       0 35.8
## 9             0      91            68            32     210 39.9
## 10            0      91            80             0       0 32.4
## 11            0      93            60            25      92 28.7
## 12            0      93            60             0       0 35.3
## 13            0      93           100            39      72 43.4
## 14            0      94             0             0       0  0.0
## 15            0      94            70            27     115 43.5
## 16            0      95            64            39     105 44.6
## 17            0      95            80            45      92 36.5
## 18            0      95            85            25      36 37.4
## 19            0      97            64            36     100 36.8
```

Usando a função summarise()

- Essa função calcula algumas medidas resumo no conjunto de dados. Normalmente, ela é utilizada em conjunto com a função `group_by()`.
- Vamos calcular a frequencia absoluta dos grupos que testaram positivo e o grupo que testaram negativo

```
group_by(dat, Outcome) %>% summarise(freq.abs = n())
```

- Vamos calcular a média do Indice de Massa Corporal para os grupos que testaram positivo e o grupo que testaram negativo

```
group_by(dat, Outcome) %>% summarise(IMC_medio = mean(BMI))
```

Usando a função summarise()

- Vamos calcular a frequencia absoluta dos grupos que testaram positivo e o grupo que testaram negativo

```
## # A tibble: 2 x 2
##   Outcome freq.abs
##     <int>    <int>
## 1       0      500
## 2       1      268
```

Usando a função summarise()

- Vamos calcular a média do Indice de Massa Corporal para os grupos que testaram positivo e o grupo que testaram negativo

```
## # A tibble: 2 x 2
##   Outcome IMC_medio
##     <int>     <dbl>
## 1       0      30.3
## 2       1      35.1
```

Visualizar dados

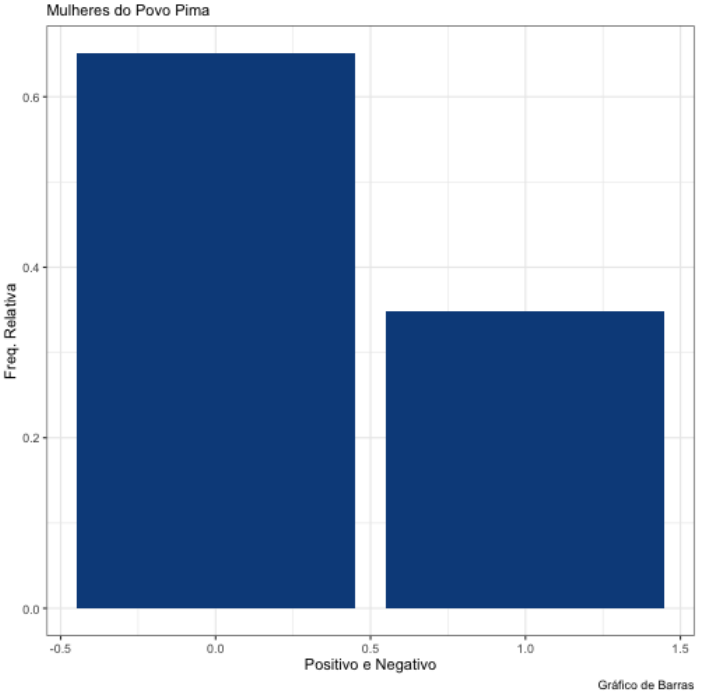
- "Um gráfico simples traz mais informações à mente do analista de dados do que qualquer outro tipo de análise." - John Tukey
- Uma boa visualização mostrará coisas que você não esperava ou levantará novas questões sobre os dados.
- Uma boa visualização também pode indicar que você está fazendo a pergunta errada ou precisa coletar dados diferentes.

Fonte:

Alguns tipos de gráficos

Gráficos de Barras

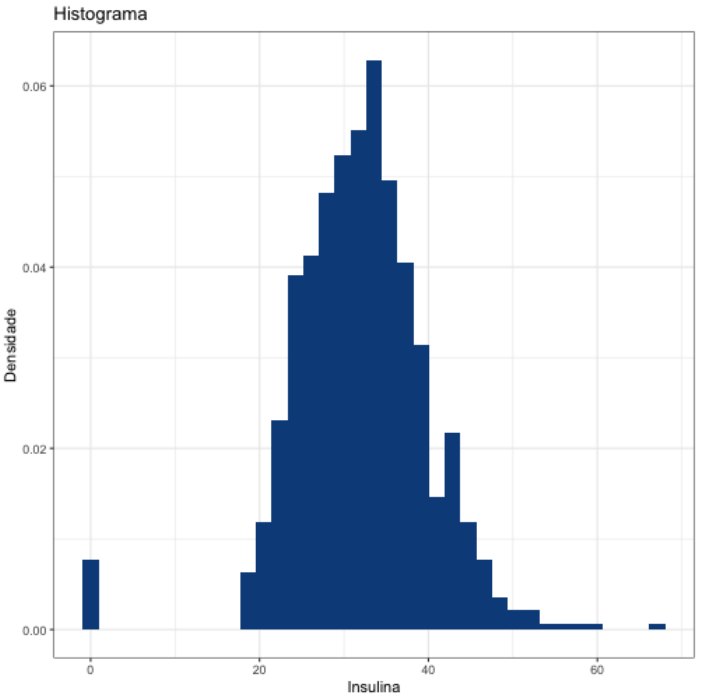
- Gráficos de Barras são apropriados para apresentar dados qualitativos ou quantitativos discretos (poucas classes).



Alguns tipos de gráficos

Histogramas

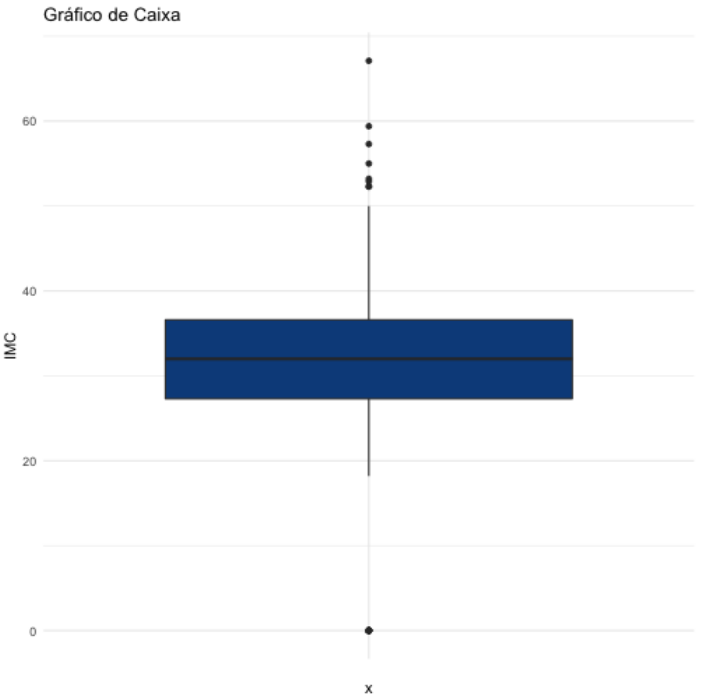
- Um tipo especial de gráfico para representam a frequência de dados quantitativos contínuos que foram organizados em intervalos.



Alguns tipos de gráficos

Gráfico de Caixas

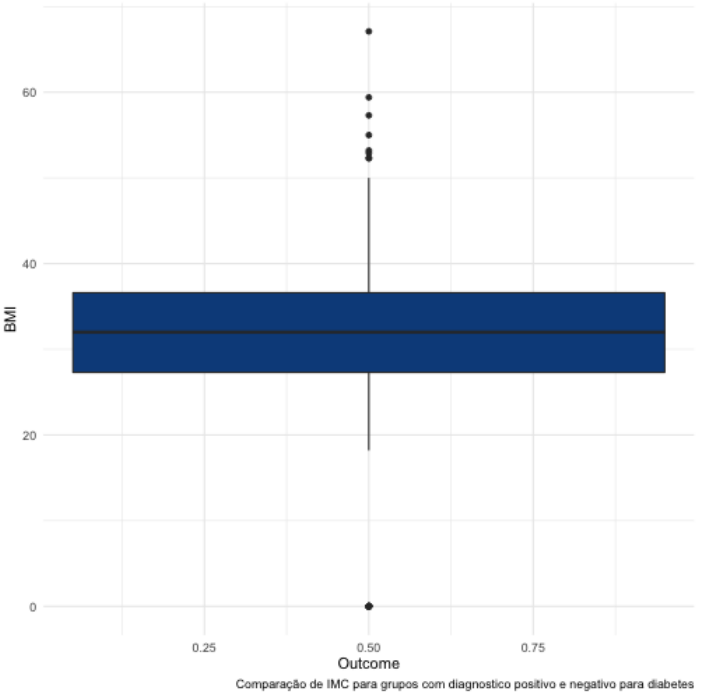
- Um gráfico de caixa é um diagrama que resume os dados por dividindo-o em quatro partes (quartis).



Alguns tipos de gráficos

Gráficos de Caixa para verficar relação entre variáveis quantitativas e qualitativas

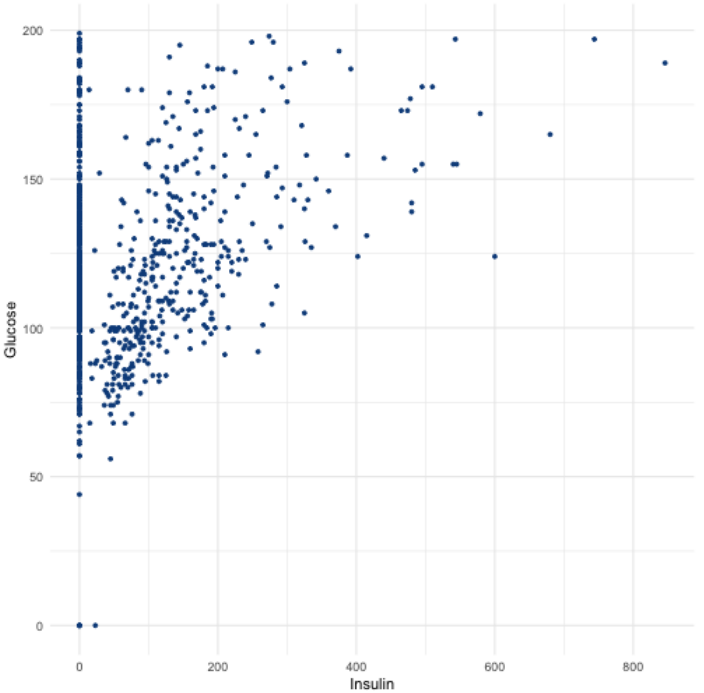
- Agora vamos visualizar os dados para responder a seguinte pergunta:
- Existe uma diferença no valor da mediana para o índice de massa corporal (IMC) e número de gestações para aqueles que testaram positivo e aqueles que testaram negativo para diabetes ?



Alguns tipos de gráficos

Gráfico de Dispersão

- Um gráfico de dispersão é um tipo de gráfico que usa coordenadas cartesianas para exibir valores de duas variáveis. Útil para verificar a relação entre as variáveis quantitativas.



Visualização de dados no R

- Existem inúmeras formas de fazer visualização de dados no R.
- Nesta aula, apresentamos a biblioteca `esquisse`.
- Para instalar a biblioteca, podemos usar `install.package("esquisse")`
- Para usar a biblioteca, devemos carregar a biblioteca `library(esquisse)`, e para inicializar a biblioteca usamos `esquisser()`.
- Maiores detalhes, podem ser vistos aqui e aqui.

Referências

- WICKHAM, H.; GROMULRMUND, G. R for Data Science, 2017. O'Reilly Media. Disponível em: https://r4ds.had.co.nz/explore-intro.html. Acesso em: 26 de nov. de 2020.
- Instalação. CURSO-R. [2018?] Disponível em: http://material.curso-r.com/instalacao/. Acesso em: 20 de nov. de 2020.
- VIEIRA, S. Bioestatística. Tópicos Avançados. 4 edição. 2018