Um Projeto de Análise Exploratória de Dados em R

Rashida Nasrin Sucky

Dec 22 2020

Objetivos:

Resumir e visualizar os dados para uma melhor compreensão, mesmo que todas as variáveis não sejam compreensíveis

Este é um Notebook R [R Markdown] (http://rmarkdown.rstudio.com). Quando você executa o código no notebook, os resultados aparecem abaixo do código.

Para executar os bloco clique no botão *Executar* dentro do bloco ou posicione o cursor dentro do bloco e pressione *Cmd + Shift + Enter*.

Introdução

Análise exploratória de dados (AED) é uma operação essencial para a análise dos conjuntos de dados (data sets). Às vezes é necessário apenas entender bem os dados estão configurados. Muitas vezes é a etapa realizada antes de se mergulhar na modelagem.

De qualquer forma, um grande conjunto de dados não terá utilidade se não for possível extrair dele as informações necessárias.

Este notebook buscará mostrar algumas técnicas e códigos de visualização para extrair informações importantes de um conjunto de dados.

O conjunto de dados utilizado neste notebook é um conjunto de dados de doenças cardíacas originário da plataforma [Kaggle] (http://kaggle.com/johnsmith88/heart-disease-dataset?select=heart.csv).

Vários termos deste conjunto de dados podem não ser tão compreensíveis para muitas pessoas, tais como depressão de ST ou grandes vasos. Mesmo assim, pode-se obter boas informações a partir dele.

Veremos como.

Visão geral do conjunto de dados

Este conjunto de dados é sobre doenças cardíacas. Cada linha representa os diferentes dados de saúde de uma pessoa.

O conjunto de dados foi baixado e colocado na pasta "data" do projeto.

Para importar o conjunto de dados no ambiente RStudio utiliza-se este pedaço de código (code chunk):

```
heart = read.csv("./data/heart.csv")
```

Estrutura do conjunto de dados

O conjunto de dados possui 14 colunas. Portanto, é muito grande para fazer uma captura de tela e mostrá-lo aqui.

Aqui está o nome das colunas e a explicação de cada variável conforme descrito no Kaggle.

1. age: a idade de uma pessoa
2. sex: o gênero da pessoa (1 = masculino, 0 = feminino)
3. cp: os tipos de dor no peito experimentados (Valor 1: angina típica, Valor 2: angina atípica, Valor 3: dor não anginosa, Valor 4: assintomática)
4. trestbps: pressão arterial em repouso (mm Hg na admissão ao hospital)
5. col: medição de colesterol em mg / dl
6. fbs: açúcar no sangue em jejum (se> 120 mg / dl, 1 = verdadeiro; 0 = falso)
7. restecg: medição eletrocardiográfica em repouso (0 = normal, 1 = tendo anormalidade da onda ST-T, 2 = mostrando hipertrofia ventricular esquerda provável ou definitiva pelos critérios de Estes)
8. thalach: frequência cardíaca máxima alcançada
9. exang: angina induzida por exercício (1 = sim; 0 = não) 10.oldpeak: (pico anterior) depressão de ST induzida por exercício em relação ao repouso ("ST" refere-se às posições no gráfico de ECG)
10. slope: (inclinação) a inclinação do segmento ST de pico do exercício (Valor 1: inclinação para cima, Valor 2: plana, Valor 3: inclinação para baixo) 12.ca: O número dos principais vasos (0–3) 13.thal: Uma doença do sangue chamada talassemia (1 = normal; 2 = defeito fixo; 3 = defeito reversível) 14.target: (alvo) doenças cardíacas (0 = não, 1 = sim)

Preparação de dados

O conjunto de dados já está limpo e bem organizado. Não foi necessário realizar muita "limpeza". Muitas coisas podem ser feitas com um conjunto de dados como este. O dataset pode ser analisado de muitas maneiras diferentes. Muitos gráficos e tabelas diferentes podem ser gerados para explicar de maneiras diferentes.

Para este notebook, optou-se por encontrar qualquer correlação entre doenças cardíacas e os outros parâmetros dos dados.

Se se obervar o conteúdo da variável 14 (target), a mesma mostra se uma pessoa tem uma doença cardíaca ou não. Nos concentraremos bastante nesta variável.

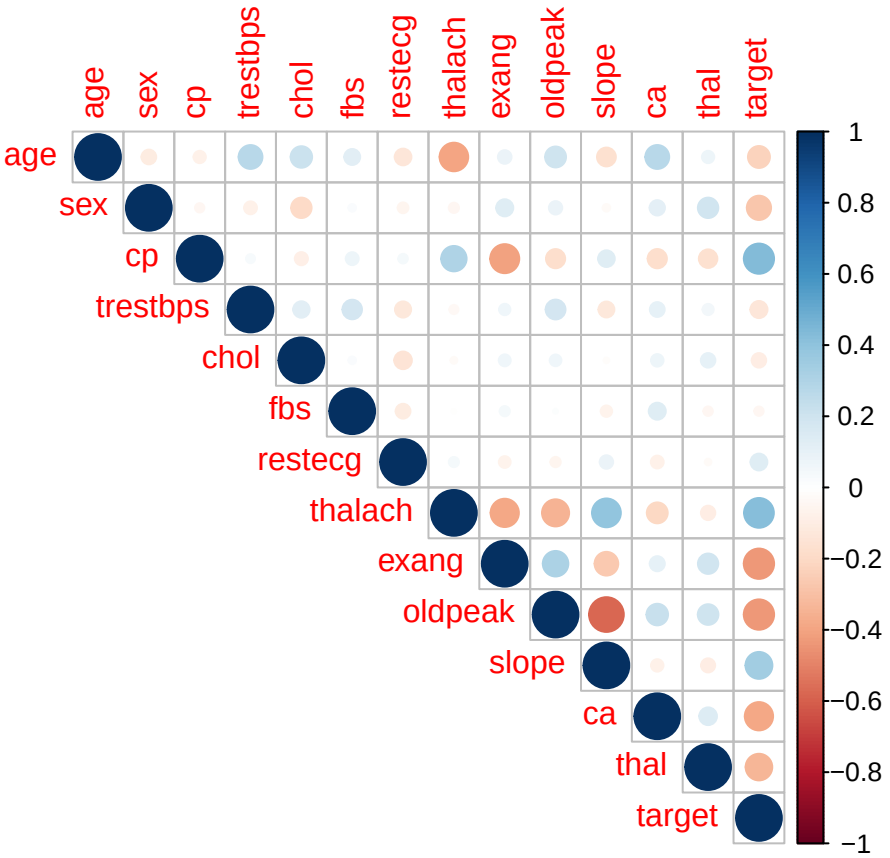
Para começar, será importante verificar a correlação entre as doenças cardíacas e as outras variáveis no conjunto de dados.

Para tanto, será utilizada a biblioteca 'corrplot' para gerar um gráfico de correlação que mostrará a correlação de cada variável com todas as outras.

```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
corrplot(cor(heart), type="upper")
```



Como o foco nesta AED é o de descobrir a relação entre a doença cardíaca e outros parâmetros, observaremos a correlação entre a variável target ("alvo") com as outras variáveis.

O tamanho dos pontos no gráfico mostra o quão forte é a correlação.

O gráfico de correlação mostra que os parâmetros 'restecg', 'fbs' e 'chol' estão vagamente correlacionados com a 'variável alvo. Portanto, podemos excluí-las com segurança do conjunto de dados para este estudo específico.

```
# Excluindo as váriáveis 'restecg', 'fbs' e 'chol'
heart = subset(heart, select=c(-restecg, -chol,-fbs))
```

Temos agora 11 variáveis. Demonstraremos a análise de algumas relações de variáveis discretas e algumas categóricas com a variável "alvo".

As variáveis categóricas são denotadas como 0, 1, 2, 3. Foram alteradas para valores de string mais significativos de acordo com a descrição acima.

Aqui estão os códigos para alterar as variáveis categóricas para as strings correspondentes:

```
heart$sex[heart$sex == 0] = "female"
heart$sex[heart$sex == 1] = "male"

heart$cp[heart$cp == 0] = "typical angina"
heart$cp[heart$cp == 1] = "atypical angina"
heart$cp[heart$cp == 2] = "non-anginal pain"
heart$cp[heart$cp == 3] = "asymptomatic"
```

```
heart$exang[heart$exang == 0] = "no"
heart$exang[heart$exang == 1] = "yes"

heart$slope[heart$slope == 0] = "upsloping"
heart$slope[heart$slope == 1] = "flat"
heart$slope[heart$slope == 2] = "downsloping"

heart$thal[heart$thal == 1] = "normal"
heart$thal[heart$thal == 2] = "fixed defect"
heart$thal[heart$thal == 3] = "reversible defect"

heart$target1 = heart$target
heart$target1[heart$target1 == 0] = "no heart disease"
heart$target1[heart$target1 == 1] = "heart disease"
```

O conjunto de dados agora está pronto !

Vamos agora mergulhar na análise exploratória.

Análise Exploratória de Dados (EAD)

A análise exploratória de dados começa à partir das perguntas, curiosidades e necessidades. Iniciaremos com as perguntas relacionadas abaixo e vamos buscar respondê-las.

Como nos concentraremos principalmente nas doenças cardíacas, é intuitivo começar com a proporção de pessoas *com doenças cardíacas* e *sem doenças cardíacas* no conjunto de dados.

```
round(prop.table(table(heart$target1)),2)
```

```
##
##    heart disease no heart disease
##             0.51             0.49
```

Os resultados mostram que 51% das pessoas neste conjunto de dados possuem com doenças cardíacas e que 49% das pessoas não apresentam doenças cardíacas. São valores muito próximos.

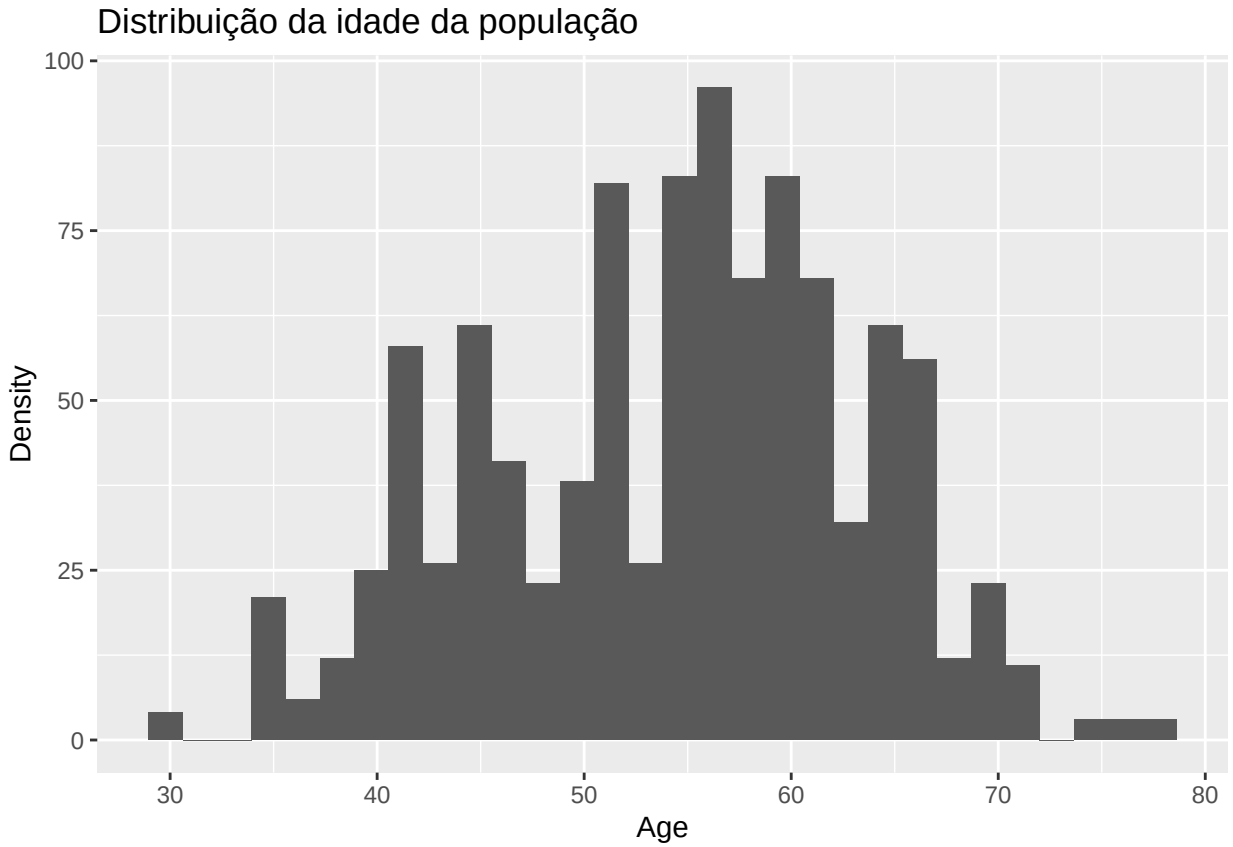
Idade da População

Há o senso comum de que os idosos são mais propensos a apresentar doenças cardíacas.

No gráfico a seguir mostra-se a distribuição da idade da população no conjunto de dados.

```
library(ggplot2)
ggplot(heart, aes(x=age)) +
  geom_histogram() + ggtitle("Distribuição da idade da população")+
  xlab("Age") + ylab("Density")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



A distribuição é quase normal e ligeiramente enviesada para a direita. A maioria da população encontra-se na faixa etária de 50 a 65 anos. Muito poucas pessoas estão na faixa dos trinta e muito poucas pessoas na faixa dos 70 anos.

Em vez de olhar para cada idade em particular, olhar para a faixa etária pode ser mais significativo em termos de taxa de doenças cardíacas.

Por esta razão, a a variável age ('idade') foi dividida para formar diferentes grupos de idade e uma coluna (variável) foi criada com o nome 'idade_grp'.

```
heart$age_grp = cut(heart$age, breaks = seq(25, 77, 4))
```

Agora, poderemos descobrir o número de pessoas com doenças cardíacas para cada faixa etária.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

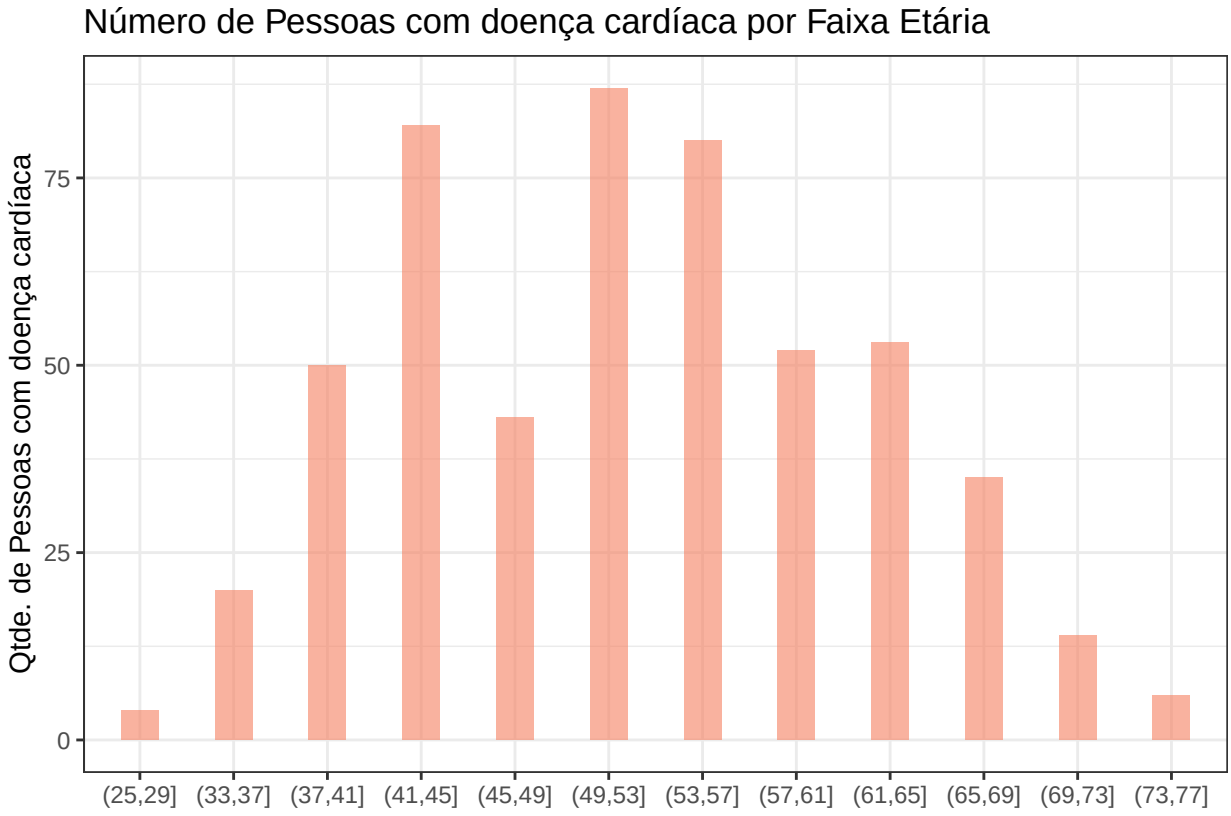
```
target_by_age = heart %>%
        group_by(age_grp) %>%
        summarise(heart_disease = sum(target))
target_by_age
```

```
## # A tibble: 12 x 2
##    age_grp heart_disease
##    <fct>           <int>
##  1 (25,29]             4
##  2 (33,37]            20
##  3 (37,41]            50
##  4 (41,45]            82
##  5 (45,49]            43
##  6 (49,53]            87
##  7 (53,57]            80
##  8 (57,61]            52
##  9 (61,65]            53
## 10 (65,69]            35
## 11 (69,73]            14
## 12 (73,77]             6
```

O resultado mostra a faixa etária das pessoas e a quantidade de pessoas cardíacas em cada grupo.

Para descobrir a frequência de pessoas com doenças cardíacas em cada faixa etária, o gráfico de barras a seguir mostra a distribuição com esses dados.

```
target_by_age %>%
  ggplot(aes(x=age_grp, y=heart_disease)) +
    geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
    xlab("") + ylab("Qtde. de Pessoas com doença cardíaca") + ggtitle("Número de Pessoas com doença car
    theme_bw()
```



Este gráfico mostra que a faixa etária entre 49 e 57 anos possui o maior número de pessoas com doença cardíaca. Este número é ainda maior do que as pessoas acima de 57 anos.

Por outro lado, pessoas com menos de 30 anos e acima de 73 anos têm um número semelhante de pacientes com doenças cardíacas. Esse fenômeno pode ocorrer porque há muito menos pessoas na faixa dos 30, 40 e 70 anos na amostra.

Para entender um pouco este fenômeno um pouco melhor, a proporção de pacientes com doenças cardíacas em cada faixa etária poderá ajudar. *Para isso, vamos encontrar a proporção de pessoas com doenças cardíacas em cada grupo.*

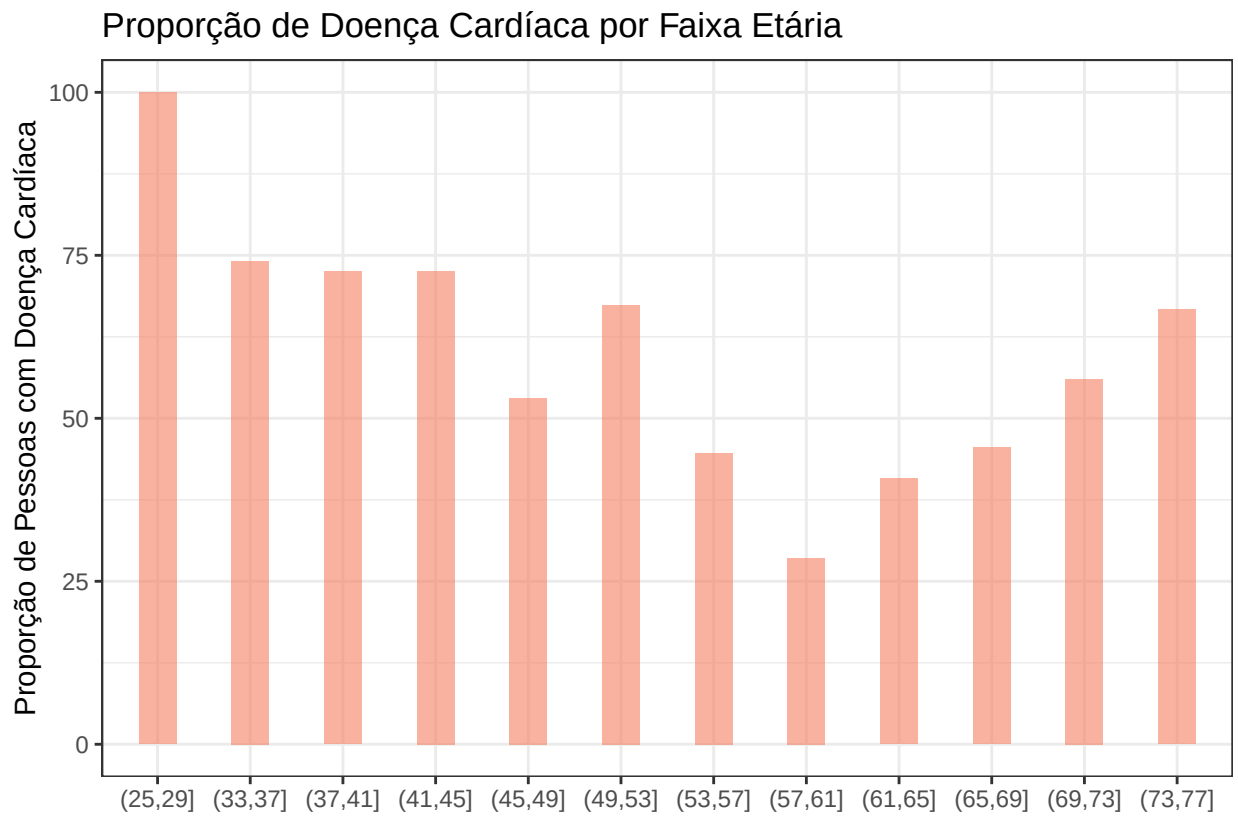
```
prop_in_age = heart %>%
        group_by(age_grp) %>%
        summarise(heart_disease_proportion = round(sum(target)/n(), 3)*100)
prop_in_age
```

```
## # A tibble: 12 x 2
##    age_grp heart_disease_proportion
##    <fct>                      <dbl>
##  1 (25,29]                    100
##  2 (33,37]                     74.1
##  3 (37,41]                     72.5
##  4 (41,45]                     72.6
##  5 (45,49]                     53.1
##  6 (49,53]                     67.4
##  7 (53,57]                     44.7
##  8 (57,61]                     28.6
##  9 (61,65]                     40.8
## 10 (65,69]                     45.5
```

```
## 11 (69,73]                    56
## 12 (73,77]                    66.7
```

E o gráfico de barras correspondente a esses dados é mostrado a seguir:

```
prop_in_age %>%
  ggplot(aes(x=age_grp, y=heart_disease_proportion)) +
    geom_bar(stat="identity", fill="#f68060", alpha=.6, width=.4) +
    xlab("") + ylab("Proporção de Pessoas com Doença Cardíaca") + ggtitle("Proporção de Doença Cardíaca
    theme_bw()
```



Na faixa etária abaixo de 30 anos, 100% das pessoas têm doenças cardíacas! Definitivamente, esse não é o caso na vida real. É claro que esta não é uma amostra representativa.

Infelizmente, não é possível inferir qualquer conclusão sobre como a idade contribui para as doenças cardíacas a partir deste conjunto de dados...

Gênero ou sexo

Antes de examinar a relação entre doenças cardíacas e gênero, é importante saber a proporção de homens e mulheres neste conjunto de dados.

```
round(prop.table(table(heart$sex)),2)
```

```
##
## female   male
##    0.3    0.7
```

A proporção de homens e mulheres não são semelhantes. 30% do conjunto de dados é feminina e 70% da amostra é masculina.

A proporção de homens e mulheres com doenças cardíacas pode ser a próxima descoberta importante.

```
round(prop.table(table(heart$sex, heart$target1)), 2)
```

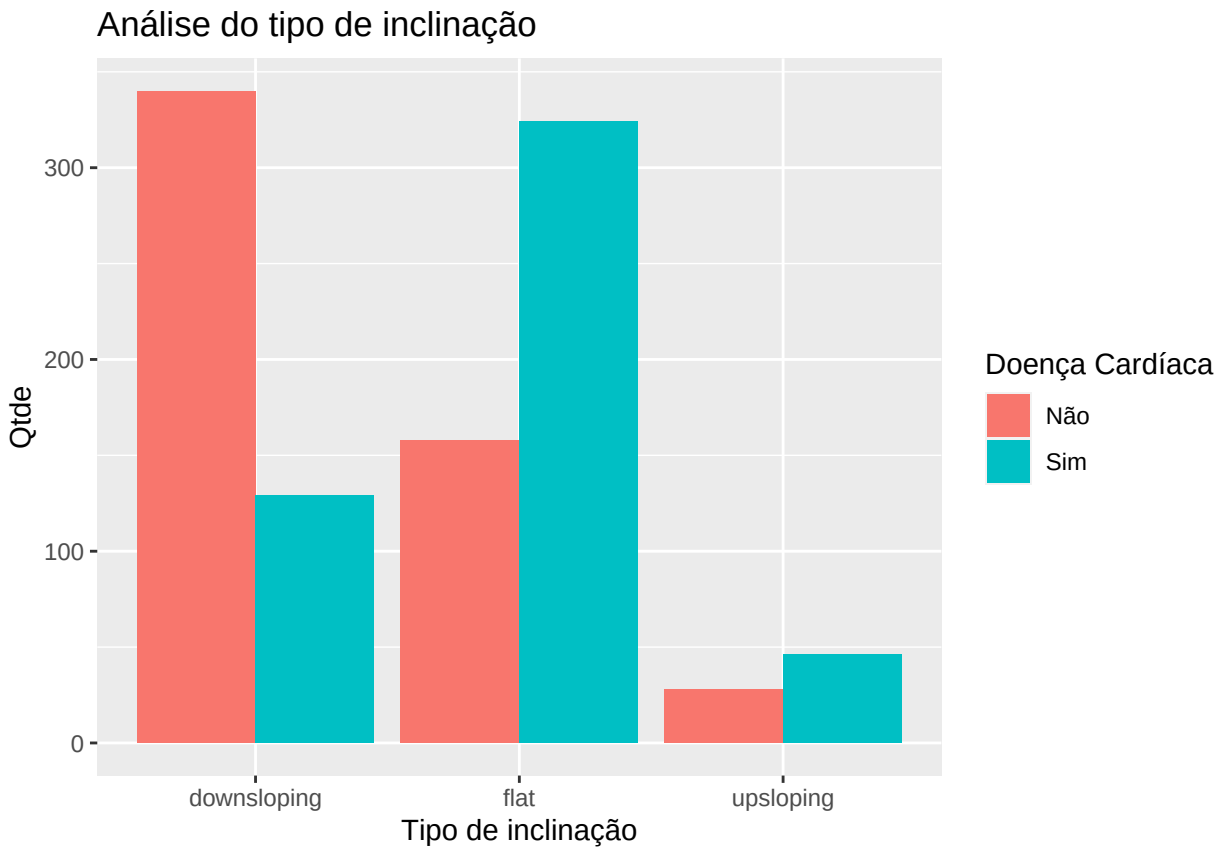
```
##
##          heart disease no heart disease
##   female          0.22             0.08
##   male            0.29             0.40
```

Pessoas com doenças cardíacas são muito mais numerosas na população feminina do que as pessoas sem doenças cardíacas. Ao mesmo tempo, na população masculina, 40% não têm doenças cardíacas enquanto apenas 29% têm doenças cardíacas.

A inclinação do segmento ST do exercício de pico (slope)

Um gráfico de barras de diferentes tipos de declive e condições de doenças cardíacas poderá ser mais apropriado para compreendê-lo.

```
ggplot(heart, aes(x= slope, fill=target1)) +
  geom_bar(position = 'dodge') +
  xlab("Tipo de inclinação") +
  ylab("Qtde") +
  ggtitle("Análise do tipo de inclinação") +
  scale_fill_discrete(name = "Doença Cardíaca", labels = c("Não", "Sim"))
```



Claramente, com diferentes tipos de inclinação, a taxa de doenças cardíacas parece diferente. Com declive, o número de sem doenças cardíacas é muito maior (cerca de 340) do que o número de pacientes com doenças cardíacas (cerca de 125). Mas com uma superfície plana é quase o oposto. O número de cardiopatias é de cerca de 325 e o número de não cardiopatas é de cerca de 160. Na tendência ascendente, não há muitas diferenças, mas o número de cardiopatias é maior do que o número de casos sem cardiopatia.

Essa tendência é a mesma na população masculina e feminina?

Outra pergunta válida para responder. O mesmo tipo de gráfico de barras para a parte masculina e feminina do conjunto de dados ajudará a entender isso.

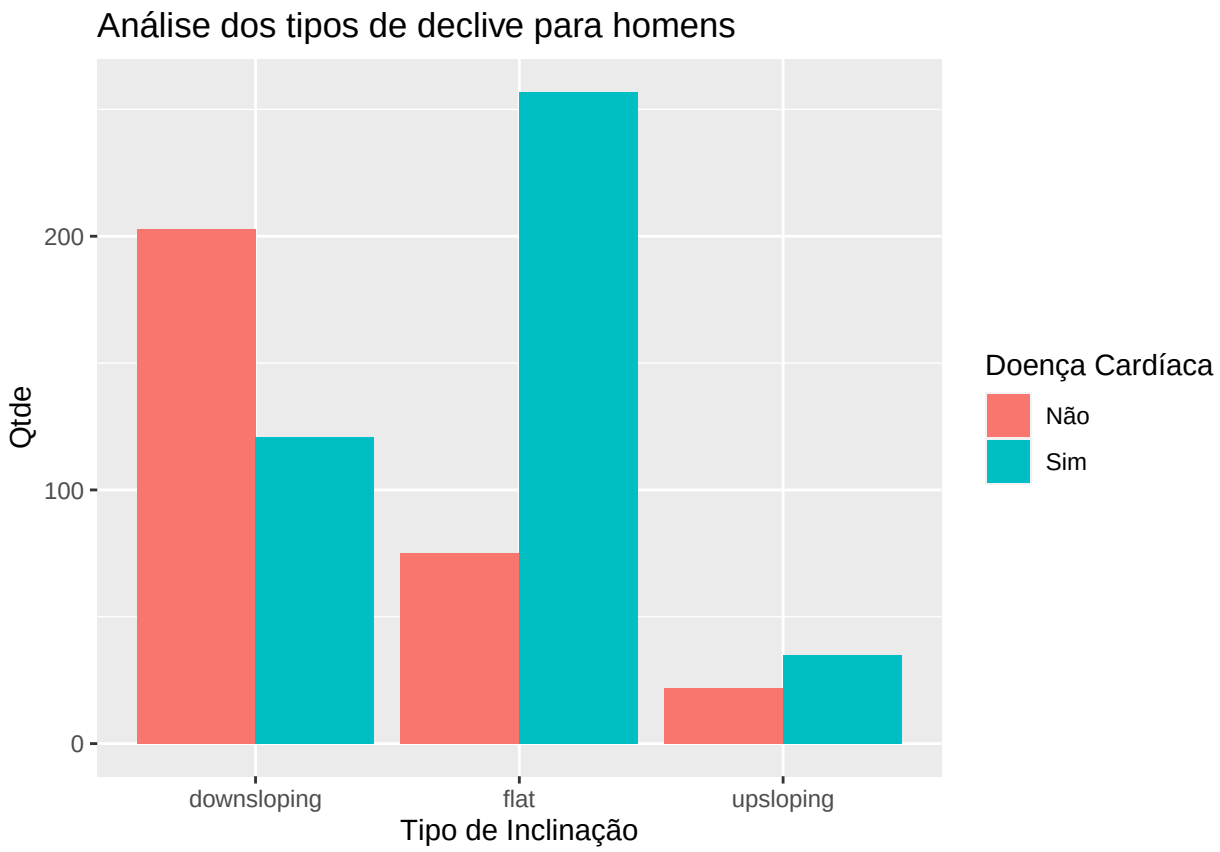
Primeiro, o conjunto de dados será separado para a população masculina e feminina:

```
male_data = heart[heart$sex=="male",]

female_data = heart[heart$sex=="female",]
```

Agora, faremos o mesmo gráfico de barras para a população masculina e feminina.

```
ggplot(male_data, aes(x= slope, fill=target1)) +
  geom_bar(position = 'dodge') +
  xlab("Tipo de Inclinação") +
  ylab("Qtde") +
  ggtitle("Análise dos tipos de declive para homens") +
  scale_fill_discrete(name = "Doença Cardíaca", labels = c("Não", "Sim"))
```



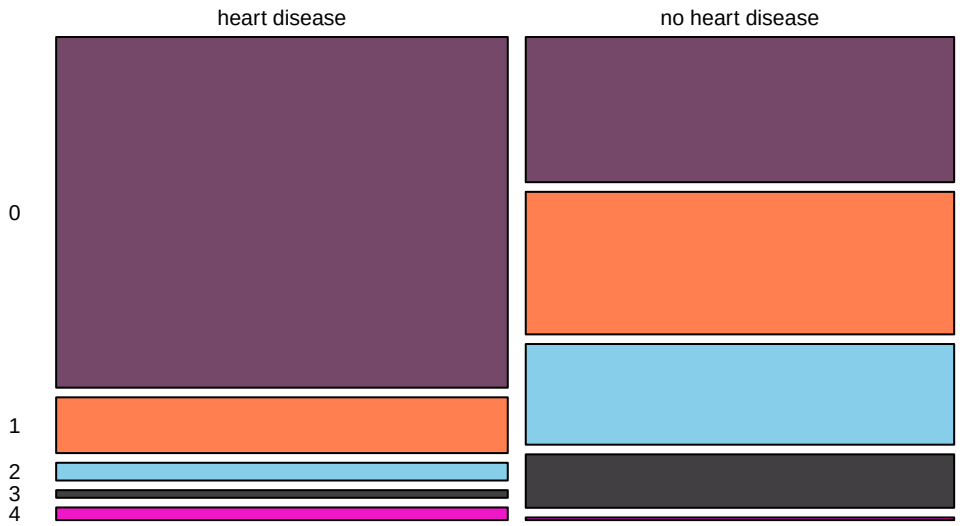
```
ggplot(female_data, aes(x= slope, fill=target1)) +
  geom_bar(position = 'dodge') +
  xlab("Tipo de Inclinação") +
  ylab("Qtde") +
  ggtitle("Análise dos tipos de declive para mulheres") +
  scale_fill_discrete(name = "Doença Cardíaca", labels = c("Não", "Sim"))
```

O gráfico da população masculina segue a mesma tendência do gráfico de barra geral para análise de declive. Mas na população feminina, a tendência é muito diferente. O número decrescente de nenhuma doença cardíaca é muito maior (180) do que o número de doenças cardíacas (25). Novamente, para a inclinação plana, ambos os casos são próximos, mas o número de casos sem doença cardíaca é um pouco maior.

Número de principais vasos (ca)

O conjunto de dados mostra que pode haver 0, 1, 2, 3 ou 4 vasos principais do coração em uma pessoa. De acordo com o gráfico de correlação, o número de vasos tem uma boa correlação com as doenças cardíacas. A representação visual de quão diferente o número de vasos principais se relaciona com as doenças cardíacas é mostrada no gráfico a seguir:

```
mosaicplot(table(heart$target1, heart$ca), col=c("#754869", "coral", "skyblue", "#423f42", "#ed18c6"),
```

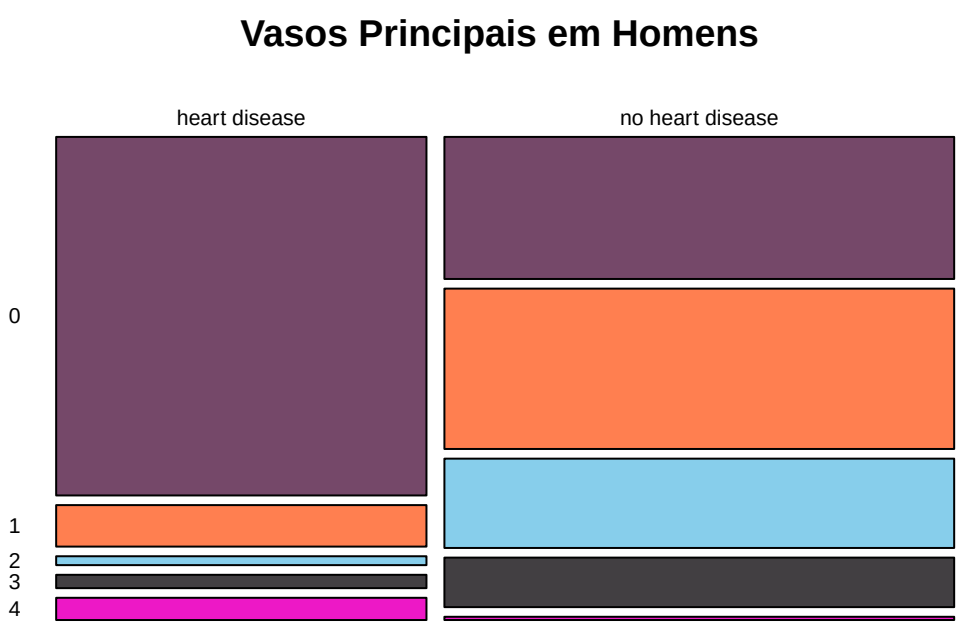


Cerca de 2/3 das pessoas com doença cardíaca não apresentam nenhum vaso importante. Muito poucas pessoas têm 4 vasos principais. Portanto, é difícil saber o impacto disso.

As populações masculinas e femininas podem ter um número diferente de vasos principais ou níveis diferentes de relacionamento entre os vasos principais e as doenças cardíacas.

O gráfico a seguir mostra os principais vasos vs doenças cardíacas em homens:

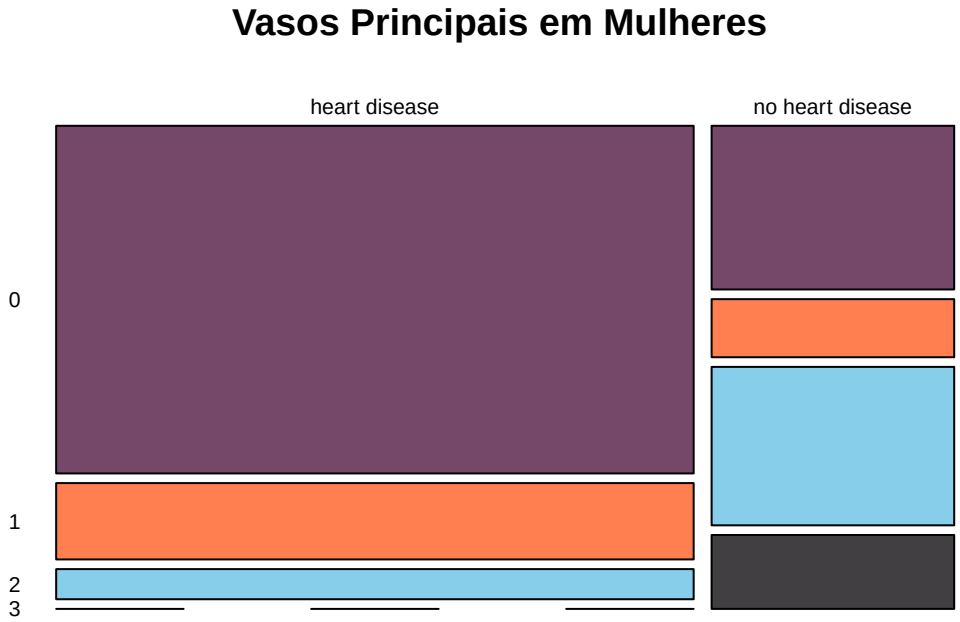
```
mosaicplot(table(male_data$target1, male_data$ca), col=c("#754869", "coral", "skyblue", "#423f42", "#ed
```



O gráfico dos dados masculinos da amostra parece seguir uma tendência muito semelhante à da população total dos vasos principais.

A seguir apresentamos o gráfico que mostra a correlação do número de vasos principais e doenças cardíacas na população feminina:

```
mosaicplot(table(female_data$target1, female_data$ca), col=c("#754869", "coral", "skyblue", "#423f42",
```

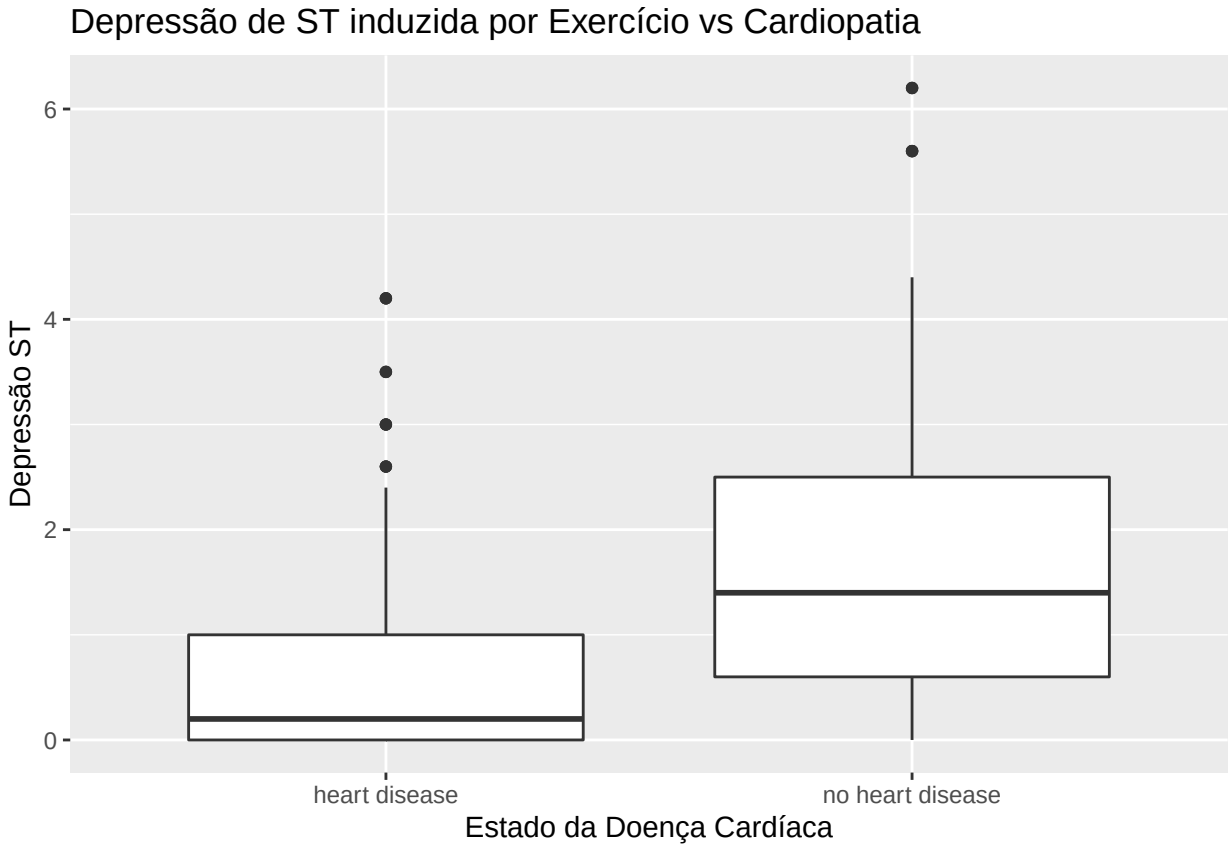


Na população feminina, existem 0, 1, 2 ou 3 vasos principais. Nenhuma mulher tem 4 vasos principais. Como na população masculina, o máximo de mulheres com doenças cardíacas não tem vasos principais. Mais uma vez, na zona sem doença cardíaca, a maioria das mulheres tem 0 ou 2 vasos principais.

Depressão de ST induzida por exercício em relação ao repouso (oldpeak)

Aqui estão os boxplots que mostram a distribuição da depressão do segmento ST para pessoas com doenças cardíacas e sem doenças cardíacas.

```
ggplot(heart, aes(x = target1, y = oldpeak)) + ylab("Depressão ST") + xlab("Estado da Doença Cardíaca")+
        geom_boxplot()
```



No lado sem doenças cardíacas, o intervalo interquartil é maior (cerca de 2) do que no lado das doenças cardíacas (1).

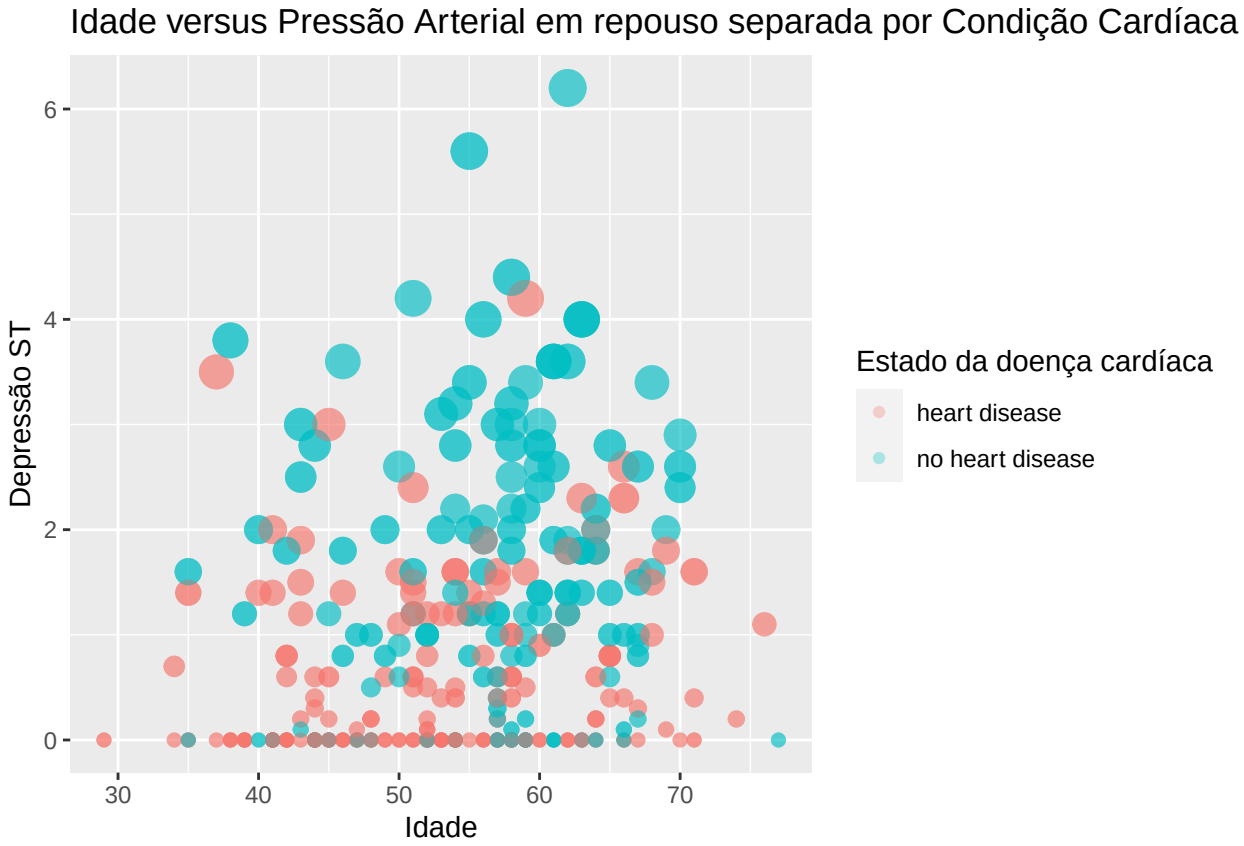
Esse tipo de depressão muda com a idade e, juntos, eles têm impactos diferentes nas doenças cardíacas?

Um gráfico de dispersão combinado pode fornecer alguns insights sobre isso.

```
ggplot(heart, aes(x = age, y = oldpeak,color=target1, size = factor(oldpeak))) +
    geom_point(alpha=0.3) + labs(color = "Estado da doença cardíaca")+guides(size=FALSE) + xlab("Idade")
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

```
## Warning: Using size for a discrete variable is not advised.
```

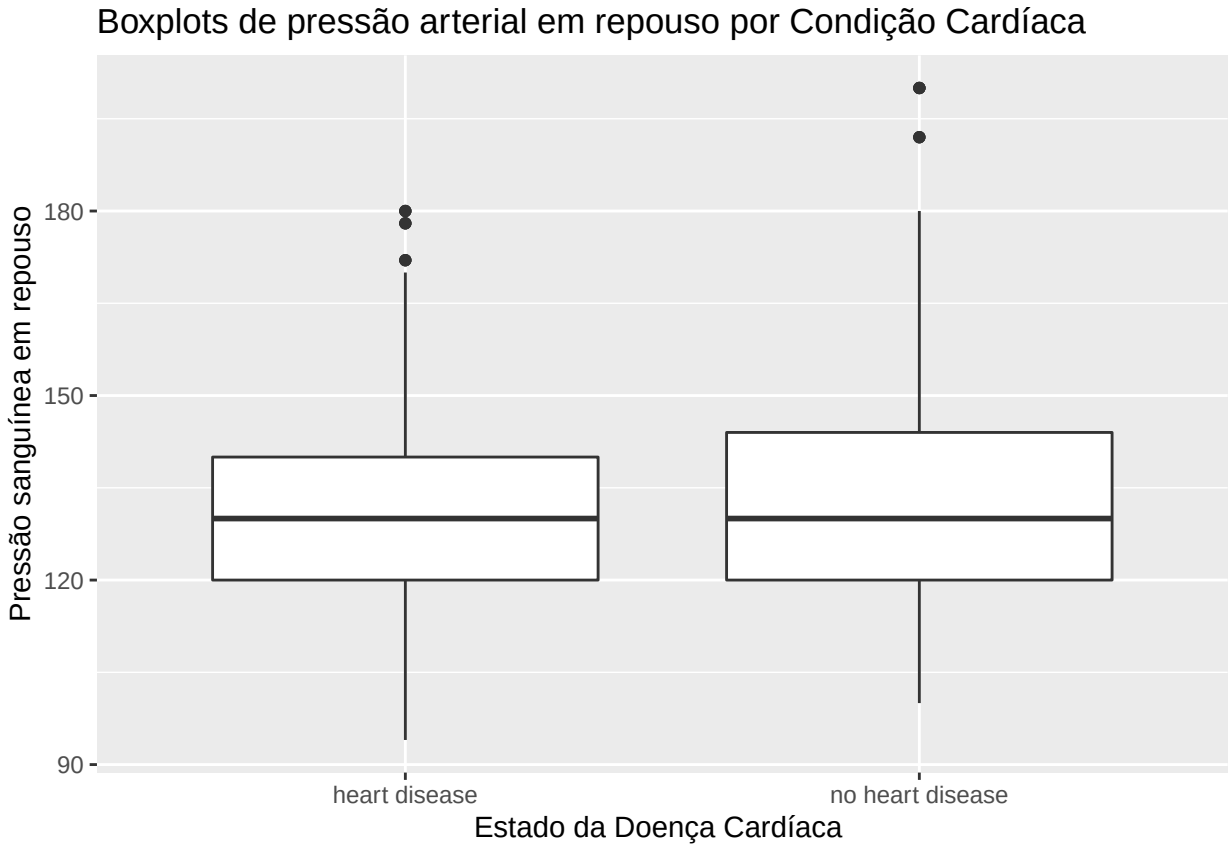


Conforme discutido no início, este conjunto de dados é diferente. Ele mostra as doenças cardíacas diminuem quanto mais elevada a idade. Parece que quando a depressão do ST aumenta, os casos de doenças cardíacas diminuem. O tamanho dos pontos muda com o açúcar no sangue em repouso. Mas, a partir dessa imagem, é difícil derivar qualquer relação entre idade e depressão de ST.

Pressão sanguínea em repouso

Boxplots de açúcares no sangue em repouso separados por estado de doença cardíaca poderão fornecer uma ideia inicial, como mostra o gráfico a seguir.

```
ggplot(heart, aes(x = target1, y = trestbps)) +
        geom_boxplot() + xlab("Estado da Doença Cardíaca ") + ylab("Pressão sanguínea em repouso") + gg
```



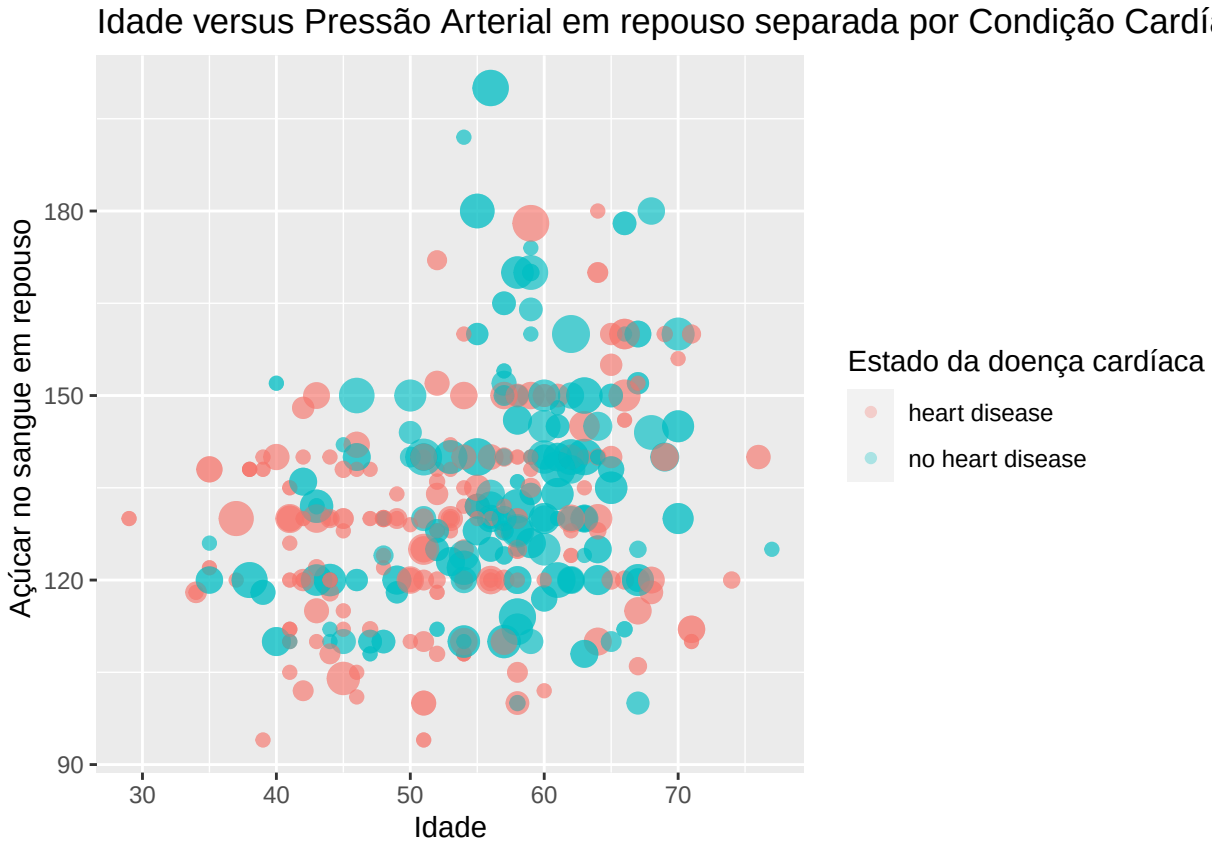
O gráfico acima mostra que uma faixa interquartil de açúcar no sangue em repouso é ligeiramente maior para o gráfico sem doença cardíaca. Mas as medianas de ambos os gráficos de caixa parecem iguais.

O próximo gráfico é um gráfico de dispersão de idade versus pressão arterial em repouso, que inclui **cores diferentes para o estado de doença cardíaca e o tamanho do ponto depende da depressão St**. Este gráfico deve revelar mais algumas informações.

```
ggplot(data=heart,aes(x=age,y=trestbps,color=target1,size=factor(oldpeak)))+
geom_point(alpha=0.3)+
  xlab("Idade")+
  ylab("Açúcar no sangue em repouso ") +
  labs(color="Estado da doença cardíaca ") +
  guides(size=FALSE)+
  ggtitle("Idade versus Pressão Arterial em repouso separada por Condição Cardíaca ")
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

```
## Warning: Using size for a discrete variable is not advised.
```



Este gráfico mostra algo muito interessante. Quando o açúcar no sangue em repouso está realmente baixo, como 100 ou menos, os casos com doença cardíaca são mais elevados do que os casos sem doença cardíaca. Quando a pressão arterial em repouso está acima de 165, os casos sem doença cardíaca é maior do que os casos com doença cardíaca.

Como vimos antes, a maioria dos pontos grandes são azuis. Isso significa que mais pessoas com depressão de ST não têm doenças cardíacas. Ao mesmo tempo, um número maior de pontos maiores está na faixa etária mais alta. Portanto, a depressão do ST é maior em pessoas mais velhas.

Conclusão

Tentamos mostrar algumas visualizações e técnicas para resumir um conjunto de dados.

Este conjunto de dados não é muito grande. Possui apenas 14 colunas. Mesmo assim, há muito mais que poderia ser explorado. Existem variáveis que não foram estudadas.

Portanto, sinta-se à vontade para explorar um pouco mais por conta própria.