Análise Probabilística
Detecção de Patologia Cardíaca Pediátrica

Christophe Ferreira (up201003838)
Pedro Bastos (up201002595)

Novembro 2015

Conteúdo

Dados Originais

A preparação dos dados consiste em construir um *dataset* para uma ou mais origens de dados, de modo a que estes possam ser modulados e analisados. Uma prática habitual, é a criação de um *dataset* inicial para uma aprendizagem dos dados tratados e uma melhor percepção de possíveis problemas com a qualidade dos mesmo. Esta preparação é um processo longo e propício a erros de eliminação de dados que parecem ser irrelevantes (dados inválidos, fora do intervalo ou ausentes), esta informação é normalmente resultado de uma má recolha de dados. Uma análise de dados que não é realizada cuidadosamente pode levar a uma interpretação errada dos resultados obtidos, assim, o sucesso para uma boa análise de dados depende fundamentalmente da qualidade do pré-processamento da informação [1].

Descrição das Variáveis

Para efectuar o pré-processamento de dados é necessário estar familiarizado com estes, o conhecimento sobre a informação é uma das primeiras e mais importantes tarefas a ser realizadas na análise de dados. A informação é geralmente o resultado de medições (numérica) ou de contagem (categórica). As variáveis servem para nomear um conjunto de dados e são normalmente divididas em dois tipos (ver figura 1), discretas e contínuas.

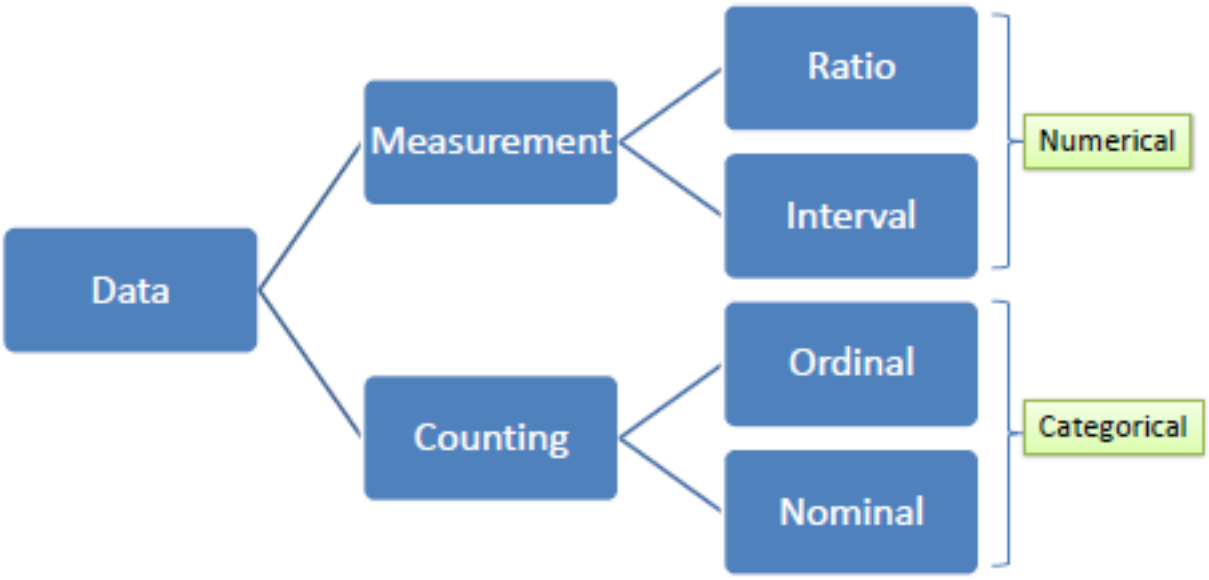

Figura 1: Tipos de variáveis

Uma variável discreta ou categórica pode aceitar dois ou mais valores (categorias). Existem dois tipos de categorias, a nominal e a ordinal. Dados nominais não têm uma ordenação intrínseca nas suas categorias. Por exemplo, o sexo, que pode ser tanto masculino como feminino. Já os dados ordinais têm uma ordenação associada a sua categoria. Por exemplo, o nível de combustível, classificado ordenadamente em três categorias (baixo, médio e alto).

Uma variável contínua ou numérica permite qualquer valor que pertença a um conjunto finito ou infinito, por exemplo, peso, altura, temperatura. No entanto, é possível encontrar dois tipos de dados numéricos, intervalos e rácios. Dados pertencentes a um determinado intervalo podem ser adicionados ou subtraídos, mas não podem ser multiplicados ou divididos porque não possuem um verdadeiro zero, ou seja, não têm uma origem definida. Neste tipo de dados a diferença entre valores é relevante, existe também uma unidade de medição. Por exemplo, não podemos dizer que um dia foi duas vezes mais quente que outro dia. Por outro lado, dados numa escala de rácio têm um zero e podem ser adicionados, subtraídos, multiplicados e divididos. Temos como exemplos, as unidades monetárias, a idade ou o comprimento.

Com base no que anteriormente foi referido, será feita agora uma descrição das variáveis contidas nesta análise de dados.

- **Variáveis Discretas**

Convénio: Variável do tipo nominal. Esta guarda a companhia responsável pelo seguro de saúde do paciente.

Pulsos: Esta variável é do tipo ordinal. Esta especifica o tipo de pulso registado (normal, diminuído ou amplo).

***PPA*:** A variável é do tipo ordinal. A amplificação da pressão do pulso (PPA) serve para auxiliar na detecção de riscos cardiovasculares.

Normal X Anormal: Variável do tipo nominal. Indica se o paciente sofre ou não de alguma patologia cardíaca.

***B2*:** Variável do tipo nominal. Esta classifica o tipo do segundo som cardíaco.

Sopro: A variável é do tipo nominal. Especifica o grau do sopro escutado no paciente.

***HDA* 1:** Esta variável é do tipo nominal. Indica o histórico de doenças cardíacas do paciente.

***HDA* 2:** Variável do tipo nominal. Mostra mais alguma doença cardíaca que o paciente padece/-padeceu.

Sexo: A variável é do tipo nominal. Representa o género do paciente, ou seja, se este é masculino ou feminino.

Motivo 1: Esta variável é do tipo nominal. Especifica o motivo pelo qual o paciente foi encaminhado para a cardiologia.

Motivo 2: A variável é do tipo nominal. Indica outro motivo pelo qual o paciente foi encaminhado para a cardiologia.

- **Variáveis Contínuas**

***ID*:** Identificador do paciente. Esta é uma variável do tipo rácio, guarda números inteiros sequencialmente.

Peso: Variável do tipo rácio. Armazena o peso dos pacientes, e a sua unidade corresponde à do sistema internacional (quilograma).

Altura: Esta variável é do tipo rácio. Guarda a altura dos pacientes, e a sua unidade está em conformidade com sistema internacional (metro).

***IMC*:** A variável é do tipo rácio. O Índice de Massa Corporal (IMC) é calculado com base no peso e na altura, sendo que será considerado que um paciente tem excesso de peso se tiver um ICM entre os 85 e os 95, aqueles com mais de 95 serão classificados como obesos. Fórmula para o cálculo do índice de massa corporal:

$$ICM = \frac{Peso}{(Altura)^2}$$

Atendimento: Variável do tipo intervalo. Corresponde a data em que o paciente foi atendido, como tal, coincide com a data em que foi feita a recolha dos dados.

***DN*:** A variável é do tipo intervalo. A Data de Nascimento (DN) do paciente esta formatada como dia/mês/ano.

Idade: Esta variável é do tipo rácio. Guarda a idade do paciente, está num formato decimal, pois existem crianças com menos de um ano de idade nos dados recolhidos. A unidade utilizada é o ano. Fórmula para o cálculo da idade:

$$Idade = Atendimento - DataDeNascimento$$

***PA* Sistólica:** Variável do tipo intervalo. A Pressão arterial sistólica refere-se à força criada pela tensão nas artérias quando o coração se contrai e bombeia sangue para estas [2], é medida em milímetros de mercúrio ($mmHg$).

***PA* Diastólica:** A variável é do tipo intervalo. A Pressão Arterial Diastólica indica que tensão está presente nas artérias entre batimentos cardíacos, quando o coração está em repouso. É apresentada como o valor mais baixo quando a tensão arterial é medida [2], a unidade também é *mmHg*.

***FC*:** A variável é do tipo intervalo. Aqui é guardado a frequência cardíaca do paciente, ou seja, a quantidade de vezes que o coração bate por minuto (Batimentos/minuto).

Análise Preliminar

Nesta secção vamos analisar os dados em bruto, ou seja, vamos dissecar a informação sem que tenha sido aplicado qualquer tipo de tratamento nos dados. A informação é apresentada como foi recolhida, vamos usar diversos tipos de gráficos para cada variável de modo a que seja mais percetível as oscilações desta. É possível ver algumas situações bizarras nos gráficos exibidos, este facto deve-se à erros cometidos na recolha dos dados. Nesta secção não vamos aprofundar esses detalhes, isso será feito na próxima secção, demonstrando também as medidas tomadas para eliminar essa imperfeições na informação.

- **Peso**

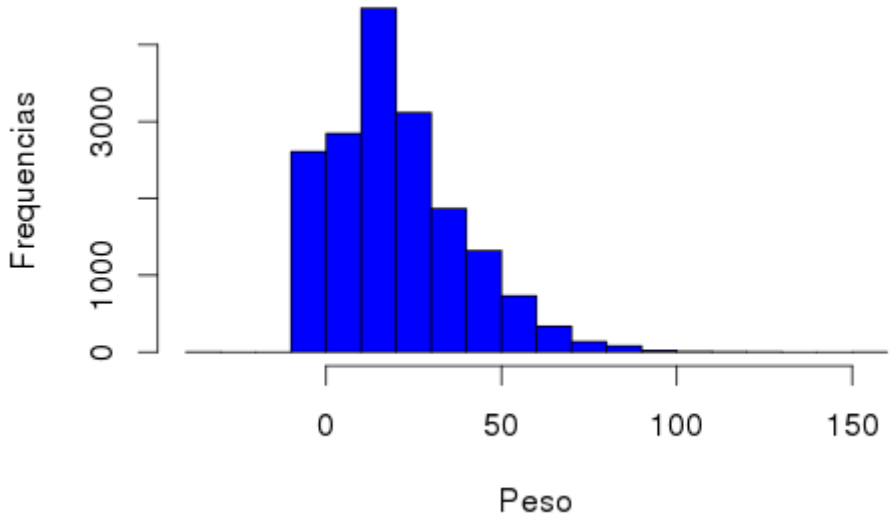

Figura 2: Histograma do peso

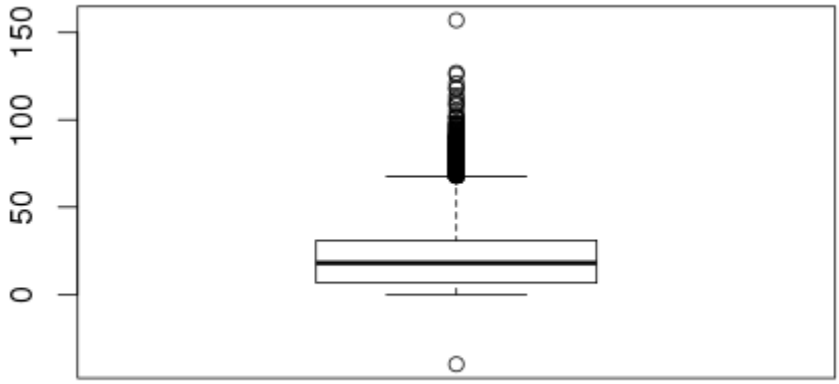

Figura 3: Diagrama de caixa e bigodes do peso

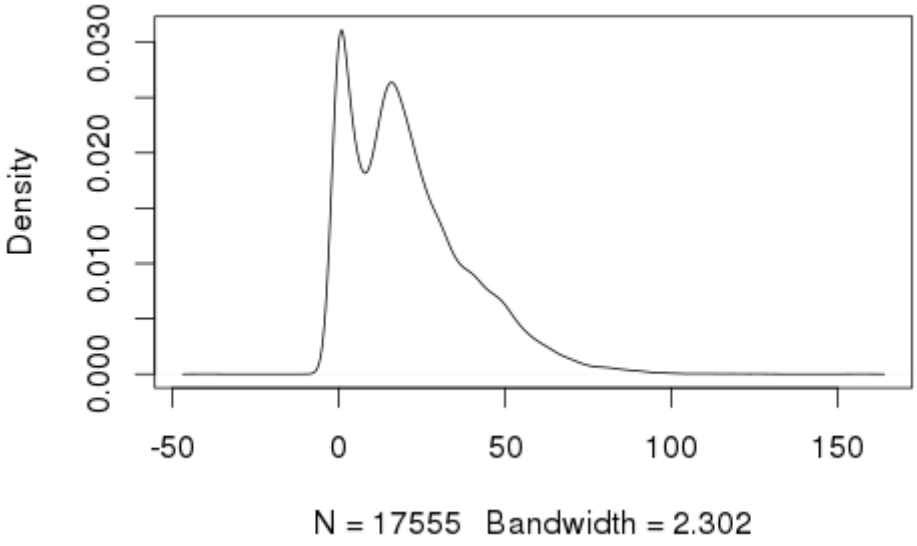

Figura 4: Gráfico de densidades do peso

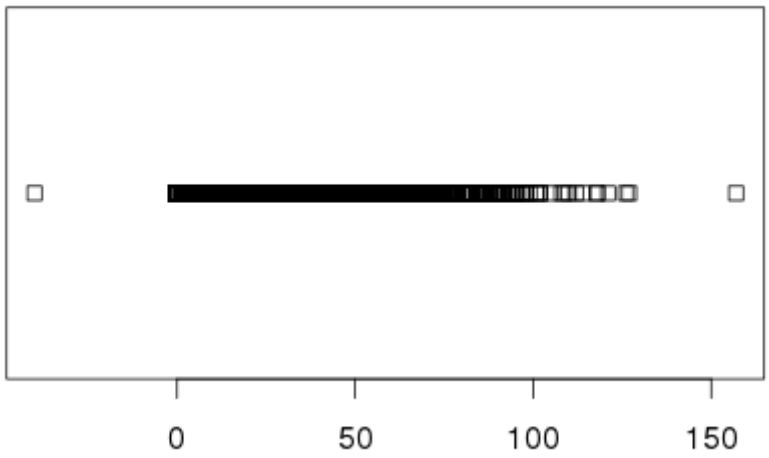

Figura 5: Gráfico de dispersão do peso

É possível verificar com a ajuda dos gráficos apresentados (figuras 2, 3, 4 e 5) que existem pesos negativos, algo que como sabemos é impossível.

- **Altura**

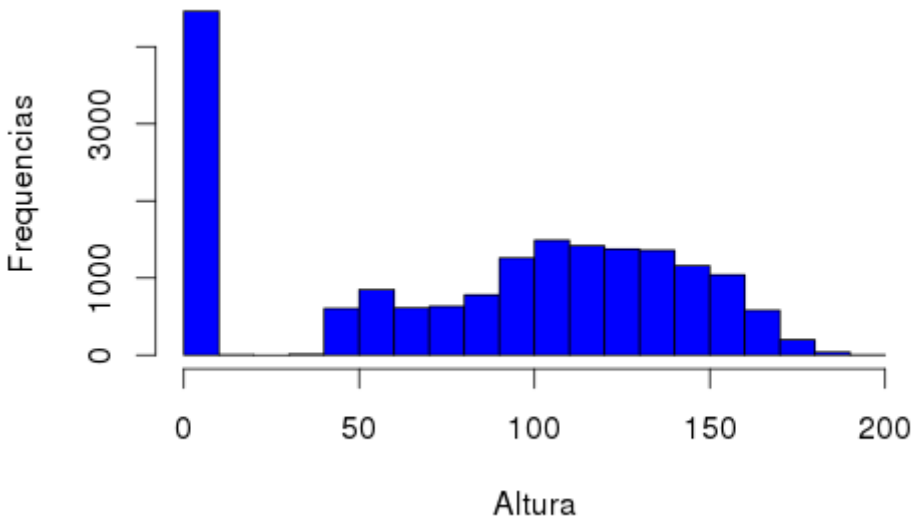

Figura 6: Histograma da altura

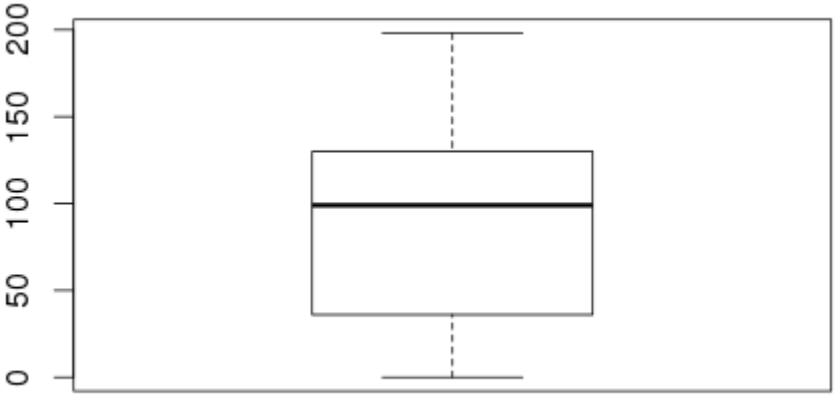

Figura 7: Diagrama de caixa e bigodes da altura

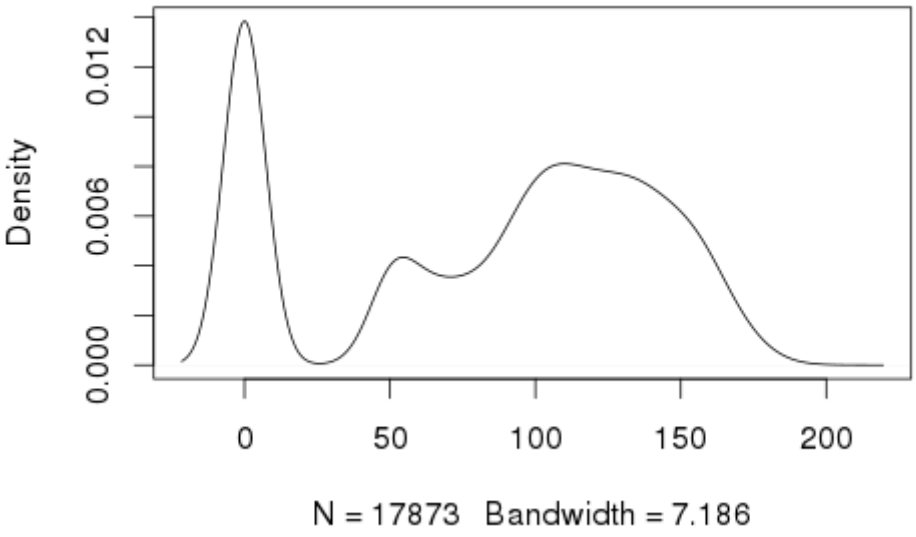

Figura 8: Gráfico de densidades da altura

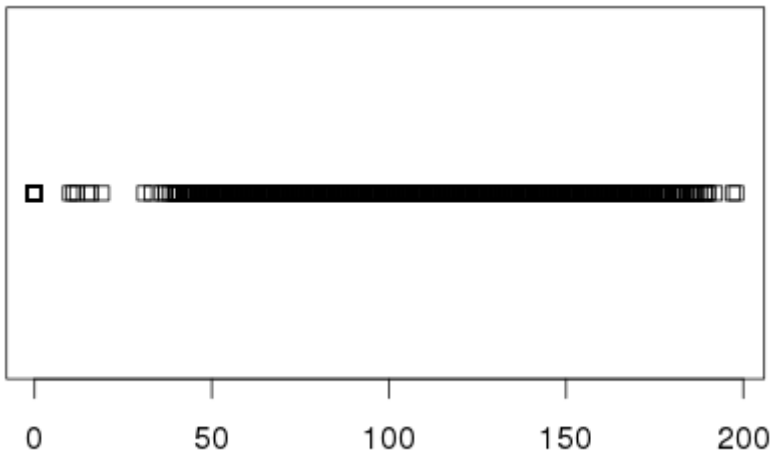

Figura 9: Gráfico de dispersão da altura

É possível observar com a ajuda dos gráficos anteriores (figuras 6, 7, 8 e 9) que ocorrem alturas negativas, como sabemos, isso é irreal.

- ***IMC***

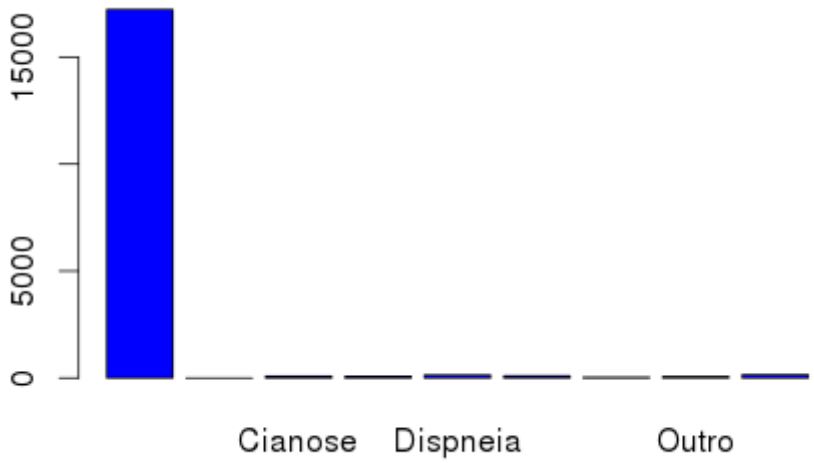

Figura 10: Histograma do *IMC*

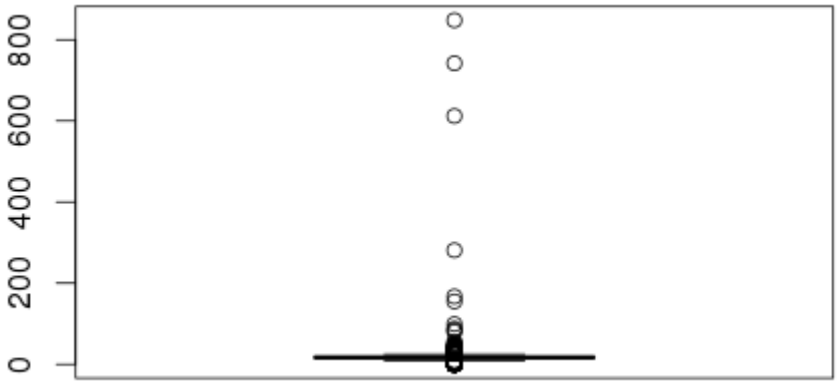

Figura 11: Diagrama de caixa e bigodes do *IMC*

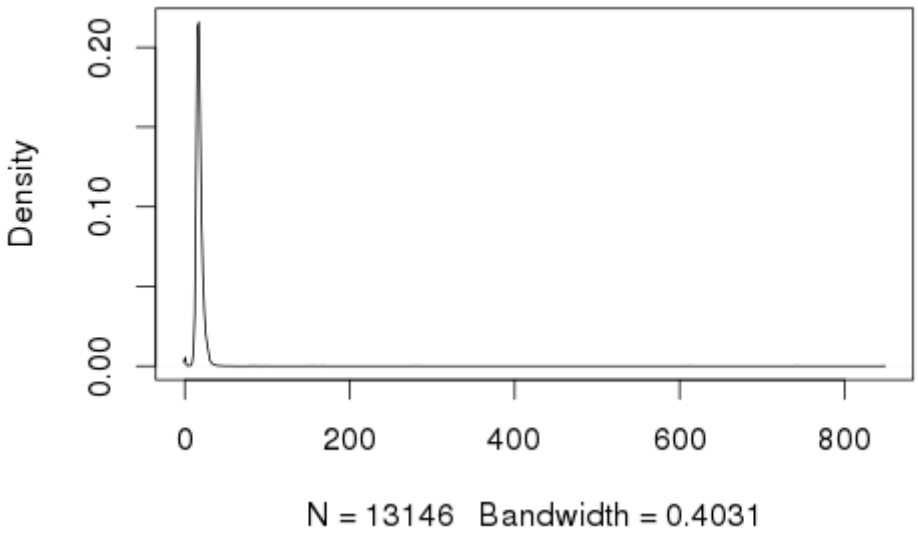

Figura 12: Gráfico de densidades do *IMC*

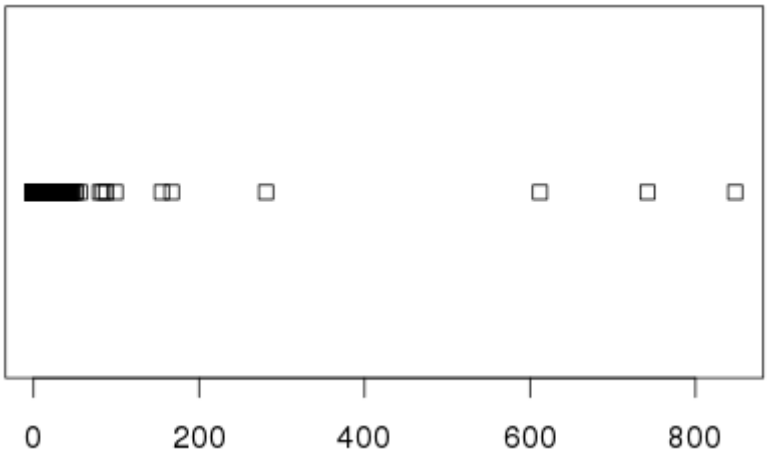

Figura 13: Gráfico de dispersão do *IMC*

Verificamos sem auxílio à qualquer ferramenta de análise, que existem laguns valores extremamente invulgares no *IMC*, estes normalmente deveriam estar situados entre os valores 10 e 50, o que não se verifica neste caso (figuras 11 e 13).

- **Idade**

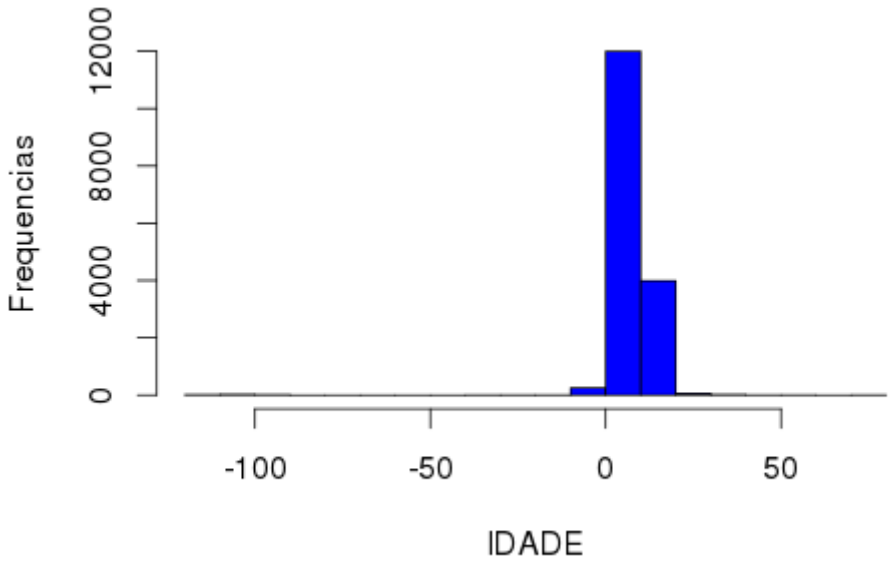

Figura 14: Histograma da idade

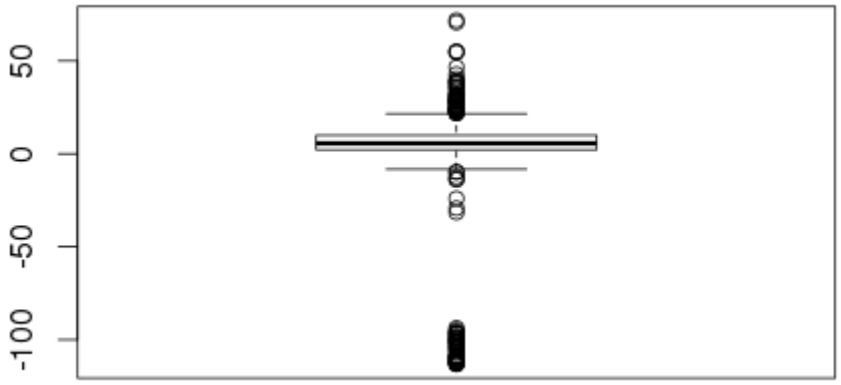

Figura 15: Diagrama de caixa e bigodes da idade

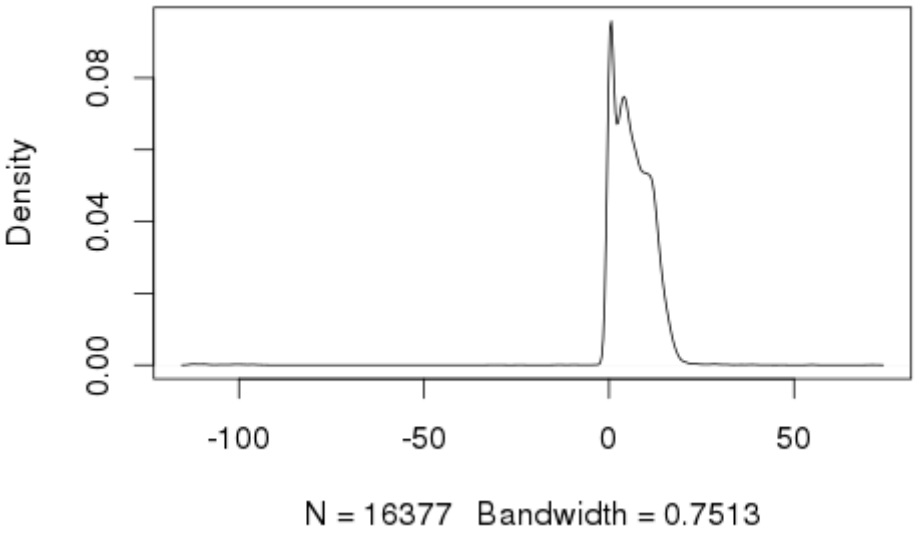

Figura 16: Gráfico de densidades da idade

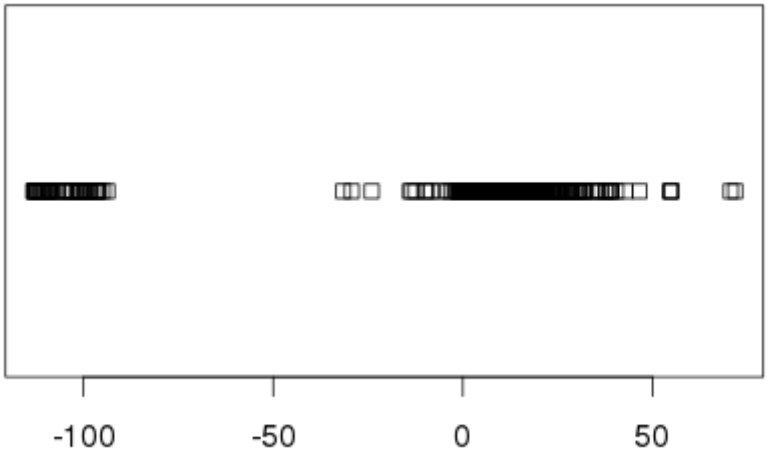

Figura 17: Gráfico de dispersão da idade

Nesta variável podemos verificar que existem idades negativas, que indicam que algo não está bem com os dados originais. Assim como idades superiores a 19 anos que não estão no estudo em questão.

- **Pulsos**

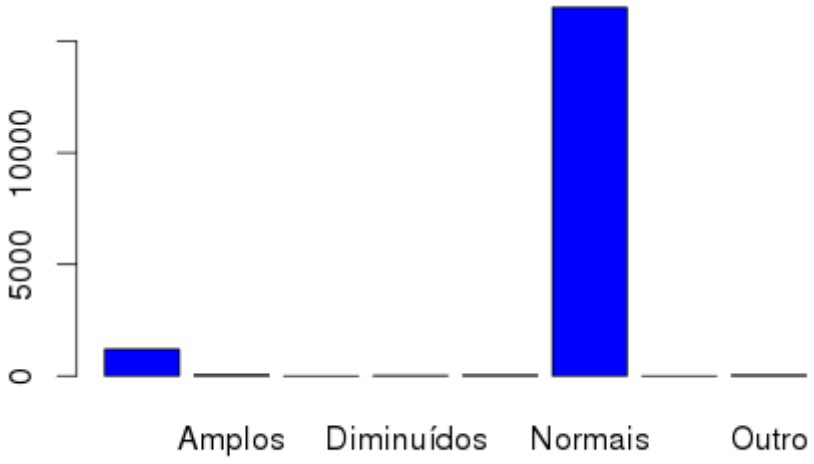

Figura 18: Histograma dos pulsos

Aqui podemos ver diversas barras (figura 18) com valores muito pequenos, existe cinco tipos de pulsos em que o valor mais alto deles era 57, que em comparação com o valor mais elevado (16.509) é bastante mais baixo.

- **PA Sistólica**

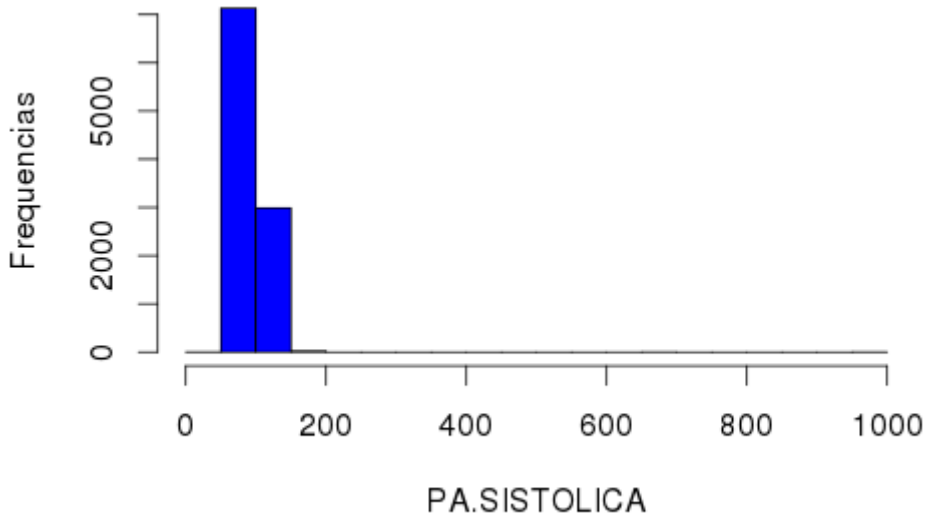

Figura 19: Histograma da pressão arterial sistólica

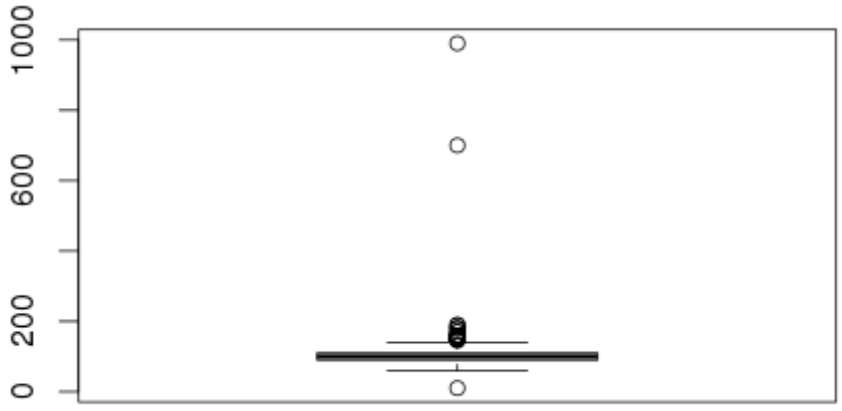

Figura 20: Diagrama de caixa e bigodes da pressão arterial sistólica

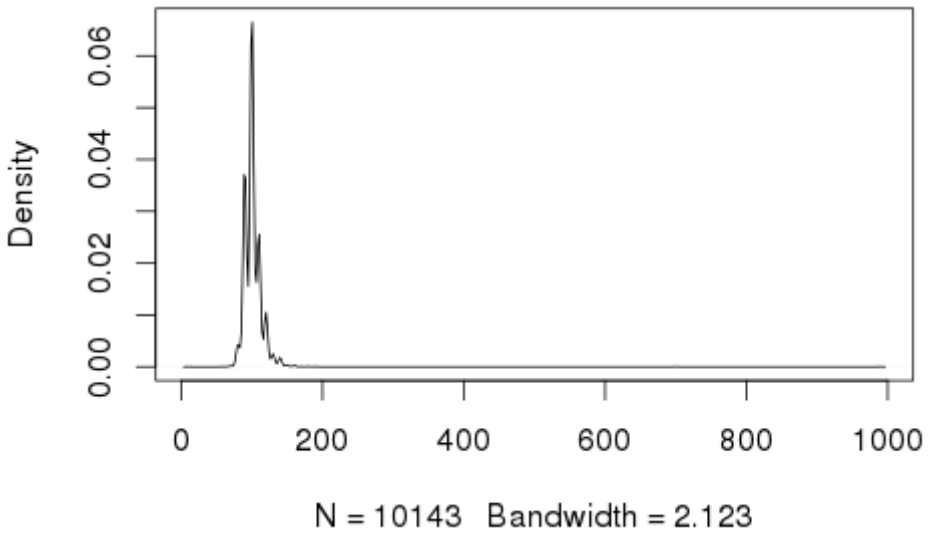

Figura 21: Gráfico de densidades da pressão arterial sistólica

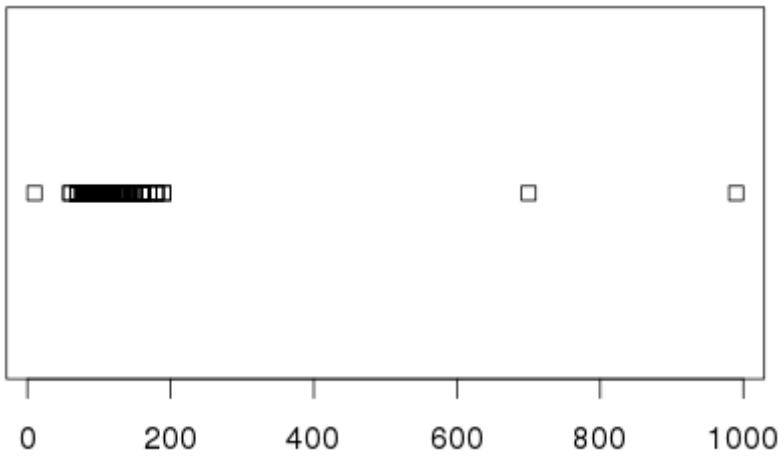

Figura 22: Gráfico de dispersão da pressão arterial sistólica

Podemos observar nos gráficos (figuras 19, 20, 21 e 22) da pressão arterial sistólica alguns *outliers*, como por exemplo, os valores 700 e 990 que são completamente absurdos.

- **PA Diastólica**

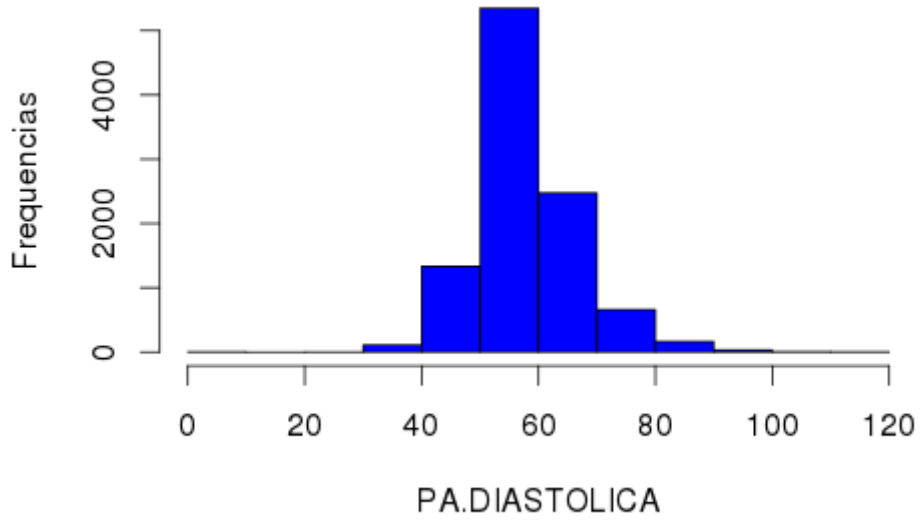

Figura 23: Histograma da pressão arterial diastólica

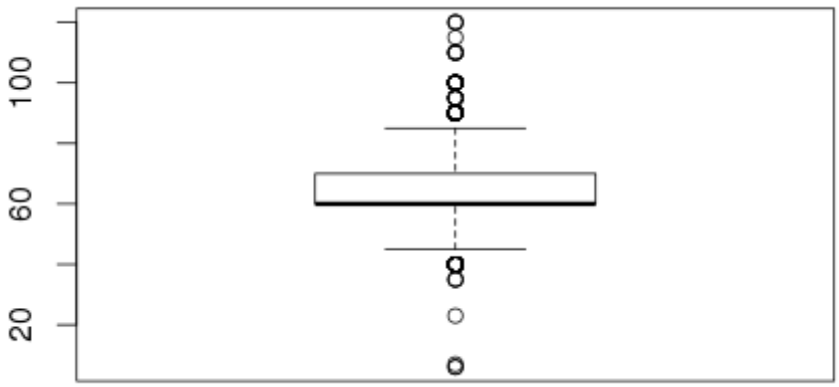

Figura 24: Diagrama de caixa e bigodes da pressão arterial diastólica

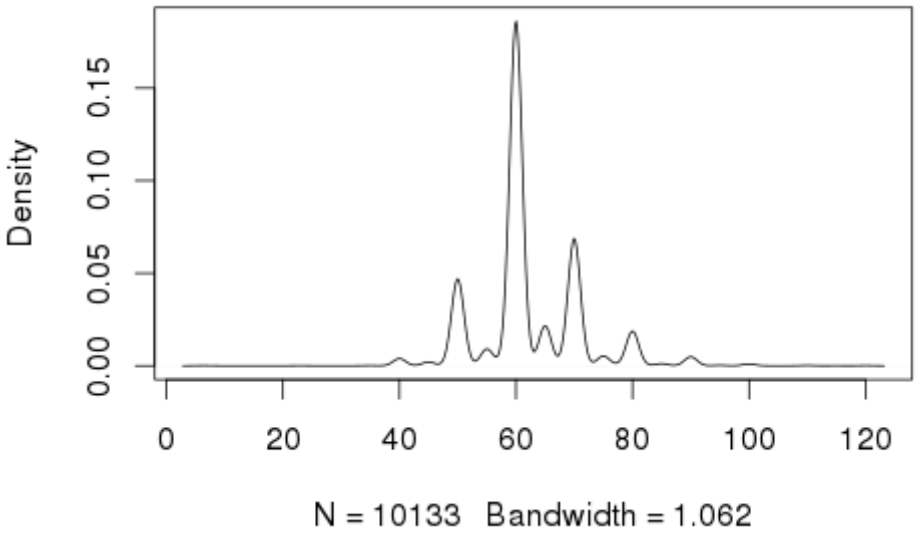

Figura 25: Gráfico de densidades da pressão arterial diastólica

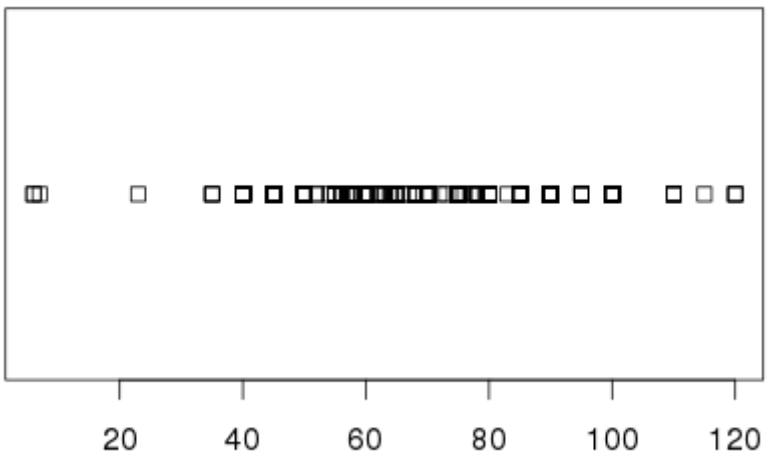

Figura 26: Gráfico de dispersão da pressão arterial diastólica

Nesta variável podemos verificar que existe alguns *outliers* (figuras 24 e 26), é pouco provável que estejam correctos, mas estes não tomam valores exageradamente grandes.

- ***PPA***

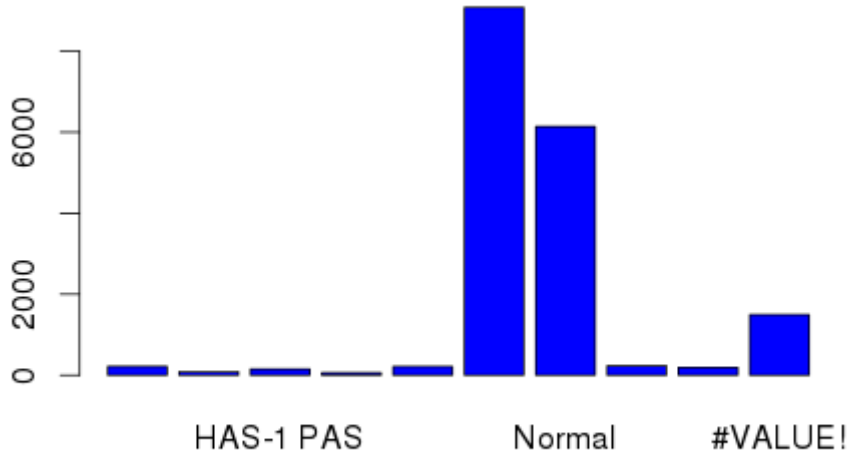

Figura 27: Histograma da amplificação da pressão do pulso

Neste gráfico (figura 27) podemos observar que existe um número substancial de valores do tipo *#VALUE!*, normalmente o *Microsoft Excel* atribuí este tipo quando são inseridos valores diferentes (numéricos, texto, data, entre outros) para uma categoria que foi formatada para um determinado tipo de dados. Com isto, queremos dizer que quem recolheu os dados inseriu valores que nada tinham a ver com esta variável.

- **Normal.x.Anormal**

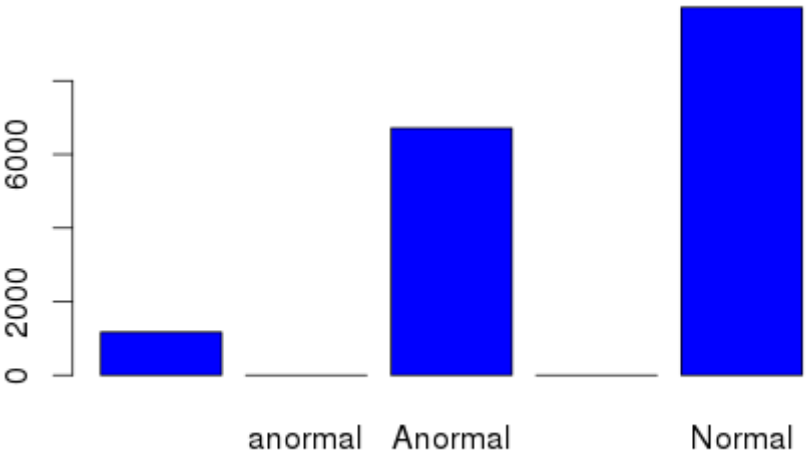

Figura 28: Histograma do Normal.x.Anormal

Através do histograma anterior (figura 28) é possível ver outro problema com os dados de origem, as maiúsculas e minúsculas, pode parecer um pormenor (neste caso é, porque os valores das minúsculas é baixo), mas pode alterar significativamente os resultados induzindo em erro quem analisa os resultados.

- ***B2***

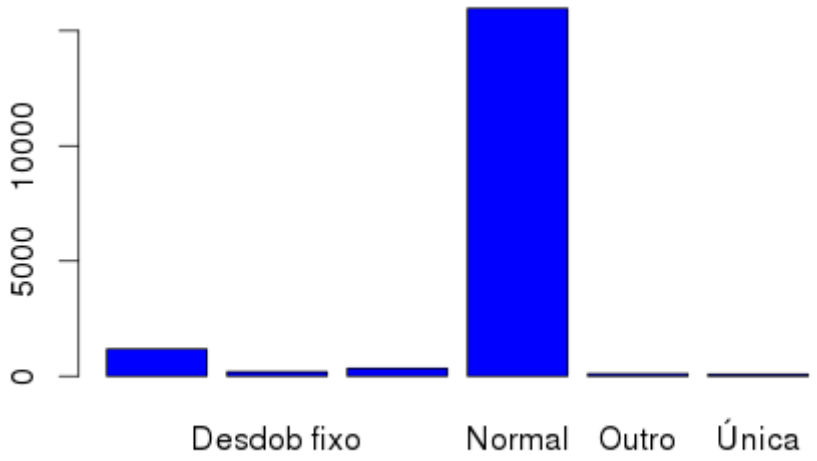

Figura 29: Histograma do segundo batimento

- **Sopro**

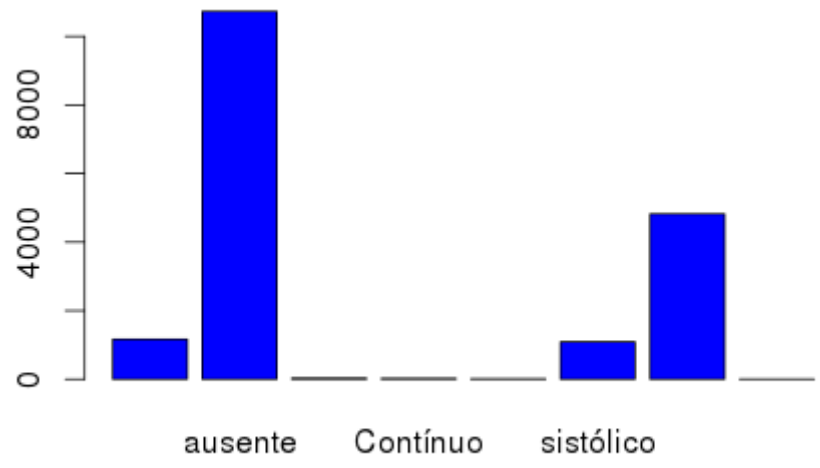

Figura 30: Histograma do sopro

Com esta variável passa-se o mesmo problema do que a "Normal.x.Anormal", o mesmo tipo é escrito de maneira diferente, isto é, as vezes é começado com maiúsculas e outras vezes com minúsculas.

- FC

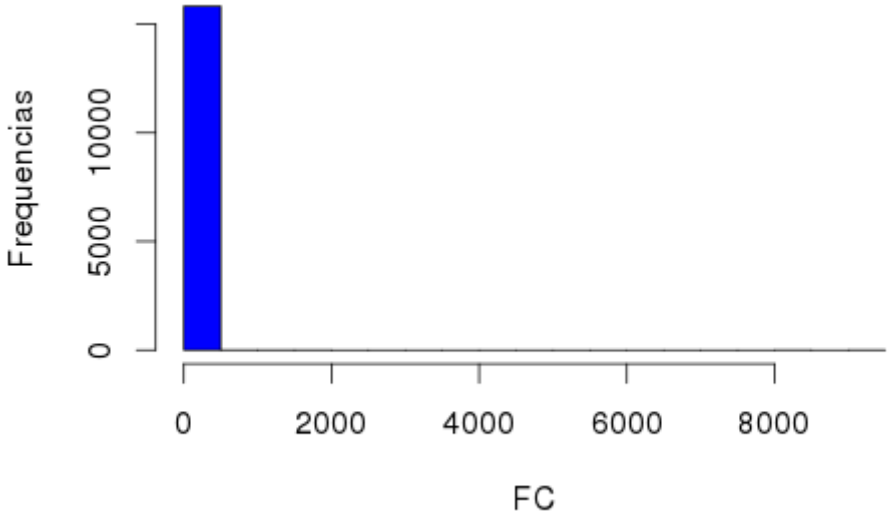

Figura 31: Histograma da frequência cardíaca

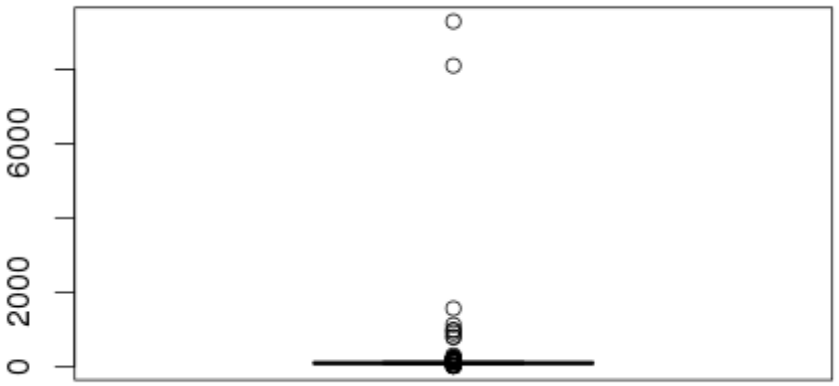

Figura 32: Diagrama de caixa e bigodes da frequência cardíaca

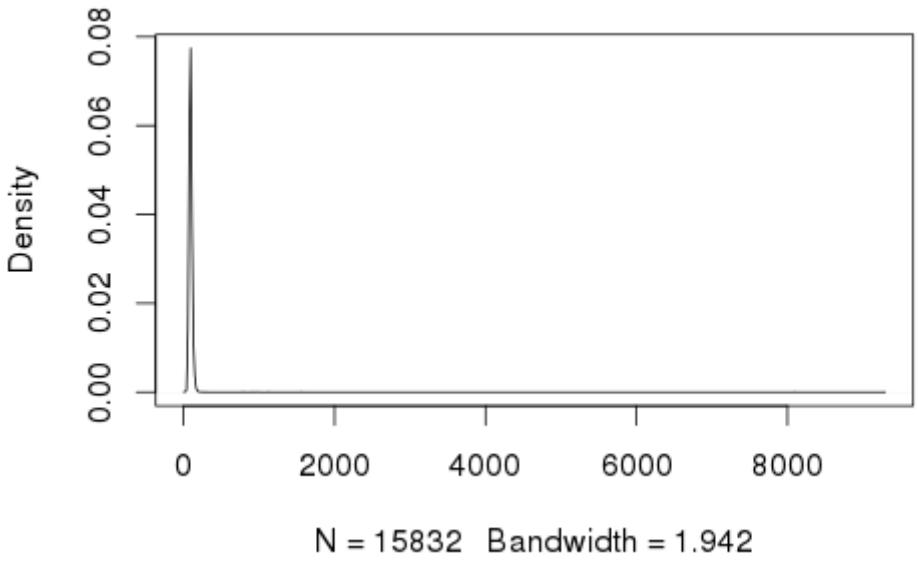

Figura 33: Gráfico de densidades da frequência cardíaca

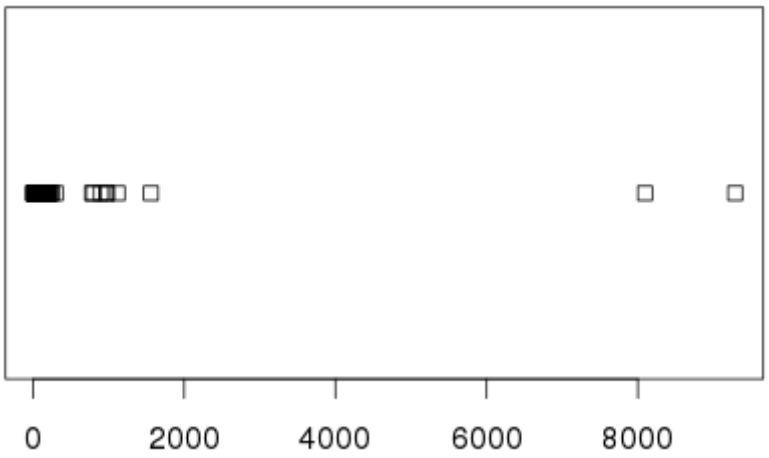

Figura 34: Gráfico de dispersão da frequência cardíaca

Nos dados originais é possível observar (figuras 32 e 34) valores fora do normal em termos de frequência cardíaca, apresentando ainda dois *outliers* com valores que são completamente ridículos.

- ***HDA1***

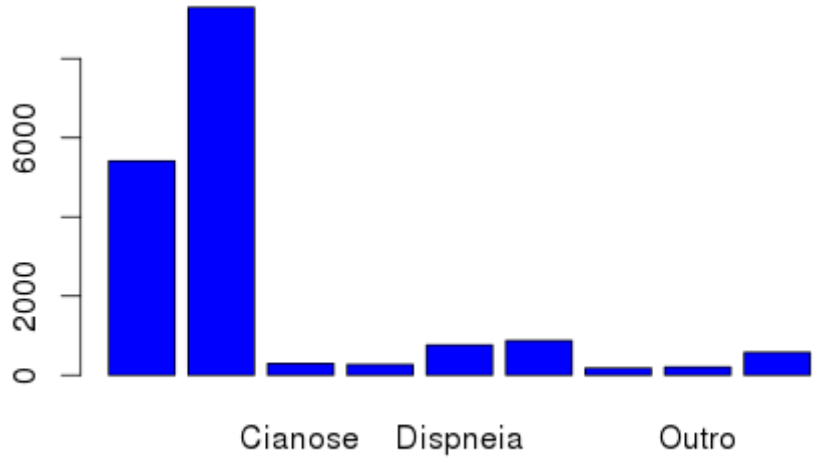

Figura 35: Histograma da primeira parte do histórico de doenças do paciente

- ***HDA2***

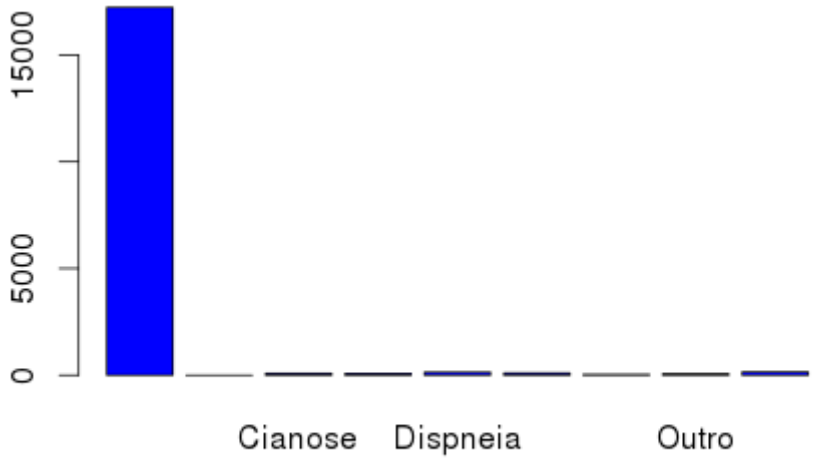

Figura 36: Histograma da segunda parte do histórico de doenças do paciente

No gráfico (figura 36) não é muito percetível os tipos de doenças, porque o tipo de doença com mais ocorrências é o vazio, ou seja, o campo não foi preenchido. Aqui podemos ver mais um exemplo em como o tratamento de dados é fundamental para a análise da informação.

- **Sexo**

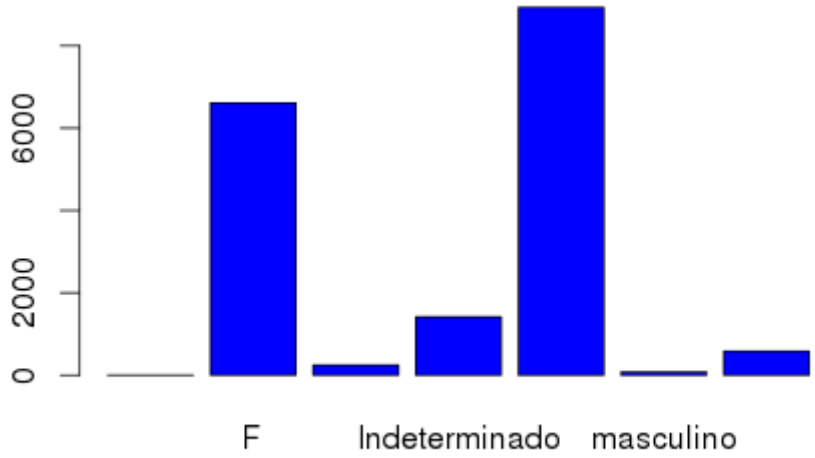

Figura 37: Histograma do sexo do paciente

A variável "sexo"apresenta uma falha anteriormente detectada noutras variáveis, a questão das palavras inicializadas com maiúsculas e minúsculas.

- **Sumário**

```
       ID              Peso           Altura           IMC          Atendimento          DN
Min.   :    1   0      : 2609   Min.   :  0.00   Min.   :  0.00          :  983          : 1376
1st Qu.: 4469   15     :  346   1st Qu.: 36.00   1st Qu.: 15.00   26/05/09:   31   09/05/04:   24
Median : 8937   20     :  341   Median : 99.00   Median : 17.00   09/07/07:   28   21/06/02:   16
Mean   : 8937          :  318   Mean   : 83.87   Mean   : 17.81   18/01/10:   28   19/05/08:   14
3rd Qu.:13405   22     :  315   3rd Qu.:130.00   3rd Qu.: 19.00   07/07/09:   27   10/04/04:   12
Max.   :17873   16     :  310   Max.   :198.00   Max.   :848.00   12/01/10:   27   15/06/06:   12
                (Other):13634                    NA's   :4727     (Other) :16749   (Other) :16419

     IDADE          Convenio                      PULSOS       PA.SISTOLICA    PA.DIASTOLICA
        : 1376          :5304   Normais            :16509   Min.   : 10.0   Min.   :  6.0
0,01    :  195   GS     :2763                      : 1198   1st Qu.: 90.0   1st Qu.: 60.0
0       :  184   UR     :2287   Amplos             :   57   Median :100.0   Median : 60.0
#VALUE!:  120   SB     : 718   Outro              :   45   Mean   :101.3   Mean   : 62.3
0,03    :   82   GRUPO  : 553   Femorais diminuidos:   43   3rd Qu.:110.0   3rd Qu.: 70.0
0,04    :   79   CAMED  : 478   Diminuídos         :   18   Max.   :990.0   Max.   :120.0
(Other):15837   (Other):5770   (Other)            :    3   NA's   :7730    NA's   :7740

                 PPA       NORMAL.X.ANORMAL             B2              SOPRO           FC
Não Calculado      :9081          :1168                  : 1179   ausente  :10727   80     :2469
Normal             :6141   anormal:   1     Desdob fixo  :  190   Sistólico: 4821   100    :2043
#VALUE!            :1496   Anormal:6712     Hiperfonética:  342            : 1167          :2041
Pre-Hipertensão PAD: 233   Normais:   1     Normal       :15969   sistólico: 1090   90     :1733
                   : 217   Normal :9991     Outro        :  107   contínuo :   30   88     :1227
HAS-2 PAS          : 215                    Única        :   86   Contínuo :   23   96     :1084
(Other)            : 490                                          (Other)  :   15   (Other):7276

          HDA.1                    HDA2                  SEXO
Assintomático :9291                  :17221                :   4
              :5414   Palpitacao     :  150   F            :6612
Dor precordial: 873   Dispneia       :  138   Feminino     : 247
Dispneia      : 764   Dor precordial :   98   Indeterminado:1417
Palpitacao    : 576   Cianose        :   86   M            :8930
Cianose       : 298   Desmaio/tontura:   77   masculino    :  79
(Other)       : 657   (Other)        :  103   Masculino    : 584

                                 MOTIVO1                                MOTIVO2
                                    :1097                                  :4778
1 - Cardiopatia já estabelecida:1428   5 - Cirurgia                        :4212
2 - Check-up                    :1048   6 - Sopro                          :2997
5 - Parecer cardiológico        :7981   1 - Cardiopatia congenica          :1262
6 - Suspeita de cardiopatia     :5863   5 - Atividade física               :1137
7 - Outro                       : 456   Outro                              :1097
                                        (Other)                            :2390
```

Figura 38: Sumário dos dados originais

Alterações efectuadas nas variáveis

Nesta secção vamos explicar as alterações que efetuamos a cada uma das variáveis, de modo, a eliminar repetições, valores incorretos e dados fora dos intervalos normais. Foi imposta também uma regra que retira os dados de uma pacientes se este não tiver pelo menos 50% da sua informação preenchida. No caso de existir um campo em branco nos atributos "Idade", "DN"e "Atendimento"todos são preenchidos com "*NA*"(*Non-available*). Retiramos em todas as variáveis os valores em branco e substituímos por "*NA*". Em anexo segue um ficheiro (tratamento.r), que contém o código em linguagem *R*, utilizado para o pré-processamento dos dados.

ID: O identificador não foi alterado, pois trata-se de um número contínuo atribuído ao paciente e este não deve ser modificado.

Peso: Nesta variável apenas retiramos os valores que eram negativos ou iguais a zero, achamos que os restantes valores eram adequados.

Altura: Na "altura"retiramos todos os valores abaixo dos 24 cm, pois esse é o tamanho do recém-nascido mais pequeno de sempre.

***IMC*:** A variável foi limitada ao intervalo entre 10 e 50 [3]. Para os valores do *"IMC"* que saíam fora deste intervalo, eliminamos o peso e a altura correspondente pois uma ou ambas devem estar erradas. Posteriormente, ainda verificamos se os restantes valores do *ICM* estavam correctos, para isso, recorremos a seguinte expressão:

$$subset(data, !IMC == round(as.numeric(as.character(Peso))/((as.numeric(Altura)/100)^2)))$$

Atendimento: Aqui existem diversos valores inteiros que correspondem à datas no formato numérico, ou seja, o número de dias entre a data de origem (1899-12-30) do *Excel* e a data inserida. Estes valores foram convertidos com base na data de origem e no número apresentado.

***DN*:** Nesta variável foi aplicado o mesmo tratamento da variável anterior. Nesta variável existiam casos como: 8anos/e/3/m ou 29//11/00. Para cada paciente (cerca 120 casos) foram analisados os valores. No primeiro caso era impossível determinar a data de nascimento, já no segundo caso a data é percetível.

Idade: Na idade foi retirado todos os valores menores ou iguais a zero, eliminando a *"DN"* e o atendimento correspondente. Apagamos também alguns "#VALUE!", mas outros foram recalculados. Também foi eliminado todos os registos em que a idade do paciente era maior que 19 anos. Ao apagar o campo idade, a "DN"e o "Atendimento"eram inúteis, logo, ambas foram eliminadas ("*NA*").

Convénio: Esta variável não sofreu quaisquer alterações, pois esta não é relevante para a análise realizada.

Pulsos: No pulso removemos apenas os dados "AMPLOS"e "NORMAIS", e substituímos por "Amplos"e "Normais"respectivamente.

***PA* Sistólica:** Nesta variável existiam valores muito acima do normal (60 a 190) [4], e como tal foram removidos esses dados.

***PA* Diastólica:** Na pressão diastólica verificavam-se também valores fora do habitual (entre 40 e 100) [4], estes foram eliminados pois induziam em erro.

***PPA*:** Neste caso, decidimos agrupar em "*NA*", os dados do tipo "Não Calculado", "#VALUE!"e campos vazios.

Normal X Anormal: Nesta variável substituímos os dados com valor "anormal"por "Anormal"e "NORMAIS"por "Normal".

B2: Apenas retiramos os campos em branco, de resto, tudo pareceu-nos estar dentro do normal.

Sopro: Aqui efectuamos apenas alterações na escrita dos dados, de maneira a que valores iguais mas escritos de maneira diferente coincidissem. Alteramos os dados "ausente"e "diastólico"para que a primeira letra fosse maiúscula. Agrupamos ainda os valores "contínuo"e "Contínuo", assim como, "sistólico"e "Sistólico"para que apenas fosse visualizado "Contínuo"e "Sistólico".

***FC*:** Nesta variável existiam dados que estavam num formato de intervalo (por exemplo: 200-300), havendo ainda um valor que estava separado com um "a". Para estes casos, decidimos fazer uma média

dos intervalos encontrados. Posto isto, retiramos todos os valores que estivessem fora do intervalo (40 até 160) normal da frequência cardíaca [5].

HDA 1: Nesta variável apenas foram removidos os valores em branco e substituídos por "*NA*".

HDA 2: Nesta variável apenas foram removidos os valores em branco e substituídos por "*NA*".

Sexo: Aqui mudamos o tipo "Indeterminado"para "*NA*"e alteramos alguns valores "Masculino"que existiam para "M".

Motivo 1: Nesta variável apenas foram removidos os valores em branco e substituídos por "*NA*".

Motivo 2: Nesta variável apenas foram removidos os valores em branco e substituídos por "*NA*".

Dados pré-processados

O resultado do pré-processamento dos dados encontra-se num ficheiro (data.cvs) em anexo à este relatório. Após este tratamento, verificamos que foram eliminadas 1.411 linhas, ou seja, 1.411 pacientes dos 17.873 iniciais, estes foram descartados por erros na recolha dos dados. Outra curiosidade, é o facto de antes dos dados terem sido organizados, existiam 87.276 campos em branco (sem qualquer tipo de dados), após a realização desta tarefa observamos que passaram a existir 91.124 (incluindo campos em branco e valores com erros nos dados originais) campos com "NA".

Descrição

Nesta secção vamos poder ver, com o auxílio de gráficos, o resultado do pré-processamento dos dados. Será possível verificar que a presença de *outliers* foi minimizada e que em geral os dados estão mais homogéneos.

- **Peso**

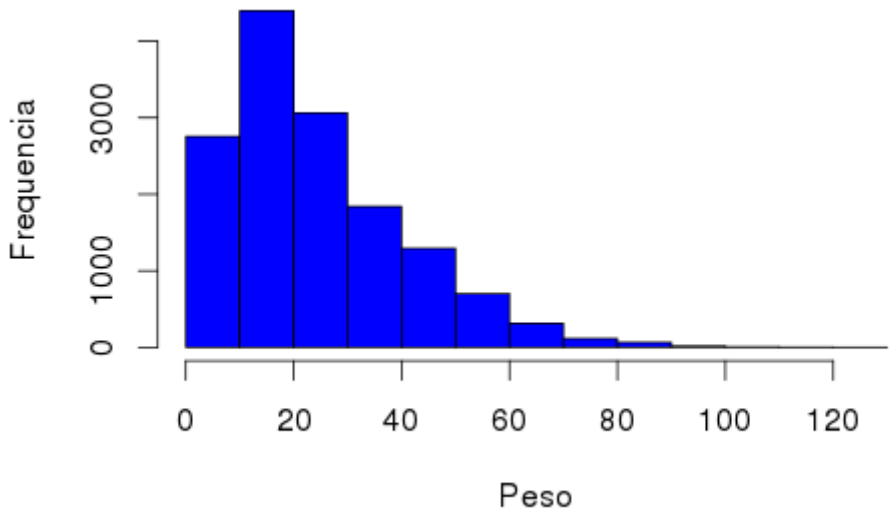

Figura 39: Histograma do peso

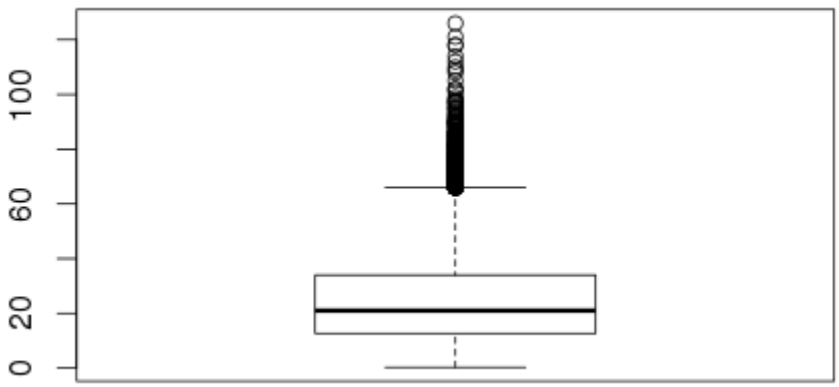

Figura 40: Diagrama de caixa e bigodes do peso

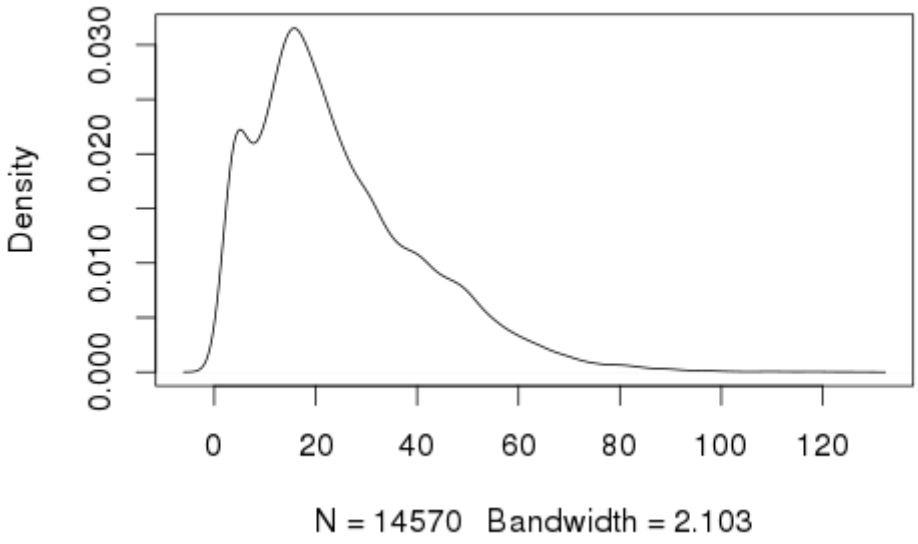

Figura 41: Gráfico de densidades do peso

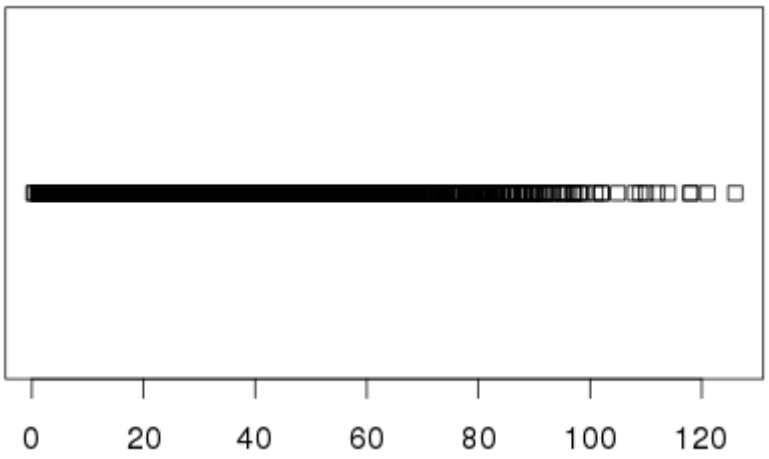

Figura 42: Gráfico de dispersão do peso

A existência de *outliers* no "Peso"terá a ver com as crianças que tem excesso de peso.

- **Altura**

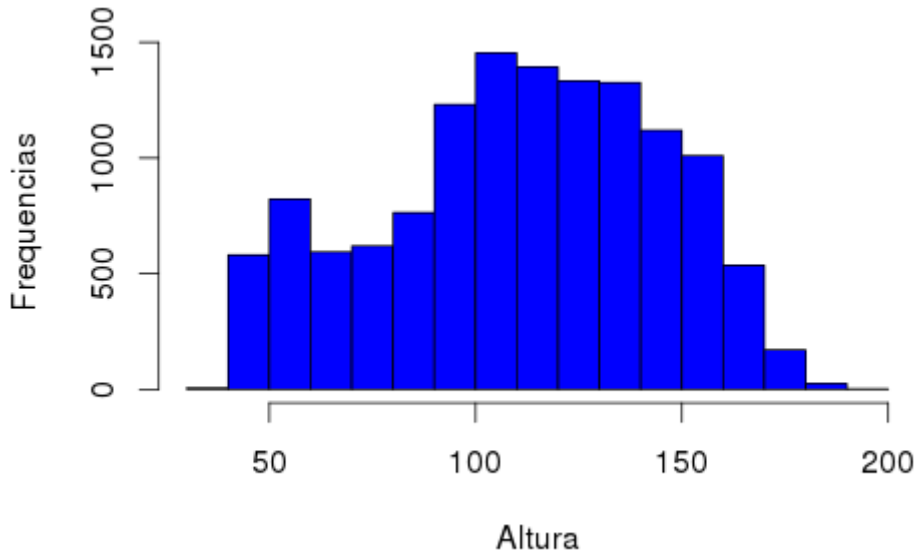

Figura 43: Histograma da altura

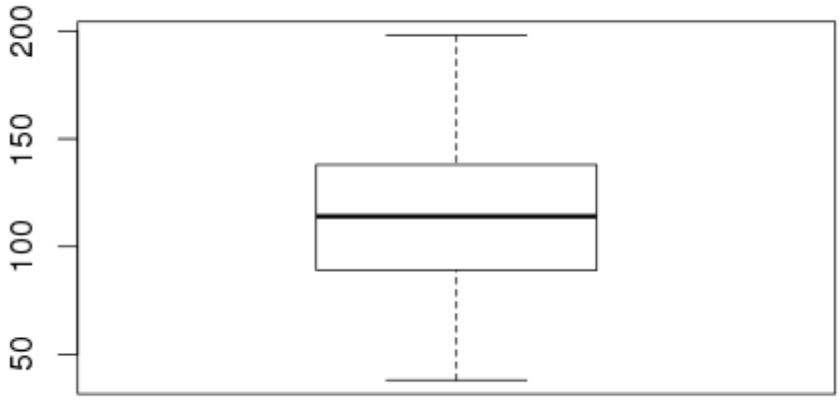

Figura 44: Diagrama de caixa e bigodes da altura

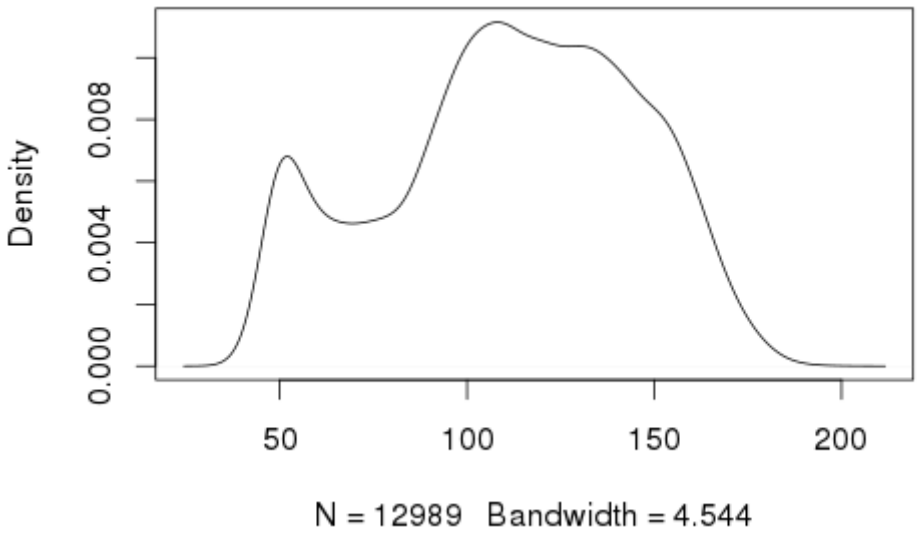

Figura 45: Gráfico de densidades da altura

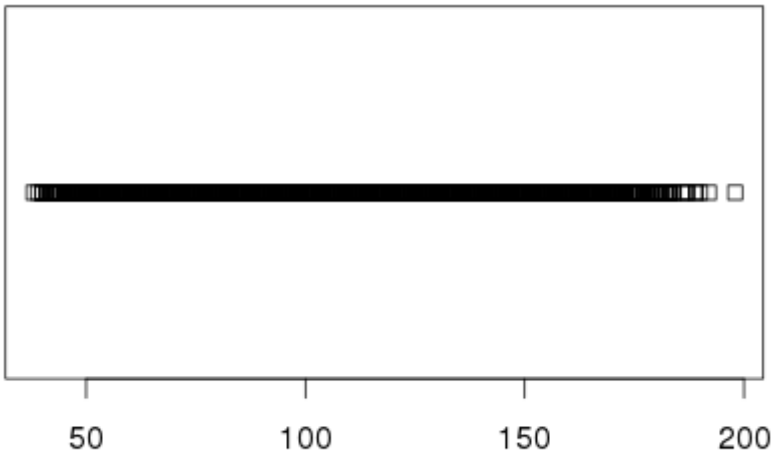

Figura 46: Gráfico de dispersão da altura

- ***IMC***

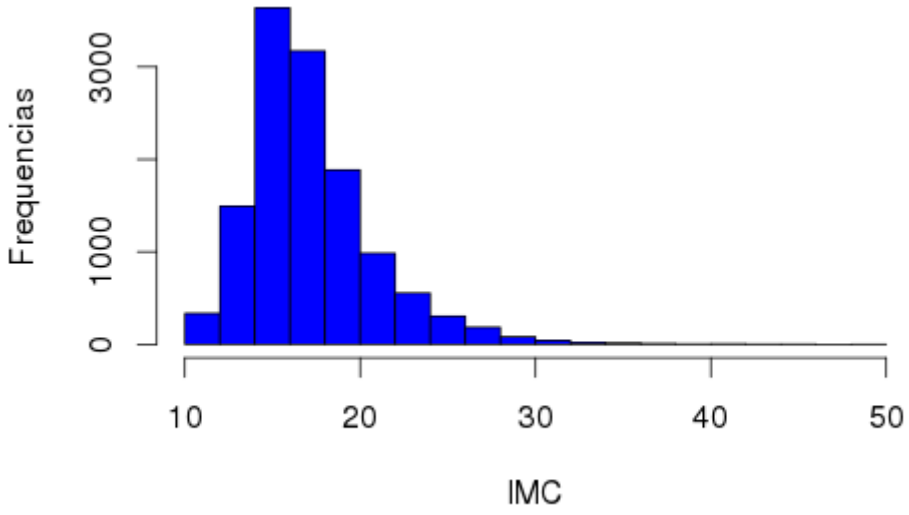

Figura 47: Histograma do *IMC*

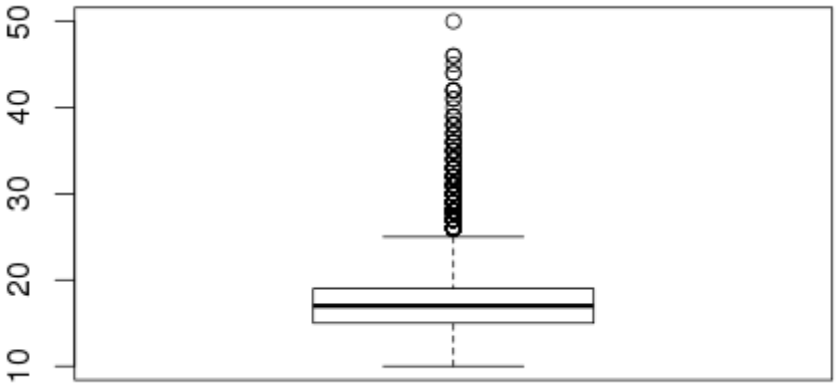

Figura 48: Diagrama de caixa e bigodes do *IMC*

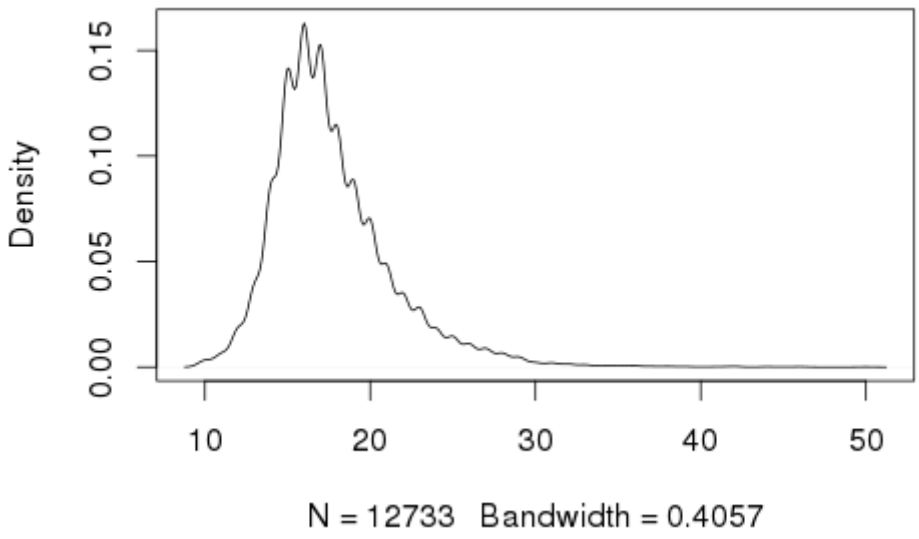

Figura 49: Gráfico de densidades do *IMC*

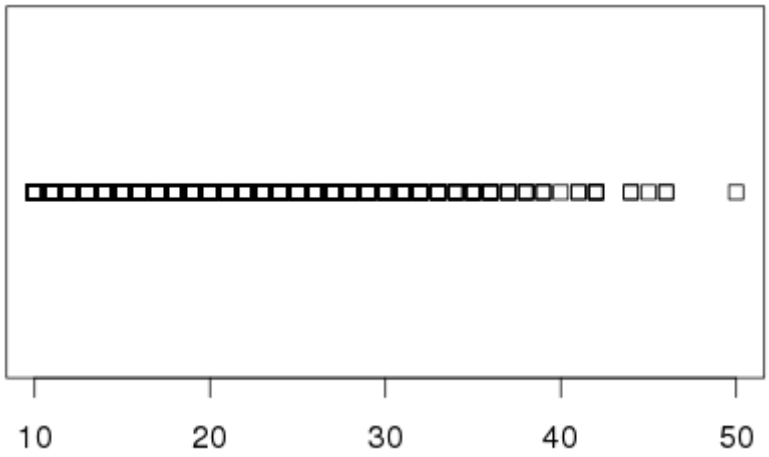

Figura 50: Gráfico de dispersão do *IMC*

A existência de *outliers* está relacionada com o facto de as crianças terem excesso de peso.

- **Idade**

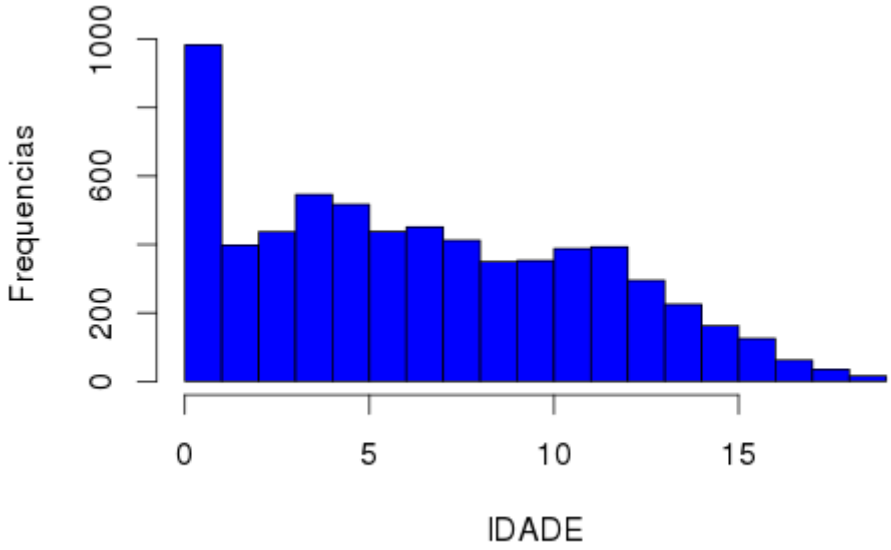

Figura 51: Histograma da idade

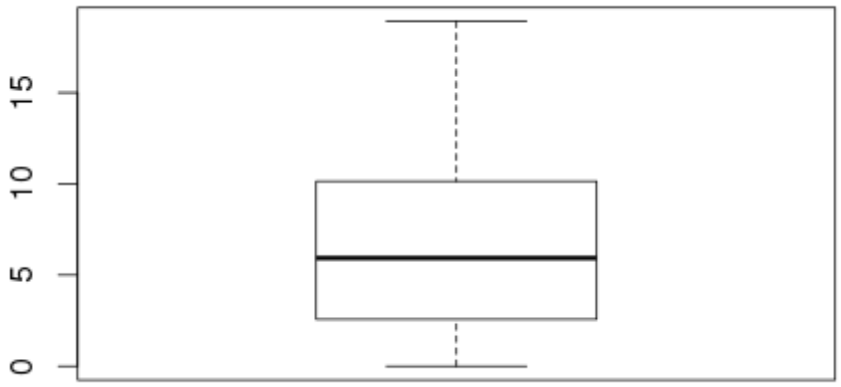

Figura 52: Diagrama de caixa e bigodes da idade

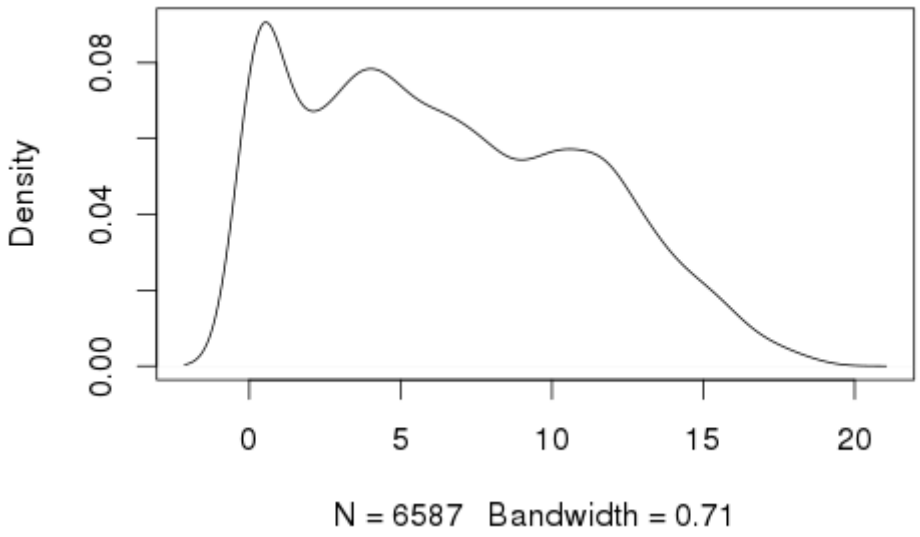

Figura 53: Gráfico de densidades da idade

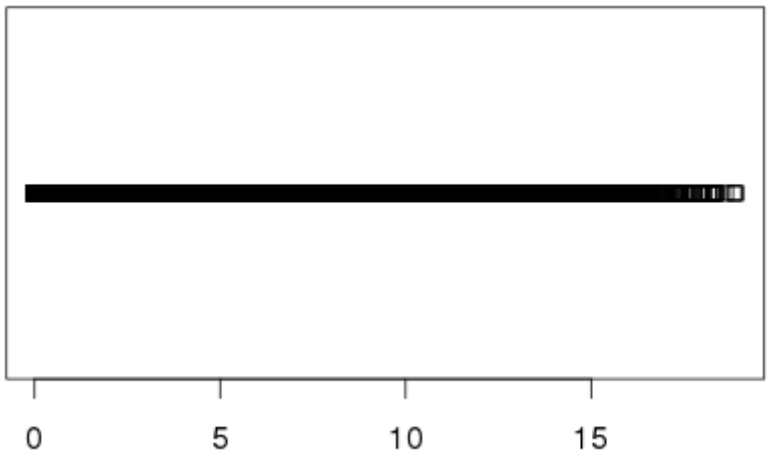

Figura 54: Gráfico de dispersão da idade

- **Pulsos**

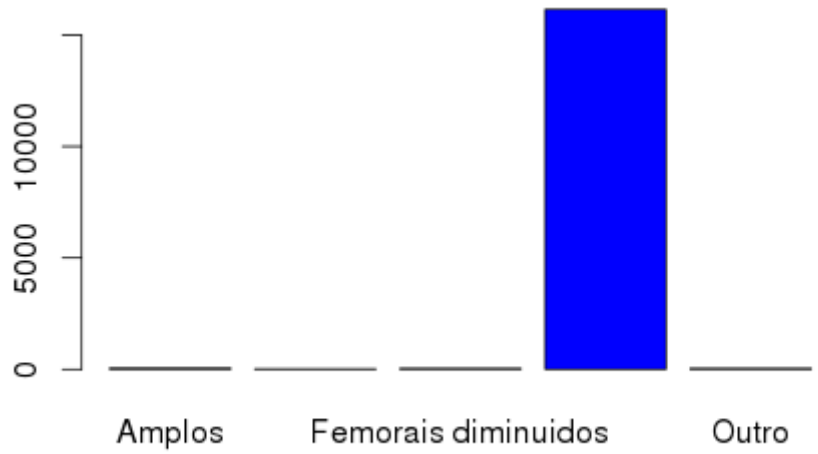

Figura 55: Histograma dos pulsos

- ***PA* Sistólica**

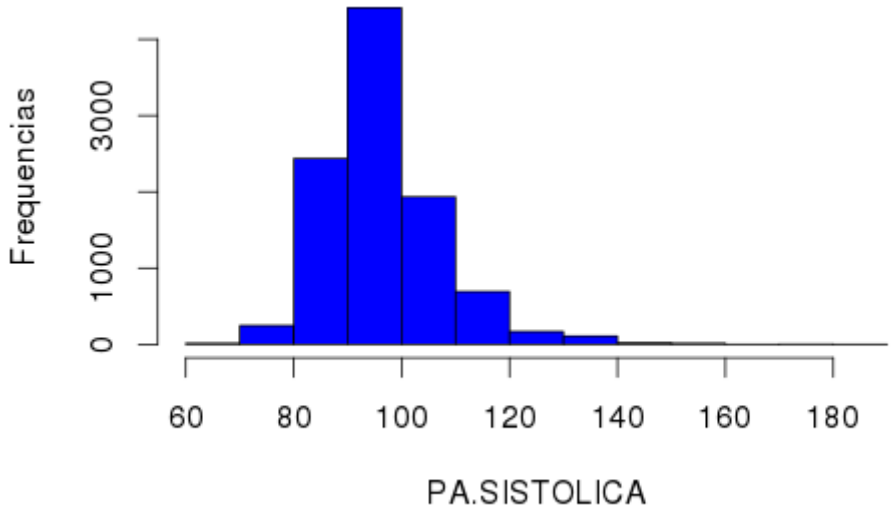

Figura 56: Histograma da pressão arterial sistólica

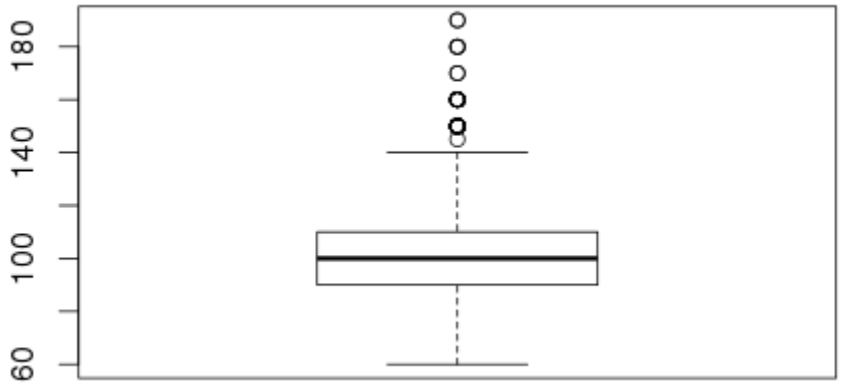

Figura 57: Diagrama de caixa e bigodes da pressão arterial sistólica

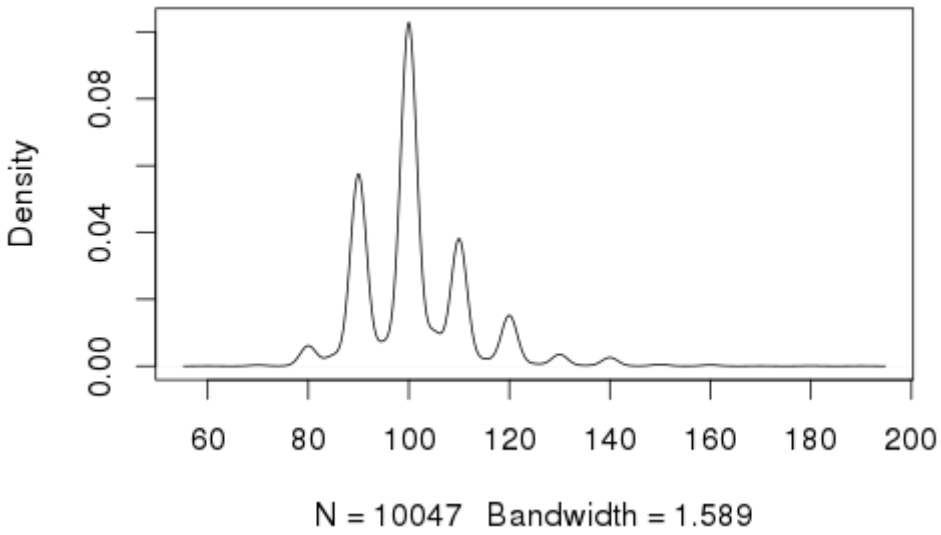

Figura 58: Gráfico de densidades da pressão arterial sistólica

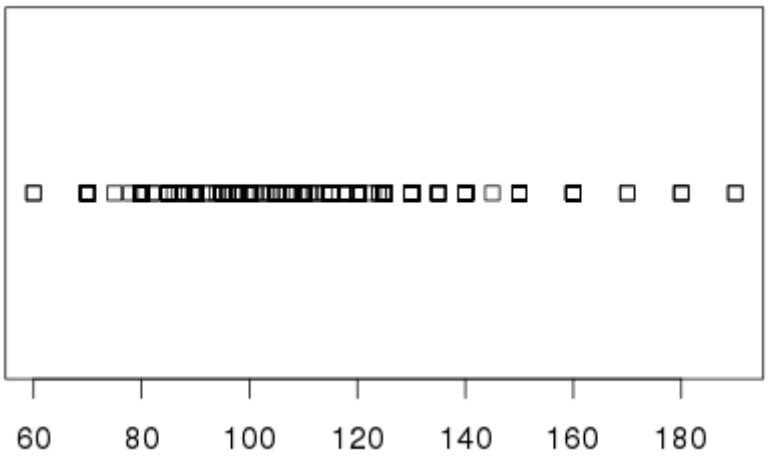

Figura 59: Gráfico de dispersão da pressão arterial sistólica

- ***PA* Diastólica**

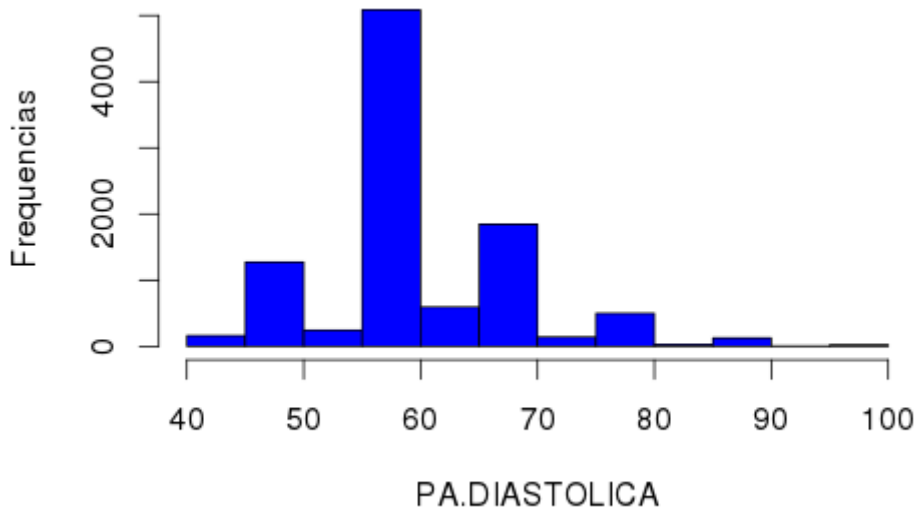

Figura 60: Histograma da pressão arterial diastólica

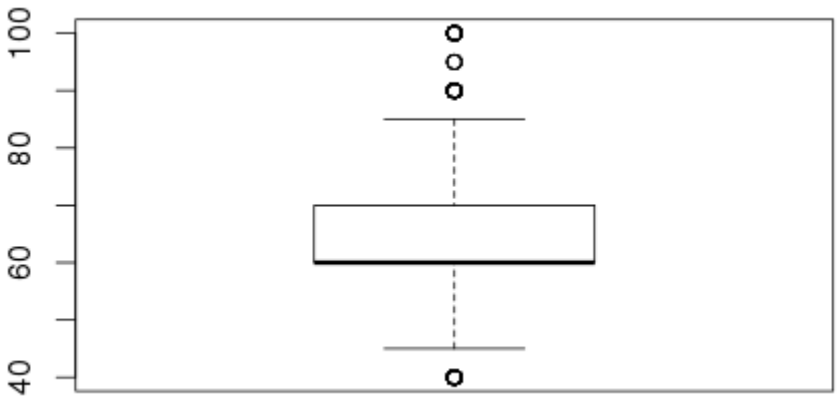

Figura 61: Diagrama de caixa e bigodes da pressão arterial diastólica

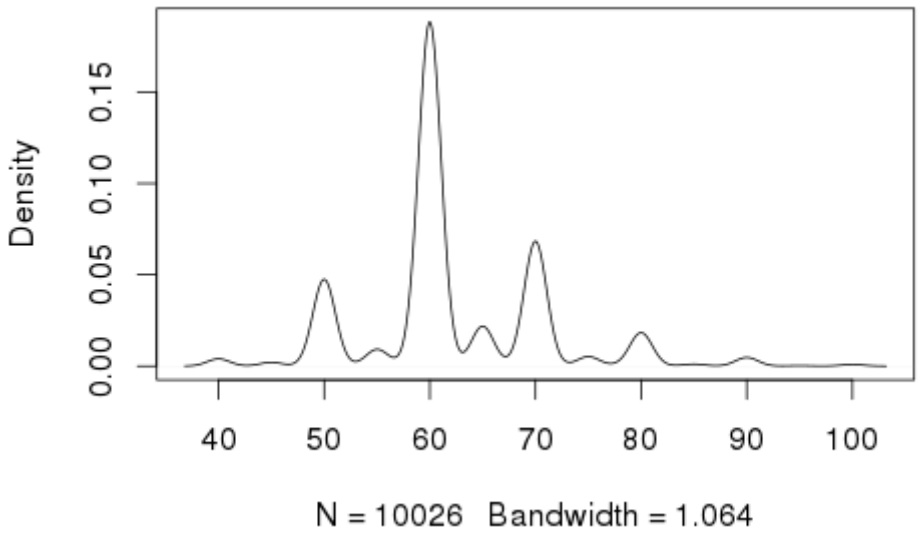

Figura 62: Gráfico de densidades da pressão arterial diastólica

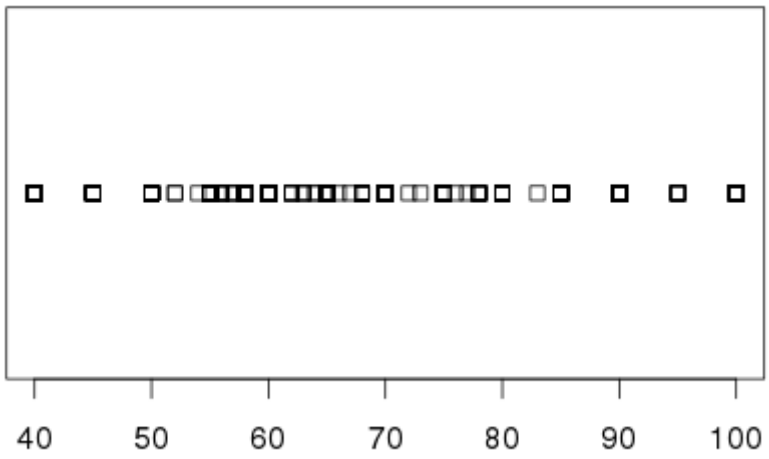

Figura 63: Gráfico de dispersão da pressão arterial diastólica

- ***PPA***

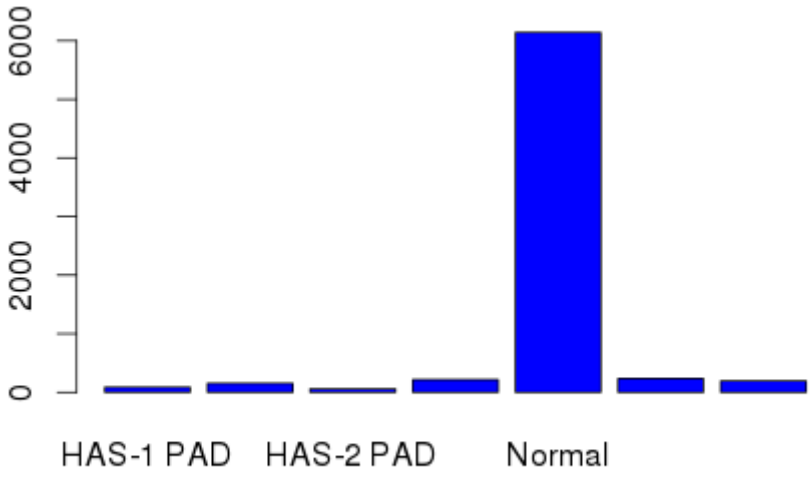

Figura 64: Histograma da amplificação da pressão do pulso

- **Normal.x.Anormal**

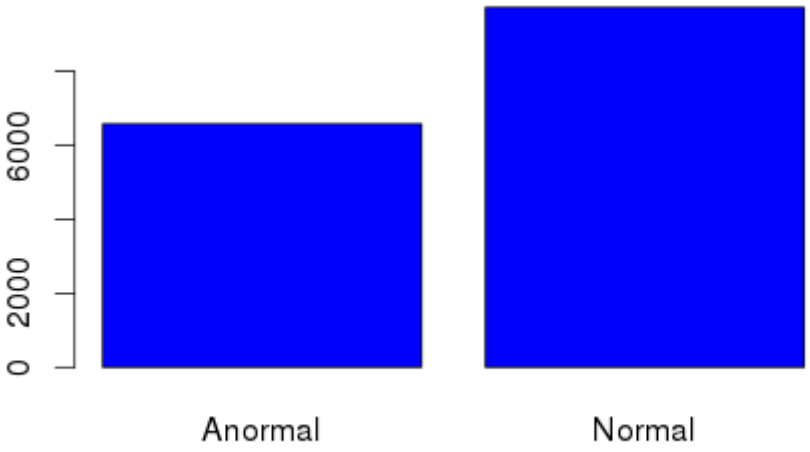

Figura 65: Histograma do Normal.x.Anormal

- ***B2***

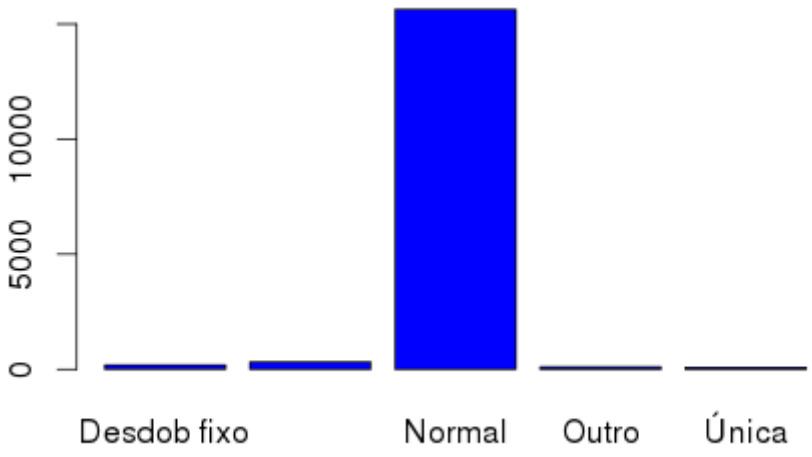

Figura 66: Histograma do segundo batimento

- **Sopro**

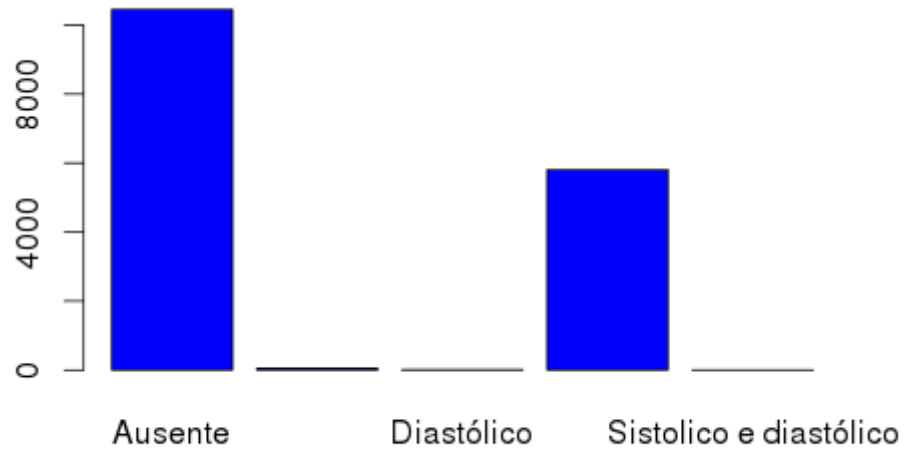

Figura 67: Histograma do sopro

- ***FC***

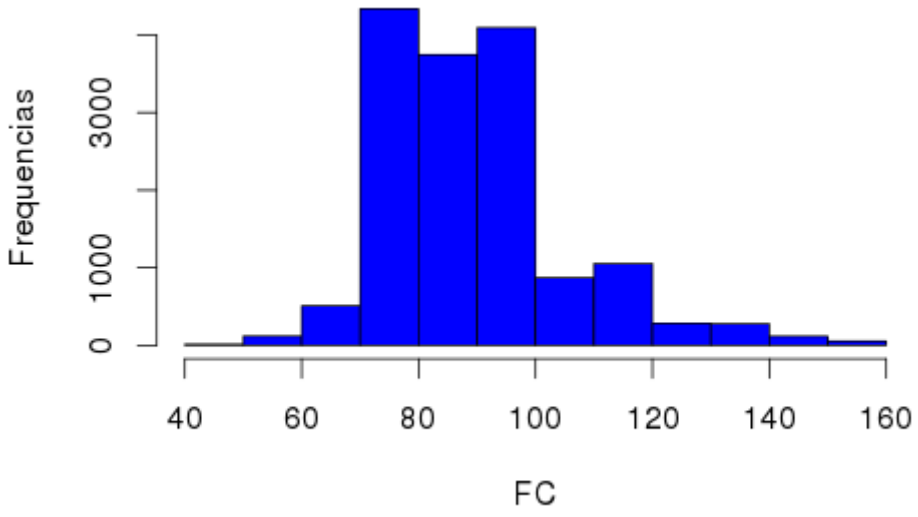

Figura 68: Histograma da frequência cardíaca

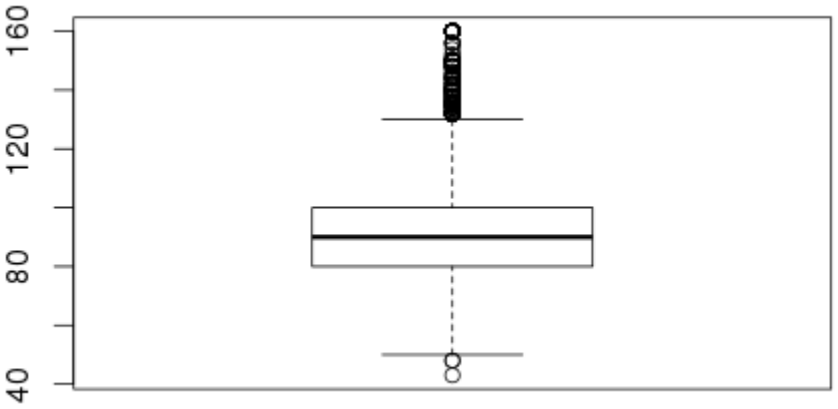

Figura 69: Diagrama de caixa e bigodes da frequência cardíaca

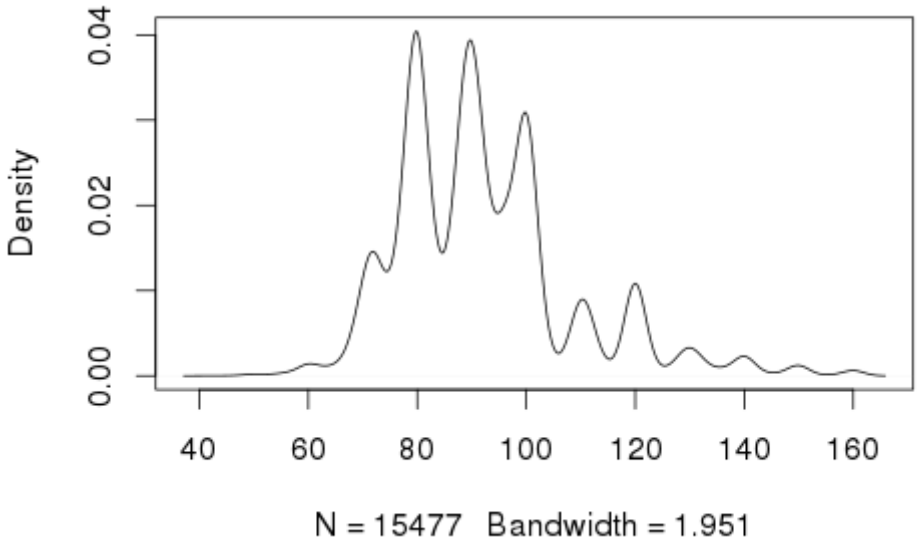

Figura 70: Gráfico de densidades da frequência cardíaca

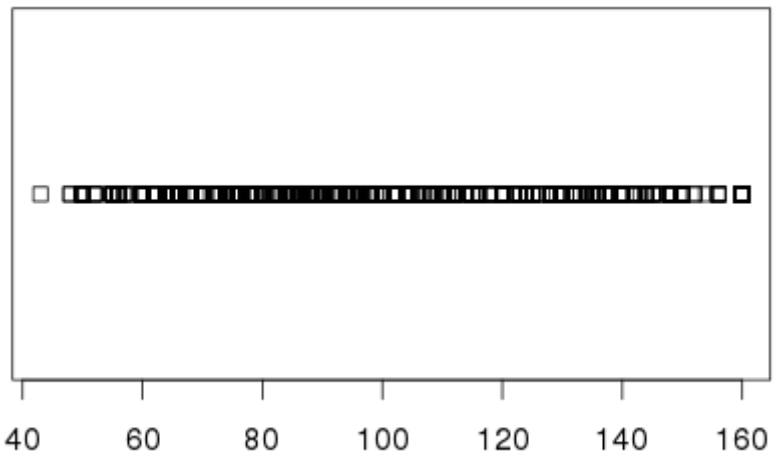

Figura 71: Gráfico de dispersão da frequência cardíaca

- ***HDA*1**

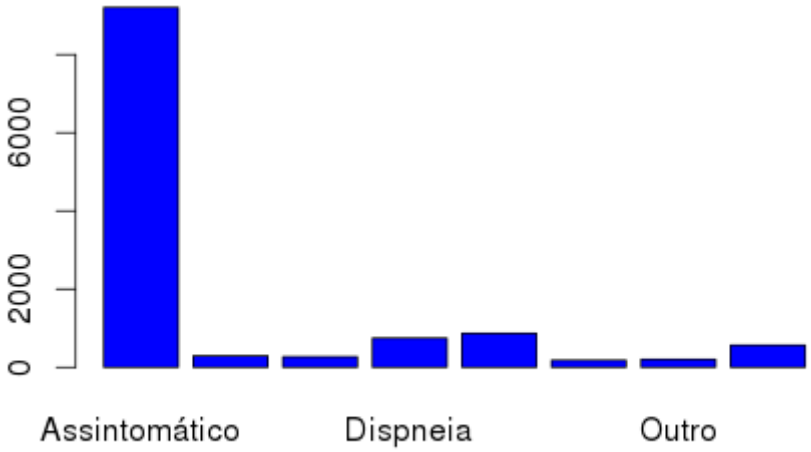

Figura 72: Histograma da primeira parte do histórico de doenças do paciente

- ***HDA2***

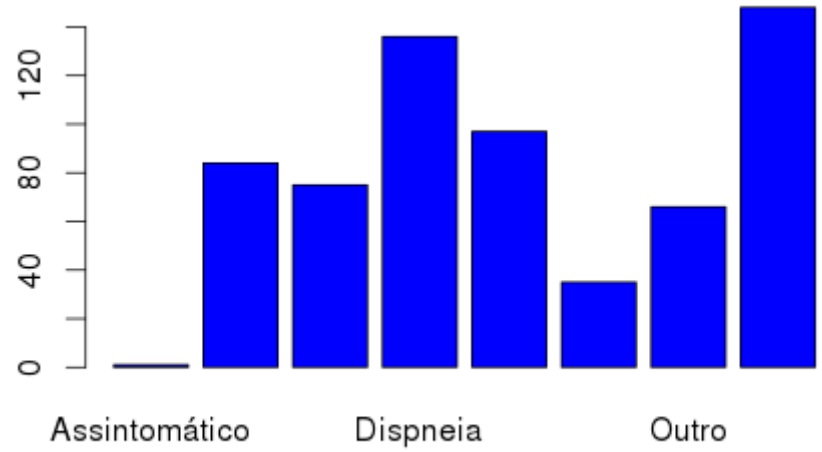

Figura 73: Histograma da segunda parte do histórico de doenças do paciente

- **Sexo**

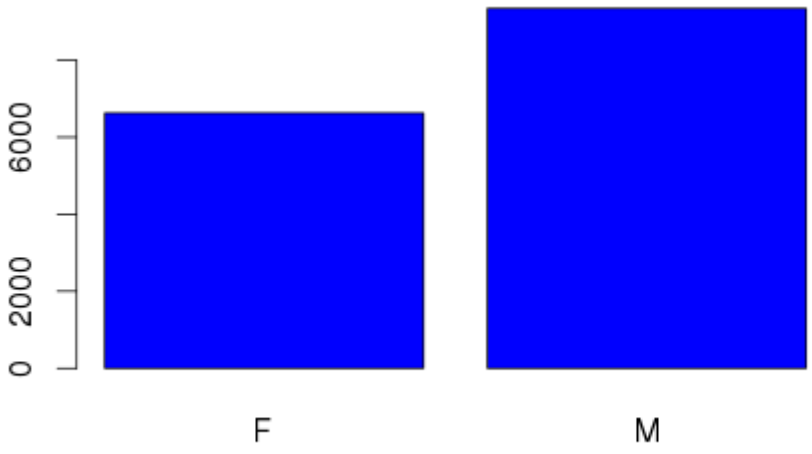

Figura 74: Histograma do sexo do paciente

- **Sumário**

```
       X                ID              Peso           Altura          IMC          Atendimento
 Min.   :    1   Min.   :    1   15      :  335   Min.   : 38.0   Min.   :10.00   09/02/10:  14
 1st Qu.: 4382   1st Qu.: 4382   20      :  333   1st Qu.: 89.0   1st Qu.:15.00   26/05/09:  14
 Median : 8842   Median : 8842   22      :  308   Median :114.0   Median :17.00   04/08/09:  13
 Mean   : 8879   Mean   : 8879   16      :  304   Mean   :111.5   Mean   :17.71   15/05/09:  13
 3rd Qu.:13342   3rd Qu.:13342   18      :  299   3rd Qu.:138.0   3rd Qu.:19.00   16/02/09:  13
 Max.   :17873   Max.   :17873   (Other):12991   Max.   :198.0   Max.   :50.00   (Other) :6520
                                 NA's    : 1892   NA's   :3473    NA's   :3729    NA's    :9875

       DN            IDADE          Convenio                      PULSOS       PA.SISTOLICA
 09/05/04:  12   0       :  41   GS      :2683   Amplos               :   54   Min.   : 60.0
 24/02/03:   9   0,01    :  33   UR      :2221   Diminuídos           :   18   1st Qu.: 90.0
 10/04/04:   7   0,13    :  29   SB      : 703   Femorais diminuidos:   38   Median :100.0
 25/02/03:   7   0,08    :  28   GRUPO   : 551   Normais              :16146   Mean   :101.1
 04/03/99:   6   0,09    :  26   CAMED   : 477   Outro                :   39   3rd Qu.:110.0
 (Other) :6546   (Other):6430   (Other):5566   NA's                 :  167   Max.   :190.0
 NA's    :9875   NA's   :9875   NA's   :4261                                  NA's   :6415

 PA.DIASTOLICA                  PPA          NORMAL.X.ANORMAL            B2
 Min.   : 40.0   Normal              :6141   Anormal:6578     Desdob fixo  :  185
 1st Qu.: 60.0   Pre-Hipertensão PAD: 233   Normal :9727     Hiperfonética:  315
 Median : 60.0   HAS-2 PAS           : 215   NA's   : 157     Normal       :15634
 Mean   : 62.2   Pre-Hipertensão PAS: 193                    Outro        :  100
 3rd Qu.: 70.0   HAS-1 PAS           : 153                    Única        :   79
 Max.   :100.0   (Other)             : 144                    NA's         :  149
 NA's   :6436    NA's                :9383

          SOPRO             FC                    HDA.1                    HDA2
 Ausente              :10451   Min.   : 43.00   Assintomático :9204   Palpitacao     :  148
 Contínuo             :   49   1st Qu.: 80.00   Dor precordial: 869   Dispneia       :  136
 Diastólico           :   11   Median : 90.00   Dispneia      : 746   Dor precordial :   97
 Sistólico            : 5807   Mean   : 92.15   Palpitacao    : 567   Cianose        :   84
 Sistolico e diastólico:    3   3rd Qu.:100.00   Cianose       : 295   Desmaio/tontura:   75
 NA's                 :  141   Max.   :160.00   (Other)       : 653   (Other)        :  102
                               NA's   :985      NA's          :4128   NA's           :15820

  SEXO                                  MOTIVO1                                MOTIVO2
 F   :6628   1 - Cardiopatia já estabelecida:1353   5 - Cirurgia                  :4128
 M   :9344   2 - Check-up                   :1019   6 - Sopro                     :2927
 NA's: 490   5 - Parecer cardiológico       :7803   1 - Cardiopatia congenica     :1193
             6 - Suspeita de cardiopatia    :5711   5 - Atividade física          :1097
             7 - Outro                      : 437   Outro                         :1061
             NA's                           : 139   (Other)                       :2322
                                                    NA's                          :3734
```

Figura 75: Sumário dos dados pré-processados

Análise dos Dados

Análise Bivariada

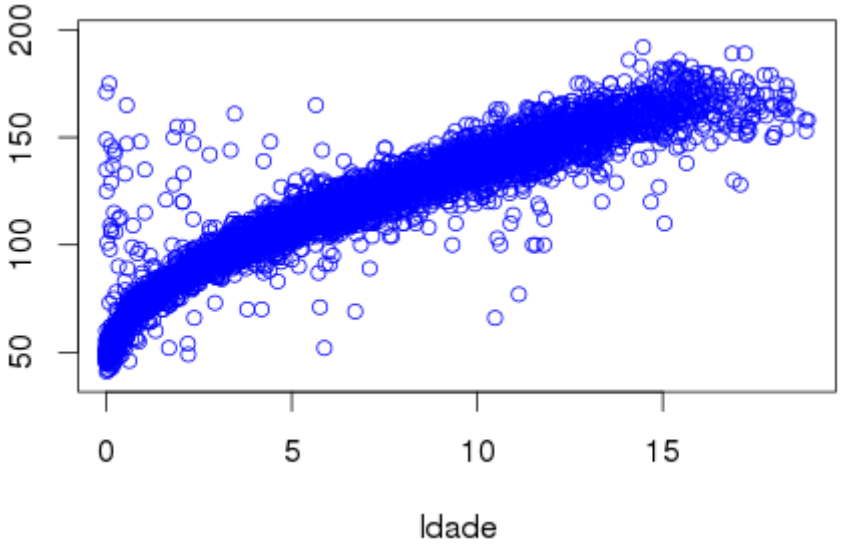

Figura 76: Gráfico da altura com a idade

Existe algumas discrepâncias que tem que ver com o facto de considerarmos idades iguais a zero, ou seja, os recém-nascidos. Poderá ser um problema pois muito provavelmente alguns campos em brancos foram considerados como zero.

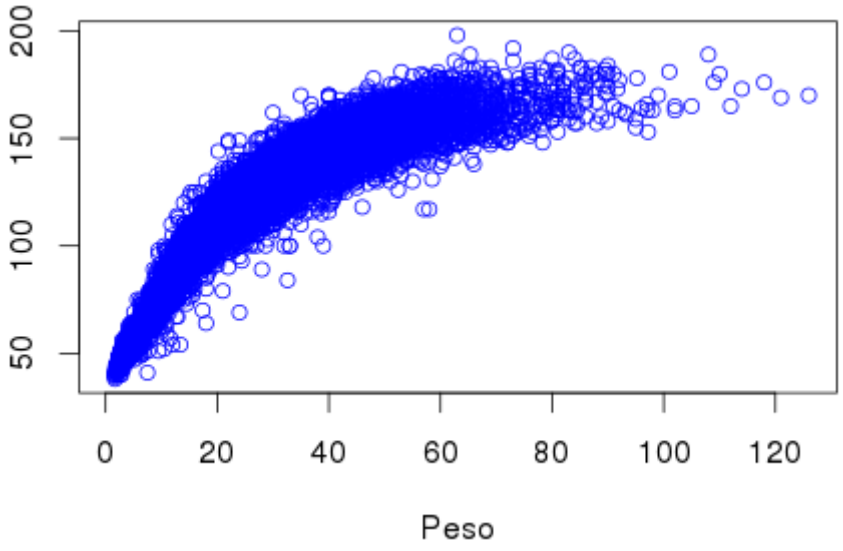

Figura 77: Gráfico da altura com o peso

Este gráfico foi adicionado para que seja possível visualizar o aumento do peso em relação à idade, verificamos que o peso aumenta principalmente a partir dos 130/150 cm, após esses valores o peso aumenta com alguma rapidez.

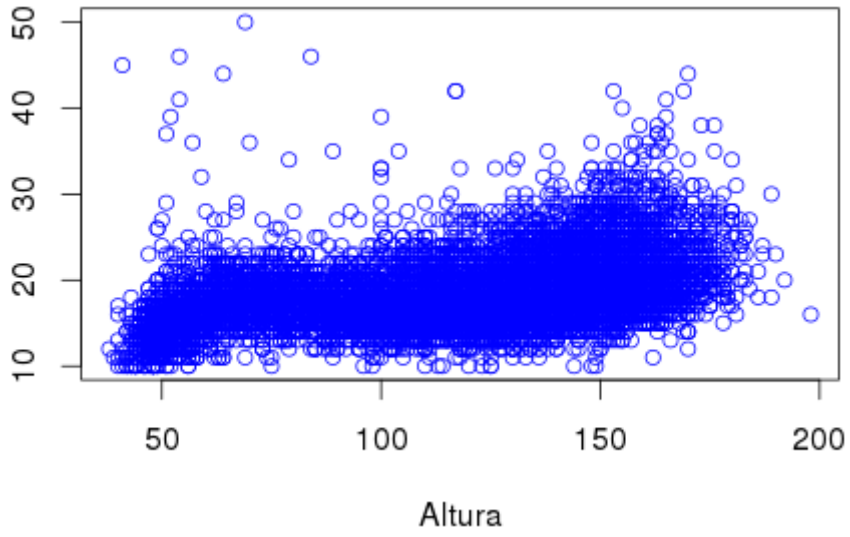

Figura 78: Gráfico da altura com o índice de massa corporal

Podemos verificar no gráfico (figura 78) anterior que o *IMC* cresce ligeiramente com o aumentar da altura. A fórmula de cálculo do *IMC* divide o peso pelo quadrado da altura, logo, se o peso e a altura aumentarem nas mesmas proporções o *IMC* vai manter-se praticamente constante. Como vimos no gráfico do peso com a altura, o peso cresce mais nas alturas superiores a 120/130cm, normalmente onde as pessoas começam a ganhar mais peso.

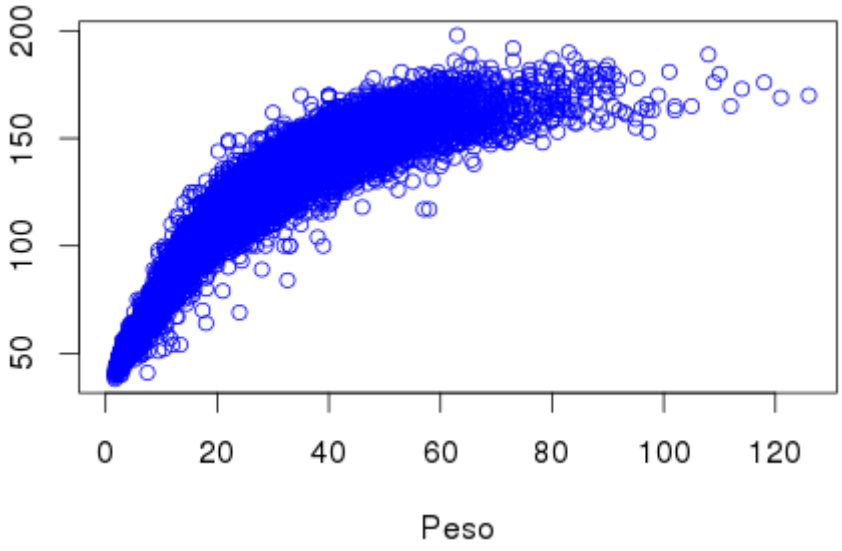

Figura 79: Gráfico do índice de massa corporal com o peso

Se compararmos os gráficos da altura com o peso (figura 77) e o gráfico anterior (figura 79), verificamos que ambos são muito semelhantes. Isto deve-se ao facto das três variáveis estarem relacionadas entre si na fórmula de cálculo do *IMC*.

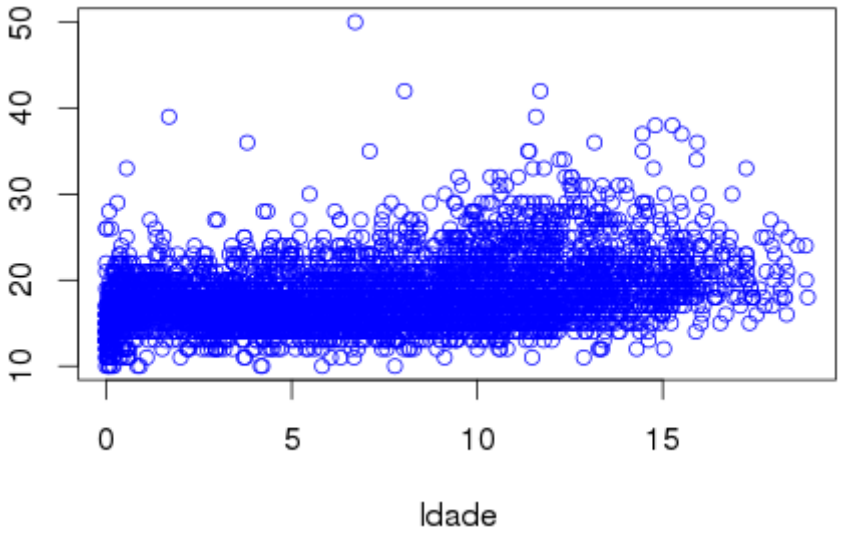

Figura 80: Gráfico do índice de massa corporal com a idade

Neste gráfico verificamos que o *IMC* cresce muito ligeiramente, após uma pesquisa constatamos que este deveria aumentar muito mais (de forma quase exponencial) [6], o que nos surpreendeu porque os restantes gráficos analisados estão dentro do normal.

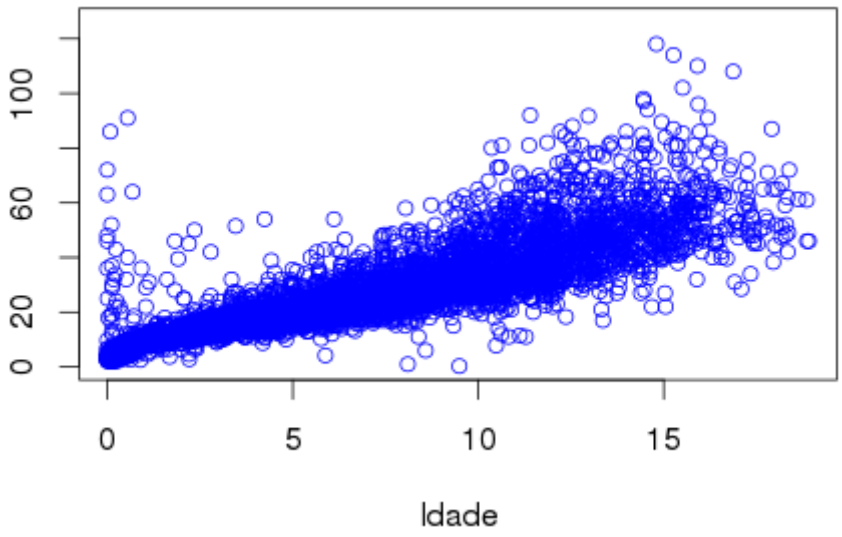

Figura 81: Gráfico do peso com a idade

Existe algumas discrepâncias que tem que ver com o facto de consideramos idades iguais a zero, ou seja, os recém-nascidos. Poderá ser um problema pois muito provavelmente alguns campos em brancos foram considerados como zero.

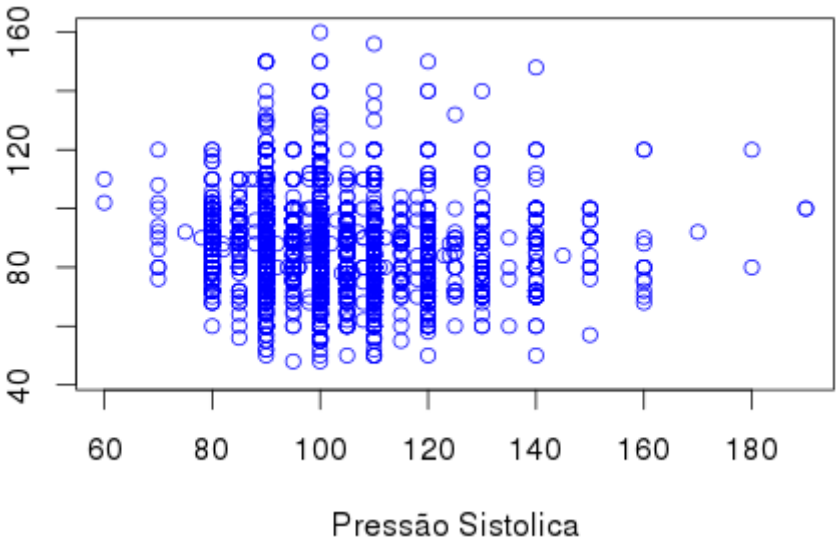

Figura 82: Gráfico da frequência cardíaca com a pressão arterial sistólica

O facto da pressão sistólica ser preenchida com múltiplos de 10 e 5 faz com que este atributo seja encarado como uma variável discreta, daí o aspeto riscado (linhas verticais) do gráfico (figura 82). É percetível a existência de alguns *outliers* nesta variável.

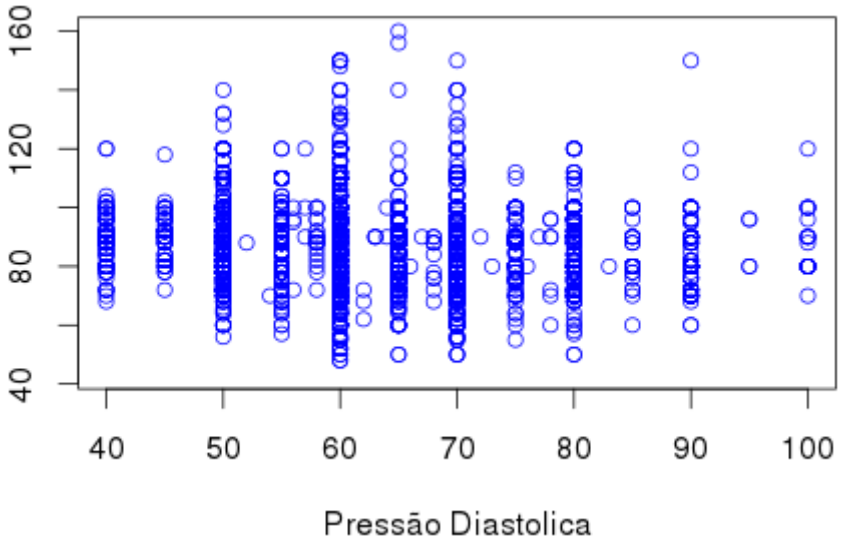

Figura 83: Gráfico da frequência cardíaca com a pressão arterial diastólica

O facto da pressão diastólica ser preenchida com múltiplos de 10 e 5 faz com que este atributo encarado como uma variável discreta daí o aspeto riscado do gráfico (figura 83).

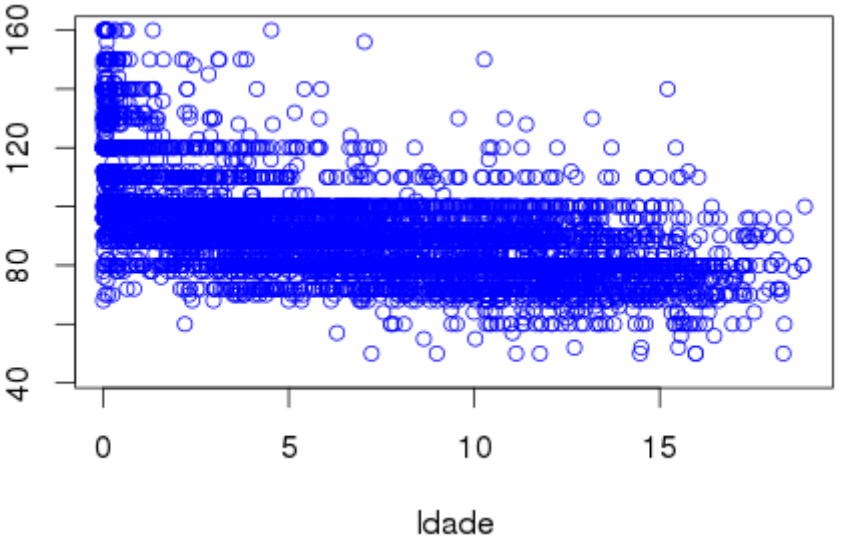

Figura 84: Gráfico da frequência cardíaca com a idade

A frequência cardíaca varia com a idade, geralmente, esta diminui ao longo da idade, o que se verifica no gráfico anterior (figura 84).

Análise Multivariada

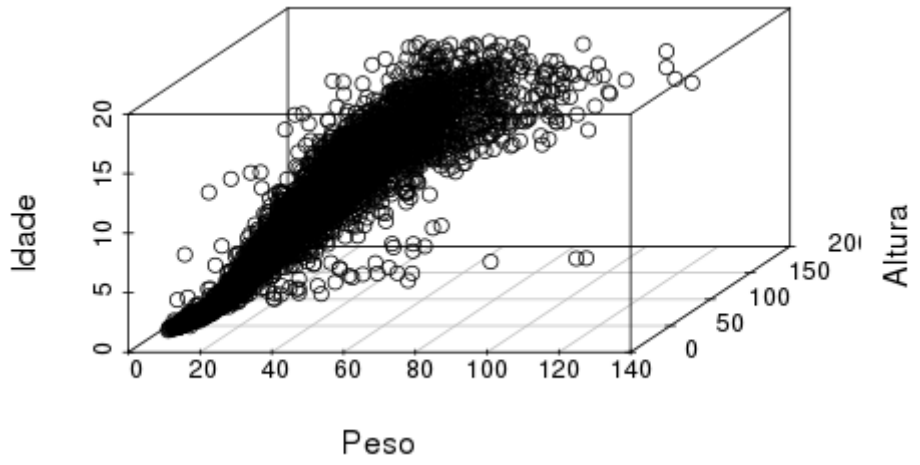

Figura 85: Gráfico da idade com o peso e a altura

No gráfico (figura 85) anterior podemos verificar que com o aumento da idade, o peso e a altura crescem exponencial até chegar aos 16/17 anos, após essa idade aumenta mais suavemente. Podemos verificar algumas ocorrências em que a idade é aproximadamente zero, isto deve-se ao facto anteriormente referido no gráfico da figura 76.

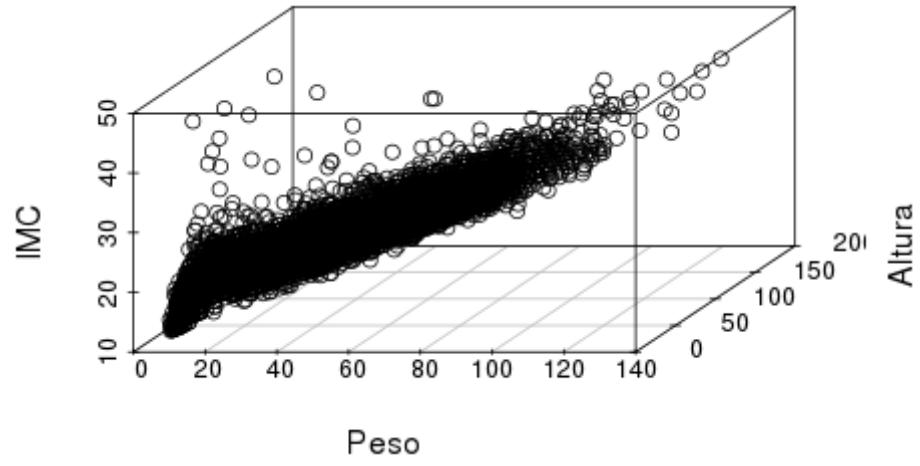

Figura 86: Gráfico do índice de massa corporal com o peso e a altura

Na nossa opinião o gráfico que agrupa o *IMC*, o peso e a altura corresponde ao esperado, um crescimento rápido do *IMC* quando o paciente tem um peso e uma altura baixa. Depois o *ICM* vai crescendo linearmente com o aumento do peso e da altura. No início é possível verificar alguns *outliers*, nós acreditamos que isso se deve à alguns valores elevados do peso com alturas mais pequenas.

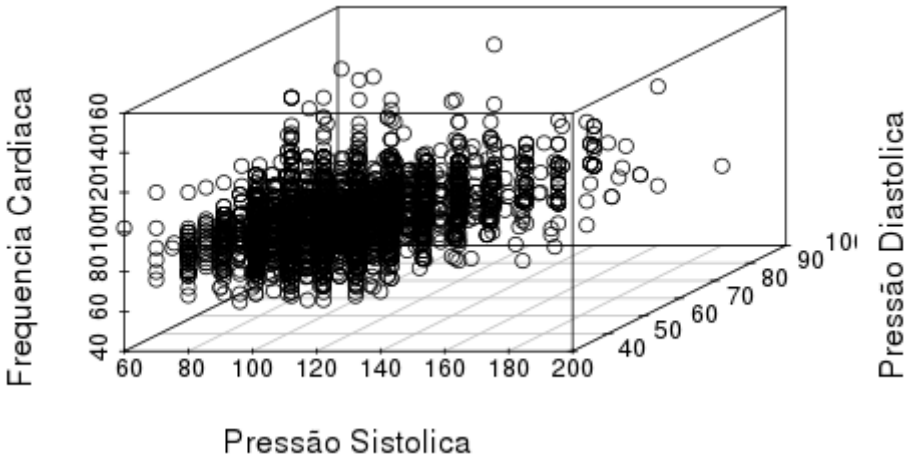

Figura 87: Gráfico da frequência cardíaca com a pressão arterial sistólica e a diastólica

Os dados parecem bem distribuídos, tendo em conta os factores referidos nas figuras 82 e 83.

Modelos Preditivos

Para o processamento de dados no *Weka* foi necessário uma conversão dos dados em "R"para este *software*. O processamento de dados no *WEKA* necessita da eliminação de todos os "*NA*", isso foi feito com o seguinte código:

```
#retira NA para o WEKA
data_WEKA <- sapply(data, as.character)
data_WEKA[is.na(data_WEKA)] <- ""
```

No *WEKA* foi necessário remover atributos, tais como, o "*ID*", a "*DN*", o "Atendimento"e o "Sexo"para que os algoritmos não detectassem a existência de qualquer relação com estes atributos. O "*ID*"foi removido pois este não apresenta informação relevante para esta análise. A "*DN*"e "Atendimento"são apenas importantes para o atributo "IDADE". Na variável "Sexo", com base no resultado da árvore produzida (árvore de decisões), chegamos a conclusão que essa informação é irrelevante e apenas servia para criar relações que eram pouco percetíveis e de credibilidade reduzida.

- **Árvore de decisões**

Na árvore de decisões optamos pelo algoritmo *J48* que é uma implementação do *C4.5* do *WEKA*. Obtivemos 15.246 (5.728+9.418) das instâncias correctamente classificadas. A *confusion matrix* (figura 88) mostra-nos que apenas 850 das 6.578 ocorrências que são classificadas como "Anormais", não são consideradas "Anormais"e 209 das 9.727 instâncias que são classificadas como "Normais", não são consideradas "Normais". O elevado valor de ocorrências classificadas correctamente, mais uma vez, prova a existência de um grau elevado de relação dos atributos no *training set*.

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     data_WEKA-weka.filters.unsupervised.attribute.Remove-R2,
6-7,9-weka.filters.unsupervised.attribute.Remove-R16
Instances:    16462
Attributes:   17
              Peso
              Altura
              IMC
              IDADE
              PULSOS
              PA.SISTOLICA
              PA.DIASTOLICA
              PPA
              NORMAL.X.ANORMAL
              B2
              SOPRO
              FC
              HDA.1
              HDA2
              MOTIVO1
              MOTIVO2
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree
------------------

SOPRO = Sistólico
|    <= 16174: Anormal (5704.36/167.36)
|    > 16174: Normal (97.42/36.0)
SOPRO = Ausente
|   B2 = Normal
|   |   PULSOS = Normais: Normal (10237.43/807.99)
|   |   PULSOS = Outro: Normal (13.63/4.0)
|   |   PULSOS = Amplos: Normal (7.01/1.0)
|   |   PULSOS = Femorais diminuidos: Anormal (11.02/2.02)
|   |   PULSOS = Diminuídos : Anormal (5.01/2.01)
|   B2 = Desdob fixo: Anormal (54.03/15.03)
|   B2 = Outro
|   |    <= 3376: Normal (10.0/1.0)
|   |    > 3376: Anormal (28.02/6.02)
|   B2 = Hiperfonética: Anormal (66.04/7.04)
|   B2 = Única: Anormal (8.01/1.01)
SOPRO = Contínuo
|    <= 16298: Anormal (46.01/7.01)
|    > 16298: Normal (3.01)
SOPRO = Diastólico: Anormal (11.0/1.0)
SOPRO = Sistolico e diastólico: Anormal (3.0/0.0)

Number of Leaves  :     16

Size of the tree :  22
```

```
Time taken to build model: 1.52 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances       15246               93.5051 %
Incorrectly Classified Instances      1059                6.4949 %
Kappa statistic                          0.8629
Mean absolute error                      0.1185
Root mean squared error                  0.2434
Relative absolute error                 24.6266 %
Root relative squared error             49.6191 %
Coverage of cases (0.95 level)          98.9758 %
Mean rel. region size (0.95 level)      82.49   %
Total Number of Instances            16305
Ignored Class Unknown Instances                157

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,871     0,021     0,965       0,871   0,915      0,866    0,928     0,900     Anormal
               0,979     0,129     0,918       0,979   0,947      0,866    0,929     0,914     Normal
Weighted Avg.  0,935     0,086     0,937       0,935   0,934      0,866    0,929     0,909

=== Confusion Matrix ===

    a    b   <-- classified as
 5728  850 |    a = Anormal
  209 9518 |    b = Normal
```

Figura 88: *Buffer* da árvore de decisões

- ***Support Vector Machines***

No *SVM* temos um conjunto bastante bom de instâncias correctamente classificadas, como podemos observar 16.214 (6.496+9.720) das instâncias são classificadas correctamente. Analisando a *confusion matrix* (figura 89) apenas 82 das 6.496 ocorrências que são classificadas como "Anormais", não são consideradas como "Anormais"e 7 das 9.720 instâncias que são classificadas como "Normais", não são consideradas como "Normais". Sendo o *SVM* um classificador binário, concluímos que estamos perante um *training set* bastante bom com uma percentagem de instâncias corectamente classificadas, atingindo um valor muito próximo dos 100%.

```
=== Run information ===

Scheme:       weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0
-R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -model /home/bastos -seed 1
Relation:     data_WEKA-weka.filters.unsupervised.attribute.Remove-R2,
6-7,9-weka.filters.unsupervised.attribute.Remove-R16
Instances:    16462
Attributes:   17
              Peso
              Altura
              IMC
              IDADE
              PULSOS
              PA.SISTOLICA
              PA.DIASTOLICA
              PPA
              NORMAL.X.ANORMAL
              B2
              SOPRO
              FC
              HDA.1
              HDA2
              MOTIVO1
              MOTIVO2
Test mode:    evaluate on training data

=== Classifier model (full training set) ===

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 363.3 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 116.53 seconds

=== Summary ===

Correctly Classified Instances       16216               99.4542 %
Incorrectly Classified Instances        89                0.5458 %
Kappa statistic                          0.9886
Mean absolute error                      0.0055
Root mean squared error                  0.0739
Relative absolute error                  1.134  %
Root relative squared error             15.0598 %
Coverage of cases (0.95 level)          99.4542 %
Mean rel. region size (0.95 level)      50      %
Total Number of Instances            16305
Ignored Class Unknown Instances                157

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,988     0,001     0,999       0,988   0,993      0,989    0,993     0,990     Anormal
               0,999     0,012     0,992       0,999   0,995      0,989    0,983     0,977     Normal
Weighted Avg.  0,995     0,008     0,995       0,995   0,995      0,989    0,987     0,982

=== Confusion Matrix ===

    a    b   <-- classified as
 6496   82 |    a = Anormal
    7 9720 |    b = Normal
```

Comparações

Ambos os modelos utilizados apresentam percentagens elevadas no que diz respeito a classificação correcta de instâncias, o que prova que temos um *training set* credível. No *SVM* temos apenas uma percentagem de 0.5% de instâncias classificadas incorretamente, enquanto que na árvore de decisões temos uma percentagem de 6.4%. Embora as árvores sejam mais adequadas para a classificação de classes, dependo do caso, aqui estas obtiveram um resultado inferior a técnica do *SVM*. Este facto está relacionado com a linearidade dos dados, sendo que uma maior linearidade faz com que se obtenha um resultado melhor num *SVM* do que numa árvore de decisões. De qualquer das maneiras, a árvore de decisões tem a vantagem de ser facilmente traduzida para a linguagem humana, produzindo um modelo que relaciona os atributos, o que facilita na interpretação dos resultados.

A árvore de decisões produziu o seguinte resultado:

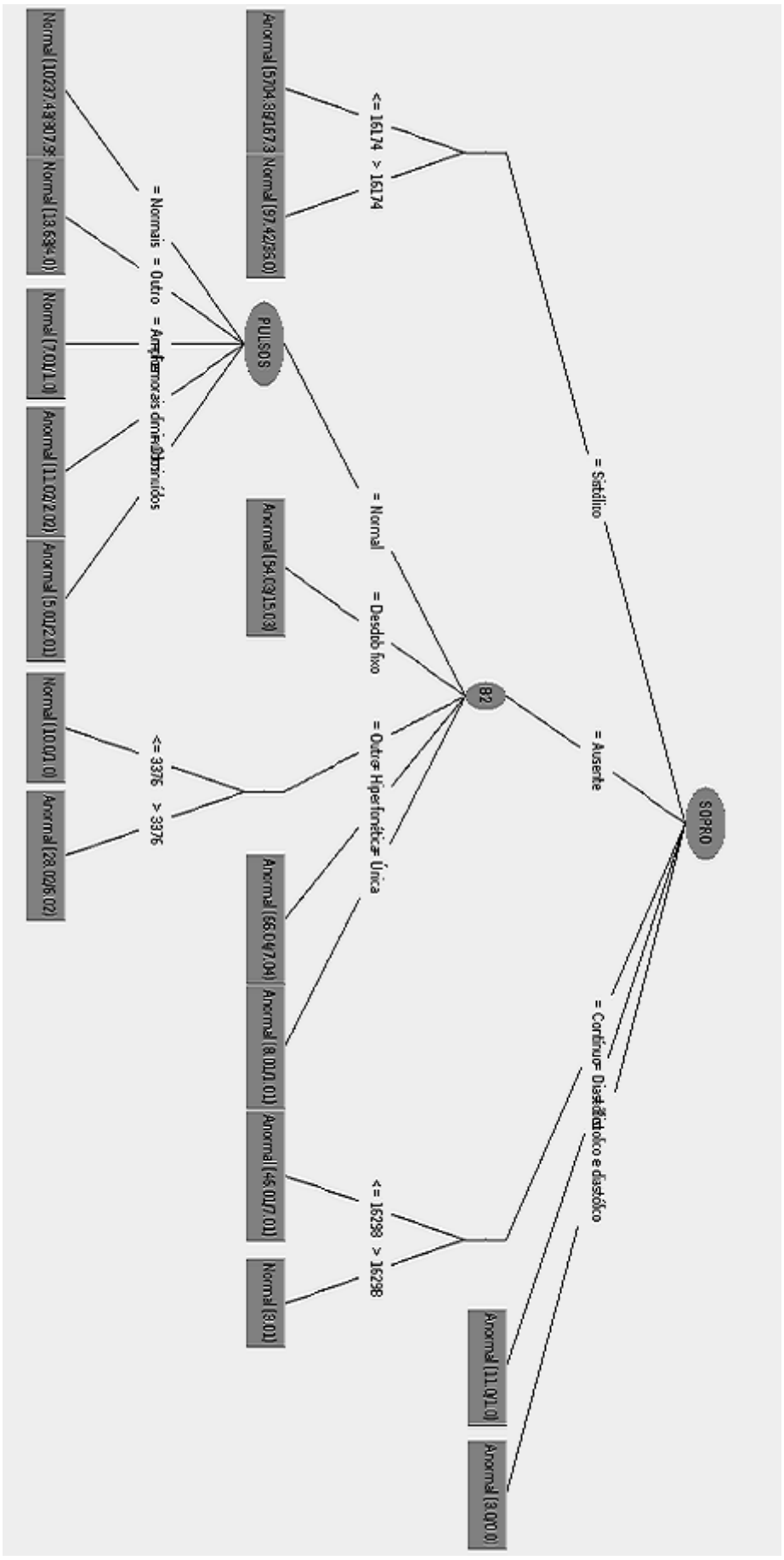

Figura 89: Árvore de decisões

Conclusão

Com base no modelo produzido pela árvore de decisões tiramos a seguintes conclusões.

A existência de um sopro sistólico determina com 98% de probabilidade a existência de uma patologia.

A existência de um sopro contínuo determina com 93% de probabilidade a existência de uma patologia.

A existência de um sopro siastólico provavelmente determina a existência de uma patologia mas devido a dimensão da amostra, apenas onze pacientes, esta afirmação não muito fiável.

A existência de um sopro sistólico e diastólico provavelmente determina a possibilidade de uma patologia, mas devido a dimensão da amostra, apenas três pacientes, não é garantida de todo a veracidade desta afirmação.

A inexistência de sopro não determina a ausência de patologia, será necessário analisar outros parâmetros, tais como o tipo do segundo batimento do coração ("*B2*"):

Um "*B2*Desdob fixo"determina a existência de uma patologia.

Um "*B2*Hiperfonética"determina a existência de uma patologia.

Um "*B2*Única"determina a existência de uma patologia.

Um "*B2*Outro"determina com 74% de probabilidade a existência de uma patologia.

Um "*B2*"'Normal' não determina a ausência da patologia, será necessário analisar outros parâmetros tais como o pulso ("PULSOS"):

Um pulso "Normal"determina a inexistência de uma patologia.

Um pulso "Outro"determina a inexistência de uma patologia.

Um pulso "Amplos"determina a inexistência de uma patologia.

Um pulso "Femorais diminuídos"determina a existência de uma patologia.

Um pulso "Diminuídos"determina a existência de uma patologia.

Inicialmente, tínhamos incluído a variável "Sexo"nos modelos preditivos, mas após uma análise mais cuidadosa e com base nos resultados obtidos, chegamos a conclusão que a influência desta na determinação da existência de patologia cardíaca era baixa. Portanto, decidimos retirar esta variável dos modelos criados para a cruzamento de informação, com o objectivo de encontrar variáveis que sejam indicadoras de patologias cardíacas em crianças e adolescentes.

Bibliografia

[1] *http://www.saedsayad.com/data$_p$reparation.htm*

[2] *http://www.euroclinix.com.pt/tensao-pressao-sistolica-diastolica.html*

[3] *https://pt.wikipedia.org/wiki/%C3%8Dndice$_d$e$_m$assa$_c$orporal*

[4] *http://www.bloodpressureuk.org/BloodPressureandyou/Thebasics/Bloodpressurechart*

[5] *http://www.tuasaude.com/frequencia-cardiaca/*

[6] *https://en.wikipedia.org/wiki/Body$_m$ass$_i$ndex/media/File : BMIGirls$_1$.svg*