

Sumário

| | |
|---|-----------|
| 1. INTRODUÇÃO E OBJETIVOS | 2 |
| 2. METODOLOGIA DE ANÁLISE | 3 |
| 2.1. Resumo Tabular | 3 |
| 2.2. Resumo Gráfico | 4 |
| 2.2.1 Histograma | 4 |
| 2.2.2 Gráfico da função densidade por Kernel | 4 |
| 2.2.3 Boxplot | 4 |
| 2.2.4 Gráfico de barras | 4 |
| 2.3. Resumo Numérico | 4 |
| 3. ANÁLISE DESCRITIVA | 5 |
| 3.1. Bibliotecas utilizadas | 5 |
| 3.2. Carregamento dos dados | 6 |
| 3.3. Definindo as Variáveis Qualitativas | 7 |
| 3.4. Validação dos dados | 8 |
| 3.5. Avaliação dos dados | 9 |
| 3.5.1 Variáveis contínuas | 9 |
| 3.5.1.1 BLUEBOOK - Valor do Veículo | 9 |
| 3.5.1.2 RETAINED - Anos como cliente | 14 |
| 3.5.1.3 CLM_AMT - Valor de cobertura solicitado | 19 |
| 3.5.1.4 AGE - Idade em anos | 24 |
| 3.5.1.5 YOJ - Anos de trabalho | 29 |
| 3.5.2 Variáveis discretas | 34 |
| 3.5.2.1 NPOLICY - Número de apólices | 34 |
| 3.5.3 Variáveis nominais | 39 |
| 3.5.3.1 MAX_EDUC - Máximo nível educacional | 39 |
| 3.5.3.2 GENDER - Sexo | 42 |
| 3.5.3.3 MARRIED - Casado | 45 |
| 4. DISCUSSÃO E CONCLUSÕES | 48 |
| 5. REFERÊNCIAS | 49 |

1. INTRODUÇÃO E OBJETIVOS

Análise exploratória de dados antigamente era chamada simplesmente de estatística descritiva. Essa abordagem consiste em apresentar dados de forma organizada para facilitar a interpretação e, por fim, retirar conclusões acerca deles. Os dados a serem analisados são coletados previamente - por censo ou por amostragem - a partir de uma população (indivíduos, objetos ou fenômenos, por exemplo, que possuem características em comum que podem ser observadas e categorizadas); gerando assim uma amostra (conjunto de dados coletados de uma parte da população) - ou censo (conjunto de dados coletados de toda a população) - . Finalmente, os dados são estruturados; expostos em forma de gráficos e tabelas; e analisados para que então conclusões possam ser feitas - ou não, nem sempre amostras são suficientemente completas - acerca da amostra. A estatística indutiva busca propor hipóteses.

O relatório apresenta informações retiradas de uma base de dados previamente coletada, bem como uma interpretação dela. Ela provém de uma grande empresa de seguros alemã, referente às reclamações dos segurados sobre sinistros associados à carteira de seguro automobilístico da empresa germânica. O conjunto de dados em estudo foi fornecido pelo Prof. Dr. Afrânio Vieira.

A análise desse acervo de dados foi feita utilizando a linguagem de programação R, cujo principal objetivo é, justamente, facilitar análises estatísticas, bem como a criação e manipulação de gráficos. Além disso, foi também utilizado o software Rstudio, que pode ser obtido em <https://www.rstudio.com/>. A linguagem R está disponível para download em <https://www.r-project.org/>. Ao longo do relatório, três métodos de análise foram utilizados: resumo tabular, análise de dados a partir de tabelas; resumo gráfico, a partir de gráficos; e resumo numérico.

Uma base de dados pode ser descrita por medidas de tendência central, como moda, média aritmética e mediana; medidas de dispersão, para identificar a variabilidade do conjunto de dados; e medidas de posição, que permite uma melhor análise se o conjunto de dados possuir outliers (valores extremos). Essas três maneiras de descrever dados fazem parte do resumo numérico.

2. METODOLOGIA DE ANÁLISE

Nesta seção serão evidenciados e explicados brevemente os métodos de análise utilizados no relatório. Eles são: resumo tabular; resumo gráfico; e resumo numérico. Os métodos de análise estatística são empregados a fim de simplificar e otimizar o processo de análise de dados.

2.1. Resumo Tabular

Tabelas são estruturas sistemáticas criadas para sintetizar um conjunto de dados. Tabelas podem ser simples (apenas uma variável) ou cruzada (duas ou mais variáveis) - em certas pesquisas pode ser interessante, além de exibir os dados coletados, mostrar o sexo da pessoa, por exemplo.

Uma tabela é composta pelo seu título, corpo e fonte. O título deve ser colocado no topo da estrutura informando o assunto; é conveniente que três perguntas sejam respondidas ao lê-lo: o que são os dados nela representados? De que lugar eles foram coletados? Quando foram coletados? O corpo é composto por linhas e colunas, e é nele que os dados são apresentados. Por fim, é na fonte que se apresenta a origem dos dados, de onde eles foram retirados.

O resumo tabular consiste em representar dados em uma tabela. As tabelas apresentadas no relatório possuem 5 colunas: a primeira informa o intervalo de valores da variável aleatória estudada; a segunda, informa a frequência, quantidade de valores dentro do intervalo; a terceira informa a frequência acumulada, a soma de todas as frequências acima com a frequência da linha atual; a quarta informa a porcentagem; e a quinta apresenta a porcentagem acumulada, soma de todas as porcentagens acima com a porcentagem da linha atual. Abaixo de cada tabela há um comentário sobre os dados apresentados na tabela.

2.2 Resumo Gráfico

Muitas vezes tabelas não são adequadas para apresentar determinado conjunto de dados e, portanto, recorre-se a outras formas de representação de dados, gráficos por exemplo. Gráficos são figuras que facilitam a visualização e interpretação dos dados. Assim como as tabelas, possuem título e fonte. Existem diversas variações de gráficos e a escolha dela normalmente está atrelada ao tipo da variável aleatória estudada. Foram utilizados quatro tipos de gráficos: histograma, gráfico de densidade por kernel, boxplot e gráfico de barras.

2.2.1 Histograma

Histograma é um gráfico composto de barras agrupadas. O eixo das abscissas indica os limites do intervalo de uma barra e outra. Já o eixo das coordenadas representa, normalmente, a frequência dos valores do intervalo.

2.2.2 Gráfico da função densidade por Kernel

aa

2.2.3 Boxplot

Boxplot é um gráfico composto por caixas e quartis. É interessante utilizar este tipo de gráfico quando se quer observar posição e dispersão dos dados. Nele é possível identificar os outliers, onde a maior parte dos dados estão concentrados, a mediana e comparar o tamanho das amostras - pela largura de cada caixa. O limite superior do primeiro quartil - e limite inferior do segundo quartil - é a base da caixa. O limite superior do segundo quartil - e limite inferior do terceiro quartil - é a linha horizontal dentro da caixa, que representa a mediana. O limite inferior do primeiro quartil é a reta imaginária paralela e abaixo à base da caixa, com coordenada obtida pela subtração do valor da coordenada da base da caixa com 1,5 vezes a **distância interquartilica** - distância entre o topo e a base da caixa. Analogamente, o limite superior do terceiro quartil é a reta imaginária paralela e acima ao topo da caixa, com coordenada obtida pela soma do valor da coordenada do topo da caixa com 1,5 vezes a distância interquartilica. Valores extremos dificultam a análise de dados com medidas de dispersão, como desvio padrão e variância; por isso, os boxplots os identificam e os isolam.

2.2.4 Gráfico de barras

Gráfico de barras é composto por barras (verticais ou horizontais) não agrupadas. Num dos eixos indica o valor ou categoria de uma variável aleatória e noutro a frequência da variável.

2.3 Resumo Numérico

Além de elementos visuais, resumos numéricos também são úteis para análise de dados. **Medidas de tendência central**, ou posição, permitem analisar a posição da concentração dos dados; moda, média e mediana são exemplos. **Medidas de dispersão** são utilizadas para analisar a variabilidade do conjunto de dados; variância, desvio-padrão e coeficiente de variação são exemplos.

3. ANÁLISE DESCRITIVA

3.1. Bibliotecas utilizadas

```
library(tidyverse)
```

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

## v ggplot2 3.3.3      v purrr  0.3.4
## v tibble  3.1.1      v dplyr  1.0.6
## v tidyr   1.1.3      v stringr 1.4.0
## v readr   1.4.0      v forcats 0.5.1

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

```
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##      src, summarize

## The following objects are masked from 'package:base':
##
##      format.pval, units
```

```
library(psych)
```

```
##
## Attaching package: 'psych'

## The following object is masked from 'package:Hmisc':
##
##      describe

## The following objects are masked from 'package:ggplot2':
##
##      %+%, alpha
```

```
library(descriptor)
library(summarytools)
```

```
## Registered S3 method overwritten by 'pryr':
##   method      from
##   print.bytes Rcpp
```

```
## For best results, restart R session and update pander using devtools:: or remotes::install_github('r')
```

```
##
## Attaching package: 'summarytools'
```

```
## The following objects are masked from 'package:Hmisc':
##
##   label, label<-
```

```
## The following object is masked from 'package:tibble':
##
##   view
```

3.2. Carregamento dos dados

```
path <- "./"
setwd(path)

Claim.Data <- read_csv2(file = "ClaimData.csv")
```

```
## i Using '\',\'' as decimal and '\'.\'' as grouping mark. Use 'read_delim()' for more control.
```

```
##
## -- Column specification -----
## cols(
##   Client = col_double(),
##   BLUEBOOK = col_double(),
##   RETAINED = col_double(),
##   NPOLICY = col_double(),
##   CLM_AMT = col_double(),
##   AGE = col_double(),
##   YOJ = col_double(),
##   GENDER = col_character(),
##   MARRIED = col_character(),
##   MAX_EDUC = col_character()
## )
```

```
glimpse(Claim.Data)
```

```
## Rows: 10,303
## Columns: 10
```

```
## $ Client    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
## $ BLUEBOOK <dbl> 9860, 1500, 30460, 16580, 23030, 20730, 27420, 24360, 36460, ~
## $ RETAINED <dbl> 6, 4, 4, 4, 4, 9, 10, 6, 1, 4, 1, 17, 6, 1, 13, 4, 4, 13, 1, ~
## $ NPOLICY   <dbl> 2, 2, 1, 2, 1, 1, 1, 3, 3, 3, 2, 2, 2, 2, 1, 1, 2, 2, 1, 1, 1~
## $ CLM_AMT   <dbl> 3336.00, 5583.00, 39103.88, 0.00, 0.00, 0.00, 5342.00, 0.00, ~
## $ AGE       <dbl> 42, 35, 58, 45, 49, 38, 60, 43, 43, 43, 42, 42, 58, 27, 38, 5~
## $ YOJ       <dbl> 13, 12, 13, 14, 13, 10, 7, 11, 11, 11, 13, 13, NA, 11, 9, 12,~
## $ GENDER    <chr> "M", "M", "M", "F", "F", "F", "F", "F", "F", "F", "F", "F", "~
## $ MARRIED   <chr> "Yes", "No", "No", "Yes", "Yes", "Yes", "No", "No", "No", "No~
## $ MAX_EDUC  <chr> "<High School", "High School", "Masters", "High School", "Hig~
```

3.3. Definindo as Variáveis Qualitativas

```
Claim.Data$GENDER <- factor(
  Claim.Data$GENDER,
  levels = c("M", "F"),
  labels = c("Male", "Female")
)
Claim.Data$MARRIED <- factor(Claim.Data$MARRIED)
Claim.Data$MAX_EDUC <- ordered(
  Claim.Data$MAX_EDUC,
  levels = c("<High School", "High School", "Bachelors", "Masters", "PhD")
)
glimpse(Claim.Data)
```

```
## Rows: 10,303
## Columns: 10
## $ Client    <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18~
## $ BLUEBOOK <dbl> 9860, 1500, 30460, 16580, 23030, 20730, 27420, 24360, 36460, ~
## $ RETAINED <dbl> 6, 4, 4, 4, 4, 9, 10, 6, 1, 4, 1, 17, 6, 1, 13, 4, 4, 13, 1, ~
## $ NPOLICY   <dbl> 2, 2, 1, 2, 1, 1, 1, 3, 3, 3, 2, 2, 2, 2, 1, 1, 2, 2, 1, 1, 1~
## $ CLM_AMT   <dbl> 3336.00, 5583.00, 39103.88, 0.00, 0.00, 0.00, 5342.00, 0.00, ~
## $ AGE       <dbl> 42, 35, 58, 45, 49, 38, 60, 43, 43, 43, 42, 42, 58, 27, 38, 5~
## $ YOJ       <dbl> 13, 12, 13, 14, 13, 10, 7, 11, 11, 11, 13, 13, NA, 11, 9, 12,~
## $ GENDER    <fct> Male, Male, Male, Female, Female, Female, Female, Female, Fem~
## $ MARRIED   <fct> Yes, No, No, Yes, Yes, Yes, No, No, No, No, Yes, Yes, Yes, No~
## $ MAX_EDUC  <ord> <High School, High School, Masters, High School, High School,~
```

3.4. Validação dos dados

```
anyNA(Claim.Data)
```

```
## [1] TRUE
```

```
Claim.Data %>% is.na() %>% sum()
```

```
## [1] 555
```

```
Claim.Data %>% is.na() %>% unique()
```

```
##      Client BLUEBOOK RETAINED NPOLICY CLM_AMT  AGE  YOJ GENDER MARRIED
## [1,] FALSE      FALSE      FALSE  FALSE  FALSE FALSE FALSE  FALSE  FALSE
## [2,] FALSE      FALSE      FALSE  FALSE  FALSE FALSE TRUE  FALSE  FALSE
## [3,] FALSE      FALSE      FALSE  FALSE  FALSE TRUE  FALSE FALSE  FALSE
##      MAX_EDUC
## [1,] FALSE
## [2,] FALSE
## [3,] FALSE
```

```
Claim.Data[is.na(Claim.Data$AGE),]
```

```
## # A tibble: 7 x 10
##      Client BLUEBOOK RETAINED NPOLICY CLM_AMT  AGE  YOJ GENDER MARRIED MAX_EDUC
##      <dbl>    <dbl>    <dbl>    <dbl>    <dbl> <dbl> <dbl> <fct>  <fct>    <ord>
## 1   1089   14500         1         2   3444   NA    0 Female No      <High Sch~
## 2   1694    3100         9         4   6142   NA    8 Female No      <High Sch~
## 3   2155    2950        10         1   4798   NA    9 Female No      <High Sch~
## 4   5206    1500        10         1   3092   NA    0 Male   No      Bachelors
## 5   9449    3180        11         2   2541   NA    0 Female No      Bachelors
## 6   9742    2600        10         1     0    NA    0 Female Yes    High Scho~
## 7   9980   20770         1         1   5640   NA   12 Male   Yes    High Scho~
```

```
Claim.Data[is.na(Claim.Data$YOJ),]
```

```
## # A tibble: 548 x 10
##      Client BLUEBOOK RETAINED NPOLICY CLM_AMT  AGE  YOJ GENDER MARRIED MAX_EDUC
##      <dbl>    <dbl>    <dbl>    <dbl>    <dbl> <dbl> <dbl> <fct>  <fct>    <ord>
## 1     13   11050         6         2     0   58   NA Male   Yes    Masters
## 2     55    8760         1         2     0   47   NA Male   Yes    High Sch~
## 3     56    8760         6         2     0   47   NA Male   Yes    High Sch~
## 4     97   14510         1         1     0   45   NA Male   No     Bachelors
## 5    100   25660         4         1  4487   27   NA Male   No     Bachelors
## 6    134    4700         7         1  4995   32   NA Female No     Bachelors
## 7    154   17190         1         1     0   33   NA Male   Yes    Bachelors
## 8    161   11910         7         3  7907   44   NA Female Yes    <High Sc~
## 9    165   19780         1         2     0   46   NA Male   No     <High Sc~
## 10   197   10020        17         2     0   45   NA Female No     <High Sc~
## # ... with 538 more rows
```


3.5. Avaliação dos dados

3.5.1 Variáveis contínuas

3.5.1.1 BLUEBOOK - Valor do Veículo

3.5.1.1.1 Estatísticas básicas do R

```
mean(Claim.Data$BLUEBOOK)      # media
```

```
## [1] 15660.37
```

```
median(Claim.Data$BLUEBOOK)    # mediana
```

```
## [1] 14400
```

```
min(Claim.Data$BLUEBOOK)       # minimo
```

```
## [1] 1500
```

```
max(Claim.Data$BLUEBOOK)       # maximo
```

```
## [1] 69740
```

```
var(Claim.Data$BLUEBOOK)       # variancia
```

```
## [1] 71039286
```

```
sd(Claim.Data$BLUEBOOK)        # desvio padrao
```

```
## [1] 8428.481
```

```
IQR(Claim.Data$BLUEBOOK)       # distancia interquartilica
```

```
## [1] 11690
```

```
summary(Claim.Data$BLUEBOOK)    # Min, Q1, Q2, media, Q3, Max
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1500   9200   14400   15660   20890   69740
```

```
quantile(Claim.Data$BLUEBOOK)   # Min, Q1, Q2, Q3, Max
```

```
##      0%    25%    50%    75%   100%
##      1500   9200  14400  20890  69740
```

```
quantile(Claim.Data$BLUEBOOK, type = 7, probs = c(.01, .05, .10, .90, .95, .99)) # percentis
```

```
##      1%      5%     10%     90%     95%     99%
## 1500.0 4801.0 5990.0 27430.0 30948.0 38899.4
```

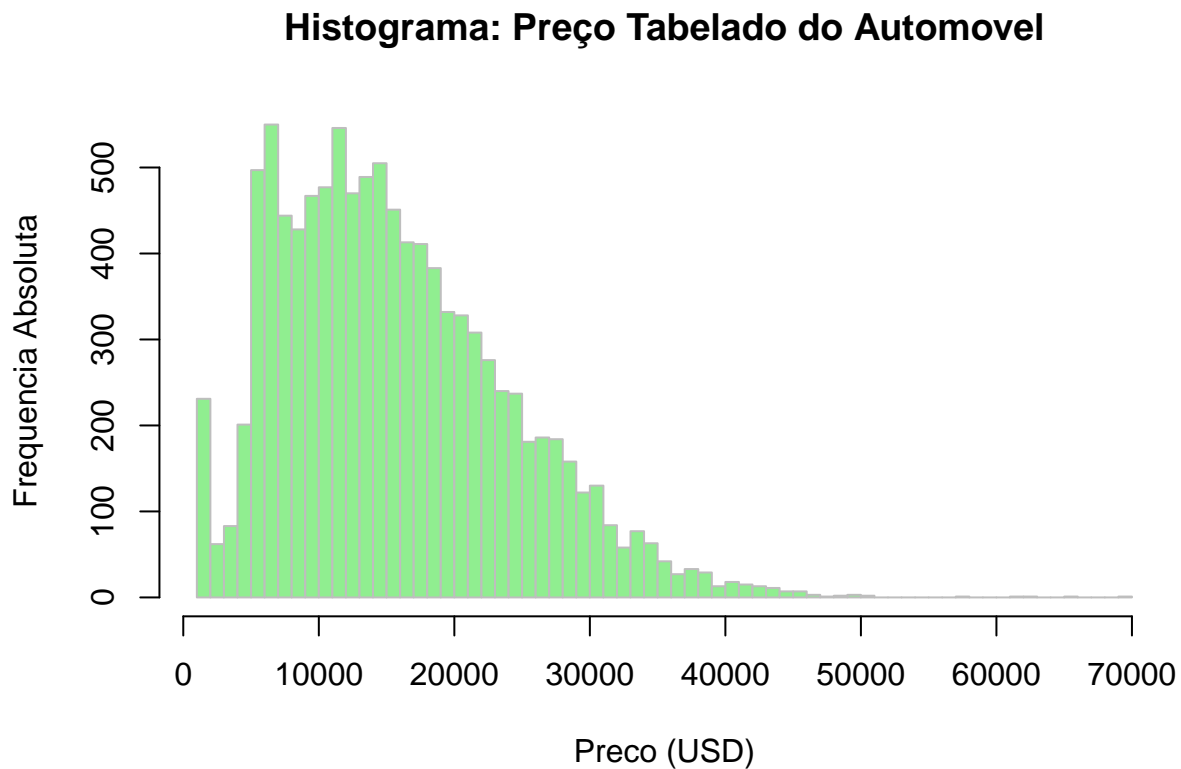
3.5.1.1.2 Resumo da biblioteca Hmisc

```
describe(Claim.Data$BLUEBOOK)
```

```
##      vars      n      mean      sd median trimmed      mad min  max range skew
## X1      1 10303 15660.37 8428.48  14400 14993.41 8539.78 1500 69740 68240 0.77
##      kurtosis      se
## X1          0.65 83.04
```

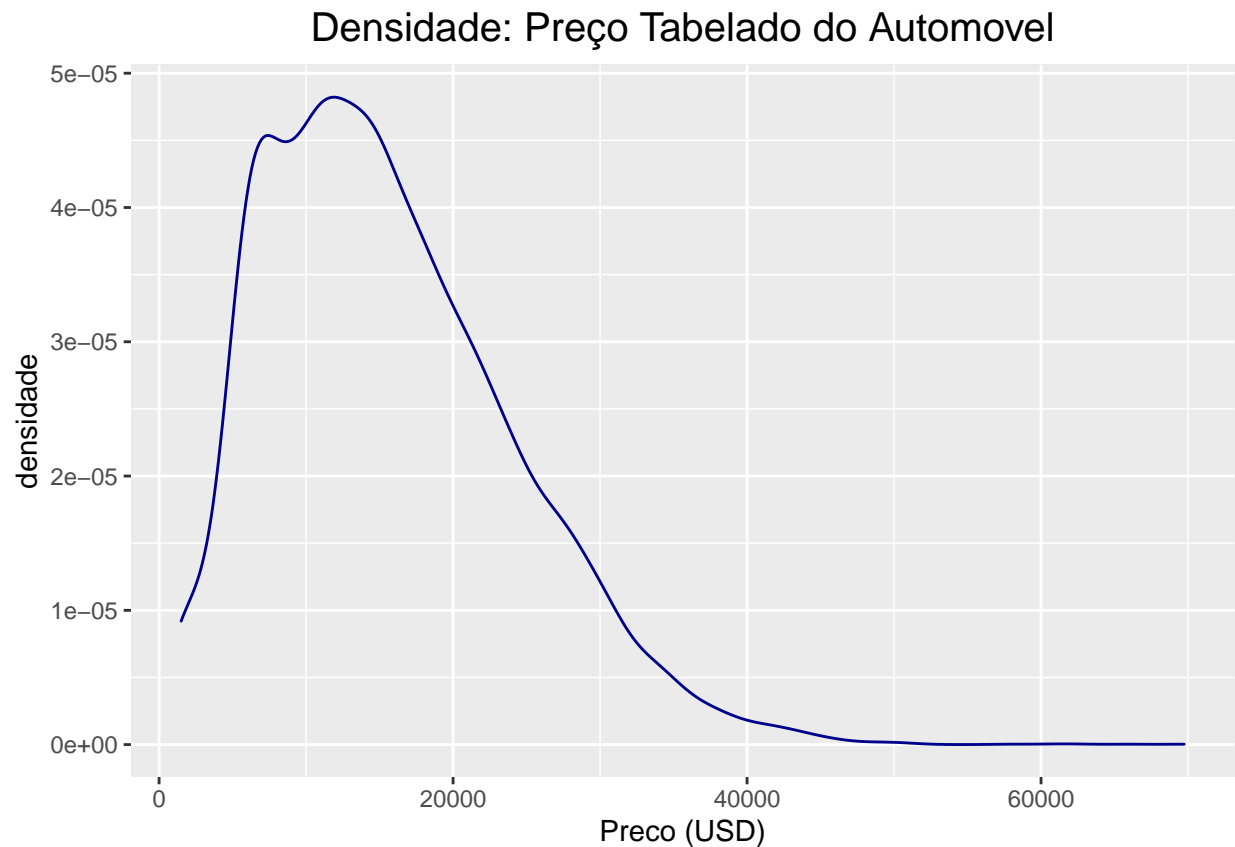
3.5.1.1.3 Histograma

```
hist(Claim.Data$BLUEBOOK, breaks = "fd",
     col = "lightgreen", border = "grey",
     main = "Histograma: Preço Tabelado do Automovel",
     xlab = "Preco (USD)", ylab = "Frequencia Absoluta"
)
```



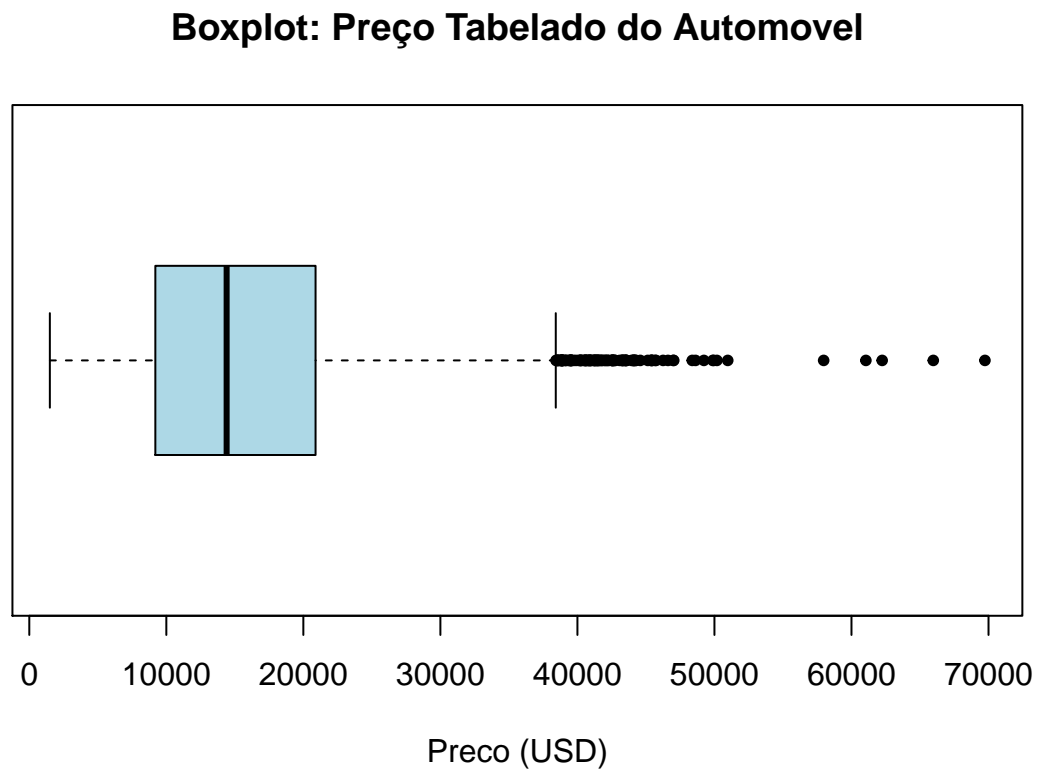
3.5.1.1.4 Gráfico de densidade por Kernel

```
graf <- ggplot(data = Claim.Data, mapping = aes(x = BLUEBOOK)) +  
  geom_density(mapping = aes(x = BLUEBOOK),  
              bw = "nrd",  
              color = "darkblue") +  
  ggtitle("Densidade: Preço Tabelado do Automovel") +  
  xlab("Preco (USD)") + ylab("densidade") +  
  theme(plot.title = element_text(hjust = 0.5, size=15) )  
graf
```



3.5.1.1.5 Boxplot

```
boxplot(  
  Claim.Data$BLUEBOOK, horizontal = T,  
  col = "lightblue", pch = 20,  
  main = "Boxplot: Preço Tabelado do Automovel",  
  xlab = "Preco (USD)"  
)
```



3.5.1.1.6 Resumo Tabular

```
ds_freq_table(Claim.Data, BLUEBOOK, bins = 20)
```

| Variable: BLUEBOOK | | | | | |
|--------------------|-----------|---------------|---------|-------------|--|
| Bins | Frequency | Cum Frequency | Percent | Cum Percent | |
| 1500 - 4912 | 545 | 545 | 5.29 | 5.29 | |
| 4912 - 8324 | 1655 | 2200 | 16.06 | 21.35 | |
| 8324 - 11736 | 1648 | 3848 | 16 | 37.35 | |
| 11736 - 15148 | 1666 | 5514 | 16.17 | 53.52 | |
| 15148 - 18560 | 1430 | 6944 | 13.88 | 67.4 | |
| 18560 - 21972 | 1123 | 8067 | 10.9 | 78.3 | |
| 21972 - 25384 | 838 | 8905 | 8.13 | 86.43 | |
| 25384 - 28796 | 609 | 9514 | 5.91 | 92.34 | |
| 28796 - 32208 | 380 | 9894 | 3.69 | 96.03 | |
| 32208 - 35620 | 206 | 10100 | 2 | 98.03 | |
| 35620 - 39032 | 109 | 10209 | 1.06 | 99.09 | |
| 39032 - 42444 | 51 | 10260 | 0.5 | 99.58 | |
| 42444 - 45856 | 33 | 10293 | 0.32 | 99.9 | |
| 45856 - 49268 | 7 | 10300 | 0.07 | 99.97 | |
| 49268 - 52680 | 4 | 10304 | 0.04 | 100.01 | |
| 52680 - 56092 | 0 | 10304 | 0 | 100.01 | |
| 56092 - 59504 | 1 | 10305 | 0.01 | 100.02 | |
| 59504 - 62916 | 2 | 10307 | 0.02 | 100.04 | |
| 62916 - 66328 | 1 | 10308 | 0.01 | 100.05 | |
| 66328 - 69740 | 1 | 10309 | 0.01 | 100.06 | |
| Total | 10303 | - | 100.00 | - | |

3.5.1.2 RETAINED - Anos como cliente

3.5.1.2.1 Estatísticas básicas do R

```
mean(Claim.Data$RETAINED)      # media

## [1] 5.329224

median(Claim.Data$RETAINED)    # mediana

## [1] 4

min(Claim.Data$RETAINED)       # minimo

## [1] 1

max(Claim.Data$RETAINED)       # maximo

## [1] 25

var(Claim.Data$RETAINED)       # variancia

## [1] 16.89704

sd(Claim.Data$RETAINED)        # desvio padrao

## [1] 4.110601

IQR(Claim.Data$RETAINED)       # distancia interquartilica

## [1] 6

summary(Claim.Data$RETAINED)    # Min, Q1, Q2, media, Q3, Max

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000   1.000   4.000   5.329   7.000  25.000

quantile(Claim.Data$RETAINED)   # Min, Q1, Q2, Q3, Max

##      0%  25%  50%  75% 100%
##      1    1    4    7   25

quantile(Claim.Data$RETAINED, type = 7, probs = c(.01, .05, .10, .90, .95, .99)) # percentis

##      1%  5% 10% 90% 95% 99%
##      1    1    1  11  13  17
```

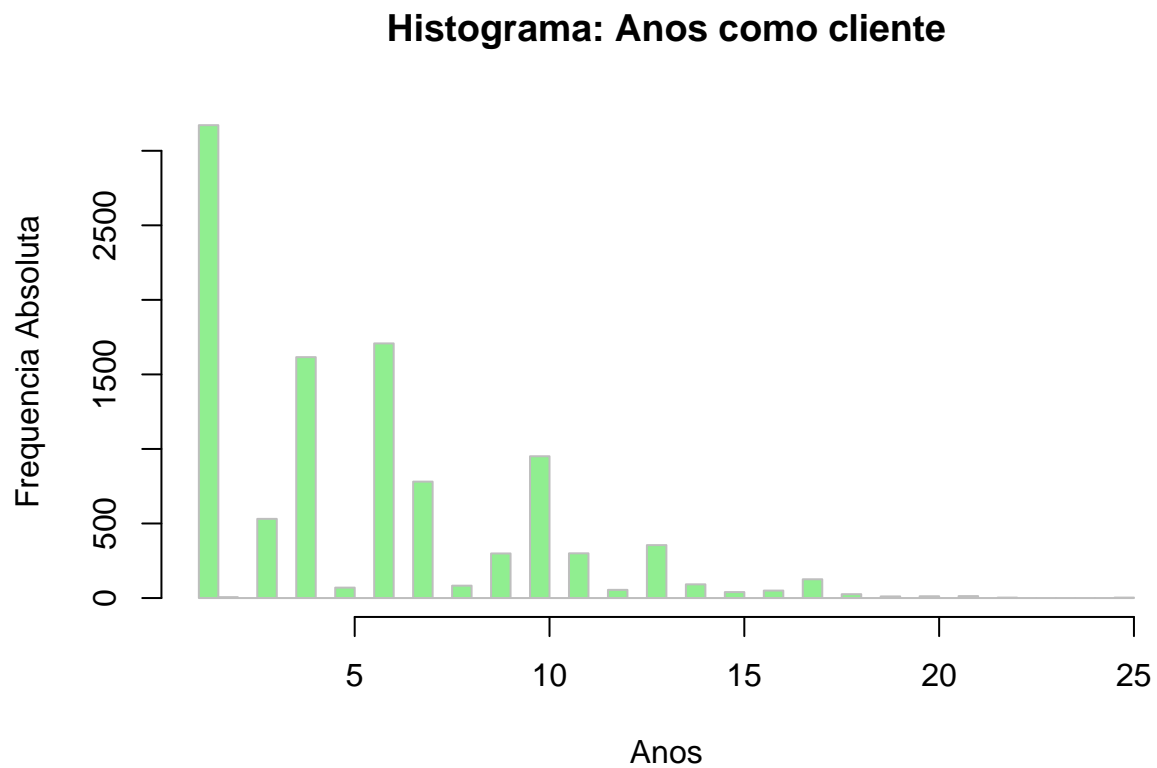
3.5.1.2.2 Resumo da biblioteca Hmisc

```
describe(Claim.Data$RETAINED)
```

```
##      vars      n mean   sd median trimmed  mad min max range skew kurtosis   se
## X1      1 10303 5.33 4.11      4    4.82 4.45   1  25   24  0.9    0.48 0.04
```

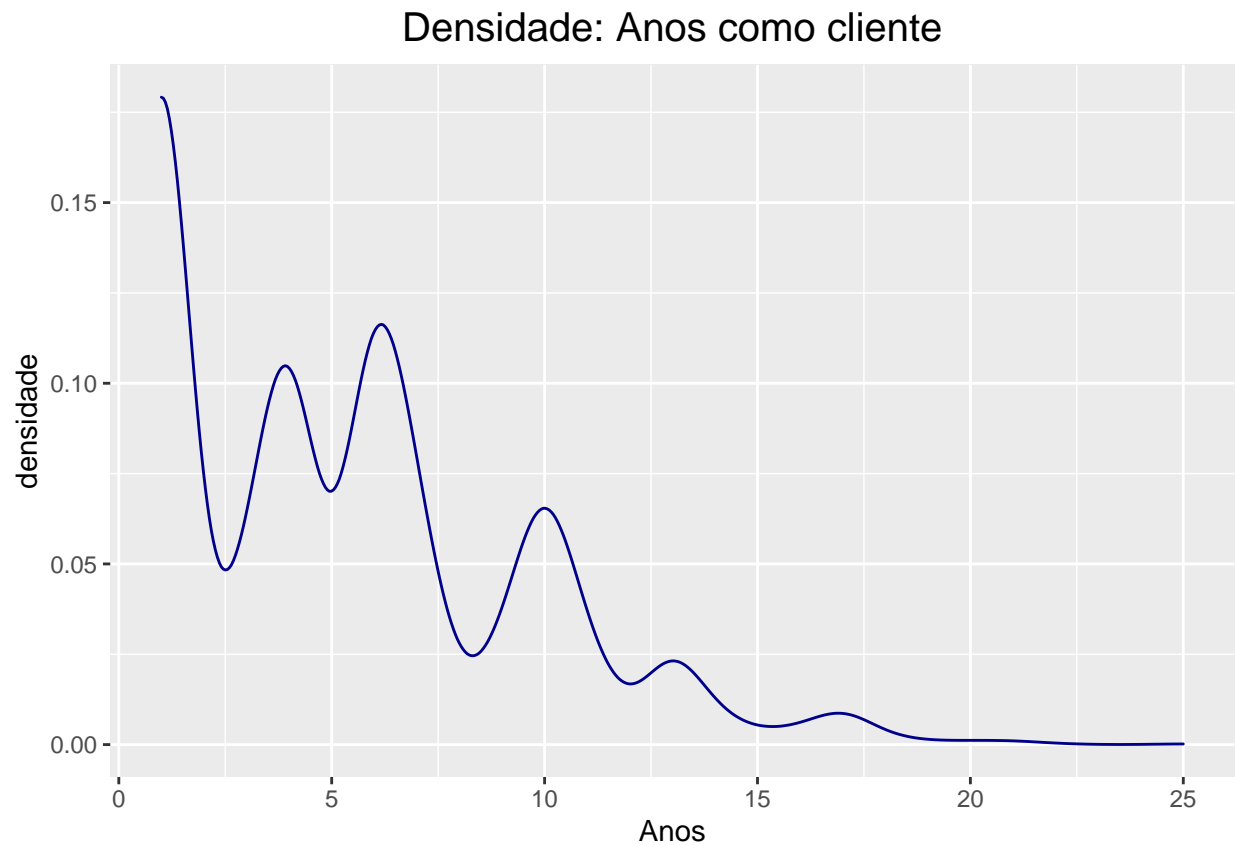
3.5.1.2.3 Histograma

```
hist(Claim.Data$RETAINED, breaks = "fd",
      col = "lightgreen", border = "grey",
      main = "Histograma: Anos como cliente",
      xlab = "Anos", ylab = "Frequencia Absoluta"
)
```



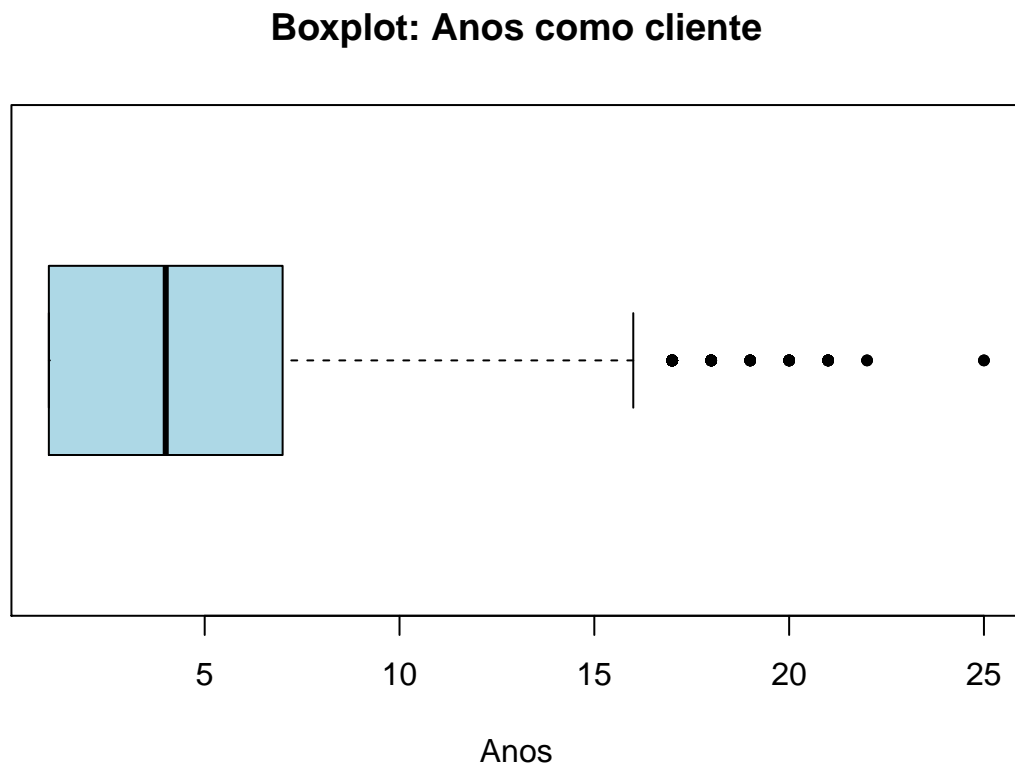
3.5.1.2.4 Grafico de densidade por Kernel

```
graf <- ggplot(data = Claim.Data, mapping = aes(x = RETAINED)) +  
  geom_density(mapping = aes(x = RETAINED),  
              bw = "nrd",  
              color = "darkblue") +  
  ggtitle("Densidade: Anos como cliente") +  
  xlab("Anos") + ylab("densidade") +  
  theme(plot.title = element_text(hjust = 0.5, size=15) )  
graf
```



3.5.1.2.5 Boxplot

```
boxplot(  
  Claim.Data$RETAINED, horizontal = T,  
  col = "lightblue", pch = 20,  
  main = "Boxplot: Anos como cliente",  
  xlab = "Anos"  
)
```



3.5.1.2.6 Resumo Tabular

```
ds_freq_table(Claim.Data, RETAINED, bins = 12)
```

```
##                               Variable: RETAINED  
## |-----|  
## | Bins   | Frequency | Cum Frequency | Percent   | Cum Percent |  
## |-----|  
## | 1 - 3   | 3709      | 3709          | 36        | 36          |  
## |-----|  
## | 3 - 5   | 2217      | 5926          | 21.52     | 57.52       |  
## |-----|  
## | 5 - 7   | 2559      | 8485          | 24.84     | 82.35       |  
## |-----|  
## | 7 - 9   | 1163      | 9648          | 11.29     | 93.64       |
```

| | | |
|----|--|--|
| ## | | ----- |
| ## | | 9 - 11 1550 11198 15.04 108.69 |
| ## | | ----- |
| ## | | 11 - 13 710 11908 6.89 115.58 |
| ## | | ----- |
| ## | | 13 - 15 487 12395 4.73 120.3 |
| ## | | ----- |
| ## | | 15 - 17 216 12611 2.1 122.4 |
| ## | | ----- |
| ## | | 17 - 19 163 12774 1.58 123.98 |
| ## | | ----- |
| ## | | 19 - 21 36 12810 0.35 124.33 |
| ## | | ----- |
| ## | | 21 - 23 16 12826 0.16 124.49 |
| ## | | ----- |
| ## | | 23 - 25 3 12829 0.03 124.52 |
| ## | | ----- |
| ## | | Total 10303 - 100.00 - |
| ## | | ----- |

3.5.1.3 CLM_AMT - Valor de cobertura solicitado

3.5.1.3.1 Estatísticas básicas do R

```
mean(Claim.Data$CLM_AMT)      # media

## [1] 1511.119

median(Claim.Data$CLM_AMT)    # mediana

## [1] 0

min(Claim.Data$CLM_AMT)      # minimo

## [1] 0

max(Claim.Data$CLM_AMT)      # maximo

## [1] 123247.1

var(Claim.Data$CLM_AMT)      # variancia

## [1] 22326069

sd(Claim.Data$CLM_AMT)       # desvio padrao

## [1] 4725.047

IQR(Claim.Data$CLM_AMT)      # distancia interquartilica

## [1] 1144.427

summary(Claim.Data$CLM_AMT)   # Min, Q1, Q2, media, Q3, Max

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##         0         0         0   1511   1144 123247

quantile(Claim.Data$CLM_AMT)   # Min, Q1, Q2, Q3, Max

##          0%          25%          50%          75%          100%
##       0.000       0.000       0.000      1144.427 123247.121

quantile(Claim.Data$CLM_AMT, type = 7, probs = c(.01, .05, .10, .90, .95, .99)) # percentis

##          1%          5%          10%          90%          95%          99%
##         0.00         0.00         0.00      4891.60      6406.80 19968.13
```

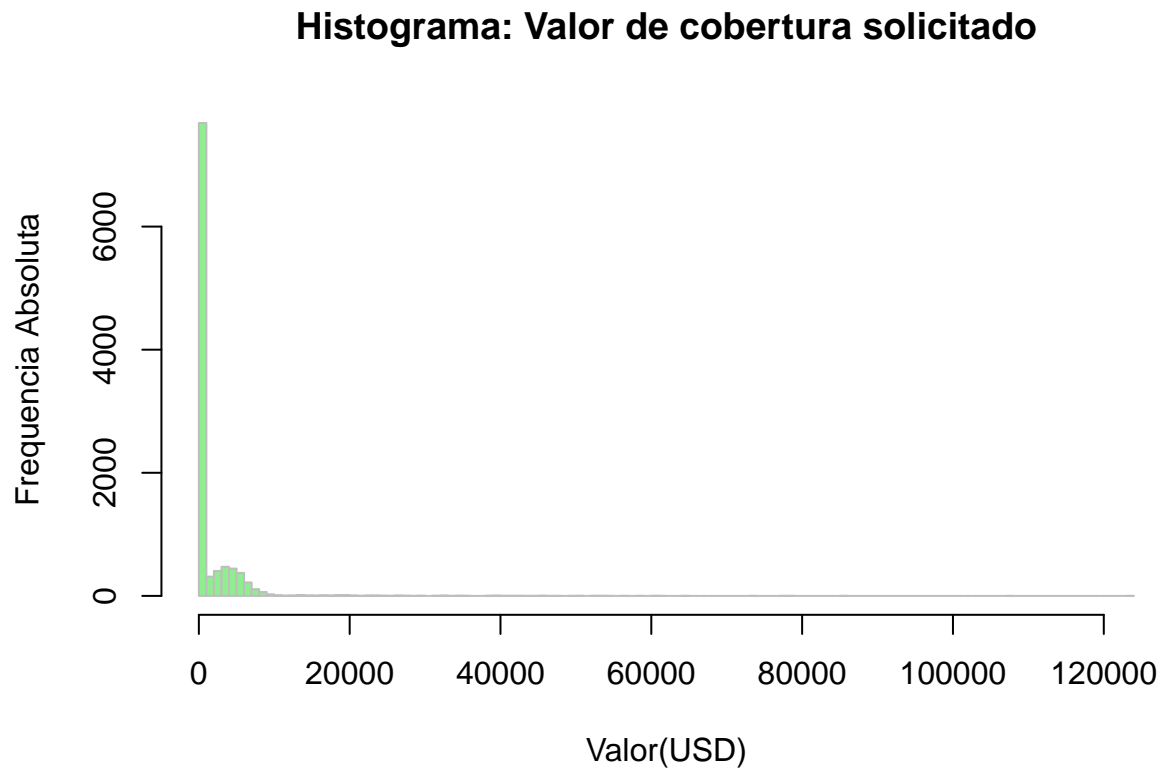
3.5.1.3.2 Resumo da biblioteca Hmisc

```
describe(Claim.Data$CLM_AMT)
```

```
##      vars      n    mean      sd median trimmed mad min      max    range skew
## X1      1 10303 1511.12 4725.05      0  607.79   0   0 123247.1 123247.1  9.29
##      kurtosis    se
## X1      136.39 46.55
```

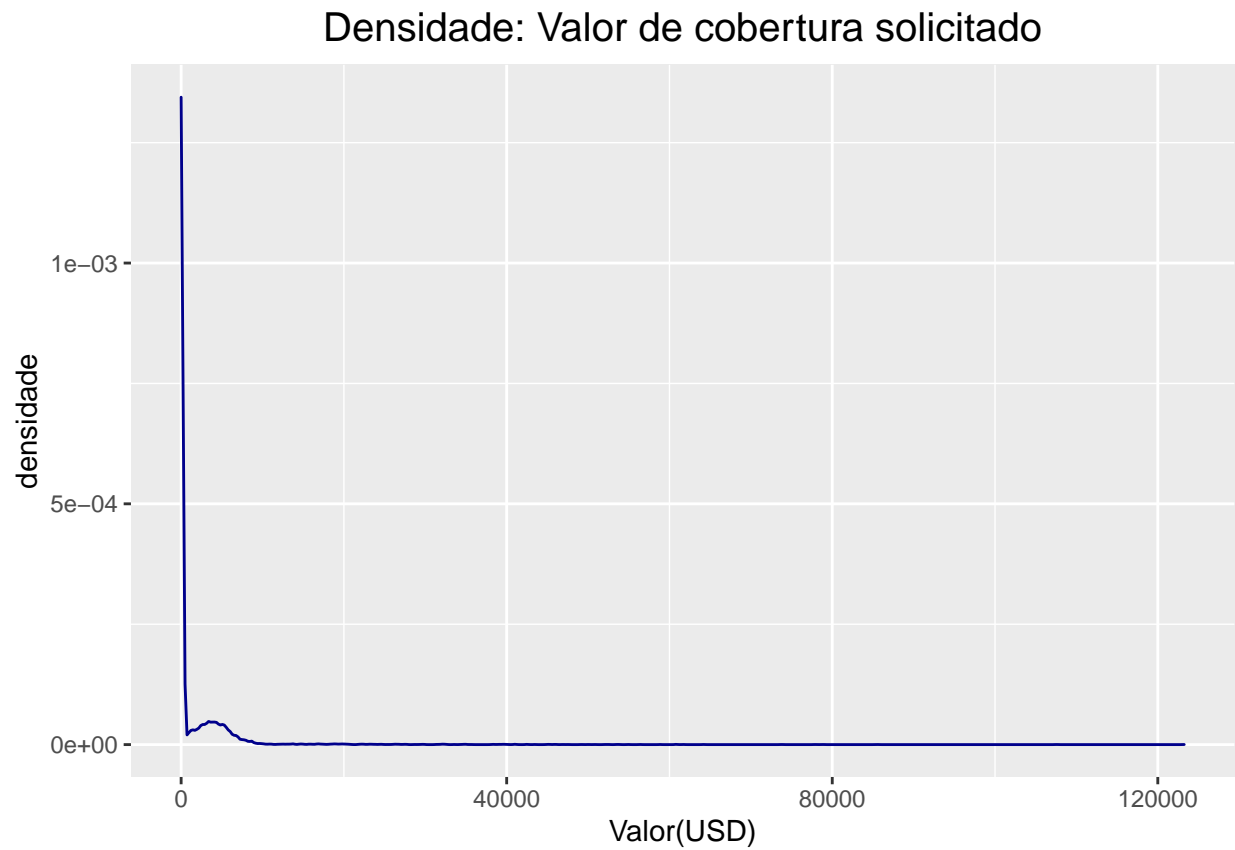
3.5.1.3.3 Histograma

```
hist(Claim.Data$CLM_AMT, breaks = "scott",
     col = "lightgreen", border = "grey",
     main = "Histograma: Valor de cobertura solicitado",
     xlab = "Valor(USD)", ylab = "Frequencia Absoluta",
)
```



3.5.1.3.4 Grafico de densidade por kernel

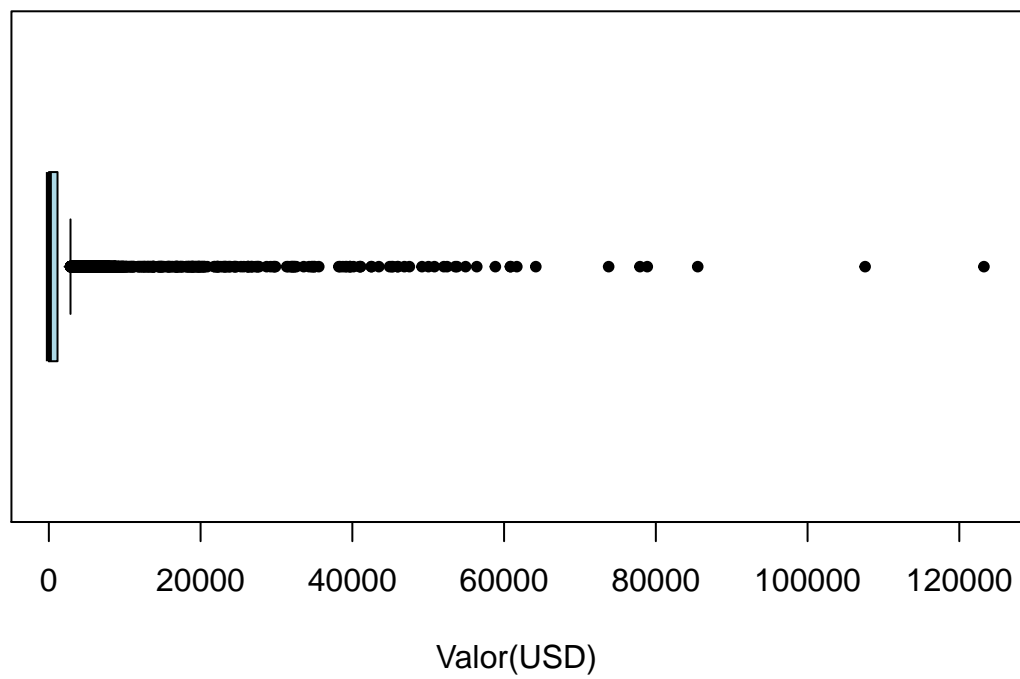
```
graf <- ggplot(data = Claim.Data, mapping = aes(x = CLM_AMT)) +  
  geom_density(mapping = aes(x = CLM_AMT),  
               bw = "nrd",  
               color = "darkblue") +  
  ggtitle("Densidade: Valor de cobertura solicitado") +  
  xlab("Valor(USD)") + ylab("densidade") +  
  theme(plot.title = element_text(hjust = 0.5, size=15) )  
graf
```



3.5.1.3.5 Boxplot

```
boxplot(Claim.Data$CLM_AMT, horizontal = T,  
        col = "lightblue", pch = 20,  
        main = "Boxplot: Valor de cobertura solicitado",  
        xlab = "Valor(USD)")
```

Boxplot: Valor de cobertura solicitado



3.5.1.3.6 Resumo Tabular

```
ds_freq_table(Claim.Data, CLM_AMT, bins = 20)
```

| Variable: CLM_AMT | | | | | | | |
|-------------------|---|----------|--|-----------|---------------|---------|-------------|
| Bins | | | | Frequency | Cum Frequency | Percent | Cum Percent |
| 0 | - | 6162.4 | | 9745 | 9745 | 94.58 | 94.58 |
| 6162.4 | - | 12324.7 | | 381 | 10126 | 3.7 | 98.28 |
| 12324.7 | - | 18487.1 | | 55 | 10181 | 0.53 | 98.82 |
| 18487.1 | - | 24649.4 | | 45 | 10226 | 0.44 | 99.25 |
| 24649.4 | - | 30811.8 | | 19 | 10245 | 0.18 | 99.44 |
| 30811.8 | - | 36974.1 | | 16 | 10261 | 0.16 | 99.59 |
| 36974.1 | - | 43136.5 | | 14 | 10275 | 0.14 | 99.73 |
| 43136.5 | - | 49298.8 | | 9 | 10284 | 0.09 | 99.82 |
| 49298.8 | - | 55461.2 | | 7 | 10291 | 0.07 | 99.88 |
| 55461.2 | - | 61623.6 | | 4 | 10295 | 0.04 | 99.92 |
| 61623.6 | - | 67785.9 | | 2 | 10297 | 0.02 | 99.94 |
| 67785.9 | - | 73948.3 | | 1 | 10298 | 0.01 | 99.95 |
| 73948.3 | - | 80110.6 | | 2 | 10300 | 0.02 | 99.97 |
| 80110.6 | - | 86273 | | 1 | 10301 | 0.01 | 99.98 |
| 86273 | - | 92435.3 | | 0 | 10301 | 0 | 99.98 |
| 92435.3 | - | 98597.7 | | 0 | 10301 | 0 | 99.98 |
| 98597.7 | - | 104760.1 | | 0 | 10301 | 0 | 99.98 |
| 104760.1 | - | 110922.4 | | 1 | 10302 | 0.01 | 99.99 |
| 110922.4 | - | 117084.8 | | 0 | 10302 | 0 | 99.99 |
| 117084.8 | - | 123247.1 | | 1 | 10303 | 0.01 | 100 |
| Total | | | | 10303 | - | 100.00 | - |

3.5.1.4 AGE - Idade em anos

3.5.1.4.1 Estatísticas básicas do R

```
mean(Claim.Data$AGE, na.rm=TRUE)      # media

## [1] 44.83664

median(Claim.Data$AGE, na.rm=TRUE)    # mediana

## [1] 45

min(Claim.Data$AGE, na.rm=TRUE)       # minimo

## [1] 16

max(Claim.Data$AGE, na.rm=TRUE)       # maximo

## [1] 81

var(Claim.Data$AGE, na.rm=TRUE)       # variancia

## [1] 74.06967

sd(Claim.Data$AGE, na.rm=TRUE)        # desvio padrao

## [1] 8.606374

IQR(Claim.Data$AGE, na.rm=TRUE)       # distancia interquartilica

## [1] 12

summary(Claim.Data$AGE, na.rm=TRUE)    # Min, Q1, Q2, media, Q3, Max

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
##    16.00   39.00   45.00   44.84   51.00   81.00         7

quantile(Claim.Data$AGE, na.rm=TRUE)   # Min, Q1, Q2, Q3, Max

##      0%   25%   50%   75%  100%
##      16    39    45    51    81

quantile(Claim.Data$AGE, na.rm=TRUE, type = 7, probs = c(.01, .05, .10, .90, .95, .99)) # percentis

##      1%   5%  10%  90%  95%  99%
##      25   30   34   56   59   64
```

Adicionado parâmetro `na.rm=TRUE` para ignorar os valores não definidos.

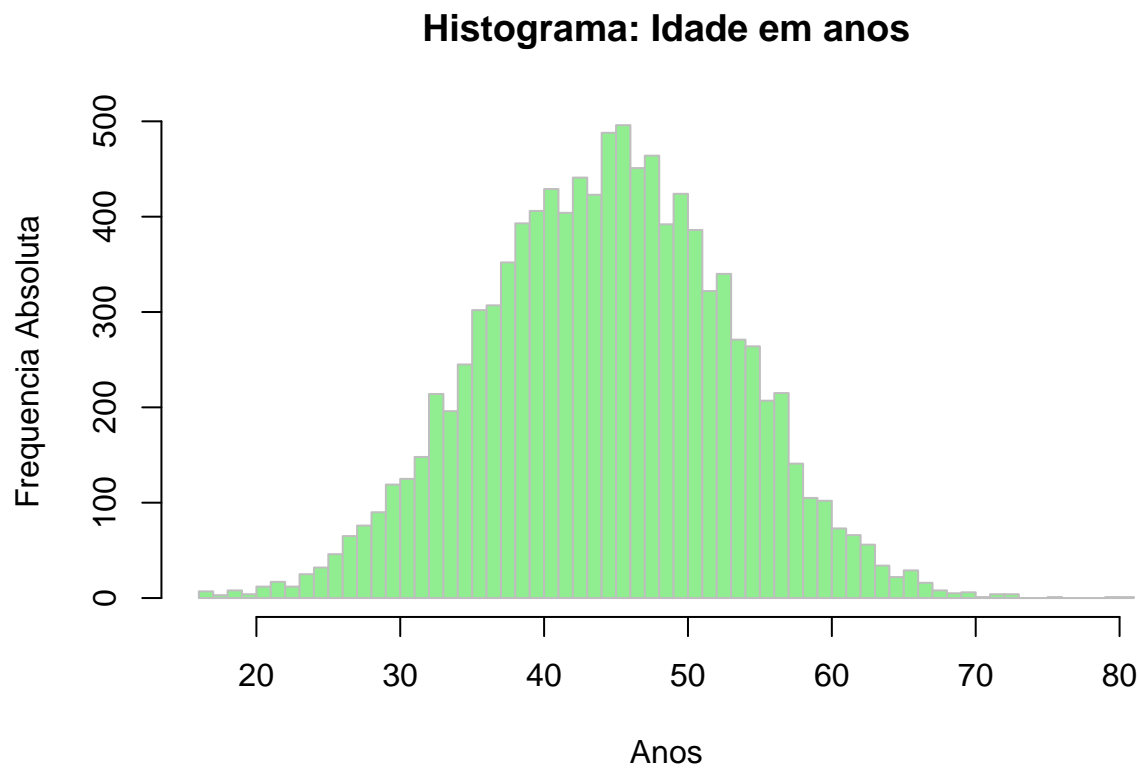
3.5.1.4.2 Resumo da biblioteca Hmisc

```
describe(Claim.Data$AGE)
```

```
##      vars      n  mean   sd median trimmed mad min max range  skew kurtosis   se
## X1      1 10296 44.84 8.61    45   44.88 8.9  16  81   65 -0.03   -0.08 0.08
```

3.5.1.4.3 Histograma

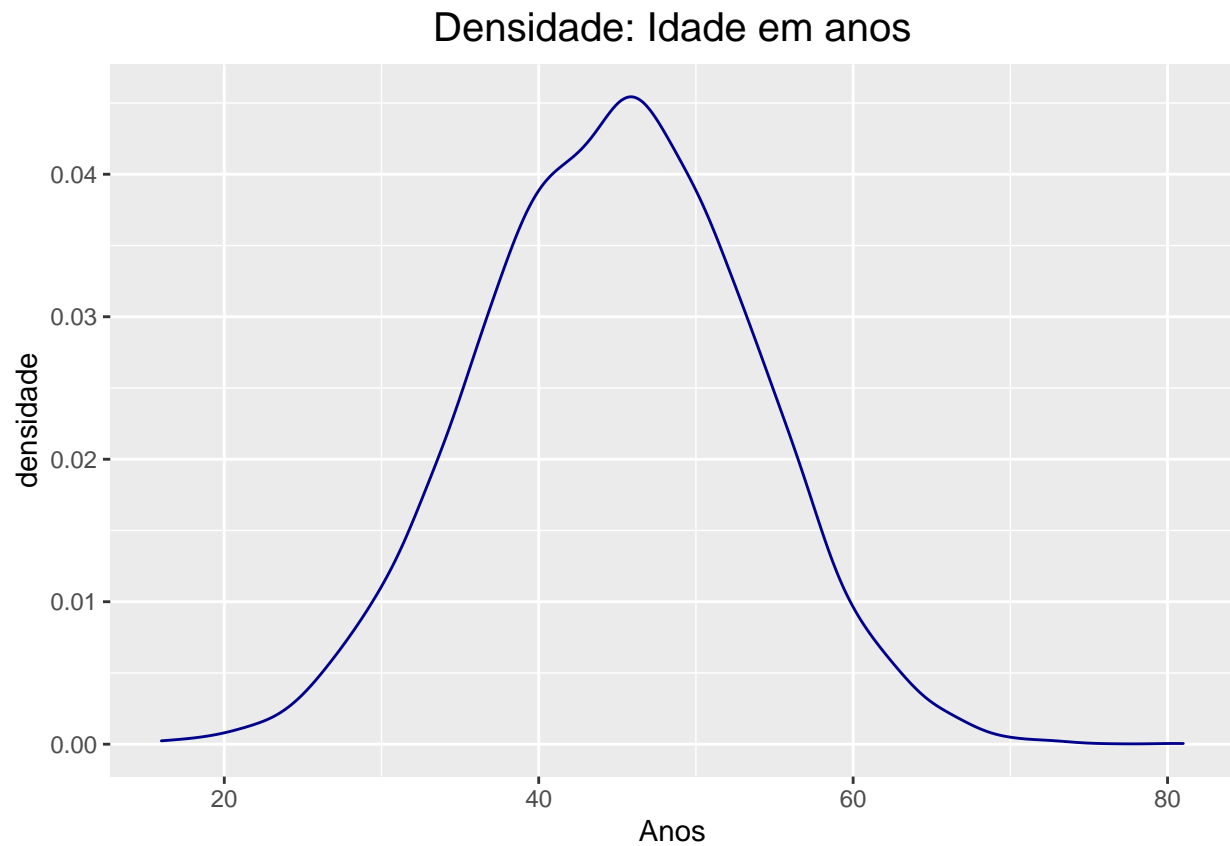
```
hist(Claim.Data$AGE, breaks = "fd",
     col = "lightgreen", border = "grey",
     main = "Histograma: Idade em anos",
     xlab = "Anos", ylab = "Frequencia Absoluta"
)
```



3.5.1.4.4 Grafico de densidade por kernel

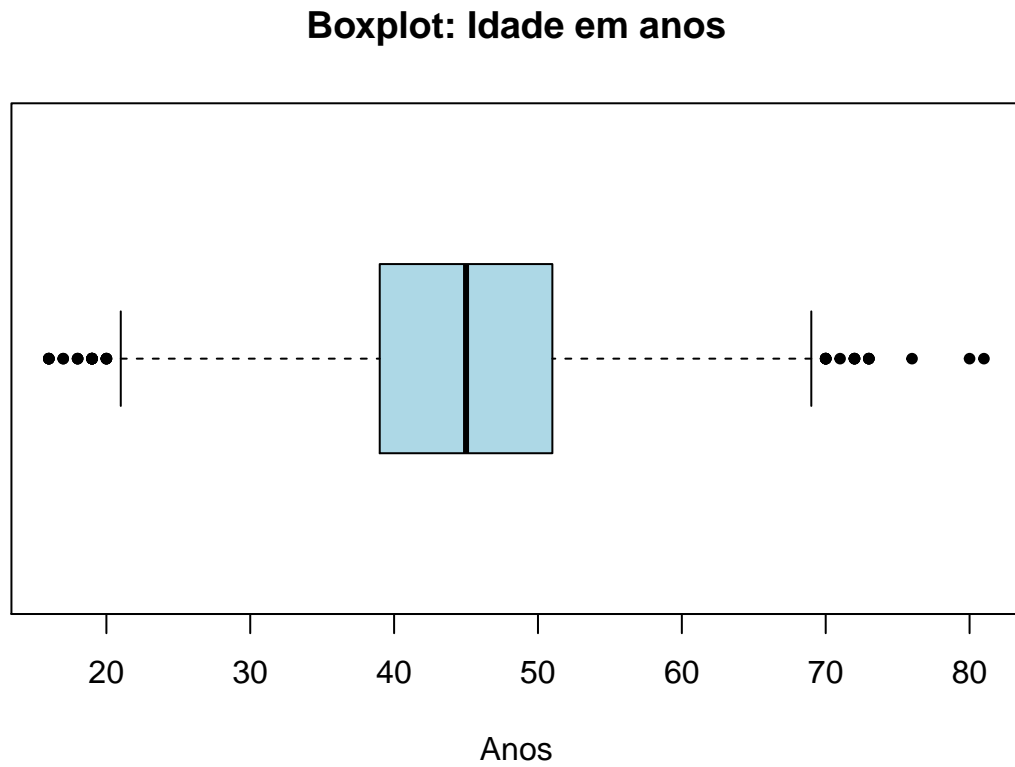
```
graf <- ggplot(data = Claim.Data, mapping = aes(x = AGE)) +  
  geom_density(mapping = aes(x = AGE),  
              bw = "nrd",  
              color = "darkblue") +  
  ggtitle("Densidade: Idade em anos") +  
  xlab("Anos") + ylab("densidade") +  
  theme(plot.title = element_text(hjust = 0.5, size=15) )  
graf
```

Warning: Removed 7 rows containing non-finite values (stat_density).



3.5.1.4.5 Boxplot

```
boxplot(Claim.Data$AGE, horizontal = T,  
        col = "lightblue", pch = 20,  
        main = "Boxplot: Idade em anos",  
        xlab = "Anos")
```



3.5.1.4.6 Resumo Tabular

```
ds_freq_table(Claim.Data, AGE, bins = 25)
```

```
##                               Variable: AGE
## |-----|
## |  Bins      | Frequency | Cum Frequency |  Percent  | Cum Percent |
## |-----|
## | 16 - 18.6 |    10     |      10       |    0.1     |    0.1       |
## |-----|
## | 18.6 - 21.2 |    24     |      34       |    0.23    |    0.33       |
## |-----|
## | 21.2 - 23.8 |    29     |      63       |    0.28    |    0.61       |
## |-----|
## | 23.8 - 26.4 |   103     |     166       |    1       |    1.61       |
## |-----|
## | 26.4 - 29   |   231     |     397       |    2.24    |    3.86       |
```

| | | | | | | |
|----|-------------|-------|-------|------------|-------|--|
| ## | ----- | | | | | |
| ## | 29 - 31.6 | 244 | 641 | 2.37 | 6.23 | |
| ## | ----- | | | | | |
| ## | 31.6 - 34.2 | 558 | 1199 | 5.42 | 11.65 | |
| ## | ----- | | | | | |
| ## | 34.2 - 36.8 | 547 | 1746 | 5.31 | 16.96 | |
| ## | ----- | | | | | |
| ## | 36.8 - 39.4 | 1052 | 2798 | 10.22 | 27.18 | |
| ## | ----- | | | | | |
| ## | 39.4 - 42 | 1239 | 4037 | 12.03 | 39.21 | |
| ## | ----- | | | | | |
| ## | 42 - 44.6 | 864 | 4901 | 8.39 | 47.6 | |
| ## | ----- | | | | | |
| ## | 44.6 - 47.2 | 1435 | 6336 | 13.94 | 61.54 | |
| ## | ----- | | | | | |
| ## | 47.2 - 49.8 | 856 | 7192 | 8.31 | 69.85 | |
| ## | ----- | | | | | |
| ## | 49.8 - 52.4 | 1132 | 8324 | 10.99 | 80.85 | |
| ## | ----- | | | | | |
| ## | 52.4 - 55 | 875 | 9199 | 8.5 | 89.35 | |
| ## | ----- | | | | | |
| ## | 55 - 57.6 | 422 | 9621 | 4.1 | 93.44 | |
| ## | ----- | | | | | |
| ## | 57.6 - 60.2 | 348 | 9969 | 3.38 | 96.82 | |
| ## | ----- | | | | | |
| ## | 60.2 - 62.8 | 139 | 10108 | 1.35 | 98.17 | |
| ## | ----- | | | | | |
| ## | 62.8 - 65.4 | 112 | 10220 | 1.09 | 99.26 | |
| ## | ----- | | | | | |
| ## | 65.4 - 68 | 53 | 10273 | 0.51 | 99.78 | |
| ## | ----- | | | | | |
| ## | 68 - 70.6 | 11 | 10284 | 0.11 | 99.88 | |
| ## | ----- | | | | | |
| ## | 70.6 - 73.2 | 9 | 10293 | 0.09 | 99.97 | |
| ## | ----- | | | | | |
| ## | 73.2 - 75.8 | 0 | 10293 | 0 | 99.97 | |
| ## | ----- | | | | | |
| ## | 75.8 - 78.4 | 1 | 10294 | 0.01 | 99.98 | |
| ## | ----- | | | | | |
| ## | 78.4 - 81 | 2 | 10296 | 0.02 | 100 | |
| ## | ----- | | | | | |
| ## | Missing | 7 | - | 0.06794138 | - | |
| ## | ----- | | | | | |
| ## | Total | 10303 | - | 100.00 | - | |
| ## | ----- | | | | | |

3.5.1.5 YOJ - Anos de trabalho

3.5.1.5.1 Estatísticas básicas do R

```
mean(Claim.Data$YOJ, na.rm=TRUE)      # media

## [1] 10.47391

median(Claim.Data$YOJ, na.rm=TRUE)    # mediana

## [1] 11

min(Claim.Data$YOJ, na.rm=TRUE)       # minimo

## [1] 0

max(Claim.Data$YOJ, na.rm=TRUE)       # maximo

## [1] 23

var(Claim.Data$YOJ, na.rm=TRUE)       # variancia

## [1] 16.88191

sd(Claim.Data$YOJ, na.rm=TRUE)        # desvio padrao

## [1] 4.10876

IQR(Claim.Data$YOJ, na.rm=TRUE)       # distancia interquartilica

## [1] 4

summary(Claim.Data$YOJ, na.rm=TRUE)    # Min, Q1, Q2, media, Q3, Max

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      0.00   9.00   11.00   10.47   13.00   23.00     548

quantile(Claim.Data$YOJ, na.rm=TRUE)   # Min, Q1, Q2, Q3, Max

##      0%  25%  50%  75% 100%
##       0    9   11   13   23

quantile(Claim.Data$YOJ, na.rm=TRUE, type = 7, probs = c(.01, .05, .10, .90, .95, .99)) # percentis

##      1%  5% 10% 90% 95% 99%
##       0    5   15   15   17
```

Adicionado parâmetro `na.rm=TRUE` para ignorar os valores não definidos.

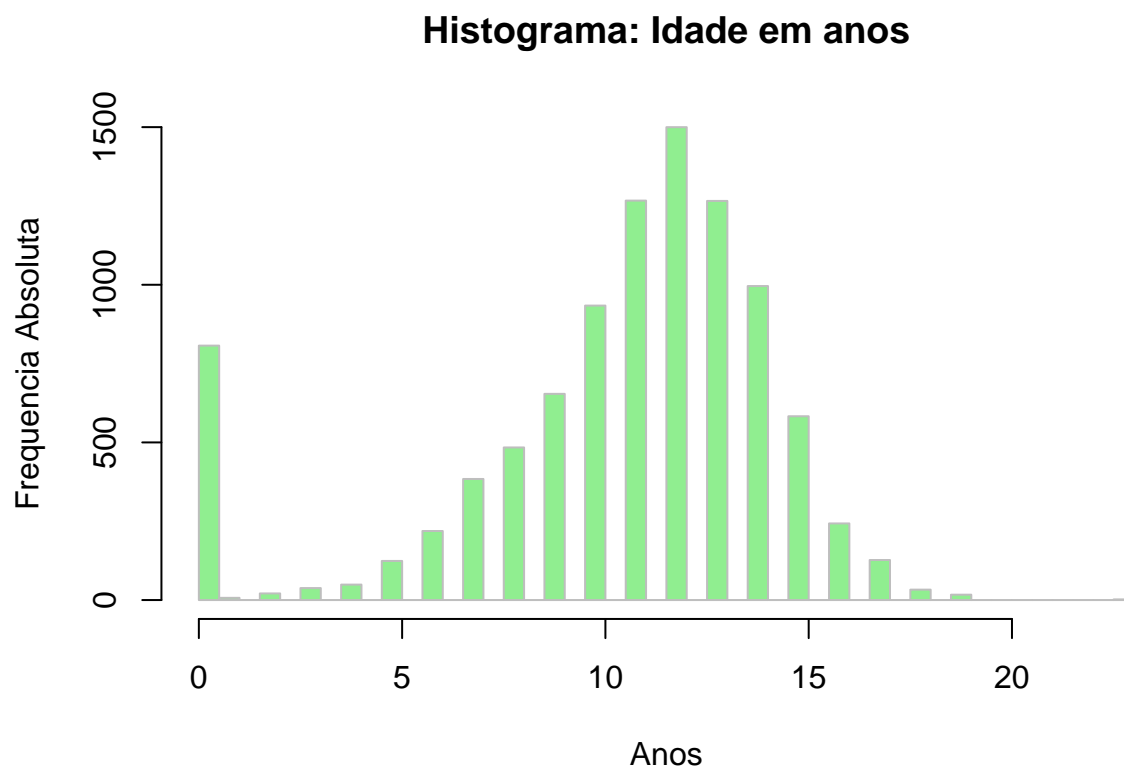
3.5.1.5.2 Resumo da biblioteca Hmisc

```
describe(Claim.Data$Y0J)
```

```
##      vars      n  mean   sd median trimmed  mad min max range skew kurtosis   se
## X1      1  9755 10.47 4.11     11   11.05 2.97   0  23   23 -1.2    1.14 0.04
```

3.5.1.5.3 Histograma

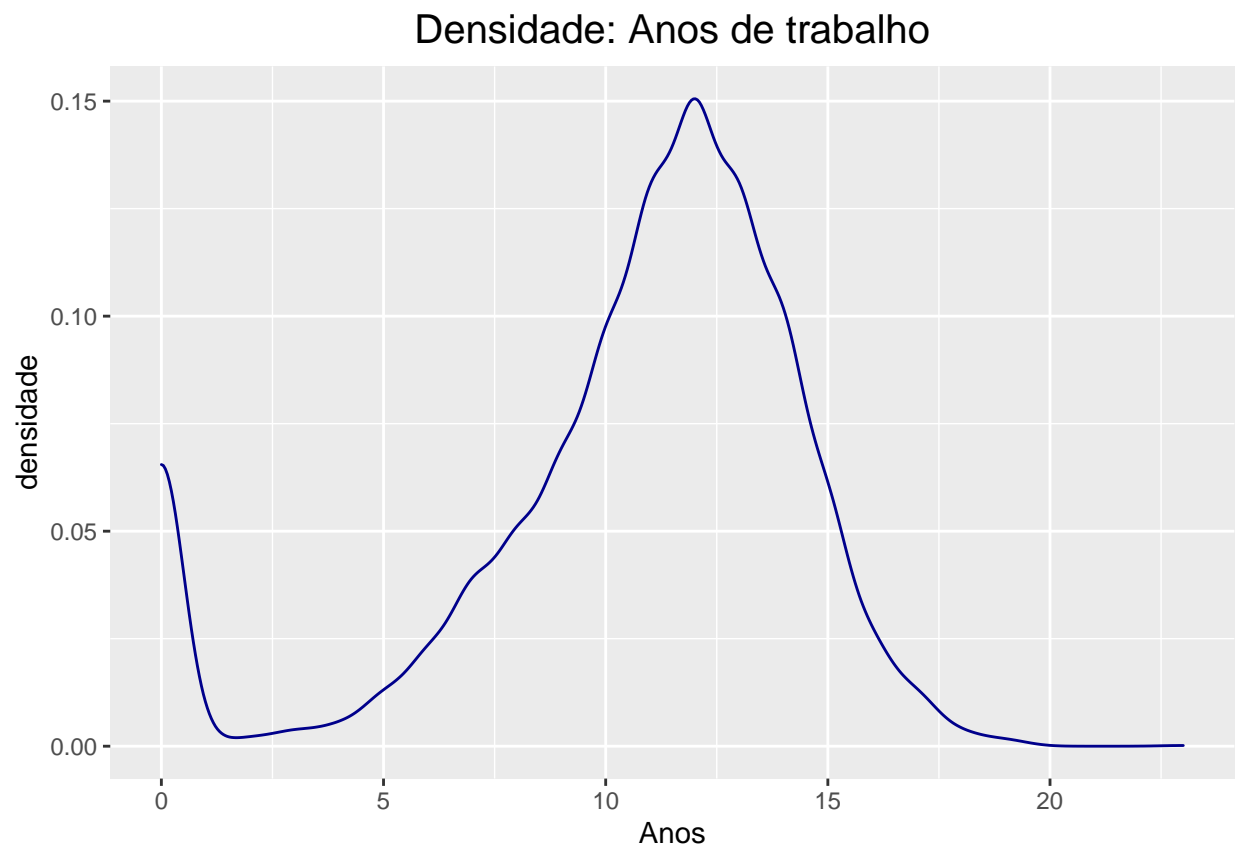
```
hist(Claim.Data$Y0J, breaks = "fd",
     col = "lightgreen", border = "grey",
     main = "Histograma: Idade em anos",
     xlab = "Anos", ylab = "Frequencia Absoluta"
)
```



3.5.1.5.4 Grafico de densidade por kernel

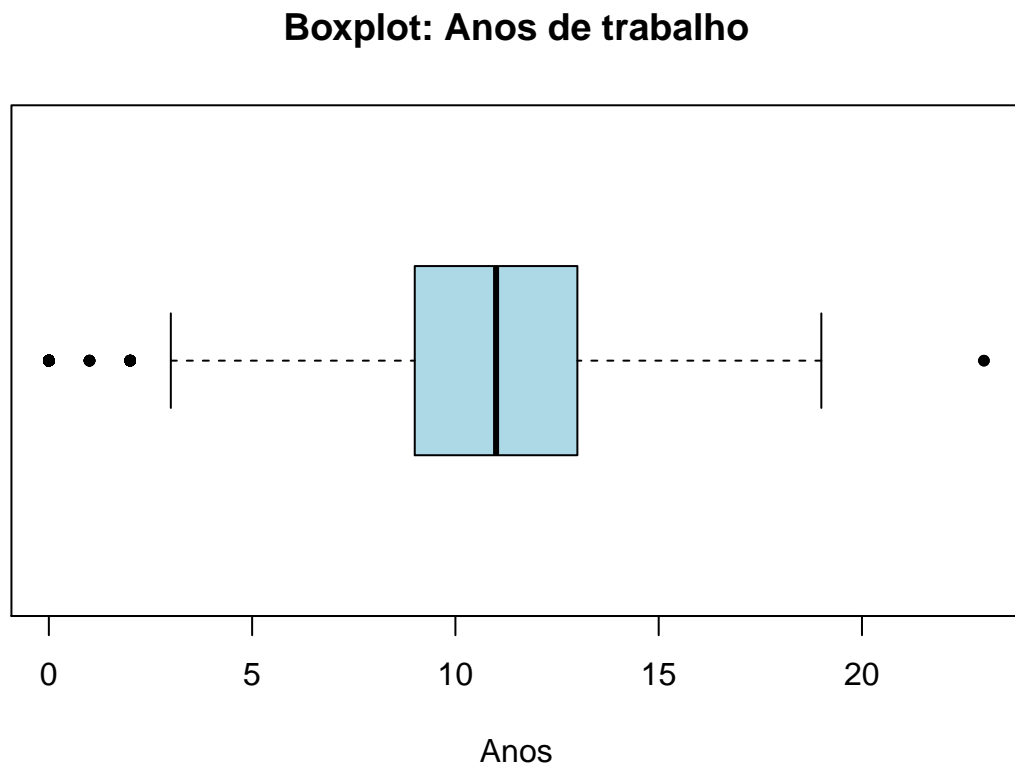
```
graf <- ggplot(data = Claim.Data, mapping = aes(x = Y0J)) +  
  geom_density(mapping = aes(x = Y0J),  
              bw = "nrd",  
              color = "darkblue") +  
  ggtitle("Densidade: Anos de trabalho") +  
  xlab("Anos") + ylab("densidade") +  
  theme(plot.title = element_text(hjust = 0.5, size=15) )  
graf
```

```
## Warning: Removed 548 rows containing non-finite values (stat_density).
```



3.5.1.5.5 Boxplot

```
boxplot(Claim.Data$Y0J, horizontal = T,  
        col = "lightblue", pch = 20,  
        main = "Boxplot: Anos de trabalho",  
        xlab = "Anos")
```



3.5.1.5.6 Resumo Tabular

```
ds_freq_table(Claim.Data, Y0J, bins = 23)
```

| Variable: Y0J | | | | | |
|---------------|-----------|---------------|---------|-------------|--|
| Bins | Frequency | Cum Frequency | Percent | Cum Percent | |
| 0 - 1 | 814 | 814 | 8.34 | 8.34 | |
| 1 - 2 | 28 | 842 | 0.29 | 8.63 | |
| 2 - 3 | 59 | 901 | 0.6 | 9.24 | |
| 3 - 4 | 87 | 988 | 0.89 | 10.13 | |
| 4 - 5 | 173 | 1161 | 1.77 | 11.9 | |
| 5 - 6 | 343 | 1504 | 3.52 | 15.42 | |
| 6 - 7 | 603 | 2107 | 6.18 | 21.6 | |
| 7 - 8 | 868 | 2975 | 8.9 | 30.5 | |
| 8 - 9 | 1138 | 4113 | 11.67 | 42.16 | |
| 9 - 10 | 1588 | 5701 | 16.28 | 58.44 | |
| 10 - 11 | 2201 | 7902 | 22.56 | 81 | |
| 11 - 12 | 2767 | 10669 | 28.36 | 109.37 | |
| 12 - 13 | 2766 | 13435 | 28.35 | 137.72 | |
| 13 - 14 | 2262 | 15697 | 23.19 | 160.91 | |
| 14 - 15 | 1579 | 17276 | 16.19 | 177.1 | |
| 15 - 16 | 826 | 18102 | 8.47 | 185.57 | |
| 16 - 17 | 370 | 18472 | 3.79 | 189.36 | |
| 17 - 18 | 160 | 18632 | 1.64 | 191 | |
| 18 - 19 | 50 | 18682 | 0.51 | 191.51 | |
| 19 - 20 | 17 | 18699 | 0.17 | 191.69 | |
| 20 - 21 | 0 | 18699 | 0 | 191.69 | |
| 21 - 22 | 0 | 18699 | 0 | 191.69 | |
| 22 - 23 | 2 | 18701 | 0.02 | 191.71 | |

```
## | Missing |    548    |      -      |  5.318839  |      -      |
## |-----|
## | Total   |   10303   |      -      |   100.00   |      -      |
## |-----|
```

3.5.2 Variáveis discretas

3.5.2.1 NPOLICY - Número de apólices

3.5.2.1.1 Estatísticas básicas do R

```
mean(Claim.Data$NPOLICY)      # media
```

```
## [1] 1.695429
```

```
median(Claim.Data$NPOLICY)    # mediana
```

```
## [1] 1
```

```
min(Claim.Data$NPOLICY)      # minimo
```

```
## [1] 1
```

```
max(Claim.Data$NPOLICY)      # maximo
```

```
## [1] 9
```

```
var(Claim.Data$NPOLICY)      # variancia
```

```
## [1] 0.8746122
```

```
sd(Claim.Data$NPOLICY)       # desvio padrao
```

```
## [1] 0.935207
```

```
IQR(Claim.Data$NPOLICY)      # distancia interquartilica
```

```
## [1] 1
```

```
summary(Claim.Data$NPOLICY)   # Min, Q1, Q2, media, Q3, Max
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      1.000   1.000   1.000   1.695   2.000   9.000
```

```
quantile(Claim.Data$NPOLICY) # Min, Q1, Q2, Q3, Max
```

```
## 0% 25% 50% 75% 100%
## 1 1 1 2 9
```

```
quantile(Claim.Data$NPOLICY, type = 7, probs = c(.01, .05, .10, .90, .95, .99)) # percentis
```

```
## 1% 5% 10% 90% 95% 99%
## 1 1 1 3 3 5
```

3.5.2.1.2 Resumo da biblioteca Hmisc

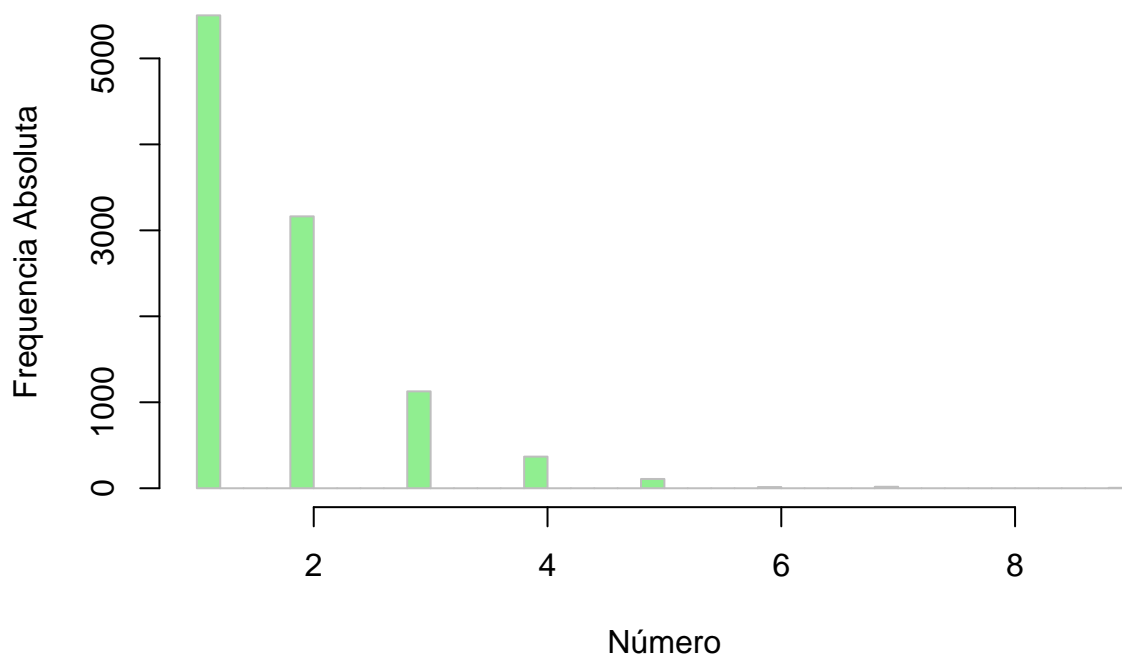
```
describe(Claim.Data$NPOLICY)
```

```
## vars n mean sd median trimmed mad min max range skew kurtosis se
## X1 1 10303 1.7 0.94 1 1.53 0 1 9 8 1.75 4.66 0.01
```

3.5.2.1.3 Histograma

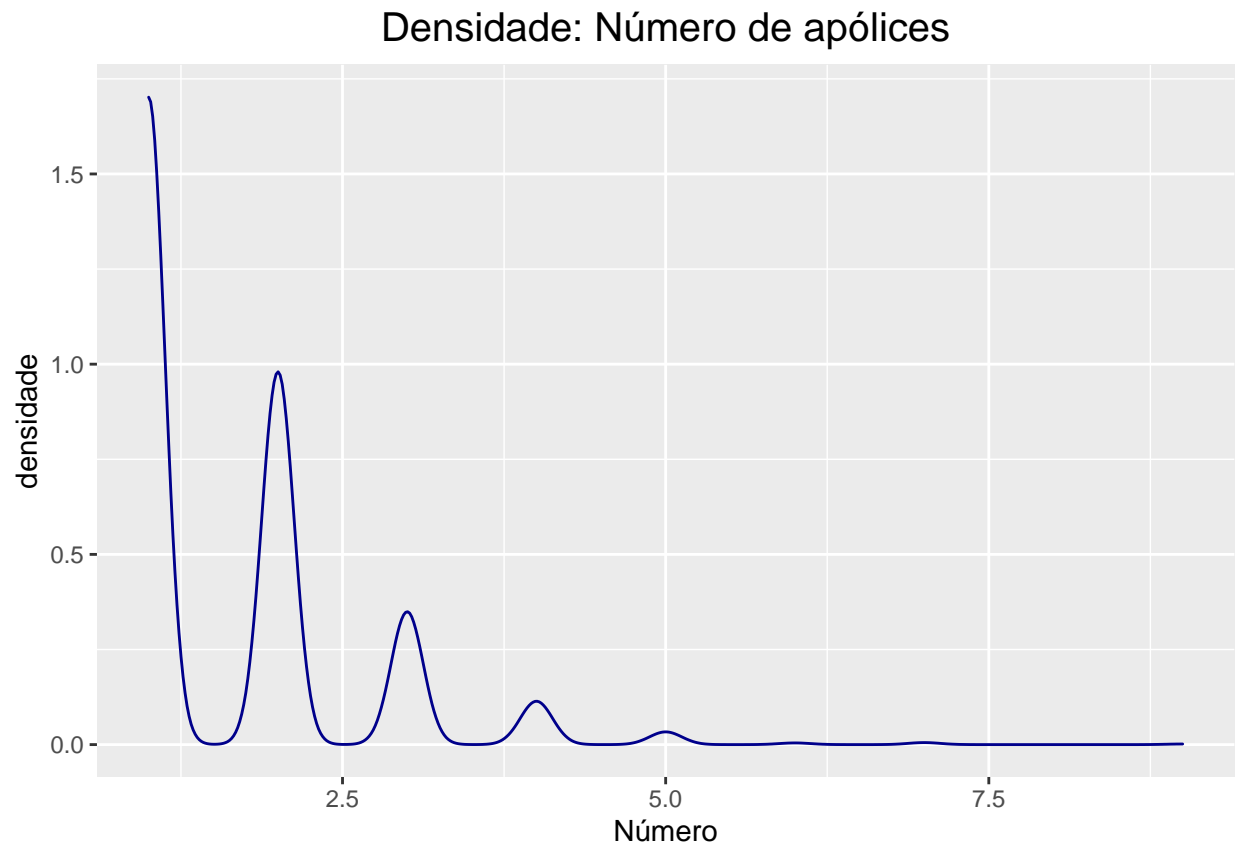
```
hist(Claim.Data$NPOLICY, breaks = "scott",
     col = "lightgreen", border = "grey",
     main = "Histograma: Número de apólices",
     xlab = "Número", ylab = "Frequencia Absoluta",
     )
```

Histograma: Número de apólices



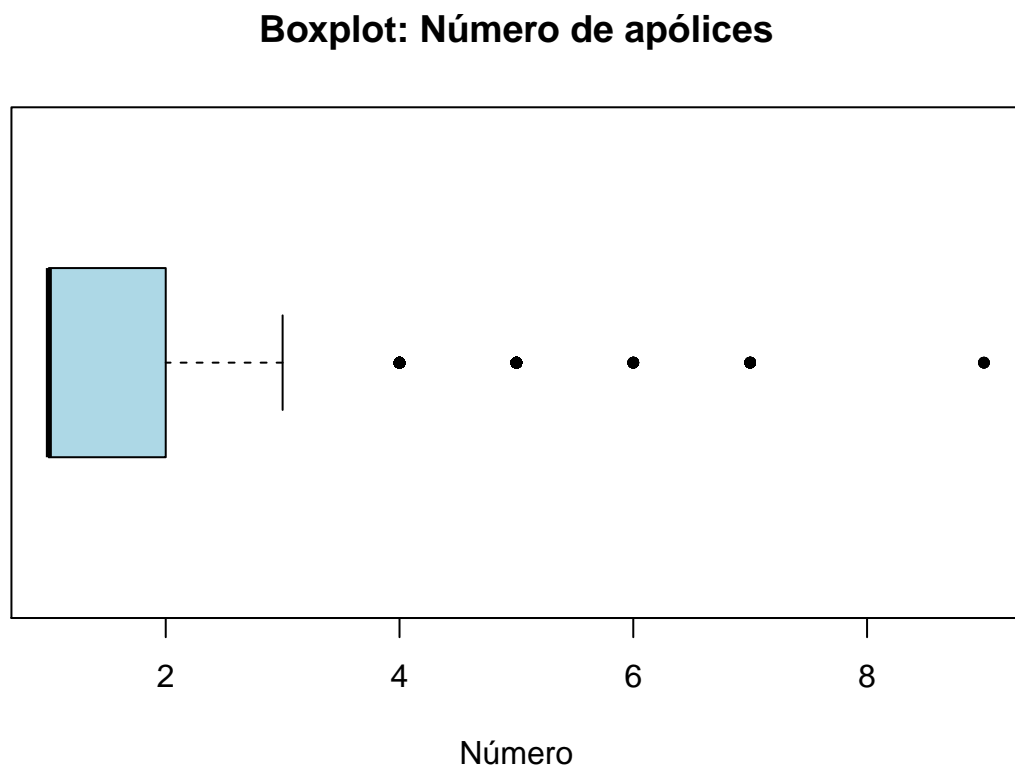
3.5.2.1.4 Grafico de densidade por kernel

```
graf <- ggplot(data = Claim.Data, mapping = aes(x = NPOLICY)) +  
  geom_density(mapping = aes(x = NPOLICY),  
               bw = "nrd",  
               color = "darkblue") +  
  ggtitle("Densidade: Número de apólices") +  
  xlab("Número") + ylab("densidade") +  
  theme(plot.title = element_text(hjust = 0.5, size=15) )  
graf
```



3.5.2.1.5 Boxplot

```
boxplot(Claim.Data$NPOLICY, horizontal = T,  
        col = "lightblue", pch = 20,  
        main = "Boxplot: Número de apólices",  
        xlab = "Número")
```



3.5.2.1.6 Resumo Tabular

```
ds_freq_table(Claim.Data, NPOLICY, bins = 8)
```

| Variable: NPOLICY | | | | | |
|-------------------|-----------|---------------|---------|-------------|--|
| Bins | Frequency | Cum Frequency | Percent | Cum Percent | |
| 1 - 2 | 8664 | 8664 | 84.09 | 84.09 | |
| 2 - 3 | 4290 | 12954 | 41.64 | 125.73 | |
| 3 - 4 | 1495 | 14449 | 14.51 | 140.24 | |
| 4 - 5 | 476 | 14925 | 4.62 | 144.86 | |
| 5 - 6 | 122 | 15047 | 1.18 | 146.04 | |
| 6 - 7 | 31 | 15078 | 0.3 | 146.35 | |
| 7 - 8 | 17 | 15095 | 0.17 | 146.51 | |
| 8 - 9 | 5 | 15100 | 0.05 | 146.56 | |
| Total | 10303 | - | 100.00 | - | |

3.5.3 Variáveis nominais

3.5.3.1 MAX_EDUC - Máximo nível educacional

3.5.3.1.1 Estatísticas básicas do R

```
median(as.numeric(Claim.Data$MAX_EDUC))           # Nível de educação Mediana
```

```
## [1] 3
```

```
quantile(as.numeric(Claim.Data$MAX_EDUC), type = 2) # Quartis
```

```
##    0%   25%   50%   75%  100%  
##     1     2     3     4     5
```

```
IQR(as.numeric(Claim.Data$MAX_EDUC), type = 2)     # Distancia interquartilica
```

```
## [1] 2
```

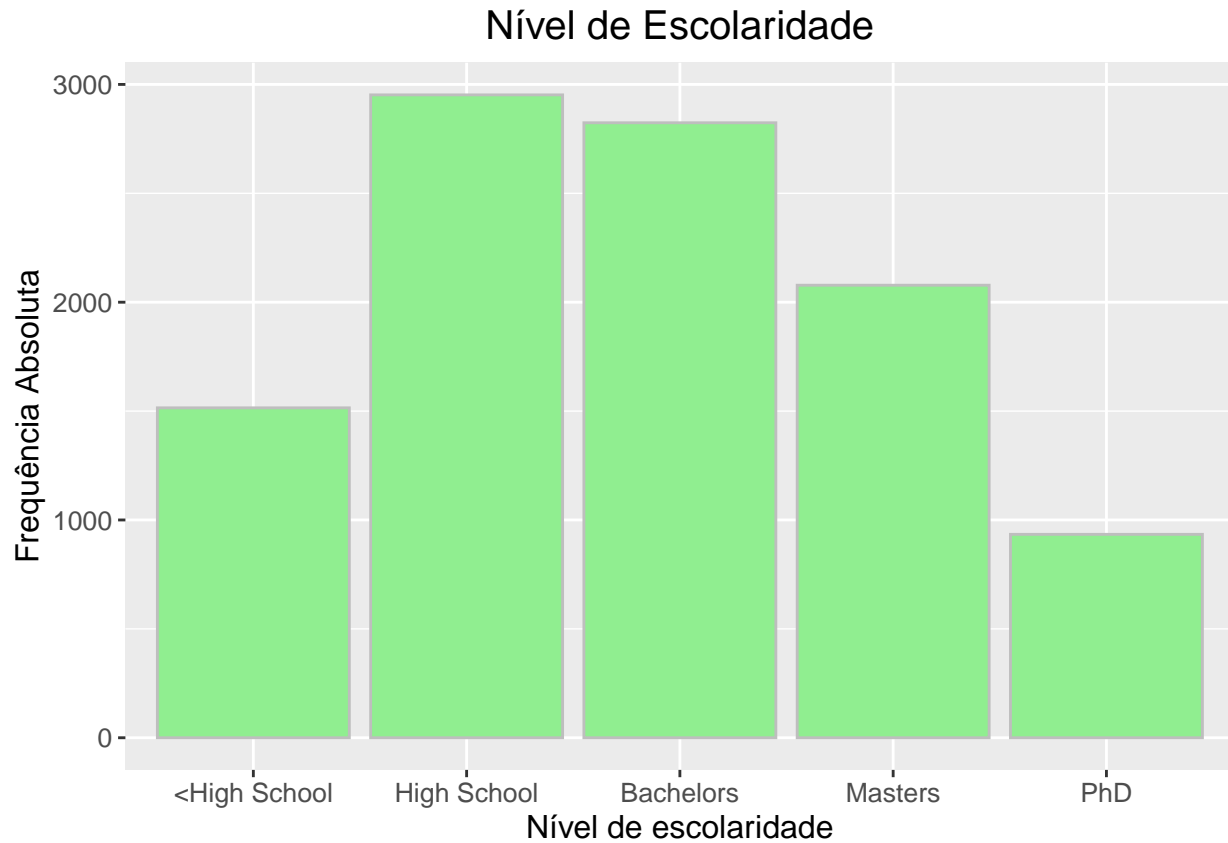
3.5.3.1.2 Resumo tabular

```
tabela <- freq(Claim.Data$MAX_EDUC, cum = TRUE, total = TRUE, valid = FALSE)  
tabela
```

```
## Frequencies  
## Claim.Data$MAX_EDUC  
## Type: Ordered Factor  
##  
##           Freq  % Valid  % Valid Cum.  % Total  % Total Cum.  
## -----  
##    <High School  1515    14.70      14.70    14.70    14.70  
##    High School   2952    28.65      43.36    28.65    43.36  
##    Bachelors     2824    27.41      70.77    27.41    70.77  
##    Masters       2078    20.17      90.93    20.17    90.93  
##    PhD           934     9.07     100.00     9.07   100.00  
##    <NA>           0      0.00     100.00     0.00   100.00  
##    Total       10303   100.00     100.00   100.00   100.00
```

3.5.3.1.3 Resumo gráfico

```
ggplot(Claim.Data,  
  aes(x = MAX_EDUC )) +  
  geom_bar(color = "grey", fill = "lightgreen") +  
  ggtitle("Nível de Escolaridade") +  
  xlab("Nível de escolaridade") +  
  ylab("Frequência Absoluta") +  
  theme(legend.position="none",  
    plot.title = element_text(hjust = 0.5, size = 15),  
    axis.title = element_text(size = 12),  
    axis.text = element_text(size = 10)  
  )
```



3.5.3.1.4 Tabela de frequências

```
dados.freq <- data.frame(
  name = rownames(table(Claim.Data$MAX_EDUC)),
  value = as.vector(table(Claim.Data$MAX_EDUC))
)
dados.freq
```

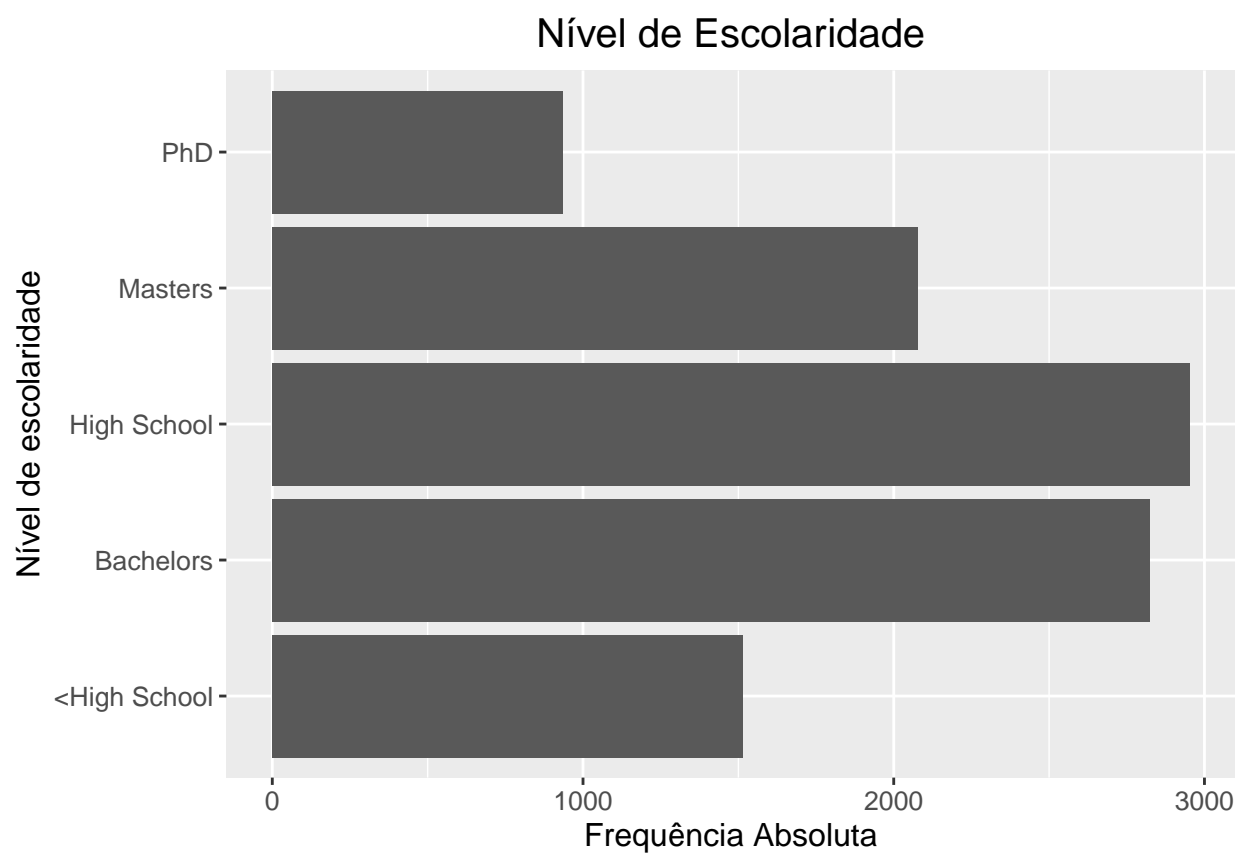
```
##           name value
## 1 <High School  1515
## 2 High School  2952
## 3 Bachelors    2824
## 4 Masters     2078
## 5 PhD          934
```

3.5.3.1.5 Barplot

```
ggplot(dados.freq, aes(x=name, y=value)) +
  geom_bar(stat = "identity") +
  ggtitle("Nível de Escolaridade") +
  xlab("Nível de escolaridade") +
  ylab("Frequência Absoluta") +
  coord_flip() +
  theme(legend.position="none",
```



```
plot.title = element_text(hjust = 0.5, size = 15),  
axis.title = element_text(size = 12),  
axis.text = element_text(size = 10)  
)
```



3.5.3.2 GENDER - Sexo

3.5.3.2.1 Estatísticas básicas do R

```
median(as.numeric(Claim.Data$GENDER))           # Nível de educação Mediana
```

```
## [1] 2
```

```
quantile(as.numeric(Claim.Data$GENDER), type = 2) # Quartis
```

```
##    0%   25%   50%   75%  100%  
##     1     1     2     2     2
```

```
IQR(as.numeric(Claim.Data$GENDER), type = 2)     # Distancia interquartilica
```

```
## [1] 1
```

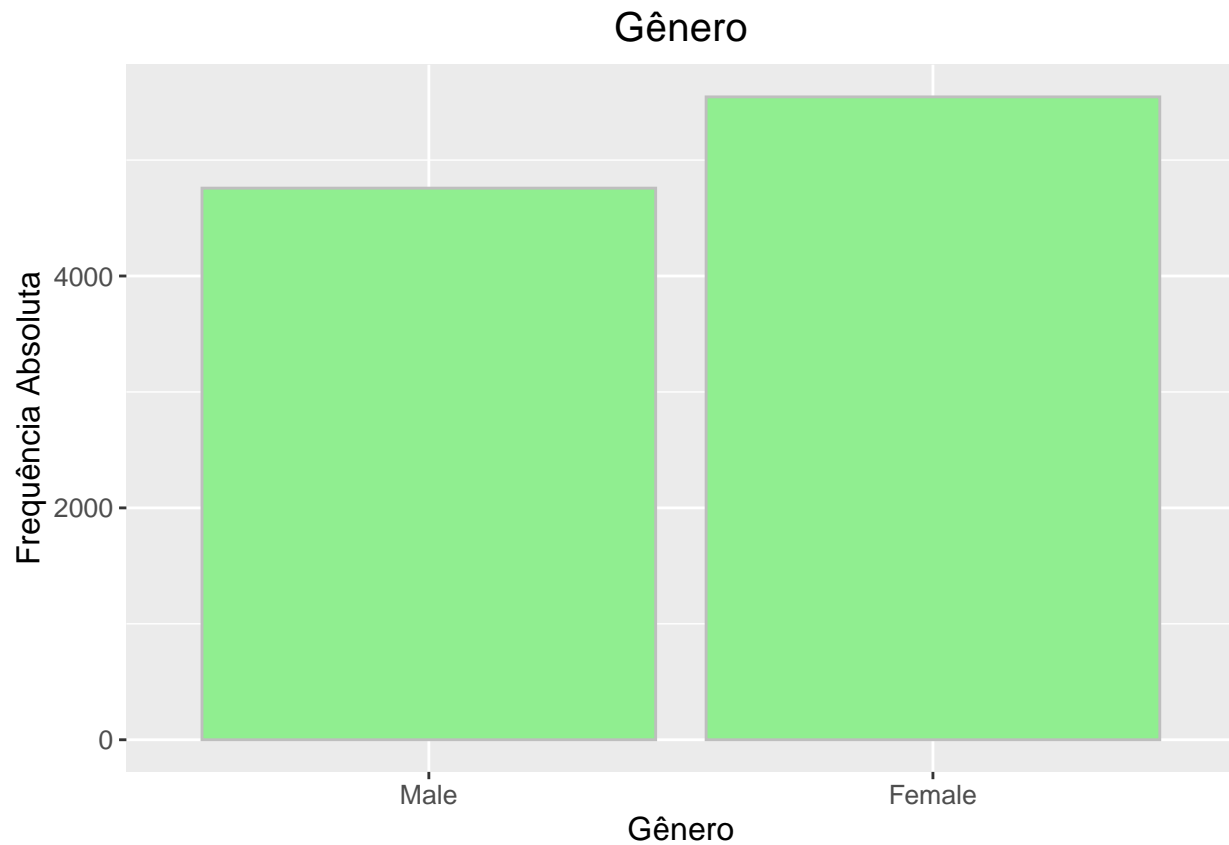
3.5.3.2.2 Resumo tabular

```
tabela <- freq(Claim.Data$GENDER, cum = TRUE, total = TRUE, valid = FALSE)  
tabela
```

```
## Frequencies  
## Claim.Data$GENDER  
## Type: Factor  
##  
##           Freq  % Valid  % Valid Cum.  % Total  % Total Cum.  
## -----  
##           Male   4758    46.18      46.18    46.18    46.18  
##           Female  5545    53.82     100.00    53.82   100.00  
##           <NA>     0      100.00     100.00    0.00   100.00  
##           Total 10303    100.00     100.00   100.00   100.00
```

3.5.3.2.3 Resumo gráfico

```
ggplot(Claim.Data,  
       aes(x = GENDER )) +  
  geom_bar(color = "grey", fill = "lightgreen") +  
  ggtitle("Gênero") +  
  xlab("Gênero") +  
  ylab("Frequência Absoluta") +  
  theme(legend.position="none",  
        plot.title = element_text(hjust = 0.5, size = 15),  
        axis.title = element_text(size = 12),  
        axis.text = element_text(size = 10)  
  )
```



3.5.3.2.4 Tabela de frequências

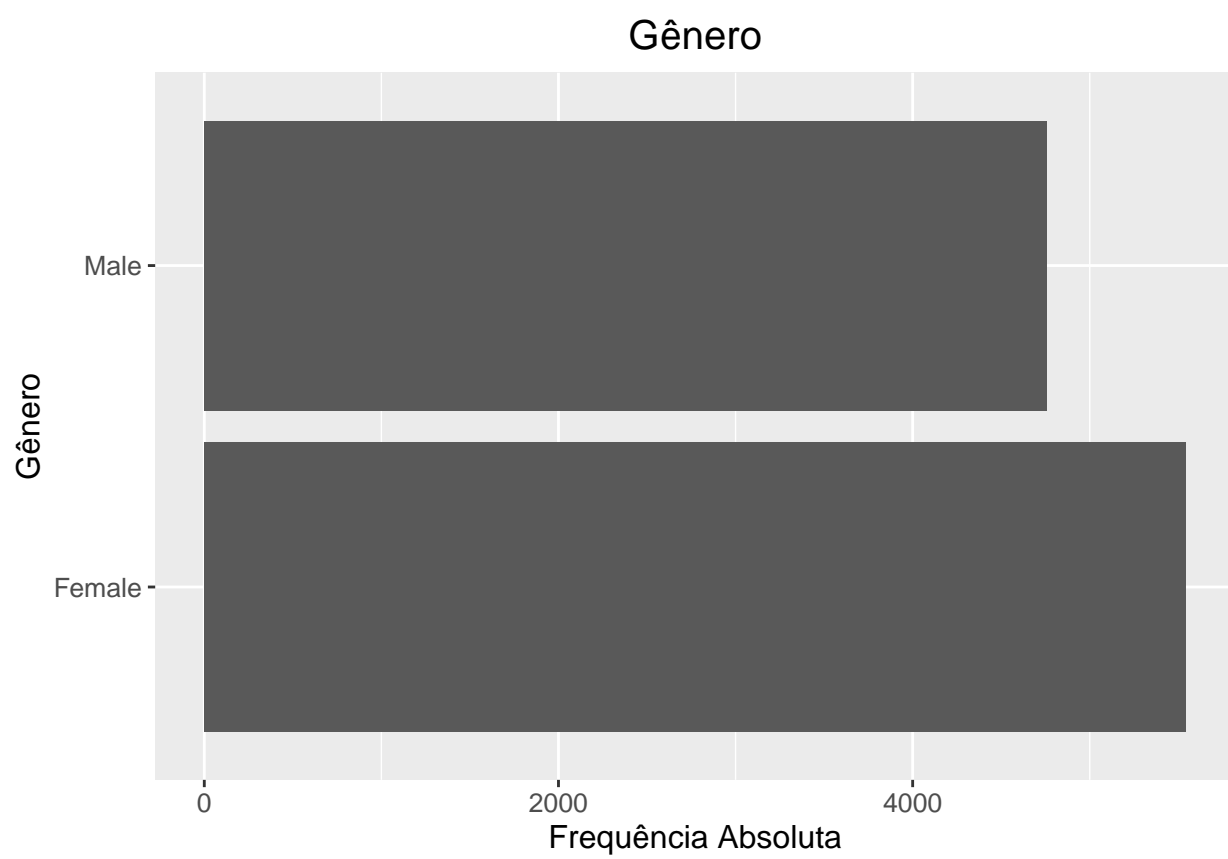
```
dados.freq <- data.frame(
  name = rownames(table(Claim.Data$GENDER)),
  value = as.vector(table(Claim.Data$GENDER))
)
dados.freq
```

```
##      name value
## 1   Male  4758
## 2 Female  5545
```

3.5.3.2.5 Barplot

```
ggplot(dados.freq, aes(x=name, y=value)) +
  geom_bar(stat = "identity") +
  ggtitle("Gênero") +
  xlab("Gênero") +
  ylab("Frequência Absoluta") +
  coord_flip() +
  theme(legend.position="none",
        plot.title = element_text(hjust = 0.5, size = 15),
        axis.title = element_text(size = 12),
```

```
)  
    axis.text = element_text(size = 10)
```



3.5.3.3 MARRIED - Casado

3.5.3.3.1 Estatísticas básicas do R Como esta variável é pelo menos ordinal, pode-se calcular as estatísticas de ordem e, portanto, calcular mediana, IQR e quantis.

```
median(as.numeric(Claim.Data$MARRIED)) # Nível de educação Mediana
```

```
## [1] 2
```

```
quantile(as.numeric(Claim.Data$MARRIED), type = 2) # Quartis
```

```
## 0% 25% 50% 75% 100%  
## 1 1 2 2 2
```

```
IQR(as.numeric(Claim.Data$MARRIED), type = 2) # Distancia interquartilica
```

```
## [1] 1
```

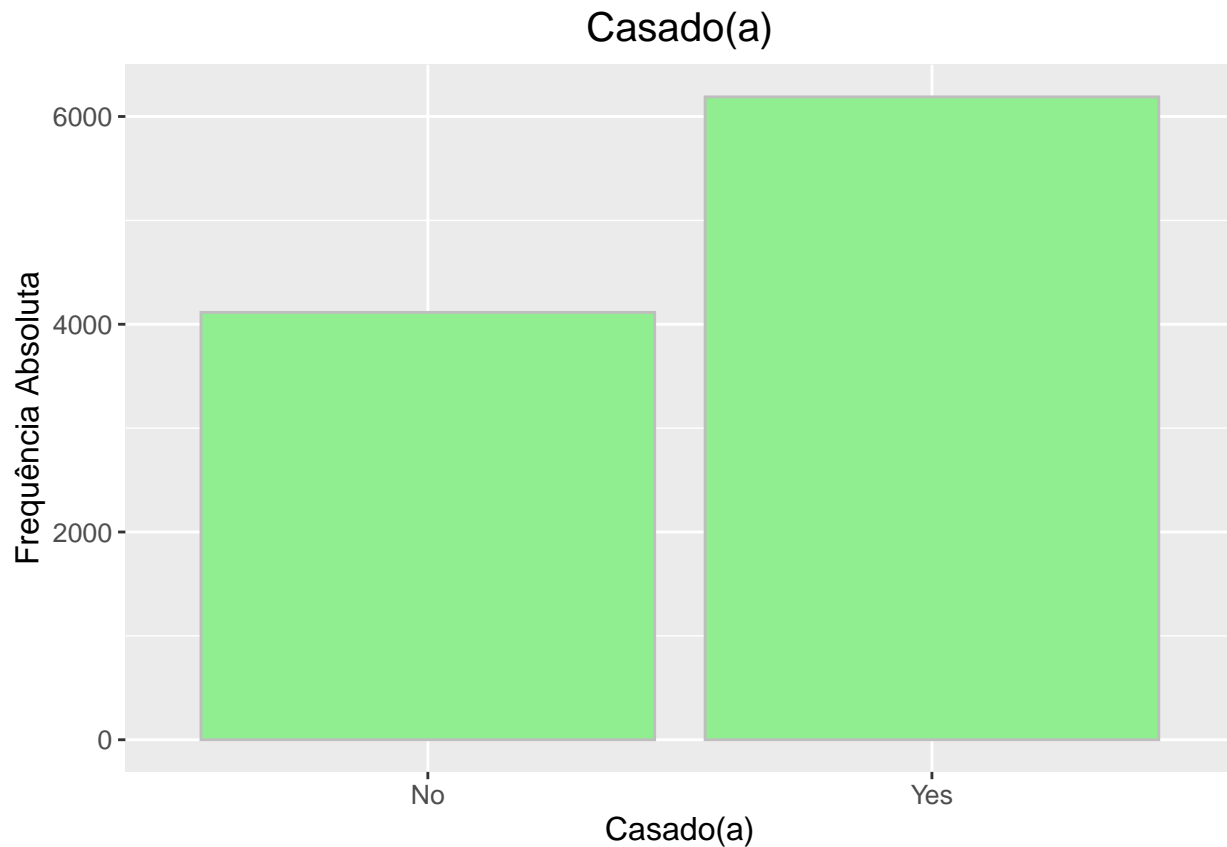
3.5.3.3.2 Resumo tabular

```
tabela <- freq(Claim.Data$MARRIED, cum = TRUE, total = TRUE, valid = FALSE)  
tabela
```

```
## Frequencies  
## Claim.Data$MARRIED  
## Type: Factor  
##  
##          Freq  % Valid  % Valid Cum.  % Total  % Total Cum.  
## -----  
##          No    4114    39.93      39.93    39.93    39.93  
##          Yes    6189    60.07     100.00    60.07    100.00  
##          <NA>     0      0.00      0.00     0.00    100.00  
##          Total 10303   100.00     100.00   100.00    100.00
```

3.5.3.3.3 Resumo gráfico

```
ggplot(Claim.Data,  
  aes(x = MARRIED )) +  
  geom_bar(color = "grey", fill = "lightgreen") +  
  ggtitle("Casado(a)") +  
  xlab("Casado(a)") +  
  ylab("Frequência Absoluta") +  
  theme(legend.position="none",  
    plot.title = element_text(hjust = 0.5, size = 15),  
    axis.title = element_text(size = 12),  
    axis.text = element_text(size = 10)  
  )
```



3.5.3.3.4 Tabela de frequências

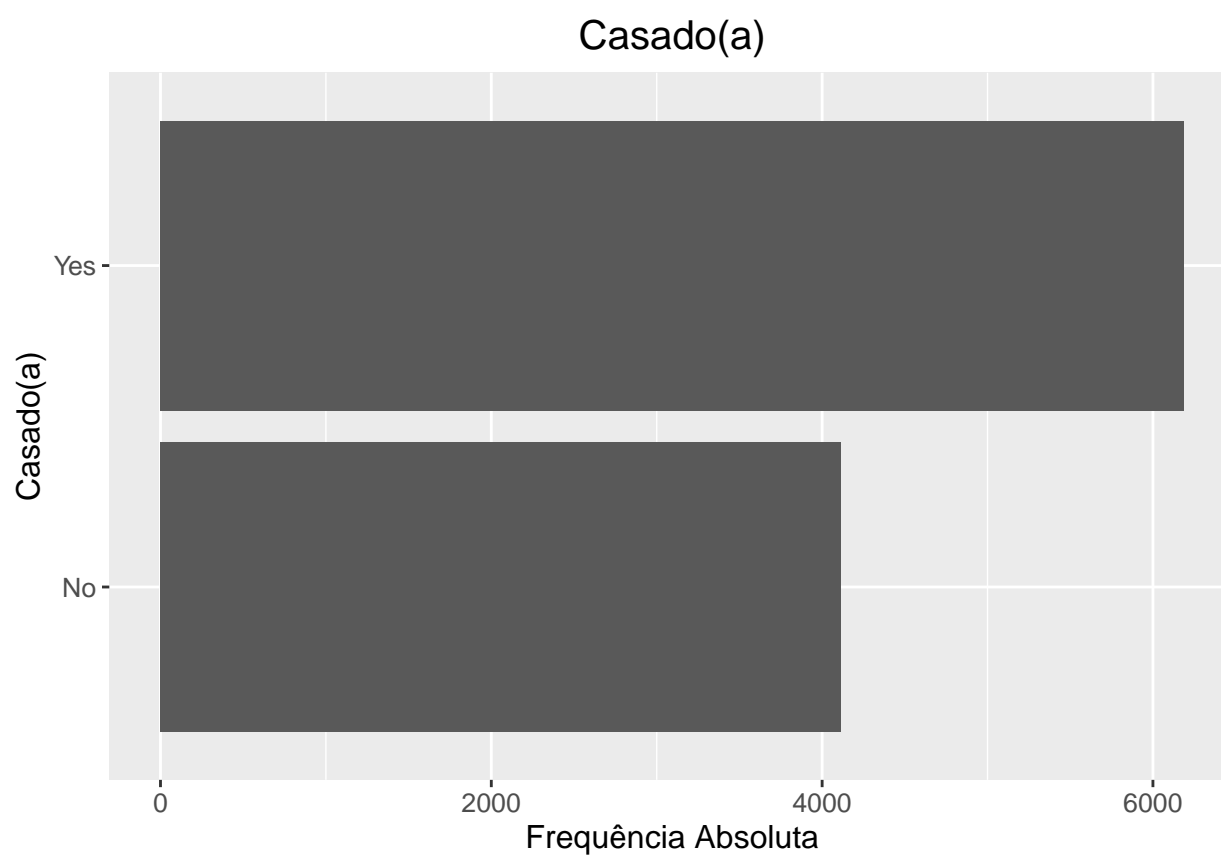
```
dados.freq <- data.frame(
  name = rownames(table(Claim.Data$MARRIED)),
  value = as.vector(table(Claim.Data$MARRIED))
)
dados.freq
```

```
##   name value
## 1  No  4114
## 2  Yes  6189
```

3.5.3.3.5 Barplot

```
ggplot(dados.freq, aes(x=name, y=value)) +
  geom_bar(stat = "identity") +
  ggtitle("Casado(a)") +
  xlab("Casado(a)") +
  ylab("Frequência Absoluta") +
  coord_flip() +
  theme(legend.position="none",
        plot.title = element_text(hjust = 0.5, size = 15),
        axis.title = element_text(size = 12),
```

```
axis.text = element_text(size = 10)  
)
```



4. DISCUSSÃO E CONCLUSÕES

5. REFERÊNCIAS

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.

Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686, <https://doi.org/10.21105/joss.01686>

JJ Allaire and Yihui Xie and Jonathan McPherson and Javier Luraschi and Kevin Ushey and Aron Atkins and Hadley Wickham and Joe Cheng and Winston Chang and Richard Iannone (2021). *rmarkdown: Dynamic Documents for R*. R package version 2.8. URL <https://rmarkdown.rstudio.com>.

Yihui Xie and J.J. Allaire and Garrett Grolemond (2018). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC. ISBN 9781138359338. URL <https://bookdown.org/yihui/rmarkdown>.

Yihui Xie and Christophe Dervieux and Emily Riederer (2020). *R Markdown Cookbook*. Chapman and Hall/CRC. ISBN 9780367563837. URL <https://bookdown.org/yihui/rmarkdown-cookbook>.