Análise descritiva de uma base de dados

Otto Tavares

2023-02-13

Introdução

Na Aula 7, temos o objetivo de abrir uma base de dados e dar os primeiros passos em análise estatística dessa base.

Como sempre, o primeiro passo é importar as bibliotecas que serão utilizadas para análise, como tydiverse, summarytools e dlookr.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dlookr)
```

```
## Registered S3 methods overwritten by 'dlookr':
##   method          from
##   plot.transform  scales
##   print.transform scales
##
## Attaching package: 'dlookr'
##
## The following object is masked from 'package:tidyr':
##
##     extract
##
## The following object is masked from 'package:base':
##
##     transform
```

```
library(summarytools)
```

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

```
library(readxl)
library(knitr)

#crimes.furtos %>% dplyr::filter(mes_ano == "2022m12") %>% diagnose()

#crimes.furtos %>% dplyr::filter(mes_ano == "2022m12") %>% dfSummary() %>% view()
```

A base trabalhada nesta aula, será a base de dados hipotética disponbilizada no livro texto dos autores Bussab e Moretim. Vamos importá-la e imprimir as primeiras observações para conhecimento das variáveis.

```
salarios <- readxl::read_excel("dados_auxiliares/dados_bussab_m.xlsx")
```

```
kable(head(salarios))
```

n	estado_civil	Grau_de_instrucao	n_filhos	salario	idade_anos	idade_meses	regiao
1	solteiro	ensino fundamental	NA	4.00	26	3	interior
2	casado	ensino fundamental	1	4.56	32	10	capital
3	casado	ensino fundamental	2	5.25	36	5	capital
4	solteiro	ensino médio	NA	5.73	20	10	outra
5	solteiro	ensino fundamental	NA	6.26	40	7	outra
6	casado	ensino fundamental	0	6.66	28	0	interior

Identificando os tipos de cada variável na base

Para identificar os tipos de cada variável na base, vamos utilizar a função diagnose do pacote dlookr e reportar o tipo de cada um para melhor trabalharmos os dados.

```
salarios %>% dlookr::diagnose() %>% kable()
```

variables	types	missing_count	missing_percent	unique_count	unique_rate
n	numeric	0	0.00000	36	1.0000000
estado_civil	character	0	0.00000	2	0.0555556
Grau_de_instrucao	character	0	0.00000	3	0.0833333
n_filhos	numeric	16	44.44444	6	0.1666667
salario	numeric	0	0.00000	36	1.0000000
idade_anos	numeric	0	0.00000	24	0.6666667
idade_meses	numeric	0	0.00000	12	0.3333333
regiao	character	0	0.00000	3	0.0833333

É fácil ver que na base há três variáveis qualitativas, sendo as variáveis Estado Civil e região nominais, enquanto a variável Grau de Instrução é ordinal.

Sobre as variáveis quantitativas, temos número de filhos e idade com variáveis discretas, equanto a variável salário é contínua.

##Análise de frequências de variáveis qualitativas

A variável região é uma das variáveis qualitativas nominais da base, sendo uma variável interessante para extraírmos as frequências. Para esse caso, vamos utilizar a função freq() do pacote summarytools

```
salarios %>% dplyr::select(regiao) %>% summarytools::freq(., style = 'rmarkdown') %>% kable()
```

```
## setting plain.ascii to FALSE
```

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
capital	11	30.55556	30.55556	30.55556	30.55556
interior	12	33.33333	63.88889	33.33333	63.88889
outra	13	36.11111	100.00000	36.11111	100.00000
	0	NA	NA	0.00000	100.00000
Total	36	100.00000	100.00000	100.00000	100.00000

Nas colunas Freq, temos a frequência absoluta, mostrando um grau de bastante homogeneidade entre as classes. Padrão esse, que é confirmado com a coluna Valid, que apresenta as frequências relativas de cada opção de região.

Podemos fazer a mesma análise para os dados de estado civil, os quais podemos estar interessados em buscar evidência se há mais funcionários casados ou solteiros na empresa. A seguir, temos a tabela destas proporções, onde é perceptível que há maior proporção de funcionários casados.

```
salarios %>% dplyr::select(estado_civil) %>% summarytools::freq(., style = 'rmarkdown') %>% kable()
```

```
## setting plain.ascii to FALSE
```

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
casado	20	55.55556	55.55556	55.55556	55.55556
solteiro	16	44.44444	100.00000	44.44444	100.00000
	0	NA	NA	0.00000	100.00000
Total	36	100.00000	100.00000	100.00000	100.00000

É importante destacar, que lemos a coluna Valid sem nos preocupar nestes casos, pois não há dados faltantes para nenhumas das duas variáveis.

Por fim, podemos criar tabelas de frequências para uma variável quantitativa discreta, como é o caso do número de filhos dos funcionários da empresa.

```
salarios %>% dplyr::select(n_filhos) %>% summarytools::freq(., style = 'rmarkdown') %>% kable()
```

```
## setting plain.ascii to FALSE
```

	Freq	% Valid	% Valid Cum.	% Total	% Total Cum.
0	4	20	20	11.111111	11.11111
1	5	25	45	13.888889	25.00000
2	7	35	80	19.444444	44.44444
3	3	15	95	8.333333	52.77778
5	1	5	100	2.777778	55.55556
	16	NA	NA	44.444444	100.00000
Total	36	100	100	100.000000	100.00000

Como há dados faltantes para essa variável, é importante o analista determinar qual o espaço amostral está interessado em focar sua análise.

A fim de ser comparável às análises pregressas, é importante que as frequências absoluta e relativa do total de dados seja considerada, isto é, leitura da coluna Total, a fim de manter o mesmo espaço amostral.

Caso, ele esteja interessado em analisar apenas os dados válidos, ele pode redefinir o espaço amostral, ler apenas a coluna Valid, porém recalculando as tabelas anteriores, considerando os indivíduos apenas com dados preenchidos para a variável filhos.

##Análise descritiva e de histogramas de uma variável contínua

Já para a variável salários, podemos analisar a centralidade dos dados, dipersão, assimetria, bem como suas estatísticas de ordem, a fim de checar se há presença de outliers.

Para realizar essa análise, podemos utilizar a função descr do pacote summarytools, e posteriormente realizar a leitura desses dados.

```
salarios %>% dplyr::select(salario) %>% summarytools::descr(., style = 'rmarkdown') %>% kable()
```

	salario
Mean	11.1222222
Std.Dev	4.5874575
Min	4.0000000
Q1	7.5150000
Median	10.1650000
Q3	14.2700000
Max	23.3000000
MAD	4.7220810
IQR	6.5075000
CV	0.4124587
Skewness	0.5997938
SE.Skewness	0.3925439
Kurtosis	-0.3291263
N.Valid	36.0000000
Pct.Valid	100.0000000

```
salarios %>% summarytools::dfSummary()
```

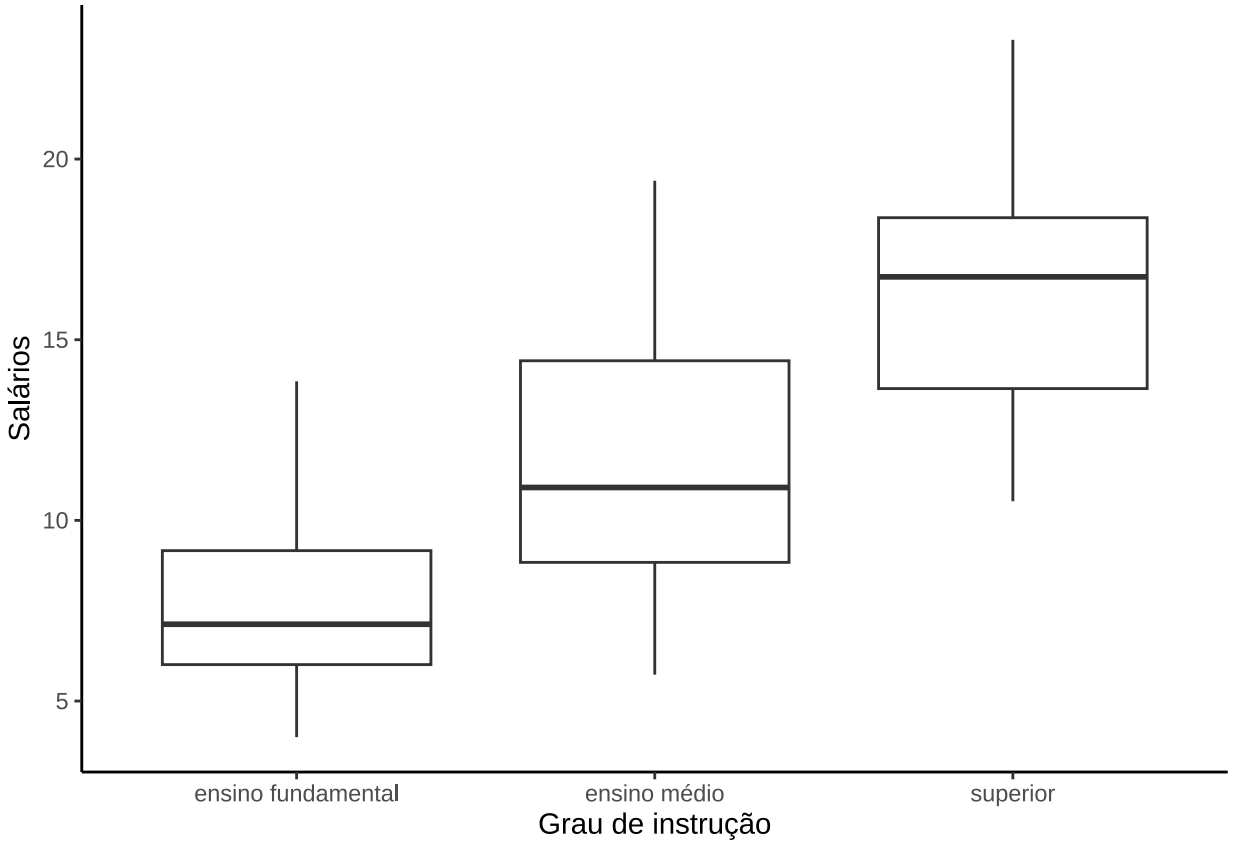
```
## Data Frame Summary
## salarios
## Dimensions: 36 x 8
```

```
## Duplicates: 0
##
## ---------------------------------------------------------------------------------------------
## No   Variable            Stats / Values            Freqs (% of Valid)   Graph                 Valid
## ---- ------------------- ------------------------- -------------------- --------------------- -------
## 1    n                   Mean (sd) : 18.5 (10.5)   36 distinct values   : : : : : : : :       36
##      [numeric]           min < med < max:                               : : : : : : : :       (100.0
##                          1 < 18.5 < 36                                  : : : : : : : :
##                          IQR (CV) : 17.5 (0.6)                          : : : : : : : :
##                                                                         : : : : : : : : :
##
## 2    estado_civil        1. casado                 20 (55.6%)           IIIIIIIIIII           36
##      [character]         2. solteiro               16 (44.4%)           IIIIIIII              (100.0
##
## 3    Grau_de_instrucao   1. ensino fundamental     12 (33.3%)           IIIIII                36
##      [character]         2. ensino médio           18 (50.0%)           IIIIIIIIII            (100.0
##                          3. superior                6 (16.7%)           III
##
## 4    n_filhos            Mean (sd) : 1.6 (1.3)     0 : 4 (20.0%)        IIII                  20
##      [numeric]           min < med < max:          1 : 5 (25.0%)        IIIII                 (55.6%
##                          0 < 2 < 5                 2 : 7 (35.0%)        IIIIIII
##                          IQR (CV) : 1 (0.8)        3 : 3 (15.0%)        III
##                                                    5 : 1 ( 5.0%)        I
##
## 5    salario             Mean (sd) : 11.1 (4.6)    36 distinct values     . :                 36
##      [numeric]           min < med < max:                                 : : .               (100.0
##                          4 < 10.2 < 23.3                                : : : : : :   :
##                          IQR (CV) : 6.5 (0.4)                           : : : : : : : : .
##                                                                         : : : : : : : : :   .
##
## 6    idade_anos          Mean (sd) : 34.6 (6.7)    24 distinct values       :                 36
##      [numeric]           min < med < max:                                 . : :               (100.0
##                          20 < 34.5 < 48                                   : : : :
##                          IQR (CV) : 10 (0.2)                            . : : : :
##                                                                         : : : : : :
##
## 7    idade_meses         Mean (sd) : 5.6 (3.3)     12 distinct values   .   . :               36
##      [numeric]           min < med < max:                               :   : :               (100.0
##                          0 < 6 < 11                                     :   : : :
##                          IQR (CV) : 4.2 (0.6)                           : : : : :
##                                                                         : : : : : :
##
## 8    regiao              1. capital                11 (30.6%)           IIIIII                36
##      [character]         2. interior               12 (33.3%)           IIIIII                (100.0
##                          3. outra                  13 (36.1%)           IIIIIII
## ---------------------------------------------------------------------------------------------
```

##Análise visual da distribuição dos indivíduos por idade

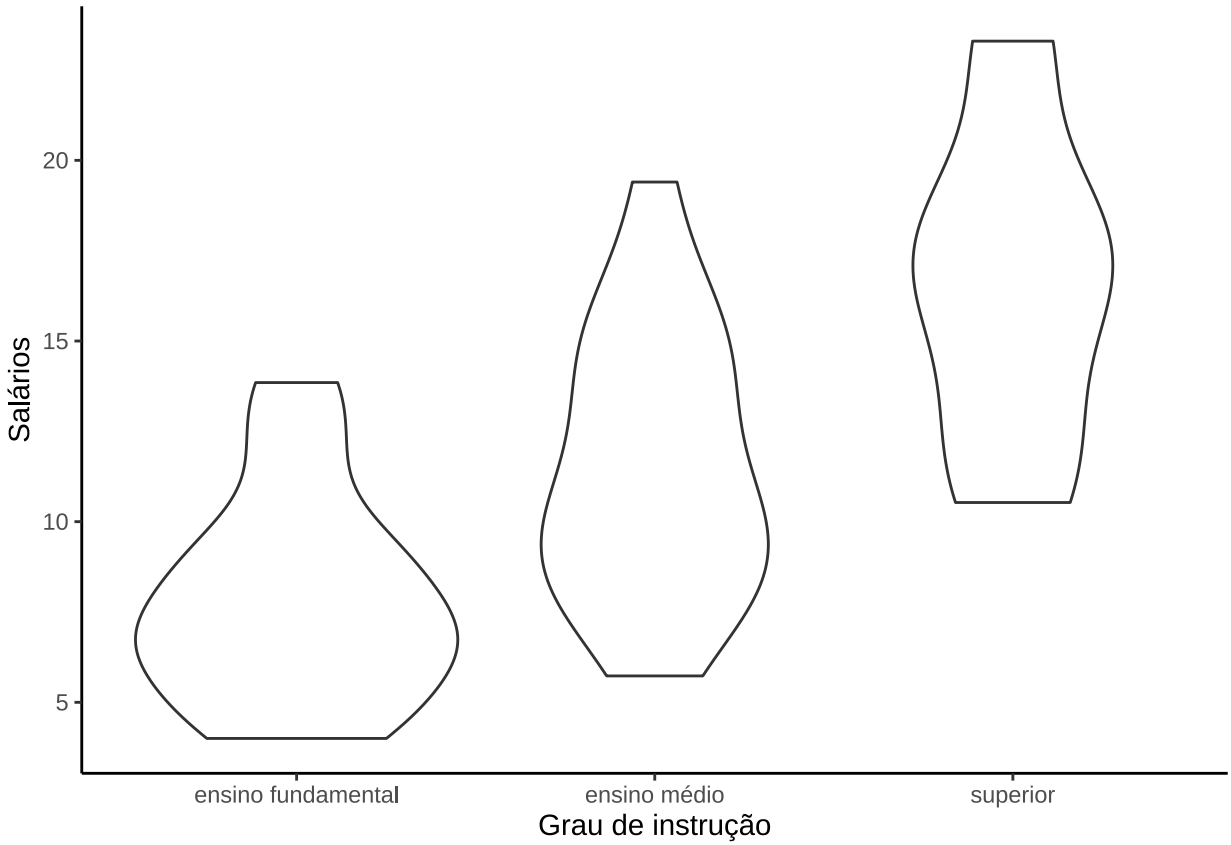
Com o boxplot

```
salarios %>% dplyr::select(Grau_de_instrucao, salario) %>% ggplot(aes(x=Grau_de_instrucao, y = salario)
```



Com o violino

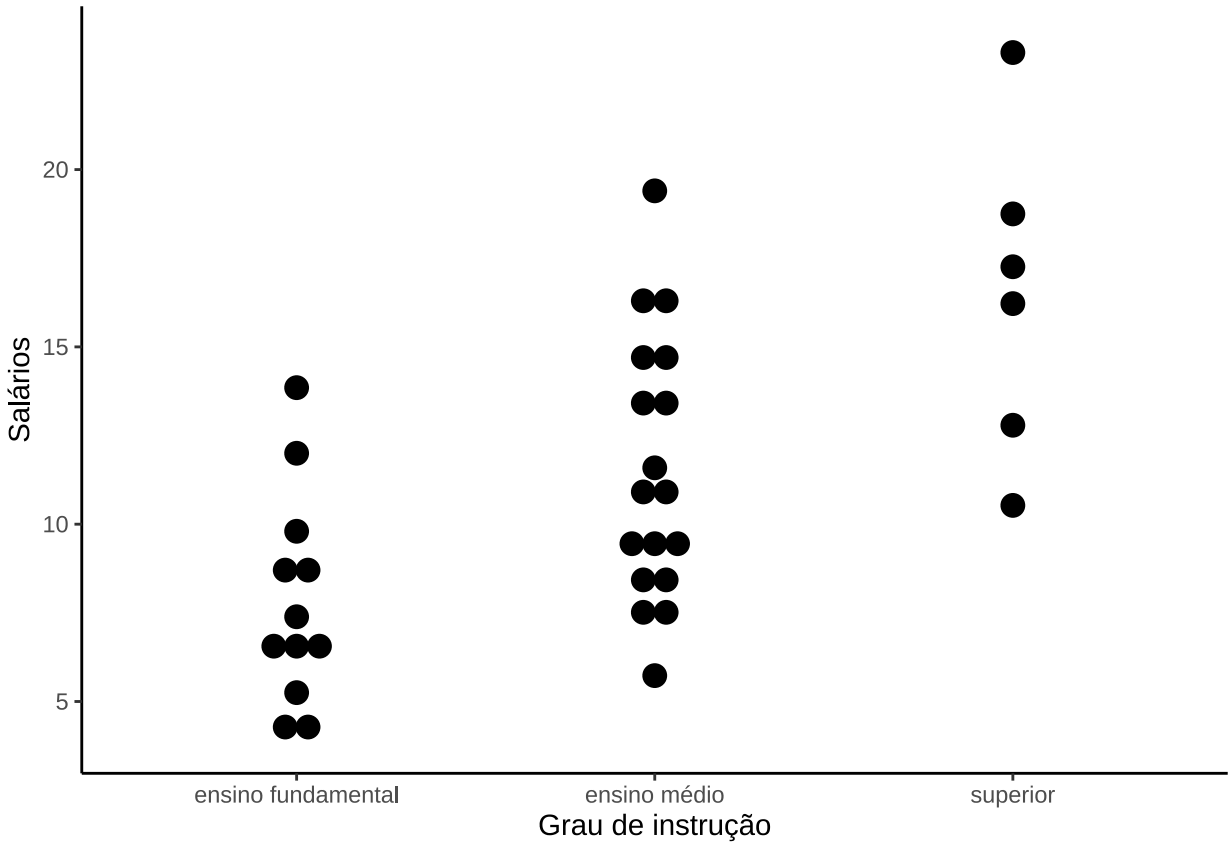
```
salarios %>% dplyr::select(Grau_de_instrucao, salario) %>% ggplot(aes(x=Grau_de_instrucao, y = salario)
```



Com o dotplot

```
salarios %>% dplyr::select(Grau_de_instrucao, salario) %>% ggplot(aes(x=Grau_de_instrucao, y = salario)
```

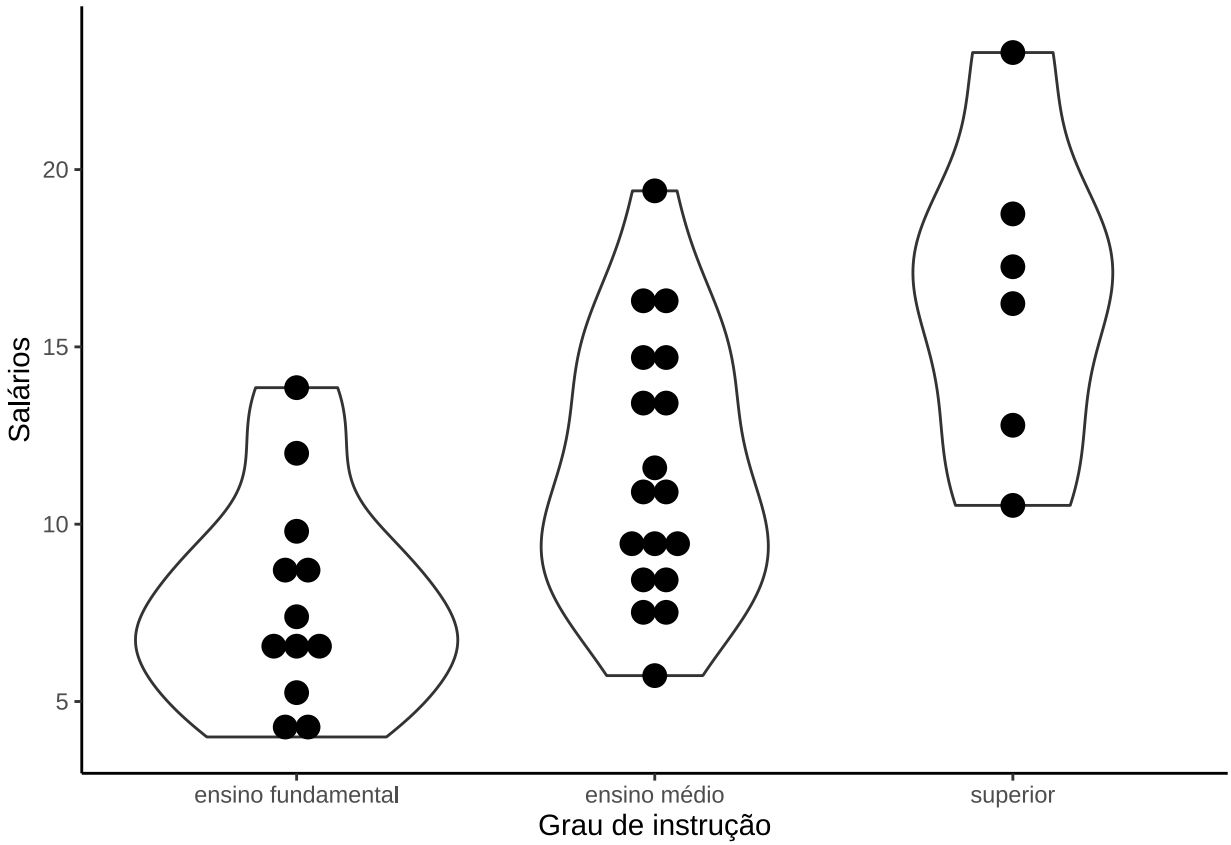
```
## Bin width defaults to 1/30 of the range of the data. Pick better value with
## `binwidth`.
```



Unindo o dotplot com o box ou violin para melhor ilustrar a análise

```
salarios %>% dplyr::select(Grau_de_instrucao, salario) %>% ggplot(aes(x=Grau_de_instrucao, y = salario)
```

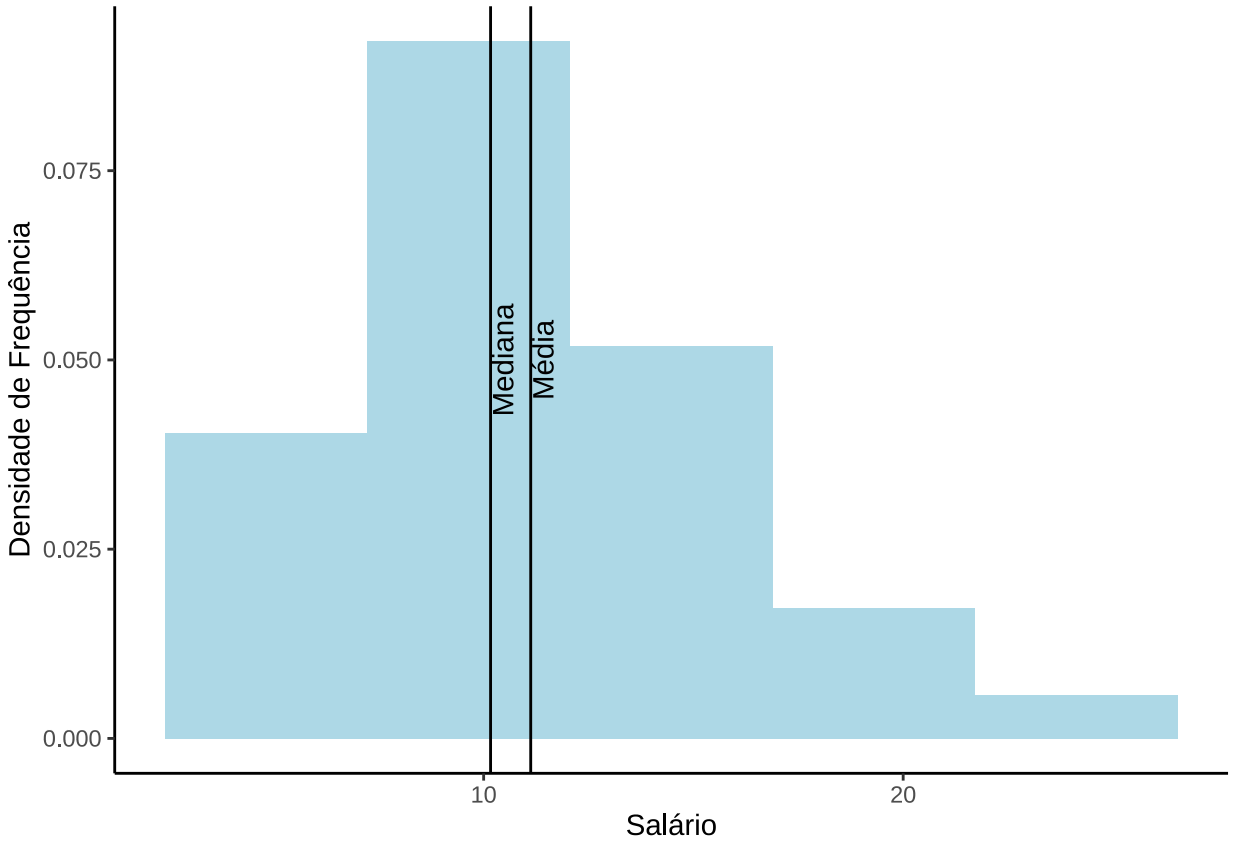
```
## Bin width defaults to 1/30 of the range of the data. Pick better value with
## `binwidth`.
```



##Análise visual da variável salário

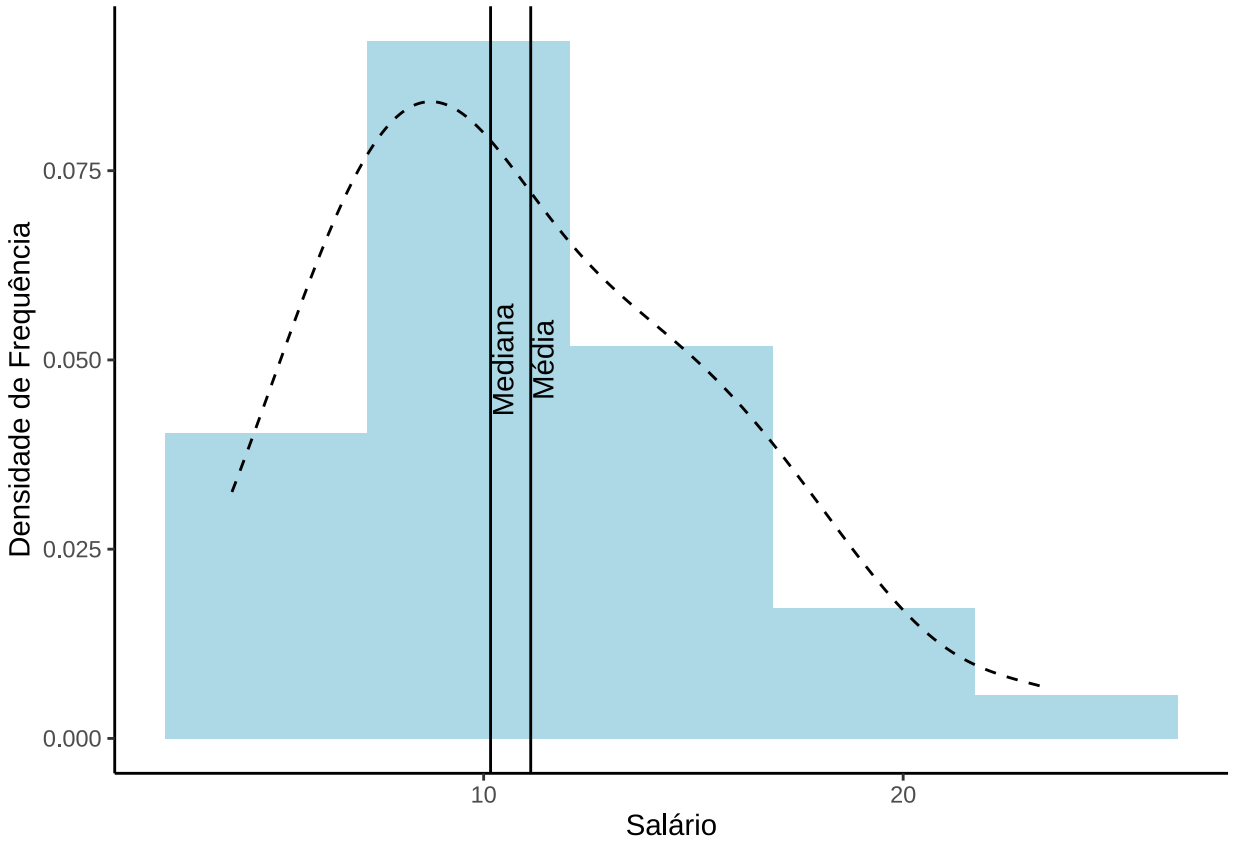
Utilizando o número de bins indicado pelos autores do livro, bins igual a 5.

```
salarios %>% dplyr::select(salario) %>% ggplot(aes(x=salario))+geom_histogram(aes(y = after_stat(density
```



Adicionando a densidade estimada via kernel à visualização

```
salarios %>% dplyr::select(salario) %>% ggplot(aes(x=salario))+geom_histogram(aes(y = after_stat(density
```



##Análise visual da variável salário, utilizando a binarização a partir de uma função customizada

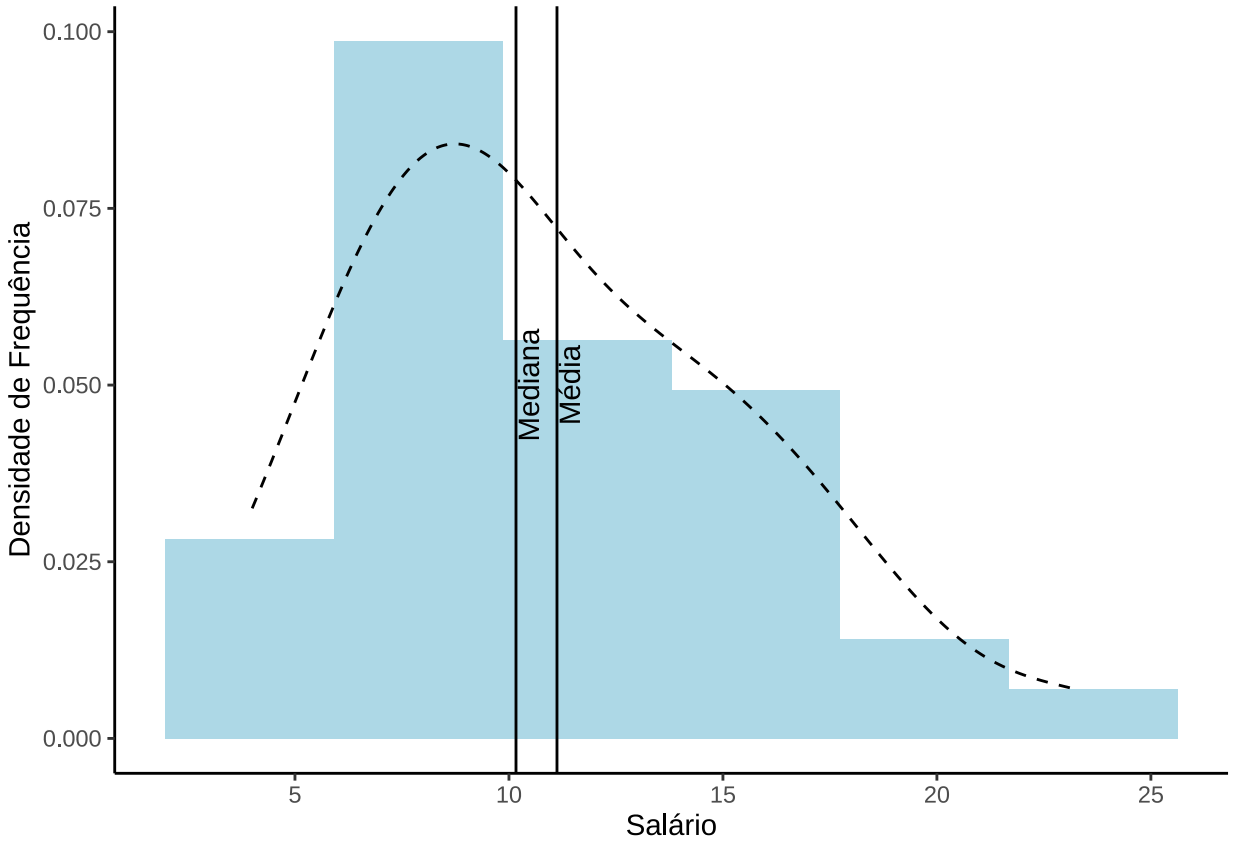
Definindo as funções gerais para criação de bins

```
#Freedman-Diaconis
fd_bins <- function(x)
{
  bins <- 2*IQR(x)/((length(x))^(1/3))
  return(bins)
}

#Sturge
s_bins <- function(x)
{
  bins <- 3.49*sd(x)/((length(x))^(1/3))
  return(bins)
}
```

Cálculo do número de bins a partir da função de Freedman-Diaconis

```
salarios %>% dplyr::select(salario) %>% ggplot(aes(x=salario))+geom_histogram(aes(y = after_stat(density
```


0.100
0.075
0.050
0.025
0.000
Densidade de Frequência
Mediana
Média
5
10
15
20
25
Salário