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   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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.

```
kable(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
7	solteiro	ensino fundamental	NA	6.86	41	0	interior
8	solteiro	ensino fundamental	NA	7.39	43	4	capital
9	casado	ensino médio	1	7.59	34	10	capital
10	solteiro	ensino médio	NA	7.44	23	6	outra
11	casado	ensino médio	2	8.12	33	6	interior
12	solteiro	ensino fundamental	NA	8.46	27	11	capital
13	solteiro	ensino médio	NA	8.74	37	5	outra
14	casado	ensino fundamental	3	8.95	44	2	outra
15	casado	ensino médio	0	9.13	30	5	interior
16	solteiro	ensino médio	NA	9.35	38	8	outra
17	casado	ensino médio	1	9.77	31	7	capital
18	casado	ensino fundamental	2	9.80	39	7	outra
19	solteiro	superior	NA	10.53	25	8	interior
20	solteiro	ensino médio	NA	10.76	37	4	interior
21	casado	ensino médio	1	11.06	30	9	outra
22	solteiro	ensino médio	NA	11.59	34	2	capital
23	solteiro	ensino fundamental	NA	12.00	41	0	outra
24	casado	superior	0	12.79	26	1	outra
25	casado	ensino médio	2	13.23	32	5	interior
26	casado	ensino médio	2	13.60	35	0	outra
27	solteiro	ensino fundamental	NA	13.85	46	7	outra
28	casado	ensino médio	0	14.69	29	8	interior
29	casado	ensino médio	5	14.71	40	6	interior
30	casado	ensino médio	2	15.99	35	10	capital
31	solteiro	superior	NA	16.22	31	5	outra
32	casado	ensino médio	1	16.61	36	4	interior
33	casado	superior	3	17.26	43	7	capital
34	solteiro	superior	NA	18.75	33	7	capital
35	casado	ensino médio	2	19.40	48	11	capital
36	casado	superior	3	23.30	42	2	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()
```

```
## # A tibble: 8 x 6
##   variables         types missing_count missing_percent unique_count unique_rate
##   <chr>             <chr>         <int>           <dbl>        <int>       <dbl>
## 1 n                 nume~             0             0            36      1
## 2 estado_civil      char~             0             0             2      0.0556
## 3 Grau_de_instrucao char~             0             0             3      0.0833
## 4 n_filhos          nume~            16            44.4           6      0.167
## 5 salario           nume~             0             0            36      1
## 6 idade_anos        nume~             0             0            24      0.667
## 7 idade_meses       nume~             0             0            12      0.333
## 8 regiao            char~             0             0             3      0.0833
```

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')
```

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

```
## ### Frequencies
## #### salarios$regiao
## **Type:** Character
##
## |         | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
## |-------------:|-----:|--------:|-------------:|--------:|-------------:|
## |  **capital** |   11 |   30.56 |        30.56 |   30.56 |        30.56 |
## | **interior** |   12 |   33.33 |        63.89 |   33.33 |        63.89 |
## |    **outra** |   13 |   36.11 |       100.00 |   36.11 |       100.00 |
## |   **\<NA\>** |    0 |         |               |    0.00 |       100.00 |
## |    **Total** |   36 |  100.00 |       100.00 |  100.00 |       100.00 |
```

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')
```

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

```
## ### Frequencies
## #### salarios$estado_civil
## **Type:** Character
##
```

```
## |         | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
## |-------------:|-----:|--------:|-------------:|--------:|-------------:|
## |   **casado** |   20 |   55.56 |        55.56 |   55.56 |        55.56 |
## | **solteiro** |   16 |   44.44 |       100.00 |   44.44 |       100.00 |
## |   **\<NA\>** |    0 |         |              |    0.00 |       100.00 |
## |    **Total** |   36 |  100.00 |       100.00 |  100.00 |       100.00 |
```

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')
```

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

```
## ### Frequencies
## #### salarios$n_filhos
## **Type:** Numeric
##
## |       | Freq | % Valid | % Valid Cum. | % Total | % Total Cum. |
## |-----------:|-----:|--------:|-------------:|--------:|-------------:|
## |      **0** |    4 |   20.00 |        20.00 |   11.11 |        11.11 |
## |      **1** |    5 |   25.00 |        45.00 |   13.89 |        25.00 |
## |      **2** |    7 |   35.00 |        80.00 |   19.44 |        44.44 |
## |      **3** |    3 |   15.00 |        95.00 |    8.33 |        52.78 |
## |      **5** |    1 |    5.00 |       100.00 |    2.78 |        55.56 |
## | **\<NA\>** |   16 |         |              |   44.44 |       100.00 |
## |  **Total** |   36 |  100.00 |       100.00 |  100.00 |       100.00 |
```

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')
```

```
## ### Descriptive Statistics
## #### salarios$salario
## **N:** 36
##
## |            | salario |
## |----------------:|--------:|
## |        **Mean** |   11.12 |
## |     **Std.Dev** |    4.59 |
```

```
## |         **Min** |    4.00 |
## |          **Q1** |    7.52 |
## |      **Median** |   10.16 |
## |          **Q3** |   14.27 |
## |         **Max** |   23.30 |
## |         **MAD** |    4.72 |
## |         **IQR** |    6.51 |
## |          **CV** |    0.41 |
## |    **Skewness** |    0.60 |
## | **SE.Skewness** |    0.39 |
## |    **Kurtosis** |   -0.33 |
## |     **N.Valid** |   36.00 |
## |   **Pct.Valid** |  100.00 |
```

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

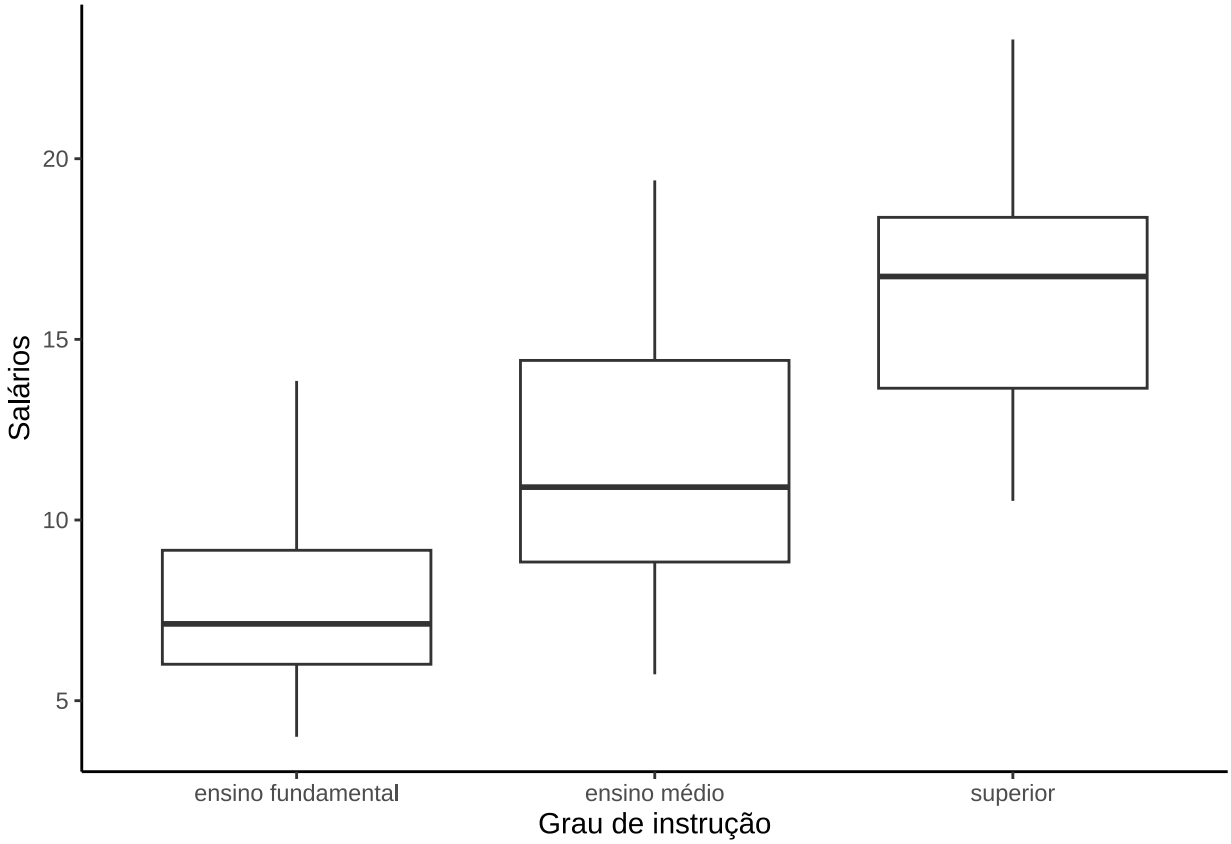
```
## Data Frame Summary
## salarios
## Dimensions: 36 x 8
## Duplicates: 0
##
## ----------------------------------------------------------------------------------------------------
## No   Variable            Stats / Values            Freqs (% of Valid)   Graph                 Valid
## ---- ------------------- ------------------------- -------------------- --------------------- -------
## 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%)           IIIIIIIII             (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

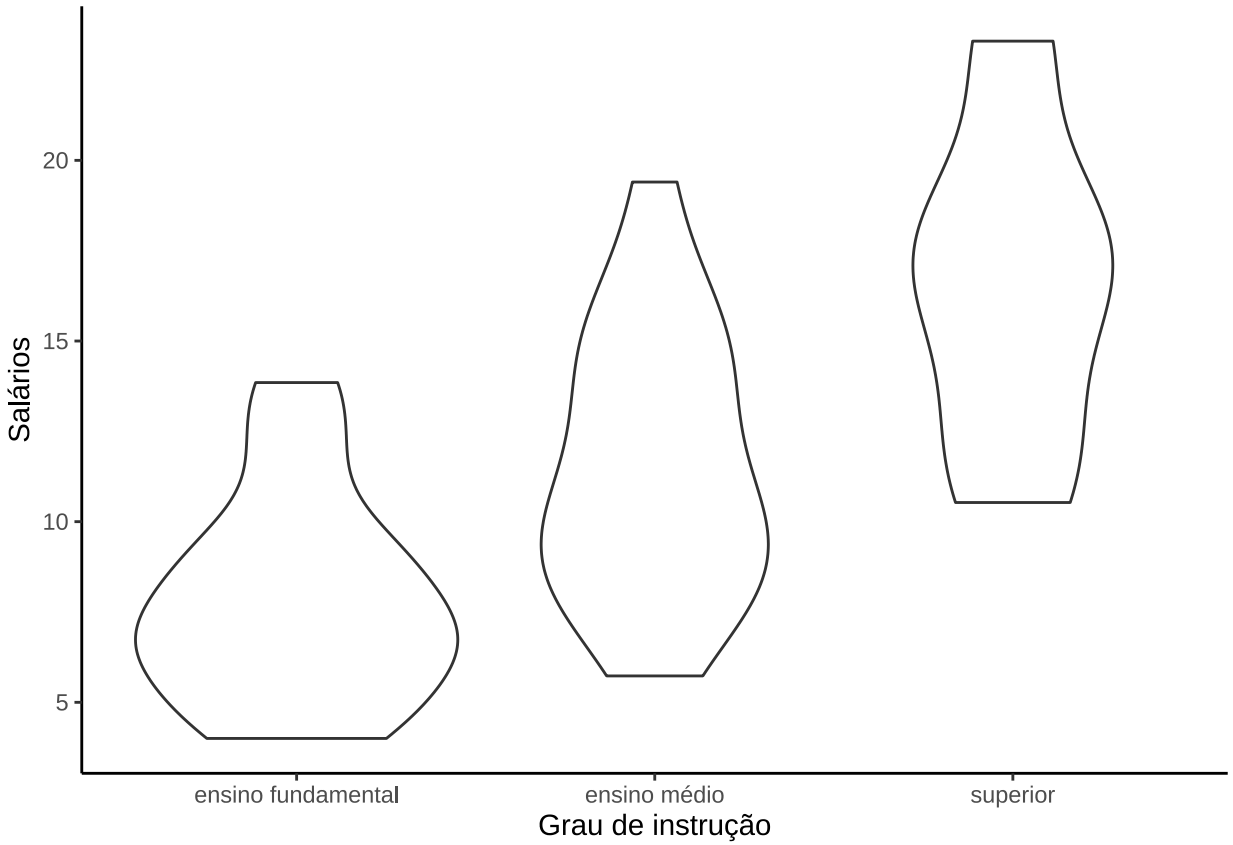
Com o boxplot

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



Com o violino

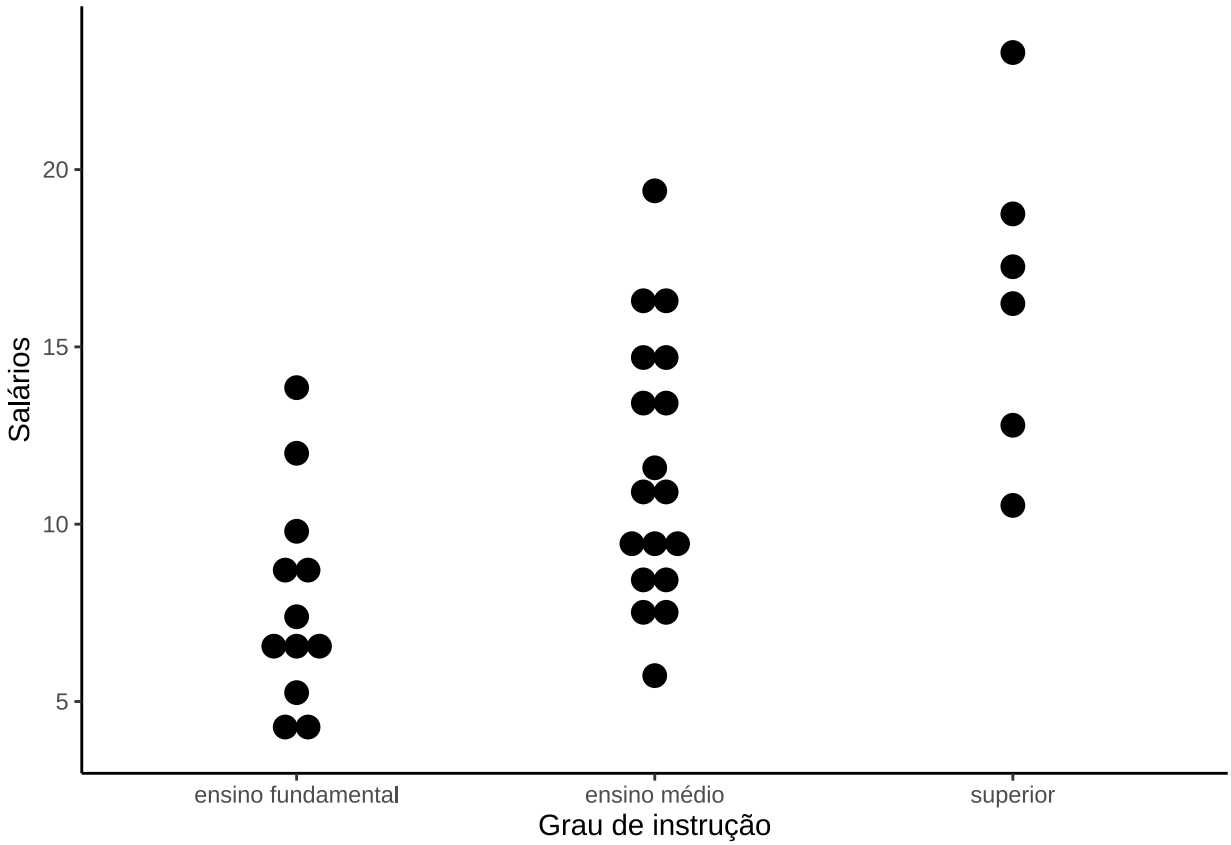
```
salarios %>% dplyr::select(Grau_de_instrucao, salario) %>% ggplot(aes(x=Grau_de_instrucao, y = salario)
```



Com o dotplot

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

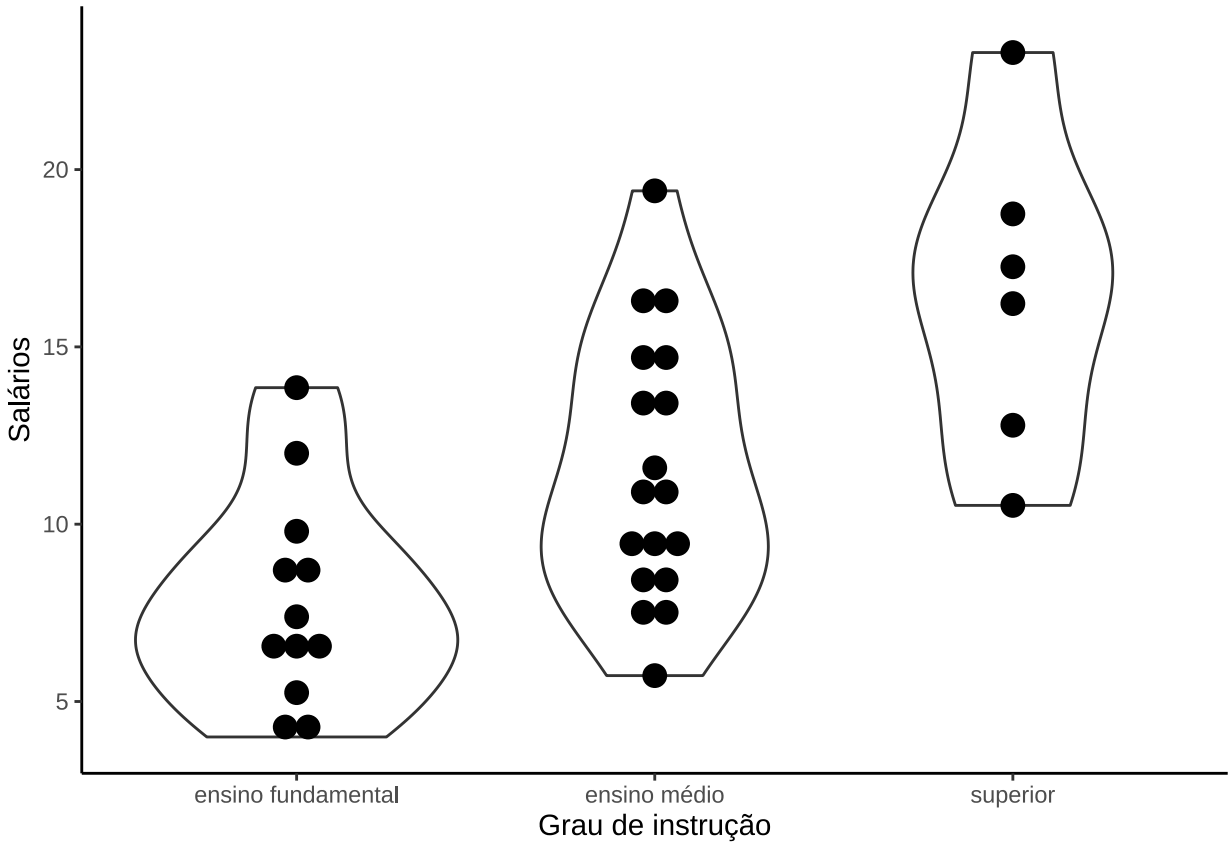
```
## Bin width defaults to 1/30 of the range of the data. Pick better value with
## `binwidth`.
```



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)
```

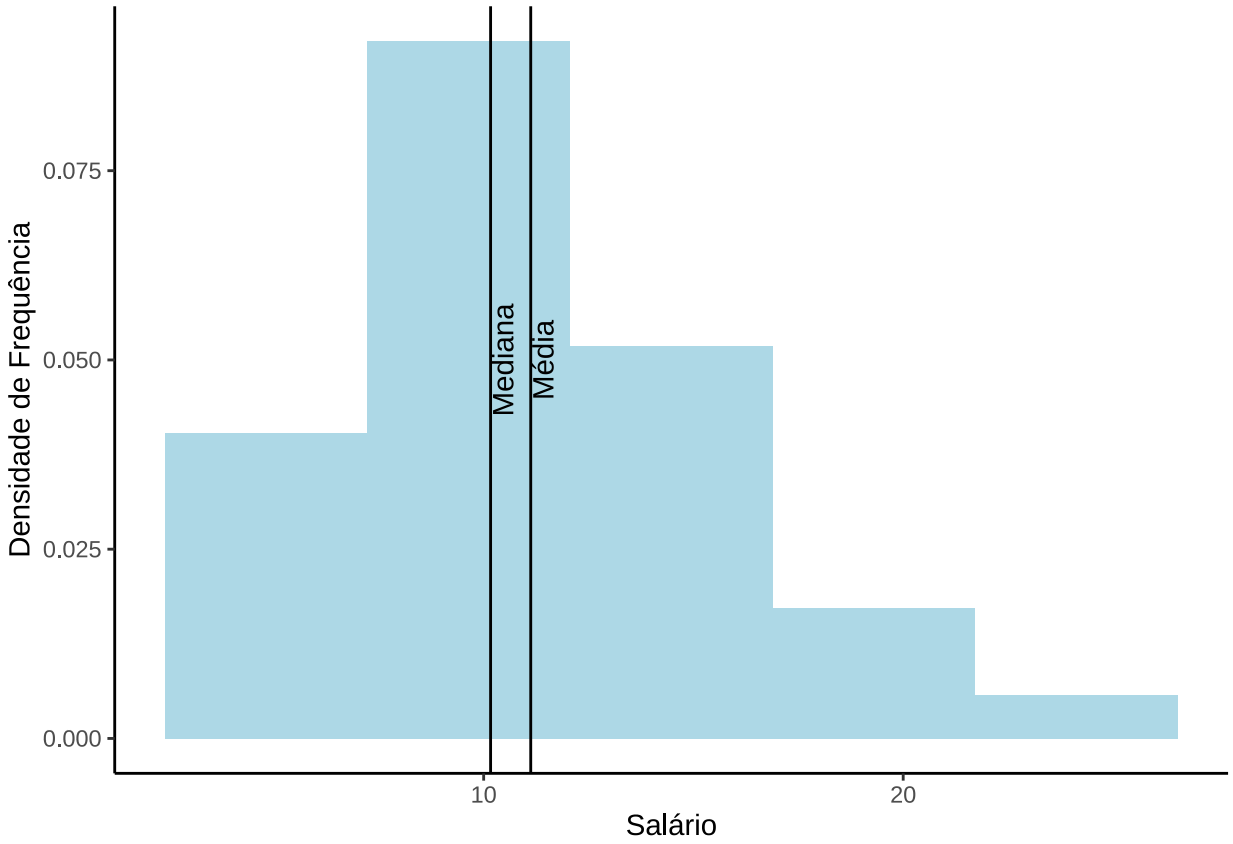
```
## Bin width defaults to 1/30 of the range of the data. Pick better value with
## `binwidth`.
```



##Análise visual da variável salário

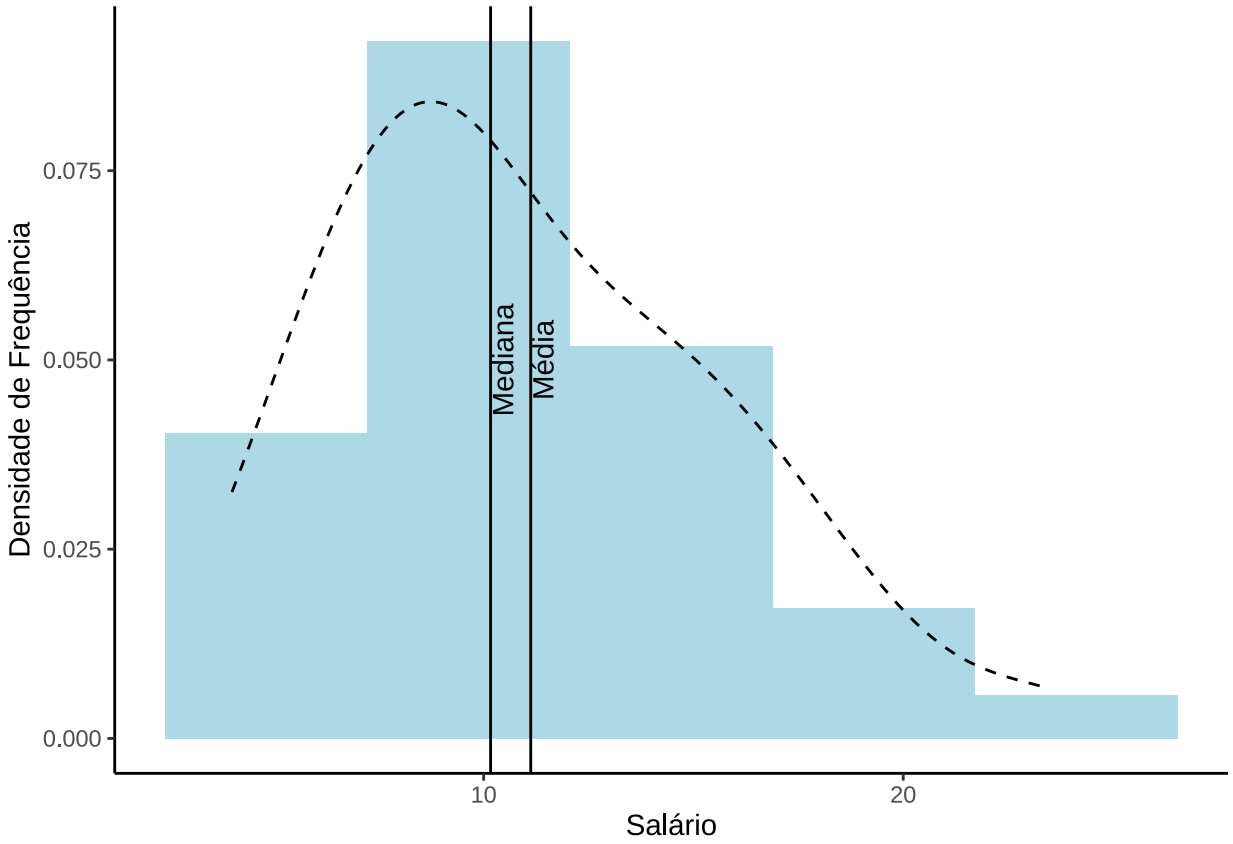
Utilizando o número de bins indicado pelos autores do livro, bins igual a 5.

```
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(densit
```



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

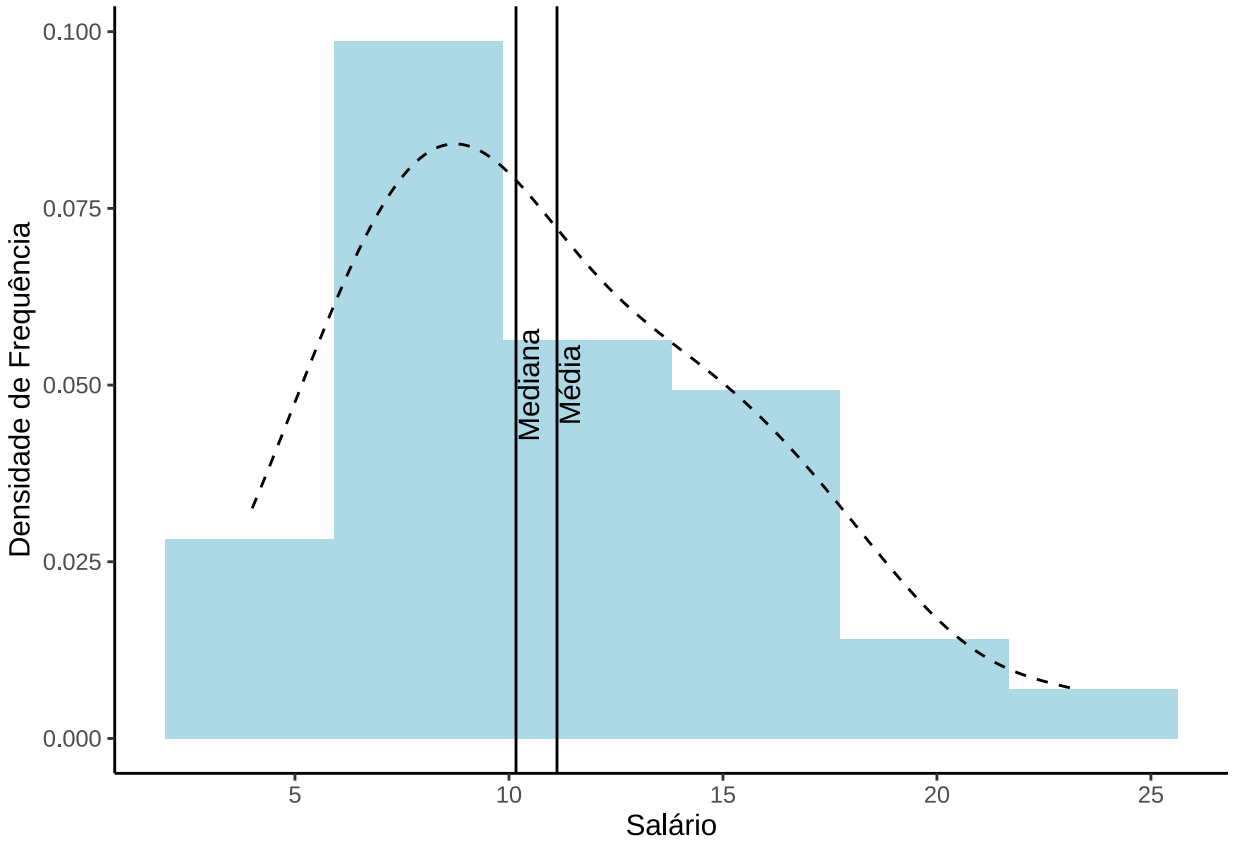
Definindo as funções gerais para criação de bins

```
#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