Análise descritiva de uma base de dados

Otto Tavares

2023-02-13

Contents

0.1 Introdução

Na Aula 8, 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 packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dlookr)
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 3 > 1' in coercion to 'logical(1)'
```

```
##
## 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 uniqu~1
##   <chr>             <chr>             <int>           <dbl>        <int>   <dbl>
## 1 n                 numeric               0             0            36  1
## 2 estado_civil      character             0             0             2  0.0556
## 3 Grau_de_instrucao character             0             0             3  0.0833
## 4 n_filhos          numeric              16            44.4           6  0.167
## 5 salario           numeric               0             0            36  1
## 6 idade_anos        numeric               0             0            24  0.667
```

```
## 7 idade_meses       numeric               0             0             12  0.333
## 8 regiao            character             0             0              3  0.0833
## # ... with abbreviated variable name 1: unique_rate
```

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()
```

```
## 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()
```

```
## 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()
```

```
## 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.

```
summarytools::ctable(x = salarios$Grau_de_instrucao,
       y = salarios$regiao,
       prop = "r")
```

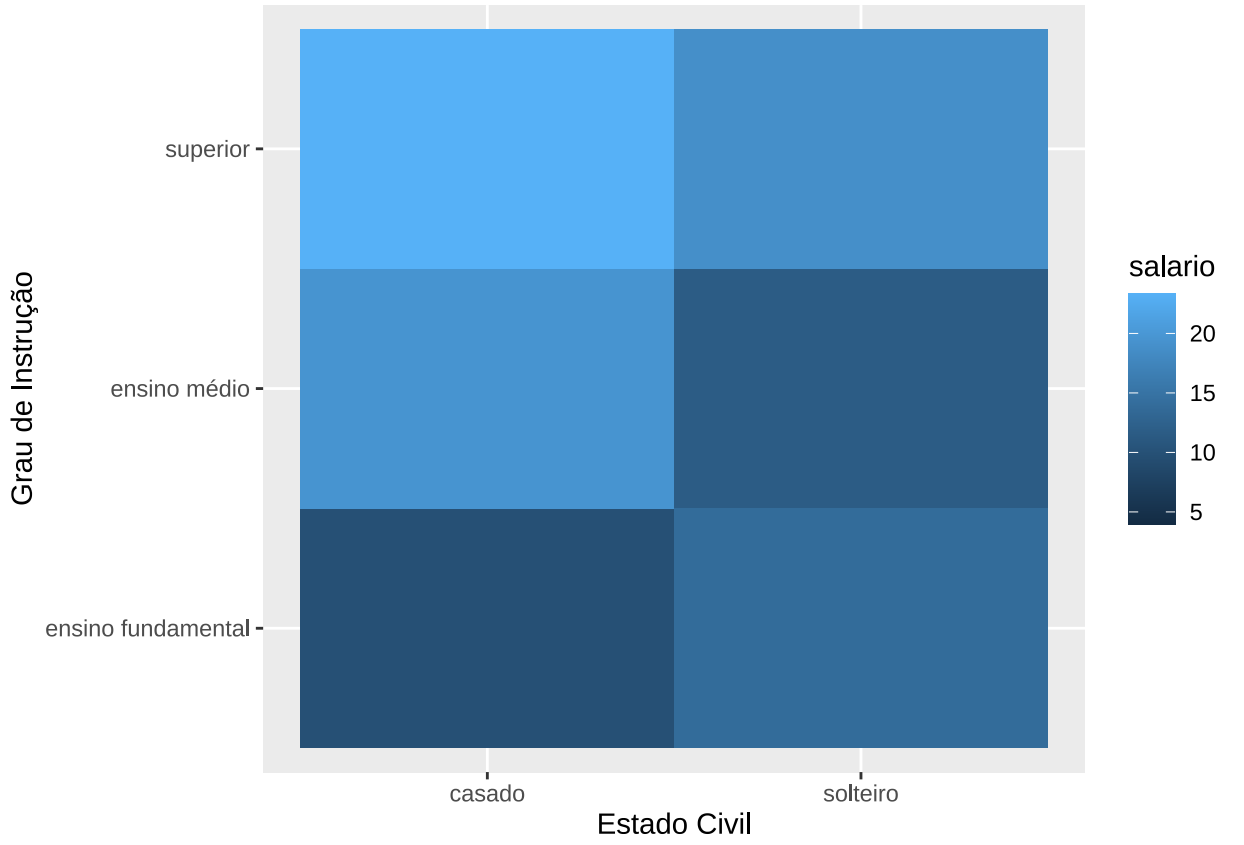
```
## Cross-Tabulation, Row Proportions
## Grau_de_instrucao * regiao
## Data Frame: salarios
##
## -------------------- -------- ------------ ------------ ------------ -------------
##                       regiao      capital     interior        outra         Total
##    Grau_de_instrucao
##   ensino fundamental             4 (33.3%)    3 (25.0%)    5 (41.7%)   12 (100.0%)
##         ensino médio             5 (27.8%)    7 (38.9%)    6 (33.3%)   18 (100.0%)
##             superior             2 (33.3%)    2 (33.3%)    2 (33.3%)    6 (100.0%)
##                Total            11 (30.6%)   12 (33.3%)   13 (36.1%)   36 (100.0%)
## -------------------- -------- ------------ ------------ ------------ -------------
```

```
summarytools::ctable(x = factor(salarios$n_filhos),
       y = salarios$estado_civil,
       prop = "r")
```

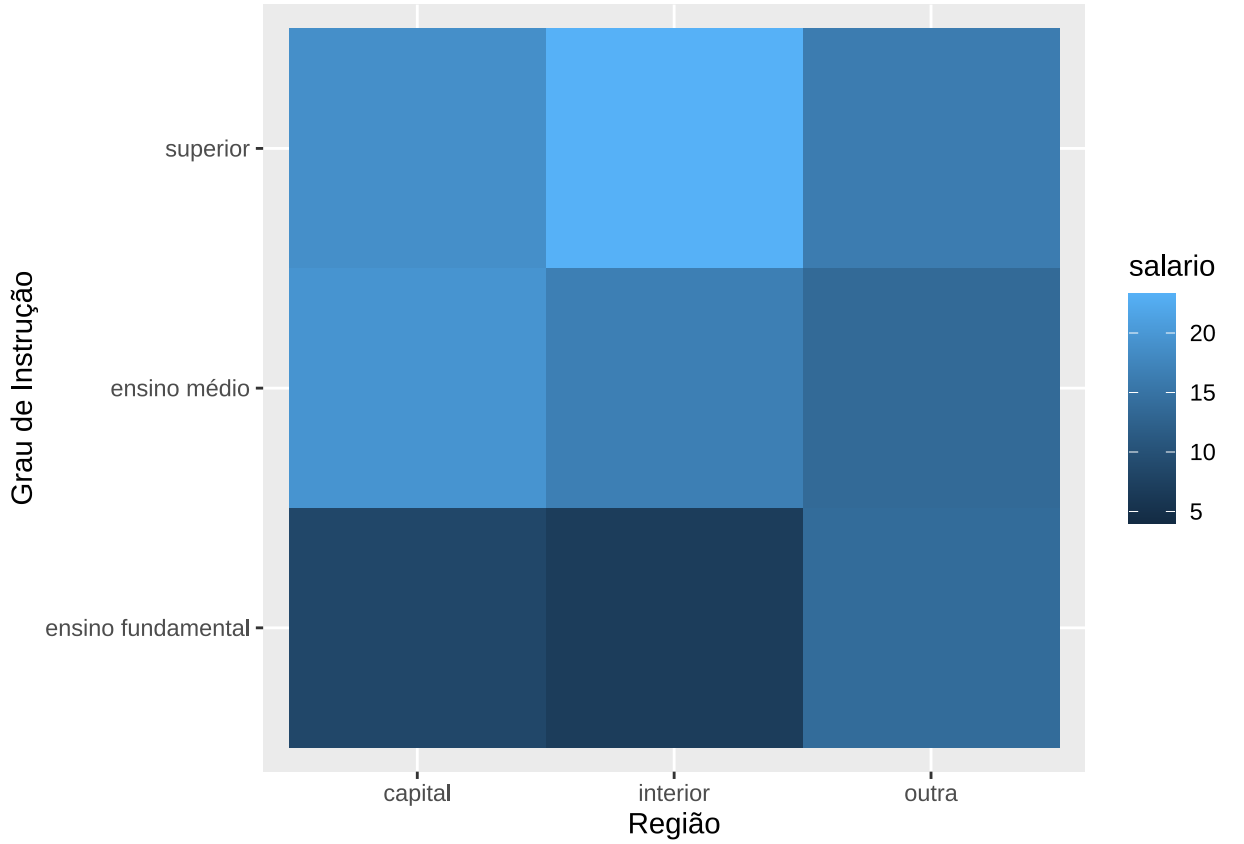
```
## Cross-Tabulation, Row Proportions
## factor(salarios$n_filhos) * estado_civil
##
## --------------------------- -------------- ------------- ------------- -------------
##                               estado_civil        casado      solteiro         Total
```

```
##   factor(salarios$n_filhos)
##                           0                  4 (100.0%)    0 (  0.0%)    4 (100.0%)
##                           1                  5 (100.0%)    0 (  0.0%)    5 (100.0%)
##                           2                  7 (100.0%)    0 (  0.0%)    7 (100.0%)
##                           3                  3 (100.0%)    0 (  0.0%)    3 (100.0%)
##                           5                  1 (100.0%)    0 (  0.0%)    1 (100.0%)
##                       <NA>                   0 (  0.0%)   16 (100.0%)   16 (100.0%)
##                       Total                 20 ( 55.6%)   16 ( 44.4%)   36 (100.0%)
## -------------------------- -------------- ------------- ------------- -------------
```

```
salarios %>% ggplot(aes(x = estado_civil, y = Grau_de_instrucao, fill = salario)) + geom_tile() + xlab(
```




```
salarios %>% ggplot(aes(x = regiao, y = Grau_de_instrucao, fill = salario)) + geom_tile() + xlab('Regiã
```



##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()
```

```
## 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
```

É possível ver pelo critério de skewness discutido em aula, que o valor de 0.6 para assimetria, nos faz interpretar essa distribução como levemente assimétrica, com cauda à direita.

Em decorrência desta assimetria, observamos que média e mediana apresentam valores distintos, com a média tendo valor levemente superior, o que aponta que os valores mais distantes do centro da distribuição puxam o valor da média pra cima.

Já a mediana por ser uma estatística de ordem, não é sensível a dados que apresentam alto valor na distribuição, o que é reforçado por seu valor levemente mais baixo que a média.

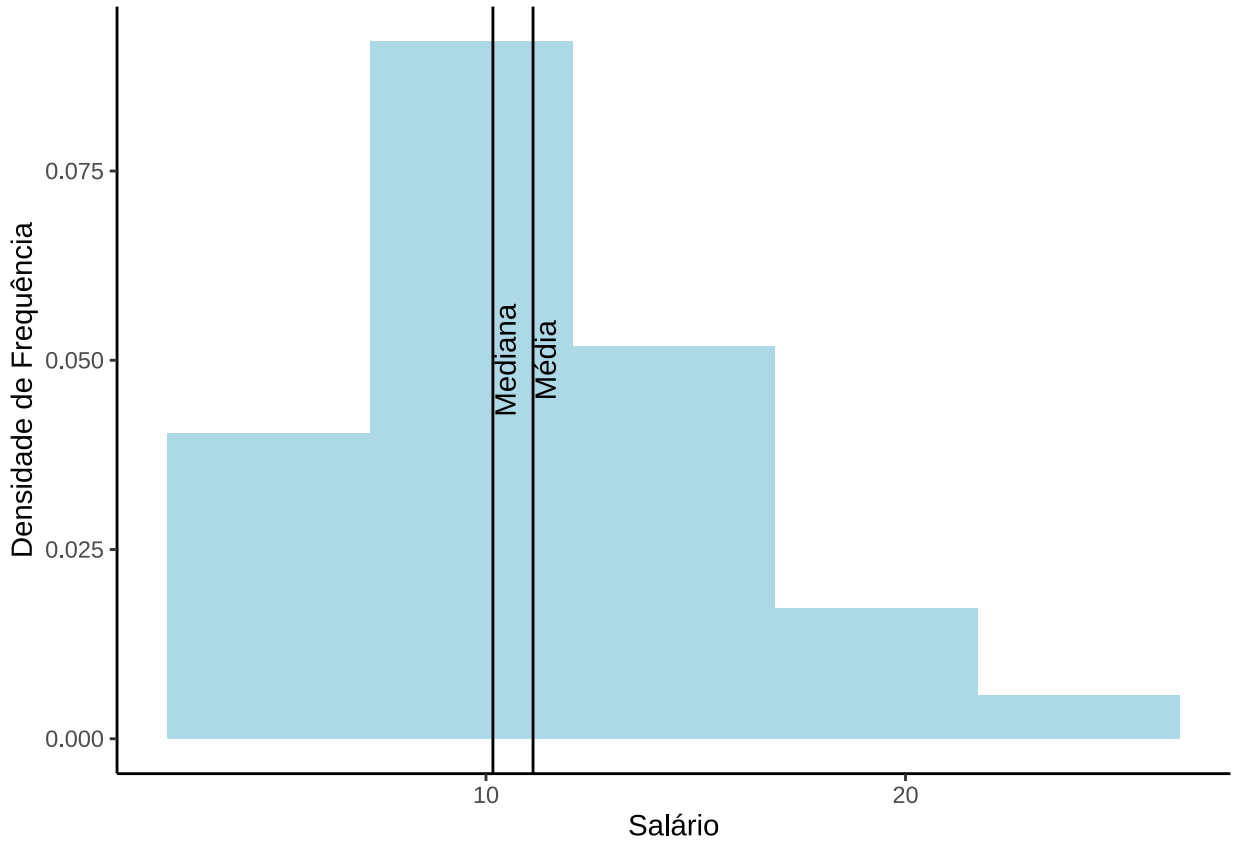
Reparem que se tivéssemos outliers nesta distribuição a média se descolaria ainda mais da mediana, pois estaria totalmente suscetível à contaminação.

##Análise visual da variável salário

Para realizar a análise visual da variável salários, seguimos o padrão de binarização recomendado pelos detentores dos dados. No entanto, reparem que se estivésses interessados em outras regras de binarização seríamos livres para escolher.

Devemos sempre ter em mente que escolher bins para aproximar a distribuição de probabilidade de uma determinada variável nos incorre em perda de informação, uma vez que estamos tratando como indiferentes eventos distintos para estarem em grupos contíguous do histograma.

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



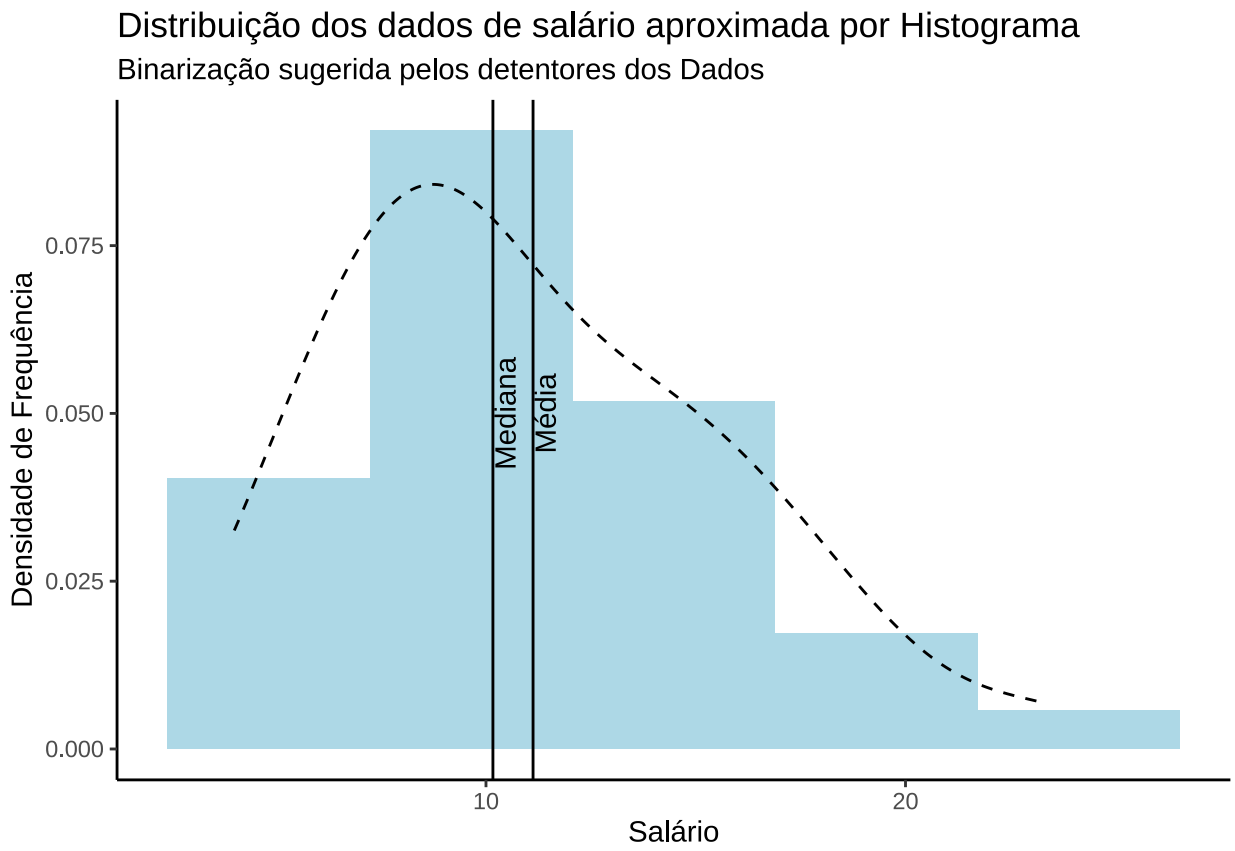
Reparem por essa visualização que a leitura visual nos leva a conclusões semelhantes a nossa leitura das estatísticas descritas, como por exemplo:

1. Leve assimetria com cauda à direita
2. Centralidade dos dados calculada pela média sofre leve contaminação dos valores mais distantes do centro da distribuição
3. Por mais que sejam poucas observações os dados não apresentam dispersão elevada, tendo a maioria dos dados concentrada próxima ao centro da distribuição.

É importante dizer, que o tamanho da perda de informação, ao aproximar a distribuição por um histograma, será proporcional ao espaço que o histograma deixa de preencher como distância da distribuição original dos dados.

Por mais que a estimativa por kernel não seja a distribuição original dos dados, ela tende a ser mais próxima da mesma. Logo, temos uma certa leitura aproximada do tamanho de informação perdida com a análise que segue.

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



Poderíamos também considerar outras regra de binarização levando em consideração regras disponíveis na literatura, como a regra de Freedman-Diaconis, bem como a regra de Sturge, como segue:

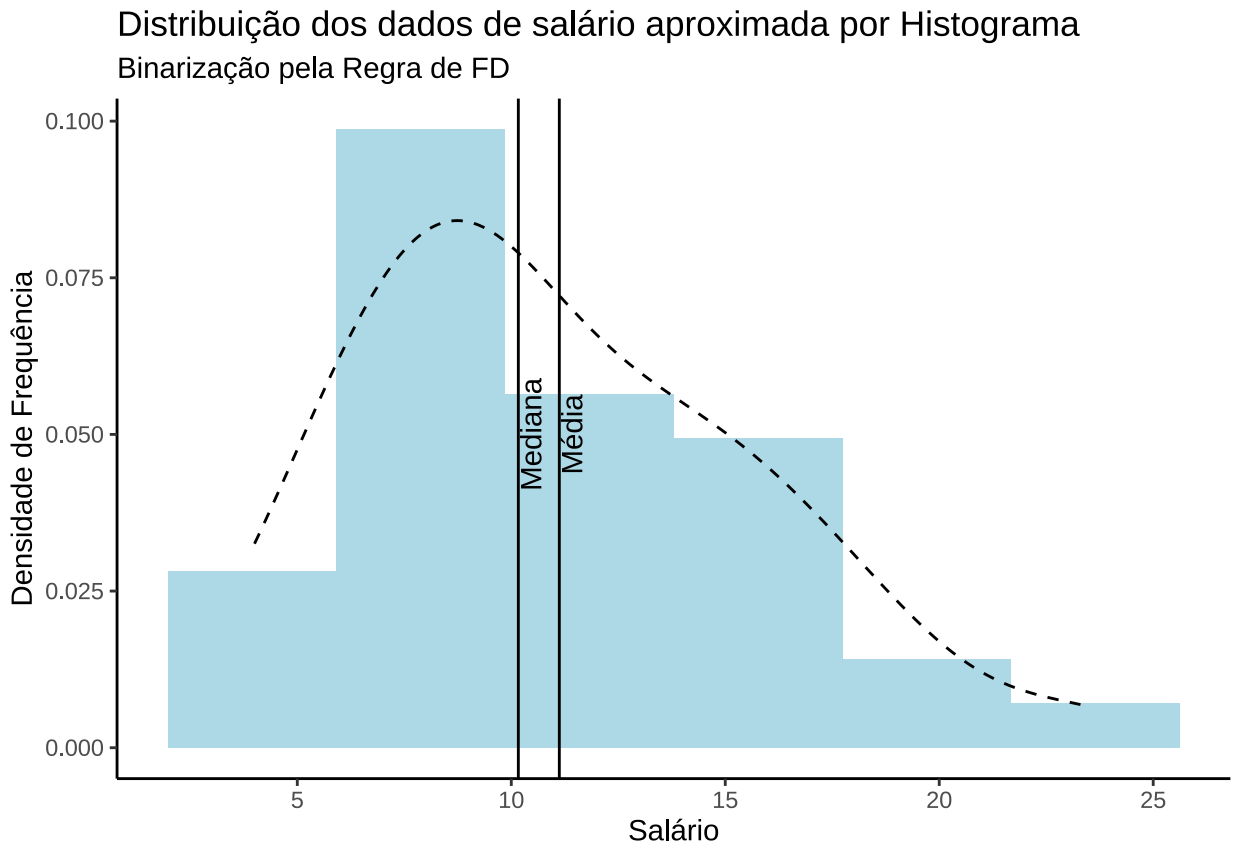
```
fd <- function(x) {
  n <-length(x)
  return((2*IQR(x))/n^(1/3))
}
```

```
sr <- function(x) {
  n <-length(x)
  return((3.49*sd(x))/n^(1/3))
}
```

Como visto em aula, a definição do intervalo do bin pela regra de Freedman-Diaconis leva em consideração o intervalo interquartil dos dados, o que impede com que eventuais outliers tenham influência na definição da amplitude do intervalo do bin.

Enquanto a regra de Sturge leva em consideração a dispersão da distribuição para definir a amplitude. Em geral, a regra de Sturge é mais recomendada quando o autor tem alguma evidência de que a distribuição dos dados se aproximará de uma distribuiçao normal, pelo menos no casso assintótico, isto é, quando a amostra dos dados é grande o suficiente.

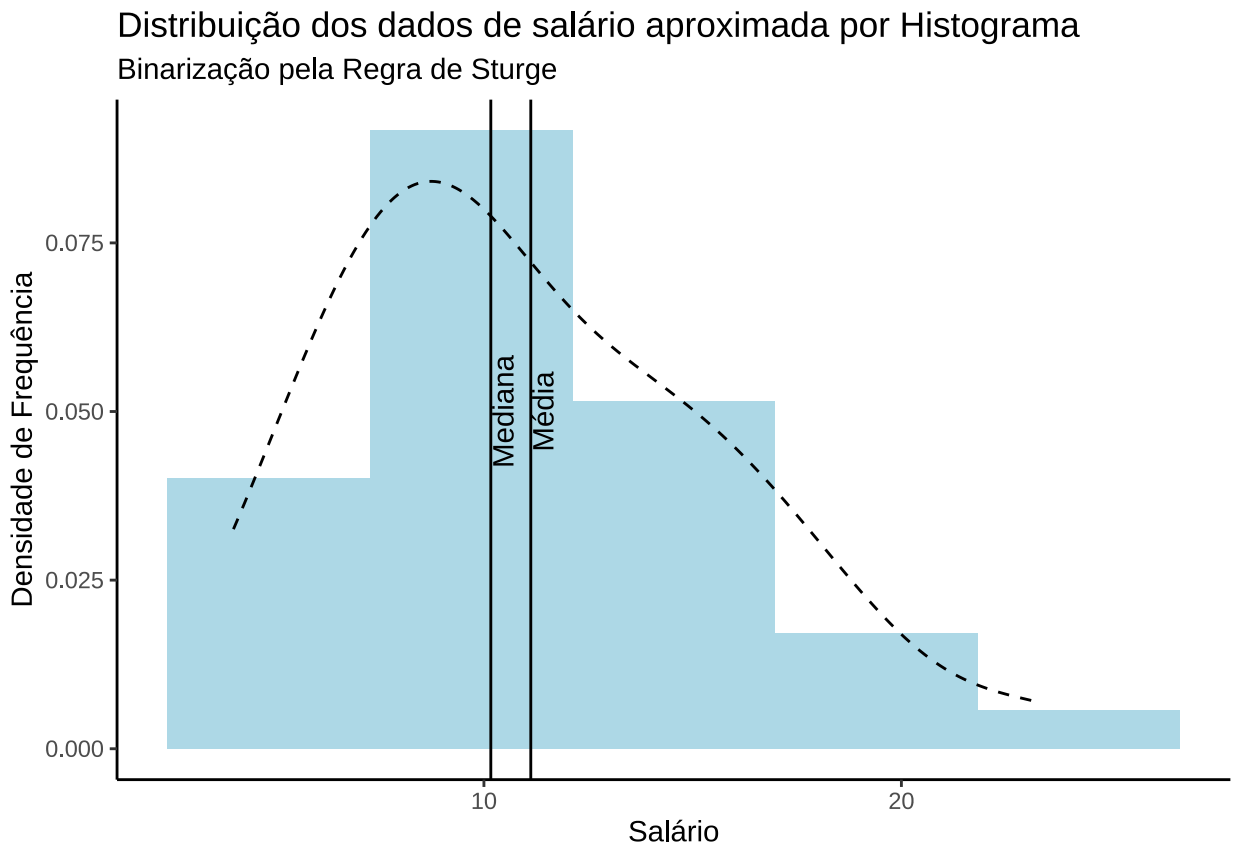
```
salarios %>% dplyr::select(salario) %>% ggplot(aes(x=salario))+geom_histogram(aes(y = after_stat(density
```



Após a aplicar a regra de Freedman-Diaconis, nosso histograma apresentou um bin a mais, o que pode ser justificado pela extração de um maior nível de detalhes da distribuição dos dados.

O que é interessante é que o padráo de assimetria fica ainda mais evidente com a Moda da distribuição aproximada claramente à esquerda mediana e da média.

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



Enquanto que ao utilizarmos a regra de Sturge, extraímos exatamente o mesmo padrão sugerido pelos autores, o que nos levanta a desconfiança de eles terem utilizado exatamente a mesma função para realizar a escolha de bins.

##Analisando a matriz de correlação da sub-amostra dos indivíduos que preencheram a variável de filhos

Vamos filtrar apenas os indivíduos de um determinado setor de uma empresa que tenham preenchido os dados de filhos no banco de dados. Aqui é importante destacar, que ao fazer esse filtro, muda-se o espaço amostral, esses valores não devem ser comparados com as tabelas anteriores.

```
kable(cor(salarios %>% dplyr::filter(!is.na(n_filhos)) %>% dplyr::select(salario, n_filhos, idade_anos))
```

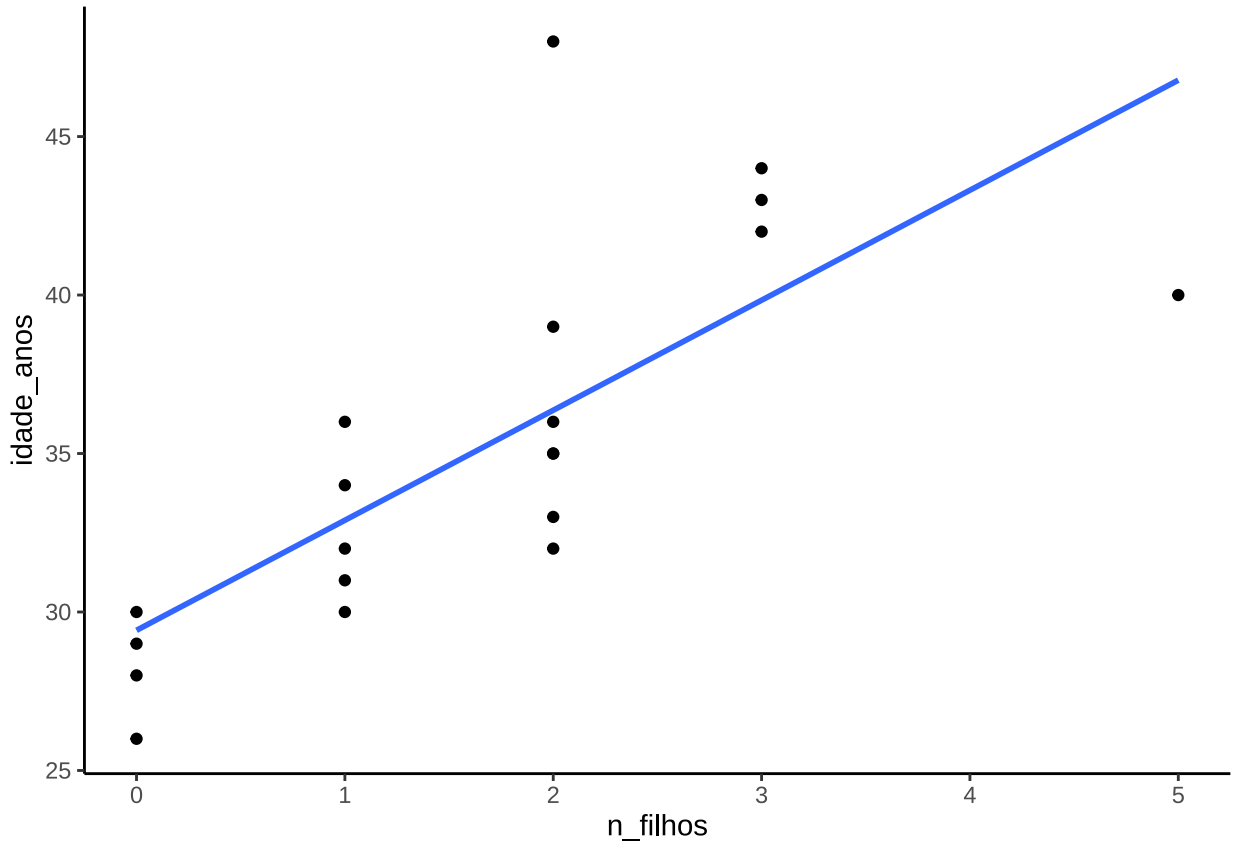
	salario	n_filhos	idade_anos
salario	1.0000000	0.3580647	0.4816920
n_filhos	0.3580647	1.0000000	0.7465385
idade_anos	0.4816920	0.7465385	1.0000000

É fácil ver que quanto maior a idade dos funcionários maior a quantidade de filhos. Relação não tão direta quando o assunto são as comparações entre salário e idade, ou salário e número de filhos.

Podemos a partir daí, contruir um scatterplot entre as variáveis idade e quantidade de filhos a fim de ver a relação positiva de crescimento propocional entre as variáveis, como segue:

```
salarios %>% dplyr::filter(!is.na(n_filhos)) %>% dplyr::select(idade_anos, n_filhos) %>% ggplot(aes(x=n
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


idade_anos
45
40
35
30
25
0
1
2
3
4
5
n_filhos