

Análise descritiva de uma base de dados

Otto Tavares

21 February, 2025

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 tidyverse, 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 disponibilizada no livro texto dos autores Bussab e Moretim. Vamos importá-la e imprimir as primeiras observações para conhecimento das variáveis.

```
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()
```

```
## # 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, enquanto 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 extraírmolas 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      0.00     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      0.00     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 nenhuma 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 = "t")
```

```
## Cross-Tabulation, Total Proportions
```

```
## Grau_de_instrucao * regiao
```

```
## Data Frame: salarios
```

```
##
```

```
## -----
##           regiao      capital      interior      outra      Total
## Grau_de_instrucao
## ensino fundamental      4 (11.1%)      3 ( 8.3%)      5 (13.9%)     12 ( 33.3%)
## ensino médio            5 (13.9%)      7 (19.4%)      6 (16.7%)     18 ( 50.0%)
## superior                2 ( 5.6%)      2 ( 5.6%)      2 ( 5.6%)      6 ( 16.7%)
## 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 = "t")
```

```
## Cross-Tabulation, Total Proportions
```

```
## factor(salarios$n_filhos) * estado_civil
```

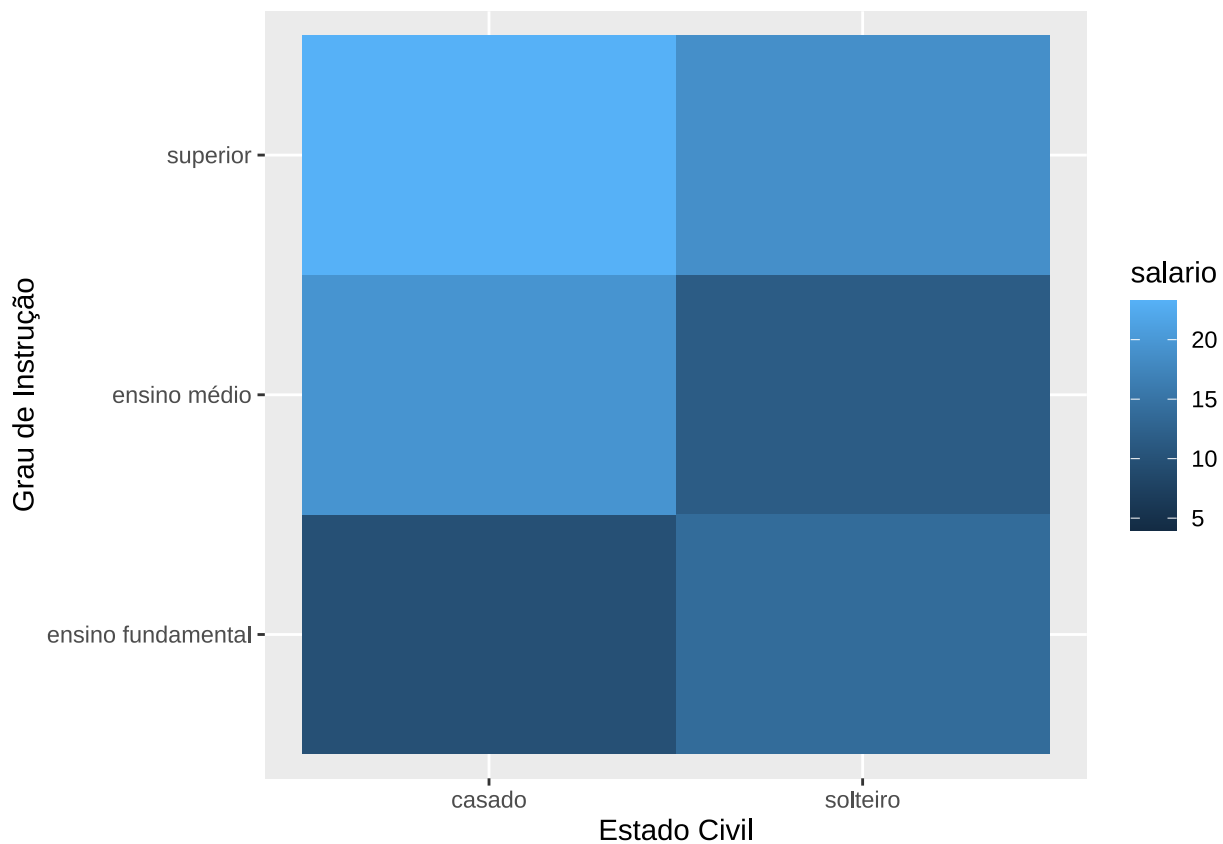
```
##
```

```
## -----
##           estado_civil      casado      solteiro      Total
## factor(salarios$n_filhos)
## 0                        4 (11.1%)      0 ( 0.0%)      4 ( 11.1%)
## 1                        5 (13.9%)      0 ( 0.0%)      5 ( 13.9%)
## 2                        7 (19.4%)      0 ( 0.0%)      7 ( 19.4%)
## 3                        3 ( 8.3%)      0 ( 0.0%)      3 (  8.3%)
## 5                        1 ( 2.8%)      0 ( 0.0%)      1 (  2.8%)
## <NA>                     0 ( 0.0%)     16 (44.4%)     16 ( 44.4%)
## Total                   20 (55.6%)     16 (44.4%)     36 (100.0%)
## -----
```

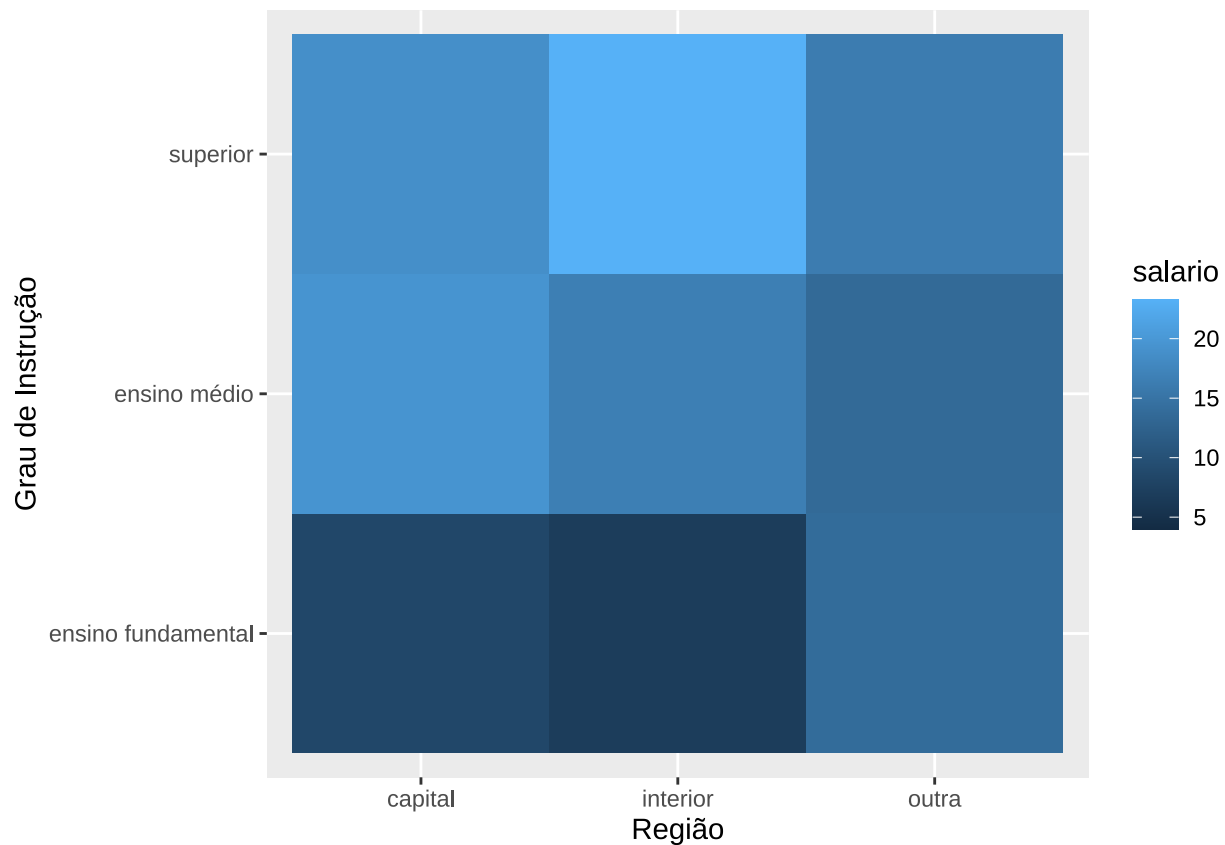
```
#Convertendo solteiro para 0 e casado para 1 com variável binária e calculando o coeficiente de correlação
cor.test(x = salarios$salario,
        y = as.numeric(as.character(factor(salarios$estado_civil, levels = c("solteiro","casado"), labels = c(0,1)))))
```

```
##
## Pearson's product-moment correlation
##
## data: salarios$salario and as.numeric(as.character(factor(salarios$estado_civil, levels = c("solteiro","casado"), labels = c(0,1))))
## t = 1.4894, df = 34, p-value = 0.1456
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
## -0.08822049 0.53271009
## sample estimates:
## cor
## 0.2474883
```

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



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



Análise descritiva e de histogramas de uma variável contínua

Já para a variável salários, podemos analisar a centralidade dos dados, dispersão, 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 distribuição 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.

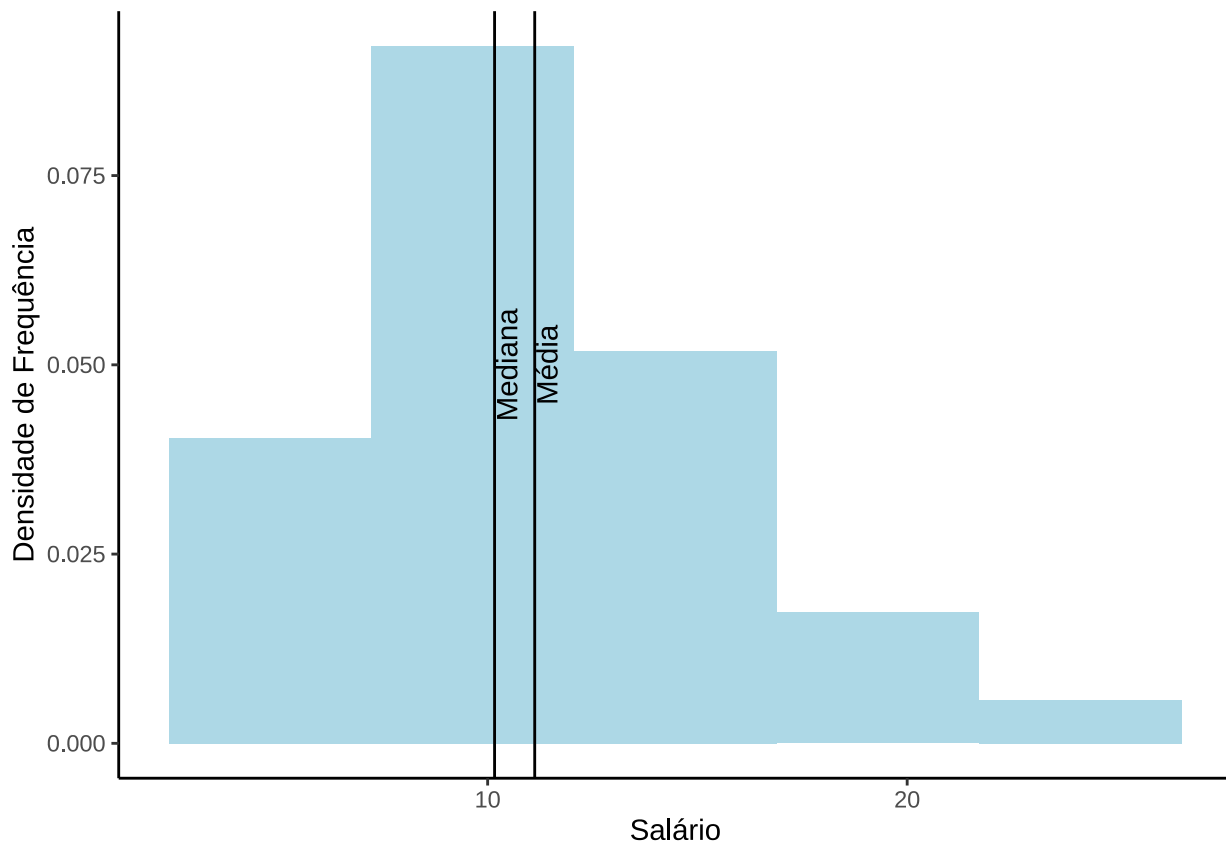
Reparem que se tivéssemos outliers nesta distribuição a média se deslocaria ainda mais da mediana, pois estaria totalmente suscetível à contaminação.

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 estivéssemos 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 contíguos do histograma.

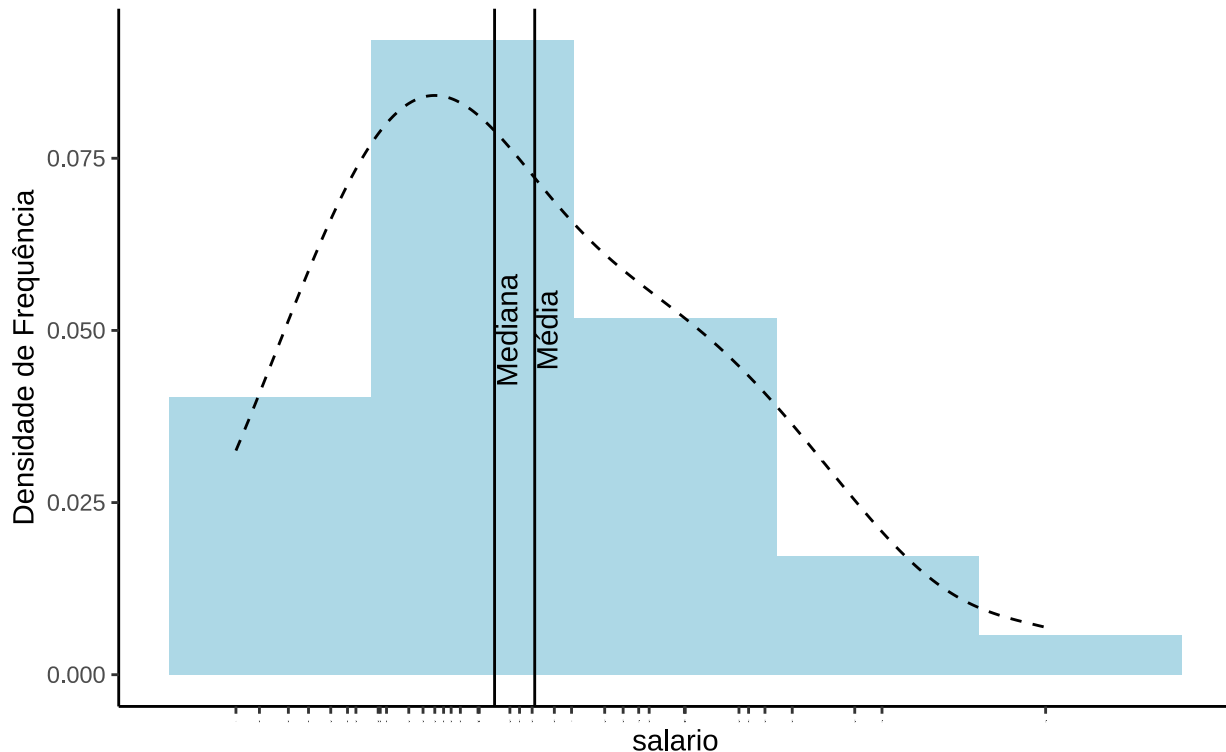
```
salarios %>% dplyr::select(salario) %>% ggplot(aes(x=salario))+geom_histogram(aes(y = after_stat(density)))
```



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

Distribuição dos dados de salário aproximada por Histograma

Binarização sugerida pelos detentores dos Dados



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.

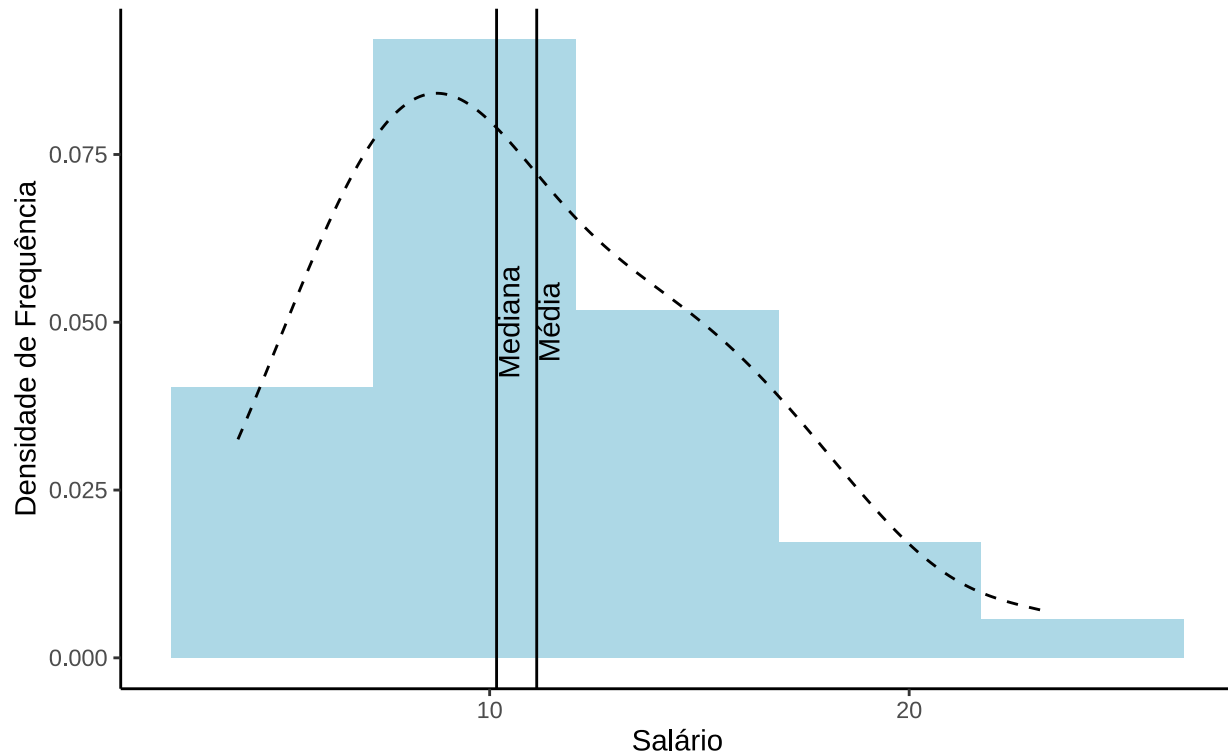
É importante dizer, que o tamanho da perda de informação, ao aproximar a distribuição por um histograma, será proporcional ao espaço que o histograma deixa de preencher como distância da distribuição original dos dados.

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
```

Distribuição dos dados de salário aproximada por Histograma

Binarização sugerida pelos detentores dos Dados



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))
}
```

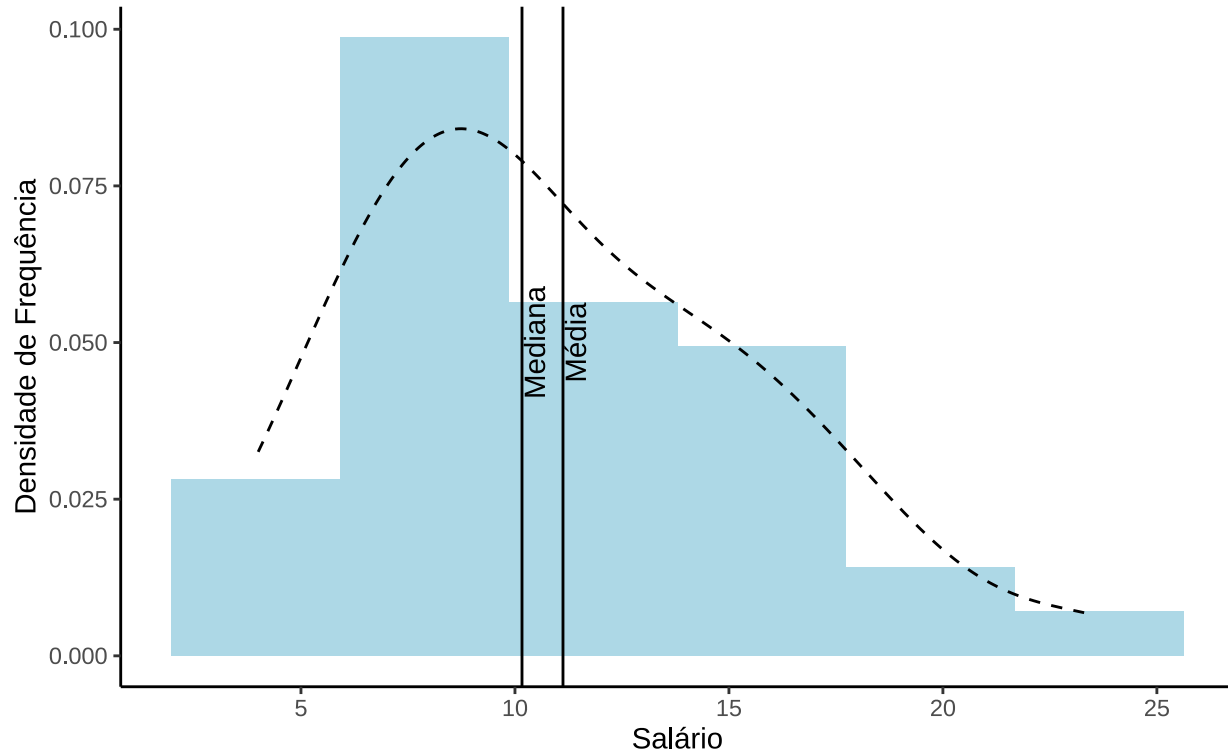
Como visto em aula, a definição do intervalo do bin pela regra de Freedman-Diaconis leva em consideração o intervalo interquartil dos dados, o que impede com que eventuais outliers tenham influência na definição da amplitude do intervalo do bin.

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 distribuição normal, pelo menos no caso 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(densit,
```

Distribuição dos dados de salário aproximada por Histograma

Binarização pela Regra de FD



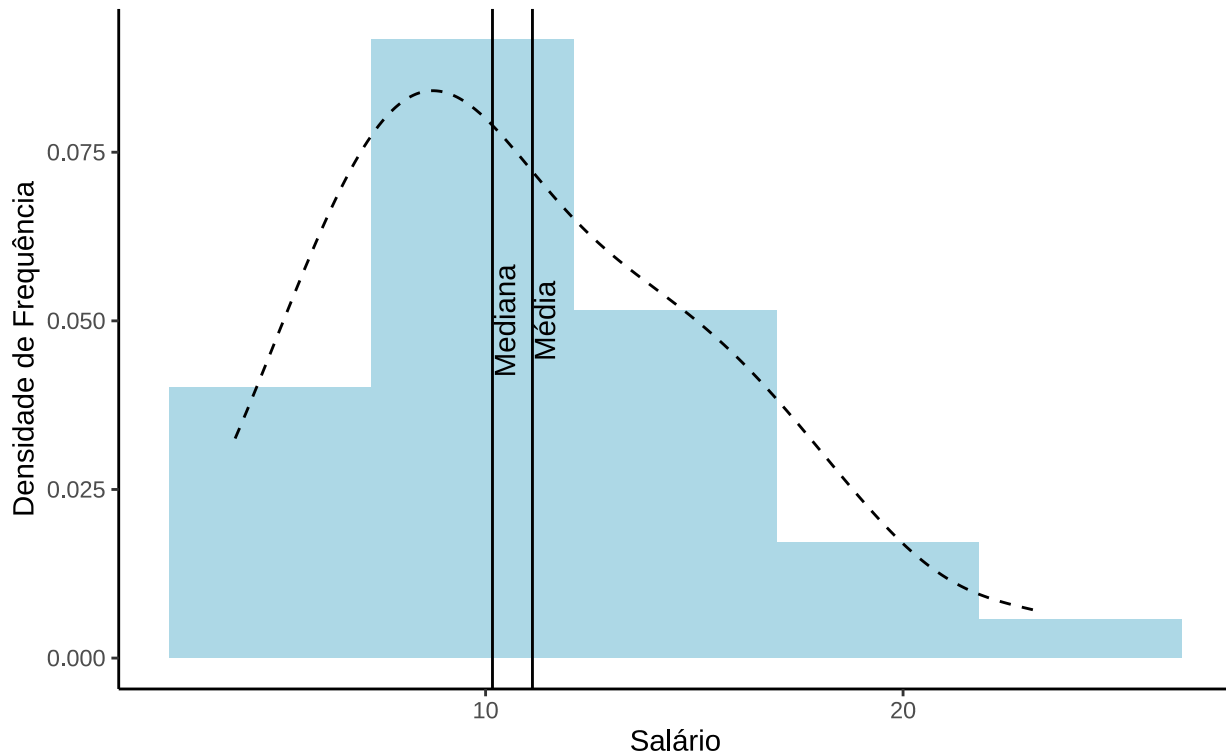
Após a aplicação da regra de Freedman-Diaconis, nosso histograma apresentou um bin a mais, o que pode ser justificado pela extração de um maior nível de detalhes da distribuição dos dados.

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

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

Distribuição dos dados de salário aproximada por Histograma

Binarização pela Regra de Sturge



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 dos indivíduos desta empresa, buscando relação entre idade e salário.

Vamos calcular a correlação entre as variáveis de salário e idade, para verificar se há evidência de relação entre as variáveis, e o quão intensa é essa relação.

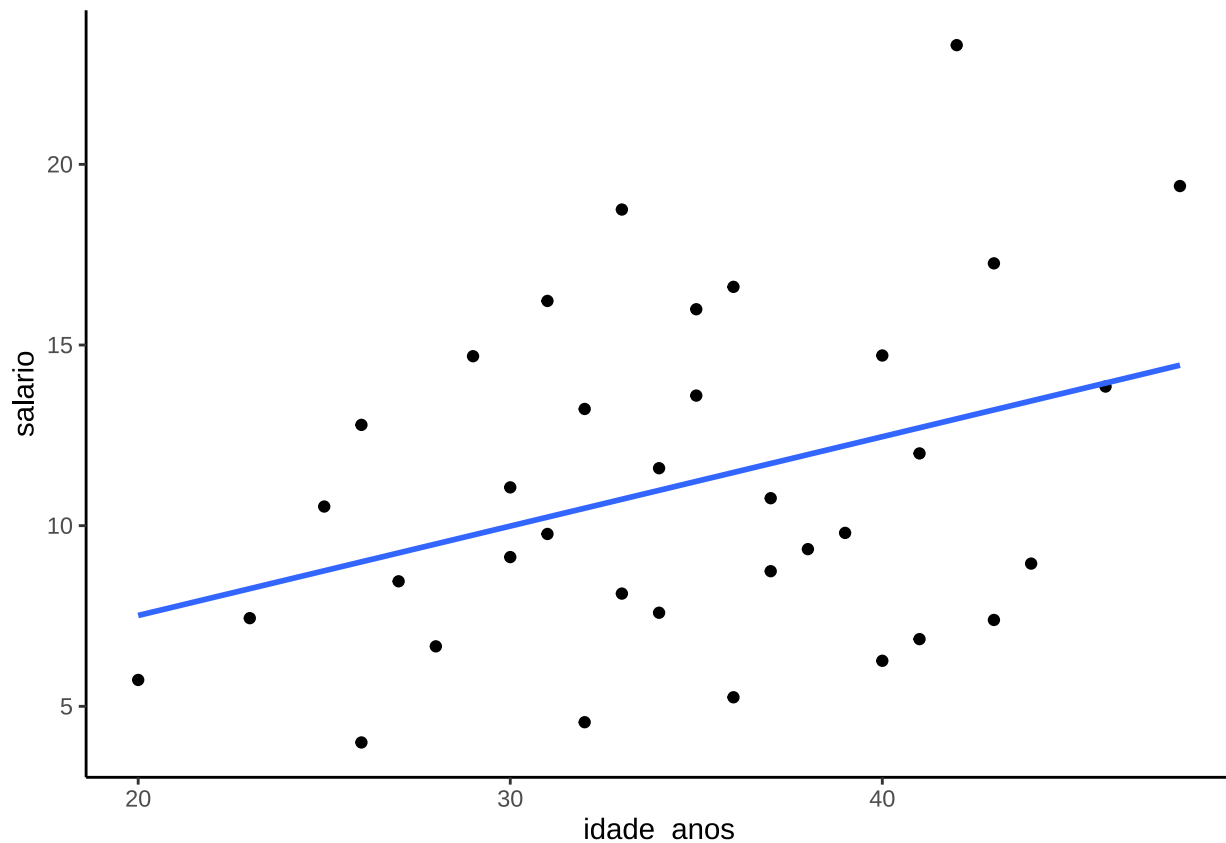
```
kable(cor(salarios %>% dplyr::select(salario, idade_anos)))
```

	salario	idade_anos
salario	1.0000000	0.3633622
idade_anos	0.3633622	1.0000000

O scatter plot pode nos auxiliar na visualização da proporcionalidade de crescimento entre ambas as variáveis.

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

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



Nessa etapa, também trazemos o número de filhos para verificar se também há uma relação entre salário, idade e número de filhos.

```
kable(cor(salarios %>% dplyr::select(salario, n_filhos, idade_anos)))
```

	salario	n_filhos	idade_anos
salario	1.0000000	NA	0.3633622
n_filhos	NA	1	NA
idade_anos	0.3633622	NA	1.0000000

O problema é que o não preenchimento da variável “n_filhos” nos impede de calcular as correlações que levam em conta essa variável. No entanto, deletar essas observações significaria mudar o espaço amostral de análise. Dessa forma, podemos utilizar o scatter plot entre as variáveis idade e salário, para nos auxiliar na visualização dessas relações, colorindo os pontos de acordo com o número de filhos dos indivíduos.

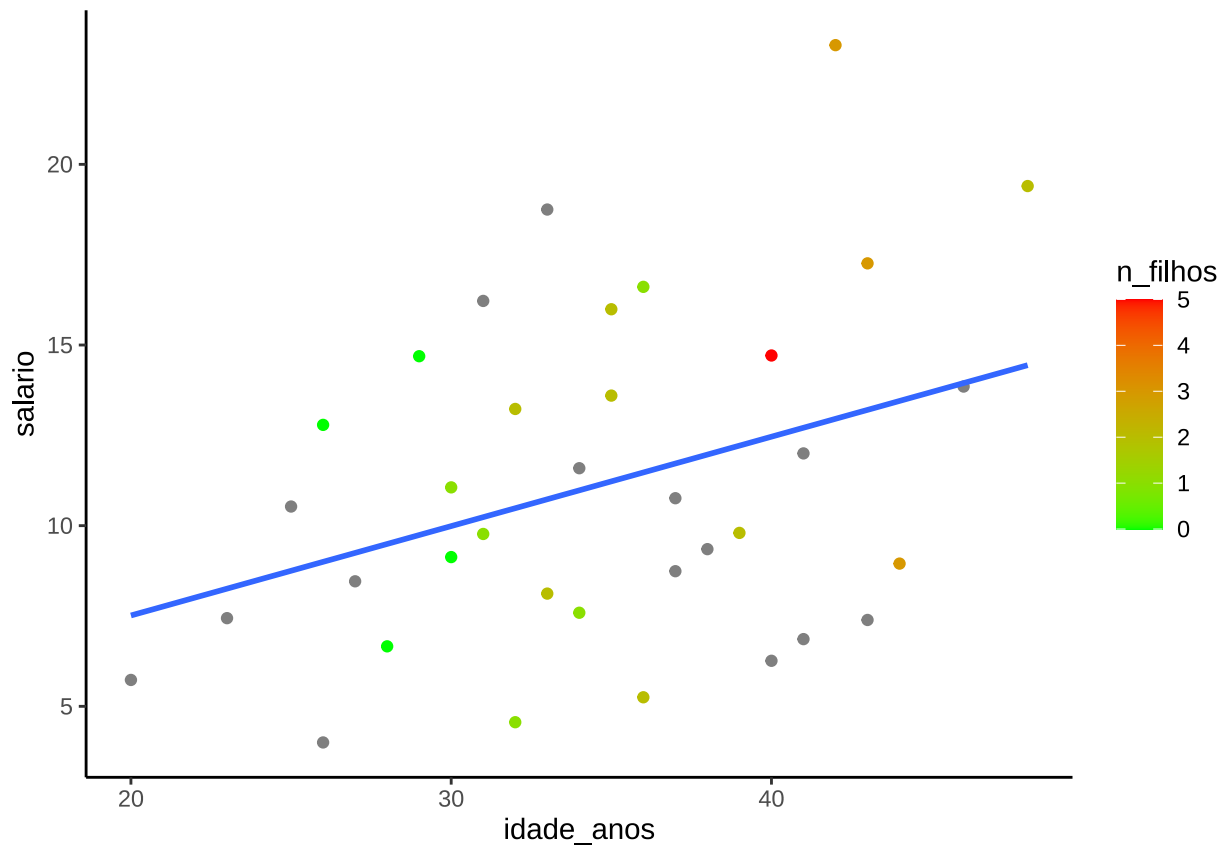
```
salarios %>% dplyr::select(idade_anos, salario, n_filhos) %>% ggplot(aes(x=idade_anos, y=salario, color=n_filhos))
```

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

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
```

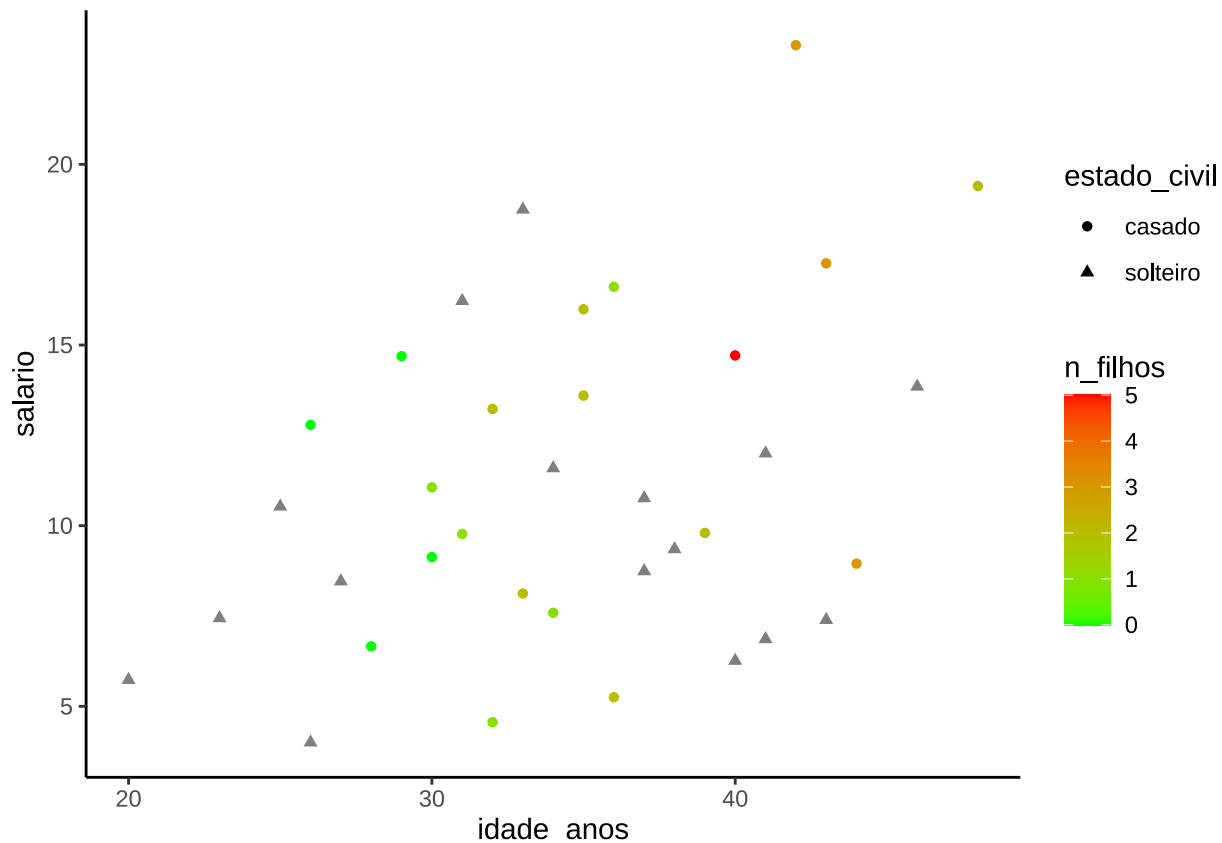
```
## i This can happen when ggplot fails to infer the correct grouping structure in
## the data.
```

```
## i Did you forget to specify a 'group' aesthetic or to convert a numerical
## variable into a factor?
```



Vimos na tabela de contingência que os solteiros não preencheram a variável de número de filhos. Podemos deixar essa distinção de forma clara no nosso gráfico com a mudança do shape entre as variáveis.

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



(Extra por curiosidade): 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

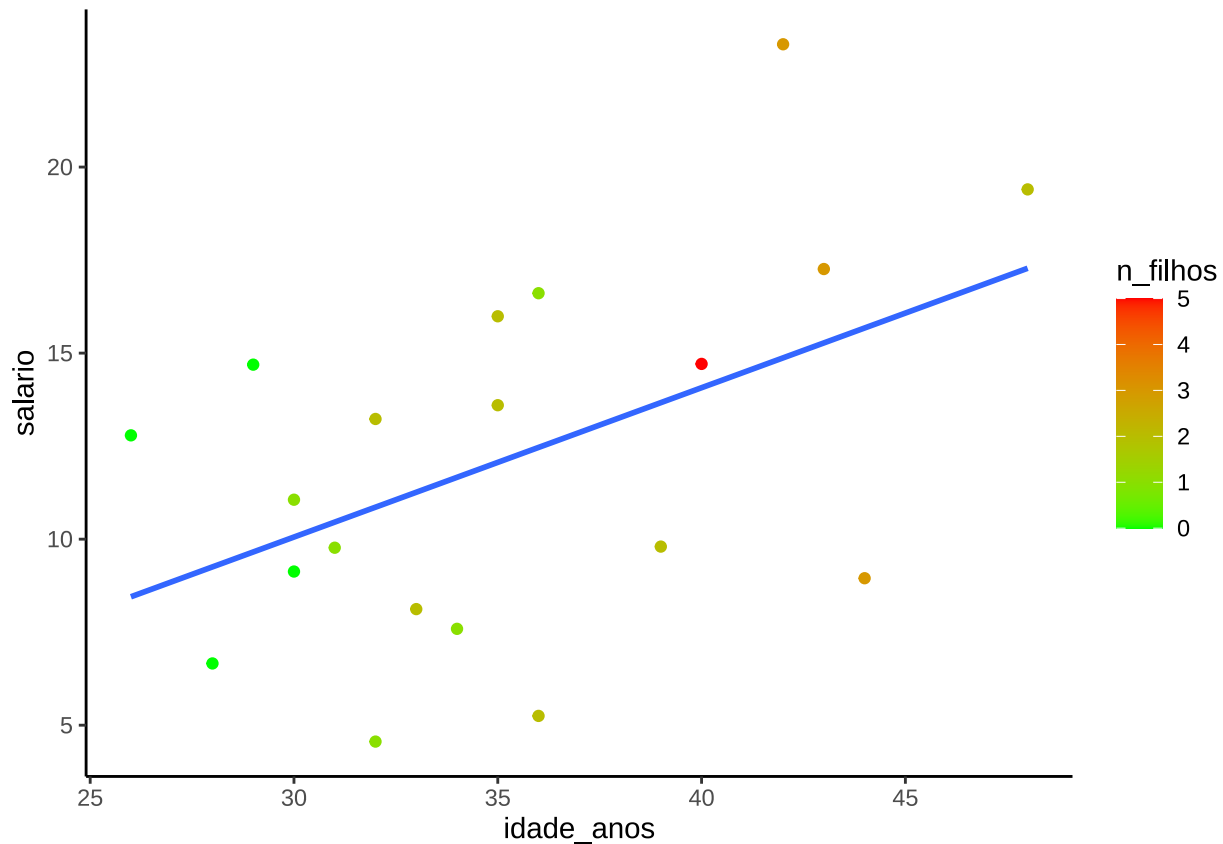
É fácil ver que quanto maior a idade dos funcionários maior a quantidade de filhos. Relação não tão direta quando o assunto são as comparações entre salário e idade, ou salário e número de filhos.

Podemos a partir daí, contruir um scatterplot entre as variáveis idade e salário, mas colorindo as observações com a quantidade de filhos de cada indivíduo, a fim de visualizar essas relações, como segue:

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

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

```
## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
## the data.
## i Did you forget to specify a 'group' aesthetic or to convert a numerical
## variable into a factor?
```



##Apêndice ### Lista com diversos gráficos de histograma acompanhados de estimativas para densidade variando a função kernel

```
hists.kernel <- lapply(c("gaussian", "epanechnikov"), function(krn) salarios %>% dplyr::select(salario)
```

Lista com diversos gráficos de histograma para testar a amplitude do bin

```
hists.bins <- lapply(1:10, function(bwdt) salarios %>% dplyr::select(salario) %>% ggplot(aes(x = salario)
```