Manipulação de Dados - Parte V

Paulo Henrique S. Guimarães

1 Análise exploratória de dados

A análise exploratória de dados (AED) não é um processo com um conjunto de regras rígidas. É uma parte importante de qualquer análise de dados de forma que a desenvolver uma compreensão dos seus dados.

Não há regras acerca de quais perguntas (indagações) você deve fazer para guiar sua pesquisa, entretanto há dois tipos de perguntas que são indispensáveis para fazer descobertas dentro de seus dados, tais como:

- Que tipo de variação ocorre dentro de minhas variáveis?
- Que tipo de covariação ocorre entre minhas variáveis?

1.1 Variação

A **variação** é a tendência à mudança dos valores de uma variável de uma medição para outra. Para tanto podemos buscar entender o padrão de variação visualizando a distribuição de valores das variáveis.

- variáveis categóricas - gráfico de barras
- variáveis numéricas - histograma (varável contínua).
- Exemplo:

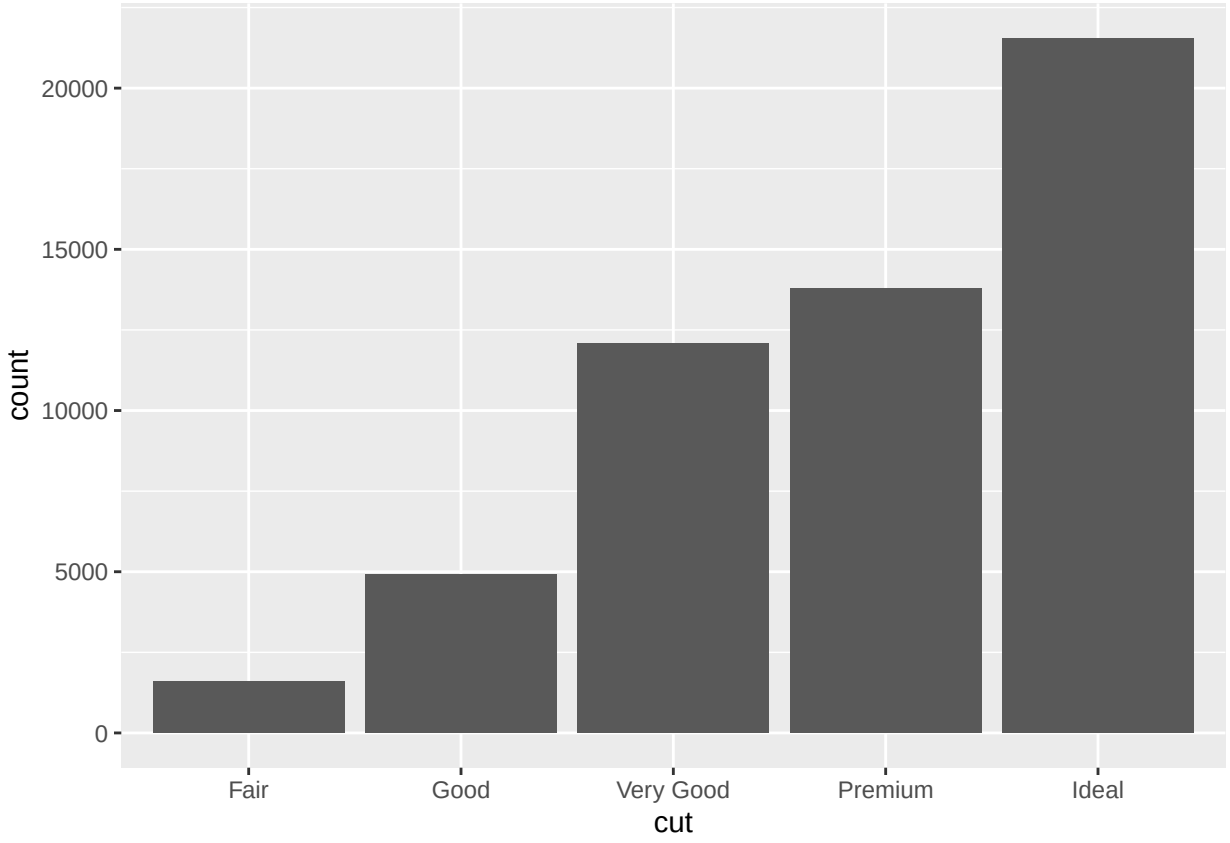
```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

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

```
library(ggplot2)

data("diamonds")

glimpse(diamonds)
```

```
## Observations: 53,940
## Variables: 10
## $ carat   <dbl> 0.23, 0.21, 0.23, 0.29, 0.31, 0.24, 0.24, 0.26, 0.22, ...
## $ cut     <ord> Ideal, Premium, Good, Premium, Good, Very Good, Very G...
## $ color   <ord> E, E, E, I, J, J, I, H, E, H, J, J, F, J, E, E, I, J, ...
## $ clarity <ord> SI2, SI1, VS1, VS2, SI2, VVS2, VVS1, SI1, VS2, VS1, SI...
## $ depth   <dbl> 61.5, 59.8, 56.9, 62.4, 63.3, 62.8, 62.3, 61.9, 65.1, ...
## $ table   <dbl> 55, 61, 65, 58, 58, 57, 57, 55, 61, 61, 55, 56, 61, 54...
## $ price   <int> 326, 326, 327, 334, 335, 336, 336, 337, 337, 338, 339,...
## $ x       <dbl> 3.95, 3.89, 4.05, 4.20, 4.34, 3.94, 3.95, 4.07, 3.87, ...
## $ y       <dbl> 3.98, 3.84, 4.07, 4.23, 4.35, 3.96, 3.98, 4.11, 3.78, ...
## $ z       <dbl> 2.43, 2.31, 2.31, 2.63, 2.75, 2.48, 2.47, 2.53, 2.49, ...
```

```r
# variável categórica

ggplot(data=diamonds) +
  geom_bar(mapping = aes(x = cut))
```

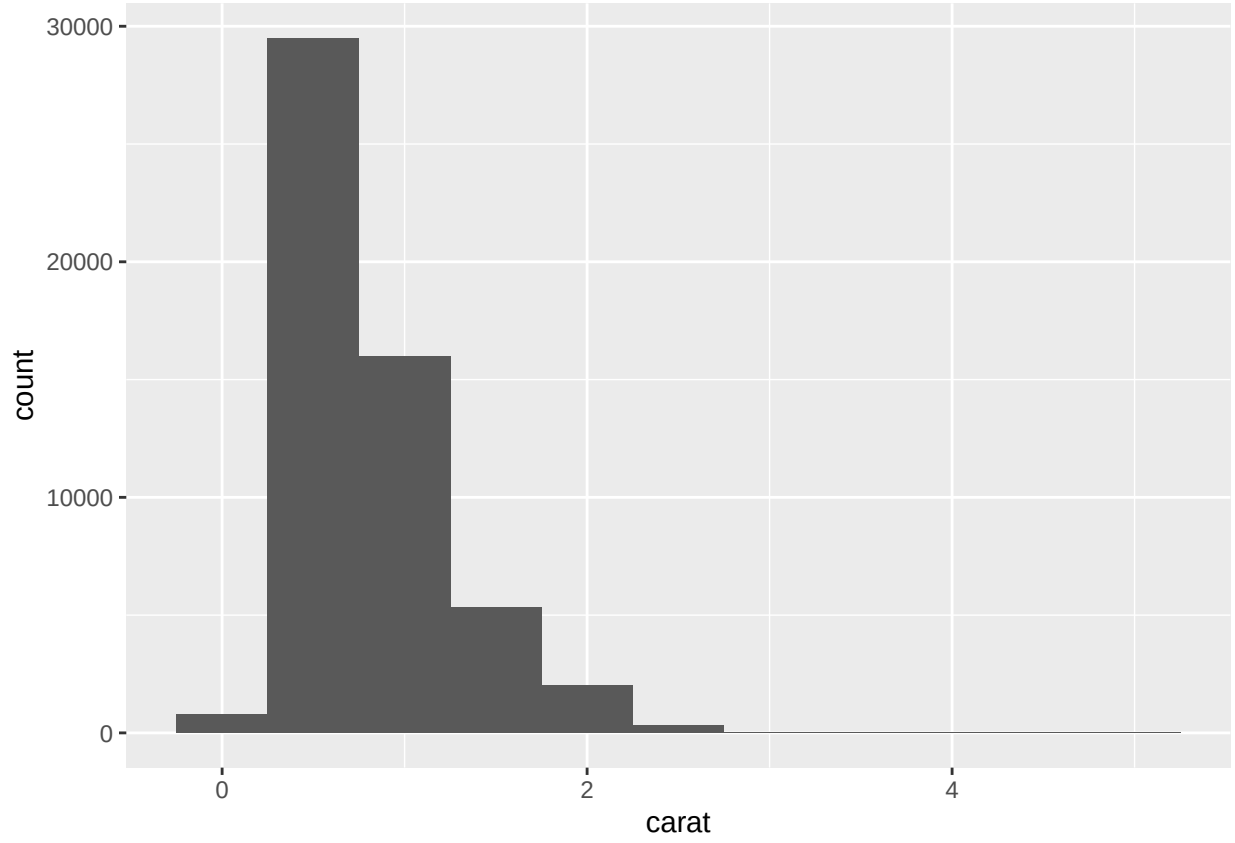



```r
diamonds %>%
  count(cut)
```

```
## # A tibble: 5 x 2
##   cut           n
##   <ord>     <int>
## 1 Fair       1610
```

```
## 2 Good       4906
## 3 Very Good 12082
## 4 Premium   13791
## 5 Ideal     21551
```

```r
# variável numérica - contínua

ggplot(data = diamonds)+
  geom_histogram(mapping = aes(x = carat), binwidth = 0.5)
```




```r
diamonds %>%
  count(cut_width(carat,0.5))
```

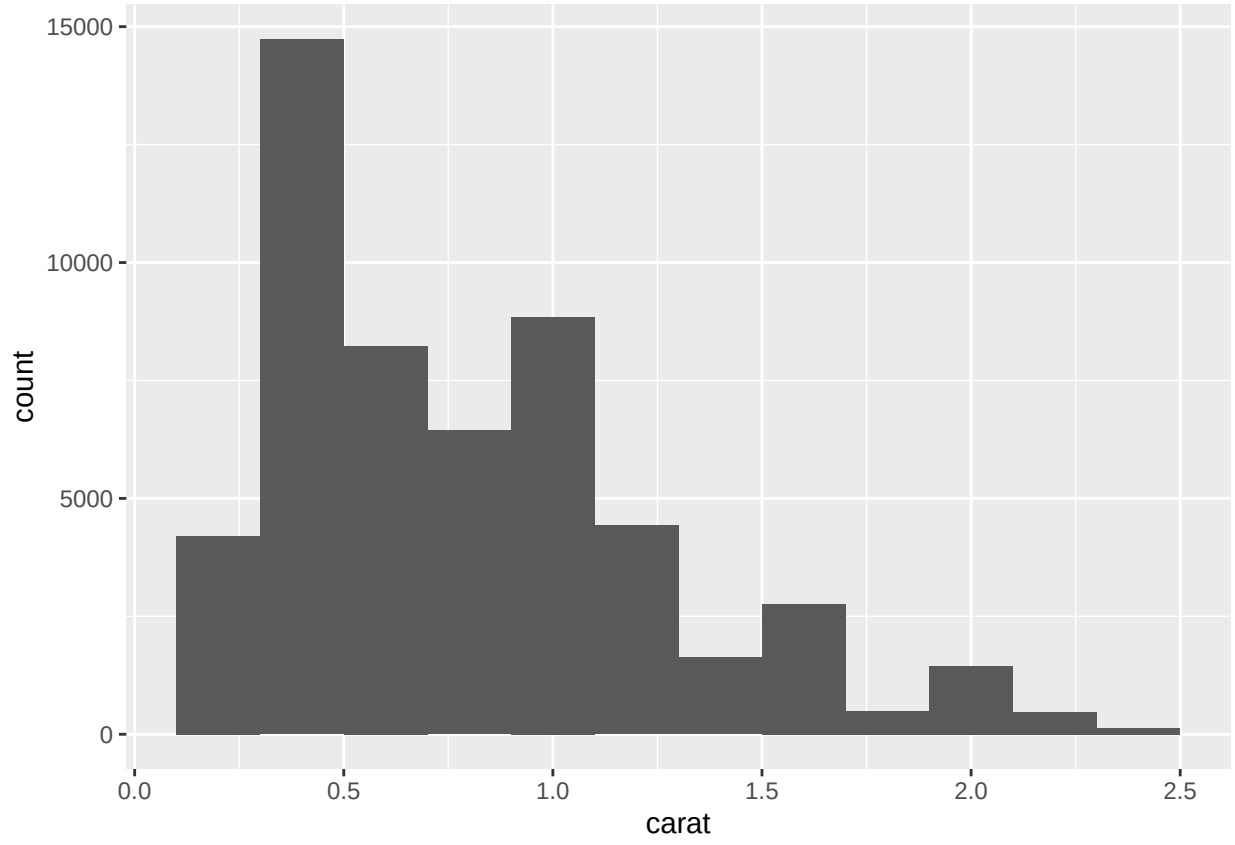
```
## # A tibble: 11 x 2
##    `cut_width(carat, 0.5)`     n
##    <fct>                   <int>
##  1 [-0.25,0.25]              785
##  2 (0.25,0.75]             29498
##  3 (0.75,1.25]             15977
##  4 (1.25,1.75]              5313
##  5 (1.75,2.25]              2002
##  6 (2.25,2.75]               322
##  7 (2.75,3.25]                32
##  8 (3.25,3.75]                 5
##  9 (3.75,4.25]                 4
```

```
## 10 (4.25,4.75]                 1
## 11 (4.75,5.25]                 1
```

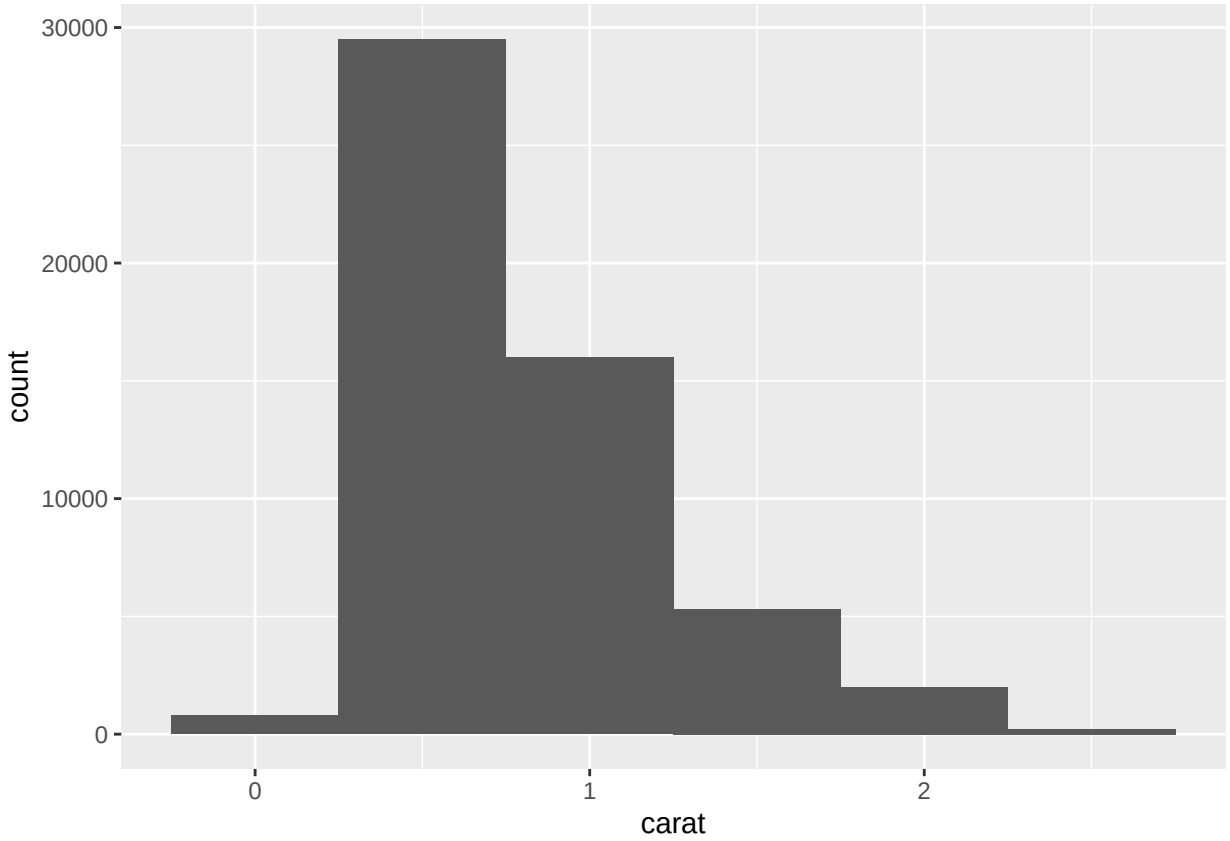
Podemos explorar nosso conjunto de dados utilizando “filtros” como o próximo exemplo que segue:

- Exemplo:

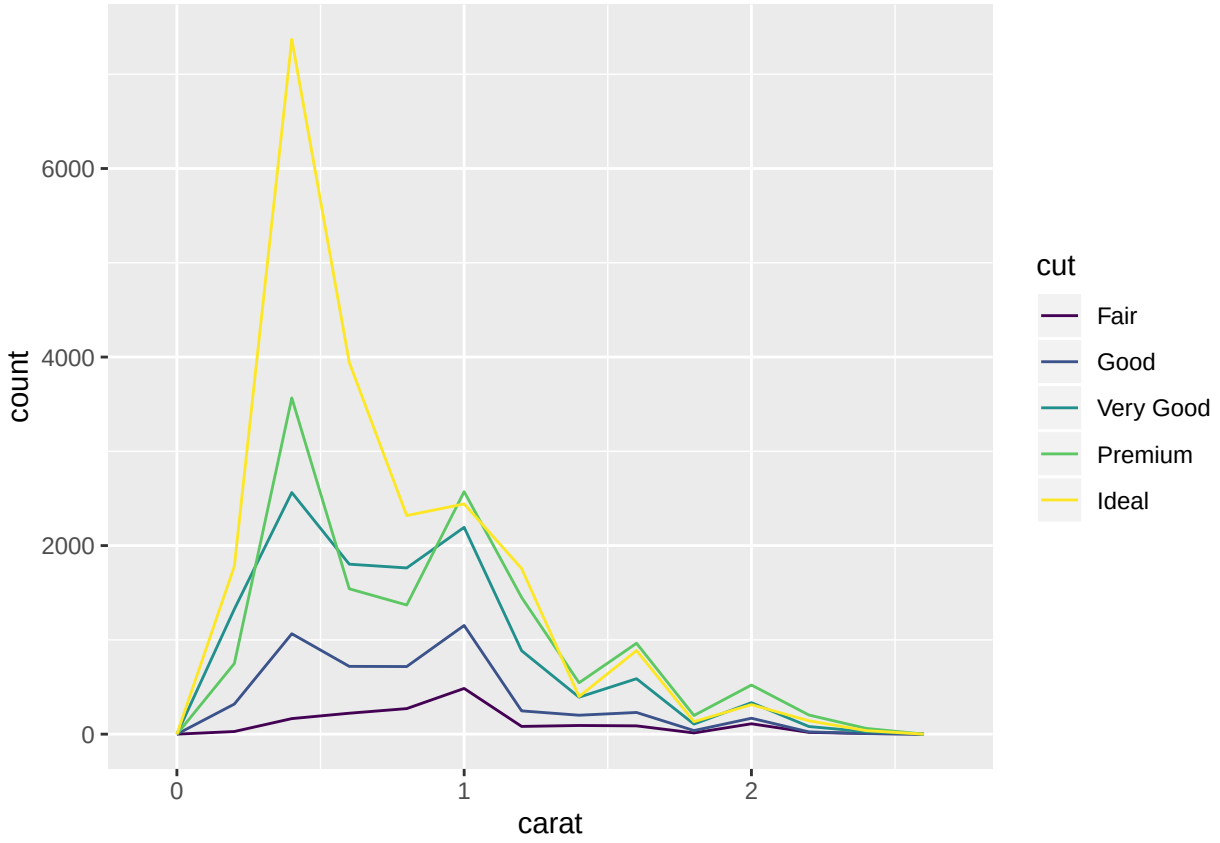
```
valor<- diamonds %>%
  filter(carat < 2.5)
ggplot(data = valor, mapping = aes(x = carat)) +
  geom_histogram(binwidth = 0.2)
```




```
## alterando a largura

ggplot(data = valor, mapping = aes(x = carat)) +
  geom_histogram(binwidth = 0.5)
```




```
###

ggplot(data = valor, mapping = aes(x = carat,color = cut)) +
  geom_freqpoly(binwidth = 0.2)
```

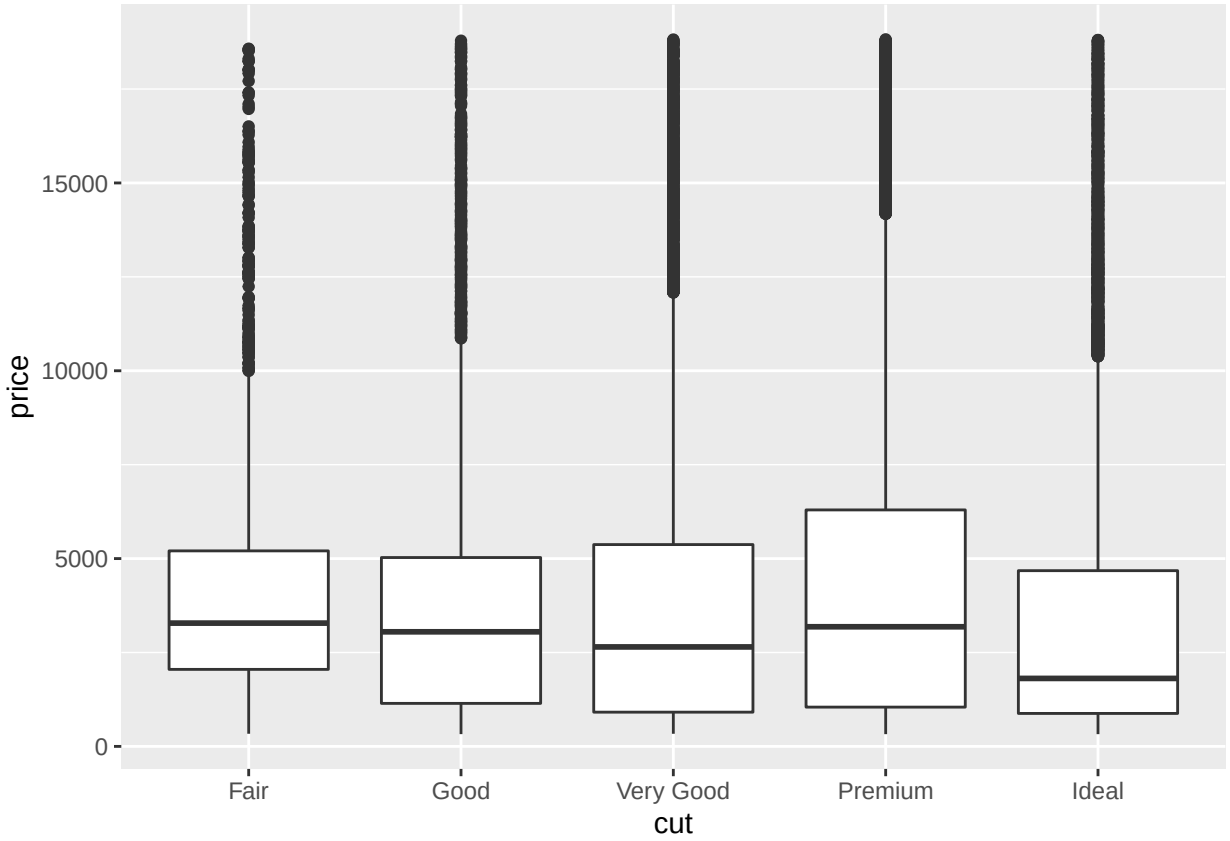


1.2 Covariação

A covariação descreve o comportamento entre as variáveis. Consiste na tendência que os valores de duas ou mais variáveis têm de variar juntas de maneira relacionada.

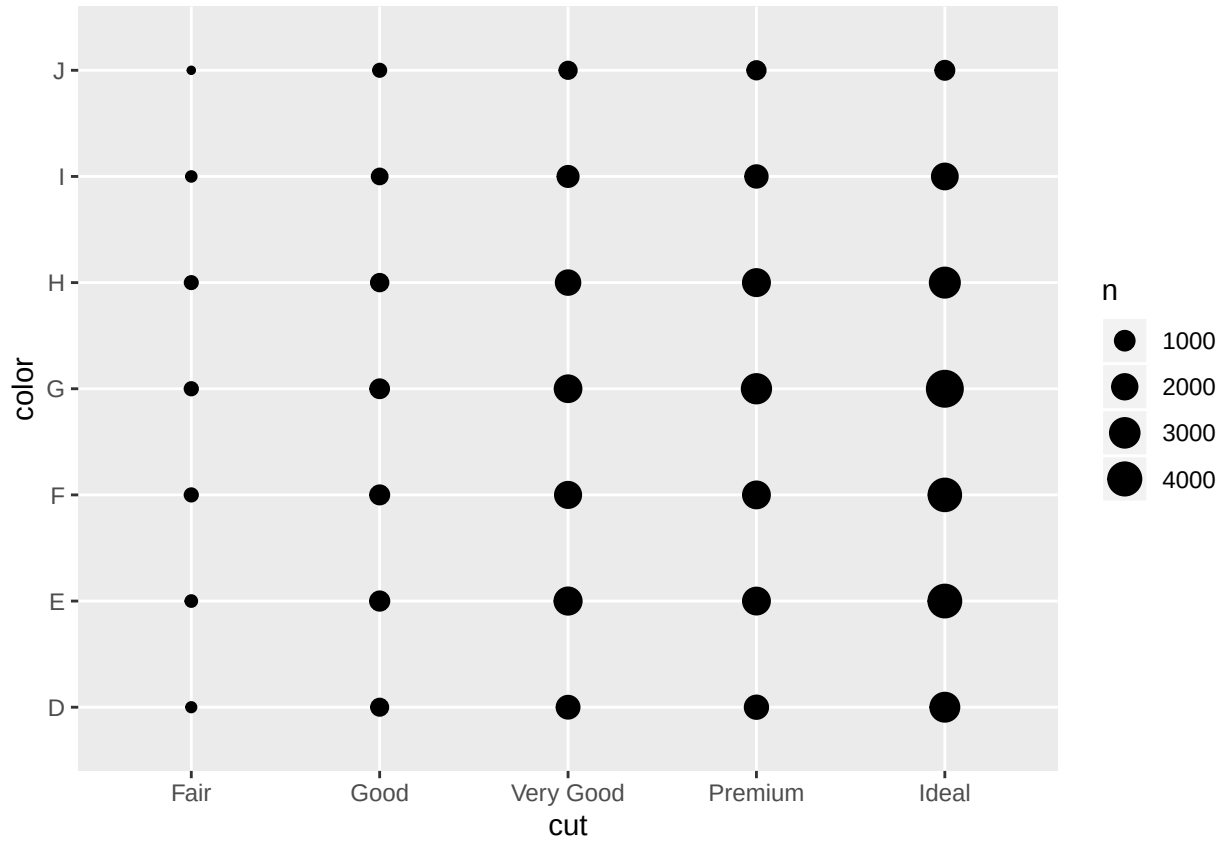
- Exemplo:

```
ggplot(data = diamonds, mapping = aes(x = cut, y = price)) +
  geom_boxplot()
```



Para visualializar a covariação entre variáveis categóricas precisamos contar o número de observações de cada combinação.

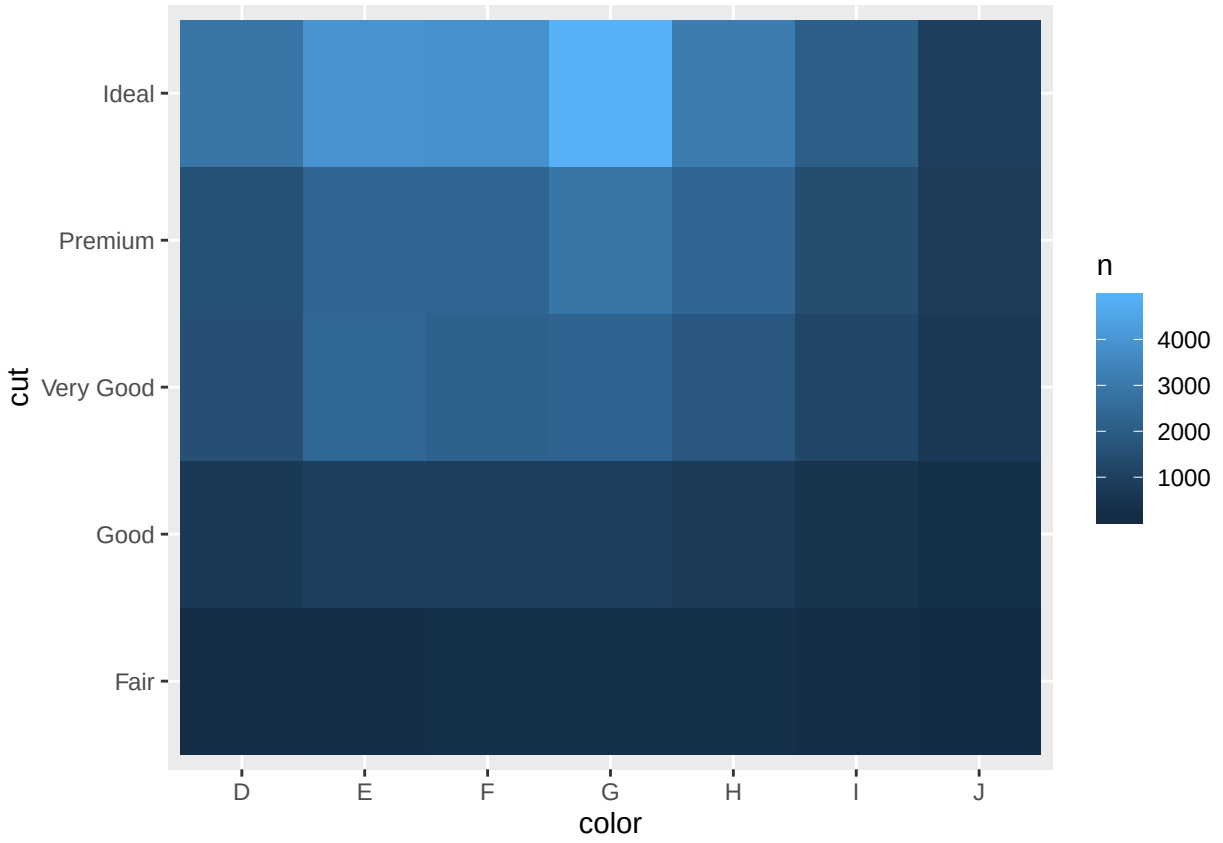
```
ggplot(data = diamonds) +
  geom_count(mapping = aes(x = cut, y = color))
```




```
# O tamanho de cada círculo denota quantas observações ocorreram em cada combinação de valores.

diamonds %>%
  count(color,cut)
```

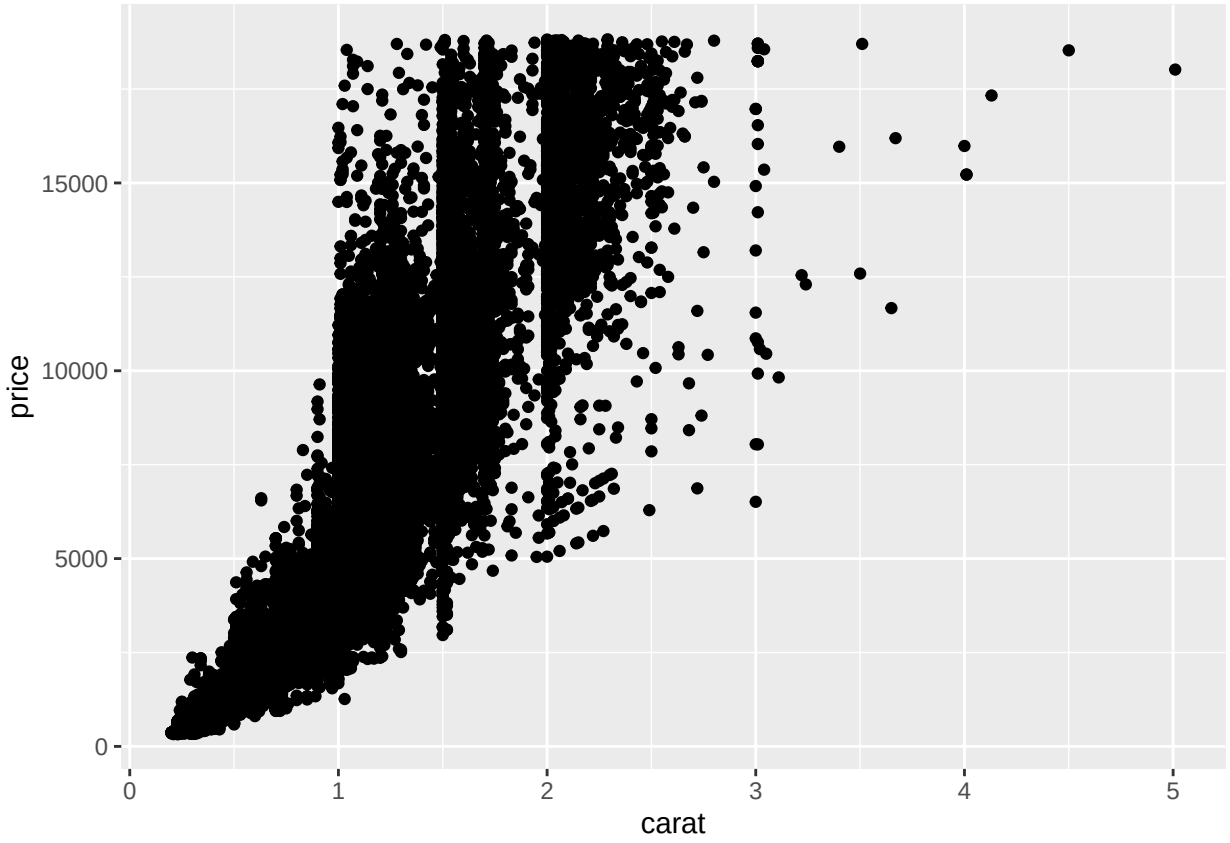
```
## # A tibble: 35 x 3
##    color cut           n
##    <ord> <ord>     <int>
##  1 D     Fair        163
##  2 D     Good        662
##  3 D     Very Good  1513
##  4 D     Premium    1603
##  5 D     Ideal      2834
##  6 E     Fair        224
##  7 E     Good        933
##  8 E     Very Good  2400
##  9 E     Premium    2337
## 10 E     Ideal      3903
## # ... with 25 more rows
```

```
##

diamonds %>%
  count(color,cut) %>%
  ggplot(mapping = aes(x = color, y = cut)) +
  geom_tile(mapping = aes(fill = n))
```

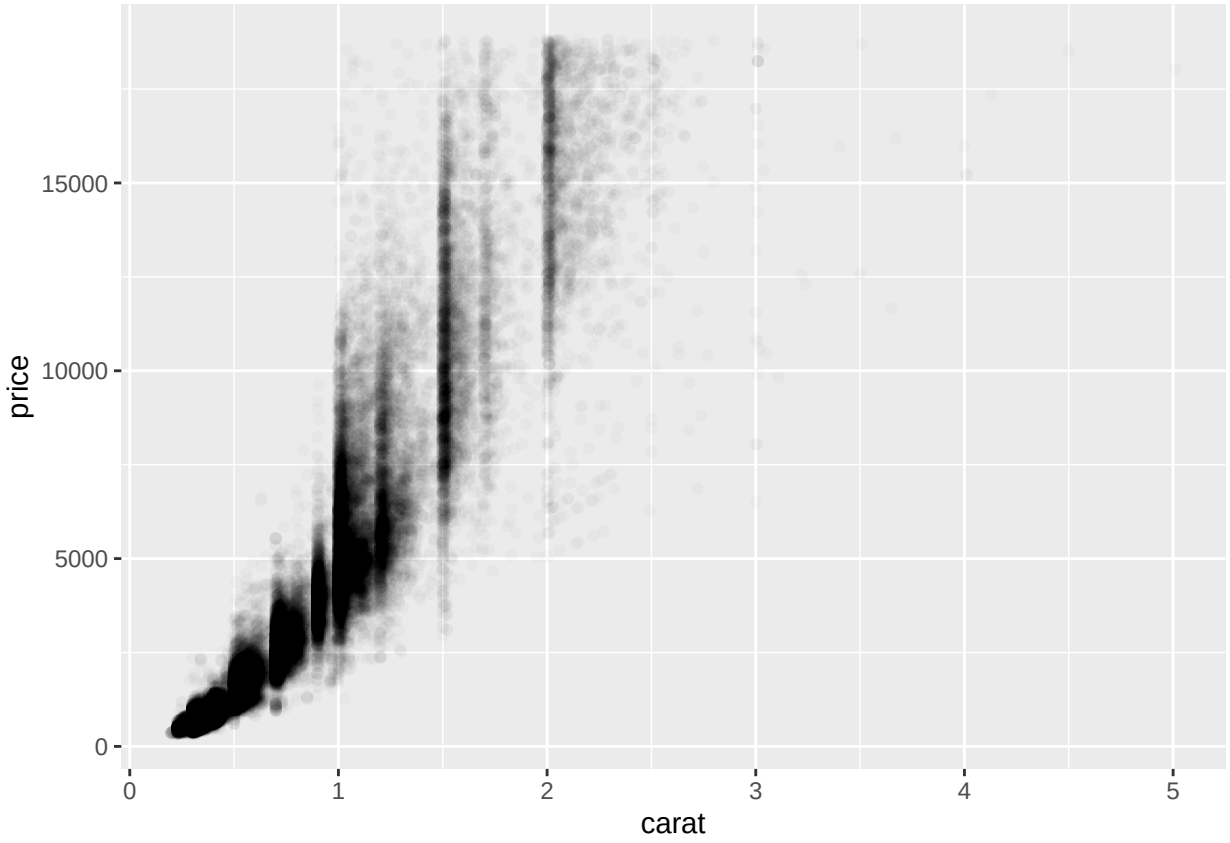


Para visualiazar a covariação entre duas variáveis contínuas podemos utilizar o diagrama de dispersão.

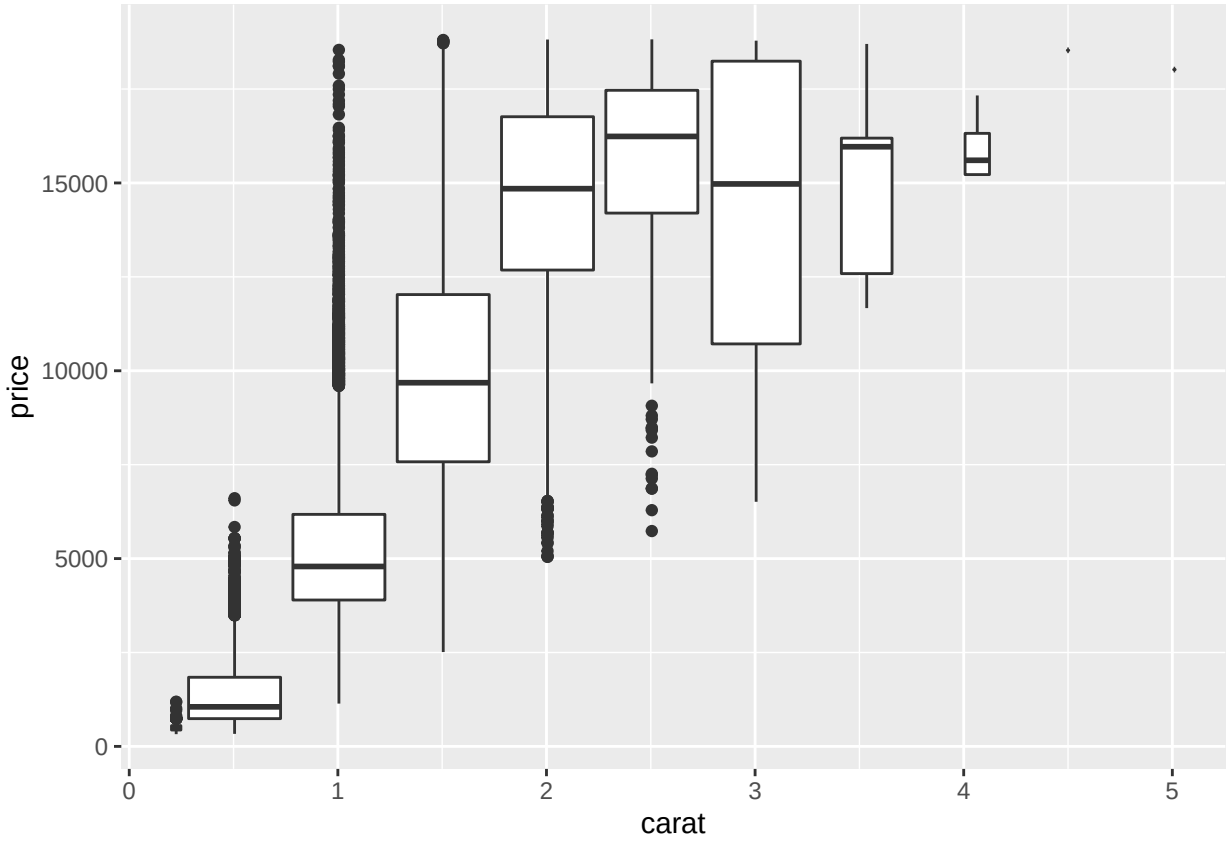
```
ggplot(data = diamonds) +
  geom_point(mapping = aes(x = carat, y = price))
```




```
##

ggplot(data = diamonds) +
  geom_point(mapping = aes(x = carat, y = price),alpha = 1/50)
```




```
ggplot(data = diamonds, mapping = aes(x = carat, y = price)) +
 geom_boxplot(mapping = aes(group = cut_width(carat,0.5)))
```



1.3 Estatística Descritiva

- Pacote pastecs

```
# antes o summary

summary(diamonds)
```

```
##      carat               cut        color        clarity
##  Min.   :0.2000   Fair      : 1610   D: 6775   SI1     :13065
##  1st Qu.:0.4000   Good      : 4906   E: 9797   VS2     :12258
##  Median :0.7000   Very Good:12082   F: 9542   SI2     : 9194
##  Mean   :0.7979   Premium   :13791   G:11292   VS1     : 8171
##  3rd Qu.:1.0400   Ideal     :21551   H: 8304   VVS2    : 5066
##  Max.   :5.0100                     I: 5422   VVS1    : 3655
##                                     J: 2808   (Other): 2531
##      depth           table           price             x
##  Min.   :43.00   Min.   :43.00   Min.   :  326   Min.   : 0.000
##  1st Qu.:61.00   1st Qu.:56.00   1st Qu.:  950   1st Qu.: 4.710
##  Median :61.80   Median :57.00   Median : 2401   Median : 5.700
##  Mean   :61.75   Mean   :57.46   Mean   : 3933   Mean   : 5.731
##  3rd Qu.:62.50   3rd Qu.:59.00   3rd Qu.: 5324   3rd Qu.: 6.540
##  Max.   :79.00   Max.   :95.00   Max.   :18823   Max.   :10.740
##
```

```
##        y                z
##  Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 4.720   1st Qu.: 2.910
##  Median : 5.710   Median : 3.530
##  Mean   : 5.735   Mean   : 3.539
##  3rd Qu.: 6.540   3rd Qu.: 4.040
##  Max.   :58.900   Max.   :31.800
##
```

```
library(pastecs)
```

```
##
## Attaching package: 'pastecs'

## The following objects are masked from 'package:dplyr':
##
##     first, last

## The following object is masked from 'package:tidyr':
##
##     extract
```

```
stat.desc(diamonds)
```

```
##                     carat cut color clarity        depth        table
## nbr.val      5.394000e+04  NA    NA      NA 5.394000e+04 5.394000e+04
## nbr.null     0.000000e+00  NA    NA      NA 0.000000e+00 0.000000e+00
## nbr.na       0.000000e+00  NA    NA      NA 0.000000e+00 0.000000e+00
## min          2.000000e-01  NA    NA      NA 4.300000e+01 4.300000e+01
## max          5.010000e+00  NA    NA      NA 7.900000e+01 9.500000e+01
## range        4.810000e+00  NA    NA      NA 3.600000e+01 5.200000e+01
## sum          4.304087e+04  NA    NA      NA 3.330763e+06 3.099241e+06
## median       7.000000e-01  NA    NA      NA 6.180000e+01 5.700000e+01
## mean         7.979397e-01  NA    NA      NA 6.174940e+01 5.745718e+01
## SE.mean      2.040954e-03  NA    NA      NA 6.168448e-03 9.621063e-03
## CI.mean.0.95 4.000286e-03  NA    NA      NA 1.209021e-02 1.885736e-02
## var          2.246867e-01  NA    NA      NA 2.052404e+00 4.992948e+00
## std.dev      4.740112e-01  NA    NA      NA 1.432621e+00 2.234491e+00
## coef.var     5.940439e-01  NA    NA      NA 2.320057e-02 3.888966e-02
##                     price            x            y            z
## nbr.val      5.394000e+04 5.394000e+04 5.394000e+04 5.394000e+04
## nbr.null     0.000000e+00 8.000000e+00 7.000000e+00 2.000000e+01
## nbr.na       0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
## min          3.260000e+02 0.000000e+00 0.000000e+00 0.000000e+00
## max          1.882300e+04 1.074000e+01 5.890000e+01 3.180000e+01
## range        1.849700e+04 1.074000e+01 5.890000e+01 3.180000e+01
## sum          2.121352e+08 3.091386e+05 3.093203e+05 1.908793e+05
## median       2.401000e+03 5.700000e+00 5.710000e+00 3.530000e+00
## mean         3.932800e+03 5.731157e+00 5.734526e+00 3.538734e+00
## SE.mean      1.717736e+01 4.829974e-03 4.917698e-03 3.038533e-03
## CI.mean.0.95 3.366776e+01 9.466787e-03 9.638727e-03 5.955549e-03
## var          1.591563e+07 1.258347e+00 1.304472e+00 4.980109e-01
## std.dev      3.989440e+03 1.121761e+00 1.142135e+00 7.056988e-01
## coef.var     1.014402e+00 1.957302e-01 1.991681e-01 1.994213e-01
```

- Pacote summarytools

```
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('ra

##
## Attaching package: 'summarytools'

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

```
descr(diamonds)
```

```
## Non-numerical variable(s) ignored: cut, color, clarity

## Descriptive Statistics
## diamonds
## N: 53940
##
##                       carat      depth      price      table          x          y          z
## ----------------- ---------- ---------- ---------- ---------- ---------- ---------- ----------
##              Mean       0.80      61.75    3932.80      57.46       5.73       5.73       3.54
##           Std.Dev       0.47       1.43    3989.44       2.23       1.12       1.14       0.71
##               Min       0.20      43.00     326.00      43.00       0.00       0.00       0.00
##                Q1       0.40      61.00     950.00      56.00       4.71       4.72       2.91
##            Median       0.70      61.80    2401.00      57.00       5.70       5.71       3.53
##                Q3       1.04      62.50    5324.50      59.00       6.54       6.54       4.04
##               Max       5.01      79.00   18823.00      95.00      10.74      58.90      31.80
##               MAD       0.47       1.04    2475.94       1.48       1.38       1.36       0.85
##               IQR       0.64       1.50    4374.25       3.00       1.83       1.82       1.13
##                CV       0.59       0.02       1.01       0.04       0.20       0.20       0.20
##          Skewness       1.12      -0.08       1.62       0.80       0.38       2.43       1.52
##       SE.Skewness       0.01       0.01       0.01       0.01       0.01       0.01       0.01
##          Kurtosis       1.26       5.74       2.18       2.80      -0.62      91.20      47.08
##           N.Valid   53940.00   53940.00   53940.00   53940.00   53940.00   53940.00   53940.00
##         Pct.Valid     100.00     100.00     100.00     100.00     100.00     100.00     100.00
```

```
dfSummary(diamonds)
```

```
## Data Frame Summary
## diamonds
## Dimensions: 53940 x 10
## Duplicates: 146
##
## ------------------------------------------------------------------------------------------------------
```

```
## No   Variable            Stats / Values                 Freqs (% of Valid)      Graph
## ---- ------------------- ------------------------------ ----------------------- ---------------------
## 1    carat               Mean (sd) : 0.8 (0.5)          273 distinct values     :
##      [numeric]           min < med < max:                                       : .
##                          0.2 < 0.7 < 5                                          : :
##                          IQR (CV) : 0.6 (0.6)                                   : : .
##                                                                                 : : : .
##
## 2    cut                 1. Fair                         1610 ( 3.0%)
##      [ordered, factor]   2. Good                         4906 ( 9.1%)           I
##                          3. Very Good                   12082 (22.4%)           IIII
##                          4. Premium                     13791 (25.6%)           IIIII
##                          5. Ideal                       21551 (40.0%)           IIIIIII
##
## 3    color               1. D                            6775 (12.6%)           II
##      [ordered, factor]   2. E                            9797 (18.2%)           III
##                          3. F                            9542 (17.7%)           III
##                          4. G                           11292 (20.9%)           IIII
##                          5. H                            8304 (15.4%)           III
##                          6. I                            5422 (10.1%)           II
##                          7. J                            2808 ( 5.2%)           I
##
## 4    clarity             1. I1                            741 ( 1.4%)
##      [ordered, factor]   2. SI2                          9194 (17.0%)           III
##                          3. SI1                         13065 (24.2%)           IIII
##                          4. VS2                         12258 (22.7%)           IIII
##                          5. VS1                          8171 (15.2%)           III
##                          6. VVS2                         5066 ( 9.4%)           I
##                          7. VVS1                         3655 ( 6.8%)           I
##                          8. IF                           1790 ( 3.3%)
##
## 5    depth               Mean (sd) : 61.7 (1.4)         184 distinct values                  :
##      [numeric]           min < med < max:                                                    :
##                          43 < 61.8 < 79                                                      :
##                          IQR (CV) : 1.5 (0)                                            . :
##                                                                                        : :
##
## 6    table               Mean (sd) : 57.5 (2.2)         127 distinct values         :
##      [numeric]           min < med < max:                                           :
##                          43 < 57 < 95                                               :
##                          IQR (CV) : 3 (0)                                           : :
##                                                                                     : :
##
## 7    price               Mean (sd) : 3932.8 (3989.4)    11602 distinct values   :
##      [integer]           min < med < max:                                       :
##                          326 < 2401 < 18823                                     :
##                          IQR (CV) : 4374.2 (1)                                  : : .
##                                                                                 : : : : . . .
##
## 8    x                   Mean (sd) : 5.7 (1.1)          554 distinct values              :
##      [numeric]           min < med < max:                                                : .
##                          0 < 5.7 < 10.7                                                  : : :
##                          IQR (CV) : 1.8 (0.2)                                            : : :
##                                                                                    . : : : :
```

```
##
## 9    y                   Mean (sd) : 5.7 (1.1)          552 distinct values     :
##      [numeric]           min < med < max:                                       : :
##                          0 < 5.7 < 58.9                                         : :
##                          IQR (CV) : 1.8 (0.2)                                   : :
##                                                                                 : :
##
## 10   z                   Mean (sd) : 3.5 (0.7)          375 distinct values       :
##      [numeric]           min < med < max:                                         :
##                          0 < 3.5 < 31.8                                         : :
##                          IQR (CV) : 1.1 (0.2)                                   : :
##                                                                                 : :
## ------------------------------------------------------------------------------------------------------------
```

- Pacote skimr

```
# library(skimr)
# skim(diamonds)
```

2 Referências

1) Livro: R for Data Science - Hadley Wickham & Garrett Grolemund. Alta Books, 2019.

2) Livro: Data Wrangling with R - Bradley C. Boehmke. Springer, 2016.

3) Livro: Practical Statistics for Data Scientists: 50 Essential Concepts - Peter Bruce & Andrew Bruce. Alta Books, 2019.

4) Livro: Exploratory Data Analysis with R - Roger D. Peng. https://bookdown.org/rdpeng/exdata/, 2016.

5) http://sillasgonzaga.com/material/curso_visualizacao/