Trilha3-Explorando a Base de Dados diamonds do pacote ggplot2

André Gustavo Silva Lovo

07/11/2020

# Trilha3

#install.packages("readr", dependencies = TRUE)  
library(readr)   
#install.packages("data.table", dependencies = TRUE)  
library(data.table)  
#install.packages("knitr")  
library(knitr)  
#install.packages("ggplot2", dependencies = TRUE)  
#install.packages("dplyr", dependencies = TRUE)  
library(dplyr)

##   
## Attaching package: 'dplyr'

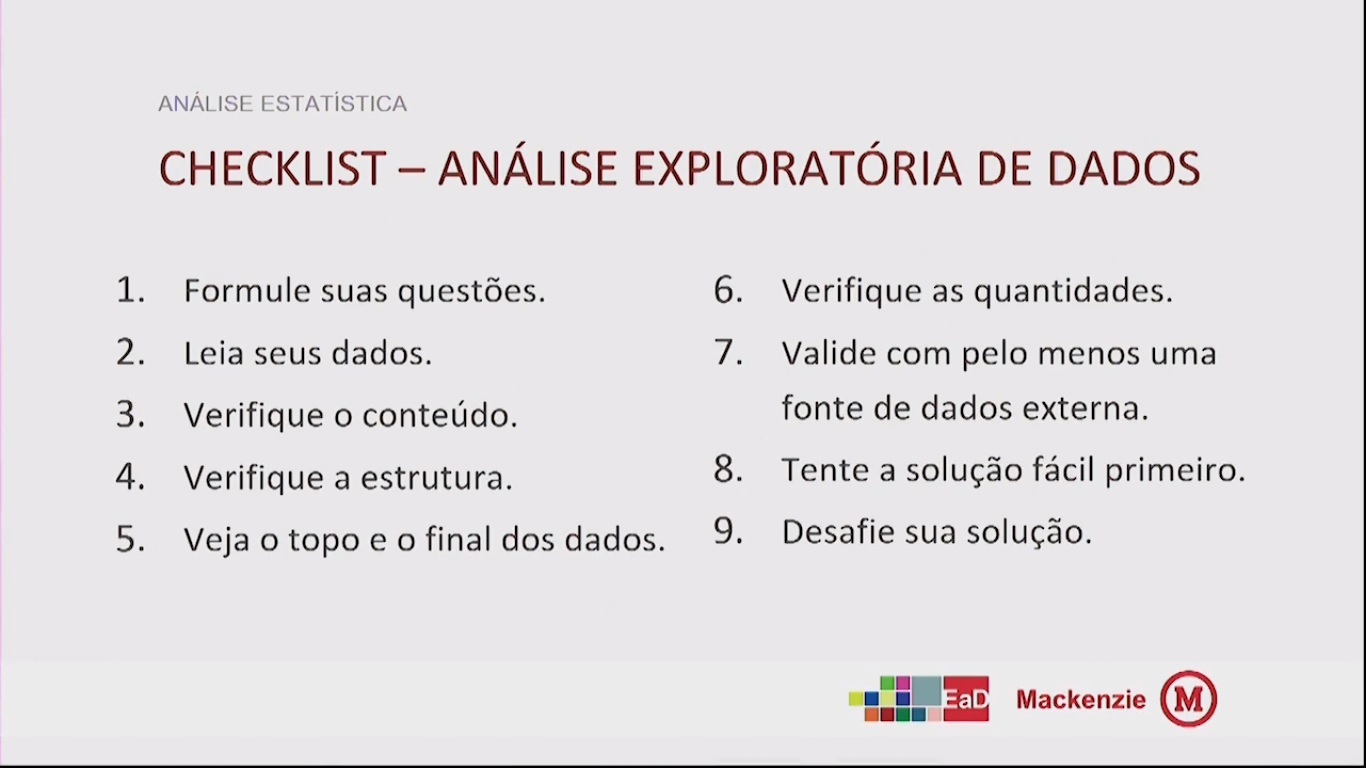
## The following objects are masked from 'package:data.table':  
##   
## between, first, last

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#install.packages("tidyr", dependencies = TRUE)  
library(tidyr)

## Para esta atividade de AED (Análise Exploratória de Dados) vamos seguir o seguinte checklist:



checklist

library(ggplot2)  
diamantes <- data(diamonds)

# Dados e informações do Banco de dados diamonds:

price: preço em dólares americanos ($326–$18,823)  
carat: peso do diamente (0.2–5.01)  
cut:qualidade do corte (Fair, Good, Very Good, Premium, Ideal)  
color: cor do diamante, indo de J (pior) a D (melhor)  
clarity: medida de quão claro é o diamante (I1 (pior), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (melhor))  
x: comprimento em mm (0–10.74)  
y: largura em mm (0–58.9)  
z: profundidade em mm (0–31.8)  
depth: percentual de profundidade total = z / mean(x, y) = 2 \* z / (x + y) (43–79)  
table: largura do topo do diamante relativo ao ponto mais largo (43–95)

Como pode ser observado a tabela apresenta as seguintes variáveis para a quantificação e qualificação de uma tabela com características para os diamantes:  
Preço - peso - qualidade do corte - claridade - comprimento - largura - profundidade - profundidade total em porcentagem - relação de largura entre o topo e a parte mais larga.

dim(diamonds)

## [1] 53940 10

## Investigando e analisando a base de dados:

### Qual é a estrutura do conjunto de dados “diamantes”?

View (diamonds)

typeof(diamonds$price)

## [1] "integer"

typeof(diamonds$carat)

## [1] "double"

typeof(diamonds$cut)

## [1] "integer"

typeof(diamonds$color)

## [1] "integer"

typeof(diamonds$clarity)

## [1] "integer"

typeof(diamonds$depth)

## [1] "double"

typeof(diamonds$table)

## [1] "double"

typeof(diamonds$x)

## [1] "double"

typeof(diamonds$y)

## [1] "double"

typeof(diamonds$z)

## [1] "double"

str(diamonds)

## tibble [53,940 x 10] (S3: tbl\_df/tbl/data.frame)  
## $ carat : num [1:53940] 0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...  
## $ cut : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...  
## $ color : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...  
## $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...  
## $ depth : num [1:53940] 61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...  
## $ table : num [1:53940] 55 61 65 58 58 57 57 55 61 61 ...  
## $ price : int [1:53940] 326 326 327 334 335 336 336 337 337 338 ...  
## $ x : num [1:53940] 3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...  
## $ y : num [1:53940] 3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...  
## $ z : num [1:53940] 2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...

View(diamonds)

### Explore a parte inicial e a final do conjunto de dados.

head(diamonds,5)

## # A tibble: 5 x 10  
## carat cut color clarity depth table price x y z  
## <dbl> <ord> <ord> <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 0.23 Ideal E SI2 61.5 55 326 3.95 3.98 2.43  
## 2 0.21 Premium E SI1 59.8 61 326 3.89 3.84 2.31  
## 3 0.23 Good E VS1 56.9 65 327 4.05 4.07 2.31  
## 4 0.290 Premium I VS2 62.4 58 334 4.2 4.23 2.63  
## 5 0.31 Good J SI2 63.3 58 335 4.34 4.35 2.75

tail(diamonds,10)

## # A tibble: 10 x 10  
## carat cut color clarity depth table price x y z  
## <dbl> <ord> <ord> <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>  
## 1 0.71 Premium E SI1 60.5 55 2756 5.79 5.74 3.49  
## 2 0.71 Premium F SI1 59.8 62 2756 5.74 5.73 3.43  
## 3 0.7 Very Good E VS2 60.5 59 2757 5.71 5.76 3.47  
## 4 0.7 Very Good E VS2 61.2 59 2757 5.69 5.72 3.49  
## 5 0.72 Premium D SI1 62.7 59 2757 5.69 5.73 3.58  
## 6 0.72 Ideal D SI1 60.8 57 2757 5.75 5.76 3.5   
## 7 0.72 Good D SI1 63.1 55 2757 5.69 5.75 3.61  
## 8 0.7 Very Good D SI1 62.8 60 2757 5.66 5.68 3.56  
## 9 0.86 Premium H SI2 61 58 2757 6.15 6.12 3.74  
## 10 0.75 Ideal D SI2 62.2 55 2757 5.83 5.87 3.64

### Faça alguns sumários estatísticos para entender melhor a base de dados.

summary(diamonds)

## carat cut color clarity depth   
## Min. :0.2000 Fair : 1610 D: 6775 SI1 :13065 Min. :43.00   
## 1st Qu.:0.4000 Good : 4906 E: 9797 VS2 :12258 1st Qu.:61.00   
## Median :0.7000 Very Good:12082 F: 9542 SI2 : 9194 Median :61.80   
## Mean :0.7979 Premium :13791 G:11292 VS1 : 8171 Mean :61.75   
## 3rd Qu.:1.0400 Ideal :21551 H: 8304 VVS2 : 5066 3rd Qu.:62.50   
## Max. :5.0100 I: 5422 VVS1 : 3655 Max. :79.00   
## J: 2808 (Other): 2531   
## table price x y   
## Min. :43.00 Min. : 326 Min. : 0.000 Min. : 0.000   
## 1st Qu.:56.00 1st Qu.: 950 1st Qu.: 4.710 1st Qu.: 4.720   
## Median :57.00 Median : 2401 Median : 5.700 Median : 5.710   
## Mean :57.46 Mean : 3933 Mean : 5.731 Mean : 5.735   
## 3rd Qu.:59.00 3rd Qu.: 5324 3rd Qu.: 6.540 3rd Qu.: 6.540   
## Max. :95.00 Max. :18823 Max. :10.740 Max. :58.900   
##   
## z   
## Min. : 0.000   
## 1st Qu.: 2.910   
## Median : 3.530   
## Mean : 3.539   
## 3rd Qu.: 4.040   
## Max. :31.800   
##

## Analisando a tabela e suas colunas:

Coluna price: preço em dólares americanos-  
R:(Atributo tipo numérico inteiro)  
Coluna carat: peso do diamente-  
R:(Atributo tipo numérico double com casas decemais)  
Coluna cut:qualidade do corte (Fair, Good, Very Good, Premium, Ideal)-  
R:(Atributo tipo Factor - categórico ou qualitativo)  
Coluna color: cor do diamante, indo de J (pior) a D (melhor)-  
R:(Atributo tipo Factor - categórico ou qualitativo)  
Coluna clarity: medida de quão claro é o diamante-  
R:(Atributo tipo Factor - categórico ou qualitativo)  
Coluna depth: percentual de profundidade total-  
R:(Atributo tipo numérico double com casas decemais)  
Coluna table: largura do topo do diamante relativo ao ponto mais largo-  
R:(Atributo tipo numérico double com casas decemais)  
Coluna x: comprimento em mm-  
R:(Atributo tipo numérico double com casas decemais)  
Coluna y: largura em mm-  
R:(Atributo tipo numérico double com casas decemais)  
Coluna z: profundidade em mm-  
R:(Atributo tipo numérico double com casas decemais)

### A saída da função summary() está de acordo com a descrição mostrada anteriormente?

As colunas que são do tipo numéricos (inteiros e/ou double com casas decimais) e seus resultados estão de acordo com a função summary ,pois apresntam tipos numéricos. Ex.:carat, price, depth, table, x,y,z.  
Já as colunas que são do tipo caracteres (factor ordinais - são do tipo categórico), não estão de acordo com a função summary.  
Ex.:cut, color, clarity.

## Resumo do Banco de Dados

Usamos esta biblioteca para visualizarmos de forma global a tabela e suas colunas(variáveis).

#install.packages("skimr")  
library(skimr)  
skim(diamonds)

Data summary

|  |  |
| --- | --- |
| Name | diamonds |
| Number of rows | 53940 |
| Number of columns | 10 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 3 |
| numeric | 7 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| cut | 0 | 1 | TRUE | 5 | Ide: 21551, Pre: 13791, Ver: 12082, Goo: 4906 |
| color | 0 | 1 | TRUE | 7 | G: 11292, E: 9797, F: 9542, H: 8304 |
| clarity | 0 | 1 | TRUE | 8 | SI1: 13065, VS2: 12258, SI2: 9194, VS1: 8171 |

**Variable type: numeric**

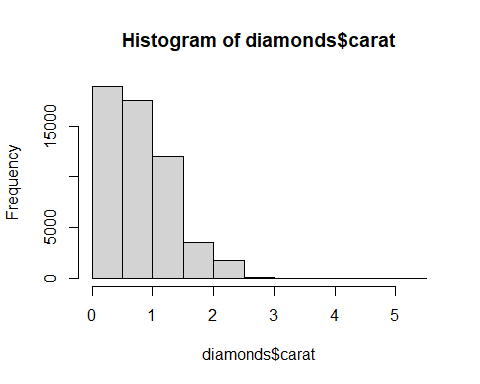
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| carat | 0 | 1 | 0.80 | 0.47 | 0.2 | 0.40 | 0.70 | 1.04 | 5.01 | ▇▂▁▁▁ |
| depth | 0 | 1 | 61.75 | 1.43 | 43.0 | 61.00 | 61.80 | 62.50 | 79.00 | ▁▁▇▁▁ |
| table | 0 | 1 | 57.46 | 2.23 | 43.0 | 56.00 | 57.00 | 59.00 | 95.00 | ▁▇▁▁▁ |
| price | 0 | 1 | 3932.80 | 3989.44 | 326.0 | 950.00 | 2401.00 | 5324.25 | 18823.00 | ▇▂▁▁▁ |
| x | 0 | 1 | 5.73 | 1.12 | 0.0 | 4.71 | 5.70 | 6.54 | 10.74 | ▁▁▇▃▁ |
| y | 0 | 1 | 5.73 | 1.14 | 0.0 | 4.72 | 5.71 | 6.54 | 58.90 | ▇▁▁▁▁ |
| z | 0 | 1 | 3.54 | 0.71 | 0.0 | 2.91 | 3.53 | 4.04 | 31.80 | ▇▁▁▁▁ |

#O comando skim é interessante, porque além do sumário numérico, também temos um pequeno histograma para vermos a distribuição dos dados.   
# Vamos analisar as variáveis contínuas (numéricas) através dos histogramas.

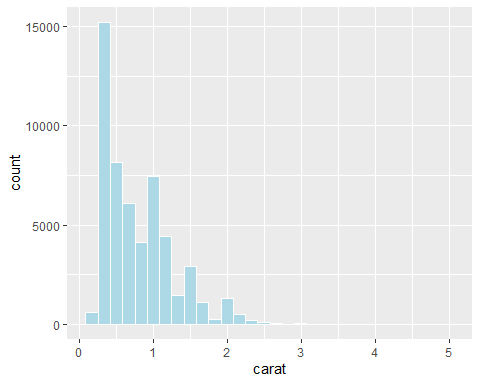
## Gráfico Histograma

### Explorando a variável carat, através do histograma.

hist(diamonds$carat)



ggplot(diamonds, aes(x = carat)) + geom\_histogram(bins=30, color = "white", fill = "lightblue")

 ### Análise: Acima fizemos a comparação da distribuição dos pesos dos diamantes conforme a faixa das quantidades, usando o histograma.

### Explorando a variável depth, através do histograma.

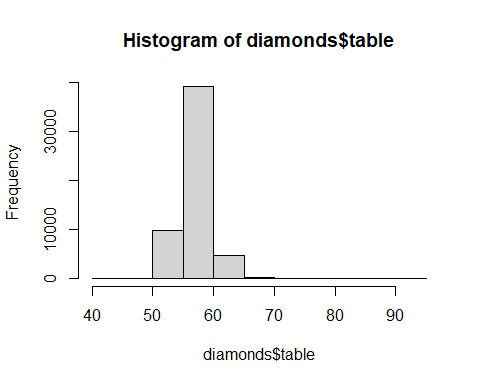
hist(diamonds$depth)

ggplot(diamonds, aes(x = depth)) + geom\_histogram(bins=30, color = "white", fill = "green")

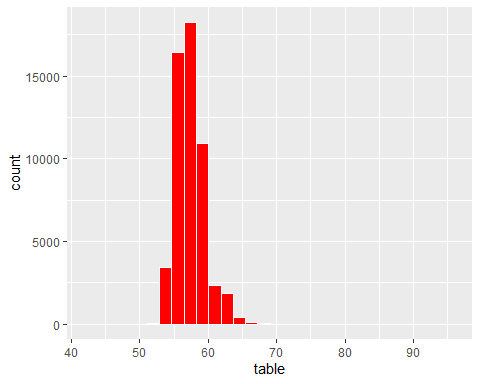
### Análise: Acima fizemos a comparação da distribuição das profundidades (em porcentagem) dos diamantes conforme a faixa das quantidades, usando o histograma.

### Explorando a variável table, através do histograma.

hist(diamonds$table)

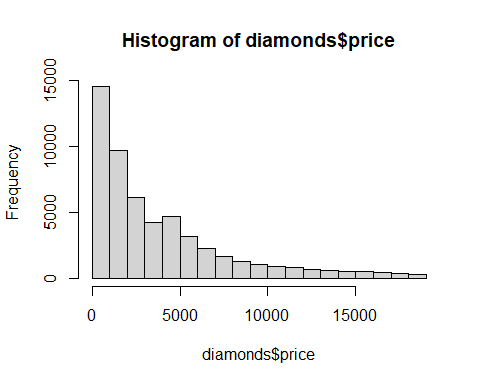


ggplot(diamonds, aes(x = table)) + geom\_histogram(bins = 30, color = "white", fill = "red")

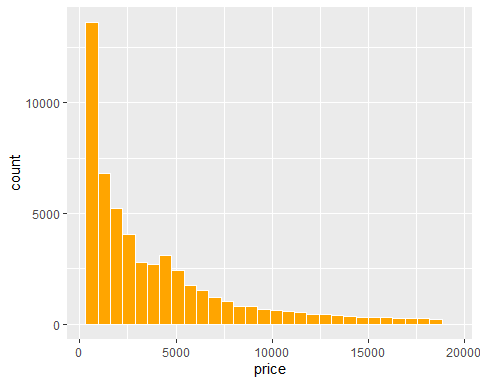
 ### Análise: Acima fizemos a comparação da distribuição das larguras (mm) dos diamantes conforme a faixa das quantidades, usando o histograma.

### Explorando a variável price, através do histograma.

hist(diamonds$price)

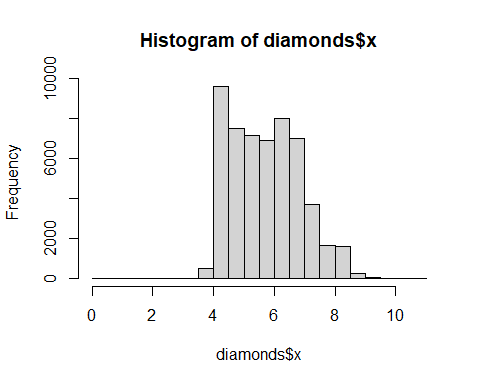


ggplot(diamonds, aes(x = price)) + geom\_histogram(bins = 30, color = "white", fill = "orange")

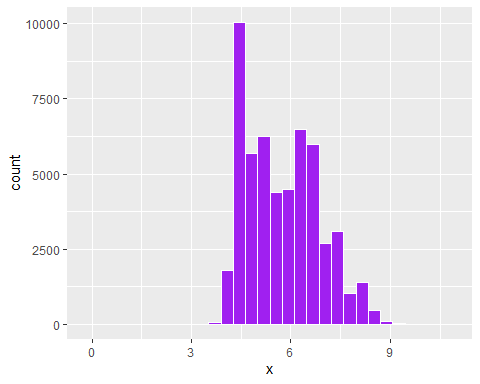
 ### Análise: Acima fizemos a comparação da distribuição dos preços (dólares americanos) dos diamantes conforme a faixa das quantidades, usando o histograma.

###Explorando a variável x, através do histograma.

hist(diamonds$x)

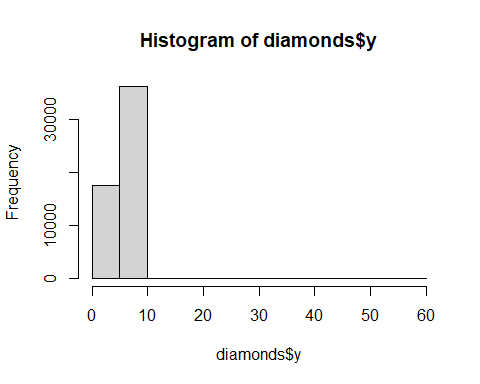


ggplot(diamonds, aes(x = x)) + geom\_histogram(bins=30, color = "white", fill = "purple")

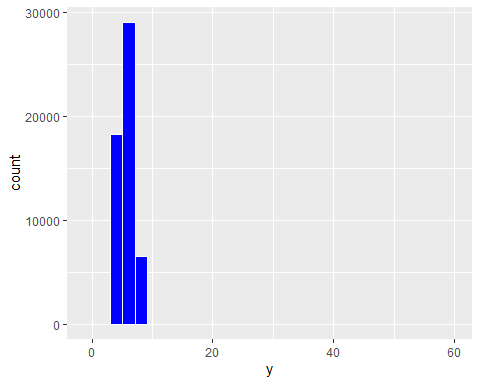
 ### Análise: Acima fizemos a comparação da distribuição do comprimento (mm) dos diamantes conforme a faixa das quantidades, usando o histograma.

###Explorando a variável y, através do histograma.

hist(diamonds$y)

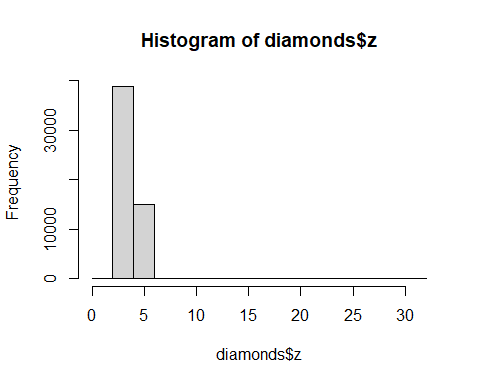


ggplot(diamonds, aes(x = y)) + geom\_histogram(bins=30, color = "white", fill = "blue")

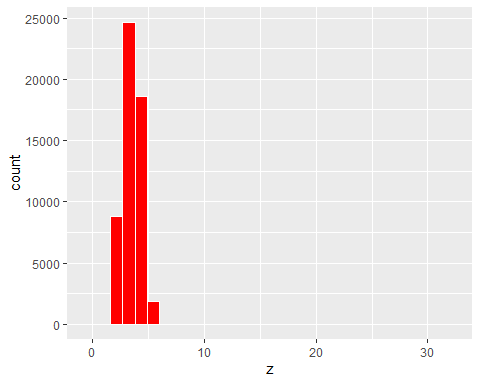
 ### Análise: Acima fizemos a comparação da distribuição da largura (mm) dos diamantes conforme a faixa das quantidades, usando o histograma.

###Explorando a variável z, através do histograma.

hist(diamonds$z)



ggplot(diamonds, aes(x = z)) + geom\_histogram(bins = 30, color = "white", fill = "red")

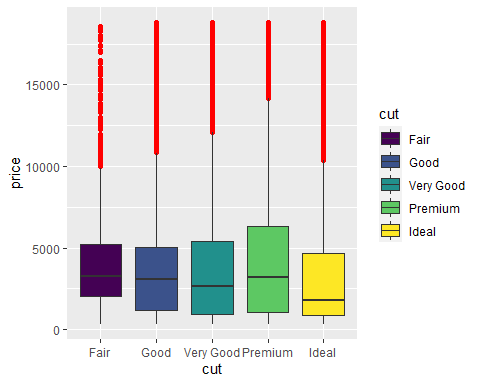
 ### Análise: Acima fizemos a comparação da distribuição da profundidade (mm) dos diamantes conforme a faixa das quantidades, usando o histograma.

## Gráfico Boxplot

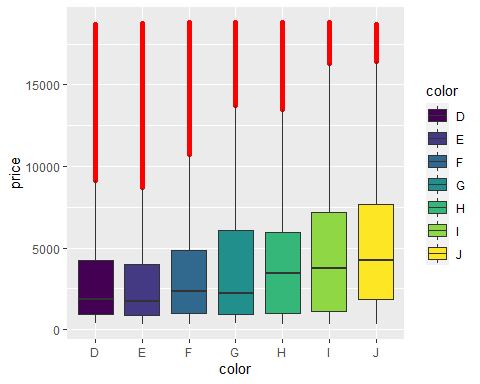
### Explore a variável price, seguindo o modelo de exploração.

Vamos analisar as variáveis de forma separada utilizando o gráfico Boxplot.

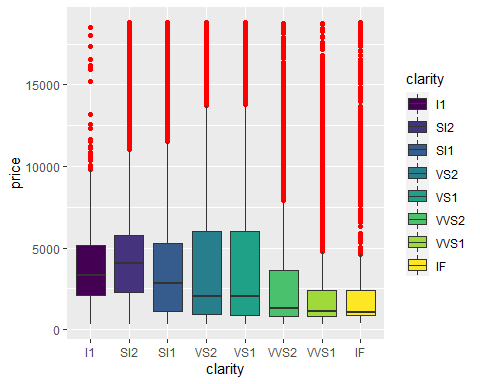
#Boxplot para variáveis numéricas (devido a separação dos quartis)  
ggplot(data = diamonds, mapping = aes(x = cut, y = price, fill =cut)) + geom\_boxplot(outlier.color = "red")



#xlab("Qualidade do corte") + ylab("Preço(Dólares americanos)")  
ggplot(data = diamonds, mapping = aes(x = color, y = price, fill =color)) + geom\_boxplot(outlier.color = "red")



#xlab("Cor") + ylab("Preço(Dólares americanos)")  
ggplot(data = diamonds, mapping = aes(x = clarity, y = price, fill =clarity)) + geom\_boxplot(outlier.color = "red")



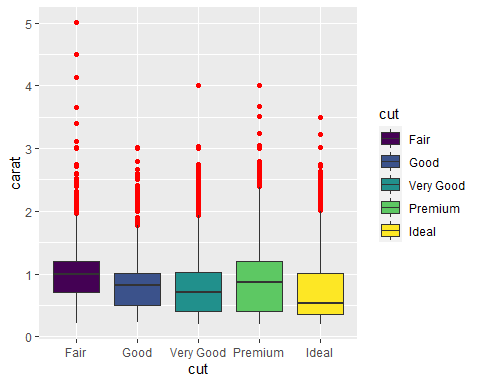
#xlab("Quão claro é o diamante") + ylab("Preço(Dólares americanos)")

### Análise:

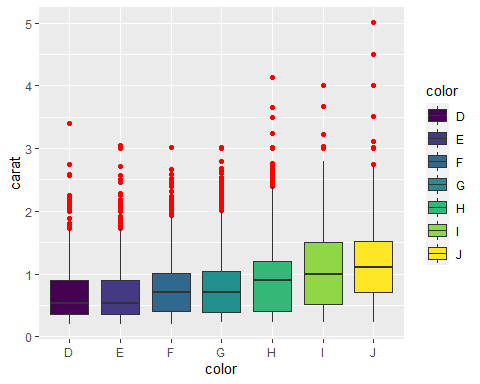
Acima fizemos a comparação dos preços dos diamantes conforme a qualidade do corte, cor e claridade (transparencia do diamante). O tipo de corte que mais influencia o preço é o corte premium. A qualidade de cor que mais influencia o preço é a cor J (pior cor). A claridade (transparência) que mais influencia o preço são as claridades VS1 e VS2.

### Explore a variável carat, seguindo o modelo de exploração.

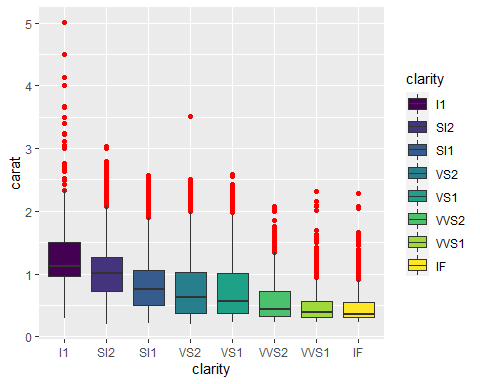
ggplot(data = diamonds, mapping = aes(x = cut, y = carat, fill =cut )) + geom\_boxplot(outlier.color = "red")



ggplot(data = diamonds, mapping = aes(x = color, y = carat, fill =color )) + geom\_boxplot(outlier.color = "red")

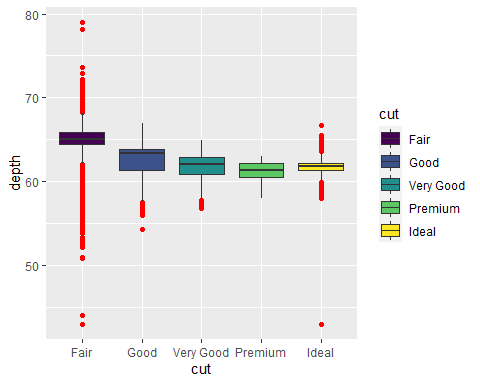


ggplot(data = diamonds, mapping = aes(x = clarity, y = carat, fill =clarity )) + geom\_boxplot(outlier.color = "red")

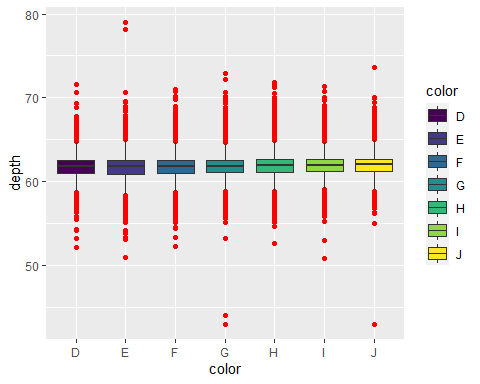
 ### Análise: Acima fizemos a comparação do peso dos diamantes conforme a qualidade do corte, cor e claridade (transparencia do diamante). O tipo de corte que mais influencia o peso é o corte premium. A qualidade de cor que mais influencia o peso é a cor J (pior cor). A claridade (transparência) que mais influencia o peso é a I1 (pior).

### Explore a variável depth, seguindo o modelo de exploração.

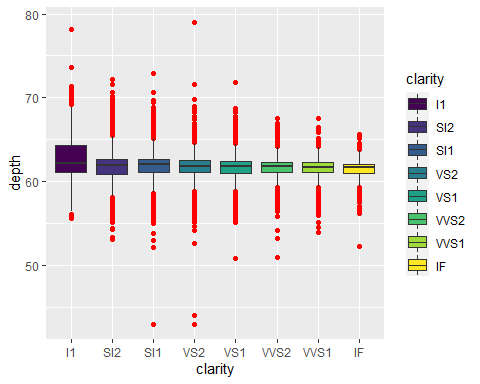
ggplot(data = diamonds, mapping = aes(x = cut, y =depth , fill =cut )) + geom\_boxplot(outlier.color = "red")



ggplot(data = diamonds, mapping = aes(x = color, y =depth , fill =color )) + geom\_boxplot(outlier.color = "red")

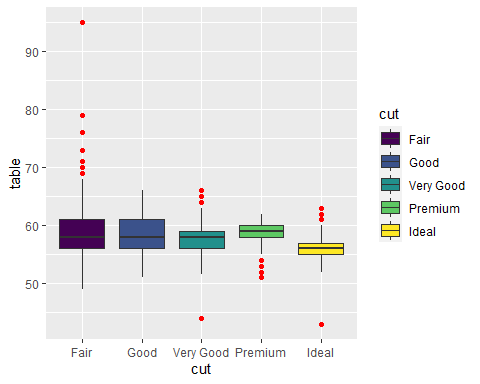


ggplot(data = diamonds, mapping = aes(x = clarity, y =depth , fill =clarity )) + geom\_boxplot(outlier.color = "red")

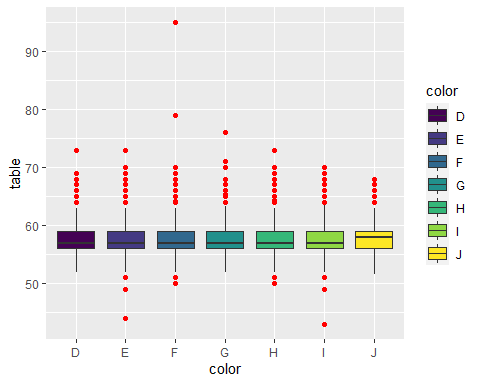
 ### Análise: Acima fizemos a comparação da porcentagem de profundidade dos diamantes conforme a qualidade do corte, cor e claridade (transparencia do diamante). O tipo de corte que mais influencia o a profundidade é o Fair. A qualidade de cor que mais influencia a profundidade E, G e J,pois apresentam valores outliers, embora todas sejam muito parecidos em sua mediana. A claridade (transparência) que mais influencia a profundidade é a I1, embora todas sejam muito parecidos em sua mediana.

### Explore a variável table, seguindo o modelo de exploração.

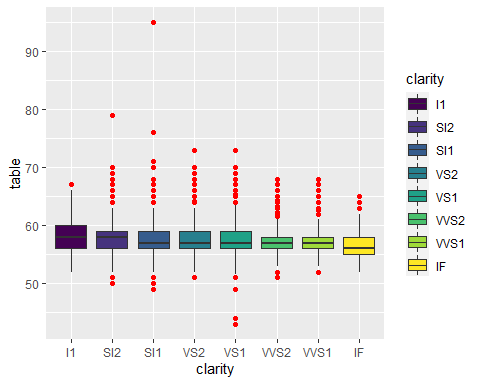
ggplot(data = diamonds, mapping = aes(x = cut, y =table , fill =cut )) + geom\_boxplot(outlier.color = "red")



ggplot(data = diamonds, mapping = aes(x = color, y =table , fill =color )) + geom\_boxplot(outlier.color = "red")

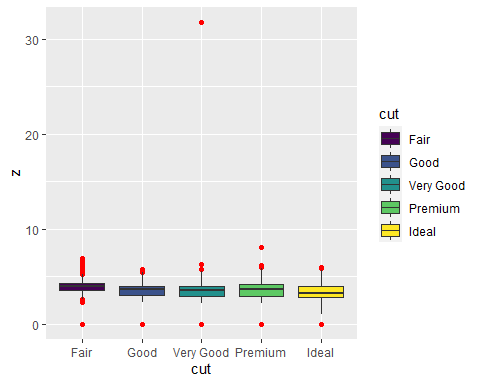


ggplot(data = diamonds, mapping = aes(x = clarity, y =table , fill =clarity )) + geom\_boxplot(outlier.color = "red")

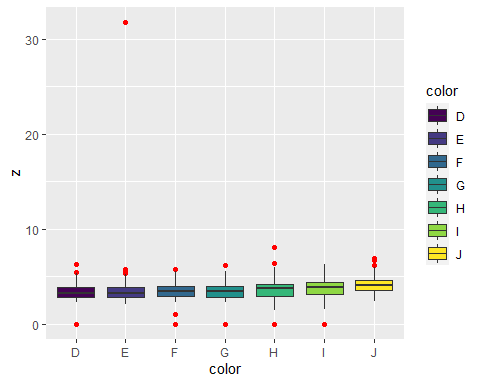
 ### Análise: Acima fizemos a comparação da largura do topo do diamante relativo ao ponto mais largo dos diamantes conforme a qualidade do corte, cor e claridade (transparencia do diamante). O tipo de corte que mais influencia a largura do topo do diamante relativo ao ponto mais largo é o corte Fair, embora as medianas sejam muito próximas. A qualidade de cor que mais influencia a largura do topo do diamante relativo ao ponto mais largo é a cor F, embora as medianas sejam muito próximas. A claridade (transparência) que mais influencia largura do topo do diamante relativo ao ponto mais largo é a transparência SI1, embora as medianas sejam muito próximas.

### Explore a variável z, seguindo o modelo de exploração.

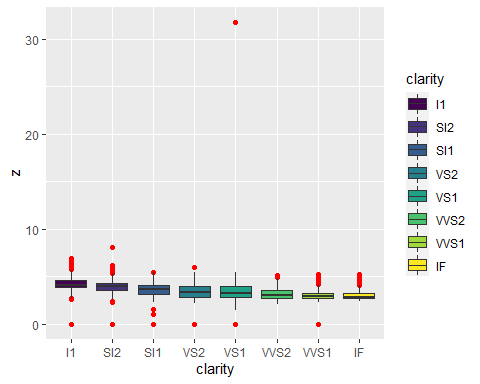
ggplot(data = diamonds, mapping = aes(x = cut, y =z , fill =cut )) + geom\_boxplot(outlier.color = "red")



ggplot(data = diamonds, mapping = aes(x = color, y =z , fill =color )) + geom\_boxplot(outlier.color = "red")

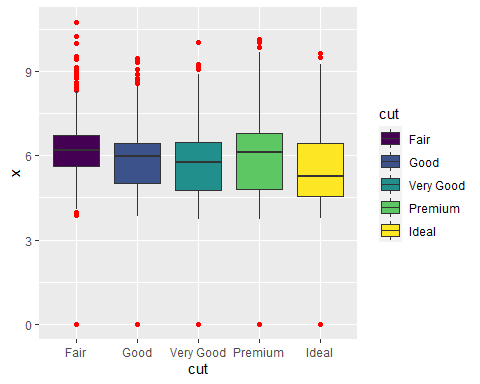


ggplot(data = diamonds, mapping = aes(x = clarity, y =z , fill =clarity )) + geom\_boxplot(outlier.color = "red")

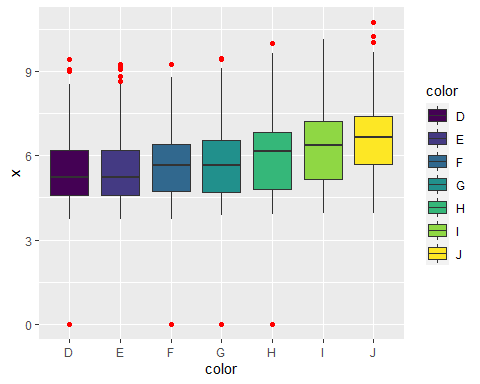
 ### Análise: Acima fizemos a comparação da profundidade dos diamantes conforme a qualidade do corte, cor e claridade (transparencia do diamante). O tipo de corte que mais influencia a profundidade é o corte premium, embora as medianas sejam muito próximas. A qualidade de cor que mais influencia a profundidade é a cor J,embora as medianas sejam muito próximas. A claridade (transparência) que mais influencia a profundidade é a SI2, embora as medianas sejam muito próximas.

### Explore a variável x, seguindo o modelo de exploração.

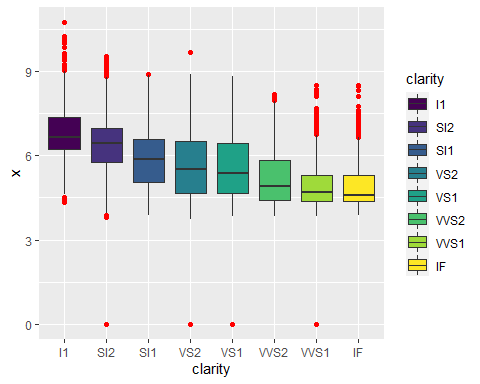
ggplot(data = diamonds, mapping = aes(x = cut, y =x , fill =cut )) + geom\_boxplot(outlier.color = "red")



ggplot(data = diamonds, mapping = aes(x = color, y =x , fill =color )) + geom\_boxplot(outlier.color = "red")

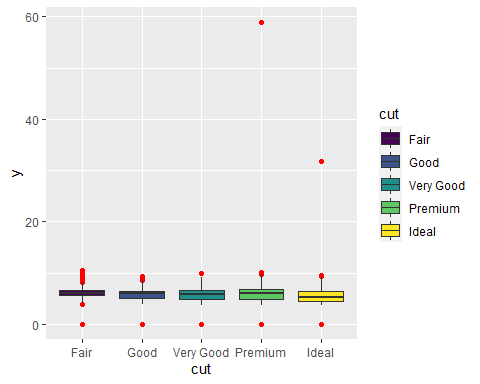


ggplot(data = diamonds, mapping = aes(x = clarity, y =x , fill =clarity )) + geom\_boxplot(outlier.color = "red")

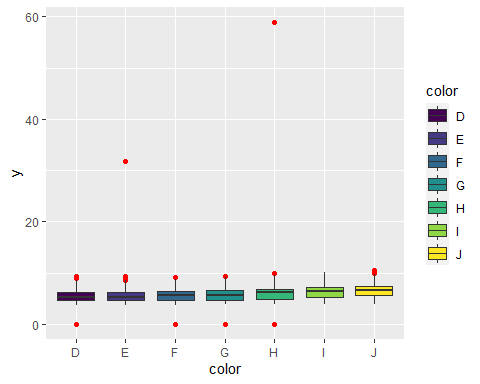
 ### Análise: Acima fizemos a comparação do comprimento dos diamantes conforme a qualidade do corte, cor e claridade (transparencia do diamante). O tipo de corte que mais influencia o comprimento dos diamantes é o premium, embora as medianas sejam muito próximas. A qualidade de cor que mais influencia o comprimento dos diamantes é a cor J (pior cor),embora as medianas sejam muito próximas. A claridade (transparência) que mais influencia o comprimento dos diamantes é a I1, embora todas sejam muito parecidos em sua mediana.

### Explore a variável y, seguindo o modelo de exploração.

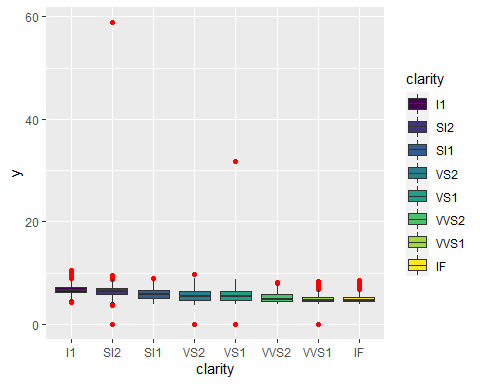
ggplot(data = diamonds, mapping = aes(x = cut, y =y , fill =cut )) + geom\_boxplot(outlier.color = "red")



ggplot(data = diamonds, mapping = aes(x = color, y =y , fill =color )) + geom\_boxplot(outlier.color = "red")

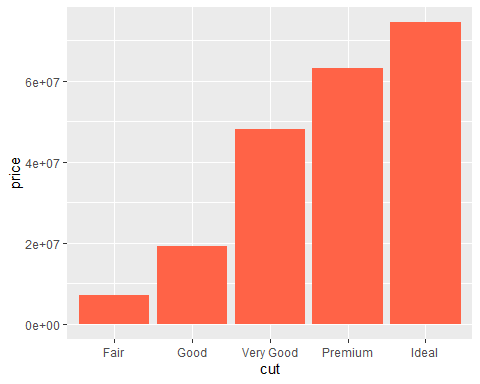


ggplot(data = diamonds, mapping = aes(x = clarity, y =y , fill =clarity )) + geom\_boxplot(outlier.color = "red")

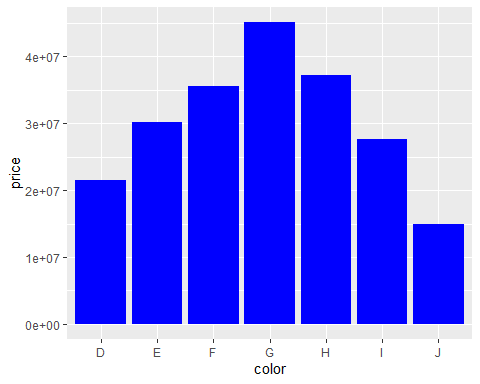
 ### Análise: Acima fizemos a comparação da largura dos diamantes conforme a qualidade do corte, cor e claridade (transparencia do diamante). O tipo de corte que mais influencia a largura dos diamantes é o premium, embora as medianas sejam muito próximas. A qualidade de cor que mais influencia a largura dos diamantes é a cor J (pior cor),embora as medianas sejam muito próximas. A claridade (transparência) que mais influencia a largura dos diamantes é a I1, embora todas sejam muito parecidos em sua mediana.

## Gráfico de Barras

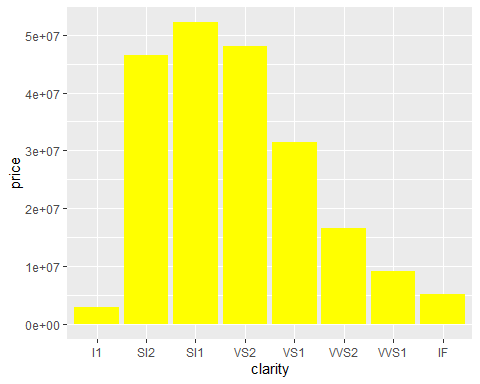
ggplot(diamonds, aes(y = price, x = cut)) + geom\_bar(stat = "identity", fill = "tomato")



ggplot(diamonds, aes(y = price, x = color)) + geom\_bar(stat = "identity", fill = "blue")

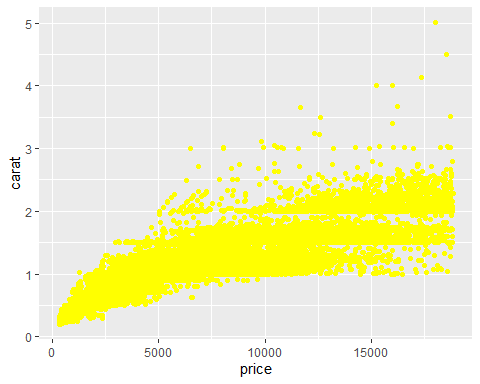


ggplot(diamonds, aes(y = price, x = clarity)) + geom\_bar(stat = "identity", fill = "yellow")

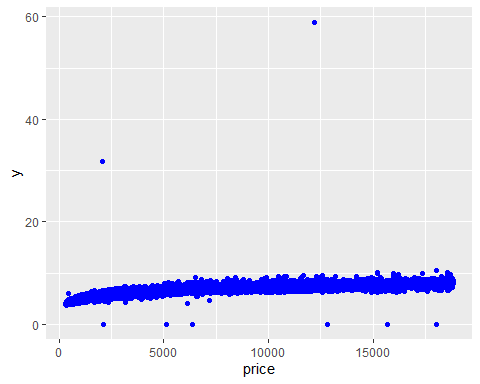
 ### Análise: Acima fizemos a comparação do preço dos diamantes conforme a qualidade do corte, cor e claridade (transparencia do diamante). O tipo de corte que mais influencia o preço dos diamantes é o Ideal. A qualidade de cor que mais influencia o preço dos diamantes é a cor G. A claridade (transparência) que mais influencia o preço dos diamantes é o SI1.

## Scatterplot (Gráfico de Dispersao)

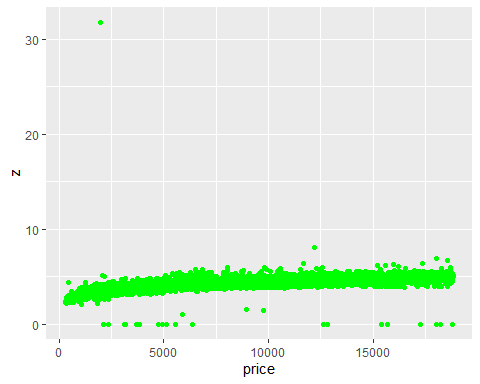
ggplot(diamonds, aes(x=price, y=carat)) + geom\_point(color="yellow")



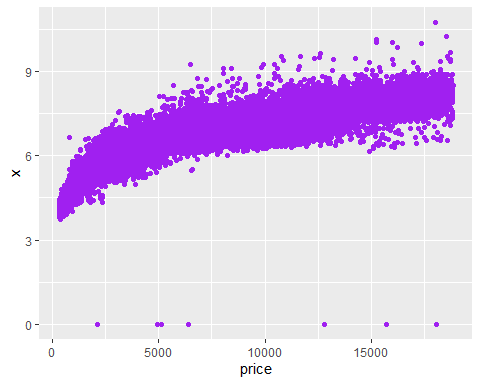
ggplot(diamonds, aes(x=price, y=y)) + geom\_point(color="blue")



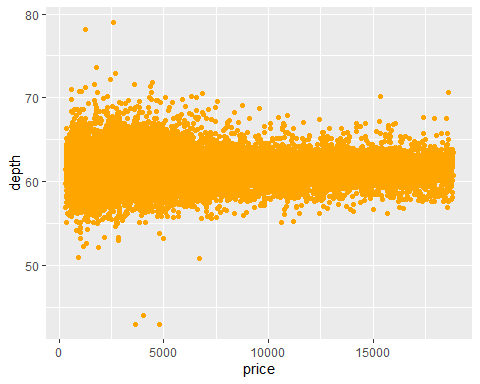
ggplot(diamonds, aes(x=price, y=z)) + geom\_point(color="green")



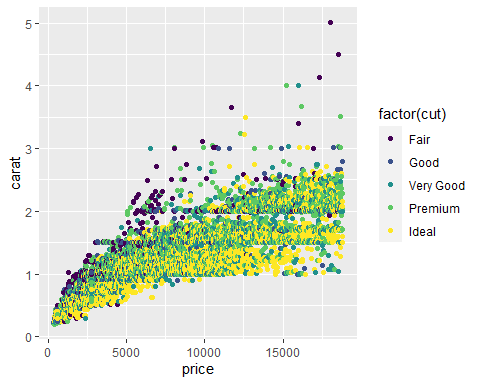
ggplot(diamonds, aes(x=price, y=x)) + geom\_point(color="purple")



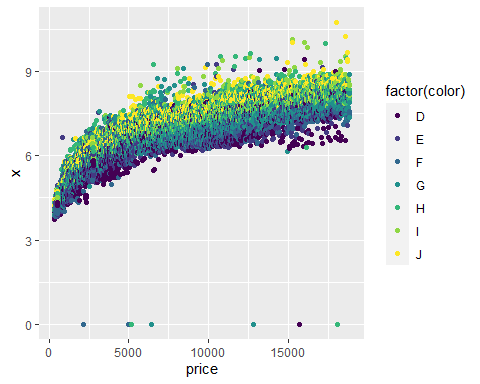
ggplot(diamonds, aes(x=price, y=depth)) + geom\_point(color="orange")



#preço x peso x corte  
ggplot (diamonds, aes (x = price, y = carat)) + geom\_point (aes (color = factor (cut)))



#preço x comprimento x cor  
ggplot (diamonds, aes (x = price, y = x)) + geom\_point (aes (color = factor (color)))



#preço x largura x "transparencia ou claridade"  
ggplot (diamonds, aes (x = price, y = table)) + geom\_point (aes (color = factor (clarity)))

