Trilha 3

Bruna Matos

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

No dicionário abaixo é possível analisarmos qual a estrutura do conjunto de dados utilizado para esta atividade

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

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

Para termos uma noção sobre os dados estudados, é interessante olharmos as primeiras linhas. No nosso caso, temos:

carat	cut	color	clarity	depth	table	price	x	y	z
<dbl>	<ord>	<ord>	<ord>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
0.23	Very Good	H	VS1	59.4	61	338	4.00	4.05	2.39

e para as últimas temos:

carat	cut	color	clarity	depth	table	price	x	y	z
<dbl>	<ord>	<ord>	<ord>	<dbl>	<dbl>	<int>	<dbl>	<dbl>	<dbl>
0.71	Premium	E	SI1	60.5	55	2756	5.79	5.74	3.49
0.71	Premium	F	SI1	59.8	62	2756	5.74	5.73	3.43
0.70	Very Good	E	VS2	60.5	59	2757	5.71	5.76	3.47
0.70	Very Good	E	VS2	61.2	59	2757	5.69	5.72	3.49
0.72	Premium	D	SI1	62.7	59	2757	5.69	5.73	3.58
0.72	Ideal	D	SI1	60.8	57	2757	5.75	5.76	3.50
0.72	Good	D	SI1	63.1	55	2757	5.69	5.75	3.61
0.70	Very Good	D	SI1	62.8	60	2757	5.66	5.68	3.56
0.86	Premium	H	SI2	61.0	58	2757	6.15	6.12	3.74
0.75	Ideal	D	SI2	62.2	55	2757	5.83	5.87	3.64

Ao analisarmos os dados apresentados, podemos observar que a leitura dos dados ocorreu bem, pois aparentemente não temos discrepâncias em relação à natureza e às ordens de grandeza das variáveis.

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

Para entendermos um pouco melhor a base de dados podemos utilizar o comando str que irá nos apresentar algumas características das variáveis, alguns valores e outras informações conforme imagem abaixo.

```
[3]  # Inspecionando os dados
     str(diamantes, strict.width = "wrap", give.attr = FALSE)
```

```
tibble [53,940 × 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 ...
```

Ainda, visando inspecionar as variáveis, podemos usar as informações fornecidas pelo comando summary como apresentado na imagem abaixo. Nós conseguimos analisar se os valores são coerentes, se os valores min e max fazem sentido e ainda, para as variáveis numéricas, conseguimos obter a moda.

No nosso caso, analisando as 4 principais características de um diamantes, que são conhecidos como os 4Cs - Clarity, Carat, Cut e Color, podemos observar por exemplo que no quesito clareza, Faz sentido que tenhamos a maioria deles sendo do tipo SI1, uma vez que diamantes com inclusões pequenas são os mais comuns.

No site https://www.tiffany.com.br/engagement/the-tiffany-guide-to-diamonds/ pode-se obter mais informações sobre os 4Cs.

```r
# Analisando os dados pelo summary
summary(diamantes)
```

```
     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  
```

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

Sim, pois para as variáveis que obtivemos as informações dos valores min, max, mediana, … havíamos classificado como quantitativas e para as que obtivemos as modas havíamos classificado como qualitativas

Explore a variável price, seguindo o modelo de exploração.

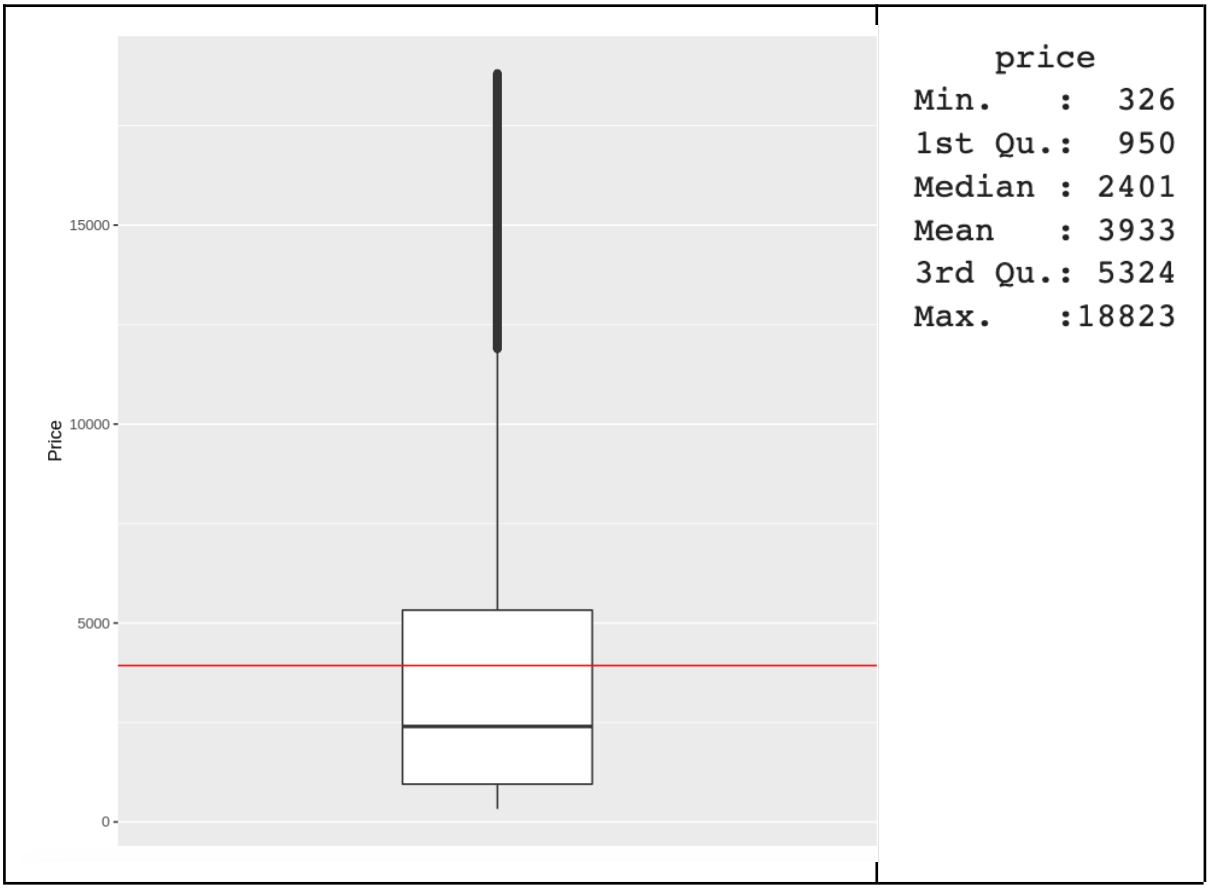
Para iniciarmos a exploração da variável "price", podemos inicialmente nos perguntar qual seriam os fatores que influenciam no preço dos diamantes?

Como já efetuamos a leitura dos dados e após uma análise inicial em cima dos valores iniciais e finais, nas grandezas das variáveis e em seus valores principais, podemos dizer que a leitura do banco ocorreu sem problemas e que os dados estão compatíveis com valores reais. Essa afirmação é feita baseada nas informações obtidas no site https://www.tiffany.com.br/engagement/the-tiffany-guide-to-diamonds/ .

Algo interessante para explorarmos ainda mais a variável preço seria olhando alguns gráficos. Vamos iniciar essa análise com o gráfico boxplot. Para o mesmo, foi utilizado o comando abaixo:

```r
# Grafico box plot
ggplot(data=diamantes) +
  geom_boxplot(
    aes(x = "Price",
        y = price),width = 0.3, varwidth = F) +
    labs(y = "Price",
         x = "") +
    geom_abline(slope=0, intercept=3933, color="red") +
    scale_x_discrete(labels = NULL, breaks = NULL)
```

Como resultado, o gráfico obtido foi:




```
     price
Min.   :  326
1st Qu.:  950
Median : 2401
Mean   : 3933
3rd Qu.: 5324
Max.   :18823
```

Analisando o gráfico, temos a média sendo apresentada pela linha vermelha. Podemos observar que a média está bem acima da mediana (característica do gráfico BoxPlot) assim como que o terceiro quarter está também um pouco afastado no valor da mediana.

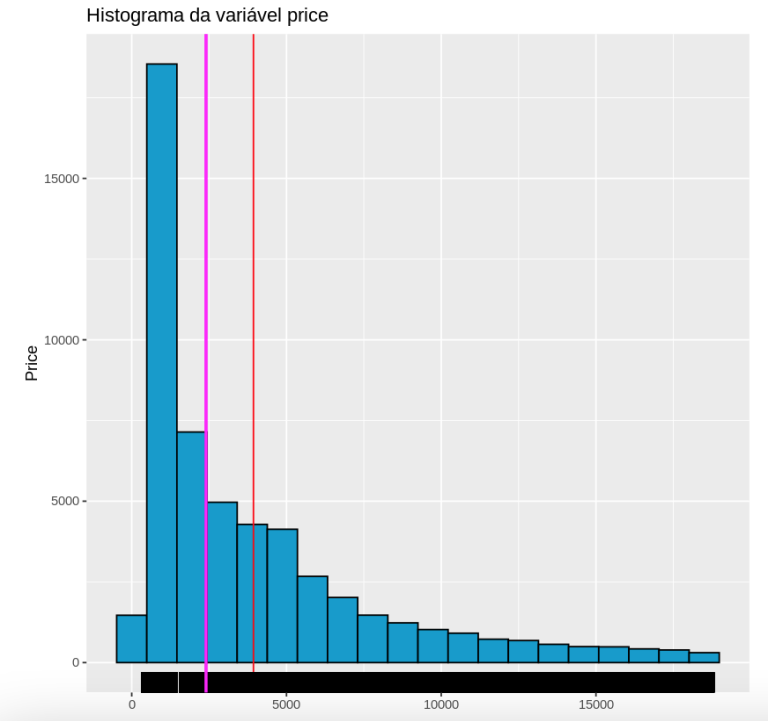
Na parte superior do gráfico podemos observar pontos que indicam valores fora do padrão. Provavelmente esses valores se referem aos casos dos diamantes mais raros e puros, que possuem os preços mais elevados. Mas poderemos tirar essa conclusão com as análises a seguir.

Veja a distribuição da variável (histograma); observe a faixa de valores da variável e também.

Para obtermos o histograma da variável price, foi utilizado o código abaixo:

```
# Grafico histograma
ggplot(data=diamantes) +
  geom_histogram(
    aes(price),bins = 20, fill = "deepskyblue3", color = "black") +
    labs(title = "Histograma da variável price", y = "Price",
         x = "") +
    geom_vline(xintercept = 3933, color = "red") +
    geom_rug(aes(price)) +
    geom_vline(xintercept = median(diamantes$price), color = "magenta", lwd = 1)
```

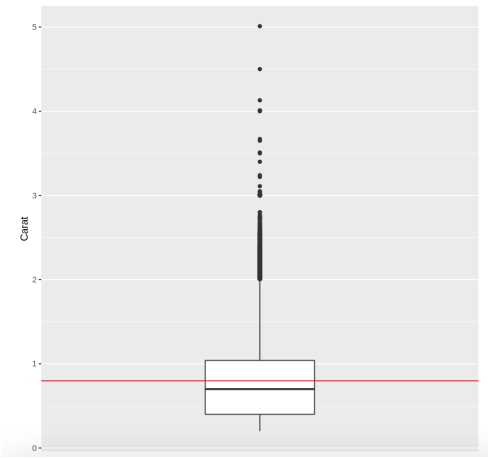
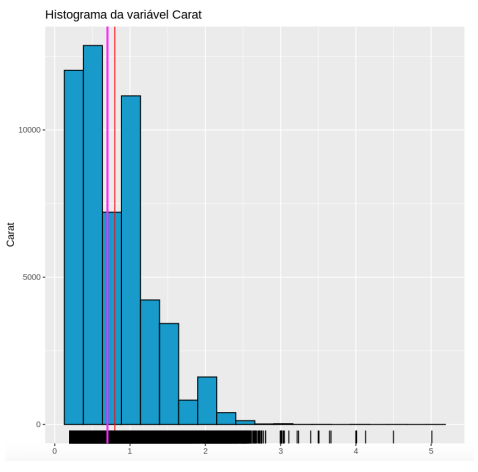
O gráfico obtido foi:



Analisando o histograma podemos confirmar a assimetria observada no BoxPlot. A média representada pela linha vermelha e a mediana representada pela linha em magenta não estão próximas entre si e também não estão próximas dos valores mais observados.

Explore também as variáveis carat, cut, color, clarity, x, y, z, depth e table, seguindo o modelo de exploração.
Crie boxplots para as variáveis numéricas; veja se existem dados anormais (outliers).

Na sequência iremos explorar as demais variáveis utilizando BoxPlot e o histograma como base para sabermos se os valores das médias, mediana, ….. fornecidos pelo comando summary estão de acordo com o que observamos.

Carat	
	

carat Min. :0.2000 1st Qu.:0.4000 Median :0.7000 Mean :0.7979 3rd Qu.:1.0400 Max. :5.0100	Em relação a variável carat, podemos observar que a média e a mediana não estão tão distantes uma da outra, porém, quando comparamos os valores com o primeiro e terceiro quartil vemos uma certa distância. Isso pode ser comprovado no histograma, onde pudemos ver a assimetria evidenciada. Pelo BoxPlot pudemos observar diversos outliers indicados pelos pontos na parte superior do gráfico.

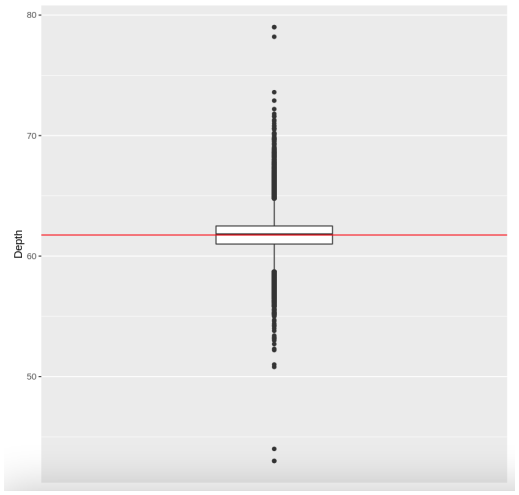
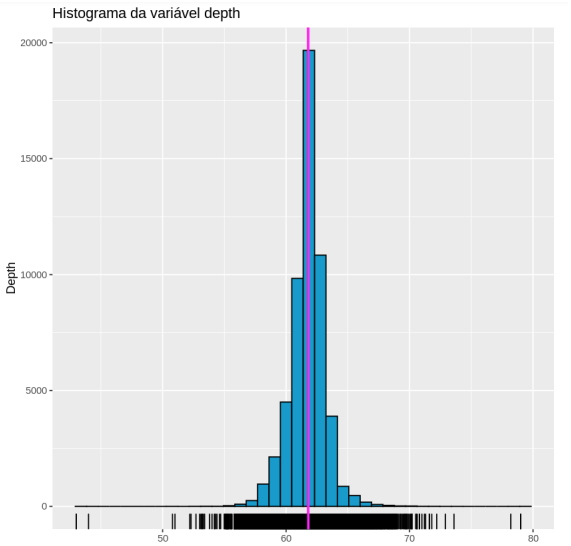
Depht	
	
depth Min. :43.00 1st Qu.:61.00 Median :61.80 Mean :61.75 3rd Qu.:62.50 Max. :79.00	Os valores de média e mediana estão bem próximos e analisando a diferença para o primeiro e terceiro quartis, podemos observar que se trata de uma distribuição simétrica. Porém, analisando o gráfico do BoxPlot, observamos a presença de bastantes outliers.

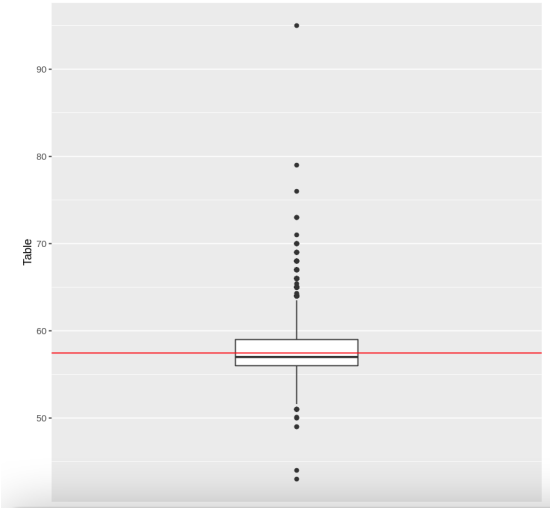
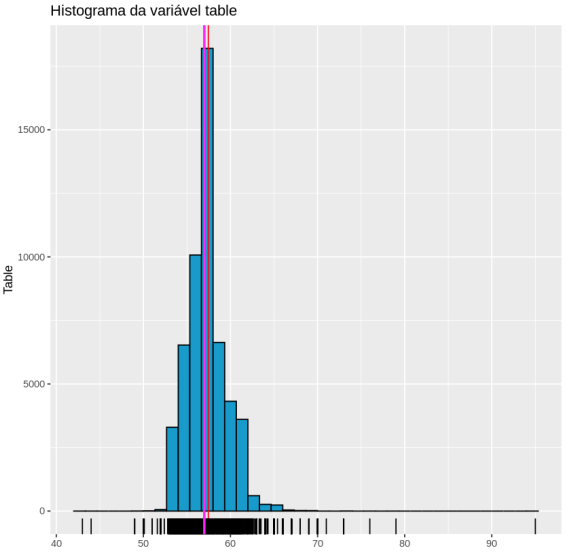
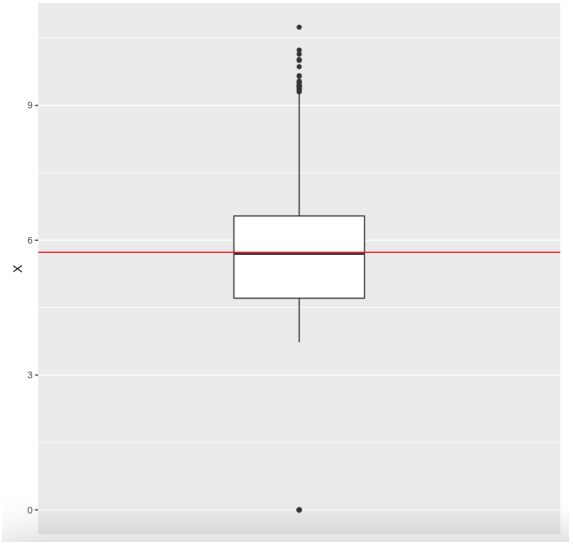
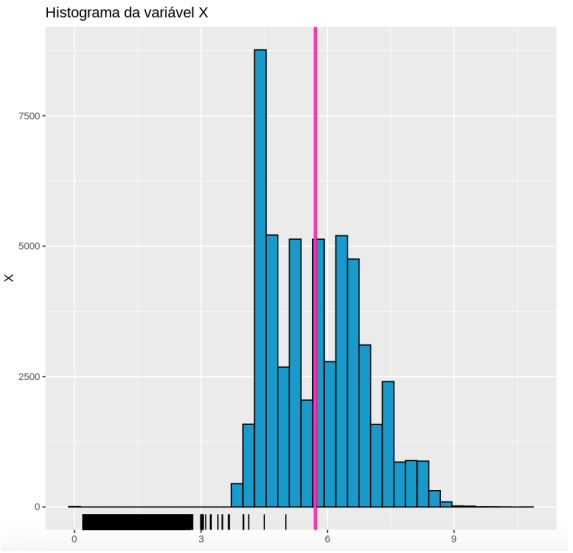
Table	
	

table Min. :43.00 1st Qu.:56.00 Median :57.00 Mean :57.46 3rd Qu.:59.00 Max. :95.00	Para a variável table temos a média e a mediana bem próximas, porém, quando observamos esses valores em relação ao terceiro quartil, vemos uma leve distorção. Analisando o BoxPlot, pudemos observar outleirs significativos.

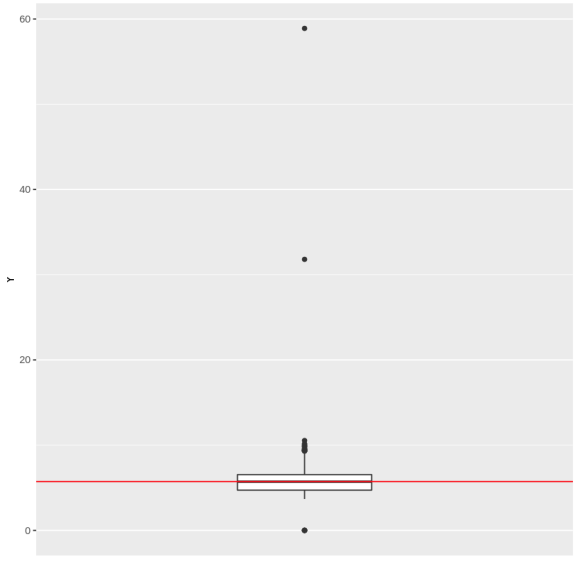
X

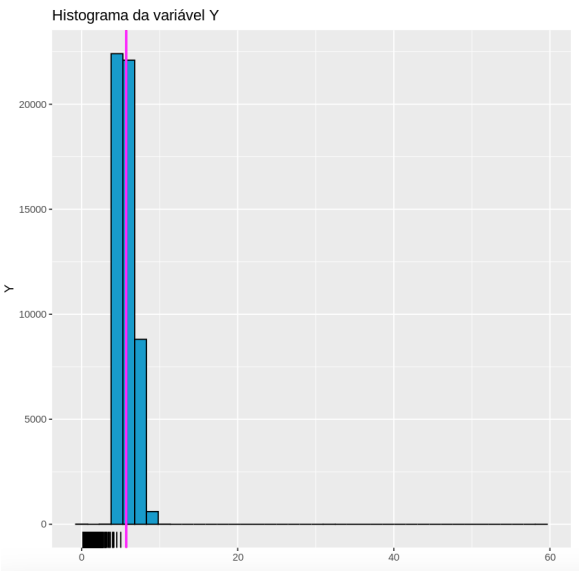




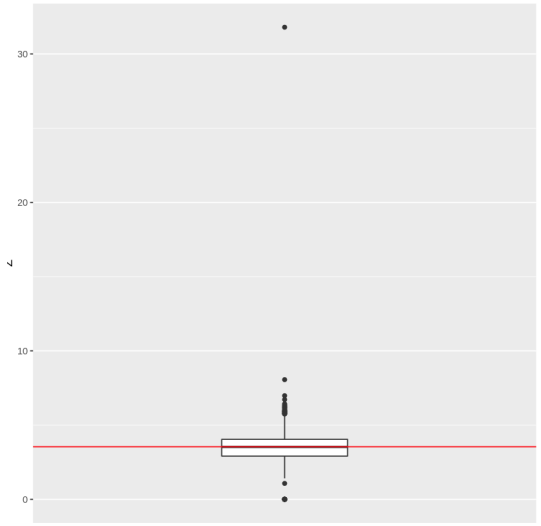
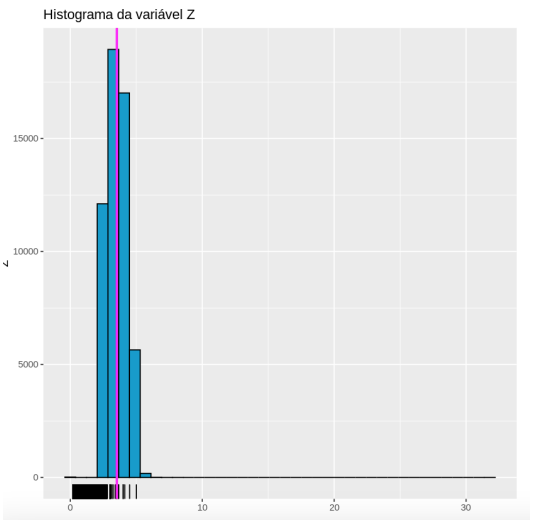
x Min. : 0.000 1st Qu.: 4.710 Median : 5.700 Mean : 5.731 3rd Qu.: 6.540 Max. :10.740	Em relação a variável X, podemos observar que a média e a mediana estão bem próximas e que considerando o primeiro e o terceiro quartis temos uma distribuição uniforme. Porém essa distribuição possui um desvio padrão baixo, sendo possível observarmos que a distribuição está achatada. Os outliers observados no boxplot não parecem ser significativos.

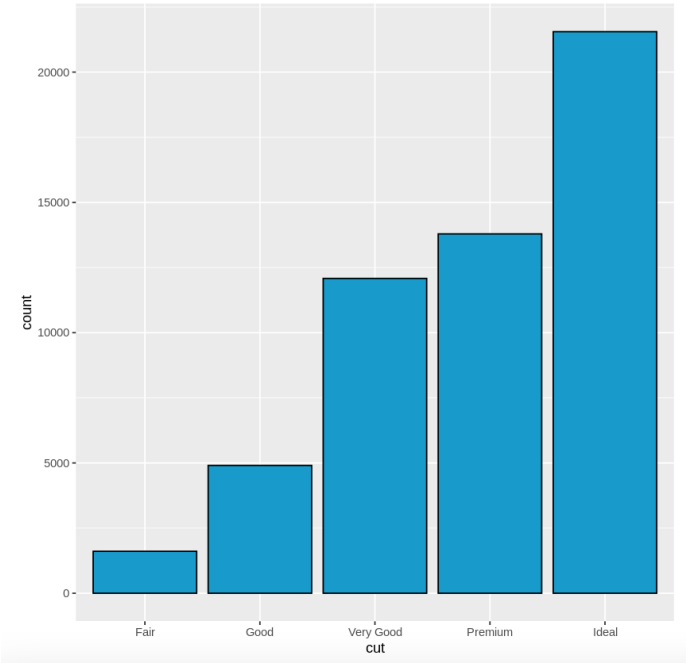
Y

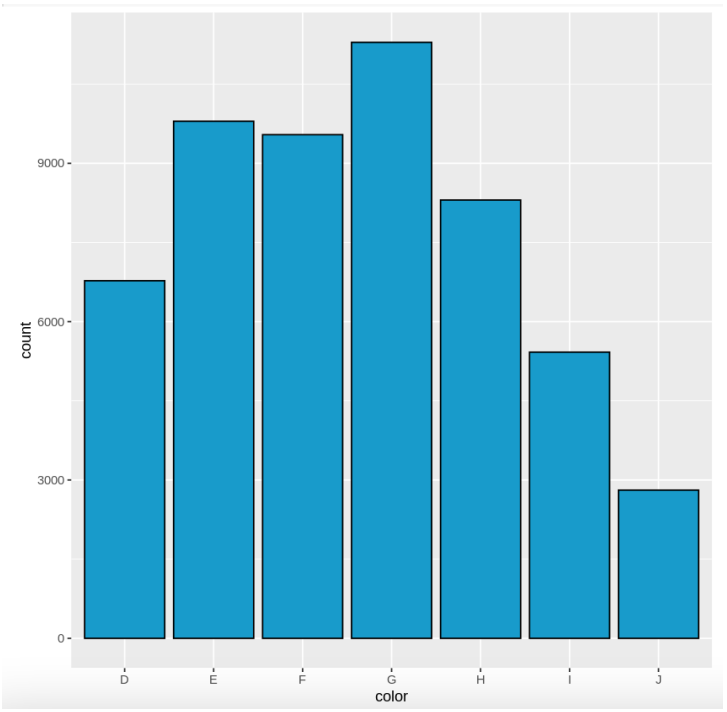


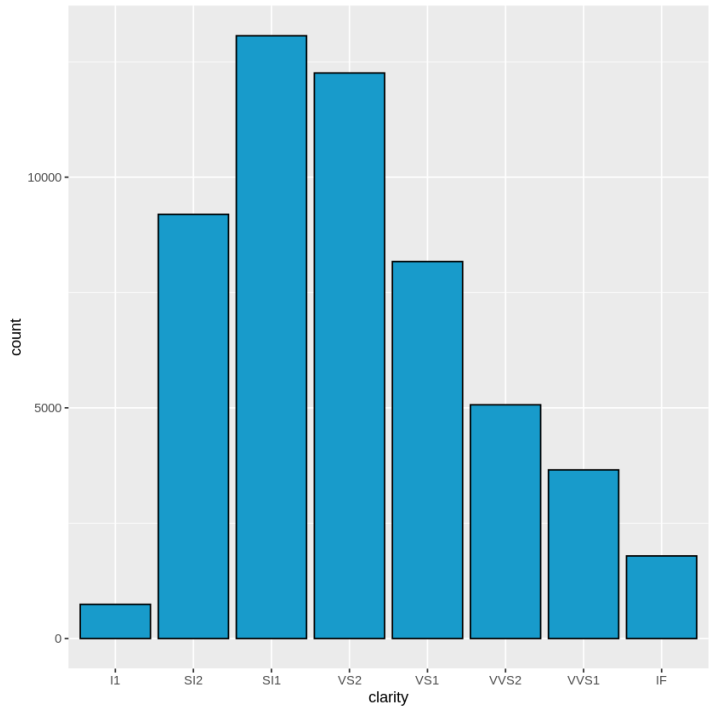


y Min. : 0.000 1st Qu.: 4.720 Median : 5.710 Mean : 5.735 3rd Qu.: 6.540 Max. :58.900	Analisando os valores de média e mediana quase iguais, assim como os quartis 1 e 3, podemos dizer que se trata de uma distribuição simétrica. Porém, no gráfico boxplot observamos alguns outliers que precisam ser estudados, pois podem causar algum distúrbio na análise.

Z	
	
z Min. : 0.000 1st Qu.: 2.910 Median : 3.530 Mean : 3.539 3rd Qu.: 4.040 Max. :31.800	Assim como para a variável Y, na variável Z pudemos observar que os valores de média e mediana são quase iguais, assim como os quartis 1 e 3, podemos dizer que se trata de uma distribuição simétrica. Porém, no gráfico boxplot observamos alguns outliers que precisam ser estudados, pois podem causar algum distúrbio na análise.

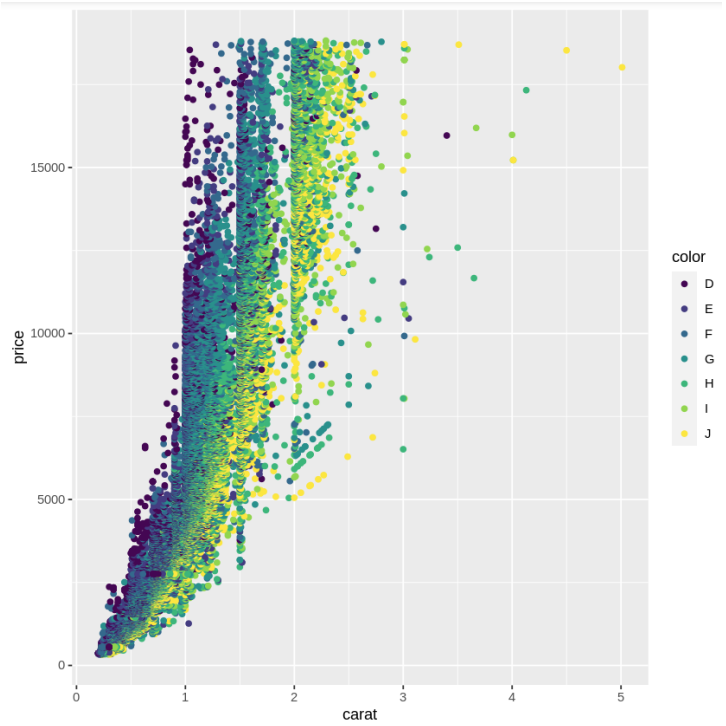
cut	
	Com o gráfico podemos observar que o corte ideal é o mais comum, ou seja, é a moda.

color	
	Com o gráfico podemos observar que a cor G é a moda da amostra

clarity	
	Com o gráfico podemos observar que a claridade SI1 é a mais comum, ou seja, é a moda da amostra.

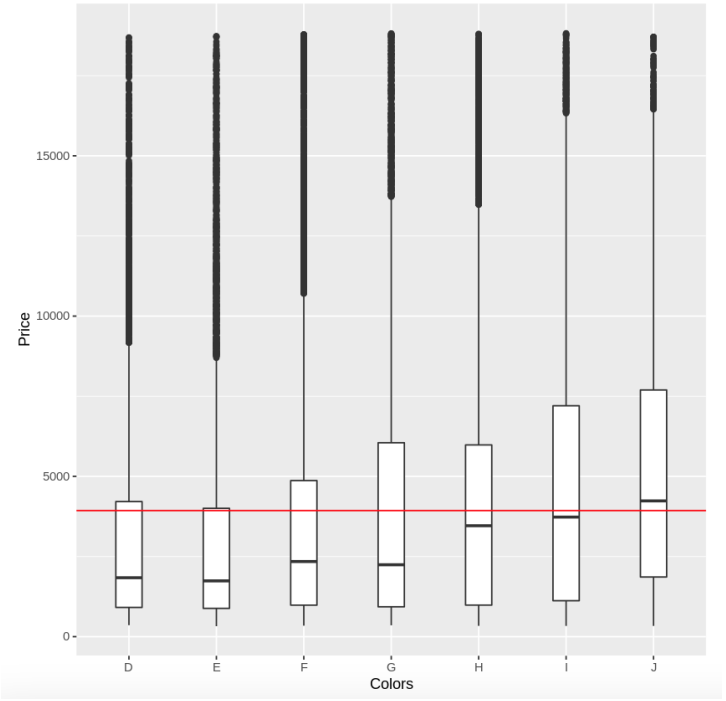
Utilize as variáveis categóricas para fazer o facetamento dos dados, mostrando alguns gráficos com 2 ou mais variáveis contínuas lado a lado. Para cada resultado/gráfico obtido, explique e discuta-os, de modo a construir um relatório de exploração dos dados, que deverá ser submetido

Análise 1



Observando o gráfico de pontos onde temos price x carat, observamos que os quilates não são decisivos para o preço dos diamantes.
Até o primeiro quilate podemos dizer que se trata de uma relação linear, mas acima de um quilate, temos diversos casos aonde diamantes de baixo quilate tem uma valor alto.
Podemos notar uma relação com as cores dos diamantes, pois os diamantes de baixo quilate de possuem a cor D (melhor) tem o preço maior que os de cor J (pior cor) apresentado em amarelo no gráfico

Análise 2



Ao analisarmos o gráfico ao lado onde a linha vermelha indica a média de preço dos diamantes, percebemos que para as piores cores o preço vai ficando perto do preço médio.
Nas melhores cores (D), por mais que o preço esteja menor que o preço médio, podemos observar que a presença de outliers é muito maior do que nos casos das piores cores (J).