Predição de Preço do Diamante com Modelos Paramétricos

“Um beijo na mão pode fazer você se sentir bem, mas uma tiara de diamantes é para sempre.”

— **Marilyn Monroe**

Objetivo e Motivação

Objetivo

Nosso objetivo é **entender** melhor esse mercado bilionário e **prever os preços ideais** para a venda dos diamantes, de acordo com algumas **características** relacionadas ao diamante, mostrando que os valores são estabelecidos conforme alguns **critérios**.

$82

Bilhões de dólares

Demanda global por joias com diamantes em 2017.

Maior diamante bruto

Maior diamante do mundo em estado bruto é vendido por **US$ 53 milhões**, a pedra preciosa de **1.109 quilates** (um quilate equivale a 0,20 gramas), do **tamanho de uma bola de tênis**, foi encontrada em Botswana (País na África Austral).

Diamante no estado bruto (sem corte)

Maior diamante lapidado

Maior diamante lapidado do mundo é vendido por **R$ 22,2 milhões**, trata-se de um raro diamante negro conhecido como **“O Enigma”**. Acredita-se que a raridade de **555,5 quilates** tenha vindo do **espaço**.

“O Enigma”, lapidada com 55 facetas

CONJUNTO DE DADOS

O dataset possui 53794 observações de diamantes e 10 atributos relacionados .

PRICE

Preço em Dólares
(**$**326 até **$**18,823)

CARAT

Peso do diamante
(0.2 até 5.01 quilates)

CUT

Qualidade do corte
Fair: razoável
Good: bom
Very Good: muito bom
Premium: premium
Ideal: ideal

COLOR

Cor do diamante
do D (melhor)
até o J (pior)

CLARITY

Grau de pureza do diamante
Do pior até o melhor: (I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF).

X

Comprimento em mm
(0 até 10.74)

Y

Largura em mm
(0 até 58.9)

Z

Espessura em mm
(0 até 31.8)

DEPTH

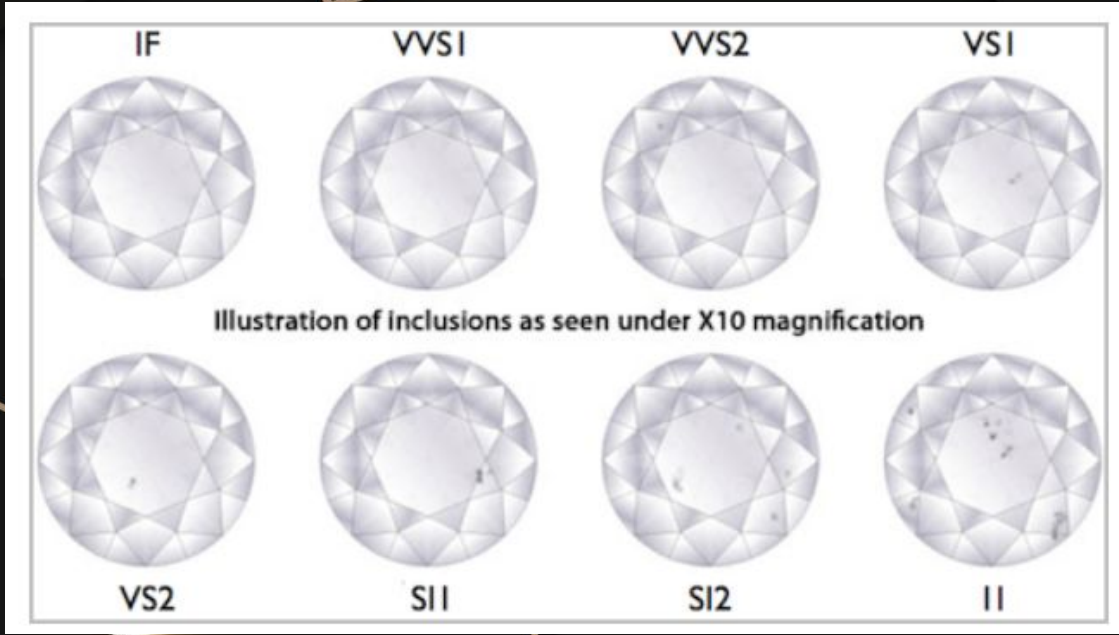
Porcentagem da espessura
= 2*Z/(X+ Y)
(43 até 79)

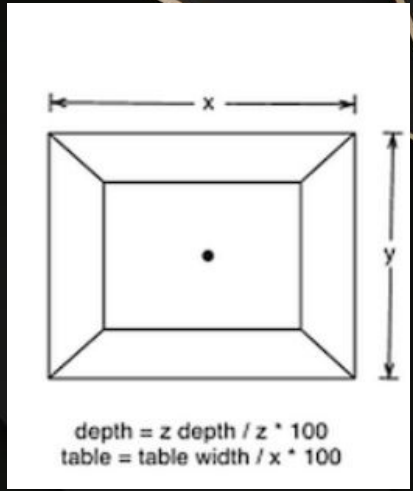
TABLE

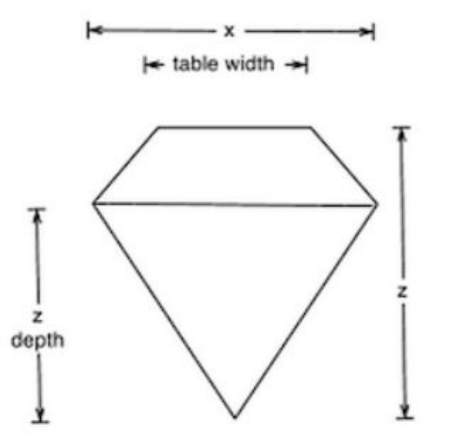
Largura do topo do diamante em relação ao ponto mais largo
(43 até 95)

COLOR GRADING SCALE

D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Colorless			Near Colorless				Faint Yellow			Very Light Yellow					Light Yellow							


IF
VVS1
VVS2
VS1
Illustration of inclusions as seen under X10 magnification
VS2
SI1
SI2
I1


x
y
depth = z depth / z * 100
table = table width / x * 100


x
table width
z
depth
z

Primeiras observações

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

VERIFICAÇÃO DOS DADOS

Tipo das variáveis

```
RangeIndex: 53940 entries, 0 to 53939
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    53940 non-null  float64
 1   cut      53940 non-null  object
 2   color    53940 non-null  object
 3   clarity  53940 non-null  object
 4   depth    53940 non-null  float64
 5   table    53940 non-null  float64
 6   price    53940 non-null  int64
 7   x        53940 non-null  float64
 8   y        53940 non-null  float64
 9   z        53940 non-null  float64
dtypes: float64(6), int64(1), object(3)
```

Os tipos das variáveis condizem com as descrições delas.

Valores Nulos

carat	0
cut	0
color	0
clarity	0
depth	0
table	0
price	0
x	0
y	0
z	0

Aqui temos a soma da quantidade de valores nulos para cada variável, como são todas 0, não há dados faltantes

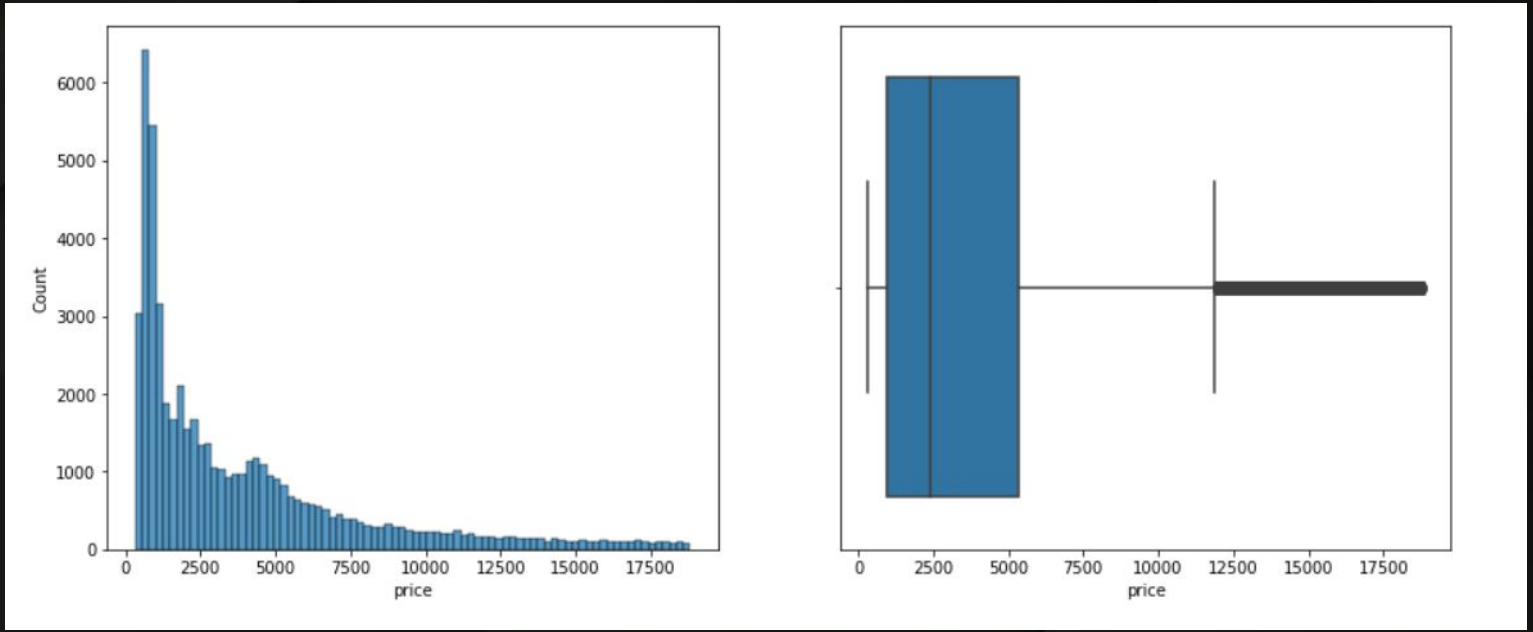
0.27% de Valores duplicados

Existem no total 146 linhas duplicadas e representam 0.27% das linhas do nosso conjunto.

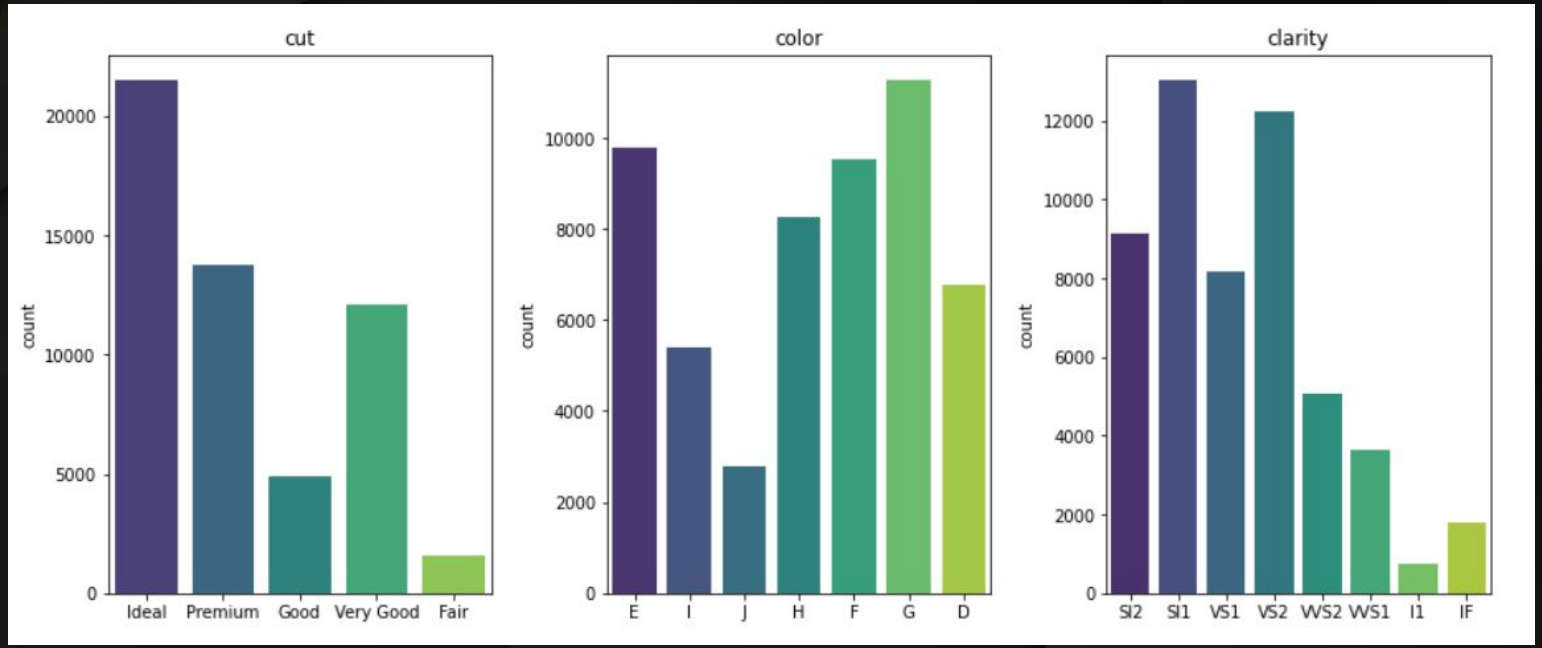
02

Análise Exploratória

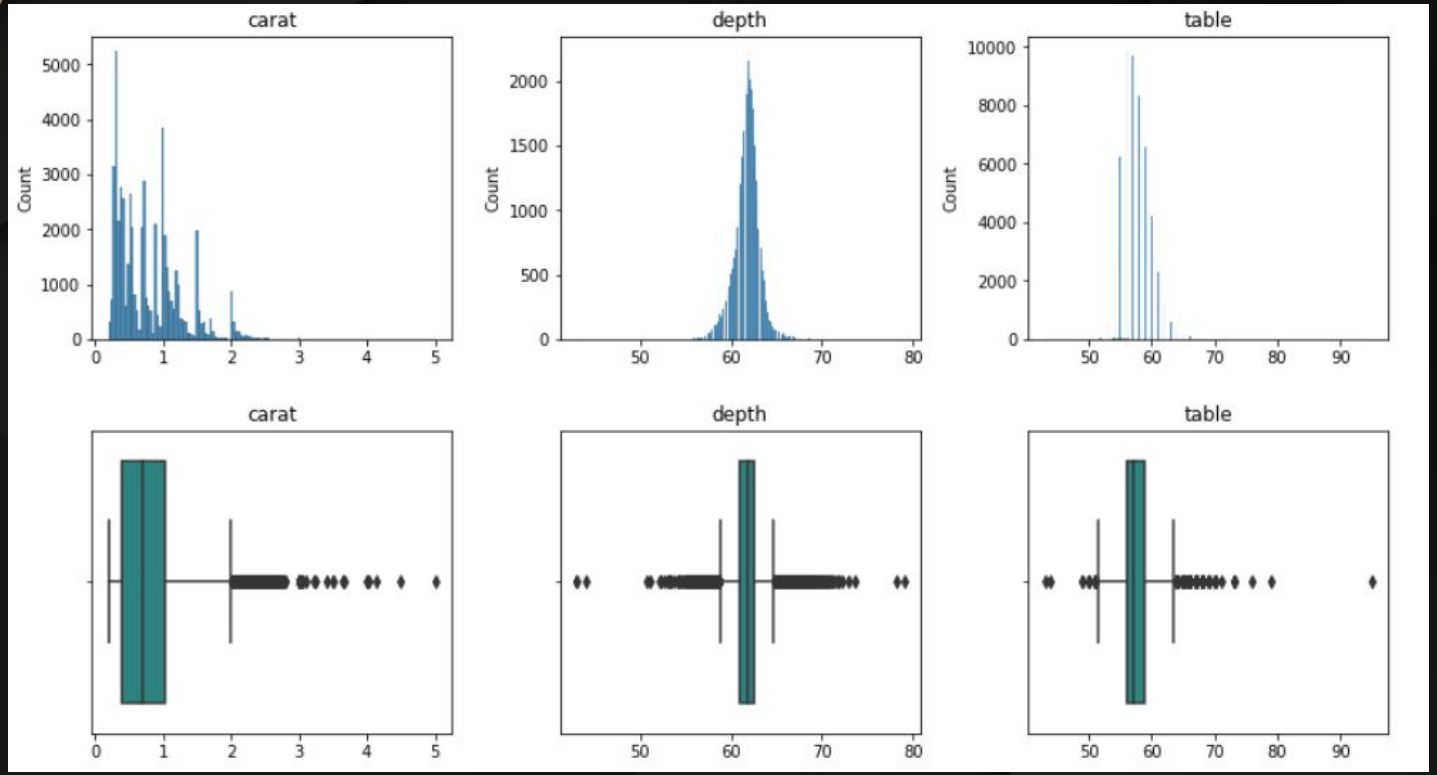
Variável Resposta (Price)



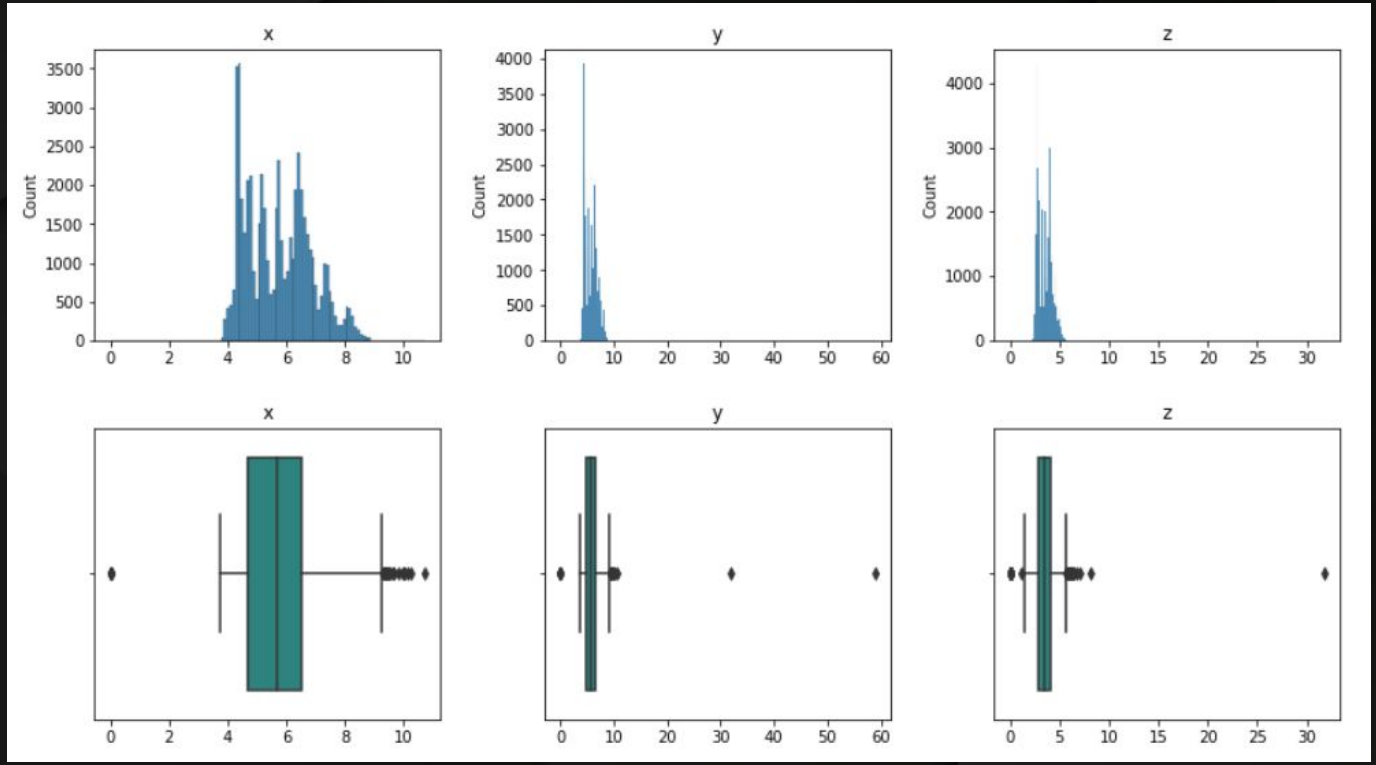
Variáveis Categóricas



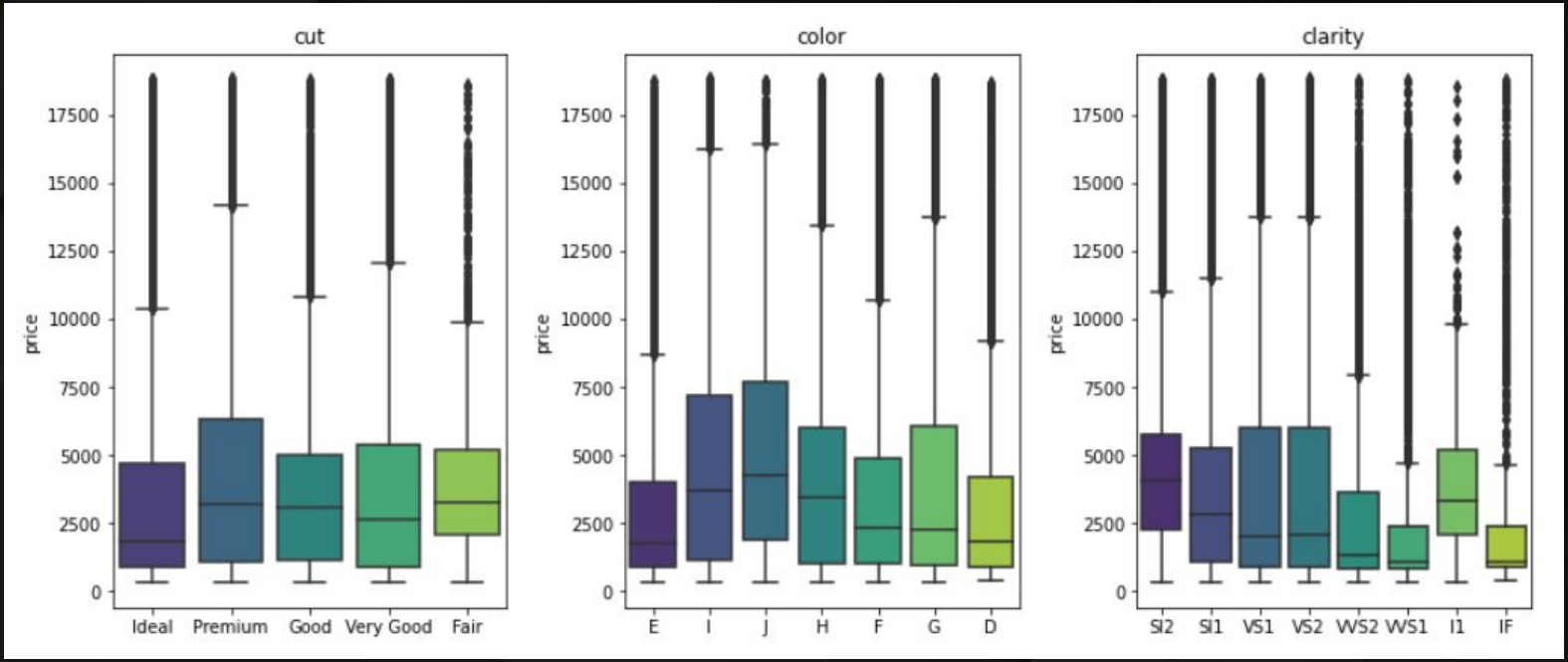
Variáveis Quantitativas



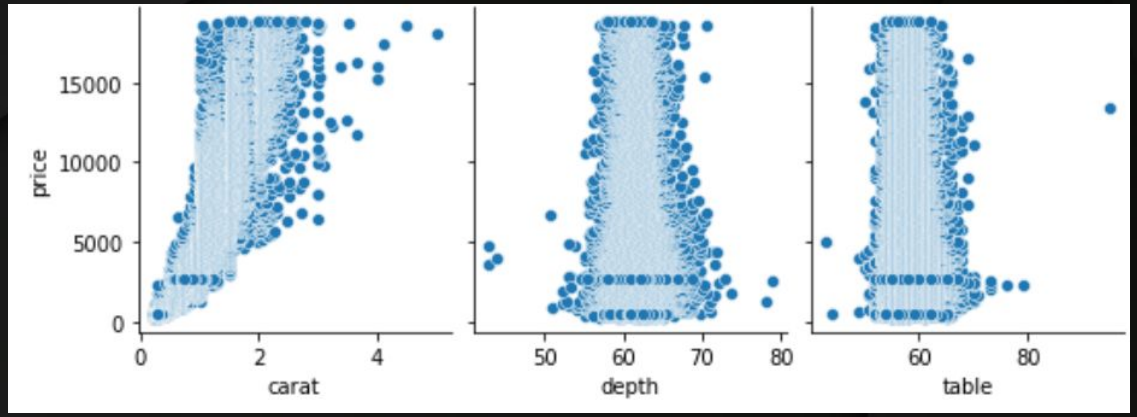
Variáveis Quantitativas

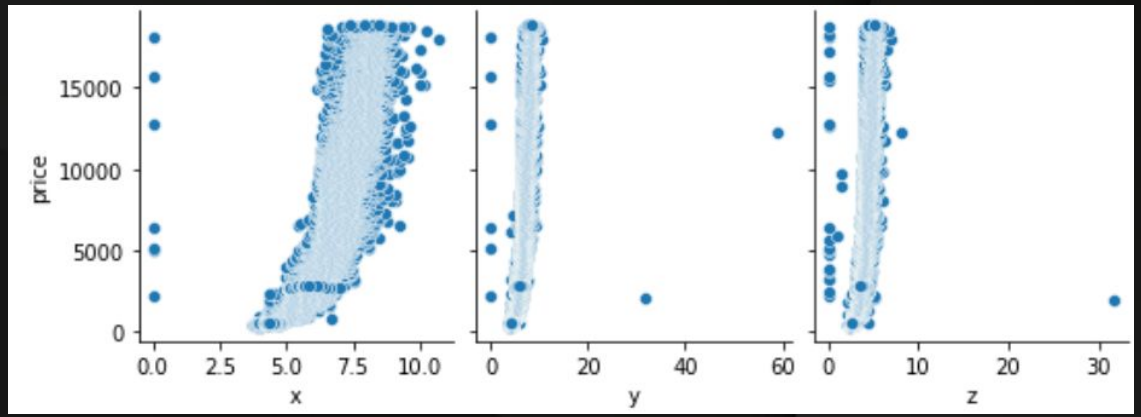


Variáveis Categóricas vs. Price

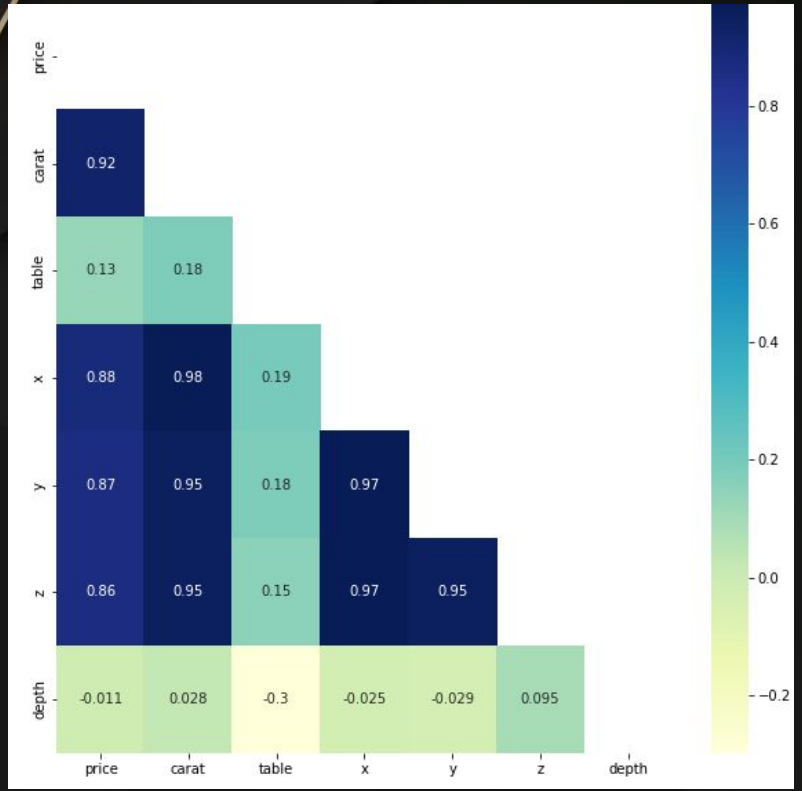


Variáveis Quantitativas vs. Price





Mapa de Correlações



Note que as variáveis ‘x’, ‘y’, ‘z’ possuem uma correlação muito forte com o ‘carat’ (peso do diamante), o que pode ser facilmente compreendido. Uma vez que o peso é influenciado diretamente pelo tamanho do diamante.

Tendo isso em mente, retiramos as variáveis ‘x’, ‘y’, ‘z’, dado que já é explicado pela variável ‘carat’.

03

Pré-processamento dos Dados

Encoding

Como essas variáveis qualitativas são ordinais, definimos de acordo com a ordem da característica:

```
# Dicionário para encoding
cut = {'Fair':1, 'Good':2, 'Very Good':3, 'Premium':4, 'Ideal':5}
color = {'J':1, 'I':2, 'H':3, 'G':4, 'F':5, 'E':6, 'D':7}
clarity = {'I1':1,  'SI2':2, 'SI1':3, 'VS2':4, 'VS1':5, 'VVS2':6, 'VVS1':7, 'IF':8}
```

Visualização

	carat	depth	table	cut	color	clarity	price
0	0.23	61.5	55.0	5	6	2	326
1	0.21	59.8	61.0	4	6	3	326
2	0.23	56.9	65.0	2	6	5	327
3	0.29	62.4	58.0	4	2	4	334
4	0.31	63.3	58.0	2	1	2	335