

Relatorio - Análise 1

Matheus Elero

14 outubro 2021

Contents

| | |
|-------------------------------------|----------|
| 1 Dados | 2 |
| 2 Análise Descritiva | 3 |
| 2.1 Variáveis | 3 |
| 2.2 Análise de Correlação | 5 |
| 3 Conclusão | 7 |

1 Dados

Os dados apresentados por esse relatório exploram o preço de diamantes e suas características. Para isso, foi extraído uma base chamada Diamonds da Biblioteca GGPlot 2. O Head dos dados está apresentado pela Tabela 1 em formato Wide, o conjunto total possui 11 colunas e 53940 linhas.

| ind | carat | cut | color | clarity | depth | table | price | x | y | z |
|-----|-------|------|-----------|---------|-------|-------|-------|-----|------|------|
| 1 | 1 | 0.23 | Ideal | E | SI2 | 61.5 | 55 | 326 | 3.95 | 3.98 |
| 2 | 2 | 0.21 | Premium | E | SI1 | 59.8 | 61 | 326 | 3.89 | 3.84 |
| 3 | 3 | 0.23 | Good | E | VS1 | 56.9 | 65 | 327 | 4.05 | 4.07 |
| 4 | 4 | 0.29 | Premium | I | VS2 | 62.4 | 58 | 334 | 4.20 | 4.23 |
| 5 | 5 | 0.31 | Good | J | SI2 | 63.3 | 58 | 335 | 4.34 | 4.35 |
| 6 | 6 | 0.24 | Very Good | J | VVS2 | 62.8 | 57 | 336 | 3.94 | 2.48 |

Tabela 1: Head dos dados extraídos

Cada coluna representa a seguinte descrição:

| Coluna | Classificação | Variável | Descrição |
|---------|---------------|----------|---|
| Price | Quantitativa | Contínua | Preço do diamante em Dólar |
| Carat | Quantitativa | Contínua | Peso do diamante |
| Cut | Qualitativa | Nominal | Qualidade do Corte |
| Color | Qualitativa | Nominal | Cor D(Melhor) para J(Pior) |
| Clarity | Qualitativa | Nominal | Medição de quão claro é o diamante (I1 (pior), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (melhor)) |
| x | Quantitativa | Contínua | Comprimento |
| y | Quantitativa | Contínua | Largura |
| z | Quantitativa | Contínua | Profundidade |
| Depth | Quantitativa | Contínua | porcentagem de profundidade total = z / média (x, y) = 2 * z / (x + y) (43-79) |
| Table | Quantitativa | Contínua | largura do topo do diamante em relação ao ponto mais largo (43-95) |

Tabela 2: Descrição dos Headers.

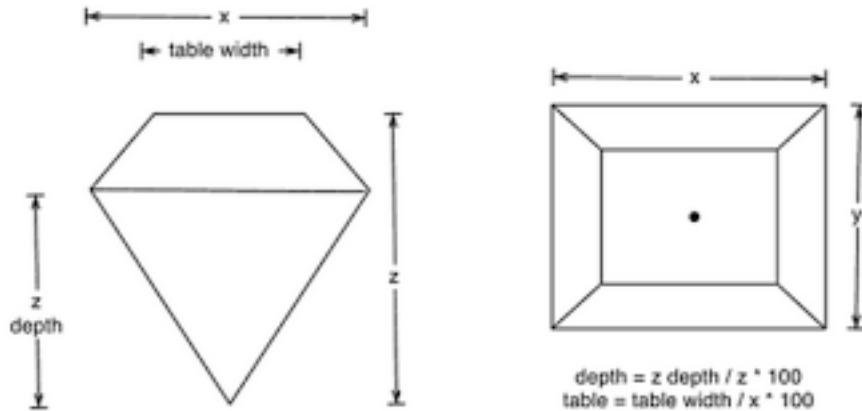


Figure 1: Dimensões do Diamante

2 Análise Descritiva

2.1 Variáveis

Aqui será apresentada uma análise geral a respeito de todos os parâmetros da base de dados. A começar pela variável principal, o preço, que por uma visão mais simplista possui média igual 3932.7997219 Dólares, e desvio padrão igual a 3989.4397381. Um histograma da Figura 2 mostra a alta variabilidade dos dados, o que indica uma dificuldade grande para prever preços de diamantes, pois estes não seguem um padrão óbvio e assertivo.

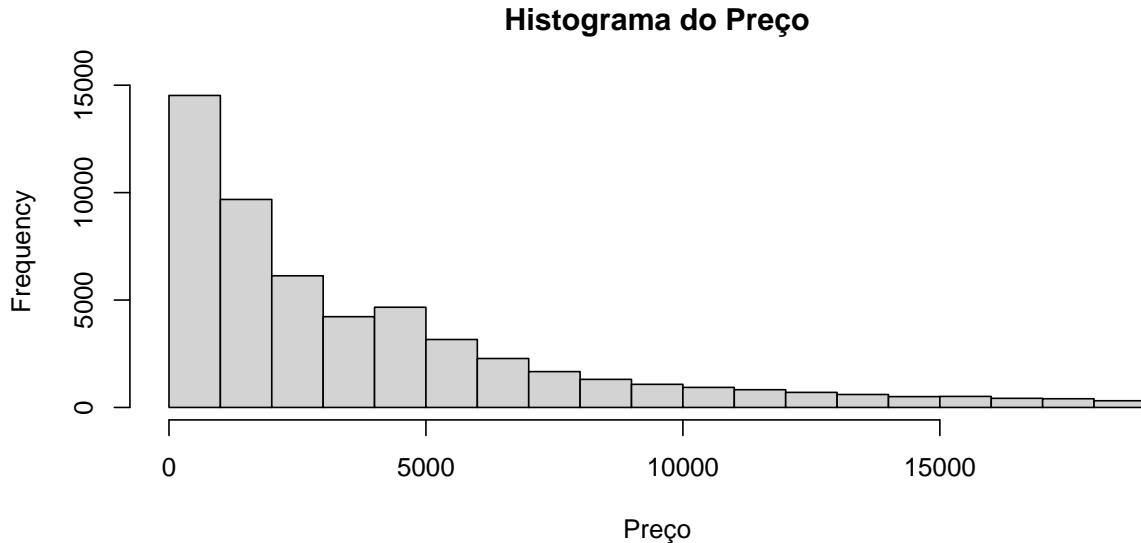


Figure 2: Histograma Preço

Para os valores dimensões x, y, z, depth, table e carat, os dados são apresentados pela tabela a seguir:

| Coluna | Média | Variância | Desvio Padrão |
|--------|------------|-----------|---------------|
| x | 5.7311572 | 1.2583472 | 1.1217607 |
| y | 5.734526 | 1.3044716 | 1.1421347 |
| z | 3.5387338 | 0.4980109 | 0.7056988 |
| Depth | 61.7494049 | 2.0524038 | 1.4326213 |
| Table | 57.4571839 | 4.9929481 | 2.2344906 |
| Carat | 0.7979397 | 0.2246867 | 0.4740112 |

Tabela 3: Média, Variância e Desvio para as grandezas quantitativas e contínuas

Uma visualização melhor pode ser vista nos histogramas da Figura 3, onde é possível observar que as dimensões x, y e z possuem uma variabilidade maior, mas em questão de Depth e Table, os dados estão mais concentrados em torno da média. Esses resultados mostram que existe diversidade nos diamantes catalogados, com várias dimensões de x, y e z. O parâmetro “Carat” também possui alta variabilidade, que é consequente aos valores dimensionais, pois essa característica é relacionado ao tamanho e peso da pedra.

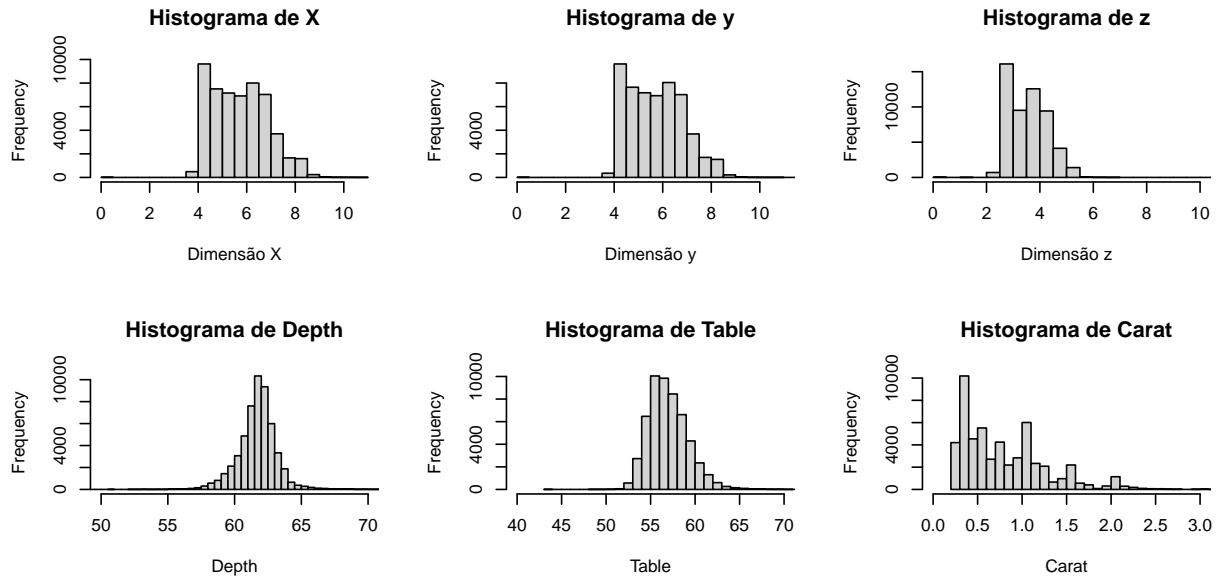


Figure 3: Histogramas das características contínuas dos dados.

Para as colunas com variáveis Qualitativas e Nominais, foi utilizado um gráfico de pizza para apresentar os resultados (Figura 4). Como é possível observar, em relação aos cortes 40% são considerados ideais, e 26% premium. Para clareza, apenas 1% tem o pior valor de “I1” e 3% do melhor “1F”, e a maioria dos diamantes estão classificados como SI1 VS2, que representam claresas intermediárias. Já para as cores, existe certo equilíbrio, porém apenas 5% dos dados são relativos a pior cor(J) e 13% como a melhor (D).

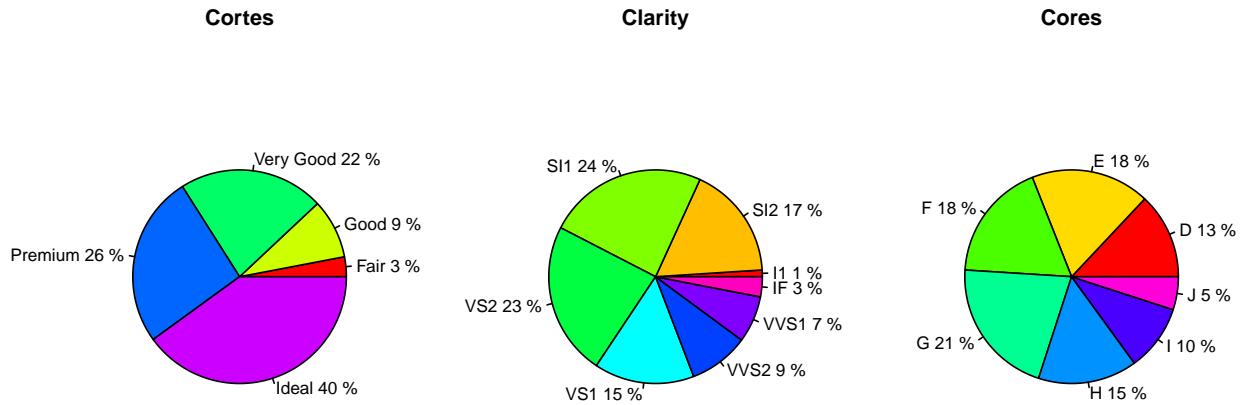


Figure 4: Gráficos de Pizza com as variáveis qualitativas nominais

2.2 Análise de Correlação

A Matriz de Correlação disponibilizada a seguir apresenta o Coeficiente de Correlação calculado para todos os parâmetros par a par. Assim é possível identificar quais deles possuem alta relação linear, como por exemplo preço com carat, x, y e z. As variáveis que tem maiores valores são entre x, y e z, o que faz sentido, pois tratam-se das dimensões do diamante, e naturalmente uma pode depender da outra. Para tornar a análise mais objetiva, serão consideradas apenas correlações absolutas maiores que 0.8.

| | ind | carat | cut | color | clarity | depth | table | price | x | y | z |
|---------|-------|-------|-------|-------|---------|-------|-------|-------|-------|-------|-------|
| ind | 1.00 | -0.38 | 0.10 | -0.10 | 0.21 | -0.03 | -0.10 | -0.31 | -0.41 | -0.40 | -0.40 |
| carat | -0.38 | 1.00 | -0.13 | 0.29 | -0.35 | 0.03 | 0.18 | 0.92 | 0.98 | 0.95 | 0.95 |
| cut | 0.10 | -0.13 | 1.00 | -0.02 | 0.19 | -0.22 | -0.43 | -0.05 | -0.13 | -0.12 | -0.15 |
| color | -0.10 | 0.29 | -0.02 | 1.00 | 0.03 | 0.05 | 0.03 | 0.17 | 0.27 | 0.26 | 0.27 |
| clarity | 0.21 | -0.35 | 0.19 | 0.03 | 1.00 | -0.07 | -0.16 | -0.15 | -0.37 | -0.36 | -0.37 |
| depth | -0.03 | 0.03 | -0.22 | 0.05 | -0.07 | 1.00 | -0.30 | -0.01 | -0.03 | -0.03 | 0.09 |
| table | -0.10 | 0.18 | -0.43 | 0.03 | -0.16 | -0.30 | 1.00 | 0.13 | 0.20 | 0.18 | 0.15 |
| price | -0.31 | 0.92 | -0.05 | 0.17 | -0.15 | -0.01 | 0.13 | 1.00 | 0.88 | 0.87 | 0.86 |
| x | -0.41 | 0.98 | -0.13 | 0.27 | -0.37 | -0.03 | 0.20 | 0.88 | 1.00 | 0.97 | 0.97 |
| y | -0.40 | 0.95 | -0.12 | 0.26 | -0.36 | -0.03 | 0.18 | 0.87 | 0.97 | 1.00 | 0.95 |
| z | -0.40 | 0.95 | -0.15 | 0.27 | -0.37 | 0.09 | 0.15 | 0.86 | 0.97 | 0.95 | 1.00 |

Tabela 4: Matriz de Correlação

As correlações analisadas a seguir são listadas abaixo:

- Preço x Carat: 0.92
- Preço x X: 0.88
- Preço x Y: 0.87
- Preço x Z: 0.86
- Carat x X: 0.98
- Carat x Y: 0.95
- Carat x Z: 0.95
- Y x X: 0.97
- Y x Z: 0.95
- X x Z: 0.97

Os gráficos de dispersão apresentados pela Figura 5 mostram claramente a relação relativa entre as variáveis, em algumas delas o comportamento é quase linear, como x em relação a y, x em relação a z, e outros. Como visto na análise anterior, o preço possui alta dispersão e desvio, sendo possível visualizar de outra forma na Figura 5, pois os pontos estão mais dispersos, devido a alta diversidade de valores.

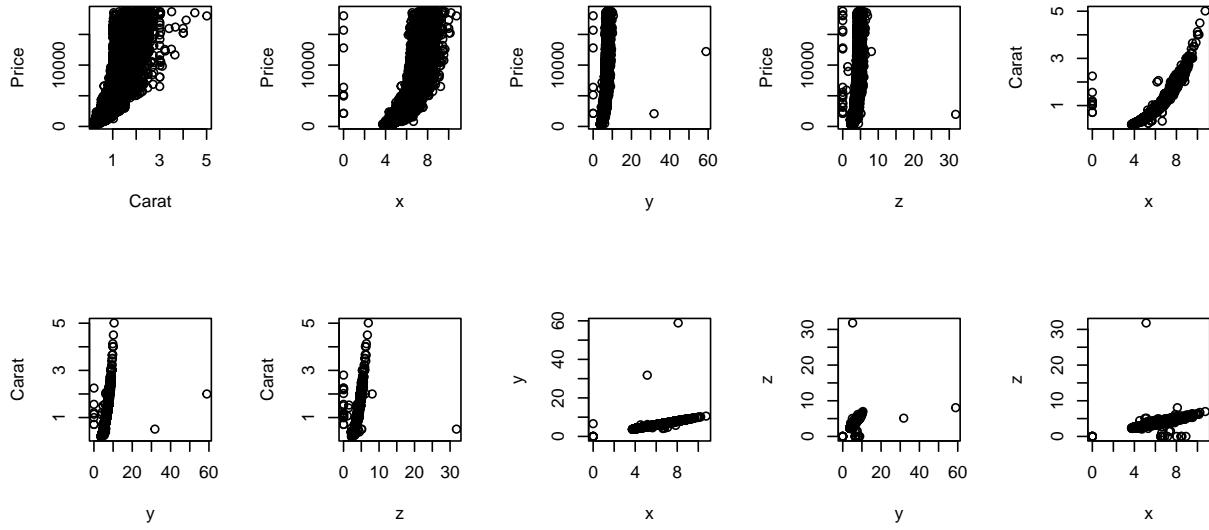


Figure 5: Gráficos de Dispersão para Correlação com Outiliers

Com uma inspeção visual dos gráficos de dispersão é possível identificar Outliers, ou valores fora do comum, que podem ser excessões, erros de coleta ou ruídos. Para tornar a análise de correlação mais efetiva, foi realizada a remoção desse pontos, que resultou em uma melhor visualização (Figura 6). Portanto, as relações de dimensões(x, y e z) possuem comportamento quase que linear, já preço e carat tem um padrão semelhante ao exponencial (Correlação Não-Linear).

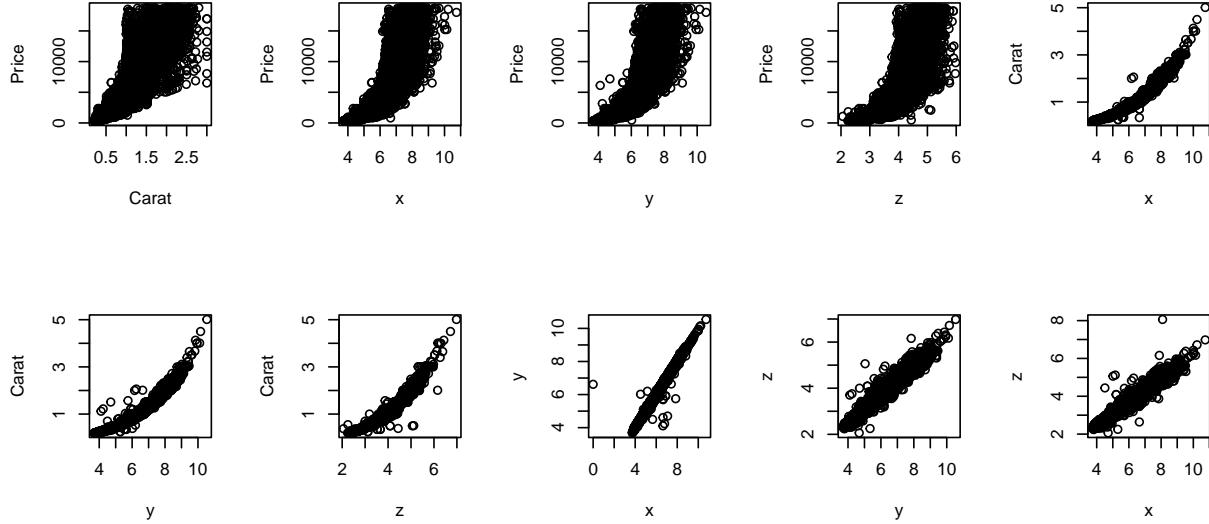


Figure 6: Gráficos de Dispersão para Correlação com remoção de Outiliers

Exceto nas relações entre Preço e Carat, e Carat e X, todas as outras apresentaram resultados correlações melhoradas após a remoção de outliers. Como os dados estão arredondados para 2 casas decimais, os casos

de x, y e z apresentaram correlação próxima de 1, um resultado bastante expressivo, porém extremamente lógico, pois tratam-se de dimensões do diamante, e podem ter relações naturalmente dependentes. Carat também possui uma boa correlação com essa grandeza, o que é também condizente, pois Carat é a principal característica utilizada para classificar o peso e dimensão de um diamante. A Tabela 5 apresenta um comparativo entre os dados de correlação.

| X e Y | com Outliers | sem Outliers |
|---------------|--------------|--------------|
| Carat e Price | 0.92 | 0.92 |
| X e Price | 0.88 | 0.89 |
| Y e Price | 0.87 | 0.89 |
| Z e Price | 0.86 | 0.88 |
| Carat e X | 0.98 | 0.98 |
| Carat e Y | 0.95 | 0.98 |
| Carat e Z | 0.95 | 0.98 |
| Y e X | 0.97 | 1 |
| Z e Y | 0.95 | 0.99 |
| Z e X | 0.97 | 0.99 |

Tabela 5: Comparação de Correlações Com e Sem Outliers

3 Conclusão

A base de dados apresentada nesse relatório possui registros de 53940 Diamantes, com atributos de preço, dimensões, Carat, Cores, Cortes e Clareza. É comum que para esse caso o objetivo principal de estudo dos dados é construir um modelo de regressão para especificar um diamante de forma mais assertiva, pois como é possível observar pela Figura 1, os valores das pedras possuem alta variabilidade, o que faz com que esses valores assumam resultados dispersos e pouco previsíveis.

Portanto, a análise de correlação é fundamental para atribuir o preço e definir padrões. Os resultados das Figuras 5 e 6 mostram a clara relação de dependência Não-Linear entre Preço e Carat, que pode ser utilizado para construção de uma função de regressão adequada. Para esse caso também pode ser utilizada a regressão linear, porém o produto final pode ser pouco preciso. Preço também possui uma alta correlação com X, Y e Z, assim como Carat, e a construção de um modelo com a união entre esses fatores pode contribuir com maior qualidade dos resultados.