Relatorio - Análise 1

Matheus Elero

09/10/2021

Dados

Os dados apresentados por esse relatório exploram o preço de diamantes baseado em algumas de suas características. Para isso, foi extrarído uma base chamada Diamonds da Biblioteca GGPLOT 2. O Head dos dados está apresentado pela Tabela 1 em format Wide, observe que existem 11 colunas e 53940 linhas.

Tabela 1: Head dos dados extraídos

	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	2.43
2	2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	4	0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
5	5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

Cada columa representa a seguinte descrição:

Tabela 2: Descrição dos Headers.

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)

Análise Descritiva

Variáveis

Aqui serão apresentadas uma análise geral a respeito de todos os parâmetos da base de dados. A começar pela variável principal, o preço, que por uma visão mais simplista possui média igual 3932.7997219 Dólares,

variância equivalente e desvio padrão 3989.4397381. Um histograma da Figura 1 mostra a alta variabilidade dos dados, o que representa uma dificuldade grande para prever preços de diamantes, pois estes não seguem um padrão óbio e assertivo.

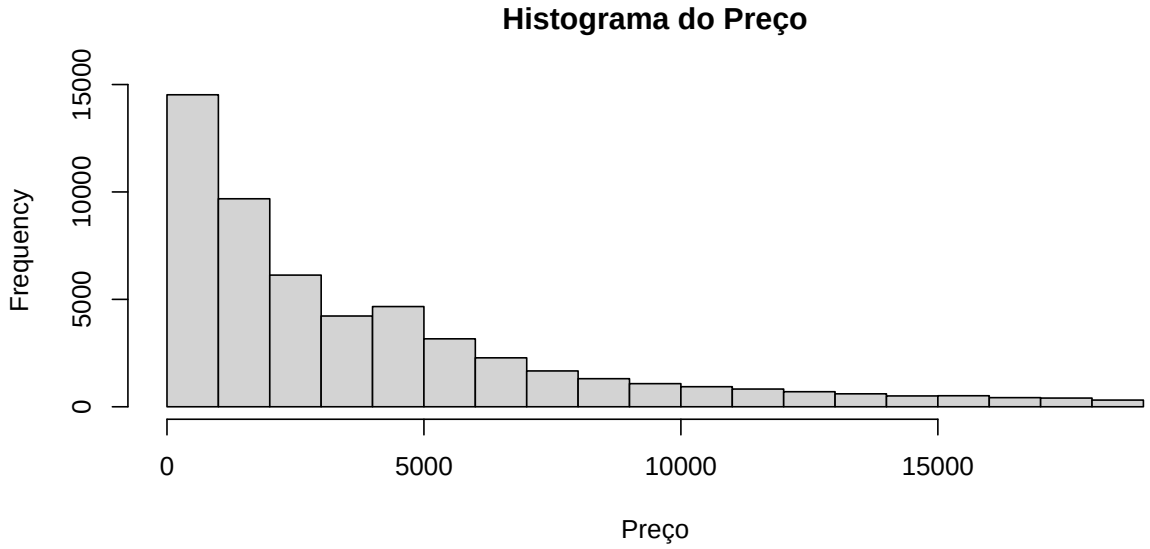

Figure 1: Histograma Preço

Para os valores dimensões x, y, z, depth e table, 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

Uma visuzalização melhor pode ser vista nos histogramas da Figura 2, 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. Isso é possível obervar pela variância e desvio padrão calculados, nota-se que para x e y o desvio padrão foi em torno de 1.1, bastante significativo próxima da média.

Para as colunas com variáveis Qualitativas e Nominais, foi utilizado um gráfico de pizza para apresentar os resultados (Figura 3). 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).

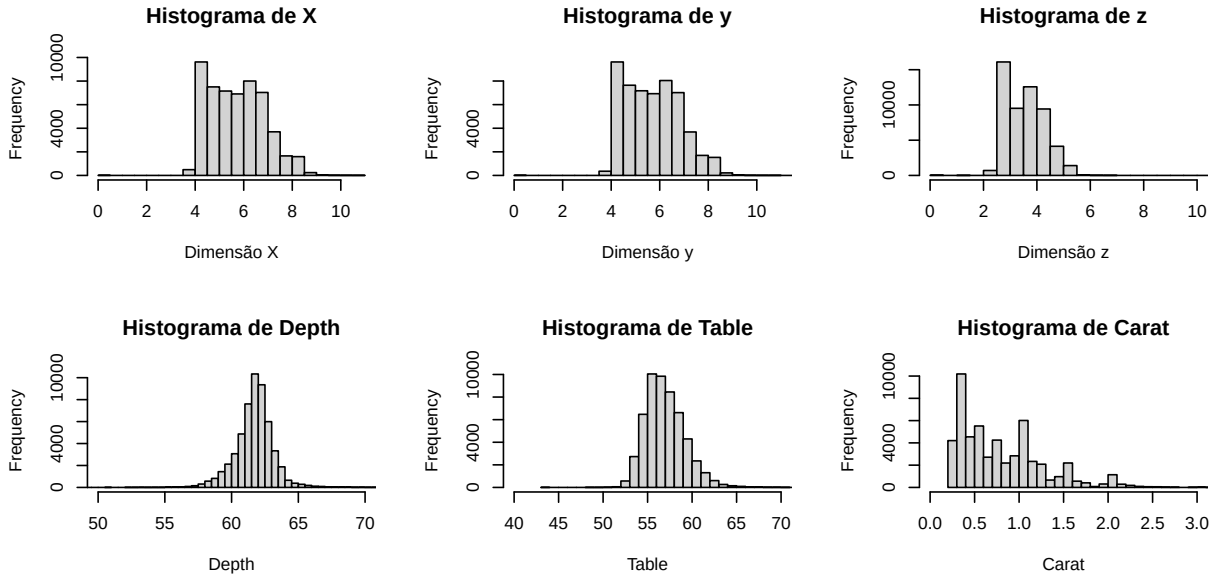

Figure 2: Histogramas das caraterísticas contínuas dos dados.

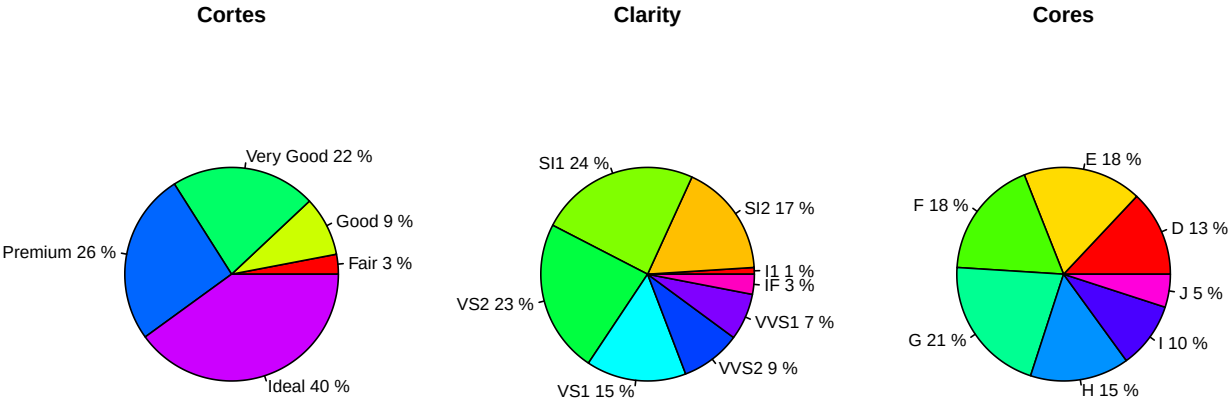

Figure 3: Gráficos de Pizza com as variáveis qualitativas nominais

Análise de Correlação