Análise de Precificação de Diamantes

Aluno: João Gabriel Caetano

1. Resumo

Este projeto teve como objetivo analisar os fatores que determinam o preço dos diamantes, utilizando o clássico dataset "Diamonds" (com ~54.000 registros). A metodologia consistiu em um processo de limpeza para tratar dados fisicamente impossíveis (dimensões zeradas) e uma etapa crucial de **Engenharia de Features**, onde variáveis categóricas (`cut`, `color`, `clarity`) foram quantificadas. Mais importante, foi criada a métrica `price_per_carat` (preço por quilate) para isolar o "valor relativo" da "qualidade" do impacto do "tamanho". O resultado principal é a confirmação de que, embora o `carat` (quilate) seja o fator dominante no preço, as qualidades de corte, cor e pureza têm um impacto claro e quantificável no valor relativo do diamante, um insight que fica oculto em análises superficiais.

2. Introdução

2.1. Contexto

A precificação de diamantes é um processo complexo, mundialmente famoso pelo modelo dos "4 Cs": *Carat* (Quilate), *Cut* (Corte), *Color* (Cor) e *Clarity* (Pureza). Enquanto o quilate é uma medida de peso, os outros três são métricas de qualidade que afetam o brilho e a raridade da pedra. Este dataset público é um padrão da indústria para treinar modelos de análise e *machine learning*, pois permite desvendar a relação matemática entre essas características e o preço final de mercado.

2.2. Objetivos do Projeto

- Realizar a limpeza e o pré-processamento do dataset "Diamonds" para corrigir dados impossíveis e garantir a qualidade da análise.
- Quantificar as variáveis categóricas ordinais (`cut`, `color`, `clarity`) para uso em análises de correlação.
- Criar novas *features* (como `volume` e `price_per_carat`) para permitir uma análise mais profunda.
- Analisar e visualizar o impacto de cada um dos "4 Cs" no preço final.
- Responder à pergunta: "As dimensões (`x`, `y`, `z`) influenciam no preço?"

3. Metodologia e Ferramentas

3.1. Fonte de Dados

O estudo utilizou o dataset "Diamonds", originalmente da biblioteca Seaborn e salvo localmente como `diamonds.csv`. O dataset bruto continha **53.940 registros** e 11 colunas (incluindo uma coluna de índice `Unnamed: 0` que foi removida).

3.2. Ferramentas Utilizadas

- **Linguagem de Programação:** Python
- **Bibliotecas de Análise:** Pandas (manipulação), Matplotlib e Seaborn (visualização).
- **Bibliotecas de Relatório:** Rich (para formatação da saída do script).
- **Ambiente de Desenvolvimento:** Visual Studio Code.

3.3. Processo de Limpeza e Tratamento dos Dados

A análise exploratória inicial (Etapa 2 do script) revelou problemas de integridade que impediam uma análise precisa. As seguintes ações foram executadas:

1. **Tratamento de Dados Impossíveis:** Foram identificados **20 registros** onde as dimensões `x`, `y` ou `z` eram iguais a 0, o que é fisicamente impossível. Esses registros foram removidos.
2. **Codificação de Variáveis Ordinais:** As colunas `cut`, `color`, e `clarity` foram mapeadas para valores numéricos que respeitam sua ordem de qualidade (ex: `cut` foi de 'Fair'=1 a 'Ideal'=5).
3. **Engenharia de Features:** Para aprofundar a análise, duas novas colunas foram criadas:
    a. `volume`: Calculado como `x * y * z`, para correlação direta com o preço.
    b. `price_per_carat`: Calculado como `price / carat`. Esta foi a métrica mais importante, pois nos permite comparar o "valor da qualidade" de um diamante de 1 quilate com um de 0.5 quilate, neutralizando o efeito do tamanho.

O dataset processado e limpo foi salvo em um novo arquivo (`diamonds_dados_limpos.csv`) para garantir a reprodutibilidade da análise.

4. Análise e Resultados

O script gerou 6 visualizações de dados e 2 análises de correlação. Destacamos os resultados mais relevantes que atendem aos requisitos do trabalho.

4.1. Análise 1: O Fator Dominante (Quilate vs. Preço)

A primeira análise investiga a relação mais óbvia: o tamanho do diamante contra seu preço. Como observado no **Gráfico 1** (ver Apêndice), a relação é positiva e exponencial. O preço não dobra quando o quilate dobra; ele cresce muito mais. Isso confirma que o `carat` é o fator que mais influencia o preço bruto, "mascarando" o efeito das outras qualidades.

4.2. Análise 2 e 3: O Valor Relativo (Qualidade vs. Preço por Quilate)

Para descobrir o impacto real do `cut` e da `color`, usamos nossa métrica `price_per_carat`. Os **Gráficos 2 e 3** (ver Apêndice) mostram como o mercado realmente precifica a "qualidade".

Interpretação: Esses gráficos são a principal descoberta do trabalho.

- No **Gráfico 2**, ao isolar o tamanho, vemos que diamantes com corte "Ideal" têm o preço mediano por quilate mais alto, provando que o mercado paga um prêmio significativo pela qualidade do corte.
- No **Gráfico 3**, vemos uma "escada" de valor clara. À medida que a cor melhora (de 'J', pior, para 'D', melhor), o preço por quilate mediano sobe consistentemente.

4.3. Correlação 1: Matriz de Correlação (Heatmap)

O heatmap (**Gráfico 6**, ver Apêndice) nos permite ver a correlação numérica (Pearson) entre todas as variáveis de uma vez.

Interpretação:

1. Confirmamos que `price` tem correlação altíssima com `carat` (**0.92**) e com nossa *feature* `volume` (**0.90**).
2. Mais importante, nossa métrica `price_per_carat` (que remove o efeito do tamanho) mostra correlações positivas fortes com as métricas de qualidade: **0.43** com `color_encoded`, **0.35** com `clarity_encoded` e **0.19** com `cut_encoded`.

4.4. Correlação 2: As dimensões (x, y, z) influenciam no preço?

Para responder à pergunta do professor, extraímos as correlações diretas do `price` com as dimensões:

Tabela 1: Correlação Direta com o Preço (price) | Variável | Coeficiente de Correlação | | :--- | :---: | | carat | 0.9216 | | volume | 0.9043 | | x | 0.8872 | | z | 0.8682 | | y | 0.8679 |

Interpretação: Sim, as dimensões (x, y, z) influenciam **fortemente** o preço. A tabela mostra que elas têm uma correlação linear quase tão alta quanto o próprio `carat`. Isso ocorre porque as três dimensões, juntas, determinam o `volume` da pedra, que está diretamente relacionado ao seu peso (`carat`).

5. Conclusão

Este projeto conseguiu dissecar os fatores de precificação dos diamantes. A análise confirma que, embora o preço final seja dominado pelo tamanho (quilate), o valor relativo da pedra (preço por quilate) é significativamente impulsionado pelas qualidades de corte, cor e pureza.

O sucesso desta análise não veio apenas da visualização, mas da **Engenharia de Features**. A criação da métrica `price_per_carat` foi a etapa-chave que permitiu "limpar" o ruído causado pelo tamanho e provar, com dados, o valor real dos outros "Cs". Isso demonstra que a preparação e transformação dos dados são fundamentais para ir além de conclusões superficiais e gerar insights estratégicos.

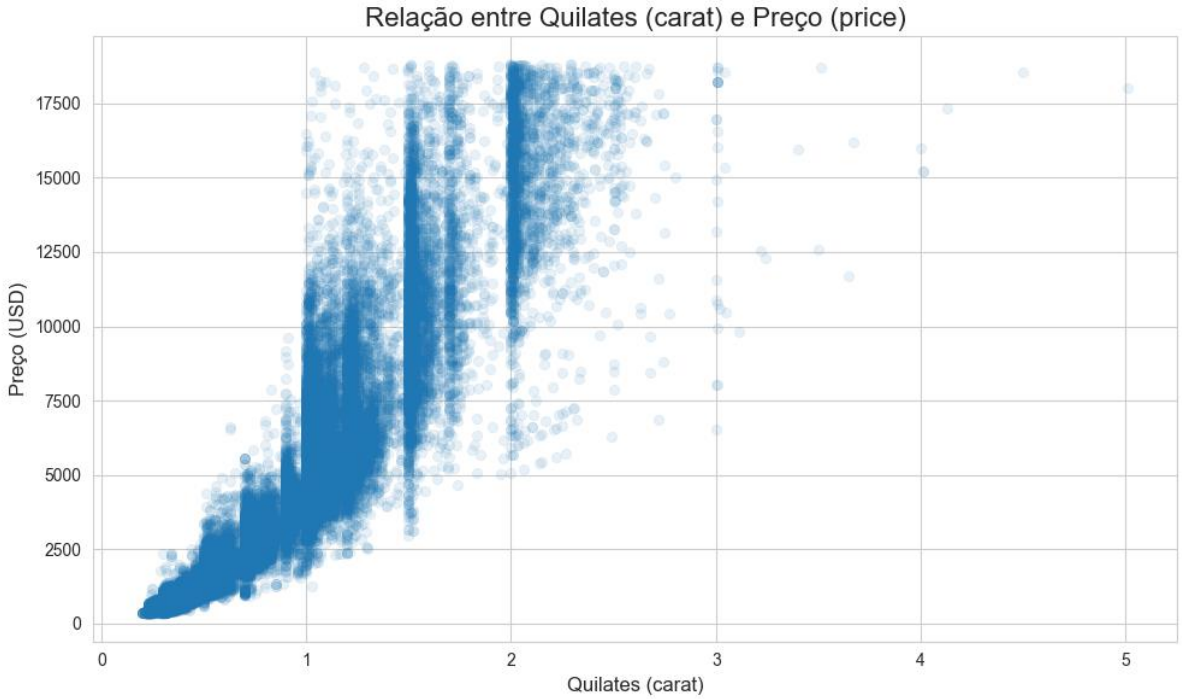
6. Limitações e Próximos Passos

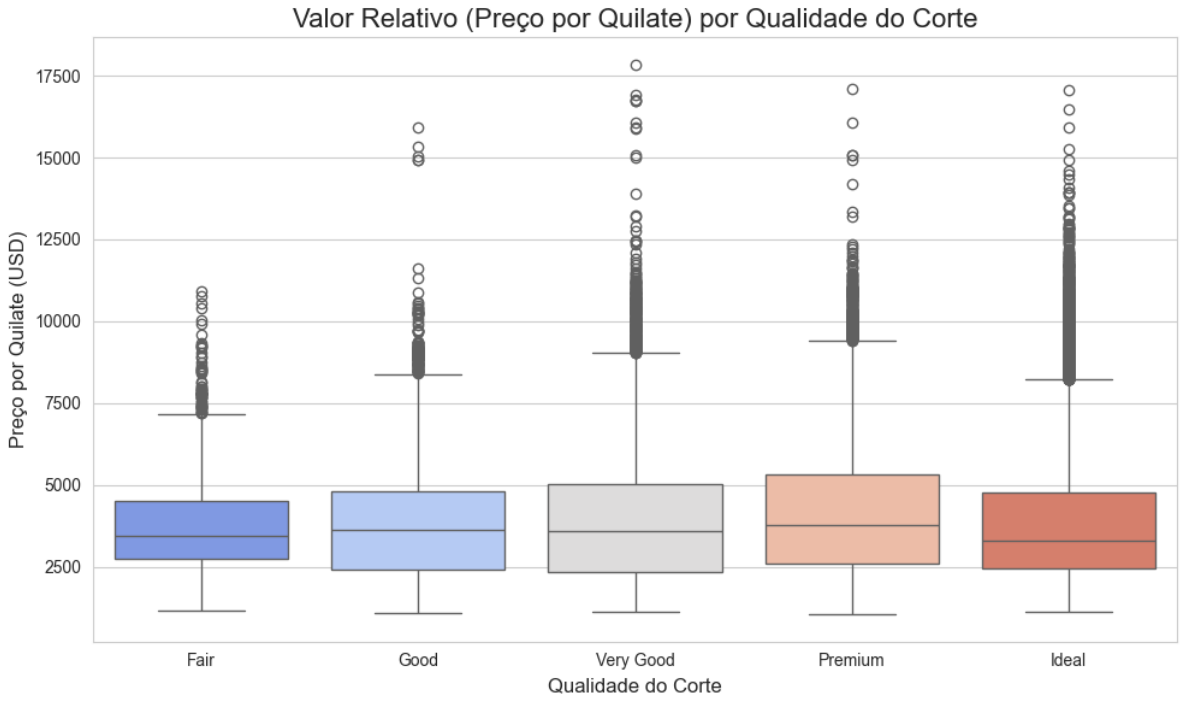
A principal limitação é que este é um dataset de referência, e os preços podem não refletir o mercado atual. Como próximos passos para aprofundar o estudo, sugere-se:

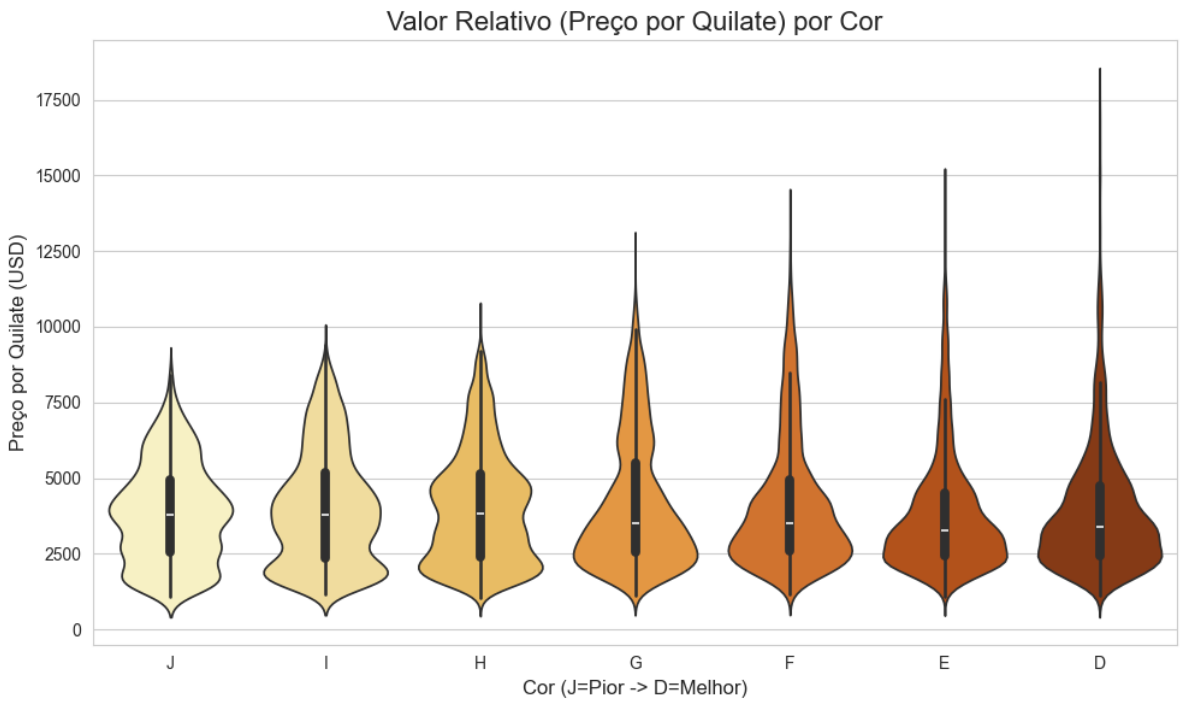
- **Modelagem Preditiva:** Utilizar este dataset limpo para treinar um modelo de *Machine Learning* (como Regressão Linear Múltipla ou Random Forest) para prever o preço de um diamante com base em suas características.
- **Análise de Proporções:** Investigar mais a fundo as colunas `depth` e `table` (como feito no **Gráfico 4**) para ver como proporções "ideais" vs. "não ideais" afetam o `price_per_carat`.
- **Análise Multivariada:** Explorar o **Gráfico 5** para entender como a combinação de múltiplas qualidades **(**ex: um diamante "Ideal" e "D" vs. "Ideal" e "J") impacta o preço.

Gráficos da Análise

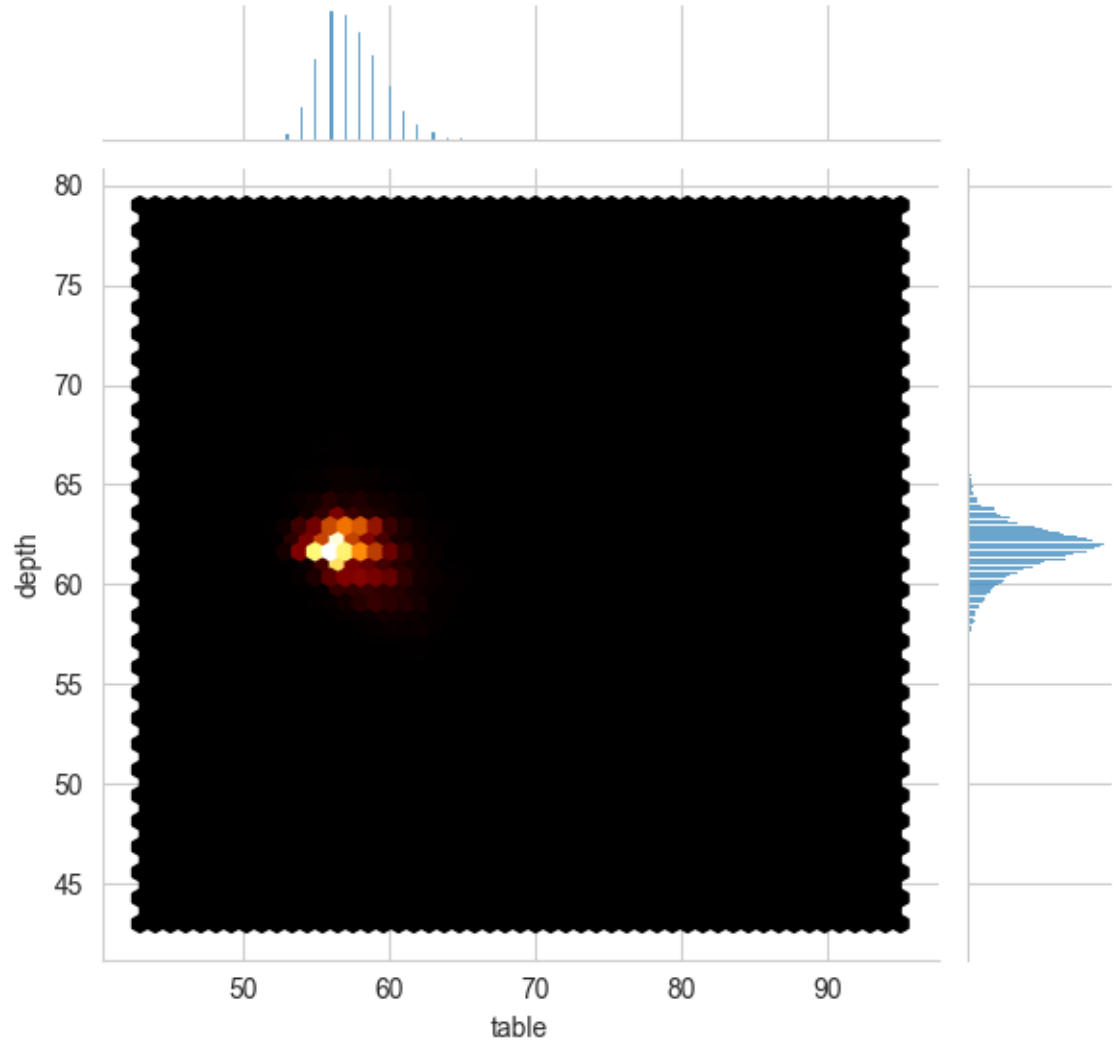
Aqui estão as visualizações de dados geradas pelo script.






Valor Relativo (Preço por Quilate) por Cor
Preço por Quilate (USD)
17500
15000
12500
10000
7500
5000
2500
0
J
I
H
G
F
E
D
Cor (J=Pior -> D=Melhor)

Relação de Densidade entre "Depth" e "Table"



Relação Preço x Quilate separada por Pureza (Clarity)