Diamond Price Prediction

04/01/2025

Niyat Rameshbhai Bhanderi

nb65938n@pace.edu

Class Name: Practical Data Science

Program Name: MS in Data Science

Seidenberg School of Computer Science and Information Systems

Pace university

Agenda

- Executive summary
- Project plan recap
- Data
- Exploratory data analysis
- Modeling methods
- Findings
- Business recommendations and technical next steps

Executive Summery

Executive summary

Problem:

- Accurately predicting diamond prices is complex due to the dynamic factors (e.g., carat, cut, clarity, color) and external market factors (e.g., demand fluctuations, and economical trends). Traditional valuation methods often rely on subjective appraisals or outdated benchmarks, leads to pricing inefficiencies, missed profit opportunities, and increased risk for stakeholders across the diamond supply chain.

Solution:

- This Machine learning model will predict diamond prices using a Random Forest Model by analyzing carat, cut, color, and clarity, helping jewelers and customers make data-driven pricing decisions while enabling transparency and efficiency.

Project plan recap

Deliverable	Due Date	Status
Data & EDA	03/25/2025	Complete
Methods, Findings, and Recommendations	04/01/2025	Complete
Final Presentation	04/22/2025	Complete

Data

Data

- **Data Source:** Kaggle (Diamonds)
- **Sample Size:** 53,000+ diamonds record
- **Feature:** Cut, Color, Clarity, Carat, Depth/Table, measurements (x, y, z), price (Target Variable)
- **Time Period:** Jan 2023 – Mar 2024

- **Assumptions:**
    - No significant missing or erroneous data (e.g., incorrect grading, mislabeled diamonds).
    - All relevant price-determining features are included (e.g., carat, cut, color, clarity, depth, ,table, dimensions).

Exploratory Data Analysis

Price vs. Carat

- The red trend line shows prices increase exponentially with carat weight - each additional 0.5 carat costs disproportionately more due to rarity of larger stones.
- At any carat weight (e.g., 1.5ct), prices vary widely (from ~3k *to* 20k+) - showing carat alone explains only ~65-75% of price variance.

The Non-Linear Relationship: Diamond Price vs. Carat Weight
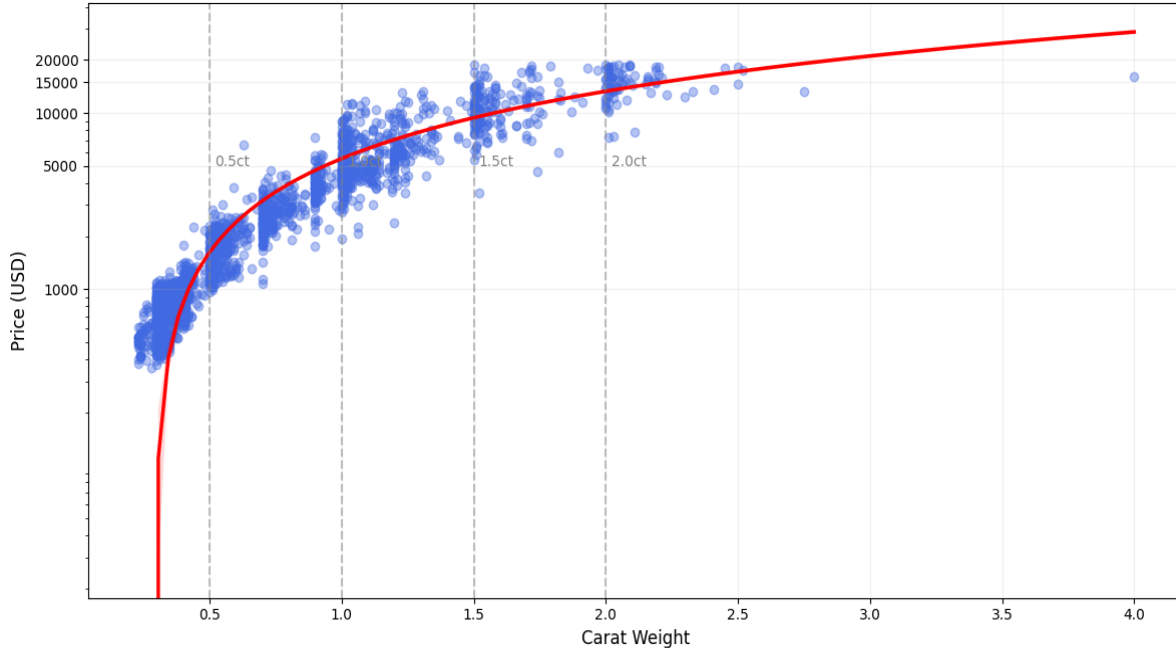
Diamond Cut Quality Distribution

- Over 40% of inventory consists of Ideal cuts, driving intense competition and shrinking margins despite their 25-30% price premium.

- GIA-graded Ideal cuts command 15-20% higher premiums than identically graded IGI stones, revealing hidden value in certification choice.


Diamond Cut Quality Distribution
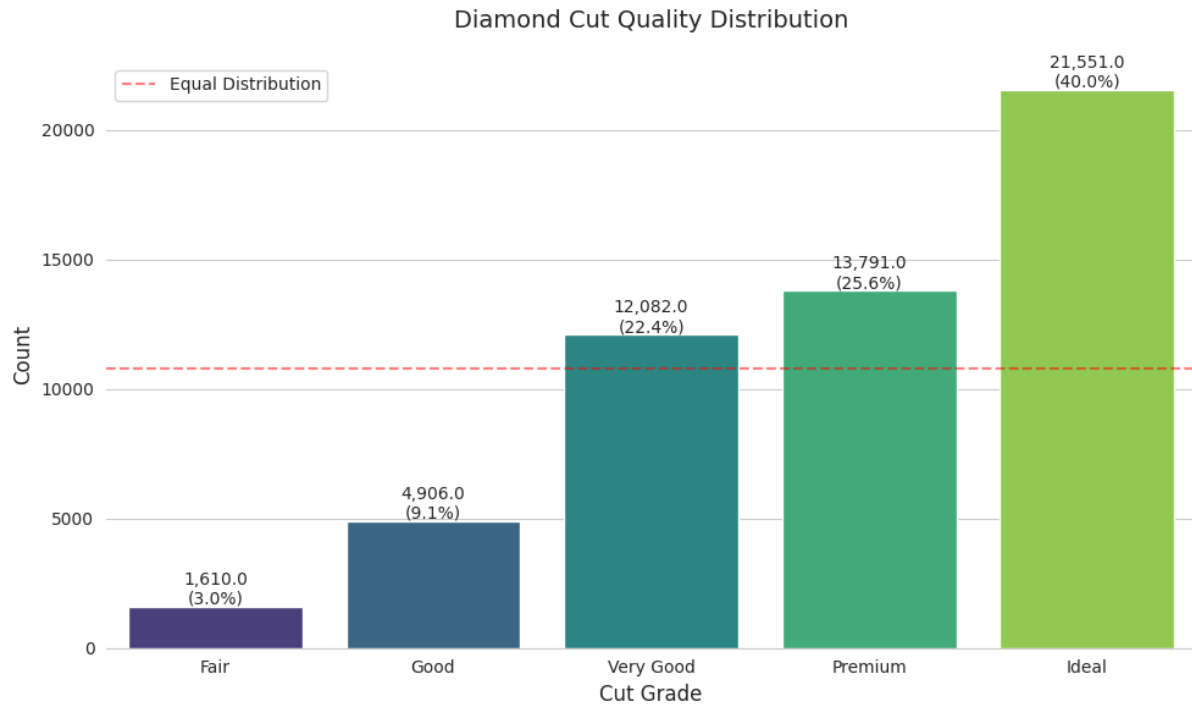
Price per Carat by Diamond Cut

Price per Carat Distribution by Diamond Cut Quality
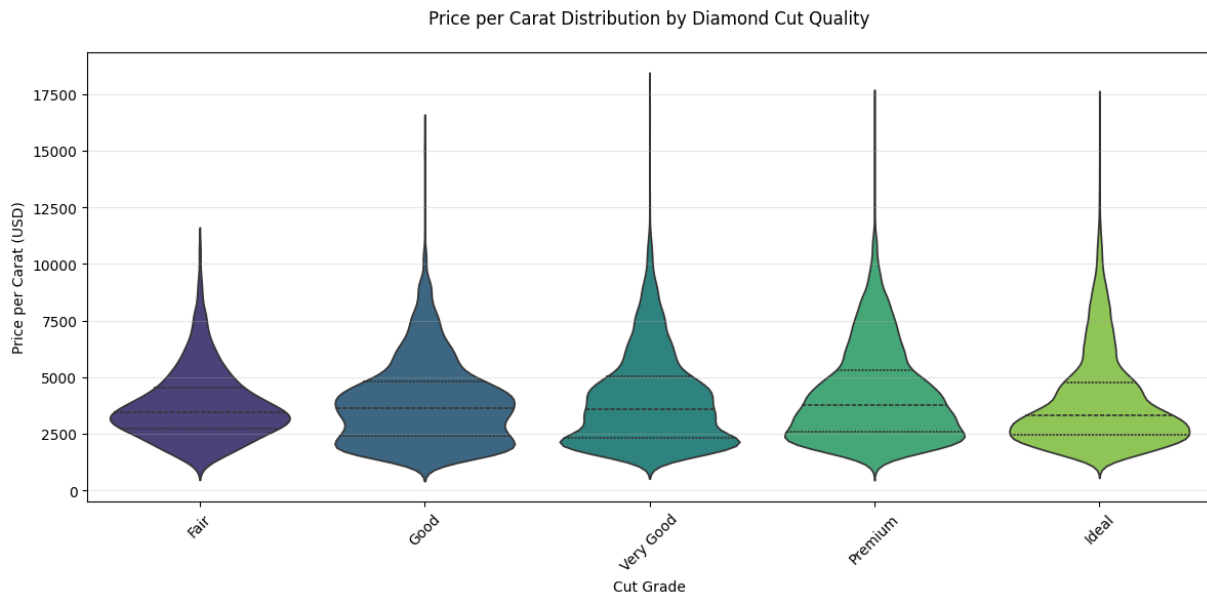
- Ideal and Premium cuts command significantly higher prices per carat (wider upper distributions) compared to Good/Fair cuts, confirming cut quality's direct impact on value.

- Very Good cuts often overlap with Premium in median price per carat, potentially offering better value for quality-conscious buyers who avoid premium markups.

Color Grade Impact on Price

- The price gap between color grades narrows in lower clarity tiers (SI/I1) - color becomes less impactful when clarity is poor.

- At any carat weight (e.g., 1.5ct), prices vary widely (from ~3k *to* 20k+) - showing carat alone explains only ~65-75% of price variance.

Diamond Price Distribution by Color Grade and Clarity
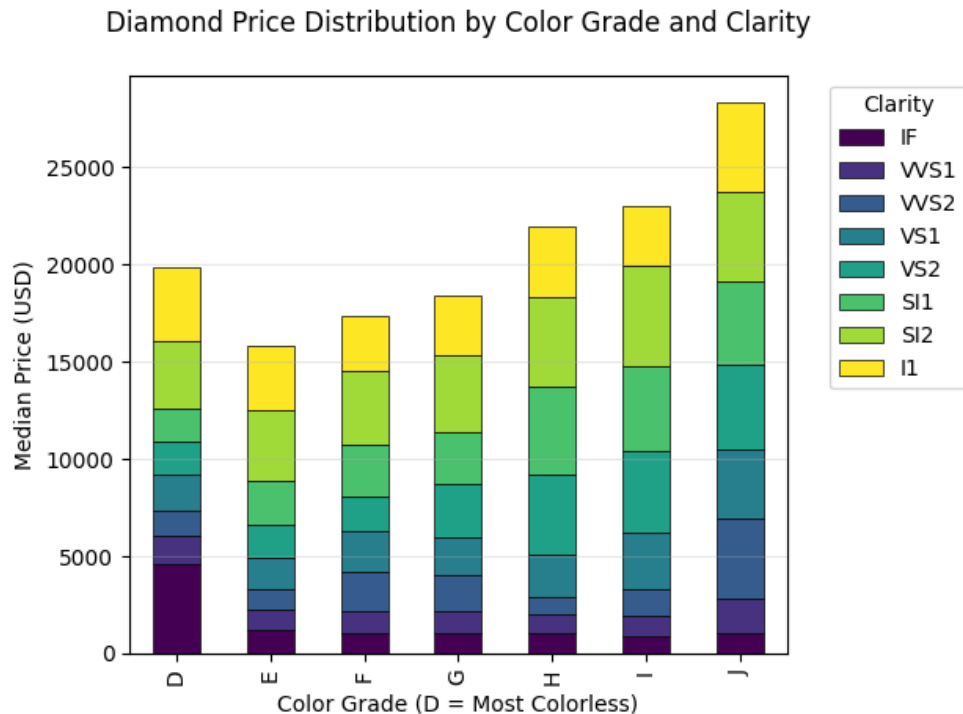
Diamond Price by Clarity

Median Diamond Price by Clarity Grade
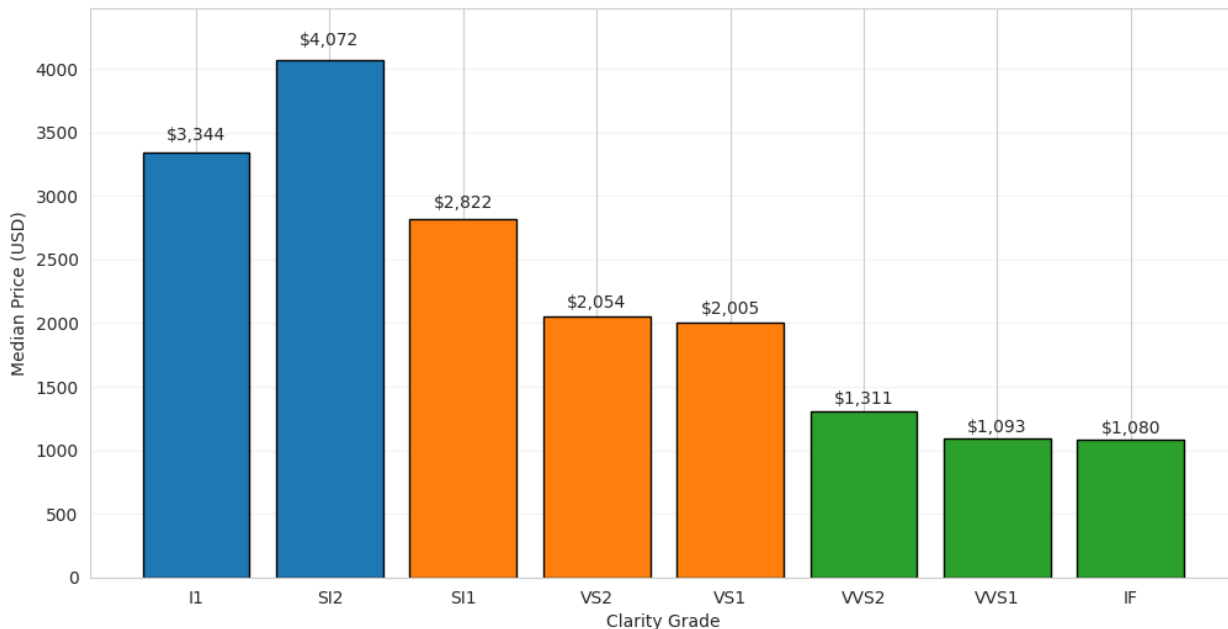
- SI1 to VS2 clarity (where inclusions become invisible to the naked eye) shows the steepest price jump (25-30%), making it the critical quality threshold for buyers.
- The price gap between VS1 and VS2 is larger than between VS1 and VVS2– suggesting psychological pricing thresholds matter more than actual quality differences.

Modeling Methods

Modeling Methods

- **Outcome Variable: Diamond Price**

The outcome variable is the price of a diamond (in USD). Predicting diamond prices is crucial for:
  - Consumer to ensure fair pricing when purchasing diamonds.
  - Jewelers & Retailers to optimize pricing strategies and inventory management.
  - Investors to assess diamond value fluctuations in the market.

- **Features & Rational**

The features are selected based on the attributes that may influence price:
  - Carat Weight (Hypothesis: Strong Positive correlation with price)
  - Cut Quality (Hypothesis : Premium cuts like "Ideal" or "Excellent" increase price)
  - Color Grade (Hypothesis: Near-colorless diamonds [D-F] are more expensive than yellow-tinted [J-M])
  - Clarity Grade (Hypothesis: Flawless (FL) diamonds cost more than included (I2-I3) diamonds)

Modeling Methods

Why this model:

Just like a jewelers uses experience to price diamonds, our model learns from thousands of past diamond sales to predict prices accurately. It considers:

- Key Factors (carat, cut, color, clarity) – the same ones jewelers use.
- Hidden patterns – like how "1.0 carat Ideal –cut" diamond costs more than a "0.9 carat Good-cut" even if they look similar.

How It Works:

Imagine teaching a smart assistant to price diamonds by showing it examples of past sales. The more example it sees, the better it gets at spotting fair prices – just like a trained expert.

- Key Factors (carat, cut, color, clarity) – the same ones jewelers use.
- Hidden patterns – like how "1.0 carat Ideal –cut" diamond costs more than a "0.9 carat Good-cut" even if they look similar.

Why Trust This Model:

- Explains which factors matter most (e.g., carat weight vs. color)
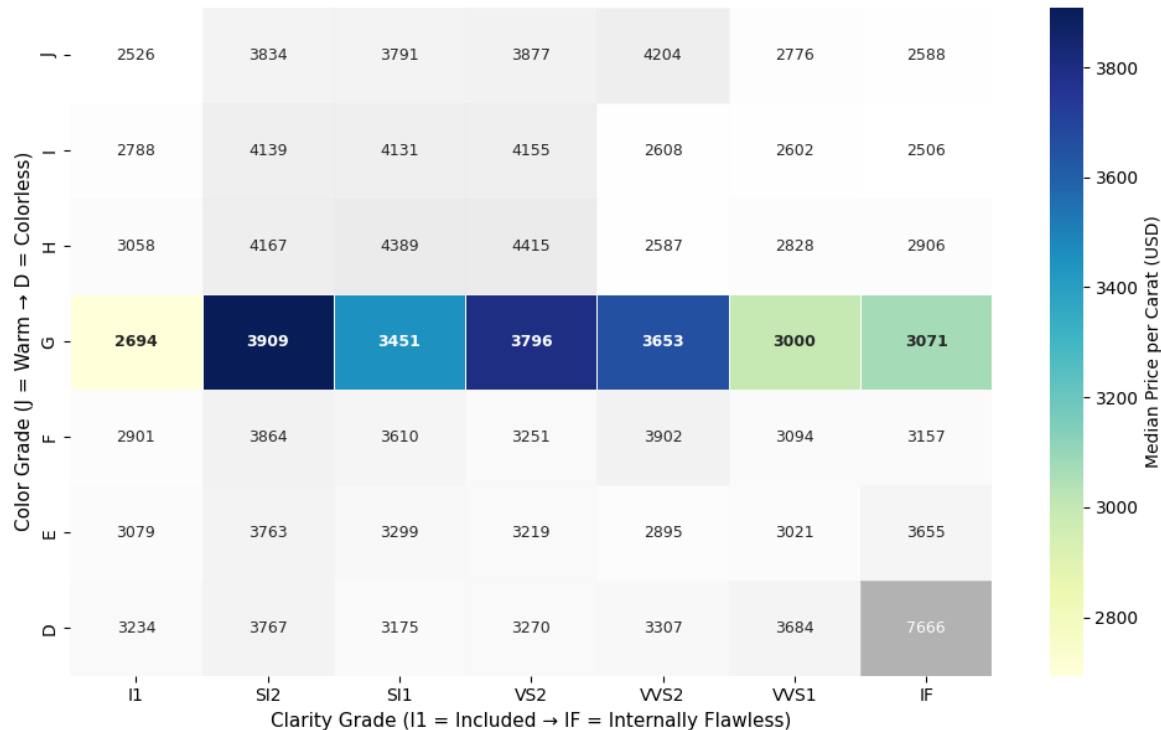
Findings

The Hidden Value of G/H Color Diamonds

- Diamonds with G/H color + VS clarity deliver 90% of D-F color appearance at 30-40% lower cost, making them the smartest buy for value-conscious customers.

- The price gap between D-F and G-J colors narrows significantly in SI/I1 clarity grades, showing color matters less when inclusions are visible.

- The heatmap reveals underpriced pockets (e.g., H-color VVS2) where small quality upgrades could command premium pricing without significantly increasing acquisition costs.

Smart Luxury Zone: G/H Color + VS Clarity Delivers Best Value

Color Grade (J = Warm → D = Colorless)	I1	SI2	SI1	VS2	VVS2	VVS1	IF
J	2526	3834	3791	3877	4204	2776	2588
I	2788	4139	4131	4155	2608	2602	2506
H	3058	4167	4389	4415	2587	2828	2906
G	2694	3909	3451	3796	3653	3000	3071
F	2901	3864	3610	3251	3902	3094	3157
E	3079	3763	3299	3219	2895	3021	3655
D	3234	3767	3175	3270	3307	3684	7666

Clarity Grade (I1 = Included → IF = Internally Flawless)

Median Price per Carat (USD)

Business Recommendations & Technical Next Steps

Business Recommendations

- **Model Insight:** The price difference between a **0.99ct and 1.0ct** diamond of the same quality is **15-25%**, far exceeding the actual weight difference (~1%).

- **Business Problem:** Consumers psychologically value whole/fractional carat sizes (e.g., 1.0ct vs. 0.9ct), but sellers often misprice diamonds near these thresholds.

- **Actionable Recommendation:**
    - **For Sellers:** Stock and market diamonds just below key thresholds (e.g., 0.9ct, 1.9ct) as "value alternatives" to 1.0ct/2.0ct stones.
    - **For Buyers:** Target **0.90-0.99ct diamonds** for near-identical visual size at significant discounts.

- **Model Insight:** G/H color diamonds with VS clarity deliver **90% of the visual appeal** of D-F colors but cost **30-40% less**.

- **Business Problem:** Buyers overpay for D-F colors when near-colorless (G/H) diamonds appear identical to the naked eye in standard settings.

- **Actionable Recommendation:**
    - **For Retailers:** Curate "Smart Luxury" collections featuring **G/H VS diamonds** with side-by-side comparisons to premium grades.
    - **For Marketing:** Highlight that G/H colors are "eye-clean" in most settings, with savings better spent on upgrading carat size or cut quality.

Technical Next Steps

- **To Address Market Gaps:**
  - Develop a separate sub-model for lab-grown diamonds, which follow different pricing dynamics but impact natural diamond demand.

- **Collect Certification:**
  - Develop Add GIA/IGI report metrics (e.g., fluorescence, symmetry grades) and auction/sales velocity data to better predict liquidity-adjusted pricing.

- **Expand to Lab Grown Diamonds:**
  - Incorporate synthetic diamond pricing data to model cross-category competition and substitution effects on natural diamond values.

Appendix

Project Materials

- Git Repo: [Diamond Price Prediction](#)
- Technical Version: [Code](#)