Diamond Price Factors Research Paper

By Lucas Su, Dheeraj Garg, William Lu, Eric Yan

Overview:
Across the world, diamonds are renowned for being the "ultimate gemstone." In the context of jewelry, they symbolize commitment and love. Because of their massive value, diamonds are a major financial investment for the average consumer. Their price is measured by four categories: carat, cut, color, and clarity. In order to help people maximize their budget, our team wanted to answer the question - "Which of the 4 C's (carat, cut, color, and clarity) affects diamond retail prices the most?" During analysis, our research question stayed consistent because of the data's simplicity and readability.

Before doing any research, our group hypothesized that carat would have the largest influence on price because larger diamonds are harder to find than smaller diamonds with better cut, color, or clarity. The other three C's seem much more difficult to predict because they don't seem to make a huge difference on how a diamond looks. Because of this, we theorized that cut, color, and clarity all made a similar impact on diamond price.

When researching this question, we found that most diamond websites don't provide a straightforward answer. For example, Gia has an article on which of the 4 C's is the most important. The issue is with the conclusion since one of the 4 C's is listed as the most valuable feature. Instead, the site avoids the question by mentioning how keeping a diamond clean is important (Gia). This research revealed that our project was necessary due to the lack of clear information online.

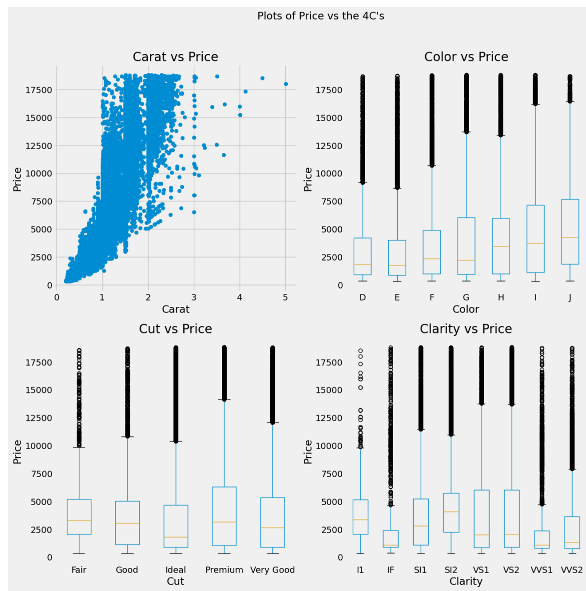
Data Description:
To answer our question, we used the Gemstone Price Dataset from Kaggle which originated from ggplot2. The dataset contains 53940 rows and 11 columns: Unnamed 0, carat, cut, color, clarity, depth, table, price, x, y, and z. First, we removed the Unnamed 0 column because it was the same as the indices. Next we cleaned the data. Any rows with dimension values of 0 or duplicates were removed for being irrelevant, leaving us with 53755 rows. Finally, all columns except for carat, cut, color, clarity, and price were removed because we only needed these columns to answer our research question.

	Unnamed: 0	carat	cut	color	clarity	depth	table	price	x	y	z
0	1	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	2	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	3	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	4	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	5	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

Each of the 4 C's have a specific value. The carat of a diamond is its weight, 0.2 grams per carat. The cut is a measure of how a diamond's shape interacts with light. It uses a scale ranging from Fair, Good, Very Good, Premium, and Ideal. The color of a diamond simply shows the color from letter D (clear) to J (faint tint). Lastly, the clarity shows how transparent and clean a diamond is. The categories are Internally Flawless (IF), Very, Very Slightly Included (VVS1 / VVS2), Very Slightly Included (VS1 / VS2), Slightly Included (SI1 / SI2), Included (I1 / I2 / I3). The last column is price, which just shows the retail price the diamond was sold in USD. This is our target variable. We originally used a scale of even intervals to make these categories numerical. However, after some research, we realized the intervals within these categories were not identical. Hence, we scaled the intervals to approximate the differences in categories. For instance, in the context of clarity, the jump from I3 to SI1 would be larger than from VS1 to VVS1. Using these intervals, we added columns which divided the categories into a 1-8 scale which ensured the multi-linear regression ran smoothly.

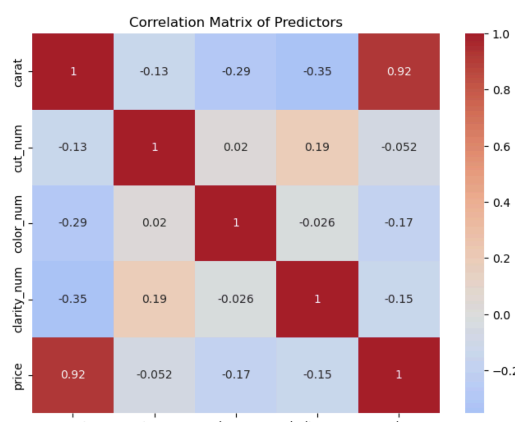
	carat	cut	color	clarity	price	cut_num	clarity_num	color_num
0	0.23	Ideal	E	SI2	326	5.00	4.49	6.47
1	0.21	Premium	E	SI1	326	4.48	5.71	6.47
2	0.23	Good	E	VS1	327	3.69	6.92	6.47
3	0.29	Premium	I	VS2	334	4.48	6.52	3.31
4	0.31	Good	J	SI2	335	3.69	4.49	1.00

Data Exploration:



The dataset ended with 53755 rows and 8 columns. In order to get a better understanding of our cleaned dataset, we explored it with some visualizations.. First, we graphed the scatter plot of carat vs price. We noticed that almost all of the diamonds had a carat less than 3, probably due to higher carat diamonds being rarer and more difficult to find. For the color, cut, and clarity graphs, we used a box plot to prevent the categorical data from clumping around a few numbers on the x-axis. Interestingly, when the color quality increased, the price actually decreased. We theorized that this happened because there weren't many diamonds with high color, causing the median to be lower. Meanwhile, the quality of the cut and clarity didn't seem to affect the price.

Next we made a correlation matrix, which shows how variables relate to each other. Carat and price had a value of 0.92, which symbolized a very high correlation. This relationship strengthened our belief that carat was the most important factor in price.



Model:

We assessed three different models to fit the data. Our first model was a multilinear regression model using the 4 C's as features. We performed linear regressions between price, carat, cut, color, and clarity, as we believed carat was a confounding variable which needs to be controlled for. After observing that the graph had an upwards curve, we used a 2-degree polynomial linear regression model since we believed it would be a better fit. A linear regression model plots the points and tries to create the best line that minimizes the sum of the squared residuals. Using data split into an 80% train set and 20% test set, we fit the linear regression model on it and outputted our equation:

Price = -5707.877 + 3773.190($carat^2$)+ 145.228(cut_num) + 358.693(color_num) + 613.223(clarity_num)

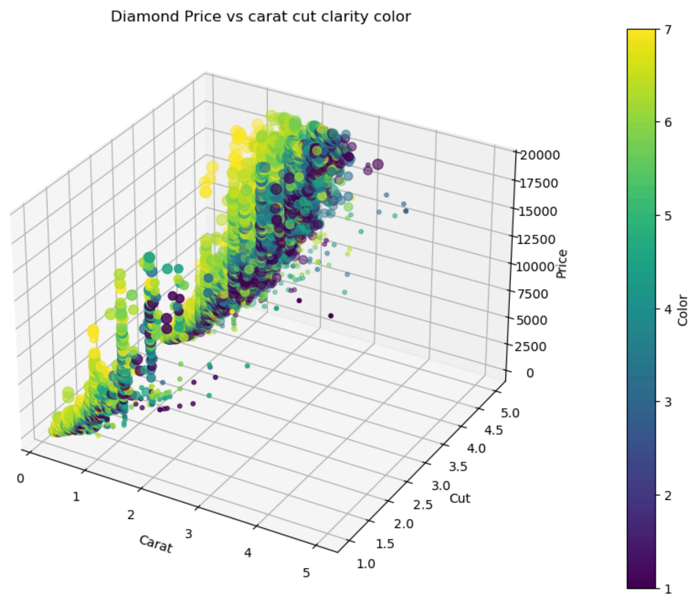
Test MSE: 2713632.035

Test R^2 Score: 0.828286

Train MSE: 2533241.078

Train R^2 Score: 0.841

The MSE (Mean Squared Error) shows the difference between the predicted values and actual values. The higher the MSE, the worse the accuracy is. R^2 tells you the amount of variability explained by the input features, giving a result from 0-1, similar to accuracy.



Therefore, we chose a random forest model from the scikit-learn random forest regressor library. The random forest model is powerful because it is able to capture non-linearities better than a standard linear regression model. For example, increasing the carat from 0.5 to 1.0 will not affect price the same way as an increase from 2.5 to 3.0, but on the linear regression model, these are treated as the same interval. It also requires little tuning and resists overfitting because each tree sees a slightly different dataset; they make different errors and averaging their results reduces variance. The random forest algorithm can also learn combinations of features. For example, large carat diamonds with high clarity get exponentially more expensive. This model works by bagging (bootstrap aggregating) and training many different trees on random subsamples of the training data (drawn with replacement). This forces each tree to be diverse. Then, the model averages each tree's predictions. Using 500 estimators, and a max depth of 10, we arrived at these metrics:

- Train MSE = 258554.893
- Train R^2 = 0.984
- Test MSE = 302140.546
- Test R^2 = 0.98

These results were much better than the linear regression model as the R^2 score was larger.

Finally, we implemented gradient boosting as a more advanced method of predicting diamond price. Gradient boosting can also predict nonlinearities and interactions similar to a random forest, but it improves sequentially by correcting previous mistakes. Since diamond prices depend on four features: carat, cut, clarity, and color, and have complex interactions with each other, gradient boosting may produce more accurate predictions than random forest. A gradient boosting model works by combining multiple small decision trees. Each tree by itself is not able to predict the price, but the subsequent tree corrects the mistakes made by the previous one. In this case, the model starts with an initial prediction for the price, then measures the error between the actual and predicted prices of diamonds in the training set. It trains a small decision tree to predict these errors and adds it to the model. Repeating this many times allows the model to become more and more accurate. The hyperparameters used were 500 iterations of the model, maximum depth of 10 and a learning rate of 0.01. This resulted in these metrics:

- Train MSE = 186602.571
- Train R^2 = 0.988
- Test MSE = 2713632.035
- Test R^2 = 0.98

These metrics turned out to be very similar to the random forest, and had no significant differences.

Results:

Based on the data, carat is by far the most important decider of price, followed by clarity, color, and cut. These numbers were decided by their coefficients: 8885.636, 787.577, 358.802, and 250.921 from the linear regression. However, this model is not the one that we decided to use at the end. We decided to use the gradient boosting model at the end, as it produced the highest R^2 score and was the best representation of the data.

Earlier, our group hypothesized that carat was paramount when determining the price. This hypothesis ended up being proven correct by our model. Meanwhile, we also predicted that the other three C's would have equal importance. This conclusion was shown to be incorrect when looking at the coefficients because all of the coefficients have a large interval between them.

Overall, we believe this study will help consumers save money when purchasing diamonds by stopping them from overpaying for invisible factors. For example, one could decide which of the 4 C's matters most to them and compare that to how expensive each of them are. There could be some improvements made to the project, such as examining other factors in price such as location or shape. Throughout the project, we didn't have many limitations. The quantity and quality of the data were very high, so we only had to add the scaled categorical values for cut, color, and clarity.

Contributions:

Dheeraj Garg - Compiling final notebook, 3d regression for color, carat, and price, correlation matrix, data cleaning, cleaning up graphs for presentation

William Lu - Most of the report and project, data exploration, carat & cut vs price graph.

Lucas Su - Multi Class Linear Regression for all features, individual feature vs price graphs, random forest model, gradient boosting model, model decisions

Eric Yan - Polynomial linear regression for individual C's, multi-linear for: clarity & carat, cut & carat vs price, ordinal scaling for categorical data

Work Cited

“The 4Cs of Diamond Quality | The Most Important C Explained.” *GIA 4Cs*, 19 April 2016, https://4cs.gia.edu/en-us/blog/4cs-diamond-quality-most-important-c/. Accessed 21 July 2026.https://www.kaggle.com/data

Reddy, S. “Gemstone_price_dataset.” Kaggle, 2024, https://www.kaggle.com/datasets/nithiin07/gemstone-price-dataset.