Diamond Attributes and Price

Unit 3 Capstone
Brandon Steed

Research Focus

- Are the features that are most reflected in pricing the same as those that are most important in creating brilliance and sparkle?
- Can a given price for a diamond be modeled and predicted using variables such as carat weight, color, and clarity

What's most important to consider when purchasing a diamond?

- Budget
- How big?
- Sparkle

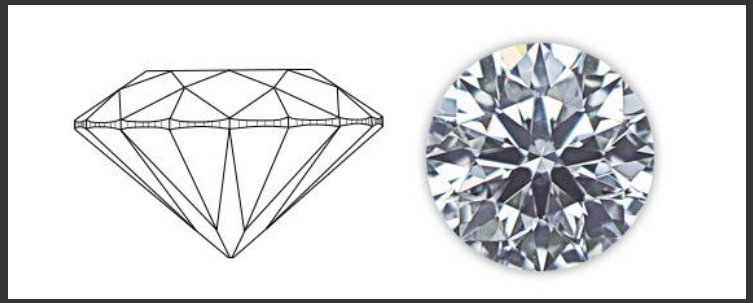
What do retailers try to sell you on?

- Size
- Clarity
- Color

Which would you rather buy?

Look at the price predictions for the two hypothetical diamonds below. Both are one-carat diamonds with no perceivable color.

The first has impeccable clarity, meaning no imperfections within the diamond. However, the dimensions of the diamond allow much of the light entering the diamond to exit through the bottom. The result is lower light return through the facets at the top of the diamond, meaning less "sparkle."



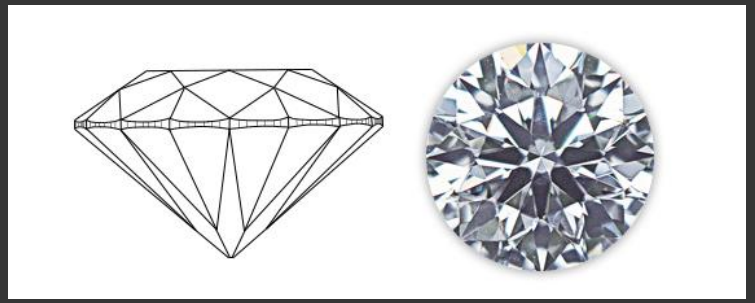
The second has S2 clarity (imperfections would be easily spotted with 10x magnification, but difficult to discern with the naked eye). However, the dimensions of the diamond allow practically all of the light entering it to be returned to the top of the diamond, resulting in impressive "sparkle."

Which would you rather buy?

Look at the price predictions for the two hypothetical diamonds below. Both are one-carat diamonds with no perceivable color.

Predicted Price:

$10,566



Predicted Price:

$5,342

Dataset

➔ Dataset Source

https://www.kaggle.com/shivam2503/diamonds

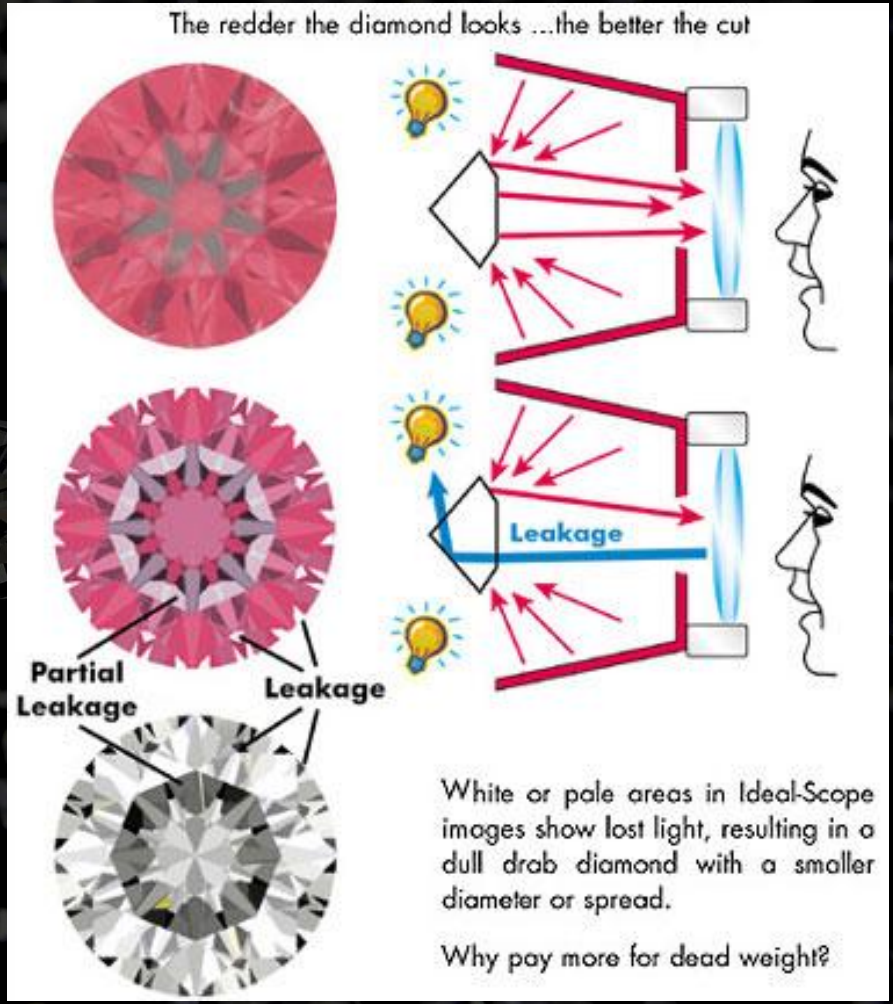
The dataset contains the prices and other attributes of 53,940 diamonds.

➔ Columns

1. price price in US dollars (\$326--\$18,823)
2. carat weight of the diamond (0.2--5.01)
3. cut quality of the cut (Fair, Good, Very Good, Premium, Ideal)
4. color diamond color, from J (worst) to D (best)
5. clarity a measurement of how clear the diamond is (I1 (worst), SI2, SI1,
   VS2, VS1, VVS2, VVS1, IF (best)
6. x length in mm (0--10.74)
7. y width in mm (0--58.9)
8. z depth in mm (0--31.8)
9. depth total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43--79)
10. table width of top of diamond relative to widest point (43--95)

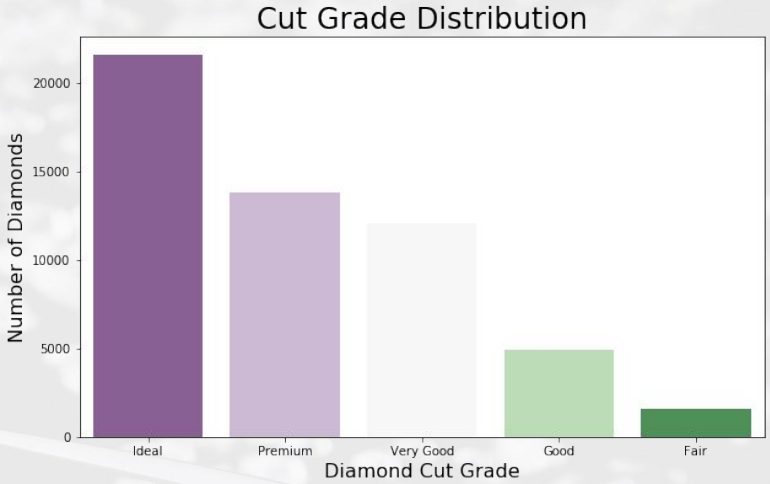
Cut Grade

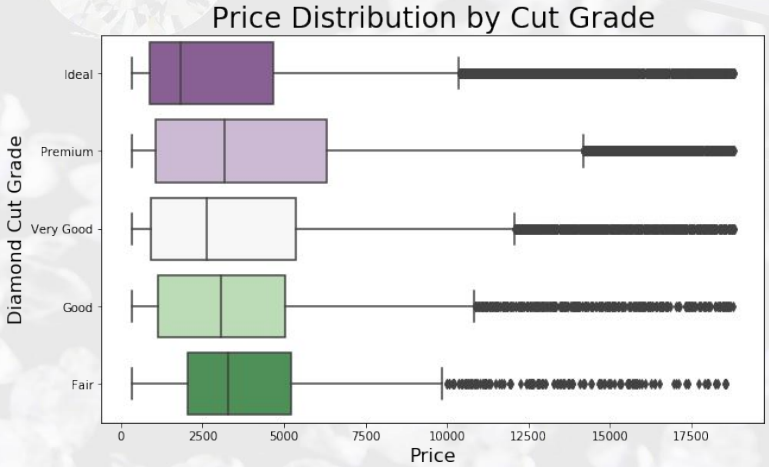
The manner in which a diamond is cut is known to be the most important factor in determining how it sparkles in the light.



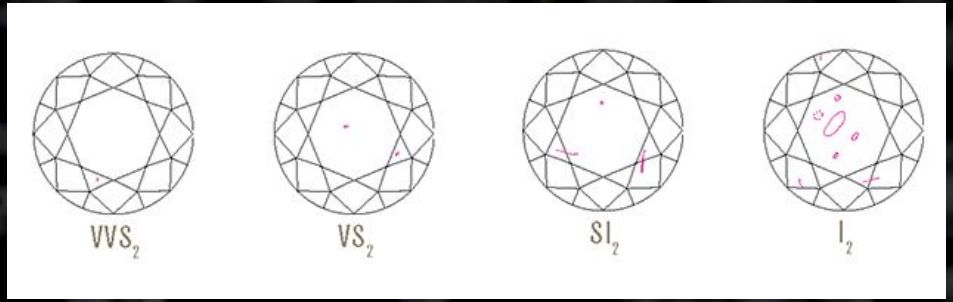
Cut Grade

Cut grades are skewed toward higher grades. Diamonds with lower-quality cuts had higher median prices due to carat weight.





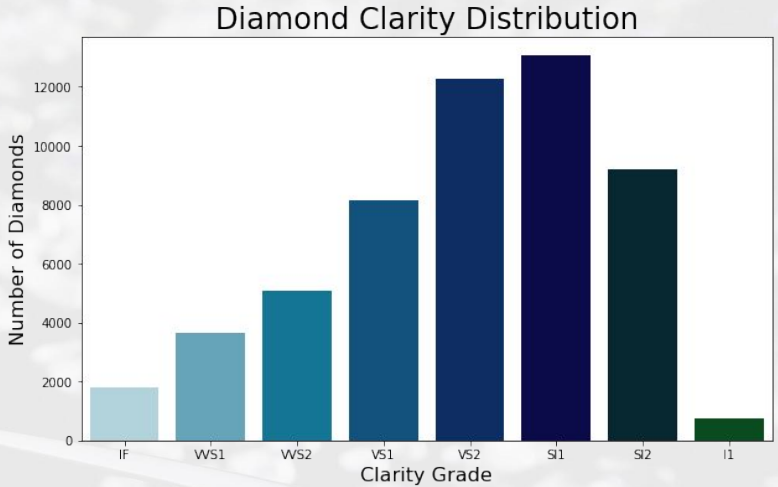
Clarity Grade

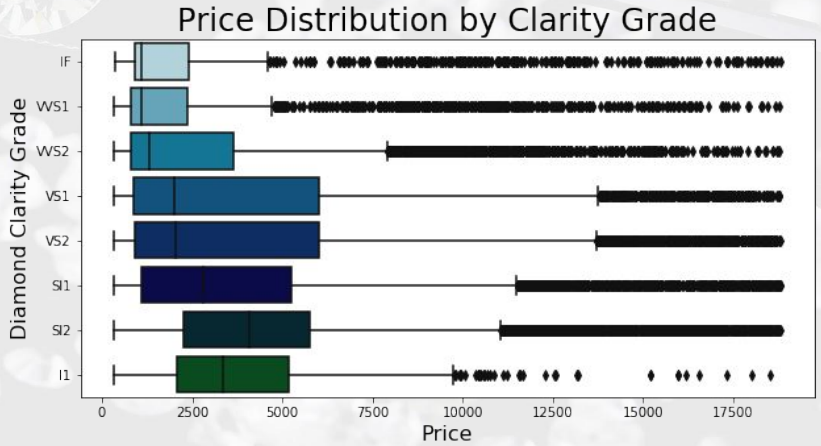


Clarity grades are given to describe the visibleness of flaws found within a diamond. Internally flawless diamonds have no flaws found within them, even with high magnification. Diamonds falling within grades labelled from "very very slightly included" to "slightly included" have flaws that are mainly seen with varying degrees of magnification. Diamonds with an "I1" grade (Included 1) possess flaws easily visible upon inspection without any magnification.

Clarity Grade

Most diamonds is the dataset fell between the grades of VS1 and S2, meaning they had flaws that would be visible with 10x magnification. The flaws in an S2 diamond would be more visible than those in the VS1. Diamonds with lower clarity grades had higher median prices due to carat weight.



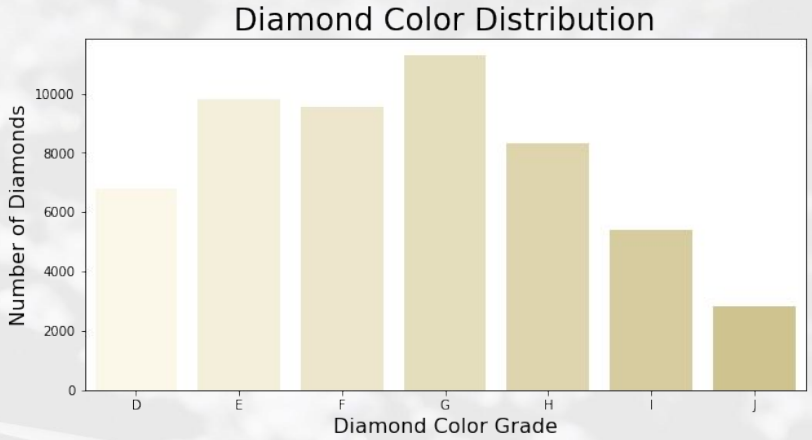


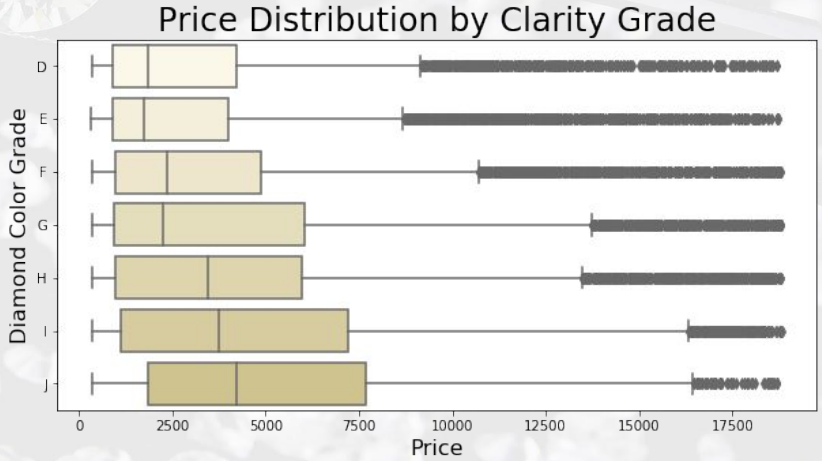
Color Grade

Diamond color is graded on a scale from D to Z. A diamond with a color grade of "D" is completely colorless, while a diamond with a color grade of "Z" is fairly yellow or brown. Diamonds with a different perceivable color or have more color than a "Z" grade are referred to as "fancy-color" and do not fall into this system. Among "white" diamonds, color as close to "D" as possible is viewed as preferable.

Color Grade

Diamonds in this dataset range in color grade from D (completely colorless), to J (last grade in the near-colorless category). Diamonds with lower color grades had higher median prices due to carat weight.





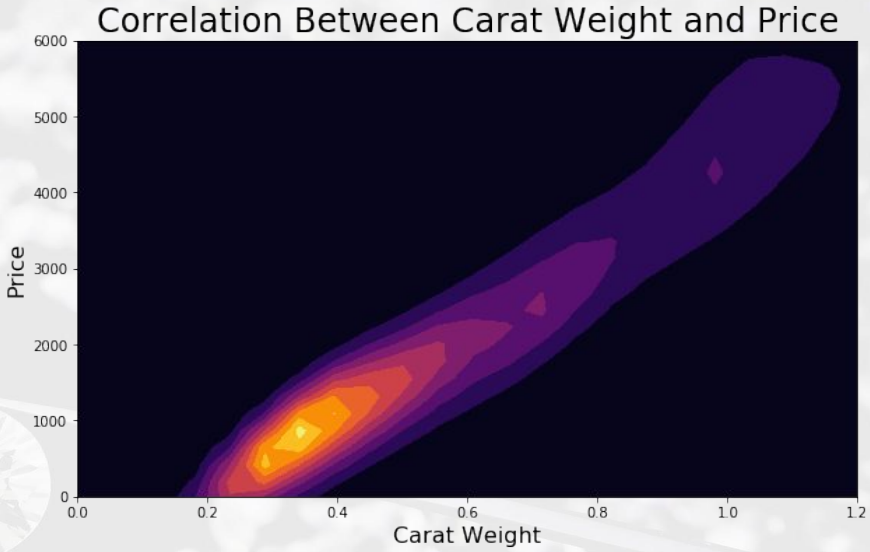
Carat Weight

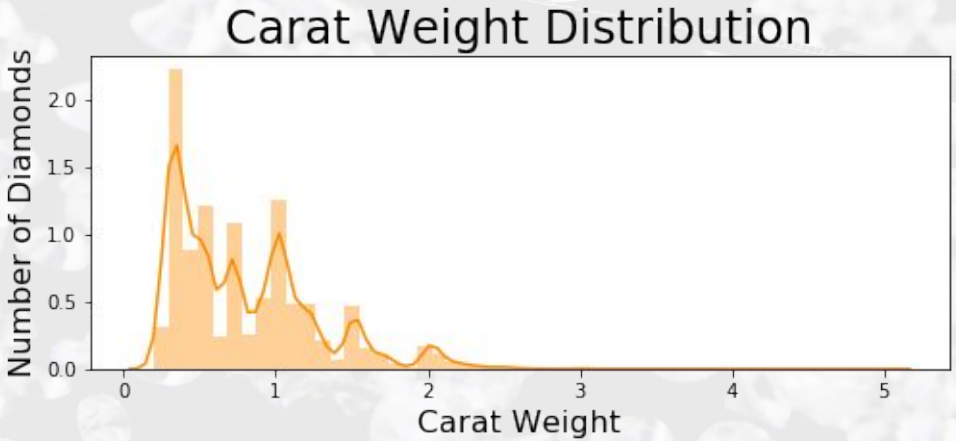
Carat weight is the standard used to compare the mass of diamonds. One carat is equivalent to 200 milligrams or 0.0070547924 ounces. High-precision scales are used to measure the carat weight of a gemstone.

I never worry about diets. The only carrots that interest me are the number you get in a diamond.
-Mae West

–Carat Weight

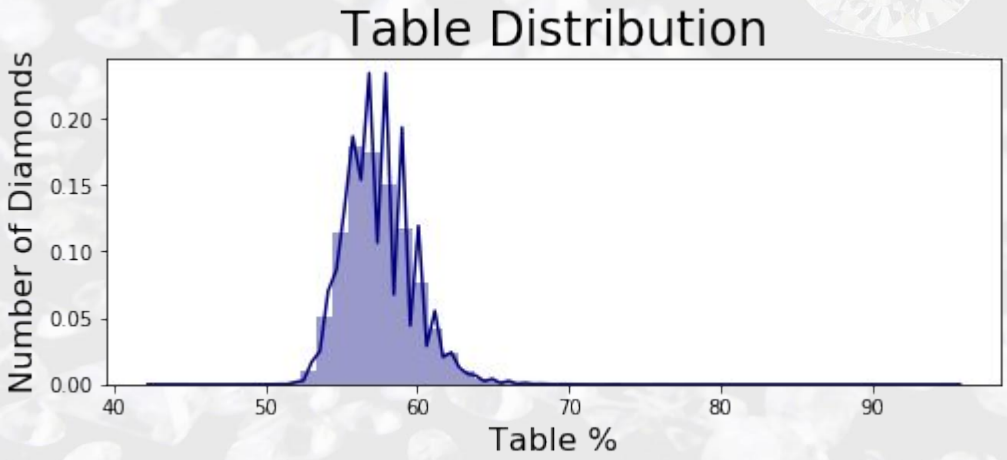
Visible spikes are shown in the distribution for carat weight at common “benchmark weights” such as ⅓, ½, ¾, and 1 carats. Carat weight accounts for most of the variation in price among the diamonds in the dataset.

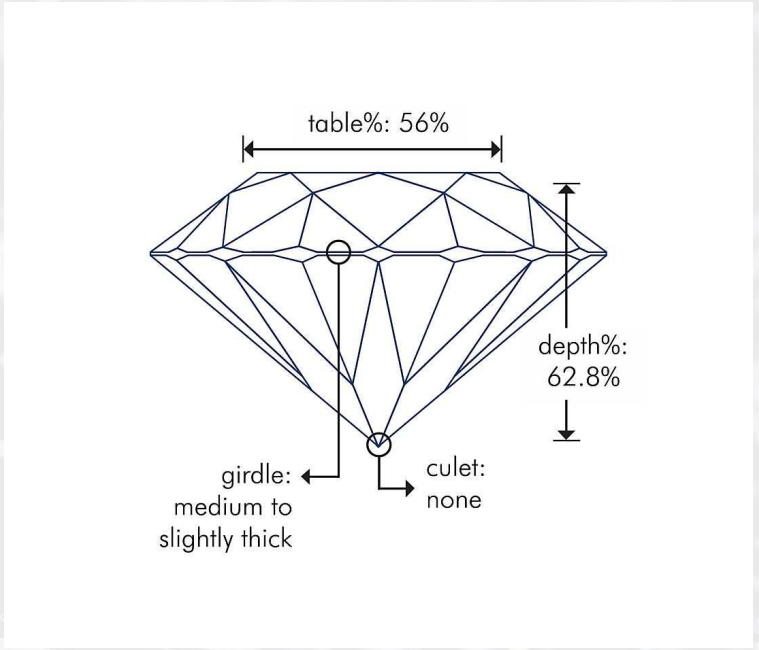




Table

A diamond's table is the flat surface on the top or the largest facet. It is measured in percentage of total width. The distribution here appears to be fairly normal.





Models

Ridge Regression
Lasso Regression
Random Forest
Gradient Boosting
K Nearest Neighbors
Support Vector Machine

Ridge Regression

R-squared Value:
0.9792243448981248

Average Cross Validation Score:
0.8907807543644193

Runtime:
0.0706410408 seconds

Lasso Regression

R-squared Value:
0.9790424546676428

Average Cross Validation Score:
0.8910724532614076

Runtime:
0.0716602802 seconds

Random Forest

R-squared Value:
0.992998352920663

Average Cross Validation Score:
0.7459212540554482

Runtime:
60.8871779442 seconds

Gradient Boosting

R-squared Value:
0.9914086356366941

Average Cross Validation Score:
0.8549451693301213

Runtime: 172.5330746174 seconds

K Nearest Neighbors

R-squared Value:
0.9956821263994483

Average Cross Validation Score:
0.42059741059005085

Runtime:
3.6659293175 seconds

Support Vector Machine

R-squared Value:
0.9880499061735765

Average Cross Validation Score:
0.9128761046328439

Runtime:
660.0739874840 seconds

Most Important Features

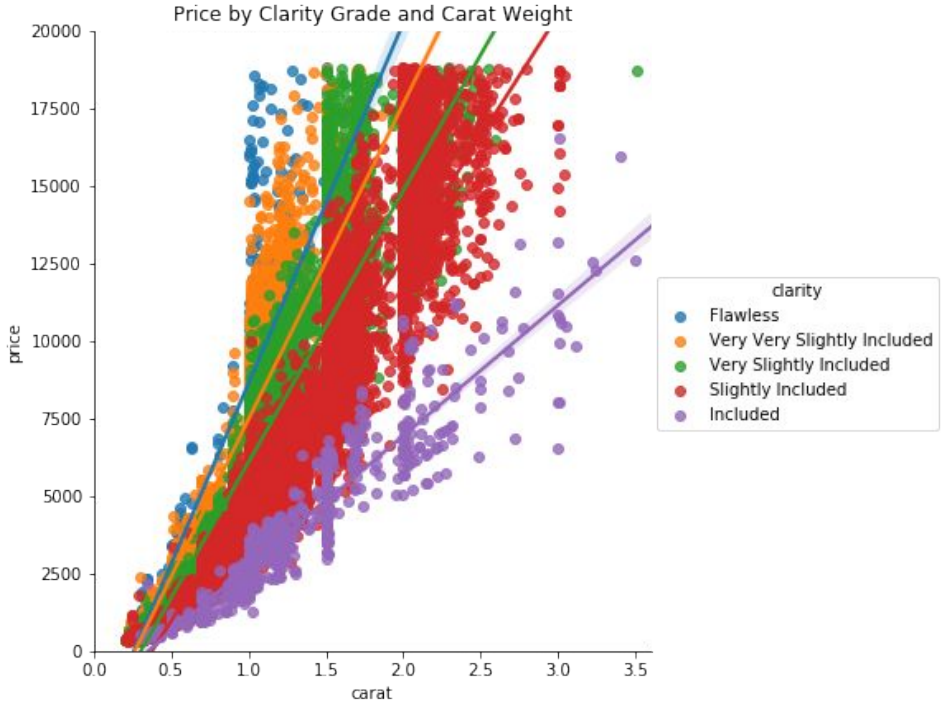
(from Gradient Boosting Model)

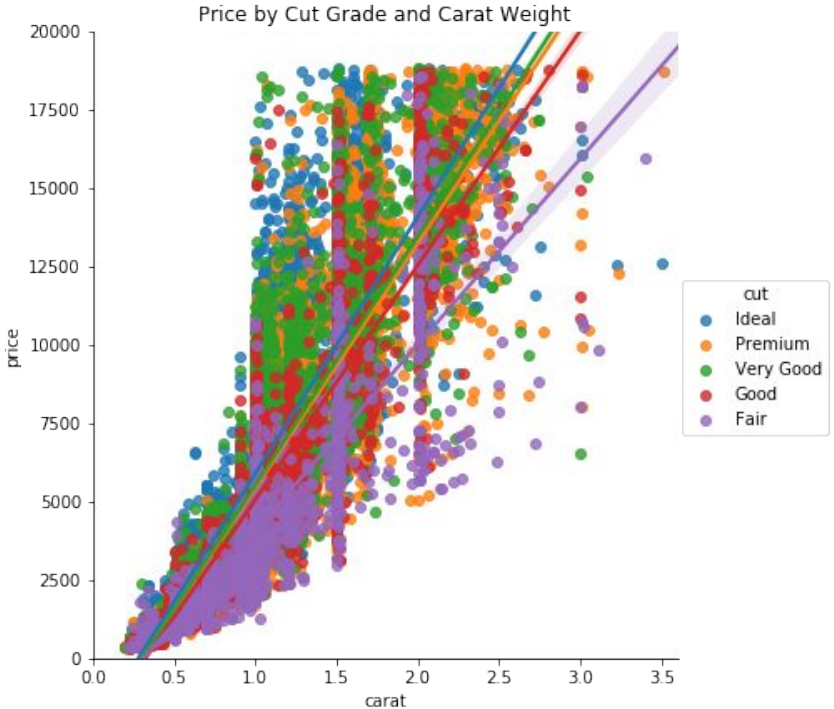
1. Carat - (0.377099)
2. Clarity - (0.189794)
3. Table - (0.171662)
4. Color - (0.144221)
5. Cut - (0.117225)

Effect of Clarity vs. Cut on Price

As previously stated, experts generally agree that cut is more important than clarity in determining the beauty of a diamond. However, that is not reflected in the price.

The plots below demonstrate how price is much more affected by clarity than cut.



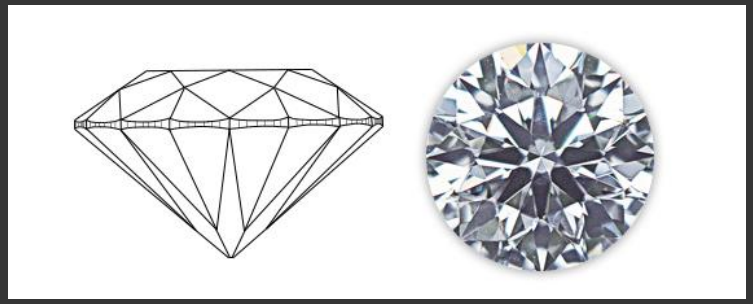


Want to save even more?

These diamonds are almost exactly the same, with one exception. The one on the left has a carat weight of 1.00 and the one on the right has a carat weight of .97.

Predicted Price:

$5,342

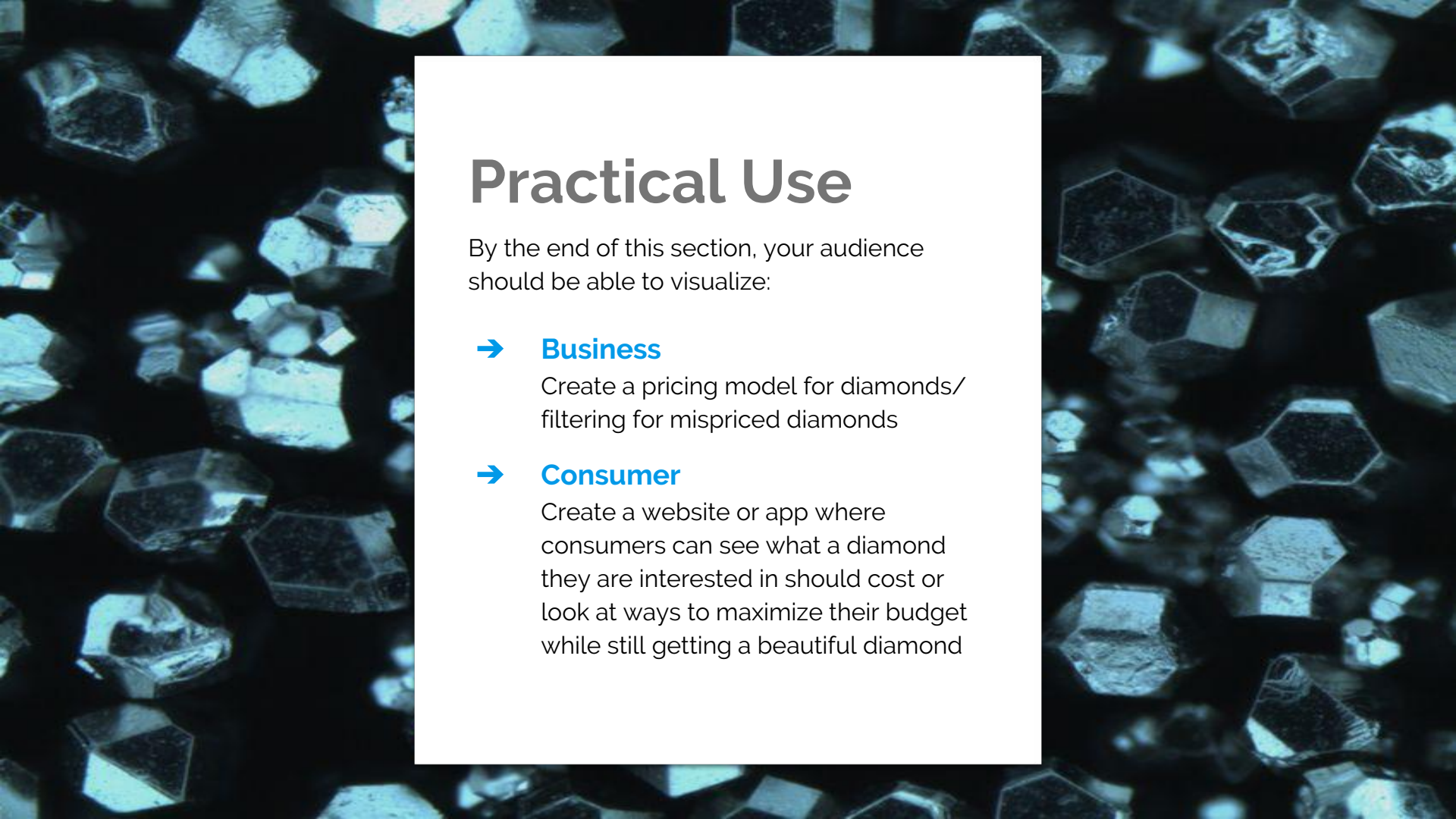


Predicted Price:

$4,738

Shortcomings

- **Other Factors**
  Some factors that can affect price were not available in the data, such as location of diamond flaws or a high level of fluorescence.
- **Quality of Data**
  The original source for this data is unknown. Is it retail prices from a single seller? Better data could be obtained.
- **Unbalanced Data**
  Oversampling could potentially improve models.



Practical Use

By the end of this section, your audience should be able to visualize:

- **Business**
  Create a pricing model for diamonds/ filtering for mispriced diamonds
- **Consumer**
  Create a website or app where consumers can see what a diamond they are interested in should cost or look at ways to maximize their budget while still getting a beautiful diamond