Examining The Relationship Between Characteristics of a Diamond and Its Price

University of Toronto JSC370 Final Project

Chia You (Benson) Chou

4/23/2023

Contents

Introduction

Growing up, it is common to hear my parent's talking about how much carat a diamond is and then proceeding to say it must be really expensive if they hear a high number. However, there are many more characteristics, such as the quality of the cut, color, clarity, table, and depth that can be taken account into when evaluating a diamond's value.

For this report, I will be using data scraped from Brilliant Earth. Brilliant Earth is a well-known jewelry company that focuses on ethically sourced and sustainable diamonds, gemstones, and metals. They offer a wide range of engagement rings, wedding bands, and other jewelry that are not only stunning but also socially conscious. On their website, they display a table of available diamonds and their characteristics such as price, shape, carat, quality of the cut, color, clarity, depth, and table. (More details of variables will be discussed in next section)

Research Question

In this report, I would like to investiage this question: What other qualities, besides from carat, influences a diamond's value the most? Do they have a positive or negative influence on a diamond's value?

Methods

Data Collection

As there are no API or directly downloadable data from Brilliant Earth, I scraped the data in Python, using Python's `Selenium Webdriver` to get the dynamic table's Json object, and using the `Json` and `pandas` packages to coerce this data into a data frame.

The table shown below is an example of how the dataset looks like.

Table 1: Diamond Table (Pre-cleaning)

price	shape	carat	cut	color	clarity	table	depth
11675	Round	1.25	Super Ideal	G	VS2	58.0	62.8
1355	Round	0.50	Super Ideal	I	SI2	60.0	59.6
7965	Round	1.02	Super Ideal	D	VS2	56.0	62.6
4255	Oval	1.00	Ideal	I	SI2	63.0	62.0
5345	Oval	1.50	Ideal	H	SI2	61.5	69.1

Table 2: Data Descriptions

Variables	Descriptions
price	Diamond's Price in CAD
shape	Shape of the diamond
carat	Weight of the diamond
cut	Quality of the cut
color	Diamond colour, from J (worst) to D (best)
clarity	A measurement of how clear the diamond is (SI2(worst), SI1, VS2, VS1, VVS2, VVS1, IF, FL(best))
table	Width of top of diamond relative to widest point
depth	Total depth percentage of the diamond

(More information on table and depth can be found here)

Data Cleaning

The data collected seems to be well-formatted and the columns are already selected to be variables of interest during the web scraping process. However, we still need to clean the data for: missing data, duplicates, factorization, and unit. With the is.na() function, we know that there are no missing data. I removed the duplicate observations ith the distinct() function. I replaced cut values 'Super Ideal' into 'Premium' for simplicity and avoid confusion. Lastly, I used as.factor() to factorize the character variables.

We can see that there are potential outliers from price, carat, table, and depth. Let's check if these observations were a mistake or not.

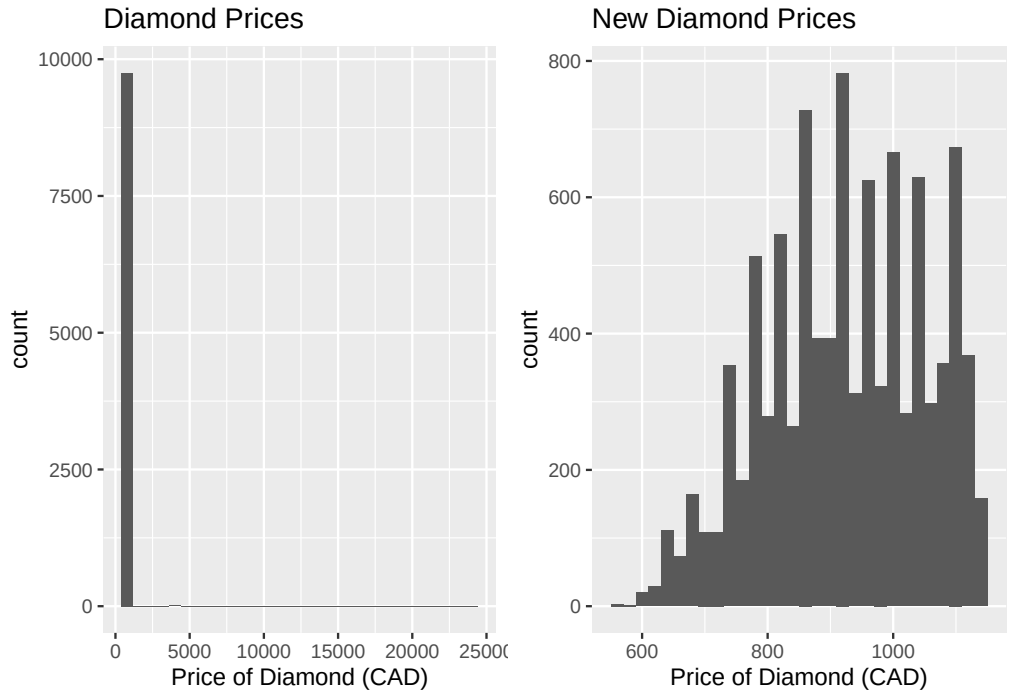
Table 3: Potential Outliers

	price	shape	carat	cut	color	clarity	table	depth
20	19805	Radiant	2.01	Premium	H	SI1	69	67.7
52	23795	Round	1.50	Premium	D	VVS2	60	62.2
853	750	Princess	0.31	Fair	F	SI2	87	73.0
8766	1095	Princess	0.40	Good	E	SI2	68	86.0

We can see that even though these observations do not seem to be recorded as a mistake or placeholder for `NA` values, from the histogram below, these outliers affects our analysis and limits the conclusion we can make. Thus, I will remove the diamonds with prices that are considered as outliers (1.5IQR + Q3 and Q1 - 1.5IQR).

Table 4: Summary Table (Post Cleaning)

price	shape	carat	cut	color	clarity	table	depth
Min. : 555.0	Round :7574	Min. :0.250	Fair : 23	D: 599	SI1 :3205	Min. :49.00	Min. :49.70
1st Qu.: 830.0	Princess: 505	1st Qu.:0.300	Good : 414	E:2515	SI2 :2387	1st Qu.:57.00	1st Qu.:61.40
Median : 925.0	Pear : 495	Median :0.310	Very Good:2042	F:2017	VS2 :1426	Median :58.00	Median :62.40
Mean : 926.6	Oval : 477	Mean :0.327	Ideal :2891	G:1695	VS1 : 979	Mean :59.19	Mean :62.95
3rd Qu.:1035.0	Emerald : 294	3rd Qu.:0.340	Premium :4374	H:1019	VVS2 : 863	3rd Qu.:60.00	3rd Qu.:63.20
Max. :1135.0	Marquise: 215	Max. :0.610	NA	I:1068	VVS1 : 711	Max. :87.00	Max. :86.00
NA	(Other) : 184	NA	NA	J: 831	(Other): 173	NA	NA



Originally, we have 9793 diamond observations, and after filtering, we reduced to 9744 diamond observations

Tools Used

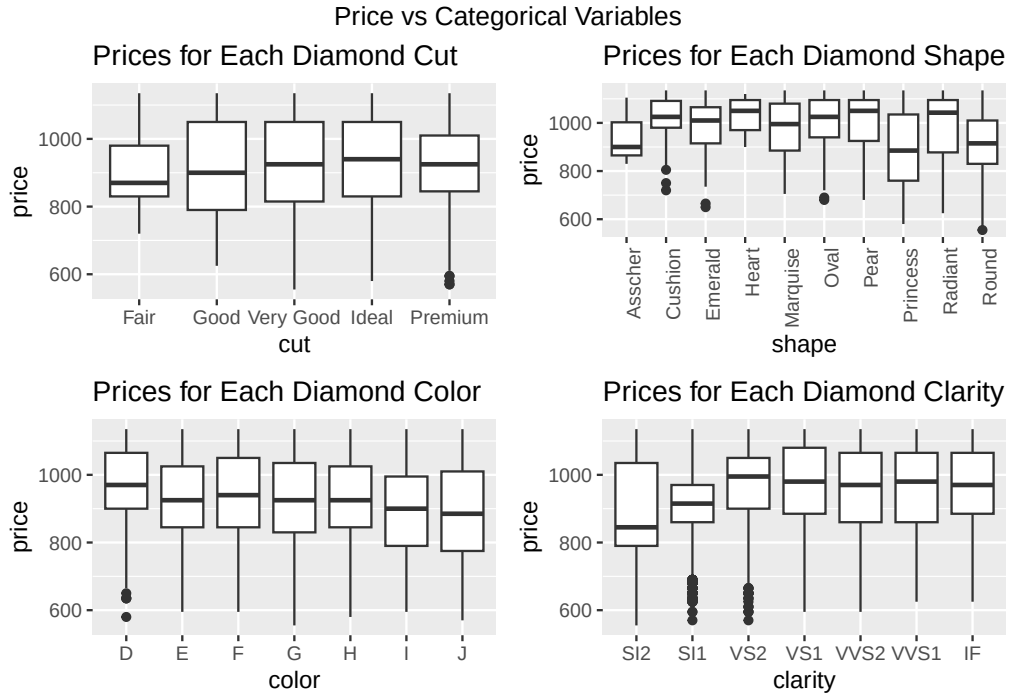
Data wrangling were completed with `tidyverse` and `dplyr`. Figures were created with `ggplot2`, interactive visuals were created using `plotly`. Tables were created with `kable` and `kableExtra`. Packages used for modelling include `rpart`, `randomForest`, `xgboost`, and `caret`.

Results

Summary Visuals

Multi-variable Relationships

First, I will create boxplots between price and categorical variables such as cut, shape, color, and clarity.



We can see that for cut and clarity, while the price range for each level are about the same, which indicates a variety of options for each quality level, the median price increases as the quality gets better. However, we see an inverse relationship between color quality and price. For prices in each diamond shape, we don't see a clear pattern. However, we do see that for the more commonly seen shape (Round), it has a wider price range with a median of about 850-900 CAD.

Next, I will create a general graph to show the relationship between carat and price for each cut and shape (Refer to website for interactive graph). We also see a similar trend among different groups, which is as carat increases, the expected diamond price also increases.

(Refer to website for interactive graph) Next, I will create scatterplot over price vs table and depth. We see that despite having diamonds with all ranges of table and depth at each price level, table does not seem to have a positive relationship with price, while the depth percentage has a positive relationship.

Modelling

Linear Model

I tried fitting a basic linear model with all the variables as price is a continuous variable that can be predicted by a linear model.

This model has an 59.5713399% which means that only about 59.5713399% of the variance of our predictors is explained by the variance of price. Notice that the insignificant variables are: `shapeCushion`, `shapeEmerald`, `shapePrincess`, `shapeRadiant`, `cutGood`, `cutVery Good` based on their p-values from Wald Z-test:

	Estimate	p-values
(Intercept)	-351.730	0.000
shapeCushion	21.054	0.655
shapeEmerald	40.939	0.382
shapeHeart	131.739	0.009
shapeMarquise	195.142	0.000
shapeOval	179.276	0.000
shapePear	162.285	0.001

(continued)

	Estimate	p-values
shapePrincess	-27.554	0.558
shapeRadiant	17.635	0.717
shapeRound	120.421	0.010
carat	2211.366	0.000
cutGood	18.947	0.278
cutVery Good	31.450	0.069
cutIdeal	55.980	0.001
cutPremium	75.387	0.000
colorE	-25.095	0.000
colorF	-27.285	0.000
colorG	-61.578	0.000
colorH	-105.997	0.000
colorI	-186.469	0.000
colorJ	-251.133	0.000
claritySI1	66.619	0.000
clarityVS2	158.542	0.000
clarityVS1	182.876	0.000
clarityVVS2	197.169	0.000
clarityVVS1	213.211	0.000
clarityIF	237.615	0.000
table	1.869	0.000
depth	3.918	0.000

With that in mind, I attempted to see if removing these insignificant factors will yield a better R^2.

This new model yields an R^2 of 0.48632, which is lower than the original model. Thus, we would want to take the full model rather than the reduced model.

Interpretation of Values

The estimates above represent the change in prices of a diamond for a unit increase in the continuous variables or for being a certain level of a categorical variable.

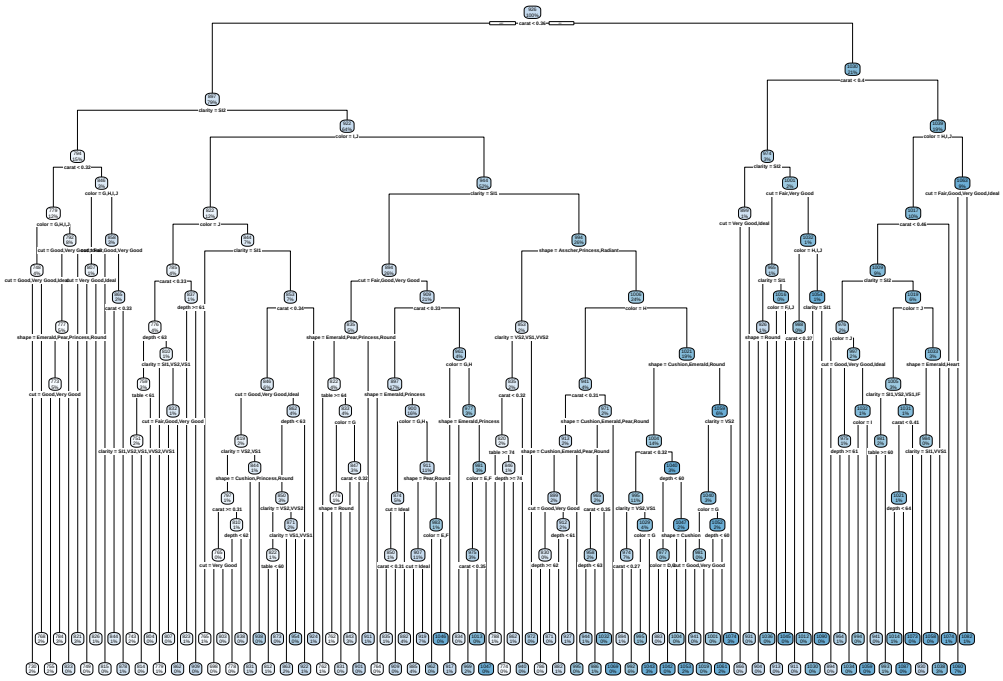
The estimated diamond price is about -351.7 CAD when all other variables are 0. Only looking at the variables with significant p-values, we see the most influential variable being carat, with an increase of 2211.366 CAD for each unit increase in carat. We also see that with the diamond having a clarity of IF (the second highest clarity), it increases the price by about 237.615 CAD. Surprisingly, our model predicts that no matter which color a diamond belongs to, it decreases the price.

Machine Learning Models

Next, I will try machine learning models such as Decision Tree, Random Forest, and XGBoost to further explore which variables are important when predicting a diamond price. RMSE will be used to compare model performance.

Decision Tree

I fit a decision tree to see what qualities can really separate the price level of a diamond. In order for the graph to be comprehensible and fit for the general cases, I set the max tree depth to 15, and pruned the tree at the minimum x-error, which has a cp-value of 0.0003427563.



Within each tree node, we have the predicted price and the percentage of that quality with the predicted price.

Table 6: Decision Tree Variable Importance

	Variable Importance
carat	30955853
clarity	29165407
color	18320085
shape	10631316
depth	7315823
table	5929165
cut	4109143

Random Forest

As the model attempts different value combinations, I didn't further specify more hyperparameters.

Table 7: Random Forest Variable Importance

	Variable Importance
carat	32794727
clarity	31350791
color	19501288
depth	13182664
table	7559503
shape	7498196
cut	5227533

XGBoost

XGBoost tunes the max depth, number of estimators, and learning rate for the optimal performance. The final model has the parameters: `nrounds = 300, max_depth = 5, eta = 0.1`

Table 8: XGBoost Variable Importance

	Overall
carat	0.3729774
colorJ	0.0674470
table	0.0614418
depth	0.0577699
clarityVS2	0.0547177
colorI	0.0458095
claritySI1	0.0422782
clarityVS1	0.0408084
clarityVVS2	0.0303189
shapeRound	0.0291391
cutPremium	0.0271444
clarityVVS1	0.0269797
clarityIF	0.0176089
colorH	0.0166674
shapeOval	0.0146190
cutVery Good	0.0141197
shapePear	0.0126653
colorE	0.0123657
shapePrincess	0.0121605
colorG	0.0102483
colorF	0.0088806
cutIdeal	0.0073893
shapeMarquise	0.0060945
shapeEmerald	0.0039101
cutGood	0.0035219
shapeCushion	0.0017272
shapeRadiant	0.0008084
shapeHeart	0.0003813
clarityFL	0.0000000

To find out which model performed the best, we will look for the model with the lowest RMSE. As we see from the table below, XGBoost yielded the best results out of the three.

Table 9: RMSE for Each Model

Model	RMSE
Decision Tree	72.93466
Random Forest	71.68729
XGBoost	69.91932

As the RMSE's are still within reasonable range, we can take the variable importance outputs and the results

from the linear model for a clear idea of the relationship between price and our predictors.

Conclusion

From our initial multi-variable relationship visuals, attributes such as cut, color, shape, clarity, carat, and depth correlate with the pricing of the diamond. In the linear model, we see that in addition to carat, attributes like: color, clarity, shape, cut, table, and depth all have factor levels that are significant. We weren't able to remove the insignificant factor levels as it decreases our R^2. Despite these insignificant varaibles and factors, things matched our expectations as carat positively influences the price the most while the other attributes, except color, also increases the price. This can be further supported by the output from our machine learning models. We noticed that carat, color, and clarity played a big role in predicting a diamond price as it appeared in the top 3 of all 3 models. In conclusion, we found that asides from carat, the next positively influential quality is clarity. And not surprisingly, the value of a diamond increases as its qualities get better.

Limitations & Future Directions

There are several limitations of this study. One would be how the data was obtained. As the table on the website was a dynamic table, the parsing process was made more difficult as it required to loop through different paths to request the data. This may caused us to miss some data that could be crucial to this project. Next, Brilliant Earth may not represent the whole diamond selling industry. Thus, there may be ways they set prices for their diamonds that are different to other vendors.

One thing that I want to further investigate that was not available was introducing the time when the prices are decided. I believe that time may also play a role when deciding the diamond prices (especially when there are high or low demands).