Diamonds

Berkay Canogullari, Stephen Embry, Youchen Zhang

December 2, 2020

1 Introduction

We are told in the movies, media, and people around us that diamonds are a symbol of love. If you've watched a diamond advertisement, it usually conveys a message that diamonds represent rare characters like clarity, durability, and longevity. Each diamond is unique. It has different sizes, shapes, and colors. However, diamonds are expensive. Buying a diamond is confusing. On the surface, it seems like nothing makes sense. Therefore, we want to learn what determines the prices of diamonds based on different features in the diamonds dataset so when we don't get ripped off and overpay when we purchase the ring.

2 Data

We use a dataset on the price of diamonds and various other characteristics to analyze the price. Table I provides description and basic descriptive statistics about each variable. The one change we made to the data was removing all the diamonds with 0 x, y, or, z values, because it is impossible for an object to have a dimension of 0.

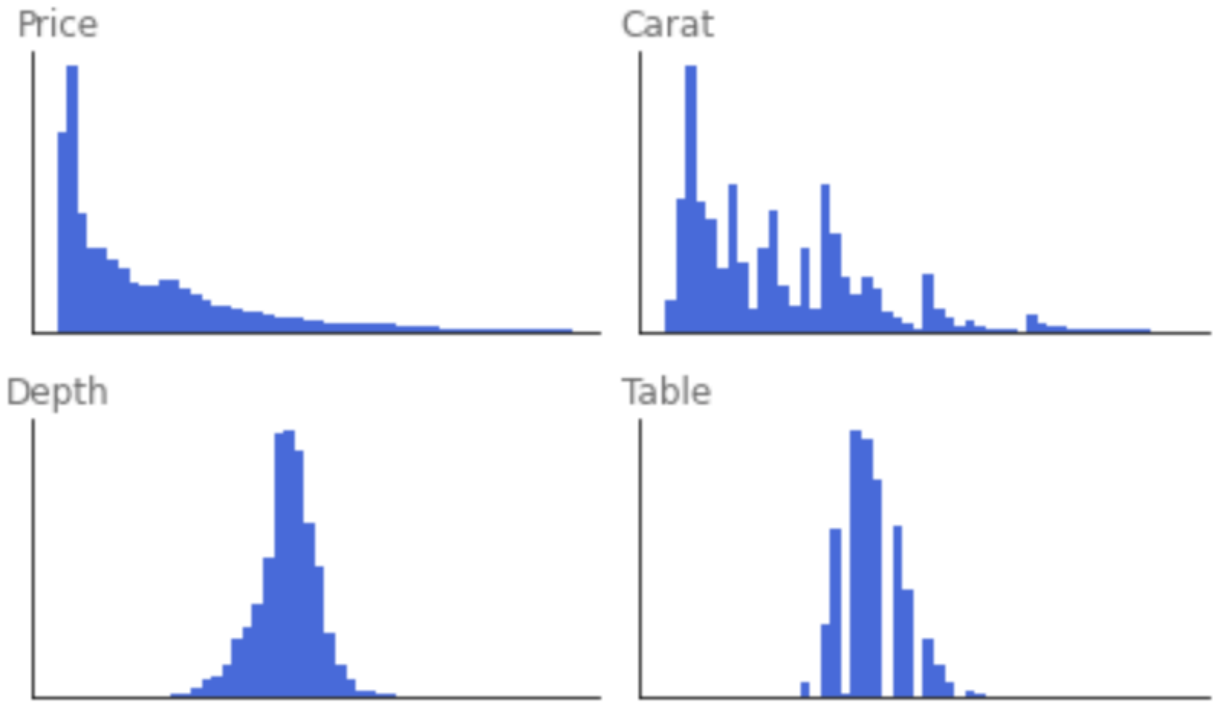
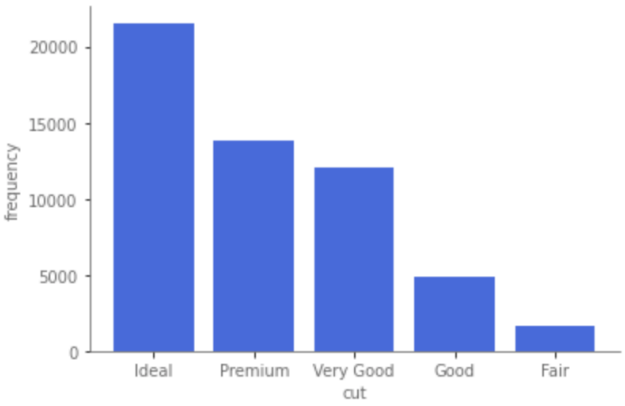


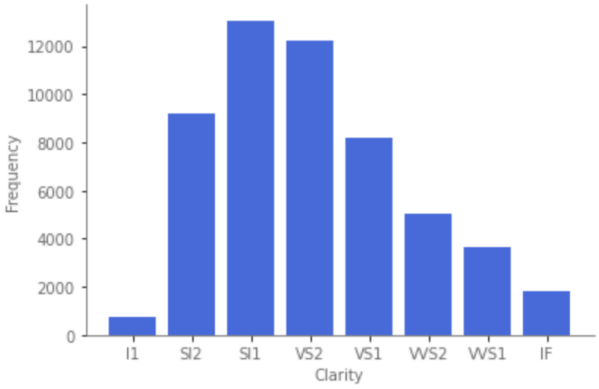
FIGURE I: Frequency histograms of price, carat, depth, and table

2.1 Exploratory Data Analysis

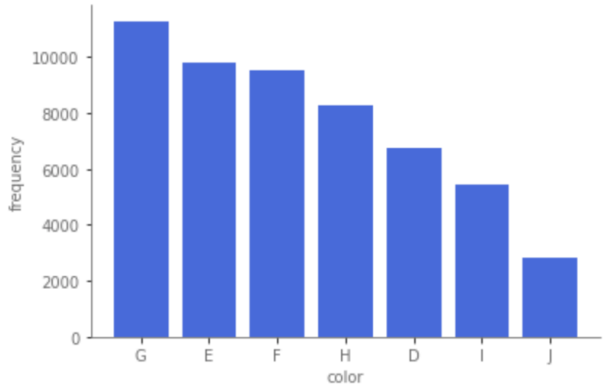
Frequency histograms are shown in I, and bar charts are shown in II. Based on the frequency histograms of price and other independent variables, we find out the histograms for price and carat are right skewed. The histograms for depth and table are centred around the mean. If we look at the bar plots for clarity, color, and cut, we find out these plots are also all right skewed. Because bar plots for clarity, color, and cut and histogram for carat are proportional to histogram of price, this may indicate the features clarity, color, cut, and carat have great impact on price. Lastly, from the description of the diamond dataset, the standard deviation for price is very high.



(A) Frequency by Cut



(B) Frequency by Clarity



(C) Frequency by Color

FIGURE II: Frequency Charts for Categorical Variables

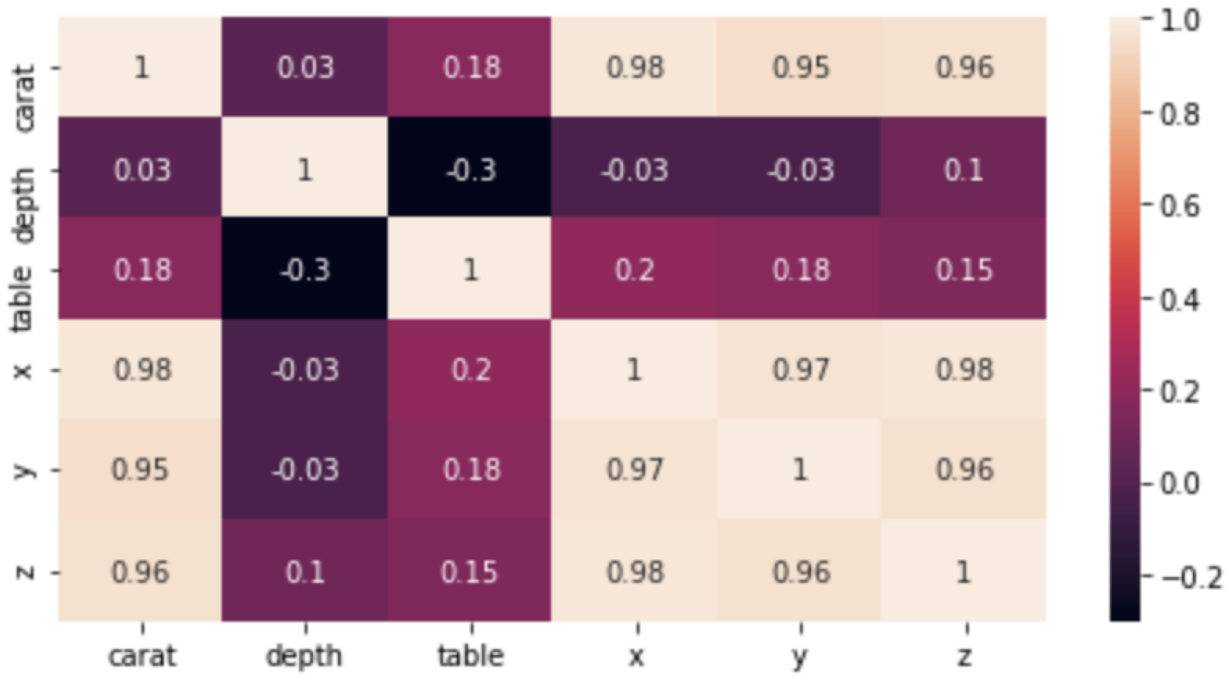


FIGURE III: Heatmap of collinearity between independent variables

3 Regression Analysis

3.1 Initial Variable Selection

To identify the independent variables to test in the regression, we ran tests on multicollinearity. Figure III shows a heat map of the possible continuous variables in the dataset. As can be seen, x, y, z, and carat are very highly correlated, indicating using these 4 independent variables would likely cause multicollinearity issues. As a result, we decided to combine x, y, and z into a volume column.

We checked the variance inflation factors using variables cut, color, clarity, carat, depth, table, and volume. Table II shows the results. Volume had an extremely high variance inflation factor of 23.6259. We took it out, and checked variance inflation factors again. Clarity has high numbers for some of the categories. However, since the predictors are still significant even after dropping the clarity from the model, we can keep clarity predictor in our model and there is no need to do pairwise chi-square tests from independence.

We believe there is a large VIF for clarity because it affects itself.

3.2 Modeling

We started off with a simple regression running price against our independent variables. (1) shows the original specification.

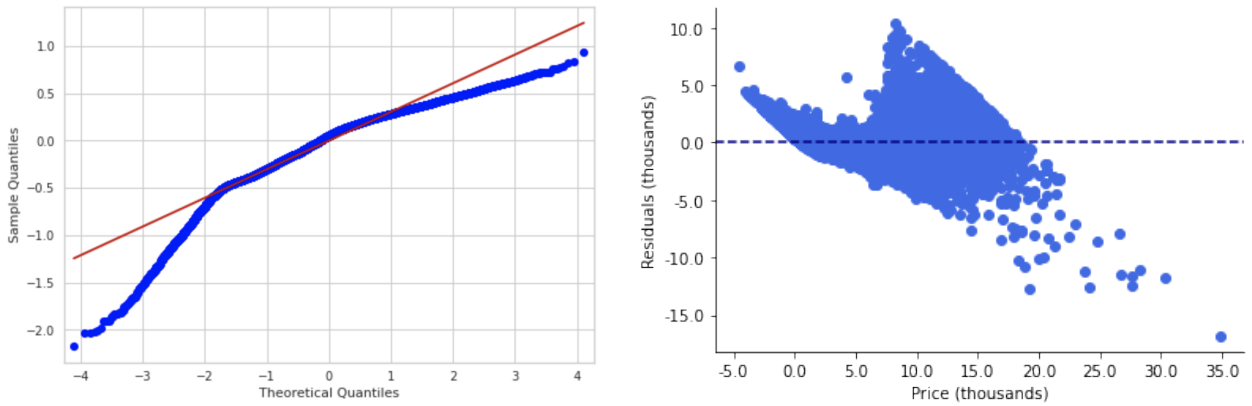
$$price = \beta_0 + \beta_1 carat + \beta_2 C(cut) + \beta_3 C(color) + \beta_4 C(clarity) + \beta_5 depth + \beta_6 table \quad (1)$$

However, we noticed issues with heteroskedasticity and non-normality in the data.

Figure IVa shows the qq-plots of the regression. The difference between the theoretical and empirical line shows that the regression is non-normal. Looking at the independent variables, we also identified carat to be non-normal.

Figure IVb shows a plot of the fitted values against their residuals. The errors are clearly correlated downward. This has two primary issues. The first is it is a sign of heteroskedasticity- the errors are smaller in the middle and larger on the outside. Our suspicions of heteroskedasticity were confirmed when the Breusch-Pagan test indicated that heteroskedasticity exists. The second is the large increase in residuals as price went up. This is likely because our dependent variable is highly skewed. This is related to the non-normality issue mentioned earlier.

To solve these issues, we took the log of the non-normal variables, removed the influential points, and used robust standard errors. Using the log of these variables helped fix the non-normality issue. In addition, it helped bring price to a smaller, less skewed value. To combat the issue of skewness, we removed the influential points in the dataset



(A) QQ plot

(B) Scatter of residuals vs fitted values

FIGURE IV: Diagnostics for Model 1

[1]. Models 2 and 3 were fit using

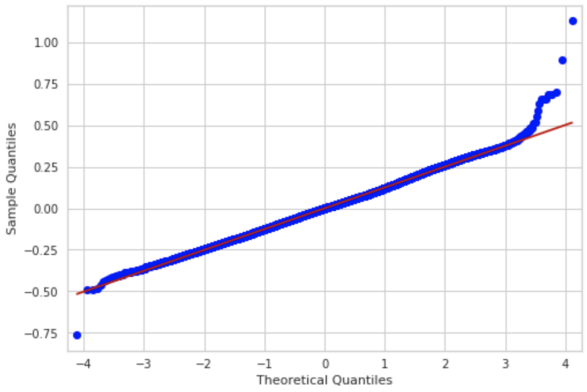
$$log(price) = \beta_0 + \beta_1 log(carat) + \beta_2 C(cut) + \beta_3 C(color) + \beta_4 C(clarity) + \beta_5 depth + \beta_6 table \tag{2}$$

with and without influential points, respectively. We provide model 2 for robustness, but our analysis will be done without influential points. The results can be found in table III. Figure V shows the qq plot and fitted vs residuals plots for model 3. Using this specification, the issues with the original model are gone. While there is still heteroskedasticity, it is corrected for in the standard errors.

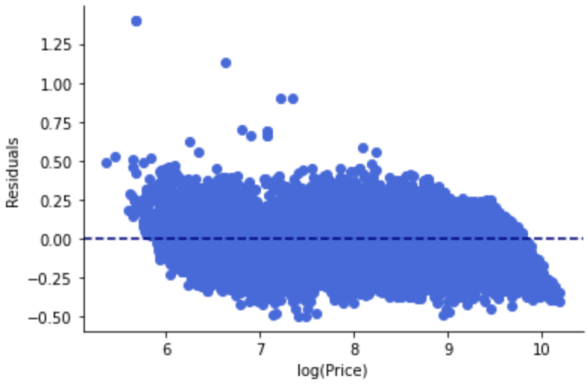
3.3 Model Selection

To select the variables to use in the regression, we use best subset selection. Table IV shows the best model for each number of predictors, selected by adjusted R^2 and Mallow's C_p. It has results almost identical. The 2 best subsets were 4 predictors and 5 predictors. Although 5 predictors performed slightly better, we chose 4 as the best subset because the improvement was too minimal to warrant another variable. The resulting

[1]Influential points were identified as points where cook's distance is greater than 4/n, where n is the number of data points we have



(A) QQ plot



(B) Scatter of residuals vs fitted values

FIGURE V: Diagnostics for Model 3

final model specification is

$$log(price) = \beta_0 + \beta_1 log(carat) + \beta_2 C(cut) + \beta_3 C(color) + \beta_4 C(clarity) \tag{3}$$

The results are listed as model 4 in Table III.

3.4 Analysis of Regression

Our model has an R^2 value of 0.983, indicating that it did a very good job explaining the variance in the data. This implies that, with these variables, we have a good idea on how much a diamond is worth.

Looking at the coefficients provide insights on how our model weighs the effect of different variables on the price of diamonds. All coefficients show up as significant, indicating that every variable used was viewed as important in predicting the diamond price. Carats had the highest correlation with price and is the single best predictor, but it does not have large effect sizes- our model predicts that a 1% increase in carats leads to a 1.878% increase in price. The variable that had the largest effect size was clarity. Comparing the lowest clarity to the highest clarity, our model predicts a 104.22% increase in price. The model also predicts very large percent changes in price between the other

clarity categories.

4 Conclusion

According to our case study, we conclude that the diamond price can be accurately predicted using certain characteristics of the diamond, such as carat, cut, color, and clarity. Based on our findings, when choosing a diamond, you don't need to check both dimensions of the diamond and carat, since there is multicollinearity between those characteristics. Thus, it's sufficient and sensible only to check the carat weight in addition to the cut, color, and clarity of a diamond when comparing prices between different diamonds. If you are in a hurry and don't have time to do a detailed analysis to pick a diamond and wondering if you are overpaying or underpaying for it, it's mostly enough to look at the carat and clarity of a diamond since these two characteristics have the highest correlation with the price of a diamond.

A Tables

TABLE I: Data Description

Variables	Mean	Description
Price	3930.99 (3987.28)	price in US dollars ($326–$18,823)
Carat	0.7977 (0.4378)	weight of the diamond (0.2–5.01)
Depth	61.7495 (1.4323)	total depth percentage
Table	57.4568 (2.2341)	width of top of diamond relative to widest point (43–95)
X	5.7316 (1.1194)	length in mm (0–10.74)
Y	5.7349 (1.1401)	width in mm (0–58.9)
Z	3.54 (0.7025)	depth in mm (0–31.8)
Cut		quality of the cut (Fair, Good, Very Good, Premium, Ideal)
Clarity		a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)
Color		diamond colour, from J (worst) to D (best)

TABLE II: Variance Inflation Factors

Variables	With Volume	Without Volume
carat	24.1178	1.323
Cut: Good	3.9355	3.9353
Cut: Very Good	7.6154	7.6143
Cut: Premium	8.3447	8.3432
Cut: Ideal	7.6154	11.2978
Clarity: SI2	11.4839	11.4826
Clarity: SI1	14.7839	14.7814
Clarity: VS2	14.3029	14.3018
Clarity: VS1	10.803	10.8017
Clarity: VVS2	7.5826	7.5819
Clarity: VVS1	5.9484	5.9479
Clarity: IF	3.5354	3.5352
Color: E	2.01	2.01
Color: F	2.01	2.01
Color: G	2.1919	2.1919
Color: H	1.9512	1.9511
Color: I	1.709	1.709
Color: J	1.4215	1.4215
Depth	1.3905	1.3784
Table	1.7939	1.7866
Volume	23.6259	

TABLE III: Descriptive Statistics

Variables	Model 2	Model 3	Model 4
log(carat)	1.8779**	1.8738**	1.8777**
	(0.001)	(0.001)	(0.001)
Cut: Good	0.0742**	0.0936**	0.0750**
	(0.004)	(0.004)	(0.004)
Cut: Very Good	0.1083**	0.1279**	0.1098
	(0.004)	(0.004)	(0.004)
Cut: Premium	0.1333**	0.1523**	0.1346**
	(0.004)	(375531.6)	(0.004)
Cut: Ideal	0.1526**	0.1732**	0.1553**
	(0.004)	(0.004)	(0.004)
Clarity: SI2	0.3827**	0.3217**	0.3828**
	(0.007)	(0.008)	(0.008)
Clarity: SI1	0.5424**	0.4728**	0.5425**
	(0.007)	(0.008)	(0.008)
Clarity: VS2	0.6895**	0.6175**	0.6896**
	(0.007)	(0.008)	(0.008)
Clarity: VS1	0.7579**	0.6851**	0.7580**
	(0.007)	(0.008)	(0.008)
Clarity: VVS2	0.8874**	0.8119**	0.8875**
	(0.008)	(0.008)	(0.008)
Clarity: VVS1	0.9545**	0.8820**	0.9546**
	(0.008)	(0.009)	(0.008)
Clarity: IF	1.0419**	0.9684**	1.0422**
	(0.008)	(0.009)	(0.008)
Color: E	-0.0513**	-0.0513**	-0.0513**
	(0.002)	(0.002)	(0.002)
Color: F	-0.0920**	-0.0919**	-0.0920**
	(0.002)	(0.002)	(0.002)
Color: G	-0.1502**	-0.1479**	-0.1501**
	(0.002)	(0.002)	(0.002)
Color: H	-0.2388**	-0.2373**	-0.2388**
	(0.002)	(0.002)	(0.002)
Color: I	-0.3603**	-0.3579**	-0.3603**
	(0.002)	(0.003)	(0.002)
Color: J	-0.4979**	-0.4946**	-0.4979
	(0.003)	(0.003)	(0.003)
Depth	-0.0004	-0.0003	
	(0.0005)	(0.0005)	
Table	-0.0006*	-0.0003	
	(0.0003)	(0.0003)	
F-stat	1.459e5	1.28e5	1.487e5
R^2	0.983	0.983	0.983
Adj R^2	0.983	0.983	0.983

Robust s.e.s are used. Cut and Color are ordered ascending, color descending

** p<.05, * p<0.1

TABLE IV: Best Subsets selection

Num predictors	Mallow's C_p	Adj R^2	AIC	BIC	log(Carat)	Cut	Clarity	Color	Depth	Table
1	128656.809	0.9396	-2461.6006	-2443.9326	x					
2	50207.8848	0.966	-31616.2442	-31536.7382	x		x			
3	3206.53	0.9818	-63398.4671	-63265.9572	x		x	x		
4	5.817	0.9829	-66496.3685	-66328.5226	x	x	x	x		
5	5.7751	0.9829	-66496.4106	-66319.7307	x	x	x	x		x
6	7	0.9829	-66495.1858	-66309.6719	x	x	x	x	x	x

Group Members	Berkay	Stephen	Victor
Proportion of Work	33%	33%	33%
List of Work	Initial check of the data (summary tables and graphs) Discussion of problems and chosen methods Data Cleaning and Modeling Discussion of Modeling Results Model Selection and Verification Final Discussion Write the Report	Initial check of the data (summary tables and graphs) Discussion of problems and chosen methods Data Cleaning and Modeling Discussion of Modeling Results Model Selection and Verification Final Discussion Write the Report	Initial check of the data (summary tables and graphs) Discussion of problems and chosen methods Data Cleaning and Modeling Discussion of Modeling Results Model Selection and Verification Final Discussion Write the Report