MA334-AU-7 ASSIGNMENT

Vamshi Beejarapu

2026-01-13

Table 1: First Six Rows of the Diamond Dataset

carat	cut	color	clarity	depth	table	price	x	y	z
0.71	Good	D	SI2	64.3	56	2215	5.64	5.59	3.61
0.51	Ideal	D	VS2	61.4	57	1716	5.13	5.16	3.16
1.00	Very Good	D	SI2	62.5	58	4661	6.35	6.44	4.00
0.70	Premium	H	VS1	62.8	60	2367	5.61	5.56	3.51
0.23	Premium	I	VVS1	60.5	61	414	3.98	3.95	2.40
1.16	Ideal	G	SI2	61.9	56	4969	6.77	6.73	4.18

1. Data Exploration

Table 2: Structure of the Dataset

	Variable	Data_Type
carat	carat	numeric
cut	cut	character
color	color	character
clarity	clarity	character
depth	depth	numeric
table	table	numeric
price	price	integer
x	x	numeric
y	y	numeric
z	z	numeric

Table 3: Dimensions of the Dataset

Rows	Columns
1000	10

Table 4: Numeric and Categorical Variables

Numeric_Variables	Categorical_Variables
carat, depth, table, price, x, y, z	cut, color, clarity

Table 5: Unique Levels in Categorical Variables

Variable	Levels
cut	Good, Ideal, Very Good, Premium, Fair
color	D, H, I, G, E, F, J
clarity	SI2, VS2, VS1, VVS1, SI1, VVS2, IF, I1

It is a medium-sized data set with 1000 data points and 10 variables. It has 7 numerical variables (carat, depth, table, price, x, y, and z) and 3 nominal variables (cut, color, and clarity) with 5, 7 (from D to J) categories, and 8 (from I1 to IF) categories, respectively, and no missing data points.

Table 6: Summary Statistics for Numerical Variables

	n	mean	sd	min	max
carat	1000	0.80	0.49	0.23	4.13
depth	1000	61.74	1.45	53.80	69.00
table	1000	57.58	2.14	52.00	66.00
price	1000	3951.52	4092.06	375.00	18680.00
x	1000	5.74	1.12	3.92	10.00
y	1000	5.74	1.12	3.95	9.85
z	1000	3.54	0.70	2.39	6.43

Table 7: Frequency Distribution of Diamond Cut

Cut	Frequency
Fair	28
Good	88
Ideal	377
Premium	265
Very Good	242

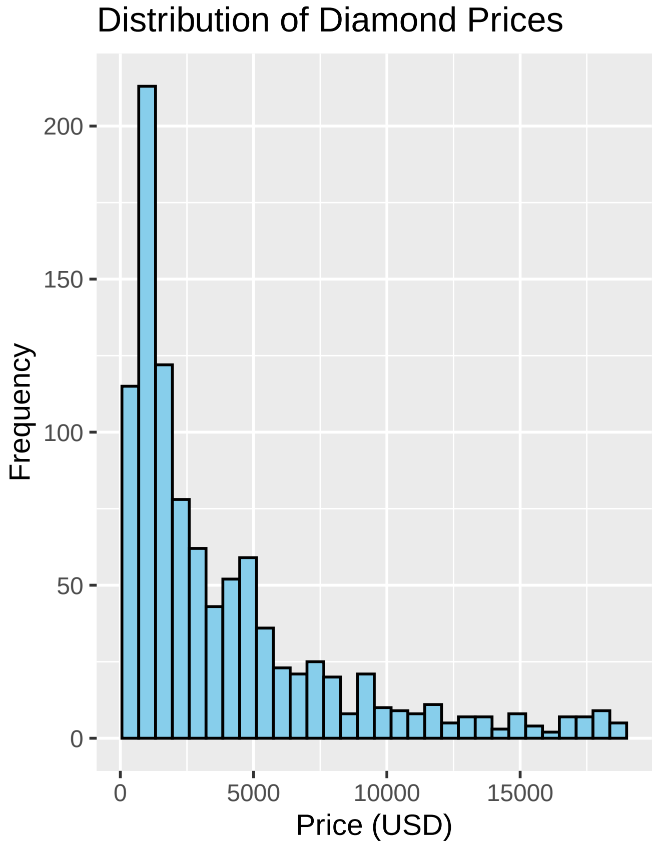
Table 8: Frequency Distribution of Diamond Color

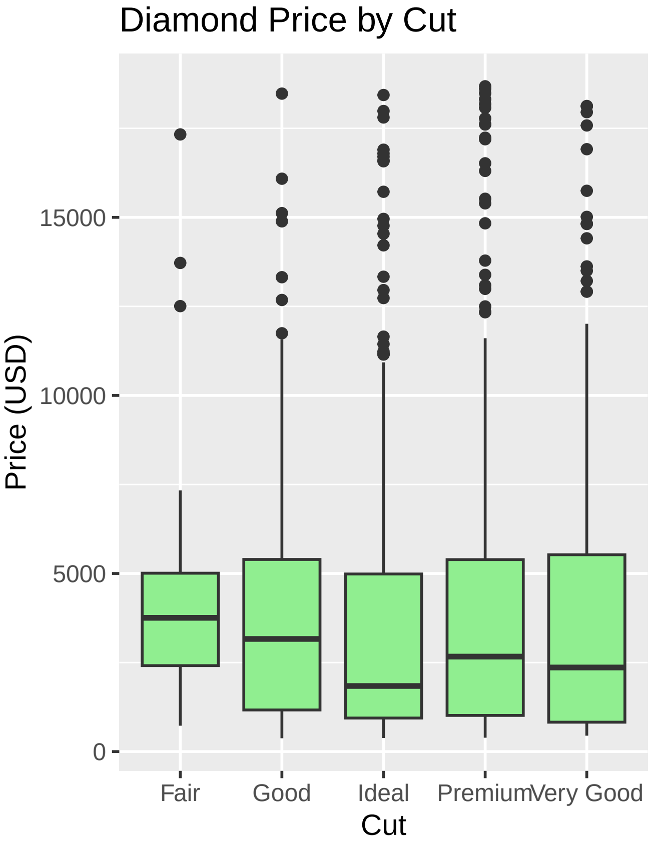
Color	Frequency
D	125
E	192
F	178
G	204
H	149
I	102
J	50

Table 9: Frequency Distribution of Diamond Clarity

Clarity	Frequency
I1	15
IF	31
SI1	246
SI2	177
VS1	151
VS2	213
VVS1	76
VVS2	91

This summary shows the average diamond carat is 0.80 with a maximum value of 4.13, indicating small and big diamonds. The average price for one is 3951.52, though with high standard deviation of 4092.06; the median is 2320.5, meaning that there is right-skewed distribution due to expensive diamonds. The average depth is 61.74 and the table value is 57.58; both are consistent. Diameter variables increase with carat value, which can be affordable to very expensive.

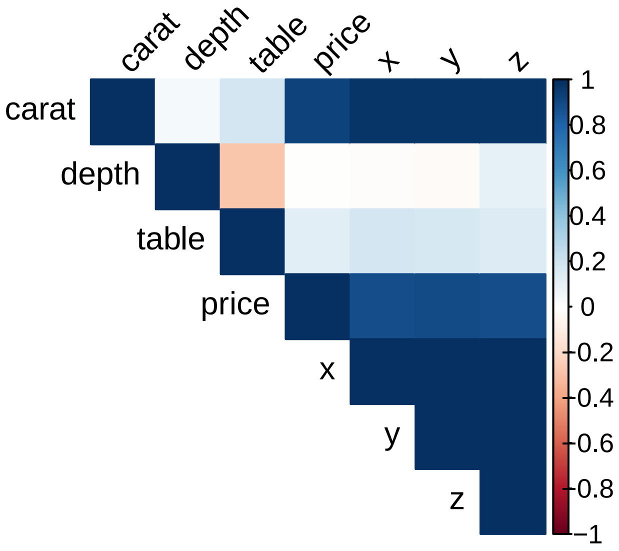

Distribution of Diamond Prices


Diamond Price by Cut

Visualizations show the distribution of values. The histogram of price will likely be dominated by low-priced diamonds, with a long tail comprising higher-priced diamonds-indicating positive skewness. Using a box plot, one could also investigate the variation of prices across different cut groups: ‘Ideal’ and ‘Premium’ diamonds have a higher price range than others. However, the overlapping nature of the groups shows that for a particular carat, clarity, and color, the prices are very different; outliers will correspond to high-priced diamonds.

Table 10: Correlation Matrix of Numerical Variables

Variable	carat	depth	table	price	x	y	z
carat	1.00	0.05	0.18	0.92	0.97	0.97	0.98
depth	0.05	1.00	-0.27	-0.01	-0.02	-0.02	0.11
table	0.18	-0.27	1.00	0.12	0.18	0.18	0.15
price	0.92	-0.01	0.12	1.00	0.89	0.89	0.88
x	0.97	-0.02	0.18	0.89	1.00	1.00	0.99
y	0.97	-0.02	0.18	0.89	1.00	1.00	0.99
z	0.98	0.11	0.15	0.88	0.99	0.99	1.00



The correlation matrix represents the linkage between the numeric attributes. Carat (0.92) has the most correlation in comparison to price, which means larger diamonds cost more. The attributes x (0.89), y (0.89), and z (0.88) of diamond size have a high correlation in relation to price. Attributes x, y, and z have a perfect correlation of +1.00, which varies directly when size is increased. Depth has less correlation (0.01) in comparison to price.

2.Probability, probability distributions and confidence intervals

[1] 0.093 [1] 0.377

The probability values are estimated by dividing the number of occurrences of events in the dataset by the number of diamonds present. Close to 9.3% of diamonds sell at above $10,000, which means approximately

93, while near or about 37.7% of them have an Ideal cut, which is estimated at approximately 377 diamonds out of 1000.

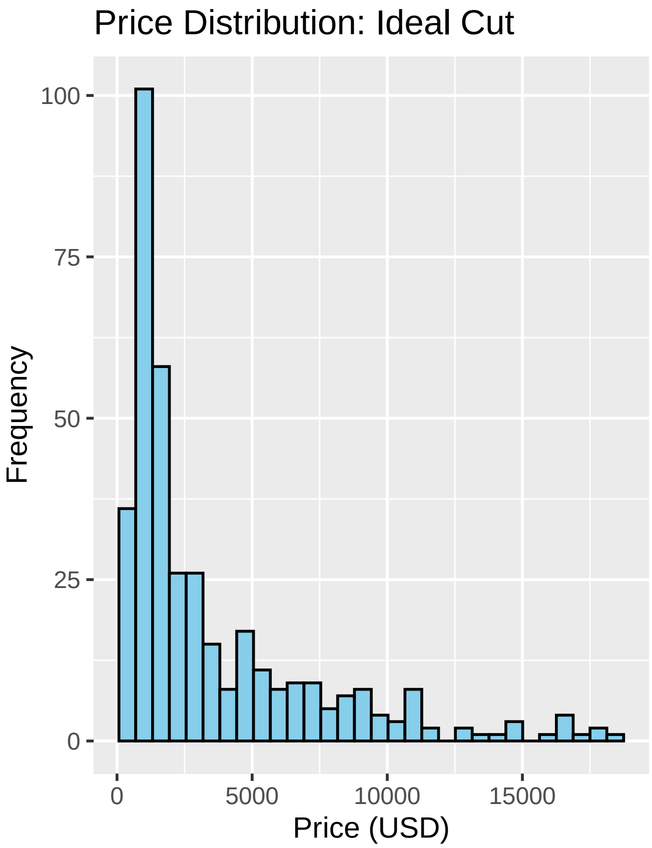
```
[1] 0.01221397
```

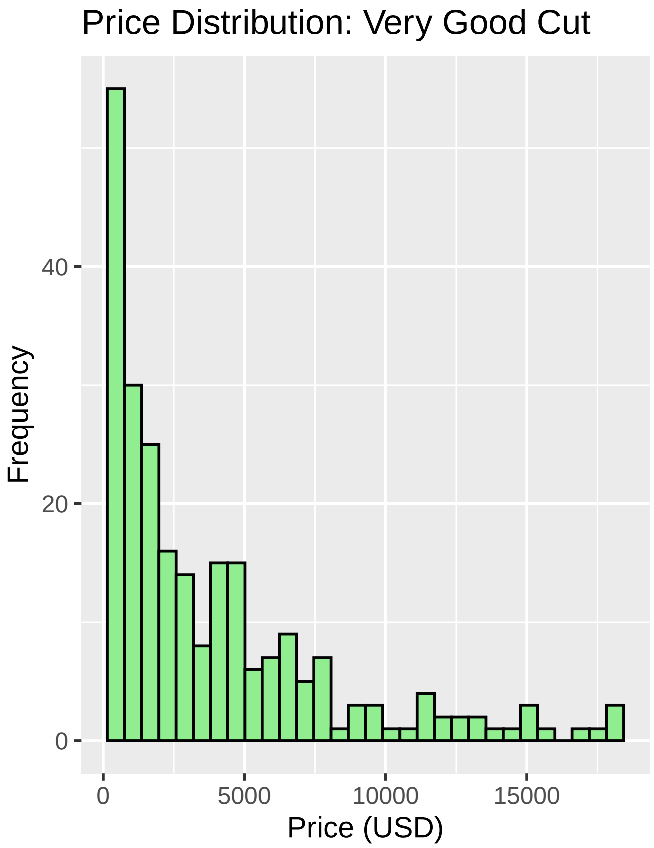
This is Binomial model and a sample of 20 diamonds is tested for Ideal cuts (p = 0.377). Let the random variable X represent the number of Ideal diamonds , where the probability of getting Ideal diamonds follows the Binomial distribution with the number of trials (n) = 20 and probability of success (p) = 0.377. This can be done through: P(X>12) = 1 - P(X <= 12). It is observed that the probability of getting more than 12 Ideal diamond is less and measures about 1.22%.

```
[1] 0.7766077 0.8272923 attr(,“conf.level”) [1] 0.9 [1] 0.7717442 0.8321558 attr(,“conf.level”) [1] 0.95
```

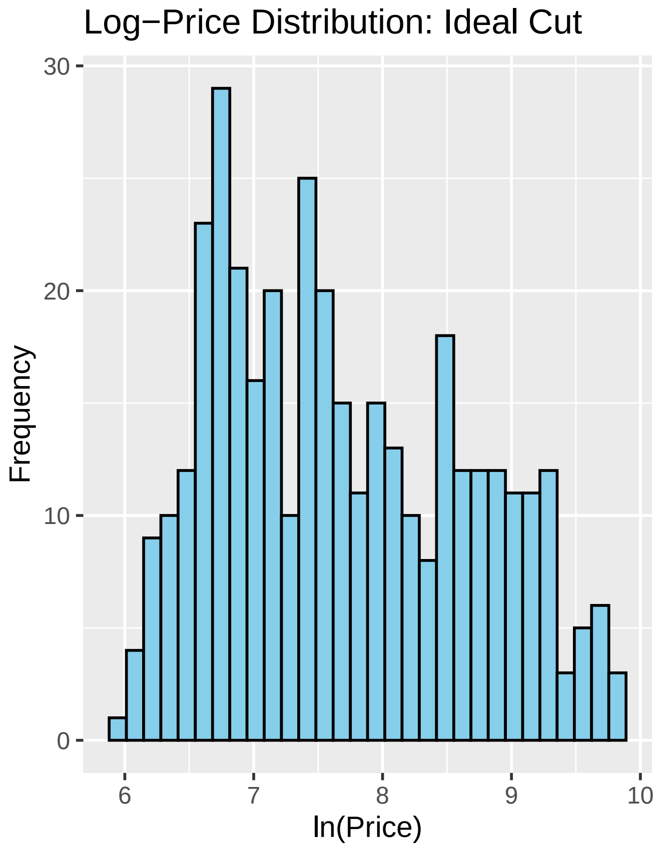
Confidence intervals use sample data to approximate the average carat weights of diamonds. From the 90% confidence interval (0.7766, 0.8273), there is a 90% chance that the actual mean will be within this interval, as compared with the 95% confidence interval (0.7717, 0.8322), with a wider interval that entails more error but with larger confidence levels. The wider interval ensures a larger error margin with the aim of being as accurate as possible with high confidence levels.

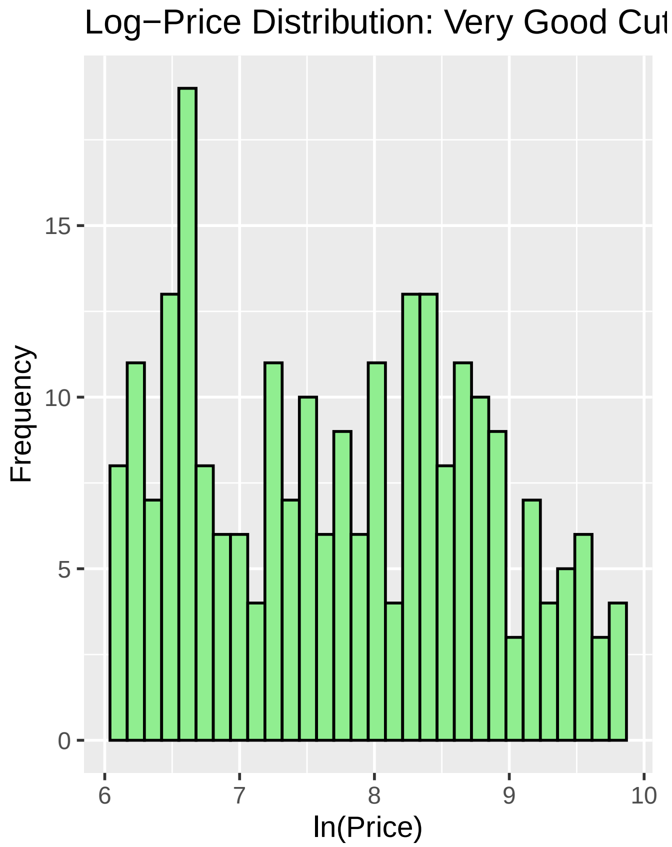
3. Hypothesis Tests





Comment: The price distributions usually appear right-skewed, so they do not look normally distributed.





Comment: ln(price) looks more symmetric and closer to normal than raw price for both cut types.

```
Welch Two Sample t-test
```

data: price by cut t = -1.1178, df = 487.79, p-value = 0.2642 alternative hypothesis: true difference in means between group Ideal and group Very Good is not equal to 0 95 percent confidence interval: -1003.7038 275.7905 sample estimates: mean in group Ideal mean in group Very Good 3552.233 3916.190

The t-test performed on diamond prices for Ideal and Very Good cut diamonds gave a test statistic of -1.118, meaning that the Ideal cut diamond price is slightly lower. However, since the p-value of 0.264 is greater than 0.05, there are no significantly different prices. This variation would be caused by sampling error, and it can be concluded that there are no significantly different prices.

```
Welch Two Sample t-test
```

data: ln_price by cut t = -0.71731, df = 483.97, p-value = 0.4735 alternative hypothesis: true difference in means between group Ideal and group Very Good is not equal to 0 95 percent confidence interval: -0.2252145 0.1047537 sample estimates: mean in group Ideal mean in group Very Good 7.694990 7.755221

This t-test on the log mean price, ln(price), is used to check the difference between the Ideal and Very Good cut diamonds. This is because the actual prices of diamonds are not normally distributed and normalized by this transformation. This value of the t-statistic again comes out to be -0.717, further justifying the interpretation that the mean log prices, ln(price), of Ideal diamonds are slightly less, but the difference is minimal. Further, the p-value comes out to be 0.474, again becomes much larger than 0.05. Therefore, there remains no statistically significant difference with respect to the log prices, ln(price), between the two groups based upon their respective cuts. This measure of the log price, ln(price), denotes proportional change in prices. Therefore, this test also justifies that there remains no significant change based upon their respective prices because of the cuts they possess.

(d)Comparison of conclusions from (b) and (c)

Graphs of both tests demonstrate no significant difference between Ideal and Very Good diamonds. The p-values of t-tests were 0.264 for the price and 0.474 for log(price); both above 0.05. Logging prices will address the skewness but does not change the test outcome: cut type is not a significant factor in the average price.

Table 11: Contingency Table: Price Category vs Clarity

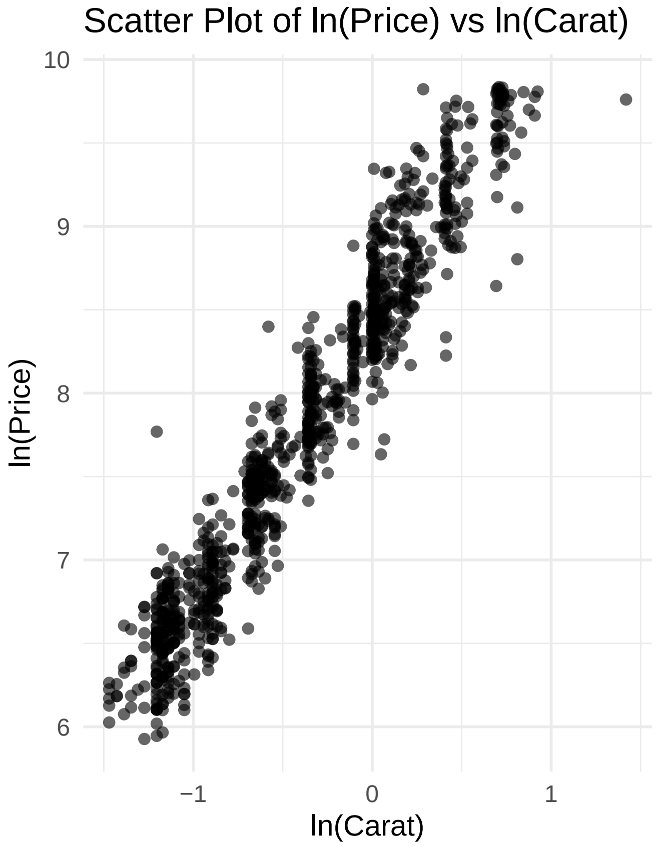
	I1	IF	SI1	SI2	VS1	VS2	VVS1	VVS2
High	4	6	57	63	39	69	6	22
Low	11	25	189	114	112	144	70	69

```
Pearson's Chi-squared test
```

data: contingency_table X-squared = 27.25, df = 7, p-value = 0.0003005

This contingency table represents the distribution of diamonds across price categories (High > $5000 or Low <= $5000) and clarity levels. The high-priced diamonds are dominant in mid-to-high clarity levels (VS2, SI2, SI1), whereas the low-price category dominates overall. Chi-square: 27.25, p-value: 0.0003. We deduce that clarity is associated with the likelihood that a given diamond is high-priced, with the lower clarity diamonds being less likely to fall in this category.

4. Linear Regression



The graph of ln(price) vs ln(carat) indicates that there is a strong positive trend, meaning that as carat, the price will also escalate. The use of natural logarithms to represent the values makes a linear relationship possible, hence easy to apply the principles of regression analysis. Moreover, the fact that the points are close together indicates a strong relationship between carats and prices. This indicates that the transformation has managed to offset the effects of those diamonds that may be highly priced, and thus draws attention to the ever-escalating prices of those that are large by a small carat difference.

Call: lm(formula = ln_price ~ ln_carat, data = diamonds_data)

Residuals: Min 1Q Median 3Q Max -1.05427 -0.16849 -0.00515 0.16216 1.33715

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.444025 0.009919 851.3 <2e-16 ***ln_carat 1.671314 0.014154 118.1 <2e-16*** — Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 0.2604 on 998 degrees of freedom Multiple R-squared: 0.9332, Adjusted R-squared: 0.9331 F-statistic: 1.394e+04 on 1 and 998 DF, p-value: < 2.2e-16

Estimated slope (ln_carat): 1.6713 Interpretation: A 1% increase in carat is associated with approximately 1.6713 % increase in price (elasticity). P-value for ln(carat): 0 Conclusion: ln(carat) is statistically significant at the 5% level. R-squared: 0.9332

A regression analysis has been performed on how carat weight affects diamond prices. Using natural logarithmic regression analysis, the effect of ln(price) on ln(carat) is that the slope is approximately 1.6713. This means that if carat weight rises by 1%, it will lead to a rise of approximately 1.67%. This implies that price is an elastic concept, which is expected to rise by a little higher amount as carat weight rises. Moreover, from this analysis, it can be seen that prices tend to be dependent on its weight to a greater extent.

The significance of ln(carat), indicated by the p-value, is less than 0.05. The R-square of .933 implies that ln(carat) accounts exactly for 93.3% of ln(price).

Call: lm(formula = ln_price ~ ln_carat + cut, data = diamonds_data)

Residuals: Min 1Q Median 3Q Max -0.83267 -0.16614 -0.00949 0.14405 1.34453

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.07837 0.04693 172.133 < 2e-16 ***ln_carat 1.69636 0.01373 123.515 < 2e-16*** cutGood 0.25184 0.05405 4.659 3.61e-06 ***cutIdeal 0.43282 0.04920 8.796 < 2e-16*** cutPremium 0.36265 0.04963 7.307 5.60e-13 ***cutVery Good 0.38843 0.05000 7.768 1.98e-14*** — Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ” 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 0.2483 on 994 degrees of freedom Multiple R-squared: 0.9395, Adjusted R-squared: 0.9392 F-statistic: 3088 on 5 and 994 DF, p-value: < 2.2e-16

Estimated Regression Model: (Intercept) ln_carat cutGood cutIdeal cutPremium cutVery Good 8.0784 1.6964 0.2518 0.4328 0.3626 0.3884

In the multiple regression equation, both ln(carat) and cut are significant. The equation is: ln(price) = 8.0784 + 1.6964 · ln(carat) + cut effects (relative to reference category

The coefficient for ln(carat) changes to 1.6964, indicating carat remains an increasing factor for the price even after adjusting for the variations in cut quality. The coefficients for cut are positive, indicating that the prices for the corresponding cuts are higher than the cutoff. For instance, the Ideal cut raises ln(price) by 0.4328, meaning the cut has higher prices than the baseline. The R-squared for the model rises to 0.940, meaning that 94% of the variations in ln(price) are explained when carat and cut are used. This indicates that the addition of cut slightly enhances the model, but carat still remains the dominant feature.

```
            Estimate Std. Error    t value     Pr(>|t|)
```

(Intercept) 8.0783706 0.04693106 172.132698 0.000000e+00 ln_carat 1.6963631 0.01373405 123.515157 0.000000e+00 cutGood 0.2518353 0.05405240 4.659096 3.607662e-06 cutIdeal 0.4328222 0.04920468 8.796362 6.141554e-18 cutPremium 0.3626473 0.04962987 7.307037 5.600821e-13 cutVery Good 0.3884296 0.05000388 7.767988 1.980202e-14 Significant predictors at 1% level: [1] "(Intercept)" "ln_carat" "cutGood" "cutIdeal" "cutPremium"
[6] "cutVery Good"

In multiple regression analysis, the significance of variables with p-values less than or equal to 0.01 is considered to be high. ln(carat), Good, Ideal, Premium, and Very Good are variables with p-values close to 0.000, making them highly significant variables. The most important variables are ln(carat) and quality, with ln(carat) being particularly important because of the high t-value of 123.5.

Hypothesis Test for ln(carat): H0: _ln(carat) = 0 (ln(carat) is NOT a significant predictor) H1: _ln(carat) 0 (ln(carat) IS a significant predictor) P-value for ln(carat) in multiple model: 0 Decision: Reject H0 at 5% level. Conclusion: ln(carat) is a significant predictor of ln(price).

(e)To determine if ln(carat) is a significant predictor of the multiple regression model, the appropriate test is a t-test on the regression coefficient for ln(carat). The hypotheses are: H : (ln_carat) = 0 (carat has no effect on ln(price) after controlling for cut) H : (ln_carat) 0 (carat does affect ln(price) even after controlling for cut) This test focuses on whether the slope associated with ln(carat) is significantly different from zero. If the p-value of ln(carat) is less than 0.05 or 0.01 then the null hypothesis is rejected. That is carat provides useful explanatory power and it should not be removed from the regression model. It is the correct test because regression coefficients are evaluated individually using t-tests.

(f)With the regression result of model (c), the p-value of ln(carat), which is 0.000, is much less than 0.05 as well as 0.01. This offers very strong evidence against the null hypothesis. Thus, the result of hypothesis testing is to reject the H : (ln_carat)=0. This indicates ln(carat) is a highly statistically significant predictor for ln(price), controlling for categories of cut. This implies diamond weight has a very significant influence upon price, irrespective of the quality of cut. The confidence interval of ln(carat) is (1.669, 1.723), which does not contain zero. Thus, the result is the same again. Carat is the most important predictor within the regression model. Its deletion will significantly affect the model's capacity for explaining the price difference.