MA334: Project Report

Nallapu Naveen

2026-01-09

1. Data Exploration

(1a) Summary of the Data Set

```
library(knitr)
# Loading the data
data <- read.csv("MA334-AU-7_2501629.csv")
data_pre <- head(data, 2)
kable(data_pre, caption = "First 2 Observations")
```

Table 1: First 2 Observations

carat	cut	color	clarity	depth	table	price	x	y	z
0.53	Premium	D	SI1	61.4	58	1566	5.19	5.23	3.20
0.85	Ideal	G	SI1	62.2	56	3403	6.06	6.09	3.78

Dataset Dimensions

Table 2: Dataset Dimensions

Item	Value
Total Observations	1000
Total Variables	10

Table 2 shows that the dataset contains 1,000 observations and 10 variables.

Variable names and types Table 3 shows all the variables in the dataset and their data types. Some variables contain numbers (such as carat, depth, price, table and x, y, z), and others contain categories (cut, color, and clarity).

(1b) Location and Spread of Variables

Numeric variables This table shows typical values (mean and median) and how spread out the values are (standard deviation and range) for each numeric variable.

Mode for categorical variables

Table 3: Mode for Categorical Variables

Variable	Mode	Frequency
cut	Ideal	394
color	G	215
clarity	SI1	228

Frequency Distribution of Cut

Table 4: Frequency Distribution of Diamond Cut

Cut	Frequency
Fair	31
Good	77
Ideal	394
Premium	274
Very Good	224

Frequency Distribution of Color

Table 5: Frequency Distribution of Diamond Color

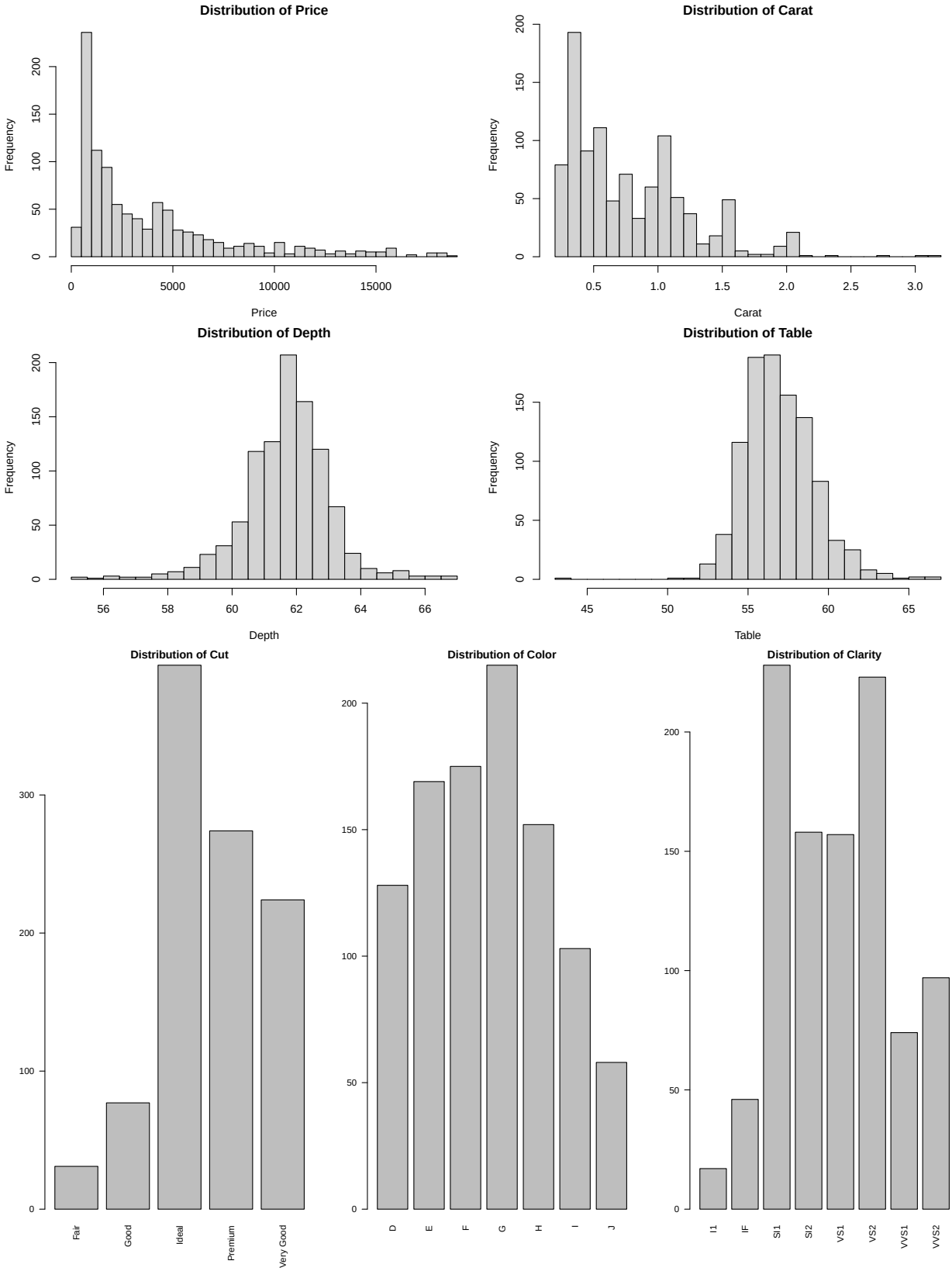
Color	Frequency
D	128
E	169
F	175
G	215
H	152
I	103
J	58

Frequency Distribution of Clarity

Table 6: Frequency Distribution of Diamond Clarity

Clarity	Frequency
I1	17
IF	46
SI1	228
SI2	158
VS1	157
VS2	223
VVS1	74
VVS2	97

(1c) Visualisations of Variable Distributions



(1d) Produce and interpret a correlation matrix between the numeric variables

Table 7: Correlation Matrix of Numeric Variables

	carat	depth	table	price	x	y	z
carat	1.00	0.05	0.21	0.91	0.96	0.96	0.96
depth	0.05	1.00	-0.31	0.00	-0.03	-0.03	0.09
table	0.21	-0.31	1.00	0.15	0.23	0.22	0.18
price	0.91	0.00	0.15	1.00	0.85	0.85	0.84
x	0.96	-0.03	0.23	0.85	1.00	1.00	0.99
y	0.96	-0.03	0.22	0.85	1.00	1.00	0.99
z	0.96	0.09	0.18	0.84	0.99	0.99	1.00

Price shows a strong positive correlation with carat (r = 0.91), making it the strongest numeric variable associated with price and it also has high correlations with the diamond dimensions x and y (r = 0.85) and z (r = 0.84). However, depth is not correlated with price (r = 0.00) and table is only weakly correlated (r = 0.15). Overall, carat and size-related variables (x, y, z) are the main numeric drivers of price in this dataset.

==========

Question 2: Probability, distributions, and confidence intervals

==========

(a) A diamond is chosen at random

(i) Probability price exceeds $10,000

```
## [1] 0.093
```

About 9.3% of the diamonds in this dataset are priced above $10,000. So if we randomly select one, there's a 9.3% chance it'll be one of those expensive ones.

(ii) Probability diamond is Ideal cut

```
## [1] 0.394
```

About 39.4% of the diamonds are of Ideal cut. Hence, the probability that a randomly chosen diamond is Ideal cut is 0.394.

(b) Sample of 20 diamonds

```
## [1] 0.01834124
```

Working:

Let X be the number of Ideal cut diamonds in 20 diamonds and each diamond can be Ideal or not, so X follows a Binomial distribution:

X ~ Bin(20, 0.394).

We need P(X > 12), so we do 1 - P(X <= 12).

That is: P(X > 12) = 1 - pbinom(12, 20, 0.394) = 0.01834.

Answer: 0.01834 = 1.83%

If we randomly select 20 diamonds, there's less than a 1.83% chance that more than 12 of them will be Ideal cuts.

(c) Confidence intervals for mean carat weight

```
## [1] 0.757763 0.805697
```

```
## [1] 0.7531634 0.8102966
```

We are 90% confident that the true mean carat weight lies between 0.758 and 0.806. We are 95% confident that the true mean carat weight lies between 0.753 and 0.810.

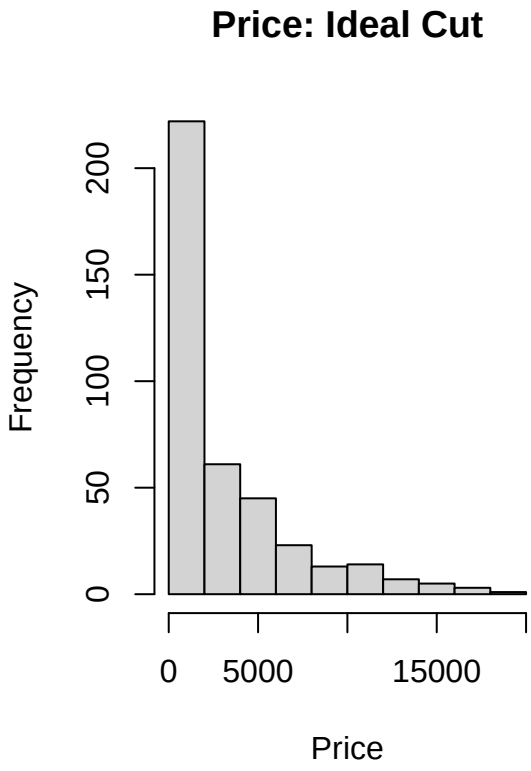
```
## [1] 0.04793403
```

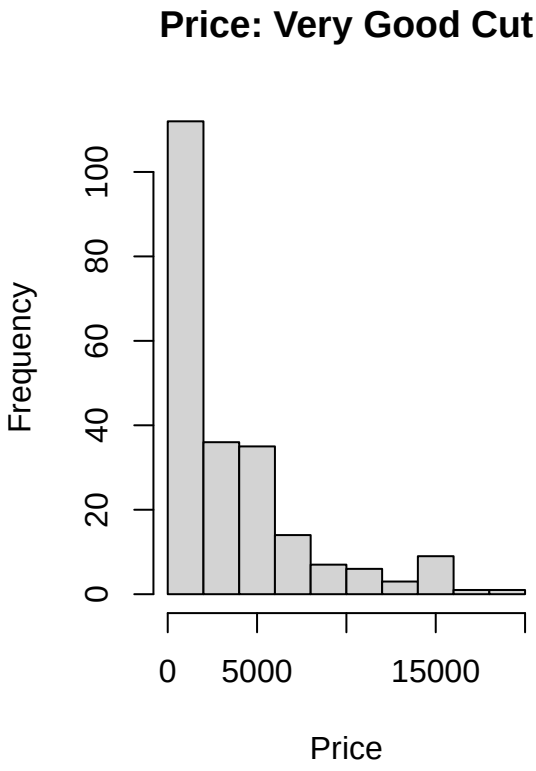
```
## [1] 0.05713317
```

The 95% confidence interval is wider than the 90% interval because higher confidence requires a broader range.

Question 3: Hypothesis Tests

(3a) i–ii: Distributions of price for Ideal and Very Good

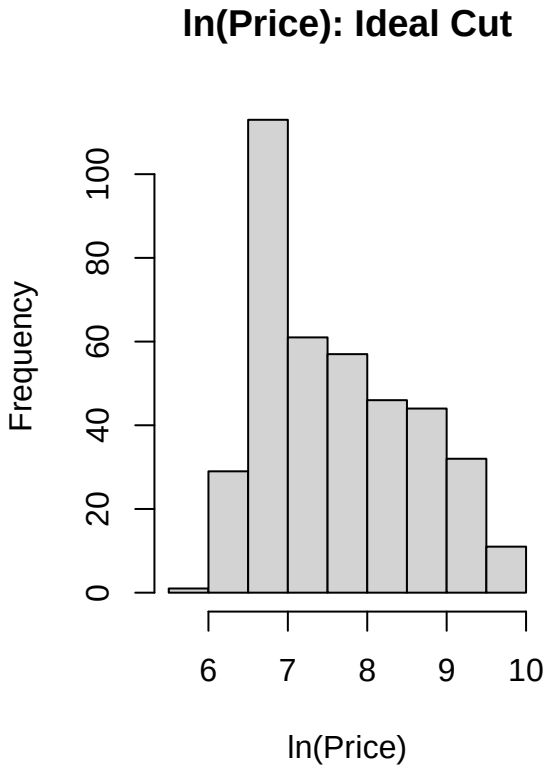


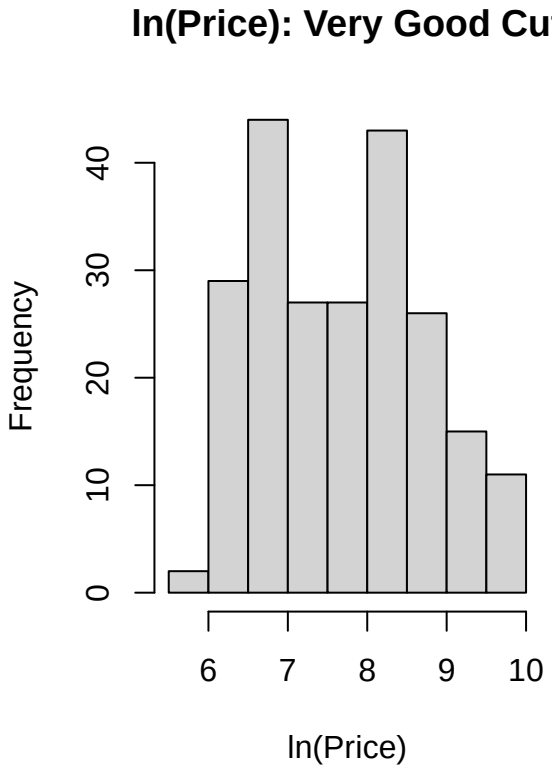


iii) Based on the plots, do you believe they are normally distributed?

The price distributions for both Ideal and Very Good cut diamonds are right-skewed and not symmetric. Therefore, the prices do not appear to be normally distributed.

(3a) iv: Create ln(price) and repeat plots





After applying ln(price), both plots look more balanced and less skewed. So, the ln(price) values appear more normally distributed than the original prices for both cuts.

(3b) t-test on price (5% level)

H0: mean price (Ideal) = mean price (Very Good)

H1: mean price (Ideal) != mean price (Very Good)

```
##
##  Welch Two Sample t-test
##
## data:  ideal_price and vg_price
## t = -1.2234, df = 431.73, p-value = 0.2219
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1015.3727   236.2943
## sample estimates:
## mean of x mean of y
##  3275.563  3665.103
```

We test whether the average price of Ideal cut diamonds is different from that of Very Good cut diamonds. The p-value is 0.2219, which is greater than 0.05, so we do not reject the null hypothesis. This means there

is no significant difference in the mean prices of Ideal and Very Good cut diamonds at the 5% significance level.

(3c) t-test on ln(price) (5% level)

H0: mean ln(price) (Ideal) = mean ln(price) (Very Good)

H1: mean ln(price) (Ideal) != mean ln(price) (Very Good)

```
##
##  Welch Two Sample t-test
##
## data:  ideal_ln and vg_ln
## t = -0.92072, df = 432.88, p-value = 0.3577
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.24054911  0.08707367
## sample estimates:
## mean of x mean of y
##  7.620497  7.697235
```

We checked if Ideal and Very Good diamonds cost different after using ln(price) and the test says they cost almost the same, so there is no real difference.

(3d) Did you obtain the same conclusion from (b) and (c)? Explain why.

```
t_price$p.value
```

```
## [1] 0.2218553
```

```
t_ln$p.value
```

```
## [1] 0.3577081
```

Yes, both tests give the same conclusion because both p-values are greater than 0.05, so we fail to reject the null hypothesis and taking ln(price) reduces skewness but does not change the mean difference between Ideal and Very Good diamonds enough to become statistically significant.

(3e) Chi-square independence test: price group (>5000) vs clarity

Hypotheses: - **H (Null):** Price group (high/low) and clarity are independent - **H (Alternative):** Price group and clarity are dependent

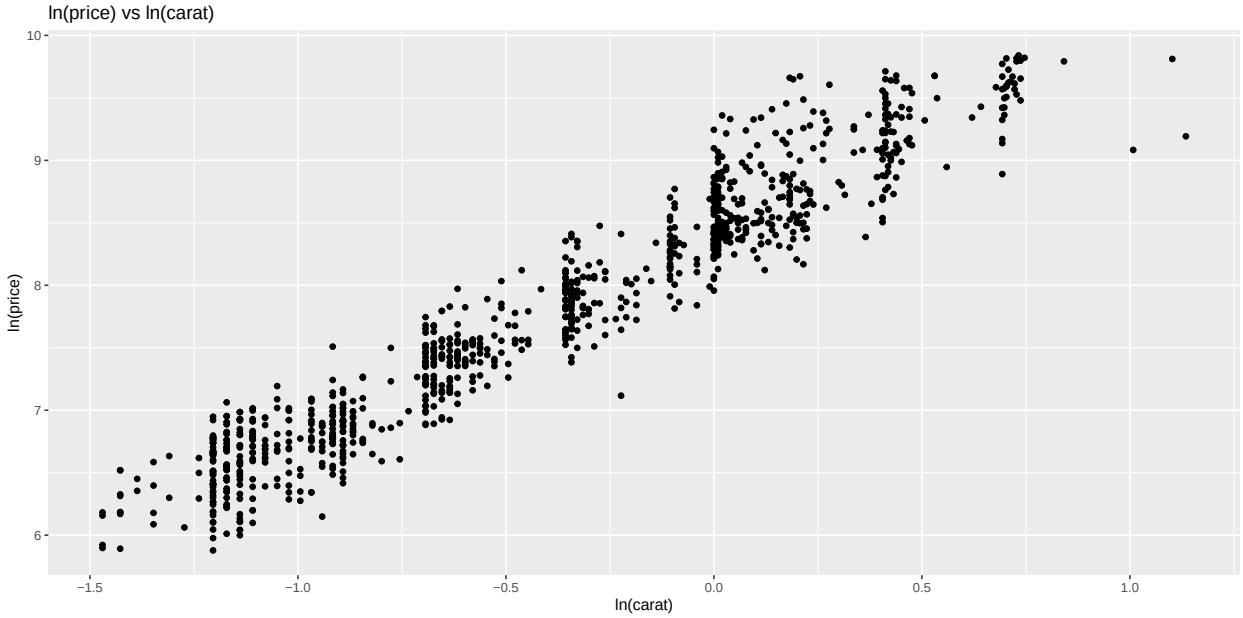
```
##
##         I1  IF SI1 SI2 VS1 VS2 VVS1 VVS2
##   High   6   8  64  37  41  76    7   13
##   Low   11  38 164 121 116 147   67   84
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 29.959, df = 7, p-value = 9.661e-05
```

The p-value is 9.661e-05 which is less than 0.05, so we reject the null hypothesis of independence and this means price group and clarity are associated, so high-price diamonds (>5000) and low-price diamonds (5000) have different clarity distributions.

Question 4: Linear Regression

(4a) Scatter plot of ln(price) vs ln(carat)



The scatter plot shows a clear upward trend. This means that as ln(carat) increases, ln(price) also increases, suggesting a strong positive relationship between carat size and price.

(4b) Simple regression: ln(price) ~ ln(carat)

(4b)(i) slope

```
## ln_carat
## 1.656087
```

The slope is 1.656, which means that a 1% increase in carat size leads to about a 1.66% increase in price on average.

(4b)(ii) p-value for ln(carat)

```
## [1] 0
```

The p-value is 0, which is much smaller than 0.05. This shows that ln(carat) is a statistically significant predictor of ln(price).

(4b)(iii) R-squared

```
## [1] 0.9279249
```

R-squared is 0.9279, so ln(carat) explains about 92.8% of the variation in ln(price) and that's really high, which means the model fits very well and carat is a strong predictor of price.

(4c) Multiple regression: ln(price) ~ ln(carat) + cut

Estimated model: Predicted ln(price) = 8.151 + 1.679 × ln(carat) + 0.168 × (Good) + 0.363 × (Ideal) + 0.307 × (Premium) + 0.286 × (Very Good)

Note: The reference category for cut is Fair, so all cut coefficients are compared to Fair.

```
##
## Call:
## lm(formula = ln_price ~ ln_carat + cut, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.86330 -0.15805 -0.01857  0.16553  0.89623
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.15080    0.04588 177.644  < 2e-16 ***
## ln_carat      1.67895    0.01448 115.937  < 2e-16 ***
## cutGood       0.16828    0.05480   3.071  0.00219 **
## cutIdeal      0.36339    0.04851   7.491 1.51e-13 ***
## cutPremium    0.30738    0.04869   6.313 4.12e-10 ***
## cutVery Good  0.28626    0.04956   5.776 1.02e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2553 on 994 degrees of freedom
## Multiple R-squared:  0.9336, Adjusted R-squared:  0.9333
## F-statistic:  2797 on 5 and 994 DF,  p-value: < 2.2e-16
```

After adding cut to the model, ln(carat) is still highly significant. All cut categories are also significant, showing that both carat size and cut quality affect diamond prices.

(4d) Significant variables at 1% level

```
## [1] "(Intercept)"  "ln_carat"     "cutGood"      "cutIdeal"     "cutPremium"
## [6] "cutVery Good"
```

At the 1% significance level, ln(carat) and all cut categories are significant. This means these variables have a strong and reliable effect on ln(price).

(4e) Hypothesis test for ln(carat) in model (c)

H0: beta_ln_carat = 0

H1: beta_ln_carat != 0

```
##     Estimate   Std. Error      t value     Pr(>|t|)
##   1.67895120   0.01448158 115.93701166   0.00000000
```

The t-statistic is very large (115.94) and the p-value is 0, confirming strong evidence that ln(carat) has a significant positive effect on ln(price).