MA334 Final Project

2026-01-06

```
## ======================================================
```

```
## 1. Data Exploration
```

```
## ======================================================
```

Table 1: Dataset Overview

Item	Value
Number of observations	1000
Number of variables	12
Numeric variables	carat, depth, table, price, x, y, z
Categorical variables	cut, color, clarity
Variable of interest	price

Table 2: Categories of Qualitative Variables

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

Table 3: First 6 Observations of the Dataset

carat	cut	color	clarity	depth	table	price	x	y	z	ln_price	ln_carat
0.71	Good	D	SI2	64.3	56	2215	5.64	5.59	3.61	7.703008	-0.3424903
0.51	Ideal	D	VS2	61.4	57	1716	5.13	5.16	3.16	7.447751	-0.6733446
1.00	Very Good	D	SI2	62.5	58	4661	6.35	6.44	4.00	8.446985	0.0000000
0.70	Premium	H	VS1	62.8	60	2367	5.61	5.56	3.51	7.769379	-0.3566749
0.23	Premium	I	VVS1	60.5	61	414	3.98	3.95	2.40	6.025866	-1.4696760
1.16	Ideal	G	SI2	61.9	56	4969	6.77	6.73	4.18	8.510974	0.1484200

```
##
## Interpretation (1a):
```

```
## The dataset has 1000 diamonds. Variables include size measures (carat, x, y, z) and quality measures

## Price is the main outcome variable we want to explain.
```

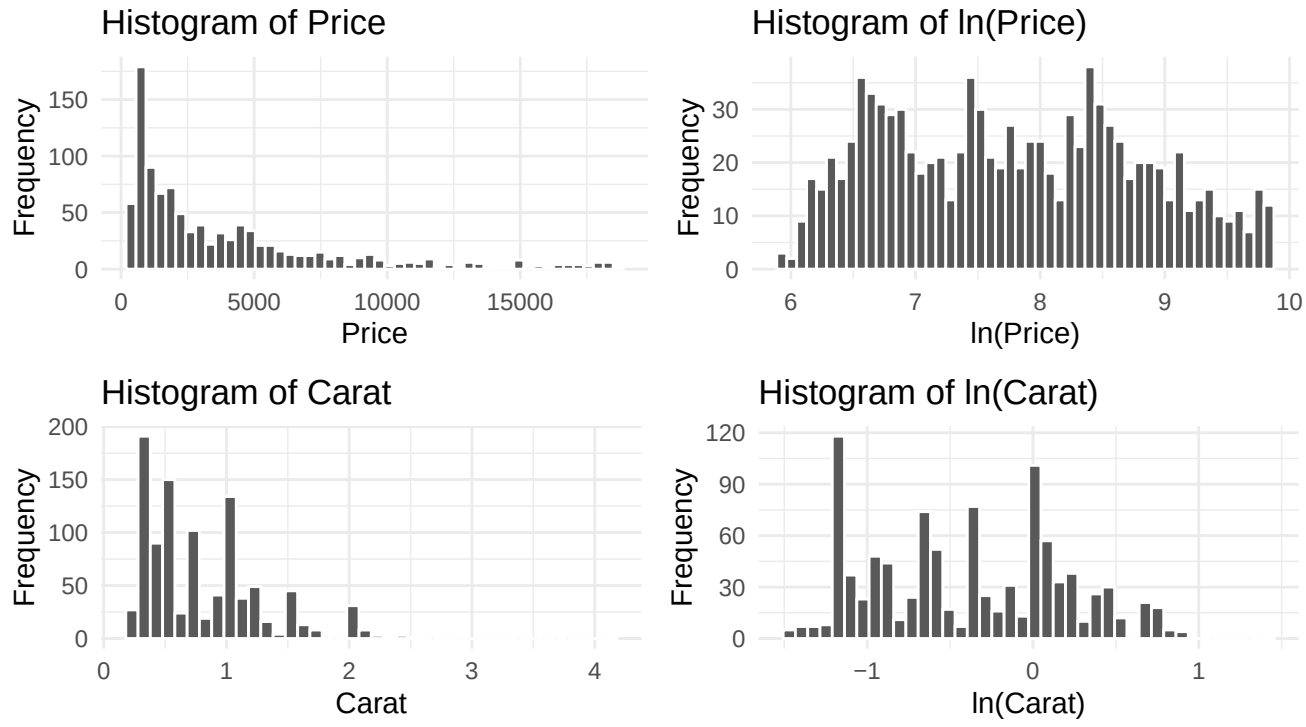
Table 4: Summary Statistics for Numeric Variables

Variable	Mean	Median	SD	Min	Max	Range
carat	0.802	0.700	0.487	0.23	4.13	3.90
depth	61.737	61.800	1.453	53.80	69.00	15.20
price	3951.520	2320.500	4092.059	375.00	18680.00	18305.00
table	57.583	57.000	2.144	52.00	66.00	14.00
x	5.738	5.640	1.124	3.92	10.00	6.08
y	5.741	5.645	1.116	3.95	9.85	5.90
z	3.543	3.490	0.695	2.39	6.43	4.04

```
##
## Modes (most common): cut = Ideal , color = G , clarity = SI1

##
## Interpretation (1b):

## Price has very large spread and is much more variable than depth/table. Carat and price have means g
```




```
##
## Interpretation (1c):
```

Price and carat are strongly right-skewed, with a few very large values. The log plots look more symm

Table 5: Correlation Matrix (Numeric Variables)

	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

```
##
## Correlation with price (highest to lowest):

## price carat     y     x     z table depth
##  1.00  0.92  0.89  0.89  0.88  0.12 -0.01

##
## Interpretation (1d):

## Price is most strongly correlated with carat and the size variables x, y, z. Depth and table have wea

## ======================================================

## 2. Probability, probability distributions and confidence intervals

## ======================================================

## (a)(i) P(price > 10000) = 0.093

## (a)(ii) P(cut = Ideal) = 0.377

## Interpretation (2a):

## Only a small proportion of diamonds cost above $10,000. Around one-third of diamonds are Ideal cut.

## (b) Let X ~ Bin(20, p) with p = 0.377

## P(X > 12) = 1 - P(X <= 12) = 0.01221

## Interpretation (2b):

## The chance of getting more than 12 Ideal diamonds out of 20 is very low (around 1%).
```

Table 6: Confidence Intervals for Mean Carat

Level	Lower	Upper	Width
90%	0.77661	0.82729	0.05068
95%	0.77174	0.83216	0.06041

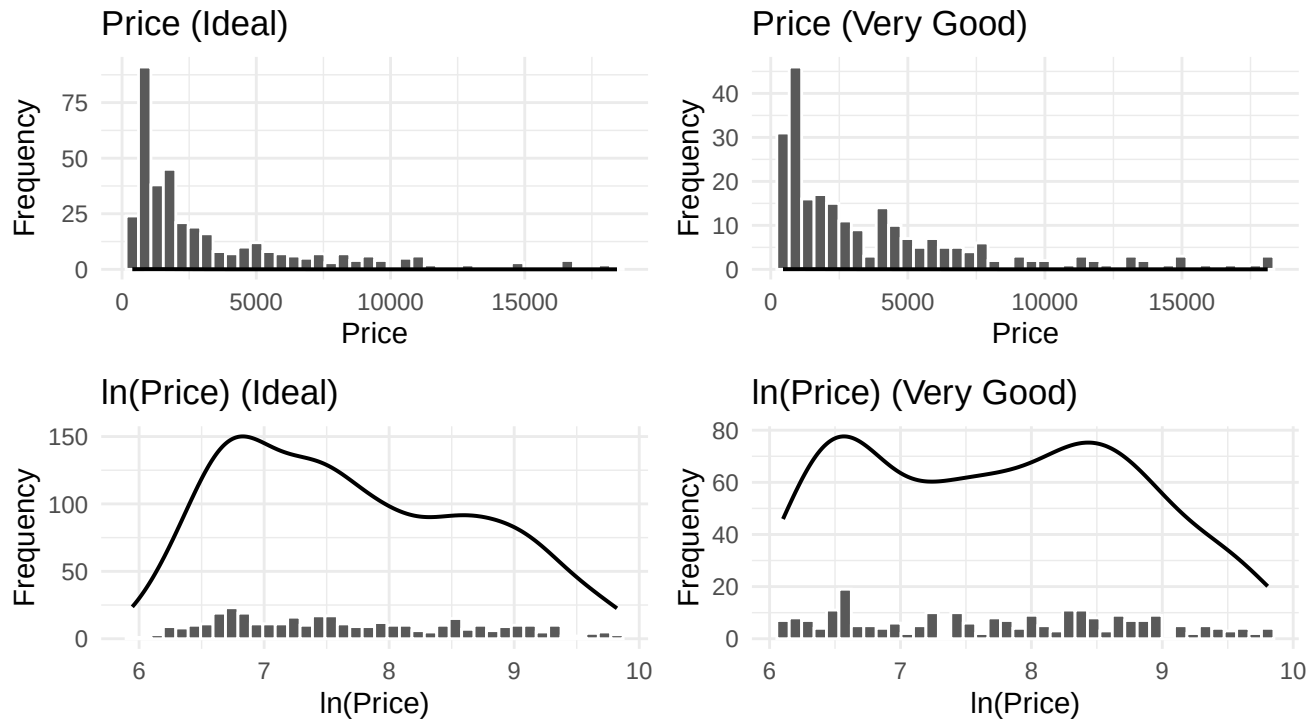
```
##
## Interpretation (2c):

## The 95% interval is wider than the 90% interval because higher confidence needs a larger margin of e

## ======================================================

## 3. Hypothesis Tests

## ======================================================
```




```
##
## Interpretation (3a):

## Raw price is right-skewed for both cuts, so it does not look normal. ln(price) is more symmetric, so

## (b) t-test (Price), alpha=0.05
## H0: mu_Ideal = mu_VeryGood
## H1: mu_Ideal != mu_VeryGood
```

Table 7: t-test Results (Price)

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-363.9567	3552.233	3916.19	-1.11781	0.2642	487.7853	-1003.704	275.7905	Welch Two Sample t-test	two.sided

```
##
## Interpretation (3b):

## The p-value is greater than 0.05, so we fail to reject H0. There is no evidence that mean price diff

## (c) t-test (ln(price)), alpha=0.05
## H0: mu_ln(Ideal) = mu_ln(VeryGood)
## H1: mu_ln(Ideal) != mu_ln(VeryGood)
```

Table 8: t-test Results (ln(price))

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-0.06023	7.69499	7.75522	-0.71731	0.47353	483.9716	-0.22521	0.10475	Welch Two Sample t-test	two.sided

```
##
## Interpretation (3c):

## The p-value is greater than 0.05, so we also fail to reject H0 for ln(price). The average log-price

## (d) Comparison

## Both tests give the same conclusion (no significant difference). Using ln(price) is still preferred

## (e) Chi-square test (Price level vs Clarity), alpha=0.05
## H0: independent
## H1: not independent
```

Table 9: Contingency Table: Price Level 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

Table 10: Chi-square Test Results

statistic	p.value	parameter	method
27.24964	3e-04	7	Pearson's Chi-squared test

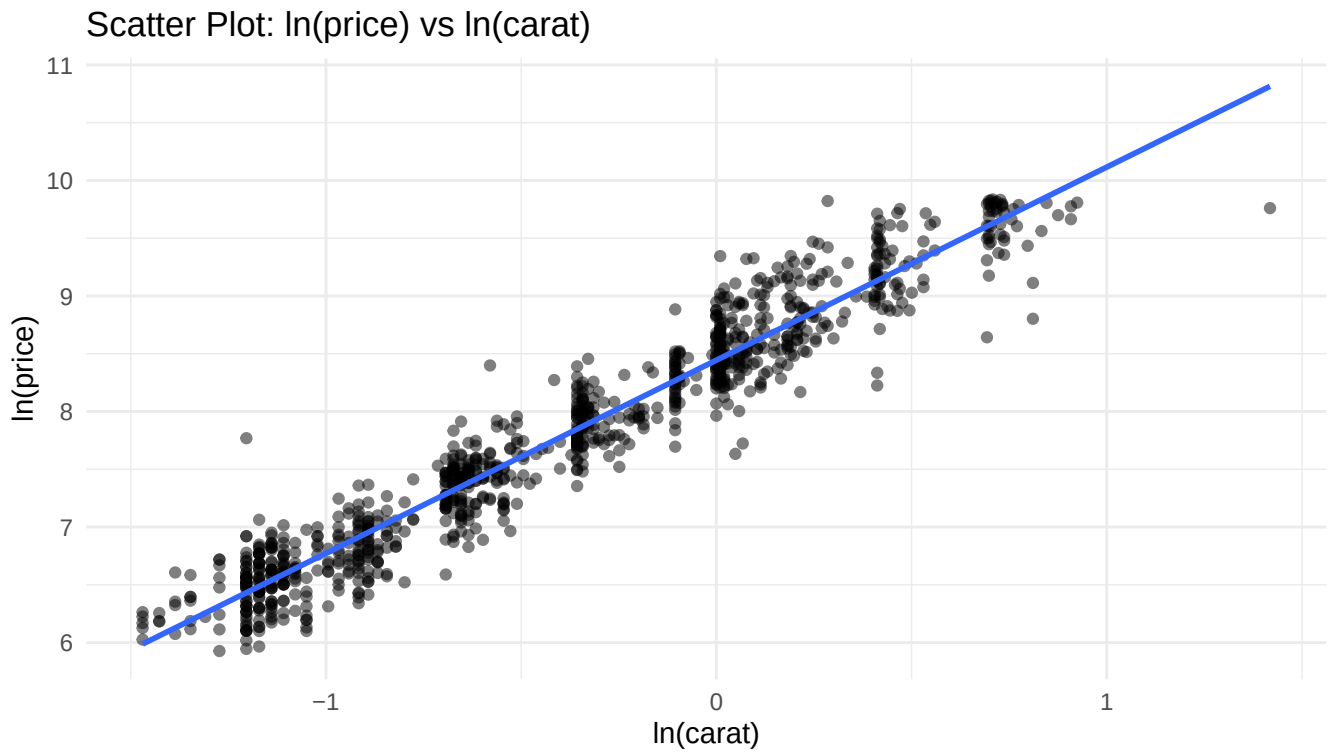
```
##
## Interpretation (3e):

## The p-value is below 0.05, so we reject H0. Clarity and price level are associated in this dataset.

## ========================================================

## 4. Linear Regression

## ========================================================
```




```
##
## Interpretation (4a):

## The plot shows a clear positive linear trend: higher carat diamonds tend to have higher prices.

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

## (i) Slope = 1.67131

## (ii) p-value = < 2.22e-16

## (iii) R-squared = 0.9332

## Interpretation (4b):
```

```
## Because this is a log-log model, the slope is elasticity: a 1% increase in carat is linked to about
```

```
## The p-value is extremely small, so carat is a strong predictor, and R-squared shows the model explai
```

Table 11: Multiple Regression Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	8.07837	0.04693	172.13270	0
ln_carat	1.69636	0.01373	123.51516	0
cutGood	0.25184	0.05405	4.65910	0
cutIdeal	0.43282	0.04920	8.79636	0
cutPremium	0.36265	0.04963	7.30704	0
cutVery Good	0.38843	0.05000	7.76799	0

Table 12: Multiple Regression Model Summary

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.93952	0.93922	0.24829	3088.282	0	5	-22.78791	59.57582	93.9301	61.28003	994	1000

```
##
## (c) Estimated model (baseline cut is reference):
```

```
## ln(price) = 8.07837 + 1.69636 * ln(carat) + cut terms
```

```
## Interpretation (4c):
```

```
## After controlling for carat, cut still affects price: better cuts have higher ln(price) compared to
```

Table 13: Significant Terms ($p < 0.01$)

term	estimate	std.error	statistic	p.value
(Intercept)	8.07837	0.04693	172.13270	0
ln_carat	1.69636	0.01373	123.51516	0
cutGood	0.25184	0.05405	4.65910	0
cutIdeal	0.43282	0.04920	8.79636	0
cutPremium	0.36265	0.04963	7.30704	0
cutVery Good	0.38843	0.05000	7.76799	0

```
##
## Interpretation (4d):
```

```
## At the 1% level, ln(carat) and all cut categories listed are statistically significant predictors of
```

```
## (e) Hypothesis test for ln(carat) in multiple regression
```

```
## H0: beta_lncarat = 0
## H1: beta_lncarat != 0
```

Table 14: Test for ln(carat) Coefficient

term	estimate	std.error	statistic	p.value
ln_carat	1.696363	0.013734	123.5152	0

```
##
## Interpretation (4f):

## The p-value is effectively zero, so we reject H0 at 1%. ln(carat) remains highly significant even af
```