DIAMOND PRICE PREDICTION USING LINEAR REGRESSION

Devansh Shah

Jennifer Indrupati

Rithika Mothukuri

Mohanapriya Subramanian

INTRODUCTION

- **Objective:** The goal of the project is to build a model that can accurately predict the price of a diamond on given factors.
- **Dataset Overview:** This dataset explores the factors shaping diamond pricing, encompassing carat weight, clarity, cut quality, and physical dimensions.
- **Price Determinants:** Highlighting these variables crucially impacts diamond market value and perceived quality.
- **Industry Significance:** Understanding how these factors influence pricing is crucial for both consumers and industry professionals, determining the value and desirability of diamonds.
- **Expected Insights:** Anticipate uncovering trends and correlations within the dataset, offering valuable insights for decision-making across the diamond industry.

DATASET

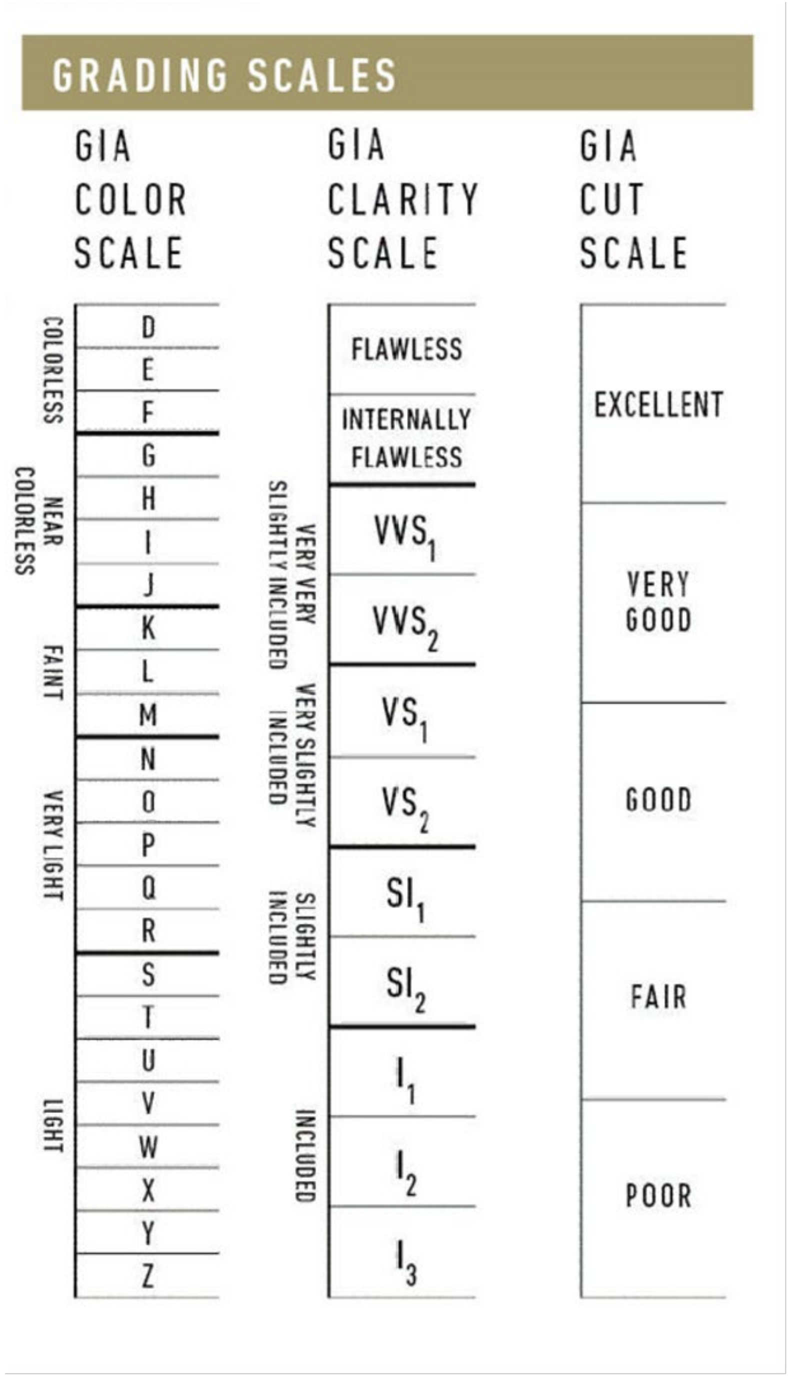
Description: Dataset analyzing diamond prices based on various factors

A	B	C	D	E	F	G	H	I	J	K
	Carat	Cut	Color	Clarity	Depth	Table	Price	X	Y	Z
1	0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
2	0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
3	0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
4	0.29	Premium	I	VS2	62.4	58	334	4.2	4.23	2.63
5	0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
6	0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
7	0.24	Very Good	I	VVS1	62.3	57	336	3.95	3.98	2.47
8	0.26	Very Good	H	SI1	61.9	55	337	4.07	4.11	2.53
9	0.22	Fair	E	VS2	65.1	61	337	3.87	3.78	2.49
10	0.23	Very Good	H	VS1	59.4	61	338	4	4.05	2.39
11	0.3	Good	J	SI1	64	55	339	4.25	4.28	2.73
12	0.23	Ideal	J	VS1	62.8	56	340	3.93	3.9	2.46
13	0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
14	0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
15	0.2	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
16	0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
17	0.3	Ideal	I	SI2	62	54	348	4.31	4.34	2.68
18	0.3	Good	J	SI1	63.4	54	351	4.23	4.29	2.7
19	0.3	Good	J	SI1	63.8	56	351	4.23	4.26	2.71
20	0.3	Very Good	J	SI1	62.7	59	351	4.21	4.27	2.66
21	0.3	Good	I	SI2	63.3	56	351	4.26	4.3	2.71
22	0.23	Very Good	E	VS2	63.8	55	352	3.85	3.92	2.48
23	0.23	Very Good	H	VS1	61	57	353	3.94	3.96	2.41
24	0.31	Very Good	J	SI1	59.4	62	353	4.39	4.43	2.62
25	0.31	Very Good	J	SI1	58.1	62	353	4.44	4.47	2.59

4C'S DETERMINING DIAMOND PRICE

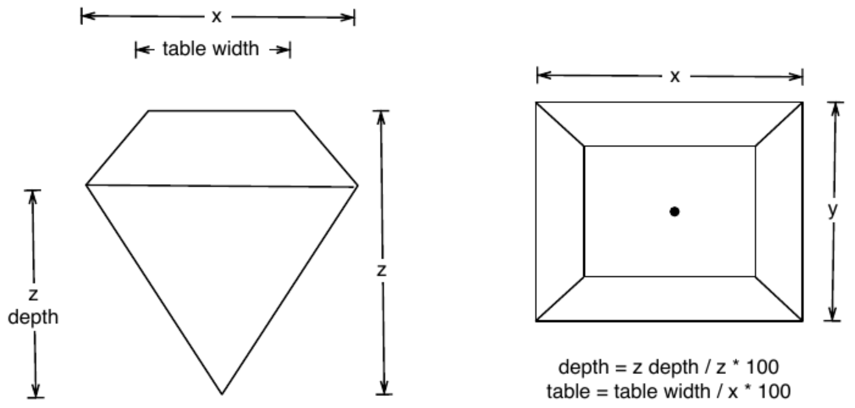
- Carat: Weight of the diamond (1 carat = 0.200 grams)

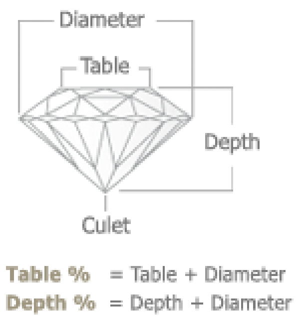
- Clarity: Diamond clarity is a measure of the purity and rarity of the stone. Flaws or inclusions in diamond
- Cut: Quality of the diamond's cut affecting its brilliance and sparkle.
- Color: Color of the diamond (Graded from D to J)



GRADING SCALES

- x length in mm (0--10.74)
- y width in mm (0--58.9)
- z depth in mm (0--31.8)
- depth total depth percentage = z / mean(x, y)
  = 2 * z / (x + y) (43--79)
- table width of top of diamond relative to widest point (43--95)





	Depth %	Table %
Excellent	59.0% - 61.0%	53% - 60%
Very Good	58.0% - 62.0%	61% - 62%
Good	56% - 64%	62% – 64%
Fair	64% - 70%	64% - 66%
Poor	over 70%	over 66% or under 53%

DATA PREPROCESSING

CORRELATION – NUMERICAL PREDICTORS

Carat vs. Price (0.922)

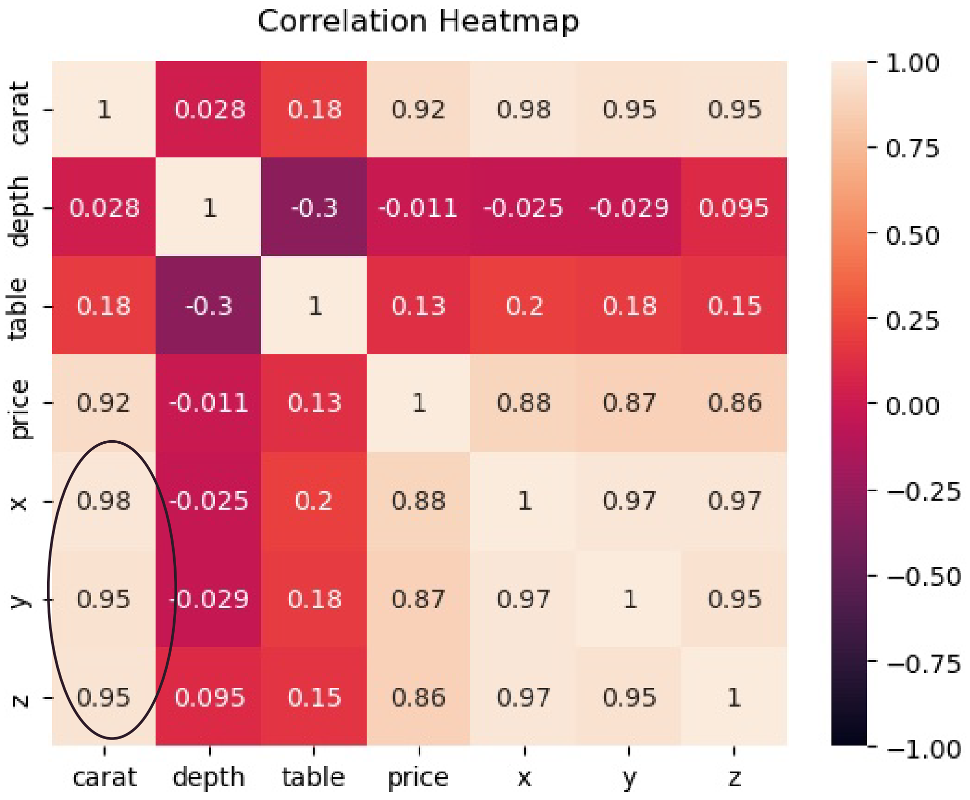
The Carat and the price have a positive correlation

Table vs. Price (0.127)

Weak positive correlation

Depth vs. Price (0.028)

Weak positive correlation



SUMMARY STATISTICS

	carat	depth	table	price	x	y	z
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000
mean	0.797940	61.749405	57.457184	3932.799722	5.731157	5.734526	3.538734
std	0.474011	1.432621	2.234491	3989.439738	1.121761	1.142135	0.705699
min	0.200000	43.000000	43.000000	326.000000	0.000000	0.000000	0.000000
25%	0.400000	61.000000	56.000000	950.000000	4.710000	4.720000	2.910000
50%	0.700000	61.800000	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	1.040000	62.500000	59.000000	5324.250000	6.540000	6.540000	4.040000
max	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

DATA WRANGLING

0

Missing Values

Outliers

Dropped 169 records

x=0, y=0, z=0 and y>30, z>30, table>80

x	y	z
8.09	58.9	8.06
5.12	5.15	31.8
5.15	31.8	5.12

→

x	y	z
8.09	5.89	8.06
5.12	5.15	3.18
5.15	3.18	5.12

Dummy Variables

Cut, Color & Clarity

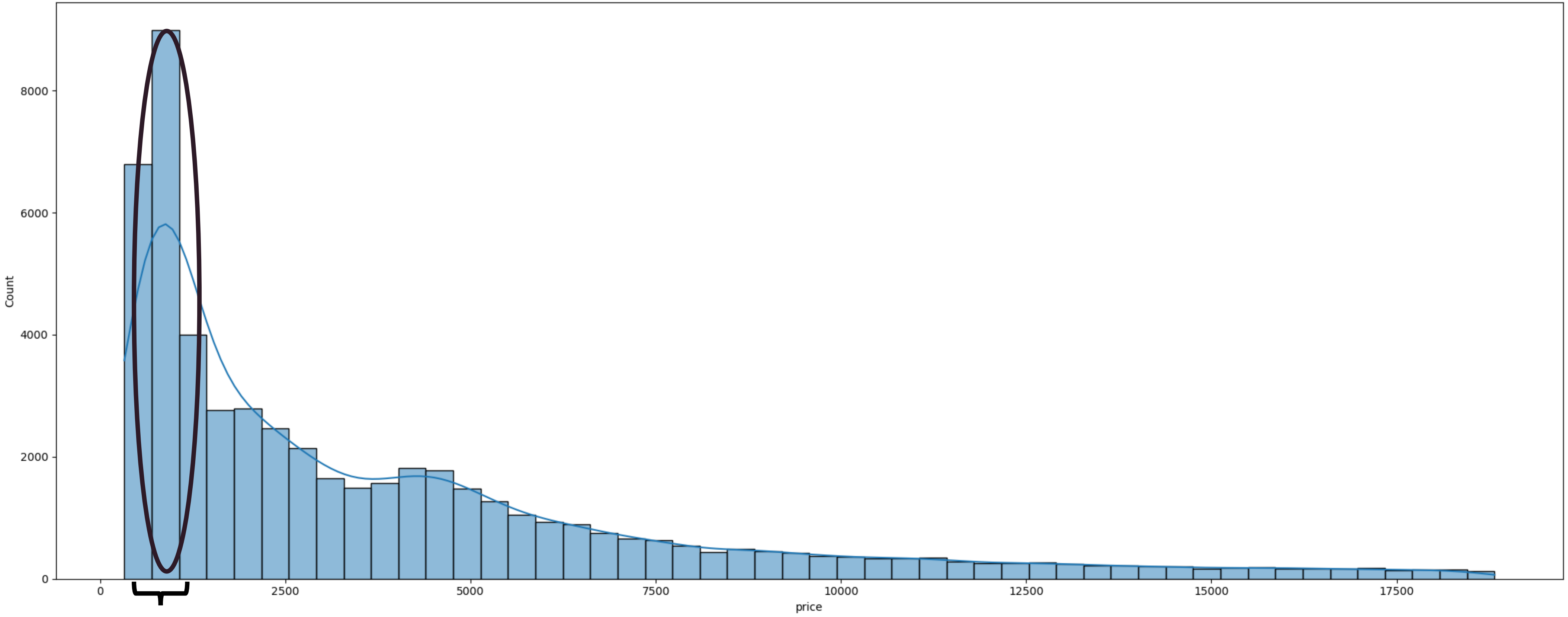
Splitting the Data

70% Train & 30% Test

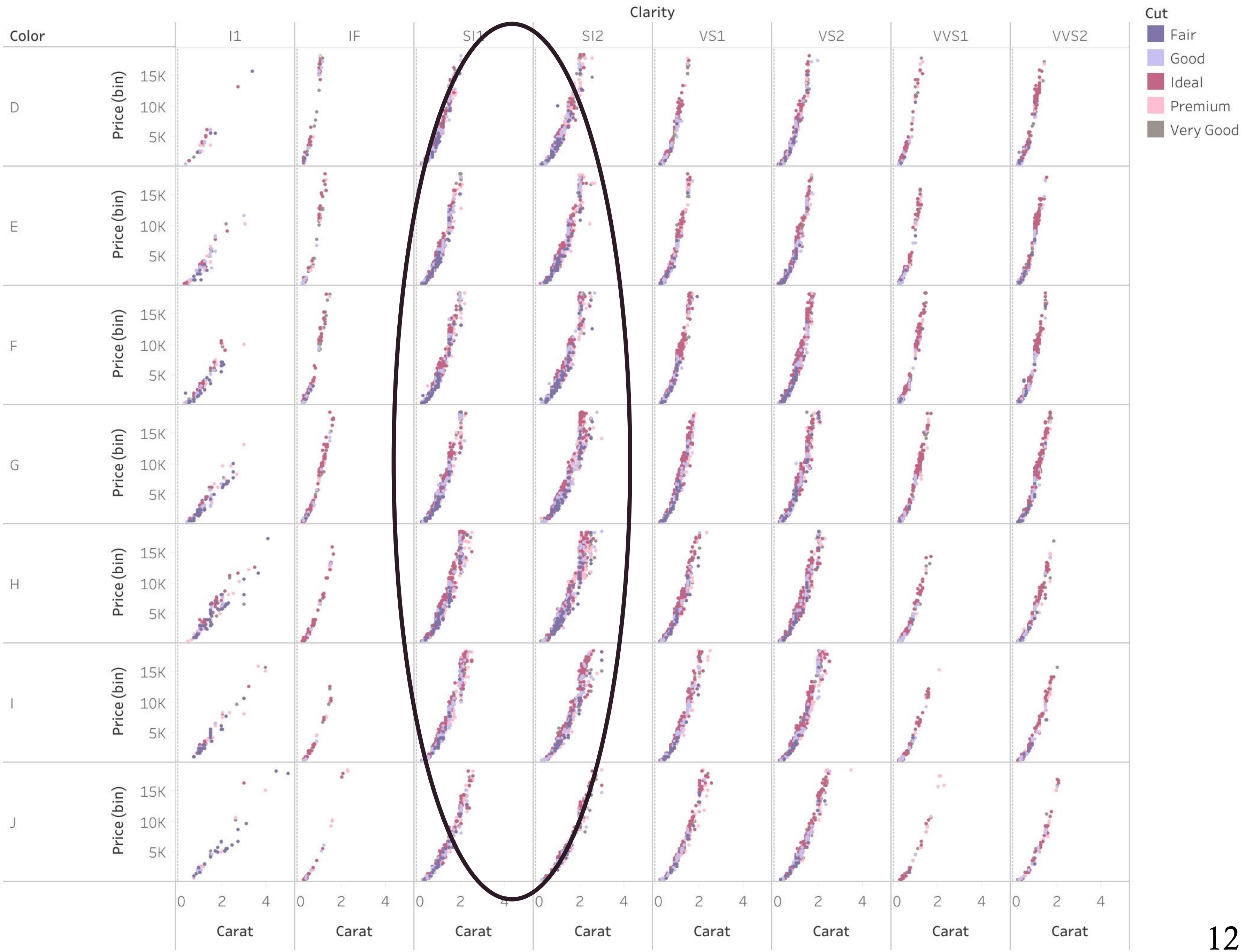
DATA EXPLORATION – EDA INSIGHTS

PRICE DISTRIBUTION

Most of the purchase is between $500 – $1500, and the range drops gradually which is a typical behavior

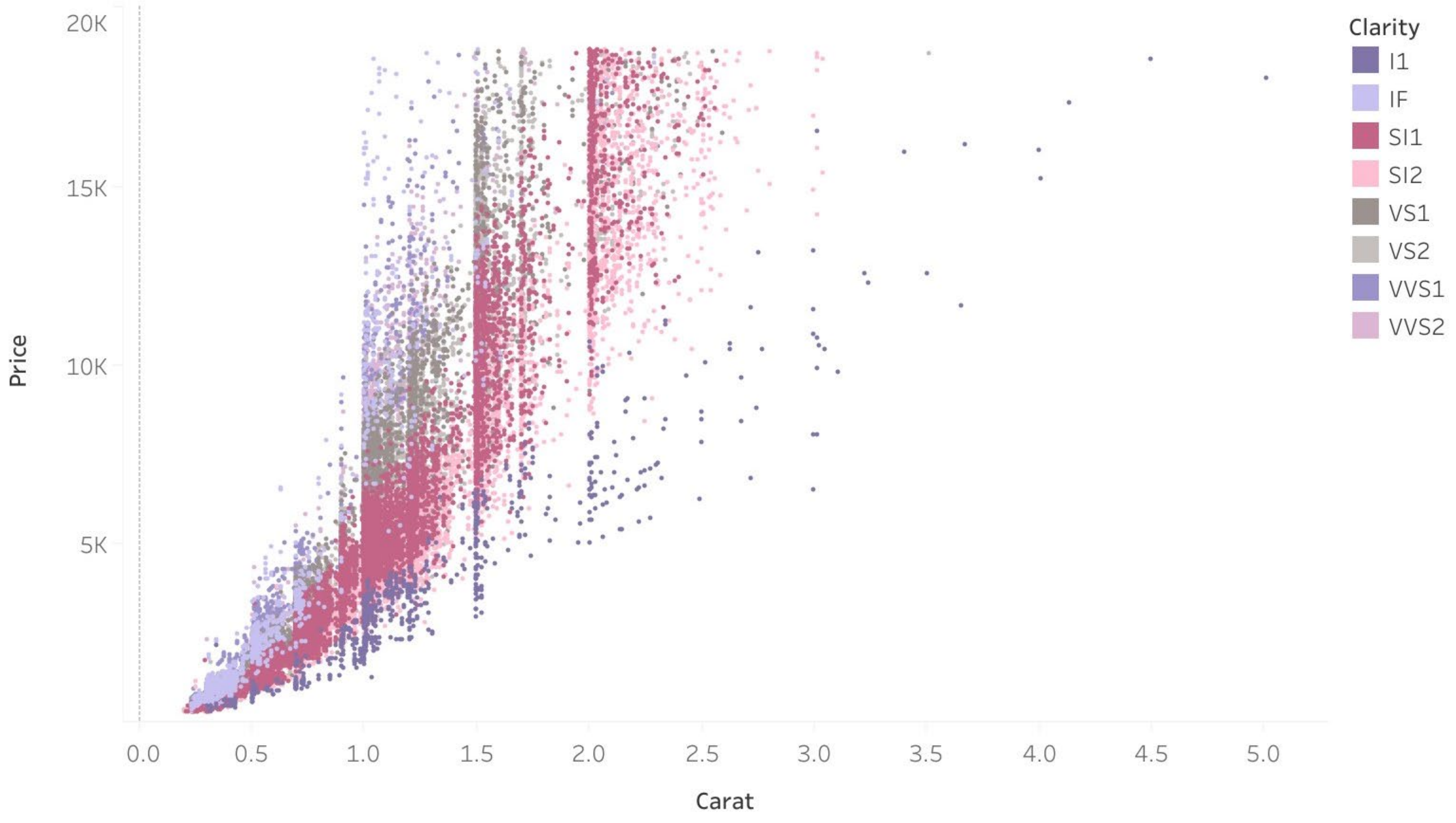


PRICE VS 4C'S



Visualizations are made with Tableau

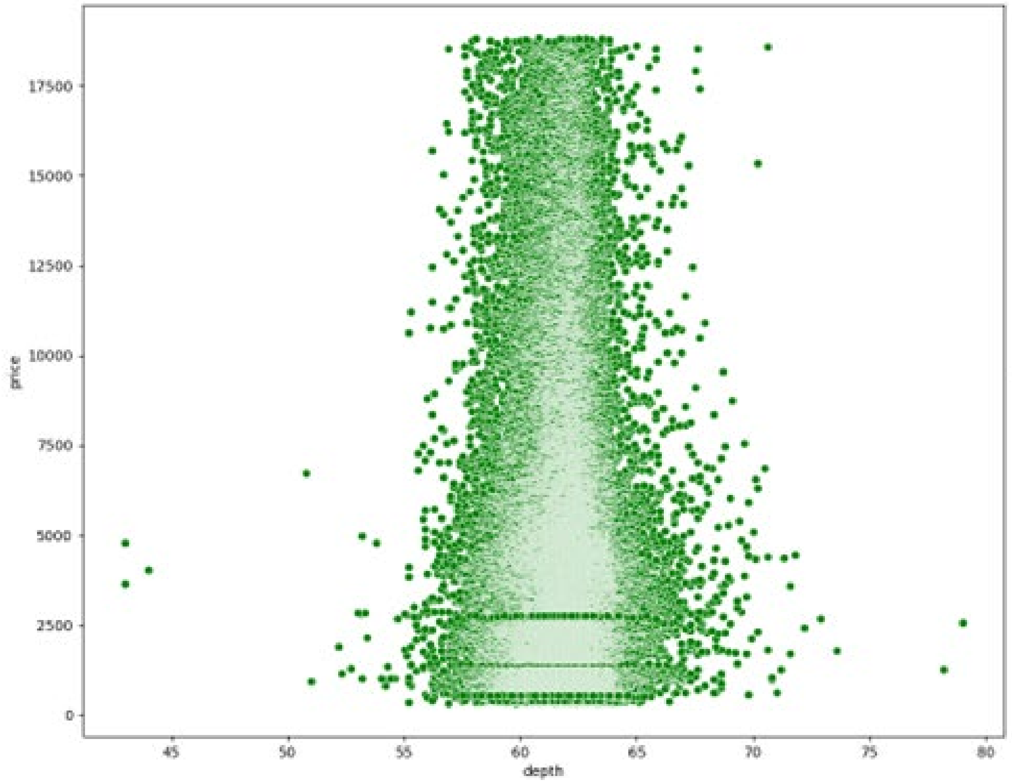
PRICE VS CARAT AND CLARITY

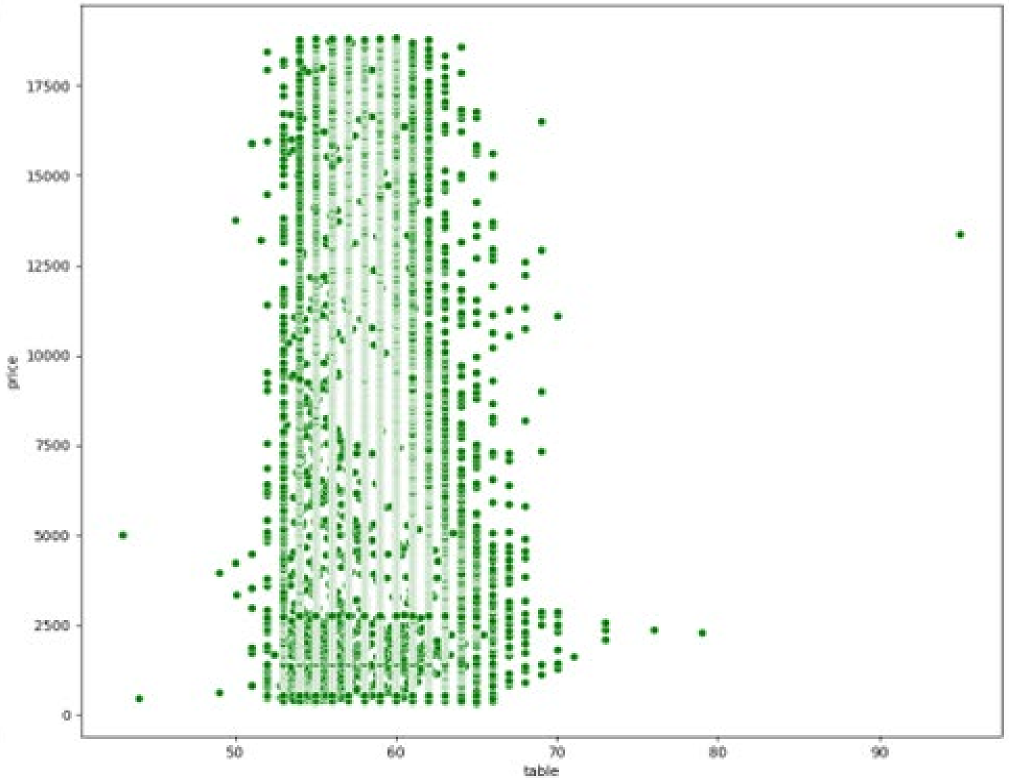


Strong Positive correlation between Price – Carat weight and Clarity

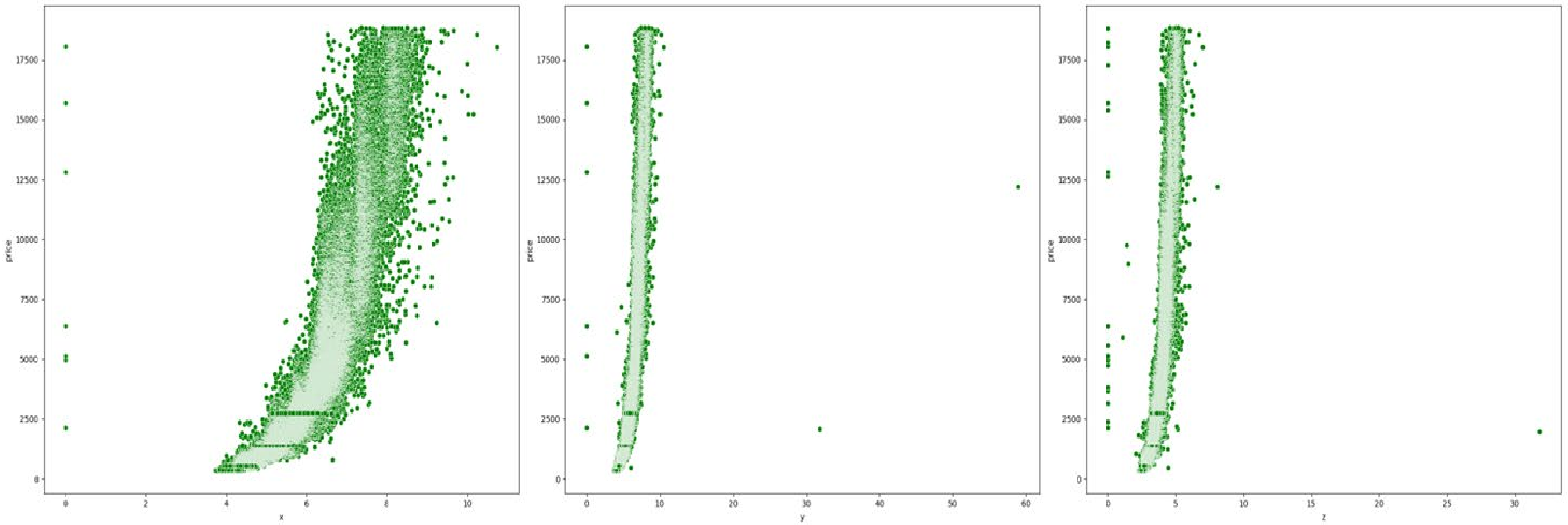
Visualizations are made with Tableau

CORRELATION BETWEEN PRICE – SIZE OF THE DIAMONDS



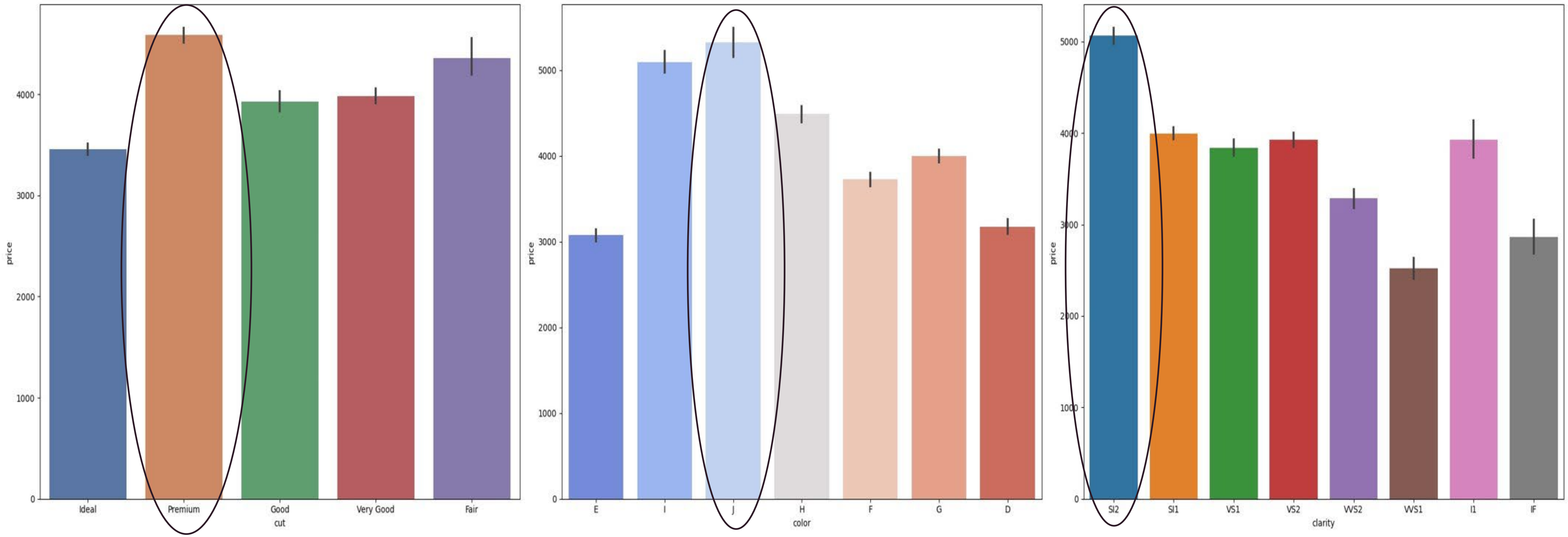


CORRELATION BETWEEN PRICE – SIZE OF THE DIAMONDS CONTD.



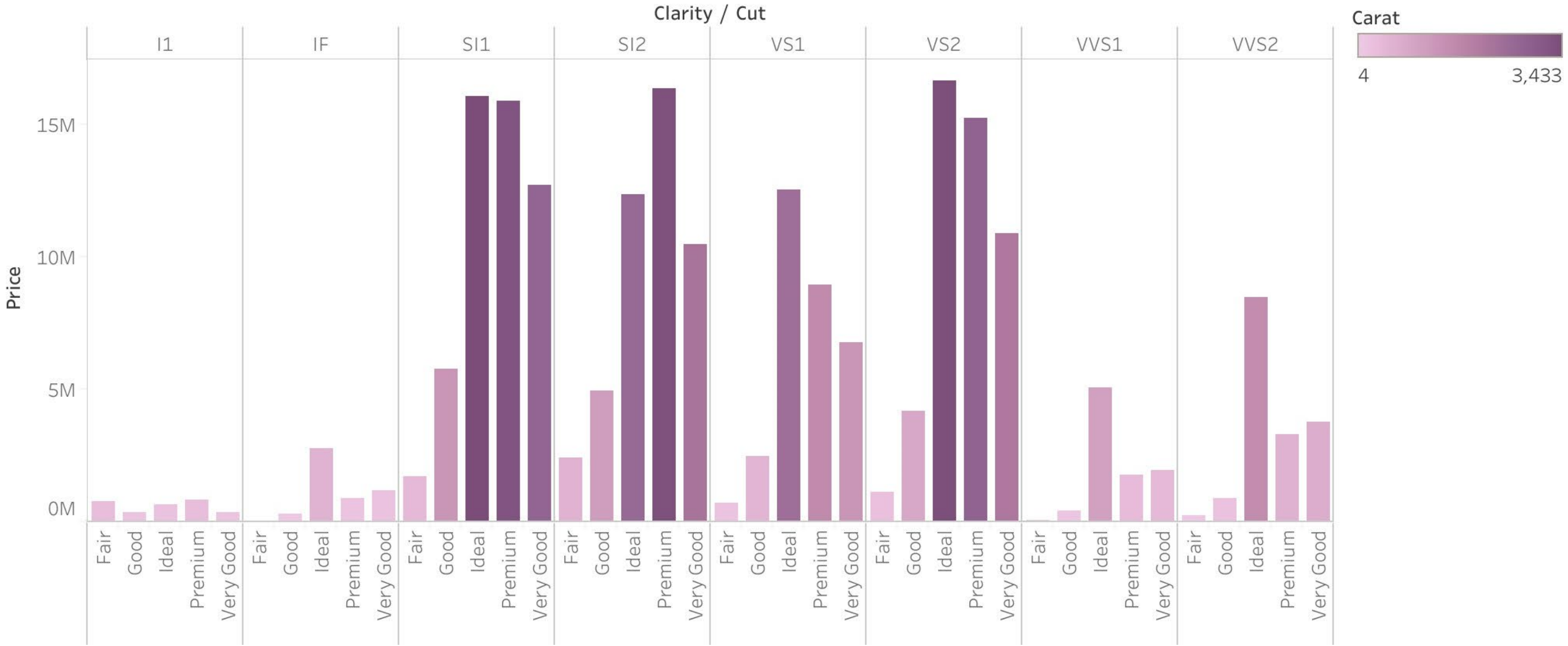
PRICE CORRELATION WITH CUT, COLOR AND CLARITY

- Highest price in terms of cut is Premium
- In terms of color is J,
- In terms of clarity is SI1



PRICE VS CUT AND CLARITY FOR DIFFERENT RANGE OF CARAT WEIGHT

High price diamonds bought are of average clarity but high carat



PREDICTIVE ANALYTICS

LINEAR REGRESSION MODELING – FORWARD SELECTION

- Linear Regression with all predictors after eliminating outliers, dropping empty values
- Split Train and test data in 7:3 ratio
- Predictor y and z has p value more than 0.05, these are less significant
- Perform Linear Regression again removing less significant predictors again
- Multiple R-squared: 91.98%
- Adjusted R-squared: 91.98%

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2184.477	408.197	5.352	8.76E-08
carat	11256.98	48.628	231.494	< 2e-16
cutGood	579.751	33.592	17.259	< 2e-16
cutIdeal	832.912	33.407	24.932	< 2e-16
cutPremium	762.144	32.228	23.649	< 2e-16
cutVery Good	726.783	32.241	22.542	< 2e-16
colorE	-209.118	17.893	-11.687	< 2e-16
colorF	-272.854	18.093	-15.081	< 2e-16
colorG	-482.039	17.716	-27.209	< 2e-16
colorH	-980.267	18.836	-52.043	< 2e-16
colorI	-1466.24	21.162	-69.286	< 2e-16
colorJ	-2369.4	26.131	-90.674	< 2e-16
clarityIF	5345.102	51.024	104.757	< 2e-16
claritySI1	3665.472	43.634	84.005	< 2e-16
claritySI2	2702.586	43.818	61.677	< 2e-16
clarityVS1	4578.398	44.546	102.779	< 2e-16
clarityVS2	4267.224	43.853	97.306	< 2e-16
clarityVVS1	5007.759	47.16	106.187	< 2e-16
clarityVVS2	4950.814	45.855	107.967	< 2e-16
depth	-63.806	4.535	-14.071	< 2e-16
table	-26.474	2.912	-9.092	< 2e-16
x	-1008.26	32.898	-30.648	< 2e-16
y	9.609	19.333	0.497	0.619
z	-50.119	33.486	-1.497	0.134

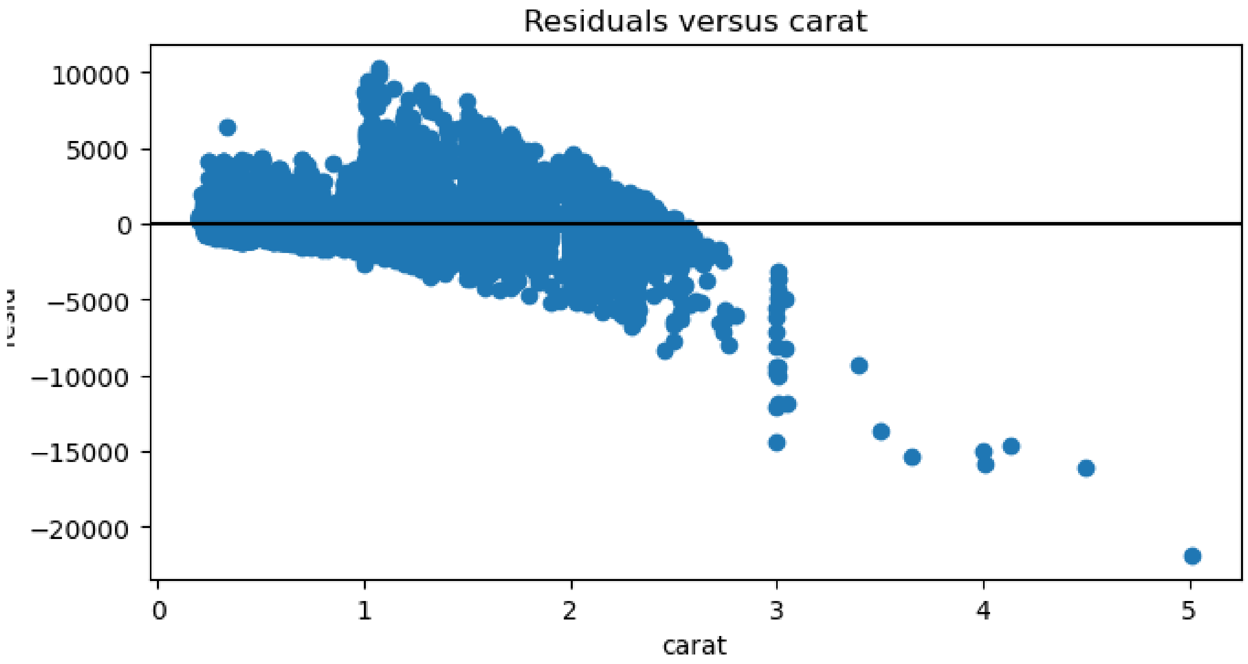
LINEAR REGRESSION MODELING BACKWARD ELIMINATION

- Linear Regression with only significant predictors
- Split Train and test data in 7:3 ratio
- All Predictors have p value less than 0.05
- Multiple R-squared: 92.02%
- Adjusted R-squared: 92.02%
- RMSE : 1130

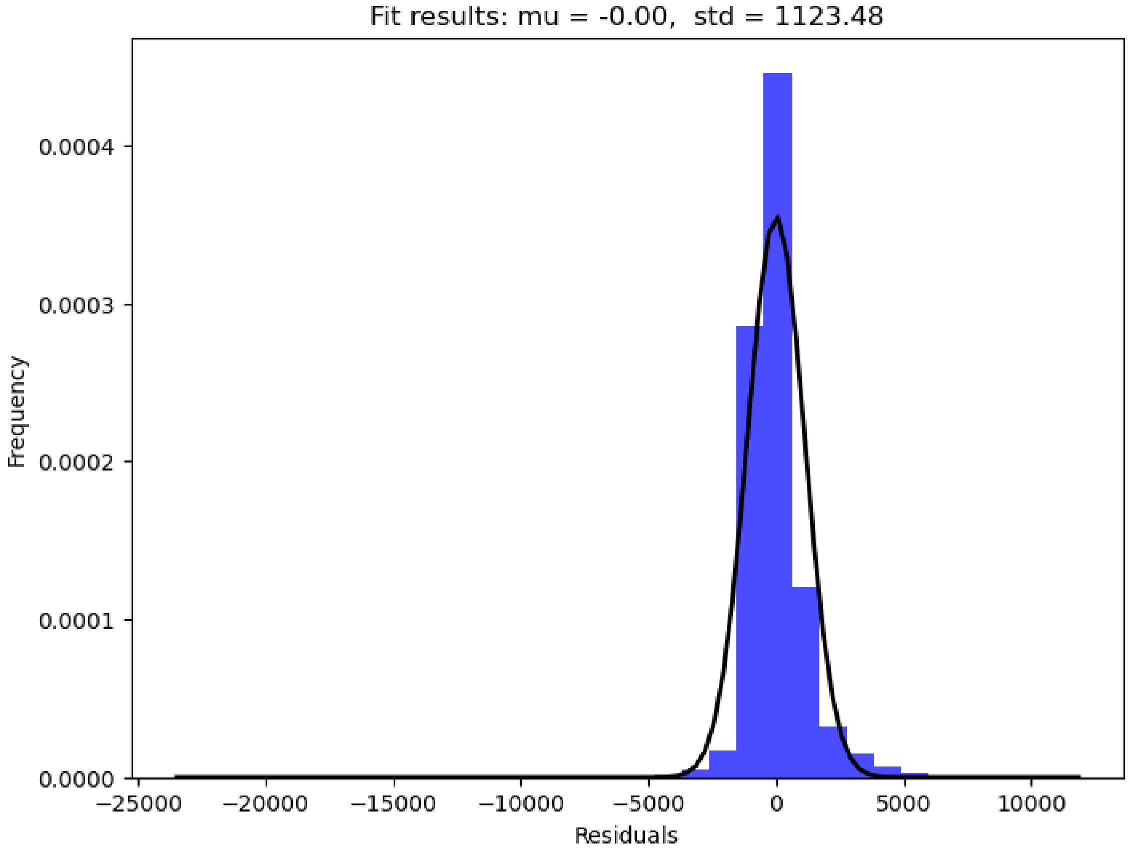
	coef	std err	t	P>|t|	[0.025	0.975]
const	3273.858	467.079	7.009	0	2358.37	4189.345
carat	11540.258	61.707	187.042	0	1.14E+04	1.17E+04
depth	-76.4341	4.863	-15.716	0	-85.967	-66.902
table	-24.9481	3.478	-7.173	0	-31.765	-18.131
x	-1148.72	26.132	-43.958	0	-1199.94	-1097.5
cut_Good	598.4588	39.983	14.968	0	520.091	676.827
cut_Ideal	851.728	39.872	21.362	0	773.578	929.878
cut_Premium	777.475	38.411	20.241	0	702.189	852.761
cut_Very Good	746.3731	38.423	19.425	0	671.064	821.682
color_E	-211.947	21.281	-9.959	0	-253.658	-170.235
color_F	-267.418	21.492	-12.443	0	-309.543	-225.293
color_G	-480.689	21.077	-22.806	0	-522.002	-439.377
color_H	-986.74	22.503	-43.849	0	-1030.85	-942.633
color_I	-1474.02	25.275	-58.319	0	-1523.56	-1424.48
color_J	-2381.16	31.164	-76.407	0	-2442.24	-2320.08
clarity_IF	5395.299	60.904	88.587	0	5275.925	5514.672
clarity_SI1	3714.577	51.85	71.641	0	3612.95	3816.204
clarity_SI2	2756.479	52.065	52.943	0	2654.43	2858.528
clarity_VS1	4629.5	52.918	87.485	0	4525.78	4733.22
clarity_VS2	4302.175	52.128	82.531	0	4200.003	4404.348
clarity_VVS1	5049.809	56.109	89.999	0	4939.833	5159.784
clarity_VVS2	4991.068	54.5	91.58	0	4884.248	5097.889

RESIDUALS

Fitted Model Residual

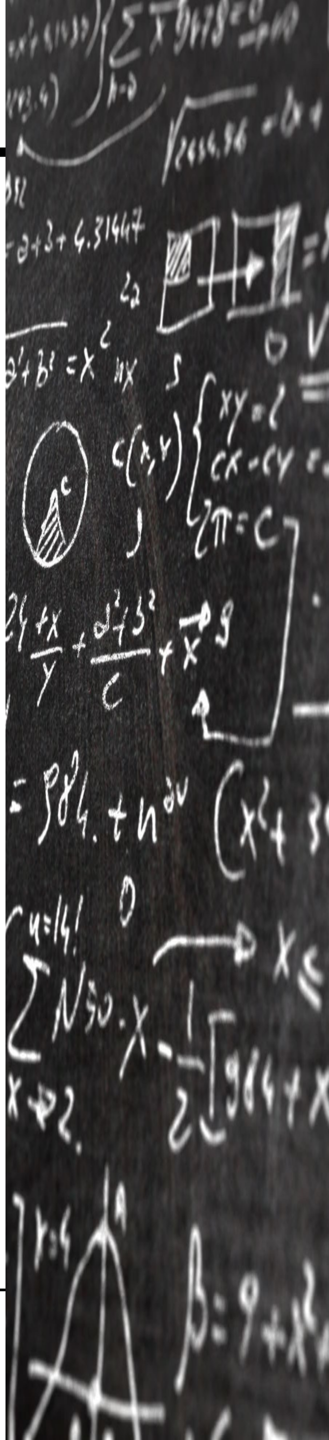


Normal Distribution of the Residuals



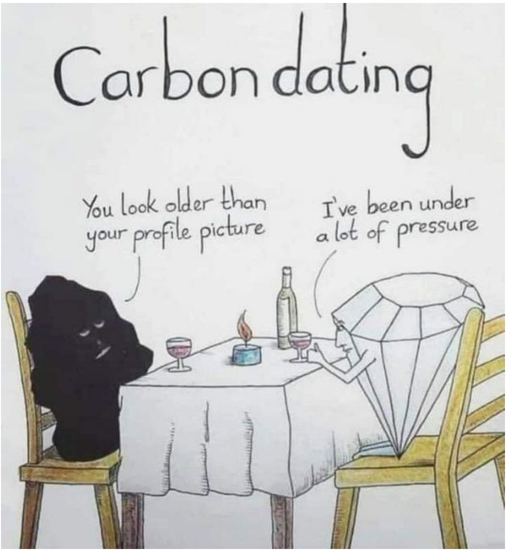
MODEL EVALUATION AND ANALYSIS

- Regression model = 3273.857 + 11540Carat + 5395.2985Clarity_IF +….. + 851.7280Cut_Ideal + ….. - 211.9465ColourE +….. - 1148.7170x - 24.9481table - 76.4341depth
- The model's intercept is 3273.857 and the coefficient for carat is 11540, indicating a strong positive relationship between carat and the dependent variable (price)
- Model has a strong fit - R-squared of 0.92
- $\beta0$, $\beta1$……., βn have meaning intervals between 25^{th} to 95^{th} percent
- Statistically significant ($p < 0.05$) predictor variables
- Residual analysis - symmetric distribution with a median residual of 0
- Categorical variables such as cut, color, and clarity have varying effects
- The residual standard error reduced to 1130 after backward elimination



FURTHER ENHANCEMENTS

- **Feature Engineering Opportunities:** Explore feature engineering possibilities, such as creating combined features or relationship between existing ones. For instance, consider combining information about "Color" and "True Colors" to improve model
- **Include Additional Predictors:** Introduce new predictors like "Shape," "True Colors," and "Origin" to capture additional dimensions of information.
- **Evaluate Various Regression Methods:** Test alternative regression methods (e.g., Decision Tree, Random forest, LGBM, GB etc.,) to assess their performance.
- **Guard Against Overfitting**: Implement measures to avoid overfitting, such as cross-validation and regularization. Balancing model accuracy with generalization ensures the model performs well on new data and doesn't overly tailor itself to the training set.



REFERENCES

- Kigo SN, Omondi EO, Omolo BO. Assessing predictive performance of supervised machine learning algorithms for a diamond pricing model. Sci Rep. 2023 Oct 12;13(1):17315. doi: 10.1038/s41598-023-44326-w. PMID: 37828360; PMCID: PMC10570374.
- Singfat Chu (2001) Pricing the C's of Diamond Stones, Journal of Statistics Education, 9:2, DOI: 10.1080/10691898.2001.11910659

THANK YOU