PROJECT TECHNICAL REPORT

“Diamonds Price Prediction”

Data Science Project using SAS (DSPS)

By

Fatemeh Kiaie

2019

TABLE OF CONTENTS

PROJECT OVERVIEW

A jewelry company wants to put in a bid to purchase a large set of diamonds but is unsure how much it should bid. In this project, results from a predictive model to make recommendation on how much the jewelry company should bid for the diamonds. In other words, this project is meant to predict price of diamonds so that jewelry company management can make informed decisions about how much to bid a large set of diamonds in the auction.

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

In this project, which is a part of data science problem domain to quantitatively estimate the unknown price of diamonds by leveraging predictive analytics and extensively utilizing team data science processes to implement a supervised algorithm appropriate for Machine Learning to predict diamonds price through data management in SAS environment. The data set is described as below

Data element	*Description*
Price	price in US dollars ($326--$18,823)
Carat	weight of the diamond (0.2--5.01)
Cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
Color	diamond colour, from D (best) to J (worst)
Clarity	measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
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)

Project in-Scope:

- SAS for Windows 9.4M6
- SAS Enterprise Guide 8.1

- Descriptive Statistics and Exploratory Data Analysis (EDA)
- Linear Regression Model
- Predictive Modeling and Model Performance Evaluation
- Statistical Data Visualization (DV)
- Exception Report in Excel format for Data Quality

Project Out-of-Scope:

- Python/R
- Excel Data Analysis and graphics
- Tableau or other visualization tools
- Unsupervised Machine Learning

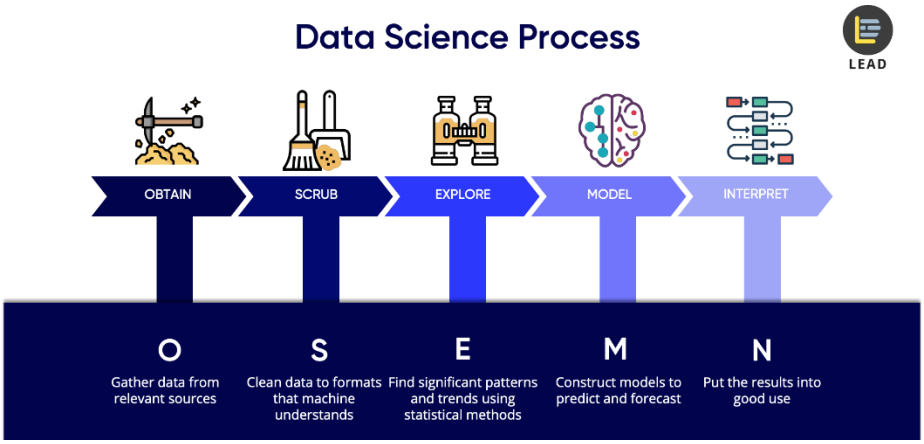
OBJECTIVES

The main objective of this project is conducting an Exploratory Data Analysis (EDA) for the diamond data set and predicting the prices of Diamonds.

DATA SCIENCE PROCESS DETAILS

As can be seen in flow below the Data Science Process Detail can be described in the following steps:

1. Setting the research goal
    a) Understanding the goal and context of your research
    b) Project Charter
2. Data Extraction
    a) Collect Data
3. Data Preparation
    a) Data Cleaning
    b) Data Transformation
4. Data Exploration and Visualization
    a) Exploratory Data Analysis (EDA)
5. Build the Model/Predictive Modeling
    a) Model and variable selection
    b) Model execution
    c) Model validation
6. Presentation and Deployment
    a) Presenting data
    b) Automating data analysis


Data Science Process
LEAD
OBTAIN
SCRUB
EXPLORE
MODEL
INTERPRET
O
Gather data from relevant sources
S
Clean data to formats that machine understands
E
Find significant patterns and trends using statistical methods
M
Construct models to predict and forecast
N
Put the results into good use

Data Cleaning

Following the data science process after importing the dataset, the cleaning data set showing that although there is no missing value in the data, 20 records are invalid, as they have x=0, or (and) y=0, or (and) z=0. Moreover, there are 289 records duplicated that needed to be removed to avoid the bias model prediction. The non-valid records and duplicated records are presented in the Exception report to the client.

Data Exploration and Visualization

The next step in the data science process is the exploration of data that is exploratory data analysis (EDA), which is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

The effects of independent "continues" feature such as cut, color, clarity, and carat on the price (continues target) are shown in Fig.1 to Fig.4 .As can be seen in Fig. 1. the maximum price is for Premium cut, and the Ideal cut gets the minimum price.

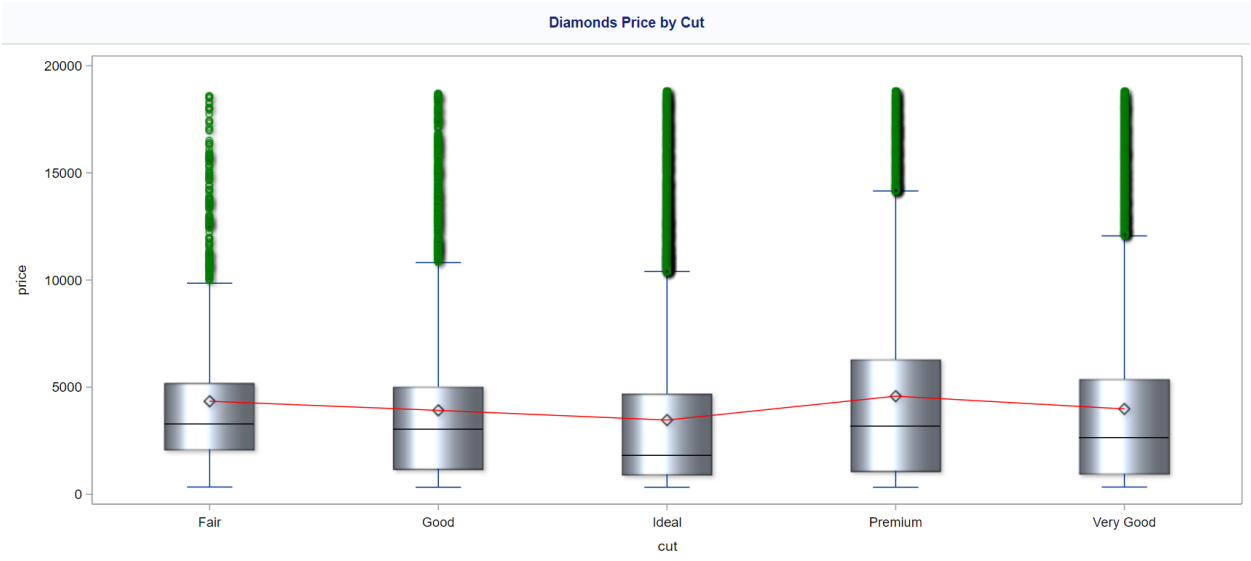


Fig. 1. Diamond Price by Cut

Similarly, comparing the color with the price in Fig. 2, shows that the “J” color is the most expensive color, and the median price for “J” color is almost twice the median for color “E” that has the minimum price.

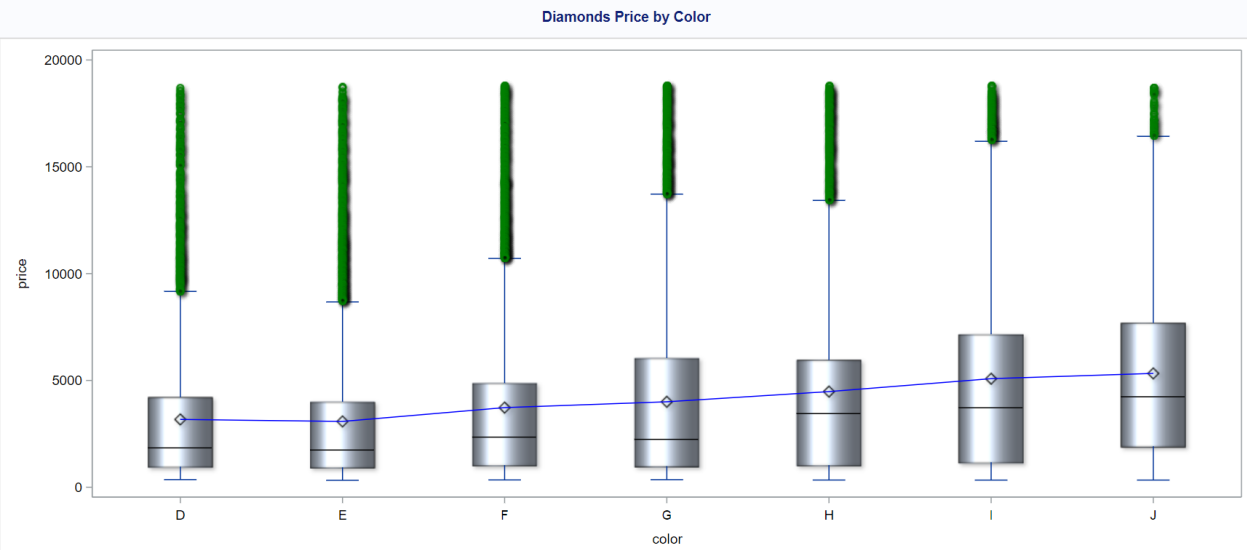


Fig. 2. Diamond Price by Color

Moreover, comparing the clarity with the price in Fig. 3 shows that the “SI2” is the most expensive, and the median price for “IF” and “WS1” are minimum. Moreover, it is clear that although the maximum price for the “VS1” and “VS2” and “SI2” are almost the same, the median price is much higher in “SI2”.

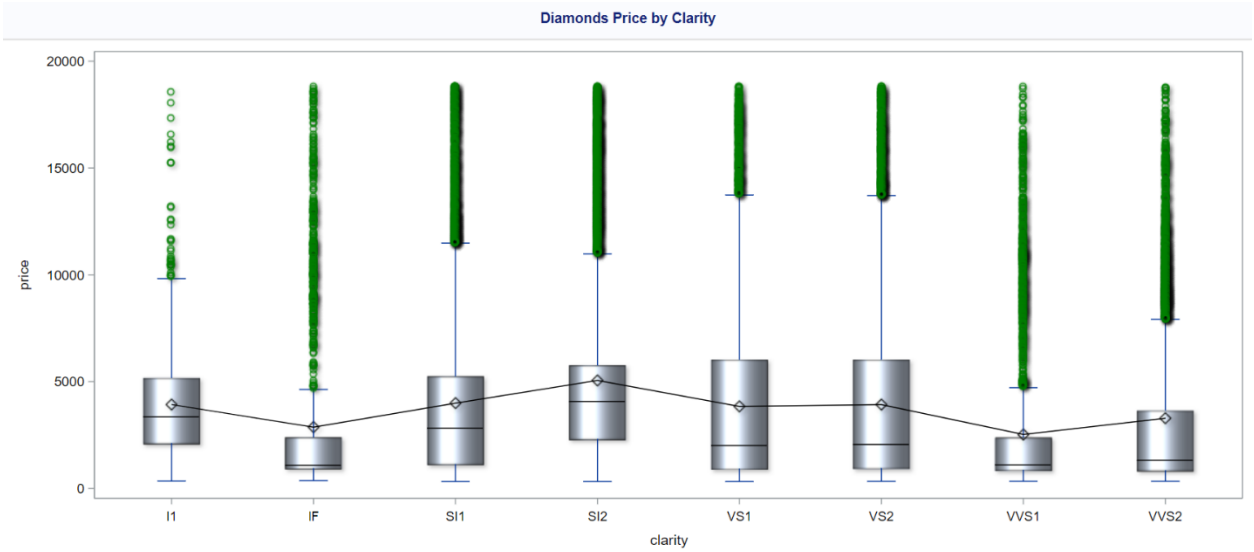


Fig. 3. Diamond Price by Clarity

As we are expecting and shown in Fig. 4, the diamond price in general increases by the carat increment; however, in the very large carat, we may have some low price, and that can be due to the lack of customer or other features, like cut and clarity of the diamond in the large carat.

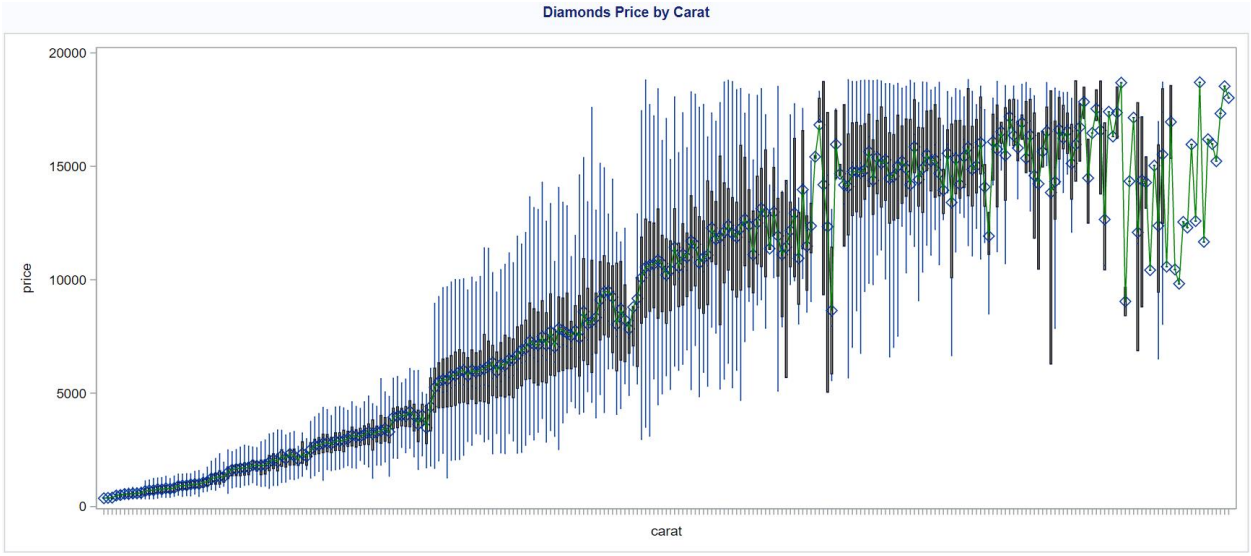


Fig. 4. Diamond Price by Carat

The population of the price for different carats shown in Fig. 5 and revealing that 5.8% of price proportion coming from carat 1.01 and 4.0% of price proportion belongs to carat 1.51, while the rest is for others.

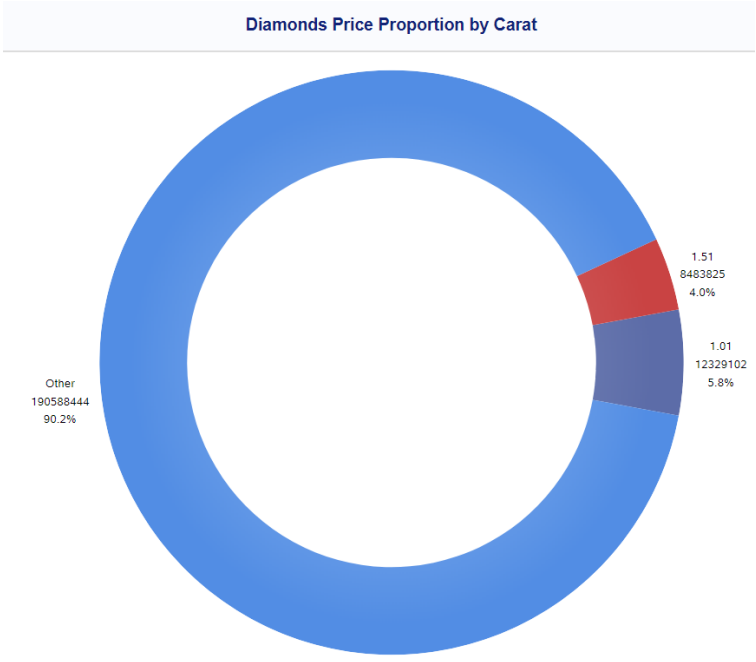


Fig. 5. Diamond Price by Carat

Moreover, as can be seen in Fig.6, the most proportion of price is for the ideal cut and premium. Comparing the figure below and Fig.1 that the minimum price belonged to the ideal cut, we can conclude that the customers are buying the cheaper Diamond more frequently, and it is evident that the cheaper cut diamond is more popular. On the other side, due to the high price for the premium cut, it will automatically have a large portion of the price.

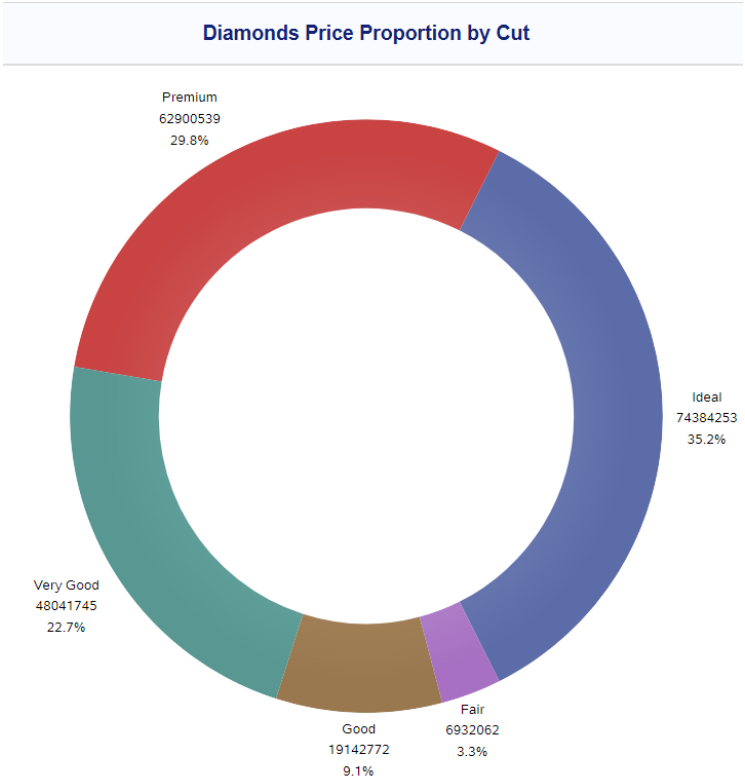


Fig. 6. Diamond Price by Cut

From Fig.7, it can be seen that the most price portion is for the “G” color, and comparing with Fig.2, color “G” is in the middle expense category and therefore, it has a chance of being the most popular to bring the most portion of the price. While color ‘J,’ the most expensive color from Fig.2, has a small portion in the price. The Price Population by color result has also been declared in Table 1, including the cutting categories with respect to a different color. The statistics information of Table1 is shown in Table2.

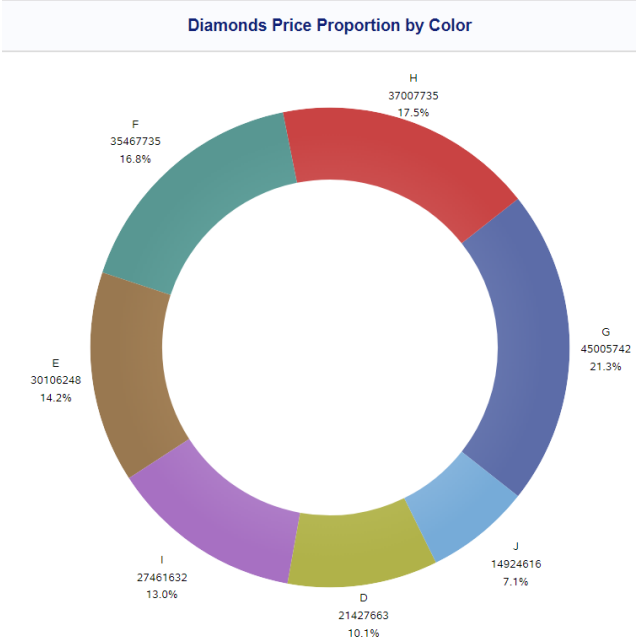


Fig. 7. Diamond Price by Color

Diamonds Price Proportion by Color

Frequency
Percent
Row Pct
Col Pct

Table of cut by color

cut	color							
	D	E	F	G	H	I	J	Total
Fair	163 0.30 10.21 2.41	222 0.41 13.90 2.27	309 0.57 19.35 3.25	311 0.58 19.47 2.76	299 0.56 18.72 3.62	174 0.32 10.90 3.22	119 0.22 7.45 4.25	1597 2.97
Good	660 1.23 13.50 9.77	931 1.73 19.05 9.52	907 1.69 18.56 9.53	867 1.61 17.74 7.70	699 1.30 14.30 8.46	518 0.96 10.60 9.58	306 0.57 6.26 10.92	4888 9.09
Ideal	2823 5.25 13.14 41.80	3893 7.24 18.12 39.82	3818 7.10 17.77 40.12	4863 9.04 22.63 43.21	3104 5.77 14.45 37.55	2090 3.89 9.73 38.66	894 1.66 4.16 31.91	21485 39.95
Premium	1598 2.97 11.63 23.66	2331 4.33 16.97 23.84	2320 4.31 16.89 24.38	2915 5.42 21.22 25.90	2346 4.36 17.08 28.38	1421 2.64 10.34 26.29	806 1.50 5.87 28.77	13737 25.55
Very Good	1510 2.81 12.51 22.36	2399 4.46 19.88 24.54	2163 4.02 17.92 22.73	2298 4.27 19.04 20.42	1818 3.38 15.06 21.99	1203 2.24 9.97 22.25	677 1.26 5.61 24.16	12068 22.44
Total	6754 12.56	9776 18.18	9517 17.70	11254 20.93	8266 15.37	5406 10.05	2802 5.21	53775 100.00

Table 1. Diamond Price Proportion by Color

Statistics for Table of cut by color

Statistic	DF	Value	Prob
Chi-Square	24	306.7023	<.0001
Likelihood Ratio Chi-Square	24	308.5476	<.0001
Mantel-Haenszel Chi-Square	1	0.0067	0.9349
Phi Coefficient		0.0755	
Contingency Coefficient		0.0753	
Cramer's V		0.0378	

Table 2. Statistic for Table of cut by color

Fig. 8 reveals the distribution of different cuts among different colors. It can be seen that the most popular cut and the color is ideal cut in the "G" color, and in general, Fair cut has the minimum popularity among all colors.

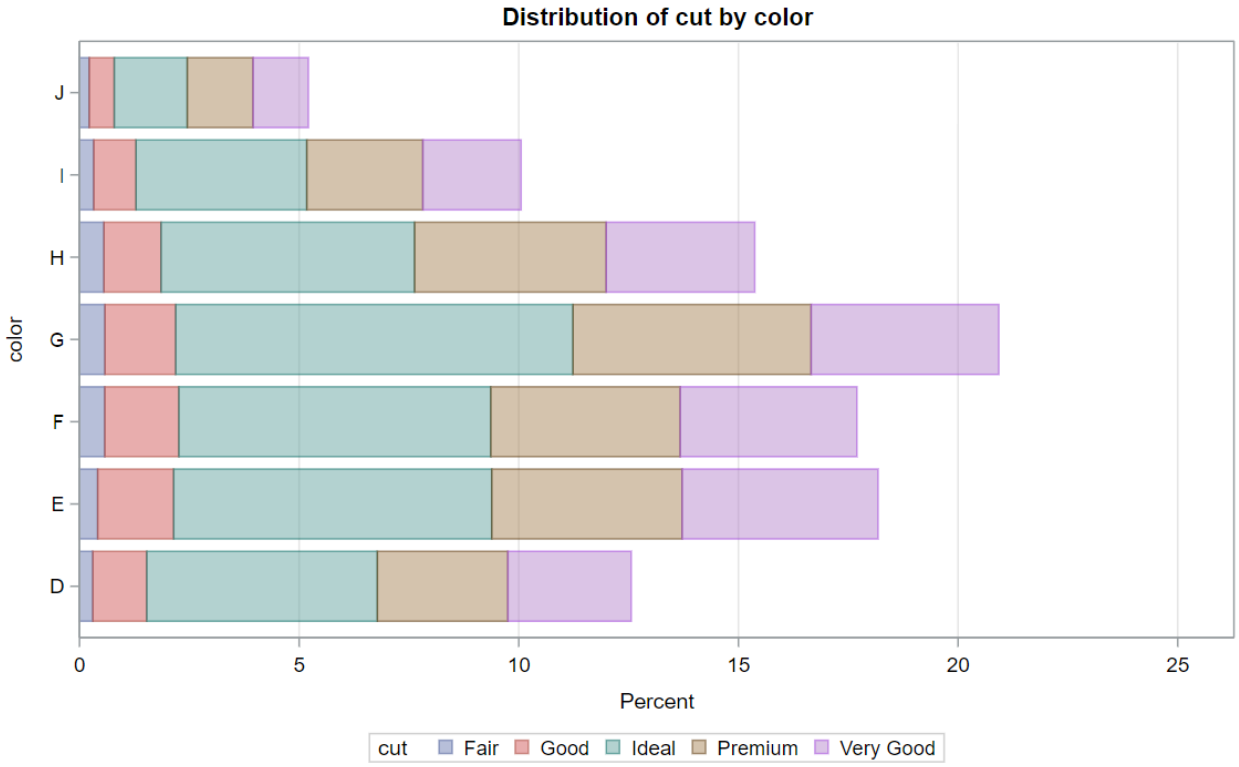


Fig. 8. Distribution of cuts by color

The population of different clarity among different cut is shown in Table 3, and corresponding graph is shown in Fig.9. Ideal cut with the VS2 clarity is the most popular combination among the other combinations. Table 4 presents the statistical information in table 3.

Frequency Percent Row Pct Col Pct	Table of cut by clarity								
	clarity								
cut	I1	IF	SI1	SI2	VS1	VS2	VVS1	VVS2	Total
Fair	210 0.39 13.15 28.49	9 0.02 0.56 0.50	406 0.75 25.42 3.12	459 0.85 28.74 5.02	169 0.31 10.58 2.07	258 0.48 16.16 2.11	17 0.03 1.06 0.47	69 0.13 4.32 1.36	1597 2.97
Good	94 0.17 1.92 12.75	71 0.13 1.45 3.98	1555 2.89 31.81 11.93	1072 1.99 21.93 11.73	646 1.20 13.22 7.92	978 1.82 20.01 8.00	186 0.35 3.81 5.10	286 0.53 5.85 5.66	4888 9.09
Ideal	146 0.27 0.68 19.81	1206 2.24 5.61 67.60	4268 7.94 19.87 32.76	2593 4.82 12.07 28.36	3582 6.66 16.67 43.92	5054 9.40 23.52 41.34	2039 3.79 9.49 55.92	2597 4.83 12.09 51.36	21485 39.95
Premium	203 0.38 1.48 27.54	230 0.43 1.67 12.89	3565 6.63 25.95 27.36	2922 5.43 21.27 31.96	1984 3.69 14.44 24.33	3349 6.23 24.38 27.39	615 1.14 4.48 16.87	869 1.62 6.33 17.19	13737 25.55
Very Good	84 0.16 0.70 11.40	268 0.50 2.22 15.02	3236 6.02 26.81 24.83	2096 3.90 17.37 22.93	1774 3.30 14.70 21.75	2586 4.81 21.43 21.15	789 1.47 6.54 21.64	1235 2.30 10.23 24.43	12068 22.44
Total	737 1.37	1784 3.32	13030 24.23	9142 17.00	8155 15.17	12225 22.73	3646 6.78	5056 9.40	53775 100.00

Table 3. The proportion of cut by clarity

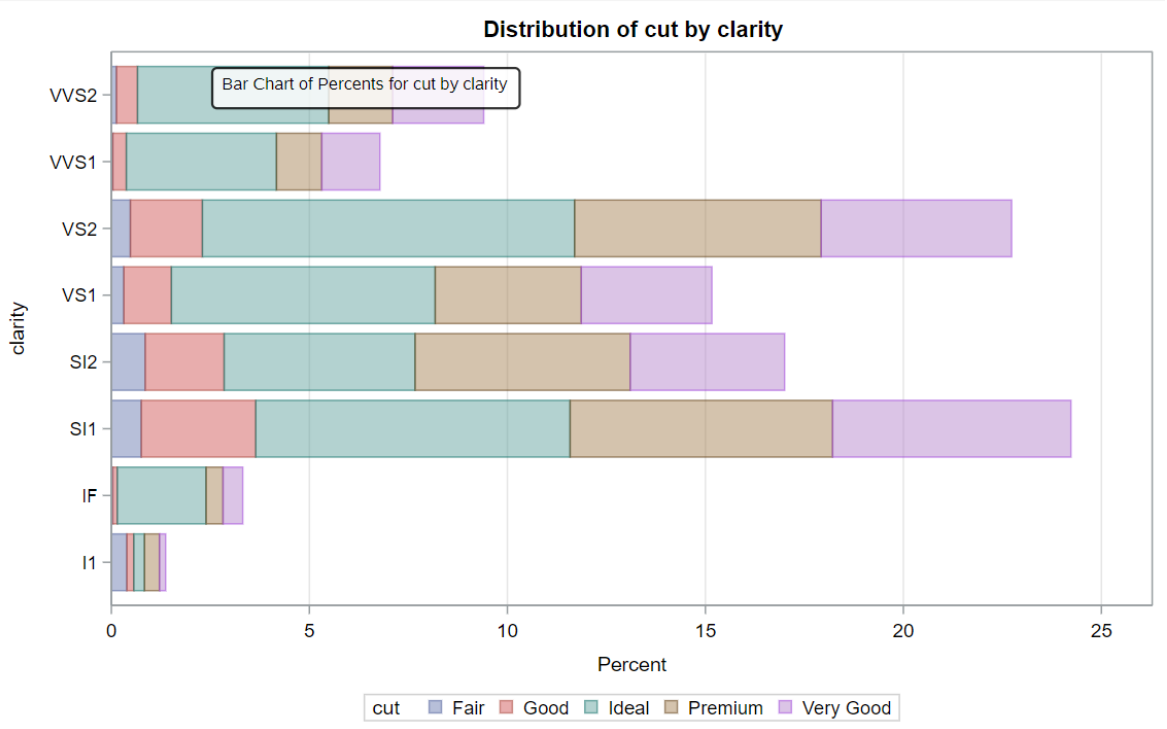


Fig. 9. Distribution of cuts by clarity

Statistics for Table of cut by clarity

Statistic	DF	Value	Prob
Chi-Square	28	4373.3795	<.0001
Likelihood Ratio Chi-Square	28	3430.0189	<.0001
Mantel-Haenszel Chi-Square	1	42.5048	<.0001
Phi Coefficient		0.2852	
Contingency Coefficient		0.2742	
Cramer's V		0.1426	

Table 4. Statistic for Table of cut by color

The population of different colors among different clarities is shown in Table 5, and the corresponding graph is shown in Fig.10.The population is almost uniform for a different color in VS2 and SI1 clarities. Table 6 presents the statistical information in table 5.

Frequency
Percent
Row Pct
Col Pct

Table of color by clarity

color	clarity								
	I1	IF	SI1	SI2	VS1	VS2	VVS1	VVS2	Total
D	42 0.08 0.62 5.70	73 0.14 1.08 4.09	2074 3.86 30.71 15.92	1367 2.54 20.24 14.95	705 1.31 10.44 8.65	1691 3.14 25.04 13.83	249 0.46 3.69 6.83	553 1.03 8.19 10.94	6754 12.56
E	101 0.19 1.03 13.70	158 0.29 1.62 8.86	2423 4.51 24.79 18.60	1704 3.17 17.43 18.64	1281 2.38 13.10 15.71	2464 4.58 25.20 20.16	656 1.22 6.71 17.99	989 1.84 10.12 19.56	9776 18.18
F	143 0.27 1.50 19.40	383 0.71 4.02 21.47	2127 3.96 22.35 16.32	1598 2.97 16.79 17.48	1362 2.53 14.31 16.70	2198 4.09 23.10 17.98	734 1.36 7.71 20.13	972 1.81 10.21 19.22	9517 17.70
G	148 0.28 1.32 20.08	678 1.26 6.02 38.00	1969 3.66 17.50 15.11	1540 2.86 13.68 16.85	2141 3.98 19.02 26.25	2340 4.35 20.79 19.14	995 1.85 8.84 27.29	1443 2.68 12.82 28.54	11254 20.93
H	161 0.30 1.95 21.85	298 0.55 3.61 16.70	2267 4.22 27.43 17.40	1550 2.88 18.75 16.95	1166 2.17 14.11 14.30	1637 3.04 19.80 13.39	583 1.08 7.05 15.99	604 1.12 7.31 11.95	8266 15.37
I	92 0.17 1.70 12.48	143 0.27 2.65 8.02	1420 2.64 26.27 10.90	904 1.68 16.72 9.89	961 1.79 17.78 11.78	1166 2.17 21.57 9.54	355 0.66 6.57 9.74	365 0.68 6.75 7.22	5406 10.05
J	50 0.09 1.78 6.78	51 0.09 1.82 2.86	750 1.39 26.77 5.76	479 0.89 17.09 5.24	539 1.00 19.24 6.61	729 1.36 26.02 5.96	74 0.14 2.64 2.03	130 0.24 4.64 2.57	2802 5.21
Total	737 1.37	1784 3.32	13030 24.23	9142 17.00	8155 15.17	12225 22.73	3646 6.78	5056 9.40	53775 100.00

Table 5. The proportion of color by clarity

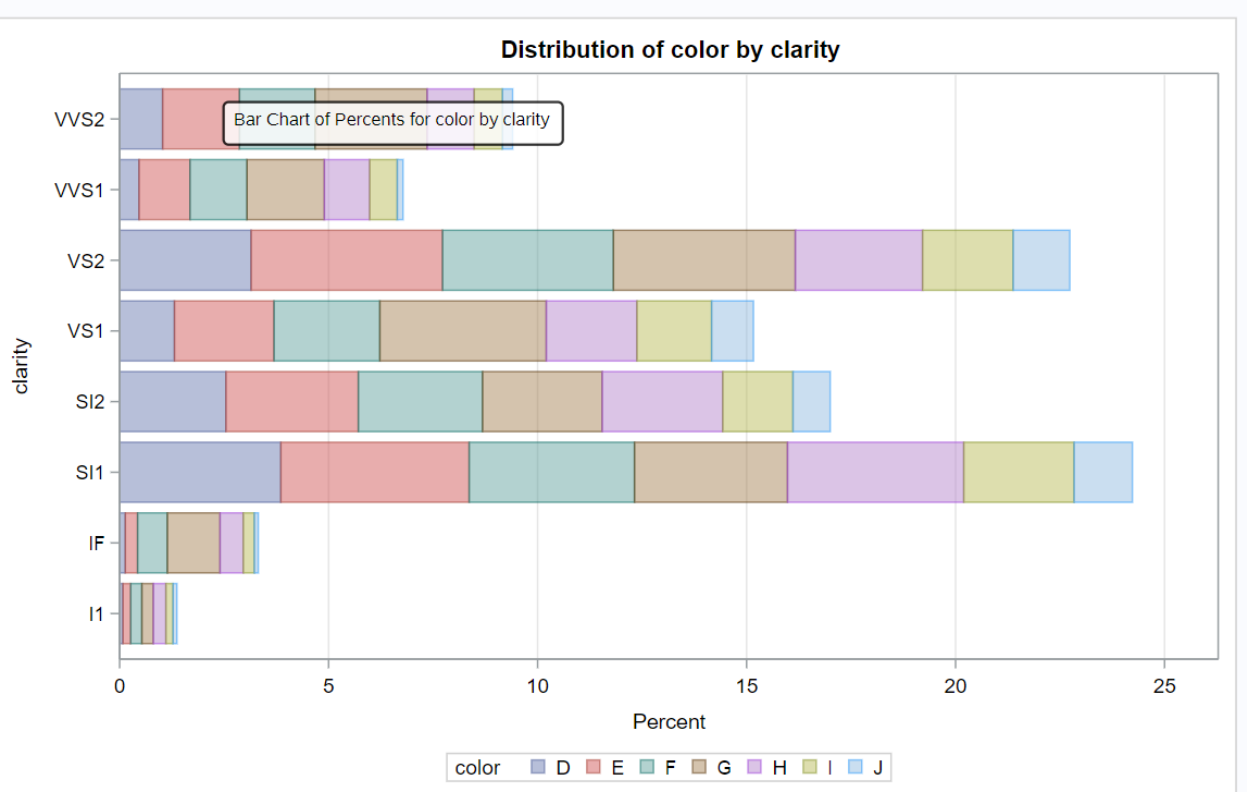


Fig. 10. Distribution of color by clarity

Statistics for Table of color by clarity

Statistic	DF	Value	Prob
Chi-Square	42	2040.1004	<.0001
Likelihood Ratio Chi-Square	42	2115.3916	<.0001
Mantel-Haenszel Chi-Square	1	41.8320	<.0001
Phi Coefficient		0.1948	
Contingency Coefficient		0.1912	
Cramer's V		0.0795	

Table 6. Statistic for Table of color by clarity

In order to reveal the correlation between continues and continues variables, the correlation matrix is presented in Fig. 10. As can be seen, there is a strong correlation relationship between the dimensions parameters of the diamond (x ,y, z) and its price. Carat is also have a strong correlation relating to the price of the diamond.

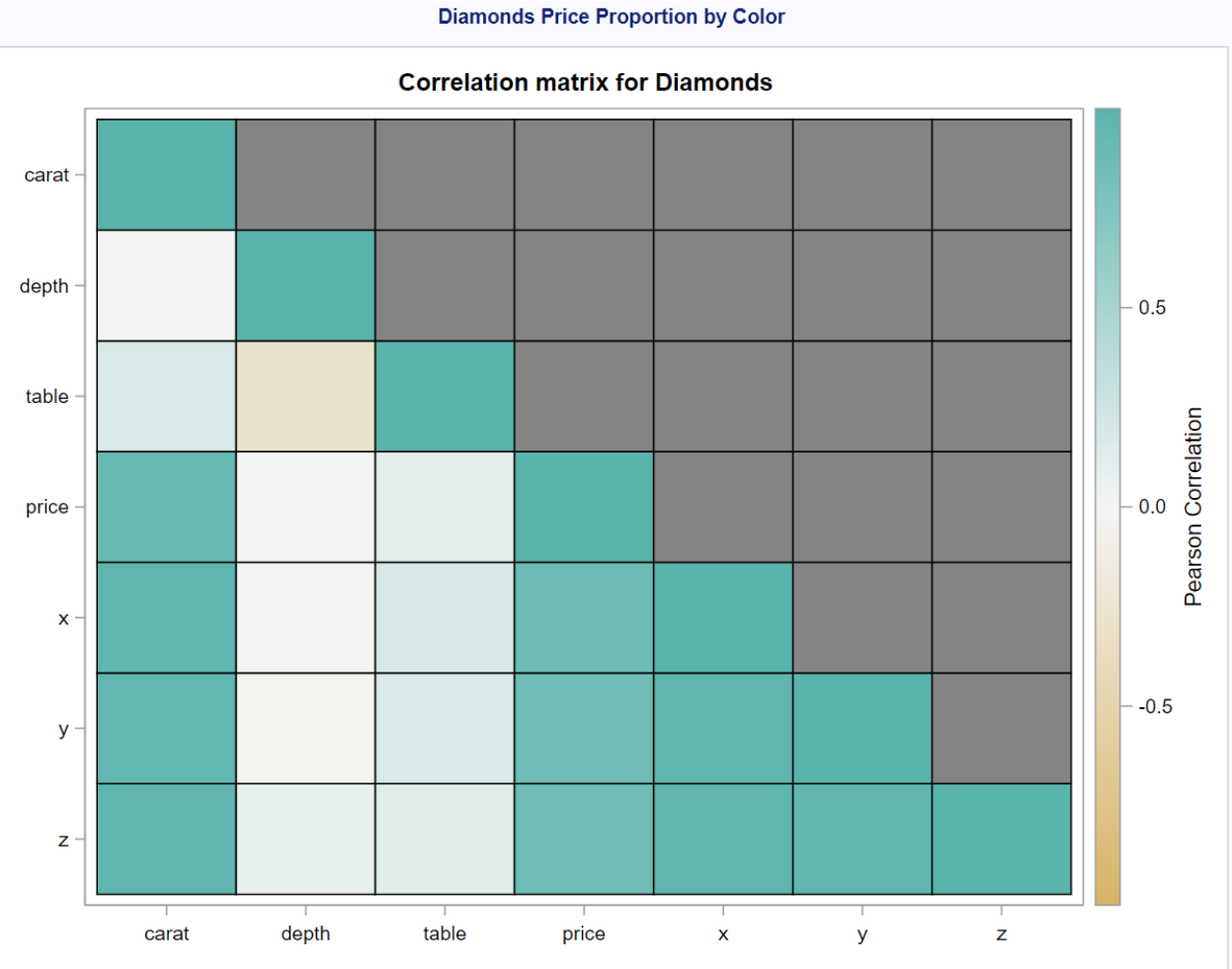


Fig. 11 Correlation matrix

The distribution of the price is the right-skewed graph shown in Fig. 11, and the Mean price of the Diamond is 3931.22$. The minimum price is 326$, with a maximum of 18823$. However, as can be seen in Fig. 12, the carat distribution fluctuates in some values, and it is not following the normal distribution. Fig.13 also reveals that the depth follows the normal symmetric distribution with a mean of 61.78mm in the range of 43 to 79. Moreover, Fig 13, 14, 15, 16 and 17 show that except the x dimension, all other physical dimensions following a normal distribution. The mean of x, y, and z are 5.73, 5.74, and 3.34, respectively.

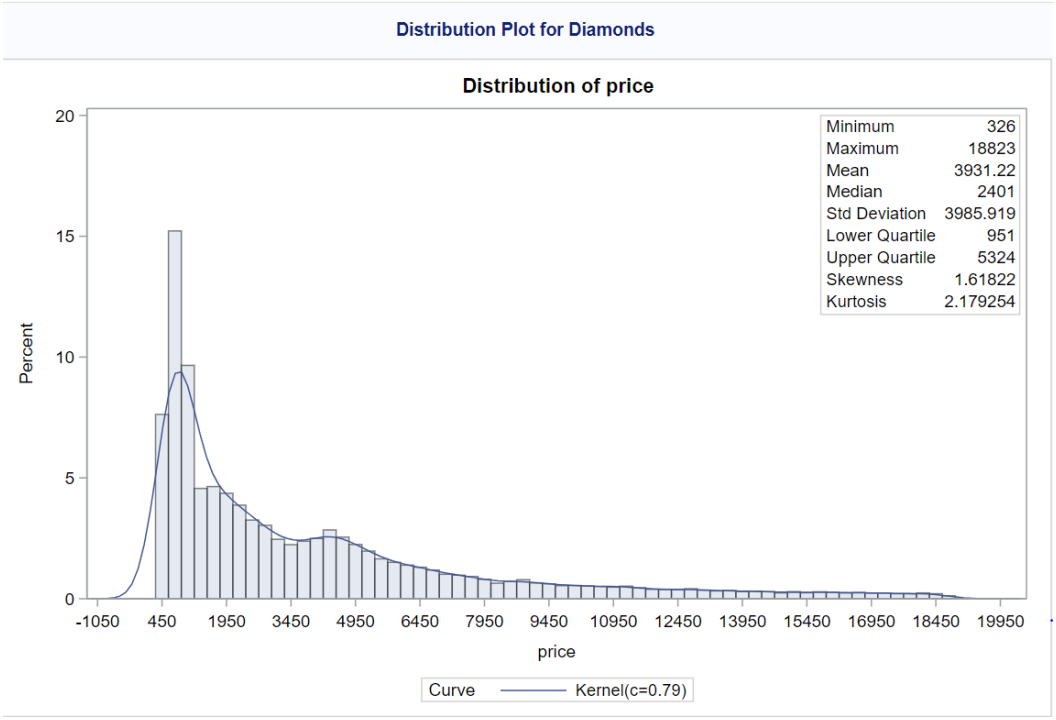


Fig. 12. Distribution of price

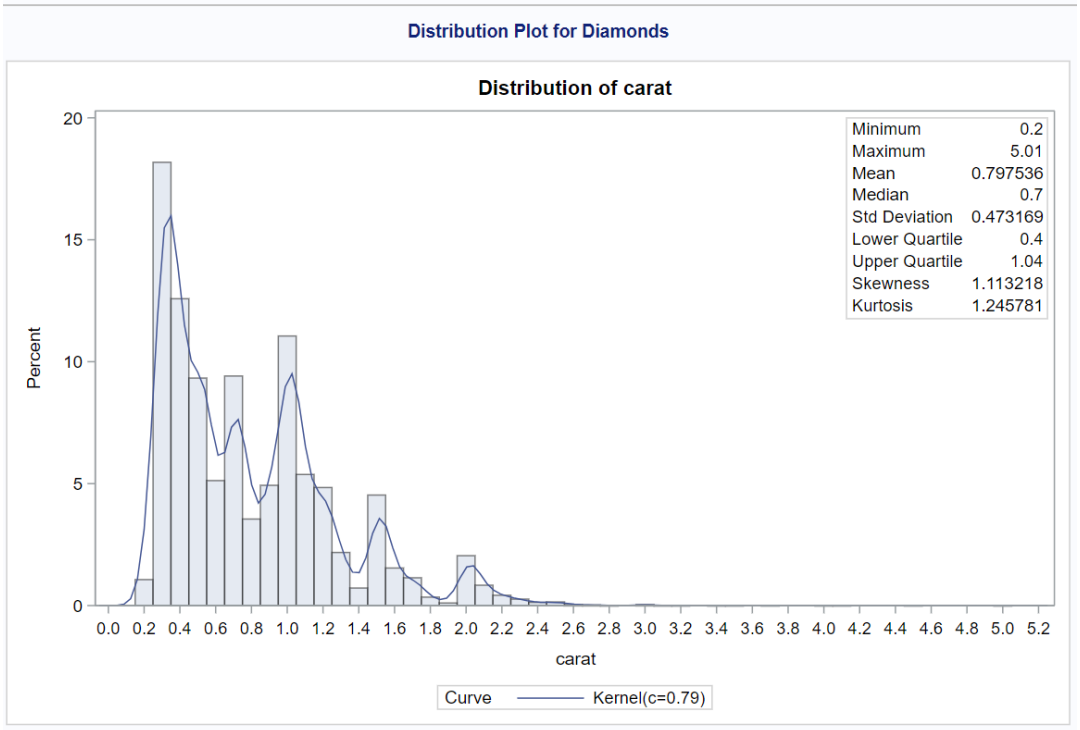


Fig. 13. Distribution of carat

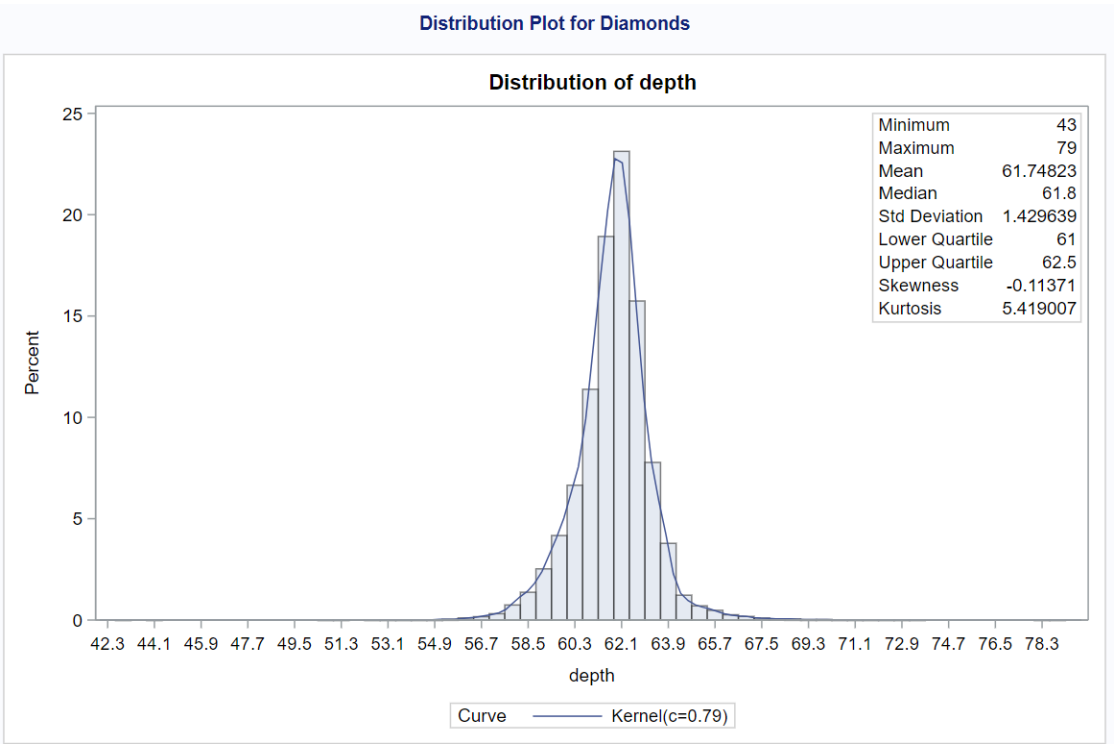


Fig. 14. Distribution of depth

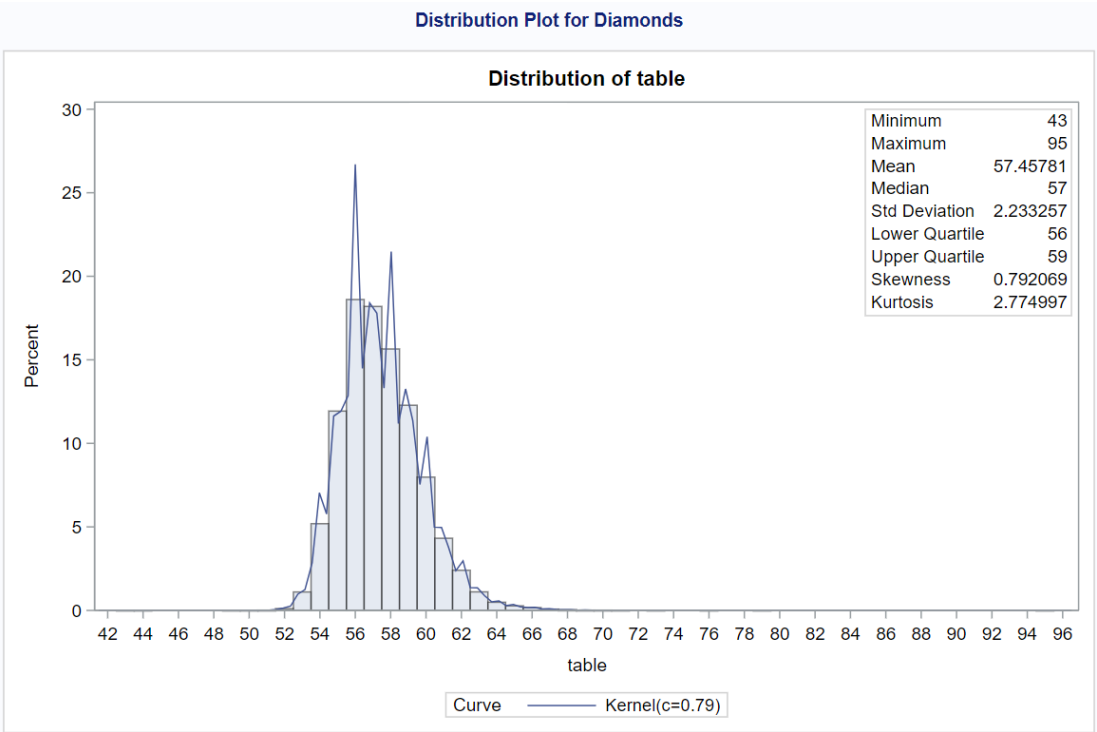


Fig. 15. Distribution of table

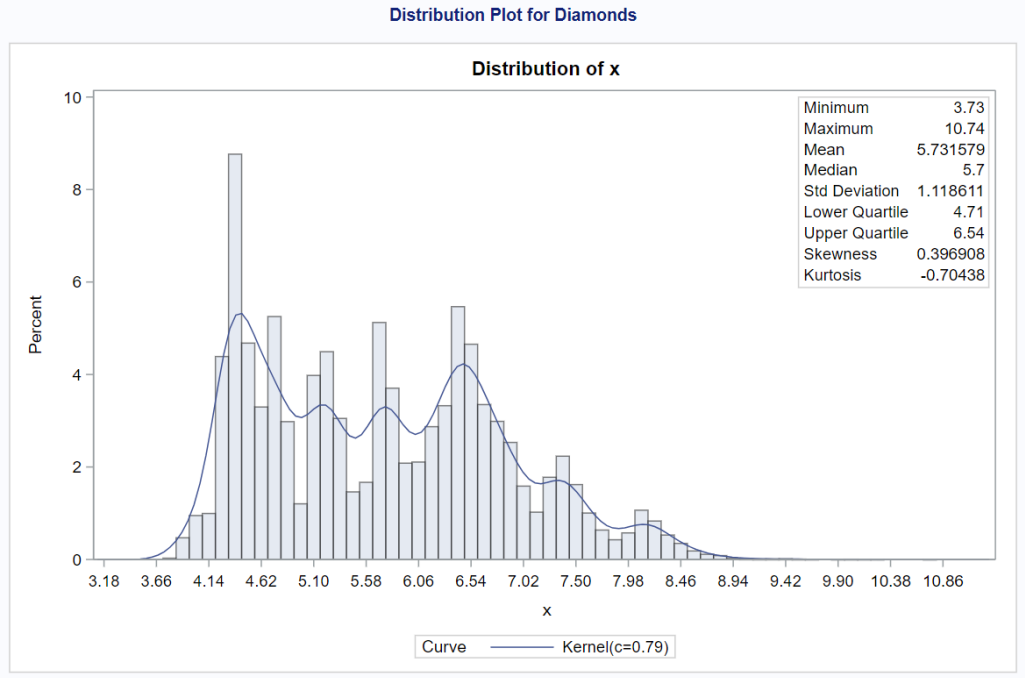


Fig. 16. Distribution of x

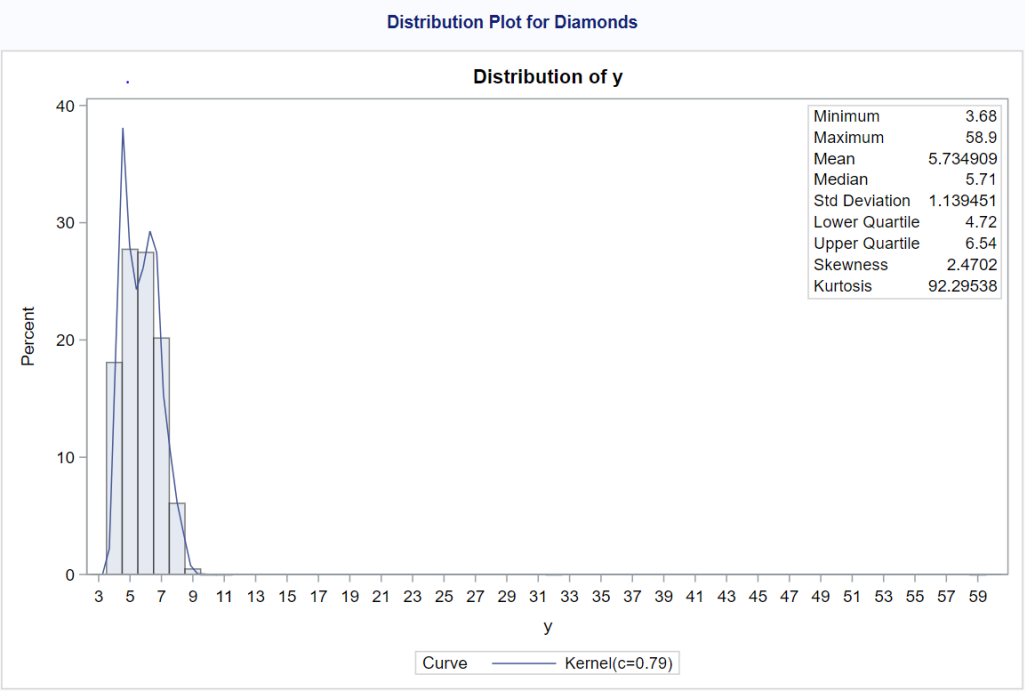


Fig. 17. Distribution of y

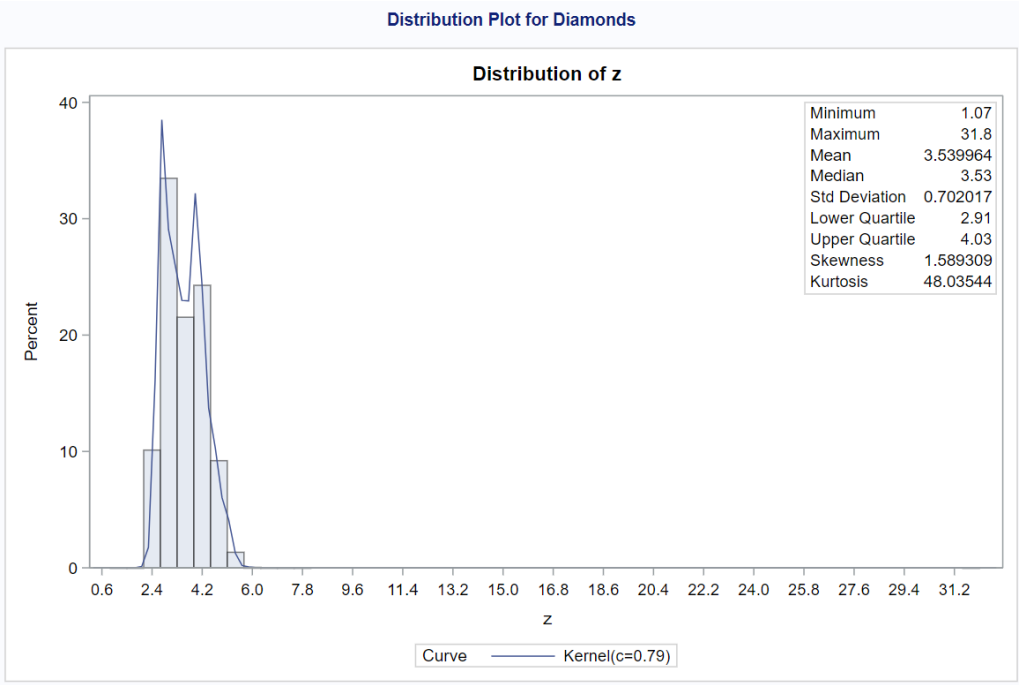


Fig. 18. Distribution of z

Feature Engineering

We are implementing the hot encoding, dummy variables to change the categorical variables to numeric variables. Another feature engineering method that been implemented is to normalize the features that are not normally distributed like carat and price. The results are shown in Fig. 19 and Fig.20.

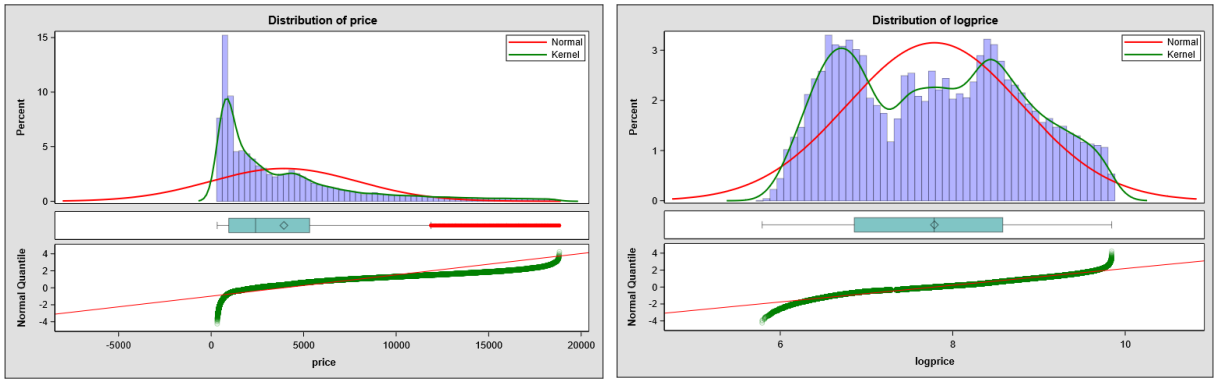


Fig. 19. Normalizing the price

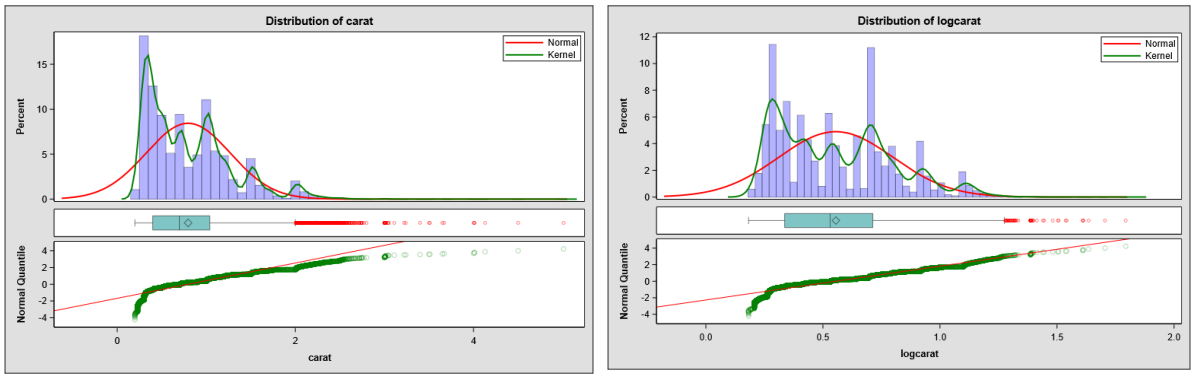


Fig. 20. Normalizing the carat

Model and variable selection

The next step after EDA is feature engineering, and we use dummy encoding that needed to be done due to the categorical variables. After EDA we split the training dataset to the validation and test, and that is required to evaluate the performance of the model. After data splitting, we have used C model, using the GLMSELECT model, to train the model. The results are shown as follows.

Distribution Plot for Diamonds

Data Set	FKIAIE.DIAMONDSFE_TRAINING
Dependent Variable	logprice
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	SBC
Effect Hierarchy Enforced	None

Number of Observations Read	37643
Number of Observations Used	37643

Class Level Information		
Class	**Levels**	**Values**
cut	5	Fair Good Ideal Premium Very Good
color	7	D E F G H I J
clarity	8	I1 IF SI1 SI2 VS1 VS2 VVS1 VVS2

Dimensions	
Number of Effects	11
Number of Parameters	173

Distribution Plot for Diamonds

Stepwise Selection Summary					
Step	**Effect Entered**	**Effect Removed**	**Number Effects In**	**Number Parms In**	**SBC**
0	**Intercept**		1	1	1004.03
1	**logcarat*clarity**		2	9	-106766.85
2	**logcarat*color**		3	15	-126544.60
3	**color*clarity**		4	70	-130939.66
4	**logcarat*cut**		5	74	-132171.76
5	**cut**		6	78	-133199.47*
* Optimal Value of Criterion					

Selection stopped at a local minimum of the SBC criterion.

Stop Details				
Candidate For	**Effect**	**Candidate SBC**		**Compare SBC**
Entry	cut*color	-133110.83	>	-133199.47
Removal	cut	-132171.76	>	-133199.47

Table 7. Model information

Distribution Plot for Diamonds

Selected Model

The selected model is the model at the last step (Step 5).

Effects:	Intercept cut logcarat*cut logcarat*color logcarat*clarity color*clarity

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	77	37580	488.04608	17131.2
Error	37565	1070.17606	0.02849	
Corrected Total	37642	38650		

Root MSE	0.16879
Dependent Mean	7.78464
R-Square	0.9723
Adj R-Sq	0.9723
AIC	-96220
AICC	-96220
SBC	-133199

Table 8. distribution plot of the selected model

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	5.094577	0.022838	223.08
cut Fair	1	0.265818	0.017585	15.12
cut Good	1	0.005359	0.009178	0.58
cut Ideal	1	0.083646	0.005802	14.42
cut Premium	1	0.188076	0.006460	29.12
cut Very Good	0	0	.	.
logcarat*cut Fair	1	-0.531567	0.024819	-21.42
logcarat*cut Good	1	-0.059013	0.014750	-4.00
logcarat*cut Ideal	1	-0.060207	0.009833	-6.12
logcarat*cut Premium	1	-0.297934	0.010271	-29.01
logcarat*cut Very Good	0	0	.	.
logcarat*color D	1	1.097973	0.019713	55.70
logcarat*color E	1	1.032514	0.018306	56.40
logcarat*color F	1	0.904996	0.017898	50.56
logcarat*color G	1	0.705739	0.016968	41.59
logcarat*color H	1	0.468214	0.017394	26.92
logcarat*color I	1	0.257683	0.018184	14.17
logcarat*color J	0	0	.	.
logcarat*clarity I1	1	2.564877	0.036243	70.77
logcarat*clarity IF	1	4.356121	0.031396	138.75
logcarat*clarity SI1	1	3.855686	0.017541	219.82
logcarat*clarity SI2	1	3.303263	0.018440	179.13
logcarat*clarity VS1	1	4.217303	0.018279	230.72
logcarat*clarity VS2	1	4.053238	0.017504	231.56
logcarat*clarity VVS1	1	4.481726	0.025391	176.51
logcarat*clarity VVS2	1	4.388022	0.021468	204.39
color*clarity D I1	1	0.223994	0.046855	4.78
color*clarity D IF	1	0.675778	0.036511	18.51
color*clarity D SI1	1	0.018813	0.024307	0.77
color*clarity D SI2	1	0.124001	0.025128	4.93
color*clarity D VS1	1	0.097382	0.025249	3.86
color*clarity D VS2	1	0.112893	0.024216	4.66
color*clarity D VVS1	1	0.277963	0.027669	10.05
color*clarity D VVS2	1	0.169716	0.022954	7.39
color*clarity E I1	1	0.248386	0.038461	6.46
color*clarity E IF	1	0.326412	0.030127	10.83
color*clarity E SI1	1	0.003274	0.024142	0.14
color*clarity E SI2	1	0.136638	0.024851	5.50
color*clarity E VS1	1	0.067942	0.024391	2.79
color*clarity E VS2	1	0.080280	0.023911	3.36

color*clarity E VS1	1	0.067942	0.024391	2.79
color*clarity E VS2	1	0.080280	0.023911	3.36
color*clarity E VVS1	1	0.160200	0.025192	6.36
color*clarity E VVS2	1	0.115105	0.022118	5.20
color*clarity F I1	1	0.314887	0.037460	8.41
color*clarity F IF	1	0.297075	0.027015	11.00
color*clarity F SI1	1	0.024694	0.024350	1.01
color*clarity F SI2	1	0.202022	0.025047	8.07
color*clarity F VS1	1	0.075048	0.024603	3.05
color*clarity F VS2	1	0.077535	0.024119	3.21
color*clarity F VVS1	1	0.181671	0.025325	7.17
color*clarity F VVS2	1	0.140248	0.022061	6.36
color*clarity G I1	1	0.356989	0.038458	9.28
color*clarity G IF	1	0.281571	0.026287	10.71
color*clarity G SI1	1	0.059510	0.024344	2.44
color*clarity G SI2	1	0.258218	0.025063	10.30
color*clarity G VS1	1	0.101481	0.024313	4.17
color*clarity G VS2	1	0.093107	0.024240	3.84
color*clarity G VVS1	1	0.151604	0.025113	6.04
color*clarity G VVS2	1	0.129359	0.021601	5.99
color*clarity H I1	1	0.601241	0.040304	14.92
color*clarity H IF	1	0.262763	0.027783	9.46
color*clarity H SI1	1	0.140546	0.024388	5.76
color*clarity H SI2	1	0.407516	0.025320	16.09
color*clarity H VS1	1	0.069605	0.024648	2.82
color*clarity H VS2	1	0.095742	0.024491	3.91
color*clarity H VVS1	1	0.125603	0.025476	4.93
color*clarity H VVS2	1	0.067932	0.022594	3.01
color*clarity I I1	1	0.751843	0.042465	17.70
color*clarity I IF	1	0.174073	0.030459	5.72
color*clarity I SI1	1	0.156933	0.025081	6.26
color*clarity I SI2	1	0.484146	0.026484	18.28
color*clarity I VS1	1	0.044825	0.025226	1.78
color*clarity I VS2	1	0.081611	0.025213	3.24
color*clarity I VVS1	1	0.063799	0.026694	2.39
color*clarity I VVS2	1	0.047815	0.023631	2.02
color*clarity J I1	1	0.881930	0.049343	17.87
color*clarity J IF	1	0.077069	0.037491	2.06
color*clarity J SI1	1	0.230365	0.022430	10.27
color*clarity J SI2	1	0.624123	0.023912	26.10
color*clarity J VS1	1	0.038732	0.023096	1.68

Tab. 9. The model results parameter

RESULTS AND DISCUSSIONS

In order to evaluate the model, we need to implement the feature engineering on the test data set as well. The model is validated based on the validation and the test data set.

Distribution Plot for Diamonds

Data Set	WORK.UNKNOWN_TESTDIAMONDS
Dependent Variable	logprice
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	SBC
Choose Criterion	Adj R-Sq
Effect Hierarchy Enforced	None

Number of Observations Read	19132
Number of Observations Used	16132

Class Level Information		
Class	**Levels**	**Values**
cut	5	Fair Good Ideal Premium Very Good
color	7	D E F G H I J
clarity	8	I1 IF SI1 SI2 VS1 VS2 VVS1 VVS2

Dimensions	
Number of Effects	11
Number of Parameters	173

Distribution Plot for Diamonds

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0	Intercept		1	1	0.0000	0.0000	16612.670	16612.671	476.833	573120.417	486.359	16617.8426	1.0300	0.00	1.0000
1	logcarat*clarity		2	9	0.9430	0.9430	-29583.223	-29583.209	-45725.118	17476.422	-45648.026	949.0578	0.0587	33338.8	<.0001
2	logcarat*color		3	15	0.9662	0.9662	-38025.713	-38025.679	-54163.451	3786.845	-54044.385	562.4929	0.0348	1850.52	<.0001
3	color*clarity		4	70	0.9707	0.9706	-40199.703	-40199.067	-56340.886	1271.210	-55795.504	492.1642	0.0302	44.42	<.0001
4	logcarat*cut		5	74	0.9719	0.9717	-40844.678	-40843.968	-56980.665	594.400	-56409.725	472.9556	0.0290	165.83	<.0001
5	cut		6	78	0.9725	0.9724*	-41219.808*	-41219.021*	-57352.360*	213.307*	-56754.100*	462.3219*	0.0283	96.46	<.0001

* Optimal Value of Criterion

Selection stopped at a local minimum of the SBC criterion.

Stop Details

Candidate For	Effect	Candidate SBC		Compare SBC
Entry	cut*color	-56644.583	>	-56754.100
Removal	cut	-56409.725	>	-56754.100

Tab. 10 The evaluation model

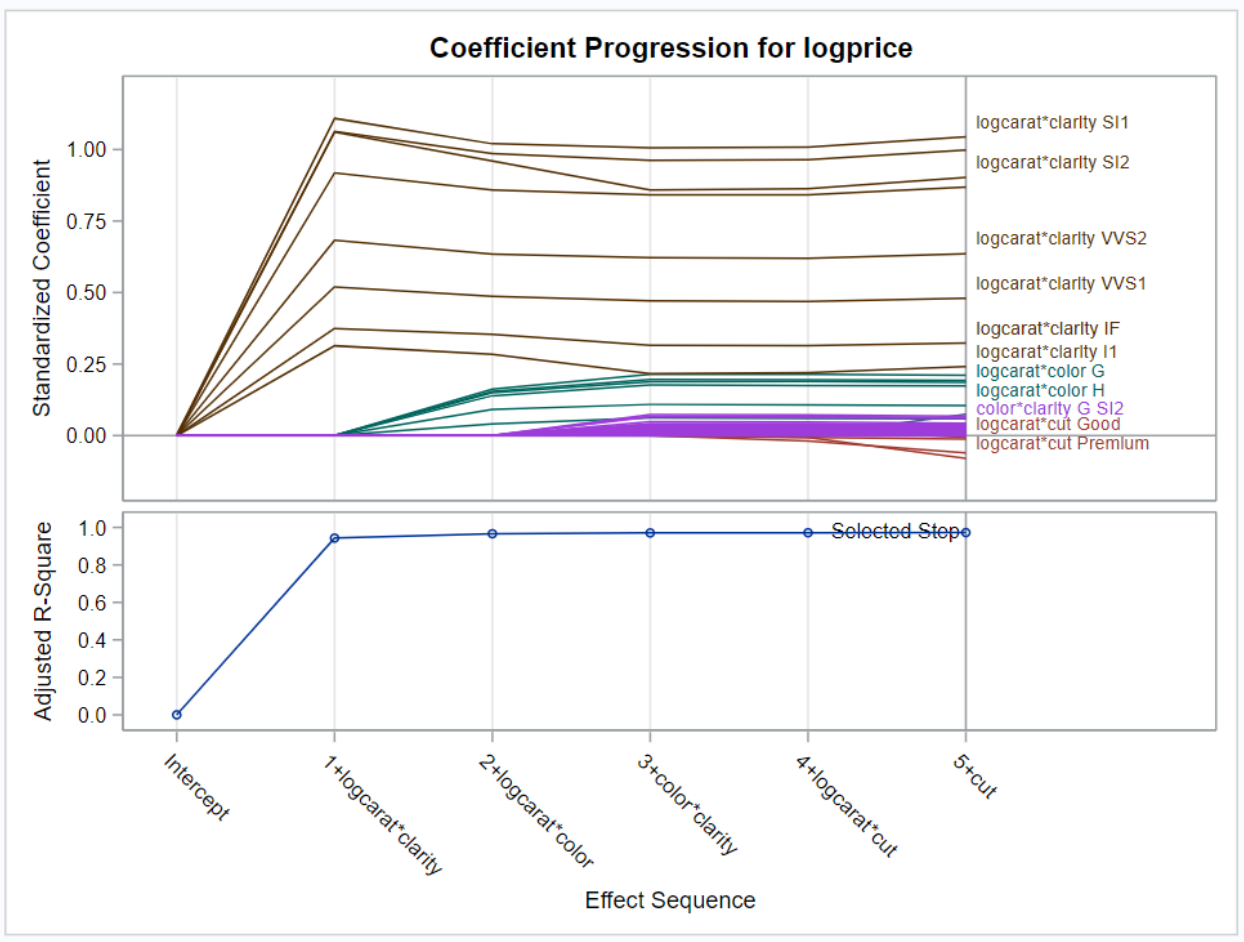


Fig. 21. Model Results- Coefficient Progression for logarice

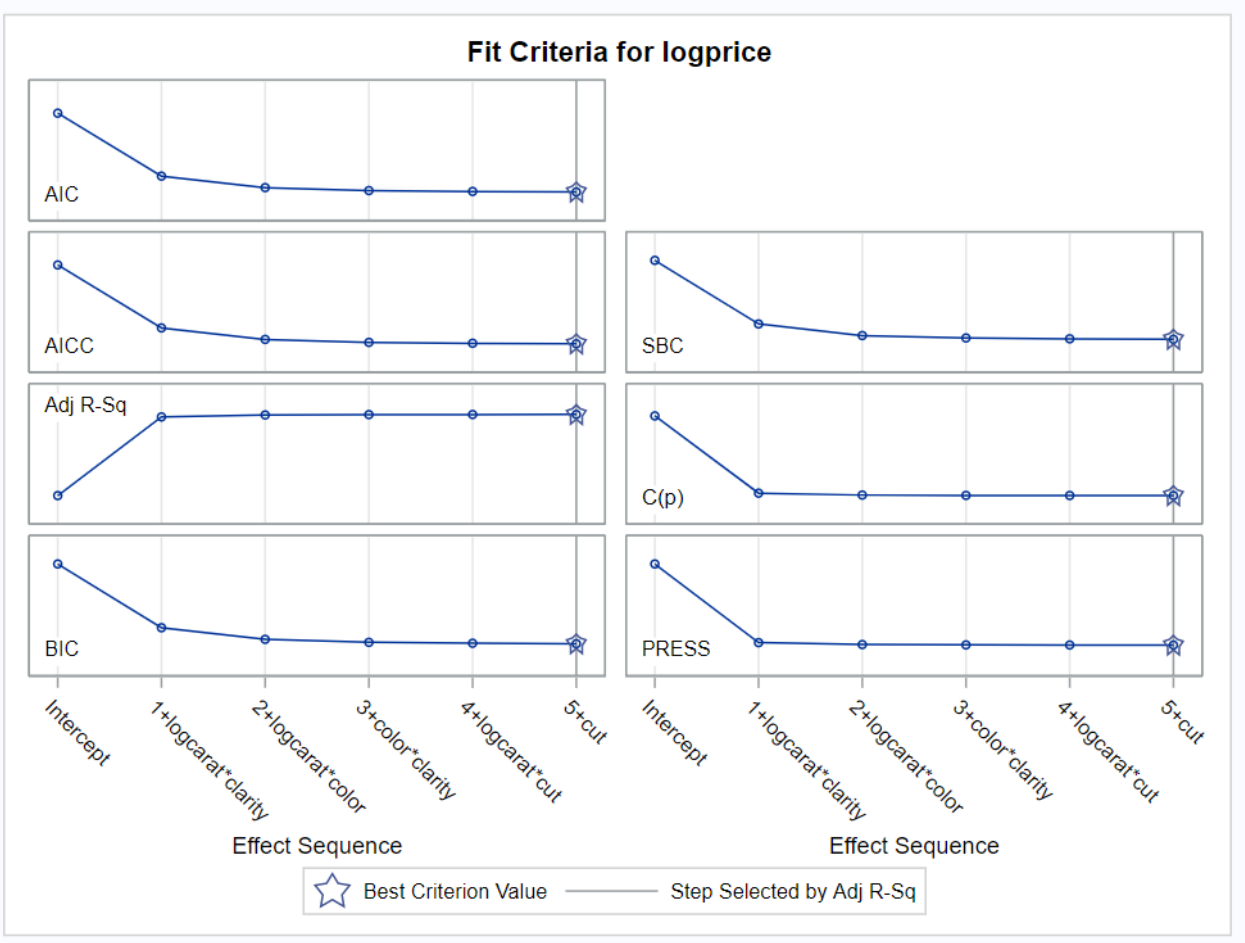


Fig. 22. Fit Criteria for logprice

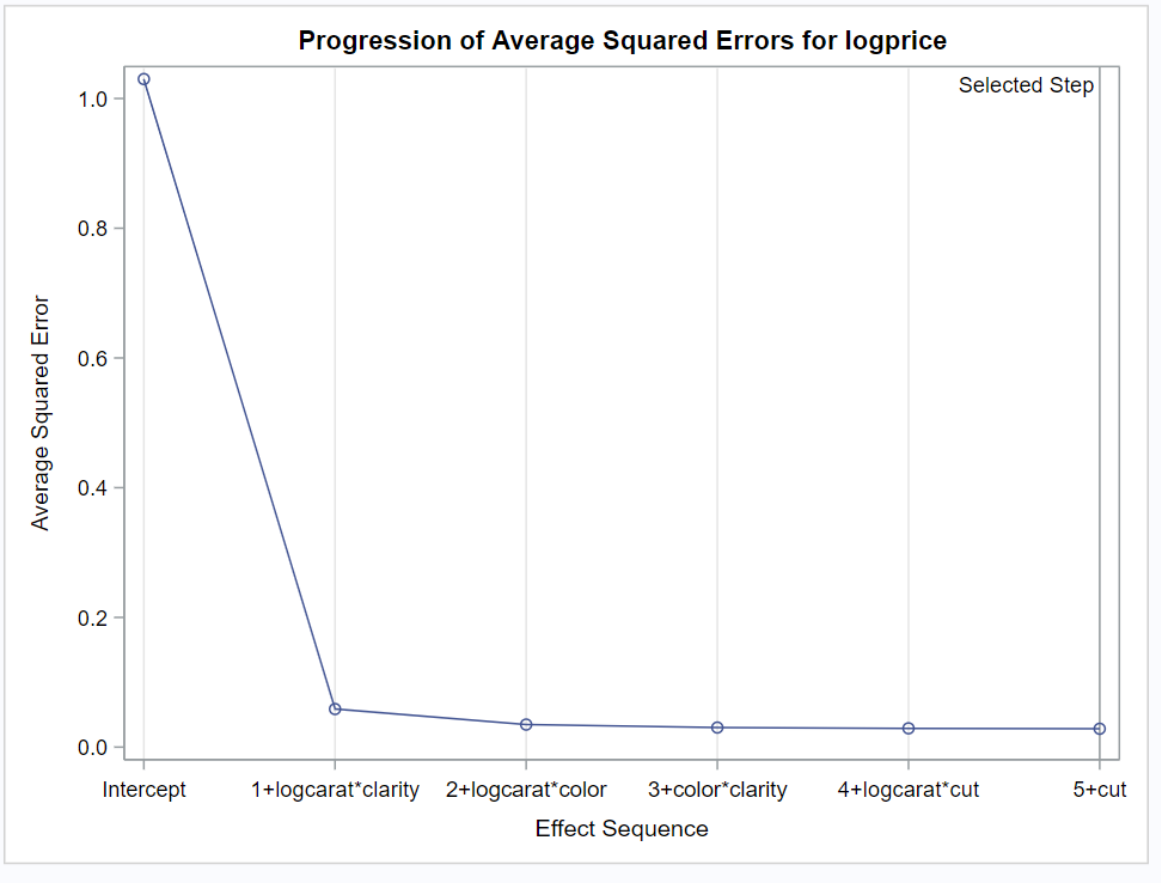


Fig. 23. Average Square Errors for logprice

The ASE plot visualizes the prediction accuracy of the models. The horizontal axis of the ASE plot shows how the models are formed from the previous model. The vertical axis tracks the change of the ASE for each successive model. The model-building process stops when it can no longer decrease the ASE on the validation data. For this example, that happens at Step=5.

Distribution Plot for Diamonds

Selected Model

The selected model, based on Adj R-Sq, is the model at Step 5.

Effects:	Intercept cut logcarat*cut logcarat*color logcarat*clarity color*clarity

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	77	16159	209.86070	7380.15	<.0001
Error	16054	456.50881	0.02844		
Corrected Total	16131	16616			

Root MSE	0.16863
Dependent Mean	7.79389
R-Square	0.9725
Adj R-Sq	0.9724
AIC	-41220
AICC	-41219
BIC	-57352
C(p)	213.30741
PRESS	462.32189
SBC	-56754
ASE	0.02830

Tab. 11 The evaluation model result

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	5.097932	0.034660	147.08	<.0001
cut Fair	1	0.238439	0.026233	9.09	<.0001
cut Good	1	0.020343	0.014163	1.44	0.1509
cut Ideal	1	0.072398	0.008819	8.21	<.0001
cut Premium	1	0.173979	0.009788	17.77	<.0001
cut Very Good	0	0	.	.	.
logcarat*cut Fair	1	-0.485966	0.037386	-13.00	<.0001
logcarat*cut Good	1	-0.065128	0.022428	-2.90	0.0037
logcarat*cut Ideal	1	-0.023682	0.014865	-1.59	0.1111
logcarat*cut Premium	1	-0.274924	0.015411	-17.84	<.0001
logcarat*cut Very Good	0	0	.	.	.
logcarat*color D	1	1.083718	0.030719	35.28	<.0001
logcarat*color E	1	1.038935	0.028709	36.19	<.0001
logcarat*color F	1	0.884209	0.028085	31.48	<.0001
logcarat*color G	1	0.714710	0.026753	26.72	<.0001
logcarat*color H	1	0.439254	0.027355	16.06	<.0001
logcarat*color I	1	0.265605	0.028113	9.45	<.0001
logcarat*color J	0	0	.	.	.
logcarat*clarity I1	1	2.442772	0.055556	43.97	<.0001
logcarat*clarity IF	1	4.284853	0.048151	88.99	<.0001
logcarat*clarity SI1	1	3.852379	0.027483	140.17	<.0001
logcarat*clarity SI2	1	3.275759	0.029224	112.09	<.0001
logcarat*clarity VS1	1	4.148536	0.028178	147.23	<.0001
logcarat*clarity VS2	1	4.030428	0.027649	145.77	<.0001
logcarat*clarity VVS1	1	4.460860	0.038421	116.10	<.0001
logcarat*clarity VVS2	1	4.363721	0.032737	133.30	<.0001
color*clarity D I1	1	0.387964	0.068678	5.65	<.0001
color*clarity D IF	1	0.753198	0.054349	13.86	<.0001
color*clarity D SI1	1	0.022194	0.036893	0.60	0.5475
color*clarity D SI2	1	0.127427	0.038041	3.35	0.0008
color*clarity D VS1	1	0.127951	0.038083	3.36	0.0008
color*clarity D VS2	1	0.118473	0.036702	3.23	0.0012
color*clarity D VVS1	1	0.264116	0.041236	6.41	<.0001
color*clarity D VVS2	1	0.175139	0.035487	4.94	<.0001
color*clarity E I1	1	0.394825	0.064003	6.17	<.0001
color*clarity E IF	1	0.379225	0.045389	8.35	<.0001
color*clarity E SI1	1	-0.004917	0.036711	-0.13	0.8934
color*clarity E SI2	1	0.159101	0.037811	4.21	<.0001
color*clarity E VS1	1	0.087489	0.037198	2.35	0.0187
color*clarity E VS2	1	0.074513	0.036401	2.05	0.0407

color*clarity E VVS1	1	0.157171	0.038035	4.13	<.0001
color*clarity E VVS2	1	0.119780	0.033945	3.53	0.0004
color*clarity F I1	1	0.362934	0.057204	6.34	<.0001
color*clarity F IF	1	0.326166	0.041124	7.93	<.0001
color*clarity F SI1	1	0.027421	0.037209	0.74	0.4612
color*clarity F SI2	1	0.212539	0.038135	5.57	<.0001
color*clarity F VS1	1	0.117171	0.037431	3.13	0.0017
color*clarity F VS2	1	0.092874	0.036623	2.54	0.0112
color*clarity F VVS1	1	0.181285	0.038233	4.74	<.0001
color*clarity F VVS2	1	0.141873	0.033841	4.19	<.0001
color*clarity G I1	1	0.482166	0.058933	8.18	<.0001
color*clarity G IF	1	0.299382	0.040162	7.45	<.0001
color*clarity G SI1	1	0.050599	0.037028	1.37	0.1718
color*clarity G SI2	1	0.260322	0.038134	6.83	<.0001
color*clarity G VS1	1	0.117655	0.037033	3.18	0.0015
color*clarity G VS2	1	0.107356	0.036850	2.91	0.0036
color*clarity G VVS1	1	0.146510	0.038059	3.85	0.0001
color*clarity G VVS2	1	0.130795	0.033178	3.94	<.0001
color*clarity H I1	1	0.642438	0.061176	10.50	<.0001
color*clarity H IF	1	0.283728	0.042474	6.68	<.0001
color*clarity H SI1	1	0.144053	0.037239	3.87	0.0001
color*clarity H SI2	1	0.432011	0.038743	11.15	<.0001
color*clarity H VS1	1	0.126367	0.037563	3.36	0.0008
color*clarity H VS2	1	0.123501	0.037261	3.31	0.0009
color*clarity H VVS1	1	0.163631	0.038591	4.24	<.0001
color*clarity H VVS2	1	0.080445	0.034361	2.34	0.0192
color*clarity I I1	1	0.748817	0.065959	11.35	<.0001
color*clarity I IF	1	0.159917	0.046597	3.43	0.0006
color*clarity I SI1	1	0.150246	0.037858	3.97	<.0001
color*clarity I SI2	1	0.464800	0.040049	11.61	<.0001
color*clarity I VS1	1	0.074572	0.038462	1.94	0.0525
color*clarity I VS2	1	0.083319	0.038311	2.17	0.0297
color*clarity I VVS1	1	0.063449	0.040731	1.56	0.1193
color*clarity I VVS2	1	0.026248	0.036107	0.73	0.4673
color*clarity J I1	1	1.020498	0.066556	15.33	<.0001
color*clarity J IF	1	0.186161	0.056474	3.30	0.0010
color*clarity J SI1	1	0.218085	0.034155	6.39	<.0001
color*clarity J SI2	1	0.662245	0.036134	18.33	<.0001
color*clarity J VS1	1	0.058879	0.034964	1.68	0.0922

Tab. 12 The evaluation model coefficient result

CONCLUSION AND RECOMMENDATIONS

With SAS using Proc GLMSELECT and Proc Log and following Data Science Process Flow we can predict the price of diamond with the accuracy above 97.5% . The Model scores are higher. The predicted prices are on the higher side with actual dataset.

1. If you were interested in a 1.5 carat diamond with a Very Good cut and a VS2 clarity rating, how much would the model predict you should pay for it?
   Logprice = 5.114068879 + 4.348410395 * logcarat + 0 * cut (very good) + 0.133169096 * clarity(VS2)
   Logprice = 5.114068879 + 4.348410395 * log(1.5) + 0 * 1 + 0.133169096 * 1
   Price = 10,214.34

2. What price do you recommend the jewelry company to bid?
   Total predicted price for 3000 diamonds is 12,261,458.95
   I recommend that the bid price should be 8,583,021.26

3. What strikes you about this comparison? After seeing this plot, do you feel confident in the model's ability to predict prices?
   Predicted prices are distributed linearly but old diamond prices are non-linear.

4. According to the model, if a diamond is 1 carat heavier than another with the same cut, how much more should I expect to pay? Why?
   For example for the 0.54 carat, D color, VS1 clarity and good cut we have e^(5.318127 + ln(.54+1)(-0.044844 + 0.871416 + 4.101698))-1
   which comes out to $1712.11277234 and for a 1.54 carat we get
   e^(5.318127 + ln(1.54+1)(-0.044844 + 0.871416 + 4.101698))-1
   and it'll be valued at $20171.6554478

APPENDIX A: SAS CODE

```
/*--------Importing the training and test data-----*/
/*--------------------------------*/
proc import datafile="C:\Users\fkiai\OneDrive\Desktop\DSPS\DIAMONDS\DATA\TRAINING\diamonds.csv"
     out=diamonds_raw
     dbms=csv
     replace;
     guessingrows=Max; /* or guessingrows=100 to improve load; performance */
     getnames=yes;     /* datarow=1 if no header name */
run;
/*labeling the indexes with rec_id and move to fkiaie lib*/
Data fkiaie.diamonds;
     Format rec_id BEST12.;
     Set diamonds_raw;
     rec_id = input(VAR1, BEST12.);
     Drop VAR1;
Run;

proc import datafile="C:\Users\fkiai\OneDrive\Desktop\DSPS\DIAMONDS\DATA\TEST\new-diamonds.csv"
     out=testdiamonds_raw
     dbms=csv
     replace;
     guessingrows=Max; /* or guessingrows=100 to improve load; performance */
     getnames=yes;     /* datarow=1 if no header name */
run;
/*labeling the indexes with rec_id and move to fkiaie lib*/
Data fkiaie.testdiamonds;
     Format rec_id BEST12.;
     Set testdiamonds_raw;
     rec_id = input(VAR1, BEST12.);
     Drop VAR1;
Run;

/*-----------Data Cleaning-------*/
/*--------------------------------*/
/***remove duplicate record and separate to dup_diamonds*/
/**exception report1 */
proc sort data=fkiaie.diamonds(drop=rec_id) nodupkey
out=sorted_daimonds
dupout=dup_diamonds;/*separate to dup_diamonds*/
by _all_;
run;

/*using SQL to create duplication**/
proc sql;
Create table fkiaie.diamins_sql_dup as
select carat, cut, color,
           clarity, depth, table,
           price, x, y, z, count(*) as dupcount
from fkiaie.diamonds
group by carat, cut, color,
                 clarity, depth, table,
                 price, x,y,z
having count(*) >1
order by dupcount desc
;
quit;

/*next step is joining this result with the diamond */
PROC SQL;
   CREATE TABLE WORK.QUERY_FOR_DIAMINS_SQL_DUP_0000 AS
   SELECT t2.rec_id,
          t2.carat,
          t2.cut,
          t2.color,
          t2.clarity,
          t2.depth,
          t2.table,
          t2.price,
          t2.x,
          t2.y,
          t2.z
      FROM FKIAIE.DIAMINS_SQL_DUP t1, FKIAIE.DIAMONDS t2
      WHERE (t1.carat = t2.carat AND t1.cut = t2.cut AND t1.color = t2.color AND t1.clarity = t2.clarity AND
t1.depth =
```

```
          t2.depth AND t1.table = t2.table AND t1.price = t2.price AND t1.x = t2.x AND t1.y = t2.y AND t1.z =
t2.z);
QUIT;

/*Query to find No Duplicated and distinct*/
PROC SQL;
   CREATE TABLE WORK.QUERY_NO_DUPLICATED AS
   SELECT DISTINCT t1.carat,
          t1.cut,
          t1.color,
          t1.clarity,
          t1.depth,
          t1.table,
          t1.price,
          t1.x,
          t1.y,
          t1.z
      FROM FKIAIE.DIAMONDS t1;
QUIT;

/***remove invalid data (x=y=z=0)***/
/**exception report 2 */
PROC SQL;
   CREATE TABLE WORK.QUERY_FOR_QUERY_NO_DUPLICATED AS
   SELECT t1.carat,
          t1.cut,
          t1.color,
          t1.clarity,
          t1.depth,
          t1.table,
          t1.price,
          t1.x,
          t1.y,
          t1.z
      FROM WORK.QUERY_NO_DUPLICATED t1
      WHERE t1.x = 0 OR t1.y = 0 OR t1.z = 0;
QUIT;

PROC SQL;
   CREATE TABLE WORK.QUERY_FOR_QUERY_NO_DUPLICAT_0001 AS
   SELECT t2.rec_id,
          t2.carat,
          t2.cut,
          t2.color,
          t2.clarity,
          t2.depth,
          t2.table,
          t2.price,
          t2.x,
          t2.y,
          t2.z
      FROM WORK.QUERY_FOR_QUERY_NO_DUPLICATED t1, FKIAIE.DIAMONDS t2
      WHERE (t1.carat = t2.carat AND t1.cut = t2.cut AND t1.color = t2.color AND t1.clarity = t2.clarity AND
t1.depth =
          t2.depth AND t1.table = t2.table AND t1.price = t2.price AND t1.x = t2.x AND t1.y = t2.y AND t1.z =
t2.z);
QUIT;

/*Removing the not valid and duplicate to get clean*/
PROC SQL;
   CREATE TABLE WORK.QUERY_FOR_QUERY_NO_DUPLICAT_0004 AS
   SELECT t1.carat,
          t1.cut,
          t1.color,
          t1.clarity,
          t1.depth,
          t1.table,
          t1.price,
          t1.x,
          t1.y,
          t1.z
      FROM WORK.QUERY_NO_DUPLICATED t1
      WHERE t1.x > 0 AND t1.y > 0 AND t1.z > 0;
QUIT;

/***remove missing values***/
PROC SQL;
   CREATE TABLE WORK.QUERY_FOR_SORTED_DAIMONDS AS
   SELECT /* NMISS_DISTINCT_of_carat */
            (NMISS(DISTINCT(t1.carat))) AS NMISS_DISTINCT_of_carat,
          /* NMISS_of_cut */
```

```
            (NMISS(t1.cut)) AS NMISS_of_cut,
          /* NMISS_of_color */
            (NMISS(t1.color)) AS NMISS_of_color,
          /* NMISS_of_clarity */
            (NMISS(t1.clarity)) AS NMISS_of_clarity,
          /* NMISS_DISTINCT_of_depth */
            (NMISS(DISTINCT(t1.depth))) AS NMISS_DISTINCT_of_depth,
          /* NMISS_of_table */
            (NMISS(t1.table)) AS NMISS_of_table,
          /* NMISS_of_price */
            (NMISS(t1.price)) AS NMISS_of_price,
          /* NMISS_of_x */
            (NMISS(t1.x)) AS NMISS_of_x,
          /* NMISS_of_y */
            (NMISS(t1.y)) AS NMISS_of_y,
          /* NMISS_of_z */
            (NMISS(t1.z)) AS NMISS_of_z
      FROM WORK.SORTED_DAIMONDS t1;
QUIT;
/*Results show there is no missing value to remove*/

/**Visualizing Missing**/
proc format ;
    value $missfmt ' '="Missing" other="Not Missing";
    value nmissfmt .  ="Missing" other="Not Missing";
run;

%MACRO Gen_MissingVariable_Bar(ds=);

        ods exclude all;
        ods output onewayfreqs=temp;
        title;

        Proc freq data=&ds;
                tables _all_ / MISSING;
                format _numeric_ nmissfmt. _character_ $missfmt.;
        Run;

        Data Wanted;
            length variable $32. variable_value $50.;
            set temp;
            Variable=scan(table, 2);
            Variable_Value=strip(trim(vvaluex(variable)));
            keep variable variable_value frequency percent;
            label variable='Variable'
                  variable_value='Variable Value';
        Run;

        Data Missing_Report(keep=variable NMiss pct_NMiss);
            do until (last.variable);
                set Wanted;
                by variable notsorted;
                select (variable_value);
                        when ('Missing')
                            do;
                               NMiss = frequency;
                               pct_NMiss = percent;
                            end;
                        when ('Not Missing')
                            do;
                               NN = frequency;
                               pct_NN = percent;
                            end;
                        Otherwise;
                end;
            end;
            NMiss = coalesce(NMiss, 0);
            pct_NMiss = coalesce(pct_NMiss, 0);
            NN = coalesce(NN, 0);
            pct_NN = coalesce(pct_NN, 0);

            Label pct_NMIss = "% Missing";
        Run;

        %let softgreen=cx8faf7f;
        ods exclude none;
        ods graphics on / height=5in width=10in;

        title "Missing Values by Variable Visualization";
        Proc sgplot data=Missing_report;
            Format pct_NMiss PERCENT8.4;
```

```
            Format NMiss COMMA16.;
            vbar variable / response=pct_NMiss
                            dataskin=sheen
                            /*other option: gloss, matte */
                            barwidth=0.8
                            fillattrs=(color=green)
                            datalabel=pct_NMiss;
            xaxis discreteorder=data;
            yaxis display=(noline) grid;

        run;

%MEND Gen_MissingVariable_Bar;

%Gen_MissingVariable_Bar(ds=fkiaie.diamonds);
/**no missing, therefore nothing in the graph**/
ods _all_ close;

/*------------------EDA----------------*/
/*-------------------------------------*/
/**continues and categorical relation**/
ODS graphics on /reset width=12in height=5in;

title 'Diamonds Price by Cut';

proc sgplot data=fkiaie.Cleaned_data;
    vbox price / category=cut
                 dataskin=sheen
                 outlierattrs=(color=green)
                 meanattrs=(color=black)
                 medianattrs=(color=black)
                 connect=mean connectattrs=(color=red);
run;

title 'Diamonds Price by Color';
proc sgplot data=fkiaie.Cleaned_data;
    vbox price / category=color
                 dataskin=sheen
                 outlierattrs=(color=green)
                 meanattrs=(color=black)
                 medianattrs=(color=black)
                 connect=mean connectattrs=(color=blue);
run;

title 'Diamonds Price by Clarity';
proc sgplot data=fkiaie.Cleaned_data;
    vbox price / category=clarity
                 dataskin=sheen
                 outlierattrs=(color=green)
                 meanattrs=(color=black)
                 medianattrs=(color=black)
                 connect=mean connectattrs=(color=black);
run;

title 'Diamonds Price by Carat';
proc sgplot data=fkiaie.Cleaned_data;
    vbox price / category=carat
                 dataskin=sheen
                 nooutliers
                 connect=mean connectattrs=(color=green);
    xaxis display=(novalues);
run;

/** Plotting the relationship**/

ods graphics / reset width=8in height=5in;
title 'Diamonds Price Proportion by Carat';

proc sgpie data=fkiaie.Cleaned_data;
    donut carat / response=price
                  datalabeldisplay=all
                  datalabelloc=outside
                  sliceorder=respdesc;
run;


ods graphics / reset width=8in height=5in;
title 'Diamonds Price Proportion by Cut';

proc sgpie data=fkiaie.Cleaned_data;
    donut cut / response=price
```

```
                datalabeldisplay=all
                datalabelloc=outside
                sliceorder=respdesc;
run;

ods graphics / reset width=6in height=3in;
title 'Diamonds Price Proportion by Color';

proc sgpie data=fkiaie.Cleaned_data;
    donut color / response=price
                  datalabeldisplay=all
                  datalabelloc=outside
                  sliceorder=respdesc;
run;


/***categorical and categorical relations**/
ODS graphics on;

proc freq data=fkiaie.Cleaned_data;
    tables cut*(color clarity) / chisq plots=freqplot(orient=horizontal twoway=stacked scale=percent);
    tables color*clarity / chisq plots=freqplot(orient=horizontal twoway=stacked scale=percent);

Run;

/**Heat map and Correlations_Type2**/
%macro prepCorrData(in=,out=);
  /* Run corr matrix for input data, all numeric vars */
  proc corr data=&in. noprint
    pearson
    outp=work._tmpCorr
    vardef=df
  ;
  run;

  /* prep data for heat map */
data &out.;
  keep cartesian_x cartesian_y r;
  set work._tmpCorr(where=(_TYPE_="CORR"));
  array v{*} _numeric_;
  cartesian_x = _NAME_;
  do i = dim(v) to 1 by -1;
    cartesian_y = vname(v(i));
    r = v(i);
    /* creates a lower triangular matrix */
    if (i<_n_) then
      r=.;
    output;
  end;
run;
/*
proc datasets lib=work nolist nowarn;
  delete _tmpcorr;
quit;
*/
%mend;

  /* Create a heat map implementation of a correlation matrix */
ods path work.mystore(update) sashelp.tmplmst(read);

proc template;
  define statgraph corrHeatmap;
   dynamic _Title;
    begingraph;
      entrytitle _Title;
      rangeattrmap name='map';
      /* select a series of colors that represent a "diverging"  */
      /* range of values: stronger on the ends, weaker in middle */
      /* Get ideas from http://colorbrewer.org                   */
      range -1 - 1 / rangecolormodel=(cxD8B365 cxF5F5F5 cx5AB4AC);
      endrangeattrmap;
      rangeattrvar var=r attrvar=r attrmap='map';
      layout overlay /
        xaxisopts=(display=(line ticks tickvalues))
        yaxisopts=(display=(line ticks tickvalues));
        heatmapparm x = cartesian_x y = cartesian_y colorresponse = r /
          xbinaxis=false ybinaxis=false
          name = "heatmap" display=all;
        continuouslegend "heatmap" /
          orient = vertical location = outside title="Pearson Correlation";
      endlayout;
```

```
    endgraph;
  end;
run;

/* Build the graphs */
ODS graphics on / reset height=600 width=800;

%prepCorrData(in=fkiaie.Cleaned_data(drop=cut color clarity),
              out=Diamonds_r);
proc sgrender data=Diamonds_r template=corrHeatmap;
   dynamic _title= "Correlation matrix for Diamonds";
run;

ODS graphics off;

/**Histogram**/
ODS graphics on / reset width=8in height=5in;

ODS noproctitle;
Title;

*ODS select histogram ParameterEstimates GoodnessOfFit FitQuantiles;
ODS select histogram;
Title 'Distribution Plot for Diamonds';

proc univariate data=fkiaie.Cleaned_data;
    var price carat depth table x y z;

    histogram price carat depth table x y z / kernel;
    inset Min Max mean median std Q1 Q3 skewness kurtosis / pos=ne;
Run;

/*---------Feature Engineering------------*/
/*------------------------------------*/

/* Apply Feature Engineering and Log Transformation */
Data fkiaie.Diamonds_FE (drop=i j k);
    Set fkiaie.Cleaned_data;
    logprice = log(1+price);
    logcarat = log(1+carat);

    Array cut_grade[5]     $9 _temporary_ ('Fair' 'Good' 'Very Good' 'Premium' 'Ideal');
    Array color_grade[7]   $1 _temporary_ ('J' 'I' 'H' 'G' 'F' 'E' 'D');
    Array clarity_grade[8] $4 _temporary_ ('I1' 'SI2' 'SI1' 'VS2' 'VS1' 'VVS2' 'VVS1' 'IF');


    Array dummy_cut[*] cut1 - cut5;
    Array dummy_color[*] color1 - color7;
    Array dummy_clarity[*] clarity1 - clarity8;

    Do i = 1 to dim(dummy_cut);
        Do j = 1 to dim(dummy_color);
            Do k = 1 to dim(dummy_clarity);
                dummy_cut(i) = 0;
                dummy_color(j) = 0;
                dummy_clarity(k) = 0;
            End;
        End;
    End;

    Do i = 1 to dim(cut_grade);
       If cut_grade[i] = Strip(cut) Then dummy_cut[i] = 1;
    End;

    Do i = 1 to dim(color_grade);
       If color_grade[i] = Strip(color) Then dummy_color[i] = 1;
    End;

    Do I = 1 to dim(clarity_grade);
       If clarity_grade[i] = Strip(clarity) Then dummy_clarity[i] = 1;
    End;

    Select (cut);
        when ('Fair')      cut_ord = 1;     /* Lowest level of fire and brilliance */
        when ('Good')      cut_ord = 2;
        when ('Very Good') cut_ord = 3;
        when ('Premium')   cut_ord = 4;
        when ('Ideal')     cut_ord = 5;     /* Highest level of fire and brilliance */
        otherwise
        ;
    End;
```

```
    Select (color);
        when ('J') color_ord = 1;
        when ('I') color_ord = 2;
        when ('H') color_ord = 3;
        when ('G') color_ord = 4;           /* G-J = Nearly Colorless */
        when ('F') color_ord = 5;
        when ('E') color_ord = 6;
        when ('D') color_ord = 7;           /* D-F = Colorless is highest color grade */
        otherwise
        ;
    End;

    Select (clarity);
        when ('I1')   clarity_ord = 1;      /* Inclusions 1 is the worst */
        when ('SI2')  clarity_ord = 2;      /* Small Inclusions 1 */
        when ('SI1')  clarity_ord = 3;      /* Small Inclusions 2 */
        when ('VS2')  clarity_ord = 4;      /* Very Small Inclusions 1 */
        when ('VS1')  clarity_ord = 5;      /* Very Small Inclusions 2 */
        when ('VVS2') clarity_ord = 6;      /* Very Very Small Inclusions 1 */
        when ('VVS1') clarity_ord = 7;      /* Very Very Small Inclusions 2 */
        when ('IF')   clarity_ord = 8;      /* Internally Flawless is the best */
        otherwise
        ;
    End;
Run;



/*----------------Splitting------------*/
/*-------------------------------------*/
proc surveyselect data=fkiaie.diamonds_FE
    out=Diamonds_Train_Valid
    method=SRS
    samprate=0.7     /* Wanted Training Dataset 70% */
    seed=1357924
    outall;
run;

Data fkiaie.DiamondsFE_Training fkiaie.DiamondsFE_Validation;
    Set Diamonds_Train_Valid;
    If Selected Then
        output fkiaie.DiamondsFE_Training;
    Else
        output fkiaie.DiamondsFE_Validation;
Run;


/*----------------Modeling------------*/
/*-------------------------------------*/
/**building the model**/
ODS graphics on;
Proc GLMSELECT data=fkiaie.DiamondsFE_Training
               outdesign=DiamondsFE_Design;

     Class Cut Color Clarity;
     Model logprice = logcarat logcarat|cut|color|clarity @2;

Run;

Proc REG Data=DiamondsFE_Design;
     ODS Output ParameterEstimates=fkiaie.Model_C_Parameters;
     MODEL_C: Model logprice = &_GLSMOD;

Run;
ODS graphics off;


/**Apply Feature Engineering To Test Diamonds**/
/*
   -Natural Log Transformation for Carat
   -One Hot Encoding for Categorical Variables
   -Ordinal Variable for Categorical Variables
*/
Data fkiaie.TestDiamonds_FE (drop=i j k);
    Set fkiaie.testdiamonds;
    logcarat = log(1+carat);

    Array cut_grade[5]     $9 _temporary_ ('Fair' 'Good' 'Very Good' 'Premium' 'Ideal');
    Array color_grade[7]   $1 _temporary_ ('J' 'I' 'H' 'G' 'F' 'E' 'D');
    Array clarity_grade[8] $4 _temporary_ ('I1' 'SI2' 'SI1' 'VS2' 'VS1' 'VVS2' 'VVS1' 'IF');
```

```
    Array dummy_cut[*] cut1 - cut5;
    Array dummy_color[*] color1 - color7;
    Array dummy_clarity[*] clarity1 - clarity8;

    Do i = 1 to dim(dummy_cut);
        Do j = 1 to dim(dummy_color);
            Do k = 1 to dim(dummy_clarity);
                dummy_cut(i) = 0;
                dummy_color(j) = 0;
                dummy_clarity(k) = 0;
            End;
        End;
    End;

    Do i = 1 to dim(cut_grade);
       If cut_grade[i] = Strip(cut) Then dummy_cut[i] = 1;
    End;

    Do i = 1 to dim(color_grade);
       If color_grade[i] = Strip(color) Then dummy_color[i] = 1;
    End;

    Do I = 1 to dim(clarity_grade);
       If clarity_grade[i] = Strip(clarity) Then dummy_clarity[i] = 1;
    End;

    Select (cut);
        when ('Fair')      _cut_ord = 1;     /* Lowest level of fire and brilliance */
        when ('Good')      _cut_ord = 2;
        when ('Very Good') _cut_ord = 3;
        when ('Premium')   _cut_ord = 4;
        when ('Ideal')     _cut_ord = 5;     /* Highest level of fire and brilliance */
        otherwise
        ;
    End;

    Select (color);
        when ('J') _color_ord = 1;
        when ('I') _color_ord = 2;
        when ('H') _color_ord = 3;
        when ('G') _color_ord = 4;           /* G-J = Nearly Colorless */
        when ('F') _color_ord = 5;
        when ('E') _color_ord = 6;
        when ('D') _color_ord = 7;           /* D-F = Colorless is highest color grade */
        otherwise
        ;
    End;

    Select (clarity);
        when ('I1')   _clarity_ord = 1;      /* Inclusions 1 is the worst */
        when ('SI2')  _clarity_ord = 2;      /* Small Inclusions 1 */
        when ('SI1')  _clarity_ord = 3;      /* Small Inclusions 2 */
        when ('VS2')  _clarity_ord = 4;      /* Very Small Inclusions 1 */
        when ('VS1')  _clarity_ord = 5;      /* Very Small Inclusions 2 */
        when ('VVS2') _clarity_ord = 6;      /* Very Very Small Inclusions 1 */
        when ('VVS1') _clarity_ord = 7;      /* Very Very Small Inclusions 2 */
        when ('IF')   _clarity_ord = 8;      /* Internally Flawless is the best */
        otherwise
        ;
    End;
Run;



/**model prediction and evaluation**/
ODS graphics on;

/* Prepare Unknown dataset for prediction */
Data DiamondsFE_Validation (keep=DS logprice logcarat cut color clarity);
    Retain DS logprice logcarat cut color clarity;
    Set fkiaie.DiamondsFE_Validation;
    Format DS $10.;
    Format cut $11.;
    Format color $3.;
    Format clarity $6.;

    DS = "VALIDATION";
Run;
```

```
Data TestDiamonds_FE (Keep=DS logcarat carat cut color clarity);
    Retain DS logcarat carat cut color clarity;
    Set fkiaie.TestDiamonds_FE;
    Format DS $10.;
    Format cut $11.;
    Format color $3.;
    Format clarity $6.;

    DS = "TEST";
Run;

Data Unknown_TestDiamonds;
    Set DiamondsFE_Validation TestDiamonds_FE;
Run;

Proc GLMSELECT data=Unknown_TestDiamonds
     PLOTS=ALL;
     Class Cut Color Clarity;
     Model logprice = logcarat|cut|color|clarity @2 / choose =adjrsq
                                                      showpvalues
                                                      stats=all;
     output out=Diamonds_Predicted_Model_C
            predicted=Predicted_Price_C;
Run;

ODS Graphics off;

Data Diamonds_Predicted_Model_C (drop=logprice logcarat);
    Set Diamonds_Predicted_Model_C;
    If DS = "TEST";
    Predicted_Price_C = exp(Predicted_Price_C)-1;
Run;

Proc Sort Data=Diamonds_Predicted_Model_C;
    By carat clarity color cut;
Run;
```