Building a Multiple Linear Regression Model to predict wholesale prices of diamonds

Data Preparation

The diamonds data set is extracted from the ICE online store database during one-week period in March of a recent year.

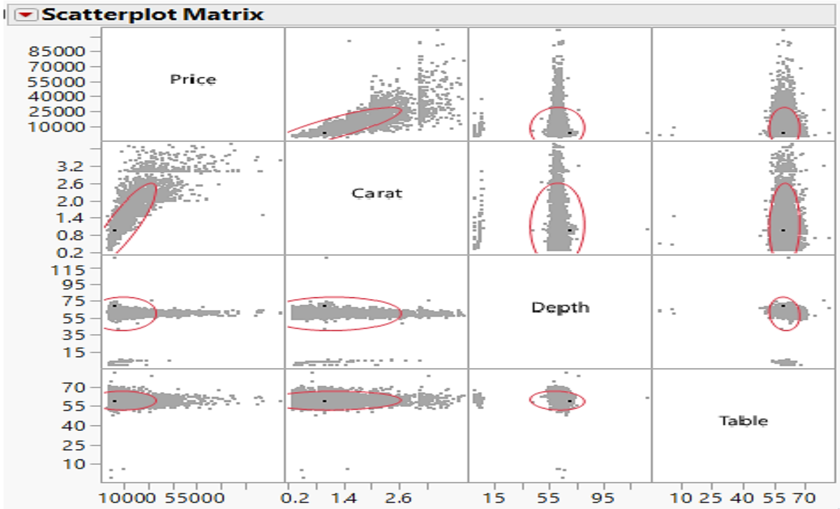
Data modeling types

Price, Carat, Table and Depth – Continuous Variables.
Clarity, Color - Ordinal
Lab, Dealer - Nominal.

Analysis of each Variable:

Price - Ranges between 0 - $35000 but for row number 1298 the price is $105876.
Since the diamond is of 3.39 carts, color E, depth and table very close to 60 the price may not be an outlier. The column is right skewed. No missing values found.
Carat - Carat is below 4 for all values. It is right skewed. No missing values found.
Clarity- The categories used here are (IF, VVS1, VVS2, VS1, VS2, SI1, SI2, I1).
Highest density- SL1, lowest I1. No missing values found.
Color - Categories used here are (K, J, I, H, G, F, E, D)
highest H colored, lowest K colored. No missing values found.
Depth - Outlier at row 10009 where depth is 126.3. Deleted this value and imputed.
10084, 8628, 12580, 1515, 5490, 11324, 7205, 2703, 9671 - all these rows have depth 0 which is not meaningful. So, removing these values from table and imputing along with missing values.
Table - Table also has one row with zero value in row number 11308, so removed that 0 value as well. It's been imputed with the average table value along with other missing values. It has 2180 missing values initially in this column.

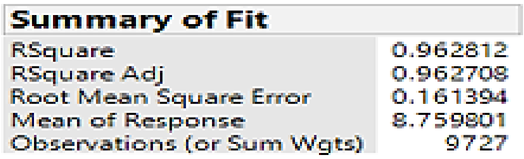
Multi-variate relationships



Carat and Price are highly correlated here. Price values show dependency on the variable 'carat'.

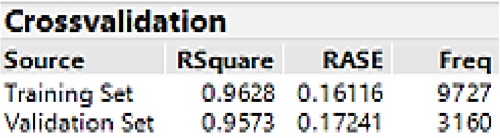
Model comparisons

Models	**Variables**	**Training &Validation R^2 values**	**Method**	**Comments**
Model A	Y-price X-carat, depth, table	T-values-0.7446 V-values-0.7340	Standard least squares	Low Validation value with heteroskedasticity
Model B	Y-log(price) X-carat, depth, table	T-values-0.7397 V-values-0.7373	Standard least squares	Reduced heteroskedasticity
Model C	Y-log(price) X-color, clarity carat, depth, table	T-values-0.8143 V-values-0.8206	Standard least squares	Finds overfitting
Model E	Y-log(price) X-color, clarity, carat, depth, table	T-values-0.8143 V-value-0.8206	Forward stepwise	No overfitting
Model F	Y-log(price) X-All metric &predictor variables	T-values-0.8624 V-values-0.8750	Forward stepwise	No overfitting with high R square value
Chosen Model G	Y-log(price) X-log(carat), clarity, color, clarity*color, depth, table, lab	T values-0.9628 V values-0.9573	Forward Stepwise	Higher validation R square value



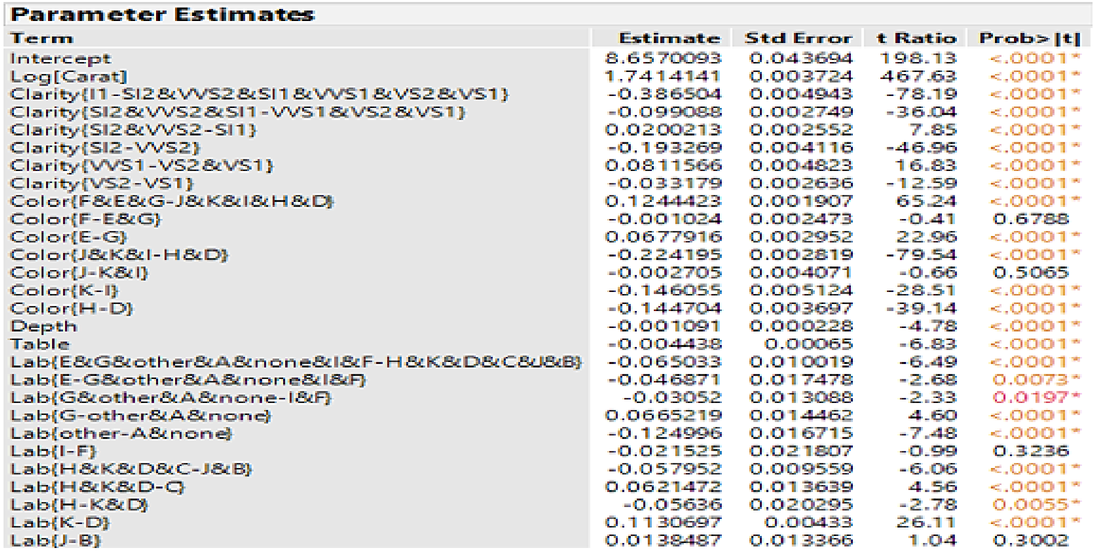
Summary of Fit

RSquare	0.962812
RSquare Adj	0.962708
Root Mean Square Error	0.161394
Mean of Response	8.759801
Observations (or Sum Wgts)	9727



Crossvalidation

Source	RSquare	RASE	Freq
Training Set	0.9628	0.16116	9727
Validation Set	0.9573	0.17241	3160



Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	8.6570093	0.043694	198.13	<.0001*
Log[Carat]	1.7414141	0.003724	467.63	<.0001*
Clarity{I1-SI2&VVS2&SI1&VVS1&VS2&VS1}	-0.386504	0.004943	-78.19	<.0001*
Clarity{SI2&VVS2&SI1-VVS1&VS2&VS1}	-0.099088	0.002749	-36.04	<.0001*
Clarity{SI2&VVS2-SI1}	0.0200213	0.002552	7.85	<.0001*
Clarity{SI2-VVS2}	-0.193269	0.004116	-46.96	<.0001*
Clarity{VVS1-VS2&VS1}	0.0811566	0.004823	16.83	<.0001*
Clarity{VS2-VS1}	-0.033179	0.002636	-12.59	<.0001*
Color{F&E&G-J&K&I&H&D}	0.1244423	0.001907	65.24	<.0001*
Color{F-E&G}	-0.001024	0.002473	-0.41	0.6788
Color{E-G}	0.0677916	0.002952	22.96	<.0001*
Color{J&K&I-H&D}	-0.224195	0.002819	-79.54	<.0001*
Color{J-K&I}	-0.002705	0.004071	-0.66	0.5065
Color{K-I}	-0.146055	0.005124	-28.51	<.0001*
Color{H-D}	-0.144704	0.003697	-39.14	<.0001*
Depth	-0.001091	0.000228	-4.78	<.0001*
Table	-0.004438	0.00065	-6.83	<.0001*
Lab{E&G&other&A&none&I&F-H&K&D&C&J&B}	-0.065033	0.010019	-6.49	<.0001*
Lab{E-G&other&A&none&I&F}	-0.046871	0.017478	-2.68	0.0073*
Lab{G&other&A&none-I&F}	-0.03052	0.013088	-2.33	0.0197*
Lab{G-other&A&none}	0.0665219	0.014462	4.60	<.0001*
Lab{other-A&none}	-0.124996	0.016715	-7.48	<.0001*
Lab{I-F}	-0.021525	0.021807	-0.99	0.3236
Lab{H&K&D&C-J&B}	-0.057952	0.009559	-6.06	<.0001*
Lab{H&K&D-C}	0.0621472	0.013639	4.56	<.0001*
Lab{H-K&D}	-0.05636	0.020295	-2.78	0.0055*
Lab{K-D}	0.1130697	0.00433	26.11	<.0001*
Lab{J-B}	0.0138487	0.013366	1.04	0.3002