Final Report

Prashant Karki*

STAT - 450

April 27, 2020

* Minnesota, USA.

Abstract

We are looking at an issue of predicting our dependent variables from the same sets of regressor variables. We are starting with looking at the nature of data, performing Exploratory Data Analysis (EDA) to find the inconsistencies and to see the relationship between regressors and response variable using multiple visualization techniques. Also, we are solving how we can enhance the predictive performance of our model compared to the model generated with the regular process of doing individual regressions of each response variable on the common set of regressors. In terms of model selection, stepwise regression for robust model was introduced and implemented to our existing model to enhance the prediction accuracy. Each step was implemented and compared to see how models in different steps are acting and to observe how the changes have had occurred during this model selection process. This project also makes sure that the Multiple Linear Regression (MLR) assumptions are met and significant tests were performed for all regressors to pick only significant regressors.

Introduction

Multiple Linear Regression (MLR) is one of the most common regression analyses performed in the field of data science and statistics. MLR as a predictive analysis attempts to generate the model between one response variable (y) and one or many regressors or independent variables (x/x_i) by fitting straight line or simply linear equation to the observation points of the dataset used. In simple terms, MLR is the extension of PLS (ordinary least squares) regression that consists more than one regressor variables. In MLR each observation of independent variable (x) is associated with response variable (y). Let's consider MLR with i regressors x_i (where i - 1,2,3..) and 1 response variable (y) in the model. The MLR model can be expressed as follows,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \epsilon$$

Where,

- y = Dependent Variable
- x_i = Regressor Variables
- β_0 = y-intercept (constant term)
- β_i = Estimated parameters or regressor coefficients for each regressors
- ϵ = Error term (Residuals)

A simple linear regression allows the analysts to make predictions about one variable based on the information available for another variable. Linear regression is possible only when there are regressor and response variables which are on continuous form. Here response variable is the one that

model is predicting and the regressor is the variable that is being used to predict the response. Extension of simple linear regression with more than on regressor variable is Multiple Linear Regression. Multiple Linear Regressions are based on the assumptions below:

- **Linear Relationship** - There must be linear relationship between response variable and independent variable.
- **Homoscedasticity** - the variance of the residuals needs to be constant or same throughout each level of explanatory variables.
- **Multivariate Normality** - MLR assumes that residuals are normally distributed with mean of 0 and variance of σ.
- **No Multicollinearity** - assumes that there should not be high correlation between independent variables.

Dataset Description

I have used this classic diamond dataset that contains the price and other attributes of almost 54,000 diamonds. I believe it is a great dataset to work with for MLR model and to perform Exploratory Data Analysis (EDA).

Data Dictionary:

- **Caret** - caret weight of diamonds
- **Cut** - cut quality of diamond (fair, good, very good, premium, ideal)
- **Color** - color of diamonds with d being the best and j being the **worst.**
- **Clarity** - how obvious the inclusions are within the diamonds
- **Depth** - depth % (the height of diamonds measured from the culet to the table, divided by its average girdle diameter)
- **Table** - table % (the width of the diamonds table expressed as a percentage of its average diameter.
- **Price** - price of the diamonds.
- **X** - length in mm
- **Y** - width in mm
- **Z** - depth in mm.

Along with the data dictionary it is important to understand the type of variable you are dealing with i.e. Categorical or Continuous.

Categorical Variables	Continuous Variable
1. Cut: a. 1 = Fair b. 2 = Good c. 3 = Very Good d. 4 = Premium e. 5 = Ideal 2. Color: a. D, E, F, G, H, I, J - (1-7) 3. Clarity: a. I1, IF, SI1, S12, VS1, VS2, VVS1, VVS2	1. Caret 2. Depth (%) 3. Table (%) 4. Price 5. Length (in mm) 6. Width (in mm) 7. Depth (in mm)

Based on our dataset Price is our response variable or dependent variable as we are trying to predict price of diamond based on its color, clarity, cut, depth and more regressors associated to it.

Exploratory Data Analysis (EDA)

EDA refers to the critical process of performing early exploration and investigation on the dataset so as to discover hidden patters, to spot anomalies, for hypothesis testing and to check assumptions with the help of summary statistics, graphs and various other visualizations. Firstly, it is always better practice to understand the dataset you are working with and trying to explore as many insights from the data. EDA is all about making sense of data in hand before getting your hands dirty with it.

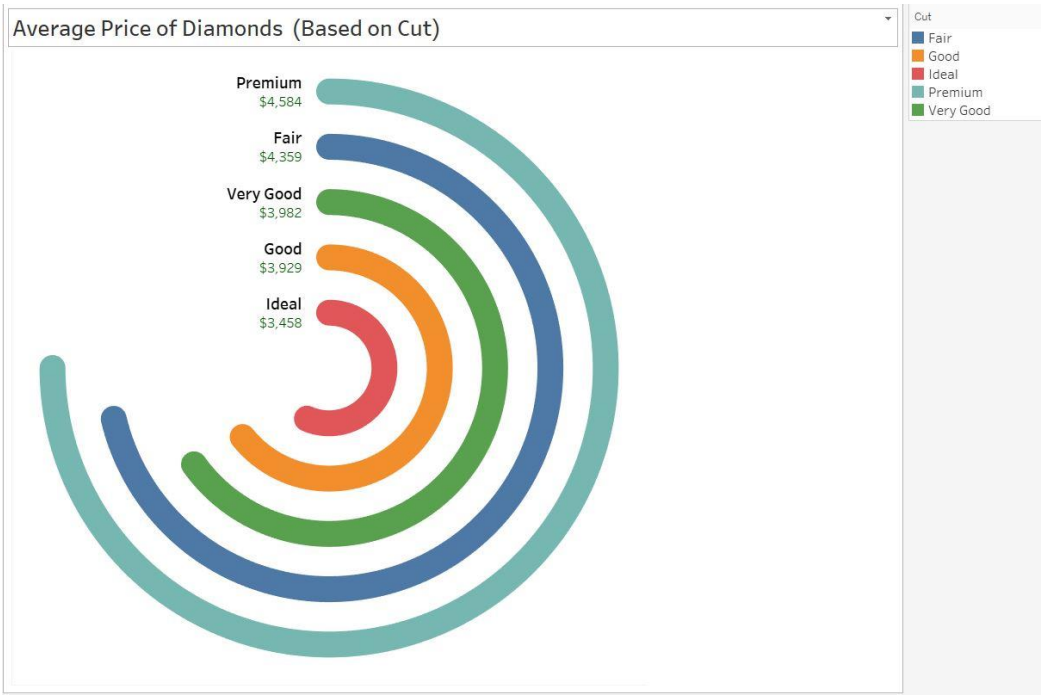
Purpose of EDA:

- To maximize insight into the dataset.
- To extract important variables.
- To test underlying assumptions.
- To uncover underlying structure.
- To detect outliers and anomalies; and more.

Average Price of Diamonds (EDA):

In our first step we can start by plotting scatterplots, bar plots or histograms to see the distribution of variables as well as to see the relationship between any variables as our major goal to perform EDA to explore as much insight as possible. As you can see the figure below, we have created

following radial bar chart to see the price distribution of diamonds based on its quality of cut. The average price of diamonds being in a range of $3,458 for Ideal and $4,548 for Premium.



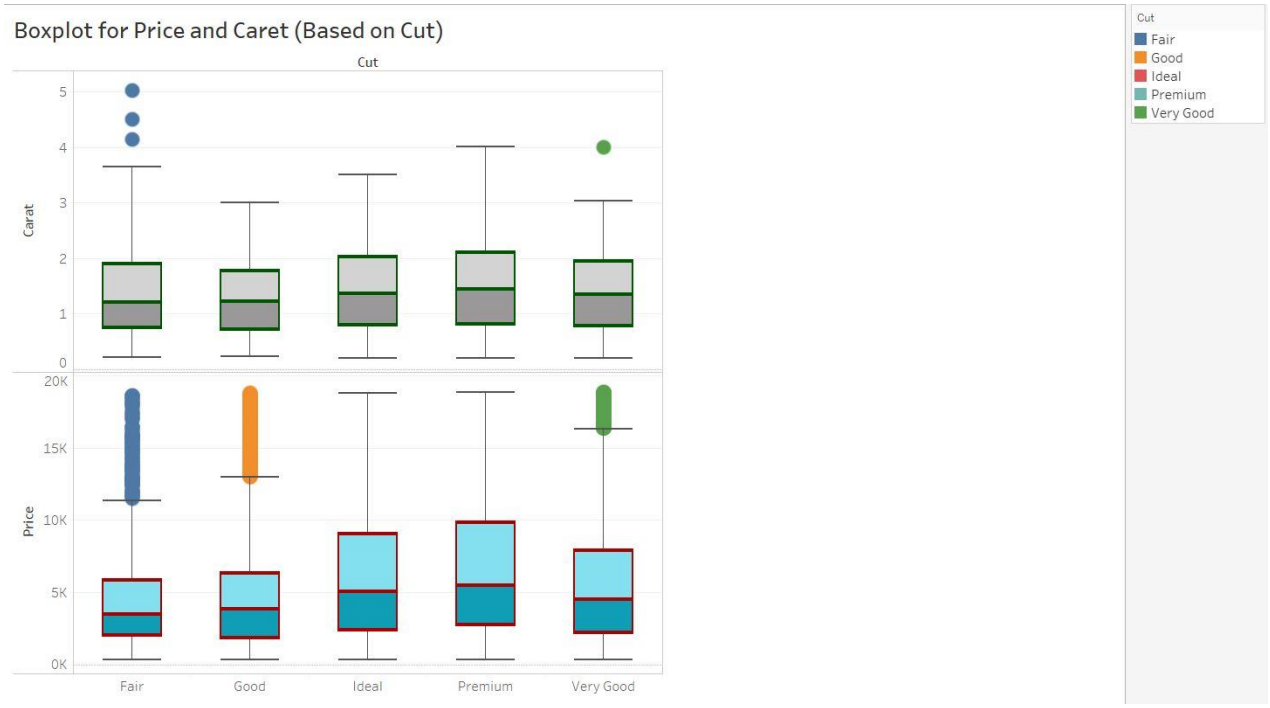
We can see the category Fair having more average price compare to Good and Very good category which simply suggests us that the cut of diamonds is not the only category that helps to determine the price of diamonds.

Detecting Outliers:

Outliers is basically a data point that is present significantly away from the other observation points. An outlier may occur due to variability in the measurements or due to experimental errors during data entry phase. Such outliers can cause serious problems to our model therefore, it is

extremely crucial to take care of such data points by removing them or by replacing them with relevant statistics such as mode and medians. However, it is equally important to research and understand the existence of such extreme points in the dataset before removing or replacing them.

There are numerous techniques and visualizations that we can perform to detect outliers. Z - Score, Standard Dev., Scatterplot, Boxplot, etc. are some EDAs we can perform to see if any outliers are present in the Dataset. Here we have created boxplot based on the Cut category of diamonds to observe the nature of outliers.



The figure above represents the various boxplot created based on the different category of cut. We can see most of the observation points are within

the upper and lower whisker range for Ideal and Premium category on both rows of boxplots.

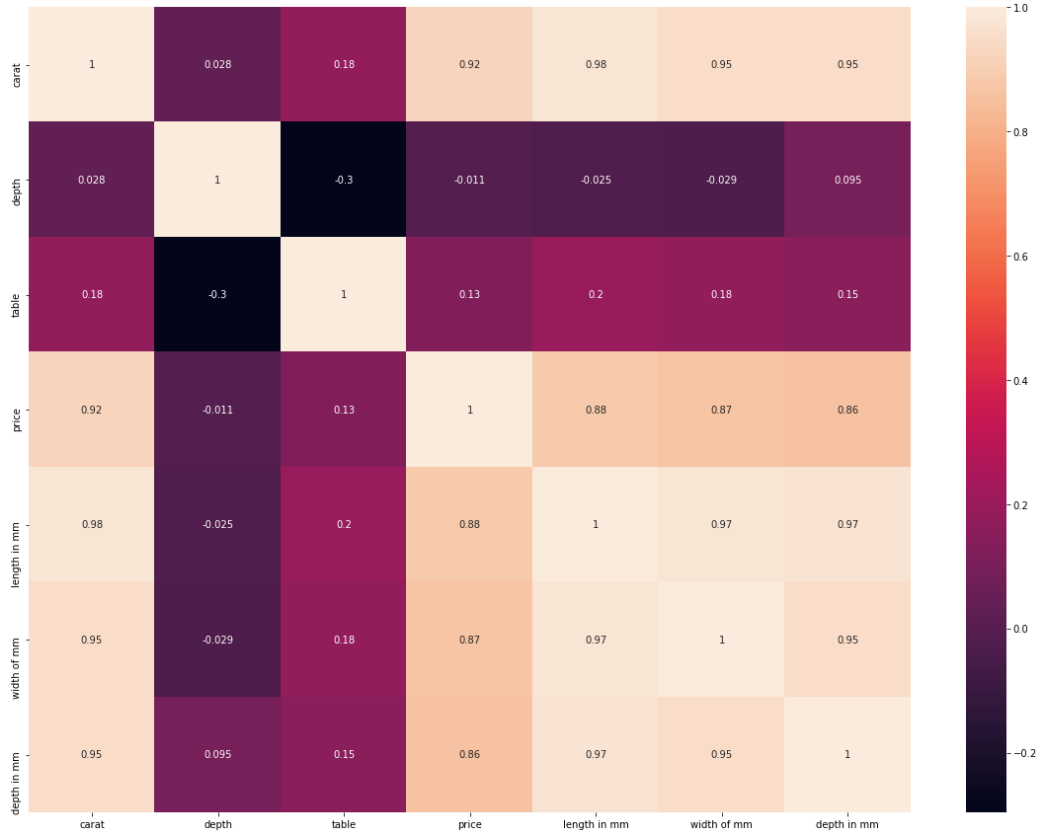
In the first row of boxplot, there are few possible outliers for Fair and Very Good category. However, the outliers are basically suggesting high value of caret for those cut categories which is possible, and they are not far from the upper IQR so, we will not eliminate the outliers from out dataset.

In the same way we can see there are a lot of observations out of the upper IQR suggesting us high price of diamonds despite of its cut. If we observe the IQR for each category we can see most of the data points are between \$12k- \$0k for Fair category and with increasing order followed by Good (\$13.5k - \$0k), Very Good (\$16k - \$0k), Ideal and Premium (\$20k - 0k).

As the price and caret does not necessarily depends on the cut of diamonds, we are not removing out outliers based on this visualization and also, we can see that the existing outliers are not that far from their IQR.

Correlation Matrix:

As you can see the Figure below, correlation matrix is the table that consists the information of correlation coefficient between the sets of variables present in our dataset. Each random variable (X_i) is correlated with every other existing variable in the dataset. This table helps you figure out the variables with highest correlation based on the correlation coefficient to tackle multicollinearity problems.



Before making assumptions based on the visualization above it is important to understand the indicators and the definition of correlation coefficient. Here, shades of white color are representing high correlation, for example there is 0.98 correlation between caret and length in mm indicated by 5^{th} block in first row of visualization. We can see more categories that are highly correlated with each other especially on the bottom right corner (length in mm vs depth in mm, width in mm vs length in mm). As these high correlation between regressors can cause serious multicollinearity issues, therefore in order to avoid them, we have to eliminate one of such correlated variables from each relationship with higher correlation coefficient.

Here, shades of purple color indicate the low correlation or no relationship between variables and the shades of black color represents the negative relationship between variables. As these 2 categories are not a serious threat to our model so, we mostly focus on the regressors with high correlation with each other and will be removed some variables while creating our model.

Multiple Linear Regression

It is extremely important to understand the dataset, to perform EDA to explore as much insight as possible with the help of numerous visualization techniques, and to implement necessary data preprocessing steps such as, dealing with missing values, taking care of outliers, creating dummy variable for categorical variables and more, before we implement MLR to predict our response. As we do not have to deal with missing value in this dataset and we have already taken care of our outliers, now we will create dummy variable for our categorical variable.

Dummy Variables:

Dummy variables are newly formed columns used in regression analysis to represent different categories of categorical variable. For an example in our dataset, we know cut is a categorical variable and one of our regressor that we are using to predict the price of diamonds. We have 5 different category in this variable and if we were to simply replace those values with random numbers, 1,2,3, etc. then our model will think the value 3 is greater than 1 but this is not the case as they are simply representing their respective

category. To tackle this issue, we will create multiple columns each corresponds to on category with only values 0 and 1, where 1 being the observation based on that category and 0 representing some other category for that observation.

1. Example: Categorical variable in our dataset, **Cut** having following categories
    a. 1 = Fair
    b. 2 = Good
    c. 3 = Very Good
    d. 4 = Premium
    e. 5 = Ideal

If observation 1, 2, 3 and 4 have cut quality fair, good, ideal and premium respectively then the dummy variable representation is as follows,

Fair	Good	Very Good	Premium	Ideal
1	0	0	0	0
0	1	0	0	0
0	0	0	0	1
0	0	0	1	0

Here, once the dummy variable is created one category should be removed to avoid dummy variable trap, as we can still preserve information about Ideal in our data frame above even if we remove the entire column.

In the same way dummy variables were created for color category as well.

Parameter Estimates:

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	909.42346	452.49840	2.01	0.0445
carat	1	9746.23452	44.04810	221.26	<.0001
depth	1	-23.48347	5.05532	-4.65	<.0001
table	1	-40.79251	3.63517	-11.22	<.0001
depth_in_mm	1	-1085.59792	29.37411	-36.96	<.0001
cut_Good	1	1007.29715	41.58296	24.22	<.0001
cut_Ideal	1	1530.57690	41.26402	37.09	<.0001
cut_Premium	1	1240.98547	39.86220	31.13	<.0001
cut_Very_Good	1	1335.47915	39.76450	33.58	<.0001
color_D	1	1938.50311	32.43831	59.76	<.0001
color_E	1	1846.62661	31.00958	59.55	<.0001
color_F	1	1873.94997	30.89437	60.66	<.0001
color_G	1	1832.82950	30.27278	60.54	<.0001
color_H	1	1196.07264	31.02329	38.55	<.0001
color_I	1	838.23213	32.88843	25.49	<.0001

Fitted line:

$$y = 909.4 + 9746.23\ carat - 23.483\ depth - 40.79\ table - 1085.59\ depth\ in\ mm + 1007.29\ cut\ Good + 1530.576\ cut\ Ideal + \cdots \ldots + 838.23213\ color\ I$$

The table above shows the parameter estimate/coefficient for each regressor using backward selection method. In general, the value of parameter estimate is the price increase in diamonds with one unit increase in that relative regressor. Where negative value shows the decrement in price with one unit increase on that regressor. For an example, the model suggests us that $9746.23 will increase in price with each caret increase. In the same way, with one unit increase of depth % price drop of $23.483 will occur.

Significance Test:

We know that a low p-value of regressor shows the greater impact to the model hence, we have a cut off margin with p-value 0.05. As the regressor being insignificant to the model if the p value is greater than 0.05 and the regressor being significant if the p-value is less than 0.05. Based on these criteria, all regressors are significant to the model generated using diamonds dataset.

ANOVA (Analysis of Variance):

Number of Observations Read	53940
Number of Observations Used	53940

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	7.509389E11	53638494089	26898.0	<.0001
Error	53925	1.075342E11	1994144		
Corrected Total	53939	8.584731E11			

Root MSE	1412.14166	R-Square	0.8747
Dependent Mean	3932.79972	Adj R-Sq	0.8747
Coeff Var	35.90678		

The table above contains the summary statistic from the multiple linear regression model we have created for the diamonds dataset with 53, 940 observations to predict the price of diamonds based on various features such as caret, cut, color, depth and more.

Our model has low SS (Errors) which is a good sign with the model being significant as the p-value is less than 0.05. Also, another important factor to look at is R-Squared values as it explains the variability percentage of response explained by its regressors. We can see observe in ANOVA table that our model has the R^2 value of 0.8747 which can be interpreted as,

"87.47 % variability of diamond price is explained by the associated regressors or independent variable." As we know R^2 value can be only in the range of 0 - 1 where the value closer to 1 being better and R^2 value of 1 being ideal case. In our model we have 0.875 which is considered as strong model.

Conclusion:

We have successfully created and implemented our multiple linear regression model to the diamond's dataset. We have tested the regressors for significance and eliminated the regressors which are not significant to our model by using model selection technique called backward selection, where the model is started with all regressors and insignificant regressors are being dropped off from the model along the elimination process.

Our model suggests that the carat has highest impact on the price of diamonds with almost $10k increase in price with one unit of carat increased. Also, Ideal cut, Color type D & F, etc. have significant impact on the price of diamonds as well.

There are also some other regressors having negative relationship with response, meaning as the value increase the price of the diamonds decrease. Especially, one unit increase of depth mm can lower the price of diamonds by $1,085.59.