Project: Predicting the Optimal Bid Price for Diamonds

Summary: In this project, based on the historical data which contains specific information about diamonds sold in the market such as carat, cut, clarity and price (53001 diamonds), I predicted the optimal total bid price for a new set of diamonds (3000 diamonds).

Step 1: Understanding the Model

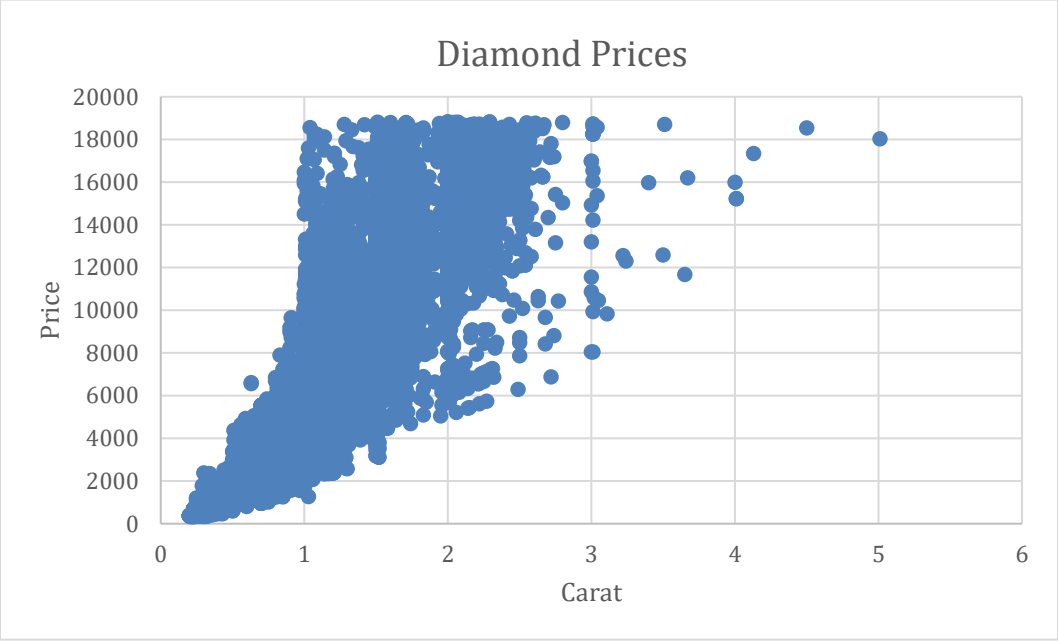
First, I created a regression model to predict the prices for the new diamonds.

1. *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?*
   According to the regression coefficient of carat variable, each unit increase in carat would result in an additional **$8413** in price.
2. *If you were interested in a 1.5 carat diamond with a **Very Good** cut (represented by a 3 in the model) and a **VS2** clarity rating (represented by a 5 in the model), how much would the model predict you should pay for it?*

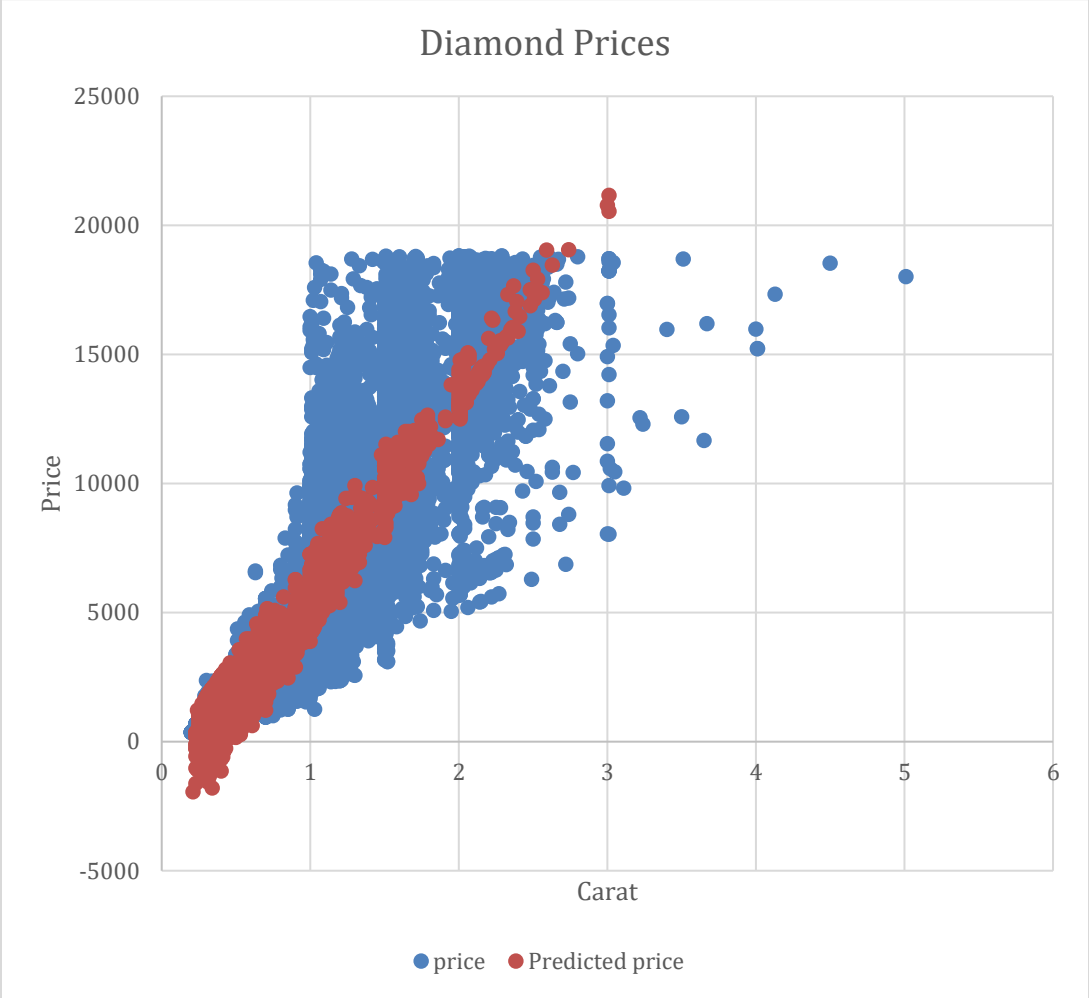
   In accordance with the regression formula the price of such a diamond should be as follows:
   Price = -5,255.2 + 8,363.4 x Carat + 160.4 x Cut + 457.8 x Clarity
   -5255.2+8363.4*1.5+ 160.4*3+457.8*5 = **$10.060.1**

Step 2: Visualizing the Data

In Plot 1, I plotted the data for the diamonds in the database, with carat on the x-axis and price on the y-axis.



In Plot 2, I plotted the data for the diamonds for which I am predicting prices with carat on the x-axis and predicted price on the y-axis. Here I plotted both sets of data on the same chart in different colors.



What I understand from my analysis is that Carat is the most important predictor of diamond prices. The model seems to predict the diamond prices well.

As can be seen from the scatterplots; both the actual diamond prices and the predicted prices seem to have similar pattern on the scatterplot.

One important point is that we do not have new diamonds over 3 carat (only two being 3.01) while we had carats in the first sample (actual diamonds) which had diamonds with carats up to 5.01.

The new diamonds data seem to have a better linear regression line than the actual diamond data.

Step 3: Making a Recommendation

Report for Linear Model Diamond_Prices

Basic Summary

Call:

lm(formula = price ~ carat + cut_ord + clarity_ord, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-19246	-693	-105	543	10956

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5255.2	30.320	-173.33	< 2.2e-16***
carat	8363.4	13.565	616.55	< 2.2e-16***
cut_ord	160.4	5.513	29.09	< 2.2e-16***
clarity_ord	457.8	3.901	117.37	< 2.2e-16***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1348 on 49996 degrees of freedom

Multiple R-squared: 0.8862, Adjusted R-Squared: 0.8862

F-statistic: 129779 on 3 and 49996 degrees of freedom (DF), p-value < 2.2e-16

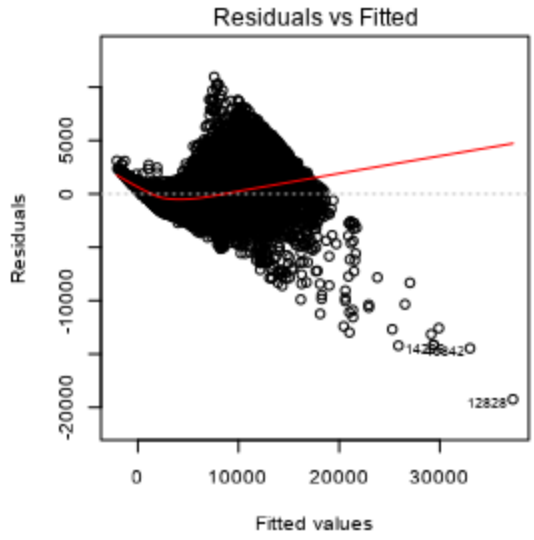
Type II ANOVA Analysis

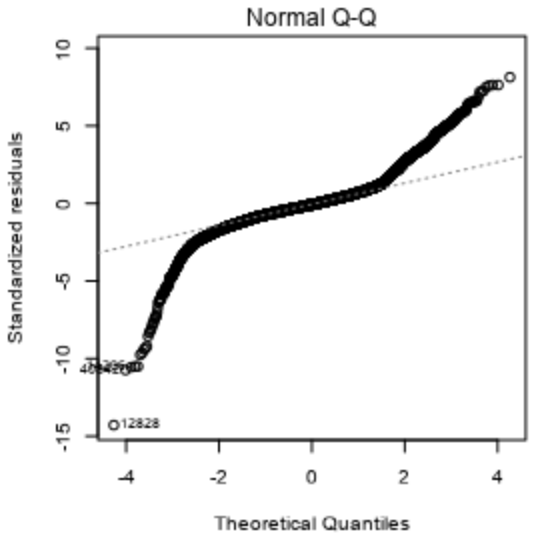
Response: price

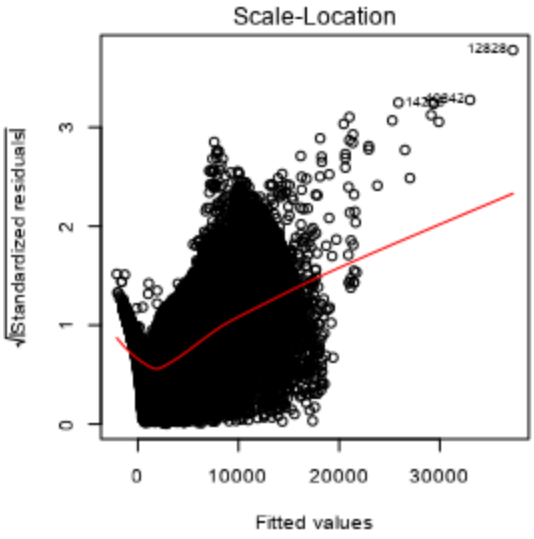
	Sum Sq	DF	F value	Pr(>F)
carat	690754945263.16	1	380130.35	< 2.2e-16***
cut_ord	1538039869.18	1	846.4	< 2.2e-16***
clarity_ord	25030996994.94	1	13774.84	< 2.2e-16***
Residuals	90850372351.38	49996		

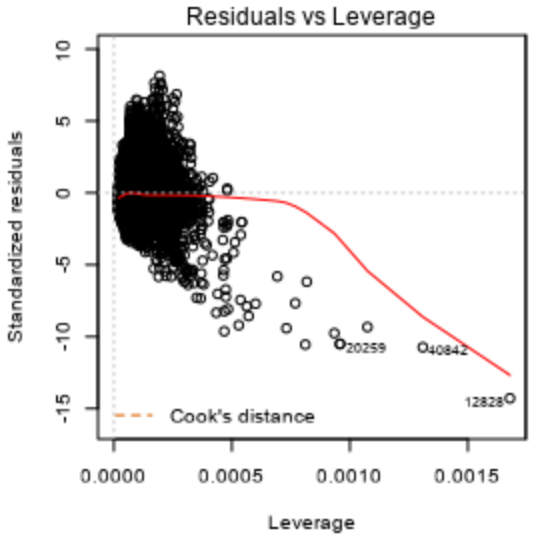
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

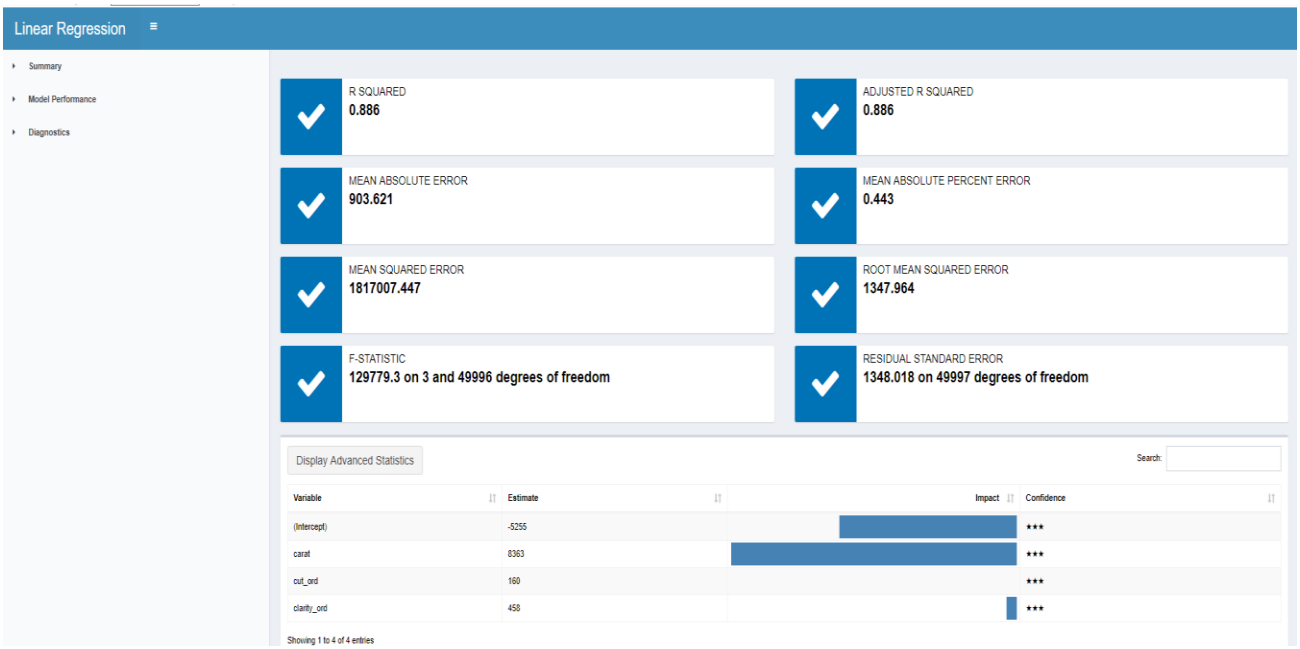
Basic Diagnostic Plots


Residuals vs Fitted
Residuals
Fitted values
12828


Normal Q-Q
Standardized residuals
Theoretical Quantiles
12828


Scale-Location
√|Standardized residuals|
Fitted values
12828

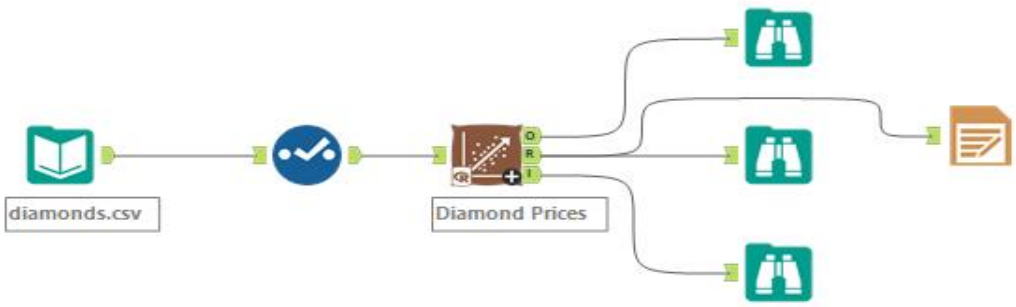

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
20259
40842
12828



Based on the regression model I created I calculated the predicted price for every single diamond in our new dataset of diamonds and then calculated the total predicted price.

The total predicted price for the new set of diamonds is $11,732877.81. The company usually buys the diamonds from distributors at 70% of the retail price. We need to multiply the predicted price with .70. to find out the bid.

Therefore, I recommend a bid of (11,733522.76*0.70) = $**8,213465.932**



Prepared by Kadir AKYUZ