3/17/2021

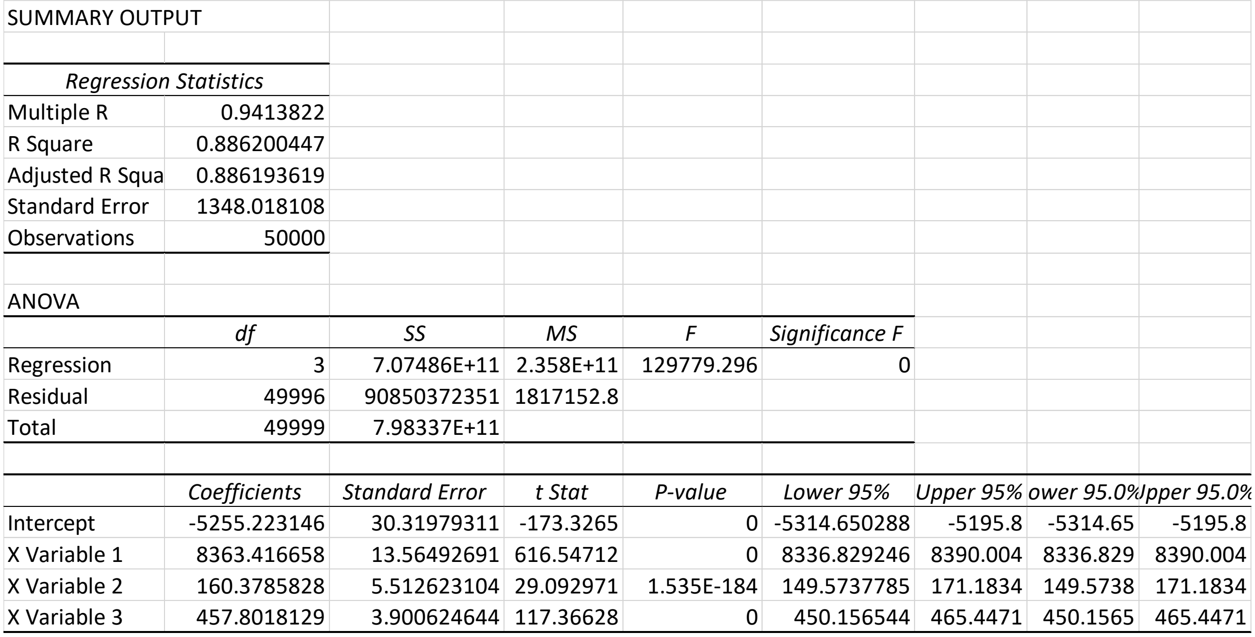
Future Seekers - Business Analytics Nanodegree Program
**Project 3**

**Atheer Alhabeeb**

# Problem Solving with Analytics
# Diamond Prices

## Step 1: Understanding the Model

Depending on the decision of the diamond distributor (ex. X), to exit the market and put up a batch of 3000 diamonds for auction. I am Atheer, as a future business analyst in the analytics team of a company (for example Y). I will develop a proposal to submit a proposal that determines the bid amount through the use of a database of diamond prices, and build a regression model to predict the price of diamonds based on the data.



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| SUMMARY OUTPUT | | | | | | | | |
| | | | | | | | | |
| *Regression Statistics* | | | | | | | | |
| Multiple R | 0.9413822 | | | | | | | |
| R Square | 0.886200447 | | | | | | | |
| Adjusted R Squa | 0.886193619 | | | | | | | |
| Standard Error | 1348.018108 | | | | | | | |
| Observations | 50000 | | | | | | | |
| | | | | | | | | |
| ANOVA | | | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Significance F* | | | |
| Regression | 3 | 7.07486E+11 | 2.358E+11 | 129779.296 | 0 | | | |
| Residual | 49996 | 90850372351 | 1817152.8 | | | | | |
| Total | 49999 | 7.98337E+11 | | | | | | |
| | | | | | | | | |
| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* | *ower 95.0%* | *Jpper 95.0%* |
| Intercept | -5255.223146 | 30.31979311 | -173.3265 | 0 | -5314.650288 | -5195.8 | -5314.65 | -5195.8 |
| X Variable 1 | 8363.416658 | 13.56492691 | 616.54712 | 0 | 8336.829246 | 8390.004 | 8336.829 | 8390.004 |
| X Variable 2 | 160.3785828 | 5.512623104 | 29.092971 | 1.535E-184 | 149.5737785 | 171.1834 | 149.5738 | 171.1834 |
| X Variable 3 | 457.8018129 | 3.900624644 | 117.36628 | 0 | 450.156544 | 465.4471 | 450.1565 | 465.4471 |

*Source: Retrieved from diamonds.csv.*

From the previous summary of the output that I extracted from the data available via Excel (for note only I can extract more information through the Eviews program that economists use to build economic models and analogize them), a multiple linear regression equation can be written that is similar to the equation written in the project details.

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X1 + \hat{\beta}X2 + \hat{\beta}X3$$

Price = -5,269 + 8,413 x Carat + 158.1 x Cut + 454 x Clarity

Where:

$\hat{Y}$: The dependent variable is price.

$\hat{\alpha}$: Is Y intercept.

$\hat{\beta}$X1: It expresses the estimated Carat value.

$\hat{\beta}$X2: It expresses the estimated Cut value.

$\hat{\beta}$X3: It expresses the estimated Clarity value.

Now that I have written the multiple linear regression equation, I can answer these important questions.

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? Why?**

   Based on the previous model, the more the carat increases with the other factors remaining constant, the higher the price (the dependent variable Y). This becomes apparent if the values are replaced by the previous equation, as I would expect to pay more. In other words, if the carat increased by one over the value of 8,413, that would naturally lead to an increase in the price by the amount of the carat coefficient.

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?**
   If my interest falls on a specific diamond with all its details, the amount expected from this form is:

   Price = -5,269 + (8,413 x Carat) + (158.1 x Cut) + (454 x Clarity)

   Where:

   - Carat equal 1.5
   - Cut equal 3
   - Clarity equal 5

   By substituting the values for this diamond, I can predict the price or which must be paid.

   Price = -5,269 + (8,413 x 1.5) + (158.1 x 3) + (454 x 5)

   Price = -5,269 + 12619.5 + 474.3 + 2270

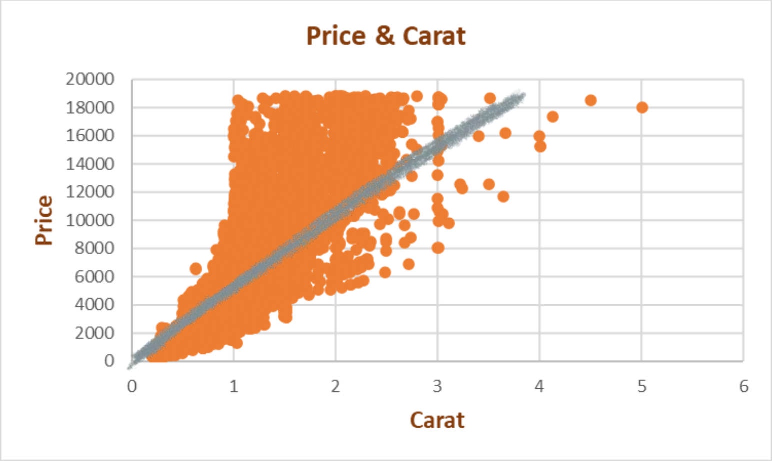
   Price = 10094.8

So, the amount that this model expects to pay for this diamond is 10,094.8$.

## Step 2: Visualize the Data

1. **Plot 1:** Plot the data for the diamonds in the database, with carat on the x-axis and price on the y-axis.

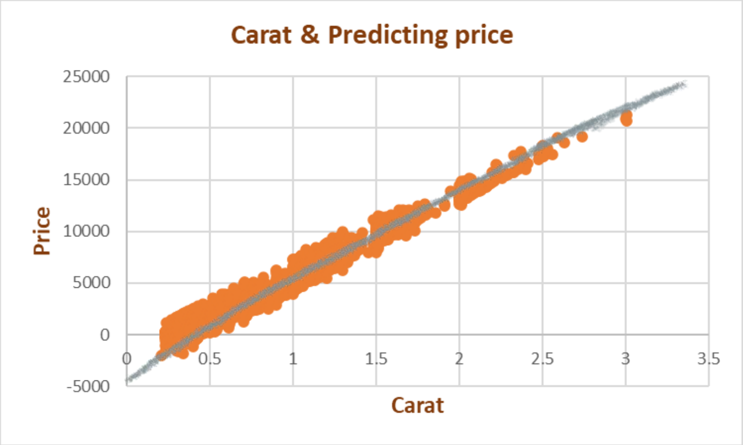
   The straight line in gray that I drew is to clarify only about the spread of the data, as it is noticeable that the data is converging in the beginning and then extends in the spread away from the line.



*Source: Retrieved from diamonds.csv.*

2. **Plot 2:** Plot the data for the diamonds for which you are predicting prices with carat on the x-axis and predicted price on the y-axis.

   The gray straight line I drew is for illustration only about the spread of the data, as it is noticeable that the data converge towards the line (which shows multiple linear regression).



*Source: Retrieved from new_diamonds.csv.*

3. What strikes you about this comparison? After seeing this plot, do you feel confident in the model’s ability to predict prices?

   After I drew these two plots, it becomes clear that the actual data on the prices of diamonds initially appears symmetrical and walks in one direction, but then they begin to disperse and move away from each other clearly and become non-linear (The range is large). This may put us in a position to see that there are many factors that affect the price other than the carat. It can be said in another way that the correlation coefficient appears weak.

But when looking at the expected prices, the matter may seem more disciplined, as the points do not disperse much, but rather are more appropriate when we bear in mind that we are studying multiple linear regression. In other words, we can say that the correlation coefficient is strong. Hence, I think the model (multiple linear regression model) does not seem very suitable for predicting prices.

## Step 3: Make a Recommendation

1. What price do you recommend the jewelry company to bid? Please explain how you arrived at that number.

It is clear from the above that I think that the multiple linear regression model does not seem very appropriate in our case, because some diamonds have a negative price and this does not make sense. Therefore, I recommend submitting the bid proposal at a price of $ 8213,465.932.
I arrived at this amount using the SUM formula for expected prices in the new_diamonds.csv file, after which I took into account 70% of the total expected prices and was multiplied by the total expected prices to obtain the final offer of $ 8213,465,932.