Project 01

Predicting Diamond Prices

Rosana Ferreira Soares dos Santos

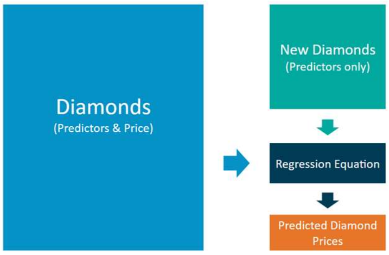
Project Overview

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, the results from a predictive model will be used to make a recommendation on how much the jewelry company should bid for the diamonds.

Project Details

A diamond distributor has recently decided to exit the market and has put up a set of 3,000 diamonds up for auction. Seeing this as a great opportunity to expand its inventory, a jewelry company has shown interest in making a bid. To decide how much to bid, the company's analytics team used a large database of diamond prices to build a linear regression model to predict the price of a diamond based on its attributes. **I, as** the **business analyst**, am **tasked to apply that model** to **make a recommendation** for how much the company should **bid** for the entire **set of 3,000 diamonds**.

The following diagram represents the analysis at a high level. Since the model is already built, my analysis will **focus** on the **right side of the diagram**.



- The linear regression model provides an equation that you can use to predict diamond prices for the set of 3,000 diamonds. The equation is below:
  Price = -5,269 + 8,413 x **Carat** + 158.1 x **Cut** + 454 x **Clarity**

Both datasets contain carat, cut, and clarity data for each diamond. Only
the diamonds.csv dataset has prices.

Complete each section. When you are ready, save your file as a PDF document and submit it in your classroom.

Step 1: Understanding the Model

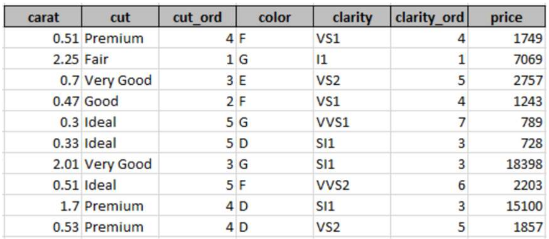
Answer the following questions:

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?

   diamonds.csv contains the data used to build the regression model.

   new_diamonds.csv contains the data for the diamonds the company would like to purchase

   Both datasets contain carat, cut, and clarity data for each diamond. Only
the diamonds.csv dataset has prices.



carat	cut	cut_ord	color	clarity	clarity_ord	price
0.51	Premium	4	F	VS1	4	1749
2.25	Fair	1	G	I1	1	7069
0.7	Very Good	3	E	VS2	5	2757
0.47	Good	2	F	VS1	4	1243
0.3	Ideal	5	G	VVS1	7	789
0.33	Ideal	5	D	SI1	3	728
2.01	Very Good	3	G	SI1	3	18398
0.51	Ideal	5	F	VVS2	6	2203
1.7	Premium	4	D	SI1	3	15100
0.53	Premium	4	D	VS2	5	1857

   Carat represents the weight of the diamond, and is a numerical variable.

   Cut represents the quality of the cut of the diamond, and falls into 5 categories: fair, good, very good, ideal, and premium. Each of these categories are represented by a number, 1-5, in the Cut_Ord variable.

   Clarity represents the internal purity of the diamond, and falls into 8 categories: I1, SI2, SI1, VS1, VS2, VVS2, VVS1, and IF. Each of these categories are represented by a number, 1-8, in the Clarity_Ord variable
   It was provided the linear regression model equation.

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

Reference Diamond = -5,269 + 8,413 + 158.1 + 454 = 3,756.1

Diamond 1 carat heavier = -5,269 + 8,413 x 2 + 158.1 + 454 = 12,169.1

If a diamond is 1 carat heavier than another with the same cut you should expect to pay ***8,413*** *more (12,169.1 – 3,756.1).*

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?

   Using the regression equation we have …

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

   Reference Diamond = -5,269 + 8,413 + 158.1 + 454

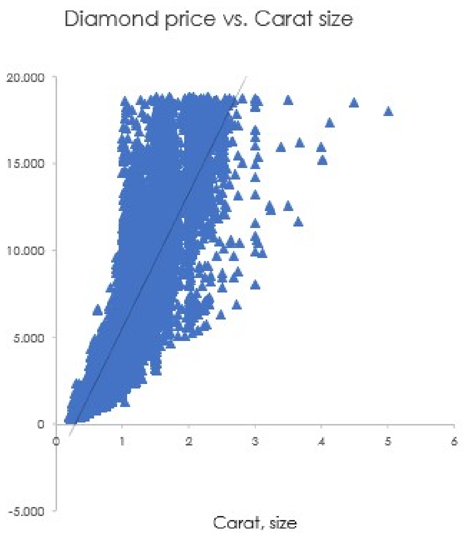
   = -5,269 + 8,413 x 1,5 + 158.1 x 3 + 454 x 5

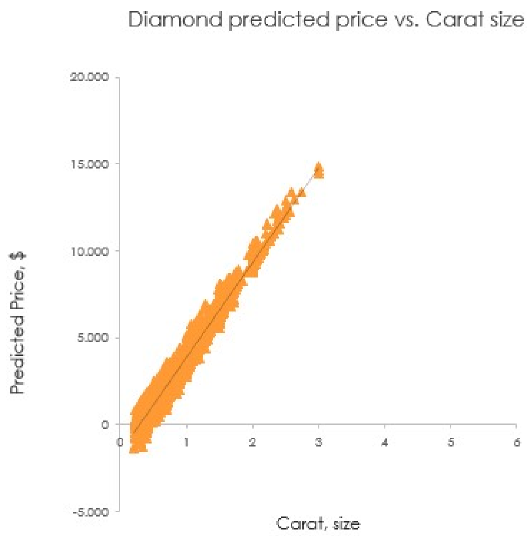
   Substituting the values in equation, results in Price = ***$10,094.8****.*

Step 2: Visualize the Data

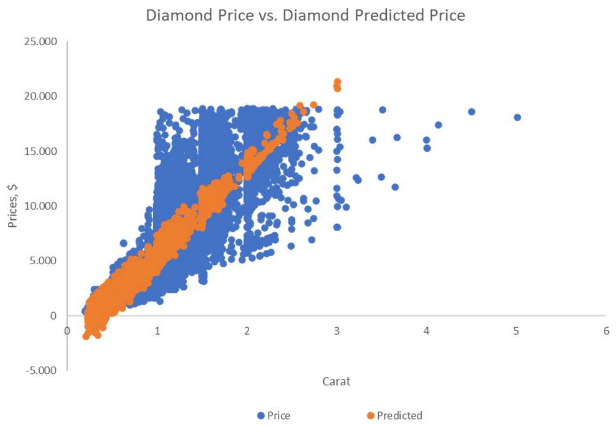
Make sure to plot and include the visualizations in this report. For example, you can create graphs in Excel and copy and paste the graphs into this Word document.

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





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.
    - **Note**: You can also plot both sets of data on the same chart in different colors.



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 <u>old diamond prices are non linear</u>. And there is the fact that negative prices were predicted for 291 diamonds. So I am not confident about the model.

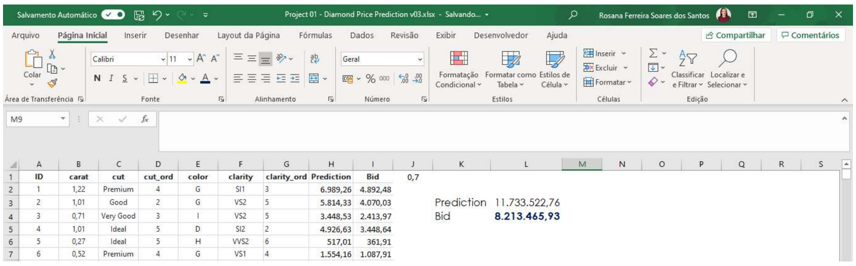
Step 3: Make a Recommendation

Answer the following questions:

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

   Bid sum should be 70% of predicted prices.

   Predicted prices sum was $11,733,522.76 So I multiplied by 0.7.



	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	carat	cut	cut_ord	color	clarity	clarity_ord	Prediction	Bid	0,7		
2	1	1,22	Premium	4	G	SI1	3	6.989,26	4.892,48			
3	2	1,01	Good	2	G	VS2	5	5.814,33	4.070,03		Prediction	11.733.522,76
4	3	0,71	Very Good	3	I	VS2	5	3.448,53	2.413,97		Bid	**8.213.465,93**
5	4	1,01	Ideal	5	D	SI2	2	4.926,63	3.448,64			
6	5	0,27	Ideal	5	H	VVS2	6	517,01	361,91			
7	6	0,52	Premium	4	G	VS1	4	1.554,16	1.087,91			

I recommend a bid of $8,213,565.93. I arrived at this number by using a formula from the regression model provided that was based on previous diamonds that were saled and applied it to the diamonds were up for bid.