PROJECT 1: PREDICTING DIAMOND PRICES

By Samuel Otisi

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, I will use the results from a predictive model 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. Now as a business analyst I will apply the model to make a recommendation for how much the company should bid for the entire set of 3,000 diamonds.

The linear regression model built have provided an equation I will use to predict diamond prices for the set of 3,000 diamonds. Below is the equation calculated using Alteryx software:

$$\textbf{Price} = -5{,}269 + 8{,}413 \times \textbf{Carat} + 158.1 \times \textbf{Cut} + 454 \times \textbf{Clarity}$$

Step 1 - Understand the data:

There are two datasets provided.

- **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 diamonds.csv dataset has prices, whereas new_diamonds.csv has not since I will be predicting the prices for them. The following are the brief description of the datasets features or predictors:

- *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.

Step 2 - Calculate the predicted price for diamond:

For each diamond, I’ve plugged in the values for each of the variables into the equation to get the estimated, or predicted, diamond price.

Step 3 - Make a recommendation:

Now that I have the predicted price for each diamond, its time to calculate the bid price for the new_diamonds.csv set, which is the 70% of that price.

Project Submission and Recommendation

Step 1 - Understanding the Model:

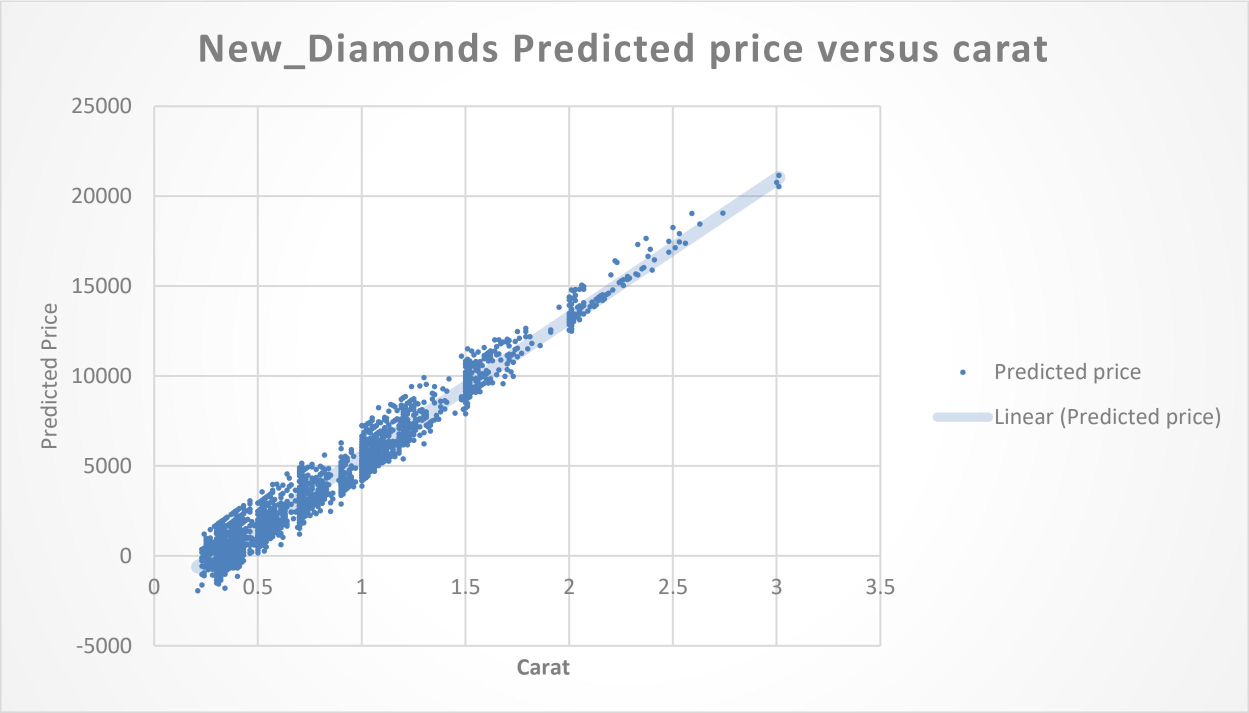
1. According to the linear regression model I built, to answer the following question, *“If a diamond is 1 carat heavier than another with the same cut and clarity, how much more would the retail price of the heavier diamond be”?* I will use a diamond with *Carat of 1.22, premium Cut* represented by *Cut_ord* feature as 4 and *SI1 Clarity* represented by *Clarity_ord* as 3 as a case study. Now plugging these values into the above model will give a predicted price of **$6,989.3.** when another diamond is 1 heavier than the former (i.e 1.22 + 1 = 2.22) keeping *Cut* and *Clarity* constant, the predicted price will change drastically to a whopping **$15,402.3**. This is due to high predictive power of the *Carat* feature compared to *cut* and *Clarity.* A plot in step 2 will verify this claim.

2. If I 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 I will expect the predicted price to be **$10,094.8**.

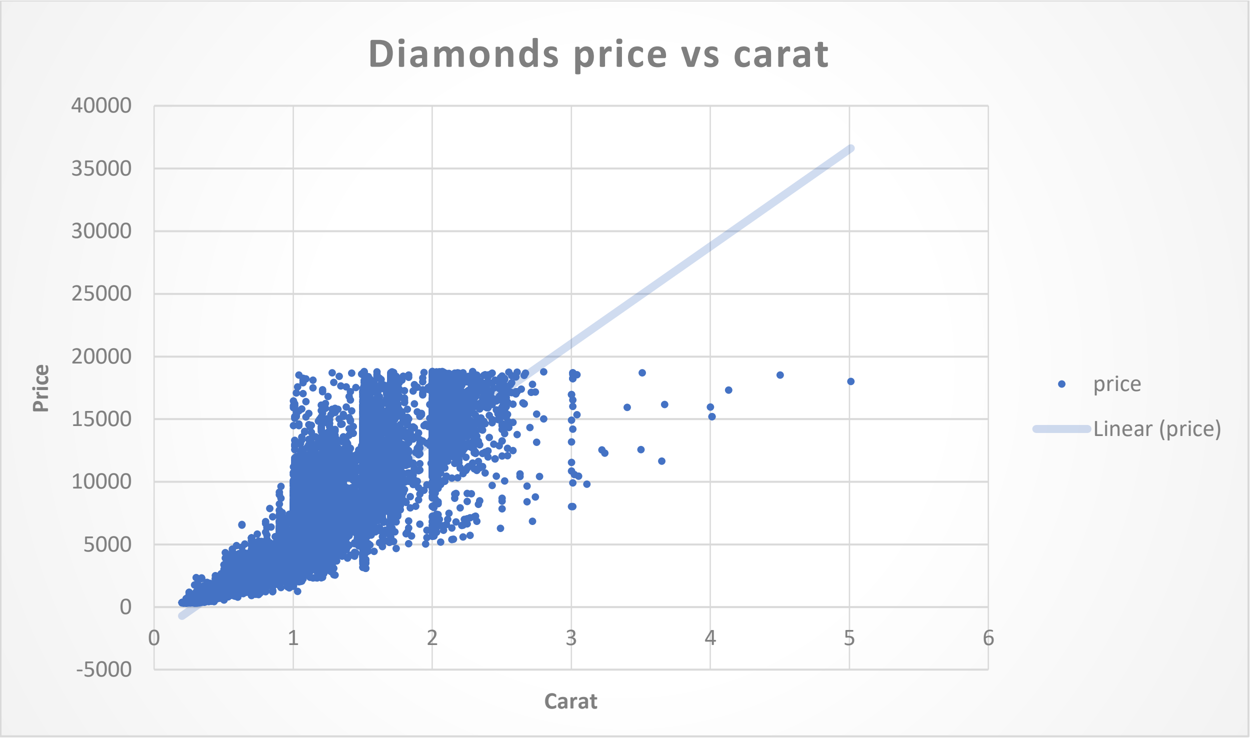
Step 2 – Visualize the Data:

The Alteryx workflow and Excel plots for these projects are found inside the zip file I created alongside with this pdf file.

- **Plot 1.**



- **Plot 2.**



- Comparing these two plots I can say the predicted diamond prices have a very strong correlation with *Carat of all weights,* while diamonds prices have a fair positive correlation especially for diamonds with weight more than 1. The positive correlations of these two plots established a good relationship between the chief predictor(carat) and expected prices thus feeling very confident in the model's ability to predict prices.

Step 3 – The Recommendation:

I have just used a predictive analytical technique (linear regression) to predict the price of these 3,000 new diamonds price. Due to proper validation of this model I recommend this jewel company to pay a sum **$8,211,163.2** (solution to this problem has been attached to the zip file), that is 70% of the total predicted prices to this diamond distributor. It is very important to note that any payment above this margin is over pricing.