**Predicting Diamond Prices**

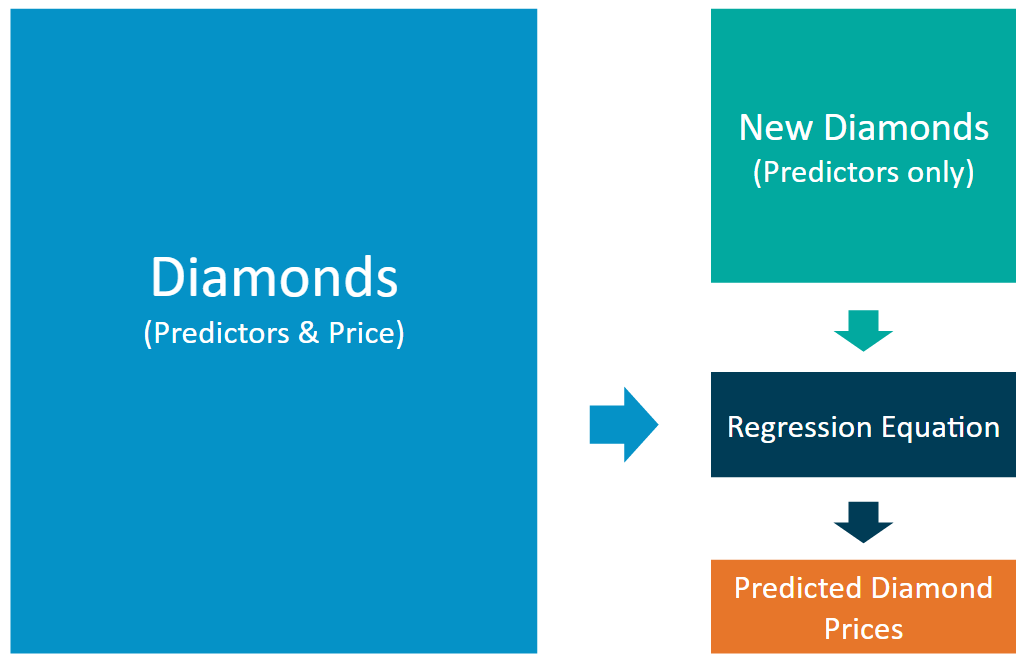
**Project Overview**

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, we 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 distributer 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. We are 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 we will be completing:

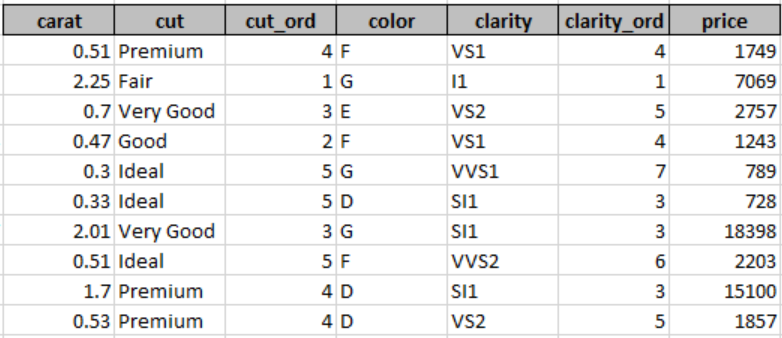


The linear regression model provides an equation that we can use to predict diamond prices for the set of 3,000 diamonds. The equation is as follows:

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

**Step 1 – Understand the data:** There are two datasets.

* **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. We’ll be predicting prices for the *new\_diamonds.csv* dataset.

* *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
* **Note**: Transforming category variables to ordinal variables like this is not always appropriate, but we’ve done it here for simplicity

**Step 2 – Calculate the predicted price for diamond:** For each diamond, plug in the values for each of the variables into the linear model (equation). Then solve the equation to get the estimated, or predicted, diamond price. We can use a spreadsheet tool like Excel, Numbers, or Google sheets to do this. We can also use Alteryx/Tableau if we have licenses for these programs.

**Step 3 – Make a recommendation:** Now that we have the predicted price for each diamond, it’s time to calculate the bid price for the whole set. Note: The diamond price that the model predicts represents the final retail price the consumer will pay. The company generally purchases diamonds from distributors at 70% of that price, so our recommended bid price should represent that.