Predicting Diamond Prices

This project is designed for three main reasons:

- To give you a feel for what you'll be doing throughout the Nanodegree Program
- To introduce you to Udacity's project submission and review process
- To make sure you feel comfortable with the basics before you begin. If it feels too easy, don't worry. We have some great stuff in store for you.

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

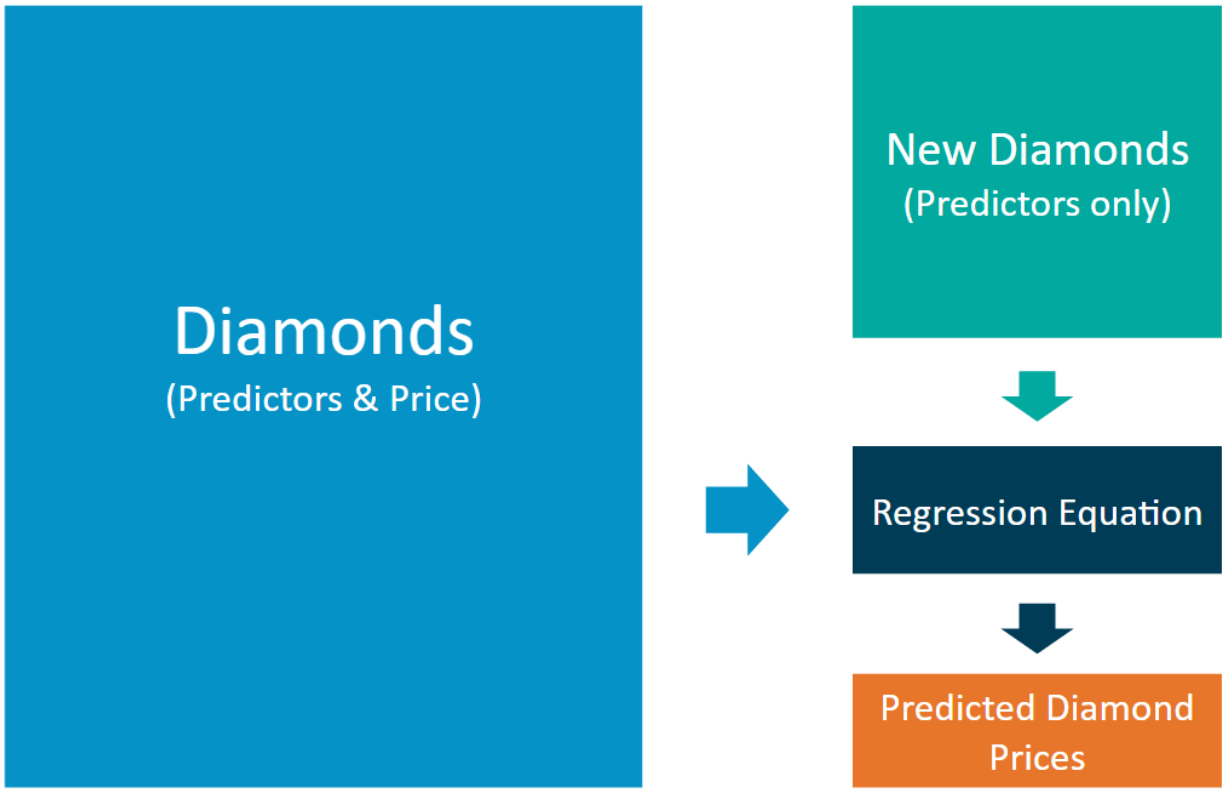
US Number System

All numbers that will be presented in this Nanodegree program will be based on the US numbering system where 5,269 is "five thousand two hundred sixty nine" and 158.1 is "one hundred fifty eight point one" where 1 is a decimal number. This is **very** important so please take note of this.

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. You, as the business analysts, 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 at a high level. Since the model is already built, your 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**

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.

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

Both datasets contain carat, cut, and clarity data for each diamond. Only the *diamonds.csv* dataset has prices. You'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 suggest using a spreadsheet tool like Excel, Numbers, or Google Sheets. You could also do it in Alteryx and/or Tableau if you already have a license. If you don't have a license yet, you'll receive one after your free trial.

Step 3 – Make a recommendation: Now that you 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 your recommended bid price should represent that.

Project Submission

To complete this project, you will be submitting a file in pdf format that contains the answers to the following questions across three steps.

Step 1 - Understanding the Model:

1. According to the linear model provided, 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? Why?
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), what retail price would the model predict for the diamond?

Step 2 - Visualize the Data: Create two scatter plots. If you're not sure what a scatter plot is, see here.

- Plot 1 - Plot the data for the diamonds in the database, with carat on the x-axis and price on the y-axis.
- 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.
- What strikes you about this comparison? After seeing this plot, do you feel confident in the model's ability to predict prices?

Step 3 - The Recommendation: What bid do you recommend for the jewelry company? Please explain how you arrived at that number.