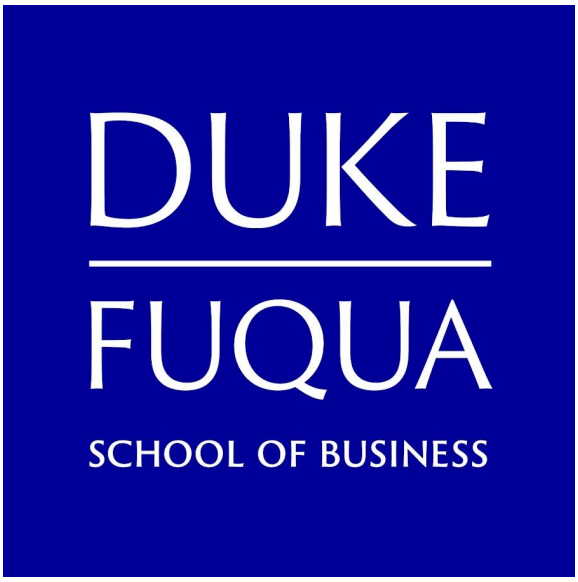


DECISION 520Q: Data Science

Final Project: Predicting Diamond Prices

Section C Team 46

Nan Wang, Yingli Sun, Jingwen Liu, Maya Rotman, Baicheng Wang

Business Understanding

The business problem we wanted to address was to be able to quickly, efficiently, and consistently price new diamonds coming into the stock of diamond retailer. We also wanted to determine which characteristics of diamonds contribute the most to the price. More specifically, we assumed that we were a group of data analysts in a luxury brand jewelry company who want to build a model for predicting the price of diamonds. Our task was to utilize the information and data on hand to come up with a formula that uses the attributes of diamonds to easily and accurately calculate the value of each diamond. In order to stay competitive in the market, goods must be reasonably priced to compete with competitors, but must also maximize firm profits. Our company can use the model to identify good quality diamonds and fully understand which variables contribute most to diamond price.

Data Understanding

The dataset we chose contains descriptive details on almost 54,000 diamonds. There are 11 variables in this data set to describe each of the 53,940 observations. The variables include carat, cut, color, clarity, depth, and price (see full list in appendix) as well as an identifier variable. The data was extracted from Kaggle (https://www.kaggle.com/shivam2503/diamonds). We decided to first investigated the correlation between each variable by performing exploratory analysis and plotting. We would then generate several models to predict how certain variables can affect the price of diamonds. This data mining solution, consisting of five models, addressed the defined business problem by helping the company better understand how the characteristics of a diamond related to price.

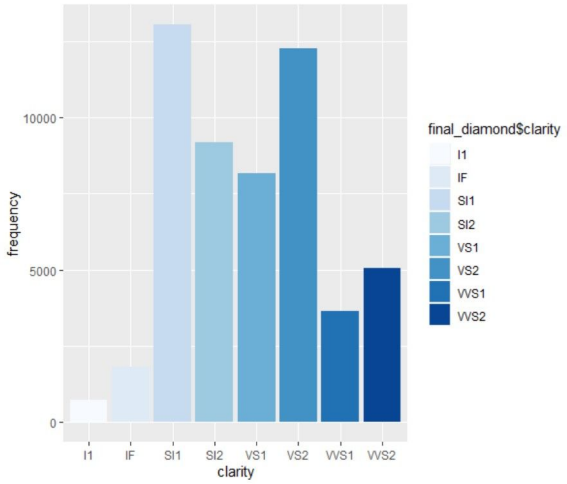
Data Preparation

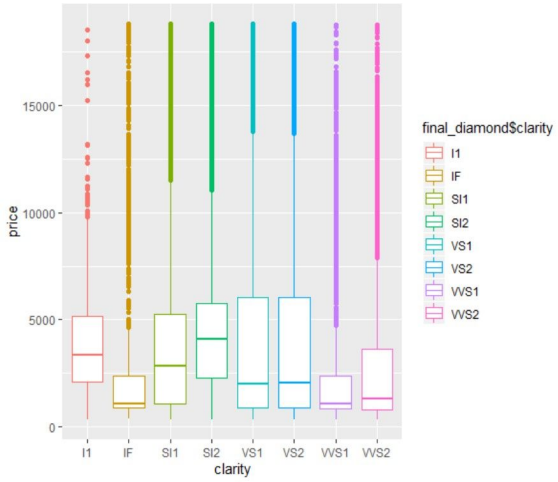
Our data preparation began with a search for null values. While there were no variables with explicitly null values, there were several observations in which the variables of x, y, and z, the measurements of diamonds, were classified as zero, which is impossible. These zero values would negatively impact our analysis, and there were very few of these observations, so we dropped the observation if any of those variables equalled zero. Next, we generated histograms and plots to investigate individual variables. Outliers were identified and dropped: we excluded observations when depth of diamond is less than 50 or greater than 75 or when table is less than 50 or greater than 90. Finally, we excluded the observations where y is greater than 20 or z is greater than 30. These changes resulted in a total of 36 dropped observations and a final cleaned data set of 53,904 observations. In order to prepare for the modeling, we decided to randomly divide the data into 2 subsets: train and test. We split the data 60% - 40% to the train set and the test set respectively, in order to provide out-of-sample performance results for each of the models we developed.

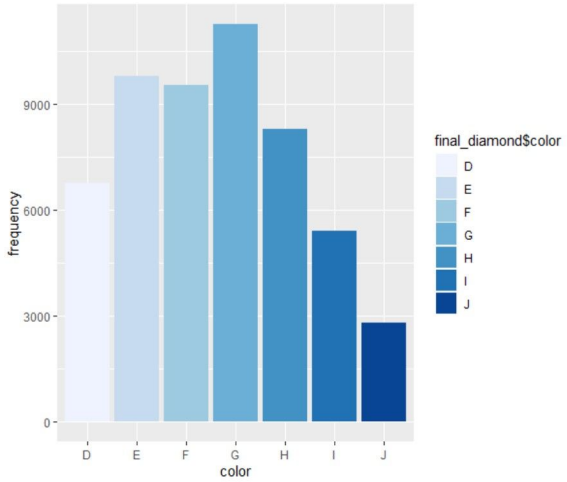
Data Visualization

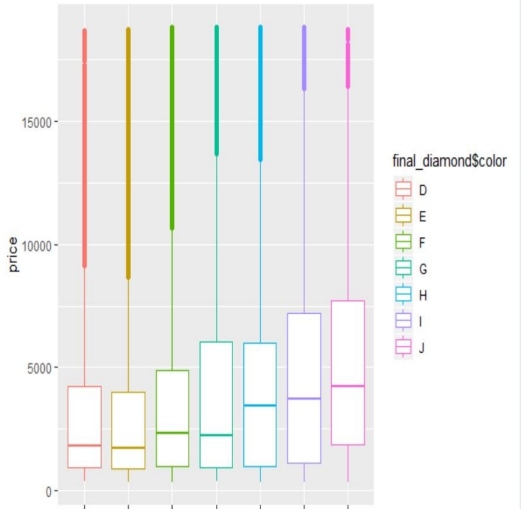
For data visualization, we create bar graphs for each one of the categorical variables in our dataset and box plots of the relationship between each categorical variable and price. There are fewest diamonds of clarity I1 and IF. The most common levels of clarity are SI1 and VS2. These results are consistent with our expectations, as I1 and IF are the lowest and highest level of diamond clarities, and SI1 and VS2 are in the middle of clarity levels. Through our correlation matrix and correlation chart, we can obtain the correlation coefficients and visualize the relationships between the variables. According to these graphs, Carat, x, y, and z all have

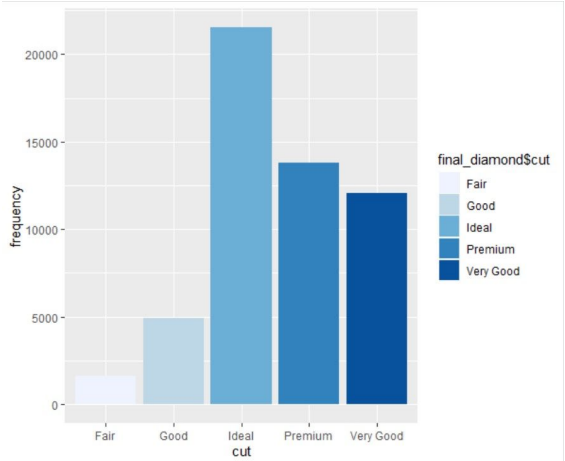
significant relationships with price. However, these variables are highly correlated, which can potentially decrease the R-Squared of our model.

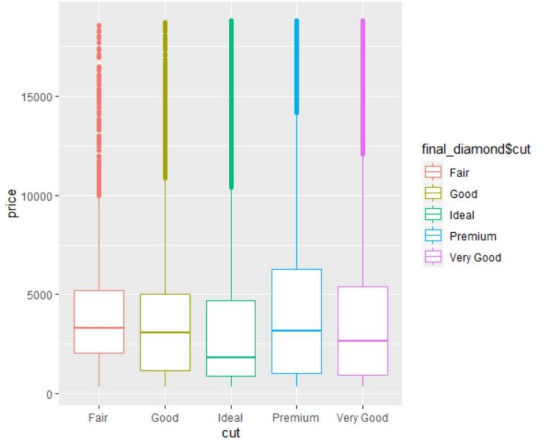


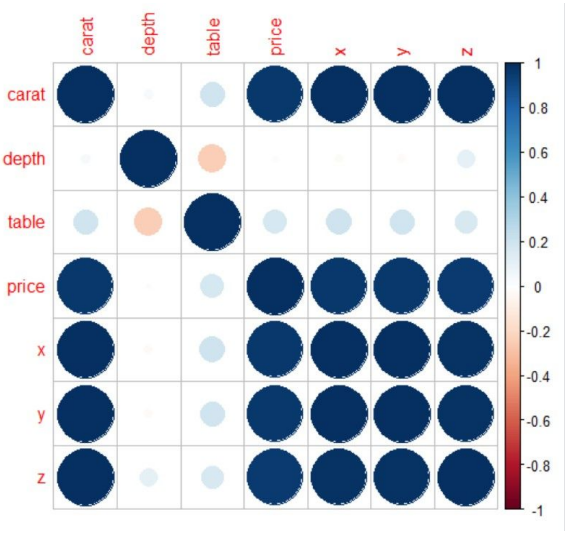


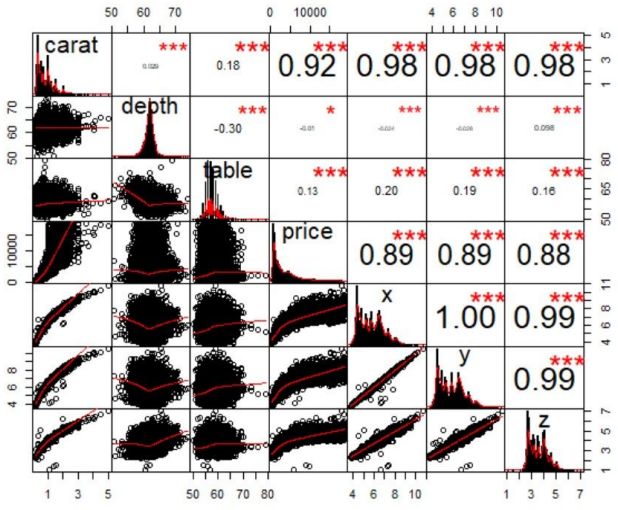








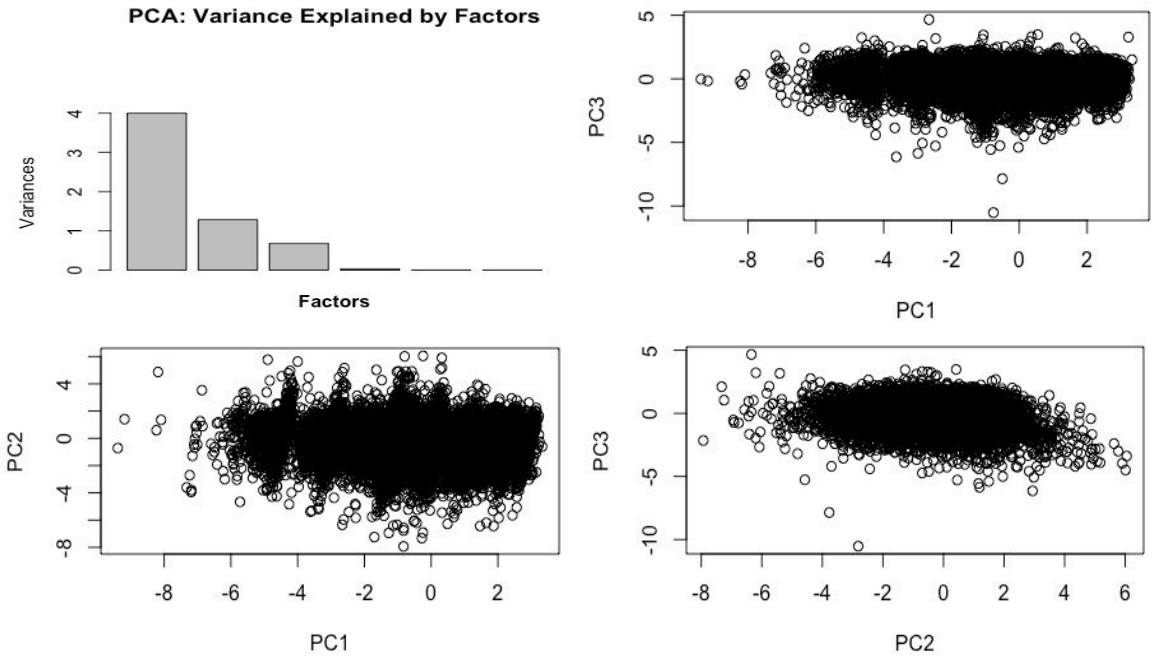




Modeling

In order to find the best model for determining the price of diamonds we developed five different models using different data mining techniques. The first model utilized variable clustering to facilitate variable reduction, the next used lasso to select variables, the third used quantile regression, the fourth used stepwise variable selection, and the last model used a regression tree.

1) PCA Clustering

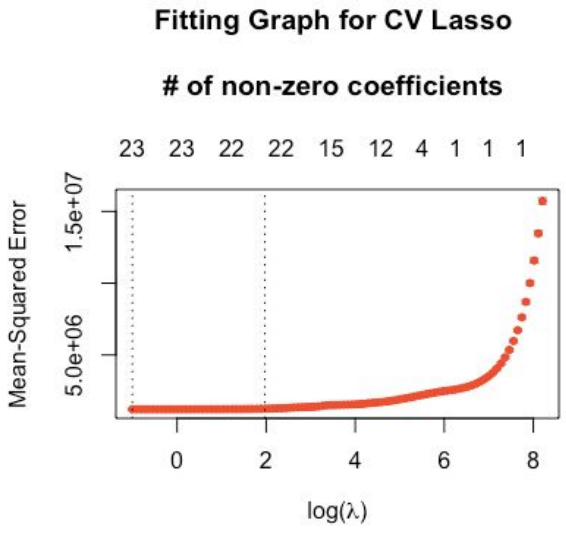


We used PCA to reduce the number of variables and intended to use the PCA factors in our regression model. Since Principle Component Analysis (PCA) only applies to the numeric variable, we first dropped the 3 categorical variables, cut, color and clarity, from the train dataset and then applied the PCA model. Based on the PCA: Variance Explained by Factors graph, we picked the first 3 factors, which explained most of the variation in the dataset. For each factor picked, we display the top features that are responsible for 3/4 of the squared norm of the loadings, includes x and y in PC1, depth in PC2 and table in PC3. The PCA model did not include the other two numeric variables, z and carat, in the first three factors.

Because each factor contained either a single quantitative variable or a combination of two quantitative variables, we found that it would be nearly as effective, and much simpler, to simply include the variables selected through the PCA factors (x, y, depth, table), add back the categorical variables we dropped previously, and run a linear model that included these variables.

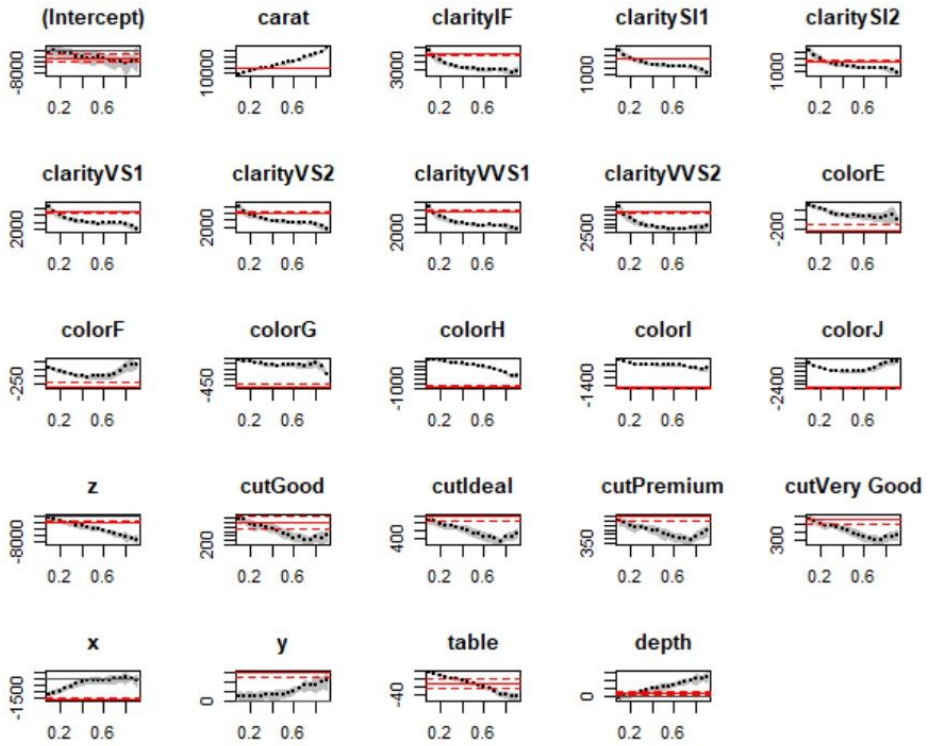
2) Linear with Lasso

We ran Lasso and created 2 linear models based on the variables selected. As the below graph shows, the minimum lasso value gave us 23 variables and 1 standard error lasso value gave us 22 variables. We then created 2 linear model: the minimum model with 23 variables and the 1 standard error model with 22 variables. We picked the minimum model as it has lower AIC score.



3) Quantile regression

We used 0.1, 0.5, and 0.9 as thresholds for quantile regressions. The R-Squared of 10% quantile regression is 0.41, 50% quantile is 0.78, and 90% quantile is 0.19. All the explanatory variables had heterogeneous impacts across different quantiles. For example, ClarityIF and ClaritySI2 both had positive effects on price, but these coefficients decrease across quantiles, which means their impacts are diminishing. In addition, Carat had increasing positive coefficients across quantiles, and we can see that the impact of carat on high quantiles was twice as large as that on lower quantiles.

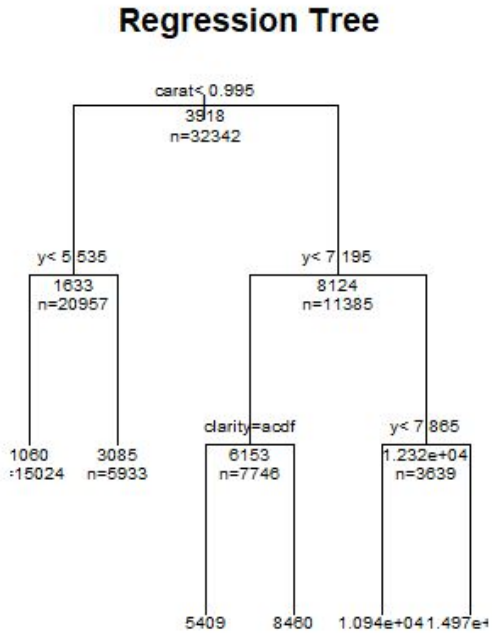


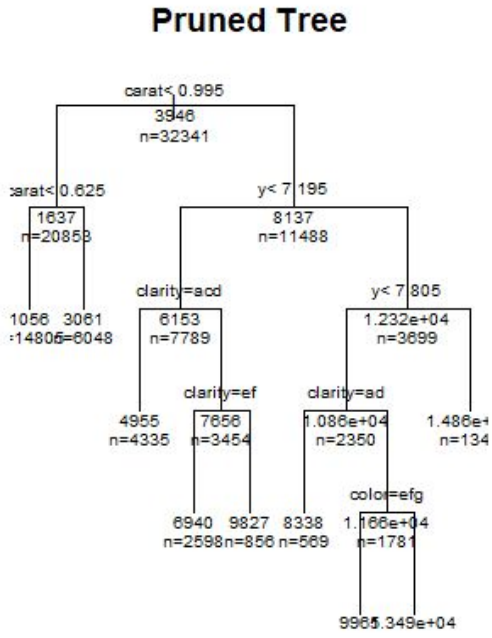
4) Stepwise linear model

We ran stepwise linear regression on price in our training dataset. The resulting model included carat, clarity, color, z, cut, x, y, table, and depth as explanatory variables and results in

an in-sample multiple R-Squared of 0.923. All of the explanatory variables are significant at the 0.05 level.

5) Regression tree





In order to predict price, we created a regression tree model. We built the original model and found that the ideal tree, with the lowest relative error and cp value, had six levels. We then pruned the tree to the height specified. The regression tree found that carat was the most important variable in determining the eventual price of the diamond, followed by the three dimension variables (y,x, and z), and then clarity.

Evaluation

We tested the out-of-sample performance and used R-squared to compare each of the models. We found the linear model that used Lasso variable selection and the linear model that used stepwise variable selection both had the lowest value of OOS R-squared (0.916736). We decided to use the stepwise model, as the Lasso model requires data transformation, as the

categorical variables must be in “dummy” form. Thus, the stepwise function requires less effort and is easily interpretable.

Deployment

The goal of our project is to predict diamond price more accurately over the long term to maximize our company profit. According to our Stepwise linear model, the variables that have a significant impact on price were carat, cut, color, clarity, depth, table and x, y ,z. We could calculate the price by entering each of the variable into a database, that automatically applies the stepwise formula, when new diamonds come in stock. However, this dataset does not include the costs related to acquiring and selling the diamonds, the sales volume for each kind of diamond, or the customer demand. Since diamonds are a precious and scarce gemstone, price predicting of diamond is related to diamond source. Also, regions with different cultures and populations will have their own demands on diamonds, which will impact diamond prices differently. Our dataset does not include variables about diamond source and regions, so the predicted price may fail to capture the true market trend of diamond. Additional data is needed to better estimate the selling price and market trends to maximize profit in certain locations.

However, there is also serious ethical considerations when participating in the sale of diamonds. The diamond industry has long been haunted by social criticism that its business is fueling illegal and inhumane diamond mining in Africa. Although focusing on profit maximization for our own company, we want to make sure our diamonds are coming from a clean ethical source.

Supplementary Material

Appendix:

Variable explanations of original data

Variable Number	Variable Name	Variable Description
1	price	price in US dollars ($326 ~ $18,823)
2	carat	weight of the diamond (0.2 ~ 5.01)
3	cut	quality of the cut (Fair, Good, Very Good, Premium, Ideal)
4	colour	diamond colour, from J (worst) to D (best)
5	clarity	a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
6	x	length in mm (0 ~ 10.74)
7	y	width in mm (0 ~ 58.9)
8	z	depth in mm (0 ~ 31.8)
9	depth	total depth percentage = z / mean(x, y) = 2 * z / (x + y) (43 ~ 79)
10	table	width of top of diamond relative to widest point (43 ~ 95)