Hillary Mahoney
Northwestern University
MSDS 454 – Sec 55
April 2019

"I never worry about diets. The only carrots that interest me are the number you get in a diamond." – Mae West

1. Data Quality Check

The two_months_salary dataset contains a total of 425 rows (individual diamond observations), seven variables including the four main diamond measures of cut, color, clarity, and carat, as well as the price of the diamond, the store, and channel which is the type of jeweler. For more information on the variables, Table 1.1 displays the variable name, data type, and a brief description. The color and clarity values are numeric, but were coded as such to represent the levels and types of diamond color and clarity characteristics.

Diamond color refers to the to how white, or colorless, the diamond appears. The more colorless, the better and more expensive the diamond. Color is typically represented on a graded scale ranging from D to Z with D being the best and having descriptions such as colorless and near colorless. In the dataset, color takes a value from one to ten where one is equivalent to a D grading and ten is equivalent to an M grading. Clarity is a measure of imperfections in the diamond; the fewer imperfections, even if invisible by the naked eye, the rarer and more expensive the diamond. Clarity is represented by a scale that rates flaws by their visibility to the naked eye. The scale ranges from a flawless (FL) diamond to one that has slight inclusions (SI1 and SI2). Again, this scale was converted to a numeric range in the dataset where one represents a flawless diamond and 11 represents a diamond with more than slight inclusions.

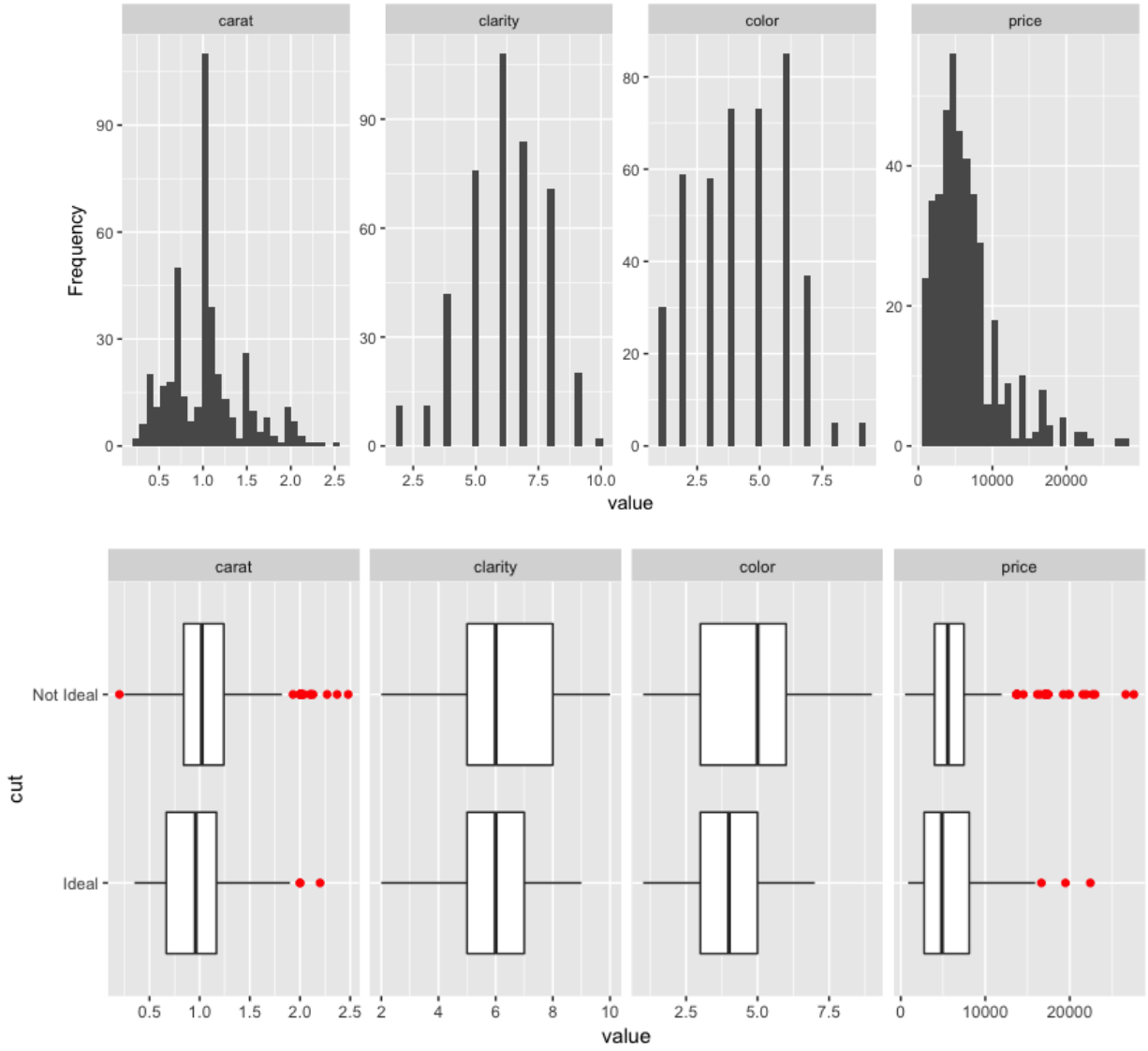
Table 1.1: Variable information

Variable Name	Data Type	Description
Carat	Double	Standard unit of weight used for gemstones
Color	Integer	Measures the visible hue of the stone
Clarity	Integer	Measures the purity of the stone
Cut	Character	Shape, or design, of the polished stone
Channel	Character	Type of jeweler
Store	Character	Store where diamond is priced
Price	Integer	Price in U.S. dollars

During the quality check of the data, it was discovered that there were not any missing values, but a number of outliers were identified for diamond price and carat. The distributions for these variables are displayed in Figure 1.2; note the skewed right distribution for price. There were 25 outliers detected and identified using boxplots, they are represented by the red dots in the boxplots below; values beyond the extremes of the whiskers are considered outliers. Outliers can have an adverse impact on statistical models such as ordinary least squares regression because they can skew results or they could be influential observations. Rather than removing the outliers, a new column was created that flagged the observation as an outlier or not so additional exploration could be done on these observations if necessary.

For diamond price, an outlier is any diamond worth more than $13,820 which is the extreme high-end whisker for price in Figure 1.2. For diamond weight (carat), anything weighing more than 1.93 ct. is considered an outlier. Many of the diamonds that were higher in price were also heavier, meaning many of the diamonds that were considered outliers for price were the same diamonds that were considered outliers given their weight. Five of the diamonds that were identified as outliers for carat were not abnormally expensive. In other words, even though they weighed more, their prices weren't identified as outliers. A full list of outliers can be found in the appendix.

Figure 1.2: Distributions for numeric variables



2. Exploratory Data Analysis

This section is broken down into two subsections, the first section focuses on traditional exploratory data analysis that encompasses checking of assumptions for a preliminary selection of appropriate models, univariate EDA to assess population distribution using the observed sample, as well as multivariate EDA to determine relationships among the explanatory variables. The second section is focused on model exploratory analysis which includes experimenting with several naïve models to further understand the direction and strength of existing relationships within the data.

2.1 Univariate and Multivariate Exploratory Data Analysis

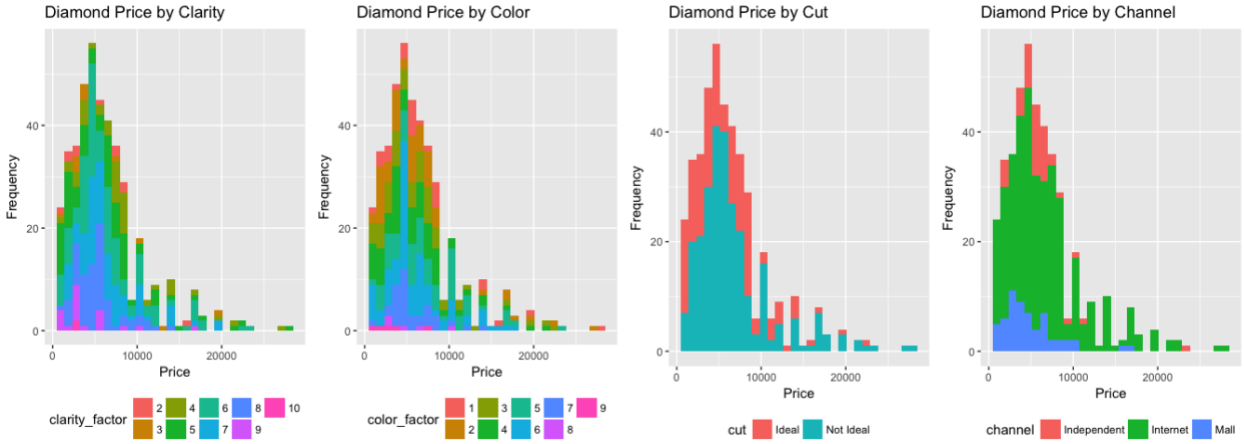
During the exploratory analysis phase, the variables were examined individually (univariate) as well as together (multivariate) to understand any relationships or how different diamond attributes contribute to the price. Because predicting diamond price is the objective, price was the first variable inspected. The distribution for price was briefly discussed in the data quality section and it was noted that it was right-skewed. Since a linear regression model may be a viable modeling solution for

predicting price, it would be wise to explore different transformations to the price variable to normalize the distribution. These transformations are discussed in greater detail in the modeling section.

When price was explored in combination with the other attributes in the dataset, it was noticed that price truly varied across clarity, color, cut and channel. That is, even some of the more expensive diamonds had poor clarity and color, as well as non-ideal cut types. However, as suspected, the better the color and clarity (i.e. the less color and imperfections a diamond has), the more expensive the diamond. Figure 2.1 shows histograms for diamond price broken down by different attributes.

Since a majority of the diamonds were sourced from Blue Nile, an internet-based jeweler, the Blue Nile store varied the most in price and contained many of the expensive and heavier diamonds that were identified as outliers in the data quality section. Ashford, which is also an internet-based jeweler, also contained a lot of pricier and heavier diamonds. Kay jewelers and Goodmans were the other two stores that sourced more expensive diamonds, but only a handful of diamonds were observed from these two stores so their price distributions were sparse and disparate compared to the internet-based stores. A full list of stores and the number of diamonds sourced from each can be found in the appendix.

Figure 2.1: Price distribution by clarity, color, cut, and channel



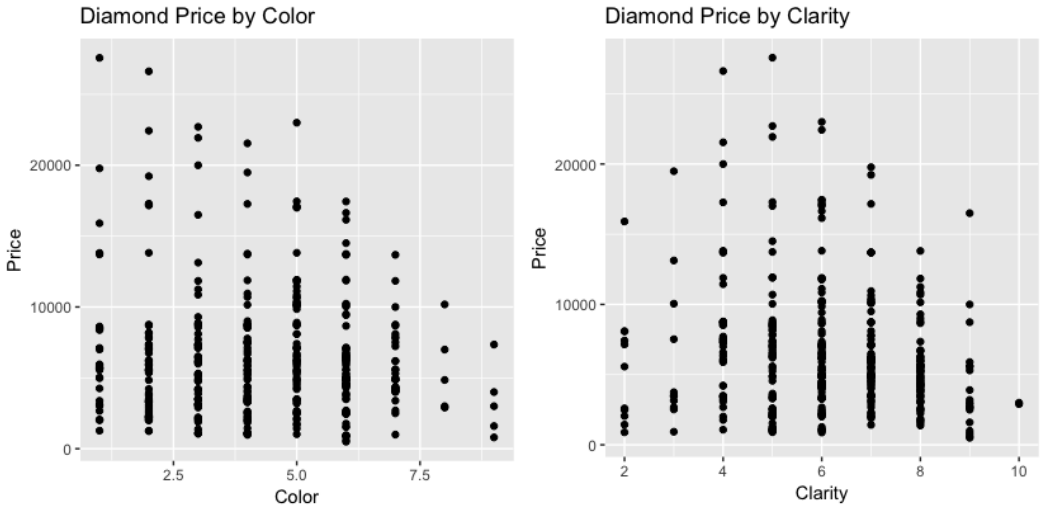
Since a diamond's worth increases based on its color (or lack thereof) and clarity, the relationships between price and color and price and clarity were explored next. From the histograms in Figure 2.1, we could see that the range of prices varied by color and clarity, meaning an expensive diamond could have imperfections and/or have a tint. Alternatively, there could be relatively inexpensive diamonds with great clarity and color. Therefore, it was crucial that these relationships were explored in greater detail to truly understand how they might impact price.

Figure 2.2 displays scatterplots of price by color and clarity. Starting with color, there is a noticeable decrease in price as the color value increases, which is expected since color is a graded scale ranging from one to ten with one being the best and ten being the worst. So, as a diamond's color quality decreases, the less it will typically be worth. Notice that the plot is quite dense with observations between $0 and $10,000 across the various color values. It does get less dense as color nears the high end of the scale, but a majority of the diamonds are less than $10,000 with color values between three and six.

Similar to color, diamond clarity also seems to vary with price. That is, diamonds with poor clarity (slight imperfections) can still be somewhat expensive, relatively speaking. Additionally, notice that the few diamonds with great clarity (clarity value of 2) are comparatively less expensive than diamonds with worse clarity. Based on the scatterplot, most diamonds are of adequate or moderate clarity, meaning they may contain some slight imperfections that aren't visible to the naked eye. It seems that clarity

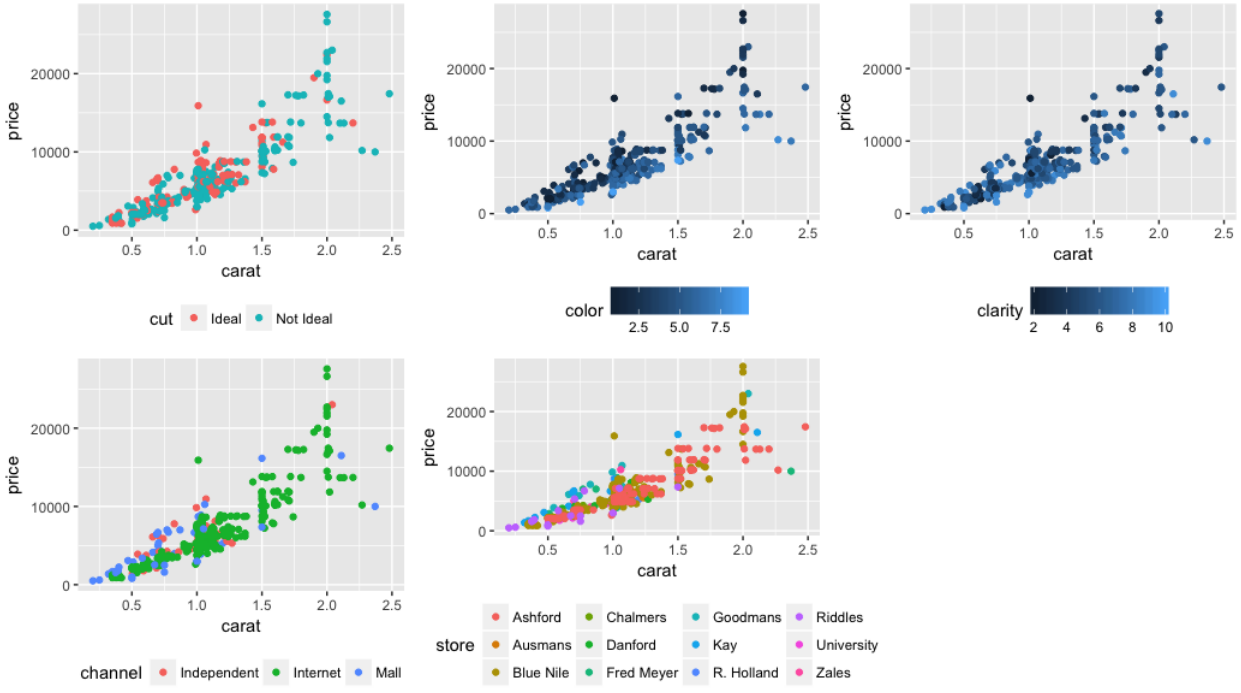
may have less impact on price than color does, but this will be covered in greater detail in the modeling section where variable importance is discussed.

Figure 2.2: Scatterplots of price by color and clarity



Lastly, diamond price was explored in conjunction with weight (carat). As assumed, price and carat are highly correlated (r=0.88), but their relationship isn't explicitly linear. Figure 2.3 displays scatterplots of price and carat grouped by the different attributes in the dataset; note that the points form a slight fan shape that could be curvilinear. As diamond weight increases, the price range increases, too, giving it that fan shape.

Figure 2.3: Scatterplots of price by carat



In addition to understanding any relationships between price and a given attribute, it is also important to know if relationships exist between two or more diamond attributes such as cut and color, etc. Therefore, time was dedicated to examining any potential associations between other variables. Charts that support this exploratory analysis are included in the appendix, but the most noteworthy findings are that diamonds sourced from a mall jeweler have higher clarity values (so, poor clarity) and tend to have smaller weights (less carats). Independent jewelers have diamonds with a wide range of color and clarity, but are smaller in carats, comparatively speaking. Additionally, all stores, regardless of channel, had more not ideal cut types than ideal diamond cuts.

2.2 Model Based Exploratory Analysis

To transition from exploratory analysis to building models, a few simple, exploratory models were created to generate insights related to some of the relationships discovered previously. For example, it's hard to understand the relationship between cut and price, because cut and carat, and carat and price are associated. So, it's possible to use a model to remove the very strong relationship between price and carat so we can explore the subtleties that remain. Furthermore, a model can quantitatively assess variable importance.

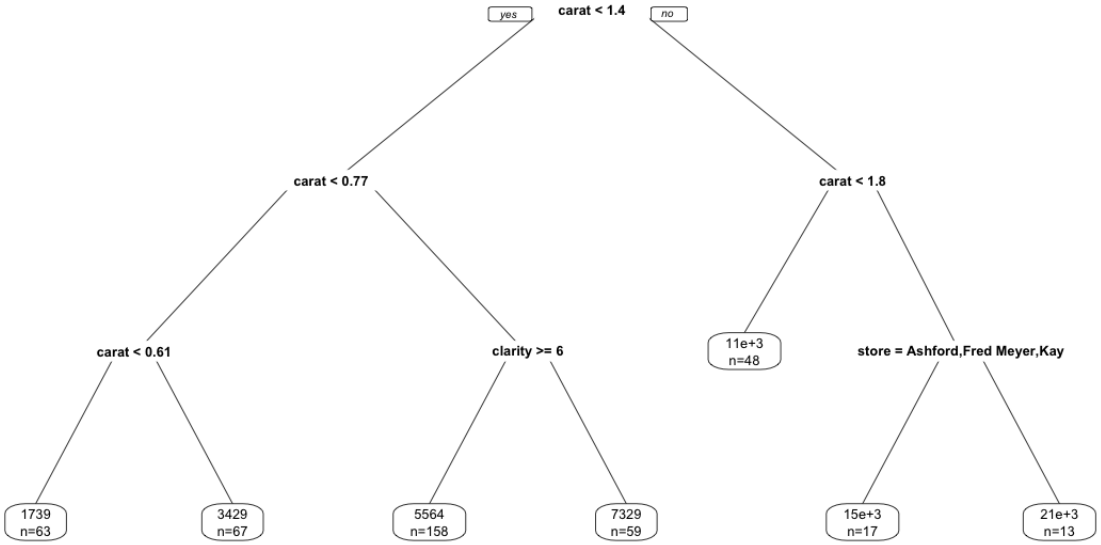
The first model was a simple regression tree to predict price using all the available explanatory variables as predictors. Regression trees are an effective way to probe the data with minimal specification in the modeling process. The tree output is displayed in Figure 2.4 and consists of a series of splitting rules for carat, clarity, and store as they were actually used in the construction of the tree. The top node in the tree contains all of the observations in the data set and the split assigns observations with less than 1.41 carats to the left branch and is further subdivided by carat size and clarity. The predicted price for these diamonds is given by the mean response value for the diamonds in the data set with less than 1.41 carats. For such diamonds, the average price is $4758. In other words, if a diamond has less than 1.41 carats, its average price is $4758 (not considering clarity). If the diamond is between 0.77 ct. and 1.41 ct., but the clarity is greater than or equal to six, then the average price is $5,564. Else, if the diamond is between 0.77 ct. and 1.41 ct., but the clarity is less than six, then the average price is $7,329. When the carat is less than 0.61, then the average price is $1,739. Whereas if the carat is greater than 0.61, but less than 0.77, then the diamond price is $3,429.

Diamonds with more than 1.41 carats are assigned to the right side of the branch where they are further subdivided by carat and store. The predicted price for these diamonds is given by the mean response value for the diamonds in the data set with more than 1.41 carats, that is, if a diamond's carat is greater than 1.41 then the average price is $13,467 (not considering the store). If the diamond is greater than 1.41 ct., but less than 1.75 ct., then the average price is $10,839. When the diamond is greater than 1.75 carats and the store is Blue Nile or Goodmans, then the average price is $21,188. Whereas when the diamond is greater than 1.75 carats, but the store is Ashford, Fred Meyer, or Kay then the average price is $14,982.

Based on the variables that were selected by the tree, it is clear that carat is an extremely important factor in determining diamond price which corroborates the relationship discovered between price and carat in the exploratory analysis phase. Diamonds with less carats, on average, have lower prices than diamonds with more carats. When a diamond has fewer carats, clarity plays a significant role in determining price. Alternatively, diamonds with more carats have higher prices and the store at which the diamond is sold plays an important role in determining diamond price.

Despite shedding light on variable importance, unfortunately, a single tree model tends to have high variance, resulting in unstable predictions. However, by bootstrap aggregating (bagging) regression trees, this technique can become quite powerful and effective. Moreover, this provides the fundamental basis of more complex tree-based models such as random forests which is a technique that will be discussed in the modeling section.

Figure 2.4: Regression Tree



Before continuing with the model-based EDA, the dataset was split into a training set and a test using a 70/30 split. That is, 70% of the dataset was randomly assigned to a training set that would be used to create models and the remaining 30% was assigned to a test set that would be used to validate the models created from the training data.

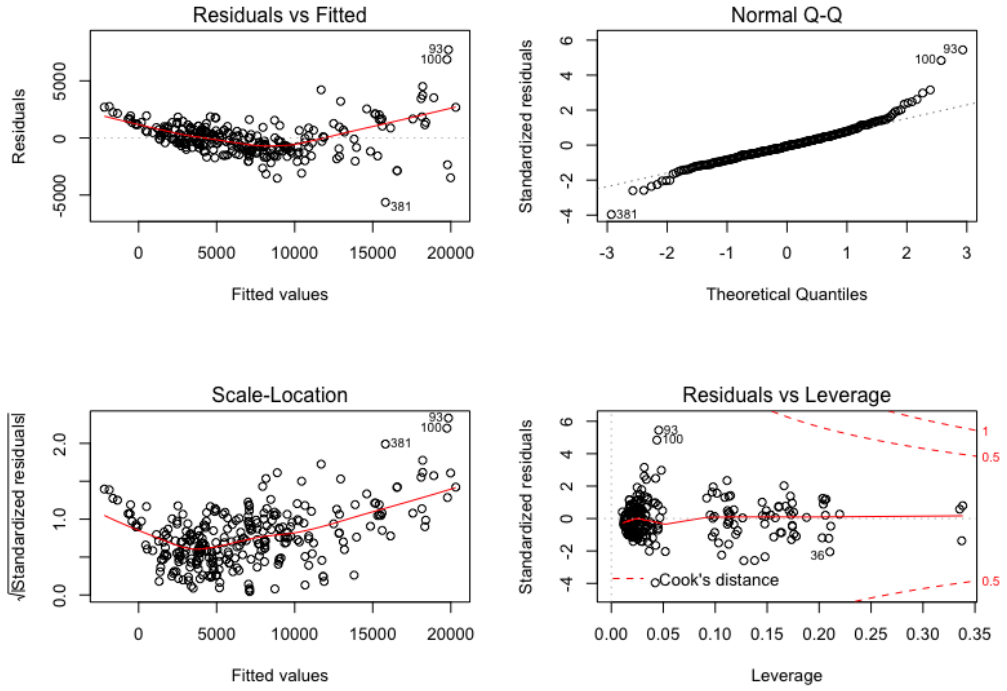
Two naïve regression models were constructed using backwards variable selection, this means it starts with a full model and removes insignificant variables one at a time until all predictors are significant. The models were fit using all of the predictor variables (untransformed), but one model used color and clarity as numeric values and the other model used them as factors. The purpose of these models was to understand which variables were significant and use them as a baseline for improvement. Surprisingly, the models didn't yield as useful as the regression tree. During the backward variable selection process, both models only removed channel. Because both models produced very similar results, only one will be discussed. The first OLS equation (keeps color and clarity as numeric values) is as follows:

price=2281.56
+ 11066.46*__carat__
– 825.85*__color__
– 728.29*__clarity__
– 614.02*__cutNotIdeal__
+ 1285.08*__storeAusmans__
+516.29*__soreBlueNile__
– 246.51*__storeChalmers__
+ 1014.97*__storeDanford__
+ 2785.14*__storeFredMeyer__
+ 4573.19*__storeGoodmans__
+ 4001.81*__storeKay__
+ 2858.27*__storeRHolland__
+ 5434.38*__storeRiddles__
+ 1466.22*__storeUniversity__
+ 3812.54*__storeZales__

The regression coefficients represent the change in the expected price with a one-unit increase for a given variable holding all others constant. For example, for every additional carat, the expected price increases by $11,066, holding all other variables constant. Similarly, if a diamond's color increases (remember, the higher the number the worse the quality of the color), the expected price decreases by about $826, holding all other variables constant. Lastly, among diamonds with similar carat size, cut, color and quality, diamonds from Ausman's typically cost an additional $1,285 compared to diamonds from Ashford. The remaining store factors can be interpreted the same way.

Figure 2.5 displays the diagnostic plots that can be used to subjectively assess goodness of fit by checking for heteroscedasticity, normality, and influential observations. In Residuals vs. Fits plot in the top, left corner, notice the residuals form a slight curve and disperse as you move along the x-axis. This is not ideal and suggests the model may not be a great fit, but potentially transforming the response or predictor variables may improve model fit. The Q-Q plot shows that it is probably safe to assume price is normally distributed, but there are observations on both ends that appear to be outliers. Additionally, these same potential outliers have high leverage meaning they may strongly influence the fitted values. Therefore, it is probably worth additional examination of these observations.

Figure 2.5: OLS Regression Diagnostic Plots



Next, a few regression models were fit using the step AIC method which performs stepwise model selection by Akaike Information Criterion (AIC). This method actually produced the same model that was fit using backward stepwise variable selection. The step AIC method attempts to minimize AIC so it adds and removes predictors until the lowest AIC is achieved. It just so happens, the model produced using backward variable selection was also the model with the lowest AIC. The models fit using variable selection attempt to explain the data in a simple way by removing redundant predictors. Unnecessary predictors add noise to the estimation of other quantities that we are interested in and can cause collinearity. The saying goes, "the best model is a simple model."

3. Modeling Results

A series of models were created based on the findings from the exploratory data analysis and model-based exploratory analysis sections. Each modeling technique has its own subsection as several different models may have been fit using the required technique.

3.1 Linear Regression without Interaction Terms

The first required model was a linear regression model without interaction terms. Several models were fit to the data with different transformations applied to different variables for each model. It was previously mentioned that there's a strong relationship between price and carat, but it appeared somewhat curvilinear so a transformation to one or both variables may improve their association. To start, price was transformed using the natural log and stepwise regression with the AIC method was applied. This resulted in little improvement over the naïve regression models created in the model-based EDA section. Therefore, different transformations were applied to carat. The natural log transformation, square root, and cubic root transformations all yielded substantially better results. The best performing model was log(price) ~ log(carat) + clarity_factor + store, where price and carat were transformed using the natural log and clarity was converted to a factor. The full equation for the model is below.

*Log(***price***)* = 8.76412
+ 1.56437**log(***carat***)*
– 0.10854***clarity_factor3**
– 0.07784***clarity_factor4**
– 0.12396***clarity_factor5**
– 0.26365***clarity_factor6**
– 0.27022***clarity_factor7**
– 0.31621***clarity_factor8**
– 0.64951***clarity_factor9**
- 1.24463***clarity_factor10**
+ 0.08667***storeAusmans**
+ 0.04879***soreBlueNile**
– 0.10713***storeChalmers**
+ 0.06186***storeDanford**
+ 0.19943***storeFredMeyer**
+ 0.57954***storeGoodmans**
+ 0.52638***storeKay**
+ 0.32600***storeRHolland**
+ 0.25479***storeRiddles**
- 0.04331***storeUniversity**
+ 0.41314***storeZales**

Because there are log transformations performed on both the response and predictor variables, the regression coefficients in this model represent the approximate percentage change in the average (geometric) price with a one-percent increase for a given variable, holding all others constant. For example, a one-percent increase in carat, results in an approximate 1.56% increase in diamond price, on average, holding all other variables constant. Conversely, diamond price decreases, on average, as clarity worsens. Diamonds with similar carat size, but slightly worse clarity (clarity=3) cost about 10.8% less, on average, compared to diamonds with a clarity level = 2. The other clarity factors can be interpreted similarly. Lastly, among diamonds with similar carat size and clarity, diamonds from Ausman's typically cost 8.67% more compared to diamonds from Ashford. Whereas, diamonds from Chalmer's cost about 10.7% less, on average, given that the diamonds are of similar carat and clarity. The remaining store factors can be interpreted the same way.

Figure 3.1: OLS Regression Diagnostic Plots

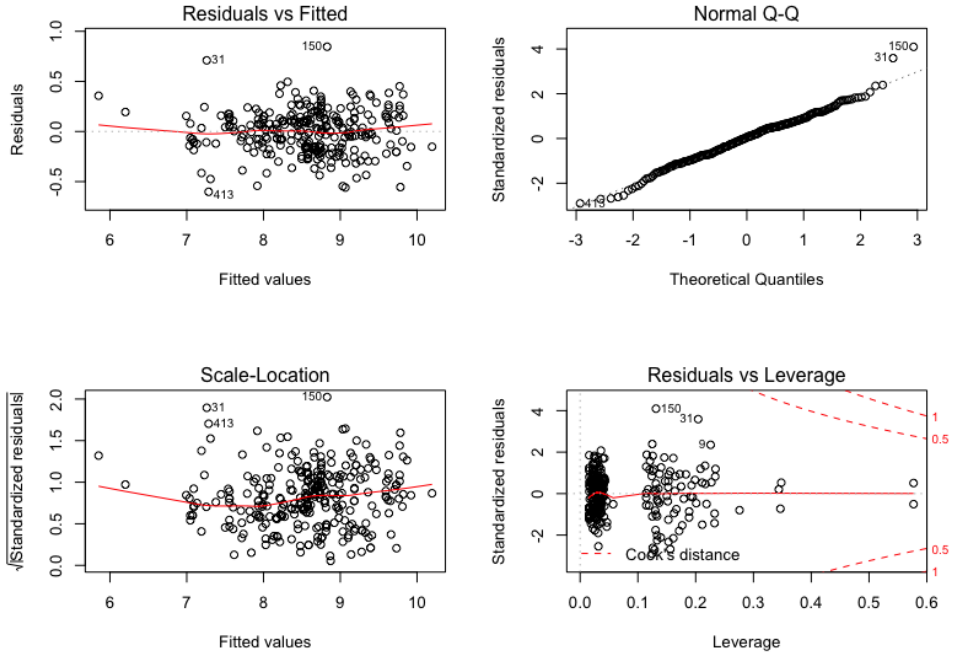


Figure 3.1 contains the regression diagnostic plots for the model. There is no longer an evident pattern in the residuals, though they are condensed into one area of the plot. The observations plotted in the Q-Q plot also show an improvement, but there are still some observations with high leverage. Overall, this model seems to be a decent fit and a substantial improvement over the original exploratory models created. When fit on the test set, the mean absolute percentage error (MAPE) was approximately 1.92%, meaning on average across all of the observations in the test set, the predicted prices deviated from the actual prices by about 1.92%.

3.2 Linear Regression with Interaction Terms

Similar to the prior section, a number of models were created and tested using different transformations on the response and predictor variables, as well as trying different interaction term combinations. After assessing the goodness of fit and predictive accuracy for each model, the following model was chosen: log(price) ~ log(carat)*clarity_factor + color_factor

Similar to the linear regression model without interaction terms, because there are log transformations performed on both the response and predictor variables, the regression coefficients in this model represent the approximate percentage change in the average (geometric) price with a one-percent increase for a given variable, holding all others constant. However, with the interaction terms there is an additional component when interpreting the coefficients. Because of the interaction, the effect of carat size is different depending on the diamond's level of clarity, therefore, there is no longer any unique effect of carat size. For example, assume the clarity level =3 and there is a 1% increase in carat, respectively, the average price would increase by approximately 1.8% (the effect is determined by the coefficient of log(carat) which is 2.244 plus the coefficient of the interaction log(carat)*clarity_factor3 which is -0.418). The other interaction terms can be interpreted similarly.

*Log(***price***)* = 9.10129
+ 2.24444**log(***carat***)**
– 0.16158***clarity_factor3**
– 0.16968***clarity_factor4**

– 0.23051*clarity_factor5
– 0.25314*clarity_factor6
– 0.31617*clarity_factor7
– 0.36466*clarity_factor8
– 0.47257*clarity_factor9
- 0.47739*clarity_factor10
- 0.04867*color_factor2
– 0.14149*color_factor3
– 0.17064*color_factor4
- 0.23959*color_factor5
- 0.36894*color_factor6
- 0.45997*color_factor7
- 0.63465*color_factor8
- 0.67801*color_factor9
- 0.41823*log(carat):clarity_factor3
- 0.58846*log(carat):clarity_factor4
- 0.52698*log(carat):clarity_factor5
- 0.61275*log(carat):clarity_factor6
- 0.59499*log(carat):clarity_factor7
- 1.11230*log(carat):clarity_factor8
- 0.90938*log(carat):clarity_factor9

Figure 3.2: OLS Regression Diagnostic Plots

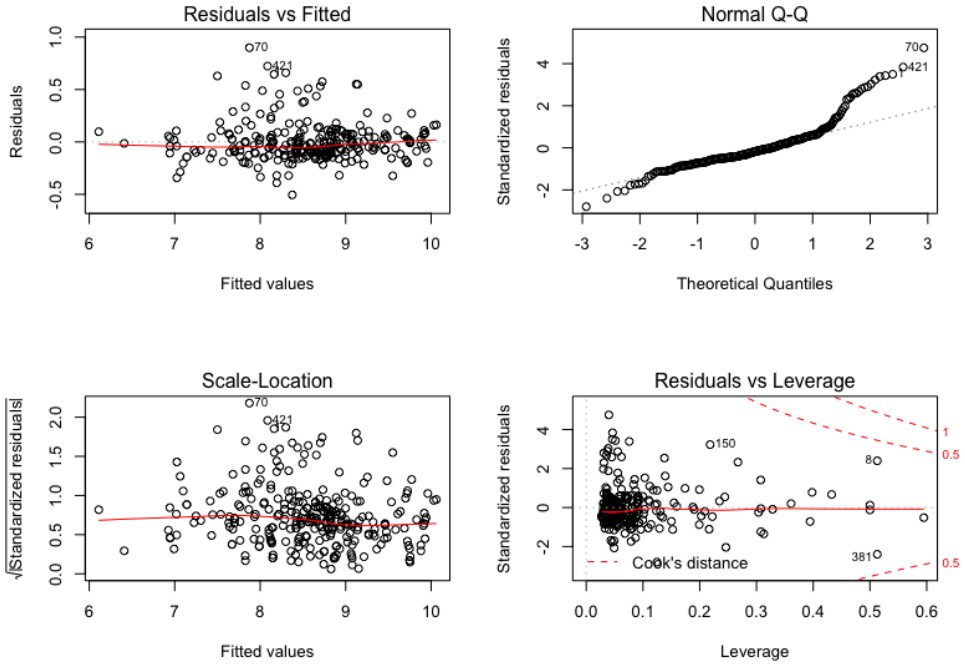


Figure 3.2 contains the regression diagnostic plots for the model. The observations plotted in the Q-Q plot show a decline over the previous model, potentially violating some of the underlying assumptions regarding normality for linear regression. The residuals also appear more condensed and there are still some observations with high leverage. Considering the error metrics, this model seems to be a decent fit on the test data. The mean absolute percentage error (MAPE) was approximately 1.55%, meaning on average across all of the observations in the test set, the predicted prices deviated from the actual prices by about 1.55%. This model actually outperformed the others in both the in-sample and out of sample error metrics, but there is some concern about the diagnostic plots and the usefulness of this

model. The Variance Inflation Factors (VIF) were calculated to check for multicollinearity to see if that explained any of the deviation and heteroscedasticity observed in the residuals versus fits plots.

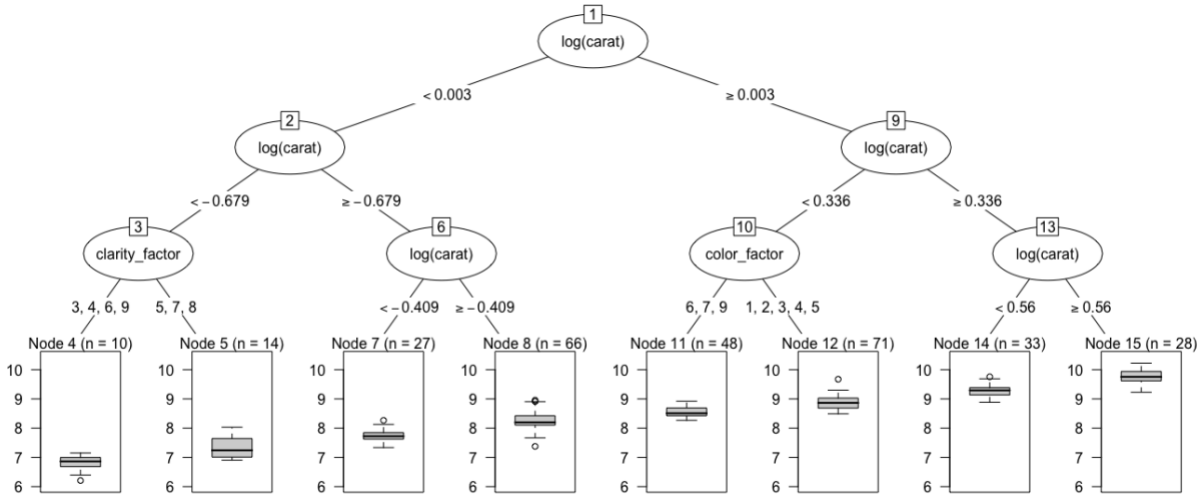
3.3 Regression Tree

The tree output is displayed in Figure 3.3 and consists of a series of splitting rules for log(carat), clarity, and color as they were actually used in the construction of the tree. The top node in the tree contains all of the observations in the training set and the split assigns observations with log(carat) less than 0.00349 carats to the left branch and is further subdivided by carat size and clarity. The predicted price for these diamonds is given by the mean response value for the diamonds in the data set with less than log(carat) of 0.00349. Diamonds with more than 0.00349 log(carat) assigned to the right side of the branch where they are further subdivided by carat and color. The predicted price for these diamonds is given by the mean response value for the diamonds in the data set with more than 0.00349 log(carat).

Based on the variables that were selected by the tree, it is clear that carat is an extremely important factor in determining diamond price which is consistent with the regression tree that was created previously and also consistent with the other models that were constructed. Diamonds with less carats, on average, have lower prices than diamonds with more carats. When a diamond has fewer carats, clarity plays a significant role in determining price. Alternatively, diamonds with more carats have higher prices and the color plays an important role in determining diamond price.

This model underperformed compared to the linear regression and random forest models. The in-sample and out of sample mean square error were greater than the other models, and it wasn't able to explain as much variation in the price. The mean absolute percentage error was also higher at 2.7%.

Figure 3.3: Regression Tree

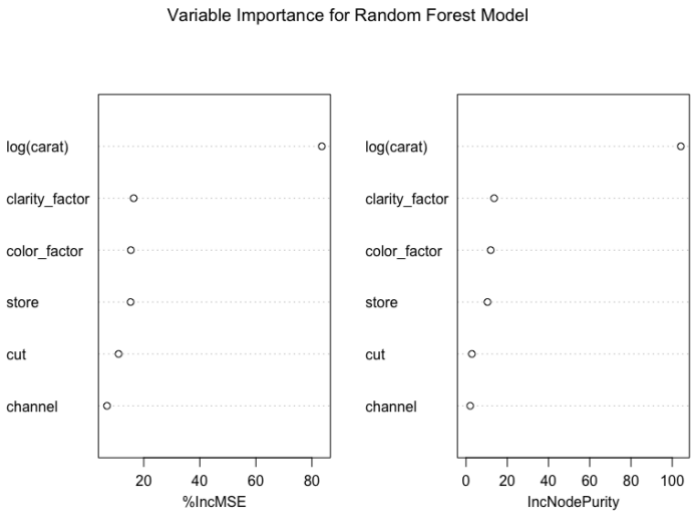


3.4 Random Forest

A different approach was taken for the random forest models. Essentially, a random forest model was trained using the log transformation of price and carat plus the remaining variables (untransformed). The model was tuned via grid search and 4-fold cross validation and the parameters that lead to the highest root means squared error (RMSE) on the cross folds were selected. This resulted in a model composed of 400 trees of varying depths with six variables available for selection by each tree. This configuration explained about 91% of the total variation in diamond price and yielded a mean absolute percentage error of 1.58% on the test data. Figure 3.4 shows the variable importance plot for

the model. As suspected, carat is the most important variable in predicting diamond price. This has now been observed consistently across modeling techniques

Figure 3.4: Random Forest Variable Importance



4. Model Comparison

The selected model from each subsection is represented in Table 4.1. Overall, the linear regression model with interaction terms outperformed the other models, but having both the response variable and a predictor variable log transformed plus interaction terms increases the difficulty of interpretation. Additionally, even though it had low predictive error rates out of sample, the diagnostic plots for the residuals were not ideal. The diagnostic plots are a subjective means of assessing goodness of fit, but they relate directly to the underlying assumptions of linear regression. Violations to these assumptions warrants further scrutiny of the model fit.

The random forest model also performed extremely well out of sample and had low prediction error. It automatically does a good job of finding interactions, as well so they don't need to be manually specified or created. Additionally, random forests don't make assumptions that the response has a linear relationship with the predictors so we're not limited to predefined, linear relationships. Despite its flexibility and immense predictive power, it is difficult to interpret. That being said, depending on the use case of this analysis, that is, if we're striving for prediction over interpretability, I would move forward with the random forest model and continue training and tuning on additional data (if possible). If interpretability is the key, then I am confident that the linear regression model without interaction terms would be a suitable model in predicting diamond price.

Table 4.1: Comparison of Modeling Results

MODEL TECHNIQUE	MODEL	MSE	MAPE	% VAR. EXP
LINEAR REGRESSION W/O INTERACTION TERMS	log(price) ~ log(carat) + clarity_factor + store	0.03809	0.019169 (~1.92 %)	0.9070 (adj. R^2)
LINEAR REGRESSION W/ INTERACTION TERMS	log(price) ~ log(carat)*clarity_factor + color_factor	0.03385	0.015495 (~1.55 %)	0.9295 (adj. R^2)
REGRESSION TREE	log(price) ~ log(carat) + clarity_factor + color_factor + cut + channel + store	0.08595	0.027030 (~2.70 %)	0.8225
RANDOM FOREST	log(price) ~ log(carat) + color_factor + clarity_factor + cut + channel + store	0.03426	0.015792 (~1.58%)	0.9133

Appendix

Table 1: Identified Outliers

	carat	color	clarity	cut	channel	store	price	outlier_column
1	2.04	5	6	Not Ideal	Independent	Goodmans	23000	carat_outliers
2	2.37	7	9	Not Ideal	Mall	Fred Meyer	9999	carat_outliers
3	2.11	3	9	Not Ideal	Mall	Kay	16500	carat_outliers
4	2	2	6	Ideal	Internet	Blue Nile	22431	carat_outliers
5	2	1	5	Not Ideal	Internet	Blue Nile	27575	carat_outliers
6	2	4	4	Not Ideal	Internet	Blue Nile	21545	carat_outliers
7	2	3	5	Not Ideal	Internet	Blue Nile	21933	carat_outliers
8	2	3	5	Not Ideal	Internet	Blue Nile	22706	carat_outliers
9	2	6	5	Not Ideal	Internet	Blue Nile	14504	carat_outliers
10	2	2	4	Not Ideal	Internet	Blue Nile	26623	carat_outliers
11	2	6	6	Ideal	Internet	Blue Nile	16641	carat_outliers
12	2	1	7	Not Ideal	Internet	Blue Nile	19776	carat_outliers
13	2	2	7	Not Ideal	Internet	Blue Nile	19227	carat_outliers
14	2.27	8	6	Not Ideal	Internet	Ashford	10180	carat_outliers
15	2.02	7	6	Not Ideal	Internet	Ashford	11840	carat_outliers
16	2.13	7	4	Not Ideal	Internet	Ashford	13680	carat_outliers
17	2.1	6	7	Not Ideal	Internet	Ashford	13680	carat_outliers
18	2.2	6	7	Ideal	Internet	Ashford	13700	carat_outliers
19	2.01	6	5	Not Ideal	Internet	Ashford	13740	carat_outliers
20	2.01	5	5	Not Ideal	Internet	Ashford	17010	carat_outliers
21	2.01	5	6	Not Ideal	Internet	Ashford	17020	carat_outliers
22	2.02	5	6	Not Ideal	Internet	Ashford	17100	carat_outliers
23	2.48	6	6	Not Ideal	Internet	Ashford	17440	carat_outliers
24	2.01	5	6	Not Ideal	Internet	Ashford	17450	carat_outliers
25	2.04	5	6	Not Ideal	Independent	Goodmans	23000	price_outliers
26	1.5	6	6	Not Ideal	Mall	Kay	16150	price_outliers
27	2.11	3	9	Not Ideal	Mall	Kay	16500	price_outliers
28	2	2	6	Ideal	Internet	Blue Nile	22431	price_outliers
29	2	1	5	Not Ideal	Internet	Blue Nile	27575	price_outliers
30	2	4	4	Not Ideal	Internet	Blue Nile	21545	price_outliers
31	2	3	5	Not Ideal	Internet	Blue Nile	21933	price_outliers
32	2	3	5	Not Ideal	Internet	Blue Nile	22706	price_outliers
33	2	6	5	Not Ideal	Internet	Blue Nile	14504	price_outliers
34	1.9	4	3	Ideal	Internet	Blue Nile	19488	price_outliers
35	1.93	3	4	Not Ideal	Internet	Blue Nile	19997	price_outliers
36	2	2	4	Not Ideal	Internet	Blue Nile	26623	price_outliers
37	2	6	6	Ideal	Internet	Blue Nile	16641	price_outliers
38	2	1	7	Not Ideal	Internet	Blue Nile	19776	price_outliers
39	2	2	7	Not Ideal	Internet	Blue Nile	19227	price_outliers
40	1.01	1	2	Ideal	Internet	Blue Nile	15909	price_outliers
41	2.01	5	5	Not Ideal	Internet	Ashford	17010	price_outliers
42	2.01	5	6	Not Ideal	Internet	Ashford	17020	price_outliers
43	2.02	5	6	Not Ideal	Internet	Ashford	17100	price_outliers
44	1.78	2	7	Not Ideal	Internet	Ashford	17160	price_outliers
45	1.76	2	6	Not Ideal	Internet	Ashford	17230	price_outliers
46	1.82	4	4	Not Ideal	Internet	Ashford	17270	price_outliers
47	1.7	2	5	Not Ideal	Internet	Ashford	17290	price_outliers
48	2.48	6	6	Not Ideal	Internet	Ashford	17440	price_outliers
49	2.01	5	6	Not Ideal	Internet	Ashford	17450	price_outliers

Figure 1: Percent of values missing per variable

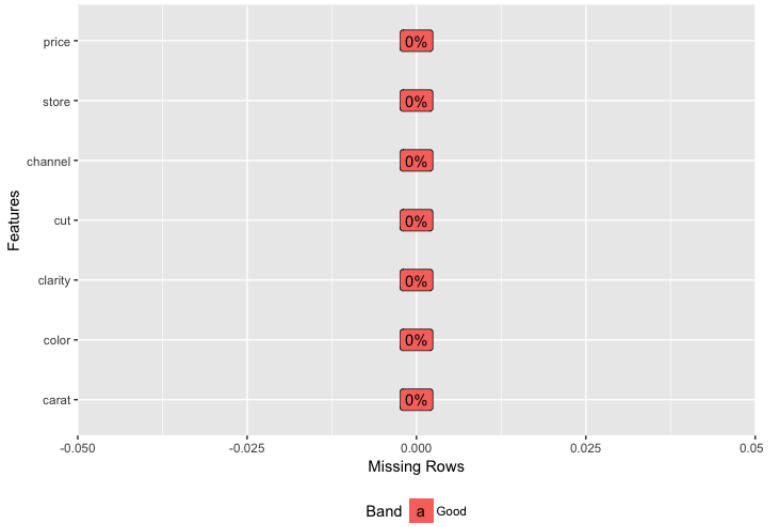


Figure 2: Density plots for numeric variables

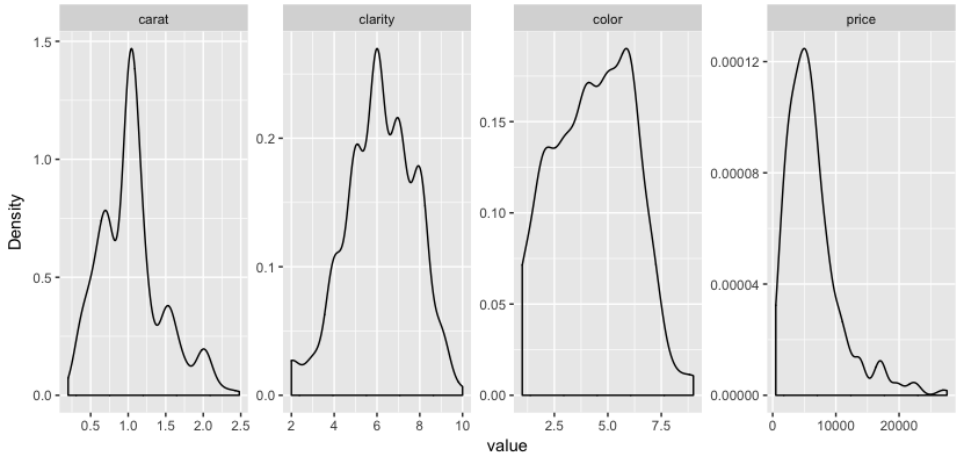


Figure 3: Bar plots for character variables

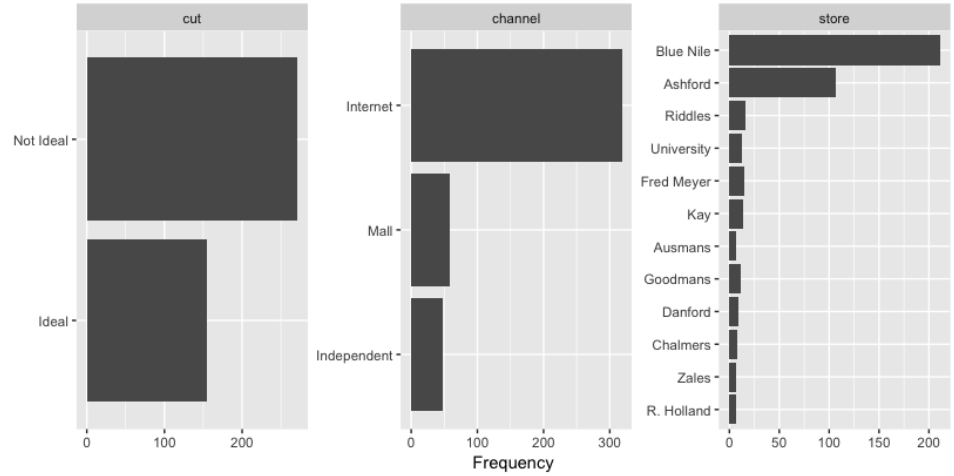


Figure 4: Correlation Matrix Plot

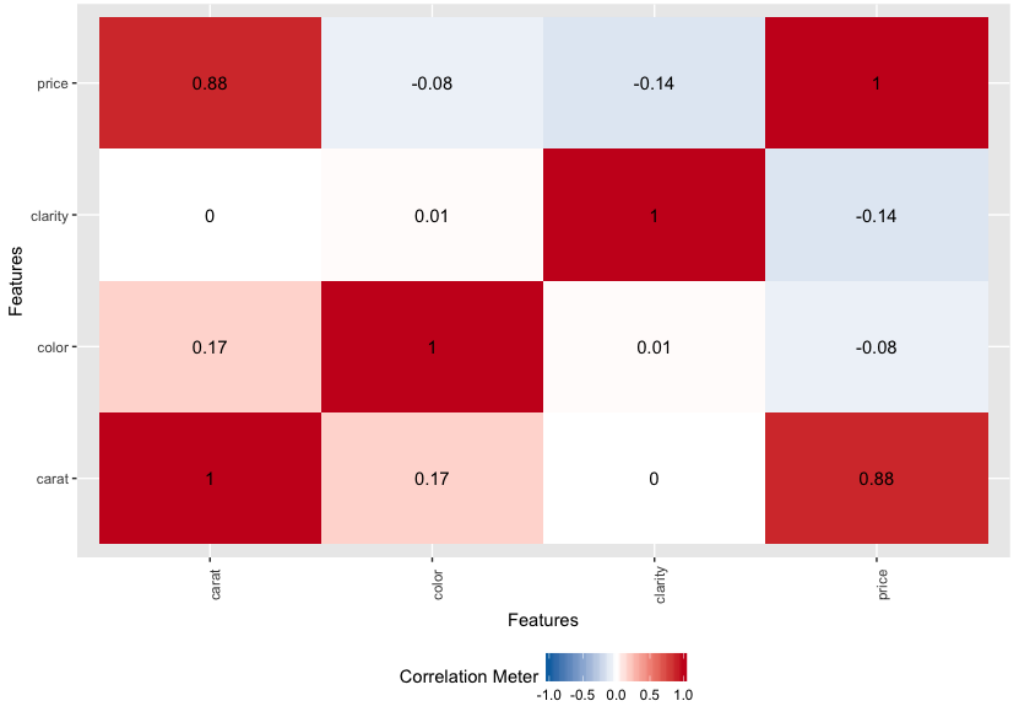
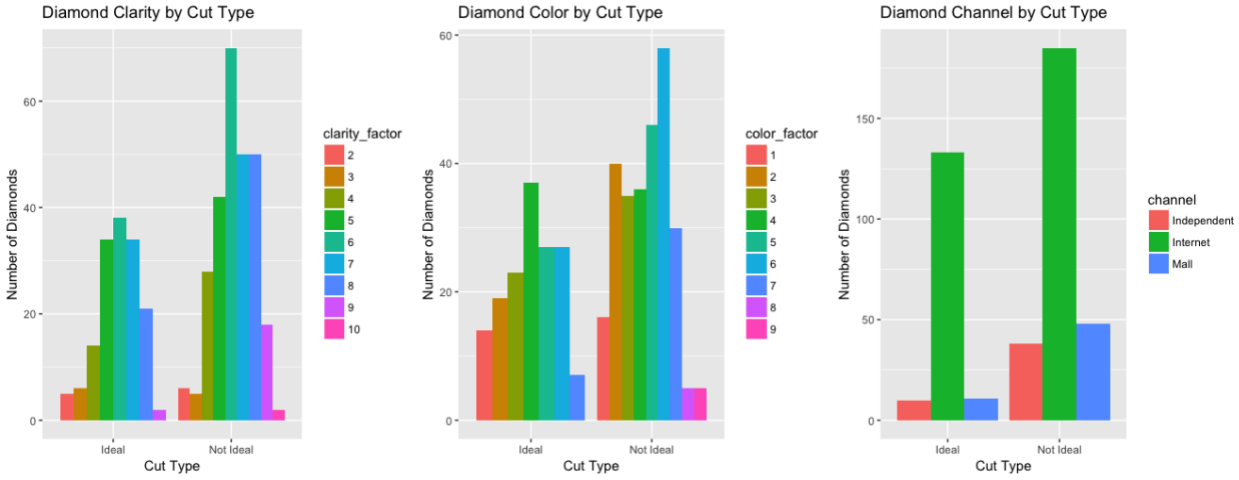
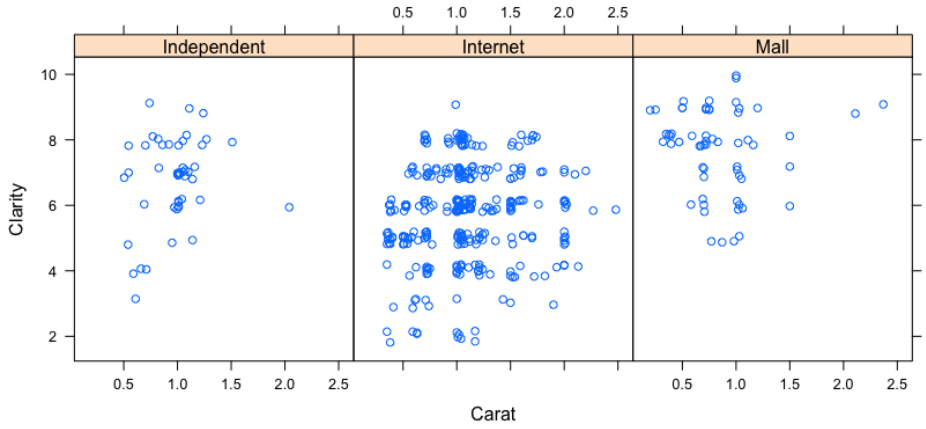
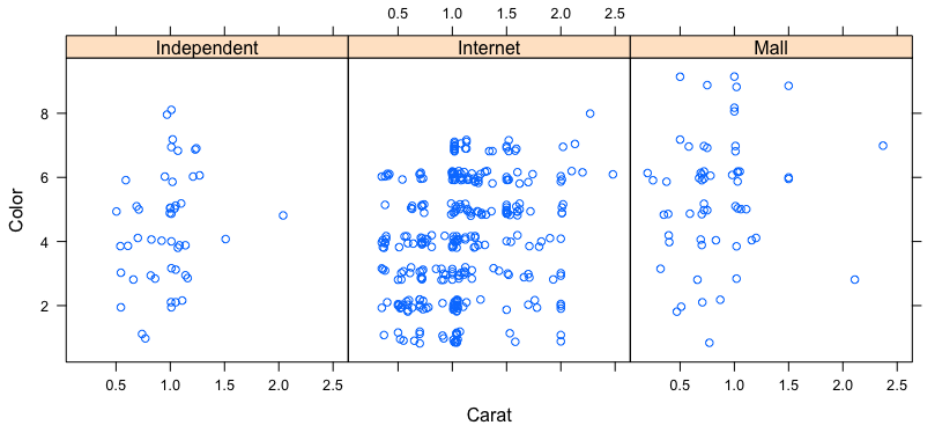
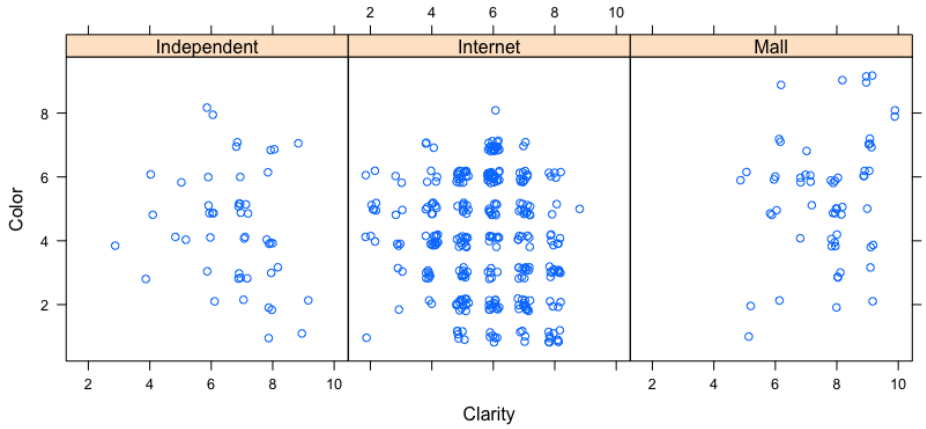


Figure 5: Bar plots for cut type broken down by various factors




Independent
Internet
Mall
Clarity
Carat
0.5
1.0
1.5
2.0
2.5
2
4
6
8
10


Independent
Internet
Mall
Color
Carat
0.5
1.0
1.5
2.0
2.5
2
4
6
8


Independent
Internet
Mall
Color
Clarity
2
4
6
8
10