Exploratory Data Analysis and Regression Models of the Two Months Salary Dataset

Prepared by Amee Amin

Assignment 4

MSPA Course 410, Summer 2017

Professors Chad R. Bhatti and Tom Miller

Introduction

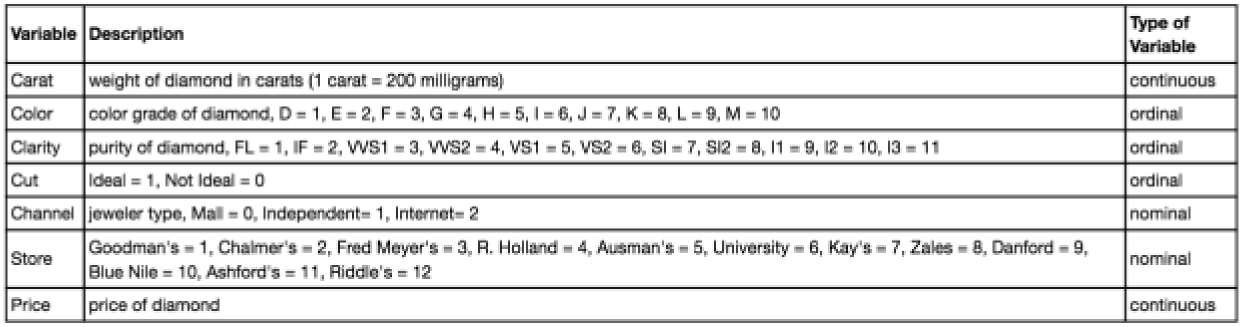
Diamond shopping is a popular consumer tradition in the United States. It is commonly expected to spend around two months salary for a diamond ring. Despite the large price tag, many people go into the ring-shopping process without previous knowledge of or experience purchasing diamonds. In this analysis we explored building a predictive model for the price of diamonds. This analysis uses the ‘two months salary’ dataset created to help students create predictive models of the price of diamonds. Because every diamond is unique, a predictive model could be immensely valuable for consumers looking to purchase and sellers looking to price diamonds. There are 425 observations of 7 variables in the dataset, including two continuous, three ordinal, and two nominal variables. We created simple and multiple linear regression models to determine the best predictors for the price of diamonds. We found that color, channel, and carat have a strong positive linear relationship with price, with carat having the strongest relationship with price, and a box-cox transformation of price improved the regression model.

Section 1: Sample Population

The dataset contains 425 observations of 7 variables: carat, color, clarity, cut, channel, store, and price. These variables are described in Table 1 below. Our predictor variable for this analysis is price. There were no missing or erroneous values in the dataset. Due to the small dataset size, we did not remove any values from this analysis.

The variables included in the dataset were chosen by the author after visiting both brick-and-mortar jewelers and online jewelers. Consumers commonly look at the characteristics included when shopping for a diamond. The dataset includes only round-cut stones, as they were the most common and held the most value, and the author was only interested in purchasing a round-cut diamond. We also assumed that the author only observed rings included in the “engagement” category.

Table 1: Predictor and Response Variables



Variable	Description	Type of Variable
Carat	weight of diamond in carats (1 carat = 200 milligrams)	continuous
Color	color grade of diamond, D = 1, E = 2, F = 3, G = 4, H = 5, I = 6, J = 7, K = 8, L = 9, M = 10	ordinal
Clarity	purity of diamond, FL = 1, IF = 2, VVS1 = 3, VVS2 = 4, VS1 = 5, VS2 = 6, SI = 7, SI2 = 8, I1 = 9, I2 = 10, I3 = 11	ordinal
Cut	Ideal = 1, Not Ideal = 0	ordinal
Channel	jeweler type, Mall = 0, Independent= 1, Internet= 2	nominal
Store	Goodman's = 1, Chalmer's = 2, Fred Meyer's = 3, R. Holland = 4, Ausman's = 5, University = 6, Kay's = 7, Zales = 8, Danford = 9, Blue Nile = 10, Ashford's = 11, Riddle's = 12	nominal
Price	price of diamond	continuous

Color, clarity, carat, and cut are the most common characteristics when diamond ring shopping. Carat represents the weight of gemstones, where one carat equals 0.200 grams. Diamond sizes typically vary between one-quarter to three carats. The dataset author noted that price increases carat, and more specifically, prices jump correspondingly with one-quarter-carat increases.

Diamond color variations result from the process by which diamonds are naturally formed. The presence of various gases and during the time in which stones undergo immense heat and pressure create diamonds of varying tints. The standard color scale for diamonds goes from D to Z, based on tint or color, where D is considered perfectly colorless. The dataset author noted that the price of a diamond decreases as you move away from a grade D color. The clarity, or purity, of a diamond is also measured through a standard scale, based on imperfections visible to the naked eye. Varying clarity results from the carbon pockets that form imperfections, called inclusions. The scale starts with a flawless diamond (FL), and decreases based on the number of inclusions. Most consumers will not come across an FL diamond while ring-shopping. Because we are only looking at round-cut diamonds, the cut variables in our dataset represents the quality of cut. Because there are many small factors, such as symmetry and depth, that influence cut quality, the dataset author categorized diamonds into two types: Ideal Cut and Not Ideal Cut.

The store brand where the diamond was purchased and method of purchase (online, mall, etc.) were also included in the dataset. From existing knowledge about diamond prices, we can expect that the price of diamonds will generally increase with carat, and decrease with poorer color and clarity.

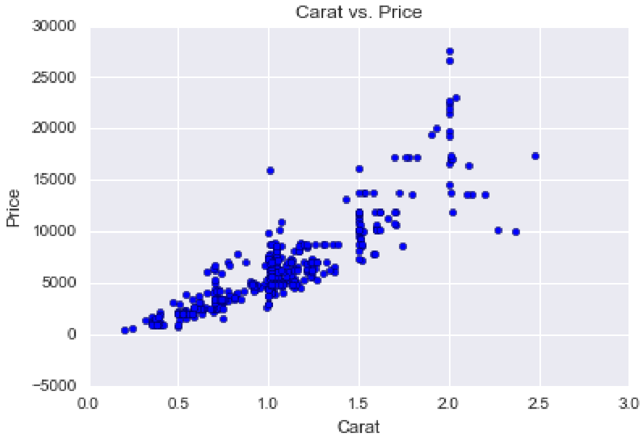
Section 2: Exploratory Data Analysis

We created scatterplots of each of the predictor variables with our response variable price to explore the linearity of their relationships, and then develop a simple linear regression model of carat versus price.

Section 2.1: Scatterplots of the Predictor Variables versus Price

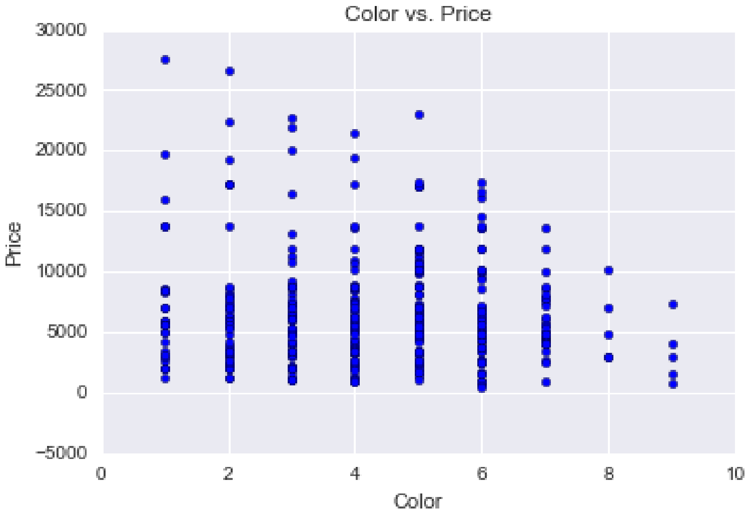
In the scatterplot of carat versus price (below), we examine that price tends to increases with carat. This is as we expected, based on common knowledge about diamond prices.

Fig 1: Scatterplot of Carat vs Price



Most of our observations fall within the carat range of approximately 0.2 to 1.5. After carat increases above 2, there is some non-linear variation in price. This may be due to other characteristics, such as poor clarity or color, that reduce the price of large diamond.

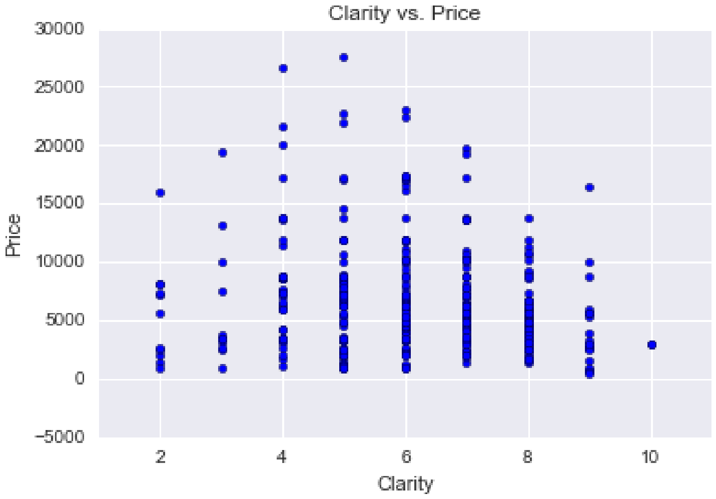
Fig 2: Scatterplot of Color vs Price



As color of a diamond moves away from colorless (above), the price tends to decrease, as seen in Figure 2 above. There are few diamonds with a color grade of D. Most of the diamonds in our dataset are below a price of $150000. There appears to be a similar price range of diamonds below $150000, across all color grades. Because of this pattern in price, we may infer that other characteristics have a larger influence on price.

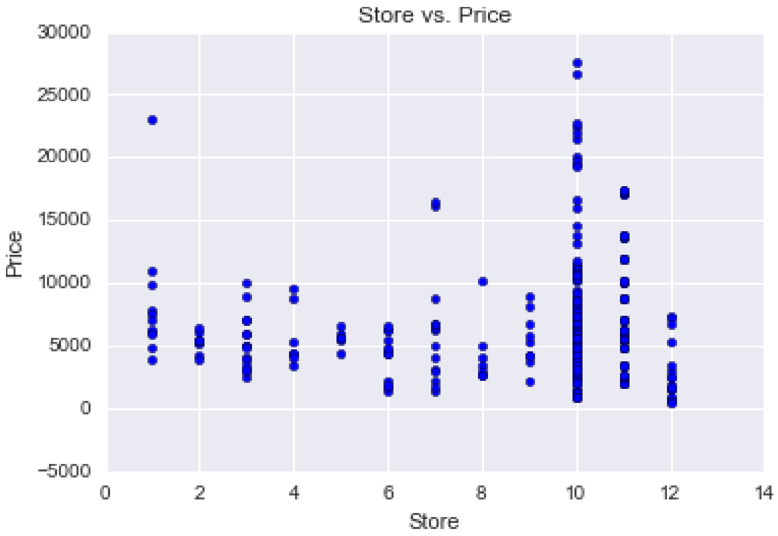
The characteristic of clarity (below) also exhibits a similar pattern with price: as the clarity moves away from flawless, the price tends to decrease, but the price of most diamonds is within a range of $0 and $100000 across all clarity grades. Most diamonds in our dataset have inclusions that are not visible to the naked eye. The increase in price from a clarity of 0 to 4, or VVS2, could be due to other characteristics such as poor color or cut quality.

Fig 3: Scatterplot of Clarity vs Price



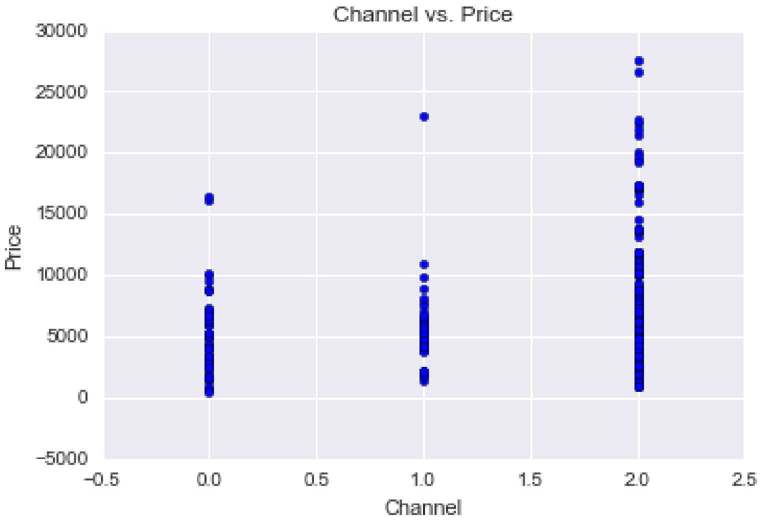
Looking at a scatterplot of store versus price, we observed that the most expensive diamonds were purchased at Blue Nile, a popular online shopping site for diamond rings. Ashford was also a common store for more expensive diamonds (above $100000). Excluding these two stores, most stores sold a similar price of diamonds, which would make it difficult to discriminate price based on store.

Fig 4: Scatterplot of Store vs Price



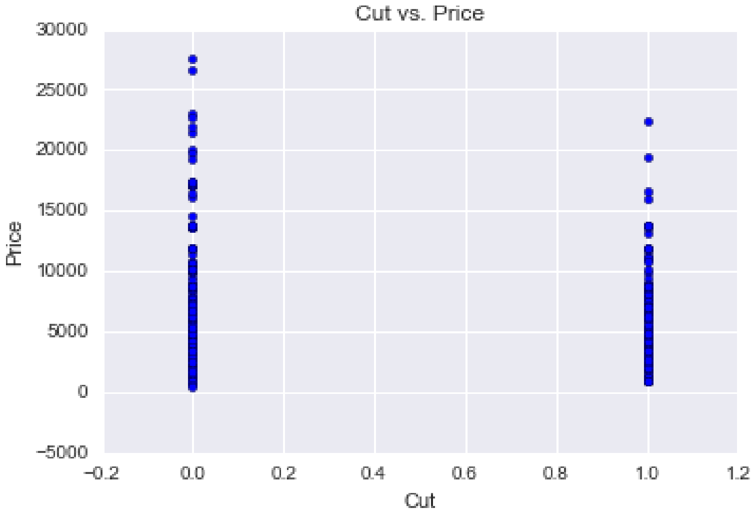
A scatterplot of channel versus price (below) shows that more expensive diamonds, above $100000, are more commonly purchased online, while diamonds below $100000 were bought at malls, independent stores, and online. Although channel may not be a good predictor for exact prices, it may indicate whether or not a diamond was in the $100000+ range.

Fig 5: Scatterplot of Channel vs Price



Last, there does not seem to be a clear pattern of diamond cut quality and price. While there are a more higher priced diamonds that are Ideal Cut, the patterns in price for Ideal and Non-Ideal cut diamonds are similarly below $150000.

Fig 6: Scatterplot of Cut vs Price

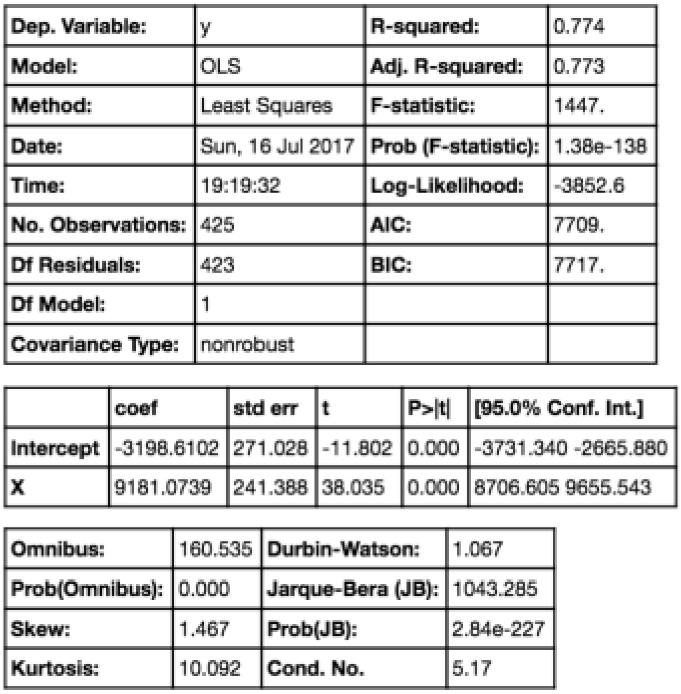


Overall, carat may be the best predictor for price, with clarity and channel type providing additional distinguishing criteria for price.

Section 2.2: Simple Linear Regression

We defined a simple linear regression model between carat, a continuous variable, and price to further explore their relationship. The OLS regression output for carat versus price is shown in the table below.

Table 2: OLS Regression Results for Carat vs Price



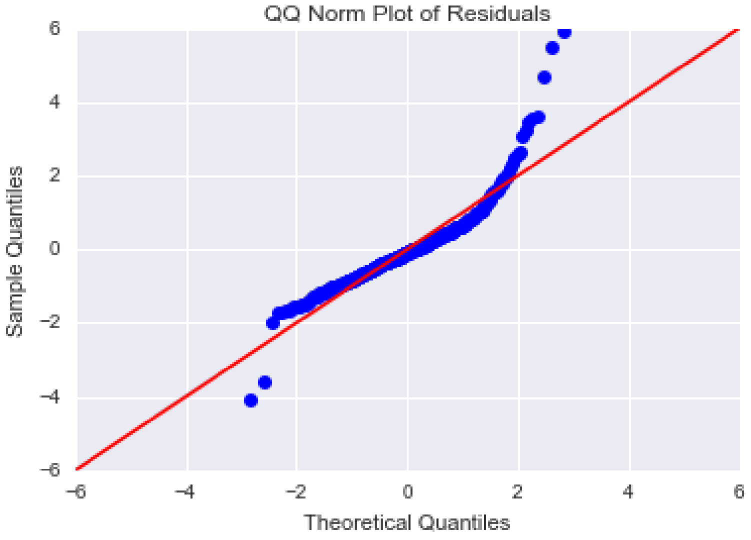
Dep. Variable:	y	**R-squared:**	0.774
Model:	OLS	**Adj. R-squared:**	0.773
Method:	Least Squares	**F-statistic:**	1447.
Date:	Sun, 16 Jul 2017	**Prob (F-statistic):**	1.38e-138
Time:	19:19:32	**Log-Likelihood:**	-3852.6
No. Observations:	425	**AIC:**	7709.
Df Residuals:	423	**BIC:**	7717.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	-3198.6102	271.028	-11.802	0.000	-3731.340 -2665.880
X	9181.0739	241.388	38.035	0.000	8706.605 9655.543

Omnibus:	160.535	**Durbin-Watson:**	1.067
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1043.285
Skew:	1.467	**Prob(JB):**	2.84e-227
Kurtosis:	10.092	**Cond. No.**	5.17

The linear model for carat versus price has a correlation coefficient of 0.774, which indicates there is a strong, positive linear relationship. The F-statistic is significant, so we can reject the null hypothesis that the regression coefficients are zero. The t-value of the coefficient for carat is also significant. However, we observed a high kurtosis and positive skew, which suggests that the distribution is not normal. We discuss how to improve the normality of our model in Section 3.

Figure 7: QQ Norm Plot of Linear Model Residuals



The QQ plot (above) shows that the residuals do not closely resemble a normal distribution. The tails especially skewed. While there does appear to be a linear relationship between carat and price, the linear model above reflects a non-normal distribution and the residuals should be normalized.

Section 3: Multiple Linear Regression Model Specification

We selected three predictor variables based on our exploratory data analysis: carat, color and channel. We also determined that transformations were needed to our response variable.

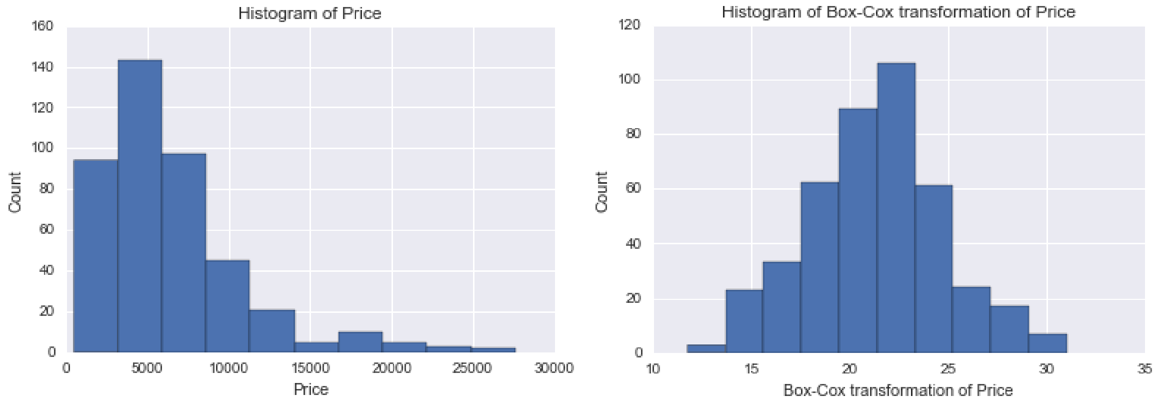
Section 3.1: Model #1 Carat, Color, Channel vs Price

Model #1 will include carat, color, and channel as the predictor variables, and price as the response variable. We created dummy variables for both color and channel, because they are both categorical variables. We selected base categories based on the most frequently observed type of color and channel: a color of 6, meaning type I, and channel 2, the Internet. These categories were not included in the model, in order to avoid collinearity effects. We did not transform the response variable, price.

Section 3.2: Model #2 Carat, Color, Channel vs Box Transformation of Price

Model #2 will include the same predictor variables as Model #1, and will apply the same dummy variable method. However, we applied a Box-Cox transformation to the response variable in Model #2.

Fig 8: Histogram Comparison of Price and Box-Cox transformation of Price



In Figure 7 (above) we examine the effect of a Box-Cox transformation. The histogram of price (left) is skewed to the right, with a long tail, and also shows a high degree of kurtosis. Box-Cox transformations help reduce skewness and kurtosis exhibited by a response variable. The histogram of Box-Cox transformed price (right) shows a much more normal distribution, with less skewness and kurtosis.

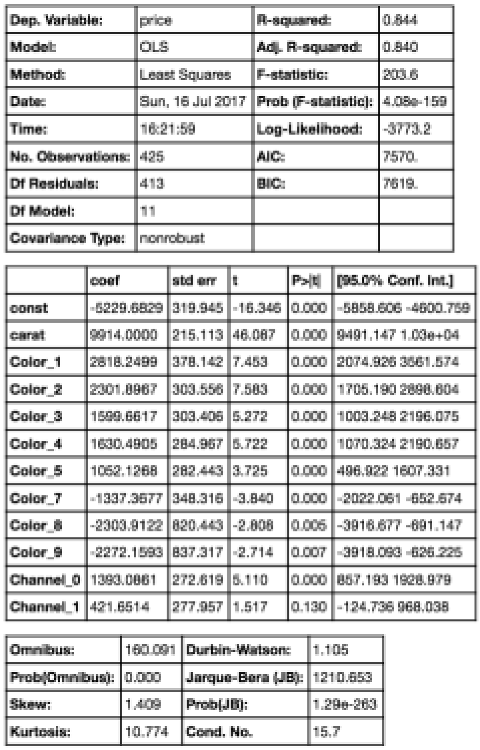
Section 4: Multiple Linear Regression Model Fitting

In this section we examine the results of Model #1 and Model #2.

Section 4.1: Model #1 Carat, Color, Channel vs Price

Looking at the regression results below, Model #1 has a correlation coefficient value of 0.844 and an adjusted correlation coefficient of 0.840. The F-statistic is significant, so we can reject the null hypothesis that the regression coefficients are zero. The t-values for all of the coefficients, excluding Channel_1 (independent), are also significant. The channel type of independent is not significantly different from our baseline. We observed that the coefficients for Color type decrease and become negative as the diamond moves away from D, which was expected based on our earlier analysis. Both carat and channel have positive coefficients.

Table 3: OLS Regression Results for Carat, Color, and Channel vs Price



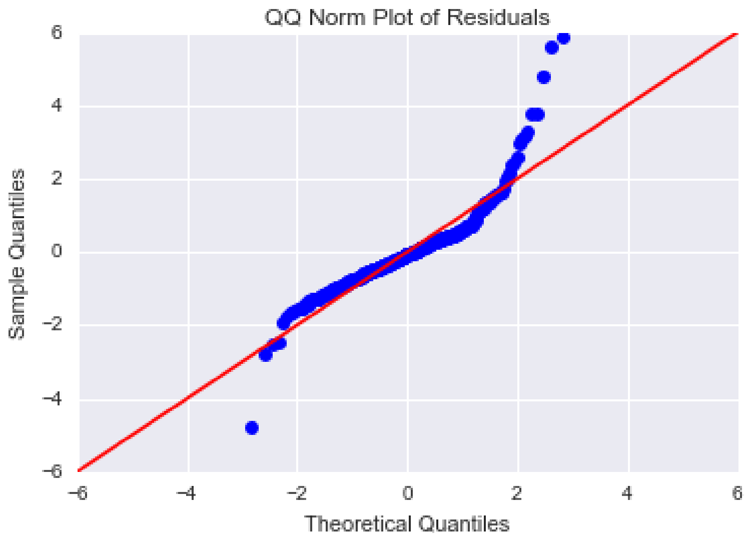
Dep. Variable:	price	**R-squared:**	0.844
Model:	OLS	**Adj. R-squared:**	0.840
Method:	Least Squares	**F-statistic:**	203.6
Date:	Sun, 16 Jul 2017	**Prob (F-statistic):**	4.08e-159
Time:	16:21:59	**Log-Likelihood:**	-3773.2
No. Observations:	425	**AIC:**	7570.
Df Residuals:	413	**BIC:**	7619.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	-5229.6829	319.945	-16.346	0.000	-5858.606 -4600.759
carat	9914.0000	215.113	46.087	0.000	9491.147 1.03e+04
Color_1	2818.2499	378.142	7.453	0.000	2074.926 3561.574
Color_2	2301.8967	303.556	7.583	0.000	1705.190 2898.604
Color_3	1599.6617	303.406	5.272	0.000	1003.248 2196.075
Color_4	1630.4905	284.967	5.722	0.000	1070.324 2190.657
Color_5	1052.1268	282.443	3.725	0.000	496.922 1607.331
Color_7	-1337.3677	348.316	-3.840	0.000	-2022.061 -652.674
Color_8	-2303.9122	820.443	-2.808	0.005	-3916.677 -691.147
Color_9	-2272.1593	837.317	-2.714	0.007	-3918.093 -626.225
Channel_0	1393.0861	272.619	5.110	0.000	857.193 1928.979
Channel_1	421.6514	277.957	1.517	0.130	-124.736 968.038

Omnibus:	160.091	**Durbin-Watson:**	1.105
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1210.653
Skew:	1.409	**Prob(JB):**	1.29e-263
Kurtosis:	10.774	**Cond. No.**	15.7

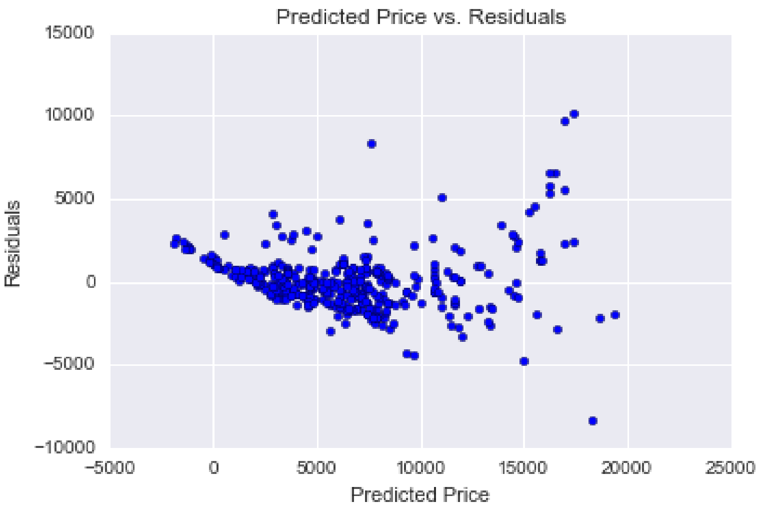
The distribution of Model #1 appears to have a high degree of kurtosis and skewness, as also observed in the linear model of carat vs price defined earlier. This is also confirmed by the QQ Norm Plot of Model #1 residuals (below). The QQ Norm Plot shows that the model is skewed, especially in the tails.

Fig 9: QQ Norm Plot of Residuals



A scatterplot of the fitted values vs residuals from Model #1 shows a few concerning characteristics. The values are clustered on the left, and exhibit heteroscedasticity, meaning residual variance increases as the predicted price increases. The residuals are not symmetrical around 0, and there appear to be more negative residuals than positive. This means that Model #1 over-predicted price more often than it under-predicted.

Fig 10: Scatterplot of Predicted Price vs Residuals

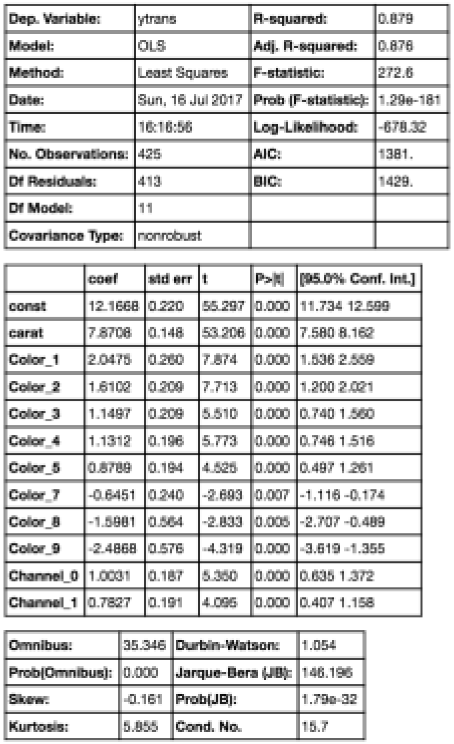


Although we observed a strong linear relationship between carat, color, and channel and price, we should be wary of applying Model #1 because the residuals follow a non-normal pattern and Model #1 does not validate the assumptions of linear regression.

Section 4.2: Model #2 Carat, Color, Channel vs Box-Cox Transformation of Price

The regression results below show that Model #2 has correlation coefficient value of 0.879 and an adjusted correlation coefficient of 0.876. The F-statistic is significant, so we can also reject the null hypothesis that the Model #2 regression coefficients are zero. All of the regression coefficients are significant. Interestingly, the regression coefficient for Channel_1 (independent) is significant in Model #2 when it was not significant in Model #1. However, the coefficient in Model #2 for Channel_1 is 1, therefore it has little influence on price. We observed that the coefficients for color also decrease and become negative as the diamond moves away from D, similar to Model #2. Both carat and channel have positive coefficients in Model #2 as well.

Table 4: OLS Regression Results for Carat, Color, and Channel vs Price



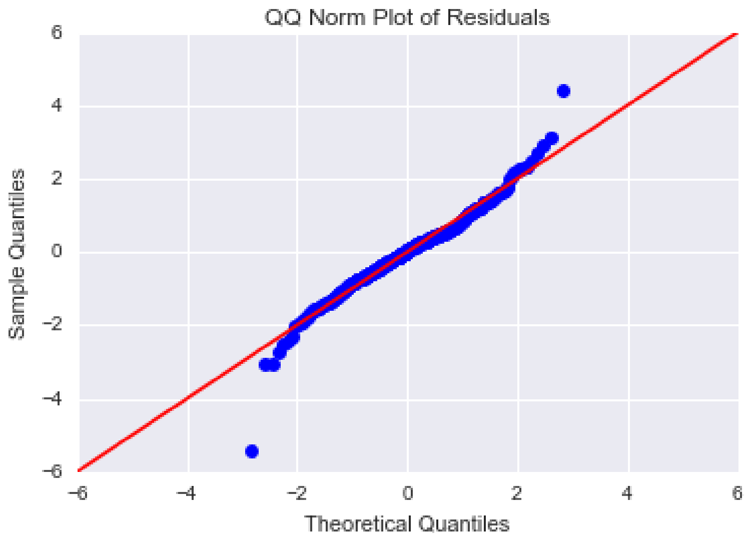
Dep. Variable:	ytrans	R-squared:	0.879
Model:	OLS	Adj. R-squared:	0.876
Method:	Least Squares	F-statistic:	272.6
Date:	Sun, 16 Jul 2017	Prob (F-statistic):	1.29e-181
Time:	16:16:56	Log-Likelihood:	-678.32
No. Observations:	425	AIC:	1381.
Df Residuals:	413	BIC:	1429.
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
const	12.1668	0.220	55.297	0.000	11.734 12.599
carat	7.8708	0.148	53.206	0.000	7.580 8.162
Color_1	2.0475	0.260	7.874	0.000	1.536 2.559
Color_2	1.6102	0.209	7.713	0.000	1.200 2.021
Color_3	1.1497	0.209	5.510	0.000	0.740 1.560
Color_4	1.1312	0.196	5.773	0.000	0.746 1.516
Color_5	0.8789	0.194	4.525	0.000	0.497 1.261
Color_7	-0.6451	0.240	-2.693	0.007	-1.116 -0.174
Color_8	-1.5981	0.564	-2.833	0.005	-2.707 -0.489
Color_9	-2.4868	0.576	-4.319	0.000	-3.619 -1.355
Channel_0	1.0031	0.187	5.350	0.000	0.635 1.372
Channel_1	0.7827	0.191	4.096	0.000	0.407 1.158

Omnibus:	35.346	Durbin-Watson:	1.054
Prob(Omnibus):	0.000	Jarque-Bera (JB):	146.196
Skew:	-0.161	Prob(JB):	1.79e-32
Kurtosis:	5.855	Cond. No.	15.7

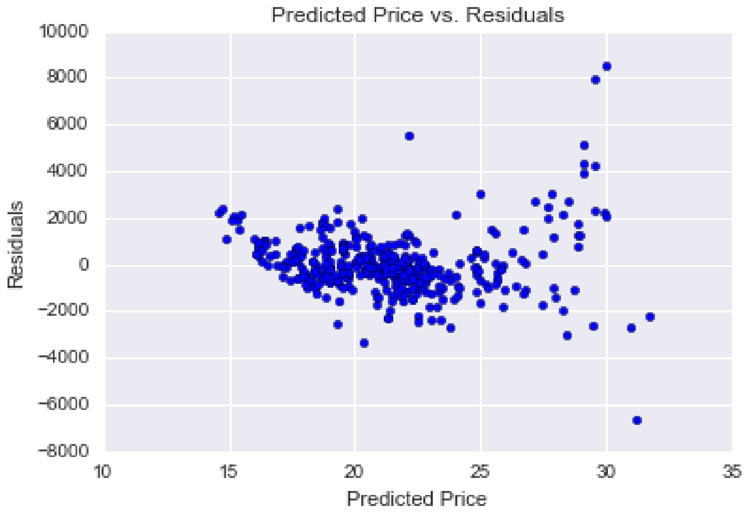
The distribution of Model #2 appears much more normal. Model #2 is slightly negatively skewed, compared to Model #1 which was slightly positively skewed. Overall, there is less absolute skewness in Model #2 than Model #1. The observed kurtosis in Model #2 is nearly half of that of Model #1. Looking at a QQ Norm Plot of Model #2 (below), the residuals more closely follow a normal distribution than Model #1. There is also less skewness in the tails of the distribution.

Fig 11: QQ Norm Plot of Residuals



A scatterplot of the fitted values vs residuals from Model #2 (below), shows more symmetry around 0 and less heteroscedasticity than Model #1. There is still a slight clustering of values on the left side of the plot, and an observable pattern of heteroscedasticity. Model #2 over-predicted values more than under-predicting, similar to Model #1 as well.

Fig 12: Scatterplot of Predicted Price vs Residuals

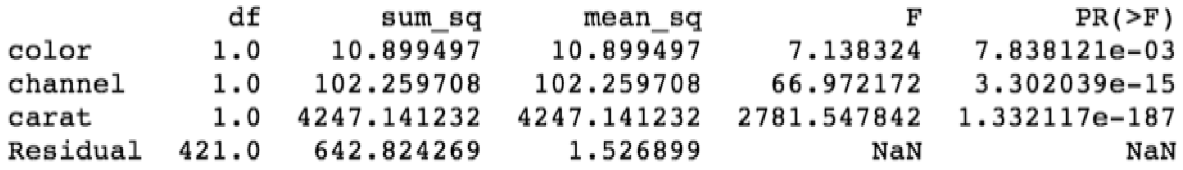


Overall, Model #2 appears to validate the assumption of linear regression better than Model #1.

Section 4.3: ANOVA Test

An analysis of variance table (below) allowed us to examine how each predictor variable accounted for variability in the response variable. The predictor variables of color, channel, and carat all have significant F-statistics. This means that we can reject the null hypothesis for each predictor, and conclude that there is a statistically significant difference among population means for each level within color, channel, and carat.

Table 5: ANOVA Test of Carat, Channel, and Color vs Price



```
             df        sum_sq     mean_sq            F         PR(>F)
color       1.0     10.899497   10.899497     7.138324   7.838121e-03
channel     1.0    102.259708  102.259708    66.972172   3.302039e-15
carat       1.0   4247.141232 4247.141232  2781.547842  1.332117e-187
Residual  421.0    642.824269    1.526899          NaN            NaN
```

Section 5: Model Comparisons and Recommendations

Overall, we would recommend Model #2 over Model #1 to be used in predicting the price of diamonds. There are two main reasons for this. First, Model #2 better validated the assumptions of linear regression. We observed that Model #2's distribution was less skewed and less leptokurtic than Model #1. The QQ norm plot confirmed that Model #2 residuals more closely followed a normal distribution than Model #1, and the scatterplot of Model #2 fitted vales vs residuals showed a less observable pattern of asymmetry and heteroscedasticity.

Second, the correlation coefficient for Model #2 (0.879) was larger than that of Model #1 (0.844). The Box-Cox transformation of price improved its linear relationship with color, channel, and carat. We would recommend that management further refines Model #2 by exploring outliers or influencing observations. The applications for Model #2 could be used by either consumers or sellers. For consumers, Model #2 may be able to give first-time ring buyers an estimate of what they should expect to spend on a diamond, given a certain color, carat, and channel they use. Being able to see an estimate may enable consumers to choose different colors or carats to get the desired diamond within a certain price range. On the seller's side, a company may be able to better price a diamond based on Model #2. If the dataset is collected throughout the industry, company's could compare prices and choose, for example, to sell more diamonds online because of the higher selling price. Model #2 could be refined for varying goals, depending on the ultimate decision that it is informing.

Conclusion

In conclusion, we observed that there is a strong linear relationship between color, carat, and channel and price. This relationship is improved after a Box-Cox transformation of price, due to its non-normal distribution. Carat had the strongest linear relationship with price, as expected based on previous knowledge. Because Model #2 is based on strictly round cut diamonds, further studies should explore the effectiveness of linear regression models in predicting price across different diamond cuts (round, princess, etc.). It would be interesting to explore whether diamond cuts introduce any collinearity to the model. Ultimately, determining the user for Model #2, and the end decision that the model is informing, will be an important step before future modifications.