Introduction

For this assignment, we are working with a data set that captures the price, grading information, and other characteristics of a population of diamonds. The data were collected by Brian A. Pope and reproduced with permission in Miller [1], which is our source. Pope collected data on the carat weight, color, clarity, cut, sales channel and price for 425 stones. Since the precise problem is not given for this assignment, I am exercising some latitude in my definition of the statistical problem.

The average engagement ring in the USA has a diamond coming in at 90-points, or just under 1 carat [2] [3]. I am electing to use this piece of information into my data selection process. Framing the statistics problem as creating models for stones I, personally would consider purchasing, I am investigating all stones between .70 and 2.0 carats. The low-end .70 cut-off allows for stones that approach the 1-carat average, and the cut-off at 2.0 carats reflects the outer edge of “normal” stone sizes. I am also factoring out certain color and clarity ratings based on what I would index on as a shopper.

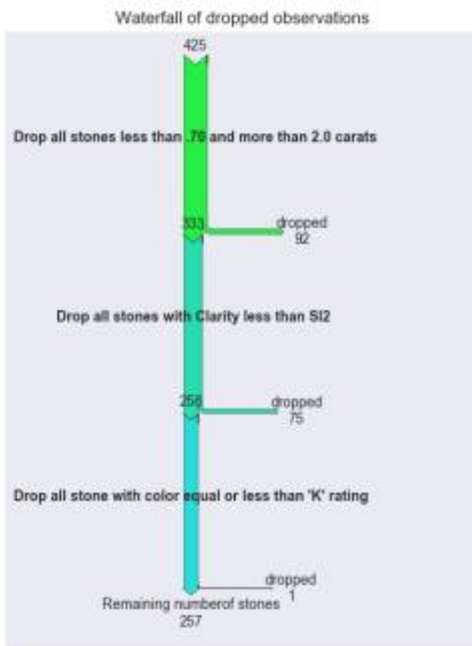
To explore the data, I will look at a range of plots such as distribution, boxplots, and scatterplot matrices to get an idea about the shape of the data, and possible correlations that could be useful predictors. I will then define a simple regression model and a pair of multiple regression models which will be fitted to the data. Results will be evaluated by looking at residuals in various ways, and by looking at the fitted values versus predicted values. Additionally, the use of dummy variables to take the place of “clarity” will be tried as a method to explore what happens when the dummy variable technique is used.

I found that it was possible to generate a linear model that is probably good enough for use, but I was not able to any model that had particularly good residual plots, in the sense that the plots conformed to a normal distribution. I had my best results from a model which used a quadratic term. It may be that this particular problem is best solved via another form of regression, or is best modeled by a formula more complicated than any I tried.

Sample Definition

For my data sample, I am electing to work with diamonds that are between .7 and 2.0 carats (inclusive). I am also restricting my data set to exclude diamonds that have a clarity in the “I” range, or colors K, L or M. I am coming from the point of view that we are evaluating diamonds to give as an engagement ring, and therefore some aspects of quality enter into the decision. The excluded diamonds are visually less appealing to most people, and as such, are less desirable for engagement rings. Figure 1 below shows the waterfall for restricting the data. The final number of observations I will be working with is 257.

FIGURE 1



Exploratory Data Analysis and Simple Linear Regression

First, I wanted to get an idea about the minimums, and maximums for the continuous variable to see if there were any suspect values. Secondly, I wanted to see if there were missing values in my set of working data. The results of this exploration are summarized in Table 1. There are no missing values to content with, and while there is a large range between the minimum and maximum price, there are no “out of bounds” values like a negative number.

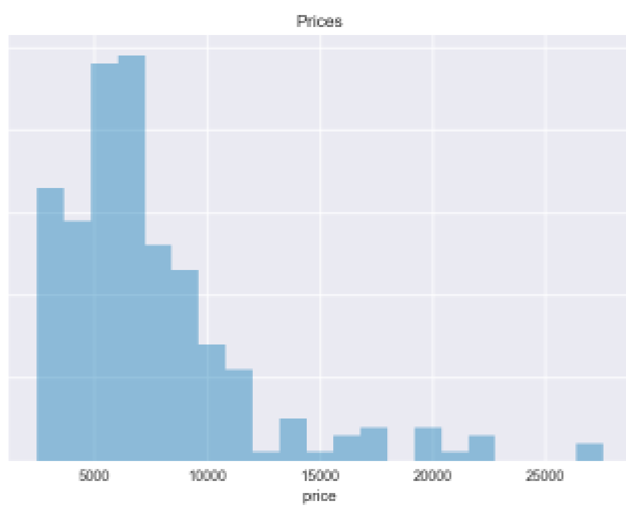
TABLE 1

Price	Carat	Missing Values
count 257.000000	count 257.000000	Data columns (total 7 columns):
mean 7460.867704	mean 1.144031	carat 257 non-null float64
std 4251.113118	std 0.318089	color 257 non-null int64
min 2450.000000	min 0.700000	clarity 257 non-null int64

25% 4850.000000 50% 6404.000000 75% 8676.000000 max 27575.000000 Name: price, dtype: float64	25% 1.000000 50% 1.040000 75% 1.260000 max 2.000000 Name: carat, dtype: float64	cut 257 non-null int64 channel 257 non-null int64 store 257 non-null int64 price 257 non-null int64 dtypes: float64(1), int64(6)

The distribution of the price data shows that it is not normally distributed. There are also gaps in the prices. The skew and kurtosis numbers listed below Figure 2, confirm that the data are positively skewed and leptokurtic.

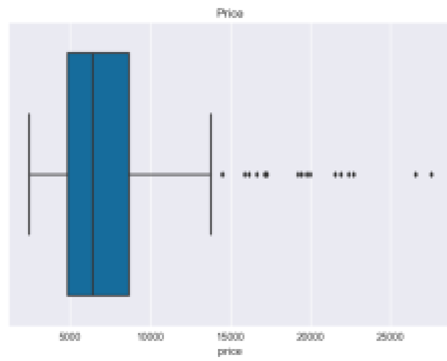
FIGURE 2



Skew = 2.02389156383
Kurtosis = 4.91634520377

Looking at the boxplot of price, we can see that there are a large number of outliers, and several extreme outliers.

FIGURE 3



Finally, looking at the scatterplot matrix for our variables, we can see that carat shows a strong linear relationship with prices, and that the other variables do not provide any clear indication of relationship to price.

FIGURE 4

Since carat is the only other continuous variable, and since it looks like it has a linear relationship with price, we will use it to fit a simple linear regression model. We get the following fitted results:

TABLE 2

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.755
Model:                            OLS   Adj. R-squared:                  0.754
Method:                 Least Squares   F-statistic:                     787.0
Date:                Wed, 12 Jul 2017   Prob (F-statistic):           6.51e-80
Time:                        20:11:42   Log-Likelihood:                -2330.5
No. Observations:                 257   AIC:                             4665.
Df Residuals:                     255   BIC:                             4672.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -5826.7447    491.545    -11.854      0.000   -6794.750   -4858.739
carat       1.161e+04    414.016     28.054      0.000    1.08e+04    1.24e+04
==============================================================================
Omnibus:                       79.521   Durbin-Watson:                   1.020
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              272.994
Skew:                           1.287   Prob(JB):                     5.25e-60
Kurtosis:                       7.344   Cond. No.                         7.46
==============================================================================
```

A plot of the fitted regression line against the data:

FIGURE 5

The QQ plot of residuals (Figure 6) and the residuals plotted against predicted (Figure 7) provide a view into the goodness-of-fit for the model. The QQ plot shows that the residuals are not normally distributed near the ends of the range. The residual plot is clustered to the left side of the chart, and does not have the well-distributed, randomness we are looking for.

FIGURE 6

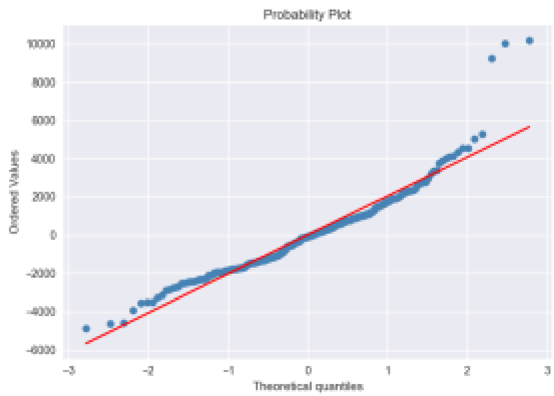
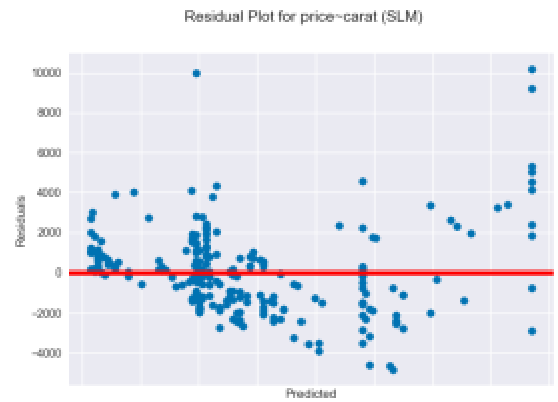


FIGURE 7



The simple linear regression model fails the assumptions of normally distributes residuals and equal variances.

Multiple Linear Regression Model Specification

Model MR1 – Response Variable: Price, Predictor Variables: Carat, Color, Clarity, Cut and Channel

I explored several different combinations of variables to create the first model. I began by adding each predictor in and assessing the resultant model. I looked at the fit and at the goodness-of-fit for each model. Details of this model fits as part of this exploration are shown in Appendix A. I also tried breaking up clarity into a set of 4 dummy variables, grouping FL and IF into a bin, the VVS1 and VVS2 into a bin, V1 and V2 into the third bin, and SI 1 and SI2 into the final bin. I used these in place of the "clarity" predictor. It didn't have much impact on the model, so I didn't use it in the final model. The residual plots using the dummy variables were about the same as without the dummy variables, and the adjusted R^2 was not as good as the good I used instead. Details are provided in the appendix.

Eventually, I landed on using "carat", "color", "cut", "clarity" and "channel" as the predictor variables. I selected this combination since it provided the second best adjusted R^2. Adding "store" did raise the adjusted R^2, but only by .001 (shown in Appendix A). So, I opted for a slightly simpler model over a slightly higher adjusted R^2.

Model MR2 – Response Variable: log(Price), Predictor Variables: Carat, Color, Clarity, Cut and Channel

We know from the simple linear regression model that the data violate the assumption of normally distributed residuals and equal variances. To address this, I transformed the price data by taking the log and used the new log(price) as the response variable in the second model. This produced an improvement to the model, which is discussed in the section below.

Multiple Linear Regression Model Fitting

Model MR1 –Price~ Carat + Color + Clarity+ Cut + Channel

Fitting the first model gives:

TABLE 3

OLS Regression Results

Dep. Variable:	price	R-squared:	0.906
Model:	OLS	Adj. R-squared:	0.904
Method:	Least Squares	F-statistic:	483.0
Date:	Thu, 13 Jul 2017	Prob (F-statistic):	1.50e-126
Time:	12:15:25	Log-Likelihood:	-2207.7
No. Observations:	257	AIC:	4427.
Df Residuals:	251	BIC:	4449.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2695.1976	623.591	4.322	0.000	1467.060	3923.335
carat	1.227e+04	265.705	46.192	0.000	1.18e+04	1.28e+04
color	-802.6394	47.264	-16.982	0.000	-895.723	-709.556
clarity	-589.5422	68.522	-8.604	0.000	-724.493	-454.591
cut	556.3022	174.583	3.186	0.002	212.467	900.137
channel	-1446.4194	157.356	-9.192	0.000	-1756.326	-1136.513

Omnibus:	97.185	Durbin-Watson:	1.210
Prob(Omnibus):	0.000	Jarque-Bera (JB):	431.742
Skew:	1.495	Prob(JB):	1.77e-94
Kurtosis:	8.602	Cond. No.	60.0

Checking for goodness-of-fit, we get:

FIGURE 8

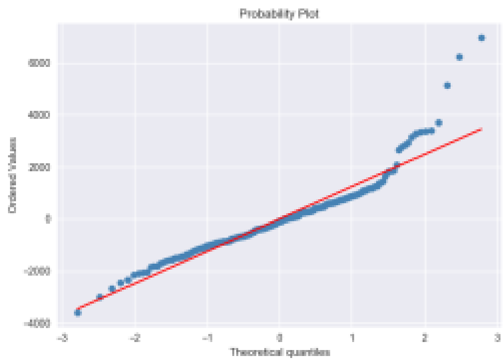
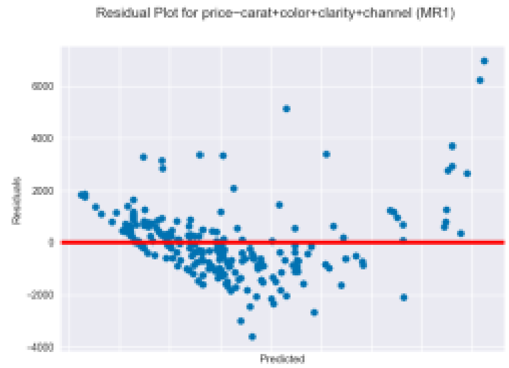


FIGURE 9



The QQ plot of residuals shows that variances are not equal, particularly for the larger theoretical quartiles. The residual scatterplot shows the multiple regression residuals do not have a normal distribution. The distribution is clustered more toward the left. I am having trouble deciphering the shape of the residuals. The plot has a shape that looks vaguely concave. The hint of curvature may indicate that we do not have an actual linear regression, but perhaps a polynomial instead. Alternately, the shape could be an extreme display of heteroscedasticity. The adjusted R^2 of .904 for model MR1 indicates the model has captured most of the variation in the data set.

Model MR2 – log(Price)~ Carat + Color + Clarity+ Cut + Channel

Fitting MR2 yields:

TABLE 4

```
  ....
                            OLS Regression Results
==============================================================================
Dep. Variable:              log_price   R-squared:                       0.922
Model:                            OLS   Adj. R-squared:                  0.921
Method:                 Least Squares   F-statistic:                     597.1
Date:                Thu, 13 Jul 2017   Prob (F-statistic):          4.19e-137
Time:                        12:13:41   Log-Likelihood:                 152.55
No. Observations:                 257   AIC:                            -293.1
Df Residuals:                     251   BIC:                            -271.8
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      8.2538      0.064    128.920      0.000       8.128       8.380
carat          1.4180      0.027     51.980      0.000       1.364       1.472
color         -0.0827      0.005    -17.045      0.000      -0.092      -0.073
clarity       -0.0681      0.007     -9.686      0.000      -0.082      -0.054
cut            0.0956      0.018      5.332      0.000       0.060       0.131
channel       -0.2050      0.016    -12.690      0.000      -0.237      -0.173
==============================================================================
Omnibus:                       27.400   Durbin-Watson:                   0.745
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               48.456
Skew:                           0.599   Prob(JB):                     3.00e-11
Kurtosis:                       4.758   Cond. No.                         60.0
==============================================================================
```

Checking the residuals, we see:

FIGURE 10

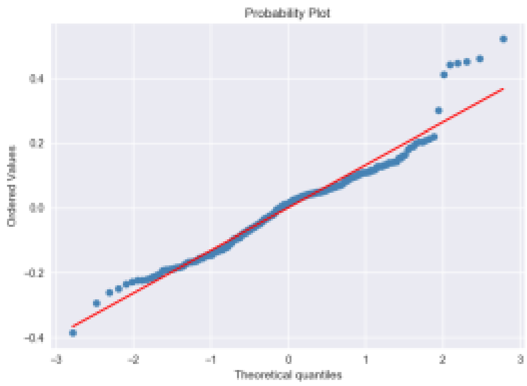
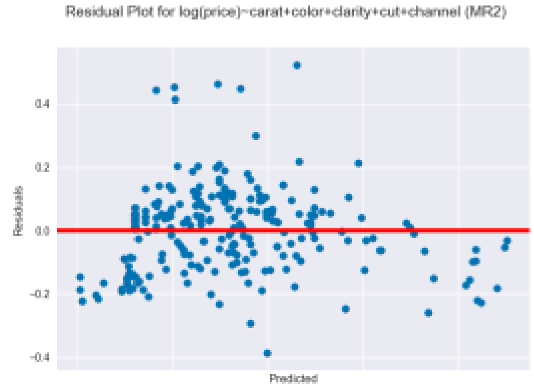


FIGURE 11



The QQ plot reveals we are doing better at having equal variance, but there is still a sharp deviation from the normal on the far-right side of the chart. The residuals versus predicted is more centered, but now has a vaguely convex shape. The points below 0 seem well distributed, the ones above 0 are still somewhat clustered to the left, and close to 0.

Model Comparisons and Recommendation

None of the models I fit above are particularly satisfying. The simple linear regression model only accounts for ~76% of the variation in price. The multiple regression models do better at capturing the price variation, but the residual plots show that both violate the assumptions for linear regression. Of the two multiple regression models, I'd tend to use MR2 over MR1 since the QQ plot conforms to the normal line for a greater distance and the adjusted R^2 is higher.

Since neither model is particularly compelling, I investigated a third model, this time adding in a quadratic term: carat**2. This third model, based on MR1, which I've named MR1_2 for this exercise, is intended to see if a quadratic shows a beneficial effect on the formula. Fitting this third model produced:

TABLE 5

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.925
Model:                            OLS   Adj. R-squared:                  0.924
Method:                 Least Squares   F-statistic:                     620.1
Date:                Thu, 13 Jul 2017   Prob (F-statistic):          5.31e-139
Time:                        13:58:35   Log-Likelihood:                -2178.4
No. Observations:                 257   AIC:                             4369.
Df Residuals:                     251   BIC:                             4390.
Df Model:                           5
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept               9250.7064    520.076     17.787      0.000    8226.437    1.03e+04
np.power(carat, 2)      4733.9690     90.332     52.406      0.000    4556.064    4911.874
np.power(color, 1)      -700.4388     42.049    -16.658      0.000    -783.252    -617.625
np.power(clarity, 1)    -547.1808     61.135     -8.950      0.000    -667.584    -426.778
np.power(cut, 1)         617.5379    155.818      3.963      0.000     310.661     924.415
channel                -1392.7327    140.083     -9.942      0.000   -1668.620   -1116.845
==============================================================================
Omnibus:                       88.594   Durbin-Watson:                   1.450
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              435.307
Skew:                           1.305   Prob(JB):                     2.98e-95
Kurtosis:                       8.818   Cond. No.                         55.9
==============================================================================
```

FIGURE 12

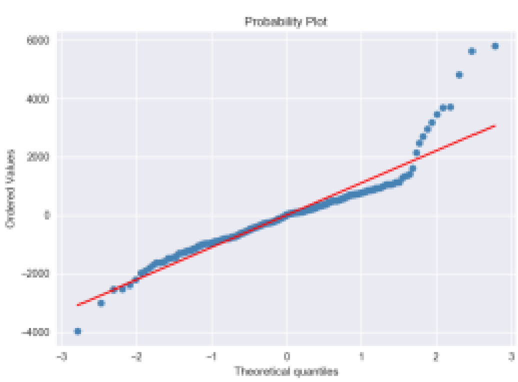
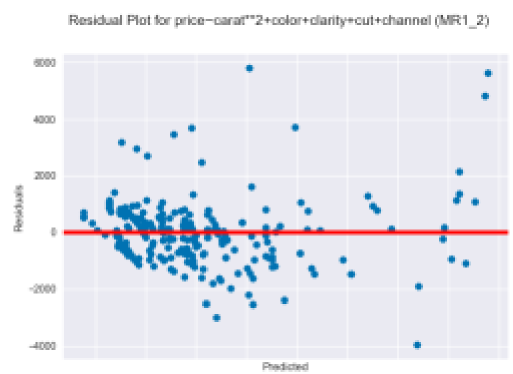


FIGURE 13



While still not perfectly normal, the residuals versus price look better distributed in Figure 13 than they do in Figure 11. If a linear equation must be used, I would use MR1_2 as my model. I believe the departures from normal are not likely to prove significant for this statistical problem.

Time and resources permitting, I suggest that more exploration into adding powers of the variables could generate additional improvements to the model. I tried a few additional models like this, one of which is shown in the appendix.

Conclusion

It is not covered in Pope's article, but the cost per carat for diamonds is typically not linear [4], [2] but more of a step function around inflection points like 1 carat and 2 carats. I was curious to see how well we could do with linear regression. Given the results from the various model fitting exercises, it appears that we can come close with linear models, but that none of them are quite right. Of the models I fit, think MR1_2 has the most promise, and if I had to use one of these, that is what I would pick. Model MR1_2, with its quadratic term, has a good adjusted R^2 value and decent residual plots. You'd expect a high a high R^2 given the use of nearly all the variables, but the adjust R^2 is also the highest of the models I created.

Appendix A

Explorations of model predictors

Price ~ Carat + Color

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.855
Model:                            OLS   Adj. R-squared:                  0.854
Method:                 Least Squares   F-statistic:                     747.9
Date:                Thu, 13 Jul 2017   Prob (F-statistic):          3.57e-107
Time:                        11:27:28   Log-Likelihood:                -2263.4
No. Observations:                 257   AIC:                             4533.
Df Residuals:                     254   BIC:                             4543.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -2686.5249    447.753     -6.000      0.000   -3568.305   -1804.745
carat       1.185e+04    319.978     37.029      0.000    1.12e+04    1.25e+04
color       -762.8391     57.795    -13.199      0.000    -876.657    -649.022
==============================================================================
Omnibus:                       97.341   Durbin-Watson:                   1.169
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              324.085
Skew:                           1.627   Prob(JB):                     4.22e-71
Kurtosis:                       7.436   Cond. No.                         25.9
==============================================================================
```

Price ~ Carat + Color + Clarity

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.873
Model:                            OLS   Adj. R-squared:                  0.872
Method:                 Least Squares   F-statistic:                     580.5
Date:                Thu, 13 Jul 2017   Prob (F-statistic):          4.41e-113
Time:                        11:27:29   Log-Likelihood:                -2246.1
No. Observations:                 257   AIC:                             4500.
Df Residuals:                     253   BIC:                             4514.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     26.2254    614.446      0.043      0.966   -1183.855    1236.305
carat       1.175e+04    300.175     39.140      0.000    1.12e+04    1.23e+04
color       -752.1066     54.165    -13.886      0.000    -858.778    -645.435
clarity     -468.7664     77.596     -6.041      0.000    -621.582    -315.951
==============================================================================
Omnibus:                      112.332   Durbin-Watson:                   1.010
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              399.685
Skew:                           1.887   Prob(JB):                     1.62e-87
Kurtosis:                       7.805   Cond. No.                         50.9
==============================================================================
```

Price ~ Carat + Color + Clarity + Cut

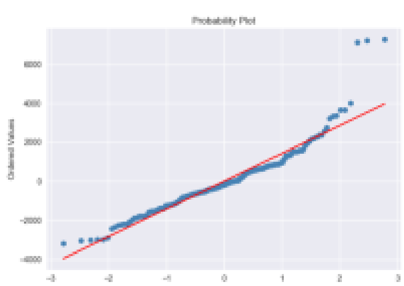
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.874
Model:                            OLS   Adj. R-squared:                  0.872
Method:                 Least Squares   F-statistic:                     437.7
Date:                Thu, 13 Jul 2017   Prob (F-statistic):          4.16e-112
Time:                        11:27:29   Log-Likelihood:                -2245.0
No. Observations:                 257   AIC:                             4500.
Df Residuals:                     252   BIC:                             4518.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -146.3626    624.883     -0.234      0.815   -1377.022    1084.297
carat       1.178e+04    300.148     39.233      0.000    1.12e+04    1.24e+04
color       -748.5384     54.110    -13.834      0.000    -855.103    -641.974
clarity     -464.5549     77.490     -5.995      0.000    -617.166    -311.944
cut          284.5258    198.530      1.433      0.153    -106.463     675.515
==============================================================================
Omnibus:                      114.302   Durbin-Watson:                   0.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              418.367
Skew:                           1.911   Prob(JB):                     1.42e-91
Kurtosis:                       7.947   Cond. No.                         51.9
==============================================================================
```

Price ~ Carat + Color + Clarity + Channel + Store

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.907
Model:                            OLS   Adj. R-squared:                  0.905
Method:                 Least Squares   F-statistic:                     406.2
Date:                Thu, 13 Jul 2017   Prob (F-statistic):          7.99e-126
Time:                        11:27:29   Log-Likelihood:                -2206.2
No. Observations:                 257   AIC:                             4426.
Df Residuals:                     250   BIC:                             4451.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   3098.2282    663.895      4.667      0.000    1790.687    4405.769
carat        1.23e+04    265.040     46.398      0.000    1.18e+04    1.28e+04
color       -806.3855     47.131    -17.110      0.000    -899.209    -713.562
clarity     -599.1532     68.484     -8.749      0.000    -734.033    -464.273
cut          546.8843    173.992      3.143      0.002     204.207     889.562
channel    -1169.0192    224.897     -5.198      0.000   -1611.953    -726.085
store        -89.7468     52.177     -1.720      0.087    -192.509      13.015
==============================================================================
Omnibus:                       87.925   Durbin-Watson:                   1.234
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              396.065
Skew:                           1.326   Prob(JB):                     9.90e-87
Kurtosis:                       8.473   Cond. No.                         101.
==============================================================================
```

Model MR1_F showing effects of using dummy variables for Clarity

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.878
Model:                            OLS   Adj. R-squared:                  0.876
Method:                 Least Squares   F-statistic:                     453.8
Date:                Sun, 16 Jul 2017   Prob (F-statistic):          7.69e-114
Time:                        14:10:40   Log-Likelihood:                -2241.0
No. Observations:                 257   AIC:                             4492.
Df Residuals:                     252   BIC:                             4510.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -569.2925    247.717     -2.298      0.022   -1057.153     -81.432
carat       1.231e+04    301.725     40.787      0.000    1.17e+04    1.29e+04
color       -804.8482     53.676    -14.995      0.000    -910.558    -699.138
F           5.588e-13   5.34e-14     10.472      0.000    4.54e-13    6.64e-13
VV          3.322e-13    4.8e-14      6.924      0.000    2.38e-13    4.27e-13
V           -569.2925    247.717     -2.298      0.022   -1057.153     -81.432
S                   0          0        nan        nan           0           0
cut          560.8456    198.271      2.829      0.005     170.367     951.324
channel    -1177.7646    175.153     -6.724      0.000   -1522.715    -832.815
==============================================================================
Omnibus:                       92.992   Durbin-Watson:                   1.339
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              399.144
Skew:                           1.434   Prob(JB):                     2.12e-87
Kurtosis:                       8.389   Cond. No.                     1.37e+34
==============================================================================
```



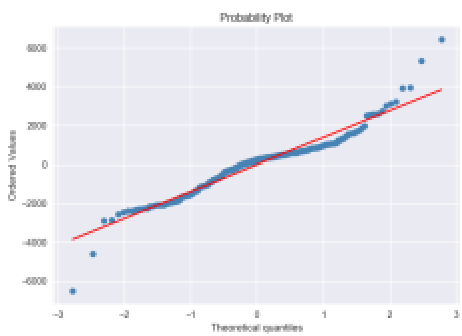
Model MR2_2 showing effects of adding quadratic terms

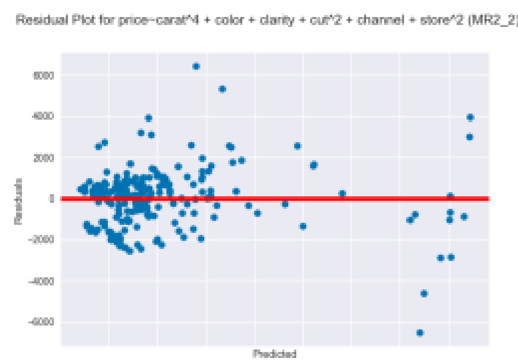
OLS Regression Results

Dep. Variable:	price	R-squared:	0.888
Model:	OLS	Adj. R-squared:	0.885
Method:	Least Squares	F-statistic:	330.9
Date:	Sat, 15 Jul 2017	Prob (F-statistic):	7.43e-116
Time:	19:56:11	Log-Likelihood:	-2229.9
No. Observations:	257	AIC:	4474.
Df Residuals:	250	BIC:	4499.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.149e+04	639.896	17.953	0.000	1.02e+04	1.27e+04
np.power(carat, 4)	1091.4507	26.117	41.791	0.000	1040.014	1142.887
np.power(color, 1)	-516.7658	51.543	-10.026	0.000	-618.279	-415.253
np.power(clarity, 1)	-496.6554	75.067	-6.616	0.000	-644.500	-348.810
np.power(cut, 2)	690.2544	191.072	3.613	0.000	313.938	1066.571
np.power(channel, 1)	-1053.6464	253.341	-4.159	0.000	-1552.602	-554.691
np.power(store, 2)	-2.3128	4.353	-0.531	0.596	-10.887	6.261

Omnibus:	27.189	Durbin-Watson:	1.199
Prob(Omnibus):	0.000	Jarque-Bera (JB):	131.630
Skew:	0.093	Prob(JB):	2.61e-29
Kurtosis:	6.501	Cond. No.	723.





Works Cited

[1] T. W. Miller, "Marketing Data Science Modeling Techniques in Predictive Analytics with R and Python," pp. 380 - 383, 2015.

[2] Beyond4Cs, "Average Engagement Ring Size," [Online]. Available: http://beyond4cs.com/carat/average-engagement-ring-size/.

[3] Your Diamond Teacher, "What is the Average Diamond Size for an Engagement Ring?," 2016. [Online]. Available: http://yourdiamondteacher.com/diamond-4cs/carat-weight/average-engagement-ring-size/.

[4] A. L. Matlins and A. C. Bonanno, Jewelry & Gems, The Buying Guide, Woodstock, VT: Gemstone Press, pp. 56 - 59.