Nicholas Keeley (ngk3pk)
Geoff Hansen (gmh9he)
Clair McLafferty (cm2rh)
John Zhang (jz5jx)

Executive Summary

Less than a century ago, diamond engagement rings were the exception, not the norm. But thanks to a hundred-year marketing campaign started by DeBeers, the association between love/commitment and diamonds endures. Over the past decade or so, trends seem to be moving away from giant stones, at least according to popular articles on the topic.

Despite these stories, demand within the U.S. diamond market is on the rise, and companies are poised to market accordingly. For Blue Nile, the world's largest online diamond market, this means a heavy emphasis on the low price of their stock, with a marked focus on cut.

When presented with the dataset representing Blue Nile's full stock list, we became curious about the factors that drove price. Common knowledge on the subject would indicate that a diamond's size is the biggest indicator of price, even though Blue Nile's marketing touts cut as the definitive factor.

With this in mind, we hypothesized that the price is driven primarily by the stone's size, with cut, clarity, and color as subsidiary factors. Additionally, we hypothesized that the diamond market operated in tiers based on customer preferences more than likely based on socioeconomic status and their intended use for the stone.

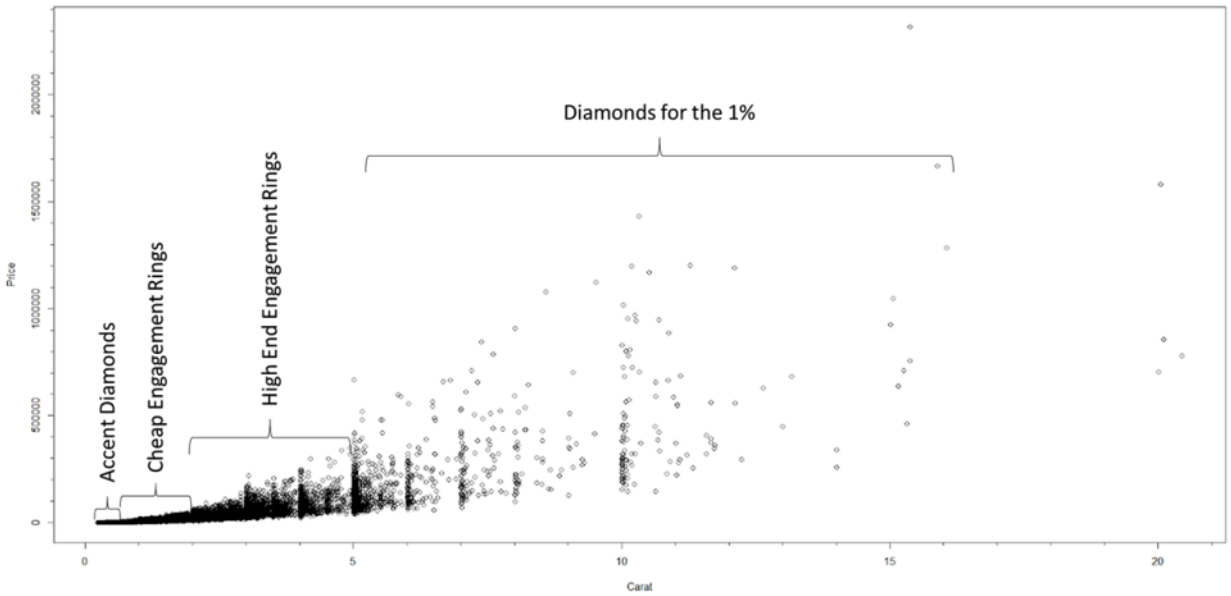
For example, individuals looking to buy a diamond of more than five carats likely care about every aspect of the stone, while those who are buying small diamonds are likely less interested in their stones being as perfect as possible. To address our hypothesis, we decided to fit and analyze a model to the full data set and build models for each predicted market to see which best represented the full set.

However, we must first note the potential bias in the data. Since this dataset represents a specific company's offerings, their interest is first and foremost to ensure their profitability. As a result, they have an incentive to appraise their wares to represent that they are as high quality as possible. Although industry standards for these measures are in place, they are necessarily measured subjectively.

To study how each factor within the dataset (size, cut, clarity, and color) affected the price, we first took a look at the dataset holistically. When we fitted a linear model that took all of these factors into account and compared it with a model that only tracked the effects of size on the price, we discovered that all of the prediction factors must be left in the model and we could produce an effective model for forecasting prices throughout the entire data set.

Based on distinctive stepwise-type jumps in the data on the graph below and our general understanding of the diamond market, we decided to conduct further analysis into the following sub-markets within the industry: accent diamonds (<0.5 carats), affordable solitaire rings/jewelry (.5 to 2 carat); expensive engagement rings (2 to 5 carats); and luxury offerings (>5 carats) to see if we could further understand customer behavior.
(FIGURE 1)

Visualizing Market Segments for Diamond Purchases



When the data was analyzed by number of carats, some interesting effects became apparent. In this analysis, we found that all factors played into pricing the luxury stones and expensive rings, with color becoming increasingly important as size increased, but for accent diamonds, price was predominantly driven by size while clarity became statistically irrelevant.

However, we couldn't prove with statistical confidence that the separate, specific models were more effective at forecasting price than the heavier, simpler model. This means that even though there were indicators of shifting preferences as size increased, these shifts weren't powerful enough to change the model with 95% statistical confidence.

Results: Ultimately, when comparing the four subset models versus the full model, we concluded that all five models accurately predicted price models, but the single model that forecasted the entire data set was ideal due to its simplicity of design and effective forecasting ability.

We recommend that future statisticians using these models as a basis for future study explore different types of non-linear modelling procedures; identify, account for, and potentially delete outliers; and use logarithmic transforms on the smaller models to more accurately fit the linear model.

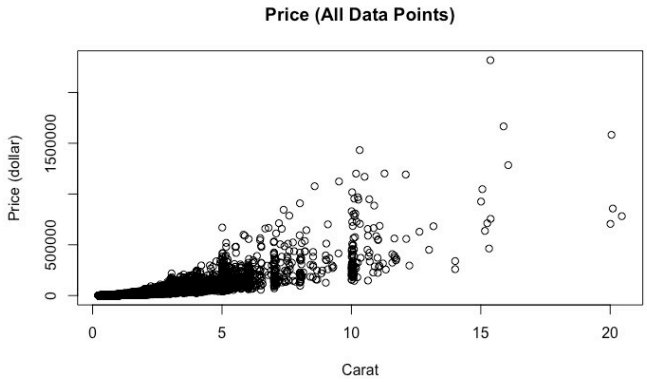
Exploratory Data Analysis

The dataset used for this analysis consisted of 210,638 observations. The response variable was the price for a given diamond (selected by Blue Nile), which is a quantitative, continuous metric in American dollars. Within this data set were five variables:

- *Price*: a quantitative variable, measured in dollars.
- *Carat*: a quantitative variable that measures diamond's weight.
- *Cut*: a categorical variable with four (4) classes indicating how the diamond reflects light due to cut and how it sparkles organized from most desirable to least (Astor Ideal, Ideal, Very Good, Good).

- *Color*: a categorical variable with seven (7) classes organized from colorless to slightly hued (D, E, F, G, H, I, J).
- *Clarity*: a categorical variable with eight (8) classes organized from flawless to "slightly included" (FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2).

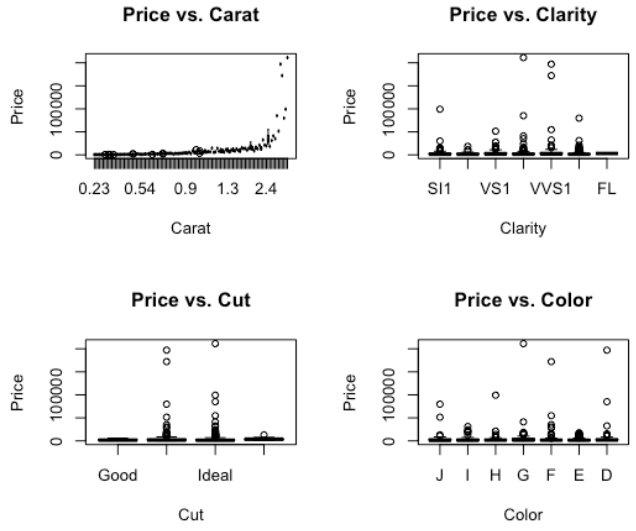
Our primary question of interest was to explore whether a comprehensive, inclusive single model with a potential transformation with the entire dataset or if a smaller model fitted to smaller samples from each market would better model Blue Nile's prices.
(FIGURE 2)



We first conducted an investigation of the full dataset by merely plotting price on the Y-axis and carat on the X-axis (see Figure 1). This prompted concerns that the extremely large number of data points could potentially -- and falsely -- support the fit of almost any model. This factor would also later drive experimentation with taking samples of the larger dataset.

We next examined each individual categorical variable's relationship to price, plotting price against variable class (Figure 2).

(FIGURE 3)



On cursory visual inspection, there did not appear to be a linear relationship between a single predictor and pricing. However, in looking at Figure 1, there appeared to be periodic "shelf-like" or "step-like" rises in prices along the exponential curve at specific carat values. This

observation prompted our hypothesis that Blue Nile's pricing structure is potentially segmented along socioeconomic lines. Further, we believe that carat-size, when studied in concert with other categorical variables, may explain price point differences across market segments. (Figure 1)

For this reason, we decided to approach the data set from two directions: first, by creating a model that attempted to address the whole range of data and second, by attempting to fit multiple linear regressions on subsections of the data based on carat size. We then compared the predictive power of the full model to the subsections'.

Detailed Analysis

Direction 1: Full Model of All Available Data

We first checked whether any variables in the data set contributed to the full model's predictive ability for all regressor variables, e.g., carat size, clarity, cut, and color (Figure 4) with an ANOVA F-Test. Results of the F-statistic and corresponding P-value for this full model led us to reject the null hypothesis that none of the regressors contributed to the model.
$H_0: \beta_1 = \beta_2 = \cdots \beta_{20}$
$H_A: \beta_j \neq 0$ for some j = (1, 2, …,20)

(Figure 4)

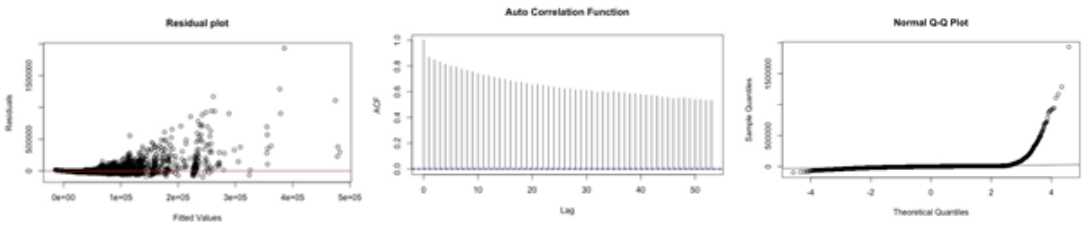
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9961.34     538.52  18.497  < 2e-16 ***
carat        24512.88      45.76 535.640  < 2e-16 ***
clarityIF   -18567.14     496.50 -37.396  < 2e-16 ***
claritySI1  -23210.90     479.61 -48.395  < 2e-16 ***
claritySI2  -23752.02     483.99 -49.076  < 2e-16 ***
clarityVS1  -21794.88     480.01 -45.405  < 2e-16 ***
clarityVS2  -22571.31     480.41 -46.984  < 2e-16 ***
clarityVVS1 -20138.46     482.18 -41.766  < 2e-16 ***
clarityVVS2 -20688.67     481.81 -42.939  < 2e-16 ***
colorE         -240.19     112.43  -2.136   0.0326 *
colorF         -846.05     111.88  -7.562 3.97e-14 ***
colorG        -2117.49     113.88 -18.594  < 2e-16 ***
colorH        -3109.97     120.80 -25.745  < 2e-16 ***
colorI        -3328.87     122.70 -27.131  < 2e-16 ***
colorJ        -5311.58     142.94 -37.159  < 2e-16 ***
cutGood       -1866.38     283.59  -6.581 4.68e-11 ***
cutIdeal       1634.42     258.44   6.324 2.55e-10 ***
cutVery Good  -1333.65     261.94  -5.091 3.56e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14790 on 210620 degrees of freedom
Multiple R-squared:  0.5829,    Adjusted R-squared:  0.5829
F-statistic: 1.732e+04 on 17 and 210620 DF,  p-value: < 2.2e-16
```

In the table, the t-tests for individual variables were all statistically significant, which we viewed with a degree of skepticism. This seemed to imply that none of the regressors could be candidates for removal. In view of these findings, our next step was to ensure that regressors within the full model weren't correlated with each other, and therefore did not introduce multicollinearity to the model. Results for pairwise correlations were as follows (figure 5).
(Figure 5)

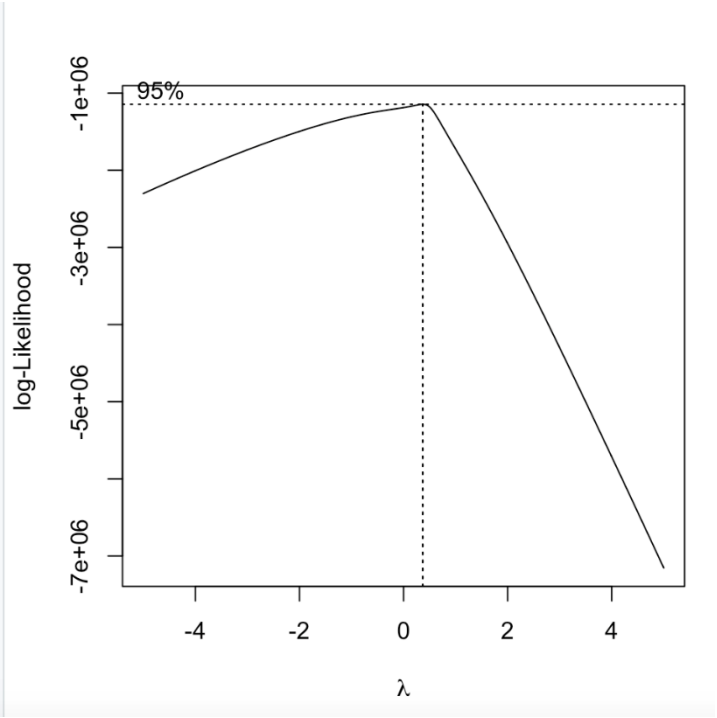
	carat	clarity	color	cut	price
carat	1.00000000	-0.018205586	0.06602168	0.07314710	0.751589247
clarity	-0.01820559	1.000000000	0.03719003	-0.01113410	0.001802124
color	0.06602168	0.037190028	1.00000000	-0.03600979	-0.021502550
cut	0.07314710	-0.011134097	-0.03600979	1.00000000	0.031281452
price	0.75158925	0.001802124	-0.02150255	0.03128145	1.000000000

Because no pairs of predictors appeared to be correlated in the pairwise chart, we calculated the VIFs for each coefficient. These results suggested that all classes within clarity, as well as two classes within the cut category, possessed linear dependencies. However, utilizing partial F-tests to compare the full model against each reduced model combination of regressors, we were unable to reasonably remove variables from the model without concern about their contribution to the model. Given that some models can possess predictive value in the presence of multicollinearity, we proceeded with our investigation, noting the likely presence of linear dependence between variables.
(Figure 6)



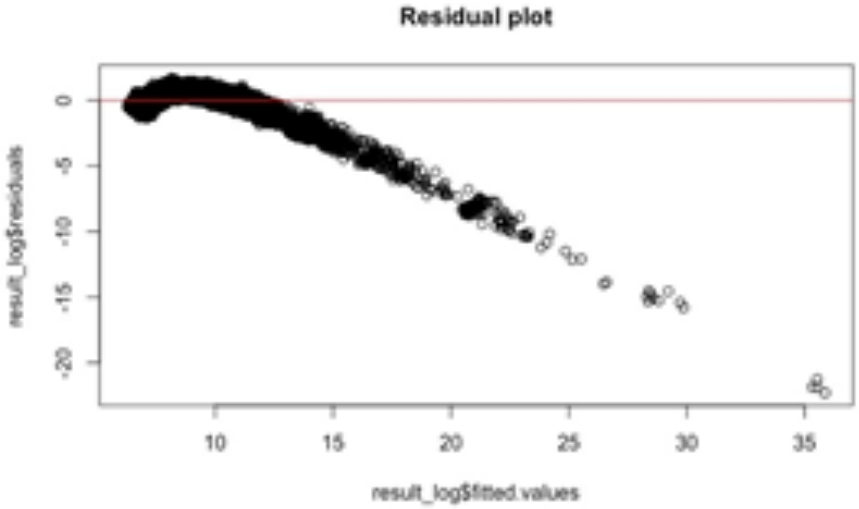
To further investigate the full data set, we checked linear model assumptions (Figure 6). A residual plot revealed that the residuals for a full model were not evenly distributed around a mean of 0 and that variance was not constant. An autocorrelation function revealed that the errors across all lags were significantly correlated, a potentially serious fault which we tentatively attributed to the multicollinearity discovered in the model. Finally, a QQ plot revealed that the residuals did not follow a normal distribution.

According to these plots, none of the linear assumptions (residual distribution about the mean and constant variance, in particular) were met. Despite these setbacks, we had a reasonable suspicion that a robust model could be formed with the existing variables given the exponential shape of the error variance and normal distribution. To that end, we decided to transform the response variable price first to stabilize variance. Conducting a Box-Cox plot of the dataset suggested that taking the natural logarithm of price would yield a suitable transformation, as 0 was in the 95% confidence interval for the log-likelihood graph (Figure 7).
(Figure 7)



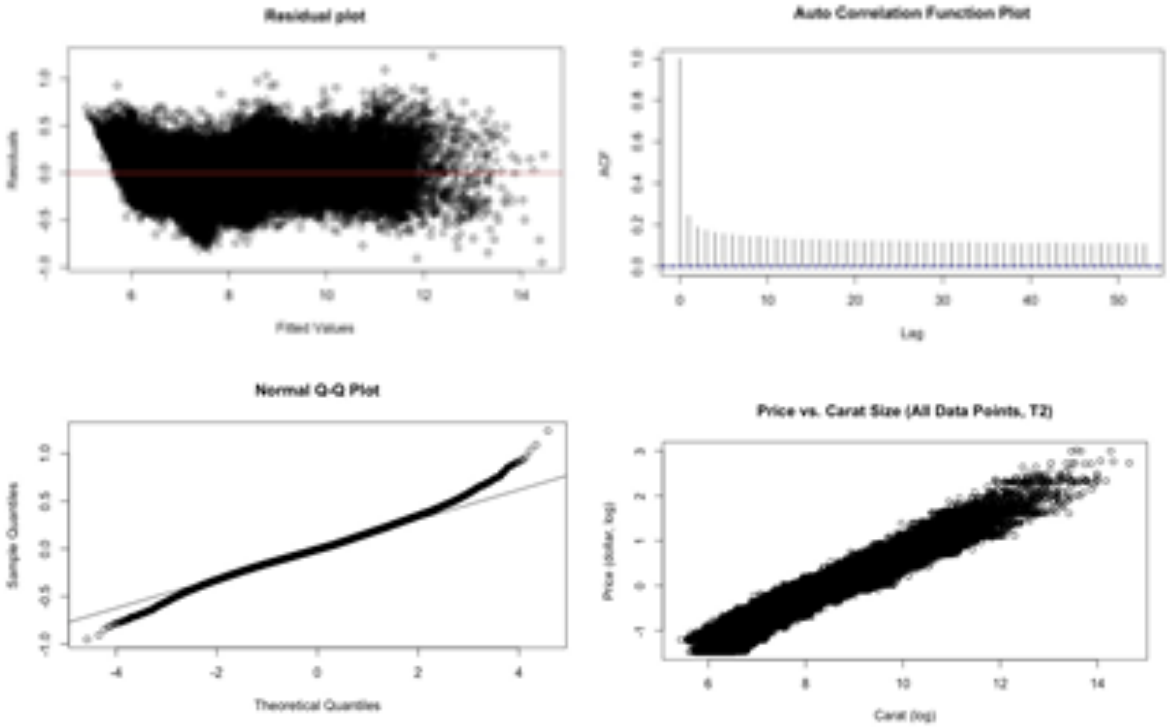
The transformed data yielded far better variance, as seen below (Figure 8). However, because the residuals still weren't distributed evenly around mean 0, we decided to conduct a second transformation of the only quantitative variable in the full model: carat size.

(Figure 8)



In light of the semi-exponential relationship between carat size and price, we decided to take the natural logarithm of the carat regressor to improve the residual distribution around mean 0 and check our assumptions (Figure 9). This subsequent transformation yielded residuals with constant variance and a far cleaner distribution of the residuals. Furthermore, the correlation across lags was reduced, and the residual distribution appeared to have normalized. Finally, the relationship between price and carat size after this second transformation also seemed to display a far more linear relationship.

(Figure 9)



Having normalized the linear model's residuals through transformation, the idealized full model for the data set distilled a linear regression model with 20 different predictor variables. The number of variables all proved to be statistically significant with the number of variables being driven up by the various factors to cut, clarity, and color (Figure 10).

(Figure 10)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.3499089  0.0060069 1556.53   <2e-16 ***
log_carat     1.9654032  0.0006011 3269.73   <2e-16 ***
cutGood      -0.2986523  0.0031754  -94.05   <2e-16 ***
cutIdeal     -0.0822843  0.0028952  -28.42   <2e-16 ***
cutVery Good -0.2527184  0.0029324  -86.18   <2e-16 ***
clarityIF    -0.2602883  0.0055576  -46.83   <2e-16 ***
claritySI1   -0.6385266  0.0053676 -118.96   <2e-16 ***
claritySI2   -0.7748139  0.0054166 -143.04   <2e-16 ***
clarityVS1   -0.4691355  0.0053727  -87.32   <2e-16 ***
clarityVS2   -0.5272827  0.0053771  -98.06   <2e-16 ***
clarityVVS1  -0.3494636  0.0053965  -64.76   <2e-16 ***
clarityVVS2  -0.4200952  0.0053925  -77.90   <2e-16 ***
colorE       -0.0622330  0.0012591  -49.43   <2e-16 ***
colorF       -0.0947113  0.0012528  -75.60   <2e-16 ***
colorG       -0.1534624  0.0012758 -120.29   <2e-16 ***
colorH       -0.2166832  0.0013534 -160.10   <2e-16 ***
colorI       -0.3172295  0.0013738 -230.92   <2e-16 ***
colorJ       -0.4458935  0.0016000 -278.68   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1656 on 210620 degrees of freedom
Multiple R-squared:  0.9813,    Adjusted R-squared:  0.9813
F-statistic: 6.517e+05 on 17 and 210620 DF,  p-value: < 2.2e-16
```

Direction 2: Subsetting the Data to Correlate to Different Markets

As previously stated, we assumed that – based on the visual "stairstep" pattern in the full model scatterplot and our knowledge of the diamond market – four different, smaller models might predict relationships in the variables better than one standardized curve. Our first model was for accent diamonds (<0.5 carats), the second for affordable engagement ring sized diamonds (0.5 to 2 carats), the third for expensive engagement ring diamonds (2 to 5 carats), and the fourth for very large diamonds used in elaborate, luxury jewelry (>5 carats).

To build comparable models, we randomly sampled 500 data points from each category to build the sub-models. We then modeled each of the variables against price to determine characteristics of the model based on the response variables (plotted in full detail in Appendix 1). To complement these visual models, we checked correlation in the data sets by building the following correlation matrices (Figures 15-18):

Accent Market (<0.5 carats)

	accentcarat	accentcolor	accentcut	accentprice
accentcarat	1.00000000	0.04550904	0.039203491	0.760398189
accentcolor	0.04550904	1.00000000	-0.066045291	-0.036290087
accentcut	0.03920349	-0.06604529	1.000000000	0.003147219
accentprice	0.76039819	-0.03629009	0.003147219	1.000000000

Cheap Engagement Ring Diamonds (0.5 to 2 carats)

	cheapcarat	cheapclarity	cheapcolor	cheapcut	cheapprice
cheapcarat	1.00000000	0.0591547363	0.1384060025	-0.05536023	0.87734356
cheapclarity	0.05915474	1.0000000000	-0.0007325359	-0.02696478	0.14306450
cheapcolor	0.13840600	-0.0007325359	1.0000000000	0.00713354	-0.08110253
cheapcut	-0.05536023	-0.0269647787	0.0071335397	1.00000000	-0.10595377
cheapprice	0.87734356	0.1430645011	-0.0811025265	-0.10595377	1.00000000

Expensive Engagement Ring Diamonds (2 to 5 carats)

	expensivecarat	expensiveclarity	expensivecolor	expensivecut	expensiveprice
expensivecarat	1.00000000	-0.034473189	0.12089287	0.045284981	0.72172987
expensiveclarity	-0.03447319	1.000000000	0.04637646	0.002523922	0.04730432
expensivecolor	0.12089287	0.046376456	1.00000000	-0.014969114	-0.32836605
expensivecut	0.04528498	0.002523922	-0.01496911	1.000000000	-0.01905326
expensiveprice	0.72172987	0.047304324	-0.32836605	-0.019053263	1.00000000

Large Diamonds (>5 carats).

	bigcarat	bigclarity	bigcolor	bigcut	bigprice
bigcarat	1.00000000	0.00640508	0.009525570	0.059843330	0.66606474
bigclarity	0.00640508	1.00000000	0.106960215	0.076728278	-0.08474591
bigcolor	0.00952557	0.10696022	1.000000000	-0.008529694	-0.45380813
bigcut	0.05984333	0.07672828	-0.008529694	1.000000000	0.05563158
bigprice	0.66606474	-0.08474591	-0.453808133	0.055631578	1.00000000

Both the models and the covariance tables determined that, throughout all five models, carat size was the biggest predictor of price. Within the cheap engagement ring market, carat size was more strongly correlated with price than in any other tier. Additionally, as size increased,

other factors began to more heavily influence price. This trend is seen most clearly in the variable for color, which shows a marked increase in correlation. It therefore seems to be a more important factor to larger diamond customers.

The next step was to do a step analysis to determine if any of the variables identified as having weak covariance factors could be dropped. Only accent diamonds generated enough statistical significance to drop a variable (cut) from the full model. This resulted in our formation of ideal functions for each subset data. All of these sub-set models contain a strong statistical significance (Figures 11).

(Figure 11)

Figure: >5 Carats

```
Call:
lm(formula = bigprice ~ bigcarat + bigclarity + bigcolor + bigcut)

Residuals:
    Min      1Q  Median      3Q     Max
-337454  -39146   -2714   35006  578424

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       101419      28025   3.619 0.000327 ***
bigcarat           64723       1761  36.762  < 2e-16 ***
bigclarityIF     -107997      31456  -3.433 0.000648 ***
bigclaritySI1    -249953      27488  -9.093  < 2e-16 ***
bigclaritySI2    -251182      28874  -8.699  < 2e-16 ***
bigclarityVS1    -181723      27389  -6.635 8.73e-11 ***
bigclarityVS2    -225940      27474  -8.224 1.84e-15 ***
bigclarityVVS1   -158448      28486  -5.562 4.41e-08 ***
bigclarityVVS2   -152971      28398  -5.387 1.12e-07 ***
bigcolorE         -76083      16788  -4.532 7.38e-06 ***
bigcolorF         -64022      15573  -4.111 4.63e-05 ***
bigcolorG        -108448      15735  -6.892 1.72e-11 ***
bigcolorH        -157885      15434 -10.230  < 2e-16 ***
bigcolorI        -170478      15111 -11.282  < 2e-16 ***
bigcolorJ        -194255      14895 -13.042  < 2e-16 ***
bigcutIdeal        33778      11582   2.916 0.003705 **
bigcutVery Good     8966      11075   0.810 0.418569
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 75960 on 483 degrees of freedom
Multiple R-squared:  0.8251,    Adjusted R-squared:  0.8193
F-statistic: 142.4 on 16 and 483 DF,  p-value: < 2.2e-16
```

Figure: 2-5 Carats

```
Call:
lm(formula = expensiveprice ~ expensivecarat + expensiveclarity +
    expensivecolor + expensivecut)

Residuals:
   Min     1Q Median     3Q    Max
-43844  -7048   -985   5027 145150

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            33343.2    15353.4   2.172  0.03036 *
expensivecarat         30797.9      945.6  32.570  < 2e-16 ***
expensiveclarityIF    -36006.6     6148.9  -5.856 8.79e-09 ***
expensiveclaritySI1   -57374.2     5661.7 -10.134  < 2e-16 ***
expensiveclaritySI2   -59837.2     5760.6 -10.387  < 2e-16 ***
expensiveclarityVS1   -46677.1     5604.3  -8.329 8.53e-16 ***
expensiveclarityVS2   -50905.6     5659.5  -8.995  < 2e-16 ***
expensiveclarityVVS1  -41205.0     5800.1  -7.104 4.38e-12 ***
expensiveclarityVVS2  -45844.7     5718.0  -8.018 8.23e-15 ***
expensivecolorE        -3070.5     2393.3  -1.283  0.20011
expensivecolorF        -7749.0     2252.0  -3.441  0.00063 ***
expensivecolorG       -12327.1     2302.9  -5.353 1.34e-07 ***
expensivecolorH       -18131.9     2273.9  -7.974 1.13e-14 ***
expensivecolorI       -22736.1     2374.2  -9.576  < 2e-16 ***
expensivecolorJ       -33837.0     2716.8 -12.455  < 2e-16 ***
expensivecutGood      -15655.5    14455.4  -1.083  0.27934
expensivecutIdeal     -11389.2    14248.8  -0.799  0.42451
expensivecutVery Good -18317.0    14245.9  -1.286  0.19914
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14070 on 482 degrees of freedom
Multiple R-squared:  0.766,     Adjusted R-squared:  0.7578
F-statistic: 92.83 on 17 and 482 DF,  p-value: < 2.2e-16
```

Figure: 0.5-2 Carats

```
Call:
lm(formula = cheapprice ~ cheapcarat + cheapclarity + cheapcolor +
    cheapcut)

Residuals:
    Min      1Q  Median      3Q     Max
-4383.2  -724.8  -125.5   538.5 13381.9

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -3430.17    1290.13  -2.659 0.008103 **
cheapcarat       10439.79     198.62  52.561  < 2e-16 ***
cheapclarityIF     365.27    1048.26   0.348 0.727650
cheapclaritySI1   -566.74    1016.02  -0.558 0.577237
cheapclaritySI2  -1490.56    1024.24  -1.455 0.146239
cheapclarityVS1    -26.46    1017.57  -0.026 0.979266
cheapclarityVS2   -143.73    1020.76  -0.141 0.888078
cheapclarityVVS1   546.49    1020.93   0.535 0.592701
cheapclarityVVS2   332.03    1019.82   0.326 0.744887
cheapcolorE       -668.86     228.65  -2.925 0.003604 **
cheapcolorF       -578.83     230.93  -2.507 0.012521 *
cheapcolorG       -856.69     228.90  -3.743 0.000204 ***
cheapcolorH      -1502.83     234.47  -6.409 3.48e-10 ***
cheapcolorI      -1767.31     247.31  -7.146 3.32e-12 ***
cheapcolorJ      -2516.05     275.03  -9.148  < 2e-16 ***
cheapcutGood     -1031.18     845.50  -1.220 0.223206
cheapcutIdeal      691.38     823.72   0.839 0.401695
cheapcutVery Good  -446.06     826.33  -0.540 0.589579
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1401 on 482 degrees of freedom
Multiple R-squared:  0.8658,    Adjusted R-squared:  0.861
F-statistic: 182.9 on 17 and 482 DF,  p-value: < 2.2e-16
```

Figure: <0.5 Carats

```
Call:
lm(formula = accentprice ~ accentcarat + accentclarity + accentcolor +
    accentcut)

Residuals:
   Min     1Q Median     3Q    Max
-25271  -2115   -158   2456 107573

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -8516.37    7489.57  -1.137 0.256063
accentcarat      19343.94     505.90  38.237  < 2e-16 ***
accentclarityIF  -1040.55    6865.52  -0.152 0.879596
accentclaritySI1 -3455.92    6739.55  -0.513 0.608338
accentclaritySI2 -4009.20    6767.03  -0.592 0.553819
accentclarityVS1 -2084.27    6754.66  -0.309 0.757783
accentclarityVS2 -1178.17    6756.46  -0.174 0.861643
accentclarityVVS1  -496.96    6770.78  -0.073 0.941520
accentclarityVVS2 -1178.40    6753.41  -0.174 0.861555
accentcolorE       1714.43    1091.28   1.571 0.116830
accentcolorF        -11.18    1037.96  -0.011 0.991409
accentcolorG      -1435.51    1038.02  -1.383 0.167326
accentcolorH      -1403.89    1058.66  -1.326 0.185437
accentcolorI      -2284.40    1160.80  -1.968 0.049647 *
accentcolorJ      -4403.31    1242.89  -3.543 0.000434 ***
accentcutGood     -2554.29    3505.16  -0.729 0.466525
accentcutIdeal       36.18    3386.10   0.011 0.991479
accentcutVery Good  -449.30    3403.53  -0.132 0.895031
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6672 on 482 degrees of freedom
Multiple R-squared:  0.7636,    Adjusted R-squared:  0.7553
F-statistic:  91.6 on 17 and 482 DF,  p-value: < 2.2e-16
```

These sub-models assumptions were checked against underlying assumptions of linear model adequacy, and each model met the assumptions of linearity, residual distribution about the

mean, and non-correlation of error. The full assumption models are in Appendix 2. Though we concluded the models were visually "good enough," the error variance was not completely constant and mean seemed to favor below 0. The normal distribution for the errors also appears to turn exponential near the tail of each subset. Further analysis could be made to transform the models to minimize these assumptions similar to the approach taken with the full model.

Full-Model vs. Subset Model Comparison

In order to conduct an equal comparison between the models we pulled a random 500-sample size subset of the data sets in order to build models in the processes we proved to be ideal above. This equal sample size allows us to use all of the regression effectiveness measures without having the number of predictors skew the model. Its important to note that this test doesn't test the actual models shown above, but the process used to develop them in a standardized procedure in order to eliminate variance based on different sizes of data sets. (Figure 23)

	df	BIC
new_model	19	-186.9555
big_full_model	18	12751.5276
expensive_full_model	19	11070.1734
cheap_full_model	19	8763.7829
ideal_accent_model	12	10297.9576

	df	AIC
new_model	19	-267.033
big_full_model	18	12675.665
expensive_full_model	19	10990.096
cheap_full_model	19	8683.705
ideal_accent_model	12	10247.382

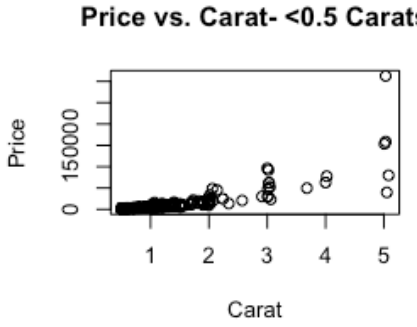
In the tables above we see that the model built from the full dataset using a more robust statistical approach (new_model) proved to have a more precise ability to predict the response variable.

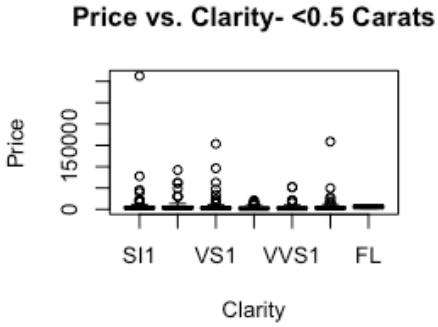
Result

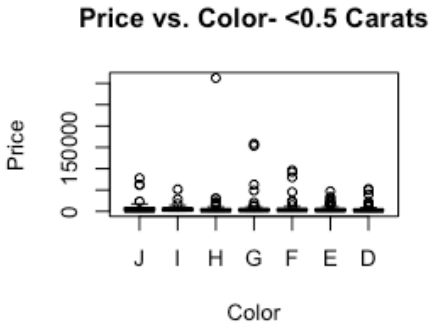
Conducting statistical analysis on the full dataset with transformations produced a statistically significant model, but so did dividing the data into subsets and analyzing it untransformed. Within the carat tiers, we determined that the importance of color increased as the carat size increased. Likewise, clarity became statistically irrelevant as the carat size approached 0. However, we did not think that the adjustments in preference were significant enough to merit significantly more complex modeling procedures.

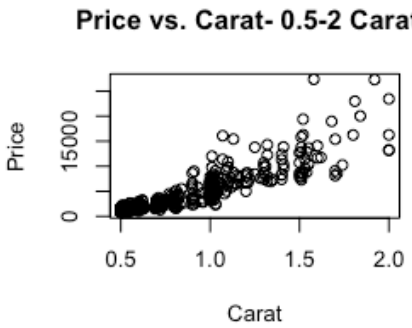
It is our hope that future statisticians will use this model as the basis for a more specific model. This analysis gives high level insight into changing customer behavior within the diamond market while providing a tangible predictive model for future research.

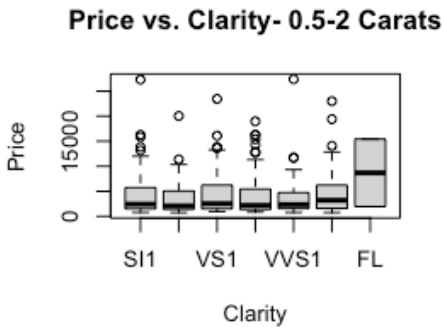
Appendix 1: Model Outputs to Show Covariance in Data

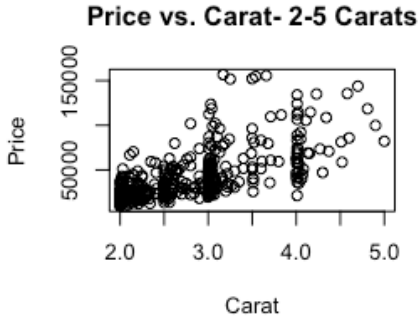


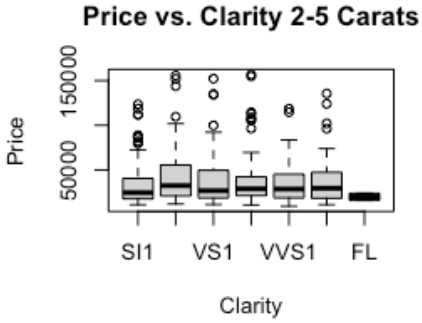


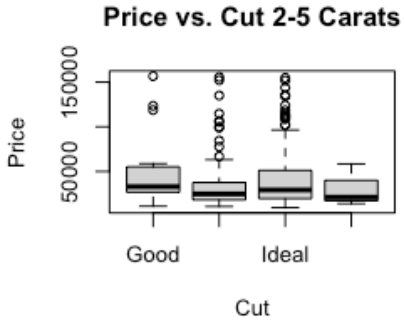


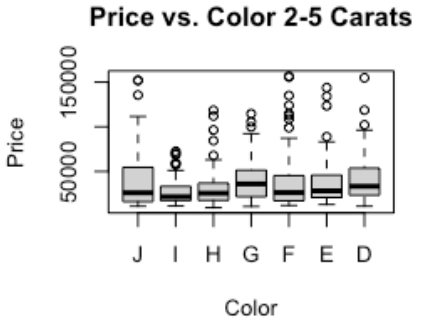


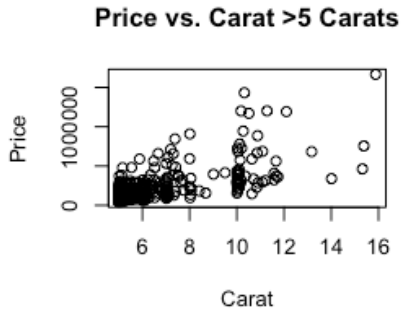


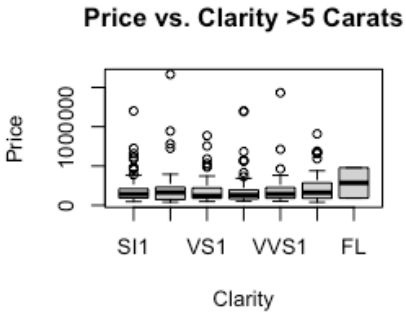

Price vs. Carat- 2-5 Carats
Price
50000
150000
2.0
3.0
4.0
5.0
Carat

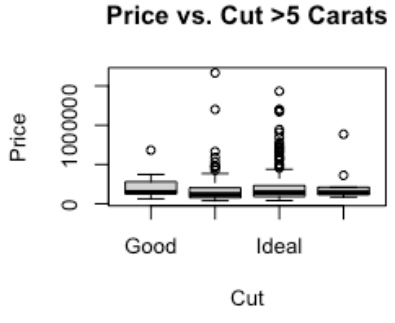

Price vs. Clarity 2-5 Carats
Price
50000
150000
SI1
VS1
VVS1
FL
Clarity


Price vs. Cut 2-5 Carats
Price
50000
150000
Good
Ideal
Cut


Price vs. Color 2-5 Carats
Price
50000
150000
J
I
H
G
F
E
D
Color

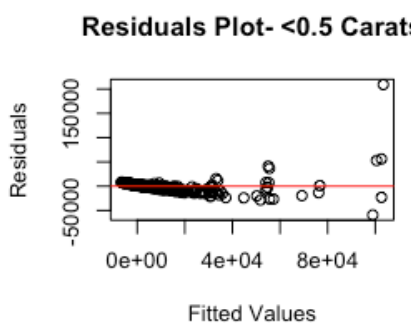

Price vs. Carat >5 Carats
Price
0
1000000
6
8
10
12
14
16
Carat

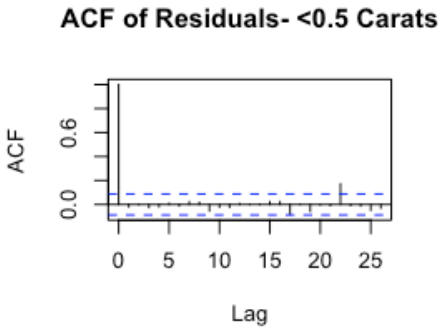

Price vs. Clarity >5 Carats
Price
0
1000000
SI1
VS1
VVS1
FL
Clarity

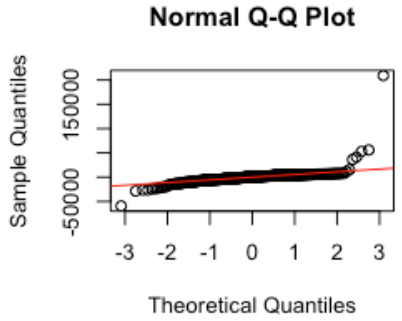

Price vs. Cut >5 Carats
Price
0
1000000
Good
Ideal
Cut

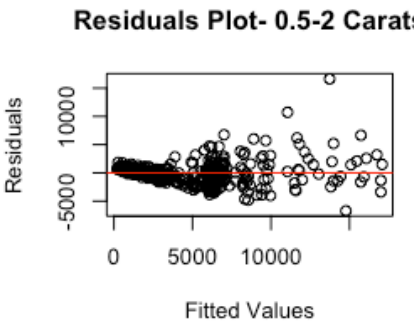

Price vs. Color >5 Carats
Price
0
1000000
J
I
H
G
F
E
D
Color

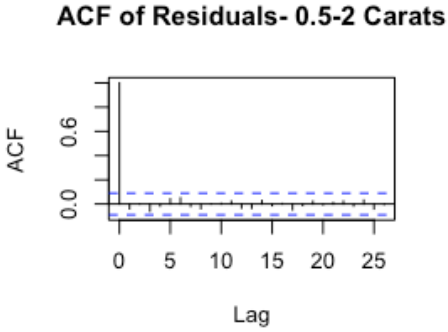
Appendix 2: Model Outputs to Validate Assumptions

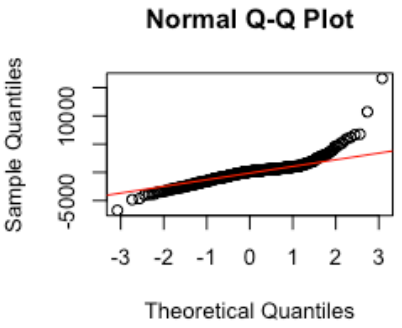


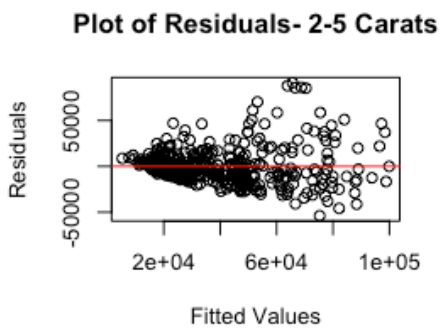


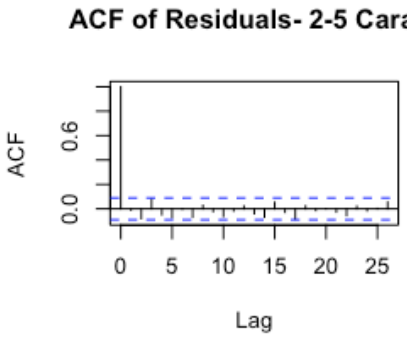


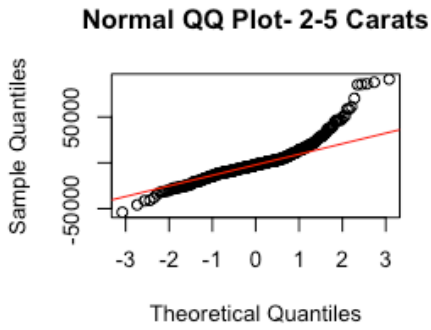


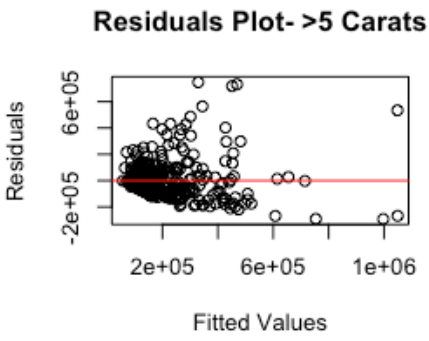


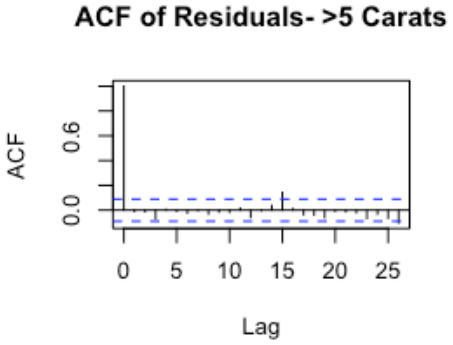


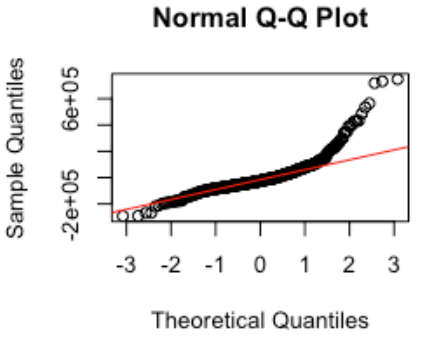

Plot of Residuals- 2-5 Carats
Residuals
-50000
50000
2e+04
6e+04
1e+05
Fitted Values


ACF of Residuals- 2-5 Carats
ACF
0.0
0.6
0
5
10
15
20
25
Lag


Normal QQ Plot- 2-5 Carats
Sample Quantiles
-50000
50000
-3
-2
-1
0
1
2
3
Theoretical Quantiles


Residuals Plot- >5 Carats
Residuals
-2e+05
6e+05
2e+05
6e+05
1e+06
Fitted Values


ACF of Residuals- >5 Carats
ACF
0.0
0.6
0
5
10
15
20
25
Lag


Normal Q-Q Plot
Sample Quantiles
-2e+05
6e+05
-3
-2
-1
0
1
2
3
Theoretical Quantiles