STAT-225 (Nonparametric Statistics) Project Report

Predicting Diamond Price

Group 8: Nicole Frontero, Anna Ballou, Alex Russell

April 22, 2020

Introduction and Exploratory Analyses

We aim to identify the best metric for predicting the price of a diamond. Diamonds vary in many ways from one another, for example, in various measurements of size, as well as in color and cut quality. Our question of interest is: Are certain qualities of diamonds better predictors for diamond price? In other words, are certain characteristics of diamonds more strongly associated with the price of a diamond than others?

The data is from the `diamonds` dataset from the `ggplot2` package in R. The data can be found on the official `ggplot2` webpage.[1] This dataset contains the prices and other attributes of nearly 54,000 diamonds. Each observation in this dataset represents a unique diamond. We randomly sampled 500 observations from the overall dataset.

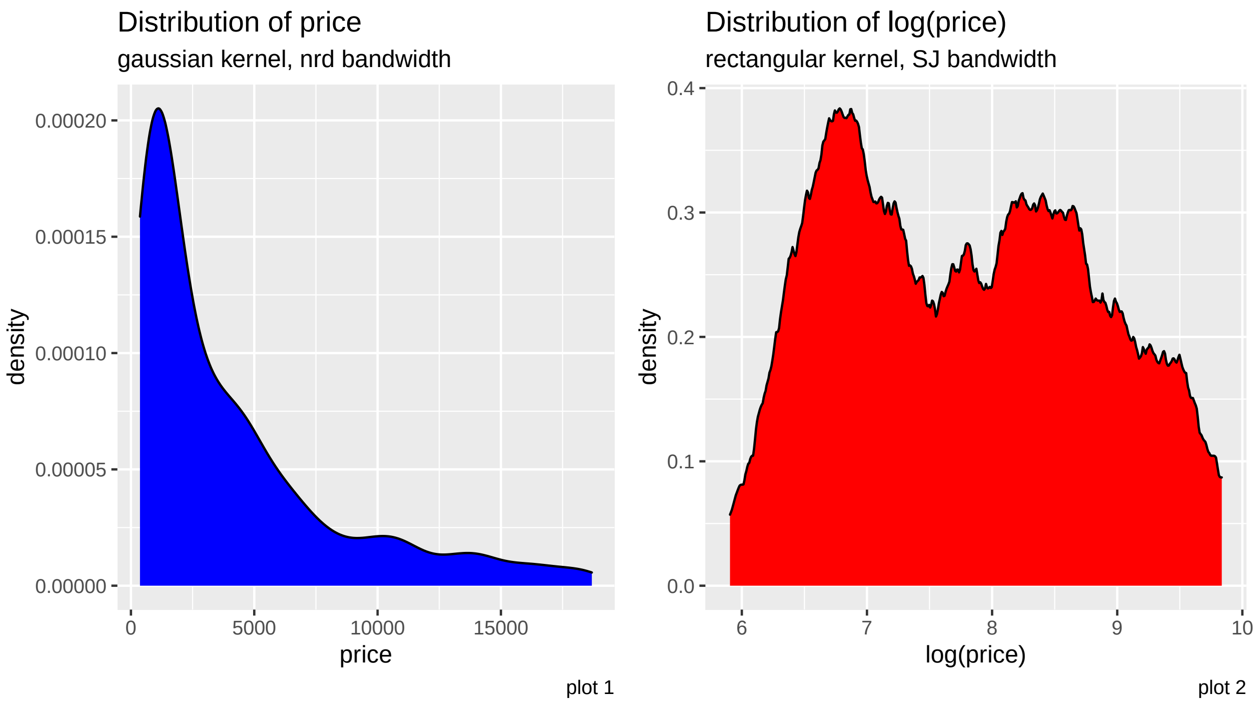
The observational units in our dataset are diamonds. Since we took a random sample from the larger dataset (`diamonds`), we would expect that our data is representative of the approximately 54,000 diamonds in `diamonds`. We are assuming that the observations in `diamonds` are representative of the world diamond population.

Below is brief introduction to our data:

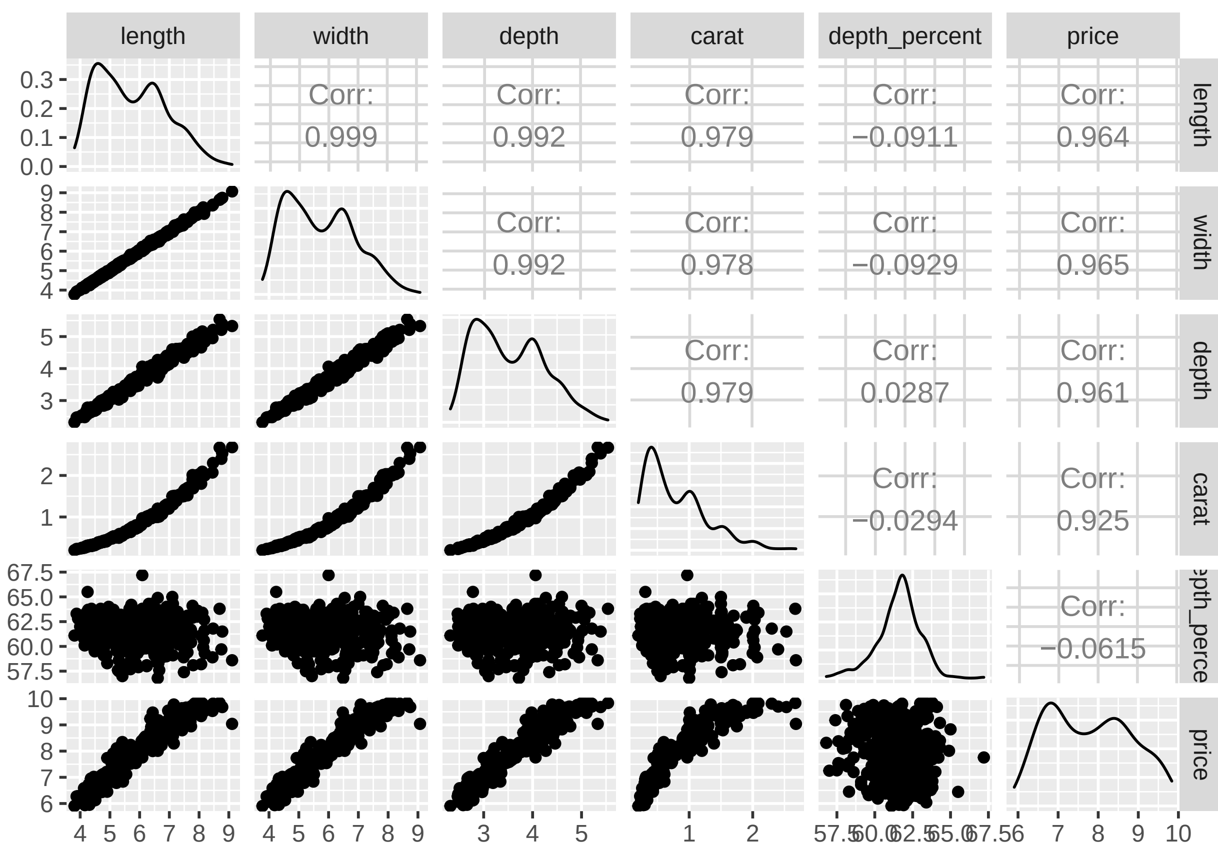
- The response variable is price in US dollars, which ranges between \$326 - \$18,823.
- The explanatory variables are as follows:
  - Carat: measure of mass (weight) of the diamond. Ranges between (0.2ct - 5.01ct). Note that 1 ct = 200 mg.
  - length: length in mm. Ranges between (0mm - 58.9mm).
  - width: width in mm. Ranges between (0mm - 58.9mm).
  - depth: depth in mm. Ranges between (0mm - 31.8mm).
  - depth: total depth percentage, which is calculated as $\frac{z}{\text{mean}(x,y)} = \frac{2z}{x+y}$, and ranges from 43% - 79%.
  - table: the width of the top of the diamond relative to the widest point in mm. Ranges between 43mm - 95mm.
  - cut: quality of the cut. Includes fair, good, very good, premium and ideal.
  - clarity: a measurment of how clear the diamond is. Ranges between "I1" (worst) to "IF" (best).
  - color: diamond color. Ranges between D (best) to J (worst.

When initially examining `price`, we noticed that it was heavily right skewed. As a result, we decided to log transform price (base e). A comparison of the non-transformed and transformed distribution of price can be seen in plots 1 and 2 below.

[1] https://github.com/tidyverse/ggplot2/blob/master/data-raw/diamonds.csv



We utilized `ggpairs()` to examine the quantitative variables in our data.



As seen above, we noticed that `length`, `width` and `depth` appear to have a strong correlation with log(`price`) and all follow a very similar trend (upward sloping and linear). Additionally, these three variables are also strongly correlated with `carat`. `depth_percent` does not appear to be correlated with log(`price`) (Pearson's correlation of -0.0615). Lastly, `carat` appeared to have the most non-linear relationship with log(`price`).

Because of this observed non-linear relationship between `carat` and `price`, we hypothesized that the correlation reported in the `ggpairs()` output above underestimated the actual correlation. `ggpairs()` reports correlation using a Pearson Coefficient, which is ideal for linear relationships, but often underestimates non-linear correlation. Thus, we performed a Spearman's test for correlation on `carat` and `price`. Spearman's method is ideal for non-linear relationships. The results of our Spearman's correlation test (outlined below) give a correlation of 0.965 (which is larger than the reported correlation of 0.925).

Hypotheses:

$H_0 : \rho = 0; H_A : \rho \neq 0$

Let $\alpha = 0.05$ and ρ represent the correlation between length and price.

Assumptions:

We will assume that all pairs of observations are independent from each other.

Test Results:

$p \approx 0$, $\rho = 0.965$

Conclusion:

Because $p \approx 0 < \alpha = 0.05$ for the Spearman's's test for correlation, we reject our null hypotheses. Thus, carat is correlated with log(price).

Methods

OLS MLR

We aimed to build a more precise model that eliminated predictors that did not enhance predictive power. We used forward stepwise regression, which systematically adds variables to a model until the adjusted R^2 (i.e. how well the model fits data) fails to increase. We utilized forward stepwise linear regression to examine all possible predictors (carat, depth, table, length, width, color, clarity, and cut). The results of the forward stepwise regression are depicted in Table 1.

Table 1: Stepwise regression results

Step	Predictors	$R^2_{\text{adj.}}$	AIC
1	width	0.931	116.286
2	clarity	0.956	-100.052
3	color	0.969	-265.746
4	carat	0.974	-360.658
5	depth	0.982	-545.237
6	cut	0.983	-563.024
7	length	0.983	-567.463

We noticed that forward stepwise regression excluded the `depth_percent` and `table` variables.

After stepwise regression, we built a multiple regression model to predict log(`price`) from the predictors included in forward stepwise regression. We exlcuded cut as it was the last categorical variable added and was largely explained by length and width. Each catgorical and quantitative variable with their respective coefficients and p-values can be seen below in Table 2.

Table 2: OLS model summary results

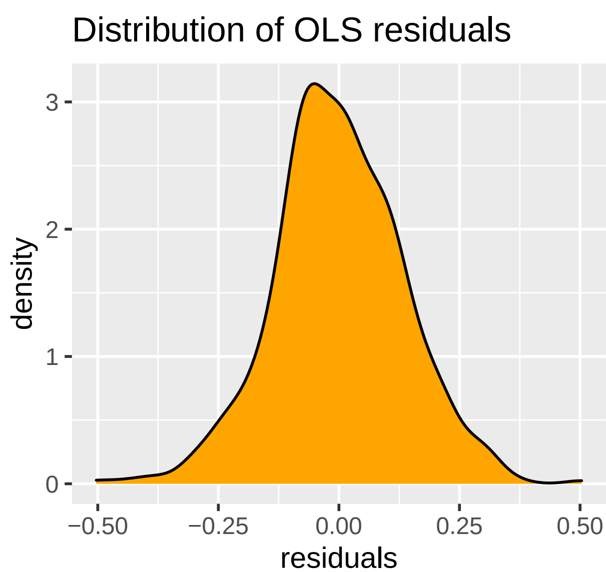
log(price) ~ width + clarity + color + carat + depth + length

Predictors	Estimate	P-value	Predictors	Estimate	P-value
(Intercept)	-0.0651831	0.59156	colorE	-0.0227264	0.30612
width	0.2949006	0.01319	colorF	-0.0803849	0.00034
clarityIF	1.0487289	< 0.0001	colorG	-0.1564346	< 0.0001
claritySI1	0.5720697	< 0.0001	colorH	-0.2358815	< 0.0001
claritySI2	0.3951465	< 0.0001	colorI	-0.3170297	< 0.0001
clarityVS1	0.7991108	< 0.0001	colorJ	-0.4649712	< 0.0001
clarityVS2	0.7011019	< 0.0001	carat	-1.1056764	< 0.0001
clarityVVS1	0.9785003	< 0.0001	depth	1.0667157	< 0.0001
clarityVVS2	0.9082077	< 0.0001	length	0.4792458	< 0.0001

Note: $R^2_{\text{adj}} = 0.983$

Is OLS the best or should we use JHM (nonparametric approach)?

In order to decide if we were able to use an OLS model to predict log(`price`), we needed to examine the model's residuals for normality. Plot 3 below depicts the density of the OLS MLR model's residuals.



plot 3

At first glance, the distribution looks relatively normal. We formally tested the residuals for normality using a Kolmogorov-Smirnov test. The details of the test are outline below:

Hypotheses:

$H_0 : F(t) = F^{star}(t); H_A : F(t) \neq F^{star}(t)$ for at least one t, where $F^{star}(t)$ is the normal distribution and $F(t)$ is the observed distribution of price.

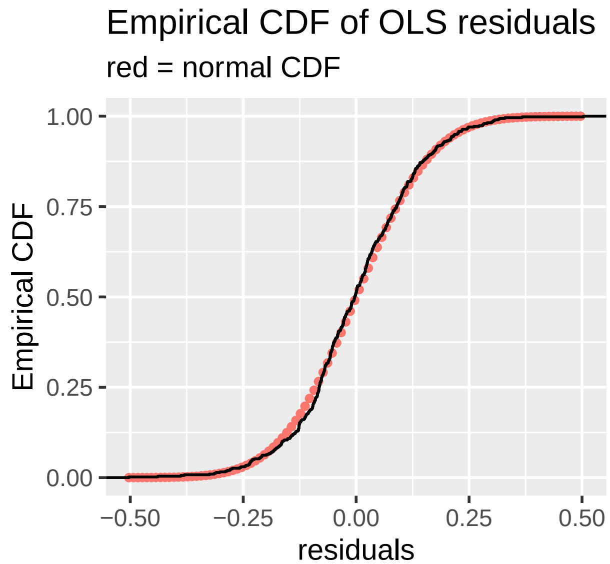
Assumptions:

Data come from a continuous distribution.

Test:

$p \approx 0$

The empirical CDF of price compared to the normal CDF with observed mean and standard deviation is below (plot 4):



plot 4

Conclusion:

Because $p \approx 0 < \alpha = 0.05$, we reject our null hypotheses. We conclude that the distribution of the OLS residuals is not normal.

JHM - best model

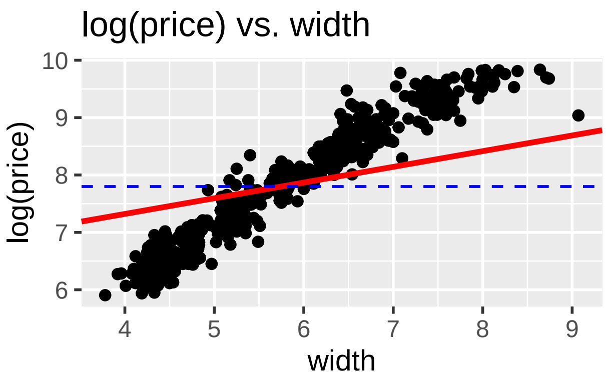
Since the OLS model residuals do not follow a normal distribution, we will create an JHM (rank based) regression model. Our JHM model included the same predictors as our OLS MLR. Each categorical and quantitative variable with their respective coefficients and p-values can be seen below in Table 3.

Table 3: JHM model summary results

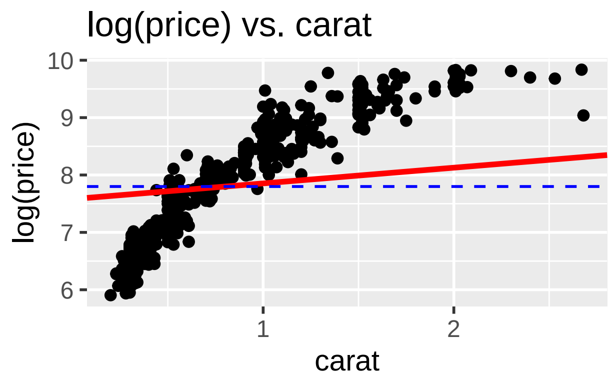
log(price) ~ width + clarity + color + carat + depth + length

Predictors	Estimate	P-value	Predictors	Estimate	P-value
(Intercept)	0.0090945	0.93893	colorE	-0.0227499	0.29299
width	0.2735470	0.01848	colorF	-0.0757285	0.00054
clarityIF	1.0108864	< 0.0001	colorG	-0.1425860	< 0.0001
claritySI1	0.5289852	< 0.0001	colorH	-0.2383103	< 0.0001
claritySI2	0.3557141	< 0.0001	colorI	-0.3201962	< 0.0001
clarityVS1	0.7575702	< 0.0001	colorJ	-0.4729935	< 0.0001
clarityVS2	0.6611988	< 0.0001	carat	-1.0753752	< 0.0001
clarityVVS1	0.9197320	< 0.0001	depth	1.0497605	< 0.0001
clarityVVS2	0.8545020	< 0.0001	length	0.5006167	< 0.0001

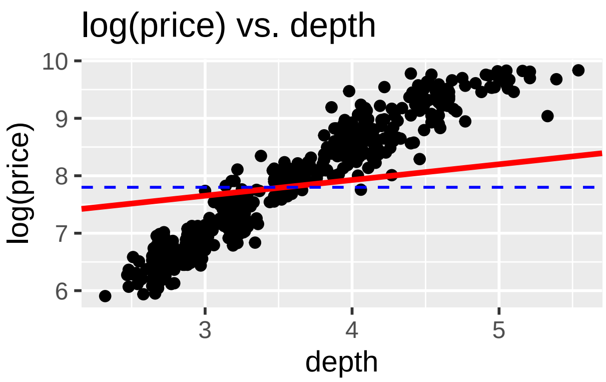
The plots below represent the JHM model fit for each quantitative variable. The red line presents the slope for that particular predictor in the presence of others. The blue dashed line indicates the mean of log(`price`). We noticed that all four variables follow a similar trend. This is potentially indicative of multicollinearity.



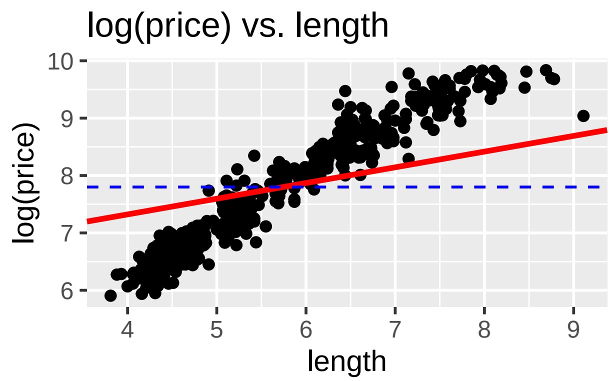
Plot 5a



Plot 5b

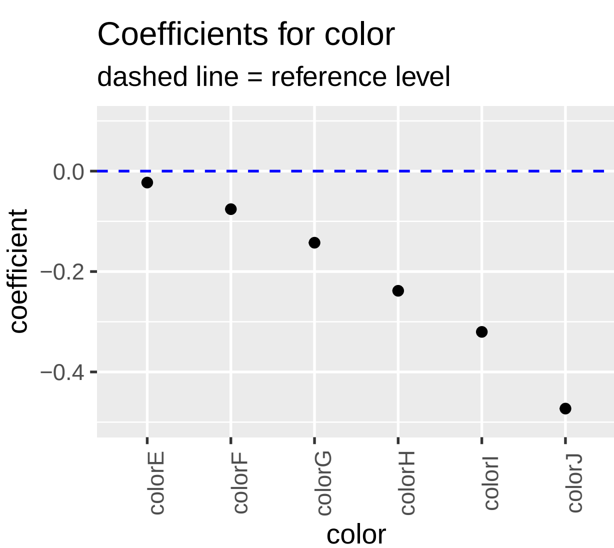


Plot 5c

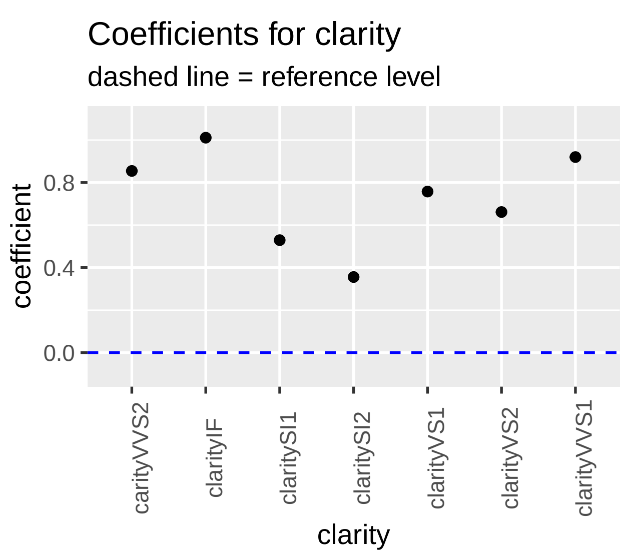


Plot 5d

The plots below represent each categorical variable in the JHM model. The height of each point represents the coefficient (β value) for that level of the categorical variable. The blue dashed line represents the baseline indicator level for each variable.



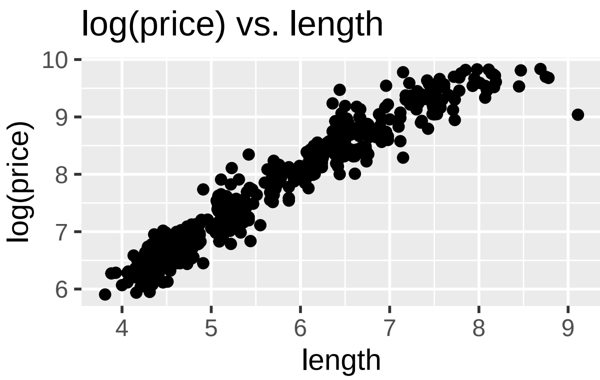
Plot 6a



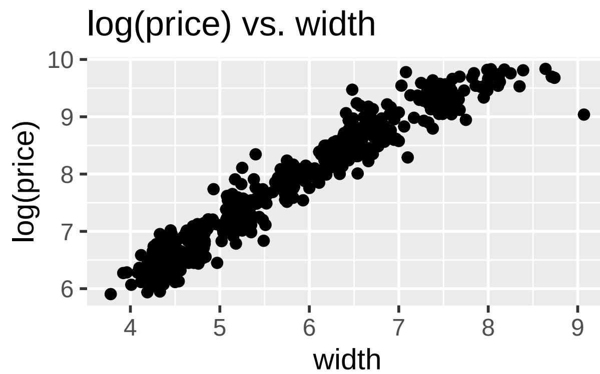
Plot 6b

GAM

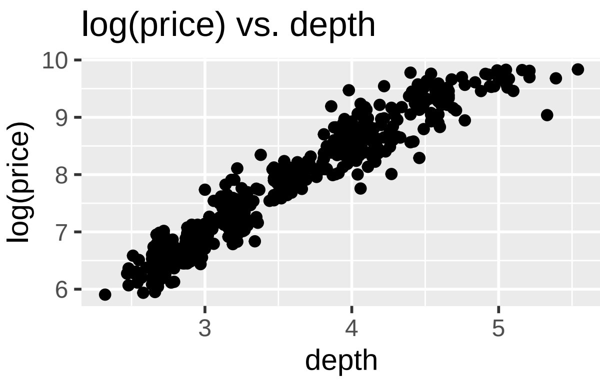
We were also interested in examining how well the same set of predictors estimate price in a Gereralized Additive Model (GAM). First, we examined the quantitative predictors (`length`, `width`, `depth` and `carat`) for any obvious relationships with log(`price`).



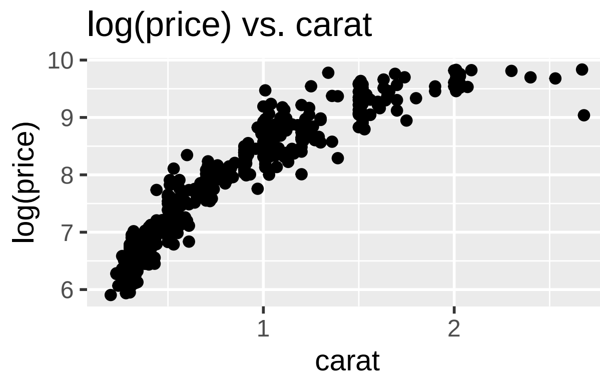
Plot 7a



Plot 7b



Plot 7c



Plot 7d

We noticed that length, width and depth all appear to follow a similar pattern. Because of this, we hypothesized that a smoother would most likely not be necessary for all three of these predictors. We also noted that `carat` is most likely co-linear with `length`, `width` and `depth`, so we noted that we may be able to explain carat's variability using other predictors.

Our first step in building the GAM incorporated comparing a SLR and a smoothing spline for each quantitative predictor. We utilized adjusted R^2 as a metric for model fit (calculated using the function below).

```r
#function for calculating adjusted r-squared for gam
gam_adjusted <- function(model){
  rsq_gam = 1 - model$deviance/model$null.deviance
  adjrsq_gam = 1 - (1 - rsq_gam)*(model$df.null/model$df.residual)
  return(adjrsq_gam)
}
```

A summary of the adjusted R^2 values for each model type can be seen in Table 4 below.

Table 4: SLR vs. smooth R^2_{adj}

Predictor	SLR	Smooth
width	0.931	0.943
length	0.929	0.942
carat	0.856	0.941
depth	0.87	0.935

As seen in Table 4, for all 4 of the quantitative predictors, the R^2_{adj} for the smoother was greater than the

SLR.

We next built several potential GAMs for these 4 quantitative predictors. We included both categorical variables (clarity and color) for all models. The following is a summary and rationale for each model:

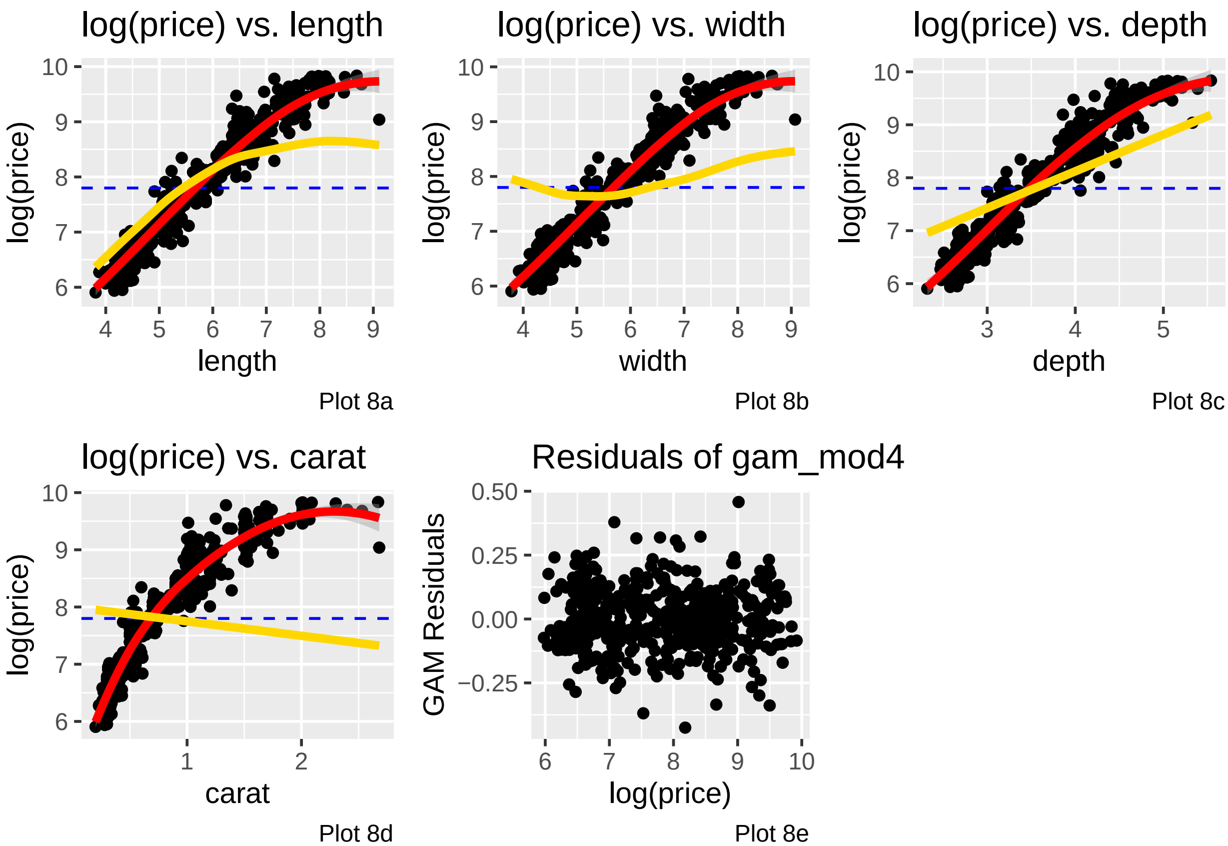
1. gam_full: a smoothing spline on a 4 quantitative predictors. We started with this GAM as our tests comparing smoothing to SLR suggested smoothing was as better metric for each variable.
2. gam_mod2: a smoothing spline on `carat` and `depth` and linear relationships for `length` and `width`. We chose to make `length` and `width` linear because a plot of their relationship showed a linear relationship with `log(price)`. We recongized that using smoothers on all 4 predictors in a GAM tends to lead to overfitting.
3. gam_mod3: a smoothing spline on `width` and linear relationship for `depth` and `carat.` We noticed that, in the presence of other predictors, `carat` was essentially linear. Switching `carat` from a smoothing spline to linear relationship would decrease the degrees of freedom and could improve the performance of our model. In this model, we also tried removing `length` as it seemed colinear with `width, depth` and `carat.`
4. gam_mod4: linear predictors on all variables. The general trend of the data appeared to be very close to linear for all predictors.
5. gam_mod5: smoothers on `length` and `width` and linear relationships for `depth` and `carat.` We decided to include all 4 quantitative variables in this model as removing one resulted in an increase in AIC.

A summary of each GAM's performance is summarized in table 5, below.

Table 5: Comparison of AIC between GAM models

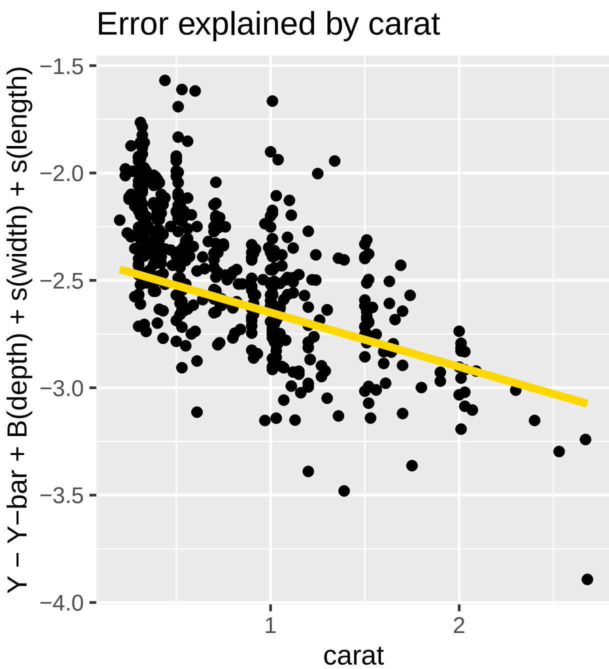
Model	AIC
color + clarity + s(length) + s(width) + s(depth) + s(carat)	-649.94
color + clarity + length + width + s(depth) + s(carat)	-612.54
color + clarity + depth + carat + s(width)	-596.29
color + clarity + depth + carat + width + length	-545.24
color + clarity + depth + carat + s(width) + s(length)	-652.97

As seen in table 5, `gam_mod5` had the lowest AIC. Plots of this GAM's performance in comparison to a smoothing spine are below.

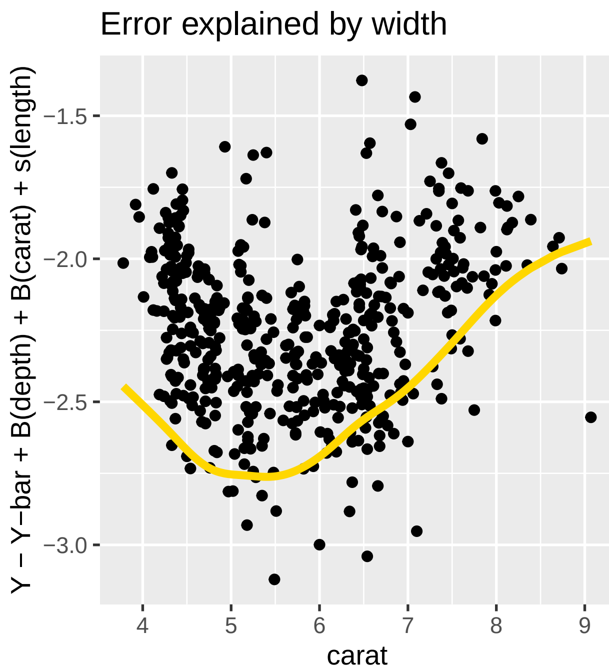


Explaining the roles of `carat` and `width` in the model

As seen in plots 8b and 8d, carat and width have very peculiar GAM lines. We hypothesized that these lines explain the variability and error introduced into the model by the other predictors. When examining `carat`, a plot of the residual error of the model with `depth`, the smoother on `width`, and the smoother on `length` against `carat` demonstrates that the GAM for carat directly explains the variability of the model caused by the 3 remaining quantitative variables. This is evident in plot 9a. Similarly, when examining `width`, a plot of the residual error of the model with `depth`, `carat`, and the smoother on `length` shows that the GAM for `width` also explains the variability of the model caused by the other 3 quantitative predictors. This is depicted on plot 9b.



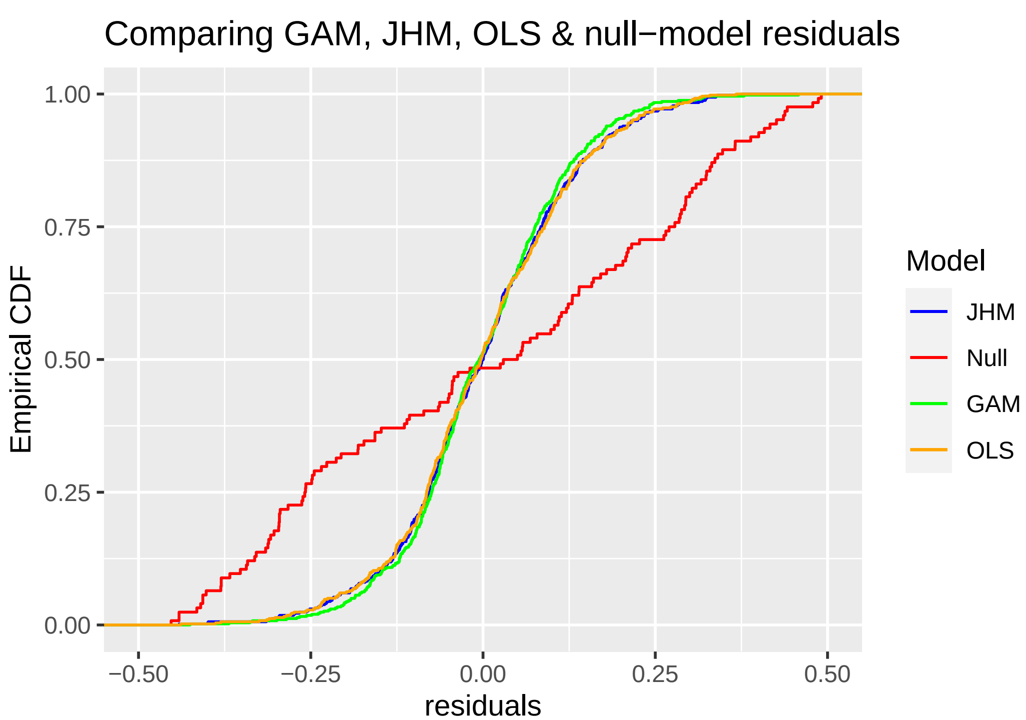
Plot 9a



plot 9b

Discussion

Proof of our models: examining residuals



plot 10

As seen in the plot 10 above, the disribution of the GAM, OLS and JHM residuals are all very similar. The null is clearly performing the worst as the eCDF of the residuals is clearly not close to matching a normal

CDF. The more vertical the eCDF, the more variability explained in the model (which is indicative of a more efffective model). Because we already found that using a parametric approach is not possible (the residuals are not normally distributed), we will focus primarily on the GAM and JHM.

As seen in plot 11, while both the JHM and GAM residuals follow a fairly normal distribution, the JHM is slighly less vertical than the GAM. This suggests that the GAM might be a slightly better model.

Using cross validation to assess model fit

We also used a cross-validation appraoch to estimate the adjusted R^2, and L1-proportion for each of the our 3 models (OLS, JHM, GAM). Table 6 below depicts the R^2, R^2_{adj} and L1-proportion on the original data as well as R^2, R^2_{adj} and L1-proportion using cross validation.

Table 6: Results from cross-validation

	Regular approach			Cross-validation approach		
	R^2	R^2_{adj}	$L1_{\text{prop}}$	R^2	R^2_{adj}	$L1_{\text{prop}}$
OLS	0.9833	0.9827	0.8842	0.9816	0.9809	0.8789
JHM	0.9832	0.9826	0.8852	0.9817	0.981	0.8796
GAM	0.9828	0.9822	0.8817	0.9847	0.9838	0.8892

Because the 3 models were generated from the same dataaset on which the 3 measures of fit (R^2, R^2_{adj} and L1-proportion) were calculated, the non-cross validation approach is potentially overfit to the data and the values are not as accurate. Cross validation, however, attempts to guard against overfitting. Thus, the values of the right side of Table 7 are how we will compare models.

All 3 models have relatively similar R^2_{adj} values. The GAM, however, has an R^2_{adj} of 0.9838, which is slightly higher than the OLS and JHM. In examining the R^2 values, OLS fails to explain 1.84% of the variability, while GAM fails to explain 1.53% of the variability. Thus, using the GAM model over OLS results in a 16% decrease in unexplained variability. It is important to note that we cannot fairly use R^2 for JHM because of its ability to “ignore” outliers. The L1 proportion, however, denotes the absolute value of the residuals over the absolute value of the deviation of the mean (i.e. proportion of error explained by the model). Notably, the L1 proportion value is also higher for the GAM. Based on our cross-validation results, the GAM is the best-performing model.

Limitations and Challenges

There are two main limitations in our work. Our main statistical concern was the high level of multi-colinearity between `carat`, `length`, `width`, and `depth`. This overlap between predictors made it difficult to create a model that didn’t incorporate variables that explained the same variation in log(`price`). This was most evident in creating our GAM. We were able to include SLR fits on two of these inter-related variables as a smoother was really only necessary on one - it explained the slight non-linear variability of the 3 remaining co-linear variables.

Our second concern arose from the dataset itself. We did not know what year (or group of years) the diamonds from within the dataset arose. Because of this, we will not be adjust for inflation when predicting diamond price in alternative years.

Overall, the main challenge we faced was dealing with the multicolinearity between our 4 main quantitative predictors. We found it difficult to create models using our standard techniques as the results of our models were almost always riddled with impacts of multicollinearity. When building our GAM, for example, while we knew `length, width, depth and carat` were likely co-linear, removing any of them from the model resulted in an increase in AIC (indicating we couldn’t remove them). Thus, our main challenge was finding ways to balance between using multicollinear variables and creating highly predictive models.

Conclusion

We suggest using a generalized additive model to predicting diamond price. We cannot use a parametric appraoch (i.e. OLS MLR) because the residuals of such a model are not normally distributed (as seen in the K-S test). The JHM model had a lower L1-proportion (from cross validation) when compared to the GAM. Because of this, we propose using a GAM to predict log(`price`) from `depth, length, width, carat, color, and clarity` (with a smoothing spline on length and width).