# STAT225: Group 2 Final Project

## Predicting Price of a Diamond Using its Defining Attributes

Dasha, Braedon, and Isabelle

05/19/2022

# Exploratory Data Analysis

# Data

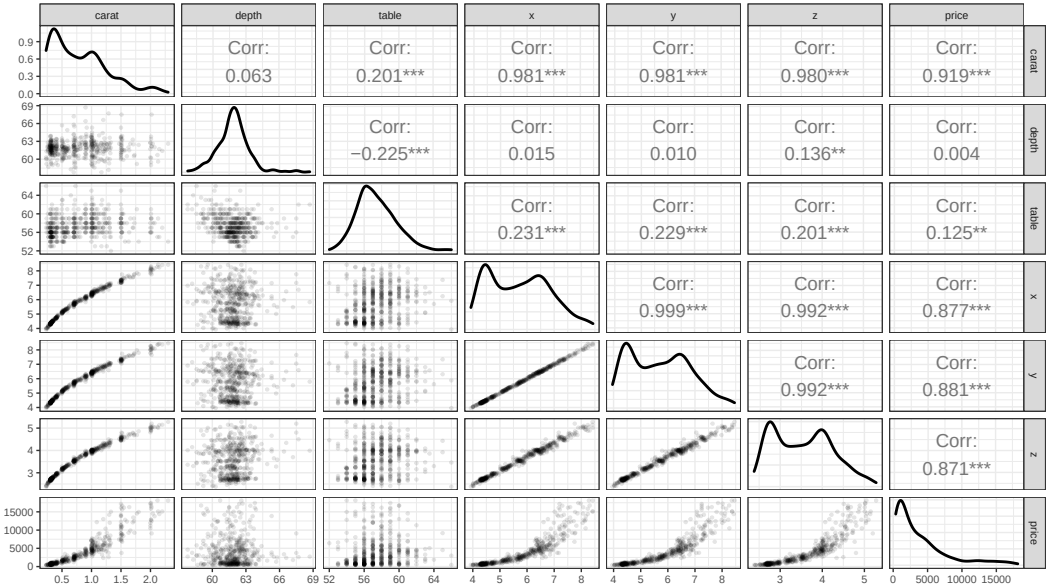
Response Variable: Price
Quantitative Predictors:

- `carat` – weight of the diamond
- `table` – width at the top of the diamond relative to its widest width
- `x`, `y`, `z` – length, width, and depth of the diamond
- `depth` – total depth percentage as a function of x, y, and z

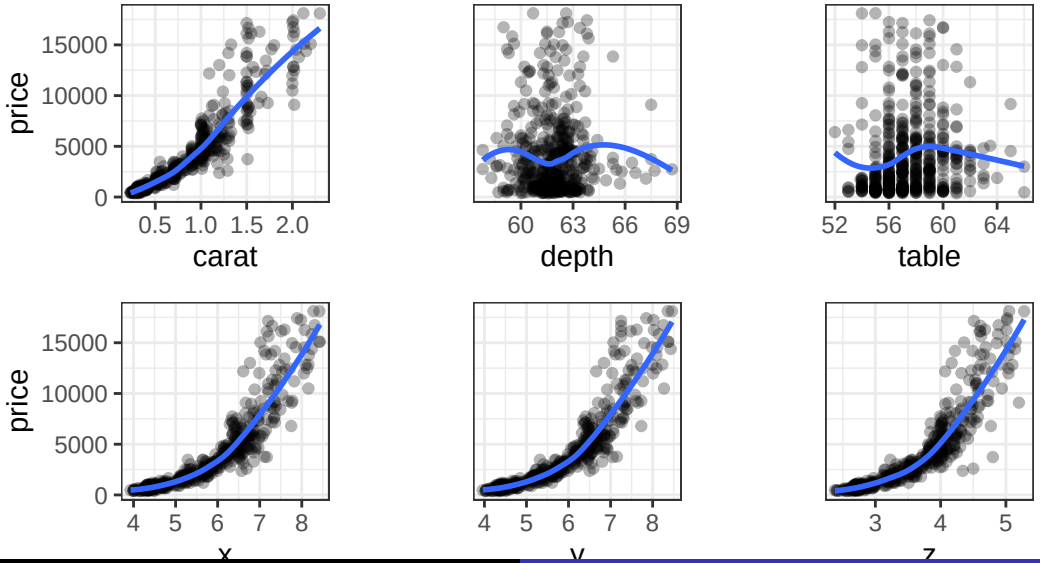
**Package**: ggplot2

**Observations**: 53,940 diamonds

**Sample for Analysis**: Random sample of 500 diamonds


carat
depth
table
x
y
z
price
Corr: 0.063
Corr: 0.201***
Corr: 0.981***
Corr: 0.981***
Corr: 0.980***
Corr: 0.919***
Corr: −0.225***
Corr: 0.015
Corr: 0.010
Corr: 0.136**
Corr: 0.004
Corr: 0.231***
Corr: 0.229***
Corr: 0.201***
Corr: 0.125**
Corr: 0.999***
Corr: 0.992***
Corr: 0.877***
Corr: 0.992***
Corr: 0.881***
Corr: 0.871***

# Response Variable Against Continuous Predictors



# Spearman's Correlation Test

$H_o : \rho_s = 0$
$H_a : \rho_s \neq 0$
Assumptions:

- Data is continuous and observations are independent.

Conclusions:

- Reject the null hypothesis in all cases except the test for `depth` and `price`.
- There is a significant correlation between `price` and `carat`, `table`, `x`, `y`, `z`.
  - `table` has the weakest correlation among the 5 significant correlations.
  - `y` has the strongest correlation among the 5 significant correlations.

| | Variable | Estimate | P-Value |
|---|---|---|---|
| rho | carat | 0.9636 | 0 |
| rho | depth | 0.0357 | 0.4253 |
| rho | table | 0.251 | 0 |
| rho | x | 0.9651 | 0 |
| rho | y | 0.9657 | 0 |
| rho | z | 0.9594 | 0 |

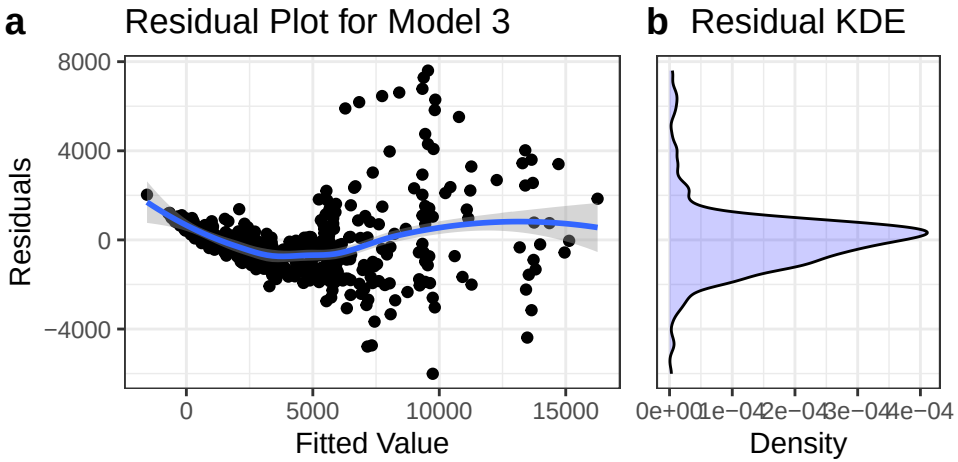
# Ordinary Least Squares Regression Models

# Final Multiple Linear Regression Models

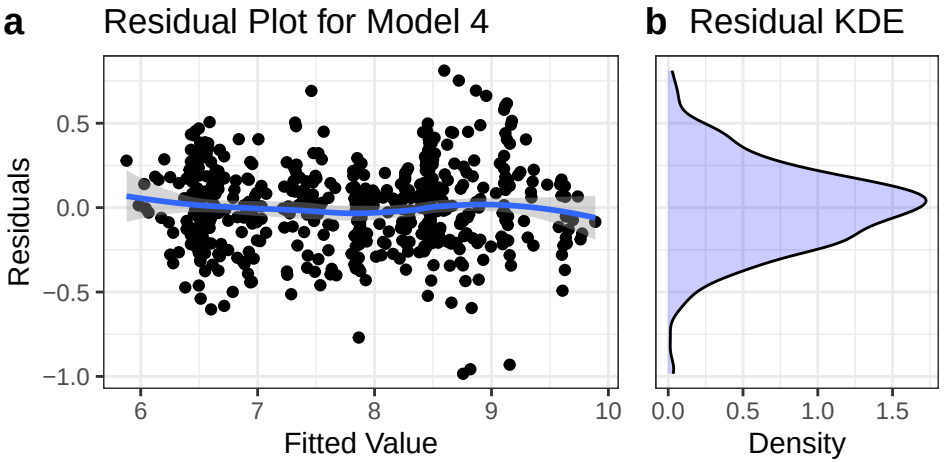
The Models we explored:

| | Predictors | $R^2$ | $R^2_{adj}$ | max($VIF$) | df | AIC |
|---|---|---|---|---|---|---|
| model1 | carat + x + y + z + depth + tabl ... | 0.87 | 0.87 | 2278.24 | 8 | 8684.67 |
| model2 | carat + x + y + z + table | 0.87 | 0.87 | 566.87 | 7 | 8682.84 |
| model3 | carat + table | 0.85 | 0.85 | 1.04 | 4 | 8771.51 |
| model4 | log(carat) + log(table) | 0.94 | 0.94 | 1.06 | 4 | 52.26 |

# Why Use the Log Function?



# Modified Log Function



# KS Test

For Model 3:

$$H_0 : F(t) = Z \sim N(0, 1549.317)$$

$$H_A : F(t) \neq Z \sim N(0, 1549.317)$$

| statistic | p.value | method | alternative |
|---|---|---|---|
| 0.1513382 | 0 | One-sample Kolmogorov-Smirnov test | two-sided |

Therefore there is significant evidence to suggest that the KDE of the residuals of Model 3 does not fit the normal distribution

# KS Test Continued

For Model 4:

$$H_0 : F(t) = Z \sim N(0, 0.2531749)$$

$$H_A : F(t) \neq Z \sim N(0, 0.2531749)$$

| statistic | p.value | method | alternative |
|---|---|---|---|
| 0.0324261 | 0.6691997 | One-sample Kolmogorov-Smirnov test | two-sided |

Therefore there is no significant evidence to suggest that the KDE of the residuals of Model 4 does not fit the normal distribution.

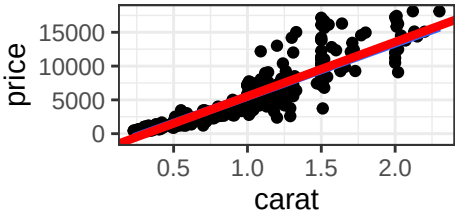
# Rank-Based Regression Models

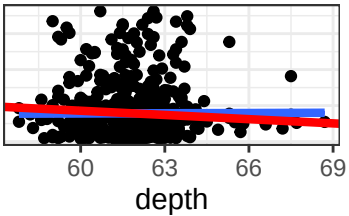
# JHM Regression Models

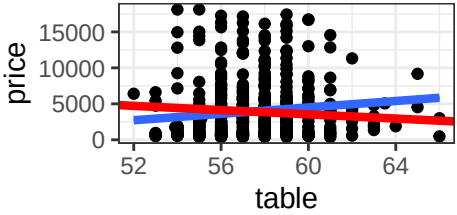
- Full Model:
  `price ~ carat + depth + table + x + y + z`
- Best Model at Predicting Response (Diamonds1):
  `price ~ carat + depth + table + y + z`
- Best Model at Explaining Relationships (Diamonds3):
  `price ~ carat + depth + table`

| Full Model | Reduced Model | F-Statistic | P-Value |
|---|---|---|---|
| DiamondsFull | Diamonds1 | 1.97 | 0.16 |
| Diamonds1 | Diamonds3 | 120.13 | 0 |

# Visualizing Models



# General Additive Models

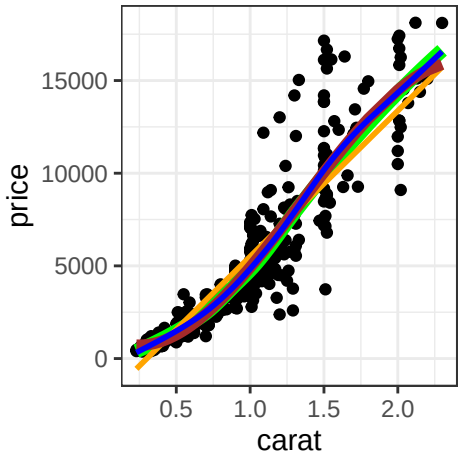
# Final GAM Models

Our GAM models:

- Full Model:
  `price ~ s(carat) +s(depth) + s(x) + s(y) + s(z) + lo(table)`
- Final Models:
  - Most Predictive Model:
    `price ~ s(carat) + s(depth) + s(x) + s(y) + s(table)`
  - Most Explainable and Simplest Model: `price ~ s(carat) + s(depth)`

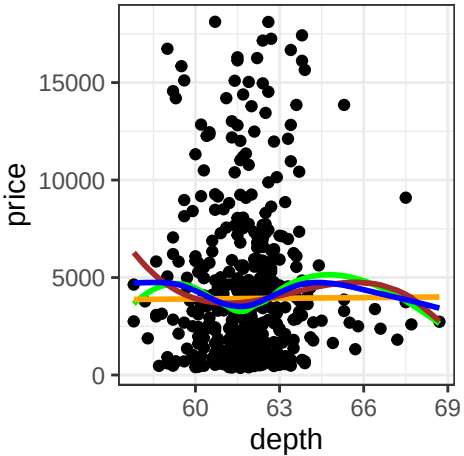
# Examining `carat`

| | CaratModels | df | AIC |
|---|---|---|---|
| caratmod1 | price ~ carat | 3 | 8781.3 |
| caratmod3 | price ~ bs(carat) | 5 | 8702.3 |
| caratmod2 | price ~ lo(carat) | 3 | 8696.8 |
| caratmod4 | price ~ s(carat) | 3 | 8695.9 |



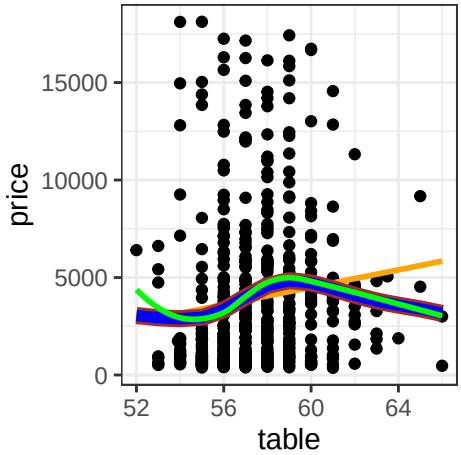
# Examining `depth`

| | DepthModels | df | AIC |
|---|---|---|---|
| depthmod3 | price ~ bs(depth) | 5 | 9711.3 |
| depthmod1 | price ~ depth | 3 | 9710.6 |
| depthmod2 | price ~ lo(depth) | 3 | 9707.4 |
| depthmod4 | price ~ s(depth) | 3 | 9707.3 |



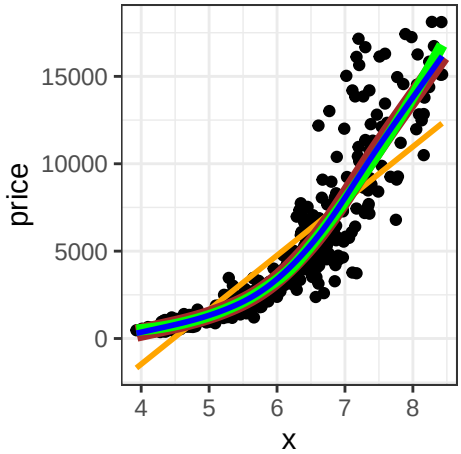
# Examining `table`

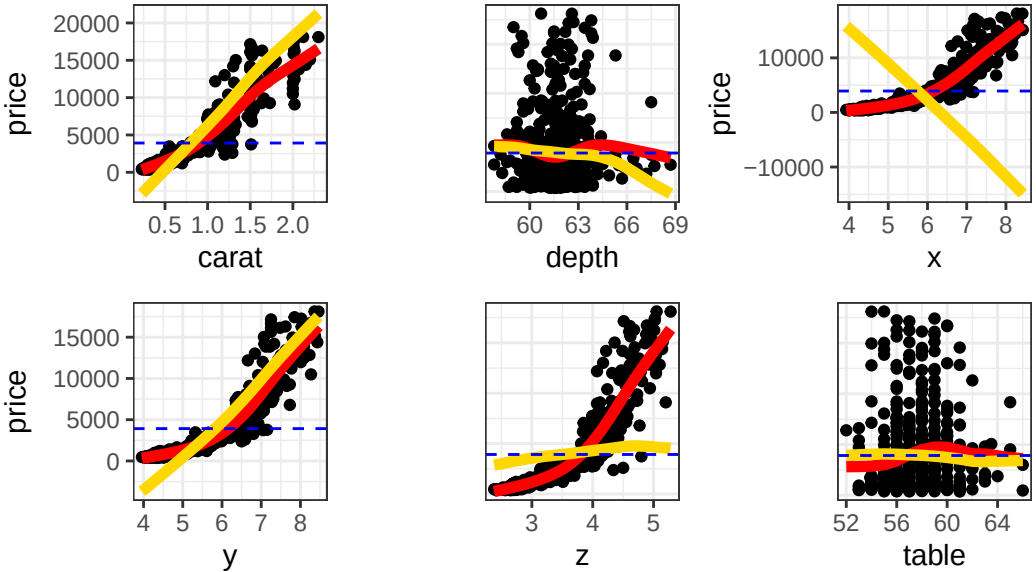
| | TableModels | df | AIC |
|---|---|---|---|
| tablemod1 | price ~ table | 3 | 9702.7 |
| tablemod3 | price ~ bs(table) | 5 | 9700.8 |
| tablemod4 | price ~ s(table) | 3 | 9698.0 |
| tablemod2 | price ~ lo(table) | 3 | 9697.9 |



# Examining x, y, and z

| | Models | df | AIC |
|---|---|---|---|
| xmod1 | price ~ x | 3 | 8979.2 |
| xmod3 | price ~ bs(x) | 5 | 8734.3 |
| xmod2 | price ~ lo(x) | 3 | 8723.6 |
| xmod4 | price ~ s(x) | 3 | 8721.9 |




price
20000
15000
10000
5000
0
0.5 1.0 1.5 2.0
carat
60 63 66 69
depth
price
10000
0
−10000
4 5 6 7 8
x
price
15000
10000
5000
0
4 5 6 7 8
y
3 4 5
z
52 56 60 64
table

# Most Predictive Model

Using the best smooths for each of the predictors, we fit a full model:

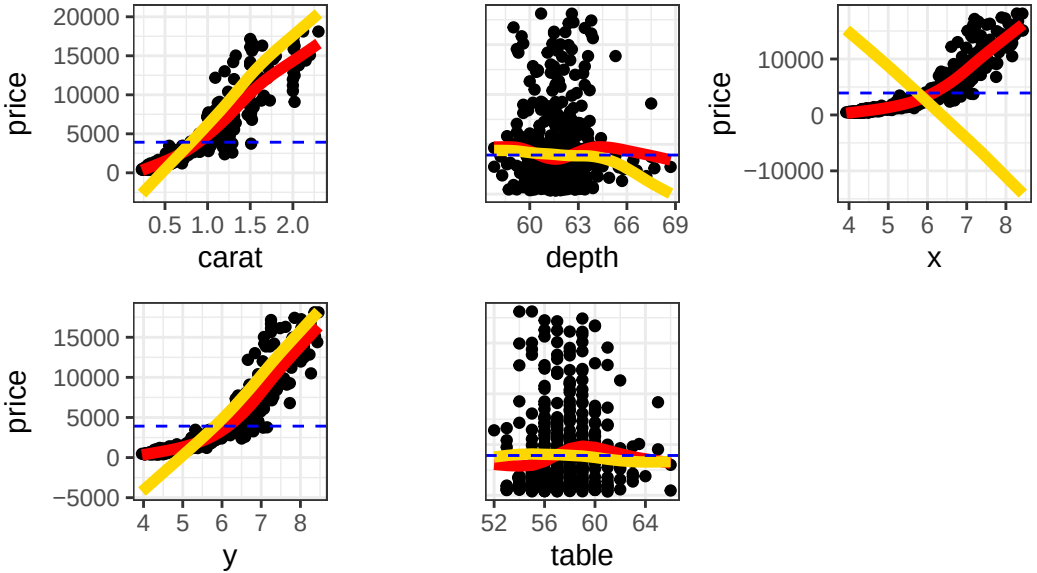
- AIC: 8661.414.
- `s(z)` is not significant in the presence of the other predictors regarding parametric effects.
- Significant non-parametric terms:
  - `s(carat)`
  - `s(depth)`

We continued to fit more models based on this information.

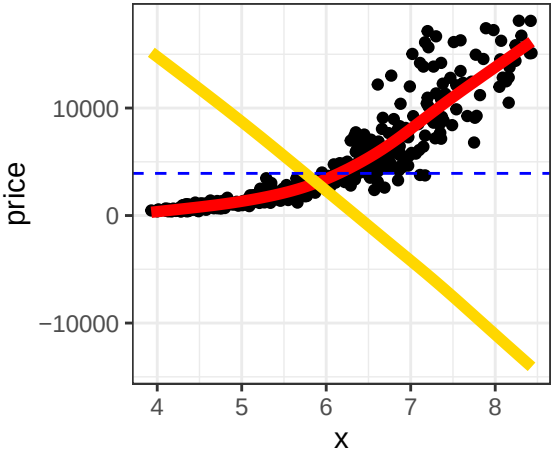
- Total Models Fitted: 18

Most Predictive Model:

- `price ~ s(carat) + s(depth) + s(x) + s(y) + s(table)`
- AIC: 8655.444
- All terms have a smoothing spline .


price
20000
15000
10000
5000
0
0.5 1.0 1.5 2.0
carat
60 63 66 69
depth
price
10000
0
−10000
4 5 6 7 8
x
price
15000
10000
5000
0
−5000
4 5 6 7 8
y
52 56 60 64
table

# Checking VIF



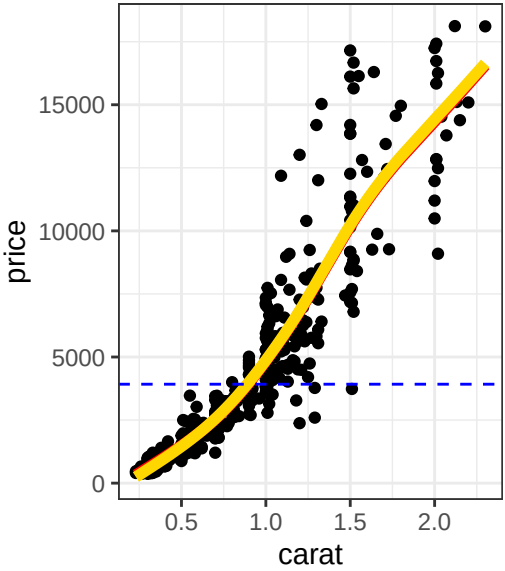

| | x |
|---|---|
| s(carat) | 28.662 |
| s(depth) | 1.139 |
| s(x) | 530.346 |
| s(y) | 539.075 |
| s(table) | 1.131 |

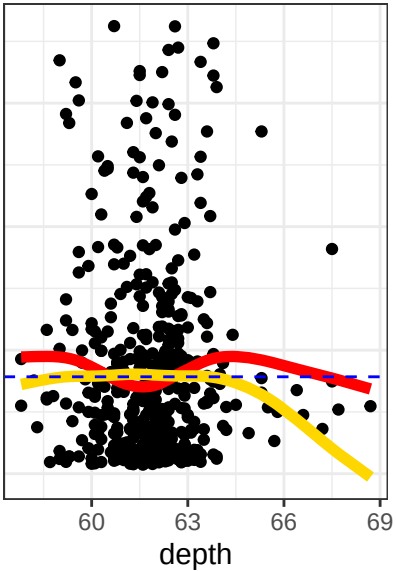
# Fitting a Simpler and More Explainable Model

We proceeded to fit simpler models that did not have issues of multi-collinearity.
Models fitted: 10
Model with lowest AIC:

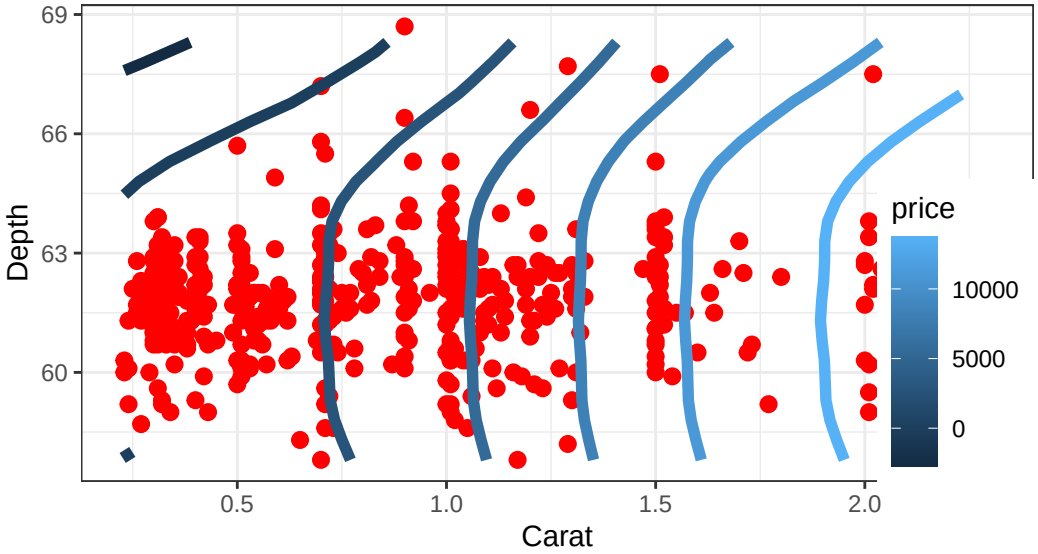
- `price ~ s(carat) + s(depth)`
  - AIC: 8664.078.
  - Significant non-parametric terms:
    - `s(carat)`
    - `s(depth)`

| | x |
|---|---|
| s(carat) | 1.003935 |
| s(depth) | 1.003935 |

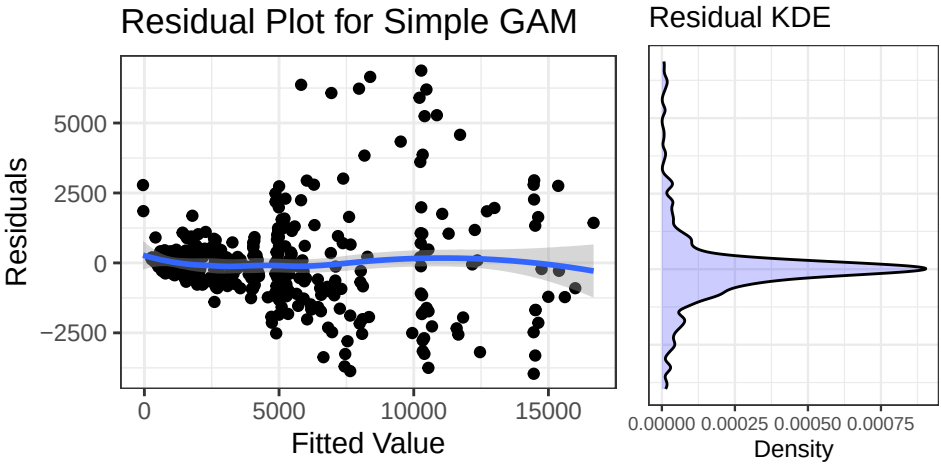

15000
10000
5000
0
price
0.5
1.0
1.5
2.0
carat


60
63
66
69
depth

# Surface Plot



# Checking GAM Residuals



# Performing Cross-Validation

# Most Predictive Models

OLS:

```
price ~ carat + x + y + z + table
```

JHM:

```
price ~ carat + depth + table
+ y + z
```

GAM:

```
price ~ s(carat) + s(depth)
+ s(x) + s(y) + s(table)
```

Table 2: Regular Fit

| R Squared | Adj R Squared | Prop L1 |
|---|---|---|
| 0.8740 | 0.8727 | 0.7187 |
| 0.8643 | 0.8629 | 0.7364 |
| 0.8877 | 0.8830 | 0.7286 |

Table 3: Cross Validation

| | R Squared | Adj R Squared | Prop L1 |
|---|---|---|---|
| OLS | 0.8740 | 0.8727 | 0.7187 |
| JHM | 0.8610 | 0.8596 | 0.7307 |
| GAM | 0.8739 | 0.8686 | 0.7139 |

# Simplest Models

OLS:

```
price ~ carat + table
```

JHM:

```
price ~ carat + depth + table
```

GAM:

```
price ~ s(carat) + s(depth)
```

Table 4: Regular Fit

| R Squared | Adj R Squared | Prop L1 |
|---|---|---|
| 0.8477 | 0.8471 | 0.6582 |
| 0.8411 | 0.8401 | 0.6891 |
| 0.8801 | 0.8781 | 0.7346 |

Table 5: Cross Validation

| | R Squared | Adj R Squared | Prop L1 |
|---|---|---|---|
| OLS | 0.8477 | 0.8471 | 0.6582 |
| JHM | 0.8387 | 0.8377 | 0.6860 |
| GAM | 0.8704 | 0.8683 | 0.7247 |

# References

Wickham, H. (2016), *Ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York. https://ggplot2.tidyverse.org.