STAT-225 Group 8 Final Project Presentation

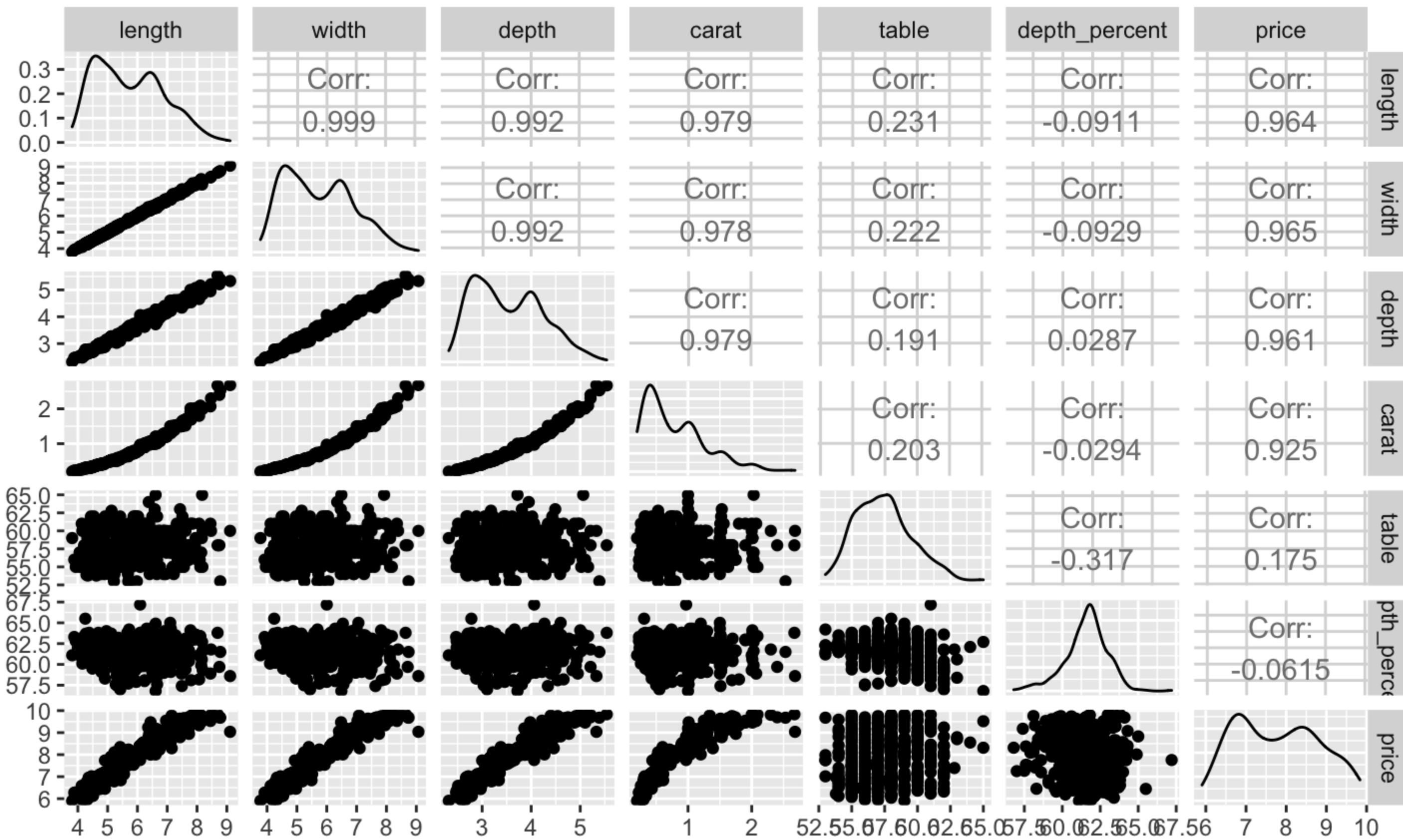
Investigating Diamond Price

Anna Ballou, Nicole Frontero, Alex Russell
5/3/2020

Introduction

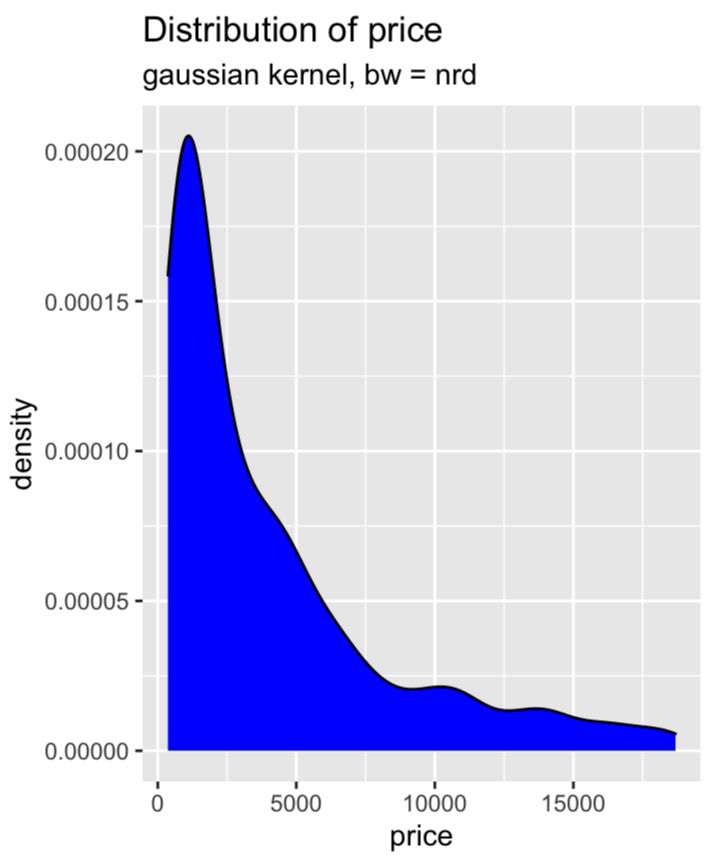
- Dataset: `diamonds`
- Random sample: (500 observations from 54,000)
- The observational unit: a diamond
- Response variable: price in US dollars
- Explanatory variables: `carat + cut + color + clarity + depth + table + x + y + z + depth_percent`
- Note: `x` = length; `y` = width; `z` = depth; `carat` = mass; `table` = width of top of diamond

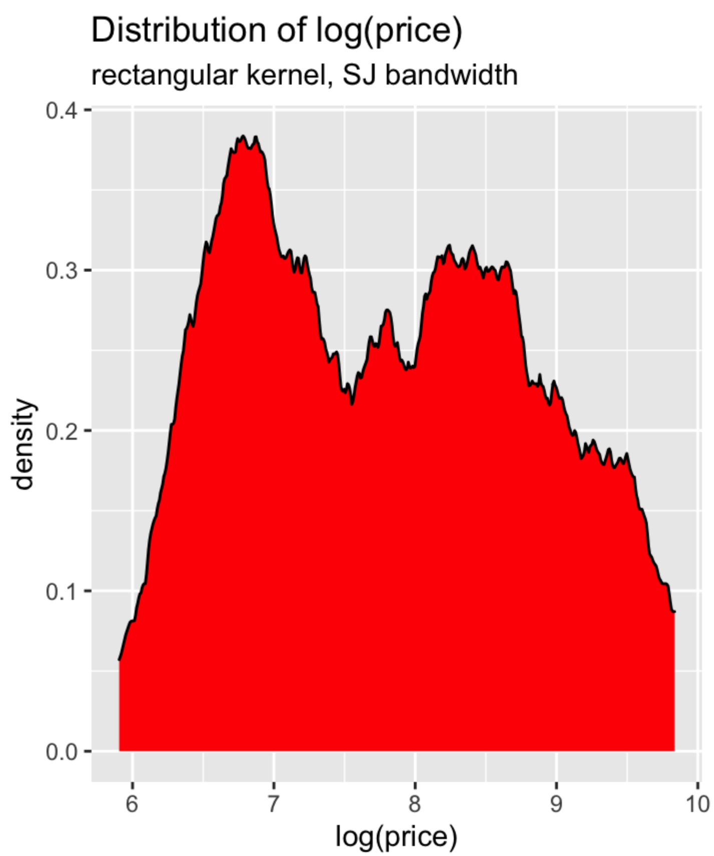
EDA: `ggpairs`



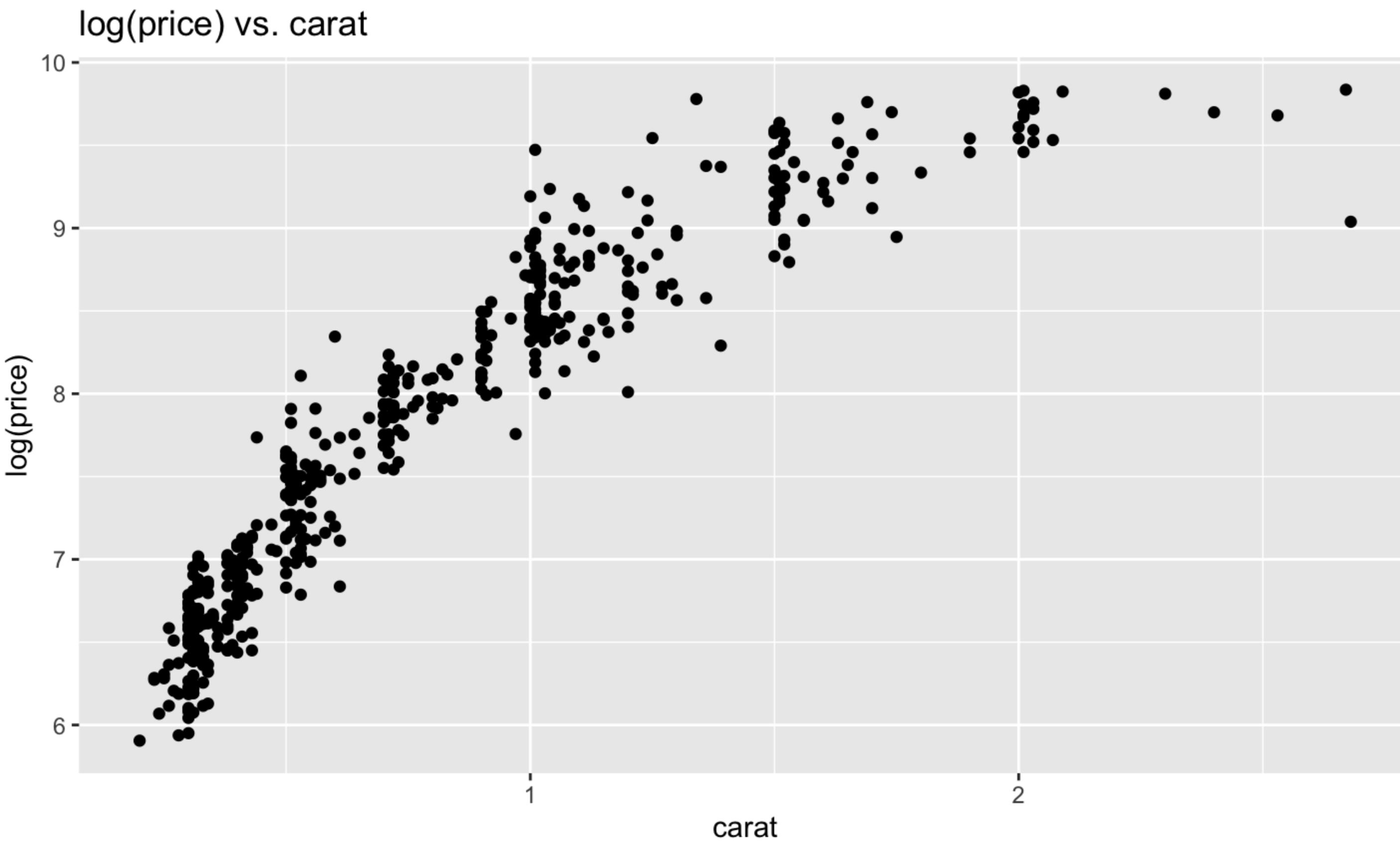
length
width
depth
carat
table
depth_percent
price
Corr: 0.999
Corr: 0.992
Corr: 0.979
Corr: 0.231
Corr: -0.0911
Corr: 0.964
Corr: 0.992
Corr: 0.978
Corr: 0.222
Corr: -0.0929
Corr: 0.965
Corr: 0.979
Corr: 0.191
Corr: 0.0287
Corr: 0.961
Corr: 0.203
Corr: -0.0294
Corr: 0.925
Corr: -0.317
Corr: 0.175
Corr: -0.0615

Response variable: `price`





Spearman’s test for correlation on `carat` vs. `log(price)`



OLS model: forward stepwise regression

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

- Note that forward stepwise regression excluded `depth_percent` and `table`.

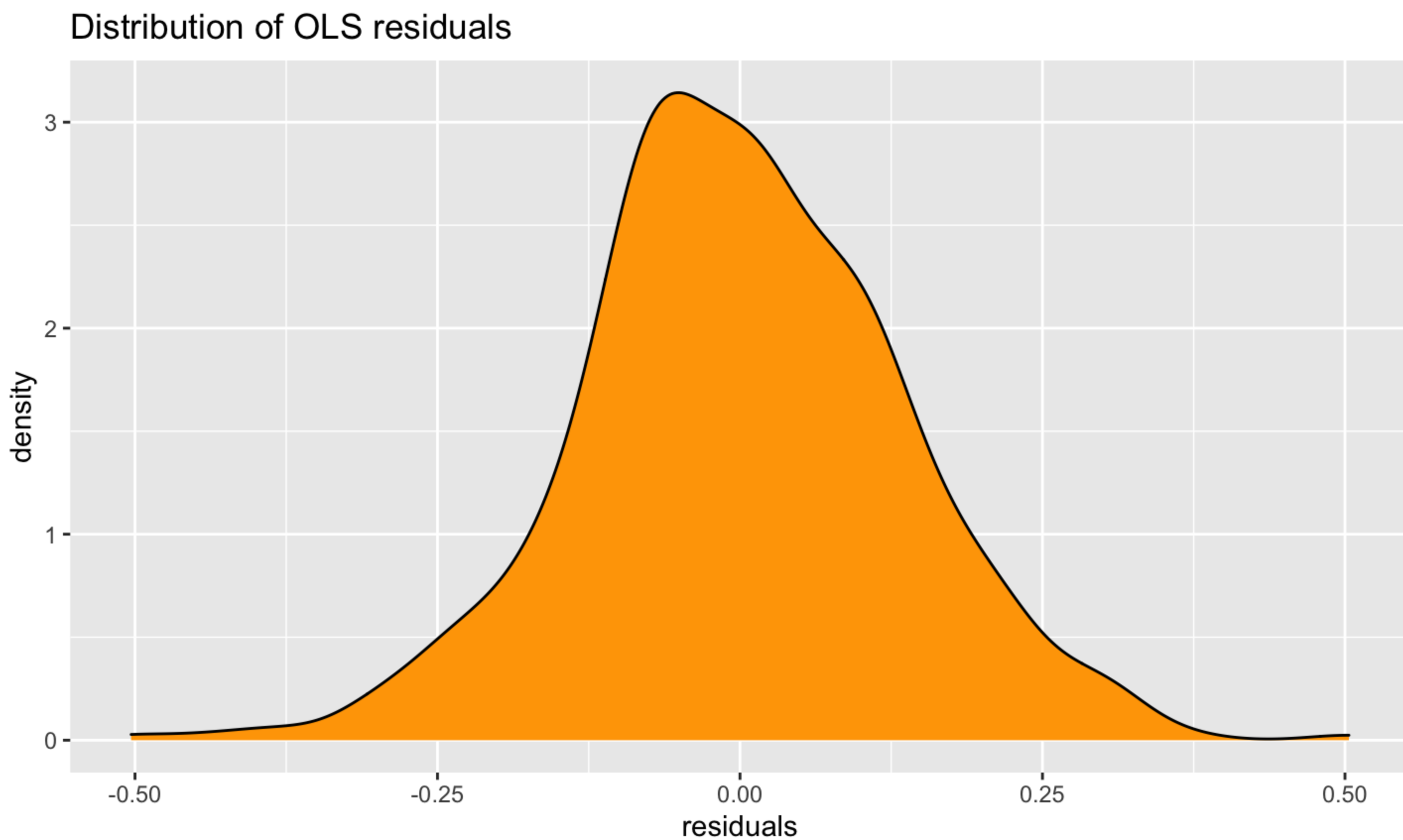
OLS model: final model summary

Table 2: OLS model summary results

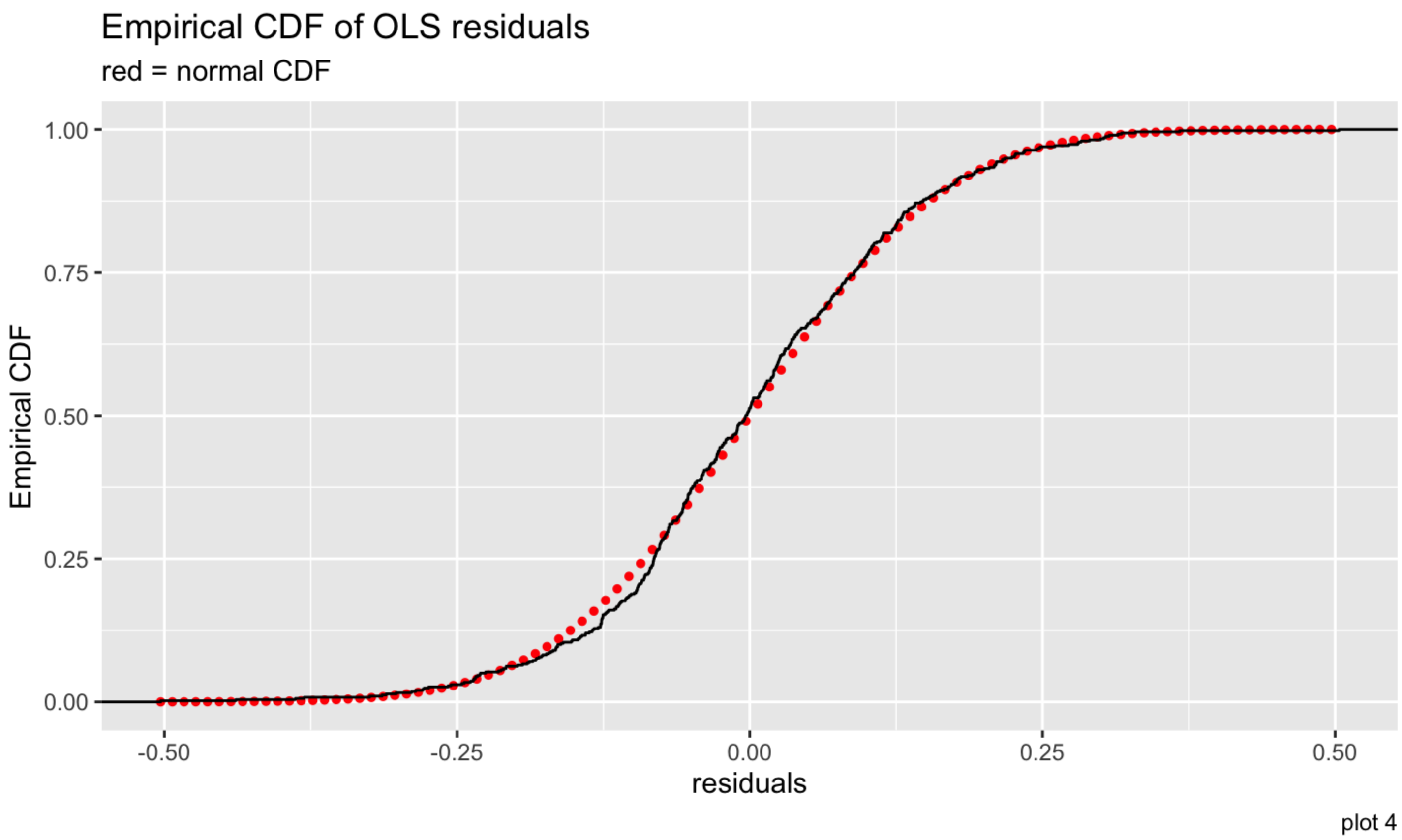
log(price) ~ width + clarity + color + carat + depth + length					
Predictors	**Estimate**	**P-value**	**Predictors**	**Estimate**	**P-value**
(Intercept)	-0.0651831	0.59156	colorE	-0.0227264	0.30612
width	0.2949006	0.01319	colorF	-0.0803849	0.00034
clarityIF	1.0487289	< 0.0001	colorG	-0.1564346	< 0.0001
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

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

OLS or JHM? Visually examine OLS residuals for normality



OLS or JHM? Test OLS residuals for normality (KS)



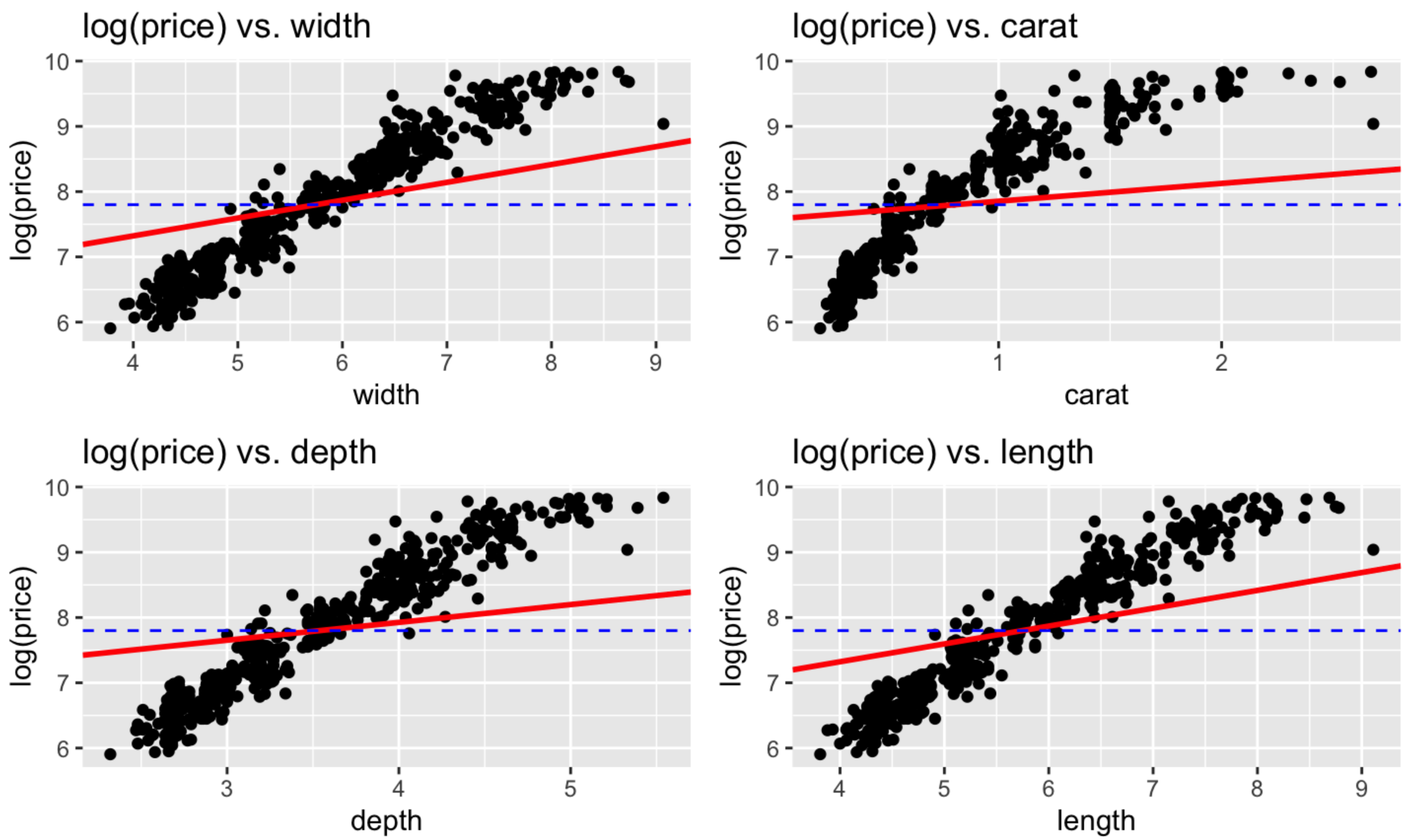
Building a JHM model

Table 3: JHM model summary results

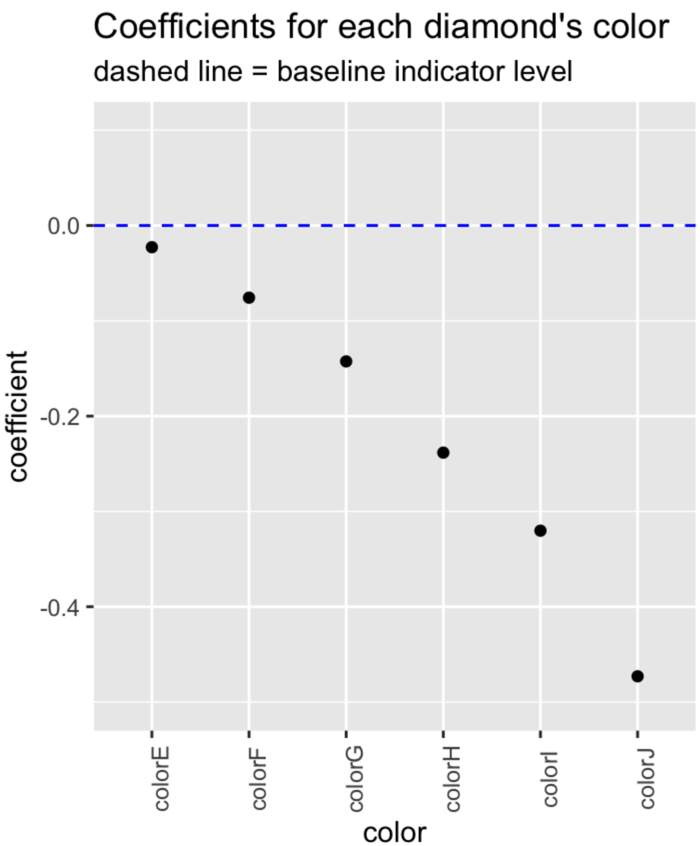
log(price) ~ width + clarity + color + carat + depth + length					
Predictors	**Estimate**	**P-value**	**Predictors**	**Estimate**	**P-value**
(Intercept)	0.0090945	0.93893	colorE	-0.0227499	0.29299
width	0.2735470	0.01848	colorF	-0.0757285	0.00054
clarityIF	1.0108864	< 0.0001	colorG	-0.1425860	< 0.0001
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

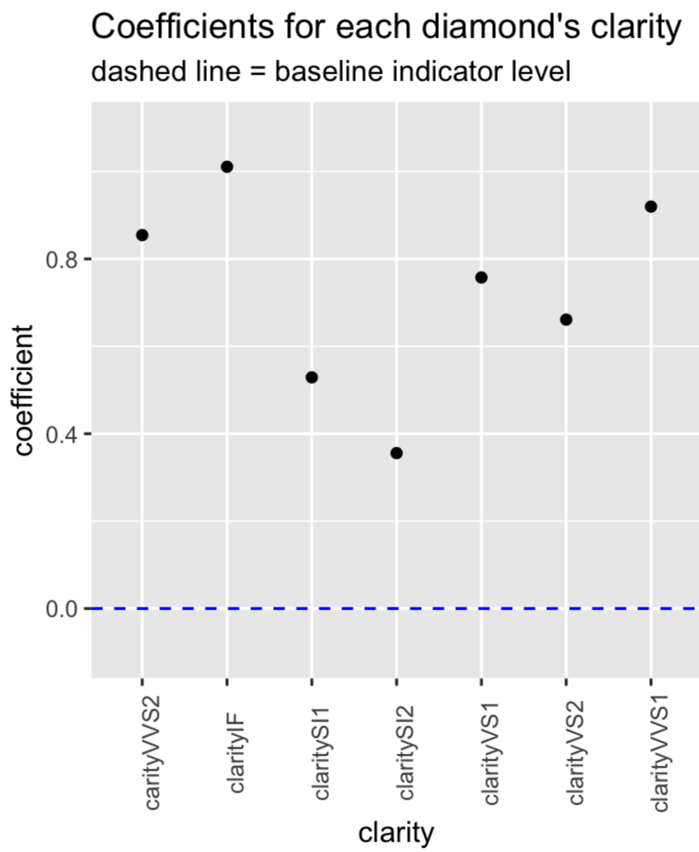
$$R^2_{\mathrm{adj}} = 0.9811$$

Plotting JHM model - quantitative predictors



Plotting JHM model - categorical predictors





GAM: smoother or SLR for quantitative predictors?

- For each quantitative predictor, we created two models - a GAM with a s-spline smoother, and a SLR - and compared their AICs.

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

- Note: R^2_{adj} for smoothing spline was higher for all quantitative predictors.

GAM: choosing the best model

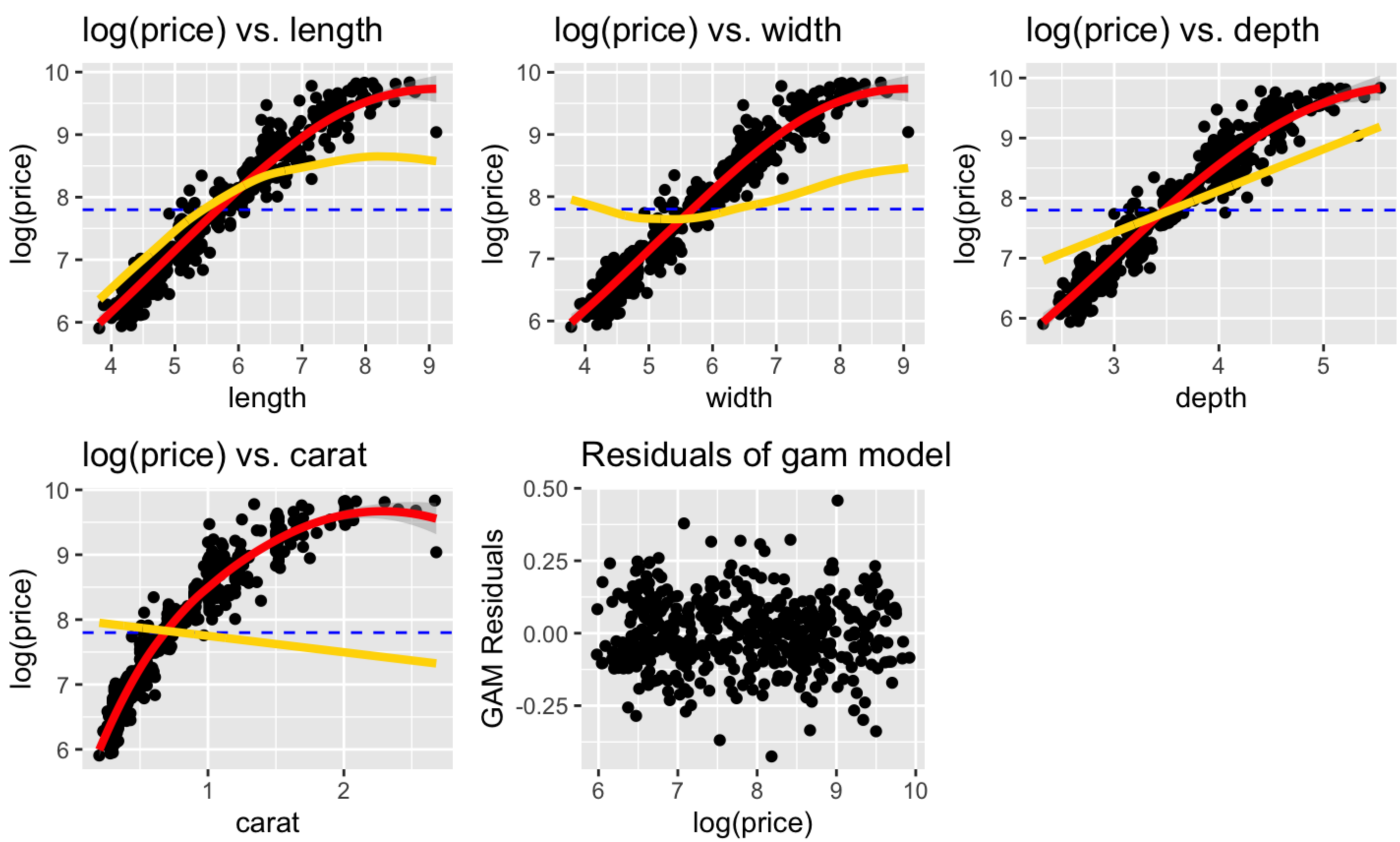
Table 5: Comparison of AIC between GAM models

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

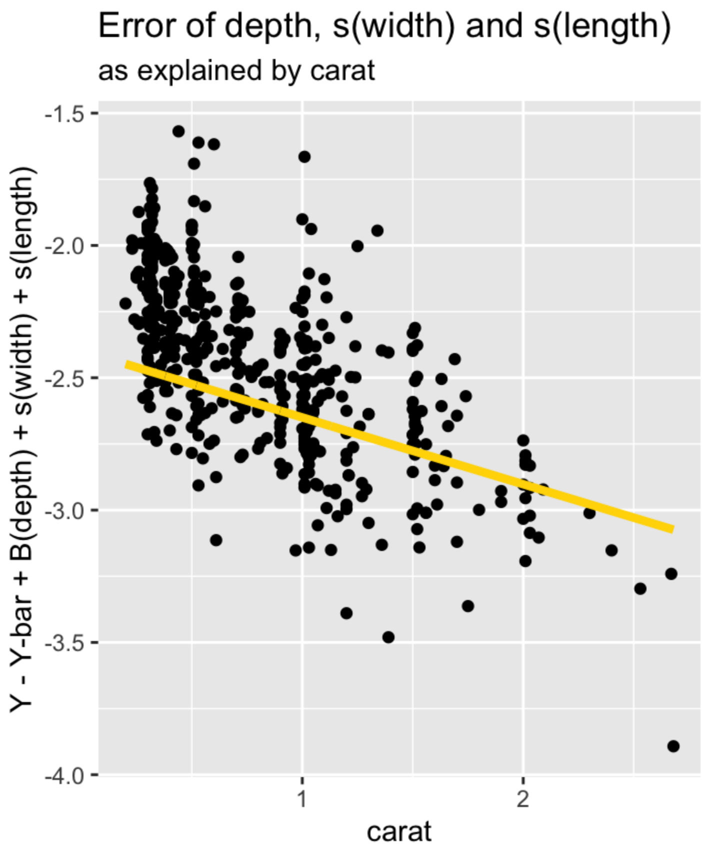
- The model with predictors `clarity + color + depth + carat + s(width) + s(length)` had the lowest AIC (-652).

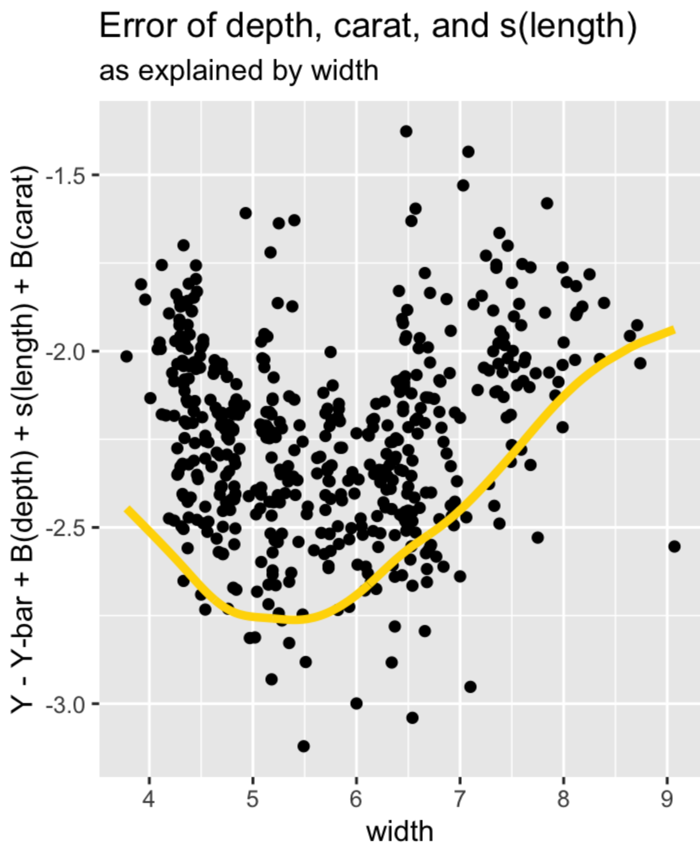
GAM: plotting chosen model

- Note: red = smoother; gold = GAM

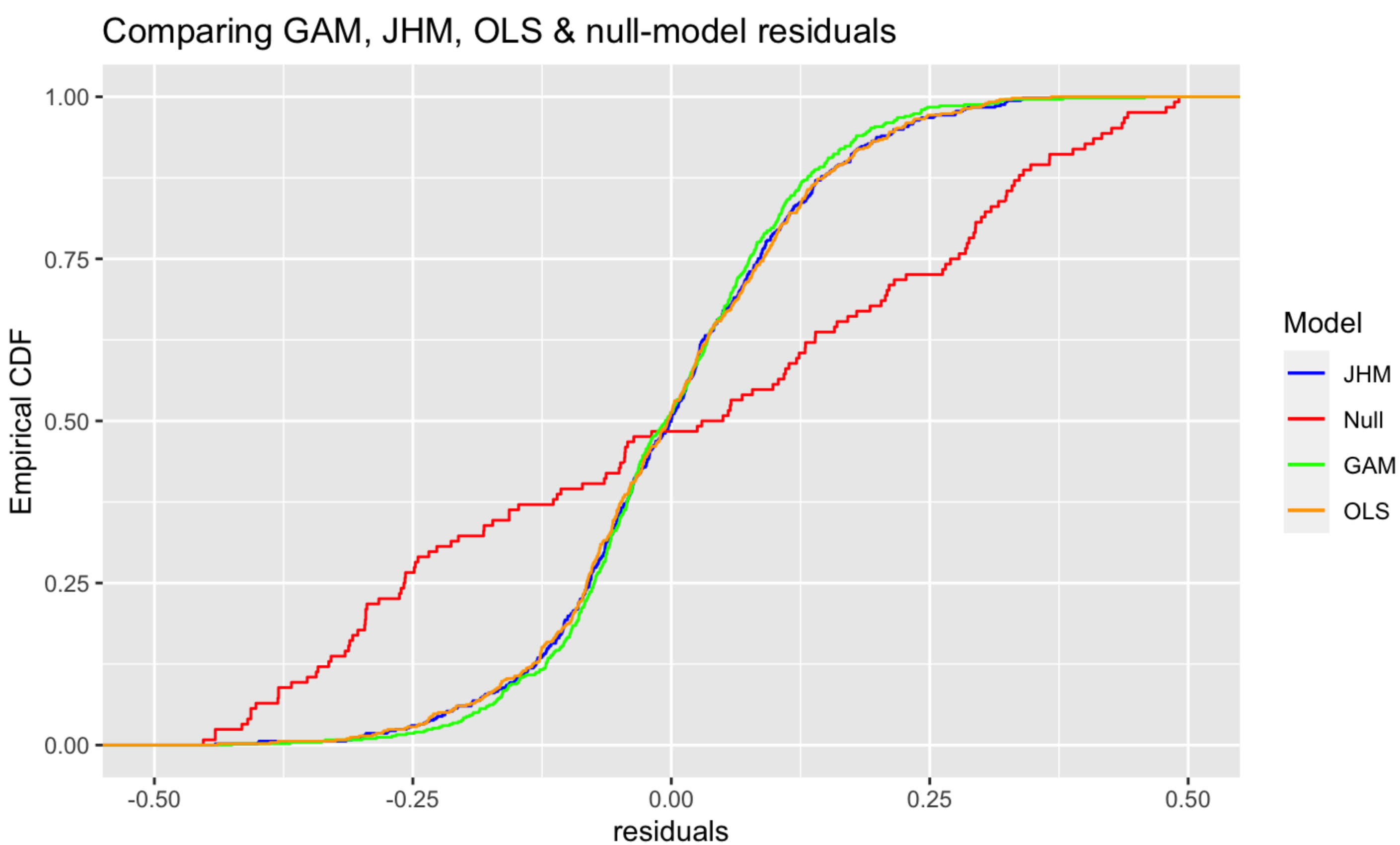


GAM: explaining the roles of `carat` and `width` in the model





Examining residuals: all attempted models



Assessing model fit: 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

- GAM outperforms the other models (look at cross-validation)
- R^2 values: OLS fails to explain 1.84% of the variability, while GAM fails to explain 1.53% of the variability
- Using the GAM model results in a 16% decrease in unexplained variability (relative to OLS).

Limitations

- Multicollinearity between `carat`, `length`, `width`, and `depth`.
- We don't know what year this dataset is from
    - If we did, we could use our model to predict diamond price and adjust for inflation.

Conclusions

- **Recall:** How can we predict diamond price?
- **Best model:** utilizes a GAM
  - Predictors: `clarity + color + depth + carat + s(width) + s(length)`