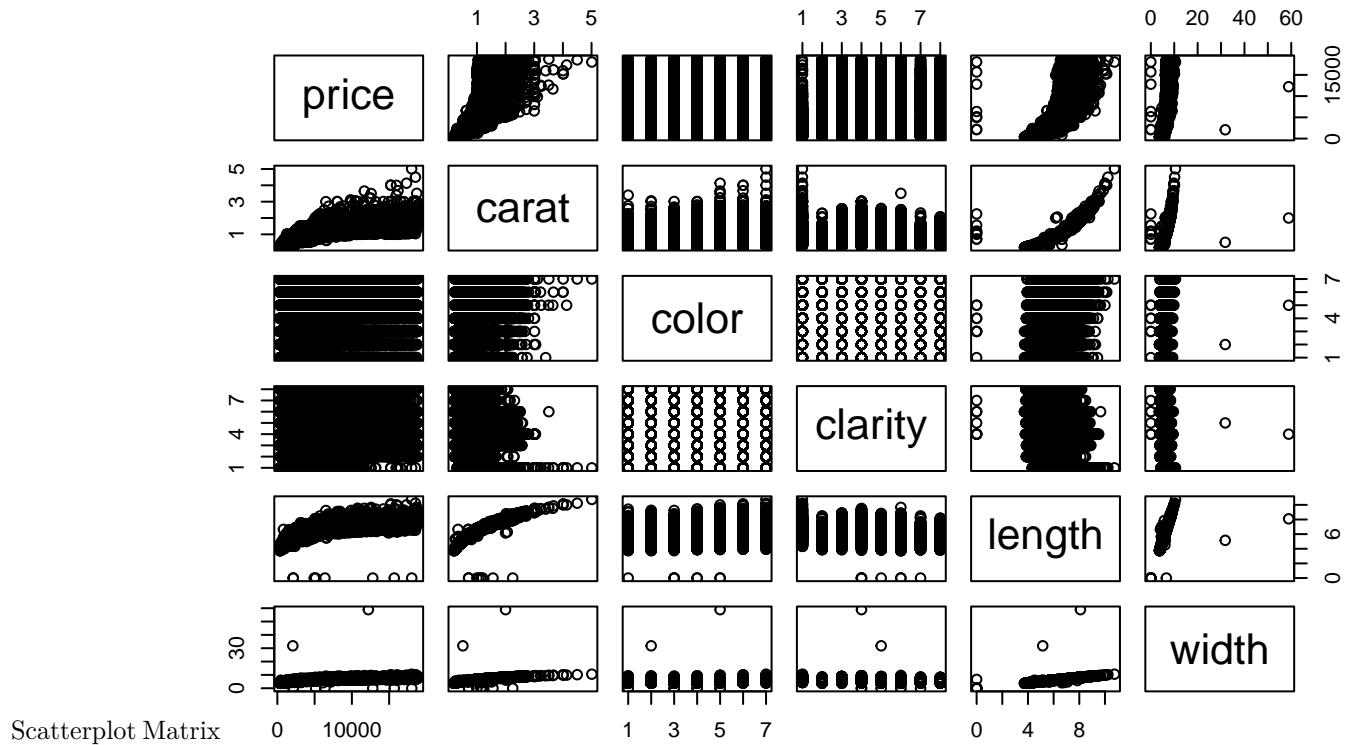


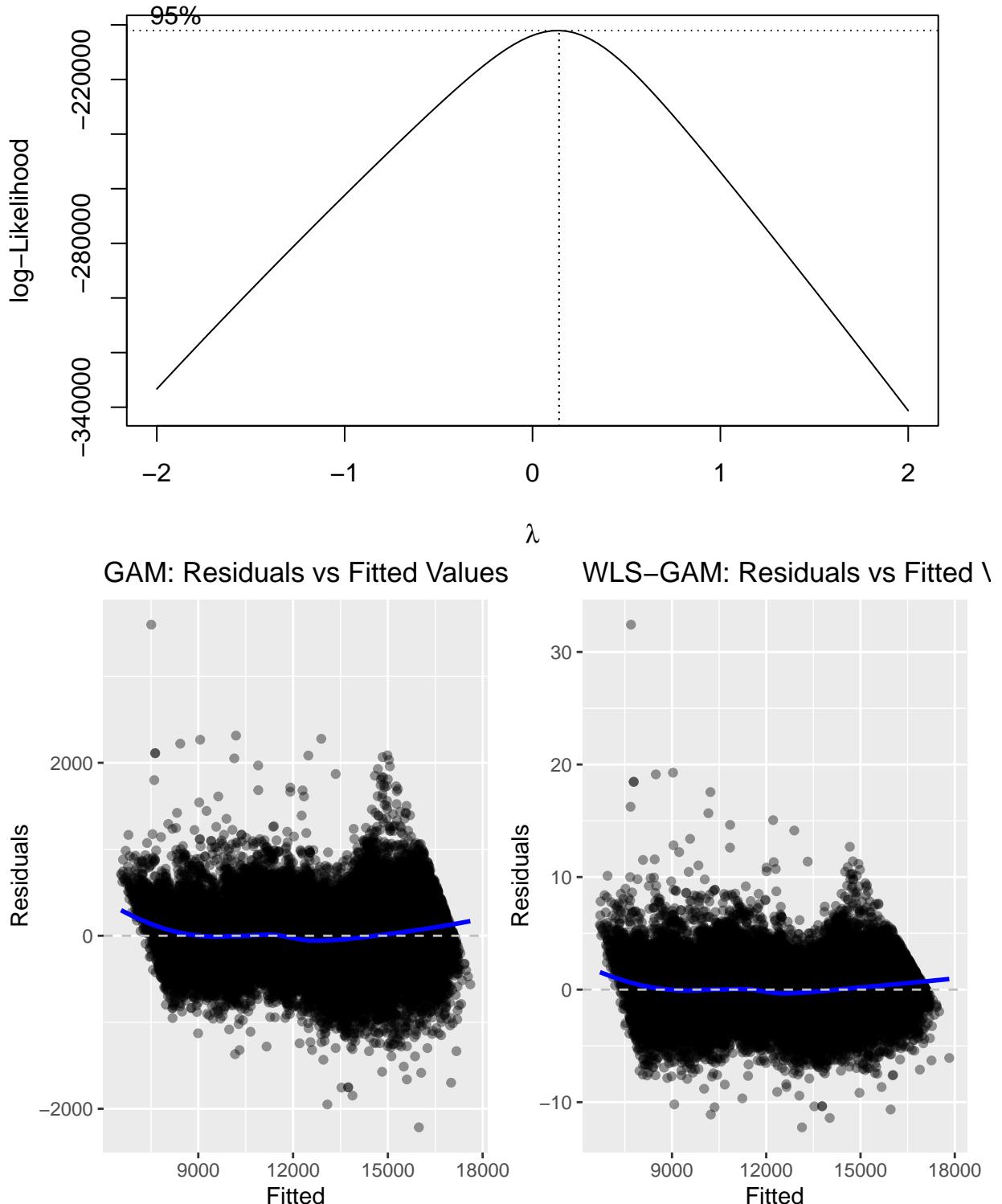
Section 1: Introduction

Section 2: Data description

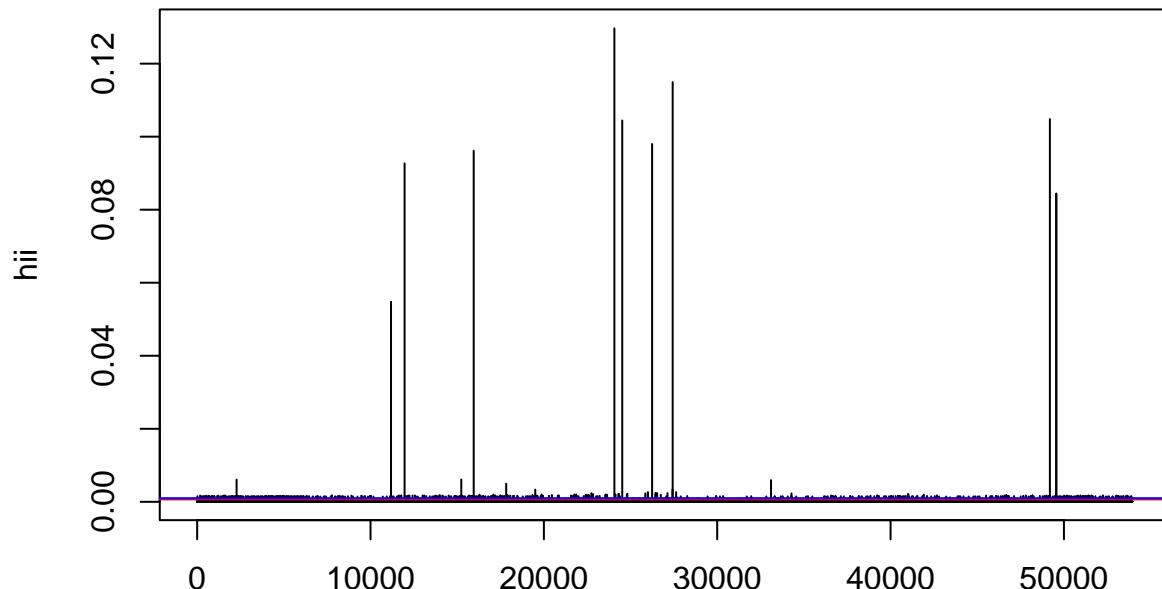


Section 3: Primary model results and diagnostics

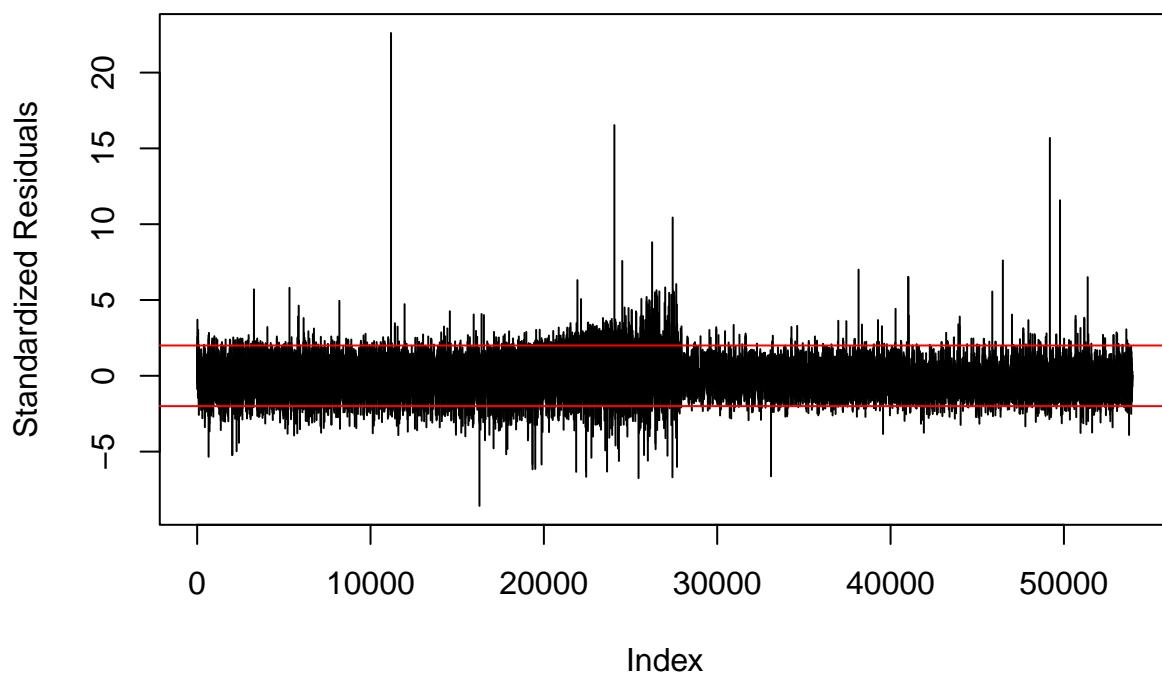
Section 4: Model Selection



Leverage Points



Standardized Residuals



Cook's Distance

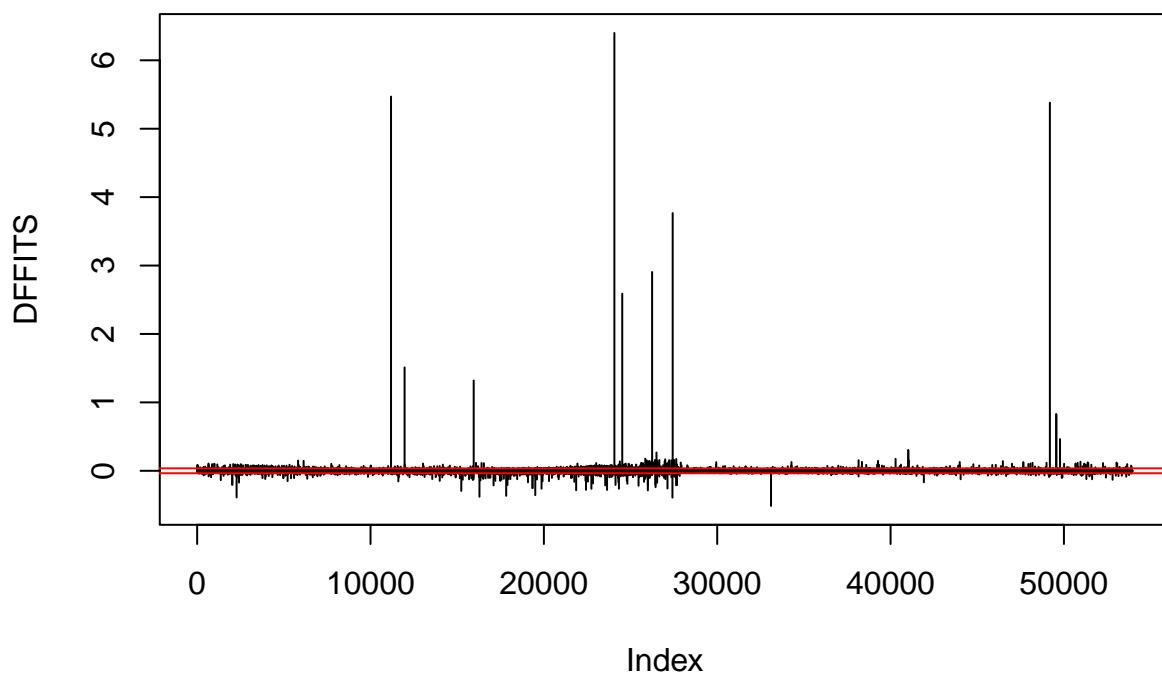
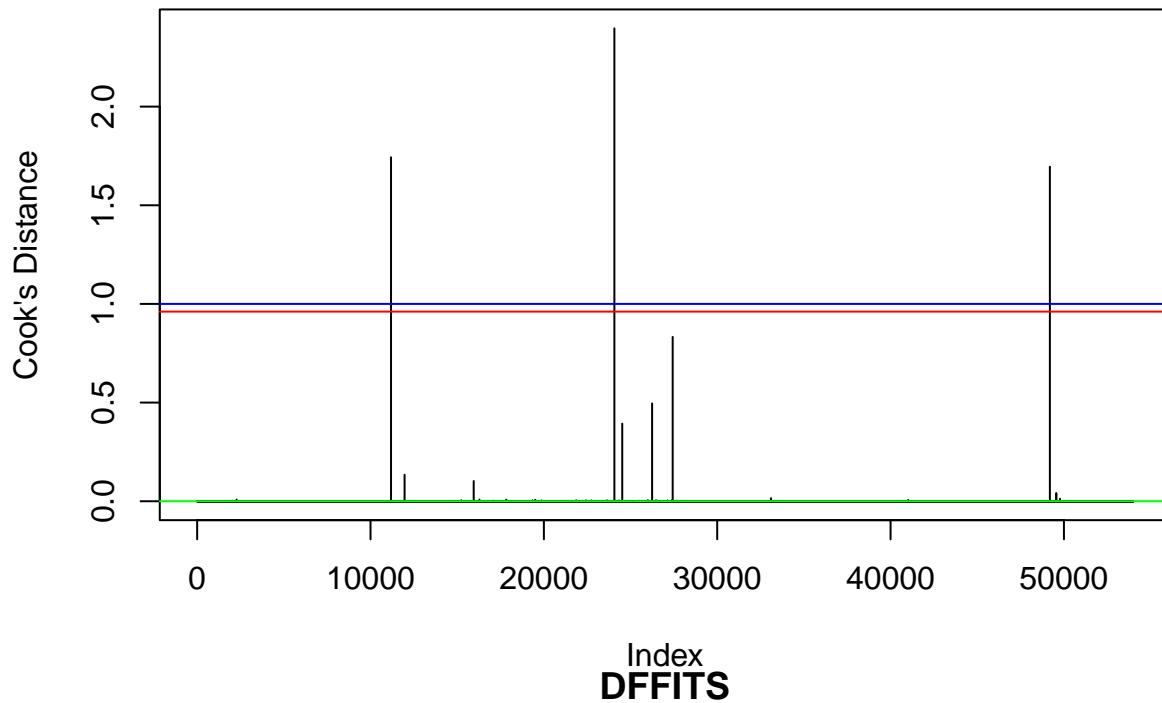


Table 1: Model Selection Table

Model	Adj_R.2	AIC	BIC
1	0.9198	911551.9	911774.3
2	0.9171	913284.5	913444.6

Model	Adj_R.2	AIC	BIC
3	0.8771	934525.8	934668.1
4	0.9789	788582.8	788742.9

Variation Inflation Factor (VIF)

```
##                               GVIF Df GVIF^(1/(2*Df))
## log(carat)      55.354350 1    7.440050
## color          1.157645 6    1.012274
## clarity         1.268696 7    1.017145
## length          84.288491 1    9.180876
## width_transformed 54.683706 1    7.394843
```

Variable Selection

```
## Start: AIC=635497.2
## y_transformed ~ log(carat) + color + clarity + length + width_transformed
##
##                               Df  Sum of Sq      RSS      AIC
## <none>                      7.0476e+09 635497
## - width_transformed  1 2.4997e+08 7.2976e+09 637375
## - length            1 1.1357e+09 8.1833e+09 643555
## - log(carat)        1 3.3802e+09 1.0428e+10 656630
## - color             6 5.3083e+09 1.2356e+10 665771
## - clarity           7 1.2388e+10 1.9436e+10 690204
##
## Call:
## lm(formula = y_transformed ~ log(carat) + color + clarity + length +
##       width_transformed, data = diamonds_cleaned)
##
## Residuals:
##   Min     1Q Median     3Q    Max 
## -3099.1 -227.4 -10.6  219.5 7947.1
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 10234.901    110.302   92.79 <2e-16 ***
## log(carat)  3184.716    19.802  160.82 <2e-16 ***
## colorE     -138.364    5.722  -24.18 <2e-16 ***
## colorF     -227.214    5.787  -39.27 <2e-16 ***
## colorG     -374.700    5.664  -66.16 <2e-16 ***
## colorH     -606.448    6.018 -100.77 <2e-16 ***
## colorI     -900.656    6.754 -133.36 <2e-16 ***
## colorJ     -1274.656   8.331 -153.01 <2e-16 ***
## clarityIF  2928.879   16.094  181.98 <2e-16 ***
## claritySI1 1660.902   13.747  120.82 <2e-16 ***
## claritySI2 1216.971   13.837   87.95 <2e-16 ***
## clarityVS1 2217.238   14.023  158.12 <2e-16 ***
## clarityVS2 2039.547   13.803  147.76 <2e-16 ***
## clarityVVS1 2696.780   14.857  181.51 <2e-16 ***
## clarityVVS2 2543.294   14.443  176.09 <2e-16 ***
## length      1187.588   12.740   93.22 <2e-16 ***
## width_transformed -3047.885  69.690  -43.73 <2e-16 ***
```

```

## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 361.5 on 53926 degrees of freedom
## Multiple R-squared: 0.9789, Adjusted R-squared: 0.9789
## F-statistic: 1.561e+05 on 16 and 53926 DF, p-value: < 2.2e-16

```

Section 5: Final Model Inference & Results

Final Model Summary with Confidence Intervals

```

##              2.5 %    97.5 %
## (Intercept) 10018.7072 10451.0945
## log(carat)   3145.9026 3223.5285
## colorE      -149.5798 -127.1483
## colorF      -238.5558 -215.8722
## colorG      -385.8010 -363.5990
## colorH      -618.2435 -594.6523
## colorI      -913.8928 -887.4188
## colorJ     -1290.9844 -1258.3283
## clarityIF   2897.3340 2960.4235
## claritySI1  1633.9578 1687.8453
## claritySI2  1189.8506 1244.0914
## clarityVS1  2189.7527 2244.7224
## clarityVS2  2012.4927 2066.6012
## clarityVVS1 2667.6594 2725.9004
## clarityVVS2 2514.9861 2571.6020
## length       1162.6180 1212.5573
## width_transformed -3184.4787 -2911.2913

```

Table 2: Final Model Summary with Confidence Intervals

Term	Estimate	Std. Error	t value	p-value	CI Lower	CI Upper
(Intercept)	10234.901	110.302	92.789	<2e-16	10018.707	10451.095
log(carat)	3184.716	19.802	160.824	<2e-16	3145.903	3223.528
colorE	-138.364	5.722	-24.180	<2e-16	-149.580	-127.148
colorF	-227.214	5.787	-39.265	<2e-16	-238.556	-215.872
colorG	-374.700	5.664	-66.158	<2e-16	-385.801	-363.599
colorH	-606.448	6.018	-100.770	<2e-16	-618.244	-594.652
colorI	-900.656	6.754	-133.360	<2e-16	-913.893	-887.419
colorJ	-1274.656	8.331	-153.009	<2e-16	-1290.984	-1258.328
clarityIF	2928.879	16.094	181.984	<2e-16	2897.334	2960.423
claritySI1	1660.902	13.747	120.821	<2e-16	1633.958	1687.845
claritySI2	1216.971	13.837	87.951	<2e-16	1189.851	1244.091
clarityVS1	2217.238	14.023	158.116	<2e-16	2189.753	2244.722
clarityVS2	2039.547	13.803	147.760	<2e-16	2012.493	2066.601
clarityVVS1	2696.780	14.857	181.512	<2e-16	2667.659	2725.900
clarityVVS2	2543.294	14.443	176.095	<2e-16	2514.986	2571.602
length	1187.588	12.740	93.220	<2e-16	1162.618	1212.557
width_transformed	-3047.885	69.690	-43.735	<2e-16	-3184.479	-2911.291

Model Performance Summary

Table 3: Model Performance Summary

Term	Value
Residual Std. Error	361.5
Residual Std. Error df	53,926
Multiple R^2	0.9788586
Adjusted R^2	0.9788524
F-statistic	1.561e+05
F-test df	(16, 53,926)
p-value	< 2.2e-16

Model Validation: 80% Training Set & 20% Testing Set

Table 4: Model Validation: Test Set MSE Comparison

Model	Test_MSE
Model 1: Full Model (carat + cut + color + clarity + depth + table + x + y + z)	1318786.2
Model 2: Preliminary Model (carat + color + clarity + length + width)	1360844.4
Model 3: (carat + cut + color + x + y + z)	2024297.1
Model 4: Final Transformed Model (log(carat), width_transformed, Box-Cox)	210146.2

