Statistical Learning

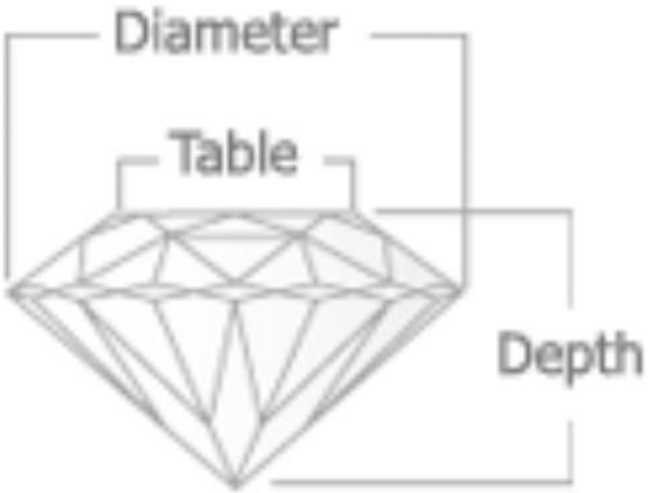
Diamonds Dataset

Project Work By:
Davide Gamba, Matr. 1053470
Aurora Zanenga, Matr. 1054891

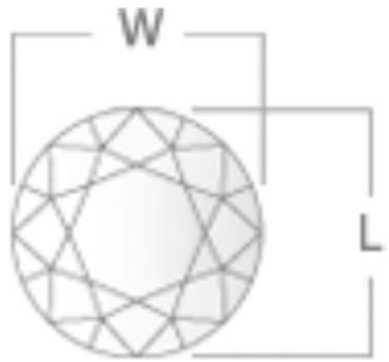
DIAMONDS DATASET

Variable Name	Description	Variable Type
PRICE	Price in US dollars ($326-$18.823)	Continuous
CARAT	Weight of the diamond (0.2--5.01)	Continuous
CUT	Quality of the cut (Fair, Good, Very Good, Premium, Ideal)	Categorical
COLOR	Diamond color, from J (coloured, worst) to D (colorless, best)	Categorical
CLARITY	Refers to how clear the material is and what surface imperfections it may have (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))	Categorical
X	Length in mm (0--10.74)	Continuous
Y	Width in mm (0--58.9)	Continuous
Z	Depth in mm (0--31.8)	Continuous
DEPTH %	Is the depth of the diamond divided by the diameter of the diamond (43%--79%)	Continuous
TABLE %	Is the length of the table facet divided by the diameter of the diamond(43%--95%)	Continuous

CONTEXT



	Depth %	Table %
Excellent	59.0% - 61.0%	53% - 60%
Very Good	58.0% - 62.0%	61% - 62%
Good	56% - 64%	62% – 64%
Fair	64% - 70%	64% - 66%
Poor	over 70%	over 66% or under 53%



CONTEXT


COLORLESS
NEAR COLORLESS
NOTICEABLE COLOR
D
E
F
G
H
I
J
K-M
N-R
S-Z
FL
IF
VVS1-VVS2
VS1-VS2
SI1-SI2-SI3
I1-I2-I3

GOALS

- Which models can most efficiently predict price?
- The relationship between features and price is linear?
- Which are the variables that have the greatest influence on price?

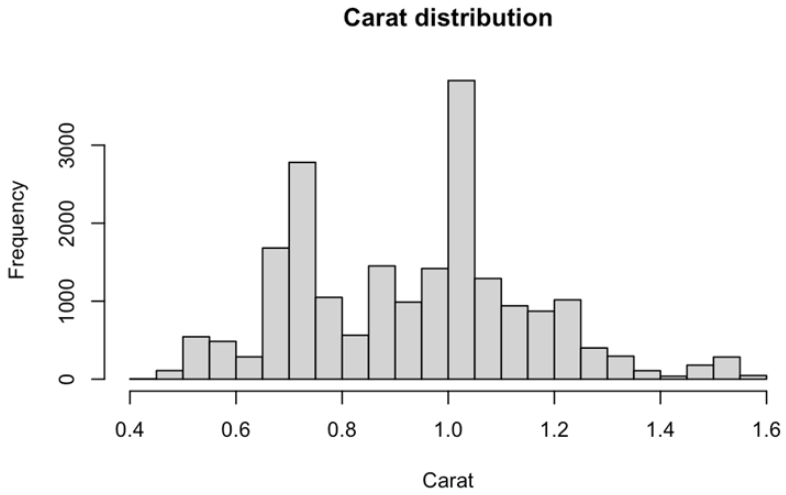
PREPROCESSING STEPS

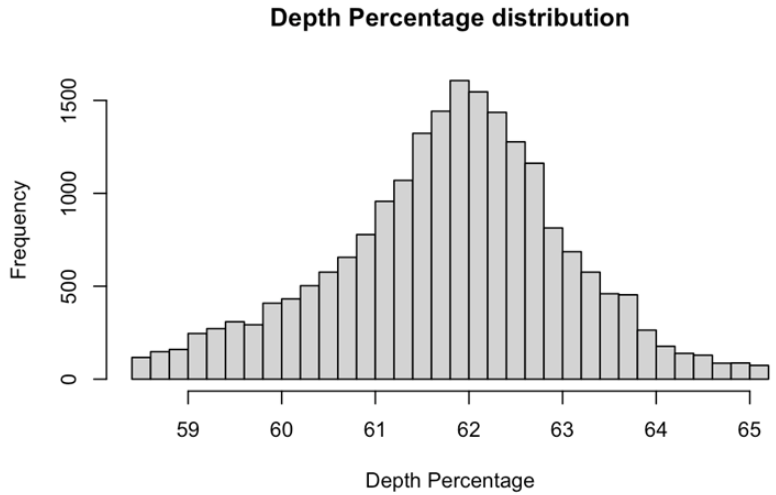
- Check for null columns or elements.
- Redefining the dataset into a range from 2000$ to 8000$.
- Converting prices in thousands dollars.
- Categorical variables into factors.
- Removing outliers.
- From 50.000 to 20.000 observations.
- Splitting the dataset in training set and test set (70% - 30%).

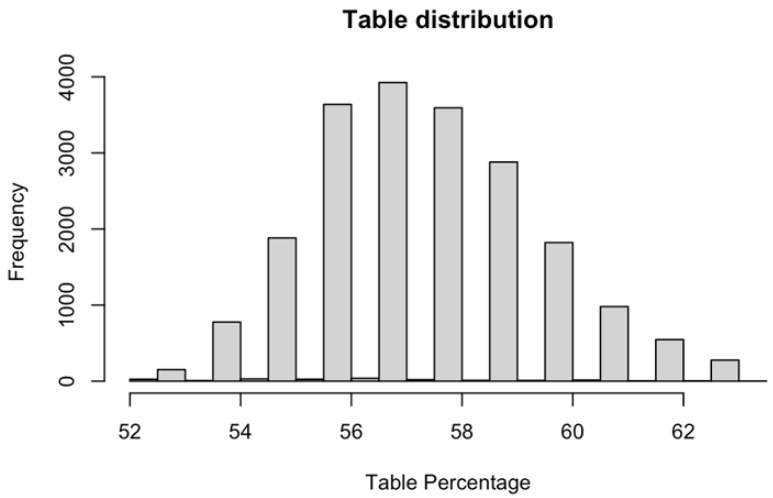
VARIABLES ANALYSIS

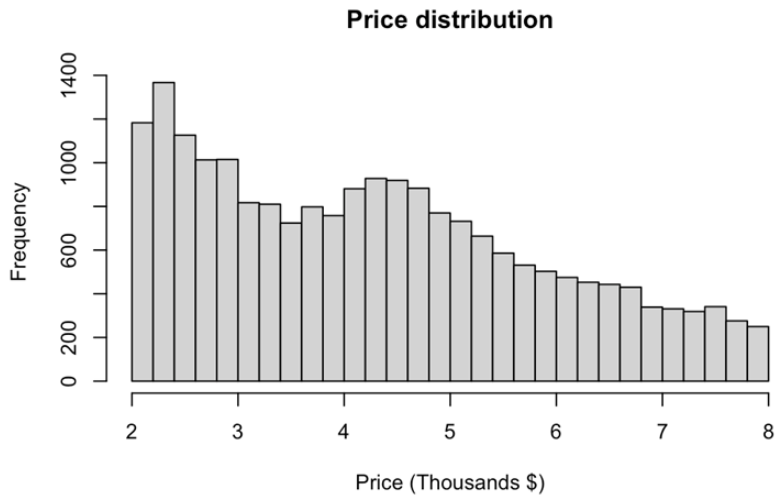
	Unit of Measure	Mean	Std_dev
Carat	Unit	0.9396	0.2189
Depth_Percentage	%	61.8425	1.2497
Table	%	57.6094	2.0321
Price	Thousand $	4.3114	1.5983
Length	mm	6.2356	0.4940
Width	mm	6.2368	0.4853
Depth	mm	3.8562	0.3063

VARIABLES DISTRIBUTION

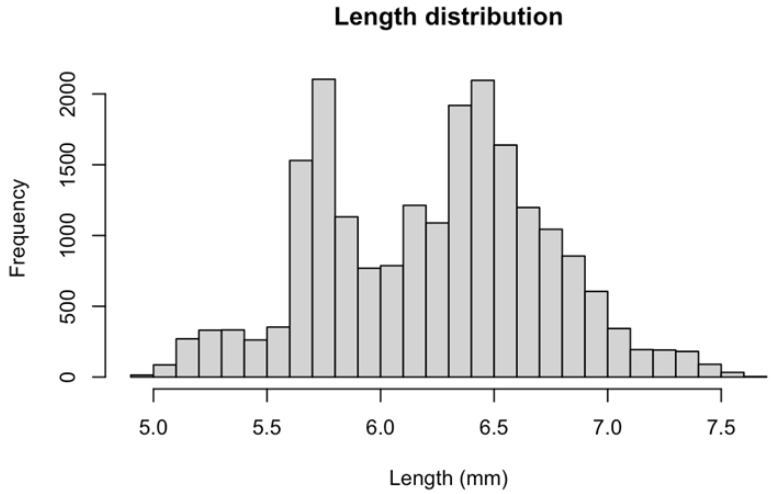


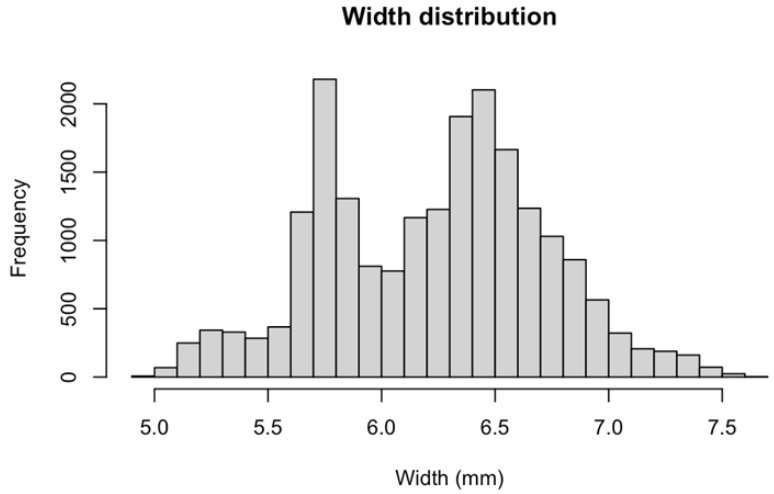


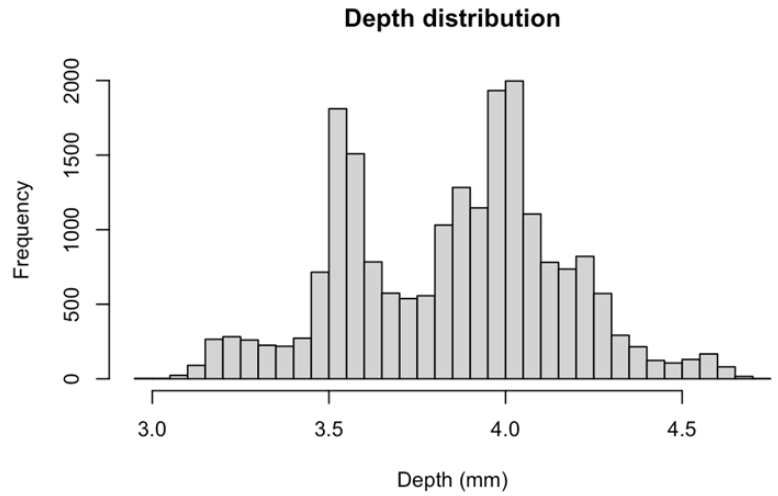




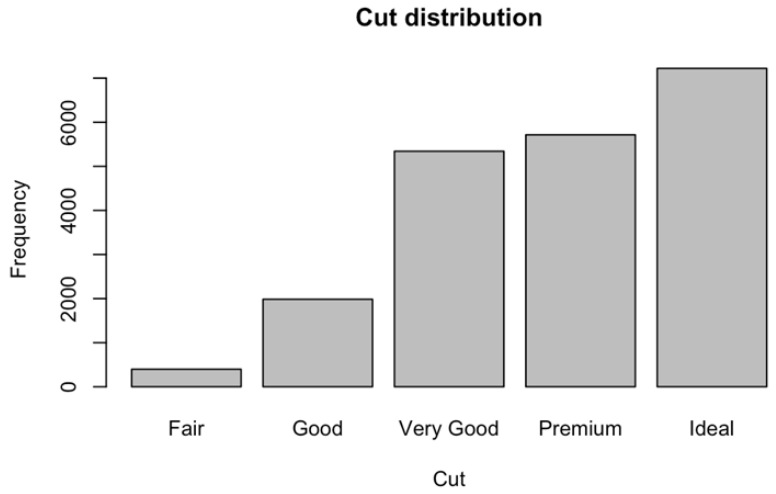
VARIABLES DISTRIBUTION

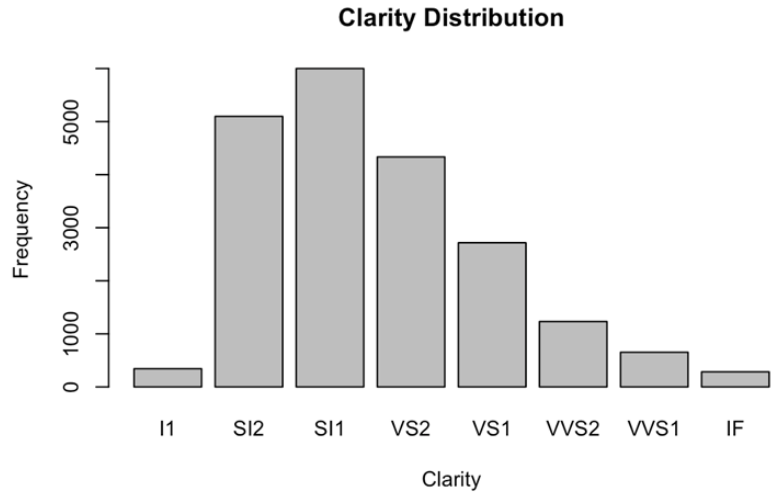




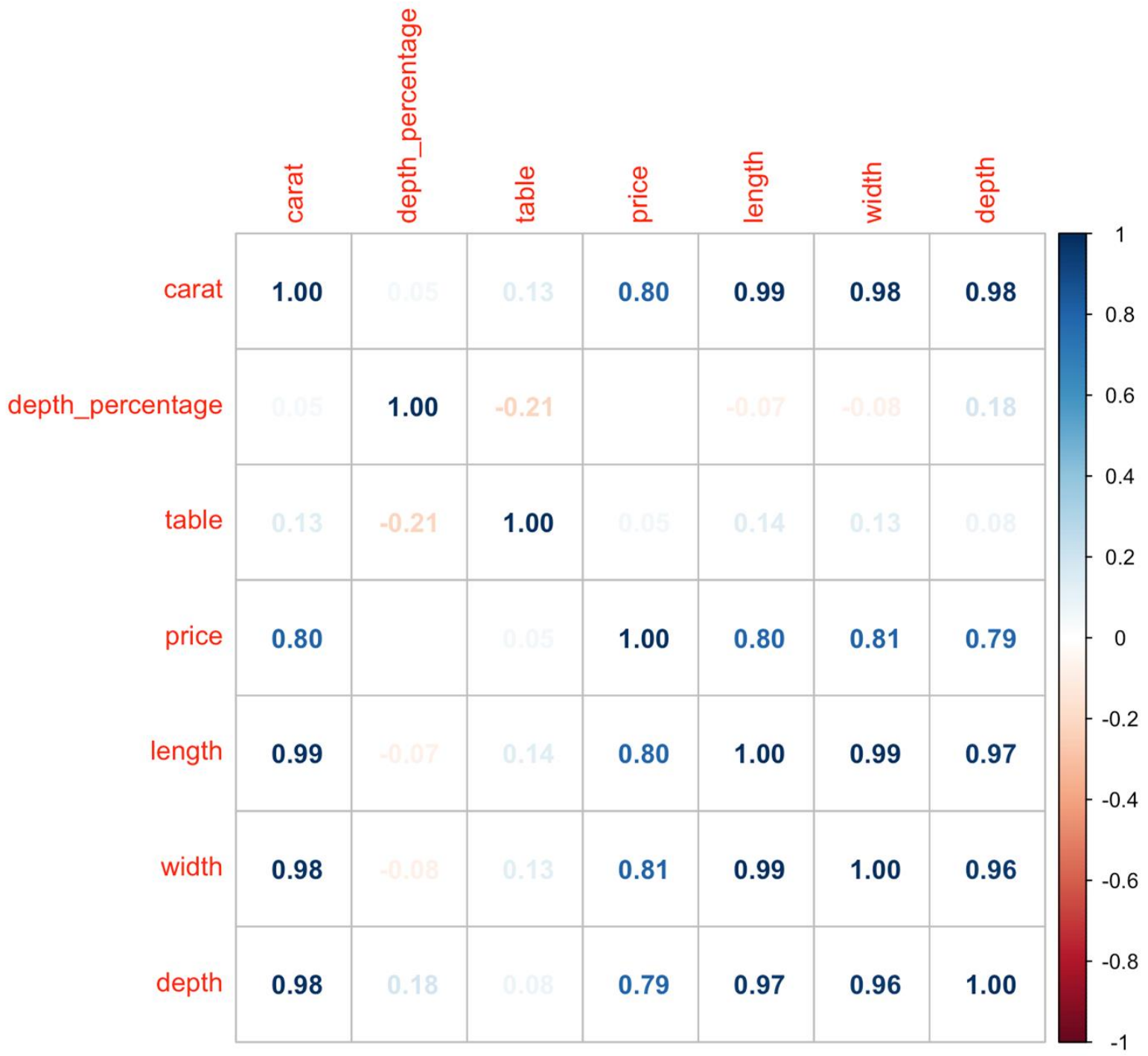


VARIABLES DISTRIBUTION





CORRELATION MATRIX



LINEAR REGRESSION

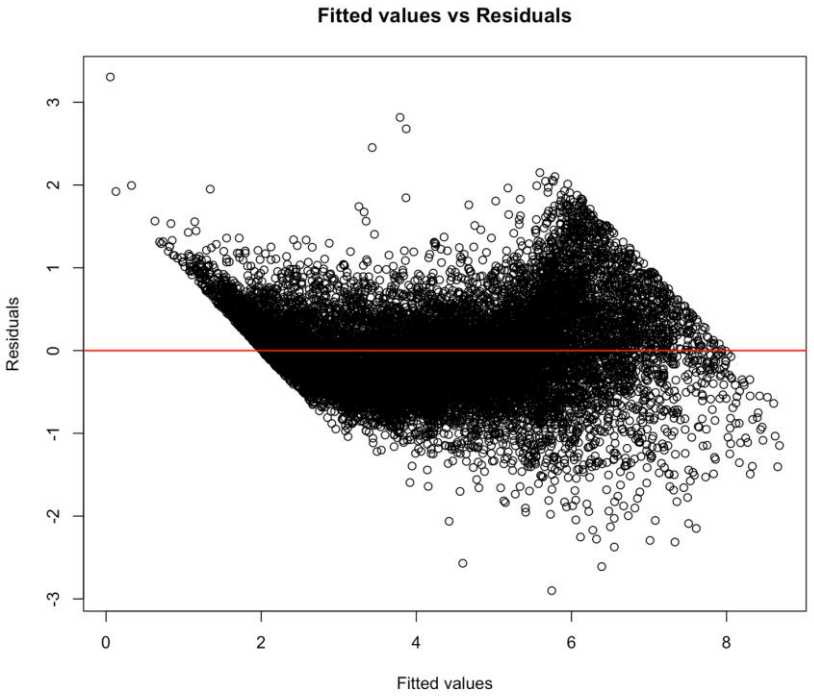
```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)      -24.028318   1.622995 -14.805  < 2e-16 ***
carat              1.630952   0.189125   8.624  < 2e-16 ***
cutGood            0.160002   0.036392   4.397 1.11e-05 ***
cutVery Good       0.265929   0.035788   7.431 1.14e-13 ***
cutPremium         0.264386   0.035952   7.354 2.03e-13 ***
cutIdeal           0.446810   0.036321  12.302  < 2e-16 ***
colorI             0.566001   0.023137  24.463  < 2e-16 ***
colorH             1.088515   0.021786  49.963  < 2e-16 ***
colorG             1.508953   0.021907  68.879  < 2e-16 ***
colorF             1.662064   0.022428  74.108  < 2e-16 ***
colorE             1.783505   0.022874  77.969  < 2e-16 ***
colorD             1.977068   0.024021  82.307  < 2e-16 ***
claritySI2         1.472192   0.036053  40.834  < 2e-16 ***
claritySI1         2.172198   0.036114  60.148  < 2e-16 ***
clarityVS2         2.808009   0.036548  76.830  < 2e-16 ***
clarityVS1         3.051263   0.037594  81.163  < 2e-16 ***
clarityVVS2        3.431726   0.040729  84.258  < 2e-16 ***
clarityVVS1        3.604959   0.044625  80.783  < 2e-16 ***
clarityIF          3.817592   0.052810  72.289  < 2e-16 ***
depth_percentage   0.057857   0.025342   2.283  0.02244 *
table              0.011924   0.002828   4.217 2.49e-05 ***
length            -0.196828   0.153418  -1.283  0.19953
width              2.452116   0.157402  15.579  < 2e-16 ***
depth              1.150811   0.404229   2.847  0.00442 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5368 on 14441 degrees of freedom
Multiple R-squared:  0.8879,    Adjusted R-squared:  0.8877
F-statistic:  4972 on 23 and 14441 DF,  p-value: < 2.2e-16
```

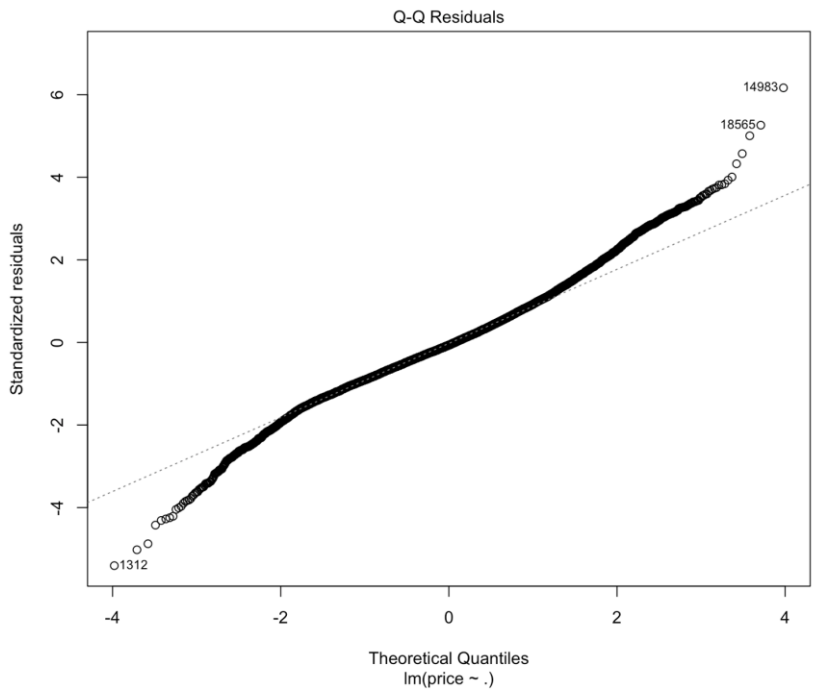
- It shows how significant the coefficients are.

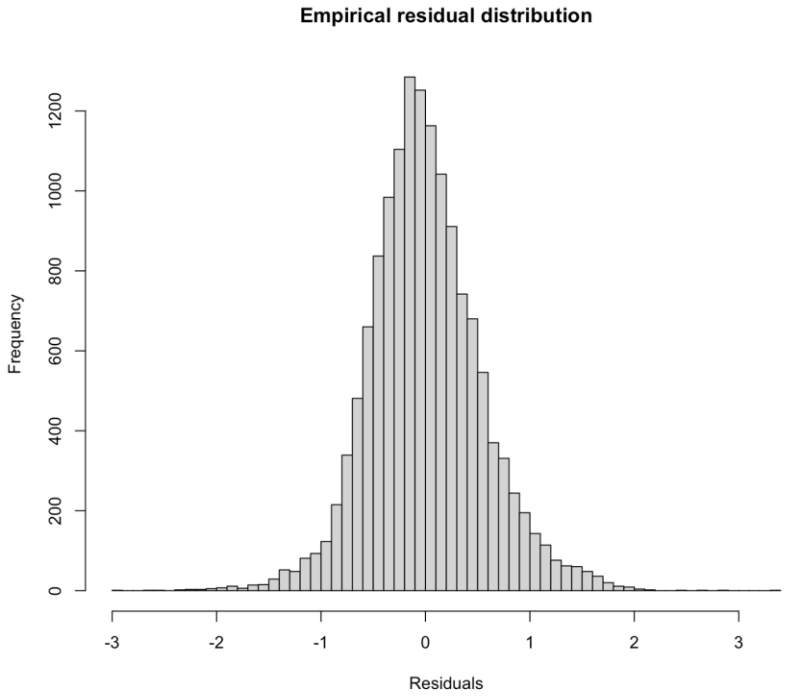
R^2	Test RMSE
0.8878	0.5548

RESIDUALS ANALYSIS

- Most of the residuals are in the [-1;1] range (which means from -1000$ to 1000$).
- The model works better on the central prices values.
- Shapiro test confirms that residuals are not normally distributed.

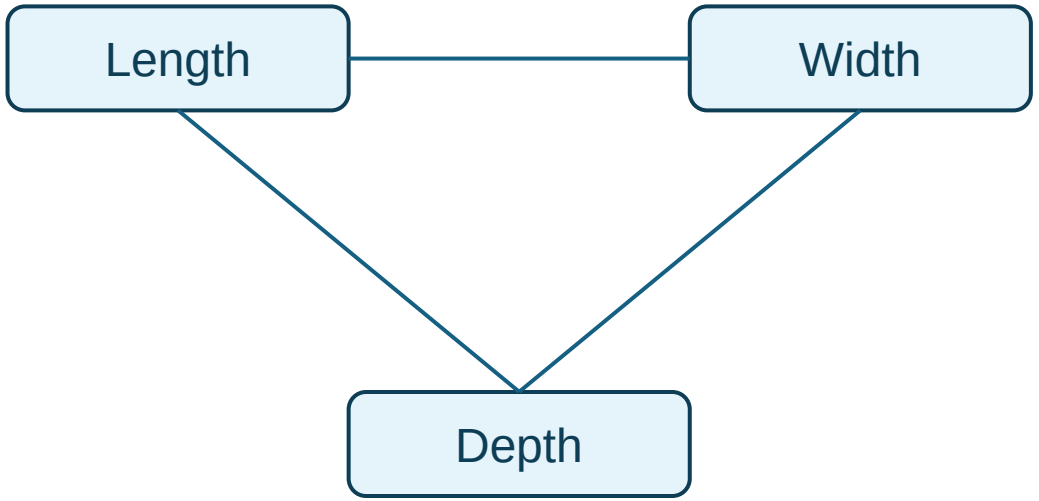






INTERACTION TERM BETWEEN DIMENSIONS

- How could the model be influenced by physical variables if considered together?
- Better results considering the interaction between variables?



LINEAR REGRESSION

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       91.785288   5.657500  16.224  < 2e-16 ***
carat              5.584447   0.315868  17.680  < 2e-16 ***
cutGood            0.124983   0.035367   3.534 0.000411 ***
cutVery Good       0.240457   0.034816   6.906 5.18e-12 ***
cutPremium         0.238701   0.034989   6.822 9.33e-12 ***
cutIdeal           0.423153   0.035401  11.953  < 2e-16 ***
colorI             0.550337   0.022491  24.470  < 2e-16 ***
colorH             1.044769   0.021264  49.133  < 2e-16 ***
colorG             1.457642   0.021401  68.111  < 2e-16 ***
colorF             1.616615   0.021879  73.889  < 2e-16 ***
colorE             1.742681   0.022281  78.214  < 2e-16 ***
colorD             1.923765   0.023431  82.105  < 2e-16 ***
claritySI2         1.462565   0.034999  41.789  < 2e-16 ***
claritySI1         2.160655   0.035059  61.629  < 2e-16 ***
clarityVS2         2.795641   0.035484  78.786  < 2e-16 ***
clarityVS1         3.045578   0.036506  83.426  < 2e-16 ***
clarityVVS2        3.391681   0.039847  85.118  < 2e-16 ***
clarityVVS1        3.563751   0.043604  81.730  < 2e-16 ***
clarityIF          3.794251   0.051737  73.337  < 2e-16 ***
depth_percentage   0.050315   0.027723   1.815 0.069557 .
table             -0.003103   0.002929  -1.059 0.289421
length            -5.699805   1.375413  -4.144 3.43e-05 ***
width            -28.276382   1.304293 -21.679  < 2e-16 ***
depth            -30.683955   1.793132 -17.112  < 2e-16 ***
length:width       2.889977   0.162463  17.788  < 2e-16 ***
length:depth       1.711700   0.347860   4.921 8.72e-07 ***
width:depth        8.275476   0.332099  24.919  < 2e-16 ***
length:width:depth -0.812760   0.036050 -22.545  < 2e-16 ***
```

- Both 2nd order interaction and 3rd order interaction are significant.
- The new model is better then the previous one (smaller RMSE).

R^2	Test RMSE
0.8945	0.5328

COMPARISON BETWEEN MODELS

Anova Test

```
Analysis of Variance Table

Model 1: price ~ carat + cut + color + clarity + depth_percentage + table +
    length + width + depth
Model 2: price ~ carat + cut + color + clarity + depth_percentage + table +
    length + width + depth + (length * width * depth)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  14441 4161.7
2  14437 3915.2  4     246.5 227.24 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The anova test shows that the second model is better compared to the first one.

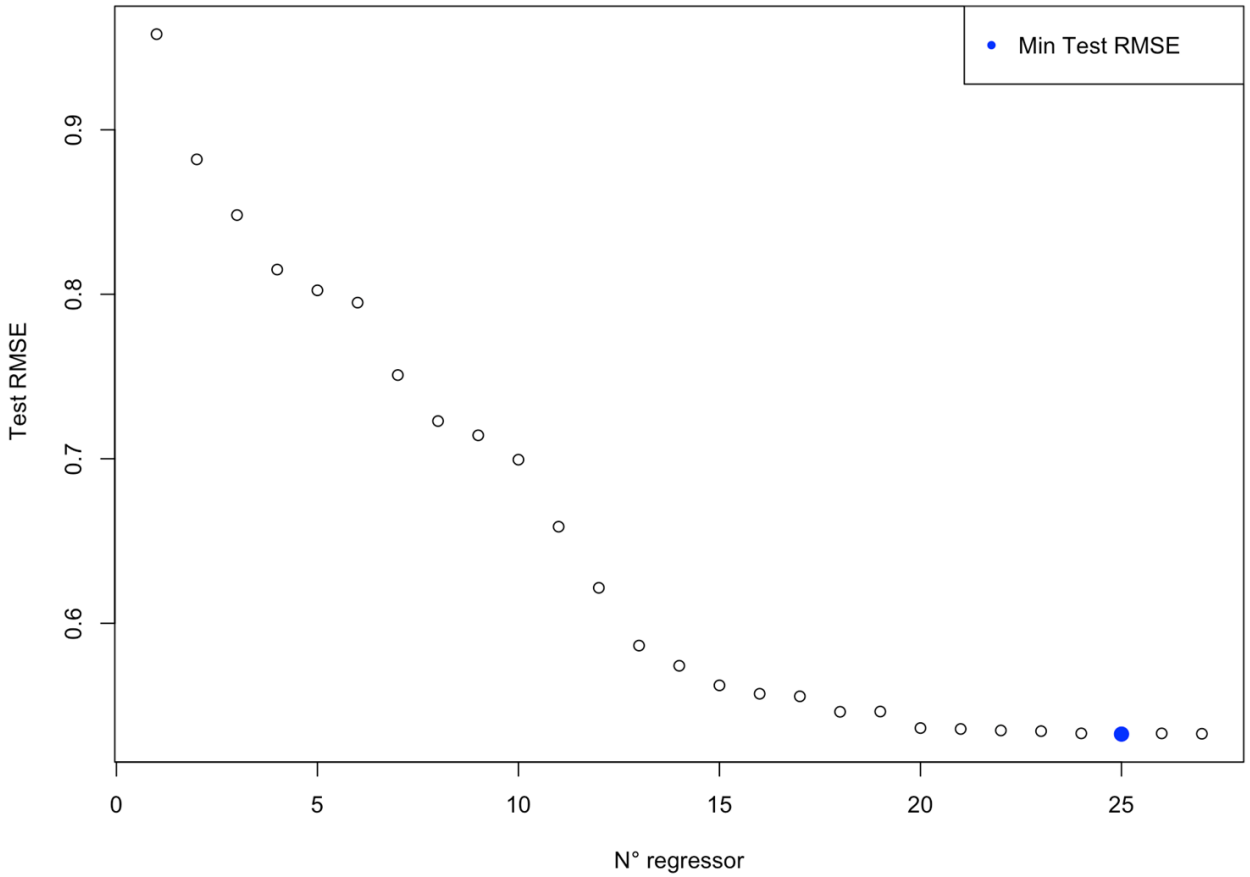
First Model:

R^2	Test RMSE
0.8878	0.5548

Second Model:

R^2	Test RMSE
0.8945	0.5328

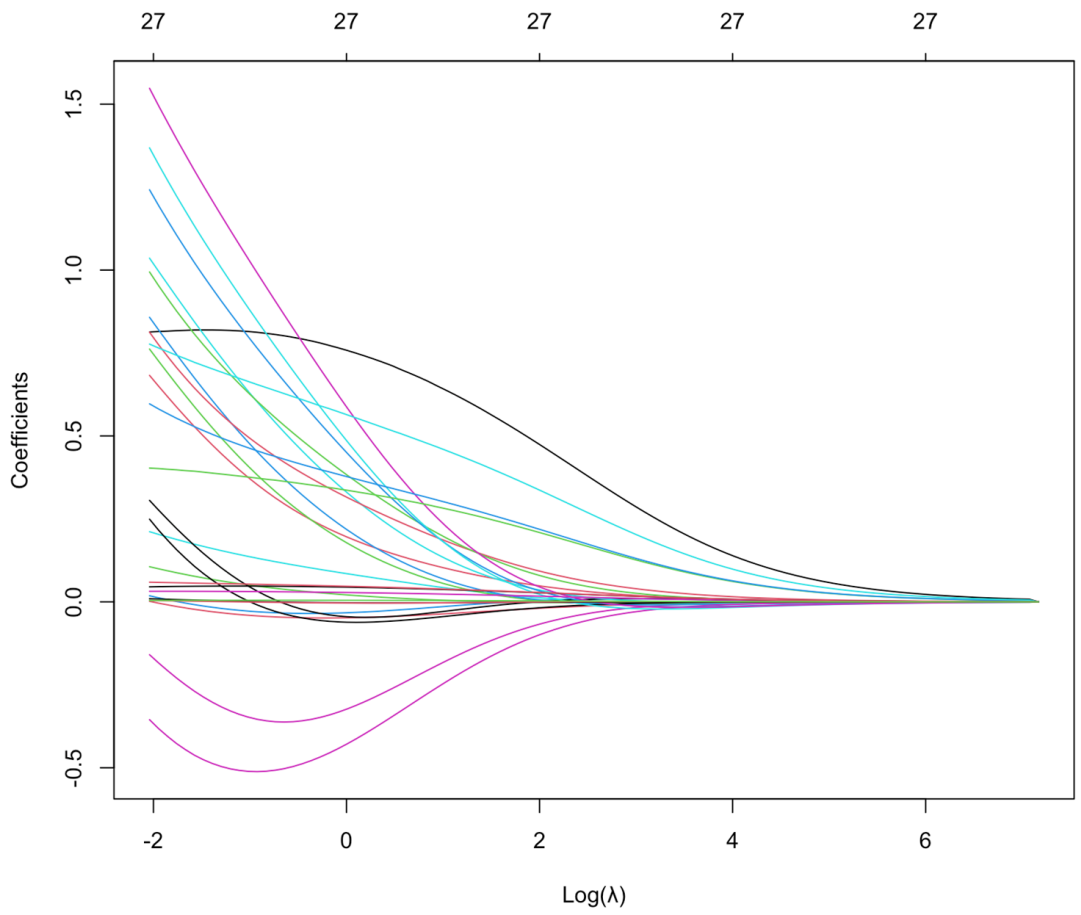
BEST SUBSET SELECTION

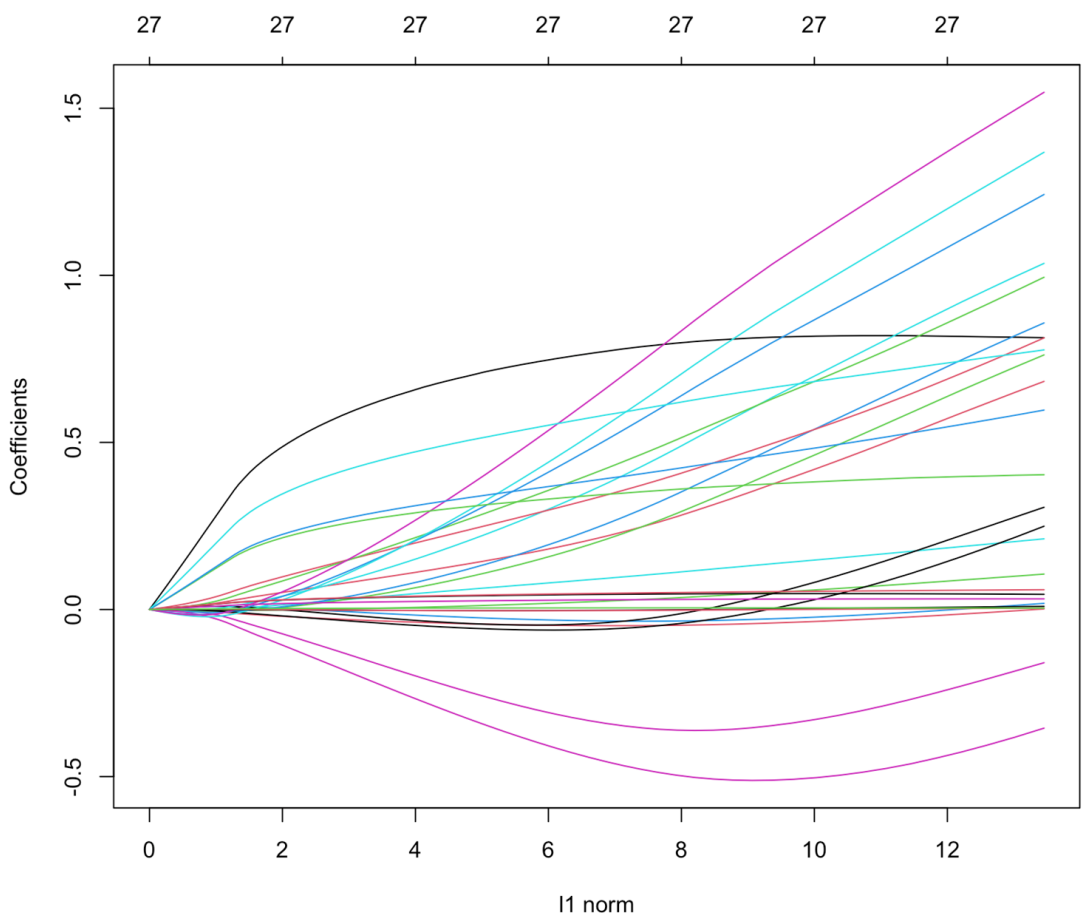


- Best subset selection with regsubset() command.
- For each model (from 1 to 27 regressors) we calculated the RMSE test.
- The best model has 25 regressors.
- The deleted regressors are table and depth%.
- This is the best achieved RMSE.

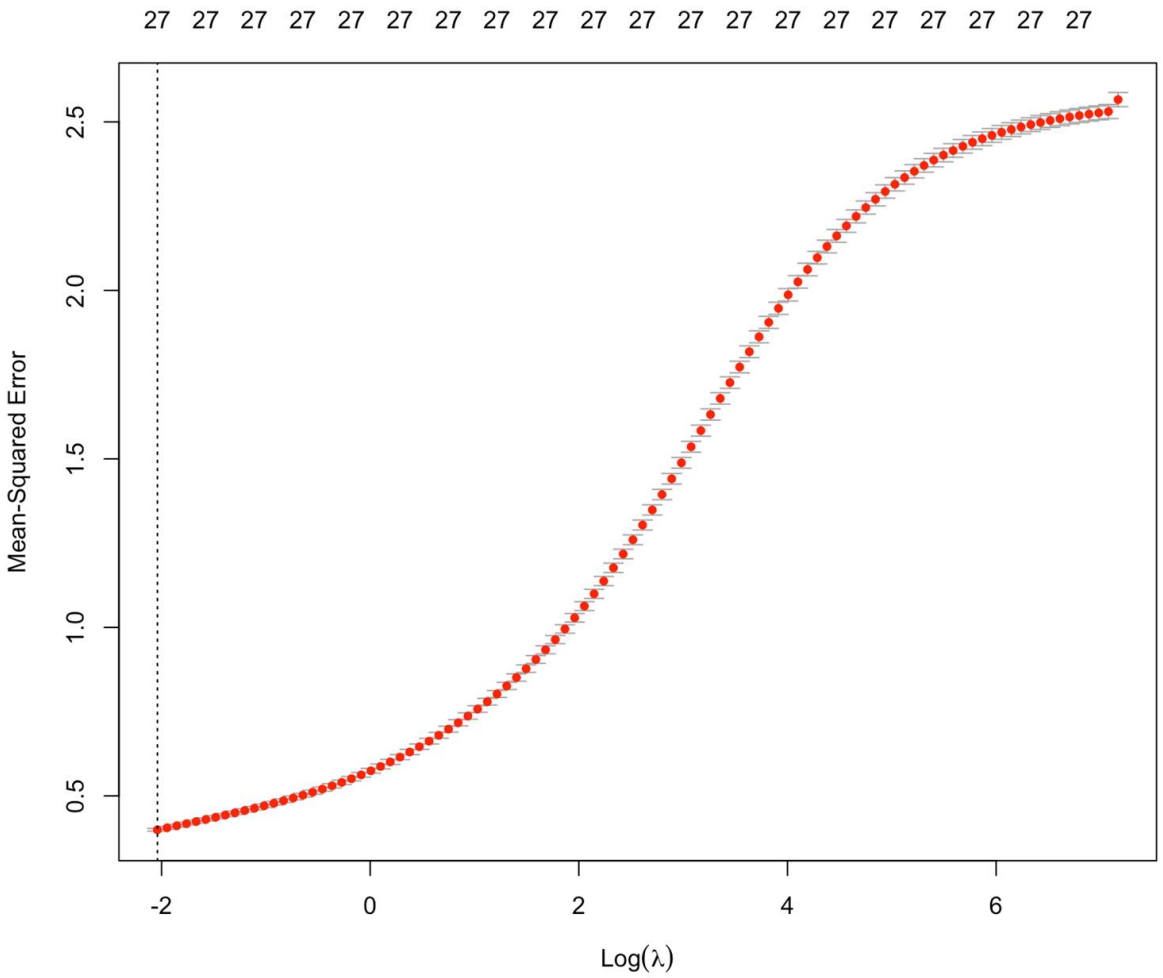
Test RMSE
0.5326

RIDGE REGRESSIONS





RIDGE REGRESSIONS



- Ridge regression model perform worst then the linear regression model.
- The reduction of the variance doesn't worth the increasing of the bias.

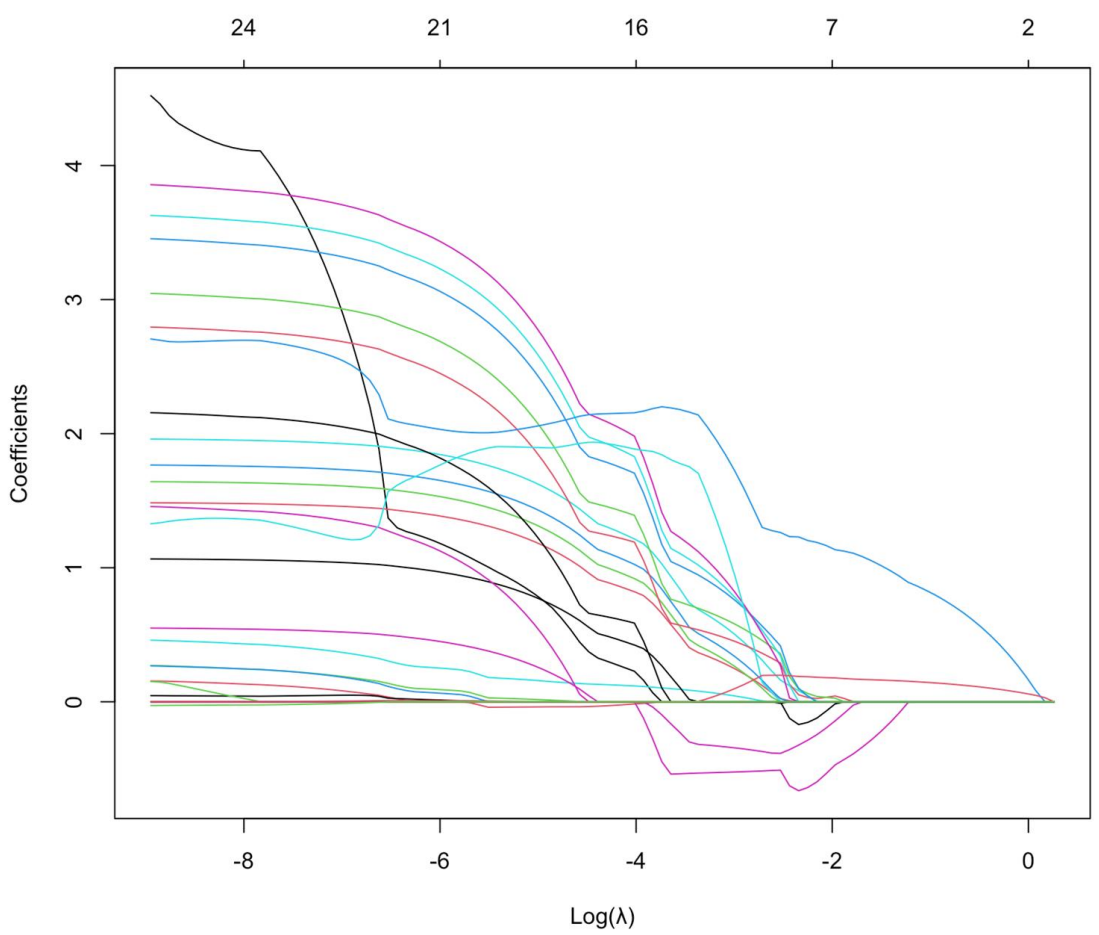
Test RMSE	Best Lambda
0.6561	0.1297

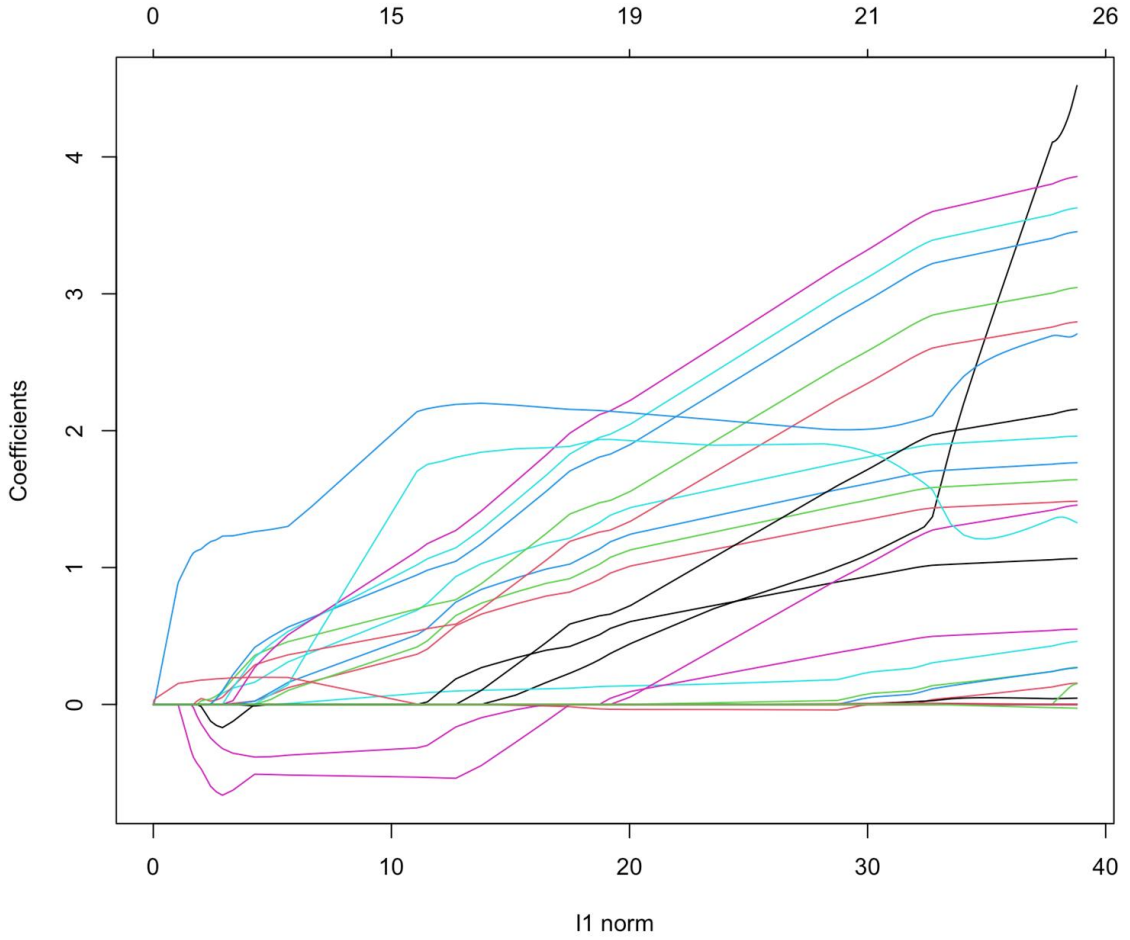
RIDGE REGRESSIONS

(Intercept)	-11.995703958
carat	0.781426147
cutGood	0.001953673
cutVery Good	0.106252853
cutPremium	0.018943217
cutIdeal	0.211956939
colorI	-0.158297174
colorH	0.306809962
colorG	0.683651996
colorF	0.763147513
colorE	0.858851954
colorD	1.037153314
claritySI2	-0.351379294
claritySI1	0.252856042
clarityVS2	0.816350151
clarityVS1	0.998008570
clarityVVS2	1.246661377
clarityVVS1	1.373172602
clarityIF	1.552750051
depth_percentage	0.008646613
table	0.004146025
length	0.385713956
width	0.594554568
depth	0.800182699
length:width	0.033571121
length:depth	0.046829528
width:depth	0.059182201
length:width:depth	0.004050587

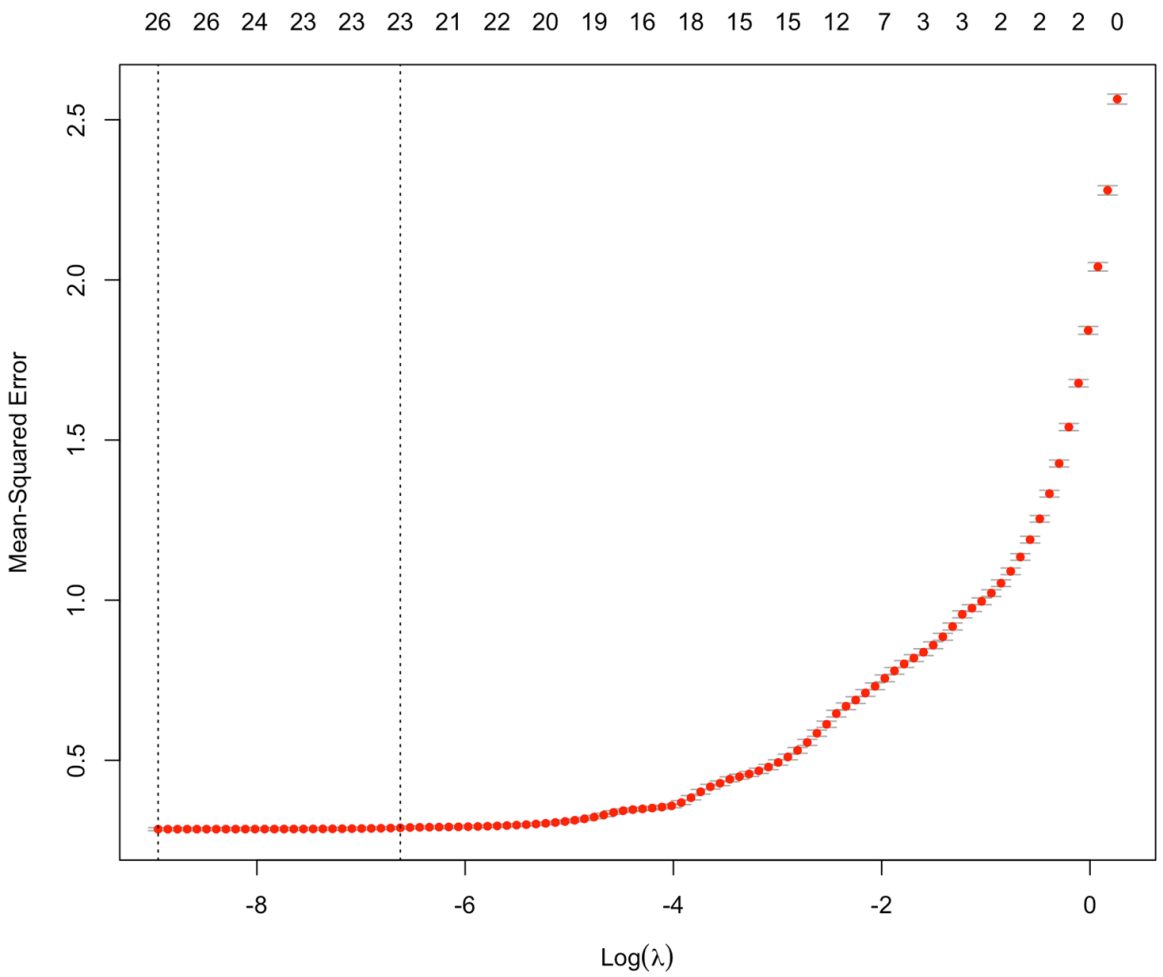
- With ridge regression coefficients has values closer to zero.

LASSO REGRESSIONS





LASSO REGRESSIONS



- The lasso regression model performed better with lower values of lambda.

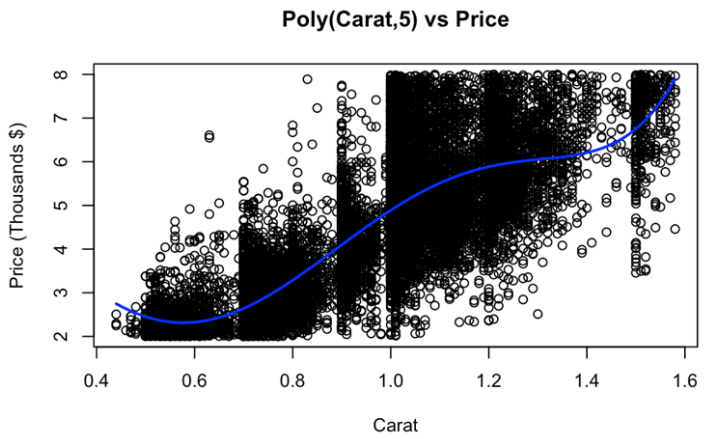
Test RMSE	Best Lambda
0.5466	0.0001

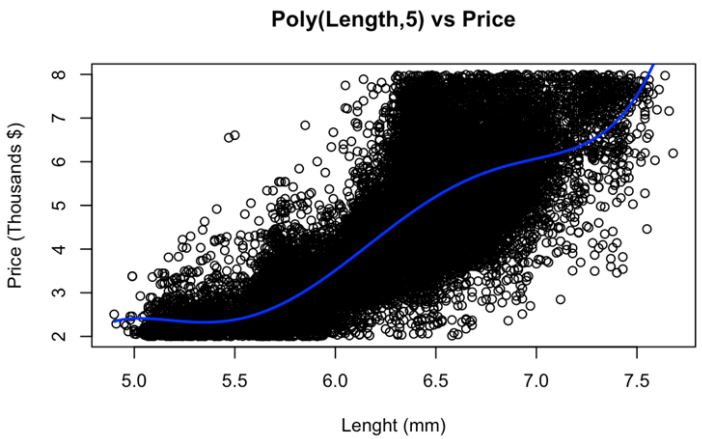
LASSO REGRESSIONS

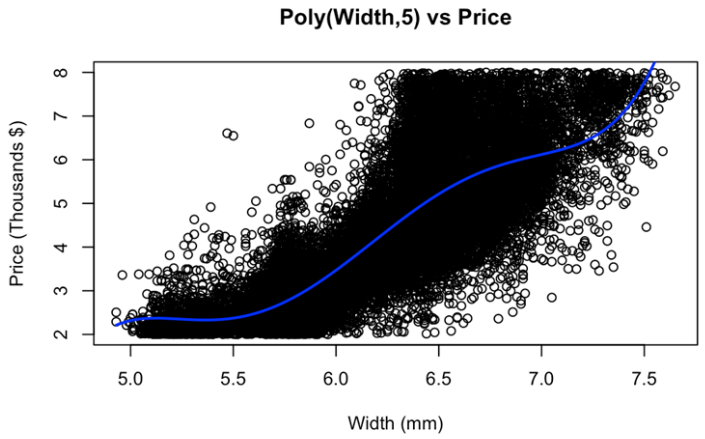
(Intercept)	-2.831053e+01
carat	5.095065e+00
cutGood	1.487364e-01
cutVery Good	2.649041e-01
cutPremium	2.725313e-01
cutIdeal	4.591402e-01
colorI	5.485587e-01
colorH	1.062152e+00
colorG	1.480463e+00
colorF	1.639001e+00
colorE	1.763329e+00
colorD	1.957580e+00
claritySI2	1.455067e+00
claritySI1	2.154946e+00
clarityVS2	2.794565e+00
clarityVS1	3.047025e+00
clarityVVS2	3.463769e+00
clarityVVS1	3.637685e+00
clarityIF	3.868710e+00
depth_percentage	6.912160e-02
table	-2.237530e-03
length	1.018746e+00
width	3.247245e+00
depth	-1.324668e-01
length:width	-1.437414e-01
length:depth	-5.625516e-06
width:depth	1.548645e-01
length:width:depth	-2.973441e-02

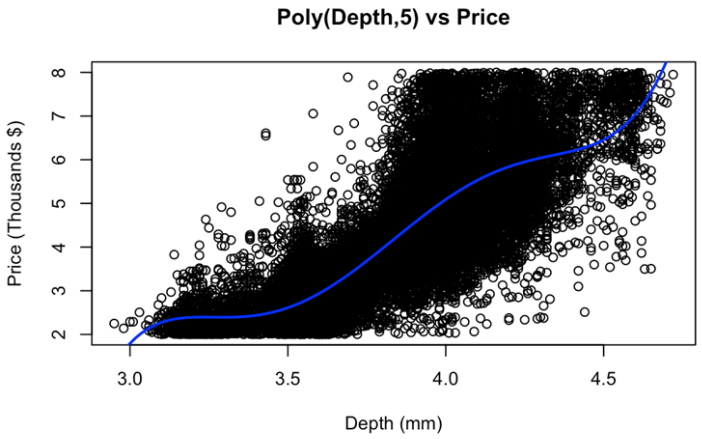
- The value of lambda chosen by the model doesn't brings coefficients to zero.

POLYNOMIAL REGRESSION



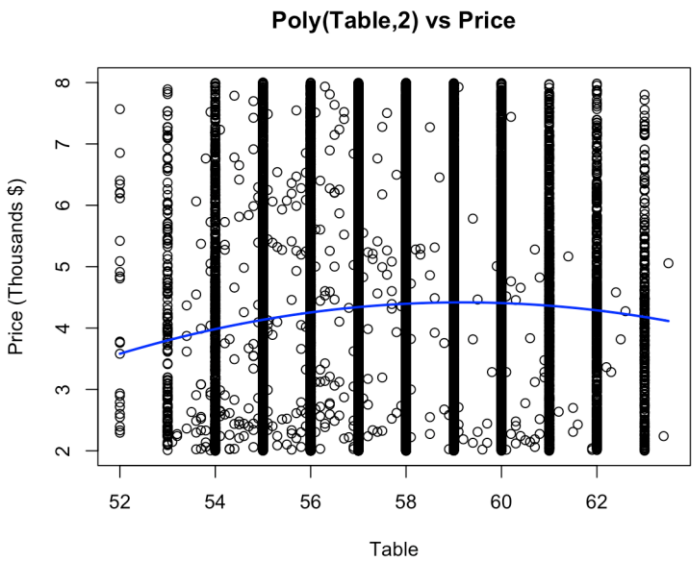


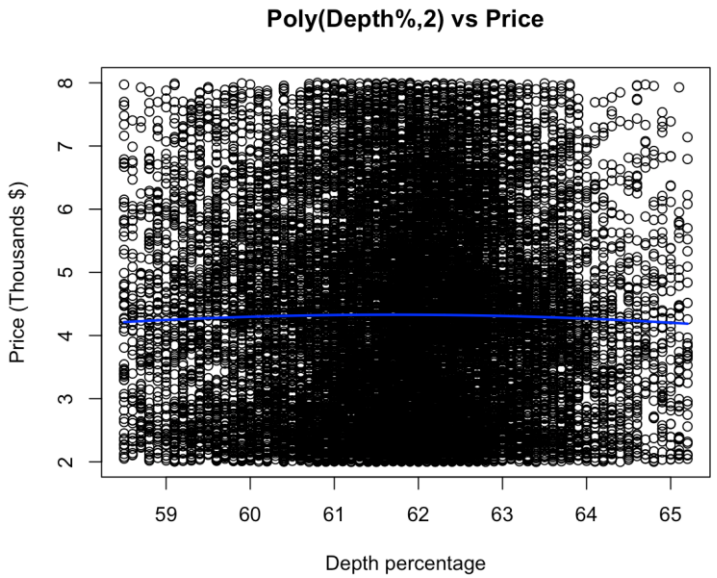




- To perform a multiple polynomial regression, searching best poly grade for each feature.
- Comparison of polynomials from 1 to 5 and select the best by Anova test.
- Plots variables vs response with chosen poly degree.

POLYNOMIAL REGRESSION





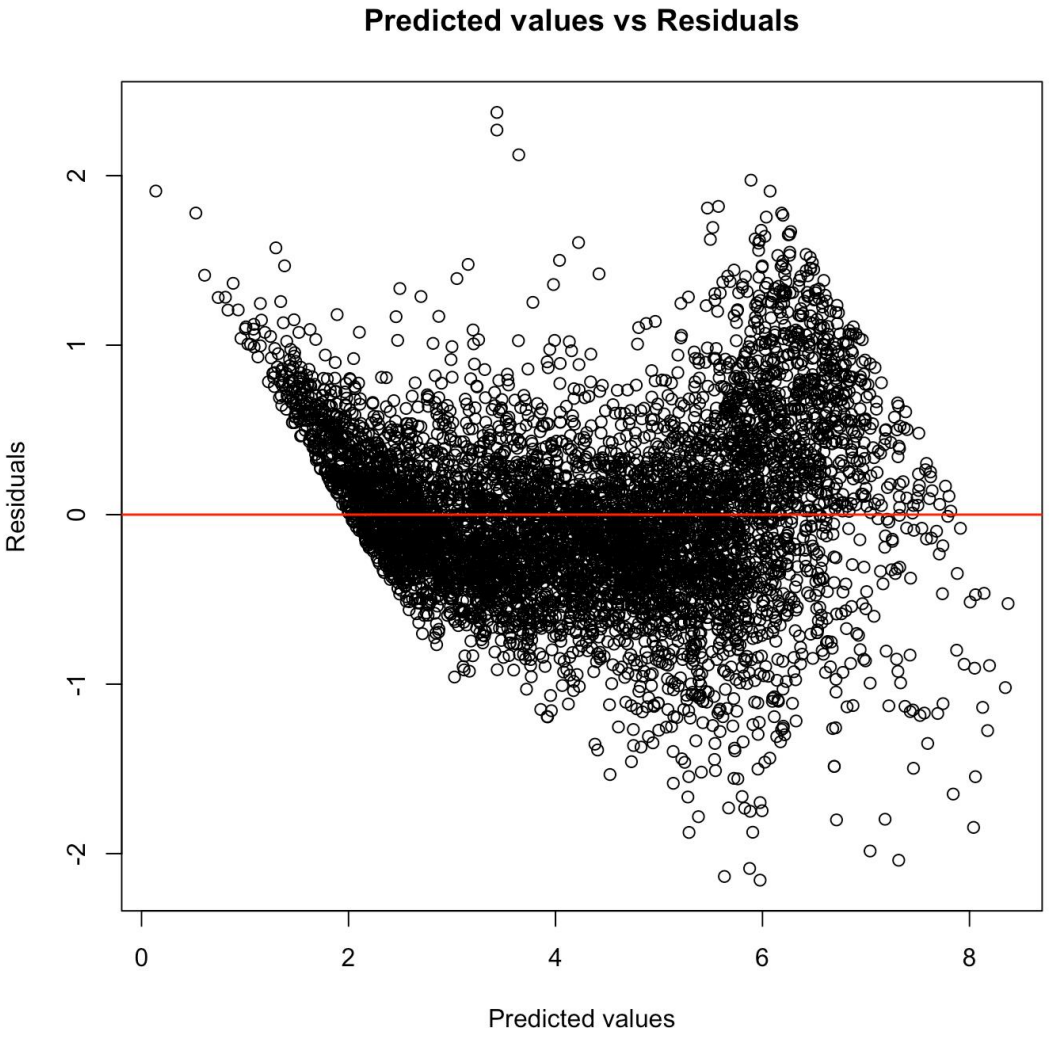
- The test RMSE is better compared to the linear models.

To fit multiple polynomial regression, the chosen orders are:

- Carat, Length, Depth and Width ~ ord(5)
- Table and Depth% ~ ord(2)

Test RMSE
0.5216

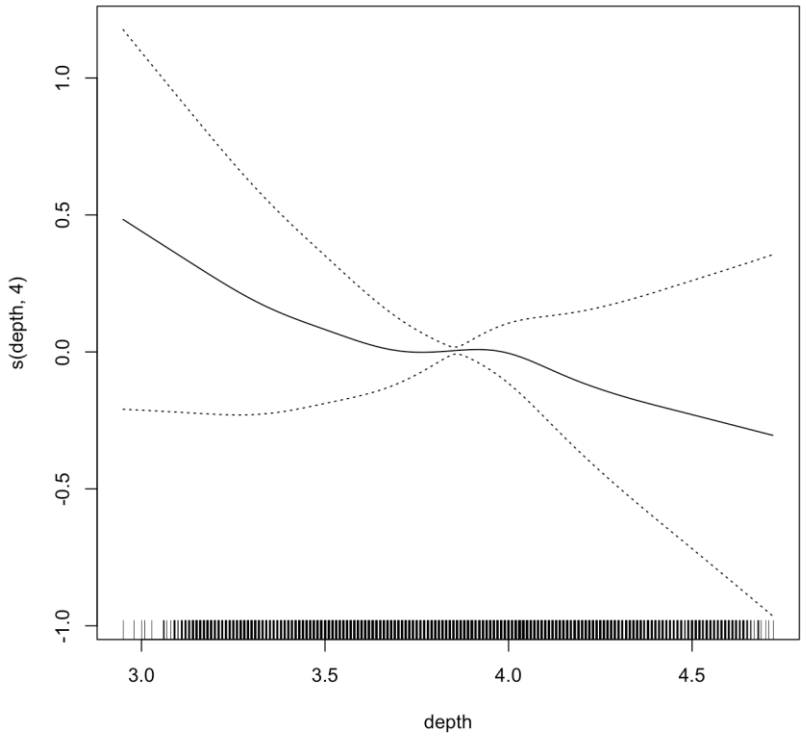
GENERALIZED ADDITIVE MODELS

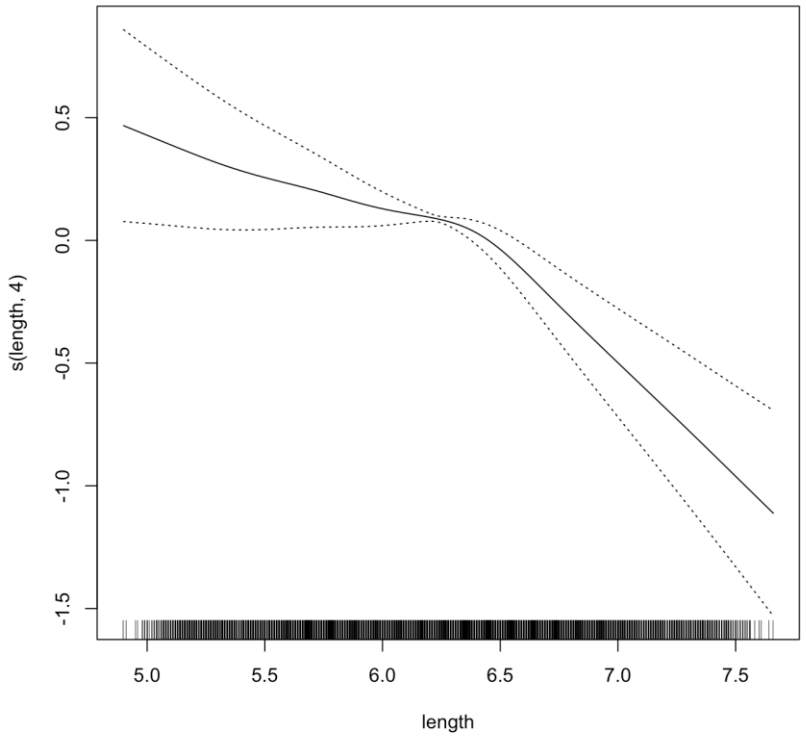


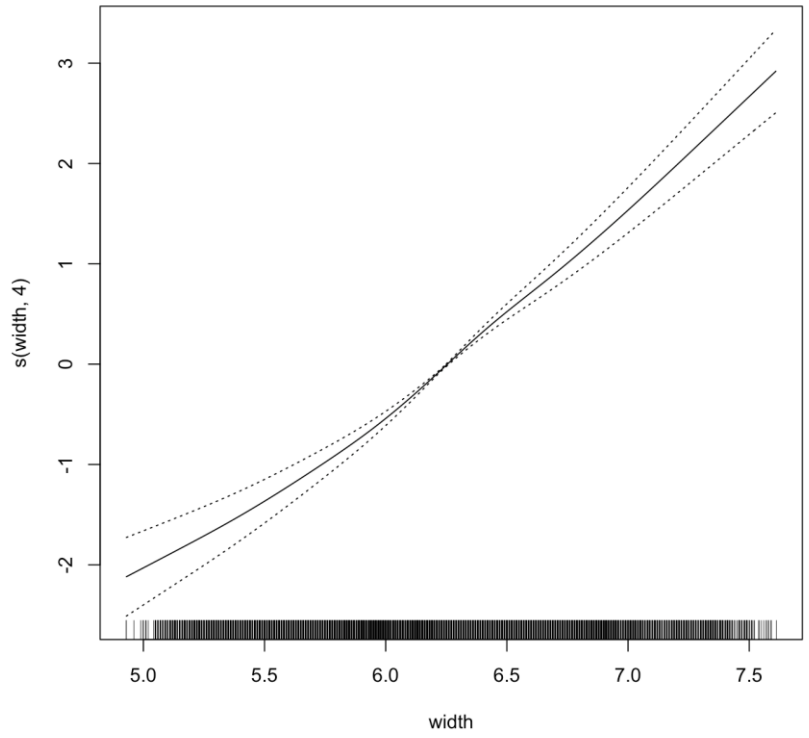
- GAM model with smooth spline.
- Test RMSE better then the one obtained with polynomial and linear regression.

Test RMSE
0.5151

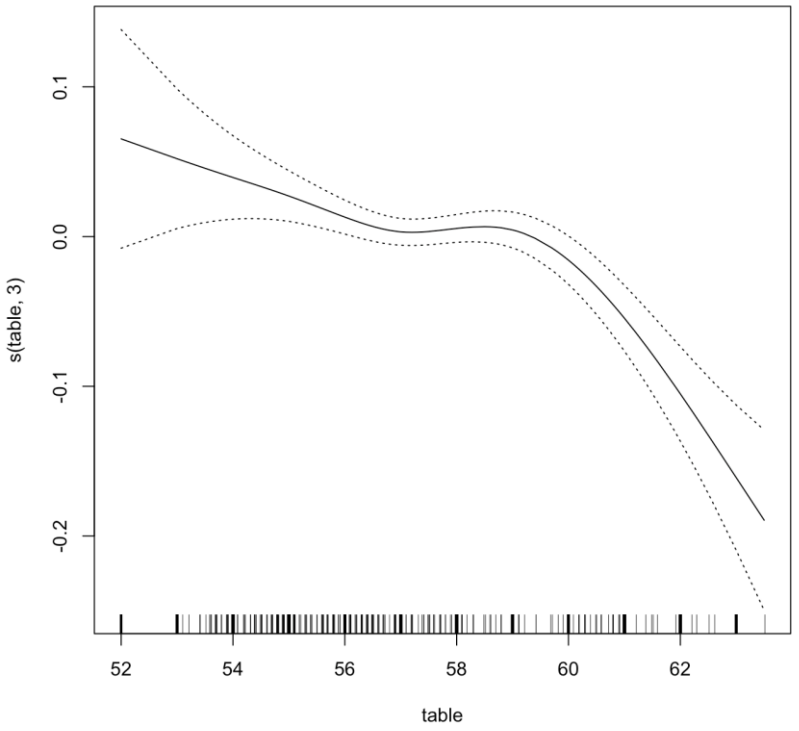
GENERALIZED ADDITIVE MODELS

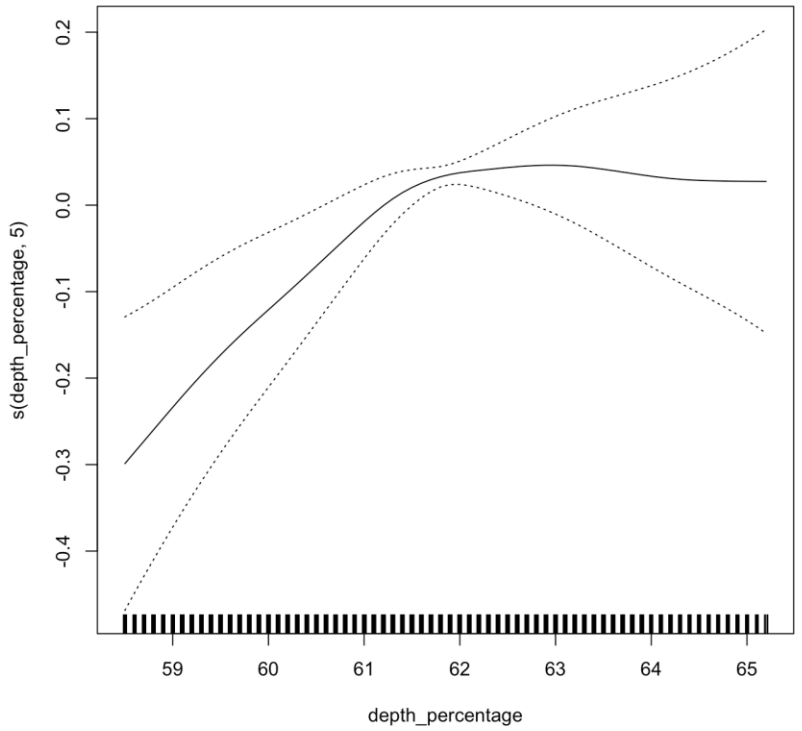


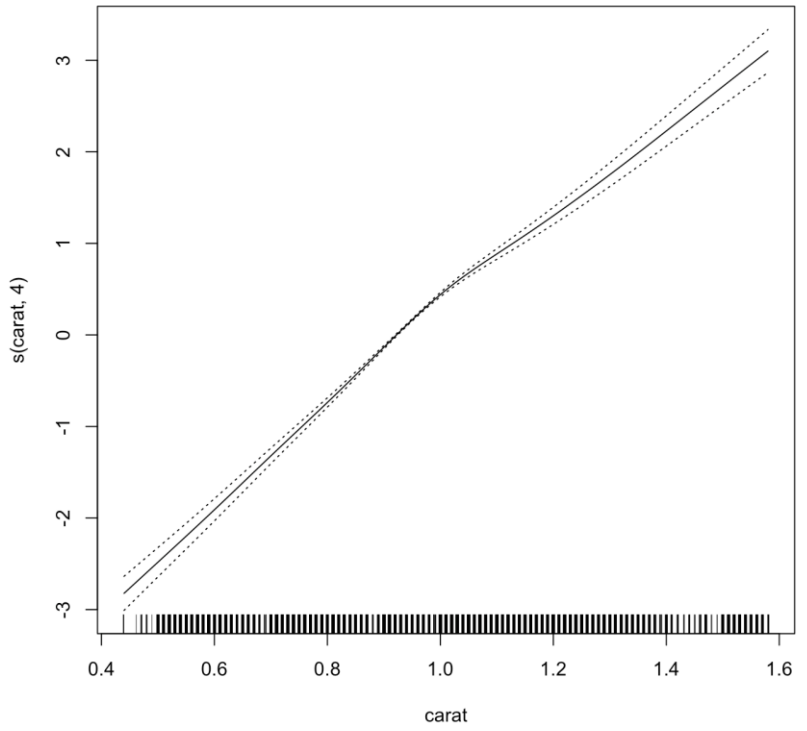




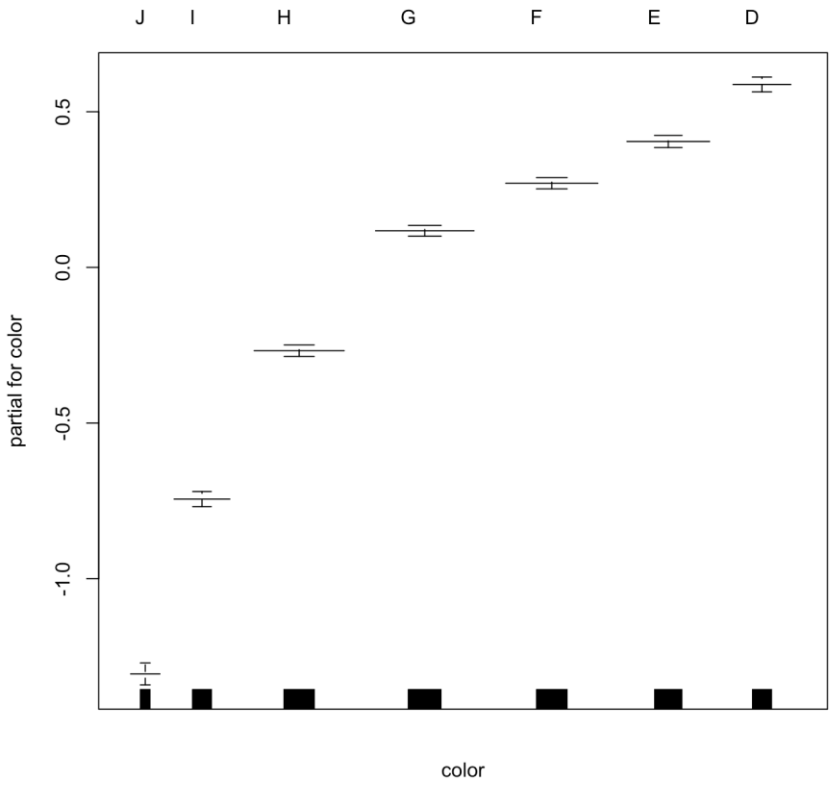
GENERALIZED ADDITIVE MODELS

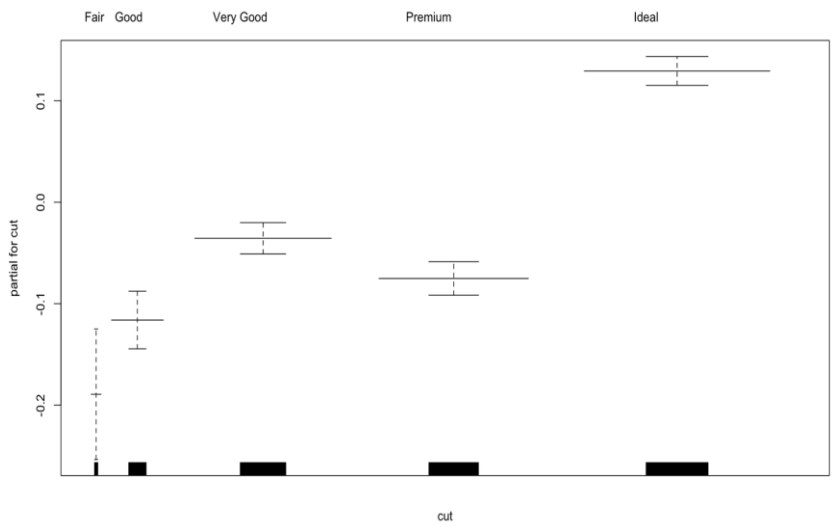


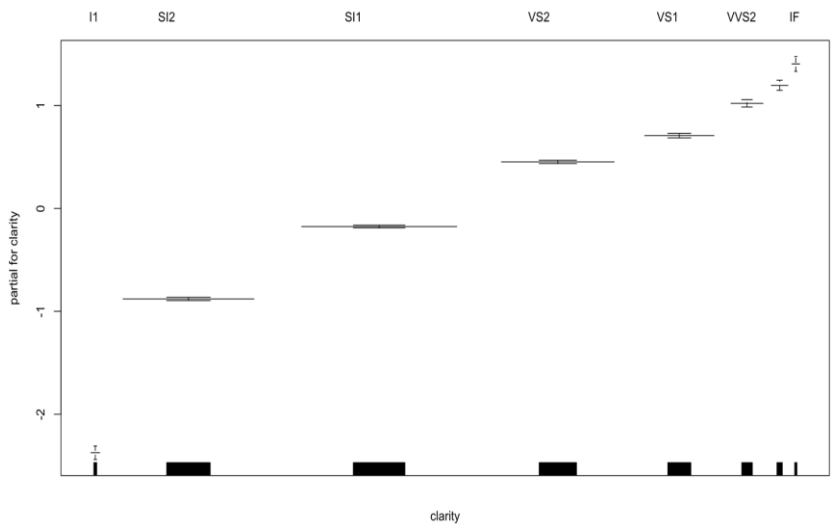




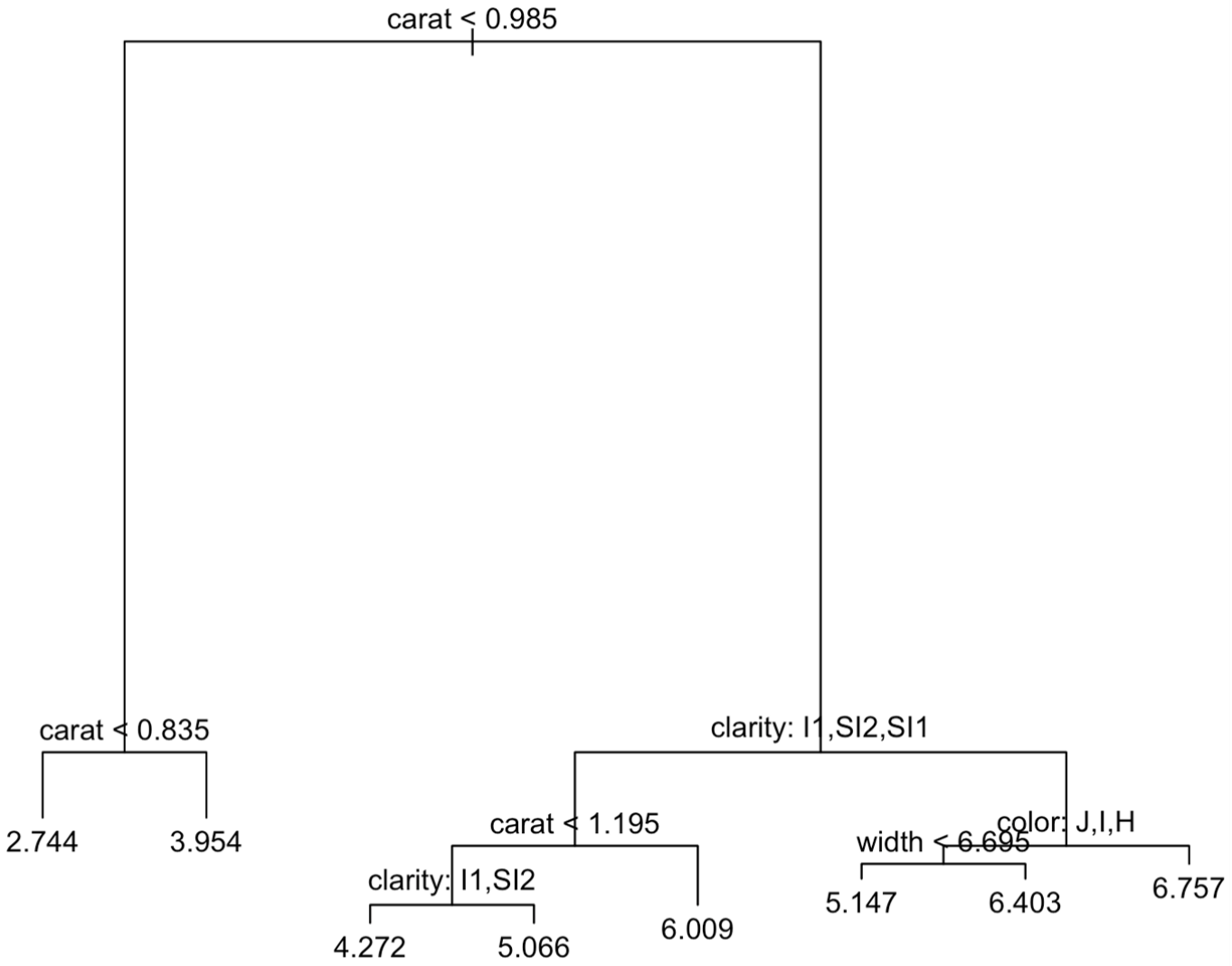
GENERALIZED ADDITIVE MODELS







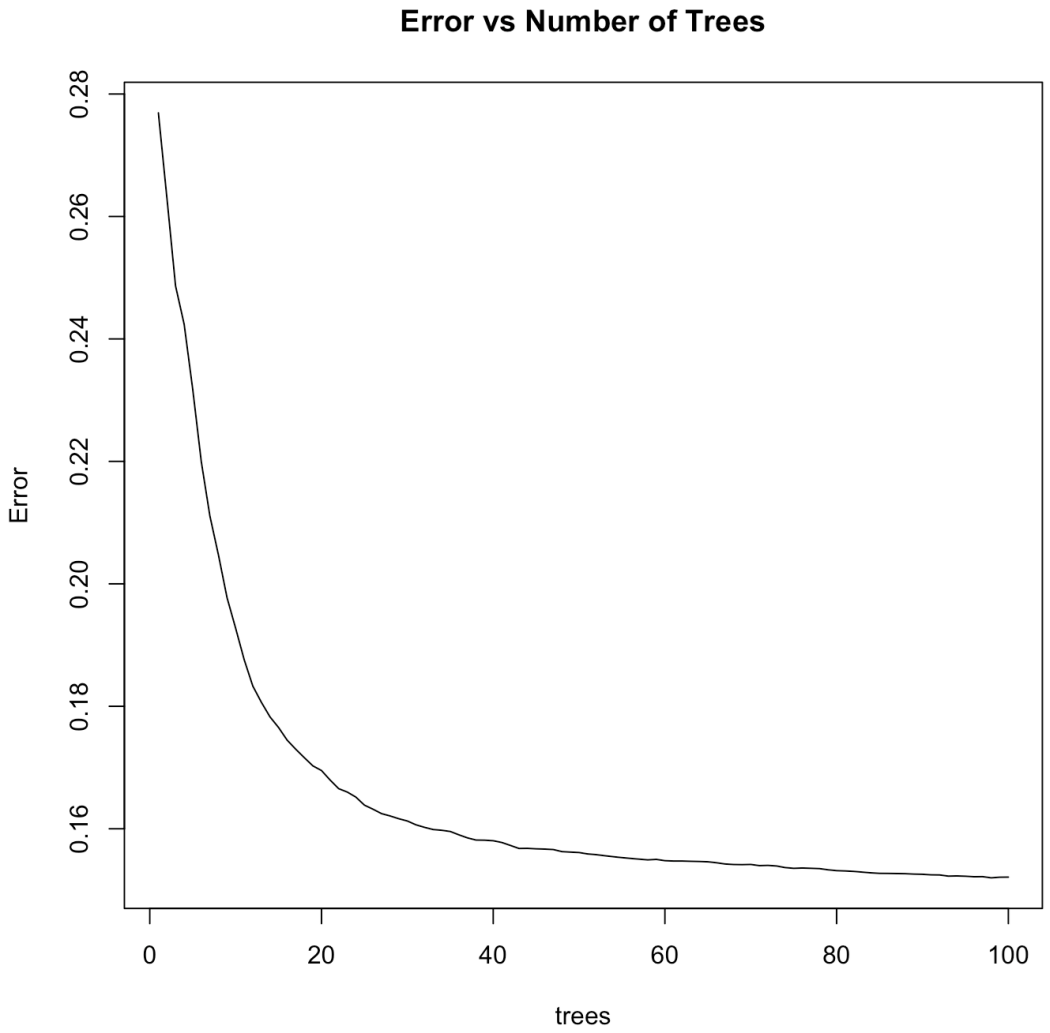
REGRESSION TREES



- Regression tree.
- Tried with pruning (crossvalidation pruning) but didn't changed anything cause we already had the best number of leaves.

Test RMSE
0.7177

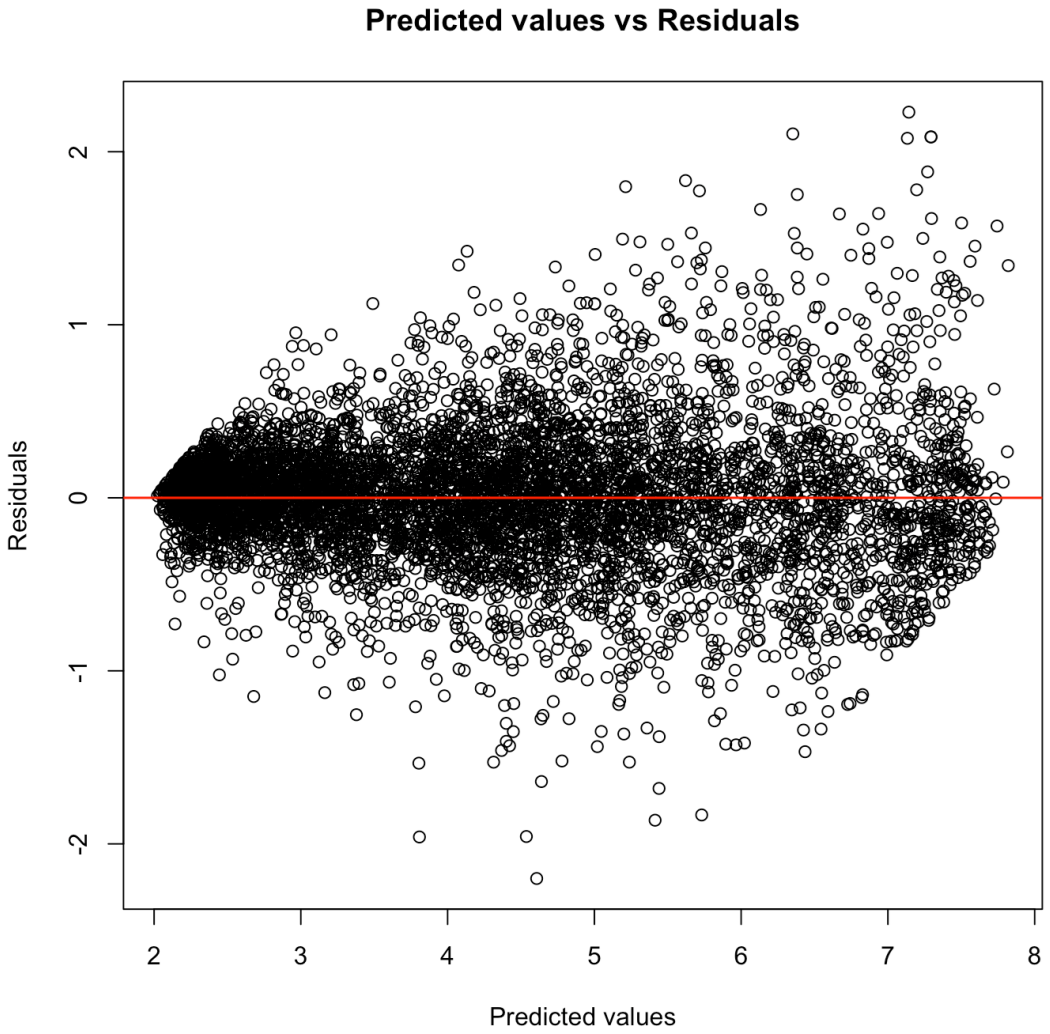
BAGGING



- Performed bagging with 100 trees.
- The test RMSE is significantly better compared to the other models.

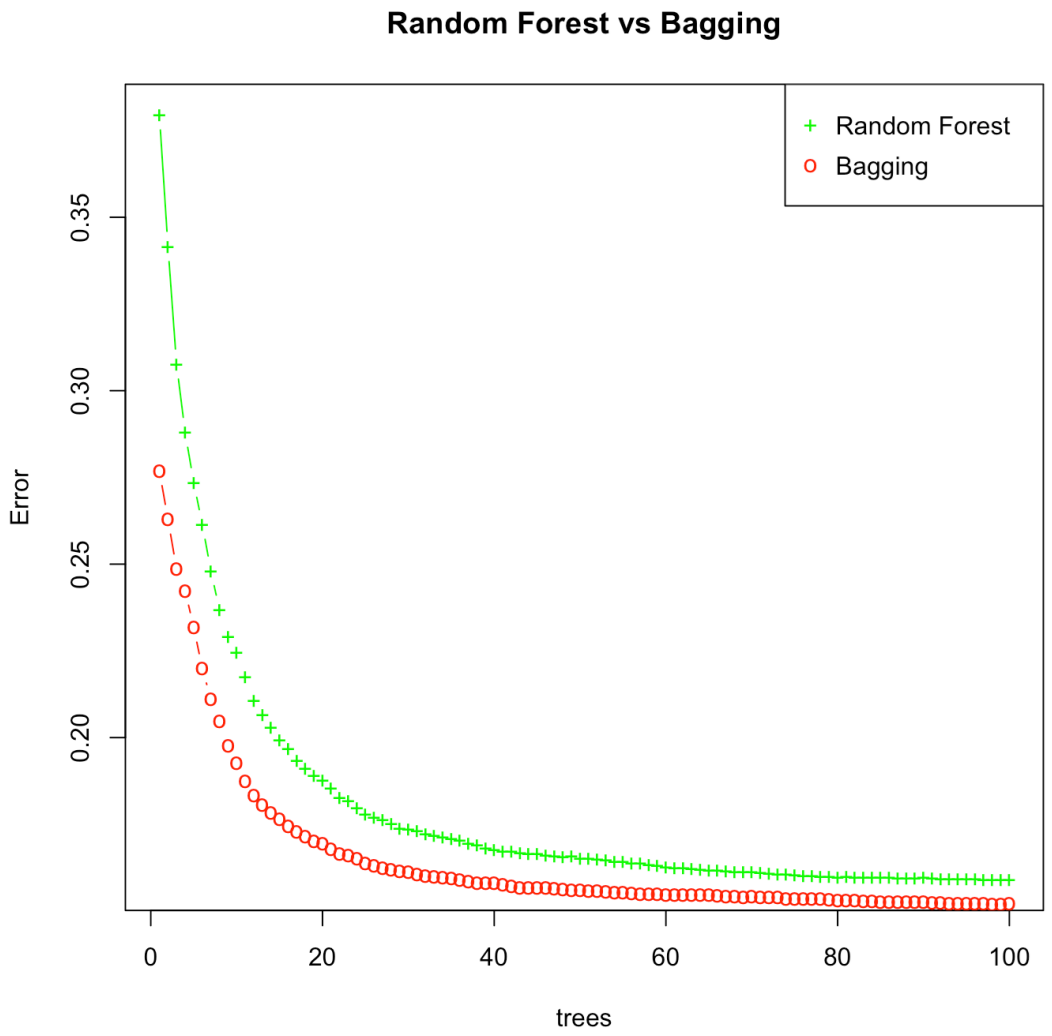
Test RMSE
0.4015

BAGGING



•The graph shows how more accurate are the predictions compared with the linear models.

BAGGING VS RANDOM FORESTS

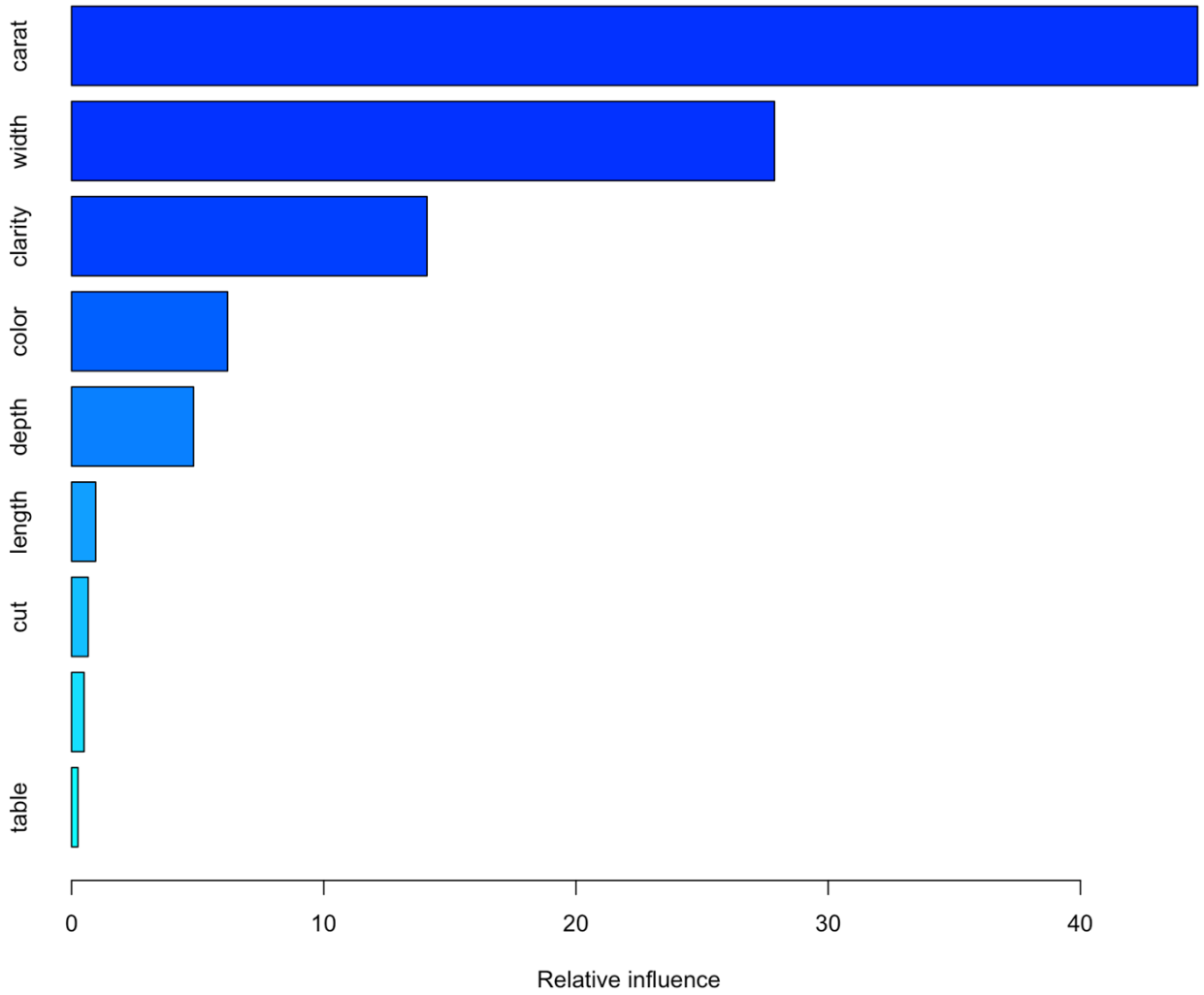


- Random forest performed with m = sqrt(p)
- Comparison between bagging and Random forest

Test RMSE - Bagging
0.4015

Test RMSE – Random Forest
0.4064

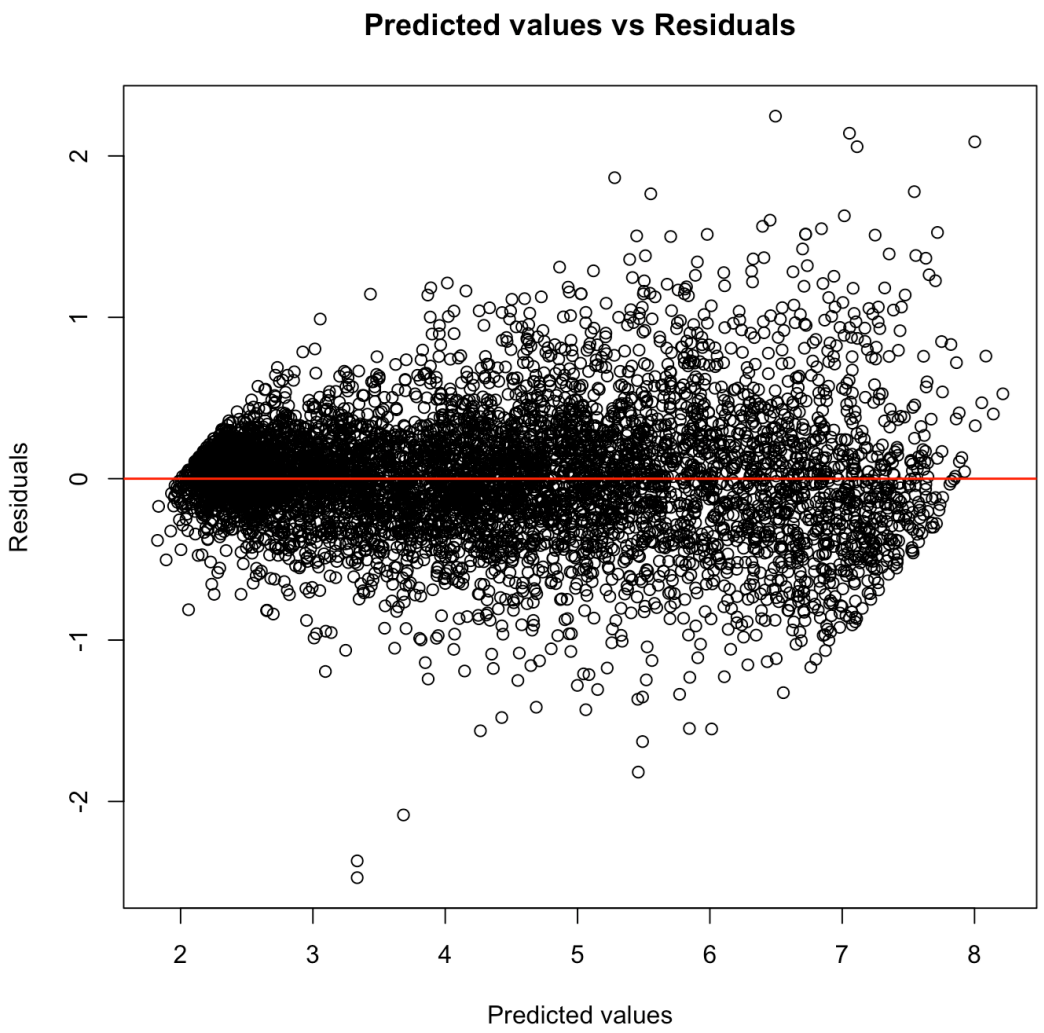
BOOSTING



- Boosting performed with 5000 trees with lambda = 0.02
- The graph shows the importance of the variables
- Best test RMSE obtained

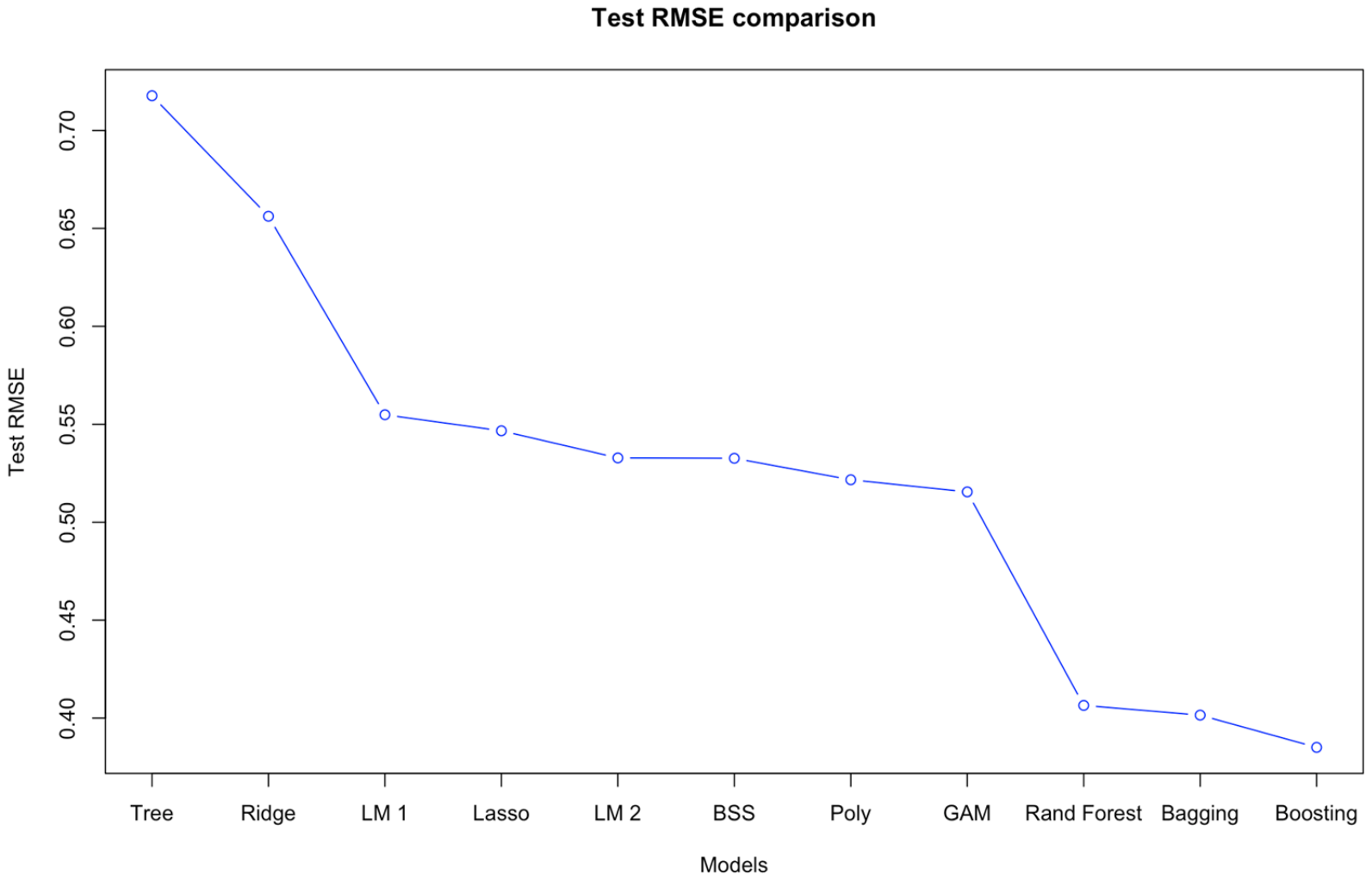
Test RMSE
0.3850

BOOSTING



•The graph shows how more accurate are the predictions compared with the linear models.

CONCLUSIONS



The graph shows the different Test RMSE obtained for each model.

CONCLUSIONS

- The best model is Boosting with an RMSE of 385 $
- Nonlinear models perform significantly better than linear models.
- The most significant variables to predict price are Carat, Clarity, Width and Color.

CLASSIFICATION

- Cut, Clarity and Color have a significant impact on the price of the diamond
- With these 3 categorical variables it’s been created a new binary variable:

Quality (High, Low)

- A diamond will have a "High" value only if it simultaneously presents high values of cut, color, clarity.

CLASSIFICATION GOALS

- Idea: Using only the remaining regressor try to predict for each diamond, it it belongs to the ‘high’ field or to the ‘low’ field
- The remaining regressors are: carat, price, Depth_Percentage, Table, Width, Length, Depth)

- Which predictor has the greatest impact on classification?
- What is the boundary (linear, radial or polynomial) that separete the two classes?

PREPROCESSING STEPS

- Creation of the new "quality" variable starting from the categorical variables color, cut, clarity.
- Since the 2 classes are strongly unbalanced towards low, we apply an **undersampling technique** with ovun.sample().

The final dataset has 6037 entry.

- Splitting dataset in training set and test set (70%-30%).

LOGISTIC REGRESSION

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        7.98075   23.08351   0.346   0.7295
carat            -30.41822    4.13093  -7.364 1.79e-13 ***
depth_percentage   0.22945    0.36128   0.635   0.5254
table             -0.01858    0.03092  -0.601   0.5479
price              2.99360    0.10708  27.955  < 2e-16 ***
length             4.26744    2.36592   1.804   0.0713 .
width             -2.37006    1.84681  -1.283   0.1994
depth             -5.25391    6.20313  -0.847   0.3970
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5854.3  on 4222  degrees of freedom
Residual deviance: 2301.9  on 4215  degrees of freedom
AIC: 2317.9

Number of Fisher Scoring iterations: 7
```

- The only significant coefficients are carat and price.

	Low	High
Low	786	79
High	119	827

Accuracy
89.06%

KNN

- Since dataset still has a lot of observations it's been used a **k-nearest neighbors**
- 10-fold cross-validation approach

best k = 5

- KNN performs worse than logistic regression

	Low	High
Low	768	107
High	137	799

Accuracy
86.52%

SUPPORT VECTOR MACHINES

- SVM with linear kernel
- To find the best cost parameter, it's been implemented cross-validation with tune()

best cost = 5

	Low	High
Low	784	69
High	121	837

Accuracy
89.50%

SUPPORT VECTOR MACHINES

- SVM with radial kernel.
- To find the best cost and gamma, we have implemented cross-validation with tune().
- The best parameters are:

cost = 5

gamma = 0.5

	Low	High
Low	788	64
High	117	842

Accuracy
90.00%

SUPPORT VECTOR MACHINES

- SVM with polynomial kernel.
- Better result with

degree = 3

cost = 5

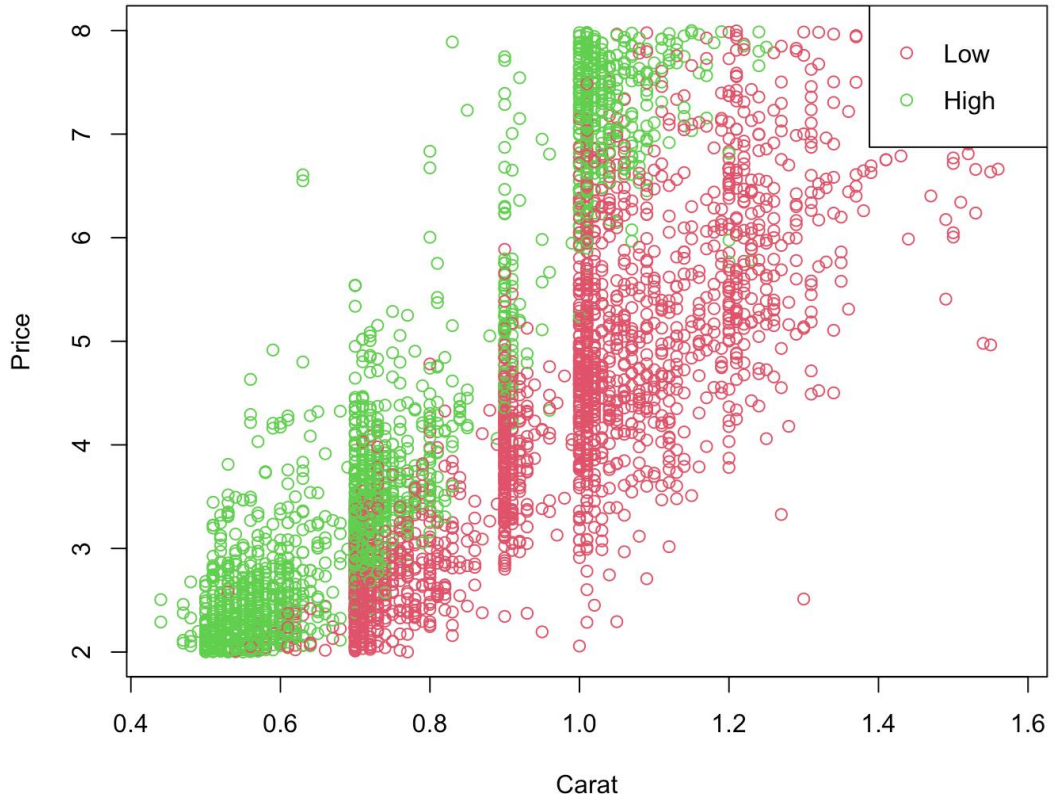
gamma = 0.5

- The model perform worse than previous.

	Low	High
Low	792	105
High	113	801

Accuracy
87.96%

SUPPORT VECTOR MACHINES



- Thanks to Logistic Regression it's clear that **Carat** and **Price** are the most significant variables
- Can a SVM model (just with these two predictors) classify diamond's quality?
- Is linear kernel precise enough?

SUPPORT VECTOR MACHINES

- The SVM model works better compared to linear SVM with all predictors.
- The linear kernel is precise enough so it’s not necessary to use radial or polynomial kernel.

	Low	High
Low	788	71
High	117	835

Accuracy
89.61%

CONCLUSIONS

- The most important predictors required to correctly classify the quality of each diamond are **price** and **carat**
- SVM models with linear boundaries are sufficiently accurate.

THANK YOU FOR THE ATTENTION!