Daniel Prusinski
Predict 412

Appendix 1: Data Visualizations and Important Outputs - Evaluating Regression Models in R

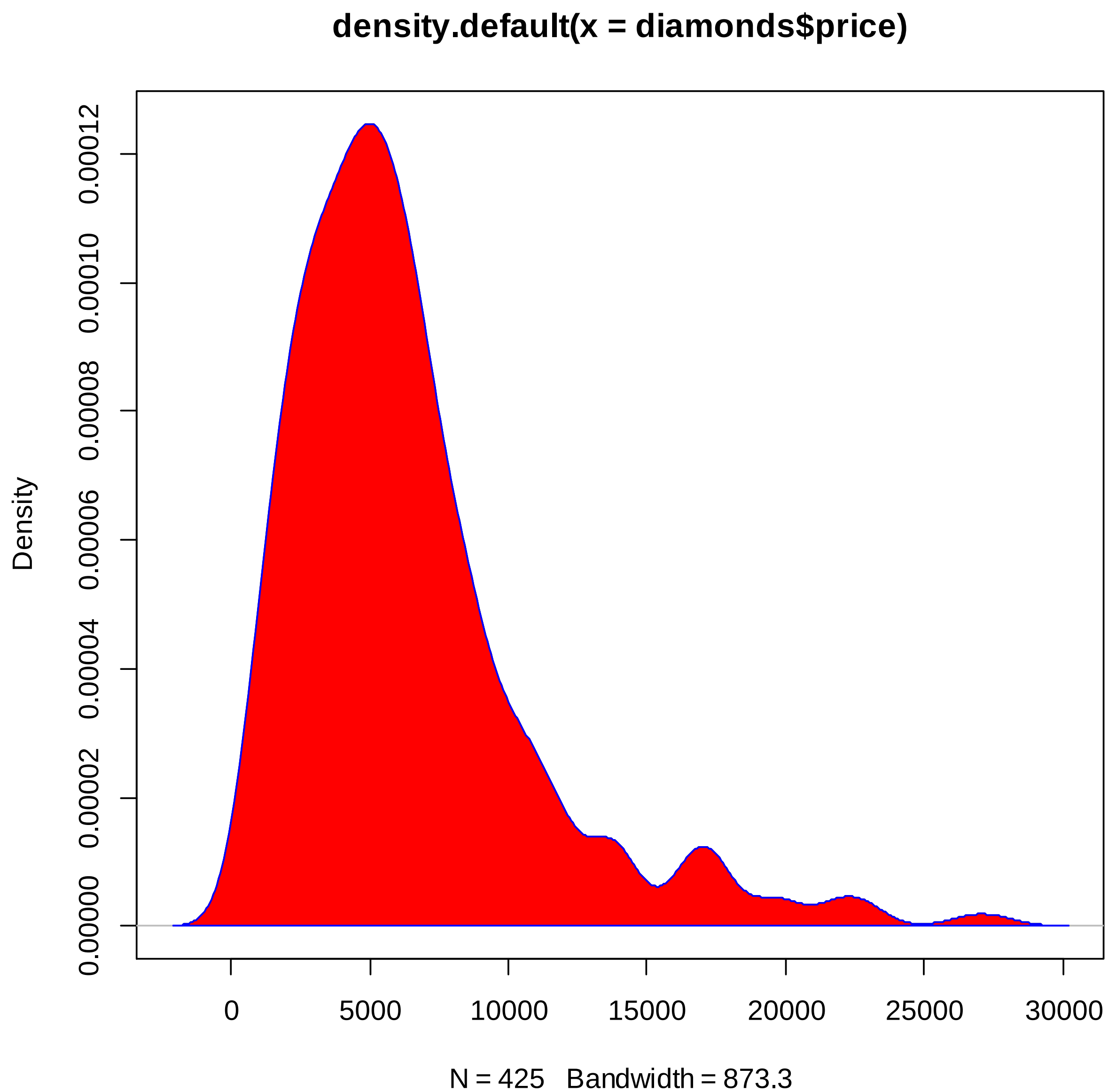
Output 1: Results from > print(str(diamonds))

```
'data.frame':   425 obs. of  7 variables:
 $ carat  : num  0.826 0.996 1.07 1.07 1.01 0.66 0.701 0.97 0.74 2.04 ...
 $ color  : int  4 5 4 7 8 3 4 8 1 5 ...
 $ clarity: int  7 6 7 7 6 4 8 6 9 6 ...
 $ cut    : Factor w/ 2 levels "Ideal","Not Ideal": 1 1 1 2 2 1 1 2 2 2 ...
 $ channel: Factor w/ 3 levels "Independent",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ store  : Factor w/ 12 levels "Ashford","Ausmans",..: 7 7 7 7 7 7 7 7 7 7 ...
 $ price  : int  7775 9850 10950 7500 6995 6100 6300 4850 5895 23000 ...
```

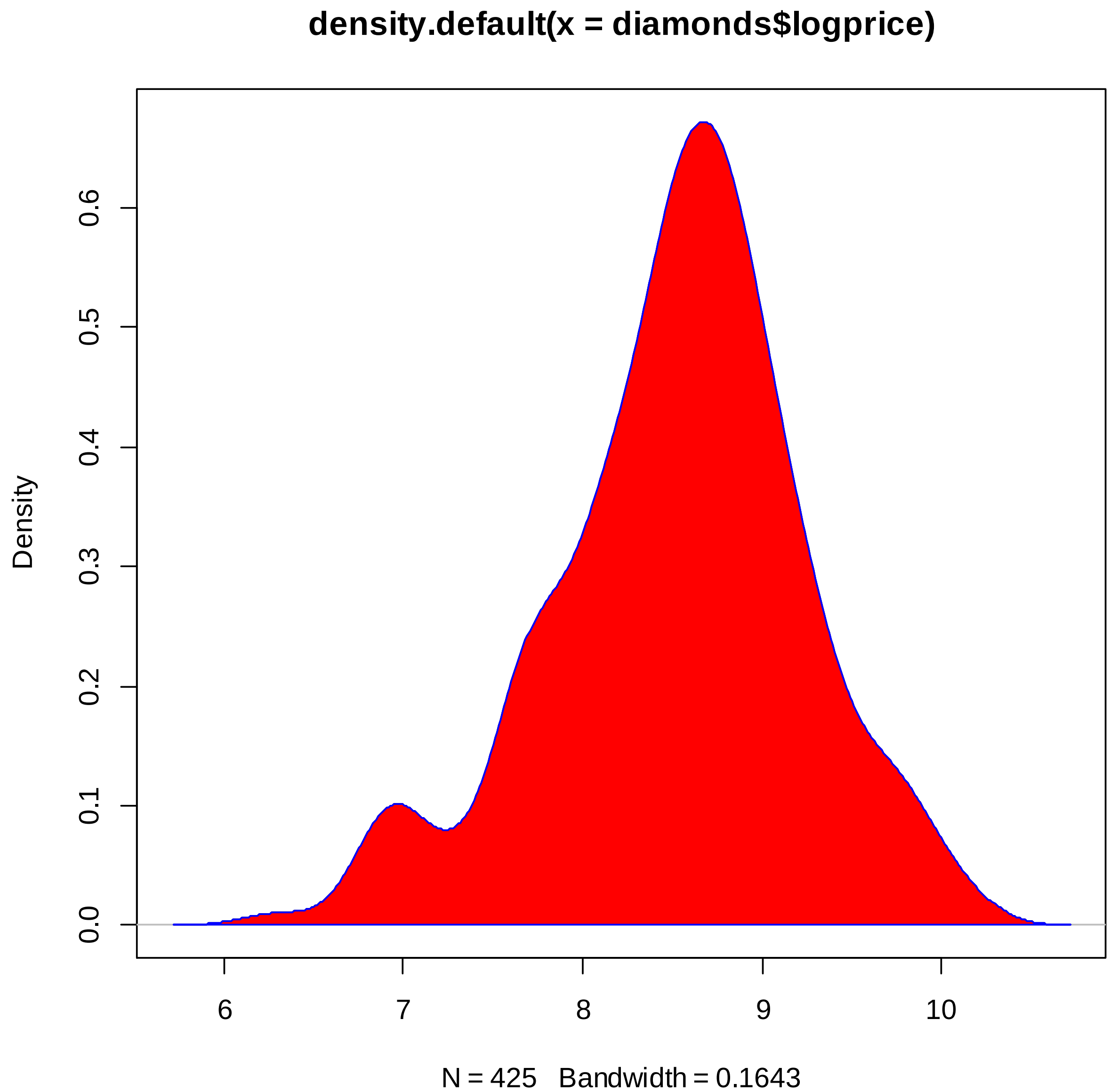
Output 2: Results from > summary(diamonds)

```
 carat            color            clarity            cut
 Min.   :0.200   Min.   :1.000   Min.   : 2.000   Ideal    :154
 1st Qu.:0.720   1st Qu.:3.000   1st Qu.: 5.000   Not Ideal:271
 Median :1.020   Median :4.000   Median : 6.000
 Mean   :1.041   Mean   :4.313   Mean   : 6.134
 3rd Qu.:1.210   3rd Qu.:6.000   3rd Qu.: 7.000
 Max.   :2.480   Max.   :9.000   Max.   :10.000

        channel          store        price
 Independent: 48   Blue Nile :211   Min.   :  497
 Internet   :318   Ashford   :107   1st Qu.: 3430
 Mall       : 59   Riddles   : 16   Median : 5476
                   Fred Meyer: 15   Mean   : 6356
                   Kay       : 14   3rd Qu.: 7792
                   University: 13   Max.   :27575
                   (Other)   : 49
```

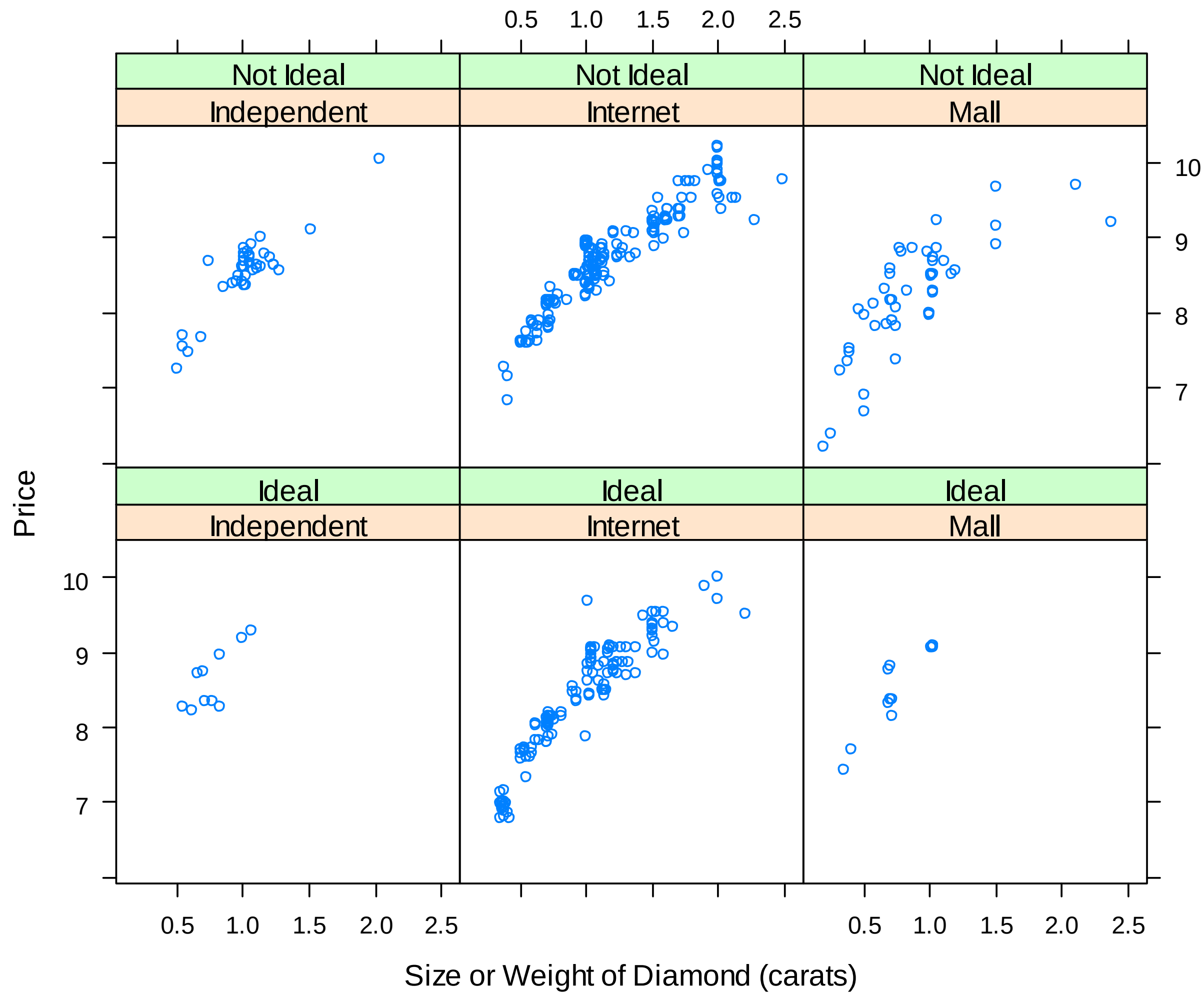
Output 3: Density Plot of Price



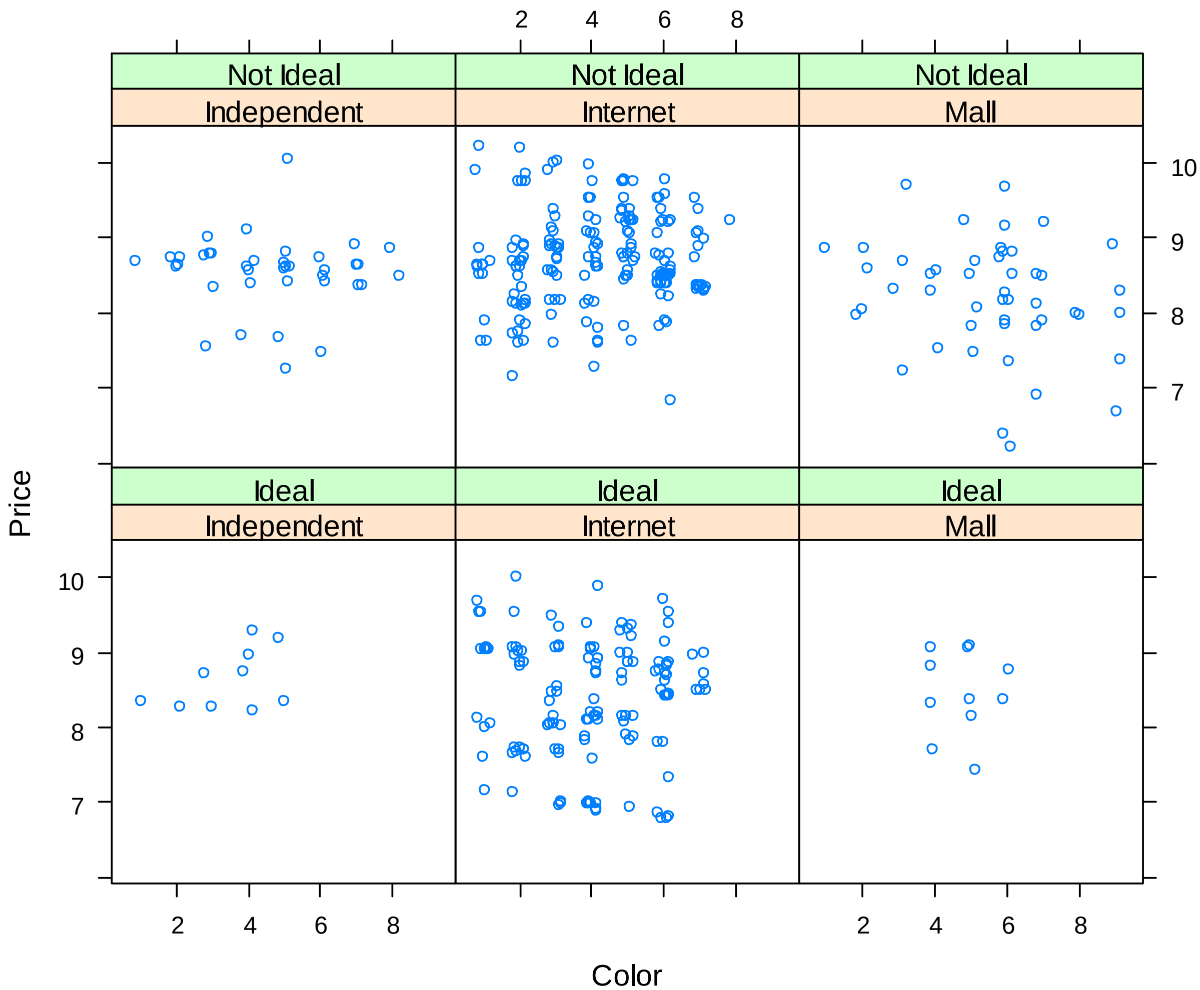
Output 4: Density Plot after Log Transformation



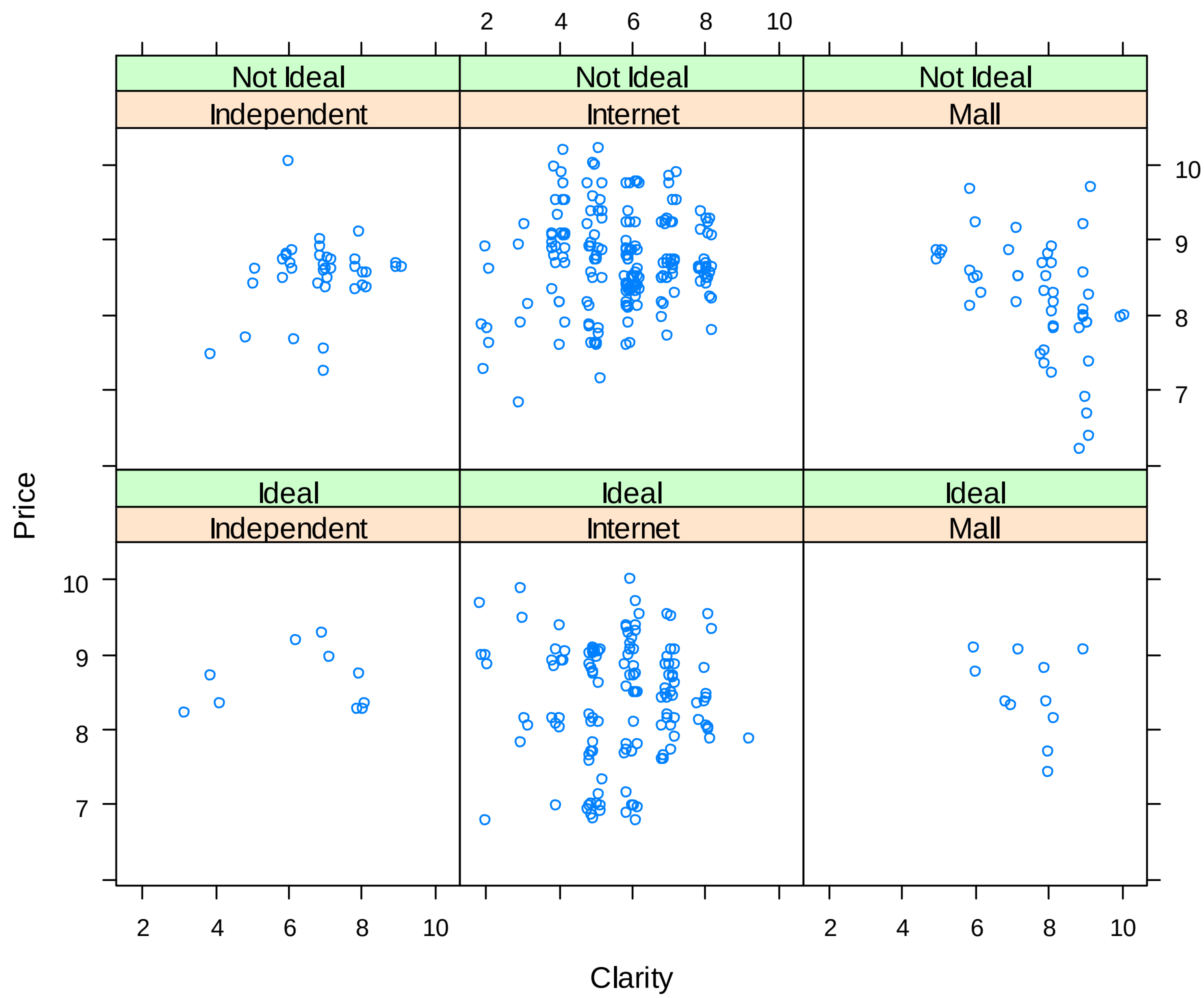
Output 5: Scatter plot of Price and Carat



Output 6: Scatter plot of Price and Color



Output 7: Scatter plot of Price and Clarity



Output 8: Dividing the data into training and testing

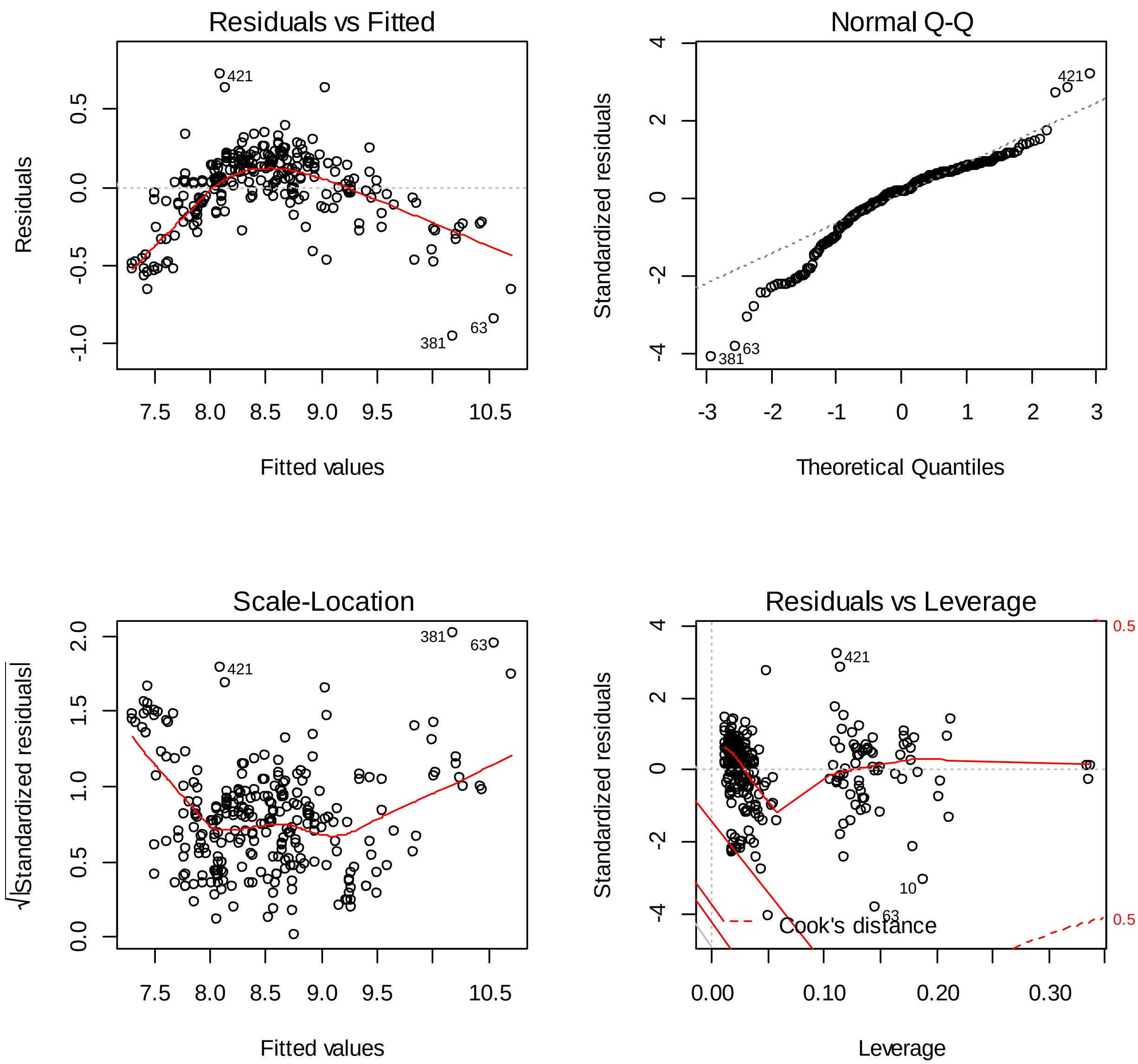
```
> print(str(diamonds.train))
'data.frame':   283 obs. of  9 variables:
 $ carat   : num  0.996 1.07 1.07 1.01 0.66 ...
 $ color   : int  5 4 7 8 3 5 4 5 6 1 ...
 $ clarity : int  6 7 7 6 4 6 8 7 6 8 ...
 $ cut     : Factor w/ 2 levels "Ideal","Not Ideal": 1 1 2 2 1 2 2 2 2 1 ...
 $ channel : Factor w/ 3 levels "Independent",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ store   : Factor w/ 12 levels "Ashford","Ausmans",..: 7 7 7 7 7 7 4 4 4 4 ...
 $ price   : int  9850 10950 7500 6995 6100 23000 5234 5375 6171 4256 ...
 $ Group   : Factor w/ 2 levels "TRAIN","TEST": 1 1 1 1 1 1 1 1 1 1 ...
```

```
 $ logprice: num  9.2 9.3 8.92 8.85 8.72 ...
NULL
> diamonds.test <- diamonds[(diamonds$Group == "TEST"),]
> print(str(diamonds.test))
'data.frame':   142 obs. of  9 variables:
 $ carat   : num  0.826 0.701 0.97 0.74 0.545 0.82 1.01 1.02 0.87 0.59 ...
 $ color   : int  4 4 8 1 2 3 7 3 2 5 ...
 $ clarity : int  7 8 6 9 8 8 7 8 5 8 ...
 $ cut     : Factor w/ 2 levels "Ideal","Not Ideal": 1 1 2 2 1 1 2 2 2 2 ...
 $ channel : Factor w/ 3 levels "Independent",..: 1 1 1 1 1 1 3 3 3 3 ...
 $ store   : Factor w/ 12 levels "Ashford","Ausmans",..: 7 7 7 7 7 4 6 6 6 6 ...
 $ price   : int  7775 6300 4850 5895 3895 3878 5000 5999 6999 2495 ...
 $ Group   : Factor w/ 2 levels "TRAIN","TEST": 2 2 2 2 2 2 2 2 2 2 ...
 $ logprice: num  8.96 8.75 8.49 8.68 8.27 ...
NULL
```

<u>Output 9: Multiple Regression Model</u>

```
Call:
lm(formula = logprice ~ color + carat + clarity + cut + channel + 
    store, data = diamonds.train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.94997 -0.09176  0.04162  0.15439  0.72376 

Coefficients: (2 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)      7.657577   0.116103  65.955  < 2e-16 ***
color           -0.091983   0.008337 -11.034  < 2e-16 ***
carat            1.708077   0.039613  43.119  < 2e-16 ***
clarity         -0.067096   0.010143  -6.615 2.02e-10 ***
cutNot Ideal    -0.076257   0.033469  -2.278 0.023491 *  
channelInternet -0.142065   0.091736  -1.549 0.122657    
channelMall      0.402564   0.139231   2.891 0.004152 ** 
storeAusmans     0.227316   0.164387   1.383 0.167879    
storeBlue Nile   0.020562   0.036579   0.562 0.574508    
storeChalmers    0.041003   0.124407   0.330 0.741973    
storeDanford     0.090549   0.120583   0.751 0.453356    
storeFred Meyer -0.133345   0.134661  -0.990 0.322960    
storeGoodmans    0.496660   0.131991   3.763 0.000207 ***
storeKay        -0.151547   0.132568  -1.143 0.253993    
storeR. Holland -0.018980   0.147130  -0.129 0.897456    
storeRiddles    -0.141564   0.133824  -1.058 0.291084    
storeUniversity        NA         NA      NA       NA    
storeZales             NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.2397 on 267 degrees of freedom
Multiple R-squared:  0.8884,    Adjusted R-squared:  0.8821
F-statistic: 141.6 on 15 and 267 DF,  p-value: < 2.2e-16

confint(multiple.r.train)
                  2.5 %      97.5 %
(Intercept)     7.42898348  7.88617143
color          -0.10839684 -0.07556891
carat           1.63008240  1.78607125
clarity        -0.08706610 -0.04712683
cutNot Ideal    -0.14215365 -0.01036021
channelInternet -0.32268300  0.03855360
channelMall      0.12843464  0.67669404
storeAusmans    -0.09634414  0.55097586
storeBlue Nile  -0.05145896  0.09258286
storeChalmers   -0.20394142  0.28594664
storeDanford    -0.14686486  0.32796323
storeFred Meyer -0.39847868  0.13178806
storeGoodmans    0.23678409  0.75653625
storeKay        -0.41255914  0.10946441
storeR. Holland -0.30866143  0.27070228
storeRiddles    -0.40504798  0.12191985
storeUniversity          NA          NA
storeZales               NA          NA

          Sum Sq  Df   F value    Pr(>F)
color      6.995   1  121.7389 < 2.2e-16 ***
carat    106.829   1 1859.2166 < 2.2e-16 ***
clarity    2.515   1   43.7624 2.017e-10 ***
cut        0.298   1    5.1913   0.02349 *
channel    1.295   2   11.2724 1.996e-05 ***
store      1.143   9    2.2100   0.02175 *
Residuals  15.342 267
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

r.rmse <- sqrt(mean(multiple.r.train$residuals^2)) # Root Mean Square Error Calculation
> print (r.rmse) # I will compare this to the other models.
[1] 0.2328316
```

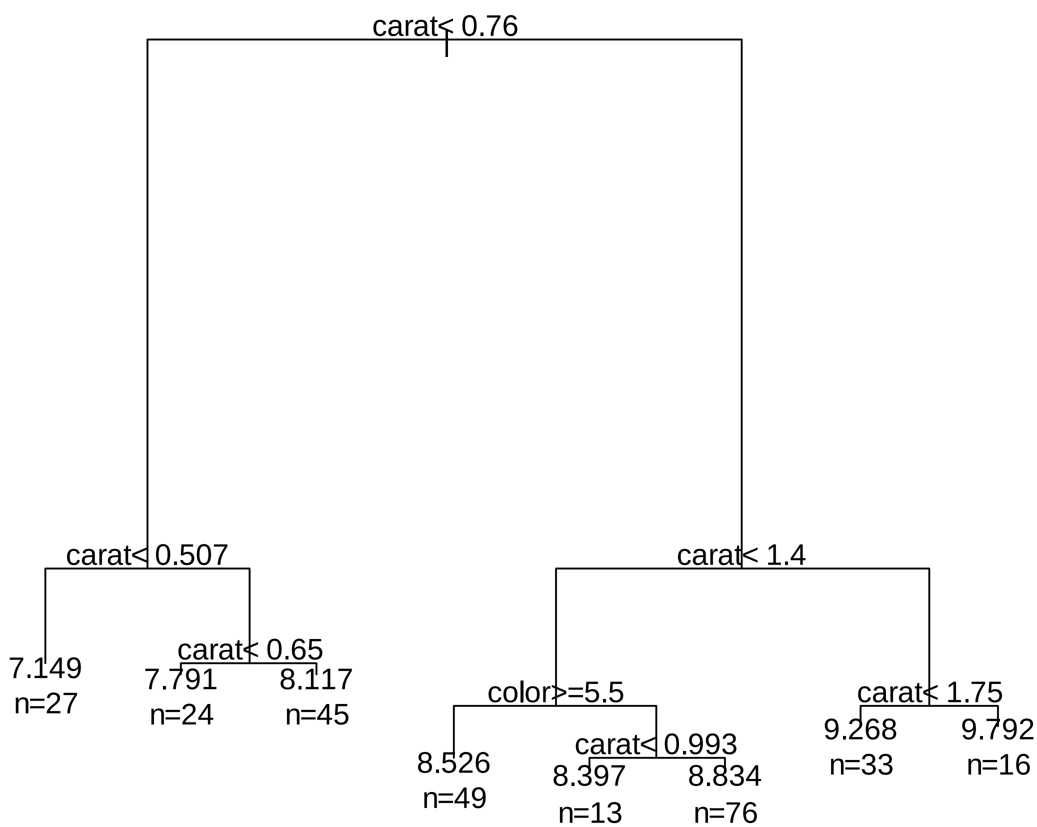



```
> Anova(stepwise.lm.model) # Anova with type II sum of squares from car package
Anova Table (Type II tests)

Response: price
            Sum Sq  Df  F value    Pr(>F)
color     663673643   1  357.566 < 2.2e-16 ***
carat    6646375072   1 3580.857 < 2.2e-16 ***
clarity   371755480   1  200.290 < 2.2e-16 ***
cut        28006360   1   15.089 0.0001196 ***
store     493849471  11   24.188 < 2.2e-16 ***
Residuals 759138804 409
---
```

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
> vif(stepwise.lm.model) # variance inflation factors with car package vif() function
           GVIF Df GVIF^(1/(2*Df))
color   1.195067  1        1.093191
carat   1.275069  1        1.129189
clarity 1.309103  1        1.144160
cut     1.199057  1        1.095015
store   1.855911 11        1.028507
```

Output 9: Regression Tree




```
Call:
rpart(formula = logprice ~ color + carat + clarity + cut, data = diamonds.train)
  n= 283

          CP nsplit rel error    xerror       xstd
1 0.57729864      0 1.0000000 1.0052860 0.08551161
```

```
2 0.14997111      1 0.4227014 0.4334143 0.03621219
3 0.10323231      2 0.2727303 0.2889702 0.02642087
4 0.02152543      3 0.1694979 0.1953513 0.01826373
5 0.01458491      4 0.1479725 0.1783661 0.01753766
6 0.01216563      6 0.1188027 0.1707751 0.01718685
7 0.01000000      7 0.1066371 0.1619750 0.01669010

Variable importance
  carat clarity   color
    95       3       2

Node number 1: 283 observations,    complexity param=0.5772986
  mean=8.502187, MSE=0.4856018
  left son=2 (96 obs) right son=3 (187 obs)
  Primary splits:
      carat   < 0.76  to the left,  improve=0.577298600, (0 missing)
      cut     splits as  LR,        improve=0.038383260, (0 missing)
      clarity < 8.5   to the right, improve=0.009793914, (0 missing)
      color   < 4.5   to the left,  improve=0.006704304, (0 missing)
  Surrogate splits:
      clarity < 3.5   to the left,  agree=0.675, adj=0.042, (0 split)

Node number 2: 96 observations,    complexity param=0.1032323
  mean=7.763219, MSE=0.2281522
  left son=4 (27 obs) right son=5 (69 obs)
  Primary splits:
      carat   < 0.507 to the left,  improve=0.64771860, (0 missing)
      color   < 5.5   to the right, improve=0.01791967, (0 missing)
      clarity < 4.5   to the right, improve=0.01568576, (0 missing)
      cut     splits as  LR,        improve=0.01056864, (0 missing)

Node number 3: 187 observations,    complexity param=0.1499711
  mean=8.88155, MSE=0.1935147
  left son=6 (138 obs) right son=7 (49 obs)
  Primary splits:
      carat   < 1.4   to the left,  improve=0.56953280, (0 missing)
      clarity < 5.5   to the right, improve=0.10860080, (0 missing)
      color   < 5.5   to the right, improve=0.09384341, (0 missing)
      cut     splits as  RL,        improve=0.00124113, (0 missing)

Node number 4: 27 observations
  mean=7.148682, MSE=0.09057353

Node number 5: 69 observations,    complexity param=0.01216563
  mean=8.00369, MSE=0.07638257
  left son=10 (24 obs) right son=11 (45 obs)
  Primary splits:
      carat   < 0.65  to the left,  improve=0.31721810, (0 missing)
```

```
      clarity < 4.5   to the right, improve=0.04212999, (0 missing)
      cut     splits as  RL,        improve=0.03177598, (0 missing)
      color   < 2.5   to the left,  improve=0.02965566, (0 missing)
  Surrogate splits:
      clarity < 3.5   to the left,  agree=0.710, adj=0.167, (0 split)
      color   < 2.5   to the left,  agree=0.681, adj=0.083, (0 split)

Node number 6: 138 observations,    complexity param=0.01458491
  mean=8.683728, MSE=0.07485294
  left son=12 (49 obs) right son=13 (89 obs)
  Primary splits:
      color   < 5.5   to the right, improve=0.1825256, (0 missing)
      clarity < 5.5   to the right, improve=0.1747292, (0 missing)
      carat   < 1.026 to the left,  improve=0.1482785, (0 missing)
      cut     splits as  RL,        improve=0.1183715, (0 missing)
  Surrogate splits:
      carat   < 1.19  to the right, agree=0.688, adj=0.122, (0 split)
      clarity < 9.5   to the right, agree=0.659, adj=0.041, (0 split)

Node number 7: 49 observations,    complexity param=0.02152543
  mean=9.438682, MSE=0.1070963
  left son=14 (33 obs) right son=15 (16 obs)
  Primary splits:
      carat   < 1.75  to the left,  improve=0.56369980, (0 missing)
      color   < 4.5   to the right, improve=0.28059940, (0 missing)
      clarity < 5.5   to the right, improve=0.14808430, (0 missing)
      cut     splits as  LR,        improve=0.00326069, (0 missing)
  Surrogate splits:
      color < 2.5   to the right, agree=0.714, adj=0.125, (0 split)

Node number 10: 24 observations
  mean=7.790545, MSE=0.03264052

Node number 11: 45 observations
  mean=8.117368, MSE=0.06255909

Node number 12: 49 observations
  mean=8.526198, MSE=0.04566017

Node number 13: 89 observations,    complexity param=0.01458491
  mean=8.770458, MSE=0.06974069
  left son=26 (13 obs) right son=27 (76 obs)
  Primary splits:
      carat   < 0.993 to the left,  improve=0.342075400, (0 missing)
      clarity < 6.5   to the right, improve=0.201197200, (0 missing)
      cut     splits as  RL,        improve=0.081611130, (0 missing)
      color   < 3.5   to the right, improve=0.009102182, (0 missing)
```

Node number 14: 33 observations
mean=9.267596, MSE=0.03489791

Node number 15: 16 observations
mean=9.791547, MSE=0.07112186

Node number 26: 13 observations
mean=8.397002, MSE=0.03806358

Node number 27: 76 observations
mean=8.834339, MSE=0.04722183