5205ProjectFinal.R

wuyin

2022-01-13

## import data  
#install.packages("tidyverse")  
#install.packages("ggplot2")  
setwd('/Users/wuyin')  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.5.9000 ✓ purrr 0.3.4   
## ✓ tibble 3.1.6 ✓ dplyr 1.0.5   
## ✓ tidyr 1.1.3 ✓ stringr 1.4.0   
## ✓ readr 1.4.0 ✓ forcats 0.5.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(ggplot2)  
library(corrplot)

## corrplot 0.84 loaded

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

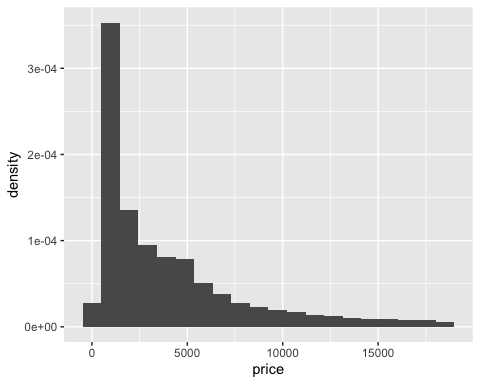
## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

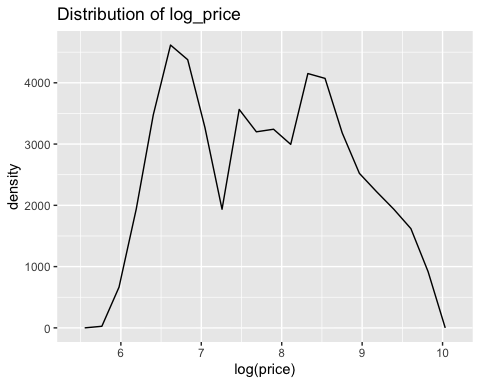
devtools::install\_github("tidyverse/ggplot2")

## Skipping install of 'ggplot2' from a github remote, the SHA1 (c89c265a) has not changed since last install.  
## Use `force = TRUE` to force installation

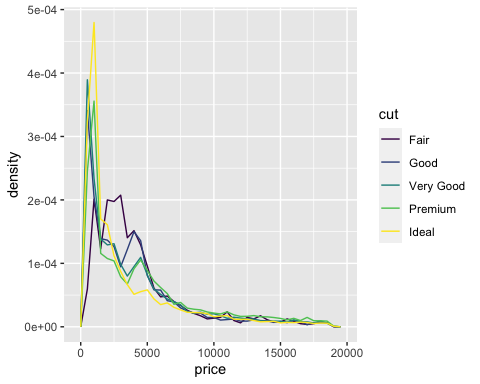
df\_diamond <- ggplot2::diamonds  
##independent variable changes to log\_price  
g1 <- ggplot(df\_diamond, aes(price)) +   
 geom\_histogram(bins = 20, aes(y = ..density..))  
g1



g\_lnprice <- ggplot(df\_diamond, aes(log(price))) +   
 geom\_histogram(bins = 20, aes(y = ..density..)) +  
 geom\_freqpoly(bins = 20) +  
 labs(title = "Distribution of log\_price")  
g\_lnprice



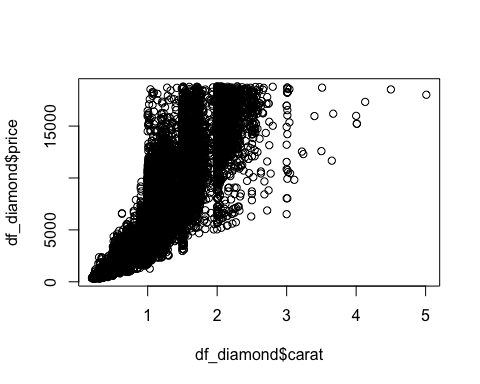
ggplot(df\_diamond, aes(price, after\_stat(density), colour = cut)) +  
 geom\_freqpoly(binwidth = 500)



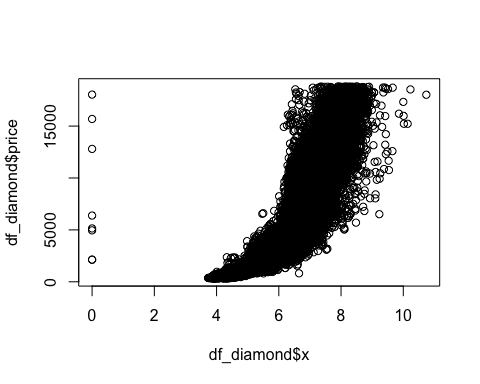
summary(diamonds$color)

## D E F G H I J   
## 6775 9797 9542 11292 8304 5422 2808

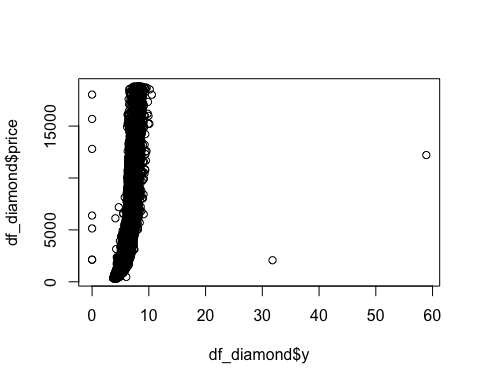
##divide diamonds into two categories, one is that carat is smaller than 1,   
##the other one is that carat is bigger than 1.   
plot(df\_diamond$carat, df\_diamond$price)



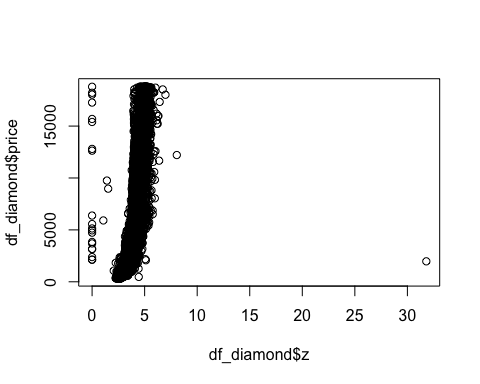
## delete outliers  
plot(df\_diamond$x, df\_diamond$price)



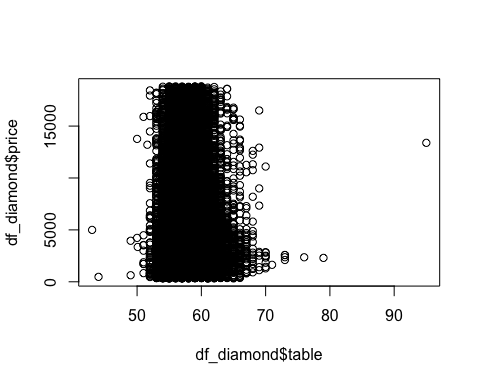
plot(df\_diamond$y, df\_diamond$price) ## filter y>30



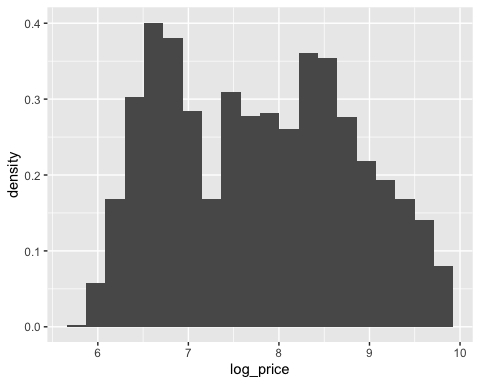
plot(df\_diamond$z, df\_diamond$price) ## filter z>30



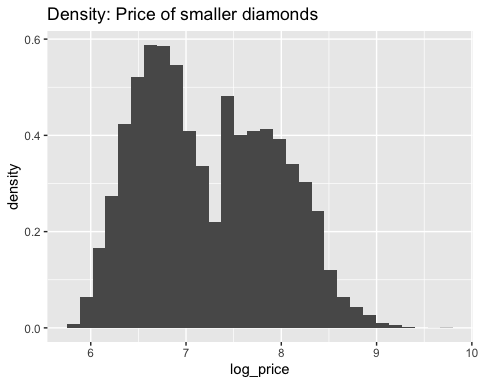
plot(df\_diamond$table, df\_diamond$price) ## filter table>90



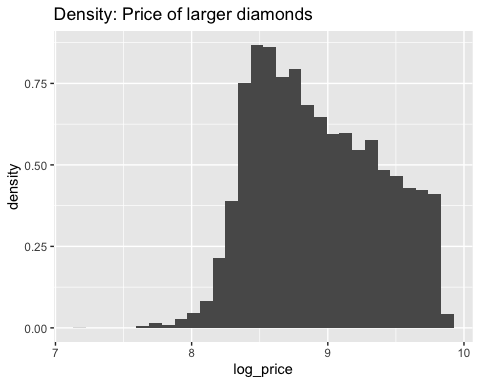
## data  
df\_diamond <- df\_diamond %>%  
 filter(y <= 30 & z <= 30 & table <= 90) %>%  
 mutate(if\_bigger = ifelse(carat>1, 1, 0),   
 log\_price = log(price))  
g2 <- ggplot(df\_diamond, aes(log\_price)) +   
 geom\_histogram(bins = 20, aes(y = ..density..))  
g2



g\_carat <- ggplot(df\_diamond, aes(carat)) +   
 geom\_histogram(bins = 20, aes(y = ..density..))  
  
## split the dataset  
diamond\_big <- df\_diamond %>%  
 filter(if\_bigger==1) %>%  
 select(-price,-x,-y,-z,-if\_bigger)  
diamond\_small <- df\_diamond %>%  
 filter(if\_bigger==0) %>%  
 select(-price,-x,-y,-z,-if\_bigger)  
  
g\_smallprice <- ggplot(diamond\_small, aes(log\_price)) +   
 geom\_histogram(bins = 30, aes(y = ..density..)) +  
 labs(title = "Density: Price of smaller diamonds")  
g\_smallprice



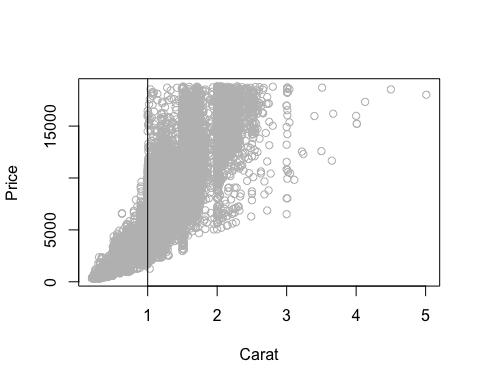
g\_bigprice <- ggplot(diamond\_big, aes(log\_price)) +   
 geom\_histogram(bins = 30, aes(y = ..density..)) +  
 labs(title = "Density: Price of larger diamonds")  
g\_bigprice



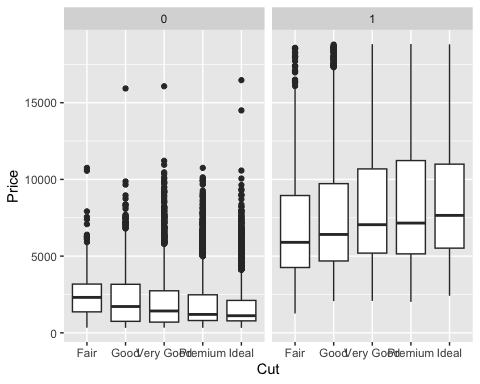
# correlation  
diamond\_num <- df\_diamond %>%  
 select(price, carat,depth,table,x,y,z)  
corr\_diamond <- cor(diamond\_num)  
corrplot(corr\_diamond, method = "color",  
 addCoef.col = "gray",  
 tl.cex = 0.8,  
 tl.col = "black")



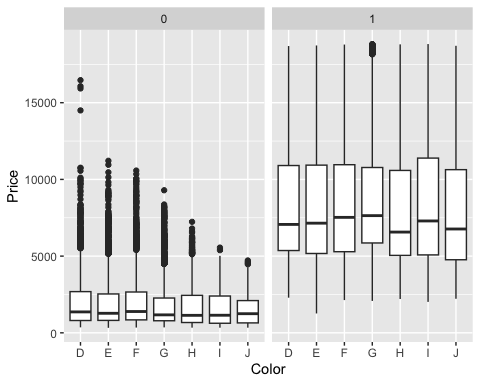
#car::scatterplotMatrix(diamond\_num, spread=FALSE, smoother.args=list(lty=2), main="Scatter Plot Matrix")  
  
plot(df\_diamond$carat, df\_diamond$price, type = "p",col = "grey",lwd = 1,  
 xlab = "Carat",  
 ylab = "Price")  
abline(v=1)



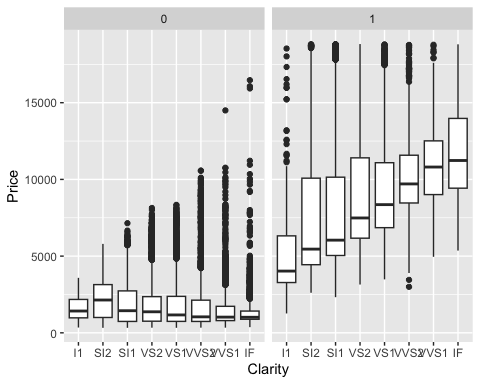
g\_box\_cut <- ggplot(data = df\_diamond, aes(x = cut, y = price)) +  
 geom\_boxplot() +  
 facet\_grid(cols = vars(if\_bigger)) +  
 labs(x = "Cut", y="Price")  
g\_box\_cut



g\_box\_color <- ggplot(data = df\_diamond, aes(x = color, y = price)) +  
 geom\_boxplot() +  
 facet\_grid(cols = vars(if\_bigger)) +  
 labs(x = "Color", y="Price")  
g\_box\_color



g\_box\_clarity <- ggplot(data = df\_diamond, aes(x = clarity, y = price)) +  
 geom\_boxplot() +  
 facet\_grid(cols = vars(if\_bigger)) +  
 labs(x = "Clarity", y="Price")  
g\_box\_clarity



### modeling  
library(olsrr)

##   
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':  
##   
## rivers

fit\_all\_big <- lm(log\_price~., data = diamond\_big)  
summary(fit\_all\_big)

##   
## Call:  
## lm(formula = log\_price ~ ., data = diamond\_big)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.19670 -0.07544 0.00643 0.08868 0.86235   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.9501365 0.0744237 106.823 < 2e-16 \*\*\*  
## carat 1.1573848 0.0030783 375.981 < 2e-16 \*\*\*  
## cut.L 0.1271199 0.0045858 27.720 < 2e-16 \*\*\*  
## cut.Q -0.0423229 0.0037130 -11.398 < 2e-16 \*\*\*  
## cut.C 0.0405610 0.0031748 12.776 < 2e-16 \*\*\*  
## cut^4 0.0188424 0.0026476 7.117 1.15e-12 \*\*\*  
## color.L -0.4330556 0.0035829 -120.868 < 2e-16 \*\*\*  
## color.Q -0.1054408 0.0032522 -32.421 < 2e-16 \*\*\*  
## color.C -0.0100612 0.0031137 -3.231 0.00123 \*\*   
## color^4 0.0007831 0.0029426 0.266 0.79014   
## color^5 -0.0133570 0.0027611 -4.838 1.33e-06 \*\*\*  
## color^6 0.0026300 0.0025056 1.050 0.29389   
## clarity.L 1.0071650 0.0068936 146.101 < 2e-16 \*\*\*  
## clarity.Q -0.2819281 0.0063322 -44.523 < 2e-16 \*\*\*  
## clarity.C 0.1374011 0.0057721 23.804 < 2e-16 \*\*\*  
## clarity^4 -0.0786119 0.0051876 -15.154 < 2e-16 \*\*\*  
## clarity^5 0.0279739 0.0045350 6.168 7.05e-10 \*\*\*  
## clarity^6 0.0015269 0.0038071 0.401 0.68837   
## clarity^7 0.0316179 0.0030098 10.505 < 2e-16 \*\*\*  
## depth -0.0053454 0.0008268 -6.465 1.04e-10 \*\*\*  
## table -0.0033654 0.0006410 -5.250 1.54e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1405 on 17479 degrees of freedom  
## Multiple R-squared: 0.9048, Adjusted R-squared: 0.9047   
## F-statistic: 8305 on 20 and 17479 DF, p-value: < 2.2e-16

fit\_all\_small <- lm(log\_price~., data = diamond\_small)  
summary(fit\_all\_small)

##   
## Call:  
## lm(formula = log\_price ~ ., data = diamond\_small)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.58381 -0.10454 0.00202 0.10697 1.86018   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.7377550 0.0575273 99.740 < 2e-16 \*\*\*  
## carat 3.4004642 0.0039477 861.374 < 2e-16 \*\*\*  
## cut.L 0.1352665 0.0037639 35.938 < 2e-16 \*\*\*  
## cut.Q -0.0145781 0.0029911 -4.874 1.10e-06 \*\*\*  
## cut.C -0.0017346 0.0025623 -0.677 0.49843   
## cut^4 -0.0163077 0.0020011 -8.150 3.77e-16 \*\*\*  
## color.L -0.4303875 0.0030728 -140.065 < 2e-16 \*\*\*  
## color.Q -0.0745163 0.0028930 -25.757 < 2e-16 \*\*\*  
## color.C -0.0071991 0.0026396 -2.727 0.00639 \*\*   
## color^4 0.0186220 0.0023484 7.930 2.26e-15 \*\*\*  
## color^5 -0.0023219 0.0021620 -1.074 0.28285   
## color^6 -0.0038978 0.0019051 -2.046 0.04076 \*   
## clarity.L 0.8420229 0.0059398 141.760 < 2e-16 \*\*\*  
## clarity.Q -0.2384260 0.0056919 -41.889 < 2e-16 \*\*\*  
## clarity.C 0.1267022 0.0047328 26.771 < 2e-16 \*\*\*  
## clarity^4 -0.0463938 0.0035518 -13.062 < 2e-16 \*\*\*  
## clarity^5 0.0175224 0.0026342 6.652 2.93e-11 \*\*\*  
## clarity^6 -0.0064111 0.0021361 -3.001 0.00269 \*\*   
## clarity^7 0.0223963 0.0018657 12.005 < 2e-16 \*\*\*  
## depth -0.0045804 0.0006546 -6.997 2.65e-12 \*\*\*  
## table -0.0028218 0.0004681 -6.028 1.68e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1481 on 36415 degrees of freedom  
## Multiple R-squared: 0.9555, Adjusted R-squared: 0.9555   
## F-statistic: 3.913e+04 on 20 and 36415 DF, p-value: < 2.2e-16

## choose a better model  
  
train\_sub\_big <- sample(nrow(diamond\_big),8/10\*nrow(diamond\_big))  
train\_big <- diamond\_big[train\_sub\_big,]  
test\_big <- diamond\_big[-train\_sub\_big,]  
  
train\_sub\_small <- sample(nrow(diamond\_small),8/10\*nrow(diamond\_small))  
train\_small <- diamond\_small[train\_sub\_small,]  
test\_small <- diamond\_small[-train\_sub\_small,]  
  
# stepwise selection  
step\_fit\_big <- lm(log\_price ~ ., data = train\_big)  
summary(step\_fit\_big)

##   
## Call:  
## lm(formula = log\_price ~ ., data = train\_big)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.22643 -0.07498 0.00673 0.08847 0.86405   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.9240317 0.0837391 94.628 < 2e-16 \*\*\*  
## carat 1.1629870 0.0034465 337.443 < 2e-16 \*\*\*  
## cut.L 0.1288300 0.0050770 25.375 < 2e-16 \*\*\*  
## cut.Q -0.0419703 0.0040998 -10.237 < 2e-16 \*\*\*  
## cut.C 0.0402171 0.0035252 11.409 < 2e-16 \*\*\*  
## cut^4 0.0188892 0.0029610 6.379 1.83e-10 \*\*\*  
## color.L -0.4345370 0.0039987 -108.671 < 2e-16 \*\*\*  
## color.Q -0.1067345 0.0036267 -29.430 < 2e-16 \*\*\*  
## color.C -0.0139027 0.0034712 -4.005 6.23e-05 \*\*\*  
## color^4 -0.0002241 0.0032932 -0.068 0.946   
## color^5 -0.0135659 0.0030735 -4.414 1.02e-05 \*\*\*  
## color^6 0.0026994 0.0027899 0.968 0.333   
## clarity.L 1.0004745 0.0076802 130.267 < 2e-16 \*\*\*  
## clarity.Q -0.2740171 0.0070344 -38.954 < 2e-16 \*\*\*  
## clarity.C 0.1347587 0.0064285 20.963 < 2e-16 \*\*\*  
## clarity^4 -0.0702105 0.0058194 -12.065 < 2e-16 \*\*\*  
## clarity^5 0.0290418 0.0050960 5.699 1.23e-08 \*\*\*  
## clarity^6 0.0037954 0.0042622 0.890 0.373   
## clarity^7 0.0312933 0.0033630 9.305 < 2e-16 \*\*\*  
## depth -0.0049582 0.0009309 -5.327 1.02e-07 \*\*\*  
## table -0.0034471 0.0007155 -4.817 1.47e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1402 on 13979 degrees of freedom  
## Multiple R-squared: 0.9055, Adjusted R-squared: 0.9053   
## F-statistic: 6695 on 20 and 13979 DF, p-value: < 2.2e-16

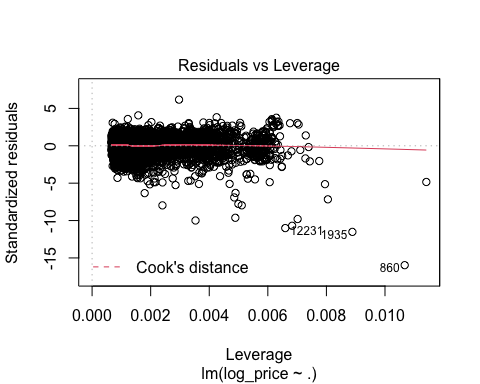
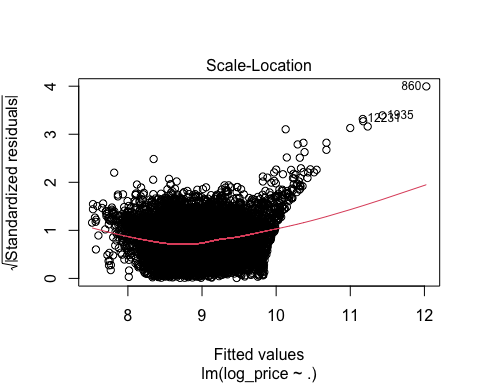
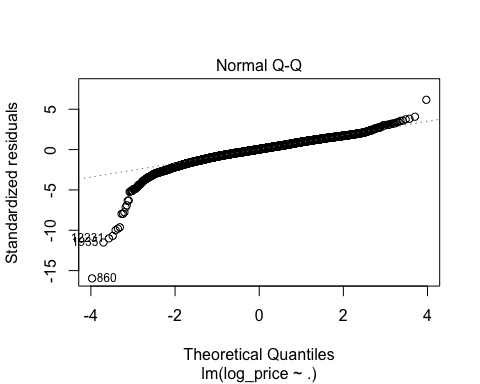
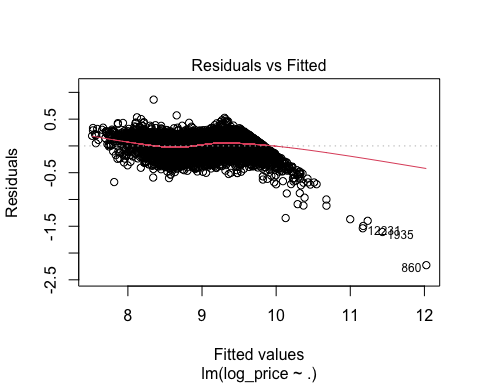
step\_aic\_forward\_big <- ols\_step\_forward\_aic(step\_fit\_big, details = TRUE)

## Forward Selection Method   
## ------------------------  
##   
## Candidate Terms:   
##   
## 1 . carat   
## 2 . cut   
## 3 . color   
## 4 . clarity   
## 5 . depth   
## 6 . table   
##   
## Step 0: AIC = 17725.41   
## log\_price ~ 1   
##   
## --------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## --------------------------------------------------------------------------  
## carat 1 6183.549 1632.286 1274.347 0.562 0.562   
## clarity 1 15835.046 369.666 2536.967 0.127 0.127   
## cut 1 17545.417 38.770 2867.863 0.013 0.013   
## color 1 17648.530 18.395 2888.238 0.006 0.006   
## depth 1 17669.695 11.958 2894.675 0.004 0.004   
## table 1 17715.198 2.535 2904.098 0.001 0.001   
## --------------------------------------------------------------------------  
##   
##   
## - carat   
##   
##   
## Step 1 : AIC = 6183.549   
## log\_price ~ carat   
##   
## -------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## -------------------------------------------------------------------------  
## clarity 1 -5154.670 707.938 566.409 0.805 0.805   
## color 1 3250.099 241.783 1032.563 0.645 0.645   
## cut 1 5044.739 100.227 1174.120 0.596 0.596   
## table 1 5914.133 24.468 1249.879 0.570 0.570   
## depth 1 5982.837 18.319 1256.028 0.568 0.568   
## -------------------------------------------------------------------------  
##   
## - clarity   
##   
##   
## Step 2 : AIC = -5154.67   
## log\_price ~ carat + clarity   
##   
## -------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## -------------------------------------------------------------------------  
## color 1 -13902.377 263.442 302.967 0.896 0.896   
## cut 1 -5760.462 24.296 542.113 0.813 0.813   
## table 1 -5295.872 5.764 560.645 0.807 0.807   
## depth 1 -5277.699 5.036 561.373 0.807 0.807   
## -------------------------------------------------------------------------  
##   
## - color   
##   
##   
## Step 3 : AIC = -13902.38   
## log\_price ~ carat + clarity + color   
##   
## ------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## cut 1 -15225.943 27.488 275.479 0.905 0.905   
## table 1 -14174.949 5.884 297.083 0.898 0.898   
## depth 1 -14044.914 3.112 299.855 0.897 0.897   
## ------------------------------------------------------------------------  
##   
## - cut   
##   
##   
## Step 4 : AIC = -15225.94   
## log\_price ~ carat + clarity + color + cut   
##   
## ------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## depth 1 -15237.839 0.273 275.206 0.905 0.905   
## table 1 -15232.676 0.172 275.307 0.905 0.905   
## ------------------------------------------------------------------------  
##   
## - depth   
##   
##   
## Step 5 : AIC = -15237.84   
## log\_price ~ carat + clarity + color + cut + depth   
##   
## ------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## table 1 -15259.062 0.456 274.749 0.905 0.905   
## ------------------------------------------------------------------------  
##   
## - table   
##   
##   
## Variables Entered:   
##   
## - carat   
## - clarity   
## - color   
## - cut   
## - depth   
## - table   
##   
##   
## Final Model Output   
## ------------------  
##   
## Model Summary   
## -------------------------------------------------------------  
## R 0.952 RMSE 0.140   
## R-Squared 0.905 Coef. Var 1.570   
## Adj. R-Squared 0.905 MSE 0.020   
## Pred R-Squared 0.905 MAE 0.104   
## -------------------------------------------------------------  
## RMSE: Root Mean Square Error   
## MSE: Mean Square Error   
## MAE: Mean Absolute Error   
##   
## ANOVA   
## -------------------------------------------------------------------------  
## Sum of   
## Squares DF Mean Square F Sig.   
## -------------------------------------------------------------------------  
## Regression 2631.884 20 131.594 6695.391 0.0000   
## Residual 274.749 13979 0.020   
## Total 2906.633 13999   
## -------------------------------------------------------------------------  
##   
## Parameter Estimates   
## ------------------------------------------------------------------------------------------  
## model Beta Std. Error Std. Beta t Sig lower upper   
## ------------------------------------------------------------------------------------------  
## (Intercept) 7.924 0.084 94.628 0.000 7.760 8.088   
## carat 1.163 0.003 0.941 337.443 0.000 1.156 1.170   
## clarity.L 1.000 0.008 0.492 130.267 0.000 0.985 1.016   
## clarity.Q -0.274 0.007 -0.143 -38.954 0.000 -0.288 -0.260   
## clarity.C 0.135 0.006 0.081 20.963 0.000 0.122 0.147   
## clarity^4 -0.070 0.006 -0.056 -12.065 0.000 -0.082 -0.059   
## clarity^5 0.029 0.005 0.025 5.699 0.000 0.019 0.039   
## clarity^6 0.004 0.004 0.003 0.890 0.373 -0.005 0.012   
## clarity^7 0.031 0.003 0.028 9.305 0.000 0.025 0.038   
## color.L -0.435 0.004 -0.306 -108.671 0.000 -0.442 -0.427   
## color.Q -0.107 0.004 -0.080 -29.430 0.000 -0.114 -0.100   
## color.C -0.014 0.003 -0.011 -4.005 0.000 -0.021 -0.007   
## color^4 0.000 0.003 0.000 -0.068 0.946 -0.007 0.006   
## color^5 -0.014 0.003 -0.012 -4.414 0.000 -0.020 -0.008   
## color^6 0.003 0.003 0.003 0.968 0.333 -0.003 0.008   
## cut.L 0.129 0.005 0.099 25.375 0.000 0.119 0.139   
## cut.Q -0.042 0.004 -0.041 -10.237 0.000 -0.050 -0.034   
## cut.C 0.040 0.004 0.039 11.409 0.000 0.033 0.047   
## cut^4 0.019 0.003 0.019 6.379 0.000 0.013 0.025   
## depth -0.005 0.001 -0.017 -5.327 0.000 -0.007 -0.003   
## table -0.003 0.001 -0.016 -4.817 0.000 -0.005 -0.002   
## ------------------------------------------------------------------------------------------

step\_aic\_forward\_big

##   
## Selection Summary   
## -----------------------------------------------------------------------  
## Variable AIC Sum Sq RSS R-Sq Adj. R-Sq   
## -----------------------------------------------------------------------  
## carat 6183.549 1632.286 1274.347 0.56157 0.56154   
## clarity -5154.670 2340.224 566.409 0.80513 0.80502   
## color -13902.377 2603.666 302.967 0.89577 0.89566   
## cut -15225.943 2631.154 275.479 0.90522 0.90510   
## depth -15237.839 2631.428 275.206 0.90532 0.90519   
## table -15259.062 2631.884 274.749 0.90548 0.90534   
## -----------------------------------------------------------------------

plot(step\_fit\_big)



step\_fit\_small <- lm(log\_price ~ ., data = train\_small)  
summary(step\_fit\_small)

##   
## Call:  
## lm(formula = log\_price ~ ., data = train\_small)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5822 -0.1043 0.0027 0.1071 1.8747   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.7117470 0.0640041 89.240 < 2e-16 \*\*\*  
## carat 3.3967694 0.0044290 766.934 < 2e-16 \*\*\*  
## cut.L 0.1368837 0.0041860 32.700 < 2e-16 \*\*\*  
## cut.Q -0.0147449 0.0033349 -4.421 9.84e-06 \*\*\*  
## cut.C -0.0014096 0.0028553 -0.494 0.62154   
## cut^4 -0.0164039 0.0022328 -7.347 2.08e-13 \*\*\*  
## color.L -0.4298907 0.0034407 -124.944 < 2e-16 \*\*\*  
## color.Q -0.0737243 0.0032426 -22.736 < 2e-16 \*\*\*  
## color.C -0.0063807 0.0029537 -2.160 0.03076 \*   
## color^4 0.0189469 0.0026247 7.219 5.37e-13 \*\*\*  
## color^5 -0.0011898 0.0024177 -0.492 0.62263   
## color^6 -0.0029132 0.0021380 -1.363 0.17302   
## clarity.L 0.8502866 0.0065522 129.770 < 2e-16 \*\*\*  
## clarity.Q -0.2455954 0.0062722 -39.156 < 2e-16 \*\*\*  
## clarity.C 0.1342345 0.0052210 25.710 < 2e-16 \*\*\*  
## clarity^4 -0.0531342 0.0039322 -13.512 < 2e-16 \*\*\*  
## clarity^5 0.0194035 0.0029346 6.612 3.86e-11 \*\*\*  
## clarity^6 -0.0074192 0.0023960 -3.097 0.00196 \*\*   
## clarity^7 0.0249569 0.0020940 11.918 < 2e-16 \*\*\*  
## depth -0.0045591 0.0007260 -6.280 3.44e-10 \*\*\*  
## table -0.0023908 0.0005248 -4.555 5.25e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1485 on 29127 degrees of freedom  
## Multiple R-squared: 0.9553, Adjusted R-squared: 0.9552   
## F-statistic: 3.109e+04 on 20 and 29127 DF, p-value: < 2.2e-16

step\_aic\_forward\_small <- ols\_step\_forward\_aic(step\_fit\_small, details = TRUE)

## Forward Selection Method   
## ------------------------  
##   
## Candidate Terms:   
##   
## 1 . carat   
## 2 . cut   
## 3 . color   
## 4 . clarity   
## 5 . depth   
## 6 . table   
##   
## Step 0: AIC = 62098.76   
## log\_price ~ 1   
##   
## ----------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ----------------------------------------------------------------------------  
## carat 1 2806.402 12486.881 1878.708 0.869 0.869   
## table 1 61486.410 299.612 14065.977 0.021 0.021   
## clarity 1 61519.851 289.261 14076.328 0.020 0.020   
## cut 1 61592.049 251.447 14114.142 0.018 0.017   
## color 1 61848.040 128.898 14236.691 0.009 0.009   
## depth 1 62095.848 2.419 14363.170 0.000 0.000   
## ----------------------------------------------------------------------------  
##   
##   
## - carat   
##   
##   
## Step 1 : AIC = 2806.402   
## log\_price ~ carat   
##   
## --------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## --------------------------------------------------------------------------  
## clarity 1 -11740.969 738.720 1139.988 0.921 0.921   
## color 1 -2380.702 306.918 1571.790 0.891 0.891   
## cut 1 395.368 149.622 1729.086 0.880 0.880   
## depth 1 2357.839 28.817 1849.891 0.871 0.871   
## table 1 2407.119 25.687 1853.021 0.871 0.871   
## --------------------------------------------------------------------------  
##   
## - clarity   
##   
##   
## Step 2 : AIC = -11740.97   
## log\_price ~ carat + clarity   
##   
## --------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## --------------------------------------------------------------------------  
## color 1 -25645.929 432.787 707.201 0.951 0.951   
## cut 1 -13635.129 72.019 1067.969 0.926 0.926   
## depth 1 -12080.588 13.283 1126.705 0.922 0.922   
## table 1 -11947.919 8.143 1131.845 0.921 0.921   
## --------------------------------------------------------------------------  
##   
## - color   
##   
##   
## Step 3 : AIC = -25645.93   
## log\_price ~ carat + clarity + color   
##   
## ------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## cut 1 -28380.683 63.511 643.690 0.955 0.955   
## table 1 -26028.048 9.259 697.942 0.951 0.951   
## depth 1 -25876.598 5.623 701.578 0.951 0.951   
## ------------------------------------------------------------------------  
##   
## - cut   
##   
##   
## Step 4 : AIC = -28380.68   
## log\_price ~ carat + clarity + color + cut   
##   
## ------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## depth 1 -28401.901 0.513 643.177 0.955 0.955   
## table 1 -28383.219 0.100 643.590 0.955 0.955   
## ------------------------------------------------------------------------  
##   
## - depth   
##   
##   
## Step 5 : AIC = -28401.9   
## log\_price ~ carat + clarity + color + cut + depth   
##   
## ------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## table 1 -28420.658 0.458 642.719 0.955 0.955   
## ------------------------------------------------------------------------  
##   
## - table   
##   
##   
## Variables Entered:   
##   
## - carat   
## - clarity   
## - color   
## - cut   
## - depth   
## - table   
##   
##   
## Final Model Output   
## ------------------  
##   
## Model Summary   
## -------------------------------------------------------------  
## R 0.977 RMSE 0.149   
## R-Squared 0.955 Coef. Var 2.052   
## Adj. R-Squared 0.955 MSE 0.022   
## Pred R-Squared 0.955 MAE 0.120   
## -------------------------------------------------------------  
## RMSE: Root Mean Square Error   
## MSE: Mean Square Error   
## MAE: Mean Absolute Error   
##   
## ANOVA   
## ---------------------------------------------------------------------------  
## Sum of   
## Squares DF Mean Square F Sig.   
## ---------------------------------------------------------------------------  
## Regression 13722.870 20 686.144 31094.906 0.0000   
## Residual 642.719 29127 0.022   
## Total 14365.590 29147   
## ---------------------------------------------------------------------------  
##   
## Parameter Estimates   
## ------------------------------------------------------------------------------------------  
## model Beta Std. Error Std. Beta t Sig lower upper   
## ------------------------------------------------------------------------------------------  
## (Intercept) 5.712 0.064 89.240 0.000 5.586 5.837   
## carat 3.397 0.004 1.056 766.934 0.000 3.388 3.405   
## clarity.L 0.850 0.007 0.310 129.770 0.000 0.837 0.863   
## clarity.Q -0.246 0.006 -0.084 -39.156 0.000 -0.258 -0.233   
## clarity.C 0.134 0.005 0.060 25.710 0.000 0.124 0.144   
## clarity^4 -0.053 0.004 -0.026 -13.512 0.000 -0.061 -0.045   
## clarity^5 0.019 0.003 0.010 6.612 0.000 0.014 0.025   
## clarity^6 -0.007 0.002 -0.004 -3.097 0.002 -0.012 -0.003   
## clarity^7 0.025 0.002 0.015 11.918 0.000 0.021 0.029   
## color.L -0.430 0.003 -0.188 -124.944 0.000 -0.437 -0.423   
## color.Q -0.074 0.003 -0.036 -22.736 0.000 -0.080 -0.067   
## color.C -0.006 0.003 -0.003 -2.160 0.031 -0.012 -0.001   
## color^4 0.019 0.003 0.011 7.219 0.000 0.014 0.024   
## color^5 -0.001 0.002 -0.001 -0.492 0.623 -0.006 0.004   
## color^6 -0.003 0.002 -0.002 -1.363 0.173 -0.007 0.001   
## cut.L 0.137 0.004 0.069 32.700 0.000 0.129 0.145   
## cut.Q -0.015 0.003 -0.010 -4.421 0.000 -0.021 -0.008   
## cut.C -0.001 0.003 -0.001 -0.494 0.622 -0.007 0.004   
## cut^4 -0.016 0.002 -0.010 -7.347 0.000 -0.021 -0.012   
## depth -0.005 0.001 -0.009 -6.280 0.000 -0.006 -0.003   
## table -0.002 0.001 -0.008 -4.555 0.000 -0.003 -0.001   
## ------------------------------------------------------------------------------------------

step\_aic\_forward\_small

##   
## Selection Summary   
## ------------------------------------------------------------------------  
## Variable AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## carat 2806.402 12486.881 1878.708 0.86922 0.86922   
## clarity -11740.969 13225.601 1139.988 0.92064 0.92062   
## color -25645.929 13658.389 707.201 0.95077 0.95075   
## cut -28380.683 13721.900 643.690 0.95519 0.95516   
## depth -28401.901 13722.412 643.177 0.95523 0.95520   
## table -28420.658 13722.870 642.719 0.95526 0.95523   
## ------------------------------------------------------------------------

#plot(step\_fit\_small)  
  
  
## poly  
train\_big2 <- train\_big %>%  
 mutate(carat2 = carat^2)  
step\_fit\_big2 <- lm(log\_price ~., data = train\_big2)  
summary(step\_fit\_big2)

##   
## Call:  
## lm(formula = log\_price ~ ., data = train\_big2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.87009 -0.06691 0.01009 0.07993 1.11537   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.2274477 0.0749728 96.401 < 2e-16 \*\*\*  
## carat 2.1234537 0.0157573 134.760 < 2e-16 \*\*\*  
## cut.L 0.1207724 0.0044962 26.861 < 2e-16 \*\*\*  
## cut.Q -0.0371858 0.0036301 -10.244 < 2e-16 \*\*\*  
## cut.C 0.0372040 0.0031210 11.920 < 2e-16 \*\*\*  
## cut^4 0.0208147 0.0026214 7.940 2.17e-15 \*\*\*  
## color.L -0.4410264 0.0035413 -124.537 < 2e-16 \*\*\*  
## color.Q -0.1022921 0.0032113 -31.853 < 2e-16 \*\*\*  
## color.C -0.0085437 0.0030741 -2.779 0.00546 \*\*   
## color^4 0.0034021 0.0029158 1.167 0.24332   
## color^5 -0.0147890 0.0027209 -5.435 5.56e-08 \*\*\*  
## color^6 0.0023220 0.0024697 0.940 0.34714   
## clarity.L 0.9739634 0.0068122 142.973 < 2e-16 \*\*\*  
## clarity.Q -0.2402646 0.0062508 -38.437 < 2e-16 \*\*\*  
## clarity.C 0.1200892 0.0056957 21.084 < 2e-16 \*\*\*  
## clarity^4 -0.0631633 0.0051529 -12.258 < 2e-16 \*\*\*  
## clarity^5 0.0254949 0.0045115 5.651 1.63e-08 \*\*\*  
## clarity^6 0.0082616 0.0037738 2.189 0.02860 \*   
## clarity^7 0.0326594 0.0029772 10.970 < 2e-16 \*\*\*  
## depth -0.0047981 0.0008240 -5.823 5.92e-09 \*\*\*  
## table -0.0037498 0.0006334 -5.920 3.30e-09 \*\*\*  
## carat2 -0.2988801 0.0048106 -62.129 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1241 on 13978 degrees of freedom  
## Multiple R-squared: 0.9259, Adjusted R-squared: 0.9258   
## F-statistic: 8321 on 21 and 13978 DF, p-value: < 2.2e-16

step\_aic\_forward\_big2 <- ols\_step\_forward\_aic(step\_fit\_big2, details = TRUE)

## Forward Selection Method   
## ------------------------  
##   
## Candidate Terms:   
##   
## 1 . carat   
## 2 . cut   
## 3 . color   
## 4 . clarity   
## 5 . depth   
## 6 . table   
## 7 . carat2   
##   
## Step 0: AIC = 17725.41   
## log\_price ~ 1   
##   
## --------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## --------------------------------------------------------------------------  
## carat 1 6183.549 1632.286 1274.347 0.562 0.562   
## carat2 1 8220.869 1432.667 1473.966 0.493 0.493   
## clarity 1 15835.046 369.666 2536.967 0.127 0.127   
## cut 1 17545.417 38.770 2867.863 0.013 0.013   
## color 1 17648.530 18.395 2888.238 0.006 0.006   
## depth 1 17669.695 11.958 2894.675 0.004 0.004   
## table 1 17715.198 2.535 2904.098 0.001 0.001   
## --------------------------------------------------------------------------  
##   
##   
## - carat   
##   
##   
## Step 1 : AIC = 6183.549   
## log\_price ~ carat   
##   
## -------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## -------------------------------------------------------------------------  
## clarity 1 -5154.670 707.938 566.409 0.805 0.805   
## color 1 3250.099 241.783 1032.563 0.645 0.645   
## carat2 1 4989.790 104.325 1170.022 0.597 0.597   
## cut 1 5044.739 100.227 1174.120 0.596 0.596   
## table 1 5914.133 24.468 1249.879 0.570 0.570   
## depth 1 5982.837 18.319 1256.028 0.568 0.568   
## -------------------------------------------------------------------------  
##   
## - clarity   
##   
##   
## Step 2 : AIC = -5154.67   
## log\_price ~ carat + clarity   
##   
## -------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## -------------------------------------------------------------------------  
## color 1 -13902.377 263.442 302.967 0.896 0.896   
## carat2 1 -6557.088 54.063 512.346 0.824 0.824   
## cut 1 -5760.462 24.296 542.113 0.813 0.813   
## table 1 -5295.872 5.764 560.645 0.807 0.807   
## depth 1 -5277.699 5.036 561.373 0.807 0.807   
## -------------------------------------------------------------------------  
##   
## - color   
##   
##   
## Step 3 : AIC = -13902.38   
## log\_price ~ carat + clarity + color   
##   
## ------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## carat2 1 -17069.271 61.369 241.598 0.917 0.917   
## cut 1 -15225.943 27.488 275.479 0.905 0.905   
## table 1 -14174.949 5.884 297.083 0.898 0.898   
## depth 1 -14044.914 3.112 299.855 0.897 0.897   
## ------------------------------------------------------------------------  
##   
## - carat2   
##   
##   
## Step 4 : AIC = -17069.27   
## log\_price ~ carat + clarity + color + carat2   
##   
## ------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## cut 1 -18625.526 25.541 216.057 0.926 0.926   
## table 1 -17424.506 6.087 235.511 0.919 0.919   
## depth 1 -17218.764 2.600 238.997 0.918 0.918   
## ------------------------------------------------------------------------  
##   
## - cut   
##   
##   
## Step 5 : AIC = -18625.53   
## log\_price ~ carat + clarity + color + carat2 + cut   
##   
## ------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## table 1 -18639.054 0.240 215.817 0.926 0.926   
## depth 1 -18637.914 0.222 215.835 0.926 0.926   
## ------------------------------------------------------------------------  
##   
## - table   
##   
##   
## Step 6 : AIC = -18639.05   
## log\_price ~ carat + clarity + color + carat2 + cut + table   
##   
## ------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## depth 1 -18670.969 0.522 215.295 0.926 0.926   
## ------------------------------------------------------------------------  
##   
## - depth   
##   
##   
## Variables Entered:   
##   
## - carat   
## - clarity   
## - color   
## - carat2   
## - cut   
## - table   
## - depth   
##   
##   
## Final Model Output   
## ------------------  
##   
## Model Summary   
## -------------------------------------------------------------  
## R 0.962 RMSE 0.124   
## R-Squared 0.926 Coef. Var 1.390   
## Adj. R-Squared 0.926 MSE 0.015   
## Pred R-Squared 0.925 MAE 0.094   
## -------------------------------------------------------------  
## RMSE: Root Mean Square Error   
## MSE: Mean Square Error   
## MAE: Mean Absolute Error   
##   
## ANOVA   
## -------------------------------------------------------------------------  
## Sum of   
## Squares DF Mean Square F Sig.   
## -------------------------------------------------------------------------  
## Regression 2691.338 21 128.159 8320.704 0.0000   
## Residual 215.295 13978 0.015   
## Total 2906.633 13999   
## -------------------------------------------------------------------------  
##   
## Parameter Estimates   
## ------------------------------------------------------------------------------------------  
## model Beta Std. Error Std. Beta t Sig lower upper   
## ------------------------------------------------------------------------------------------  
## (Intercept) 7.227 0.075 96.401 0.000 7.080 7.374   
## carat 2.123 0.016 1.719 134.760 0.000 2.093 2.154   
## clarity.L 0.974 0.007 0.479 142.973 0.000 0.961 0.987   
## clarity.Q -0.240 0.006 -0.125 -38.437 0.000 -0.253 -0.228   
## clarity.C 0.120 0.006 0.072 21.084 0.000 0.109 0.131   
## clarity^4 -0.063 0.005 -0.050 -12.258 0.000 -0.073 -0.053   
## clarity^5 0.025 0.005 0.022 5.651 0.000 0.017 0.034   
## clarity^6 0.008 0.004 0.007 2.189 0.029 0.001 0.016   
## clarity^7 0.033 0.003 0.029 10.970 0.000 0.027 0.038   
## color.L -0.441 0.004 -0.311 -124.537 0.000 -0.448 -0.434   
## color.Q -0.102 0.003 -0.076 -31.853 0.000 -0.109 -0.096   
## color.C -0.009 0.003 -0.007 -2.779 0.005 -0.015 -0.003   
## color^4 0.003 0.003 0.003 1.167 0.243 -0.002 0.009   
## color^5 -0.015 0.003 -0.013 -5.435 0.000 -0.020 -0.009   
## color^6 0.002 0.002 0.002 0.940 0.347 -0.003 0.007   
## carat2 -0.299 0.005 -0.792 -62.129 0.000 -0.308 -0.289   
## cut.L 0.121 0.004 0.092 26.861 0.000 0.112 0.130   
## cut.Q -0.037 0.004 -0.036 -10.244 0.000 -0.044 -0.030   
## cut.C 0.037 0.003 0.036 11.920 0.000 0.031 0.043   
## cut^4 0.021 0.003 0.021 7.940 0.000 0.016 0.026   
## table -0.004 0.001 -0.018 -5.920 0.000 -0.005 -0.003   
## depth -0.005 0.001 -0.016 -5.823 0.000 -0.006 -0.003   
## ------------------------------------------------------------------------------------------

step\_aic\_forward\_big2

##   
## Selection Summary   
## -----------------------------------------------------------------------  
## Variable AIC Sum Sq RSS R-Sq Adj. R-Sq   
## -----------------------------------------------------------------------  
## carat 6183.549 1632.286 1274.347 0.56157 0.56154   
## clarity -5154.670 2340.224 566.409 0.80513 0.80502   
## color -13902.377 2603.666 302.967 0.89577 0.89566   
## carat2 -17069.271 2665.036 241.598 0.91688 0.91679   
## cut -18625.526 2690.576 216.057 0.92567 0.92557   
## table -18639.054 2690.816 215.817 0.92575 0.92564   
## depth -18670.969 2691.338 215.295 0.92593 0.92582   
## -----------------------------------------------------------------------

#plot(step\_fit\_big2)  
  
  
train\_small2 <- train\_small %>%  
 mutate(carat2 = carat^2)  
step\_fit\_small2 <- lm(log\_price ~., data = train\_small2)  
summary(step\_fit\_small2)

##   
## Call:  
## lm(formula = log\_price ~ ., data = train\_small2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.45391 -0.08532 -0.00768 0.08079 1.88748   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.8241382 0.0539876 89.356 < 2e-16 \*\*\*  
## carat 5.7169458 0.0209159 273.330 < 2e-16 \*\*\*  
## cut.L 0.1199624 0.0034964 34.310 < 2e-16 \*\*\*  
## cut.Q -0.0207264 0.0027834 -7.447 9.85e-14 \*\*\*  
## cut.C -0.0045085 0.0023828 -1.892 0.058487 .   
## cut^4 -0.0194080 0.0018634 -10.415 < 2e-16 \*\*\*  
## color.L -0.4289012 0.0028712 -149.382 < 2e-16 \*\*\*  
## color.Q -0.0788961 0.0027063 -29.153 < 2e-16 \*\*\*  
## color.C -0.0119554 0.0024653 -4.849 1.24e-06 \*\*\*  
## color^4 0.0169333 0.0021903 7.731 1.10e-14 \*\*\*  
## color^5 0.0041748 0.0020180 2.069 0.038579 \*   
## color^6 0.0016346 0.0017845 0.916 0.359681   
## clarity.L 0.8636184 0.0054690 157.912 < 2e-16 \*\*\*  
## clarity.Q -0.2188624 0.0052394 -41.773 < 2e-16 \*\*\*  
## clarity.C 0.1240468 0.0043578 28.466 < 2e-16 \*\*\*  
## clarity^4 -0.0559136 0.0032815 -17.039 < 2e-16 \*\*\*  
## clarity^5 0.0206619 0.0024489 8.437 < 2e-16 \*\*\*  
## clarity^6 -0.0069235 0.0019994 -3.463 0.000535 \*\*\*  
## clarity^7 0.0306306 0.0017481 17.522 < 2e-16 \*\*\*  
## depth -0.0011795 0.0006066 -1.945 0.051826 .   
## table -0.0007158 0.0004382 -1.633 0.102388   
## carat2 -1.9334738 0.0171556 -112.702 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.124 on 29126 degrees of freedom  
## Multiple R-squared: 0.9688, Adjusted R-squared: 0.9688   
## F-statistic: 4.313e+04 on 21 and 29126 DF, p-value: < 2.2e-16

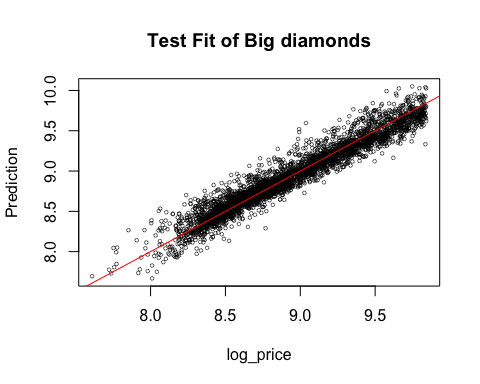
step\_aic\_forward\_small2 <- ols\_step\_forward\_aic(step\_fit\_small2, details = TRUE)

## Forward Selection Method   
## ------------------------  
##   
## Candidate Terms:   
##   
## 1 . carat   
## 2 . cut   
## 3 . color   
## 4 . clarity   
## 5 . depth   
## 6 . table   
## 7 . carat2   
##   
## Step 0: AIC = 62098.76   
## log\_price ~ 1   
##   
## ----------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ----------------------------------------------------------------------------  
## carat 1 2806.402 12486.881 1878.708 0.869 0.869   
## carat2 1 13425.587 11661.125 2704.465 0.812 0.812   
## table 1 61486.410 299.612 14065.977 0.021 0.021   
## clarity 1 61519.851 289.261 14076.328 0.020 0.020   
## cut 1 61592.049 251.447 14114.142 0.018 0.017   
## color 1 61848.040 128.898 14236.691 0.009 0.009   
## depth 1 62095.848 2.419 14363.170 0.000 0.000   
## ----------------------------------------------------------------------------  
##   
##   
## - carat   
##   
##   
## Step 1 : AIC = 2806.402   
## log\_price ~ carat   
##   
## --------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## --------------------------------------------------------------------------  
## clarity 1 -11740.969 738.720 1139.988 0.921 0.921   
## color 1 -2380.702 306.918 1571.790 0.891 0.891   
## carat2 1 -469.296 199.816 1678.892 0.883 0.883   
## cut 1 395.368 149.622 1729.086 0.880 0.880   
## depth 1 2357.839 28.817 1849.891 0.871 0.871   
## table 1 2407.119 25.687 1853.021 0.871 0.871   
## --------------------------------------------------------------------------  
##   
## - clarity   
##   
##   
## Step 2 : AIC = -11740.97   
## log\_price ~ carat + clarity   
##   
## --------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## --------------------------------------------------------------------------  
## color 1 -25645.929 432.787 707.201 0.951 0.951   
## carat2 1 -18339.626 231.011 908.977 0.937 0.937   
## cut 1 -13635.129 72.019 1067.969 0.926 0.926   
## depth 1 -12080.588 13.283 1126.705 0.922 0.922   
## table 1 -11947.919 8.143 1131.845 0.921 0.921   
## --------------------------------------------------------------------------  
##   
## - color   
##   
##   
## Step 3 : AIC = -25645.93   
## log\_price ~ carat + clarity + color   
##   
## -------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## -------------------------------------------------------------------------  
## carat2 1 -36422.115 218.603 488.598 0.966 0.966   
## cut 1 -28380.683 63.511 643.690 0.955 0.955   
## table 1 -26028.048 9.259 697.942 0.951 0.951   
## depth 1 -25876.598 5.623 701.578 0.951 0.951   
## -------------------------------------------------------------------------  
##   
## - carat2   
##   
##   
## Step 4 : AIC = -36422.11   
## log\_price ~ carat + clarity + color + carat2   
##   
## ------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## cut 1 -38967.563 40.981 447.617 0.969 0.969   
## table 1 -36635.163 3.592 485.007 0.966 0.966   
## depth 1 -36582.520 2.715 485.883 0.966 0.966   
## ------------------------------------------------------------------------  
##   
## - cut   
##   
##   
## Step 5 : AIC = -38967.56   
## log\_price ~ carat + clarity + color + carat2 + cut   
##   
## ------------------------------------------------------------------------  
## Variable DF AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## depth 1 -38967.494 0.030 447.587 0.969 0.969   
## table 1 -38966.380 0.013 447.604 0.969 0.969   
## ------------------------------------------------------------------------  
##   
##   
## No more variables to be added.  
##   
## Variables Entered:   
##   
## - carat   
## - clarity   
## - color   
## - carat2   
## - cut   
##   
##   
## Final Model Output   
## ------------------  
##   
## Model Summary   
## -------------------------------------------------------------  
## R 0.984 RMSE 0.124   
## R-Squared 0.969 Coef. Var 1.713   
## Adj. R-Squared 0.969 MSE 0.015   
## Pred R-Squared 0.969 MAE 0.098   
## -------------------------------------------------------------  
## RMSE: Root Mean Square Error   
## MSE: Mean Square Error   
## MAE: Mean Absolute Error   
##   
## ANOVA   
## ---------------------------------------------------------------------------  
## Sum of   
## Squares DF Mean Square F Sig.   
## ---------------------------------------------------------------------------  
## Regression 13917.973 19 732.525 47667.984 0.0000   
## Residual 447.617 29128 0.015   
## Total 14365.590 29147   
## ---------------------------------------------------------------------------  
##   
## Parameter Estimates   
## ------------------------------------------------------------------------------------------  
## model Beta Std. Error Std. Beta t Sig lower upper   
## ------------------------------------------------------------------------------------------  
## (Intercept) 4.709 0.006 812.191 0.000 4.698 4.720   
## carat 5.719 0.021 1.778 273.638 0.000 5.678 5.760   
## clarity.L 0.864 0.005 0.315 158.239 0.000 0.853 0.875   
## clarity.Q -0.219 0.005 -0.075 -41.791 0.000 -0.229 -0.209   
## clarity.C 0.124 0.004 0.055 28.531 0.000 0.116 0.133   
## clarity^4 -0.056 0.003 -0.027 -17.068 0.000 -0.062 -0.050   
## clarity^5 0.021 0.002 0.011 8.522 0.000 0.016 0.026   
## clarity^6 -0.007 0.002 -0.004 -3.505 0.000 -0.011 -0.003   
## clarity^7 0.031 0.002 0.019 17.535 0.000 0.027 0.034   
## color.L -0.429 0.003 -0.188 -149.594 0.000 -0.435 -0.423   
## color.Q -0.079 0.003 -0.039 -29.181 0.000 -0.084 -0.074   
## color.C -0.012 0.002 -0.006 -4.840 0.000 -0.017 -0.007   
## color^4 0.017 0.002 0.009 7.745 0.000 0.013 0.021   
## color^5 0.004 0.002 0.002 2.042 0.041 0.000 0.008   
## color^6 0.002 0.002 0.001 0.929 0.353 -0.002 0.005   
## carat2 -1.935 0.017 -0.733 -112.957 0.000 -1.969 -1.902   
## cut.L 0.123 0.003 0.062 38.999 0.000 0.117 0.129   
## cut.Q -0.021 0.003 -0.014 -7.795 0.000 -0.027 -0.016   
## cut.C -0.004 0.002 -0.002 -1.613 0.107 -0.008 0.001   
## cut^4 -0.019 0.002 -0.012 -10.396 0.000 -0.023 -0.015   
## ------------------------------------------------------------------------------------------

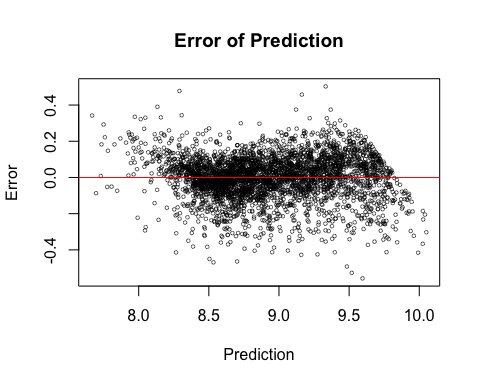
step\_aic\_forward\_small2

##   
## Selection Summary   
## ------------------------------------------------------------------------  
## Variable AIC Sum Sq RSS R-Sq Adj. R-Sq   
## ------------------------------------------------------------------------  
## carat 2806.402 12486.881 1878.708 0.86922 0.86922   
## clarity -11740.969 13225.601 1139.988 0.92064 0.92062   
## color -25645.929 13658.389 707.201 0.95077 0.95075   
## carat2 -36422.115 13876.991 488.598 0.96599 0.96597   
## cut -38967.563 13917.973 447.617 0.96884 0.96882   
## ------------------------------------------------------------------------

#plot(step\_fit\_small2)  
  
## use test dataset to predict  
test\_small2 <- test\_small %>%  
 mutate(carat2 = carat^2)  
test\_big2 <- test\_big %>%  
 mutate(carat2 = carat^2)  
  
prediction\_test\_big <- predict(step\_fit\_big2, newdata = test\_big2)  
error\_test\_big <- test\_big2$log\_price - prediction\_test\_big  
plot(x = test\_big2$log\_price, y = prediction\_test\_big,  
 main = "Test Fit of Big diamonds",  
 lwd = 0.5,  
 type = "p",  
 cex = 0.5,  
 xlab = "log\_price",  
 ylab = "Prediction")  
abline(a=0, b = 1,col = "red")



plot(x = prediction\_test\_big, y = error\_test\_big,  
 main = "Error of Prediction",  
 lwd = 0.5,  
 type = "p",  
 cex = 0.5,  
 xlab = "Prediction",  
 ylab = "Error")  
abline(h = 0,col = "red")



ssto\_big = sum((test\_big2$log\_price - mean(test\_big2$log\_price))^2)  
ssto\_big

## [1] 717.7368

sse\_big = sum(error\_test\_big^2)  
sse\_big

## [1] 51.04277

ssp\_big = sum((prediction\_test\_big - mean(test\_big2$log\_price))^2)  
ssp\_big

## [1] 672.8299

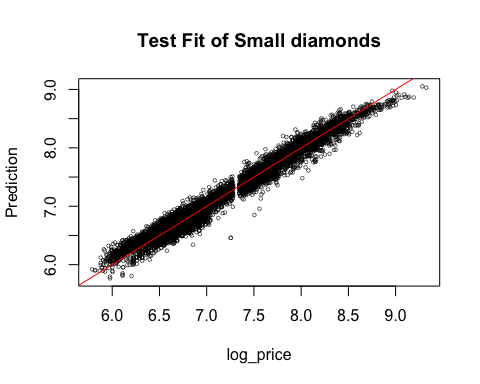
ssp\_big+sse\_big

## [1] 723.8726

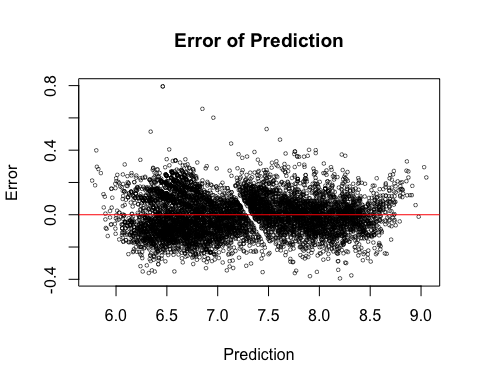
ssp\_big/ssto\_big

## [1] 0.9374325

prediction\_test\_small <- predict(step\_fit\_small2, newdata = test\_small2)  
error\_test\_small <- test\_small2$log\_price - prediction\_test\_small  
plot(x = test\_small2$log\_price, y = prediction\_test\_small,  
 main = "Test Fit of Small diamonds",  
 lwd = 0.5,  
 type = "p",  
 cex = 0.5,  
 xlab = "log\_price",  
 ylab = "Prediction")  
abline(a=0, b = 1,col = "red")



plot(x = prediction\_test\_small, y = error\_test\_small,  
 main = "Error of Prediction",  
 lwd = 0.5,  
 type = "p",  
 cex = 0.5,  
 xlab = "Prediction",  
 ylab = "Error")  
abline(h = 0,col = "red")



ssto\_small = sum((test\_small2$log\_price - mean(test\_small2$log\_price))^2)  
ssto\_small

## [1] 3611.437

sse\_small = sum(error\_test\_small^2)  
sse\_small

## [1] 107.4503

ssp\_small = sum((prediction\_test\_small - mean(test\_small2$log\_price))^2)  
ssp\_small

## [1] 3480.672

ssp\_small+sse\_small

## [1] 3588.123

ssp\_small/ssto\_small

## [1] 0.9637915

## need to assign a best model from above analysis  
best\_fit\_big <- fit\_all\_big ### 【need to choose】  
best\_fit\_small <- fit\_all\_small  
  
## residuals  
plot(fit\_all\_big$fitted.values, fit\_all\_big$residuals)

