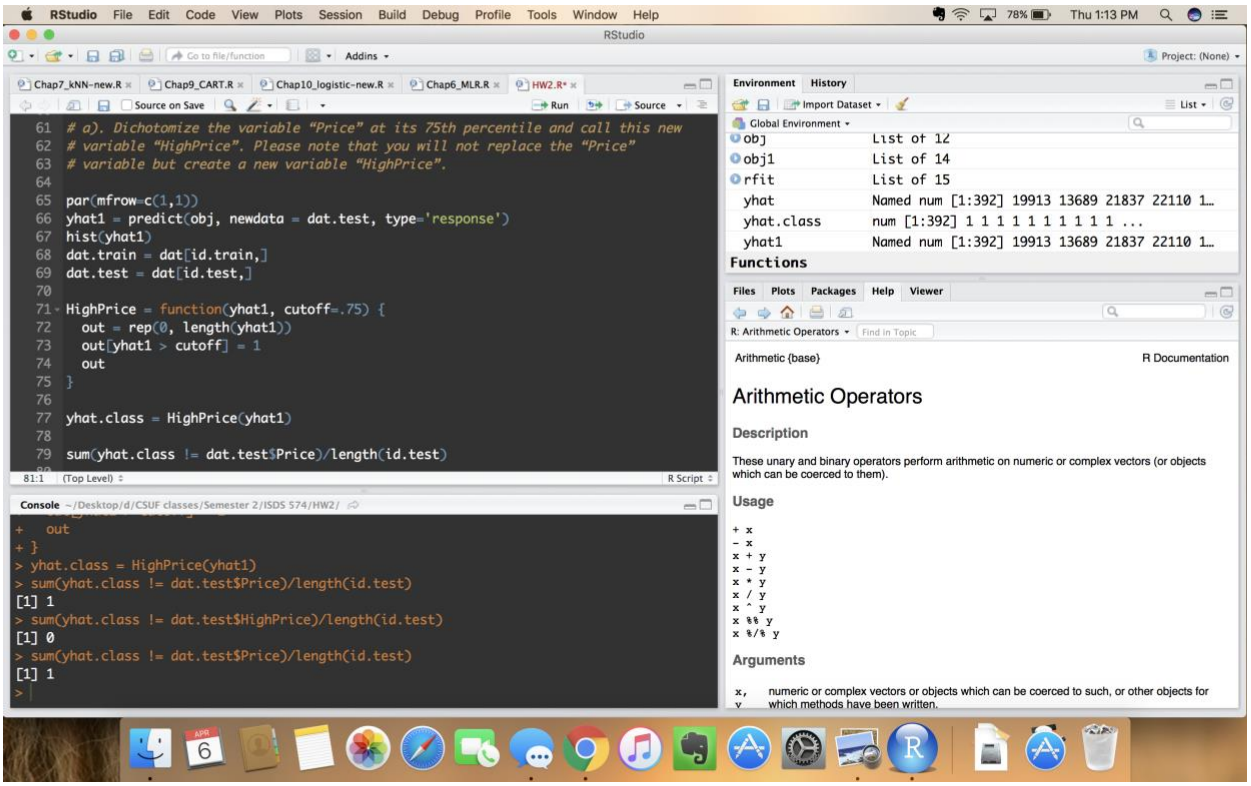
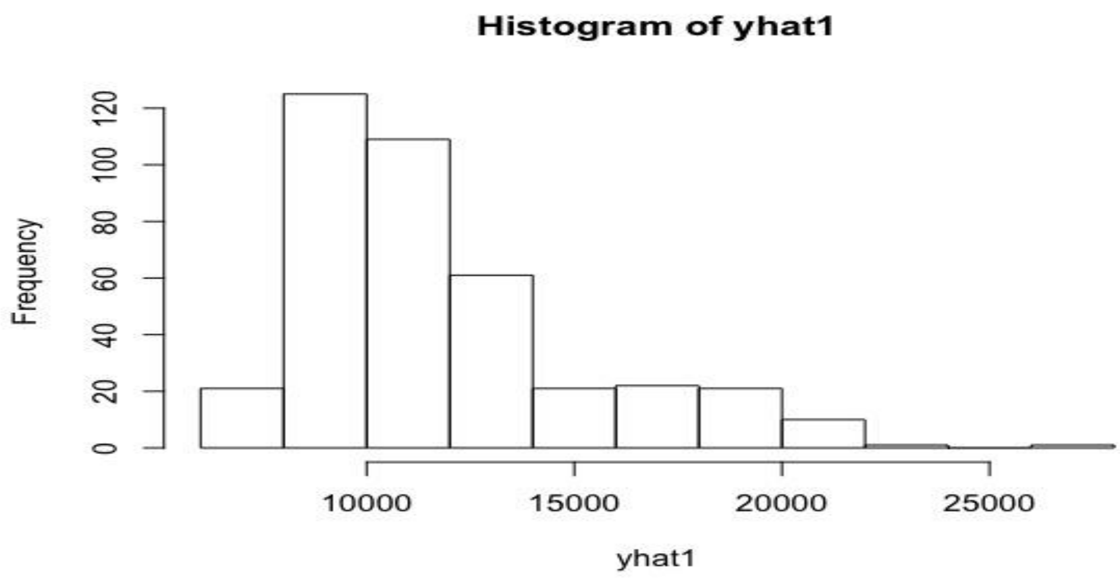
Archit Jajoo
ISDS 574

HOMEWORK 2

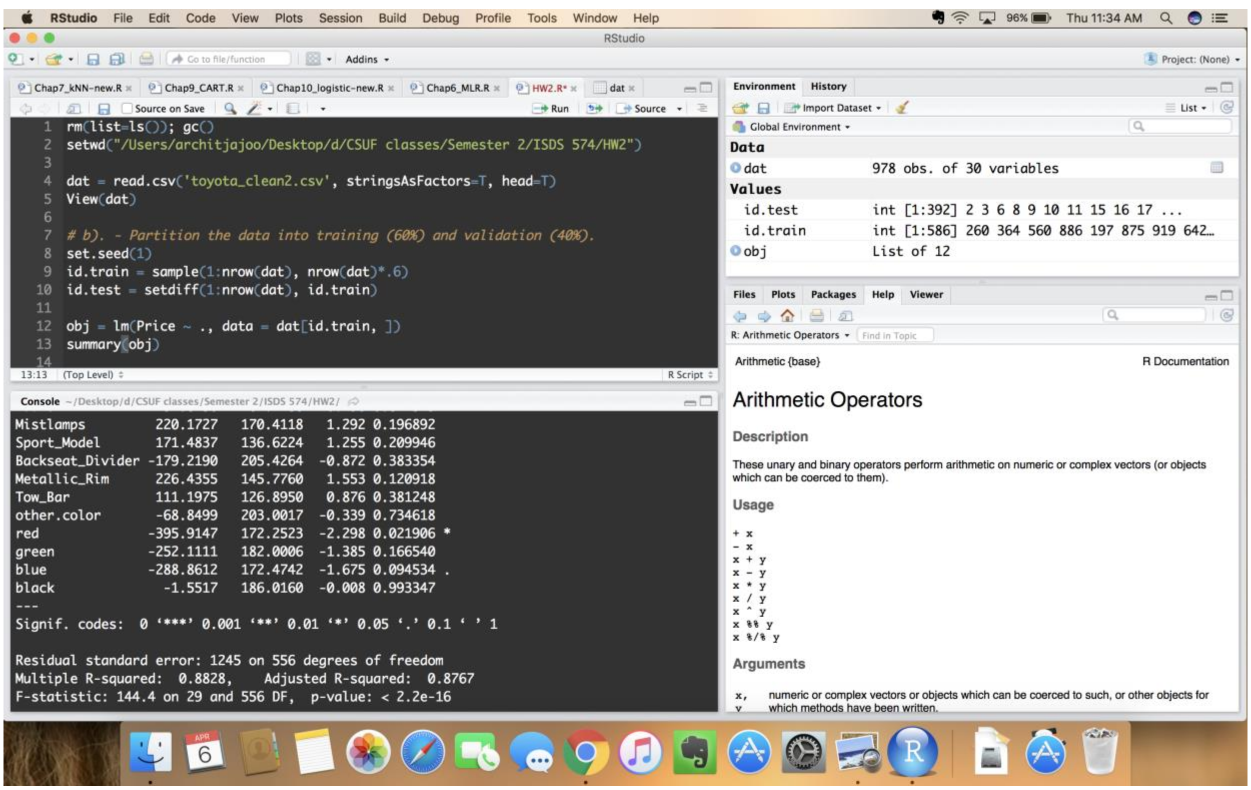
a) Dichotomize the variable “Price” at its 75th percentile and call this new variable “HighPrice”. Please note that you will not replace the “Price” variable but create a new variable “HighPrice”.



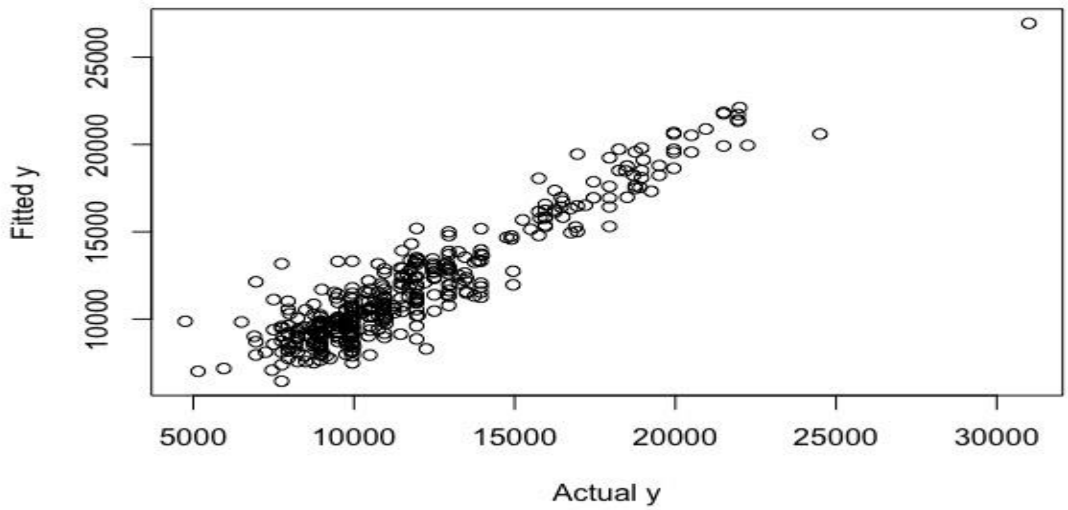
Archit Jajoo
ISDS 574

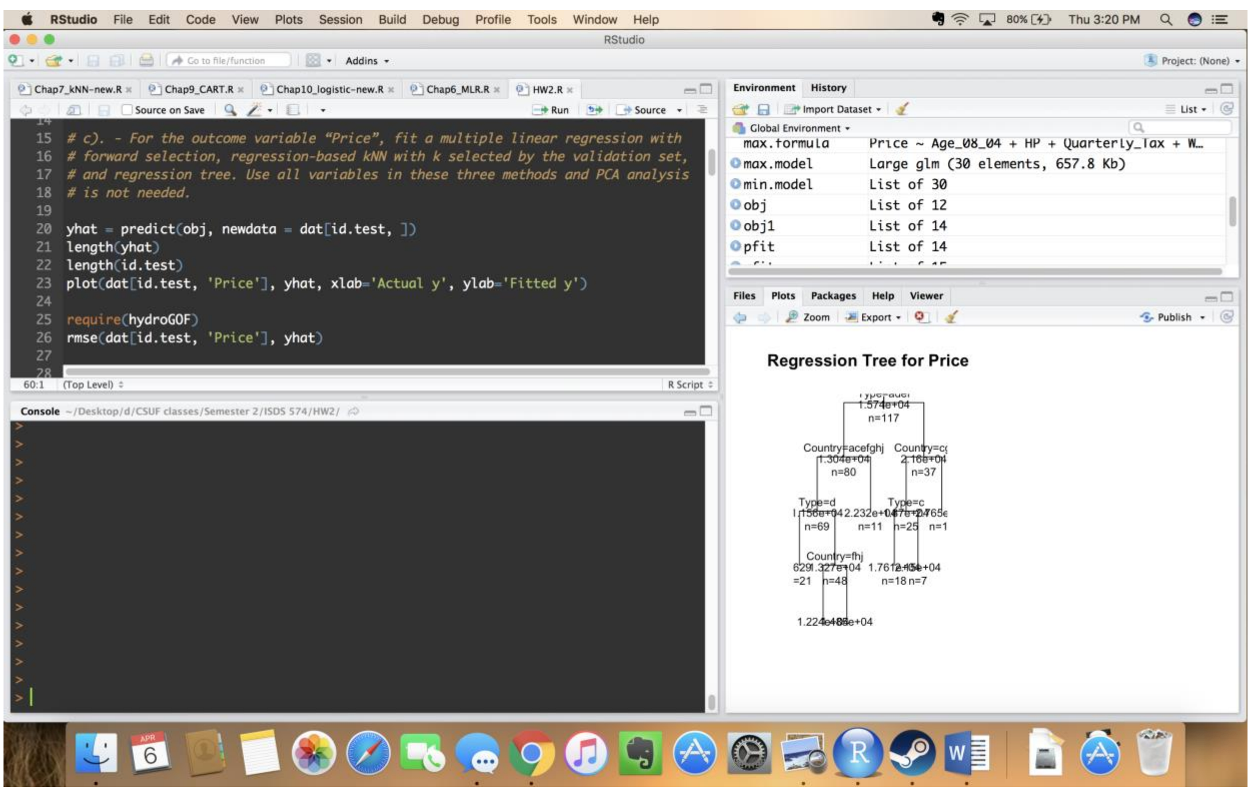


b) Partition the data into training (60%) and validation (40%).

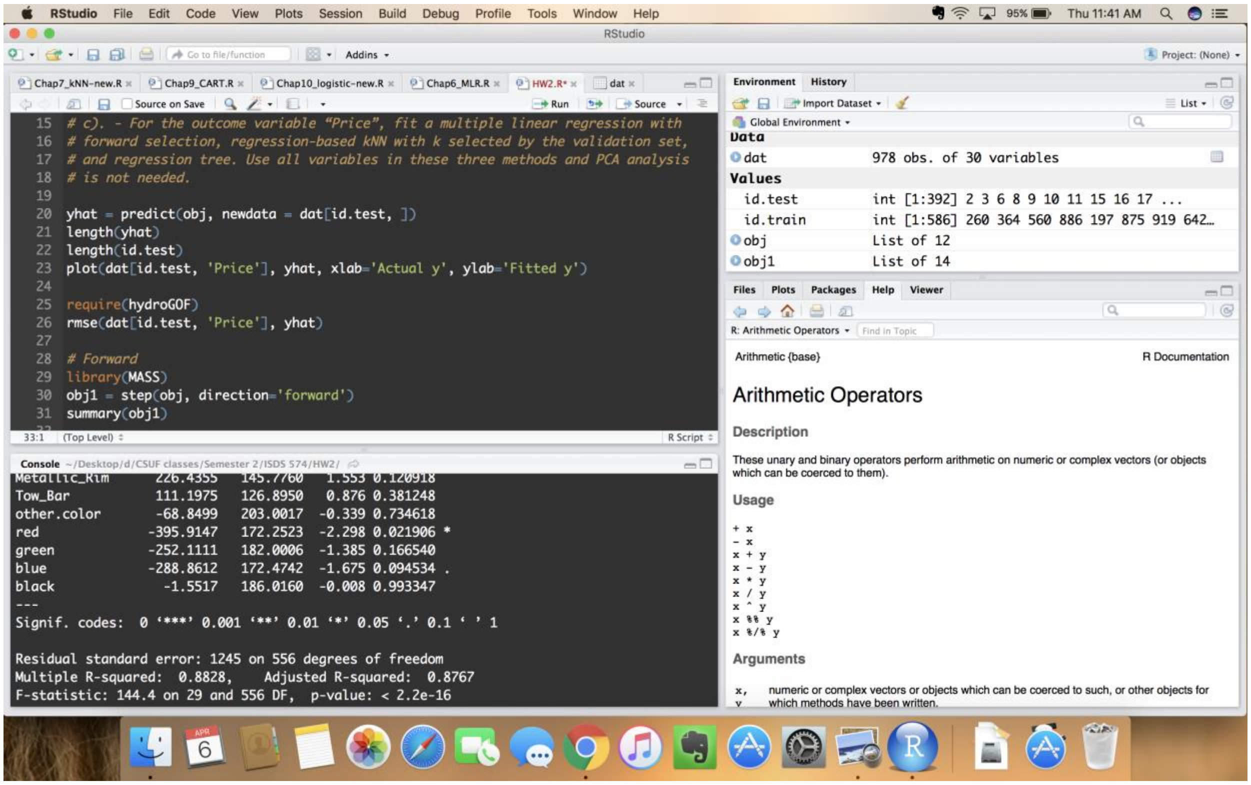


c) For the outcome variable "Price", fit a multiple linear regression with forward selection, regression-based kNN with k selected by the validation set, and regression tree. Use all variables in these three methods and PCA analysis is not needed.



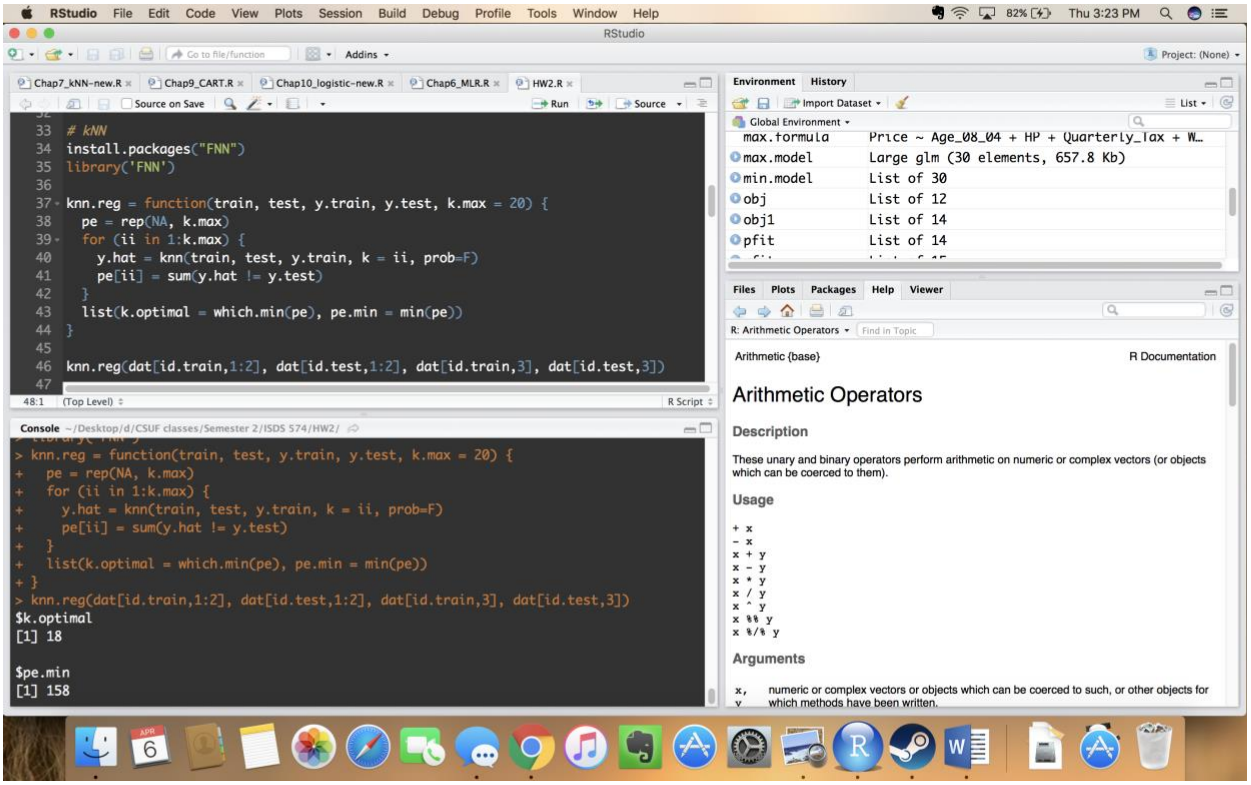


Forward selection:

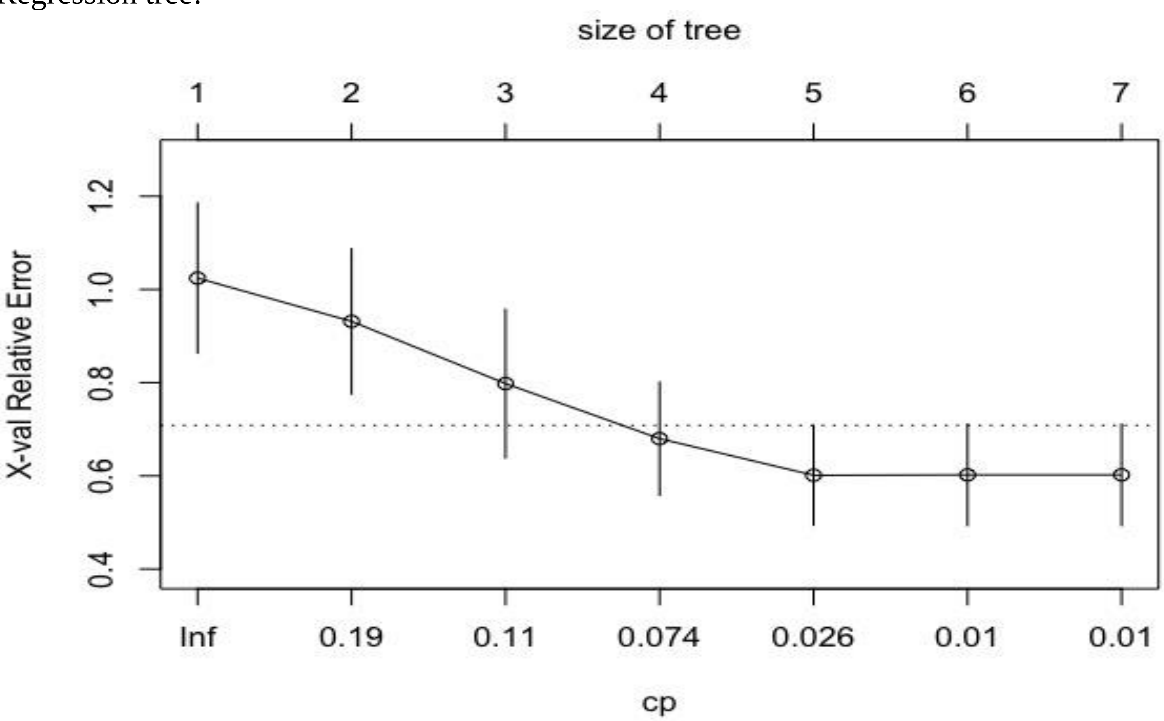

RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help
Thu 11:41 AM
RStudio
Chap7_kNN-new.R
Chap9_CART.R
Chap10_logistic-new.R
Chap6_MLR.R
HW2.R*
dat
Source on Save
Run
Source
# c). - For the outcome variable "Price", fit a multiple linear regression with
# forward selection, regression-based kNN with k selected by the validation set,
# and regression tree. Use all variables in these three methods and PCA analysis
# is not needed.
yhat = predict(obj, newdata = dat[id.test,])
length(yhat)
length(id.test)
plot(dat[id.test, 'Price'], yhat, xlab='Actual y', ylab='Fitted y')
require(hydroGOF)
rmse(dat[id.test, 'Price'], yhat)
# Forward
library(MASS)
obj1 = step(obj, direction='forward')
summary(obj1)
Console ~/Desktop/d/CSUF classes/Semester 2/ISDS 574/HW2/
Tow_Bar 111.1975 126.8950 0.876 0.381248
other.color -68.8499 203.0017 -0.339 0.734618
red -395.9147 172.2523 -2.298 0.021906 *
green -252.1111 182.0006 -1.385 0.166540
blue -288.8612 172.4742 -1.675 0.094534 .
black -1.5517 186.0160 -0.008 0.993347

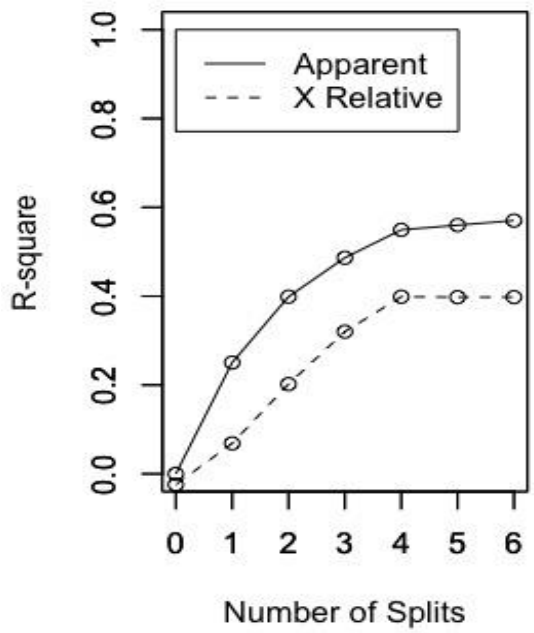
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1245 on 556 degrees of freedom
Multiple R-squared: 0.8828, Adjusted R-squared: 0.8767
F-statistic: 144.4 on 29 and 556 DF, p-value: < 2.2e-16
Environment History
Import Dataset
Global Environment
Data
dat 978 obs. of 30 variables
Values
id.test int [1:392] 2 3 6 8 9 10 11 15 16 17 ...
id.train int [1:586] 260 364 560 886 197 875 919 642…
obj List of 12
obj1 List of 14
Files Plots Packages Help Viewer
R: Arithmetic Operators
Arithmetic {base}
R Documentation
Arithmetic Operators
Description
These unary and binary operators perform arithmetic on numeric or complex vectors (or objects which can be coerced to them).
Usage
Arguments
numeric or complex vectors or objects which can be coerced to such, or other objects for which methods have been written.

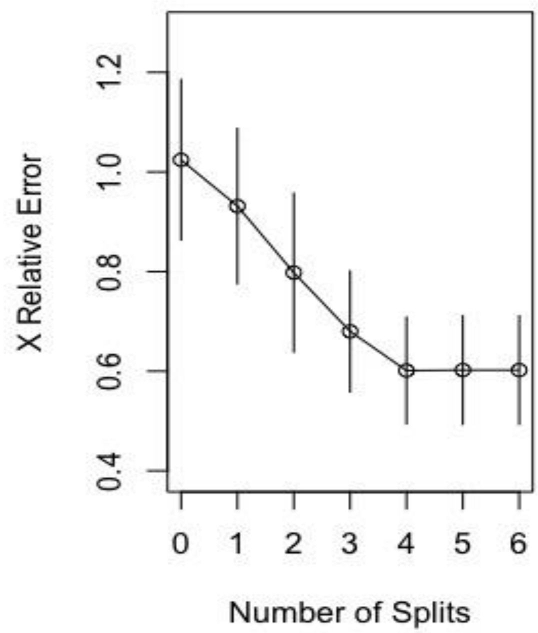
kNN:

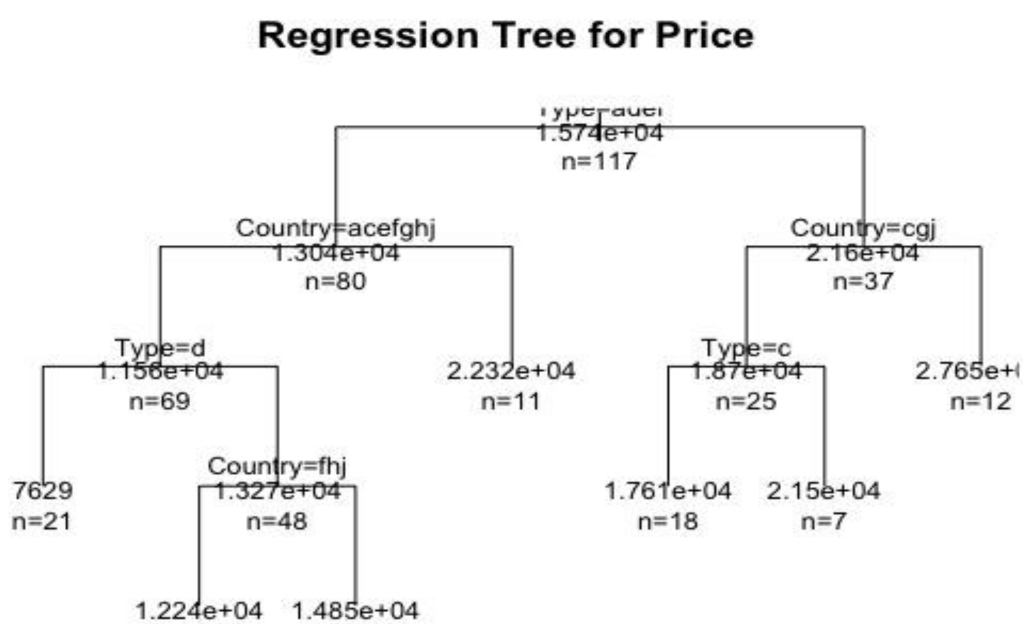


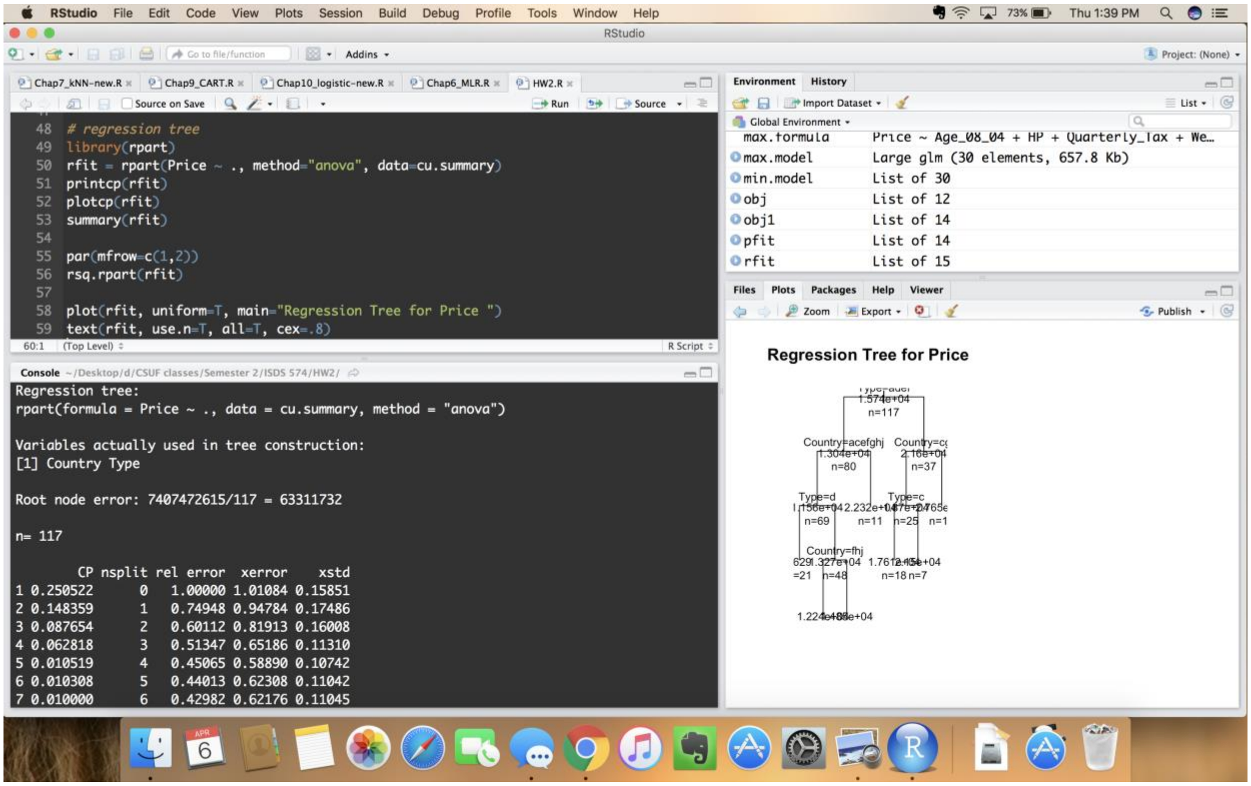
Regression tree:




Apparent
X Relative
R-square
0.0
0.2
0.4
0.6
0.8
1.0
0 1 2 3 4 5 6
Number of Splits


X Relative Error
0.4
0.6
0.8
1.0
1.2
0 1 2 3 4 5 6
Number of Splits


Regression Tree for Price
1.574e+04
n=117
Country=acefghj
1.304e+04
n=80
Country=cgj
2.16e+04
n=37
Type=d
1.156e+04
n=69
2.232e+04
n=11
Type=c
1.87e+04
n=25
2.765e+
n=12
7629
n=21
Country=fhj
1.327e+04
n=48
1.761e+04
n=18
2.15e+04
n=7
1.224e+04
1.485e+04

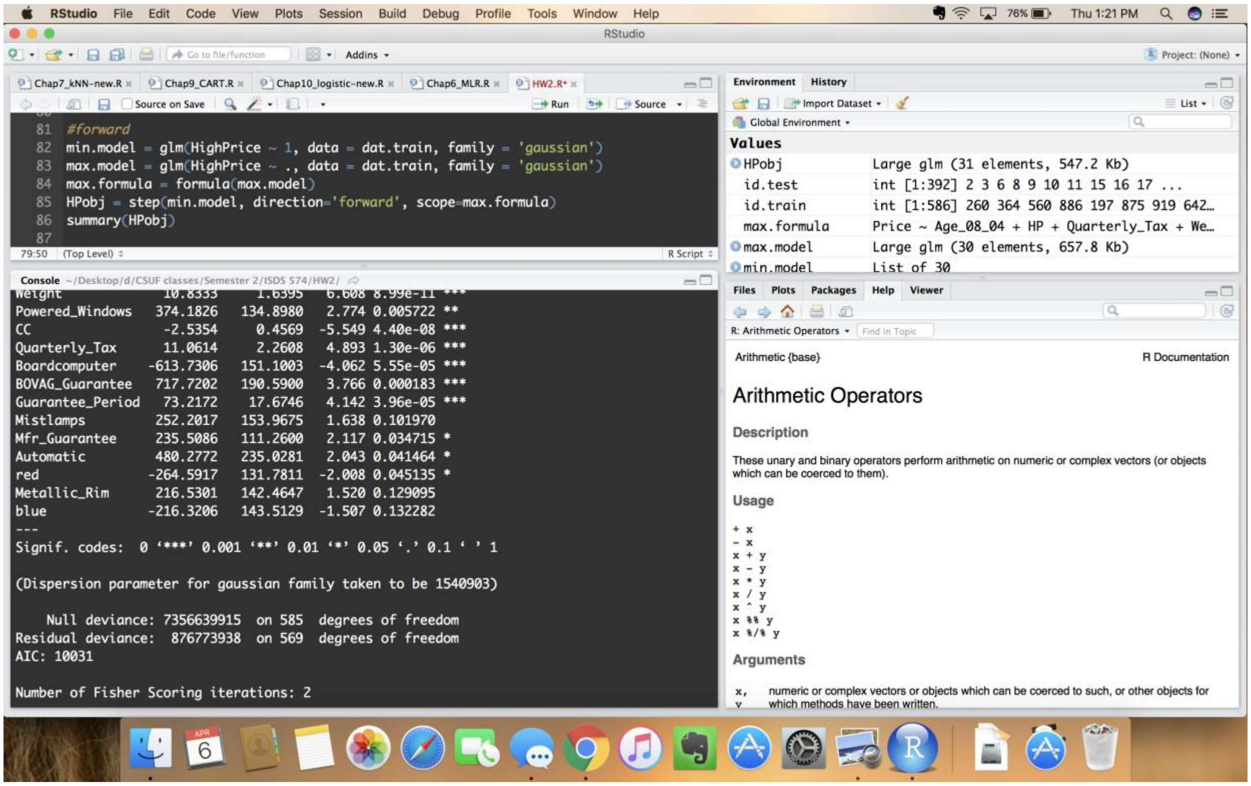


d) Compare the above three models obtained and pick up the best one. Explain why you think it is the best one. Note that I already took out the "grey" as the reference group for color. You don't have to do any data cleaning in this homework.

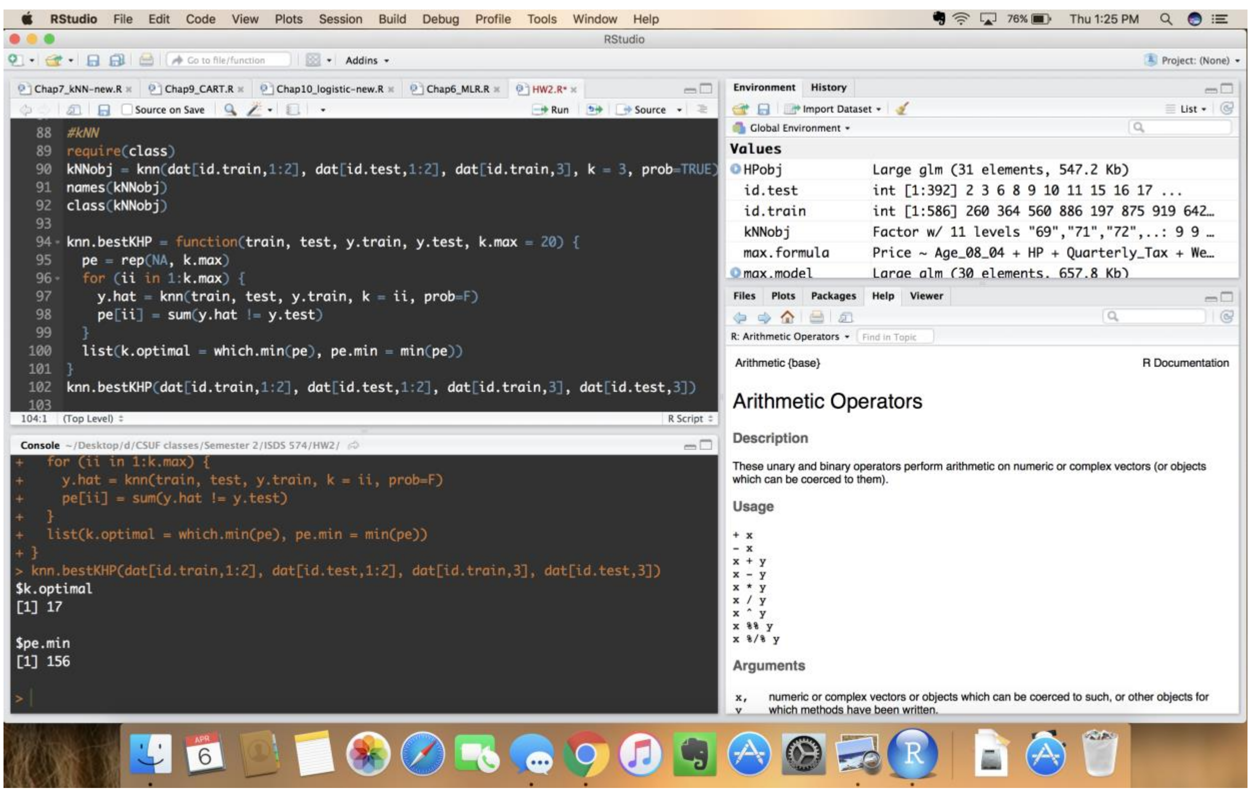
Multiple Linear Regression through forward selection is the best model because it gives lowest RMSE.

e) For the outcome variable "HighPrice", fit a logistic regression with forward regression, classification-based kNN with k selected by the validation set, and classification tree.

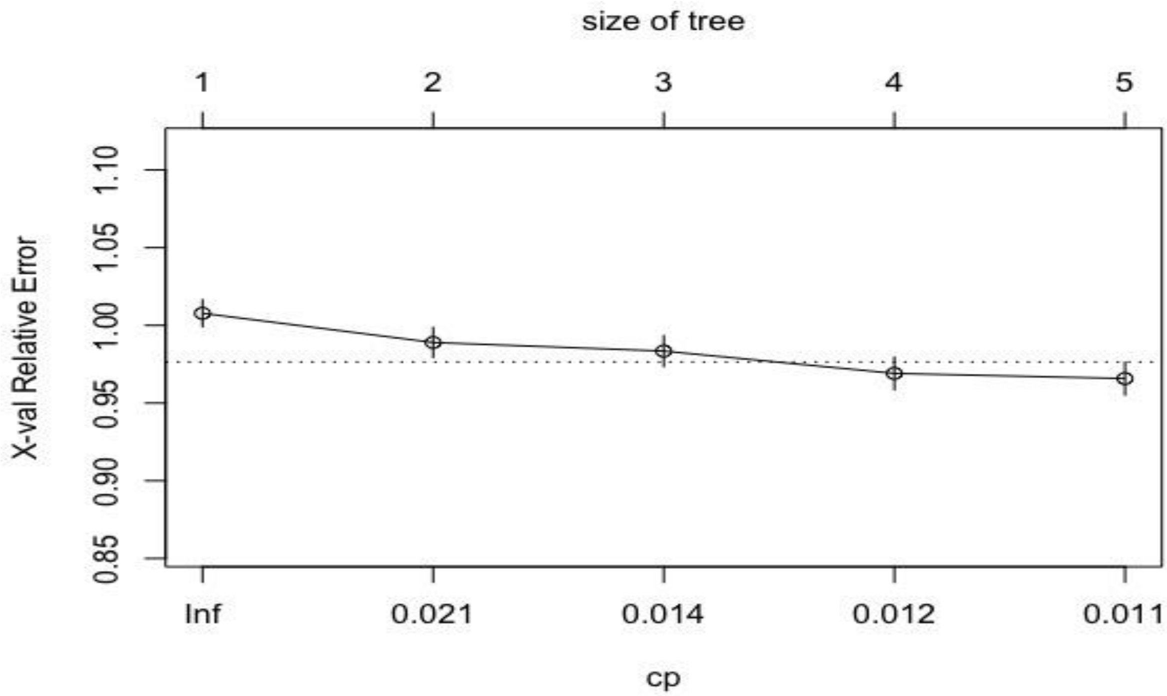
Forward:


#forward
min.model = glm(HighPrice ~ 1, data = dat.train, family = 'gaussian')
max.model = glm(HighPrice ~ ., data = dat.train, family = 'gaussian')
max.formula = formula(max.model)
HPobj = step(min.model, direction='forward', scope=max.formula)
summary(HPobj)
Powered_Windows 374.1826 134.8980 2.774 0.005722 **
CC -2.5354 0.4569 -5.549 4.40e-08 ***
Quarterly_Tax 11.0614 2.2608 4.893 1.30e-06 ***
Boardcomputer -613.7306 151.1003 -4.062 5.55e-05 ***
BOVAG_Guarantee 717.7202 190.5900 3.766 0.000183 ***
Guarantee_Period 73.2172 17.6746 4.142 3.96e-05 ***
Mistlamps 252.2017 153.9675 1.638 0.101970
Mfr_Guarantee 235.5086 111.2600 2.117 0.034715 *
Automatic 480.2772 235.0281 2.043 0.041464 *
red -264.5917 131.7811 -2.008 0.045135 *
Metallic_Rim 216.5301 142.4647 1.520 0.129095
blue -216.3206 143.5129 -1.507 0.132282
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for gaussian family taken to be 1540903)
Null deviance: 7356639915 on 585 degrees of freedom
Residual deviance: 876773938 on 569 degrees of freedom
AIC: 10031
Number of Fisher Scoring iterations: 2
Values
HPobj Large glm (31 elements, 547.2 Kb)
id.test int [1:392] 2 3 6 8 9 10 11 15 16 17 ...
id.train int [1:586] 260 364 560 886 197 875 919 642...
max.formula Price ~ Age_08_04 + HP + Quarterly_Tax + We...
max.model Large glm (30 elements, 657.8 Kb)
min.model List of 30
Arithmetic Operators

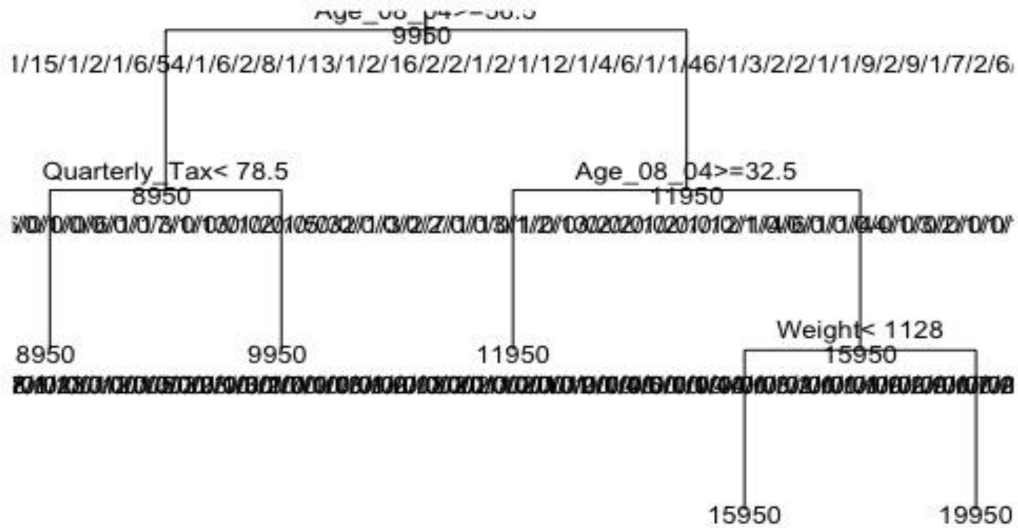
kNN:

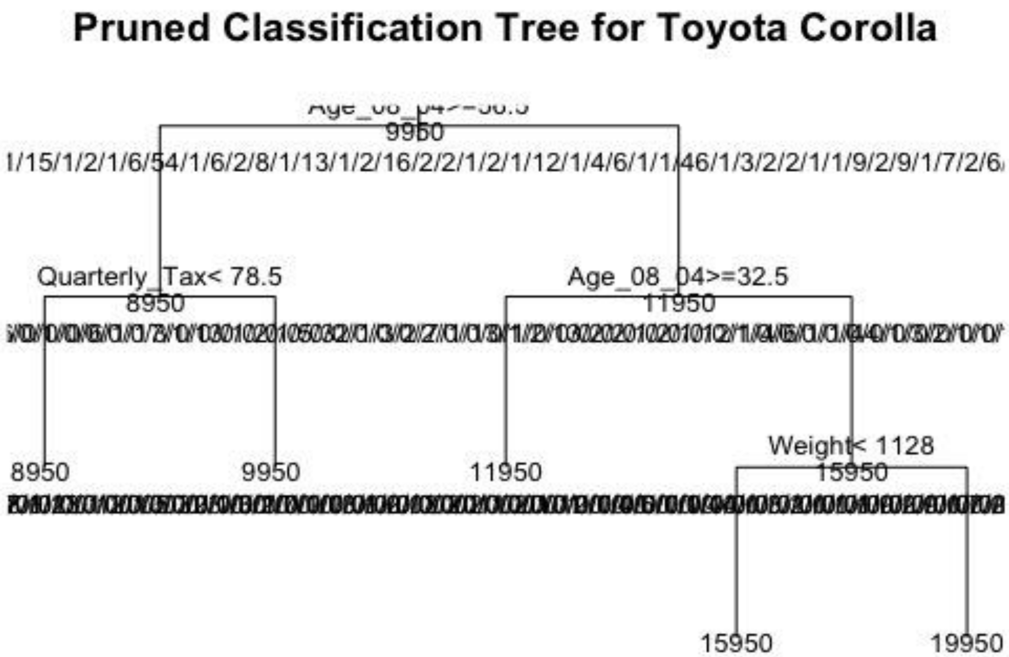

#kNN
require(class)
kNNobj = knn(dat[id.train,1:2], dat[id.test,1:2], dat[id.train,3], k = 3, prob=TRUE)
names(kNNobj)
class(kNNobj)
knn.bestKHP = function(train, test, y.train, y.test, k.max = 20) {
pe = rep(NA, k.max)
for (ii in 1:k.max) {
y.hat = knn(train, test, y.train, k = ii, prob=F)
pe[ii] = sum(y.hat != y.test)
}
list(k.optimal = which.min(pe), pe.min = min(pe))
}
knn.bestKHP(dat[id.train,1:2], dat[id.test,1:2], dat[id.train,3], dat[id.test,3])
$k.optimal
[1] 17
$pe.min
[1] 156
Values
HPobj Large glm (31 elements, 547.2 Kb)
id.test int [1:392] 2 3 6 8 9 10 11 15 16 17 ...
id.train int [1:586] 260 364 560 886 197 875 919 642...
kNNobj Factor w/ 11 levels "69","71","72",..: 9 9 ...
max.formula Price ~ Age_08_04 + HP + Quarterly_Tax + We...
Arithmetic Operators

Classification tree:

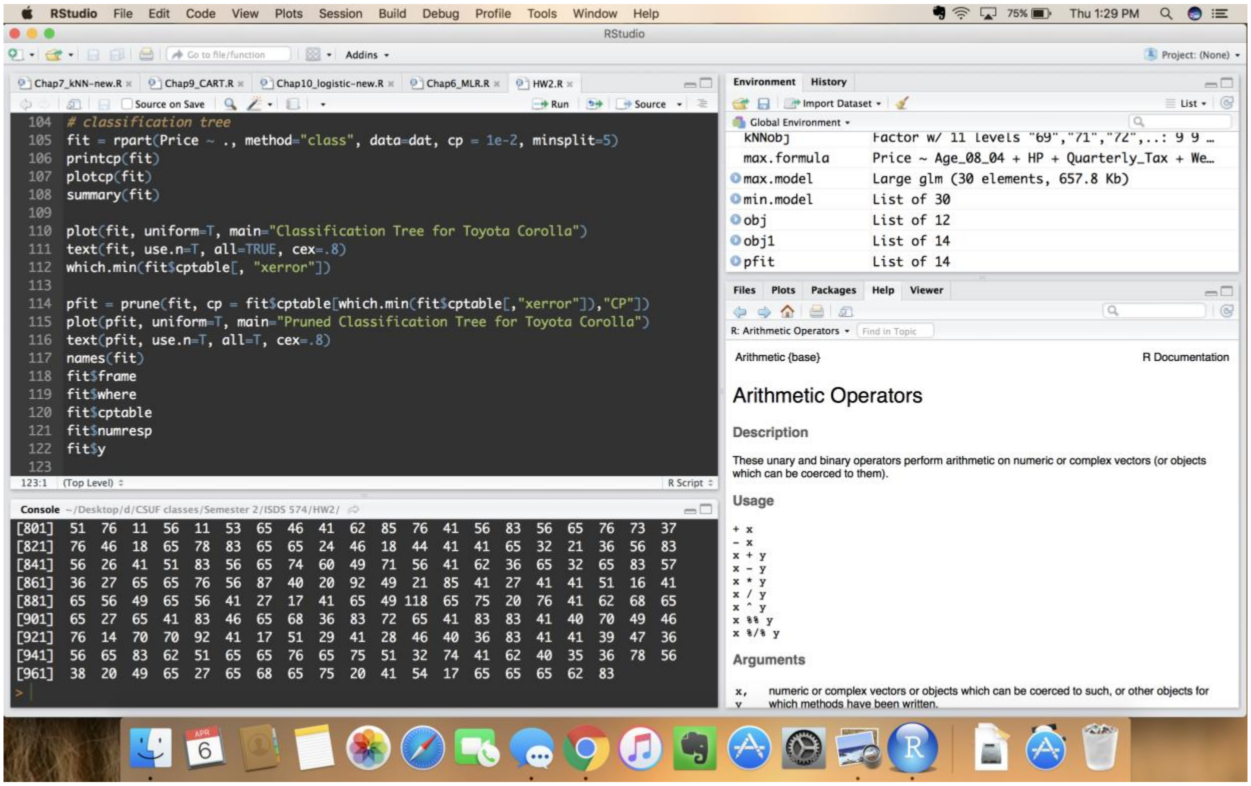


Classification Tree for Toyota Corolla




Pruned Classification Tree for Toyota Corolla
9950
Quarterly_Tax< 78.5
8950
Age_08_04>=32.5
11950
Weight< 1128
15950
8950
9950
11950
15950
19950

Archit Jajoo
ISDS 574



. **f)** Compare the models obtained and pick up the best one. Explain why you think it is the best one.

Classification tree is the best model because it gives lowest misclassification rate.