M6L2 Homework Assignment

Joshua Conte

October 29, 2017

1 M6L2 Homework Assignment

R studio was configured with the following parameters before beginning the project:

```
# clears the console in RStudio
cat("\014")
```

```
# clears environment
rm(list = ls())

# Load required packages
require("ggplot2");
require("C50");
require("gmodels");
require("rpart");
require("RColorBrewer");
require("tree");
require("party");
```

1.1 Load Data.

I opened the Wholesale customers Data Set using read.csv2 and downloaded it directly from the UC Irvine Machine Learning Repository.

To format the data, the data is separated by ‘,’, stringsAsFactors = FALSE so that the strings in a data frame will be treated as plain strings and not as factor variables. I set na strings for missing data. Once the data was loaded I added the column names and changed the data types to numeric and finally removed the text data type.

Below is my R code:

```
# Some csv files are really big and take a while to open.  This command checks to
# see if it is already opened, if it is, it does not open it again.
# I also omitted the first column
if (!exists("dfWCD")) {
dfWCD <-
  read.csv2("Wholesale customers data.csv",
    sep = ",",
    stringsAsFactors = FALSE,
    na.strings=c("","NA")
  )
}

# Download directly from site (unreliable from Ecuador)
# if (!exists("dfWCD")) {
# dfWCD <-
#   read.csv2(
#     url(
#       "https://archive.ics.uci.edu/ml/machine-learning-databases/00292/Wholesale customers data.csv"
#     ),
#     sep = ",",
#     stringsAsFactors = FALSE,
#     na.strings=c("","NA")
#   )
# # Add a column so I know which study the data is referring to
```

```
# study <- sprintf("study_%s",seq(1:440))
# dfWCD$study<-study
# }

# change 2 to 24 to numeric
dfWCD[1:8] <- sapply(dfWCD[1:8], as.numeric)

# Print first lines
str(dfWCD)
```

```
## 'data.frame':    440 obs. of  8 variables:
##  $ Channel         : num  2 2 2 1 2 2 2 2 1 2 ...
##  $ Region          : num  3 3 3 3 3 3 3 3 3 3 ...
##  $ Fresh           : num  12669 7057 6353 13265 22615 ...
##  $ Milk            : num  9656 9810 8808 1196 5410 ...
##  $ Grocery         : num  7561 9568 7684 4221 7198 ...
##  $ Frozen          : num  214 1762 2405 6404 3915 ...
##  $ Detergents_Paper: num  2674 3293 3516 507 1777 ...
##  $ Delicassen      : num  1338 1776 7844 1788 5185 ...
```

1.1.1 Understanding the data

The data set refers to clients of a wholesale distributor in Portugal. It includes the annual spending in monetary units (m.u.) on diverse product categories. The data has the following attribute information:

1. FRESH: annual spending (m.u.) on fresh products (Continuous);
2. MILK: annual spending (m.u.) on Fresh products (Continuous);
3. GROCERY: annual spending (m.u.)on grocery products (Continuous);
4. FROZEN: annual spending (m.u.)on frozen products (Continuous)
5. DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)
6. DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);
7. CHANNEL: customer channel - 1 = Horeca (Hotel/Restaurant/Cafe) or 2 = Retail
8. REGION: Customers Region - 1= Lisnon 2 = Oporto or 3 = Other (Nominal)

1.2 Decision Trees in R

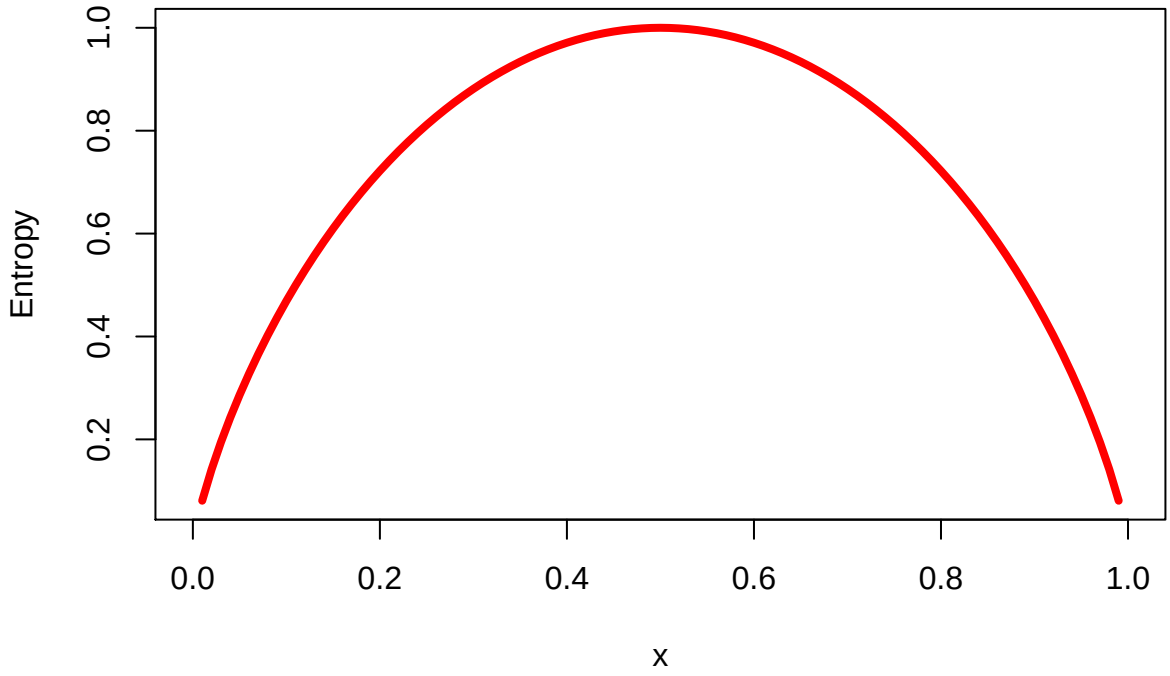
Top-down: Which attribute should ne the root?

We construct a tree from the top down starting with the question: which attribute should be tested at the root of the tree? That is, which attribute best splits/separates the labeled training data.

Then build subtrees recursively, asking the same question on the remaining attributes.

This model will predict the customers Region. **Step 1 - Decision Trees:**

```
#####  Decision Trees  ------------------

## Understanding Decision Trees ----
# calculate entropy of a two-class segment


curve(-x * log2(x) - (1 - x) * log2(1 - x),
      col="red", xlab = "x", ylab = "Entropy", lwd=4)
```




```
## Example: Identifying Mushroom Type: Either 'poisonous' or 'edible' ----
```

Step 2 - Exploring and preparing the data: This makes sure the data is random and uses the first 390 data points to train the last 50 data points.

```
str(dfWCD)
```

```
## 'data.frame':    440 obs. of  8 variables:
##  $ Channel         : num  2 2 2 1 2 2 2 2 1 2 ...
##  $ Region          : num  3 3 3 3 3 3 3 3 3 3 ...
##  $ Fresh           : num  12669 7057 6353 13265 22615 ...
##  $ Milk            : num  9656 9810 8808 1196 5410 ...
##  $ Grocery         : num  7561 9568 7684 4221 7198 ...
##  $ Frozen          : num  214 1762 2405 6404 3915 ...
##  $ Detergents_Paper: num  2674 3293 3516 507 1777 ...
##  $ Delicassen      : num  1338 1776 7844 1788 5185 ...
```

```
# look at the class variable
table(dfWCD$Channel)
```

```
##
##   1   2
## 298 142
```

```
# create a random sample for training and test data
set.seed(12345)
dfWCD_rand <- dfWCD[order(runif(440)), ]

# compare the original and random order data frames
```

```
summary(dfWCD$Channel)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.323   2.000   2.000
```

```
summary(dfWCD_rand$Channel)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.000   1.000   1.323   2.000   2.000
```

```
head(dfWCD$Channel)
```

```
## [1] 2 2 2 1 2 2
```

```
head(dfWCD_rand$Channel)
```

```
## [1] 2 1 1 2 1 1
```

```
# split the data frames
dfWCD_train <- dfWCD_rand[1:390, ]
dfWCD_test  <- dfWCD_rand[391:440, ]

# check the proportion of class variable
prop.table(table(dfWCD_train$Region))
```

```
##
##         1         2         3
## 0.1666667 0.1128205 0.7205128
```

```
prop.table(table(dfWCD_test$Region))
```

```
##
##    1    2    3
## 0.24 0.06 0.70
```

Step 3 - Training a model on the data: This uses the C5.0 algorithm:

```
# First convert this to a factor
dfWCD_train$Region<-as.factor(dfWCD_train$Region)
model <- C5.0(dfWCD_train[-1], dfWCD_train$Region)

# display simple facts about the tree
model
```

```
##
## Call:
## C5.0.default(x = dfWCD_train[-1], y = dfWCD_train$Region)
##
## Classification Tree
## Number of samples: 390
## Number of predictors: 7
##
## Tree size: 3
##
## Non-standard options: attempt to group attributes
```

```
# display detailed information about the tree
# This prints out a lot of lines of information that is not needed for the report.
summary(model)
```

```
##
## Call:
## C5.0.default(x = dfWCD_train[-1], y = dfWCD_train$Region)
##
##
## C5.0 [Release 2.07 GPL Edition]      Sat Oct 28 08:41:00 2017
## -------------------------------
##
## Class specified by attribute `outcome'
##
## Read 390 cases (8 attributes) from undefined.data
##
## Decision tree:
##
## Region = 1: 1 (65)
## Region = 2: 2 (44)
## Region = 3: 3 (281)
##
##
## Evaluation on training data (390 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##       3    0( 0.0%)   <<
##
##
##     (a)   (b)   (c)    <-classified as
##    ----  ----  ----
##      65                (a): class 1
##            44          (b): class 2
##                 281    (c): class 3
##
##
##  Attribute usage:
##
##  100.00% Region
##
##
## Time: 0.0 secs
```

Step 4 - Evaluating model performance: This evaluates how well the training model did:

```
# create a factor vector of predictions(model) on test data

dfWCD_Region_pred <- predict(model, dfWCD_test)

# cross tabulation of predicted versus actual classes
length(dfWCD_test$Region)
```

```
## [1] 50
```

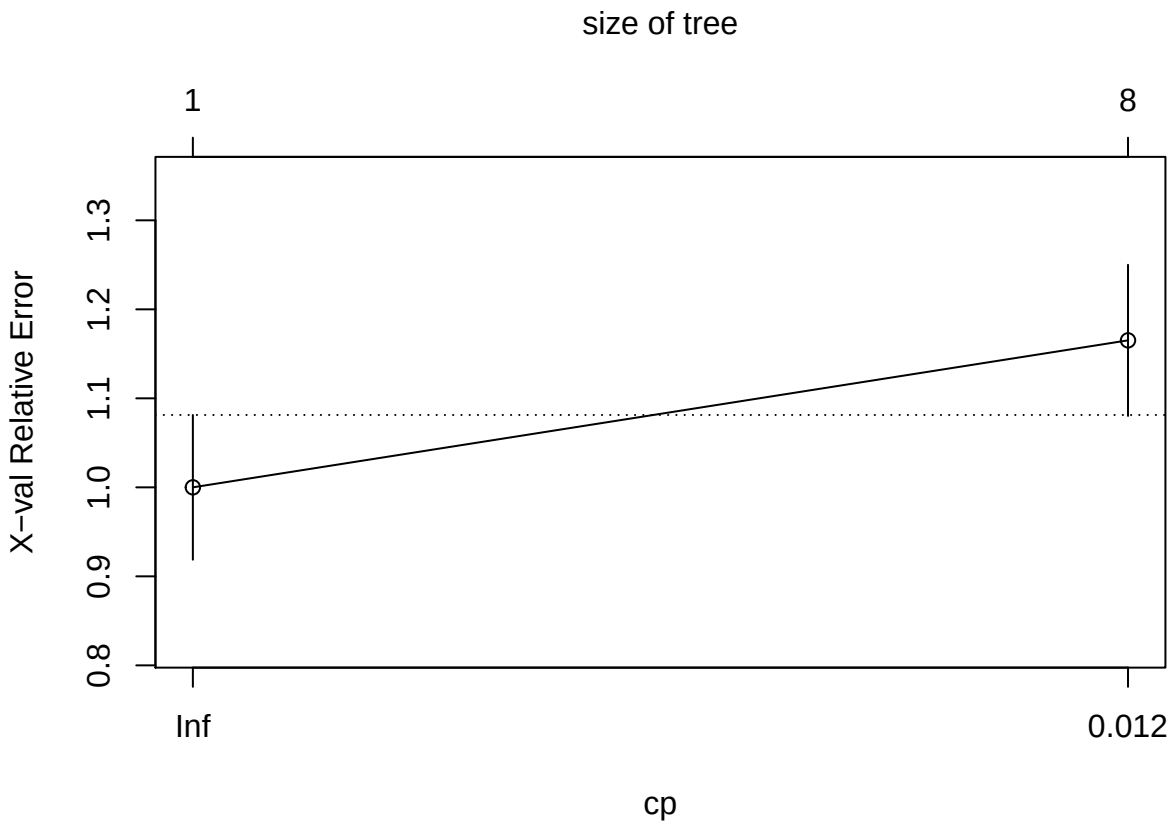
```
CrossTable(dfWCD_test$Region, dfWCD_Region_pred,
           prop.chisq = FALSE, prop.c = FALSE, prop.r = FALSE,
           dnn = c('actual Region', 'predicted Region'))
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  50
##
##
##                | predicted Region
## actual Region |         1 |         2 |         3 | Row Total |
## --------------|-----------|-----------|-----------|-----------|
##             1 |        12 |         0 |         0 |        12 |
##               |     0.240 |     0.000 |     0.000 |           |
## --------------|-----------|-----------|-----------|-----------|
##             2 |         0 |         3 |         0 |         3 |
##               |     0.000 |     0.060 |     0.000 |           |
## --------------|-----------|-----------|-----------|-----------|
##             3 |         0 |         0 |        35 |        35 |
##               |     0.000 |     0.000 |     0.700 |           |
## --------------|-----------|-----------|-----------|-----------|
##  Column Total |        12 |         3 |        35 |        50 |
## --------------|-----------|-----------|-----------|-----------|
##
##
```

```
formula<-Region  ~ Fresh + Milk + Grocery + Frozen + Detergents_Paper + Delicassen

fit = rpart(formula, method="class", data=dfWCD_train)

printcp(fit) # display the results
```

```
##
## Classification tree:
## rpart(formula = formula, data = dfWCD_train, method = "class")
##
## Variables actually used in tree construction:
## [1] Fresh  Frozen Milk
##
## Root node error: 109/390 = 0.27949
##
## n= 390
##
##         CP nsplit rel error xerror     xstd
## 1 0.014417      0   1.00000 1.0000 0.081303
## 2 0.010000      7   0.89908 1.1651 0.084903
```

```
plotcp(fit) # visualize cross-validation results
```




```
summary(fit) # detailed summary of splits
```

```
## Call:
## rpart(formula = formula, data = dfWCD_train, method = "class")
##   n= 390
##
##           CP nsplit rel error   xerror       xstd
## 1 0.01441678      0 1.0000000 1.000000 0.08130319
## 2 0.01000000      7 0.8990826 1.165138 0.08490259
##
## Variable importance
##           Frozen            Fresh             Milk Detergents_Paper
##               34               23               19               10
##          Grocery       Delicassen
##                8                6
##
## Node number 1: 390 observations,    complexity param=0.01441678
##   predicted class=3  expected loss=0.2794872  P(node) =1
##     class counts:    65    44   281
##    probabilities: 0.167 0.113 0.721
##   left son=2 (333 obs) right son=3 (57 obs)
##   Primary splits:
##       Frozen     < 409.5   to the right, improve=2.502173, (0 missing)
##       Milk       < 2753.5  to the left,  improve=1.764103, (0 missing)
##       Fresh      < 11233   to the left,  improve=1.545527, (0 missing)
##       Delicassen < 3868    to the left,  improve=1.259707, (0 missing)
```

```
##       Grocery    < 2126    to the left,  improve=1.161477, (0 missing)
##   Surrogate splits:
##       Fresh < 150.5   to the right, agree=0.862, adj=0.053, (0 split)
##
## Node number 2: 333 observations,    complexity param=0.01441678
##   predicted class=3  expected loss=0.3063063  P(node) =0.8538462
##     class counts:    61    41   231
##    probabilities: 0.183 0.123 0.694
##   left son=4 (192 obs) right son=5 (141 obs)
##   Primary splits:
##       Fresh      < 11233   to the left,  improve=1.928816, (0 missing)
##       Delicassen < 3868    to the left,  improve=1.650582, (0 missing)
##       Grocery    < 2178    to the left,  improve=1.591273, (0 missing)
##       Milk       < 2753.5  to the left,  improve=1.570898, (0 missing)
##       Frozen     < 530     to the left,  improve=1.235812, (0 missing)
##   Surrogate splits:
##       Frozen           < 2970.5  to the left,  agree=0.643, adj=0.156, (0 split)
##       Delicassen       < 1782    to the left,  agree=0.628, adj=0.121, (0 split)
##       Milk             < 27899   to the left,  agree=0.598, adj=0.050, (0 split)
##       Grocery          < 511.5   to the right, agree=0.592, adj=0.035, (0 split)
##       Detergents_Paper < 46.5    to the right, agree=0.592, adj=0.035, (0 split)
##
## Node number 3: 57 observations
##   predicted class=3  expected loss=0.122807  P(node) =0.1461538
##     class counts:     4     3    50
##    probabilities: 0.070 0.053 0.877
##
## Node number 4: 192 observations,    complexity param=0.01441678
##   predicted class=3  expected loss=0.359375  P(node) =0.4923077
##     class counts:    41    28   123
##    probabilities: 0.214 0.146 0.641
##   left son=8 (17 obs) right son=9 (175 obs)
##   Primary splits:
##       Frozen           < 6328    to the right, improve=2.425760, (0 missing)
##       Detergents_Paper < 94      to the left,  improve=1.555544, (0 missing)
##       Milk             < 1955    to the left,  improve=1.385795, (0 missing)
##       Grocery          < 4573    to the right, improve=1.375873, (0 missing)
##       Delicassen       < 3868    to the left,  improve=1.264634, (0 missing)
##
## Node number 5: 141 observations
##   predicted class=3  expected loss=0.2340426  P(node) =0.3615385
##     class counts:    20    13   108
##    probabilities: 0.142 0.092 0.766
##
## Node number 8: 17 observations
##   predicted class=1  expected loss=0.5882353  P(node) =0.04358974
##     class counts:     7     4     6
##    probabilities: 0.412 0.235 0.353
##
## Node number 9: 175 observations,    complexity param=0.01441678
##   predicted class=3  expected loss=0.3314286  P(node) =0.4487179
##     class counts:    34    24   117
##    probabilities: 0.194 0.137 0.669
##   left son=18 (93 obs) right son=19 (82 obs)
```

```
##   Primary splits:
##       Frozen           < 1487.5  to the left,  improve=1.778505, (0 missing)
##       Grocery          < 4573    to the right, improve=1.480733, (0 missing)
##       Milk             < 11099   to the right, improve=1.332381, (0 missing)
##       Fresh            < 9714.5  to the right, improve=1.300833, (0 missing)
##       Detergents_Paper < 3566.5  to the right, improve=1.300168, (0 missing)
##   Surrogate splits:
##       Fresh            < 2739.5  to the left,  agree=0.640, adj=0.232, (0 split)
##       Milk             < 8556    to the left,  agree=0.577, adj=0.098, (0 split)
##       Delicassen       < 2486.5  to the left,  agree=0.577, adj=0.098, (0 split)
##       Grocery          < 2016    to the right, agree=0.566, adj=0.073, (0 split)
##       Detergents_Paper < 5964.5  to the left,  agree=0.560, adj=0.061, (0 split)
##
## Node number 18: 93 observations,    complexity param=0.01441678
##   predicted class=3  expected loss=0.4086022  P(node) =0.2384615
##     class counts:    22    16    55
##    probabilities: 0.237 0.172 0.591
##   left son=36 (11 obs) right son=37 (82 obs)
##   Primary splits:
##       Milk             < 11099   to the right, improve=3.043845, (0 missing)
##       Detergents_Paper < 2116    to the right, improve=2.439359, (0 missing)
##       Grocery          < 5176    to the right, improve=2.366778, (0 missing)
##       Frozen           < 1107.5  to the right, improve=2.042503, (0 missing)
##       Fresh            < 6695.5  to the right, improve=1.041491, (0 missing)
##   Surrogate splits:
##       Grocery          < 16625   to the right, agree=0.903, adj=0.182, (0 split)
##       Detergents_Paper < 5890    to the right, agree=0.892, adj=0.091, (0 split)
##       Delicassen       < 2959    to the right, agree=0.892, adj=0.091, (0 split)
##
## Node number 19: 82 observations
##   predicted class=3  expected loss=0.2439024  P(node) =0.2102564
##     class counts:    12     8    62
##    probabilities: 0.146 0.098 0.756
##
## Node number 36: 11 observations
##   predicted class=2  expected loss=0.4545455  P(node) =0.02820513
##     class counts:     2     6     3
##    probabilities: 0.182 0.545 0.273
##
## Node number 37: 82 observations,    complexity param=0.01441678
##   predicted class=3  expected loss=0.3658537  P(node) =0.2102564
##     class counts:    20    10    52
##    probabilities: 0.244 0.122 0.634
##   left son=74 (24 obs) right son=75 (58 obs)
##   Primary splits:
##       Frozen           < 1107.5  to the right, improve=2.6595880, (0 missing)
##       Milk             < 4981.5  to the right, improve=1.2690540, (0 missing)
##       Fresh            < 6336.5  to the right, improve=0.9411971, (0 missing)
##       Grocery          < 5062    to the right, improve=0.8915709, (0 missing)
##       Detergents_Paper < 94      to the left,  improve=0.7715781, (0 missing)
##
## Node number 74: 24 observations,    complexity param=0.01441678
##   predicted class=1  expected loss=0.5416667  P(node) =0.06153846
##     class counts:    11     2    11
```

```
##    probabilities: 0.458 0.083 0.458
##   left son=148 (9 obs) right son=149 (15 obs)
##   Primary splits:
##       Fresh            < 3372.5  to the left,  improve=4.505556, (0 missing)
##       Delicassen       < 940     to the right, improve=3.022727, (0 missing)
##       Milk             < 6361.5  to the right, improve=2.372222, (0 missing)
##       Grocery          < 5085.5  to the right, improve=1.721429, (0 missing)
##       Detergents_Paper < 2522    to the right, improve=1.665966, (0 missing)
##   Surrogate splits:
##       Milk             < 4190    to the right, agree=0.833, adj=0.556, (0 split)
##       Detergents_Paper < 3836.5  to the right, agree=0.833, adj=0.556, (0 split)
##       Grocery          < 10092.5 to the right, agree=0.750, adj=0.333, (0 split)
##       Delicassen       < 2179    to the right, agree=0.708, adj=0.222, (0 split)
##       Frozen           < 1139.5  to the left,  agree=0.667, adj=0.111, (0 split)
##
## Node number 75: 58 observations
##   predicted class=3  expected loss=0.2931034  P(node) =0.1487179
##     class counts:     9     8    41
##    probabilities: 0.155 0.138 0.707
##
## Node number 148: 9 observations
##   predicted class=1  expected loss=0.1111111  P(node) =0.02307692
##     class counts:     8     0     1
##    probabilities: 0.889 0.000 0.111
##
## Node number 149: 15 observations
##   predicted class=3  expected loss=0.3333333  P(node) =0.03846154
##     class counts:     3     2    10
##    probabilities: 0.200 0.133 0.667
```

The diagonals are good, what is predicted is what it actually is.

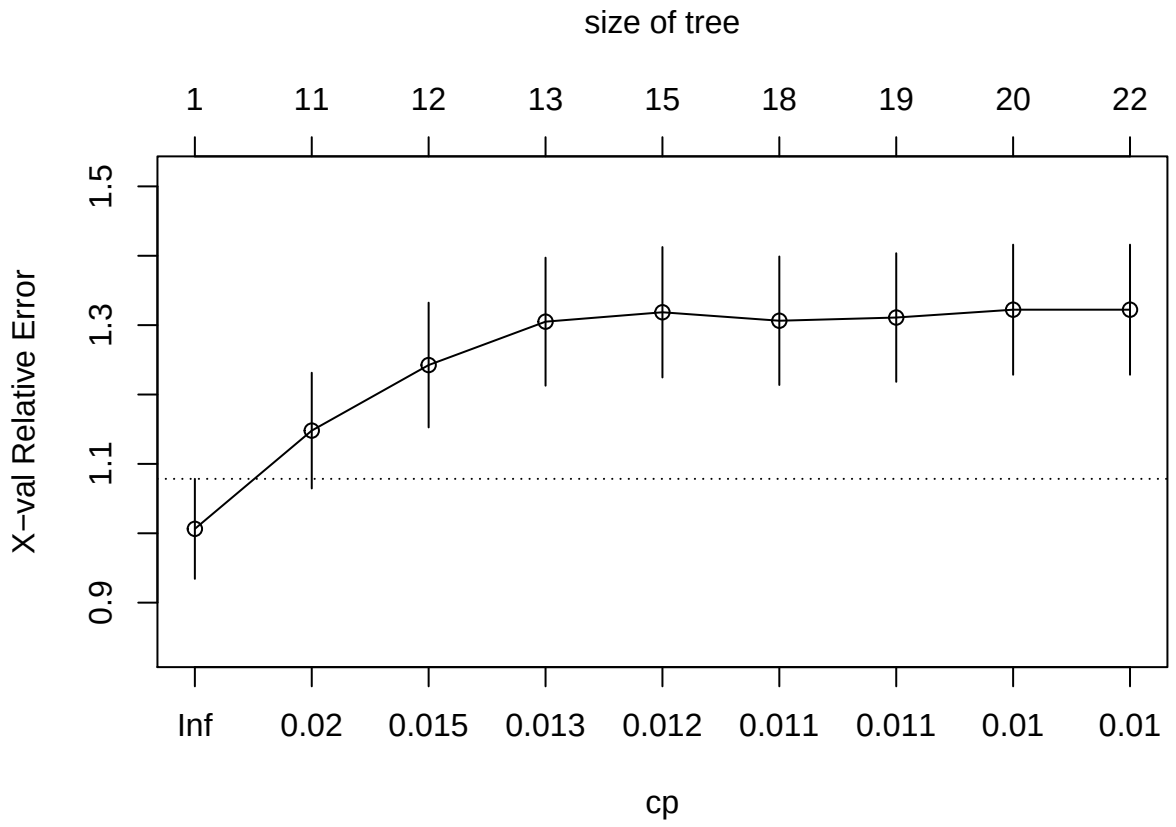
Step 5 - Growing the tree and plotting:

```
###- Regression Tree Example

# grow tree
fit <- rpart(formula, method="anova", data=dfWCD_train)

printcp(fit) # display the results
```

```
##
## Regression tree:
## rpart(formula = formula, data = dfWCD_train, method = "anova")
##
## Variables actually used in tree construction:
## [1] Delicassen       Detergents_Paper Fresh            Frozen
## [5] Grocery          Milk
##
## Root node error: 226.37/390 = 0.58043
##
## n= 390
##
##         CP nsplit rel error xerror     xstd
## 1 0.022626      0   1.00000 1.0064 0.072059
## 2 0.017153     10   0.77225 1.1480 0.083371
```

```
## 3 0.013981     11   0.75509 1.2425 0.089880
## 4 0.012484     12   0.74111 1.3050 0.092253
## 5 0.011619     14   0.71614 1.3185 0.093964
## 6 0.010860     17   0.68129 1.3064 0.092592
## 7 0.010668     18   0.67043 1.3110 0.092635
## 8 0.010101     19   0.65976 1.3223 0.093739
## 9 0.010000     21   0.63956 1.3223 0.093739
```

```
plotcp(fit) # visualize cross-validation results
```




```
summary(fit) # detailed summary
```

```
## Call:
## rpart(formula = formula, data = dfWCD_train, method = "anova")
##   n= 390
##
##           CP nsplit rel error   xerror       xstd
## 1 0.02262605      0 1.0000000 1.006387 0.07205867
## 2 0.01715263     10 0.7722459 1.147993 0.08337051
## 3 0.01398102     11 0.7550932 1.242483 0.08988042
## 4 0.01248384     12 0.7411122 1.305020 0.09225257
## 5 0.01161890     14 0.7161445 1.318550 0.09396435
## 6 0.01086019     17 0.6812879 1.306351 0.09259166
## 7 0.01066800     18 0.6704277 1.310983 0.09263472
## 8 0.01010135     19 0.6597597 1.322257 0.09373946
## 9 0.01000000     21 0.6395570 1.322257 0.09373946
##
```

```
## Variable importance
##           Frozen       Delicassen            Fresh             Milk
##               26               19               18               14
## Detergents_Paper          Grocery
##               13               11
##
## Node number 1: 390 observations,    complexity param=0.02262605
##   mean=2.553846, MSE=0.5804339
##   left son=2 (333 obs) right son=3 (57 obs)
##   Primary splits:
##       Frozen           < 409.5   to the right, improve=0.018901970, (0 missing)
##       Fresh            < 11233   to the left,  improve=0.011066340, (0 missing)
##       Milk             < 2753.5  to the left,  improve=0.009786598, (0 missing)
##       Detergents_Paper < 1004    to the left,  improve=0.007313139, (0 missing)
##       Grocery          < 6916.5  to the left,  improve=0.007303949, (0 missing)
##   Surrogate splits:
##       Fresh < 150.5   to the right, agree=0.862, adj=0.053, (0 split)
##
## Node number 2: 333 observations,    complexity param=0.02262605
##   mean=2.510511, MSE=0.6162559
##   left son=4 (192 obs) right son=5 (141 obs)
##   Primary splits:
##       Fresh      < 11233   to the left,  improve=0.015379280, (0 missing)
##       Delicassen < 3868    to the left,  improve=0.010866140, (0 missing)
##       Milk       < 2753.5  to the left,  improve=0.010098030, (0 missing)
##       Frozen     < 8343.5  to the left,  improve=0.009894533, (0 missing)
##       Grocery    < 7025.5  to the left,  improve=0.007778715, (0 missing)
##   Surrogate splits:
##       Frozen           < 2970.5  to the left,  agree=0.643, adj=0.156, (0 split)
##       Delicassen       < 1782    to the left,  agree=0.628, adj=0.121, (0 split)
##       Milk             < 27899   to the left,  agree=0.598, adj=0.050, (0 split)
##       Grocery          < 511.5   to the right, agree=0.592, adj=0.035, (0 split)
##       Detergents_Paper < 46.5    to the right, agree=0.592, adj=0.035, (0 split)
##
## Node number 3: 57 observations,    complexity param=0.01010135
##   mean=2.807018, MSE=0.2960911
##   left son=6 (23 obs) right son=7 (34 obs)
##   Primary splits:
##       Detergents_Paper < 4131    to the right, improve=0.13357850, (0 missing)
##       Grocery          < 13571.5 to the right, improve=0.11909920, (0 missing)
##       Milk             < 4866.5  to the right, improve=0.05106582, (0 missing)
##       Fresh            < 8577.5  to the left,  improve=0.04492129, (0 missing)
##       Frozen           < 265     to the left,  improve=0.03732230, (0 missing)
##   Surrogate splits:
##       Grocery    < 9278.5  to the right, agree=0.912, adj=0.783, (0 split)
##       Milk       < 6095    to the right, agree=0.772, adj=0.435, (0 split)
##       Frozen     < 63      to the left,  agree=0.684, adj=0.217, (0 split)
##       Delicassen < 422.5   to the right, agree=0.667, adj=0.174, (0 split)
##       Fresh      < 412.5   to the left,  agree=0.632, adj=0.087, (0 split)
##
## Node number 4: 192 observations,    complexity param=0.02262605
##   mean=2.427083, MSE=0.6717665
##   left son=8 (17 obs) right son=9 (175 obs)
##   Primary splits:
```

```
##       Frozen           < 6328    to the right, improve=0.03414276, (0 missing)
##       Detergents_Paper < 94      to the left,  improve=0.02301044, (0 missing)
##       Delicassen       < 3868    to the left,  improve=0.01383783, (0 missing)
##       Grocery          < 25365   to the left,  improve=0.01298519, (0 missing)
##       Milk             < 1143.5  to the left,  improve=0.01178289, (0 missing)
##
## Node number 5: 141 observations,    complexity param=0.02262605
##   mean=2.624113, MSE=0.5182838
##   left son=10 (94 obs) right son=11 (47 obs)
##   Primary splits:
##       Detergents_Paper < 1004    to the left,  improve=0.05944293, (0 missing)
##       Frozen           < 5293    to the left,  improve=0.03898986, (0 missing)
##       Milk             < 2737.5  to the left,  improve=0.02615489, (0 missing)
##       Grocery          < 4164.5  to the left,  improve=0.02606749, (0 missing)
##       Delicassen       < 565     to the right, improve=0.01861369, (0 missing)
##   Surrogate splits:
##       Grocery    < 5652.5  to the left,  agree=0.823, adj=0.468, (0 split)
##       Milk       < 5371    to the left,  agree=0.794, adj=0.383, (0 split)
##       Delicassen < 2728.5  to the left,  agree=0.723, adj=0.170, (0 split)
##       Frozen     < 1350    to the right, agree=0.709, adj=0.128, (0 split)
##       Fresh      < 42937   to the left,  agree=0.674, adj=0.021, (0 split)
##
## Node number 6: 23 observations,    complexity param=0.01010135
##   mean=2.565217, MSE=0.5935728
##   left son=12 (15 obs) right son=13 (8 obs)
##   Primary splits:
##       Fresh            < 1046.5  to the right, improve=0.16985140, (0 missing)
##       Delicassen       < 1273.5  to the left,  improve=0.07143557, (0 missing)
##       Frozen           < 231.5   to the left,  improve=0.06281278, (0 missing)
##       Detergents_Paper < 7947.5  to the right, improve=0.05758076, (0 missing)
##       Milk             < 7603.5  to the left,  improve=0.03537482, (0 missing)
##   Surrogate splits:
##       Milk       < 4502.5  to the right, agree=0.739, adj=0.250, (0 split)
##       Frozen     < 376     to the left,  agree=0.739, adj=0.250, (0 split)
##       Delicassen < 197.5   to the right, agree=0.739, adj=0.250, (0 split)
##       Grocery    < 17665.5 to the left,  agree=0.696, adj=0.125, (0 split)
##
## Node number 7: 34 observations
##   mean=2.970588, MSE=0.02854671
##
## Node number 8: 17 observations
##   mean=1.941176, MSE=0.7612457
##
## Node number 9: 175 observations,    complexity param=0.02262605
##   mean=2.474286, MSE=0.6379102
##   left son=18 (89 obs) right son=19 (86 obs)
##   Primary splits:
##       Frozen           < 1455.5  to the left,  improve=0.02574387, (0 missing)
##       Milk             < 1037    to the left,  improve=0.02174994, (0 missing)
##       Detergents_Paper < 94      to the left,  improve=0.01986291, (0 missing)
##       Fresh            < 2822    to the left,  improve=0.01848245, (0 missing)
##       Delicassen       < 61      to the left,  improve=0.01469296, (0 missing)
##   Surrogate splits:
##       Fresh            < 2739.5  to the left,  agree=0.640, adj=0.267, (0 split)
```

```
##       Delicassen       < 675     to the left,  agree=0.571, adj=0.128, (0 split)
##       Milk             < 8556    to the left,  agree=0.566, adj=0.116, (0 split)
##       Detergents_Paper < 5964.5  to the left,  agree=0.560, adj=0.105, (0 split)
##       Grocery          < 2016    to the right, agree=0.554, adj=0.093, (0 split)
##
## Node number 10: 94 observations,    complexity param=0.02262605
##   mean=2.5, MSE=0.6542553
##   left son=20 (64 obs) right son=21 (30 obs)
##   Primary splits:
##       Frozen           < 5293    to the left,  improve=0.06448171, (0 missing)
##       Delicassen       < 565     to the right, improve=0.04728176, (0 missing)
##       Detergents_Paper < 241.5   to the right, improve=0.04596858, (0 missing)
##       Grocery          < 3621.5  to the right, improve=0.02367423, (0 missing)
##       Milk             < 4872.5  to the right, improve=0.01778117, (0 missing)
##   Surrogate splits:
##       Milk       < 4386.5  to the left,  agree=0.745, adj=0.200, (0 split)
##       Fresh      < 11585   to the right, agree=0.723, adj=0.133, (0 split)
##       Delicassen < 2657    to the left,  agree=0.713, adj=0.100, (0 split)
##       Grocery    < 8496.5  to the left,  agree=0.702, adj=0.067, (0 split)
##
## Node number 11: 47 observations
##   mean=2.87234, MSE=0.1539158
##
## Node number 12: 15 observations
##   mean=2.333333, MSE=0.7555556
##
## Node number 13: 8 observations
##   mean=3, MSE=0
##
## Node number 18: 89 observations,    complexity param=0.02262605
##   mean=2.348315, MSE=0.7213736
##   left son=36 (24 obs) right son=37 (65 obs)
##   Primary splits:
##       Frozen           < 1107.5  to the right, improve=0.07784400, (0 missing)
##       Milk             < 4981.5  to the right, improve=0.03745223, (0 missing)
##       Delicassen       < 376     to the left,  improve=0.02044836, (0 missing)
##       Detergents_Paper < 2116    to the right, improve=0.02012639, (0 missing)
##       Grocery          < 5062    to the right, improve=0.01736365, (0 missing)
##   Surrogate splits:
##       Milk       < 18606.5 to the right, agree=0.742, adj=0.042, (0 split)
##       Grocery    < 25365   to the right, agree=0.742, adj=0.042, (0 split)
##       Delicassen < 4842    to the right, agree=0.742, adj=0.042, (0 split)
##
## Node number 19: 86 observations,    complexity param=0.0116189
##   mean=2.604651, MSE=0.5181179
##   left son=38 (25 obs) right son=39 (61 obs)
##   Primary splits:
##       Frozen           < 3484    to the right, improve=0.04734522, (0 missing)
##       Delicassen       < 205     to the left,  improve=0.04554260, (0 missing)
##       Fresh            < 5147    to the left,  improve=0.03757829, (0 missing)
##       Milk             < 2225.5  to the left,  improve=0.03034658, (0 missing)
##       Detergents_Paper < 263.5   to the left,  improve=0.02449548, (0 missing)
##   Surrogate splits:
##       Fresh < 9635    to the right, agree=0.721, adj=0.04, (0 split)
```

```
##
## Node number 20: 64 observations,    complexity param=0.02262605
##   mean=2.359375, MSE=0.7614746
##   left son=40 (45 obs) right son=41 (19 obs)
##   Primary splits:
##       Delicassen       < 565     to the right, improve=0.10257040, (0 missing)
##       Detergents_Paper < 238     to the right, improve=0.06999272, (0 missing)
##       Milk             < 2093    to the right, improve=0.06778868, (0 missing)
##       Grocery          < 3621.5  to the right, improve=0.05713370, (0 missing)
##       Frozen           < 730     to the left,  improve=0.04067961, (0 missing)
##   Surrogate splits:
##       Milk             < 882     to the right, agree=0.766, adj=0.211, (0 split)
##       Frozen           < 617     to the right, agree=0.719, adj=0.053, (0 split)
##       Detergents_Paper < 43.5    to the right, agree=0.719, adj=0.053, (0 split)
##
## Node number 21: 30 observations,    complexity param=0.010668
##   mean=2.8, MSE=0.2933333
##   left son=42 (7 obs) right son=43 (23 obs)
##   Primary splits:
##       Delicassen       < 2552    to the right, improve=0.27442120, (0 missing)
##       Fresh            < 30176.5 to the right, improve=0.14313950, (0 missing)
##       Detergents_Paper < 567     to the right, improve=0.14313950, (0 missing)
##       Milk             < 4487    to the right, improve=0.05284906, (0 missing)
##       Grocery          < 2093    to the right, improve=0.04958678, (0 missing)
##   Surrogate splits:
##       Frozen < 17624.5 to the right, agree=0.867, adj=0.429, (0 split)
##       Milk   < 980     to the left,  agree=0.833, adj=0.286, (0 split)
##       Fresh  < 30176.5 to the right, agree=0.800, adj=0.143, (0 split)
##
## Node number 36: 24 observations,    complexity param=0.02262605
##   mean=1.958333, MSE=0.8732639
##   left son=72 (8 obs) right son=73 (16 obs)
##   Primary splits:
##       Fresh            < 3372.5  to the left,  improve=0.5258449, (0 missing)
##       Milk             < 6361.5  to the right, improve=0.1651768, (0 missing)
##       Delicassen       < 1158.5  to the right, improve=0.1610338, (0 missing)
##       Detergents_Paper < 2522    to the right, improve=0.1114645, (0 missing)
##       Grocery          < 5085.5  to the right, improve=0.1004463, (0 missing)
##   Surrogate splits:
##       Milk             < 4190    to the right, agree=0.708, adj=0.125, (0 split)
##       Frozen           < 1127.5  to the left,  agree=0.708, adj=0.125, (0 split)
##       Detergents_Paper < 3836.5  to the right, agree=0.708, adj=0.125, (0 split)
##       Delicassen       < 2179    to the right, agree=0.708, adj=0.125, (0 split)
##
## Node number 37: 65 observations,    complexity param=0.01398102
##   mean=2.492308, MSE=0.5884024
##   left son=74 (7 obs) right son=75 (58 obs)
##   Primary splits:
##       Milk             < 804.5   to the left,  improve=0.08275010, (0 missing)
##       Delicassen       < 1072    to the left,  improve=0.06837487, (0 missing)
##       Fresh            < 6292    to the right, improve=0.06570126, (0 missing)
##       Frozen           < 530     to the left,  improve=0.05155863, (0 missing)
##       Detergents_Paper < 256.5   to the right, improve=0.02745571, (0 missing)
##   Surrogate splits:
```

```
##       Grocery          < 1056    to the left,  agree=0.954, adj=0.571, (0 split)
##       Detergents_Paper < 79      to the left,  agree=0.954, adj=0.571, (0 split)
##
## Node number 38: 25 observations,    complexity param=0.01086019
##   mean=2.36, MSE=0.7104
##   left son=76 (7 obs) right son=77 (18 obs)
##   Primary splits:
##       Detergents_Paper < 1341    to the right, improve=0.13842410, (0 missing)
##       Grocery          < 4456    to the right, improve=0.12162160, (0 missing)
##       Frozen           < 5501    to the left,  improve=0.10075520, (0 missing)
##       Fresh            < 4465    to the right, improve=0.06481000, (0 missing)
##       Milk             < 1960.5  to the left,  improve=0.04904905, (0 missing)
##   Surrogate splits:
##       Grocery    < 6147.5  to the right, agree=0.88, adj=0.571, (0 split)
##       Milk       < 5145.5  to the right, agree=0.84, adj=0.429, (0 split)
##       Delicassen < 1393    to the right, agree=0.76, adj=0.143, (0 split)
##
## Node number 39: 61 observations,    complexity param=0.0116189
##   mean=2.704918, MSE=0.4047299
##   left son=78 (24 obs) right son=79 (37 obs)
##   Primary splits:
##       Fresh            < 4335    to the left,  improve=0.09744879, (0 missing)
##       Delicassen       < 205     to the left,  improve=0.05628456, (0 missing)
##       Detergents_Paper < 263.5   to the left,  improve=0.04573397, (0 missing)
##       Milk             < 2225.5  to the left,  improve=0.03565437, (0 missing)
##       Grocery          < 1326.5  to the right, improve=0.02788845, (0 missing)
##   Surrogate splits:
##       Frozen           < 1597    to the left,  agree=0.672, adj=0.167, (0 split)
##       Milk             < 549     to the left,  agree=0.639, adj=0.083, (0 split)
##       Grocery          < 894.5   to the left,  agree=0.639, adj=0.083, (0 split)
##       Detergents_Paper < 282.5   to the left,  agree=0.639, adj=0.083, (0 split)
##       Delicassen       < 124     to the left,  agree=0.639, adj=0.083, (0 split)
##
## Node number 40: 45 observations,    complexity param=0.02262605
##   mean=2.177778, MSE=0.857284
##   left son=80 (12 obs) right son=81 (33 obs)
##   Primary splits:
##       Delicassen       < 847     to the left,  improve=0.19485760, (0 missing)
##       Detergents_Paper < 238     to the right, improve=0.07463875, (0 missing)
##       Grocery          < 2023    to the right, improve=0.06915323, (0 missing)
##       Frozen           < 1037    to the left,  improve=0.04665899, (0 missing)
##       Milk             < 2089    to the right, improve=0.04233871, (0 missing)
##   Surrogate splits:
##       Frozen < 730     to the left,  agree=0.778, adj=0.167, (0 split)
##       Milk   < 514.5   to the left,  agree=0.756, adj=0.083, (0 split)
##
## Node number 41: 19 observations
##   mean=2.789474, MSE=0.2714681
##
## Node number 42: 7 observations
##   mean=2.285714, MSE=0.7755102
##
## Node number 43: 23 observations
##   mean=2.956522, MSE=0.0415879
```

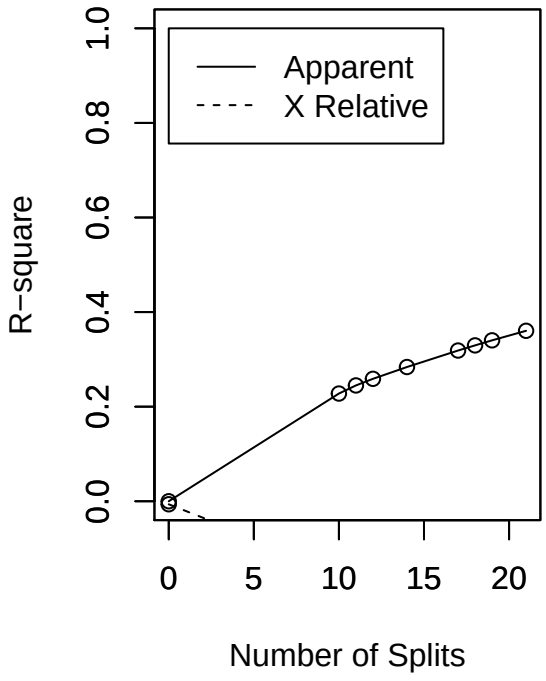
```
##
## Node number 72: 8 observations
##   mean=1, MSE=0
##
## Node number 73: 16 observations
##   mean=2.4375, MSE=0.6210938
##
## Node number 74: 7 observations
##   mean=1.857143, MSE=0.9795918
##
## Node number 75: 58 observations,    complexity param=0.01248384
##   mean=2.568966, MSE=0.4866231
##   left son=150 (7 obs) right son=151 (51 obs)
##   Primary splits:
##       Milk             < 11099   to the right, improve=0.09130763, (0 missing)
##       Frozen           < 530     to the left,  improve=0.06550932, (0 missing)
##       Delicassen       < 1072    to the left,  improve=0.05214022, (0 missing)
##       Fresh            < 9423.5  to the right, improve=0.05121242, (0 missing)
##       Detergents_Paper < 7840    to the right, improve=0.05121242, (0 missing)
##   Surrogate splits:
##       Delicassen < 2959    to the right, agree=0.897, adj=0.143, (0 split)
##
## Node number 76: 7 observations
##   mean=1.857143, MSE=0.9795918
##
## Node number 77: 18 observations
##   mean=2.555556, MSE=0.4691358
##
## Node number 78: 24 observations,    complexity param=0.0116189
##   mean=2.458333, MSE=0.7482639
##   left son=156 (12 obs) right son=157 (12 obs)
##   Primary splits:
##       Delicassen       < 886.5   to the left,  improve=0.18793500, (0 missing)
##       Detergents_Paper < 5958    to the left,  improve=0.08752364, (0 missing)
##       Frozen           < 1699    to the right, improve=0.08752364, (0 missing)
##       Fresh            < 3147    to the right, improve=0.05476808, (0 missing)
##       Grocery          < 14554   to the left,  improve=0.03605062, (0 missing)
##   Surrogate splits:
##       Grocery          < 1898    to the left,  agree=0.750, adj=0.500, (0 split)
##       Fresh            < 2890    to the right, agree=0.708, adj=0.417, (0 split)
##       Milk             < 2711    to the left,  agree=0.667, adj=0.333, (0 split)
##       Frozen           < 1500    to the left,  agree=0.625, adj=0.250, (0 split)
##       Detergents_Paper < 308.5   to the left,  agree=0.625, adj=0.250, (0 split)
##
## Node number 79: 37 observations
##   mean=2.864865, MSE=0.1168736
##
## Node number 80: 12 observations
##   mean=1.5, MSE=0.75
##
## Node number 81: 33 observations,    complexity param=0.01715263
##   mean=2.424242, MSE=0.6685032
##   left son=162 (18 obs) right son=163 (15 obs)
##   Primary splits:
```

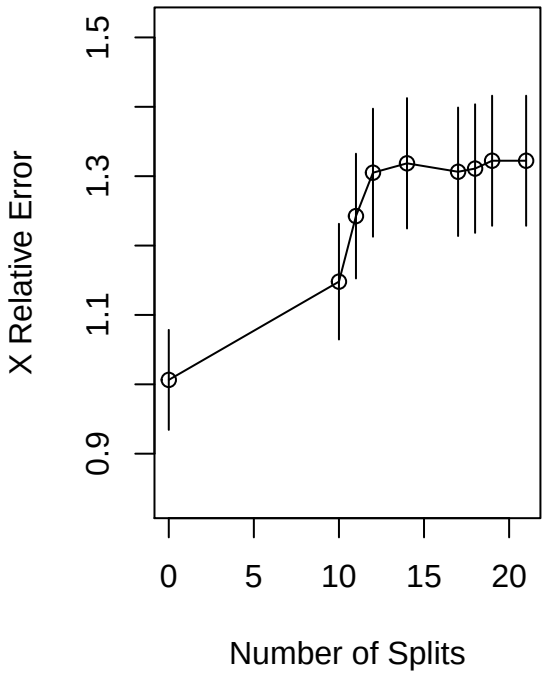
```
##       Grocery          < 2023    to the right, improve=0.17600730, (0 missing)
##       Milk             < 1211    to the right, improve=0.13350590, (0 missing)
##       Detergents_Paper < 237.5   to the right, improve=0.11909340, (0 missing)
##       Delicassen       < 1836.5  to the right, improve=0.08727175, (0 missing)
##       Frozen           < 3709    to the right, improve=0.08615385, (0 missing)
##   Surrogate splits:
##       Detergents_Paper < 264     to the right, agree=0.848, adj=0.667, (0 split)
##       Milk             < 1964.5  to the right, agree=0.758, adj=0.467, (0 split)
##       Fresh            < 15020.5 to the right, agree=0.697, adj=0.333, (0 split)
##       Delicassen       < 2234    to the right, agree=0.667, adj=0.267, (0 split)
##       Frozen           < 2277.5  to the right, agree=0.636, adj=0.200, (0 split)
##
## Node number 150: 7 observations
##   mean=2, MSE=0.5714286
##
## Node number 151: 51 observations,    complexity param=0.01248384
##   mean=2.647059, MSE=0.4244521
##   left son=302 (15 obs) right son=303 (36 obs)
##   Primary splits:
##       Frozen           < 530     to the left,  improve=0.14204410, (0 missing)
##       Fresh            < 9423.5  to the right, improve=0.09528515, (0 missing)
##       Delicassen       < 376     to the left,  improve=0.09078557, (0 missing)
##       Detergents_Paper < 442.5   to the left,  improve=0.05902437, (0 missing)
##       Grocery          < 10218   to the left,  improve=0.03934447, (0 missing)
##   Surrogate splits:
##       Milk             < 1246    to the left,  agree=0.765, adj=0.200, (0 split)
##       Detergents_Paper < 94      to the left,  agree=0.745, adj=0.133, (0 split)
##
## Node number 156: 12 observations
##   mean=2.083333, MSE=0.9097222
##
## Node number 157: 12 observations
##   mean=2.833333, MSE=0.3055556
##
## Node number 162: 18 observations
##   mean=2.111111, MSE=0.7654321
##
## Node number 163: 15 observations
##   mean=2.8, MSE=0.2933333
##
## Node number 302: 15 observations
##   mean=2.266667, MSE=0.5955556
##
## Node number 303: 36 observations
##   mean=2.805556, MSE=0.2677469
```

```
# create additional plots
par(mfrow=c(1,2)) # two plots on one page
rsq.rpart(fit) # visualize cross-validation results
```

```
##
## Regression tree:
## rpart(formula = formula, data = dfWCD_train, method = "anova")
##
## Variables actually used in tree construction:
```

```
## [1] Delicassen       Detergents_Paper Fresh            Frozen
## [5] Grocery          Milk
##
## Root node error: 226.37/390 = 0.58043
##
## n= 390
##
##         CP nsplit rel error xerror     xstd
## 1 0.022626      0   1.00000 1.0064 0.072059
## 2 0.017153     10   0.77225 1.1480 0.083371
## 3 0.013981     11   0.75509 1.2425 0.089880
## 4 0.012484     12   0.74111 1.3050 0.092253
## 5 0.011619     14   0.71614 1.3185 0.093964
## 6 0.010860     17   0.68129 1.3064 0.092592
## 7 0.010668     18   0.67043 1.3110 0.092635
## 8 0.010101     19   0.65976 1.3223 0.093739
## 9 0.010000     21   0.63956 1.3223 0.093739
```






```
# plot tree
plot(fit, uniform=TRUE,
     main="Regression Tree for 'type' ")
text(fit, use.n=TRUE, all=TRUE, cex=.8)


### ----------------  plot tree

plot(fit, uniform=T, main="Classification Tree for Customer Channels")
```

```
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

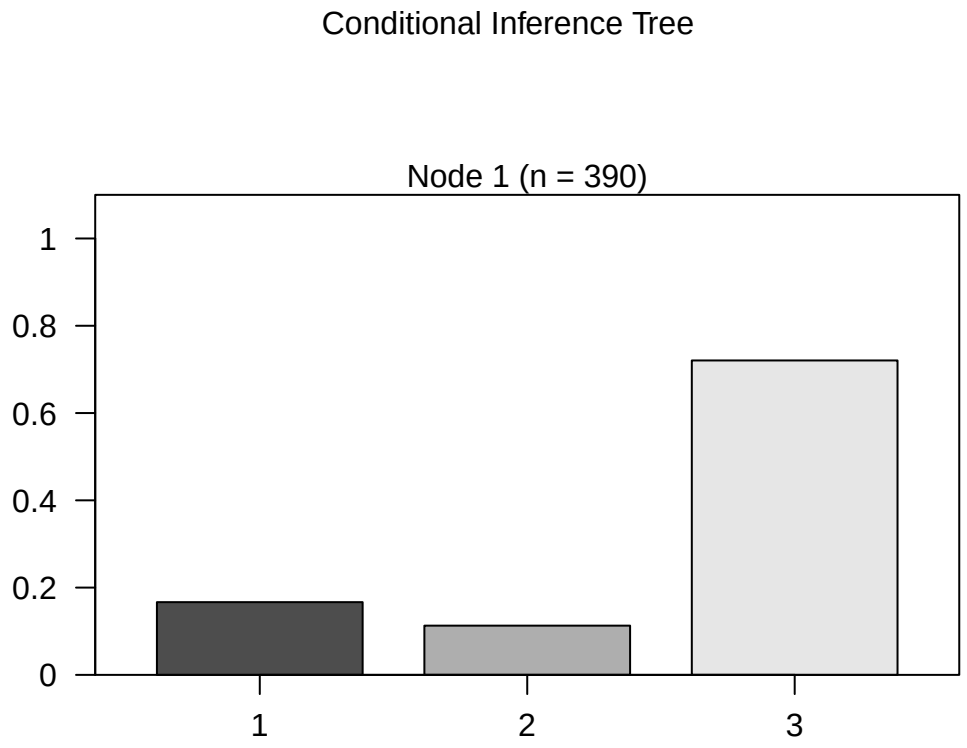

```
##------------------ TREE package

tr = tree(formula, data=dfWCD_train)
summary(tr)
```

```
##
## Classification tree:
## tree(formula = formula, data = dfWCD_train)
## Number of terminal nodes:  23
## Residual mean deviance:  1.075 = 394.7 / 367
## Misclassification error rate: 0.2154 = 84 / 390
```

```
plot(tr); text(tr)


##-------------------Party package


ct = ctree(formula, data = dfWCD_train)
plot(ct, main="Conditional Inference Tree")


# Estimated class probabilities
tr.pred = predict(ct, newdata=dfWCD_train, type="prob")
```

```
#Table of prediction errors
table(predict(ct), dfWCD_train$Channel)
```

```
##
##        1   2
##   1    0   0
##   2    0   0
##   3  262 128
```



1.3 Questions

1. Does the size of the data set make a difference?
   - Yes, with more training data the predictive data will have better results with these algorithms. For the C4.5 (and C5.0) algorithms, the equation is:

$$p = e + z \times \sqrt{e \times \frac{1-e}{n}}$$

   where: p = true error rate, e = the observed error rate, z = level of confidence, and n is the number of trials. If n is low, p will not be very close to e; as n gets higher p will become closer to e. So in this case, the more data the better the results.
2. Do the rules make sense? If so why did the algorithm generate good rules? If not, why not?
   - The rules look like they make sense, nothing seems odd. The results of more points in region 3 makes sense since a majority of the trained data was there. The contents look like they fit too, I think this algorithm generated good rules for this data set.
3. Does scaling, normalization or leaving the data unscaled make a difference?

- For this dataset it does not matter because all of the data used were at a similar scale. However, looking at the equations, I don't think scaling the data is possible, a majority of these equations take the log of a value, which would be invalid for negative numbers and the log function would essentially scale the data anyway. I tried to normalize this data and run it to see what would happen, but I ended up getting an error so I do not think it is possible with this dataset.