## Module 4 Example Classification Trees

options(tidyverse.quiet = TRUE)  
library(titanic)

## Warning: package 'titanic' was built under R version 3.5.2

library(tidyverse)  
library(mice)

## Warning: package 'mice' was built under R version 3.5.2

## Loading required package: lattice

##   
## Attaching package: 'mice'

## The following object is masked from 'package:tidyr':  
##   
## complete

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(VIM)

## Warning: package 'VIM' was built under R version 3.5.2

## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

## Warning: package 'data.table' was built under R version 3.5.2

##   
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':  
##   
## between, first, last

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

## VIM is ready to use.   
## Since version 4.0.0 the GUI is in its own package VIMGUI.  
##   
## Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

##   
## Attaching package: 'VIM'

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

library(rpart)

## Warning: package 'rpart' was built under R version 3.5.2

library(rattle)

## Warning: package 'rattle' was built under R version 3.5.2

## Rattle: A free graphical interface for data science with R.  
## Version 5.2.0 Copyright (c) 2006-2018 Togaware Pty Ltd.  
## Type 'rattle()' to shake, rattle, and roll your data.

library(RColorBrewer)

## Warning: package 'RColorBrewer' was built under R version 3.5.2

titanic = titanic::titanic\_train

str(titanic)

## 'data.frame': 891 obs. of 12 variables:  
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : int 0 1 1 1 0 0 0 0 1 1 ...  
## $ Pclass : int 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...  
## $ Sex : chr "male" "female" "female" "female" ...  
## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : chr "" "C85" "" "C123" ...  
## $ Embarked : chr "S" "C" "S" "S" ...

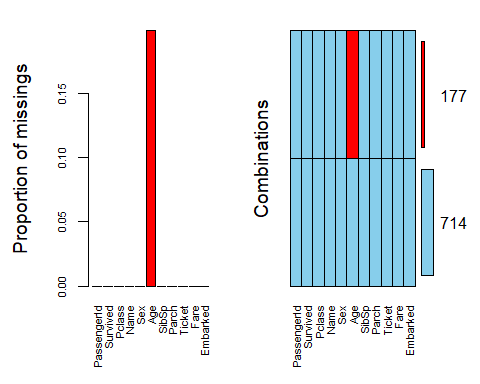
summary(titanic)

## PassengerId Survived Pclass Name   
## Min. : 1.0 Min. :0.0000 Min. :1.000 Length:891   
## 1st Qu.:223.5 1st Qu.:0.0000 1st Qu.:2.000 Class :character   
## Median :446.0 Median :0.0000 Median :3.000 Mode :character   
## Mean :446.0 Mean :0.3838 Mean :2.309   
## 3rd Qu.:668.5 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :891.0 Max. :1.0000 Max. :3.000   
##   
## Sex Age SibSp Parch   
## Length:891 Min. : 0.42 Min. :0.000 Min. :0.0000   
## Class :character 1st Qu.:20.12 1st Qu.:0.000 1st Qu.:0.0000   
## Mode :character Median :28.00 Median :0.000 Median :0.0000   
## Mean :29.70 Mean :0.523 Mean :0.3816   
## 3rd Qu.:38.00 3rd Qu.:1.000 3rd Qu.:0.0000   
## Max. :80.00 Max. :8.000 Max. :6.0000   
## NA's :177   
## Ticket Fare Cabin Embarked   
## Length:891 Min. : 0.00 Length:891 Length:891   
## Class :character 1st Qu.: 7.91 Class :character Class :character   
## Mode :character Median : 14.45 Mode :character Mode :character   
## Mean : 32.20   
## 3rd Qu.: 31.00   
## Max. :512.33   
##

titanic = titanic %>% mutate(Survived = as.factor(Survived)) %>%  
 mutate(Survived = fct\_recode(Survived, "NO" = "0", "Yes" = "1")) %>%  
 mutate(Pclass = as.factor(Pclass)) %>% mutate(Sex = as.factor(Sex)) %>%  
 mutate(Embarked = as.factor(Embarked)) %>%  
 mutate(Embarked = fct\_recode(Embarked, "Unknown" = "", "Cherbourg" = "C", "Southampton" = "S", "Queenstown" = "Q"))  
  
titanic$Cabin[titanic$Cabin==""] = NA  
  
str(titanic)

## 'data.frame': 891 obs. of 12 variables:  
## $ PassengerId: int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : Factor w/ 2 levels "NO","Yes": 1 2 2 2 1 1 1 1 2 2 ...  
## $ Pclass : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...  
## $ Name : chr "Braund, Mr. Owen Harris" "Cumings, Mrs. John Bradley (Florence Briggs Thayer)" "Heikkinen, Miss. Laina" "Futrelle, Mrs. Jacques Heath (Lily May Peel)" ...  
## $ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...  
## $ Age : num 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : int 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : int 0 0 0 0 0 0 0 1 2 0 ...  
## $ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : chr NA "C85" NA "C123" ...  
## $ Embarked : Factor w/ 4 levels "Unknown","Cherbourg",..: 4 2 4 4 4 3 4 4 4 2 ...

titanic = titanic %>% select(-Cabin)  
vim\_plot = aggr(titanic, numbers = TRUE, prop = c(TRUE, FALSE), cex.axis=.7)



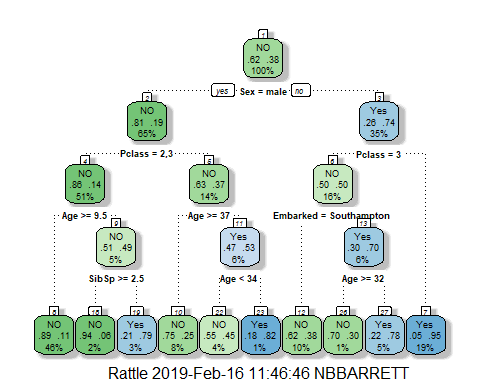
titanic = titanic %>% select(c("Survived", "Pclass", "Sex", "Age", "SibSp", "Parch", "Embarked"))  
  
imp\_age = mice(titanic, m=1, method='pmm', printFlag = FALSE)  
summary(imp\_age)

## Class: mids  
## Number of multiple imputations: 1   
## Imputation methods:  
## Survived Pclass Sex Age SibSp Parch Embarked   
## "" "" "" "pmm" "" "" ""   
## PredictorMatrix:  
## Survived Pclass Sex Age SibSp Parch Embarked  
## Survived 0 1 1 1 1 1 1  
## Pclass 1 0 1 1 1 1 1  
## Sex 1 1 0 1 1 1 1  
## Age 1 1 1 0 1 1 1  
## SibSp 1 1 1 1 0 1 1  
## Parch 1 1 1 1 1 0 1

titanic\_complete = complete(imp\_age)  
summary(titanic\_complete)

## Survived Pclass Sex Age SibSp   
## NO :549 1:216 female:314 Min. : 0.42 Min. :0.000   
## Yes:342 2:184 male :577 1st Qu.:20.00 1st Qu.:0.000   
## 3:491 Median :28.00 Median :0.000   
## Mean :29.47 Mean :0.523   
## 3rd Qu.:38.00 3rd Qu.:1.000   
## Max. :80.00 Max. :8.000   
## Parch Embarked   
## Min. :0.0000 Unknown : 2   
## 1st Qu.:0.0000 Cherbourg :168   
## Median :0.0000 Queenstown : 77   
## Mean :0.3816 Southampton:644   
## 3rd Qu.:0.0000   
## Max. :6.0000

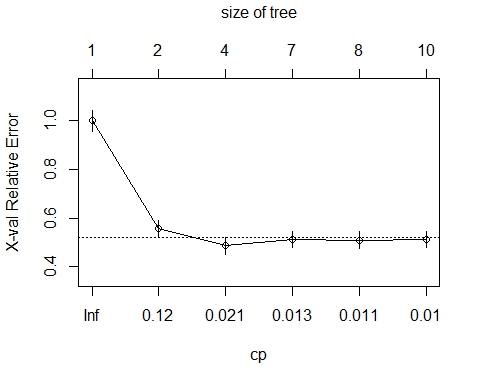
tree1 = rpart(Survived ~., titanic\_complete, method = "class")  
fancyRpartPlot(tree1)



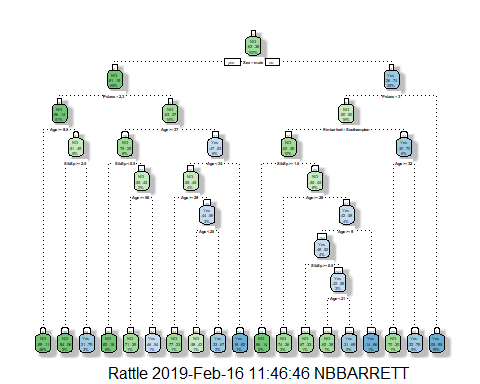
printcp(tree1)

##   
## Classification tree:  
## rpart(formula = Survived ~ ., data = titanic\_complete, method = "class")  
##   
## Variables actually used in tree construction:  
## [1] Age Embarked Pclass Sex SibSp   
##   
## Root node error: 342/891 = 0.38384  
##   
## n= 891   
##   
## CP nsplit rel error xerror xstd  
## 1 0.444444 0 1.00000 1.00000 0.042446  
## 2 0.032164 1 0.55556 0.55556 0.035750  
## 3 0.013645 3 0.49123 0.48538 0.033983  
## 4 0.011696 6 0.45029 0.51170 0.034675  
## 5 0.010234 7 0.43860 0.50877 0.034599  
## 6 0.010000 9 0.41813 0.51170 0.034675

plotcp(tree1)



tree2 = rpart(Survived ~., titanic\_complete, cp = 0.0001, method = "class")  
fancyRpartPlot(tree2)



printcp(tree2)

##   
## Classification tree:  
## rpart(formula = Survived ~ ., data = titanic\_complete, method = "class",   
## cp = 1e-04)  
##   
## Variables actually used in tree construction:  
## [1] Age Embarked Pclass Sex SibSp   
##   
## Root node error: 342/891 = 0.38384  
##   
## n= 891   
##   
## CP nsplit rel error xerror xstd  
## 1 0.4444444 0 1.00000 1.00000 0.042446  
## 2 0.0321637 1 0.55556 0.55556 0.035750  
## 3 0.0136452 3 0.49123 0.47953 0.033824  
## 4 0.0116959 6 0.45029 0.47953 0.033824  
## 5 0.0102339 7 0.43860 0.47953 0.033824  
## 6 0.0087719 9 0.41813 0.46491 0.033419  
## 7 0.0058480 12 0.39181 0.46784 0.033501  
## 8 0.0043860 13 0.38596 0.46199 0.033336  
## 9 0.0014620 16 0.37135 0.44737 0.032916  
## 10 0.0001000 18 0.36842 0.44737 0.032916

plotcp(tree2)

