Flights - Decision Trees

Code

Hide

```
library(hflights)
library(rpart)
library(rattle)
library(rpart.plot)
library(ggplot2)
```

1. Cleaning data

Hide

```
hflights <- hflights %>% na.omit()
```

2. Exploring data

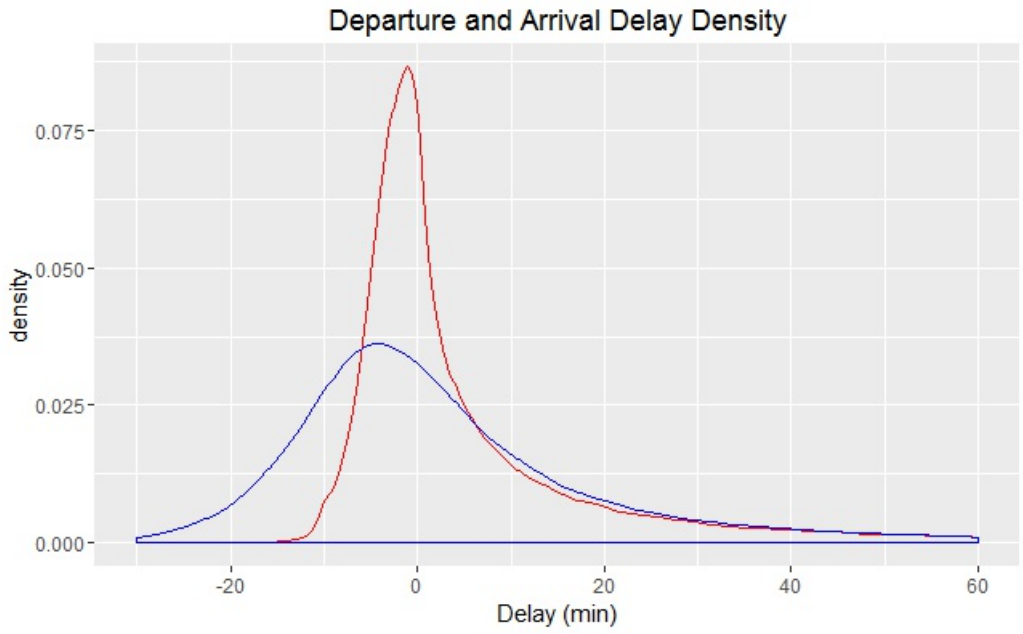
Hide

```
str(hflights)
```

```
'data.frame':   223874 obs. of  23 variables:
 $ Year             : int  2011 2011 2011 2011 2011 2011 2011 2011 2011 2011 ...
 $ Month            : int  1 1 1 1 1 1 1 1 1 1 ...
 $ DayofMonth       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ DayOfWeek        : int  6 7 1 2 3 4 5 6 7 1 ...
 $ DepTime          : int  1400 1401 1352 1403 1405 1359 1359 1355 1443 1443 ...
 $ ArrTime          : int  1500 1501 1502 1513 1507 1503 1509 1454 1554 1553 ...
 $ UniqueCarrier    : chr  "AA" "AA" "AA" "AA" ...
 $ FlightNum        : int  428 428 428 428 428 428 428 428 428 428 ...
 $ TailNum          : chr  "N576AA" "N557AA" "N541AA" "N403AA" ...
 $ ActualElapsedTime: int  60 60 70 70 62 64 70 59 71 70 ...
 $ AirTime          : int  40 45 48 39 44 45 43 40 41 45 ...
 $ ArrDelay         : int  -10 -9 -8 3 -3 -7 -1 -16 44 43 ...
 $ DepDelay         : int  0 1 -8 3 5 -1 -1 -5 43 43 ...
 $ Origin           : chr  "IAH" "IAH" "IAH" "IAH" ...
 $ Dest             : chr  "DFW" "DFW" "DFW" "DFW" ...
 $ Distance         : int  224 224 224 224 224 224 224 224 224 224 ...
 $ TaxiIn           : int  7 6 5 9 9 6 12 7 8 6 ...
 $ TaxiOut          : int  13 9 17 22 9 13 15 12 22 19 ...
 $ Cancelled        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ CancellationCode : chr  "" "" "" "" ...
 $ Diverted         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ DepDelay60       : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ ArrDelay60       : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 - attr(*, "na.action")=Class 'omit'  Named int [1:3622] 195 211 253 284 324 336 348 416 425 535 ...
  .. ..- attr(*, "names")= chr [1:3622] "33074" "35264" "39463" "50174" ...
```

Hide

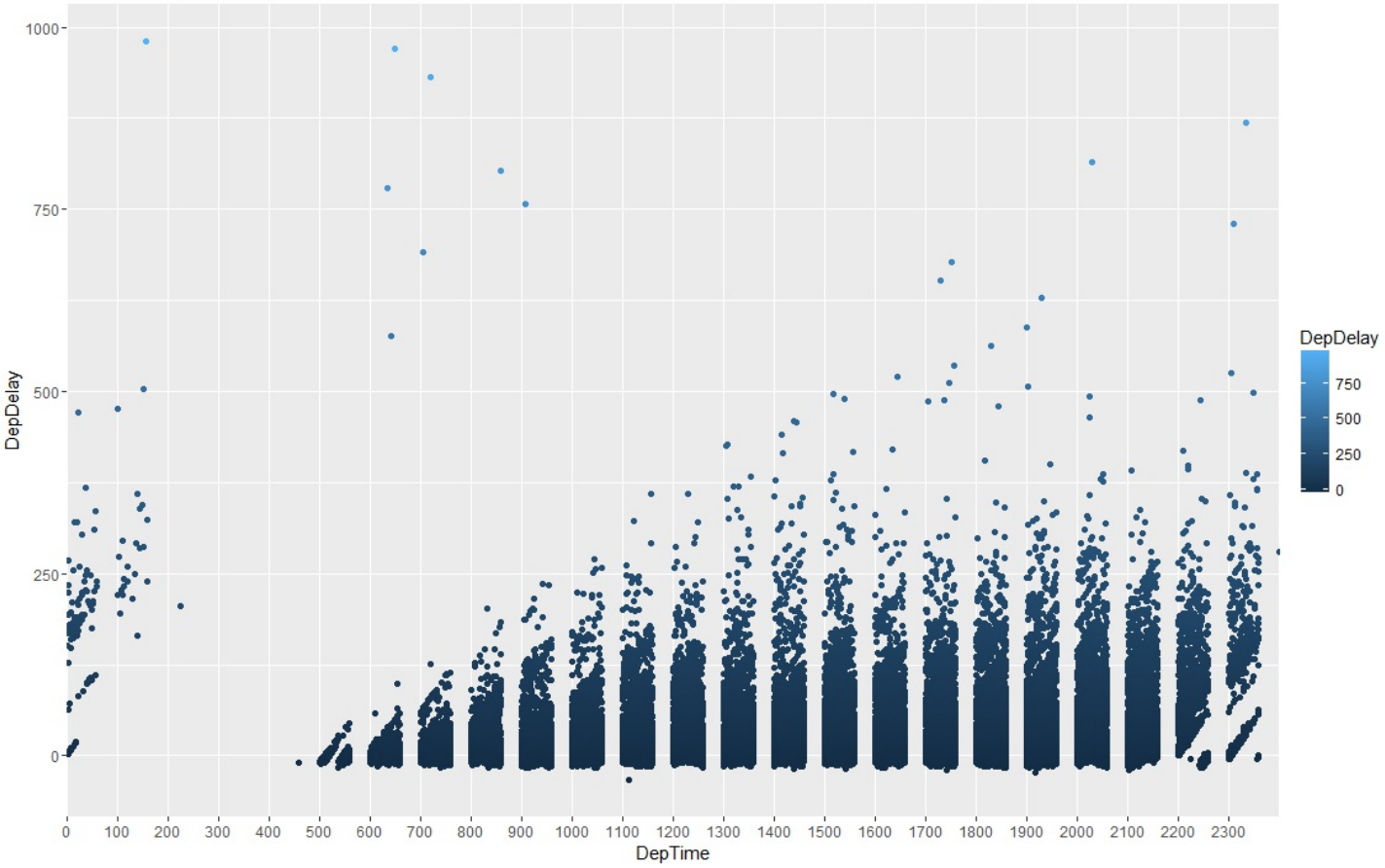
```
ggplot() +
    geom_density(aes(x=DepDelay), colour="red", data=hflights[(hflights$DepDelay<=60 & hflights$DepDelay >= -30
),]) +
    geom_density(aes(x=ArrDelay), colour="blue", data=hflights[(hflights$ArrDelay<=60 & hflights$ArrDelay >= -3
0),]) +
    labs(x="Delay (min)") +
    ggtitle("Departure and Arrival Delay Density")
```



Departure delay plot

Hide

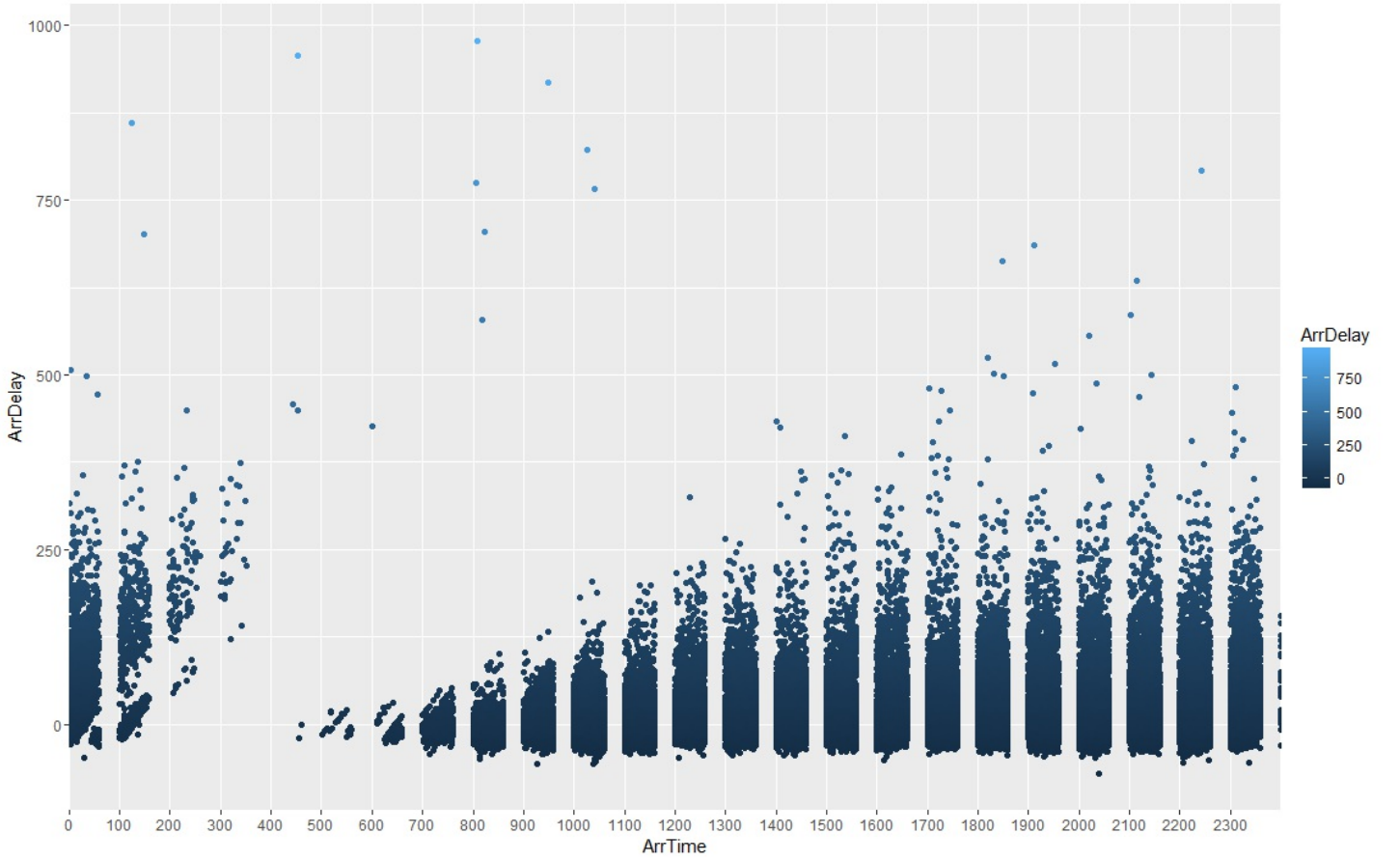
```
ggplot(hflights, aes(DepTime, DepDelay)) + geom_point(aes(colour = DepDelay)) +
  scale_x_discrete(limits=c(0:2359), breaks = seq(0, 2359, 100))
```



Arrival delay plot

Hide

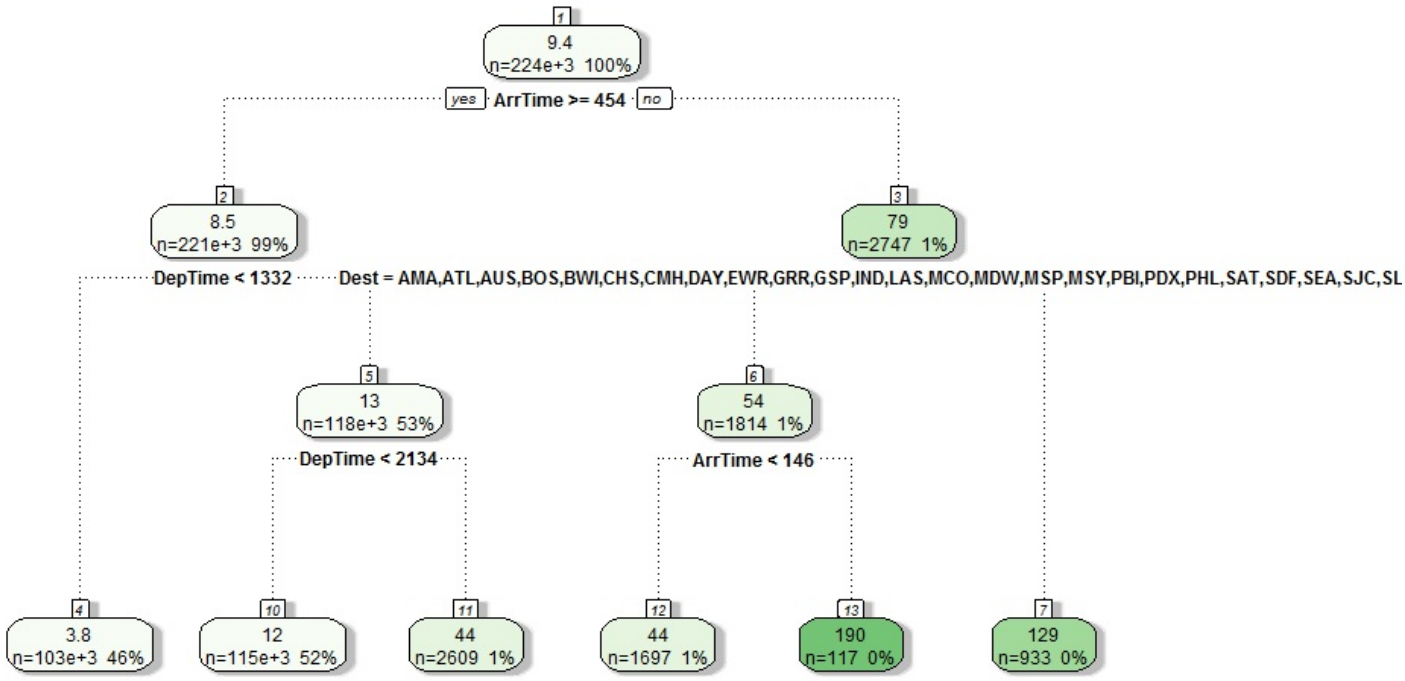
```
ggplot(hflights, aes(ArrTime, ArrDelay)) + geom_point(aes(colour = ArrDelay)) +
  scale_x_discrete(limits=c(0:2359), breaks = seq(0, 2359, 100)) +
  scale_fill_distiller(palette = "Spectral")
```



3. Departure delay – regresion tree

Hide

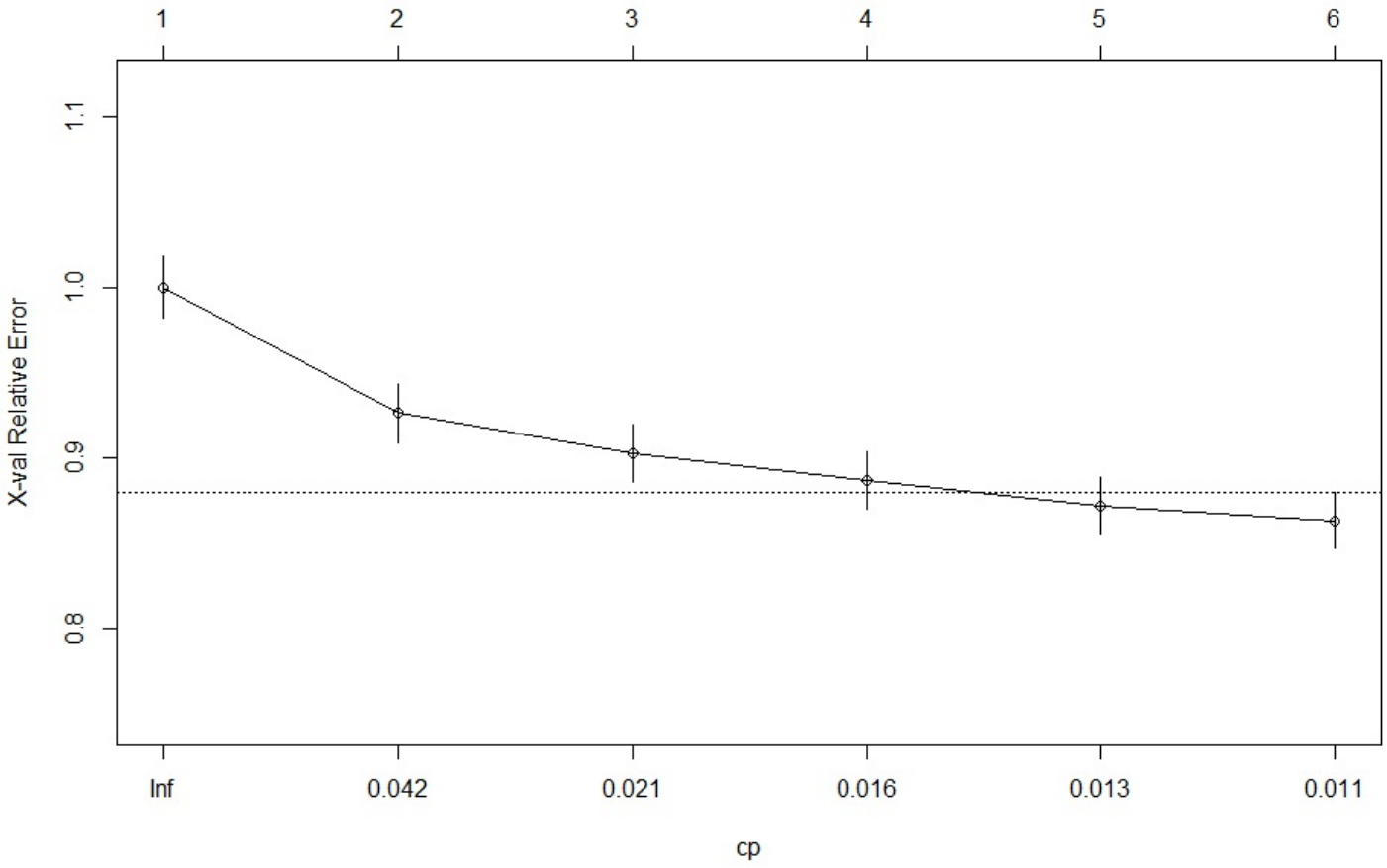
```
formula =  DepDelay ~ Month + DayofMonth +DayOfWeek + DepTime + ArrTime + UniqueCarrier + FlightNum +  ActualEl
apsedTime + AirTime + Origin + Dest + Distance + TaxiIn + TaxiOut
set.seed(7)
fit <- rpart(formula, data=hflights, method="anova")
fancyRpartPlot(fit)
```



Rattle 2016-Nov-16 20:45:28 frank

Hide

```
plotcp(fit)
```



Hide

```
fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"]
```

```
[1] 0.01
```

Hide

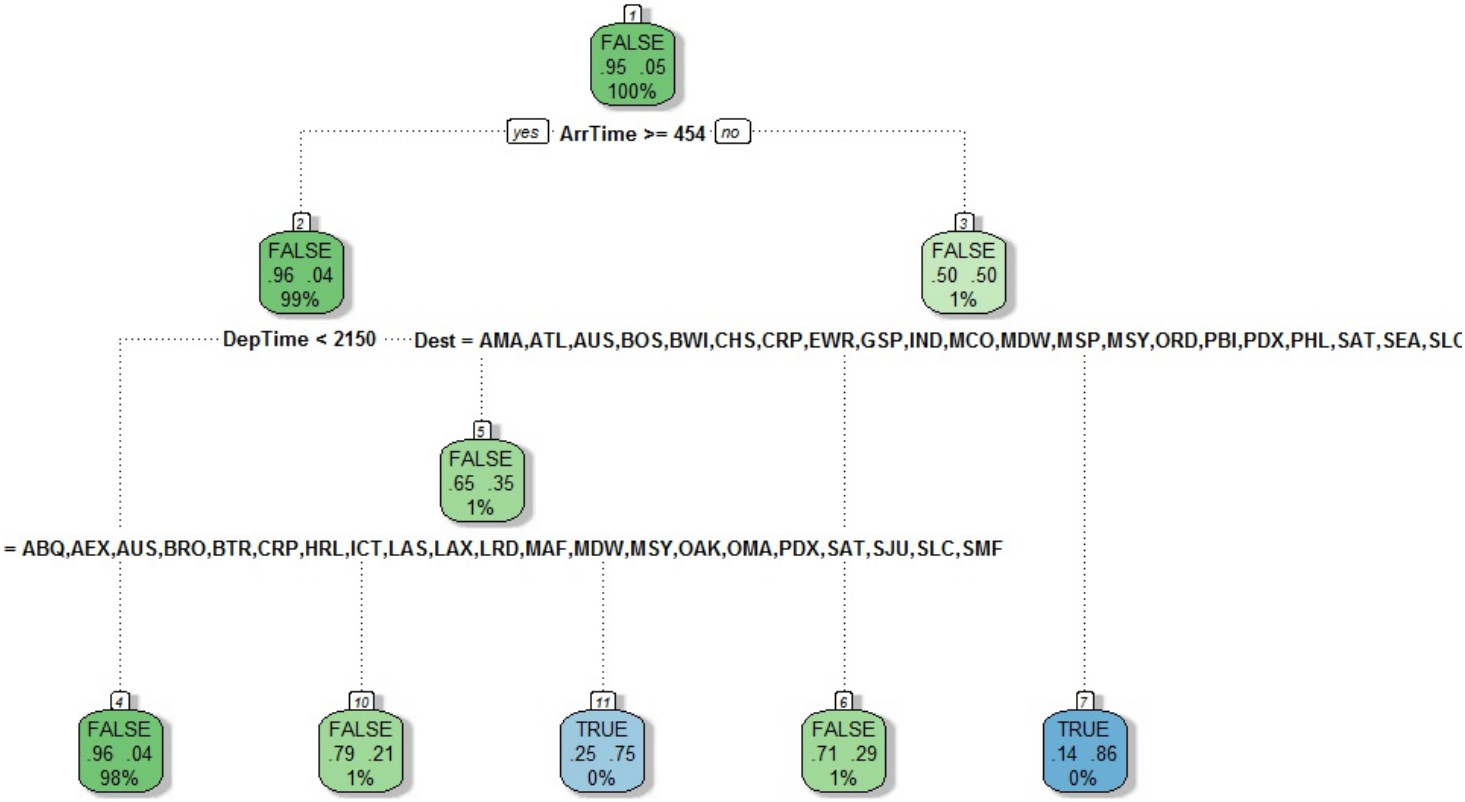
```
printcp(fit)
```

```
Regression tree:
rpart(formula = formula, data = hflights, method = "anova")

Variables actually used in tree construction:
[1] ArrTime DepTime Dest

Root node error: 184941788/223874 = 826.1

n= 223874

        CP nsplit rel error  xerror     xstd
1 0.073730      0   1.00000 1.00001 0.018194
2 0.023598      1   0.92627 0.92638 0.017066
3 0.018898      2   0.90267 0.90282 0.016942
4 0.014286      3   0.88377 0.88718 0.016945
5 0.012465      4   0.86949 0.87223 0.016848
6 0.010000      5   0.85702 0.86349 0.016399
```

4. Departure delay longer than 1 hour – classification tree

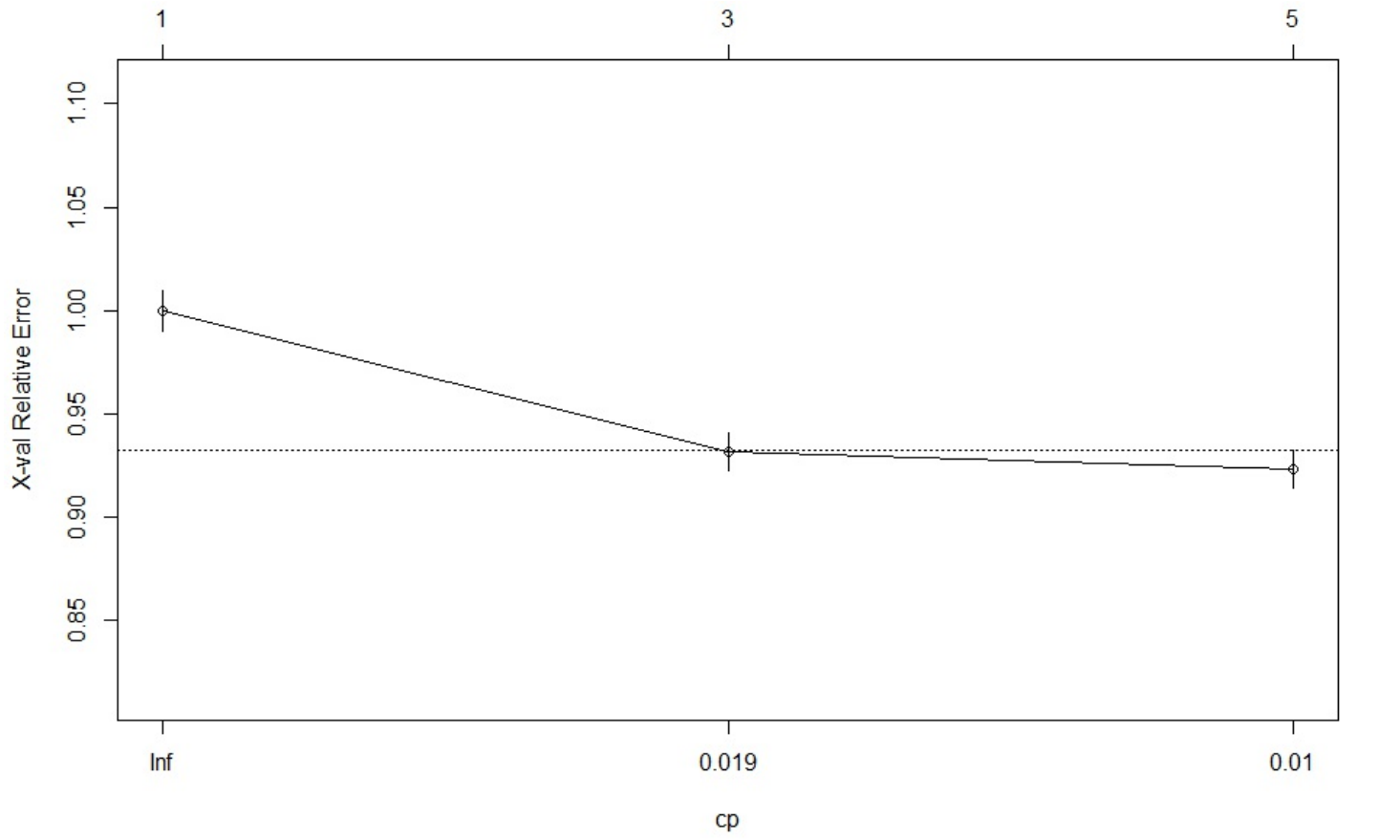
Hide

```
hflights$DepDelay60 <-hflights$DepDelay > 60
formula =  DepDelay60 ~ Month + DayofMonth +DayOfWeek + DepTime + ArrTime + UniqueCarrier + FlightNum +  Actual
ElapsedTime + AirTime + Origin + Dest + Distance + TaxiIn + TaxiOut
set.seed(7)
fit <- rpart(formula, data=hflights, method="class")
fancyRpartPlot(fit)
```



Hide

```
plotcp(fit)
```



Hide

```
fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"]
```

```
[1] 0.01
```

Hide

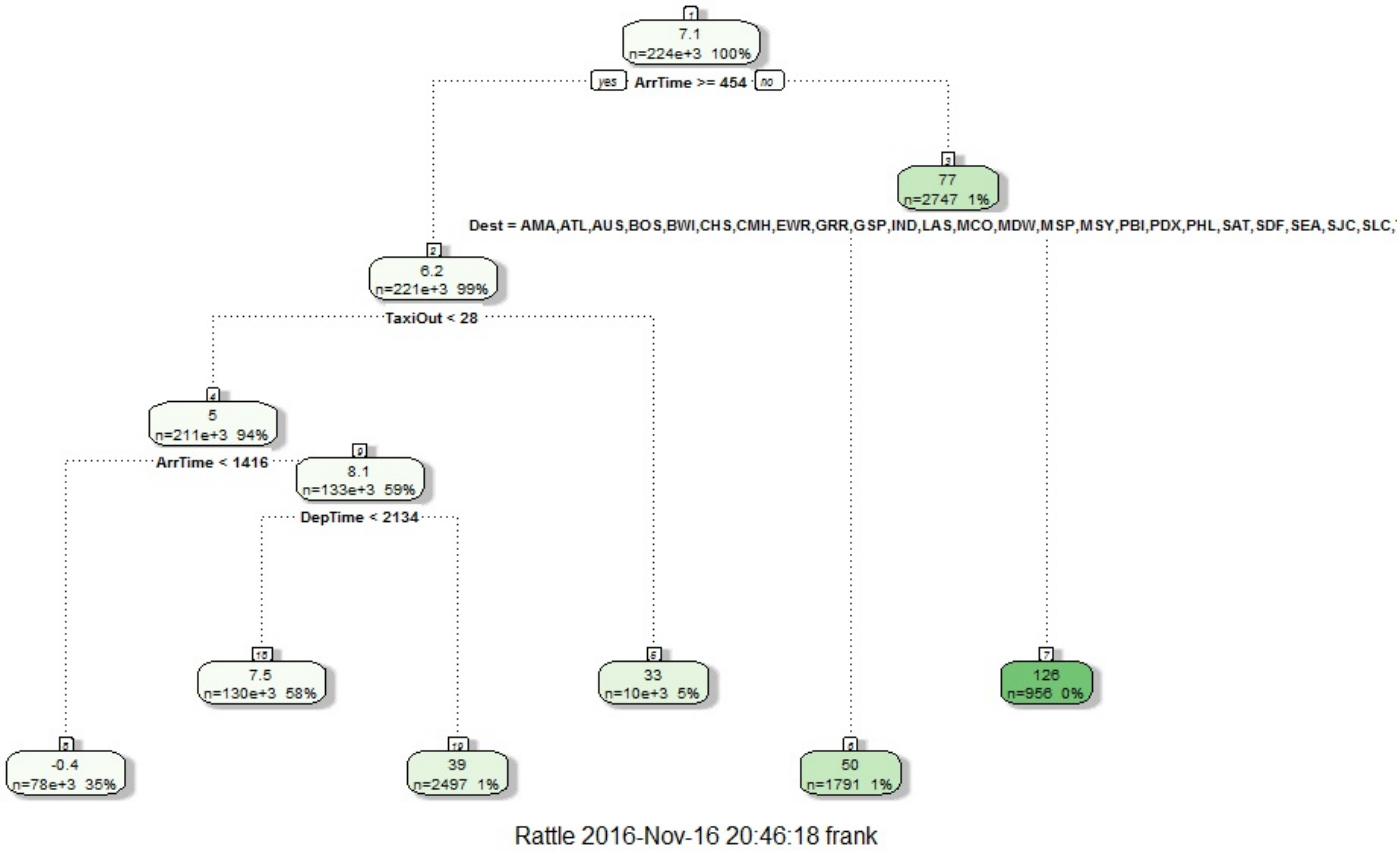
```
printcp(fit)
```

```
Classification tree:
rpart(formula = formula, data = hflights, method = "class")

Variables actually used in tree construction:
[1] ArrTime DepTime Dest

Root node error: 10164/223874 = 0.045401

n= 223874

        CP nsplit rel error  xerror      xstd
1 0.034878      0   1.00000 1.00000 0.0096912
2 0.010527      2   0.93024 0.93172 0.0093697
3 0.010000      4   0.90919 0.92326 0.0093289
```

5. Arrival delay – regresion tree

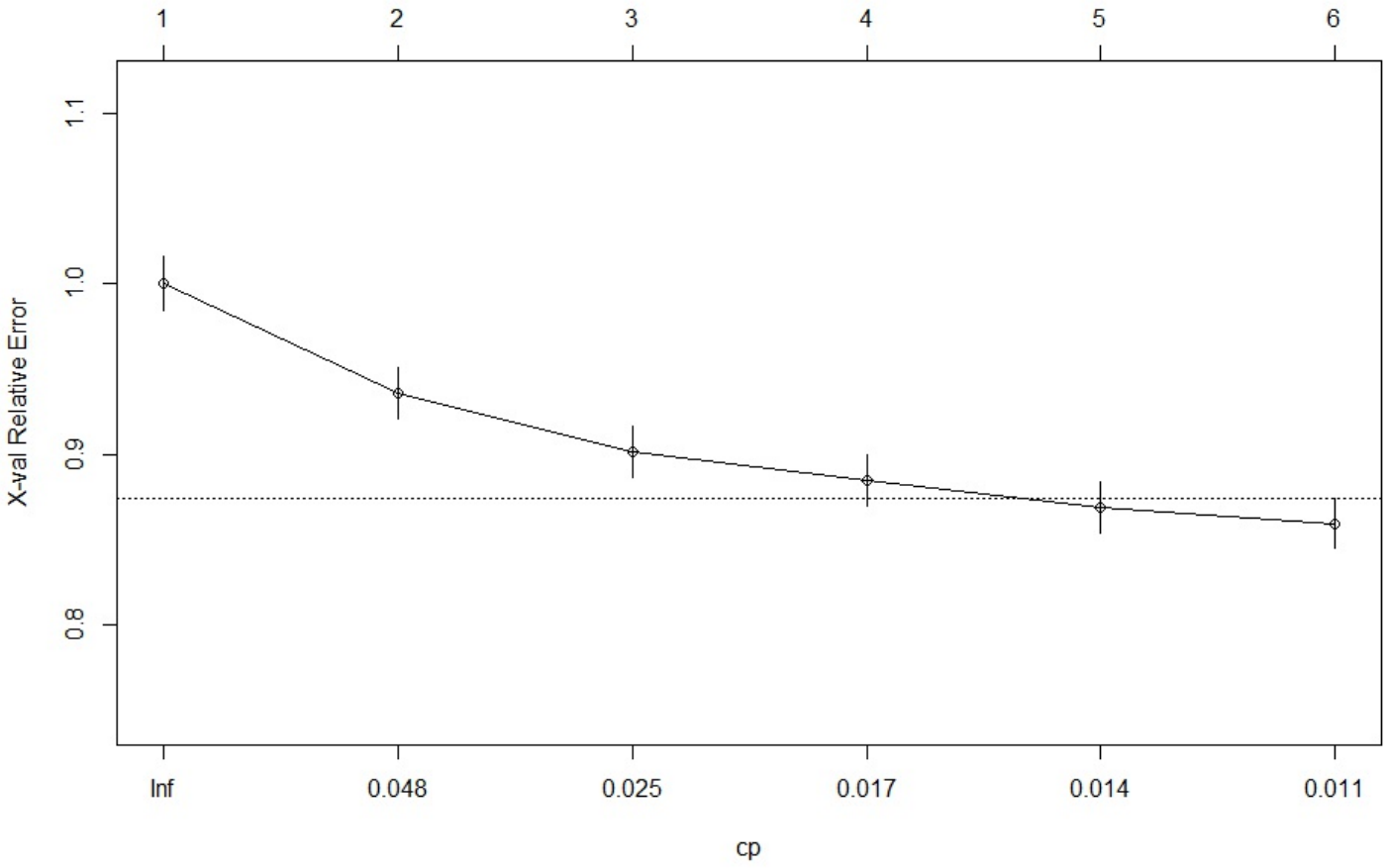
Hide

```
formula =  ArrDelay ~ Month + DayofMonth +DayOfWeek + DepTime + ArrTime + UniqueCarrier + FlightNum +  ActualEl
apsedTime + AirTime + Origin + Dest + Distance + TaxiIn + TaxiOut
set.seed(7)
fit <- rpart(formula, data=hflights, method="anova")
fancyRpartPlot(fit)
```



Hide

```
plotcp(fit)
```



Hide

```
fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"]
```

```
[1] 0.01
```

Hide

```
printcp(fit)
```

```
Regression tree:
rpart(formula = formula, data = hflights, method = "anova")

Variables actually used in tree construction:
[1] ArrTime DepTime Dest    TaxiOut

Root node error: 211115143/223874 = 943.01

n= 223874

        CP nsplit rel error  xerror     xstd
1 0.064126      0   1.00000 1.00000 0.016036
2 0.035341      1   0.93587 0.93597 0.015133
3 0.017024      2   0.90053 0.90101 0.014982
4 0.016949      3   0.88351 0.88435 0.014937
5 0.011297      4   0.86656 0.86897 0.014888
6 0.010000      5   0.85526 0.85932 0.014821
```

6. Arrival delay longer than 1 hour – classification tree

Hide

```
hflights$ArrDelay60 <-hflights$ArrDelay > 60
formula =  ArrDelay60 ~ Month + DayofMonth +DayOfWeek + DepTime + ArrTime + UniqueCarrier + FlightNum +  Actual
ElapsedTime + AirTime + Origin + Dest + Distance + TaxiIn + TaxiOut
set.seed(7)
fit <- rpart(formula, data=hflights, method="class")
```