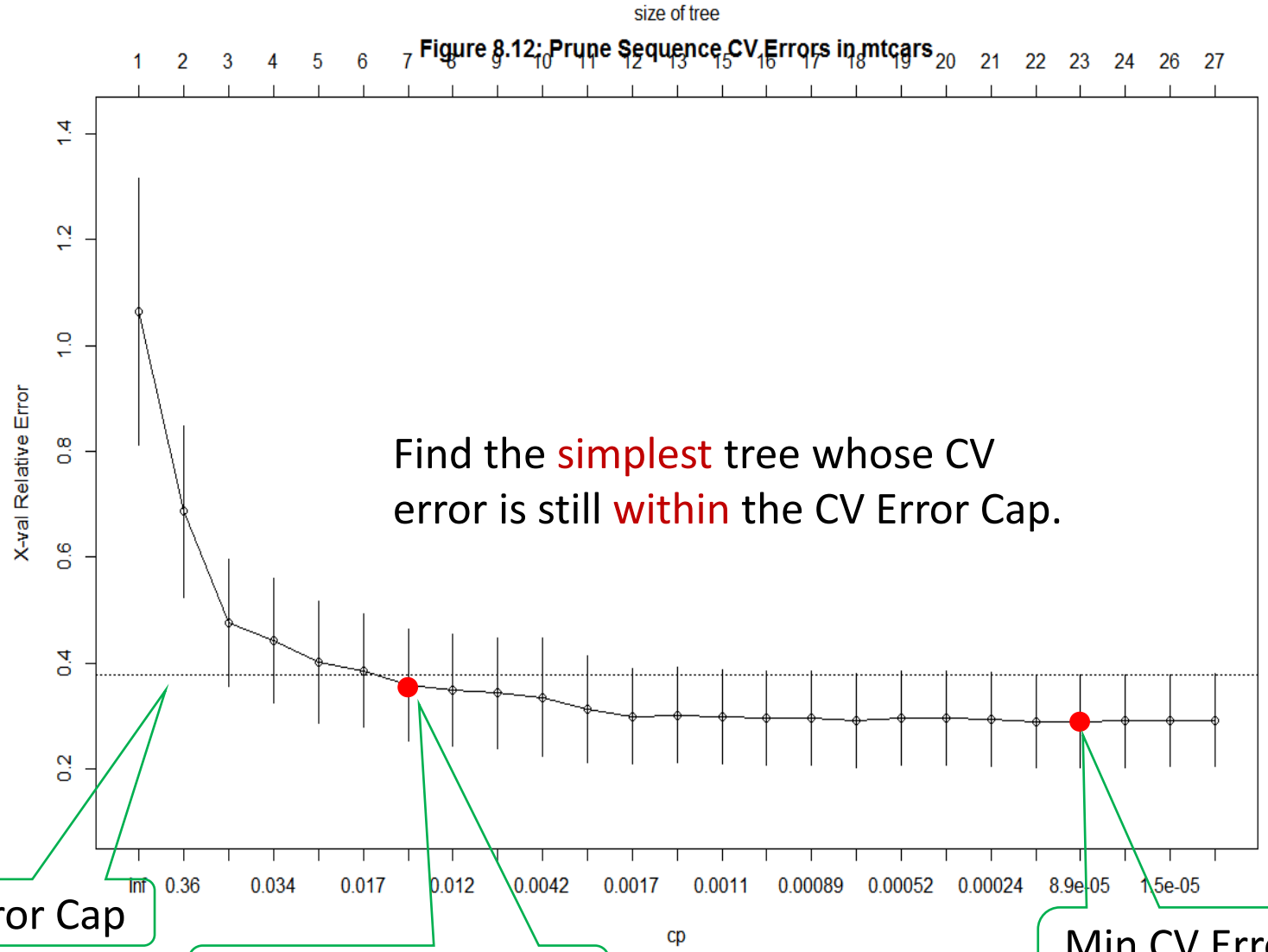
(Optional) Automating the Search for Optimal Tree

Rscript: mtcars CART.R

CART

Based on Chew C. H. (2020) textbook: AI, Analytics and Data Science. Vol 1., Chap 8.

CART on mtcars dataset shows 25 trees
[32 cases, 10 X variables, continuous Y variable]



Find the simplest tree whose CV error is still within the CV Error Cap.

CV Error Cap

Optimal Tree #7

Min CV Error Tree #22

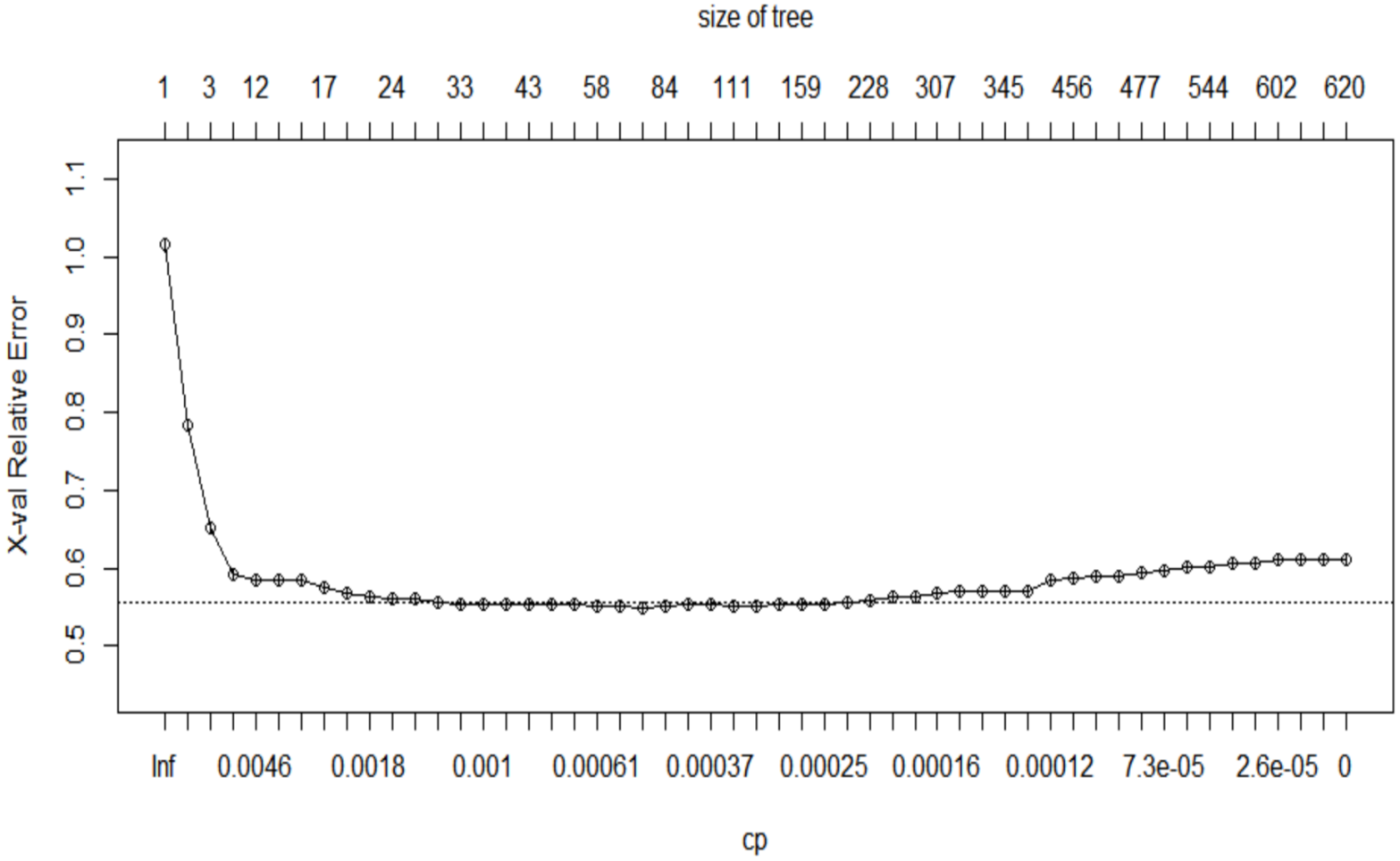
CP Table can be used to search for Optimal Tree too

```
Root node error: 1126/32 = 35.189

n= 32

           CP nsplit  rel error  xerror     xstd
1  6.5266e-01      0 1.0000e+00 1.06389 0.252198
2  1.9470e-01      1 3.4734e-01 0.68629 0.160870
3  4.5774e-02      2 1.5264e-01 0.47604 0.119551
4  2.5328e-02      3 1.0686e-01 0.44324 0.117846
5  2.3250e-02      4 8.1534e-02 0.40281 0.115329
6  1.2488e-02      5 5.8285e-02 0.38559 0.106955
7  1.2149e-02      6 4.5796e-02 0.35818
8  1.1647e-02      7 3.3648e-02 0.34943 0.105751
9  9.6700e-03      8 2.2000e-02 0.34357 0.104871
10 1.8010e-03      9 1.2330e-02 0.33605 0.112381
11 1.8010e-03     10 1.0529e-02 0.31304 0.099915
12 1.5156e-03     11 8.7282e-03 0.29965 0.090460
13 1.2868e-03     12 7.2125e-03 0.30216 0.090476
14 9.9907e-04     14 4.6389e-03 0.29853 0.089054
15 9.2506e-04     15 3.6399e-03 0.29704 0.089144
16 8.5254e-04     16 2.7148e-03 0.29628 0.089205
17 7.5041e-04     17 1.8623e-03 0.29221 0.089117
18 3.5967e-04     18 1.1119e-03 0.29642 0.088779
19 2.8418e-04     19 7.5219e-04 0.29642 0.088779
20 2.0011e-04     20 4.6801e-04 0.29479 0.088862
21 1.1101e-04     21 2.6790e-04 0.29055 0.086900
22 7.1045e-05     22 1.5689e-04 0.29013 0.086937
23 3.9963e-05     23 8.5846e-05 0.29080 0.086896
24 5.9204e-06     25 5.9204e-06 0.29159 0.086831
25 0.0000e+00     26 0.0000e+00 0.29237 0.087231
```

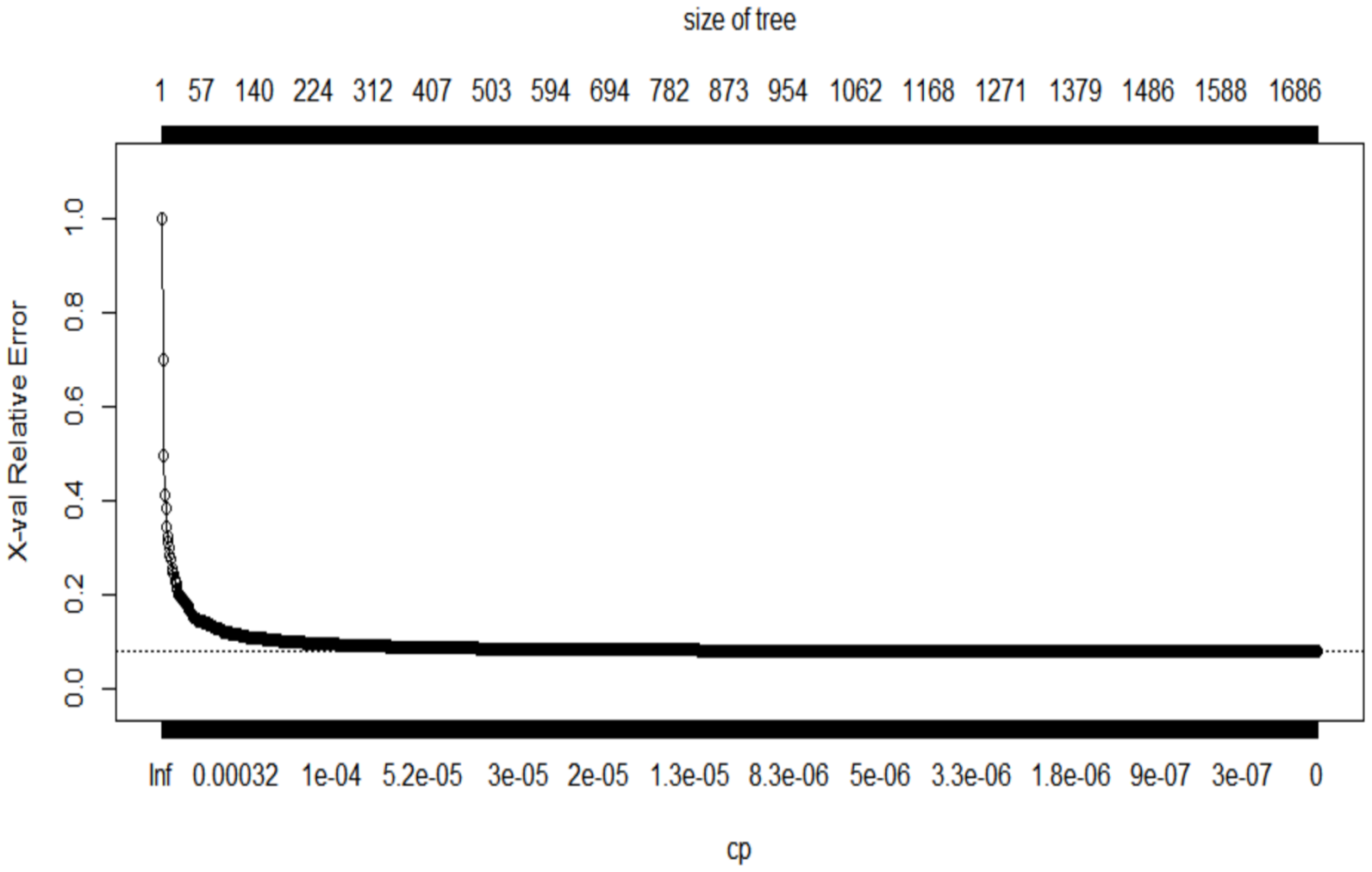
```
> 0.29013 + 0.086937
[1] 0.377067
```

Min CV Error Tree

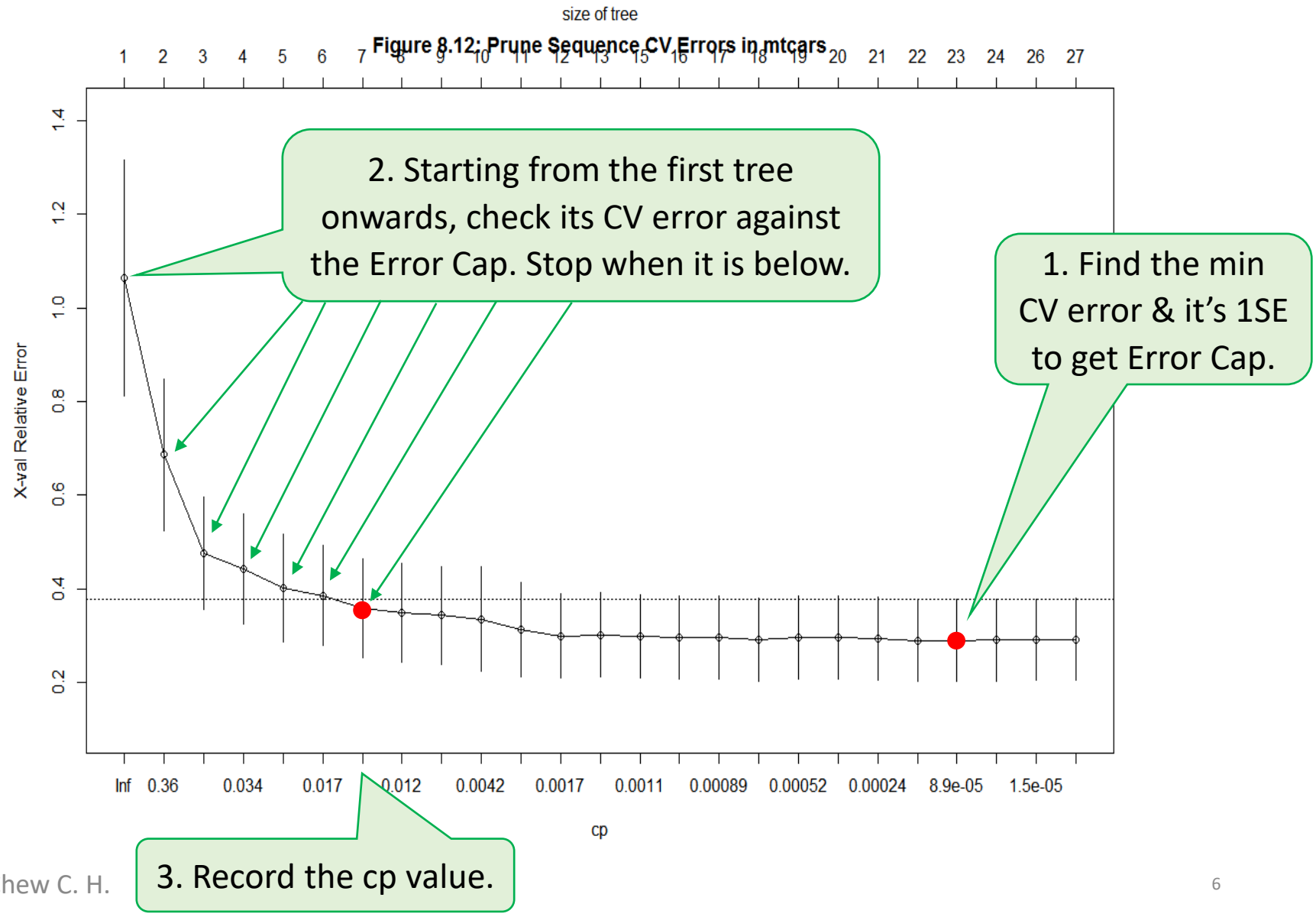
Census data sample: 17,296 rows. Categorical Y. Easy to find optimal tree?



HDB Resale Flat data: 22,204 cases, Continuous Y. Easy to find optimal tree?



How did we identify the optimal tree's CP value?



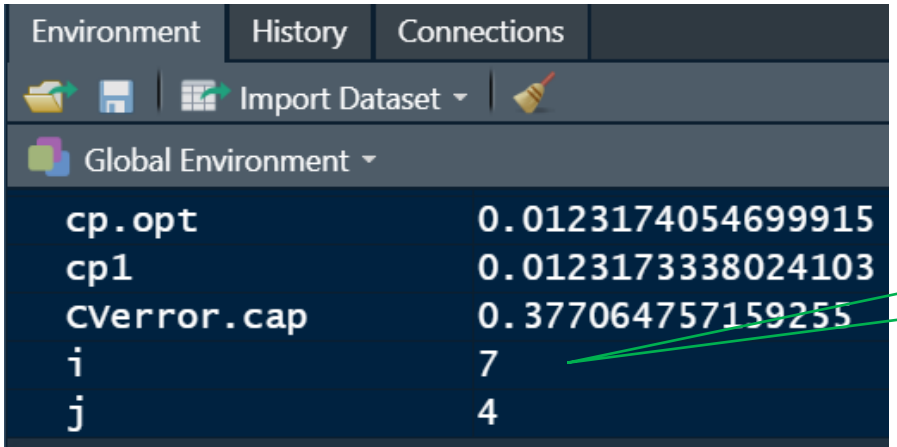
Rscript to Automate the Search for Optimal Tree

2. Starting from the first tree onwards, check its CV error against the Error Cap. Stop when it is below.

1. Find the min CV error & it's 1SE to get Error Cap.

```
35  # [Optional] Extract the Optimal Tree via code instead of eye power ----------
36  # Compute min CVerror + 1SE in maximal tree cart1.
37  CVerror.cap <- cart1$cptable[which.min(cart1$cptable[,"xerror"]), "xerror"] +
38    cart1$cptable[which.min(cart1$cptable[,"xerror"]), "xstd"]
39
40  # Find the optimal CP region whose CV error is just below CVerror.cap in maximal tree cart1.
41  i <- 1; j<- 4
42  while (cart1$cptable[i,j] > CVerror.cap) {
43    i <- i + 1
44  }
45
46  # Get geometric mean of the two identified CP values in the optimal region
47  # if optimal tree has at least one split.
48  cp.opt = ifelse(i > 1, sqrt(cart1$cptable[i,1] * cart1$cptable[i-1,1]), 1)
49  # ------------------------------------------------------------------------------
```

3. Record the cp value.



The 7th tree is optimal.

Important: Remember to adjust the Rscript to use the name of your maximal CART model

```
35  # [Optional] Extract the Optimal Tree via code instead of eye power ------------
36  # Compute min CVerror + 1SE in maximal tree cart1.
37  CVerror.cap <- cart1$cptable[which.min(cart1$cptable[,"xerror"]), "xerror"] +
38    cart1$cptable[which.min(cart1$cptable[,"xerror"]), "xstd"]
39
40  # Find the optimal CP region whose CV error is just below CVerror.cap in maximal tree cart1.
41  i <- 1; j<- 4
42  while (cart1$cptable[i,j] > CVerror.cap) {
43    i <- i + 1
44  }
45
46  # Get geometric mean of the two identified CP values in the optimal region
47  # if optimal tree has at least one split.
48  cp.opt = ifelse(i > 1, sqrt(cart1$cptable[i,1] * cart1$cptable[i-1,1]), 1)
49  # ---------------------------------------------------------------------------
```

- The maximal tree in this script was named cart1.
- Replace it with the name of your maximal tree.
- Select the 3 blocks of code and use <CTRL> + <F> on windows machines to find and replace all occurrence of cart1.

Advice on using this automation

- Use plotcp() chart if there are not too many trees and the optimal cp can be obviously identified from the chart.
- Else, use my R script to extract the cp value of the optimal tree.
- Remember to correct the name of the maximal tree before use.