Assignment 4

Group 2

2023-10-27

Task 1

Task 2

Task 3

Task 4

Task 5

```
suppressMessages(library("ISLR"))
suppressMessages(library("rpart"))
suppressMessages(library("tree"))
suppressMessages(library("dplyr"))
suppressMessages(library("rpart.plot"))
suppressMessages(library("knitr"))
```

```
data("Carseats", package="ISLR")
df=Carseats
```

a) Split the data set into a training set and a test set. – (70% train-30% test)

```
set.seed(123)
df$id = 1:nrow(df)
train = df %>% dplyr::sample_frac(0.70)
test = dplyr::anti_join(df, train, by = 'id')
train = train[-c(12)]
test = test[-c(12)]
```

b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

```
set.seed(123)
tree = rpart(Sales ~ ., data = train)
summary(tree)
```

```
## Call:
## rpart(formula = Sales ~ ., data = train)
##   n= 280
##
##            CP nsplit rel error    xerror       xstd
## 1  0.25462228      0 1.0000000 1.0091773 0.08510095
## 2  0.09215223      1 0.7453777 0.7573932 0.05919359
## 3  0.07090167      2 0.6532255 0.7416373 0.05767966
## 4  0.04324517      3 0.5823238 0.6393493 0.04881898
## 5  0.03604927      4 0.5390787 0.6672966 0.05248835
## 6  0.03227129      5 0.5030294 0.6551342 0.05261532
## 7  0.02428572      7 0.4384868 0.5957057 0.04893045
## 8  0.01748177      8 0.4142011 0.5958479 0.05110351
## 9  0.01591543      9 0.3967193 0.6162069 0.05494734
## 10 0.01562578     10 0.3808039 0.6195996 0.05473044
## 11 0.01413317     11 0.3651781 0.6135908 0.05355575
## 12 0.01354372     12 0.3510449 0.6083060 0.05298598
## 13 0.01265304     14 0.3239575 0.6031406 0.05311808
## 14 0.01046095     15 0.3113044 0.6012050 0.05337364
## 15 0.01000000     16 0.3008435 0.6055665 0.05434016
##
## Variable importance
##   ShelveLoc       Price   CompPrice         Age  Population Advertising
##          35          31          17           5           4           3
##   Education      Income
##           3           3
##
## Node number 1: 280 observations,    complexity param=0.2546223
##   mean=7.437786, MSE=7.954364
##   left son=2 (222 obs) right son=3 (58 obs)
##   Primary splits:
##       ShelveLoc   splits as  LRL,       improve=0.25462230, (0 missing)
##       Price       < 105.5 to the right, improve=0.13694340, (0 missing)
##       Age         < 65.5  to the right, improve=0.10010780, (0 missing)
##       Advertising < 7.5   to the left,  improve=0.06180591, (0 missing)
##       Income      < 61.5  to the left,  improve=0.03311595, (0 missing)
##
## Node number 2: 222 observations,    complexity param=0.09215223
##   mean=6.71036, MSE=5.627182
##   left son=4 (150 obs) right son=5 (72 obs)
##   Primary splits:
##       Price       < 105.5 to the right, improve=0.16429540, (0 missing)
##       ShelveLoc   splits as  L-R,       improve=0.11604670, (0 missing)
##       Age         < 68.5  to the right, improve=0.09259253, (0 missing)
##       Advertising < 7.5   to the left,  improve=0.08661964, (0 missing)
##       Income      < 61.5  to the left,  improve=0.07886769, (0 missing)
##   Surrogate splits:
##       CompPrice  < 109.5 to the right, agree=0.761, adj=0.264, (0 split)
##       Population < 507.5 to the left,  agree=0.685, adj=0.028, (0 split)
##       Income     < 22.5  to the right, agree=0.680, adj=0.014, (0 split)
##
## Node number 3: 58 observations,    complexity param=0.07090167
##   mean=10.22207, MSE=7.084261
##   left son=6 (38 obs) right son=7 (20 obs)
```

```
##   Primary splits:
##       Price       < 109.5 to the right, improve=0.38432390, (0 missing)
##       Age         < 61.5  to the right, improve=0.15967180, (0 missing)
##       Education   < 11.5  to the right, improve=0.11849500, (0 missing)
##       Advertising < 13.5  to the left,  improve=0.11063440, (0 missing)
##       CompPrice   < 131.5 to the left,  improve=0.08607235, (0 missing)
##   Surrogate splits:
##       Population  < 92.5  to the right, agree=0.741, adj=0.25, (0 split)
##       Education   < 11.5  to the right, agree=0.707, adj=0.15, (0 split)
##       CompPrice   < 102   to the right, agree=0.672, adj=0.05, (0 split)
##       Advertising < 15.5  to the left,  agree=0.672, adj=0.05, (0 split)
##
## Node number 4: 150 observations,    complexity param=0.04324517
##   mean=6.0442, MSE=4.453584
##   left son=8 (44 obs) right son=9 (106 obs)
##   Primary splits:
##       ShelveLoc   splits as  L-R,       improve=0.14417840, (0 missing)
##       CompPrice   < 124.5 to the left,  improve=0.11790140, (0 missing)
##       Advertising < 7.5   to the left,  improve=0.10645280, (0 missing)
##       Age         < 65.5  to the right, improve=0.08327840, (0 missing)
##       Income      < 61.5  to the left,  improve=0.08264313, (0 missing)
##   Surrogate splits:
##       Population < 15    to the left,  agree=0.720, adj=0.045, (0 split)
##       Age        < 28.5  to the left,  agree=0.720, adj=0.045, (0 split)
##       Price      < 162.5 to the right, agree=0.713, adj=0.023, (0 split)
##
## Node number 5: 72 observations,    complexity param=0.03227129
##   mean=8.098194, MSE=5.221573
##   left son=10 (51 obs) right son=11 (21 obs)
##   Primary splits:
##       CompPrice  < 123.5 to the left,  improve=0.19013200, (0 missing)
##       Age        < 54.5  to the right, improve=0.18899550, (0 missing)
##       ShelveLoc  splits as  L-R,       improve=0.16987950, (0 missing)
##       Price      < 88    to the right, improve=0.09985730, (0 missing)
##       Population < 162   to the right, improve=0.08620326, (0 missing)
##   Surrogate splits:
##       Price      < 103.5 to the left,  agree=0.750, adj=0.143, (0 split)
##       Income     < 34.5  to the right, agree=0.722, adj=0.048, (0 split)
##       Population < 494   to the left,  agree=0.722, adj=0.048, (0 split)
##
## Node number 6: 38 observations,    complexity param=0.02428572
##   mean=9.025, MSE=4.931751
##   left son=12 (7 obs) right son=13 (31 obs)
##   Primary splits:
##       Price       < 144   to the right, improve=0.2886222, (0 missing)
##       Age         < 63.5  to the right, improve=0.1833129, (0 missing)
##       US          splits as  LR,        improve=0.1796395, (0 missing)
##       Advertising < 0.5   to the left,  improve=0.1796395, (0 missing)
##       CompPrice   < 121.5 to the left,  improve=0.1434922, (0 missing)
##   Surrogate splits:
##       Income < 104.5 to the right, agree=0.842, adj=0.143, (0 split)
##
## Node number 7: 20 observations
##   mean=12.4965, MSE=3.278343
```

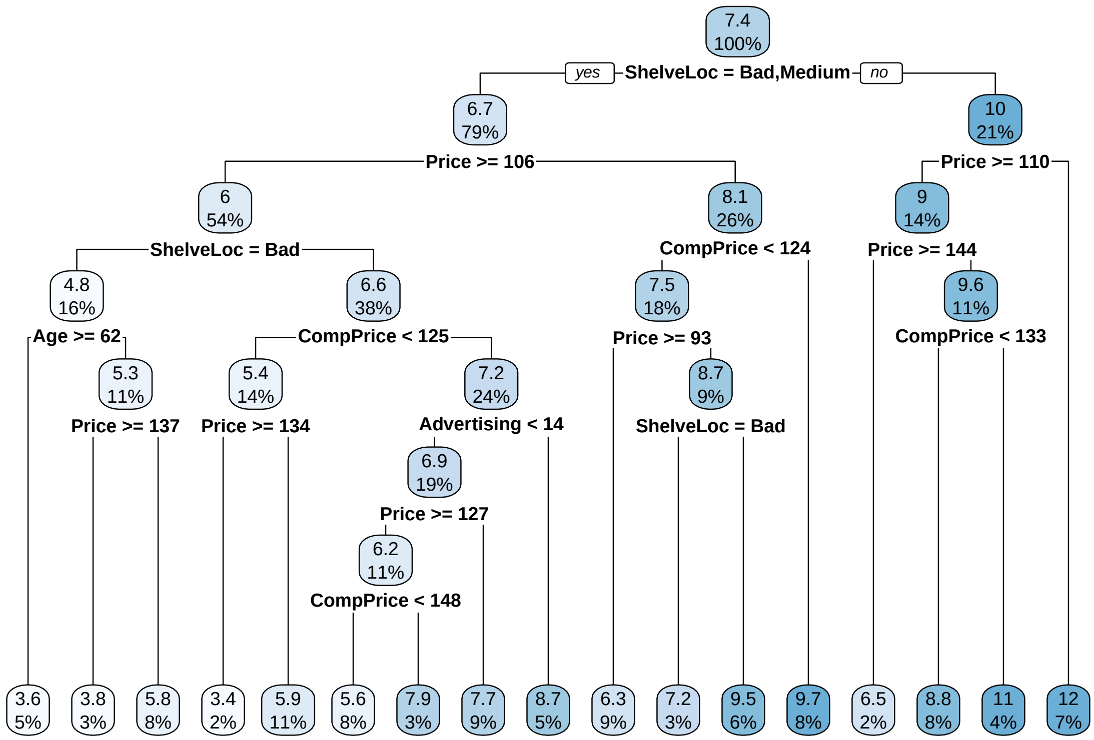
```
##
## Node number 8: 44 observations,    complexity param=0.01265304
##   mean=4.800455, MSE=3.601359
##   left son=16 (13 obs) right son=17 (31 obs)
##   Primary splits:
##       Age        < 61.5  to the right, improve=0.17784400, (0 missing)
##       CompPrice  < 144   to the left,  improve=0.16485990, (0 missing)
##       Population < 283   to the left,  improve=0.11292850, (0 missing)
##       Price      < 132.5 to the right, improve=0.11036090, (0 missing)
##       Income     < 101   to the left,  improve=0.09771273, (0 missing)
##   Surrogate splits:
##       CompPrice < 124.5 to the left,  agree=0.773, adj=0.231, (0 split)
##       Income    < 33.5  to the left,  agree=0.727, adj=0.077, (0 split)
##       Price     < 119   to the left,  agree=0.727, adj=0.077, (0 split)
##
## Node number 9: 106 observations,    complexity param=0.03604927
##   mean=6.560472, MSE=3.898691
##   left son=18 (39 obs) right son=19 (67 obs)
##   Primary splits:
##       CompPrice   < 124.5 to the left,  improve=0.19428320, (0 missing)
##       Income      < 61.5  to the left,  improve=0.15169500, (0 missing)
##       Advertising < 6.5   to the left,  improve=0.12475180, (0 missing)
##       Age         < 49.5  to the right, improve=0.12023020, (0 missing)
##       Price       < 135.5 to the right, improve=0.07821028, (0 missing)
##   Surrogate splits:
##       Price      < 111.5 to the left,  agree=0.736, adj=0.282, (0 split)
##       Population < 499.5 to the right, agree=0.651, adj=0.051, (0 split)
##       Income     < 29.5  to the left,  agree=0.642, adj=0.026, (0 split)
##       Age        < 78.5  to the right, agree=0.642, adj=0.026, (0 split)
##
## Node number 10: 51 observations,    complexity param=0.03227129
##   mean=7.458824, MSE=4.963912
##   left son=20 (26 obs) right son=21 (25 obs)
##   Primary splits:
##       Price     < 92.5  to the right, improve=0.2854717, (0 missing)
##       Income    < 100.5 to the left,  improve=0.2085942, (0 missing)
##       ShelveLoc splits as  L-R,       improve=0.1890332, (0 missing)
##       Age       < 35.5  to the right, improve=0.1583237, (0 missing)
##       Education < 11.5  to the left,  improve=0.1570051, (0 missing)
##   Surrogate splits:
##       CompPrice  < 99.5  to the right, agree=0.667, adj=0.32, (0 split)
##       Education  < 13.5  to the left,  agree=0.667, adj=0.32, (0 split)
##       Population < 271   to the right, agree=0.647, adj=0.28, (0 split)
##       Age        < 49    to the right, agree=0.627, adj=0.24, (0 split)
##       Income     < 50.5  to the left,  agree=0.588, adj=0.16, (0 split)
##
## Node number 11: 21 observations
##   mean=9.650952, MSE=2.443475
##
## Node number 12: 7 observations
##   mean=6.514286, MSE=2.881396
##
## Node number 13: 31 observations,    complexity param=0.01748177
##   mean=9.591935, MSE=3.649906
```

```
##   left son=26 (21 obs) right son=27 (10 obs)
##   Primary splits:
##       CompPrice   < 132.5 to the left,  improve=0.3441166, (0 missing)
##       Age         < 61.5  to the right, improve=0.1886891, (0 missing)
##       Advertising < 12.5  to the left,  improve=0.1215375, (0 missing)
##       US          splits as  LR,        improve=0.1184813, (0 missing)
##       Income      < 41.5  to the left,  improve=0.1115535, (0 missing)
##   Surrogate splits:
##       Price < 138   to the left,  agree=0.742, adj=0.2, (0 split)
##
## Node number 16: 13 observations
##   mean=3.564615, MSE=1.418609
##
## Node number 17: 31 observations,    complexity param=0.01046095
##   mean=5.31871, MSE=3.607637
##   left son=34 (8 obs) right son=35 (23 obs)
##   Primary splits:
##       Price       < 136.5 to the right, improve=0.2083292, (0 missing)
##       Population  < 283   to the left,  improve=0.1647044, (0 missing)
##       CompPrice   < 144   to the left,  improve=0.1201417, (0 missing)
##       Advertising < 8.5   to the left,  improve=0.1047802, (0 missing)
##       Income      < 87    to the left,  improve=0.1020453, (0 missing)
##   Surrogate splits:
##       Age       < 27.5  to the left,  agree=0.871, adj=0.500, (0 split)
##       Education < 11.5  to the left,  agree=0.774, adj=0.125, (0 split)
##
## Node number 18: 39 observations,    complexity param=0.01562578
##   mean=5.419744, MSE=3.270602
##   left son=36 (7 obs) right son=37 (32 obs)
##   Primary splits:
##       Price       < 133.5 to the right, improve=0.2728430, (0 missing)
##       Advertising < 6     to the left,  improve=0.2590152, (0 missing)
##       Income      < 83.5  to the left,  improve=0.1940935, (0 missing)
##       US          splits as  LR,        improve=0.1728556, (0 missing)
##       Age         < 68    to the right, improve=0.1210741, (0 missing)
##
## Node number 19: 67 observations,    complexity param=0.01591543
##   mean=7.224478, MSE=3.065941
##   left son=38 (54 obs) right son=39 (13 obs)
##   Primary splits:
##       Advertising < 13.5  to the left,  improve=0.1725612, (0 missing)
##       Price       < 127   to the right, improve=0.1665229, (0 missing)
##       Income      < 57.5  to the left,  improve=0.1333498, (0 missing)
##       Age         < 54.5  to the right, improve=0.1172588, (0 missing)
##       Education   < 16.5  to the right, improve=0.1134184, (0 missing)
##   Surrogate splits:
##       CompPrice < 127.5 to the right, agree=0.821, adj=0.077, (0 split)
##
## Node number 20: 26 observations
##   mean=6.291538, MSE=3.624328
##
## Node number 21: 25 observations,    complexity param=0.01413317
##   mean=8.6728, MSE=3.466284
##   left son=42 (9 obs) right son=43 (16 obs)
```

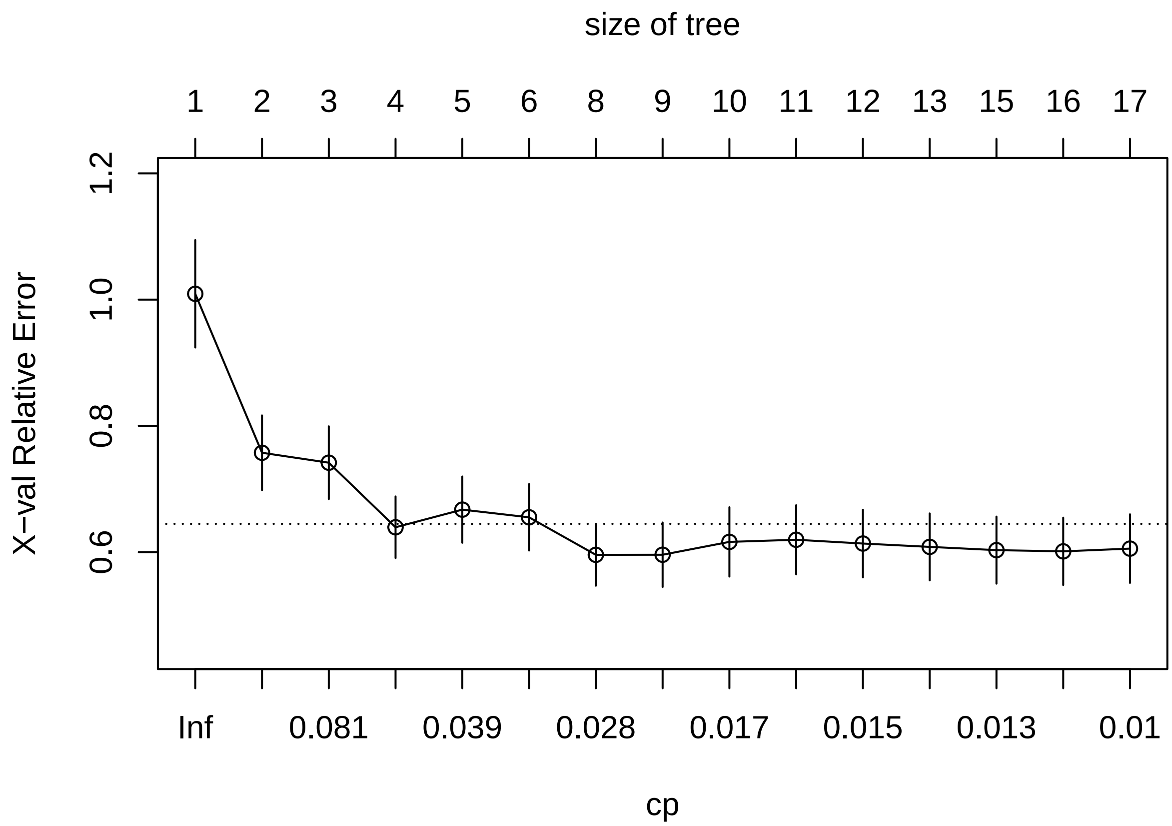
```
##   Primary splits:
##       ShelveLoc   splits as  L-R,       improve=0.36324440, (0 missing)
##       Price       < 75.5  to the right, improve=0.17532440, (0 missing)
##       Income      < 62    to the left,  improve=0.15169510, (0 missing)
##       Population  < 336   to the left,  improve=0.08664599, (0 missing)
##       Advertising < 11.5  to the left,  improve=0.05254227, (0 missing)
##   Surrogate splits:
##       Income      < 47.5  to the left,  agree=0.76, adj=0.333, (0 split)
##       Age         < 45    to the left,  agree=0.76, adj=0.333, (0 split)
##       CompPrice   < 90.5  to the left,  agree=0.68, adj=0.111, (0 split)
##       Advertising < 15    to the right, agree=0.68, adj=0.111, (0 split)
##       Population  < 296   to the right, agree=0.68, adj=0.111, (0 split)
##
## Node number 26: 21 observations
##   mean=8.818571, MSE=2.168469
##
## Node number 27: 10 observations
##   mean=11.216, MSE=2.867344
##
## Node number 34: 8 observations
##   mean=3.84875, MSE=2.911736
##
## Node number 35: 23 observations
##   mean=5.83, MSE=2.836696
##
## Node number 36: 7 observations
##   mean=3.4, MSE=2.206314
##
## Node number 37: 32 observations
##   mean=5.861562, MSE=2.415851
##
## Node number 38: 54 observations,    complexity param=0.01354372
##   mean=6.867593, MSE=2.775033
##   left son=76 (30 obs) right son=77 (24 obs)
##   Primary splits:
##       Price     < 127   to the right, improve=0.19144250, (0 missing)
##       Age       < 65    to the right, improve=0.14766490, (0 missing)
##       CompPrice < 147.5 to the left,  improve=0.12784600, (0 missing)
##       Education < 16.5  to the right, improve=0.10972240, (0 missing)
##       Income    < 41    to the left,  improve=0.09634824, (0 missing)
##   Surrogate splits:
##       CompPrice   < 132.5 to the right, agree=0.685, adj=0.292, (0 split)
##       Income      < 41    to the left,  agree=0.667, adj=0.250, (0 split)
##       Advertising < 4.5   to the right, agree=0.667, adj=0.250, (0 split)
##       Age         < 38    to the right, agree=0.611, adj=0.125, (0 split)
##       Education   < 16.5  to the right, agree=0.611, adj=0.125, (0 split)
##
## Node number 39: 13 observations
##   mean=8.706923, MSE=1.547621
##
## Node number 42: 9 observations
##   mean=7.176667, MSE=2.858156
##
## Node number 43: 16 observations
```

```
##   mean=9.514375, MSE=1.841
##
## Node number 76: 30 observations,    complexity param=0.01354372
##   mean=6.215667, MSE=2.711025
##   left son=152 (22 obs) right son=153 (8 obs)
##   Primary splits:
##       CompPrice < 147.5 to the left,  improve=0.38905020, (0 missing)
##       Age       < 65    to the right, improve=0.13371790, (0 missing)
##       Income    < 83.5  to the left,  improve=0.08814781, (0 missing)
##       Price     < 142.5 to the right, improve=0.08113575, (0 missing)
##       Education < 15.5  to the right, improve=0.05311452, (0 missing)
##   Surrogate splits:
##       Age        < 33.5  to the right, agree=0.833, adj=0.375, (0 split)
##       Population < 358.5 to the left,  agree=0.800, adj=0.250, (0 split)
##       Income     < 30.5  to the right, agree=0.767, adj=0.125, (0 split)
##       Price      < 158.5 to the left,  agree=0.767, adj=0.125, (0 split)
##
## Node number 77: 24 observations
##   mean=7.6825, MSE=1.65971
##
## Node number 152: 22 observations
##   mean=5.596364, MSE=1.783096
##
## Node number 153: 8 observations
##   mean=7.91875, MSE=1.307611
```

```r
rpart.plot(tree)
```


7.4
100%
yes
ShelveLoc = Bad,Medium
no
6.7
79%
Price >= 106
10
21%
Price >= 110
6
54%
ShelveLoc = Bad
8.1
26%
CompPrice < 124
9
14%
Price >= 144
4.8
16%
Age >= 62
6.6
38%
CompPrice < 125
7.5
18%
Price >= 93
9.6
11%
CompPrice < 133
5.3
11%
Price >= 137
5.4
14%
Price >= 134
7.2
24%
Advertising < 14
8.7
9%
ShelveLoc = Bad
6.9
19%
Price >= 127
6.2
11%
CompPrice < 148
3.6
5%
3.8
3%
5.8
8%
3.4
2%
5.9
11%
5.6
8%
7.9
3%
7.7
9%
8.7
5%
6.3
9%
7.2
3%
9.5
6%
9.7
8%
6.5
2%
8.8
8%
11
4%
12
7%

```
plotcp(tree)
```




```
test$preds = predict(tree, test)
mse = mean((test$preds - test$Sales)^2)
cat("Mean squared error is: ", mse)
```

```
## Mean squared error is:  3.784419
```

- From the plot, we can see that the root node, the variable with the highest feature importance value is ShelveLoc. It is the best predictor of the model.
- Price has the second highest value for feature importance.
- The decision tree only used 5 features out of 10.
- The algorithm splits the data based on "ShelveLoc" into two categories: Bad-Medium or not. If it's either bad or medium, the next node checks if the "Price" is higher than 106. If it is, in the next node the algorithm again goes back to "ShelveLoc" and checks if the value is bad or not.

c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

```
set.seed(123)
cv_min = tree$cptable[which.min(tree$cptable[,"xerror"]),"xerror"]
cat("Lowest cross validated error is: ", cv_min)
```

```
## Lowest cross validated error is:  0.5957057
```

```r
tc_min = tree$cptable[which.min(tree$cptable[,"xerror"]),"CP"]
cat("Optimal level of tree complexity is: ",tc_min)
```

```
## Optimal level of tree complexity is:  0.02428572
```

```r
### Pruning the tree
imin = which.min(tree$cptable[, "xerror"])
select = which(
  tree$cptable[, "xerror"] <
    sum(tree$cptable[imin, c("xerror", "xstd")]))[1]
ptree = prune(tree, cp = tree$cptable[select, "CP"])

test$pruned_preds =  predict(ptree, test)
mse_pruned = mean((test$pruned_preds - test$Sales)^2)
cat("Mean squared error is: ", mse_pruned)
```

```
## Mean squared error is:  4.979248
```

According to the results from part b, we can say that pruning the tree did not improve the test MSE. There may be different reasons for such a case. One possible explanation is that pruning simplifies the tree by removing some of the complex branches, reducing the model's overfitting problem. However, if the tree was suffering from sever overfitting, pruning may decrease the predictive power, increasing test MSE.

Another reason may be that pruning can remove important splits that were important and the removed splits might have been capturing meaningful patterns or relationships in the data, and when we eliminate them via pruning, the model may become less accurate, which explains the increased test MSE.

Task 6

Task 7

```r
set.seed(123)
x_1 = runif(100)
x_2 = rnorm(100)
x_3 = as.integer(rbernoulli(100))
```

```
## Warning: `rbernoulli()` was deprecated in purrr 1.0.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

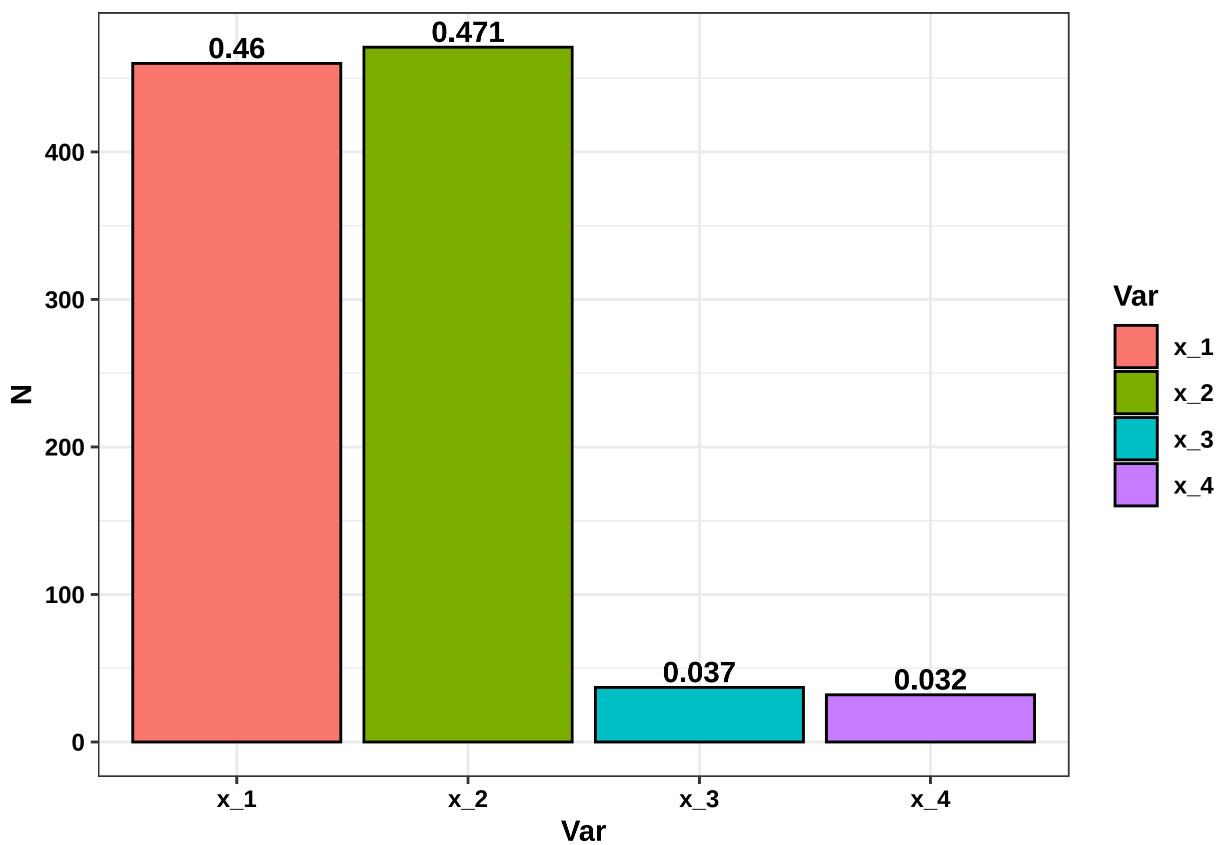
```r
x_4 = as.integer(rbernoulli(100 , p=0.1))

df = data_frame(x_1=x_1, x_2=x_2,x_3=x_3, x_4=x_4)
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## i Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```r
result = NULL
for (i in 1:1000)
{
  y=rnorm(100)
  tree = rpart(y ~ ., data = df, control = list(maxdepth = 1))
  temp = paste0(".*(", paste(colnames(df), collapse="|"), ").*")
  result = rbind(result, unique(sub(temp,"\\1", labels(tree)[-1])))
}

result %>% as.data.frame %>%
  rename(Var='V1') %>%
  group_by(Var) %>% summarise(N=n()) %>%
  ggplot(aes(x=Var,y=N,fill=Var))+
  geom_bar(stat = 'identity',color='black')+
  scale_y_continuous(labels = scales::comma_format(accuracy = 2))+
  geom_text(aes(label=N/sum(N)),vjust=-0.25,fontface='bold')+
  theme_bw()+
  theme(axis.text = element_text(color='black',face='bold'),
        axis.title = element_text(color='black',face='bold'),
        legend.text = element_text(color='black',face='bold'),
        legend.title = element_text(color='black',face='bold'))
```




```r
df_results=table(result)
kable(df_results)
```

result	Freq
x_1	460
x_2	471
x_3	37
x_4	32

According to the results, it's clear that variables X_1 and X_2 were selected more frequently for splitting in comparison to X_3 and X_4. This observation is in line with the fundamental behavior of decision trees, which tend to choose independent variables with distributions resembling that of the dependent variable y. Decision trees try to discover splits that minimize the variance, and given that y follows a normal distribution, it makes sense for the decision tree to favor independent variables with distributions closer to the normal distribution. Keeping that in mind, since X_2 follows a standard normal distribution which is more similar to the normal distribution compared to Bernoulli(Binomial) distribution the model also chooses X_1 more often for splitting. We would expect X_2 to be chosen more frequently than the other independent variables and the results are aligned with our expectations.

To summarize, the frequencies of X_1 and X_2, being higher than X_3 and X_4, can be explained by the distributional similarities to y.

Task 8

Task 9

Task 10