Data Science II: Homework 5

Name: Jasmin Martinez (JRM2319) Date: 05/03/25

Question 1: In this problem, we will apply support vector machines to predict whether a given car gets high or low gas mileage based on the dataset "auto.csv" (used in Homework 3; see Homework 3 for more details of the dataset). The response variable is mpg cat. The predictors are cylinders, displacement, horsepower, weight, acceleration, year, and origin. Split the dataset into two parts: training data (70%) and test data (30%).

```
auto = read.csv("auto.csv")
head(auto)
```

```
##   cylinders displacement horsepower weight acceleration year origin mpg_cat
## 1         8          307        130   3504         12.0   70      1     low
## 2         8          350        165   3693         11.5   70      1     low
## 3         8          318        150   3436         11.0   70      1     low
## 4         8          304        150   3433         12.0   70      1     low
## 5         8          302        140   3449         10.5   70      1     low
## 6         8          429        198   4341         10.0   70      1     low
```

```
set.seed(111111)
datSplit = initial_split(data = auto, prop = 0.7)
trainData = training(datSplit)
testData = testing(datSplit)
head(trainData)
```

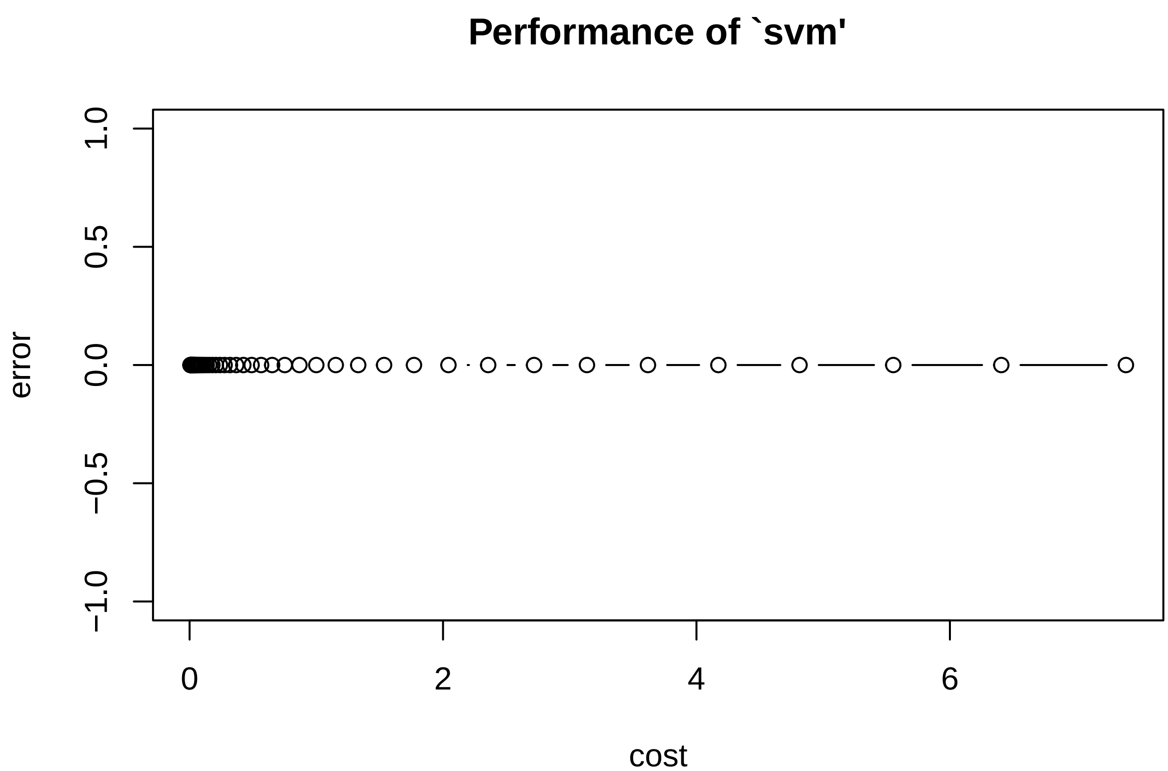
```
##   cylinders displacement horsepower weight acceleration year origin mpg_cat
## 1         4          134         95   2515         14.8   78      3     low
## 2         4          156         92   2585         14.5   82      1    high
## 3         6          168        120   3820         16.7   76      2     low
## 4         4          151         90   2670         16.0   79      1    high
## 5         6          258        110   3632         18.0   74      1     low
## 6         4           98         68   2135         16.6   78      3    high
```

```
trainData$mpg_cat = as.factor(trainData$mpg_cat)
testData$mpg_cat = as.factor(testData$mpg_cat)
```

(a): Fit a support vector classifier to the training data. What are the training and test error rates?

```r
testData$mpg_cat_n <- as.numeric(testData$mpg_cat)
trainData$mpg_cat_n <- as.numeric(trainData$mpg_cat)

set.seed(1)
linear.tune <- tune.svm(mpg_cat ~ .,
                        data = trainData,
                        kernel = "linear",
                        cost = exp(seq(-5,2, len = 50)),
                        scale = TRUE)
plot(linear.tune)
```




```r
linear.tune$best.parameters
```

```
##          cost
## 1 0.006737947
```

```r
best.linear <- linear.tune$best.model
summary(best.linear)
```

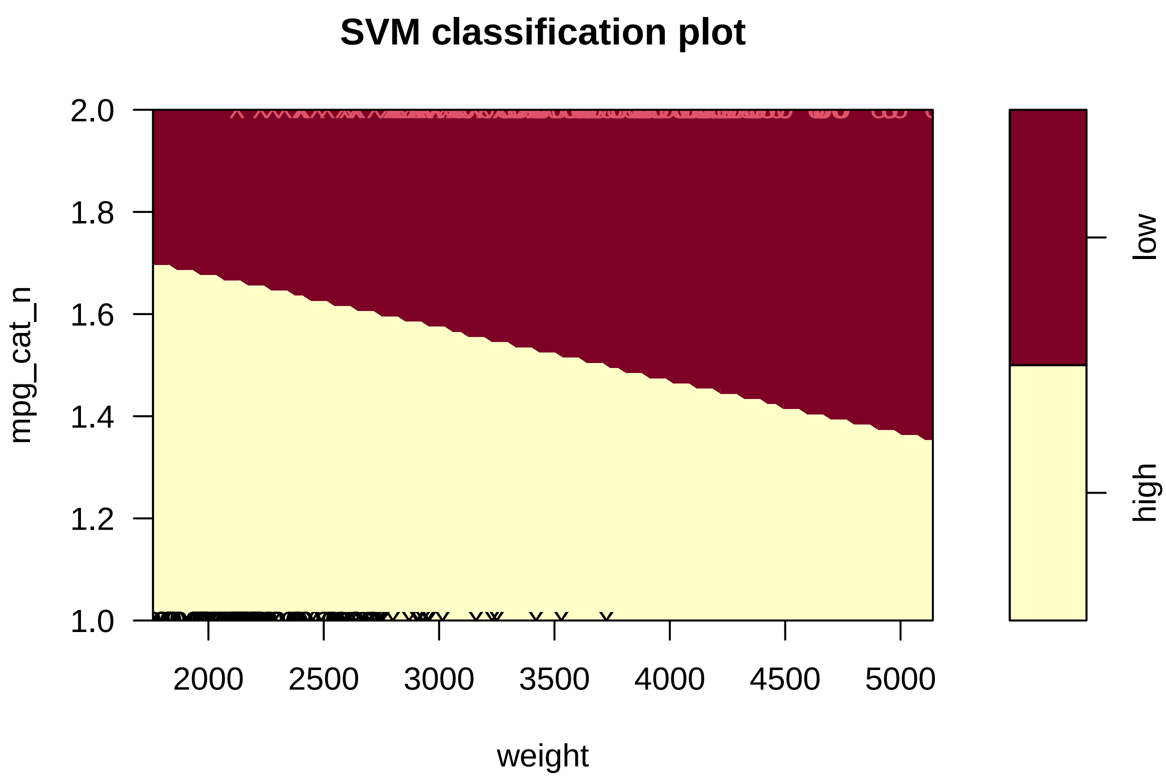
```
##
```

```
## Call:
## best.svm(x = mpg_cat ~ ., data = trainData, cost = exp(seq(-5, 2,
##     len = 50)), kernel = "linear", scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  0.006737947
##
## Number of Support Vectors:  108
##
##  ( 53 55 )
##
##
## Number of Classes:  2
##
## Levels:
##  high low
```

```r
pred.linear <- predict(best.linear, newdata = testData)
confusionMatrix(data = pred.linear,
reference = testData$mpg_cat)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction high low
##       high   59   0
##       low     0  59
##
##                Accuracy : 1
##                  95% CI : (0.9692, 1)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.0
##             Specificity : 1.0
##          Pos Pred Value : 1.0
##          Neg Pred Value : 1.0
##              Prevalence : 0.5
##          Detection Rate : 0.5
##    Detection Prevalence : 0.5
##       Balanced Accuracy : 1.0
##
##        'Positive' Class : high
##
```

```r
plot(best.linear, trainData,
      mpg_cat_n ~ weight,
slice = list(
       cylinders     = median(trainData$cylinders, na.rm = TRUE),
       displacement  = median(trainData$displacement, na.rm = TRUE),
       horsepower = median(trainData$horsepower, na.rm = TRUE),
       weight = median(trainData$weight, na.rm = TRUE),
       acceleration  = median(trainData$acceleration, na.rm = TRUE),
       year          = median(trainData$year, na.rm = TRUE),
       origin        = median(trainData$origin, na.rm = TRUE)),
      grid = 100)
```




```r
best.linear.model = linear.tune$best.model
test.pred = predict(best.linear.model, newdata = testData)
test.error = mean(test.pred != testData$mpg_cat)
cat("Test Data Error Rate:", test.error, "\n")
```

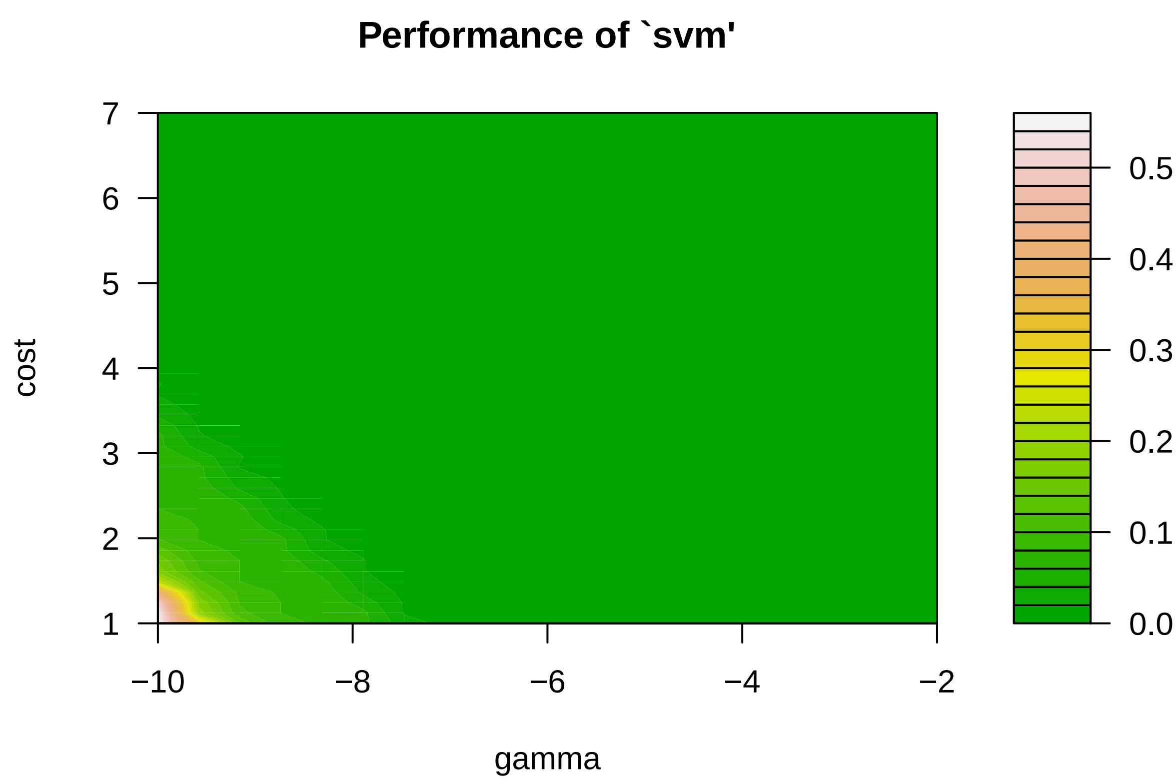
```
## Test Data Error Rate: 0
```

```r
train.pred = predict(best.linear.model, newdata = trainData)
train.error = mean(train.pred != trainData$mpg_cat)
cat("Training Data Error Rate:", train.error, "\n")
```

```
## Training Data Error Rate: 0
```

(b): Fit a support vector machine with a radial kernel to the training data. What are the training and test error rates?

```
set.seed(1)
radial.tune <- tune.svm(mpg_cat ~ . ,
                data = trainData,
                kernel = "radial",
                cost = exp(seq(1, 7, len = 50)),
                gamma = exp(seq(-10, -2,len = 20)))
plot(radial.tune, transform.y = log, transform.x = log,
color.palette = terrain.colors)
```




```
radial.tune$best.parameters
```

```
##         gamma     cost
## 9 0.00131808 2.718282
```

```
best.radial = radial.tune$best.model
summary(best.radial)
```

```
##
## Call:
## best.svm(x = mpg_cat ~ ., data = trainData, gamma = exp(seq(-10,
```

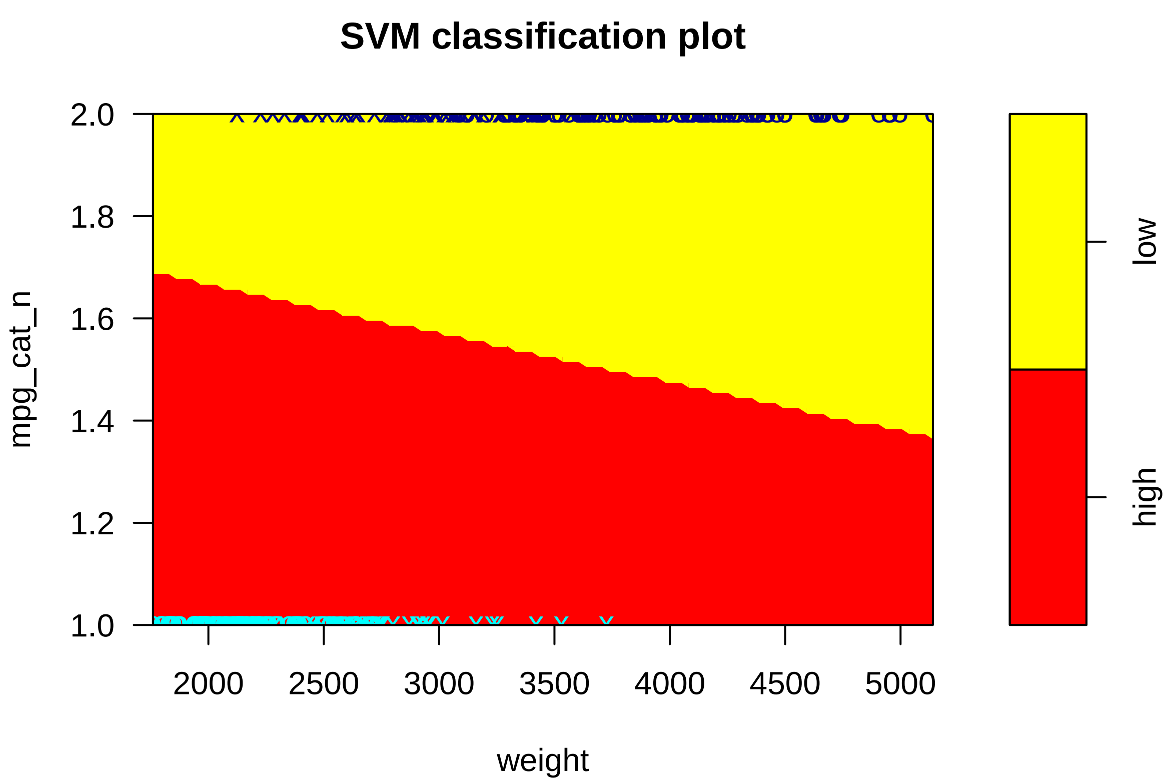
```
##     -2, len = 20)), cost = exp(seq(1, 7, len = 50)), kernel = "radial")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  2.718282
##
## Number of Support Vectors:  105
##
##  ( 52 53 )
##
##
## Number of Classes:  2
##
## Levels:
##  high low
```

```r
pred.radial = predict(best.radial, newdata = testData)
confusionMatrix(data = pred.radial,
reference = testData$mpg_cat)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction high low
##       high   59   0
##       low     0  59
##
##                Accuracy : 1
##                  95% CI : (0.9692, 1)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
##             Sensitivity : 1.0
##             Specificity : 1.0
##          Pos Pred Value : 1.0
##          Neg Pred Value : 1.0
##              Prevalence : 0.5
##          Detection Rate : 0.5
##    Detection Prevalence : 0.5
##       Balanced Accuracy : 1.0
##
##        'Positive' Class : high
##
```

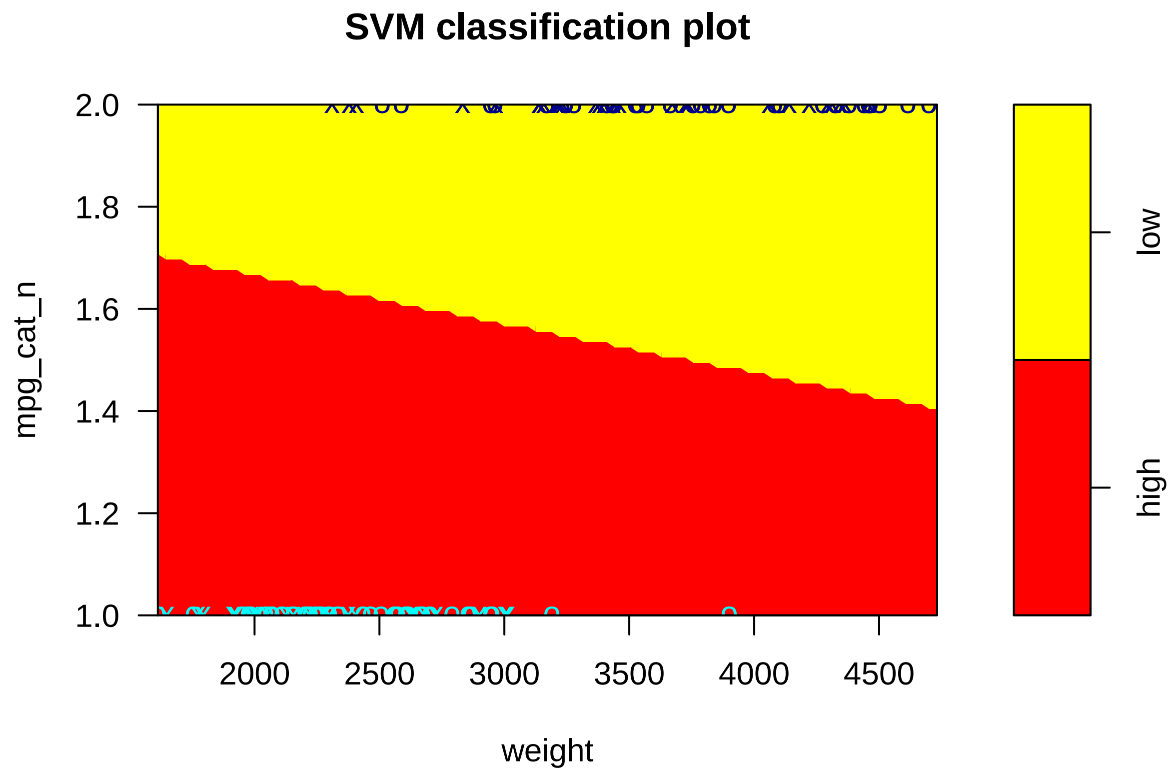
```r
plot(best.radial, trainData,
    mpg_cat_n ~ weight,
slice = list(
```

```
        cylinders     = median(trainData$cylinders, na.rm = TRUE),
        displacement  = median(trainData$displacement, na.rm = TRUE),
        horsepower = median(trainData$horsepower, na.rm = TRUE),
        weight = median(trainData$weight, na.rm = TRUE),
        acceleration  = median(trainData$acceleration, na.rm = TRUE),
        year          = median(trainData$year, na.rm = TRUE),
        origin        = median(trainData$origin, na.rm = TRUE)
      ),
    grid = 100,
    symbolPalette = c("cyan","darkblue"),
    color.palette = heat.colors)
```




```
plot(best.radial, testData,
    mpg_cat_n ~ weight,
slice = list(
        cylinders     = median(trainData$cylinders, na.rm = TRUE),
        displacement  = median(trainData$displacement, na.rm = TRUE),
        horsepower = median(trainData$horsepower, na.rm = TRUE),
        weight = median(trainData$weight, na.rm = TRUE),
        acceleration  = median(trainData$acceleration, na.rm = TRUE),
        year          = median(trainData$year, na.rm = TRUE),
        origin        = median(trainData$origin, na.rm = TRUE)
      ),
    grid = 100,
    symbolPalette = c("cyan","darkblue"),
```

```
    color.palette = heat.colors)
```



Question 2: In this problem, we perform hierarchical clustering on the states using the USArrests data in the ISLR package. For each of the 50 states in the United States, the dataset contains the number of arrests per 100,000 residents for each of three crimes: Assault, Murder, and Rape. The dataset also contains the percent of the population in each state living in urban areas, UrbanPop. The four variables will be used as features for clustering.

```
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```r
library(RColorBrewer)
library(gplots)
```

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:plotrix':
##
##     plotCI

## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(jpeg)

data("USArrests")
str(USArrests)
```

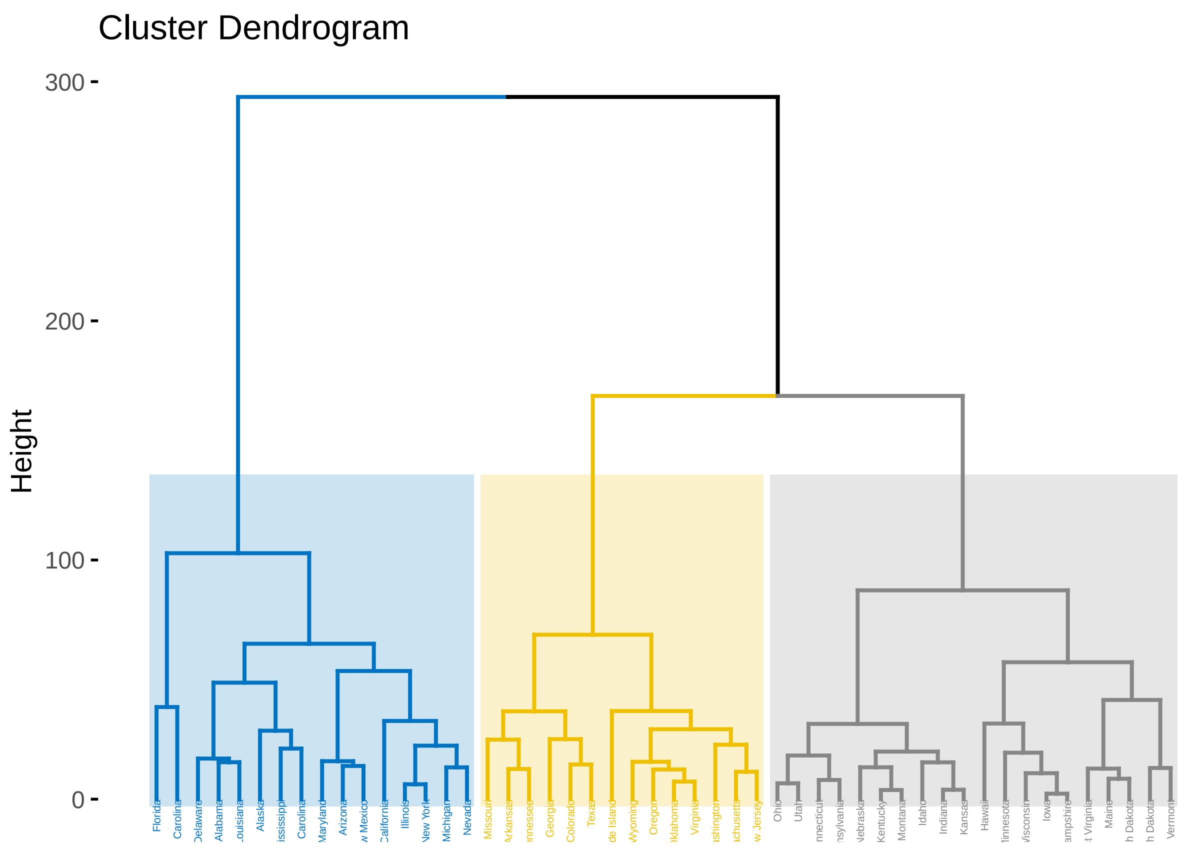
```
## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

(a): Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

```r
hc.complete <- hclust(dist(USArrests), method = "complete")
hc.average <- hclust(dist(USArrests), method = "average")
hc.single <- hclust(dist(USArrests), method = "single")
hc.centroid <- hclust(dist(USArrests), method = "centroid")
```

```r
fviz_dend(hc.complete, k = 3,
cex = 0.3,
palette = "jco", # color scheme; other palettes:"npg","aaas"...
color_labels_by_k = TRUE,
rect = TRUE, # whether to add a rectangle around groups.
rect_fill = TRUE,
rect_border = "jco",
labels_track_height = 2.5)
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
ind4.complete <- cutree(hc.complete, 3)
USArrests[ind4.complete == 1,]
```

```
##                Murder Assault UrbanPop Rape
## Alabama          13.2     236       58 21.2
## Alaska           10.0     263       48 44.5
## Arizona           8.1     294       80 31.0
## California        9.0     276       91 40.6
## Delaware          5.9     238       72 15.8
## Florida          15.4     335       80 31.9
## Illinois         10.4     249       83 24.0
## Louisiana        15.4     249       66 22.2
## Maryland         11.3     300       67 27.8
## Michigan         12.1     255       74 35.1
## Mississippi      16.1     259       44 17.1
## Nevada           12.2     252       81 46.0
## New Mexico       11.4     285       70 32.1
```

```
## New York         11.1     254       86 26.1
## North Carolina   13.0     337       45 16.1
## South Carolina   14.4     279       48 22.5
```

```
USArrests[ind4.complete == 2,]
```

```
##                Murder Assault UrbanPop Rape
## Arkansas          8.8     190       50 19.5
## Colorado          7.9     204       78 38.7
## Georgia          17.4     211       60 25.8
## Massachusetts     4.4     149       85 16.3
## Missouri          9.0     178       70 28.2
## New Jersey        7.4     159       89 18.8
## Oklahoma          6.6     151       68 20.0
## Oregon            4.9     159       67 29.3
## Rhode Island      3.4     174       87  8.3
## Tennessee        13.2     188       59 26.9
## Texas            12.7     201       80 25.5
## Virginia          8.5     156       63 20.7
## Washington        4.0     145       73 26.2
## Wyoming           6.8     161       60 15.6
```

```
USArrests[ind4.complete == 3,]
```

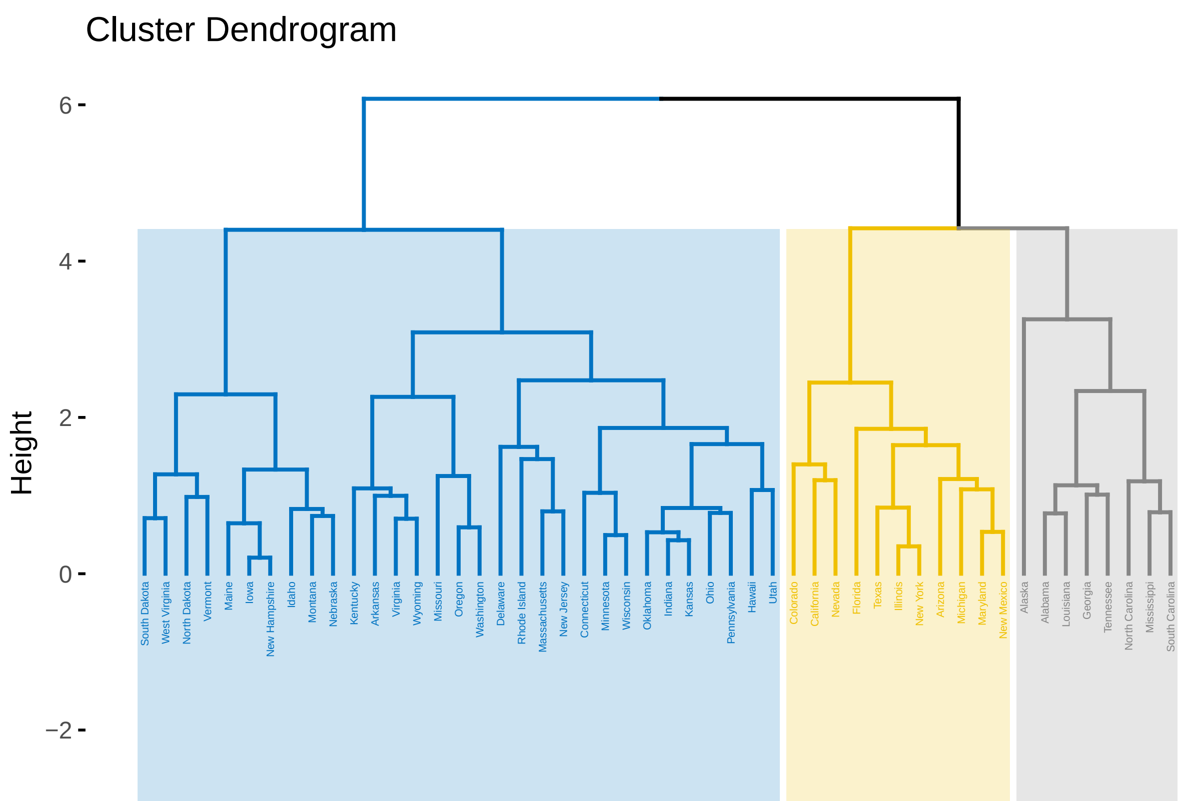
```
##                Murder Assault UrbanPop Rape
## Connecticut       3.3     110       77 11.1
## Hawaii            5.3      46       83 20.2
## Idaho             2.6     120       54 14.2
## Indiana           7.2     113       65 21.0
## Iowa              2.2      56       57 11.3
## Kansas            6.0     115       66 18.0
## Kentucky          9.7     109       52 16.3
## Maine             2.1      83       51  7.8
## Minnesota         2.7      72       66 14.9
## Montana           6.0     109       53 16.4
## Nebraska          4.3     102       62 16.5
## New Hampshire     2.1      57       56  9.5
## North Dakota      0.8      45       44  7.3
## Ohio              7.3     120       75 21.4
## Pennsylvania      6.3     106       72 14.9
## South Dakota      3.8      86       45 12.8
## Utah              3.2     120       80 22.9
## Vermont           2.2      48       32 11.2
## West Virginia     5.7      81       39  9.3
## Wisconsin         2.6      53       66 10.8
```

(b): Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one. Does scaling the variables change the clustering results? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed?

Yes, clustering with scaling affects the number of states in each cluster-there are more states is cluster 3 when scaling compared to when there is not scaling. Yes, the variables be scaled before the inter-observation dissimilarities are computed.

```r
USArrests.scaled <- scale(USArrests)
hc.complete.scaled <- hclust(dist(USArrests.scaled), method = "complete")

fviz_dend(hc.complete.scaled, k = 3,
          cex = 0.3,
          palette = "jco",
          color_labels_by_k = TRUE,
          rect = TRUE,
          rect_fill = TRUE,
          rect_border = "jco",
          labels_track_height = 2.5)
```




```r
ind4.scaled <- cutree(hc.complete.scaled, 3)

USArrests[ind4.scaled == 1, ]
```

```
##                Murder Assault UrbanPop Rape
## Alabama          13.2     236       58 21.2
## Alaska           10.0     263       48 44.5
## Georgia          17.4     211       60 25.8
## Louisiana        15.4     249       66 22.2
## Mississippi      16.1     259       44 17.1
## North Carolina   13.0     337       45 16.1
```

```
## South Carolina   14.4     279       48 22.5
## Tennessee        13.2     188       59 26.9
```

```
USArrests[ind4.scaled == 2, ]
```

```
##            Murder Assault UrbanPop Rape
## Arizona       8.1     294       80 31.0
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
## Florida      15.4     335       80 31.9
## Illinois     10.4     249       83 24.0
## Maryland     11.3     300       67 27.8
## Michigan     12.1     255       74 35.1
## Nevada       12.2     252       81 46.0
## New Mexico   11.4     285       70 32.1
## New York     11.1     254       86 26.1
## Texas        12.7     201       80 25.5
```

```
USArrests[ind4.scaled == 3, ]
```

```
##               Murder Assault UrbanPop Rape
## Arkansas         8.8     190       50 19.5
## Connecticut      3.3     110       77 11.1
## Delaware         5.9     238       72 15.8
## Hawaii           5.3      46       83 20.2
## Idaho            2.6     120       54 14.2
## Indiana          7.2     113       65 21.0
## Iowa             2.2      56       57 11.3
## Kansas           6.0     115       66 18.0
## Kentucky         9.7     109       52 16.3
## Maine            2.1      83       51  7.8
## Massachusetts    4.4     149       85 16.3
## Minnesota        2.7      72       66 14.9
## Missouri         9.0     178       70 28.2
## Montana          6.0     109       53 16.4
## Nebraska         4.3     102       62 16.5
## New Hampshire    2.1      57       56  9.5
## New Jersey       7.4     159       89 18.8
## North Dakota     0.8      45       44  7.3
## Ohio             7.3     120       75 21.4
## Oklahoma         6.6     151       68 20.0
## Oregon           4.9     159       67 29.3
## Pennsylvania     6.3     106       72 14.9
## Rhode Island     3.4     174       87  8.3
## South Dakota     3.8      86       45 12.8
## Utah             3.2     120       80 22.9
## Vermont          2.2      48       32 11.2
## Virginia         8.5     156       63 20.7
## Washington       4.0     145       73 26.2
## West Virginia    5.7      81       39  9.3
## Wisconsin        2.6      53       66 10.8
## Wyoming          6.8     161       60 15.6
```