Homework 5

Yiying Wu (yw3996)

R packages

```
library(tidyverse)
library(caret)
library(tidymodels)
library(ISLR)
library(e1071)
library(factoextra)
```

1. auto.csv data

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%).

Input dataset

```
dat<-read_csv("./data/auto.csv")%>%
  mutate(
    mpg_cat = as.factor(mpg_cat),
    origin = as.factor(origin))
dat <- dat%>%
  na.omit()
```

Response: mpg_cat

```
contrasts(dat$mpg_cat)
```

```
##      low
## high   0
## low    1
```

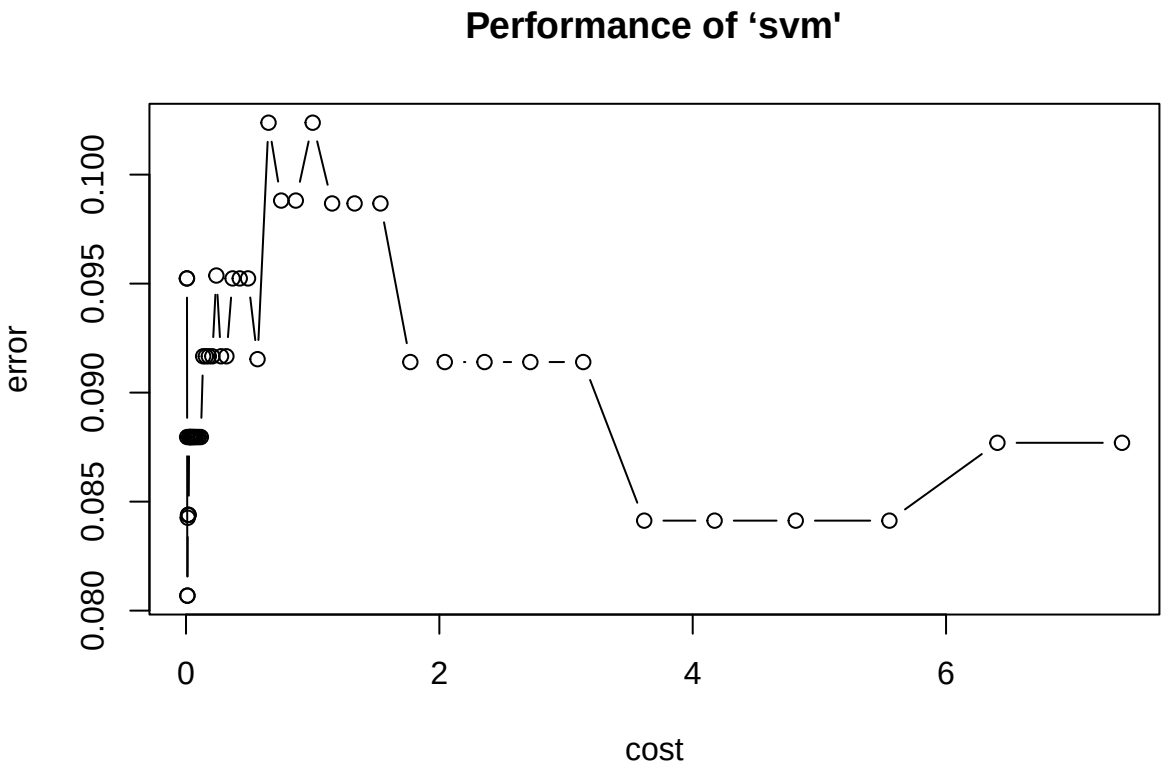
Split the dataset into two parts: training data (70%) and test data (30%).

```
set.seed(1)
data_split <- initial_split(dat, prop = 0.7)

# Extract the training and test data
training_data <- training(data_split)
```

```
testing_data <- testing(data_split)

ctrl <- trainControl(method = "cv", number = 10,
                     summaryFunction = twoClassSummary,
                     classProbs = TRUE)
```

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

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




```
# summary(linear.tune)
linear.tune$best.parameters
```

```
##         cost
## 5 0.01193152
```

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

```
##
## Call:
## best.svm(x = mpg_cat ~ ., data = training_data, cost = exp(seq(-5,
##     2, len = 50)), kernel = "linear", scale = TRUE)
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  0.01193152
##
## Number of Support Vectors:  123
##
##  ( 62 61 )
##
##
## Number of Classes:  2
##
## Levels:
##  high low
```

```
pred.linear <- predict(best.linear, newdata = testing_data)

confusionMatrix(data = pred.linear,
                reference = testing_data$mpg_cat)
```

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction high low
##       high   56  10
##       low     1  51
##
##                Accuracy : 0.9068
##                  95% CI : (0.8393, 0.9525)
##     No Information Rate : 0.5169
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8143
##
##  Mcnemar's Test P-Value : 0.01586
##
##             Sensitivity : 0.9825
##             Specificity : 0.8361
##          Pos Pred Value : 0.8485
##          Neg Pred Value : 0.9808
##              Prevalence : 0.4831
##          Detection Rate : 0.4746
##    Detection Prevalence : 0.5593
##       Balanced Accuracy : 0.9093
##
##        'Positive' Class : high
##
```

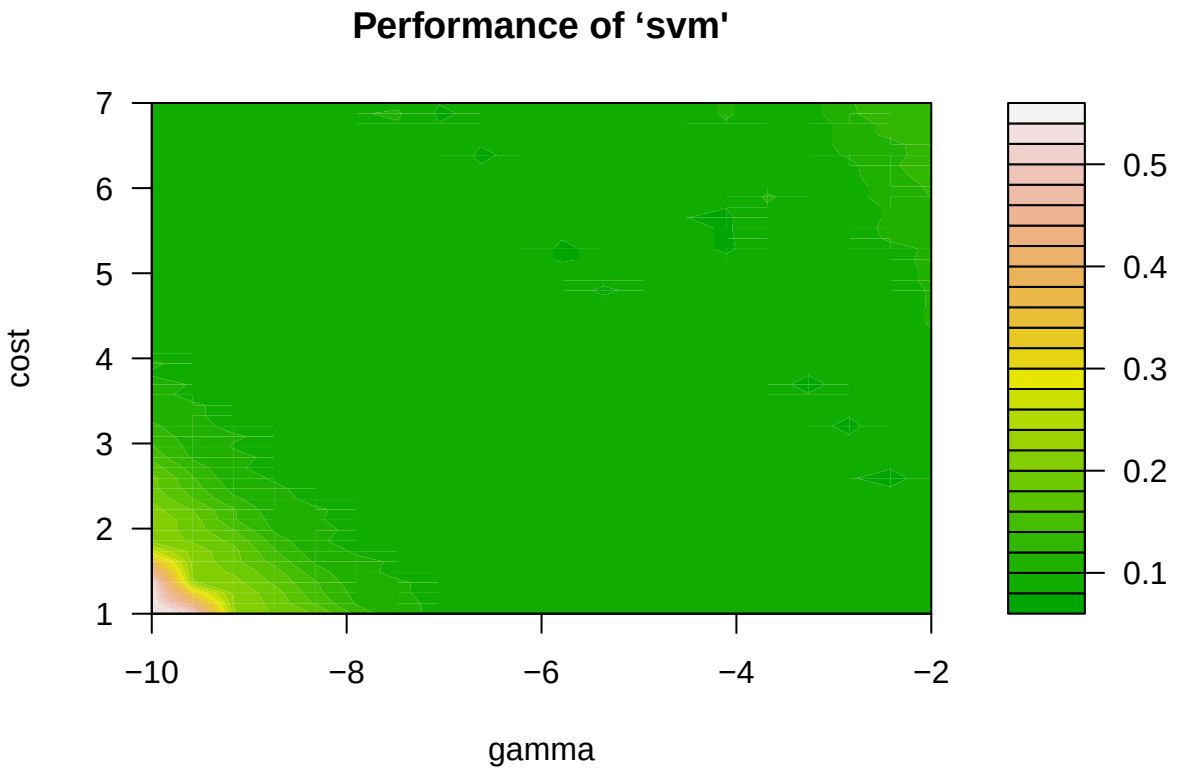
The **training error rate** is 0.0119.

Test Error Rate $= 1 - Accuracy = 1 - 0.9068 = 0.0932$

The **test error rate** for the model on the testing data is approximately 0.0932.

(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 = training_data,
                        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)
```




```
# summary(radial.tune)

radial.tune$best.parameters
```

```
##          gamma     cost
## 715 0.01648568 197.4952
```

```
best.radial <- radial.tune$best.model
summary(best.radial)
```

```
##
## Call:
## best.svm(x = mpg_cat ~ ., data = training_data, 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:  197.4952
##
## Number of Support Vectors:  63
##
##  ( 32 31 )
##
##
## Number of Classes:  2
##
## Levels:
##  high low
```

```
# Predict on the training data using the best model
pred.radial.train <- predict(best.radial, newdata = training_data)

# Calculate the confusion matrix for the training predictions
conf.matrix.train <- confusionMatrix(data = pred.radial.train,
                                     reference = training_data$mpg_cat)

# Extract and print the training error rate
train.error.rate <- 1 - conf.matrix.train$overall['Accuracy']
print(train.error.rate)
```

```
##  Accuracy
## 0.0620438
```

```
pred.radial <- predict(best.radial, newdata = testing_data)

confusionMatrix(data = pred.radial,
                reference = testing_data$mpg_cat)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction high low
##       high   56   9
##       low     1  52
##
##                Accuracy : 0.9153
##                  95% CI : (0.8497, 0.9586)
```

```
##     No Information Rate : 0.5169
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8311
##
##  Mcnemar's Test P-Value : 0.02686
##
##             Sensitivity : 0.9825
##             Specificity : 0.8525
##          Pos Pred Value : 0.8615
##          Neg Pred Value : 0.9811
##              Prevalence : 0.4831
##          Detection Rate : 0.4746
##    Detection Prevalence : 0.5508
##       Balanced Accuracy : 0.9175
##
##        'Positive' Class : high
##
```

The **training error rate** is 0.062.

Test Error Rate $= 1 - Accuracy = 1 - 0.9153 = 0.0847$

The **test error rate** for the model on the testing data is approximately 0.0847.

2. USArrests data

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.

```
data(USArrests)
dat2 <- na.omit(USArrests)
set.seed(1)
```

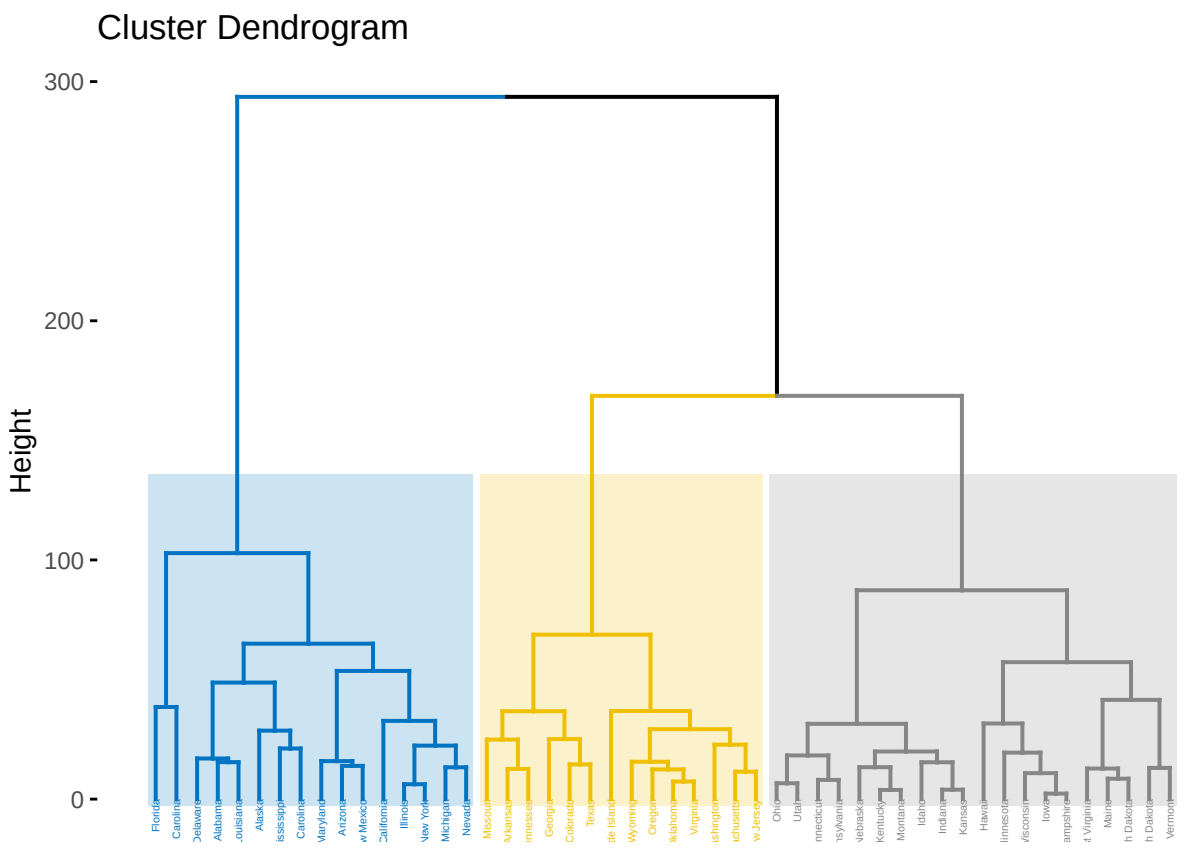
(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states. Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

Use hierarchical clustering with complete linkage and Euclidean distance, cluster the states

```
hc.complete <- hclust(dist(dat2), method = "complete")
```

Visualize the dendrogram and cut the dendrogram at a height that results in three distinct clusters

```
fviz_dend(hc.complete, 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)
```




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

Which states belong to which clusters

```
# States in the first cluster
dat2[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
```

```
# States in the second cluster
dat2[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
```

```
# States in the third cluster
dat2[ind4.complete == 3,]
```

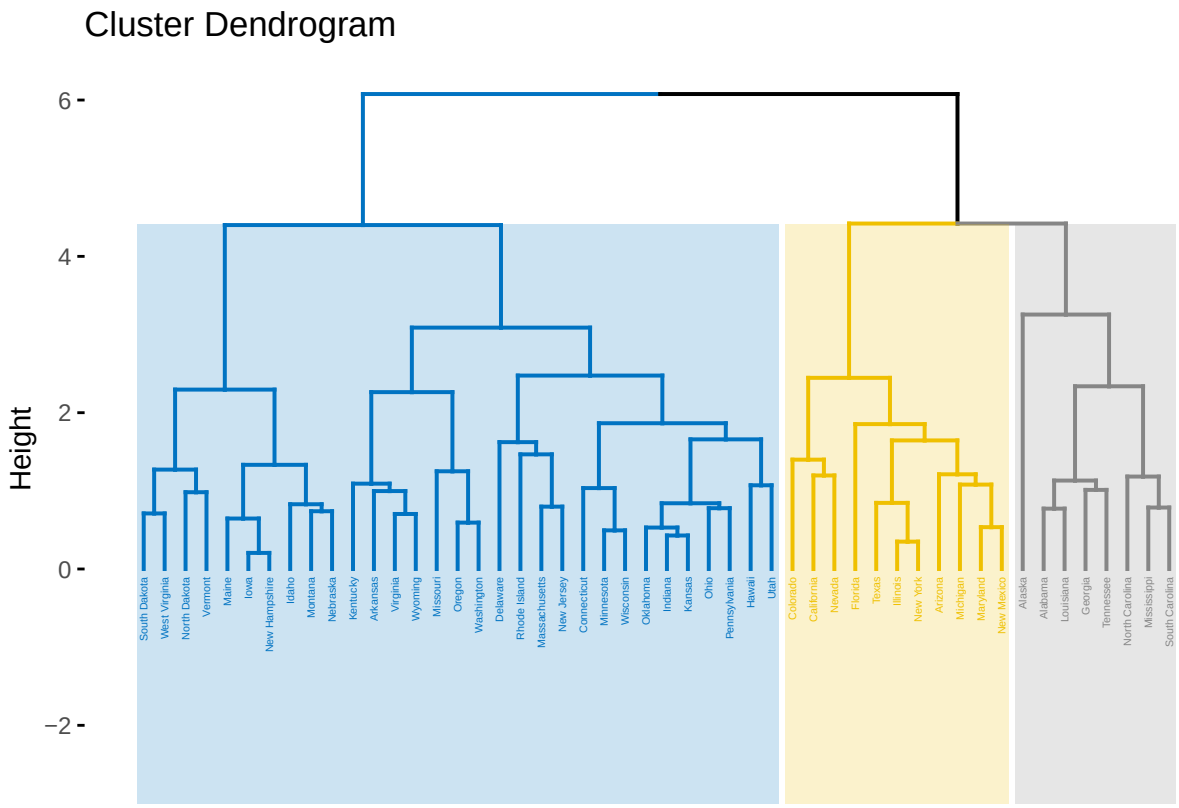
```
##               Murder Assault UrbanPop Rape
## Connecticut      3.3     110       77 11.1
## Hawaii           5.3      46       83 20.2
## Idaho            2.6     120       54 14.2
## Indiana          7.2     113       65 21.0
## Iowa             2.2      56       57 11.3
## Kansas           6.0     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.

```
dat2_scale <- scale(dat2)
hc.complete_scale <- hclust(dist(dat2_scale), method = "complete")
```

Visualize the dendrogram and cut the dendrogram at a height that results in three distinct clusters

```
fviz_dend(hc.complete_scale, 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)
```




```
ind4.complete_scale <- cutree(hc.complete_scale, 3)
```

Which states belong to which clusters

```
# States in the first cluster
dat2_scale[ind4.complete_scale == 1,]
```

```
##                  Murder   Assault    UrbanPop         Rape
## Alabama        1.2425641 0.7828393 -0.52090661 -0.003416473
## Alaska         0.5078625 1.1068225 -1.21176419  2.484202941
## Georgia        2.2068599 0.4828549 -0.38273510  0.487701523
## Louisiana      1.7476714 0.9388312  0.03177945  0.103348309
## Mississippi    1.9083874 1.0588250 -1.48810723 -0.441152078
## North Carolina 1.1966452 1.9947764 -1.41902147 -0.547916860
## South Carolina 1.5180772 1.2988126 -1.21176419  0.135377743
## Tennessee      1.2425641 0.2068693 -0.45182086  0.605142783
```

```r
# States in the second cluster
dat2_scale[ind4.complete_scale == 2,]
```

```
##                Murder   Assault  UrbanPop      Rape
## Arizona    0.07163341 1.4788032 0.9989801 1.0428784
## California 0.27826823 1.2628144 1.7589234 2.0678203
## Colorado   0.02571456 0.3988593 0.8608085 1.8649672
## Florida    1.74767144 1.9707777 0.9989801 1.1389667
## Illinois   0.59970018 0.9388312 1.2062373 0.2955249
## Maryland   0.80633501 1.5507995 0.1008652 0.7012311
## Michigan   0.99001041 1.0108275 0.5844655 1.4806140
## Nevada     1.01296983 0.9748294 1.0680658 2.6443501
## New Mexico 0.82929443 1.3708088 0.3081225 1.1603196
## New York   0.76041616 0.9988281 1.4134946 0.5197310
## Texas      1.12776696 0.3628612 0.9989801 0.4556721
```

```r
# States in the third cluster
dat2_scale[ind4.complete_scale == 3,]
```

```
##                   Murder     Assault    UrbanPop        Rape
## Arkansas       0.23234938  0.23086801 -1.07359268 -0.18491660
## Connecticut   -1.03041900 -0.72908214  0.79172279 -1.08174077
## Delaware      -0.43347395  0.80683810  0.44629400 -0.57994629
## Hawaii        -0.57123050 -1.49704226  1.20623733 -0.11018125
## Idaho         -1.19113497 -0.60908837 -0.79724965 -0.75076995
## Indiana       -0.13500142 -0.69308401 -0.03730631 -0.02476943
## Iowa          -1.28297267 -1.37704849 -0.58999237 -1.06038781
## Kansas        -0.41051452 -0.66908525  0.03177945 -0.34506377
## Kentucky       0.43898421 -0.74108152 -0.93542116 -0.52656390
## Maine         -1.30593210 -1.05306531 -1.00450692 -1.43406455
## Massachusetts -0.77786532 -0.26110644  1.34440885 -0.52656390
## Minnesota     -1.16817555 -1.18505846  0.03177945 -0.67603460
## Missouri       0.27826823  0.08687549  0.30812248  0.74393700
## Montana       -0.41051452 -0.74108152 -0.86633540 -0.51588743
## Nebraska      -0.80082475 -0.82507715 -0.24456358 -0.50521095
## New Hampshire -1.30593210 -1.36504911 -0.65907813 -1.25256442
## New Jersey    -0.08908257 -0.14111267  1.62075188 -0.25965195
## North Dakota  -1.60440462 -1.50904164 -1.48810723 -1.48744694
## Ohio          -0.11204199 -0.60908837  0.65355127  0.01793648
## Oklahoma      -0.27275797 -0.23710769  0.16995096 -0.13153421
## Oregon        -0.66306820 -0.14111267  0.10086521  0.86137826
## Pennsylvania  -0.34163624 -0.77707965  0.44629400 -0.67603460
## Rhode Island  -1.00745957  0.03887798  1.48258036 -1.38068216
## South Dakota  -0.91562187 -1.01706718 -1.41902147 -0.90024064
## Utah          -1.05337842 -0.60908837  0.99898006  0.17808366
## Vermont       -1.28297267 -1.47304350 -2.31713632 -1.07106429
## Virginia       0.16347111 -0.17711080 -0.17547783 -0.05679886
## Washington    -0.86970302 -0.30910395  0.51537975  0.53040744
## West Virginia -0.47939280 -1.07706407 -1.83353601 -1.27391738
## Wisconsin     -1.19113497 -1.41304662  0.03177945 -1.11377020
## Wyoming       -0.22683912 -0.11711392 -0.38273510 -0.60129925
```

(c) Does scaling the variables change the clustering results? Why? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed?

Yes, scaling the variables affects the clustering outcomes.

This is because when variables are on different scales, particularly in methods like hierarchical clustering that depend on distance calculations, variables with larger scales (eg. Assault) disproportionately influence the results.

I think the variables should be scaled before the inter-observation dissimilarities are computed to ensure that no single variable influences the outcome purely because of its scale.