Homework 4

name1 and name2 and name3

2017-12-01

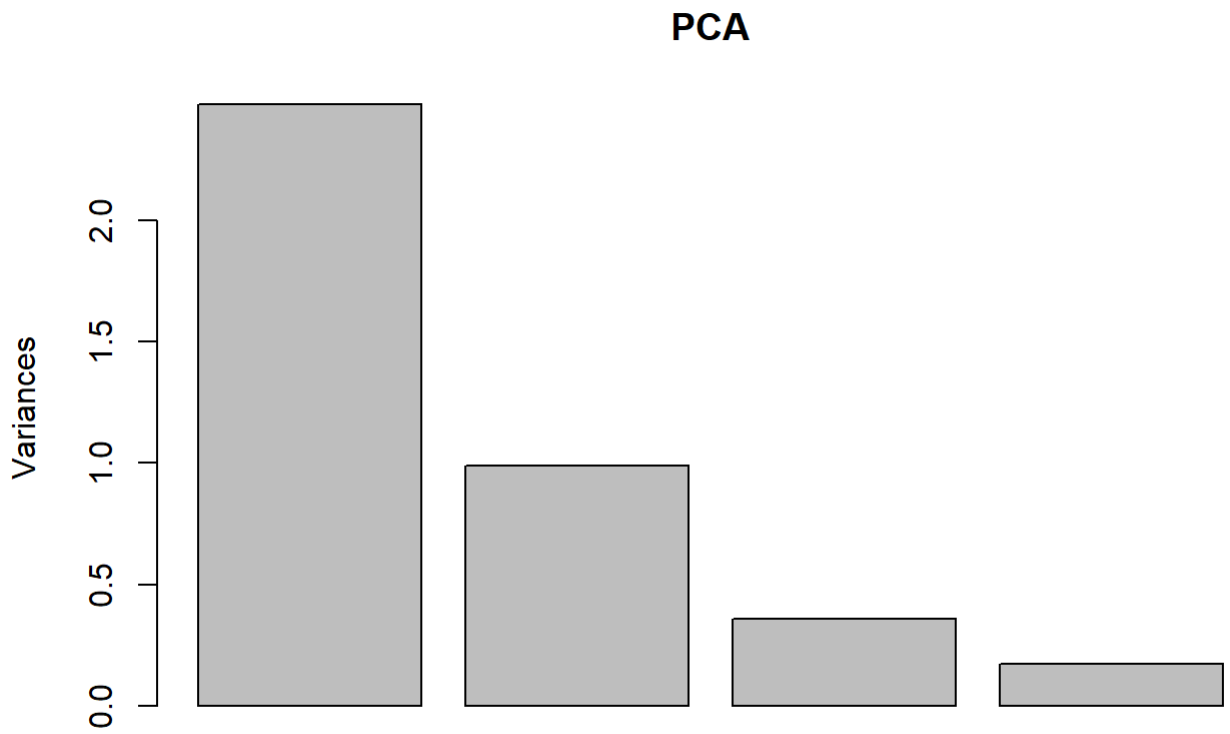
Question 1 (Chapter 10, #8, 4 marks)

a. (1 mark)

```
library(ISLR)
data("USArrests")
dim(USArrests)
```

```
## [1] 50  4
```

```
PCA = prcomp(USArrests, scale = TRUE)
screeplot(PCA)
```



```
# Results don't match findings in textbook on page 397/440, check
# Matches Brad McNeney's findings
prop = (PCA$sdev)^2 / sum((PCA$sdev)^2) * 100
prop
```

```
## [1] 62.006039 24.744129  8.914080  4.335752
```

b. (3 marks)

```
# Calculate PVE using the formula with loadings
# Each column of the rotation is phi1, phi2, etc
comp = matrix(nrow = dim(USArrests)[1], ncol = 4)
data = scale(USArrests)

res = sum(data^2)
for(j in 1:dim(data)[1]){
  for(i in 1:4){
    #print(j )
    #print(i)
    #print(PCA$rotation[,i])
    #print(USArrests[j,])
    comp[j, i] = sum(PCA$rotation[,i]*data[j,])^2

  }
}
#comp
PCA$x[, 1]
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##    -0.97566045    -1.93053788    -1.74544285     0.13999894    -2.49861285
##       Colorado    Connecticut       Delaware        Florida        Georgia
##    -1.49934074     1.34499236    -0.04722981    -2.98275967    -1.62280742
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##     0.90348448     1.62331903    -1.36505197     0.50038122     2.23099579
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##     0.78887206     0.74331256    -1.54909076     2.37274014    -1.74564663
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##     0.48128007    -2.08725025     1.67566951    -0.98647919    -0.68978426
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##     1.17353751     1.25291625    -2.84550542     2.35995585    -0.17974128
##     New Mexico       New York North Carolina   North Dakota           Ohio
##    -1.96012351    -1.66566662    -1.11208808     2.96215223     0.22369436
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##     0.30864928    -0.05852787     0.87948680     0.85509072    -1.30744986
##   South Dakota      Tennessee          Texas           Utah        Vermont
##     1.96779669    -0.98969377    -1.34151838     0.54503180     2.77325613
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##     0.09536670     0.21472339     2.08739306     2.05881199     0.62310061
```

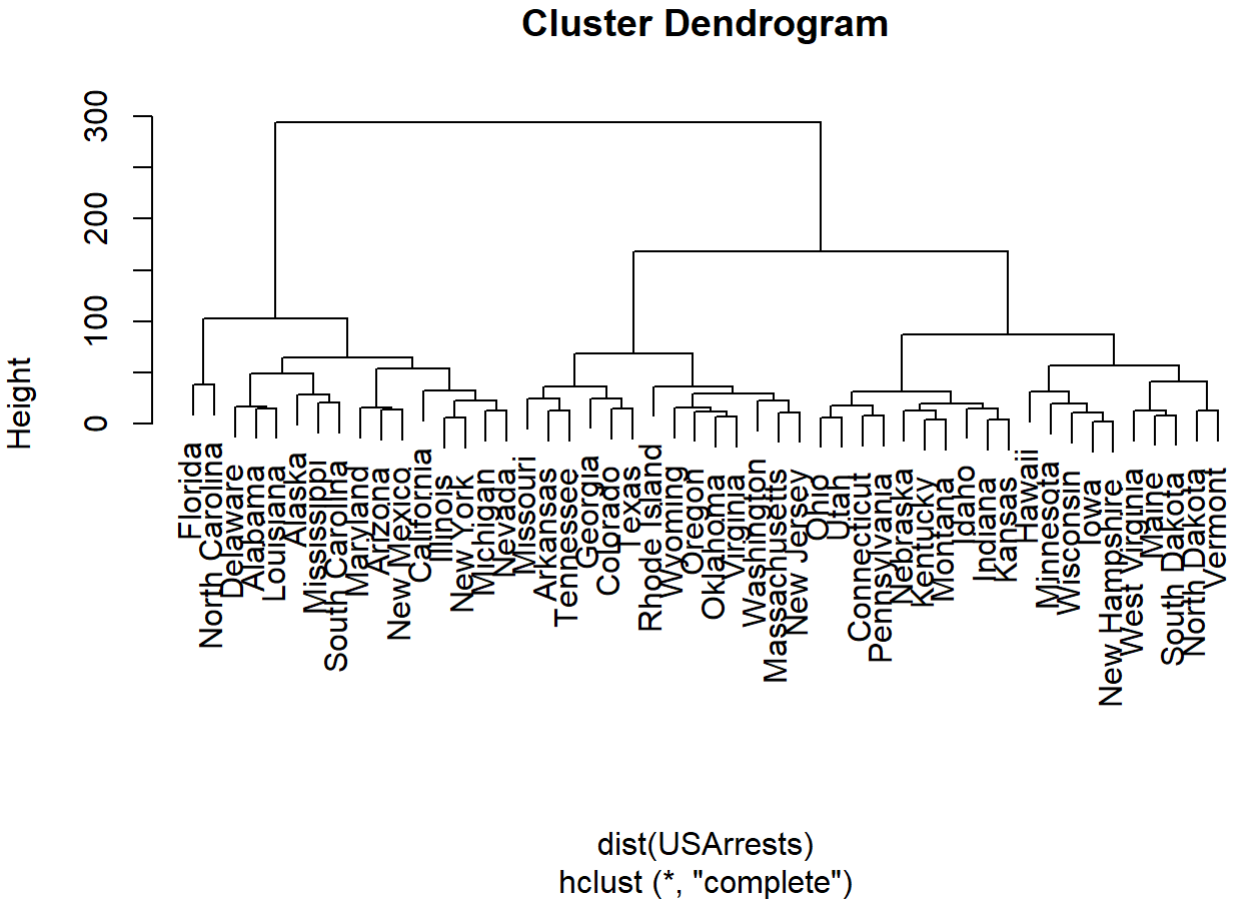
```
firstComp = sum(comp[, 1]) / res *100
secondComp= sum(comp[, 2]) / res *100
thirdComp = sum(comp[, 3]) / res *100
fourthComp= sum(comp[, 4]) / res *100

print(c(firstComp, secondComp, thirdComp, fourthComp))
```

```
## [1] 62.006039 24.744129  8.914080  4.335752
```

Question 2 (Chapter 10, #9, 7 marks)

a. (1 mark)

```
data(USArrests)
hc.complete = hclust(dist(USArrests), method = "complete")
plot(hc.complete) # Dendogram
```



b. (2 marks)

```
groups = cutree(hc.complete, 3) # Specify number of clusters I am interested in, identifies wher
e the observations go

groups[groups == 1]
```

```
##        Alabama         Alaska        Arizona     California       Delaware
##              1              1              1              1              1
##        Florida       Illinois      Louisiana       Maryland       Michigan
##              1              1              1              1              1
##    Mississippi         Nevada     New Mexico       New York North Carolina
##              1              1              1              1              1
## South Carolina
##              1
```

```
groups[groups == 2]
```

```
##      Arkansas      Colorado       Georgia Massachusetts      Missouri
##             2             2             2             2             2
##    New Jersey      Oklahoma        Oregon  Rhode Island     Tennessee
##             2             2             2             2             2
##         Texas      Virginia    Washington       Wyoming
##             2             2             2             2
```

```
groups[groups == 3]
```

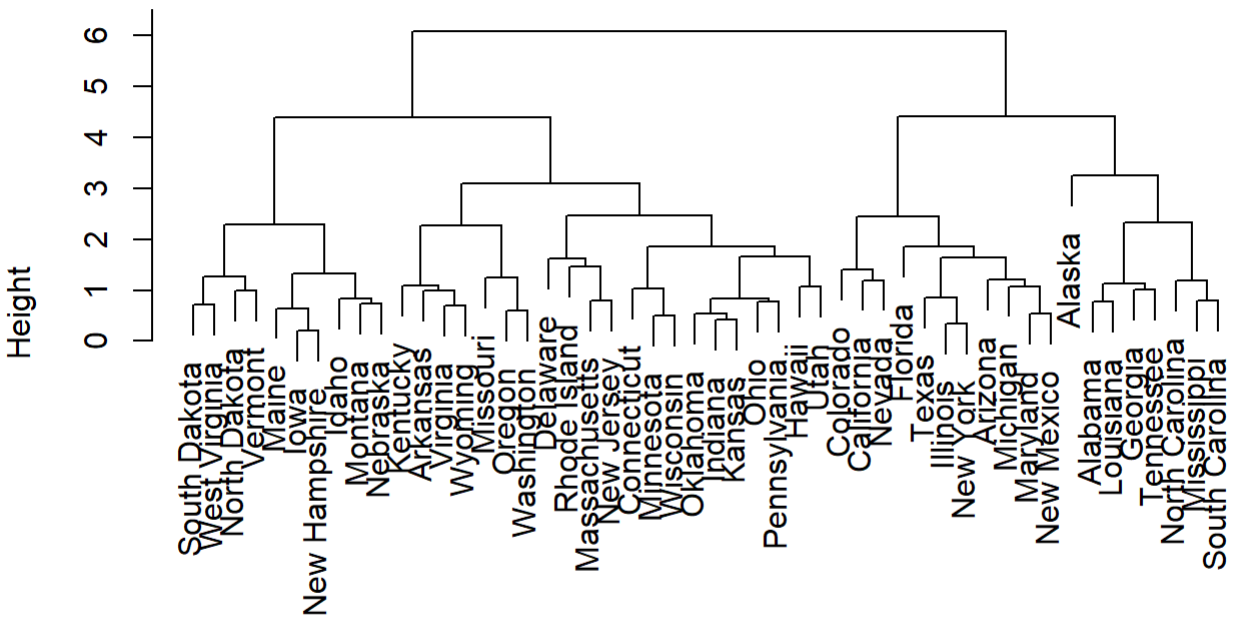
```
##   Connecticut        Hawaii         Idaho       Indiana          Iowa
##             3             3             3             3             3
##        Kansas      Kentucky         Maine     Minnesota       Montana
##             3             3             3             3             3
##      Nebraska New Hampshire  North Dakota          Ohio  Pennsylvania
##             3             3             3             3             3
##  South Dakota          Utah       Vermont West Virginia     Wisconsin
##             3             3             3             3             3
```

c. (2 marks)

```
USArrestsScale = scale(USArrests)
hc.complete.scale = hclust(dist(USArrestsScale), method = "complete")
```

d. (2 marks)

```
plot(hc.complete.scale)
```




```
groupsScale = cutree(hc.complete.scale, 3)
identical = sum(groups == groupsScale)
identical
```

```
## [1] 28
```

The results of the two trees differ in almost 50% of assignments (28 of the states received the same cluster assignment). Part of this may be due to the labelling of groups so we will check for similarities between each group.

```
holder = matrix(nrow = 3, ncol = 3)
for(i in 1:3){
  for(j in 1:3){
   holder[i,j] = length(setdiff(names(groupsScale[groupsScale == i]),
                 names(groups[groups == j])))
  }
}
holder
```

```
##      [,1] [,2] [,3]
## [1,]    2    6    8
## [2,]    2    9   11
## [3,]   30   21   11
```

Using the results of the Holder matrix above, we see that the groups were likely adequately labelled as 1 is most similar with 1, 2 is most similar with 2 (as 1 is already paired), and 3 is most similar with 3. We can now proceed with the notion that 28 of the states are clustered in the same group with or without scaling and that 22 of them are not. This represents a 44% different classification rate. These different classifications are occurring because of the scaling. Reasons behind this are the fact that Euclidean distance is NOT scale invariant. Consider the Euclidean distance between the two following measurements, done in different scales. Distance between 6ft2in and 6ft = 2in = 4 Euclidean distance Distance between 1879.6mm and 1828.8mm = 50.8 = 2580.64 Euclidean distance Both sets of measurements consider the same two people. The first uses the American standard of feet and inches while the latter uses the metric system in millimetres. Note that although the people are equally different in height, the Euclidean distance of the second example is significantly larger. By changing the units, we have artificially changed the Euclidean distance. However, by using a scaling process, we can remove the influence of these unit differences. (x - mean(x)) / sd(x) allows for identical distances to be calculated across units. For this reason we should be using scaling processes before doing Hierarchical clustering.