Week 8 Assignment

David Russo

3/9/2017

```
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(ISLR)
library(maps)
```

```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

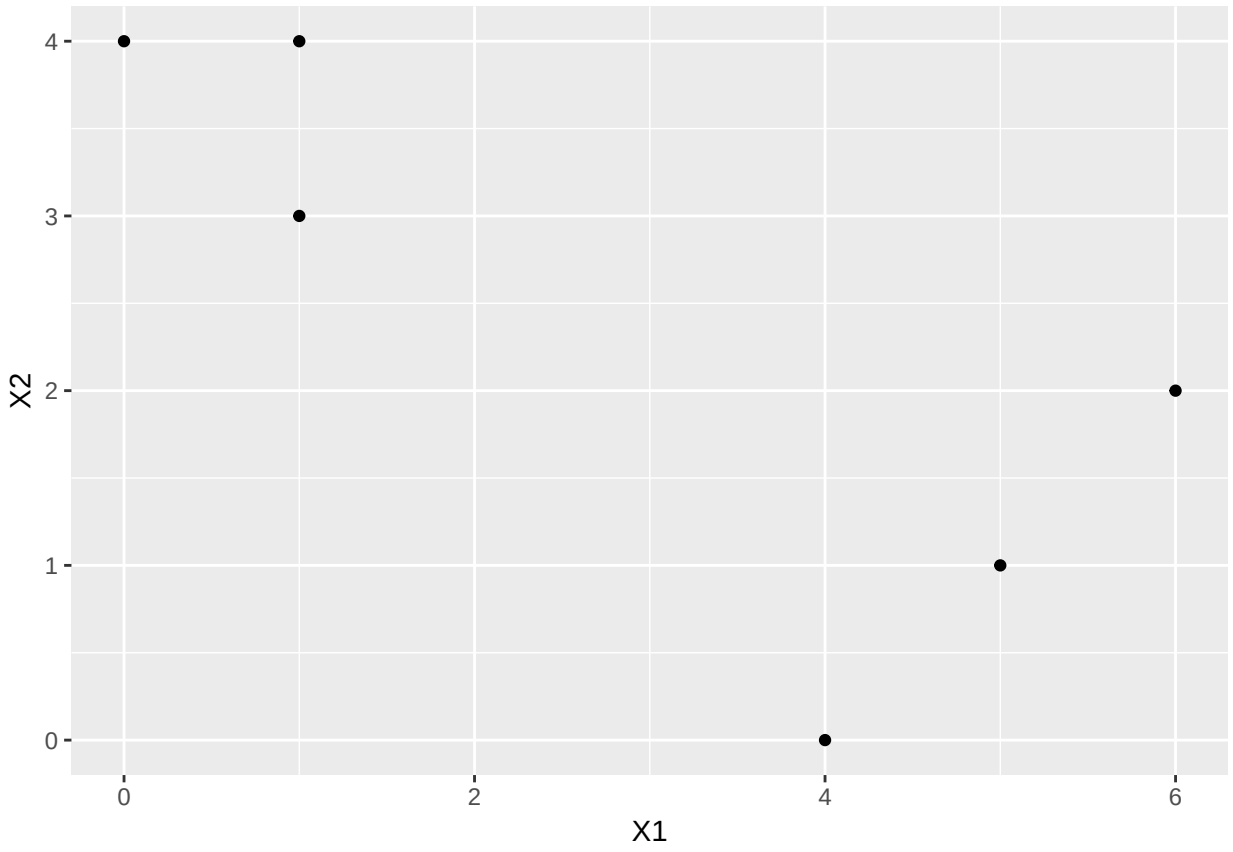
```
library(ggmap)
library(mapdata)
```

10.7 #3

- a)

```
obs <- data.frame(
  X1 = c(1, 1, 0, 5, 6, 4),
  X2 = c(4, 3, 4, 1, 2, 0)
)

obs %>%
  ggplot(aes(x = X1, y = X2)) +
  geom_point()
```



- b)

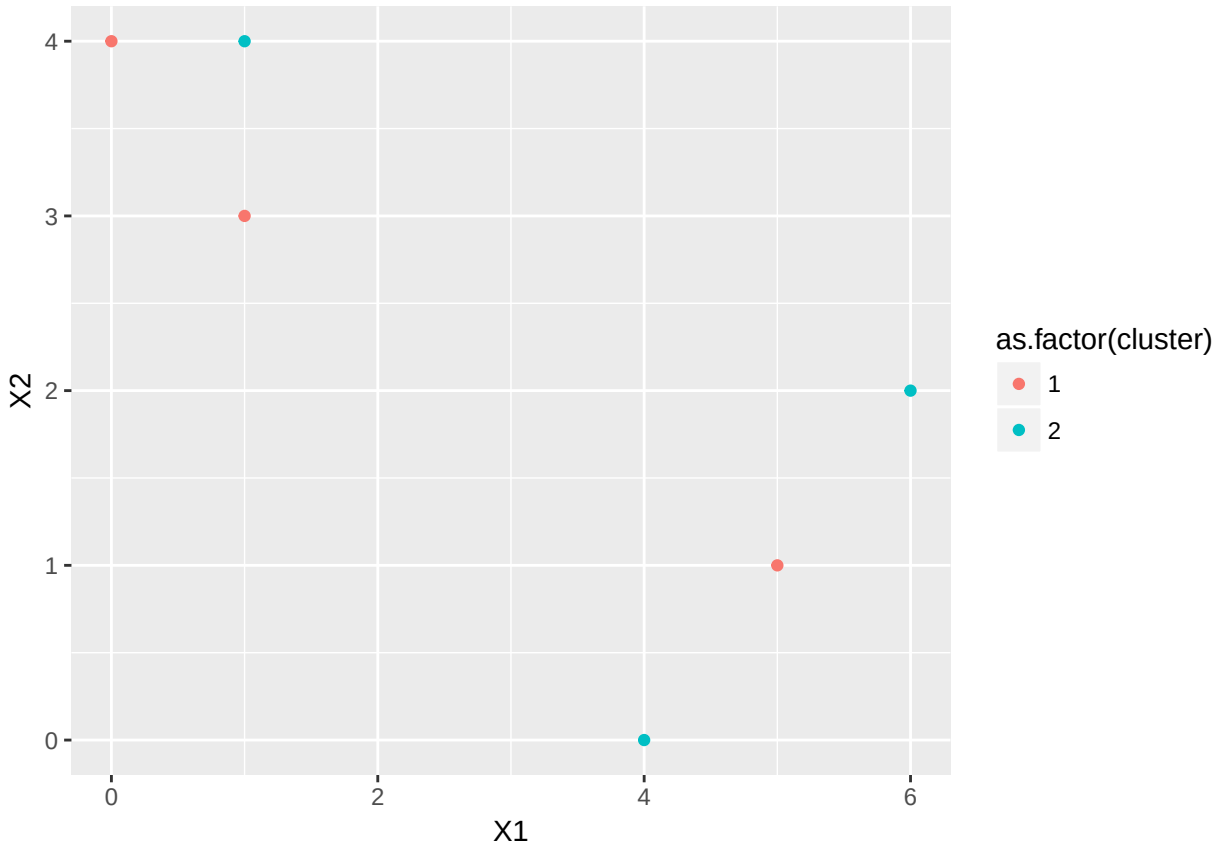
```
# set random number seed for replicable results
set.seed(2017)

# assign the labels
obs$cluster <- sample(rep(c(1, 2), each = 3), 6, replace = FALSE)

# print the labels
obs
```

```
##   X1 X2 cluster
## 1  1  4       2
## 2  1  3       1
## 3  0  4       1
## 4  5  1       1
## 5  6  2       2
## 6  4  0       2
```

```
# plot the labels
obs %>%
  dplyr::select(X1, X2, cluster) %>%
  ggplot(aes(x = X1, y = X2, color = as.factor(cluster))) +
  geom_point()
```



- c)

```
centroids <-
  obs %>%
  dplyr::group_by(cluster) %>%
  dplyr::summarise(centroid_X1 = round(mean(X1), 2), centroid_X2 = round(mean(X2), 2)) %>%
  as.data.frame()

centroids
```

```
##   cluster centroid_X1 centroid_X2
## 1       1        2.00        2.67
## 2       2        3.67        2.00
```

- d)

```
# create distance from cluster1 centroid
obs$dist_from_cluster1_centroid <- round(
  sqrt((obs$X1 - centroids$centroid_X1[centroids$cluster == 1])^2 +
       (obs$X2 - centroids$centroid_X2[centroids$cluster == 1])^2), 4)

# create distance from cluster2 centroid
obs$dist_from_cluster2_centroid <- round(
  sqrt((obs$X1 - centroids$centroid_X1[centroids$cluster == 2])^2 +
       (obs$X2 - centroids$centroid_X2[centroids$cluster == 2])^2), 4)

obs$new_cluster <-
  ifelse(obs$dist_from_cluster1_centroid <= obs$dist_from_cluster2_centroid, 1, 2)
```

```
obs
```

```
##   X1 X2 cluster dist_from_cluster1_centroid dist_from_cluster2_centroid
## 1  1  4       2                     1.6640                     3.3360
## 2  1  3       1                     1.0530                     2.8511
## 3  0  4       1                     2.4019                     4.1796
## 4  5  1       1                     3.4335                     1.6640
## 5  6  2       2                     4.0557                     2.3300
## 6  4  0       2                     3.3360                     2.0270
##   new_cluster
## 1           1
## 2           1
## 3           1
## 4           2
## 5           2
## 6           2
```

- e)

```
cluster_difference <- FALSE

while(cluster_difference == FALSE){

# reset obs$cluster to obs$new_cluster
obs$cluster <- obs$new_cluster

# create distance from cluster1 centroid
obs$dist_from_cluster1_centroid <- round(
  sqrt((obs$X1 - centroids$centroid_X1[centroids$cluster == 1])^2 +
       (obs$X2 - centroids$centroid_X2[centroids$cluster == 1])^2), 4)

# create distance from cluster2 centroid
obs$dist_from_cluster2_centroid <- round(
  sqrt((obs$X1 - centroids$centroid_X1[centroids$cluster == 2])^2 +
       (obs$X2 - centroids$centroid_X2[centroids$cluster == 2])^2), 4)

obs$new_cluster <-
  ifelse(obs$dist_from_cluster1_centroid <= obs$dist_from_cluster2_centroid, 1, 2)

cluster_difference <- all(obs$cluster == obs$new_cluster)

print(obs)

}
```

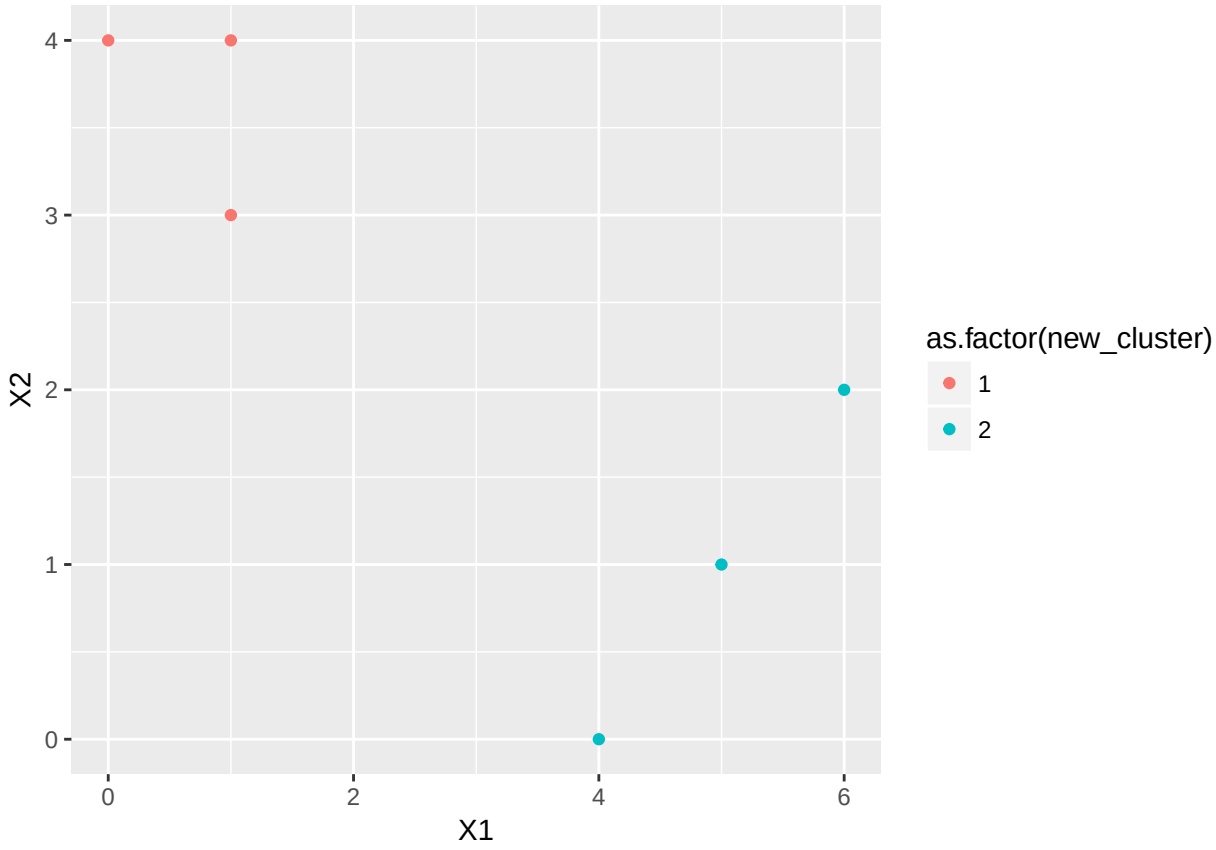
```
##   X1 X2 cluster dist_from_cluster1_centroid dist_from_cluster2_centroid
## 1  1  4       1                     1.6640                     3.3360
## 2  1  3       1                     1.0530                     2.8511
## 3  0  4       1                     2.4019                     4.1796
## 4  5  1       2                     3.4335                     1.6640
## 5  6  2       2                     4.0557                     2.3300
## 6  4  0       2                     3.3360                     2.0270
##   new_cluster
## 1           1
## 2           1
```

```
## 3           1
## 4           2
## 5           2
## 6           2
```

The clustering only took one iteration until the algorithm converged.

- f)

```
obs %>%
  dplyr::select(X1, X2, new_cluster) %>%
  ggplot(aes(x = X1, y = X2, color = as.factor(new_cluster))) +
  geom_point()
```

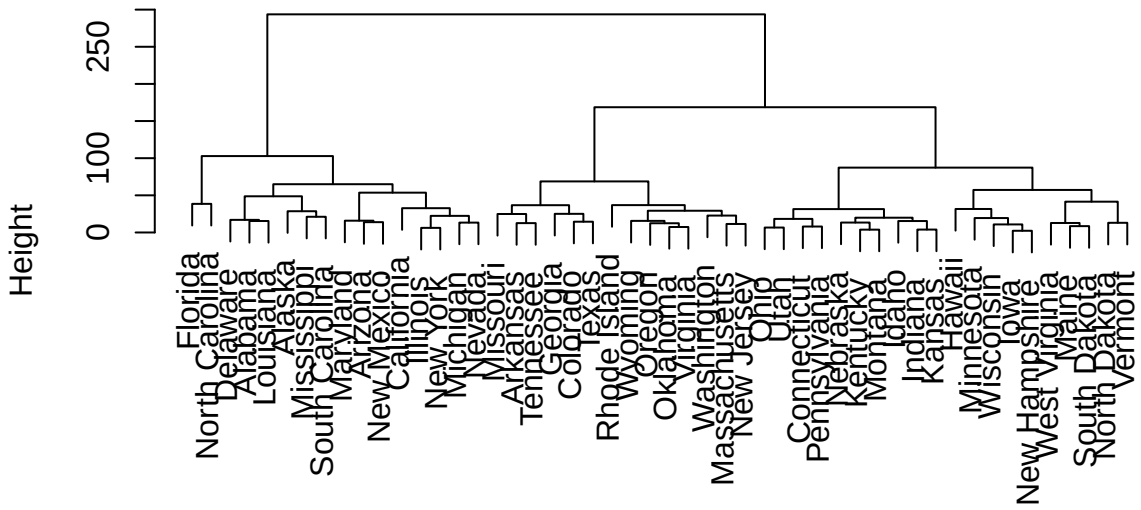


10.7 #9

- a)

```
# load data
data("USArrests")

# perform clustering and plot
hc_complete_no_scale <- hclust(dist(USArrests), method = "complete")
plot(hc_complete_no_scale, main = "Complete Linkage \n No Scaling")
```



- b)

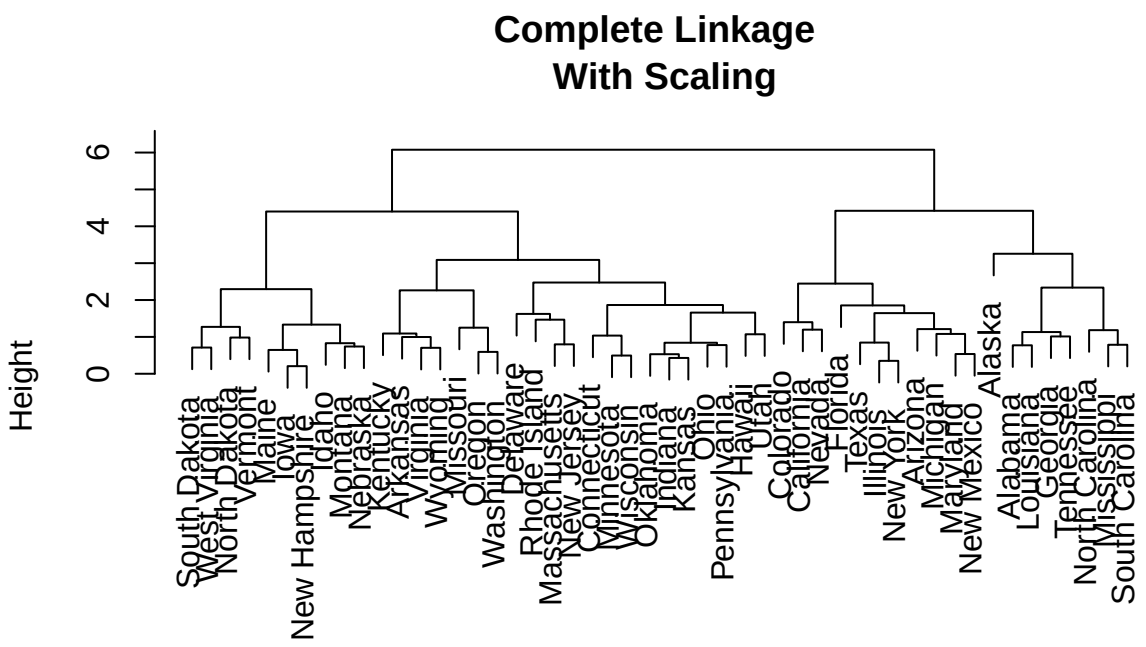
```
# use cutree, specifying 3 clusters
cutree(hc_complete_no_scale, k = 3)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              1              1              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              1              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              3              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              1              3              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              3              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              3              3              2
```

- c)

```r
# scale the USArrests data set
USArrests_scaled <- scale(USArrests)

# perform clustering on scaled data
hc_complete_scaled <- hclust(dist(USArrests_scaled), method = "complete")
plot(hc_complete_scaled, main = "Complete Linkage \n With Scaling")
```



dist(USArrests_scaled)
hclust (*, "complete")

```r
# use cutree, specifying 3 clusters
cutree(hc_complete_scaled, k = 3)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              2              3              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              3              2              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              2              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              2              3              1              3
##        Montana       Nebraska         Nevada New Hampshire     New Jersey
##              3              3              2              3              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              1              2              3              3
```

```
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              3              3              3
```

- d)

As can be seen from the results below, clustering on scaled data can create much different results than clustering on non-scaled data. The non-scaled data more evenly distributes the cluster assignments between 1, 2, and 3 while the scaled data creates far more cluster 3 labels (31) than cluster 1 or 2 labels (8 and 11 respectively). Furthermore, in examining in the scale of the variables, we see that Murder ranges from 0.80 to 17.40, Assault ranges from 45.0 to 337.0, UrbanPop ranges from 32.00 to 91.00, and Rape ranges from 7.30 to 46.00. Given the varying scales of these variables and because we are using Euclidean Distance, we expect Assault to dominate the clustering, followed by UrbanPop and Rape. Murder will have less of an effect because it is on a much smaller scale than the other variables. For example, a difference of 15 is significant for the Murder variable while it is fairly small for the Assault variable. In fact, 15 almost covers the entire range of Murder while it covers hardly any of the range for Assault.

```
# create data frame for comparison
no_scale <- data.frame(not_scaled_clusters = cutree(hc_complete_no_scale, k = 3))
with_scale <- data.frame(with_scaled_clusters = cutree(hc_complete_scaled, k = 3))
both_clusters <- cbind(no_scale, with_scale)
both_clusters$clusters_match <-
  ifelse(both_clusters$not_scaled_clusters == both_clusters$with_scaled_clusters,
         TRUE,
         FALSE)

# determine percentage of states whose clusters match between scaled and non-scaled data
round(mean(both_clusters$clusters_match), 4)
```

```
## [1] 0.56
```

```
# get counts of clusters per method
table(both_clusters$not_scaled_clusters)
```

```
##
##  1  2  3
## 16 14 20
```

```
table(both_clusters$with_scaled_clusters)
```

```
##
##  1  2  3
##  8 11 31
```

```
# create a table comparison of the two clustering methods
table(both_clusters$not_scaled_clusters, both_clusters$with_scaled_clusters)
```

```
##
##      1  2  3
##   1  6  9  1
##   2  2  2 10
##   3  0  0 20
```

```
# create summary of USArrests data set
summary(USArrests)
```

```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
```

```
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```