Week12_Data Clustering

Call useful libraries

```
library(tidyverse)  # data manipulation
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

```
## -- Attaching packages ----------------------------------------------------------------
------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## -- Conflicts -------------------------------------------------------------------------
------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(cluster)    # clustering algorithms
```

```
## Warning: package 'cluster' was built under R version 3.5.2
```

```
library(factoextra)  # clustering algorithms & visualization{r}
```

```
## Warning: package 'factoextra' was built under R version 3.5.2
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

```
library(gridExtra)
```

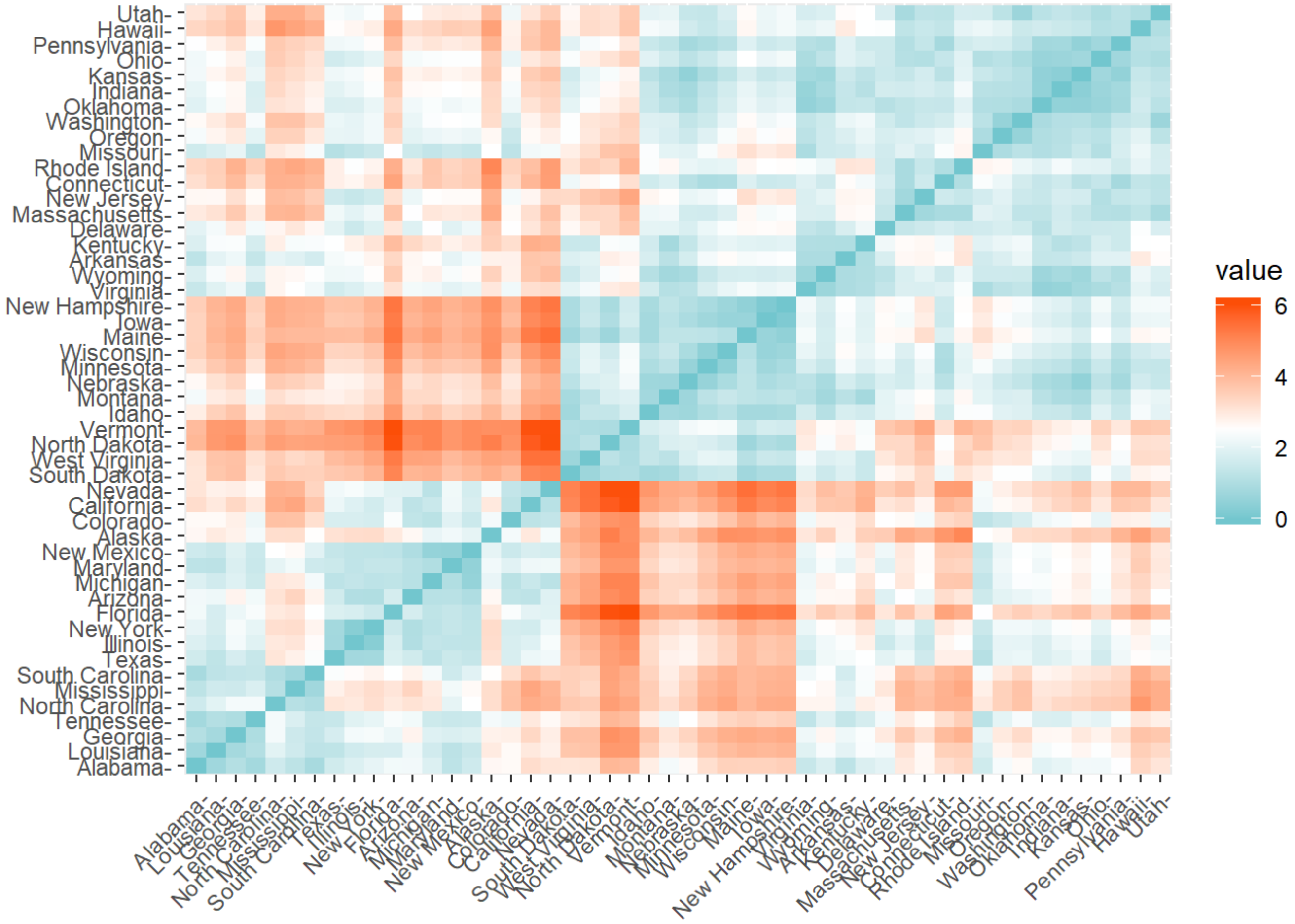
```
## Warning: package 'gridExtra' was built under R version 3.5.2
```

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

```
df <- USArrests # read USArrests data
df <- na.omit(df) #Remove any missing values
df <- scale(df)
head(df)
```

```
##                 Murder   Assault   UrbanPop         Rape
## Alabama    1.24256408 0.7828393 -0.5209066 -0.003416473
## Alaska     0.50786248 1.1068225 -1.2117642  2.484202941
## Arizona    0.07163341 1.4788032  0.9989801  1.042878388
## Arkansas   0.23234938 0.2308680 -1.0735927 -0.184916602
## California 0.27826823 1.2628144  1.7589234  2.067820292
## Colorado   0.02571456 0.3988593  0.8608085  1.864967207
```

```
distance <- get_dist(df)  # distance matrix between the rows of a data matrix
fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07")) # It perf
orms principle component analysis and Visualization of distance matrix
```

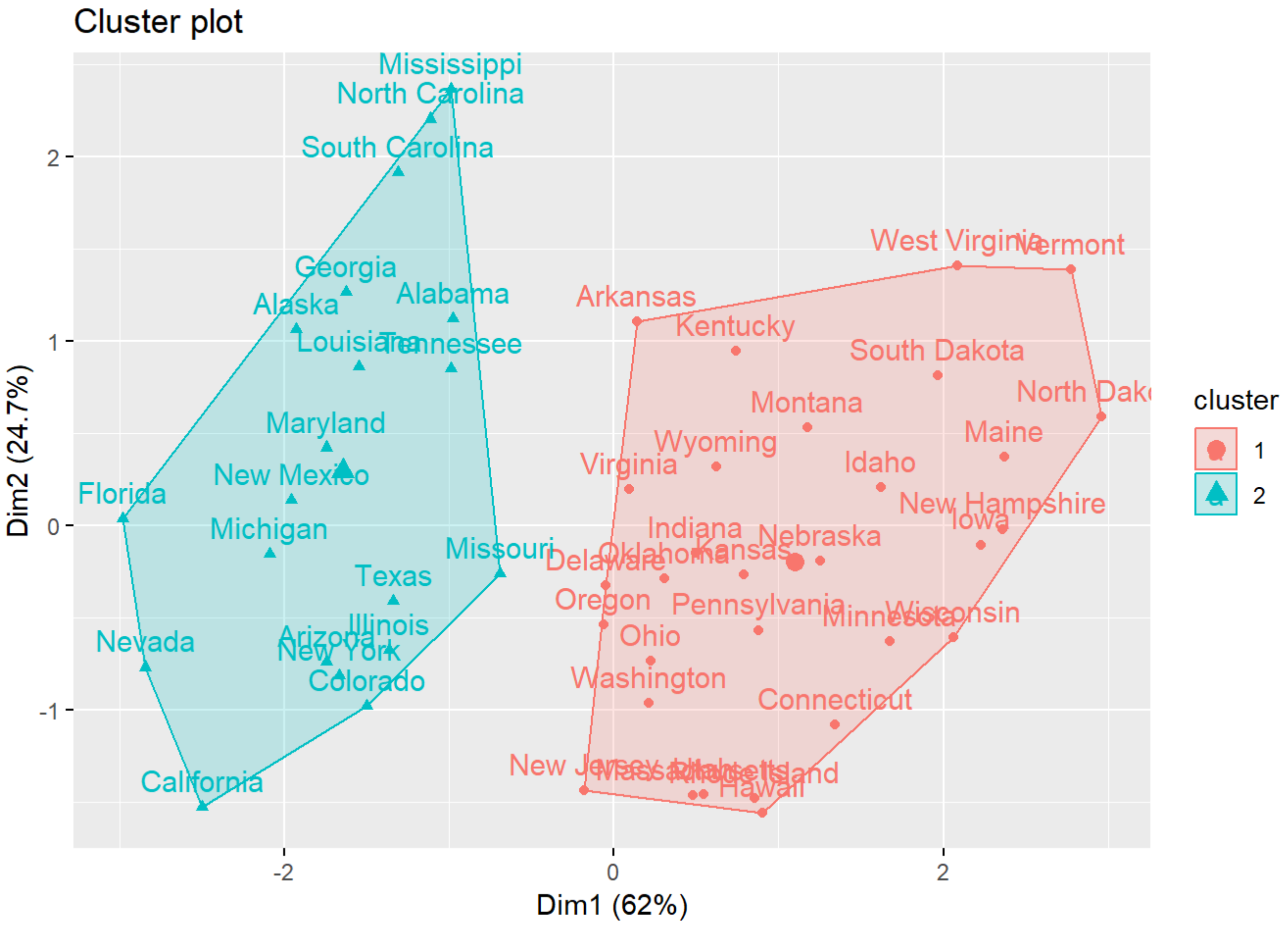


Kemans clustering

```
k2 <- kmeans(df, centers = 2, nstart = 25)
str(k2)
```

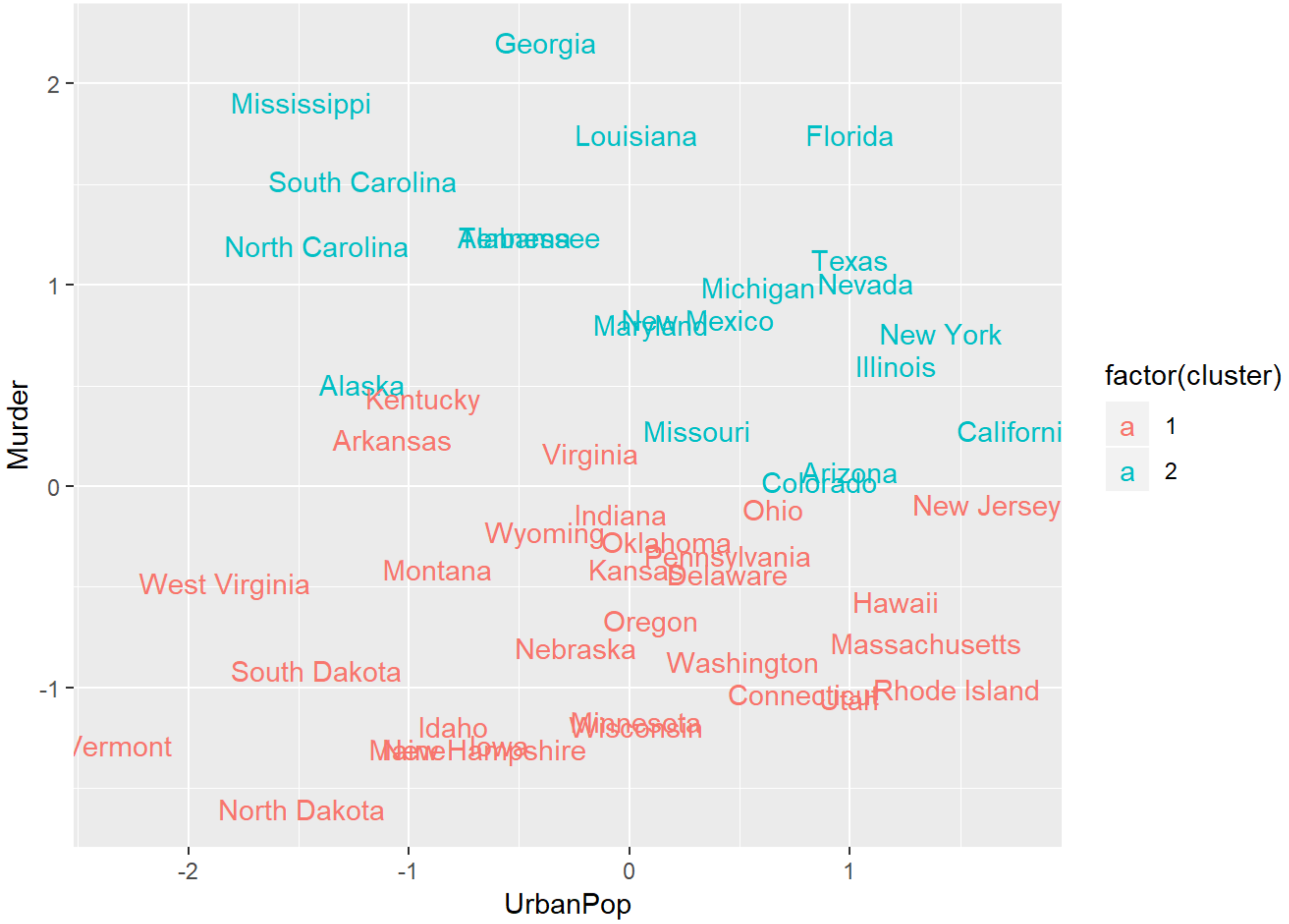
```
## List of 9
##  $ cluster     : Named int [1:50] 2 2 2 1 2 2 1 1 2 2 ...
##   ..- attr(*, "names")= chr [1:50] "Alabama" "Alaska" "Arizona" "Arkansas" ...
##  $ centers     : num [1:2, 1:4] -0.67 1.005 -0.676 1.014 -0.132 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:4] "Murder" "Assault" "UrbanPop" "Rape"
##  $ totss       : num 196
##  $ withinss    : num [1:2] 56.1 46.7
##  $ tot.withinss: num 103
##  $ betweenss   : num 93.1
##  $ size        : int [1:2] 30 20
##  $ iter        : int 1
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
fviz_cluster(k2, data = df)
```



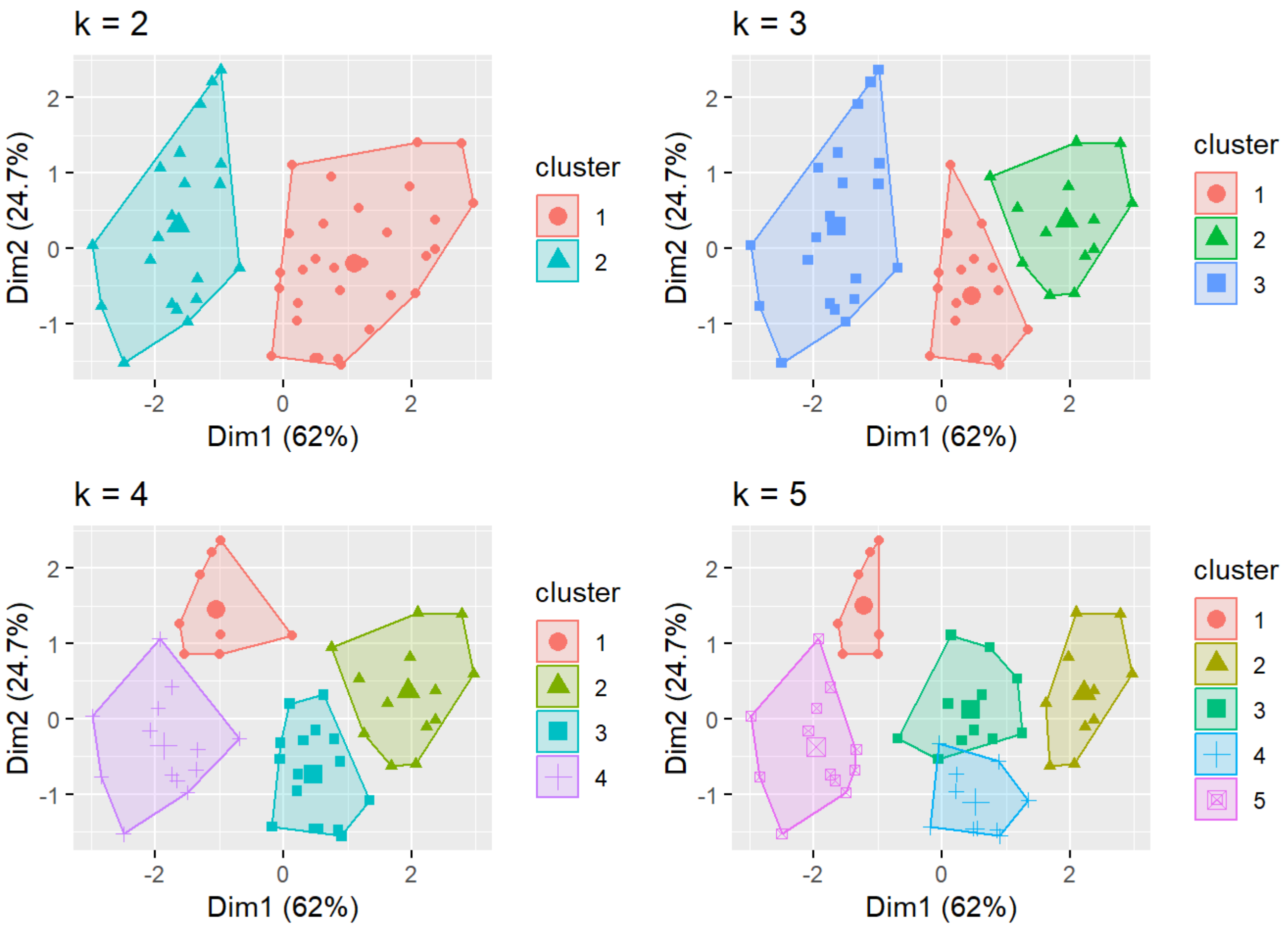
Pairwise scatter plot

```
df %>%
  as_tibble() %>%
  mutate(cluster = k2$cluster,
         state = row.names(USArrests)) %>%
  ggplot(aes(UrbanPop, Murder, color = factor(cluster), label = state)) +
  geom_text()
```



Comparision plots

```
k3 <- kmeans(df, centers = 3, nstart = 25)
k4 <- kmeans(df, centers = 4, nstart = 25)
k5 <- kmeans(df, centers = 5, nstart = 25)
p1 <- fviz_cluster(k2, geom = "point", data = df) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = df) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = df) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = df) + ggtitle("k = 5")
grid.arrange(p1, p2, p3, p4, nrow = 2)
```


k = 2
Dim2 (24.7%)
Dim1 (62%)
2
1
0
-1
-2
0
2
cluster
1
2
k = 3
Dim2 (24.7%)
Dim1 (62%)
cluster
1
2
3
k = 4
Dim2 (24.7%)
Dim1 (62%)
cluster
1
2
3
4
k = 5
Dim2 (24.7%)
Dim1 (62%)
cluster
1
2
3
4
5