**Problem Statement**: Analyze the monthly seasonally adjusted unemployment rates for the given employment data of the U.S. “unemp.csv” file, covering the period January 1976 through August 2010 for the 50 U.S. states($n$ = 50).

The requirement is to cluster the states into groups that are alike. Here, each state is characterized by a feature vector of a very large dimension ($p$ = 416). Its components represent 416 monthly observations.

For the purpose of an illustration, assume that New York and California form a cluster.

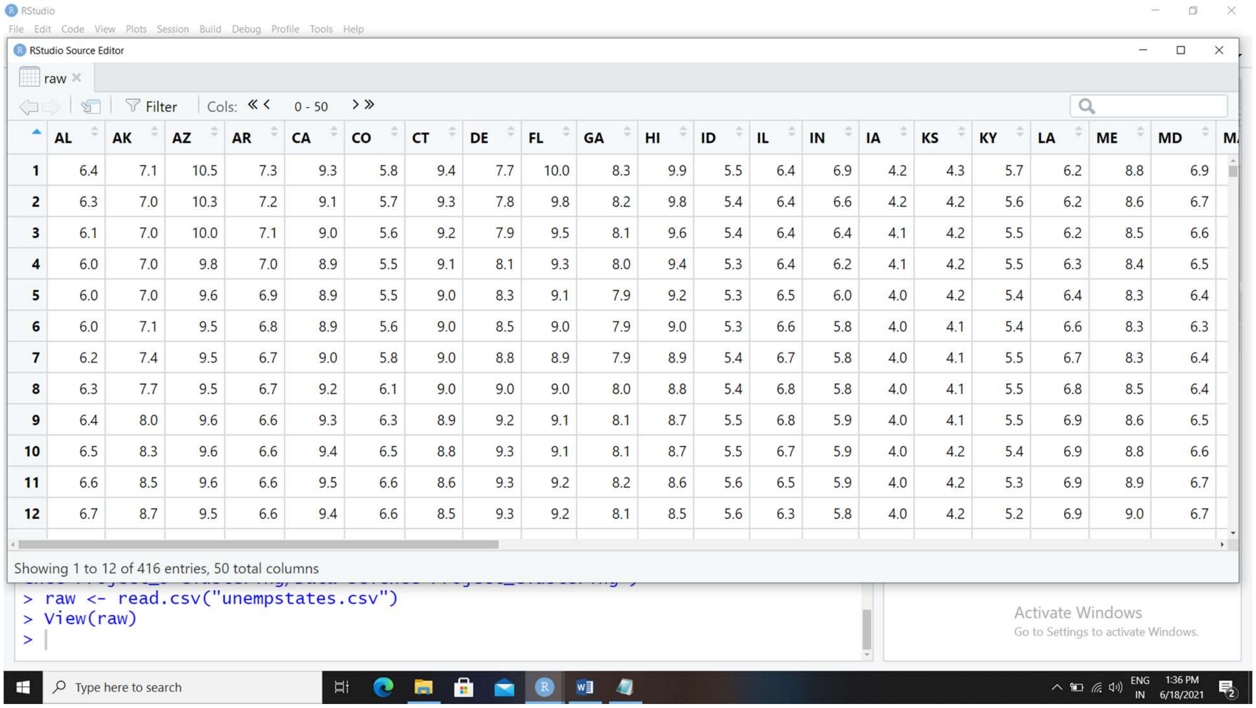
You will need to calculate 416 monthly averages (of two observations each). This vector of averages is called the centroid for that cluster.

Note that the sum of the squared distances from the centroid of this cluster expresses the within-cluster sum of squares. To explain it better, in the mentioned example, for New York and California, there are 2*(416)* = 832 such distances.

Code

```
raw <- read.csv("unempstates.csv")
```

```
View(raw)
```



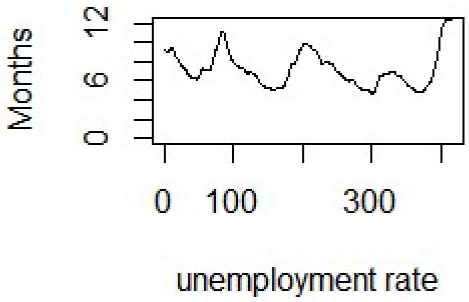
| | AL | AK | AZ | AR | CA | CO | CT | DE | FL | GA | HI | ID | IL | IN | IA | KS | KY | LA | ME | MD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 6.4 | 7.1 | 10.5 | 7.3 | 9.3 | 5.8 | 9.4 | 7.7 | 10.0 | 8.3 | 9.9 | 5.5 | 6.4 | 6.9 | 4.2 | 4.3 | 5.7 | 6.2 | 8.8 | 6.9 |
| 2 | 6.3 | 7.0 | 10.3 | 7.2 | 9.1 | 5.7 | 9.3 | 7.8 | 9.8 | 8.2 | 9.8 | 5.4 | 6.4 | 6.6 | 4.2 | 4.2 | 5.6 | 6.2 | 8.6 | 6.7 |
| 3 | 6.1 | 7.0 | 10.0 | 7.1 | 9.0 | 5.6 | 9.2 | 7.9 | 9.5 | 8.1 | 9.6 | 5.4 | 6.4 | 6.4 | 4.1 | 4.2 | 5.5 | 6.2 | 8.5 | 6.6 |
| 4 | 6.0 | 7.0 | 9.8 | 7.0 | 8.9 | 5.5 | 9.1 | 8.1 | 9.3 | 8.0 | 9.4 | 5.3 | 6.4 | 6.2 | 4.1 | 4.2 | 5.5 | 6.3 | 8.4 | 6.5 |
| 5 | 6.0 | 7.0 | 9.6 | 6.9 | 8.9 | 5.5 | 9.0 | 8.3 | 9.1 | 7.9 | 9.2 | 5.3 | 6.5 | 6.0 | 4.0 | 4.2 | 5.4 | 6.4 | 8.3 | 6.4 |
| 6 | 6.0 | 7.1 | 9.5 | 6.8 | 8.9 | 5.6 | 9.0 | 8.5 | 9.0 | 7.9 | 9.0 | 5.3 | 6.6 | 5.8 | 4.0 | 4.1 | 5.4 | 6.6 | 8.3 | 6.3 |
| 7 | 6.2 | 7.4 | 9.5 | 6.7 | 9.0 | 5.8 | 9.0 | 8.8 | 8.9 | 7.9 | 8.9 | 5.4 | 6.7 | 5.8 | 4.0 | 4.1 | 5.5 | 6.7 | 8.3 | 6.4 |
| 8 | 6.3 | 7.7 | 9.5 | 6.7 | 9.2 | 6.1 | 9.0 | 9.0 | 9.0 | 8.0 | 8.8 | 5.4 | 6.8 | 5.8 | 4.0 | 4.1 | 5.5 | 6.8 | 8.5 | 6.4 |
| 9 | 6.4 | 8.0 | 9.6 | 6.6 | 9.3 | 6.3 | 8.9 | 9.2 | 9.1 | 8.1 | 8.7 | 5.5 | 6.8 | 5.9 | 4.0 | 4.1 | 5.5 | 6.9 | 8.6 | 6.5 |
| 10 | 6.5 | 8.3 | 9.6 | 6.6 | 9.4 | 6.5 | 8.8 | 9.3 | 9.1 | 8.1 | 8.7 | 5.5 | 6.7 | 5.9 | 4.0 | 4.2 | 5.4 | 6.9 | 8.8 | 6.6 |
| 11 | 6.6 | 8.5 | 9.6 | 6.6 | 9.5 | 6.6 | 8.6 | 9.3 | 9.2 | 8.2 | 8.6 | 5.6 | 6.5 | 5.9 | 4.0 | 4.2 | 5.3 | 6.9 | 8.9 | 6.7 |
| 12 | 6.7 | 8.7 | 9.5 | 6.6 | 9.4 | 6.6 | 8.5 | 9.3 | 9.2 | 8.1 | 8.5 | 5.6 | 6.3 | 5.8 | 4.0 | 4.2 | 5.2 | 6.9 | 9.0 | 6.7 |

Showing 1 to 12 of 416 entries, 50 total columns

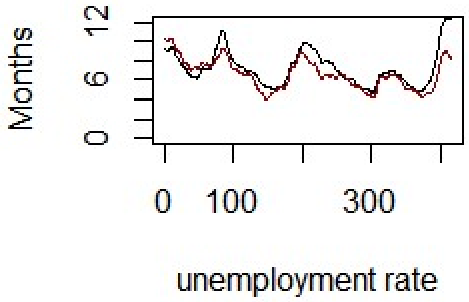
```
> raw <- read.csv("unempstates.csv")
> View(raw)
>
```

```
## time sequence plots of three series
```

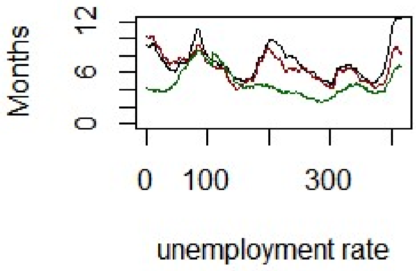
```
plot(raw[,5],type="l",ylim=c(0,12),xlab="unemployment rate",ylab="Months")  #CA
```



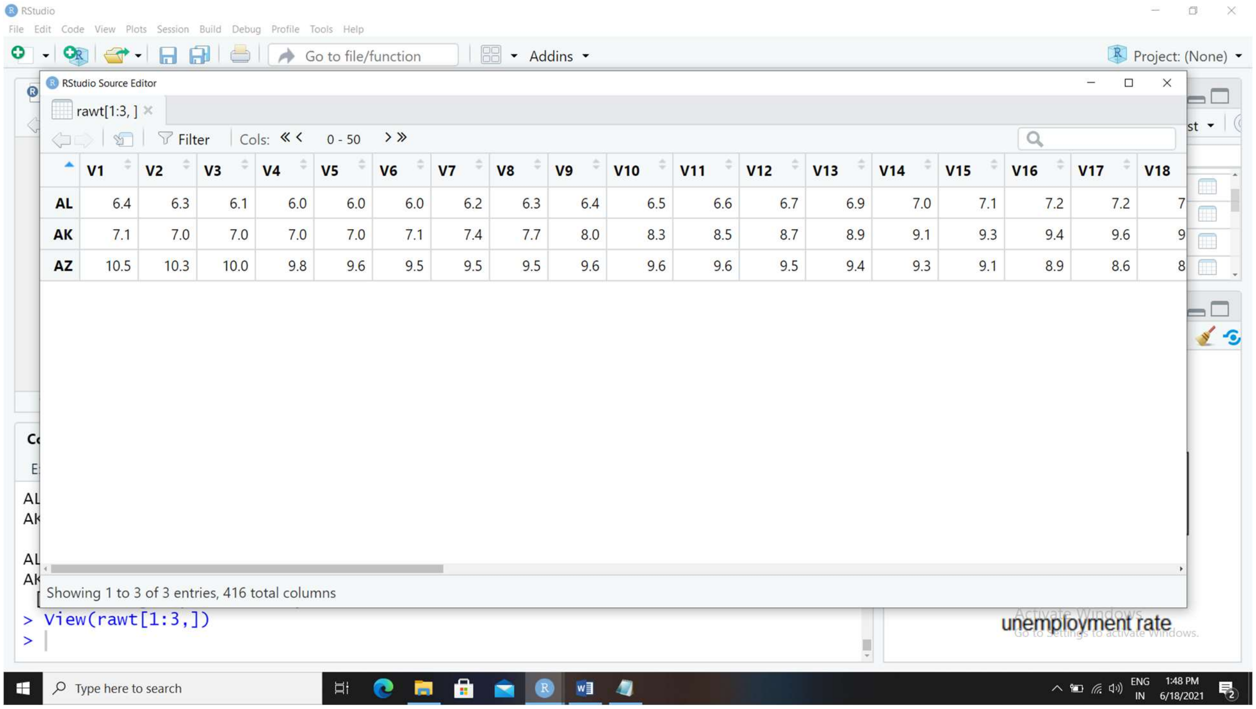
```
points(raw[,32],type="l", cex = .5, col = "dark red") ## New York\
```



```
points(raw[,15],type="l", cex = .5, col = "dark green") ## Iowa
```



```
## transpose the data
## then we have 50 rows (states) and 416 columns (time periods)

rawt=matrix(nrow=50,ncol=416)
rawt=t(raw)
View(rawt[1:3,])
```



|  | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 | V11 | V12 | V13 | V14 | V15 | V16 | V17 | V18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AL | 6.4 | 6.3 | 6.1 | 6.0 | 6.0 | 6.0 | 6.2 | 6.3 | 6.4 | 6.5 | 6.6 | 6.7 | 6.9 | 7.0 | 7.1 | 7.2 | 7.2 | 7 |
| AK | 7.1 | 7.0 | 7.0 | 7.0 | 7.0 | 7.1 | 7.4 | 7.7 | 8.0 | 8.3 | 8.5 | 8.7 | 8.9 | 9.1 | 9.3 | 9.4 | 9.6 | 9 |
| AZ | 10.5 | 10.3 | 10.0 | 9.8 | 9.6 | 9.5 | 9.5 | 9.5 | 9.6 | 9.6 | 9.6 | 9.5 | 9.4 | 9.3 | 9.1 | 8.9 | 8.6 | 8 |

Showing 1 to 3 of 3 entries, 416 total columns

```
> View(rawt[1:3,])
>
```

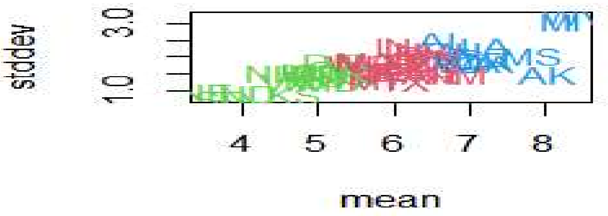
```
> ## k-means clustering in 416 dimensions

> set.seed(1)

> grpunemp2 <- kmeans(rawt, centers=2, nstart=10)

> sort(grpunemp2$cluster)

          AL AK AZ AR CA FL ID IL IN KY LA MI MS MO NV NJ NM NY OH OR PA RI SC TN TX

           1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1

          WA WV CO CT DE GA HI IA KS ME MD MA MN MT NE NH NC ND OK SD UT VT VA WI WY

          1  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
```

```
> grpunemp3 <- kmeans(rawt, centers=3, nstart=10)

> sort(grpunemp3$cluster)

        AZ CA CT DE FL GA ID IN ME MA MO MT NV NJ NY NC PA RI SC TX WI CO HI IA KS

         1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  2  2  2  2

        MD MN NE NH ND OK SD UT VT VA WY AL AK AR IL KY LA MI MS NM OH OR TN WA WV

         2  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3  3  3  3  3  3  3


> grpunemp4 <- kmeans(rawt, centers=4, nstart=10)

> sort(grpunemp4$cluster)

        AL AR CA ID IL IN KY MO MT NV NM OH OR PA SC TN TX WA WI AK LA MI MS WV AZ

         1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  2  2  2  2  2  3

        CT DE FL GA ME MD MA NJ NY NC RI CO HI IA KS MN NE NH ND OK SD UT VT VA WY

         3  3  3  3  3  3  3  3  3  3  3  4  4  4  4  4  4  4  4  4  4  4  4  4  4

> grpunemp5 <- kmeans(rawt, centers=5, nstart=10)

> sort(grpunemp5$cluster)

        HI KS NE NH ND SD VT VA CO ID IA MN MO MT OK TX UT WI WY AK LA MI MS WV AL

         1  1  1  1  1  1  1  1  2  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  4

        AR CA IL IN KY NV NM OH OR PA SC TN WA AZ CT DE FL GA ME MD MA NJ NY NC RI

         4  4  4  4  4  4  4  4  4  4  4  4  4  5  5  5  5  5  5  5  5  5  5  5  5




> ## another analysis

> ## data set unemp.csv with means and standard deviations for each state

> ## k-means clustering on 2 dimensions (mean, stddev)

> unemp <- read.csv("unemp.csv")
> unemp[1:3,]

  state    mean              stddev

1    AL    6.644952          2.527530

2    AK    8.033173          1.464966

3    AZ    6.120673          1.743672


> set.seed(1)

> grpunemp <- kmeans(unemp[,c("mean","stddev")], centers=3, nstart=10)

> sort(grpunemp3$cluster)

          AZ CA CT DE FL GA ID IN ME MA MO MT NV NJ NY NC PA RI SC TX WI CO HI IA KS

          1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  2  2  2  2

          MD MN NE NH ND OK SD UT VT VA WY AL AK AR IL KY LA MI MS NM OH OR TN WA WV

          2  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3  3  3  3  3  3  3  3  3


## list of cluster assignments

> o=order(grpunemp$cluster)

> data.frame(unemp$state[o],grpunemp$cluster[o])

   unemp.state.o. grpunemp.cluster.o.

1            AZ                1

2            AR                1

3            FL                1

4            GA                1

5            ID                1

6            IN                1

7            ME                1
```

| | | |
|---|---|---|
| 8 | MA | 1 |
| 9 | MO | 1 |
| 10 | MT | 1 |
| 11 | NV | 1 |
| 12 | NJ | 1 |
| 13 | NM | 1 |
| 14 | NY | 1 |
| 15 | NC | 1 |
| 16 | PA | 1 |
| 17 | RI | 1 |
| 18 | SC | 1 |
| 19 | TN | 1 |
| 20 | TX | 1 |
| 21 | WI | 1 |
| 22 | CO | 2 |
| 23 | CT | 2 |
| 24 | DE | 2 |
| 25 | HI | 2 |
| 26 | IA | 2 |
| 27 | KS | 2 |
| 28 | MD | 2 |
| 29 | MN | 2 |
| 30 | NE | 2 |
| 31 | NH | 2 |
| 32 | ND | 2 |

|  |  |  |
|---|---|---|
| 33 | OK | 2 |
| 34 | SD | 2 |
| 35 | UT | 2 |
| 36 | VT | 2 |
| 37 | VA | 2 |
| 38 | WY | 2 |
| 39 | AL | 3 |
| 40 | AK | 3 |
| 41 | CA | 3 |
| 42 | IL | 3 |
| 43 | KY | 3 |
| 44 | LA | 3 |
| 45 | MI | 3 |
| 46 | MS | 3 |
| 47 | OH | 3 |
| 48 | OR | 3 |
| 49 | WA | 3 |
| 50 | WV | 3 |

```
text(x=unemp$mean,y=unemp$stddev,labels=unemp$state, col=grpunemp$cluster+1)
```


stddev
mean
MS
AK