# 535 Exam1

Minghan

2023-03-07

```
library(readxl)
library(tibble)
cities0 = read_xlsx("cities1.xlsx")
cities0 = column_to_rownames(cities0, var = "Metropolitan_Area")
cities0['Metropolitan_Area'] = NULL
cities0['Unemployment_Threat'] = NULL
cities0['Crime_Trend'] = NULL
cities = scale(cities0)
head(cities)
```

```
##                              Cost_Living Transportation        Jobs   Education
## Abilene, TX                    1.5375841     -0.4278150 -1.17109558 -0.02189764
## Akron, OH                     -0.1592646      0.7216188  1.21773326  0.75761314
## Albany, GA                     1.1844346     -0.7233242 -0.65987302 -0.80175243
## Albany-Schenectady-Troy, NY   -0.9240755      1.1735537  0.06756926  1.70293338
## Albuquerque, NM               -0.2572462      1.2228052  1.37529948  0.74798106
## Alexandria, LA                 1.4004790     -0.2214441 -1.10237727 -1.31810093
##                                 Climate       Crime       Arts Health_Care
## Abilene, TX                   0.1217642  0.02160855 -0.7787573 -0.09942403
## Akron, OH                    -1.0275913  0.17567503  1.1159913 -0.83809718
## Albany, GA                    0.8251586 -1.13440021 -0.5714811 -0.97758911
## Albany-Schenectady-Troy, NY -1.5130768  0.85009850  1.0470153  1.03799876
## Albuquerque, NM               0.9143508 -1.56803103  0.8989609  1.05807714
## Alexandria, LA                0.4883268 -1.42348742 -0.3049831  0.49905268
##                              Recreation Population_2000 Total_Violent
## Abilene, TX                  -1.6178059     -0.49804629    0.05087827
## Akron, OH                     0.9251346      0.01424981   -0.15393457
## Albany, GA                   -1.4826375     -0.50064748    0.62371418
## Albany-Schenectady-Troy, NY   0.9350333      0.19192785   -0.64356527
## Albuquerque, NM               0.6834699      0.05473629    1.81418882
## Alexandria, LA               -0.9409400     -0.49449352    1.70858220
##                              Total_Property Past_Job_Growth
## Abilene, TX                      -0.3703589      -0.5754658
## Akron, OH                        -0.2780023       0.1144370
## Albany, GA                        1.4908717      -0.5127474
## Albany-Schenectady-Troy, NY      -0.9801939      -0.6883590
## Albuquerque, NM                   1.6494993       1.3437184
## Alexandria, LA                    0.4650799      -0.5754658
##                              Fcast_Future_Job_Growth Fcast_Blue_Collar_Jobs
## Abilene, TX                                -0.5869163             -0.3326589
## Akron, OH                                   0.6135103              0.6154288
## Albany, GA                                  0.4134392             -0.3145416
## Albany-Schenectady-Troy, NY                -0.9070300             -0.3731472
```

```
## Albuquerque, NM                           0.8535956             0.5339796
## Alexandria, LA                           -0.4668736            -0.2882320
##                             Fcast_White_Collar_Jobs Fcast_High_Jobs
## Abilene, TX                              -0.5579673      -0.4183905
## Akron, OH                                 0.1095455       0.4460329
## Albany, GA                               -0.5121827      -0.4422484
## Albany-Schenectady-Troy, NY              -0.0380763      -0.3212547
## Albuquerque, NM                           0.2960859       0.6457010
## Alexandria, LA                           -0.5797517      -0.4003551
##                             Fcast_Average_Jobs
## Abilene, TX                       -0.549087327
## Akron, OH                          0.133627625
## Albany, GA                        -0.450492367
## Albany-Schenectady-Troy, NY       -0.008084422
## Albuquerque, NM                    0.219945036
## Alexandria, LA                    -0.563862004
```

```
dist = dist(cities)
head(dist)
```

```
## [1] 5.220089 2.924855 5.445753 6.487337 2.910852 3.882701
```

```
length(dist)
```

```
## [1] 52650
```

##K-MEANS CLUSTERING

```
library(cluster)
library(factoextra)
```

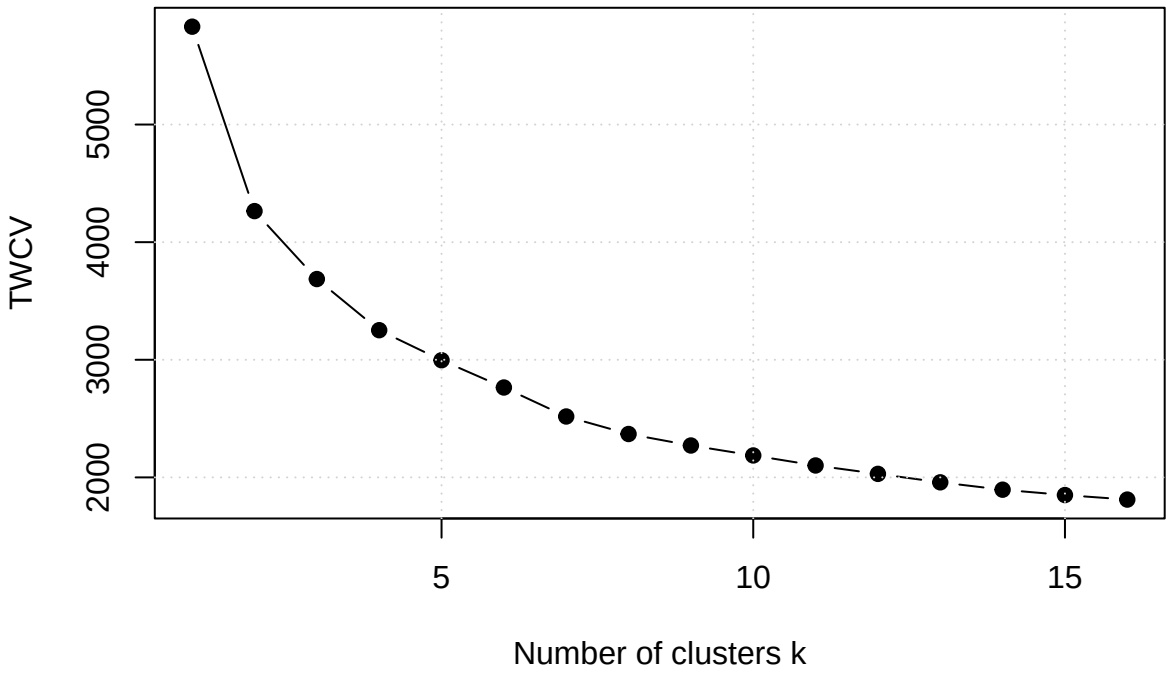
```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

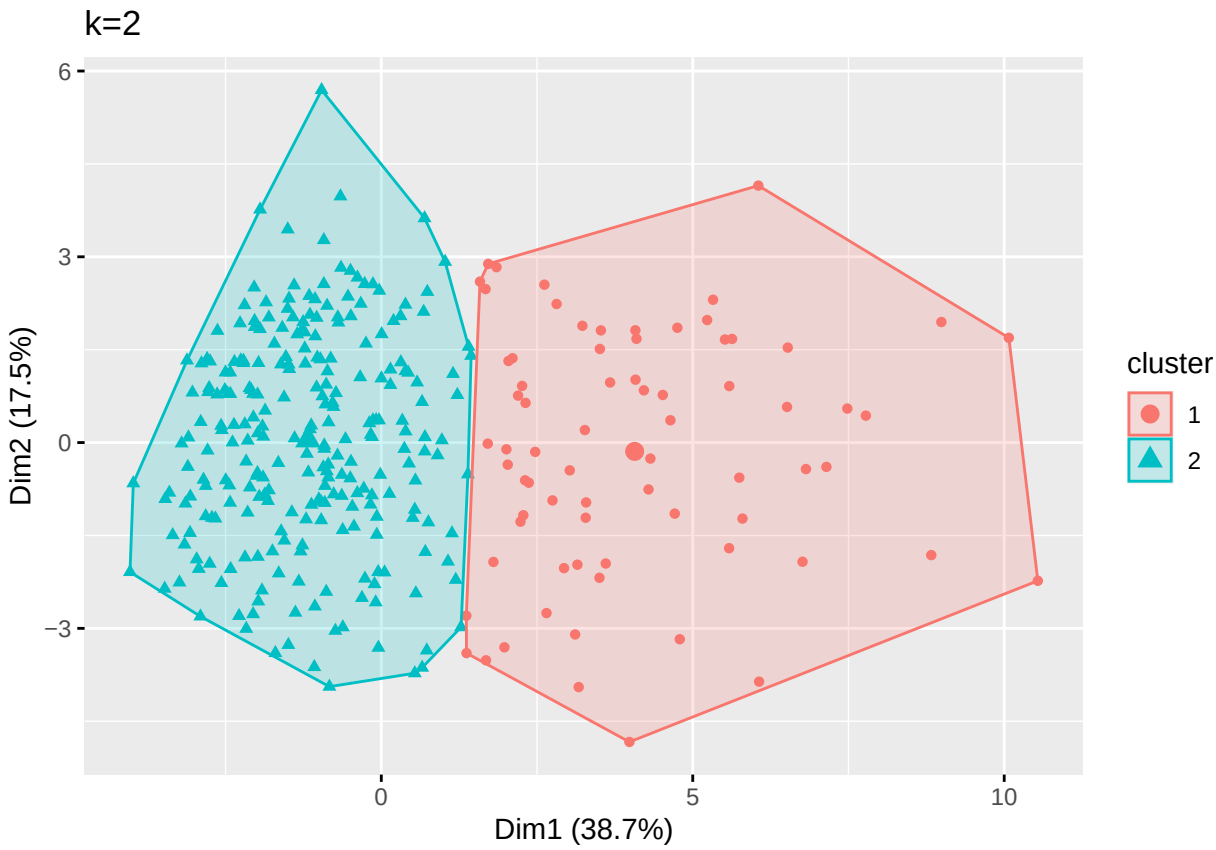
```
library(ggplot2)

set.seed(123)
twcv = function(k) kmeans(cities,k,nstart = 25)$tot.withinss
k = 1:16
twcv_value = sapply(k,twcv)
head(twcv_value)
```

```
## [1] 5832.000 4264.026 3686.381 3251.504 2996.052 2764.309
```

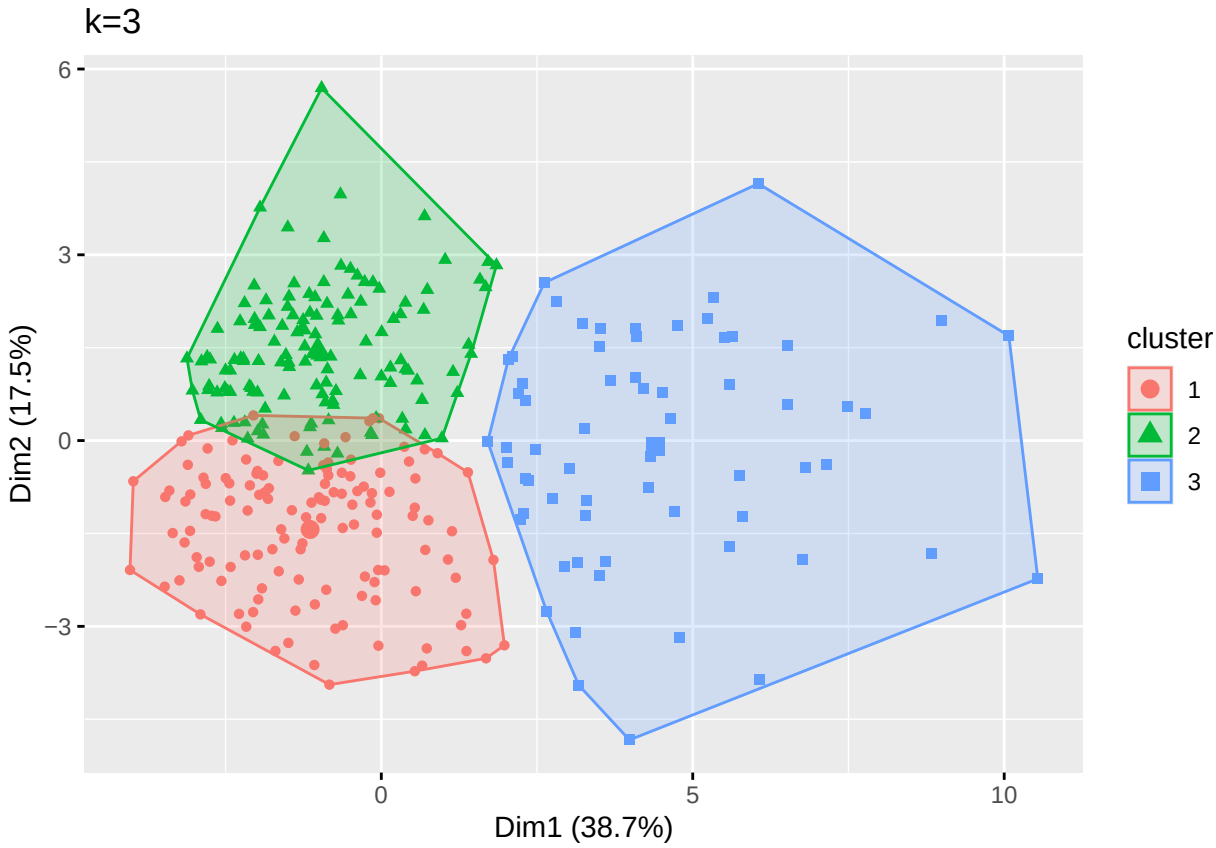
```
#write a elbow chart, Identify the point where the TWCV starts to decrease slowly
plot(k,twcv_value, type = 'b', pch = 19, xlab = 'Number of clusters k', ylab = 'TWCV')
grid()
```



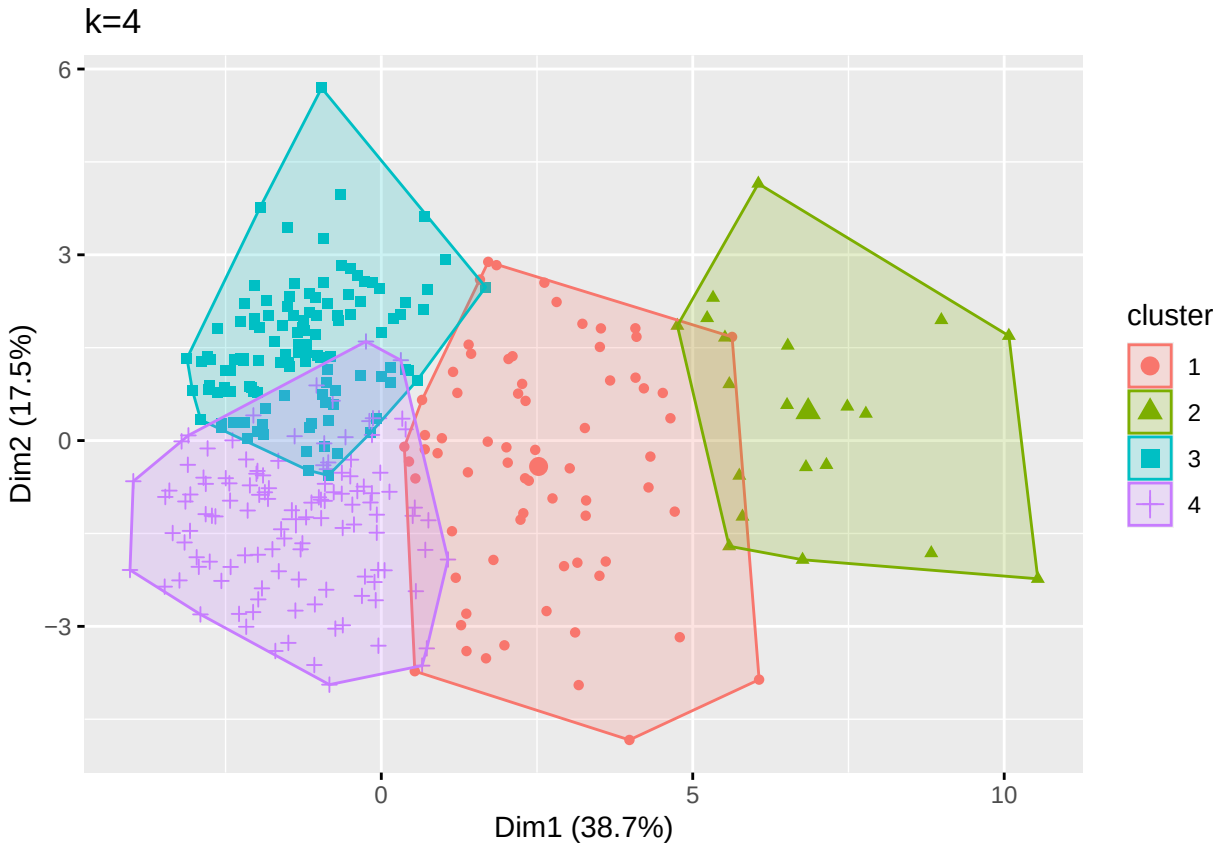

```
k2 = kmeans(cities, 2, nstart = 25)
k3 = kmeans(cities, 3, nstart = 25)
k4 = kmeans(cities, 4, nstart = 25)
k5 = kmeans(cities, 5, nstart = 25)
k6 = kmeans(cities, 6, nstart = 25)
library(ggplot2)
library(ggrepel)
fviz_cluster(k2, data = cities,geom='point') + ggtitle("k=2")
```



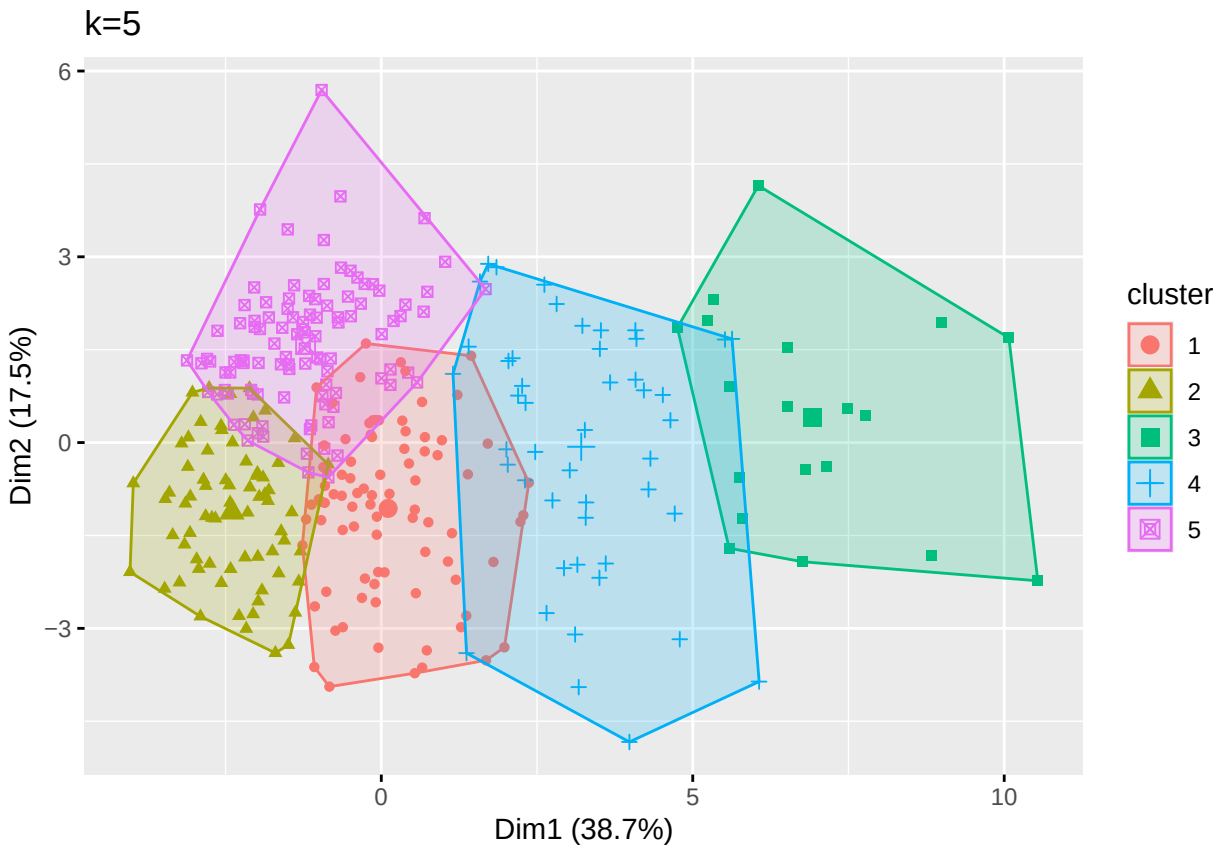

```
fviz_cluster(k3, data = cities,geom='point') + ggtitle("k=3")
```



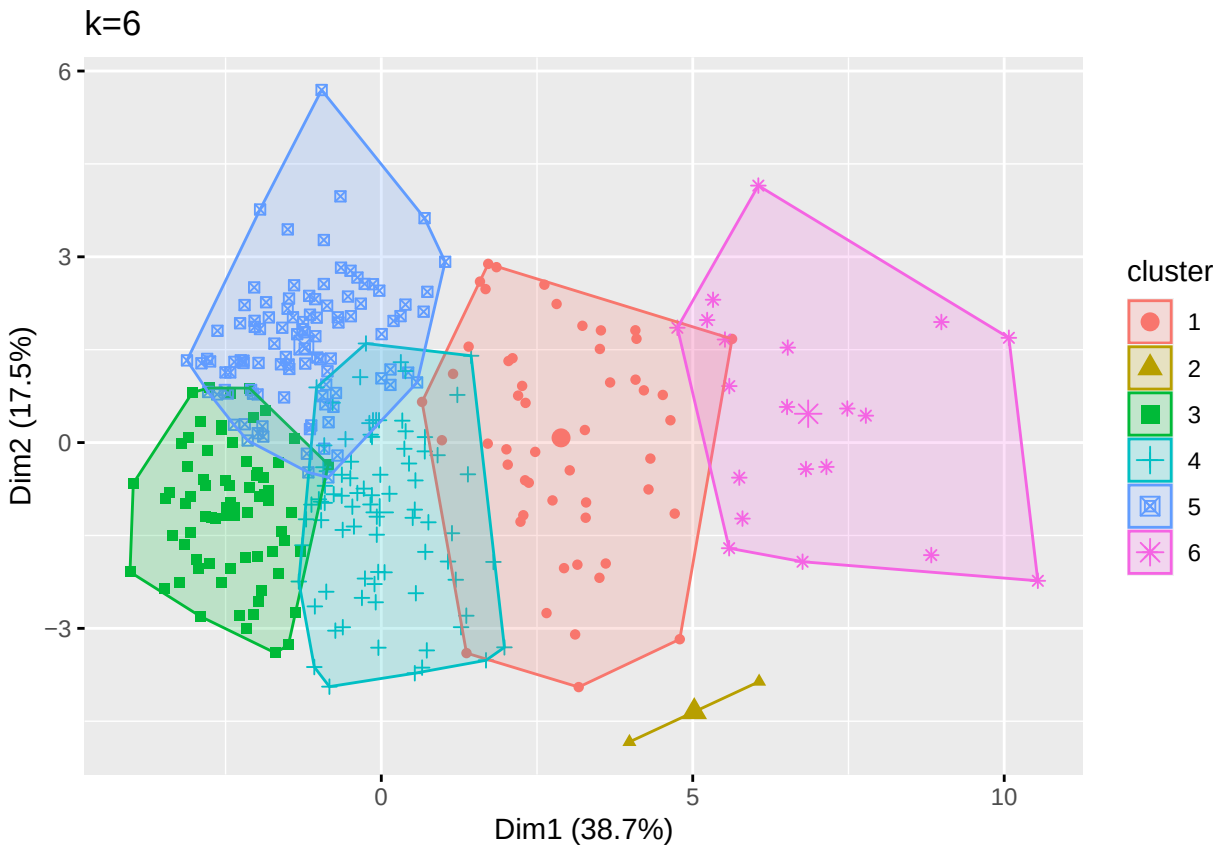

```
fviz_cluster(k4, data = cities,geom='point') + ggtitle("k=4")
```




```
fviz_cluster(k5, data = cities,geom='point') + ggtitle("k=5")
```




```
fviz_cluster(k6, data = cities,geom='point') + ggtitle("k=6")
```




## the best k is

```
cluster_number_k = as.factor(k4$cluster)
cities0$cluster_k = cluster_number_k
# the number of cities in each cluster.
table(cluster_number_k)
```

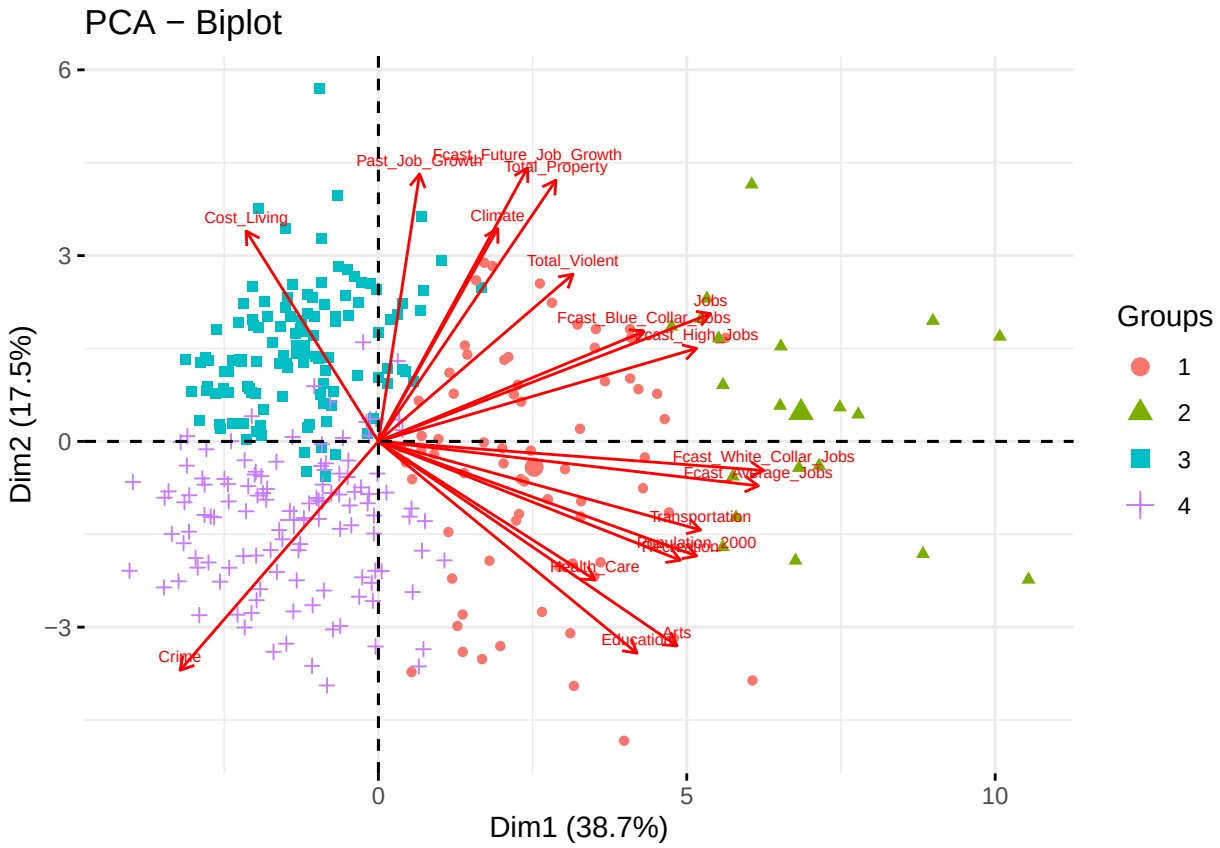
```
## cluster_number_k
##   1   2   3   4
##  69  20 116 120
```

```
aggregate( .~ cluster_k,FUN=median,data = cities0)
```

```
##   cluster_k Cost_Living Transportation   Jobs Education Climate  Crime  Arts
## 1         1      47.310         80.730 81.010    80.730   64.58 27.200 80.46
## 2         2      26.920         92.065 97.305    82.005   70.82 22.665 91.65
## 3         3      76.070         30.730 43.760    24.215   67.13 31.305 23.94
## 4         4      45.615         40.785 30.450    52.830   32.15 80.315 51.28
##   Health_Care Recreation Population_2000 Total_Violent Total_Property
## 1      76.480     78.750       1059044.0           696         5436.0
## 2      65.290     90.365       2818808.5           753         5878.5
## 3      29.175     23.790        179977.5           653         5472.0
## 4      43.055     46.880        227733.5           273         3645.0
##   Past_Job_Growth Fcast_Future_Job_Growth Fcast_Blue_Collar_Jobs
## 1            10.9                    5.9                3388.0
## 2            15.6                    8.3               20447.5
## 3            11.9                    6.0                 797.5
## 4             8.3                    4.8                 436.0
##   Fcast_White_Collar_Jobs Fcast_High_Jobs Fcast_Average_Jobs
```

```
## 1                33198.0          4976.0           23990.0
## 2               119533.5         23248.0           83826.0
## 3                 6020.0          1367.5            3721.0
## 4                 6518.5           796.5            4489.5
```

```
m = prcomp(cities, scale=T)
fviz_pca_biplot(m,labelsize = 2,col.var = "red",
                habillage = cluster_number_k,geom='point')
```

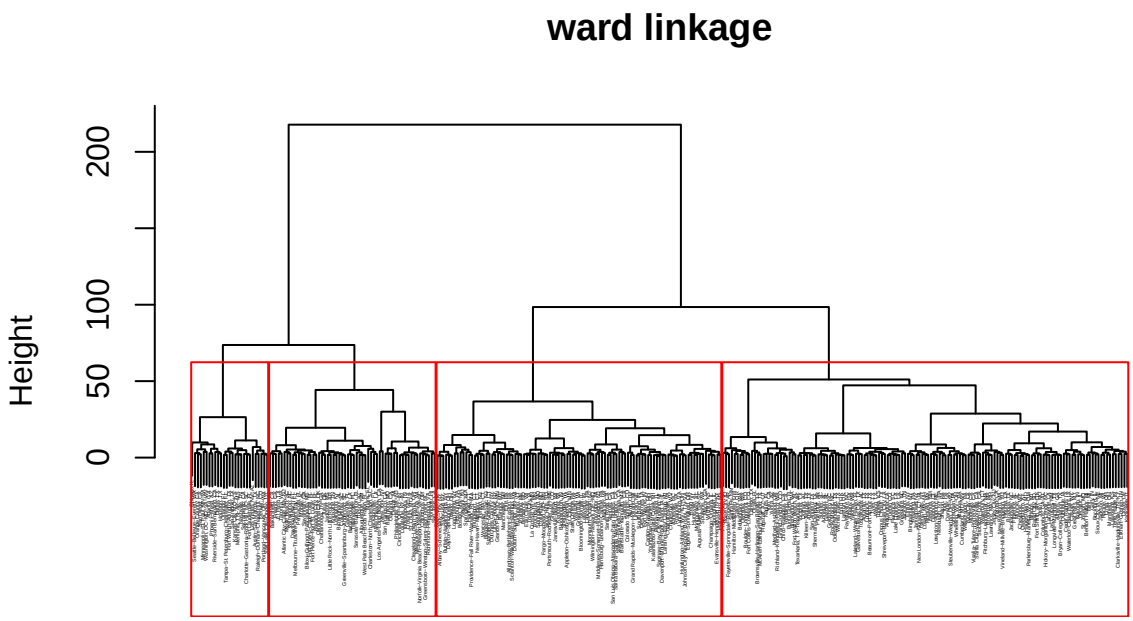



```
# Group (cluster) 1 has high rates on transportation, jobs, education, arts, health care,
# recreation and total violent, has low rates on crime.

# Group (cluster) 2 has low rates on cost living and crime, high rates on others.

# Group (cluster) 3 has high rates on cost living, climate, past job growth
#and fcast future job growth.
# has low rates on transprtation, jobs, education, Arts, health care, recreation and population 2000

# Group (cluster) 4 has low rates on transportation, jobs, climate, population 2000, total violent,
# past job growth and fcast future job growth, has high rates on crime
```

##HIERARCHICAL CLUSTERING

```
# display the dendrogram.
h1 <- hclust(dist, method = "ward.D")

plot(h1,cex=0.2,main="ward linkage")
```

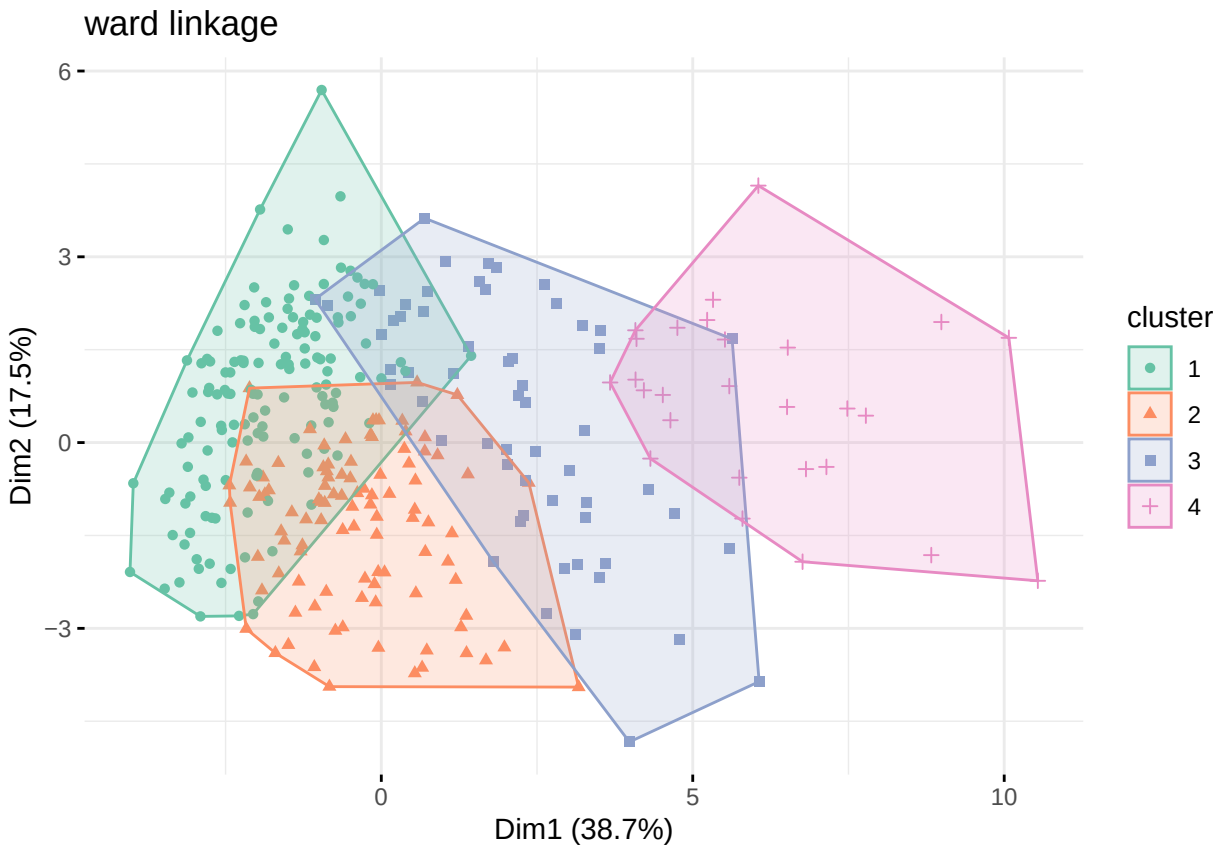
```
rect.hclust(h1,k=4,border='red')
```




```
#find the clusters
cut1 = cutree(h1, k = 4)

#Find the number of cities in each cluster.
table(cut1)
```

```
## cut1
##   1   2   3   4
## 141  99  58  27
```

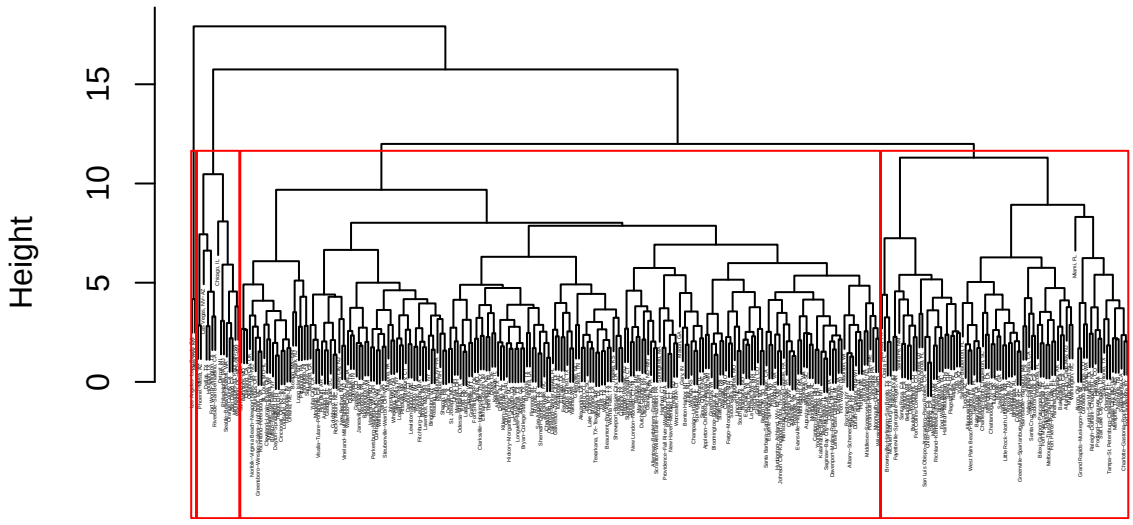
```
#display the cluster plots
fviz_cluster(list(data = cities, cluster = cut1),main="ward linkage",
             palette = "Set2",show.clust.cent = F, labelsize = 10,geom='point',
             ggtheme = theme_minimal()
              )
```




```
# Find the CCPC
c1 = cophenetic(h1)
cor(dist,c1)
```

```
## [1] 0.5079247
```

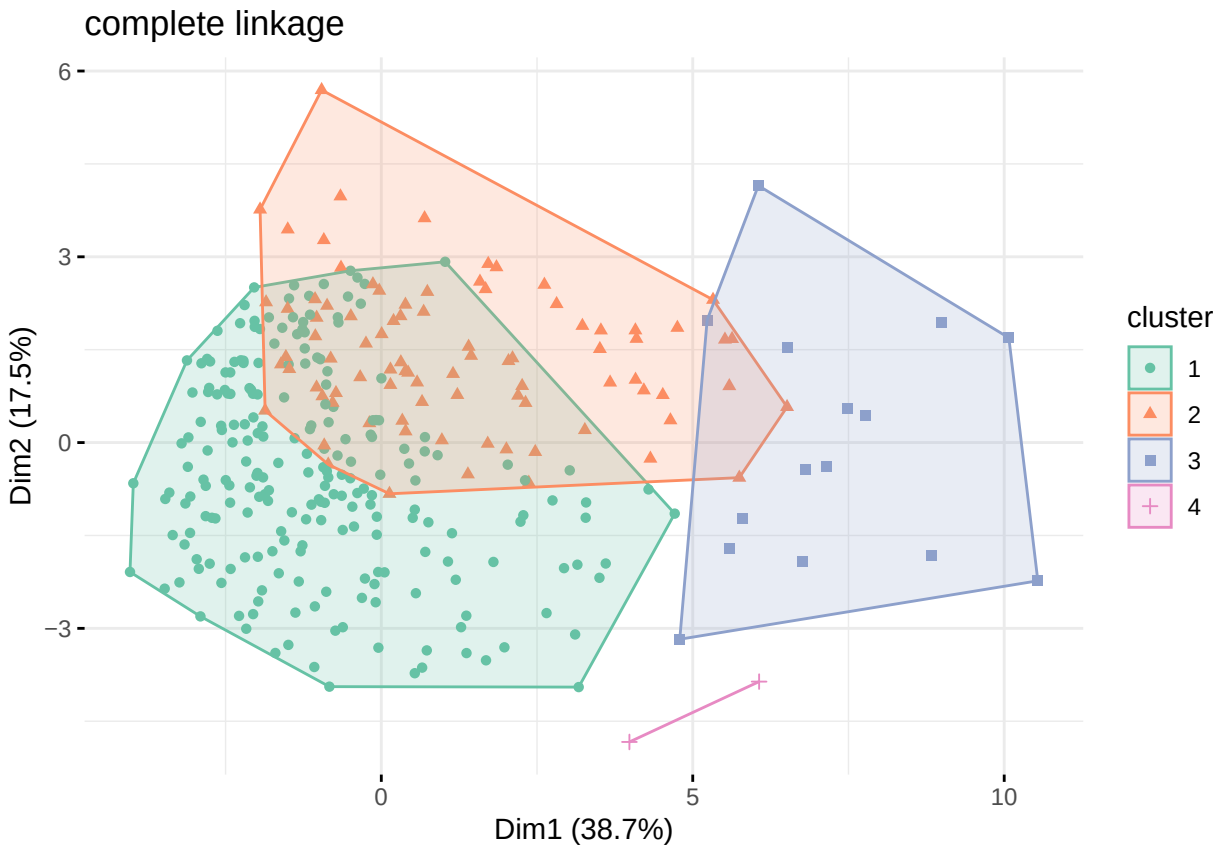
```
# display the dendrogram.
h2 <- hclust(dist, method = "complete")

plot(h2,cex=0.2,main="complete linkage")

rect.hclust(h2,k=4,border='red')
```




```
#find the clusters
cut2 = cutree(h2, k = 4)

#Find the number of cities in each cluster.
table(cut2)
```

```
## cut2
##   1   2   3   4
## 222  86  15   2
```

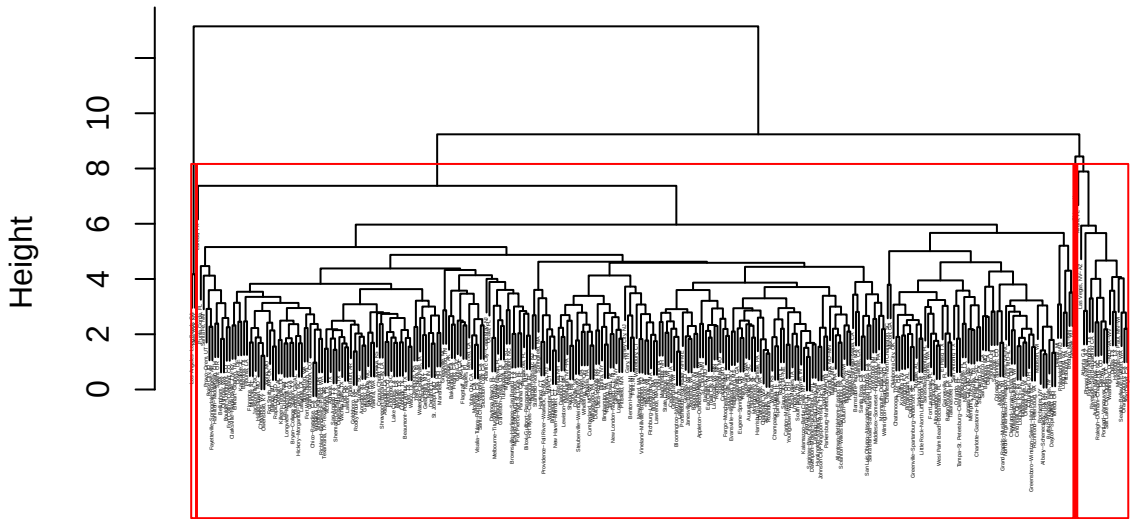
```
#display the cluster plots
fviz_cluster(list(data = cities, cluster = cut2),main="complete linkage",
             palette = "Set2",show.clust.cent = F, labelsize = 10,geom='point',
             ggtheme = theme_minimal()
              )
```




```
# Find the CCPC
c2 = cophenetic(h2)
cor(dist,c2)
```

```
## [1] 0.6848473
```

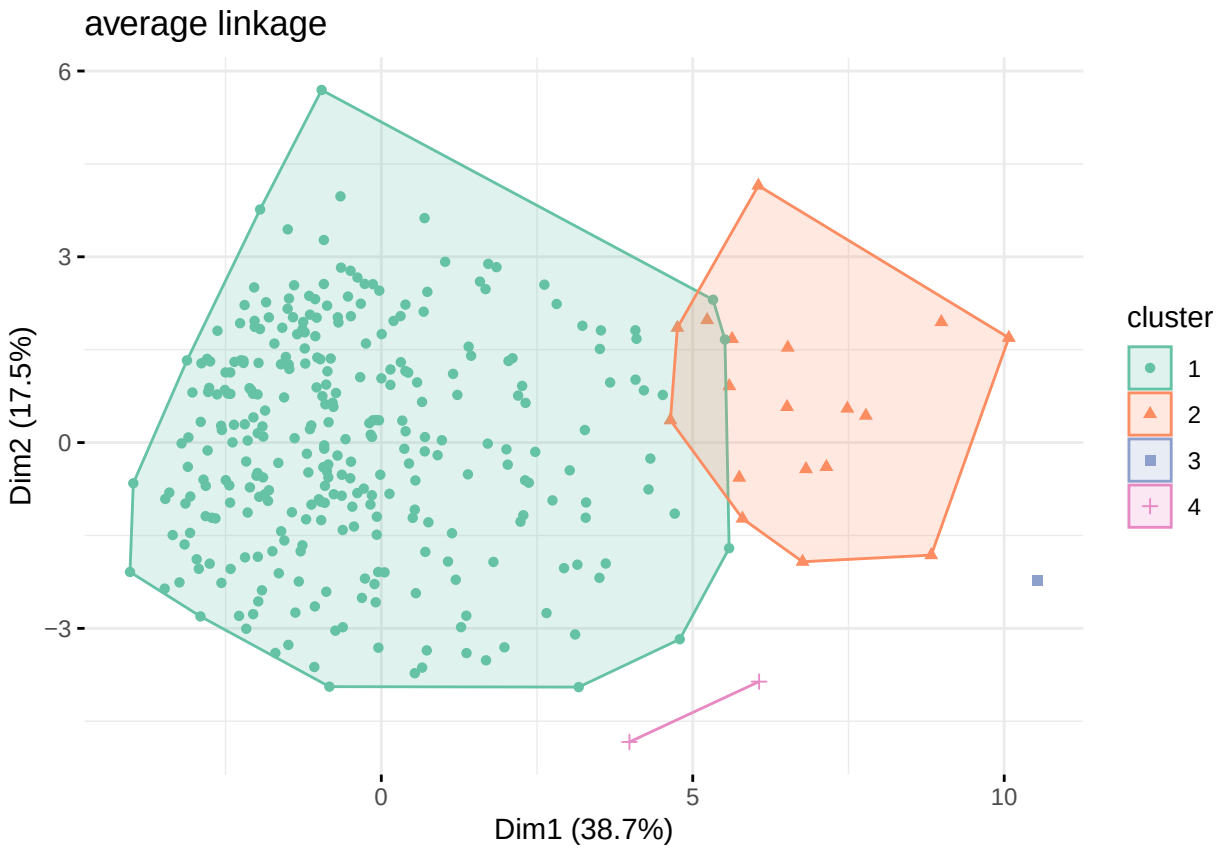
```
# display the dendrogram.
h3 <- hclust(dist, method = "average")

plot(h3,cex=0.2,main="average linkage")

rect.hclust(h3,k=4,border='red')
```




```
#find the clusters
cut3 = cutree(h3, k = 4)

#Find the number of cities in each cluster.
table(cut3)
```

```
## cut3
##   1   2   3   4
## 304  18   1   2
```

```
#display the cluster plots
fviz_cluster(list(data = cities, cluster = cut3),main="average linkage",
             palette = "Set2",show.clust.cent = F, labelsize = 10,geom='point',
             ggtheme = theme_minimal()
              )
```




```
# Find the CCPC
c3 = cophenetic(h3)
cor(dist,c3)
```

```
## [1] 0.8047003
```

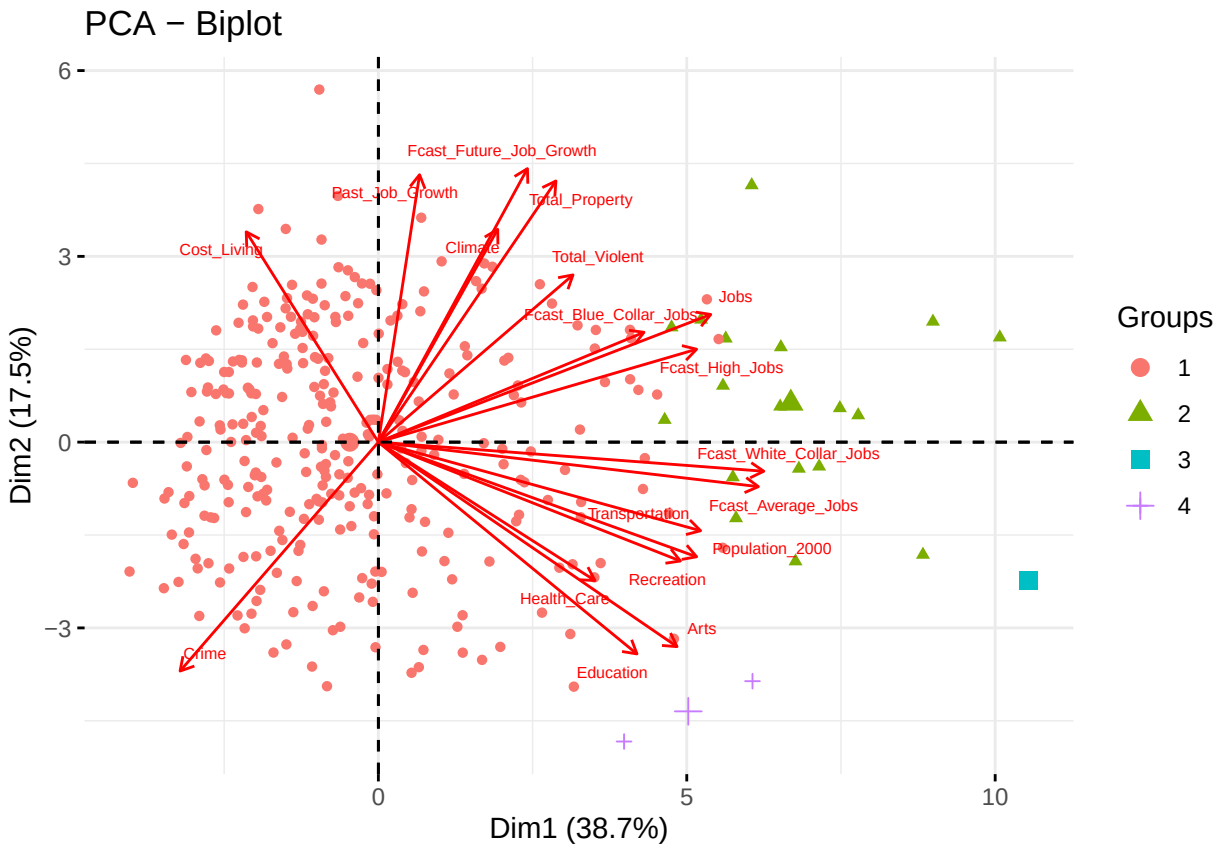
### I prefer the average linkage

```
cluster_number_h = as.factor(cut3)
cities0$cluster_h = cluster_number_h
# the number of cities in each cluster.
aggregate( .~ cluster_h,FUN=median,data = cities0)
```

```
##   cluster_h Cost_Living Transportation   Jobs Education Climate  Crime   Arts
## 1         1      55.670         45.180 49.145    47.445  51.555 50.570 46.040
## 2         2      26.920         91.355 97.445    83.845  71.245 29.045 91.365
## 3         3       9.350        100.000 86.960    98.860  16.140  2.270 99.160
## 4         4       2.835         96.455 45.035    85.830  84.840  0.855 99.720
##   Health_Care Recreation Population_2000 Total_Violent Total_Property
## 1       45.18     47.305          258587         531.5         4891.0
## 2       66.99     88.805         2567279         693.5         5878.5
## 3       81.30     97.160         7864846        1386.0         5676.0
## 4       80.02     92.490         8912152        1570.0         5082.0
##   Past_Job_Growth Fcast_Future_Job_Growth Fcast_Blue_Collar_Jobs
## 1            10.3                    5.60                 877.5
## 2            15.6                    8.85               20447.5
## 3             5.3                    4.40               21442.0
```

```
## 4           -6.1                  1.80              -32786.5
##   Fcast_White_Collar_Jobs Fcast_High_Jobs Fcast_Average_Jobs cluster_k
## 1                  8219.5         1483.0             5606.5         3
## 2                119533.5        25695.0            80787.5         2
## 3                195150.0        21334.0           170426.0         2
## 4                123941.5       -14965.5            98620.5         1
```

```
m2 = prcomp(cities, scale=T)
fviz_pca_biplot(m2,labelsize = 2,col.var = "red",repel = T,geom='point',
                habillage = cut3)
```




```
# Group (cluster) 1 has high rates on cost living, crime, past job growth
# has low rates on transportation, jobs, education, arts, health care, recreation,
# population 2000, total violent, total property.

# Group (cluster) 2 has low rates on total violent.
# has high rates on jobs, climate, total property, past job growth, fcast future job growth.


# Group (cluster) 3 has low rates on cost living, climate, crime, past job growth
# and fcast future job growth.
# has high rates on transprtation, jobs, education, arts, health care, recreation,
# population 2000 and total violent.

# Group (cluster) 4 has low rates on cost living, jobs, crime, past job growth
# and fcast future job growth,

# has high rates on transportation, cilmate, arts, health care, recreation,
```

```
# population 2000 and total violent,
```