```
df = read.csv("cities1.csv")
# Metropolitan_Area becomes the row names of dataframe
rownames(df) <- df[,1]
# Remove Crime_Trend and Unemployment_Threat
df0 = df[,-c(1,14,15)]
df1 = scale(df0)
distance = dist(df1)
head(distance)
```

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

```
distmat = as.matrix(distance)
```

K-means Clustering

```
# Question 1
library(cluster)
library(factoextra)
```

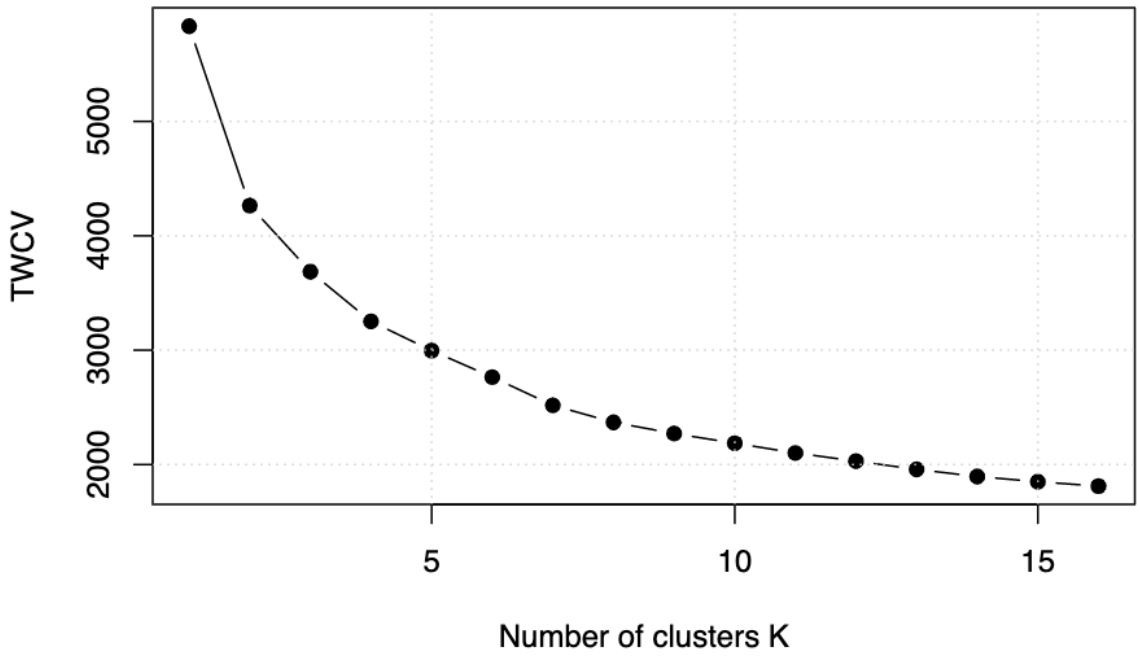
```
## Loading required package: ggplot2
```

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

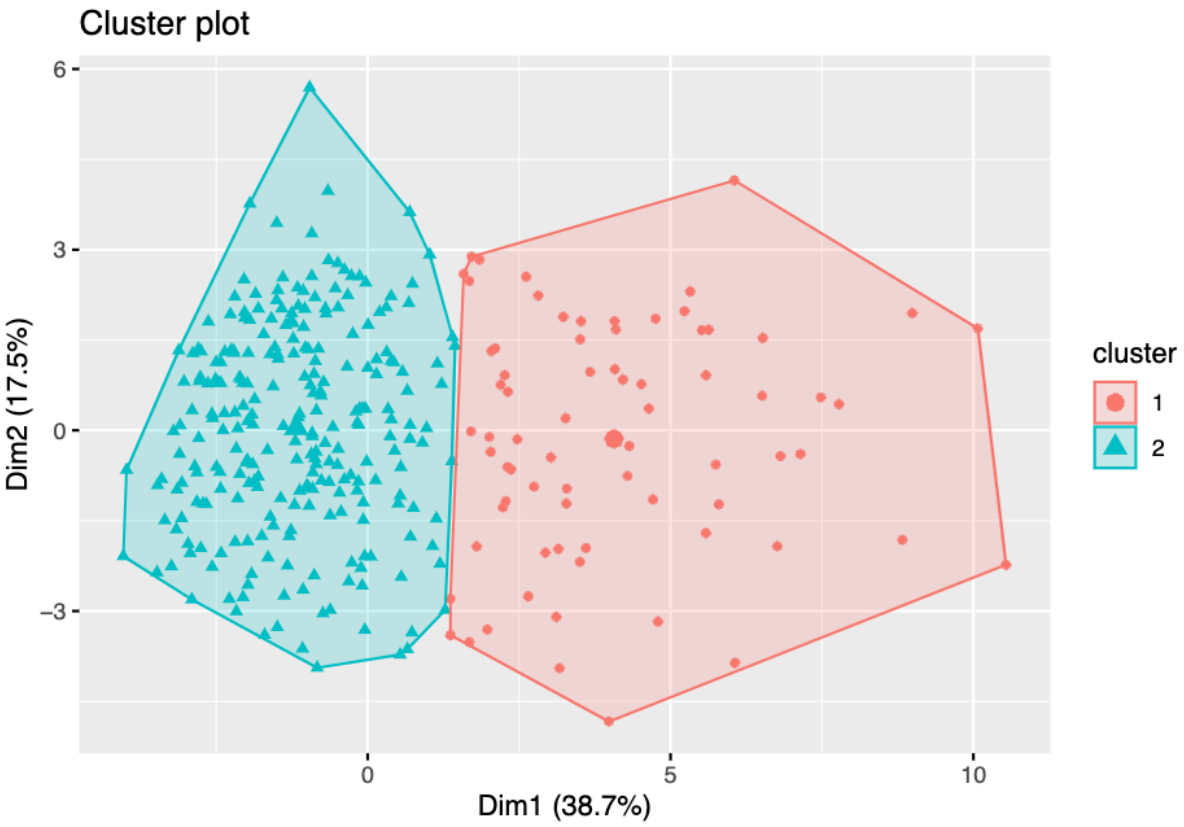
```
set.seed(123)
twcv = function(k) kmeans(df1, k, nstart = 25)$tot.withinss
k = 1:16
twcv_values = sapply(k,twcv)
head(twcv_values)
```

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

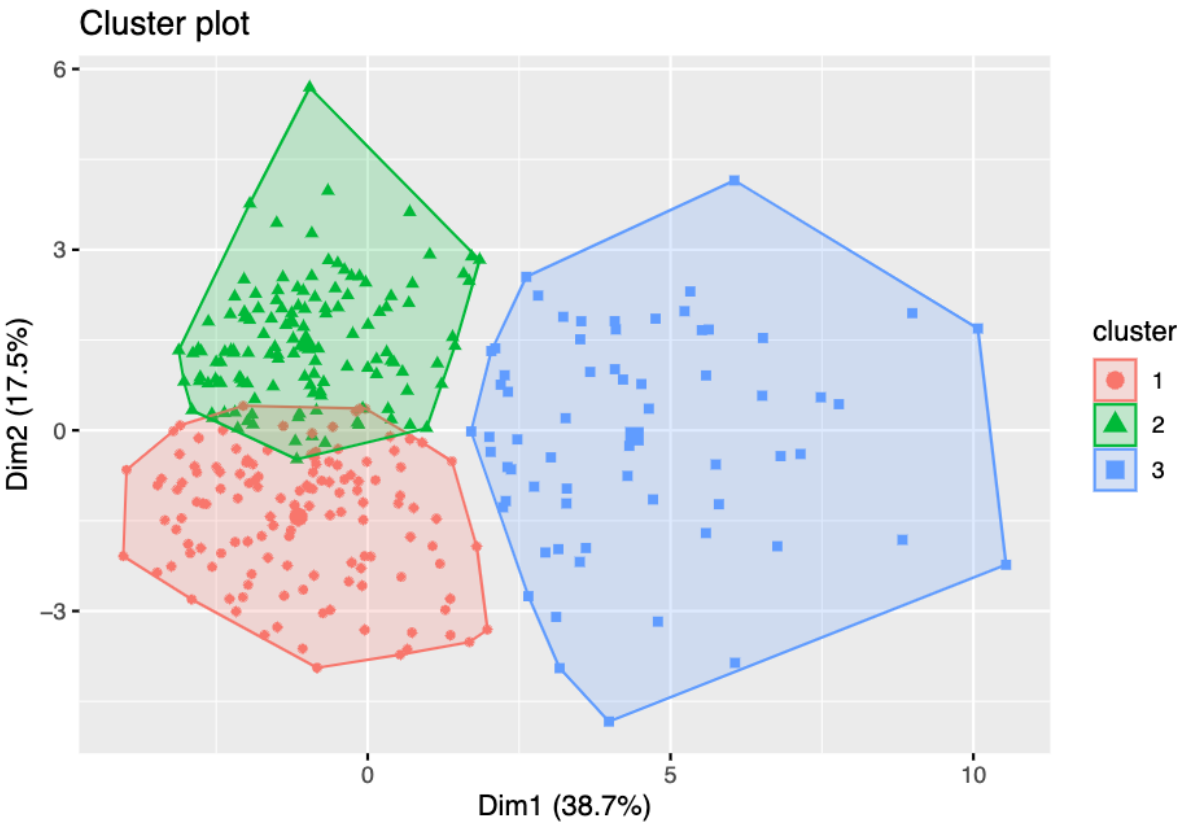
```
plot(k, twcv_values,type="b",pch = 19, xlab="Number of clusters K",ylab="TWCV")
grid()
```



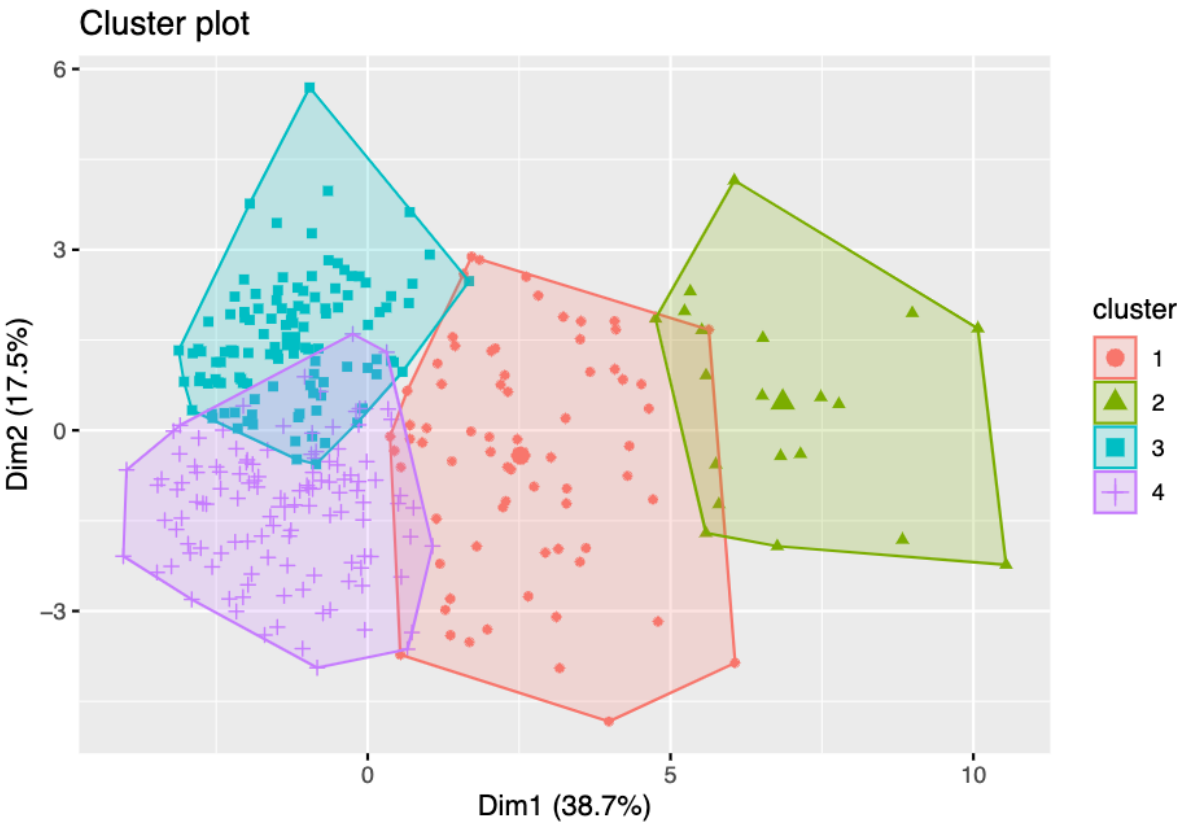

```r
# Question 2
k2 = kmeans(df1, centers = 2, nstart = 25)
k3 = kmeans(df1, centers = 3, nstart = 25)
k4 = kmeans(df1, centers = 4, nstart = 25)
k5 = kmeans(df1, centers = 5, nstart = 25)
fviz_cluster(k2, geom = "point", data = df1)
```




```
fviz_cluster(k3, geom = "point",  data = df1)
```




```
fviz_cluster(k4, geom = "point",  data = df1)
```




```
fviz_cluster(k5, geom = "point",  data = df1)
```


```
# The best K is 4.

table(k4$cluster)
```

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

```
# Question 3
cluster_number = as.factor(k4$cluster)
df0$cluster = cluster_number
aggregate( .~ cluster,FUN=median,data = df0)
```

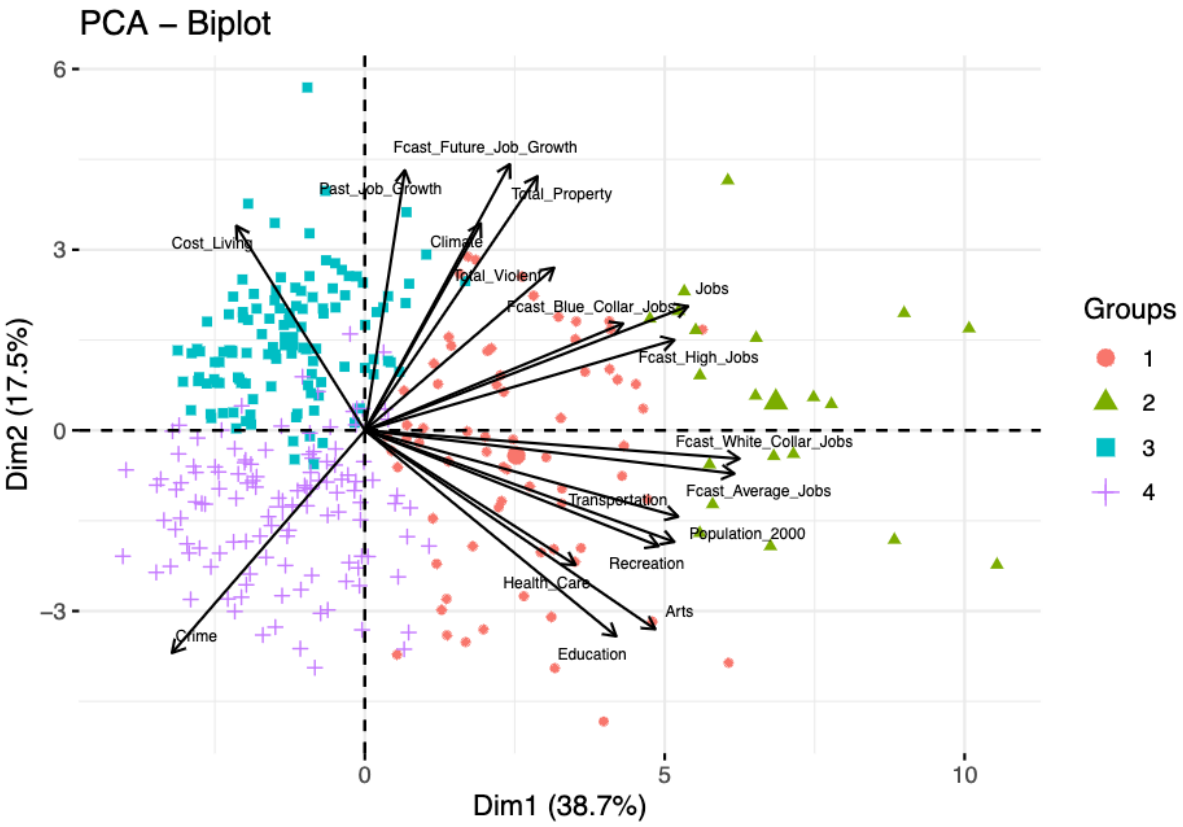
```
##   cluster Cost_Living Transportation   Jobs Education Climate  Crime  Arts
## 1       1      47.310         80.730 81.010    80.730   64.58 27.200 80.46
## 2       2      26.920         92.065 97.305    82.005   70.82 22.665 91.65
## 3       3      76.070         30.730 43.760    24.215   67.13 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
```

```
aggregate( .~ cluster,FUN=mean,data = df0)
```

```
##   cluster Cost_Living Transportation     Jobs Education  Climate    Crime
## 1       1    44.18913       80.03000 76.64609  76.58942 59.03420 29.96333
## 2       2    32.44100       91.03000 95.76050  78.33750 66.86550 27.27150
## 3       3    67.17103       31.68267 43.63336  27.32681 63.08750 35.04190
## 4       4    44.84208       40.55317 35.97708  51.70667 34.86483 76.91042
##       Arts Health_Care Recreation Population_2000 Total_Violent Total_Property
## 1 78.26971    72.50797   77.83348       1409618.9      783.3043       5708.725
## 2 87.66850    64.32950   89.04550       2932061.5      752.0000       5867.400
## 3 26.01345    32.06491   31.26371        239501.6      664.8190       5536.284
## 4 49.44792    46.21792   46.21342        294146.8      314.8000       3716.425
##   Past_Job_Growth Fcast_Future_Job_Growth Fcast_Blue_Collar_Jobs
## 1        9.871014                6.133333               2668.507
## 2       14.300000                8.840000              21047.900
## 3       12.576724                6.415517               1051.629
## 4        8.729167                4.958333                594.225
##   Fcast_White_Collar_Jobs Fcast_High_Jobs Fcast_Average_Jobs
## 1               39565.420        4979.000          29492.304
## 2              120170.100       25333.750          87683.400
## 3                7117.819        1616.767           4586.259
## 4                7928.308        1067.175           5776.775
```

```
prcomp1 = prcomp(df1, scale=T)
fviz_pca_biplot(prcomp1, label = "var",
                labelsize = 2,col.var = "black",
                habillage = cluster_number, repel = TRUE)
```



From median table,

Group (cluster) 1 has largest value in "Health_Care".

Group (cluster) 2 has largest value in "Transportation", "Jobs", "Education", "Climate", "Arts", "Recreation", "Population_2000", "Total_Violent", "Total_Property", "Past_Job_Growth", "Fcast_Future_Job_Growth", "Fcast_Blue_Collar_Jobs", "Fcast_White_Collar_Jobs", "Fcast_High_Jobs" and "Fcast_Average_Jobs"; has smallest value in "Cost_Living", "Crime".

Group (cluster) 3 has largest value in "Cost_Living"; has smallest value in "Transportation", "Education", "Arts", "Health_Care", "Recreation", "Population_2000", "Fcast_White_Collar_Jobs", "Fcast_Average_Jobs".

Group (cluster) 4 has largest value in "Crime"; has smallest value in "Jobs", "Climate", "Total_Violent", "Total_Property", "Past_Job_Growth", "Fcast_Future_Job_Growth", "Fcast_Blue_Collar_Jobs", "Fcast_High_Jobs".

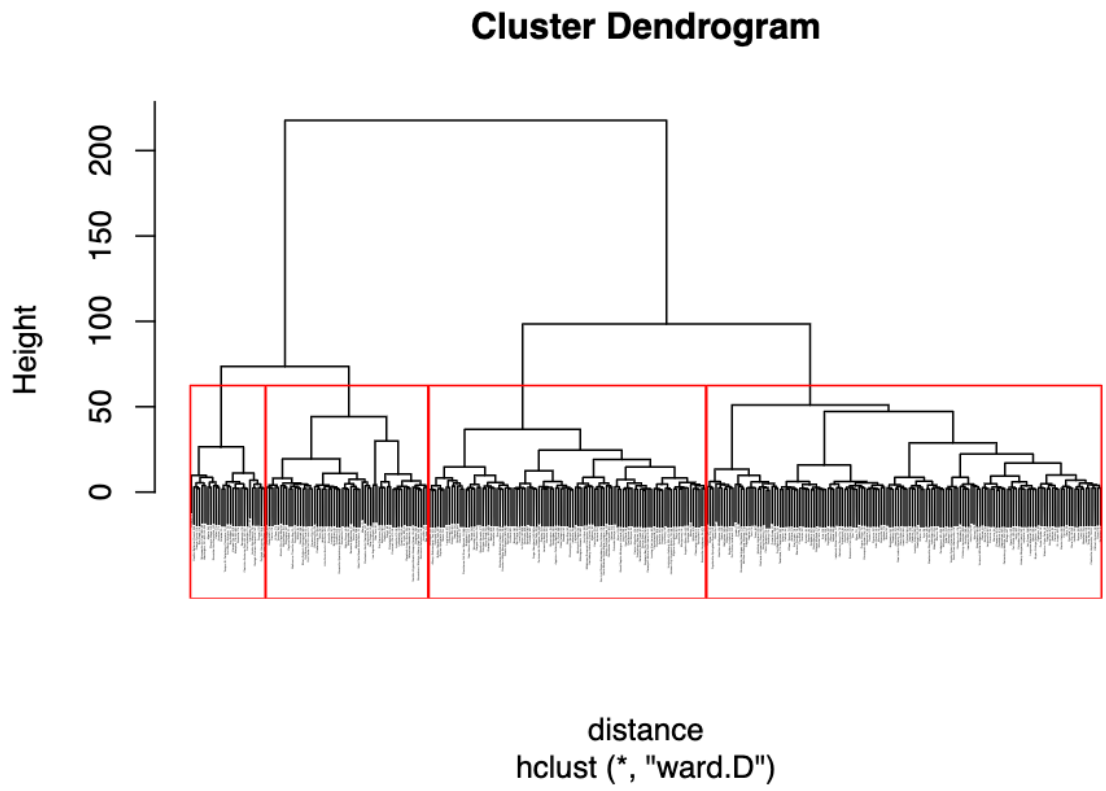
Hierarchical Clustering

```
# Question 4
h1 = hclust(distance, method ='ward.D')
str(h1)
```

```
## List of 7
##  $ merge      : int [1:324, 1:2] -58 -223 -56 -4 -14 -73 -35 -114 -41 -258 ...
```

```
##  $ height     : num [1:324] 0.975 1.181 1.225 1.282 1.308 ...
##  $ order      : int [1:325] 57 271 217 260 192 308 16 225 244 69 ...
##  $ labels     : chr [1:325] "Abilene, TX" "Akron, OH" "Albany, GA" "Albany-Schenectady-Troy, NY" ...
##  $ method     : chr "ward.D"
##  $ call       : language hclust(d = distance, method = "ward.D")
##  $ dist.method: chr "euclidean"
##  - attr(*, "class")= chr "hclust"
```

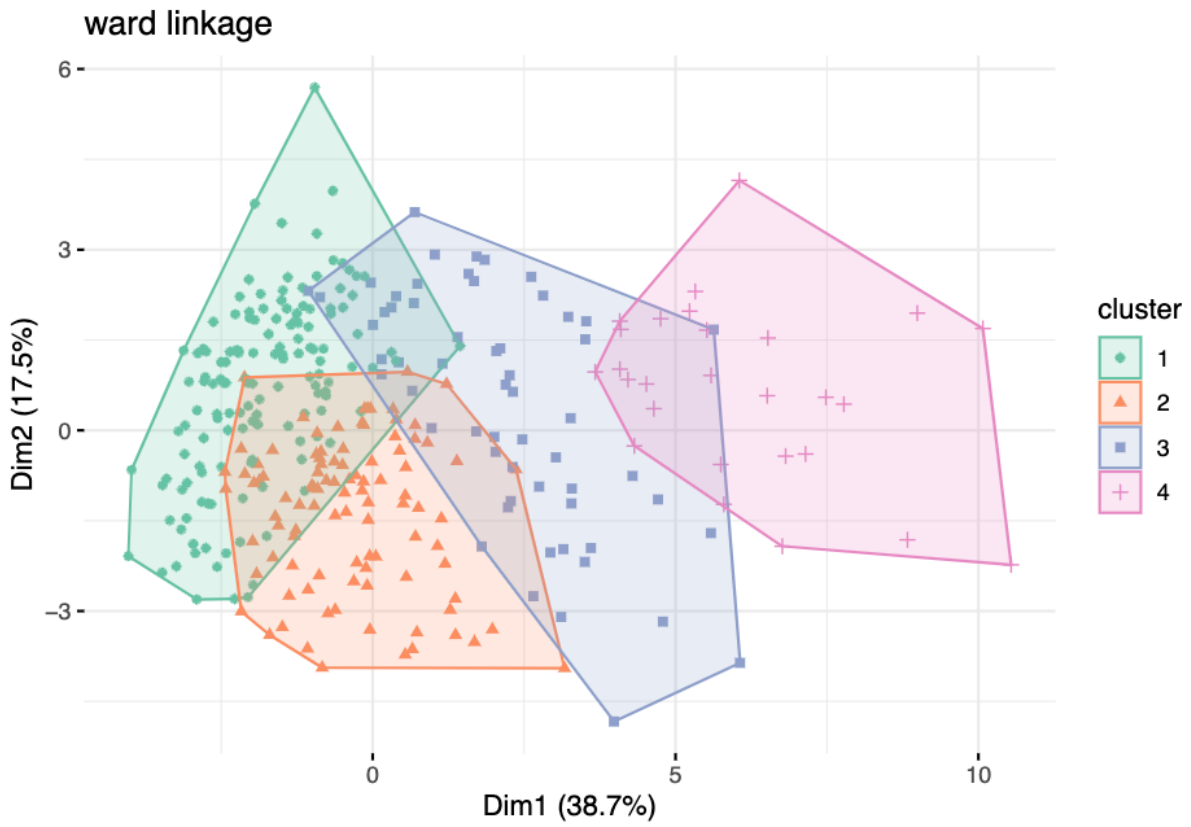
```
plot(h1,cex=0.1)
rect.hclust(h1,k=4,border="red")
```




```
cut1 = cutree(h1,k=4)
table(cut1)
```

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

```
fviz_cluster(list(data = df1, cluster = cut1),main="ward linkage",
             palette = "Set2",show.clust.cent = F, labelsize = 10,
             repel = T,
             ggtheme = theme_minimal(), geom = "point"
             )
```



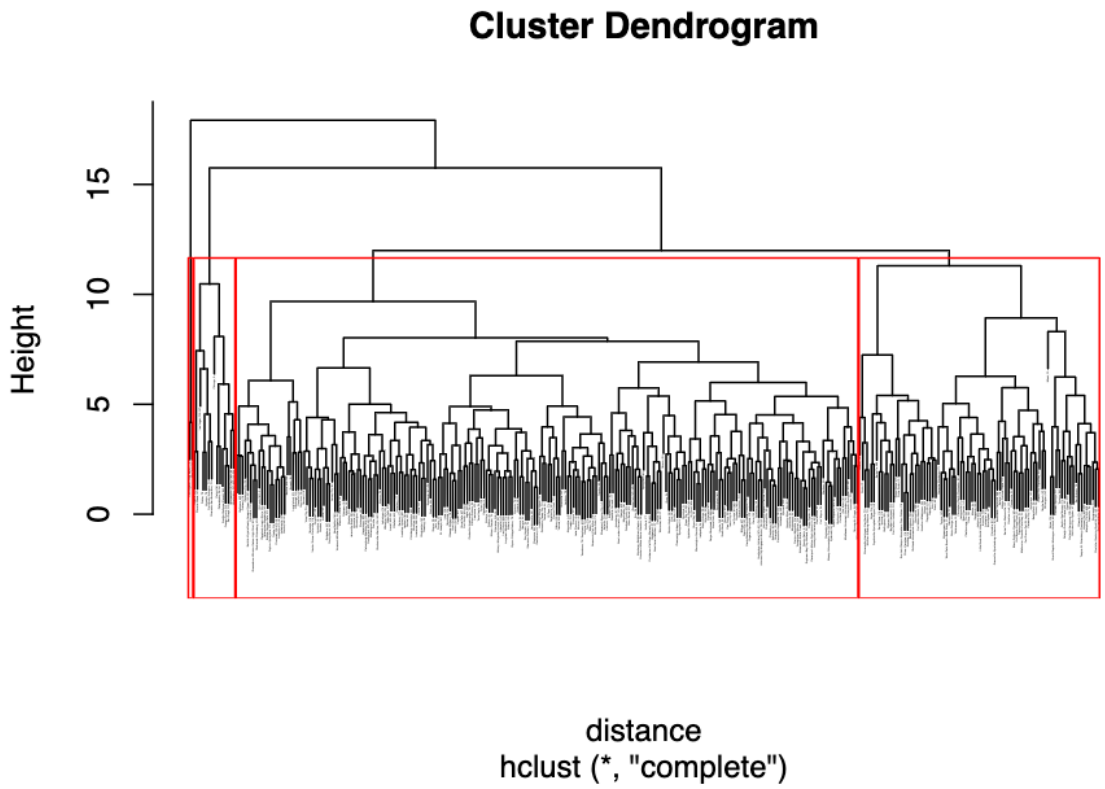

```
c1 = cophenetic(h1)
CPCC1 = cor(distance,c1)
CPCC1
```

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

```
# Question 5
h2 = hclust(distance, method ='complete')
str(h2)
```

```
## List of 7
##  $ merge      : int [1:324, 1:2] -58 -223 -56 -4 -14 -73 -35 -114 -41 -258 ...
##  $ height     : num [1:324] 0.975 1.181 1.225 1.282 1.308 ...
##  $ order      : int [1:325] 175 207 16 225 164 69 132 107 244 57 ...
##  $ labels     : chr [1:325] "Abilene, TX" "Akron, OH" "Albany, GA" "Albany-Schenectady-Troy, NY" ...
##  $ method     : chr "complete"
##  $ call       : language hclust(d = distance, method = "complete")
##  $ dist.method: chr "euclidean"
##  - attr(*, "class")= chr "hclust"
```

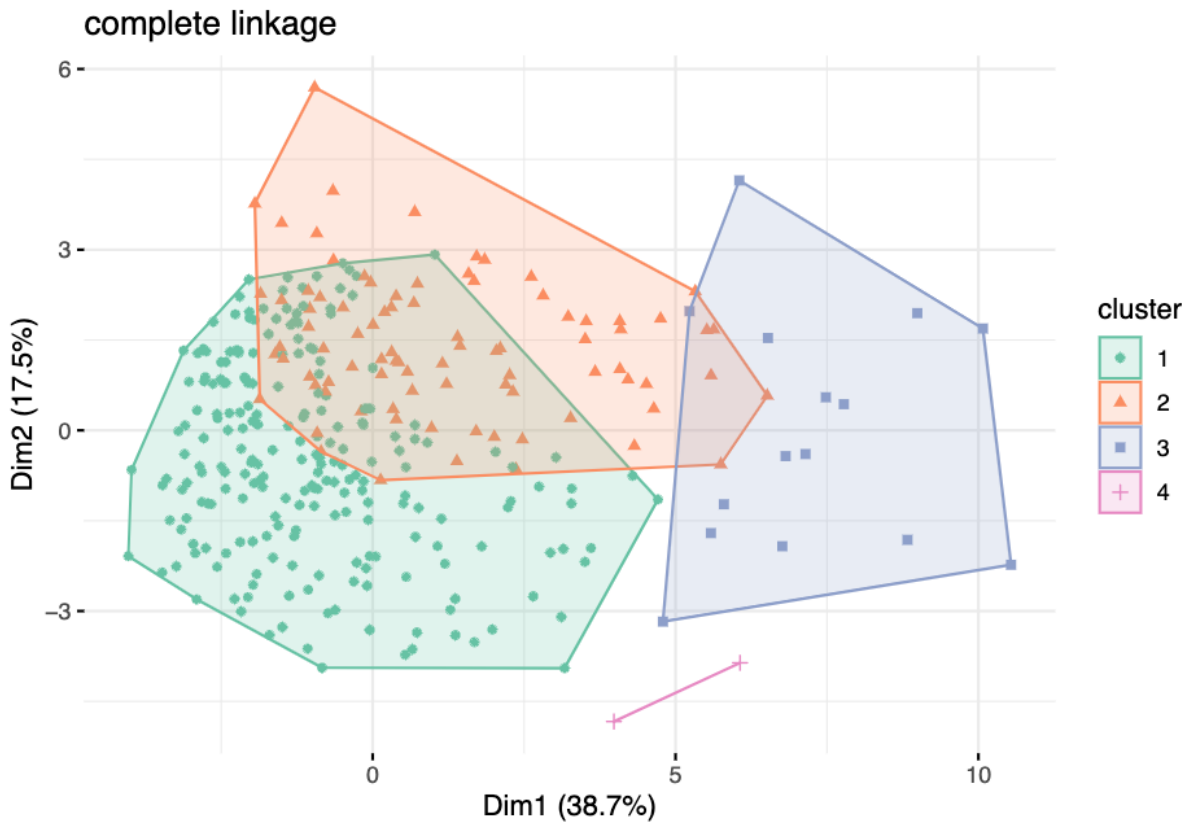
```
plot(h2,cex=0.1)
rect.hclust(h2,k=4,border="red")
```




```
cut2 = cutree(h2,k=4)
table(cut2)
```

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

```
fviz_cluster(list(data = df1, cluster = cut2),main="complete linkage",
             palette = "Set2",show.clust.cent = F, labelsize = 10,
             repel = T,
             ggtheme = theme_minimal(), geom = "point"
             )
```



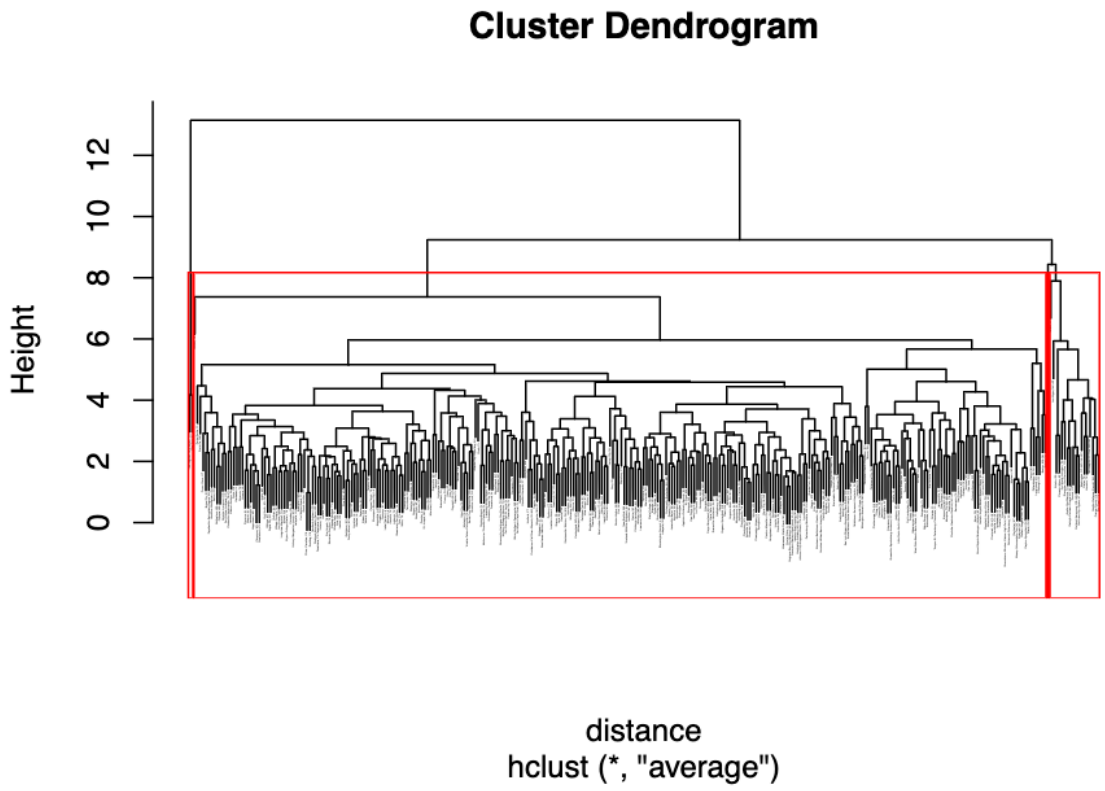

```
c2 = cophenetic(h2)
CPCC2 = cor(distance,c2)
CPCC2
```

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

```
# Question 6
h3 = hclust(distance, method ='average')
str(h3)
```

```
## List of 7
##  $ merge      : int [1:324, 1:2] -58 -223 -56 -4 -14 -73 -35 -114 -41 -247 ...
##  $ height     : num [1:324] 0.975 1.181 1.225 1.282 1.308 ...
##  $ order      : int [1:325] 175 207 162 235 266 233 36 94 125 255 ...
##  $ labels     : chr [1:325] "Abilene, TX" "Akron, OH" "Albany, GA" "Albany-Schenectady-Troy, NY" ...
##  $ method     : chr "average"
##  $ call       : language hclust(d = distance, method = "average")
##  $ dist.method: chr "euclidean"
##  - attr(*, "class")= chr "hclust"
```

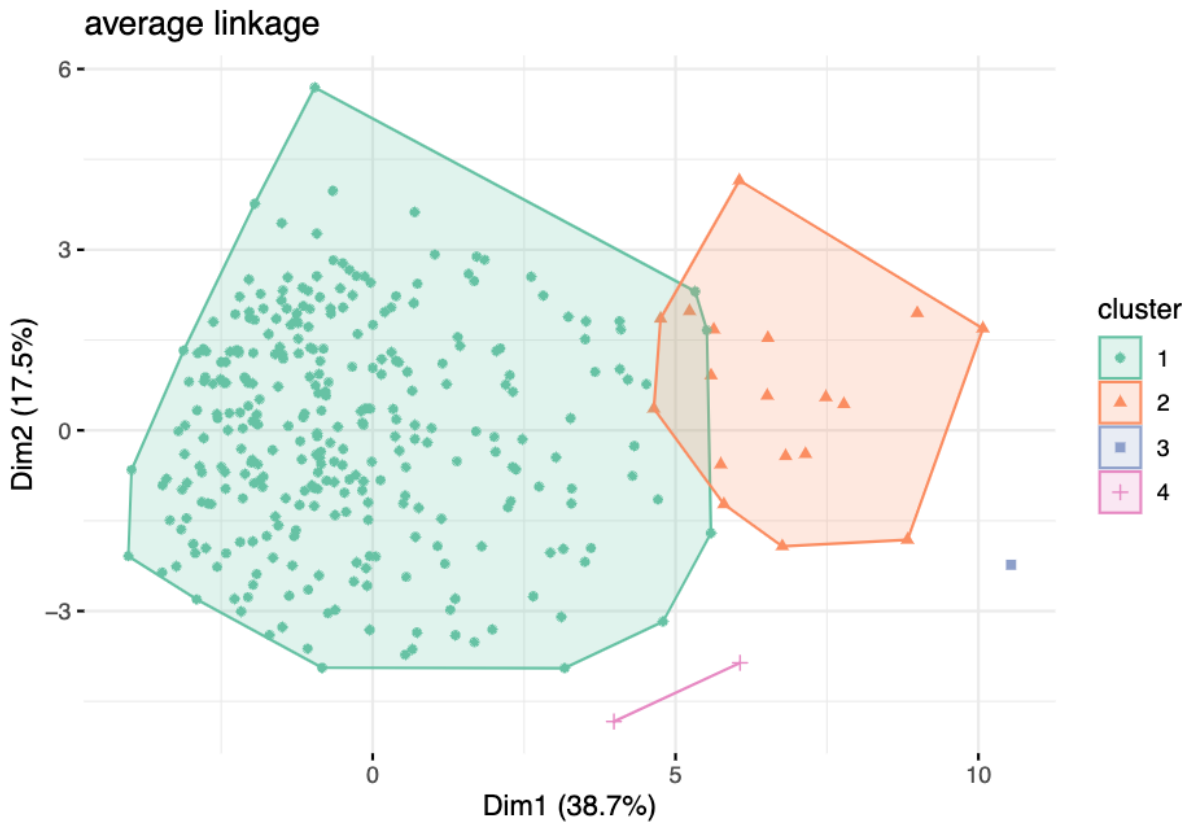
```
plot(h3,cex=0.1)
rect.hclust(h3,k=4,border="red")
```




```
cut3 = cutree(h3,k=4)
table(cut3)
```

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

```
fviz_cluster(list(data = df1, cluster = cut3),main="average linkage",
             palette = "Set2",show.clust.cent = F, labelsize = 10,
             repel = T,
             ggtheme = theme_minimal(), geom = "point"
             )
```




```
c3 = cophenetic(h3)
CPCC3 = cor(distance,c3)
CPCC3
```

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

```
# Question 7
# I prefer average linkage
# because its cluster plot has fewer overlaps
# than the cluster plots of ward linkage and complete linkage.
# Also the CPCC value of the average linkage
# is the highest (0.8047003) among the three.
cluster_number3 = as.factor(cut3)
df0$cluster = cluster_number3
aggregate(.~ cluster,FUN=median,data = df0)
```

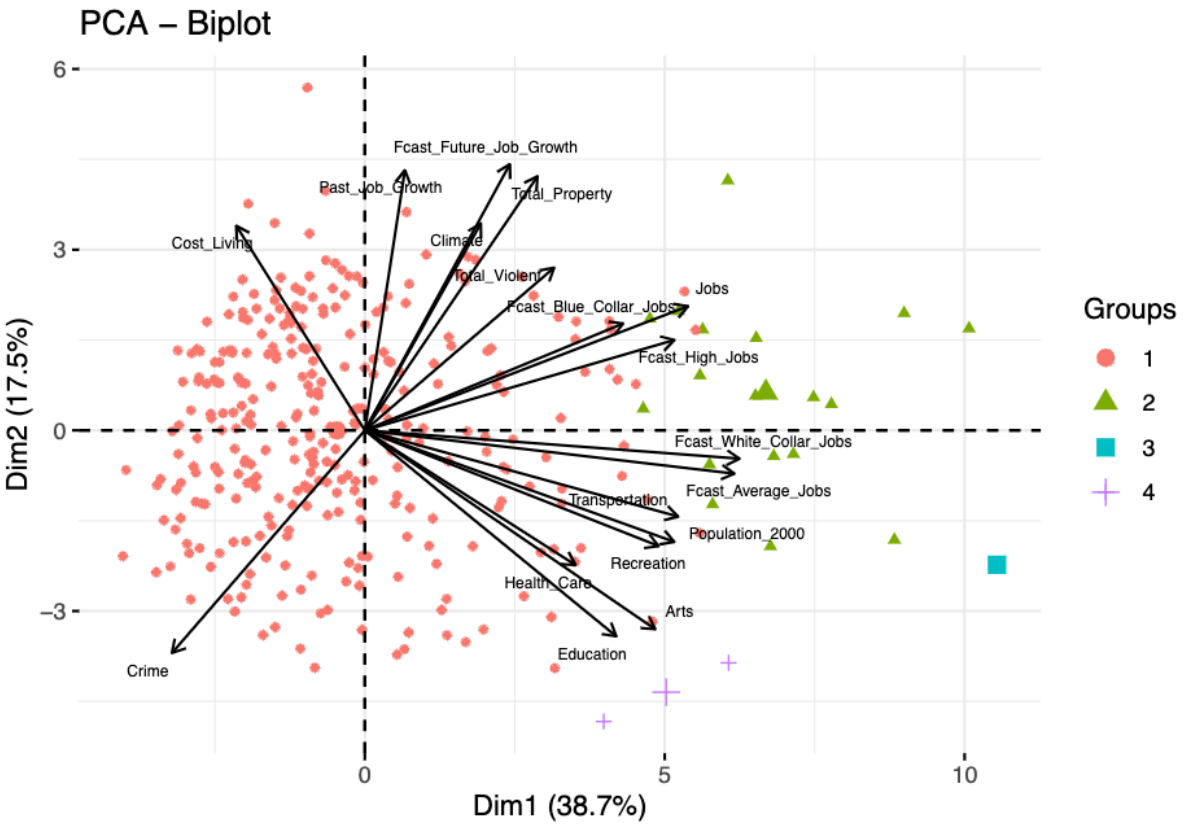
```
##   cluster Cost_Living Transportation   Jobs Education Climate  Crime   Arts
## 1       1      55.670         45.180 49.145    47.445  51.555 50.570 46.040
## 2       2      26.920         91.355 97.445    83.845  71.245 29.045 91.365
## 3       3       9.350        100.000 86.960    98.860  16.140  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
## 1                  8219.5          1483.0             5606.5
## 2                119533.5         25695.0            80787.5
## 3                195150.0         21334.0           170426.0
## 4                123941.5        -14965.5            98620.5
```

```
aggregate(.~ cluster,FUN=mean,data = df0)
```

```
##   cluster Cost_Living Transportation     Jobs Education  Climate    Crime
## 1       1    53.72694       45.93957 48.25332  47.72493 50.82822 50.48398
## 2       2    29.04167       90.31667 96.47056  80.40167 70.83333 30.88333
## 3       3     9.35000      100.00000 86.96000  98.86000 16.14000  2.27000
## 4       4     2.83500       96.45500 45.03500  85.83000 84.84000  0.85500
##       Arts Health_Care Recreation Population_2000 Total_Violent Total_Property
## 1 46.76766    46.38872   47.65615        485867.4      547.1184       4850.095
## 2 88.29556    67.32278   86.33389       2532884.1      729.6111       6064.500
## 3 99.16000    81.30000   97.16000       7864846.0     1386.0000       5676.000
## 4 99.72000    80.02000   92.49000       8912152.0     1570.0000       5082.000
##   Past_Job_Growth Fcast_Future_Job_Growth Fcast_Blue_Collar_Jobs
## 1        10.55197                5.809868                1528.414
## 2        15.14444                9.166667               20993.000
## 3         5.30000                4.400000               21442.000
## 4        -6.10000                1.800000              -32786.500
##   Fcast_White_Collar_Jobs Fcast_High_Jobs Fcast_Average_Jobs
## 1                 14523.1        2290.997            10507.10
## 2                114023.6       26553.667            80668.44
## 3                195150.0       21334.000           170426.00
## 4                123941.5      -14965.500            98620.50
```

```
prcomp1 = prcomp(df1, scale=T)
fviz_pca_biplot(prcomp1, label = "var", labelsize = 2,col.var = "black",
                habillage = cluster_number3, repel = T)
```



From median table,

Group (cluster) 1 has largest value in “Cost_Living” and “Crime”; has smallest value in “Transportation”, “Education”, “Arts”, “Health_Care”, “Recreation”, “Population_2000”, “Total_Violent”, “Total_Property”, “Fcast_White_Collar_Jobs”, “Fcast_Average_Jobs”.

Group (cluster) 2 has largest value in “Jobs”, “Total_Property”, “Past_Job_Growth”, “Fcast_Future_Job_Growth”, “Fcast_High_Jobs”.

Group (cluster) 3 has largest value in “Transportation”, “Education”, “Health_Care”, “Recreation”, “Fcast_Blue_Collar_Jobs”, “Fcast_White_Collar_Jobs” and “Fcast_Average_Jobs”; has smallest value in “Climate”.

Group (cluster) 4 has largest value in “Climate”, “Arts”, “Population_2000”, “Total_Violent”; has smallest value in “Cost_Living”, “Jobs”, “Crime”, “Past_Job_Growth”, “Fcast_Future_Job_Growth”, “Fcast_Blue_Collar_Jobs”, “Fcast_High_Jobs”.