535_Midterm

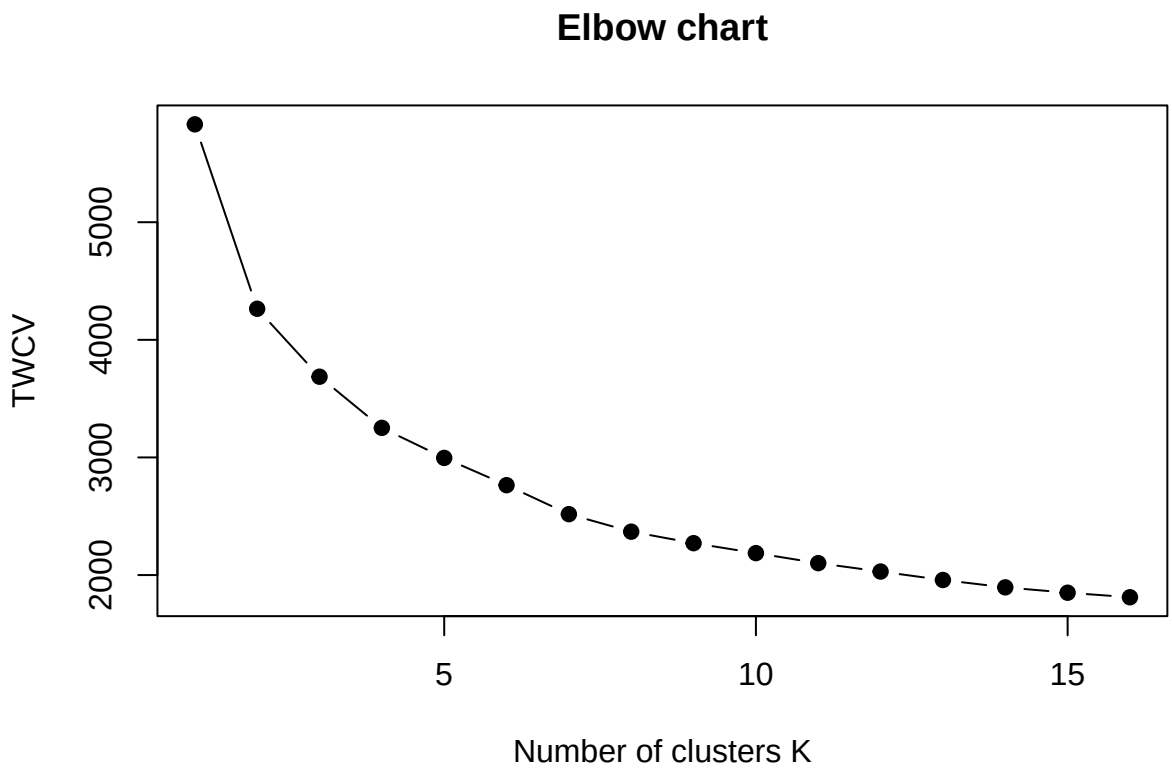
Shijie Gao, USC ID:6037-6293-25

2023-02-22

```
setwd("C:/Users/GAOSHIJIE/Desktop")
library(openxlsx)
d = read.xlsx("cities1.xlsx", sheet = 1)

#head(d)
library(cluster)
library(ggplot2)
library(factoextra)

#preprocessing
d$Crime_Trend = NULL
d$Unemployment_Threat = NULL
rownames(d) = d$Metropolitan_Area
d$Metropolitan_Area = NULL
#head(d)

df = scale(d) #make all attributes scaled
#head(df)

#1 K-MEANS CLUSTERING

#(1)
set.seed(123)

#TWCV for K = 1:16
twcv = function(k) kmeans(df, k, nstart = 25)$tot.withinss
k = 1:16
twcv_values = sapply(k, twcv)
twcv_values
```

```
##  [1] 5832.000 4264.026 3686.381 3251.504 2996.052 2764.309 2517.799 2368.927
##  [9] 2271.226 2186.526 2101.381 2029.885 1958.112 1895.399 1850.013 1811.432
```

```
#elbow chart
plot(k, twcv_values, type = "b", pch = 19, xlab = "Number of clusters K",
     ylab = "TWCV", main = "Elbow chart")
```

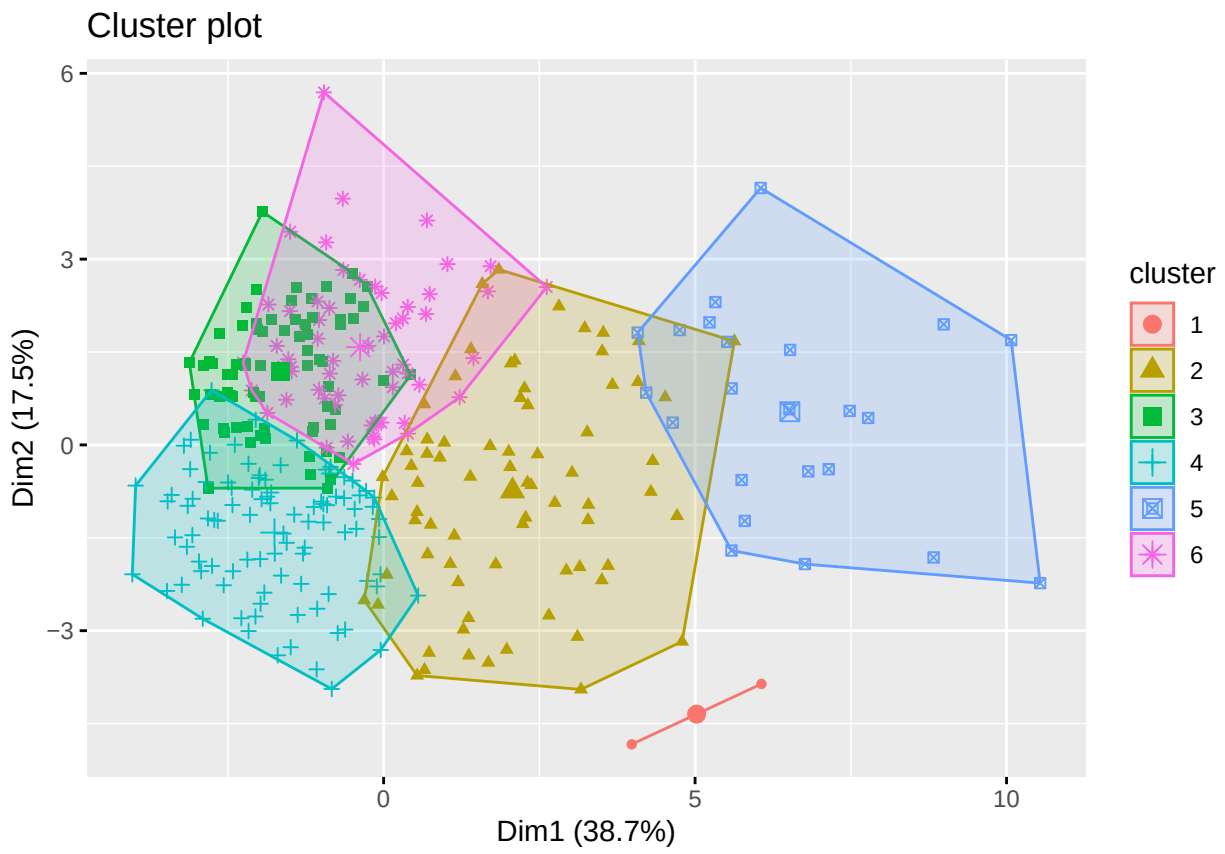



```
#Through elbow chart, we think K = 4, 6, 8 might be the optimal number of clusters

#(2)
#We try K = 4, 6, 8
#Use K-means to find the clusters with K = 4
k4 = kmeans(df, centers = 4, nstart = 25)

#Cluster plot
fviz_cluster(k4, data = df, geom = "point")
```


```
#Number of cities in each cluster
table(k4$cluster)
```

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

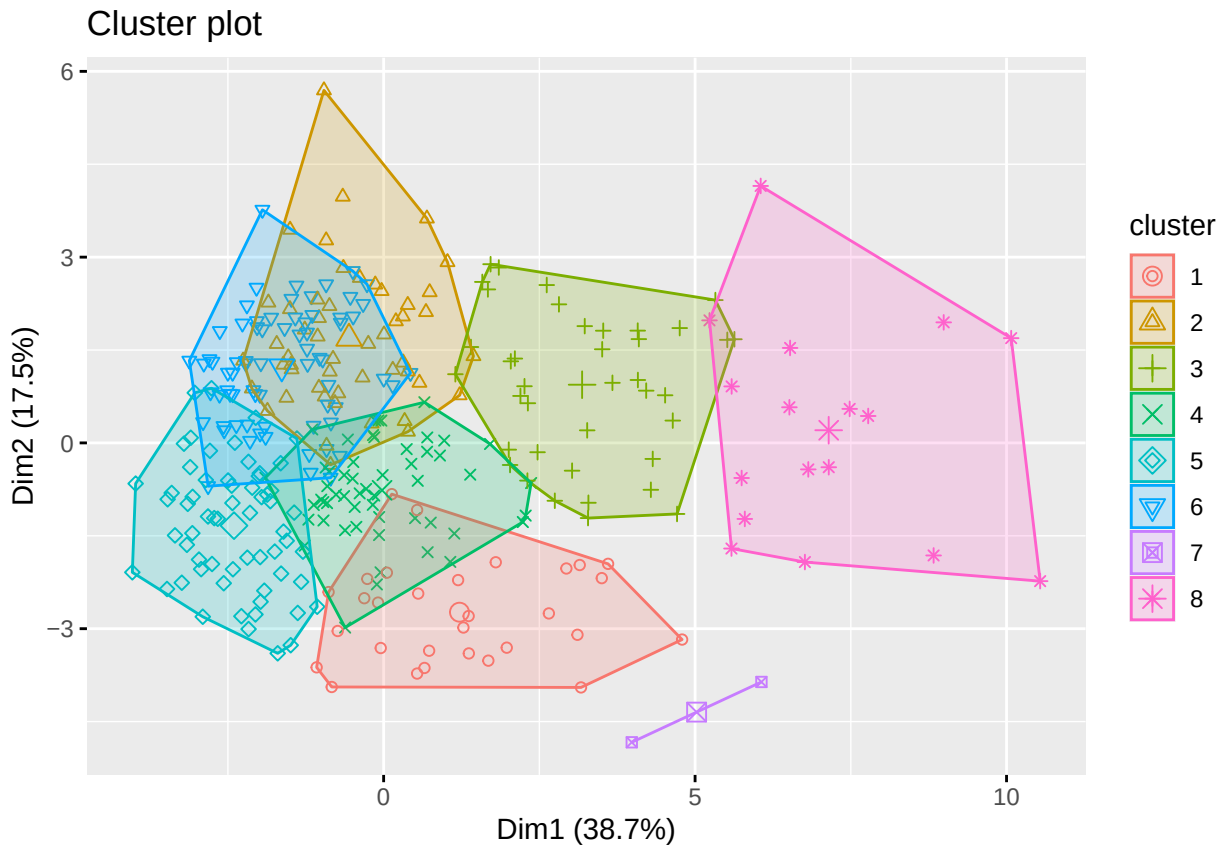
```
#Use K-means to find the clusters with K = 6
k6 = kmeans(df, centers = 6, nstart = 25)

#Cluster plot
fviz_cluster(k6, data = df, geom = "point")
```




```
#Number of cities in each cluster
table(k6$cluster)
```

```
##
##  1  2  3  4  5  6
##  2 72 73 94 23 61
```

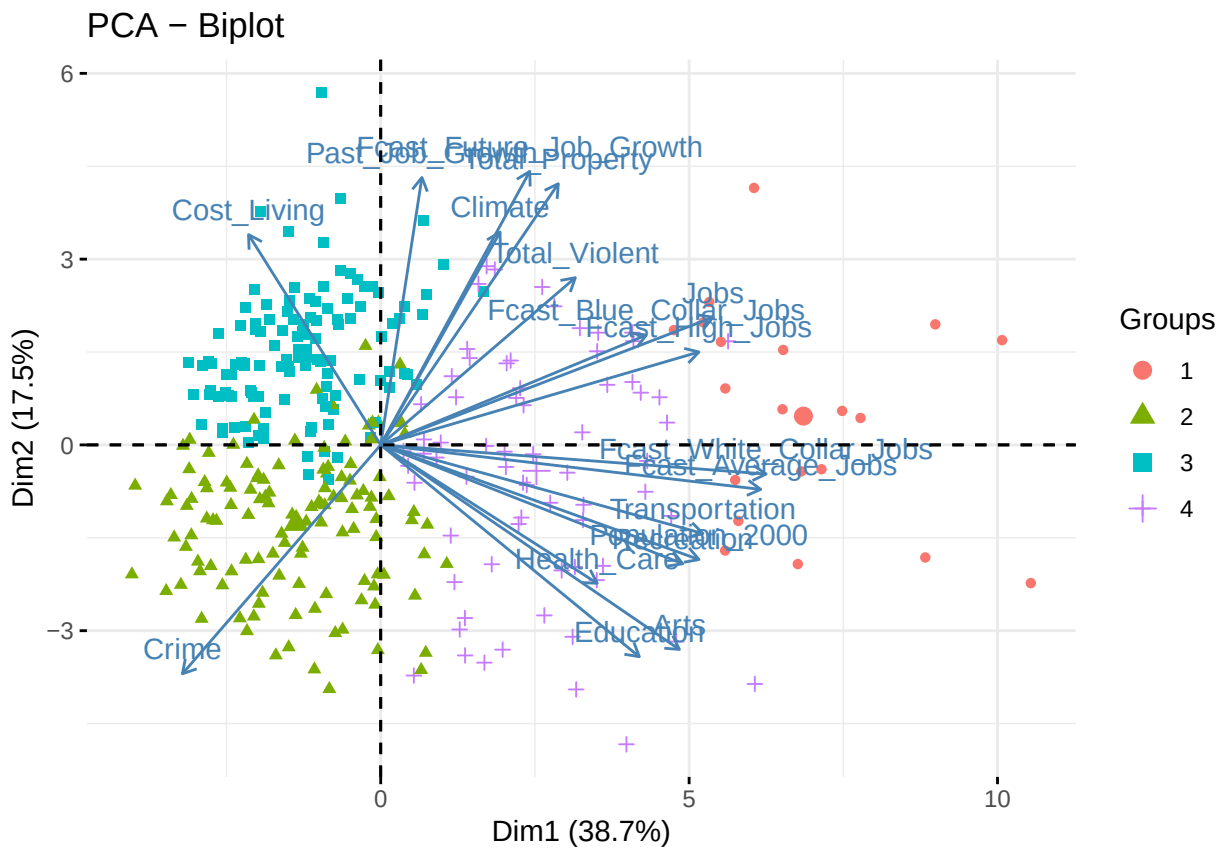
```
#Use K-means to find the clusters with K = 8
k8 = kmeans(df, centers = 8, nstart = 25)

#Cluster plot
fviz_cluster(k8, data = df, geom = "point")
```




```
#Number of cities in each cluster
table(k8$cluster)
```

```
##
##  1  2  3  4  5  6  7  8
## 30 51 39 54 61 71  2 17
```

```
#K = 6 and 8 have too much overlap,
#So we decide K = 4 be the optimal number of clusters as TWCV starts
#to decrease slowly and there are not too much overlap between each cluster

#(3)

#biplot
#add cluster to dataframe
df1 = df #copy df1, make df1 = df + column of cluster

cluster_number = as.data.frame(k4$cluster)
df1 = cbind(df1,cluster_number)
colnames(df1)[19] = "cluster"

m1 = prcomp(df)
fviz_pca_biplot(m1, label = "var", habillage = df1$cluster)
```

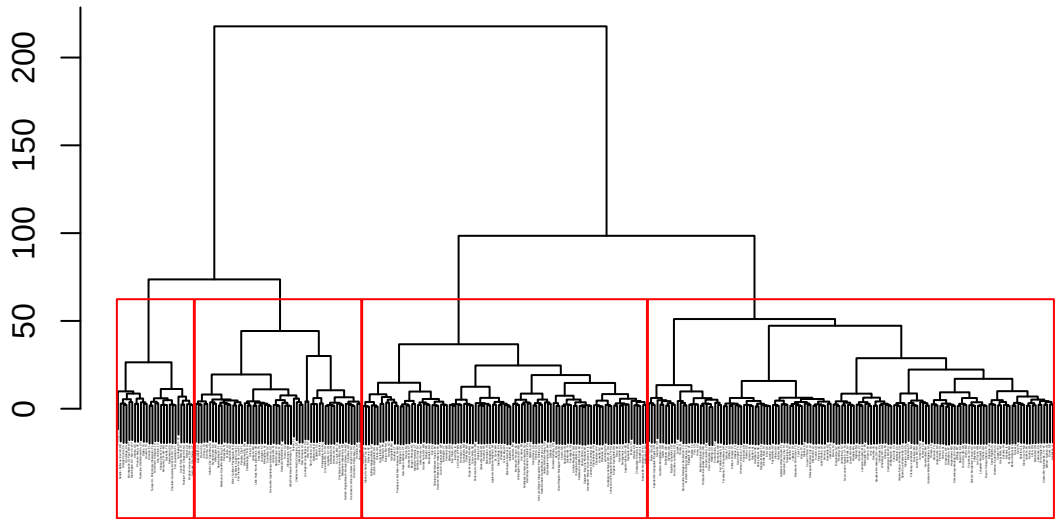



```
#median of each numerical column(on unscaled dataset)
aggregate(d, list(k4$cluster), median)
```

```
##   Group.1 Cost_Living Transportation   Jobs Education Climate  Crime  Arts
## 1       1      26.920         92.065 97.305    82.005   70.82 22.665 91.65
## 2       2      45.615         40.785 30.450    52.830   32.15 80.315 51.28
## 3       3      76.070         30.730 43.760    24.215   67.13 31.305 23.94
## 4       4      47.310         80.730 81.010    80.730   64.58 27.200 80.46
##   Health_Care Recreation Population_2000 Total_Violent Total_Property
## 1      65.290     90.365       2818808.5           753         5878.5
## 2      43.055     46.880        227733.5           273         3645.0
## 3      29.175     23.790        179977.5           653         5472.0
## 4      76.480     78.750       1059044.0           696         5436.0
##   Past_Job_Growth Fcast_Future_Job_Growth Fcast_Blue_Collar_Jobs
## 1            15.6                    8.3                20447.5
## 2             8.3                    4.8                  436.0
## 3            11.9                    6.0                  797.5
## 4            10.9                    5.9                 3388.0
##   Fcast_White_Collar_Jobs Fcast_High_Jobs Fcast_Average_Jobs
## 1                119533.5         23248.0            83826.0
## 2                  6518.5           796.5             4489.5
## 3                  6020.0          1367.5             3721.0
## 4                 33198.0          4976.0            23990.0
```

```
# Conclusion:
# Cluster 1 has highest rate in Transportation, Jobs, Education, Climate, Arts,
# Recreation, Population_2000, Total_Violent,Total_Property, Past_Job_Growth,
```

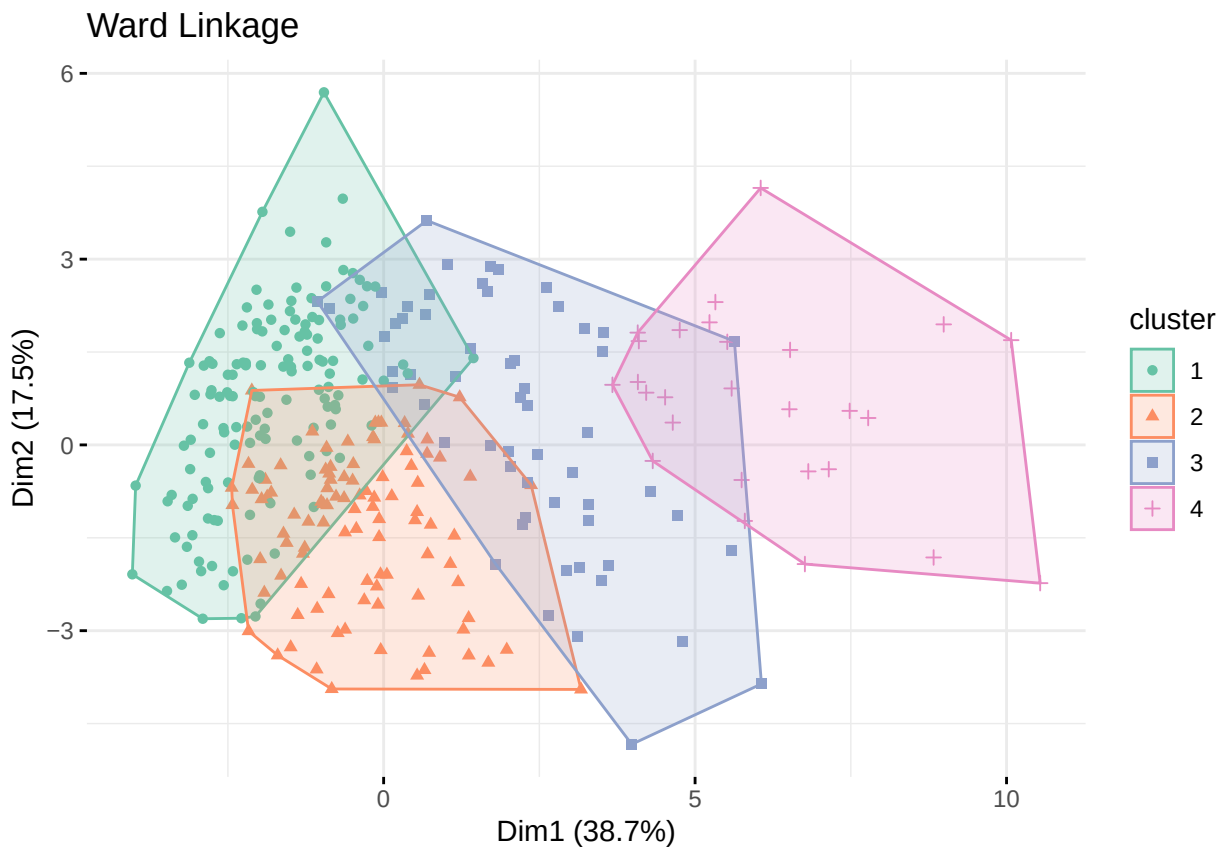
```r
# Fcast_Future_Job_Growth, Fcast_Blue_Collar_Jobs, Fcast_White_Collar_Jobs,
# Fcast_high_Jobs, Fcast_Average_Jobs, lowest rate in Cost_Living, Crime
# and Health_Care are in middle among four clusters

# Cluster 2 has highest rate in Crime, lowest rate in Jobs, Climate,
# Total_Violent, Total_Property, Past_Job_Growth, Fcast_Future_Job_Growth,
# Fcast_Blue_Collar_Jobs, Fcast_high_Jobs and Cost_Living, Transportation,
# Education, Arts, Health_Care, Recreation, Population_2000,
# Fcast_White_Collar_Jobs, Fcast_Average_Jobs are in middle among four clusters

# Cluster 3 has highest rate in Cost_Living, lowest rate in Transportation,
# Education, Arts, Health_Care, Recreation, Population_2000,
# Fcast_White_Collar_Jobs, Fcast_Average_Jobs and Jobs, Climate, Crime,
# Total_Violent, Total_Property, Past_Job_Growth, Fcast_Future_Job_Growth,
# Fcast_Blue_Collar_Jobs, Fcast_High_Jobs are in middle among four clusters

# Cluster 4 has highest rate in Health_Care and Cost_Living,
# Transportation, Jobs, Education, Climate, Crime, 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, Fcast_Average_Jobs are in middle among four clusters


#2 HIERARCHICAL CLUSTERING with K = 4
distance = dist(df)

#(1) Ward.D linkage
#dendrogram
h1 = hclust(distance, method = "ward.D")
plot(h1, cex = 0.1, xlab = "", ylab = "", sub = "",
     main = "Ward Linkage Dengrogram")
rect.hclust(h1, k = 4, border = "red")
```




```
cut1 = cutree(h1, k=4)
#number of cities in each cluster
table(cut1)
```

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

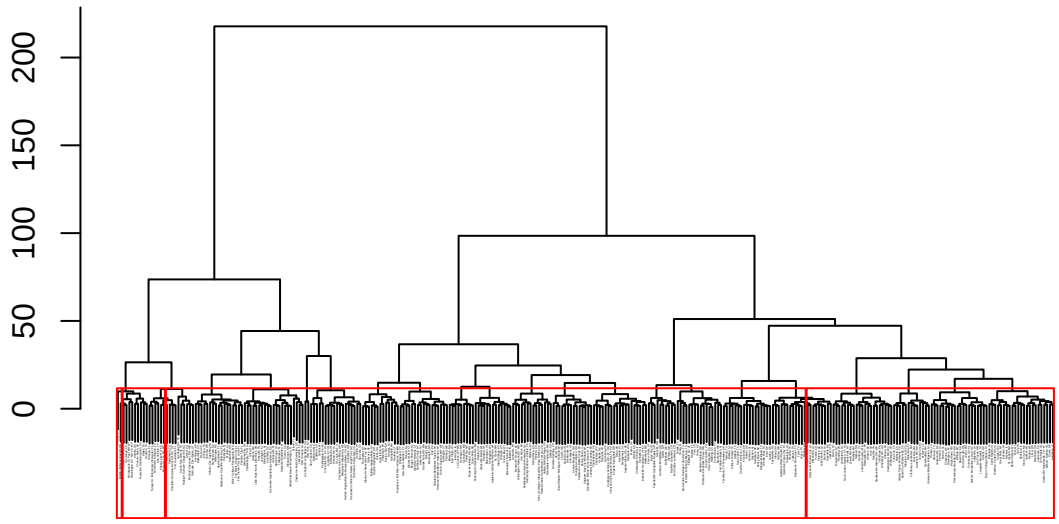
```
#cluster plot
fviz_cluster(list(data = df, cluster = cut1),main="Ward Linkage",
             palette = "Set2",show.clust.cent = F, geom = "point",
             repel = T, # Avoid label overplotting (slow)
             ggtheme = theme_minimal()
)
```




```
# CCPC for ward.D
c1 = cophenetic(h1)
cor(distance, c1)
```

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

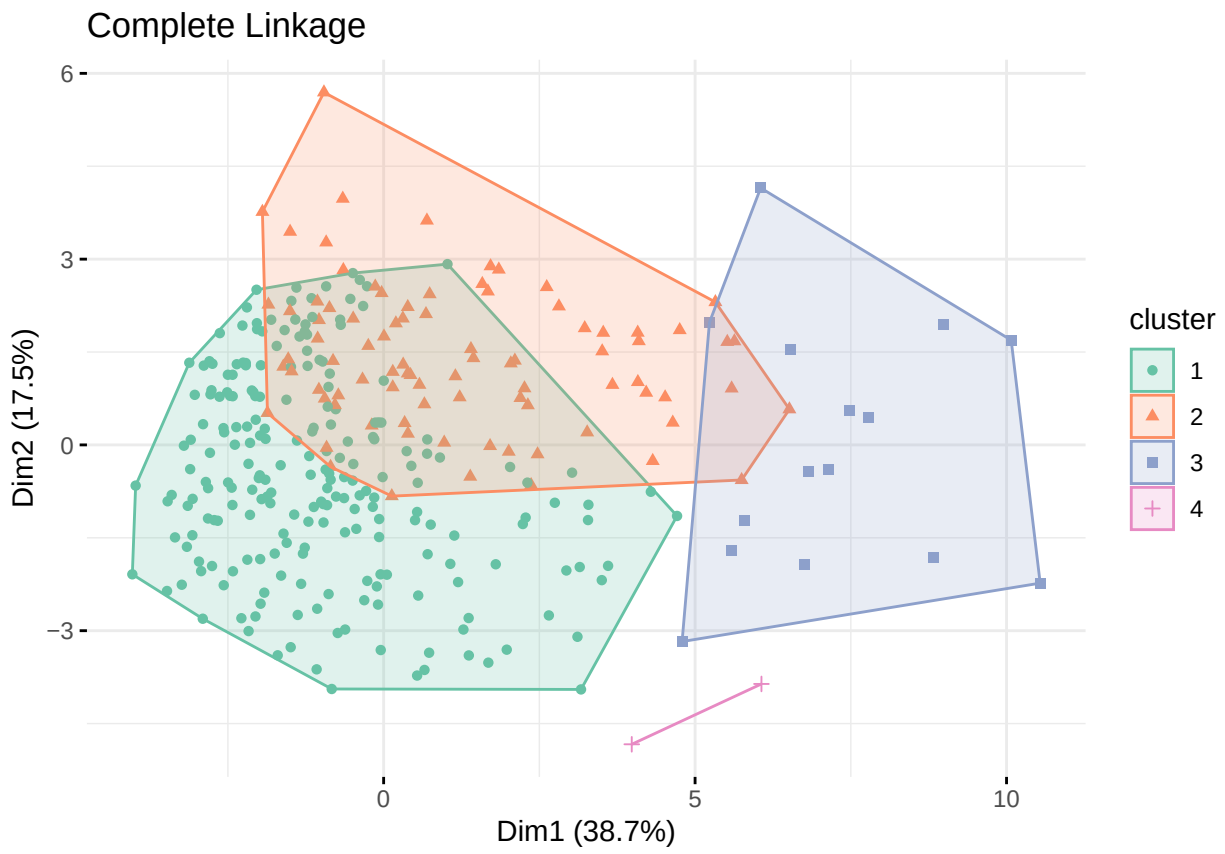
```
#(2) Complete linkage
#dendrogram
h2 = hclust(distance, method = "complete")
plot(h1, cex = 0.1, xlab = "", ylab = "", sub = "",
     main = "Complete Linkage Dengrogram")
rect.hclust(h2, k = 4, border = "red")
```




```
cut2 = cutree(h2, k=4)
#number of cities in each cluster
table(cut2)
```

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

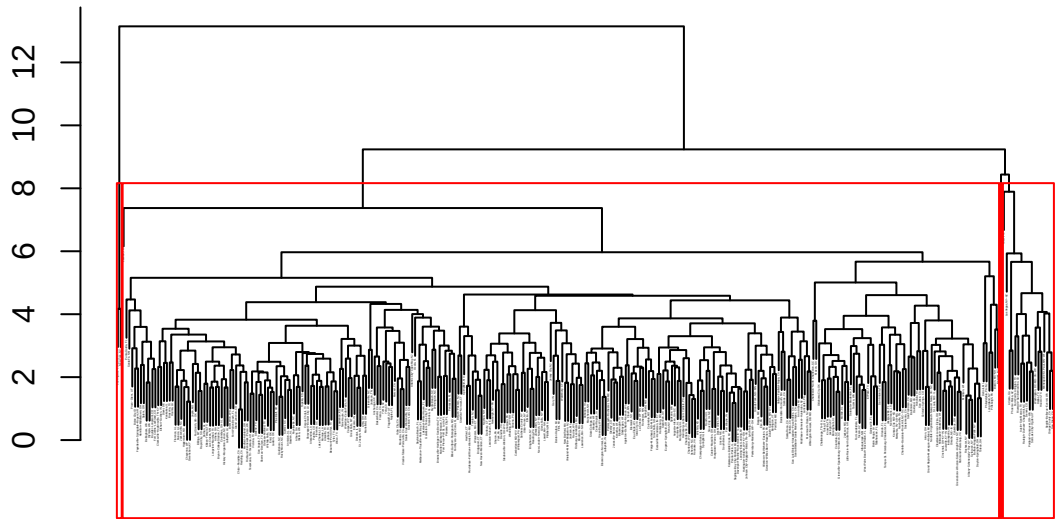
```
#cluster plot
fviz_cluster(list(data = df, cluster = cut2),main="Complete Linkage",
             palette = "Set2",show.clust.cent = F, geom = "point",
             repel = T, # Avoid label overplotting (slow)
             ggtheme = theme_minimal()
)
```




```
# CCPC for complete linkage
c2 = cophenetic(h2)
cor(distance, c2)
```

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

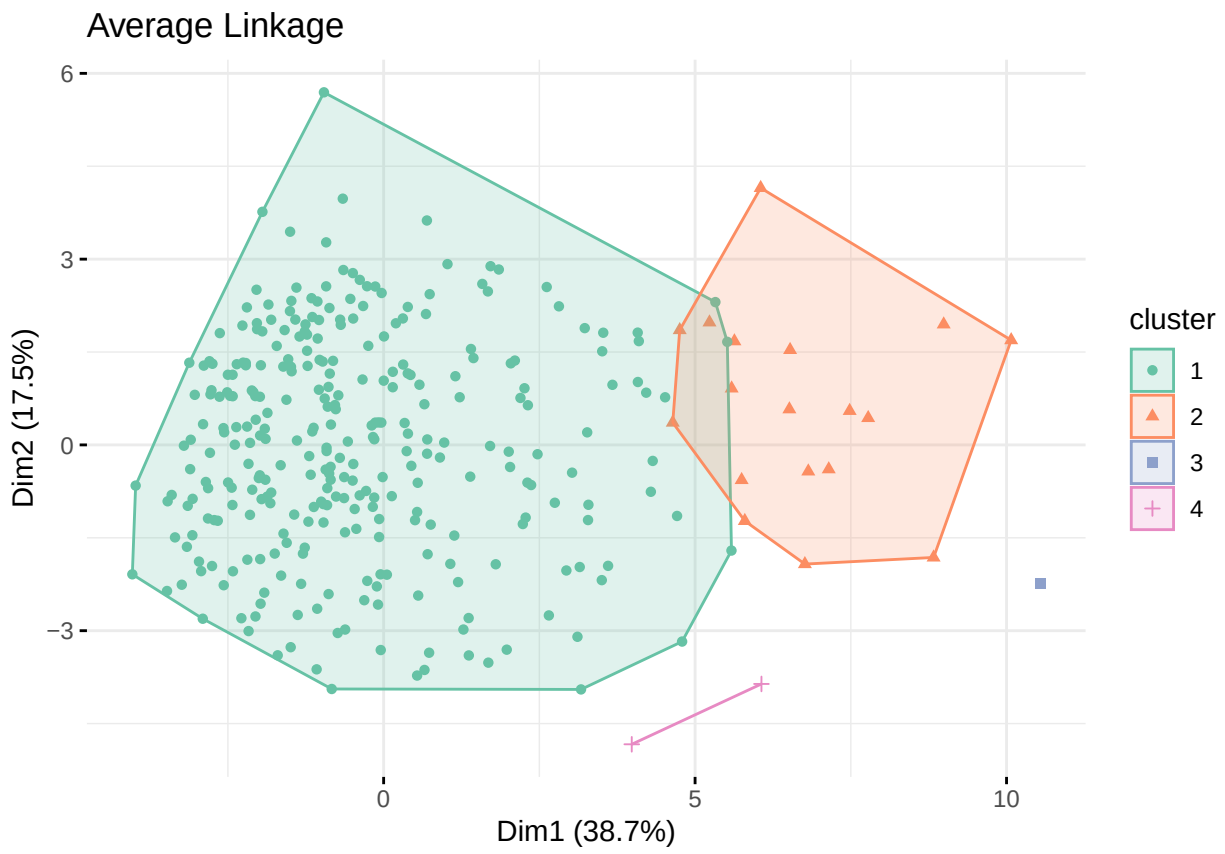
```
#(3) Average linkage
#dendrogram
h3 = hclust(distance, method = "average")
plot(h3, cex = 0.1, xlab = "", ylab = "", sub = "",
     main = "Average Linkage Dengrogram")
rect.hclust(h3, k = 4, border = "red")
```




```
cut3 = cutree(h3, k=4)
#number of cities in each cluster
table(cut3)
```

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

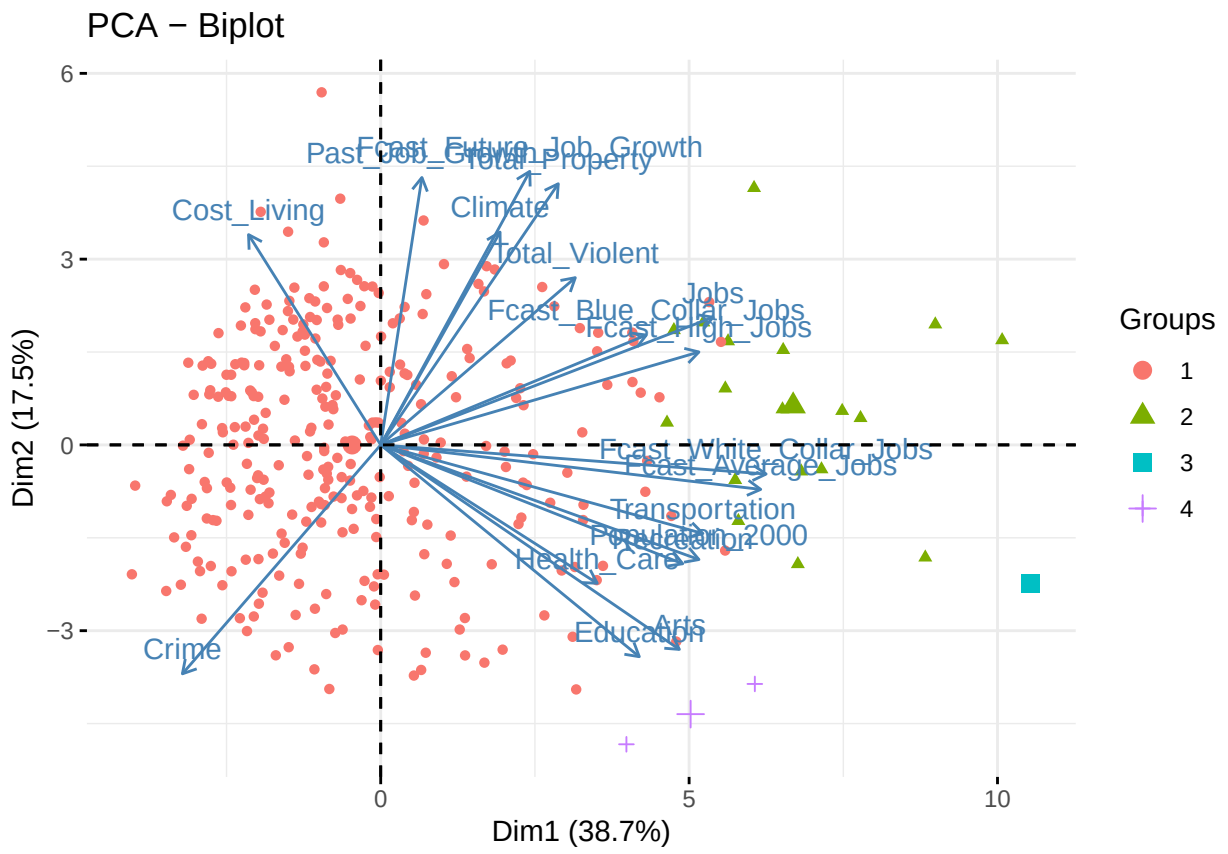
```
#cluster plot
fviz_cluster(list(data = df, cluster = cut3),main="Average Linkage",
             palette = "Set2",show.clust.cent = F, geom = "point",
             repel = T, # Avoid label overplotting (slow)
             ggtheme = theme_minimal()
)
```




```
# CCPC for average linkage.
c3 = cophenetic(h3)
cor(distance, c3)
```

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

```
#(4)
#I prefer average linkage to do hierarchical clustering since the
#clustering using average linkage has the minimum overlap between each clusters
#and it also has the largest CCPC

#biplot
#add cluster to dataframe
df2 = df #copy df2, make df2 = df + column of cluster

cluster_number = as.data.frame(cut3)
df2 = cbind(df2,cluster_number)
colnames(df2)[19] = "cluster"

m2 = prcomp(df)
fviz_pca_biplot(m2, label = "var", habillage = df2$cluster)
```




```
#median of each numerical column(on unscaled dataset)
aggregate(d, list(cut3), median)
```

```
##   Group.1 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
```

```
# Conclusion:
# Cluster 1 has the highest rate in Cost_Living, Crime,
# lowest rate in Transportation, Education, Arts, Health_care,
```

```
# Recreation, Population_2000, Total_Violent, Total_Property,
# Fcast_White_Collar_Jobs, Fcast_Average_Jobs and Jobs, Climate,
# Past_Job_Growth, Fcast_Future_Job_Growth, Fcast_Blue_Collar_Jobs,
# Fcast_High_Jobs are in middle among four clusters

# Cluster 2 has the highest rate in Jobs, Total_Property, Past_Job_Growth,
# Fcast_Future_Job_Growth, Fcast_High_Jobs, lowest rate in and Cost_Living,
# Transportation, Education, Climate, Crime, Arts, Health_Care, Recreation,
# Population_2000, Total_Violent, Fcast_Blue_Collar_Jobs,
# Fcast_White_Collar_Jobs, Fcast_Average_Jobs are in middle among four clusters

# Cluster 3 has the highest rate in Transportation, Education, Health_Care,
# Recreation, Fcast_Blue_Collar_Jobs, Fcast_White_Collar_Jobs,
# Fcast_Average_Jobs, lowest rate in Climate and Cost_Living, Jobs, Crime,
# Arts, Population_2000, Total_Violent, Total_Property, Past_Job_Growth,
# Fcast_Future_Job_Growth, Fcast_High_Jobs are in middle among four clusters

# Cluster 4 has the highest rate in Climate, Arts, Population_2000,
# Total_Violent, lowest rate in Cost_Living, Jobs, Crime, Past_Job_Growth,
# Fcast_Future_Job_Growth, Fcast_Blue_Collar_Jobs, Fcast_High_Jobs
# and Transportation, Education, Health_Care, Recreation,Total_Property,
# Fcast_White_Collar_Jobs, Fcast_Average_Jobs are in middle among four clusters
```