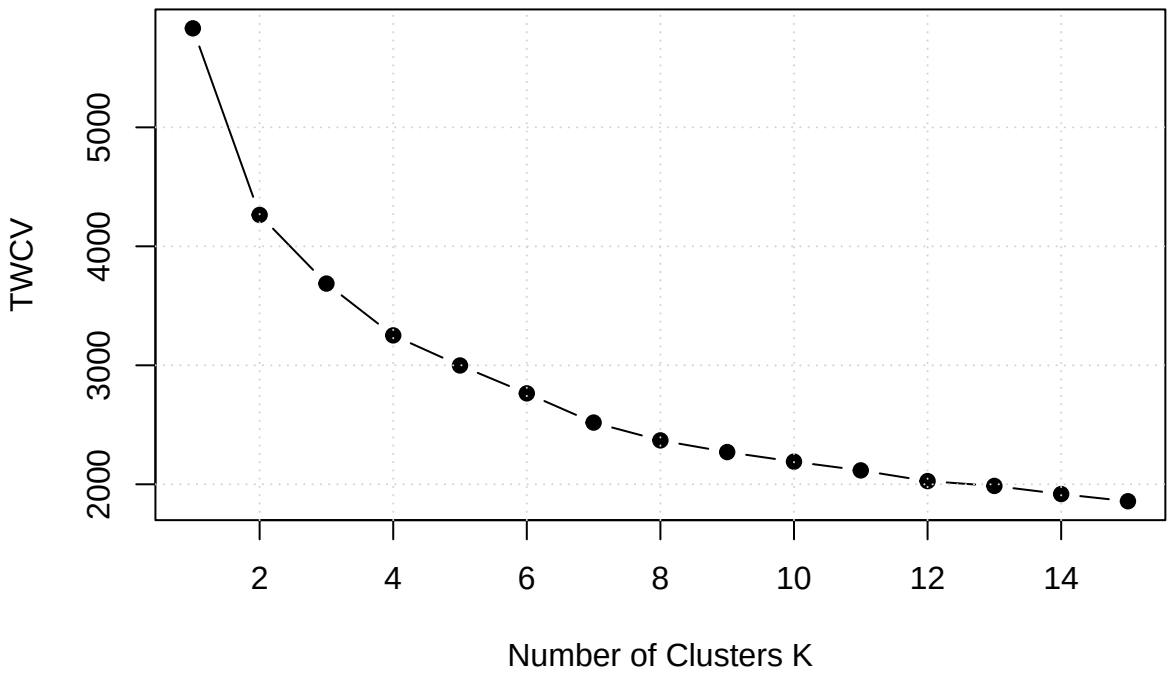
Siyu 535 hw4

Siyu Wu | USC ID: 2350-0175-97

```
library(readxl)
library(tibble)
df0 = read_excel("cities1(1).xlsx",sheet="Data")
df0$Crime_Trend = NULL
df0$Unemployment_Threat = NULL
df = column_to_rownames(df0,"Metropolitan_Area")
df = scale(df)
```

```
df_prep = df0
df_prep = column_to_rownames(df_prep,"Metropolitan_Area")
df_mm = data.frame(apply(df_prep,2,function(x) (x-min(x))/(max(x)-min(x))))
```

```
set.seed(123)
twcv = function(k) kmeans(df,k,nstart=10)$tot.withinss
k = 1:15
twcv_values = sapply(k,twcv)
plot(k,twcv_values,type="b",pch=19,
     xlab="Number of Clusters K", ylab="TWCV")
grid()
```



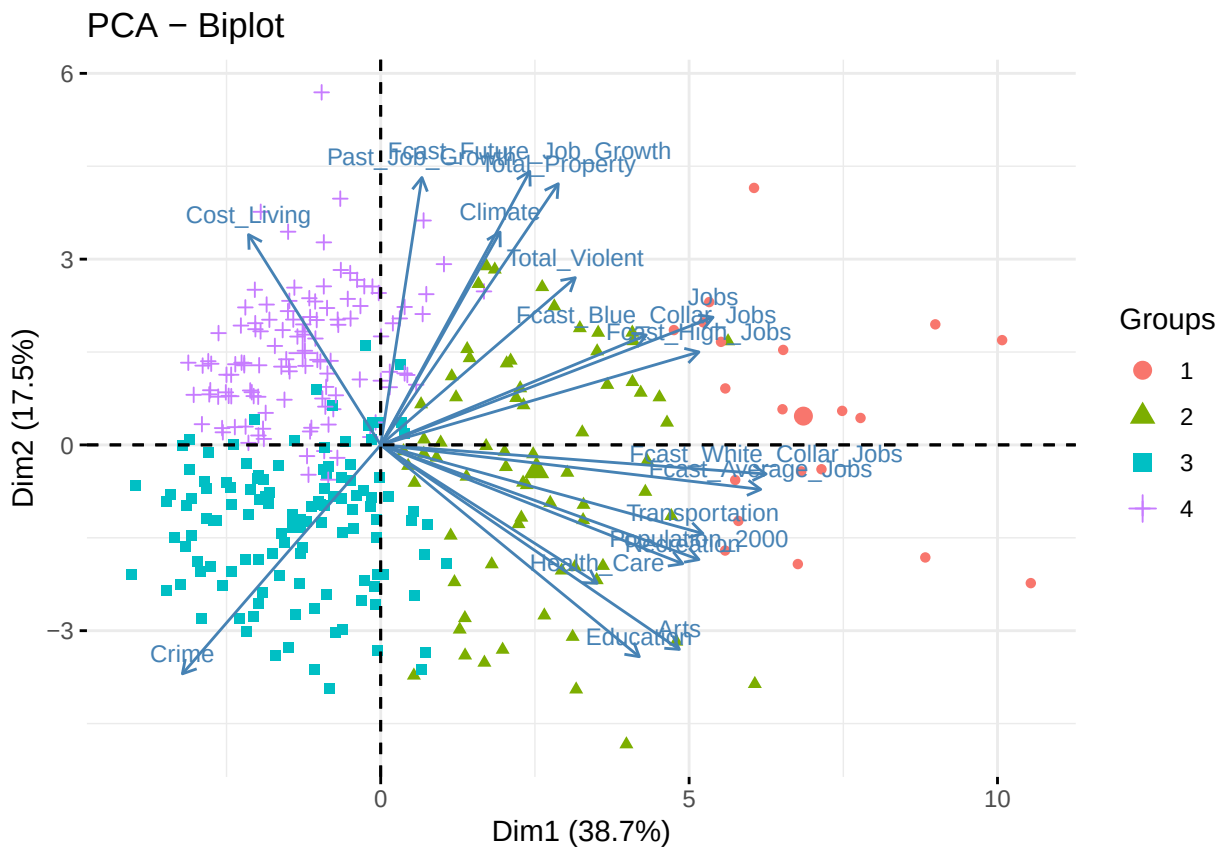

```
set.seed(123)
library(cluster)
library(factoextra)
```

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

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

```
final = kmeans(df,centers=4,nstart=25)
fviz_cluster(final,data=df,geom = "point")
```


```
cluster_number = as.factor(final$cluster)
m1 = prcomp(df,scale=T)
fviz_pca_biplot(m1,habillage=cluster_number,geom="point",labelsize=3)
```




```
dfc = df_mm
dfc$cluster = final$cluster[rownames(dfc)]
dfc$Metropolitan_Area = NULL
library(data.table)
df_avg = setDT(dfc)[, lapply(.SD, mean), keyby = cluster]

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df_avg = data.frame(apply(df_avg %>% select(2:19),2,function(x) x / sum(x)))
df_avg['cluster'] = c(1,2,3,4)

library(tidyr)
df_td = gather(df_avg,key='type',value='value',1:18)
```

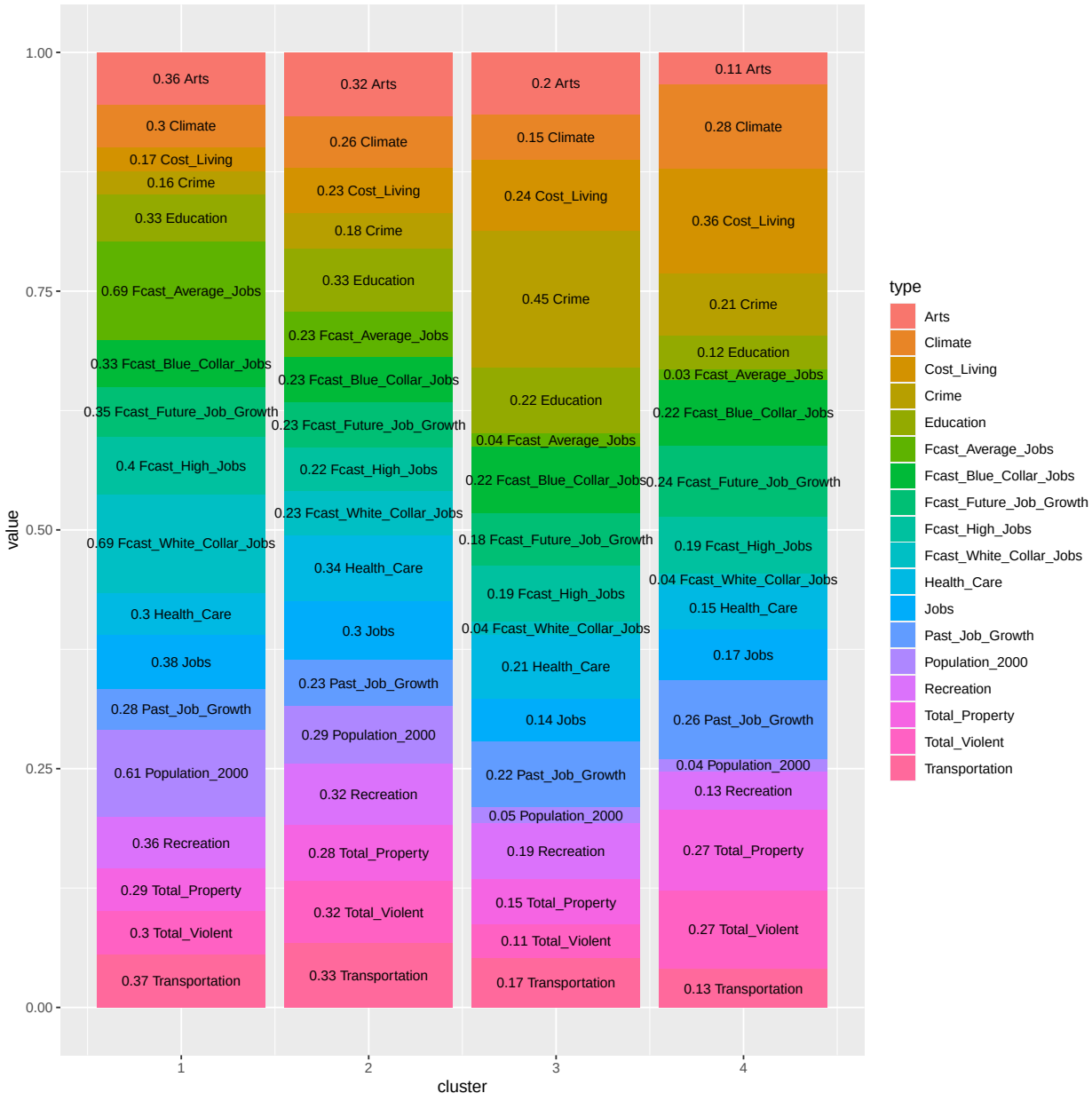
```
df_min = df_td %>% group_by(type) %>% slice(which.min(value))
for (i in 1:4){
  cat('-----------')
  cat('kCluster',i,'Minimum Attribute\n')
  print(filter(df_min, cluster==i)$type)
}
```

```
## -----------kCluster 1 Minimum Attribute
## [1] "Cost_Living" "Crime"
## -----------kCluster 2 Minimum Attribute
## character(0)
## -----------kCluster 3 Minimum Attribute
## [1] "Climate"                 "Fcast_Blue_Collar_Jobs"
## [3] "Fcast_Future_Job_Growth" "Fcast_High_Jobs"
## [5] "Jobs"                    "Past_Job_Growth"
## [7] "Total_Property"          "Total_Violent"
## -----------kCluster 4 Minimum Attribute
## [1] "Arts"                    "Education"
## [3] "Fcast_Average_Jobs"      "Fcast_White_Collar_Jobs"
## [5] "Health_Care"             "Population_2000"
## [7] "Recreation"              "Transportation"
```

```
df_max = df_td %>% group_by(type) %>% slice(which.max(value))
for (i in 1:4){
  cat('-----------')
  cat('kCluster',i,'Maximum Attribute\n')
  print(filter(df_max, cluster==i)$type)
}
```

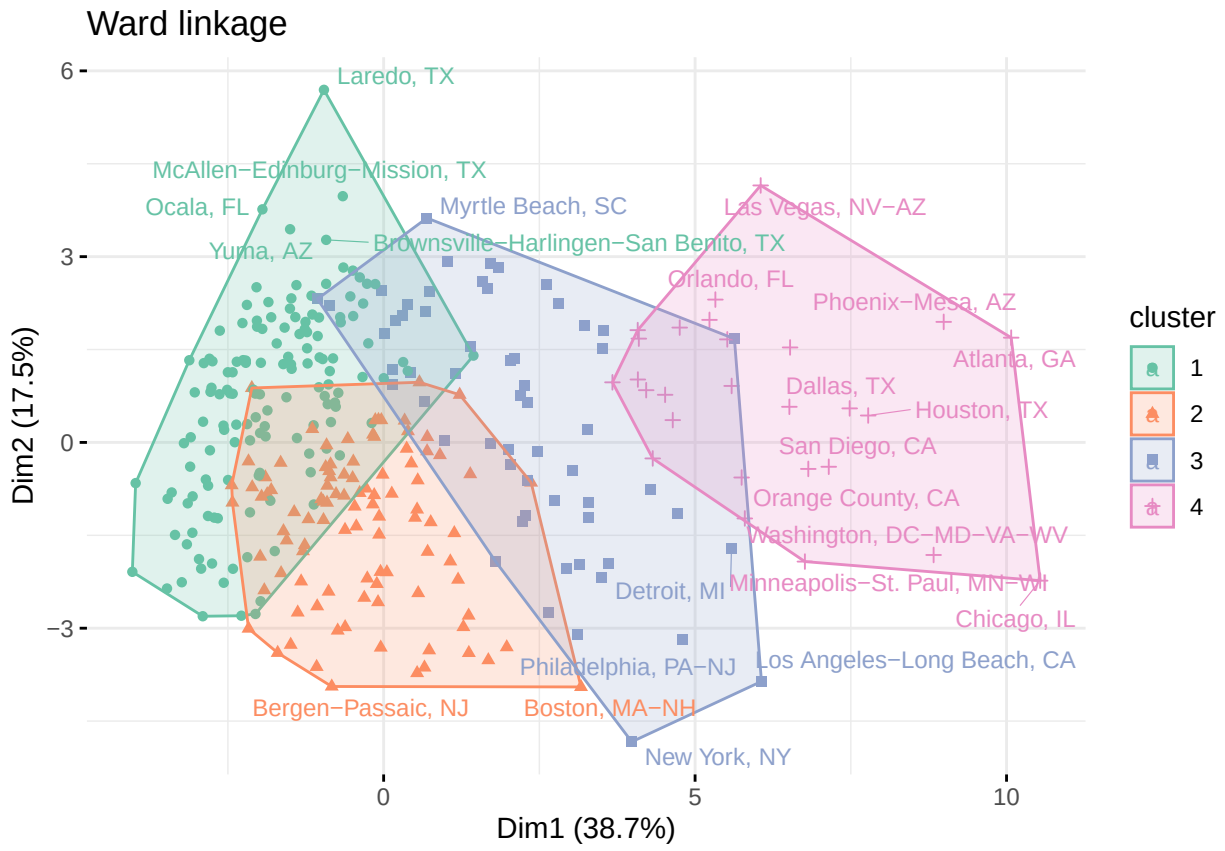
```
## -----------kCluster 1 Maximum Attribute
##  [1] "Arts"                    "Climate"
##  [3] "Education"               "Fcast_Average_Jobs"
##  [5] "Fcast_Blue_Collar_Jobs"  "Fcast_Future_Job_Growth"
##  [7] "Fcast_High_Jobs"         "Fcast_White_Collar_Jobs"
##  [9] "Jobs"                    "Past_Job_Growth"
## [11] "Population_2000"         "Recreation"
## [13] "Total_Property"          "Transportation"
## -----------kCluster 2 Maximum Attribute
## [1] "Health_Care"   "Total_Violent"
## -----------kCluster 3 Maximum Attribute
## [1] "Crime"
## -----------kCluster 4 Maximum Attribute
## [1] "Cost_Living"
```

```
ggplot(df_td,aes(x = cluster, y = value, fill = type)) +
  geom_col(position = "fill") +
  geom_text(aes(label = paste(round(value,2),type)),
                position = position_fill(vjust=0.5),
                size=3)
```




```
distance=dist(df)
h1 = hclust(distance,method="ward.D")
cut1 = cutree(h1,k=4)
fviz_cluster(list(data=df,cluster=cut1),main="Ward linkage",
             palette = "Set2", show.clust.cent=F,labelsize=10,
             repel=T,
             ggtheme=theme_minimal())
```

```
## Warning: ggrepel: 302 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

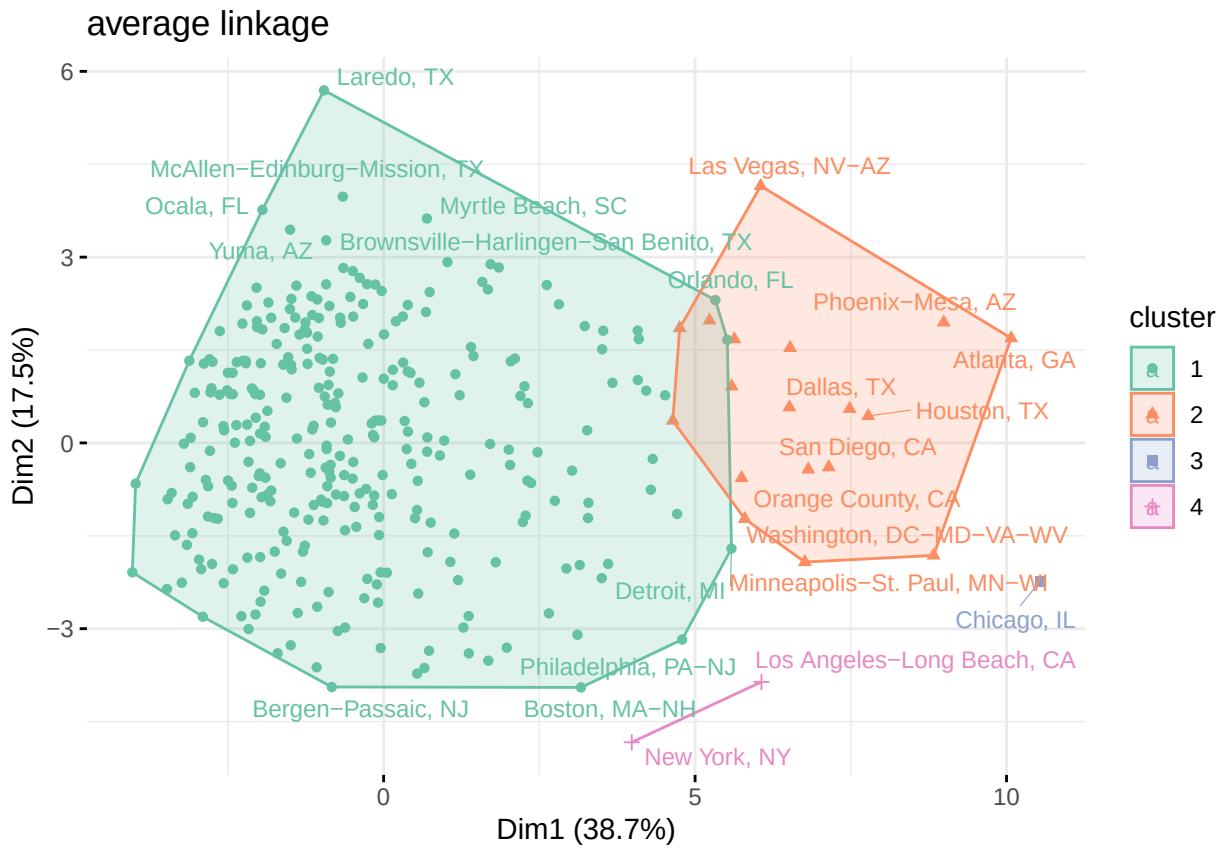



```
c1 = cophenetic(h1)
cor(distance,c1)
```

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

```
h2 = hclust(distance, method ='average')
cut2 = cutree(h2,k=4)
fviz_cluster(list(data = df, cluster = cut2),main="average linkage",
             palette = "Set2",show.clust.cent = F, labelsize = 10,
             repel = T, # Avoid label overlap (slow)
             ggtheme = theme_minimal()
             )
```

```
## Warning: ggrepel: 302 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```

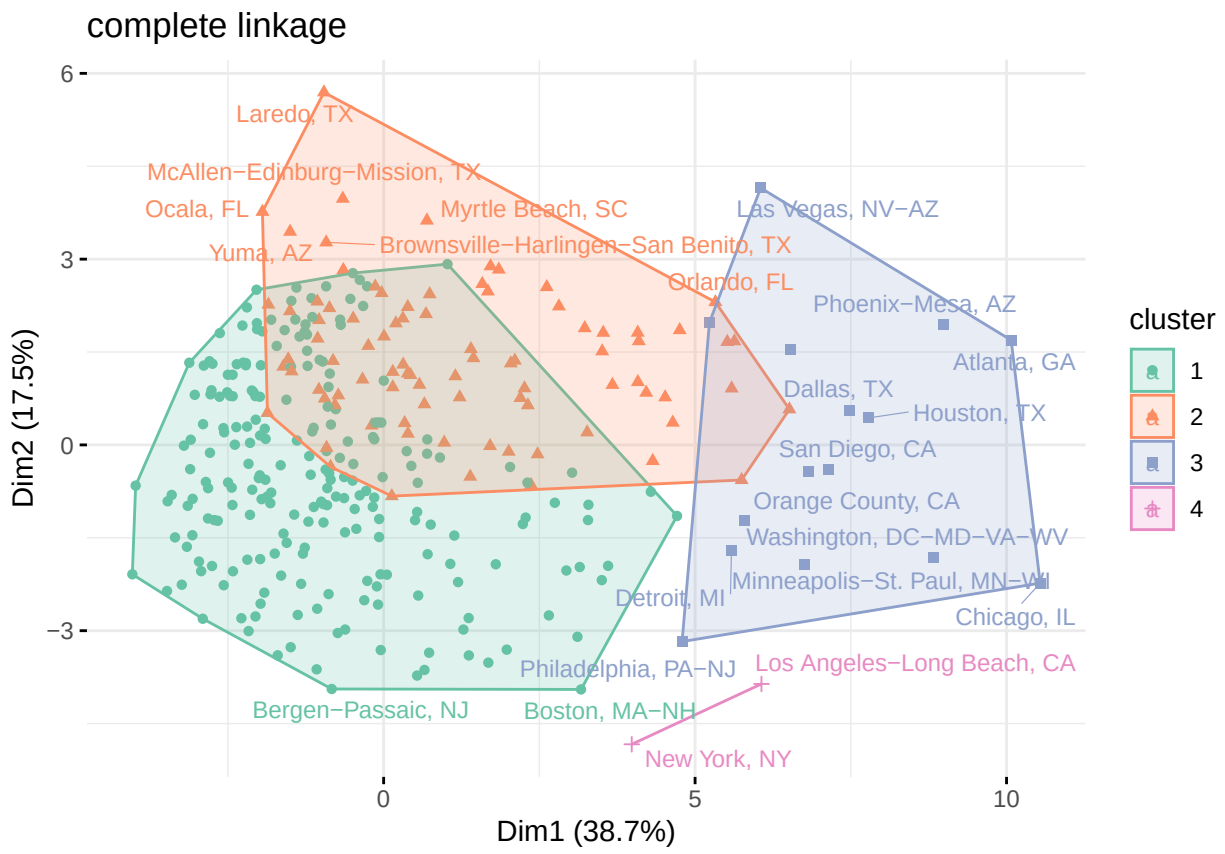



```
c2 = cophenetic(h2)
cor(distance,c2)
```

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

```
h3 = hclust(distance, method ='complete')
cut3 = cutree(h3,k=4)
fviz_cluster(list(data = df, cluster = cut3),main="complete linkage",
             palette = "Set2",show.clust.cent = F, labelsize = 10,
             repel = T,
             ggtheme = theme_minimal()
             )
```

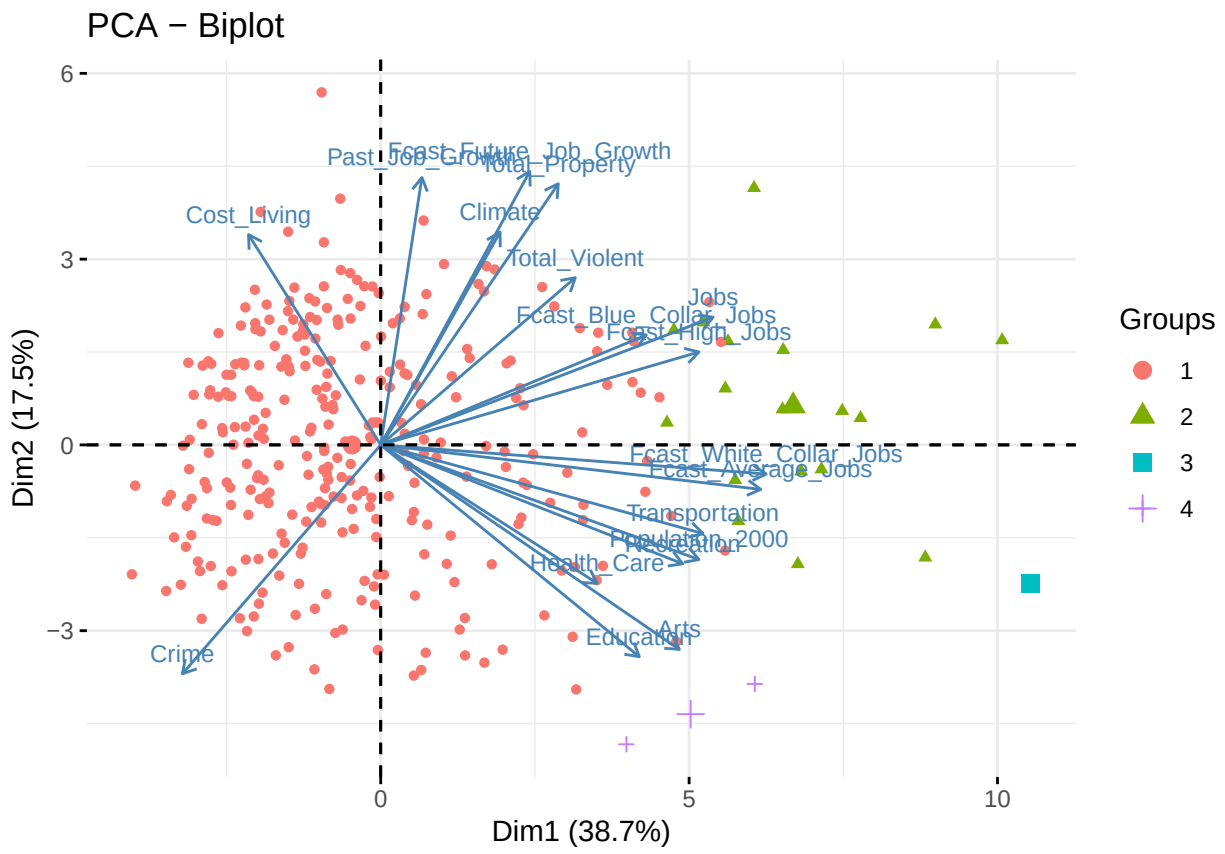
```
## Warning: ggrepel: 302 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




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

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

```
fviz_pca_biplot(m1,habillage=cut2,geom="point",labelsize=3)
```




```
dfh = df_mm
dfh$cluster = cut2
dfh$Metropolitan_Area = NULL
dfh_avg = setDT(dfh)[, lapply(.SD, mean), keyby = cluster]

dfh_avg = data.frame(apply(dfh_avg %>% select(2:19),2,function(x) x / sum(x)))
dfh_avg['cluster'] = c(1,2,3,4)

dfh_td = gather(dfh_avg,key='type',value='value',1:18)
```

```
dfh_min = dfh_td %>% group_by(type) %>% slice(which.min(value))
for (i in 1:4){
  cat('-----------')
  cat('hCluster',i,'Minimum Attribute\n')
  print(filter(dfh_min, cluster==i)$type)
}
```

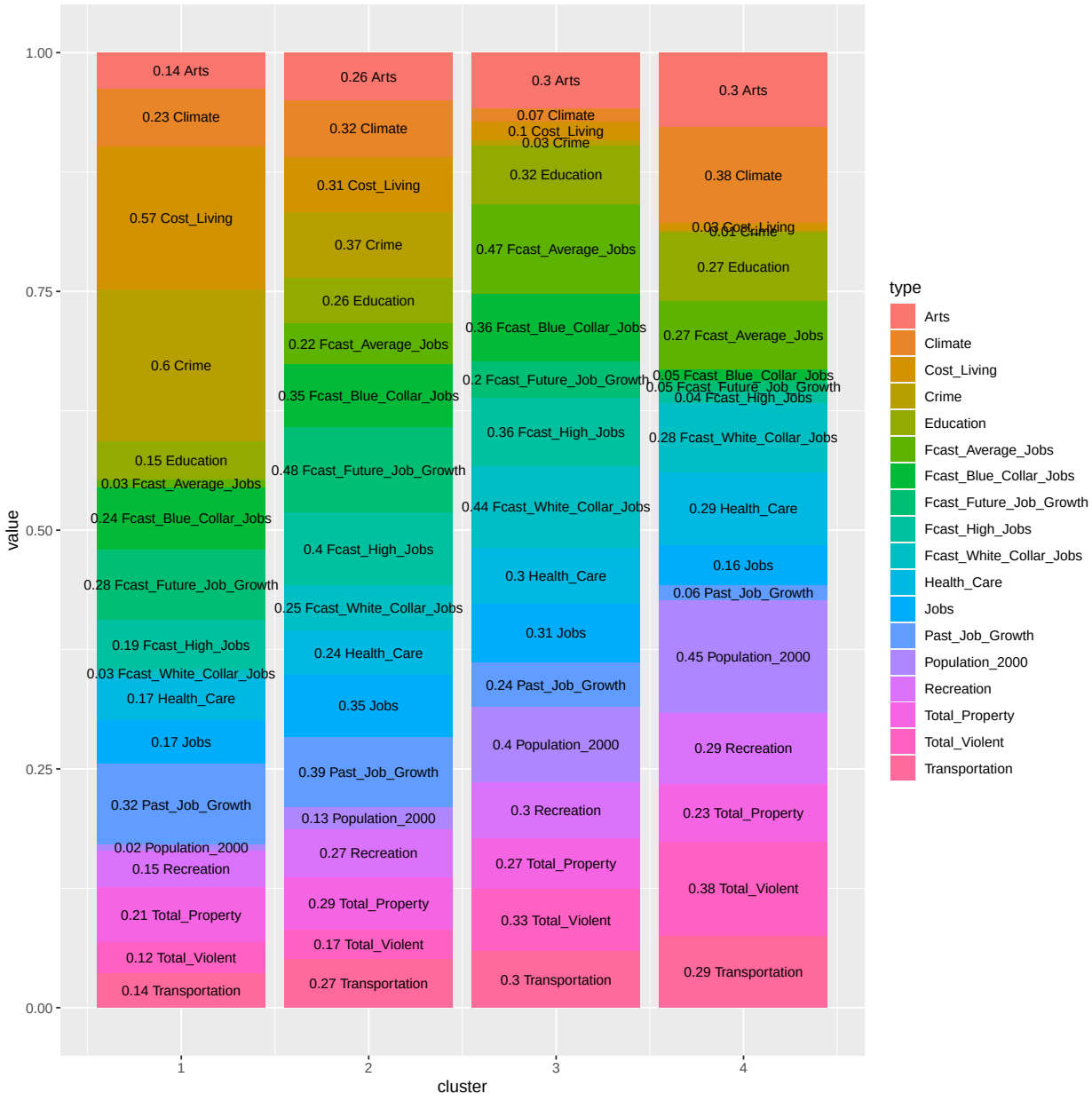
```
## -----------hCluster 1 Minimum Attribute
##  [1] "Arts"                    "Education"
##  [3] "Fcast_Average_Jobs"      "Fcast_White_Collar_Jobs"
##  [5] "Health_Care"             "Population_2000"
##  [7] "Recreation"              "Total_Property"
##  [9] "Total_Violent"           "Transportation"
## -----------hCluster 2 Minimum Attribute
## character(0)
```

```
## -----------hCluster 3 Minimum Attribute
## [1] "Climate"
## -----------hCluster 4 Minimum Attribute
## [1] "Cost_Living"             "Crime"
## [3] "Fcast_Blue_Collar_Jobs"  "Fcast_Future_Job_Growth"
## [5] "Fcast_High_Jobs"         "Jobs"
## [7] "Past_Job_Growth"
```

```
dfh_max = dfh_td %>% group_by(type) %>% slice(which.max(value))
for (i in 1:4){
  cat('-----------')
  cat('hCluster',i,'Maximum Attribute\n')
  print(filter(dfh_max, cluster==i)$type)
}
```

```
## -----------hCluster 1 Maximum Attribute
## [1] "Cost_Living" "Crime"
## -----------hCluster 2 Maximum Attribute
## [1] "Fcast_Future_Job_Growth" "Fcast_High_Jobs"
## [3] "Jobs"                    "Past_Job_Growth"
## [5] "Total_Property"
## -----------hCluster 3 Maximum Attribute
## [1] "Education"               "Fcast_Average_Jobs"
## [3] "Fcast_Blue_Collar_Jobs"  "Fcast_White_Collar_Jobs"
## [5] "Health_Care"             "Recreation"
## [7] "Transportation"
## -----------hCluster 4 Maximum Attribute
## [1] "Arts"            "Climate"         "Population_2000" "Total_Violent"
```

```
ggplot(dfh_td,aes(x = cluster, y = value, fill = type)) +
  geom_col(position = "fill") +
  geom_text(aes(label = paste(round(value,2),type)),
                position = position_fill(vjust=0.5),
                size=3)
```


value
1.00
0.75
0.50
0.25
0.00
cluster
1
2
3
4
0.14 Arts
0.23 Climate
0.57 Cost_Living
0.6 Crime
0.15 Education
0.03 Fcast_Average_Jobs
0.24 Fcast_Blue_Collar_Jobs
0.28 Fcast_Future_Job_Growth
0.19 Fcast_High_Jobs
0.03 Fcast_White_Collar_Jobs
0.17 Health_Care
0.17 Jobs
0.32 Past_Job_Growth
0.02 Population_2000
0.15 Recreation
0.21 Total_Property
0.12 Total_Violent
0.14 Transportation
0.26 Arts
0.32 Climate
0.31 Cost_Living
0.37 Crime
0.26 Education
0.22 Fcast_Average_Jobs
0.35 Fcast_Blue_Collar_Jobs
0.48 Fcast_Future_Job_Growth
0.4 Fcast_High_Jobs
0.25 Fcast_White_Collar_Jobs
0.24 Health_Care
0.35 Jobs
0.39 Past_Job_Growth
0.13 Population_2000
0.27 Recreation
0.29 Total_Property
0.17 Total_Violent
0.27 Transportation
0.3 Arts
0.07 Climate
0.1 Cost_Living
0.03 Crime
0.32 Education
0.47 Fcast_Average_Jobs
0.36 Fcast_Blue_Collar_Jobs
0.2 Fcast_Future_Job_Growth
0.36 Fcast_High_Jobs
0.44 Fcast_White_Collar_Jobs
0.3 Health_Care
0.31 Jobs
0.24 Past_Job_Growth
0.4 Population_2000
0.3 Recreation
0.27 Total_Property
0.33 Total_Violent
0.3 Transportation
0.3 Arts
0.38 Climate
0.03 Cost_Living
0.01 Crime
0.27 Education
0.27 Fcast_Average_Jobs
0.05 Fcast_Blue_Collar_Jobs
0.05 Fcast_Future_Job_Growth
0.04 Fcast_High_Jobs
0.28 Fcast_White_Collar_Jobs
0.29 Health_Care
0.16 Jobs
0.06 Past_Job_Growth
0.45 Population_2000
0.29 Recreation
0.23 Total_Property
0.38 Total_Violent
0.29 Transportation
type
Arts
Climate
Cost_Living
Crime
Education
Fcast_Average_Jobs
Fcast_Blue_Collar_Jobs
Fcast_Future_Job_Growth
Fcast_High_Jobs
Fcast_White_Collar_Jobs
Health_Care
Jobs
Past_Job_Growth
Population_2000
Recreation
Total_Property
Total_Violent
Transportation