clustering and PCA

Jiatao Wang

11/3/2021

Clustering

```
library(readr)
library(readxl)
setwd("C:/Users/CKA/Documents/CKA/the-green-chair-project")
data <- read_csv("cleaned_STATCOM_data.csv")
tgcp_demog <- read_excel("C:/Users/CKA/Downloads/STATCOM_data.xlsx", col_types = "text")



anyNA(data$Homeincome)
```

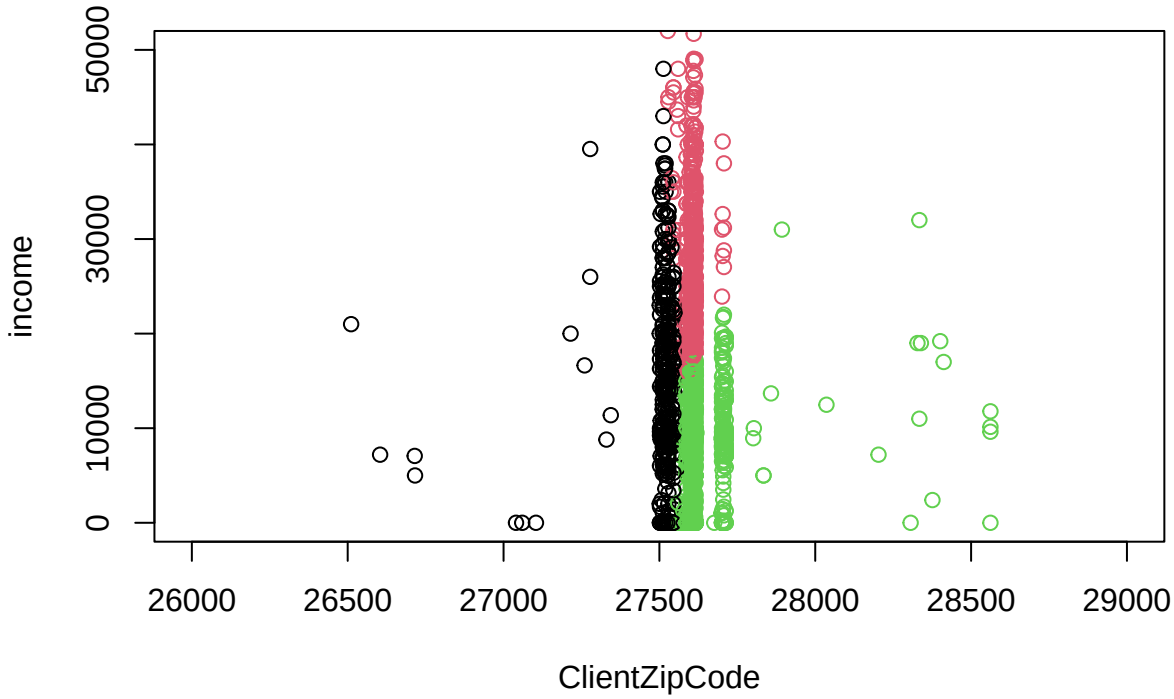
```
## [1] FALSE
```

```
#data$Homeincome

income_amount <- tgcp_demog$AnnualIncomeAmount

all <- cbind(data,income_amount)
anyNA(all$income_amount)
```

```
## [1] TRUE
```

```
library(tidyverse)

cleaned <- all%>% filter(!is.na(ClientZipCode),!is.na(income_amount))
final <- cleaned%>% select(ClientZipCode,income_amount)
#normalizerd data varaibles zipcode and income_amount.
#str(final)
income<-as.numeric(final$income_amount)
Z <-cbind(final,income)
last<-Z[,-2]

means <- apply(last,2,mean)
sds <- apply(last,2,sd)
get <- scale(last,center=means,scale=sds)

set.seed(123)
```

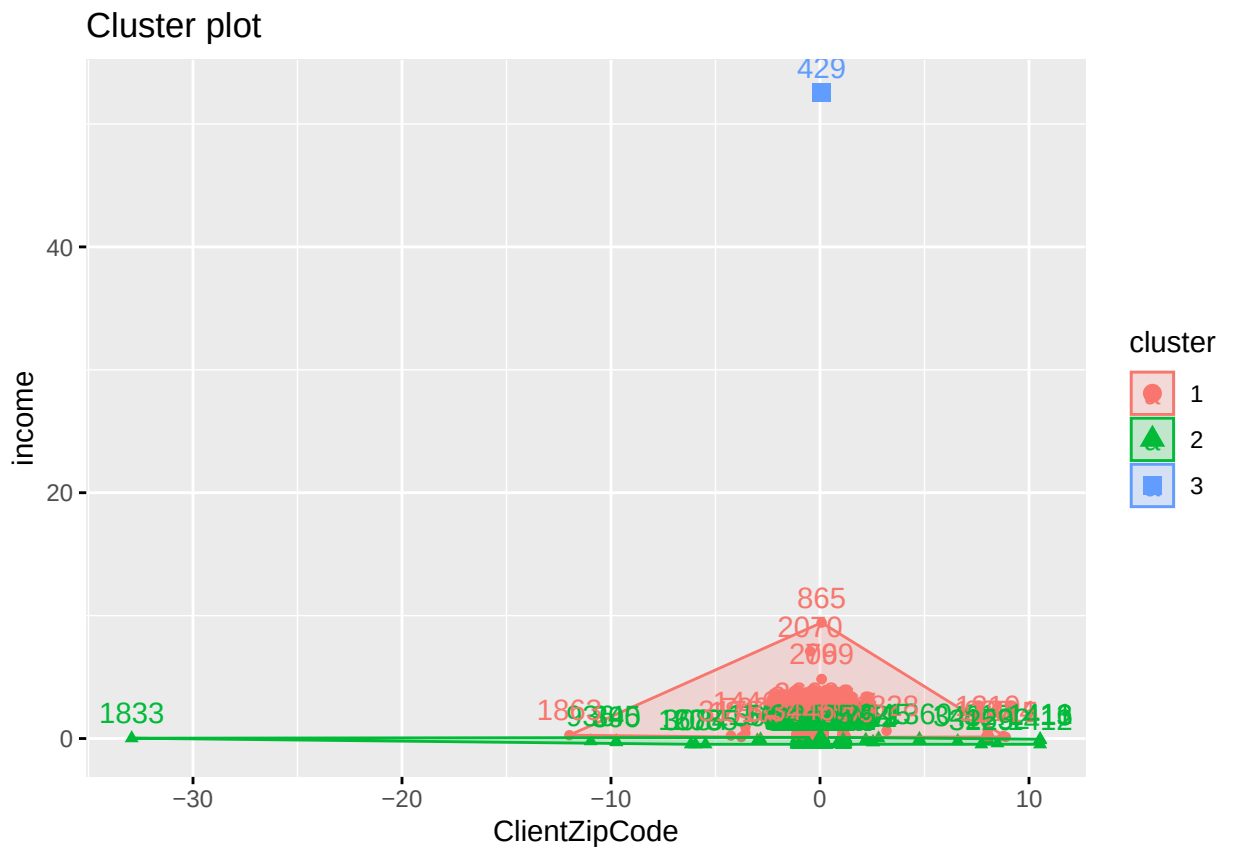
```
cluster<-kmeans(get,3)

plot(last,col = (cluster$cluster),xlim = c(26000, 29000), ylim = c(0, 50000))
```




```
library(factoextra)

k3 <- kmeans(last, centers = 3,nstart = 25)

k3$centers
```

```
##   ClientZipCode      income
## 1      27597.99   25409.151
## 2      27605.46    6326.125
## 3      27610.00 1500000.000
```

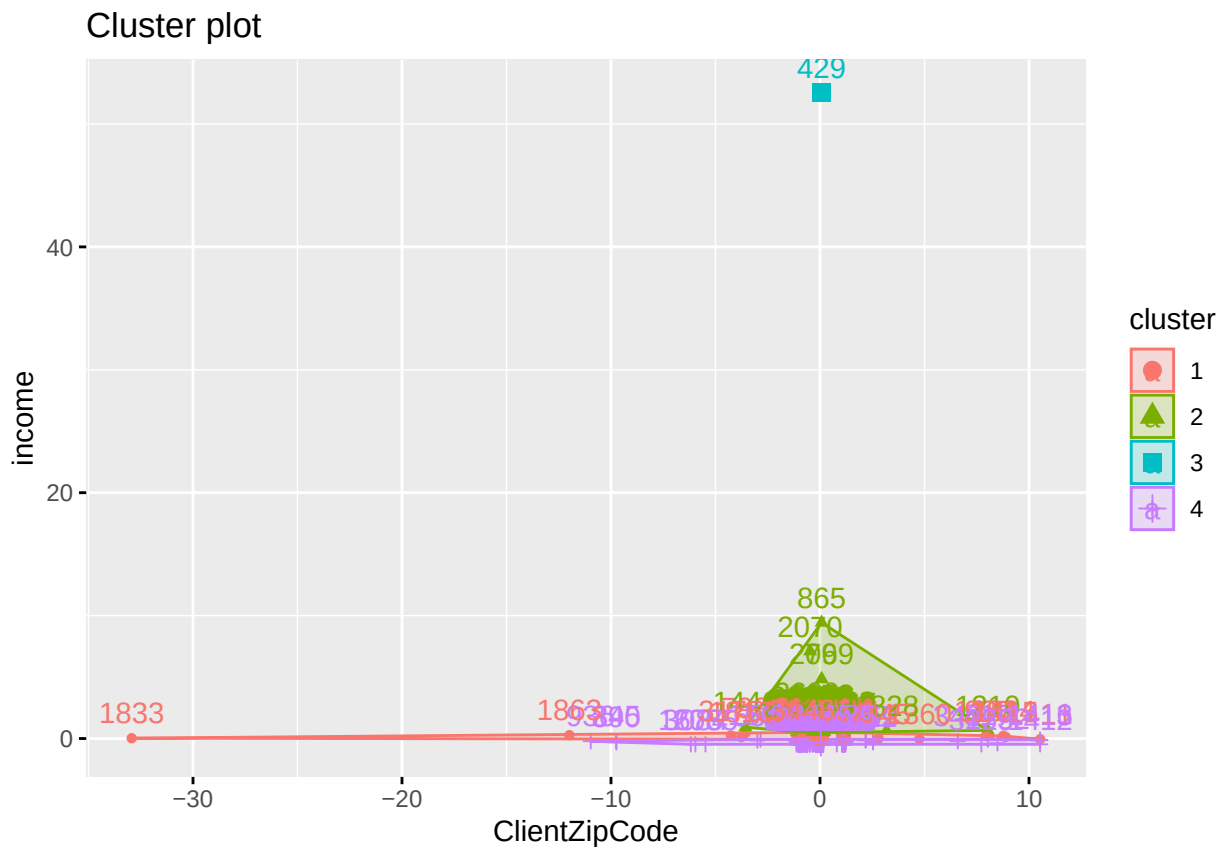
```
fviz_cluster(k3, data = last)
```




```
k4 <- kmeans(last, centers = 4,nstart = 25)

k4$centers
```

```
##   ClientZipCode      income
## 1      27599.40   18335.201
## 2      27596.51   36298.926
## 3      27610.00 1500000.000
## 4      27606.55    4574.016
```

```
fviz_cluster(k4, data = last)
```




```
#hier.cluster<-dist(get,method = "euclidean")




#hc1 <- hclust(hier.cluster, method = "complete" )
#pam <- pam(hier.cluster,4, diss = FALSE)
#clusplot(pam, shade = FALSE,labels=2,col.clus="blue",col.p="red",span=FALSE,main="Cluster Mapping",cex



#seasonal trend?
#str(cleaned_STATCOM_data$Timestamp)

cluster_level <- k3$cluster
cluster_data <- cbind(cleaned,cluster_level)

hh <- cluster_data %>% group_by(cluster_level,Agency_Clean_Short) %>%
  summarize(percent = 100*n()/nrow(cluster_data))
hh
```

```
## # A tibble: 179 x 3
## # Groups:   cluster_level [3]
```

```
##    cluster_level Agency_Clean_Short               percent
##            <int> <chr>                              <dbl>
##  1             1 A Doorway to Hope                 0.377
##  2             1 Alliance Health                   0.551
##  3             1 Alliance Medical Ministry         0.0290
##  4             1 Alliance of Disability Advocates  0.116
##  5             1 Arc of the Triangle               0.0290
##  6             1 Caring Connections Ministry       0.0580
##  7             1 Cary Church of God                0.0290
##  8             1 CASA                              0.842
##  9             1 Catholic Charities                1.48
## 10             1 CCWJC                             1.02
## # ... with 169 more rows
```

```
#str(F)


# indicating the clients from different cluster may have different number of referals from  agency

#low income cluster
hh1 <- hh %>% filter(percent >=1,cluster_level==1)
knitr::kable(hh1)
```

cluster_level	Agency_Clean_Short	percent
1	Catholic Charities	1.479977
1	CCWJC	1.015670
1	Families Together	3.424260
1	Passage Home	1.247824
1	Salvation Army	1.073709
1	WCHS-Middle Class Express	1.567034
1	WCHS-Wake Prevent!	1.015670
1	WCPSS	8.908880

```
#lowest income cluster
hh2<-hh %>% filter(percent >=1 & cluster_level==2)
knitr::kable(hh2)
```

cluster_level	Agency_Clean_Short	percent
2	Alliance Health	9.866512
2	CASA	2.234475
2	CCWJC	2.524666
2	Durham VA	1.479977
2	Families Together	1.944283
2	Haven House	1.654092
2	InterAct	1.305862
2	Passage Home	4.962275
2	Salvation Army	1.392919
2	Triangle Family Services	4.033662
2	USCRI	1.073709
2	Wake County Human Services	1.712130

cluster_level	Agency_Clean_Short	percent
2	Wake FS&CPS	1.044690
2	Wake Supportive Housing	2.988973
2	WCHS-Maternal Child Health	4.439930
2	WCHS-Middle Class Express	1.567034
2	WCPSS	8.183401

```
# using the within 2/1 standard deviation to select the zip code.
# table or graphs involving other categoriacal variables,
# or clustering using SVI variables, in it. ;lets do it!!
#missing values need to take into account
#local host mapping in the useful code folder, need to try it,
```

Using data from cleaned original green chair data merged with SVI

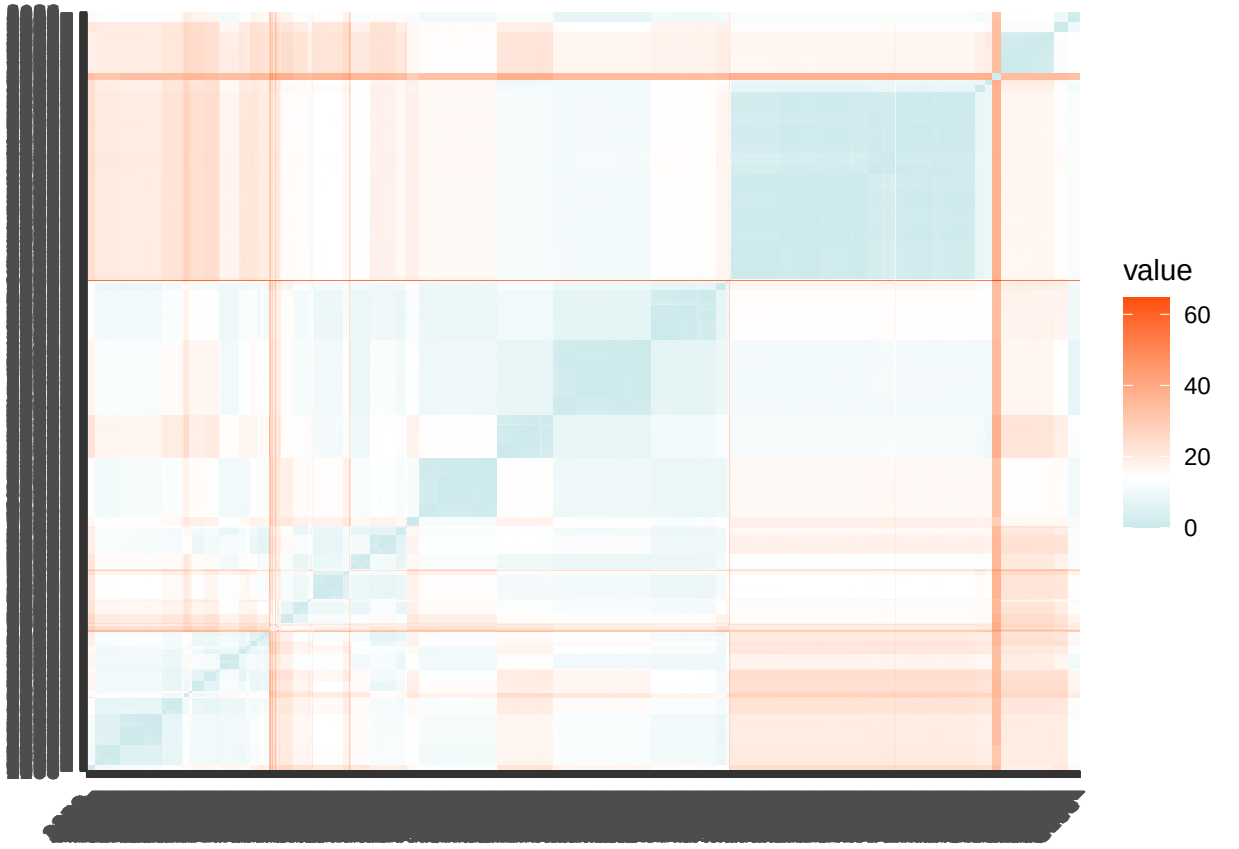
```
library(tidyverse)
library(cluster)    # clustering algorithms
library(factoextra) # clustering algorithms & visualization
merge <- read.csv("C:/Users/CKA/Documents/CKA/the-green-chair-project/merged_SVI/merged_CDC_GC_clean.csv
str(merge)
```

```
## 'data.frame':    5448 obs. of  220 variables:
##  $ X                    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID                   : int  36182 36183 36184 36185 36186 36187 36188 36189 36190 36191 ...
##  $ Timestamp            : chr  "2010-12-27" "2010-12-27" "2010-12-27" "2010-12-27" ...
##  $ Agency_Clean_Short   : chr  "StepUp Ministry" "StepUp Ministry" "StepUp Ministry" "StepUp Ministry"
##  $ Agency_Clean_Full    : chr  "StepUp Ministry" "StepUp Ministry" "StepUp Ministry" "StepUp Ministry"
##  $ ClientZipCode        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ ClientAge            : chr  "Adult" "Adult" "Adult" "Adult" ...
##  $ ClientGender         : chr  "Male" "Male" "Male" "Female" ...
##  $ Ethnicity            : chr  "N/A" "N/A" "N/A" "N/A" ...
##  $ Race                 : chr  "Other or Unknown" "Other or Unknown" "Other or Unknown" "Other or Unkn
##  $ Veteran              : chr  "N/A" "N/A" "N/A" "N/A" ...
##  $ Incarcerated         : chr  "N/A" "N/A" "N/A" "Yes" ...
##  $ Disability           : chr  "N/A" "N/A" "N/A" "N/A" ...
##  $ AnnualIncomeAmount   : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TotalHHNumber        : int  2 2 4 2 3 2 1 1 2 1 ...
##  $ NumAdultFemales      : int  1 NA 1 1 1 1 1 NA 1 1 ...
##  $ NumAdultMales        : int  1 1 1 NA NA NA NA 1 NA NA ...
##  $ NumChildren          : int  NA 1 2 1 2 1 NA NA 1 NA ...
##  $ MoreThan1HHMember    : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ HHMember1Age         : chr  "Adult (age 18 or over) Female" "Child (under age 18)" "Adult (age 18 o
##  $ HHMember1Gender      : chr  "Adult (age 18 or over) Female" "Child (under age 18)" "Adult (age 18 o
##  $ HHMember1Ethnicity   : chr  NA NA NA NA ...
##  $ HHMember1Race        : chr  NA NA NA NA ...
##  $ HHMember1BedReq      : chr  NA NA NA NA ...
##  $ HHMember1School      : chr  NA NA NA NA ...
##  $ HHMember1GradeLevel  : chr  NA NA NA NA ...
##  $ MoreThan2HHMember    : chr  NA NA "Yes" "Yes" ...
##  $ HHMember2Age         : chr  NA NA "Child (under age 18)" NA ...
##  $ HHMember2Gender      : chr  NA NA "Child (under age 18)" NA ...
##  $ HHMember2Ethnicity   : chr  NA NA NA NA ...
```

```
##  $ HHMember2Race       : chr  NA NA NA NA ...
##  $ HHMember2BedReq     : chr  NA NA NA NA ...
##  $ HHMember2School     : chr  NA NA NA NA ...
##  $ HHMember2GradeLevel : chr  NA NA NA NA ...
##  $ MoreThan3HHMember   : chr  NA NA "Yes" NA ...
##  $ HHMember3Age        : chr  NA NA "Child (under age 18)" NA ...
##  $ HHMember3Gender     : chr  NA NA "Child (under age 18)" NA ...
##  $ HHMember3Ethnicity  : chr  NA NA NA NA ...
##  $ HHMember3Race       : chr  NA NA NA NA ...
##  $ HHMember3BedReq     : chr  NA NA NA NA ...
##  $ HHMember3School     : chr  NA NA NA NA ...
##  $ HHMember3GradeLevel : chr  NA NA NA NA ...
##  $ MoreThan4HHMember   : chr  NA NA NA NA ...
##  $ HHMember4Age        : chr  NA NA NA NA ...
##  $ HHMember4Gender     : chr  NA NA NA NA ...
##  $ HHMember4Ethnicity  : chr  NA NA NA NA ...
##  $ HHMember4Race       : chr  NA NA NA NA ...
##  $ HHMember4BedReq     : chr  NA NA NA NA ...
##  $ HHMember4School     : chr  NA NA NA NA ...
##  $ HHMember4GradeLevel : chr  NA NA NA NA ...
##  $ MoreThan5HHMember   : chr  NA NA NA NA ...
##  $ HHMember5Age        : chr  NA NA NA NA ...
##  $ HHMember5Gender     : chr  NA NA NA NA ...
##  $ HHMember5Ethnicity  : chr  NA NA NA NA ...
##  $ HHMember5Race       : chr  NA NA NA NA ...
##  $ HHMember5BedReq     : chr  NA NA NA NA ...
##  $ HHMember5School     : chr  NA NA NA NA ...
##  $ HHMember5GradeLevel : chr  NA NA NA NA ...
##  $ MoreThan6HHMember   : chr  NA NA NA NA ...
##  $ HHMember6Age        : chr  NA NA NA NA ...
##  $ HHMember6Gender     : chr  NA NA NA NA ...
##  $ HHMember6Ethnicity  : chr  NA NA NA NA ...
##  $ HHMember6Race       : chr  NA NA NA NA ...
##  $ HHMember6BedReq     : chr  NA NA NA NA ...
##  $ HHMember6School     : chr  NA NA NA NA ...
##  $ HHMember6GradeLevel : chr  NA NA NA NA ...
##  $ MoreThan7HHMember   : chr  NA NA NA NA ...
##  $ HHMember7Age        : chr  NA NA NA NA ...
##  $ HHMember7Gender     : chr  NA NA NA NA ...
##  $ HHMember7Ethnicity  : chr  NA NA NA NA ...
##  $ HHMember7Race       : chr  NA NA NA NA ...
##  $ HHMember7BedReq     : chr  NA NA NA NA ...
##  $ HHMember7School     : chr  NA NA NA NA ...
##  $ HHMember7GradeLevel : chr  NA NA NA NA ...
##  $ MoreThan8HHMember   : chr  NA NA NA NA ...
##  $ HHMember8Age        : chr  NA NA NA NA ...
##  $ HHMember8Gender     : chr  NA NA NA NA ...
##  $ HHMember8Ethnicity  : chr  NA NA NA NA ...
##  $ HHMember8Race       : chr  NA NA NA NA ...
##  $ HHMember8BedReq     : chr  NA NA NA NA ...
##  $ HHMember8School     : chr  NA NA NA NA ...
##  $ HHMember8GradeLevel : chr  NA NA NA NA ...
##  $ MoreThan9HHMember   : logi  NA NA NA NA NA NA ...
##  $ HHMember9Age        : logi  NA NA NA NA NA NA ...
```

```
##  $ HHMember9Gender     : logi  NA NA NA NA NA NA ...
##  $ HHMember9Ethnicity  : logi  NA NA NA NA NA NA ...
##  $ HHMember9Race       : logi  NA NA NA NA NA NA ...
##  $ HHMember9BedReq     : logi  NA NA NA NA NA NA ...
##  $ HHMember9School     : logi  NA NA NA NA NA NA ...
##  $ HHMember9GradeLevel : logi  NA NA NA NA NA NA ...
##  $ Morethan10HHInfo    : logi  NA NA NA NA NA NA ...
##  $ QueenBeds           : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ Assistance          : chr  "Transition" "Transition" "Transition" "Transition" ...
##  $ Circumstance        : chr  "Homelessness" "Homelessness" "Addiction / Recovery" "N/A" ...
##  $ HomeSize            : chr  "1 Bedroom" "2 Bedrooms" "3 Bedrooms" "2 Bedrooms" ...
##  $ FurnishingFeePayment: chr  "N/A" "N/A" "N/A" "N/A" ...
##  $ COVID.19            : chr  "N/A" "N/A" "N/A" "N/A" ...
##  $ Cribs               : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ TwinBeds            : int  NA NA NA NA NA NA NA NA NA NA ...
##   [list output truncated]
```

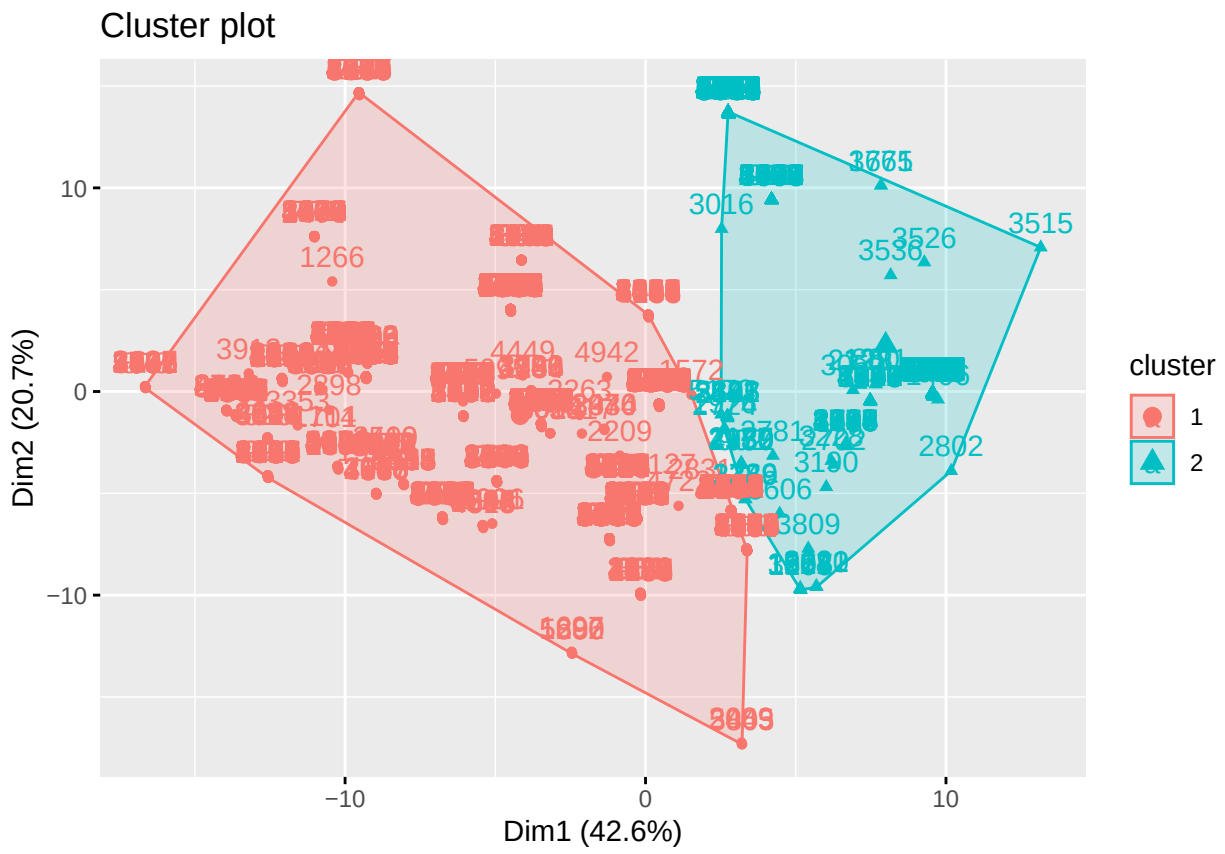
```
View(merge)

# get rid of the Household number information for this analysis
short <- merge %>% select(-starts_with(c("HH","More")))
last2 <- short %>% select(Agency_Clean_Short,Race,ClientZipCode,AnnualIncomeAmount,TotalHHNumber,31:147)
View(last2)


# return the objects that does not contain any NA values in the last dataset.


# next step : clustering:
df <- na.omit(last2)
View(df)


# center and scale the matrix
scaled <- scale(df[,-1:-2])

#computing Euclidean distance between the rows of this data
distance <- get_dist(scaled)

fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))
```



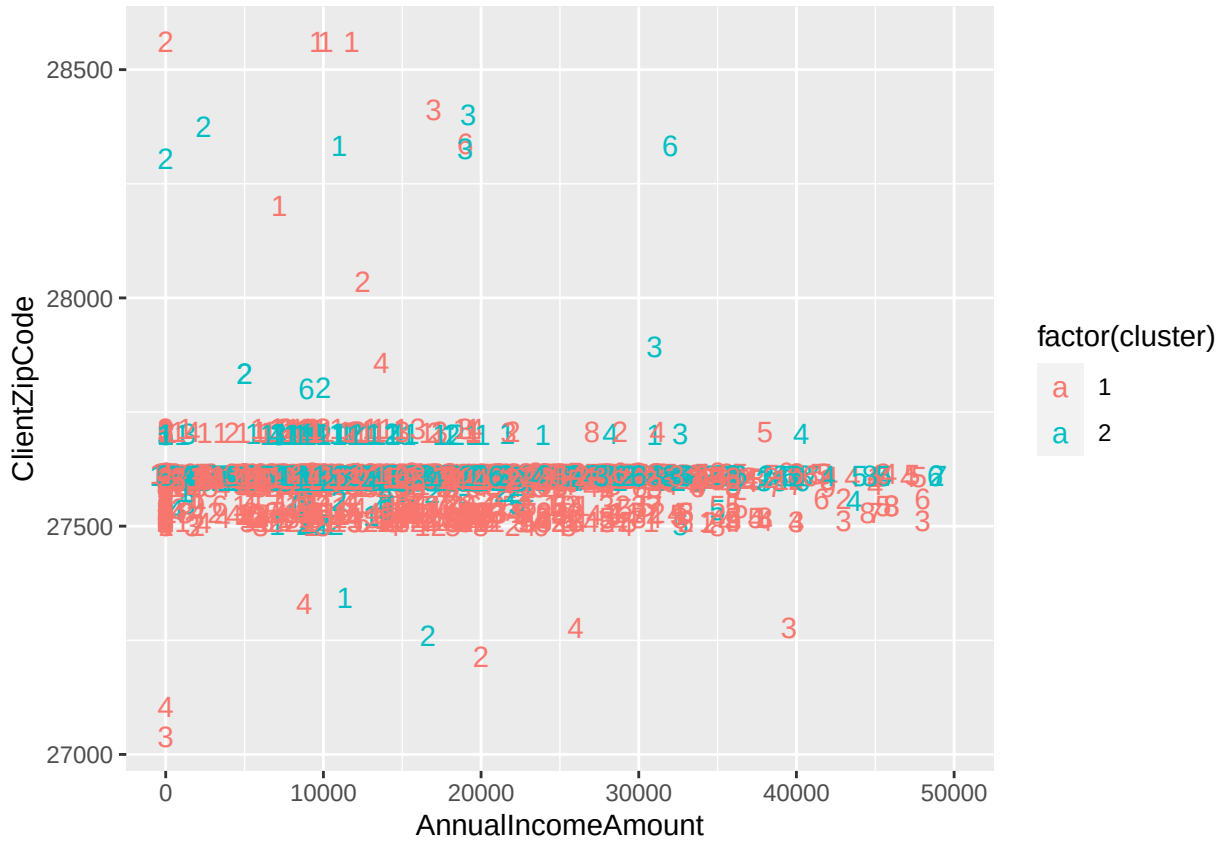

```
try2 <- kmeans(scaled, centers = 2, nstart = 25)
str(try2)
```

```
## List of 9
##  $ cluster     : Named int [1:3402] 1 1 1 1 2 1 1 1 1 1 ...
##   ..- attr(*, "names")= chr [1:3402] "321" "676" "677" "678" ...
##  $ centers     : num [1:2, 1:120] -0.1077 0.2063 -0.0191 0.0365 -0.0116 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:120] "ClientZipCode" "AnnualIncomeAmount" "TotalHHNumber" "AREA_SQMI" ...
##  $ totss       : num 408120
##  $ withinss    : num [1:2] 208249 73630
##  $ tot.withinss: num 281879
##  $ betweenss   : num 126241
##  $ size        : int [1:2] 2235 1167
##  $ iter        : int 1
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

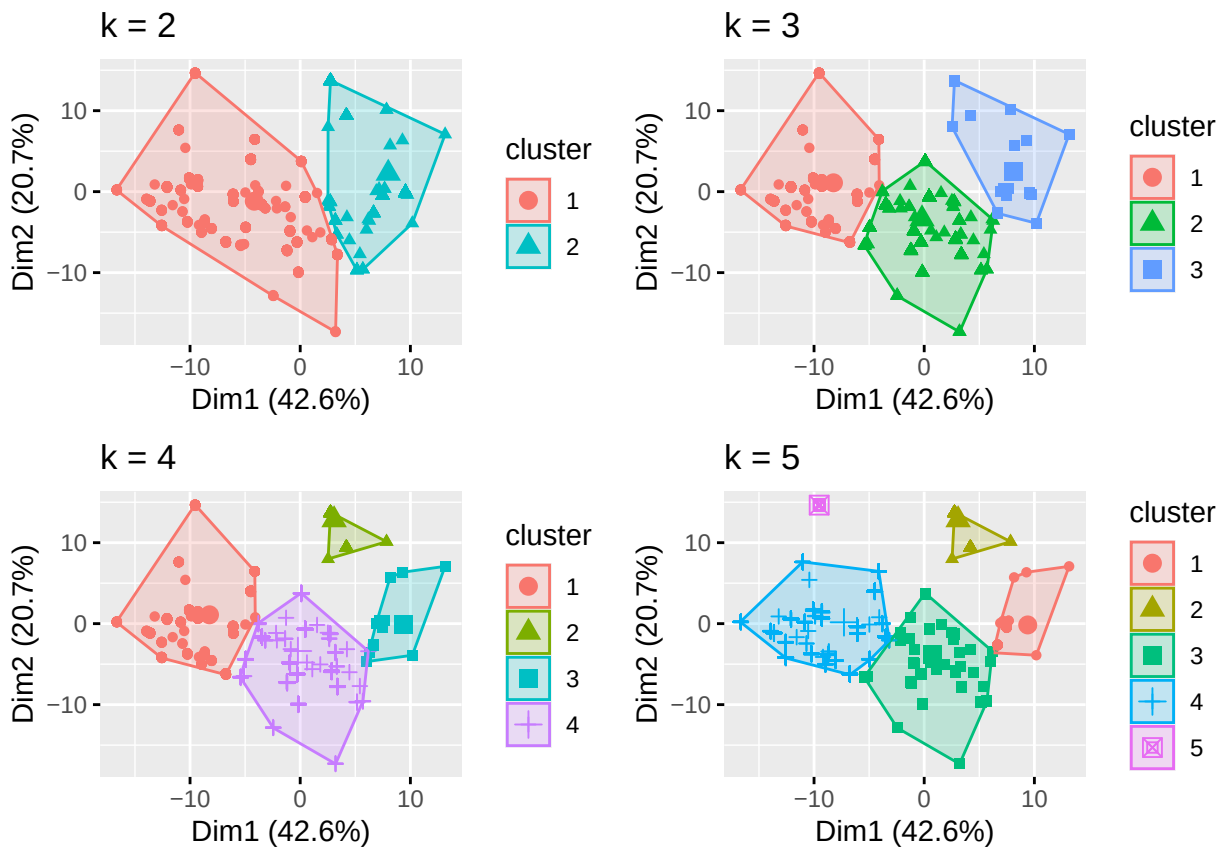
```
fviz_cluster(try2, data = scaled)
```



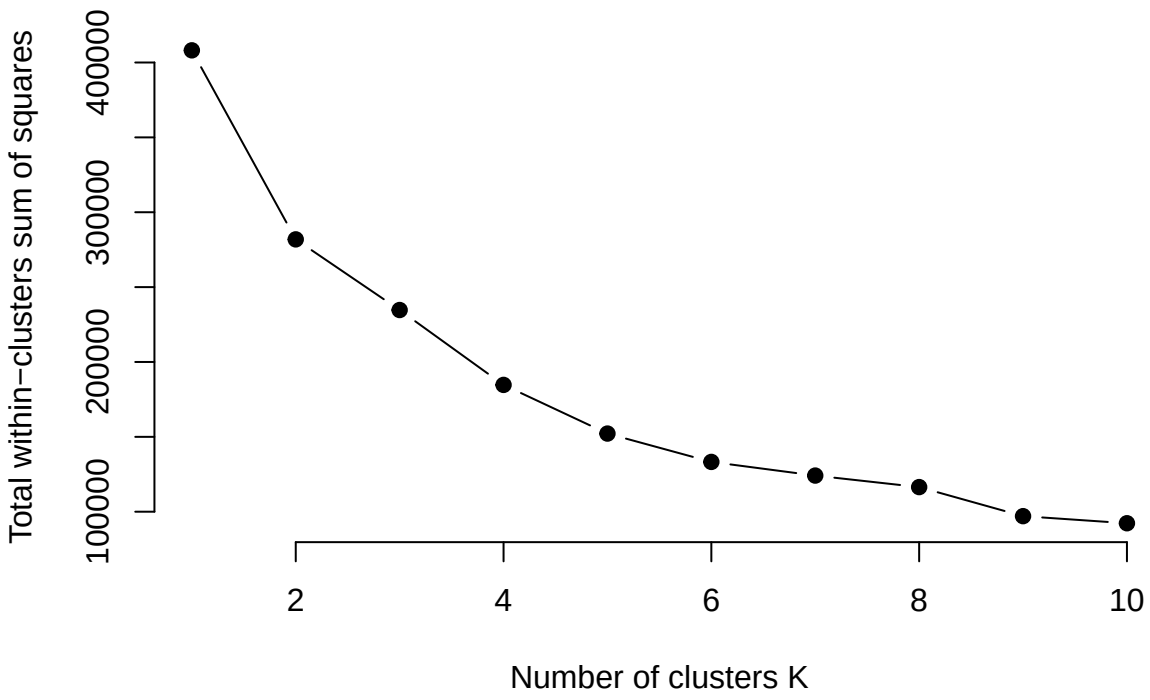

```
df %>%
  as_tibble() %>%
  mutate(cluster = try2$cluster) %>%
  ggplot(aes(AnnualIncomeAmount,ClientZipCode,  color = factor(cluster),label = TotalHHNumber)) +
  geom_text() + xlim(0,50000)
```



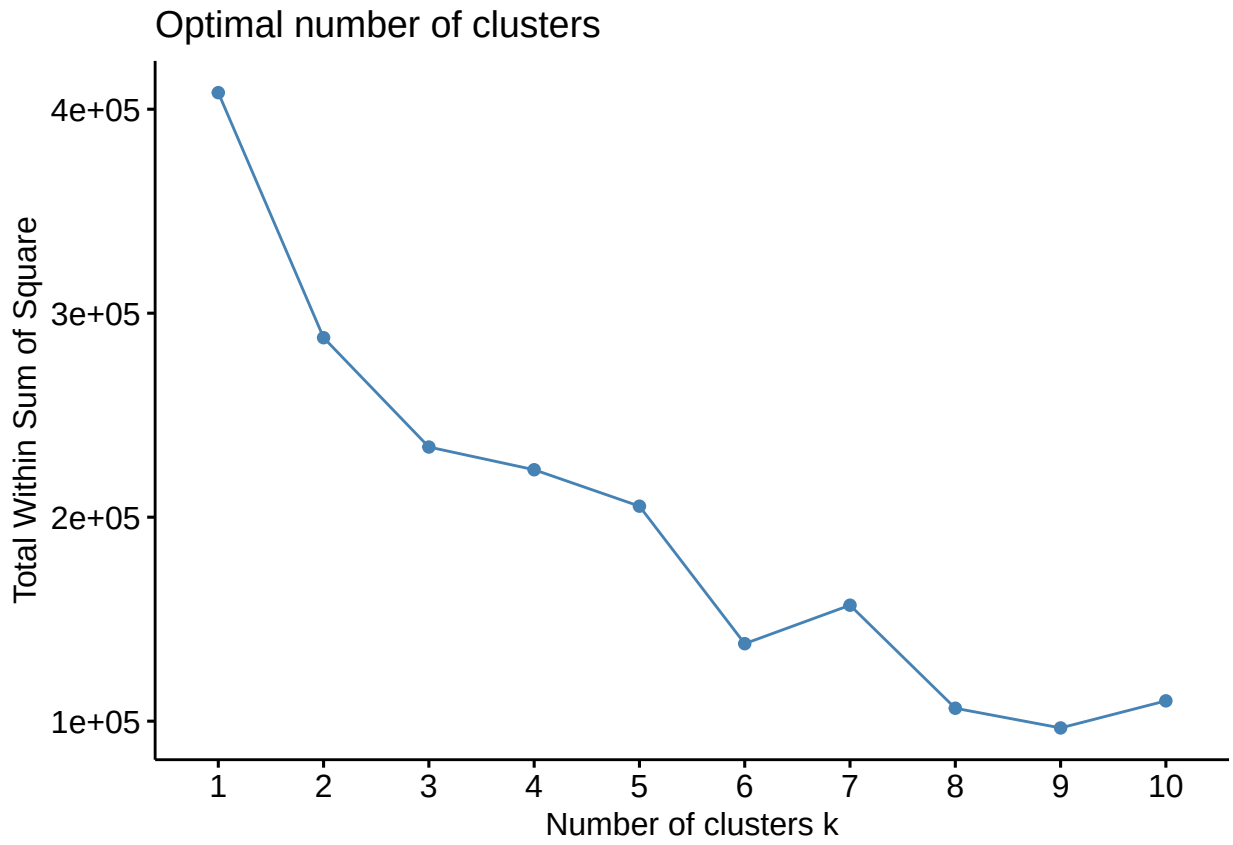

```
try3 <- kmeans(scaled, centers = 3, nstart = 25)
try4 <- kmeans(scaled, centers = 4, nstart = 25)
try5 <- kmeans(scaled, centers = 5, nstart = 25)

p1 <- fviz_cluster(try2, geom = "point", data = scaled) + ggtitle("k = 2")
p2 <- fviz_cluster(try3, geom = "point",  data = scaled) + ggtitle("k = 3")
p3 <- fviz_cluster(try4, geom = "point",  data = scaled) + ggtitle("k = 4")
p4 <- fviz_cluster(try5, geom = "point",  data = scaled) + ggtitle("k = 5")

library(gridExtra)
grid.arrange(p1, p2, p3, p4, nrow = 2)
```



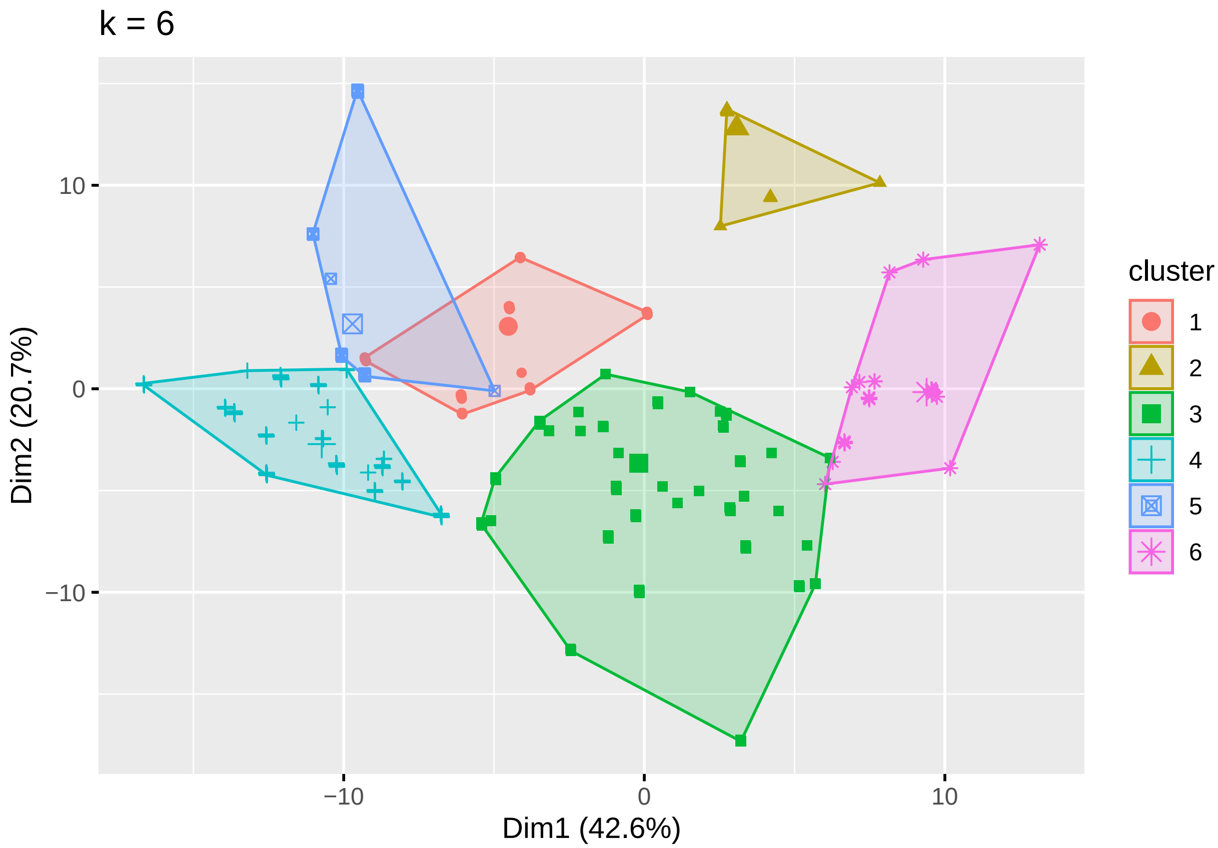

```
# when k = 4 or 5 it captures some outliers within the cluster

# determine the number of cluster to use in this case:

set.seed(12345)
within_sum_squares <- function(k) {
  kmeans(scaled, k, nstart = 10 )$tot.withinss
}

# Compute and plot wss for k = 1 to k = 15
k.values <- 1:10

# extract wss for 2-15 clusters
wss_values <- map_dbl(k.values, within_sum_squares)

plot(k.values, wss_values,
       type="b", pch = 19, frame = FALSE,
       xlab="Number of clusters K",
       ylab="Total within-clusters sum of squares")
```




```
# 5 looks like a good one



# or use this method
fviz_nbclust(scaled, kmeans, method = "wss")
```



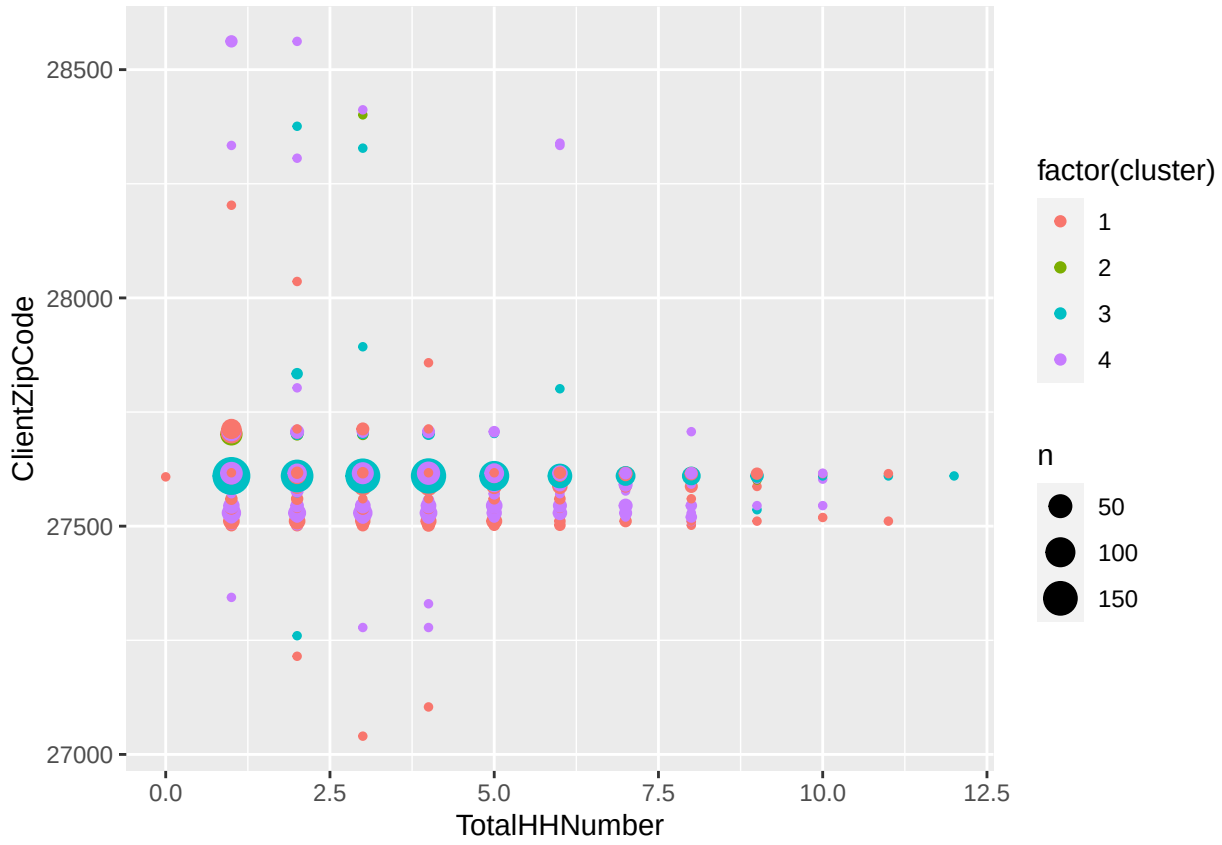

```
#there is a back shift of within sum of squares at k = 6.


# we want the within sums of squares to be as small as possible,
# also want to control the number of k used.
# so tried  within 2 standard deviation method. to select the best k




#fviz_nbclust(scaled, kmeans, method = "silhouette")



try6 <- kmeans(scaled, centers = 6, nstart = 25)
fviz_cluster(try6, geom = "point",  data = scaled) + ggtitle("k = 6")
```



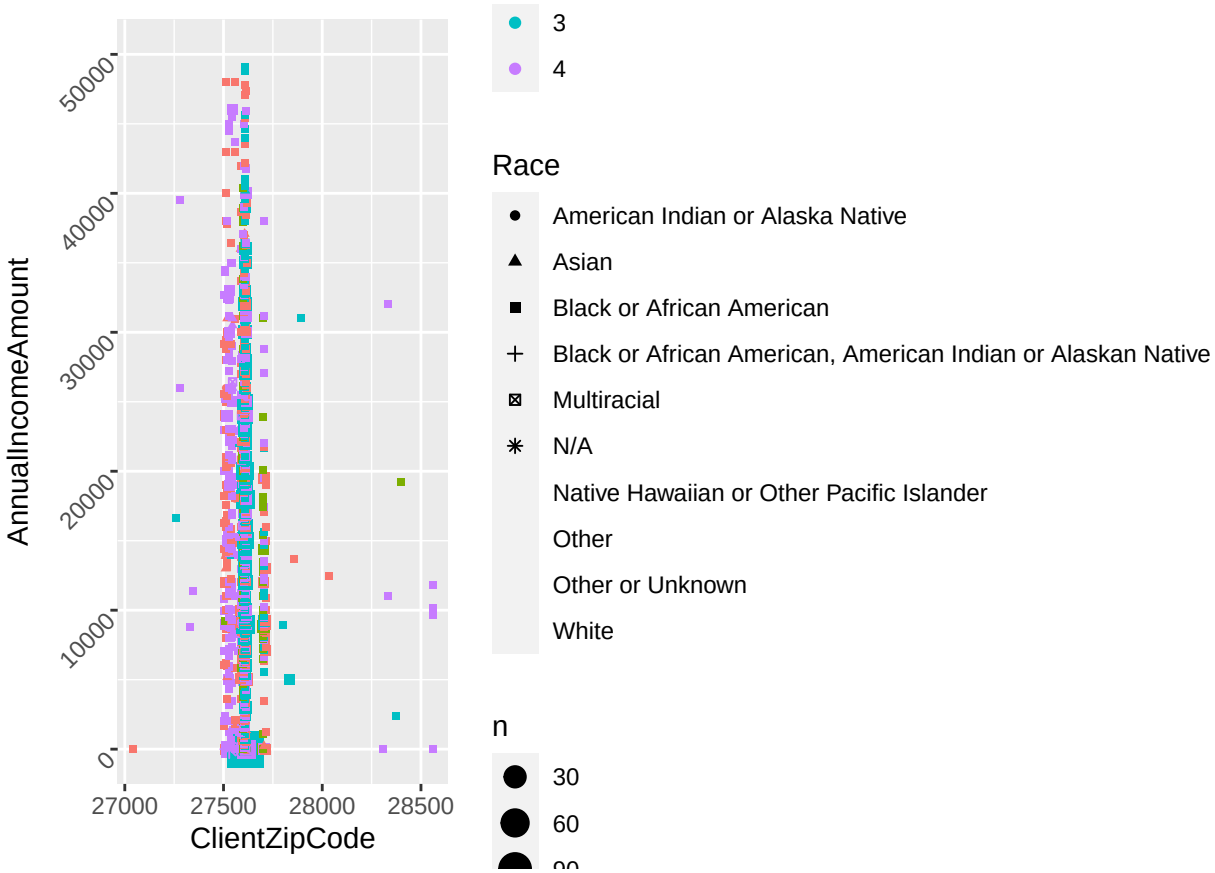

```r
# using k = 4 or 5
# summarize by mean for each cluster using the df dataset
EE <- df %>%
  mutate(Cluster = try4$cluster) %>%
  group_by(Cluster) %>%
  summarise_all("mean")

knitr::kable(EE)
```

Clu[illegible]	[illegible]	[illegible]
1	NA NA	[illegible]
2	NA NA	[illegible]
3	NA NA	[illegible]
4	NA NA	[illegible]

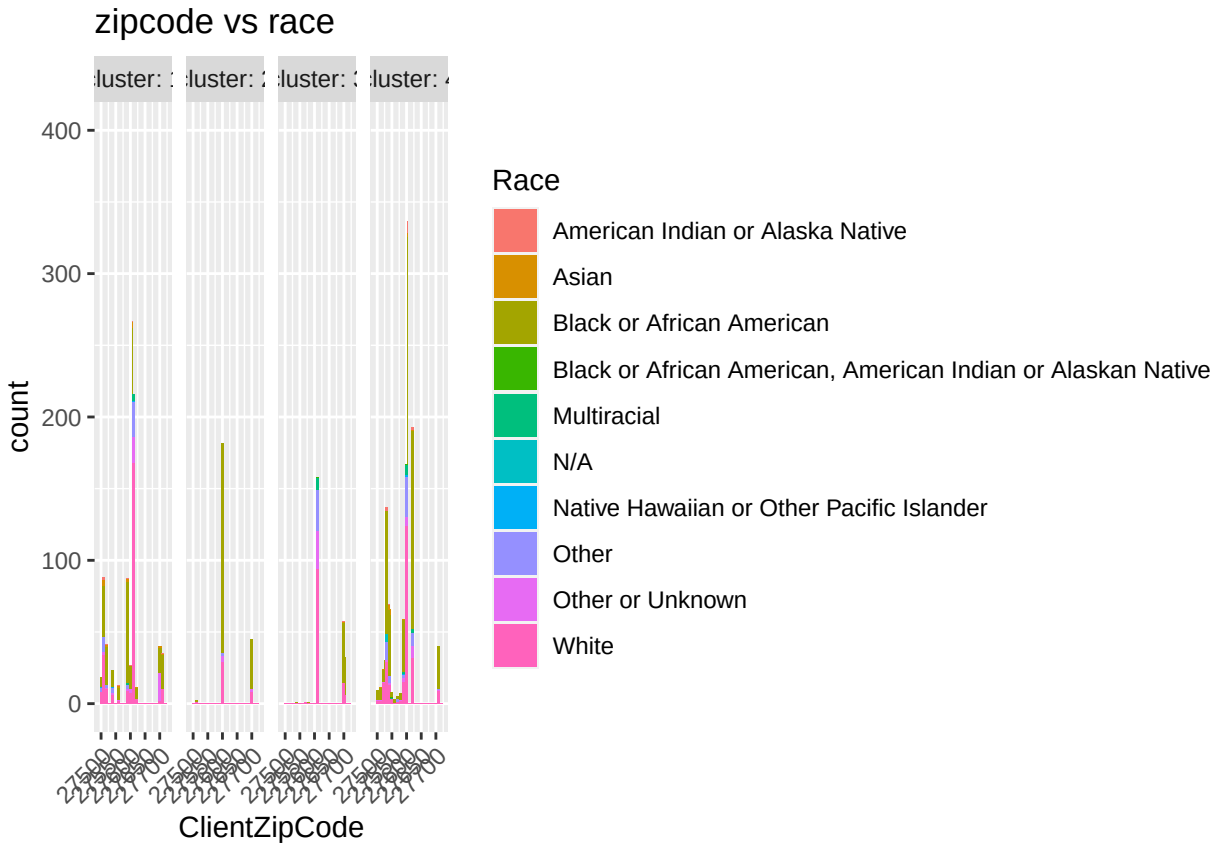
```r
# plot household number versus zipcode
df %>%
  mutate(cluster = try4$cluster) %>%
  ggplot(aes(TotalHHNumber,ClientZipCode,color = factor(cluster))) +
  geom_count()
```




```
#client zipcode vs cluster
df %>%
  mutate(cluster = try4$cluster) %>%
  ggplot(aes(ClientZipCode,AnnualIncomeAmount,color = factor(cluster))) +
  geom_count(aes(shape = Race))+
  theme(axis.text.y  = element_text(angle = 45, vjust = 1, hjust = 1)) +
  ylim(0, 50000)
```




```
# zipcode counts for each cluster (race category)
df %>%
  mutate(cluster = try4$cluster) %>%
  ggplot(aes(ClientZipCode, fill = as.factor(Race))) + geom_bar(position = "stack")+
    facet_grid(cols = vars(cluster),labeller = label_both)+
  scale_fill_discrete(name = "Race") +
  labs (title = "zipcode vs race")+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))+
  xlim(27490, 27730) + ylim(0,400) + stat_bin(binwidth = 10)
```




```
cluster_level2 <- try4$cluster
cluster_data2 <- cbind(df,cluster_level2)

hh2 <- cluster_data2 %>% group_by(cluster_level2,Agency_Clean_Short) %>%
  summarize(percent = 100*n()/nrow(cluster_data2))
WW <- hh2 %>% filter(percent >= 0.5)
knitr::kable(WW)
```

cluster_level2	Agency_Clean_Short	percent
1	Alliance Health	2.7924750
1	CASA	1.2639624
1	Catholic Charities	0.7936508
1	CCWJC	1.1757790
1	Families Together	1.4109347
1	Haven House	1.0875955
1	InterAct	1.0288066
1	Passage Home	1.3815403
1	StepUp Ministry	0.5291005
1	Triangle Family Services	1.9988242
1	USCRI	0.7054674
1	Wake County Human Services	0.6466784
1	Wake FS&CPS	0.5878895
1	Wake Supportive Housing	1.2639624
1	WCHS-Maternal Child Health	1.3227513

cluster_level2	Agency_Clean_Short	percent
1	WCHS-Middle Class Express	0.7936508
1	WCPSS	5.8788948
2	Alliance Health	0.9112287
2	Passage Home	0.7936508
2	Wake Supportive Housing	0.5878895
2	WCPSS	0.7348618
3	Alliance Health	2.7336861
3	CASA	1.1169900
3	Catholic Charities	0.7642563
3	CCWJC	0.9406232
3	Families Together	1.8812463
3	Passage Home	1.8518519
3	Salvation Army	0.8818342
3	Triangle Family Services	1.4109347
3	Wake County Human Services	0.8230453
3	Wake FS&CPS	0.5878895
3	WCHS-Maternal Child Health	1.2639624
3	WCHS-Middle Class Express	1.0875955
3	WCPSS	4.6149324
4	Alliance Health	3.9976484
4	Catholic Charities	0.8524397
4	CCWJC	1.2051734
4	Durham VA	0.6466784
4	Families Together	1.9106408
4	Family Promise	0.6466784
4	InterAct	0.6466784
4	Passage Home	2.1751911
4	Salvation Army	0.9700176
4	StepUp Ministry	0.5878895
4	Triangle Family Services	1.4109347
4	Wake County Human Services	0.7642563
4	Wake FS&CPS	0.5878895
4	Wake Supportive Housing	1.4109347
4	WCHS-Maternal Child Health	1.9694297
4	WCHS-Middle Class Express	1.1757790
4	WCHS-Wake Prevent!	0.8818342
4	WCPSS	5.9670782