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")
```

```
## Warning: One or more parsing issues, see `problems()` for details
```

```
## Rows: 5448 Columns: 102
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (66): Agency_Clean_Short, Agency_Clean_Full, ClientAge, ClientGender, E...
## dbl  (11): ID, ClientZipCode, AnnualIncomeAmount, TotalHHNumber, NumAdultFem...
## lgl  (24): HHMember6School, HHMember7Gender, HHMember7Ethnicity, HHMember7Ra...
## date  (1): Timestamp
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
tgcp_demog <- read_excel("C:/Users/CKA/Downloads/STATCOM_data.xlsx", col_types = "text")
```

```
## New names:
## * `` -> ...1
```

```
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)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.3     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

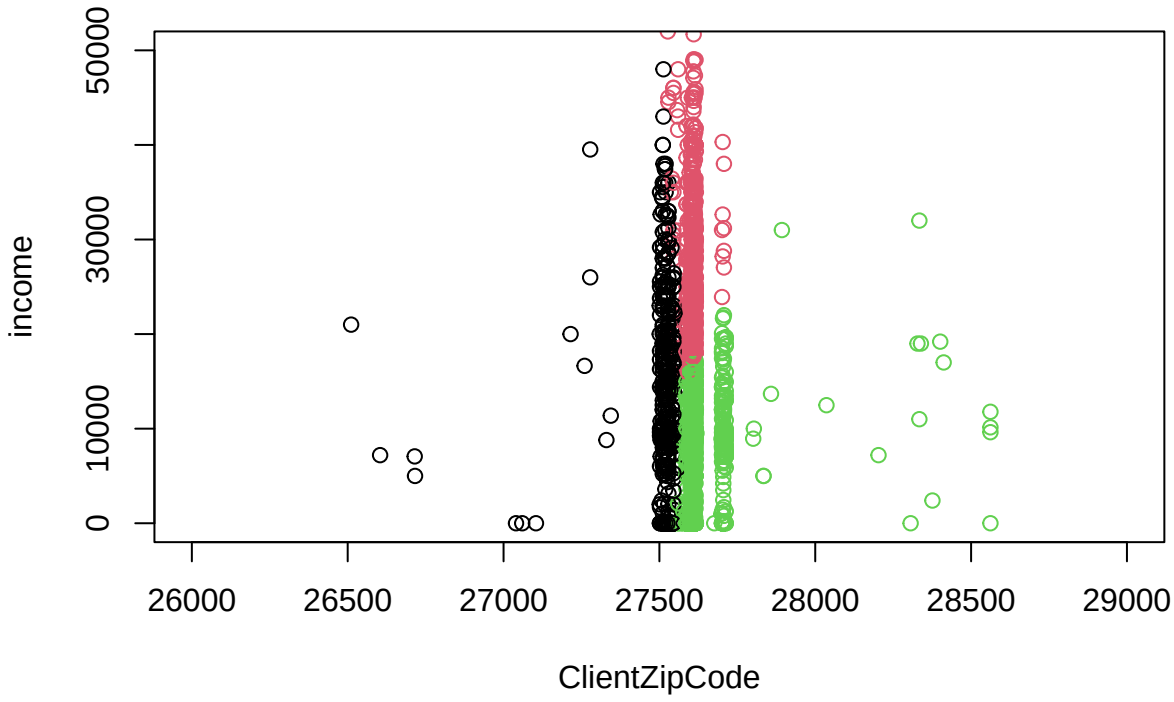
```
cleaned <- all%>% filter(!is.na(ClientZipCode),!is.na(income_amount))
final <- cleaned%>% select(ClientZipCode,income_amount)
#normalizerd data varaibles zipcode and income_amount.
#str(final)
income<-as.numeric(final$income_amount)
Z <-cbind(final,income)
last<-Z[,-2]

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)
```

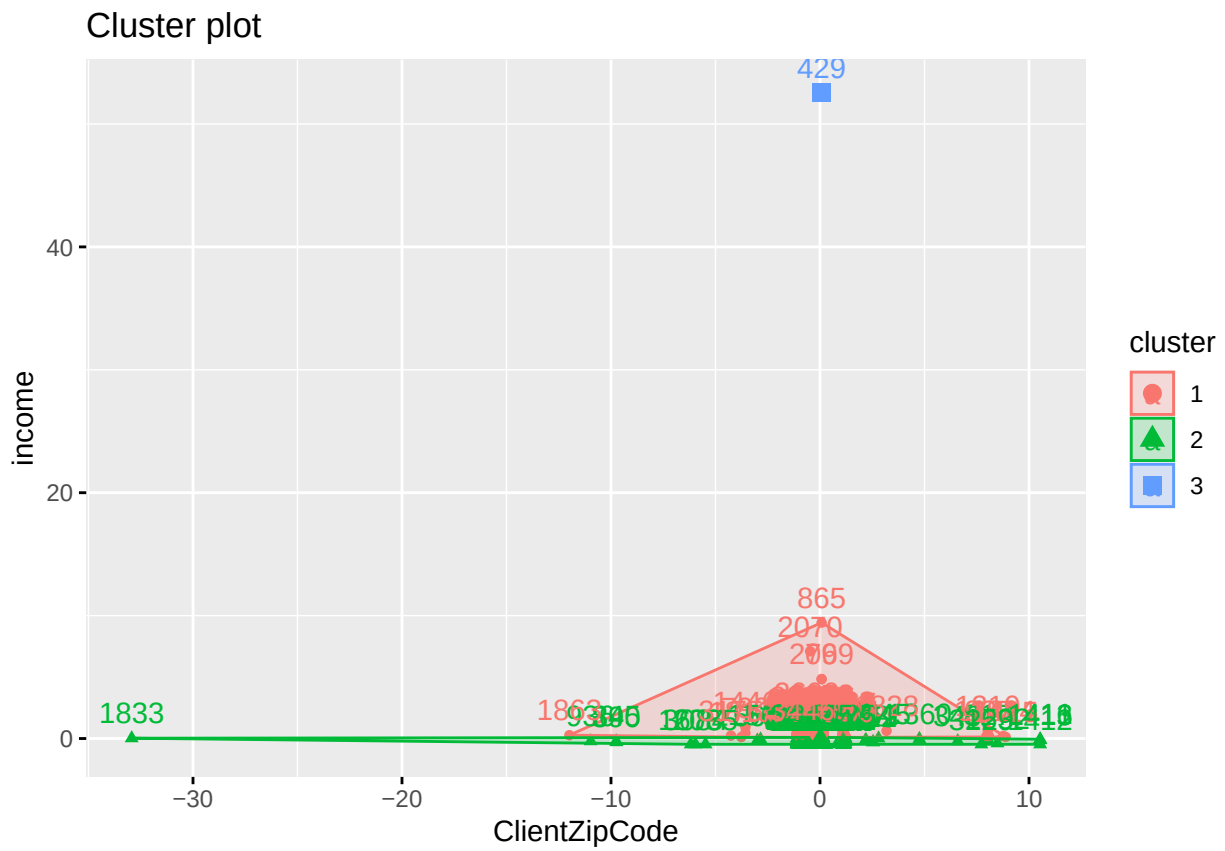
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
k2 <- kmeans(last, centers = 3,nstart = 25)

k2$centers
```

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

```
fviz_cluster(k2, 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 <- k2$cluster
cluster_data <- cbind(cleaned,cluster_level)

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

```
## `summarise()` has grouped output by 'cluster_level'. You can override using the `.groups` argument.
```

```
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

cluster_level	Agency_Clean_Short	percent
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
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

“‘