clustering exploration

Jiatao Wang

12/1/2021

information from the dataset cleaned_STATCOM_data_SVI.rds

```
cleaned_last <- readRDS("~/CKA/the-green-chair-project/cleaned_STATCOM_data_SVI.rds")
```

Including Plots

You can also embed plots, for example:

```
library(tidyverse)
```

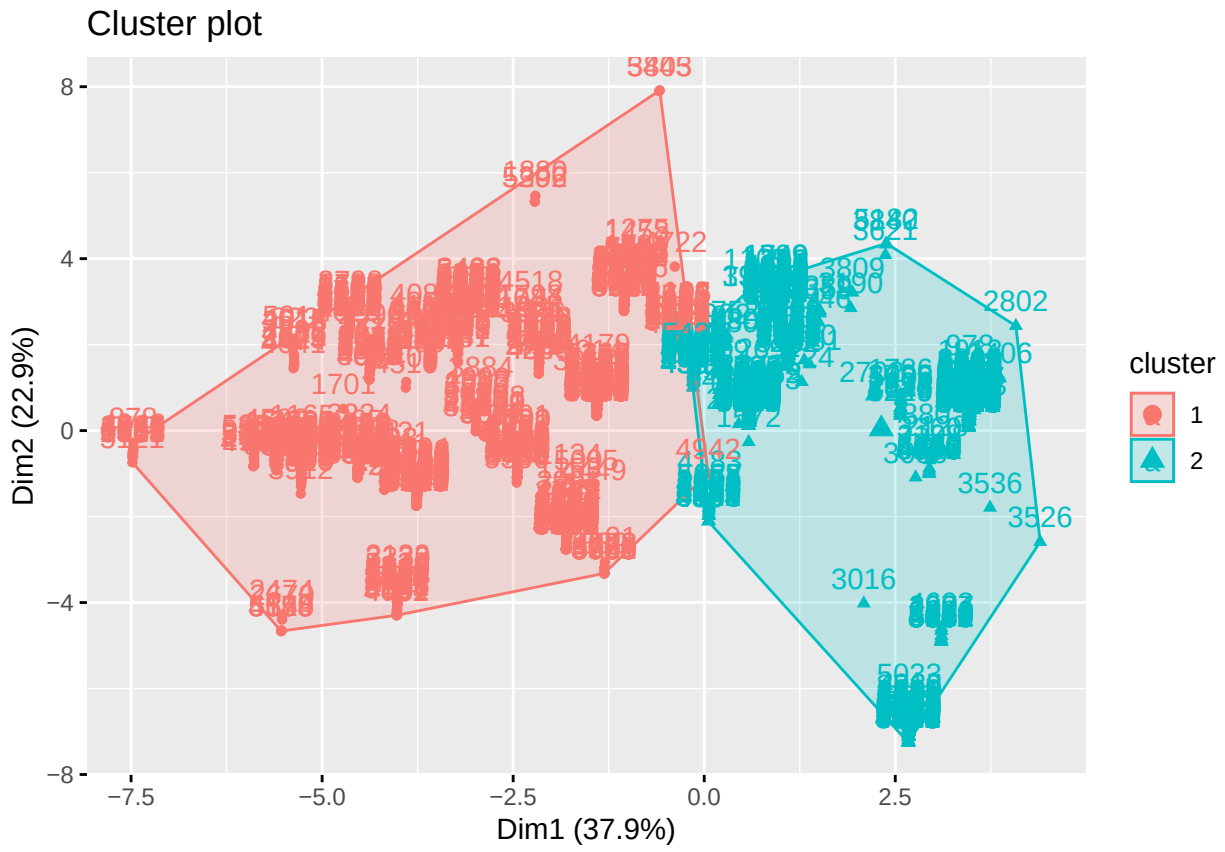
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.3     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

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

```
library(cluster)    # clustering algorithms
library(factoextra) # clustering algorithms & visualization
```

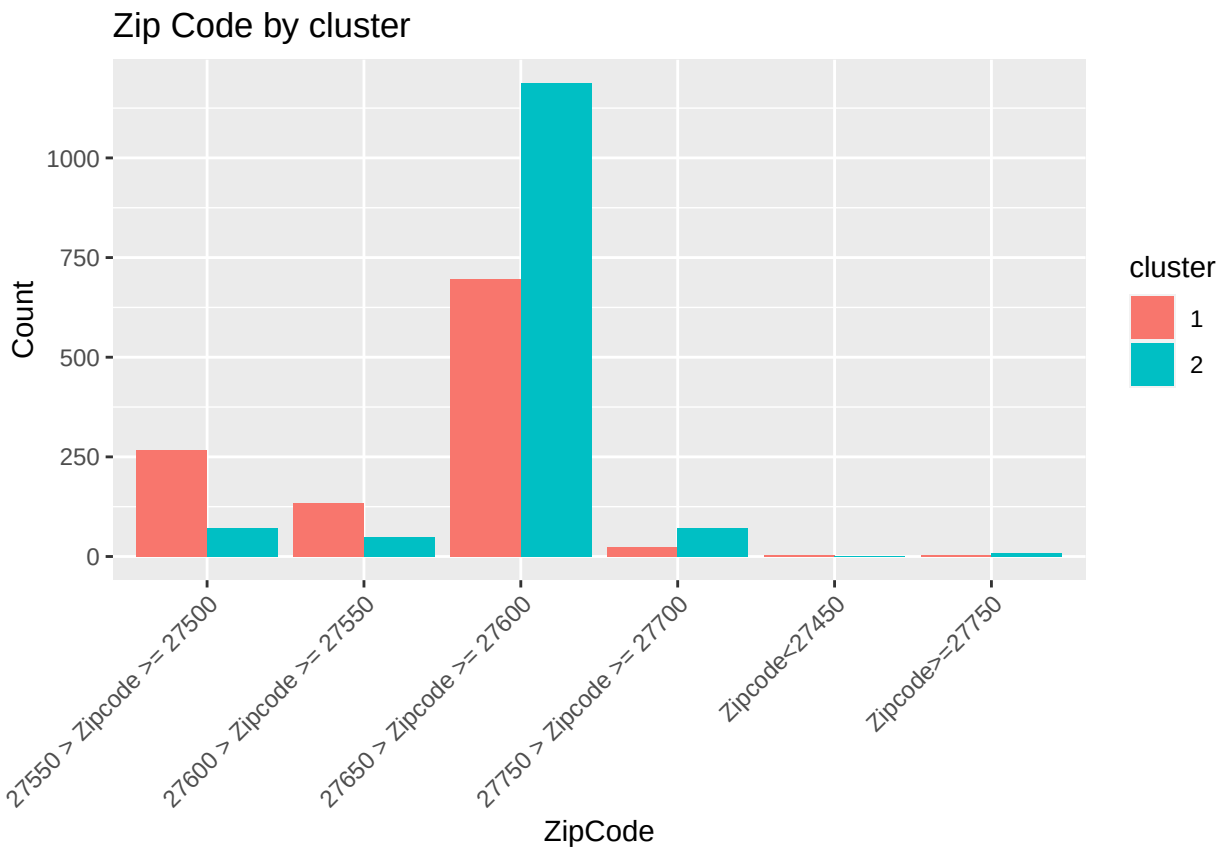
```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
# get rid of the Household number information for this analysis
short <- cleaned_last %>% select(-starts_with(c("HH","More")))
last2 <- short %>% select(ClientZipCode,Circumstance,Agency_Clean_Short,AnnualIncomeAmount,TotalHHNumbe
View(last2)


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


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

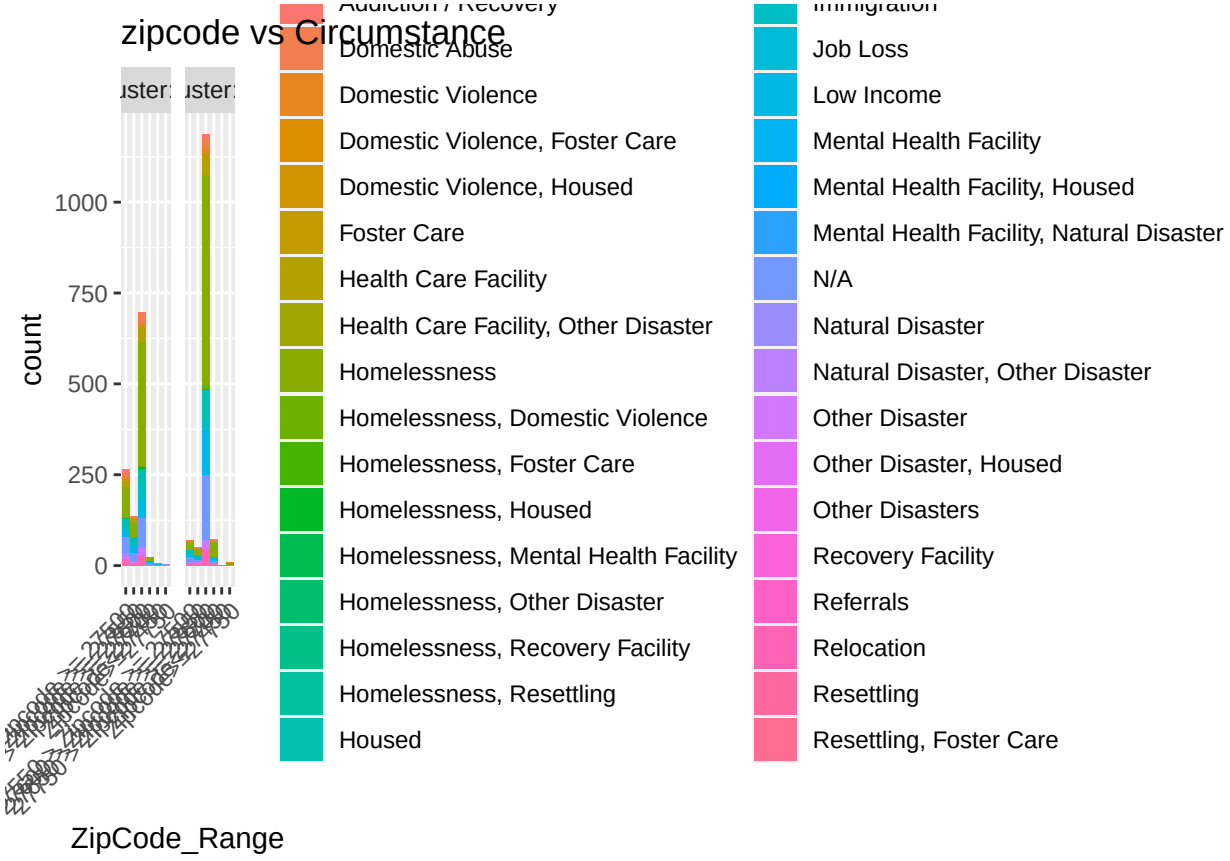
```
View(df)


v<- ifelse(df$ClientZipCode >= 27750, "Zipcode>=27750",
       ifelse(df$ClientZipCode >= 27700, "27750 > Zipcode >= 27700",
              ifelse(df$ClientZipCode >= 27650, "27700 > Zipcode >= 27650",
                  ifelse(df$ClientZipCode >= 27600, "27650 > Zipcode >= 27600",
                     ifelse(df$ClientZipCode >= 27550, "27600 > Zipcode >= 27550",
                            ifelse(df$ClientZipCode >= 27500, "27550 > Zipcode >= 27500",
                                   ifelse(df$ClientZipCode >= 27450, "27550 > Zipcode >= 27450","Zipcod
       ))))))

df$ZipCode_Range <- v




TotalHHnumbers <- as.numeric(df[,5])

Z <- cbind(df,TotalHHnumbers)

final <- Z[,c(-1:-3,-5,-27,-28)]

# center and scale the matrix
scaled <- scale(final)

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



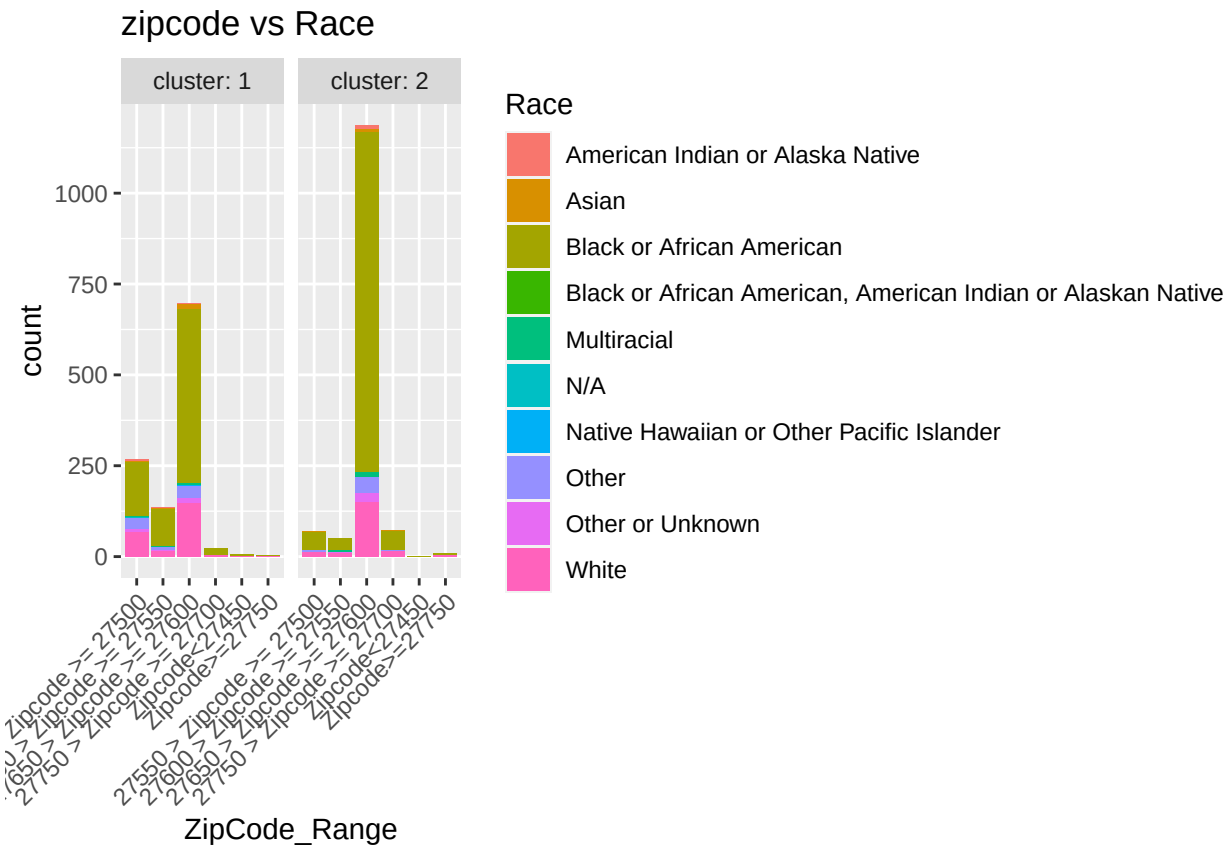

```
# explore the zipcode based on the cluster.


df %>%
  as_tibble() %>%
  mutate(cluster = try2$cluster) %>%
  ggplot(aes(x = ZipCode_Range)) +
    geom_bar(aes(fill = as.factor(cluster)),position = "dodge") +
    labs(x = "ZipCode", y = "Count", title = "Zip Code by cluster") +
    scale_fill_discrete(name = "cluster") +
    theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```




```
df %>%
  mutate(cluster = try2$cluster) %>%
  ggplot(aes(ZipCode_Range, fill = as.factor(Circumstance))) + geom_bar(position = "stack")+
    facet_grid(cols = vars(cluster),labeller = label_both)+
  scale_fill_discrete(name = "Circumstance") +
  labs (title = "zipcode vs Circumstance")+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1))
```




```
df %>%
  mutate(cluster = try2$cluster) %>%
  ggplot(aes(ZipCode_Range, 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))
```




```
cluster_level2 <- try2$cluster
cluster_data2 <- cbind(df,cluster_level2)

hh2 <- cluster_data2 %>% group_by(ZipCode_Range,Agency_Clean_Short) %>%
  summarize(percent = 100*n()/nrow(cluster_data2))
```

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

```
WW <- hh2 %>% filter(percent >= 0.5)
knitr::kable(WW)
```

ZipCode_Range	Agency_Clean_Short	percent
27550 > Zipcode >= 27500	CCWJC	0.5952381
27550 > Zipcode >= 27500	Families Together	0.6746032
27550 > Zipcode >= 27500	InterAct	0.5555556
27550 > Zipcode >= 27500	WCHS-Maternal Child Health	0.7936508
27550 > Zipcode >= 27500	WCHS-Middle Class Express	0.5158730
27550 > Zipcode >= 27500	WCHS-Wake Prevent!	0.5952381
27550 > Zipcode >= 27500	WCPSS	3.5317460
27600 > Zipcode >= 27550	WCHS-Maternal Child Health	0.7142857
27600 > Zipcode >= 27550	WCHS-Middle Class Express	0.7142857
27600 > Zipcode >= 27550	WCPSS	2.6984127
27650 > Zipcode >= 27600	Alliance Health	2.1428571

ZipCode_Range	Agency_Clean_Short	percent
27650 > Zipcode >= 27600	CASA	0.9920635
27650 > Zipcode >= 27600	Catholic Charities	2.2619048
27650 > Zipcode >= 27600	CCWJC	2.0634921
27650 > Zipcode >= 27600	Community Partnerships, Inc.	0.5158730
27650 > Zipcode >= 27600	Families Together	5.9523810
27650 > Zipcode >= 27600	Family Promise	1.3888889
27650 > Zipcode >= 27600	Haven House	1.6666667
27650 > Zipcode >= 27600	InterAct	1.8650794
27650 > Zipcode >= 27600	Lutheran Services Carolinas	0.8333333
27650 > Zipcode >= 27600	NC Recovery Support Services	0.7539683
27650 > Zipcode >= 27600	Passage Home	5.8333333
27650 > Zipcode >= 27600	Salvation Army	2.6984127
27650 > Zipcode >= 27600	StepUp Ministry	1.3095238
27650 > Zipcode >= 27600	Telamon North Carolina	0.8730159
27650 > Zipcode >= 27600	Triangle Family Services	5.2777778
27650 > Zipcode >= 27600	USCRI	1.2698413
27650 > Zipcode >= 27600	Wake County Human Services	1.9047619
27650 > Zipcode >= 27600	Wake FS&CPS	1.6269841
27650 > Zipcode >= 27600	Wake Supportive Housing	1.0317460
27650 > Zipcode >= 27600	WCHS-Maternal Child Health	4.0873016
27650 > Zipcode >= 27600	WCHS-Middle Class Express	2.8571429
27650 > Zipcode >= 27600	WCHS-Wake Prevent!	1.5079365
27650 > Zipcode >= 27600	WCPSS	16.4285714
27750 > Zipcode >= 27700	Alliance Health	1.3492063
27750 > Zipcode >= 27700	CASA	0.7936508

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.