Untitled

2024-04-10

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

```
library(readr)
read.csv('/Users/hebeyuan/Desktop/bc/7900 spring/project.data/project.sales.cities.csv')
read.csv('/Users/hebeyuan/Desktop/bc/7900 spring/project.data/project.acs.cities.csv')
total <- merge(project.acs.cities,project.sales.cities)
```

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

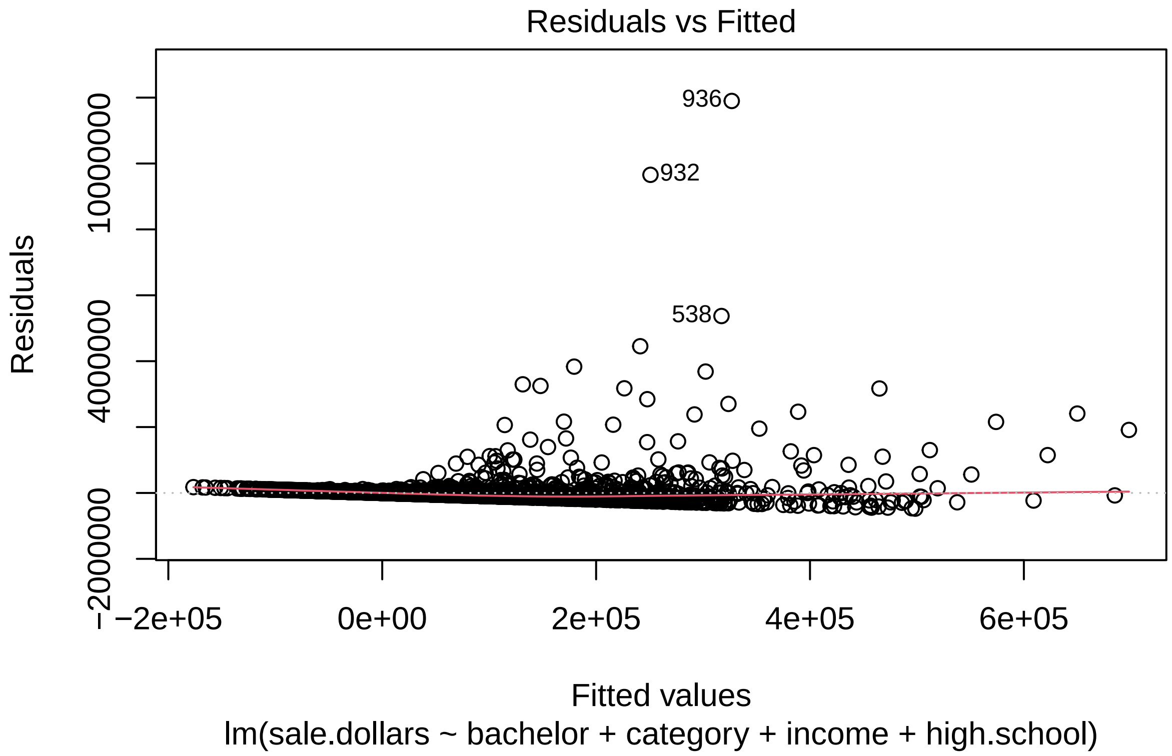
```
library(tidyverse)
library(dplyr)
library(grid)
library(ggplot2)
library(ggpubr)
```

#merge the data

```
knitr::opts_chunk$set(cache =TRUE)
total <- merge(project.acs.cities,project.sales.cities)
```

#run the regression between total sales and relating education level variables

```
liquor1 <- lm(sale.dollars ~ bachelor + category + income + high.school, data = total)
plot(liquor1, 1)
```



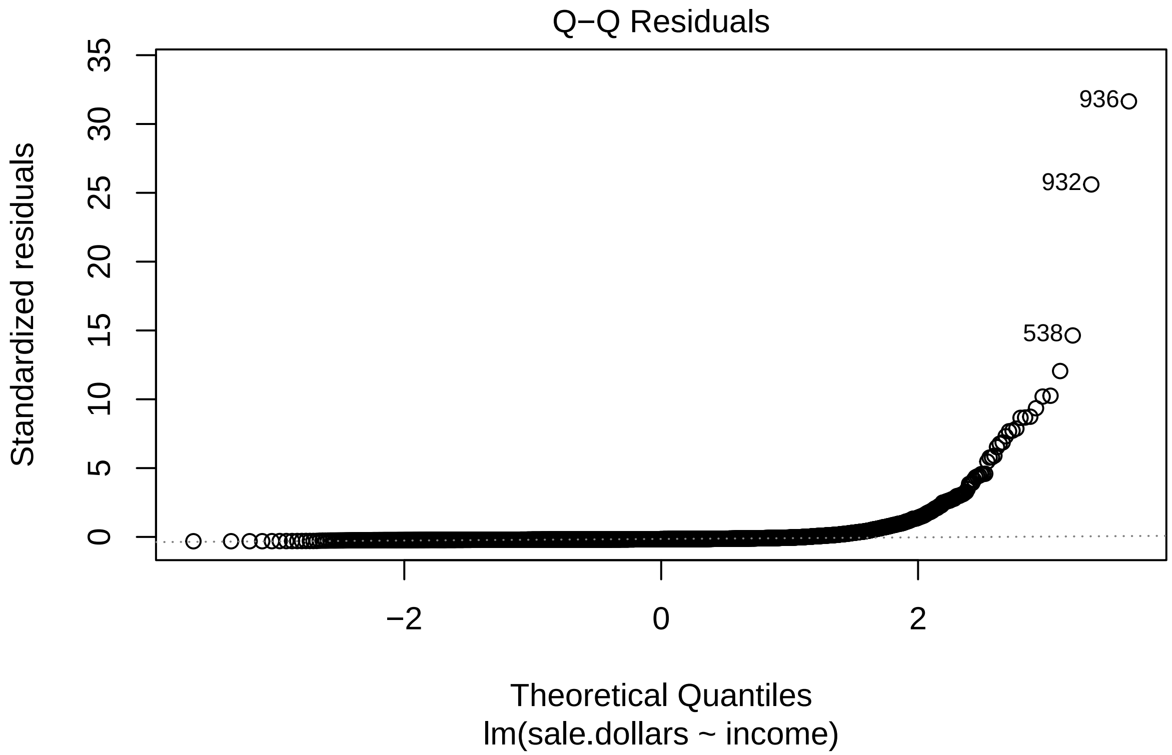
```r
summary(liquor1)
```

```
##
## Call:
## lm(formula = sale.dollars ~ bachelor + category + income + high.school,
##     data = total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##  -475850   -85447   -21739    29444 11898391
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 667687.336 128162.207   5.210 2.00e-07 ***
## bachelor                      9466.487    684.185  13.836  < 2e-16 ***
## categoryBrandy               30375.066  27638.268   1.099 0.271830
## categoryDistilled Spirits     1213.679  28881.010   0.042 0.966482
## categoryGin                  21744.040  27638.562   0.787 0.431492
## categoryMisc                 63795.320  27587.959   2.312 0.020809 *
## categoryRum                  95289.204  27571.445   3.456 0.000554 ***
## categorySchnapps             33137.890  27621.576   1.200 0.230329
## categoryTequila              47236.560  27604.465   1.711 0.087130 .
## categoryVodka               166659.121  27571.445   6.045 1.65e-09 ***
## categoryWhisky              242655.322  27554.629   8.806  < 2e-16 ***
## income                          -5.267      1.083  -4.862 1.21e-06 ***
## high.school                 -7689.087   1513.021  -5.082 3.92e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 368000 on 3680 degrees of freedom
## Multiple R-squared:  0.0854, Adjusted R-squared:  0.08241
## F-statistic: 28.63 on 12 and 3680 DF,  p-value: < 2.2e-16
```

```
liquor2 <- lm(sale.dollars ~ income, data = total)
summary(liquor2)
```

```
##
## Call:
## lm(formula = sale.dollars ~ income, data = total)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
##  -119956   -71501   -62434   -41673 12150163
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 133436.306  31662.513   4.214 2.57e-05 ***
## income          -2.028      1.014  -2.001   0.0455 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 384000 on 3691 degrees of freedom
## Multiple R-squared:  0.001083,   Adjusted R-squared:  0.0008128
## F-statistic: 4.003 on 1 and 3691 DF,  p-value: 0.04548
```

```
plot(liquor2, 2)
```



#correlation between the liquor sales and high school degree, bachelor degree

```
cor(total$sale.dollars,total$bachelor)
```

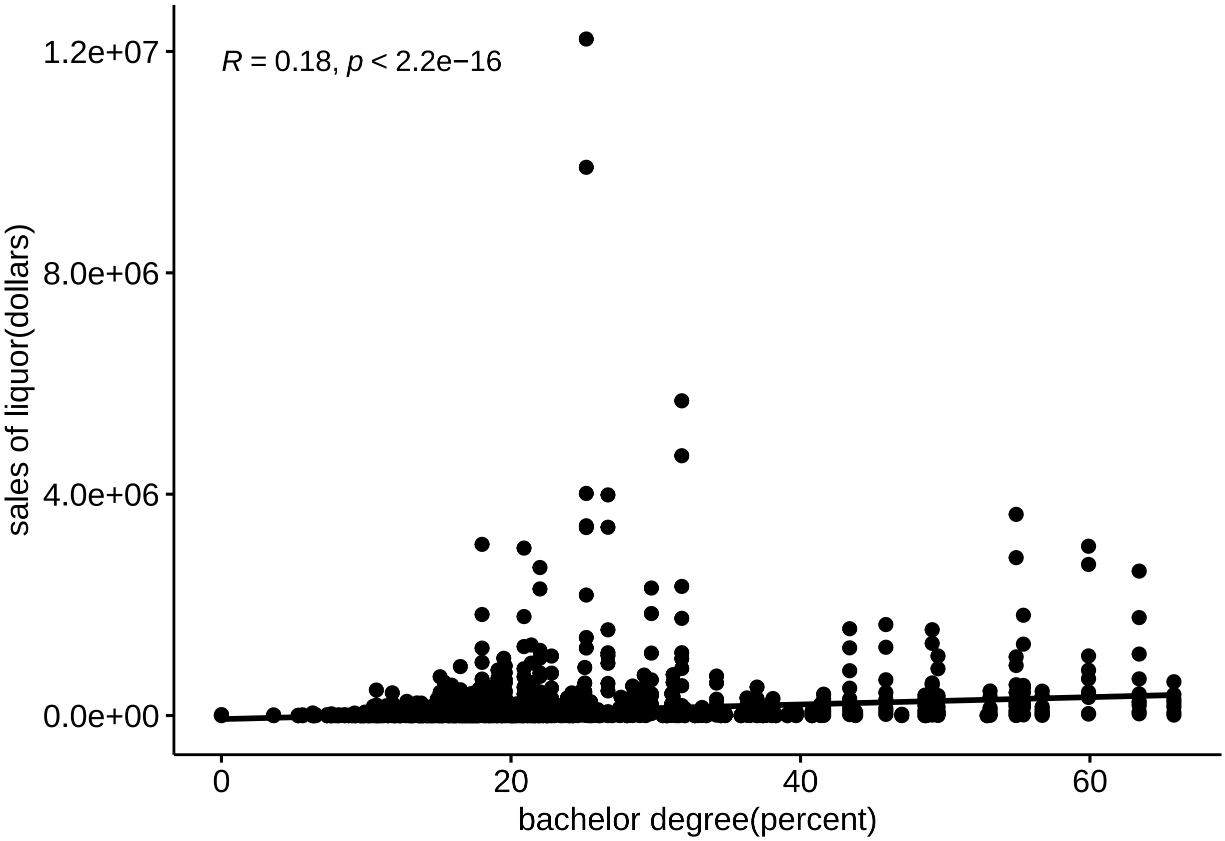
```
## [1] 0.1764605
```

```
cor(total$sale.dollars,total$unemployment)
```

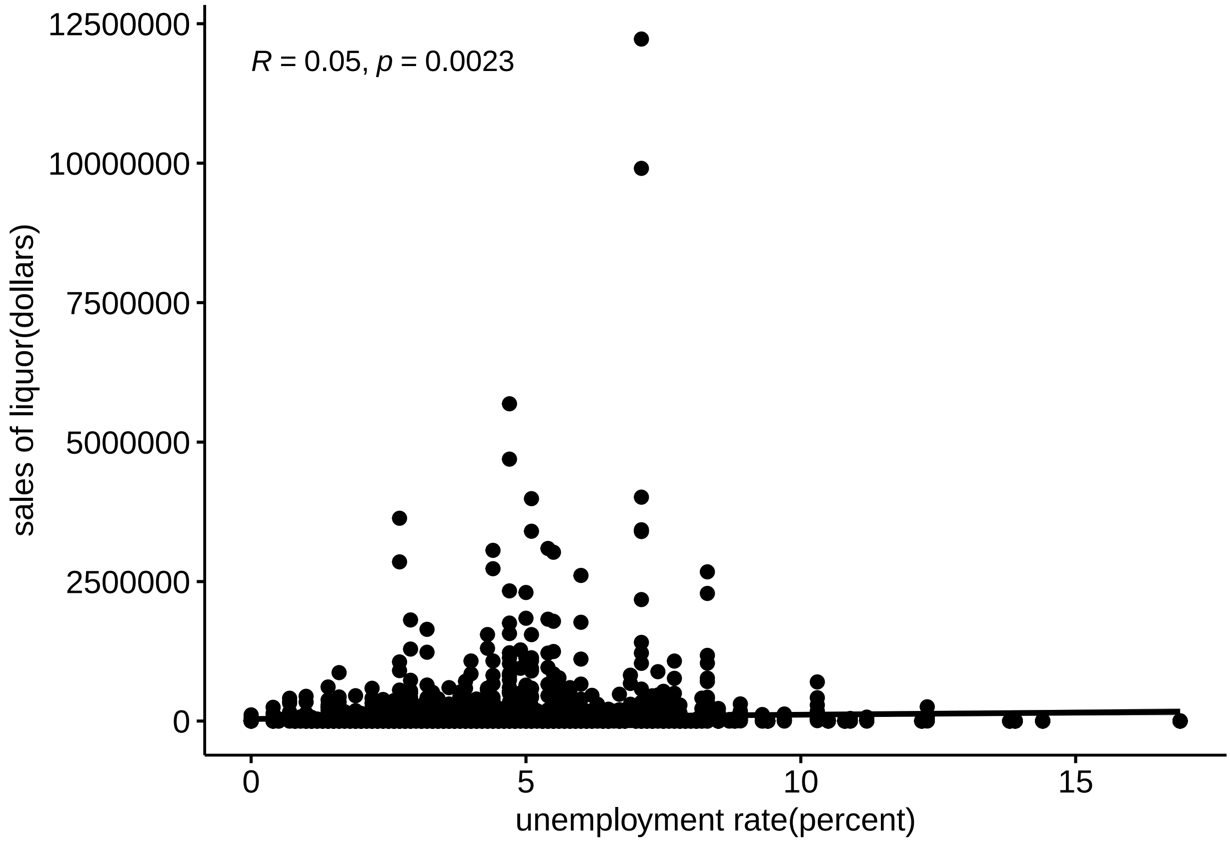
```
## [1] 0.05011339
```

#visualize by scatter plots

```
ggscatter(total, x = "bachelor", y = "sale.dollars",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "bachelor degree(percent)", ylab = "sales of liquor(dollars)")
```



```
ggscatter(total, x = "unemployment", y = "sale.dollars",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "unemployment rate(percent)", ylab = "sales of liquor(dollars)")
```



k means clustering: #standardize and remove unnumerical variables

```
data1<- total %>% select(- city, - category) %>% scale()
```

#create clusters with k-means

```
kmeans(data1, centers = 4, iter.max = 100, nstart = 100)
```

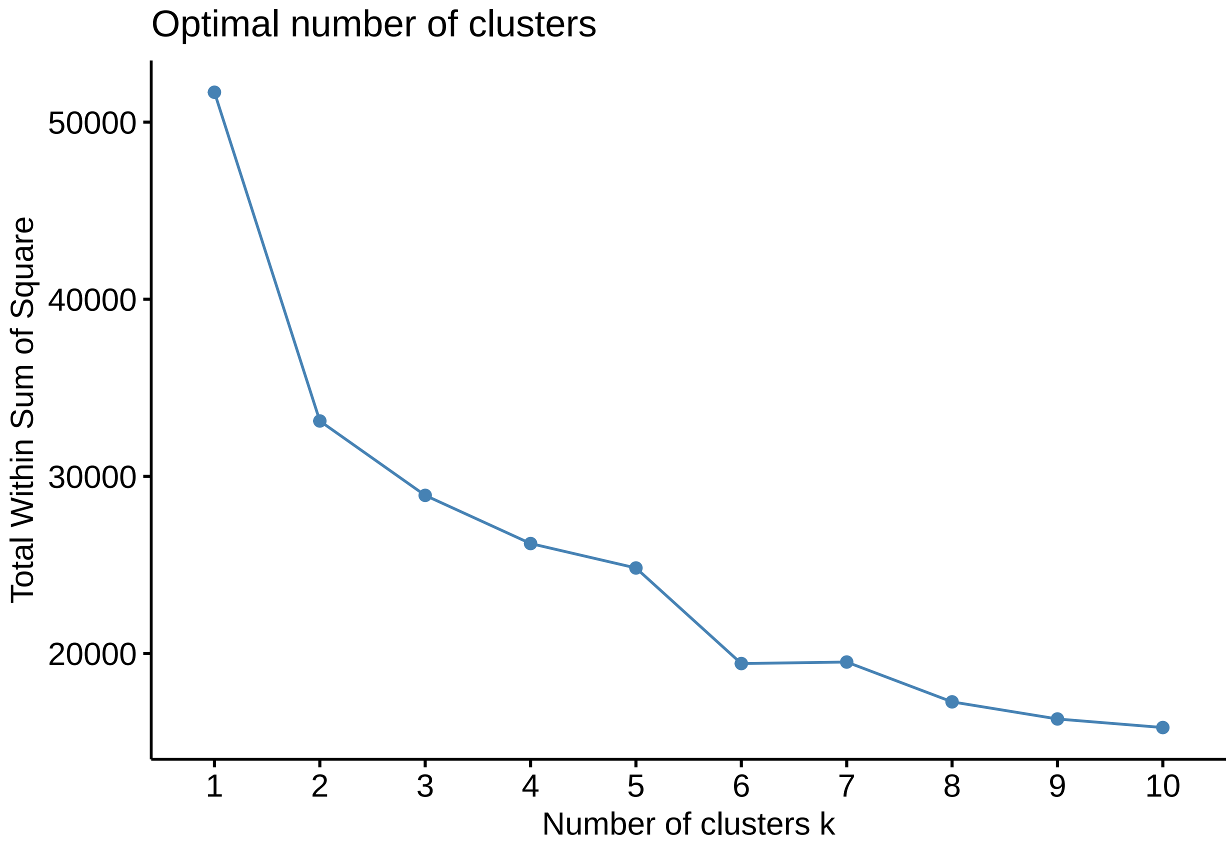
```
## K-means clustering with 4 clusters of sizes 924, 10, 87, 2672
##
## Cluster means:
##   high.school   bachelor unemployment     income  population    pop.white   pop.black pop.indian   pc
## 1   0.8520051  1.0119956   -0.6125104  1.0392163 -0.01251435  0.004327699 -0.08609643 -0.1119733 -0.0
## 2  -0.8671015  0.4555008    1.0515049 -0.2934683 11.94693033 11.152961925 14.70227855 10.2886446 13.7
## 3   0.0644713  1.6087224    0.3190681 -0.5101276  4.15941936  4.226122172  3.19407354  3.2775151  3.4
## 4  -0.2934846 -0.4040411    0.1974871 -0.3416617 -0.17581420 -0.180838713 -0.12924928 -0.1064996 -0.1
##
## Clustering vector:
##    [1] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
##   [88] 1 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [175] 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
##  [262] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
##  [349] 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4
##  [436] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [523] 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
##  [610] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4
##  [697] 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 4 4 4 4
##  [784] 1 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1
##  [871] 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1
##  [958] 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 4 4 4 4
##  [ reached getOption("max.print") -- omitted 2693 entries ]
##
## Within cluster sum of squares by cluster:
## [1]  4427.706  2008.036  6725.883 10141.315
##  (between_SS / total_SS =  54.9 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
```

#check the clusters between different methods

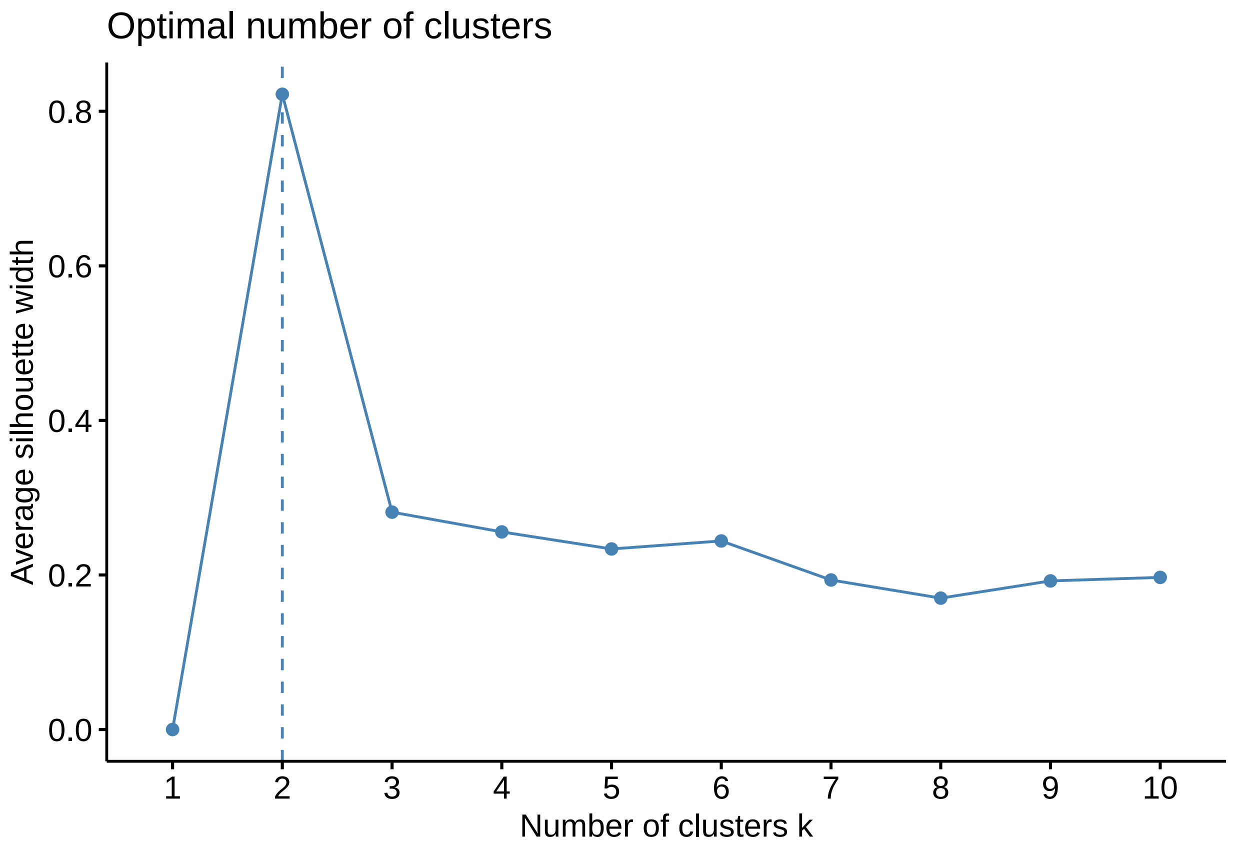
```
install.packages("factoextra")
```

```
## Error in install.packages : Updating loaded packages
```

```
library(factoextra)
fviz_nbclust(data1, kmeans, method = "wss")
```

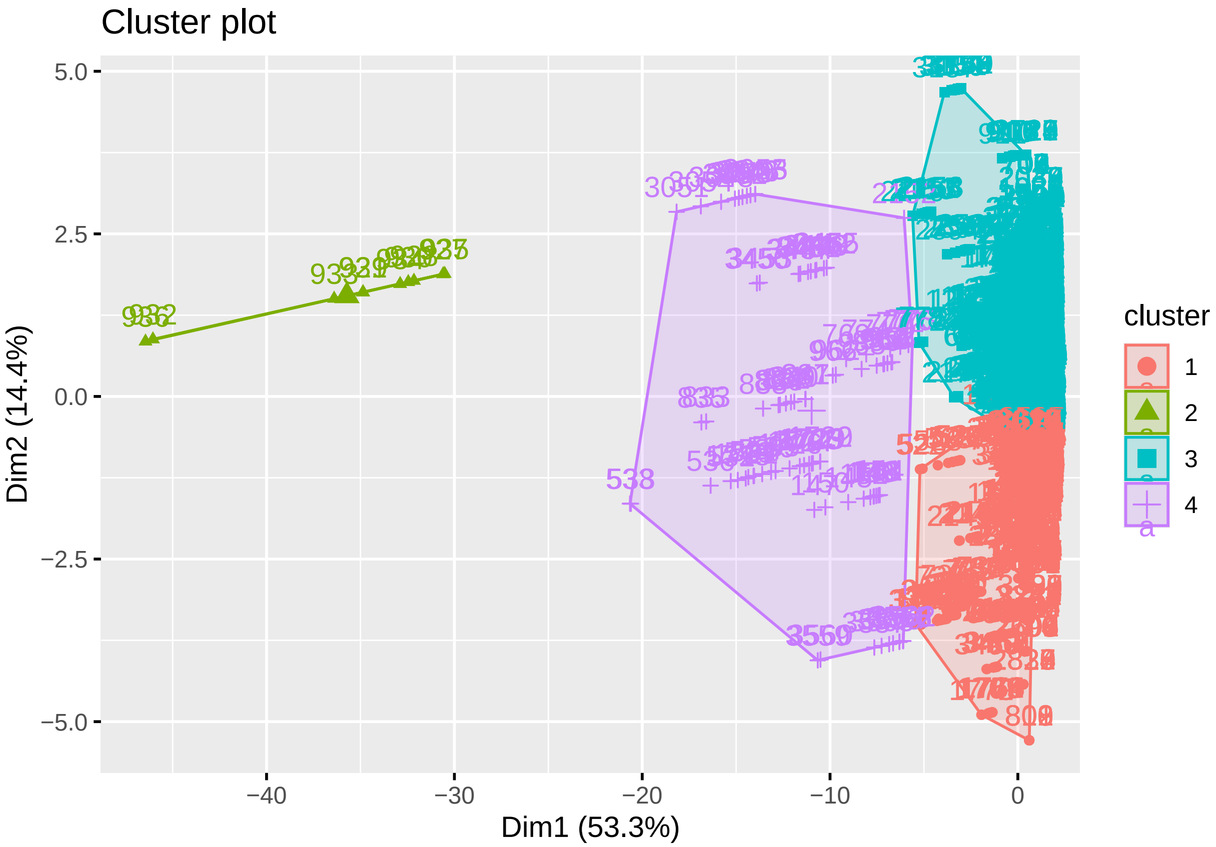


```r
fviz_nbclust(data1, kmeans, method = "silhouette")
```



#change the number of clusters

```r
fviz_cluster(kmeans(data1, centers = 4, iter.max = 100, nstart = 100)
,data = data1)
```


Cluster plot
Dim2 (14.4%)
Dim1 (53.3%)
5.0
2.5
0.0
−2.5
−5.0
−40
−30
−20
−10
0
cluster
1
2
3
4
932
933
929
927
3059
3455
833
538
530
1150
3559
962
767
2215
809
1755
2838
523