Problemset9

Matt He

2022-11-28

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.6      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(factoextra)
```

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

```
library(stringr)
```

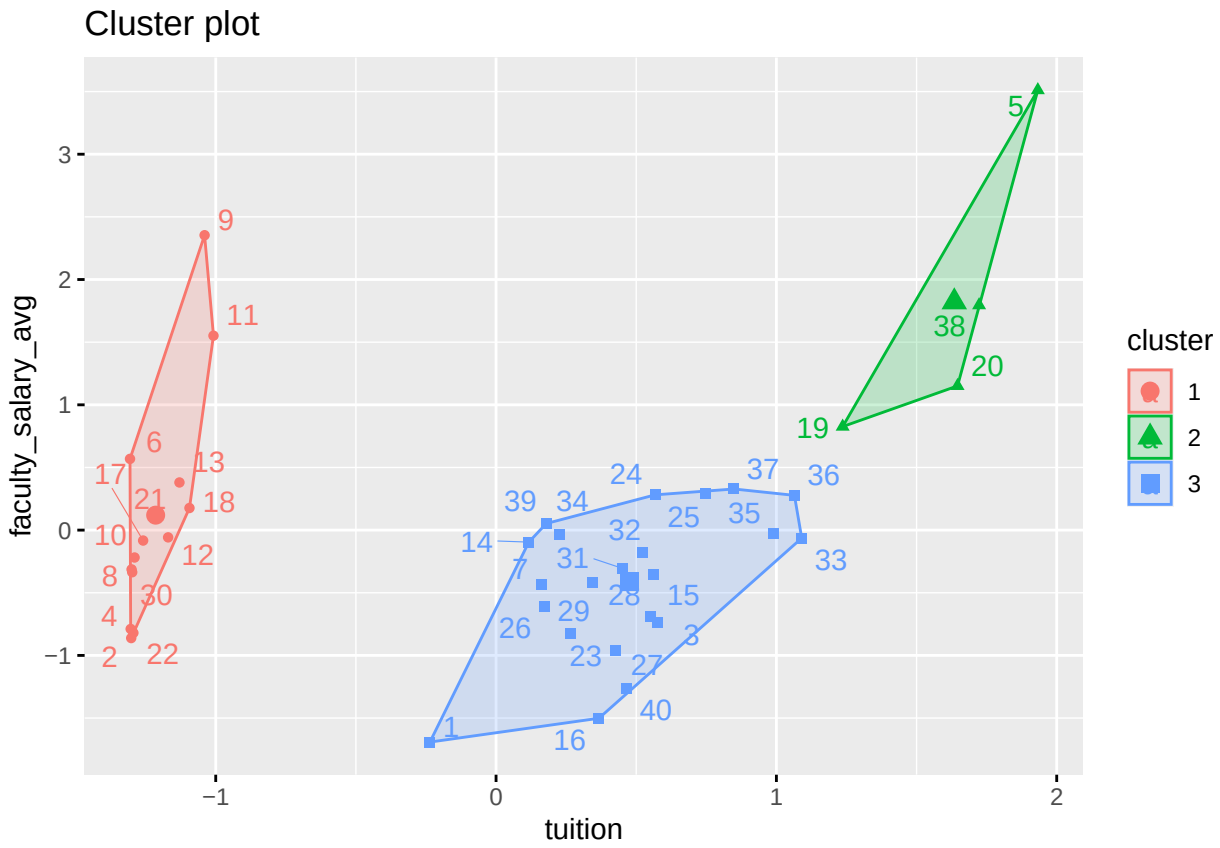
##Part 1

```
college = read_csv('~/Downloads/college.csv')
```

```
## Rows: 1270 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): name, city, state, region, highest_degree, control, gender, loan_de...
## dbl (9): id, admission_rate, sat_avg, undergrads, tuition, faculty_salary_av...
##
## 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.
```

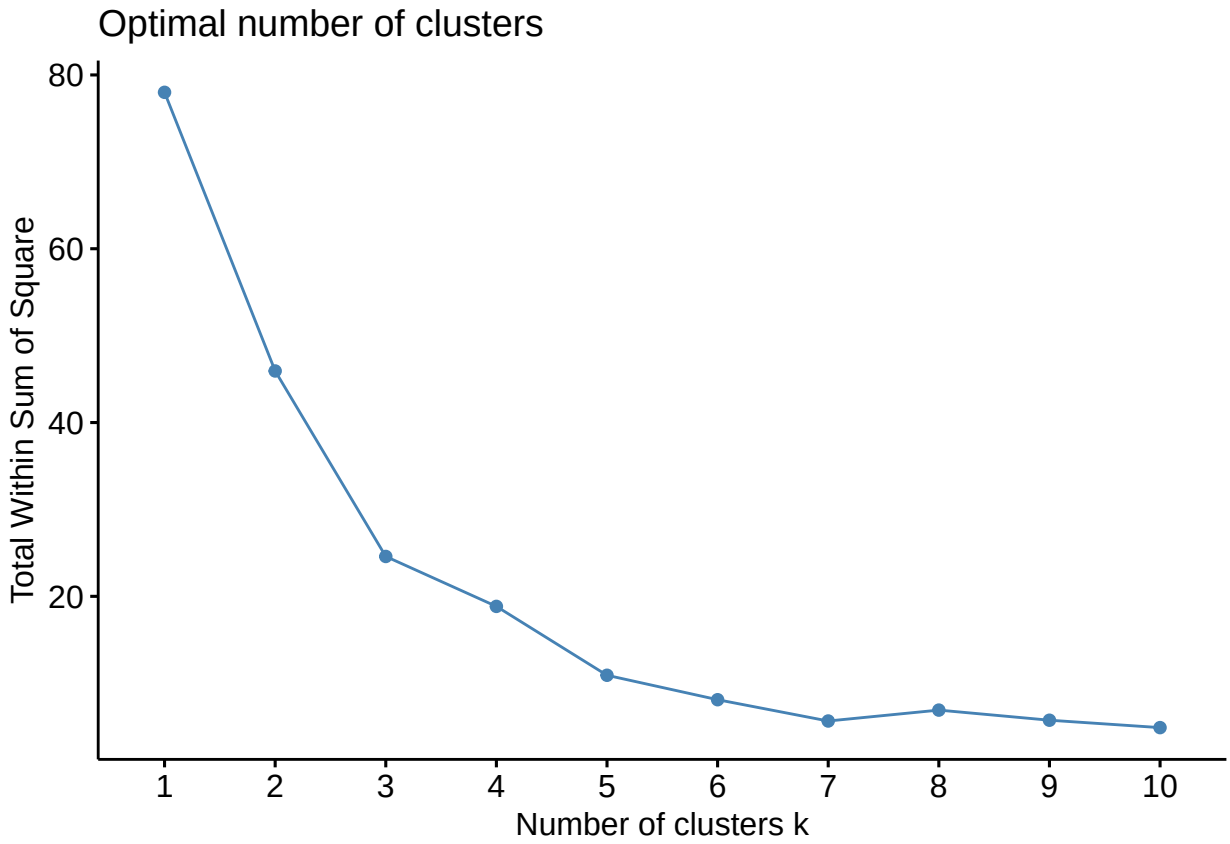
```
college_IN <- college %>%
  filter(state == 'IN')%>%
  select(tuition, faculty_salary_avg)%>%
  scale()
##filtering all people from Indiana
```

```
set.seed(1122)
k_clust = kmeans(college_IN,
                 centers = 3,
                 nstart = 30)
fviz_cluster(
  k_clust,
  data = college_IN,
  repel = TRUE
)
```



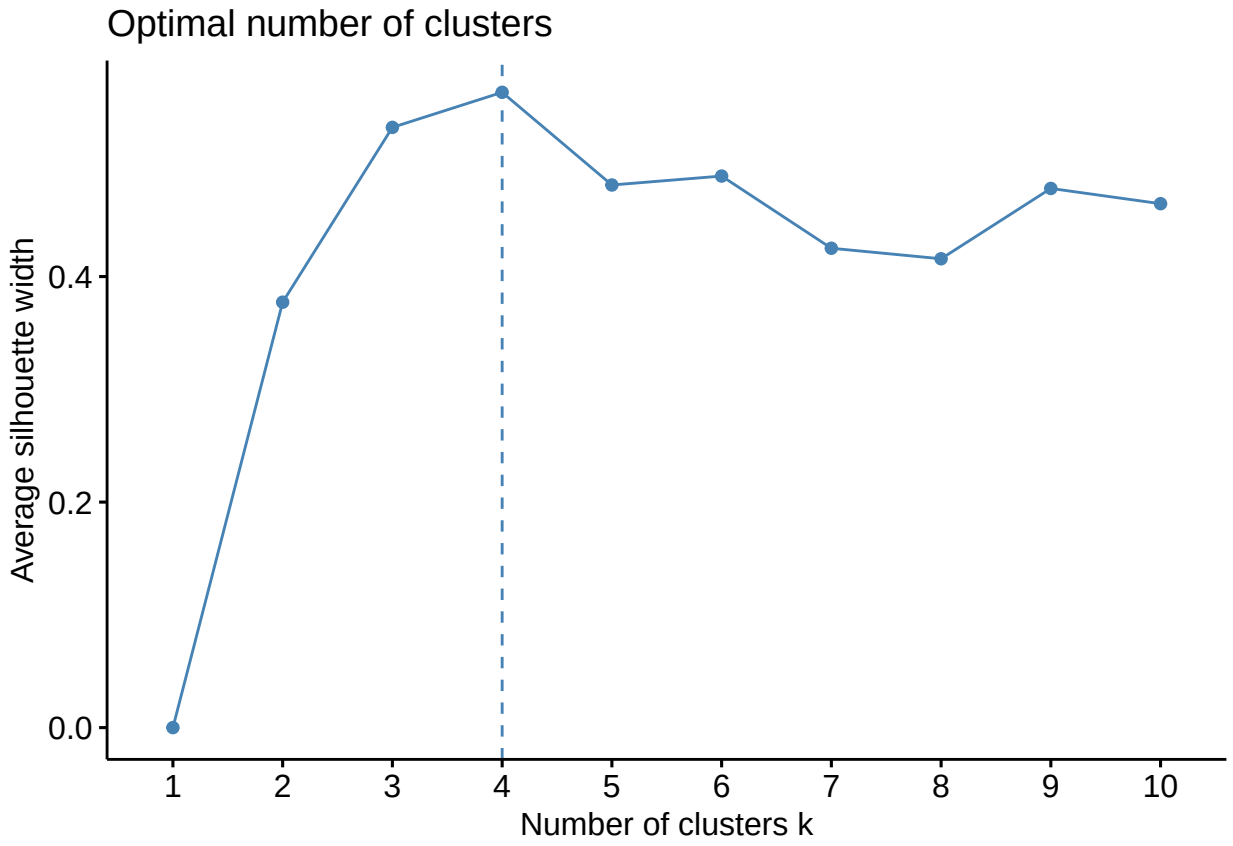
Exercise 2:

```
fviz_nbclust(college_IN, kmeans, method = "wss")
```



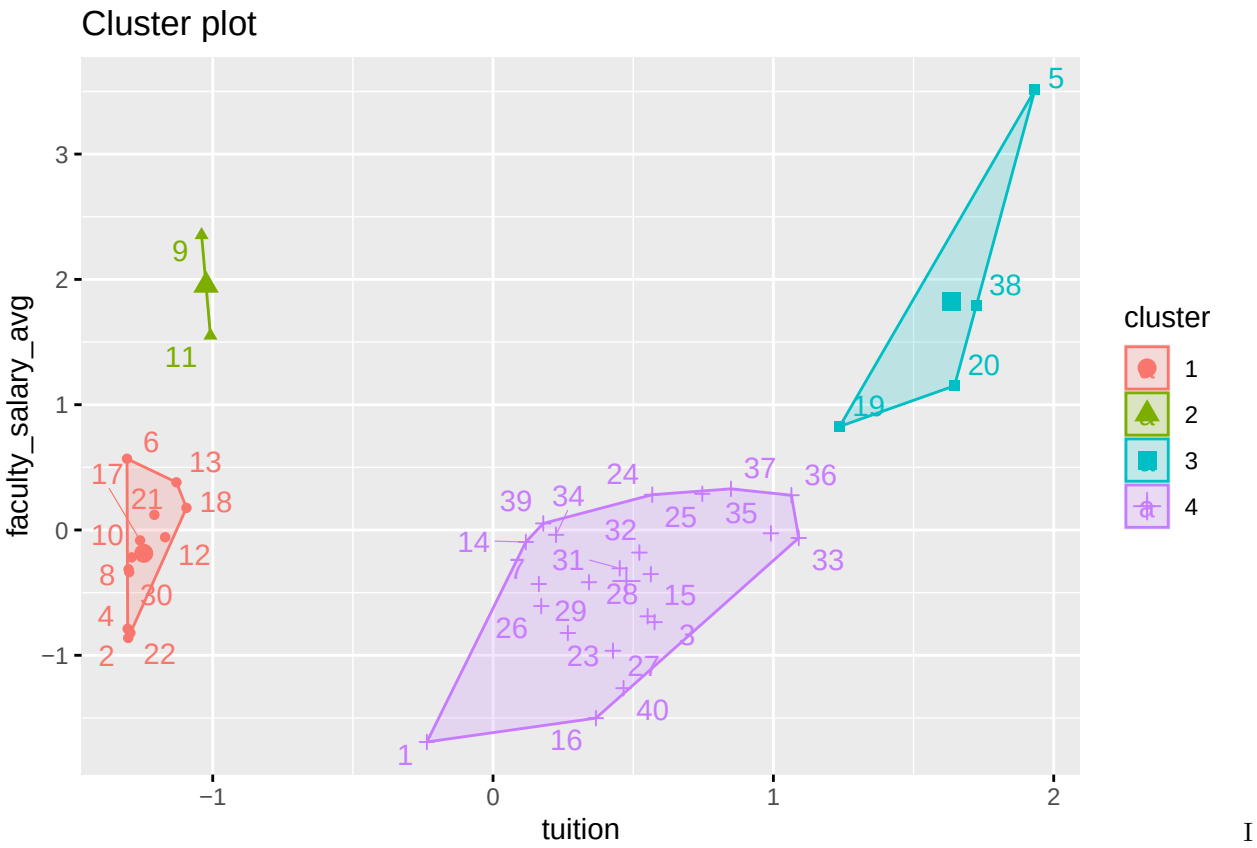
According to the elbow method, the curve is relatively flat when k=3, so we choose 3 as number of cluster.

```
fviz_nbclust(college_IN, kmeans, method = "silhouette")
```



Due to the average silhouette width reaches the highest at k=4, 4 clusters can be also considered another option.

```
set.seed(1122)
k_clust = kmeans(college_IN,
                 centers = 4,
                 nstart = 30)
fviz_cluster(
  k_clust,
  data = college_IN,
  repel = TRUE
)
```



I believe 3 clustering group is better. In the scatter plot for 4 groups, the second group only have two semple in there, I believe it is over fitting.

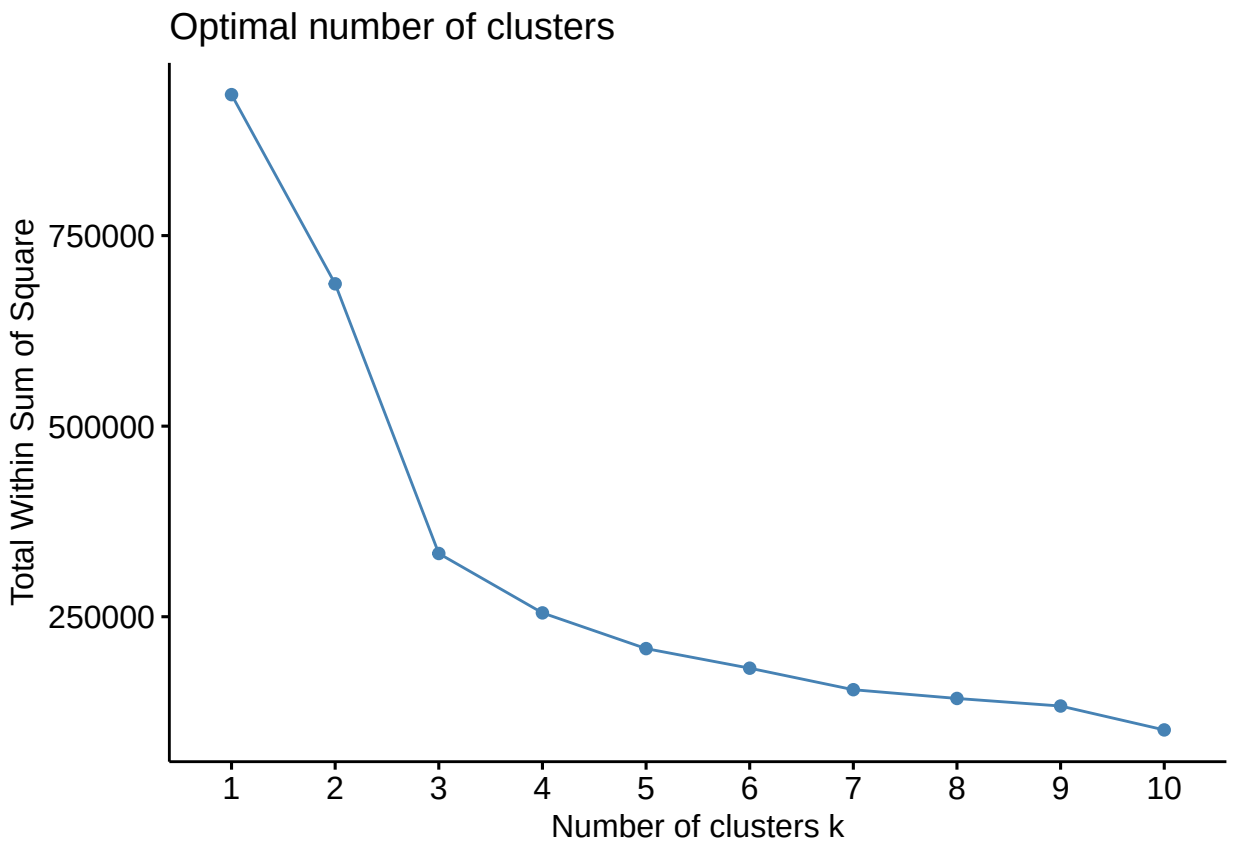
##Part 2

```
cereal <- read_csv("~/Downloads/Cereals.csv")
```

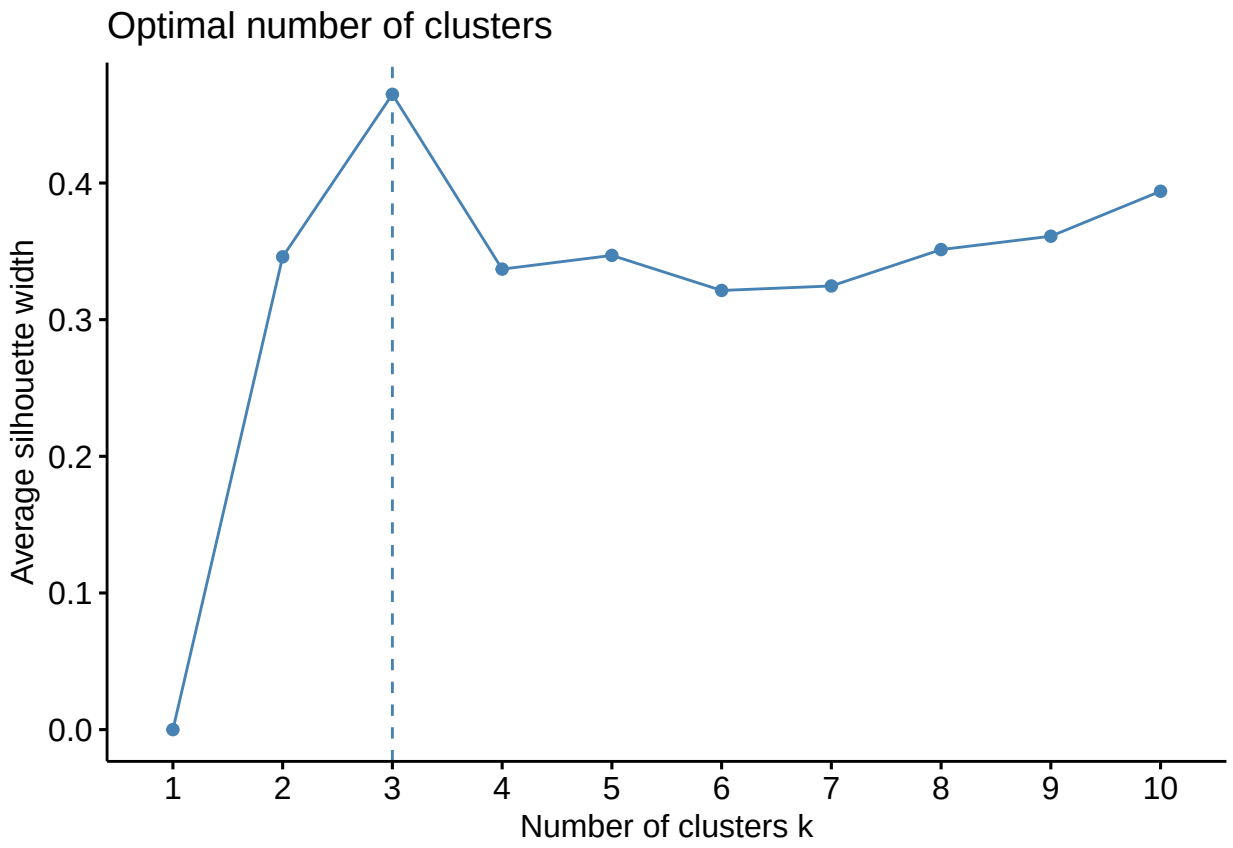
```
## Rows: 77 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): name, mfr, type
## dbl (13): calories, protein, fat, sodium, fiber, carbo, sugars, potass, vita...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
cereal <- cereal%>%
  select(-name, -mfr, -type,-weight, -shelf,-cups,-rating)%>%
  drop_na()
```

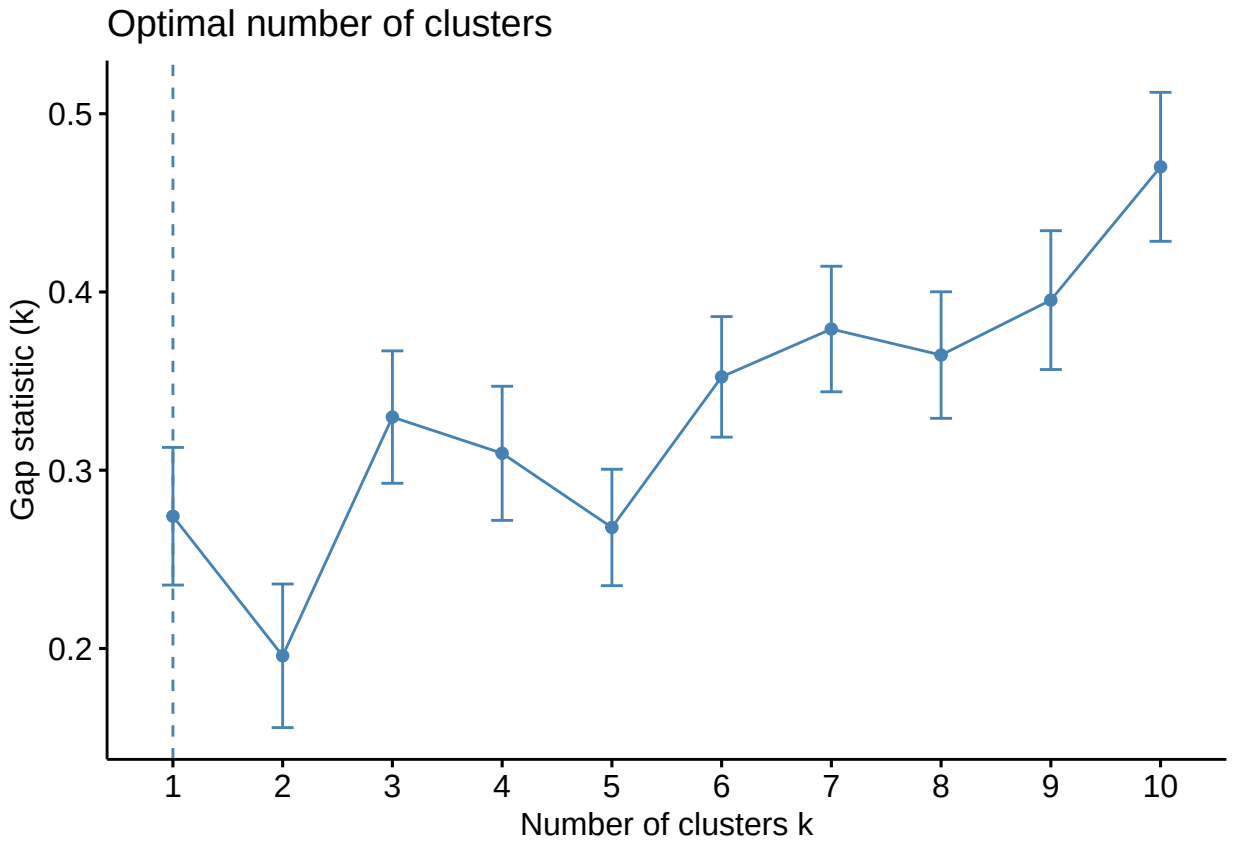
```
fviz_nbclust(cereal, kmeans, method = "wss")
```




```
fviz_nbclust(cereal, kmeans, method = "silhouette")
```



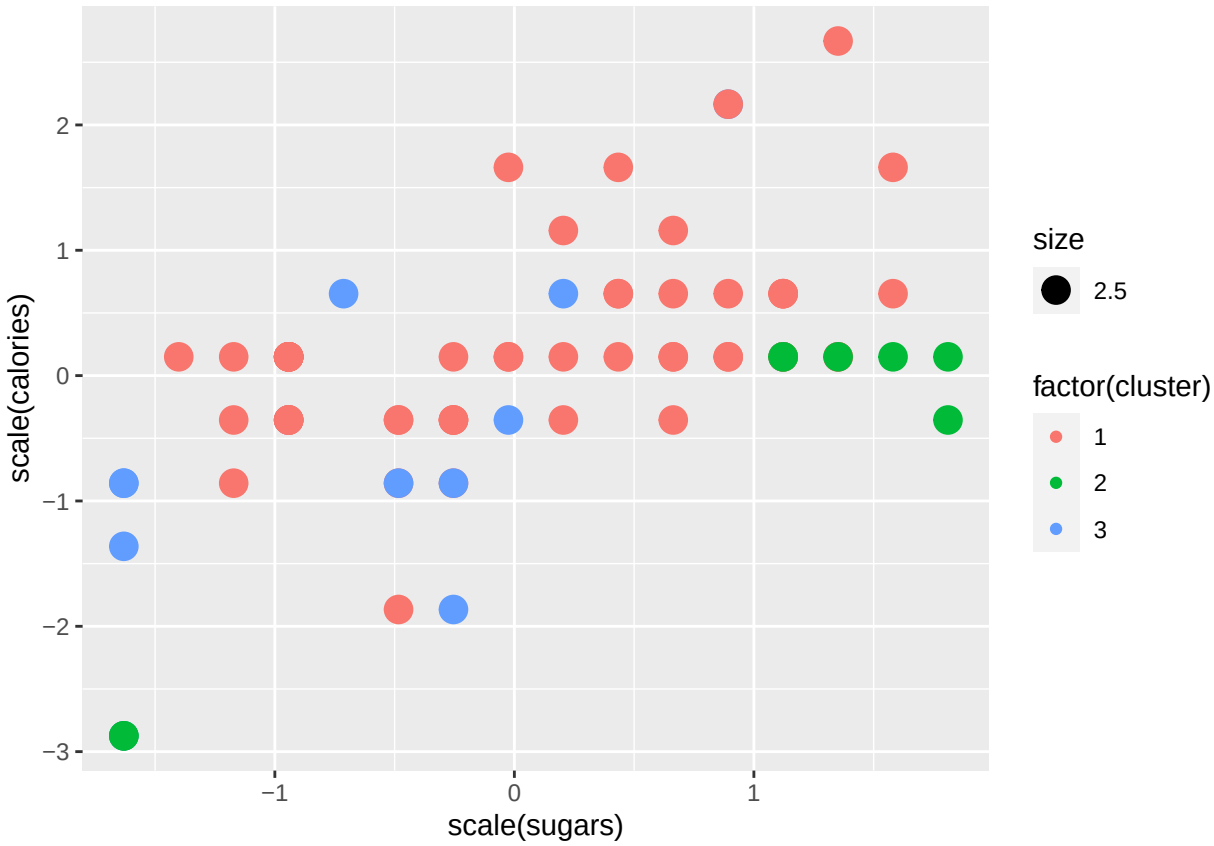

```
fviz_nbclust(cereal, kmeans, method = "gap_stat")
```



According to these three plots, we can see the the number 3 and number 1 could be possible numbers for clusters. However, if we only have 1 group, there is no point for the cluster. So we can give up 1 as number for clustering.

Exercise 5:

```
set.seed(1122)
cluster = kmeans(cereal,
                 centers = 3)
cereal$cluster <- cluster$cluster
ggplot(data = cereal,
       aes(x = scale(sugars), y = scale(calories), color=factor(cluster),size = 2.5))+
  geom_point()
```



Exercise 6: Group one is high calories-average sugar. Group two is high sugar-average calories. Group 3 is low sugars-average calories.

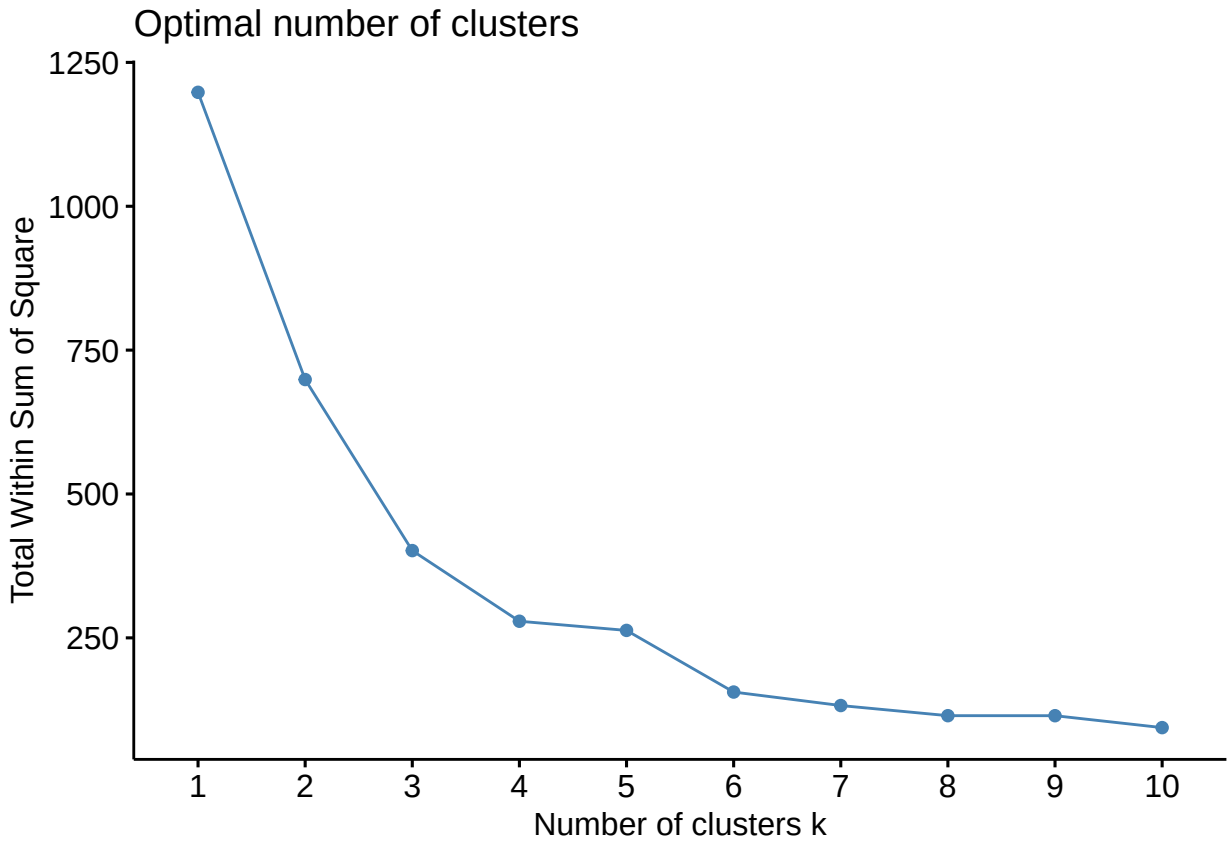
##Part 3 Exercise 7:

```
soap_og <- read_csv('~/Downloads/BathSoapHousehold.csv')
```

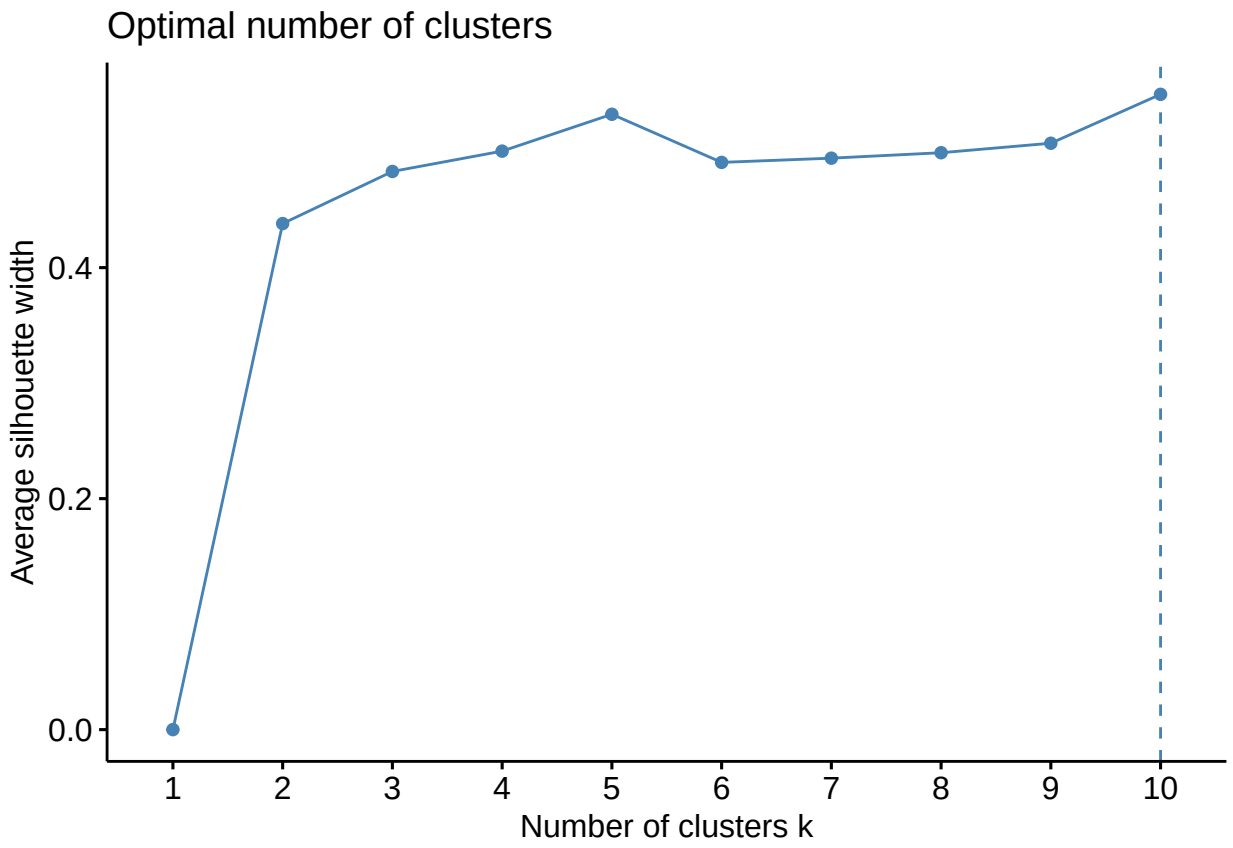
```
## Rows: 600 Columns: 46
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (46): Member id, SEC, FEH, MT, SEX, AGE, EDU, HS, CHILD, CS, Affluence I...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
soap <- soap_og%>%
  select(CHILD, `Affluence Index`)%>%
  scale()
```

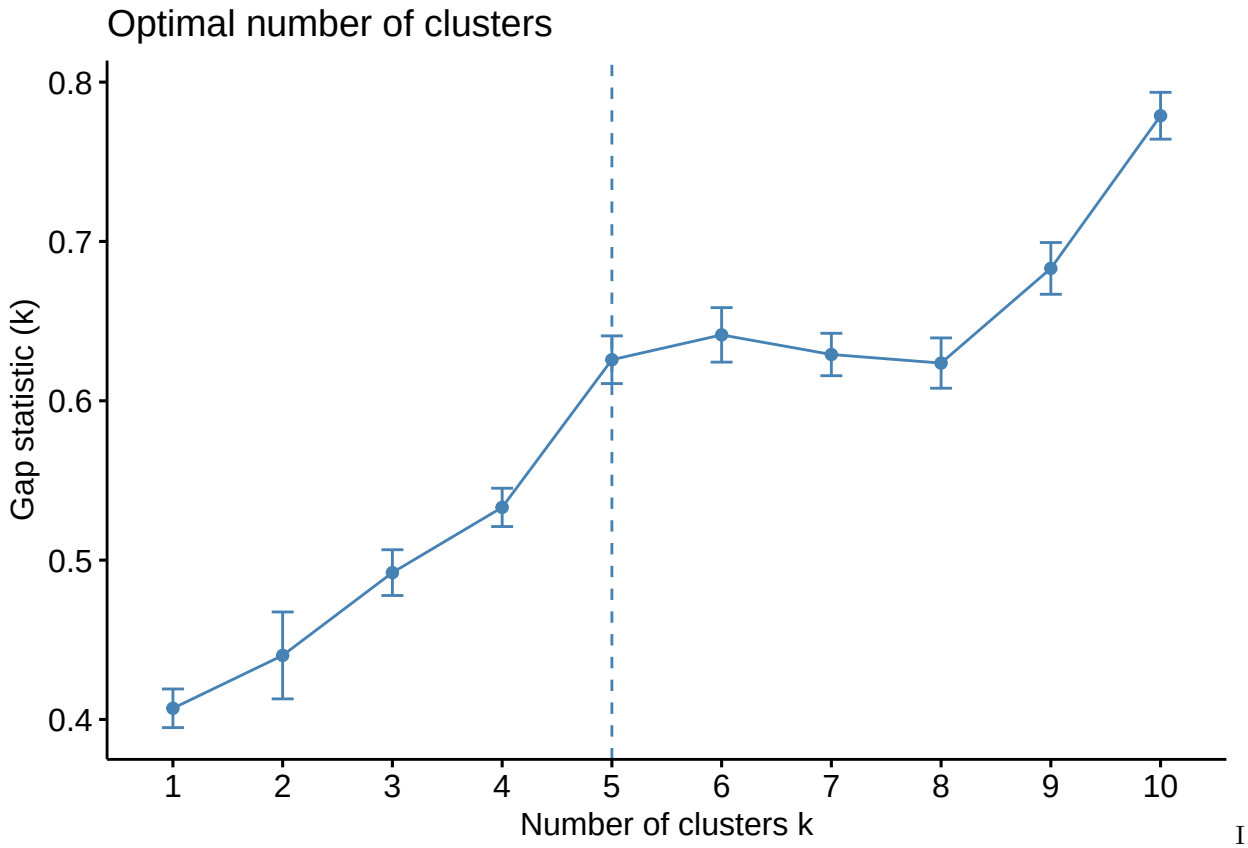
```
fviz_nbclust(soap, kmeans, method = "wss")
```




```
fviz_nbclust(soap, kmeans, method = "silhouette")
```



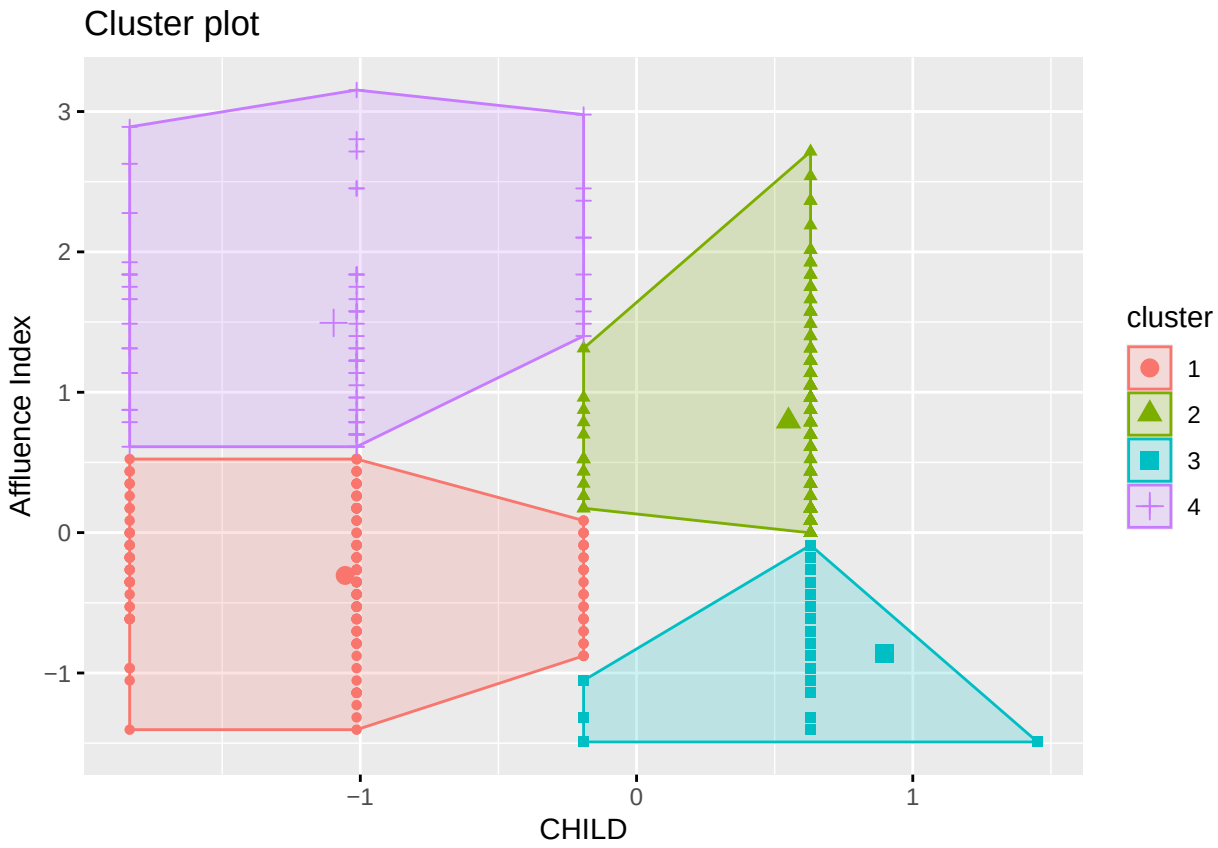

```
fviz_nbclust(soap, kmeans, method = "gap_stat")
```



I think 4 is a good number for clustering group due to elbow method. 5 could also be considered according to gap stat.

Exercise 8:

```
set.seed(123)
k_clust_soap = kmeans(soap,centers = 4)
fviz_cluster(
  k_clust_soap,
  data = soap,
  geom = c('point')
)
```



Group 1, low wealth-less number child. This kinds of household can not afford to raise too many kids, so they do not have too many children. Group 2, high-wealth-higher number child. This kinds of household are more wealthy and have more kids. Group 3, low wealth-higher number child. Group 4, high wealth- less number child.

Exercise 9:

```
soap_og%>%
  mutate(cluster = k_clust_soap$cluster)%>%
  group_by(cluster)%>%
  summarise_at(vars(Value, 'Total Volume'), funs(mean))
```

```
## # A tibble: 4 x 3
##   cluster Value `Total Volume`
##     <int> <dbl>          <dbl>
## 1       1 1416.         13512.
## 2       2 1495.         11919.
## 3       3 1035.         10261.
## 4       4 1653.         12555.
```

Number 4 cluster group has highest value. The total volume highest group is cluster number one. They are not the same group. In my understanding, group 4 has high wealth so their spending availability is better and there are more likely going to have a potential spending in the future. One the other hand, because group one is less wealthy now, they have spend more at this moment to avoid possible inflation in the future.