title: “DSC20 week10 assignment” author: “xin tang” date: “2023-08-08” output: pdf_document: default editor_options: markdown: wrap: 72

use the nearest neighbors algorithm to fit a model on two simplified datasets.

Plot the data from each dataset using a scatter plot.

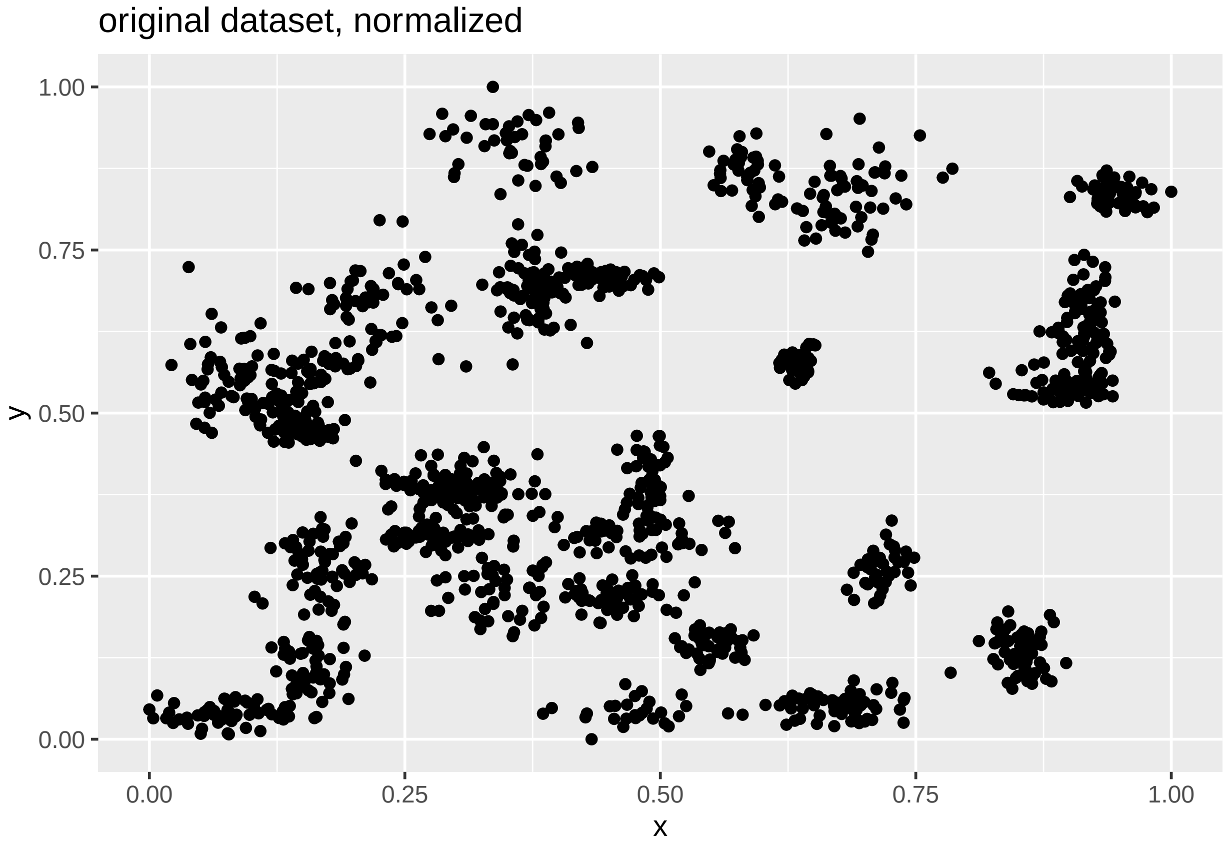
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##
## Attaching package: 'scales'

## The following object is masked from 'package:readr':
##
##     col_factor
```


original dataset, normalized

from data, the dataset is too scattered to use a linear regression

now start to fit a nearest neighbor model using K =3, 5,15, 20 and 25.

```
## [1] "0.711 is the accuracy when k =3"
```

```
## [1] "0.708 is the accuracy when k =5"
```

```
## [1] "0.691 is the accuracy when k = 10"
```

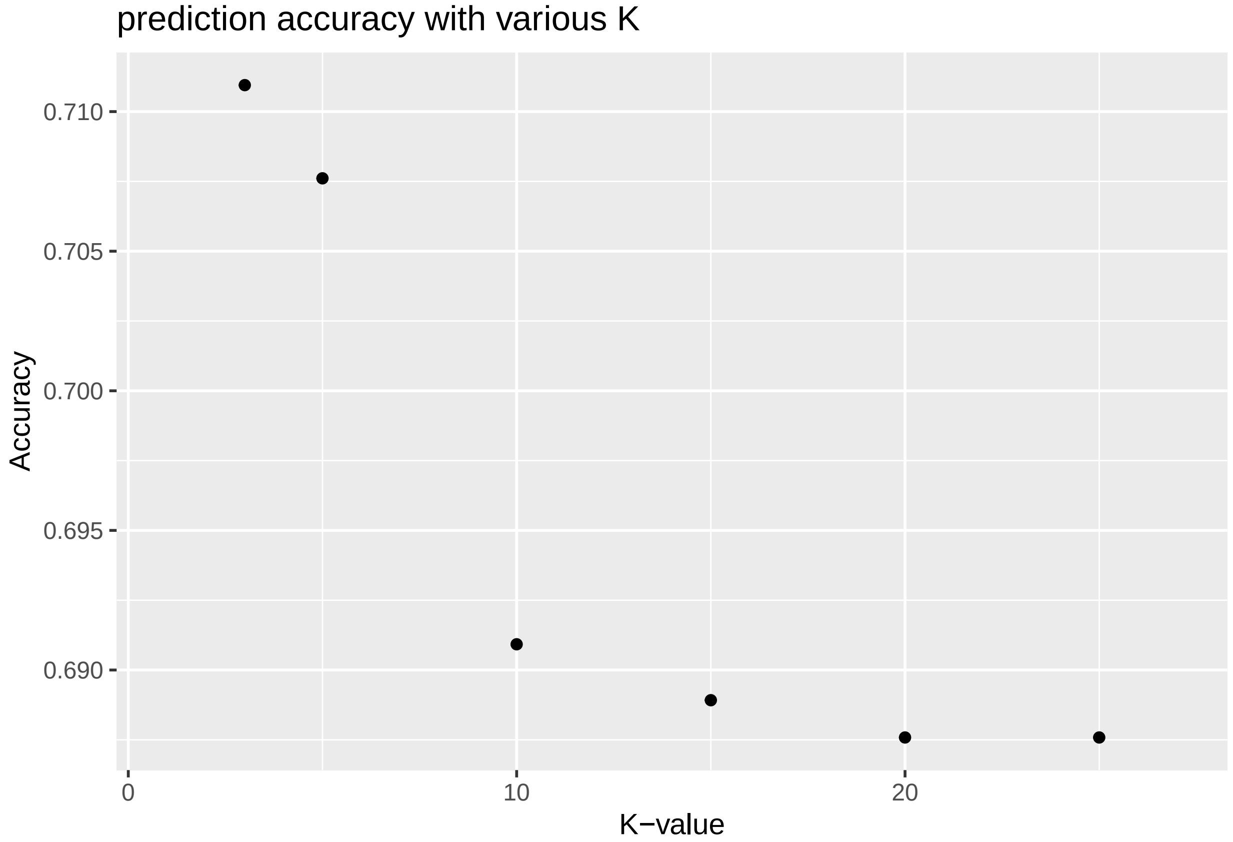
```
## [1] "0.689 is the accuracy when k = 15"
```

```
## [1] "0.688 is the accuracy when k = 20"
```

```
## [1] "0.688 is the accuracy when k = 25"
```

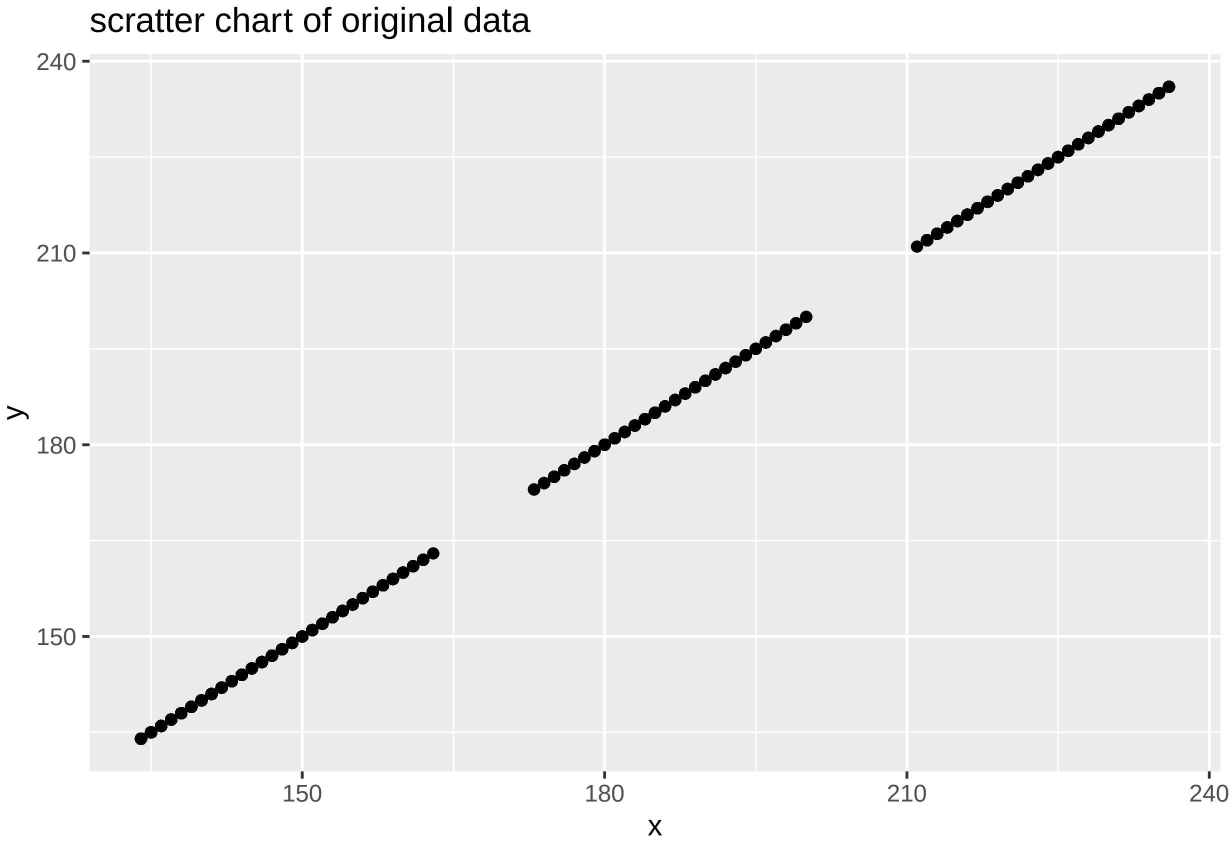
From the output, the accuracy greatly improved (last week is only 58%), which prove this is a better model.

Plot a comparision chart for different K value

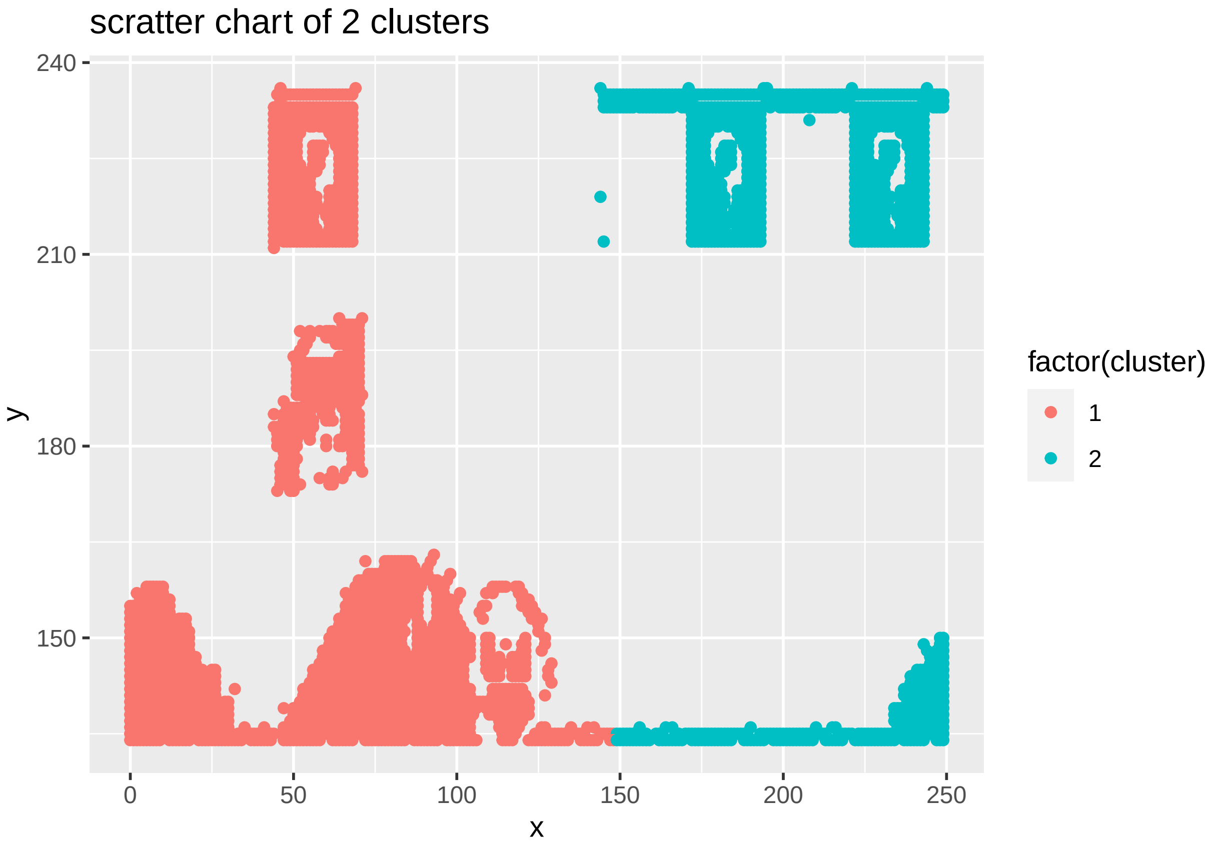


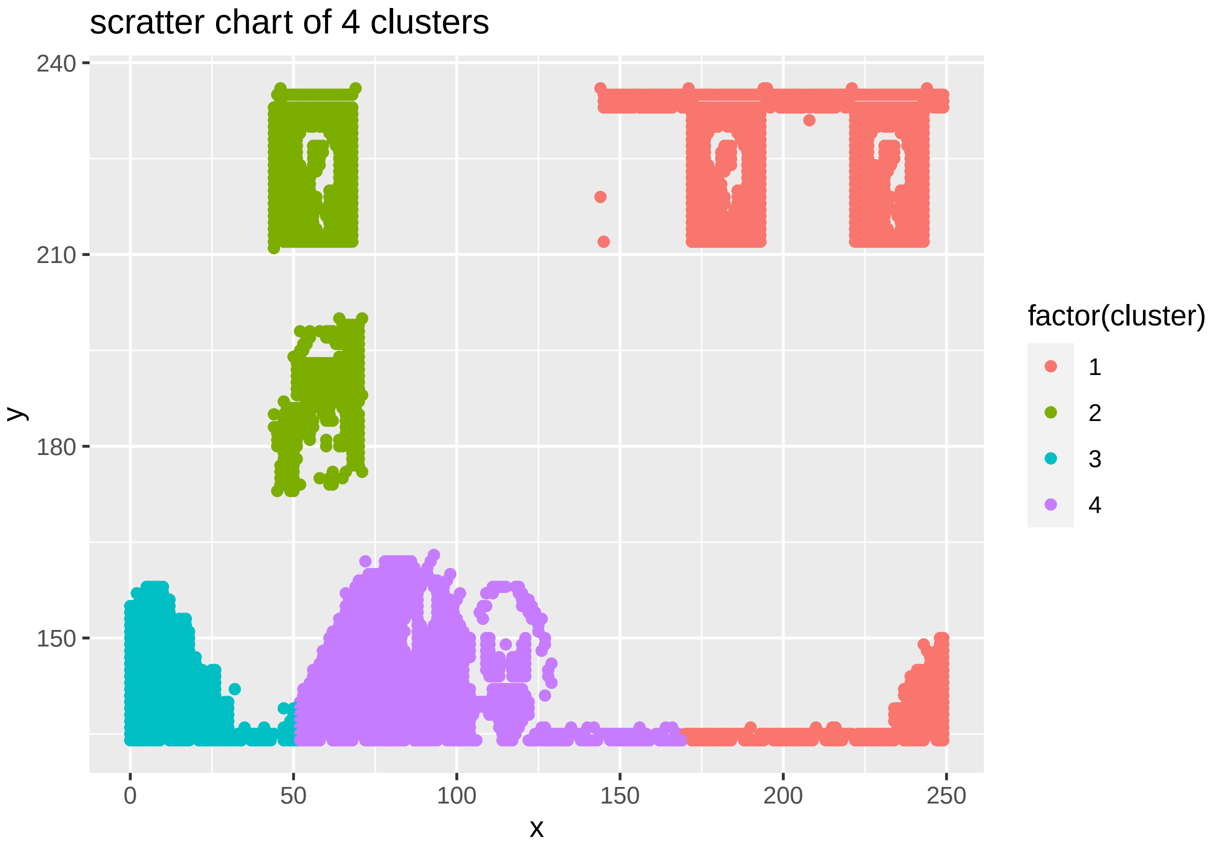
Begin of the cluster data analysis using K means clustering

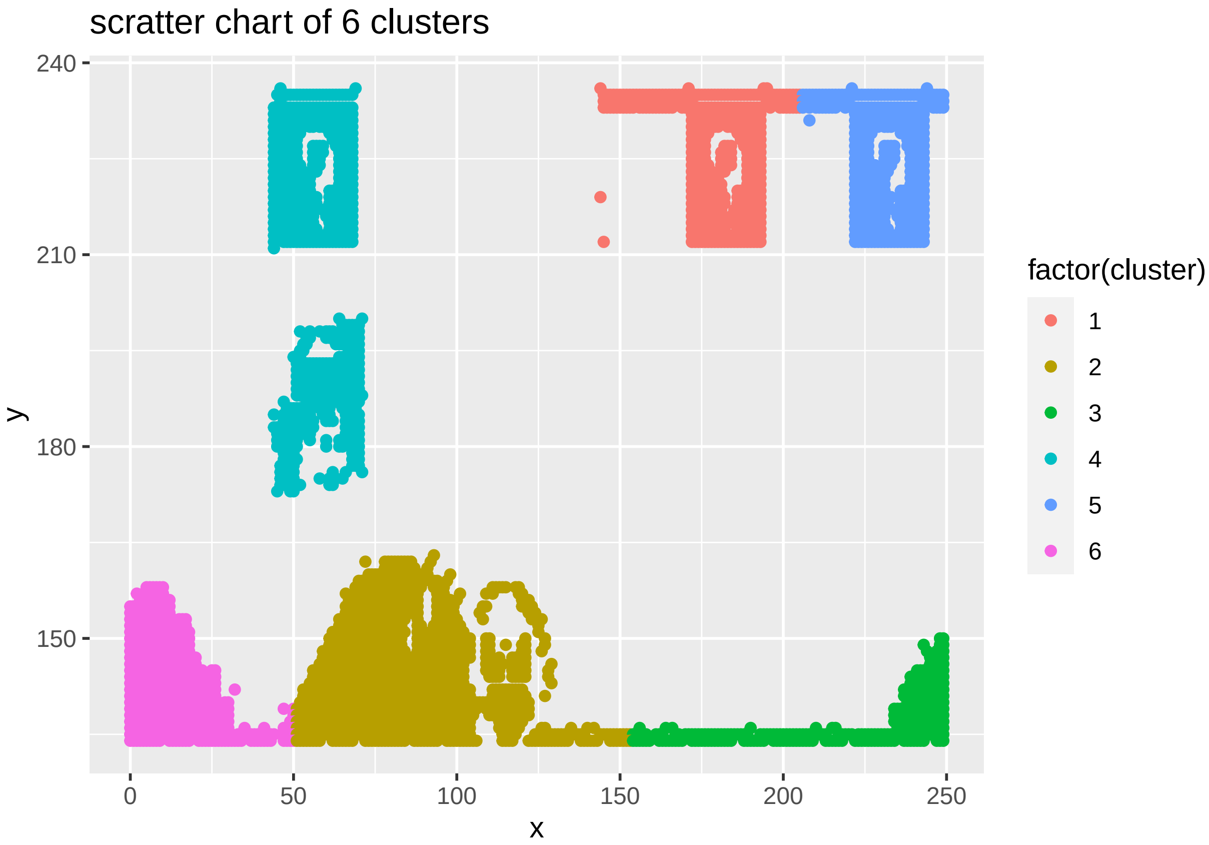
##plot the original data in scatter plot

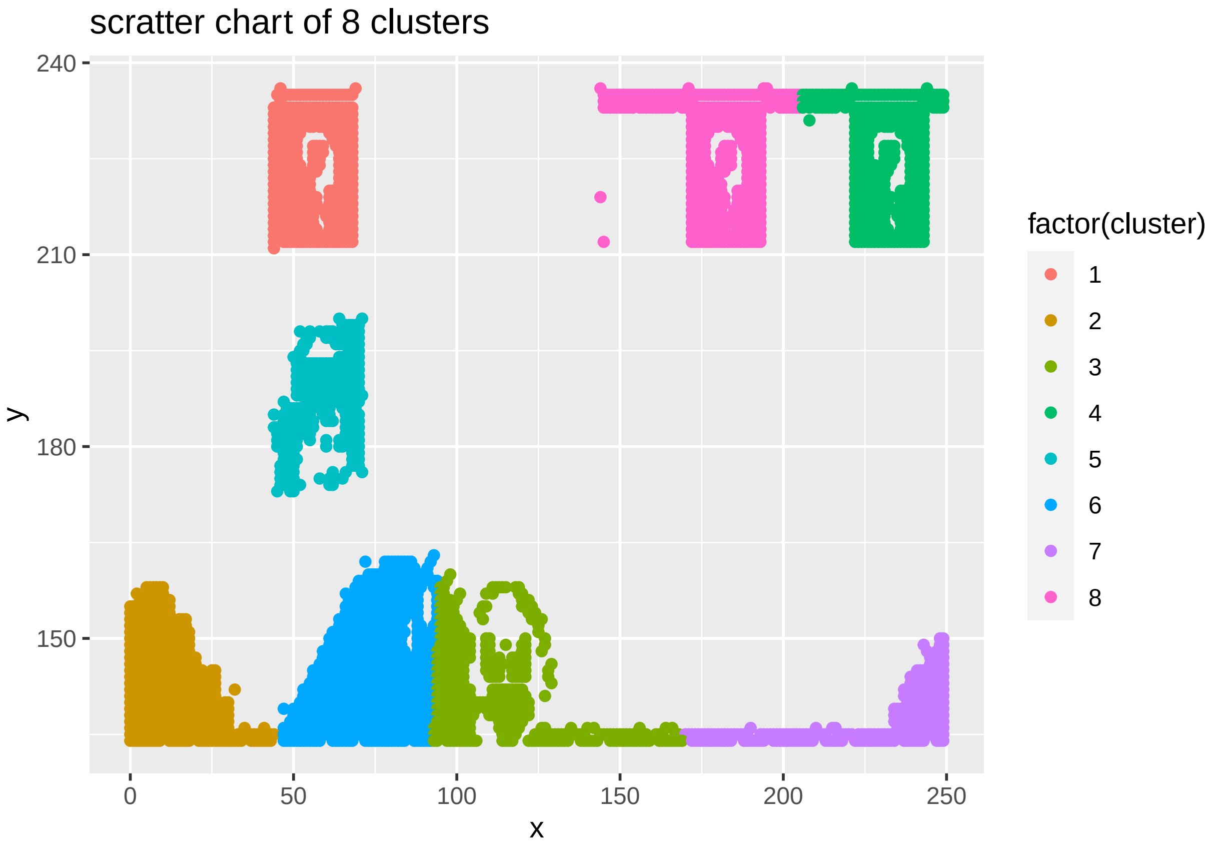


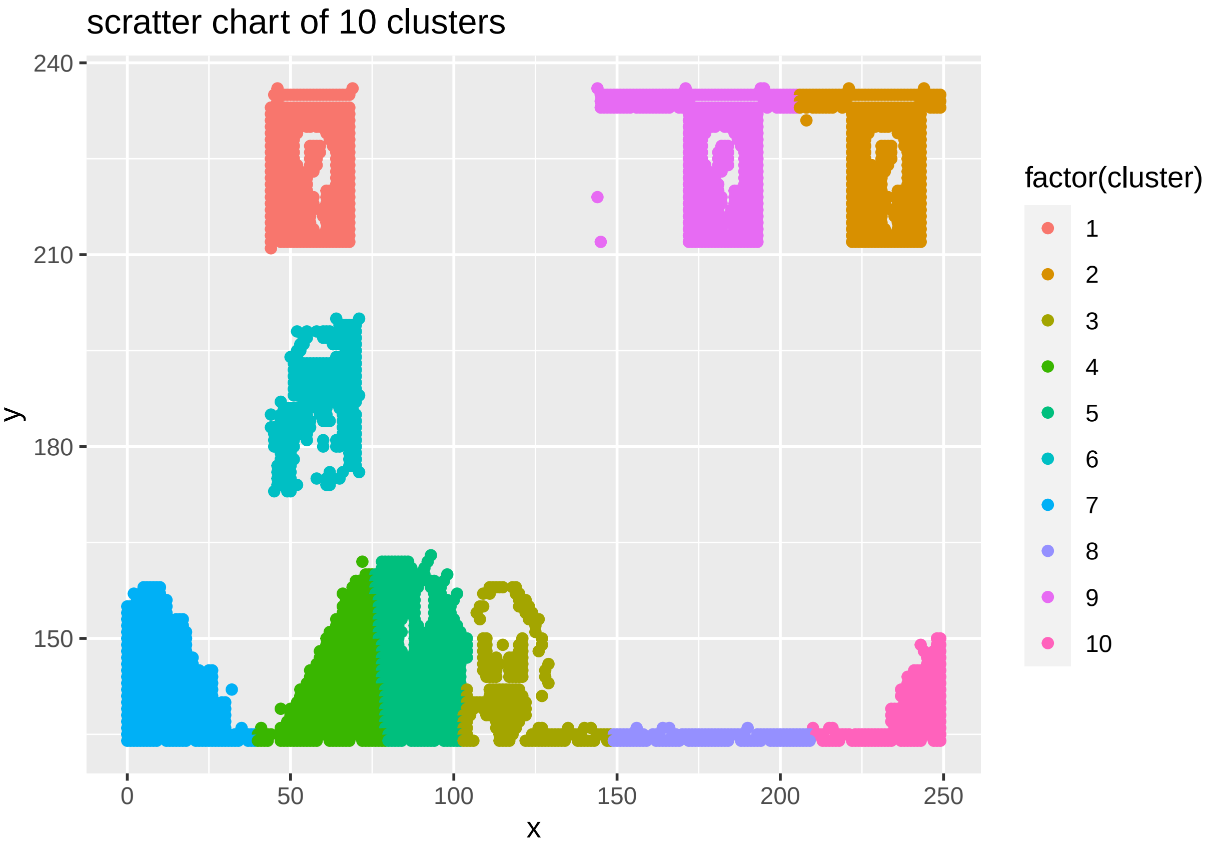
start to model using different K-value, visualize using scatter chart

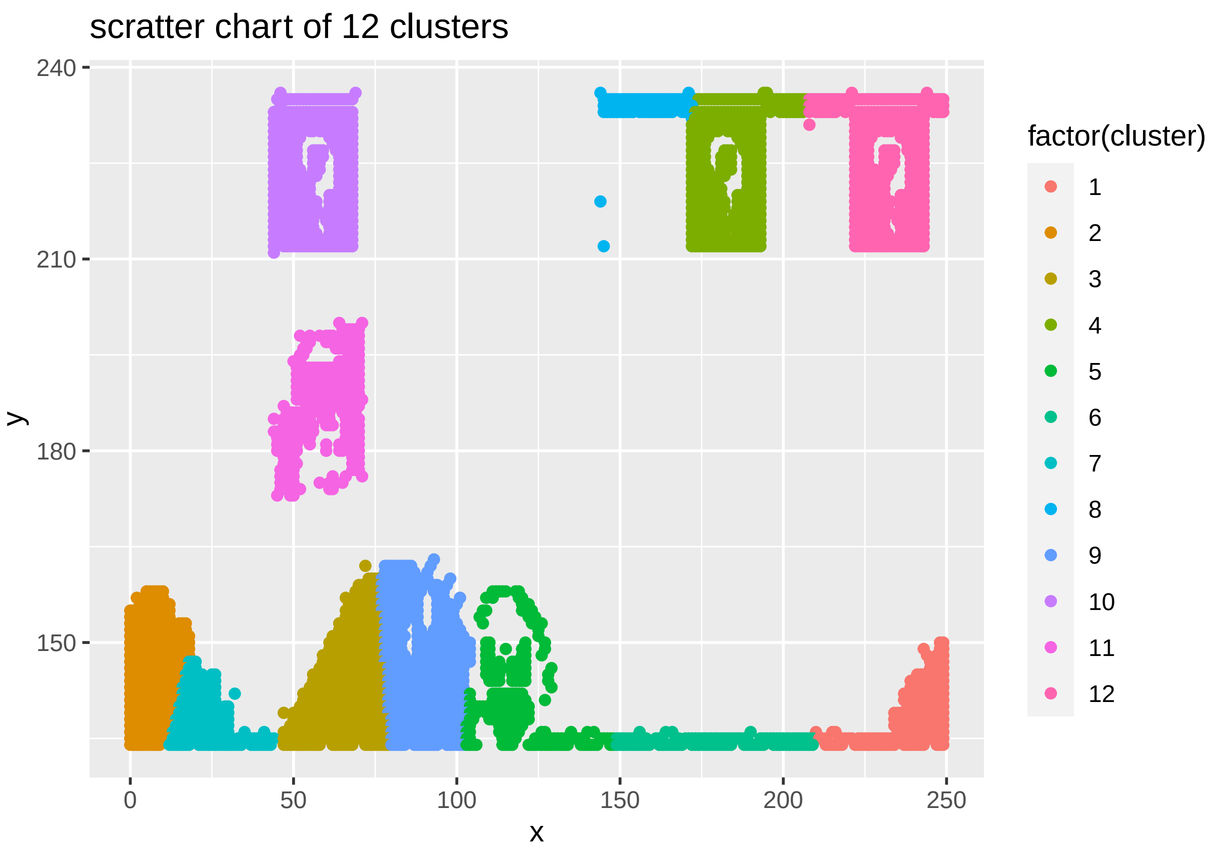




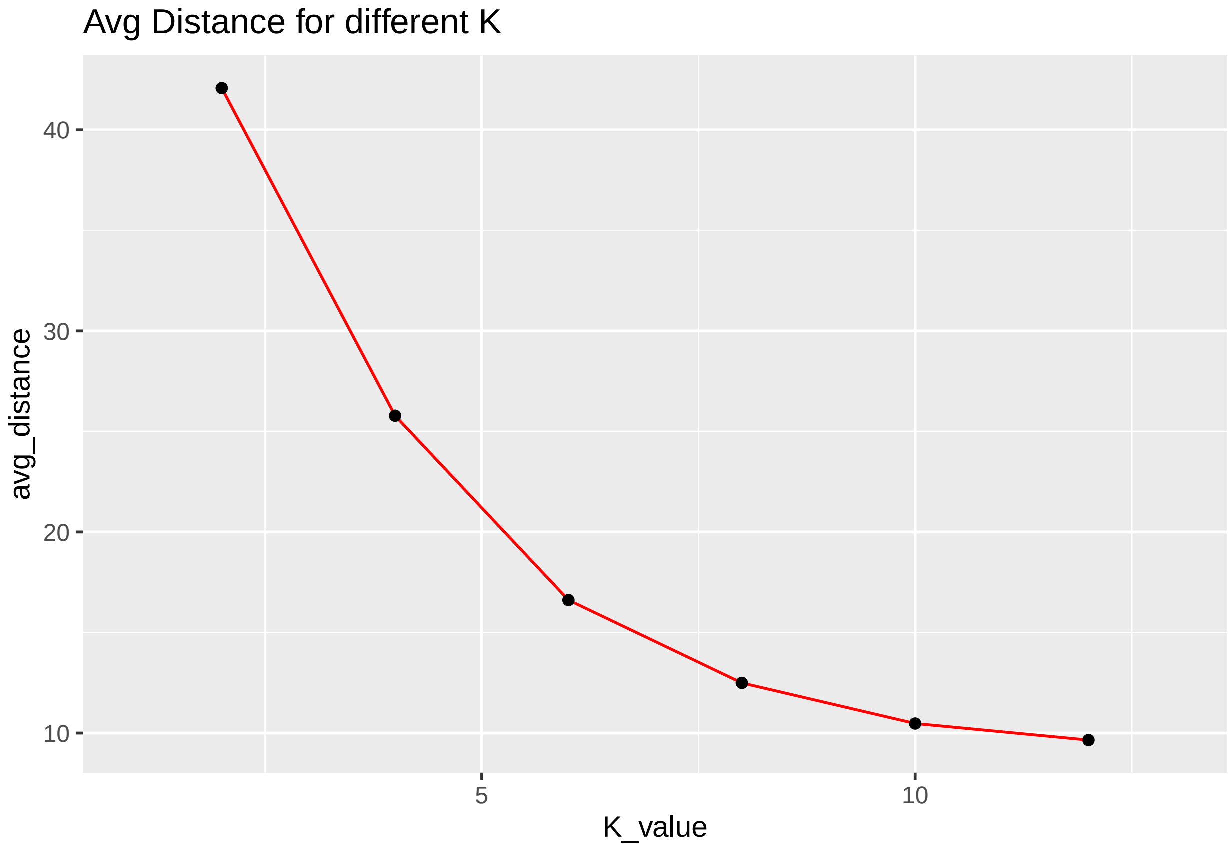

scratter chart of 6 clusters
240
210
180
150
y
0
50
100
150
200
250
x
factor(cluster)
● 1
● 2
● 3
● 4
● 5
● 6


scratter chart of 8 clusters
240
210
180
150
y
0
50
100
150
200
250
x
factor(cluster)
1
2
3
4
5
6
7
8


scratter chart of 10 clusters
240
210
180
150
y
0
50
100
150
200
250
x
factor(cluster)
1
2
3
4
5
6
7
8
9
10

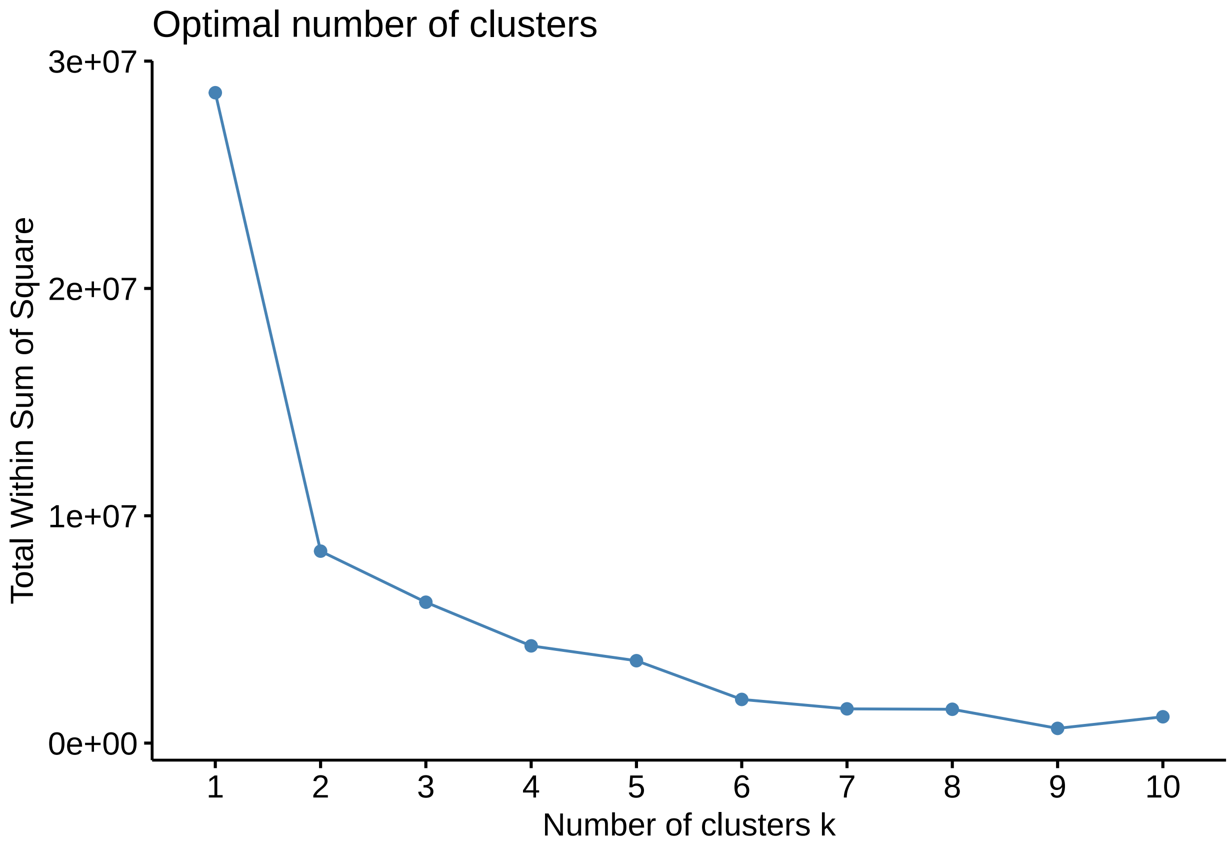

scratter chart of 12 clusters
240
210
180
150
y
0
50
100
150
200
250
x
factor(cluster)
1
2
3
4
5
6
7
8
9
10
11
12

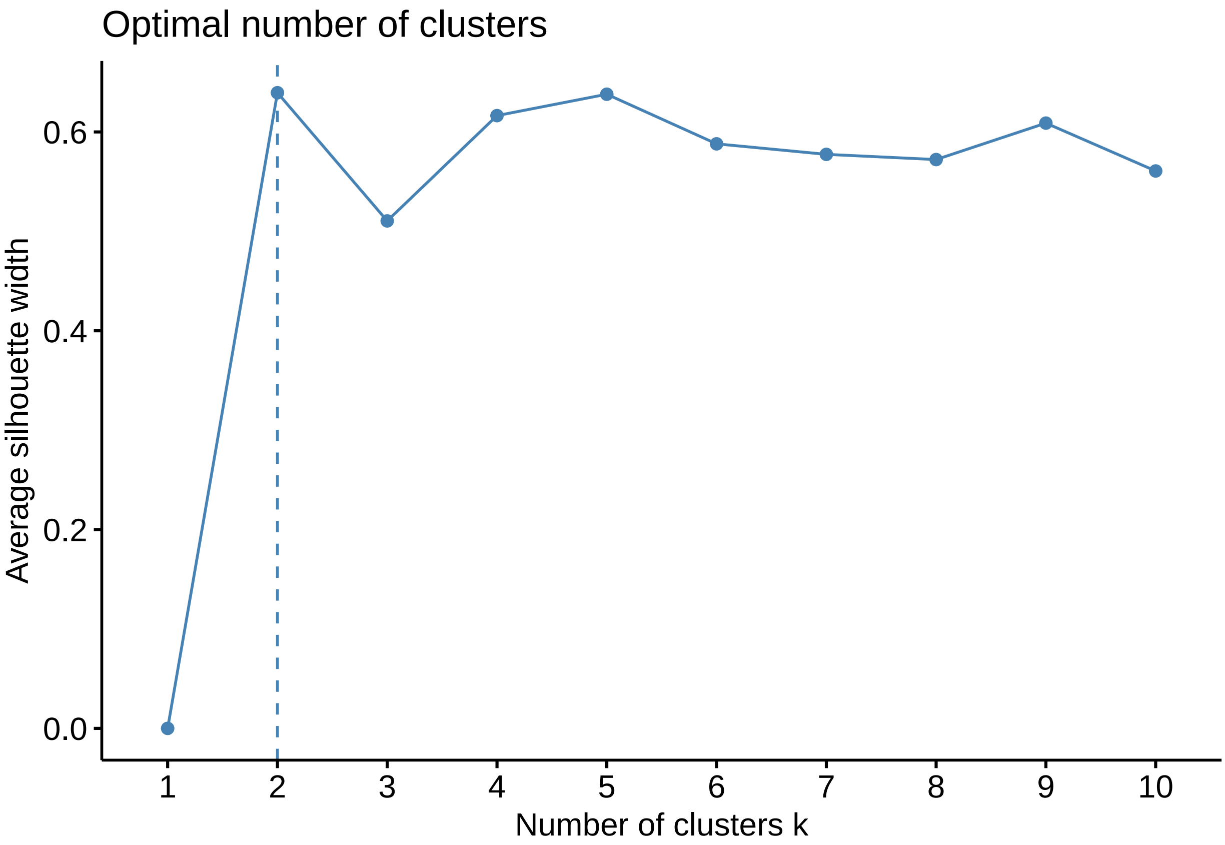
Finally do a comparison of average distance under different K

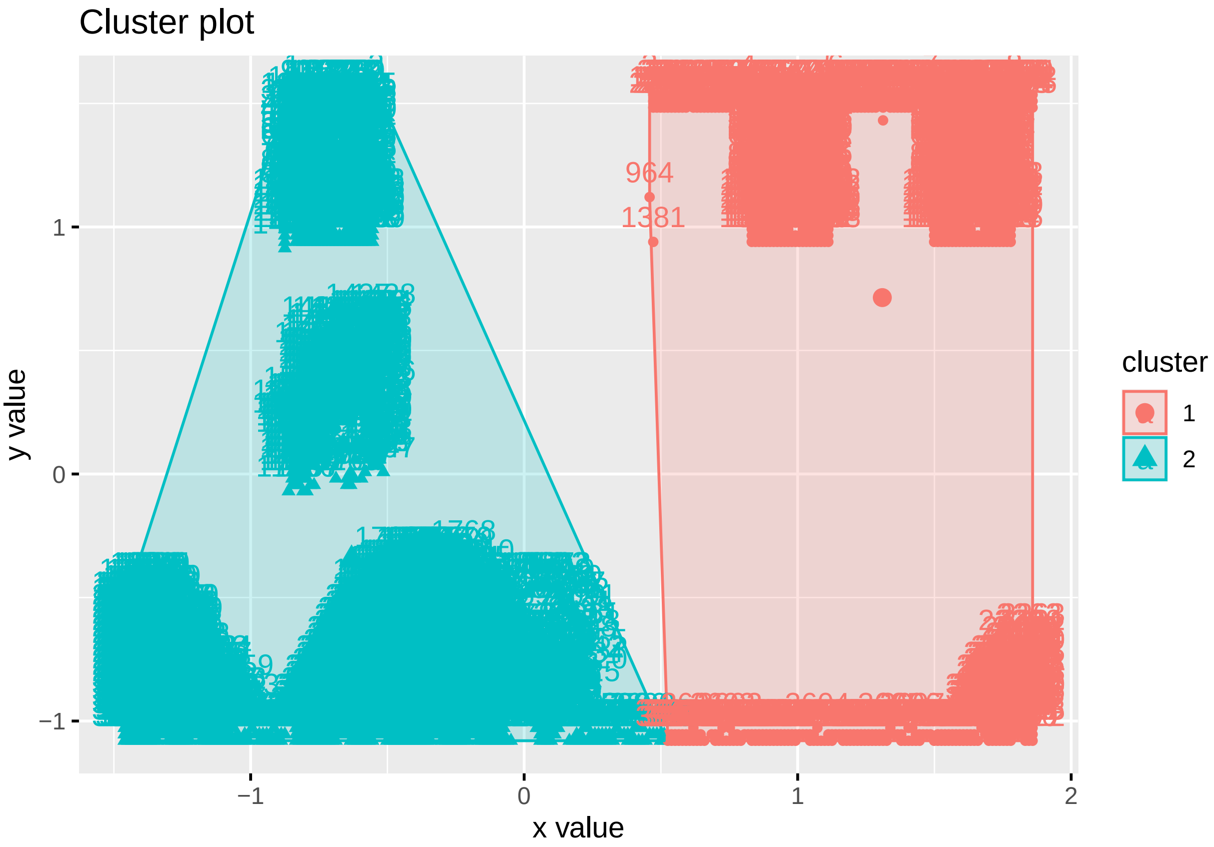


Finally, using Factoextra package to visualize the optimal K-value

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


Optimal number of clusters
Total Within Sum of Square
3e+07
2e+07
1e+07
0e+00
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Optimal number of clusters
Average silhouette width
0.0
0.2
0.4
0.6
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Cluster plot
y value
x value
1
0
−1
−1
0
1
2
964
1381
cluster
1
2