```
---
title: "ASSIGNMENT 8.3_Clustering"
author: "Bhushan Suryawanshi"
date: '2020-07-13'

---
```

Labeled data is not always available. For these types of datasets, you can use unsupervised algorithms to extract structure. The k-means clustering algorithm and the k nearest neighbor algorithm both use the Euclidean distance between points to group data points. The difference is the k-means clustering algorithm does not use labeled data.

In this problem, you will use the k-means clustering algorithm to look for patterns in an unlabeled dataset. The dataset for this problem is found at data/clustering-data.csv.

a. Plot the dataset using a scatter plot.

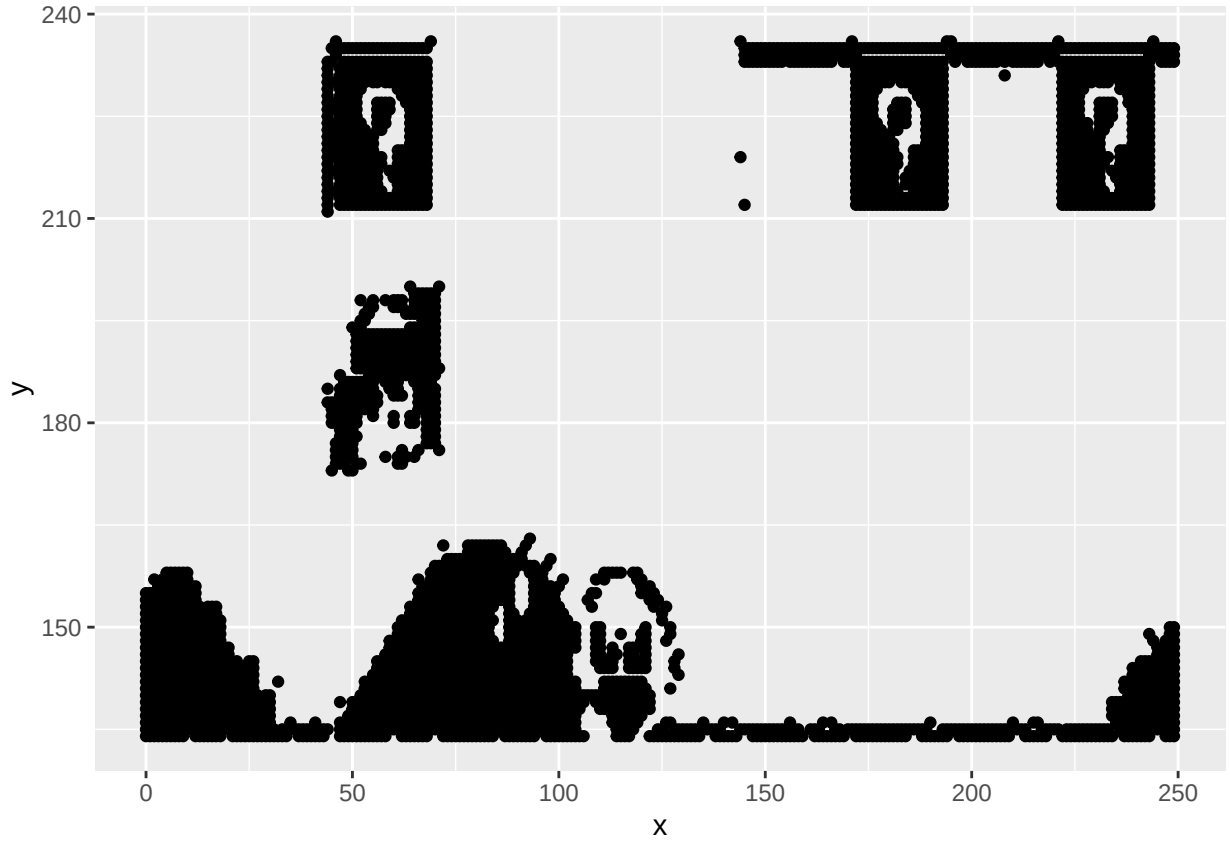
```
library(ggplot2)
clustering_df <- read.csv("clustering-data.csv")
head(clustering_df)
```

```
##     x   y
## 1  46 236
## 2  69 236
## 3 144 236
## 4 171 236
## 5 194 236
## 6 195 236
```

```
str(clustering_df)
```

```
## 'data.frame':    4022 obs. of  2 variables:
##  $ x: int  46 69 144 171 194 195 221 244 45 47 ...
##  $ y: int  236 236 236 236 236 236 236 236 235 235 ...
```

```
ggplot(clustering_df, aes(x=x, y=y)) + geom_point()
```



b. Fit the dataset using the k-means algorithm from k=2 to k=12. Create a scatter plot of the resultant clusters for each value of k.

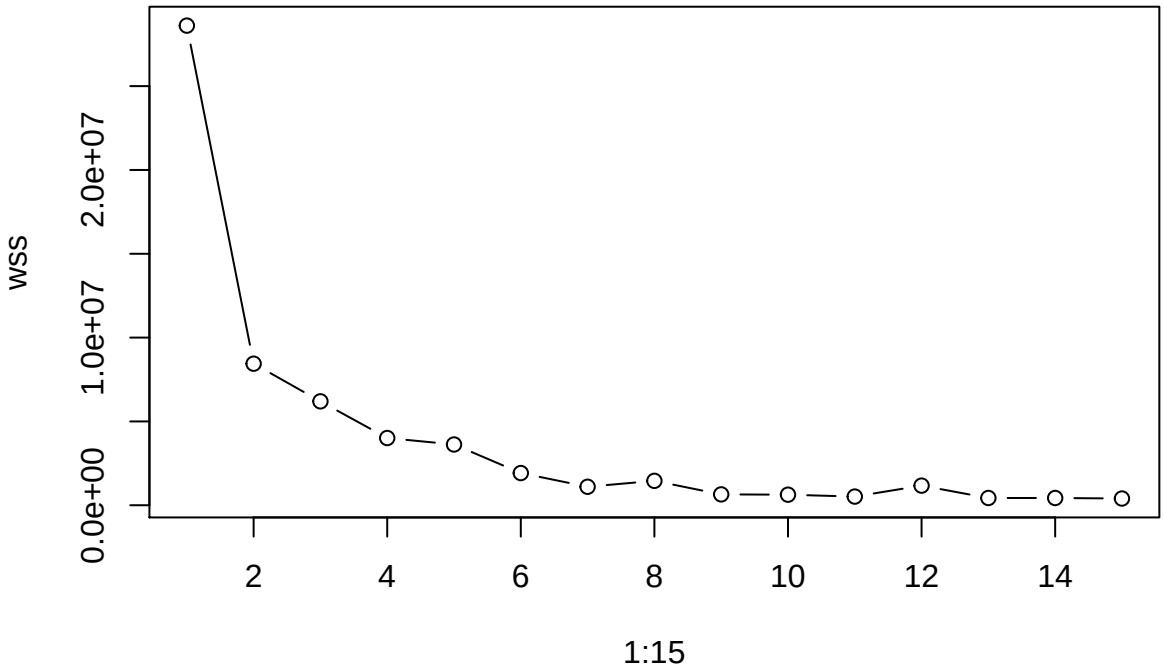
```
matrix_data <- data.matrix(clustering_df)

wss <- (nrow(matrix_data) -1)*sum(apply(matrix_data,2,var))
for (i in 2:15) wss[i]<-sum(kmeans(matrix_data, centers = i)$withinss)
wss
```

```
##  [1] 28608984.7  8443681.1  6196566.0  4009678.4  3623864.0  1920708.1
##  [7]  1102869.9  1456085.6   647331.9   629241.4   517272.6  1170142.0
## [13]   432710.8   432463.7   401373.7
```

```
#Plot to find Elbow point
plot(1:15,wss, type = "b", color="red")
```

```
## Warning in plot.window(...): "color" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "color" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "color" is not a
## graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "color" is not a
## graphical parameter
```

```
## Warning in box(...): "color" is not a graphical parameter

## Warning in title(...): "color" is not a graphical parameter
```



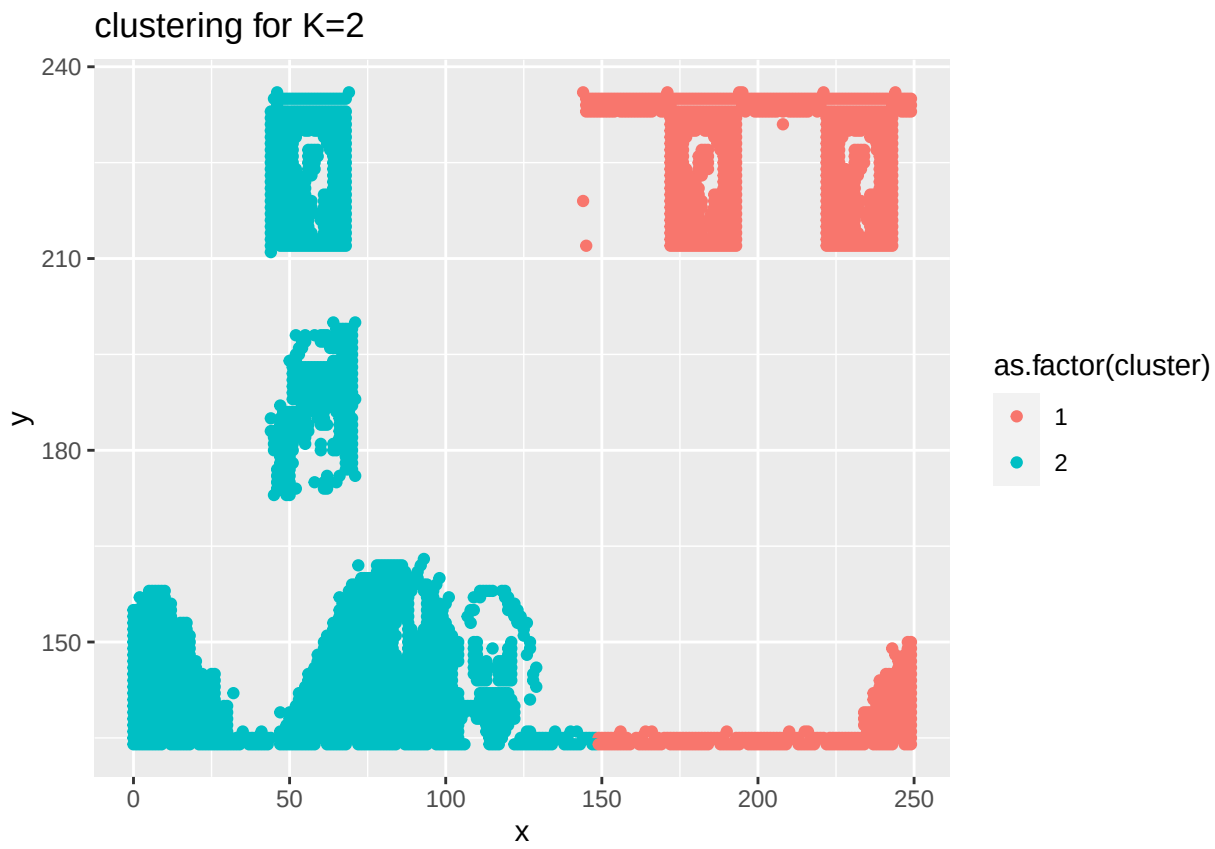
c. As k-means is an unsupervised algorithm, you cannot compute the accuracy as there are no correct values to compare the output to. Instead, you will use the average distance from the center of each cluster as a measure of how well the model fits the data. To calculate this metric, simply compute the distance of each data point to the center of the cluster it is assigned to and take the average value of all of those distances.

Calculate this average distance from the center of each cluster for each value of k and plot it as a line chart where k is the x-axis and the average distance is the y-axis.

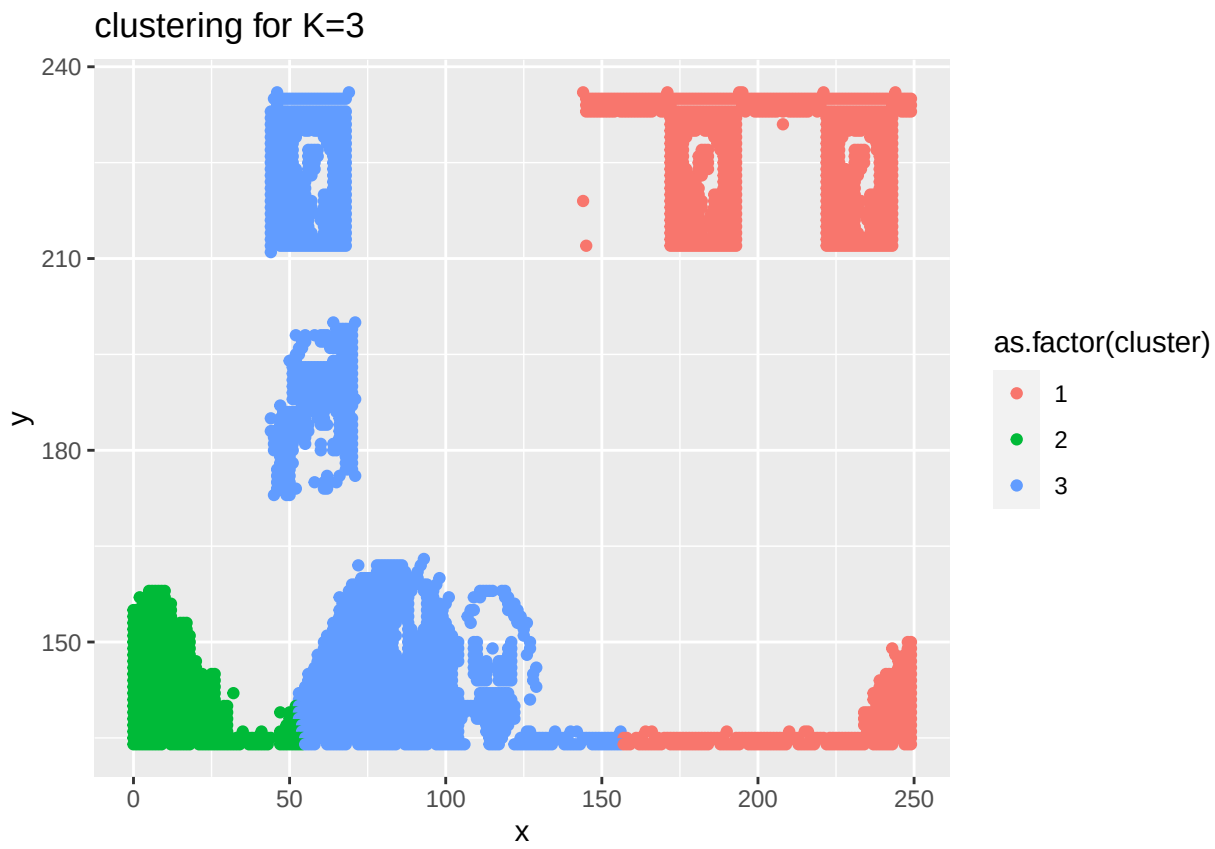
One way of determining the "right" number of clusters is to look at the graph of k versus average distance and finding the "elbow point". Looking at the graph you generated in the previous example, what is the elbow point for this dataset?

Answer - Looking at the k-vs-Average distance graph the "elbow-point" is at K=6. Also with value K=6 the cluster graph shows more accuracy.

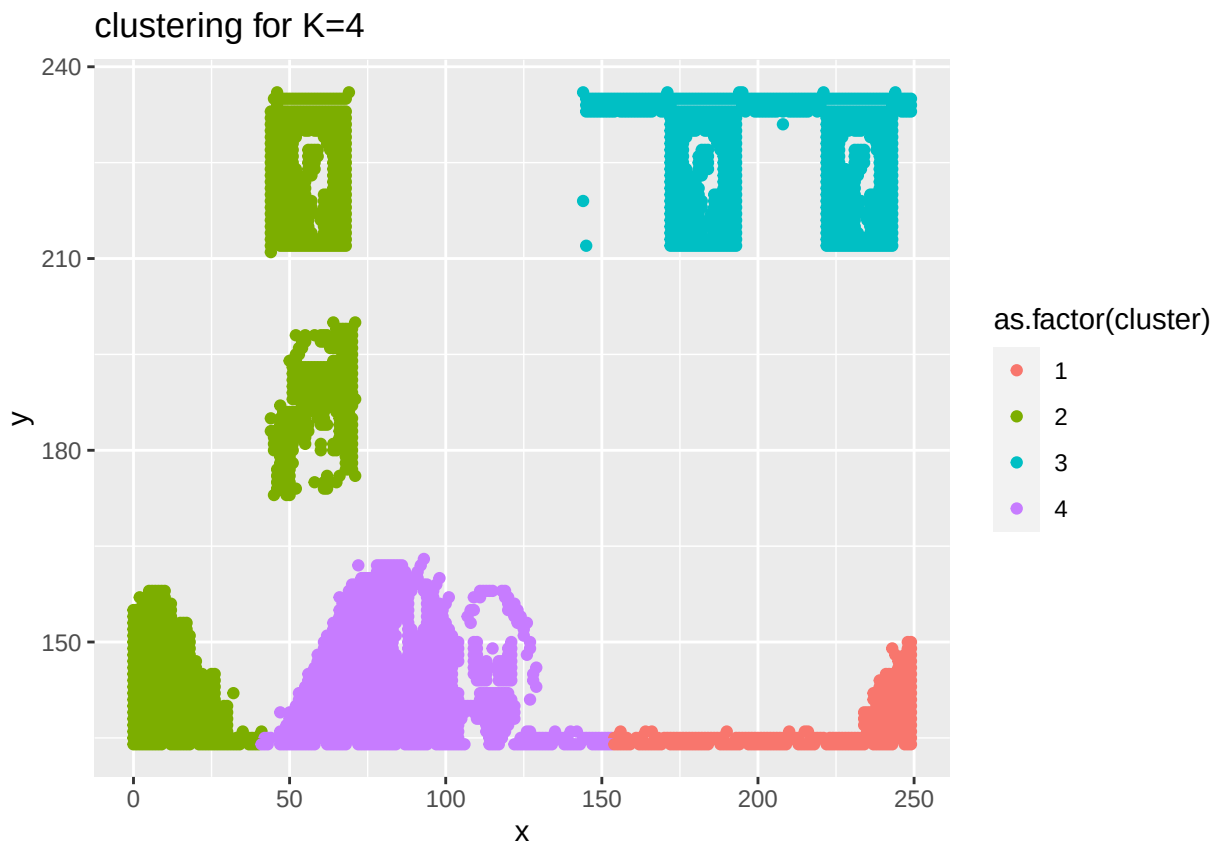
```
clusters <- kmeans(clustering_df, 2)

clustering_df$cluster <- as.factor(clusters$cluster)

ggplot(clustering_df) + geom_point(aes(x = x, y = y, colour = as.factor(cluster)),data = clustering_df)
  ggtitle("clustering for K=2")
```



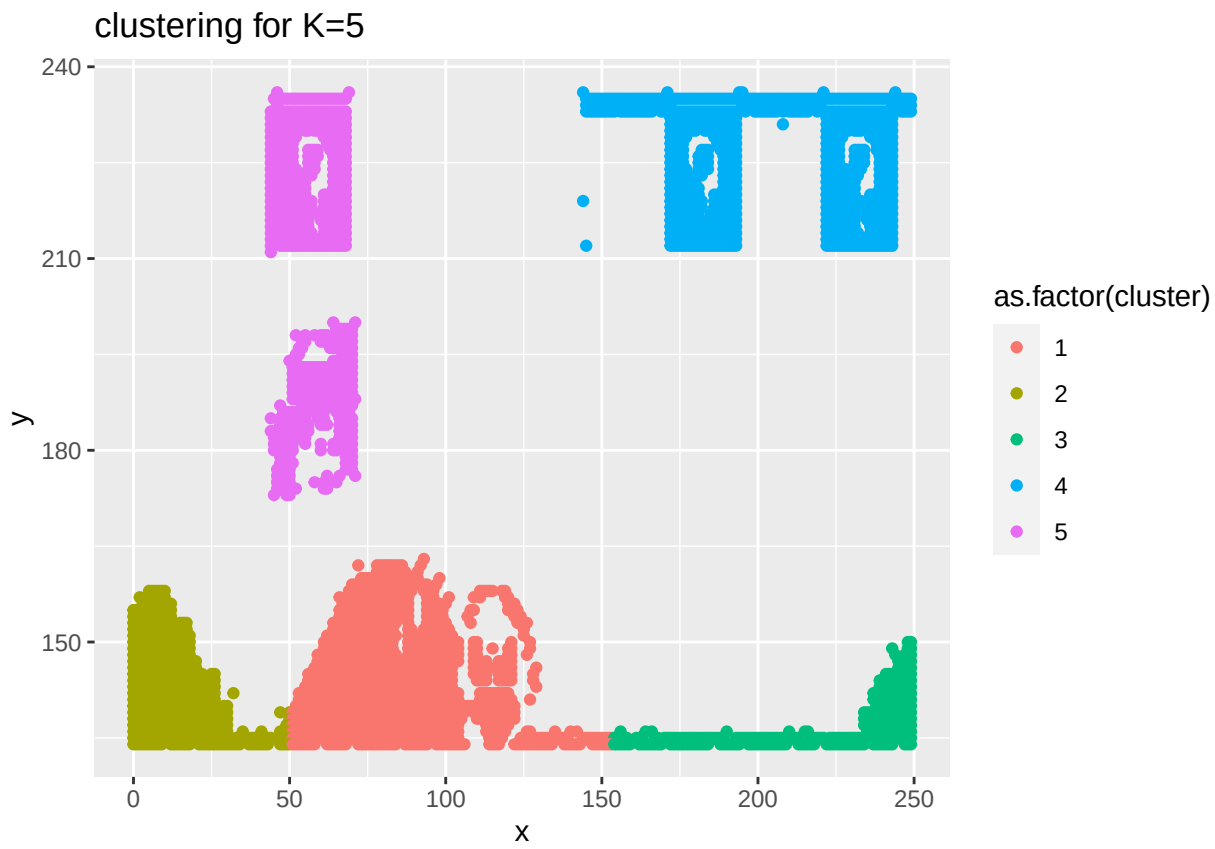

```
clusters <- kmeans(clustering_df, 3)

clustering_df$cluster <- as.factor(clusters$cluster)

ggplot(clustering_df) + geom_point(aes(x = x, y = y, colour = as.factor(cluster)),data = clustering_df)
  ggtitle("clustering for K=3")
```



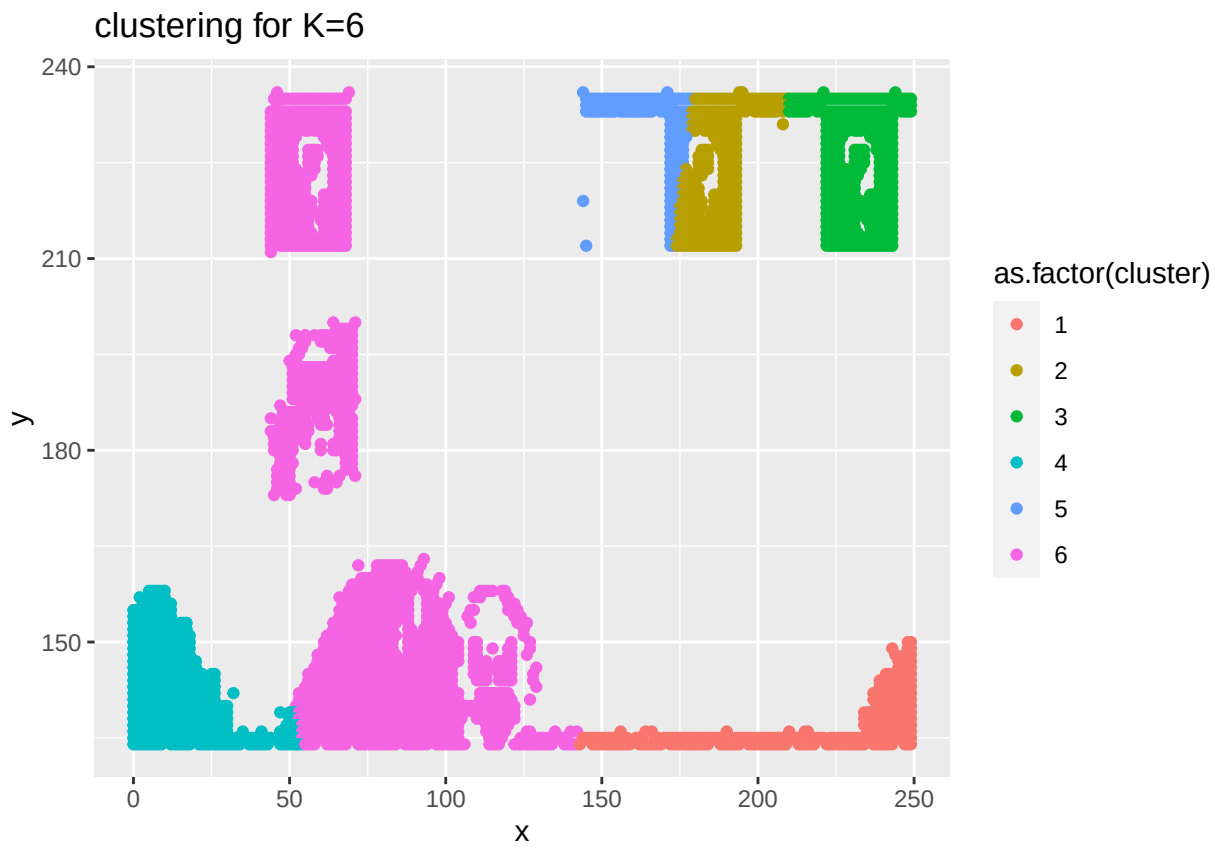

```
clusters <- kmeans(clustering_df, 4)

clustering_df$cluster <- as.factor(clusters$cluster)

ggplot(clustering_df) + geom_point(aes(x = x, y = y, colour = as.factor(cluster)),data = clustering_df)
  ggtitle("clustering for K=4")
```



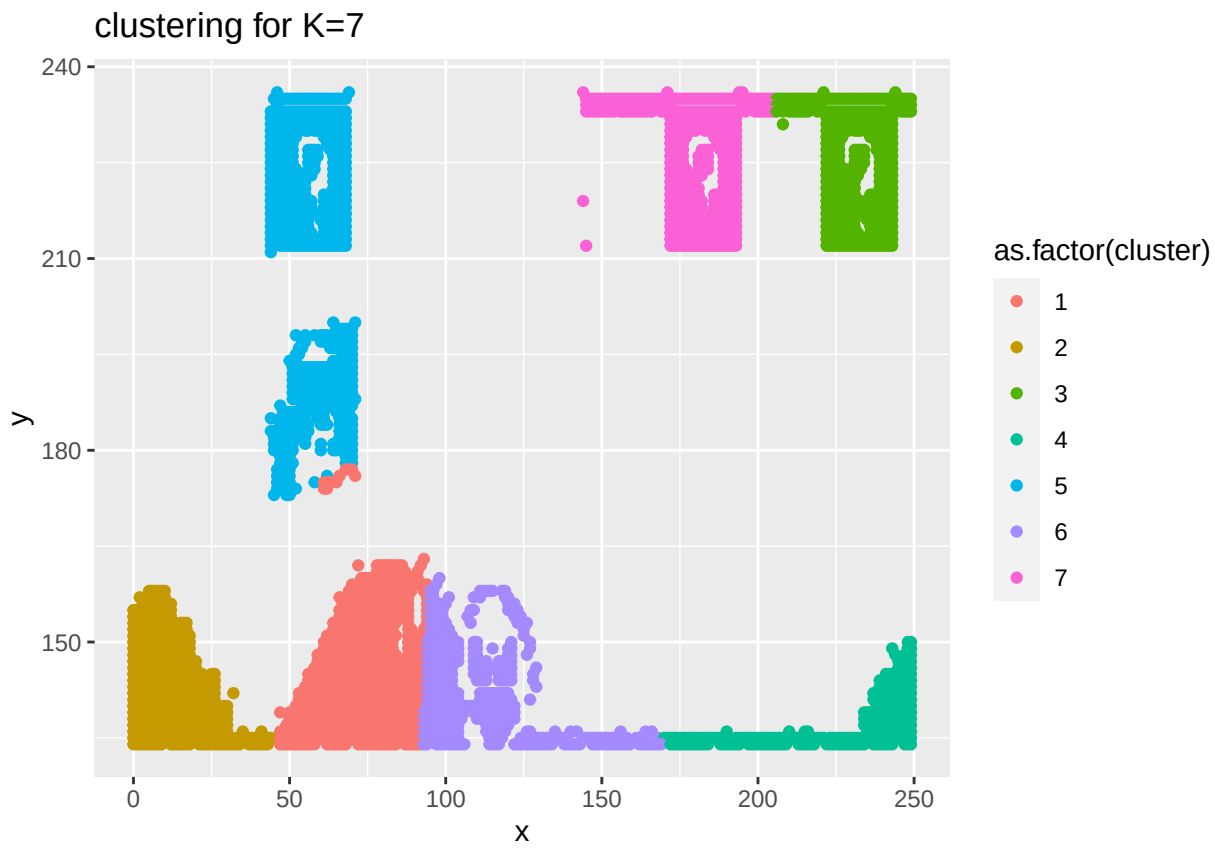

```
clusters <- kmeans(clustering_df, 5)

clustering_df$cluster <- as.factor(clusters$cluster)

ggplot(clustering_df) + geom_point(aes(x = x, y = y, colour = as.factor(cluster)),data = clustering_df)
  ggtitle("clustering for K=5")
```



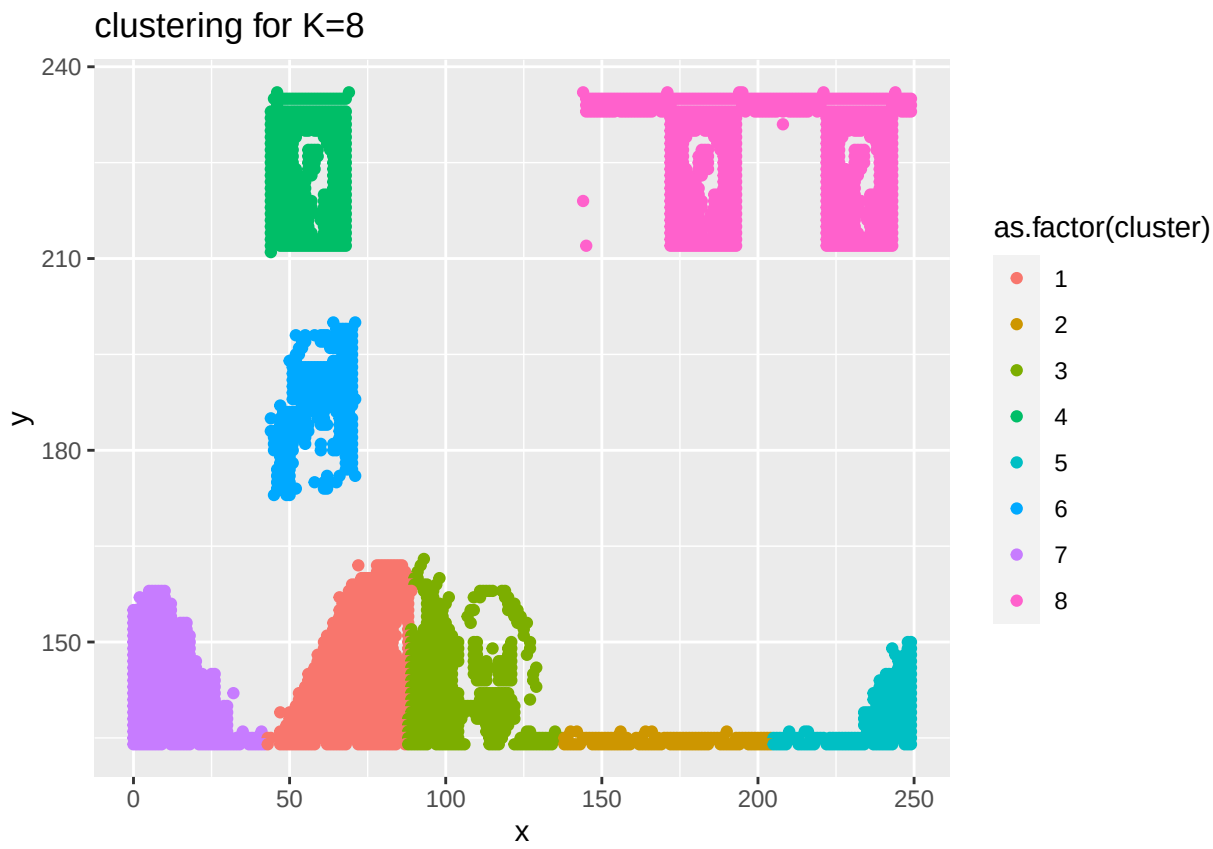

```
clusters <- kmeans(clustering_df, 6)

clustering_df$cluster <- as.factor(clusters$cluster)

ggplot(clustering_df) + geom_point(aes(x = x, y = y, colour = as.factor(cluster)),data = clustering_df)
  ggtitle("clustering for K=6")
```



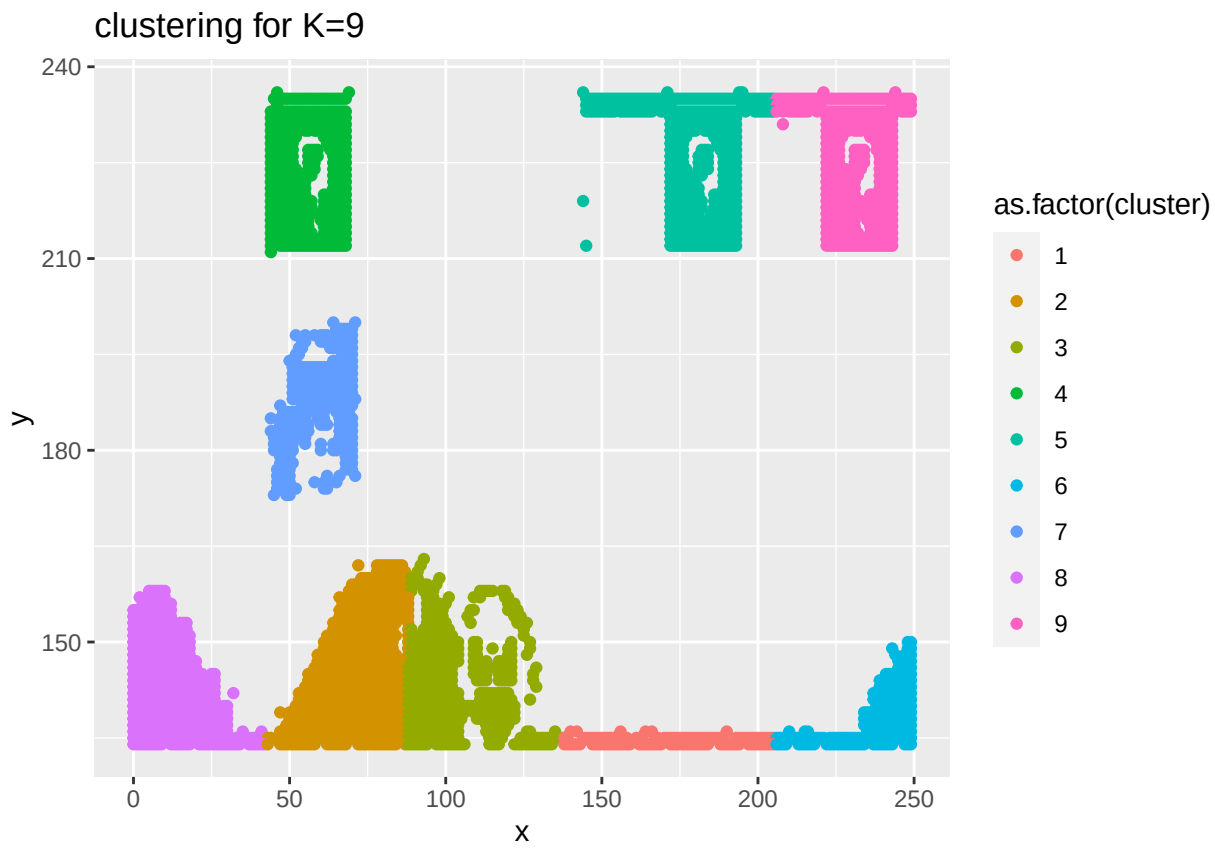

```
clusters <- kmeans(clustering_df, 7)

clustering_df$cluster <- as.factor(clusters$cluster)

ggplot(clustering_df) + geom_point(aes(x = x, y = y, colour = as.factor(cluster)),data = clustering_df)
  ggtitle("clustering for K=7")
```



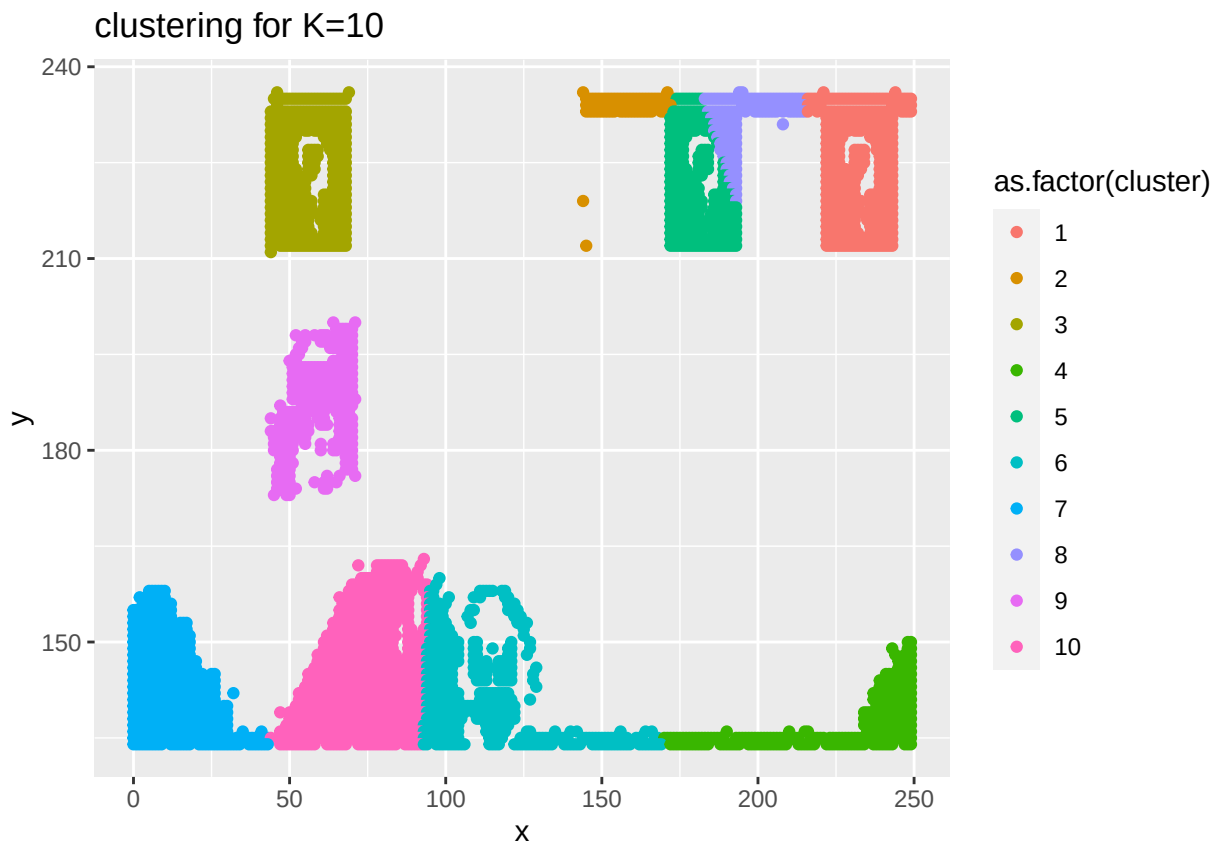

```
clusters <- kmeans(clustering_df, 8)

clustering_df$cluster <- as.factor(clusters$cluster)

ggplot(clustering_df) + geom_point(aes(x = x, y = y, colour = as.factor(cluster)),data = clustering_df)
  ggtitle("clustering for K=8")
```



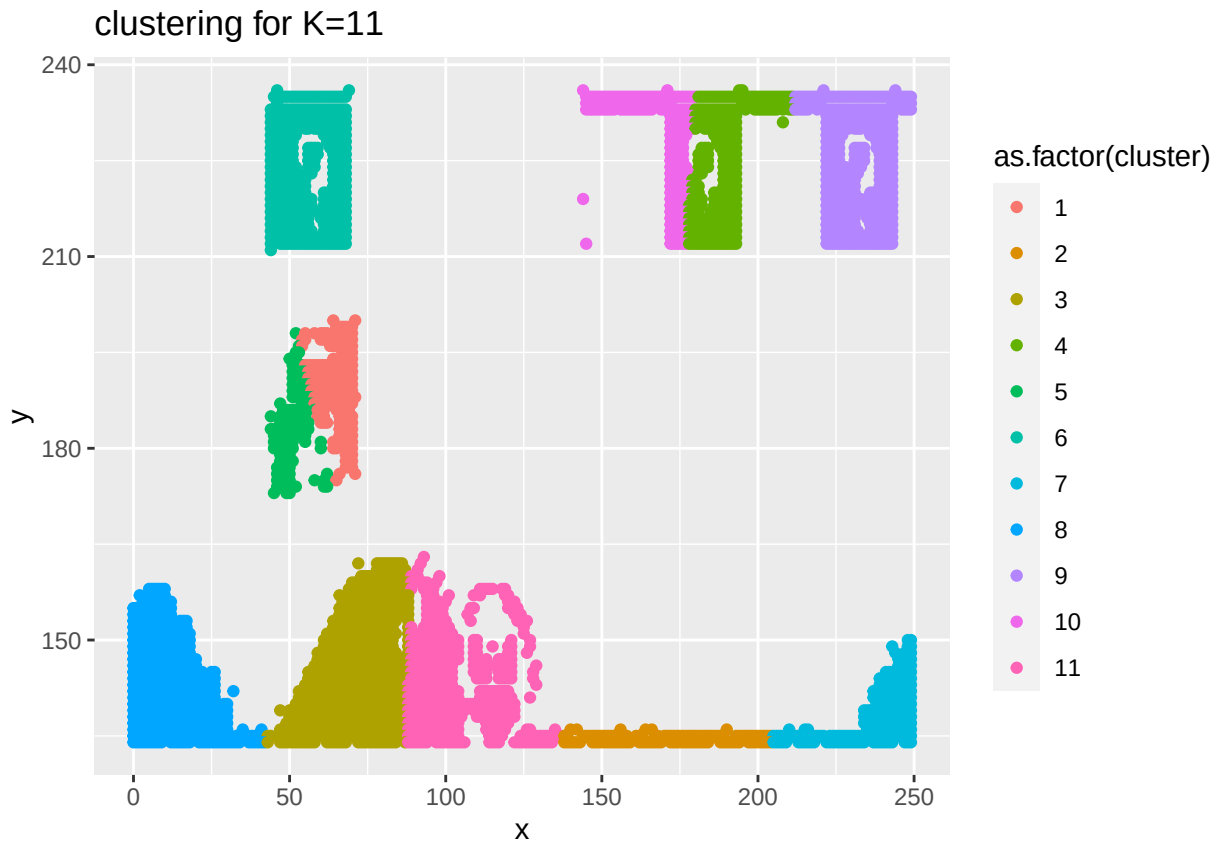

```
clusters <- kmeans(clustering_df, 9)

clustering_df$cluster <- as.factor(clusters$cluster)

ggplot(clustering_df) + geom_point(aes(x = x, y = y, colour = as.factor(cluster)),data = clustering_df)
  ggtitle("clustering for K=9")
```



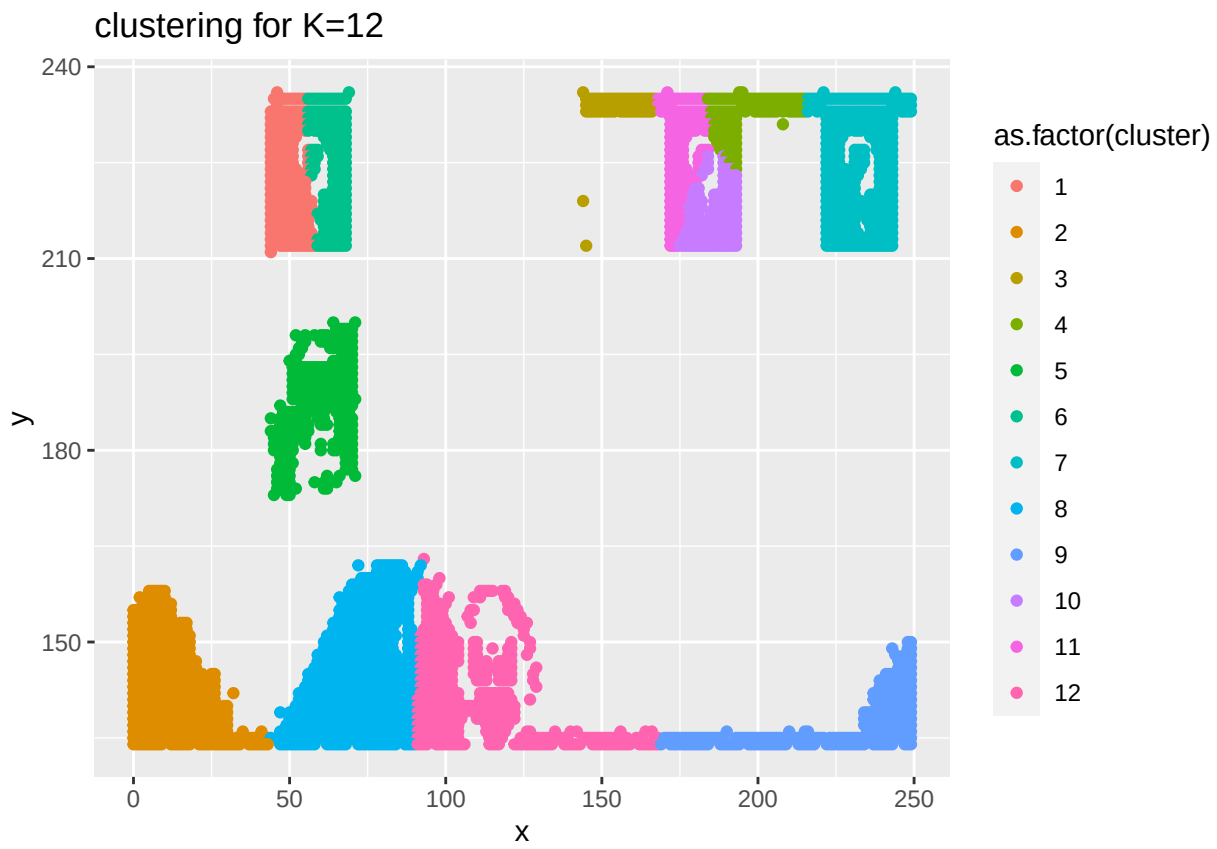

```
clusters <- kmeans(clustering_df, 10)

clustering_df$cluster <- as.factor(clusters$cluster)

ggplot(clustering_df) + geom_point(aes(x = x, y = y, colour = as.factor(cluster)),data = clustering_df)
  ggtitle("clustering for K=10")
```




```
clusters <- kmeans(clustering_df, 11)

clustering_df$cluster <- as.factor(clusters$cluster)

ggplot(clustering_df) + geom_point(aes(x = x, y = y, colour = as.factor(cluster)),data = clustering_df)
  ggtitle("clustering for K=11")
```




```
clusters <- kmeans(clustering_df, 12)

clustering_df$cluster <- as.factor(clusters$cluster)

ggplot(clustering_df) + geom_point(aes(x = x, y = y, colour = as.factor(cluster)),data = clustering_df)
  ggtitle("clustering for K=12")
```


clustering for K=12
240
210
180
150
y
0
50
100
150
200
250
x
as.factor(cluster)
1
2
3
4
5
6
7
8
9
10
11
12