Assignment 9gw : Exercise 16: Clustering

Ragunath Gunasekaran

2020-11-01

Footnote

This is a Footnote test.

Citations

- R for Everyone
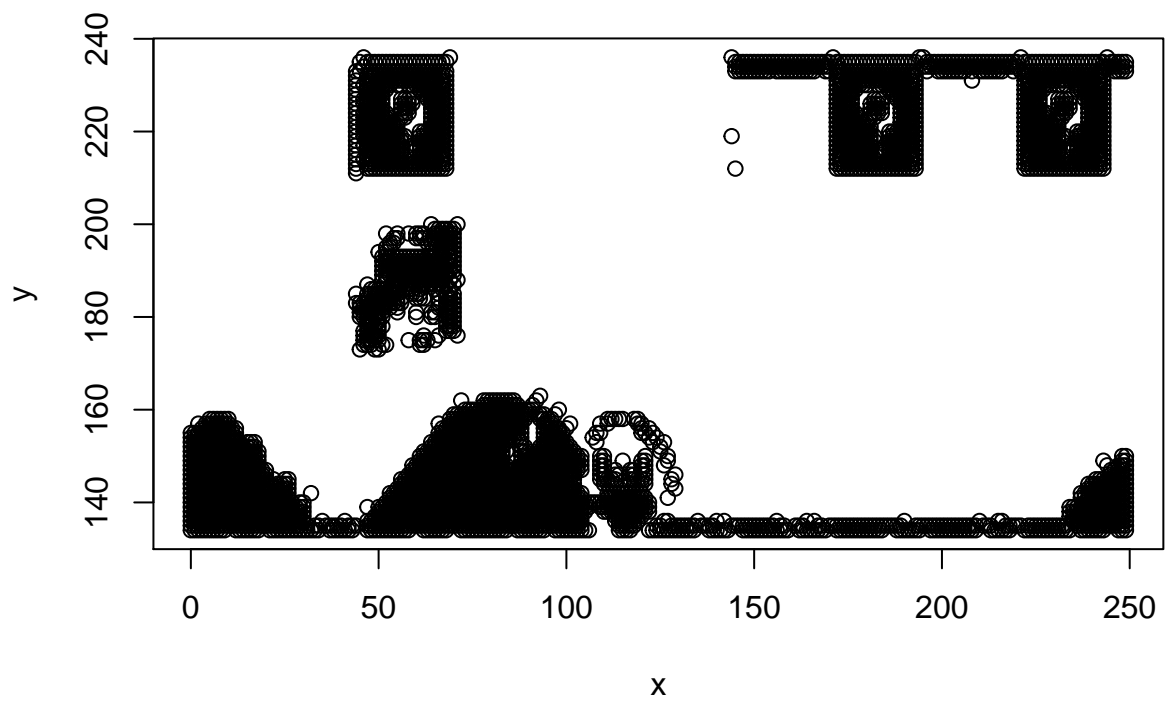- Discovering Statistics Using R

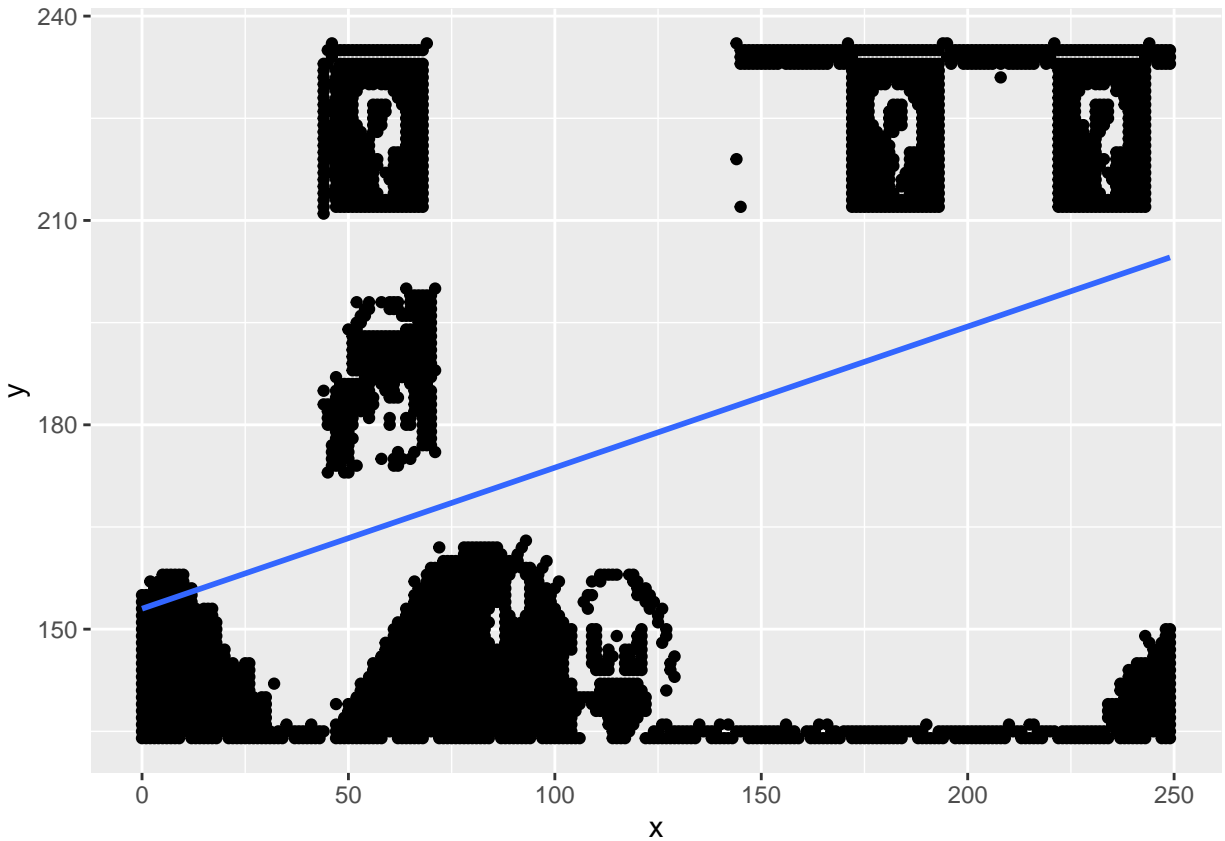## Data Load - Data from CSV file load into Dataframe

Structure Of An Arbitrary R Object

```
## '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 ...
```

Scatter plot of clustering dataset by using ggplot2

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

Observation : Based on this plot, I have seen various clusters and the high groups are 7 and few small groups are there too.

Fiting 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.

Cluster Data for each K Value

#Ref - Source 5,4,6

```
## Warning: package 'cluster' was built under R version 4.0.2
```

```
##          x         y
## 1   62.2815 162.4090
## 2 207.8005 203.2898
```

```
##            x         y
## 1 209.16575 205.1791
## 2  87.29886 144.3702
## 3  38.73348 180.4610
```

```
##           x         y
## 1  57.42711 207.3043
## 2 209.19733 205.2351
```

```
## 3   14.47973 142.9662
## 4   88.12209 144.5305


##              x        y
## 1   14.10580 143.0324
## 2  221.61981 137.6677
## 3   57.42711 207.3043
## 4   86.55457 144.6858
## 5  203.98579 224.8406


##              x        y
## 1  179.55019 225.2664
## 2   57.42711 207.3043
## 3  221.61981 137.6677
## 4   14.10580 143.0324
## 5   86.55457 144.6858
## 6  231.08994 224.3683


##              x        y
## 1  112.54167 142.3063
## 2   74.28817 145.8355
## 3   13.07206 143.2460
## 4  227.08362 137.9477
## 5  231.08994 224.3683
## 6   57.31518 207.7575
## 7  179.55019 225.2664


##              x        y
## 1  112.74316 142.3032
## 2  231.08994 224.3683
## 3   58.82698 186.9824
## 4  179.55019 225.2664
## 5   74.49892 145.4643
## 6   13.07206 143.2460
## 7  227.08362 137.9477
## 8   56.34467 223.0181


##              x        y
## 1   58.82698 186.9824
## 2  171.17391 134.5739
## 3  179.55019 225.2664
## 4   56.34467 223.0181
## 5  231.08994 224.3683
## 6  103.68796 143.7500
## 7   12.91166 143.2915
## 8   71.78875 145.4550
## 9  237.85841 138.8628


##              x        y
## 1   89.41301 146.7431
## 2  179.35514 134.5514
## 3  118.62931 141.4440
```

```
## 4   231.08994 224.3683
## 5    56.34467 223.0181
## 6   239.28704 139.0648
## 7    12.61071 143.3839
## 8    65.81143 143.4457
## 9   179.55019 225.2664
## 10   58.82698 186.9824


##             x        y
## 1    58.82698 186.9824
## 2   157.85714 233.6310
## 3    56.34467 223.0181
## 4   184.00683 223.7677
## 5   231.35714 224.2619
## 6   178.21495 134.5514
## 7    65.72169 143.4299
## 8    89.27110 146.7792
## 9    12.61071 143.3839
## 10  118.17167 141.4549
## 11  239.00917 139.0229


##             x        y
## 1    77.35546 149.4625
## 2    56.34467 223.0181
## 3    60.05246 139.9115
## 4   180.79808 134.5673
## 5    95.02488 143.8458
## 6   166.78378 228.9946
## 7    58.82698 186.9824
## 8    12.29295 143.4955
## 9   120.97129 141.4019
## 10  231.66009 224.1382
## 11  187.30814 223.5378
## 12  239.42326 139.0791
```
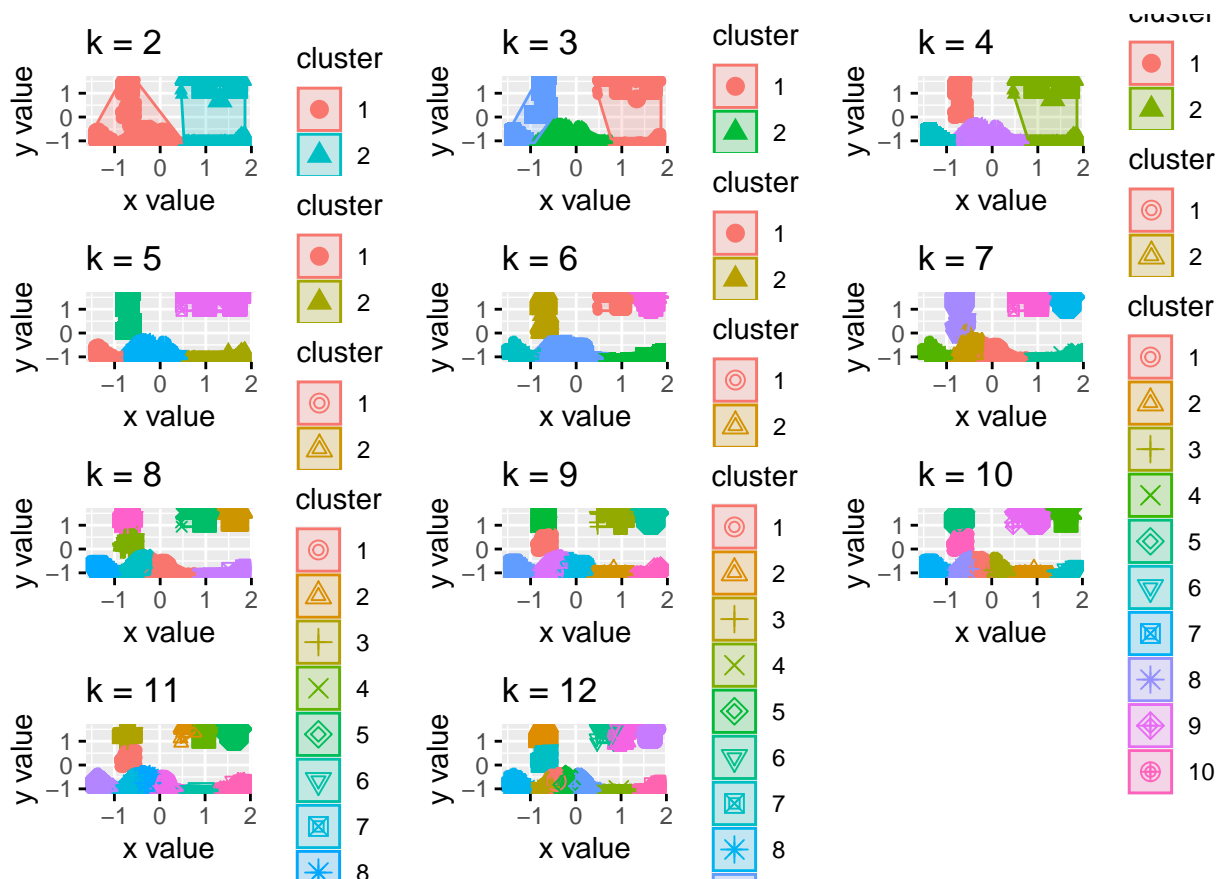
#install.package("factoextra")

Plots Comparision

#Ref Source -6,3,7

```
## Warning: package 'factoextra' was built under R version 4.0.2

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

## Warning: package 'gridExtra' was built under R version 4.0.2
```
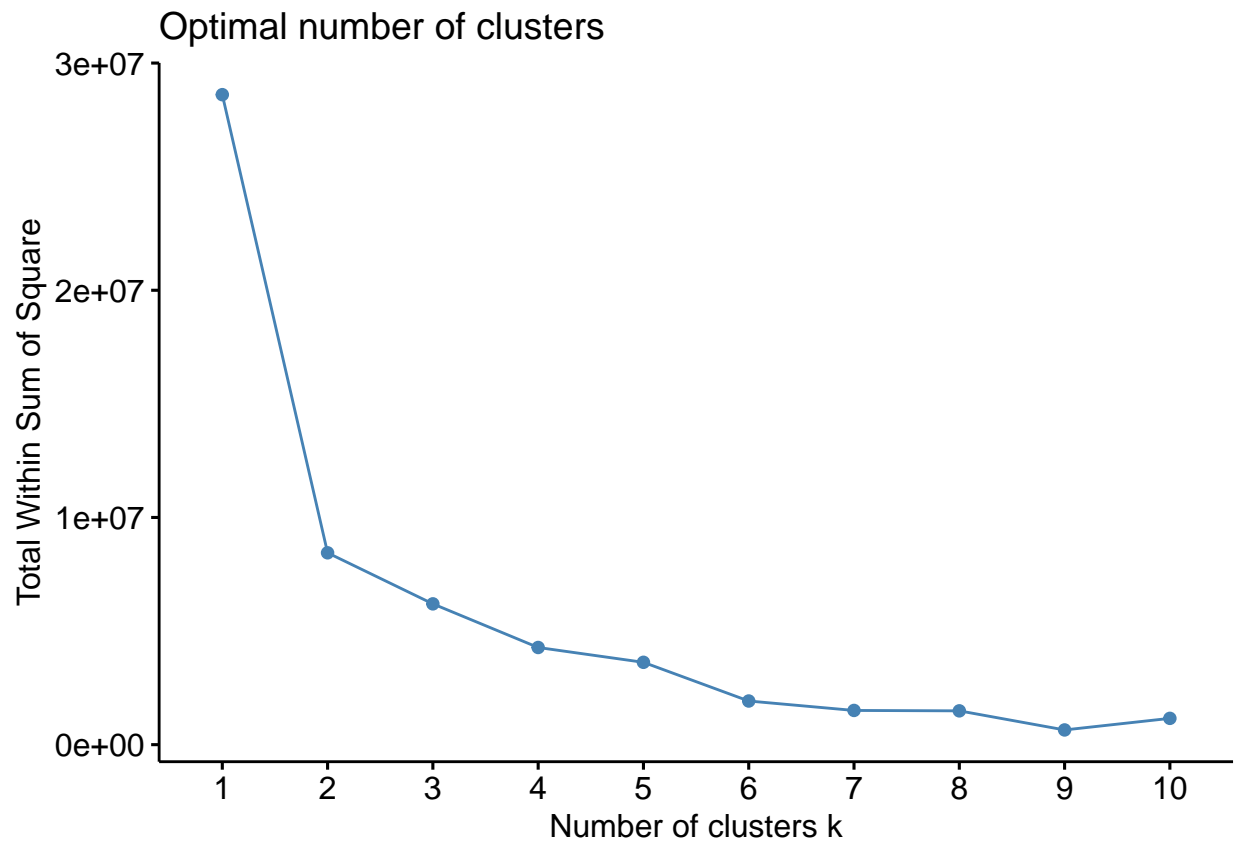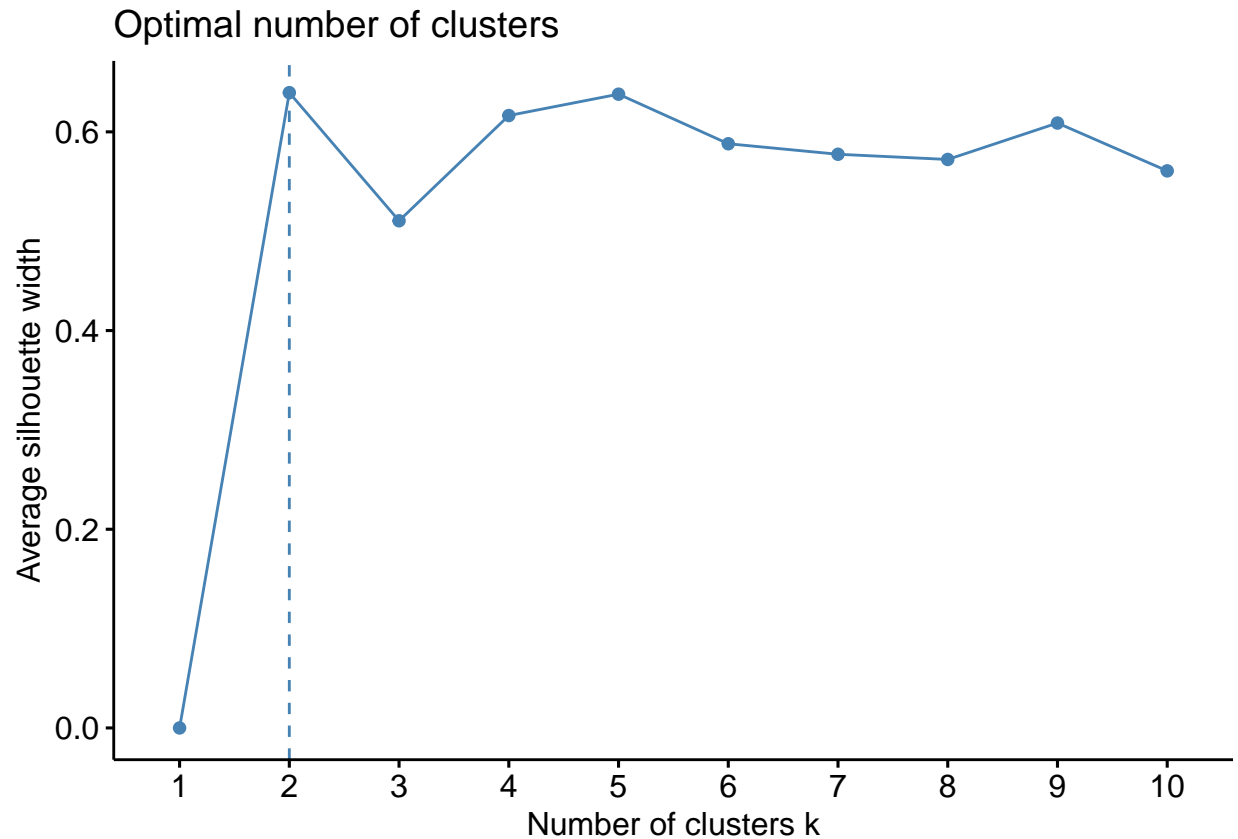
k = 2

y value
x value

cluster
- 1
- 2

k = 3

y value
x value

cluster
- 1
- 2

k = 4

y value
x value

cluster
- 1
- 2

cluster
- 1
- 2

k = 5

y value
x value

cluster
- 1
- 2

k = 6

y value
x value

cluster
- 1
- 2

k = 7

y value
x value

cluster
- 1
- 2

cluster
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8
- 9
- 10

k = 8

y value
x value

cluster
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

k = 9

y value
x value

cluster
- 1
- 2
- 3
- 4
- 5
- 6
- 7
- 8

k = 10

y value
x value

k = 11

y value
x value

k = 12

y value
x value

Average Silhouette Method

#Ref Source -6,3,7

Optimal number of clusters

Optimal number of clusters

Conclusion:

A high average silhouette width indicates a good clustering. In my view, The disadvantage of K-means clustering is that need to specify cluster details.

References

1. Sejal Jaiswal. (2018). K-Means Clustering in R Tutorial
2. https://rpubs.com/dnchari/ggplot2
3. https://towardsdatascience.com/clustering-with-k-means-1e07a8bfb7ca
4. https://rpubs.com/dnchari/kmeans
5. https://www.youtube.com/watch?v=3GorGZgTTEk
6. https://rpubs.com/abdul_yunus/Kmeans_Clustering
7. https://cran.r-project.org/web/packages/factoextra/factoextra.pdf