Assignment 9gw : Exercise 16: Clustering

Ragunath Gunasekaran

2020-11-01

## Footnote

This is a Footnote test.

## Citations

* R for Everyone
* Discovering Statistics Using R

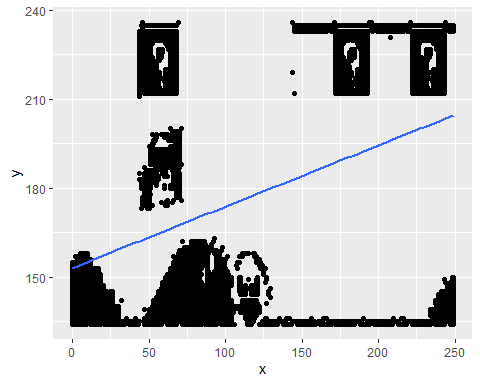
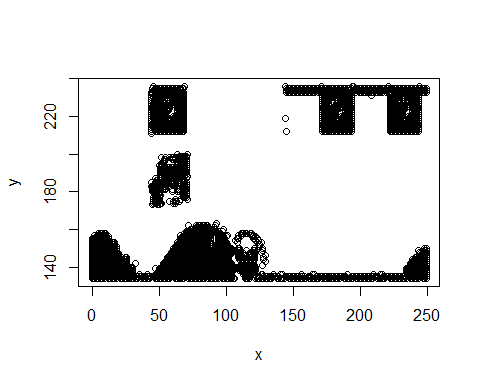
## Data Load - Data from CSV file load into Dataframe

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  
## 1 207.8005 203.2898  
## 2 62.2815 162.4090

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

## x y  
## 1 88.12209 144.5305  
## 2 209.19733 205.2351  
## 3 14.47973 142.9662  
## 4 57.42711 207.3043

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

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

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

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

## x y  
## 1 56.34467 223.0181  
## 2 231.08994 224.3683  
## 3 103.77737 143.7245  
## 4 71.80899 145.4594  
## 5 179.55019 225.2664  
## 6 58.82698 186.9824  
## 7 12.91166 143.2915  
## 8 238.15179 138.9018  
## 9 172.05172 134.5690

## x y  
## 1 231.08994 224.3683  
## 2 179.35514 134.5514  
## 3 179.55019 225.2664  
## 4 65.81143 143.4457  
## 5 58.82698 186.9824  
## 6 12.61071 143.3839  
## 7 56.34467 223.0181  
## 8 239.28704 139.0648  
## 9 89.41301 146.7431  
## 10 118.62931 141.4440

## x y  
## 1 56.34467 223.0181  
## 2 184.00683 223.7677  
## 3 179.35514 134.5514  
## 4 65.81143 143.4457  
## 5 157.85714 233.6310  
## 6 12.61071 143.3839  
## 7 58.82698 186.9824  
## 8 231.35714 224.2619  
## 9 89.41301 146.7431  
## 10 118.62931 141.4440  
## 11 239.28704 139.0648

## x y  
## 1 58.82698 186.9824  
## 2 184.00683 223.7677  
## 3 231.35714 224.2619  
## 4 94.84158 143.7327  
## 5 239.28704 139.0648  
## 6 60.12052 139.9218  
## 7 157.85714 233.6310  
## 8 56.34467 223.0181  
## 9 120.57895 141.6077  
## 10 77.33189 149.5922  
## 11 179.93333 134.5524  
## 12 12.29295 143.4955

#install.package(“factoextra”)

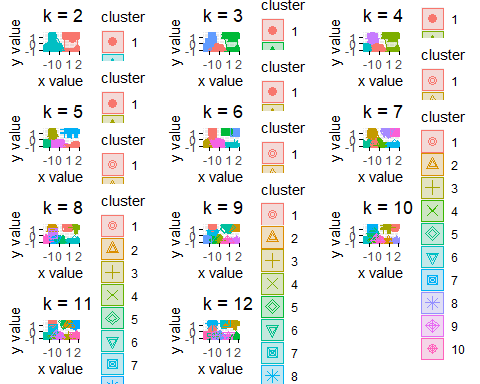
# Plots Comparision

#Ref Source -6,3,7

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

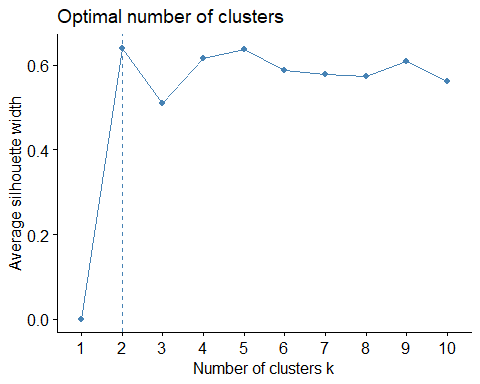
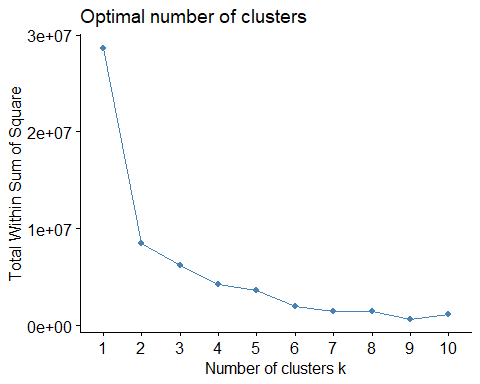
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

## Warning: package 'gridExtra' was built under R version 4.0.2



## Average Silhouette Method

#Ref Source -6,3,7



## 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>