Karthikeyan Chellamuthu 11.2.2 Exercise Clustering

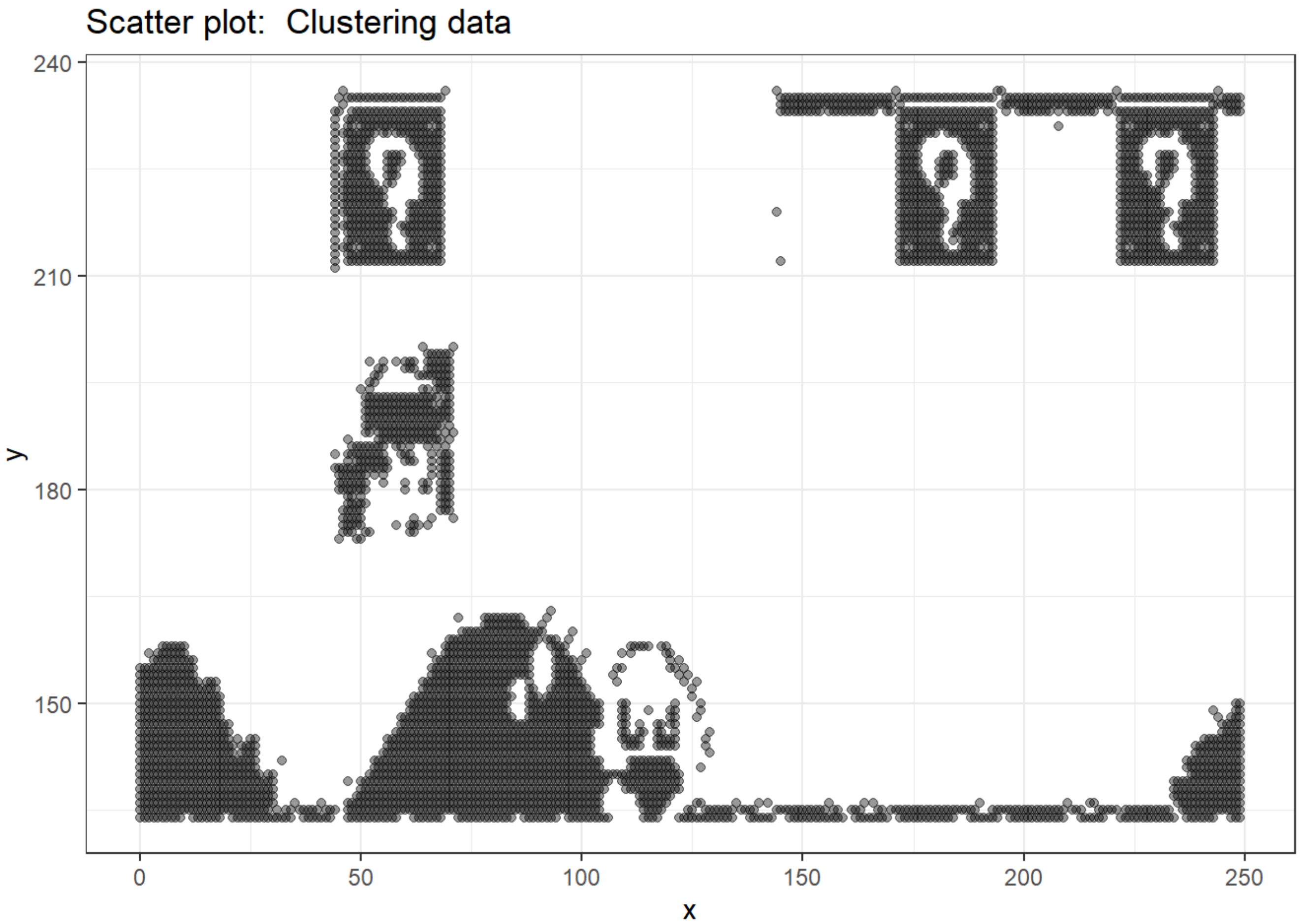
Karthikeyan Chellamuthu

03/05/2022

Structure of Clustering data

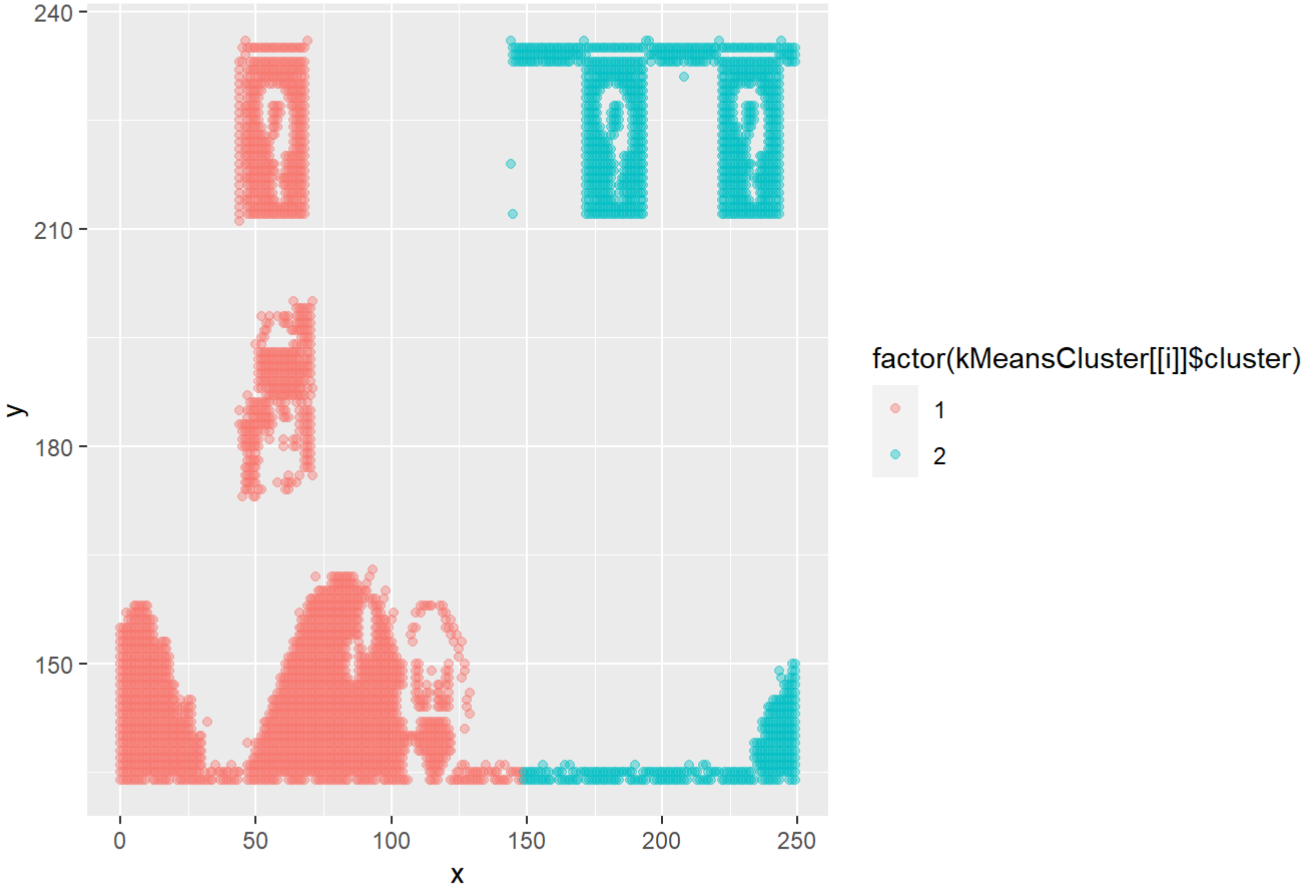
```
## 'data.frame':    4022 obs. of  2 variables:
##  $ x: int  46 69 144 171 194 195 221 244 45 47 ...
##  $ y: int  236 236 236 236 236 236 236 236 235 235 ...
```

9.3.a Plot the dataset using a scatter plot.

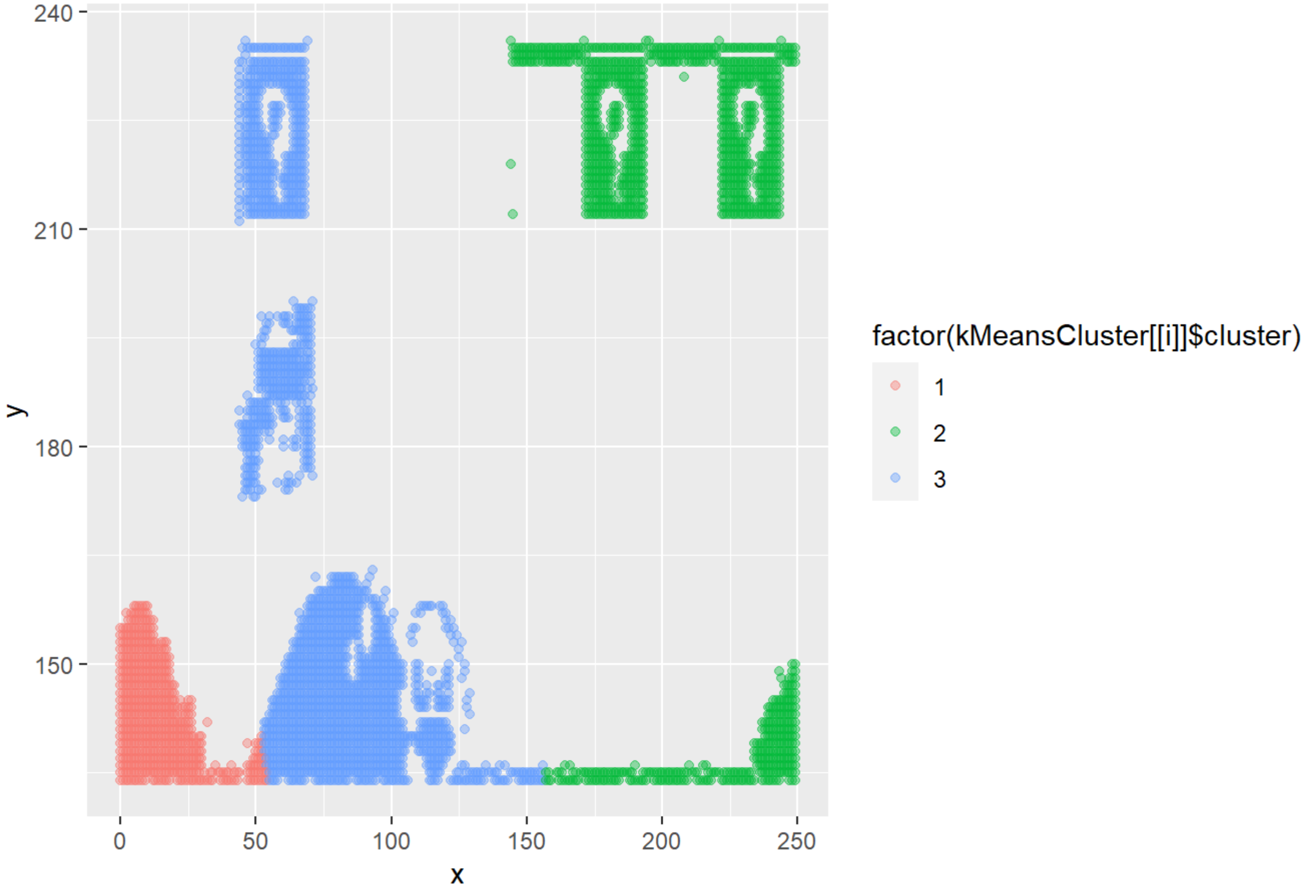


9.3.b Fit the dataset using the k-means algorithm from k=2 to k=12. Create a scatter plot of the resultant clusters for each value of k.

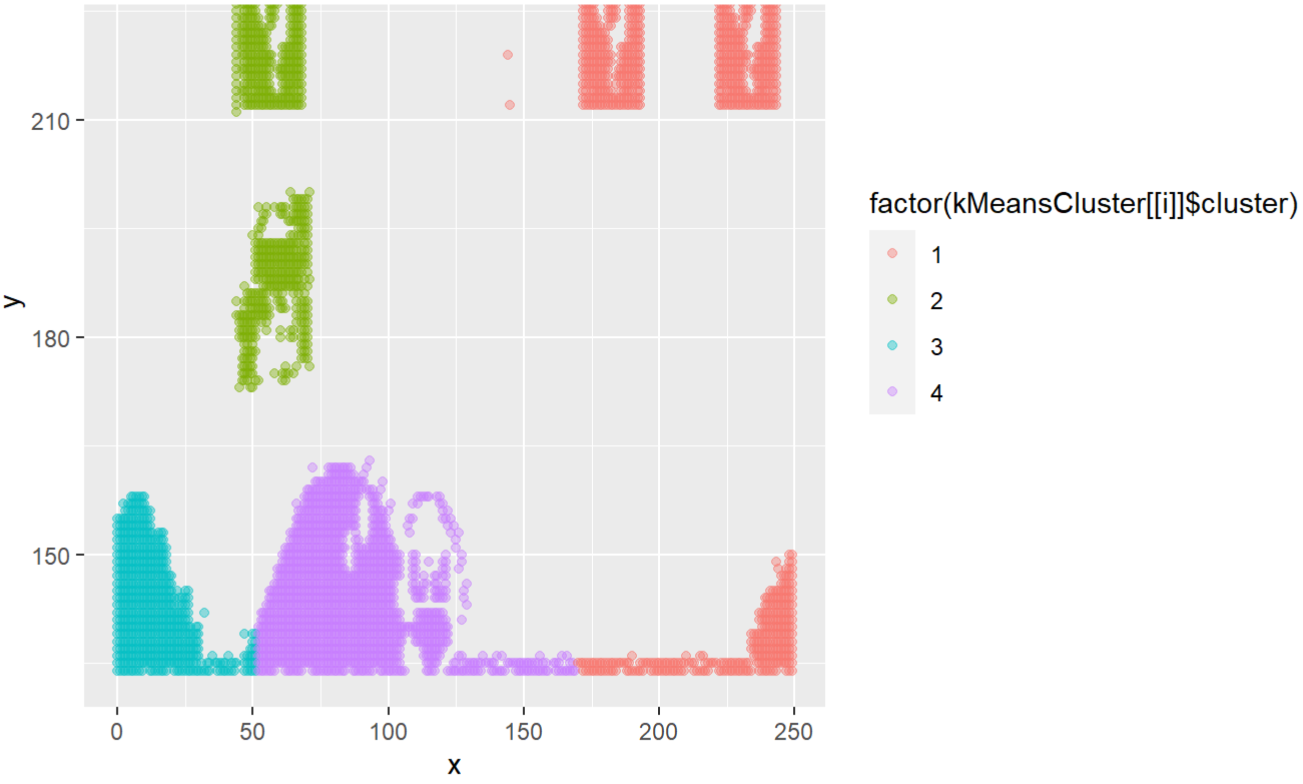
kMeans Cluster Plot: Clustering data - 2 centroids



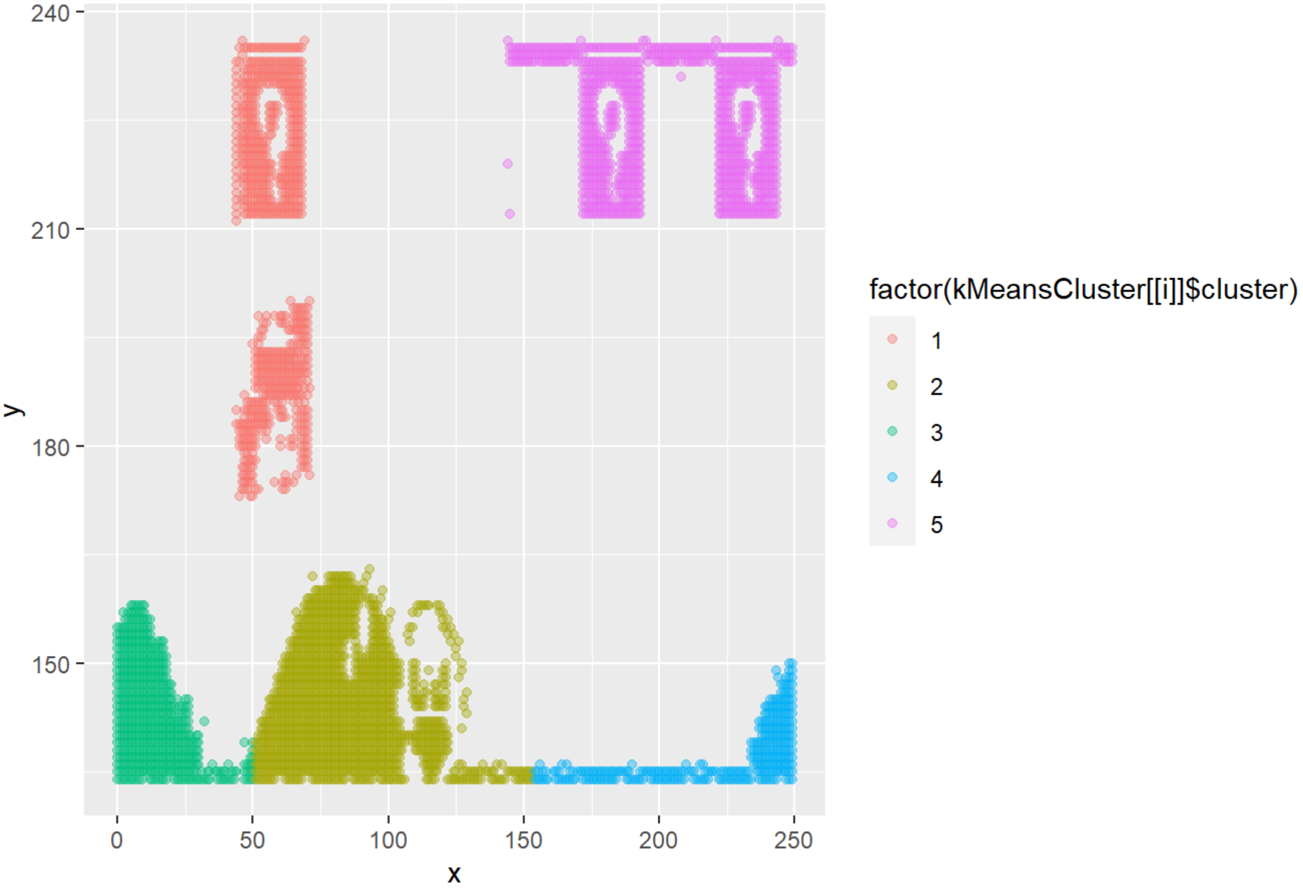
kMeans Cluster Plot: Clustering data - 3 centroids

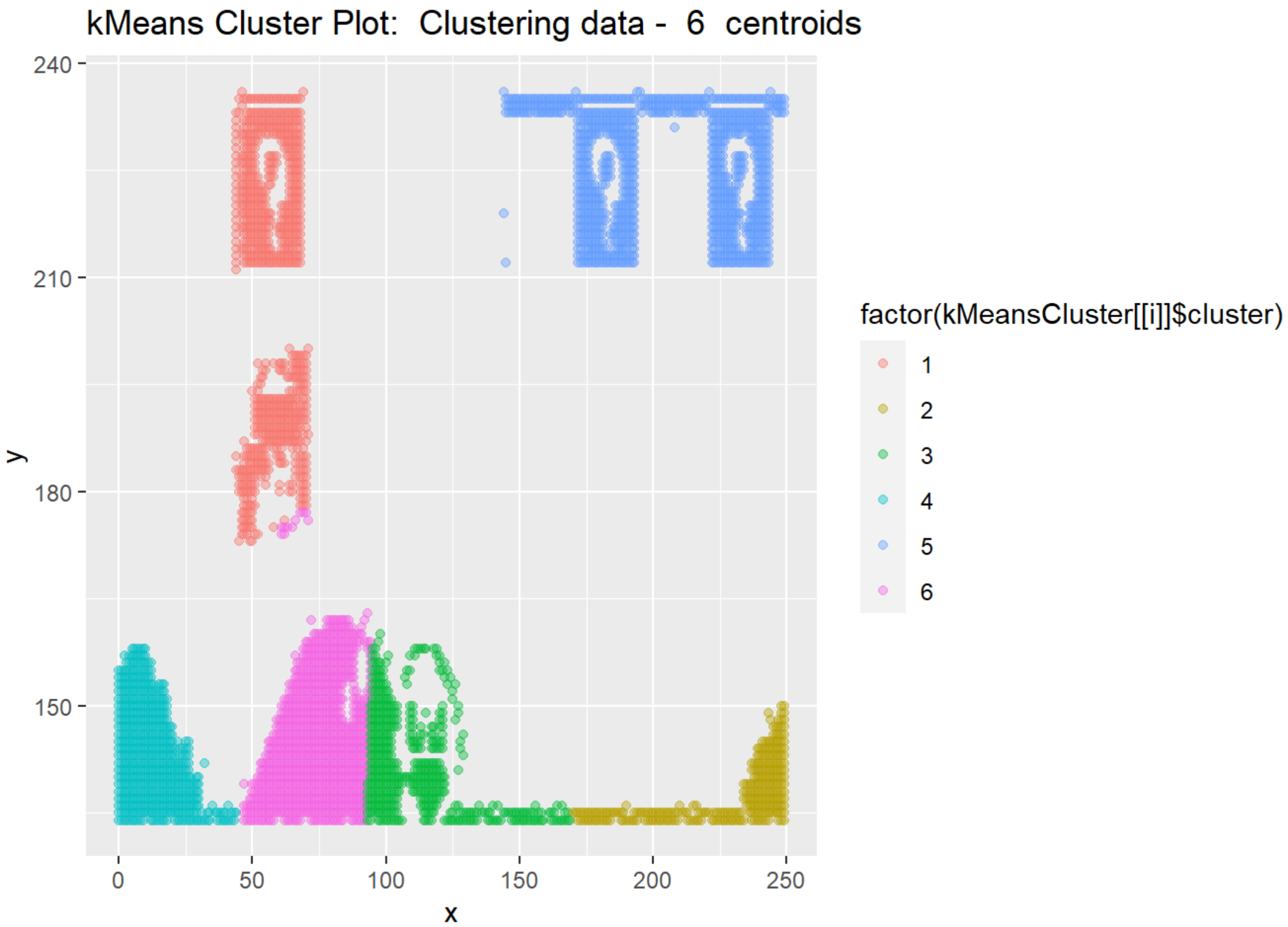


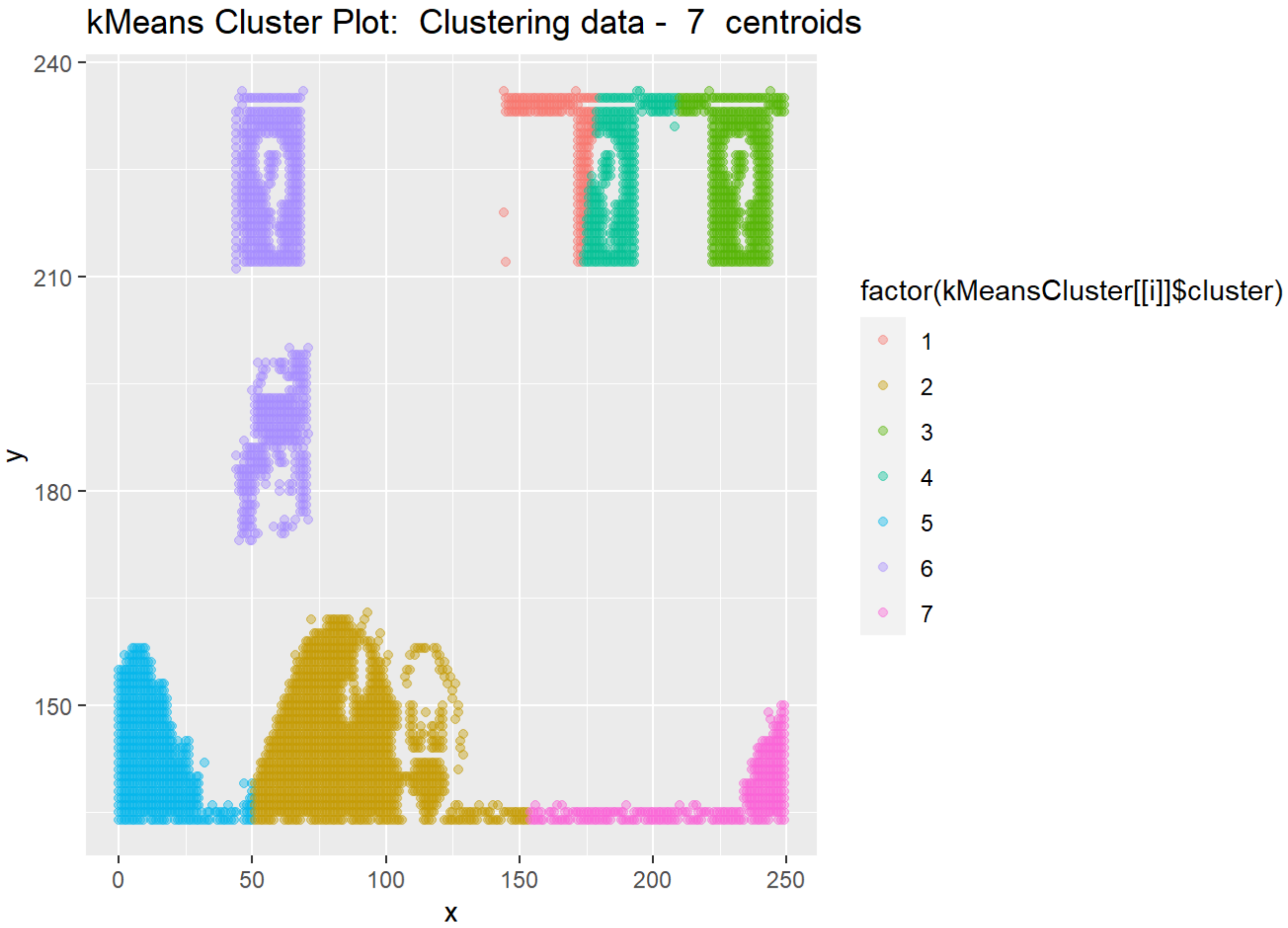
kMeans Cluster Plot: Clustering data - 4 centroids

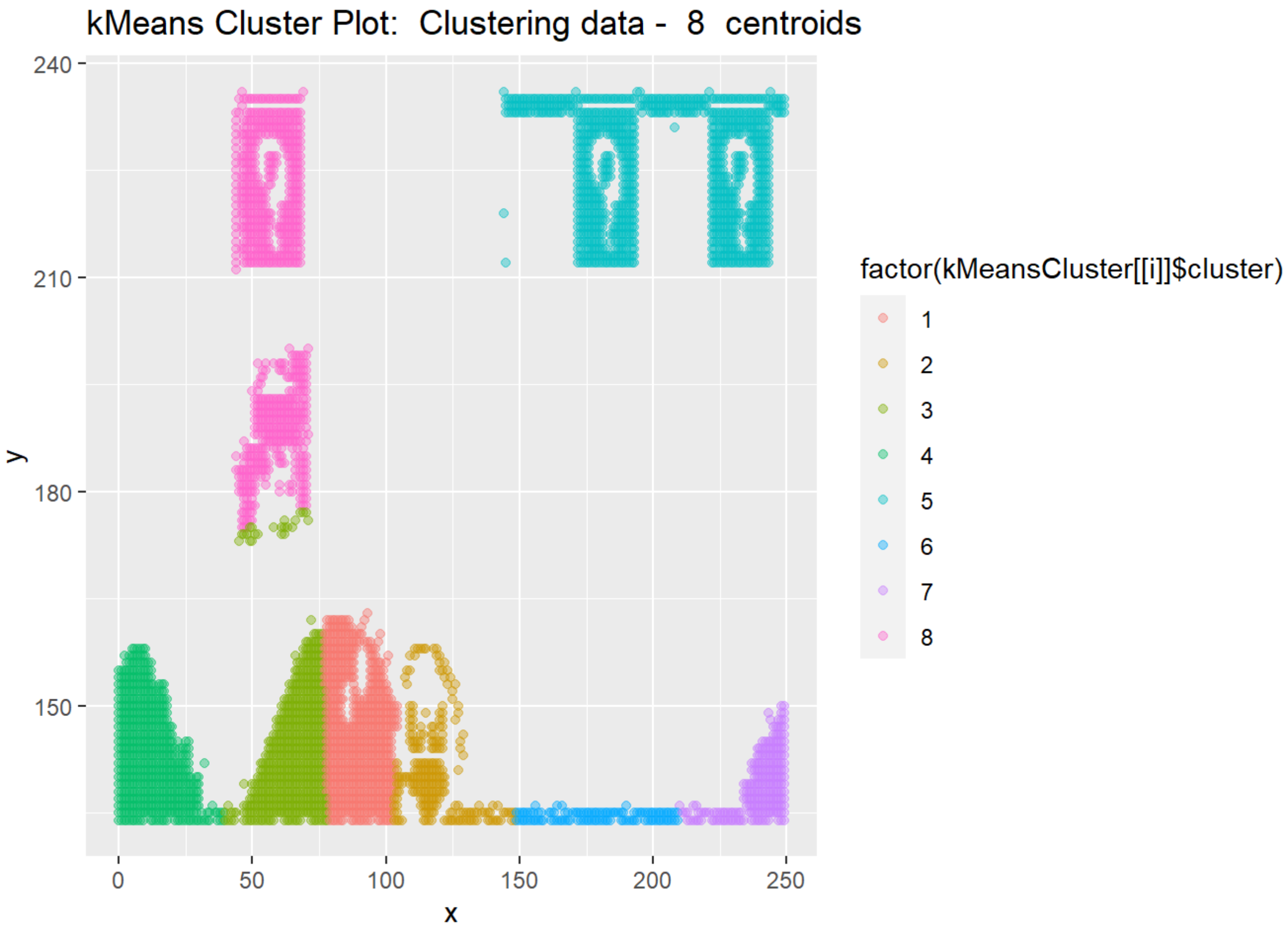


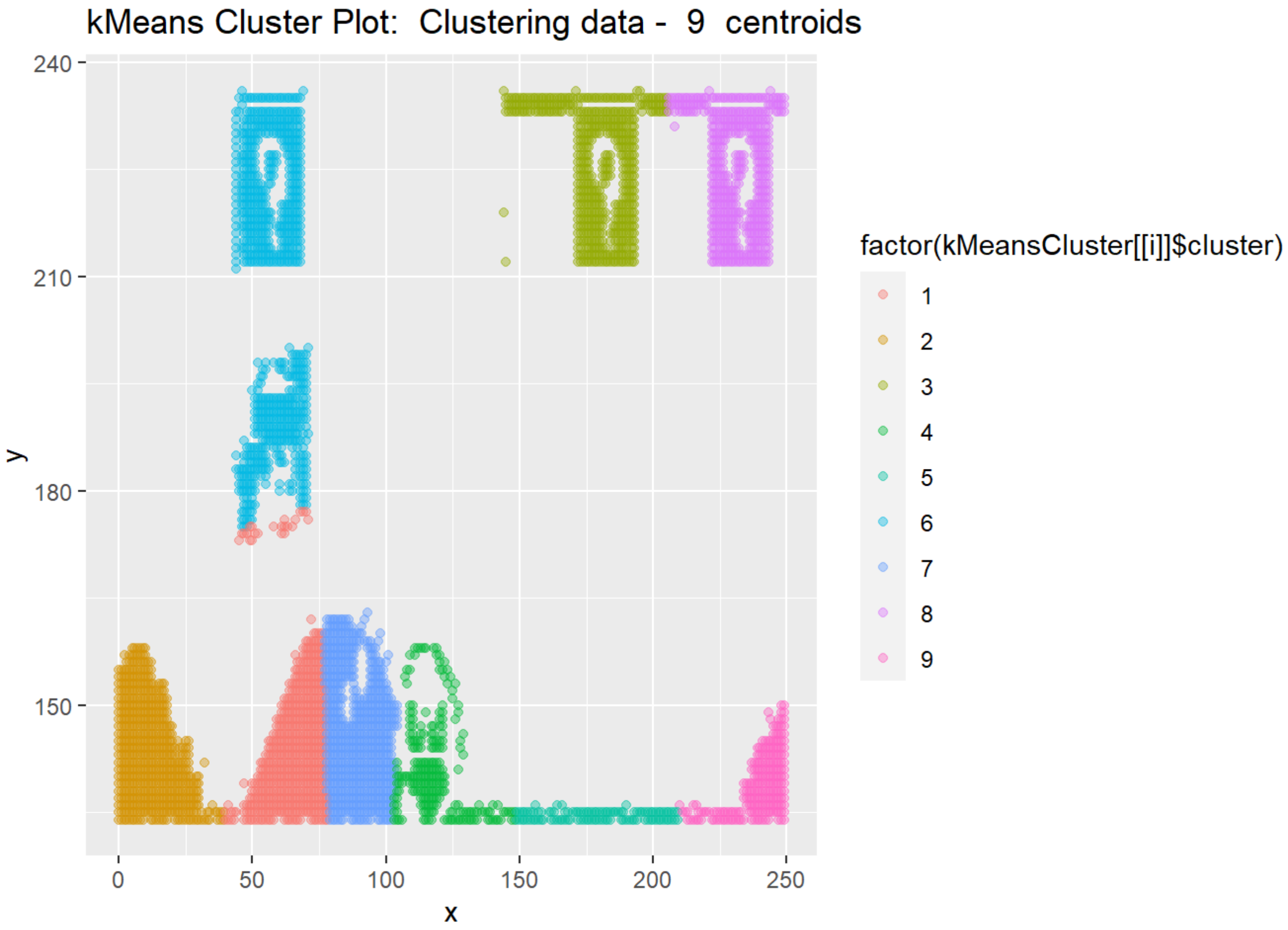
kMeans Cluster Plot: Clustering data - 5 centroids

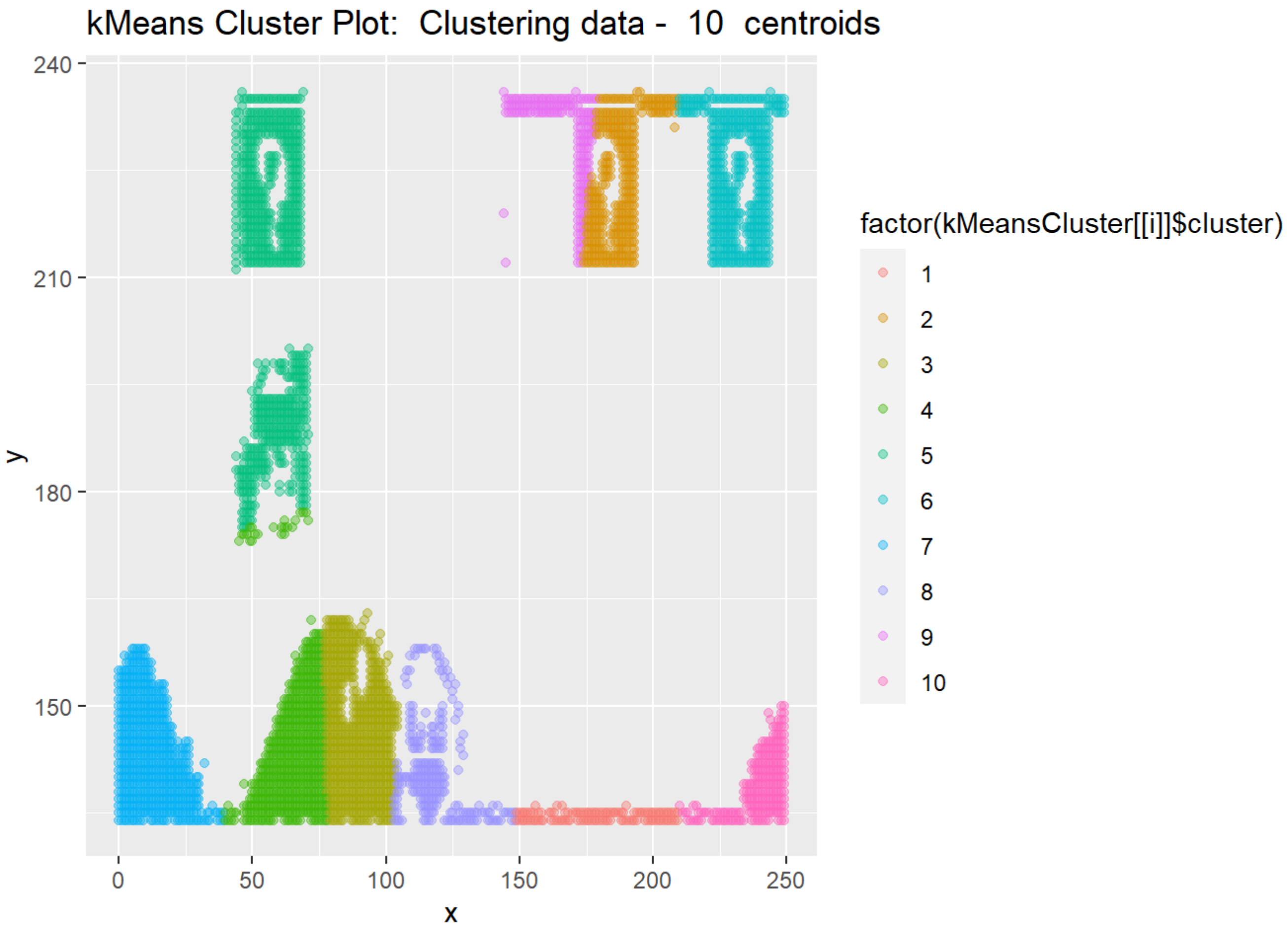


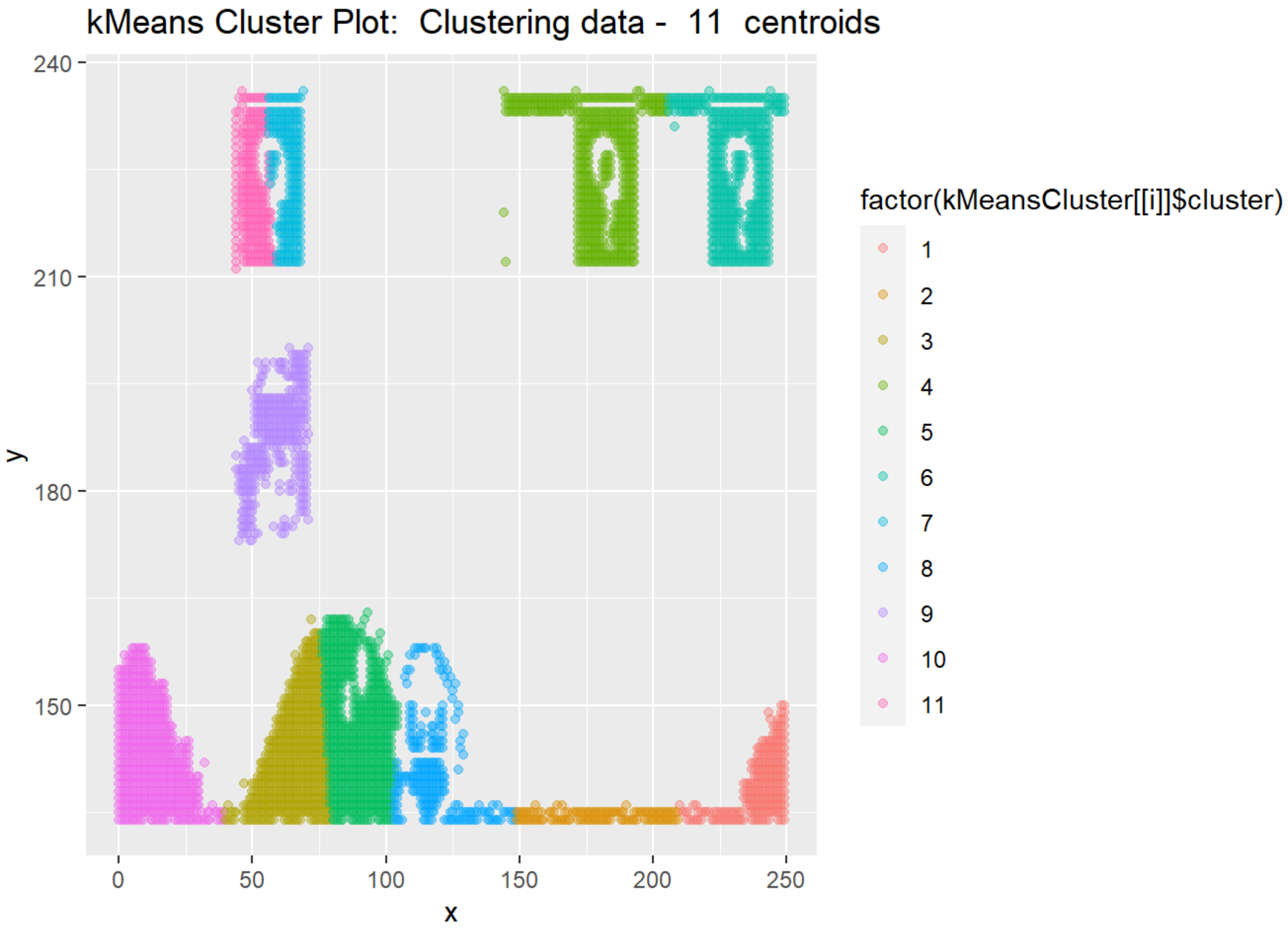

kMeans Cluster Plot: Clustering data - 6 centroids
240
210
180
150
y
0
50
100
150
200
250
x
factor(kMeansCluster[[i]]$cluster)
1
2
3
4
5
6

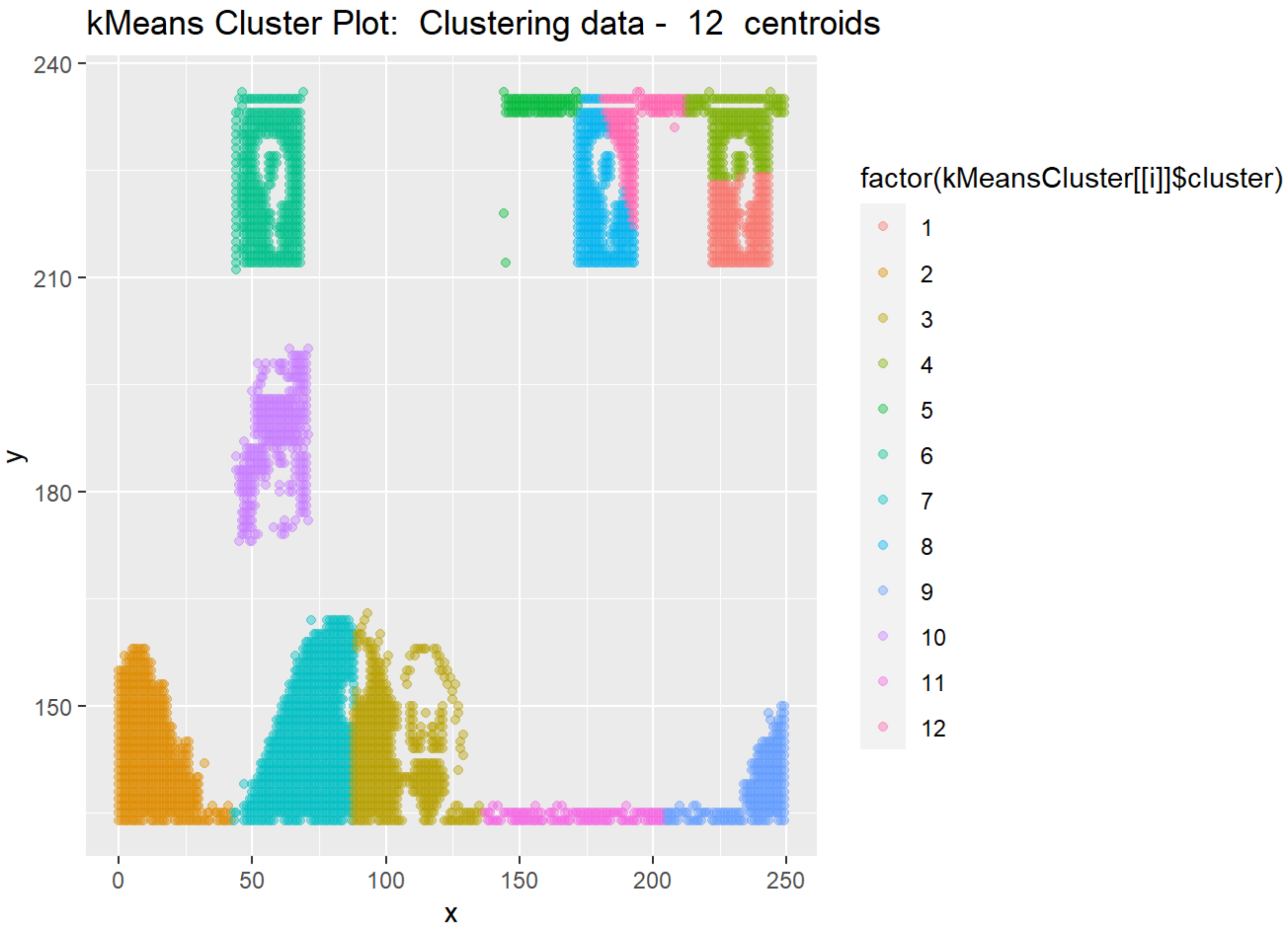

kMeans Cluster Plot: Clustering data - 7 centroids
240
210
180
150
y
0
50
100
150
200
250
x
factor(kMeansCluster[[i]]$cluster)
1
2
3
4
5
6
7


kMeans Cluster Plot: Clustering data - 8 centroids
y
240
210
180
150
x
0
50
100
150
200
250
factor(kMeansCluster[[i]]$cluster)
1
2
3
4
5
6
7
8

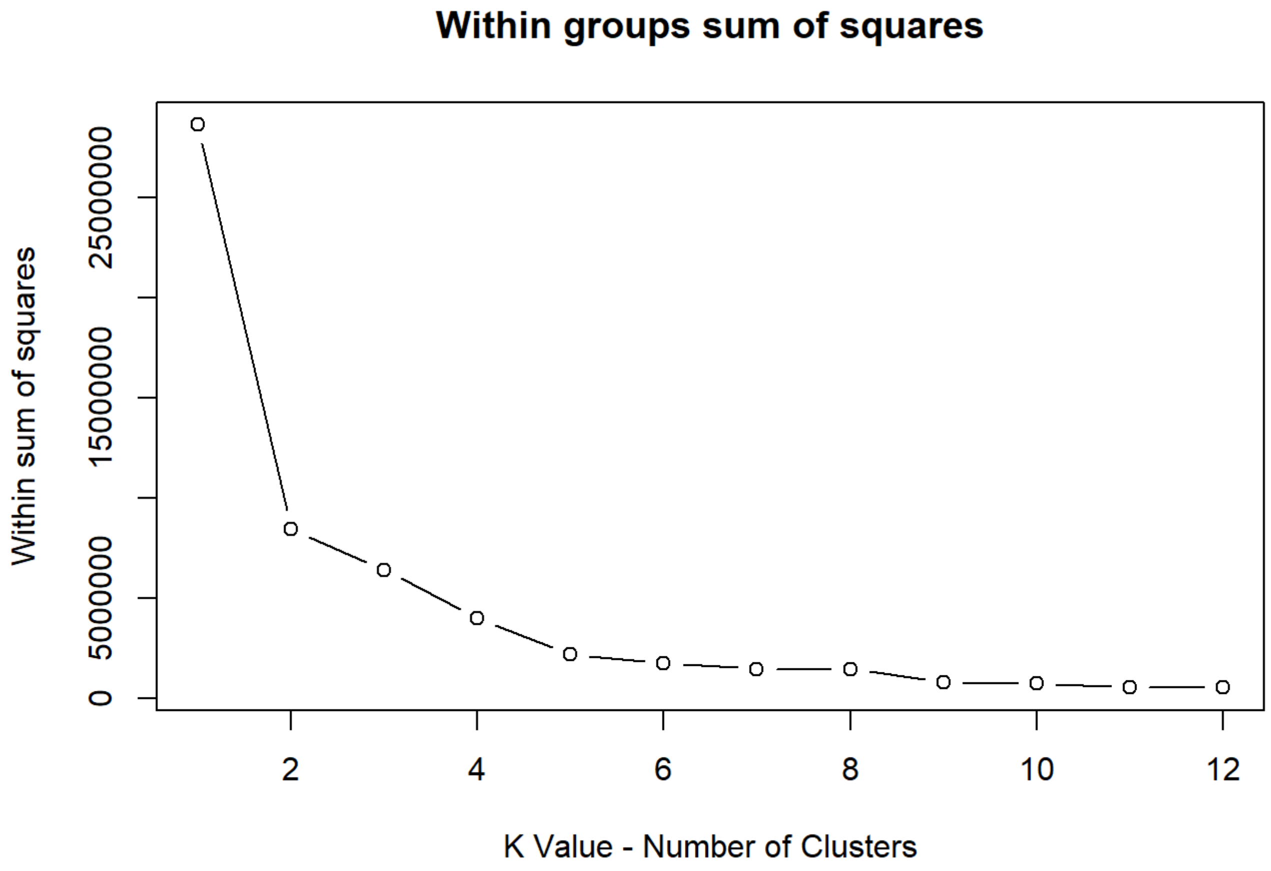

kMeans Cluster Plot: Clustering data - 9 centroids
y
240
210
180
150
x
0
50
100
150
200
250
factor(kMeansCluster[[i]]$cluster)
1
2
3
4
5
6
7
8
9


kMeans Cluster Plot: Clustering data - 10 centroids
y
240
210
180
150
x
0
50
100
150
200
250
factor(kMeansCluster[[i]]$cluster)
1
2
3
4
5
6
7
8
9
10


kMeans Cluster Plot: Clustering data - 11 centroids
y
240
210
180
150
x
0
50
100
150
200
250
factor(kMeansCluster[[i]]$cluster)
1
2
3
4
5
6
7
8
9
10
11



9.3.c 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.



9.3.d 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?

For k=5, between_ss/total_ss ratio tends to change slowly and remain less changing as compared to other k’s. Hence data k=5 should be a good choice for number of clusters.