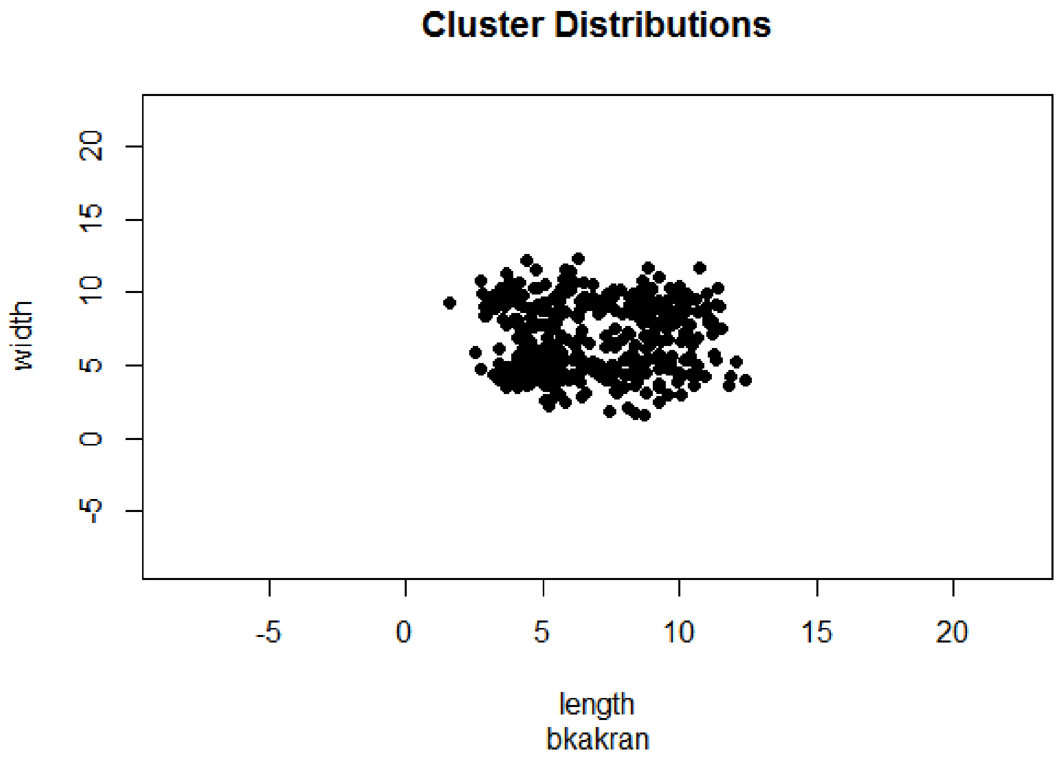
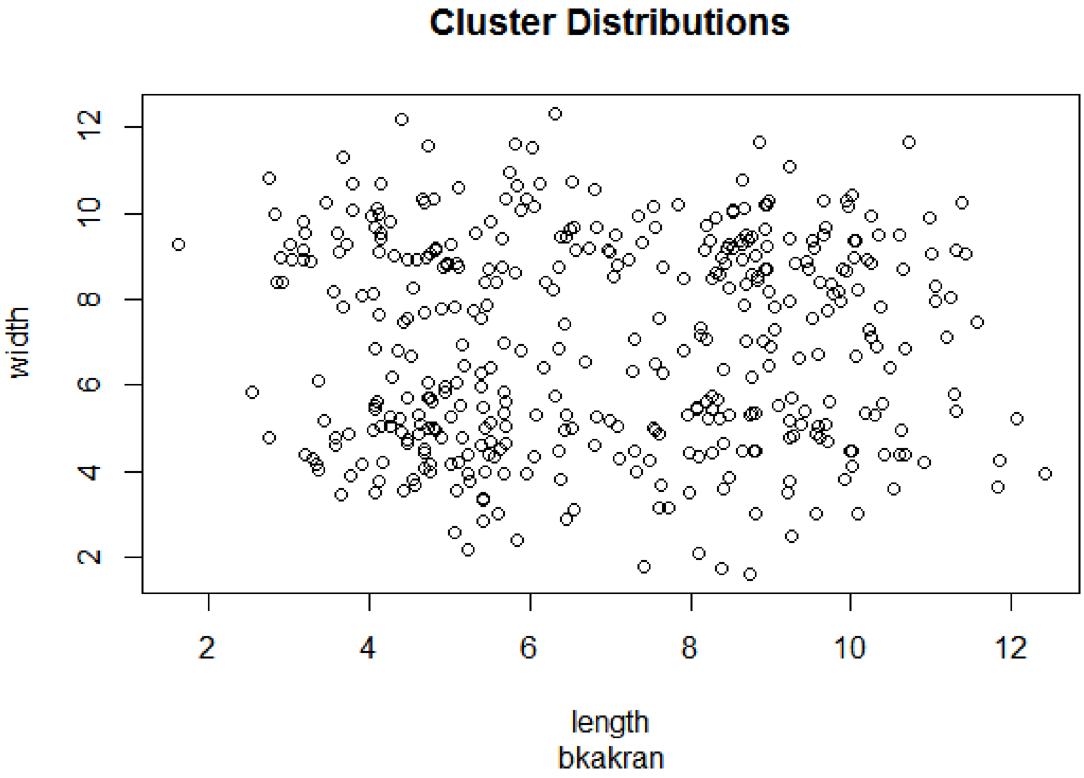
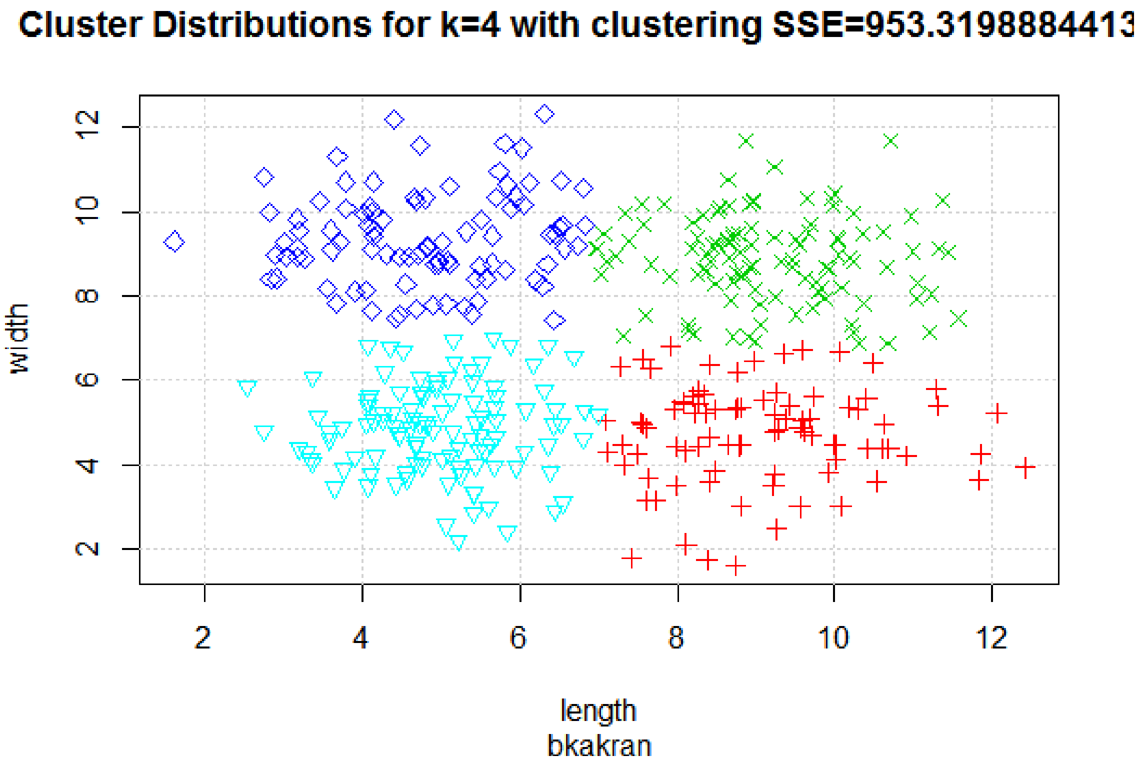
(1) Plot the given data (single color – as you haven't clustered yet!), example is shown below



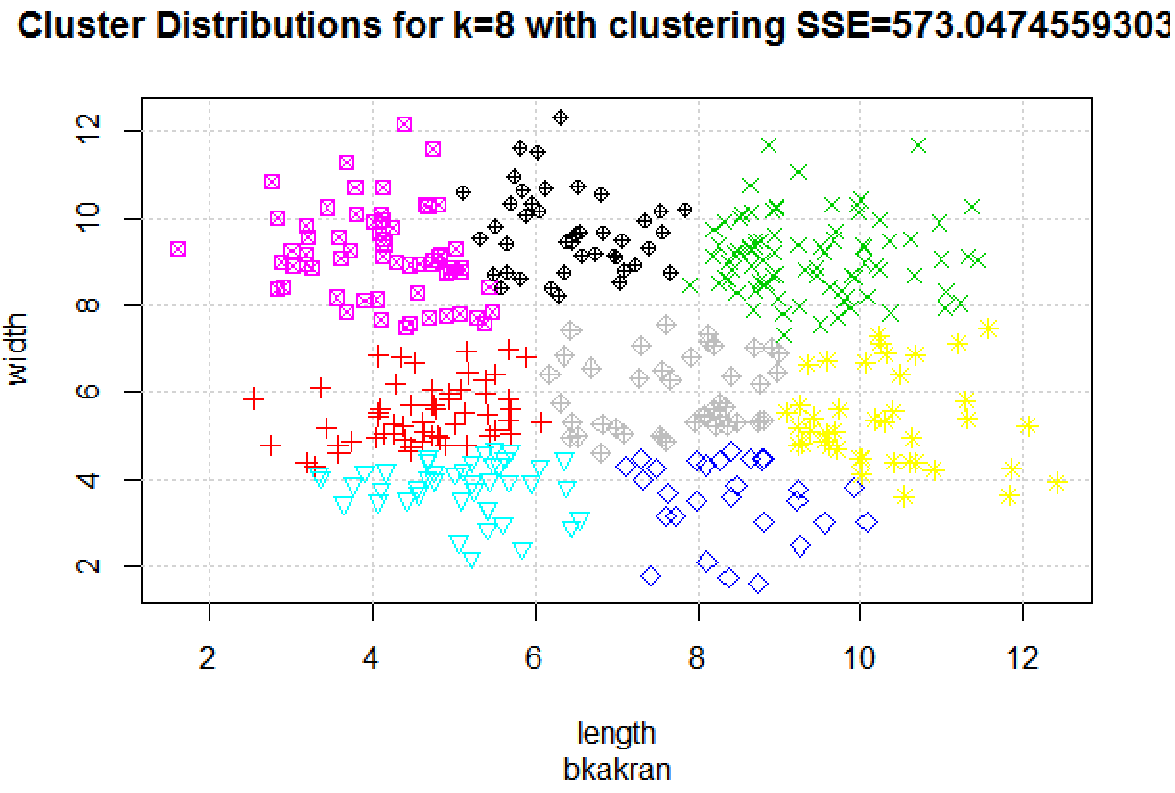


(2) Apply your own K-means clustering, with K=4 and K=8. Generate clustering plots (see examples below), use random colors or symbols to distinguish clusters. Comment on which clustering result is better (just by visual comparison).

```
data = read.table(file = "C:/Users/pejakalabhargava/Desktop/d-c4hw2.csv", header=T, sep=",")
data = data[,-c(3)]
result = myKMeans(data,4)
SSE = result$SSE;
clusteringSSE = colSums(SSE)
result = result$result
x1 = min(data[,1])
x2 = max(data[,1])
y1 = min(data[,2])
y2 = max(data[,2])
du.cl1 = subset(result, result$clusters == 1)
du.cl2 = subset(result, result$clusters == 2)
du.cl3 = subset(result, result$clusters == 3)
du.cl4 = subset(result, result$clusters == 4)
title= paste("Cluster Distributions for k=4 with clustering SSE", clusteringSSE, sep="=")
plot(0, col=c(1), main = title,xlim=c(x1,x2), ylim=c(y1,y2),sub="bkakran", xlab = "length", ylab = "width")
grid()
i=1
points(du.cl1, col=c(i+1), pch=c(2+i))
points(du.cl2, col=c(i+2), pch=c(3+i))
points(du.cl3, col=c(i+3), pch=c(4+i))
points(du.cl4, col=c(i+4), pch=c(5+i))
```



```
data = read.table(file = "C:/Users/pejakalabhargava/Desktop/d-c4hw2.csv", header=T, sep=",")
data = data[,-c(3)]
result  = myKMeans(data,8)
SSE = result$SSE;
clusteringSSE = colSums(SSE)
result = result$result
x1 = min(data[,1])
x2 = max(data[,1])
y1 = min(data[,2])
y2 = max(data[,2])
du.cl1 = subset(result, result$clusters == 1)
du.cl2 = subset(result, result$clusters == 2)
du.cl3 = subset(result, result$clusters == 3)
du.cl4 = subset(result, result$clusters == 4)
du.cl5 = subset(result, result$clusters == 5)
du.cl6 = subset(result, result$clusters == 6)
du.cl7 = subset(result, result$clusters == 7)
du.cl8 = subset(result, result$clusters == 8)
title= paste("Cluster Distributions for k=8 with clustering SSE", clusteringSSE, sep="=")
plot(0, col=c(1), main = title,xlim=c(x1,x2), ylim=c(y1,y2),sub="bkakran", xlab = "length", ylab = "width")
grid()
i=1
points(du.cl1, col=c(i+1), pch=c(2+i))
points(du.cl2, col=c(i+2), pch=c(3+i))
points(du.cl3, col=c(i+3), pch=c(4+i))
points(du.cl4, col=c(i+4), pch=c(5+i))
points(du.cl5, col=c(i+5), pch=c(6+i))
points(du.cl6, col=c(i+6), pch=c(7+i))
points(du.cl7, col=c(i+7), pch=c(8+i))
points(du.cl8, col=c(i+8), pch=c(9+i))
```



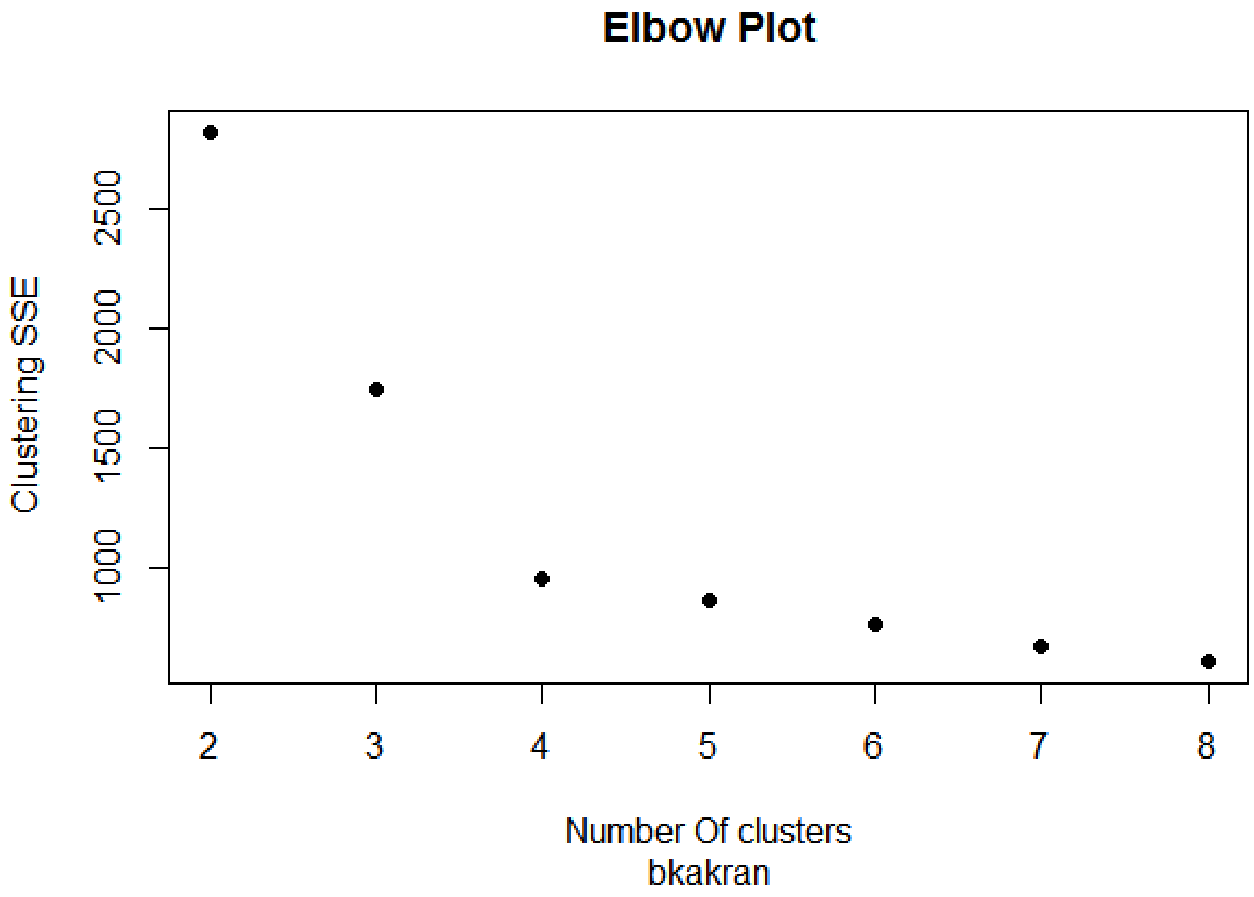
Results:

We see that when k=4 it forms natural good clusters. K=8 will be the case when we are splitting natural sub clusters. Visually looking k=4 looks better since each cluster has globular structure and looks natural. If we compare purely based on SSE Clusters with k=8 is better since it has a lesser clustering SSE as calculated and clearly seen from the plot. With k=8 all the points are more close to its cluster centroids hence making it a better fit as compared to clusters with k=4 if we compare based on SSE.

(3) Apply your own bisecting k-means on the given data for K=2 to 8, and answer the following sub-questions

a. Plot clustering SEE (y-axis) against number of clusters (x-axis). (e.g., is shown below)

```
data = read.table(file = "C:/Users/pejakalabhargava/Desktop/d-c4hw2.csv", header=T, sep=",")
colNames=c('Number Of clusters','Clustering SSE')
SSEmat = matrix(nrow=7,ncol=2,0)
colnames(SSEmat) <- colNames
for(i in 2:8) {
        result = myBKMeans(data,i)
        SSE = result$SSE;
        clusteringSSE = colSums(SSE)
        SSEmat[i-1,1] = i
        SSEmat[i-1,2] = clusteringSSE
}
plot(SSEmat, col=c(1), main = "Elbow Plot", pch=c(19), sub="bkakran")
```

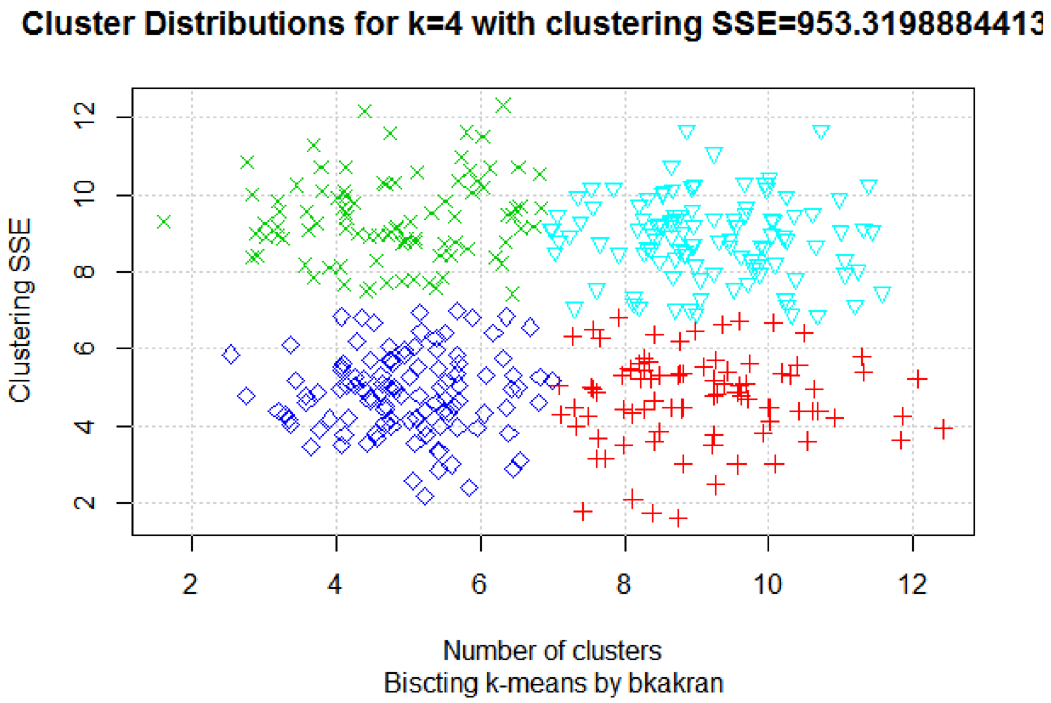


b. From the plot identify optimal K

The number of natural clusters in the data set are estimated by looking for the number of clusters at which there is a knee, peak or dip in the plot of measure when it is plotted against the number of clusters. From the plot above it is clear that there is a dip at k=4, which means **optimal k is 4.**

c. For identified optimal K, show clustering plot

```
data = read.table(file = "C:/Users/pejakalabhargava/Desktop/d-c4hw2.csv", header=T, sep=",")
data = data[,-c(3)]
result = myBKMeans(data,4)
SSE = result$SSE;
clusteringSSE = colSums(SSE)
result = result$result
x1 = min(data[,1])
x2 = max(data[,1])
y1 = min(data[,2])
y2 = max(data[,2])
du.cl1 = subset(result, result$clusters == 1)
du.cl2 = subset(result, result$clusters == 2)
du.cl3 = subset(result, result$clusters == 3)
du.cl4 = subset(result, result$clusters == 4)
title= paste("Cluster Distributions for k=4 with clustering SSE", clusteringSSE, sep="=")
plot(0, col=c(1), main = title,xlim=c(x1,x2), ylim=c(y1,y2),sub="Biscting k-means by bkakran", xlab = "Number of clusters", ylab = "Clustering SSE")
grid()
i=1
points(du.cl1, col=c(i+1), pch=c(2+i))
points(du.cl2, col=c(i+2), pch=c(3+i))
points(du.cl3, col=c(i+3), pch=c(4+i))
points(du.cl4, col=c(i+4), pch=c(5+i))
```


Cluster Distributions for k=4 with clustering SSE=953.3198884413
Clustering SSE
Number of clusters
Biscting k-means by bkakran