Q1: For nstart being 1

>R kmeans(clusterdata1, centers=3, nstart=1)

1st:

```
K-means clustering with 3 clusters of sizes 22, 100, 28

Cluster means:
          V1          V2
1 -0.9676830  0.00352443
2  5.7968425  2.90795207
3  0.7219445 -0.03544639

Clustering vector:
  [1] 1 3 3 3 3 1 3 3 1 1 1 3 3 3 3 3 3 3 1 3 3 3 1 3 1 1 3 3 1 1 1 1 1 1 1 1 3 1 3 3 3 3 1 3 3 1 1
 [48] 3 1 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [95] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[142] 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1]   24.59478 1239.93075   25.34034
 (between_SS / total_SS =  52.9 %)
```

2nd:

```
K-means clustering with 3 clusters of sizes 100, 22, 28

Cluster means:
          V1          V2
1  5.7968425  2.90795207
2 -0.9676830  0.00352443
3  0.7219445 -0.03544639

Clustering vector:
  [1] 2 3 3 3 3 2 3 3 2 2 2 3 3 3 3 3 3 3 2 3 3 3 2 3 2 2 3 3 2 2 2 2 2 2 2 2 3 2 3 3 3 3 2 3 3 2 2
 [48] 3 2 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [95] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[142] 1 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 1239.93075   24.59478   25.34034
 (between_SS / total_SS =  52.9 %)
```

3rd:

```
K-means clustering with 3 clusters of sizes 50, 70, 30

Cluster means:
           V1          V2
1 -0.02149164 -0.01829923
2  5.83669823  0.72313385
3  5.70384586  8.00586126

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [48] 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [95] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[142] 3 3 3 3 3 3 3 3 3

Within cluster sum of squares by cluster:
[1] 85.12547 46.91125 78.84837
 (between_SS / total_SS =  92.3 %)
```

4th:

```
K-means clustering with 3 clusters of sizes 50, 30, 70

Cluster means:
           V1          V2
1 -0.02149164 -0.01829923
2  5.70384586  8.00586126
3  5.83669823  0.72313385

Clustering vector:
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [48] 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [95] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[142] 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 85.12547 78.84837 46.91125
 (between_SS / total_SS =  92.3 %)
```

5th:

```
K-means clustering with 3 clusters of sizes 70, 50, 30

Cluster means:
           V1          V2
1  5.83669823  0.72313385
2 -0.02149164 -0.01829923
3  5.70384586  8.00586126

Clustering vector:
  [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [48] 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [95] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[142] 3 3 3 3 3 3 3 3 3

Within cluster sum of squares by cluster:
[1] 46.91125 85.12547 78.84837
 (between_SS / total_SS =  92.3 %)
```

For nstart = 1:

The results for 'K-means clustering with 3 clusters of sizes' & 'within cluster sum of squares by cluster' and 'between_SS / total_SS' have changed over times. This may be due to the changed of the centroids. So that the stability builds up as we runs more times of trying.

For nstart=100:

Within cluster sum of squares are unchanged for 10 times and the between_SS / total_SS remain stable at 92.3%

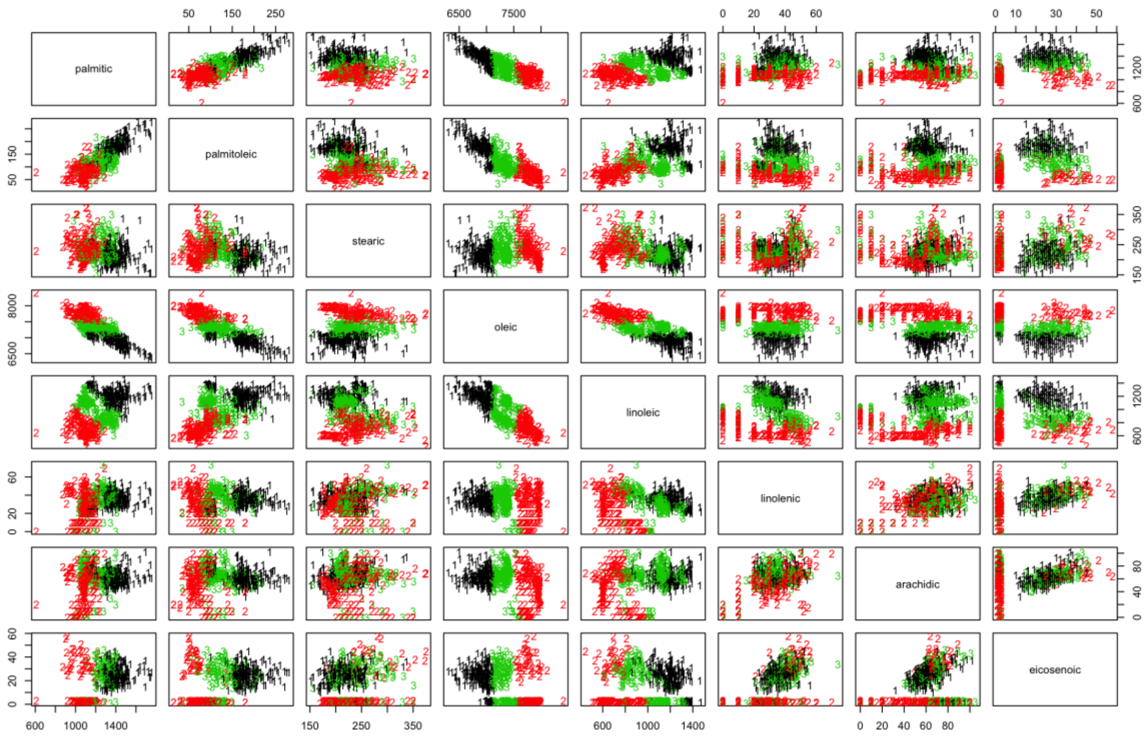
Q2:

```
> R:
k3<-kmeans(olive, centers = 3, nstart = 100)
plot(olive, col=k3$cluster,pch=clusym[k3$cluster])

### After scale
solive<- scale(olive)
pairs(solive, cex=0.3)
sk3<- kmeans(solive, 3, 100)
plot(olive, col=sk3$cluster,pch=clusym[sk3$cluster])
table(sk3$cluster, oliveoil$macro.area)

sk9 <- kmeans(solive, 9, 100)
plot(solive, col=sk9$cluster,pch=clusym[sk9$cluster])
table(sk9$cluster, oliveoil$macro.area)
```

With K=3: before scale



After scale:

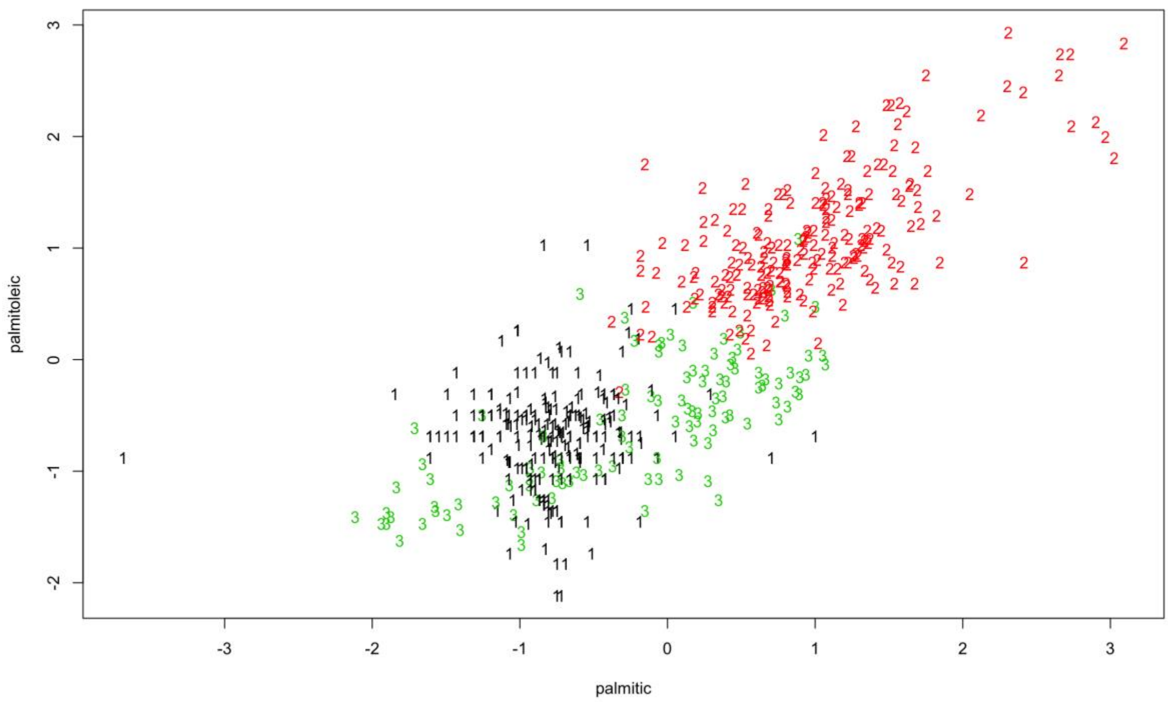


table:

	South	Sardinia	Centre.North
1	4	97	142
2	217	1	0
3	102	0	9

Looks not bad…
I don’t see quiet much how the criteria of how well this clustering is … I could see that South has been allocated to two major parts of clusters 2 and 3, and Centre.North and Sardinia are allocated to 1.

k=9

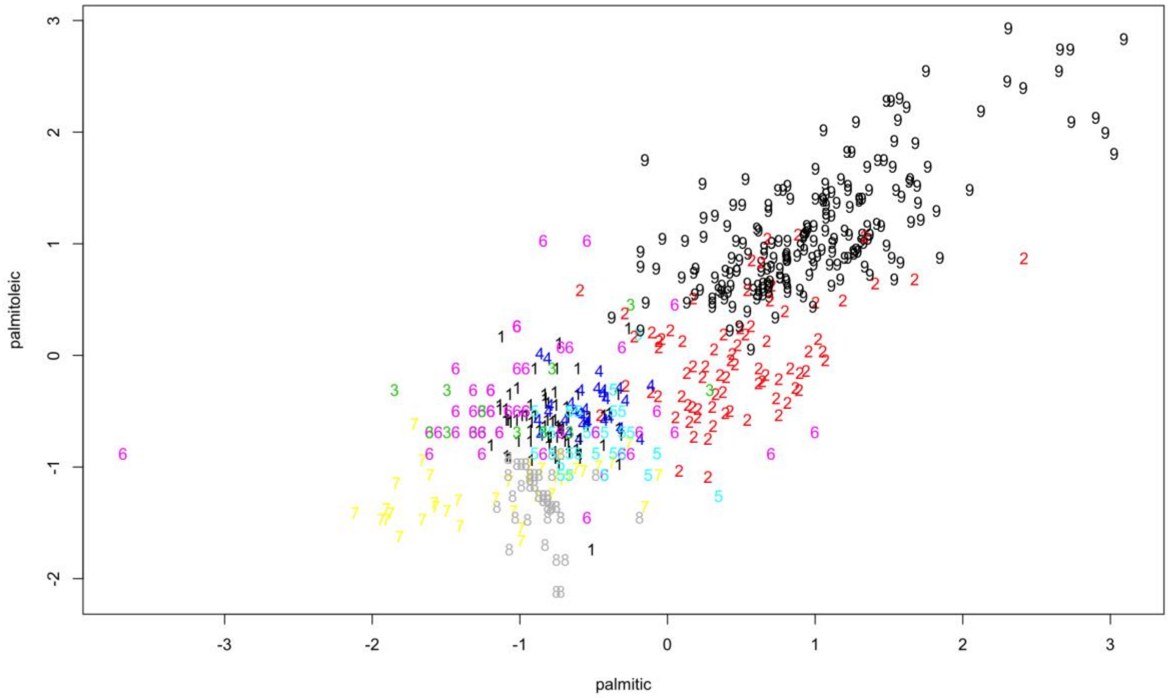


table:

	South	Sardinia	Centre.North
1	1	68	0
2	79	0	0
3	0	0	15
4	0	30	0
5	1	0	35
6	0	0	48
7	38	0	0
8	2	0	53
9	202	0	0

I think this is just an expand of K=3. But at least this trial makes Sardinia and Centre.North in different clusters. So that they are sperated when k=9.

Q3:
Multiplying the variables by the same constant q won't change the variation of the data. Therefore, the Euclidean distances between each xi and the centroids $m_{i...k}^{km}$ will not change. Thus, the K-means clustering of D which is defined as choosing the ms and cs to minimise:

$$S(\mathcal{C}, \mathbf{m}_1, \ldots, \mathbf{m}_K) = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}_{c(i)}\|^2.$$

will not change.

However, multiplying the variables by a same constant does not affect the variation but the value of the data. The new centroids are affected and are multiplied by the q.