Homework 10

STAT 4510/7510

Instructions: Please list your name and student number clearly. In order to receive credit for a problem, your solution must show sufficient detail so that the grader can determine how you obtained your answer.

Submit a single pdf generated using R Markdown. All R code should be included, as well as all output produced. Upload your work to the Canvas course site.

We will once again revisit the `tumor` dataset from Chapter 4. Recall that the rows of `tumor.csv` indicate if the tumor was deemed to be malignant or benign, along with information which describes the characteristics of the tumor.

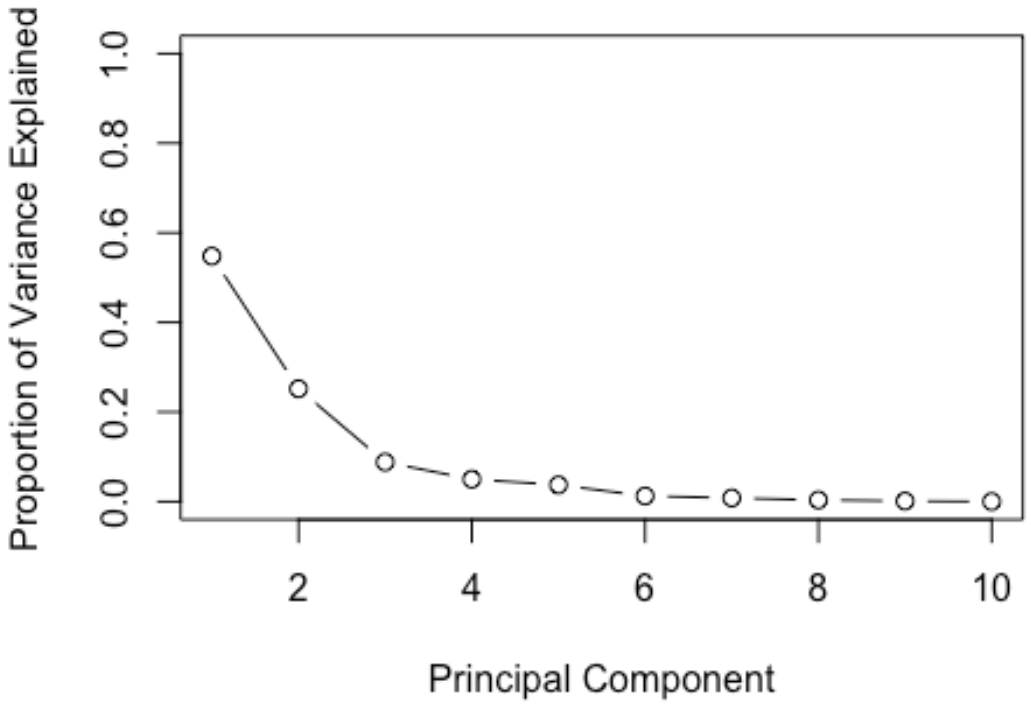
Problem 1

a) Read the data in and remove the first column, which corresponds to the diagnosis. We now have a dataset with only X variables.

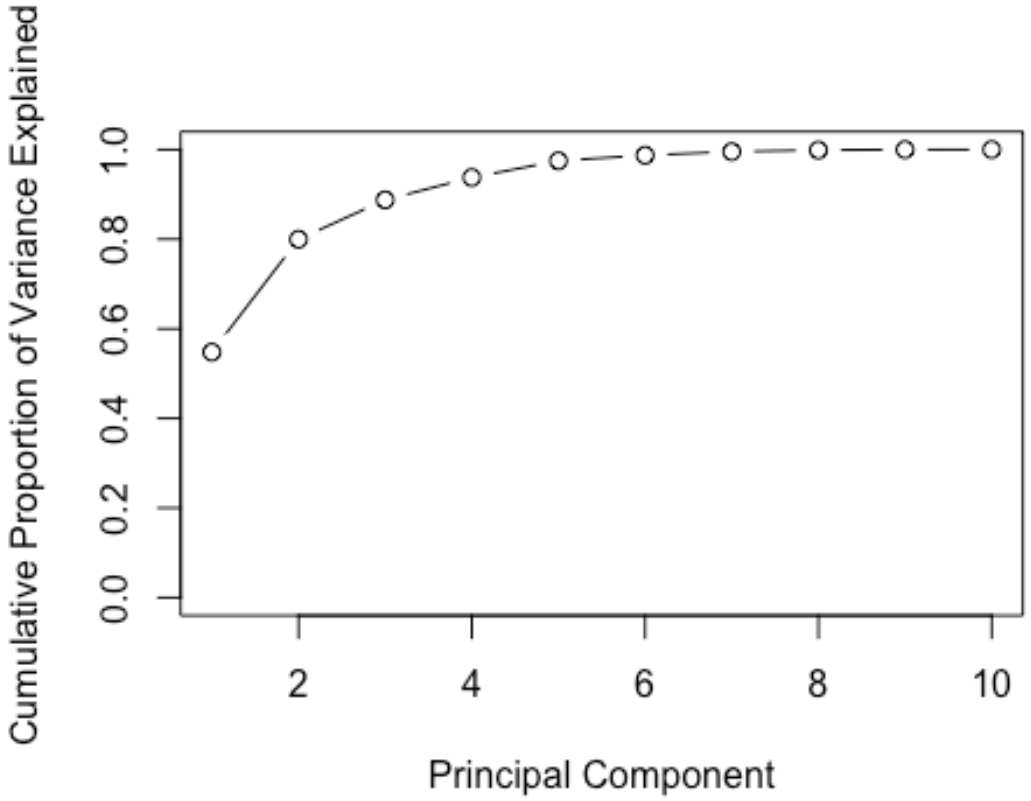
```
set.seed(1)
data<-read.csv("tumor.csv")
data<-data[,-1]
```

b) Conduct a PCA analysis, being sure to center and scale the variables. Provide a plot of the variation explained by each PC. How many principal components do you think should be kept?

```
set.seed(1)
pca<-prcomp(data, scale = TRUE, center = TRUE)
pca_var<-pca$sdev^2
pve<-pca_var/sum(pca_var)
plot(pve, xlab = "Principal Component",
ylab = "Proportion of Variance Explained", ylim = c(0, 1),
type = "b")
```




```
plot(cumsum(pve), xlab = "Principal Component",
ylab = "Cumulative Proportion of Variance Explained", ylim = c(0, 1), type =
"b")
```



According to the plot, we can take 4 principal components if the threshold is 90% of Proportion of Variance Explained.

c) How much of the total variation in the data is explained by 2 principal components? By 3?

```
print(cumsum(pve))
```

```
##  [1] 0.5478588 0.7997302 0.8877917 0.9376926 0.9749465 0.9873607 0.9953692
##  [8] 0.9988582 0.9999718 1.0000000
```

According to the result, 2 principal components has 80% of Proportion Variance Explained, and 3 principal components has about 89% of Proportion of Variance Explained.

d) Print the principal component loadings, which describe how all of the variables combine to form the PC variables. Interpret the first 2 principal components. (*Hint: See Lab 12.5.1 in the book and the interpretations in the second half of Section 12.2.1 in the ISLR 2e text.*)

```
print(pca$rotation)
```

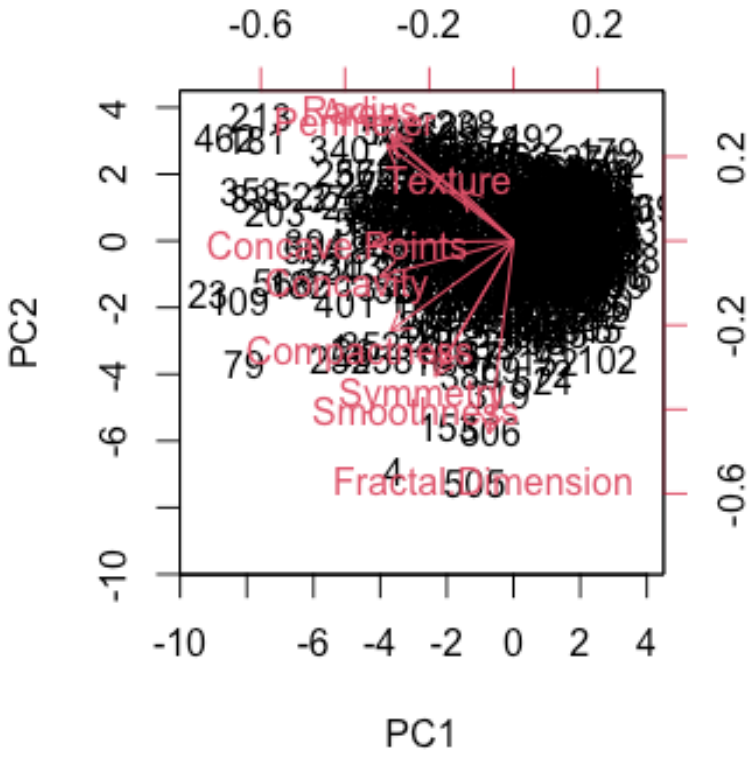
```
##                          PC1          PC2         PC3          PC4
## Radius            -0.36393793  0.313929073 -0.12442759  0.029558858
## Texture           -0.15445113  0.147180909  0.95105659  0.008916084
## Perimeter         -0.37604434  0.284657885 -0.11408360  0.013458069
```

```
## Area               -0.36408585  0.304841714 -0.12337786  0.013442682
## Smoothness         -0.23248053 -0.401962324 -0.16653247 -0.107802033
## Compactness        -0.36444206 -0.266013147  0.05827786 -0.185700413
## Concavity          -0.39574849 -0.104285968  0.04114649 -0.166653523
## Concave.Points     -0.41803840 -0.007183605 -0.06855383 -0.072983951
## Symmetry           -0.21523797 -0.368300910  0.03672364  0.892998475
## Fractal.Dimension  -0.07183744 -0.571767700  0.11358395 -0.349331790
##                            PC5          PC6         PC7          PC8
## Radius             -0.031067022  0.264180150 -0.04418839  0.084834062
## Texture            -0.219922761  0.032206572  0.02055748 -0.007126797
## Perimeter          -0.005945081  0.237819464 -0.08336923  0.089258879
## Area               -0.019341222  0.331707454  0.26118796  0.144609749
## Smoothness         -0.843745292 -0.062225368  0.01129197  0.170503128
## Compactness         0.240182967 -0.005271104 -0.80380484  0.063980134
## Concavity           0.312533244 -0.601467155  0.36713629  0.449573315
## Concave.Points     -0.009180198 -0.265613395  0.14131308 -0.850918762
## Symmetry            0.112888068  0.061957003  0.04790201  0.016455606
## Fractal.Dimension   0.264878077  0.567918997  0.34521359 -0.065259461
##                            PC9          PC10
## Radius              0.474425305 -0.6690714888
## Texture             0.004212629  0.0002497826
## Perimeter           0.380167210  0.7404905337
## Area               -0.747347357 -0.0323589585
## Smoothness          0.005847386  0.0036904058
## Compactness        -0.218732407 -0.0527527802
## Concavity           0.081170670 -0.0103668020
## Concave.Points     -0.022024652 -0.0037475480
## Symmetry            0.009067850  0.0014669472
## Fractal.Dimension   0.129667491  0.0070573477
```

The table shows the loadings of the component loadings. The PC1 has strong negative loadings for several feature. This suggests that these variables have a significant impact on PCA1 and tend to move in the same direction. The PC2 shows different patterns of loadings. Features like Fractal.Dimension (-0.572), Smoothness (-0.402), Symmetry (-0.368), and Compactness (-0.266) have significant negative contributions. This indicates that PC2 might represent a “smoothness/irregularity” dimension, focusing more on the irregularity or roughness of the objects in question.

e) Make a biplot of the first 2 principal components and interpret.

```
biplot(pca, scale = 0)
```



According to the biplot, we can see that the Fractal Dimension has significant effect on PC2 but not as much on PC1. We can see that the Concave Point has the strongest effect on PC1 but not as much on PC2. We can also see that all the variables has is pointing left, which means that the variables have a negative loading or contribution to the PC1. We can also see the most of the variables have negative loadings to the PC2.

Problem 2

Now, we will use *K*-means and hierarchical clustering on the data.

a) How many clusters do you think we should specify? Why?

The original data has a 2 diagnosis as the response variable, so there should be only 2 clusters.

b) Conduct *K*-means clustering with 2 clusters. What are the cluster means?

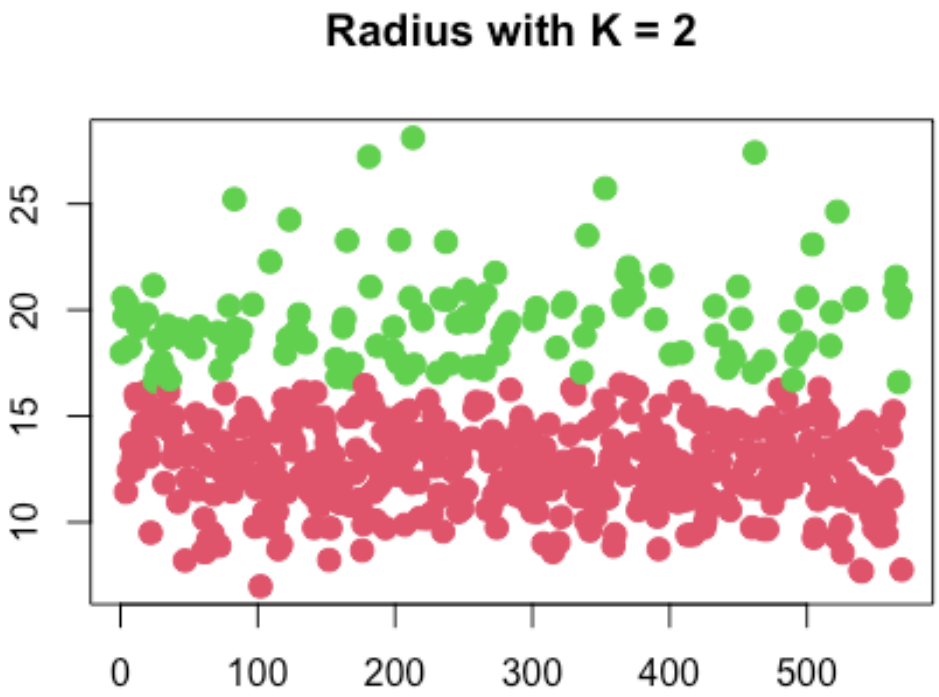
```
set.seed(1)
kmean<-kmeans(data, 2)
print("Cluster Mean:")
```

```
## [1] "Cluster Mean:"
```

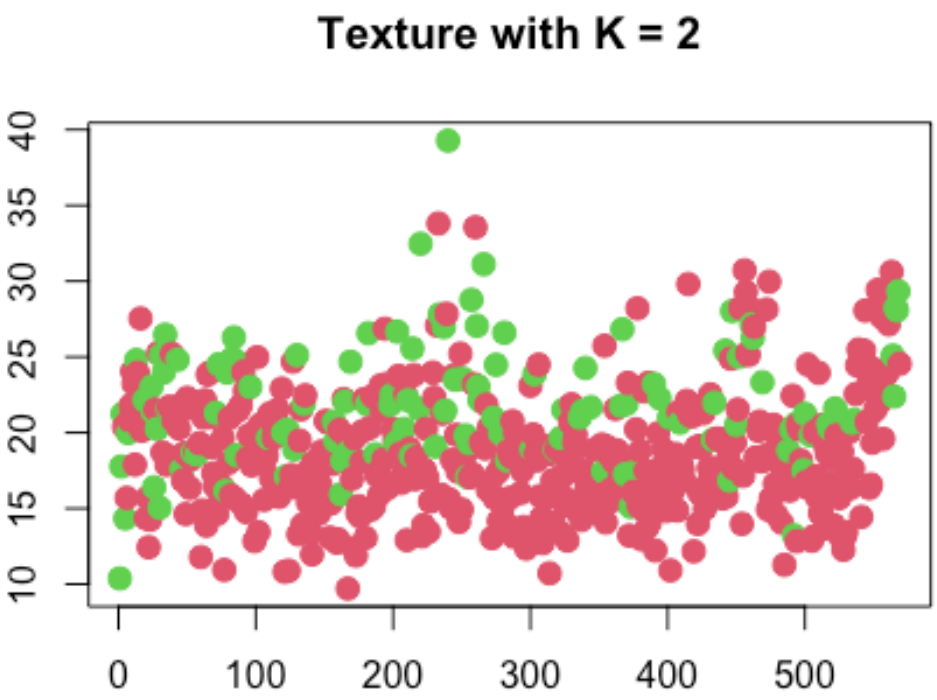
```
print(kmean$centers)

##     Radius  Texture Perimeter     Area Smoothness Compactness Concavity
## 1 12.59721 18.57845  81.45276  499.667 0.09525933  0.09277371 0.0645051
## 2 19.61831 21.84194 129.70887 1211.936 0.10031129  0.14585258 0.1759842
##   Concave.Points  Symmetry Fractal.Dimension
## 1     0.03459259 0.1786782        0.06359333
## 2     0.10033298 0.1900750        0.05994202
```

c) Make a scatterplot matrix between all variables and color the points by cluster. Describe your findings.

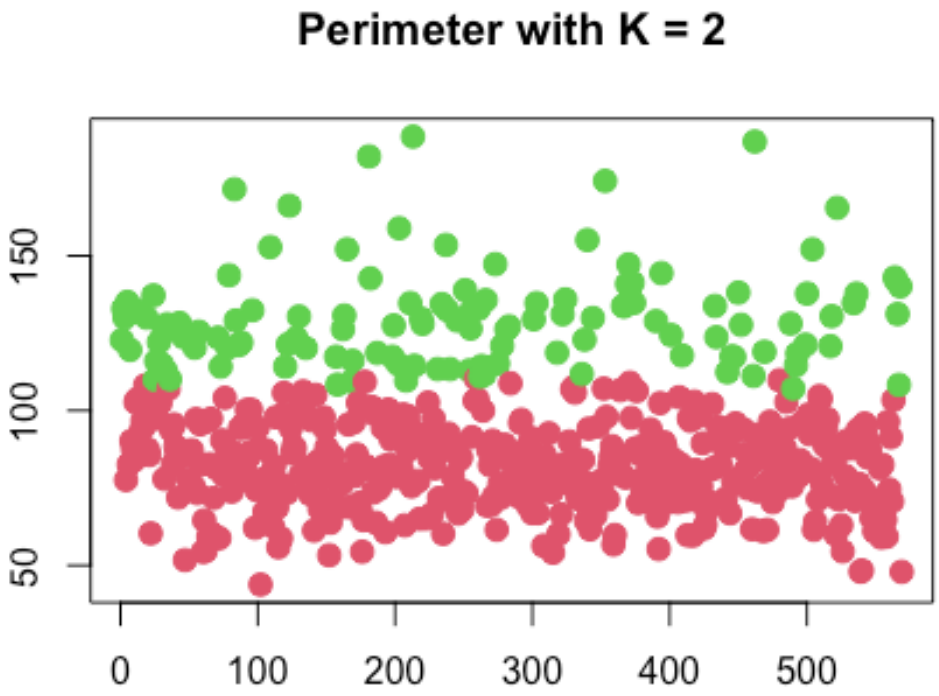
```
plot(data$Radius, col = (kmean$cluster + 1),
main = "Radius with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



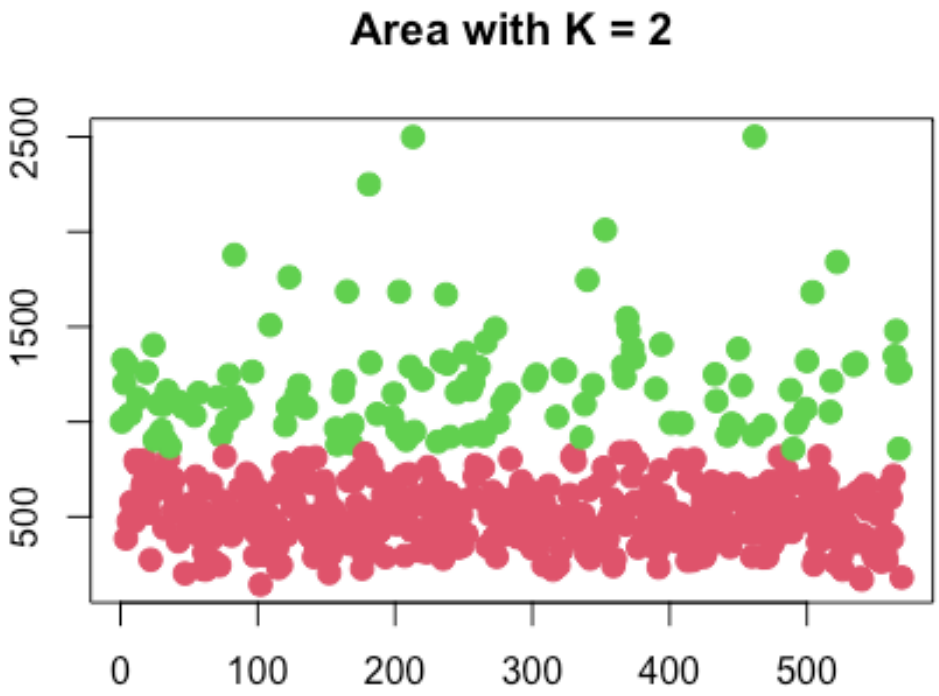

```
plot(data$Texture, col = (kmean$cluster + 1),
main = "Texture with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



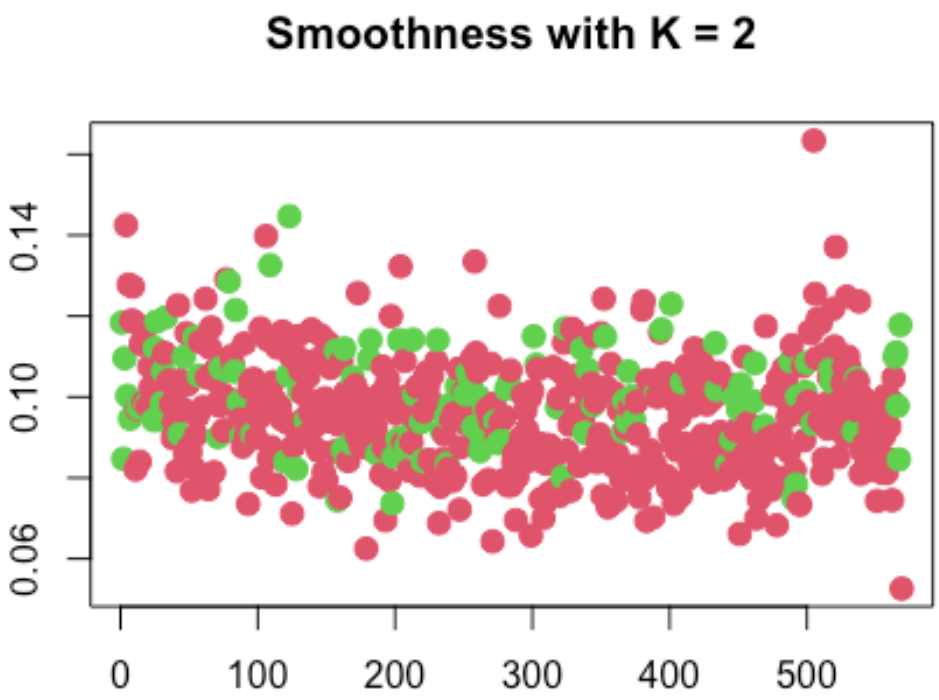

```
plot(data$Perimeter, col = (kmean$cluster + 1),
main = "Perimeter with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



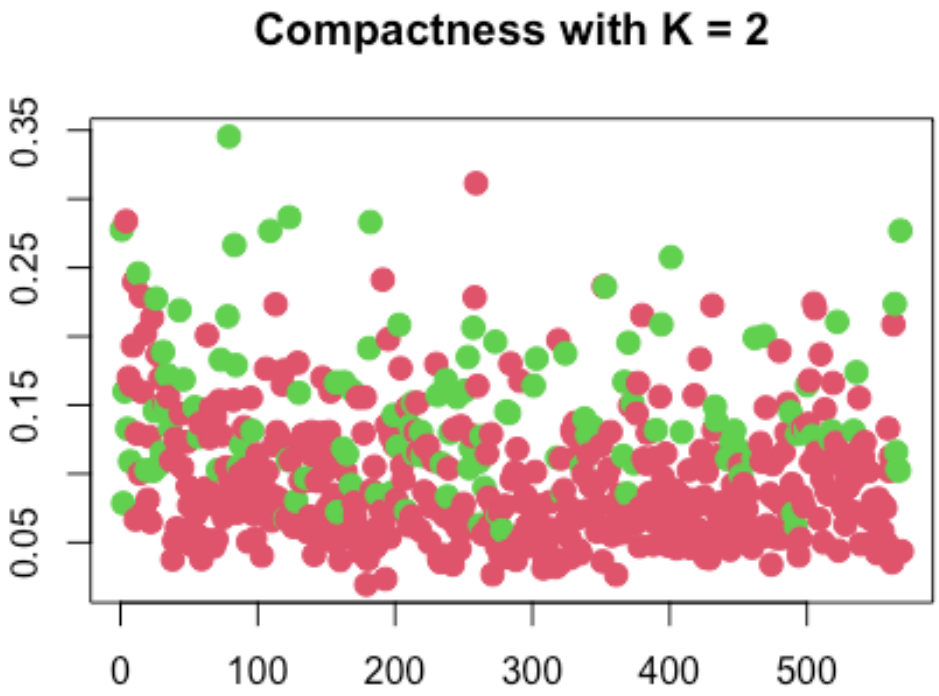

```
plot(data$Area, col = (kmean$cluster + 1),
main = "Area with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



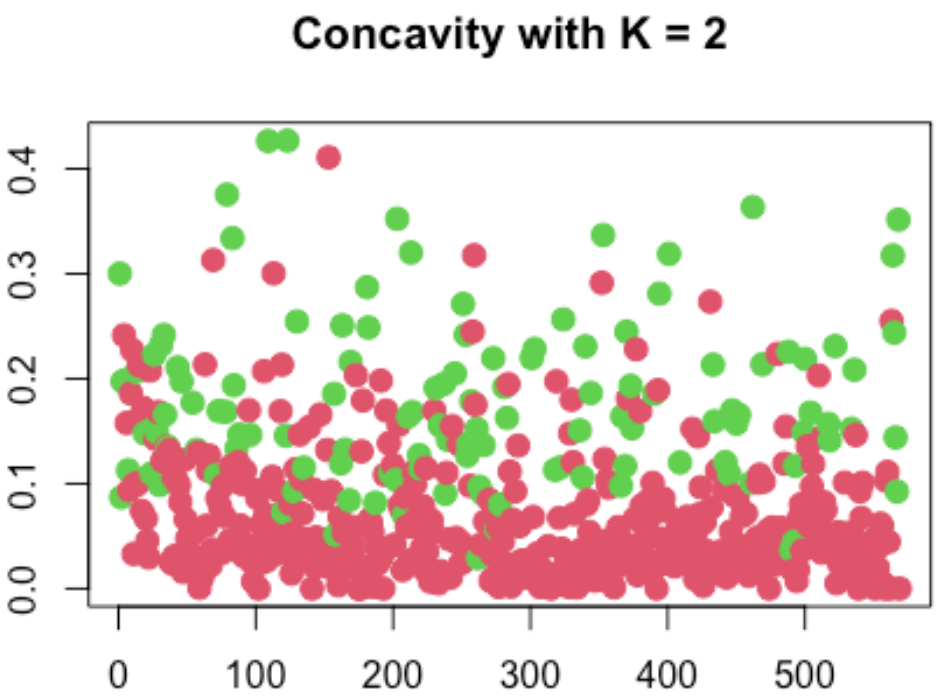

```
plot(data$Smoothness, col = (kmean$cluster + 1),
main = "Smoothness with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



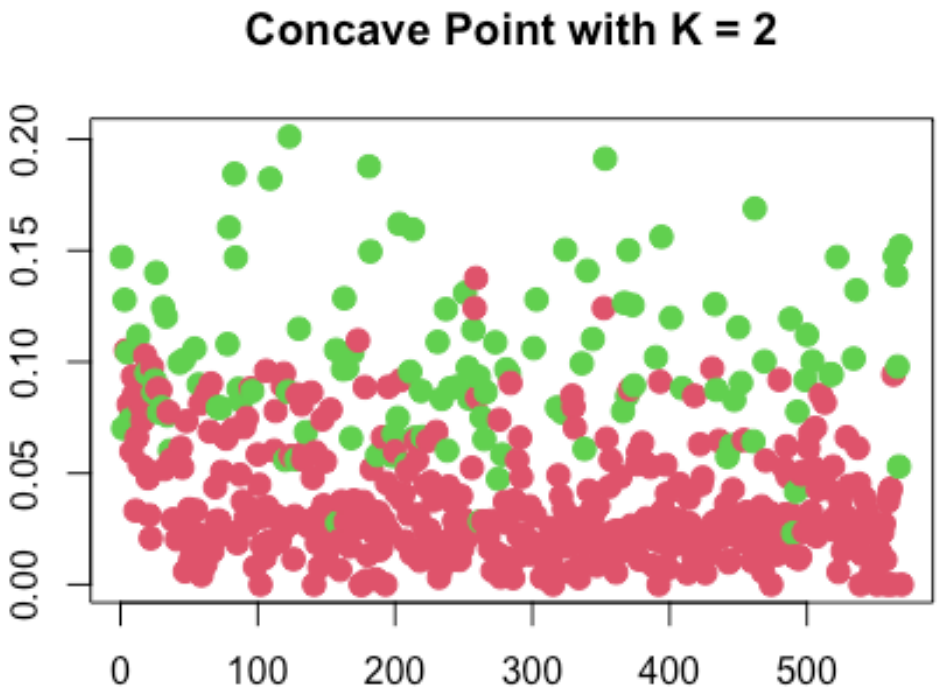

```
plot(data$Compactness, col = (kmean$cluster + 1),
main = "Compactness with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



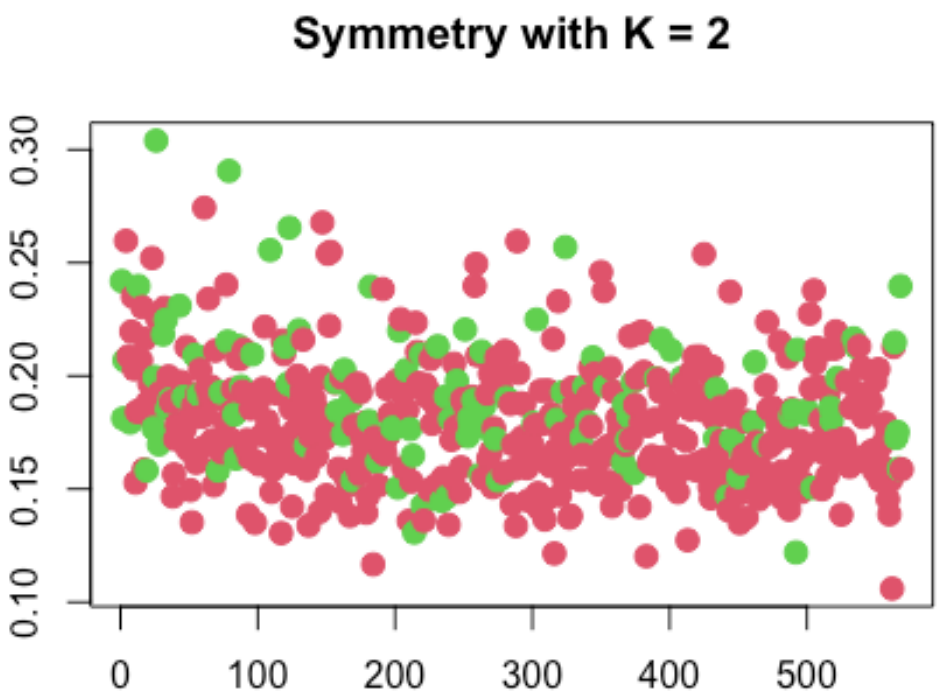

```
plot(data$Concavity, col = (kmean$cluster + 1),
main = "Concavity with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



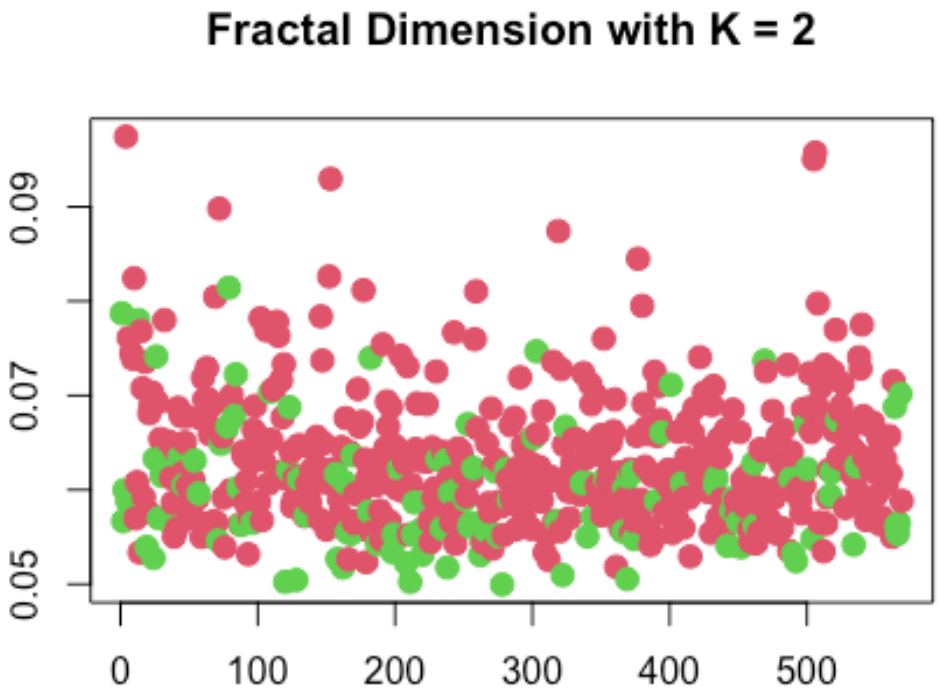

```
plot(data$Concave.Points, col = (kmean$cluster + 1),
main = "Concave Point with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```




```
plot(data$Symmetry, col = (kmean$cluster + 1),
main = "Symmetry with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



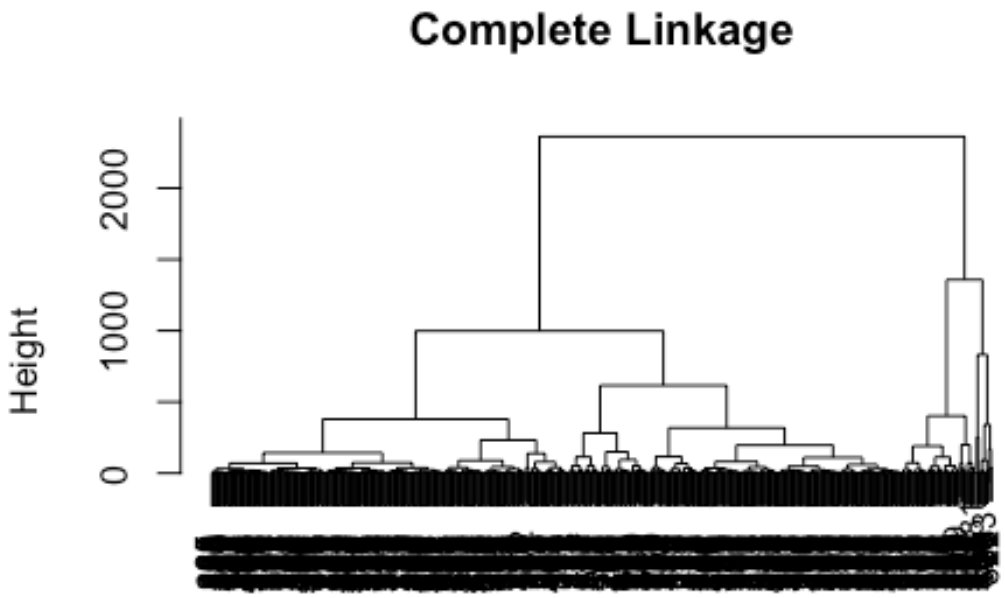

```
plot(data$Fractal.Dimension, col = (kmean$cluster + 1),
main = "Fractal Dimension with K = 2", xlab = "", ylab = "", pch = 20, cex =
2)
```



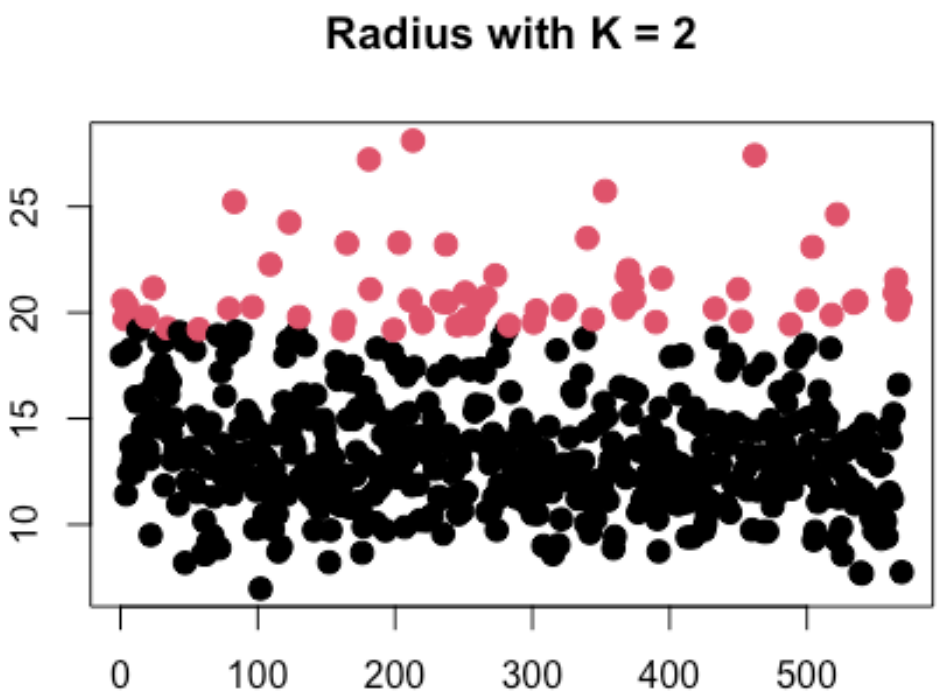
According to these plots, we can see that there's a clear cluster distinction for the variables Radius, Perimeter, and Area. This means that there's a clear separation in the data. In addition, we can see that red cluster is on the bottom and green point is on the top. In conclusion, there are several variables that have have clear difference in clusters, but some variables is difficult to tell the difference.

d) Using complete linkage, perform hierarchical clustering and color the points on the scatterplot matrix as in the previous question. Is the result much different?

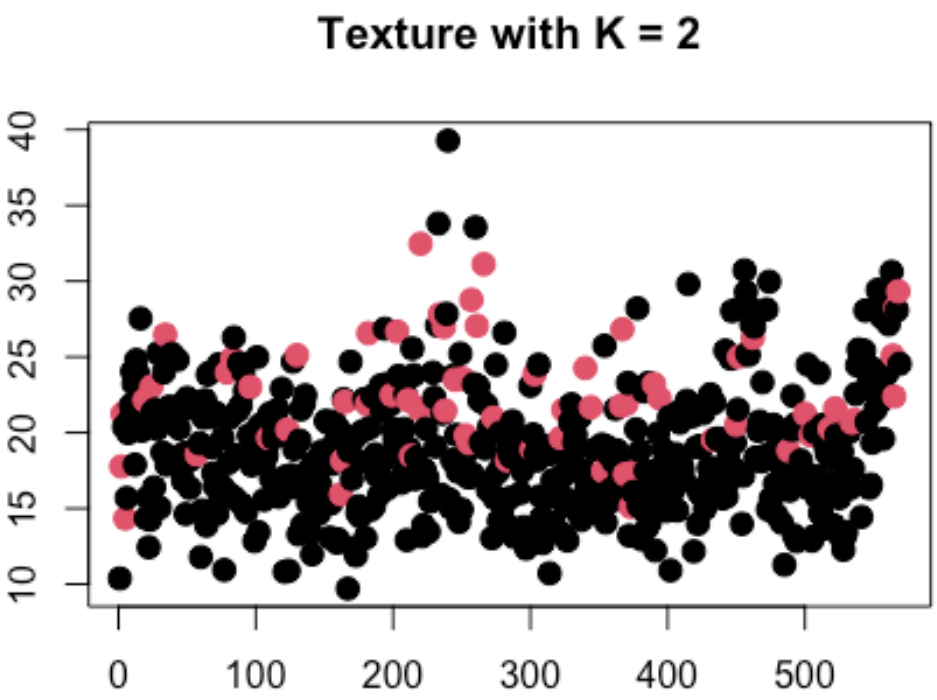
```
hc.complete<-hclust(dist(data), method = "complete")

plot(hc.complete, main = "Complete Linkage", xlab = "", sub = "", cex = .9)
```



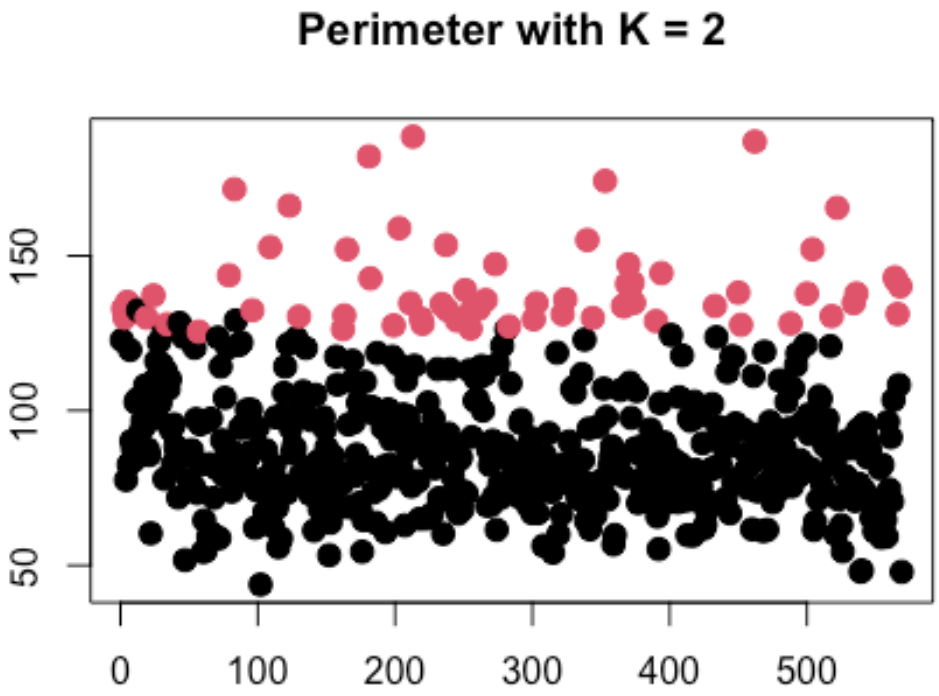

```
clusters<-cutree(hc.complete, k=2)

data$cluster<-as.factor(clusters)

plot(data$Radius, col = (data$cluster),
main = "Radius with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



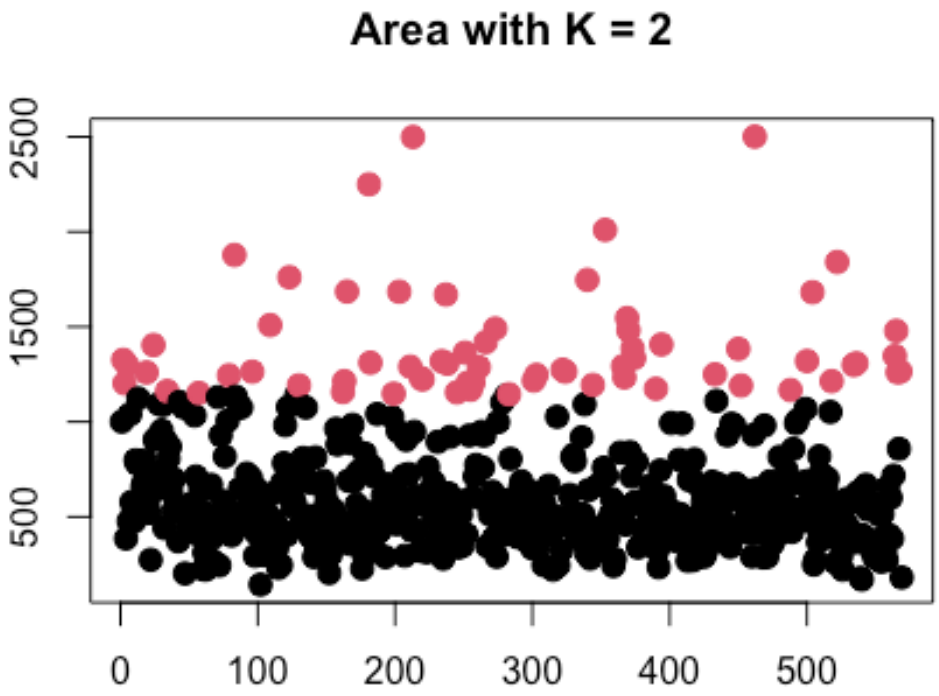

```
plot(data$Texture, col = (data$cluster),
main = "Texture with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



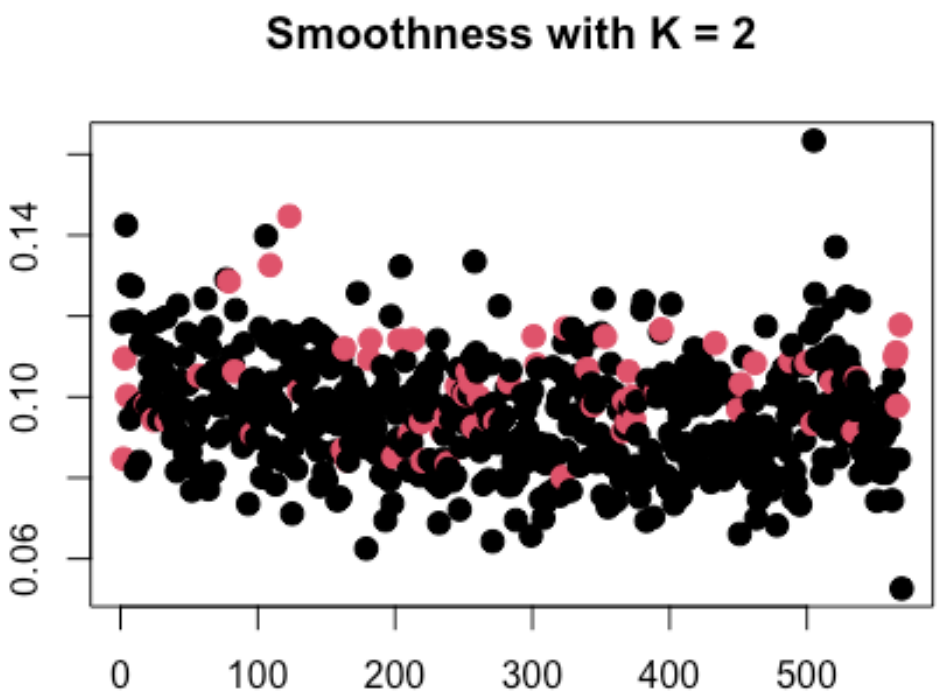

```
plot(data$Perimeter, col = (data$cluster),
main = "Perimeter with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



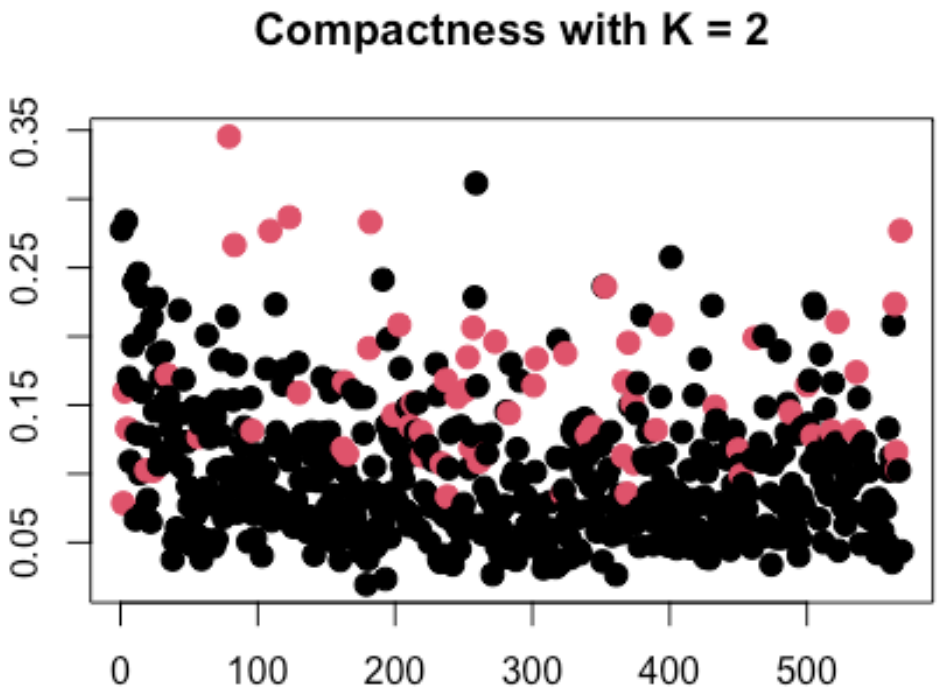

```
plot(data$Area, col = (data$cluster),
main = "Area with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



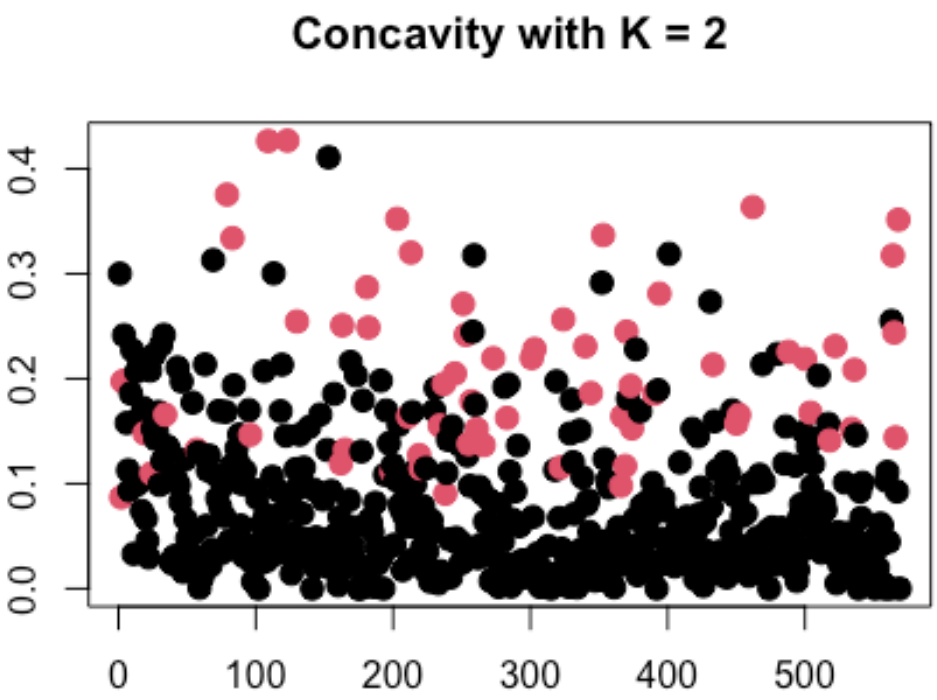

```
plot(data$Smoothness, col = (data$cluster),
main = "Smoothness with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



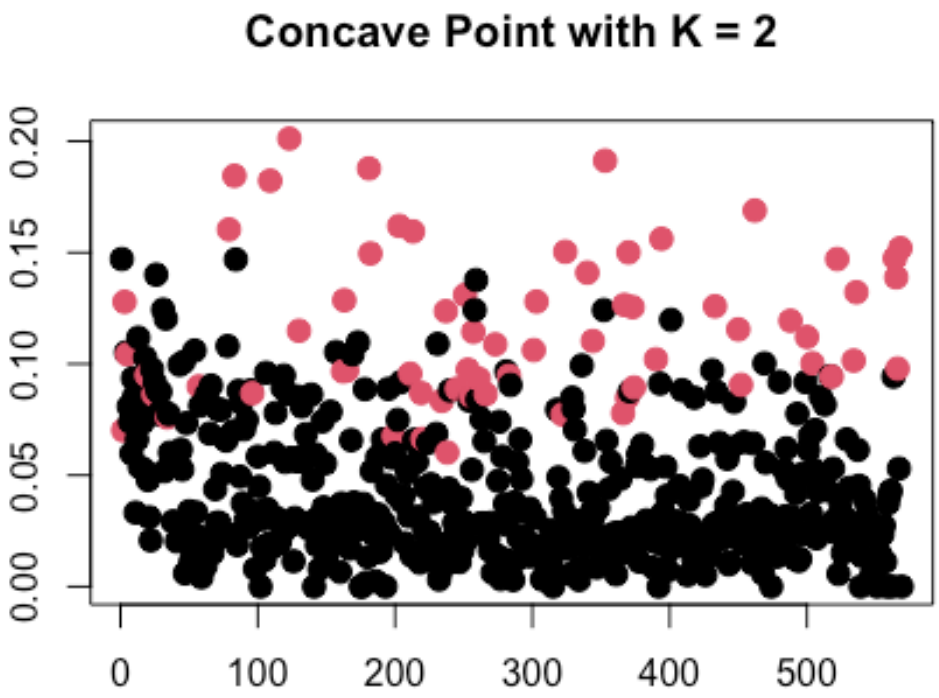

```
plot(data$Compactness, col = (data$cluster),
main = "Compactness with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



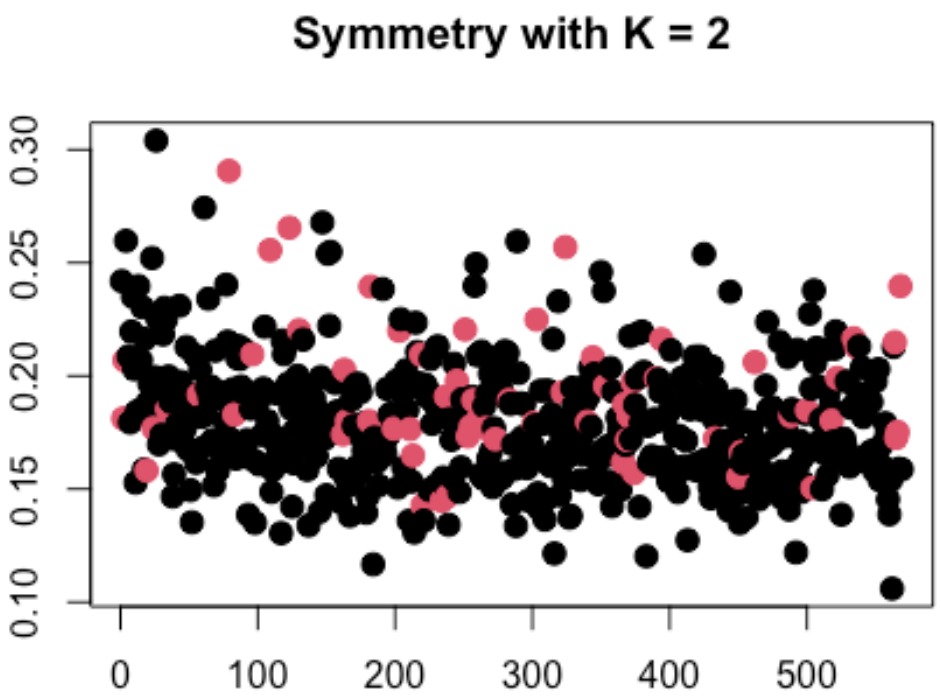

```
plot(data$Concavity, col = (data$cluster),
main = "Concavity with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```



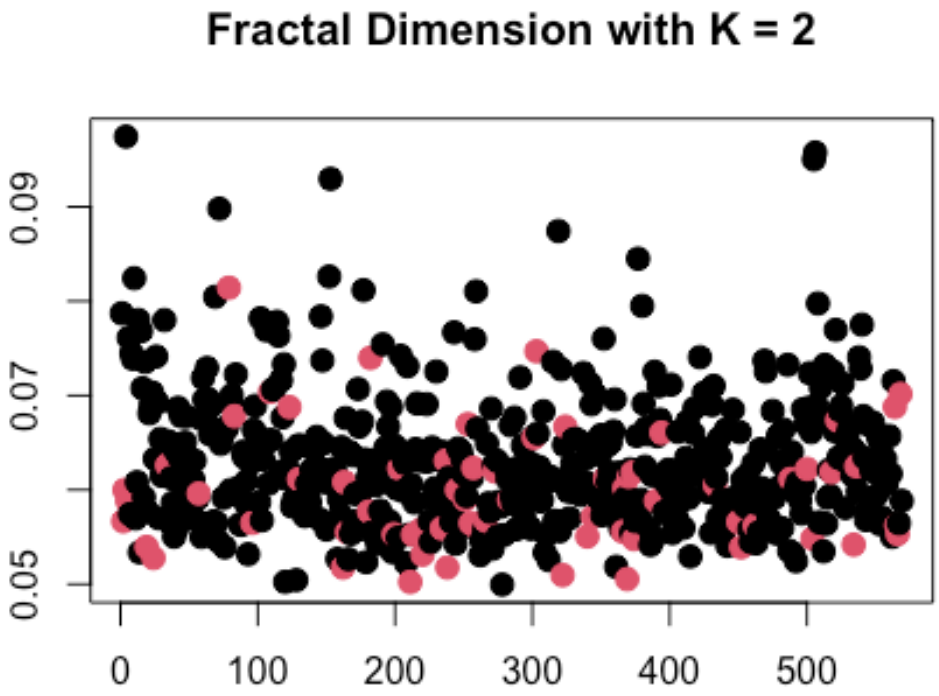

```
plot(data$Concave.Points, col = (data$cluster),
main = "Concave Point with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```




```
plot(data$Symmetry, col = (data$cluster),
main = "Symmetry with K = 2", xlab = "", ylab = "", pch = 20, cex = 2)
```




```
plot(data$Fractal.Dimension, col = (data$cluster),
main = "Fractal Dimension with K = 2", xlab = "", ylab = "", pch = 20, cex =
2)
```



According to the plots, there is not that much different than K-Means clustering, except the proportion of the cluster of black point (corresponds to red for K-Means) is higher compare to the Pink (correspond to green for K-Means).