class09_mini_project

Jaquish (PID: A59010386)

10/27/2021

```
fna.data <- "WisconsinCancer.csv"
wisc.df <- read.csv(fna.data, row.names=1)
```

```
# wisc.df
```

```
wisc.data <- wisc.df[,-1]
```

```
diagnosis <- as.factor(wisc.df$diagnosis)
#diagnosis
```

> Q1. How many observations are in this dataset?

```
nrow(wisc.data)
```

```
## [1] 569
```

> Q2. How many of the observations have a malignant diagnosis?

```
sum(diagnosis == "M")
```

```
## [1] 212
```

```
table(diagnosis)
```

```
## diagnosis
##   B   M
## 357 212
```

> Q3. How many variables/features in the data are suffixed with _mean?

```
grep("_mean", colnames(wisc.data), value=TRUE)
```

```
##  [1] "radius_mean"            "texture_mean"           "perimeter_mean"
##  [4] "area_mean"              "smoothness_mean"        "compactness_mean"
##  [7] "concavity_mean"         "concave.points_mean"    "symmetry_mean"
## [10] "fractal_dimension_mean"
```

```r
length(grep("_mean", colnames(wisc.data), TRUE))
```

```
## [1] 10
```

```r
#colMeans(wisc.data)
```

```r
#apply(wisc.data,2,sd)
```

```r
wisc.pr <- prcomp( wisc.data, scale=TRUE)
```

```r
summary(wisc.pr)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.6444 2.3857 1.67867 1.40735 1.28403 1.09880 0.82172
## Proportion of Variance 0.4427 0.1897 0.09393 0.06602 0.05496 0.04025 0.02251
## Cumulative Proportion  0.4427 0.6324 0.72636 0.79239 0.84734 0.88759 0.91010
##                            PC8    PC9    PC10   PC11    PC12    PC13    PC14
## Standard deviation     0.69037 0.6457 0.59219 0.5421 0.51104 0.49128 0.39624
## Proportion of Variance 0.01589 0.0139 0.01169 0.0098 0.00871 0.00805 0.00523
## Cumulative Proportion  0.92598 0.9399 0.95157 0.9614 0.97007 0.97812 0.98335
##                           PC15    PC16    PC17    PC18    PC19    PC20   PC21
## Standard deviation     0.30681 0.28260 0.24372 0.22939 0.22244 0.17652 0.1731
## Proportion of Variance 0.00314 0.00266 0.00198 0.00175 0.00165 0.00104 0.0010
## Cumulative Proportion  0.98649 0.98915 0.99113 0.99288 0.99453 0.99557 0.9966
##                           PC22    PC23   PC24    PC25    PC26    PC27    PC28
## Standard deviation     0.16565 0.15602 0.1344 0.12442 0.09043 0.08307 0.03987
## Proportion of Variance 0.00091 0.00081 0.0006 0.00052 0.00027 0.00023 0.00005
## Cumulative Proportion  0.99749 0.99830 0.9989 0.99942 0.99969 0.99992 0.99997
##                           PC29    PC30
## Standard deviation     0.02736 0.01153
## Proportion of Variance 0.00002 0.00000
## Cumulative Proportion  1.00000 1.00000
```

> Q4. From your results, what proportion of the original variance is captured by the first principal components (PC1)?

The first PC1 captures 44% of the original variance.

> Q5. How many principal components (PCs) are required to describe at least 70% of the original variance in the data?

To get 70% you need to go up to PC3.

> Q6. How many principal components (PCs) are required to describe at least 90% of the original variance in the data?

To hit 90% you need PC7.

> Q7. What stands out to you about this plot? Is it easy or difficult to understand? Why?

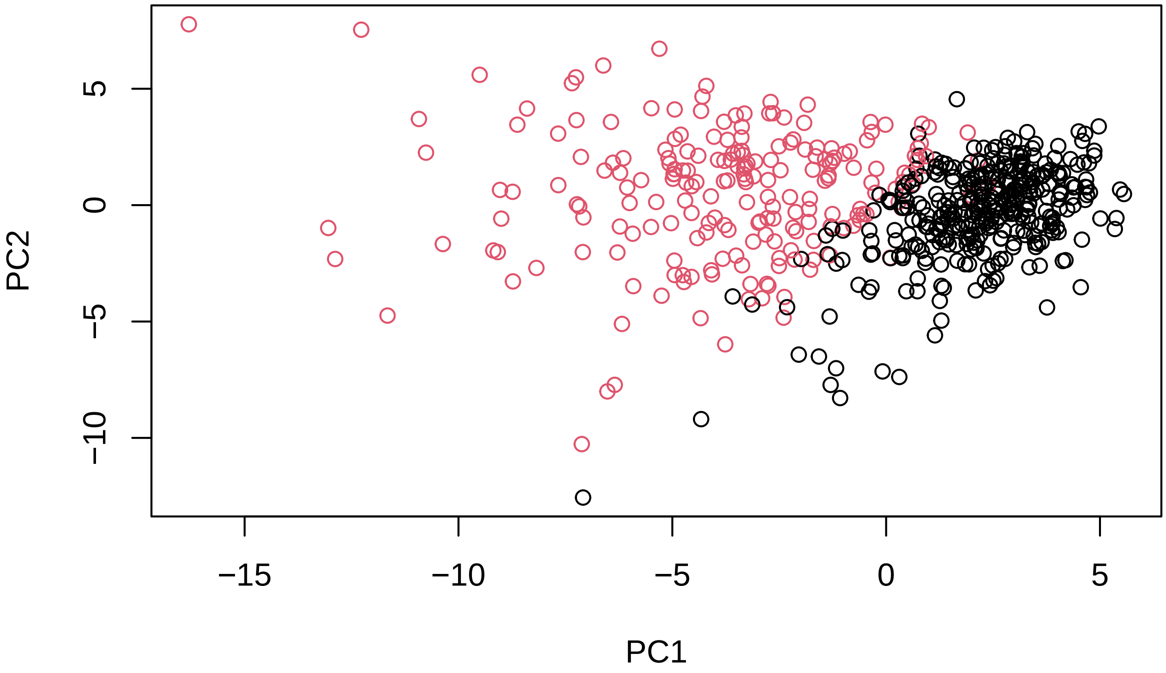
Very messy and difficult to understand because there is just too much data.

```r
#biplot(wisc.pr)
```

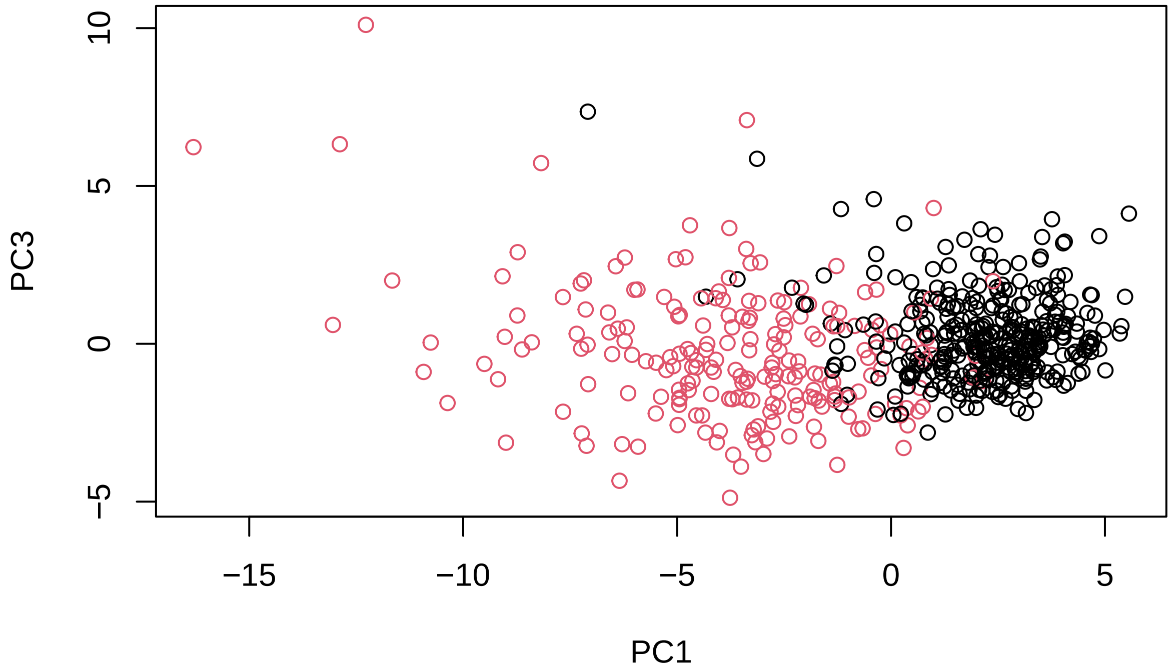
> Q8. Generate a similar plot for principal components 1 and 3. What do you notice about these plots?

They are much prettier and look pretty similar. The plot with 1 and 3 look to be tighter together because they cover a small variance.

```r
plot(wisc.pr$x[,1:2], col= diagnosis, xlab = "PC1", ylab = "PC2")
```

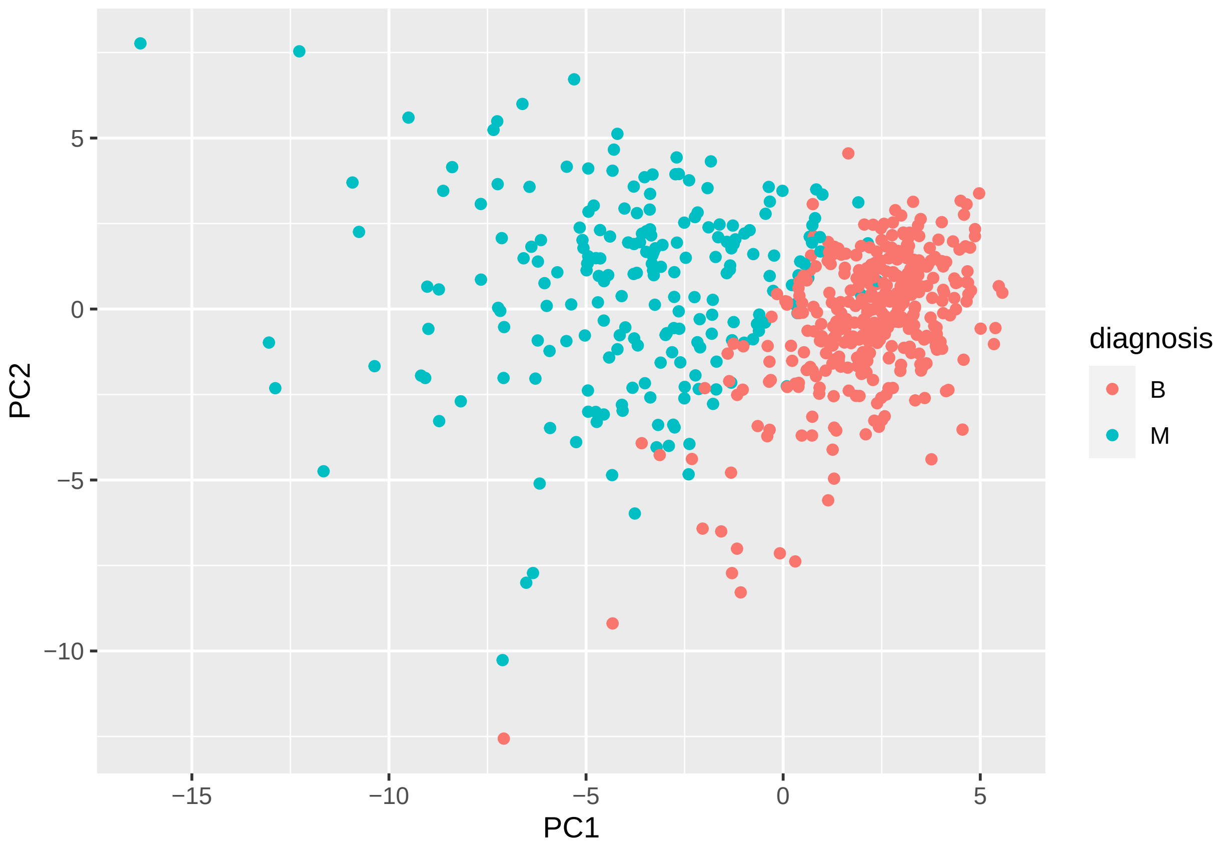


```r
plot(wisc.pr$x[,1], wisc.pr$x[,3], col= diagnosis, xlab = "PC1", ylab = "PC3")
```



```
df <- as.data.frame(wisc.pr$x)
df$diagnosis <- diagnosis
library(ggplot2)
```

```
ggplot(df) +
  aes(PC1, PC2, col= diagnosis) +
  geom_point()
```



```
pr.var <- wisc.pr$sdev^2
pr.var
```

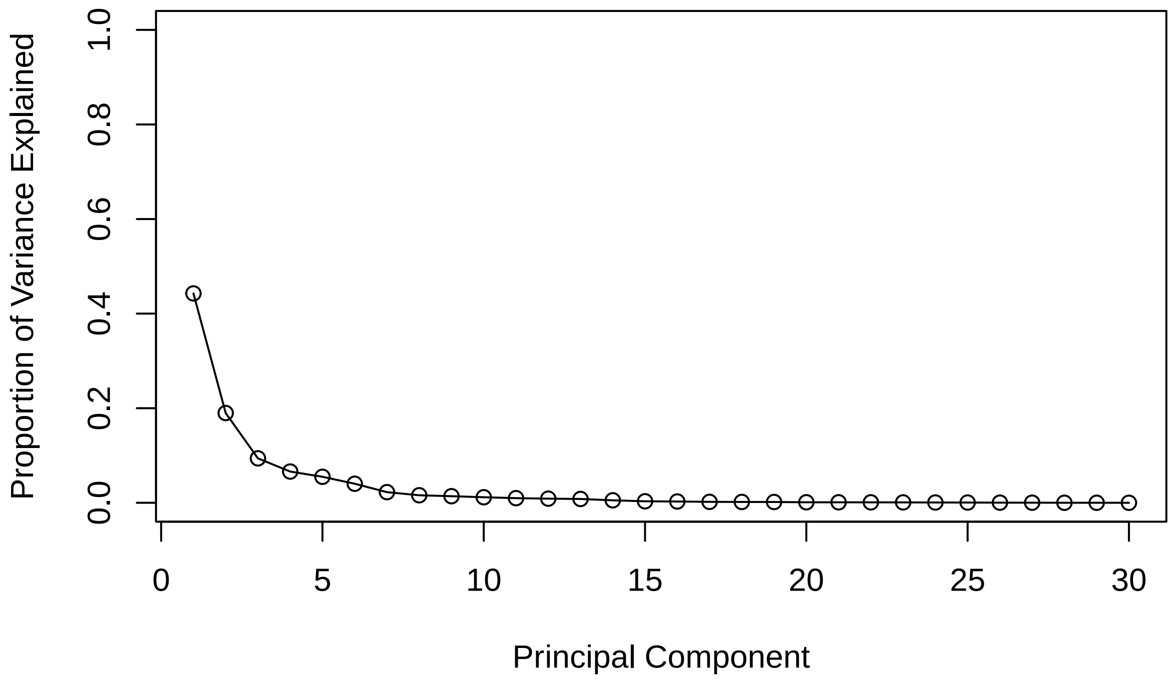
```
##  [1] 1.328161e+01 5.691355e+00 2.817949e+00 1.980640e+00 1.648731e+00
##  [6] 1.207357e+00 6.752201e-01 4.766171e-01 4.168948e-01 3.506935e-01
## [11] 2.939157e-01 2.611614e-01 2.413575e-01 1.570097e-01 9.413497e-02
## [16] 7.986280e-02 5.939904e-02 5.261878e-02 4.947759e-02 3.115940e-02
## [21] 2.997289e-02 2.743940e-02 2.434084e-02 1.805501e-02 1.548127e-02
## [26] 8.177640e-03 6.900464e-03 1.589338e-03 7.488031e-04 1.330448e-04
```

```
?head(pr.var)
```

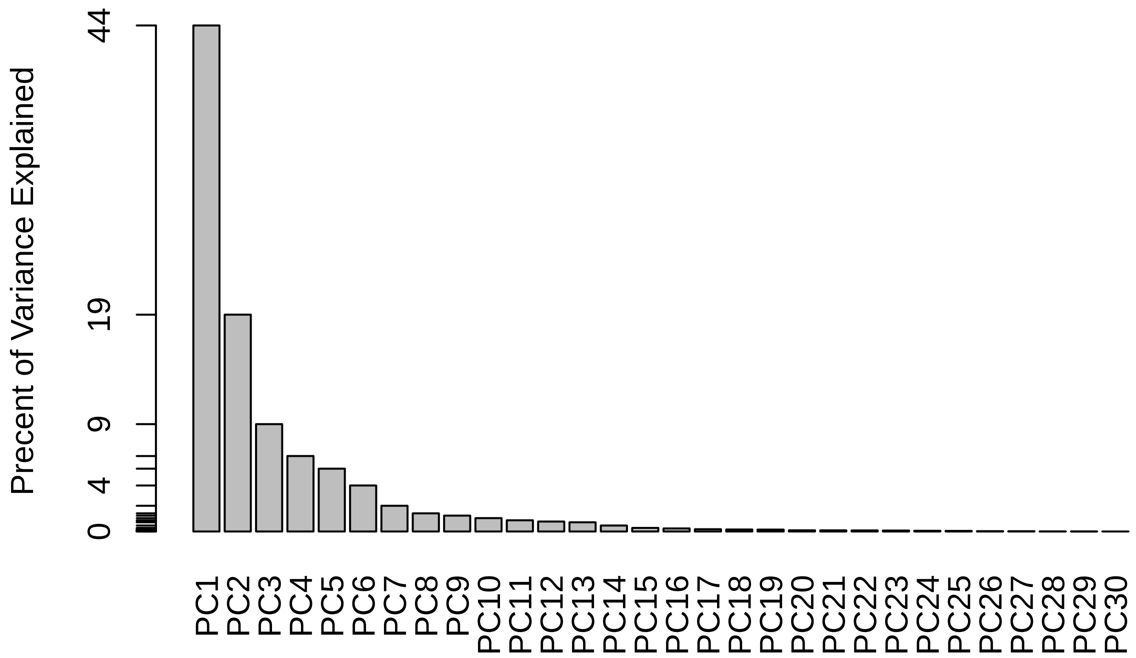
```
pve <-  pr.var/ sum(pr.var)
pve
```

```
##  [1] 4.427203e-01 1.897118e-01 9.393163e-02 6.602135e-02 5.495768e-02
##  [6] 4.024522e-02 2.250734e-02 1.588724e-02 1.389649e-02 1.168978e-02
## [11] 9.797190e-03 8.705379e-03 8.045250e-03 5.233657e-03 3.137832e-03
## [16] 2.662093e-03 1.979968e-03 1.753959e-03 1.649253e-03 1.038647e-03
## [21] 9.990965e-04 9.146468e-04 8.113613e-04 6.018336e-04 5.160424e-04
## [26] 2.725880e-04 2.300155e-04 5.297793e-05 2.496010e-05 4.434827e-06
```

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



```
barplot(pve, ylab = "Precent of Variance Explained",
     names.arg=paste0("PC",1:length(pve)), las=2, axes = FALSE)
axis(2, at=pve, labels=round(pve,2)*100 )
```



Optional:

```
#install.packages("factoextra")
#library(factoextra)
#fviz_eig(wisc.pr, addlabels = TRUE)
```

> Q9. For the first principal component, what is the component of the loading vector (i.e. wisc.pr$rotation[,1]) for the feature concave.points_mean?

The concave.points_mean is -0.26. This value is the average of the downward curve PC1 took because of the distance from this point.

```
#wisc.pr$rotation[,1]
```

> Q10. What is the minimum number of principal components required to explain 80% of the variance of the data?

You need 4 PC's to describe 80% of the variance.

```
var <- summary(wisc.pr)
sum(var$importance[3,] <0.8)
```

```
## [1] 4
```

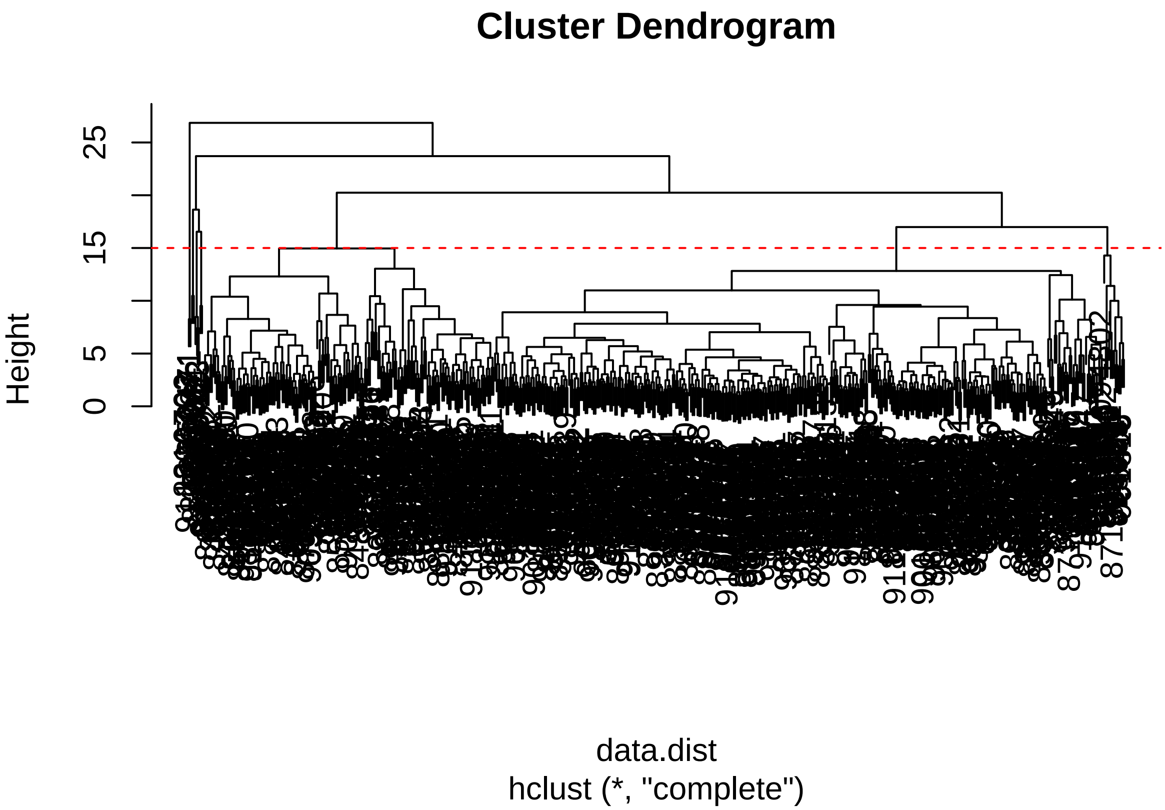
```
data.scaled <- scale(wisc.data)
```

```
data.dist <- dist(data.scaled)
```

```
#data.dist
```

```
wisc.hclust <- hclust(data.dist, "complete")
```

```
plot(wisc.hclust)
abline( h=15, col="red", lty=2)
```



```
wisc.hclust.clusters <- cutree(wisc.hclust,k=5)
table(wisc.hclust.clusters, diagnosis)
```

```
##                     diagnosis
## wisc.hclust.clusters   B   M
##                    1  12 165
##                    2   0   5
##                    3 343  40
##                    4   2   0
##                    5   0   2
```

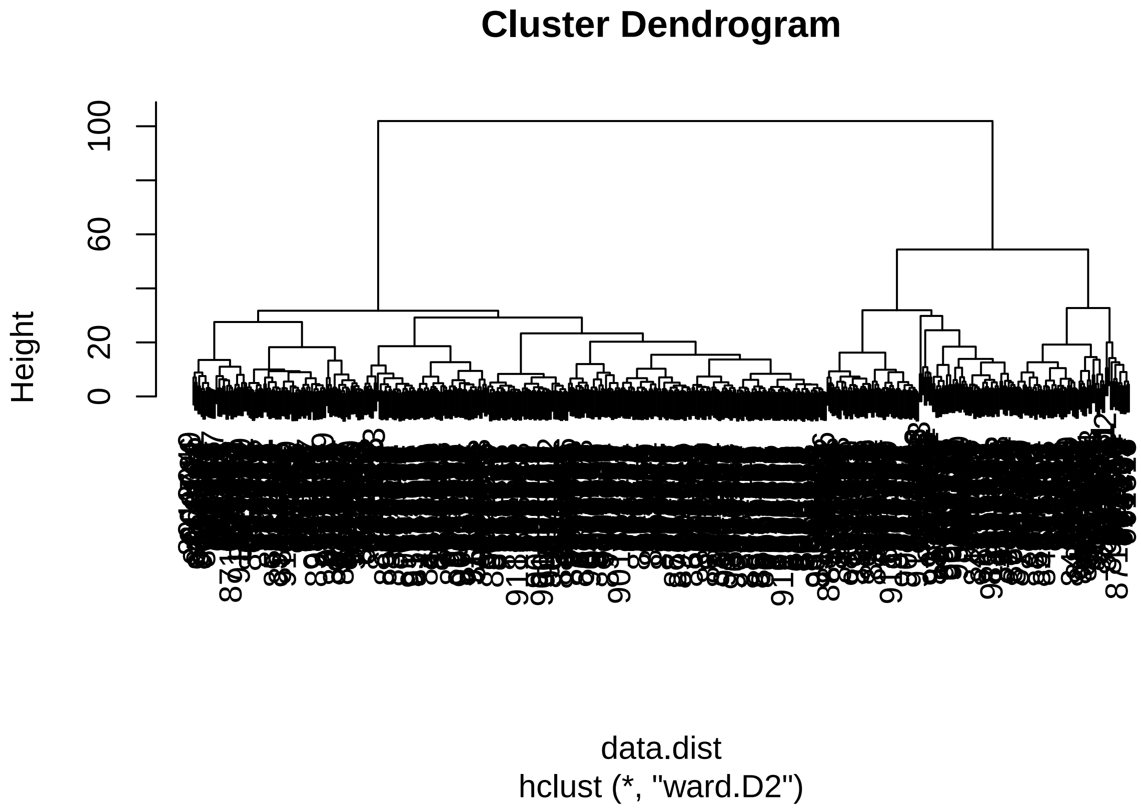
Q12. Can you find a better cluster vs diagnoses match by cutting into a different number of clusters between 2 and 10?

Cluster 5 seems to be good because there is a nice separation between the diagnoses of benign and malignant.

> Q13. Which method gives your favorite results for the same data.dist dataset? Explain your reasoning.

I also like the ward.D2 because the groups look very clean and it builds it so the variance is minimized. This is similar to why I chose 5 clusters above because it groups them together into a single category best as possible.

```
wisc.hclust <- hclust(data.dist, "ward.D2")
plot(wisc.hclust)
```


Cluster Dendrogram
Height
data.dist
hclust (*, "ward.D2")