class08:Breast Cancer_mini_project

Angie(PID:69028746)

2024-02-02

1.Exploratory data analysis

– Complete the following code to input the data and store as wisc.df

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

– Save the diagnosis for reference later

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

– and remove or exclude this column from any of our analysis

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

> **Q1.** How many observations/samples/patients/rows are in this dataset?
>
> **A1.** There are 569 observations in this dataset

```
dim(wisc.data)
```

```
[1] 569  30
```

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

```
[1] 569
```

> **Q2.** How many of the observations have a malignant diagnosis?
>
> **A2.** There are 212 observations with a malignant diagnosis.

```r
sum(wisc.df$diagnosis == "M")
```

```
[1] 212
```

```r
table(wisc.df$diagnosis)
```

```

  B   M 
357 212 
```

> **Q3.** How many variables/features in the data are suffixed with _mean?
>
> **A3.** There are 10 variables in the data are suffixed with _mean.

```r
length(grep("_mean", colnames(wisc.df), value=TRUE))
```

```
[1] 10
```

2.Principal Component Analysis

– Let's try PCA on this data. Before doing any analysis like this we should check if our input data needs to be scaled first? – Do we need to scale this data set? Yes, we do, because the spread is very different

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

– How well do the PCs capture the variants in the original data?

```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
```

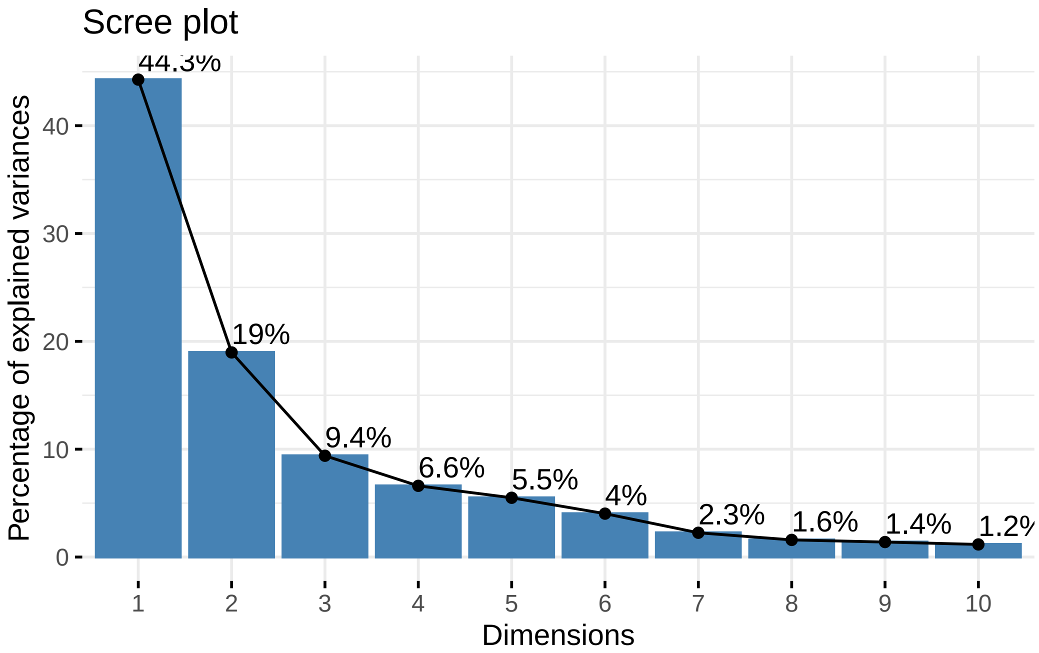
```r
v<- summary(wisc.pr)
v$importance[2,]
```

```
    PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10
0.44272 0.18971 0.09393 0.06602 0.05496 0.04025 0.02251 0.01589 0.01390 0.01169
   PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18    PC19    PC20
0.00980 0.00871 0.00805 0.00523 0.00314 0.00266 0.00198 0.00175 0.00165 0.00104
   PC21    PC22    PC23    PC24    PC25    PC26    PC27    PC28    PC29    PC30
0.00100 0.00091 0.00081 0.00060 0.00052 0.00027 0.00023 0.00005 0.00002 0.00000
```

```r
library(ggplot2)
library(factoextra)
```

```
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```r
fviz_eig(wisc.pr, addlabels = TRUE)
```



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

A4. 44.27% of the total variance in the original data is captured by PC1.

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

A5. The first 3 principle components together capture approximately 72.64% of the original variance in the data. Therefore, the first 3 principal components are required to describe at least 70% of the original variance in the data.

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

A6. The first 7 principle components together capture approximately 91.01% of the original variance in the data. Therefore, the first 7 principal components are required to describe at least 90% of the original variance in the data.

our main PC score plot (aka. PC plot, PC1 vs PC2, ordination plot)

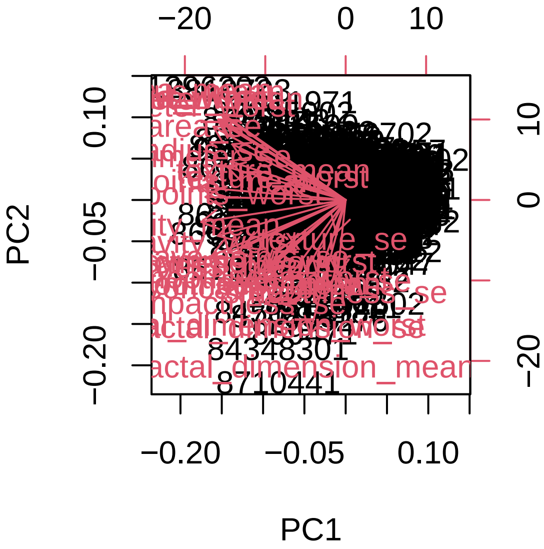
```
attributes(wisc.pr)
```

```
$names
[1] "sdev"     "rotation" "center"   "scale"    "x"

$class
[1] "prcomp"
```

```
#wisc.pr$x
```

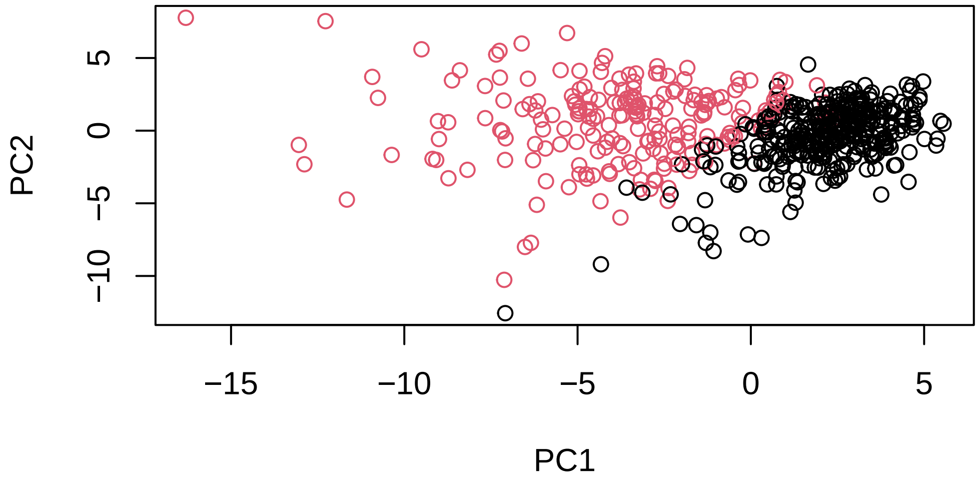
```
biplot(wisc.pr)
```



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

A7. This is a hot mess of a plot. It is very hard to understand since it contains too much overlapped information. Rownames are used as the plotting character for biplots like this one which can make trends rather hard to see. We will need to generate our own plots to make sense of this PCA result.

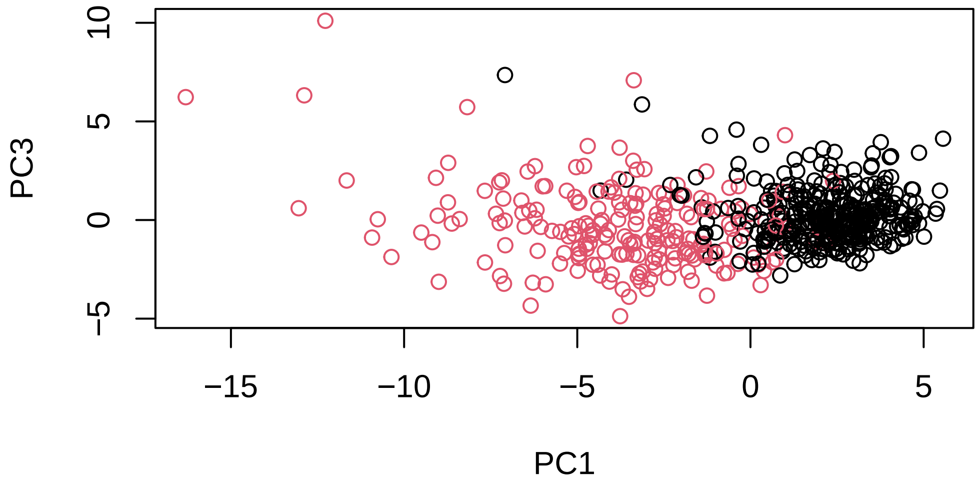
```
plot(wisc.pr$x[,1], wisc.pr$x[,2], col=diagnosis,
    xlab = "PC1", ylab = "PC2")
```



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

A8. Because PC 2 explains more variance in the original data than PC 3, I can see that the first plot of PC1 VS PC2 has a cleaner cut separating the two subgroups. Overall, the plots indicate that PC 1 is capturing a separation of malignant (red) from benign (black) samples.

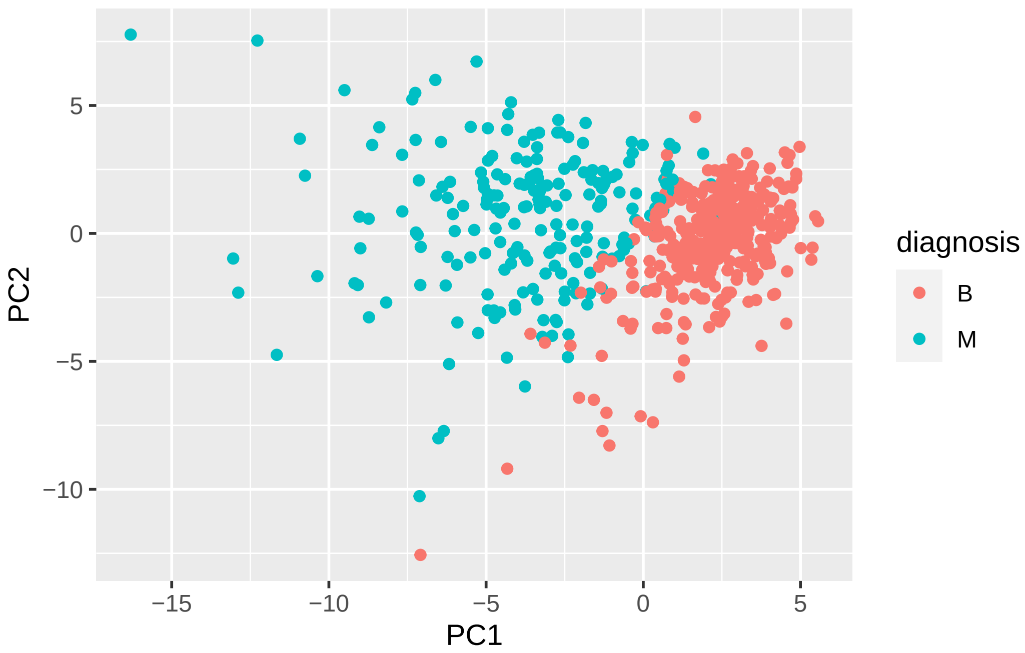
```
plot(wisc.pr$x[,1], wisc.pr$x[,3], col = diagnosis,
     xlab = "PC1", ylab = "PC3")
```



make a nice ggplot version

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

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



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? This tells us how much this original feature contributes to the first PC.

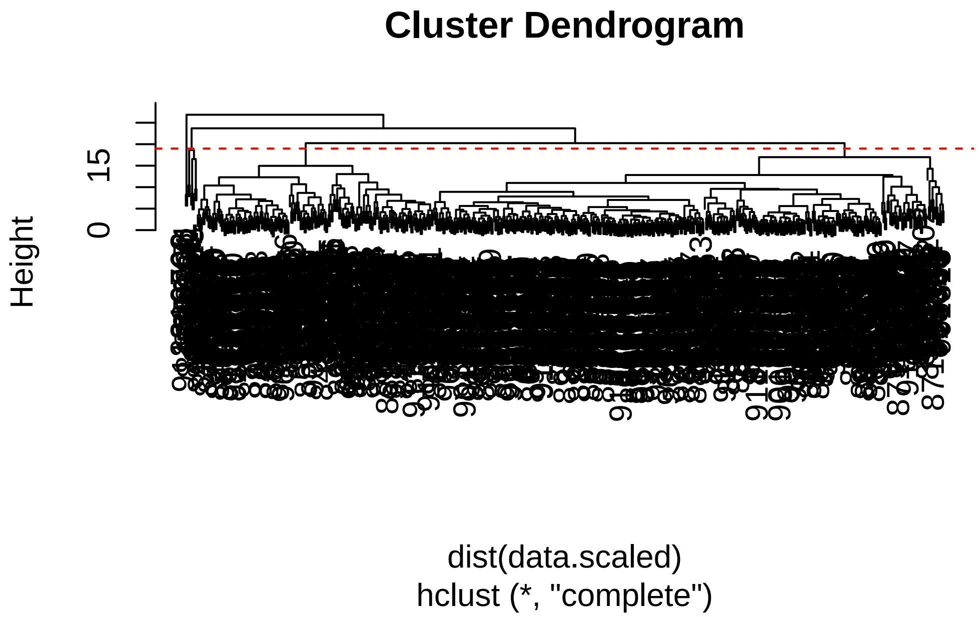
```
wisc.pr$rotation["concave.points_mean", 1]
```

```
[1] -0.2608538
```

Let's try clustering this data: ## **3.Hierarchical Clustering** > **Q10.** Using the plot() and abline() functions, what is the height at which the clustering model has 4 clusters?

A10. At height of around 19, the clustering model has 4 clusters

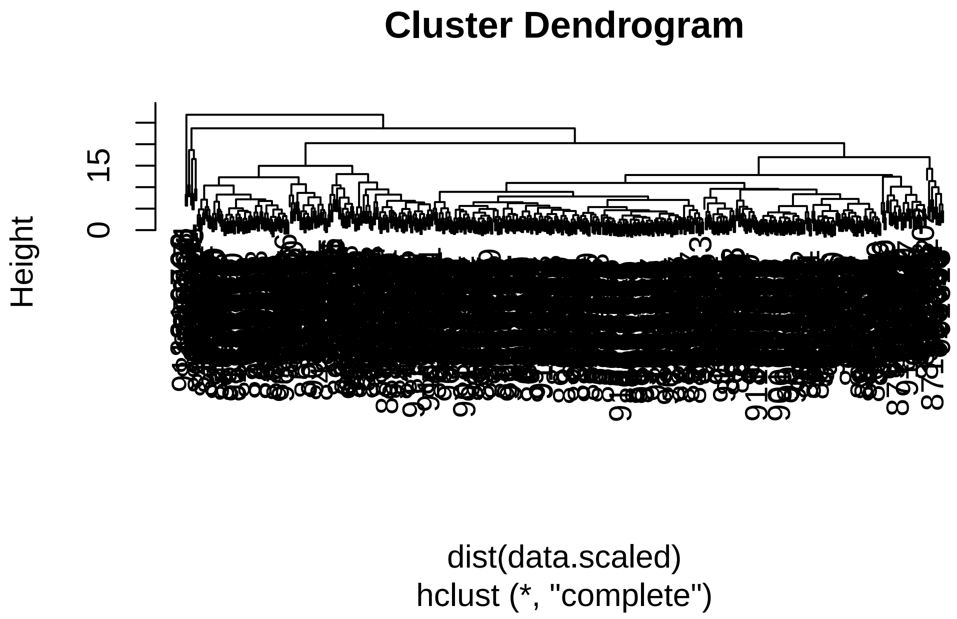
```
data.scaled <- scale(wisc.data)
wisc.hc <- hclust(dist(data.scaled))
plot(wisc.hc)
abline(h=19, col="red", lty=2)
```



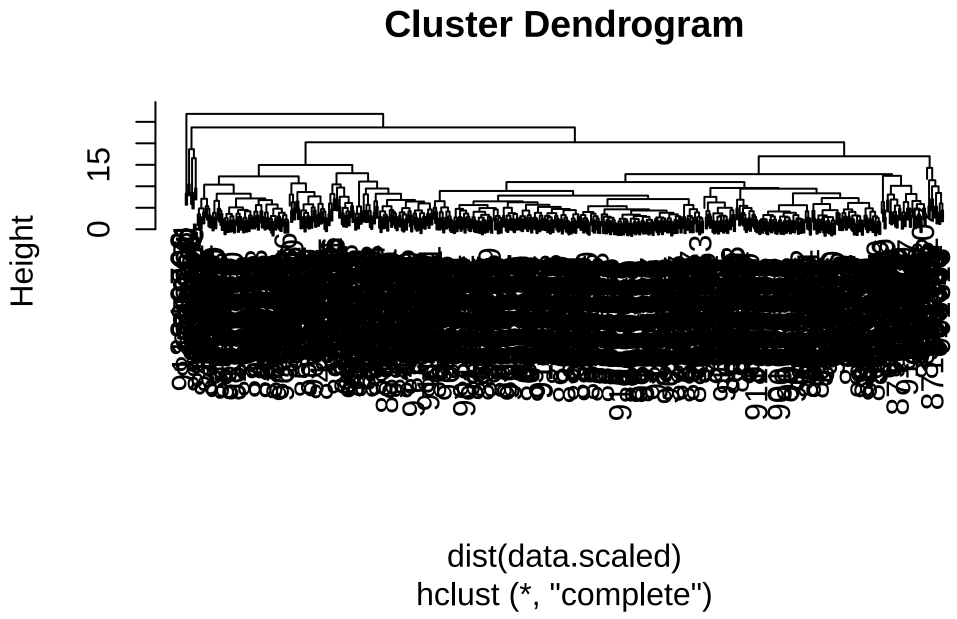
Q12. Which method gives your favorite results for the same data.dist dataset? Explain your reasoning.

A12. "ward.D2" method gives me favorite results. Because it minimizes the total within-cluster variance, aiming to create compact, spherical clusters.It is particularly effective when clusters are assumed to be spherical and evenly sized.

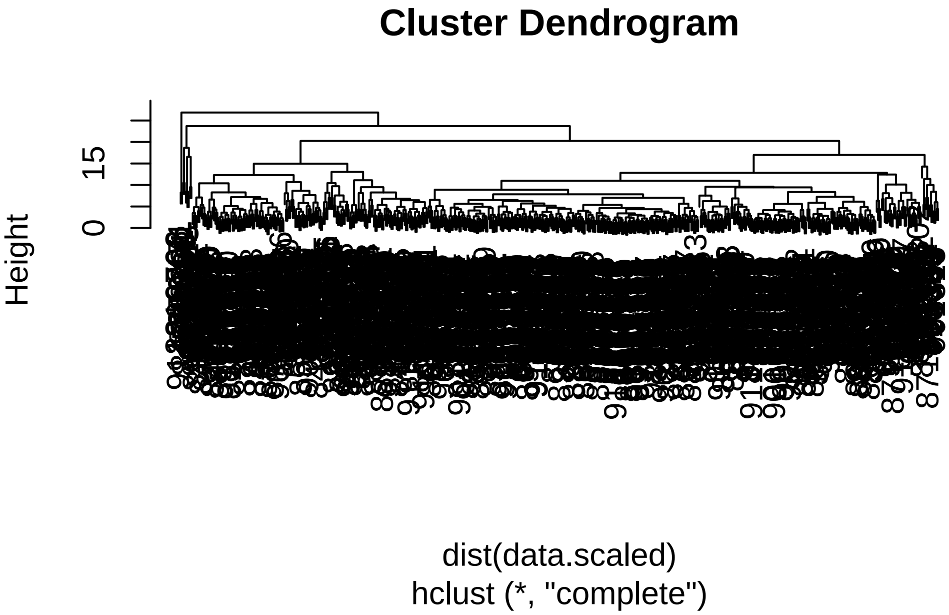
```
wisc.hc.2 <- hclust(dist(data.scaled), method = "single")
plot(wisc.hc)
```



```
wisc.hc.3 <- hclust(dist(data.scaled), method = "average")
plot(wisc.hc)
```



```
wisc.hc.4 <- hclust(dist(data.scaled), method = "ward.D2")
plot(wisc.hc)
```

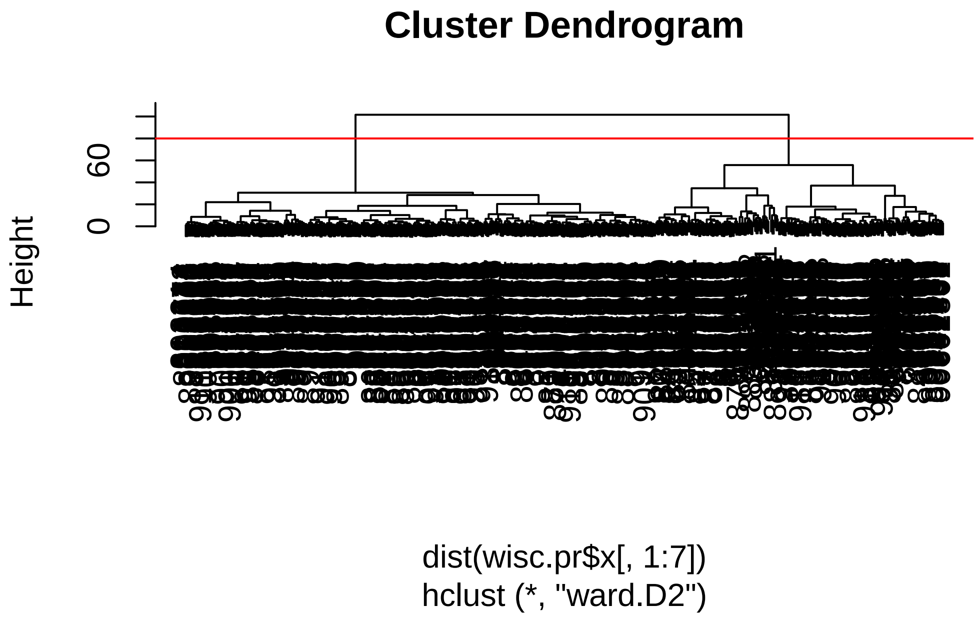


4.Combining methods

Here we will use the results of PCA as the input to a clustering analysis.

```
wisc.pr.hclust <- hclust(dist(wisc.pr$x[,1:7]), method = "ward.D2")
```

```
plot(wisc.pr.hclust)
abline(h=80, col="red")
```



```
wisc.pr.hclust.clusters <- cutree(wisc.pr.hclust, k=2)
```

```
wisc.hclust.clusters <- cutree(wisc.hc, k=4)
```

Q13. How well does the newly created model with four clusters separate out the two diagnoses?

A13. Ideally, I want to see clusters that are dominated by one diagnosis, indicating a clear separation. Both Cluster 1 and Cluster 2 have a mix of both diagnoses but are skewed more towards M or B diagnosis, which means the new model can roughly separate out the two diagnoses.

Compare to actual diagnoses

```
table(wisc.pr.hclust.clusters, diagnosis)
```

```
                       diagnosis
wisc.pr.hclust.clusters   B   M
                      1  28 188
                      2 329  24
```

```
wisc.km <- kmeans(wisc.data, centers = 4)
```

> **Q14.** How well do the hierarchical clustering models you created in previous sections (i.e. before PCA) do in terms of separating the diagnoses? Again, use the table() function to compare the output of each model (wisc.km$cluster and wisc.hclust.clusters) with the vector containing the actual diagnoses.

> **A14.** The hierarchical clustering models I created in previous sections do not work perfectly in terms of separating the diagnoses. Most clusters have a mix of both diagnoses although with a higher count of either M or B cases.

```
table(wisc.km$cluster, diagnosis)
```

```
   diagnosis
      B   M
  1   1 100
  2 262   6
  3  94  87
  4   0  19
```

```
table(wisc.hclust.clusters, diagnosis)
```

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

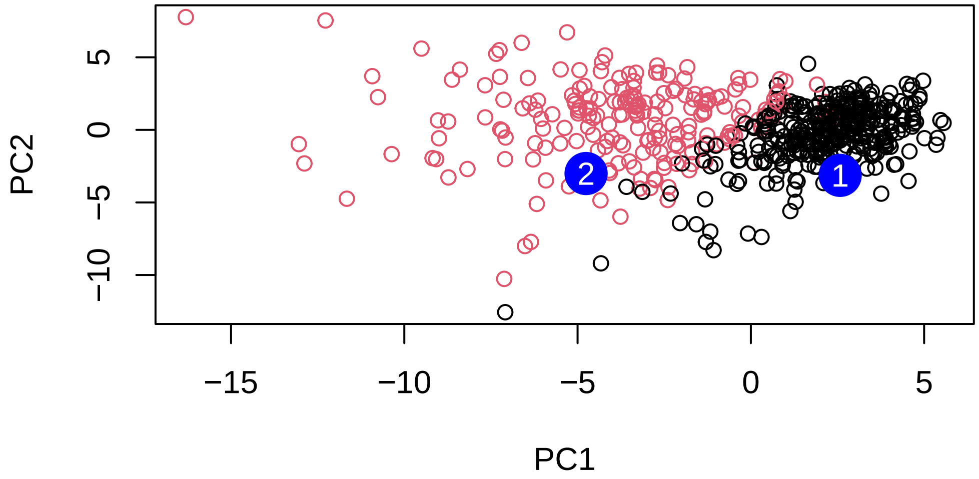
6.Prediction

#url <- “new_samples.csv”

```
url <- "https://tinyurl.com/new-samples-CSV"
new <- read.csv(url)
npc <- predict(wisc.pr, newdata=new)
npc
```

```
           PC1       PC2        PC3        PC4       PC5        PC6        PC7
[1,]  2.576616 -3.135913  1.3990492 -0.7631950  2.781648 -0.8150185 -0.3959098
[2,] -4.754928 -3.009033 -0.1660946 -0.6052952 -1.140698 -1.2189945  0.8193031
            PC8       PC9       PC10      PC11      PC12      PC13     PC14
[1,] -0.2307350 0.1029569 -0.9272861 0.3411457  0.375921 0.1610764 1.187882
[2,] -0.3307423 0.5281896 -0.4855301 0.7173233 -1.185917 0.5893856 0.303029
          PC15       PC16        PC17        PC18        PC19       PC20
[1,] 0.3216974 -0.1743616 -0.07875393 -0.11207028 -0.08802955 -0.2495216
[2,] 0.1299153  0.1448061 -0.40509706  0.06565549  0.25591230 -0.4289500
           PC21       PC22       PC23       PC24        PC25         PC26
[1,]  0.1228233 0.09358453 0.08347651  0.1223396  0.02124121  0.078884581
[2,] -0.1224776 0.01732146 0.06316631 -0.2338618 -0.20755948 -0.009833238
             PC27        PC28         PC29         PC30
[1,]  0.220199544 -0.02946023 -0.015620933  0.005269029
[2,] -0.001134152  0.09638361  0.002795349 -0.019015820
```

```r
plot(wisc.pr$x[,1:2], col=diagnosis)
points(npc[,1], npc[,2], col="blue", pch=16, cex=3)
text(npc[,1], npc[,2], c(1,2), col="white")
```



Q16. Which of these new patients should we prioritize for follow up based on your results?

A16. Patient 2 should be prioritized for follow up.Because all the sample data from patient 2 falls into cluster 1 for benign diagnosis ##