class09_mini-project

Rui Huang (PID: A15606522)

10/26/2021

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

```
wisc.data <- wisc.df[,-1]
diagnosis <- as.factor(wisc.df[,1])
```

#Q1: there are 569 observations

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

```
## [1] 569  30
```

#Q2: 212 observations of malignant diagnosis

```
table(diagnosis)
```

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

#Q3:there is 10 variables are sufficed with mean

```
length(grep("mean", colnames(wisc.df)))
```

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

```
colMeans(wisc.data)
```

```
##             radius_mean            texture_mean          perimeter_mean
##            1.412729e+01            1.928965e+01            9.196903e+01
##               area_mean         smoothness_mean        compactness_mean
##            6.548891e+02            9.636028e-02            1.043410e-01
##          concavity_mean     concave.points_mean           symmetry_mean
##            8.879932e-02            4.891915e-02            1.811619e-01
##  fractal_dimension_mean               radius_se              texture_se
##            6.279761e-02            4.051721e-01            1.216853e+00
##            perimeter_se                 area_se           smoothness_se
##            2.866059e+00            4.033708e+01            7.040979e-03
```

```
##          compactness_se            concavity_se       concave.points_se
##            2.547814e-02            3.189372e-02            1.179614e-02
##             symmetry_se    fractal_dimension_se            radius_worst
##            2.054230e-02            3.794904e-03            1.626919e+01
##           texture_worst         perimeter_worst              area_worst
##            2.567722e+01            1.072612e+02            8.805831e+02
##        smoothness_worst       compactness_worst         concavity_worst
##            1.323686e-01            2.542650e-01            2.721885e-01
##    concave.points_worst          symmetry_worst fractal_dimension_worst
##            1.146062e-01            2.900756e-01            8.394582e-02
```

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

```
##             radius_mean            texture_mean          perimeter_mean
##            3.524049e+00            4.301036e+00            2.429898e+01
##               area_mean         smoothness_mean        compactness_mean
##            3.519141e+02            1.406413e-02            5.281276e-02
##          concavity_mean     concave.points_mean           symmetry_mean
##            7.971981e-02            3.880284e-02            2.741428e-02
##  fractal_dimension_mean               radius_se              texture_se
##            7.060363e-03            2.773127e-01            5.516484e-01
##            perimeter_se                 area_se           smoothness_se
##            2.021855e+00            4.549101e+01            3.002518e-03
##          compactness_se            concavity_se       concave.points_se
##            1.790818e-02            3.018606e-02            6.170285e-03
##             symmetry_se    fractal_dimension_se            radius_worst
##            8.266372e-03            2.646071e-03            4.833242e+00
##           texture_worst         perimeter_worst              area_worst
##            6.146258e+00            3.360254e+01            5.693570e+02
##        smoothness_worst       compactness_worst         concavity_worst
##            2.283243e-02            1.573365e-01            2.086243e-01
##    concave.points_worst          symmetry_worst fractal_dimension_worst
##            6.573234e-02            6.186747e-02            1.806127e-02
```

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

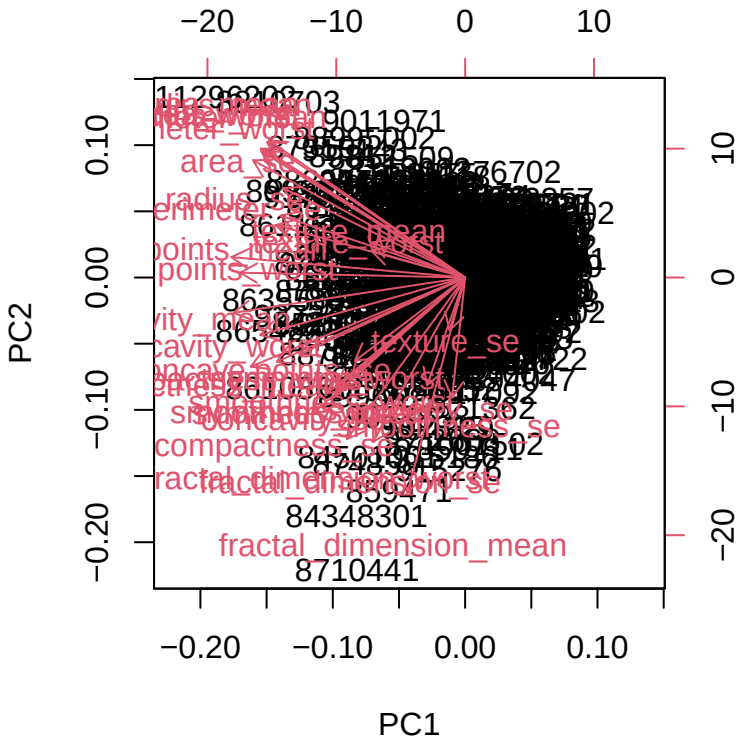
```
## Cumulative Proportion  0.99749 0.99830 0.9989 0.99942 0.99969 0.99992 0.99997
##                            PC29    PC30
## Standard deviation     0.02736 0.01153
## Proportion of Variance 0.00002 0.00000
## Cumulative Proportion  1.00000 1.00000
```

#Q4:the original variance captured by the PC1 has a proportion of 0.4427.

#Q5: 3 principle components are required.

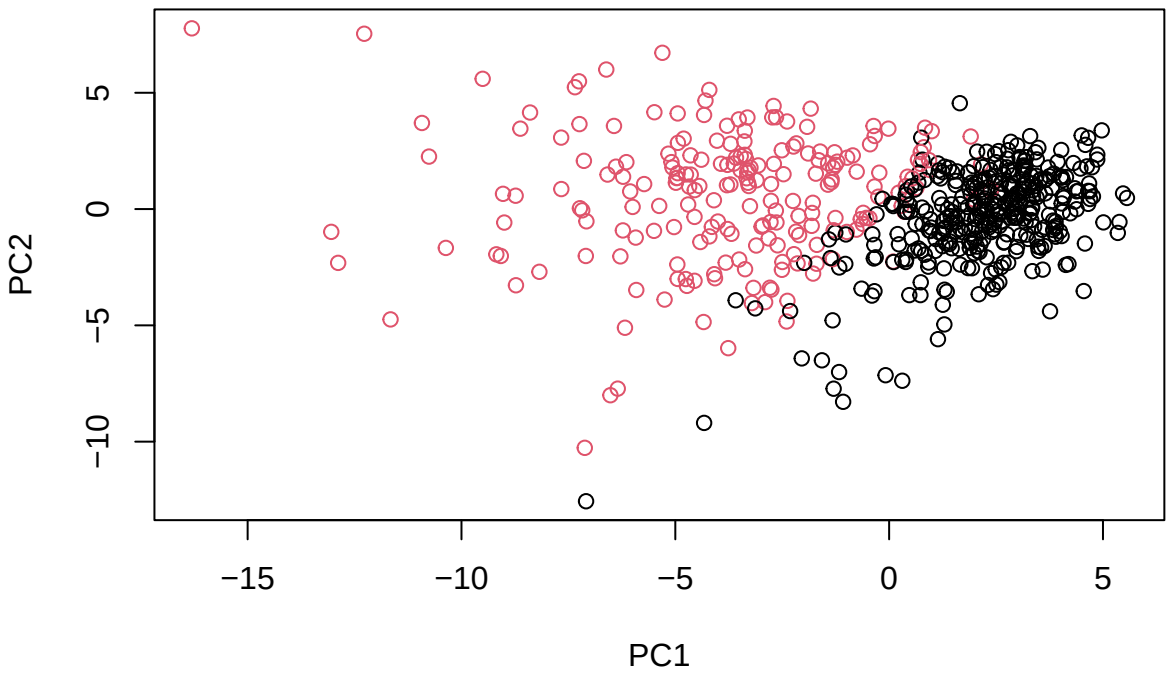
#Q6: 7 principle components are required.

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



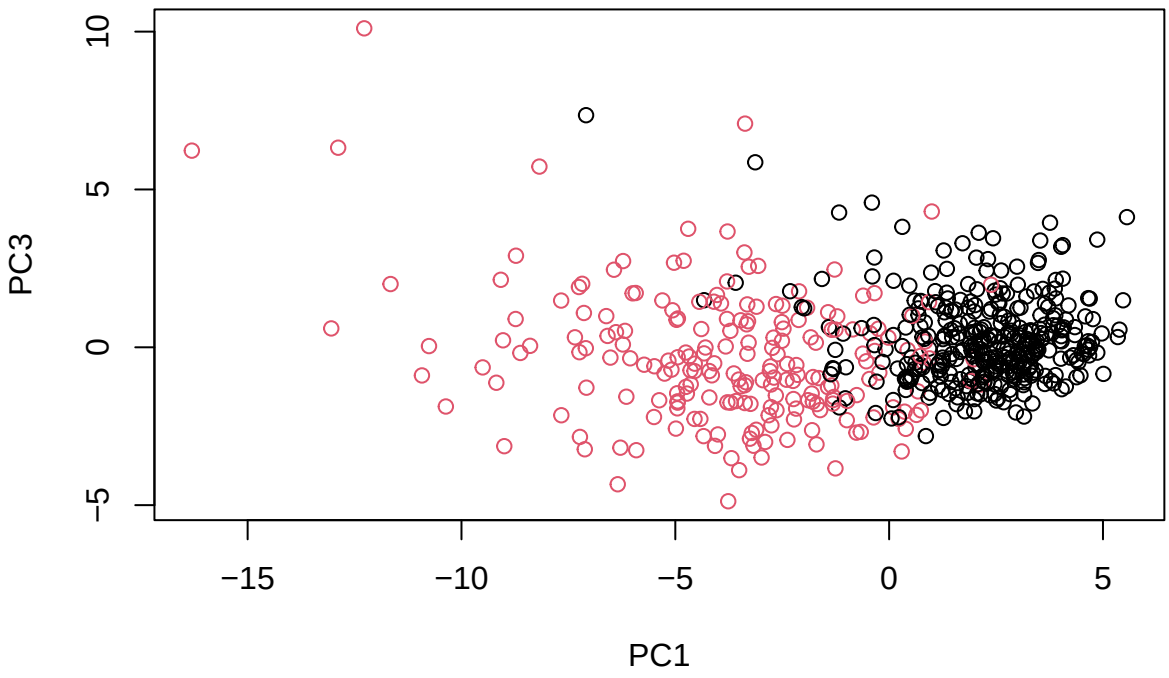
#Q7: the plot is crowded as there are way too many data on it, it is hard to understand since there is too much information with many overlapping scripts.

```
plot(wisc.pr$x[,1:2], col=diagnosis)
```

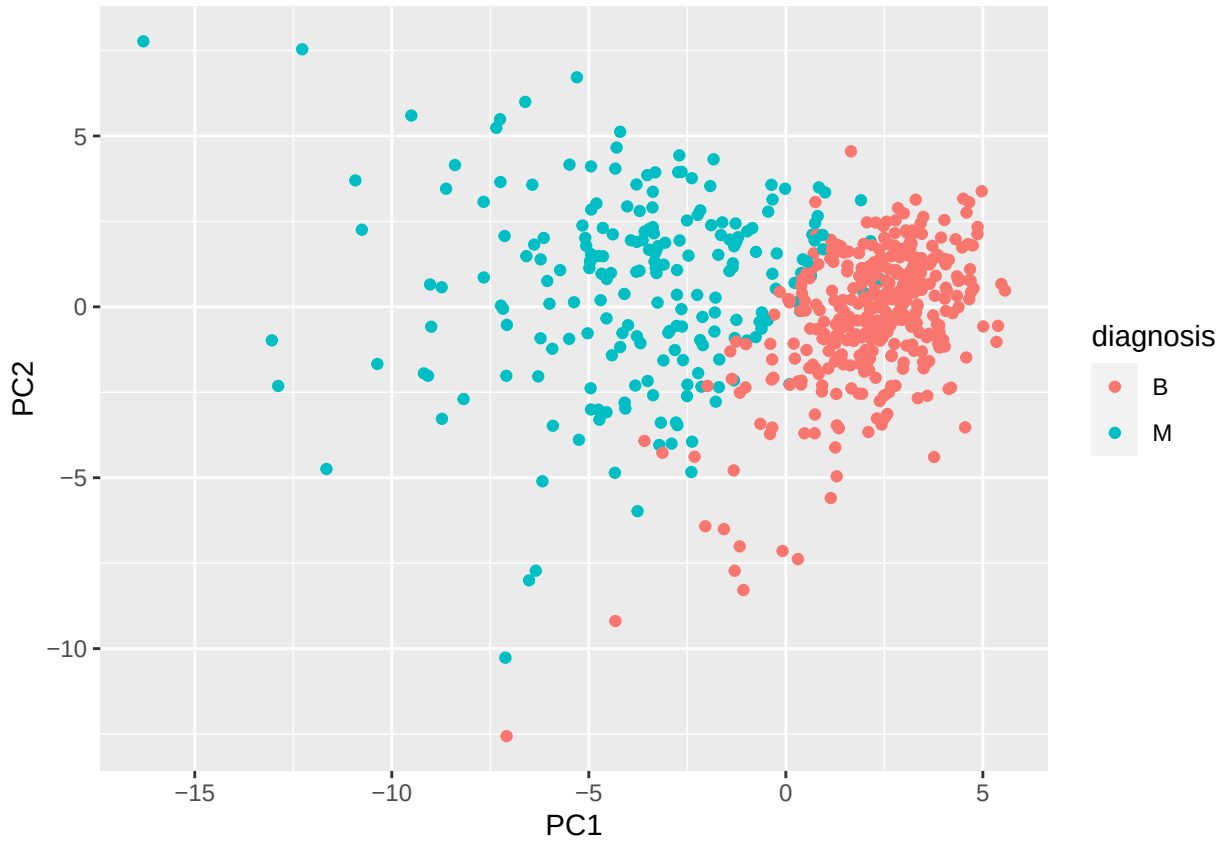


#Q8:the shown y-axis window range is different, and also the distribution of the dots are different.

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



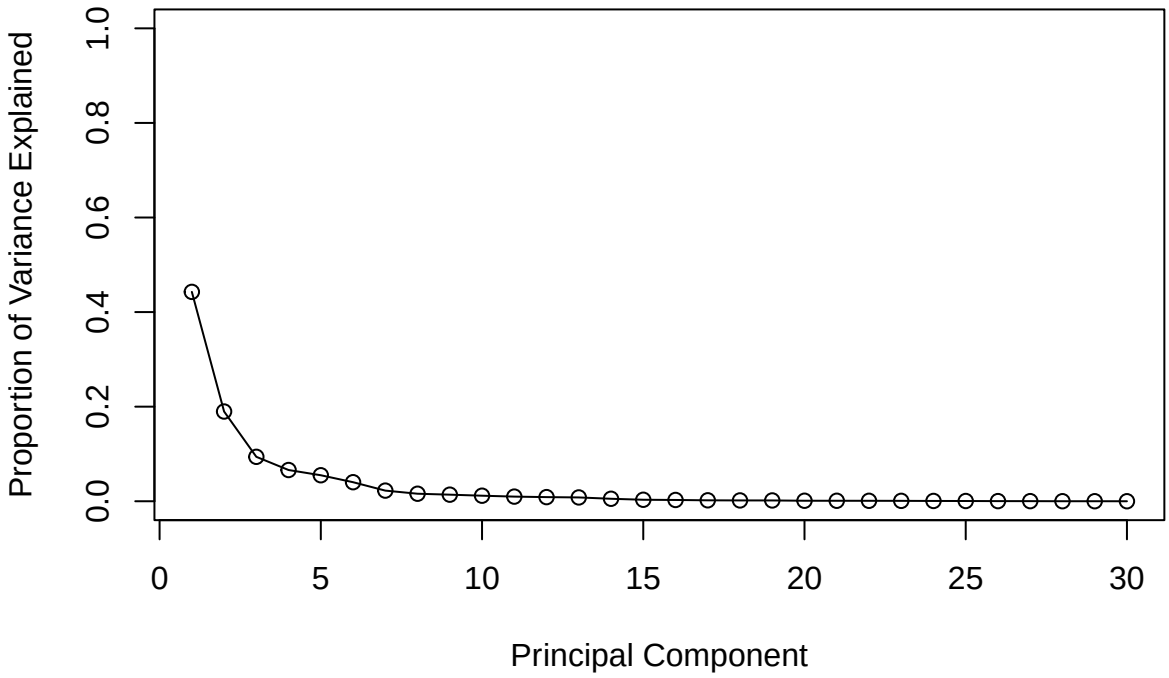

```
df<- as.data.frame(wisc.pr$x)
df$diagnosis<- diagnosis
library(ggplot2)
ggplot(df)+aes(PC1,PC2, col=diagnosis)+geom_point()
```



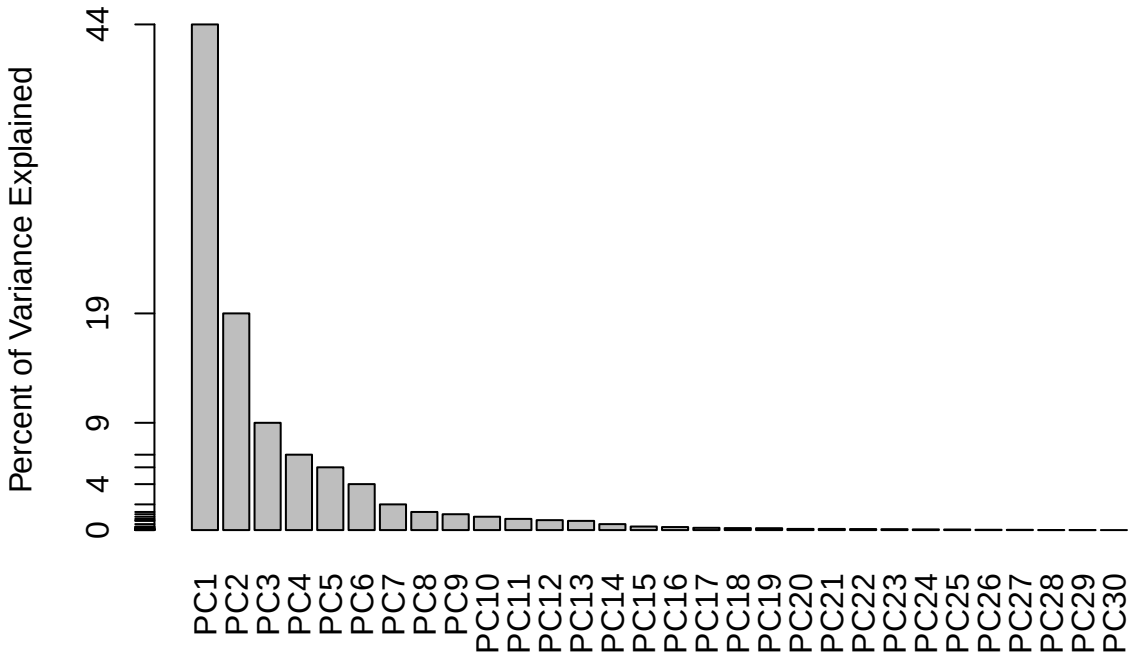

```
pr.var <- wisc.pr$sdev^2
head(pr.var)
```

```
## [1] 13.281608  5.691355  2.817949  1.980640  1.648731  1.207357
```

```
pve<-pr.var/sum(pr.var)
plot(pve,xlab="Principal Component", ylab="Proportion of Variance Explained", ylim=c(0,1), type="o")
```



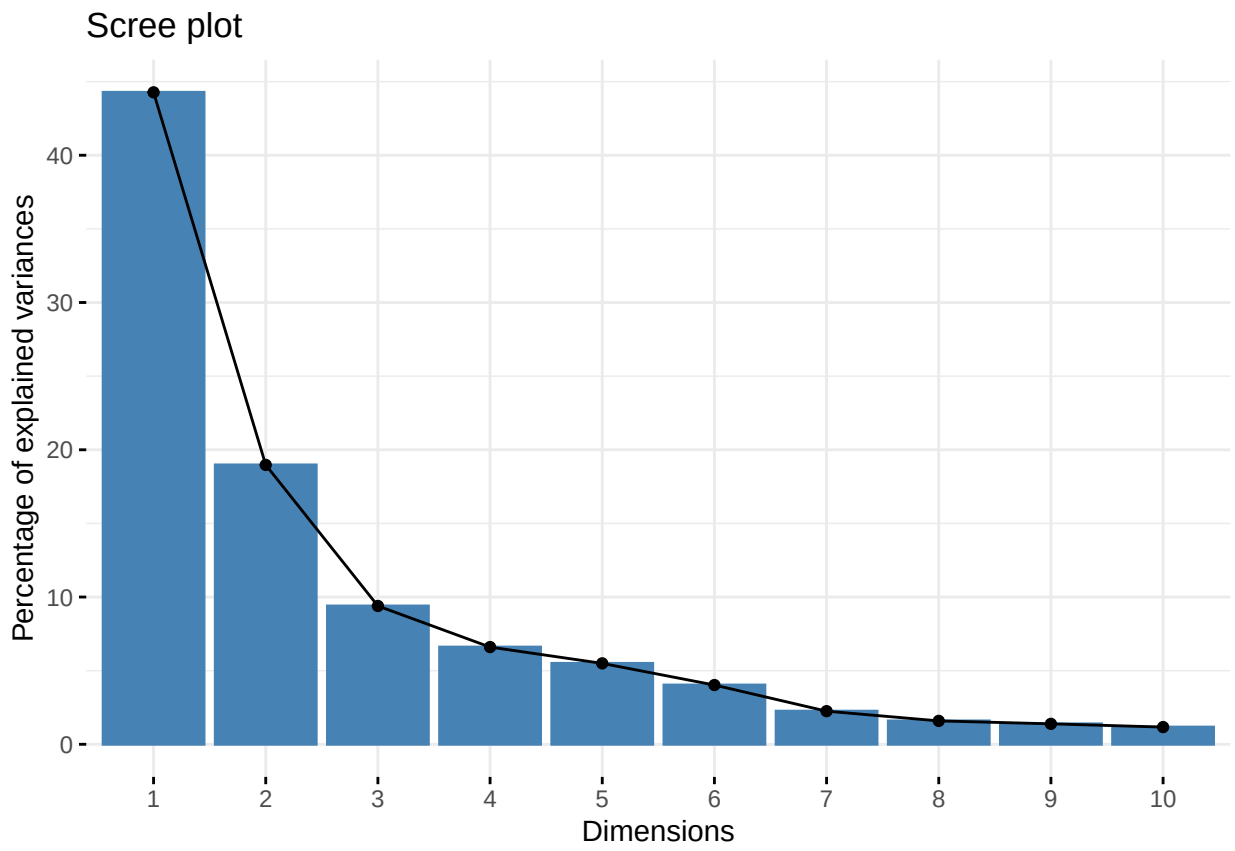

```
barplot(pve,ylab="Percent of Variance Explained", names.arg=paste0("PC",1:length(pve)),las=2,axes=F)
axis(2, at=pve, labels=round(pve,2)*100)
```




```
library(factoextra)
```

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

```
fviz_eig(wisc.pr, addlables=T)
```



#Q9: it is -0.2608538

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

```
## [1] -0.2608538
```

#Q10: the minimum is 5

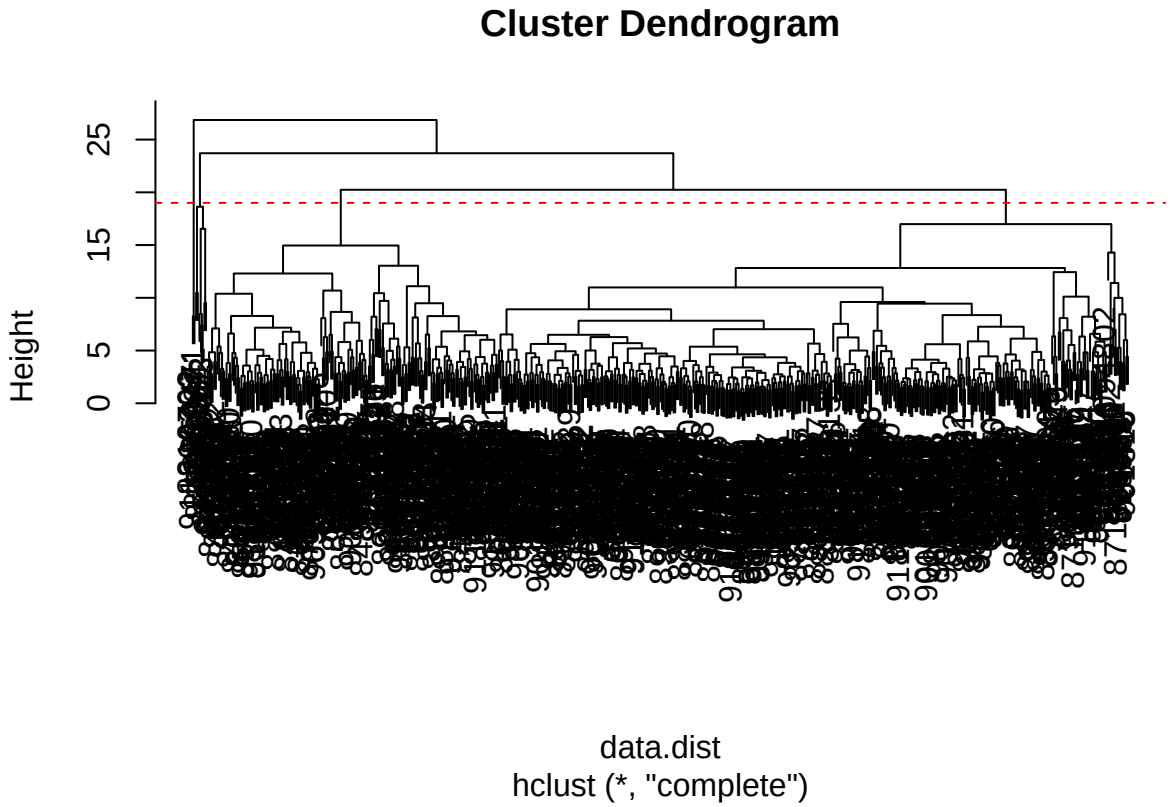
```
var<-summary(wisc.pr)
var$importance[3,]>=0.8
```

```
##   PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10  PC11  PC12  PC13
## FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  PC14  PC15  PC16  PC17  PC18  PC19  PC20  PC21  PC22  PC23  PC24  PC25  PC26
##  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##  PC27  PC28  PC29  PC30
##  TRUE  TRUE  TRUE  TRUE
```

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

#Q11:at the height of 19, the clustering model has 4 clusters.

```
plot(wisc.hclust)
abline(h=19,col="red",lty=2)
```




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

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

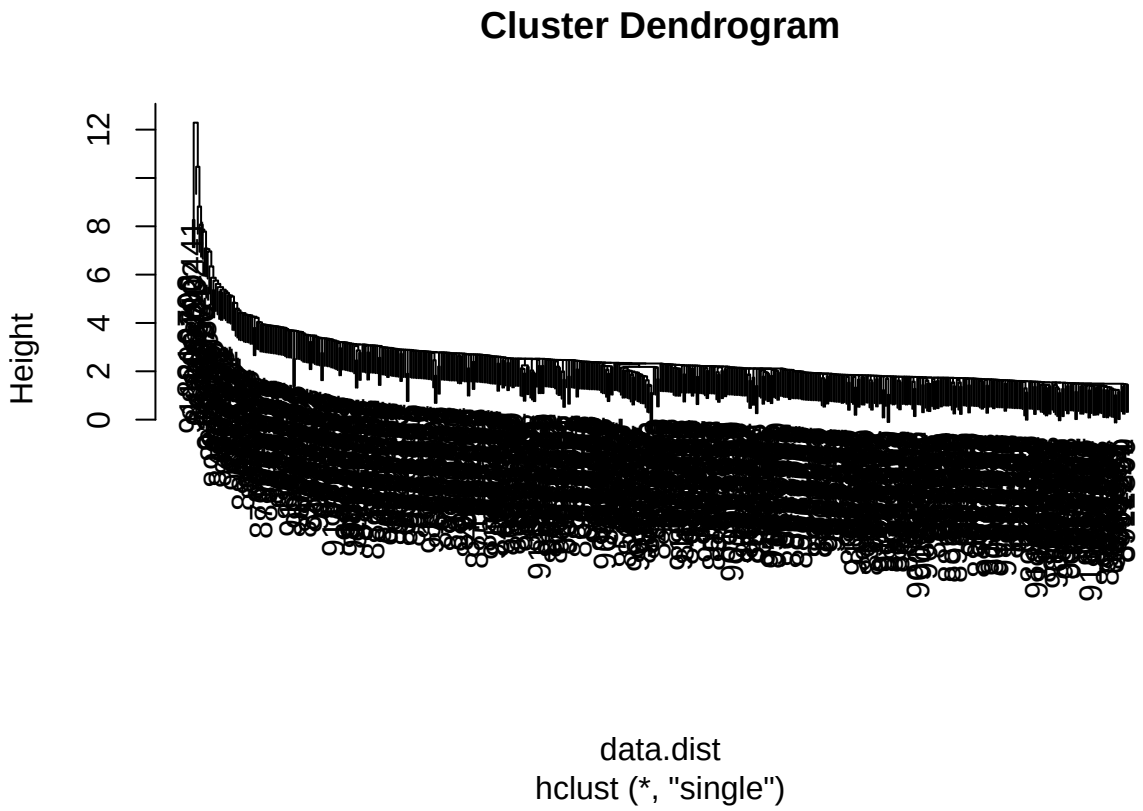
#Q12:cutting them into 2 clusters is a better way.

```
wisc.hclust.clusters1<- cutree(wisc.hclust, k=2)
table(wisc.hclust.clusters1,diagnosis)
```

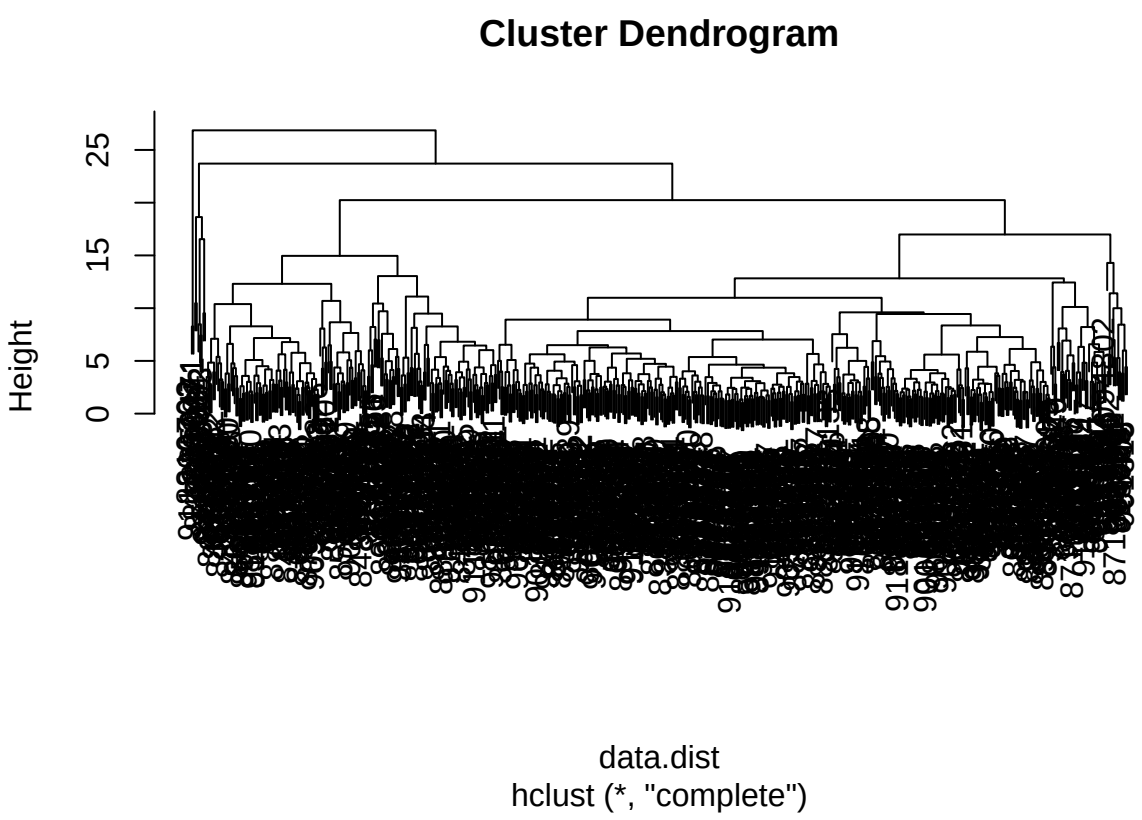
```
##                      diagnosis
## wisc.hclust.clusters1   B   M
##                     1 357 210
##                     2   0   2
```

#Q13:My favorite method is “word.D2”, because the clusters branching corresponds to greater height range, so that it is better presented and organized and it is easier to see the precise height of the branching clusters.

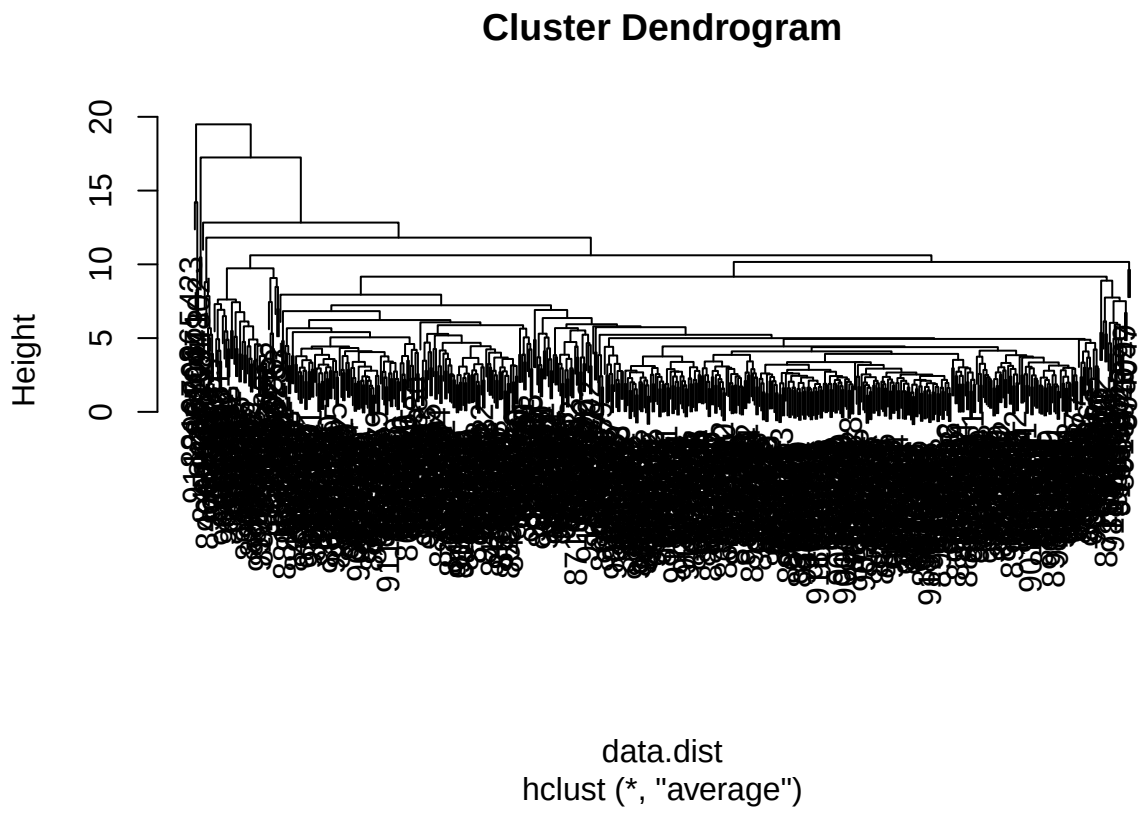
```
hc.single<-hclust(data.dist, method="single")
hc.complete<-hclust(data.dist, method="complete")
hc.average<-hclust(data.dist, method="average")
hc.ward<-hclust(data.dist, method="ward.D2")
plot(hc.single)
```



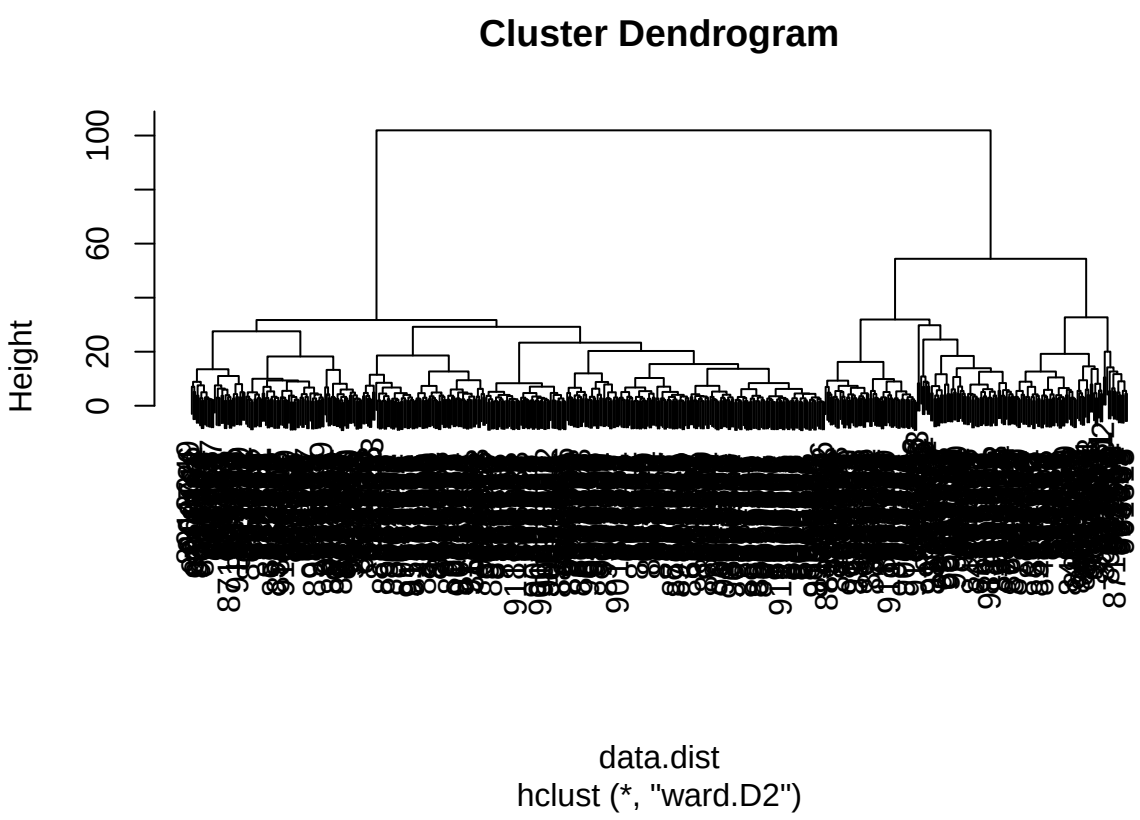

```
plot(hc.complete)
```




```
plot(hc.average)
```


Cluster Dendrogram
Height
0
5
10
15
20
data.dist
hclust (*, "average")

```
plot(hc.ward)
```



#kmeans

```
wisc1<-scale(wisc.data, center=T,scale=T)
wisc.km<- kmeans(wisc1,centers=2, nstart=20)
table(wisc.km$cluster, diagnosis)
```

```
##    diagnosis
##       B   M
##   1 343  37
##   2  14 175
```

```
table(wisc.hclust.clusters,wisc.km$cluster)
```

```
##
## wisc.hclust.clusters   1   2
##                    1  17 160
##                    2   0   7
##                    3 363  20
##                    4   0   2
```

#Q14: k-means separate the two diagnosis well, its separation of 2 diagnosis is as good as hclust result.

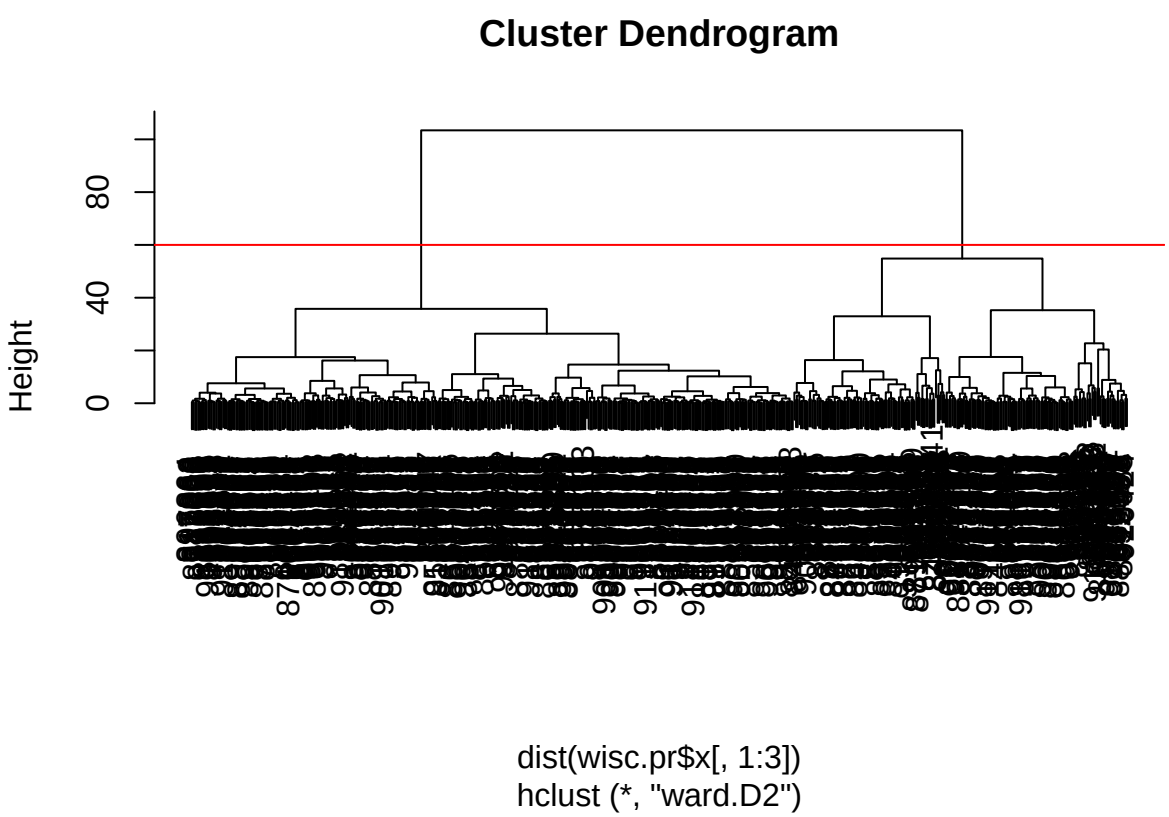
#5.Combing methods

```
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
```

```
wisc.pc.hclust<-hclust(dist(wisc.pr$x[,1:3]),
                       method="ward.D2")
```

```
plot(wisc.pc.hclust)
abline(h=60,col="red")
```




```
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
```

```
wisc.pr.hclust<-hclust(dist(wisc.pr$x[,1:7]),
                       method="ward.D2")
wisc.pr.hclust.cluster <- cutree(wisc.pr.hclust, k=2)
```

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

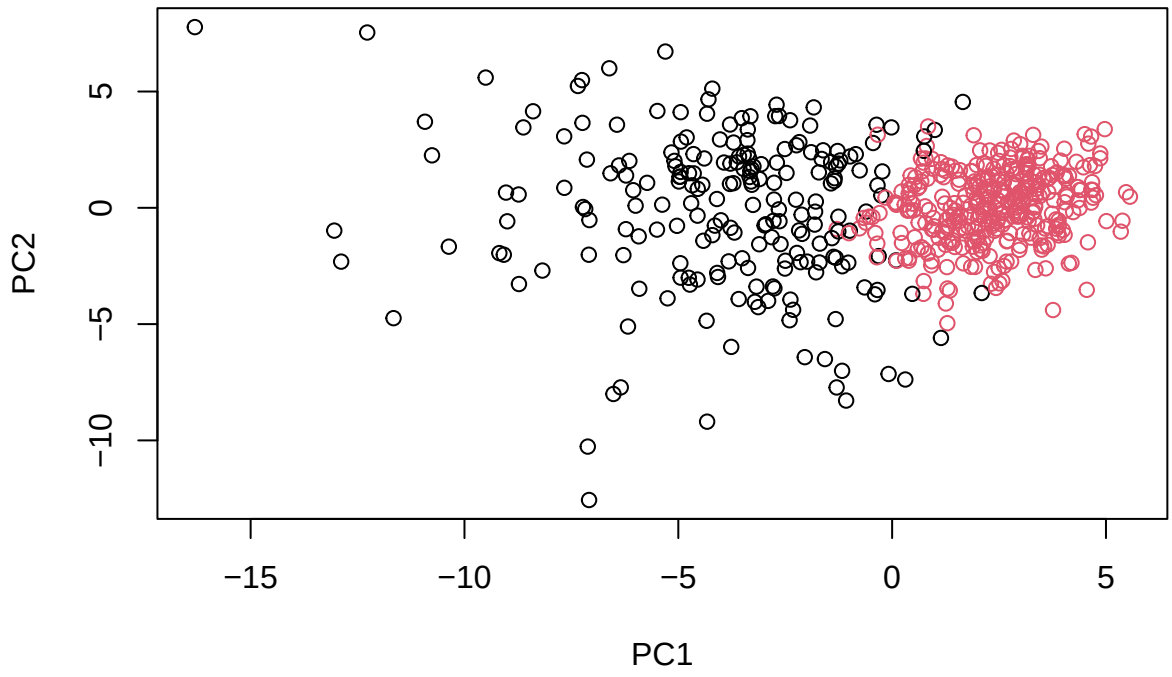
```
## grps
##   1   2
## 216 353
```

#cross table compare of diagnosis and my cluster group

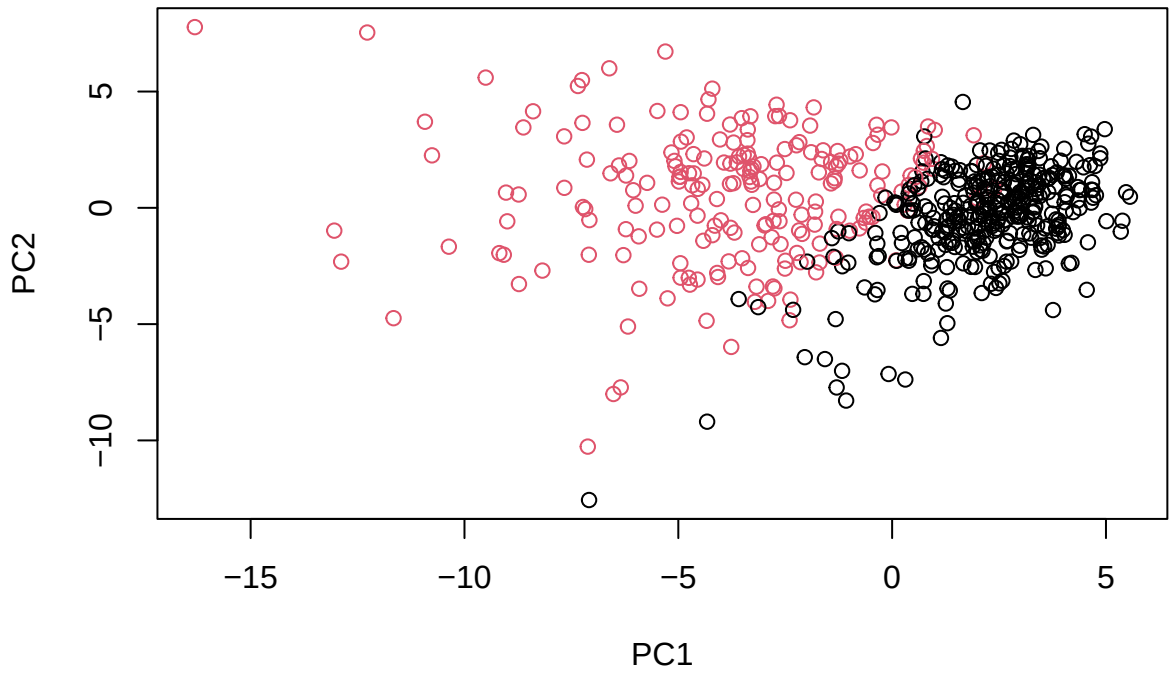
```
table(diagnosis,grps)
```

```
##          grps
## diagnosis   1   2
##         B  28 329
##         M 188  24
```

```
plot(wisc.pr$x[,1:2], col=grps)
```



```
plot(wisc.pr$x[,1:2], col=diagnosis)
```



#Q15: the separation of diagnosis outcomes is good,as the cluster 1 mainly corresponds to malignant, and cluster 2 corresponds to benign diagnosis.

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

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

#Q16:Both the k-means and the hierarchial clustering separate the diagnosis well, and both of them are eqaully good in terms of separating the diagnosis.

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

```
##    diagnosis
##       B   M
##   1 343  37
##   2  14 175
```

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

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

#sensitivity

```
188/(188+24)
```

```
## [1] 0.8867925
```

```
175/(175+37)
```

```
## [1] 0.8254717
```

```
165/(165+47)
```

```
## [1] 0.7783019
```

#specificity

```
329/(329+28)
```

```
## [1] 0.9215686
```

```
343/(343+14)
```

```
## [1] 0.9607843
```

```
343/(343+14)
```

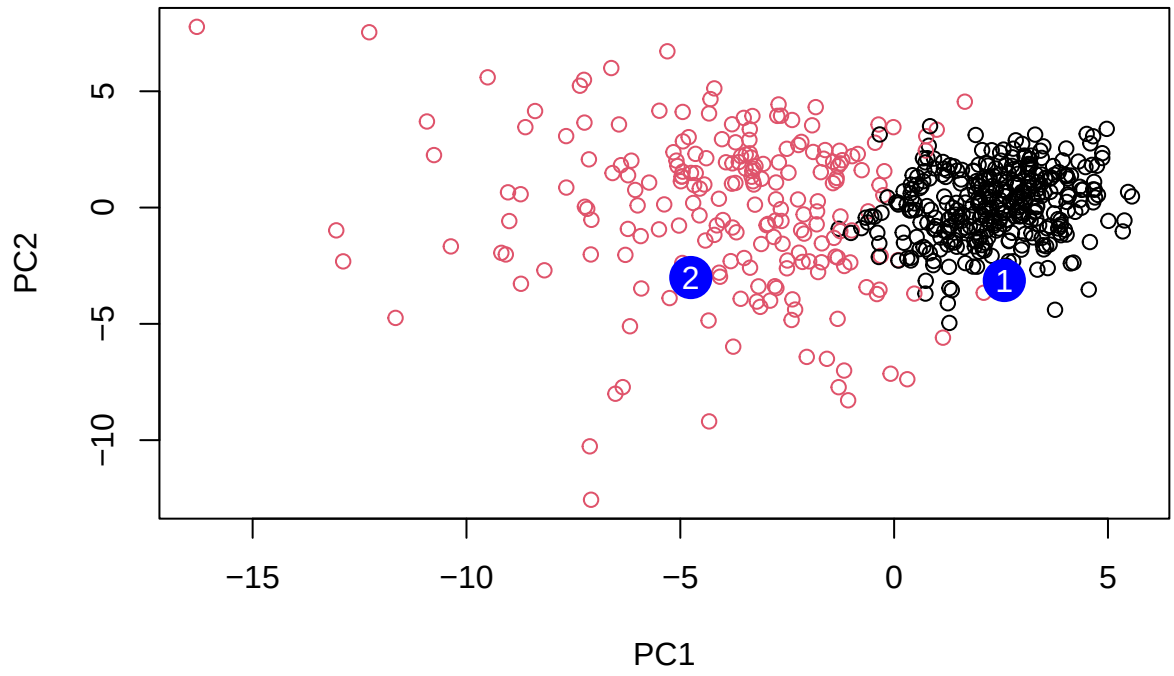
```
## [1] 0.9607843
```

#Q17:wisc.pr.hclust.cluster' result has better sensitivity, k means and wisc.hclust.clusters' result have better specificity.

```
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
```

```
g <- as.factor(grps)
g <- relevel(g,2)
plot(wisc.pr$x[,1:2], col=g)
points(npc[,1], npc[,2], col="blue", pch=16, cex=3)
text(npc[,1], npc[,2], c(1,2), col="white")
```



#Q18: we should prioritize patient 2.

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
```

```
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] factoextra_1.0.7 ggplot2_3.3.5
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.1  xfun_0.26         purrr_0.3.4       haven_2.4.3
##  [5] carData_3.0-4     colorspace_2.0-2  vctrs_0.3.8       generics_0.1.1
##  [9] htmltools_0.5.2   yaml_2.2.1        utf8_1.2.2        rlang_0.4.11
## [13] pillar_1.6.3      ggpubr_0.4.0      foreign_0.8-81    glue_1.4.2
## [17] withr_2.4.2       readxl_1.3.1      lifecycle_1.0.1   stringr_1.4.0
## [21] cellranger_1.1.0  munsell_0.5.0     ggsignif_0.6.3    gtable_0.3.0
## [25] zip_2.2.0         evaluate_0.14     labeling_0.4.2    knitr_1.36
## [29] rio_0.5.27        forcats_0.5.1     fastmap_1.1.0     curl_4.3.2
## [33] fansi_0.5.0       highr_0.9         broom_0.7.9       Rcpp_1.0.7
## [37] scales_1.1.1      backports_1.2.1   abind_1.4-5       farver_2.1.0
## [41] hms_1.1.1         digest_0.6.28     stringi_1.7.5     openxlsx_4.2.4
## [45] rstatix_0.7.0     dplyr_1.0.7       ggrepel_0.9.1     grid_4.1.1
## [49] tools_4.1.1       magrittr_2.0.1    tibble_3.1.5      crayon_1.4.1
## [53] tidyr_1.1.4       car_3.0-11        pkgconfig_2.0.3   ellipsis_0.3.2
## [57] data.table_1.14.2 rmarkdown_2.11    R6_2.5.1          compiler_4.1.1
```