Class 8: PCA Mini Project

Pamelina Lo (PID: A16735368)

It is important to condiser scrolling our data before analsysis.

For example:

```
head(mtcars)
```

```
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
colMeans(mtcars)
```

```
       mpg        cyl       disp         hp       drat         wt       qsec 
 20.090625   6.187500 230.721875 146.687500   3.596563   3.217250  17.848750 
        vs         am       gear       carb 
  0.437500   0.406250   3.687500   2.812500 
```

```
apply(mtcars,2,sd)
```

```
       mpg        cyl       disp         hp       drat         wt 
 6.0269481  1.7859216 123.9386938  68.5628685  0.5346787  0.9784574 
      qsec         vs         am       gear       carb 
 1.7869432  0.5040161  0.4989909  0.7378041  1.6152000 
```

```
x <- scale(mtcars)
head(x)
```

```
                        mpg        cyl        disp         hp       drat
Mazda RX4          0.1508848 -0.1049878 -0.57061982 -0.5350928  0.5675137
Mazda RX4 Wag      0.1508848 -0.1049878 -0.57061982 -0.5350928  0.5675137
Datsun 710         0.4495434 -1.2248578 -0.99018209 -0.7830405  0.4739996
Hornet 4 Drive     0.2172534 -0.1049878  0.22009369 -0.5350928 -0.9661175
Hornet Sportabout -0.2307345  1.0148821  1.04308123  0.4129422 -0.8351978
Valiant           -0.3302874 -0.1049878 -0.04616698 -0.6080186 -1.5646078
                          wt       qsec         vs         am       gear
Mazda RX4         -0.610399567 -0.7771651 -0.8680278  1.1899014  0.4235542
Mazda RX4 Wag     -0.349785269 -0.4637808 -0.8680278  1.1899014  0.4235542
Datsun 710        -0.917004624  0.4260068  1.1160357  1.1899014  0.4235542
Hornet 4 Drive    -0.002299538  0.8904872  1.1160357 -0.8141431 -0.9318192
Hornet Sportabout  0.227654255 -0.4637808 -0.8680278 -0.8141431 -0.9318192
Valiant            0.248094592  1.3269868  1.1160357 -0.8141431 -0.9318192
                         carb
Mazda RX4           0.7352031
Mazda RX4 Wag       0.7352031
Datsun 710         -1.1221521
Hornet 4 Drive     -1.1221521
Hornet Sportabout  -0.5030337
Valiant            -1.1221521
```

```
round(colMeans(x),2)
```

```
 mpg  cyl disp   hp drat   wt qsec   vs   am gear carb
   0    0    0    0    0    0    0    0    0    0    0
```

Preparing the data

Values in this data set describe characterisitics of the cell nuclei present in digitized images of a fine needle aspiration of a breast mass.

```
# Save your input data file into your Project directory
fna.data <- "WisconsinCancer.csv"

# Complete the following code to input the data and store as wisc.df
wisc.df <- read.csv(fna.data,row.names=1)
```

```
head(wisc.df)
```

```
         diagnosis radius_mean texture_mean perimeter_mean area_mean
842302           M       17.99        10.38         122.80    1001.0
842517           M       20.57        17.77         132.90    1326.0
84300903         M       19.69        21.25         130.00    1203.0
84348301         M       11.42        20.38          77.58     386.1
84358402         M       20.29        14.34         135.10    1297.0
843786           M       12.45        15.70          82.57     477.1
         smoothness_mean compactness_mean concavity_mean concave.points_mean
842302           0.11840          0.27760         0.3001             0.14710
842517           0.08474          0.07864         0.0869             0.07017
84300903         0.10960          0.15990         0.1974             0.12790
84348301         0.14250          0.28390         0.2414             0.10520
84358402         0.10030          0.13280         0.1980             0.10430
843786           0.12780          0.17000         0.1578             0.08089
         symmetry_mean fractal_dimension_mean radius_se texture_se perimeter_se
842302          0.2419                0.07871    1.0950     0.9053        8.589
842517          0.1812                0.05667    0.5435     0.7339        3.398
84300903        0.2069                0.05999    0.7456     0.7869        4.585
84348301        0.2597                0.09744    0.4956     1.1560        3.445
84358402        0.1809                0.05883    0.7572     0.7813        5.438
843786          0.2087                0.07613    0.3345     0.8902        2.217
         area_se smoothness_se compactness_se concavity_se concave.points_se
842302    153.40      0.006399        0.04904      0.05373           0.01587
842517     74.08      0.005225        0.01308      0.01860           0.01340
84300903   94.03      0.006150        0.04006      0.03832           0.02058
84348301   27.23      0.009110        0.07458      0.05661           0.01867
84358402   94.44      0.011490        0.02461      0.05688           0.01885
843786     27.19      0.007510        0.03345      0.03672           0.01137
         symmetry_se fractal_dimension_se radius_worst texture_worst
842302       0.03003             0.006193        25.38         17.33
842517       0.01389             0.003532        24.99         23.41
84300903     0.02250             0.004571        23.57         25.53
84348301     0.05963             0.009208        14.91         26.50
84358402     0.01756             0.005115        22.54         16.67
843786       0.02165             0.005082        15.47         23.75
         perimeter_worst area_worst smoothness_worst compactness_worst
842302             184.60     2019.0           0.1622            0.6656
842517             158.80     1956.0           0.1238            0.1866
84300903           152.50     1709.0           0.1444            0.4245
```

```
84348301           98.87      567.7           0.2098           0.8663
84358402          152.20     1575.0           0.1374           0.2050
843786            103.40      741.6           0.1791           0.5249
          concavity_worst concave.points_worst symmetry_worst
842302             0.7119               0.2654         0.4601
842517             0.2416               0.1860         0.2750
84300903           0.4504               0.2430         0.3613
84348301           0.6869               0.2575         0.6638
84358402           0.4000               0.1625         0.2364
843786             0.5355               0.1741         0.3985
          fractal_dimension_worst
842302                    0.11890
842517                    0.08902
84300903                  0.08758
84348301                  0.17300
84358402                  0.07678
843786                    0.12440
```

```
diagnosis <-wisc.df[,1]
table(diagnosis)
```

```
diagnosis
  B   M
357 212
```

Remove this first `diagnosis` column from the data set because I don't want to pass this to PCS etc. It is essentially the expert "answer" that we will compare our analysis results to.

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

Exploratory data analysis

Q1. How many observations are in this dataset?

```
ncol(wisc.df)
```

```
[1] 31
```

There are 31 observations in this dataset.

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

```
table(diagnosis)
```

```
diagnosis
  B   M
357 212
```

There are 212 observations have a malignant diagnosis.

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

```
length(grep("_mean", colnames(wisc.data), value =1 ))
```

```
[1] 10
```

Performing PCA

```
# Check column means and standard deviations
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
```

```
# Perform PCA on wisc.data by completing the following code
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
```

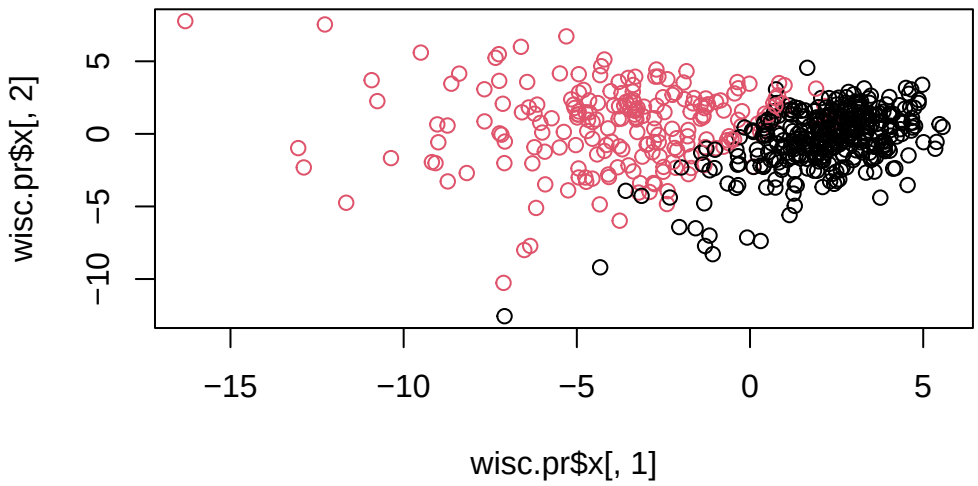
```
Standard deviation     0.16565 0.15602 0.1344 0.12442 0.09043 0.08307 0.03987
Proportion of Variance 0.00091 0.00081 0.0006 0.00052 0.00027 0.00023 0.00005
Cumulative Proportion  0.99749 0.99830 0.9989 0.99942 0.99969 0.99992 0.99997
                          PC29    PC30
Standard deviation     0.02736 0.01153
Proportion of Variance 0.00002 0.00000
Cumulative Proportion  1.00000 1.00000
```

Main "PC score plot", "PC1 vs PC2 plot" PCA result object:

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

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

$class
[1] "prcomp"
```

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



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

```
pca_summary <- summary(wisc.pr)
prop_var <- pca_summary$importance[2,]
```

```
cat(prop_var[1])
```

```
0.44272
```

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

```
pca_result <- summary(wisc.pr)
cum_var_explained <- cumsum(pca_result$sdev^2) / sum(pca_result$sdev^2)

num_pcs_70 <- which(cum_var_explained >= 0.70)[1]

cat("Number of PCs required to explain at least 70% of the variance:", num_pcs_70, "\n")
```

```
Number of PCs required to explain at least 70% of the variance: 3
```

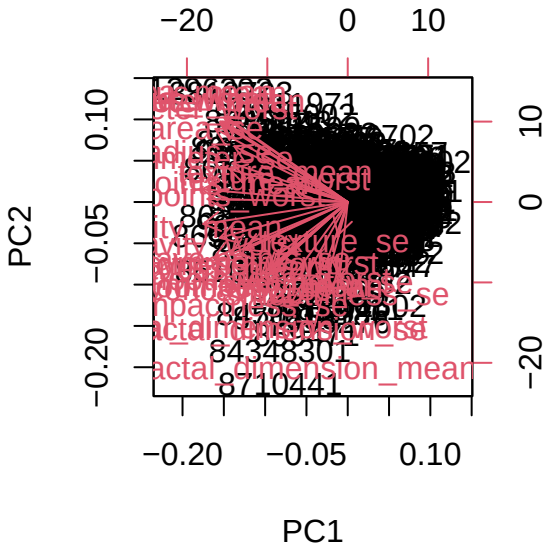
Q6. How many principal components (PCs) are required to describe at least 90% of the original variance in the data?

```
pca_result <- summary(wisc.pr)
cum_var_explained <- cumsum(pca_result$sdev^2) / sum(pca_result$sdev^2)

num_pcs_90 <- which(cum_var_explained >= 0.90)[1]

cat("Number of PCs required to explain at least 90% of the variance:", num_pcs_90, "\n")
```

```
Number of PCs required to explain at least 90% of the variance: 7
```

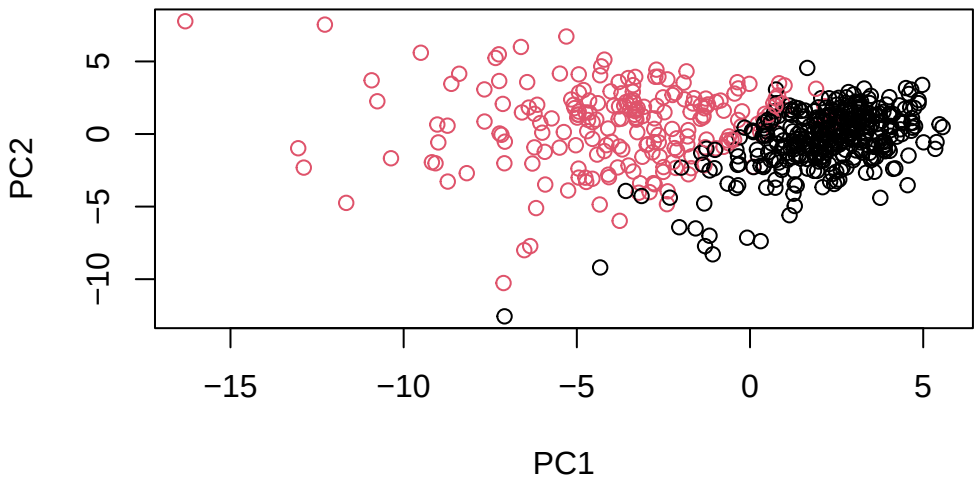
Q7. What stands out to you about this plot? Is it easy or difficult to understand? Why?

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



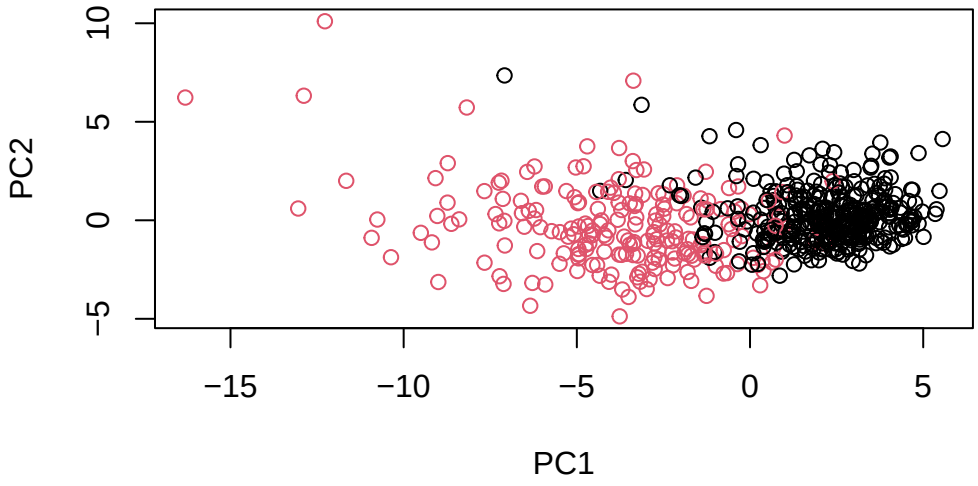
Yes, this is very difficult to read and understand because making an observation is really hard to see. All the components (the black and red data) are very close and overlapped together which is difficult to look at trends, make conclusions, and form analysis. We would need to generate another plot to understand the PCA result.

```
# Scatter plot observations by components 1 and 2
wisc.pr <- prcomp(wisc.data, scale =T)
plot(wisc.pr$x[,1], wisc.pr$x[,2], col = as.factor(diagnosis), xlab = "PC1", ylab = "PC2")
```



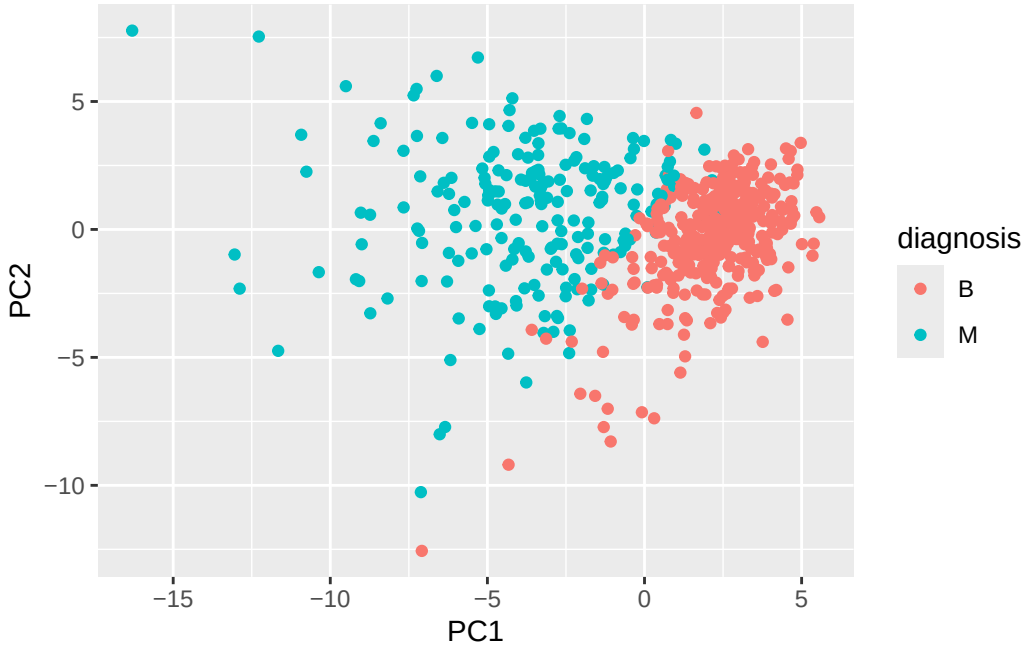
Q8. Generate a similar plot for principal components 1 and 3. What do you notice about these plots?

```
wisc.pr <- prcomp(wisc.data, scale =T)
plot(wisc.pr$x[,1], wisc.pr$x[,3], col = as.factor(diagnosis), xlab = "PC1", ylab = "PC2")
```



By looking at these plots, it looks much cleaner and its easier to read because its not too messy. There is more of a seperation between the variances. You can make observations and analysis of this data.

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

library(ggplot2)

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

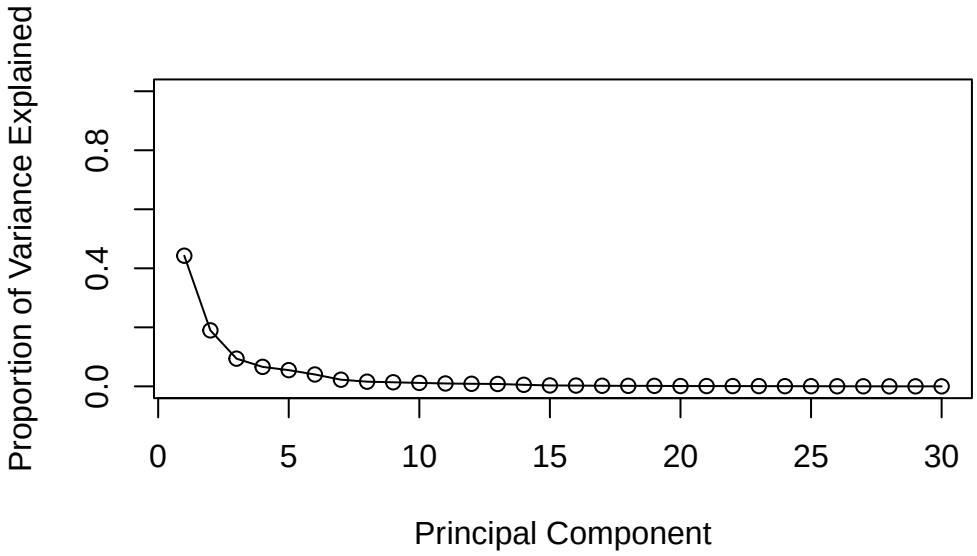


Variance Explained

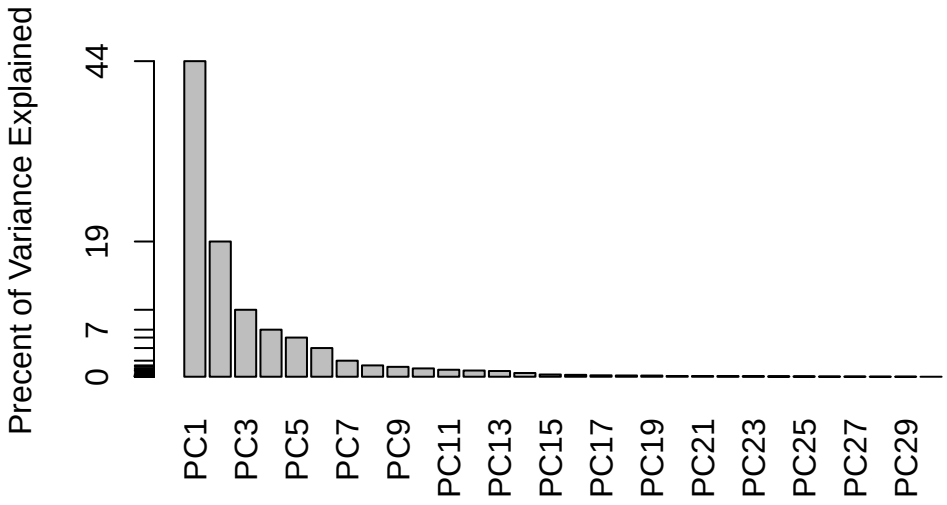
```
pr.var <- wisc.pr$sdev^2
head(pr.var)
```

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



Communicating PCA results

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?

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

cat("Component of the loading vector for concave.points_mean in the first PC:", loading_conca
```

```
Component of the loading vector for concave.points_mean in the first PC: -0.2608538
```

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

```
pca_result <- summary(wisc.pr)
cum_var_explained <- cumsum(pca_result$sdev^2) / sum(pca_result$sdev^2)

num_pcs_80 <- which(cum_var_explained >= 0.80)[1]

cat("Number of PCs required to explain at least 90% of the variance:", num_pcs_80, "\n")
```

```
Number of PCs required to explain at least 90% of the variance: 5
```

Hierarchical Clustering

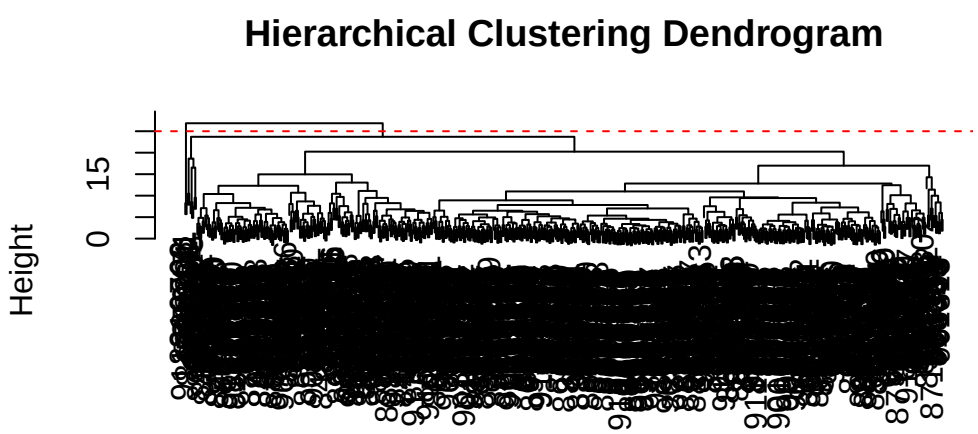
```
data.scaled <- scale(wisc.data)
```

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

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

Q11. Using the plot() and abline() functions, what is the height at which the clustering model has 4 clusters?

```
plot(wisc.hclust, main = "Hierarchical Clustering Dendrogram", xlab = "data.dist", sub = "Hei
abline(h=25, col="red", lty=2)
```



The height is 25.

Selecting number of clusters

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

```
wisc.hclust.clusters <- cutree(wisc.hclust, k=2:10)
print(wisc.hclust.clusters)
```

	2	3	4	5	6	7	8	9	10
842302	1	1	1	1	1	1	1	1	1
842517	1	1	1	1	1	1	2	2	2
84300903	1	1	1	1	1	1	2	2	2
84348301	1	2	2	2	2	2	3	3	3
84358402	1	1	1	1	1	1	2	2	2
843786	1	1	1	1	1	1	1	1	1
844359	1	1	1	1	1	1	1	1	1
84458202	1	1	1	1	1	1	1	1	1
844981	1	1	1	1	1	1	1	1	1
84501001	1	2	2	2	2	2	3	3	3
845636	1	1	3	3	3	3	4	4	4
84610002	1	1	1	1	1	1	1	1	1
846226	1	1	1	1	1	1	2	2	5
846381	1	1	3	3	3	3	4	4	4
84667401	1	1	1	1	1	1	1	1	1
84799002	1	1	1	1	1	1	1	1	1
848406	1	1	3	3	3	3	4	4	4
84862001	1	1	1	1	1	1	1	1	1
849014	1	1	1	1	1	1	2	2	2
8510426	1	1	3	3	3	3	4	4	4
8510653	1	1	3	3	3	3	4	4	4
8510824	1	1	3	3	3	3	4	4	4
8511133	1	1	1	1	1	1	1	1	1
851509	1	1	1	1	1	1	2	2	2
852552	1	1	1	1	1	1	2	2	2
852631	1	1	1	1	1	1	1	1	1
852763	1	1	1	1	1	1	1	1	1
852781	1	1	1	1	1	1	2	2	2
852973	1	1	1	1	1	1	1	1	1
853201	1	1	3	3	3	3	4	4	4
853401	1	1	1	1	1	1	1	1	1
853612	1	1	1	1	1	1	1	1	1
85382601	1	1	1	1	1	1	1	1	1
854002	1	1	1	1	1	1	1	1	1
854039	1	1	1	1	1	1	1	1	1
854253	1	1	1	1	1	1	1	1	1
854268	1	1	1	1	1	1	1	1	1
854941	1	1	3	3	3	3	4	4	4
855133	1	1	3	3	3	3	4	4	4
855138	1	1	1	1	1	1	1	1	1
855167	1	1	3	3	3	3	4	4	4
855563	1	1	1	1	1	1	1	1	1

855625	1	1	1	1	1	1	2	2	5
856106	1	1	1	1	1	1	1	1	1
85638502	1	1	1	1	1	1	1	1	1
857010	1	1	1	1	1	1	2	2	2
85713702	1	1	3	3	3	3	4	4	4
85715	1	1	1	1	1	1	1	1	1
857155	1	1	3	3	3	3	4	4	4
857156	1	1	3	3	3	3	4	4	4
857343	1	1	3	3	3	3	4	4	4
857373	1	1	3	3	3	3	4	4	4
857374	1	1	3	3	3	3	4	4	4
857392	1	1	1	1	1	1	2	2	2
857438	1	1	3	3	3	3	4	4	4
85759902	1	1	3	3	3	3	4	4	4
857637	1	1	1	1	1	1	2	2	2
857793	1	1	1	1	1	1	1	1	1
857810	1	1	3	3	3	3	4	4	4
858477	1	1	3	3	3	3	4	4	4
858970	1	1	3	3	3	3	4	4	4
858981	1	1	3	3	3	3	4	4	4
858986	1	1	1	1	1	1	1	1	1
859196	1	1	3	3	3	3	4	4	4
85922302	1	1	1	1	1	1	1	1	1
859283	1	1	1	1	1	1	1	1	1
859464	1	1	3	3	3	3	4	4	4
859465	1	1	3	3	3	3	4	4	4
859471	1	2	2	4	4	4	5	5	6
859487	1	1	3	3	3	3	4	4	4
859575	1	1	1	1	1	1	2	2	2
859711	1	1	3	3	5	5	6	6	7
859717	1	1	1	1	1	1	1	1	1
859983	1	1	1	1	1	1	1	1	1
8610175	1	1	3	3	3	3	4	4	4
8610404	1	1	3	3	3	3	4	4	4
8610629	1	1	3	3	3	3	4	4	4
8610637	1	1	1	1	1	1	2	2	5
8610862	1	2	2	2	2	6	7	7	8
8610908	1	1	3	3	3	3	4	4	4
861103	1	1	3	3	3	3	4	4	4
8611161	1	1	1	1	1	1	1	1	1
8611555	1	1	1	1	1	1	2	2	5
8611792	1	1	1	1	1	1	2	2	5
8612080	1	1	3	3	3	3	4	4	4

8612399	1	1	1	1	1	1	2	2	2
86135501	1	1	3	3	3	3	4	4	4
86135502	1	1	1	1	1	1	2	2	2
861597	1	1	3	3	3	3	4	4	4
861598	1	1	1	1	1	1	1	1	1
861648	1	1	3	3	3	3	4	4	4
861799	1	1	3	3	3	3	4	4	4
861853	1	1	3	3	3	3	4	4	4
862009	1	1	3	3	3	3	4	4	4
862028	1	1	1	1	1	1	1	1	1
86208	1	1	1	1	1	1	2	2	2
86211	1	1	3	3	3	3	4	4	4
862261	1	1	3	3	3	3	4	4	4
862485	1	1	3	3	3	3	4	4	4
862548	1	1	3	3	3	3	4	4	4
862717	1	1	3	3	3	3	4	4	4
862722	1	1	3	3	3	3	4	4	4
862965	1	1	3	3	3	3	4	4	4
862980	1	1	3	3	3	3	4	4	4
862989	1	1	3	3	3	3	4	4	4
863030	1	1	1	1	1	1	1	1	1
863031	1	1	1	1	1	1	1	1	1
863270	1	1	3	3	3	3	4	4	4
86355	1	1	1	1	1	1	2	2	5
864018	1	1	3	3	3	3	4	4	4
864033	1	1	3	3	3	3	4	4	4
86408	1	1	3	3	3	3	4	4	4
86409	1	1	3	3	5	5	6	6	7
864292	1	1	3	3	3	3	4	4	4
864496	1	1	3	3	3	3	4	4	4
864685	1	1	3	3	3	3	4	4	4
864726	1	1	3	3	3	3	4	4	4
864729	1	1	1	1	1	1	1	1	1
864877	1	1	1	1	1	1	1	1	1
865128	1	1	3	3	3	3	4	4	4
865137	1	1	3	3	3	3	4	4	4
86517	1	1	1	1	1	1	2	2	2
865423	1	2	2	2	2	6	7	7	8
865432	1	1	3	3	3	3	4	4	4
865468	1	1	3	3	3	3	4	4	4
86561	1	1	3	3	3	3	4	4	4
866083	1	1	1	1	1	1	1	1	1
866203	1	1	3	3	3	3	4	4	4

866458	1	1	1	1	1	1	1	1	1
866674	1	1	1	1	1	1	1	1	1
866714	1	1	3	3	3	3	4	4	4
8670	1	1	1	1	1	1	1	1	1
86730502	1	1	1	1	1	1	1	1	1
867387	1	1	3	3	3	3	4	4	4
867739	1	1	1	1	1	1	1	1	1
868202	1	1	3	3	3	3	4	4	4
868223	1	1	3	3	3	3	4	4	4
868682	1	1	3	3	3	3	4	4	4
868826	1	1	1	1	1	1	1	1	1
868871	1	1	3	3	3	3	4	4	4
868999	1	1	3	3	3	3	4	4	4
869104	1	1	3	3	3	3	4	4	4
869218	1	1	3	3	3	3	4	4	4
869224	1	1	3	3	3	3	4	4	4
869254	1	1	3	3	3	3	4	4	4
869476	1	1	3	3	3	3	4	4	4
869691	1	1	1	1	1	1	1	1	1
86973701	1	1	3	3	3	3	4	4	4
86973702	1	1	3	3	3	3	4	4	4
869931	1	1	3	3	3	3	4	4	4
871001501	1	1	3	3	3	3	4	4	4
871001502	1	1	3	3	5	5	6	6	7
8710441	1	2	2	4	4	4	5	5	6
87106	1	1	3	3	3	3	4	4	4
8711002	1	1	3	3	3	3	4	4	4
8711003	1	1	3	3	3	3	4	4	4
8711202	1	1	1	1	1	1	2	2	2
8711216	1	1	3	3	3	3	4	4	4
871122	1	1	3	3	3	3	4	4	4
871149	1	1	3	3	3	3	4	4	4
8711561	1	1	3	3	3	3	4	4	4
8711803	1	1	1	1	1	1	2	2	2
871201	1	1	1	1	1	1	1	1	1
8712064	1	1	3	3	3	3	4	4	4
8712289	1	1	1	1	1	1	2	2	2
8712291	1	1	3	3	3	3	4	4	4
87127	1	1	3	3	3	3	4	4	4
8712729	1	1	3	3	3	3	4	4	4
8712766	1	1	1	1	1	1	2	2	5
8712853	1	1	3	3	3	3	4	4	4
87139402	1	1	3	3	3	3	4	4	4

87163	1	1	3	3	3	3	4	4	4
87164	1	1	1	1	1	1	1	1	1
871641	1	1	3	3	3	3	4	4	4
871642	1	1	3	3	3	3	4	4	4
872113	1	1	3	3	3	3	4	4	4
872608	1	1	3	3	5	5	6	6	7
87281702	1	1	1	1	1	1	1	1	1
873357	1	1	3	3	3	3	4	4	4
873586	1	1	3	3	3	3	4	4	4
873592	1	1	1	1	1	1	2	2	2
873593	1	1	1	1	1	1	2	2	5
873701	1	1	1	1	1	1	1	1	1
873843	1	1	3	3	3	3	4	4	4
873885	1	1	1	1	1	1	1	1	1
874158	1	1	3	3	3	3	4	4	4
874217	1	1	3	3	3	3	4	4	4
874373	1	1	3	3	3	3	4	4	4
874662	1	1	3	3	3	3	4	4	4
874839	1	1	3	3	3	3	4	4	4
874858	1	2	2	2	2	2	3	3	3
875093	1	1	3	3	3	3	4	4	4
875099	1	1	3	3	3	3	4	4	4
875263	1	1	1	1	1	1	1	1	1
87556202	1	1	1	1	1	1	1	1	1
875878	1	1	3	3	3	3	4	4	4
875938	1	1	1	1	1	1	1	1	1
877159	1	1	3	3	3	3	4	4	4
877486	1	1	1	1	1	1	2	2	2
877500	1	1	1	1	1	1	1	1	1
877501	1	1	3	3	3	3	4	4	4
877989	1	1	3	3	3	3	4	4	4
878796	1	1	1	1	1	1	2	2	5
87880	1	1	1	1	1	1	1	1	1
87930	1	1	3	3	3	3	4	4	4
879523	1	1	3	3	3	3	4	4	4
879804	1	1	3	3	3	3	4	4	4
879830	1	1	3	3	3	3	4	4	4
8810158	1	1	1	1	1	1	1	1	1
8810436	1	1	3	3	3	3	4	4	4
881046502	1	1	1	1	1	1	2	2	2
8810528	1	1	3	3	3	3	4	4	4
8810703	2	3	4	5	6	7	8	8	9
881094802	1	1	3	3	5	5	6	9	10

8810955	1	1	1	1	1	1	1	1	1
8810987	1	1	1	1	1	1	1	1	1
8811523	1	1	3	3	3	3	4	4	4
8811779	1	1	3	3	3	3	4	4	4
8811842	1	1	1	1	1	1	2	2	2
88119002	1	1	1	1	1	1	2	2	2
8812816	1	1	3	3	3	3	4	4	4
8812818	1	1	3	3	3	3	4	4	4
8812844	1	1	3	3	3	3	4	4	4
8812877	1	1	1	1	1	1	1	1	1
8813129	1	1	3	3	3	3	4	4	4
88143502	1	1	3	3	3	3	4	4	4
88147101	1	1	3	3	3	3	4	4	4
88147102	1	1	3	3	3	3	4	4	4
88147202	1	1	3	3	3	3	4	4	4
881861	1	1	1	1	1	1	1	1	1
881972	1	1	1	1	1	1	1	1	1
88199202	1	1	3	3	3	3	4	4	4
88203002	1	1	3	3	3	3	4	4	4
88206102	1	1	1	1	1	1	2	2	2
882488	1	1	3	3	3	3	4	4	4
88249602	1	1	3	3	3	3	4	4	4
88299702	1	1	1	1	1	1	2	2	2
883263	1	1	1	1	1	1	2	2	2
883270	1	1	3	3	3	3	4	4	4
88330202	1	1	1	1	1	1	2	2	2
88350402	1	1	3	3	3	3	4	4	4
883539	1	1	3	3	3	3	4	4	4
883852	1	1	3	3	5	5	6	6	7
88411702	1	1	3	3	3	3	4	4	4
884180	1	1	1	1	1	1	2	2	2
884437	1	1	3	3	3	3	4	4	4
884448	1	1	3	3	3	3	4	4	4
884626	1	1	3	3	3	3	4	4	4
88466802	1	1	3	3	3	3	4	4	4
884689	1	1	3	3	3	3	4	4	4
884948	1	1	1	1	1	1	2	2	5
88518501	1	1	3	3	3	3	4	4	4
885429	1	1	1	1	1	1	1	1	1
8860702	1	1	3	3	3	3	4	4	4
886226	1	1	1	1	1	1	2	2	2
886452	1	1	3	3	3	3	4	4	4
88649001	1	1	1	1	1	1	2	2	5

886776	1	1	1	1	1	1	1	1	1
887181	1	1	1	1	1	1	2	2	5
88725602	1	1	1	1	1	1	1	1	1
887549	1	1	1	1	1	1	2	2	2
888264	1	1	3	3	3	3	4	4	4
888570	1	1	1	1	1	1	2	2	2
889403	1	1	3	3	3	3	4	4	4
889719	1	1	1	1	1	1	1	1	1
88995002	1	1	1	1	1	1	2	2	2
8910251	1	1	3	3	3	3	4	4	4
8910499	1	1	3	3	3	3	4	4	4
8910506	1	1	3	3	3	3	4	4	4
8910720	1	1	3	3	3	3	4	4	4
8910721	1	1	3	3	3	3	4	4	4
8910748	1	1	3	3	3	3	4	4	4
8910988	1	1	1	1	1	1	2	2	5
8910996	1	1	3	3	3	3	4	4	4
8911163	1	1	3	3	3	3	4	4	4
8911164	1	1	3	3	3	3	4	4	4
8911230	1	1	3	3	3	3	4	4	4
8911670	1	1	3	3	3	3	4	4	4
8911800	1	1	3	3	3	3	4	4	4
8911834	1	1	3	3	3	3	4	4	4
8912049	1	1	1	1	1	1	1	1	1
8912055	1	1	3	3	3	3	4	4	4
89122	1	1	1	1	1	1	2	2	2
8912280	1	1	1	1	1	1	1	1	1
8912284	1	1	3	3	3	3	4	4	4
8912521	1	1	3	3	3	3	4	4	4
8912909	1	1	3	3	3	3	4	4	4
8913	1	1	3	3	3	3	4	4	4
8913049	1	1	3	3	3	3	4	4	4
89143601	1	1	3	3	3	3	4	4	4
89143602	1	1	3	3	5	5	6	6	7
8915	1	1	3	3	3	3	4	4	4
891670	1	1	3	3	3	3	4	4	4
891703	1	1	3	3	3	3	4	4	4
891716	1	1	3	3	3	3	4	4	4
891923	1	1	3	3	3	3	4	4	4
891936	1	1	3	3	3	3	4	4	4
892189	1	1	3	3	3	3	4	4	4
892214	1	1	3	3	3	3	4	4	4
892399	1	1	3	3	3	3	4	4	4

892438	1	1	1	1	1	1	2	2	5
892604	1	1	3	3	3	3	4	4	4
89263202	1	1	1	1	1	1	2	2	5
892657	1	1	3	3	3	3	4	4	4
89296	1	1	3	3	3	3	4	4	4
893061	1	1	3	3	3	3	4	4	4
89344	1	1	3	3	3	3	4	4	4
89346	1	1	3	3	3	3	4	4	4
893526	1	1	3	3	3	3	4	4	4
893548	1	1	3	3	3	3	4	4	4
893783	1	1	3	3	3	3	4	4	4
89382601	1	1	3	3	3	3	4	4	4
89382602	1	1	3	3	3	3	4	4	4
893988	1	1	3	3	3	3	4	4	4
894047	1	1	3	3	3	3	4	4	4
894089	1	1	3	3	3	3	4	4	4
894090	1	1	3	3	3	3	4	4	4
894326	1	1	1	1	1	1	1	1	1
894329	1	1	3	3	5	5	6	6	7
894335	1	1	3	3	3	3	4	4	4
894604	1	1	3	3	3	3	4	4	4
894618	1	1	3	3	3	3	4	4	4
894855	1	1	3	3	3	3	4	4	4
895100	1	1	1	1	1	1	1	1	1
89511501	1	1	3	3	3	3	4	4	4
89511502	1	1	3	3	3	3	4	4	4
89524	1	1	3	3	3	3	4	4	4
895299	1	1	3	3	3	3	4	4	4
8953902	1	1	1	1	1	1	1	1	1
895633	1	1	1	1	1	1	1	1	1
896839	1	1	1	1	1	1	1	1	1
896864	1	1	1	1	1	1	1	1	1
897132	1	1	3	3	3	3	4	4	4
897137	1	1	3	3	3	3	4	4	4
897374	1	1	3	3	3	3	4	4	4
89742801	1	1	1	1	1	1	2	2	2
897604	1	1	3	3	3	3	4	4	4
897630	1	1	1	1	1	1	2	2	2
897880	1	1	3	3	3	3	4	4	4
89812	1	1	1	1	1	1	2	2	2
89813	1	1	3	3	3	3	4	4	4
898143	1	1	3	3	3	3	4	4	4
89827	1	1	3	3	3	3	4	4	4

898431	1	1	1	1	1	1	2	2	2
89864002	1	1	3	3	3	3	4	4	4
898677	1	1	3	3	3	3	4	4	4
898678	1	1	3	3	3	3	4	4	4
89869	1	1	3	3	3	3	4	4	4
898690	1	1	3	3	3	3	4	4	4
899147	1	1	3	3	3	3	4	4	4
899187	1	1	3	3	3	3	4	4	4
899667	1	1	1	1	1	1	1	1	1
899987	1	1	1	1	1	1	2	2	2
9010018	1	1	1	1	1	1	1	1	1
901011	1	1	3	3	3	3	4	4	4
9010258	1	1	3	3	3	3	4	4	4
9010259	1	1	3	3	3	3	4	4	4
901028	1	1	3	3	3	3	4	4	4
9010333	1	1	3	3	3	3	4	4	4
901034301	1	1	3	3	3	3	4	4	4
901034302	1	1	3	3	3	3	4	4	4
901041	1	1	3	3	3	3	4	4	4
9010598	1	1	3	3	3	3	4	4	4
9010872	1	1	3	3	3	3	4	4	4
9010877	1	1	3	3	3	3	4	4	4
901088	1	1	1	1	1	1	2	2	2
9011494	1	1	1	1	1	1	2	2	2
9011495	1	1	3	3	3	3	4	4	4
9011971	1	1	1	1	1	1	2	2	2
9012000	1	1	1	1	1	1	2	2	2
9012315	1	1	1	1	1	1	1	1	1
9012568	1	1	3	3	3	3	4	4	4
9012795	1	1	1	1	1	1	2	2	2
901288	1	1	1	1	1	1	2	2	2
9013005	1	1	3	3	3	3	4	4	4
901303	1	1	3	3	3	3	4	4	4
901315	1	1	3	3	5	5	6	6	7
9013579	1	1	3	3	3	3	4	4	4
9013594	1	1	3	3	3	3	4	4	4
9013838	1	1	1	1	1	1	1	1	1
901549	1	1	3	3	3	3	4	4	4
901836	1	1	3	3	3	3	4	4	4
90250	1	1	3	3	3	3	4	4	4
90251	1	1	3	3	3	3	4	4	4
902727	1	1	3	3	3	3	4	4	4
90291	1	1	3	3	3	3	4	4	4

902975	1	1	3	3	3	3	4	4	4
902976	1	1	3	3	3	3	4	4	4
903011	1	1	3	3	3	3	4	4	4
90312	1	1	1	1	1	1	2	2	5
90317302	1	1	3	3	3	3	4	4	4
903483	1	1	3	3	3	3	4	4	4
903507	1	1	1	1	1	1	1	1	1
903516	1	1	1	1	1	1	1	1	1
903554	1	1	3	3	3	3	4	4	4
903811	1	1	3	3	3	3	4	4	4
90401601	1	1	3	3	3	3	4	4	4
90401602	1	1	3	3	3	3	4	4	4
904302	1	1	3	3	3	3	4	4	4
904357	1	1	3	3	3	3	4	4	4
90439701	1	1	1	1	1	1	1	1	1
904647	1	1	3	3	3	3	4	4	4
904689	1	1	3	3	3	3	4	4	4
9047	1	1	3	3	3	3	4	4	4
904969	1	1	3	3	3	3	4	4	4
904971	1	1	3	3	3	3	4	4	4
905189	1	1	3	3	3	3	4	4	4
905190	1	1	3	3	3	3	4	4	4
90524101	1	1	1	1	1	1	1	1	1
905501	1	1	3	3	3	3	4	4	4
905502	1	1	3	3	3	3	4	4	4
905520	1	1	3	3	3	3	4	4	4
905539	1	1	3	3	3	3	4	4	4
905557	1	1	3	3	3	3	4	4	4
905680	1	1	3	3	3	3	4	4	4
905686	1	1	3	3	3	3	4	4	4
905978	1	1	3	3	3	3	4	4	4
90602302	1	1	1	1	1	1	2	2	5
906024	1	1	3	3	3	3	4	4	4
906290	1	1	3	3	3	3	4	4	4
906539	1	1	3	3	3	3	4	4	4
906564	1	1	1	1	1	1	1	1	1
906616	1	1	3	3	3	3	4	4	4
906878	1	1	3	3	3	3	4	4	4
907145	1	1	3	3	3	3	4	4	4
907367	1	1	3	3	3	3	4	4	4
907409	1	1	3	3	3	3	4	4	4
90745	1	1	3	3	3	3	4	4	4
90769601	1	1	3	3	3	3	4	4	4

90769602	1	1	3	3	3	3	4	4	4
907914	1	1	1	1	1	1	1	1	1
907915	1	1	3	3	3	3	4	4	4
908194	1	1	1	1	1	1	2	2	2
908445	1	1	1	1	1	1	2	2	2
908469	1	1	3	3	3	3	4	4	4
908489	1	1	1	1	1	1	1	1	1
908916	1	1	3	3	3	3	4	4	4
909220	1	1	3	3	3	3	4	4	4
909231	1	1	3	3	3	3	4	4	4
909410	1	1	3	3	3	3	4	4	4
909411	1	1	3	3	3	3	4	4	4
909445	1	1	3	3	3	3	4	4	4
90944601	1	1	3	3	3	3	4	4	4
909777	1	1	3	3	3	3	4	4	4
9110127	1	1	3	3	3	3	4	4	4
9110720	1	1	3	3	3	3	4	4	4
9110732	1	1	1	1	1	1	2	2	2
9110944	1	1	3	3	3	3	4	4	4
911150	1	1	3	3	3	3	4	4	4
911157302	1	1	1	1	1	1	2	2	2
9111596	1	1	3	3	3	3	4	4	4
9111805	1	1	1	1	1	1	2	2	2
9111843	1	1	3	3	3	3	4	4	4
911201	1	1	3	3	3	3	4	4	4
911202	1	1	3	3	3	3	4	4	4
9112085	1	1	3	3	3	3	4	4	4
9112366	1	1	3	3	3	3	4	4	4
9112367	1	1	3	3	3	3	4	4	4
9112594	1	1	3	3	3	3	4	4	4
9112712	1	1	3	3	3	3	4	4	4
911296201	1	1	1	1	1	1	2	2	2
911296202	2	3	4	5	6	7	8	8	9
9113156	1	1	3	3	3	3	4	4	4
911320501	1	1	3	3	3	3	4	4	4
911320502	1	1	3	3	3	3	4	4	4
9113239	1	1	3	3	3	3	4	4	4
9113455	1	1	3	3	3	3	4	4	4
9113514	1	1	3	3	3	3	4	4	4
9113538	1	1	1	1	1	1	2	2	5
911366	1	1	1	1	1	1	1	1	1
9113778	1	1	3	3	3	3	4	4	4
9113816	1	1	3	3	3	3	4	4	4

911384	1	1	3	3	3	3	4	4	4
9113846	1	1	3	3	3	3	4	4	4
911391	1	1	3	3	3	3	4	4	4
911408	1	1	3	3	3	3	4	4	4
911654	1	1	3	3	3	3	4	4	4
911673	1	1	3	3	3	3	4	4	4
911685	1	1	3	3	3	3	4	4	4
911916	1	1	1	1	1	1	1	1	1
912193	1	1	3	3	3	3	4	4	4
91227	1	1	3	3	3	3	4	4	4
912519	1	1	3	3	3	3	4	4	4
912558	1	1	3	3	3	3	4	4	4
912600	1	1	3	3	3	3	4	4	4
913063	1	1	3	3	5	5	6	6	7
913102	1	1	3	3	3	3	4	4	4
913505	1	1	1	1	1	1	1	1	1
913512	1	1	3	3	3	3	4	4	4
913535	1	1	3	3	3	3	4	4	4
91376701	1	1	3	3	3	3	4	4	4
91376702	1	1	3	3	3	3	4	4	4
914062	1	1	1	1	1	1	2	2	2
914101	1	1	3	3	3	3	4	4	4
914102	1	1	3	3	3	3	4	4	4
914333	1	1	3	3	3	3	4	4	4
914366	1	1	1	1	1	1	1	1	1
914580	1	1	3	3	3	3	4	4	4
914769	1	1	1	1	1	1	2	2	2
91485	1	1	1	1	1	1	1	1	1
914862	1	1	3	3	3	3	4	4	4
91504	1	1	1	1	1	1	1	1	1
91505	1	1	3	3	3	3	4	4	4
915143	1	1	1	1	1	1	2	2	2
915186	1	1	3	3	5	5	6	6	7
915276	1	1	3	3	5	5	6	6	7
91544001	1	1	3	3	3	3	4	4	4
91544002	1	1	3	3	3	3	4	4	4
915452	1	1	3	3	3	3	4	4	4
915460	1	1	1	1	1	1	1	1	1
91550	1	1	3	3	3	3	4	4	4
915664	1	1	3	3	3	3	4	4	4
915691	1	1	1	1	1	1	1	1	1
915940	1	1	3	3	3	3	4	4	4
91594602	1	1	3	3	3	3	4	4	4

916221	1	1	3	3	3	3	4	4	4
916799	1	1	1	1	1	1	1	1	1
916838	1	1	1	1	1	1	2	2	2
917062	1	1	3	3	3	3	4	4	4
917080	1	1	3	3	3	3	4	4	4
917092	1	1	3	3	3	3	4	4	4
91762702	1	1	1	1	1	1	2	2	2
91789	1	1	3	3	3	3	4	4	4
917896	1	1	3	3	3	3	4	4	4
917897	1	1	3	3	3	3	4	4	4
91805	1	1	3	3	3	3	4	4	4
91813701	1	1	1	1	1	1	1	1	1
91813702	1	1	3	3	3	3	4	4	4
918192	1	1	3	3	3	3	4	4	4
918465	1	1	3	3	3	3	4	4	4
91858	1	1	3	3	3	3	4	4	4
91903901	1	1	3	3	3	3	4	4	4
91903902	1	1	3	3	3	3	4	4	4
91930402	1	1	1	1	1	1	2	2	2
919537	1	1	3	3	3	3	4	4	4
919555	1	1	1	1	1	1	2	2	2
91979701	1	1	3	3	3	3	4	4	4
919812	1	1	1	1	1	1	1	1	1
921092	1	1	3	3	3	3	4	4	4
921362	1	1	3	3	5	5	6	6	7
921385	1	1	3	3	3	3	4	4	4
921386	1	1	1	1	1	1	1	1	1
921644	1	1	3	3	3	3	4	4	4
922296	1	1	3	3	3	3	4	4	4
922297	1	1	3	3	3	3	4	4	4
922576	1	1	3	3	3	3	4	4	4
922577	1	1	3	3	3	3	4	4	4
922840	1	1	3	3	3	3	4	4	4
923169	1	1	3	3	3	3	4	4	4
923465	1	1	3	3	3	3	4	4	4
923748	1	1	3	3	3	3	4	4	4
923780	1	1	3	3	3	3	4	4	4
924084	1	1	3	3	3	3	4	4	4
924342	1	1	3	3	3	3	4	4	4
924632	1	1	3	3	3	3	4	4	4
924934	1	1	3	3	3	3	4	4	4
924964	1	1	3	3	3	3	4	4	4
925236	1	1	3	3	3	3	4	4	4

```
925277    1 1 3 3 3 3 4 4  4
925291    1 1 3 3 3 3 4 4  4
925292    1 1 3 3 3 3 4 4  4
925311    1 1 3 3 3 3 4 4  4
925622    1 1 1 1 1 1 1 1  1
926125    1 1 1 1 1 1 2 2  5
926424    1 1 1 1 1 1 2 2  2
926682    1 1 1 1 1 1 2 2  2
926954    1 1 3 3 3 3 4 4  4
927241    1 1 1 1 1 1 2 2  5
92751     1 1 3 3 3 3 4 4  4
```

No you can not find a better cluster vs diagnoses match by cutting into a different number of clusters between 2 and 10 because you will have a large data set which can be difficult for matching. It will need to be condensed through a different method.

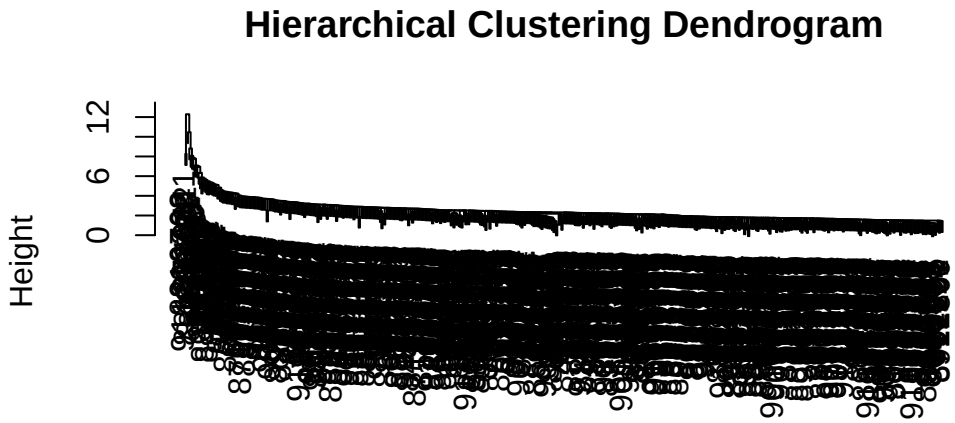
Using Different Methods

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

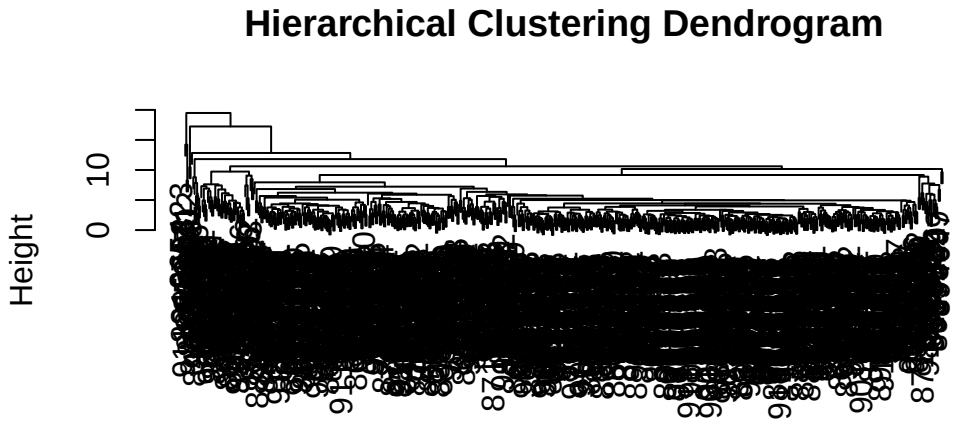
```
wisc.hclust <- hclust(data.dist, method = "single")
wisc.hclust
```

```
Call:
hclust(d = data.dist, method = "single")

Cluster method   : single
Distance         : euclidean
Number of objects: 569
```

```
plot(wisc.hclust, main = "Hierarchical Clustering Dendrogram", xlab = "data.dist", sub = "Hei
abline(h=25, col="red", lty=2)
```



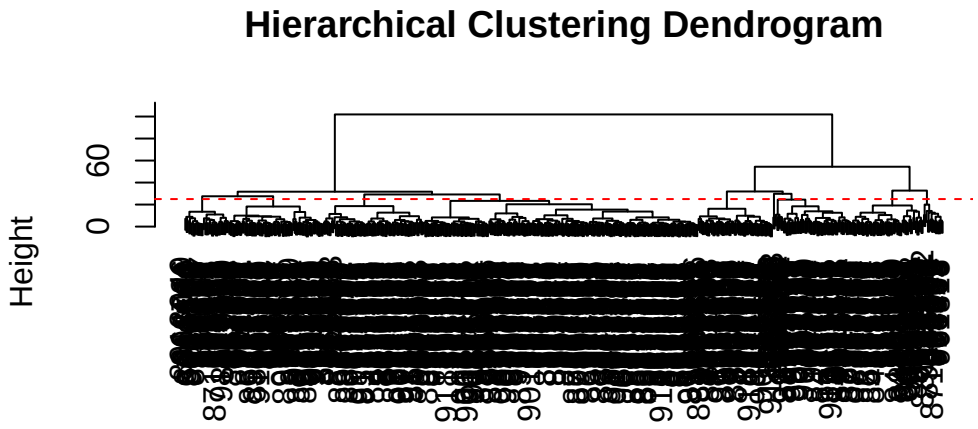

```
wisc.hclust <- hclust(data.dist, method = "average")
plot(wisc.hclust, main = "Hierarchical Clustering Dendrogram", xlab = "data.dist", sub = "Hei
abline(h=25, col="red", lty=2)
```



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

```
Call:
hclust(d = data.dist, method = "ward.D2")

Cluster method   : ward.D2
Distance         : euclidean
Number of objects: 569
```

```
plot(wisc.hclust, main = "Hierarchical Clustering Dendrogram", xlab = "data.dist", sub = "Hei
abline(h=25, col="red", lty=2)
```



The best method for the data set would be using "ward.D2" because this method creates groups to have their variance to be smaller in their clusters. This makes this easier to make observations.

K-means

```
scaled_data <- scale(wisc.data)
wisc.km <- kmeans(scaled_data, centers= 2, nstart= 20)
```

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

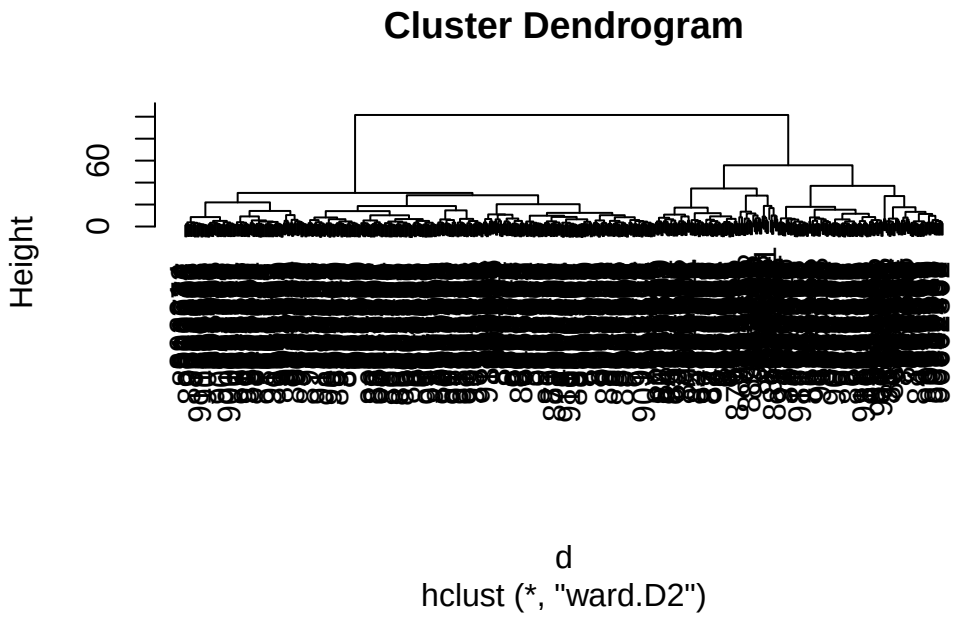
```
   diagnosis
      B   M
  1 343  37
  2  14 175
```

Q14. How well does k-means separate the two diagnoses? How does it compare to your hclust results?

The k-means does not separate the two diagnoses well. The output was too messy, long, and difficult to read. Compared to the the hclust results, the table is much shorter, easier to read, and better for making analysis since the table has seperated the results into two clusters.

##Combine PCS and clustering

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



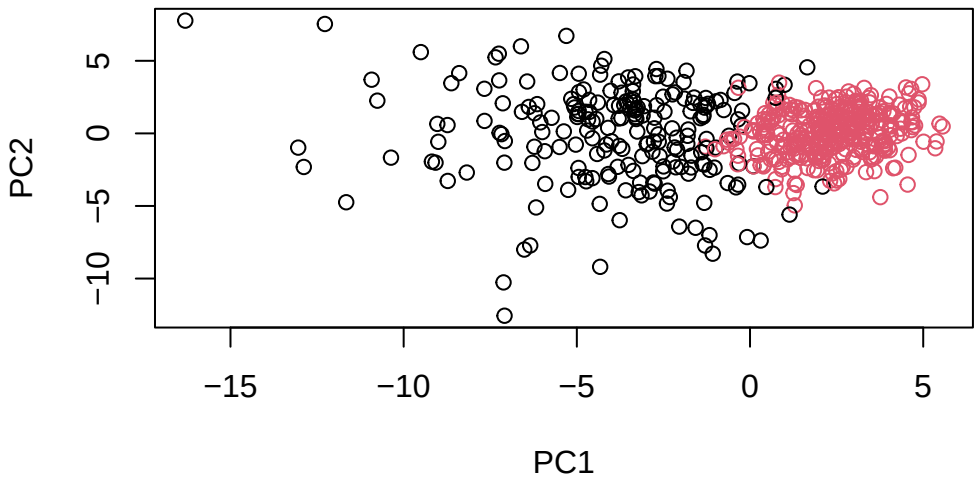
```
grps <- cutree(wisc.pr.hclust, k=2)
table(grps)
```

```
grps
  1   2
216 353
```

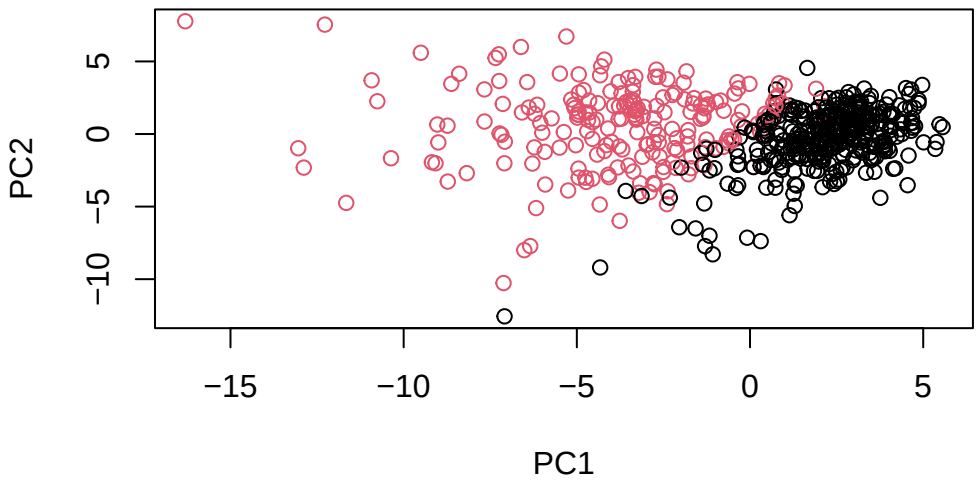
```
table(grps, diagnosis)
```

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

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




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



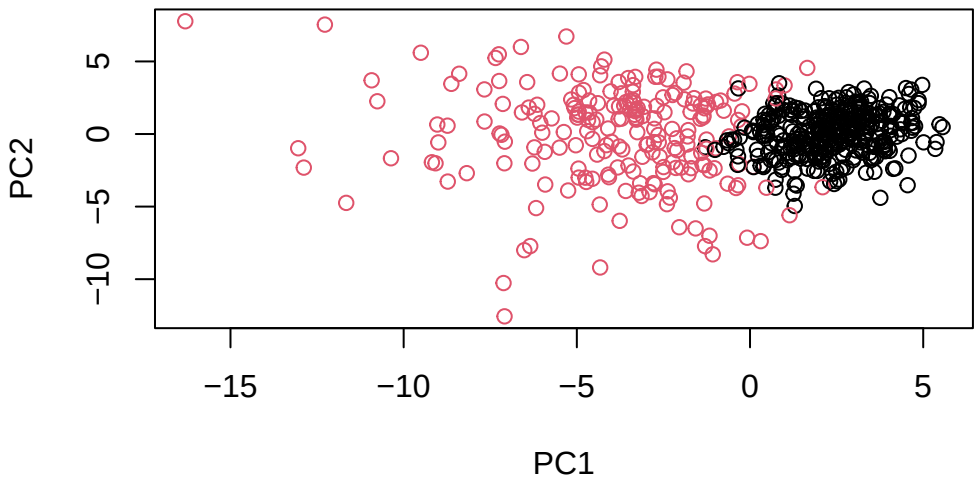

```
g <- as.factor(grps)
levels(g)
```

```
[1] "1" "2"
```

```
g <- relevel(g,2)
levels(g)
```

```
[1] "2" "1"
```

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




```
wisc.pr.hclust <- hclust(d, method="ward.D2")
wisc.pr.hclust.clusters <- cutree(wisc.pr.hclust, k=2)
```

Q15. How well does the newly created model with four clusters separate out the two diagnoses?

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

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

The four clusters separates the two diagnoses well because the outputs are different, suggesting a good separation of the two diagnoses, and there isn't any significant overlaps in the clustering.

Q16. How well do the k-means and 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.

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

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

```
scaled_data <- scale(wisc.data)
data.dist <- dist(scaled_data)
hclust_model <- hclust(data.dist, method = "complete")
wisc.hclust.clusters <- cutree(hclust_model, k = 4)
comparison_table <- table(wisc.hclust.clusters, diagnosis)
print(comparison_table)
```

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

The kmeans and hierarchical clustering models separated the diagnoses well because in k-means cluster 1 contains malignant cases and cluster 2 contains benign cases which are in good separation. Additionally, in hierarchical clustering, all of the clusters contain mixed diagnoses, indicating a good separated diagnoses.

Sensitivity/Specificity

> **Q17. Which of your analysis procedures resulted in a clustering model with the best specificity? How about sensitivity?**

k-means: Sensitivity= TP/(TP + FN) 175/(175 + 37) = 175/212 0.825

Specificity: TN/(TN+FP) 343/(343+14) = 343/357 0.961

hierarchical clustering: Sensitivity= TP/(TP + FN) 165/(165 + 47) = 165/212 0.778

Specificity: TN/(TN+FP) 343/(343+14) = 343/357 0.961

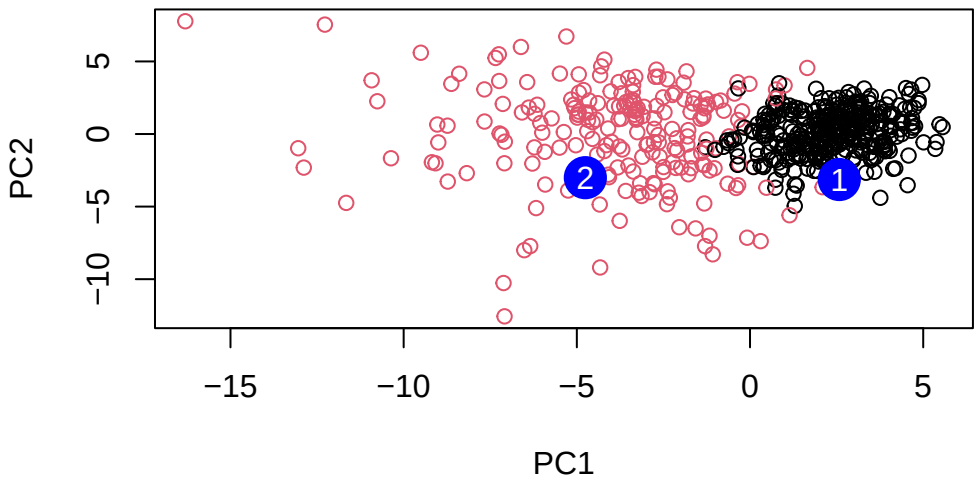
Both k-means and hierarchical clustering models have the best specificity; however, k-means model has the best sensitivity.

##Prediction >**Q18. Which of these new patients should we prioritize for follow up based on your results?**

```
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
```

```
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")
```



Patient 2 should be prioritized for a follow up result.