November 1, 2023: Class 09 Breast Cancer

Savannah Bogus A69027475

Exploration of Data

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

We're going to delete the diagnosis since that's the answer we're actually looking for.

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

Now, we're going to make a diagnosis vector for later.

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

Q1

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

```
[1] 569  31
```

There are 569 patient samples/observations in the dataset and 31 variables in the dataset (excluding diagnosis).

Q2

```
table(diagnosis)
```

```
diagnosis
  B   M
357 212
```

There are 357 benign and 212 malignant diagnoses.

Q3

```
  colnames(wisc.data)
```

```
 [1] "radius_mean"             "texture_mean"
 [3] "perimeter_mean"          "area_mean"
 [5] "smoothness_mean"         "compactness_mean"
 [7] "concavity_mean"          "concave.points_mean"
 [9] "symmetry_mean"           "fractal_dimension_mean"
[11] "radius_se"               "texture_se"
[13] "perimeter_se"            "area_se"
[15] "smoothness_se"           "compactness_se"
[17] "concavity_se"            "concave.points_se"
[19] "symmetry_se"             "fractal_dimension_se"
[21] "radius_worst"            "texture_worst"
[23] "perimeter_worst"         "area_worst"
[25] "smoothness_worst"        "compactness_worst"
[27] "concavity_worst"         "concave.points_worst"
[29] "symmetry_worst"          "fractal_dimension_worst"
[31] "X"
```

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

```
[1] 10
```

10 variables are suffixed with "_mean".

```
  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
                    X
                   NA
```

```
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
                    X
                   NA
```

Principle Component Analysis

```
wisc.pr<-prcomp(wisc.data[,colnames(wisc.data)!="X"],scale=TRUE)
summary<-summary(wisc.pr)
summary
```

```
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

~44% of the variance is captured by the first PC.

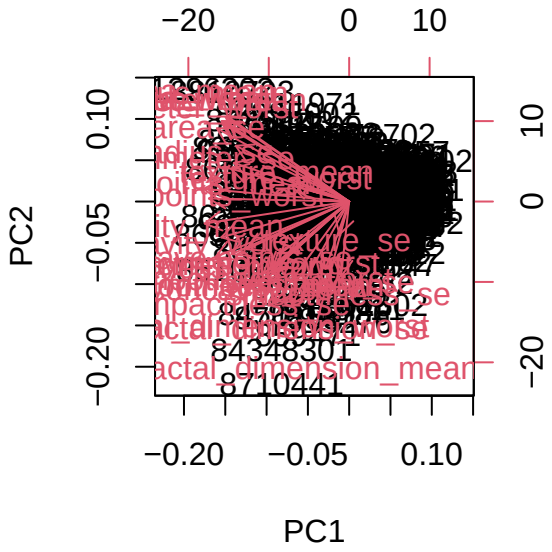
Q5

You need PC1-3 to describe at least 70% of the variance in the original data.

Q6

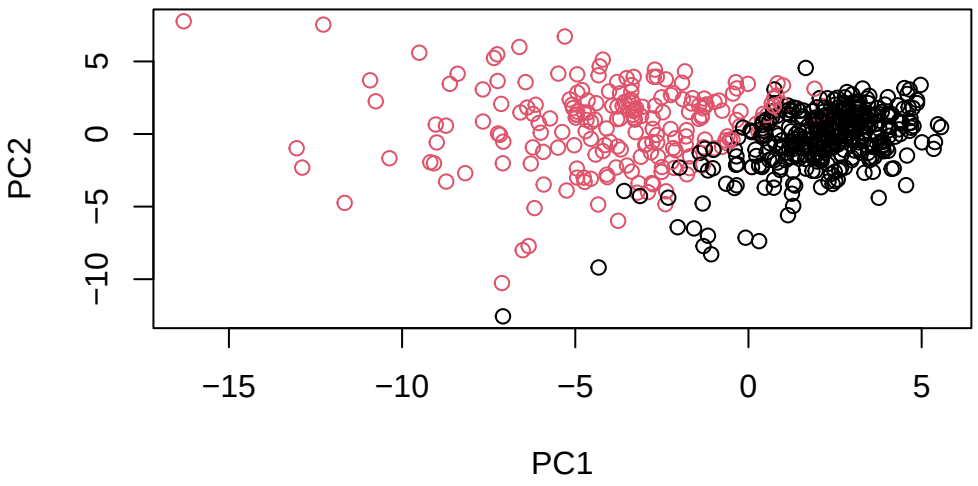
You need PC1-PC7 to describe at least 90% of the variance in the original data.

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



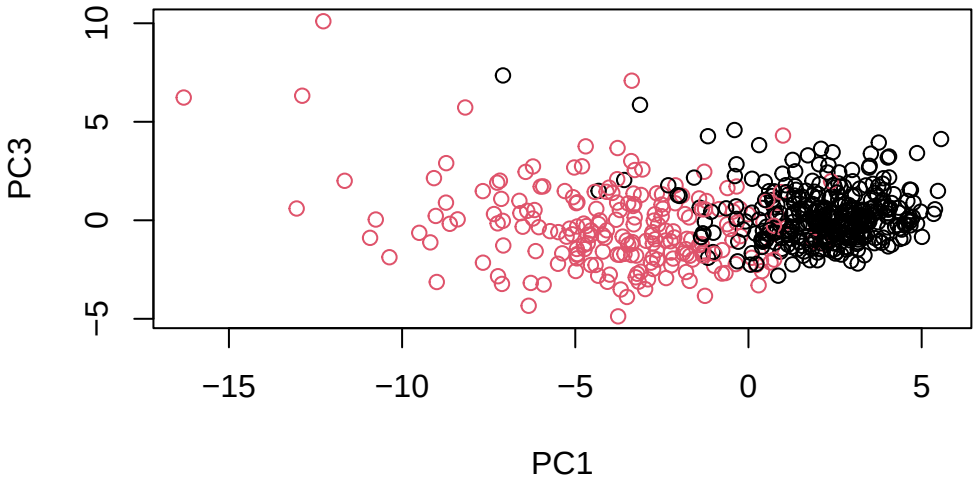
What stands out to me is a lot of the pink variables are not very centered and a lot of the black variables are very centered. This plot is a mess, so w're going to make a better one below.

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



Q8

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

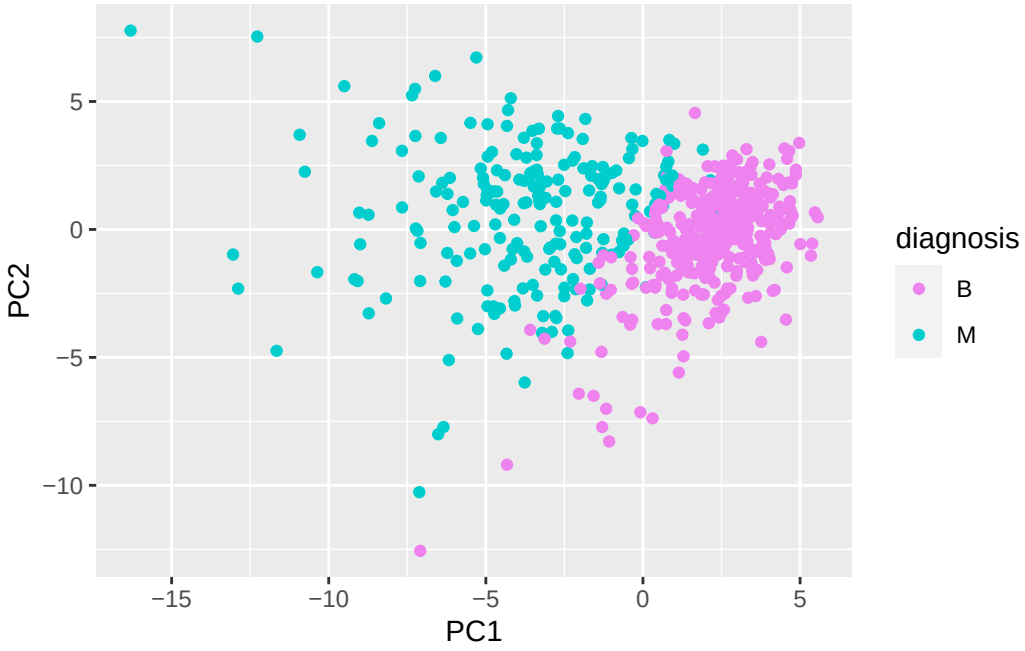


In general, these two plots seem similar in that they both have one more dense cluster and one less dense cluster that segregate the benign and malignant samples, although PC3 had a less dense cluster than PC2. There is also less overlap between red and black in the PC1 vs PC2 plot as opposed to the PC1 vs PC3 plot.

Now, we're going to move into ggplot.

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

```
ggplot(df)+
  aes(PC1, PC2, col=diagnosis)+
  geom_point()+
  scale_color_manual(values=c("M"="cyan3","B"="violet"))
```



Variance explained:

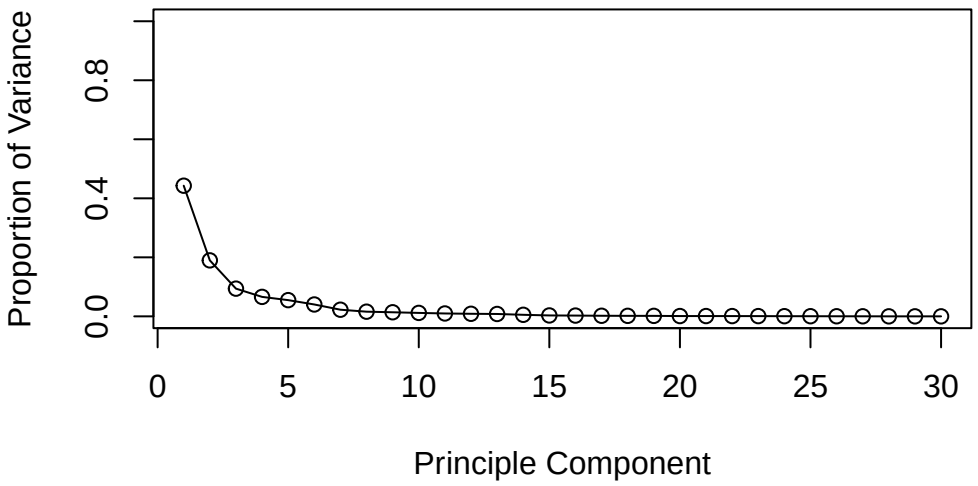
```
pr.var<-wisc.pr$sdev^2
head(pr.var)
```

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

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

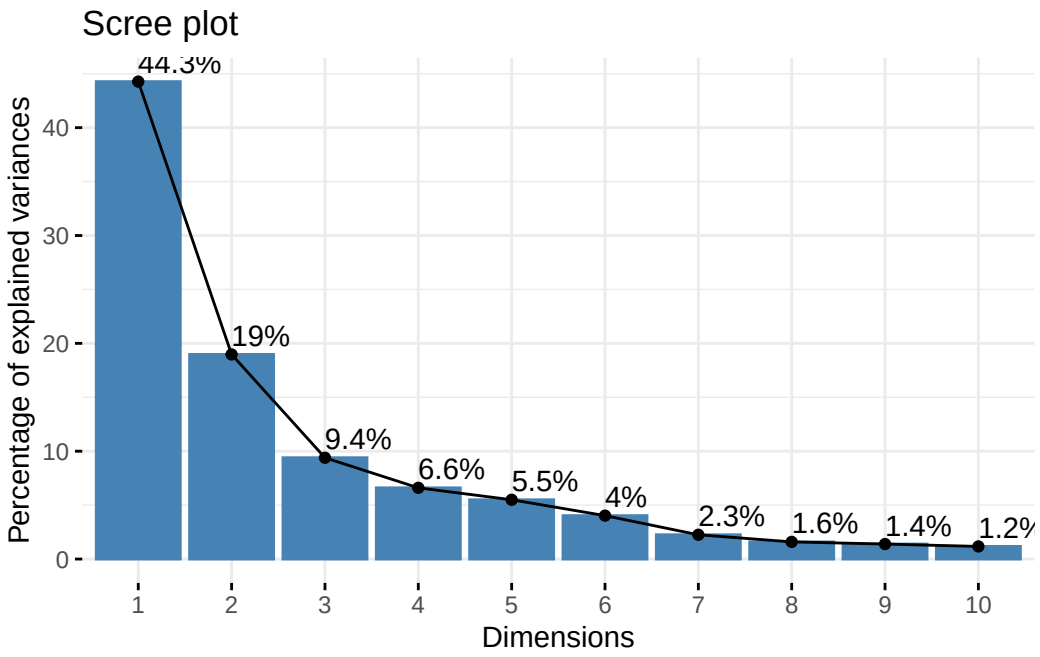
```
plot(pve,xlab="Principle Component",ylab="Proportion of Variance",ylim=c(0,1),type="o")
```




```
#install.packages("factoextra")
library(factoextra)
```

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

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



Q9

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

```
concave.points_mean
         -0.2608538
```

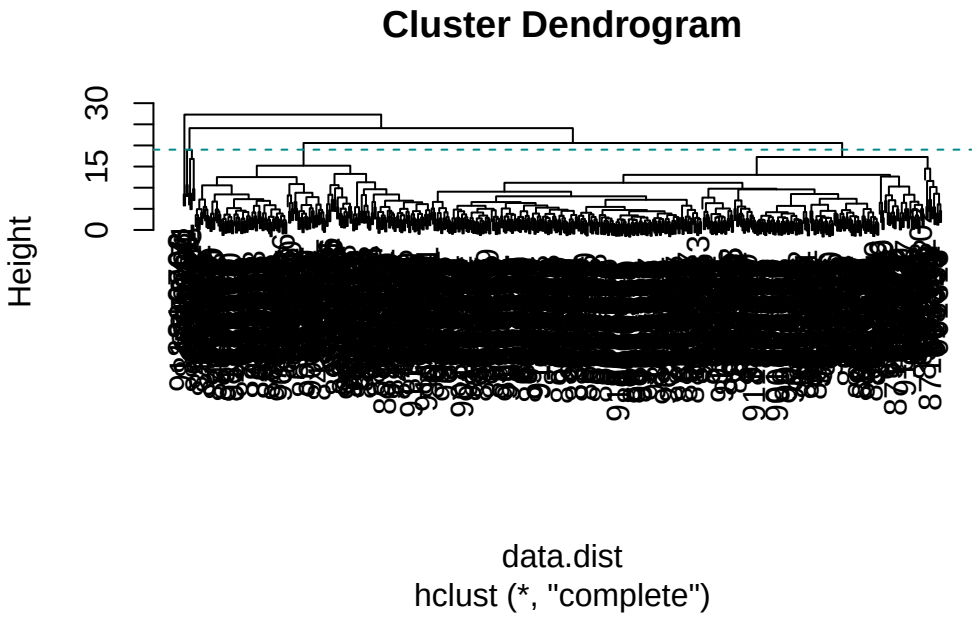
This tells us that 26% of the variance is due to concave.points_mean.

Hierarchal Clustering

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

Q10

```
plot(wisc.hclust)
abline(h=19,col="darkcyan",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
```

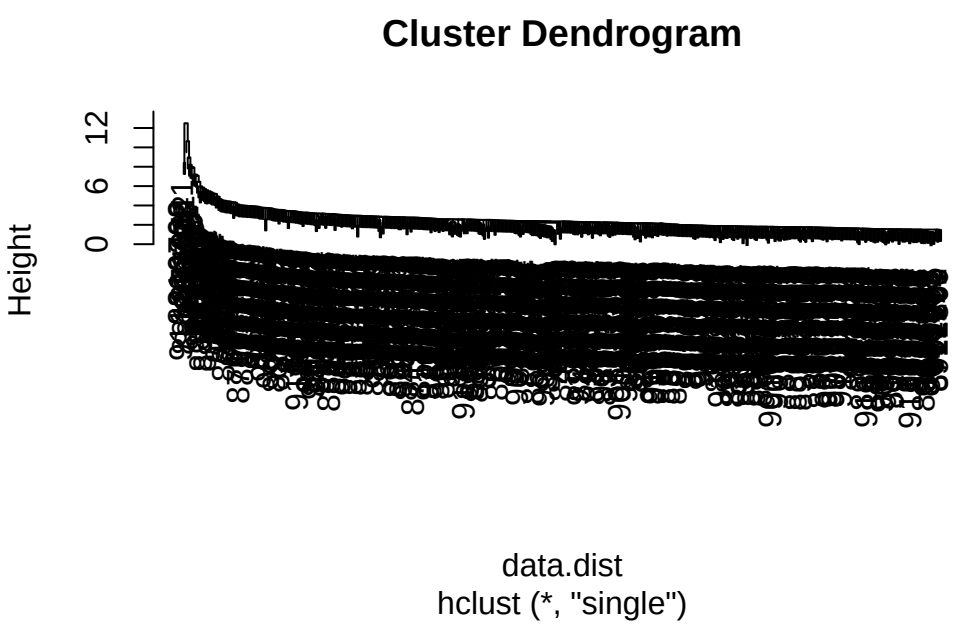
Q11

```
wisc.hclust.clust<-cutree(wisc.hclust,k=10)
table(wisc.hclust.clust,diagnosis)
```

```
                diagnosis
wisc.hclust.clust   B   M
               1   12  86
               2    0  59
               3    0   3
               4  331  39
               5    0  20
               6    2   0
               7   12   0
               8    0   2
               9    0   2
               10   0   1
```

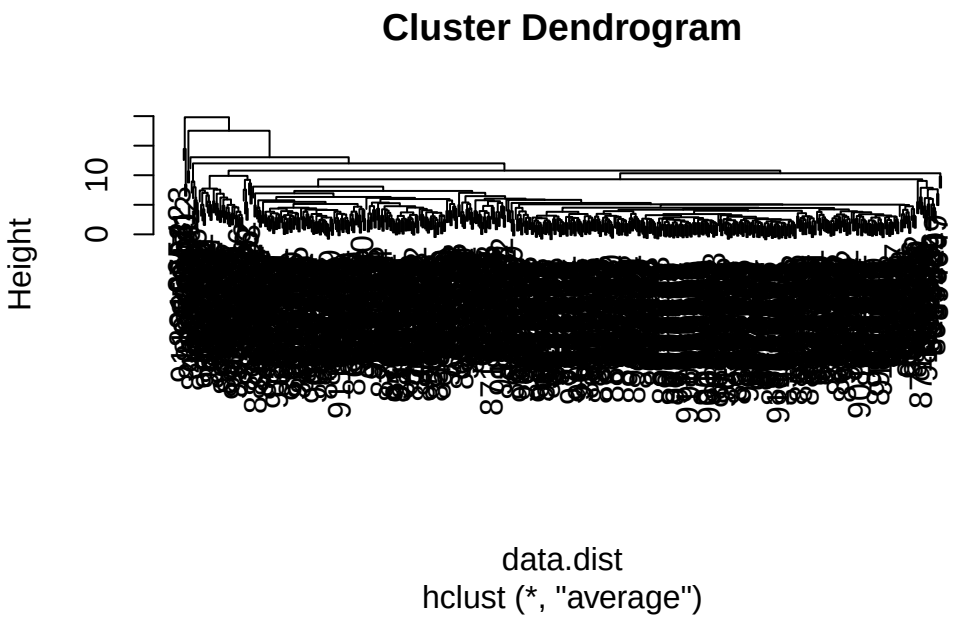
Q12

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



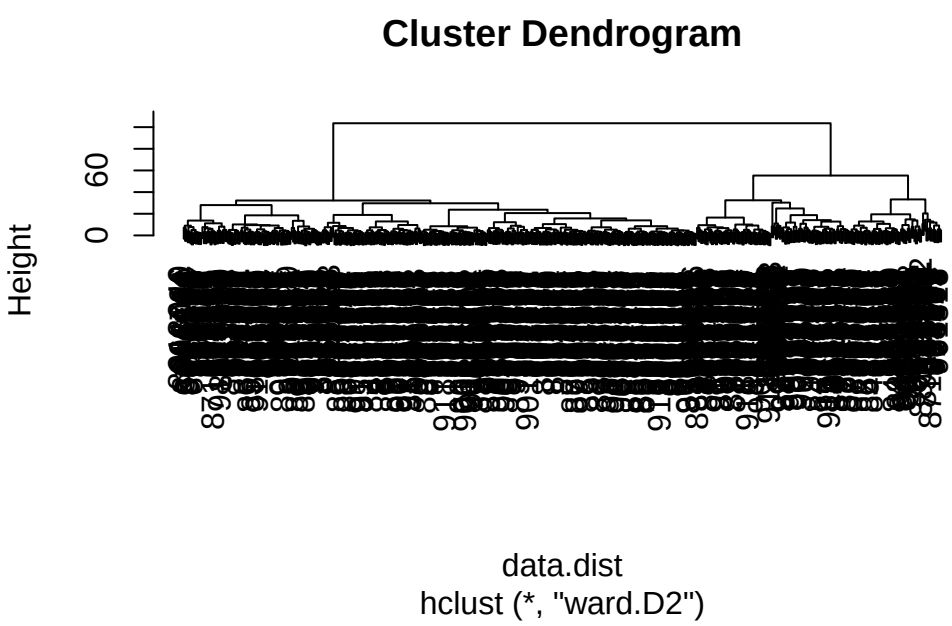
Single certainly isn't my favorite because it looks very heavy on one side.

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



Average looks like it groups better than single, for sure, and potentially slightly worse than complete.

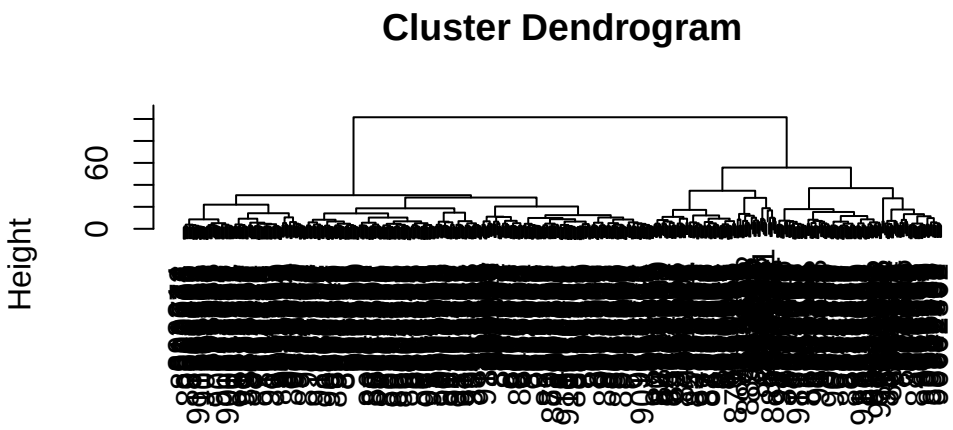
```
wisc.hclust.w<-hclust(data.dist,method="ward.D2")
plot(wisc.hclust.w)
```



Either ward.D2 or complete look like they're the most all encompassing

Combining methods

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




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

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

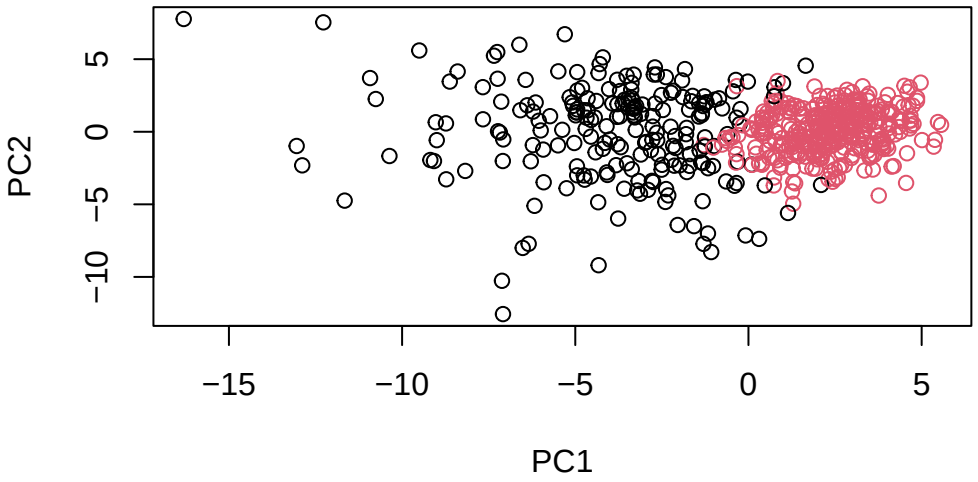
```
g<-as.factor(grps)
levels(g)
```

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

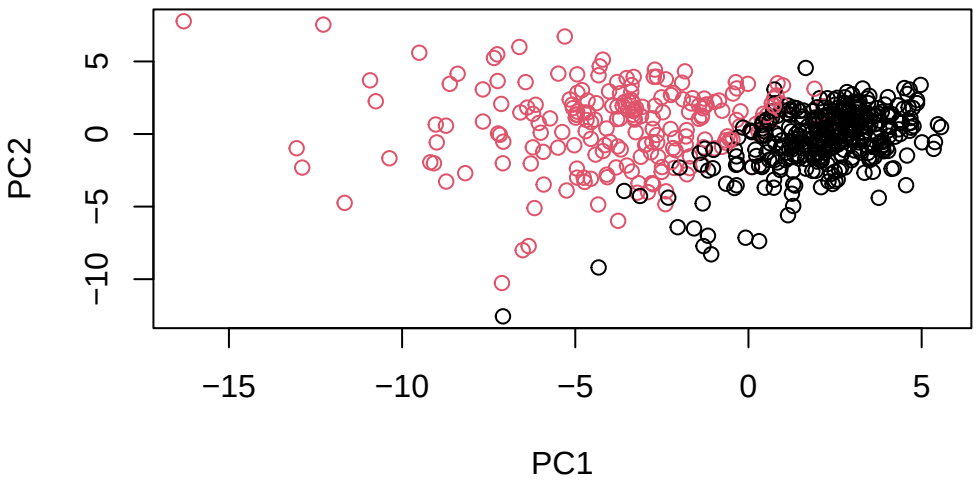
```
g<-relevel(g,2)
levels(g)
```

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

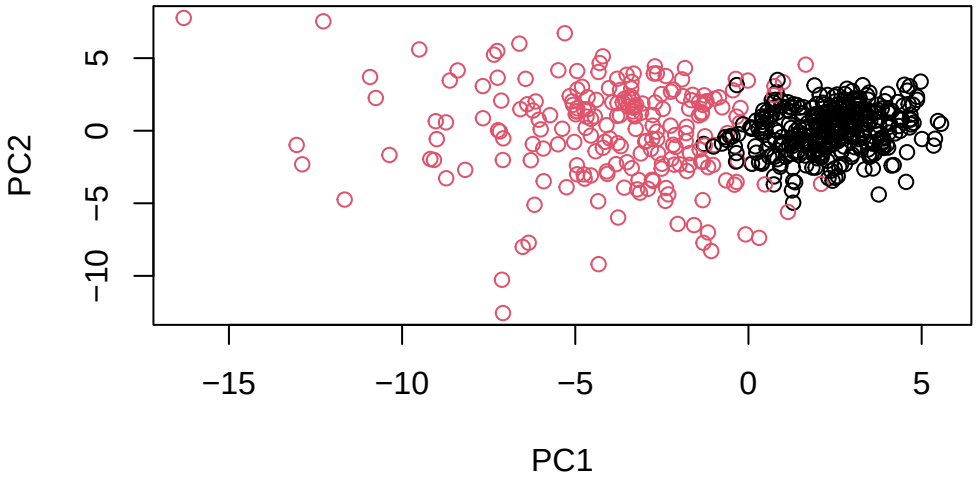
```
plot(wisc.pr$x[,1:2],col=grps)
```




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




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



I edited the group level as I went, so it's all in one code chunk instead of spread out.

Q13

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

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

There's still a lot of overlap in our clusters between diagnoses. There's around 50 people here who have the opposite diagnosis to the majority in their cluster.

Q14

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

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

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

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

It seems that the PCA has slightly better clustering. It also takes 4 clusters before PCA to get similar data to what can be done in 2 clusters after PCA.

```
table(diagnosis)
```

```
diagnosis
  B   M
357 212
```

Sensitivity pre PCA=0.877 Specificit ypre PCA=1.07 Sensitivity post PCA=1.01 Specificity post PCA=.0989

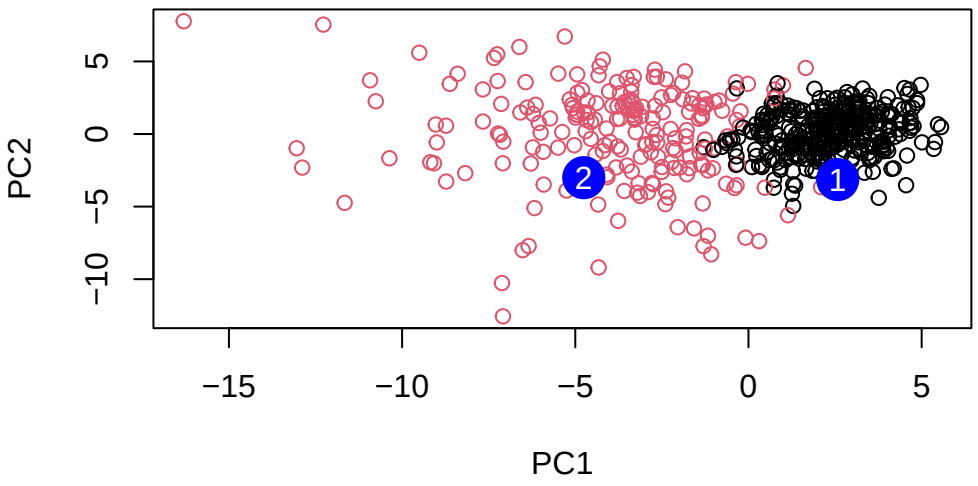
After PCA, sensitivty and specificity are both closer to 1, indicating they're better.

Prediction

```
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")
```



Q16

We should follow up with patient 2.