# Class 8: Breast Cancer Mini Project

Lilia Jimenez (PID:A16262599)

Before we get stuck into project work

read the data (lab 7)

```
url2 <- "https://tinyurl.com/expression-CSV"
rna.data <- read.csv(url2, row.names=1)
head(rna.data)
```

```
      wt1 wt2  wt3  wt4 wt5 ko1 ko2 ko3 ko4 ko5
gene1  439 458  408  429 420  90  88  86  90  93
gene2  219 200  204  210 187 427 423 434 433 426
gene3 1006 989 1030 1017 973 252 237 238 226 210
gene4  783 792  829  856 760 849 856 835 885 894
gene5  181 249  204  244 225 277 305 272 270 279
gene6  460 502  491  491 493 612 594 577 618 638
```

Q. How many genes are in this dataset?

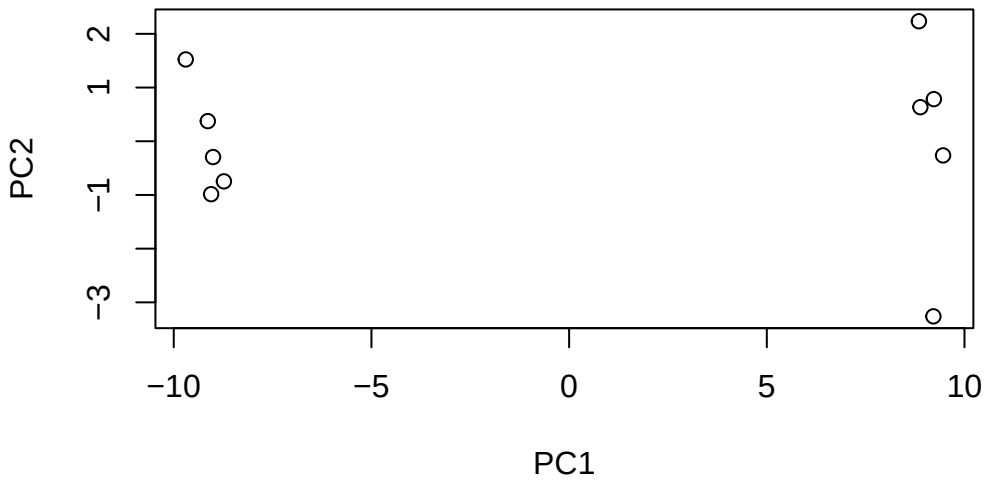
```
nrow(rna.data)
```

```
[1] 100
```

## Run PCA

```
## Again we have to take the transpose of our data
pca <- prcomp(t(rna.data), scale=TRUE)

## Simple un polished plot of pc1 and pc2
plot(pca$x[,1], pca$x[,2], xlab="PC1", ylab="PC2")
```



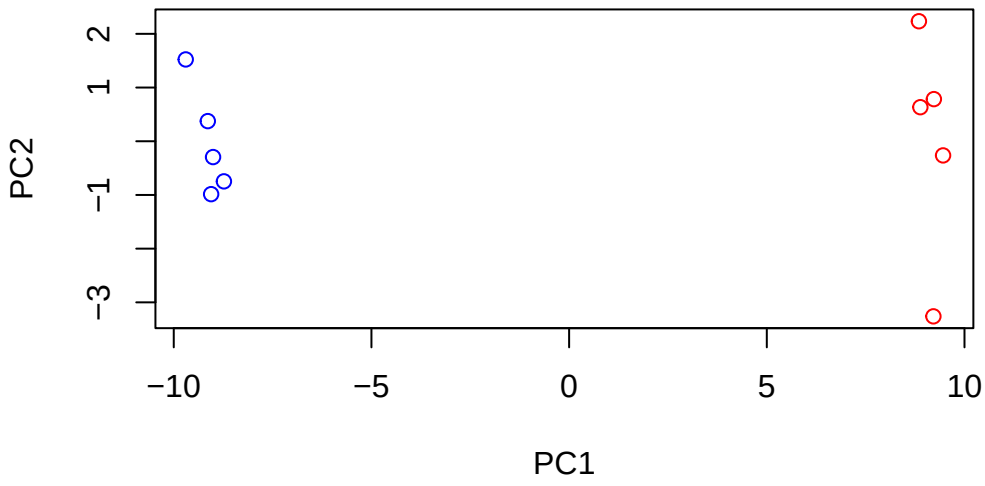

```
summary(pca)
```

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     9.6237 1.5198 1.05787 1.05203 0.88062 0.82545 0.80111
Proportion of Variance 0.9262 0.0231 0.01119 0.01107 0.00775 0.00681 0.00642
Cumulative Proportion  0.9262 0.9493 0.96045 0.97152 0.97928 0.98609 0.99251
                           PC8     PC9      PC10
Standard deviation     0.62065 0.60342 3.345e-15
Proportion of Variance 0.00385 0.00364 0.000e+00
Cumulative Proportion  0.99636 1.00000 1.000e+00
```

```
#We have 5 wt and 5 ko samples
mycols<- c(rep("blue",5), rep("red",5))
mycols
```

```
 [1] "blue" "blue" "blue" "blue" "blue" "red"  "red"  "red"  "red"  "red"
```

```
plot(pca$x[,1], pca$x[,2], xlab="PC1", ylab="PC2", col=mycols)
```




I could examibe which genes contribute the most to the first pc

```
head(sort(abs(pca$rotation[,1]),decreasing = T))
```

```
  gene100    gene66    gene45    gene68    gene98    gene60
0.1038708 0.1038455 0.1038402 0.1038395 0.1038372 0.1038055
```

## Analysis of Human Breast cells

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

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

Now I want to make sure I remove the column from my data set for analysis.

```
wisc.data<-wisc.df[,-1]
#head(wisc.data)
```

Q1. How many observations are in this dataset?

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

```
[1] 31
```

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

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

```
  B   M
357 212
```

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

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

```
[1] 10
```

##Principal Component Analysis

Here we will use 'prcomp()' on the 'wisc.data' object - the one without the diagnosis column.

We can look at the means and sd of each column. If they are similar then we are all good to go. If not, we should use 'scale=TRUE'

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

```
    radius_mean     texture_mean   perimeter_mean        area_mean
    14.12729174      19.28964851      91.96903339     654.88910369
smoothness_mean compactness_mean
     0.09636028       0.10434098
```

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

```
    radius_mean     texture_mean   perimeter_mean        area_mean
     3.52404883       4.30103577      24.29898104     351.91412918
smoothness_mean compactness_mean
     0.01406413       0.05281276
```

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

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

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

44.27%

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

3 PCs

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

7 PCs

## Plotting the PCA results

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

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

It is difficult to read. Everything is just overlayed on top of eachother so I cant read most of it.

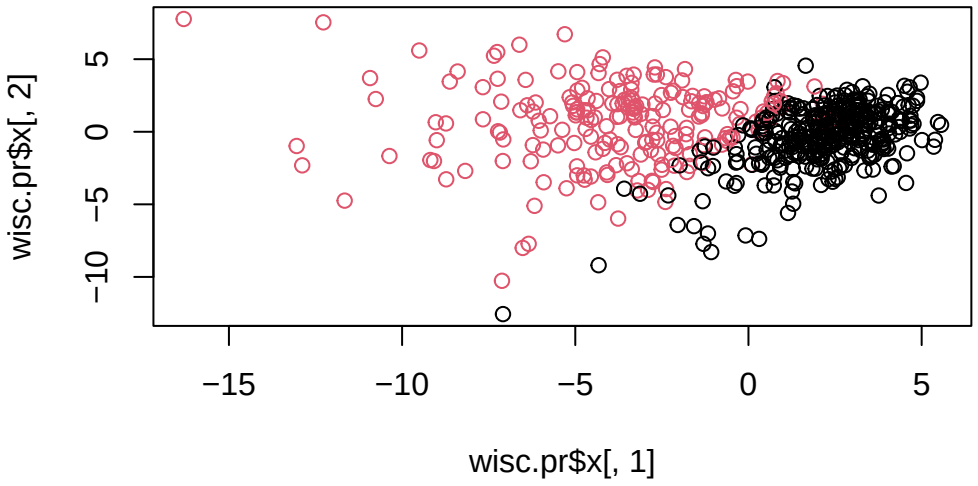
we need our own plot

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

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

$class
[1] "prcomp"
```

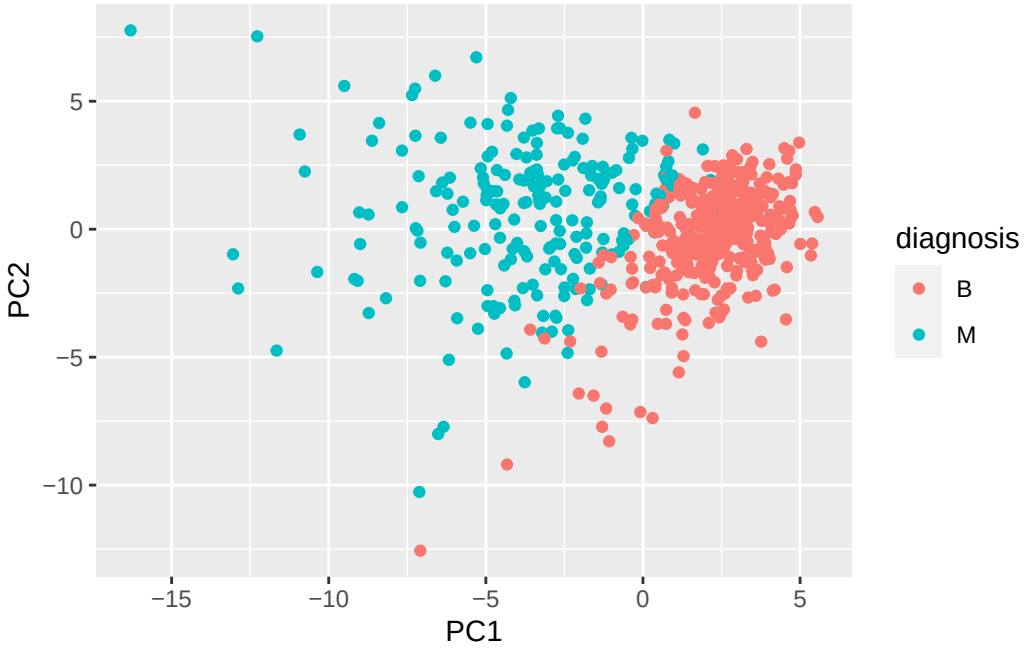
```
plot(wisc.pr$x[,1], wisc.pr$x[,2], col=diagnosis)
```



```
library(ggplot2)

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

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




#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?

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

```
[1] -0.2608538
```

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

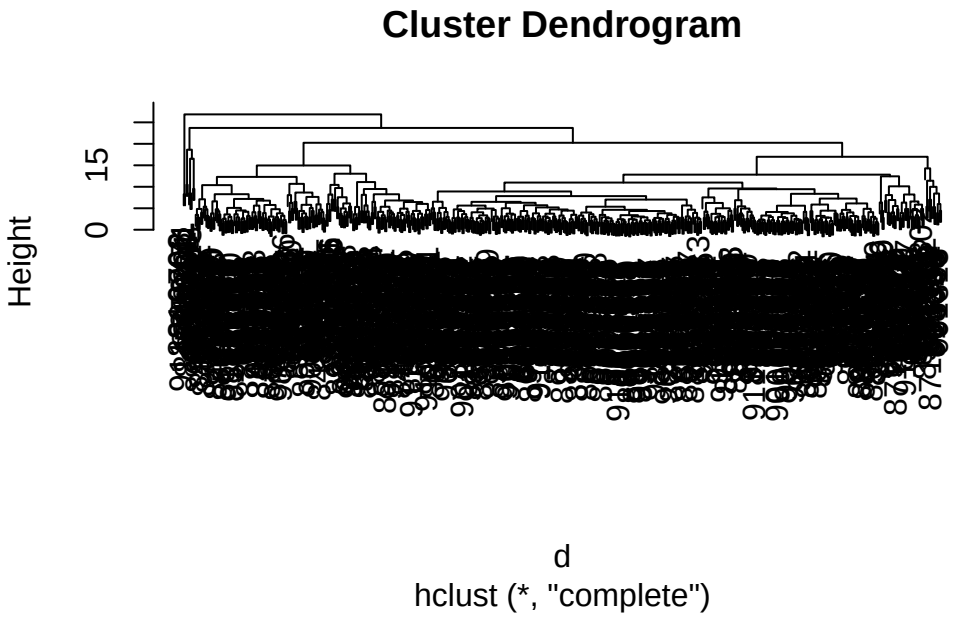
```
tbl<-summary(wisc.pr)
which(tbl$importance[3,]>0.8)[1]
```

```
PC5
  5
```

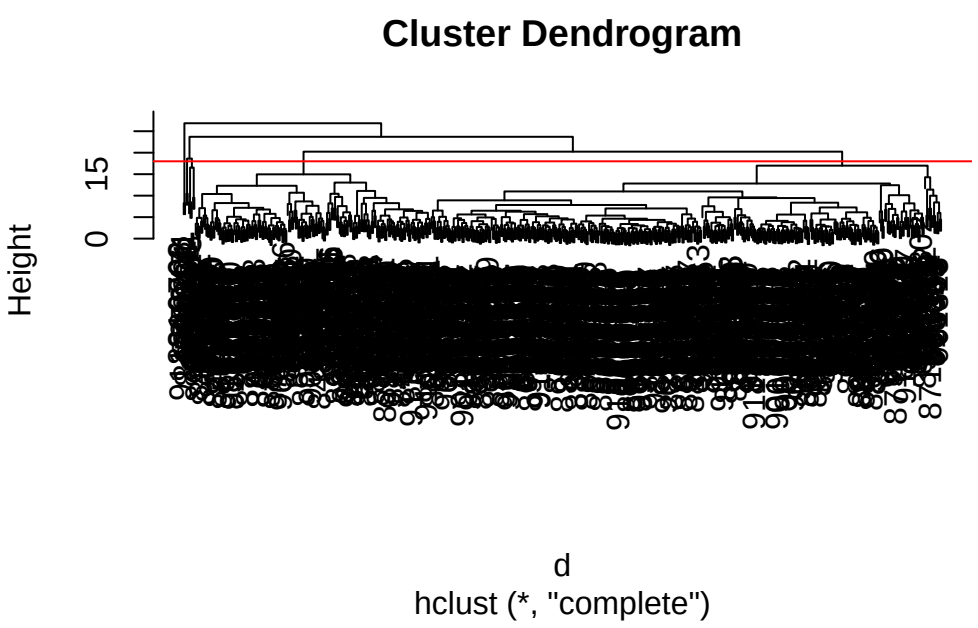
##Hierarchial clustering

The main function for hierarchial clustering is 'hclust()' it takes a distance matrix

```
d<-dist(scale(wisc.data))
wisc.hclust<-hclust(d)
plot(wisc.hclust)
```




```
plot(wisc.hclust)
abline(h=18, col="red")
```




```
grps<-cutree(wisc.hclust, h=18)
table(grps)
```

```
grps
  1   2   3   4   5
177   5 383   2   2
```

Come back here. Later to see how our cluster grps correspond to M or B groups.

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

##5. Combining methods

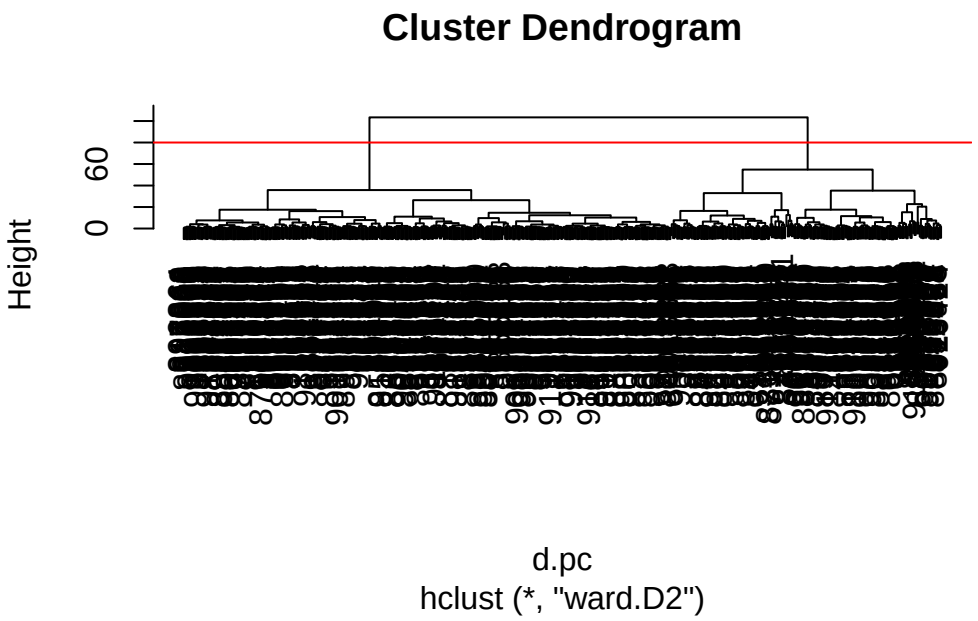
Here we will perform clustering on our PCA results rather than on original data.

In other words we will cluster using 'wisc.pr$x' - our new better variables or PCs. We can choose as many or as few PCs as we like.

```
d.pc<-dist(wisc.pr$x[,1:3])

wisc.pr.hclust<- hclust(d.pc, method = "ward.D2")
plot(wisc.pr.hclust)
```

```
abline(h=80, col="red")
```




```
grps<-cutree(wisc.pr.hclust, h=80)
table(grps)
```

```
grps
  1   2
203 366
```

we can use 'table()' function to make a cross-table ass as just a count table

```
table(diagnosis)
```

```
diagnosis
  B   M
357 212
```

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

```
    diagnosis
grps   B   M
   1  24 179
   2 333  33
```

the results indicate that our cluster 1 mostly captures cancer(M) and our cluster 2 mostly capture healthy (B) sample/individuals.

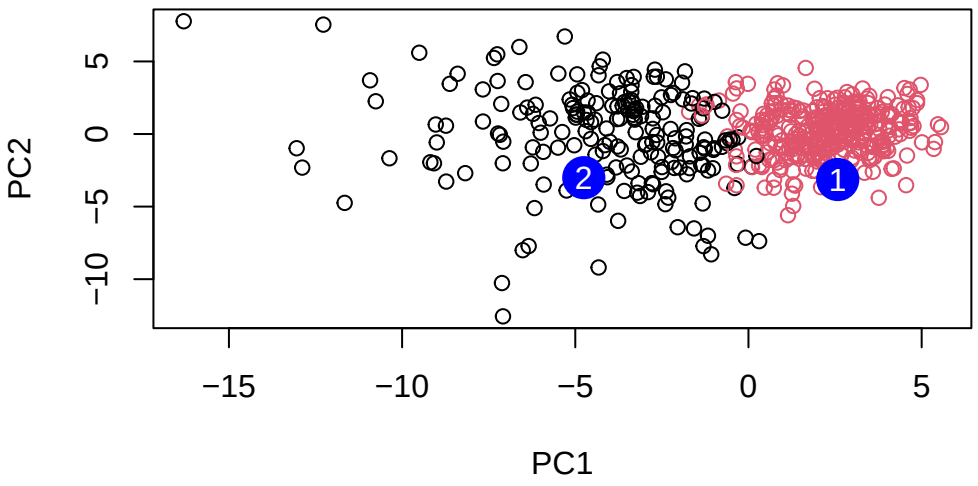
## 7. Prediction

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

and plot this up

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




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

Patients in group 2 because they are more spread out so we dont know which one truly belong in that group.