Class 08 Mini-Project

Josie (A11433761)

Side_Note: Let's look at the mean value of every column:

```
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
```

```
apply(mtcars, 2, mean)
```

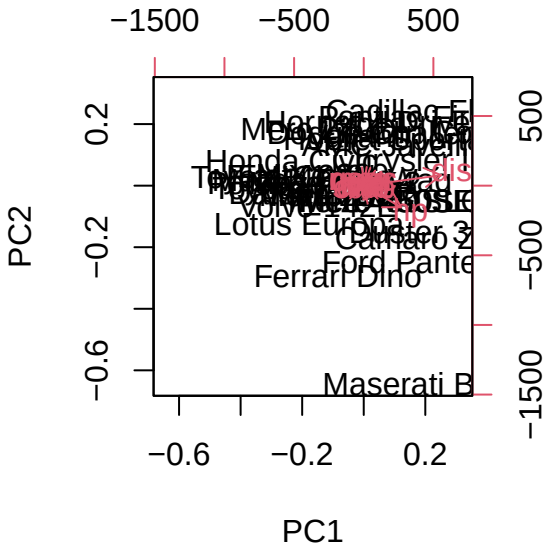
```
      mpg        cyl       disp         hp       drat         wt       qsec
20.090625   6.187500 230.721875 146.687500   3.596563   3.217250  17.848750
       vs         am       gear       carb
 0.437500   0.406250   3.687500   2.812500
```

Let's look at "spread" via `sd()`

```
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
```

```
pca<-prcomp(mtcars)
biplot(pca)
```



Let's try scalling the data:

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

```
                           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
```

What is the mean of each "dimension" /column in mtscale?

```
round(apply(mtscale, 2, mean, 3))
```

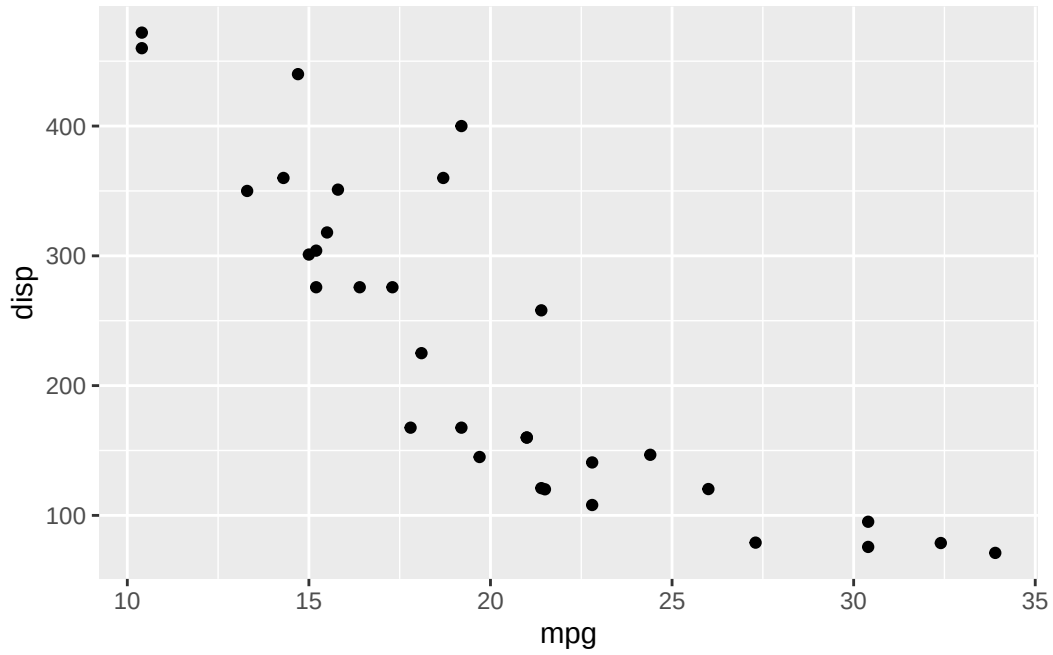
```
 mpg  cyl disp   hp drat   wt qsec   vs   am gear carb
   0    0    0    0    0    0    0   -1   -1    0   -1
```

```
round(apply(mtscale, 2, sd, 3))
```

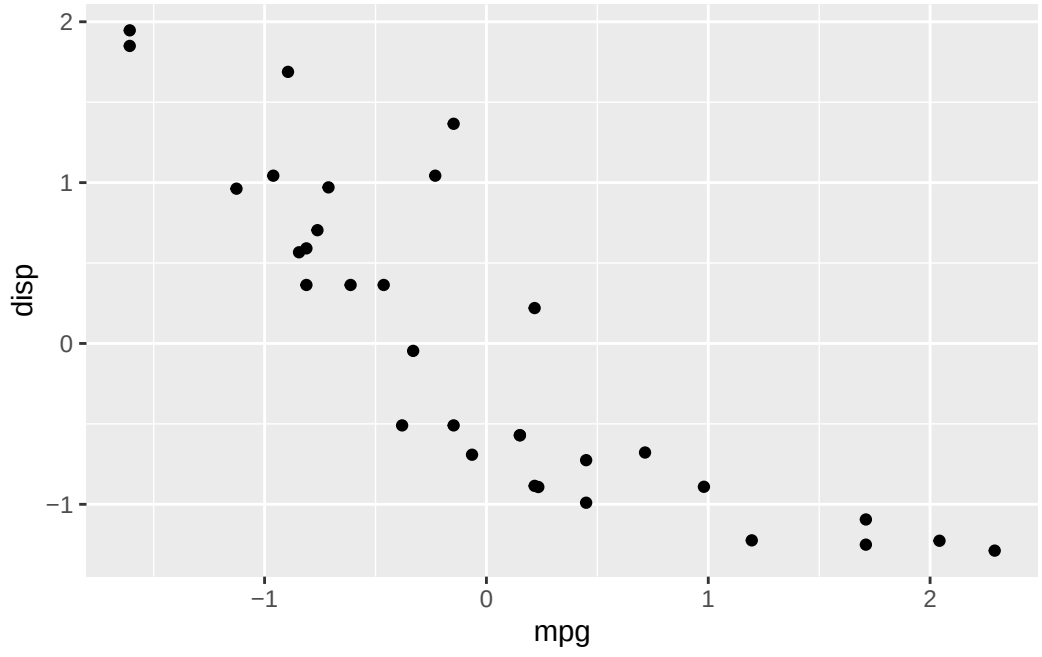
```
 mpg  cyl disp   hp drat   wt qsec   vs   am gear carb
   1    1    1    1    1    1    1    1    1    1    1
```

Let's plot `mpg` vs `disp` for both mtcars and after the scaled data in `mtscale`

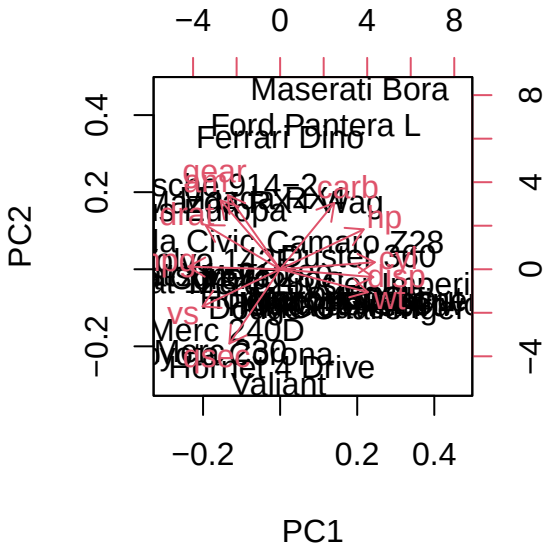
```
library(ggplot2)
ggplot(mtcars)+
  aes(mpg,disp)+
  geom_point()
```




```
ggplot(mtscale)+
  aes(mpg,disp)+
  geom_point()
```



```
pca2<-prcomp(mtscale)
biplot(pca2)
```



Breast Cancer FNA data

Download file and move it to project folder or directly download to project folder

```
fna.data <- "WisconsinCancer.csv"
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  X
842302                   0.11890 NA
842517                   0.08902 NA
84300903                 0.08758 NA
```

```
84348301                    0.17300 NA
84358402                    0.07678 NA
843786                      0.12440 NA
```

Removing `diagnosis` by creating data frame that removes the first column

```
wisc.data<-wisc.df[,-1]
wisc.data<-wisc.data[,-31]
diagnosis<-as.factor(wisc.df$diagnosis)

#How many rows? Patients
nrow(wisc.df)
```

```
[1] 569
```

```
#How many M (cancer) and B (benign)?
table(wisc.df$diagnosis)
```

```
  B   M
357 212
```

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

```
[1] 10
```

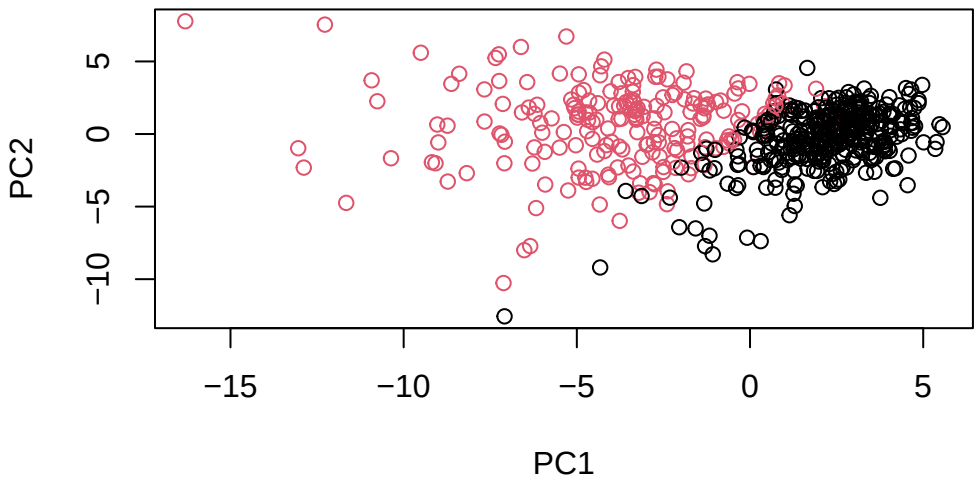
Principal Component Analysis

We want to scale our data before PCA by setting scale=TRUE

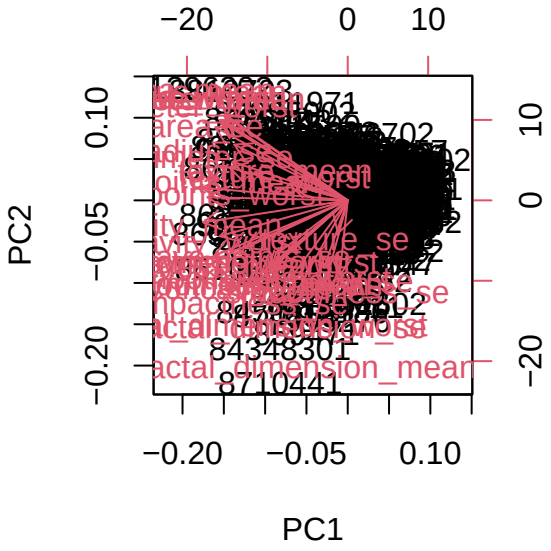
```
#colMeans(wisc.data)
#apply(wisc.data, 2,sd)
wisc.pr <- prcomp(wisc.data,scale=TRUE)
x<-summary(wisc.pr)
x$importance
```

```
                            PC1       PC2       PC3       PC4       PC5       PC6
Standard deviation     3.644394 2.385656 1.678675 1.407352 1.284029 1.098798
Proportion of Variance 0.442720 0.189710 0.093930 0.066020 0.054960 0.040250
Cumulative Proportion  0.442720 0.632430 0.726360 0.792390 0.847340 0.887590
                              PC7       PC8       PC9      PC10      PC11
Standard deviation     0.8217178 0.6903746 0.6456739 0.5921938 0.5421399
Proportion of Variance 0.0225100 0.0158900 0.0139000 0.0116900 0.0098000
Cumulative Proportion  0.9101000 0.9259800 0.9398800 0.9515700 0.9613700
                             PC12      PC13      PC14      PC15      PC16
Standard deviation     0.5110395 0.4912815 0.3962445 0.3068142 0.2826001
Proportion of Variance 0.0087100 0.0080500 0.0052300 0.0031400 0.0026600
Cumulative Proportion  0.9700700 0.9781200 0.9833500 0.9864900 0.9891500
                             PC17      PC18      PC19      PC20      PC21
Standard deviation     0.2437192 0.2293878 0.2224356 0.1765203 0.1731268
Proportion of Variance 0.0019800 0.0017500 0.0016500 0.0010400 0.0010000
Cumulative Proportion  0.9911300 0.9928800 0.9945300 0.9955700 0.9965700
                             PC22      PC23      PC24      PC25      PC26
Standard deviation     0.1656484 0.1560155 0.1343689 0.1244238 0.0904303
Proportion of Variance 0.0009100 0.0008100 0.0006000 0.0005200 0.0002700
Cumulative Proportion  0.9974900 0.9983000 0.9989000 0.9994200 0.9996900
                              PC27       PC28       PC29       PC30
Standard deviation     0.08306903 0.0398665 0.02736427 0.01153451
Proportion of Variance 0.00023000 0.0000500 0.00002000 0.00000000
Cumulative Proportion  0.99992000 0.9999700 1.00000000 1.00000000
```

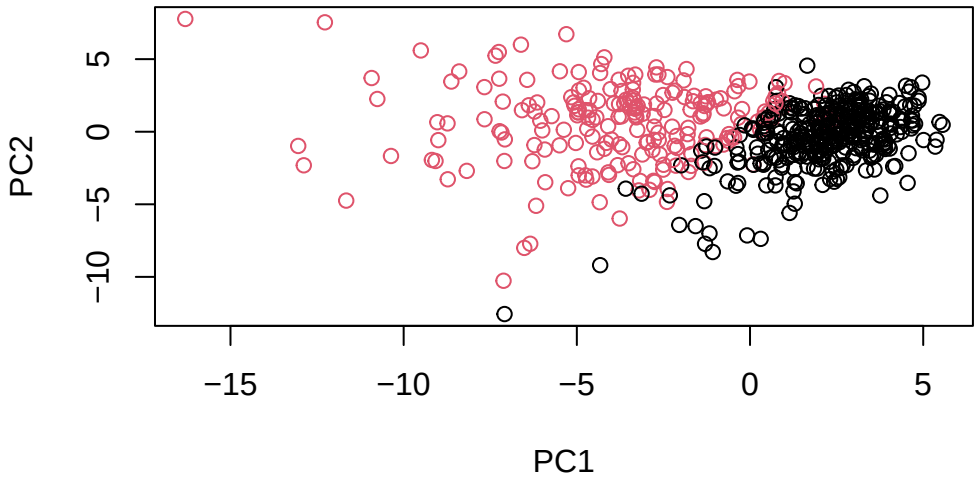
```
plot(wisc.pr$x,col=diagnosis)
```




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



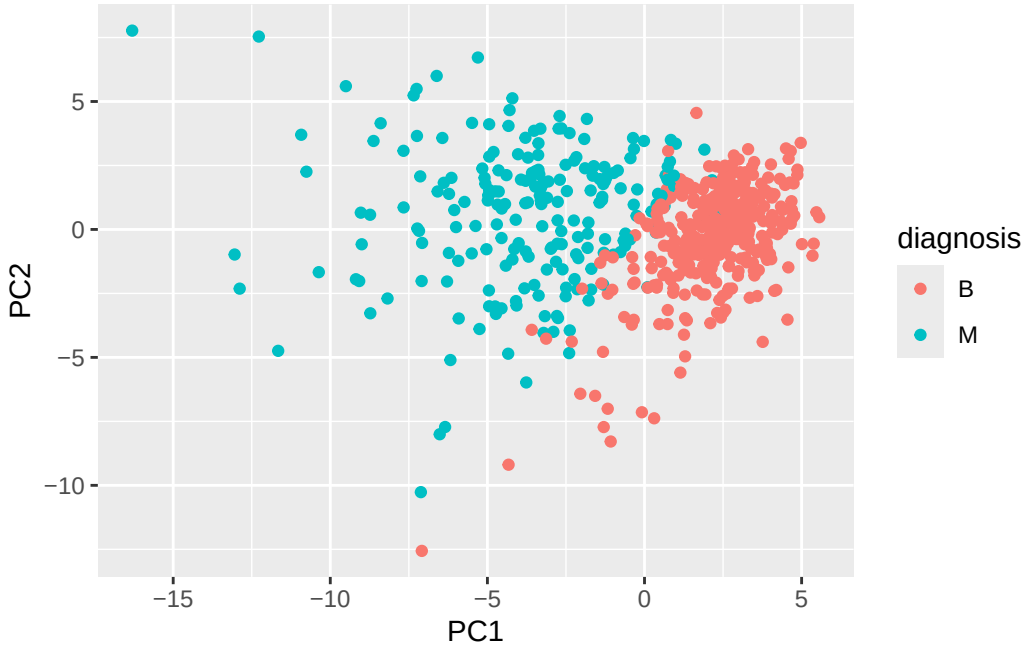
```
plot(wisc.pr$x, col=diagnosis, xlab="PC1", ylab="PC2")
```




```
# Create a data.frame for ggplot
df <- as.data.frame(wisc.pr$x)
df$diagnosis <- diagnosis

# Load the ggplot2 package
library(ggplot2)

# Make a scatter plot colored by diagnosis
ggplot(df) +
  aes(PC1, PC2, col=diagnosis) +
  geom_point()
```

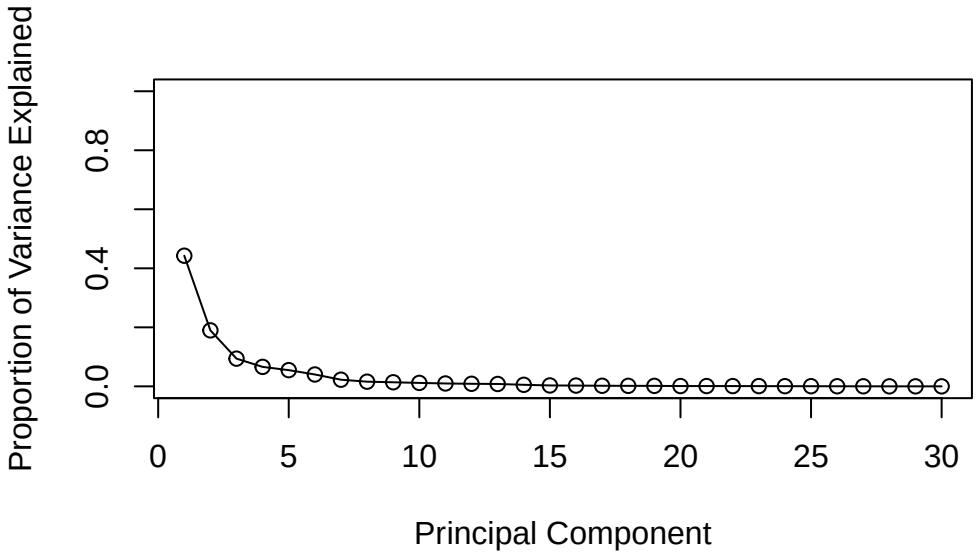


Variance explained

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

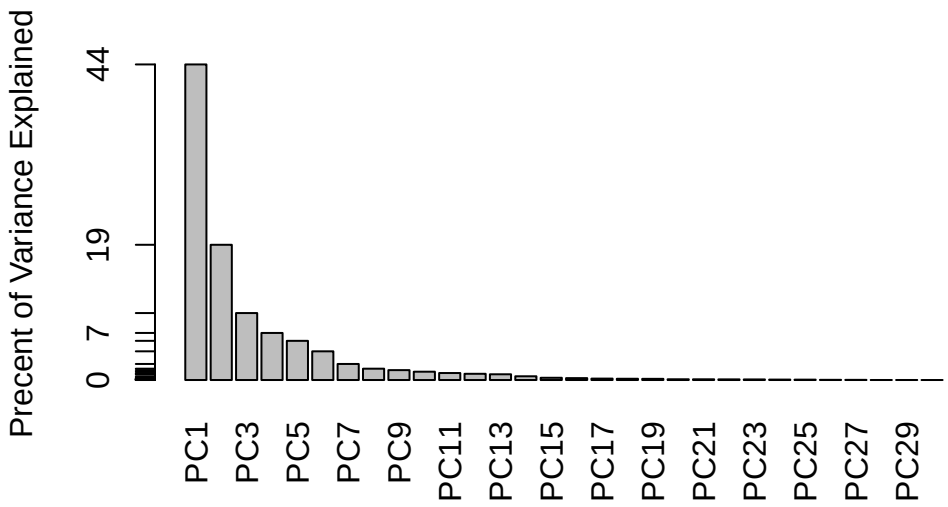
```
[1] 13.281608  5.691355  2.817949  1.980640  1.648731  1.207357
```

```
pve<-pr.var/30 #total variance divided by number of PC
plot(pve, xlab = "Principal Component",
     ylab = "Proportion of Variance Explained",
     ylim = c(0, 1), type = "o")
```



Alternative plot data driven y-axis

```
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

```
head(wisc.pr$rotation[,1])
```

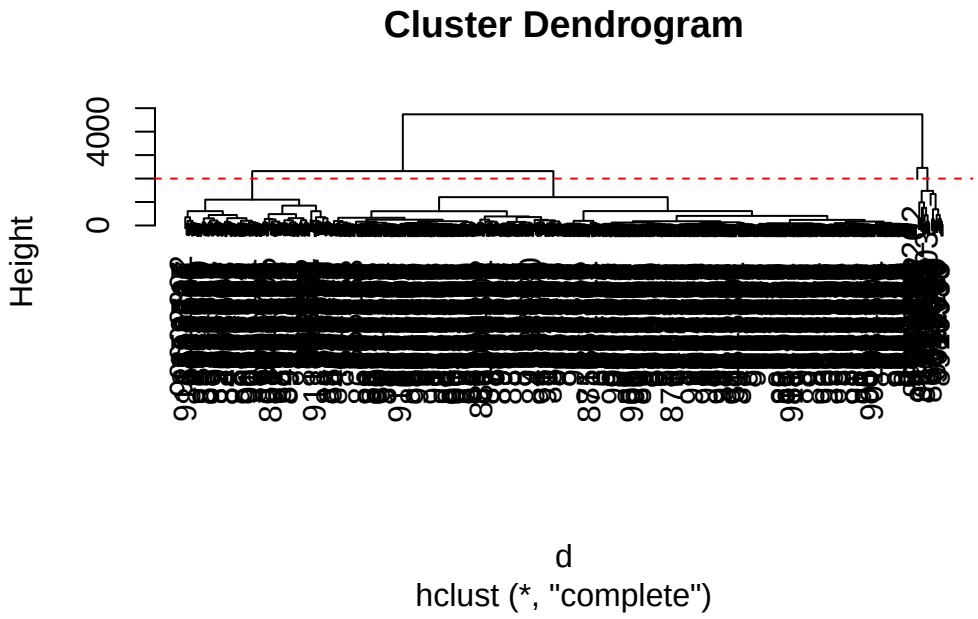
```
    radius_mean     texture_mean   perimeter_mean        area_mean
     -0.2189024       -0.1037246       -0.2275373       -0.2209950
smoothness_mean compactness_mean
     -0.1425897       -0.2392854
```

Clustering

```
km<-kmeans(wisc.data,centers=2)
table(km$cluster)
```

```
  1   2
438 131
```

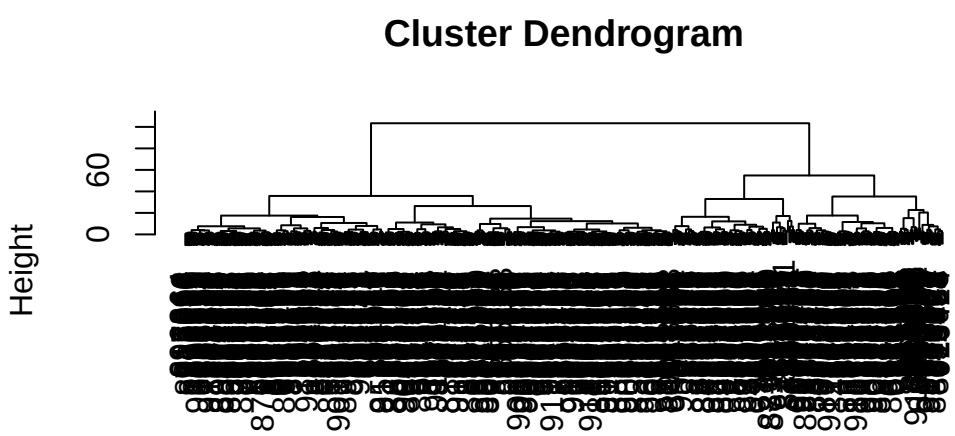
```
d<-dist(wisc.data)
hc<-hclust(d)
plot(hc)
abline(h=2000, col="red", lty=2)
```



Cluster in PC

Use my PCA results as a baisis of clustering

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



Cut this tree to yield 2 groups/clusters

```
grps<-cutree(hc,k=2)
table(grps)
```

```
grps
  1   2
203 366
```

```
table(diagnosis)
```

```
diagnosis
  B   M
357 212
```

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

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

Using different methods

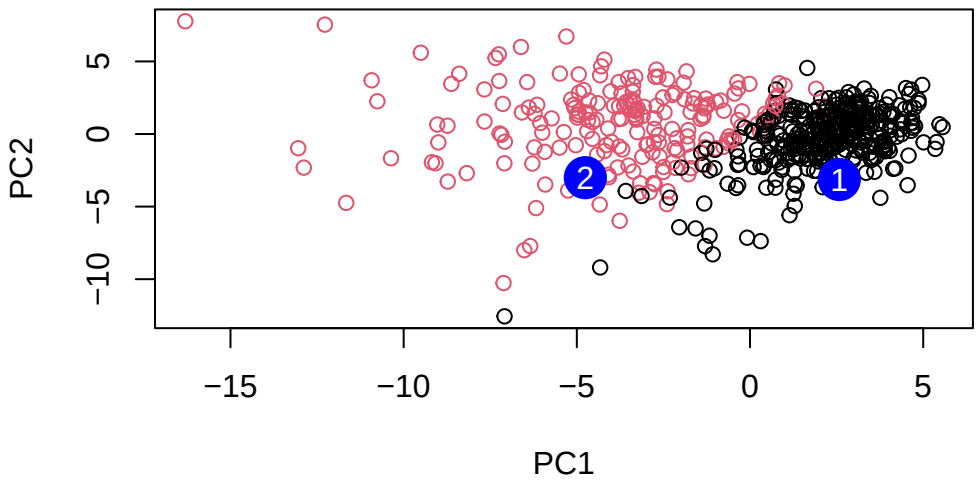
"single", "complete", "average" and "ward.D2"

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=diagnosis)
points(npc[,1], npc[,2], col="blue", pch=16, cex=3)
text(npc[,1], npc[,2], c(1,2), col="white")
```


PC2
5
0
−5
−10
−15
−10
−5
0
5
PC1
2
1