Class 8:PCA Mini Project

Troy Lee A17078296

```
# Save your input data file into your Project directory
fna.data <- "https://bioboot.github.io/bimm143_S20/class-material/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

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

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

```
diagnosis
  B   M
357 212
```

Remove this first `diagnosis` column from the dataset as I don't want to pass this to pCA etc. It is essentialy the expert "answer" that we will compare our analysis results to.

```
# We can use -1 here to remove the first column
wisc.data <- wisc.df[,-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
```

Q1. How many observations are in this dataset?

```
dim(wisc.df)
```

```
[1] 569  31
```

- 569 observation

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

- 212 diagnosis

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

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

```
[1] 10
```

- 10 _mean

Principal Component Analysis

```
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"

See what is in our PCA result object:

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

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

$class
[1] "prcomp"
```
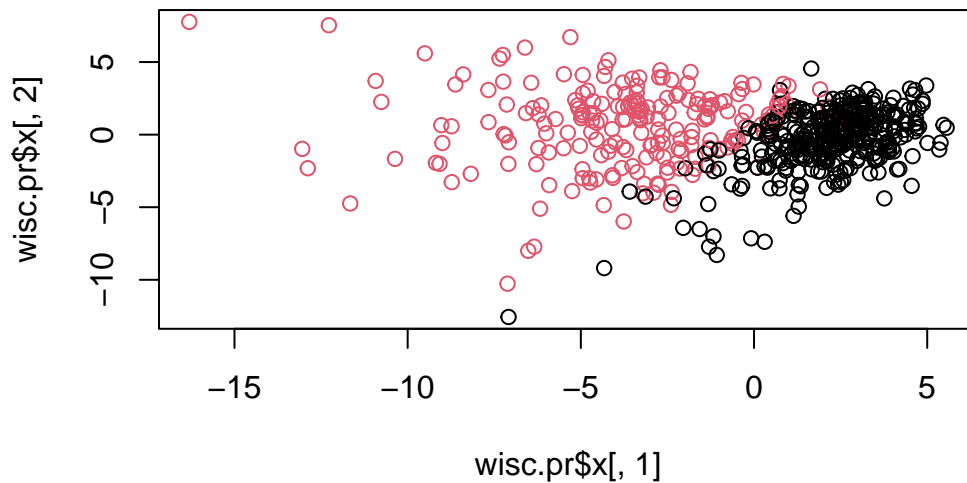
```
head(wisc.pr$x)
```

```
                 PC1        PC2        PC3       PC4        PC5         PC6
842302    -9.184755  -1.946870 -1.1221788 3.6305364  1.1940595  1.41018364
842517    -2.385703   3.764859 -0.5288274 1.1172808 -0.6212284  0.02863116
84300903  -5.728855   1.074229 -0.5512625 0.9112808  0.1769302  0.54097615
84348301  -7.116691 -10.266556 -3.2299475 0.1524129  2.9582754  3.05073750
84358402  -3.931842   1.946359  1.3885450 2.9380542 -0.5462667 -1.22541641
843786    -2.378155  -3.946456 -2.9322967 0.9402096  1.0551135 -0.45064213
                   PC7         PC8         PC9       PC10       PC11       PC12
842302      2.15747152  0.39805698 -0.15698023 -0.8766305 -0.2627243 -0.8582593
842517      0.01334635 -0.24077660 -0.71127897  1.1060218 -0.8124048  0.1577838
84300903   -0.66757908 -0.09728813  0.02404449  0.4538760  0.6050715  0.1242777
84348301    1.42865363 -1.05863376 -1.40420412 -1.1159933  1.1505012  1.0104267
84358402   -0.93538950 -0.63581661 -0.26357355  0.3773724 -0.6507870 -0.1104183
843786      0.49001396  0.16529843 -0.13335576 -0.5299649 -0.1096698  0.0813699
                   PC13        PC14        PC15        PC16        PC17
842302       0.10329677 -0.690196797  0.601264078  0.74446075 -0.26523740
842517      -0.94269981 -0.652900844 -0.008966977 -0.64823831 -0.01719707
84300903    -0.41026561  0.016665095 -0.482994760  0.32482472  0.19075064
84348301    -0.93245070 -0.486988399  0.168699395  0.05132509  0.48220960
84358402     0.38760691 -0.538706543 -0.310046684 -0.15247165  0.13302526
843786      -0.02625135  0.003133944 -0.178447576 -0.01270566  0.19671335
                   PC18       PC19        PC20         PC21        PC22
842302      -0.54907956  0.1336499  0.34526111  0.096430045 -0.06878939
842517       0.31801756 -0.2473470 -0.11403274 -0.077259494  0.09449530
84300903    -0.08789759 -0.3922812 -0.20435242  0.310793246  0.06025601
84348301    -0.03584323 -0.0267241 -0.46432511  0.433811661  0.20308706
84358402    -0.01869779  0.4610302  0.06543782 -0.116442469  0.01763433
843786      -0.29727706 -0.1297265 -0.07117453 -0.002400178  0.10108043
                   PC23         PC24         PC25         PC26        PC27
842302       0.08444429  0.175102213  0.150887294 -0.201326305 -0.25236294
842517      -0.21752666 -0.011280193  0.170360355 -0.041092627  0.18111081
84300903    -0.07422581 -0.102671419 -0.171007656  0.004731249  0.04952586
84348301    -0.12399554 -0.153294780 -0.077427574 -0.274982822  0.18330078
84358402     0.13933105  0.005327110 -0.003059371  0.039219780  0.03213957
843786       0.03344819 -0.002837749 -0.122282765 -0.030272333 -0.08438081
                    PC28         PC29         PC30
842302      -0.0338846387  0.045607590  0.0471277407
842517       0.0325955021 -0.005682424  0.0018662342
84300903     0.0469844833  0.003143131 -0.0007498749
84348301     0.0424469831 -0.069233868  0.0199198881
```

```
84358402 -0.0347556386  0.005033481 -0.0211951203
843786     0.0007296587 -0.019703996 -0.0034564331
```

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



```
head(wisc.pr$x)
```

```
                 PC1          PC2          PC3        PC4        PC5          PC6
842302    -9.184755  -1.946870  -1.1221788  3.6305364  1.1940595   1.41018364
842517    -2.385703   3.764859  -0.5288274  1.1172808 -0.6212284   0.02863116
84300903  -5.728855   1.074229  -0.5512625  0.9112808  0.1769302   0.54097615
84348301  -7.116691 -10.266556  -3.2299475  0.1524129  2.9582754   3.05073750
84358402  -3.931842   1.946359   1.3885450  2.9380542 -0.5462667  -1.22541641
843786    -2.378155  -3.946456  -2.9322967  0.9402096  1.0551135  -0.45064213
                 PC7          PC8          PC9        PC10        PC11        PC12
842302    2.15747152   0.39805698 -0.15698023 -0.8766305 -0.2627243 -0.8582593
842517    0.01334635  -0.24077660 -0.71127897  1.1060218 -0.8124048  0.1577838
84300903 -0.66757908  -0.09728813  0.02404449  0.4538760  0.6050715  0.1242777
84348301  1.42865363  -1.05863376 -1.40420412 -1.1159933  1.1505012  1.0104267
84358402 -0.93538950  -0.63581661 -0.26357355  0.3773724 -0.6507870 -0.1104183
843786    0.49001396   0.16529843 -0.13335576 -0.5299649 -0.1096698  0.0813699
```

7

```
              PC13          PC14          PC15          PC16          PC17
842302     0.10329677 -0.690196797   0.601264078   0.74446075 -0.26523740
842517    -0.94269981 -0.652900844  -0.008966977  -0.64823831 -0.01719707
84300903  -0.41026561   0.016665095  -0.482994760   0.32482472   0.19075064
84348301  -0.93245070 -0.486988399   0.168699395   0.05132509   0.48220960
84358402   0.38760691 -0.538706543  -0.310046684  -0.15247165   0.13302526
843786    -0.02625135   0.003133944  -0.178447576  -0.01270566   0.19671335
              PC18         PC19          PC20           PC21          PC22
842302    -0.54907956   0.1336499    0.34526111    0.096430045 -0.06878939
842517     0.31801756  -0.2473470   -0.11403274   -0.077259494   0.09449530
84300903  -0.08789759  -0.3922812   -0.20435242    0.310793246   0.06025601
84348301  -0.03584323  -0.0267241   -0.46432511    0.433811661   0.20308706
84358402  -0.01869779   0.4610302    0.06543782   -0.116442469   0.01763433
843786    -0.29727706  -0.1297265   -0.07117453   -0.002400178   0.10108043
              PC23         PC24          PC25          PC26          PC27
842302     0.08444429   0.175102213   0.150887294  -0.201326305 -0.25236294
842517    -0.21752666  -0.011280193   0.170360355  -0.041092627   0.18111081
84300903  -0.07422581  -0.102671419  -0.171007656   0.004731249   0.04952586
84348301  -0.12399554  -0.153294780  -0.077427574  -0.274982822   0.18330078
84358402   0.13933105   0.005327110  -0.003059371   0.039219780   0.03213957
843786     0.03344819  -0.002837749  -0.122282765  -0.030272333 -0.08438081
              PC28         PC29          PC30
842302    -0.0338846387   0.045607590   0.0471277407
842517     0.0325955021  -0.005682424   0.0018662342
84300903   0.0469844833   0.003143131  -0.0007498749
84348301   0.0424469831  -0.069233868   0.0199198881
84358402  -0.0347556386   0.005033481  -0.0211951203
843786     0.0007296587  -0.019703996  -0.0034564331
```

```
#plot(wisc.pr$x)
```

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

- .4427 > Q5. How many principal components (PCs) are required to describe at least 70% of the original variance in the data?

- 3

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

- 7

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

- There are lots of data points on this plot and is very messy and difficult to understand. This is because all the information we don't need is also included into this plot. There are other variables within the excel that are recorded and do not have correlation and these points make it hard to read.

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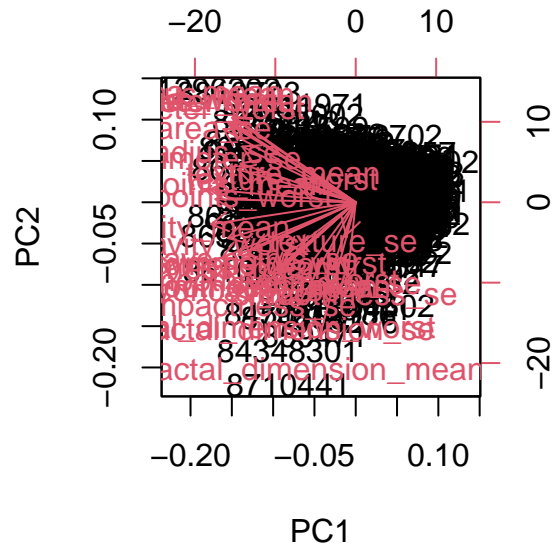


```
# Scatter plot observations by components 1 and 2
plot(wisc.pr$x[,1], wisc.pr$x[,2], col = as.factor(diagnosis) ,
     xlab = "PC1", ylab = "PC2")
```
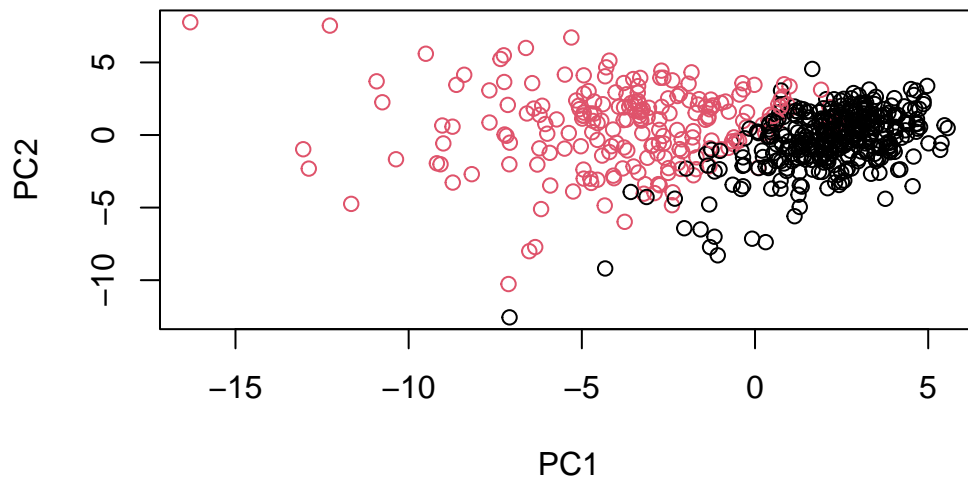
Q8. Generate a similar plot for principal components 1 and 3. What do you notice about these plots?

- These plots are cleaner compared to the original `wisc.pr` plot.

```
# Repeat for components 1 and 3
plot(wisc.pr$x[, 1 ], wisc.pr$x[,3] ,col = as.factor(diagnosis),
     xlab = "PC1", ylab = "PC3")
```
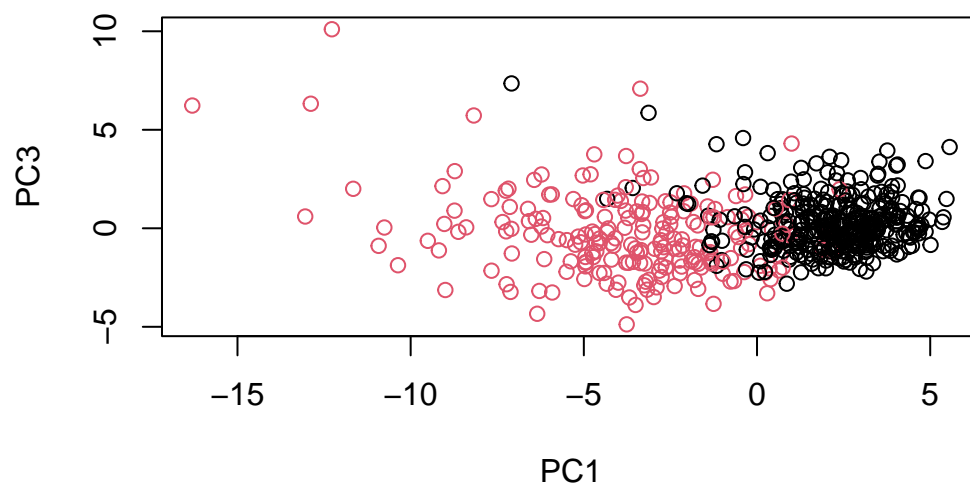
```r
# Create a data.frame for ggplot
df <- as.data.frame(wisc.pr$x)
df$diagnosis <- diagnosis

# Load the ggplot2 package
library(ggplot2)

# Make a scatter plot colored by diagnosis
ggplot(df) +
  aes(PC1, PC2, col=diagnosis) +
  geom_point()
```
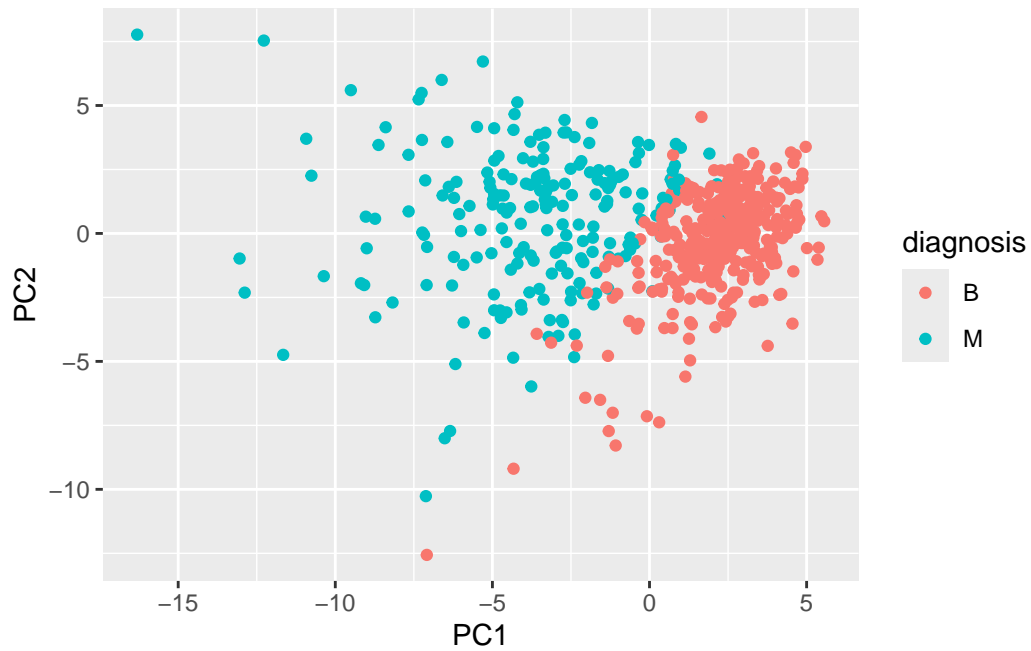
```
# Calculate variance of each component
pr.var <- wisc.pr$sdev^2
head(pr.var)
```

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

Combine PCA and clustering

Our PCA results were in 'wisc.pr$x

```
# distance matric from PCA result
d <- dist(wisc.pr$x[,1:3])
hc <- hclust(d, method="ward.D2")
plot(hc)
```
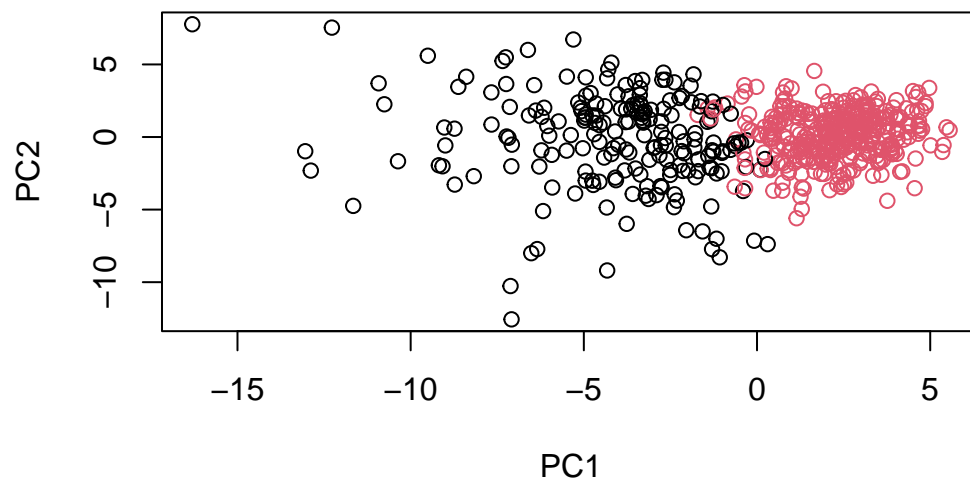
Cluster Dendrogram

Height
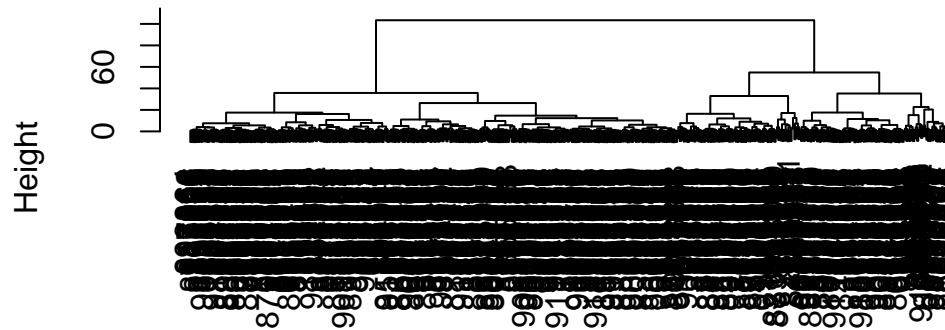
d
hclust (*, "ward.D2")

Cut tree into two groups/branches/clusters...

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

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

Compare my clustering result (my`grps`) to the expert `diagnosis`

```r
table(diagnosis)
```

```
diagnosis
  B   M
357 212
```
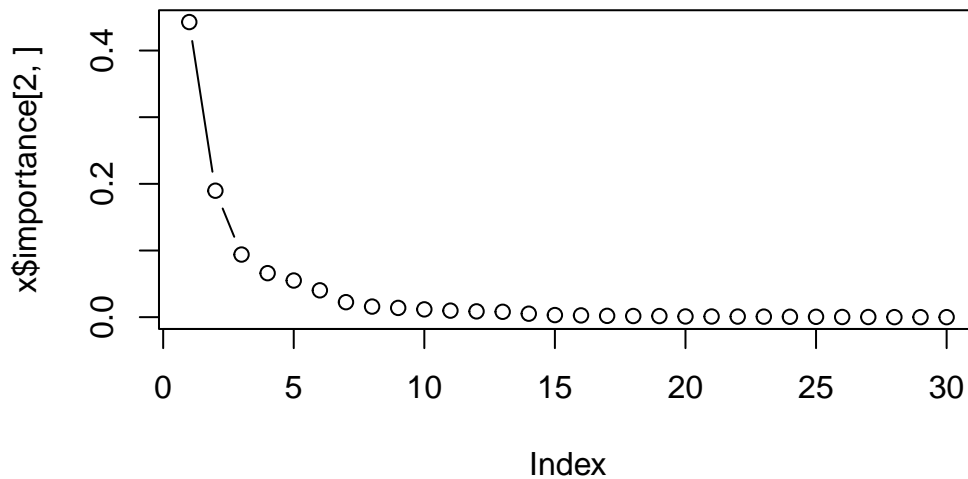
```r
table(grps)
```

```
grps
  1   2
203 366
```

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

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

```
x <- summary(wisc.pr)
  plot(x$importance[2,], typ="b")
```


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?

- -0.26085376

```
wisc.pr$rotation[,1]
```

radius_mean	texture_mean	perimeter_mean
-0.21890244	-0.10372458	-0.22753729
area_mean	smoothness_mean	compactness_mean
-0.22099499	-0.14258969	-0.23928535
concavity_mean	concave.points_mean	symmetry_mean
-0.25840048	-0.26085376	-0.13816696
fractal_dimension_mean	radius_se	texture_se
-0.06436335	-0.20597878	-0.01742803
perimeter_se	area_se	smoothness_se
-0.21132592	-0.20286964	-0.01453145
compactness_se	concavity_se	concave.points_se
-0.17039345	-0.15358979	-0.18341740

15

```
         symmetry_se      fractal_dimension_se               radius_worst
          -0.04249842               -0.10256832                -0.22799663
        texture_worst           perimeter_worst                 area_worst
          -0.10446933               -0.23663968                -0.22487053
      smoothness_worst        compactness_worst            concavity_worst
          -0.12795256               -0.21009588                -0.22876753
   concave.points_worst            symmetry_worst  fractal_dimension_worst
          -0.25088597               -0.12290456                -0.13178394
```

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

- 5 PCs

<pre>
summary(wisc.pr)
</pre>

```
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
```

Heirchal Clustering

```
# Scale the wisc.data data using the "scale()" function
data.scaled <- scale(wisc.data)
```

```
data.dist <- dist(data.scaled, method="euclidean")
head(data.dist)
```

```
[1] 10.309426  6.771675 10.463467  8.663413  8.402233  9.843286
```

```
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?

- The height at which the clustering model has 4 clusters is 19

```
plot(wisc.hclust,main = "Cluster Dendrogam", xlab = "data list")
abline(h=19, col="red", lty=2)
```


Cluster Dendrogam

data list
hclust (*, "complete")
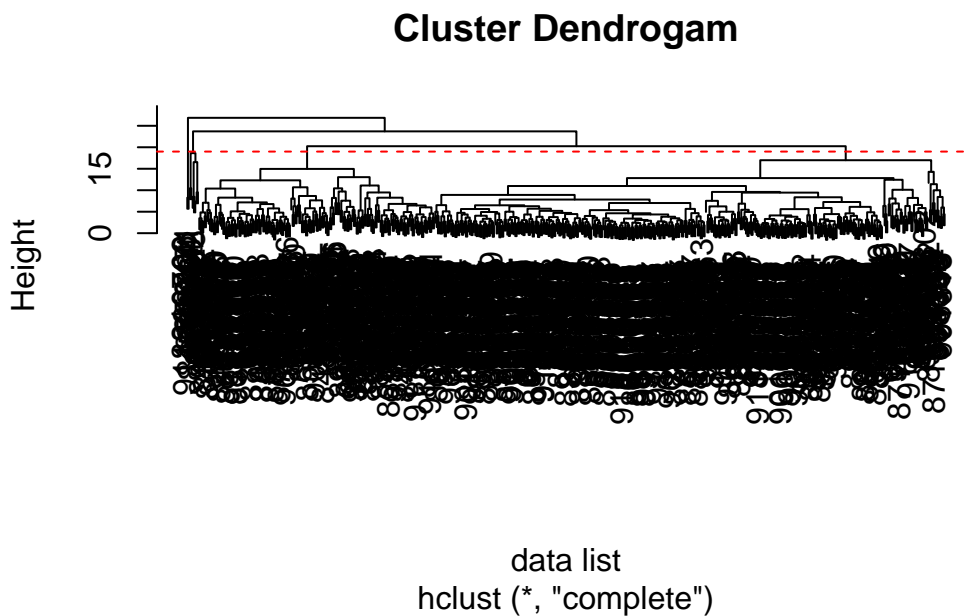
Q12. Can you find a better cluster vs diagnosis match by cutting into a different number of clusters between 2 and 10?

17

- 2 is the best cluster vs diagnosis match for maligant cells because it has the highest count of maligant cells per cluster.

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

```
                      diagnosis
wisc.hclust.clusters   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
```
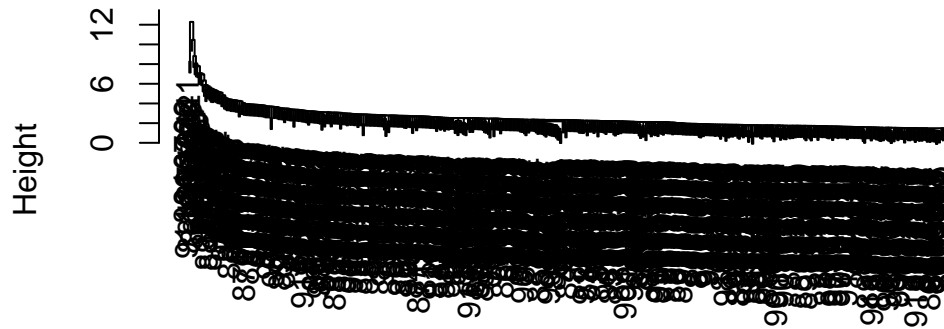
Q13. Which method gives your favorite results for the same data.dist dataset? Explain your reasoning.

```
wisc.hclust1 <- hclust(data.dist, method="complete")
plot(wisc.hclust1)
```


Cluster Dendrogram

data.dist
hclust (*, "complete")
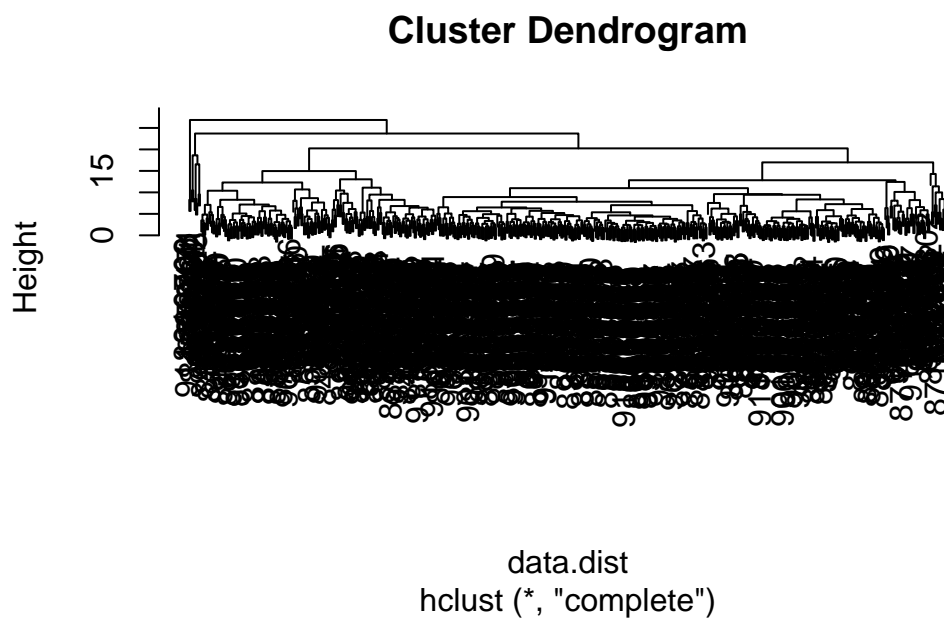
```
wisc.hclust1 <- hclust(data.dist, method="single")
plot(wisc.hclust1)
```

Cluster Dendrogram

data.dist
hclust (*, "single")

```
wisc.hclust1 <- hclust(data.dist, method="average")
plot(wisc.hclust1)
```
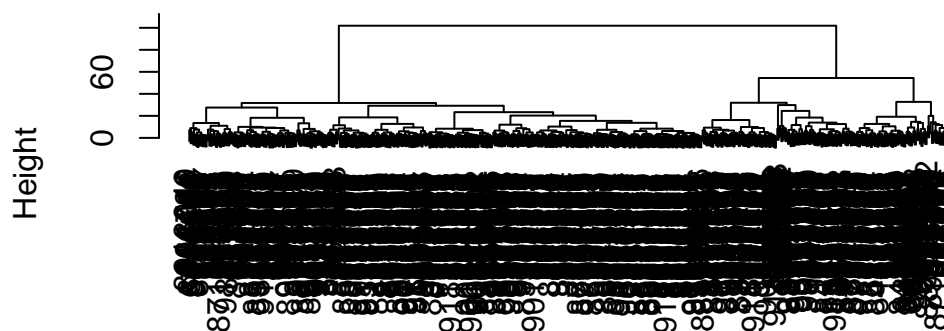
Cluster Dendrogram
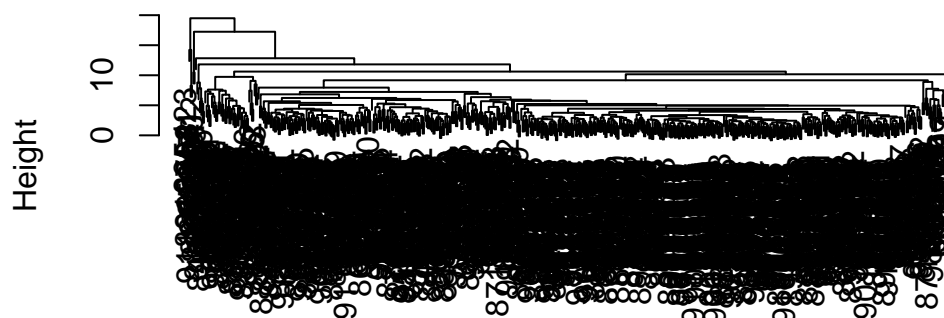
data.dist
hclust (*, "average")

```r
wisc.hclust1 <- hclust(data.dist, method="ward.D2")
plot(wisc.hclust1)
```

Cluster Dendrogram

data.dist
hclust (*, "ward.D2")

Ward.D2 is the best because of the organization pattern which makes it easier to read and the pattern is most centralized.