Class08

Chen

#Mini-Project

```
wisc.df <- read.csv("/Users/showwhale/Desktop/HsiaoinMac/HsiaoinMac/UCSD/Freshman/Winter/B
```

Q1. How many obervations/samples/patients/rows?

There are 'r nrow(wisc.df) individuals in this dataset.

```
nrow(wisc.df)
```

```
[1] 569
```

Q2 How many of the observations have a malignant diagnosis?

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

```
  B   M
357 212
```

```
sum(wisc.df$diagnosis == "M")
```

```
[1] 212
```

```
sum(wisc.df$diagnosis == "B")
```

```
[1] 357
```

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

```
cname <- colnames(wisc.df)
cname
```

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

```
cnum <- grep("_mean", cname, value = TRUE)
length(cnum)
```

```
[1] 10
```

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

```
[1] 31
```

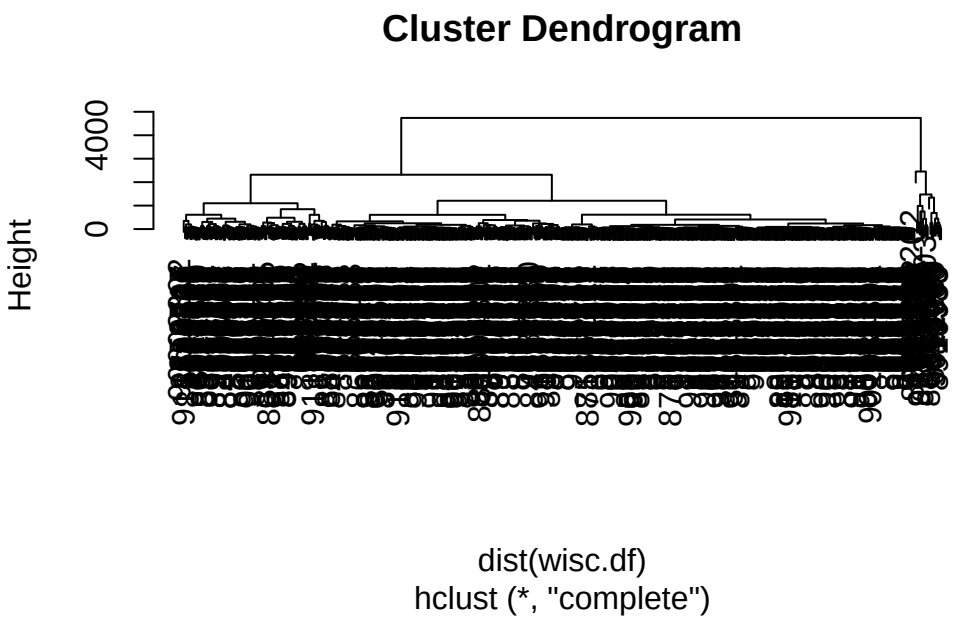
```
diagnosis <- as.factor(wisc.df$diagnosis)
```

and remove or exclude this column from any of our further analysis

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

Let's try clustering this data!

```
wisc.hc <- hclust(dist(wisc.df))
plot(wisc.hc)
```



Back to our cancer data set

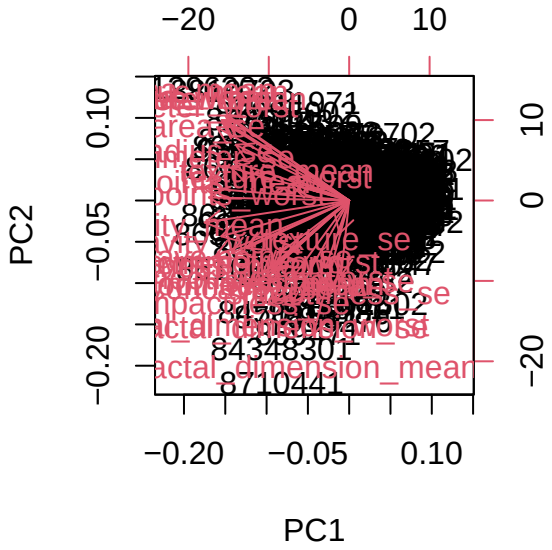
Do we need to scale this data set? Yes. Because the spread is very different in each variables.

```
wisc.pr <- prcomp(wisc.df, 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
```

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



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

```
sumwisc.pr <- summary(wisc.pr)
sumwisc.pr$importance[2,1]
```

```
[1] 0.44272
```

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

```
which(sumwisc.pr$importance[3,]>0.7)[1]
```

```
PC3
  3
```

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

```
which(sumwisc.pr$importance[3,]>0.9)[1]
```

```
PC7
  7
```

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

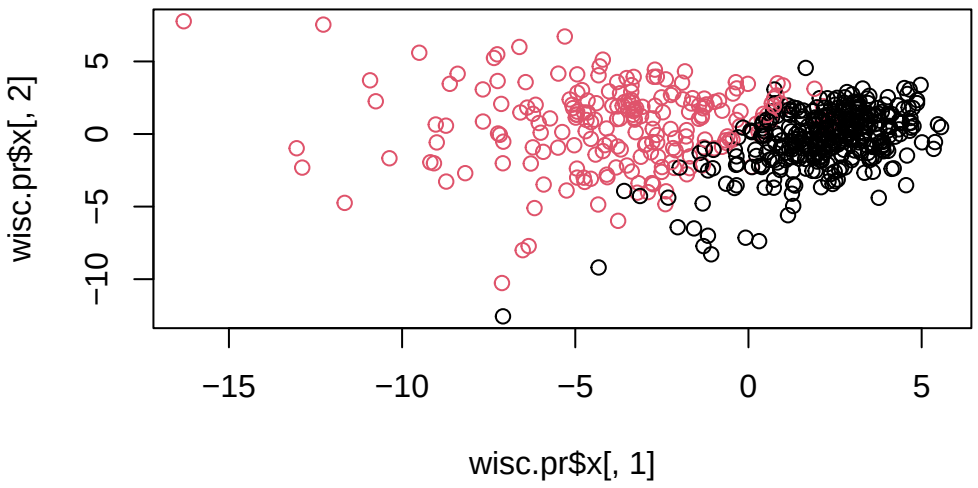
**There is no cluster pattern is shown and is difficult to understand. This is because these data comprises continuous values rather than distinct clusters.

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

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

$class
[1] "prcomp"
```

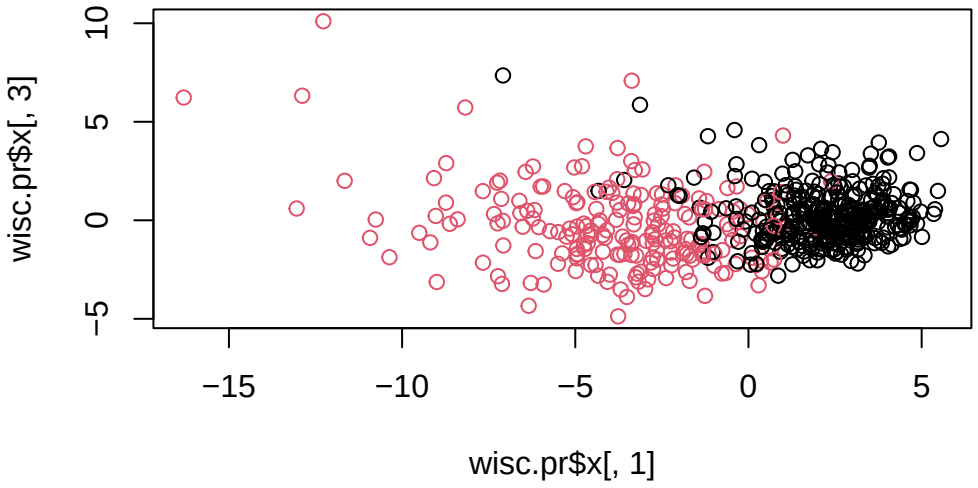
We need to build our own plot because wisc.pr is too crowded and non informative.

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

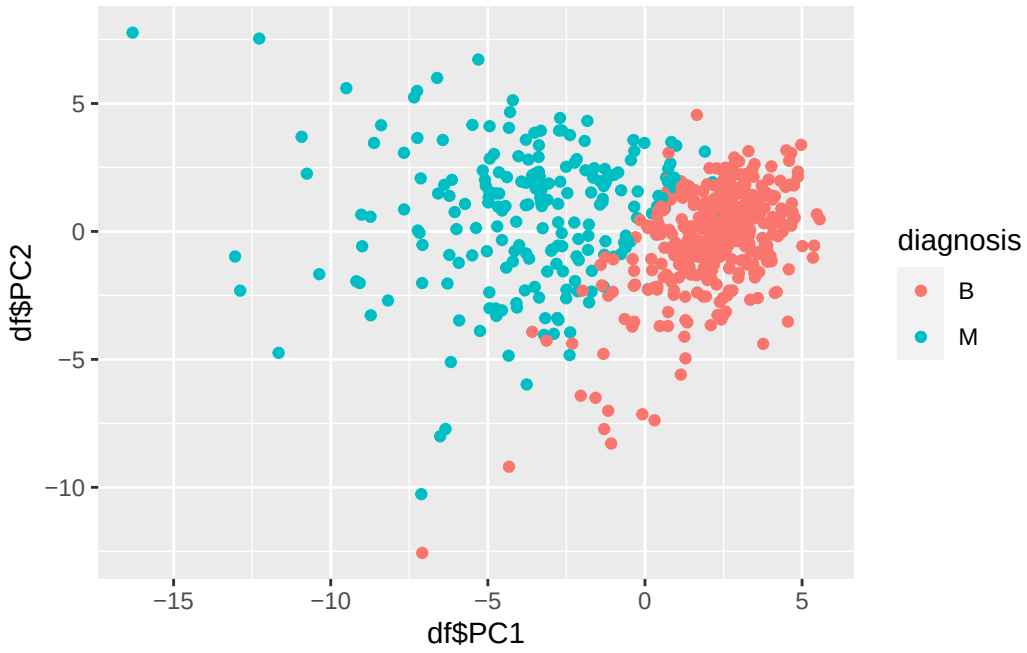


Q8. Generate a similar plot for principal components 1 and 3. What do you notice about these plots?

```
plot(wisc.pr$x[,1], wisc.pr$x[,3], col = diagnosis)
```



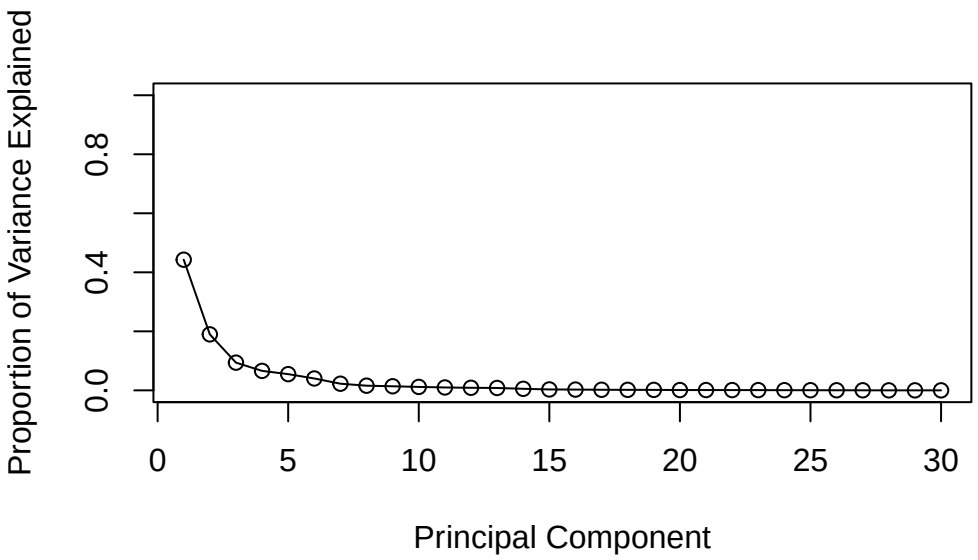

```
df <- as.data.frame(wisc.pr$x)

library(ggplot2)

ggplot(df, aes(x = df$PC1, y = df$PC2, color = diagnosis))+
  geom_point()
```




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

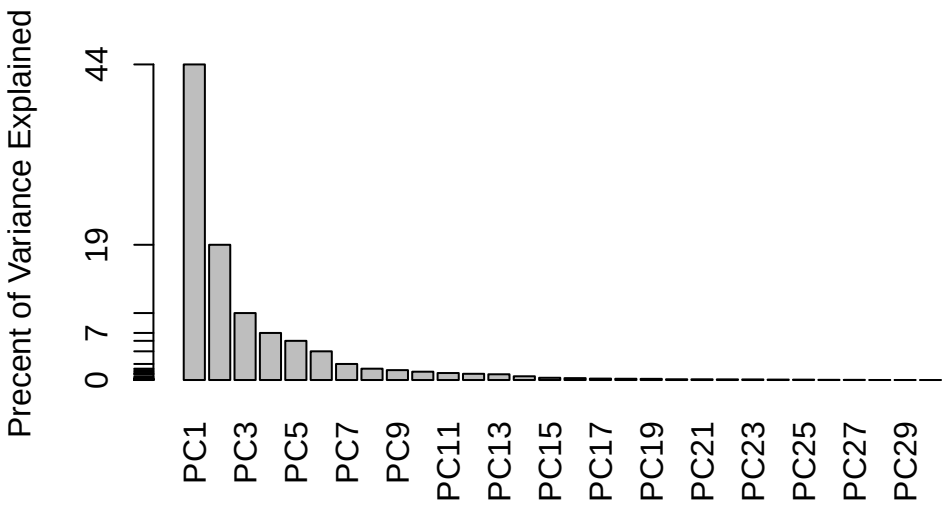
```
[1] 13.281608  5.691355  2.817949  1.980640  1.648731  1.207357
```

```
pve <- pr.var/sum(pr.var)
plot(pve, xlab = "Principal Component",
     ylab = "Proportion of Variance Explained",
     ylim = c(0, 1), type = "o")
```




```
#barplot(PCall2[2,], ylab = "Precent of Variance Explained",
     #names.arg=paste0("PC",1:length(PCall2[2,])), las=2, axes = FALSE)
```

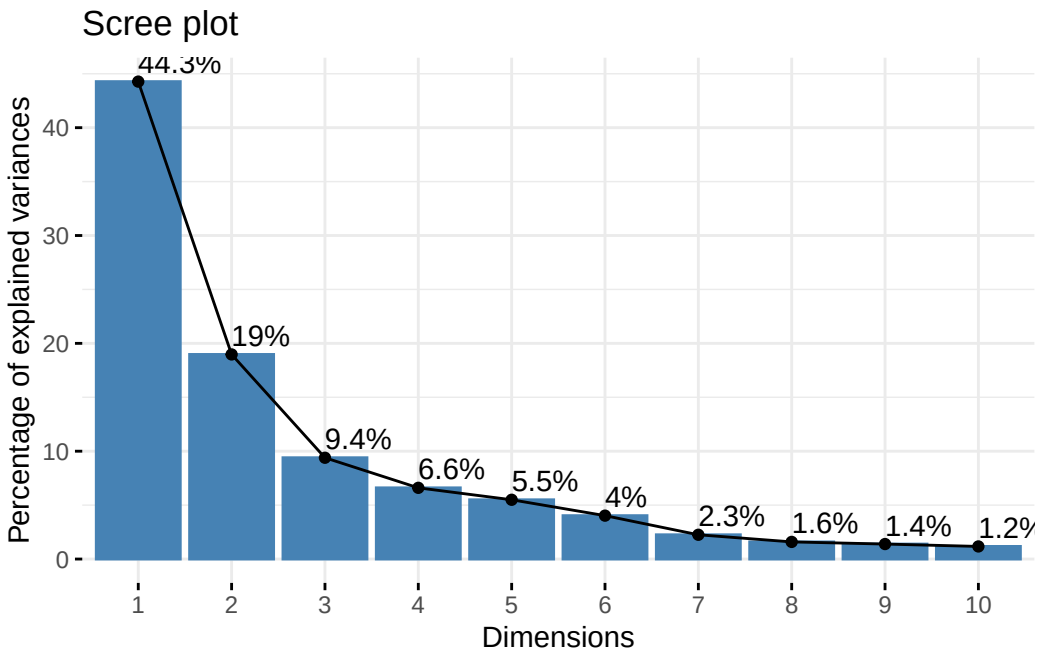
```
barplot(pve, ylab = "Precent of Variance Explained",
     names.arg=paste0("PC",1:length(pve)), las=2, axes = FALSE)
axis(2, at=pve, labels=round(pve,2)*100 )
```




```
library(factoextra)
```

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

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



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? This tells us how much this original feature contributes to the first PC.

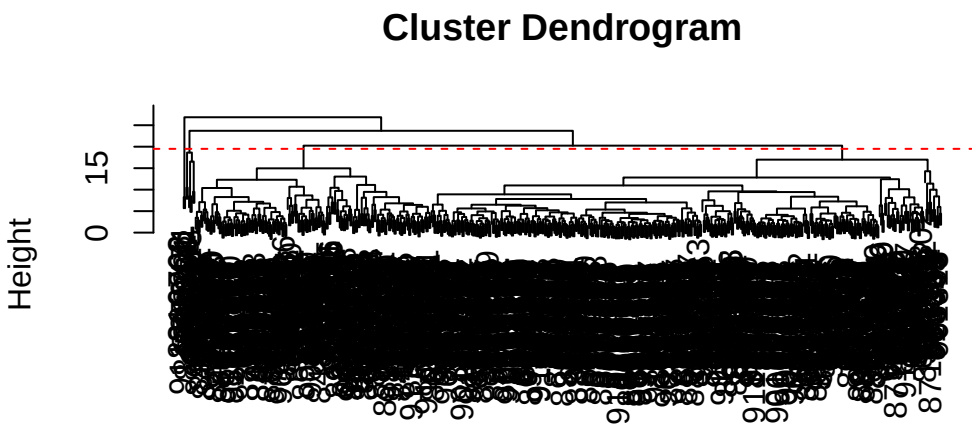
```
wisc.pr$rotation[,1]["concave.points_mean"]
```

```
concave.points_mean
         -0.2608538
```

#Hierarchical clustering

Q10. Using the plot() and abline() functions, what is the height at which the clustering model has 4 clusters?

```
data.scaled <-scale(wisc.df)
data.dist <- dist(data.scaled)
wisc.hclust <- hclust(data.dist, method = "complete")
plot(wisc.hclust)
abline(wisc.hclust, col="red", lty=2, h=19.5)
```




```
wisc.hclust.clusters <- cutree(wisc.hclust, 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 OPTIONAL: Can you find a better cluster vs diagnoses match by cutting into a different number of clusters between 2 and 10? How do you judge the quality of your result in each case?

**9 or 10 clusters will be better clustering because they have minimum false malign (39) and false benign (12)

```
wisc.hclust.clusters <- cutree(wisc.hclust, 4)
table(wisc.hclust.clusters, diagnosis)
```

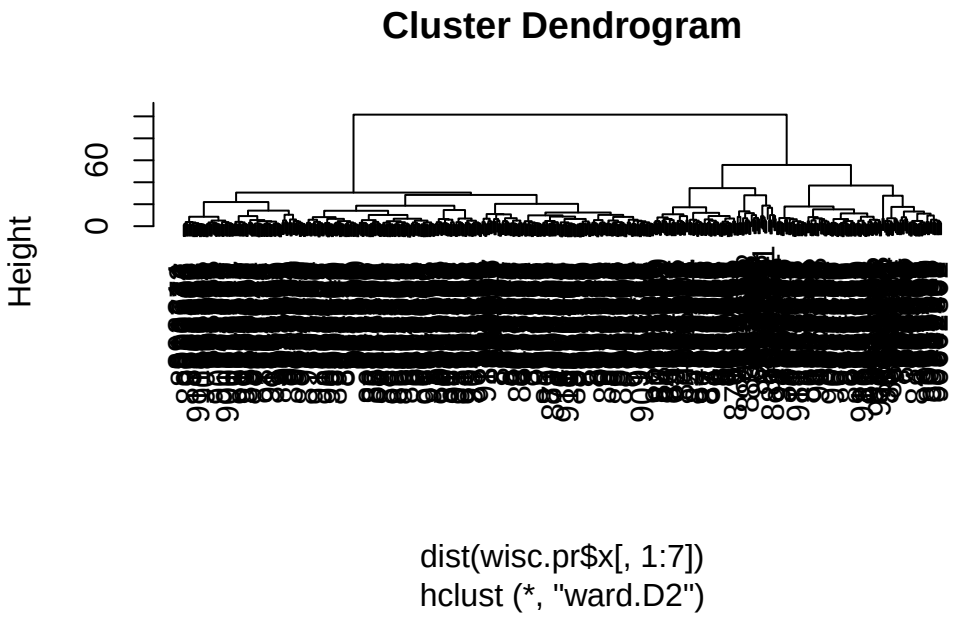
```
                    diagnosis
```

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

Q12. Which method gives your favorite results for the same data.dist dataset? Explain your reasoning.

**Average method calculates the distance between clusters based on the average distance of all pairs of points and sounds more intuitive to me.

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



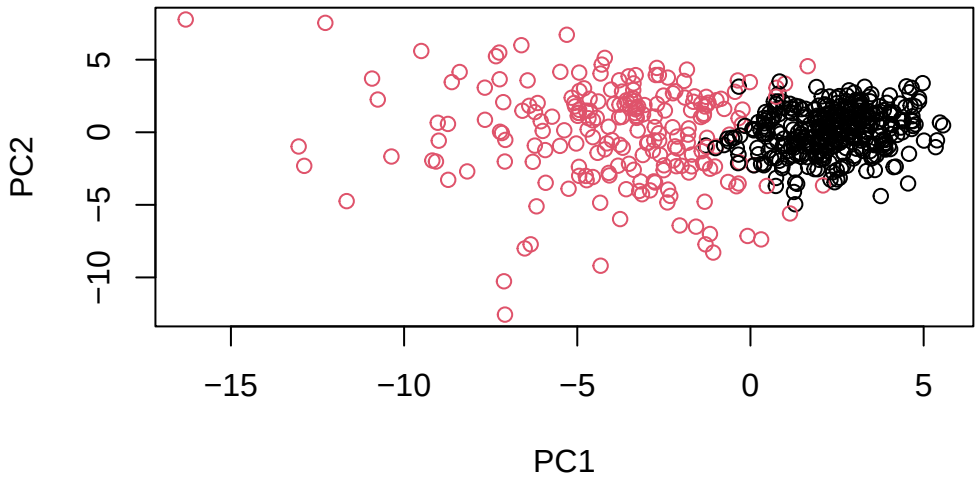

```
grps <- cutree(wisc.pr.hclust, k=2)
table(grps)
```

```
grps
  1   2
216 353
```

```
g <- as.factor(grps)
g <- relevel(g,2)
levels(g)
```

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

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



Q13. How well does the newly created model with four clusters separate out the two diagnoses?

```
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
```

Q14. How well do the 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.

**wisc.pr.hclust.clusters shows less false malign + false benign

```
wisc.hclust.clusters <- cutree(wisc.hclust, 4)
table(wisc.hclust.clusters, diagnosis)
```

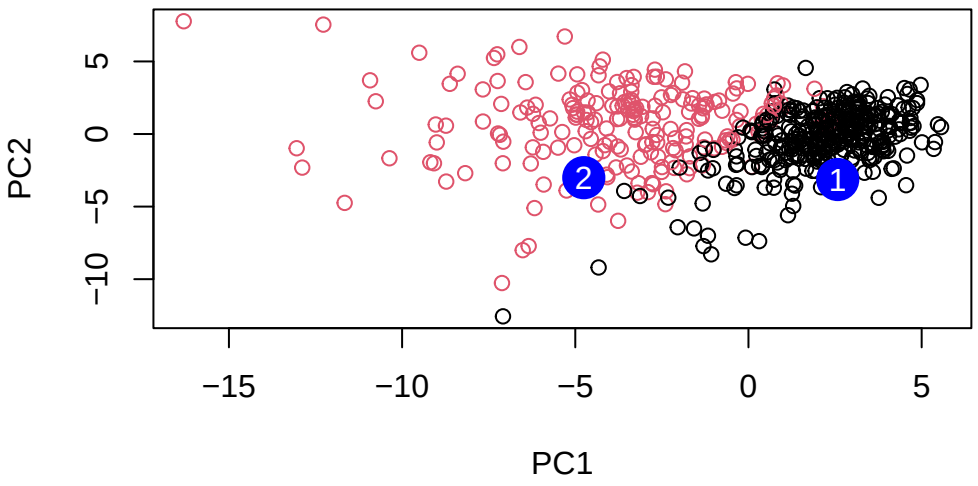
```
                    diagnosis
wisc.hclust.clusters   B   M
                   1  12 165
                   2   2   5
                   3 343  40
                   4   0   2
```

Prediction

Q16. Which of these new patients should we prioritize for follow up based on your results?

**Patient 2 should be proritized.

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

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