Class07

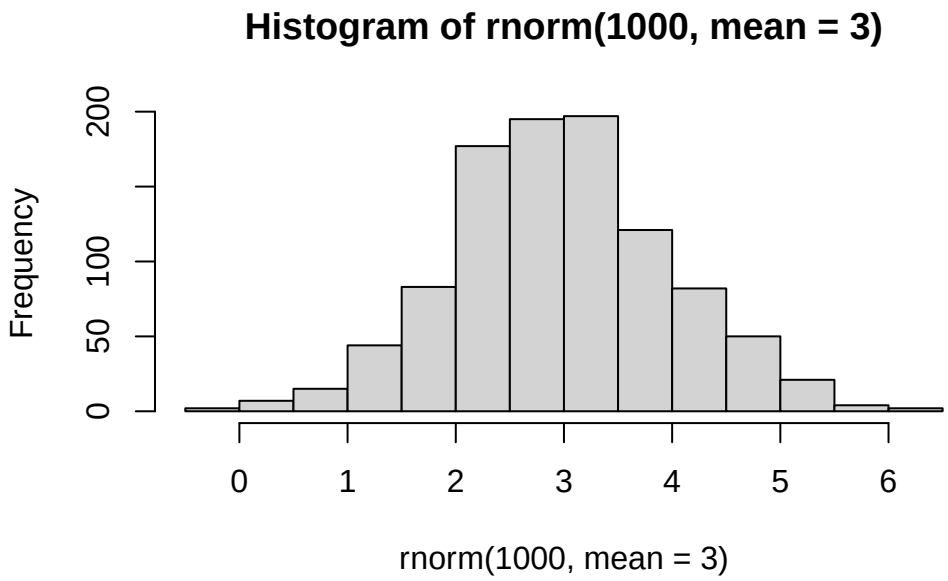
Nicholas Thiphakhinkeo

rnorm

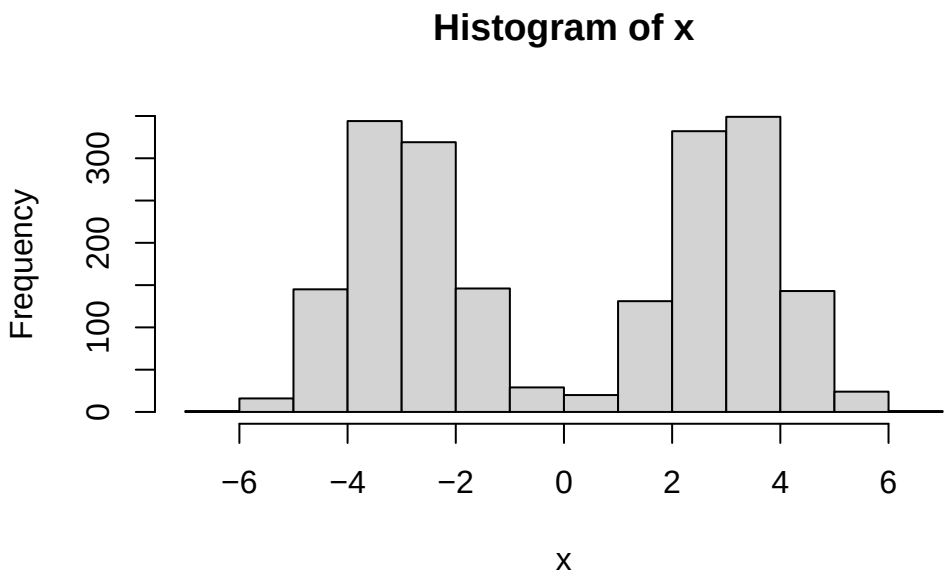
```
rnorm(10)
```

```
 [1] -0.8406972  0.7752770 -1.1336286  1.5898798  0.3645093 -2.2865040
 [7] -2.4296390  1.0881165 -0.3892030  0.3098780
```

```
hist(rnorm(1000, mean=3))
```



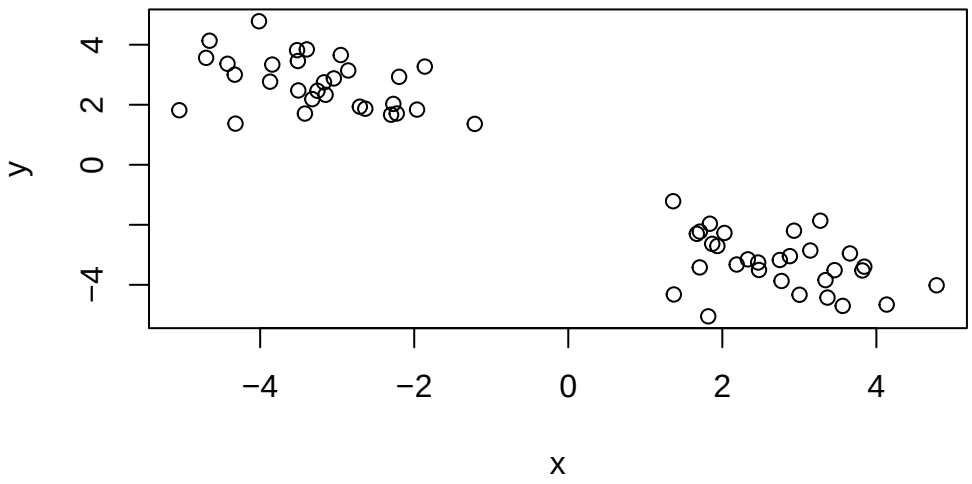
```r
n <- 1000
x <- c(rnorm(n,-3),rnorm(n,+3))
hist(x)
```




```r
n<-30
x<-c(rnorm(n,-3), rnorm(n,+3))
y <- rev(x)
z <- cbind(x,y)
head(z)
```

```
             x        y
[1,] -3.256146 2.464787
[2,] -1.963510 1.838160
[3,] -1.214137 1.361263
[4,] -2.300530 1.668796
[5,] -2.226912 1.707903
[6,] -3.043440 2.876707
```

```r
plot(z)
```




```
km <- kmeans(z,centers=2)
km
```

```
K-means clustering with 2 clusters of sizes 30, 30

Cluster means:
          x         y
1  2.715122 -3.256493
2 -3.256493  2.715122

Clustering vector:
 [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1
[39] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 48.39342 48.39342
 (between_SS / total_SS =  91.7 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

```
km$size
```

```
[1] 30 30
```

Cluster Assignment

```
km$cluster
```

```
 [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1
[39] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Cluster Center

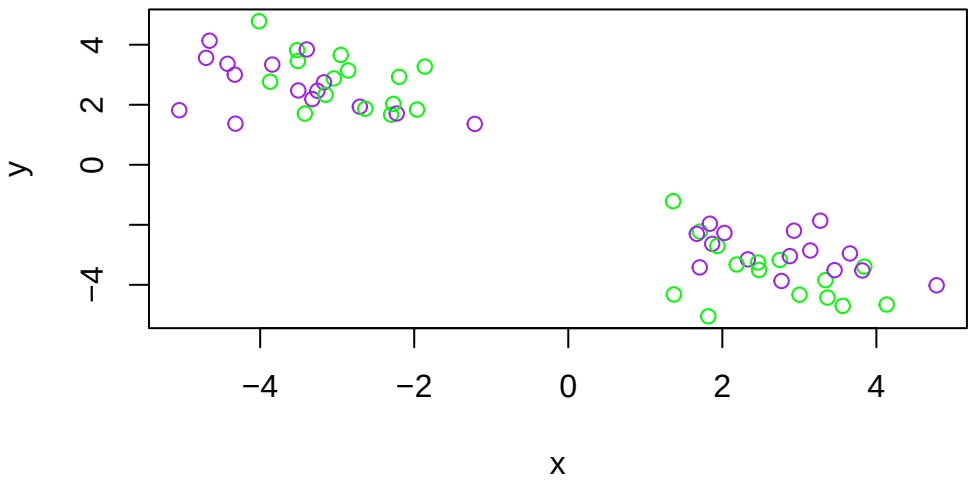
```
km$centers
```

```
          x         y
1  2.715122 -3.256493
2 -3.256493  2.715122
```

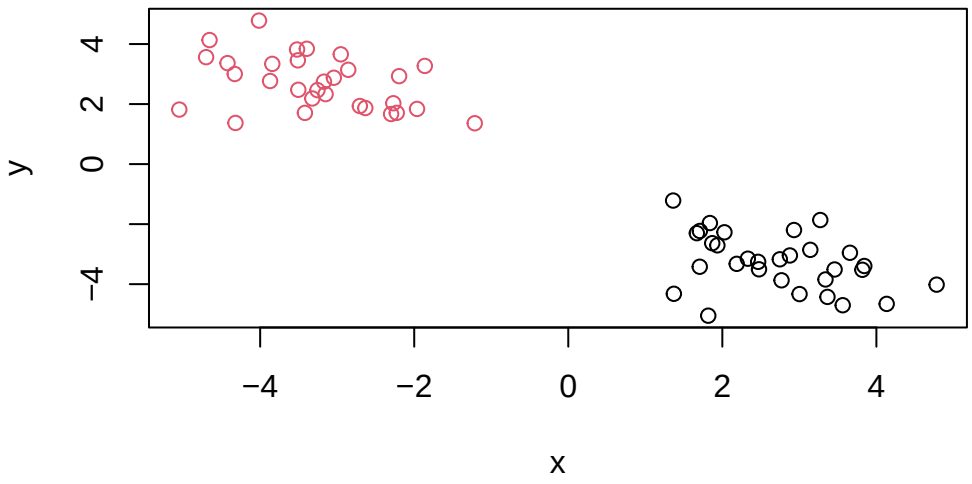
Plot z color colored by kmeans cluster assignment and add cluster centers as blue points

R recycles shorter color vector to be the same length as the longer (number of data points) in z

```
plot(z, col=c("purple","green"))
```

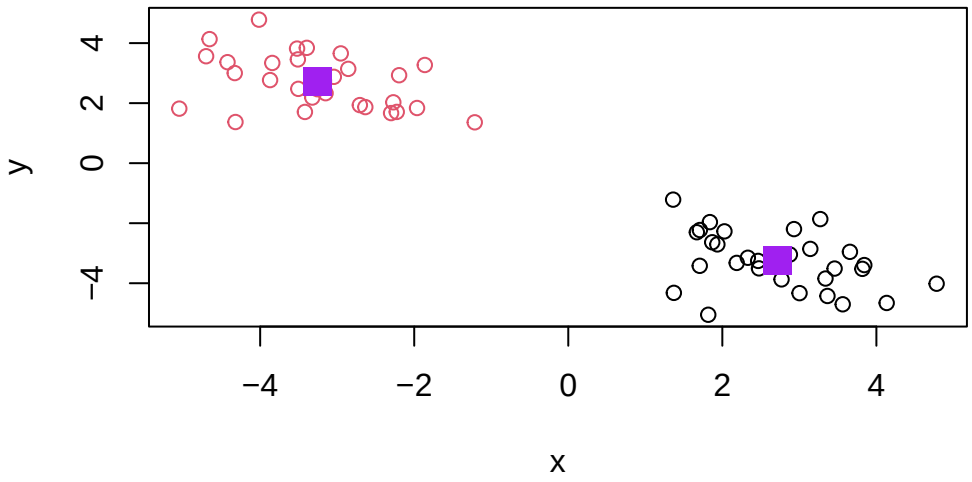



```
plot(z, col=km$cluster)
```



use points() function to add new points to an existing plot... like a cluster plot

```
plot(z, col=km$cluster)
points(km$centers, col="purple", pch=15, cex=2)
```



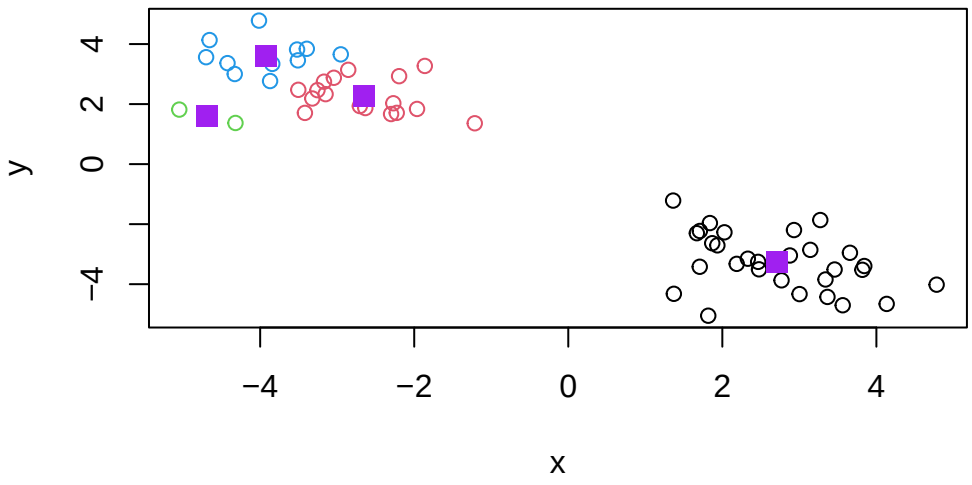
Q. Run kmeans and ask for 4 clusters and plot results

```
km2 <- kmeans(z, centers=4)
km2
```

```
K-means clustering with 4 clusters of sizes 30, 17, 2, 11

Cluster means:
          x         y
1  2.715122 -3.256493
2 -2.653050  2.267015
3 -4.685379  1.593931
```

```
4 -3.929289  3.611503

Clustering vector:
 [1] 2 2 2 2 2 2 2 4 2 4 4 4 3 2 4 4 4 2 2 2 4 2 4 4 3 2 4 2 2 2 1 1 1 1 1 1 1 1
[39] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Within cluster sum of squares by cluster:
[1] 48.3934250 11.8945093  0.3656591  6.1139684
 (between_SS / total_SS =  94.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

```r
plot(z, col=km2$cluster)
points(km2$centers, col="purple", pch=15, cex=1.5)
```



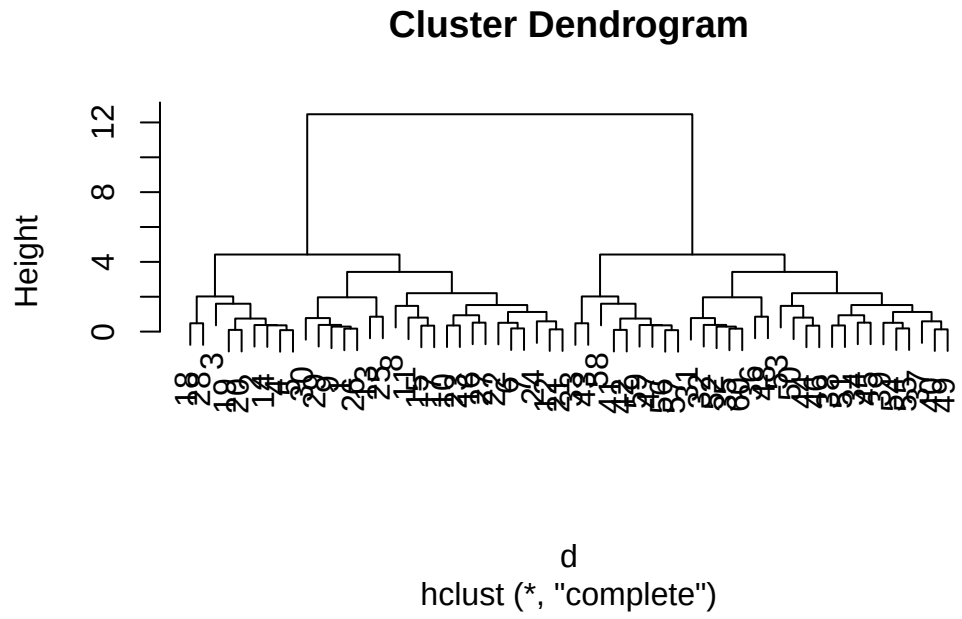
Hierarchical Clustering

#need distance matrix of data to be clustered

```
d <- dist(z)
hc<- hclust(d)
hc
```

```
Call:
hclust(d = d)

Cluster method   : complete
Distance         : euclidean
Number of objects: 60
```

```
plot(hc)
```



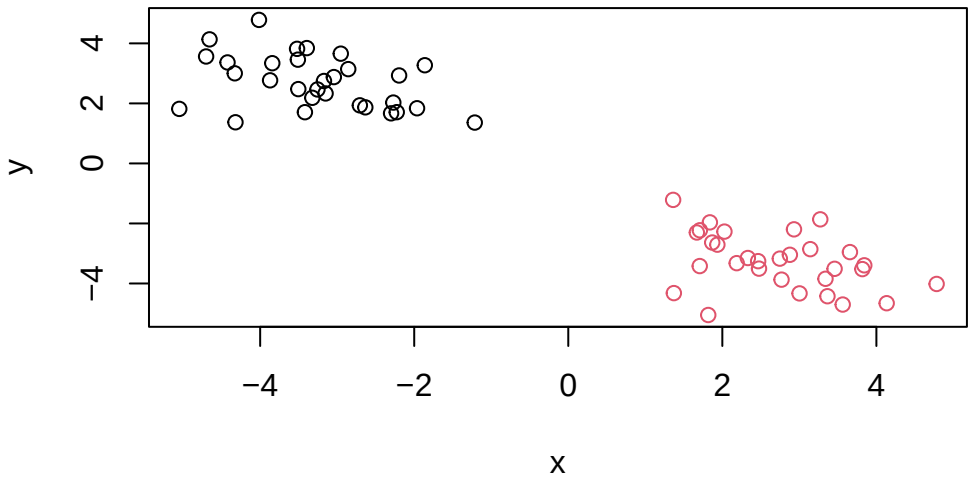
#Cluster Membership by cutting tree with cutree()

```
grps <- cutree(hc,h=8)
grps
```

```
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2
[39] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

Plot "z" colored by hclust

```
plot(z, col=grps)
```

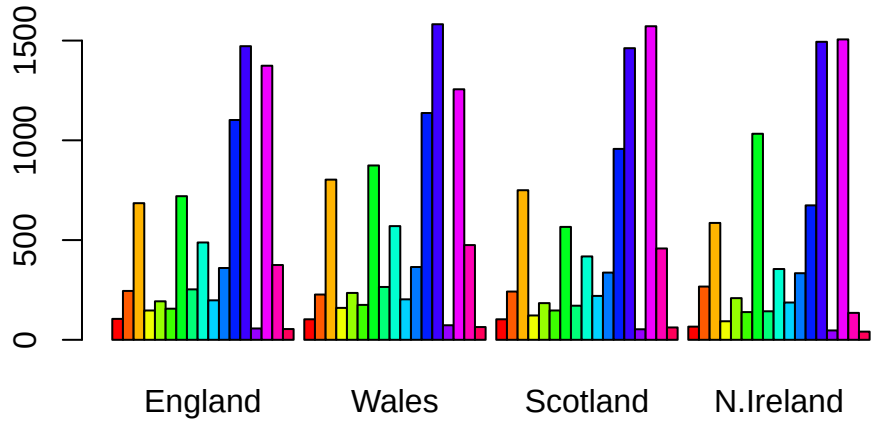


PCA Uk Foods Data

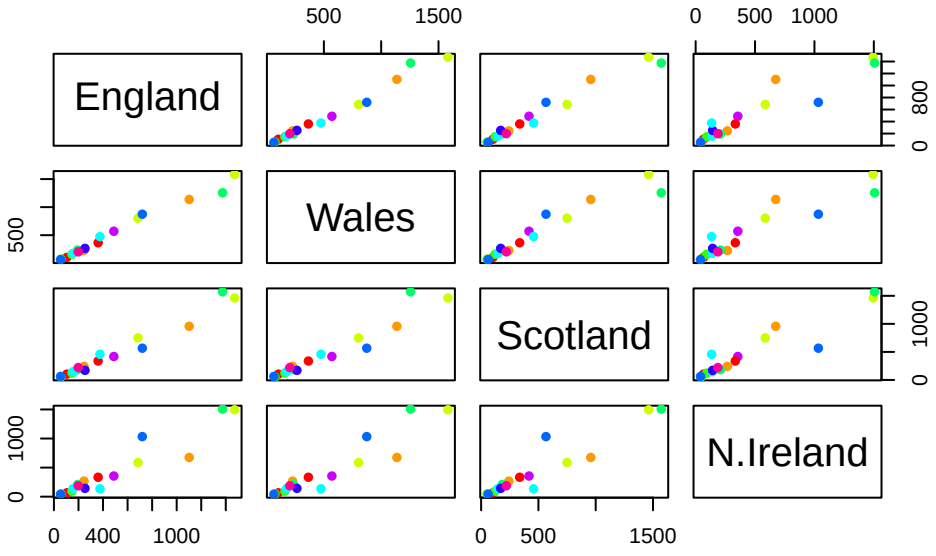
```
url <- "https://tinyurl.com/UK-foods"
x <- read.csv(url, row.names=1)
head(x)
```

```
              England Wales Scotland N.Ireland
Cheese            105   103      103        66
Carcass_meat      245   227      242       267
Other_meat        685   803      750       586
Fish              147   160      122        93
Fats_and_oils     193   235      184       209
Sugars            156   175      147       139
```

```
barplot(as.matrix(x), beside=T, col=rainbow(nrow(x)))
```




```
pairs(x, col=rainbow(10), pch=16)
```



Using PCA for Larger Data Sets

```
pca <- prcomp(t(x))
summary(pca)
```

```
Importance of components:
                            PC1      PC2      PC3       PC4
Standard deviation     324.1502 212.7478 73.87622 3.176e-14
Proportion of Variance   0.6744   0.2905  0.03503 0.000e+00
Cumulative Proportion    0.6744   0.9650  1.00000 1.000e+00
```

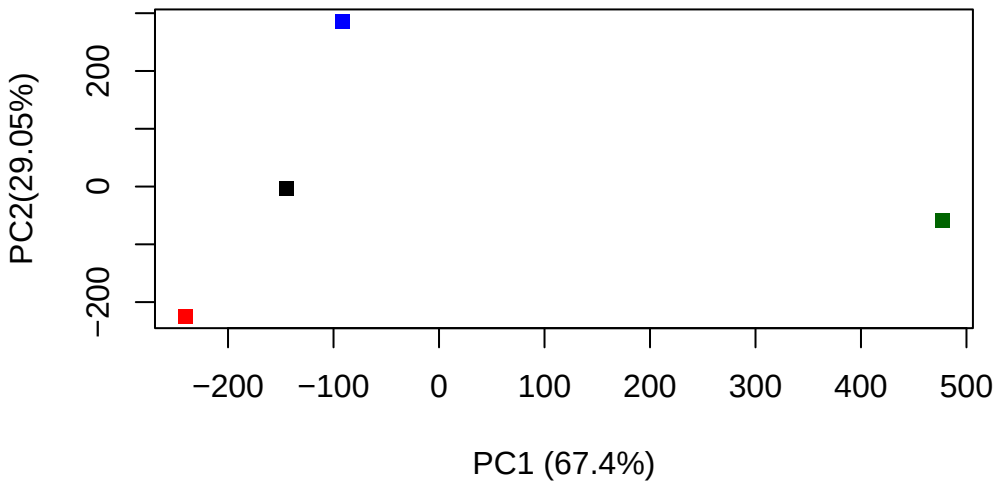
```
attributes(pca)
```

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

$class
[1] "prcomp"
```

```
pca$x
```

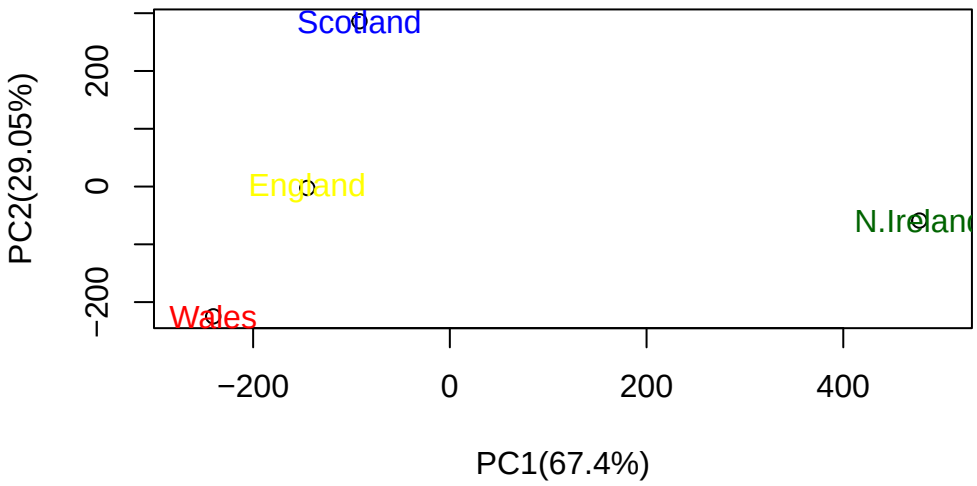
```
                PC1         PC2        PC3           PC4
England   -144.99315   -2.532999 105.768945 -4.894696e-14
Wales     -240.52915 -224.646925 -56.475555  5.700024e-13
Scotland   -91.86934  286.081786 -44.415495 -7.460785e-13
N.Ireland  477.39164  -58.901862  -4.877895  2.321303e-13
```

```
plot(pca$x[,1],pca$x[,2], col=c("black","red","blue","darkgreen"), pch=15, xlab="PC1 (67.4%)"
```



Colored Country Plot

```
plot(pca$x[,1], pca$x[,2], xlab="PC1(67.4%)", ylab="PC2(29.05%)", xlim=c(-270,500))
text(pca$x[,1], pca$x[,2], colnames(x),col=c("yellow","red","blue","darkgreen"),pch=0, cex=1)
```


Scotland
England
N.Ireland
Wales
PC2(29.05%)
PC1(67.4%)
−200
0
200
400
−200
0
200