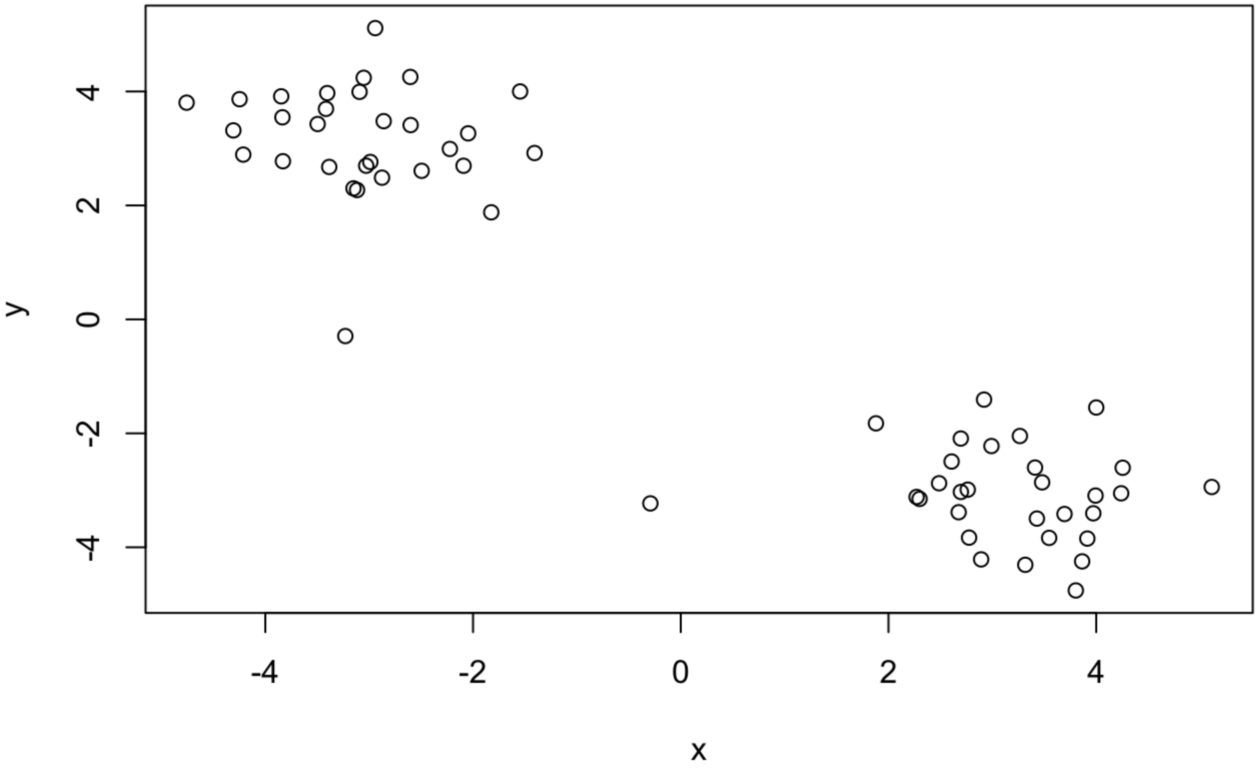
class07

AUTHOR
Ziheng Huang

Test data:

```
tmp <- c(rnorm(30,-3),rnorm(30,3))
x <- cbind(x=tmp,y=rev(tmp))
plot(x)
```



K-means Clustering

kmeans()

```
km <- kmeans(x,centers=2,nstart=20)
km
```

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

Cluster means:
          x         y
1  3.165462 -3.063958
2 -3.063958  3.165462
```

```
2 -3.063958  3.165462

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 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] 47.02053 47.02053
 (between_SS / total_SS =  92.5 %)

Available components:

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

Q: points in cluster

```
km$size
```

```
[1] 30 30
```

Q: cluster assignment, center

```
km$cluster
```

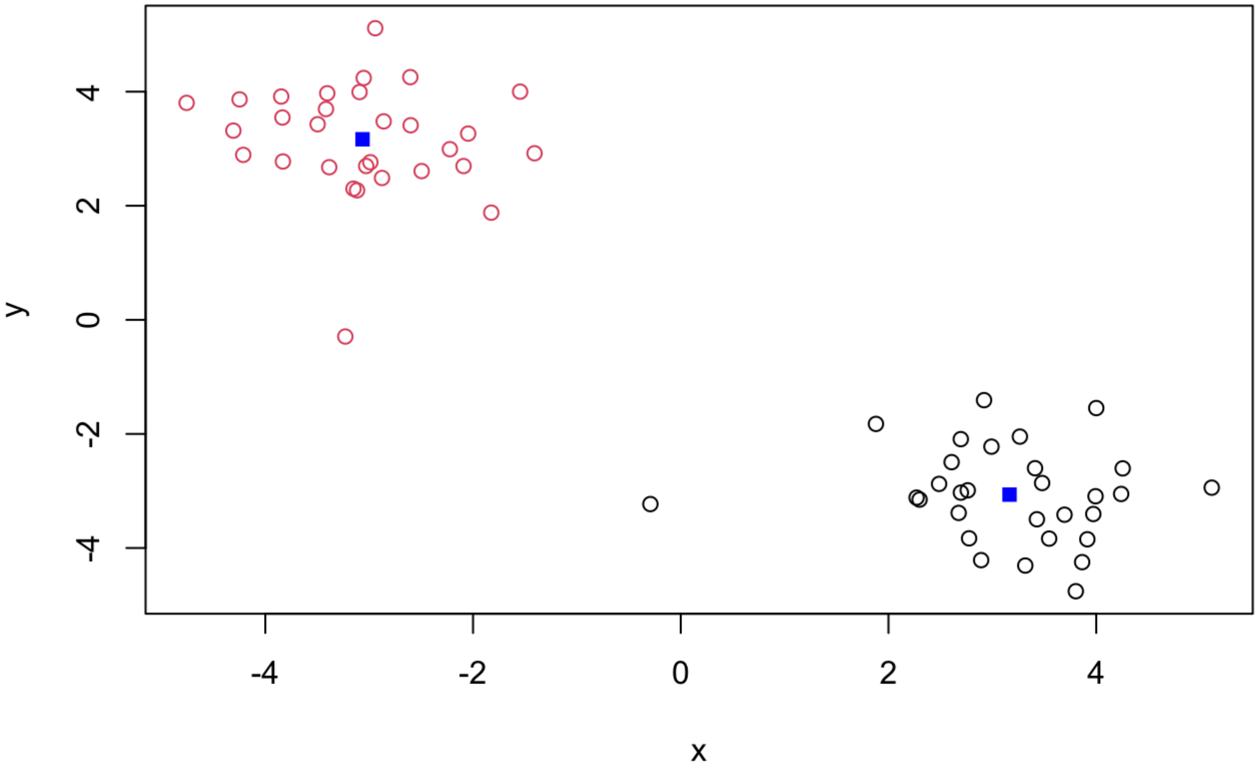
```
 [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1
[39] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
km$centers
```

```
          x         y
1  3.165462 -3.063958
2 -3.063958  3.165462
```

Q: plot clusters, centers

```
plot(x,col=km$cluster)
points(km$centers,col='blue',pch=15)
```

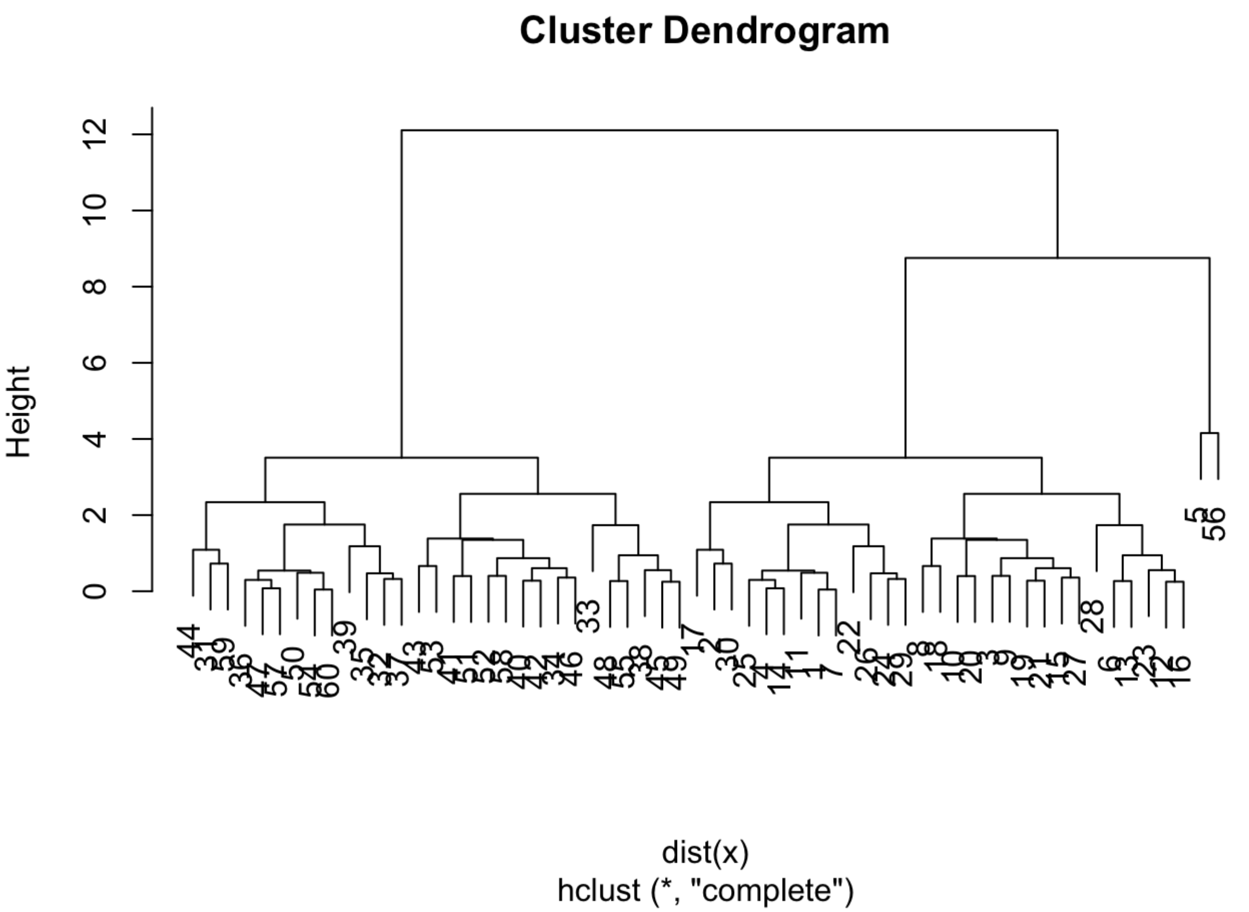


hclust()

```
hc <- hclust(dist(x))
```

plot() for hc

```
plot(hc)
```



get cluster groupings for hc, cut the tree with height

```
cutree(hc,h=8)
```

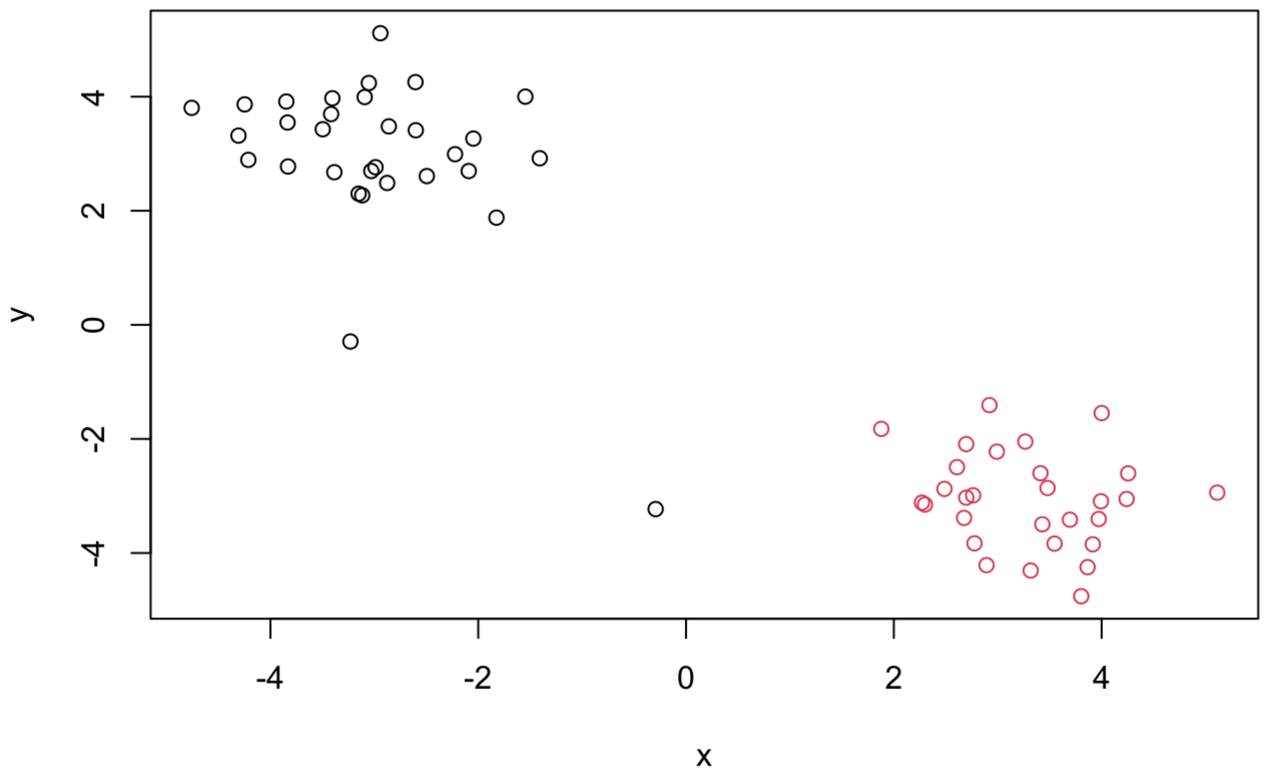
```
 [1] 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3
[39] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3
```

use cutree with k=2

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

plot

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



PCA

load data

```
url <- "https://tinyurl.com/UK-foods"
x <- read.csv(url)
```

Q1: rows/cols

```
dim(x)
```

```
[1] 17  5
```

fixed row/col num

```
x <- read.csv(url, row.names=1)
head(x)
```

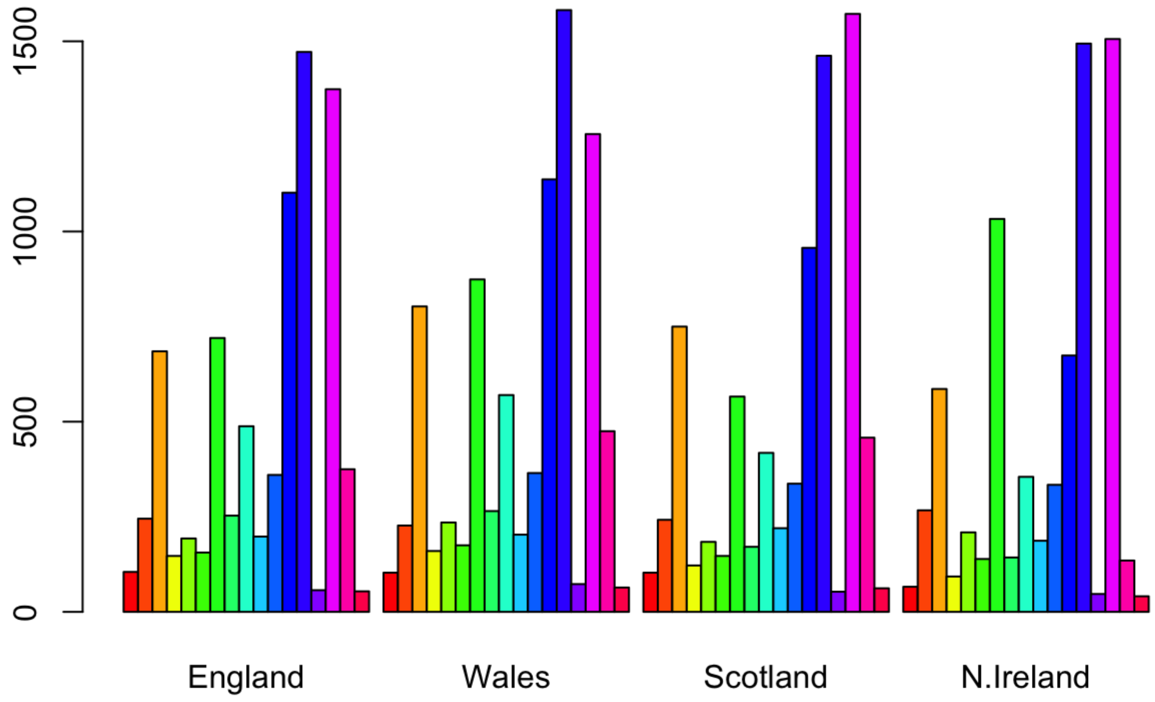
```
               England Wales Scotland N.Ireland
Cheese             105   103      103        66
Carcass_meat       245   227      242       267
Other_meat         685   803      750       586
Fish               147   160      122        93
```

```
Fats_and_oils       193   235       184        209
Sugars              156   175       147        139
```

Q2: Which approach to solving the 'row-names problem' mentioned above do you prefer and why? Is one approach more robust than another under certain circumstances?

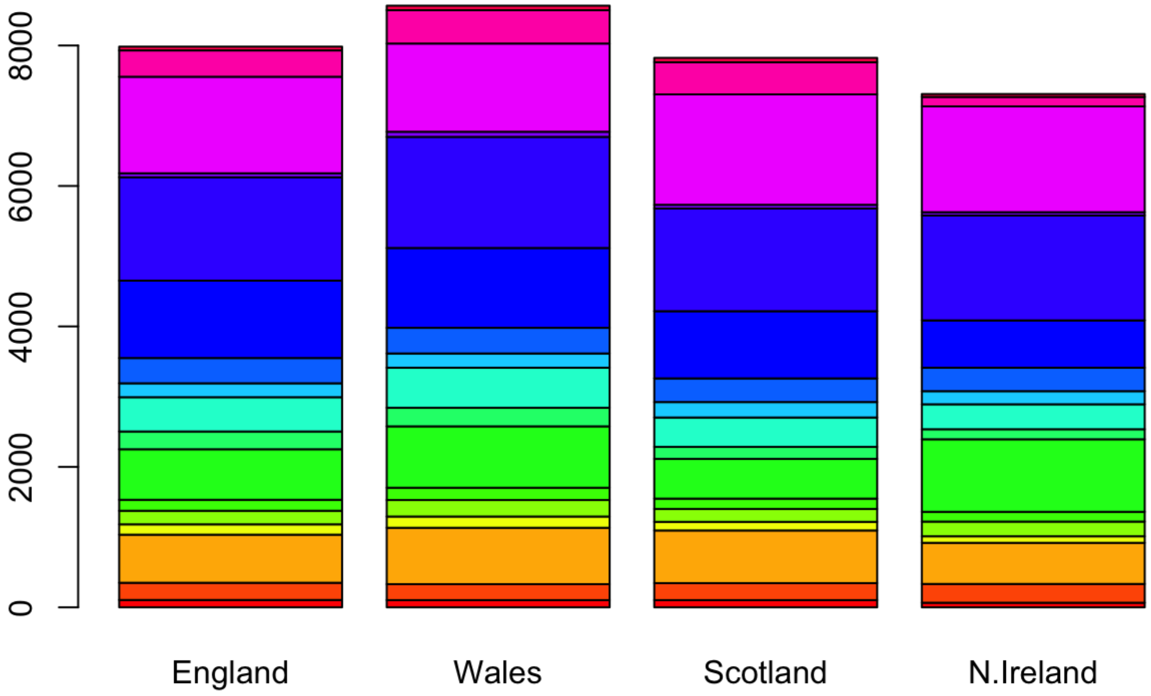
I like the second approach. first approach is a hack.

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



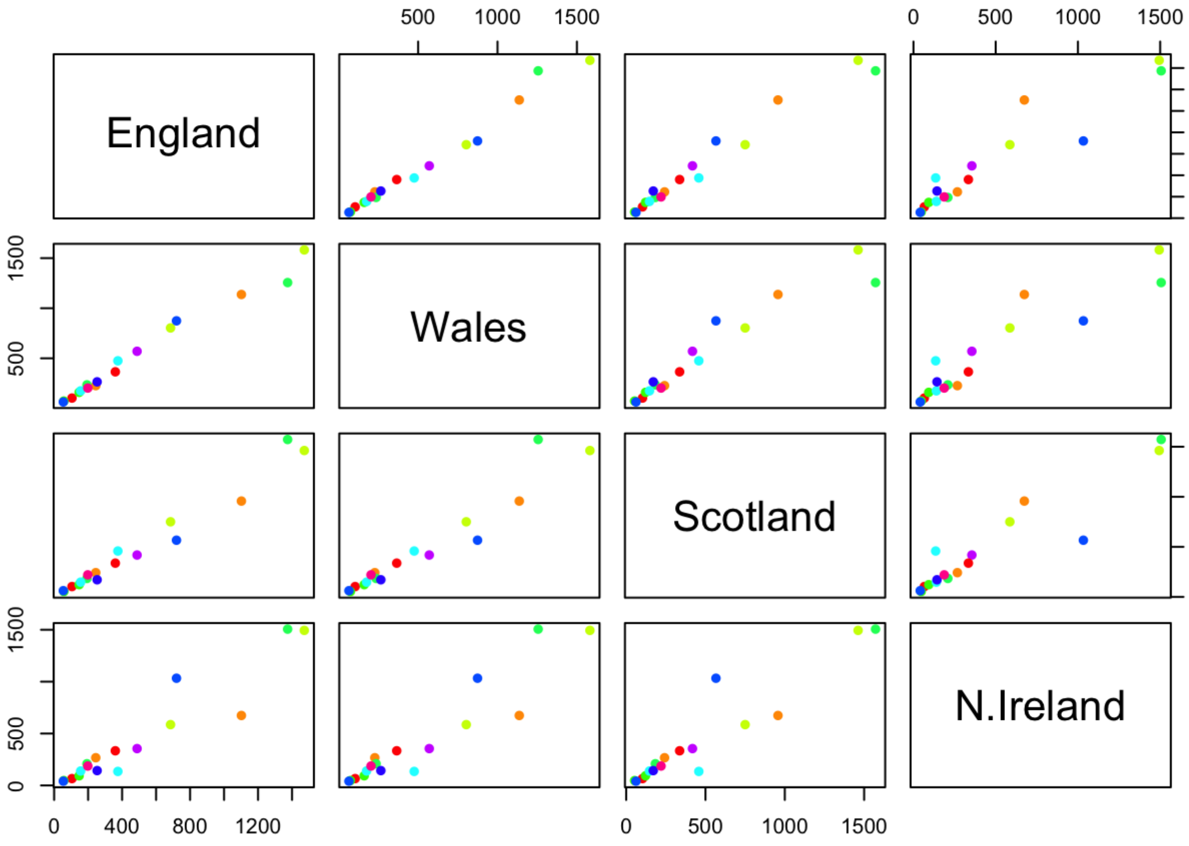
Q3: Changing what optional argument in the above **barplot()** function results in the following plot?

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



Q5: Generating all pairwise plots may help somewhat. Can you make sense of the following code and resulting figure? What does it mean if a given point lies on the diagonal for a given plot?

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



lying on diagonal means the two values are same

Q6. What is the main differences between N. Ireland and the other countries of the UK in terms of this data-set?

little fruit, lots of potato

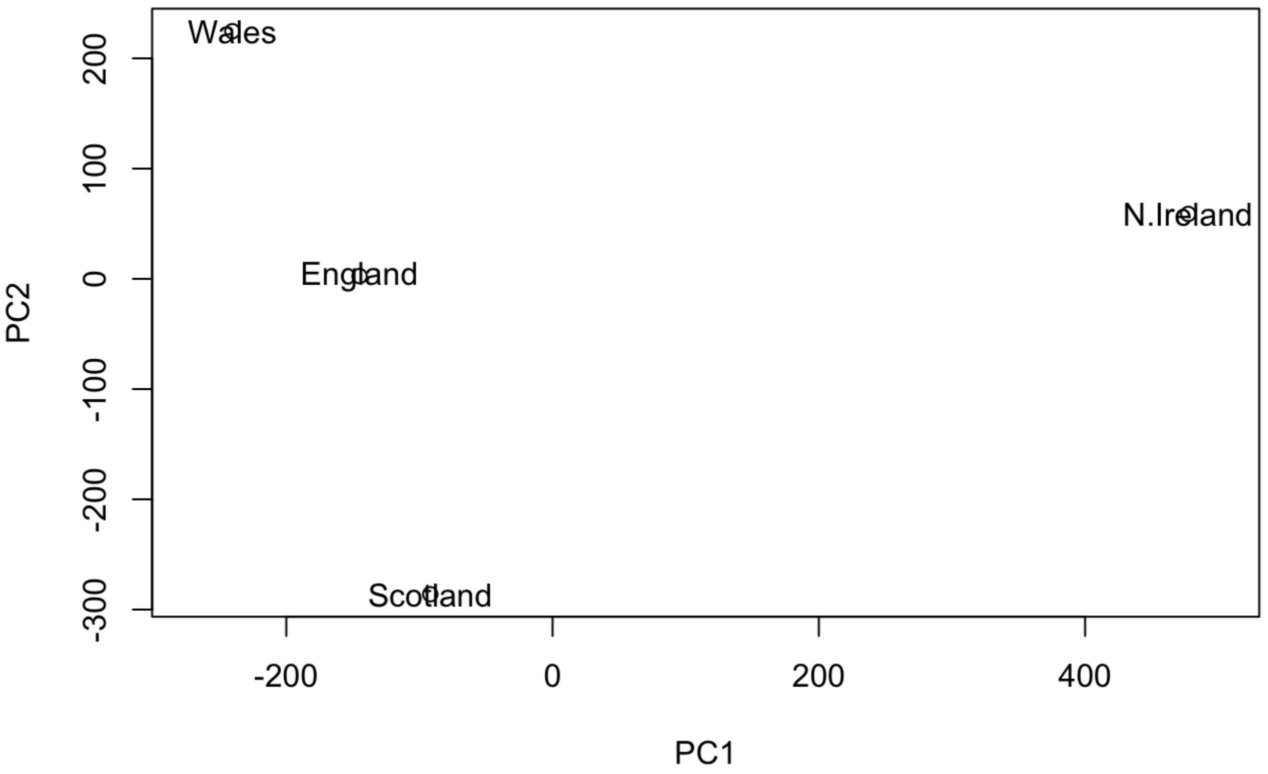
PCA starts here

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

```
Importance of components:
                            PC1      PC2      PC3       PC4
Standard deviation     324.1502 212.7478 73.87622 5.552e-14
Proportion of Variance   0.6744   0.2905  0.03503 0.000e+00
Cumulative Proportion    0.6744   0.9650  1.00000 1.000e+00
```

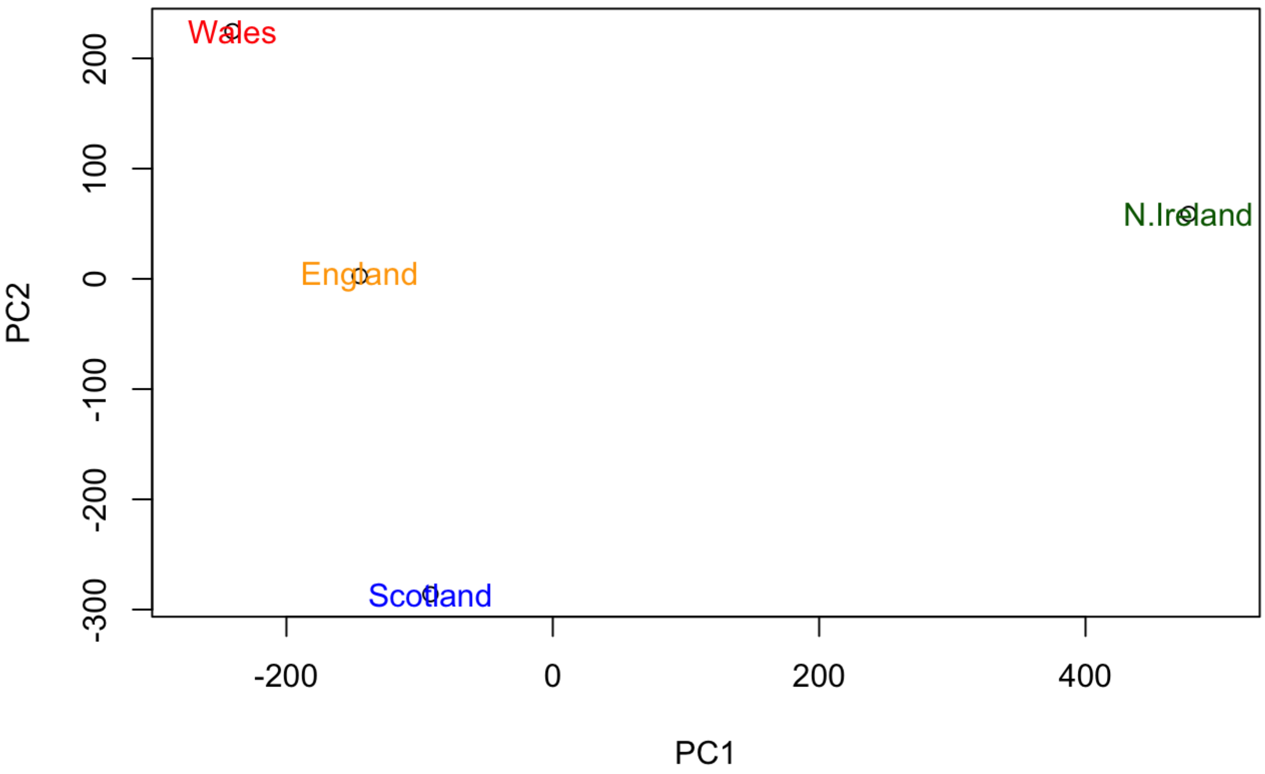
Q7. Complete the code below to generate a plot of PC1 vs PC2. The second line adds text labels over the data points.

```
plot(pca$x[,1], pca$x[,2], xlab="PC1", ylab="PC2", xlim=c(-270,500))
text(pca$x[,1], pca$x[,2], colnames(x))
```



Q8. Customize your plot so that the colors of the country names match the colors in our UK and Ireland map and table at start of this document.

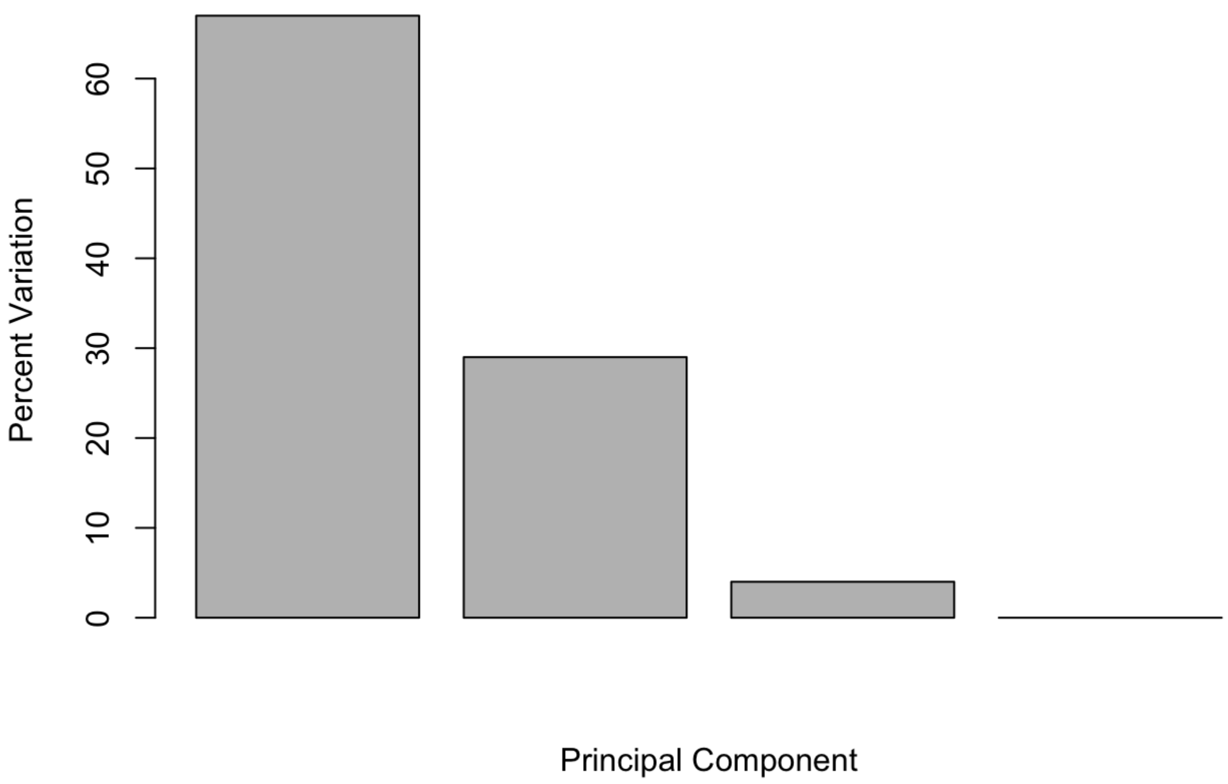
```
plot(pca$x[,1], pca$x[,2], xlab="PC1", ylab="PC2", xlim=c(-270,500))
text(pca$x[,1], pca$x[,2], colnames(x),col=c('orange','red','blue','darkgreen'))
```



```
v <- round( pca$sdev^2/sum(pca$sdev^2) * 100 )
z <- summary(pca)
z$importance
```

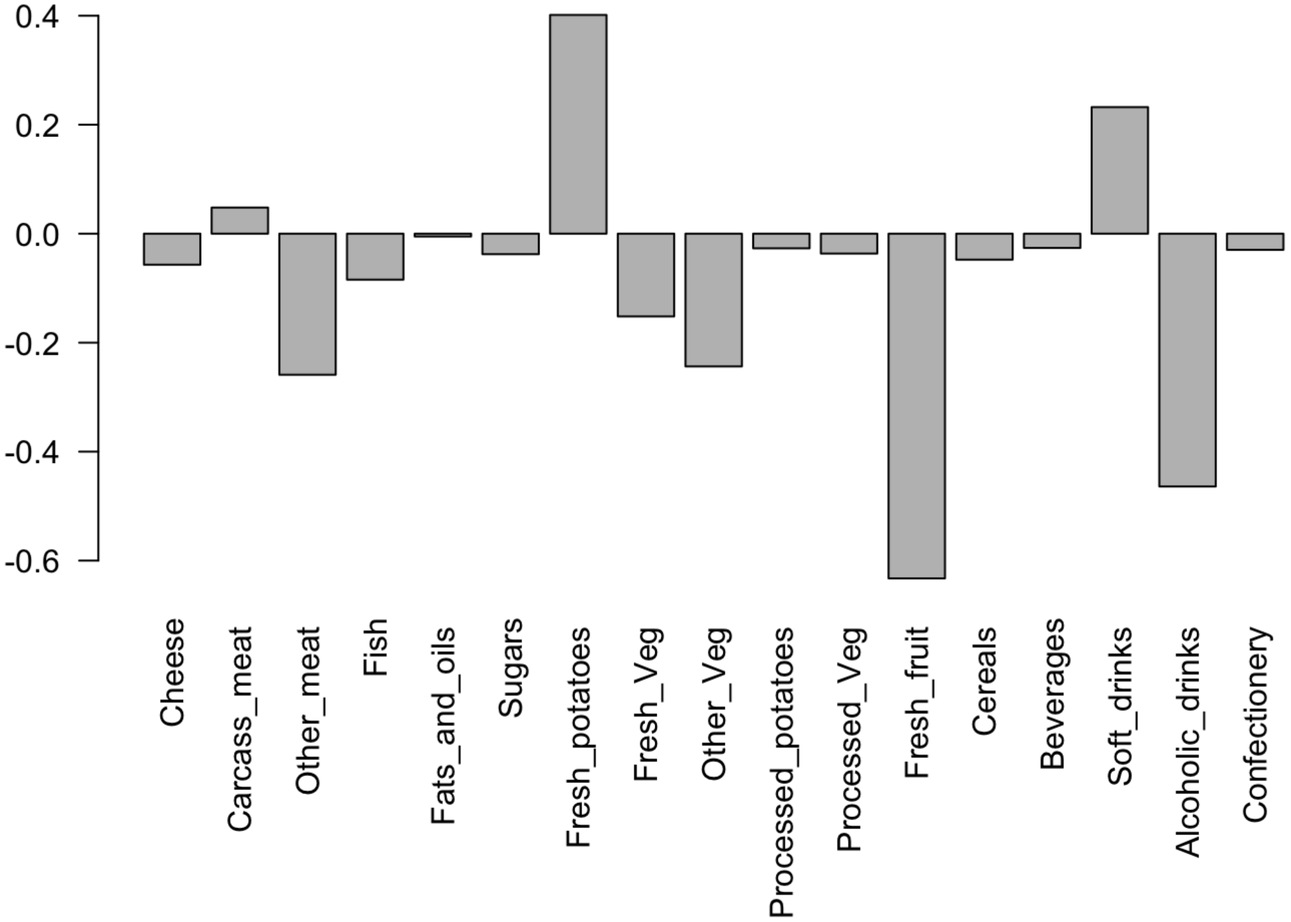
```
                             PC1       PC2      PC3          PC4
Standard deviation     324.15019 212.74780 73.87622 5.551558e-14
Proportion of Variance   0.67444   0.29052  0.03503 0.000000e+00
Cumulative Proportion    0.67444   0.96497  1.00000 1.000000e+00
```

```
barplot(v, xlab="Principal Component", ylab="Percent Variation")
```



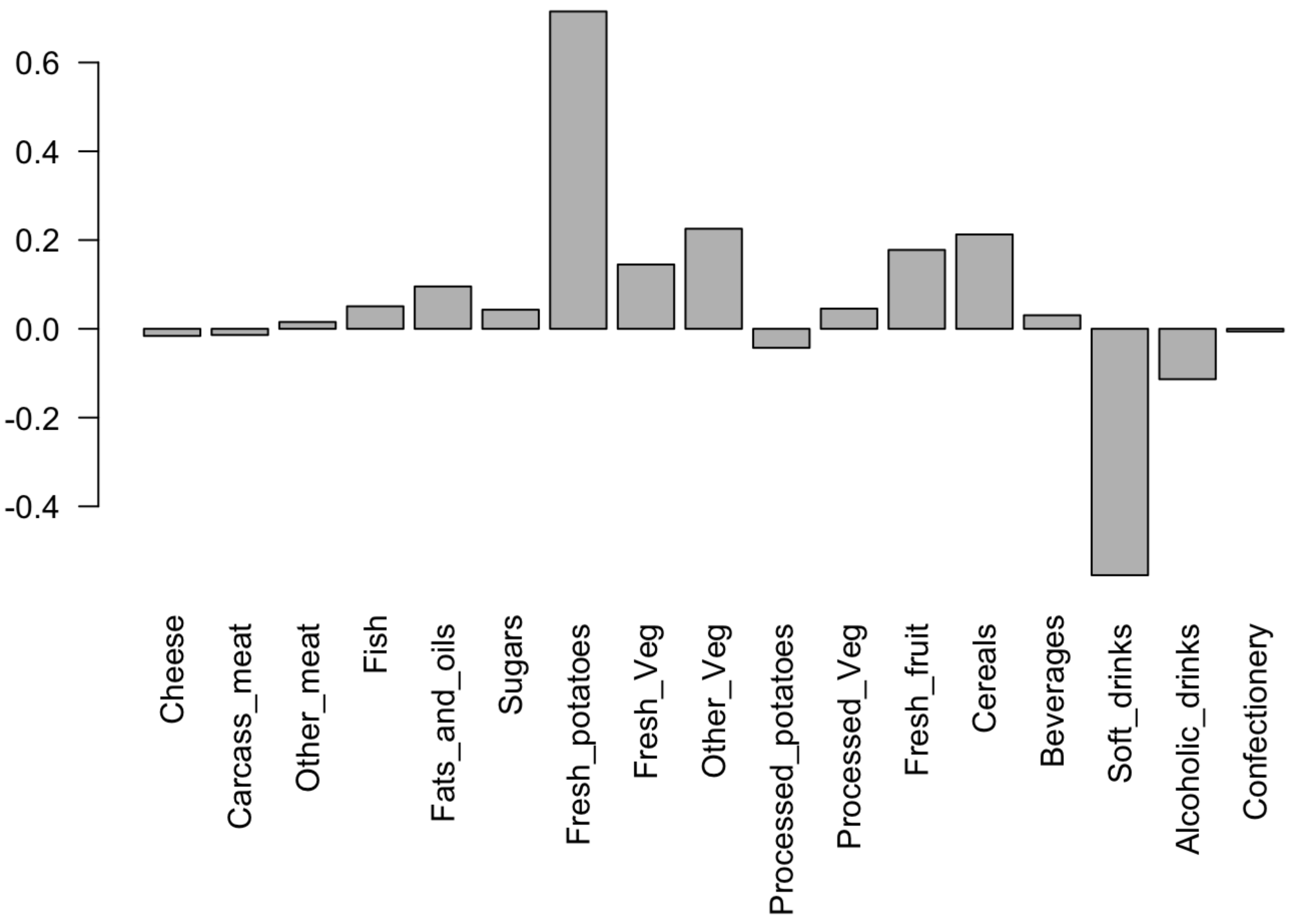
Variable loading: PCA1

```
par(mar=c(10, 3, 0.35, 0))
barplot( pca$rotation[,1], las=2 )
```



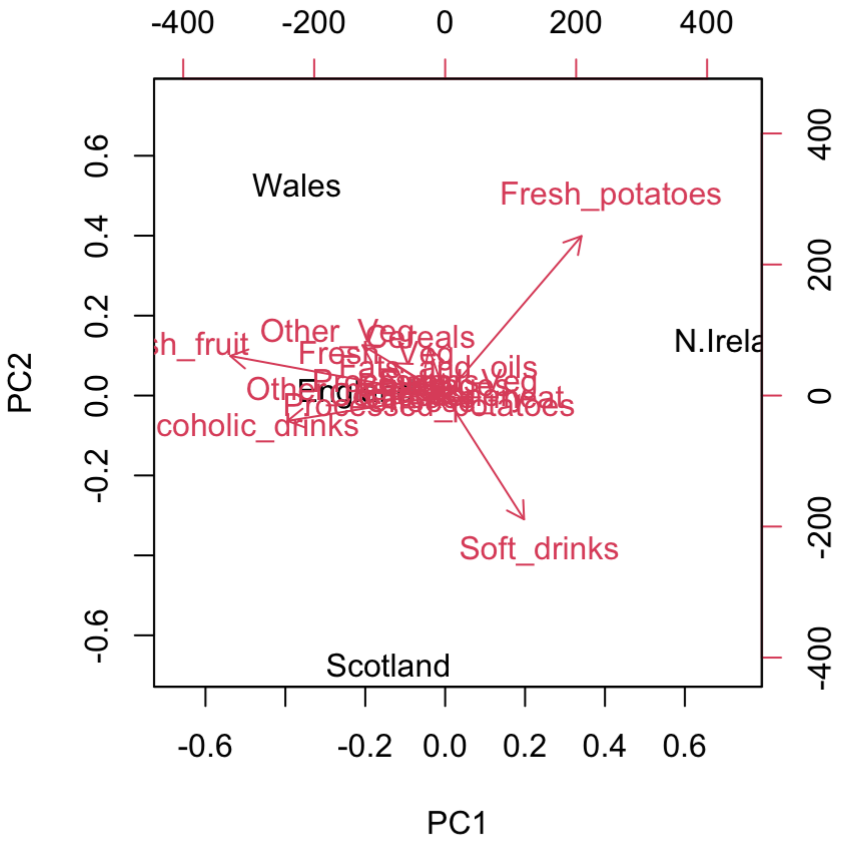
Q9: Variable loading: PCA2

```
par(mar=c(10, 3, 0.35, 0))
barplot( pca$rotation[,2], las=2 )
```



Biplot:

```
biplot(pca)
```


-400
-200
0
200
400
Wales
Fresh_potatoes
sh_fruit
Other_Veg
Cereals
Fresh_Veg
Fats_and_oils
N.Irela
Eng
Processed_potatoes
coholic_drinks
Soft_drinks
Scotland
0.6
0.4
0.2
0.0
-0.2
-0.6
PC2
400
200
0
-200
-400
-0.6
-0.2
0.0
0.2
0.4
0.6
PC1