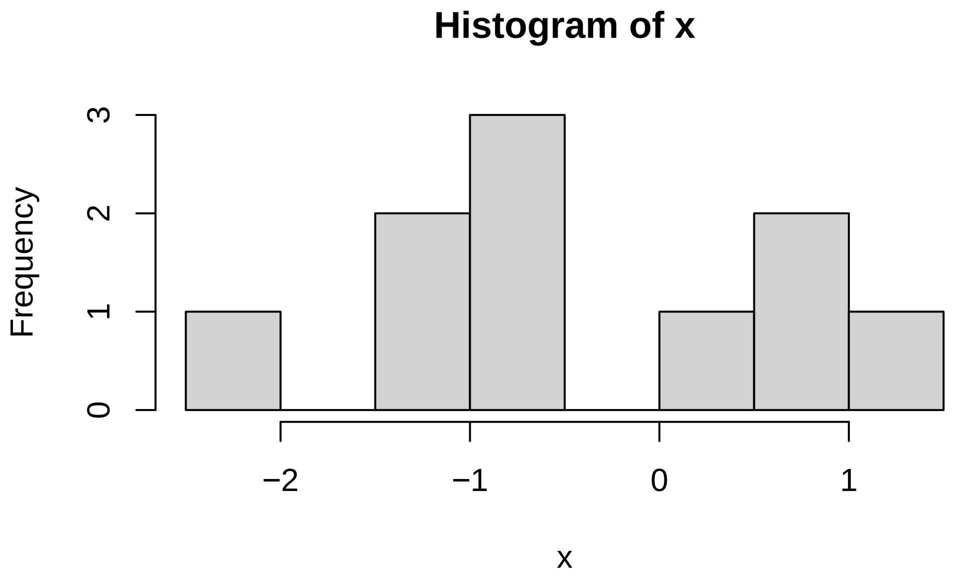
class07

Daniel Xu

#k means clustering test

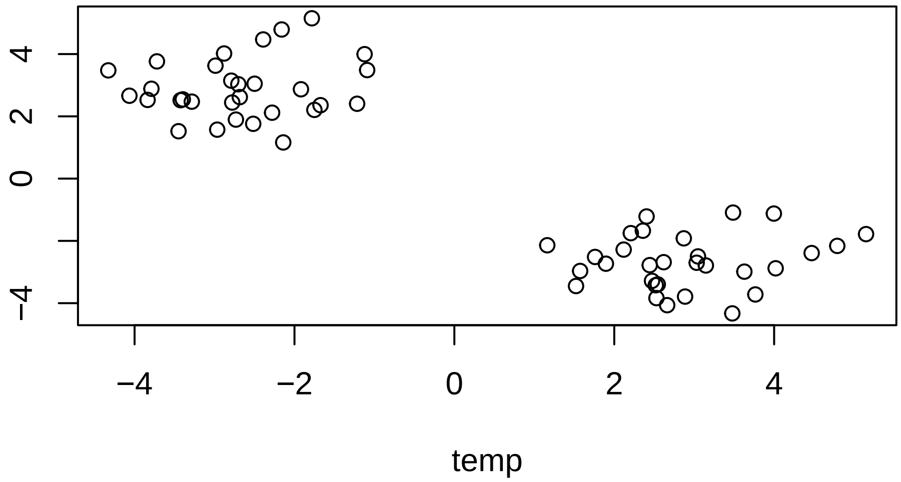
```
x<-rnorm(10)
hist(x)
```



center around -3

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

```
plot(x)
```



kmeans test

```
km<-kmeans(x,centers=3,nstart=20)
km
```

```
K-means clustering with 3 clusters of sizes 6, 30, 24

Cluster means:
       temp
1 -1.904589  4.318270
2  2.884870 -2.678534
3 -2.872020  2.526520

Clustering vector:
 [1] 3 3 1 3 3 3 1 3 3 3 3 3 3 3 3 3 3 3 3 1 1 3 3 3 3 3 1 3 1 3 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

Within cluster sum of squares by cluster:
[1]  4.36463 48.92352 24.65670
```

```
 (between_SS / total_SS =  92.4 %)

Available components:

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

```
#how many points are in each cluster
km$size
```

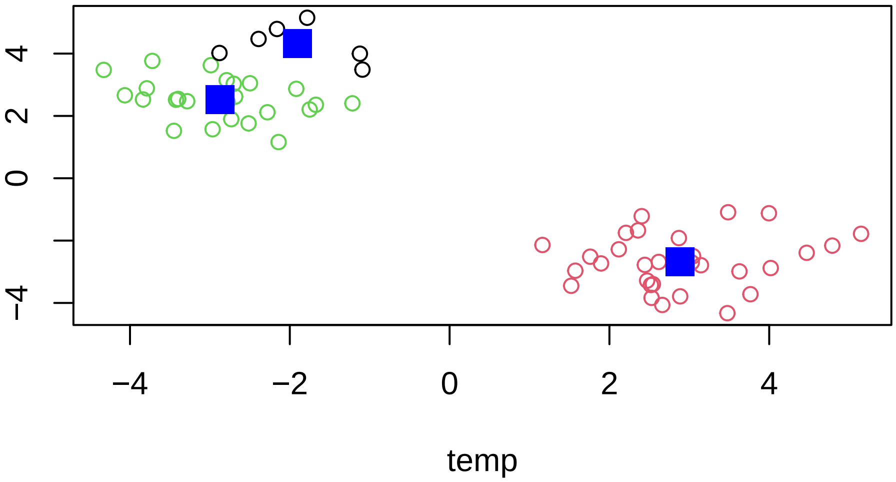
```
[1]  6 30 24
```

Q what components of your result object details

```
cluster<-km$cluster
km$center
```

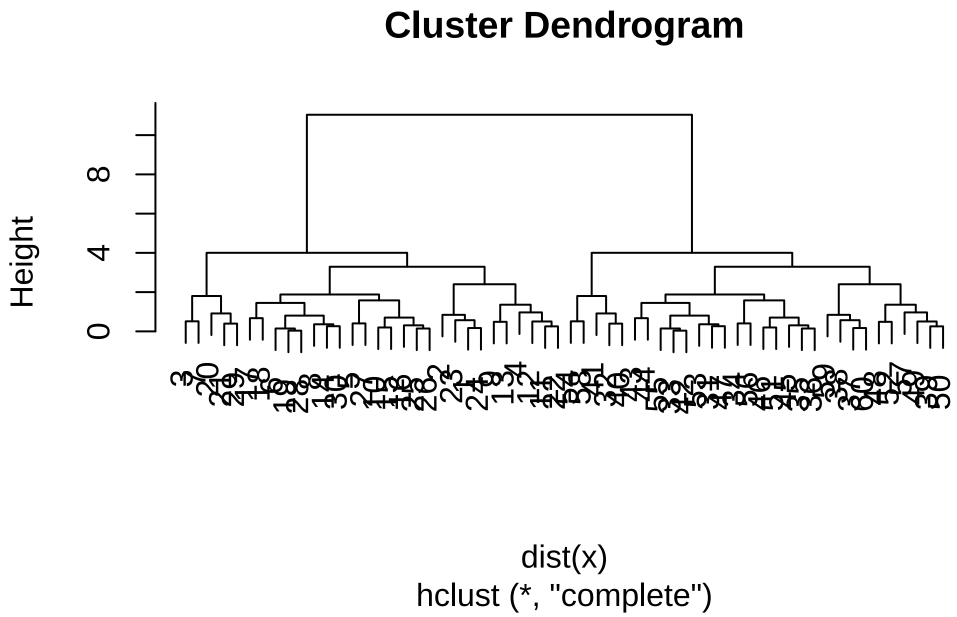
```
       temp
1 -1.904589  4.318270
2  2.884870 -2.678534
3 -2.872020  2.526520
```

```
plot(x,col=cluster)
points(km$center,col="blue",pch=15,cex=2)
```

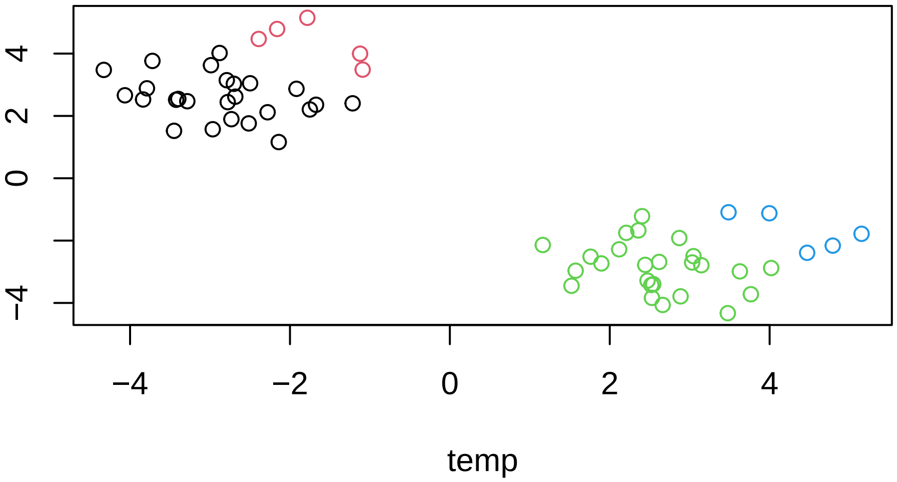


heirarchal test

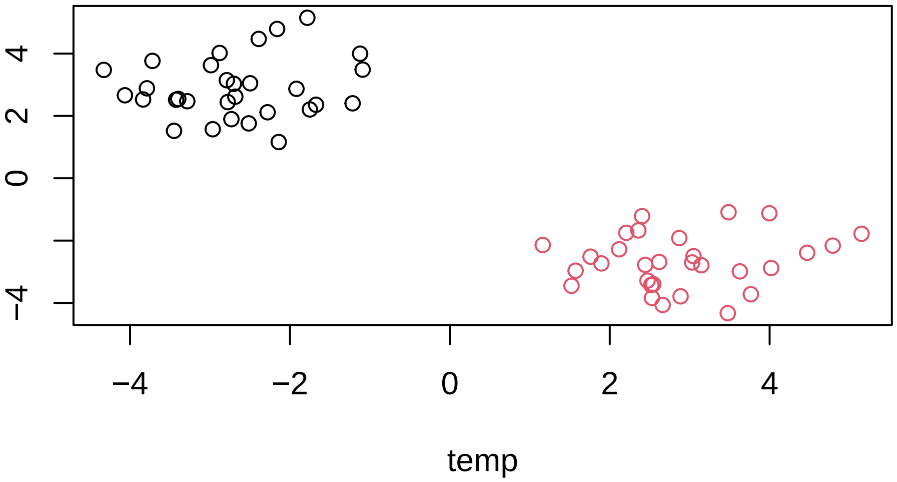
```
hc<-hclust(dist(x))
plot(hc)
```



```
cluster<-cutree(hc,h=4)
c2<-cutree(hc,k=2)
plot(x,col=cluster)
```



```
plot(x,col=c2)
```



Principal Componenet Analysis(PCA)

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

```
[1] 17  5
```

Q1) There are 17 rows with 5 columns

```
rownames(x) <- x[,1]
x<-x[,-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
```

```
dim(x)
```

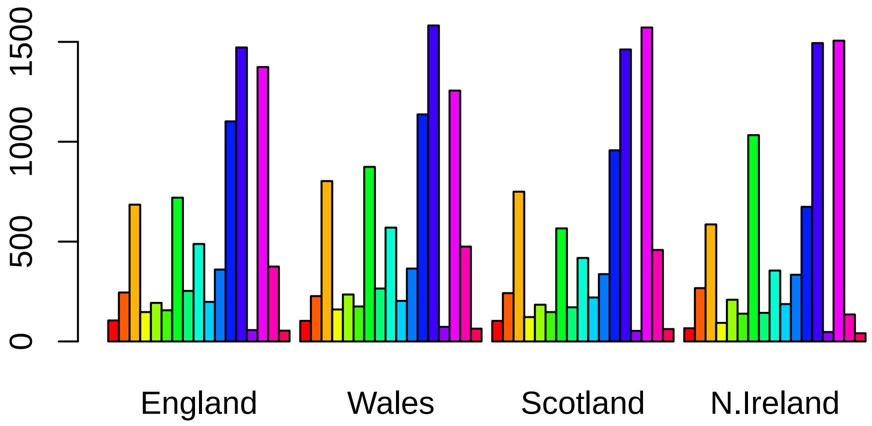
```
[1] 17  4
```

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

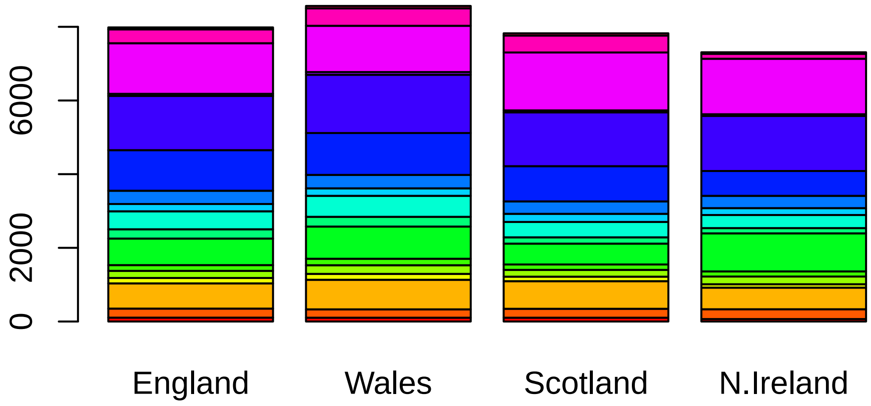
```
               England Wales Scotland N.Ireland
Cheese             105   103      103        66
Carcass_meat       245   227      242       267
Other_meat         685   803      750       586
Fish               147   160      122        93
Fats_and_oils      193   235      184       209
Sugars  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

Q2) The second method is better because there is less code. It's usually better to specify which row is the names rather than use exclusion of the names

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



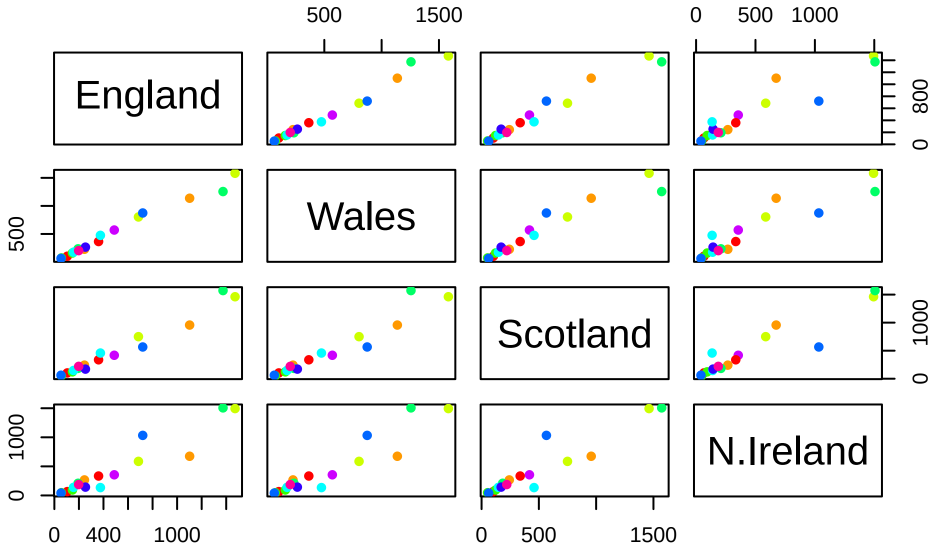
```
barplot(as.matrix(x), col=rainbow(nrow(x)))
```



Q3 You can remove beside=T which results in that plot

Q5

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



If they lie on the diagonal that means, that the two countries have similar consumption of that commodity. The pairwise plot puts each countries on one axis and plots the values for each commodity with those country axes.

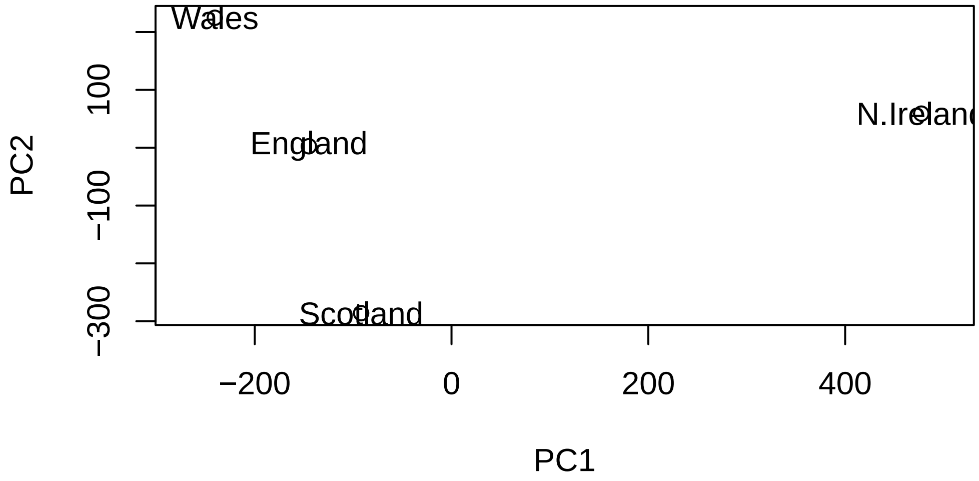
Q6 North Ireland seems to produce very different amounts when compared to other countries as many of the commodities stray from the diagonal for every single other UK country specifically potatoes and soft drinks.

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

```
Importance of components:
                            PC1      PC2      PC3       PC4
Standard deviation     324.1502 212.7478 73.87622 4.189e-14
Proportion of Variance   0.6744   0.2905  0.03503 0.000e+00
Cumulative Proportion    0.6744   0.9650  1.00000 1.000e+00
```

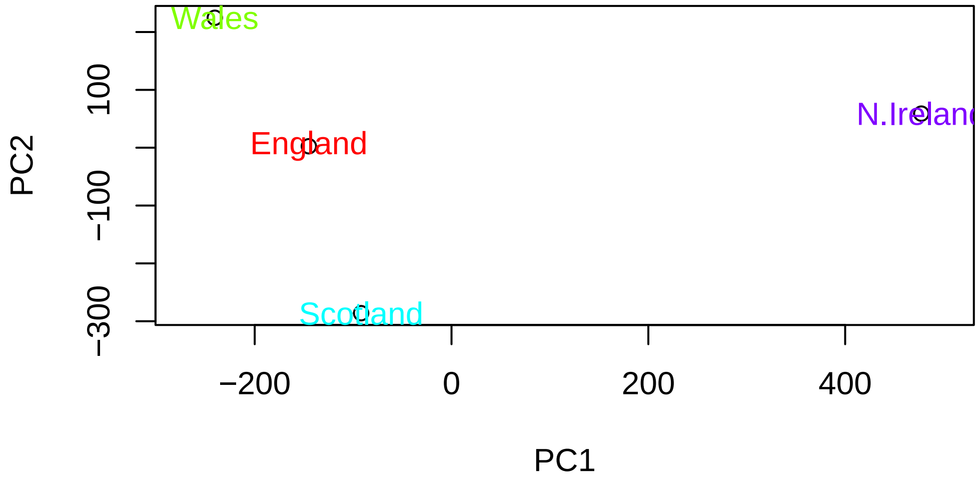
Q7

```
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

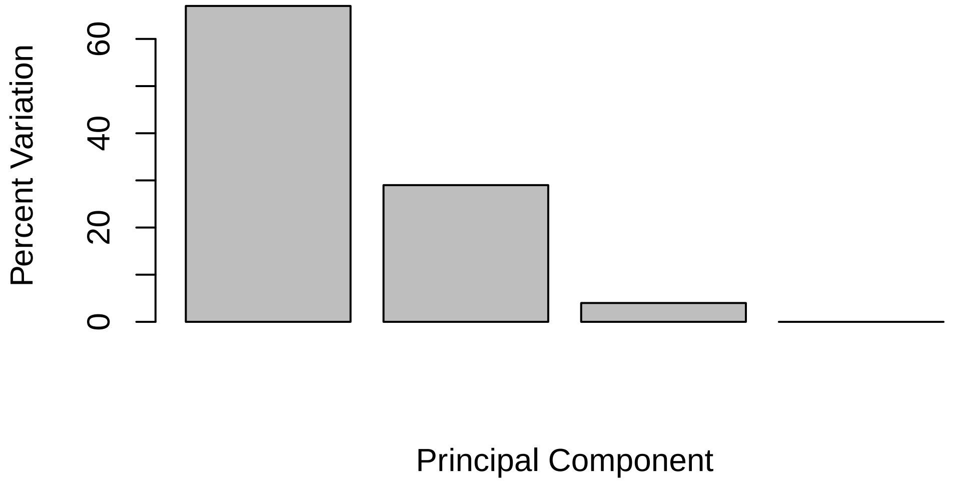
```
plot(pca$x[,1], pca$x[,2], xlab="PC1", ylab="PC2", xlim=c(-270,500))
text(pca$x[,1], pca$x[,2], colnames(x),col=rainbow(4))
```



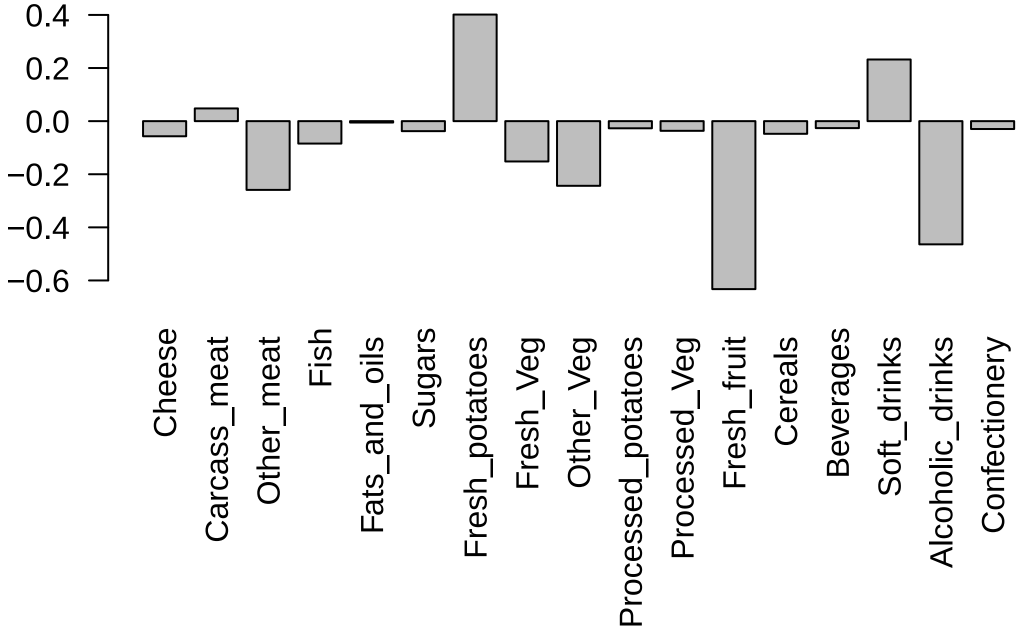
```
v <- round( pca$sdev^2/sum(pca$sdev^2) * 100 )
v
```

```
[1] 67 29  4  0
```

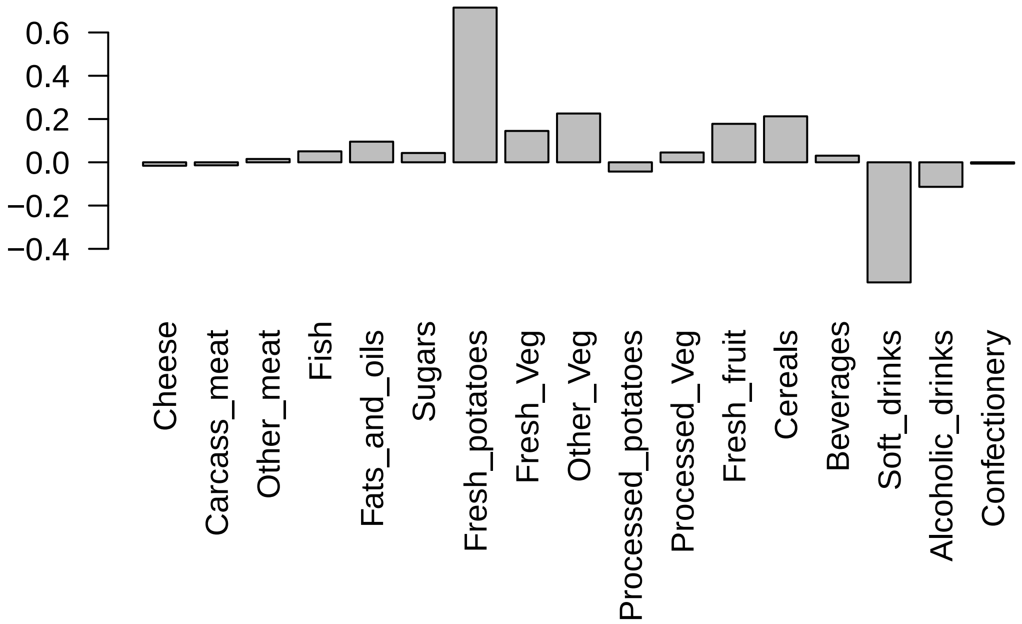
```
barplot(v, xlab="Principal Component", ylab="Percent Variation")
```



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

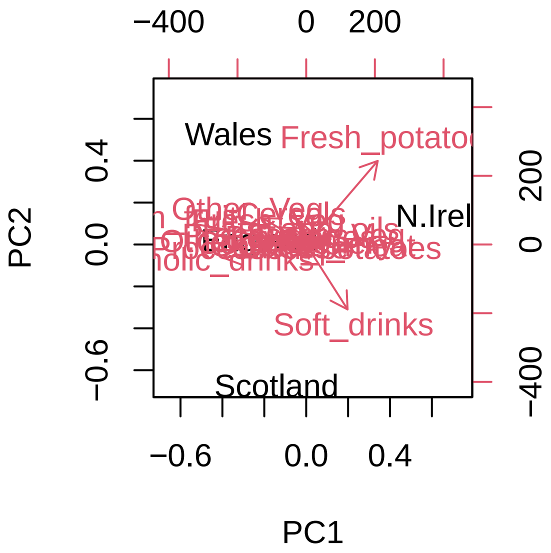


```r
par(mar=c(10, 3, 0.35, 0))
barplot( pca$rotation[,2], las=2 )
```



Q9) The two groups that are prominent are soft drinks and fresh potatoes. This tells us that fresh_potatoes pushes N. ireland up for fresh potatoes and down for soft drinks. Specifically, N.ireland consumes more fresh potatoes and consumes less soft drinks compared to other UK countries

```r
biplot(pca)
```



```r
url2 <- "https://tinyurl.com/expression-CSV"
rna.data <- read.csv(url2, row.names=1)
head(rna.data)
```

```
       wt1 wt2  wt3  wt4 wt5 ko1 ko2 ko3 ko4 ko5
gene1  439 458  408  429 420  90  88  86  90  93
gene2  219 200  204  210 187 427 423 434 433 426
gene3 1006 989 1030 1017 973 252 237 238 226 210
gene4  783 792  829  856 760 849 856 835 885 894
gene5  181 249  204  244 225 277 305 272 270 279
gene6  460 502  491  491 493 612 594 577 618 638
```

```r
dim(rna.data)
```

```
[1] 100  10
```

Q10 There are 10 genes with 100 samples