

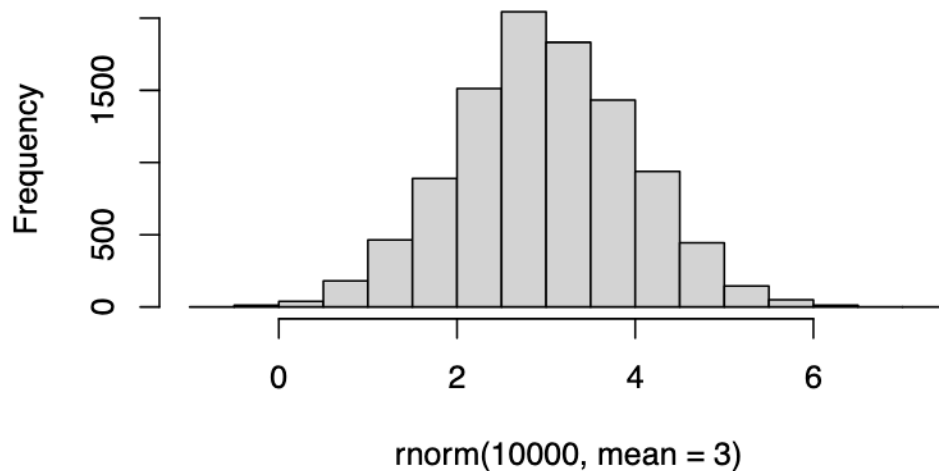
Class07

```
rmnorm(10)
```

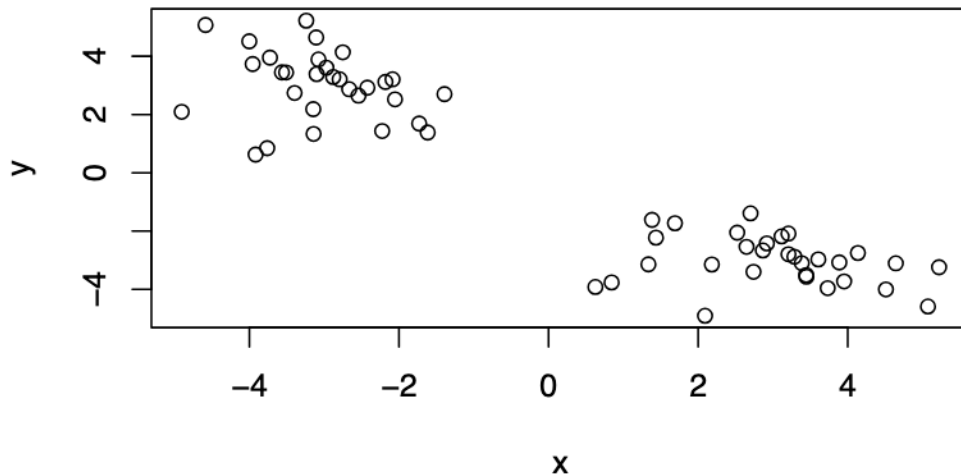
```
[1] -0.7041508 -0.3406981  1.3403915 -2.0842758  1.7103341  1.0190302  
[7] -1.0318180  0.1484586  1.2454737 -1.3385844
```

```
hist(rnorm(10000,mean=3))
```

Histogram of rnorm(10000, mean = 3)



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



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

Cluster means:

	x	y
1	2.994745	-3.014638
2	-3.014638	2.994745

Clustering vector:

[illegible]

Within cluster sum of squares by cluster:

```
[1] 61.44498 61.44498
(between_SS / total_SS = 89.8 %)
```

Available components:

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

Q. How many points are in each cluster

k\$size

[1] 30 30

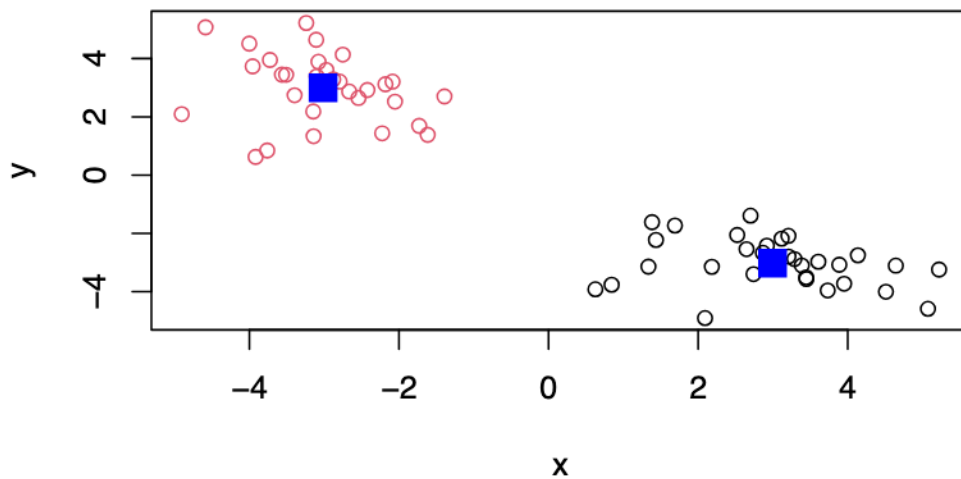
<Q2. The clustering result i.e. membership vector?

```
k$cluster
```

[illegible]

Q4. Make a plot of our data colored by clustering results with optionally the cluster centers shown

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

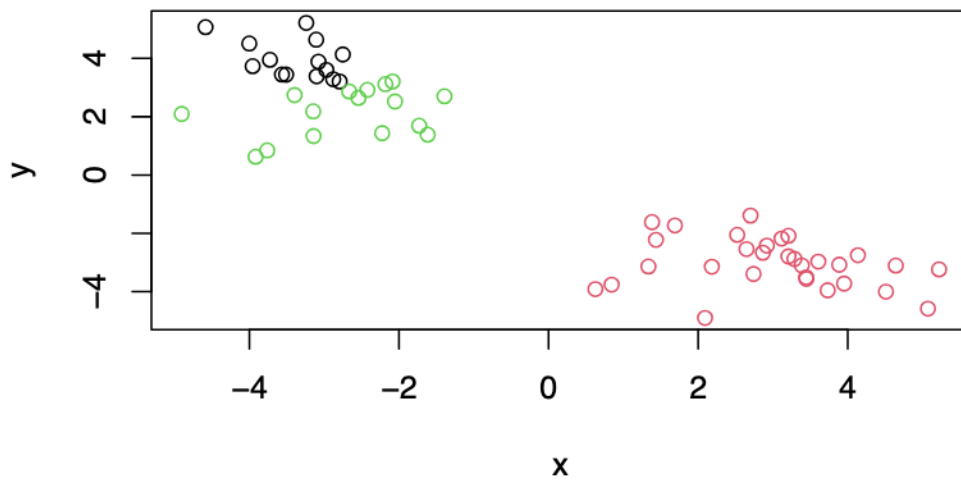


Q5. Run Kmeans again but cluster into 3 groups and plot the results

```
k<-kmeans(x,centers=3,nstart=20)
k$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 3 3 1 3 3 3 3
[39] 3 1 3 1 3 3 1 1 3 3 3 1 1 1 1 3 3 1 3 1 1 1
```

```
plot(x,col=k$cluster)
```



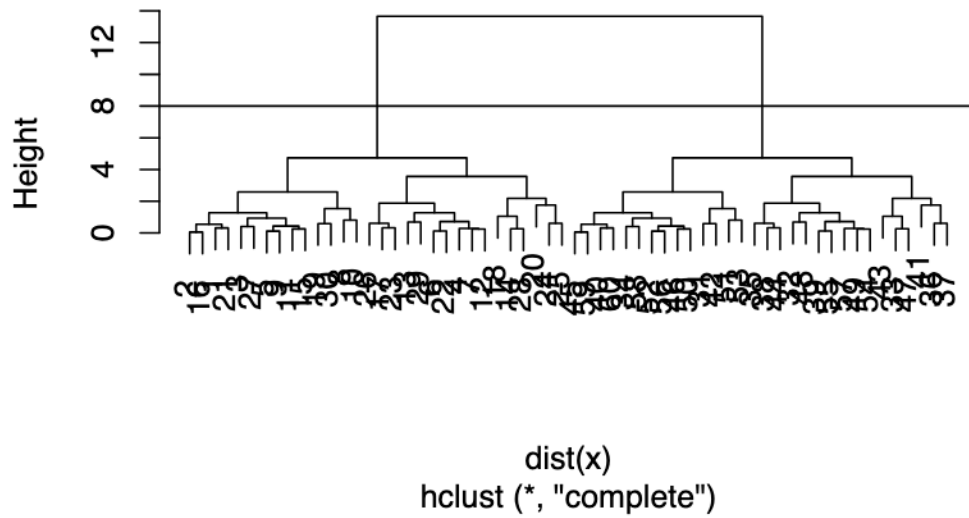
#Hierarchical Clustering

The main function in baseR is 'hclust()'

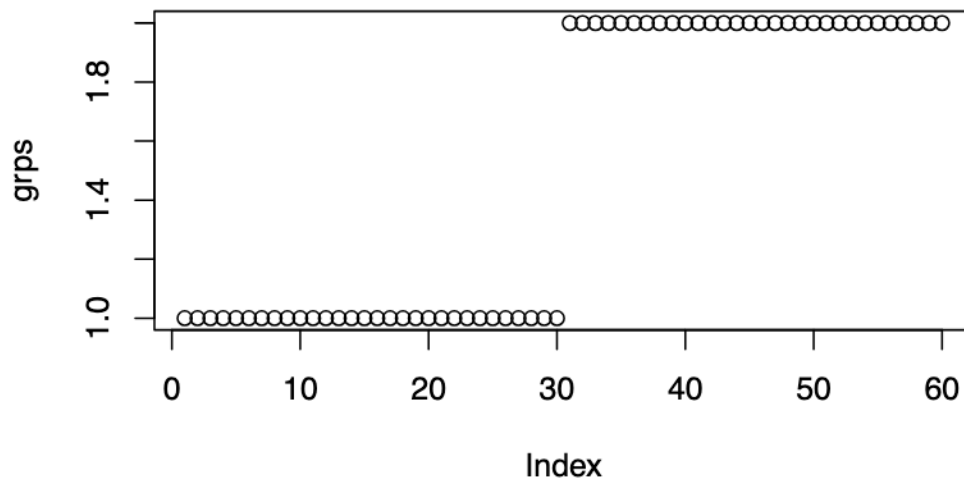
```
hc<-hclust(dist(x))
```

```
plot(hc)
abline(h=8)
```

Cluster Dendrogram

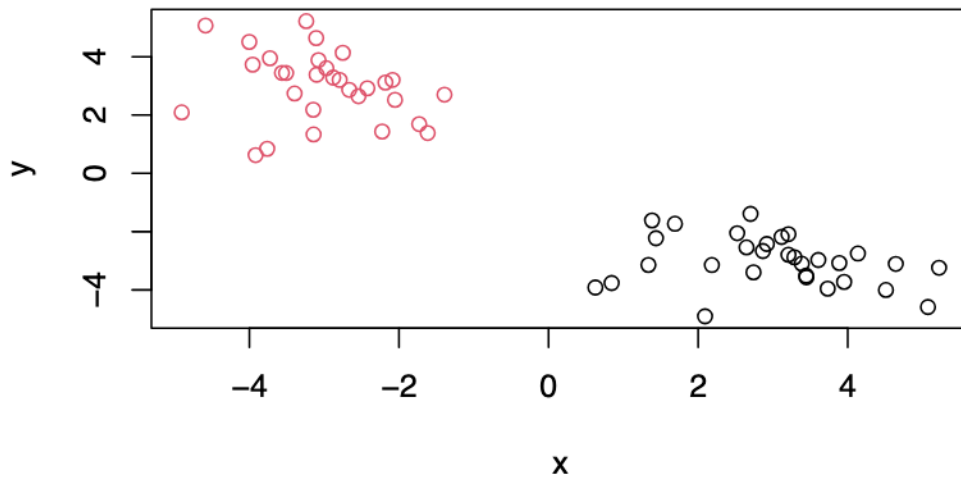


```
grps<-cutree(hc,h=8)
plot(grps)
```



Q. Plot our hclust results in terms of our data colored by cluster membership

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



Lab portion

```
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

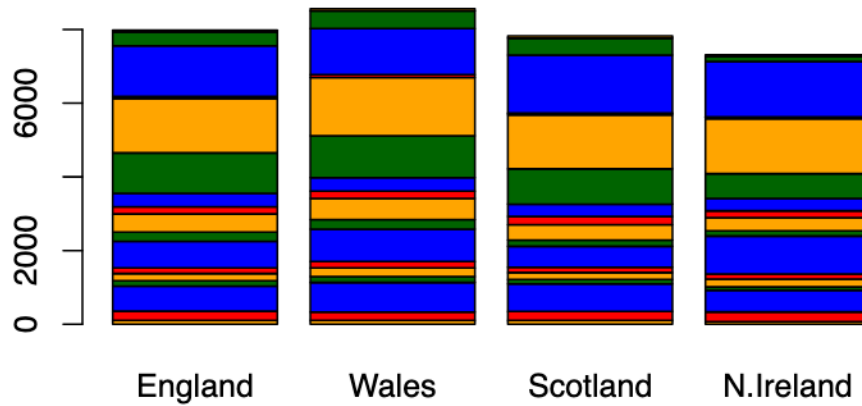
```
dim(x)
```

```
[1] 17 4
```

Q1> 5 columns and 17 rows, but if you do row.names=1, there are 17 rows and 4 columns

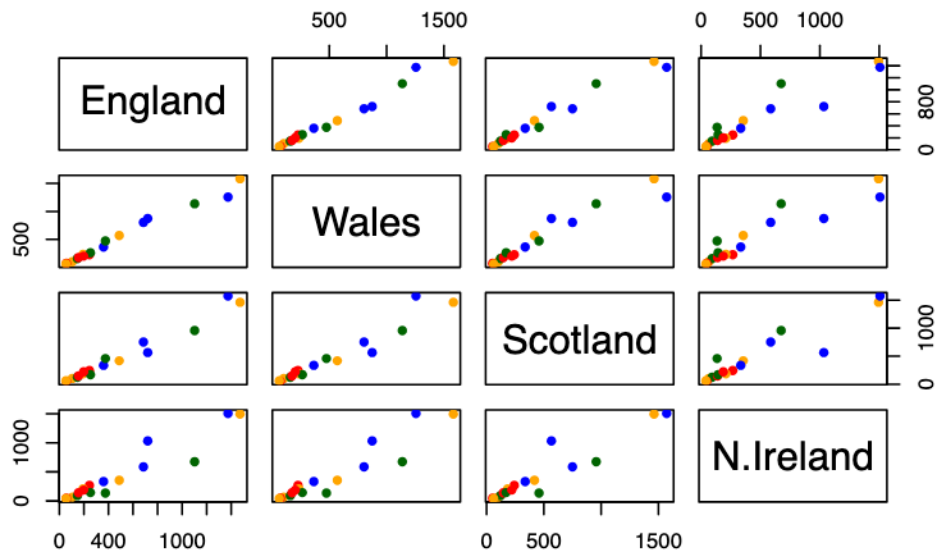
Q2> I prefer to import the df using row.names=1 so I know I am always working with the same df in following lines ‘

```
colors= (c("orange","red","blue","dark green"))
barplot(as.matrix(x), beside=F, col=colors)
```



Q5> horiz! I”f FALSE, the bars are drawn vertically with the first bar to the left. If TRUE, the bars are drawn horizontally with the first at the bottom.”

```
pairs(x, col=colors, pch=16)
```



Q6> There is a stronger outlier in the N. Ireland dataset that may affect the coorelation value

```
t(x) #transpose x
```

	Cheese	Carcass_meat	Other_meat	Fish	Fats_and_oils	Sugars
England	105	245	685	147	193	156
Wales	103	227	803	160	235	175
Scotland	103	242	750	122	184	147
N.Ireland	66	267	586	93	209	139
	Fresh_potatoes	Fresh_Veg	Other_Veg	Processed_potatoes		
England		720	253	488		198
Wales		874	265	570		203
Scotland		566	171	418		220
N.Ireland		1033	143	355		187
	Processed_Veg	Fresh_fruit	Cereals	Beverages	Soft_drinks	
England		360	1102	1472	57	1374
Wales		365	1137	1582	73	1256
Scotland		337	957	1462	53	1572
N.Ireland		334	674	1494	47	1506
	Alcoholic_drinks	Confectionery				
England		375		54		

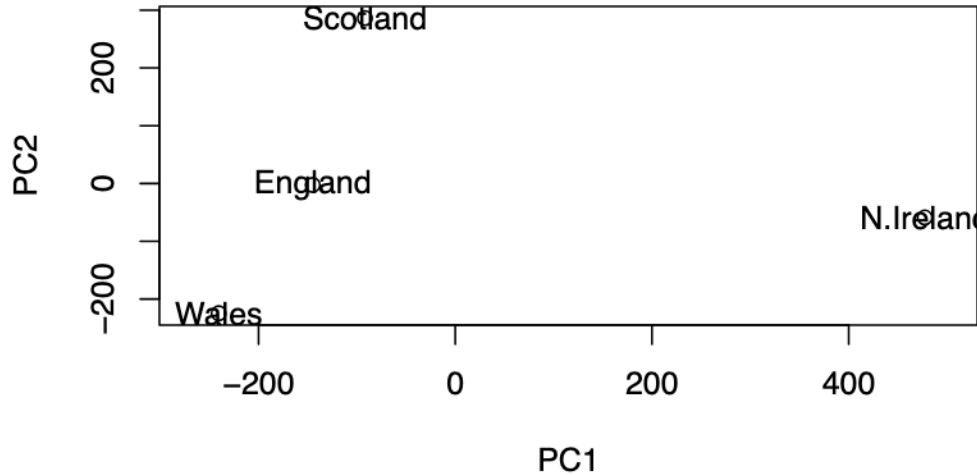
Wales	475	64
Scotland	458	62
N.Ireland	135	41

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

Importance of components:

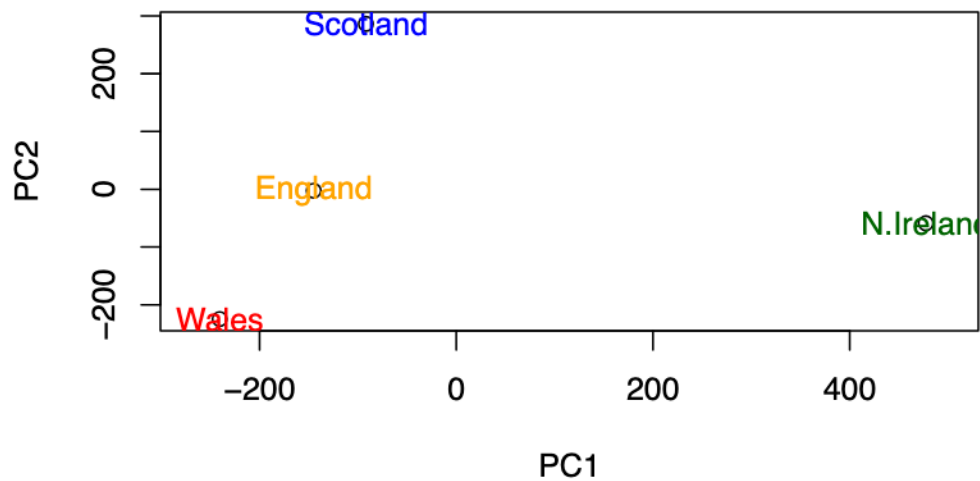
	PC1	PC2	PC3	PC4
Standard deviation	324.1502	212.7478	73.87622	3.176e-14
Proportion of Variance	0.6744	0.2905	0.03503	0.000e+00
Cumulative Proportion	0.6744	0.9650	1.00000	1.000e+00

```
# Plot PC1 vs PC2
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 PC1 vs PC2
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=colors)
```



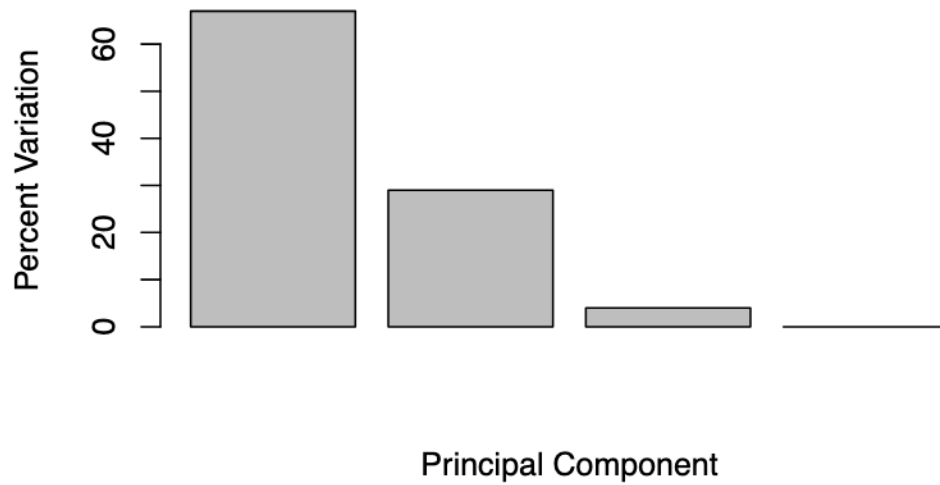
```
#variance from diff PC
v <- round( pca$sdev^2/sum(pca$sdev^2) * 100 )
v
```

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

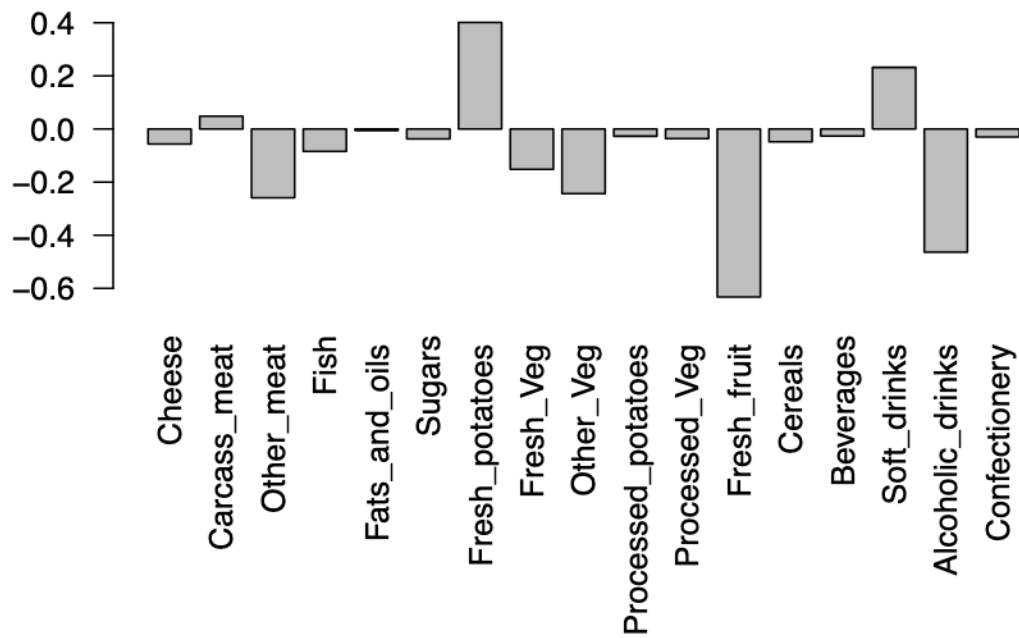
```
## or the second row here...
z <- summary(pca)
z$importance
```

	PC1	PC2	PC3	PC4
Standard deviation	324.15019	212.74780	73.87622	3.175833e-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")
```

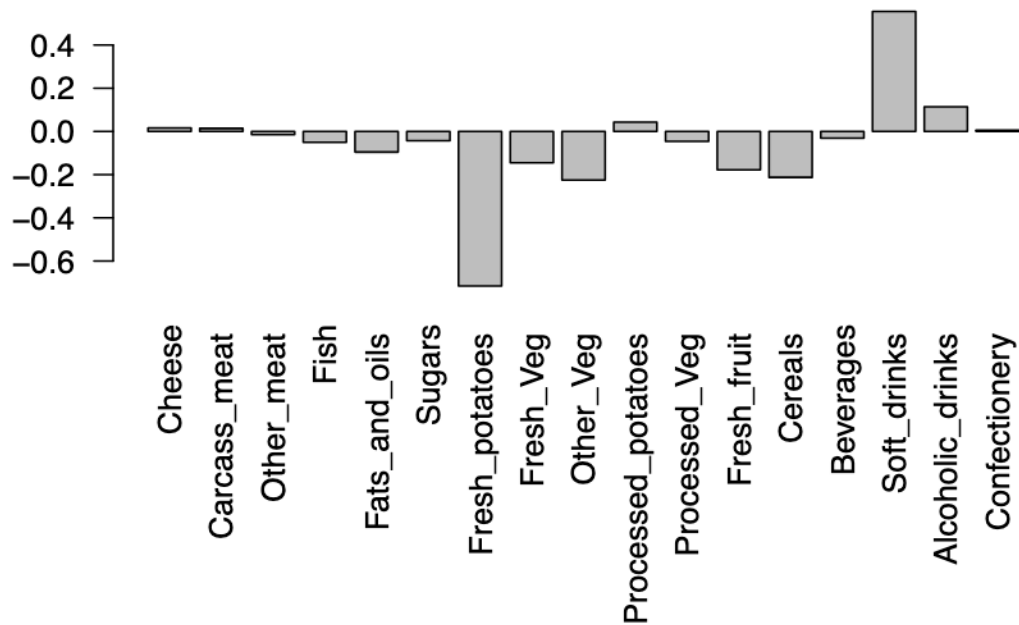


```
## Lets focus on PC1 as it accounts for > 90% of variance  
par(mar=c(10, 3, 0.35, 0))  
barplot( pca$rotation[,1], las=2 )
```



Q9: Generate a similar 'loadings plot' for PC2. What two food groups feature prominently and what does PC2 mainly tell us about?

```
##PC2
par(mar=c(10, 3, 0.35, 0))
barplot( pca$rotation[,2], las=2 )
```



It tells us about Fresh_potatoes and soft_drinks. wales eats more fresh potatoes than the rest and drinks fewer soft drinks.

Q10> How many genes and samples are in this data set?

```
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

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

```
[1] 100 10
```