

# Class 7: Machine Learning 1

Marielle Samonte (A16861951)

## Table of contents

Clustering . . . . .	1
K-means . . . . .	4
Hierarchical Clustering . . . . .	7
Principal Component Analysis (PCA) . . . . .	9
Data import . . . . .	9
PCA to the rescue . . . . .	11
Interpreting PCA results . . . . .	13

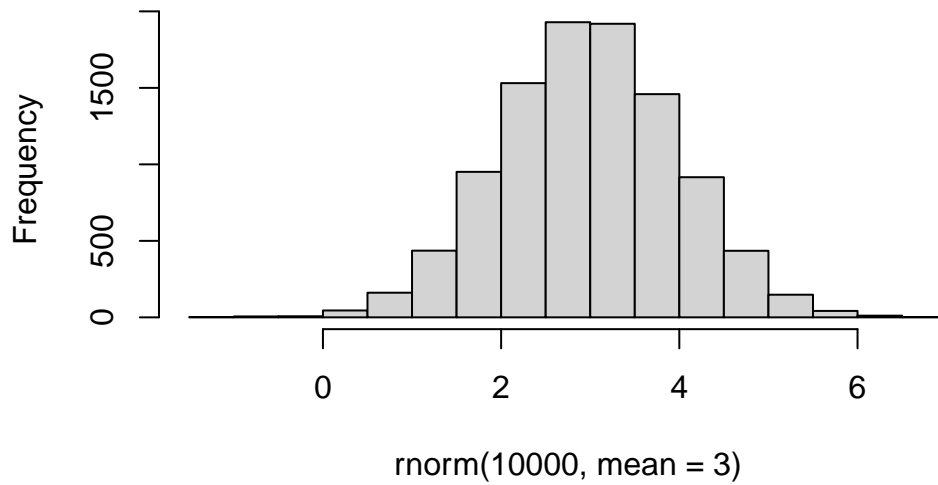
Today we will explore unsupervised machine learning methods starting with clustering and dimensionality reduction.

## Clustering

To start let's make up some data to cluster where we know what the answer should be. The `rnorm()` function will help us here.

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

## Histogram of rnorm(10000, mean = 3)



Return 30 numbers centered on -3

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

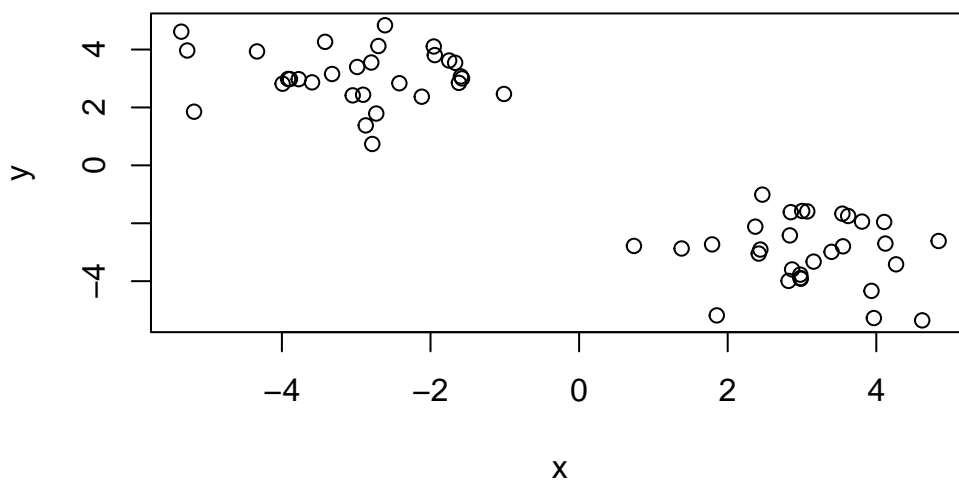
	x	y
[1,]	-2.7835745	0.7381569
[2,]	-2.7970650	3.5515869
[3,]	-2.7016577	4.1221019
[4,]	-2.9867701	3.3977425
[5,]	-5.3542573	4.6178839
[6,]	-1.6176358	2.8492845
[7,]	-1.5904533	3.0707964
[8,]	-3.9906659	2.8193665
[9,]	-1.5753369	3.0028574
[10,]	-2.6124502	4.8388794
[11,]	-3.9174350	2.9831974
[12,]	-5.1824142	1.8529387
[13,]	-1.9559443	4.1056746

[14,]	-3.8917141	2.9792577
[15,]	-2.1176416	2.3713118
[16,]	-4.3349899	3.9352457
[17,]	-3.7739223	2.9766560
[18,]	-2.4191864	2.8363971
[19,]	-1.6683595	3.5435666
[20,]	-1.7491393	3.6213667
[21,]	-2.9072707	2.4394641
[22,]	-3.0471746	2.4175385
[23,]	-1.0114522	2.4662801
[24,]	-3.5942969	2.8669826
[25,]	-5.2742634	3.9674432
[26,]	-2.7299631	1.7898808
[27,]	-3.3247053	3.1561984
[28,]	-1.9422761	3.8077817
[29,]	-3.4187287	4.2662170
[30,]	-2.8716619	1.3796645
[31,]	1.3796645	-2.8716619
[32,]	4.2662170	-3.4187287
[33,]	3.8077817	-1.9422761
[34,]	3.1561984	-3.3247053
[35,]	1.7898808	-2.7299631
[36,]	3.9674432	-5.2742634
[37,]	2.8669826	-3.5942969
[38,]	2.4662801	-1.0114522
[39,]	2.4175385	-3.0471746
[40,]	2.4394641	-2.9072707
[41,]	3.6213667	-1.7491393
[42,]	3.5435666	-1.6683595
[43,]	2.8363971	-2.4191864
[44,]	2.9766560	-3.7739223
[45,]	3.9352457	-4.3349899
[46,]	2.3713118	-2.1176416
[47,]	2.9792577	-3.8917141
[48,]	4.1056746	-1.9559443
[49,]	1.8529387	-5.1824142
[50,]	2.9831974	-3.9174350
[51,]	4.8388794	-2.6124502
[52,]	3.0028574	-1.5753369
[53,]	2.8193665	-3.9906659
[54,]	3.0707964	-1.5904533
[55,]	2.8492845	-1.6176358
[56,]	4.6178839	-5.3542573

```
[57,] 3.3977425 -2.9867701
[58,] 4.1221019 -2.7016577
[59,] 3.5515869 -2.7970650
[60,] 0.7381569 -2.7835745
```

Make a plot of x

```
plot(x)
```



## K-means

The main function in “base” R for K-means clustering is called `kmeans()`:

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

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

Cluster means:

	x	y
1	3.092391	-2.971414

```
2 -2.971414 3.092391
```

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 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Within cluster sum of squares by cluster:

```
[1] 63.47923 63.47923
(between_SS / total_SS = 89.7 %)
```

Available components:

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

The `kmeans()` function return a “list” with 9 components. You can see the named components of any list with the `attributes()` function.

```
attributes(km)
```

\$names

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

\$class

```
[1] "kmeans"
```

Q. How many points are in each cluster?

```
km$size
```

```
[1] 30 30
```

Q. Cluster assignment/membership vector?

```
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 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

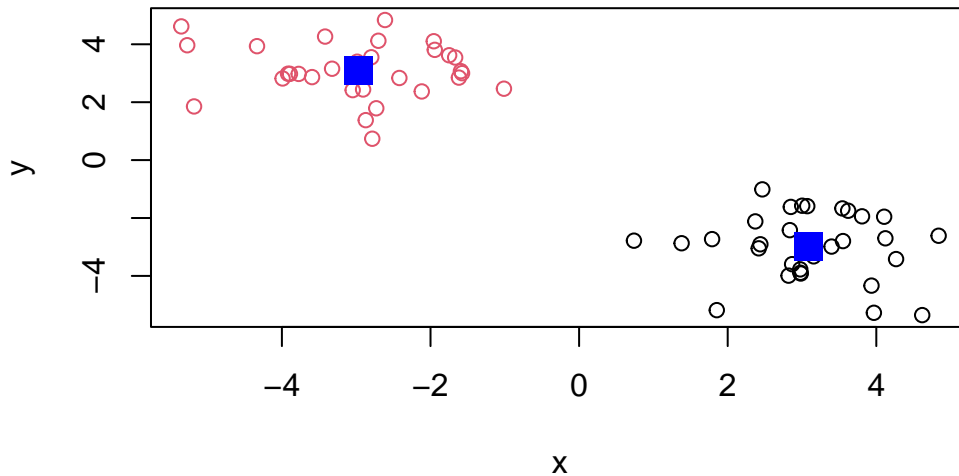
Q. Cluster centers?

```
km$centers
```

	x	y
1	3.092391	-2.971414
2	-2.971414	3.092391

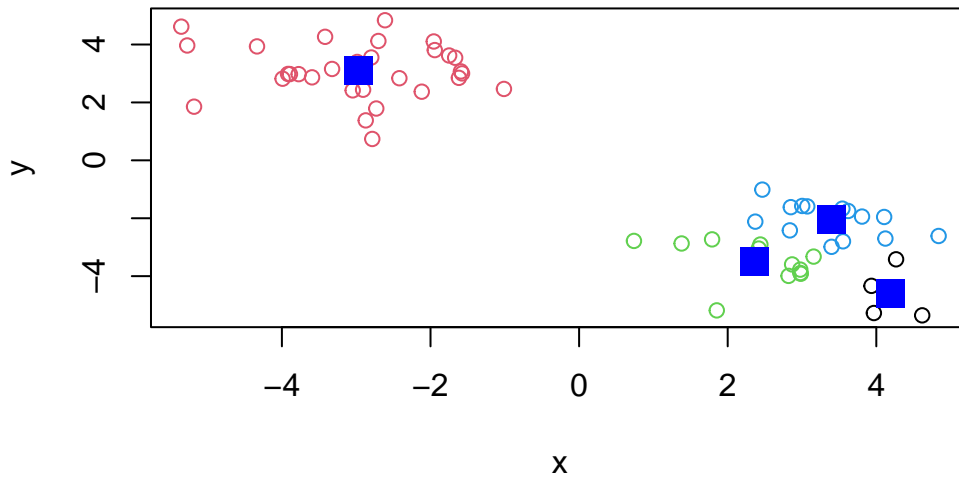
Q. Make a plot of our `kmeans()` results showing cluster assignment using different colors for each cluster/group of points and cluster centers in blue.

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



Q. Run `kmeans()` again on `x` and this cluster into 4 groups/clusters and plot the same result figure as above.

```
km4 <- kmeans(x, centers = 4)
plot(x, col= km4$cluster)
points(km4$centers, col="blue",pch=15, cex=2)
```



**key-point:** K-means clustering is super popular but can be miss-used. One big limitation is that it can impose a clustering pattern on your data even if clear natural grouping don't exist - i.e. it does what you tell it to do in terms of **centers**.

## Hierarchical Clustering

The main function in “base” R for Hierarchical Clustering is called `hclust()`.

You can't just pass our dataset as is into `hcluster()` you must give “distance matrix” as input. We can get this from the `dist()` function in R.

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

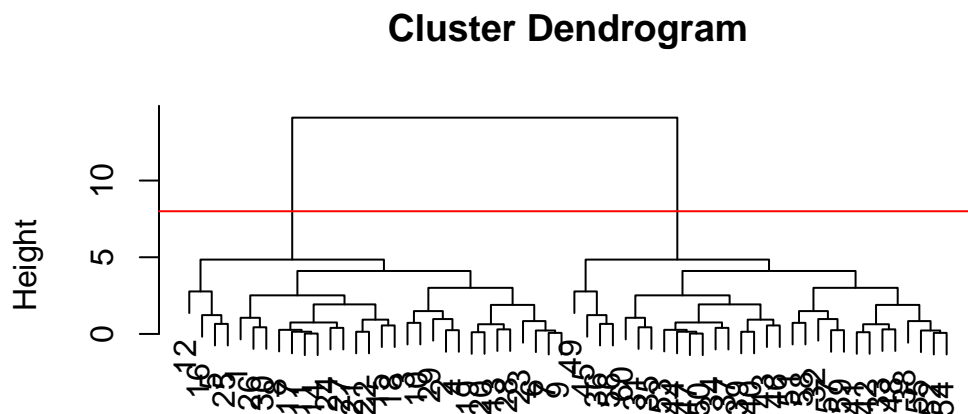
Call:

```
hclust(d = d)
```

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

The results of `hclust()` don't have a useful `print()` method but do have a special `plot()` method.

```
plot(hc)
abline(h=8, col="red")
```



```
hclust (*, "complete")
```

To get our main cluster assignment (membership vector) we need to “cut” the tree at the big goal posts...

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

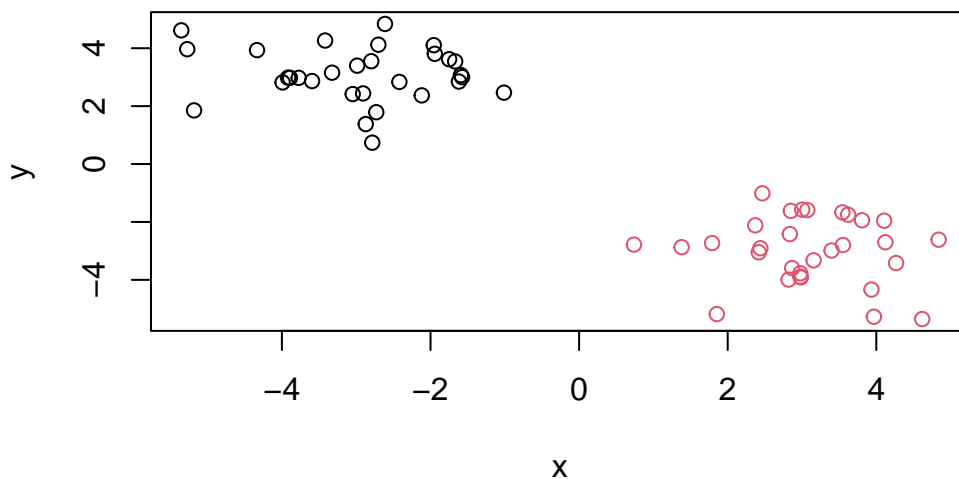
[illegible]

```
table(grps)
```

```
grps
  1  2
30 30
```



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



Hierarchical clustering is distinct in that the dendrogram (tree figure) can reveal the potential grouping in your data (unlike K-means)

## Principal Component Analysis (PCA)

PCA is a common and highly useful dimensionality reduction technique used in many fields - particularly bioinformatics

Here we will analyze some data from the UK on food consumption.

### Data import

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

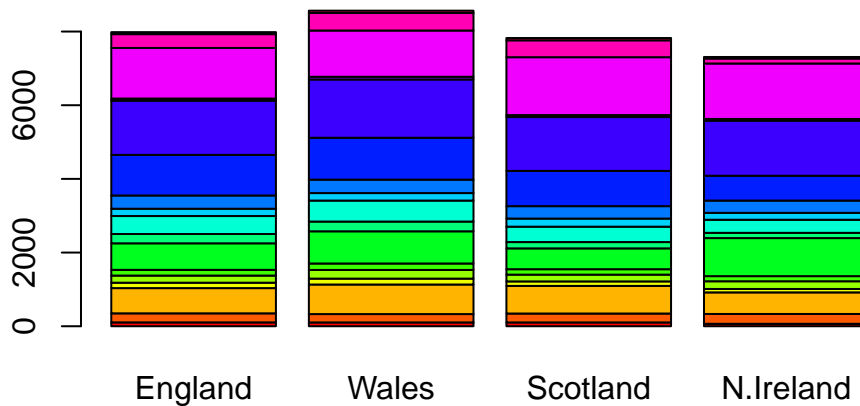
```
rownames(x) <- x[,1]  
x <- x[,1]  
head(x)
```

```
[1] "Cheese"          "Carcass_meat " "Other_meat "    "Fish"
[5] "Fats_and_oils " "Sugars"
```

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

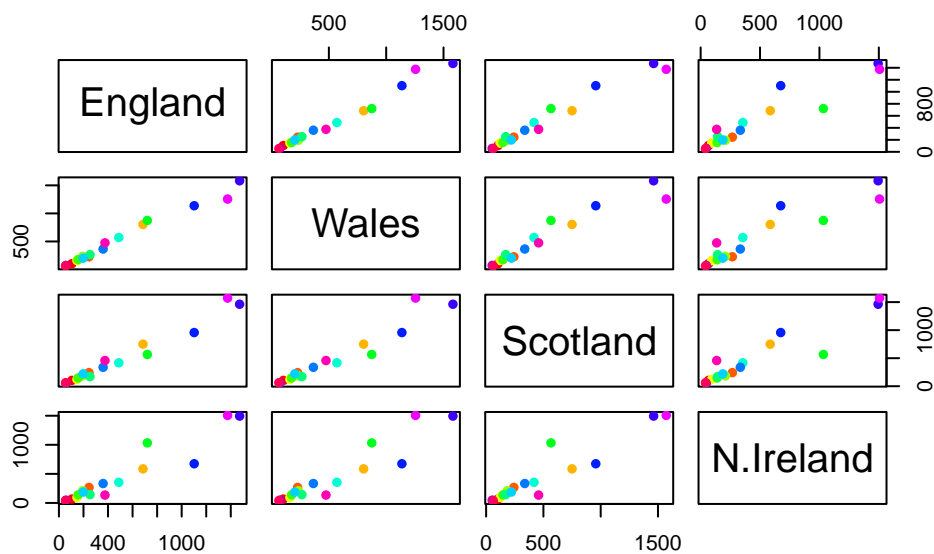
	England	Wales	Scotland	N.Ireland
Cheese	105	103	103	66
Carcass_meat	245	227	242	267
Other_meat	685	803	750	586
Fish	147	160	122	93
Fats_and_oils	193	235	184	209
Sugars	156	175	147	139

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



One conventional plot that can be useful is called a “pairs” plot.

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



## PCA to the rescue

The main function in base R for PCA is called `prcomp()`

```
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

The `prcomp()` function returns a list object of our results with five attributes/components

```
attributes(pca)
```

```
$names
```

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

```
$class
```

```
[1] "prcomp"
```

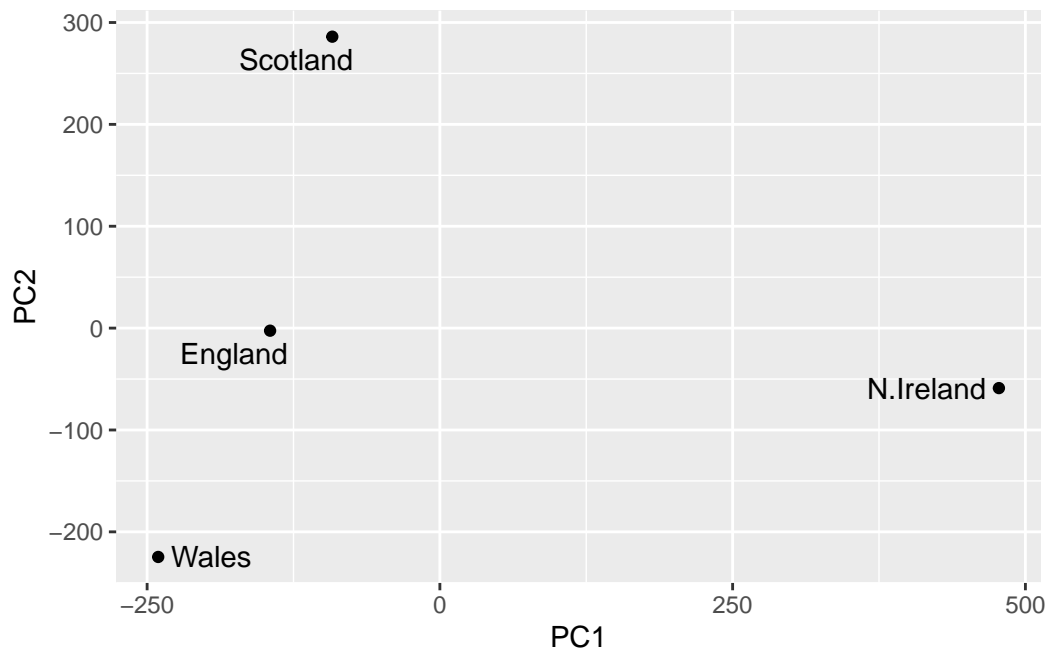
The two main “results” in here are `pca$x` and `pca$rotation`. The first of these (`pca$x`) contains the scores of the data on the new PC axis - we use these to make our “PCA plot”.

```
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

```
library(ggplot2)
library(ggrepel)

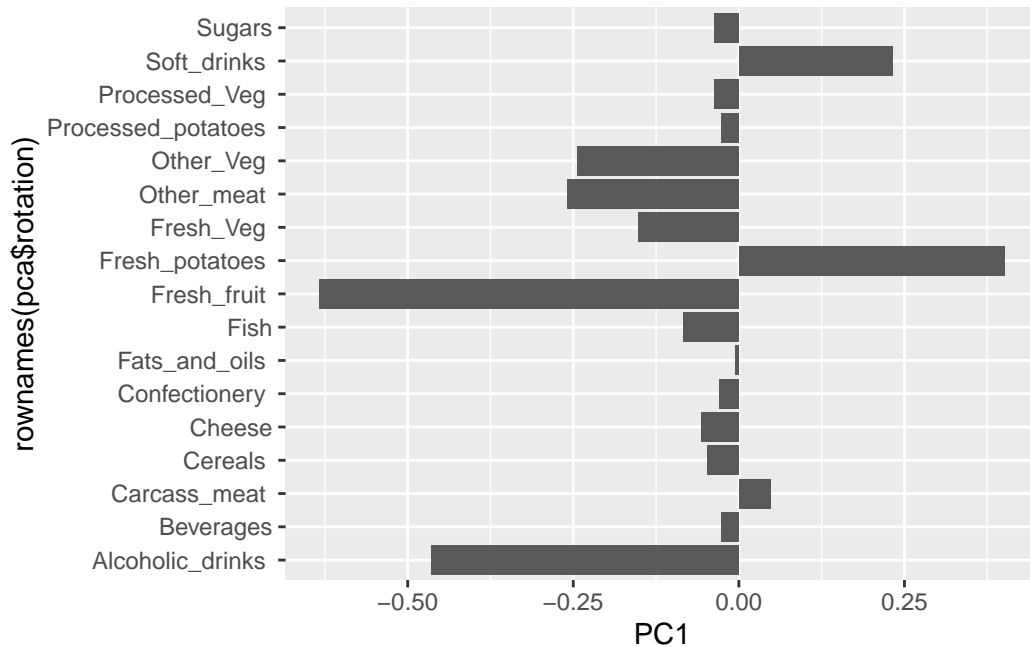
# Make a plot of pca$x with PC1 vs PC2
ggplot(pca$x) +
  aes(PC1, PC2, label = rownames(pca$x)) +
  geom_point() +
  geom_text_repel()
```



The plot maps the relationship between the countries and the major principal components reducing the variation in the data to a two dimensional plot.

The second major result is contained in the `pca$rotation` object or component. Let's plot this to see what PCA is picking up...

```
ggplot(pca$rotation) +  
  aes(PC1, rownames(pca$rotation)) +  
  geom_col()
```



### Interpreting PCA results

The plot above shows food groups with positive and negative loading scores which push the countries to the left or right side of the plot. Observing the two plots above, N. Ireland consumes more fresh potatoes than the other UK countries, while fresh fruit and alcoholic drinks are more consumed by Wales, England, and Scotland than N. Ireland.