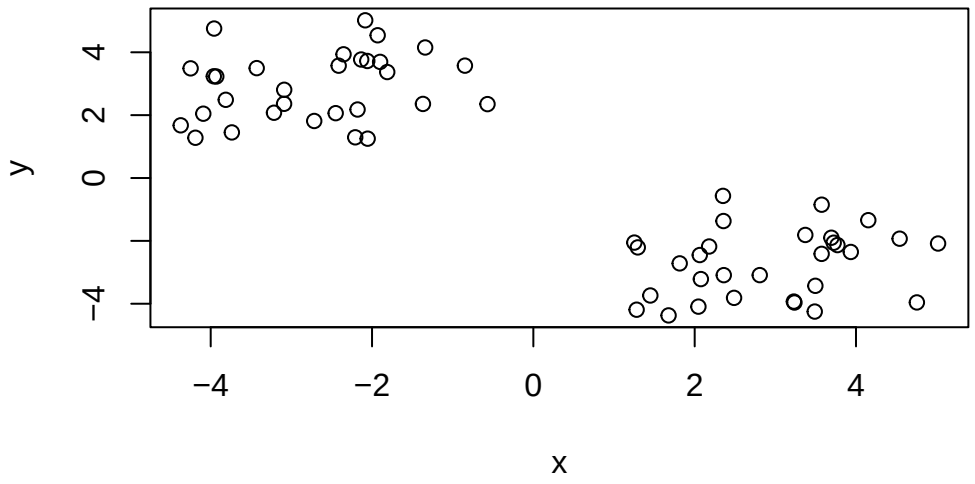
# Class7

Hetian Su

## K-means clustering

create a custom data to cluster

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




Try the kmeans clustering function in base R

```r
km <- kmeans(x, centers = 2, nstart = 30)
km
```

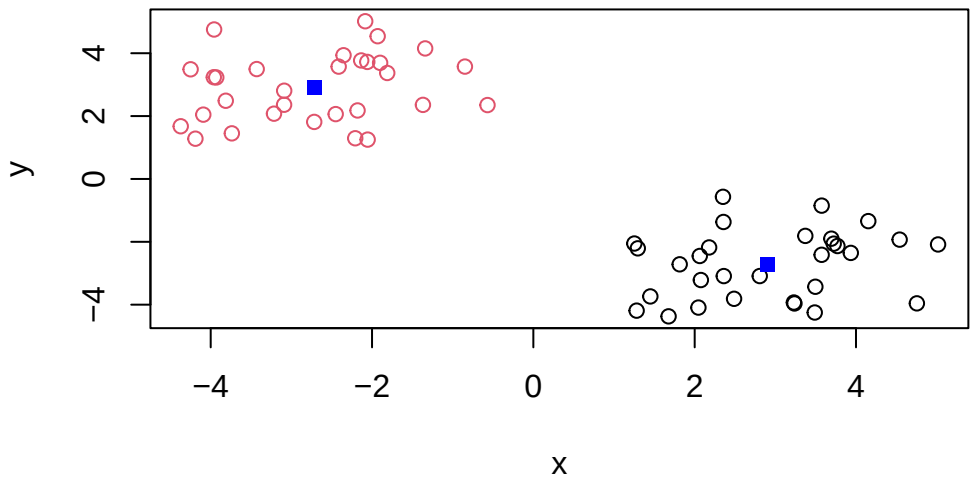
```
K-means clustering with 2 clusters of sizes 30, 30

Cluster means:
          x         y
1  2.901926 -2.718307
2 -2.718307  2.901926

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] 67.1324 67.1324
 (between_SS / total_SS =  87.6 %)

Available components:

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

## Q

What are the components of the clustering that give:

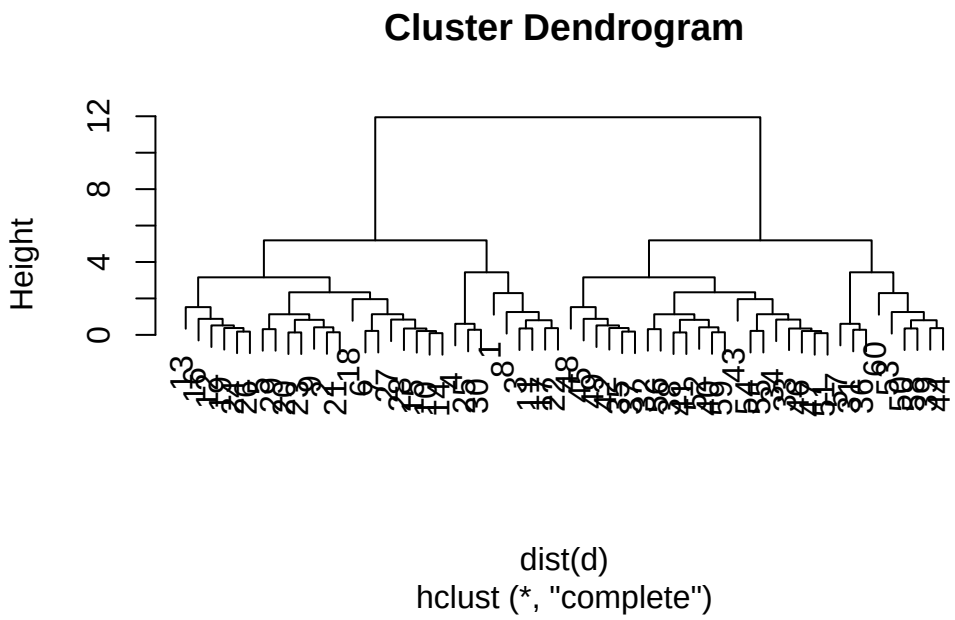
- cluster center: “centers”
- cluster size: “size”
- plot the centers as blue points on top of the scatter plot where the clusters are colored differently

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




## Hierarchical Clustering

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

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




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

## UK Food Data

### Q1

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

#inspect number of rows and columns
dim(x)
```

```
[1] 17  5
```

inspect the data to make sure the dataframe is correctly imported

```
# view(x)
head(x, 6)
```

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

Rename the row names to sample/food names

```
rownames(x) <- x[,1]
x <- x[,-1]
head(x,6)
```

```
              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
```

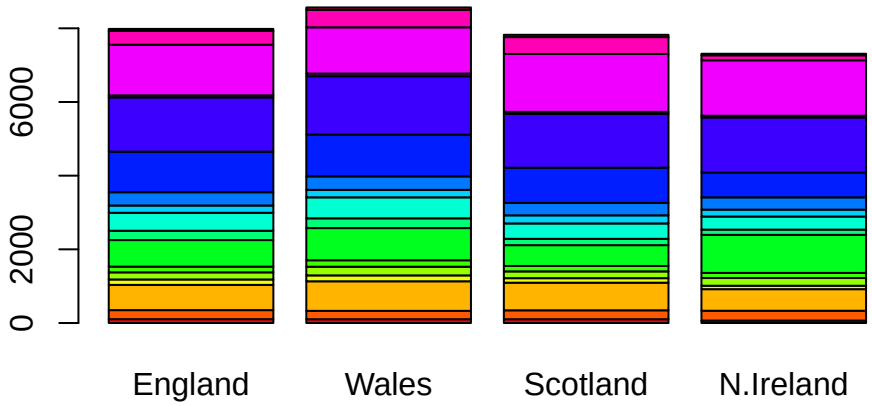
## Q2

In this case where the structure of the dataframe is known to us, we should use the setting row names while importing method to be concise with codes. However, in cases where we do not know the structure of the dataframe before inspecting it, the first method is more robust.

## Q3

Change the histogram to be stacked

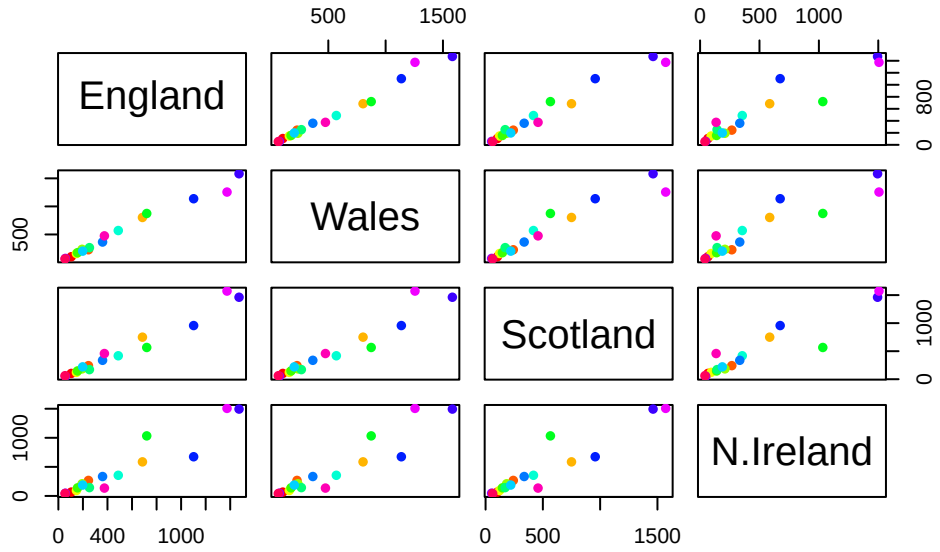
```
barplot(as.matrix(x), beside = F, col = rainbow(nrow(x)))
```




## Q5

create and interpret a pairwise plot of the dataset

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




Each pairwise scatter plot has x axis representing one region and y axis representing another region. A diagonal line indicates that the data points have similar values in both regions.

## Q6

As compared to all other 3 regions, N.Ireland has more data points that are off the diagonal line. The most evident ones are the points colored in blue, orange and cyan.

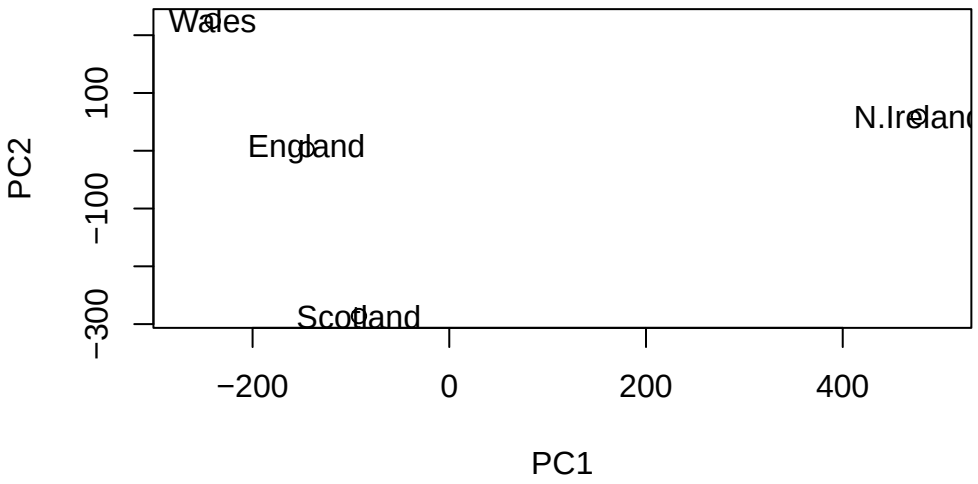
## Q7

Use base R PCA function to analyze the data.

```
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
```

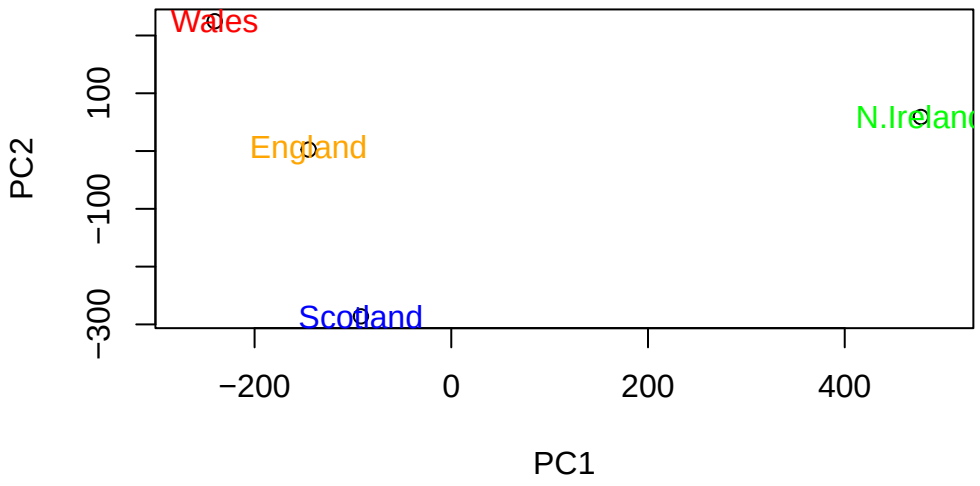
```
#plot PC1 va PC2
plot(pca$x[,1], pca$x[,2], xlab='PC1', ylab='PC2', xlim=c(-270, 500))
text(pca$x[,1], pca$x[,2], colnames(x))
```




## Q8

Change the color of the text to match the colors on the map

```
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','green'))
```




We can also retrieve the properties of the PCA from pca data

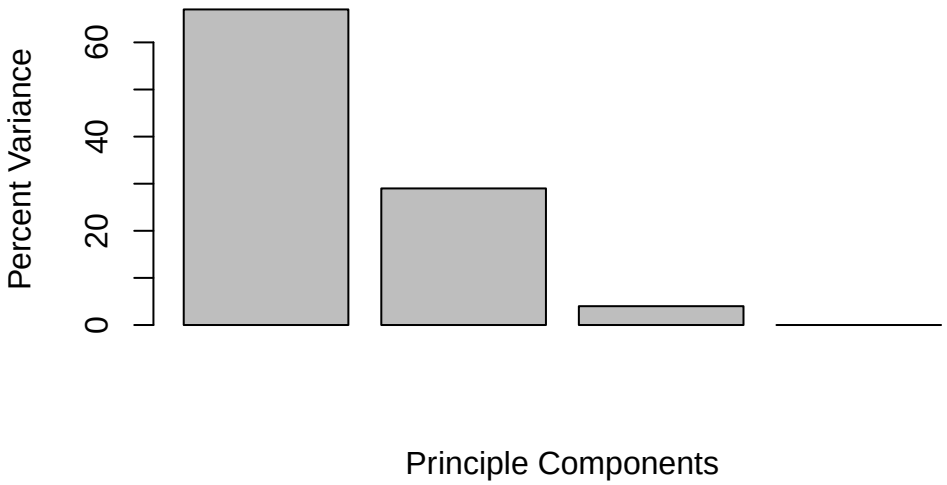
```
# retrieve the standard deviations
sd <- round(pca$sdev^2/sum(pca$sdev^2)*100)
sd
```

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

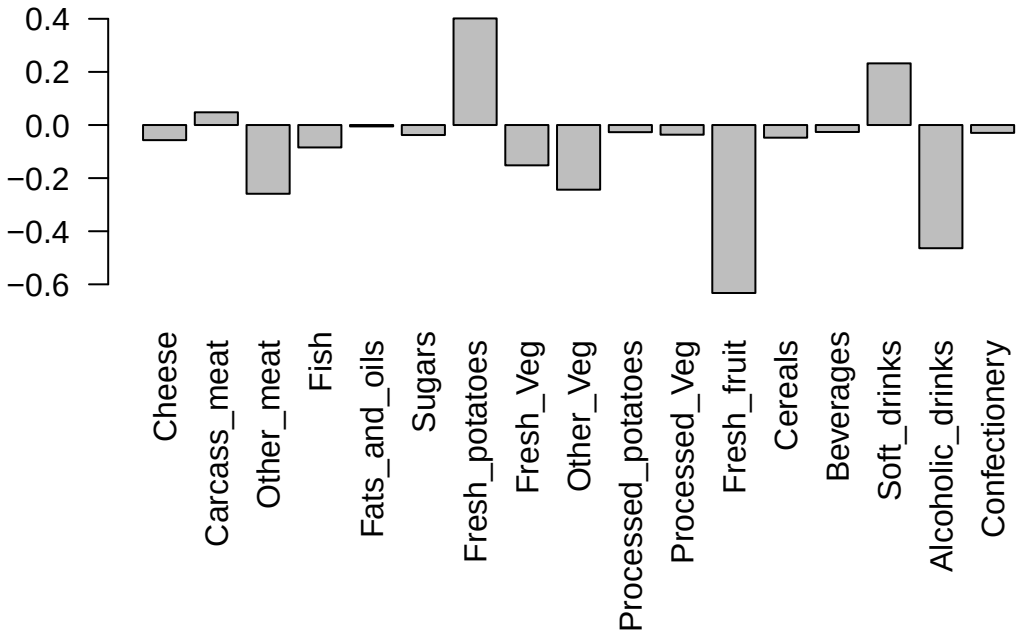
```
z <- summary(pca)
z$importance
```

```
                             PC1       PC2      PC3           PC4
Standard deviation     324.15019 212.74780 73.87622 4.188568e-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(sd, xlab = 'Principle Components', ylab = 'Percent Variance')
```

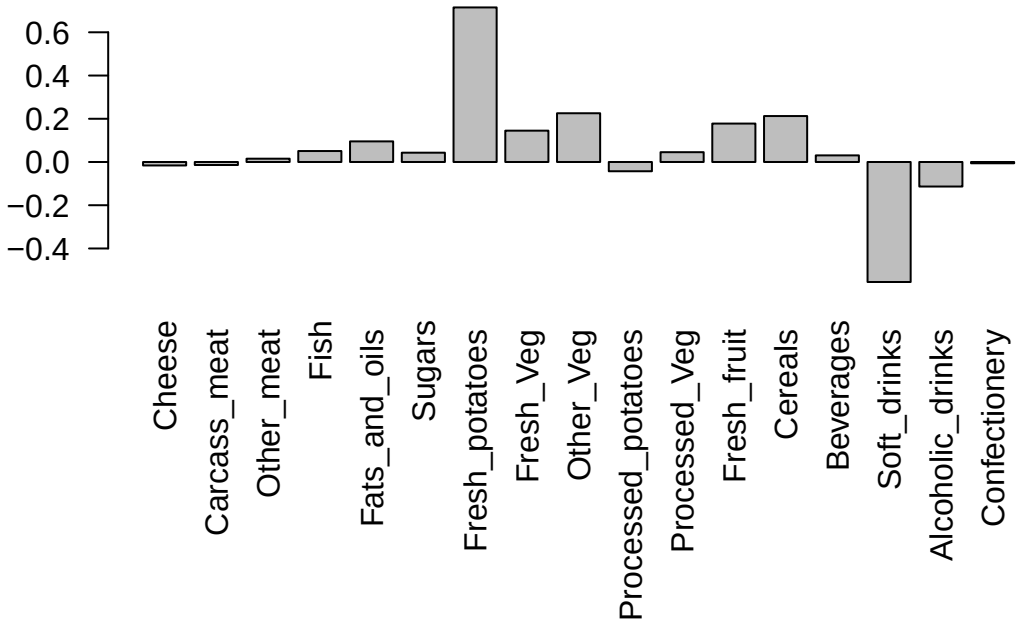



```
# retrieve the loading scores of each varianle on PC1
par(mar=c(10, 3, 0.35, 0))
barplot(pca$rotation[,1], las=2, ylab = 'PC1')
```




**Q9**

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




PC2 loading scores show that Fresh_potatoes most strongly pushes Wales up the axis from England and N.Ireland, while Soft_drinks most strongly pushes Scotland down the axis from England and N.Ireland.