Class 7

We will import the `UK_foods.csv` dataset

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

> Q1. How many rows and columns are in your new data frame named x? What R functions could you use to answer this questions?

```
dim(x)
```

```
[1] 17  5
```

We will want to check that the data has imported correctly

```
head(x)
```

```
              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
```

It seems that the names for the rows was incorrectly counted as a column, let's fix that

```
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
```

Or alternatively:

```
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

We see that the problem seems to have been resolved, let's check with `dim()`

```
dim(x)
```

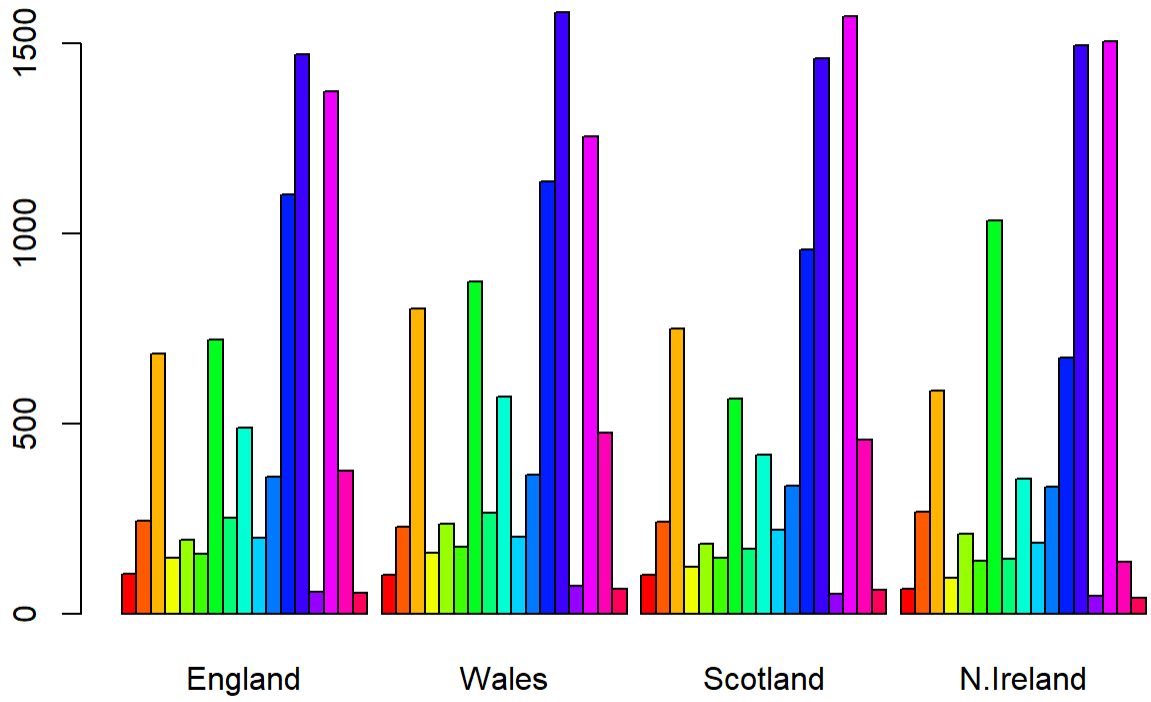
```
[1] 17  4
```

> 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?

The second approach is better because it applies the change to the entire dataset without having to redefine x, which may become a problem if there are other aspects of the imported dataset that you would like to change

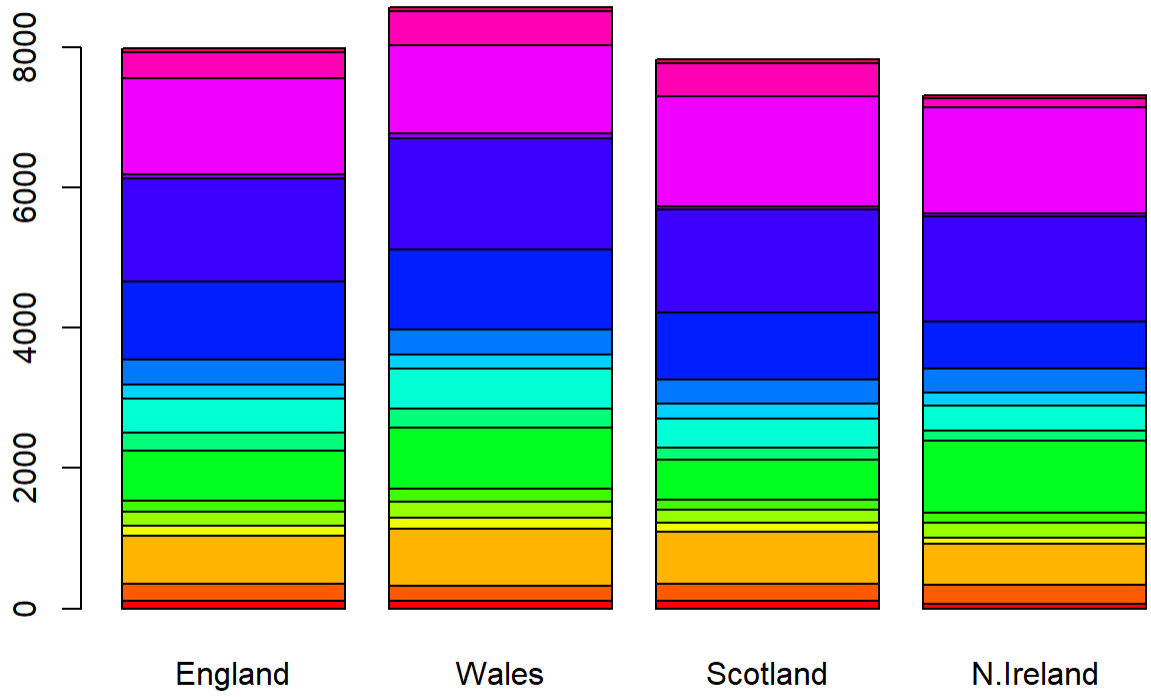
As a table this data does not show us much. Let's graph it:

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



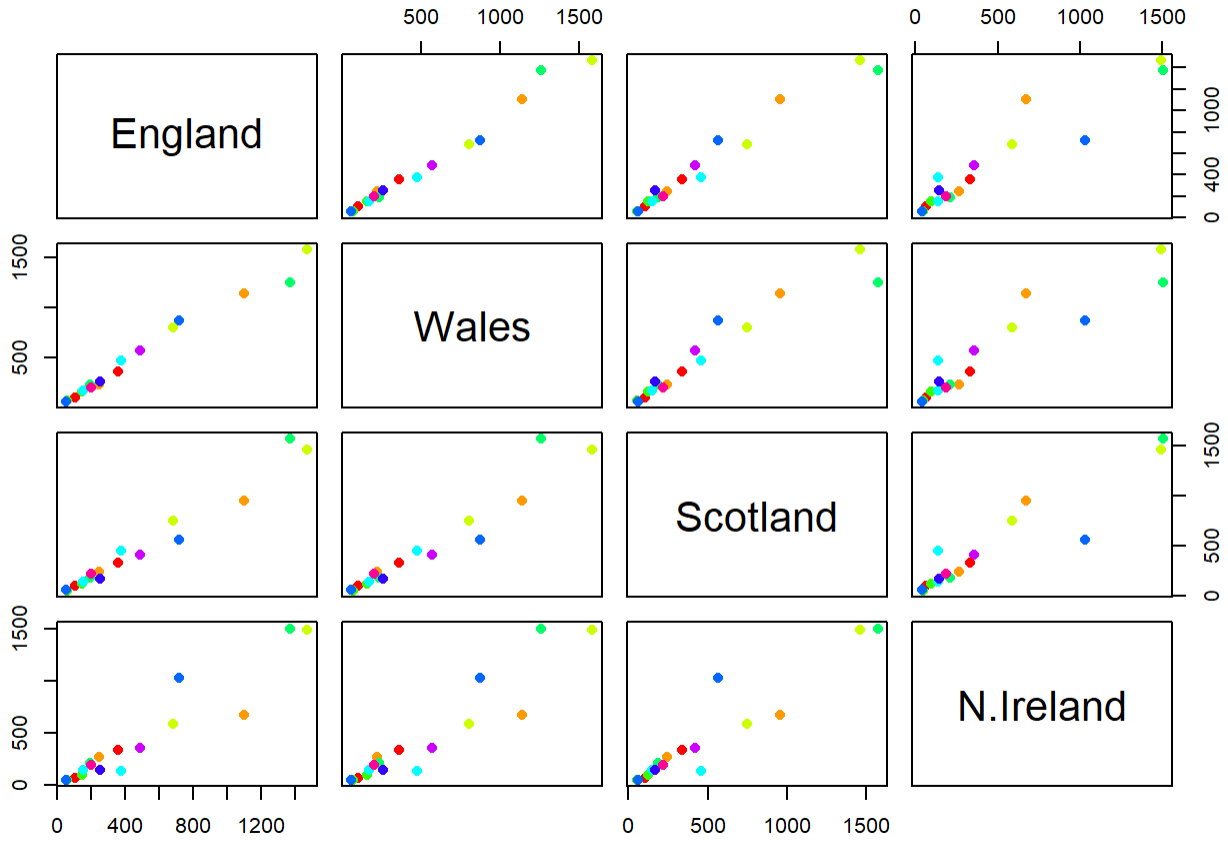
> Q3: Changing what optional argument in the above barplot() function results in a stacked barplot?

```
#the beside portion is shown as true, needs to be false to stack the bars
barplot(as.matrix(x), beside=F, 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)
```



The pairs shows graphs of each combination of observations plotted against each other: i.e England against itself, England vs Wales, etc. The closer the dots to the diagonal line the more similarities there are, the further the more differences

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

The first cluster of dots deviate from the diagonal line slightly more than the other plots, it is hard to really say what this means, or even see the differences clearly in the first place b/c we are forced to visually compare with 9 other graphs

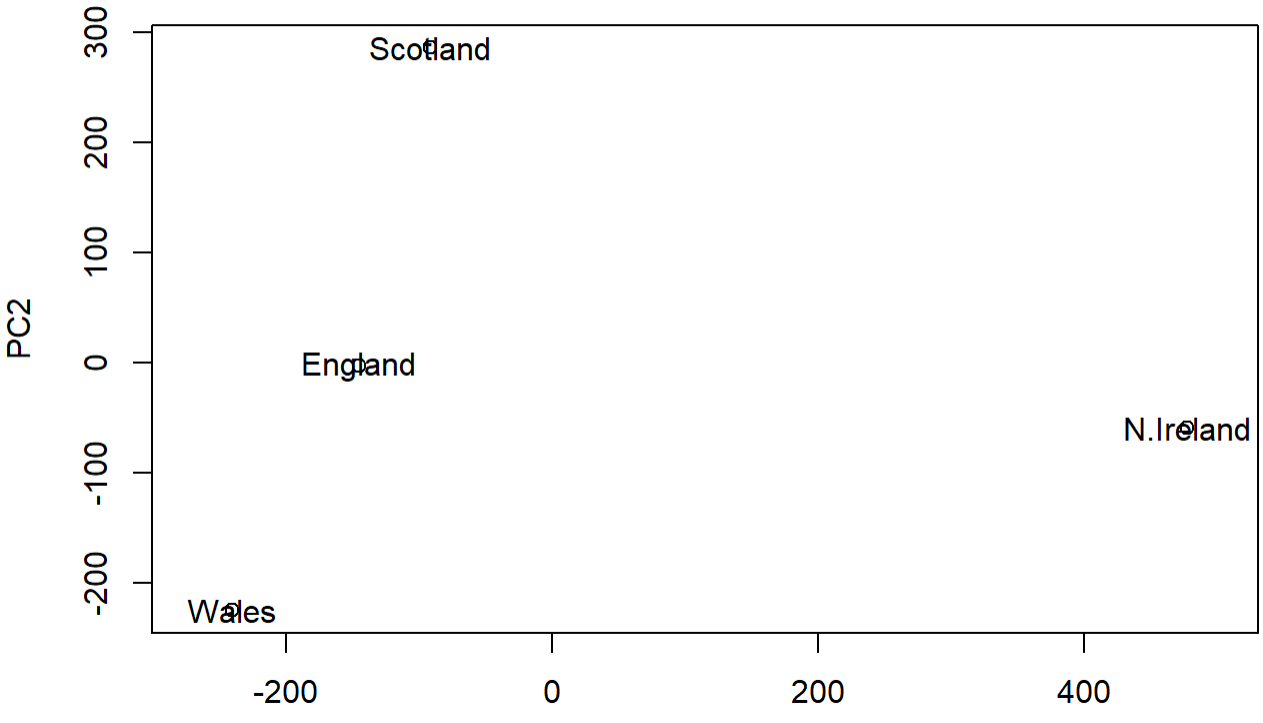
First, the data needs to be transposed to work with PCA

```
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
```

> 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))
```

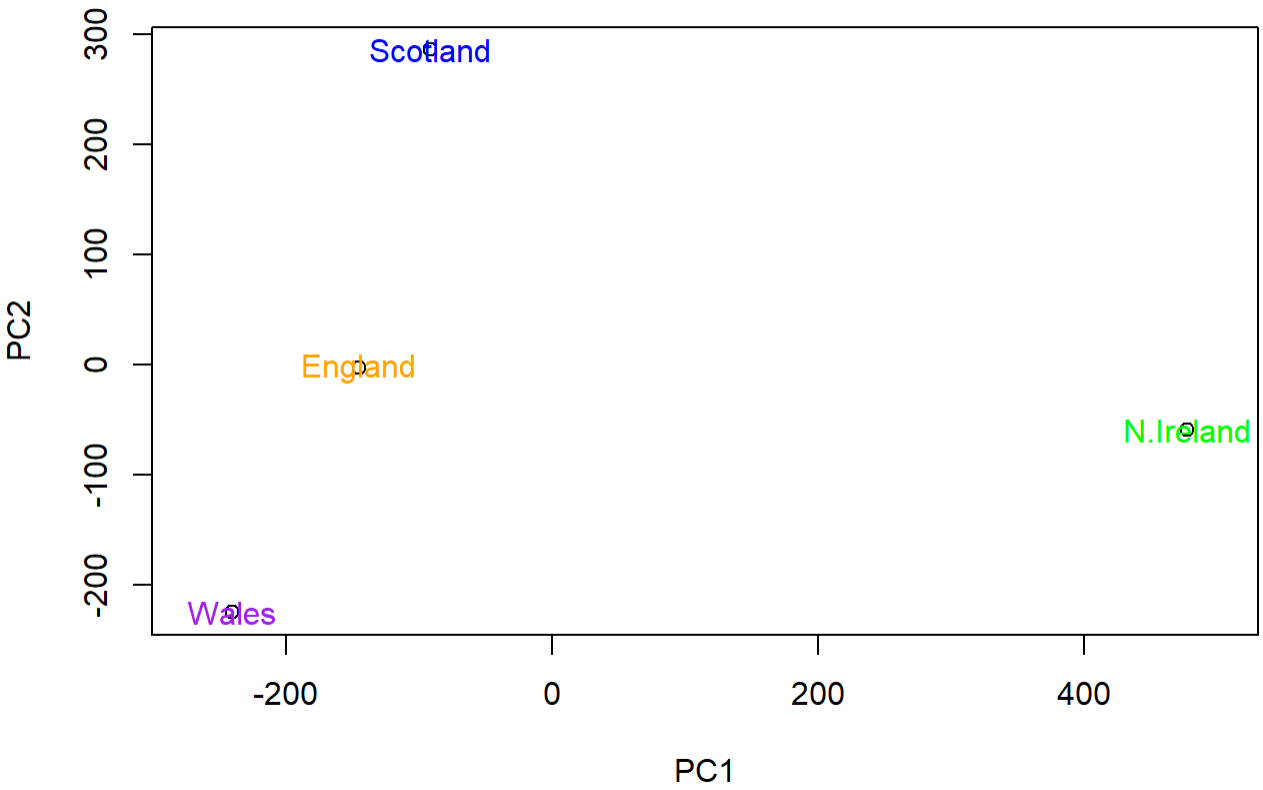


PC1

> 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.

```
my_colors <- c("orange", "purple", "blue", "green")

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 = my_colors)
```



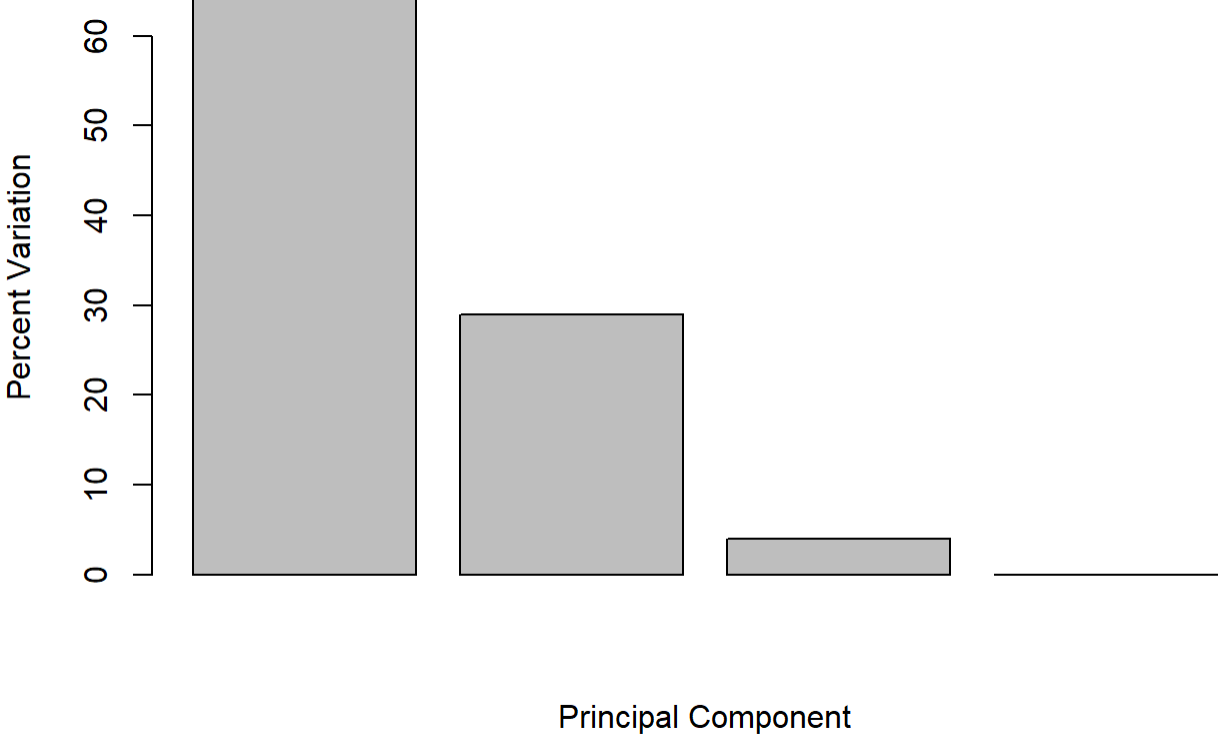
To figure out the variance in the data for each PC we need the standard deviation:

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

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

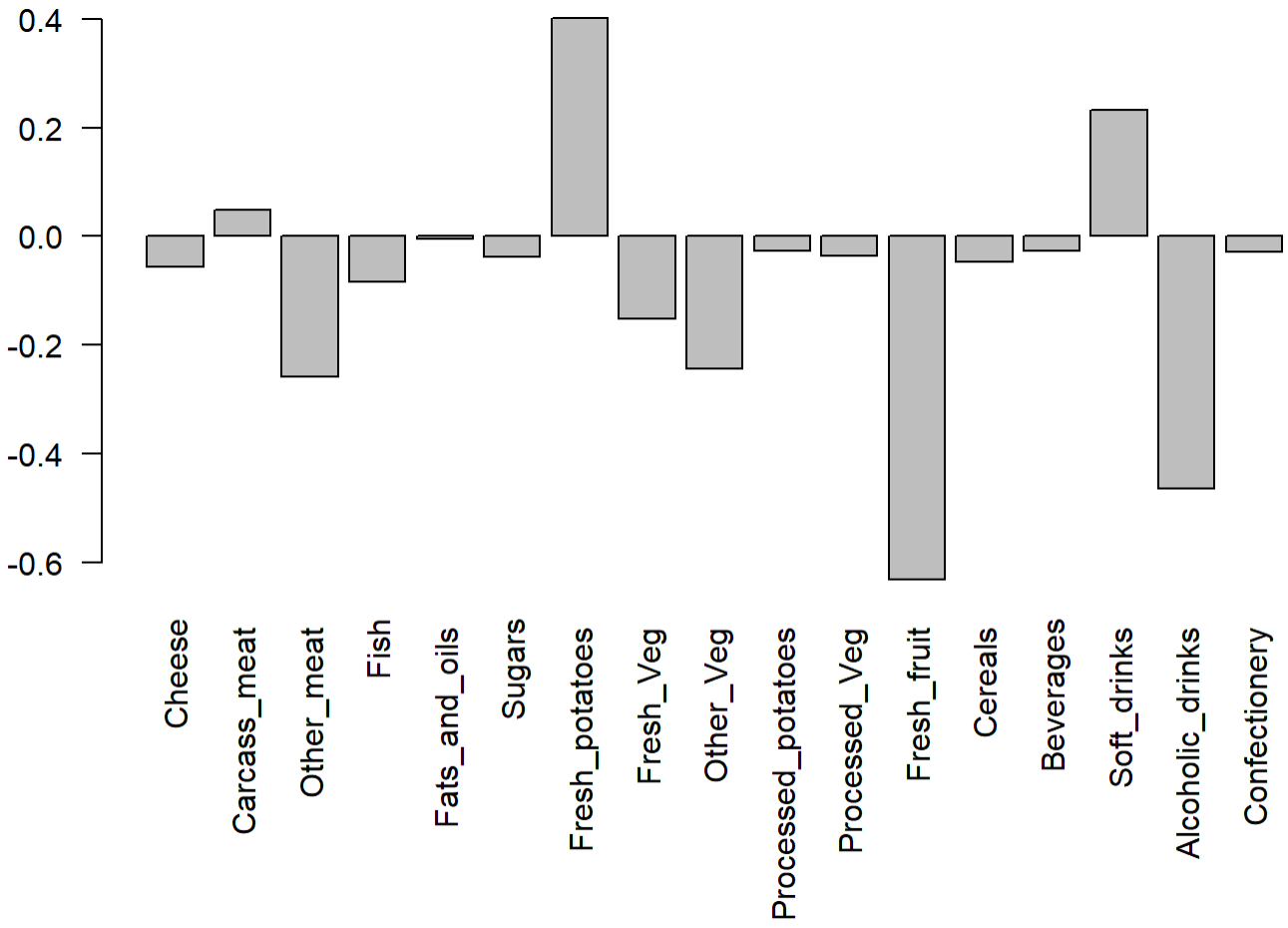
Now to plot the variances

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



To see the influence of each variable on the PCs, aka the loading scores, we use the `$rotation` component:

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

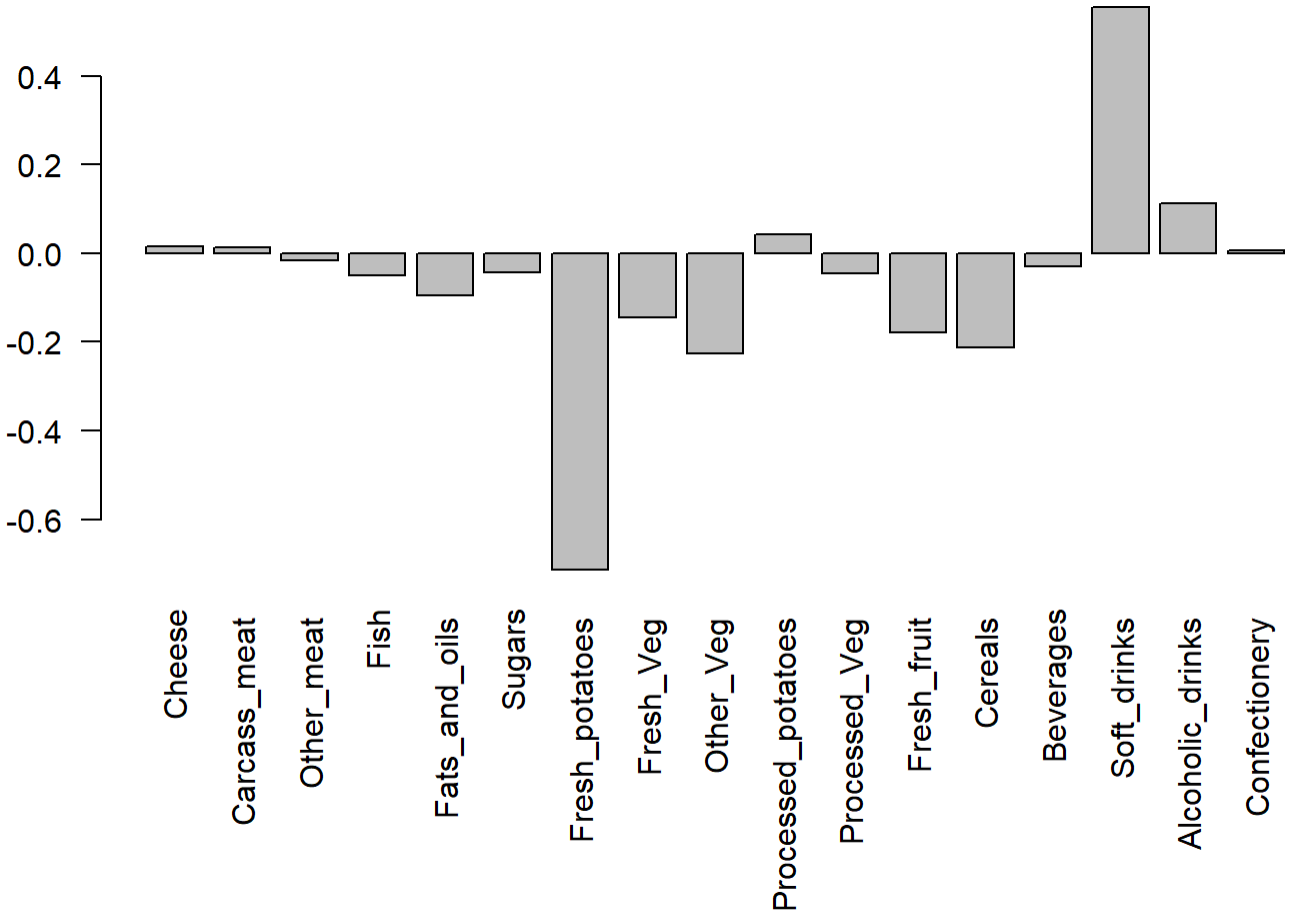


The highest positive values are what cause skewing to the right in the PC1 vs PC2 plot, which is potatoes and soft drinks here. The lowest negativee values are what cause skewing to the right in the PC1 vs PC2 plot, which is

fresh fruit and alcoholic drinks

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

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

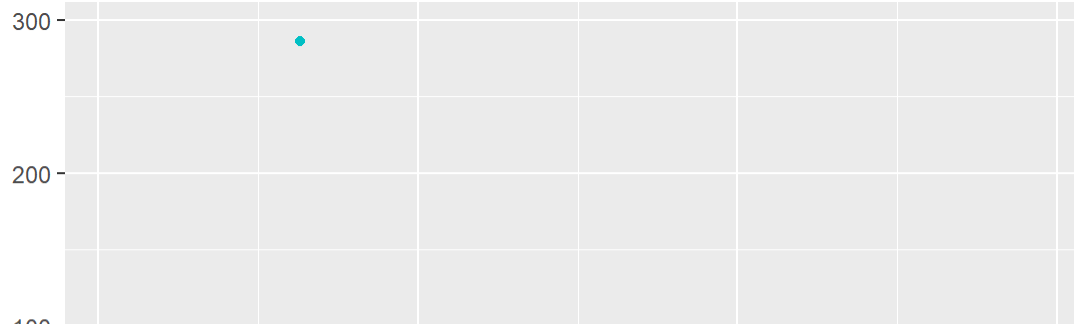


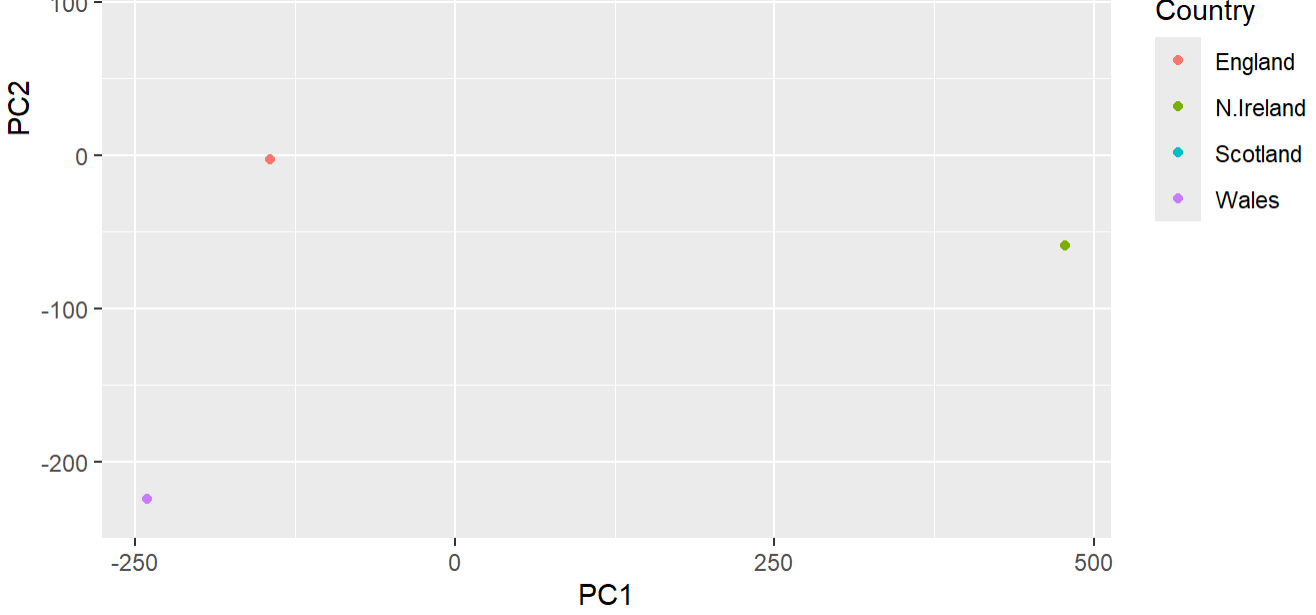
Now we will try to plot the same thing on ggplot2

```
library(ggplot2)

df <- as.data.frame(pca$x)
df_lab <- tibble::rownames_to_column(df, "Country")

ggplot(df_lab) +
  aes(PC1, PC2, col=Country) +
  geom_point()
```

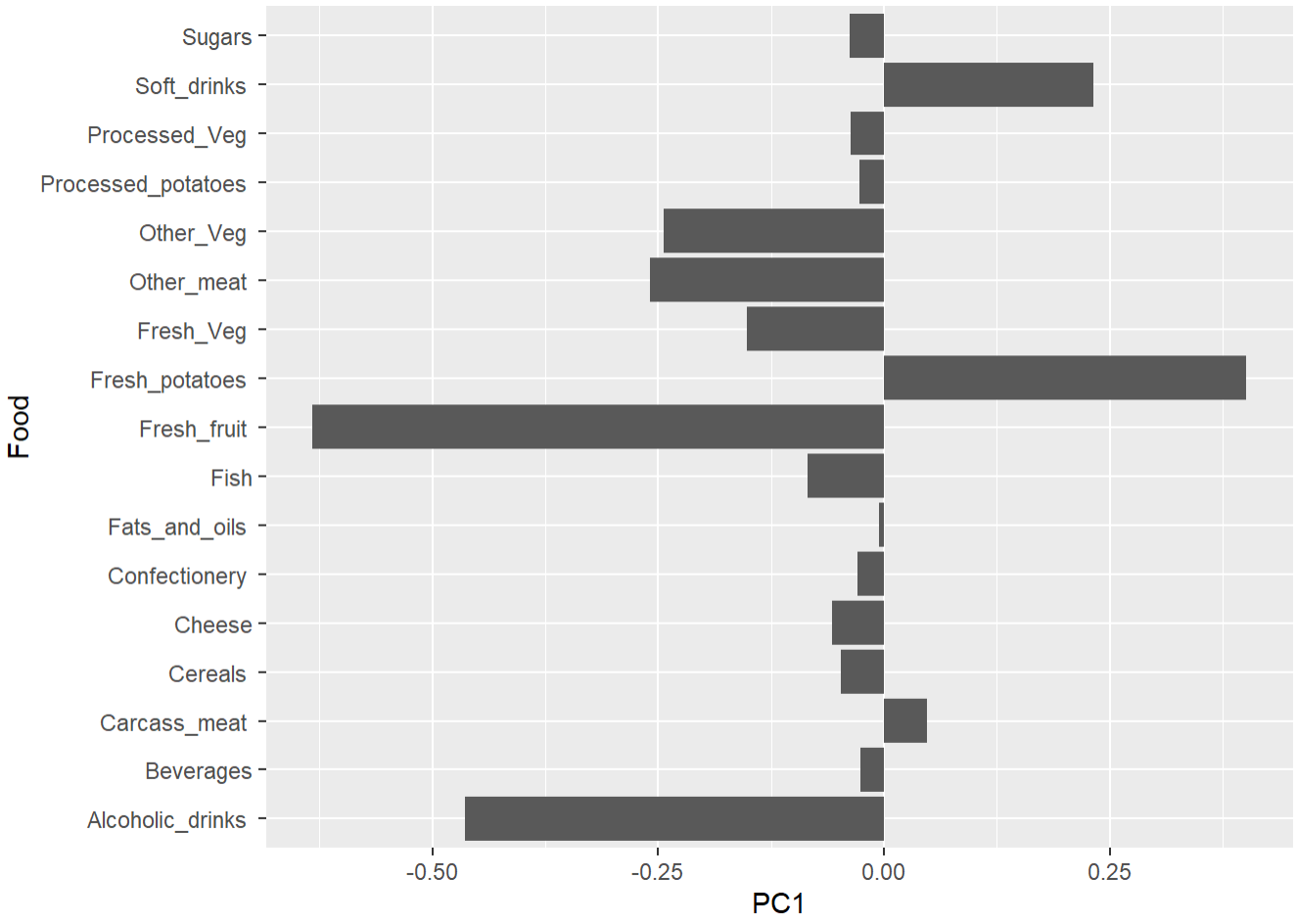




Now to plot the loading plot:

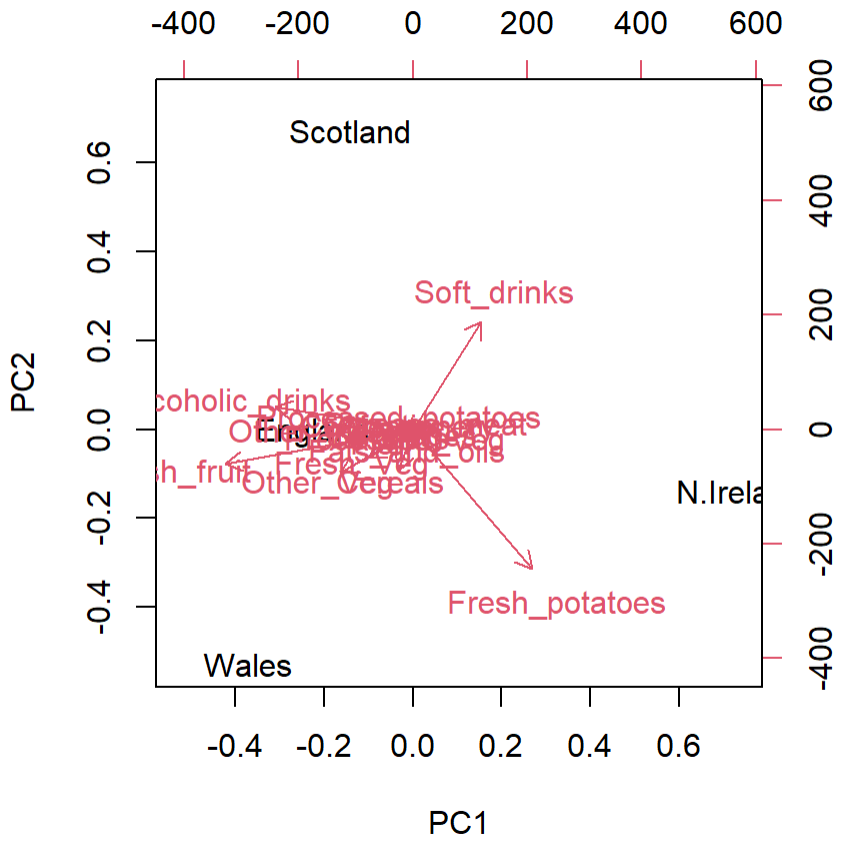
```
ld <- as.data.frame(pca$rotation)
ld_lab <- tibble::rownames_to_column(ld, "Food")

ggplot(ld_lab) +
  aes(PC1, Food) +
  geom_col()
```



Another way to see this information together with the main PCA plot is in a so-called biplot:

```
biplot(pca)
```



PCA of RNA seq data

```
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
```

```
#Again we have to take the transpose of our data
pca <- prcomp(t(rna.data), scale=TRUE)
```

```
summary(pca)
```

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
Standard deviation     9.6237 1.5198 1.05787 1.05203 0.88062 0.82545 0.80111
Proportion of Variance 0.9262 0.0231 0.01119 0.01107 0.00775 0.00681 0.00642
Cumulative Proportion  0.9262 0.9493 0.96045 0.97152 0.97928 0.98609 0.99251
```

```
                            PC8     PC9      PC10
Standard deviation     0.62065 0.60342 3.457e-15
Proportion of Variance 0.00385 0.00364 0.000e+00
Cumulative Proportion  0.99636 1.00000 1.000e+00
```

> Q How many genes in this dataset?

```
nrow(rna.data)
```

```
[1] 100
```

```
head(pca$x)
```

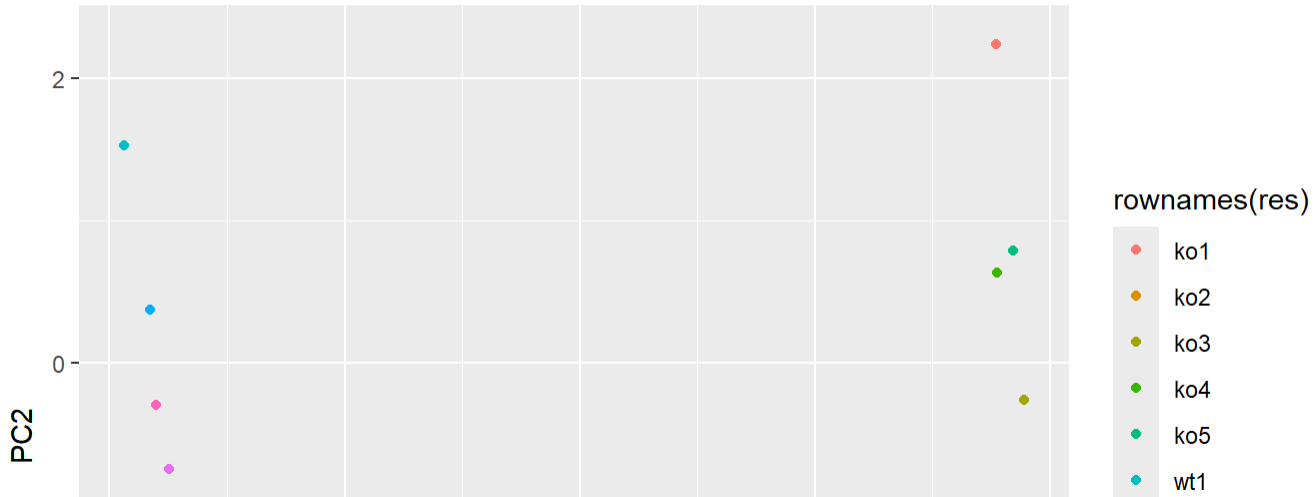
```
          PC1        PC2        PC3        PC4        PC5        PC6
wt1 -9.697374  1.5233313 -0.2753567  0.7322391 -0.6749398  1.1823860
wt2 -9.138950  0.3748504  1.0867958 -1.9461655  0.7571209 -0.4369228
wt3 -9.054263 -0.9855163  0.4152966  1.4166028  0.5835918  0.6937236
wt4 -8.731483 -0.7468371  0.5875748  0.2268129 -1.5404775 -1.2723618
wt5 -9.006312 -0.2945307 -1.8498101 -0.4303812  0.8666124 -0.2496025
ko1  8.846999  2.2345475 -0.1462750 -1.1544333 -0.6947862  0.7128021
            PC7         PC8         PC9         PC10
wt1 -0.24446614  1.03519396  0.07010231 3.073930e-15
wt2 -0.03275370  0.26622249  0.72780448 1.963707e-15
wt3 -0.03578383 -1.05851494  0.52979799 2.893519e-15
wt4 -0.52795595 -0.20995085 -0.50325679 2.872702e-15
wt5  0.83227047 -0.05891489 -0.81258430 1.693090e-15
ko1 -0.07864392 -0.94652648 -0.24613776 4.052314e-15
```

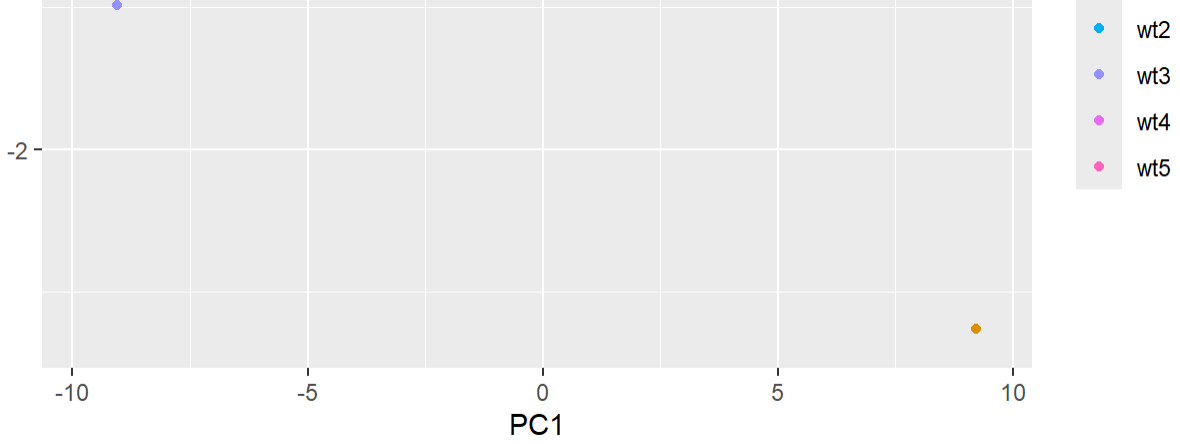
i will make a main result figure use ggplot:

```
library(ggplot2)
```

```
res <- as.data.frame(pca$x)
```

```
ggplot(res) +
  aes(x=PC1, y=PC2, col=rownames(res))+
  geom_point()
```






```
colnames(rna.data)
```

```
[1] "wt1" "wt2" "wt3" "wt4" "wt5" "ko1" "ko2" "ko3" "ko4" "ko5"
```