Machine Learning 1

David Alvarez

2023-10-29

Principal Component Analysis (PCA)

PCA of UK food data

```
url <- "https://tinyurl.com/UK-foods"
x <- read.csv(url, row.names=1)
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
Fresh_potatoes          720   874      566      1033
Fresh_Veg               253   265      171       143
Other_Veg               488   570      418       355
Processed_potatoes      198   203      220       187
Processed_Veg           360   365      337       334
Fresh_fruit            1102  1137      957       674
Cereals                1472  1582     1462      1494
Beverages                57    73       53        47
Soft_drinks            1374  1256     1572      1506
Alcoholic_drinks        375   475      458       135
Confectionery            54    64       62        41
```

Q1. How many rows and columns are in the new data frame?

```r
# Number of rows in the data frame
nrow(x)
```

```
[1] 17
```

```r
# Number of columns in the data frame
ncol(x)
```

```
[1] 4
```

> Q2. Which approach do you prefer for the 'row names problem'?

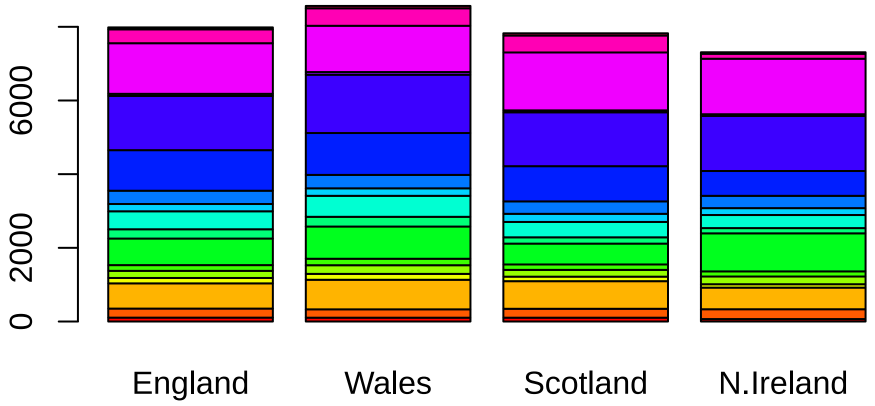
I prefer the option to set the row names equal to 1 since that instantly solves the problem of having to write out a code that would normally take you more time to do and edit.

~~~~
>Q3. Changing what optional argument in the 'barplot()' function results in the plot?

##Changing the color argument to a rainbow color and setting the bars to that color

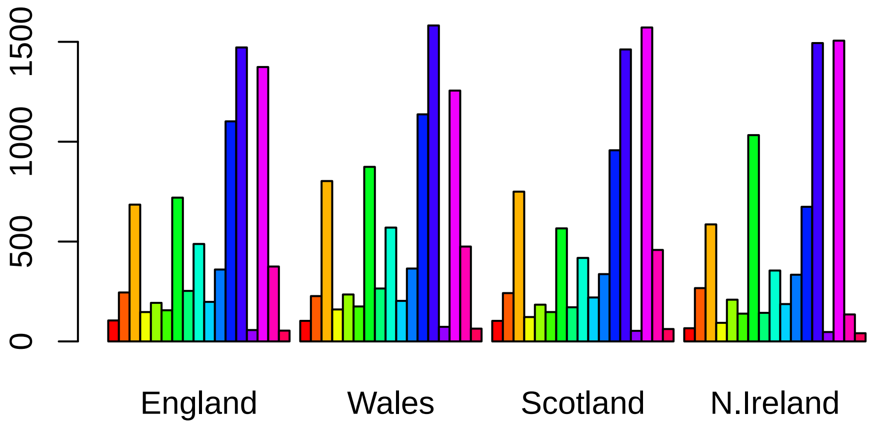
::: {.cell}

```{.r .cell-code}
cols <- rainbow(nrow(x))
barplot(as.matrix(x), col=cols )
~~~~



:::

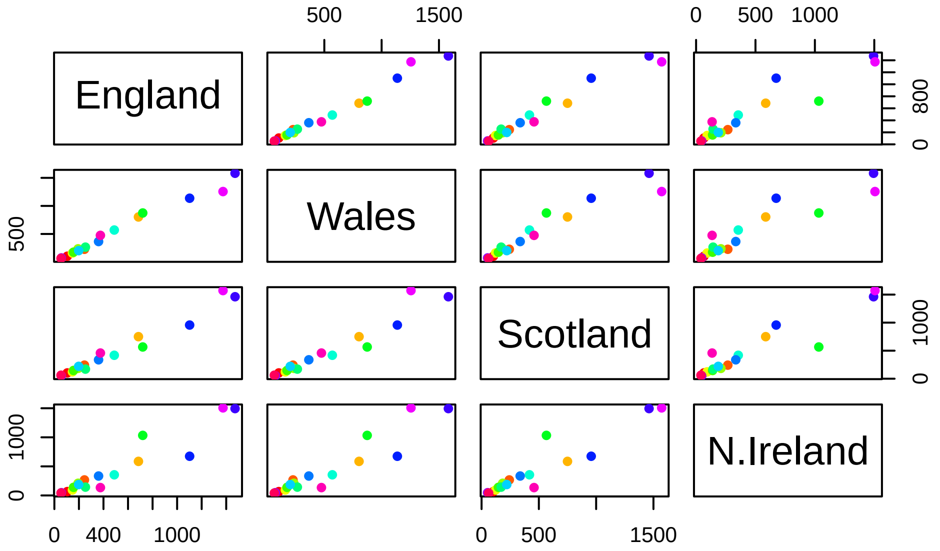
```
barplot(as.matrix(x), col=cols, beside=TRUE )
```



Q5. Can you make sense of the following pairwise plots and what does it mean if a given point lies on the diagonal for a given plot?

##The following plots show the different categories measured compared between two different countries to show how similar or deviant they are from each other. So if a point lies on the diagonal of the plot between the two countries it is most likely very similar in value among the two.

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



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

There are a couple of data values that are shown to be variant among Ireland compared to the other countries that remains consistent which would need more labeling to discover.

~~~~
::: {.cell}

```{.r .cell-code}
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


:::
~~~~

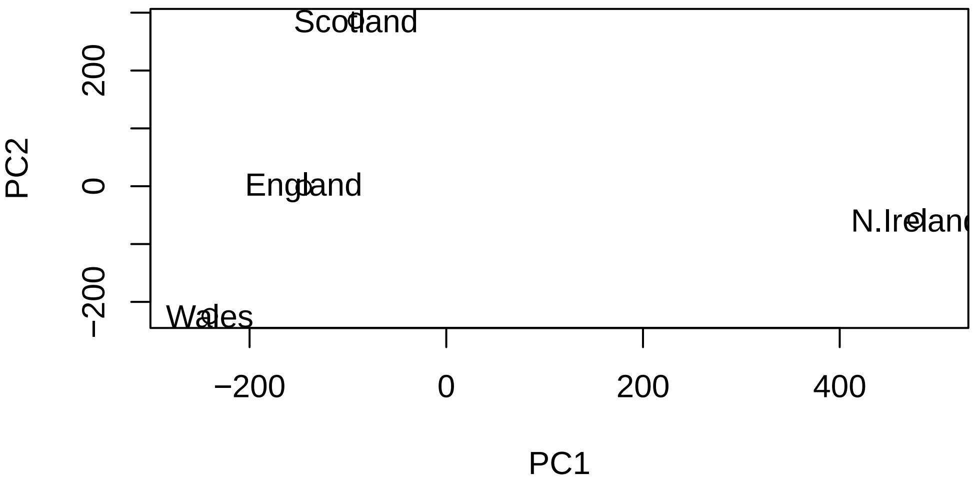
```
attributes(pca)
```

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

$class
[1] "prcomp"
```

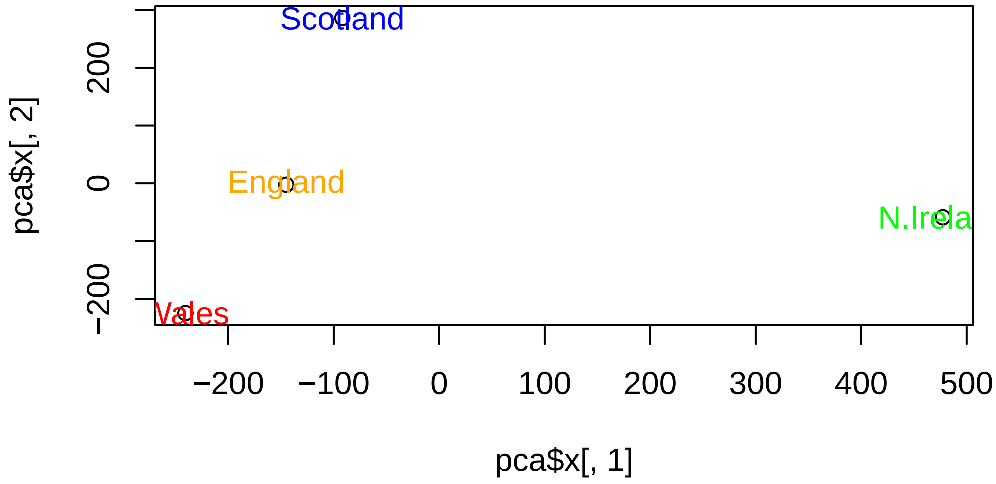
Q7. Complete the code to generate a plot of PC1 vs PC2

```
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. Customise the plot so color of the country colors in the table match.

```
country_cols <- c("orange", "red", "blue", "green")
plot(pca$x[,1], pca$x[,2])
text(pca$x[,1], pca$x[,2], colnames(x), col=country_cols)
```



Defining Variables below

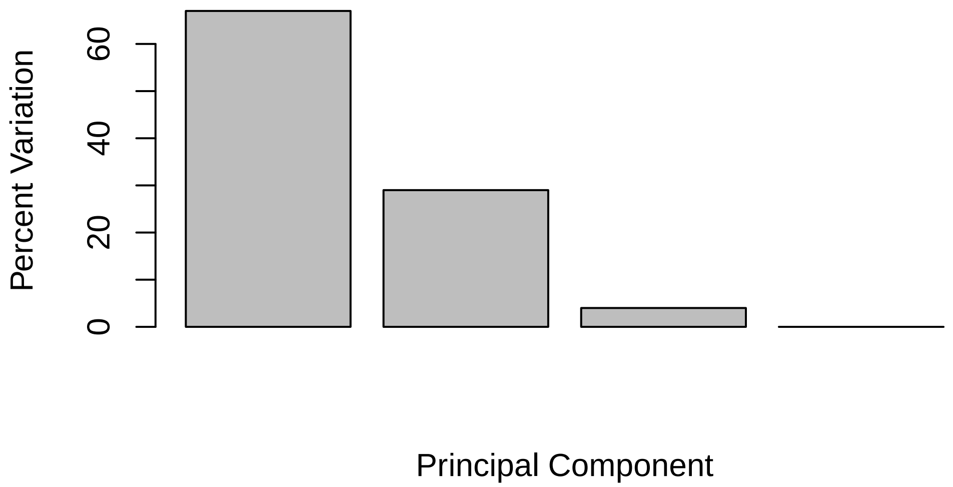
```
v <- round( pca$sdev^2/sum(pca$sdev^2) * 100)
v
```

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

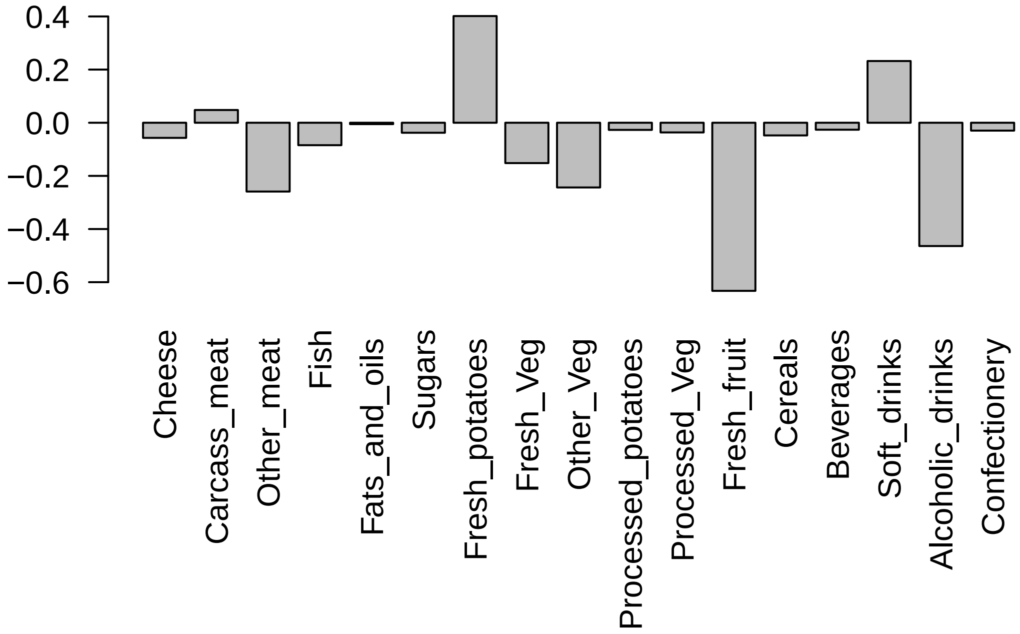
```
z <- summary(pca)
z$importance
```

```
                             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")
```

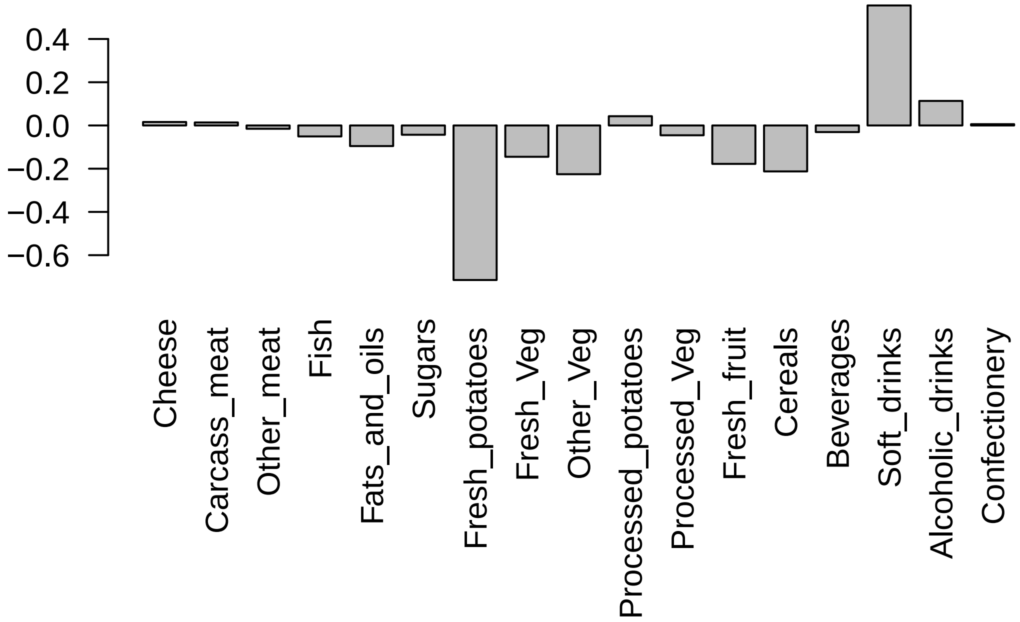


```
par(mar=c(10, 3, 0.35, 0))
barplot(pca$rotation[,1], las=2)
```



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

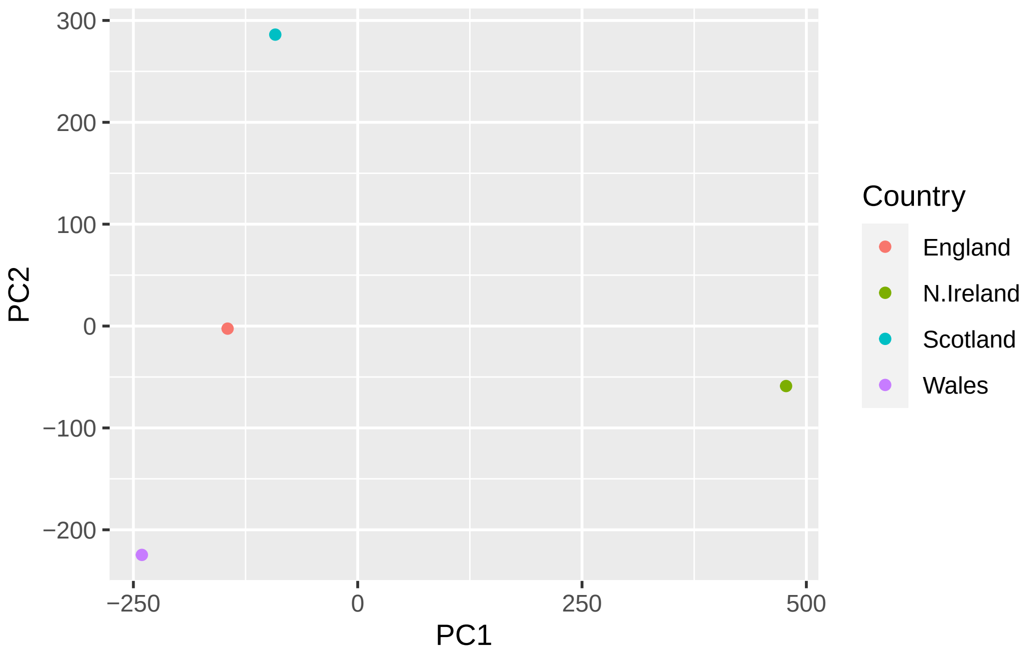
```
par(mar=c(10, 3, 0.35, 0))
barplot(pca$rotation[,2], las=2)
```



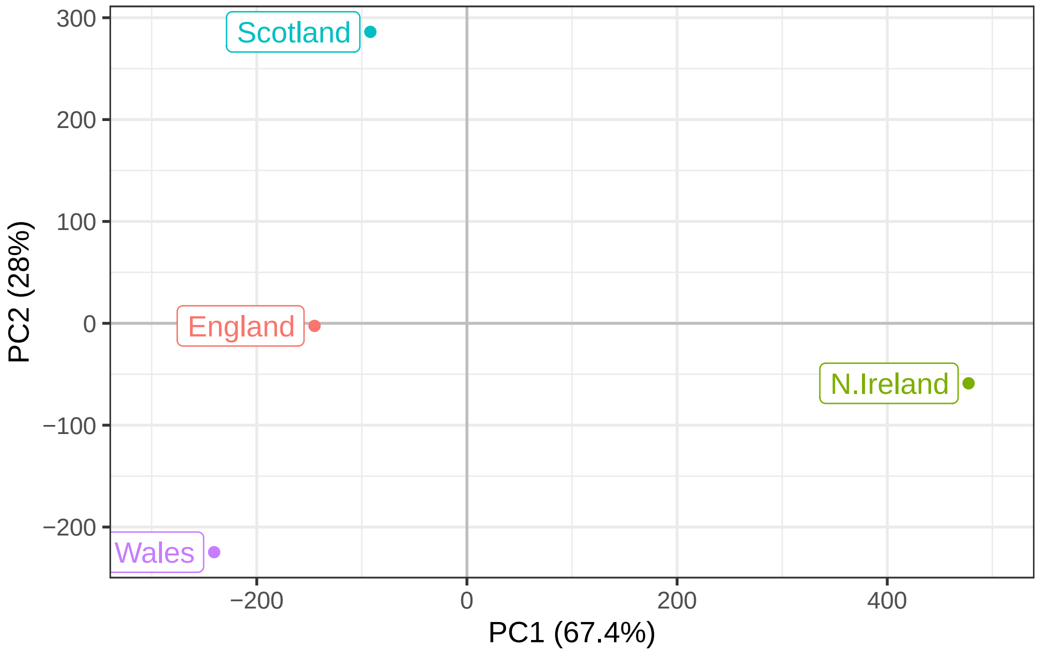
The two food groups that feature prominantely are fresh potatoes and soft drinks. This mainly tells us about the trends in quantities among the PC2 variable which accounts for 29% of the sample variance that can help us study the data set.

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

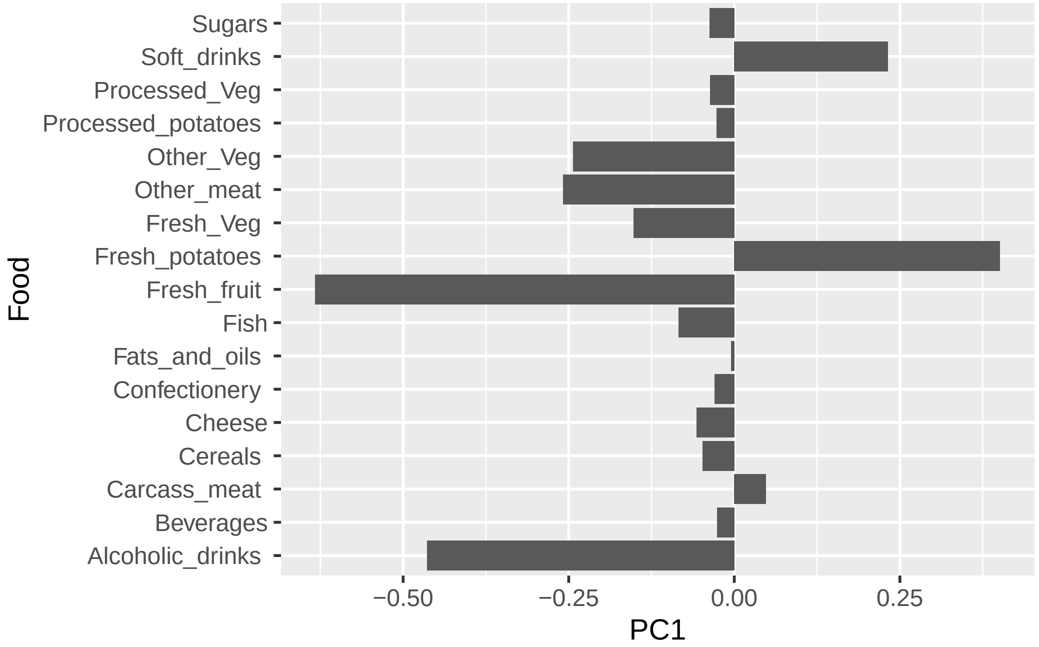
```
# First basic Plot
library(ggplot2)
ggplot(df_lab) + aes(PC1, PC2, col=Country) +geom_point()
```



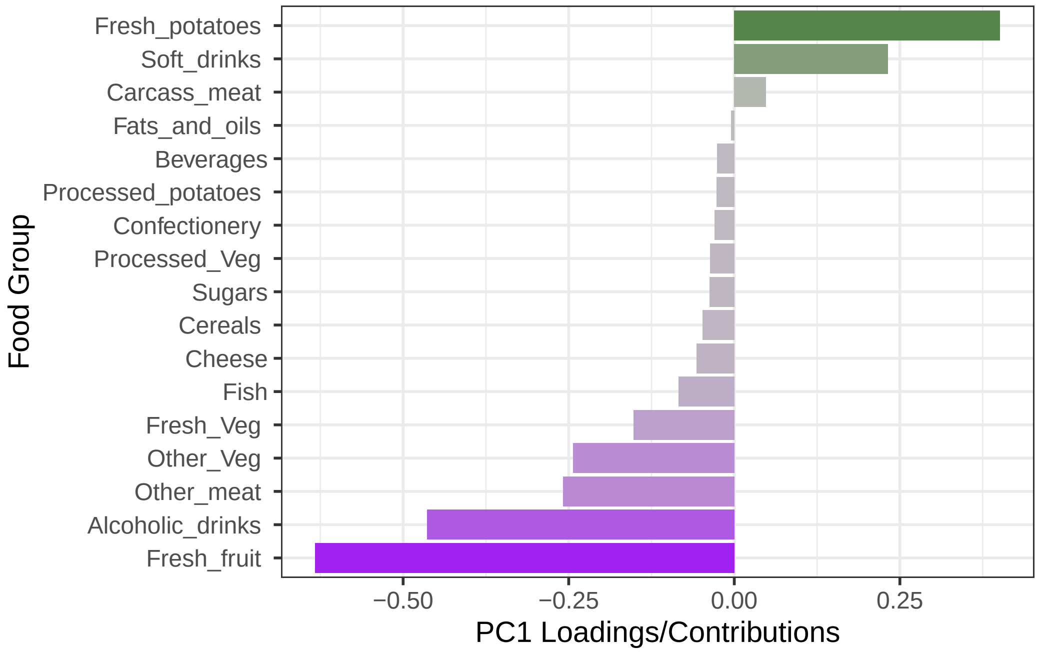
```
# A nicer plot
ggplot(df_lab) + aes(PC1, PC2, col=Country, label=Country) + geom_hline(yintercept=0, col=
```



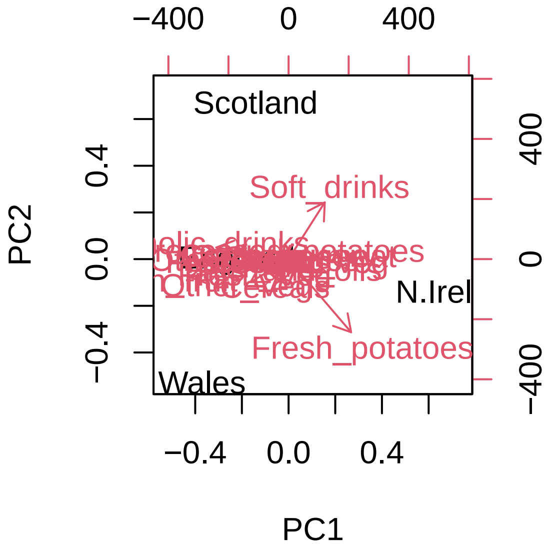
```r
ld <- as.data.frame(pca$rotation)
ld_lab <- tibble::rownames_to_column(ld, "Food")

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



```r
ggplot(ld_lab) +
  aes(PC1, reorder(Food, PC1), bg=PC1) +
  geom_col() +
  xlab("PC1 Loadings/Contributions") +
  ylab("Food Group") +
  scale_fill_gradient2(low="purple", mid="gray", high="darkgreen", guide=NULL) +
  theme_bw()
```


Food Group
Fresh_potatoes
Soft_drinks
Carcass_meat
Fats_and_oils
Beverages
Processed_potatoes
Confectionery
Processed_Veg
Sugars
Cereals
Cheese
Fish
Fresh_Veg
Other_Veg
Other_meat
Alcoholic_drinks
Fresh_fruit
−0.50
−0.25
0.00
0.25
PC1 Loadings/Contributions

```
biplot(pca)
```


−400
0
400
Scotland
Soft_drinks
N.Irel
Fresh_potatoes
Wales
0.4
0.0
−0.4
PC2
400
0
−400
−0.4
0.0
0.4
PC1

PCA of RNA-Seq data

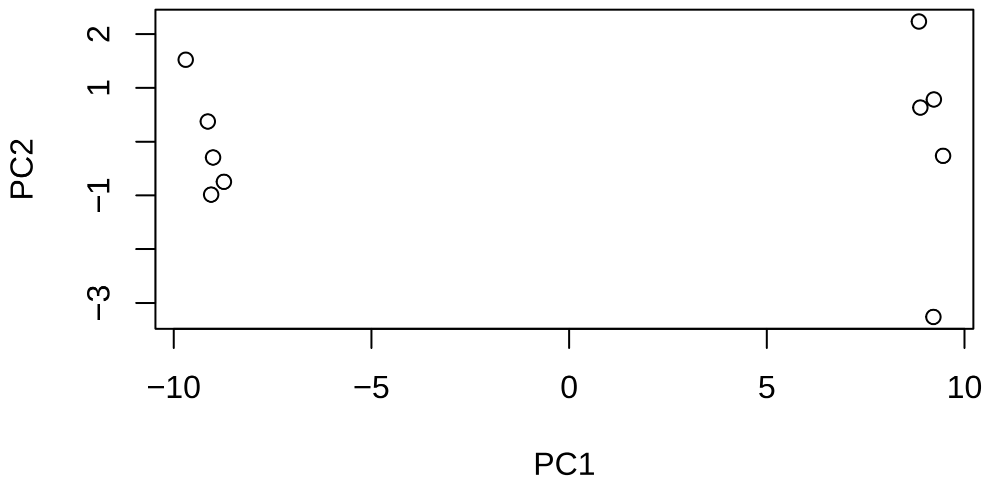
Data from the website

```
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
```

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

```
pca2 <-prcomp(t(rna.data), scale=TRUE)
plot(pca2$x[,1], pca2$x[,2], xlab="PC1", ylab="PC2")
```

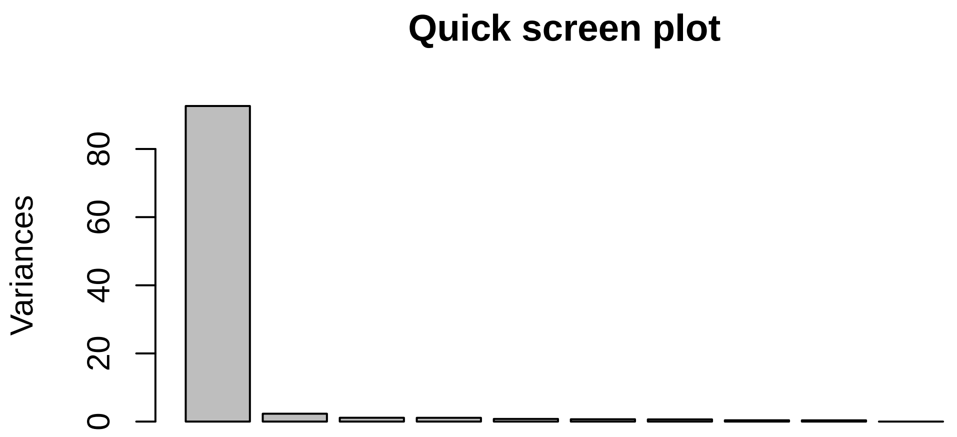


```
summary(pca2)
```

```
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
```

Plot of Variance

```
plot(pca2, main="Quick screen plot")
```

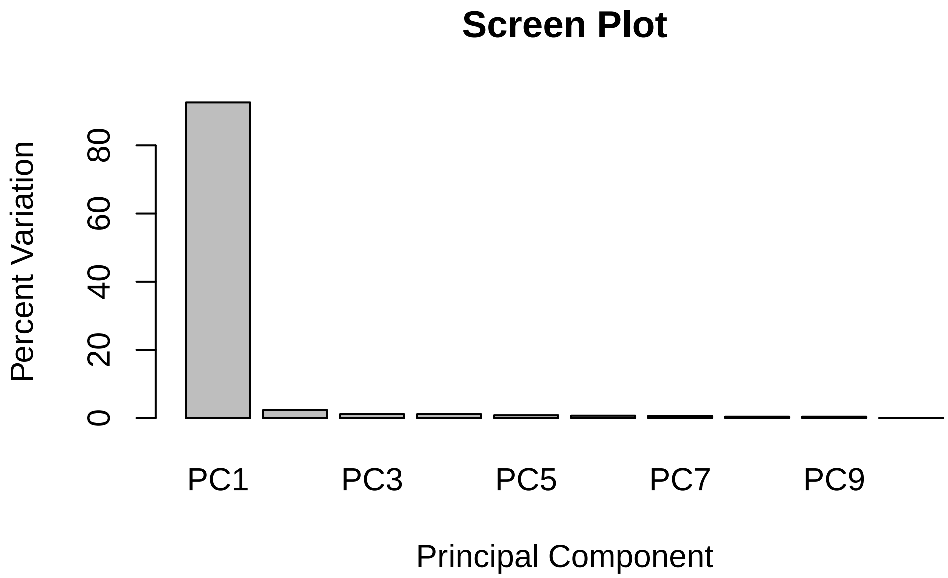


```
pca.var <- pca2$sdev^2

pca2.var.per <- round(pca.var/sum(pca.var)*100, 1)
```

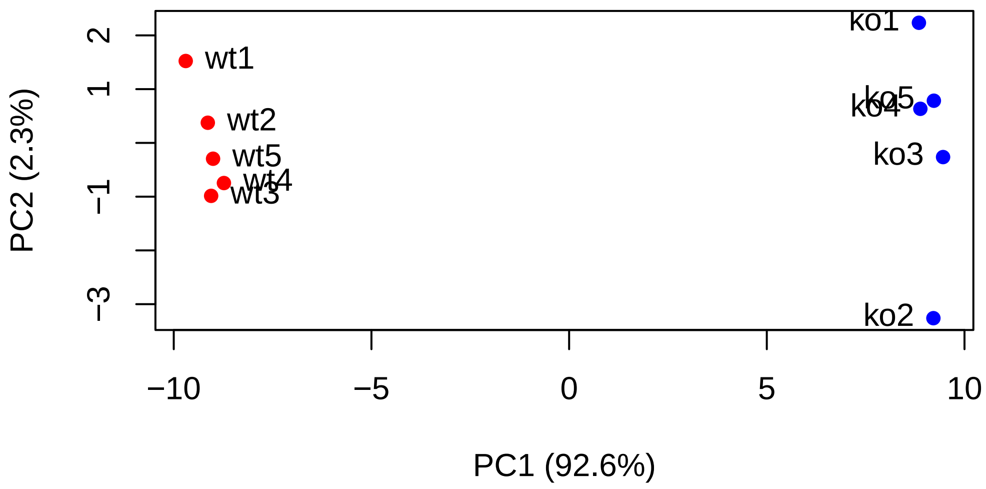
```
pca2.var.per
```

```
[1] 92.6  2.3  1.1  1.1  0.8  0.7  0.6  0.4  0.4  0.0
```

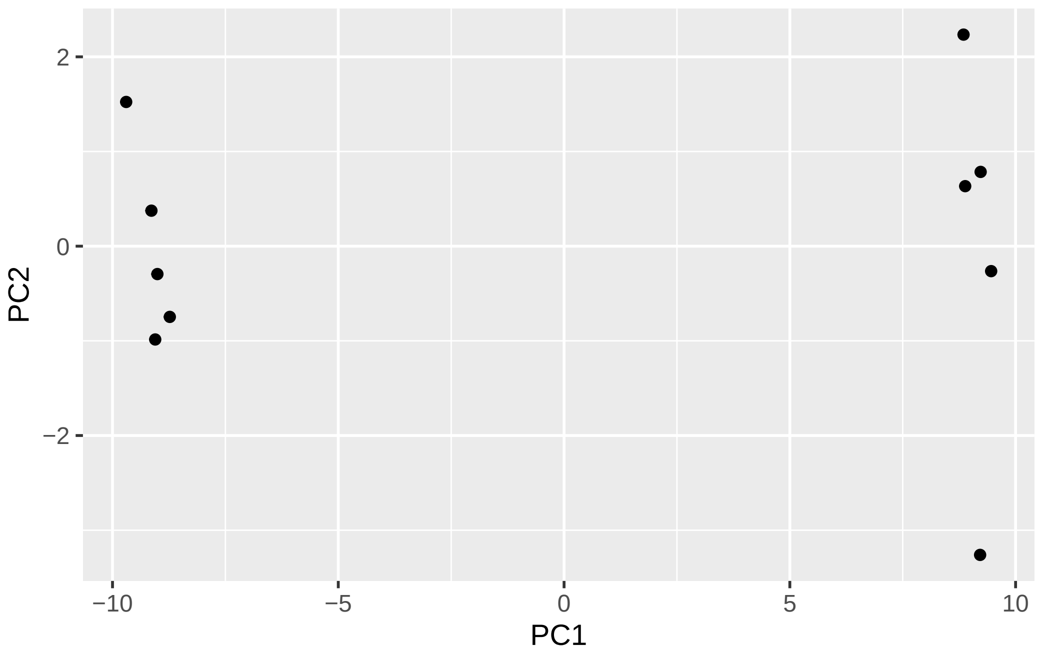
```
barplot(pca2.var.per, main="Screen Plot", names.arg=paste0("PC", 1:10), xlab= "Principal C
```



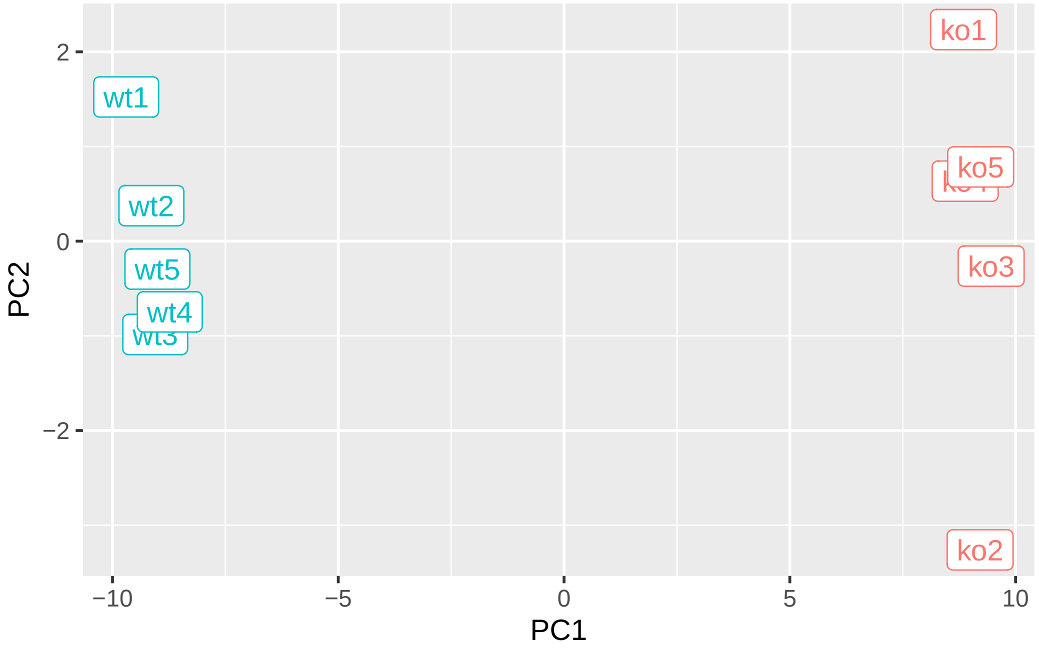
```
## A vector of colors for wt and ko samples
colvec <- colnames(rna.data)
colvec[grep("wt", colvec)] <- "red"
colvec[grep("ko", colvec)] <- "blue"

plot(pca2$x[,1], pca2$x[,2], col=colvec, pch=16,
     xlab=paste0("PC1 (", pca2.var.per[1], "%)"),
     ylab=paste0("PC2 (", pca2.var.per[2], "%)"))

text(pca2$x[,1], pca2$x[,2], labels = colnames(rna.data), pos=c(rep(4,5), rep(2,5)))
```



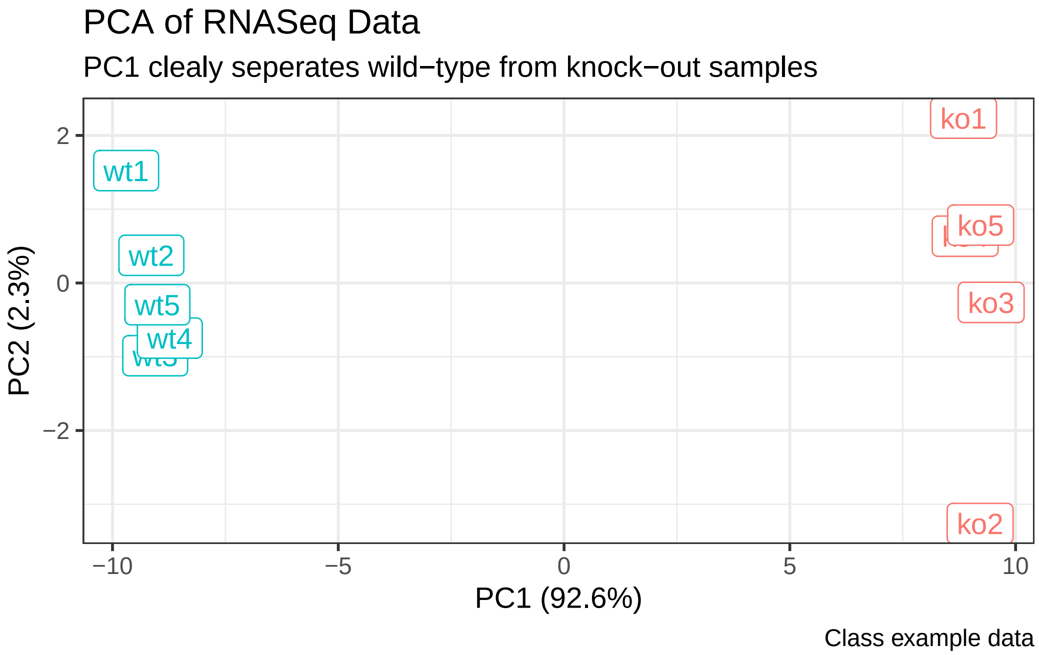
```
library(ggplot2)
df <- as.data.frame(pca2$x)
# Basic plot
ggplot(df) + aes(PC1, PC2) + geom_point()
```



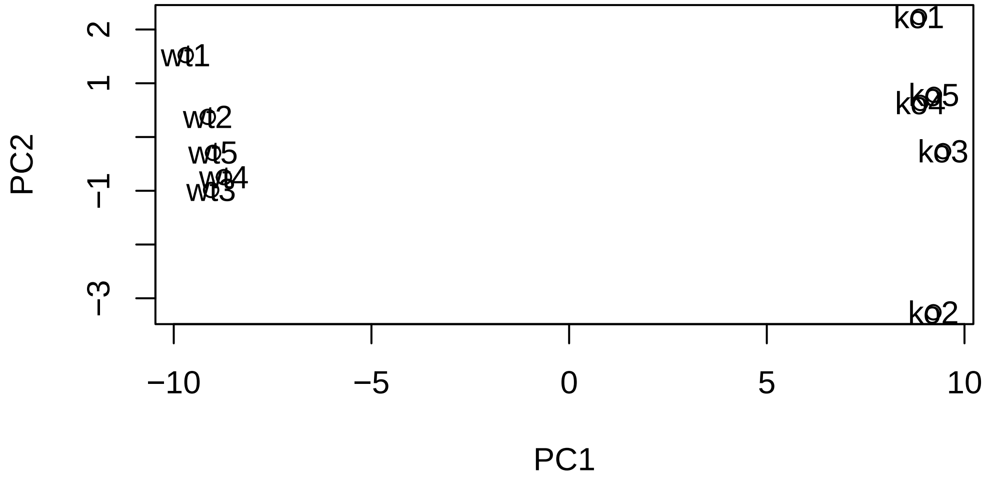
```
df$samples <- colnames(rna.data)
df$condition <- substr(colnames(rna.data),1,2)

p <- ggplot(df) +
        aes(PC1, PC2, label=samples, col=condition) +
        geom_label(show.legend = FALSE)
p
```



```
p + labs(title="PCA of RNASeq Data",
       subtitle = "PC1 clealy seperates wild-type from knock-out samples",
       x=paste0("PC1 (", pca2.var.per[1], "%)"),
       y=paste0("PC2 (", pca2.var.per[2], "%)"),
       caption="Class example data") +
     theme_bw()
```



```
plot(pca2$x[,1], pca2$x[,2], xlab="PC1", ylab="PC2")
text(pca2$x[,1], pca2$x[,2], colnames(rna.data))
```


ko1
wt1
ko5
ko4
wt2
ko3
wt5
wt4
wt3
ko2
PC2
PC1