Class07: Machine learning 1

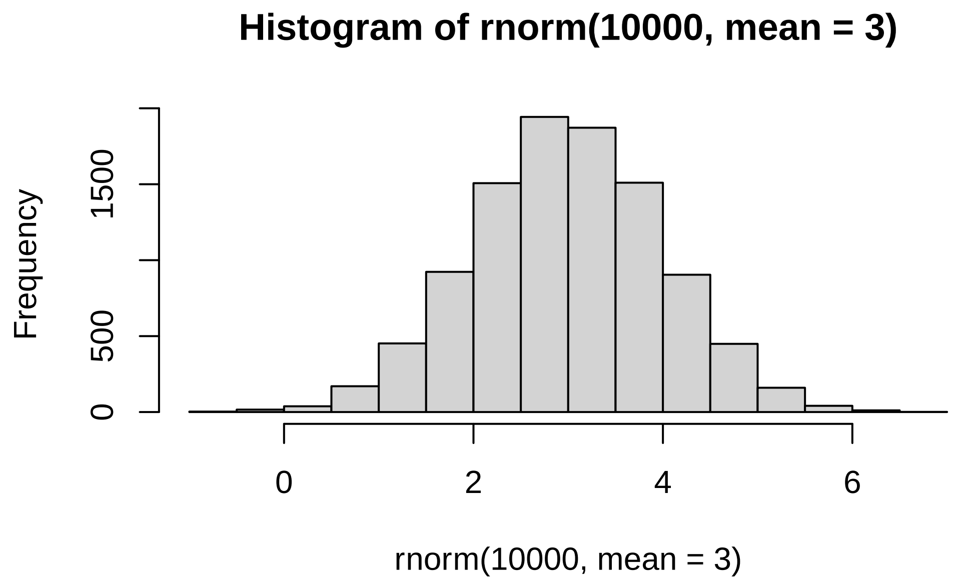
Changcheng Li (PID: A69027828)

Clustering

We will start with k-means clustering, one of the most prevalent of all clustering methods.

To get started let's make some data up:

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

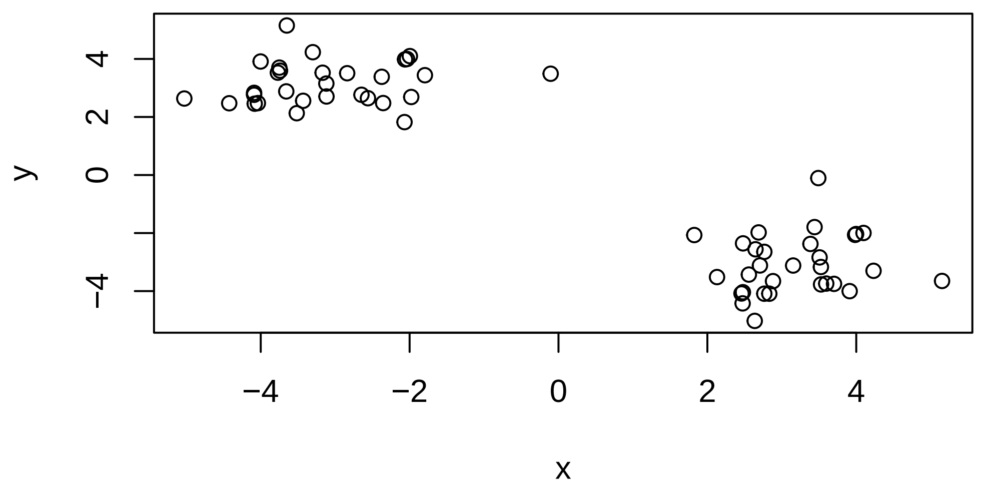


```r
tmp <- c( rnorm(30,3), rnorm(30,-3))
x <- cbind(x = tmp, y = rev(tmp))
x
```

```
              x          y
 [1,]  2.4733430 -4.4232836
 [2,]  4.2323131 -3.3005598
 [3,]  3.4401745 -1.7945127
 [4,]  3.5250546 -3.1689156
 [5,]  2.4584811 -4.0812408
 [6,]  3.1522188 -3.1186236
 [7,]  4.0976563 -1.9962970
 [8,]  2.7651893 -2.6470566
 [9,]  3.5271825 -3.7700875
[10,]  3.9979107 -2.0345065
[11,]  3.5076948 -2.8388374
[12,]  2.6884878 -1.9782140
[13,]  2.6466121 -2.5598126
[14,]  5.1527766 -3.6500956
[15,]  2.7062181 -3.1168045
[16,]  3.3846049 -2.3747892
[17,]  2.4794192 -2.3554828
[18,]  2.8818319 -3.6570434
[19,]  2.6362707 -5.0266382
[20,]  2.4775753 -4.0383965
[21,]  1.8240338 -2.0685621
[22,]  3.5969678 -3.7408185
[23,]  3.4903044 -0.1052975
[24,]  3.9846518 -2.0619940
[25,]  2.1296809 -3.5163982
[26,]  2.8320764 -4.0890048
[27,]  2.7644796 -4.0891756
[28,]  2.5572380 -3.4302671
[29,]  3.9115423 -4.0019557
[30,]  3.7013929 -3.7501856
[31,] -3.7501856  3.7013929
[32,] -4.0019557  3.9115423
[33,] -3.4302671  2.5572380
[34,] -4.0891756  2.7644796
[35,] -4.0890048  2.8320764
[36,] -3.5163982  2.1296809
[37,] -2.0619940  3.9846518
[38,] -0.1052975  3.4903044
[39,] -3.7408185  3.5969678
[40,] -2.0685621  1.8240338
[41,] -4.0383965  2.4775753
[42,] -5.0266382  2.6362707
```

```
[43,] -3.6570434  2.8818319
[44,] -2.3554828  2.4794192
[45,] -2.3747892  3.3846049
[46,] -3.1168045  2.7062181
[47,] -3.6500956  5.1527766
[48,] -2.5598126  2.6466121
[49,] -1.9782140  2.6884878
[50,] -2.8388374  3.5076948
[51,] -2.0345065  3.9979107
[52,] -3.7700875  3.5271825
[53,] -2.6470566  2.7651893
[54,] -1.9962970  4.0976563
[55,] -3.1186236  3.1522188
[56,] -4.0812408  2.4584811
[57,] -3.1689156  3.5250546
[58,] -1.7945127  3.4401745
[59,] -3.3005598  4.2323131
[60,] -4.4232836  2.4733430
```

```
plot(x)
```



The main function in R for K-means clustering is called ‘kmeans()’

```r
k <- kmeans(x, centers = 2, nstart = 20)
k
```

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

Cluster means:
          x         y
1  3.167446 -3.092829
2 -3.092829  3.167446

Clustering vector:
 [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 2 2 2 2 2 2 2 2
[39] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 45.99167 45.99167
 (between_SS / total_SS =  92.7 %)

Available components:

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

```r
#View(k)
```

Q1. How many points are in each cluster

```r
k$size
```

```
[1] 30 30
```

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

```r
k$cluster
```

```
 [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 2 2 2 2 2 2 2 2
[39] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

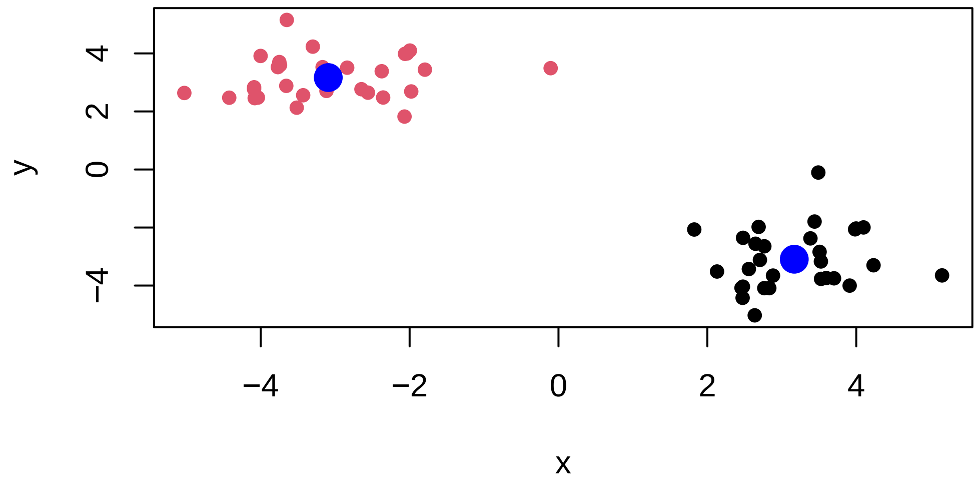
Q3. Cluster centers

```r
k$centers
```

```
          x         y
1  3.167446 -3.092829
2 -3.092829  3.167446
```

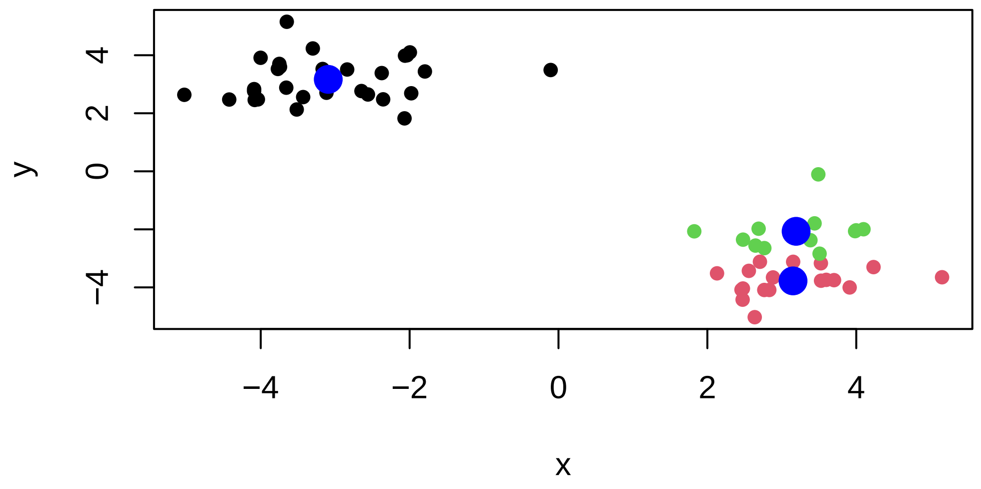
> Q4. Make a plot of our data colored by clustering results with optionally the cluster centers shown

```r
plot(x, col = k$cluster, pch = 16)
points(k$centers, col = "blue", pch = 16, cex =2)
```



> Q5. Run kmeans again but cluster into 3 groups and plot results like we did above

```r
k3 <- kmeans(x, centers = 3, nstart = 20)
plot(x, col = k3$cluster, pch = 16)
points(k3$centers, col = "blue", pch = 16, cex =2)
```



K-means will always return a clustering result - even if there is no clear groupings.

#Hierarchical Clustering

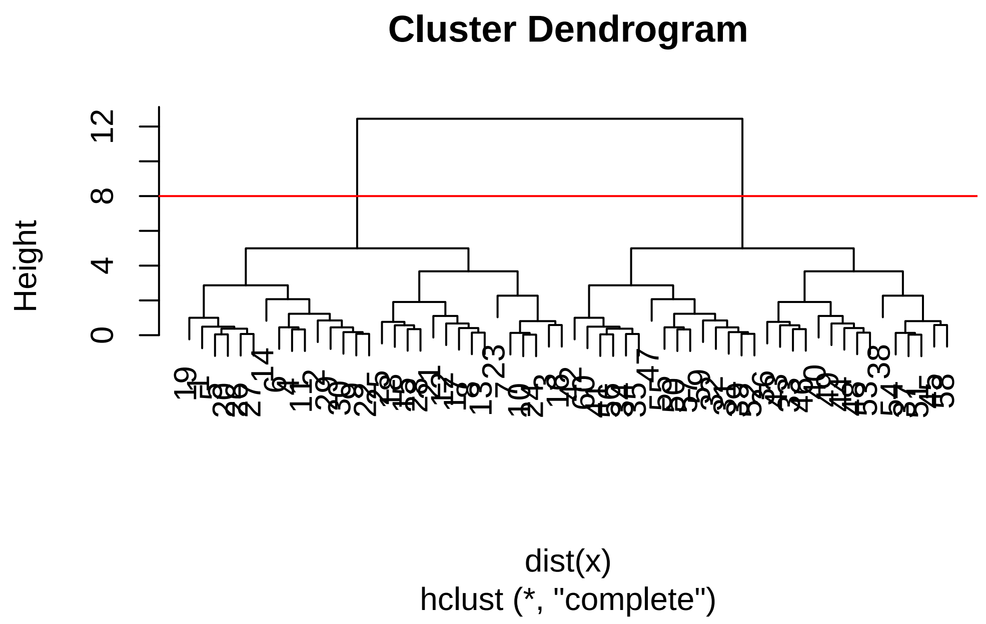
Hierarchical clustering it has an advantage in that it can reveal the structure in your data rather than imposing a structure as k-means will.

The main function in "base" R is called 'hclust()'

It requires a distance matrix input.

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

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

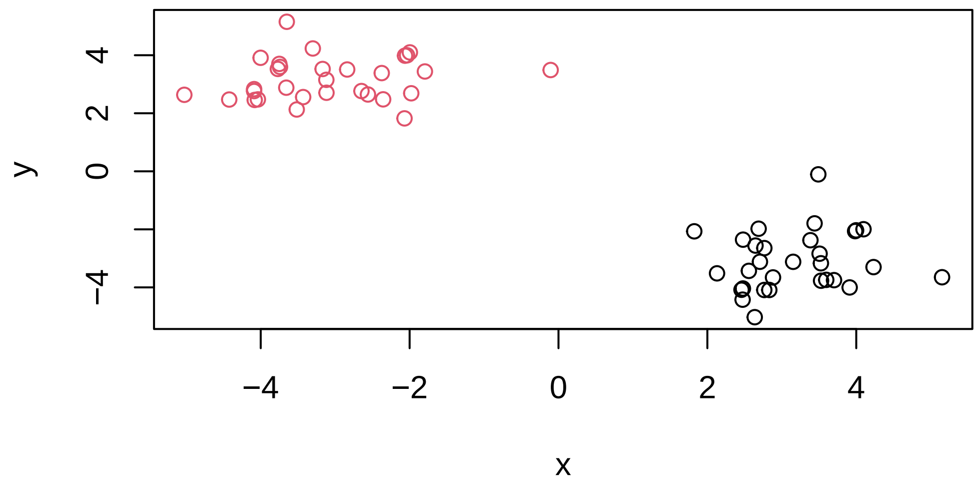


The function to get our clusters/groups fron a hclust object is called ‘cutree()’

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

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

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



#Principle component analysis

```
#Q1. How many rows and columns are in your new data frame named x? What R functions could
url <- "https://tinyurl.com/UK-foods"
x <- read.csv(url)
## Complete the following code to find out how many rows and columns are in x?
dim(x)
```

```
[1] 17  5
```

```
## Preview the first 6 rows
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
```

```r
# Note how the minus indexing works
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
```

```r
dim(x)
```

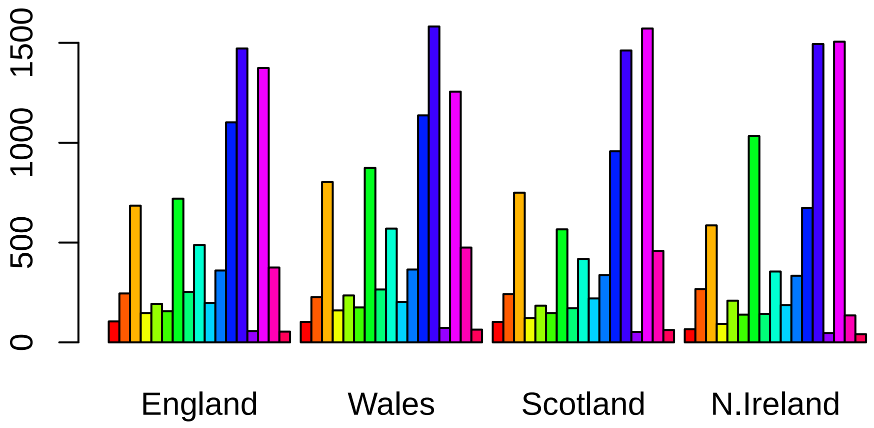
```
[1] 17  4
```

```r
#A better way for exclude rowname
x <- read.csv(url, row.names=1)
head(x)
```

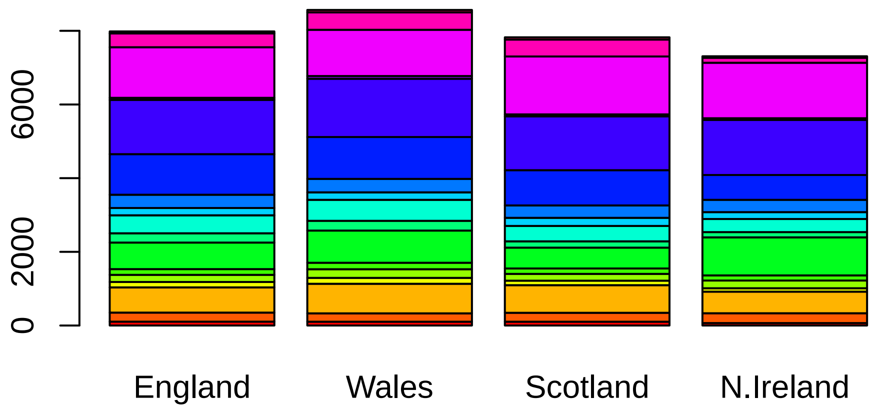
```
               England Wales Scotland N.Ireland
Cheese             105   103      103        66
Carcass_meat       245   227      242       267
Other_meat         685   803      750       586
Fish               147   160      122        93
Fats_and_oils      193   235      184       209
Sugars  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

#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? "x <- read.csv(url, row.names=1) head(x)" works more robust because when you rerun the chunk the x <- x[,-1] in the first method will shrink x

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

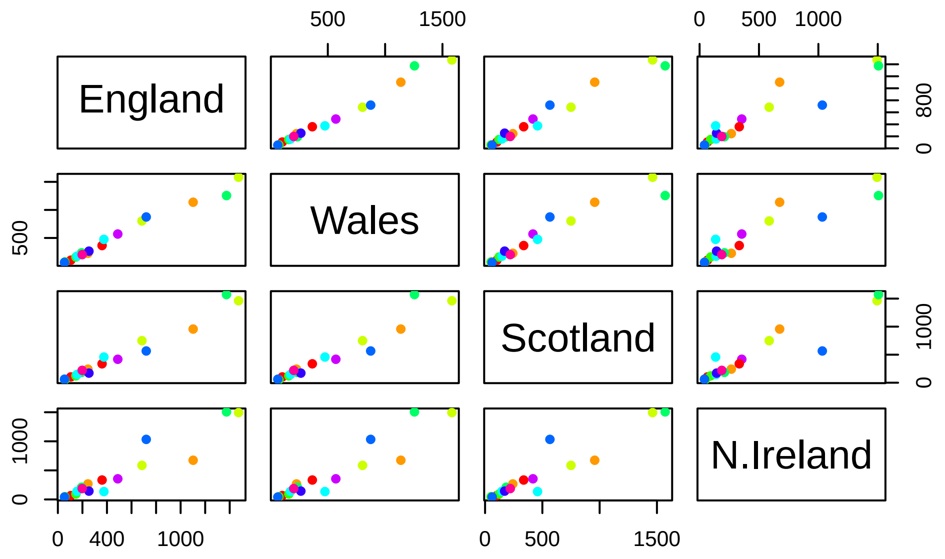


```
#Q3: Changing what optional argument in the above barplot() function results in the follow
#beside = FALSE
barplot(as.matrix(x), beside=FALSE, col=rainbow(nrow(x)))
```



#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? The code uses 17 colors to distinguish different food. The resulting figure are food consumption paired for the corresponding horizontal and vertical countries. When a given point lies on the diagonal, it represents that this food consumption is similar in the two corresponding countries.

```
pairs(x, col=rainbow(10), pch=16)
```



#Q6. What is the main differences between N. Ireland and the other countries of the UK in terms of this data-set? The points are more dispersed from the diagonal in plots comparing N.Ireland and other countries than that in other plots(like Fresh_potatoes represented by the blue point and Fresh_fruit represented by the orange point), which means food consumption is more different N.Ireland.

#PCA to rescue

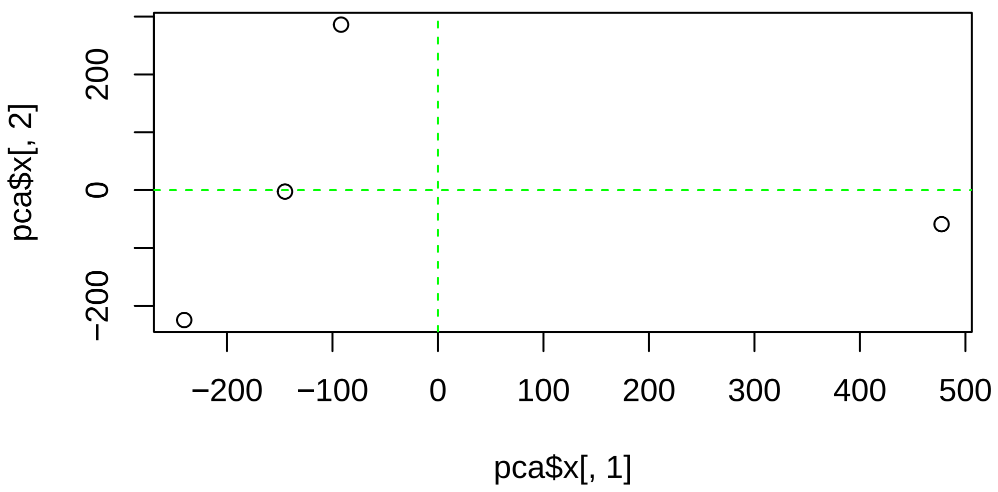
Help me make sense of this data The main function for PCA in base R is called 'prcomp()'

```
# Use the prcomp() PCA function
pca <- prcomp( t(x) )
summary(pca)
```

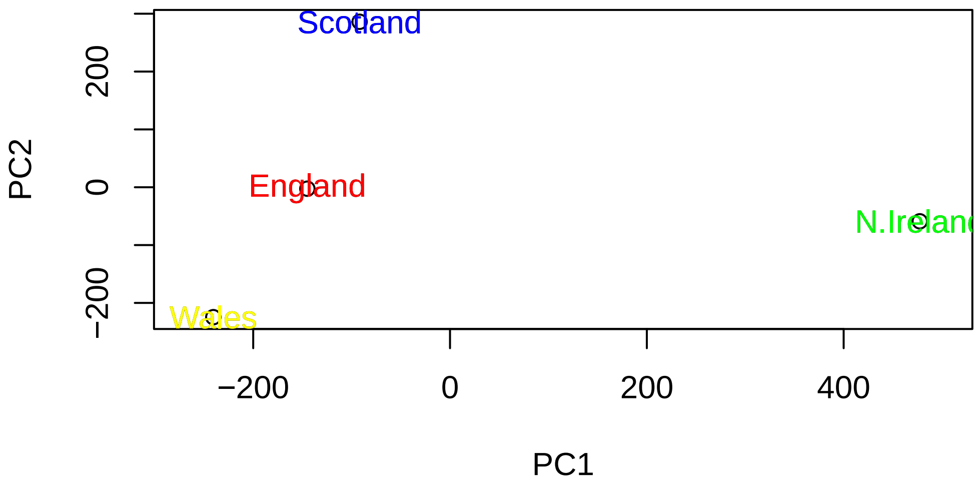
```
Importance of components:
                            PC1      PC2      PC3       PC4
Standard deviation     324.1502 212.7478 73.87622 3.176e-14
Proportion of Variance   0.6744   0.2905  0.03503 0.000e+00
Cumulative Proportion    0.6744   0.9650  1.00000 1.000e+00
```

One of the main results that folks look for is called the "score plot" a.k.a PC plot, PC1 vs PC2 plot.

```
plot(pca$x[,1], pca$x[,2])
abline(h = 0, v = 0, col = 'green', lty = "dashed")
```



```
# Plot PC1 vs PC2
#Q7. Complete the code below to generate a plot of PC1 vs PC2. The second line adds text l
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 U
text(pca$x[,1], pca$x[,2], colnames(x), col = c("red", "yellow", "blue", "green"))
```



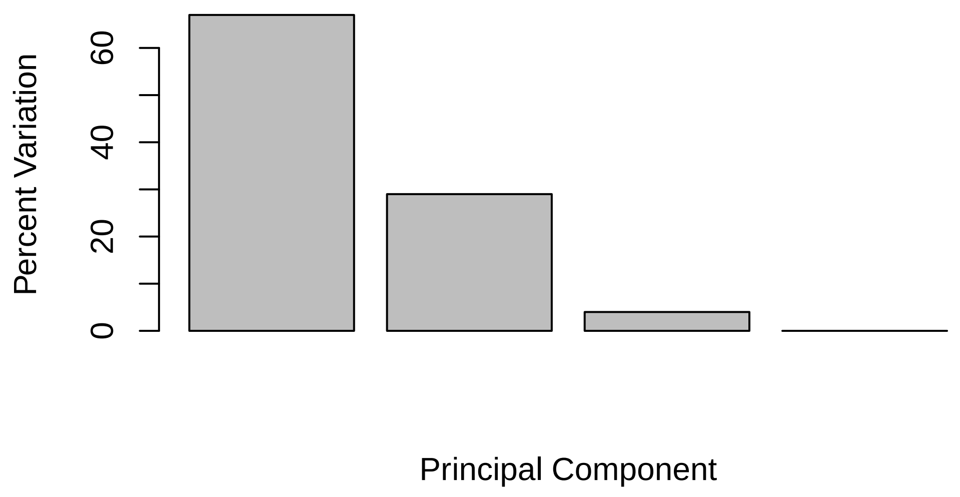
```r
v <- round( pca$sdev^2/sum(pca$sdev^2) * 100 )
v
```

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

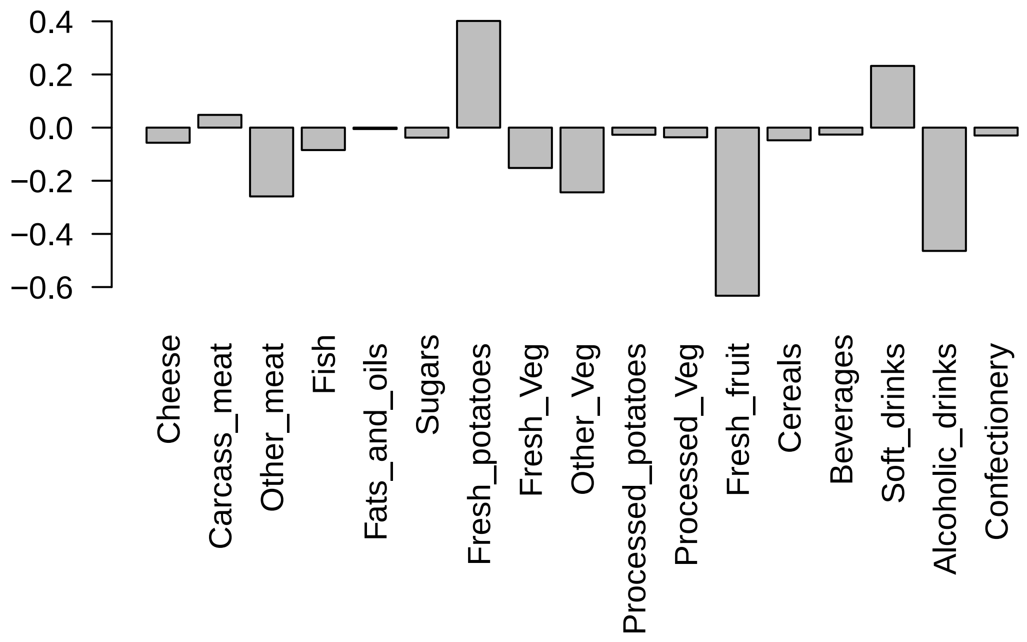
```r
## or the second row here...
z <- summary(pca)
z$importance
```

```
                             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
```

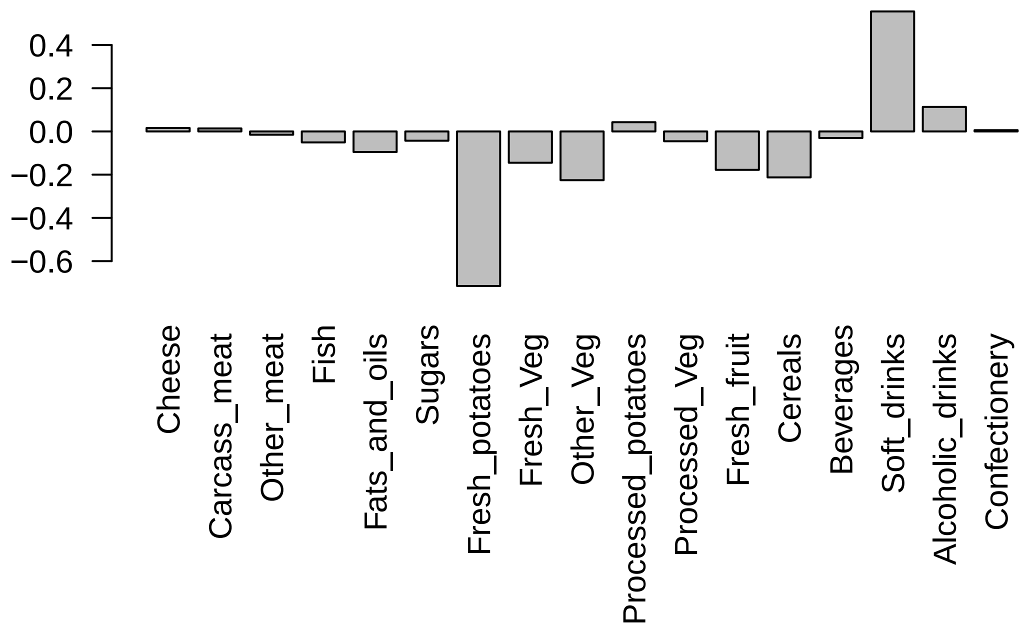
```r
barplot(v, xlab="Principal Component", ylab="Percent Variation")
```



```
## 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 )
```

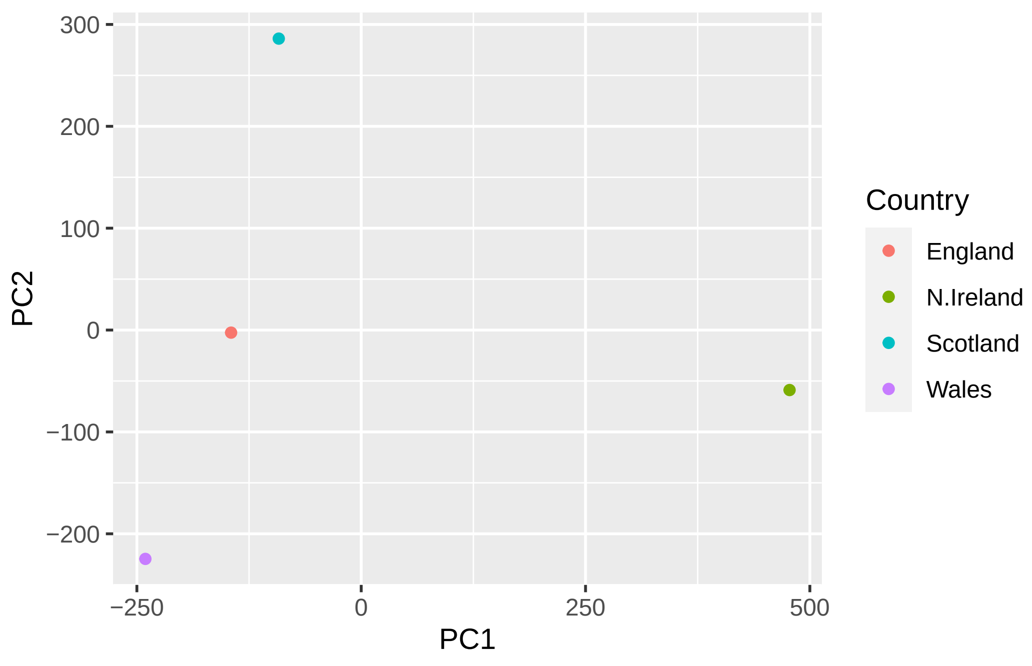


```
#Q9: Generate a similar ‘loadings plot’ for PC2. What two food groups feature prominantely
par(mar=c(10, 3, 0.35, 0))
barplot( pca$rotation[,2], las=2 )
```

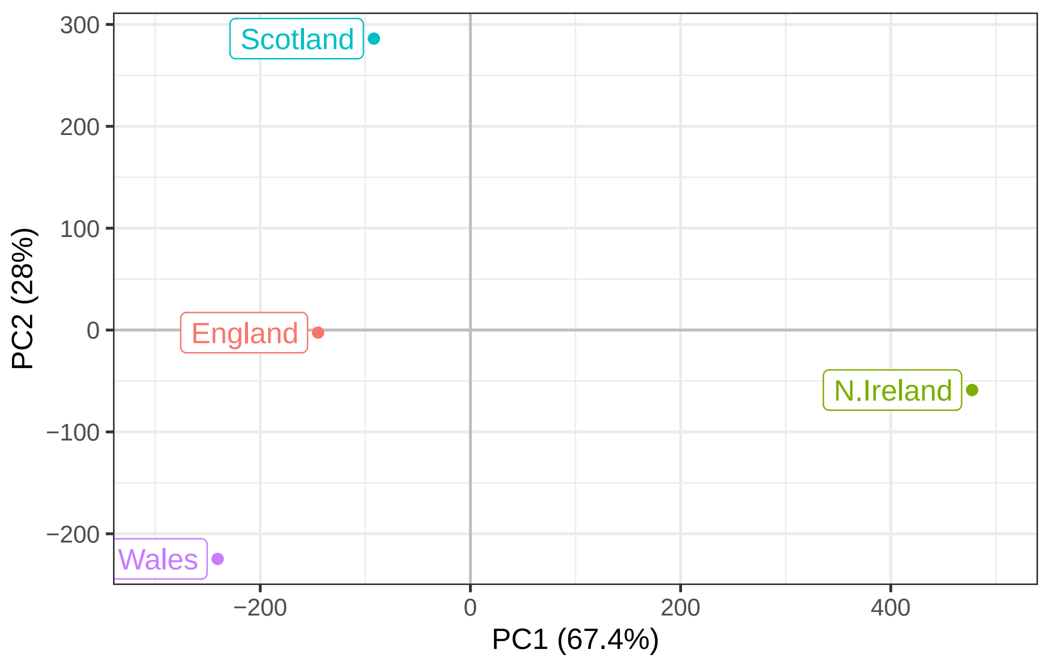


```
#Fresh_potatoes and Soft_drinks are the two food groups feature prominantely.PC2 tells tha
```

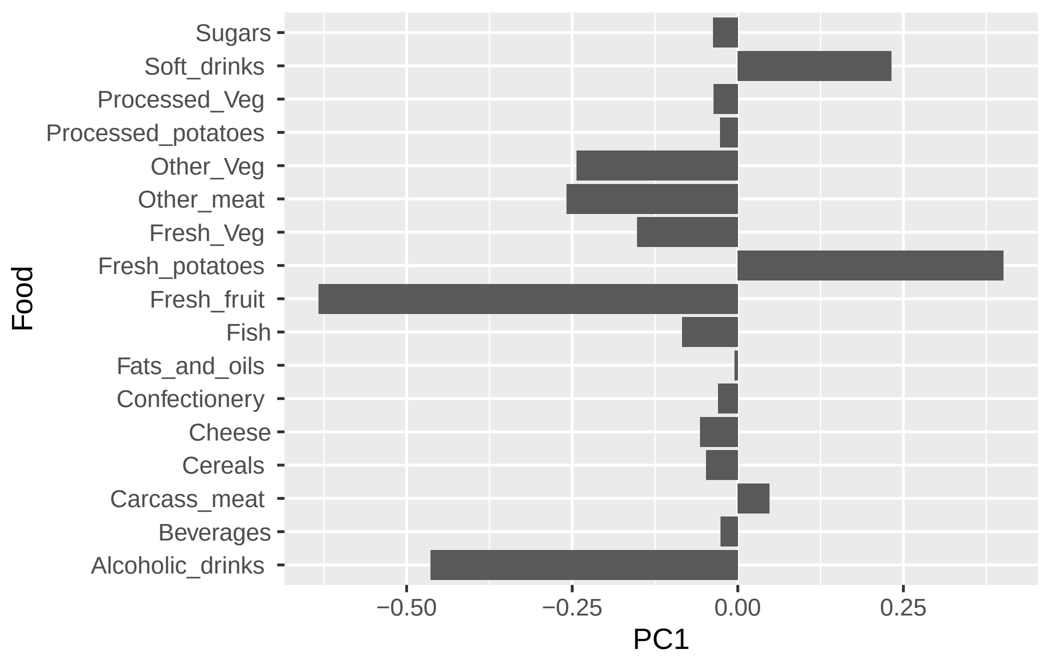
```
library(ggplot2)

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

# Our first basic plot
ggplot(df_lab) +
  aes(PC1, PC2, col=Country) +
  geom_point()
```



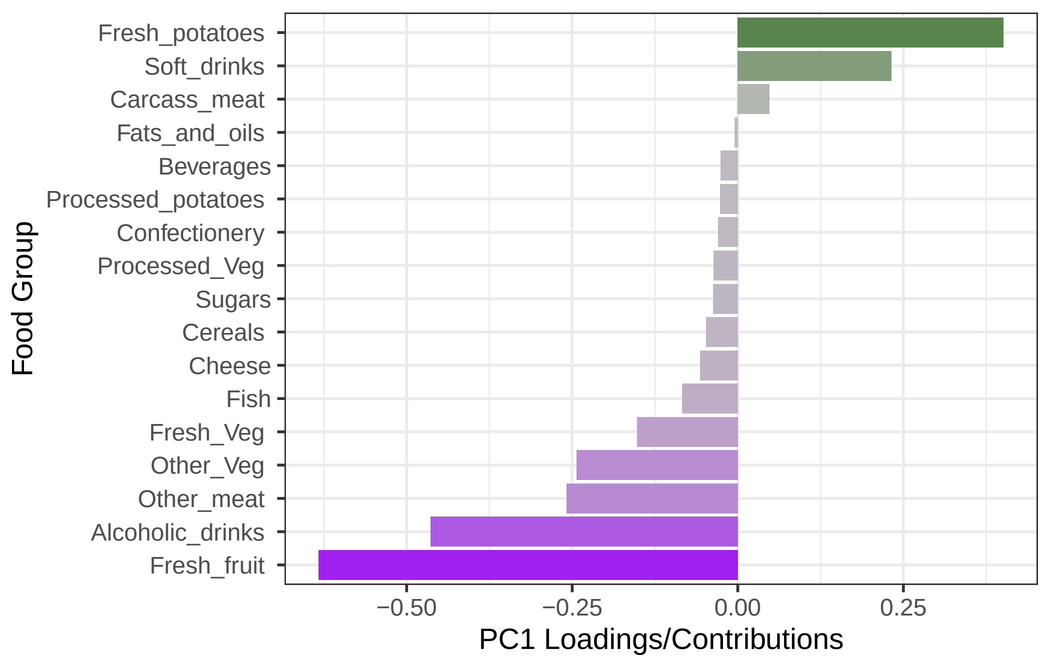
```
ggplot(df_lab) +
  aes(PC1, PC2, col=Country, label=Country) +
  geom_hline(yintercept = 0, col="gray") +
  geom_vline(xintercept = 0, col="gray") +
  geom_point(show.legend = FALSE) +
  geom_label(hjust=1, nudge_x = -10, show.legend = FALSE) +
  expand_limits(x = c(-300,500)) +
  xlab("PC1 (67.4%)") +
  ylab("PC2 (28%)") +
  theme_bw()
```



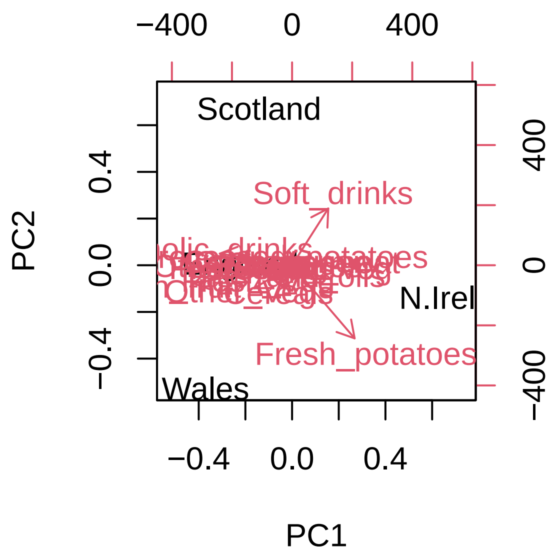
```
ld <- as.data.frame(pca$rotation)
ld_lab <- tibble::rownames_to_column(ld, "Food")

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



```
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()
```



```r
## The inbuilt biplot() can be useful for small datasets
biplot(pca)
```



```r
url2 <- "https://tinyurl.com/expression-CSV"
rna.data <- read.csv(url2, row.names=1)
head(rna.data)
```

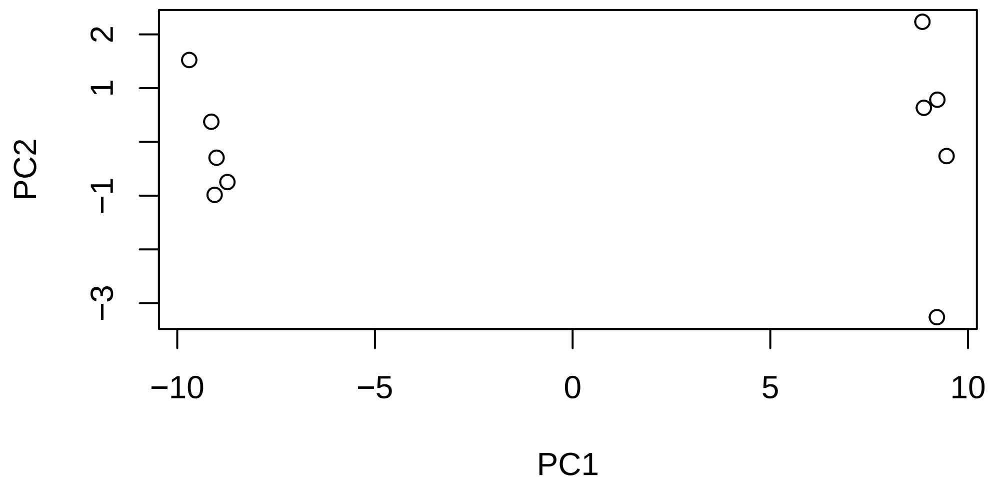
```
       wt1 wt2  wt3  wt4 wt5 ko1 ko2 ko3 ko4 ko5
gene1  439 458  408  429 420  90  88  86  90  93
gene2  219 200  204  210 187 427 423 434 433 426
gene3 1006 989 1030 1017 973 252 237 238 226 210
gene4  783 792  829  856 760 849 856 835 885 894
gene5  181 249  204  244 225 277 305 272 270 279
gene6  460 502  491  491 493 612 594 577 618 638
```

```r
#Q10: How many genes and samples are in this data set?
dim(rna.data)
```

```
[1] 100  10
```

```r
#100 genes and 10 samples
```

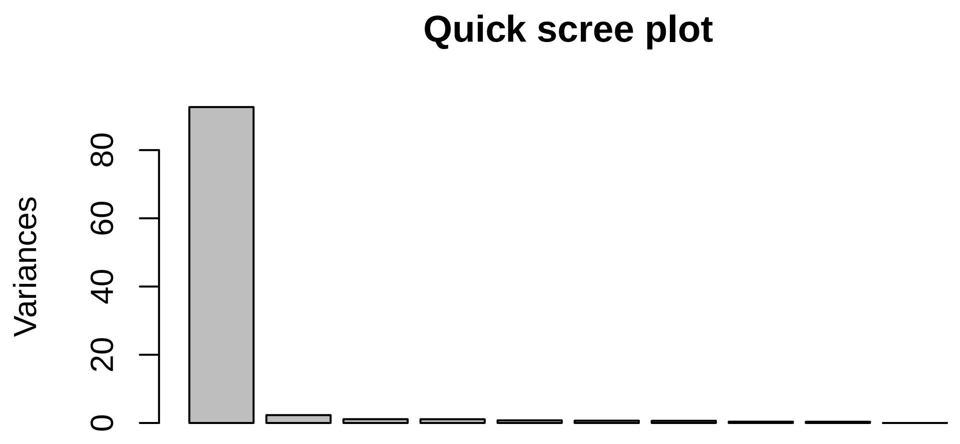
```r
## Again we have to take the transpose of our data
pca <- prcomp(t(rna.data), scale=TRUE)

## Simple un polished plot of pc1 and pc2
plot(pca$x[,1], pca$x[,2], xlab="PC1", ylab="PC2")
```



```r
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
```

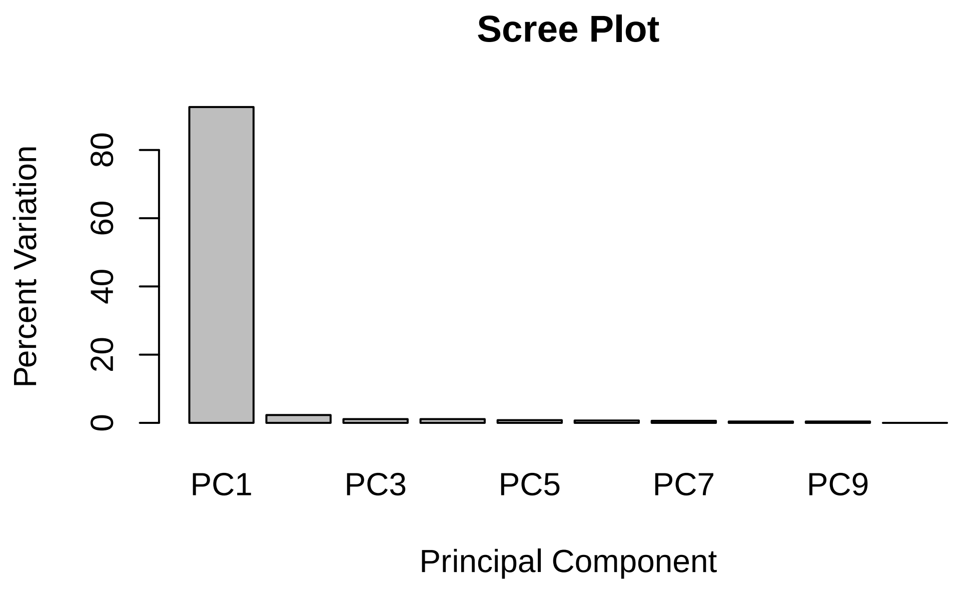
```r
plot(pca, main="Quick scree plot")
```



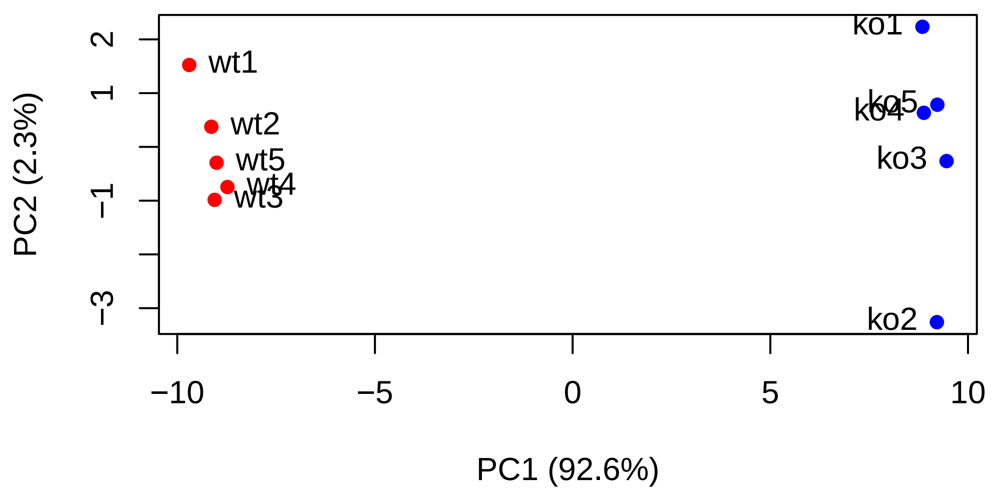
```r
## Variance captured per PC
pca.var <- pca$sdev^2

## Percent variance is often more informative to look at
pca.var.per <- round(pca.var/sum(pca.var)*100, 1)
pca.var.per
```

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

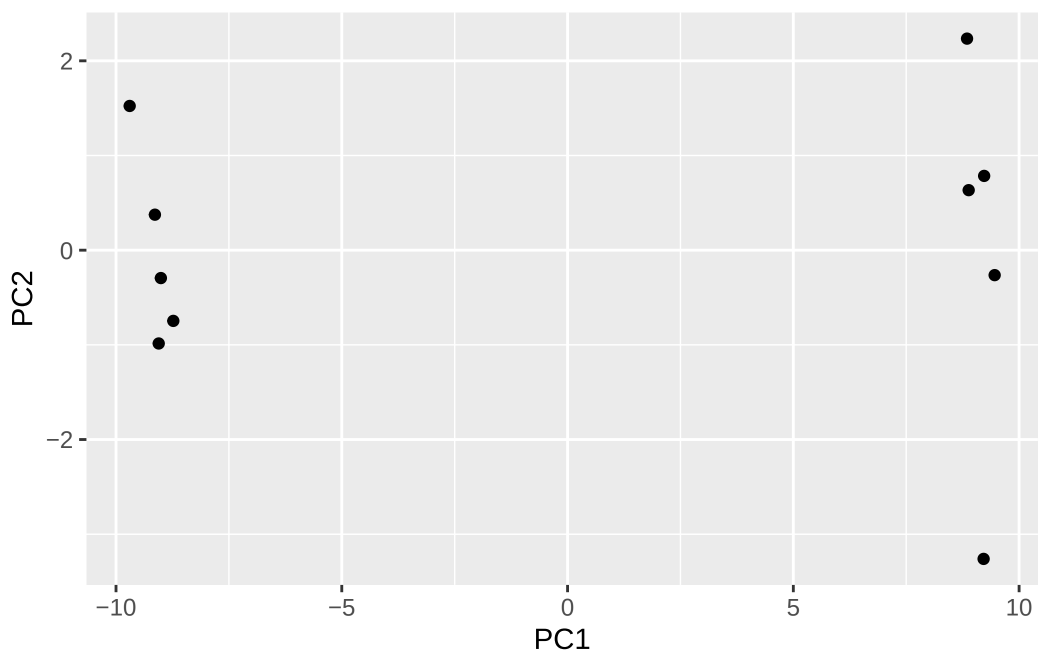
```r
barplot(pca.var.per, main="Scree Plot",
        names.arg = paste0("PC", 1:10),
        xlab="Principal Component", ylab="Percent Variation")
```



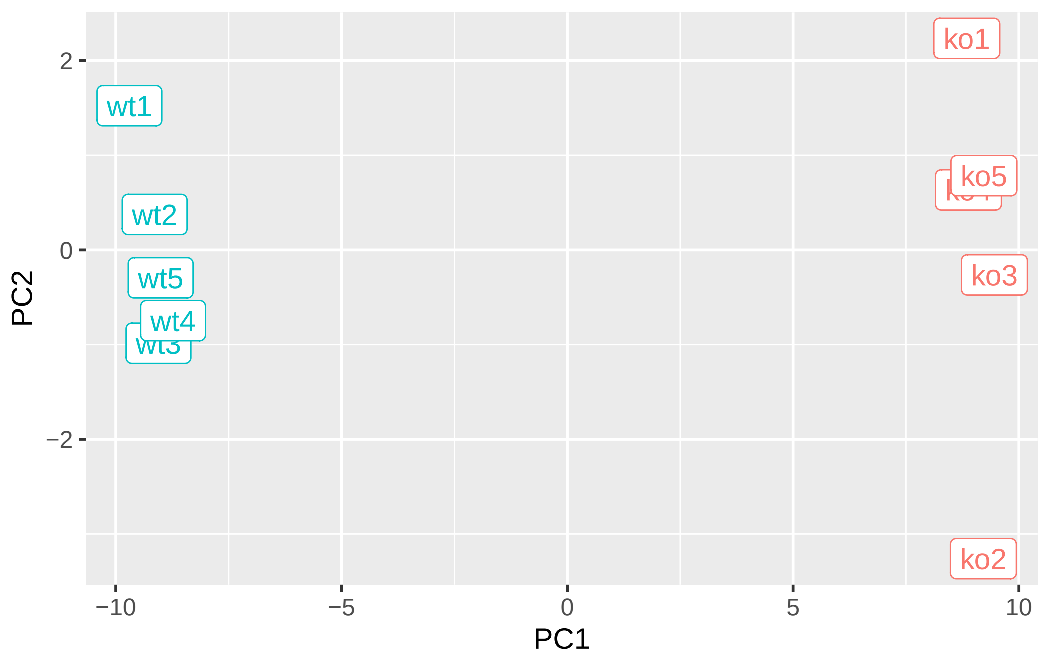
```r
## A vector of colors for wt and ko samples
colvec <- colnames(rna.data)
colvec[grep("wt", colvec)] <- "red"
colvec[grep("ko", colvec)] <- "blue"

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

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



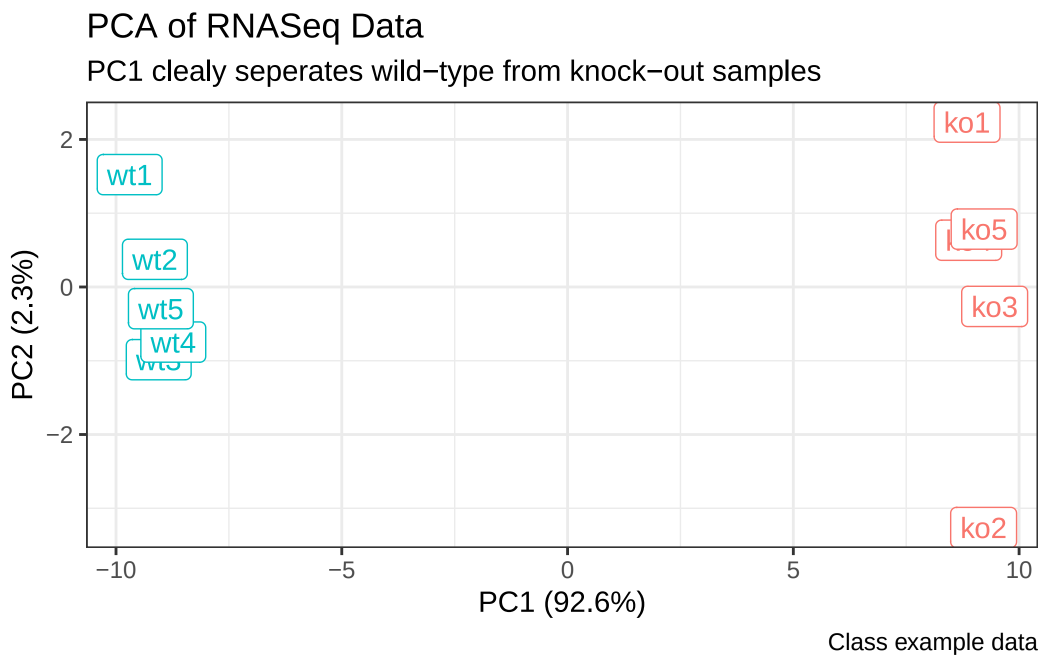
```r
library(ggplot2)

df <- as.data.frame(pca$x)

# Our first basic plot
ggplot(df) +
  aes(PC1, PC2) +
  geom_point()
```



```r
# Add a 'wt' and 'ko' "condition" column
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 (", pca.var.per[1], "%)"),
       y=paste0("PC2 (", pca.var.per[2], "%)"),
       caption="Class example data") +
     theme_bw()
```



```
loading_scores <- pca$rotation[,1]

## Find the top 10 measurements (genes) that contribute
## most to PC1 in either direction (+ or -)
gene_scores <- abs(loading_scores)
gene_score_ranked <- sort(gene_scores, decreasing=TRUE)

## show the names of the top 10 genes
top_10_genes <- names(gene_score_ranked[1:10])
top_10_genes
```

```
 [1] "gene100" "gene66"  "gene45"  "gene68"  "gene98"  "gene60"  "gene21"
 [8] "gene56"  "gene10"  "gene90"
```

```
sessionInfo()
```

```
R version 4.3.1 (2023-06-16 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19045)
```

```
Matrix products: default


locale:
[1] LC_COLLATE=Chinese (Simplified)_China.utf8
[2] LC_CTYPE=Chinese (Simplified)_China.utf8
[3] LC_MONETARY=Chinese (Simplified)_China.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=Chinese (Simplified)_China.utf8

time zone: America/Tijuana
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] ggplot2_3.4.4

loaded via a namespace (and not attached):
 [1] vctrs_0.6.4       cli_3.6.1         knitr_1.44        rlang_1.1.1
 [5] xfun_0.40         generics_0.1.3    jsonlite_1.8.7    labeling_0.4.3
 [9] glue_1.6.2        colorspace_2.1-0  htmltools_0.5.6.1 scales_1.2.1
[13] fansi_1.0.5       rmarkdown_2.25    grid_4.3.1        evaluate_0.22
[17] munsell_0.5.0     tibble_3.2.1      fastmap_1.1.1     yaml_2.3.7
[21] lifecycle_1.0.3   compiler_4.3.1    dplyr_1.1.3       pkgconfig_2.0.3
[25] farver_2.1.1      digest_0.6.33     R6_2.5.1          tidyselect_1.2.0
[29] utf8_1.2.3        pillar_1.9.0      magrittr_2.0.3    withr_2.5.1
[33] tools_4.3.1       gtable_0.3.4
```