Class07

Jiawei Xu

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

nrow(x)
```

```
[1] 17
```

```
ncol(x)
```

```
[1] 5
```

```
dim(x)
```

```
[1] 17  5
```

```
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
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
```

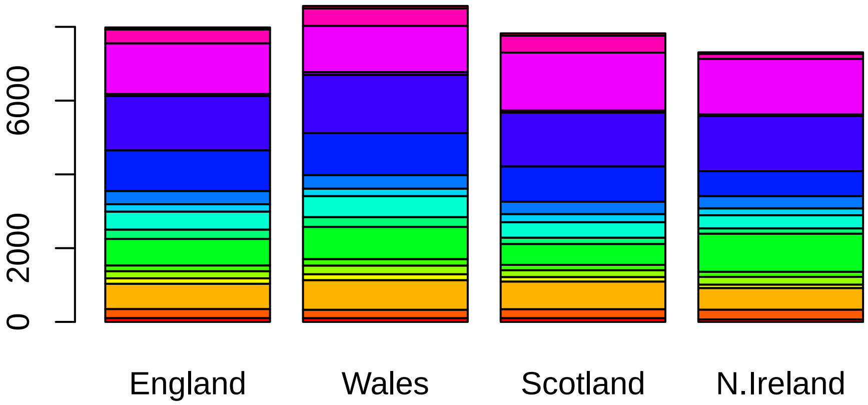
```r
dim(x)
```

```
[1] 17  4
```

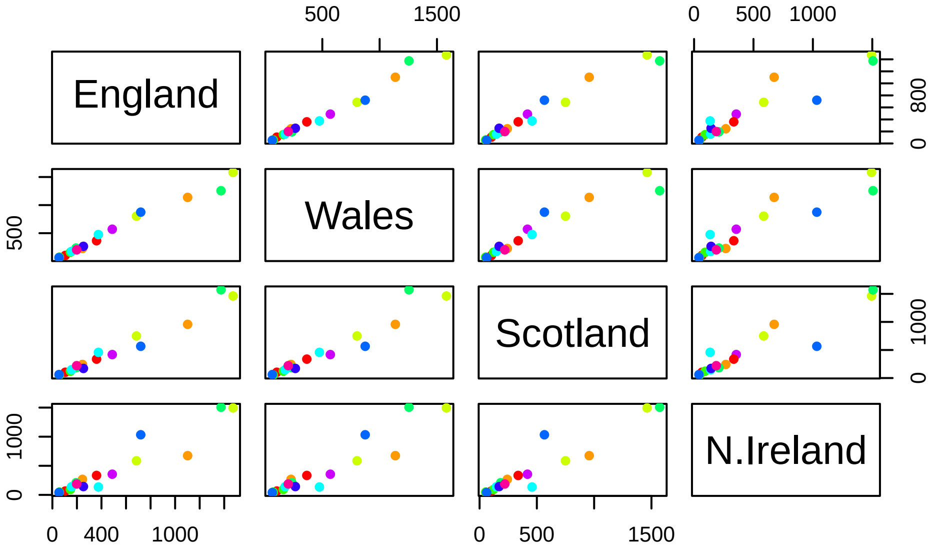
```r
x <- read.csv(url, row.names=1)
head(x)
```

```
               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
```

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



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



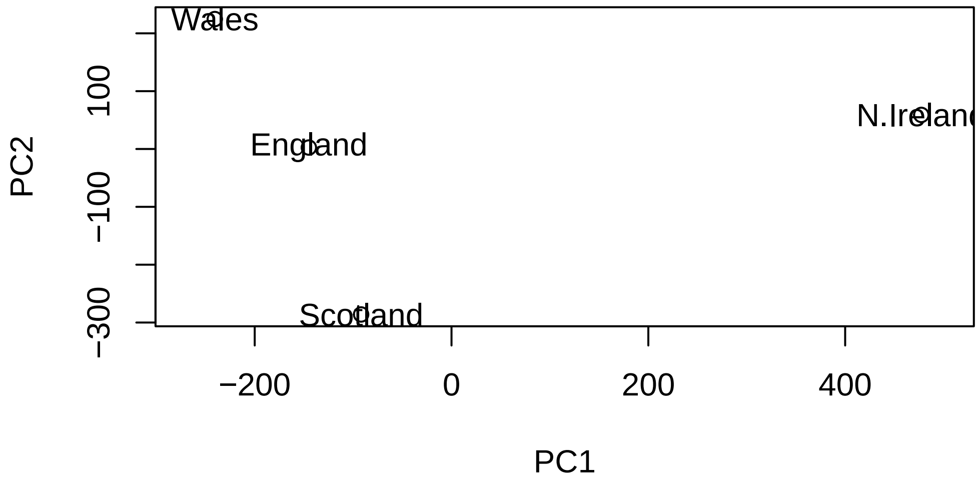
```
pca <- prcomp(t(x))
summary(pca)
```

```
Importance of components:
                            PC1      PC2      PC3       PC4
Standard deviation     324.1502 212.7478 73.87622 5.552e-14
Proportion of Variance   0.6744   0.2905  0.03503 0.000e+00
Cumulative Proportion    0.6744   0.9650  1.00000 1.000e+00
```

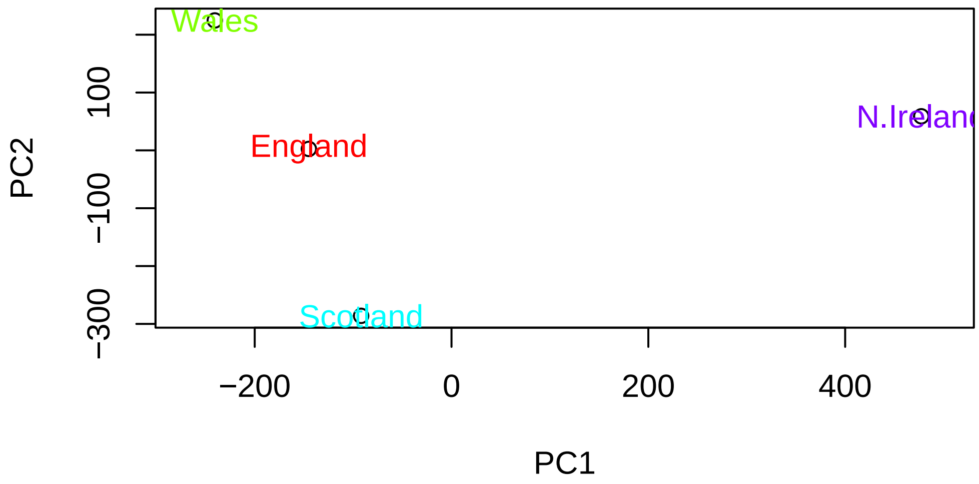
```
pca$x
```

```
                 PC1         PC2         PC3           PC4
England   -144.99315    2.532999 -105.768945  1.042460e-14
Wales     -240.52915  224.646925   56.475555  9.556806e-13
Scotland   -91.86934 -286.081786   44.415495 -1.257152e-12
N.Ireland  477.39164   58.901862    4.877895  2.872787e-13
```

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



```
plot(pca$x[,"PC1"], pca$x[,"PC2"], xlab="PC1", ylab="PC2", xlim=c(-270,500))
text(pca$x[,1], pca$x[,2], colnames(x), col = rainbow(4))
```



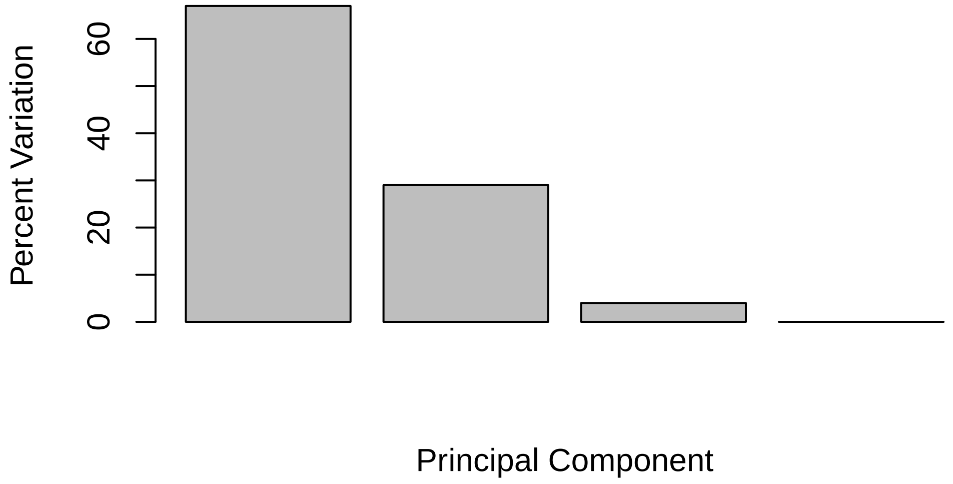
```
v <- round( pca$sdev^2/sum(pca$sdev^2) * 100 )
v
```

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

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

```
                             PC1       PC2      PC3          PC4
Standard deviation     324.15019 212.74780 73.87622 5.551558e-14
Proportion of Variance   0.67444   0.29052  0.03503 0.000000e+00
Cumulative Proportion    0.67444   0.96497  1.00000 1.000000e+00
```

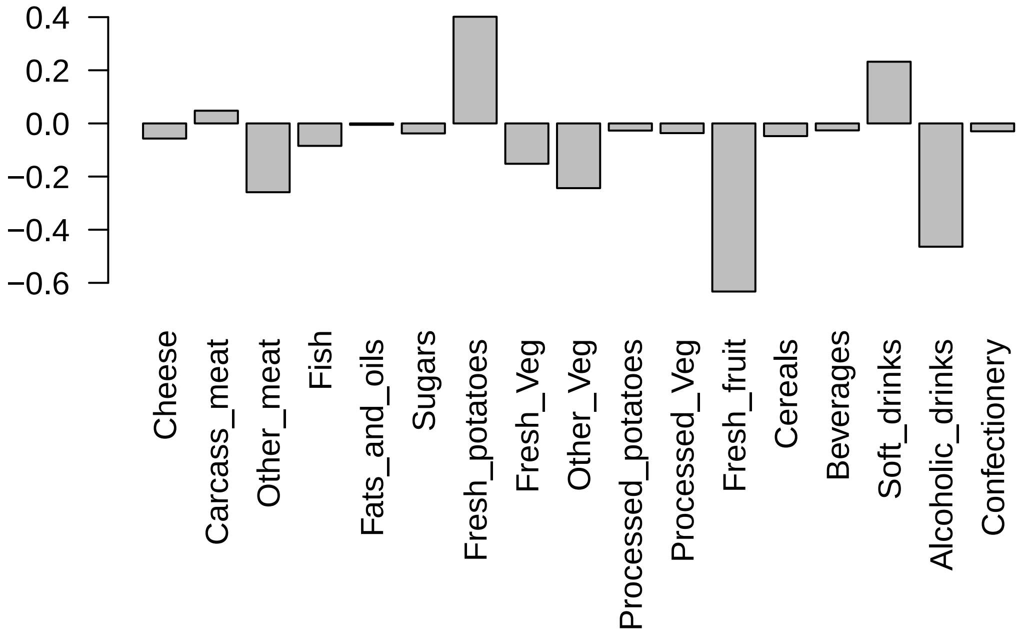
```
barplot(v, xlab="Principal Component", ylab="Percent Variation")
```



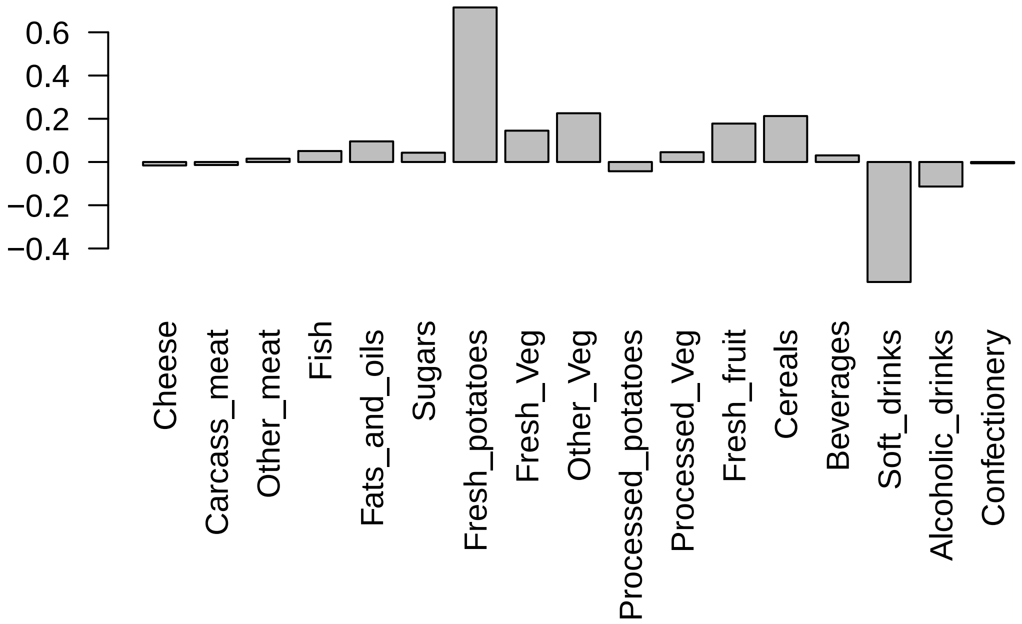
```
pca$rotation
```

```
                            PC1          PC2         PC3          PC4
Cheese             -0.056955380 -0.016012850 -0.02394295 -0.537717586
Carcass_meat        0.047927628 -0.013915823 -0.06367111  0.827327785
Other_meat         -0.258916658  0.015331138  0.55384854 -0.054885657
Fish               -0.084414983  0.050754947 -0.03906481 -0.017195729
Fats_and_oils      -0.005193623  0.095388656  0.12522257  0.039441462
Sugars             -0.037620983  0.043021699  0.03605745  0.002788534
Fresh_potatoes      0.401402060  0.715017078  0.20668248 -0.030319813
Fresh_Veg          -0.151849942  0.144900268 -0.21382237 -0.051070911
Other_Veg          -0.243593729  0.225450923  0.05332841  0.060355222
Processed_potatoes -0.026886233 -0.042850761  0.07364902  0.003645959
Processed_Veg      -0.036488269  0.045451802 -0.05289191 -0.003672450
Fresh_fruit        -0.632640898  0.177740743 -0.40012865  0.031359988
Cereals            -0.047702858  0.212599678  0.35884921  0.073618516
Beverages          -0.026187756  0.030560542  0.04135860 -0.005163295
Soft_drinks         0.232244140 -0.555124311  0.16942648 -0.009904437
Alcoholic_drinks   -0.463968168 -0.113536523  0.49858320  0.088180533
Confectionery      -0.029650201 -0.005949921  0.05232164  0.004029923
```

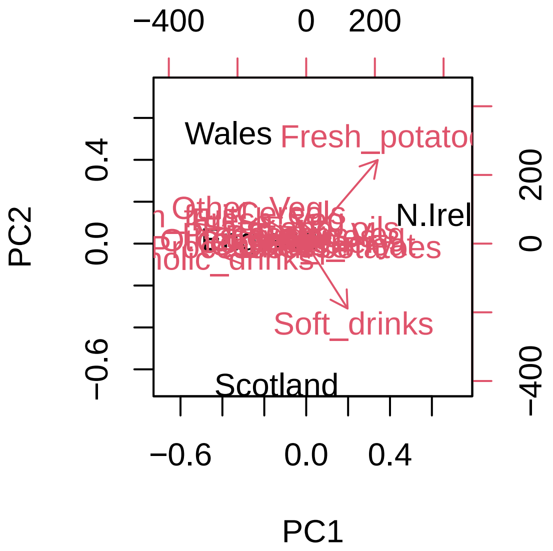
```
## Lets focus on PC1 as it accounts for > 90% of variance
par(mar=c(10, 3, 0.35, 0))
barplot( pca$rotation[,1], las=2 )
```



```
## Lets look at PC2
par(mar=c(10, 3, 0.35, 0))
barplot( pca$rotation[,2], las=2 )
```



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



now we will do some RNA seq analysis

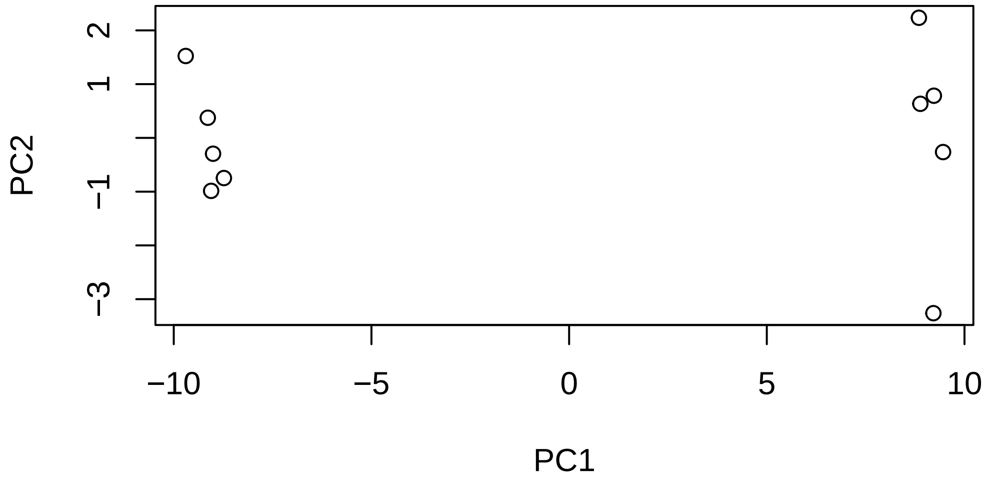
```
url2 <- "https://tinyurl.com/expression-CSV"
rna.data <- read.csv(url2, row.names=1)
head(rna.data)
```

```
       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
```

```
dim(rna.data)
```

```
[1] 100  10
```

```
## 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
pca$x
```

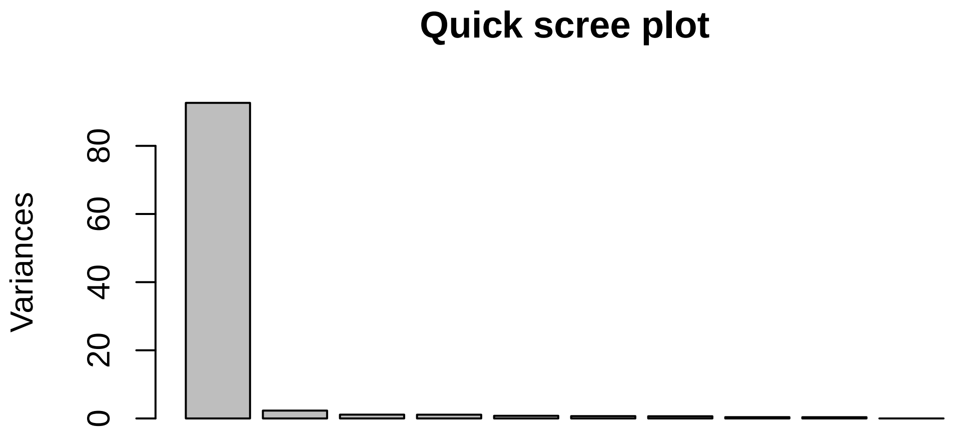
```
         PC1        PC2        PC3        PC4        PC5        PC6
wt1 -9.697374  1.5233313 -0.2753567  0.7322391 -0.6749398 -1.1823860
wt2 -9.138950  0.3748504  1.0867958 -1.9461655  0.7571209  0.4369228
wt3 -9.054263 -0.9855163  0.4152966  1.4166028  0.5835918 -0.6937236
wt4 -8.731483 -0.7468371  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
ko2  9.213885 -3.2607503  0.2287292 -0.7658122 -0.4922849 -0.9170241
ko3  9.458412 -0.2636283 -1.5778183  0.2433549  0.3654124  0.5837724
ko4  8.883412  0.6339701  1.5205064  0.7760158  1.2158376  0.1446094
ko5  9.225673  0.7845635  0.0103574  0.9017667 -0.3860869  0.8186668
            PC7         PC8         PC9         PC10
wt1  0.24446614  1.03519396  0.07010231 3.031594e-15
wt2  0.03275370  0.26622249  0.72780448 2.383634e-15
wt3  0.03578383 -1.05851494  0.52979799 3.139973e-15
wt4  0.52795595 -0.20995085 -0.50325679 3.202096e-15
wt5 -0.83227047 -0.05891489 -0.81258430 2.996904e-15
ko1  0.07864392 -0.94652648 -0.24613776 3.551480e-15
ko2 -0.30945771  0.33231138 -0.08786782 3.443451e-15
ko3  1.43723425  0.14495188  0.56617746 3.165891e-15
```

```
ko4  0.35073859  0.30381920 -0.87353886 2.978853e-15
ko5 -1.56584821  0.19140827  0.62950330 2.910146e-15
```

```
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 3.327e-15
Proportion of Variance 0.00385 0.00364 0.000e+00
Cumulative Proportion  0.99636 1.00000 1.000e+00
```

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

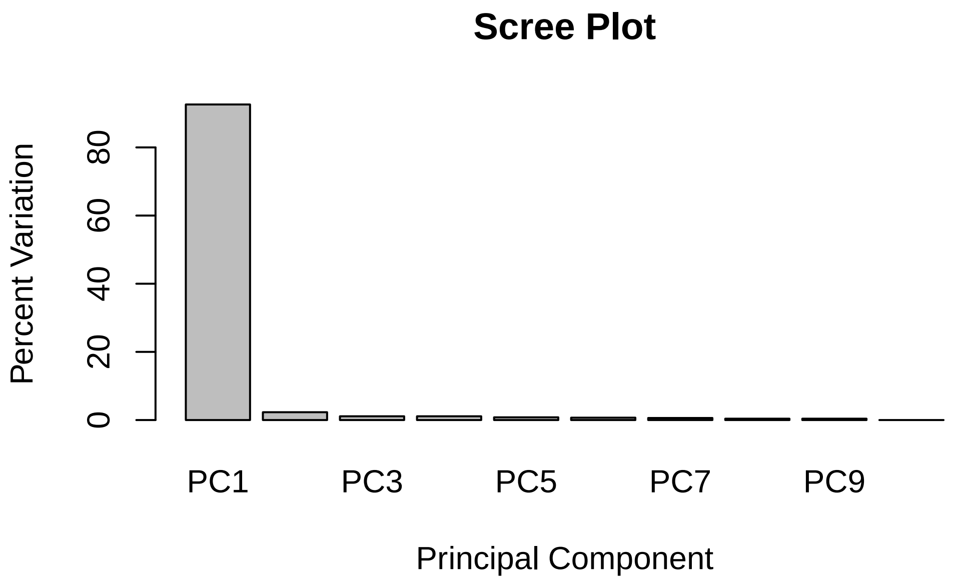


```
## Variance captured per PC
pca.var <- pca$sdev^2
```

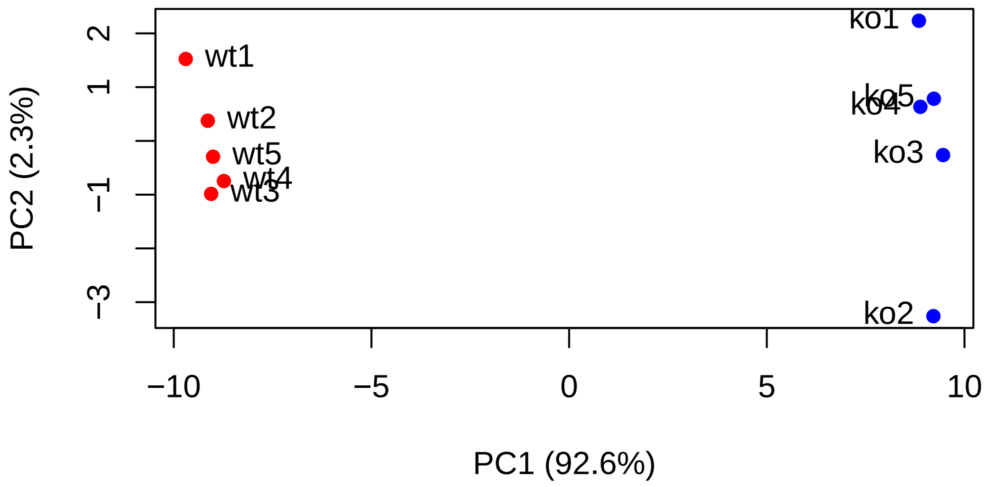
```r
## Percent variance is often more informative to look at
pca.var.per <- round(pca.var/sum(pca.var)*100, 1)
pca.var.per
```

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

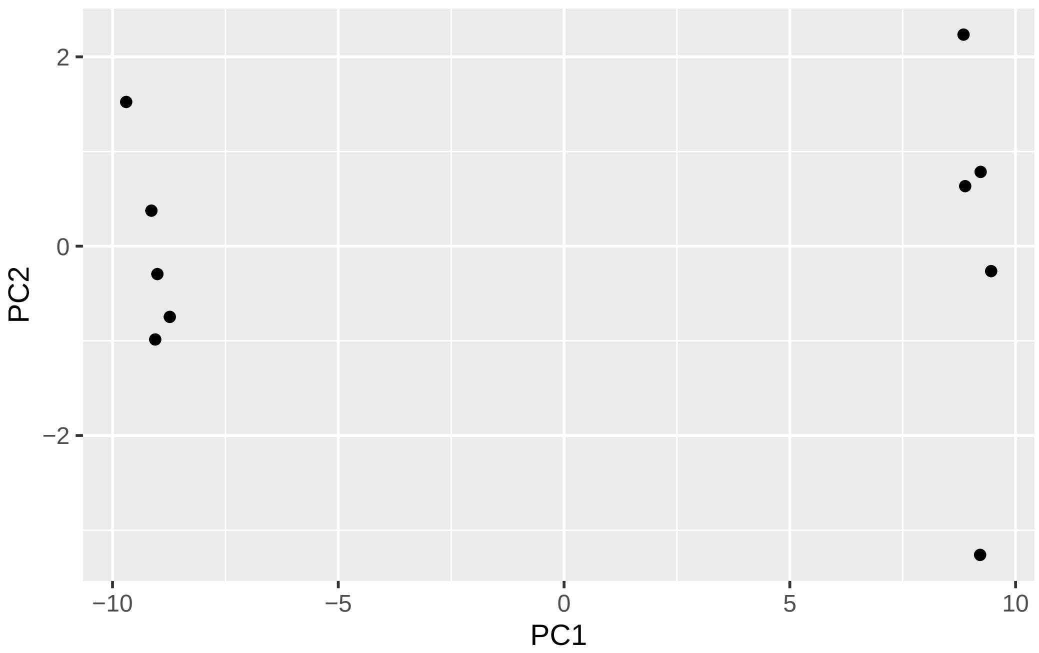
```r
barplot(pca.var.per, main="Scree Plot",
        names.arg = paste0("PC", 1:10),
        xlab="Principal Component", ylab="Percent Variation")
```



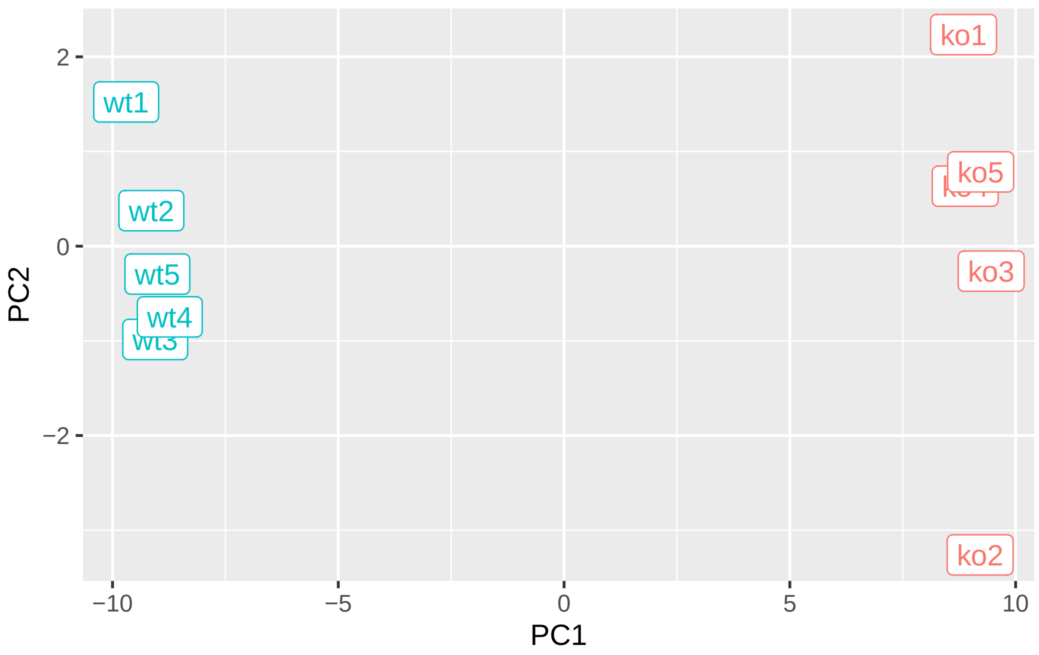
```r
# A vector of colors for wt and ko samples
colvec <- colnames(rna.data)
colvec[grep("wt", colvec)] <- "red"
colvec[grep("ko", colvec)] <- "blue"

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)))
```



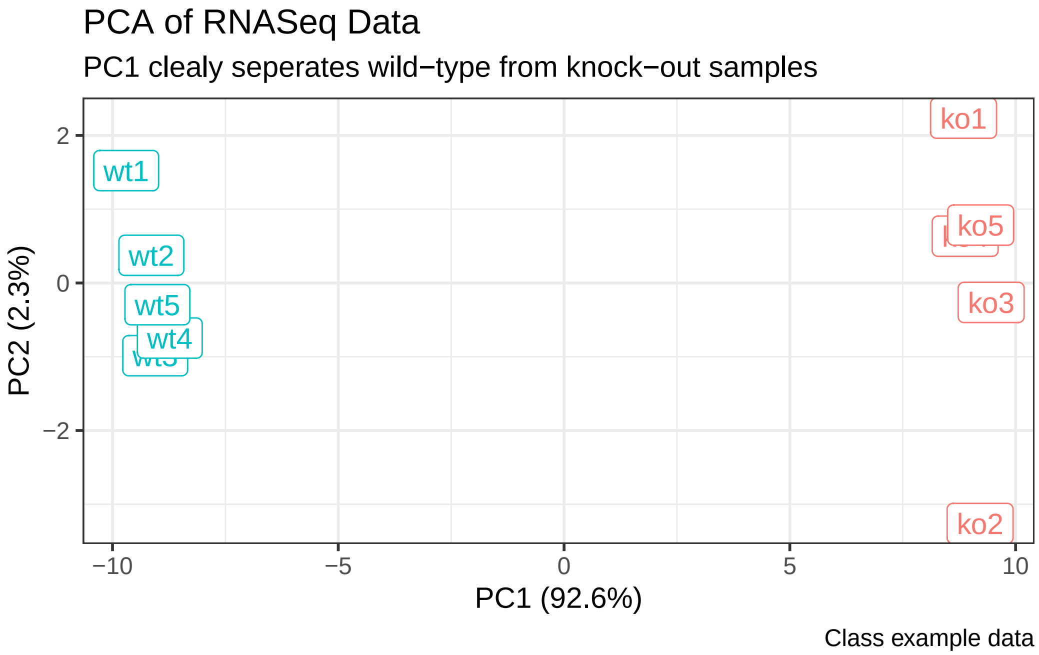
```
library(ggplot2)

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

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



```
# 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()
```



```r
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"
```