# Exploratory Data Analysis

Harlan Gillespie

29/08/2021

##Dependencies

```
library(SummarizedExperiment)
library(gplots)
library(DESeq2)
```

## Plot colour palette

```
tropical=  c('darkorange', 'dodgerblue', 'hotpink', 'limegreen', 'yellow')
palette(tropical)
par(pch=19)
```

## Load the dataset

Firstly, we can load the dataset obtained through the previous steps in the capstone project. These steps are outlined in the README.md file.

```
colData <- DataFrame(read.delim("~/Coursera Capstone/PData.txt", stringsAsFactors=TRUE))
counts <- as.matrix(read.delim("~/Coursera Capstone/featureCount-data-ENTREZ.txt"))
rownames(counts) = counts[,1]
counts = counts[,-1]
```

Next, it is good practice to organise the dataset into a Summarized Experiment object.

```
data.se = SummarizedExperiment(assays = list(counts = counts),  colData = colData)
```

Next, we can extract the count data matrix (edata) and the phenotype DataFrame (pdata)

```
edata = assays(data.se)[[1]]
pdata = colData(data.se)
```

## Exploratory Data Analysis

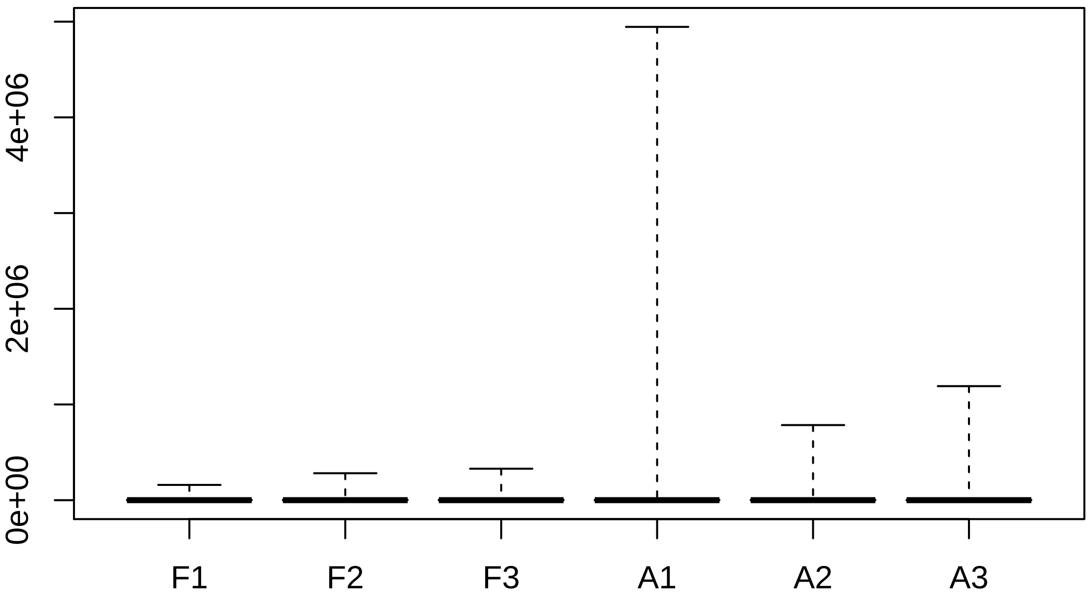
We can show the sex and race of the samples using a table. There seems to be a bias towards african american (AA) samples over hispanic (HISP).

```
table(pdata$Sex, pdata$Race)
```

```
##
##          AA HISP
##   female  2    0
##   male    3    1
```

Next we can look at the overall distribution of the dataset using boxplots.

```
boxplot(edata, col = 2, range = 0)
```



It seems that outliers in the sample A1 that may make the rest of the dataset difficult to visualise. A data transformation is therefore necessary.

```
summary(edata)
```

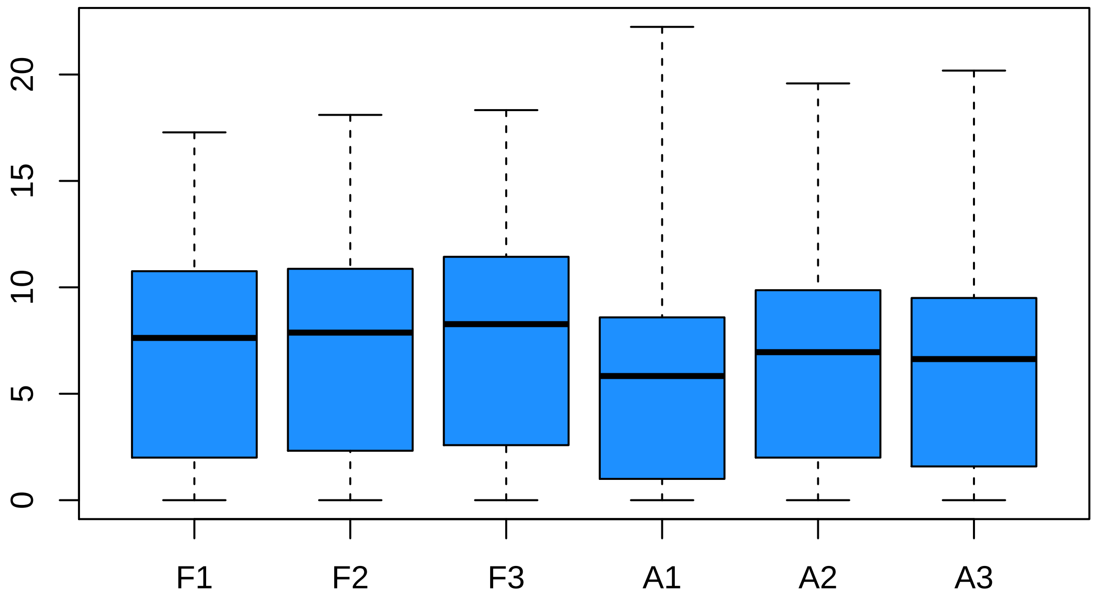
```
##       F1               F2                 F3                A1
##  Min.   :     0   Min.   :     0.0   Min.   :     0   Min.   :      0
##  1st Qu.:     3   1st Qu.:     4.0   1st Qu.:     5   1st Qu.:      1
##  Median :   196   Median :   233.5   Median :   308   Median :     56
##  Mean   :  1785   Mean   :  2048.7   Mean   :  3060   Mean   :   1179
##  3rd Qu.:  1731   3rd Qu.:  1871.8   3rd Qu.:  2768   3rd Qu.:    384
##  Max.   :159431   Max.   :281475.0   Max.   :328682   Max.   :4945442
##       A2                  A3
##  Min.   :     0.0   Min.   :      0
##  1st Qu.:     3.0   1st Qu.:      2
##  Median :   123.0   Median :     98
##  Mean   :  1311.5   Mean   :   1134
##  3rd Qu.:   931.8   3rd Qu.:    723
##  Max.   :784719.0   Max.   :1191660
```

A1’s larger variability, demonstrated by the boxplots and the summary table, may be due to its lower RNA integrity number. This will be something to monitor as the project continues.

Next, a log2 transformation can be tested and evaluated using the same boxplot method.

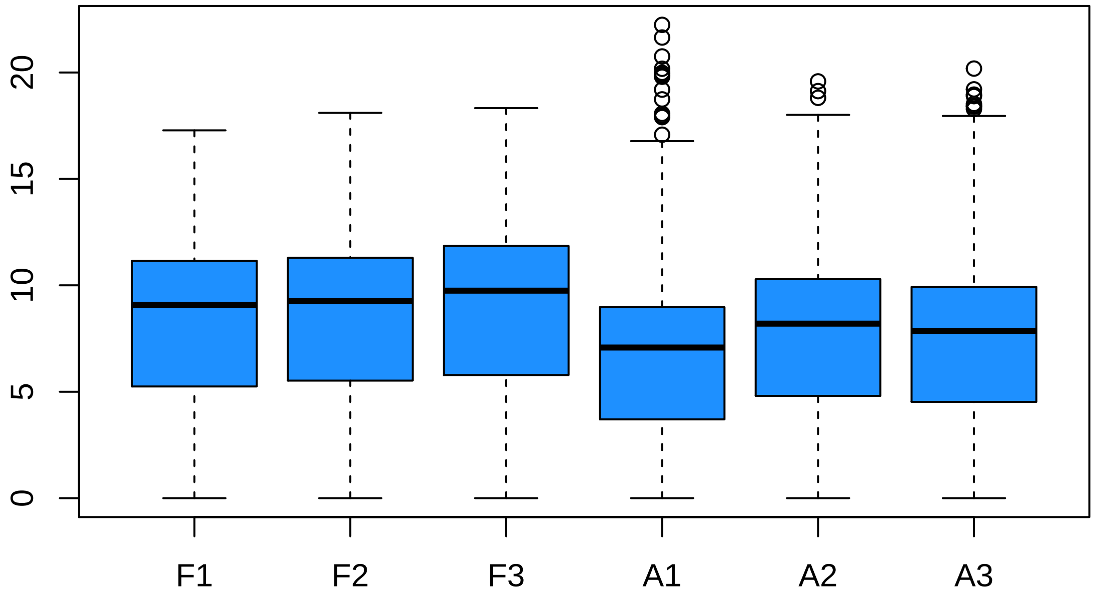
Removing genes that had little or no expression for all datasets is also a good practice.

```
boxplot(log2(edata+1),col=2,range=0)
```

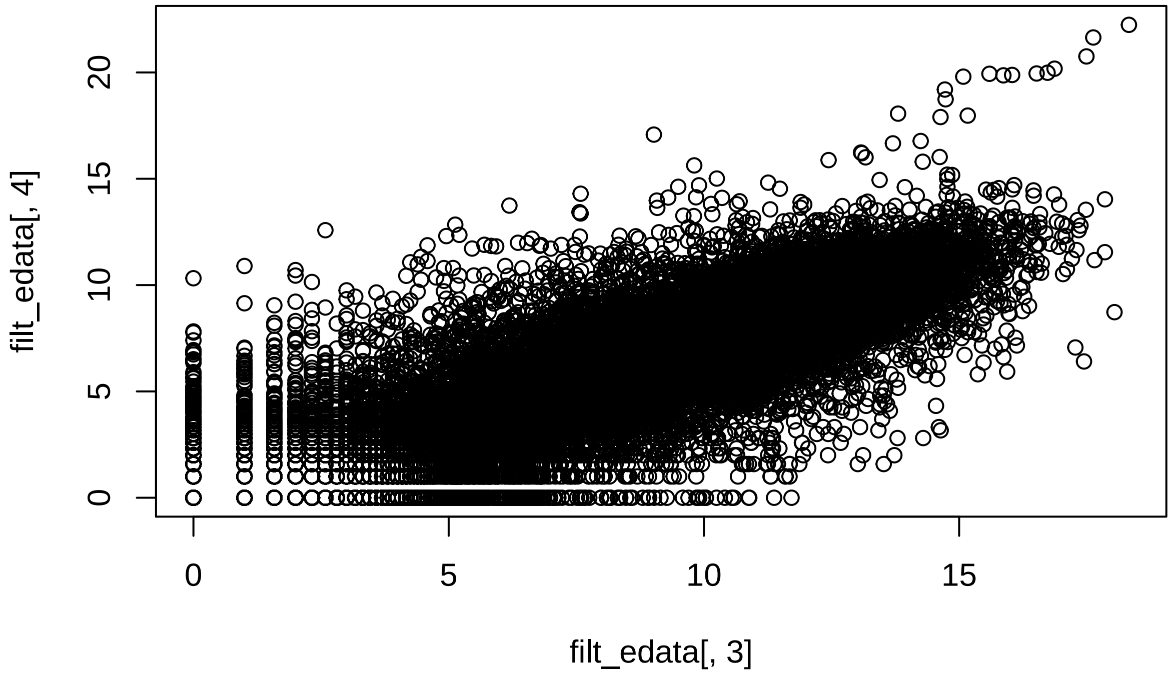


The data is now much more clear and easier to visualise.

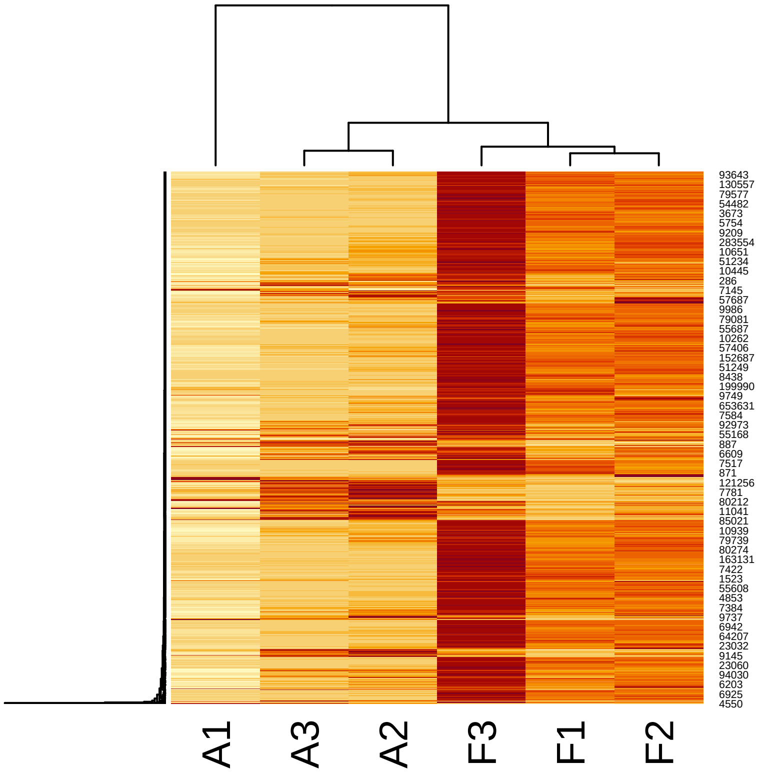
```
filt_edata = log2(edata[rowMeans(edata)>1,]+1)
boxplot(as.matrix(filt_edata),col=2)
```



```
plot(filt_edata[,3], filt_edata[,4])
```

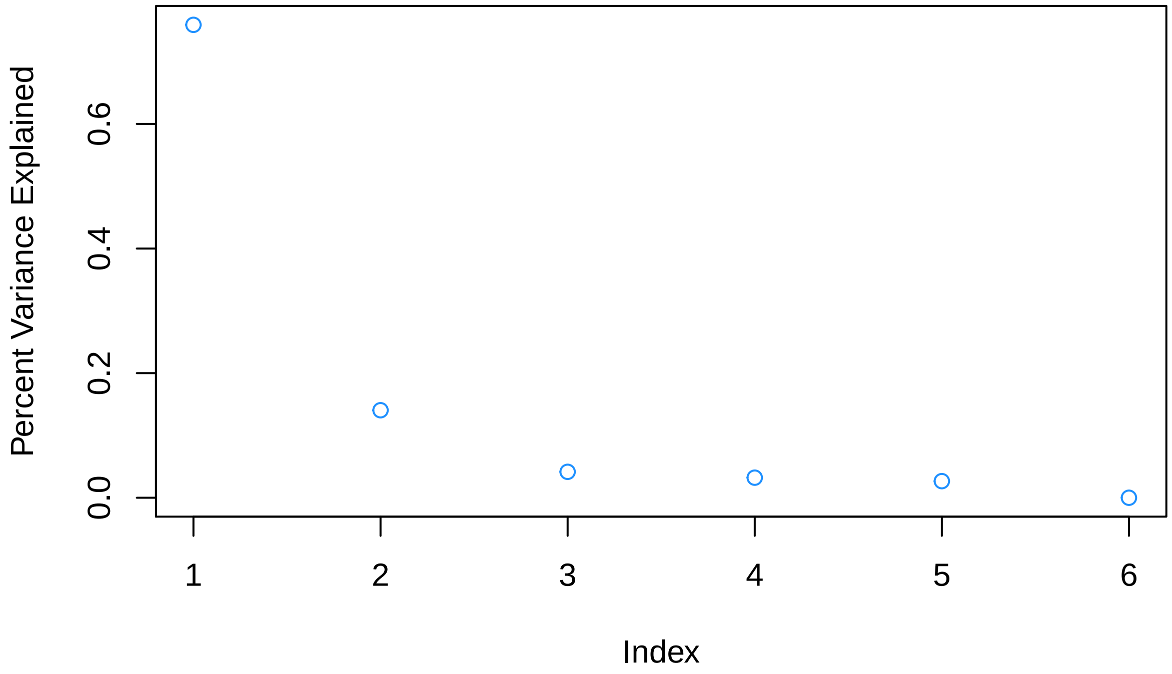


```
heatmap(edata[rowMeans(edata)>500,], Rowv = NULL)
```

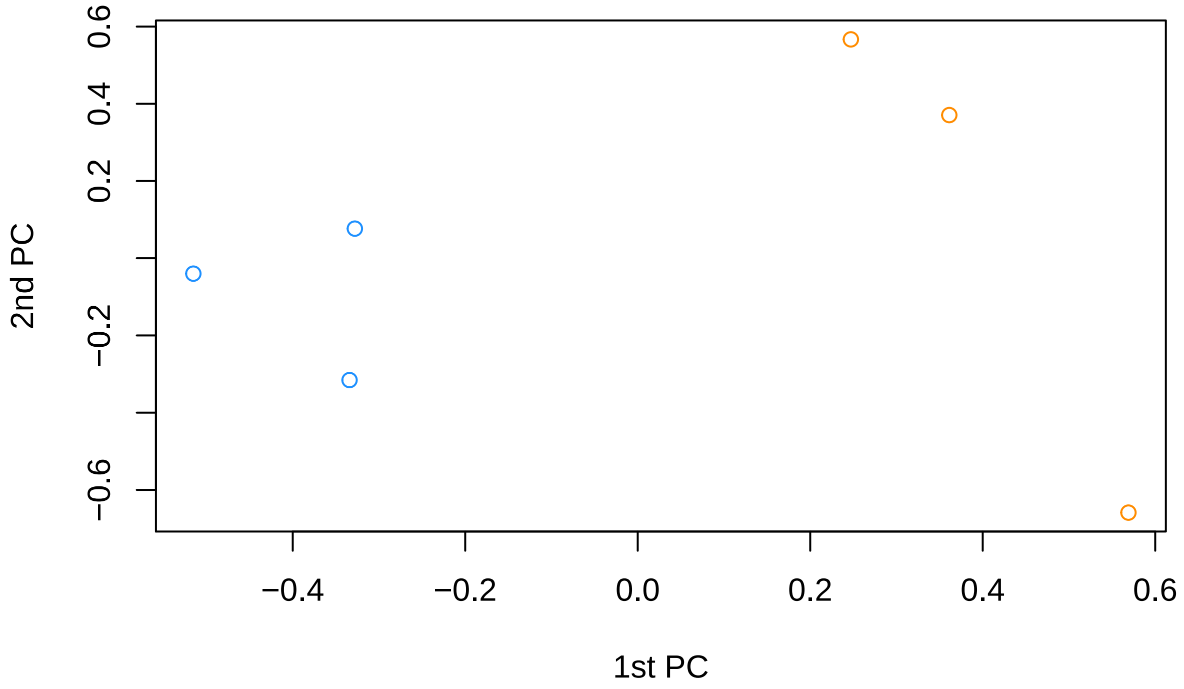


### Principal Component Analysis

```
edata_centered = filt_edata - rowMeans(filt_edata)
svd1 = svd(edata_centered)
plot(svd1$d^2/sum(svd1$d^2),ylab="Percent Variance Explained",col=2)
```

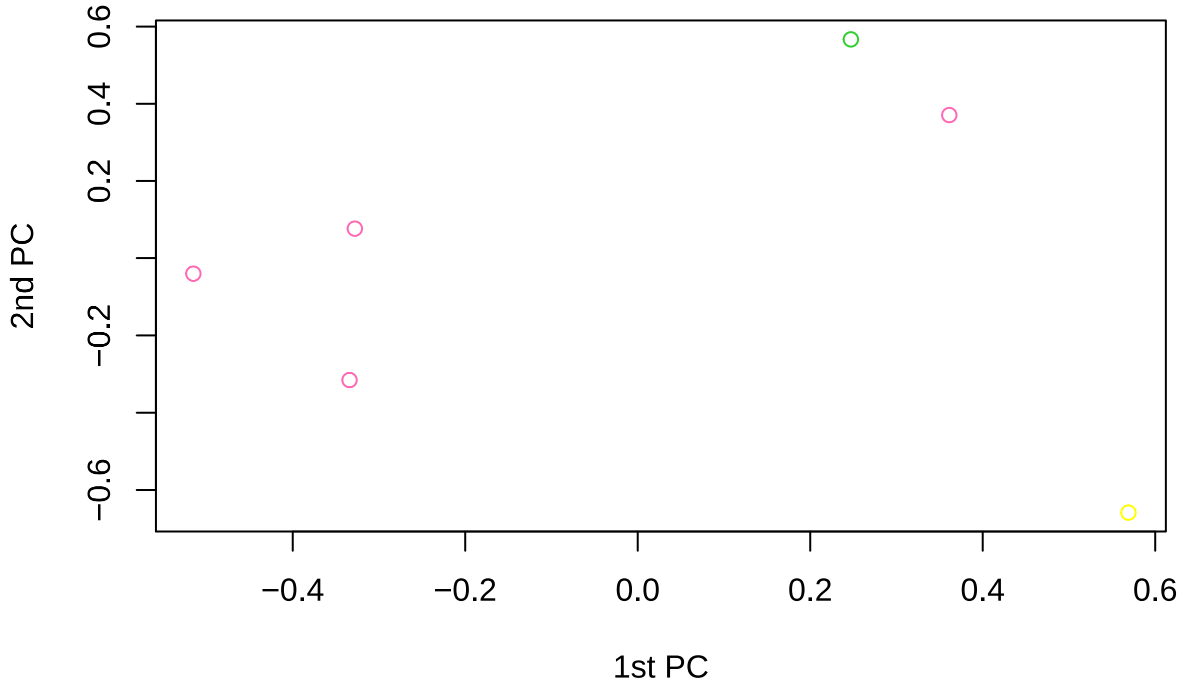


```
plot(svd1$v[,1],svd1$v[,2],ylab="2nd PC",xlab="1st PC", col=as.numeric(pdata$Group))
```



This PCA plot shows that the first PC forms two clusters predominantly explained by their phenotype (group).

```
plot(svd1$v[,1],svd1$v[,2],ylab="2nd PC",xlab="1st PC", col=as.numeric(pdata$RIN))
```



This second PCA plot shows the second PC is correlated with the RIN of each sample.