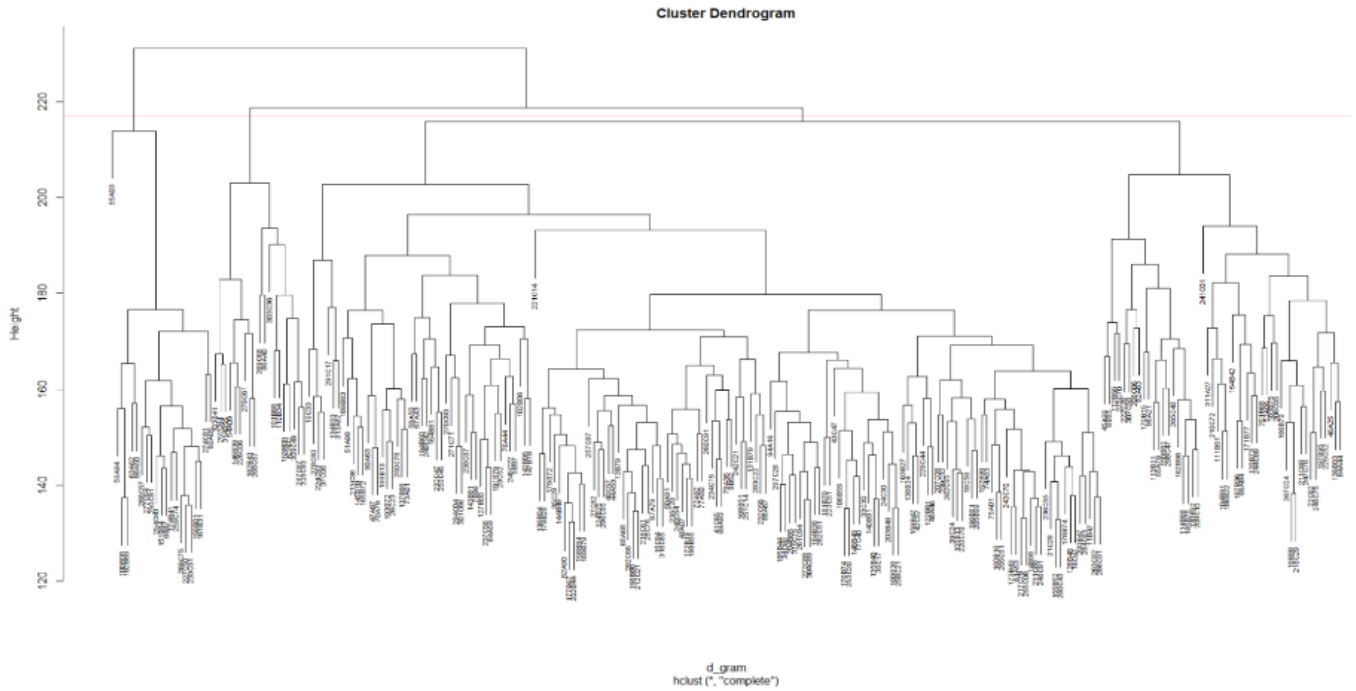
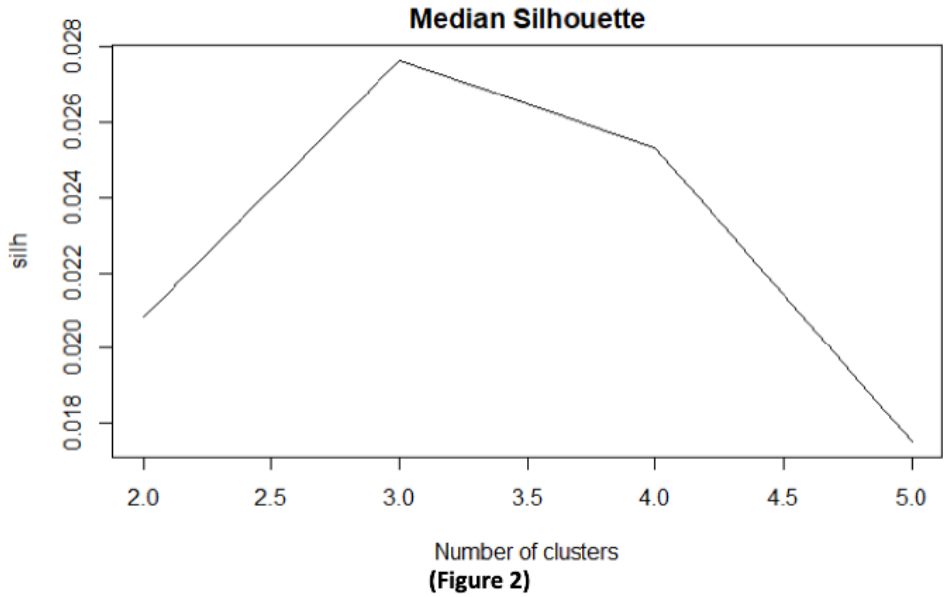
Analysis of breast cancer and subtypes (R programming):

When we initially load the breast cancer data, dendrogram is used to represent the genes in a hierarchy. If we observe Figure 1 there are three clusters which can be determined by the AB line.



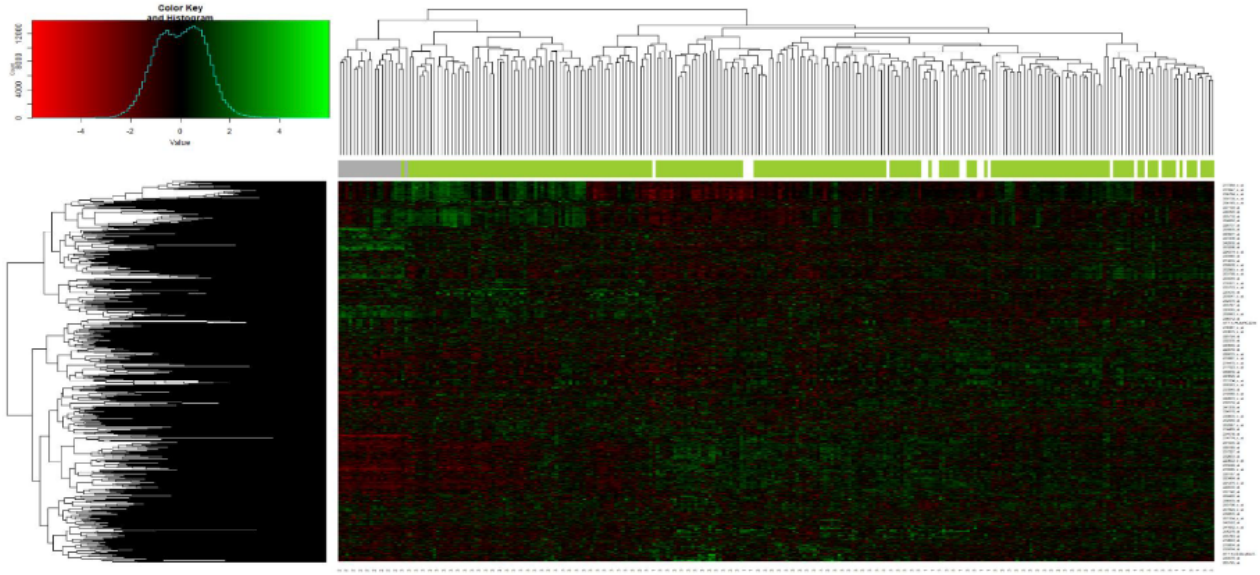
(Figure 1)

In order to obtain optimal number of clusters based on the results from dendrogram and to avoid any issue of overfitting/under-fitting, Silhouette method is Utilised. If we consider the highest data point in Figure 2 which coincides the value on y- axis, it indicates the number of clusters.



(Figure 2)

As Silhouette method lacks the capacity to display the gene clusters, we are required to generate a heat map of the breast cancer gene data. With the help of heat map, it's easier to understand the gene data of breast cancer for a person, who has insufficient knowledge regarding the statistical analysis behind it. Dark red color visualises unregulated genes and on the contrary green color visualizes down-regulated genes.

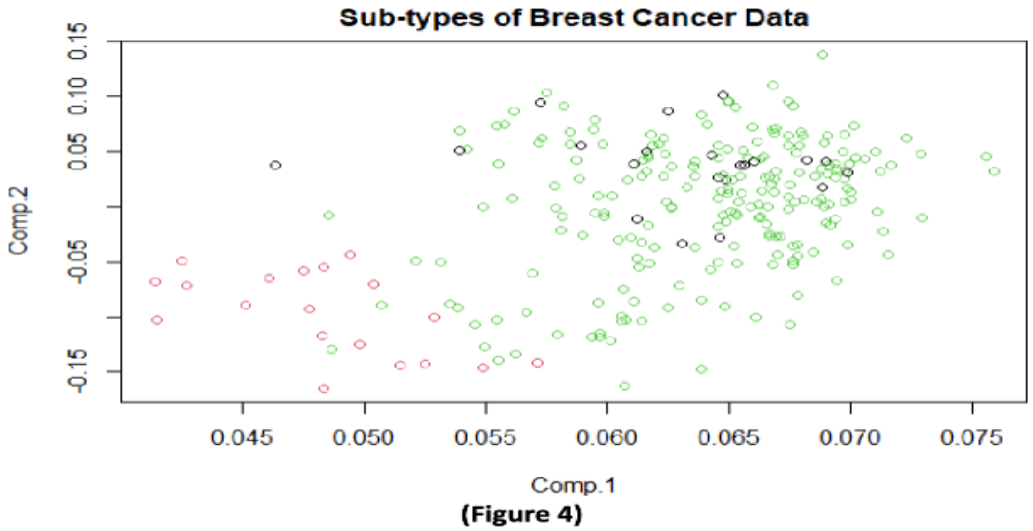



```
[1,] "100B08" "white"
[2,] "101B88" "grey68"
[3,] "102B06" "yellowgreen"
[4,] "103B41" "white"
[5,] "104B91" "grey68"
[6,] "105B13" "yellowgreen"
```

(Figure 3)

Since heat map is unable to explain which genes are highly expressed than the other genes, we need to perform principal component analysis. To gain useful information, filtering of some data attribute is necessary to decrease the dimensions.

Figure 4 enables us to differentiate between the three clusters, according to their subtypes of breast cancer. We can observe some close resemblance between few data points as they appear to be overlapped, due to similarity in the breast cancer dataset.



(Figure 4)

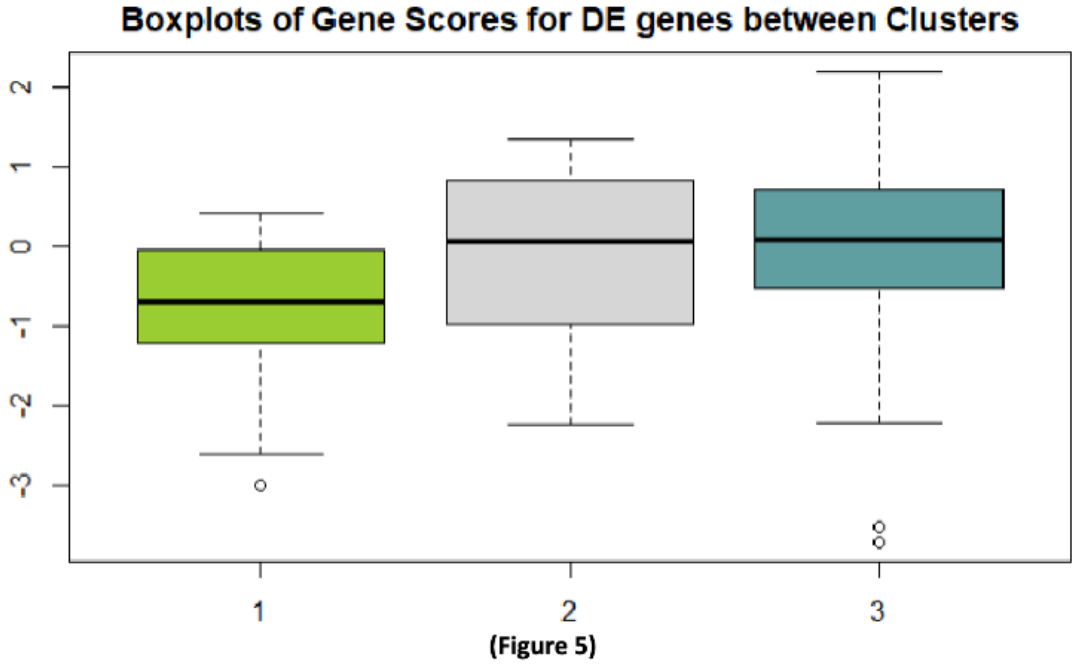
Principal component analysis doesn't provide us with q-values, so we are required to implement differential expression analysis. Firstly, we specify the design matrix model and then we construct a differential expression object which provides us with q-values containing false discovery rate equal to 0.5 and its corresponding gene score.

As differential expression analysis falls short to provide us with hazard ratio, we need to perform Cox regression. To check model accuracy concordance is a parameter which indicates, how correct our model performs based on the given breast cancer data. Figure 6 indicates it as 0.955 which is quite excellent as 0.7 and above is considered as good.

```
  n= 236, number of events= 55
   (15 observations deleted due to missingness)

                     coef  exp(coef)   se(coef)      z Pr(>|z|)
gene.score     -5.387e-02  9.476e-01  1.694e-01 -0.318  0.75047
age             2.954e-02  1.030e+00  1.250e-02  2.364  0.01808 *
tumor_size_mm   3.614e-02  1.037e+00  1.462e-02  2.472  0.01342 *
LNstatusLN?    -3.536e-02  9.653e-01  3.946e+04  0.000  1.00000
LNstatusLN+     7.975e-01  2.220e+00  2.958e-01  2.696  0.00701 **
eventTRUE       2.331e+01  1.331e+10  5.148e+03  0.005  0.99639
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

               exp(coef) exp(-coef) lower .95 upper .95
gene.score     9.476e-01  1.055e+00    0.6799     1.321
age            1.030e+00  9.709e-01    1.0051     1.056
tumor_size_mm  1.037e+00  9.645e-01    1.0075     1.067
LNstatusLN?    9.653e-01  1.036e+00    0.0000       Inf
LNstatusLN+    2.220e+00  4.505e-01    1.2433     3.964
eventTRUE      1.331e+10  7.513e-11    0.0000       Inf

Concordance= 0.955  (se = 0.008 )
Likelihood ratio test= 250.8  on 6 df,   p=<2e-16
Wald test            = 7.45  on 6 df,   p=0.3
Score (logrank) test = 363.6  on 6 df,   p=<2e-16
```

(Figure 6)

As differential expression analysis lacks to share the plot of gene expression, we are expected to generate a boxplot. Figure 5 displays the boxplot for three clusters with their gene scores which are given by differential expression genes.



(Figure 5)