STA5MB

ASSIGNMENT 2

APPLIED PROJECT REPORT BY:

ARAVIND MADINENI

20500070

NOTE: “THIS PROJECT IS SOLELY MY OWN WORK. I HAVE NOT COPIED FROM ANYONE ELSE”

BACKGROUND:

Cancer one of the most dangerous diseases is a disease of cells which occurs when the abnormal cells in any part of a body grow and multiply in an uncontrolled manner. These cells if not controlled could possibly invade and destroy the surrounding healthy and well preserved cells, tissues, and organs included. Cancer cells can survive and sustain in almost any type of cells. Cancer cells may be benign or malignant. Benign cells are not dangerous as they do not spread to the encircling areas. The other type of cells malignant cells are dangerous cells which spreads to the surrounding areas and even other parts and this process of spreading of cancer cells is called metastasis. There are different types of cancers with each type having the own method of diagnosis and treatment. Some of the most common cancers are prostrate cancer, breast cancer, skin cancer, lymphomas etc.

BREAST CANCER:

Is one of the types of cancer where the cells in the breast region grow abnormally without any control which eventually form lumps or tumors which if untreated spreads from breasts to other parts of the body such as lungs or bones. This type of cancer is more prevalent in women although in rare cases sometimes seen in men. Breast cancer is divided into four sub-types by most of the research studies. They are:

- Luminal A
- Luminal B
- Triple negative/basal-like
- HER2-enriched

Luminal A : This luminal cells look mostly like cells of breast cancers which start in the inner cells lining the mammary ducts. These tumors tend be estrogen receptor positive(ER-positive), HER2 receptor-negative(HER2-negative), tumor grade 1 or 2.

Luminal B: These cells also look like the most cells like the ones in the breast cancers

like luminal A sub type. Luminal B tumors tends to be ER-positive. They may even be HER2-negative or HER2-positive.

Triple negative/basal-like: These cancers are estrogen receptor-negative(ER-negative), Progesterone receptor-negative(PR-negative), HER2-negative. About 15-20% of breast cancers are of this sub-type.

- HER2-enriched: These tumors tend to be ER-negative, PR-negative, Lymph node-positive, Poorer tumor grade. Most of these cancers are HER2-positive and about 30% are HER2-negagtive.

METHODS:

The statistical methods used in process are:

- PCA analysis
- Survival analysis
- Hierarchial clustering

PCA ANALYSIS:

Principal Component Analysis, generally short notified as PCA, is a dimension reduction method which is used for reducing the dimensions of the given data set by transforming a large set of variables into small ones which even still contains most of the information from the large data. Generally reducing the number of variables from a data set is likely to cost us the accuracy but for the sake of simplicity a little accuracy trade of would do good for the process. This PCA analysis can be performed using the built in R functions prcomp() and princomp(). There are two general methods for performing the PCA analysis in r:

- Spectral decomposition that examines the covariances / correlations between the variables

- Singular value decomposition which examines covariances/ correlations between the variables

The princomp() function we use in the process uses the spectral decomposition method.

Finally it can be concluded that PCA is a linear projection algorithm, which means that most cardinal apprehension of variations which are captured are restricted to linear functions. Sometimes if the relationship between the features are not captured by PCA, then in such cases there is an alternate method required for dimension reduction, such as the more modern approach of auto-encoders, that are direct generalizations of PCA.

SURVIVAL ANALYSIS:

Also called as event history analysis in social sciences, deals with the time until any occurrence of an event of interest. Sometimes the failure time might not be observed within the germane time period, producing so called censored observations. This is a branch of statistics which is concerned with the estimation of the expected time until some event of interest occurs.

ESTIMATION OF SURVIVAL DISTRIBUTION:

- KAPLAN-MEIER :

The survfit function from the survival package computes the kaplan-meier estimator for the censored data

- NON-PARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION:

Generally short notified as NPMLE. The Icens package provides numerous ways for computing the NPMLE of the survival distribution for various censoring themes.

- PARAMETRIC:

Different packages work different ways. The fitdistrplus package allows to fit an univariate distribution by maximum likelihood. Data can be interval censored. The vitality package provides procedures fir fitting models

in the vitality family of mortality models.

The survival analysis can be considered as kind of regression. At one instance there is a concern with time at which something occurs, that can be considered as a response variable and there also exists some external factors such as genetics or phenotypical variables. In survival analysis the incompleteness of the data may be due to phenomenon of censoring which is a catchall term that refers to a situation where some measurement is imperfectly measured.

HIERARCHIAL CLUSTERING:

Is a kind of linkage based clustering algorithm where we decide on distance d and linkage l. For each pair of clusters we compute the linkage between the clusters, combine the pair of clusters with the smallest linkage together into a new clusters. If we have only one cluster left we stop else we perform again from step 1. If we have more number of clusters then we do require a method for choosing some optimal clusters which we consider silhouette method.

DIFFERENTIAL GENE EXPRESSION ANALYSIS:

Generally shortly written as DGE is the process where the normalized read count data is taken and then statistical analysis is performed to find or discover any quantitative changes in expression levels between the experimental groups. There are different methods for DGE such as edgeR, DESeq which are based on negative binomial distributions, bayesian approaches such as baySeq and EBSeq. For this analysis Bioconductor has lot of packages that support the analysis of high-throughput sequence data, including the RNA sequencing. In this applied project specifically LIMMA package is used for analyzing the gene expression data arising from either micro array or RNA-sequence technologies. This package provides the ability for analyzing comparisons between number of RNA targets simultaneously. Limma is able to handle both single channel and two-color micro

arrays.

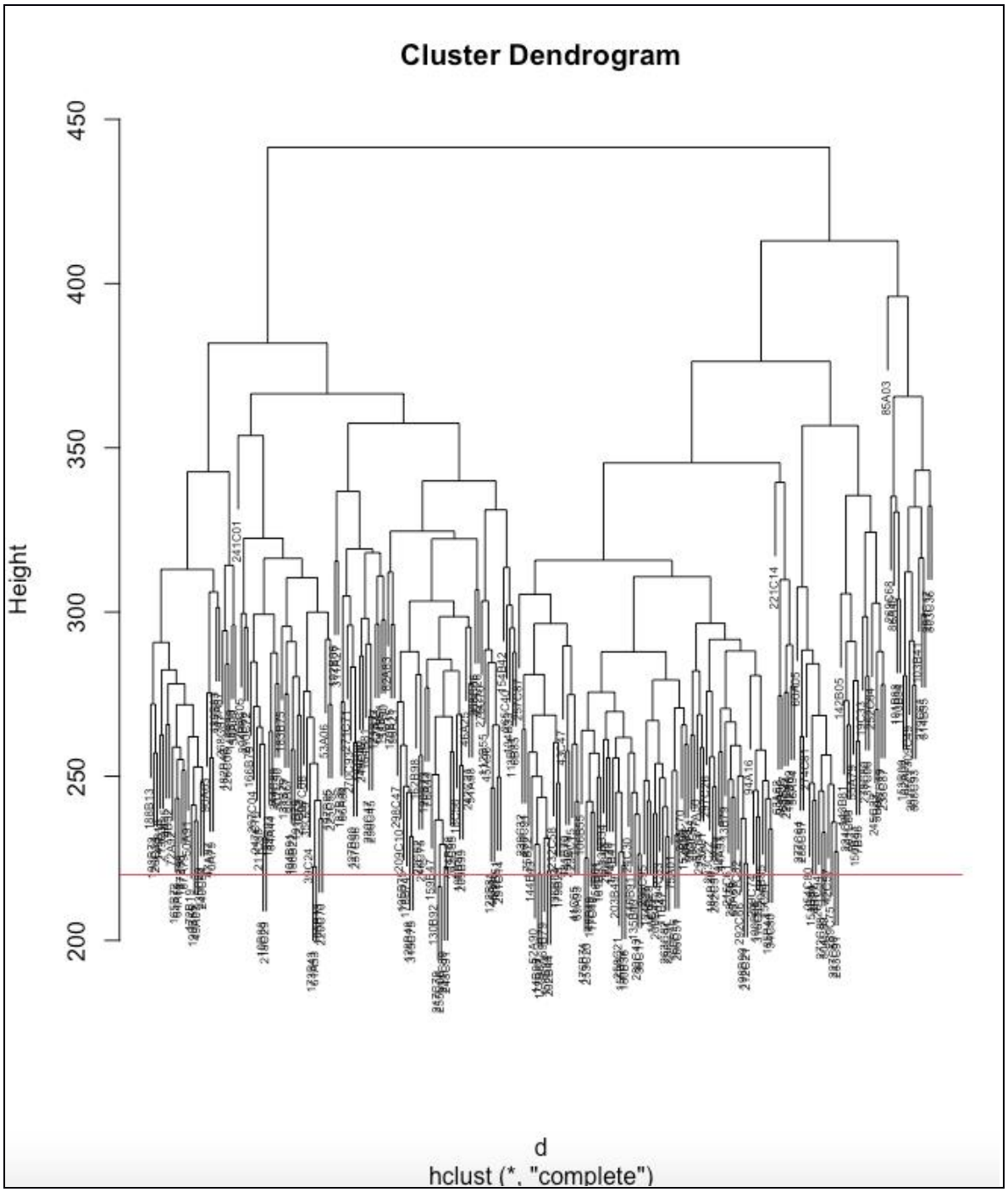
RESULTS:

After the data is loaded and inspected, data is scaled to avoid any biases and also so that the gene data roughly has the same standard deviation, also avoid the gene with large standard deviation which are likely to dominate the clustering process, while performing hierarchial clustering with complete linkage as follows:

```
table(BC_clinical$LNstatus)
head(BC_data)

#hierarchial clustering
scaled.BC_data <- t(scale(t(BC_data), center = TRUE))
head(scaled.BC_data)
d <- dist(t(scaled.BC_data))
hc <- hclust(d, method = "complete")

#plot
plot(hc, cex=0.5)
abline(h=220, col = 2)

K <- 2:5
sh <- NULL
for (i in K){
```

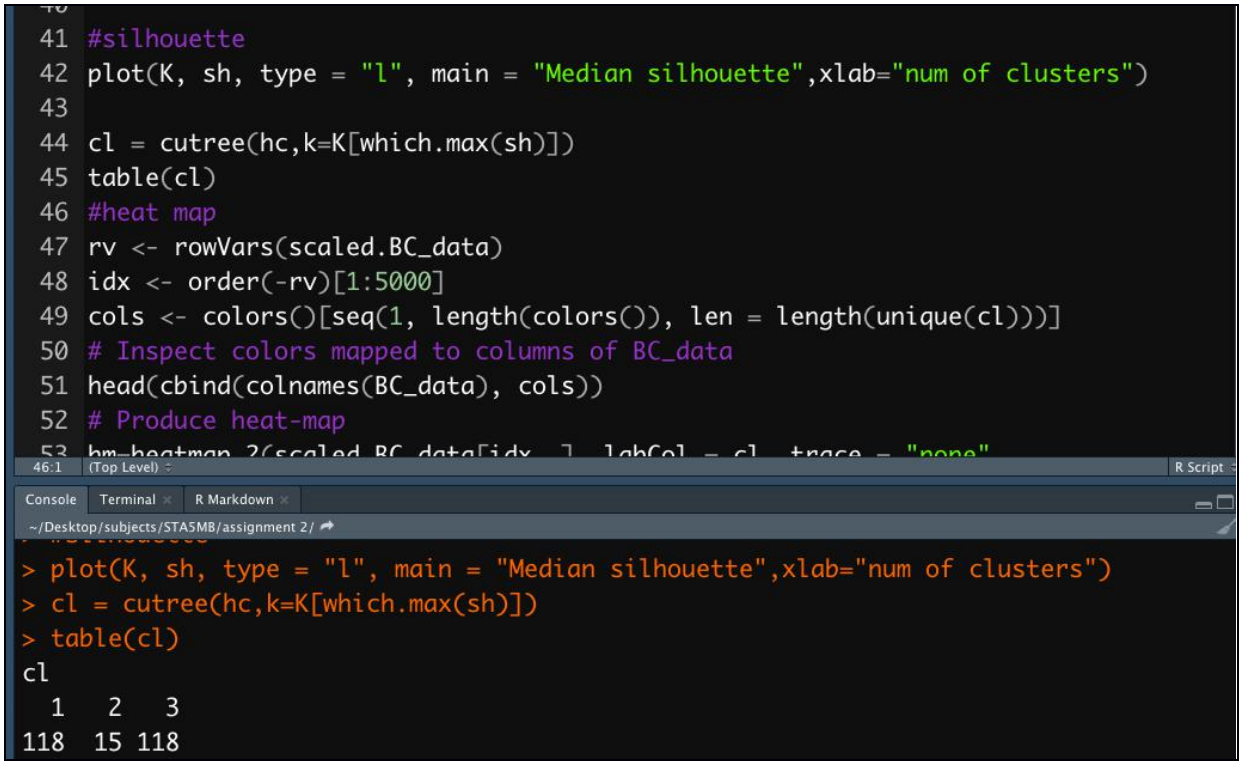
The clustering dendogram is as follows:

The horizontal axis represents the clusters where as the vertical axis represent the distance or dissimilarity. The joining of two clusters is shown in the dendogram as bifurcation of a vertical line into two vertical lines. The vertical position of the bifurcation shown by the short bar gives the dissimilarity between the two clusters.

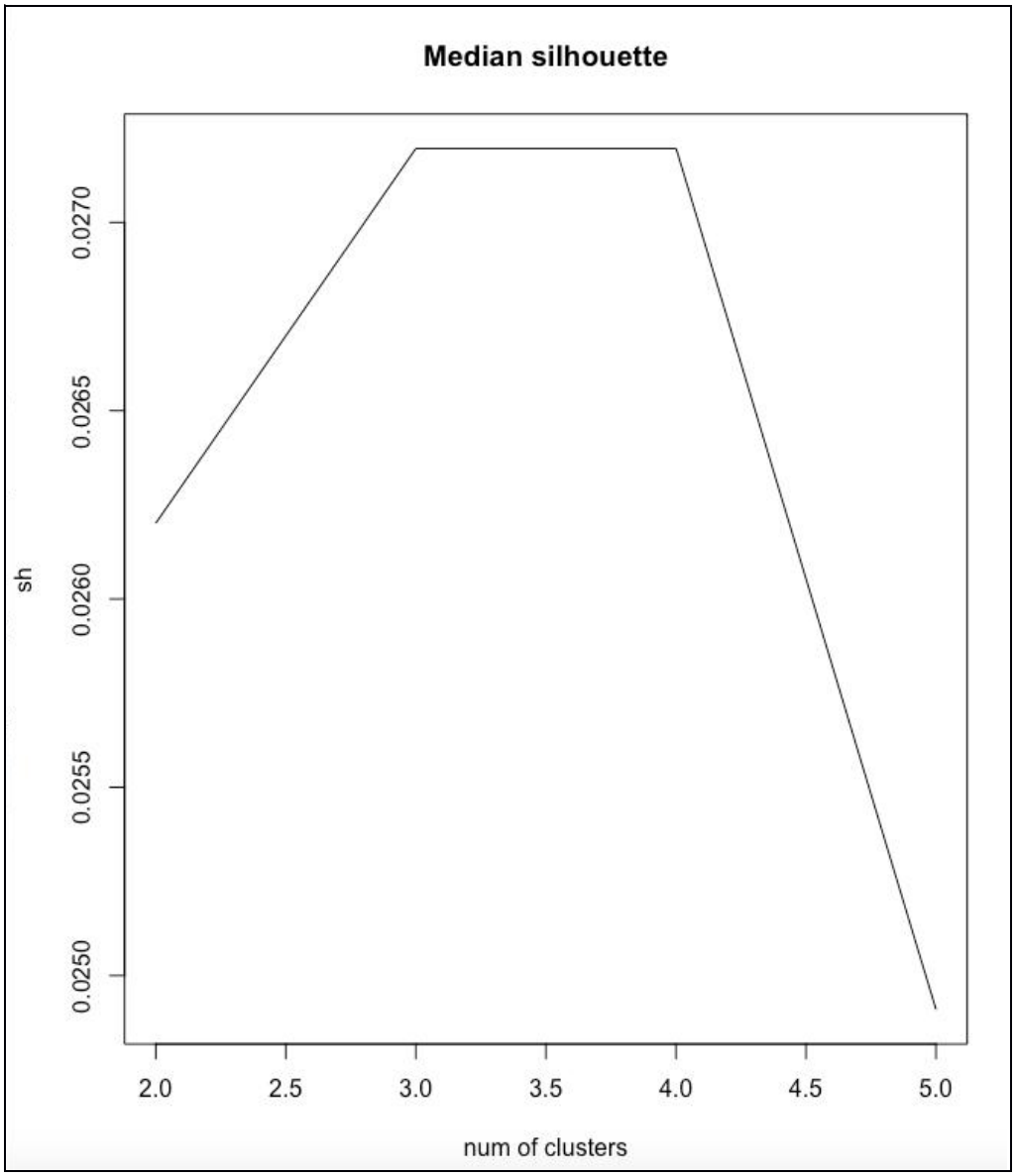


Finding the optimal number of clusters:

For finding the optimal number of clusters silhouette method is used and we can see that there are three clusters with clusters 1 and 3 having the same count and with very little data in cluster 2.



From the median Silhouette it can be clearly understood that optimal number of clusters is 3 since this the number of clusters at which the silhouette value is greatest, out of all cluster numbers considered.



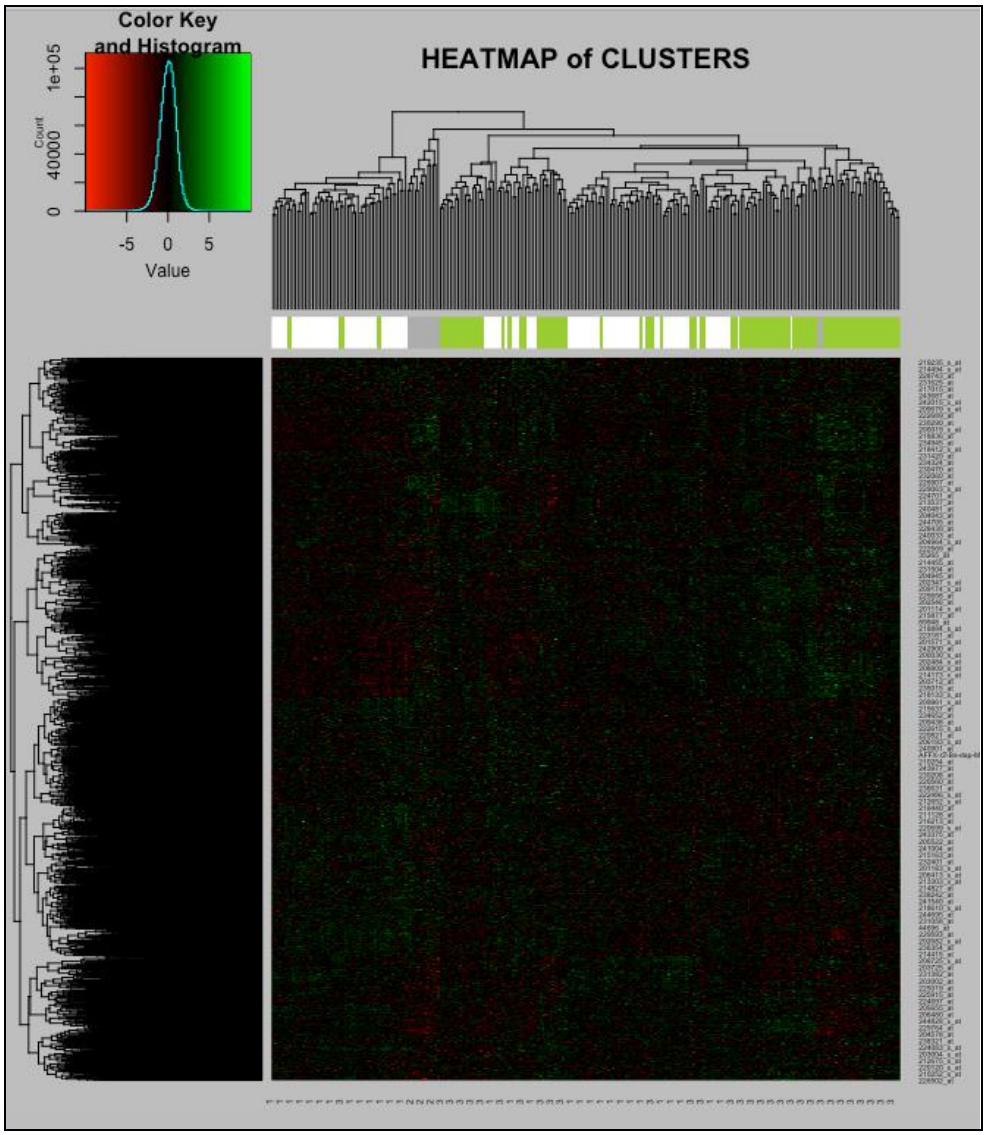
A heat map is a way to visualize hierarchial clustering. It is also sometimes called as false colored image as the data values are transformed to color scale. The heat map for the scaled expression data is found as follows:

The different colors for the column bars denote the three different clusters. The clusters colors are white, grey68 and yellow-green.

```
#heat map
rv <- rowVars(scaled.BC_data)
idx <- order(-rv)[1:5000]
cols <- colors()[seq(1, length(colors()), len = length(unique(cl)))]
# Inspect colors mapped to columns of BC_data
head(cbind(colnames(BC_data), cols))
# Produce heat-map
hm=heatmap.2(scaled.BC_data[idx, ], labCol = cl, trace = "none",
          ColSideColors = cols[cl],col = redgreen(100))
```

```
> suppressWarnings(head(cbind(colnames(BC_data),cols)))
                cols
[1,] "100B08" "white"
[2,] "101B88" "grey68"
[3,] "102B06" "yellowgreen"
[4,] "103B41" "white"
[5,] "104B91" "grey68"
[6,] "105B13" "yellowgreen"
```

In the first one third from the heat map below, in the white cluster seems to be dominating with little of green cluster having more of up-regulated(green) signals with very little of down-regulated(red) signals. Even in the next parts of heat map also more of the up-regulated signals(green) are seen than the red signals.

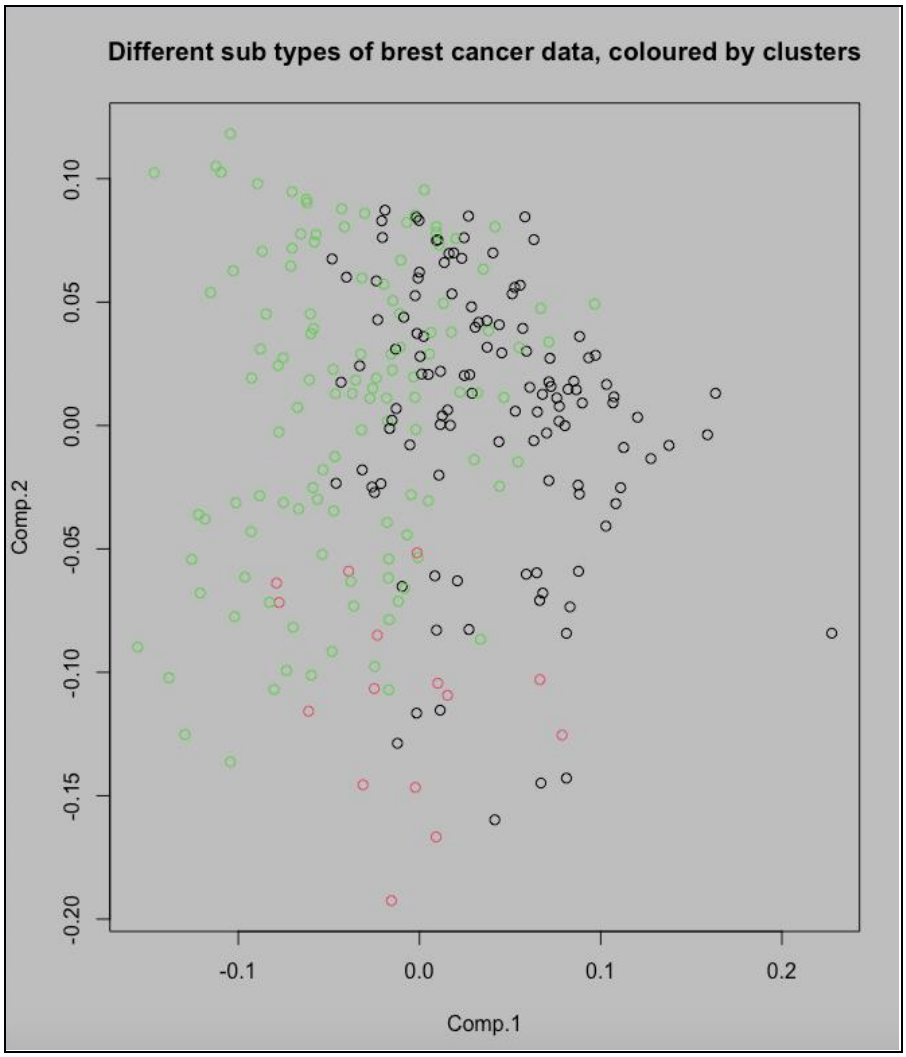


PCA analysis:

PCA analysis is performed to get the clusters plotted as follows:

```
#PCA analysis
par(bg = "grey")
pc <- princomp(scaled.BC_data[idx, ])
plot(pc$load[, 1:2], col = cl)
title("Different sub types of brest cancer data, coloured by clusters")
```

There are three clusters represented in three different colors green, black and red, splitting the data into three clear groups which highlights the efficacy of PCA method being able to visually distinguish the data into three clear groups.



DGE analysis:

For this a design matrix of factors or covariates, which we want to adjust is defined.

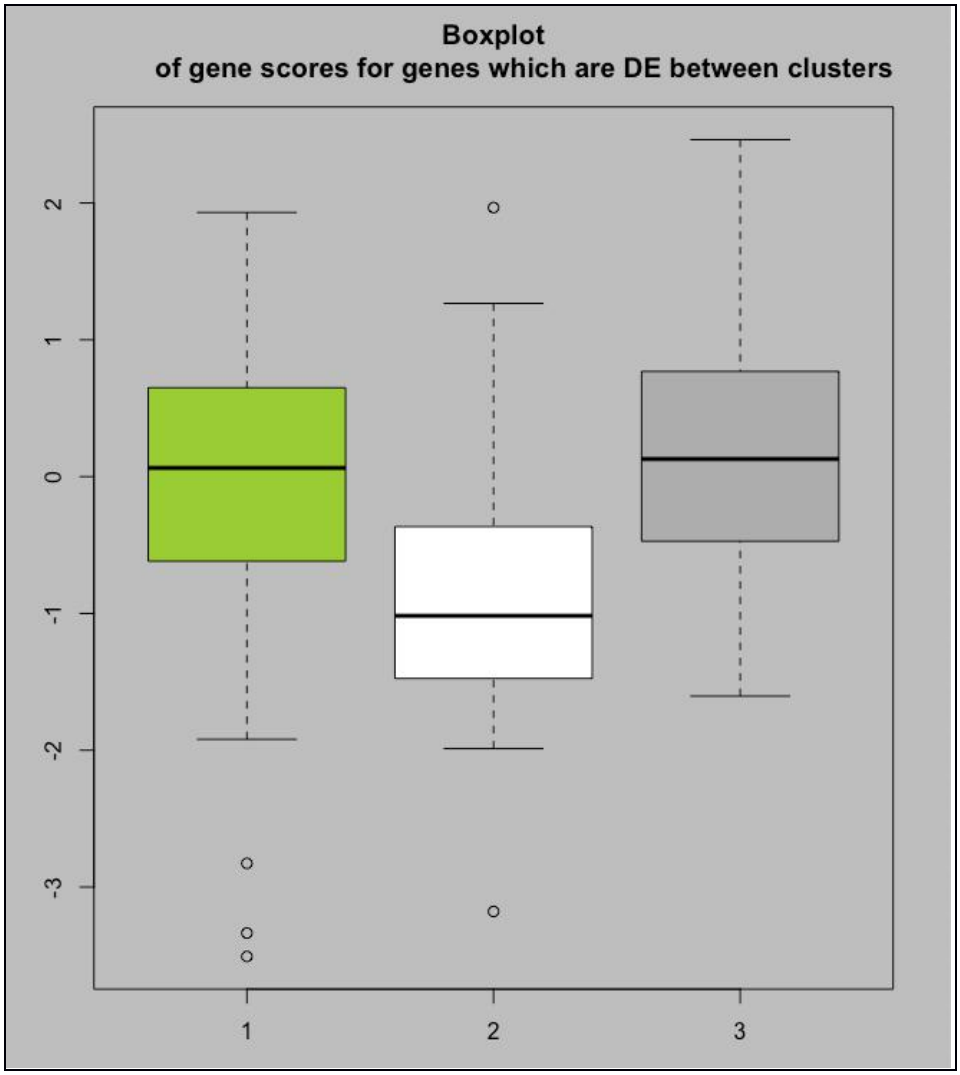
Limma package is used for this analysis.

```
#Performing the Differential gene expression analysis by comparing the clusters

design <- model.matrix(~as.factor(cl))
DE.object<- lmFit(BC_data, design)
DE.object <- eBayes(DE.object)

qval<- qvalue(DE.object$p.value[,2], fdr.level = 0.05)
```

SURVIVAL ANALYSIS:

Gene expression is formed based on the expression of significant DE genes. Then gene score is standardized for having mean=0 and standard deviation and cox regression is performed to estimate HR of gene score. The gene score for each sample is the sum of DE genes expression values for that particular sample. If we split the gene.scores values by cluster membership, we can visualize the gene scores for those genes which are differentially expressed between the three clusters using a boxplot as follows:

```
#survival
gene.score <- colSums(BC_data[qval$sig, ])
# standardize gene score (to have mean=0, SD=1)
gene.score <- scale(gene.score)

boxplot(split(gene.score,cl),col=c("yellowgreen","white","grey68"),main="Boxplot
        of gene scores for genes which are DE between clusters")
# Perform Cox regression to estimate HR of gene.score
cox.model <- coxph(Surv(Surv_time, event) ~ gene.score + histgrade +
                     ERstatus + PRstatus + age + tumor_size_mm +
                     LNstatus , data = BC_clinical)
summary(cox.model)
```



The summary command is used for inspecting the cox regression model.

```
Call:
coxph(formula = Surv(Surv_time, event) ~ gene.score + histgrade +
    ERstatus + PRstatus + age + tumor_size_mm + LNstatus, data = BC_clinical)

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

                   coef  exp(coef)   se(coef)      z Pr(>|z|)
gene.score     1.910e-01  1.210e+00  1.364e-01  1.400  0.16147
histgradeG1   -1.772e+00  1.699e-01  1.086e+00 -1.632  0.10257
histgradeG2   -1.236e+00  2.905e-01  1.033e+00 -1.197  0.23133
histgradeG3   -1.042e+00  3.527e-01  1.085e+00 -0.961  0.33680
ERstatusER?   -1.548e+01  1.891e-07  3.799e+03 -0.004  0.99675
ERstatusER+    6.438e-01  1.904e+00  5.510e-01  1.168  0.24270
PRstatusPgR+  -4.650e-01  6.282e-01  4.485e-01 -1.037  0.29991
age            3.780e-03  1.004e+00  1.041e-02  0.363  0.71667
tumor_size_mm  3.217e-02  1.033e+00  1.226e-02  2.624  0.00870 **
LNstatusLN?   -1.617e+01  9.495e-08  3.720e+03 -0.004  0.99653
LNstatusLN+    9.853e-01  2.679e+00  3.070e-01  3.209  0.00133 **
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

              exp(coef) exp(-coef) lower .95 upper .95
gene.score    1.210e+00  8.261e-01   0.92647     1.582
histgradeG1   1.699e-01  5.884e+00   0.02024     1.427
histgradeG2   2.905e-01  3.442e+00   0.03839     2.199
histgradeG3   3.527e-01  2.836e+00   0.04205     2.958
ERstatusER?   1.891e-07  5.288e+06   0.00000       Inf
ERstatusER+   1.904e+00  5.253e-01   0.64645     5.606
PRstatusPgR+  6.282e-01  1.592e+00   0.26079     1.513
age           1.004e+00  9.962e-01   0.98350     1.024
tumor_size_mm 1.033e+00  9.683e-01   1.00817     1.058
LNstatusLN?   9.495e-08  1.053e+07   0.00000       Inf
LNstatusLN+   2.679e+00  3.733e-01   1.46755     4.889

Concordance= 0.78  (se = 0.029 )
Likelihood ratio test= 46.22  on 11 df,   p=3e-06
Wald test            = 44.18  on 11 df,   p=7e-06
Score (logrank) test = 54.46  on 11 df,   p=1e-07
```

CONCLUSION:

Tumor size is an important factor in the staging of breast cancer which can affect a treatment of person. The tumors when detected early makes then easier to treat else would deprive the situation. The tumors of size 2.1 cm has a volume of 4500 mm3, tumor of size 4.9 cm has a volume of 60.000 mm3. The tumor sizes can be classified based on different sizes like TX means where the tumor size cannot be assessed, T0 mean no tumor is found, Tis is carcinoma in situ, T1 means tumor smaller or equal to 2cm and so on .Tumor size is modeled as a continuous variable by proportional hazards using a generalized addictive models procedure. It is also concluded in a study that tumor size between 3 and 50 mm, there is increase in death rates. The tumor size also seems to be a strong predictor in node-positive and node-negative sub groups. Any decline in survival is rendered to positive nodal status that increased with tumor size. The reduction in tumor size on diagnosis is important for women with node positive breast cancer and this screening is likely to benefit women even after regional spread .In the gist it can be concluded that tumor size should be given prime importance i.e., should be measured and taken into consideration when planning the treatment but eventual decisions should be made by including the inferences from the

biological and molecular characteristics of the tumor in order to achieve a personalized approach to each individual for offering the best possible treatment.

LNstatusLN+ is the lymph node status which is the most prognostic variable that guides the ER positive breast cancer treatment. Generally a positive node status is associated with the poor prognosis, subset of the patients with the breast cancer who does respond well to treatment and achieve long term survival. Lymph node status confirms whether or not the lymph nodes contain cancer. Lymph node involvement is a characteristic for molecular sub type of breast cancer . The importance of lymph node status can be observed during the growth of prognostic gene signatures as the size of tumor is positively correlated with the LN positivity.

REFERENCES:

Verschraegen, C., Vinh-Hung, V., Cserni, G., Gordon, R., Royce, M., Vlastos, G., Tai, P. and Storme, G., 2005. Modeling the Effect of Tumor Size in Early Breast Cancer. Annals of Surgery, 241(2), pp.309-318.

Kasangian, A., Gherardi, G., Biagioli, E., Torri, V., Moretti, A., Bernardin, E., Cordovana, A., Farina, G., Bramati, A., Piva, S., Dazzani, M., Paternò, E. and La Verde, N., 2017. The prognostic role of tumor size in early breast cancer in the era of molecular biology. PLOS ONE, 12(12), p.e0189127.

Provencher, L., Diorio, C., Hogue, J., Doyle, C. and Jacob, S., 2012. Does breast cancer tumor size really matter that much?. The Breast, 21(5), pp.682-685.

Cockburn, J., Hallett, R., Gillgrass, A., Dias, K., Whelan, T., Levine, M., Hassell, J. and Bane, A., 2016. The effects of lymph node status on predicting outcome in ER+ /HER2- tamoxifen treated breast cancer patients using gene signatures. BMC Cancer, 16(1).