123456

Report on Breast Cancer and its sub-type

**INRODUCTION:**

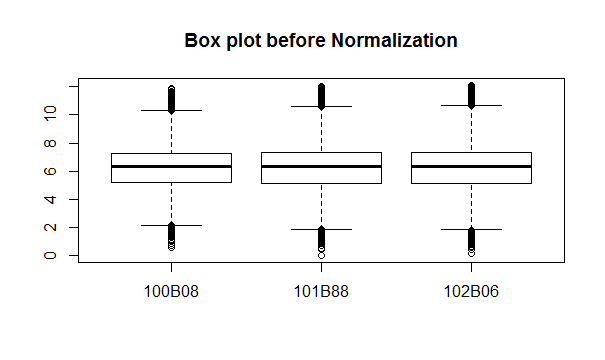
The dataset that is being used contains an array gene expression with a 251 set of data consisting of detailed information about the breast-cancer patients like survival time, LNstatus, tumor size in mm, age ,PRstatus, histgrade etc.

The given dataset is analyzed by performing the Differential Gene Expression (DGE) analysis to estimate the survival rate and breast- cancer.

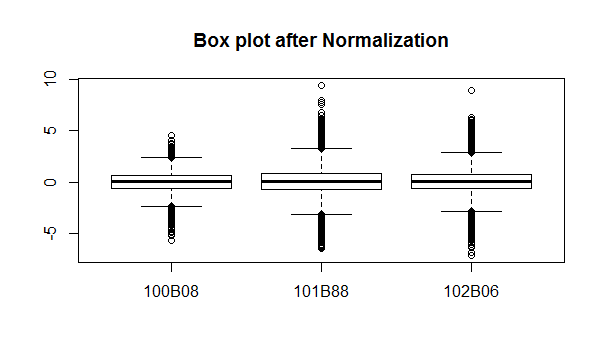
**METHODS:**

**NORMALIZATION OF DATA:**

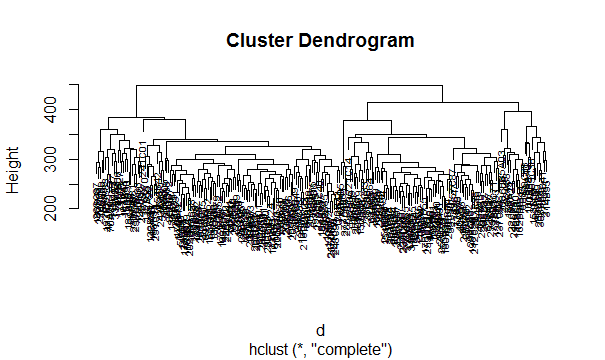
One of the standard ways to demonstrate the quartiles of the distribution of data, this has been done to evaluate the normalization that is required for such a huge data set. Moreover, RStudio tools have been used for the performance of this analysis. This can be achieved by loading each data from the directory provided.



[1]The above boxplot depicts that the data is not normalized. Hence the data can be normalized with the help of scale function then using scaling expressions for the same. Therefore, further normalization has been done with the help of normalizeBetweenArrays method.



The figure below shows the cluster dendrogram of the data. [2]The relationship is shown in the form of hierarchy between objects, which is graphically represented to illustrate the correspondence of data that has been arranged in the form of clusters.

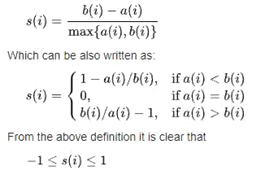


In the figure, the Y-axis is denoted by the clusters that are joint with each other. The two clusters result in the form of drawing a horizontal line around h = 251.

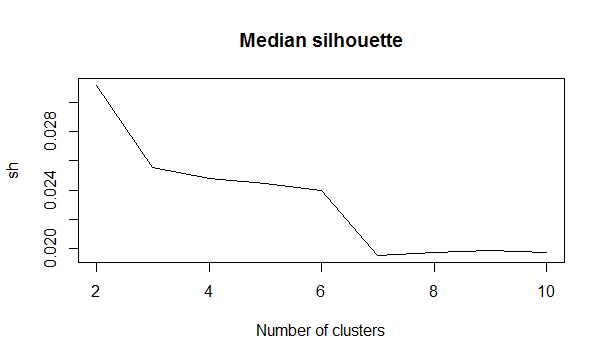
**SILHOUETTE METHOD**:

For determining the optimal clusters, *silhouette* is been used. [3]This will help in the calculation of each observation to the similarity with the clusters that need to be assigned to the other similar clusters respectively. Hence, the rate of clusters is always better with a higher *silhouette.*

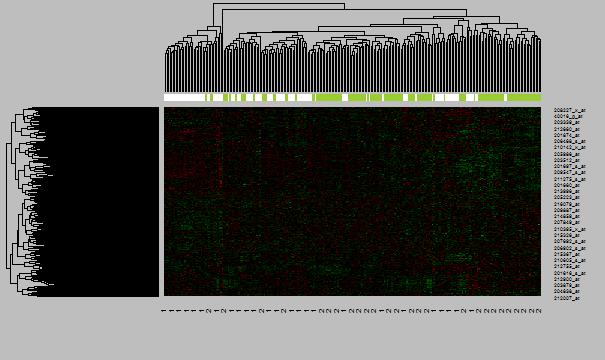
The formula used to determine it is as follows:



* a(i) determines the average distance between i and all other data within similar clusters.
* b (i) determines the average distance of i to other clusters within all points, in which i is not a member.



To consider the clusters out of all the other greater values of sh, the number of clusters has been determined. In our case, the median silhouette is plotted with the help of the evidence of the optimal number of clusters, in this case it has been observed as 2.

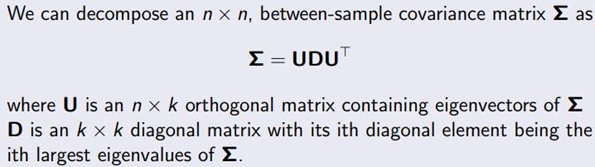


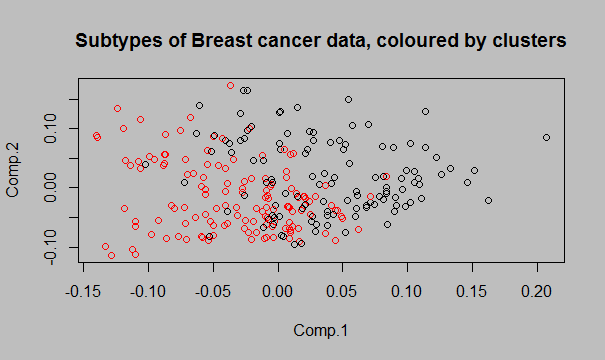
To add more information to the bi-clustering that is performed we use the Heatmap, which in other words is nothing but samples of clusters of genes. Hence, the Heatmap provided a better idea of the same.

In the above Heatmap it is observed that, the contribution of the gene can be seen in the form of green clusters (green in color) and white clusters. These are the two main clusters that are featured respectively.

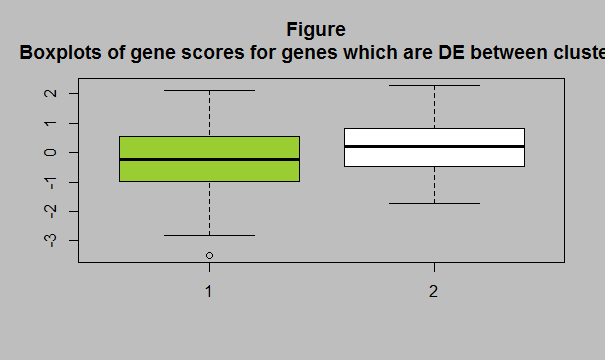
**PRINCIPAL COMPONENT ANALYSIS:**

[4] PCA is executed to avoid surpassing the computational time cased during the analysis of our data as it consists of many features. These features are not in 2-D but are of higher dimension. Executing PCA helps in preserving the distance between samples as much as possible by performing the dimensional reduction.





First, we're going to outline differentially expressed genes using genescore by performing DE analysis with cluster membership vectors. we must get a collection of differentially expressed (DE) genes with q-value= 0.05. Then Evaluate the genescore of each DE gene sample. Now we get a box plot, a green and a white box that demonstrates two clusters where we can see the genes of those genes.

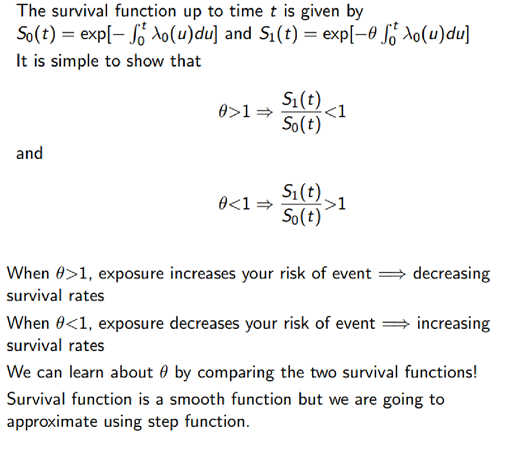


**COX REGRESSION:**

Cox Regression or (Proportional hazards regression) is one of the most popular regression techniques. It is used to link multiple risk factors or conditions, viewed at the same time, to survival time. [5]The hazard rate is the measure of effect, in other words, it’s the risk or likelihood of damage from an incident of interest, provided that the subject has existed for a specified period of time.

Hazard ratio (i.e;t) is the main factor in survival analysis, stated as the ratio of(risk of the outcome in one group)/(risk of outcome in another group), occurring at a given interval of time. It has two coefficients positive which represent the worst prognosis and a negative coefficient represents a better prognosis.

The Hazard ratio and survival function are given below:



**SUMMARY:**

> summary(cox.model)

Call:

coxph(formula = Surv(Surv\_time, event) ~ gene.score + 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.score1.809e-01 1.198e+00 1.489e-01 1.215 0.224429

ERstatusER? -1.573e+01 1.469e-07 3.842e+03 -0.004 0.996733

ERstatusER +7.672e-01 2.154e+00 5.372e-01 1.428 0.153258

PrstatusPgR+ -7.029e-01 4.952e-01 3.858e-01 -1.822 0.068472 .

age 2.972e-03 1.003e+00 1.035e-02 0.287 0.773972

tumor\_size\_mm3.564e-02 1.036e+00 1.127e-02 3.164 0.001557 \*\*

LNstatusLN? -1.613e+01 9.895e-08 3.757e+03 -0.004 0.996574

LNstatusLN+ 1.051e+00 2.860e+00 2.998e-01 3.506 0.000456 \*\*\*

---

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

exp(coef) exp(-coef) lower .95 upper .95

gene.score 1.198e+00 8.345e-01 0.8950 1.604

ERstatusER? 1.469e-07 6.808e+06 0.0000 Inf

ERstatusER+ 2.154e+00 4.643e-01 0.7515 6.172

PRstatusPgR+ 4.952e-01 2.020e+00 0.2325 1.055

age 1.003e+00 9.970e-01 0.9828 1.024

tumor\_size\_mm 1.036e+00 9.650e-01 1.0137 1.059

LNstatusLN? 9.895e-08 1.011e+07 0.0000 Inf

LNstatusLN+ 2.860e+00 3.496e-01 1.5894 5.147

Concordance= 0.772 (se = 0.028 )

Likelihood ratio test= 42.52 on 8 df, p=1e-06

Wald test = 42.1 on 8 df, p=1e-06

Score (logrank) test = 51.12 on 8 df, p=2e-08

The form of calculating the gene score which in this case is found to be 1.08. Based on the De gene we have calculated the hazard ratio of gene expression in the above figure.

It is observed that there is a small p-value in terms of the Cox Regression in the above summary, thus making it statistically significant. Moreover, the conclusion also is aligned with the support of about 95% CI in case of hazard ratio with value 1.

>coxph(formula = Surv(BC\_clinical$Surv\_time, BC\_clinical$event) ~ gene.score + as.factor(scatter),data = BC\_clinical)

Call:

coxph(formula = Surv(BC\_clinical$Surv\_time, BC\_clinical$event) ~

gene.score + as.factor(scatter), data = BC\_clinical)

coef exp(coef) se(coef) z p

gene.score 0.17364 1.18963 0.22732 0.764 0.445

as.factor(scatter) 1 -0.08883 0.91500 0.45286 -0.196 0.844

Likelihood ratio test=1.08 on 2 df, p=0.583

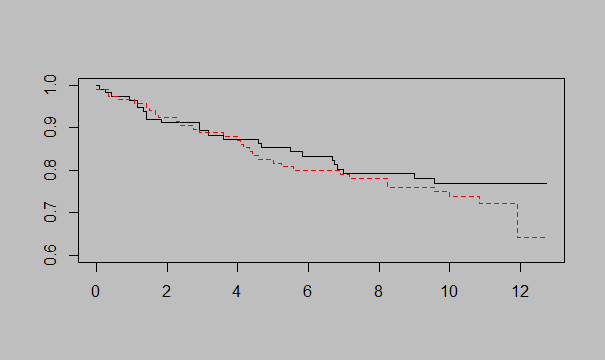
n= 236, number of events= 55

(15 observations deleted due to missingness)

**SURVIVAL RATES OF CLUSTERS:**

[6]The percentage of people in a study or treatment group who are still alive for a period of time after they have been diagnosed with or began treatment for a condition, such as cancer. It is also known as the overall survival rate.

The below graph represents the survival rate between two clusters.



**CONCLUSION:**

First normalization is implemented to the given dataset in consideration to identify any important differences and a denogram is generated. Unfortunately, the generated denogram fails to provide complete information so as to which gene adds value to the clusters. To solve this we generate heatmap and observe the up and down-regulated genes. The data then generated is insufficient for our analysis and hence we opt for PCA - principal component analysis and gain dimensionally reduced data. The DE analysis is carried out after PCA plot. The DE analysis happens with a precondition of having 0.05 as q value.DE analysis helps us find the gene score. Indicators for breast cancer metastasis are cox regression, tumor size, and lymph node metastasis. The final output of our analysis showcased that between the tumor size and lymph node metastasis there was important association statistically also these are played an important role in the survival and have got various survival prognosis.

**REFERENCE LIST:**

[1] Quackenbush, J, ‘Microarray data normalization and transformation’, “Natural Publishing Group”, vol.32, pp. 496-501, 2002

[2] J. C. Gower & G. J. S. Ross, ‘Minimum Spanning Trees and Single Linkage Cluster Analysis’, “Rothamsted Experimental Station”, 2019

[3] Raskar, C & Cohen, M, ‘Image Precision Silhouette Edges’, 2000

[4] Zou, H, Hastle, T, &Tibshirani, R, ‘Sparse Principal Component Analysis’, “American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America*”,* vol.15, pp. 265-286, 2005

[5] D. Y. Lin & Z. Ying, ‘Cox Regression WithIncompleteCovariate Measurements’, “Journal of the American Statistical Association”, vol. 88, pp. 1341-1349, 2019

[6] Grijs, R &Parmentier, G, ‘The long-term survival chances of young massive star clusters’, “Chinese Journal of Astronomy and Astrophysics”, vol.0, 2007