STA5077Z: UNSUPERVISED LEARNING - PROJECT

Corne Oosthuizen - OSTAND005

Due: 14 August 2017

# 1 - Leukemia Dataset

## Warning: package 'kohonen' was built under R version 3.3.3

## Warning: package 'tidyverse' was built under R version 3.3.3

## Warning: package 'ggplot2' was built under R version 3.3.3

## Warning: package 'tibble' was built under R version 3.3.3

## Warning: package 'tidyr' was built under R version 3.3.3

## Warning: package 'readr' was built under R version 3.3.3

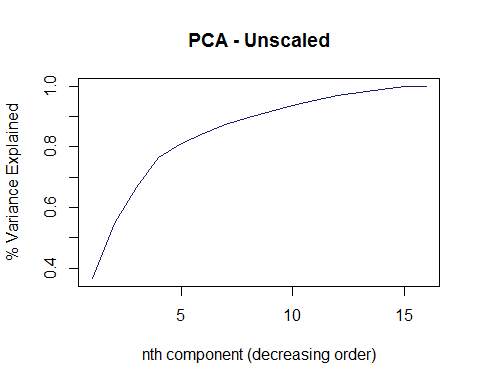
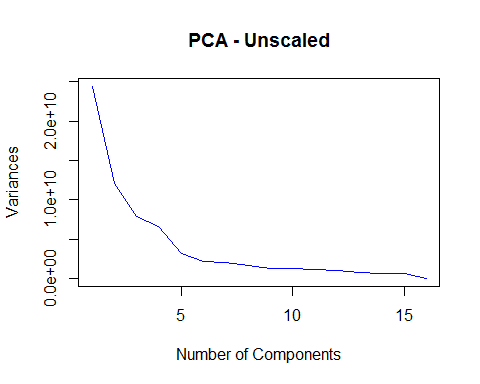
## Warning: package 'purrr' was built under R version 3.3.3

## Warning: package 'dplyr' was built under R version 3.3.3

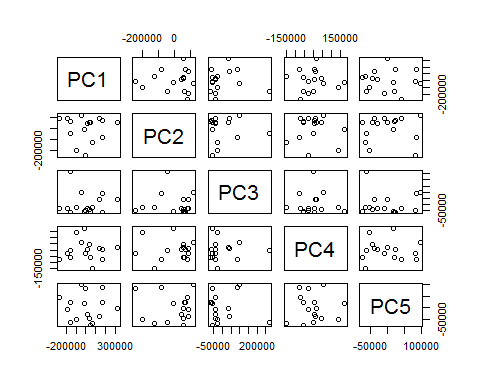
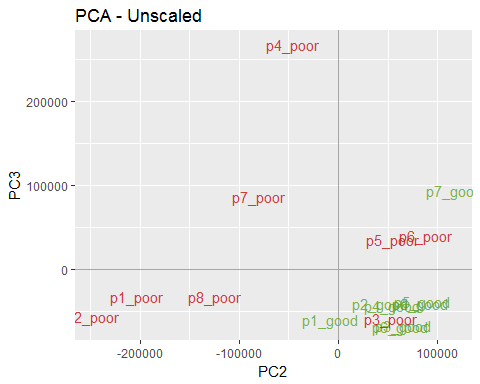
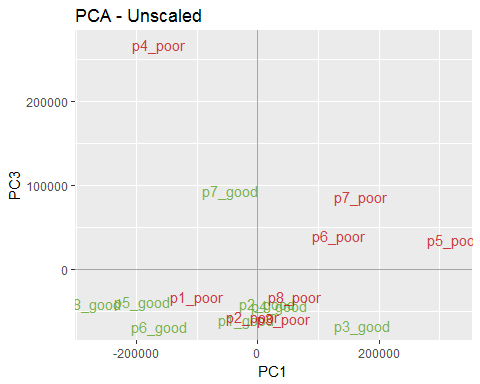
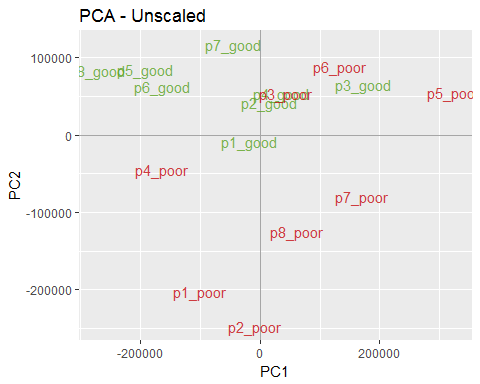
## Warning: package 'formattable' was built under R version 3.3.3

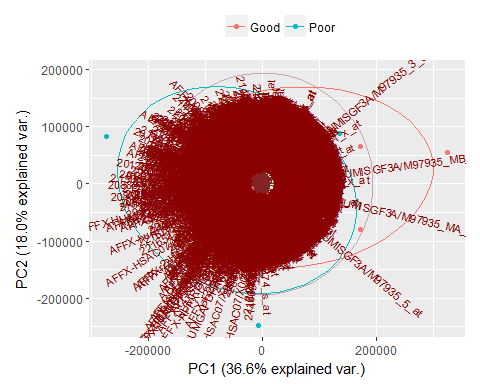
## Warning: package 'plyr' was built under R version 3.3.3

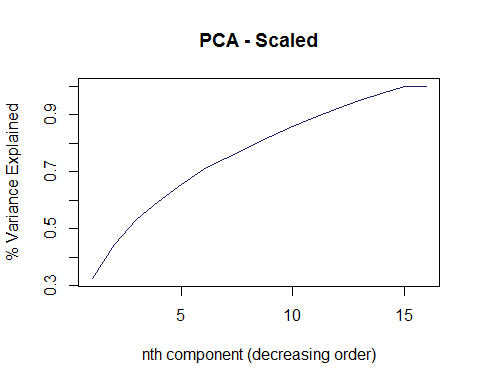
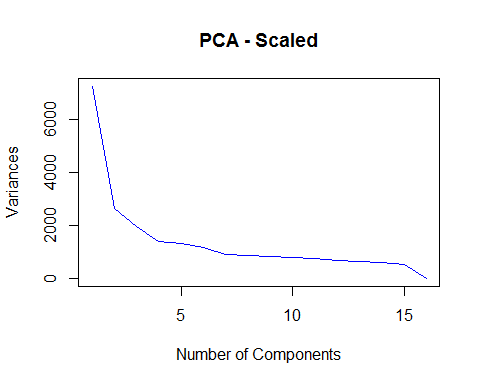
## Warning: package 'scales' was built under R version 3.3.3



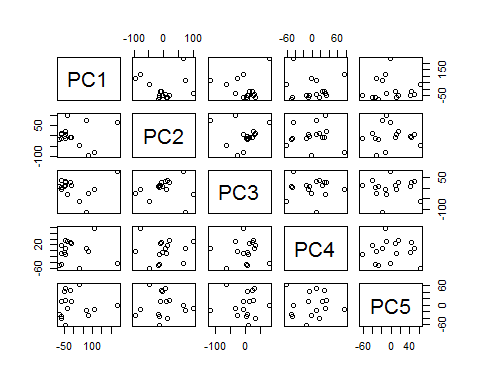
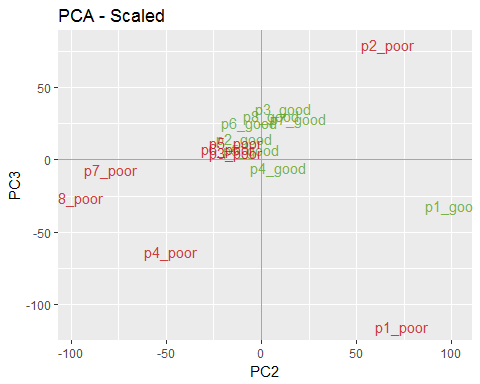
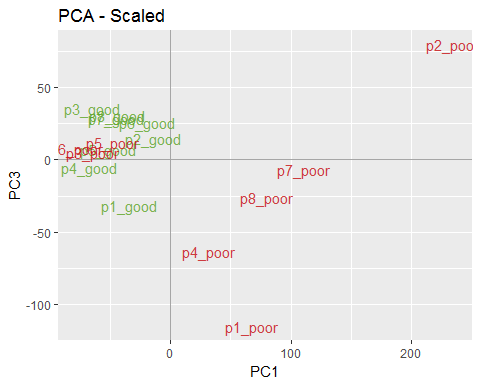
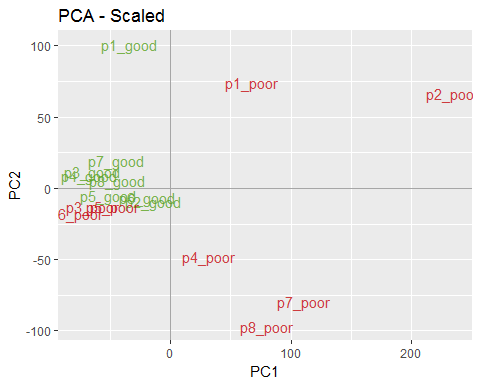
## Importance of components:  
## PC1 PC2 PC3 PC4  
## Standard deviation 156238.6335 109660.68045 89020.73562 80953.43395  
## Proportion of Variance 0.3663 0.18045 0.11892 0.09834  
## Cumulative Proportion 0.3663 0.54676 0.66568 0.76402  
## PC5 PC6 PC7 PC8  
## Standard deviation 56506.15293 46154.70483 44327.27105 40692.64712  
## Proportion of Variance 0.04791 0.03197 0.02949 0.02485  
## Cumulative Proportion 0.81193 0.84390 0.87338 0.89823  
## PC9 PC10 PC11 PC12  
## Standard deviation 36028.25444 35068.76793 33728.38573 31976.10588  
## Proportion of Variance 0.01948 0.01845 0.01707 0.01534  
## Cumulative Proportion 0.91771 0.93616 0.95323 0.96858  
## PC13 PC14 PC15 PC16  
## Standard deviation 27846.40656 26133.25873 25210.79476 3.745941e-10  
## Proportion of Variance 0.01164 0.01025 0.00954 0.000000e+00  
## Cumulative Proportion 0.98021 0.99046 1.00000 1.000000e+00

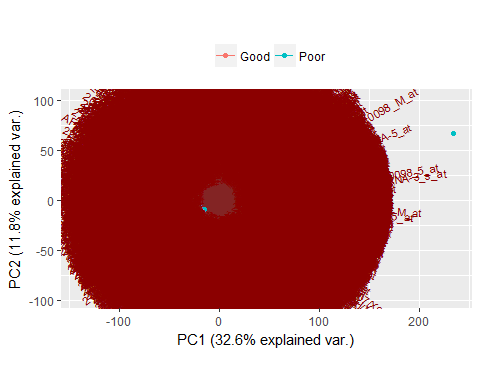


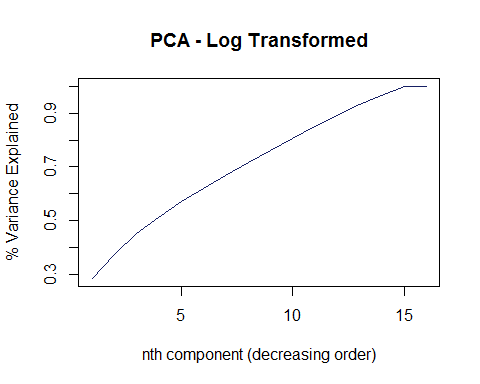
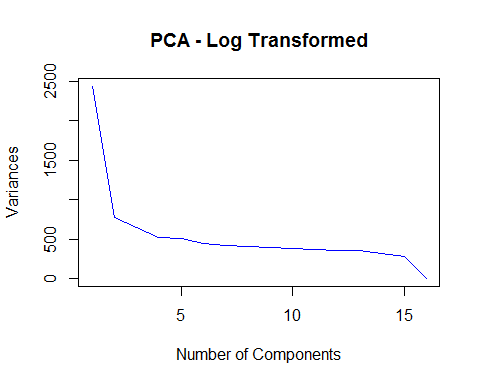




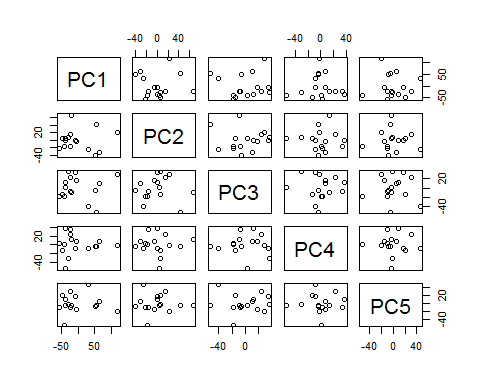
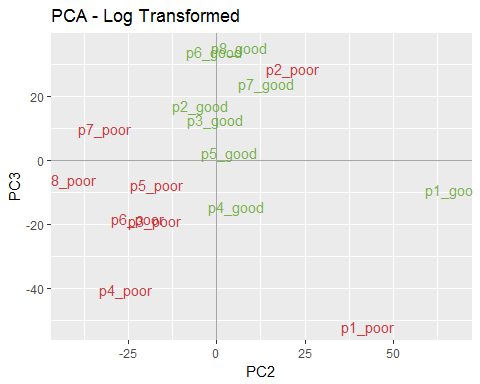
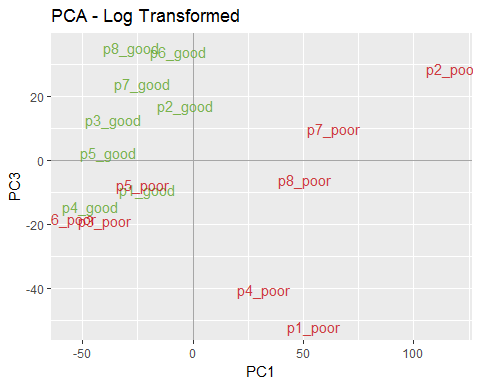
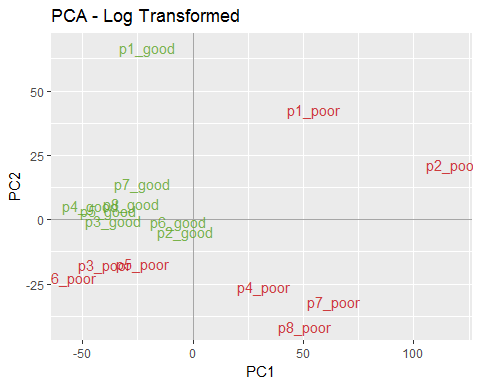
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 85.22634 51.34150 44.34176 37.21473 36.33934  
## Proportion of Variance 0.32597 0.11829 0.08824 0.06215 0.05926  
## Cumulative Proportion 0.32597 0.44426 0.53250 0.59465 0.65391  
## PC6 PC7 PC8 PC9 PC10  
## Standard deviation 34.33981 30.23176 29.39799 28.51449 27.80705  
## Proportion of Variance 0.05292 0.04102 0.03878 0.03649 0.03470  
## Cumulative Proportion 0.70683 0.74785 0.78663 0.82312 0.85782  
## PC11 PC12 PC13 PC14 PC15  
## Standard deviation 27.09373 26.27332 25.18051 24.22583 22.86498  
## Proportion of Variance 0.03294 0.03098 0.02845 0.02634 0.02346  
## Cumulative Proportion 0.89077 0.92175 0.95020 0.97654 1.00000  
## PC16  
## Standard deviation 1.691711e-13  
## Proportion of Variance 0.000000e+00  
## Cumulative Proportion 1.000000e+00

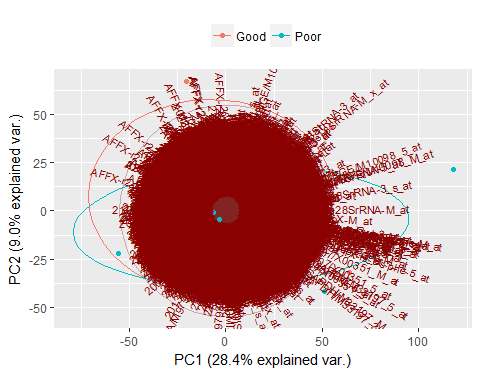
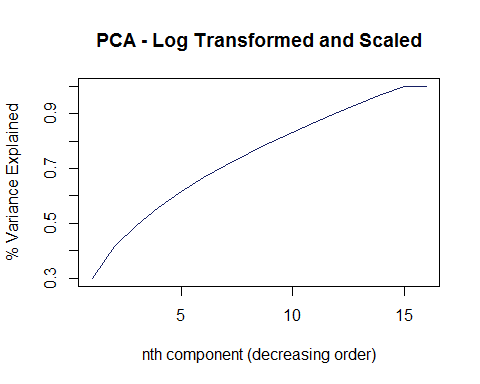
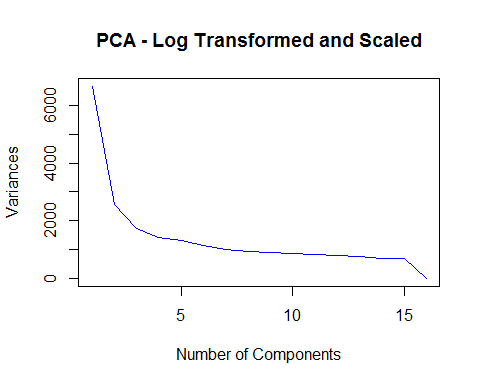




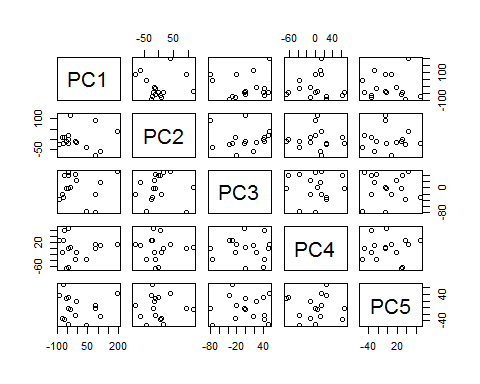
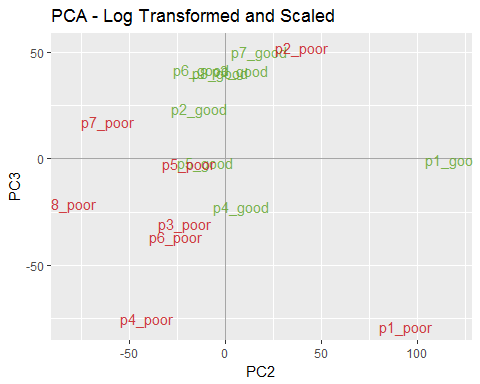
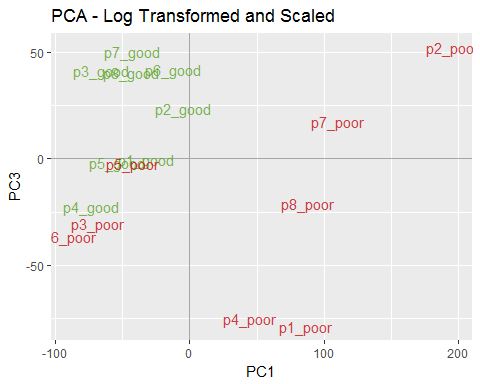
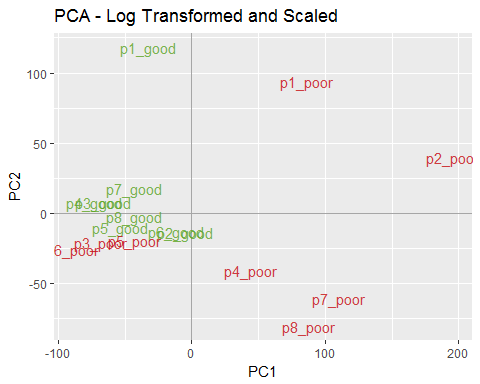


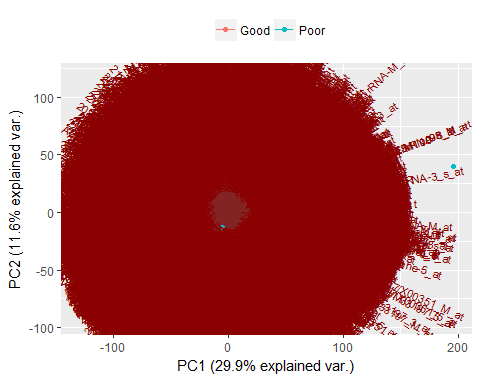
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 49.36871 27.82194 25.52599 22.75810 22.53319  
## Proportion of Variance 0.28385 0.09015 0.07588 0.06032 0.05913  
## Cumulative Proportion 0.28385 0.37399 0.44988 0.51019 0.56933  
## PC6 PC7 PC8 PC9 PC10  
## Standard deviation 20.90393 20.46430 20.17392 19.64844 19.49222  
## Proportion of Variance 0.05089 0.04877 0.04740 0.04496 0.04425  
## Cumulative Proportion 0.62022 0.66899 0.71639 0.76135 0.80560  
## PC11 PC12 PC13 PC14 PC15  
## Standard deviation 19.21435 18.93668 18.69074 17.77911 16.61446  
## Proportion of Variance 0.04300 0.04176 0.04068 0.03681 0.03215  
## Cumulative Proportion 0.84859 0.89035 0.93104 0.96785 1.00000  
## PC16  
## Standard deviation 9.516276e-14  
## Proportion of Variance 0.000000e+00  
## Cumulative Proportion 1.000000e+00



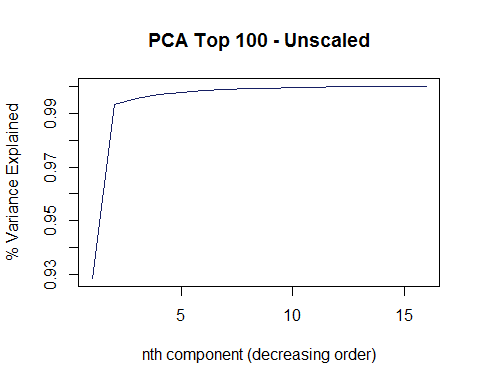
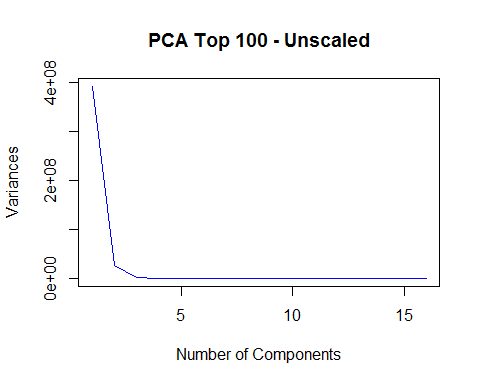
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 81.63214 50.79477 41.81428 37.86222 36.23344  
## Proportion of Variance 0.29905 0.11579 0.07846 0.06433 0.05892  
## Cumulative Proportion 0.29905 0.41484 0.49331 0.55764 0.61656  
## PC6 PC7 PC8 PC9 PC10  
## Standard deviation 33.74602 31.46372 30.65406 30.11230 29.10130  
## Proportion of Variance 0.05111 0.04443 0.04217 0.04069 0.03801  
## Cumulative Proportion 0.66766 0.71209 0.75426 0.79495 0.83296  
## PC11 PC12 PC13 PC14 PC15  
## Standard deviation 28.68284 27.93695 27.29255 26.39643 26.02576  
## Proportion of Variance 0.03692 0.03503 0.03343 0.03127 0.03040  
## Cumulative Proportion 0.86988 0.90491 0.93833 0.96960 1.00000  
## PC16  
## Standard deviation 1.920718e-13  
## Proportion of Variance 0.000000e+00  
## Cumulative Proportion 1.000000e+00



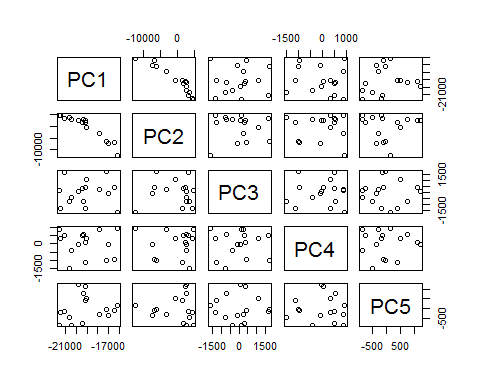
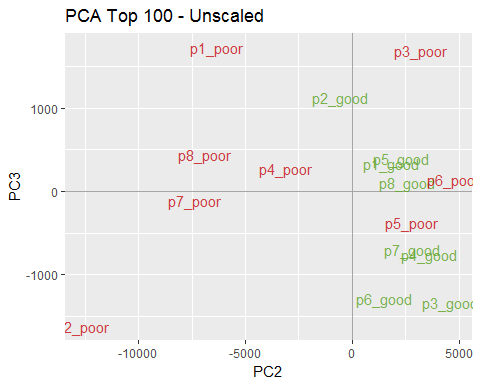
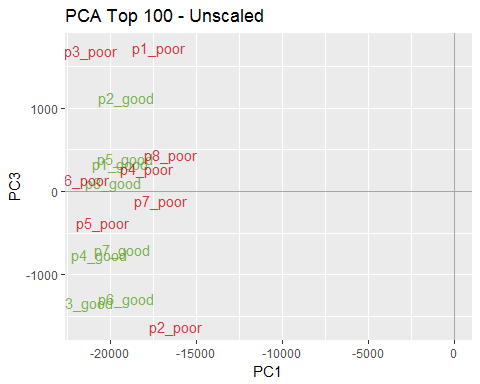
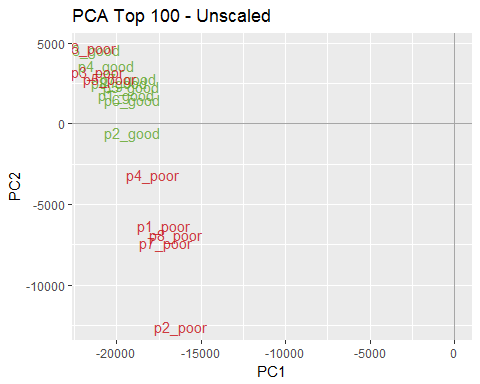


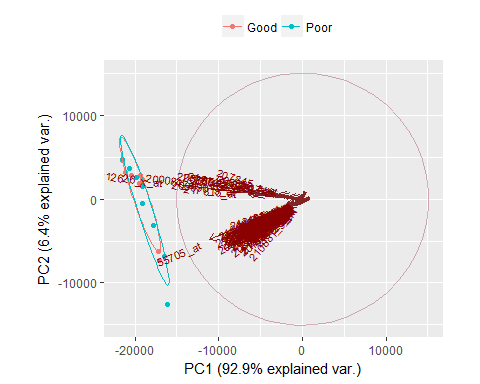
### Select the Top 100

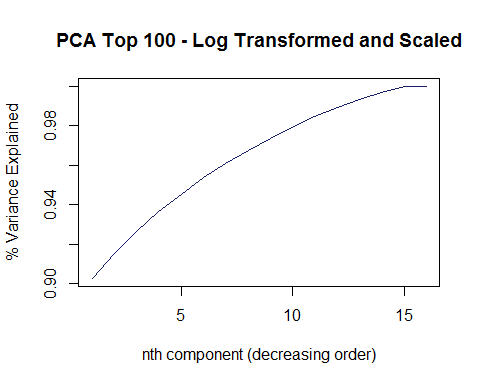
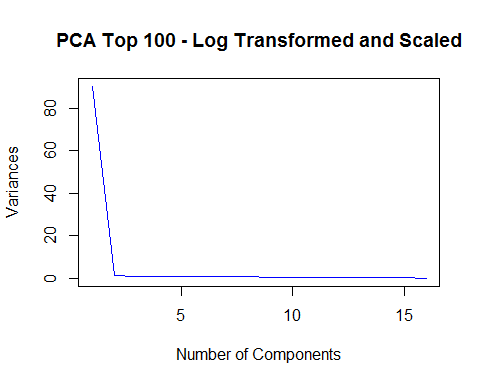
## p1\_good p2\_good p3\_good p4\_good p5\_good p6\_good p7\_good  
## 204365\_s\_at 381.6 404.7 322.8 326.9 317.1 384.7 330.2  
## 206766\_at 421.2 472.4 319.4 322.3 370.8 404.3 407.4  
## 221870\_at 21.1 22.9 22.7 18.1 22.8 32.4 24.6  
## 204542\_at 64.3 55.1 55.7 49.1 51.4 70.9 51.7  
## 215356\_at 19.7 23.6 20.2 15.9 19.7 19.1 16.2  
## 211843\_x\_at 27.3 28.6 25.5 23.0 23.4 23.0 24.7  
## p8\_good p1\_poor p2\_poor p3\_poor p4\_poor p5\_poor p6\_poor  
## 204365\_s\_at 339.5 794.6 1344.8 243.2 614.6 365.1 228.7  
## 206766\_at 397.8 673.9 1144.6 349.8 728.3 436.9 273.7  
## 221870\_at 19.8 46.2 106.5 21.5 46.5 25.1 17.8  
## 204542\_at 62.4 107.6 174.3 47.7 97.3 67.2 41.6  
## 215356\_at 18.7 40.3 102.7 16.6 49.4 18.1 14.7  
## 211843\_x\_at 27.5 49.6 137.0 24.5 41.6 27.6 19.5  
## p7\_poor p8\_poor  
## 204365\_s\_at 820.6 741.1  
## 206766\_at 863.7 714.2  
## 221870\_at 53.9 55.8  
## 204542\_at 115.2 110.0  
## 215356\_at 63.1 57.1  
## 211843\_x\_at 57.2 65.4



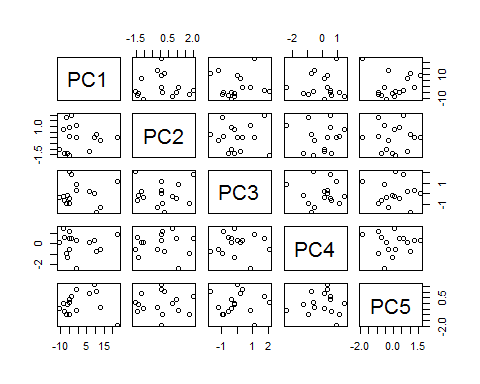
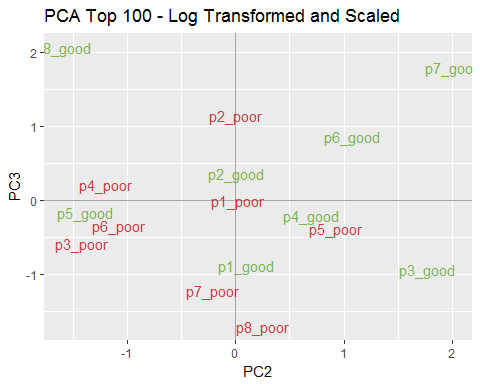
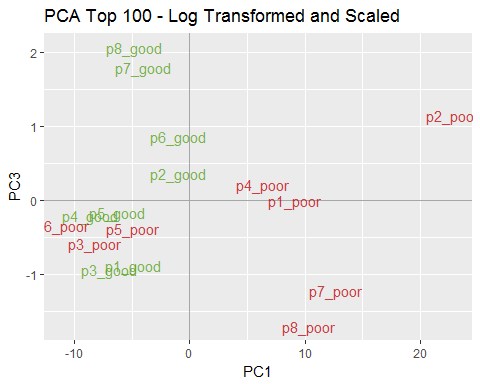
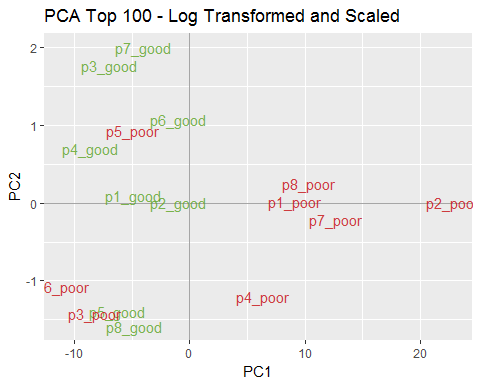
## Importance of components:  
## PC1 PC2 PC3 PC4  
## Standard deviation 19810.47483 5220.99979 998.60488 772.76111  
## Proportion of Variance 0.92859 0.06450 0.00236 0.00141  
## Cumulative Proportion 0.92859 0.99309 0.99545 0.99686  
## PC5 PC6 PC7 PC8 PC9  
## Standard deviation 646.24559 490.19201 409.49121 322.17051 312.13495  
## Proportion of Variance 0.00099 0.00057 0.00040 0.00025 0.00023  
## Cumulative Proportion 0.99785 0.99842 0.99881 0.99906 0.99929  
## PC10 PC11 PC12 PC13 PC14  
## Standard deviation 303.05889 272.19239 212.91015 180.07516 147.22423  
## Proportion of Variance 0.00022 0.00018 0.00011 0.00008 0.00005  
## Cumulative Proportion 0.99951 0.99968 0.99979 0.99987 0.99992  
## PC15 PC16  
## Standard deviation 139.72407 122.73943  
## Proportion of Variance 0.00005 0.00004  
## Cumulative Proportion 0.99996 1.00000

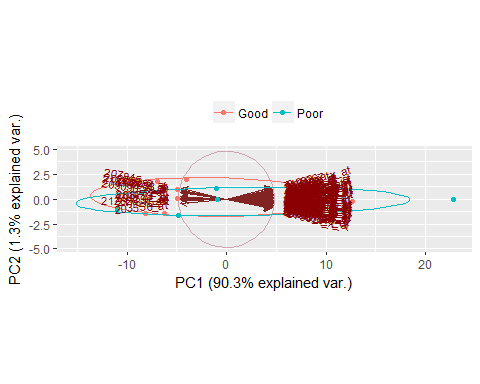






## Importance of components:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 9.501008 1.122253 1.049703 1.002219 0.9346711  
## Proportion of Variance 0.902690 0.012590 0.011020 0.010040 0.0087400  
## Cumulative Proportion 0.902690 0.915290 0.926300 0.936350 0.9450900  
## PC6 PC7 PC8 PC9 PC10  
## Standard deviation 0.9245676 0.8447119 0.8257738 0.7861957 0.7459383  
## Proportion of Variance 0.0085500 0.0071400 0.0068200 0.0061800 0.0055600  
## Cumulative Proportion 0.9536300 0.9607700 0.9675900 0.9737700 0.9793300  
## PC11 PC12 PC13 PC14 PC15  
## Standard deviation 0.7293633 0.6685855 0.6316559 0.6042104 0.5688881  
## Proportion of Variance 0.0053200 0.0044700 0.0039900 0.0036500 0.0032400  
## Cumulative Proportion 0.9846500 0.9891200 0.9931100 0.9967600 1.0000000  
## PC16  
## Standard deviation 6.579005e-15  
## Proportion of Variance 0.000000e+00  
## Cumulative Proportion 1.000000e+00





## Warning: package 'gplots' was built under R version 3.3.3

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

## Dataset

The assignment uses a leukemia dataset. The data set comprises 22283 observations (genes) and sixteen samples (patients). The leukemia dataset contains gene expression levels for patients with poor and good leukemia progonosis. Good and poor leukemia progonosis is measured in terms of the number of leukemic cells that are present in the bone marrow after a period of treatment. Patients with low numbers have a good progonosis whereas patients with high numbers have a poor progonosis. Samples (columns) 1-8 represent the patients with good leukemia prognosis while samples 9-16 represent the patients with poor leukemia prognosis.

The leukemia dataset is available on Vula under the tab: Data Science 2017 Resources/Unsupervised Learning/Assignment.

Download it and save it in your current folder. The R package will be used to carry out the assignment.

## Principal Component Analysis (PCA).

Note that when we perform PCA we assume that observations or samples are in the rows of a matrix and variables are in the columns. Here the observations are in the columns and the variables in the rows. This is the usual format in which microarray data is presented.

Carry out appropriate data transformation that will allow you to perform PCA. Consider performing PCA on:

* Original data.
* Standardised data.
* Transformed original data using the log\_transform.
* log\_transformed and standardised data.

In each case, use relevant plots and otherwise to interpret your results.

Next investigate the impact of including only the top 100 genes with the most variable expression levels across the samples.

That is excluding genes that contribute mostly noise and no information. In this case you will have to find a way of isolating the top 100 genes with the most variable expression levels across the samples. Once these genes have been isolated. Perform PCA on the reduced data.

Compare and contrast the results before and when the 100 genes with the most variable expression levels are used. What conclusions can you make.

## Clustering Analysis.

Copy the original leukemia data to a new file say data2.

Use cluster analysis to find groups of patients that have similar progonosis on the basis of the expression of (1) all genes and (2) the top 100 most variable genes as found above.

Identify and briefly describe the cluster method(s) used to formulate patient groups. That is, describe the similarity measure(s) used and whether or not the data were standardized, and briefly discuss why you chose this strategy.

Evaluate the performance of cluster analysis in meeting the first objective above. That is, evaluate the stability of the cluster solution.

## Problem Statement

## Data Description

## Conclusion

set center and scale. equal to TRUE in the call to prcomp to standardize the variables prior to the application of PCA:

But if you’re trying to combine correlated variables that all get at the size of trees, like: the trunk diameter in cm, biomass of leaves in kg, number of branches, overall height in meters–those are going to be on vastly different scales. Variables whose numbers are just larger will have much bigger variance just because the numbers are so big. (Remember that variances are squared values, so big numbers get amplified).

If you’re starting with a covariance matrix, it’s a good idea to standardize those variables before you begin so that the variables with the biggest scales don’t overwhelm the PCA.

We perform PCA to get insight of the general structure of a data set. We center, scale and sometimes log-transform to filter off some trivial effects, which could dominate our PCA. The algorithm of a PCA will in turn find the rotation of each PC to minimize the squared residuals, namely the sum of squared perpendicular distances from any sample to the PCs. Large values tend to have high leverage.

# References

* Project Description (assignment.pdf, DataScienceProject.html, leukemia\_array.txt), Author: Juwa Nyirenda. Last Accessed: 24-Jul-2017 17:51. Vula Resources.
* R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.
* Hadley Wickham (2017). tidyverse: Easily Install and Load 'Tidyverse' Packages. R package version 1.1.1. <https://CRAN.R-project.org/package=tidyverse>
* Kirill Müller and Hadley Wickham (2017). tibble: Simple Data Frames. R package version 1.3.3. <https://CRAN.R-project.org/package=tibble>
* Hadley Wickham (2017). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.6.1. <https://CRAN.R-project.org/package=tidyr>
* Hadley Wickham, Jim Hester and Romain Francois (2016). readr: Read Tabular Data. R package version 1.0.0. <https://CRAN.R-project.org/package=readr>
* Lionel Henry and Hadley Wickham (2017). purrr: Functional Programming Tools. R package version 0.2.2.2. <https://CRAN.R-project.org/package=purrr>
* Hadley Wickham and Romain Francois (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. <https://CRAN.R-project.org/package=dplyr>