# Aleksey Morozov
# Microarray Data Analysis Using Hierarchical Clustering.

# The "unsupervised learning" approach deals with data that has the features X1,X2...Xp, but does not have an associated response variable Y. As such, there is no prediction based on observations. The goal of this approach is to visualize interesting commonalities / dissimilarities, such as subgroups, between the features.

# The difficulty in using this approach is that there is no good way of assessing the validity of the results. While "supervised learning" techniques allow one to use Y values outside of those used to generate a model to check the validity of the model, these values are not available in unsupervised learning.

# Assessment of gene expression levels between several cell group types is a common application of the unsupervised technique.

# For the purpose of demonstration, the NCI60 dataset in the ISLR package will be used. The datatset comes from a microarray analysis, containing expression levels of 6,830 genes from 64 cancer cell lines, with the cancer types provided as well.

# To download the package:
```
install.packages("ISLR")
```
# To load the package:
```
library(ISLR)
```

# As mentioned in the NCI60 dataset description, each of the 64 cell lines is associated with a particular cancer type. To view the cancer types associated with each cell line, use the following command:

```
cline <- NCI60$labs # cline for "cancer line"
cline
```
```
[1] "CNS"         "CNS"         "CNS"         "RENAL"       "BREAST"
[6] "CNS"         "CNS"         "BREAST"      "NSCLC"       "NSCLC"
[11] "RENAL"       "RENAL"       "RENAL"       "RENAL"       "RENAL"
[16] "RENAL"       "RENAL"       "BREAST"      "NSCLC"       "RENAL"
[21] "UNKNOWN"     "OVARIAN"     "MELANOMA"    "PROSTATE"    "OVARIAN"
[26] "OVARIAN"     "OVARIAN"     "OVARIAN"     "OVARIAN"     "PROSTATE"
[31] "NSCLC"       "NSCLC"       "NSCLC"       "LEUKEMIA"    "K562B-repro"
[36] "K562A-repro" "LEUKEMIA"    "LEUKEMIA"    "LEUKEMIA"    "LEUKEMIA"
[41] "LEUKEMIA"    "COLON"       "COLON"       "COLON"       "COLON"
[46] "COLON"       "COLON"       "COLON"       "MCF7A-repro" "BREAST"
[51] "MCF7D-repro" "BREAST"      "NSCLC"       "NSCLC"       "NSCLC"
[56] "MELANOMA"    "BREAST"      "BREAST"      "MELANOMA"    "MELANOMA"
[61] "MELANOMA"    "MELANOMA"    "MELANOMA"    "MELANOMA"
```

# To view how many cell lines fall under each of the cancer types:
table(cline) # The results viewed here will be the basis for evaluating the validity of the results obtained using clustering, with the assumption that cells of the same cancer type will show gene expression that will cluster closer than to cells of another cancer type.

```
> table(cline)
cline
     BREAST         CNS       COLON K562A-repro K562B-repro    LEUKEMIA
          7           5           7           1           1           6
MCF7A-repro MCF7D-repro    MELANOMA       NSCLC     OVARIAN    PROSTATE
          1           1           8           9           6           2
      RENAL     UNKNOWN
          9           1
```

# Clustering is a way of finding distinct groups in a set of data. Data is clustered in such a way that observations within a particular group are closer to one another than to observations in a different group. The criterion for difference is left up to the "domain specialist" to establish.
# Many clustering methods exist, each with their own field for application. The Clustering method used for this data set will be "hierarchical clustering". In hierarchical clustering, there is no predetermined number of clusters that we expect to observe. Although we know that there are 14 distinct cancer types present in the dataset, and hypothesize that expression patterns will correlate to the cancer types, this approach will remove analysis bias. Alternatively, "K-means clustering" allows for specification of desired number of clusters.
# Hierarchical clustering generates dendrograms based on the similarity / dissimilarity of the data, with the measure of similarity being Euclidian distance.

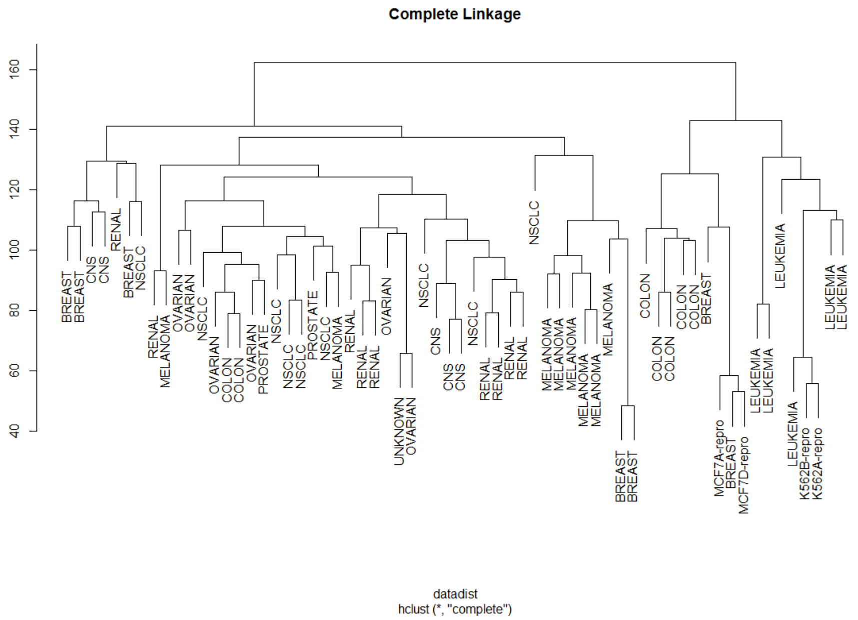
# We first want to standardize the data to have a mean of zero and a standard deviation of one.
expdata <- NCI60$data # Assigns expression data to variable expdata
stddata <- scale(expdata) # scale() standardizes the data to above parameters.

# To perform hierarchical clustering:

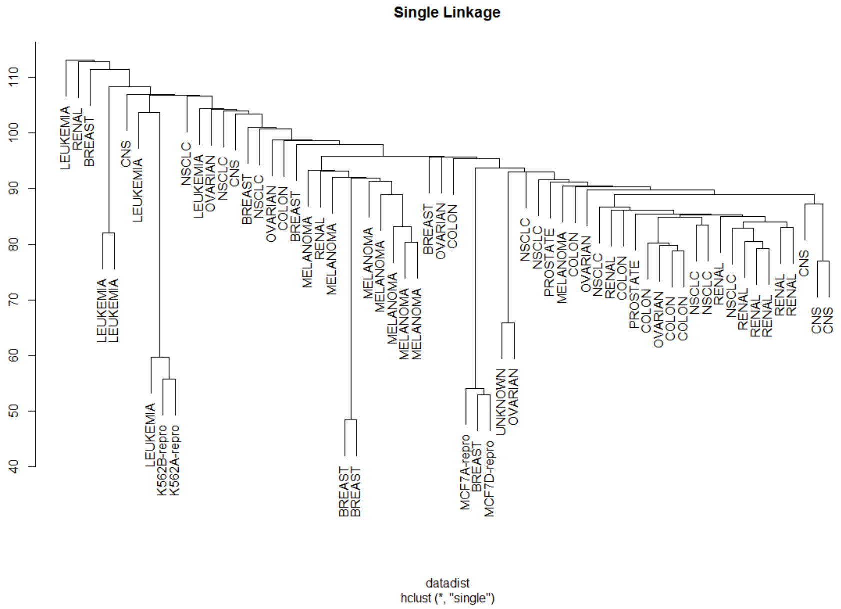
par(mfrow <- c(1,3))
datadist <- dist(stddata) # dist() function computes the distances between the rows of a data matrix.

# The three most commonly used types of linkages in hierarchical clustering are: Complete, Single, and Average. The Complete linkage computes the largest dissimilarities between clusters. The Single linkage computes the smallest dissimilarities between clusters. The Average linkage computes all dissimilarities and then records the average of them.
distcomp <- hclust(datadist, method="complete")
distsingle <- hclust(datadist, method="single")
distavg <- hclust(datadist, method="average")

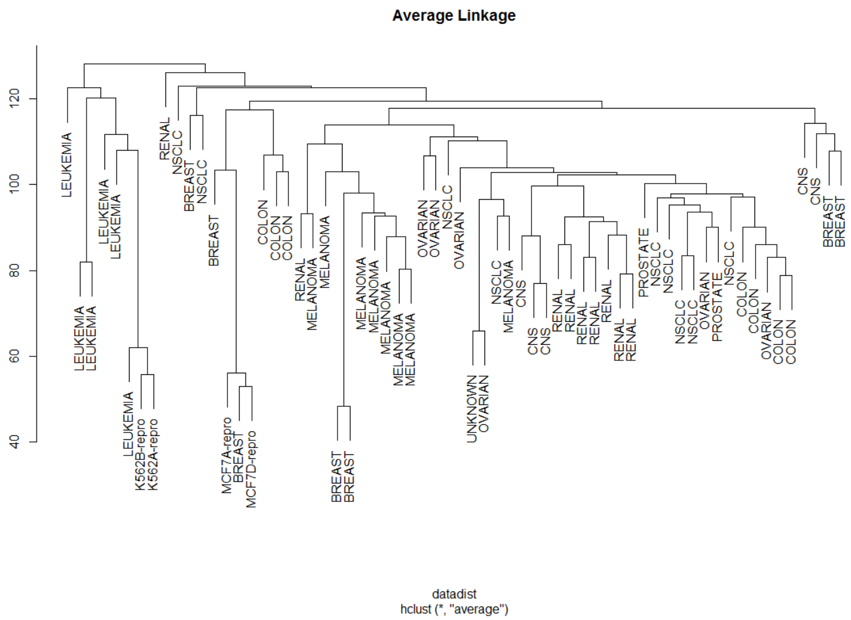
plot(distcomp, labels=cline, main="Complete Linkage", xlab = NULL, ylab = NULL) # Complete Linkage



plot(distsingle, labels=cline, main="Single Linkage", xlab = NULL, sub = NULL, ylab = NULL) # Single Linkage



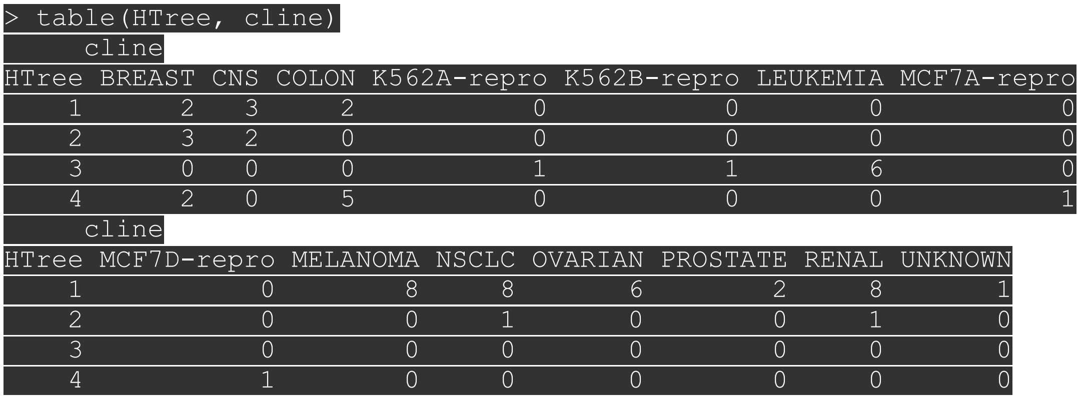
plot(distavg, labels=cline, main="Average Linkage", xlab = NULL, sub = NULL, ylab = NULL)
# Average Linkage



# The above results demonstrate that the method of linkage will affect the results. For the purpose of demonstration, Complete linkage will be explored further, as it will highlight the greatest differences between samples.

HiClust <- hclust(dist(stddata))
HTree <- cutree(HiClust, 4) # cutree() function allows the user to specify a cut off point for how many clusters they want the tree to have. The cut off of 4 is arbitrarily chosen.

table(HTree, cline) # Creates a table showing into which of the four clusters do the different cancer type cell lines fall. For leukemia, it appears that all cell lines fall into the third cluster, as do the K562A- and K562B-repro cell lines. All of the melanoma cell lines fall into cluster #1, as do all of the ovarian cell lines.

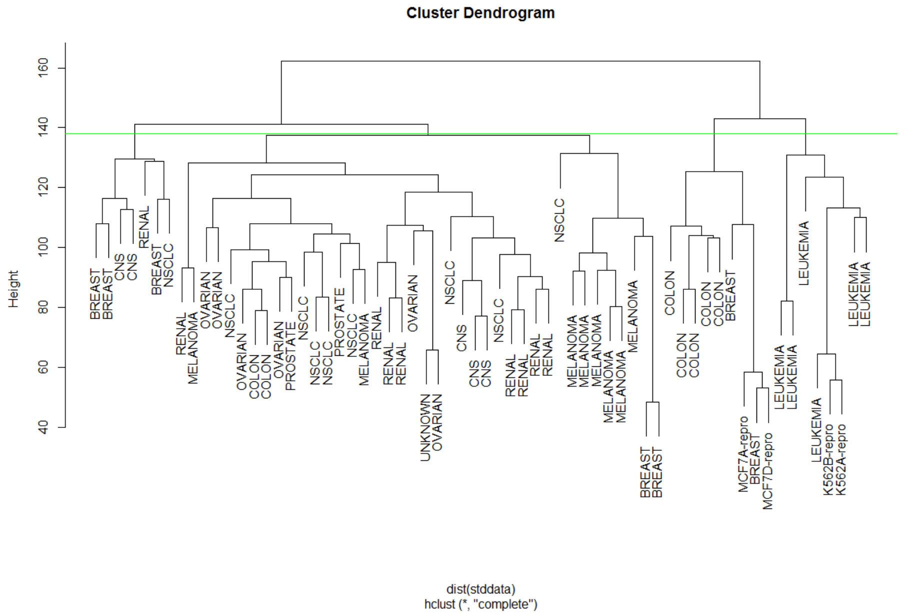


```
> table(HTree, cline)
     cline
HTree BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro
    1      2   3     2           0           0        0           0
    2      3   2     0           0           0        0           0
    3      0   0     0           1           1        6           0
    4      2   0     5           0           0        0           1
     cline
HTree MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
    1           0        8     8       6        2     8       1
    2           0        0     1       0        0     1       0
    3           0        0     0       0        0     0       0
    4           1        0     0       0        0     0       0
```

# To convert the data seen in the above table into a dengrogram, showing the clustering:

par(mfrow <- c(1,1))
plot(HiClust, labels=cline)
abline(h=138, col="green") # abline() function draws a line on an existing plot. The 138 in the argument h=138 was selected because that is the value that results in the four distinct clusters

that we want. It was obtained by looking at the plot and selecting a value at which the abline intersects the tree at only four points.



# Parameters such as the cut off point and linkage method can be changes to obtain different results and draw different conclusions. The parameters are at the discretion of the domain specialist. Based on the parameters established here, it would appear that the mutation that leads to leukemia is localized to perhaps a single family of genes, thus giving a relatively similar expression pattern. On the other hand, the gene expression observed in breast cancer cell lines is divided between three clusters, suggesting that there may be several mutations, all impacting different regulatory mechanisms, that are responsible. Lastly, due to the proximity on the tree, it would appear that the "Unknown" cancer cell line is most closely related to the ovarian cancer cell line.