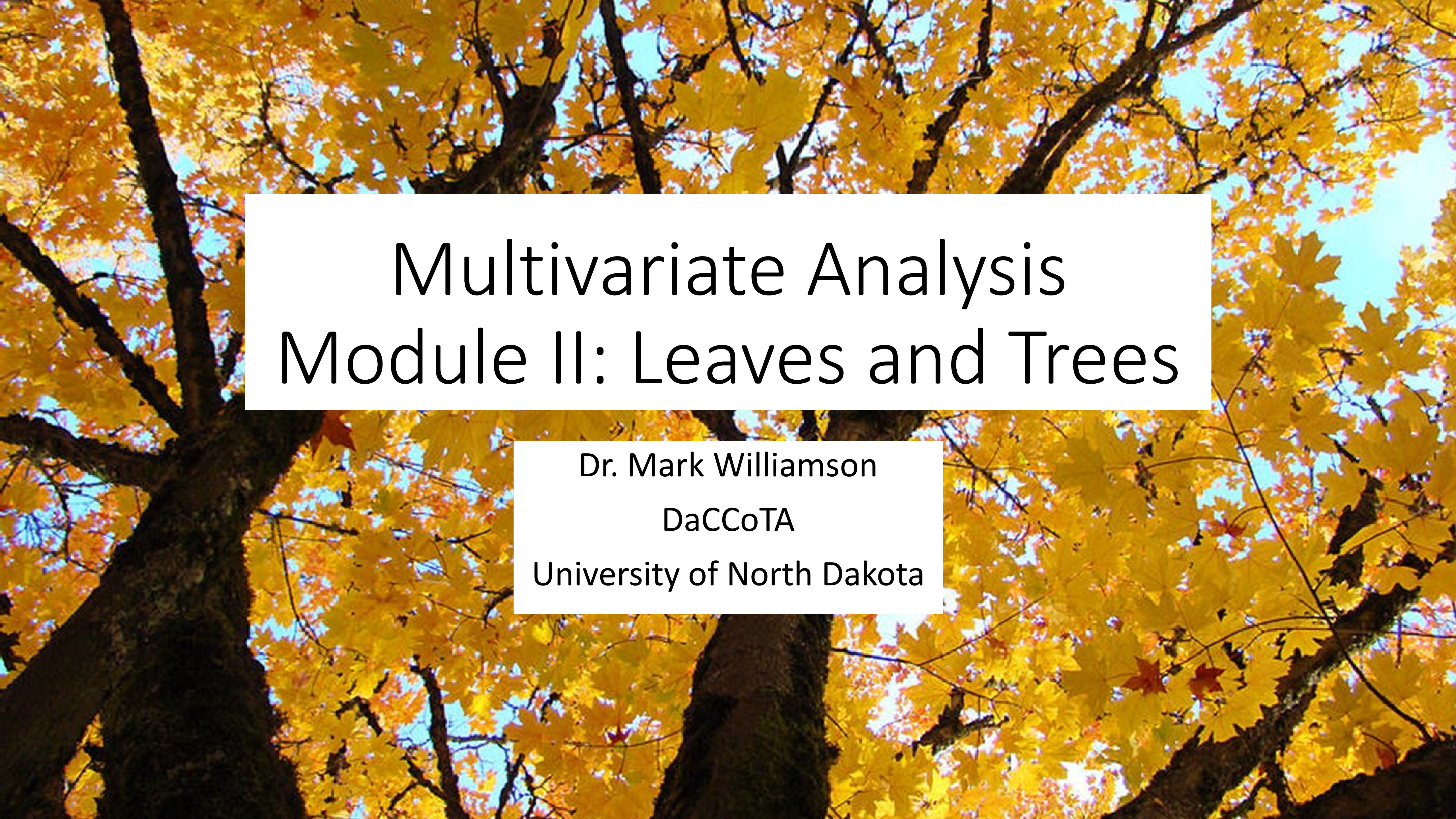


Multivariate Analysis Module II: Leaves and Trees

Dr. Mark Williamson

DaCCoTA

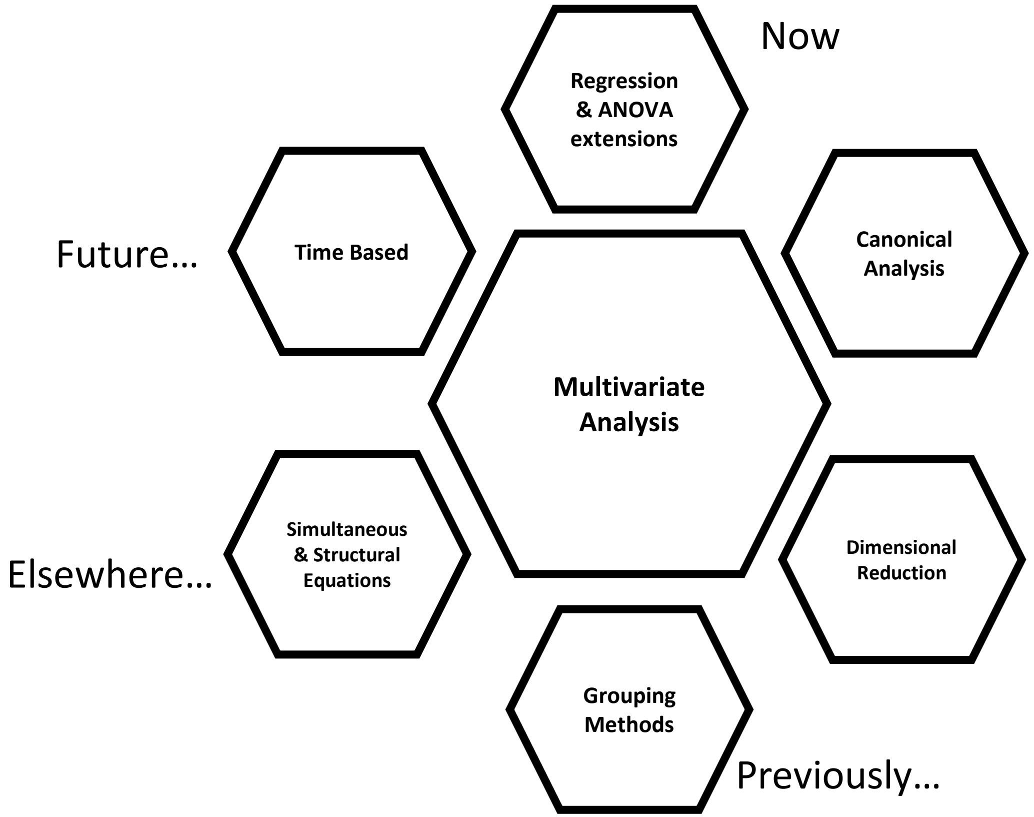
University of North Dakota

Introduction

- Last time, we covered a broad overview of multivariate analysis

<u>Multivariate:</u>
multiple dependent variables or other more complicated structures such as ordination or non-linearity

- Today we'll cover more details about specific multivariate methods



Now

Regression & ANOVA extensions

Future…

Time Based

Canonical Analysis

Multivariate Analysis

Elsewhere…

Simultaneous & Structural Equations

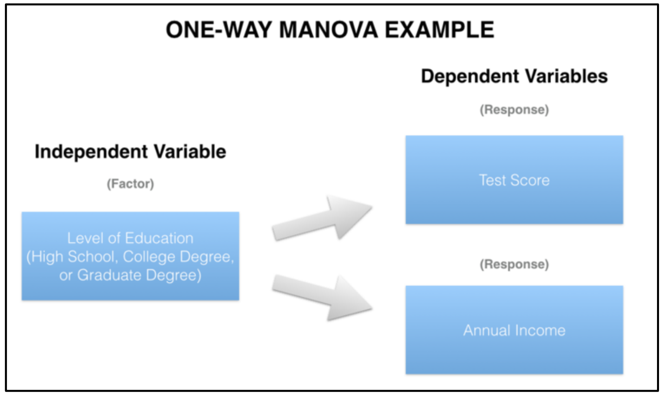
Dimensional Reduction

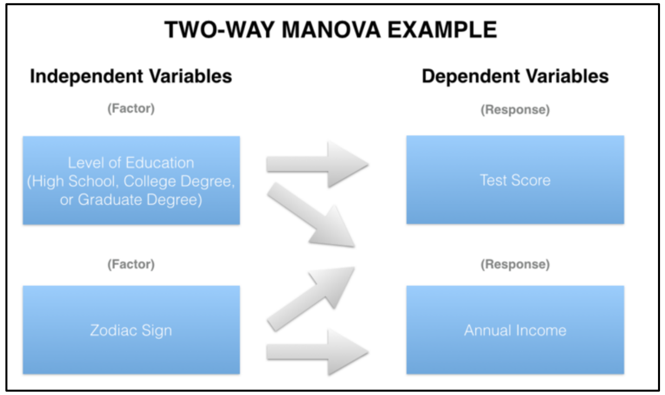
Grouping Methods

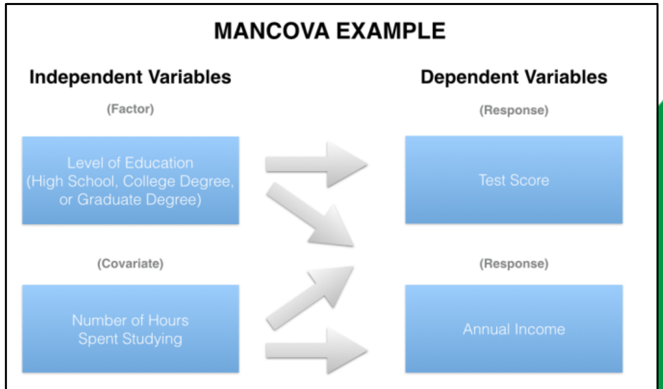
Previously…

Descriptions

Name	Description
MANOVA	Multivariate analysis of variance. Used to test the statistical significance of the effect of one or more independent variables (categorical) on a set of two or more dependent variables
MANCOVA	Multivariate analysis of covariance. Same as MANOVA but after controlling for covariate(s).
Multivariate Regression	Used to test the statistical significance of the effect of one or more independent variables (numerical) on a set of two or more dependent variables
K-means clustering	Clustering method that partitions observations into k number of clusters, where each observation belongs to the cluster with the nearest mean. Also known as centroid-based clustering
Hierarchical clustering	Clustering method that separates observations based on a measure of similarity using a tree-based approach either from the bottom-up (agglomerative) or top-down (divisive)
Density-based clustering	Clustering method that connects areas where observations are high density and allows for arbitrary-shaped clusters.
Distribution-based clustering	Clustering method that assumes observations come from a certain distribution (such as Gaussian) and groups them with decreasing probability from the distribution's center.
Classification tree	Recursive partitioning decision tree in which target variables are categorical
Regression tree	Recursive partitioning decision tree in which target variables are numerical.

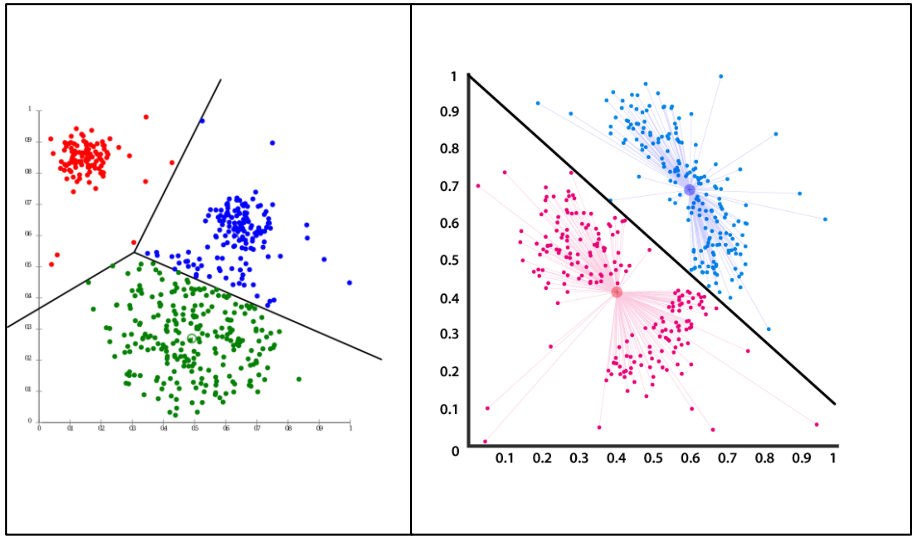

ONE-WAY MANOVA EXAMPLE

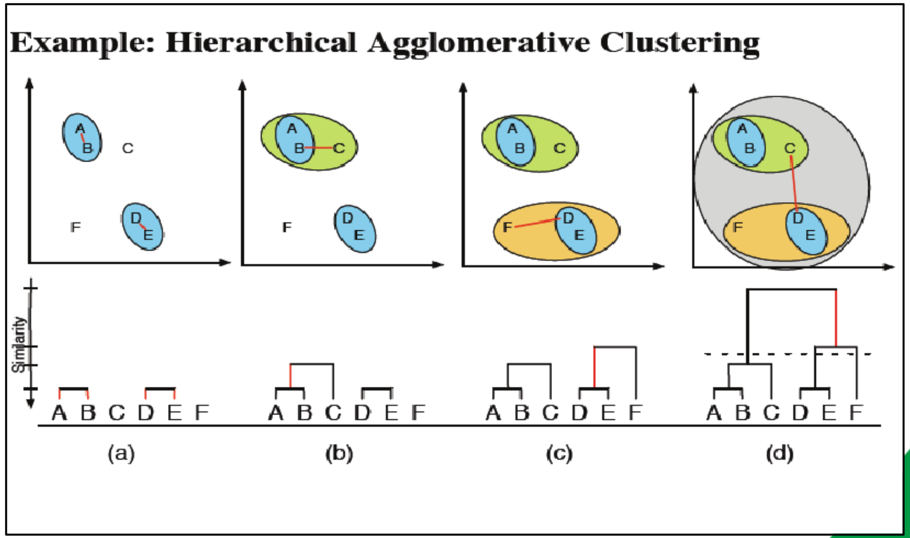

TWO-WAY MANOVA EXAMPLE


MANCOVA EXAMPLE

Descriptions

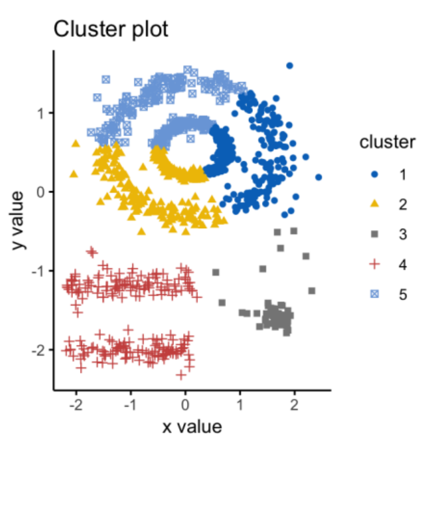
Name	Description
MANOVA	Multivariate analysis of variance. Used to test the statistical significance of the effect of one or more independent variables (categorical) on a set of two or more dependent variables
MANCOVA	Multivariate analysis of covariance. Same as MANOVA but after controlling for covariate(s).
Multivariate Regression	Used to test the statistical significance of the effect of one or more independent variables (numerical) on a set of two or more dependent variables
K-means clustering	Clustering method that partitions observations into k number of clusters, where each observation belongs to the cluster with the nearest mean. Also known as centroid-based clustering
Hierarchical clustering	Clustering method that separates observations based on a measure of similarity using a tree-based approach either from the bottom-up (agglomerative) or top-down (divisive)
Density-based clustering	Clustering method that connects areas where observations are high density and allows for arbitrary-shaped clusters.
Distribution-based clustering	Clustering method that assumes observations come from a certain distribution (such as Gaussian) and groups them with decreasing probability from the distribution's center.
Classification tree	Recursive partitioning decision tree in which target variables are categorical
Regression tree	Recursive partitioning decision tree in which target variables are numerical.

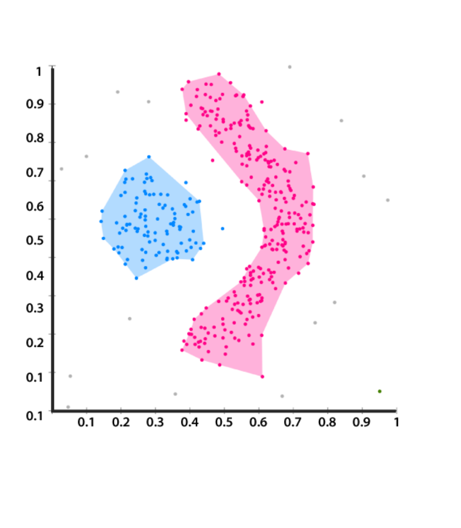


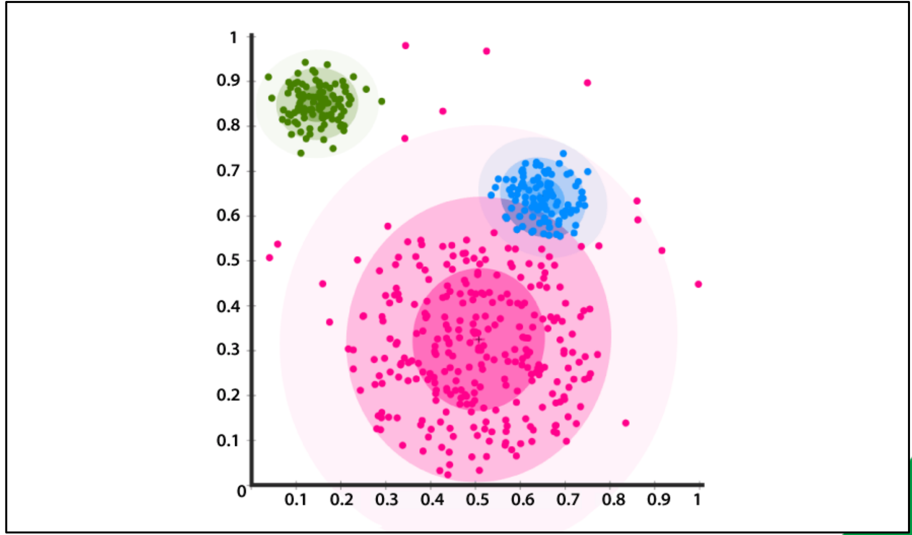

Example: Hierarchical Agglomerative Clustering

Descriptions

Name	Description
MANOVA	Multivariate analysis of variance. Used to test the statistical significance of the effect of one or more independent variables (categorical) on a set of two or more dependent variables
MANCOVA	Multivariate analysis of covariance. Same as MANOVA but after controlling for covariate(s).
Multivariate Regression	Used to test the statistical significance of the effect of one or more independent variables (numerical) on a set of two or more dependent variables
K-means clustering	Clustering method that partitions observations into k number of clusters, where each observation belongs to the cluster with the nearest mean. Also known as centroid-based clustering
Hierarchical clustering	Clustering method that separates observations based on a measure of similarity using a tree-based approach either from the bottom-up (agglomerative) or top-down (divisive)
Density-based clustering	Clustering method that connects areas where observations are high density and allows for arbitrary-shaped clusters.
Distribution-based clustering	Clustering method that assumes observations come from a certain distribution (such as Gaussian) and groups them with decreasing probability from the distribution's center.
Classification tree	Recursive partitioning decision tree in which target variables are categorical
Regression tree	Recursive partitioning decision tree in which target variables are numerical.

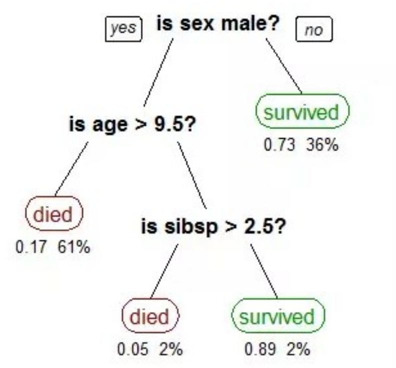


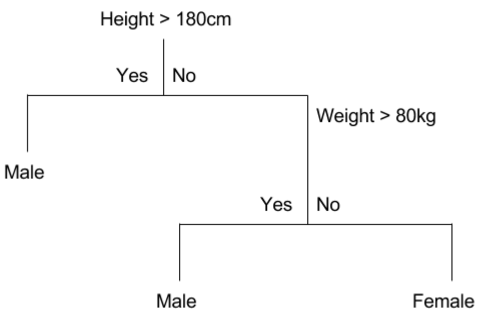


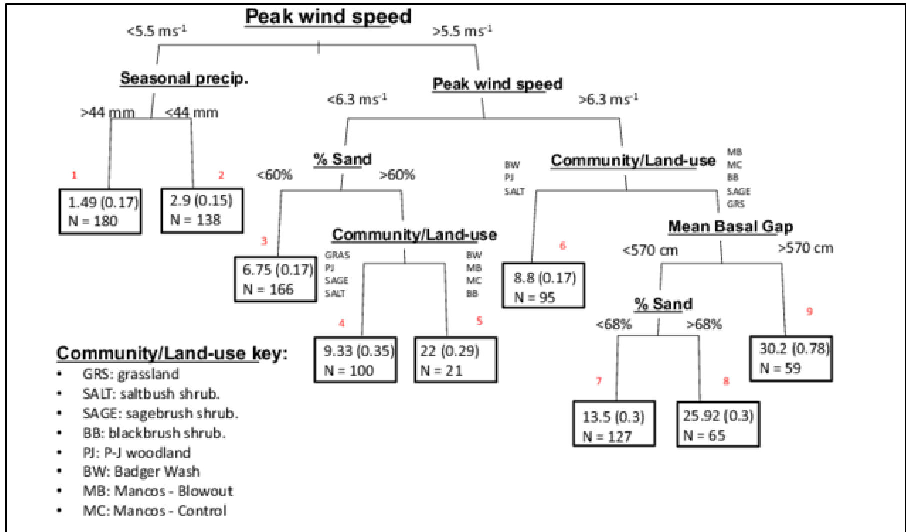


Descriptions

Name	Description
MANOVA	Multivariate analysis of variance. Used to test the statistical significance of the effect of one or more independent variables (categorical) on a set of two or more dependent variables
MANCOVA	Multivariate analysis of covariance. Same as MANOVA but after controlling for covariate(s).
Multivariate Regression	Used to test the statistical significance of the effect of one or more independent variables (numerical) on a set of two or more dependent variables
K-means clustering	Clustering method that partitions observations into k number of clusters, where each observation belongs to the cluster with the nearest mean. Also known as centroid-based clustering
Hierarchical clustering	Clustering method that separates observations based on a measure of similarity using a tree-based approach either from the bottom-up (agglomerative) or top-down (divisive)
Density-based clustering	Clustering method that connects areas where observations are high density and allows for arbitrary-shaped clusters.
Distribution-based clustering	Clustering method that assumes observations come from a certain distribution (such as Gaussian) and groups them with decreasing probability from the distribution's center.
Classification tree	Recursive partitioning decision tree in which target variables are categorical
Regression tree	Recursive partitioning decision tree in which target variables are numerical.







Rationales

When and Why should you use multivariate analysis in general?

- Complex for a complex world
- Don’t use if you don’t understand it
- Don’t use if a simpler method works

When should you use specific multivariate analysis methods?

What are the assumptions of multivariate analysis?

Name	Usage	Assumptions
1-Way MANOVA	Multiple numerical Y-variables (Responses), Single categorical X-variable (Factor)	• Independence of observations • Multivariate normality • Linearity (Y-vars) • No multicollinearity (X-vars) • Equality of variance • Equality of variance-covariance matrices
2-Way MANOVA	Multiple numerical Y-variables (Responses), Two categorical X-variables (Factors)	
1-WAY MANCOVA	Multiple numerical Y-variables (Responses), Single categorical X-variable of interest (Factor), Single numerical X-variable controlled for (Covariate)	
2-WAY MANCOVA	Multiple numerical Y-variables (Responses), Two categorical X-variable of interest (Factors), Single numerical X-variable controlled for (Covariate)	
Multivariate Regression	Multiple numerical Y-variables (Responses), numerical X-variables (Predictors)	• Same as MANOVA/MANCOVA except variance
K-means clustering	[Categorical Y-variable], Numerical X-variables, set number of clusters	• N/A
Hierarchical clustering	[Categorical Y-variable], Numerical X-variables, non-specified cluster number	
Distribution-based clustering	[Categorical Y-variable], Numerical X-variables, known distribution	
Density-based clustering	[Categorical Y-variable], Numerical X-variables, dataset with noise and/or outliers	
Classification tree	Categorical Y-variable (Outcome), Numerical or Categorical X-variables	• N/A
Regression tree	Continuous Y-variable (Outcome), Numerical or Categorical X-variables	

Step-by-step Example 1

MANOVA, MANCOVA, and multivariate regression using the mtcars dataset

A. Is there a significant effect of transmission category (automatic/manual) and gear category (3/4/5) on MPG and quarter-mile time?
B. Is there a significant effect of transmission category (automatic/manual) and gear category (3/4/5) on MPG and quarter-mile time, while accounting for weight?
C. Is there a significant effect of displacement, gross horsepower, and rear axle ratio on MPG and quarter-mile time?

Step-by-step Example 1

Set-Up

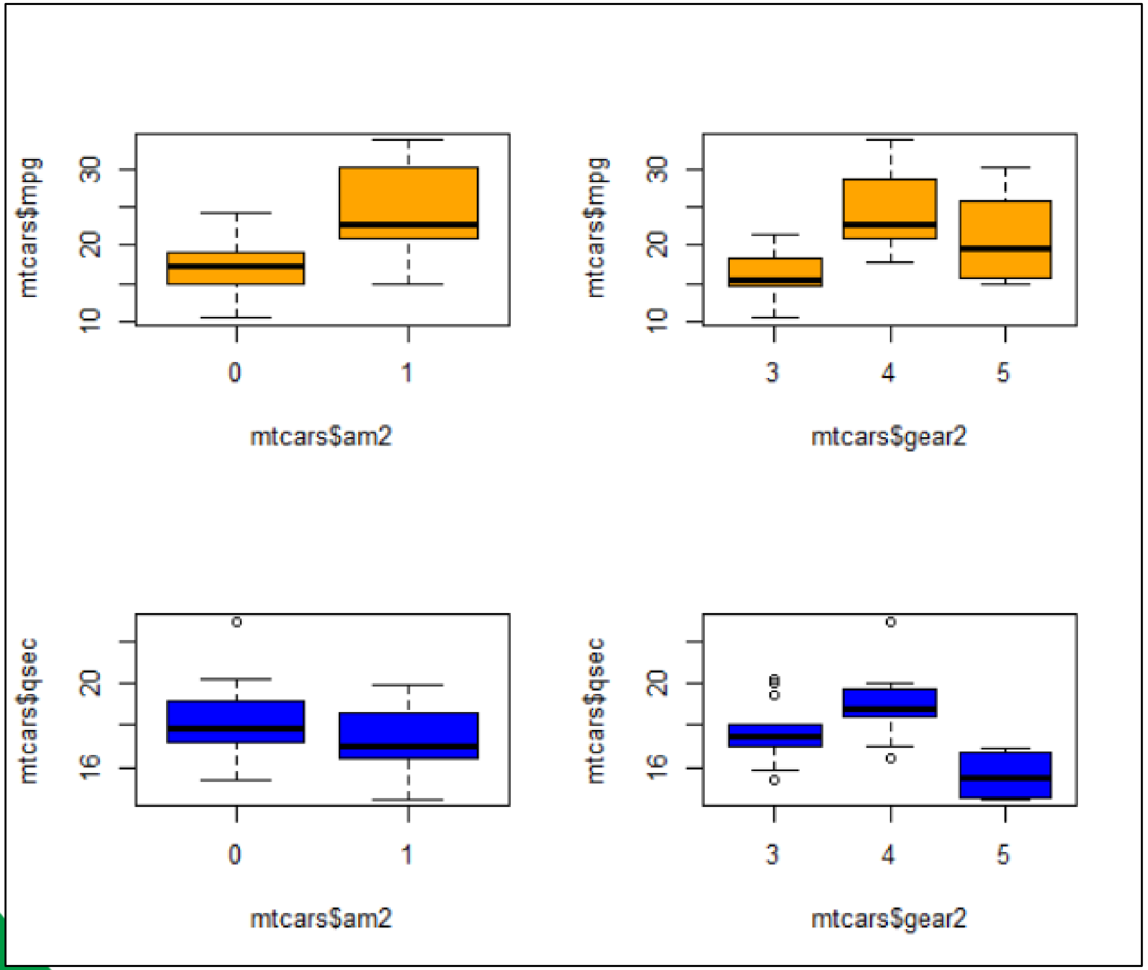
```
#Intro stuff:

library(rstatix)

library(plyr)

library(tidyverse)


head(mtcars)


#outcomes (mpg and qsec); categorical
predictors (am, gear); covariate (wt)


mtcars$am2 <-as.factor(mtcars$am)

mtcars$gear2 <-as.factor(mtcars$gear)
```

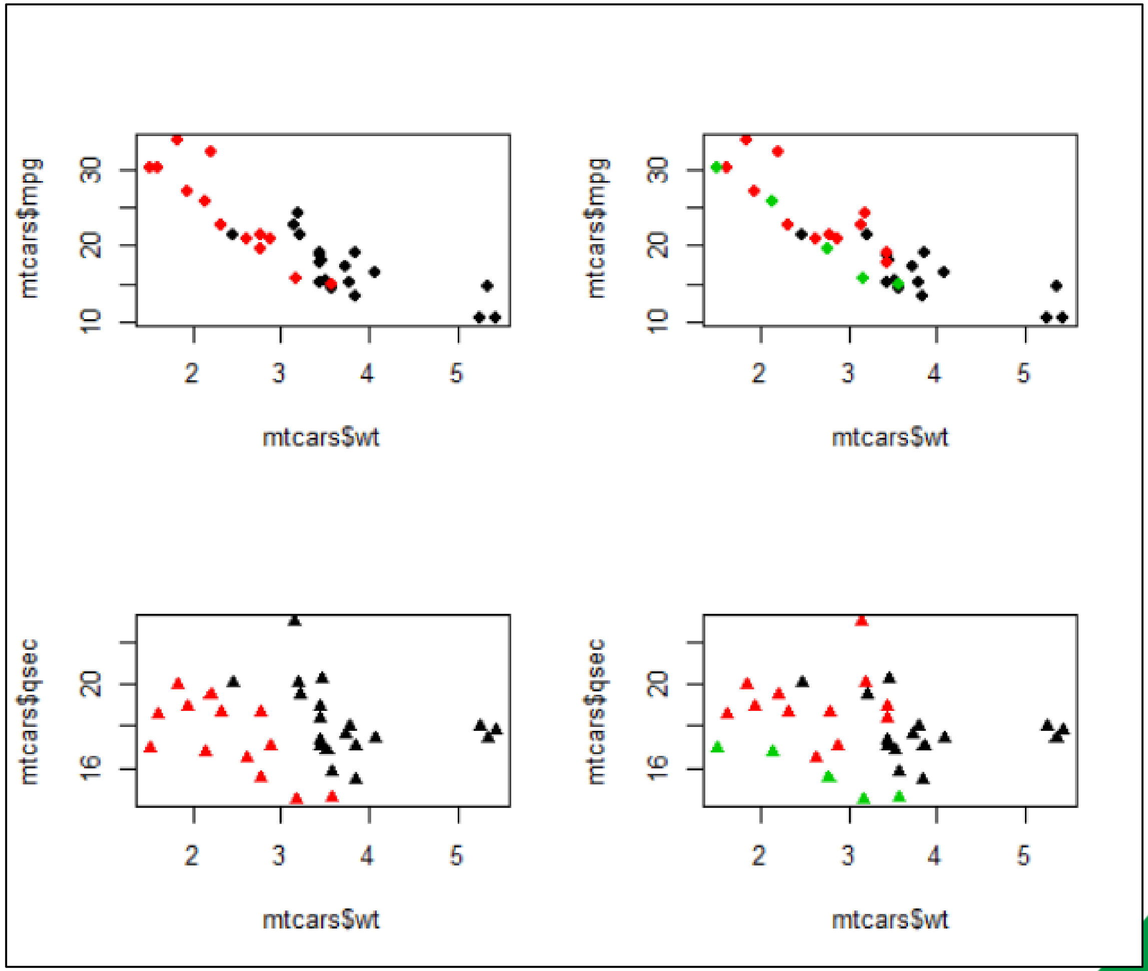
MANOVA/MANCOVA data exploration

```
#data visualization
par(mfrow=c(2,2))

plot(mtcars$mpg~mtcars$am2, col='orange')
plot(mtcars$mpg~mtcars$gear2, col='orange')
plot(mtcars$qsec~mtcars$am2, col='blue')
plot(mtcars$qsec~mtcars$gear2, col='blue')

plot(mtcars$mpg~mtcars$wt, col=mtcars$am2, pch=16)
plot(mtcars$mpg~mtcars$wt, col=mtcars$gear2, pch=16)
plot(mtcars$qsec~mtcars$wt, col=mtcars$am2, pch=17)
plot(mtcars$qsec~mtcars$wt, col=mtcars$gear2, pch=17)

par(mfrow=c(1,1))
```

https://www.datanovia.com/en/lessons/one-way-manova-in-r/

Step-by-step Example 1

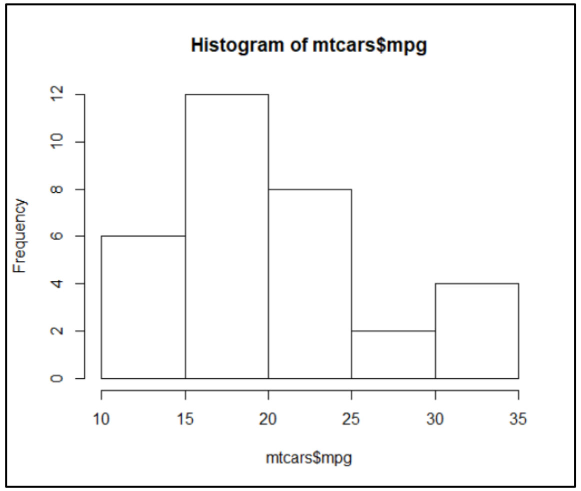


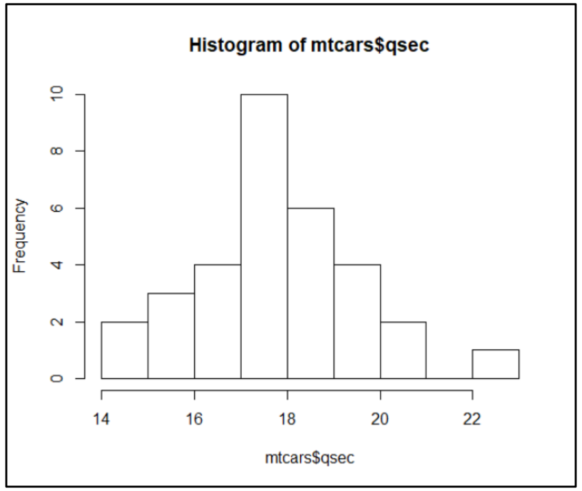


Step-by-step Example 1

Testing MANOVA/MANCOVA assumptions

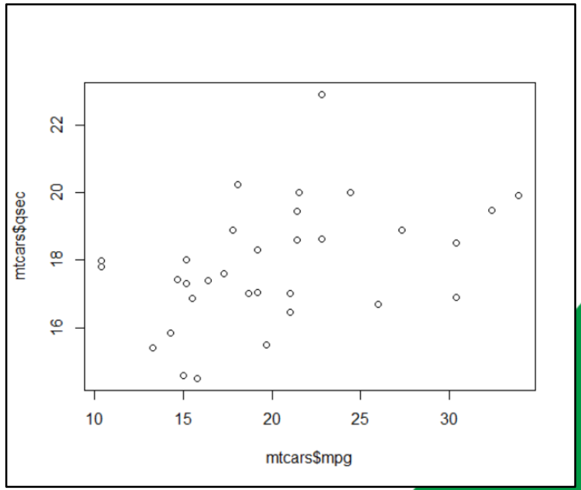
```
#normality
        hist(mtcars$mpg)
        hist(mtcars$qsec) #good enough

        mtcars %>%
          select(mpg, qsec) %>%
          mshapiro_test() #good

#multicollinearity
        cor.test(mtcars$mpg, mtcars$qsec) #good

#linearity
        plot(mtcars$mpg, mtcars$qsec) #good
```





```
 statistic p.value
     <dbl>   <dbl>
1    0.967   0.420
```

```
Pearson's product-moment correlation

data:  mtcars$mpg and mtcars$qsec
t = 2.5252, df = 30, p-value = 0.01708
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.08195487 0.66961864
sample estimates:
     cor
0.418684
```



Step-by-step Example 1

Testing MANOVA/MANCOVA assumptions cont.

```
#homogeneity of variance and covariance
      ldply(mtcars[,9:10],function(x) t(rbind(names(table(x)), table(x),
      paste0(prop.table(table(x))*100, "%"))))

      box_m(mtcars[, c("mpg", "qsec")], mtcars$am2) #good
      box_m(mtcars[, c("mpg", "qsec")], mtcars$gear2) #significant (Pillai's)

      mtcars %>%
        gather(key = "variable", value = "value", mpg, qsec) %>%
        group_by(variable) %>%
        levene_test(value ~ am2) #good for mpg, not good for qsec

      mtcars %>%
        gather(key = "variable", value = "value", mpg, qsec) %>%
        group_by(variable) %>%
        levene_test(value ~ gear2) #not good for mpg, not good for qsec
```

```
statistic p.value parameter method
    <dbl>   <dbl>     <dbl> <chr>
1    4.11   0.250         3 Box's M-test for Homogeneity of Covariance Matrices
```

```
 statistic p.value parameter method
    <dbl>   <dbl>     <dbl> <chr>
1    14.0  0.0298         6 Box's M-test for Homogeneity of Covariance Matrices
```

```
  variable   df1   df2 statistic      p
   <chr>   <int> <int>     <dbl>  <dbl>
1 mpg          1    30     4.19  0.0496
2 qsec         1    30     0.322 0.575
```

```
  variable   df1   df2 statistic     p
   <chr>   <int> <int>     <dbl> <dbl>
1 mpg          2    29    1.49   0.242
2 qsec         2    29    0.0491 0.952
```

Step-by-step Example 1

MANOVA results

```
manova1 <- manova(cbind(mpg, qsec)~ am2 + gear2, data=mtcars)
summary(manova1)
summary.aov(manova1)
```

```
 Df  Pillai approx F num Df den Df    Pr(>F)
am2        1 0.64944  25.0093      2     27 7.151e-07 ***
gear2      2 0.44423   3.9976      4     56  0.006369 **
Residuals 28
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Response mpg :
            Df Sum Sq Mean Sq F value    Pr(>F)
am2          1 405.15  405.15 19.9021 0.0001208 ***
gear2        2 150.89   75.45  3.7062 0.0373294 *
Residuals   28 570.00   20.36
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 Response qsec :
            Df Sum Sq Mean Sq F value    Pr(>F)
am2          1  5.230  5.2301  2.7877 0.1061372
gear2        2 41.225 20.6125 10.9865 0.0003006 ***
Residuals   28 52.533  1.8762
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Step-by-step Example 1

MANCOVA results

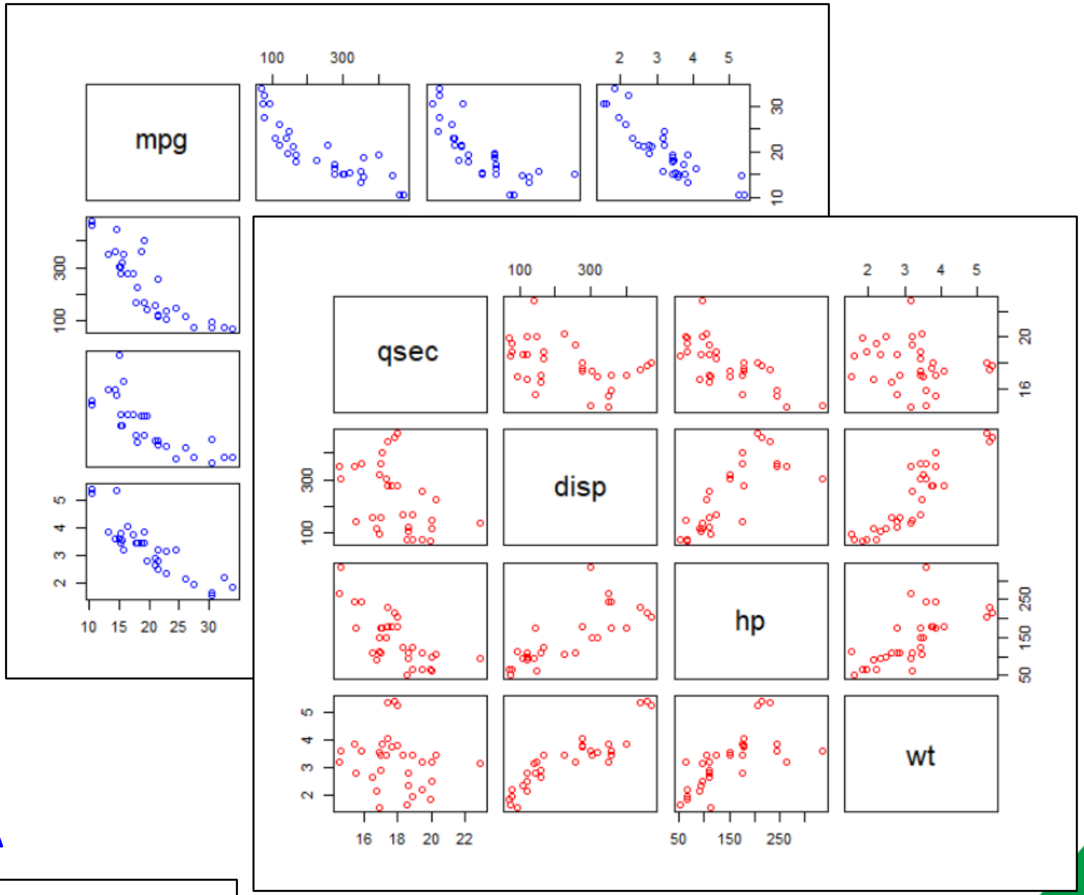
```
mancova1 <-manova(cbind(mpg, qsec)~ am2 + gear2 + wt,
data=mtcars)
summary(mancova1)
summary.aov(mancova1)
```

```
Df  Pillai approx F num Df den Df    Pr(>F)
am2        1 0.76412   42.113      2     26 6.997e-09 ***
gear2      2 0.51596    4.694      4     54 0.002526 **
wt         1 0.59288   18.931      2     26 8.442e-06 ***
Residuals 27
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response mpg :
          Df Sum Sq Mean Sq F value    Pr(>F)
am2        1 405.15  405.15 45.9872 2.778e-07 ***
gear2      2 150.89   75.45  8.5637  0.001318 **
wt         1 332.13  332.13 37.6990 1.464e-06 ***
Residuals   27 237.87    8.81
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 Response qsec :
          Df Sum Sq Mean Sq F value    Pr(>F)
am2        1  5.230  5.2301  3.1824 0.0856791 .
gear2      2 41.225 20.6125 12.5422 0.0001406 ***
wt         1  8.160  8.1598  4.9650 0.0343877 *
Residuals   27 44.373  1.6435
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Step-by-step Example 1

Multivariate Regression data exploration

```
#mpg
mtcars_mpg <- data.frame(mpg=mtcars$mpg,
disp=mtcars$disp, hp=mtcars$hp, wt=mtcars$wt)
pairs(mtcars_mpg, col=“blue”)

#qsec
mtcars_qsec <- data.frame(qsec=mtcars$qsec,
disp=mtcars$disp, hp=mtcars$hp, wt=mtcars$wt)
pairs(mtcars_qsec, col=“red”)
```



Multivariate Regression assumptions

```
#normality and linearity
  #good,based on previously done with MANOVA/MANCOVA

#multicollinearity the X-variables
cor(mtcars[,4:6])
```

```
             hp       drat         wt
hp    1.0000000 -0.4487591  0.6587479
drat -0.4487591  1.0000000 -0.7124406
wt    0.6587479 -0.7124406  1.0000000
```

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

UND
UNIVERSITY OF NORTH DAKOTA

Step-by-step Example 1

Multivariate Regression results

```
mvreg1 <-lm(cbind(mpg,qsec) ~ disp + hp + wt, data=mtcars)
summary(mvreg1)
```

```
Response mpg :

Call:
lm(formula = mpg ~ disp + hp + wt, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-3.891 -1.640 -0.172  1.061  5.861

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.105505   2.110815  17.579  < 2e-16 ***
disp        -0.000937   0.010350  -0.091  0.92851
hp          -0.031157   0.011436  -2.724  0.01097 *
wt          -3.800891   1.066191  -3.565  0.00133 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.639 on 28 degrees of freedom
Multiple R-squared:  0.8268, Adjusted R-squared:  0.8083
F-statistic: 44.57 on 3 and 28 DF,  p-value: 8.65e-11
```

```
Response qsec :

Call:
lm(formula = qsec ~ disp + hp + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8121 -0.3125 -0.0245  0.3544  3.3693

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.965050   0.849663  21.144  < 2e-16 ***
disp        -0.006622   0.004166  -1.590  0.12317
hp          -0.022953   0.004603  -4.986 2.88e-05 ***
wt           1.485283   0.429172   3.461  0.00175 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.062 on 28 degrees of freedom
Multiple R-squared:  0.6808, Adjusted R-squared:  0.6466
F-statistic: 19.91 on 3 and 28 DF,  p-value: 4.134e-07
```

Assessment 1

qualtrics XM

UNIVERSITY OF NORTH DAKOTA UND

https://und.qualtrics.com/jfe/form/SV_d7ozvbwOekrBahU

Step-by-step Example 2

Clustering and recursive partitioning using the mtcars dataset

A. Can cars be clustering into groups by using the car characteristic variables?

B. Can car MPG or engine type (V-shaped or straight) be predicted using car characteristic variables?

Step-by-step Example 2

Setup

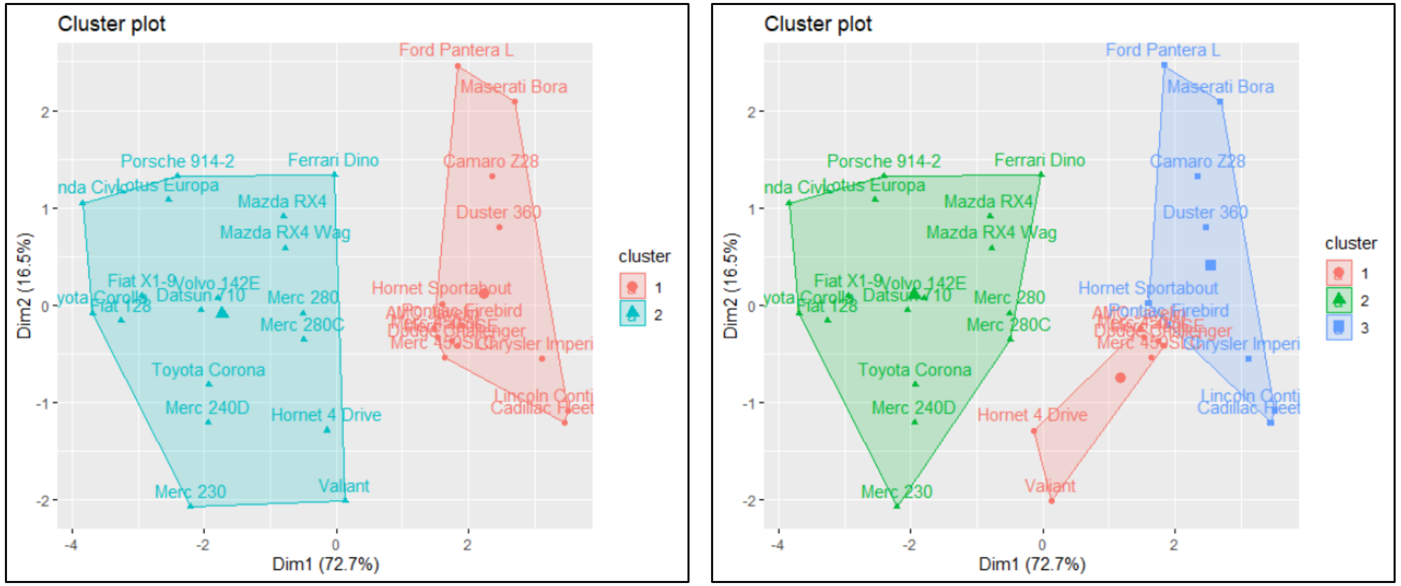
```
#Intro stuff:
library(cluster)
library(factoextra)
library(dbscan)
library(mclust)

head(mtcars)
mtcars2 <- mtcars[,1:7]
head(mtcars2)
```

https://data-flair.training/blogs/clustering-in-r-tutorial/

Step-by-step Example 2

K-means clustering

```
kmeans1 <- kmeans(mtcars2, centers=2, nstart=100)
str(kmeans1)
fviz_cluster(kmeans1, data=mtcars2)

kmeans2 <- kmeans(mtcars2, centers=3, nstart=100)
fviz_cluster(kmeans2, data=mtcars2)
```

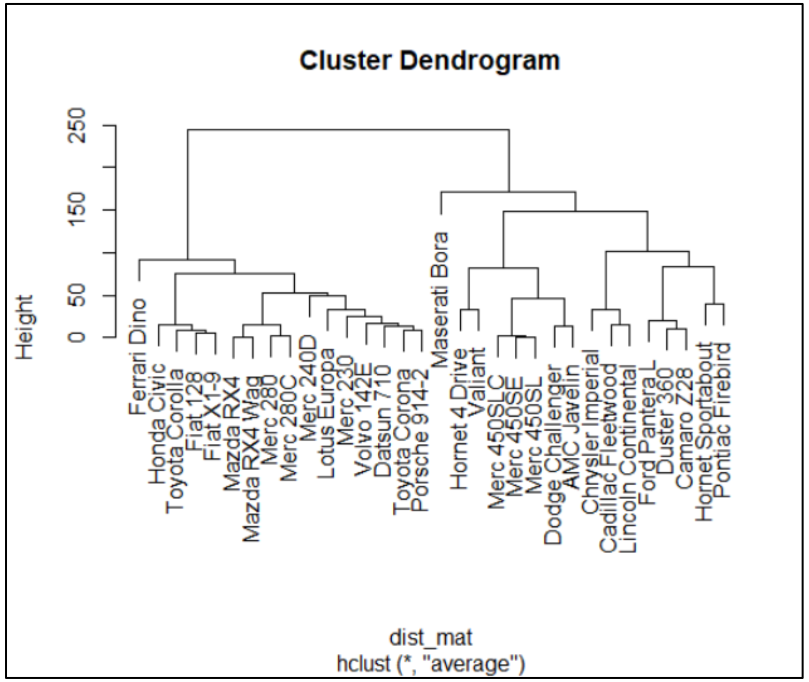


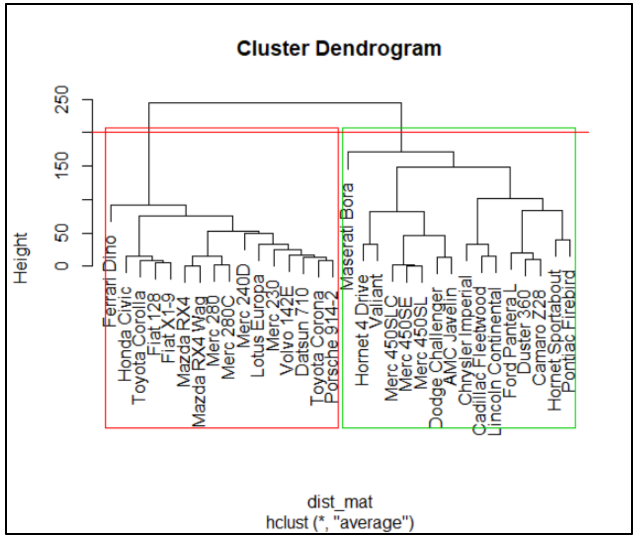
```
List of 9
 $ cluster     : Named int [1:32] 2 2 2 2 1 2 1 2 2 2 ...
  ..- attr(*, "names")= chr [1:32] "Mazda RX4" "Mazda RX4 Wag" "Datsun 710" "Hornet 4 Drive" ...
 $ centers     : num [1:2, 1:7] 15.1 23.97 8 4.78 353.1 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:2] "1" "2"
  .. ..$ : chr [1:7] "mpg" "cyl" "disp" "hp" ...
 $ totss       : num 623274
 $ withinss    : num [1:2] 93604 58870
 $ tot.withinss: num 152473
 $ betweenss   : num 470801
 $ size        : int [1:2] 14 18
 $ iter        : int 1
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
```

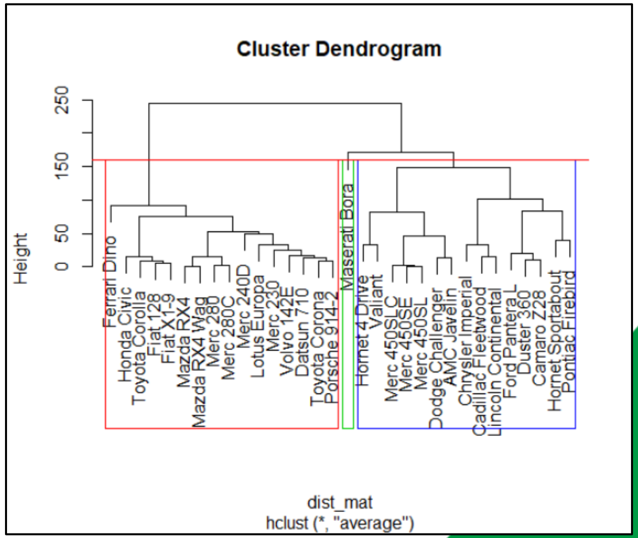
Step-by-step Example 2

Hierarchical clustering

```
dist_mat <- dist(mtcars2, method="euclidean")
hclust1 <-hclust(dist_mat, method='average')
plot(hclust1)

plot(hclust1)
rect.hclust(hclust1, k=2, border=2:6)
abline(h=200, col="red")

plot(hclust1)
rect.hclust(hclust1, k=3, border=2:6)
abline(h=160, col="red")
```



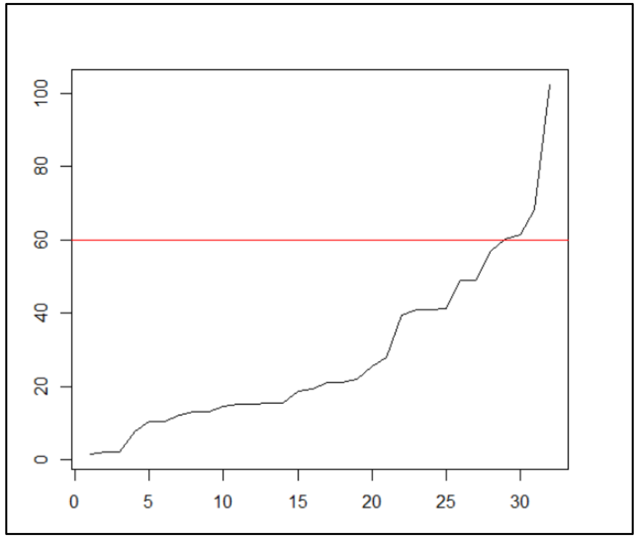


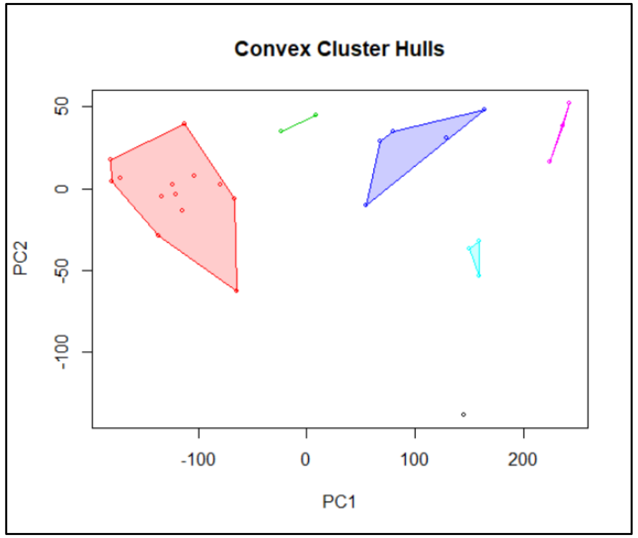


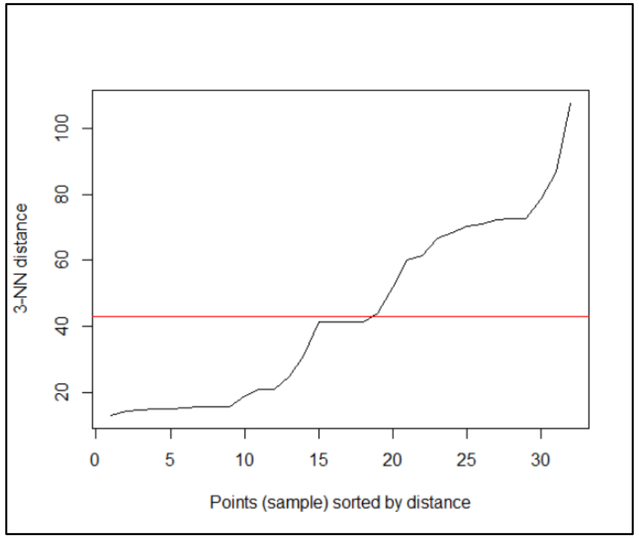
Step-by-step Example 2

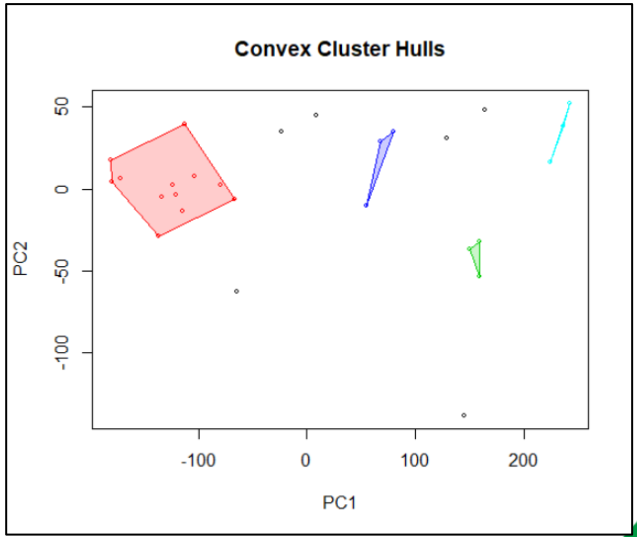
Density-based clustering

```
kNNdistplot(mtcars2, k=2)
abline(h=60, col="red")

dbclust1 <-dbscan(mtcars2, 60, 2)
hullplot(mtcars2, dbclust1$cluster)

kNNdistplot(mtcars2, k=3)
abline(h=43, col="red")

dbclust2 <-dbscan(mtcars2, 43, 3)
hullplot(mtcars2, dbclust2$cluster)
```









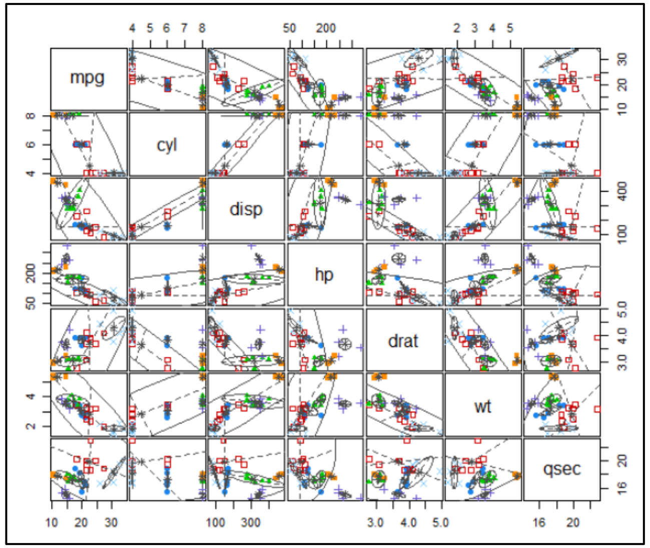
https://en.proft.me/2017/02/3/density-based-clustering-r/

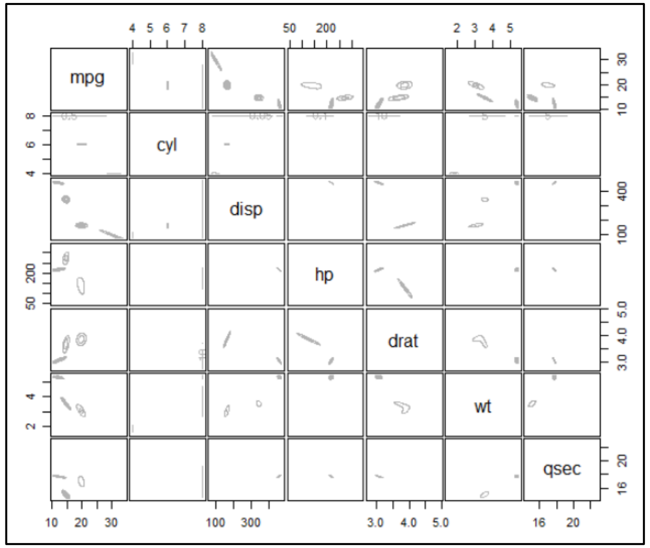
Step-by-step Example 2

Distribution-based clustering

```
mclust1 <-Mclust(mtcars2)
mclust1$modelName
mclust1$G
plot(mclust1, what=c('classification'))
plot(mclust1, "density")

mtcars3 <-mtcars2[,1:3]

mclust2 <-Mclust(mtcars3)
mclust2$modelName
mclust2$G
plot(mclust2, what=c('classification'))
plot(mclust2, "density")

mclust3 <-Mclust(mtcars3, 2)
mclust3$modelName
mclust3$G
plot(mclust3, what=c('classification'))
plot(mclust3, "density")
```

```
[1] "VEV"
[1] 6
```





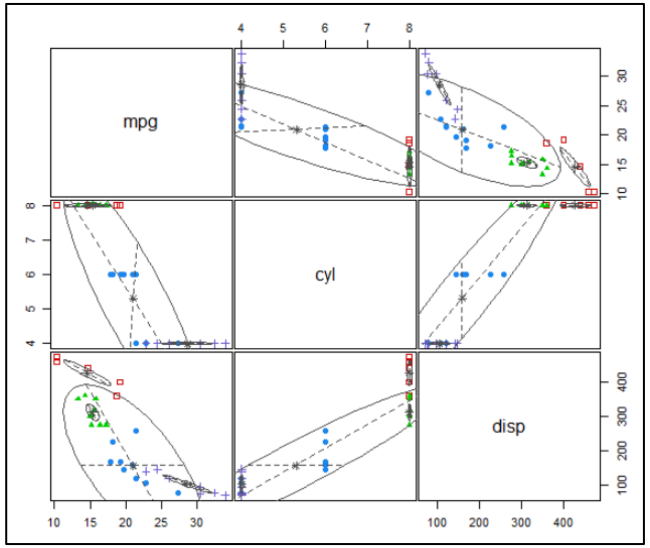
https://en.proft.me/2017/02/1/model-based-clustering-r/

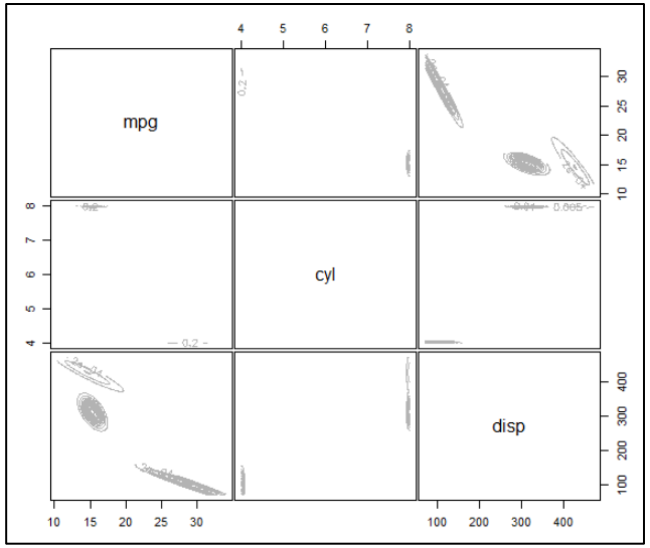
Step-by-step Example 2

Distribution-based clustering

```
mclust1 <-Mclust(mtcars2)
mclust1$modelName
mclust1$G
plot(mclust1, what=c('classification'))
plot(mclust1, "density")

mtcars3 <-mtcars2[,1:3]

mclust2 <-Mclust(mtcars3)
mclust2$modelName
mclust2$G
plot(mclust2, what=c('classification'))
plot(mclust2, "density")

mclust3 <-Mclust(mtcars3, 2)
mclust3$modelName
mclust3$G
plot(mclust3, what=c('classification'))
plot(mclust3, "density")
```

```
[1] "VEV"
[1] 4
```





https://en.proft.me/2017/02/1/model-based-clustering-r/

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

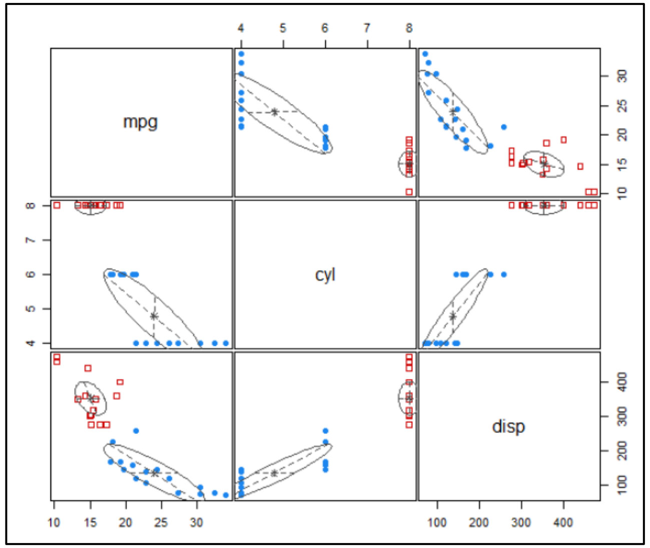
UNIVERSITY OF NORTH DAKOTA

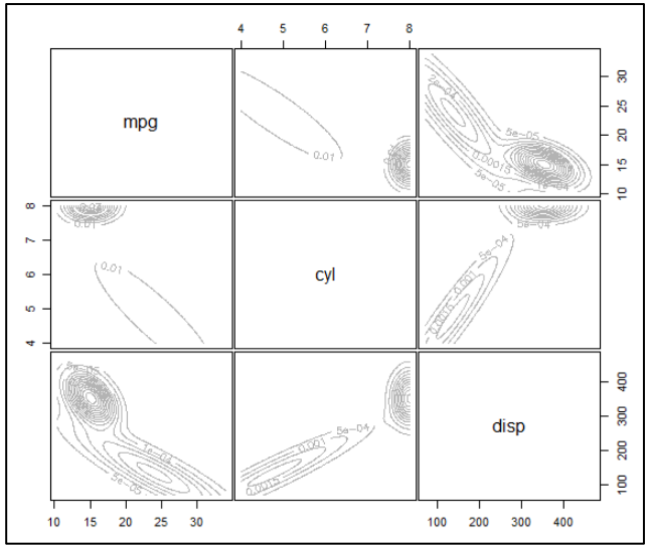
Step-by-step Example 2

Distribution-based clustering

```
mclust1 <-Mclust(mtcars2)
mclust1$modelName
mclust1$G
plot(mclust1, what=c('classification'))
plot(mclust1, "density")

mtcars3 <-mtcars2[,1:3]

mclust2 <-Mclust(mtcars3)
mclust2$modelName
mclust2$G
plot(mclust2, what=c('classification'))
plot(mclust2, "density")

mclust3 <-Mclust(mtcars3, 2)
mclust3$modelName
mclust3$G
plot(mclust3, what=c('classification'))
plot(mclust3, "density")
```

```
[1] "VEV"
[1] 2
```





https://en.proft.me/2017/02/1/model-based-clustering-r/

Step-by-step Example 2

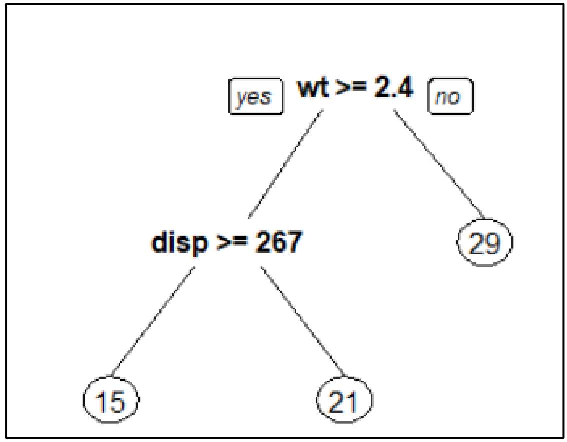
Regression Trees

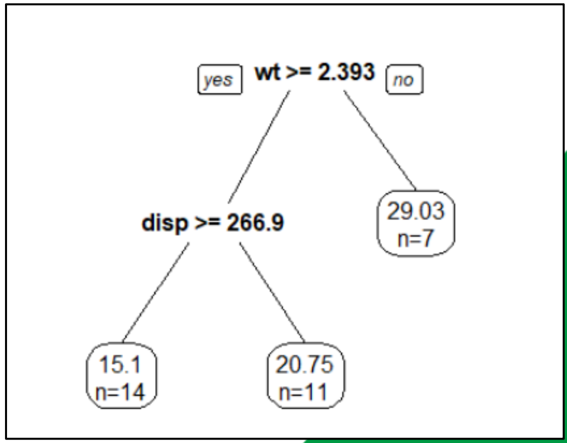
https://www.statology.org/classification-and-regression-trees-in-r/

```
library(rpart)
library(rpart.plot)
head(mtcars)

tree1 <-rpart(mpg ~ disp + hp + drat + wt + qsec, data=mtcars,
control=rpart.control(cp=0.0001))

printcp(tree1)
prp(tree1)

best <-tree1$cptable[which.min(tree1$cptable[,"xerror"]),"CP"]
pruned_tree1 <-prune(tree1, cp=best)
prp(pruned_tree1, faclen=0, extra=1, roundint=F, digits=4)

  #predict
tree2 <-rpart(mpg~disp+wt,data=mtcars, control=rpart.control(cp=0.0001))
prp(tree2)
new_car <- data.frame(wt=3, disp=300)
predict(tree2, newdata=new_car)
```

15.1

```
Regression tree:
rpart(formula = mpg ~ disp + hp + drat + wt + qsec, data = mtcars,
   control = rpart.control(cp = 1e-04))

Variables actually used in tree construction:
[1] disp wt

Root node error: 1126/32 = 35.189

n= 32
      CP nsplit rel error  xerror    xstd
1 0.63566      0   1.00000 1.09142 0.25931
2 0.17491      1   0.36434 0.74268 0.17098
3 0.00010      2   0.18943 0.59461 0.12028
```

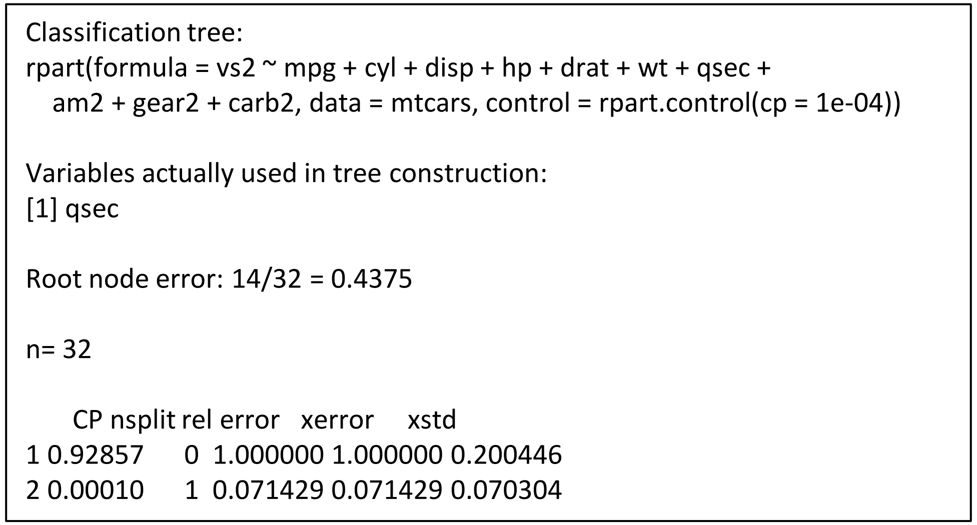




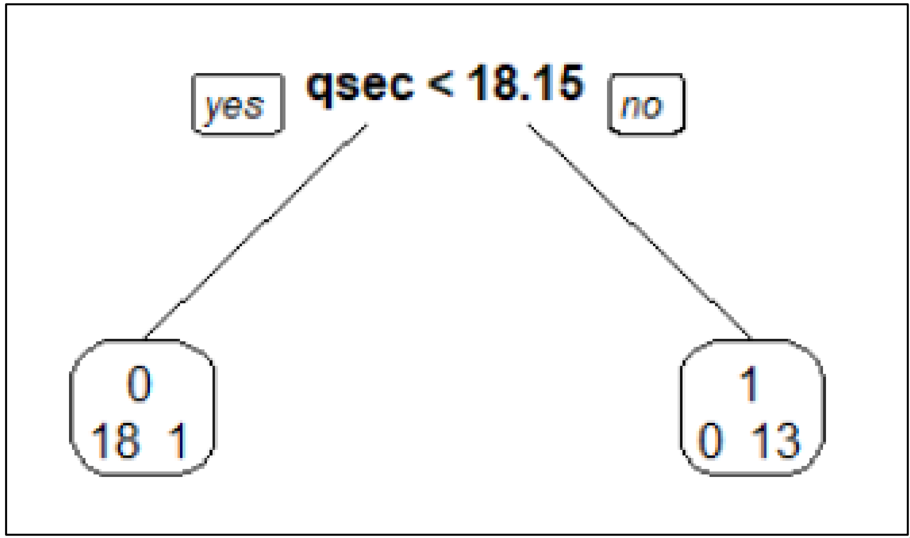
Step-by-step Example 2

Classification Trees

```
mtcars$vs2 <-as.factor(mtcars$vs)
mtcars$am2 <-as.factor(mtcars$am)
mtcars$gear2 <-as.factor(mtcars$gear)
mtcars$carb2 <-as.factor(mtcars$carb)
str(mtcars)

tree3 <-rpart(vs2 ~ mpg + cyl + disp + hp + drat +wt +qsec +am2
+gear2 +carb2, data=mtcars, control=rpart.control(cp=0.0001))

printcp(tree3)
prp(tree3,faclen=0, extra=1, roundint=F, digits=4)

  #less predictive
tree4 <-rpart(vs2 ~ drat +am2 +carb2, data=mtcars,
control=rpart.control(cp=0.0001))

printcp(tree4)
prp(tree4,faclen=0, extra=1, roundint=F, digits=4)
```



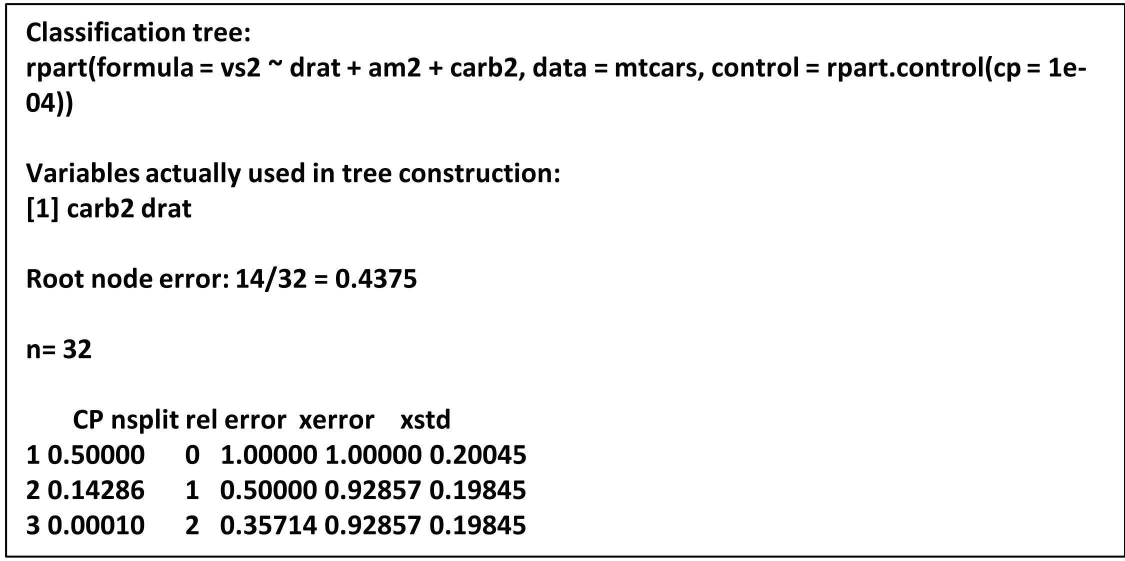
```
Classification tree:
rpart(formula = vs2 ~ mpg + cyl + disp + hp + drat + wt + qsec +
    am2 + gear2 + carb2, data = mtcars, control = rpart.control(cp = 1e-04))

Variables actually used in tree construction:
[1] qsec

Root node error: 14/32 = 0.4375

n= 32

       CP nsplit rel error   xerror     xstd
1 0.92857      0  1.000000 1.000000 0.200446
2 0.00010      1  0.071429 0.071429 0.070304
```



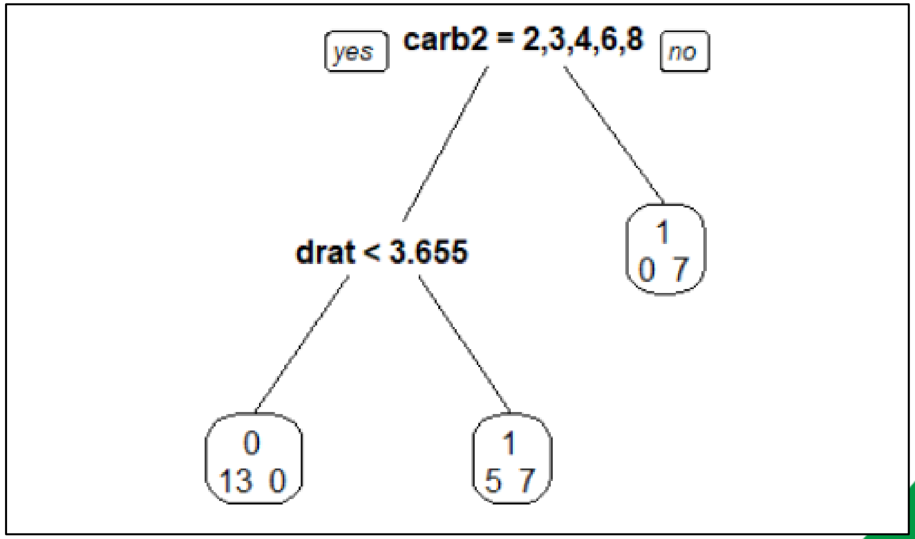
Step-by-step Example 2

Classification Trees

```
mtcars$vs2 <-as.factor(mtcars$vs)
mtcars$am2 <-as.factor(mtcars$am)
mtcars$gear2 <-as.factor(mtcars$gear)
mtcars$carb2 <-as.factor(mtcars$carb)
str(mtcars)

tree3 <-rpart(vs2 ~ mpg + cyl + disp + hp + drat +wt +qsec +am2
+gear2 +carb2, data=mtcars, control=rpart.control(cp=0.0001))

printcp(tree3)
prp(tree3,faclen=0, extra=1, roundint=F, digits=4)

  #less predictive
tree4 <-rpart(vs2 ~ drat +am2 +carb2, data=mtcars,
control=rpart.control(cp=0.0001))

printcp(tree4)
prp(tree4,faclen=0, extra=1, roundint=F, digits=4)
```



```
Classification tree:
rpart(formula = vs2 ~ drat + am2 + carb2, data = mtcars, control = rpart.control(cp = 1e-
04))

Variables actually used in tree construction:
[1] carb2 drat

Root node error: 14/32 = 0.4375

n= 32

       CP nsplit rel error  xerror    xstd
1 0.50000      0   1.00000 1.00000 0.20045
2 0.14286      1   0.50000 0.92857 0.19845
3 0.00010      2   0.35714 0.92857 0.19845
```



Assessment 2

qualtrics XM

UNIVERSITY OF NORTH DAKOTA UND

https://und.qualtrics.com/jfe/form/SV_6ustS4Q8hjLwLqK

Caveats and Concerns

MANOVA and MANCOVA

A. When to use versus ANOVA and ANCOVA
B. Assumptions, assumptions
C. Time getting involved
D. Interpretation
E. Post-hoc tests

Clustering and Trees

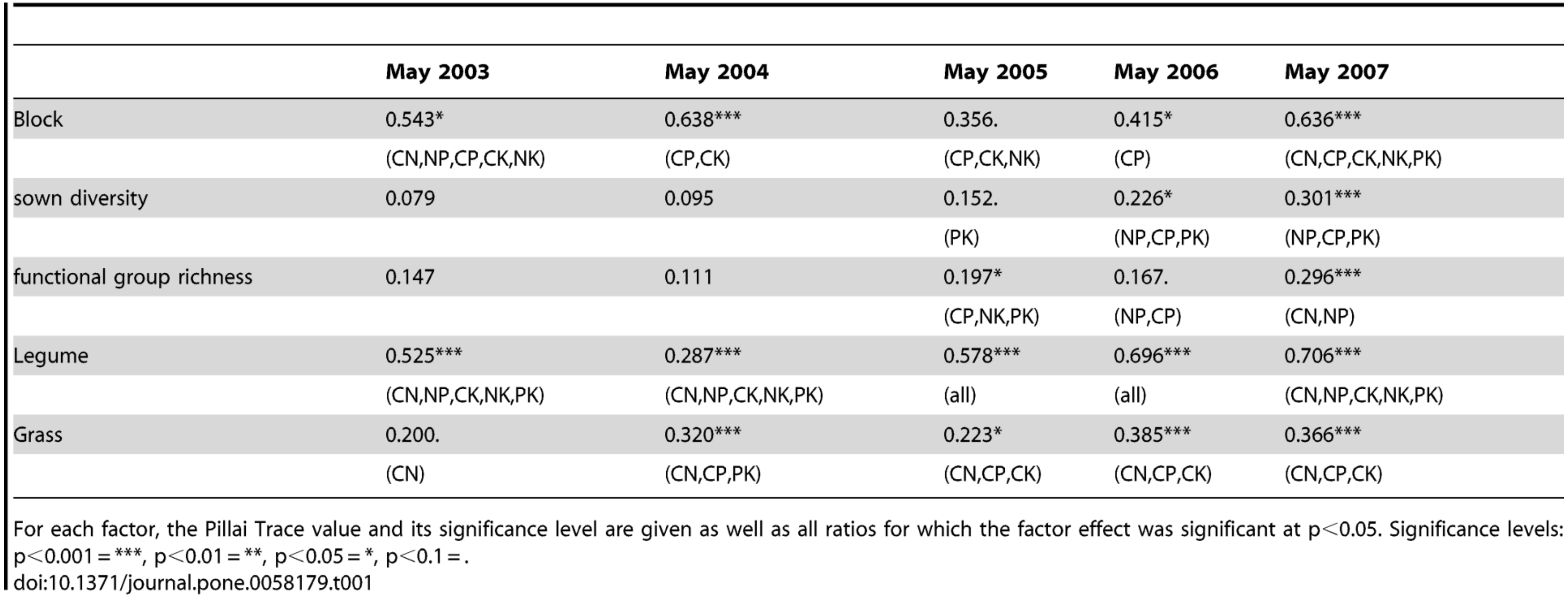
A. When to use certain types
B. Clarity and usefulness

Real World Examples

MANOVA

Abbas, M., Ebeling, A., Oelmann, Y., Ptacnik, R., Roscher, C., Weigelt, A., et al. (2013). Biodiversity Effects on Plant Stoichiometry. PloS One, 8(3), e58179, doi:10.1371/journal.pone.0058179.

Since the six different molar ratios (C:N, C:P, C:K, N:P, N:K, P:K) were not independent of each other, a multivariate analysis of variance (MANOVA) was performed with the following factors: block, sown species richness, functional group richness, legume presence, grass presence.



	May 2003	May 2004	May 2005	May 2006	May 2007
Block	0.543*	0.638***	0.356.	0.415*	0.636***
	(CN,NP,CP,CK,NK)	(CP,CK)	(CP,CK,NK)	(CP)	(CN,CP,CK,NK,PK)
sown diversity	0.079	0.095	0.152.	0.226*	0.301***
			(PK)	(NP,CP,PK)	(NP,CP,PK)
functional group richness	0.147	0.111	0.197*	0.167.	0.296***
			(CP,NK,PK)	(NP,CP)	(CN,NP)
Legume	0.525***	0.287***	0.578***	0.696***	0.706***
	(CN,NP,CK,NK,PK)	(CN,NP,CK,NK,PK)	(all)	(all)	(CN,NP,CK,NK,PK)
Grass	0.200.	0.320***	0.223*	0.385***	0.366***
	(CN)	(CN,CP,PK)	(CN,CP,CK)	(CN,CP,CK)	(CN,CP,CK)

For each factor, the Pillai Trace value and its significance level are given as well as all ratios for which the factor effect was significant at $p<0.05$. Significance levels: $p<0.001$ = ***, $p<0.01$ = **, $p<0.05$ = *, $p<0.1$ = .

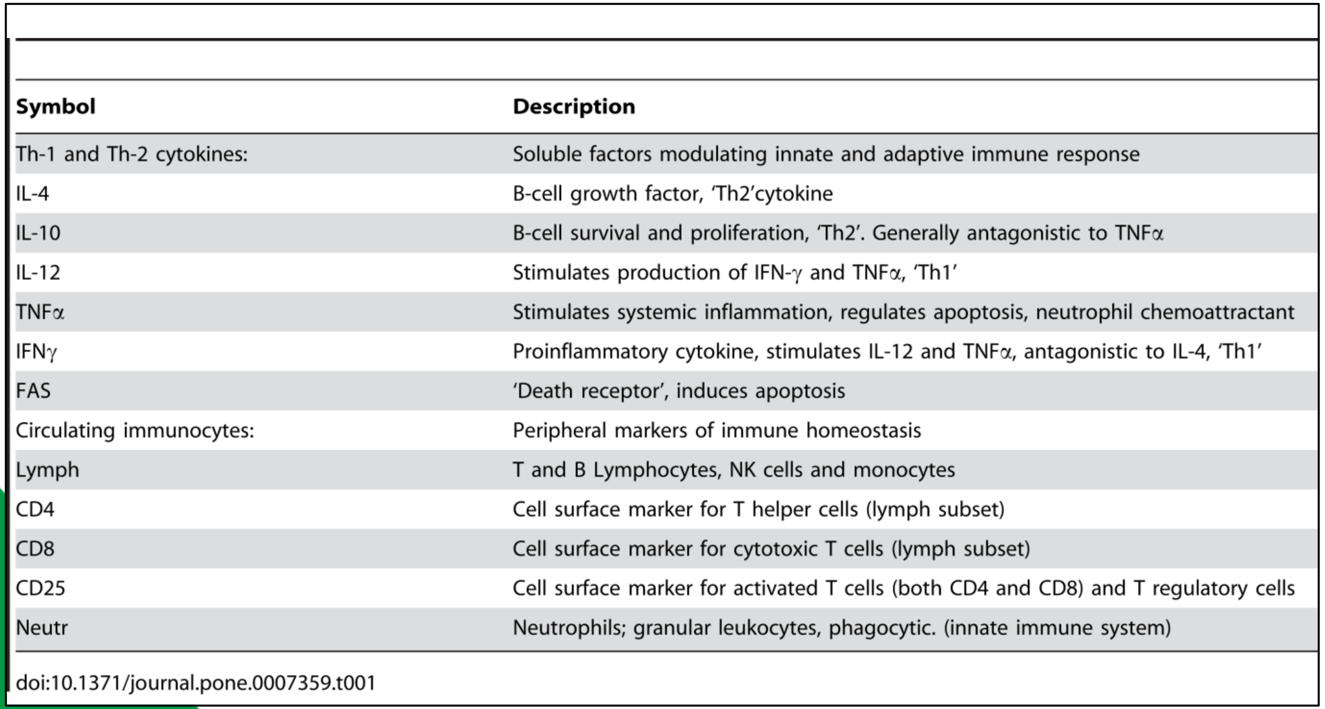
doi:10.1371/journal.pone.0058179.t001

Real World Examples

MANOVA

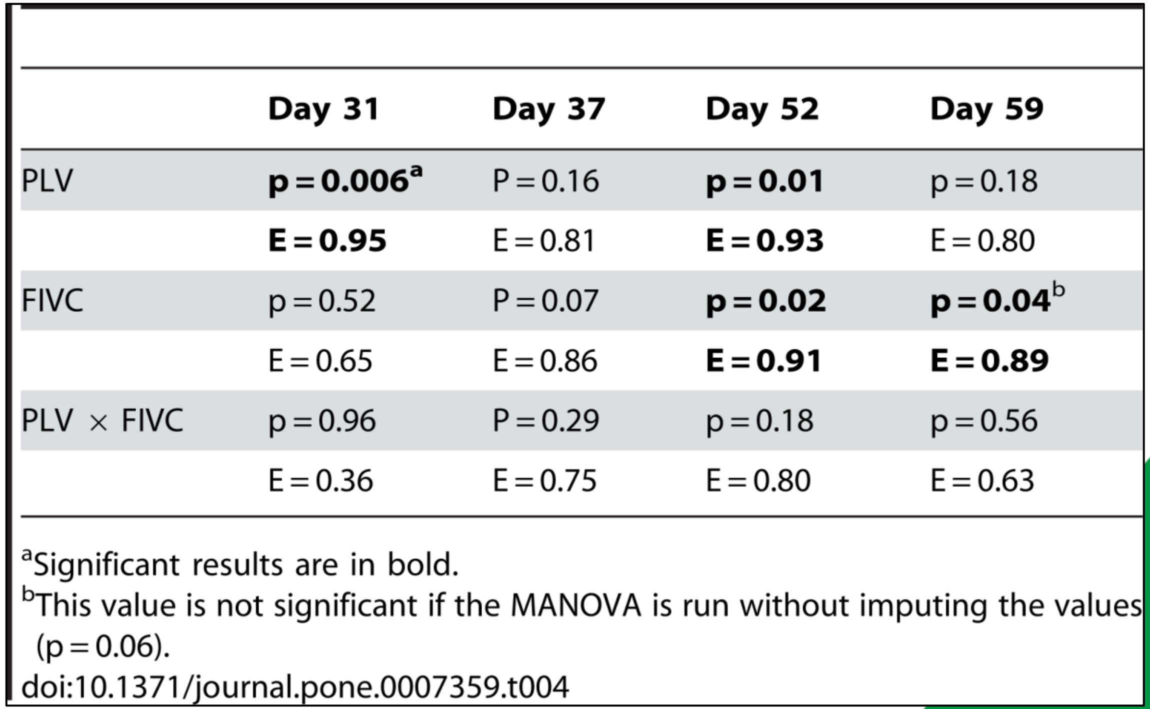
Roy, S., Lavine, J., Chiaromonte, F., Terwee, J., VandeWoude, S., Bjornstad, O., et al. (2009). Multivariate Statistical Analyses Demonstrate Unique Host Immune Responses to Single and Dual Lentiviral Infection. *PloS One, 4*(10), e7359, doi:10.1371/journal.pone.0007359.

To investigate significant effects of single and dual infection on the responses, we used Multivariate Analysis of Variance (MANOVA; [22], [23]) with tests based on Pillai's trace. Since the responses have a marked co-variation structure, these provide enhanced power relative to univariate tests assessing differences in infection group means separately for each response [24].



Symbol	Description
Th-1 and Th-2 cytokines:	Soluble factors modulating innate and adaptive immune response
IL-4	B-cell growth factor, 'Th2'cytokine
IL-10	B-cell survival and proliferation, 'Th2'. Generally antagonistic to TNFα
IL-12	Stimulates production of IFN-γ and TNFα, 'Th1'
TNFα	Stimulates systemic inflammation, regulates apoptosis, neutrophil chemoattractant
IFNγ	Proinflammatory cytokine, stimulates IL-12 and TNFα, antagonistic to IL-4, 'Th1'
FAS	'Death receptor', induces apoptosis
Circulating immunocytes:	Peripheral markers of immune homeostasis
Lymph	T and B Lymphocytes, NK cells and monocytes
CD4	Cell surface marker for T helper cells (lymph subset)
CD8	Cell surface marker for cytotoxic T cells (lymph subset)
CD25	Cell surface marker for activated T cells (both CD4 and CD8) and T regulatory cells
Neutr	Neutrophils; granular leukocytes, phagocytic. (innate immune system)

doi:10.1371/journal.pone.0007359.t001



	Day 31	Day 37	Day 52	Day 59
PLV	**p = 0.006**[a]	P = 0.16	**p = 0.01**	p = 0.18
	E = 0.95	E = 0.81	**E = 0.93**	E = 0.80
FIVC	p = 0.52	P = 0.07	**p = 0.02**	**p = 0.04**[b]
	E = 0.65	E = 0.86	**E = 0.91**	**E = 0.89**
PLV × FIVC	p = 0.96	P = 0.29	p = 0.18	p = 0.56
	E = 0.36	E = 0.75	E = 0.80	E = 0.63

[a] Significant results are in bold.

[b] This value is not significant if the MANOVA is run without imputing the values (p = 0.06).

doi:10.1371/journal.pone.0007359.t004

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY

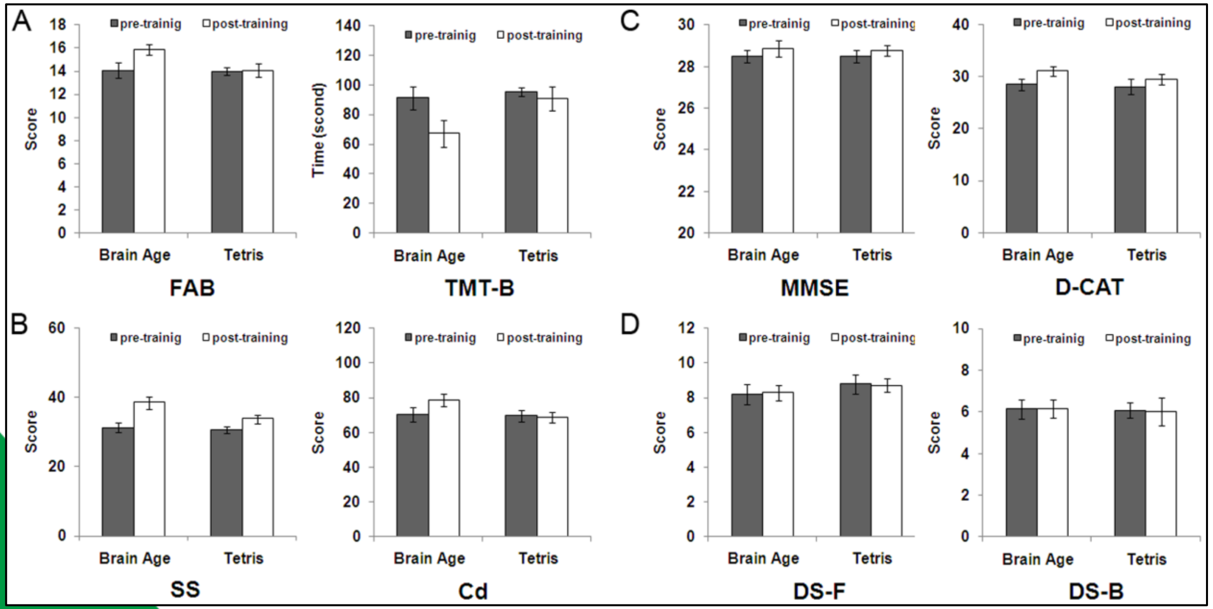
UNIVERSITY OF NORTH DAKOTA

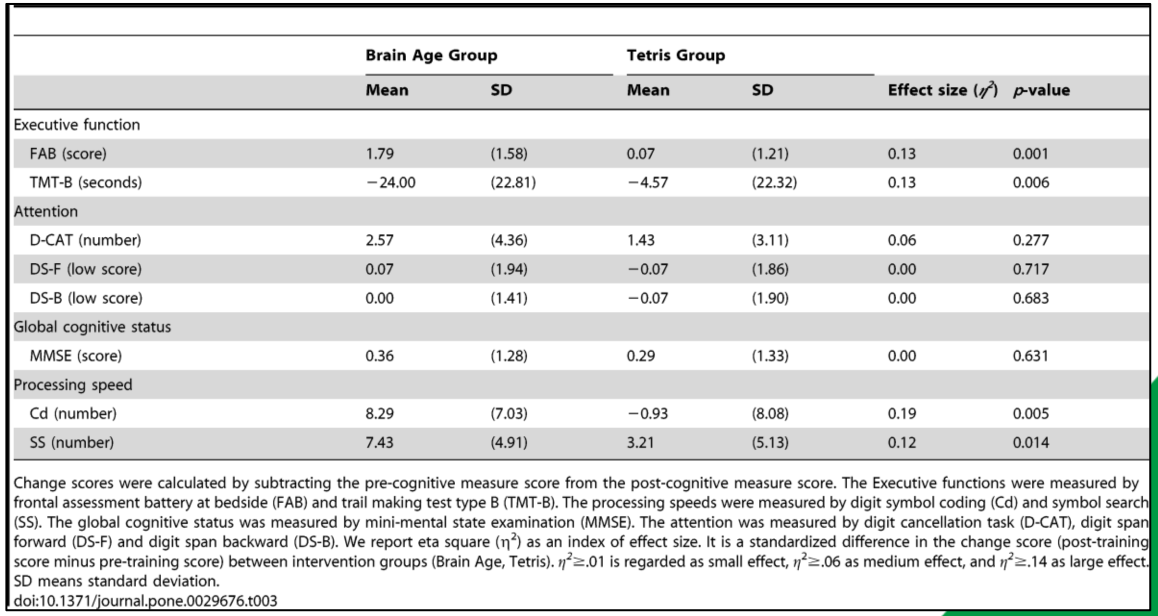
Real World Examples

MANCOVA

Nouchi, R., Taki, Y., Takeuchi, H., Hashizume, H., Akitsuki, Y., Shigemune, Y., et al. (2012). Brain Training Game Improves Executive Functions and Processing Speed in the Elderly: A Randomized Controlled Trial. PloS One, 7(1), e29676, doi:10.1371/journal.pone.0029676.

We conducted multivariate analyses of covariance (MANCOVA) for the change scores (post-training score minus pre-training score) in each of cognitive tests (Figure 2, Table 3). The change scores were the dependent variable, groups (Brain Age, Tetris) was the independent variable. Pre-training scores in all cognitive tests, sex, age, and education levels (years) were the covariate to exclude the possibility that any pre-existing difference of measure between groups affected the result of each measure and adjust for background characteristics.





	Brain Age Group		Tetris Group			
	Mean	SD	Mean	SD	Effect size (η^2)	*p*-value
Executive function						
FAB (score)	1.79	(1.58)	0.07	(1.21)	0.13	0.001
TMT-B (seconds)	−24.00	(22.81)	−4.57	(22.32)	0.13	0.006
Attention						
D-CAT (number)	2.57	(4.36)	1.43	(3.11)	0.06	0.277
DS-F (low score)	0.07	(1.94)	−0.07	(1.86)	0.00	0.717
DS-B (low score)	0.00	(1.41)	−0.07	(1.90)	0.00	0.683
Global cognitive status						
MMSE (score)	0.36	(1.28)	0.29	(1.33)	0.00	0.631
Processing speed						
Cd (number)	8.29	(7.03)	−0.93	(8.08)	0.19	0.005
SS (number)	7.43	(4.91)	3.21	(5.13)	0.12	0.014

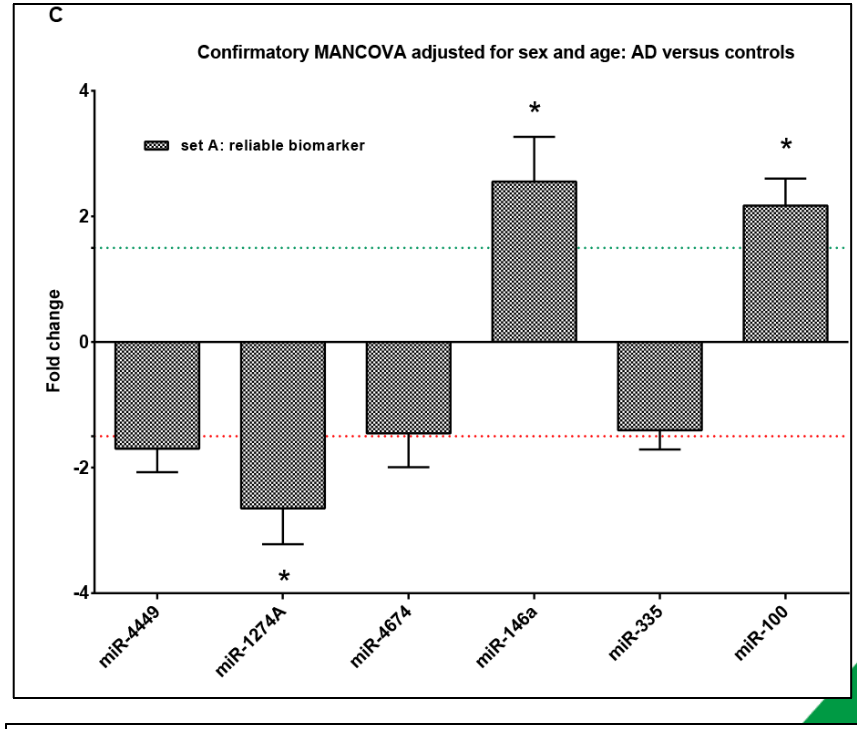
Change scores were calculated by subtracting the pre-cognitive measure score from the post-cognitive measure score. The Executive functions were measured by frontal assessment battery at bedside (FAB) and trail making test type B (TMT-B). The processing speeds were measured by digit symbol coding (Cd) and symbol search (SS). The global cognitive status was measured by mini-mental state examination (MMSE). The attention was measured by digit cancellation task (D-CAT), digit span forward (DS-F) and digit span backward (DS-B). We report eta square (η^2) as an index of effect size. It is a standardized difference in the change score (post-training score minus pre-training score) between intervention groups (Brain Age, Tetris). $\eta^2 \geq .01$ is regarded as small effect, $\eta^2 \geq .06$ as medium effect, and $\eta^2 \geq .14$ as large effect. SD means standard deviation.
doi:10.1371/journal.pone.0029676.t003

Real World Examples

MANCOVA

Denk, J., Boelmans, K., Siegismund, C., Lassner, D., Arlt, S., & Jahn, H. (2015). MicroRNA Profiling of CSF Reveals Potential Biomarkers to Detect Alzheimer`s Disease. *PloS One, 10*(5), e0126423, doi:10.1371/journal.pone.0126423.

After identifying the reliable biomarker candidates of set A and the most informative variables of set B, inferential statistics followed by applying multivariate analyses of covariance (MANCOVA) with sex and age as covariates. Those miRNAs among the biomarker candidates, which revealed significant differences between the AD and control group after Bonferroni adjustments on the confirmatory level, were designated as significant biomarkers.



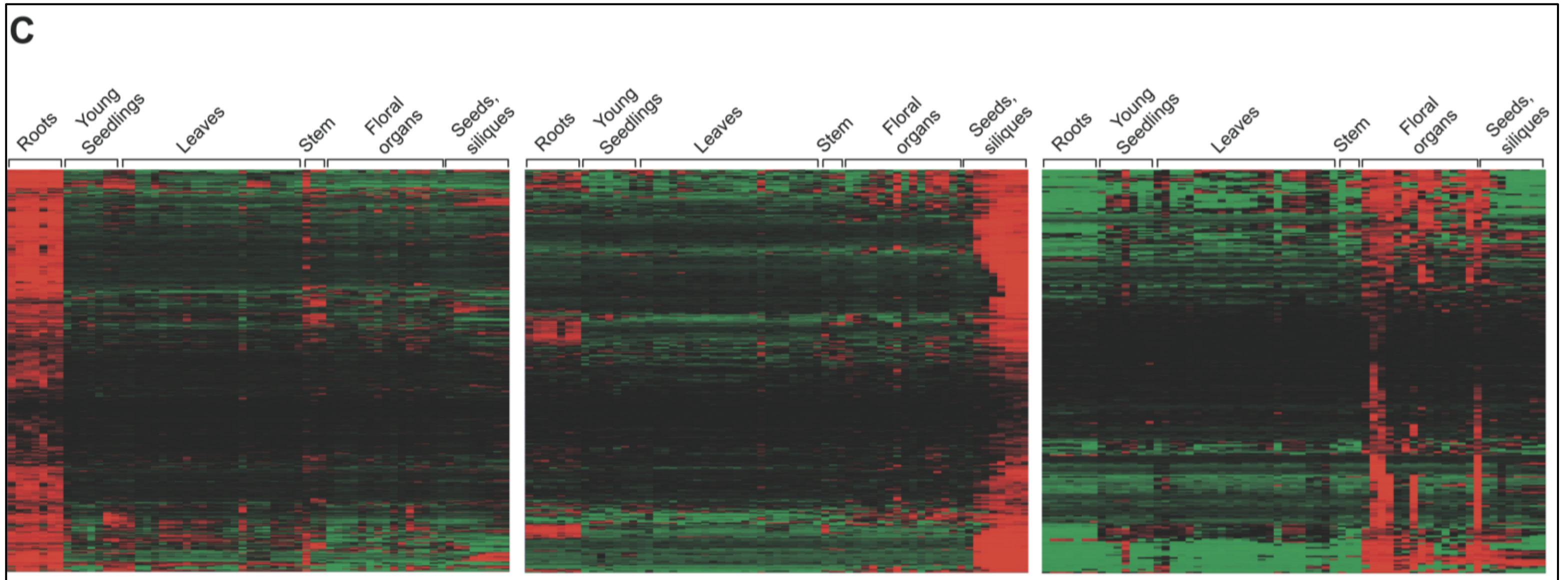
(C) Bar diagram of the reliable biomarker signals of set A. Stars (*) over the bars point to significant p-values (MANCOVA, $p < \alpha^*$, where α^* is Bonferroni corrected $\alpha = 0.05$) and therewith to significant biomarkers.

Real World Examples 🌎

K-means clustering

Zhang, X., Clarenz, O., Cokus, S., Bernatavichute, Y. V., Pellegrini, M., Goodrich, J., et al. (2007). Whole-Genome Analysis of Histone H3 Lysine 27 Trimethylation in Arabidopsis. PLoS Biology, 5(5), e129, doi:10.1371/journal.pbio.0050129.

For cluster analysis, the logarithm of the expression ratio for each gene divided by its mean value across all conditions was computed. This data was then clustered into 8–10 mutually exclusive groups using K-means clustering [50]. The genes within each cluster were then hierarchically clustered and displayed in the figures.



GO Term	% in	% in	p-value	definition
	Genome	Cluster		
GO:0016020	23.5%	38.7%	10^{-30}	membrane
GO:0051244	6.5%	13.2%	10^{-30}	regulation of cellular physiological process
GO:0003700	6.3%	13.2%	$10^{-11.7}$	transcription factor activity
GO:0050791	6.7%	13.7%	$10^{-11.6}$	regulation of physiological process
GO:0019219	5.8%	13.0%	$10^{-11.4}$	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
GO:0006350	6.0%	13.0%	$10^{-11.4}$	transcription

GO Term	% in	% in	p-value	definition
	Genome	Cluster		
GO:0045735	0.3%	3.2%	10^{-30}	nutrient reservoir activity
GO:0003700	6.3%	14.1%	$10^{-8.6}$	transcription factor activity
GO:0030528	6.8%	14.5%	$10^{-8.0}$	transcription regulator activity
GO:0007275	4.2%	10.6%	$10^{-7.9}$	development
GO:0048316	1.2%	4.6%	$10^{-6.5}$	seed development
GO:0048608	1.2%	4.6%	$10^{-6.5}$	reproductive structure development

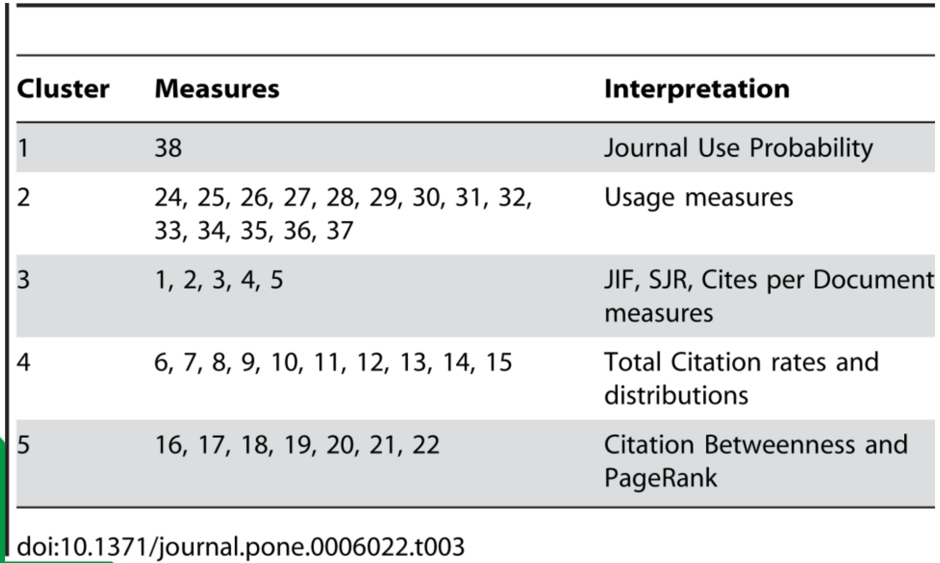
GO Term	% in	% in	p-value	definition
	Genome	Cluster		
GO:0019953	0.2%	5.0%	10^{-30}	sexual reproduction
GO:0009908	0.6%	5.4%	$10^{-9.8}$	flower development
GO:0003700	6.3%	16.8%	$10^{-9.3}$	transcription factor activity
GO:0030528	6.8%	17.6%	$10^{-9.0}$	transcription regulator activity
GO:0019915	0.1%	2.2%	$10^{-8.0}$	sequestering of lipid
GO:0048437	0.2%	3.2%	$10^{-7.6}$	floral organ development

Real World Examples

Hierarchical clustering

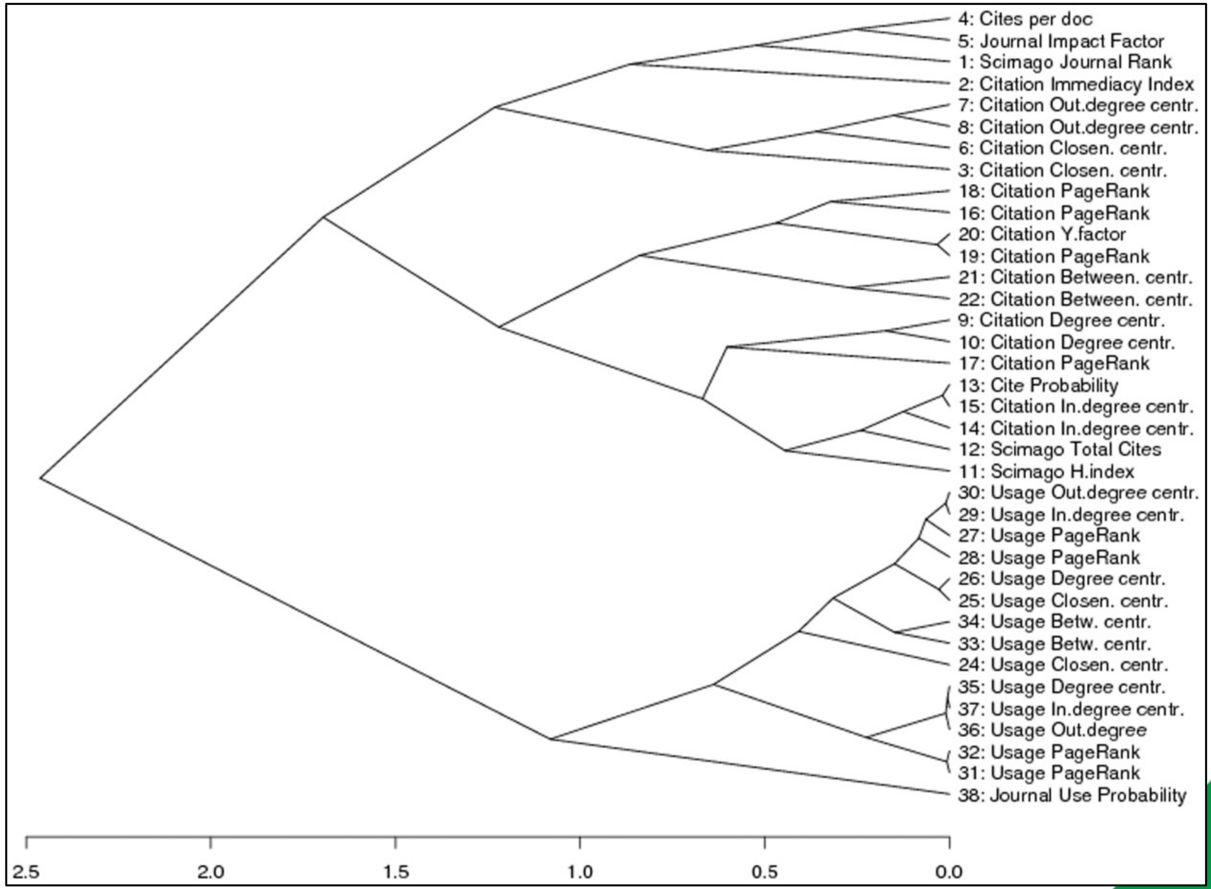
Bollen, J., Van de Sompel, H., Hagberg, A., & Chute, R. (2009). A Principal Component Analysis of 39 Scientific Impact Measures. PloS One, 4(6), e6022, doi:10.1371/journal.pone.0006022.

To cross-validate the PCA results, a hierarchical cluster analysis (single linkage, euclidean distances over row vectors) and a k-means cluster analysis were applied to the measure correlations in to identify clusters of measures that produce similar journal rankings.



Cluster	Measures	Interpretation
1	38	Journal Use Probability
2	24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37	Usage measures
3	1, 2, 3, 4, 5	JIF, SJR, Cites per Document measures
4	6, 7, 8, 9, 10, 11, 12, 13, 14, 15	Total Citation rates and distributions
5	16, 17, 18, 19, 20, 21, 22	Citation Betweenness and PageRank

doi:10.1371/journal.pone.0006022.t003



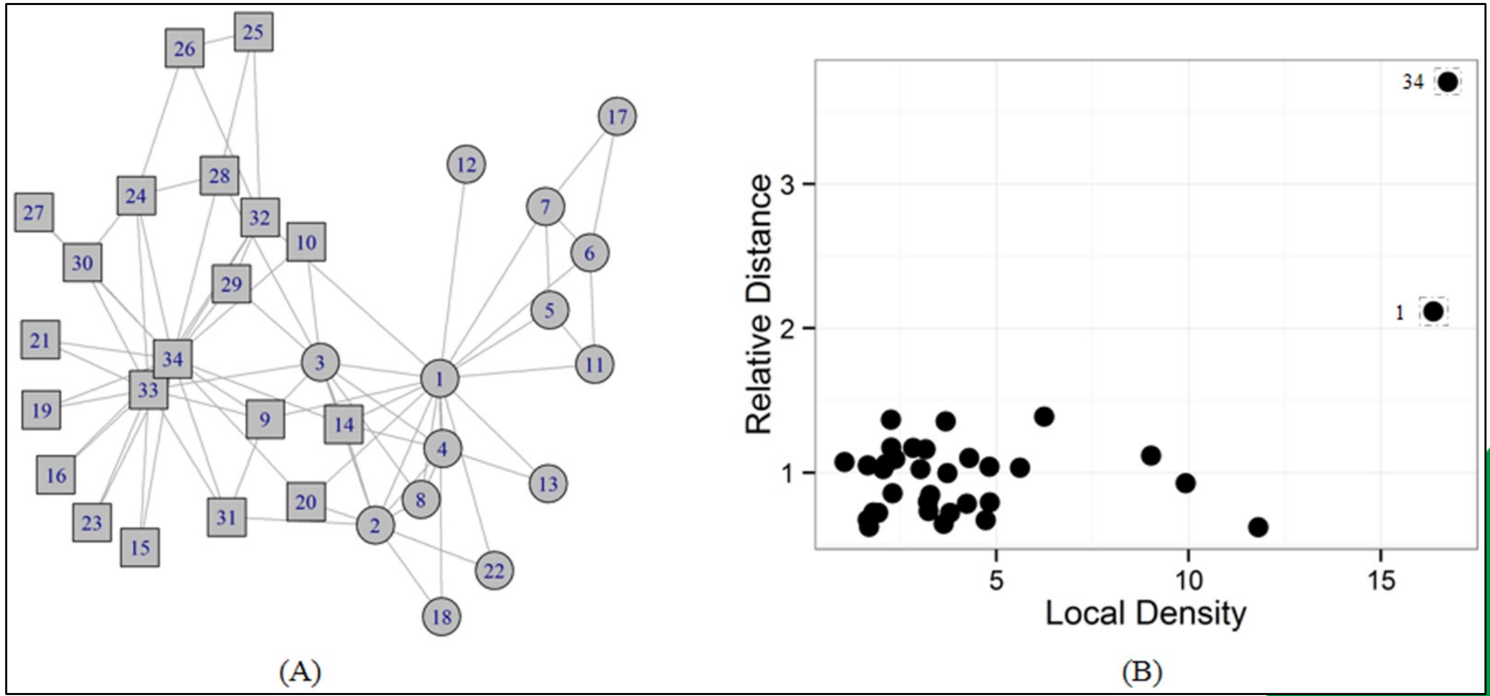
Real World Examples

Density-based clustering

Wang X, Liu G, Li J, Nees JP (2017) Locating Structural Centers: A Density-Based Clustering Method for Community Detection. PLoS ONE 12(1): e0169355. https://doi.org/10.1371/journal.pone.0169355

In this work, we present a new method for community detection which is termed as LCCD. It is a density-based clustering method, inspired by recent research on data analysis [32]where data points are clustered by finding the cluster centers.

This observation is illustrated in Fig 1 by the Zachary's karate club network [42]that is a real-world social network. This interactive network with 34 nodes, ultimately split into two distinct groups, because of a disagreement between the administrator (vertex 1) and the instructor (vertex 34), as shown in Fig 1(A).

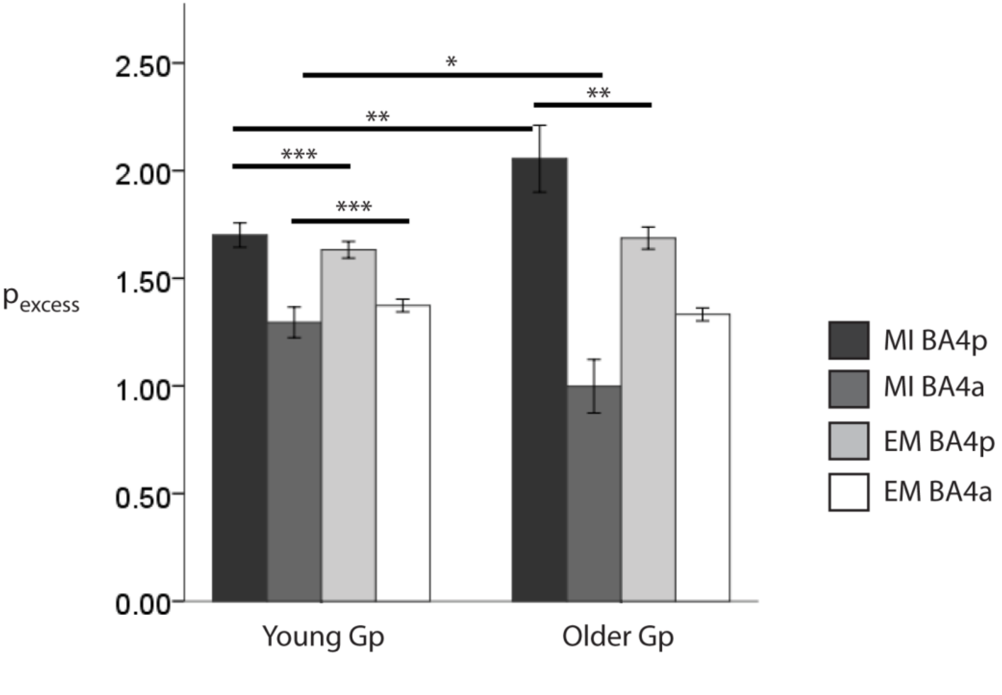


Real World Examples

Distribution-based clustering

Sharma N, Baron J-C (2014) Effects of Healthy Ageing on Activation Pattern within the Primary Motor Cortex during Movement and Motor Imagery: An fMRI Study. PLoS ONE 9(6): e88443. https://doi.org/10.1371/journal.pone.0088443

We explored the distribution-based clustering and weighted laterality index within BA4a and BA4p. The involvement of BA4p during MI (measured with distribution-based clustering) was significantly greater in the older group ($p<0.05$) than in the younger group.



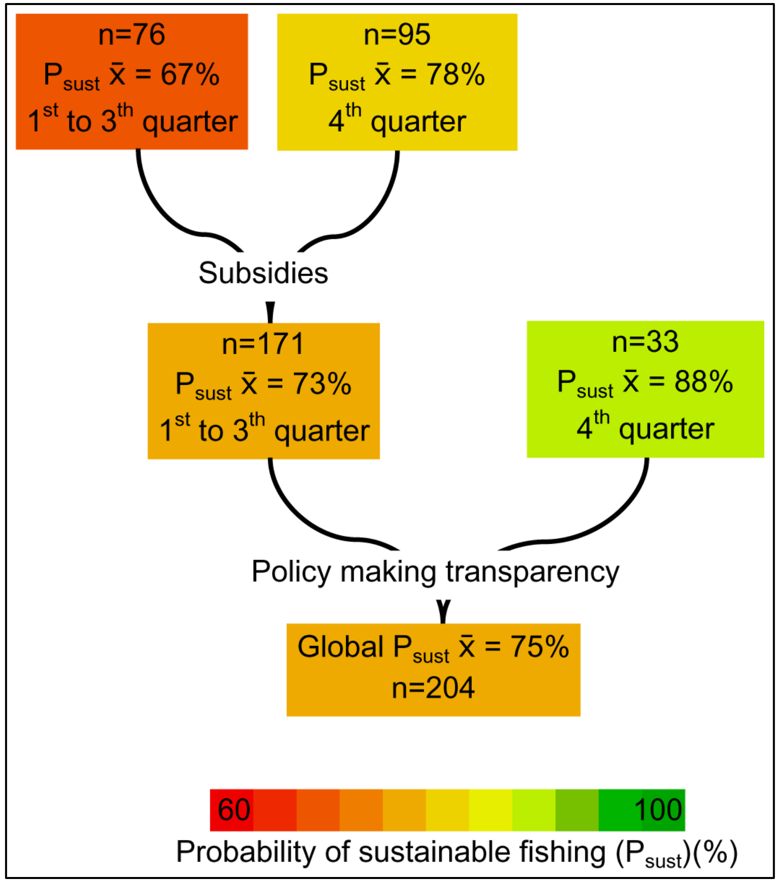
* $p = 0.05$, ** $p<0.05$ *** $p<0.001$. Error Bars SE.
In all tasks p_{excess} of BA4p was greater than BA4a ($p<0.001$) Not labelled.

Real World Examples

Classification Tree

Mora, C., Myers, R. A., Coll, M., Libralato, S., Pitcher, T. J., Sumaila, R. U., et al. (2009). Management Effectiveness of the World's Marine Fisheries. *PLoS Biology, 7*(6), e1000131, doi:10.1371/journal.pbio.1000131.

Data on fisheries sustainability was quantified for the year 2004 and linked to the effectiveness of fisheries management using a classification/regression tree. A classification tree tests for significant differences in fisheries sustainability among the quarters of each attribute (note that the first and fourth quarters are the extremes of a scale from worst- to best-case scenarios for each attribute



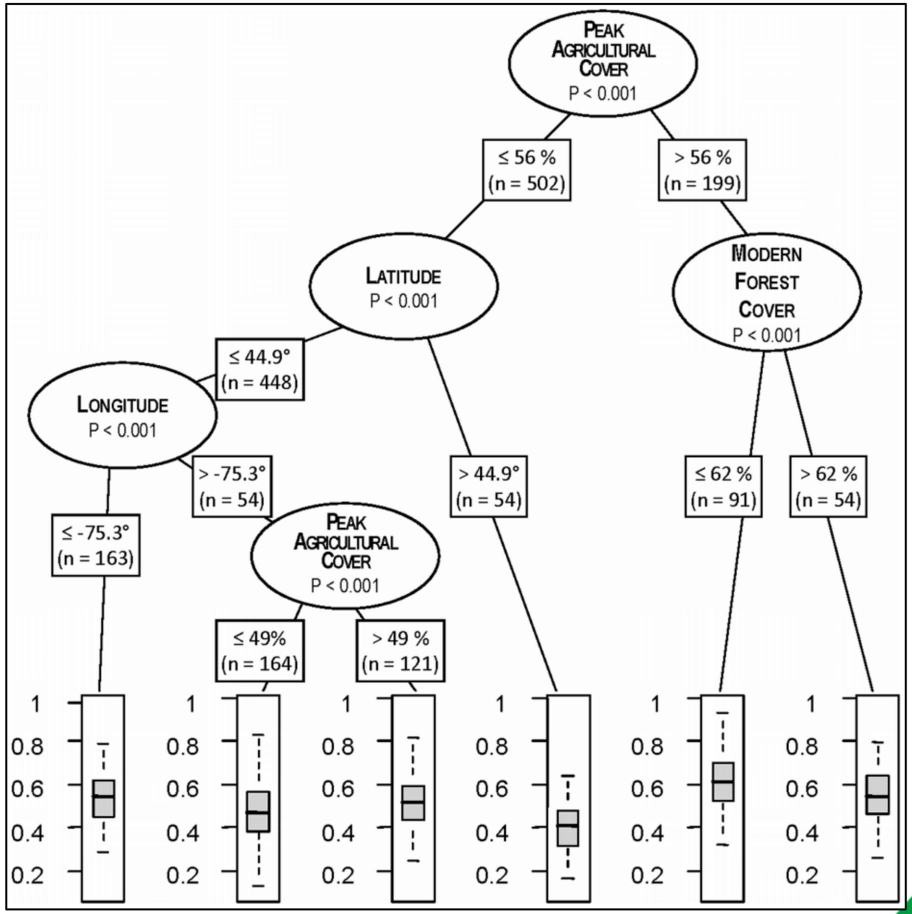
n=76
P_{sust} $\bar{x}$ = 67%
1st to 3th quarter
n=95
P_{sust} $\bar{x}$ = 78%
4th quarter
Subsidies
n=171
P_{sust} $\bar{x}$ = 73%
1st to 3th quarter
n=33
P_{sust} $\bar{x}$ = 88%
4th quarter
Policy making transparency
Global P_{sust} $\bar{x}$ = 75%
n=204
60
100
Probability of sustainable fishing (P_{sust})(%)

Real World Examples

Regression Tree

Thompson, J. R., Carpenter, D. N., Cogbill, C. V., & Foster, D. R. (2013). Four Centuries of Change in Northeastern United States Forests. PloS One, 8(9), e72540, doi:10.1371/journal.pone.0072540.

Finally, to evaluate the relationships between compositional change and the suite of predictor variables identified in Table 2 further, we used regression tree analysis (RTA) with the Sørenson's distance between time periods as the response variable.



Summary and Conclusion

- MANOVA and MANCOVA are extension of ANOVA and ANCOVA
  - Both involve several assumptions that need to be tested
  - The actual analysis is straightforward
  - Results are usually in table form
- Clustering comes in a variety of methods
  - Used to classify observations into responses based on information
  - Often can be graphed
- Tune in next time for a plunge into advanced topics of Multivariate Analysis Module III: Deep Dive

Acknowledgements

- The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729.
- For the labs that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this Module, please acknowledge us for publications. ***"Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)".***

DaCCoTA
DAKOTA CANCER COLLABORATIVE
ON TRANSLATIONAL ACTIVITY