# Cluster Analysis

Today . . .

- Hierarchical clustering applied to some MDS and factor analysis data
- K-Means Clustering

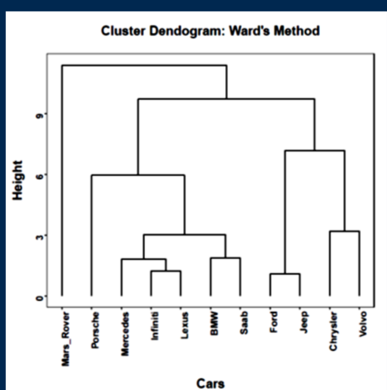
Clustering methods can be applied to the same kind of data that are examined using MDS. A proximity matrix can be used as input and the clusters identified using any of the methods.

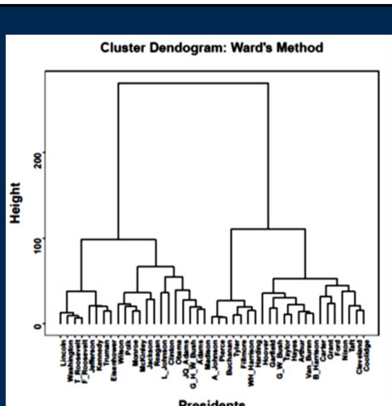
- Car ratings
- President rankings
- Trump ratings

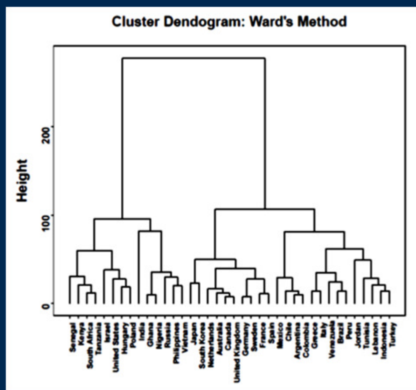
Previous methods focus on the dimensions that underlie similarity among variables or objects. Cluster analysis focuses on discernable separation of objects into groups.

| | Car | Rugged | Fun | Safe | Performance | Family | Versatile | Sports | Status | Practical |
|---|---|---|---|---|---|---|---|---|---|---|
| BMW | BMW | 1.800 | 4.133 | 3.467 | 4.000 | 2.200 | 2.500 | 3.433 | 4.300 | 3.000 |
| Chrysler | Chrysler | 2.065 | 1.613 | 3.355 | 1.677 | 4.774 | 3.323 | 1.355 | 1.710 | 3.903 |
| Ford | Ford | 4.517 | 3.833 | 3.900 | 2.533 | 3.967 | 3.967 | 3.400 | 3.033 | 3.767 |
| Infiniti | Infiniti | 1.517 | 3.367 | 3.767 | 3.500 | 3.367 | 2.667 | 2.600 | 3.567 | 3.138 |
| Jeep | Jeep | 4.300 | 3.533 | 3.400 | 2.967 | 4.000 | 4.133 | 3.700 | 3.267 | 3.900 |
| Lexus | Lexus | 1.677 | 3.633 | 4.033 | 3.800 | 3.400 | 2.667 | 2.300 | 4.100 | 3.167 |
| Mercedes | Mercedes | 2.161 | 3.710 | 4.548 | 4.129 | 3.500 | 3.129 | 2.964 | 4.387 | 3.323 |
| Porsche | Porsche | 2.500 | 4.967 | 2.633 | 4.733 | 1.233 | 1.933 | 4.967 | 4.800 | 1.800 |
| Saab | Saab | 2.400 | 3.933 | 3.700 | 4.000 | 3.167 | 3.067 | 3.433 | 3.567 | 3.100 |
| Volvo | Volvo | 2.333 | 2.133 | 4.733 | 3.133 | 4.767 | 3.500 | 2.200 | 3.033 | 4.333 |
| Mars_Rover | Mars_Rover | 5.000 | 5.000 | 1.000 | 5.000 | 1.000 | 5.000 | 1.000 | 5.000 | 1.000 |

| | Car | Exciting | Dependable | Luxurious | Outdoorsy | Powerful | Stylish | Comfortable |
|---|---|---|---|---|---|---|---|---|
| BMW | BMW | 3.700 | 3.500 | 3.900 | 1.933 | 3.483 | 4.233 | 3.633 |
| Chrysler | Chrysler | 1.452 | 3.290 | 2.097 | 2.290 | 2.258 | 1.484 | 3.419 |
| Ford | Ford | 3.167 | 3.533 | 3.000 | 4.767 | 4.033 | 3.233 | 3.800 |
| Infiniti | Infiniti | 3.067 | 3.767 | 3.933 | 1.533 | 3.533 | 3.500 | 3.967 |
| Jeep | Jeep | 3.567 | 3.167 | 2.900 | 4.667 | 3.900 | 3.333 | 3.600 |
| Lexus | Lexus | 3.323 | 4.290 | 4.387 | 1.710 | 3.613 | 3.935 | 4.226 |
| Mercedes | Mercedes | 3.484 | 4.516 | 4.484 | 1.839 | 3.806 | 4.032 | 4.290 |
| Porsche | Porsche | 4.900 | 3.333 | 3.800 | 3.067 | 4.567 | 4.733 | 2.967 |
| Saab | Saab | 3.733 | 3.033 | 3.500 | 2.333 | 3.867 | 3.533 | 3.467 |
| Volvo | Volvo | 1.800 | 4.267 | 2.900 | 2.500 | 2.967 | 2.367 | 3.633 |
| Mars_Rover | Mars_Rover | 5.000 | 5.000 | 1.000 | 5.000 | 5.000 | 1.000 | 1.000 |






An alternative to the agglomerative, hierarchical approach to clustering more closely resembles the spirit of analysis of variance.

The partitioning procedure known as *K-Means clustering* attempts to form clusters that have the smallest possible within-cluster variances.

The partitioning approach to finding clusters begins with specification of the number of clusters desired (K) and "seed" values for the initial cluster centroids.

Then, cases are assigned to clusters so that the sum of the squared distances from cases to cluster centroids are minimized.

Cases are reassigned until no further reduction in the sum of squared deviations is found.

The K-Means clustering procedure is similar to Ward's method, but is not a hierarchical approach. In Ward's method, when cases are joined in a cluster they cannot later separate and join different clusters. Reassignment is possible in K-Means clustering.

The nature of the final clusters can be heavily dependent on the seed values that are used.

By default, most software chooses an initial set of cases as the seed values, sometimes chosen randomly or to be relatively far apart from each other. Multiple random starts can be specified, with the best solution kept (minimum total within-cluster sum of squares).

To be used in K-Means clustering, variables should be quantitative at the interval or ratio level.

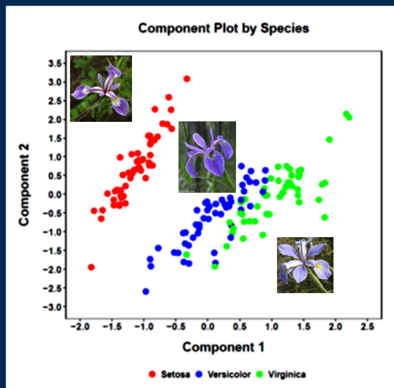
If the variables have different scales, they should be standardized.

Distances are based on simple Euclidean distance.

If the variables are binary or counts, then one of the hierarchical procedures should be used (although counts are sometimes used in K-Means clustering).

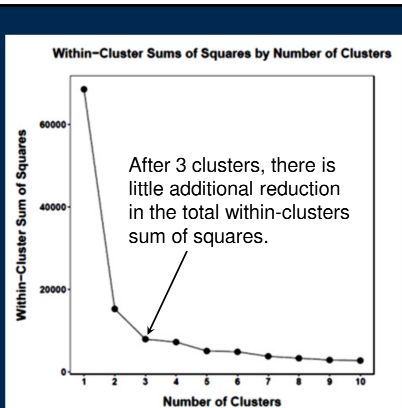
The adequacy of the K-Means approach can be tested in the same way as the hierarchical methods—by examining how well it recovers a known structure.

We'll begin by analyzing the iris data. The kmeans( ) function in the basic stats available when R starts up is a good option for most problems. It requires the raw data matrix, with the objects to be clustered on the rows.




The method attempts to minimize the within-cluster sums of squares. This can be used to help identify the optimal number of clusters. For different numbers of clusters, a point may be found, after which little improvement in the solution occurs.

Similar to a scree plot, the point at which the plot of within-cluster sums of squares reaches a discernible floor can be used as the optimal number of clusters.



```
Iris_K$totss
## [1] 68137
Iris_K$tot.withinss
## [1] 7885
Iris_K$betweenss
## [1] 60252
Iris_K$withinss
## [1] 3982 2388 1515
Iris_K$size
## [1] 62 38 50
```

The K-Means procedure finds the K clusters that minimize the total within-cluster sum of squares (tot.withinss).

The other sums of squares provide information about the relative size of between-cluster and within-cluster variability—similar to ANOVA.

```
Iris_K <- kmeans(Iris[, 1:4], centers = 3, iter.max = 1000, nstart = 10)
Iris_K$cluster

##   [1] 3 2 1 2 1 3 2 1 1 3 1 1 2 1 2 1 2 3 1 2 2 1 2 2 1 3 2 1 1 1
##  [31] 3 1 1 1 2 3 3 1 2 3 2 3 1 3 2 1 3 1 1 2 3 3 1 3 3 3 2 2 3 3
##  [61] 3 1 1 3 3 1 1 3 3 1 1 3 3 2 2 1 1 2 3 3 2 2 1 2 1 1 1 3 3 2
##  [91] 1 3 1 1 1 3 3 1 1 1 3 3 2 1 2 1 3 3 1 1 2 2 3 1 1 3 1 1 1 1
## [121] 1 1 2 2 3 3 2 2 1 1 2 2 2 1 3 3 3 1 3 3 1 1 1 3 3 3 1 2 2 3
```

K-Means

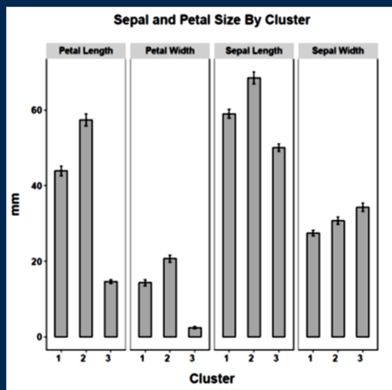
| | 1 | 2 | 3 |
|---|---|---|---|
| Setosa | 0 | 0 | 50 |
| Versicolor | 48 | 2 | 0 |
| Virginica | 14 | 36 | 0 |

Ward's

| | 1 | 2 | 3 |
|---|---|---|---|
| Setosa | 50 | 0 | 0 |
| Versicolor | 0 | 1 | 49 |
| Virginica | 0 | 35 | 15 |

The two major methods provide similar classifications.

```
summary(aov(Iris_New$Sepal_Length ~ as.factor(Iris_New$Cluster)))

##                              Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Iris_New$Cluster)   2   7378    3689     191 <2e-16
## Residuals                   147   2839      19

summary(aov(Iris_New$Sepal_Width ~ as.factor(Iris_New$Cluster)))

##                              Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Iris_New$Cluster)   2   1280     640    60.6 <2e-16
## Residuals                   147   1551      11

summary(aov(Iris_New$Petal_Length ~ as.factor(Iris_New$Cluster)))

##                              Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Iris_New$Cluster)   2  43822   21911    1234 <2e-16
## Residuals                   147   2611      18

summary(aov(Iris_New$Petal_Width ~ as.factor(Iris_New$Cluster)))

##                              Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(Iris_New$Cluster)   2   7773    3886     646 <2e-16
## Residuals                   147    884       6
```




One potential use for cluster analysis is to simplify a sample of data, perhaps when initial analyses suggest a discontinuous nature.

A sample of 150 people were surveyed concerning their opinions about four controversial issues. On a 10-point rating scale, ranging from *Completely Disapprove* (1) to *Completely Approve* (10), the respondents rated their opinions of:

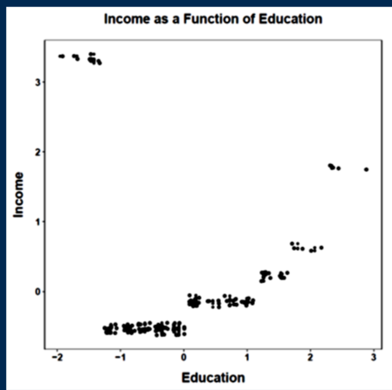
- Gun Control
- Prayer in the Schools
- Death Penalty
- Same Sex Marriage

The sample also reported their annual income and their number of years of education.

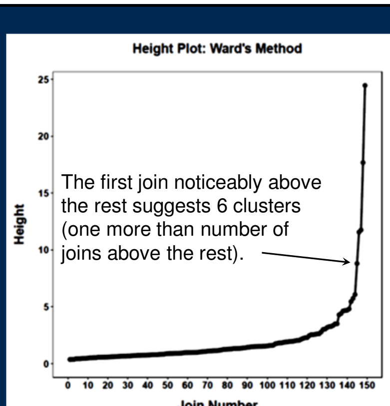
The role of socioeconomic status in shaping opinions on controversial topics was the goal of the study.

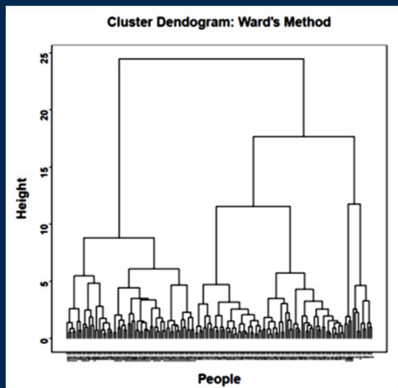
Because of the different scales, all measures are standardized.

An examination of the relationship between education and income revealed an unusual pattern.



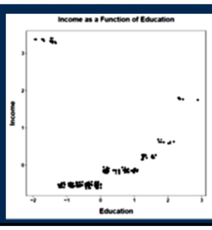

We can analyze these data using either hierarchical or K-Means clustering. Using Ward's hierarchical method, we can plot the height at which cluster joining occurs and identify the point at which clear breaks occur.



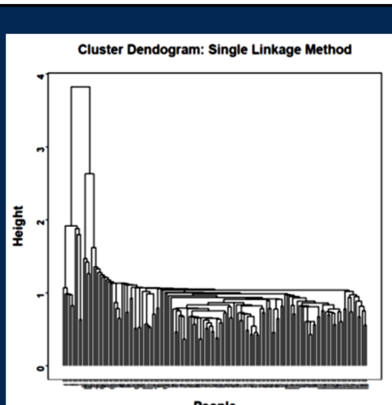


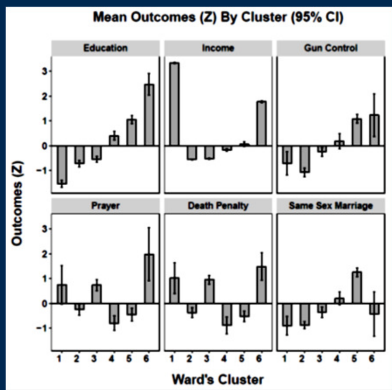

Cluster membership is strongly driven by income and education.

| | Group.1 | educate | income | gun | prayer | death | samesex |
|---|---|---|---|---|---|---|---|
| 1 | 1 | -1.5286 | 3.33375 | -0.7145 | 0.7445 | 1.0163 | -0.8991 |
| 2 | 2 | -0.7215 | -0.55074 | -1.0690 | -0.2478 | -0.3747 | -0.8877 |
| 3 | 3 | -0.5486 | -0.52267 | -0.2513 | 0.7326 | 0.9530 | -0.3638 |
| 4 | 4 | 0.3998 | -0.17050 | 0.1827 | -0.8026 | -0.8877 | 0.1952 |
| 5 | 5 | 1.0489 | 0.06428 | 1.0839 | -0.4503 | -0.5356 | 1.2516 |
| 6 | 6 | 2.4718 | 1.77354 | 1.2292 | 1.9747 | 1.4809 | -0.4295 |

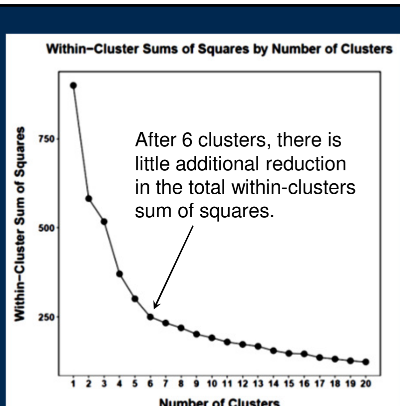



What problem will the single linkage method encounter with these data?






To determine the appropriate number of clusters in the K-Means approach, we can plot the within-cluster sums of squares for different numbers of clusters. Here, too, the plot resembles a scree plot and can be used to identify a number of clusters that produces the greatest reduction in within-cluster error.



```
summary(aov(RW_New_K$educate ~ as.factor(RW_New_K$Cluster_K)))

##                               Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(RW_New_K$Cluster_K)  5  123.3   24.67     138 <2e-16
## Residuals                    144   25.7    0.18

summary(aov(RW_New_K$income ~ as.factor(RW_New_K$Cluster_K)))

##                               Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(RW_New_K$Cluster_K)  5  145.6   29.12    1229 <2e-16
## Residuals                    144    3.4    0.02

summary(aov(RW_New_K$gun ~ as.factor(RW_New_K$Cluster_K)))

##                               Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(RW_New_K$Cluster_K)  5    109   21.81    78.6 <2e-16
## Residuals                    144     40    0.28

summary(aov(RW_New_K$prayer ~ as.factor(RW_New_K$Cluster_K)))

##                               Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(RW_New_K$Cluster_K)  5   80.2   16.03    33.5 <2e-16
## Residuals                    144   68.8    0.48

summary(aov(RW_New_K$death ~ as.factor(RW_New_K$Cluster_K)))

##                               Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(RW_New_K$Cluster_K)  5   87.7   17.53    41.2 <2e-16
## Residuals                    144   61.3    0.43

summary(aov(RW_New_K$samesex ~ as.factor(RW_New_K$Cluster_K)))

##                               Df Sum Sq Mean Sq F value Pr(>F)
## as.factor(RW_New_K$Cluster_K)  5   98.7   19.73    56.5 <2e-16
## Residuals                    144   50.3    0.35
```

Unsurprisingly the clusters are different on education and income. They are different on the other measures too.

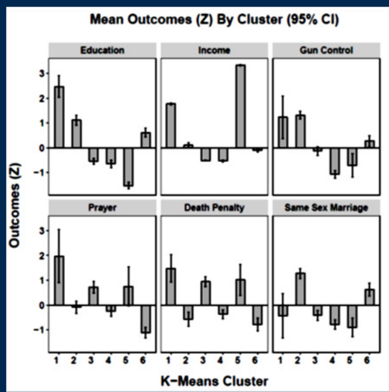
| Group.1 | educate | income | gun | prayer | death | samesex |
|---|---|---|---|---|---|---|
| 1 | 2.4718 | 1.77354 | 1.2292 | 1.97471 | 1.4809 | -0.4295 |
| 2 | 1.1115 | 0.09967 | 1.3200 | -0.09148 | -0.5743 | 1.2798 |
| 3 | -0.5419 | -0.52031 | -0.1332 | 0.72080 | 0.9424 | -0.4107 |
| 4 | -0.6520 | -0.53117 | -1.0759 | -0.25048 | -0.3681 | -0.7874 |
| 5 | -1.5286 | 3.33375 | -0.7145 | 0.74446 | 1.0163 | -0.8991 |
| 6 | 0.6156 | -0.10194 | 0.2677 | -1.10908 | -0.7908 | 0.6165 |

The highlighted clusters have similar education and income, but different attitudes about the issues.

Cluster 5 is the undereducated über-rich group; they are fairly right-wing in their attitudes.

The hierarchical and K-Means methods need not arrive at the same cluster identifications, although if there is a strong cluster structure to the data, they each ought to arrive at this underlying truth.

| | | K-Means Clusters | | | | | |
|---|---|---|---|---|---|---|---|
| | | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 | Cluster 6 |
| Ward's Clusters | Cluster 1 | 0 | 0 | 0 | 0 | 10 | 0 |
| | Cluster 2 | 0 | 0 | 1 | 32 | 0 | 0 |
| | Cluster 3 | 0 | 0 | 37 | 3 | 0 | 0 |
| | Cluster 4 | 0 | 2 | 2 | 2 | 0 | 17 |
| | Cluster 5 | 0 | 28 | 0 | 0 | 0 | 12 |
| | Cluster 6 | 4 | 0 | 0 | 0 | 0 | 0 |




Next time . . .

Additional issues in cluster analysis