Session 11 12, C ust
and Segmentation (
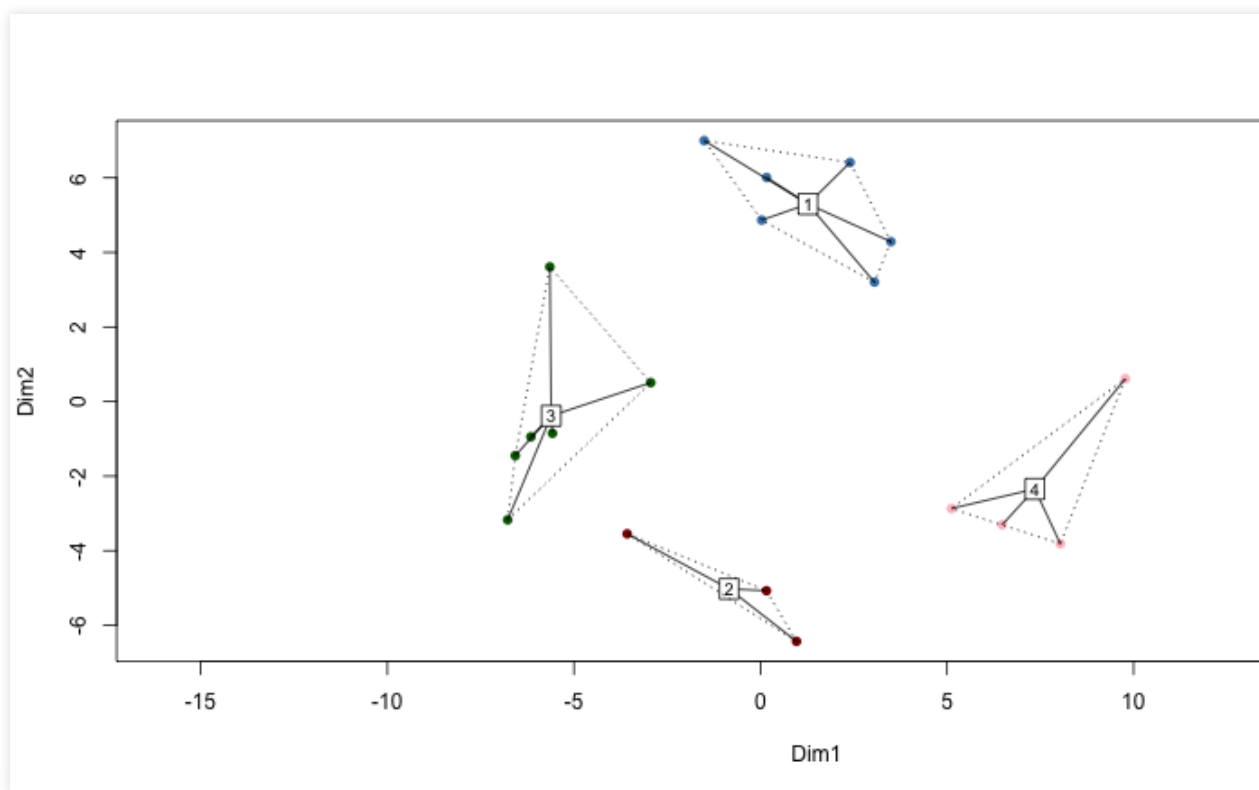nica S ides)

Clustering and Segmentat

What is Clustering and Segmer

Example Usage

-

-

-

-

-

A Segmentation Process

1. Confirm data is metric
2. Scale the data (Optional)
3. Select Segmentation Variables
4. Define similarity measure
5. Visualize Pair-wise Distances
6. Method and Number of Segments
7. Profile and interpret the segments
8. Robustness Analysis

Example Data: Market Research

Step 1: Confirm data is me

	Variables	V1	V2	V3	V4	V5
1	1	6	4	7	3	2
2	2	2	3	1	4	5
3	3	7	2	6	4	1
4	4	4	6	4	5	3
5	5	1	3	2	2	6
6	6	6	4	6	3	3
7	7	5	3	6	3	3
8	8	7	3	7	4	1

Step 2: Scale the data (Optio

	Variables	min	X25.percent	median	mean	X75.percent
1	V1	1	2	4	3.85	5.25
2	V2	2	3	4	4.1	5
3	V3	1	2	4	3.95	6
4	V4	2	3	4	4.1	5.25
5	V5	1	2	3.5	3.45	4.25
6	V6	2	3	4	4.35	5.25

Data Standardization: Example

```
ProjectData_segment_scaled=apply(Proj
 function(r) {
  if (sd(r)!=0) {
    res=(r-mean(r))/sd(r)
    } else {
    res=0*r; res
    }
 })
```

Standardized Data: Summary St

	Variables	min	X25.percent	median	mean	X75.percent
1	V1	-1.52	-0.99	0.08	0	0.75
2	V2	-1.51	-0.79	-0.07	0	0.65
3	V3	-1.49	-0.98	0.03	0	1.03
4	V4	-1.4	-0.73	-0.07	0	0.77
5	V5	-1.41	-0.83	0.03	0	0.46
6	V6	-1.59	-0.91	-0.24	0	0.61

Step 3. Select Segmentation Va

Step 4. Define similarity mea

Distances across our data using clidean distance

	1	2	3	4	5
	0.0				
	8.0	0.0			
	2.8	8.2	0.0		
	5.6	5.6	6.6	0.0	
	8.3	2.6	9.1	6.6	0.0

Distances across our data using hattan distance

1	2	3	4	5
0.0				
16.0	0.0			
6.0	16.0	0.0		
13.0	13.0	15.0	0.0	
17.0	5.0	19.0	16.0	0.0

Manually Defined Distances: an

```r
My_Distance_function<-function(x,y)
{sum(abs(x-y)>2)}
```

Manually Defined Distances: an

1	2	3	4	5
0.0				
3.0	0.0			
0.0	3.0	0.0		
2.0	2.0	3.0	0.0	
3.0	0.0	3.0	4.0	0.0

Step 5. Visualize Pair-wise Dis

Histogram of all pairwise dist

Histogram of all pairwise Distances between observtions
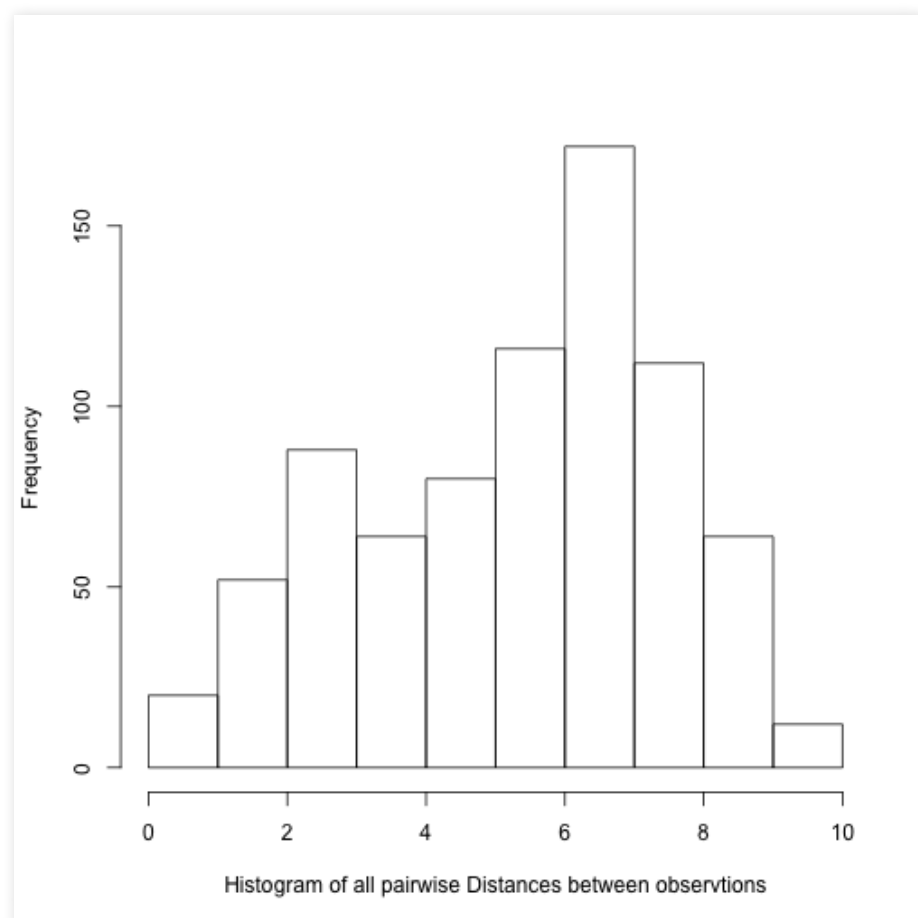
Step 6. Method and Number of S

-

-

Hierarchical Clustering: Dendr
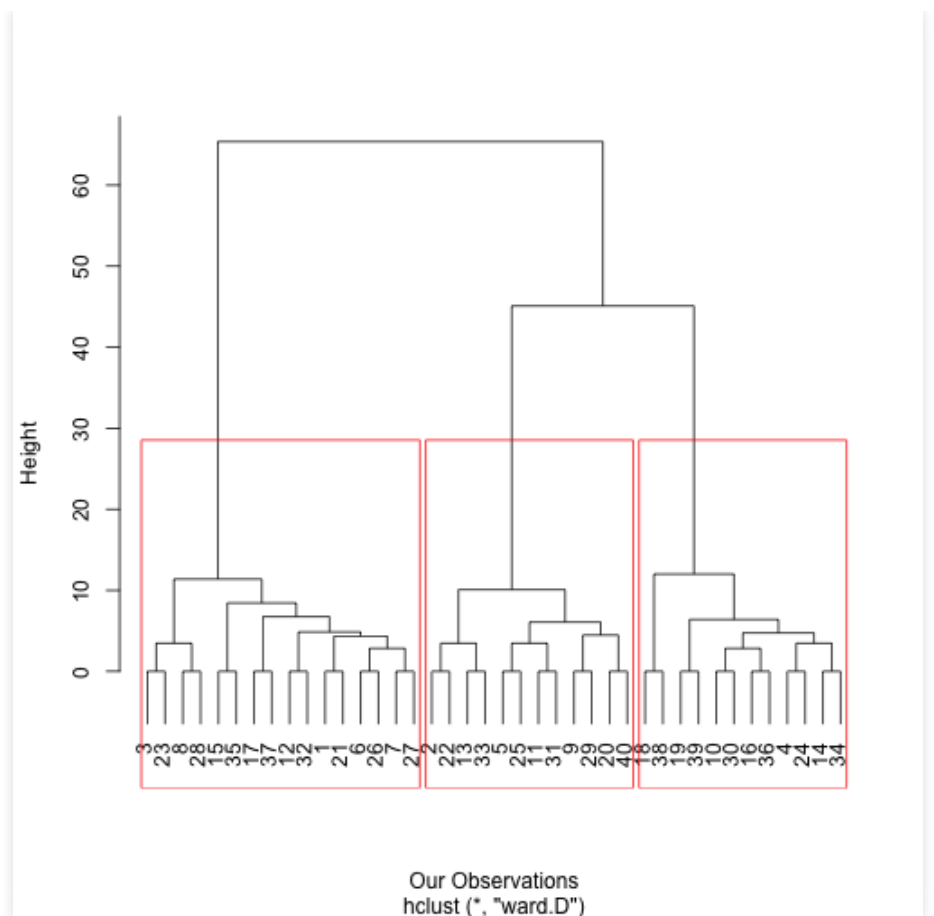
Hierarchical Clustering Dendr
Heights Plot

Distances plot
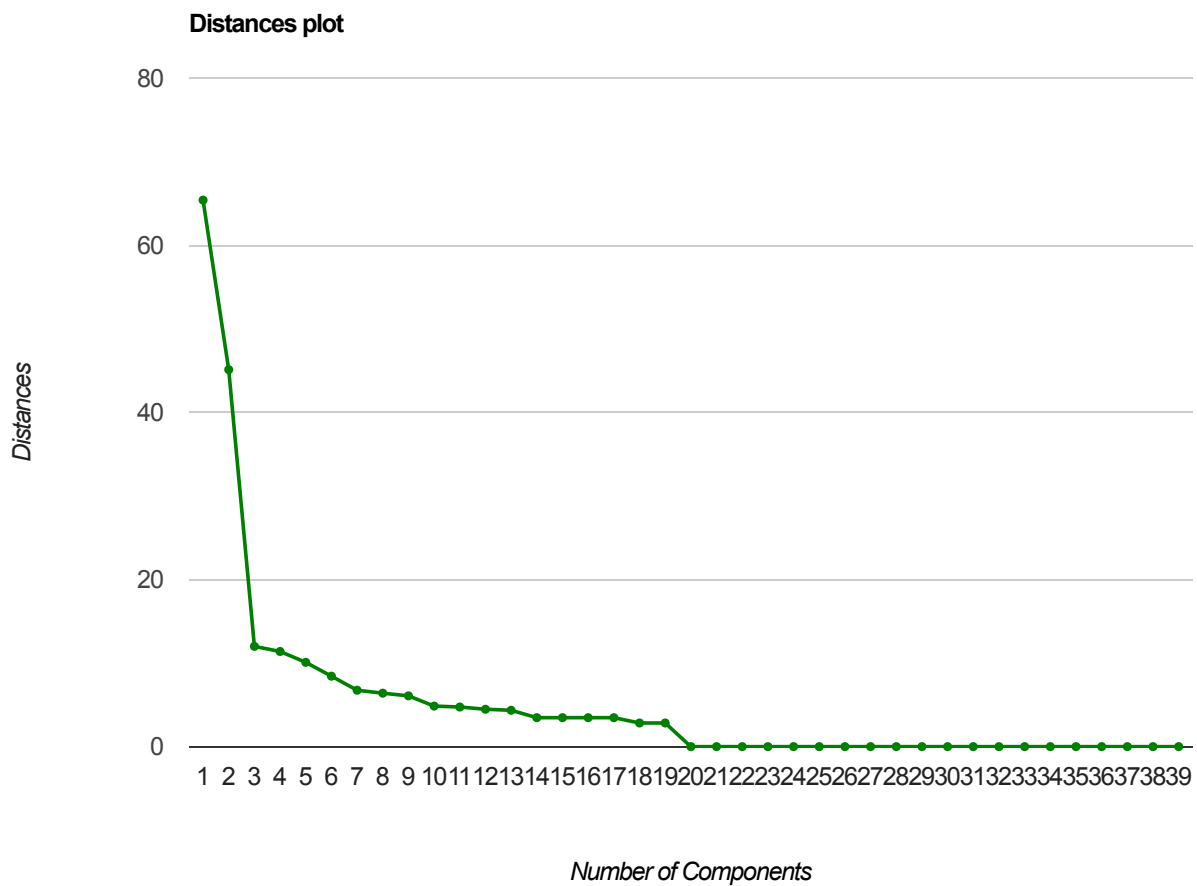
Cluster Membership: Hierarchical Clustering

	Observation	Observation.Number	Cluster_Membership
1	1	1	1
2	2	2	2
3	3	3	1
4	4	4	3
5	5	5	2
6	6	6	1
7	7	7	1
8	8	8	1

Cluster Membership: K-means C

|	Observation	Observation.Number	Cluster_Membership
1	1	1	3
2	2	2	2
3	3	3	3
4	4	4	1
5	5	5	2
6	6	6	3
7	7	7	3
8	8	8	3

Step 7. Profile and interpret the s

Cluster Profiling: Cluster Cent[...]
Profling Variables

	Variables	Population	Segment.1	Segment.2
1	V1	3.85	5.75	1.67
2	V2	4.1	3.62	3
3	V3	3.95	6	1.83
4	V4	4.1	3.12	3.5
5	V5	3.45	1.88	5.5
6	V6	4.35	3.88	3.33
7	Income	46,000	60,000	42,500
8	Mall.Visits	3.25	3.25	1

Interpretation: Snake Plot

Snake plot for each cluster
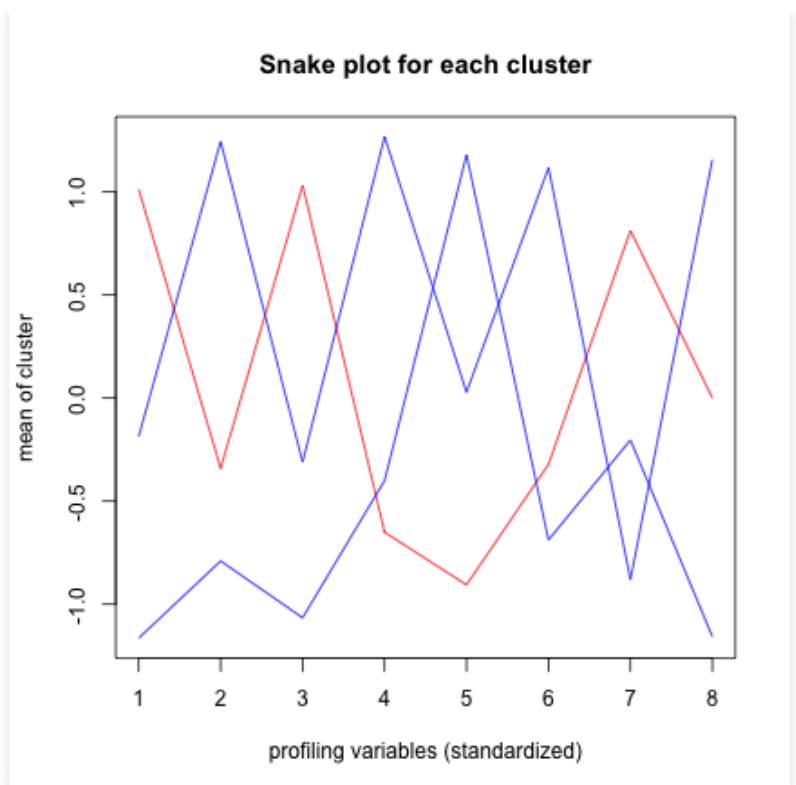
Interpretation: Ratio to Average Population -1 (0 = Averag

Segment.1	Segment.2	Segment.3
0.49	-0.57	-0.09
-0.12	-0.27	0.42
0.52	-0.54	-0.16
-0.24	-0.15	0.46

The data

Step 8. Robustness Analy

- using different subsets of the original data
- using variations of the original segmentatio
- using different distance metrics
- using different segmentation methods
- using different numbers of clusters

Example Robustness Test: Di[f]
Methods

100 %.

Segment 1	Segment 2	Segment 3
100.0	100.0	100.0

Key Technical Terms and Les

- Segmentation Variables
- Profiling Variables
- Distance Metrics
- Hierarchical Clustering
- Dendrogram
- K-means Clustering
- Robustness: Statistics, Interpretation, Dec
- Actionability, Interpretability, Statistical R

Group Work

1.

2.

3.