सिद्धिमूलं प्रबन्धनम्
भा. प्र. सं. इन्दौर
IIM INDORE

Hierarchical Clustering

An Application on Wholesale Market Data

Aarjav Sethi
Apurv Chaudhary
Harshvardhan
Pradeep Charan
Tamanna Gupta

IPM 2016-21 Batch

Table of contents

Introduction

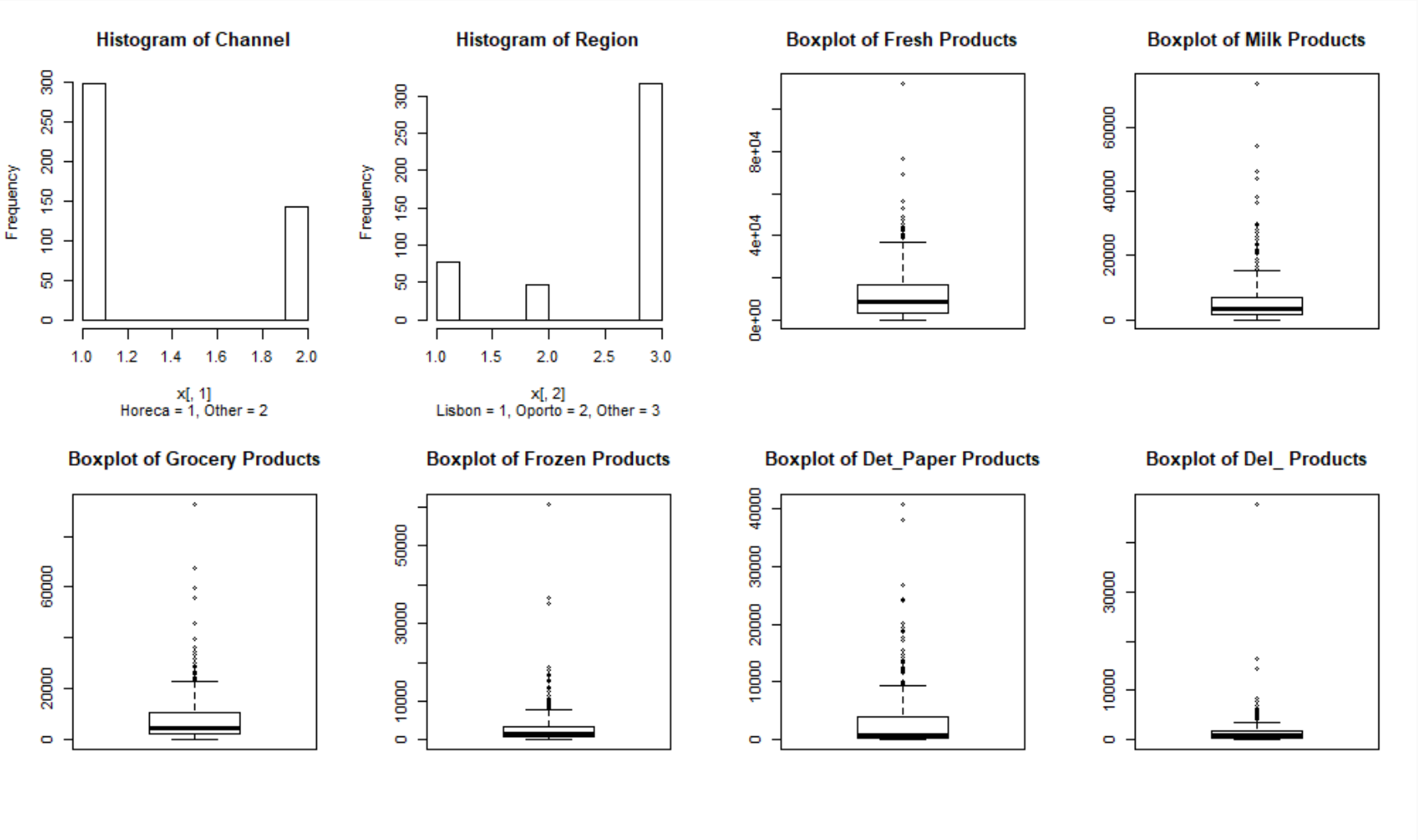
Description of Data

Data Source: University of California, Irvine's Machine Learning Repository

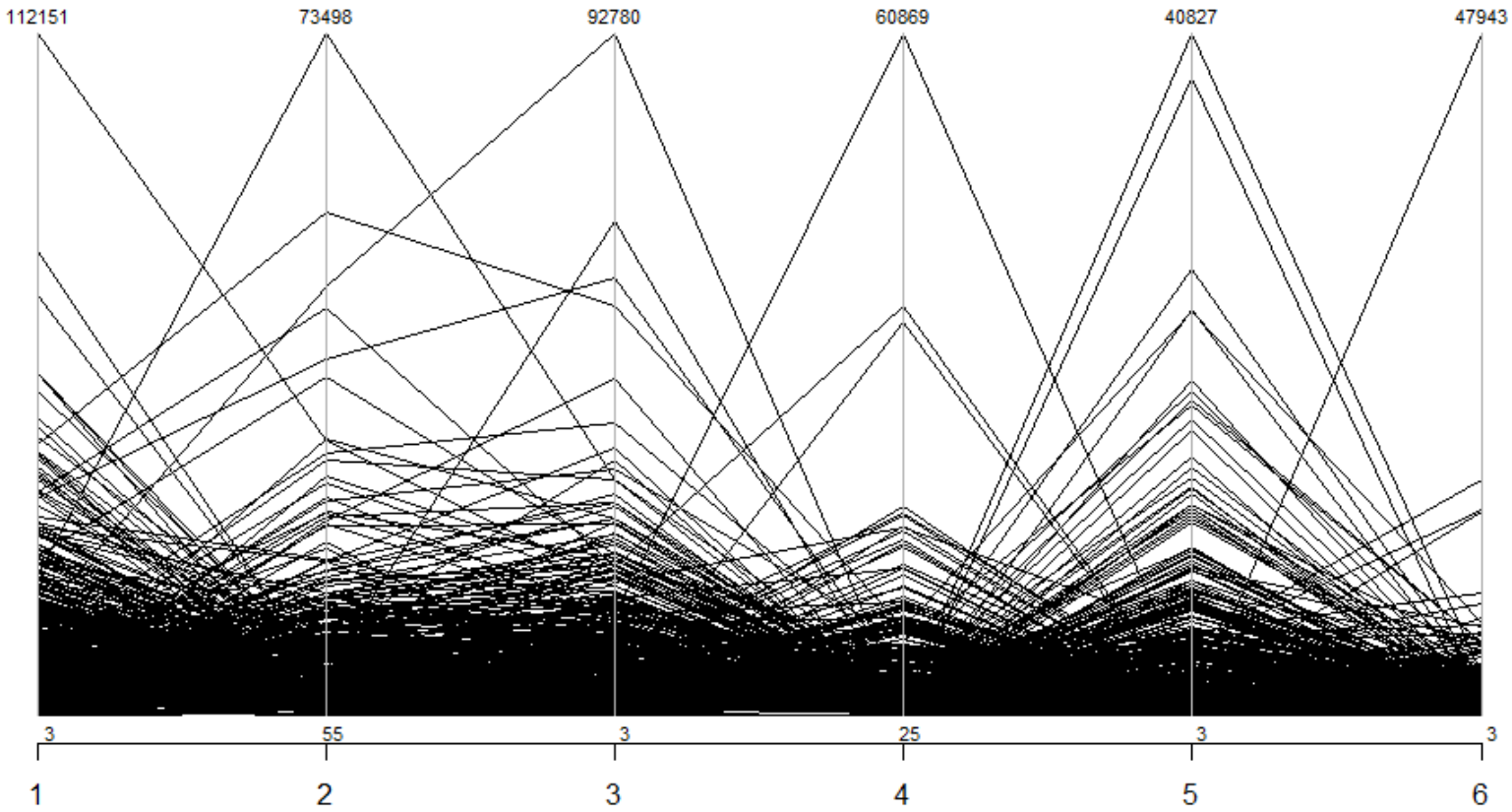
Attribute	Annual Spending/Description	Type
FRESH	Fresh Products	Continuous
MILK	Milk Products	Continuous
GROCERY	Grocery Products	Continuous
FROZEN	Frozen Products	Continuous
DETERGENTS	Detergents and Paper Products	Continuous
DELICATESSEN	Delicatessen Products	Continuous
CHANNEL	Horeca or Retail Channel	Nominal
REGION	Lisnon, Oporto or Other	Nominal

Plots

Histograms and Boxplots for descriptive knowhow of data



Parallel Coordinate Plot



Problem

Statement

Problem: *Group* wholesale customers on the basis of their buying patterns

Method: Use *Hierarchical Clustering* to group based on “similar” buying patterns

Motivation

Using *Unsupervised Learning Methods* to group data without any previously known attributes of such groups

Assume these values are output of some internal function unknown to us

K-means VS Hierarchical

- Complexity: $O(n)$ and $O(n^2)$
- Consistency: *K-means* renders different results with every run
- No previously known k!

Hierarchical Clustering

Method

Builds *clusters*, i.e. groups that have maximum similarity with each other in the same cluster and maximum dissimilarity with other clusters

Results in *hierarchy of clusters*

Agglomerative

"bottom up" approach; each starts in its own cluster and then merged as we move up the hierarchy

Divisive

"top down" approach; all observations start in one cluster, and splits as one moves down the hierarchy

Measuring Similarity

How to measure similarity between two clusters?

- Euclidean Distance
  $d_0^2 = (\mathbf{x}_1 - \mathbf{x}_2)^T(\mathbf{x}_1 - \mathbf{x}_2)$
- Correlation Distance
  $d_0 = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1) \cdot Var(X_2)}}$ (Pearson's Correlation Coefficient)

Measuring Similarity

Examples

- **Similarity based on Sizes:** grouping based on buying sizes; grouping which fruits are larger in size and which are smaller
- **Similarity based on Attributes:** grouping based on buying patterns; grouping which fruits are edible and which aren't

Measuring Distance - Linkages

How to measure distance between a cluster and a point of another cluster?

- **Single** - minimum of all possible distances; chaining effects
- **Complete** - maximum of all possible distances; no chaining; affected by outliers
- **Average** - unweighted average; compromise b/w single and complete
- **Centroid** - average of points in the cluster then distance is calculated

An Example

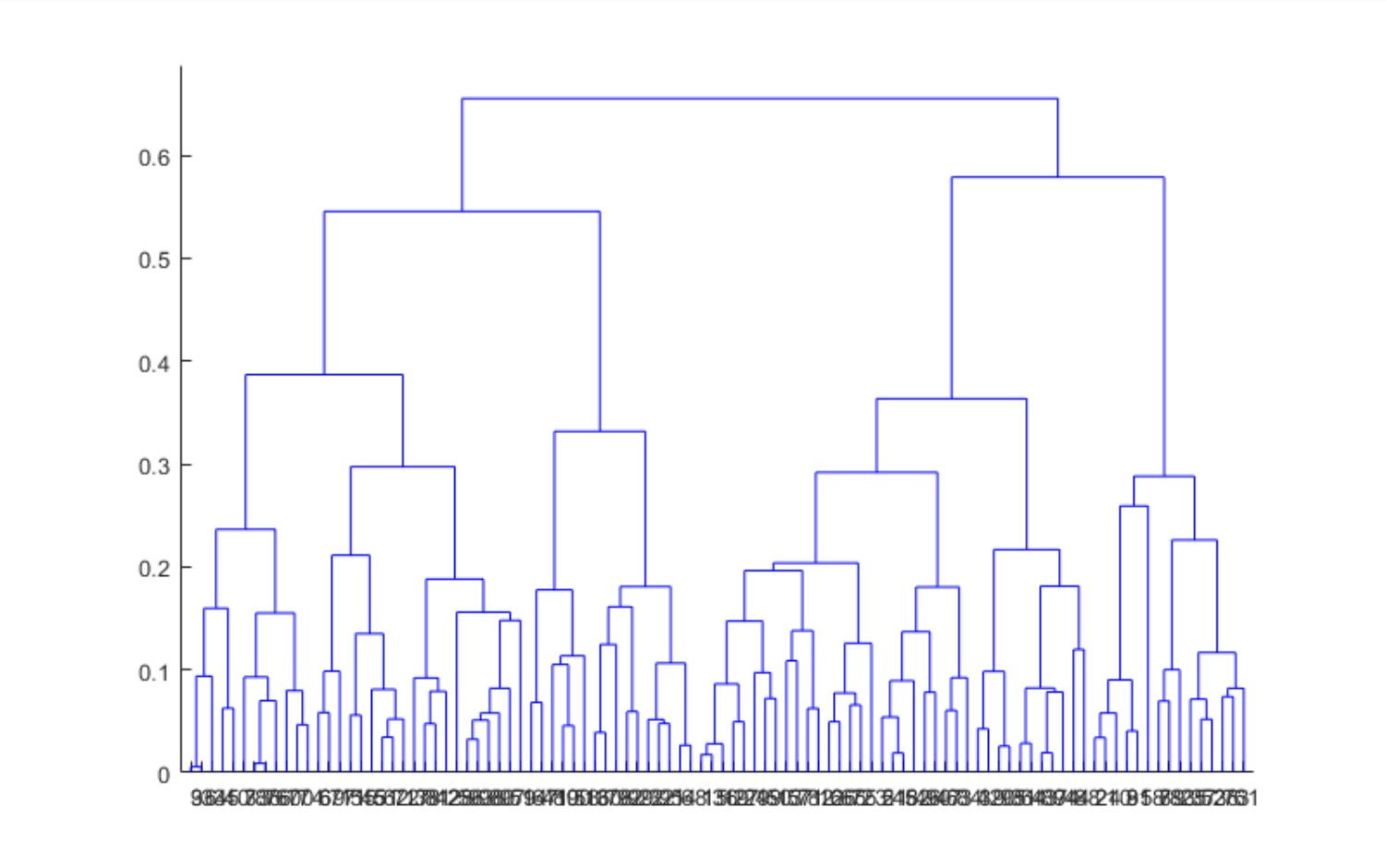
Consider clusters - $\{1, 4\}$ and $\{2\}$

Linkage	Calculation	Distance
Single	$\min\{(2-1), (4-2)\}$	1
Complete	$\max\{(2-1), (4-2)\}$	2
Average	$\frac{(2-1)+(4-2)}{2}$	1.5
Centroid	$\frac{1+4}{2} - 2$	0.5

DENDROGRAMS

A *tree based* diagram to represent *hierarchies* of the hierarchical clustering

Example

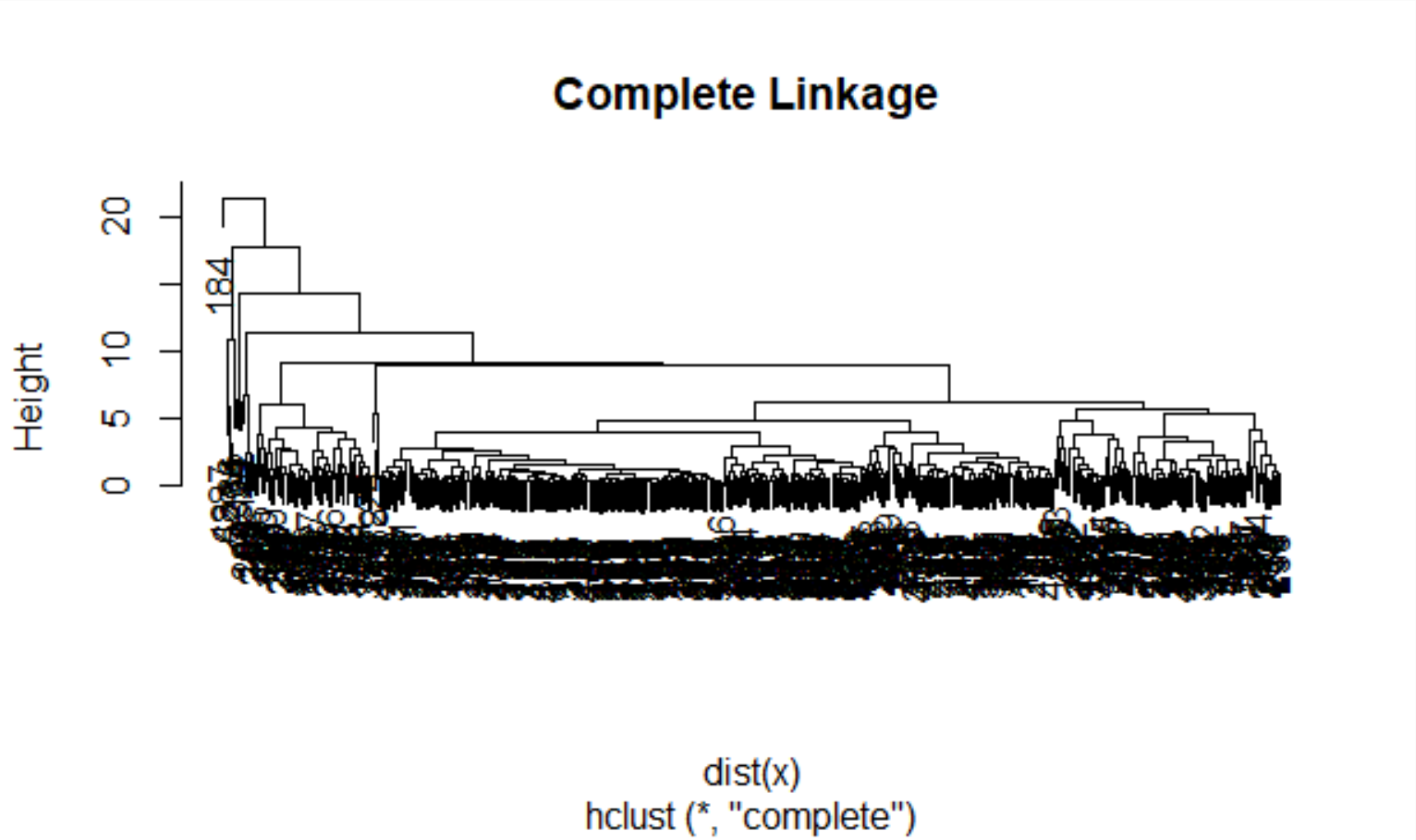


Application

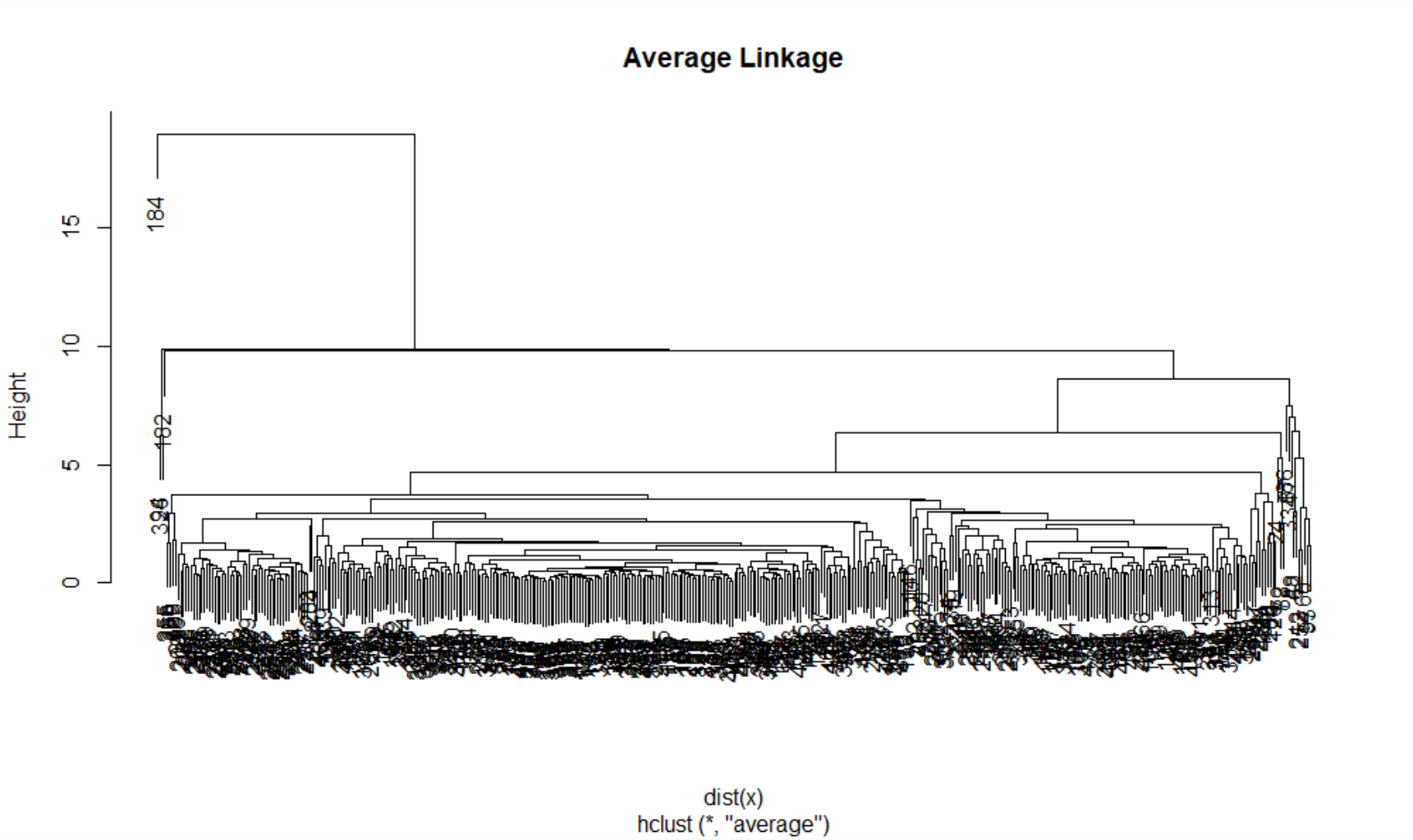
Dendrograms

Dendrogram for Single-Euclidean

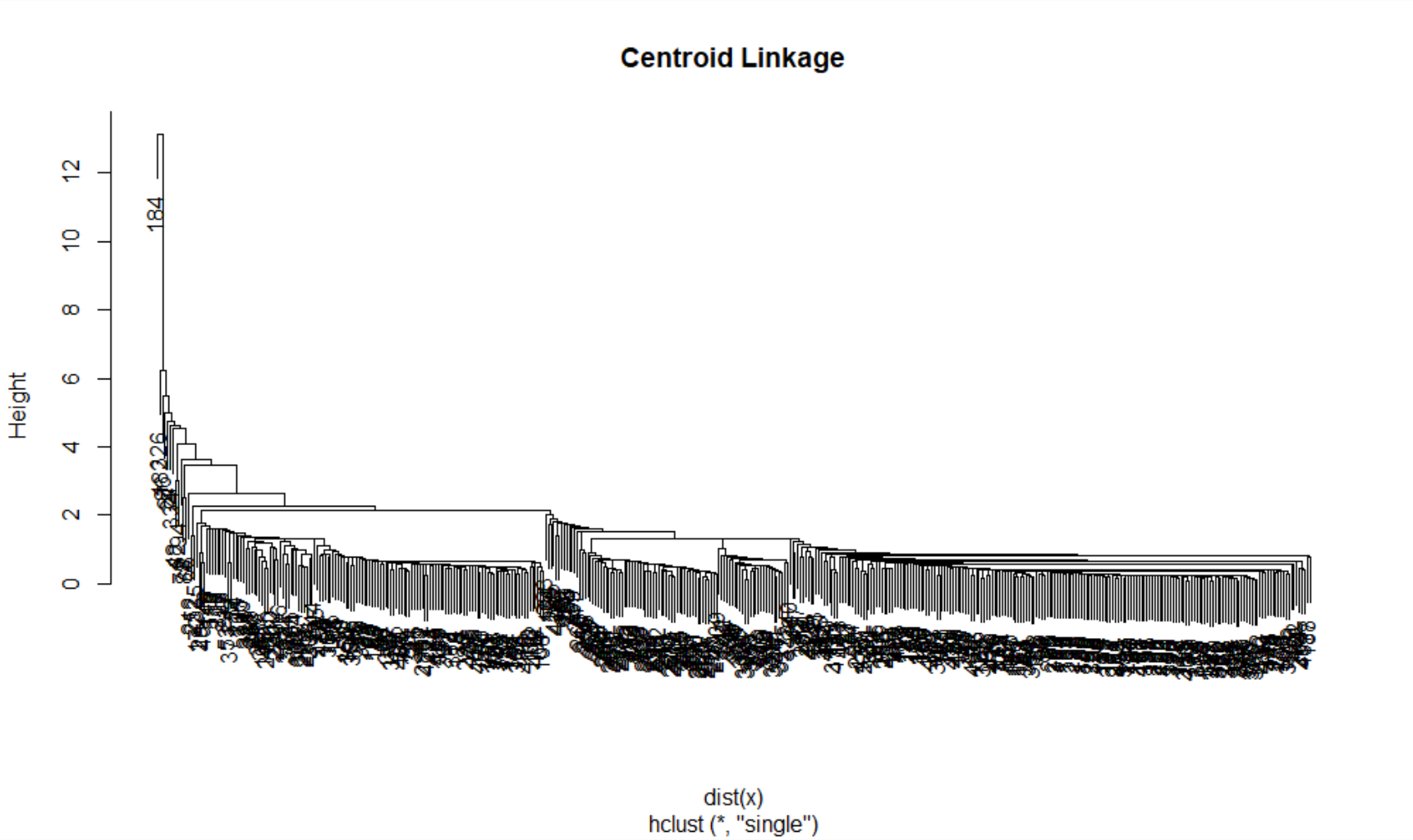
Dendrogram for Complete-Euclidean



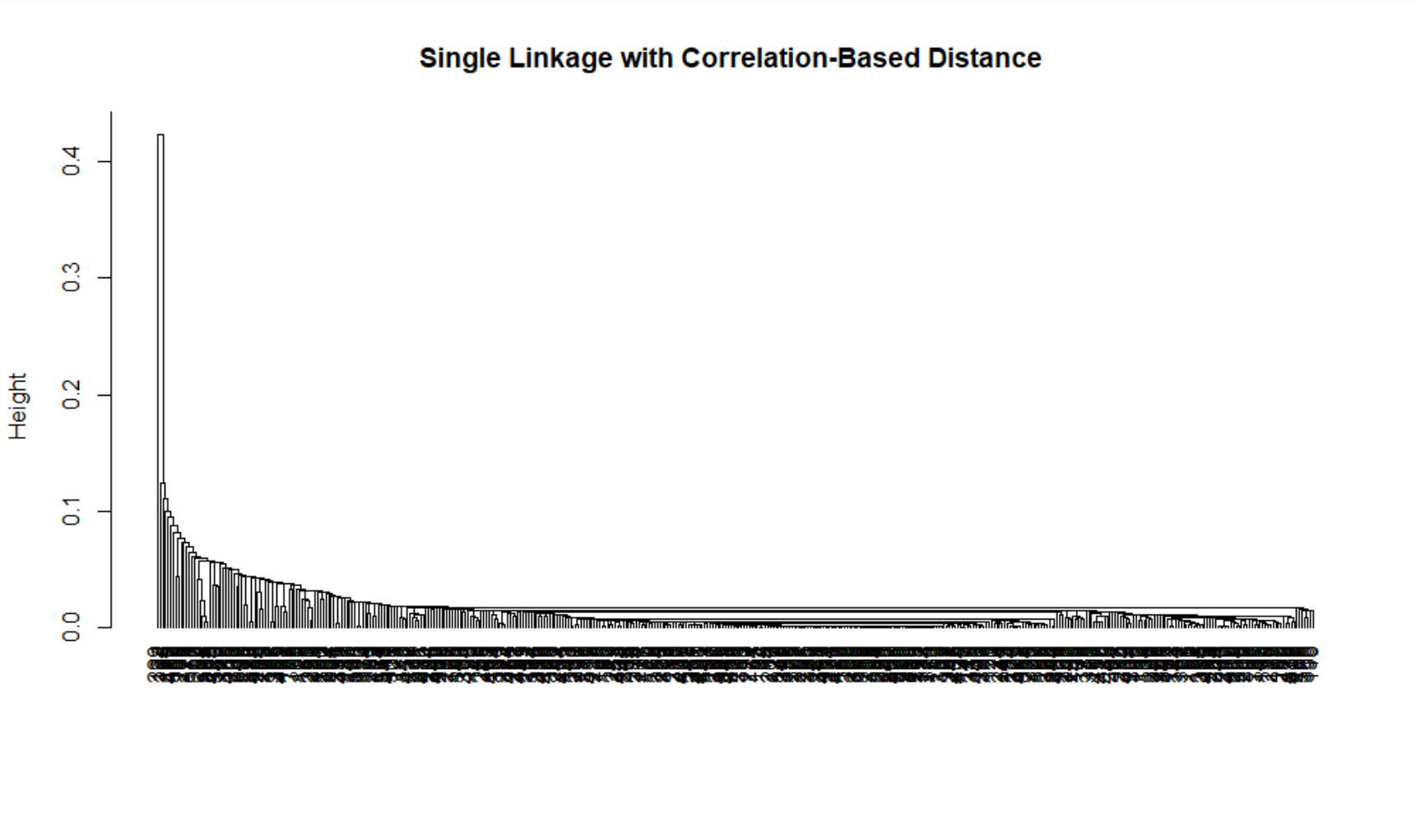
Dendrogram for Average-Euclidean



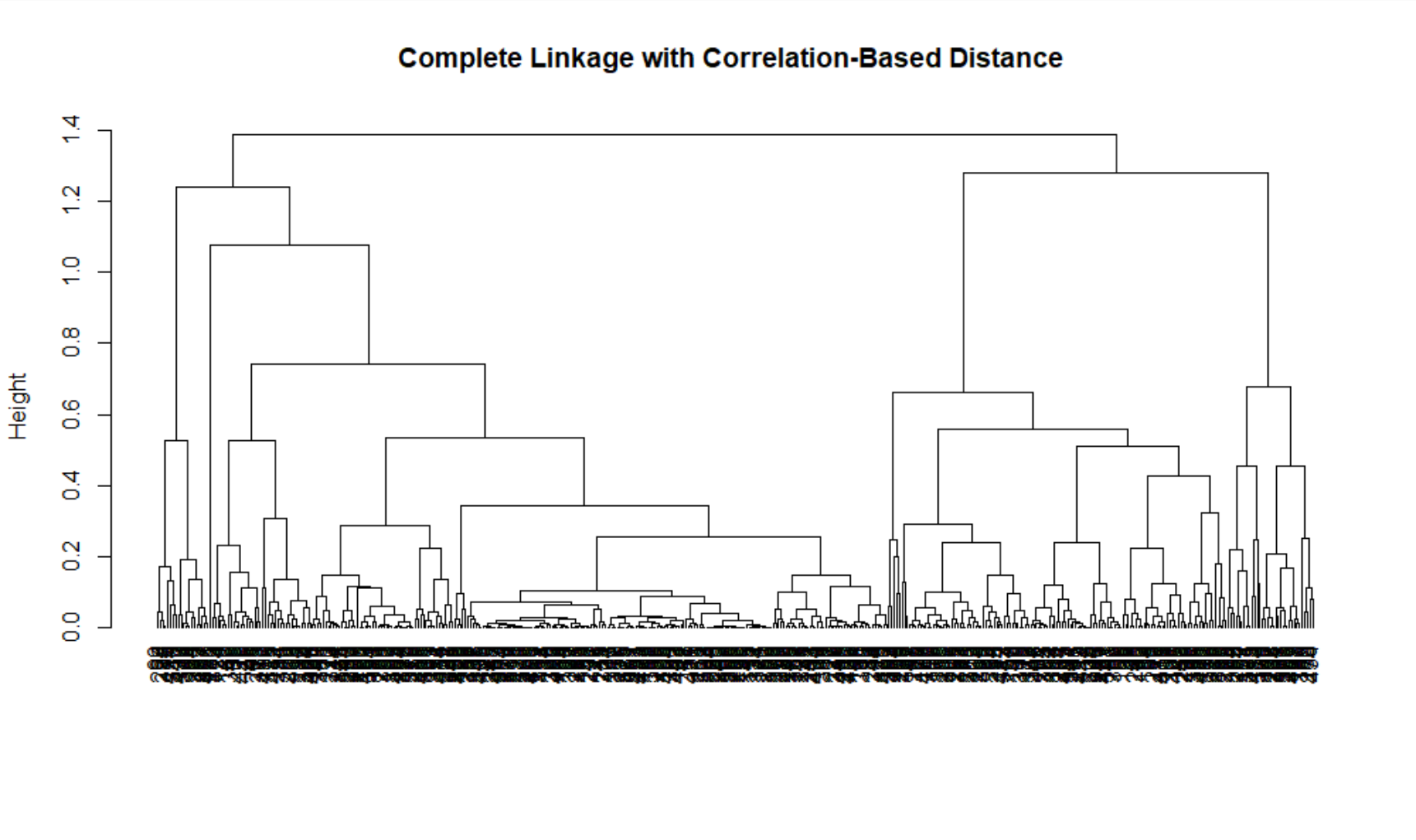
Dendrogram for Centroid-Euclidean



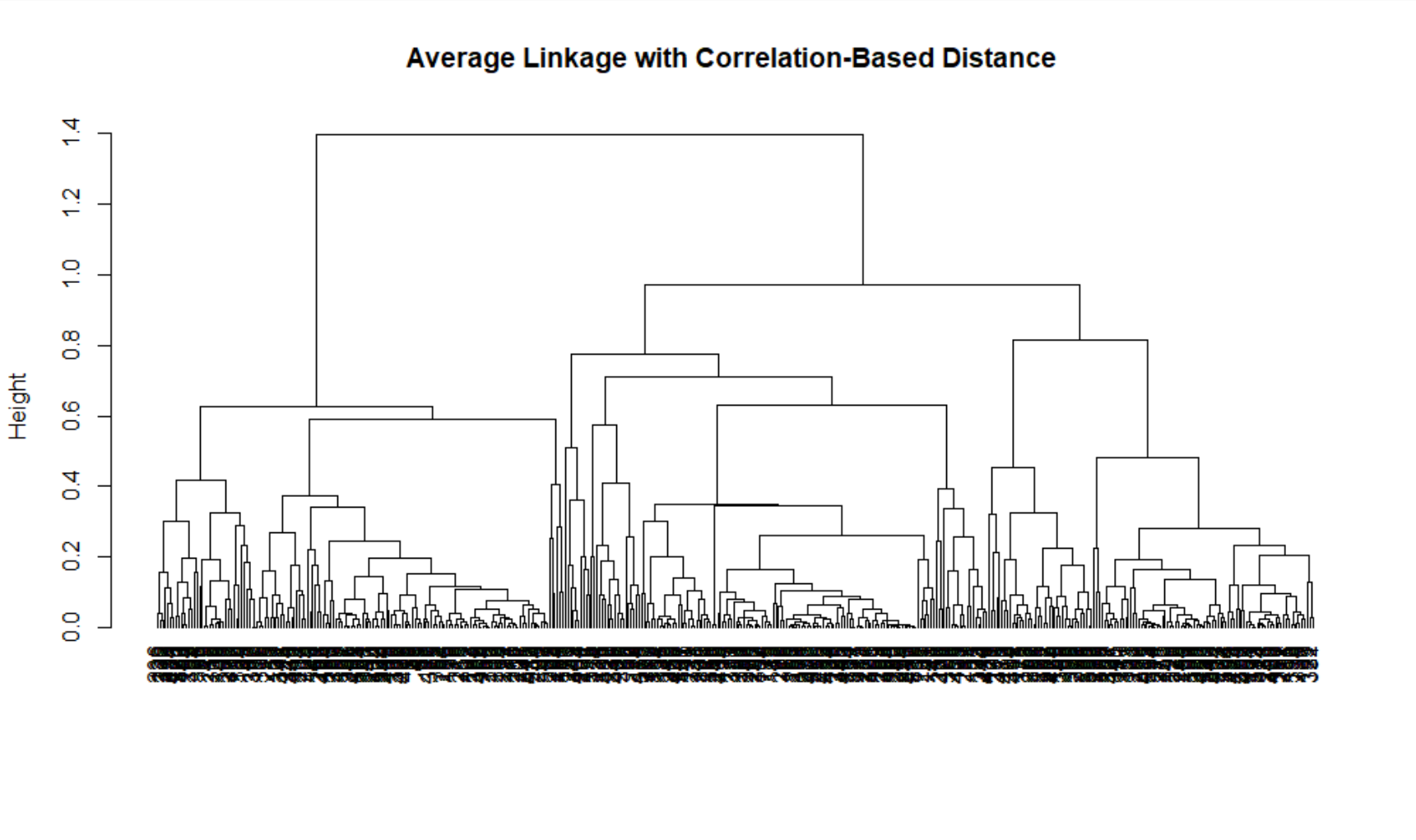
Dendrogram for Single-Correlation



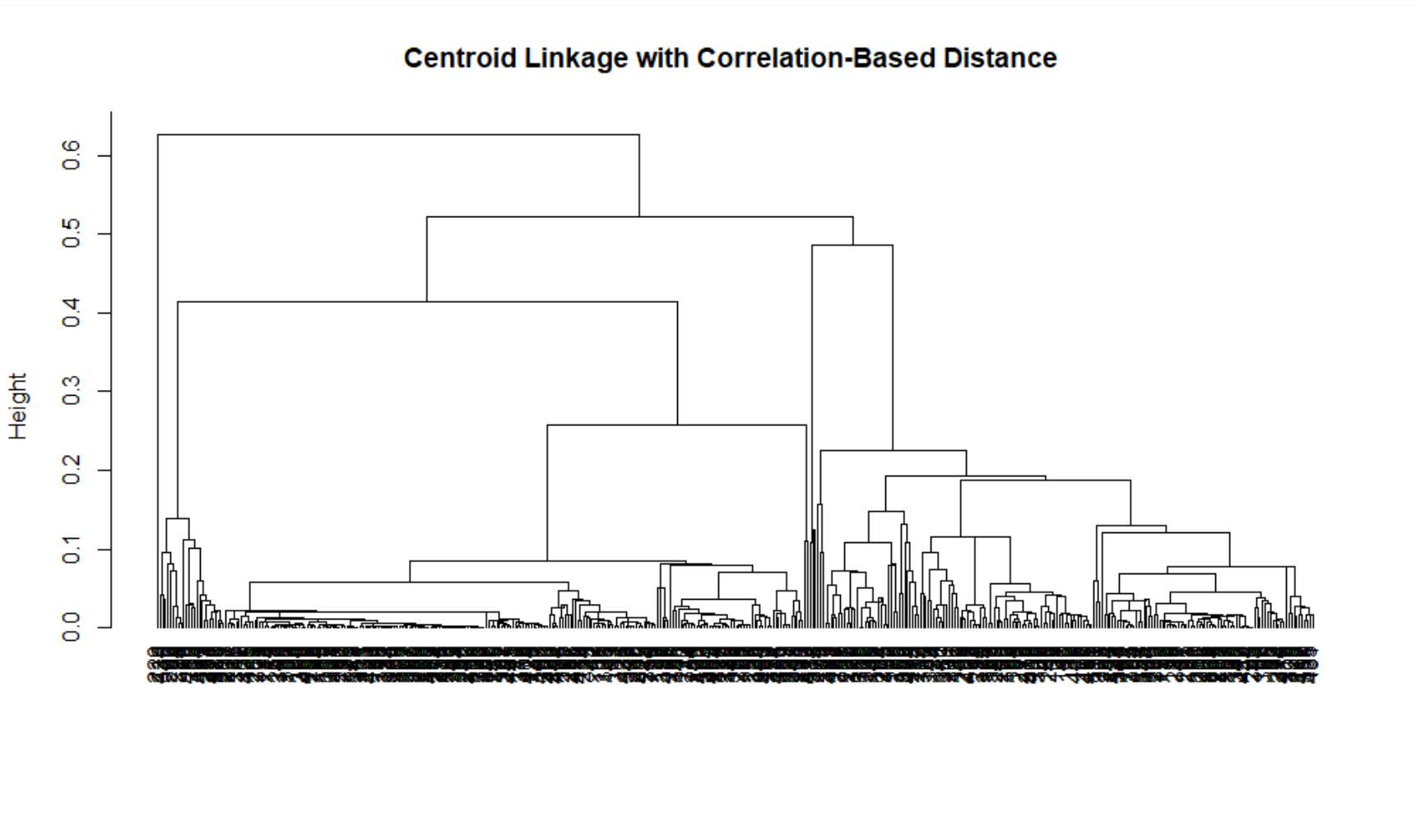
Dendrogram for Complete-Correlation



Dendrogram for Average-Correlation



Dendrogram for Centroid-Correlation



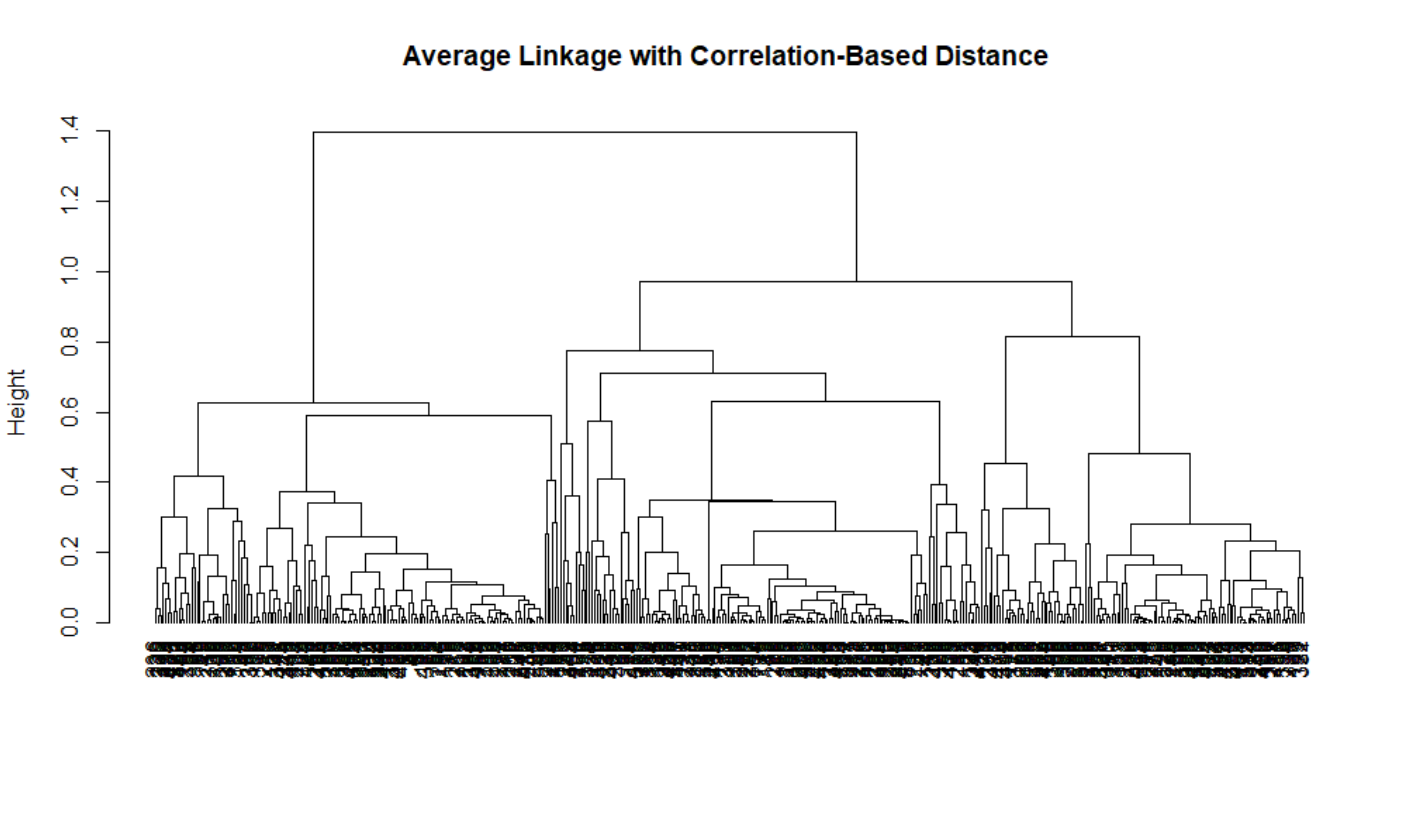
Choosing Distance and Similarity Measure for Post Analysis

Distance Measure *average* over single, complete and centroid

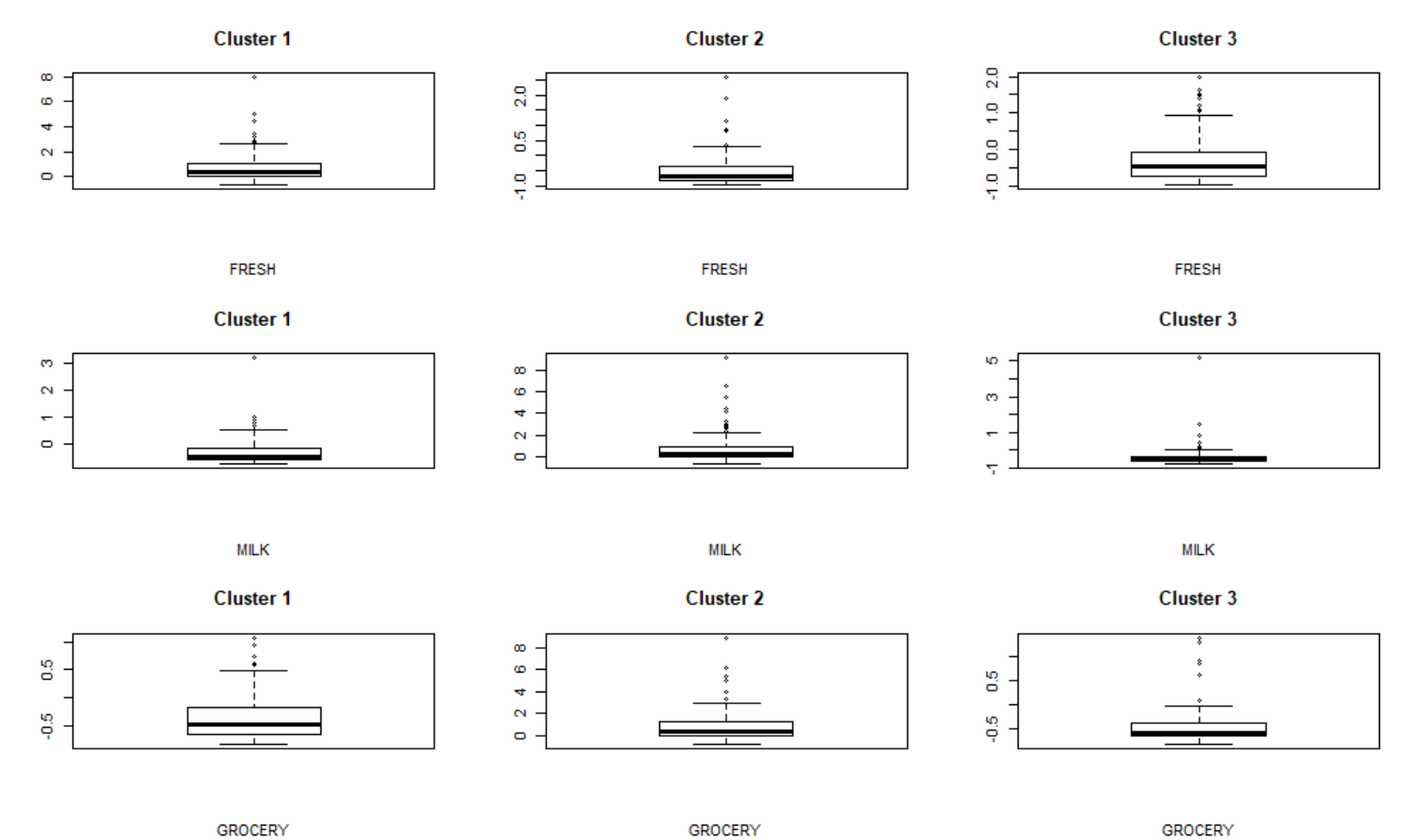
Similarity Measure *correlation* as we wanted to find customers with similar buying pattern rather than size of customers

Number of Clusters

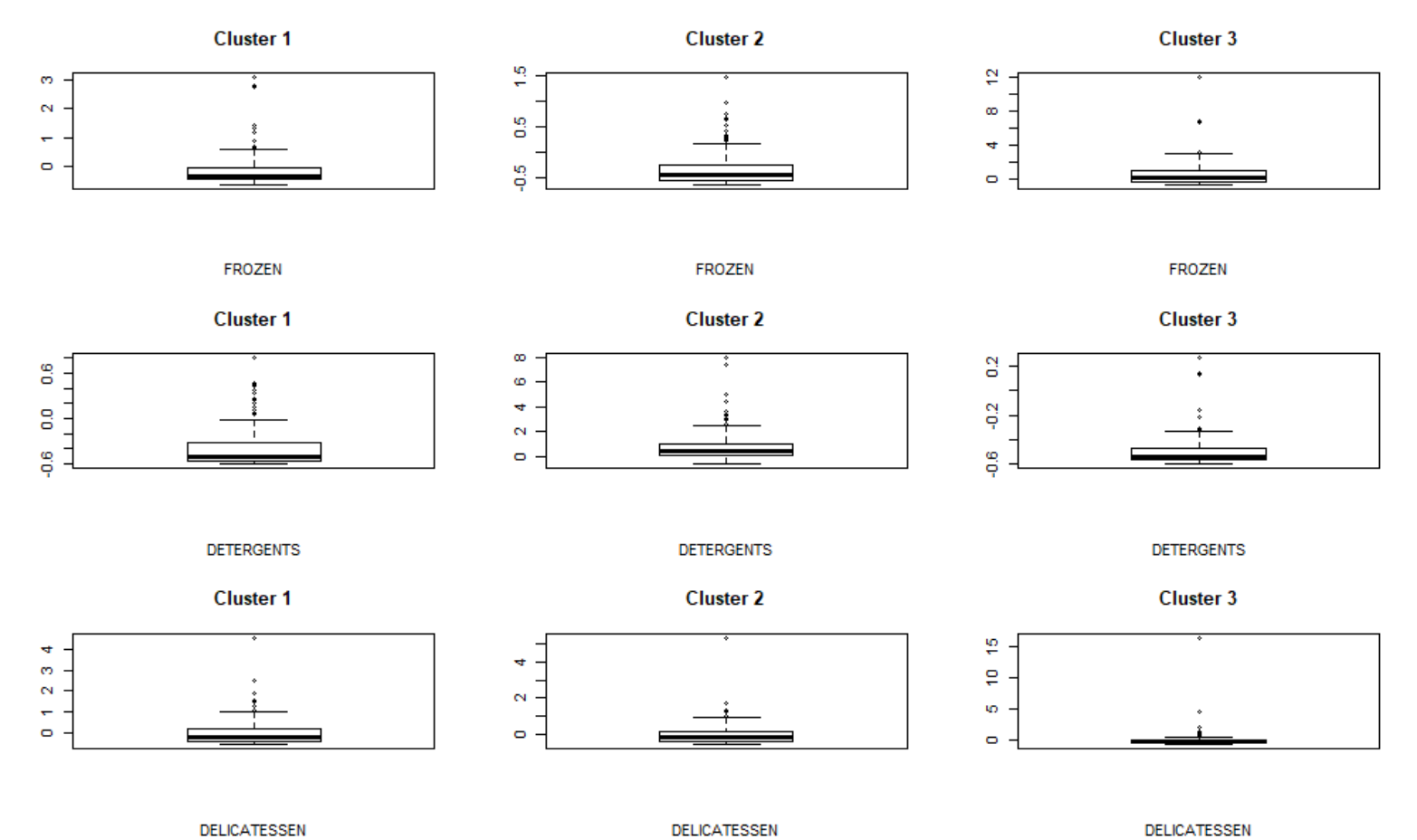
Choose *three* clusters for further analysis.



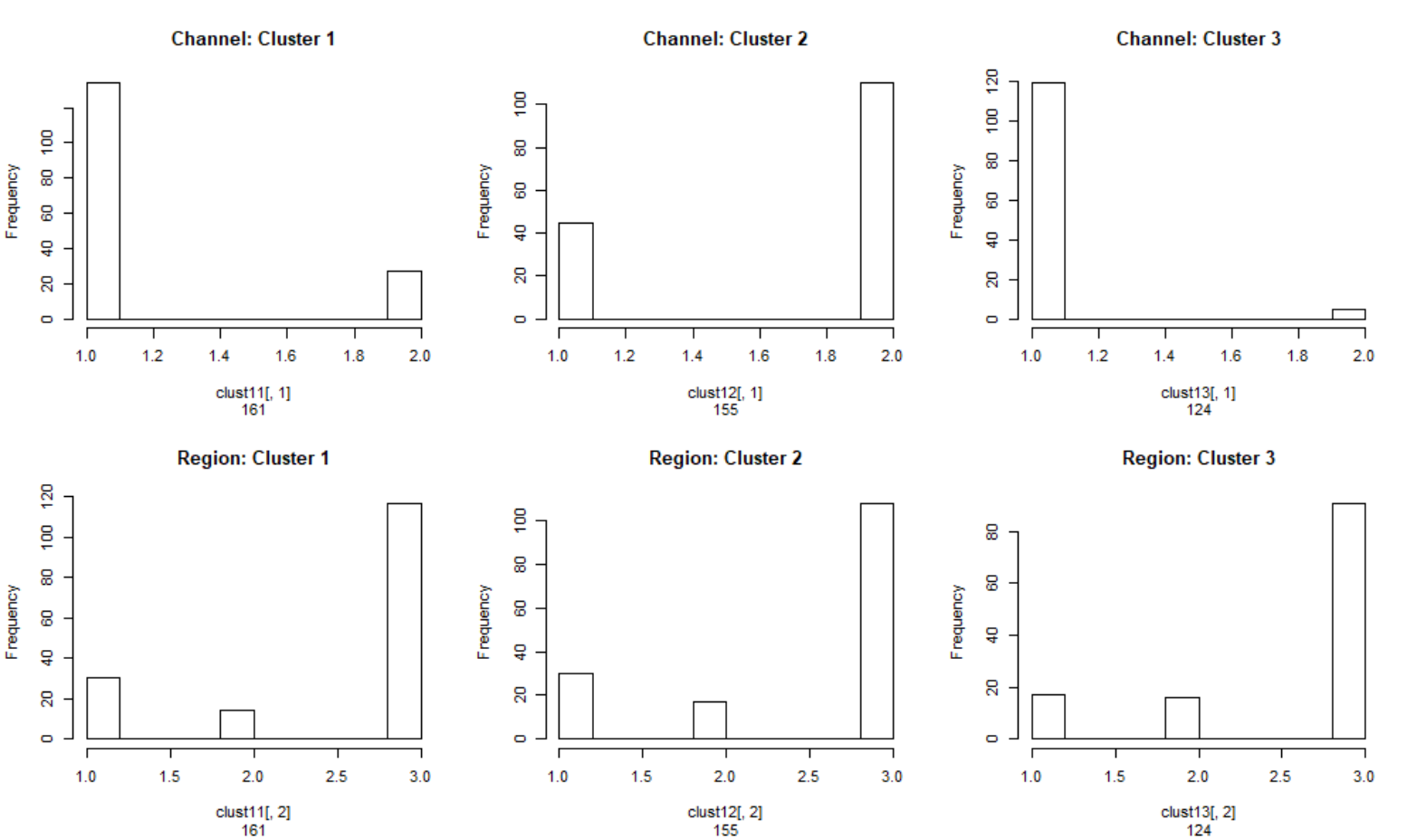
Interpretating Clusters...



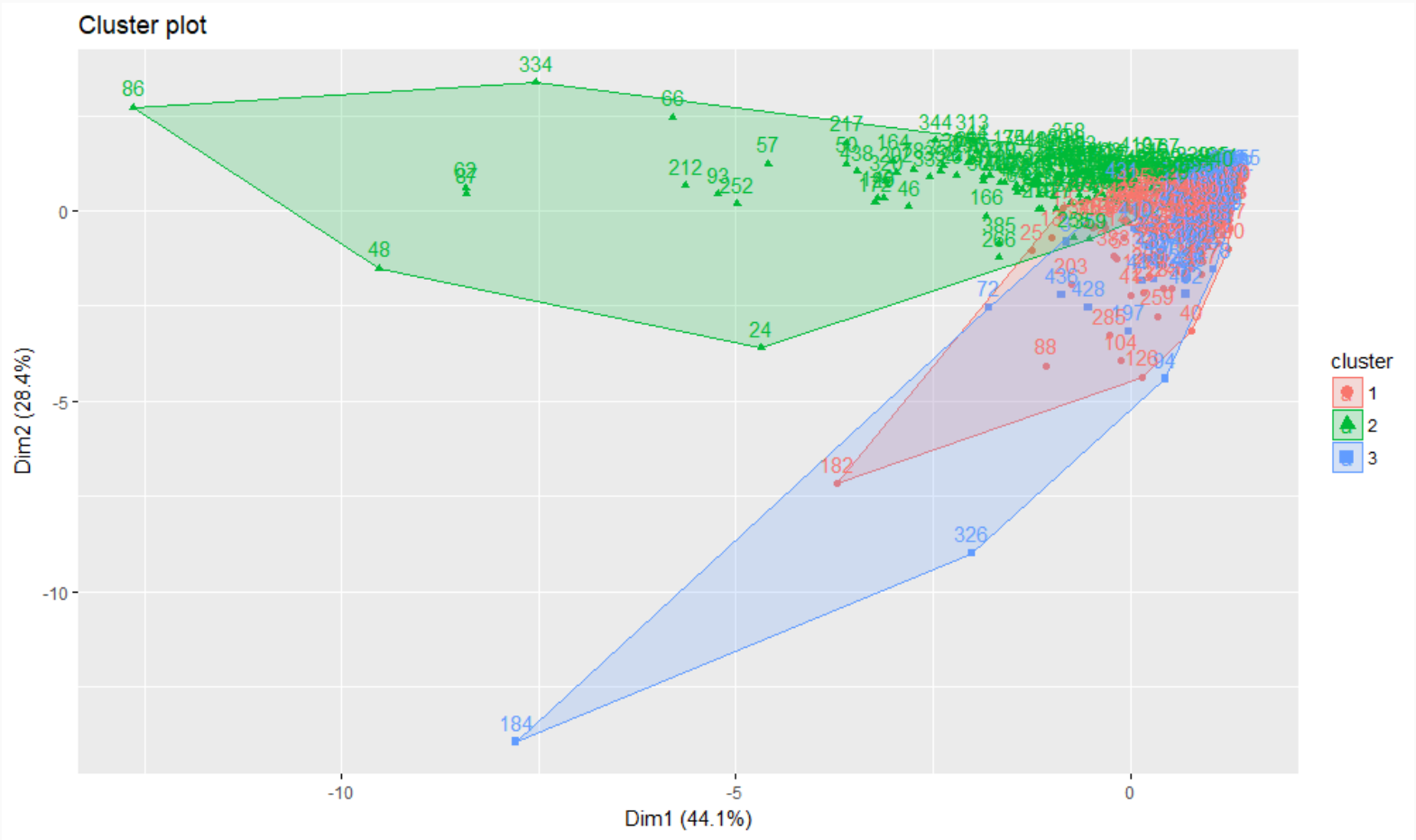
Interpretating Clusters...



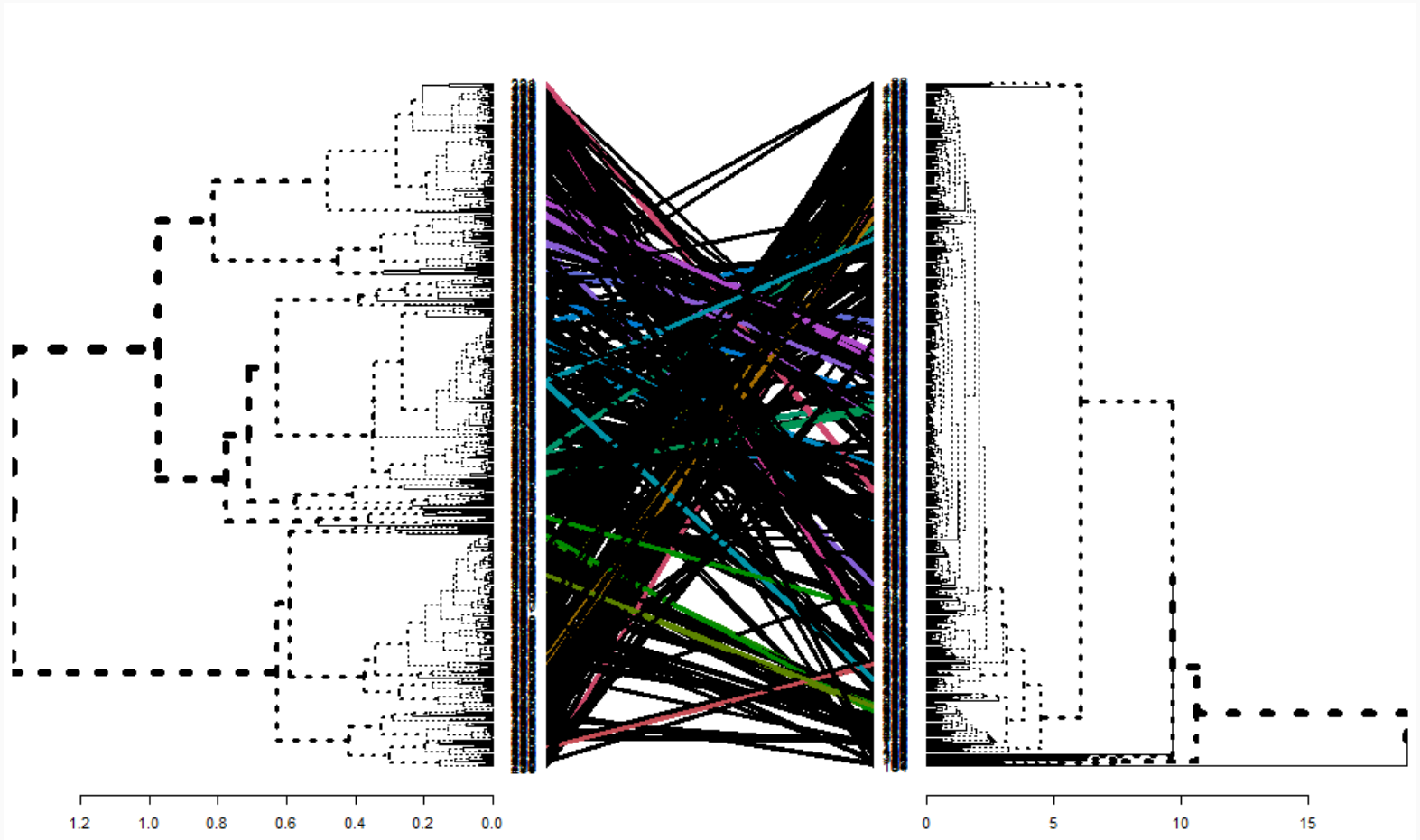
Interpretating Clusters...



Interpretating Clusters...



Tanglegram



Conclusion

Thank You!

Thank you all for you presence and sir, for this opportunity!

Questions?