Отчёт к лабораторной работе

по дисциплине  
«Интеллектуальный анализ данных»

выполнил   
студент гр. ИС/б-18-1-з Демиденко А. А.  
зачётная книжка № 481483  
принял Шумейко И. П.

Лабораторной работа № 4  
«Кластерный анализ. Основные этапы и задачи кластерного анализа данных»

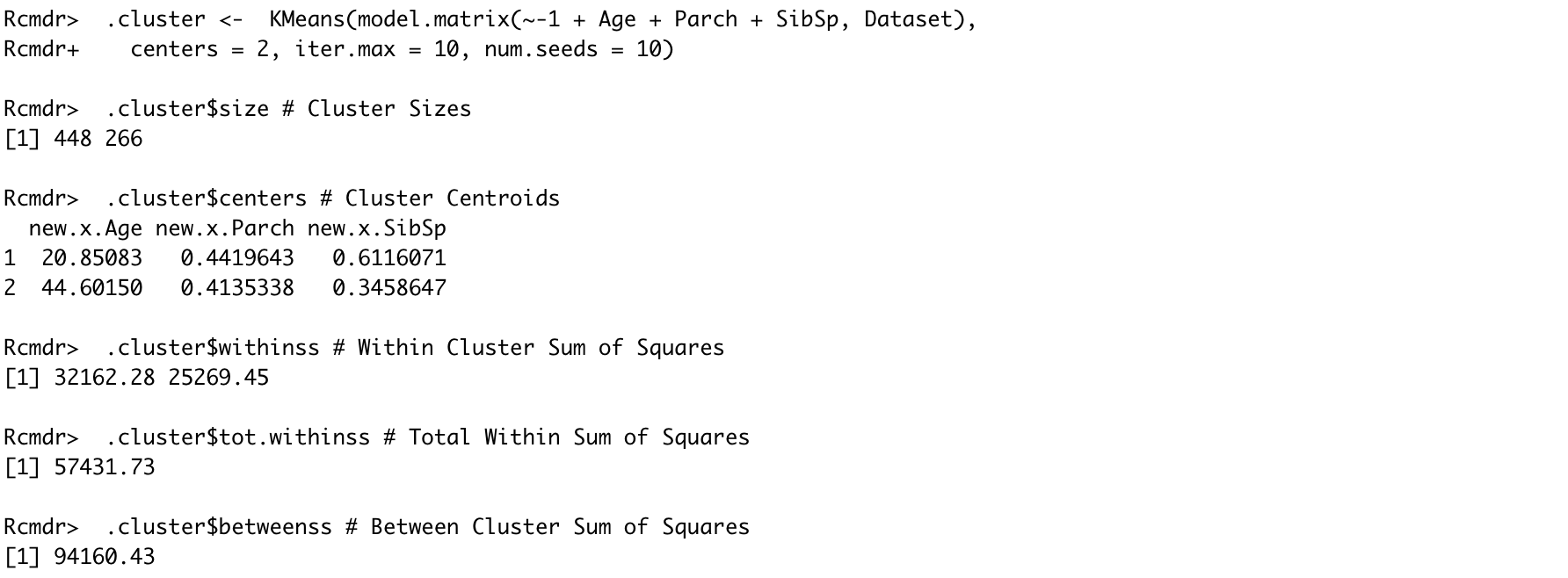
## Цель работы

* закрепить теоретические знания и приобрести практические навыки в проведении кластерного анализа по экспериментальным данным
* исследовать возможности языка R для проведения кластерного анализа

## Ход работы

1. Импортируем экспериментальные данные в пакет Rcmdr. Произведём кластерный анализ методом k-средних несколько раз.

Анализ данных выполним с помощью функции кластерного анализа языка R `kmeans()`:

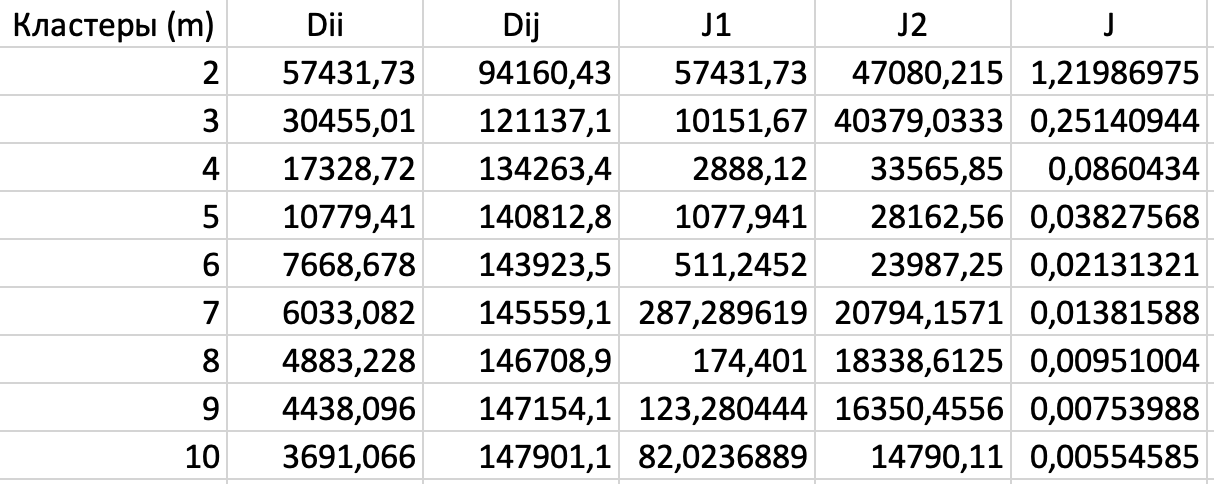


Из анализа выданных результатов разбиения выделим следующие данные:

* Количество элементов в кластерах: первый содержит – 448 объектов, второй – 266
* Сумма квадратов внутри кластера 1 – 32162.28, 2 – 25269.45
* Общая сумма квадратов расстояний внутри кластеров –57431.73
* Сумма квадратов расстояний между кластерами – 94160.43

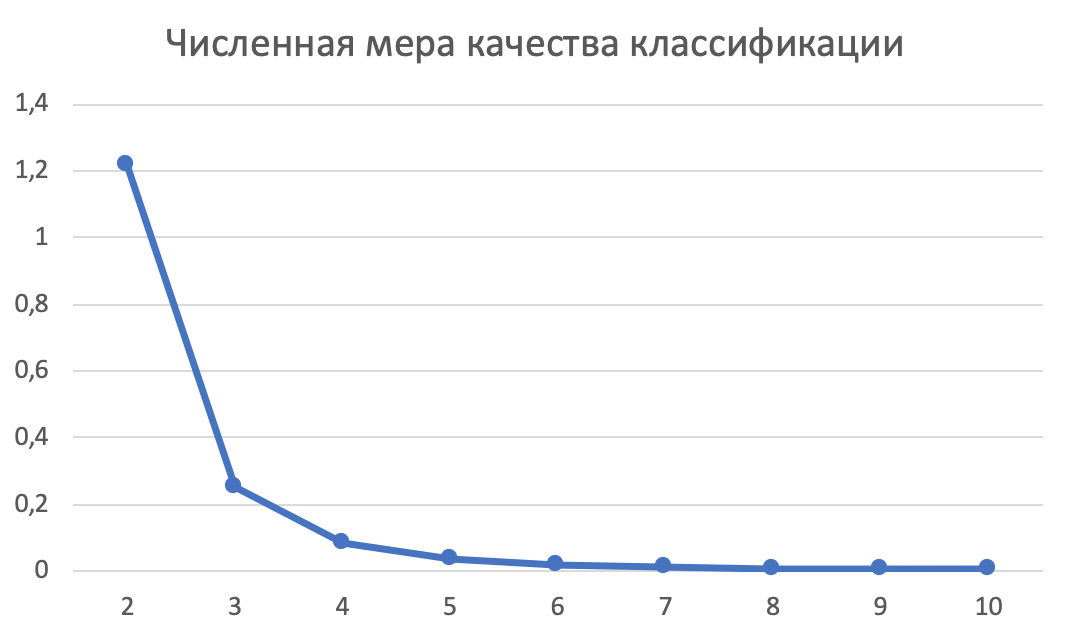
Чтобы выбрать лучшую группировку в смысле критерия минимума отношений средних внутри кластерных и меж кластерных расстояний было проведено деление на 2 – 10 кластеров и заполнена таблица в Excel.

Результаты процедур кластеризации представлены на рисунке 2.

  
Рисунок 1 – Результаты разбиений на кластеры экспериментальных данных

Значения Dii b Dij соответственно обозначают общую сумму квадратов расстояний внутри кластеров и сумму квадратов расстояний между кластерами, withinss и betweenеss на рисунке 1. J – одна из оценок качества классификации, предназначенная для сравнения нескольких типизаций и выбора наиболее оптимальной из них.

Для графической интерпретации значения показателя качества классификации используем критерий «каменистой осыпи», который заключается в нахождении такой абсциссы на графике, в которой уменьшение стресса максимально замедляется. График изображён на рисунке 2.

  
Рисунок 2 – График численной меры качества классификации

По рисунку видно, что в соответствии с этим критерием оптимальным разбиением экспериментальных данных является разбиение на 3 кластера.

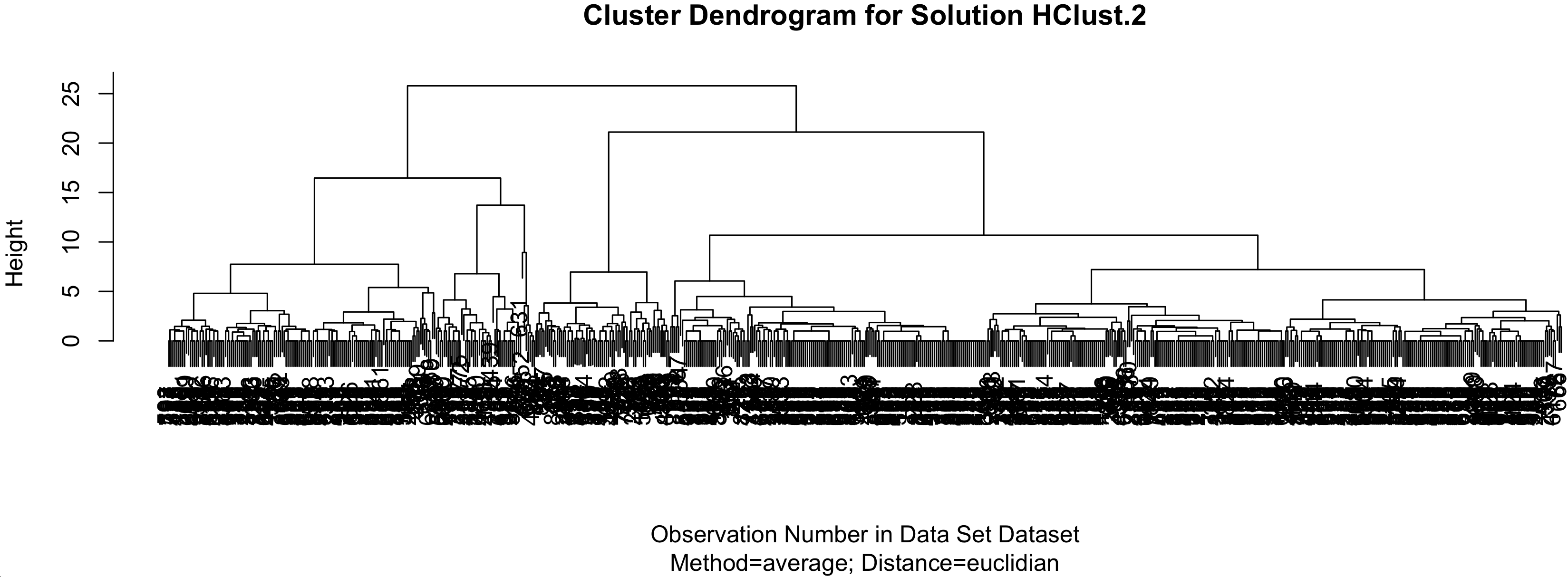
Произведём иерархический кластерный анализ.

Существуют иерархические агломеративные методы, идея которых – последовательное объединение исходных элементов и соответствующее уменьшение числа кластеров и дивизимные методы, смысл которых – делить один общий кластер на меньшие кластеры, образуя последовательность расщепляющих групп.

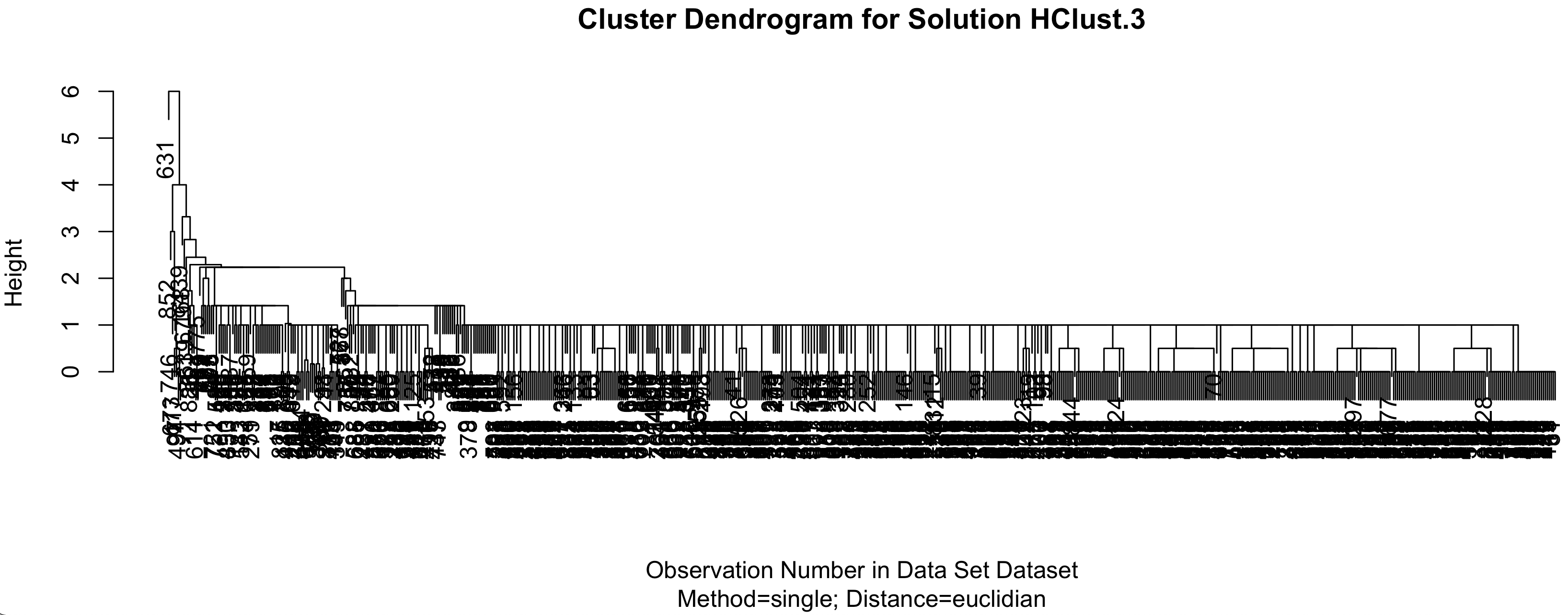
В языке R для иерархического кластерного анализа агломеративными методами используются функции `hclust()` и `agnes()`, для дивизимного – функция `diana()`.

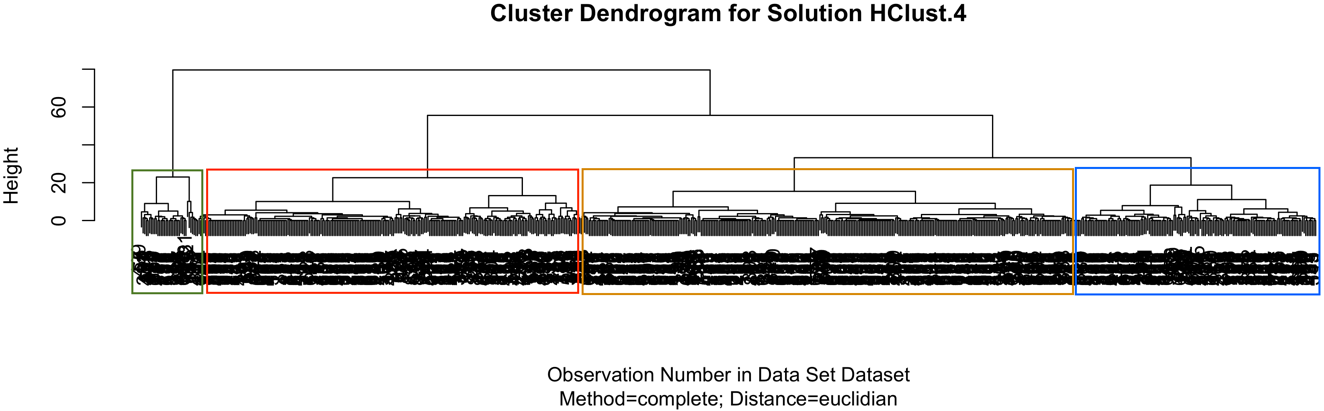
В зависимости от правил вычисления расстояния между кластерами выделяют несколько различных алгоритмов, таких как алгоритм средней связи, полной связи и простой связи.

Алгоритм средней связи (см. рисунок 3), на каждом следующем шаге объединяет два ближайших кластера, рассчитывая среднюю арифметическую дистанцию между всеми парами объектов.

  
Рисунок 3 – Дендрограмма агломеративного метода алгоритма средней связи

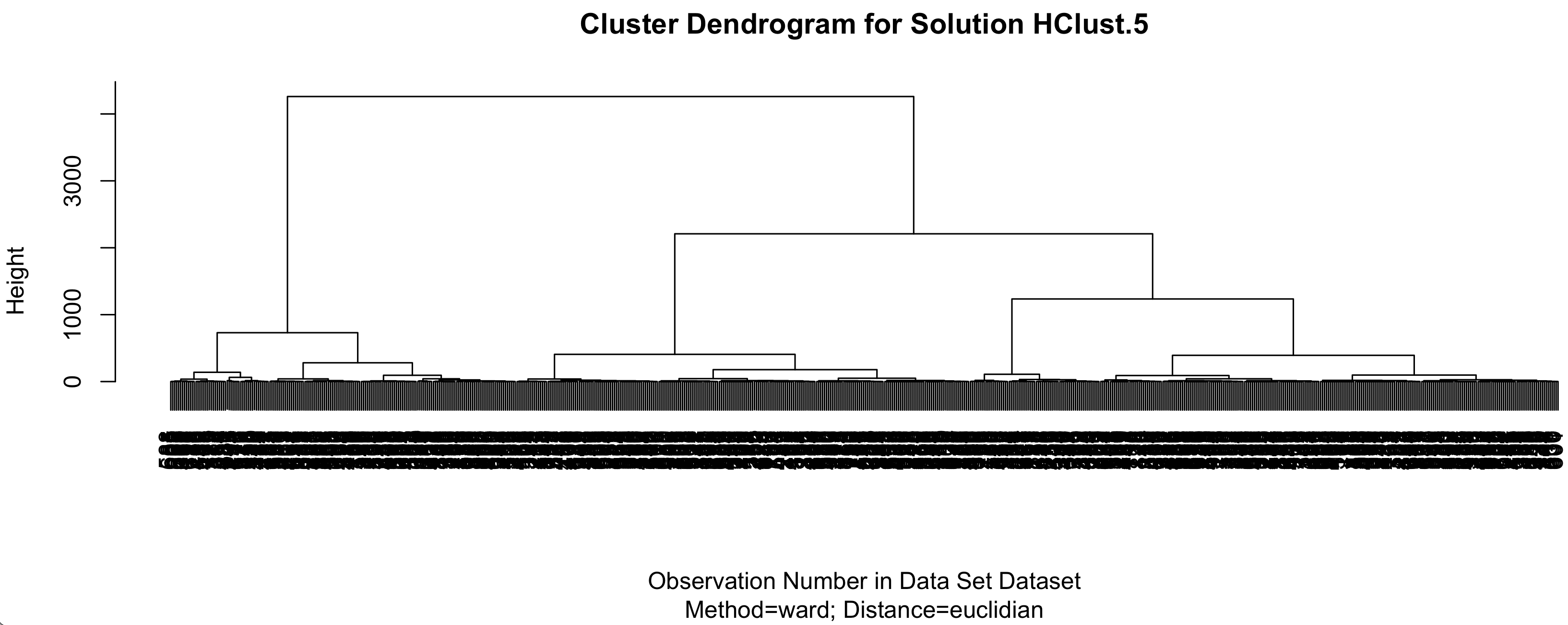
При использовании алгоритма простой связи (см. рисунок 4) расстояние между кластерами оценивается как минимальное из дистанций между парами объектов из разных кластеров, обратно алгоритму полной связи (см. рисунок 5), когда вычисляется расстояние между наиболее удалёнными объектами.

  
Рисунок 4 – Дендрограмма алгоритма одиночной (простой) связи

  
Рисунок 5 – Дендрограмма алгоритма полной связи

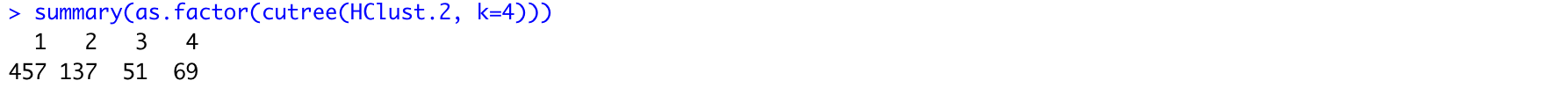
Диаграммы на рисунках 3-5 построены с помощью базовой функции plot.hclust(). Для варианта «complete», визуализирующий алгоритм полной связи, произведено разрезание дерева на 4 группы, которые на дендрограмме принимают вид прямоугольников различного цвета.

На рисунке 6 представлена дендрограмма, полученная методом минимума дисперсий Уорда, в котором в качестве расстояния между кластерами берётся прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения. В отличие от других методов кластерного анализа, для оценки расстояний между кластерами здесь используются методы дисперсионного анализа.

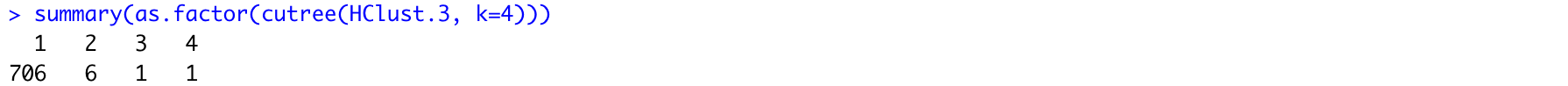
  
Рисунок 6 – Дендрограмма методом Уорда

Разделим все вышеперечисленные дендрограмы на 4 части и посчитаем количество элементов в каждом кластере.

Количество элементов в кластере при использовании метода средней связи:



Количество элементов в кластере при использовании метода простой связи:



Количество элементов в кластере при использовании метода полной связи:



Количество элементов в кластере при использовании метода Уорда:



На основе данных расчётов можно сделать вывод, что наихудший результат показал метод простой связи, наиболее адекватные результаты показали методы средней связи и Уорда.

1. Ответим на контрольные вопросы.

* В чем заключается задача кластерного анализа? Для каких задач обработки экспериментальных данных используются методы иерархического кластерного анализа?

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы).

Иерархические методы кластерного анализа используются при небольших объемах наборов данных.

* Перечислите основные меры сравнения объектов между собой. Что такое дендрограмма?

Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

* Что представляют собой иерархические агломеративные методы кластерного анализа?

Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

* Что представляют собой иерархические дивизимные методы кластерного анализа?

Эти методы являются логической противоположностью агломеративным методам.

* В чем заключается задача неиерархического кластерного анализа? Для каких задач обработки экспериментальных данных используются методы неиерархического кластерного анализа?

При большом количестве наблюдений используют неиерархические методы, основанные на разделении, которые представляют собой итеративные методы дробления исходной совокупности. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки.

* В чем суть алгоритма k-средних? Перечислите основные этапы иерархического кластерного анализа по методу k-средних. Каким образом оценить число кластеров в алгоритме k-средних?

Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

**Выводы**

В ходе выполнения данной лабораторной работы были закреплены теоретические знания и приобретены практические навыки в проведении кластерного анализа по экспериментальным данным. А также были исследованы возможности языка R для проведения кластерного анализа.