Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**ЛАБОРАТОРНАЯ РАБОТА №6.1**

**Дисциплина: Обработка больших данных**

Работу выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_А.М. Нагалевский

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В.И. Шиян

**Цель работы:** закрепить знания об алгоритмах классификации и кластеризации данных, ознакомиться с некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Научиться визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты. Научиться выполнять классификацию на основе формулы Байеса и деревьев решений

**Ход работы:**

**Часть 1.** **Задачи классификации и кластеризации**

Из исходного csv файла данные были импортированы в RStduio. Ниже представлен фрагмент таблицы, всего в ней содержится 97 строк.



Рисунок 1 – Фрагмент исходной таблицы с данными

Для анализа нам необходимо удалить последний столбец region, а также заметим, что значения -9999 соответствуют пропущенным значениям. Так как количество строк с пропущенными значениями незначительно, удалим их из исходной таблицы.



Рисунок 2 – Удалены пропущенные значения.

В данной задаче переменные существенно различны, поэтому необходимо выполнить нормализацию данных.

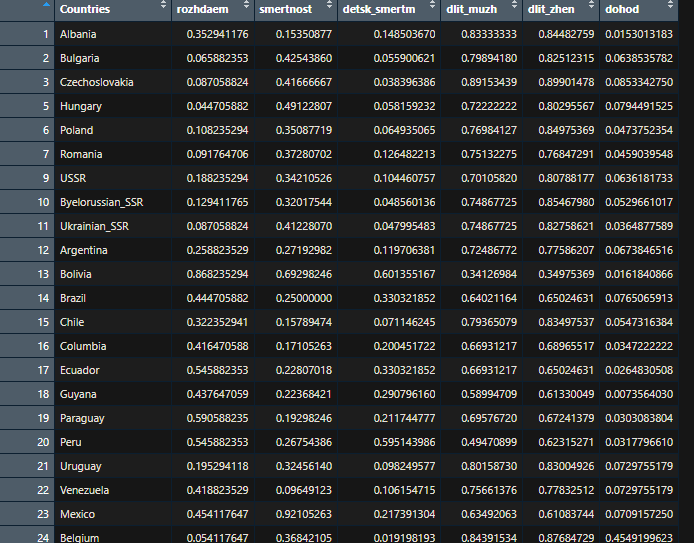


Рисунок 3 – Фрагмент стандартизированной таблицы с данными

По полученным данным построим дендрограмму.

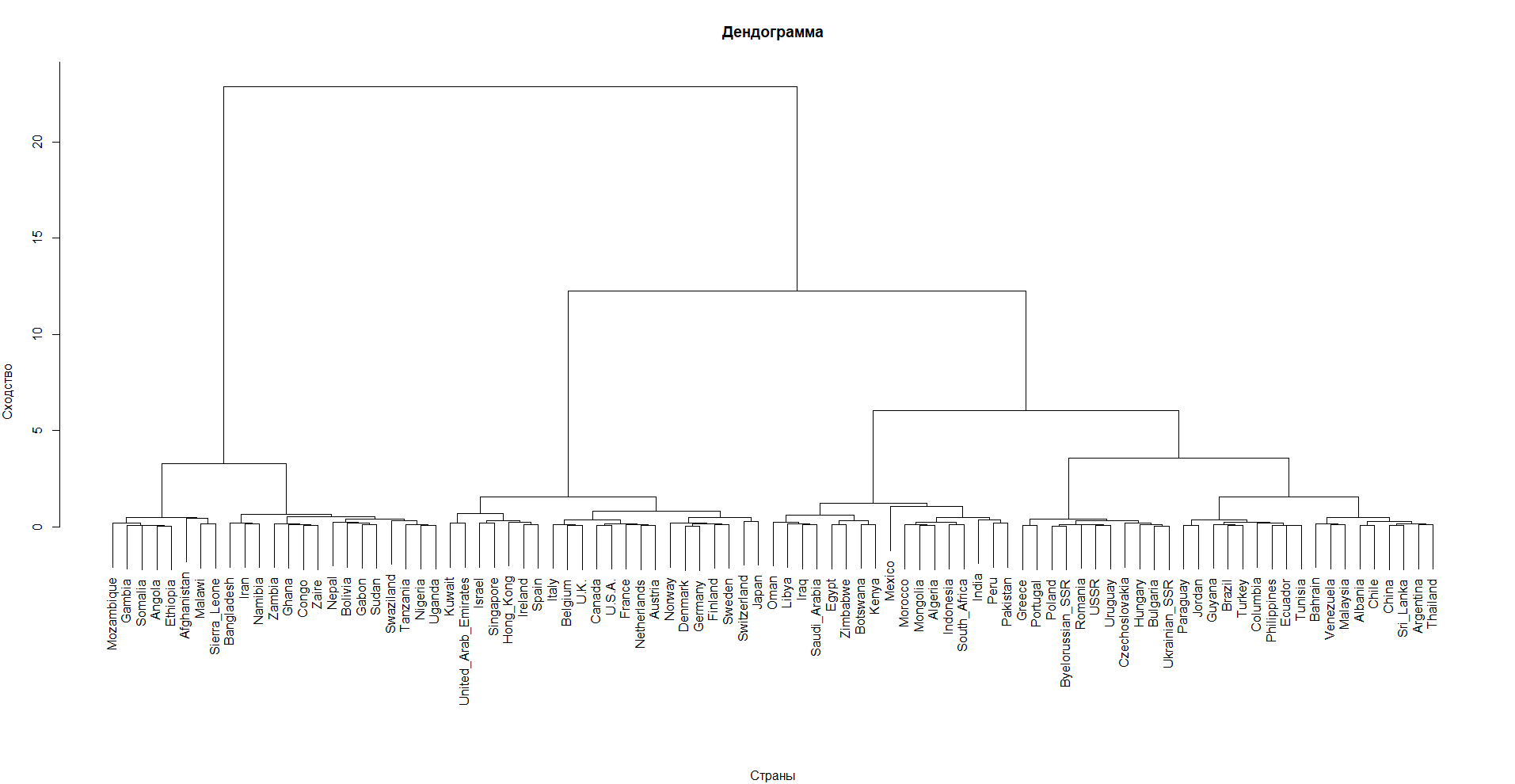


Рисунок 4 – Дендограмма

Целесообразнее выделить 5 кластеров, выделим их красным цветом.

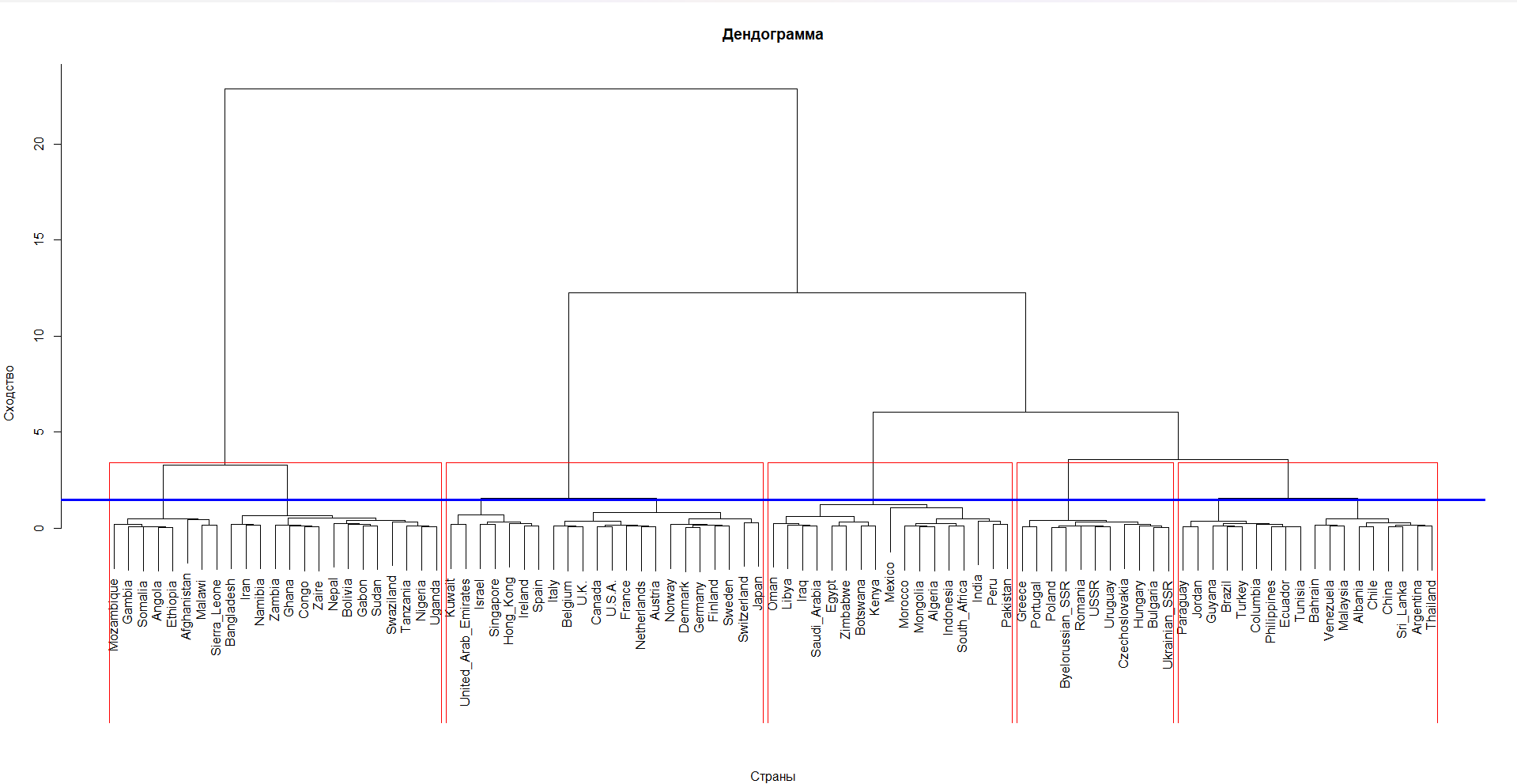


Рисунок 5 – Дендограмма с кластерами.

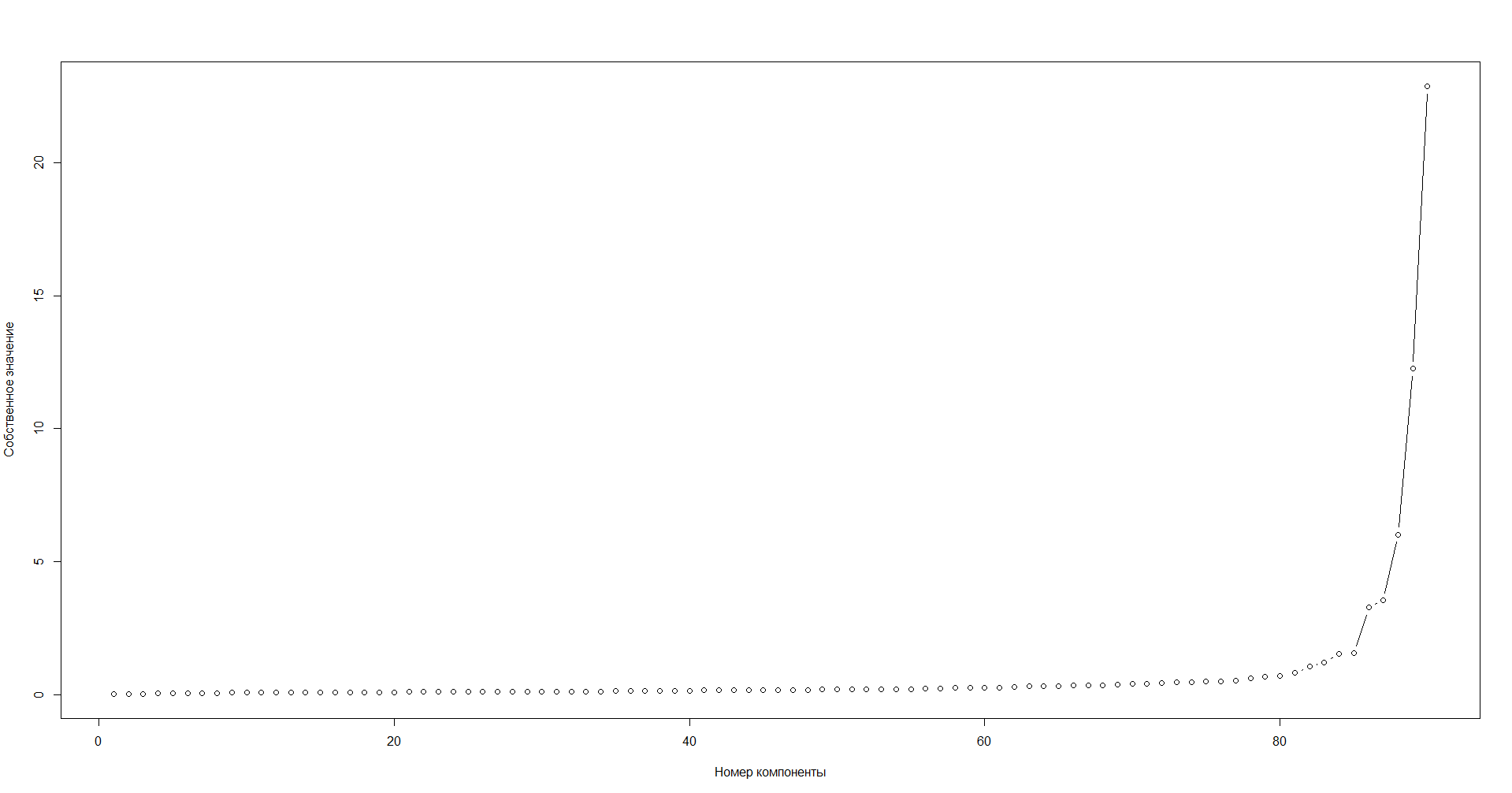


Рисунок 6 – Диаграмма «Каменная осыпь».

Разрезаем дерево, например, полученное в результате hclust, на несколько групп путем указания желаемого количества групп или высоты среза.

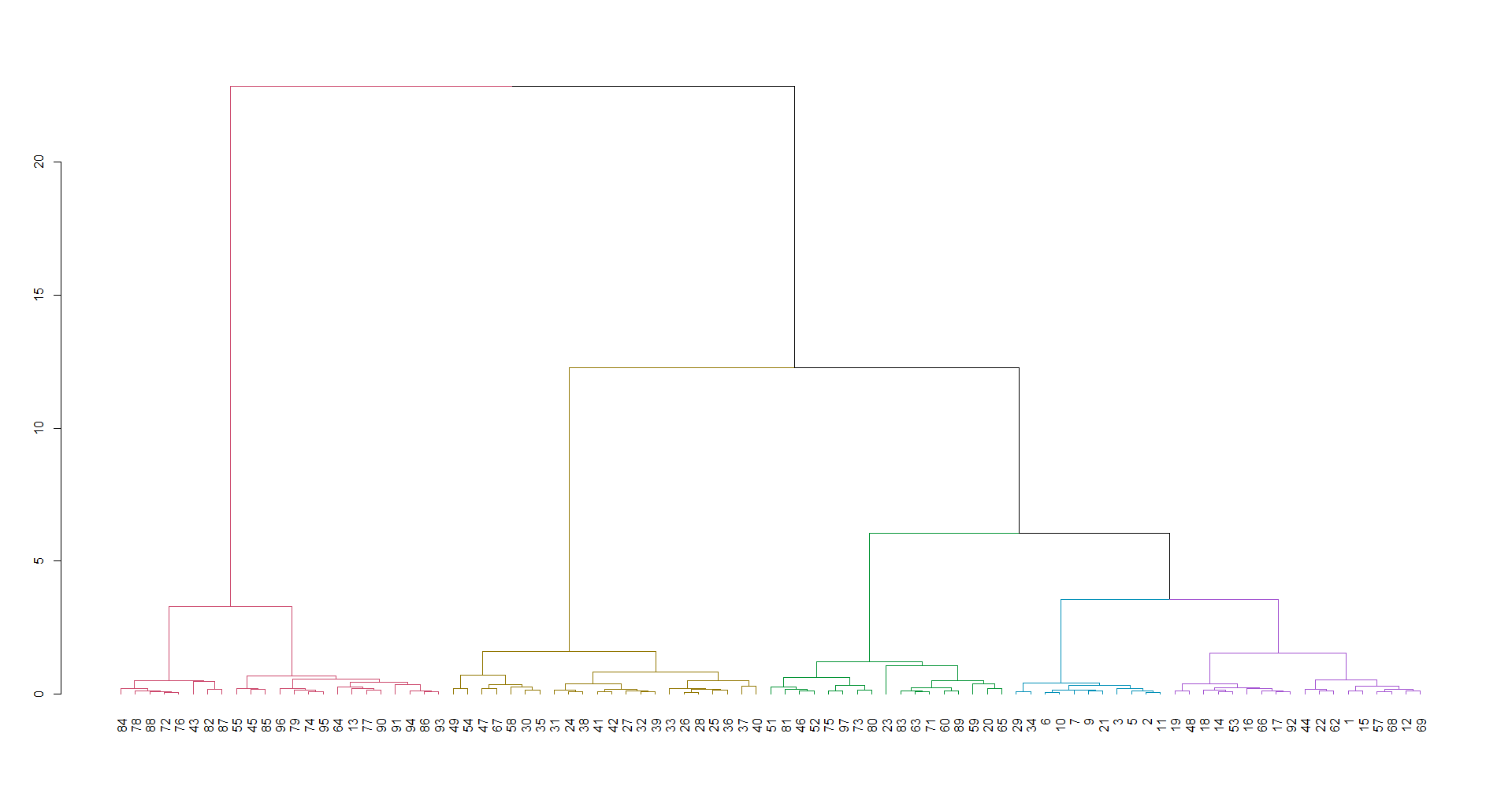


Рисунок 7 – Группы, разбитые по кластерам.

Далее представим списки стран, разбитых по кластерам.

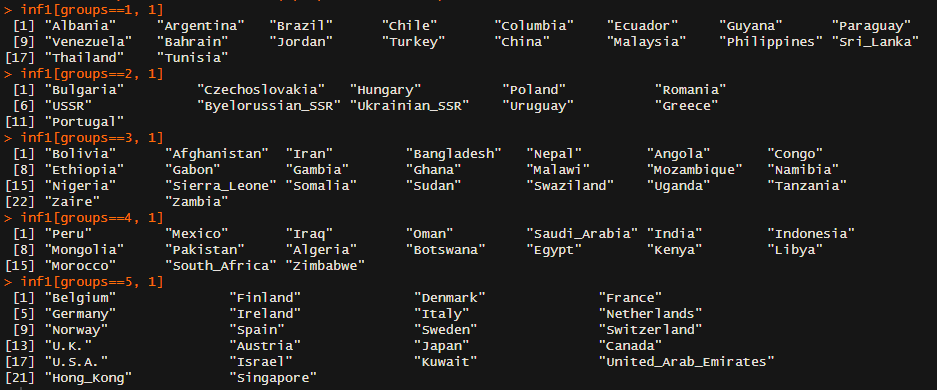


Рисунок 8 – Города, разбитые по кластерам

Далее представим распределения характеристик по всем кластерам.

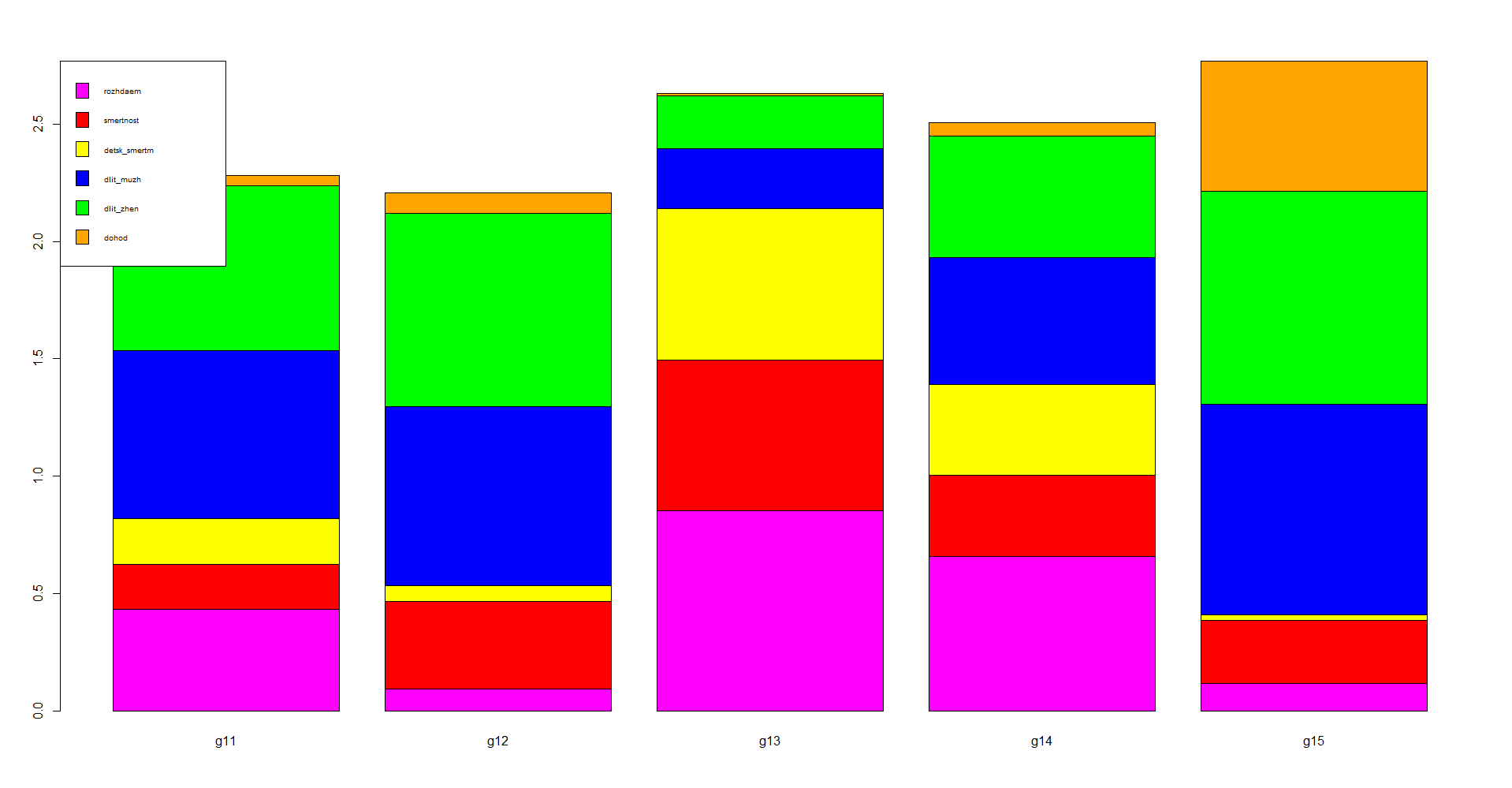


Рисунок 9 – Распределение характеристик в каждом кластере.

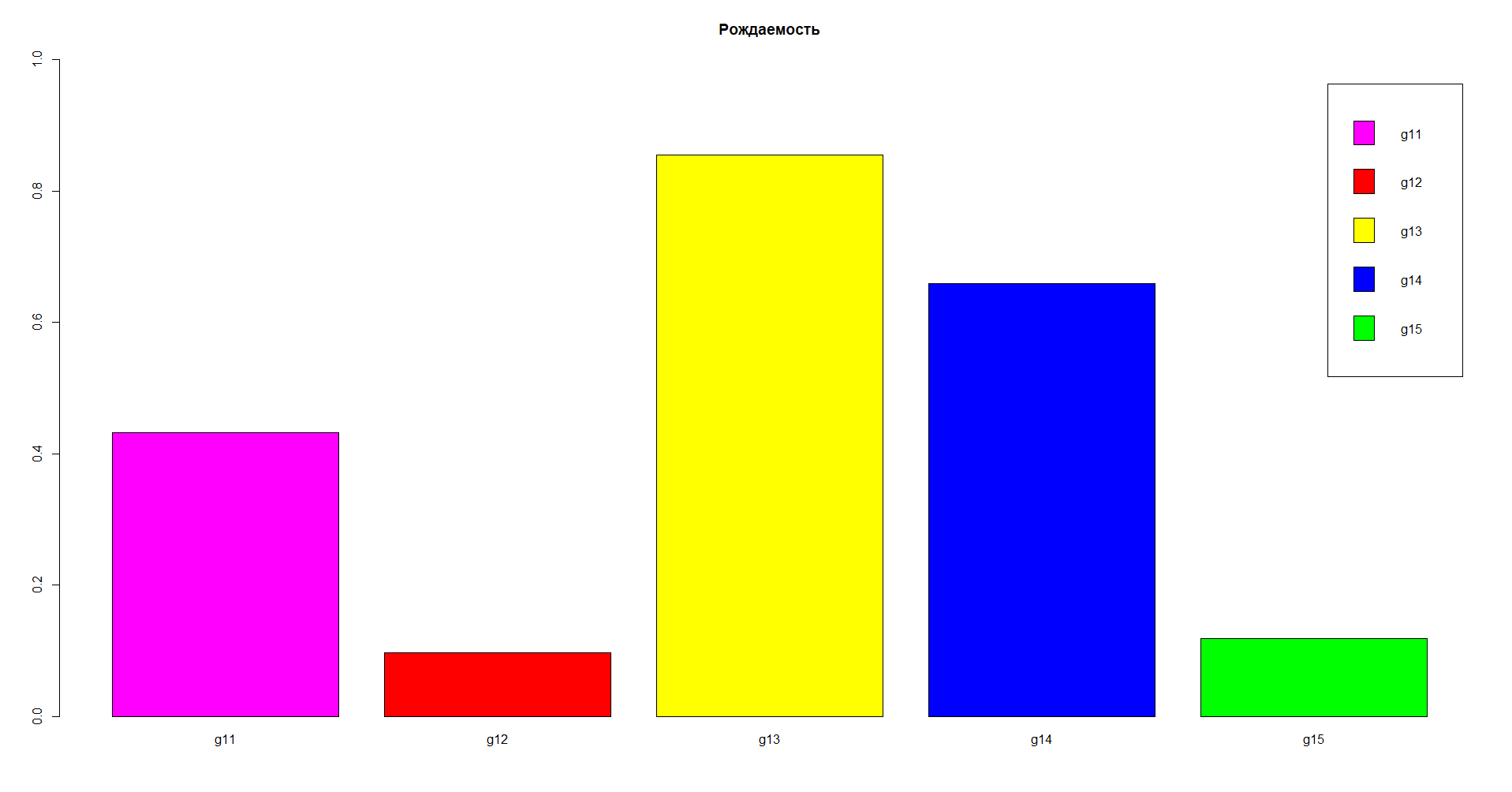


Рисунок 10 – Распределение характеристики Рождаемость.

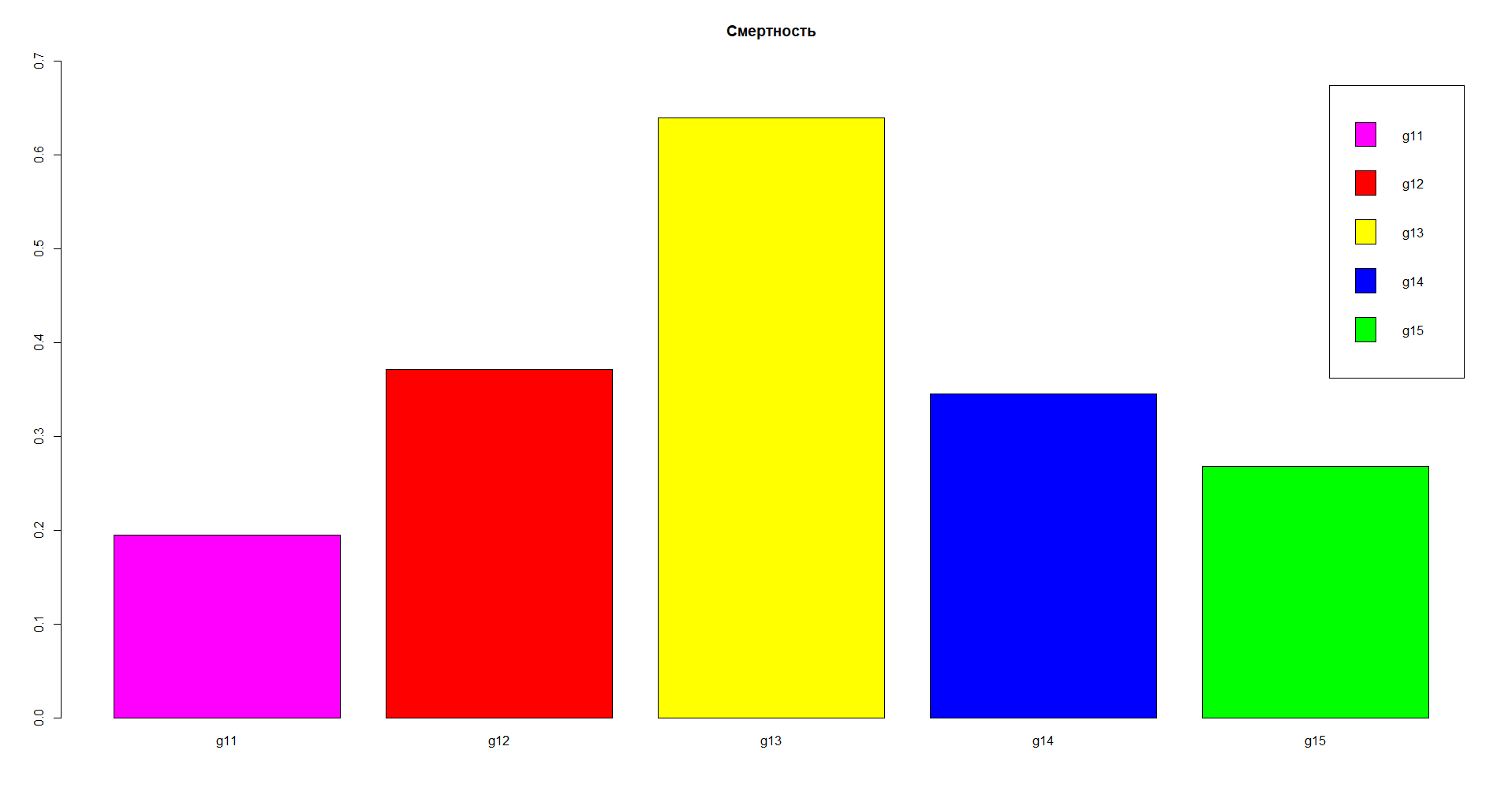


Рисунок 11 – Распределение характеристики Смертность.

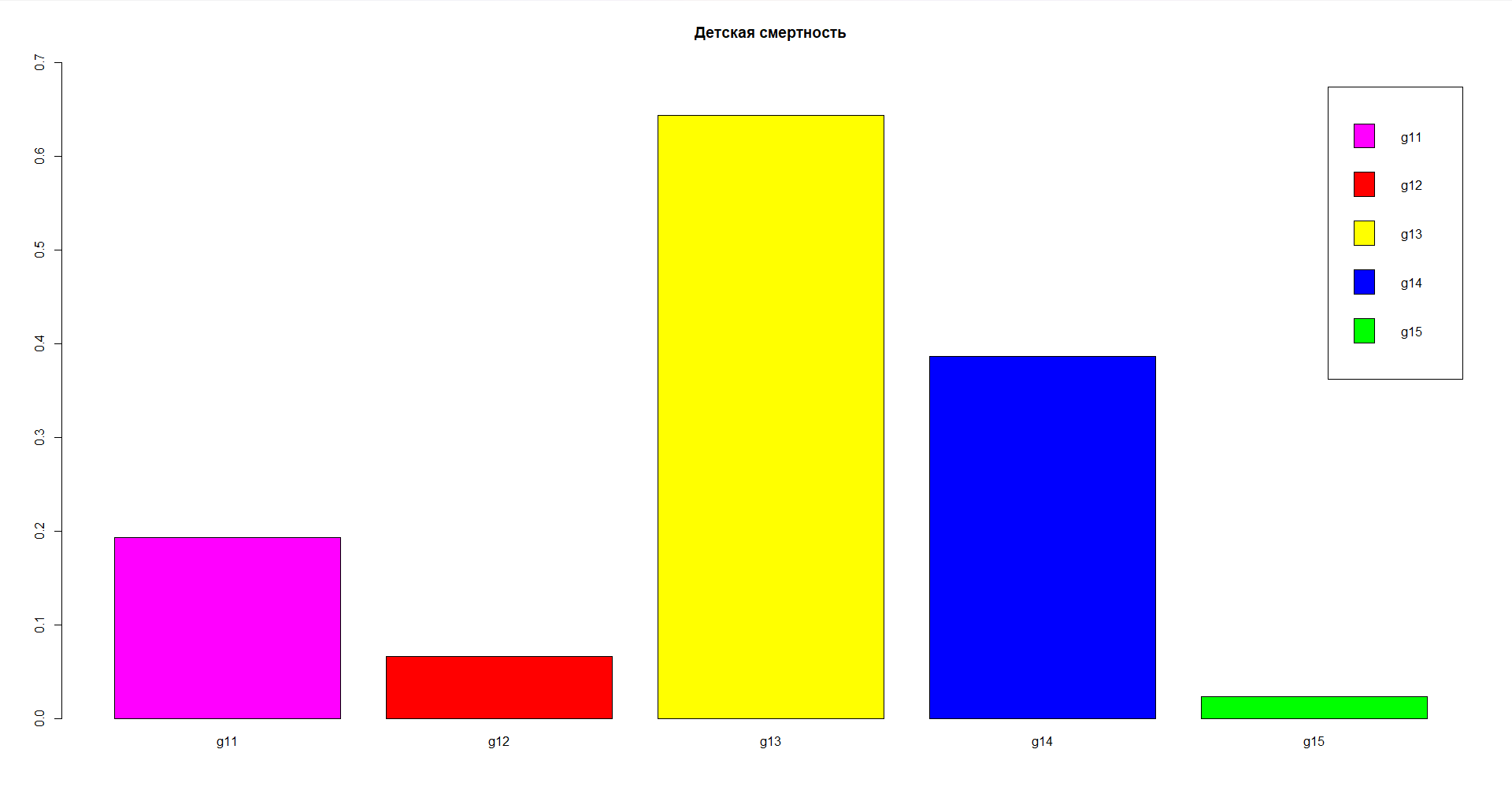


Рисунок 12 – Распределение характеристики Детская смертность.

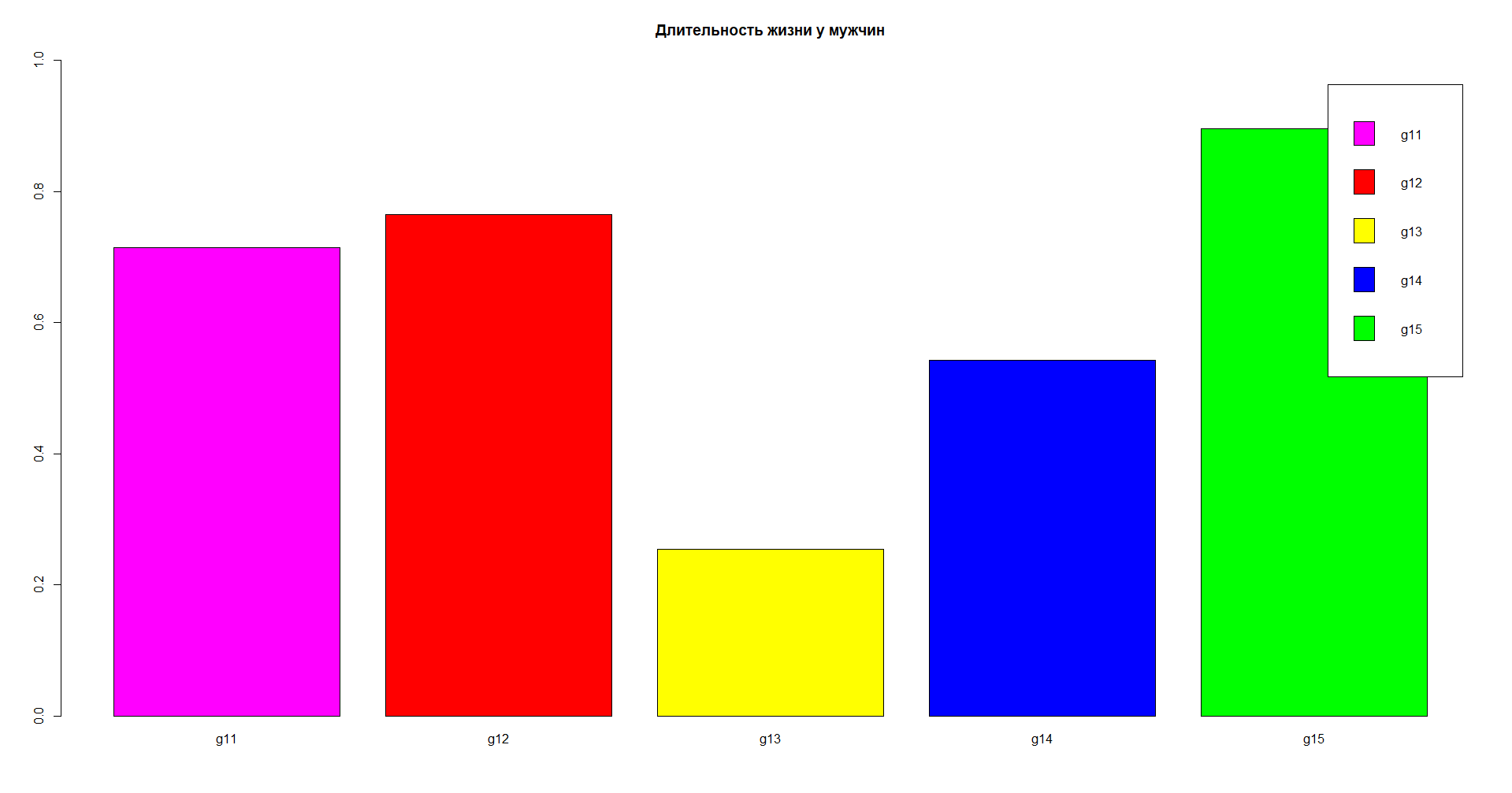


Рисунок 13 – Распределение характеристики Длительность жизни у мужчин.

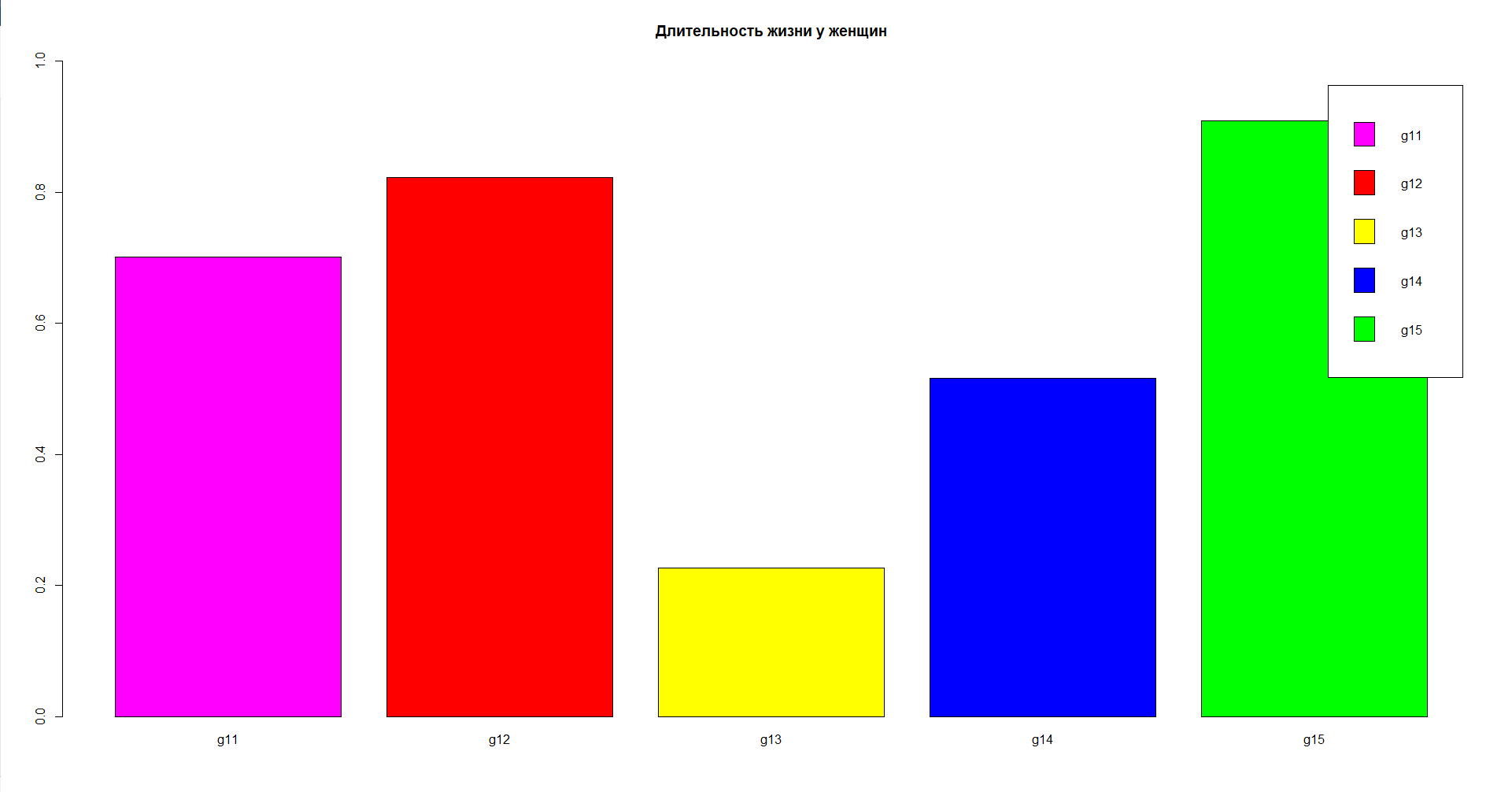


Рисунок 14 – Распределение характеристики Длительность жизни у женщин.

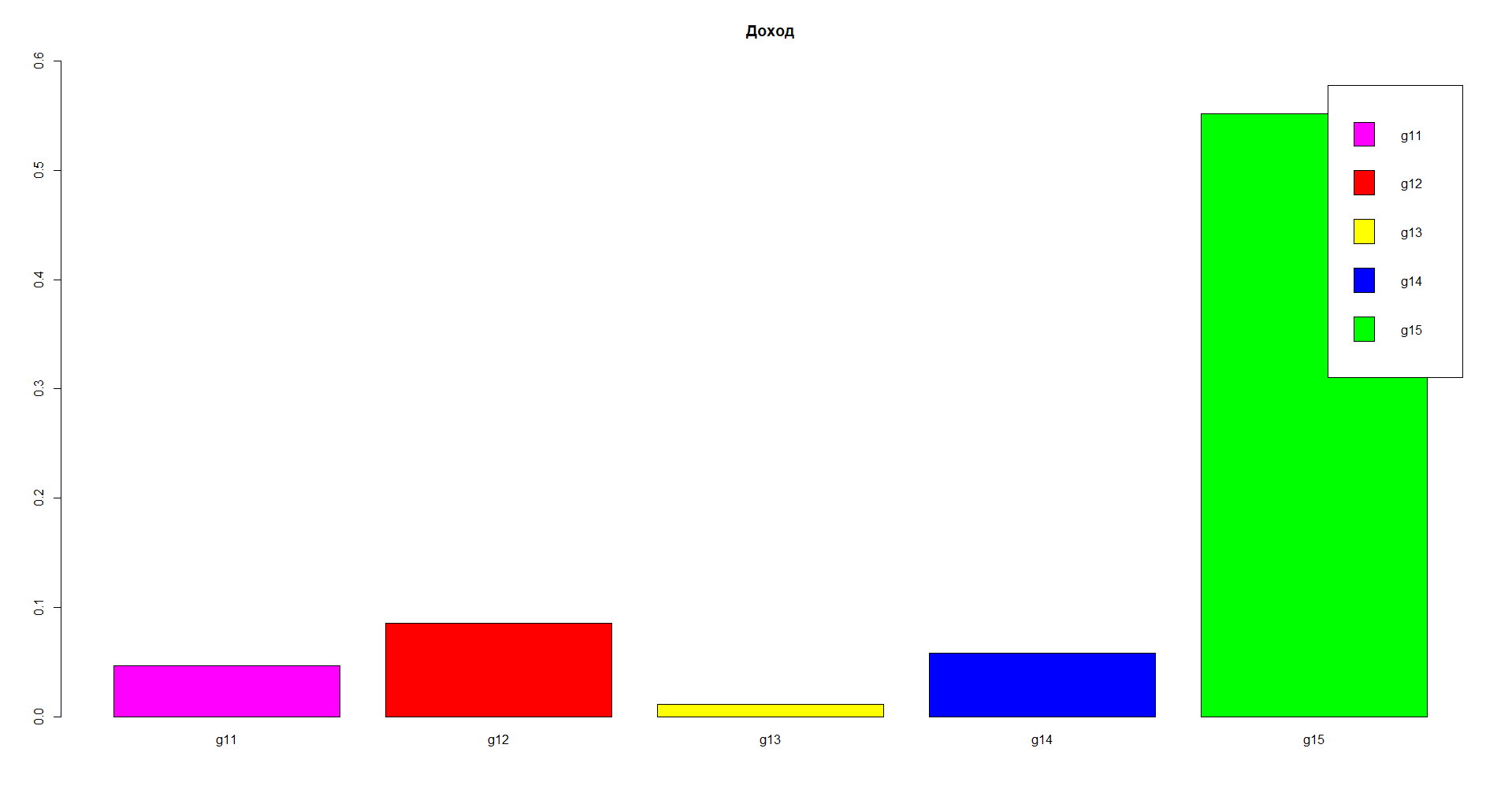


Рисунок 15 – Распределение характеристики Доход.

Подводя итоге по характеристикам по всем кластерам, можно увидеть, что кластер g13 по всем пунктам намного проигрывать другим, что ясно показывает не высокий уровень жизни среди населения стран этого кластера.

На рисунке 16 изображена диаграмма рассеивания, выражающая зависимость между индексами Смертности и Рождаемости.

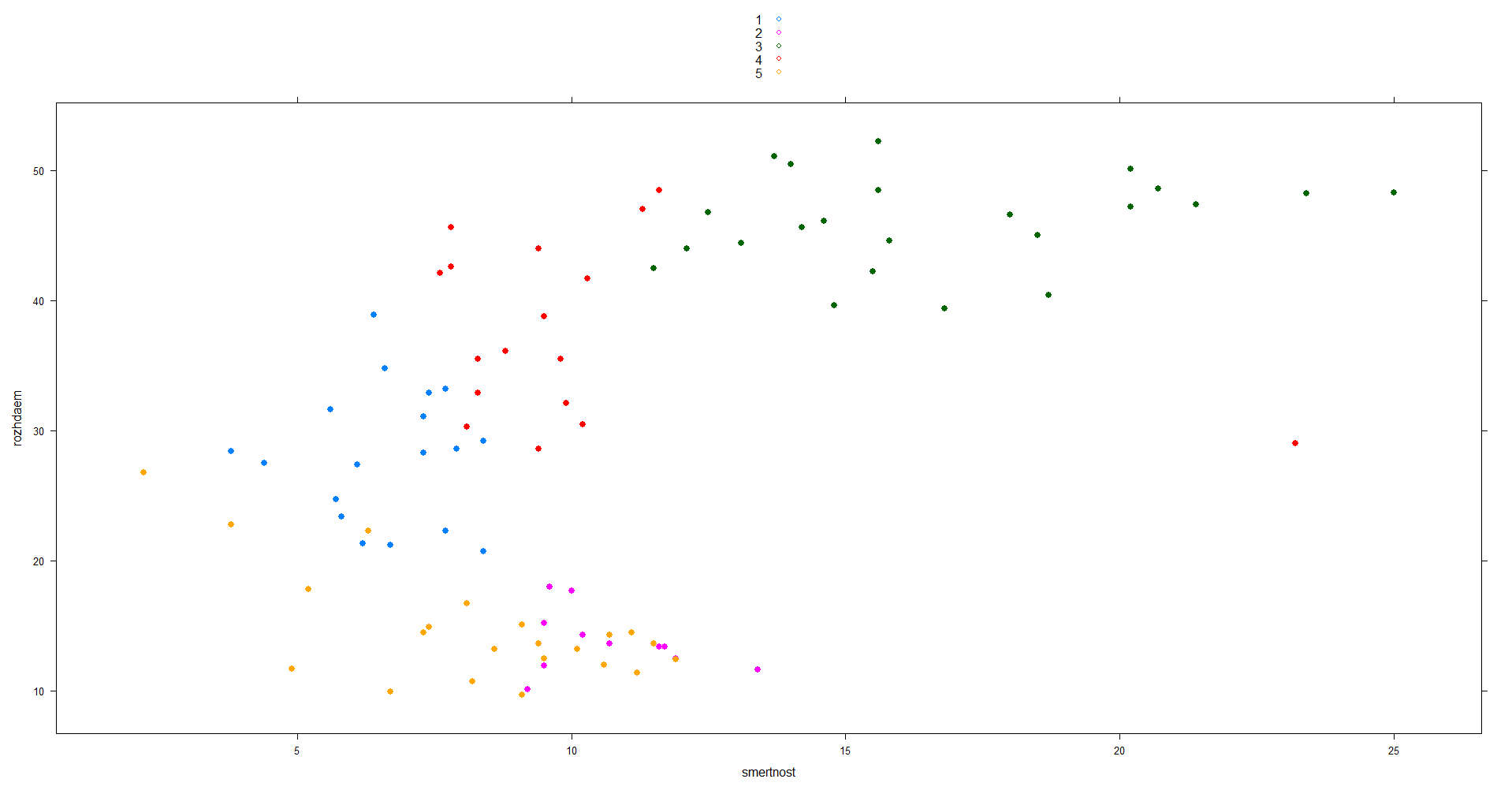


Рисунок 16 – Диаграмма рассеивания для рождаемости и смертности.

После были построены «ящики с усами», показанные на рисунке 17.

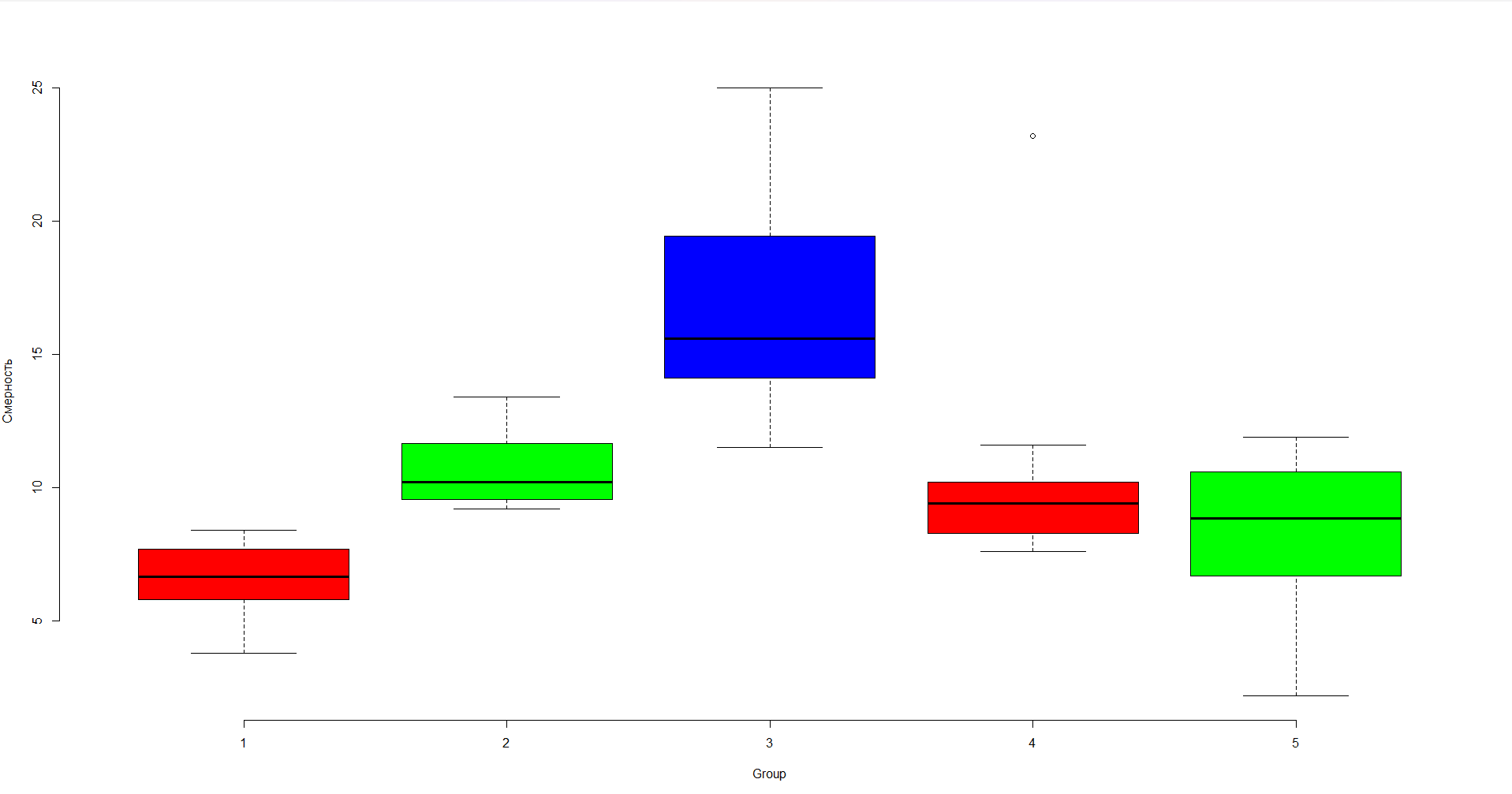


Рисунок 17 – Диаграмма ящик с усами.

На нем мы можем видеть, как распределяются индексы смертности по кластерам. Можно заметить, что в третьем кластере этот индекс превышает все остальные.

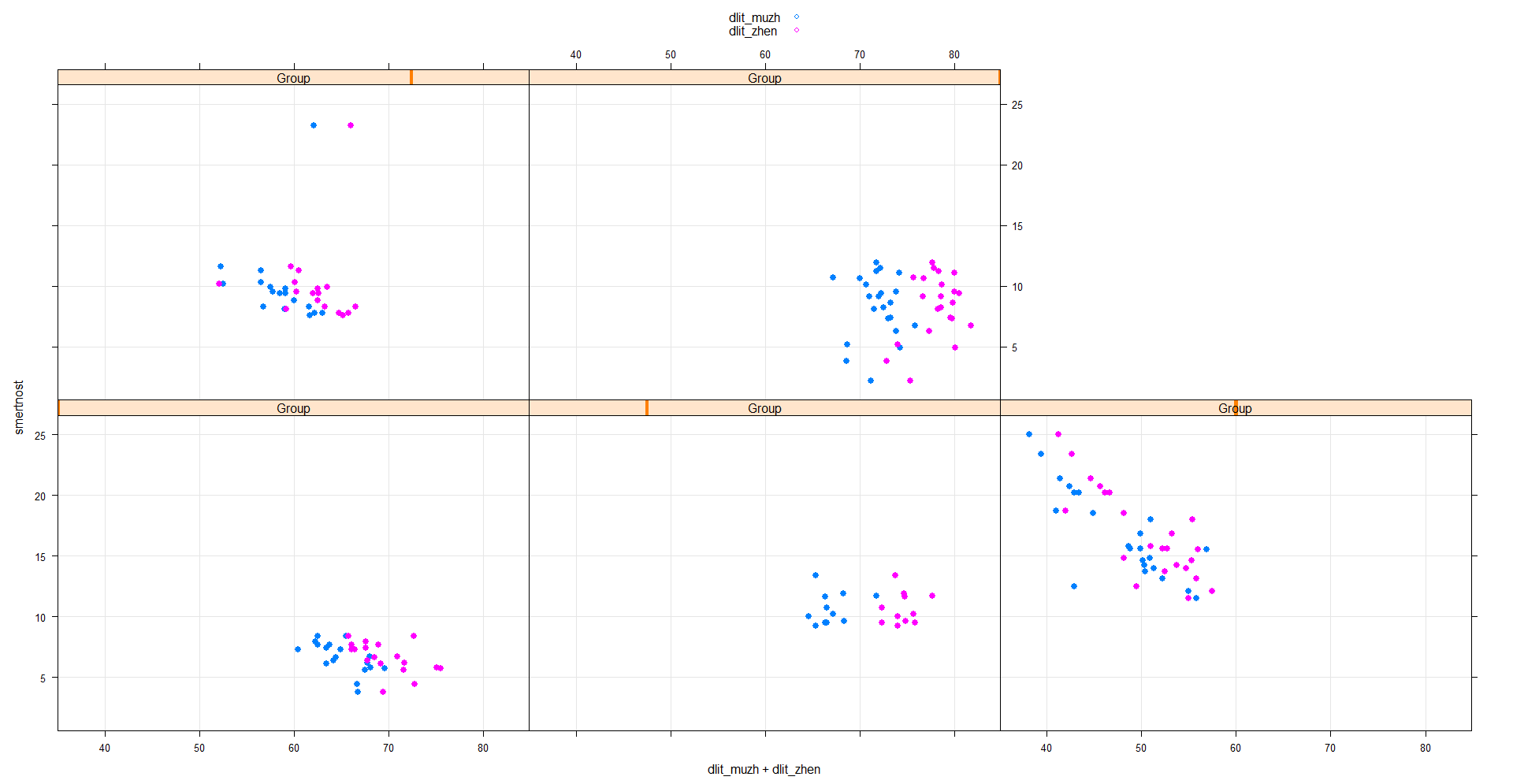


Рисунок 18 – График, показывающий распределение смертности и длительности мужской и женской жизни.

Можно сделать вывод, что прослеживается закономерность, чем короче длительность жизни мужчин и женщин, тем выше индекс смертности.

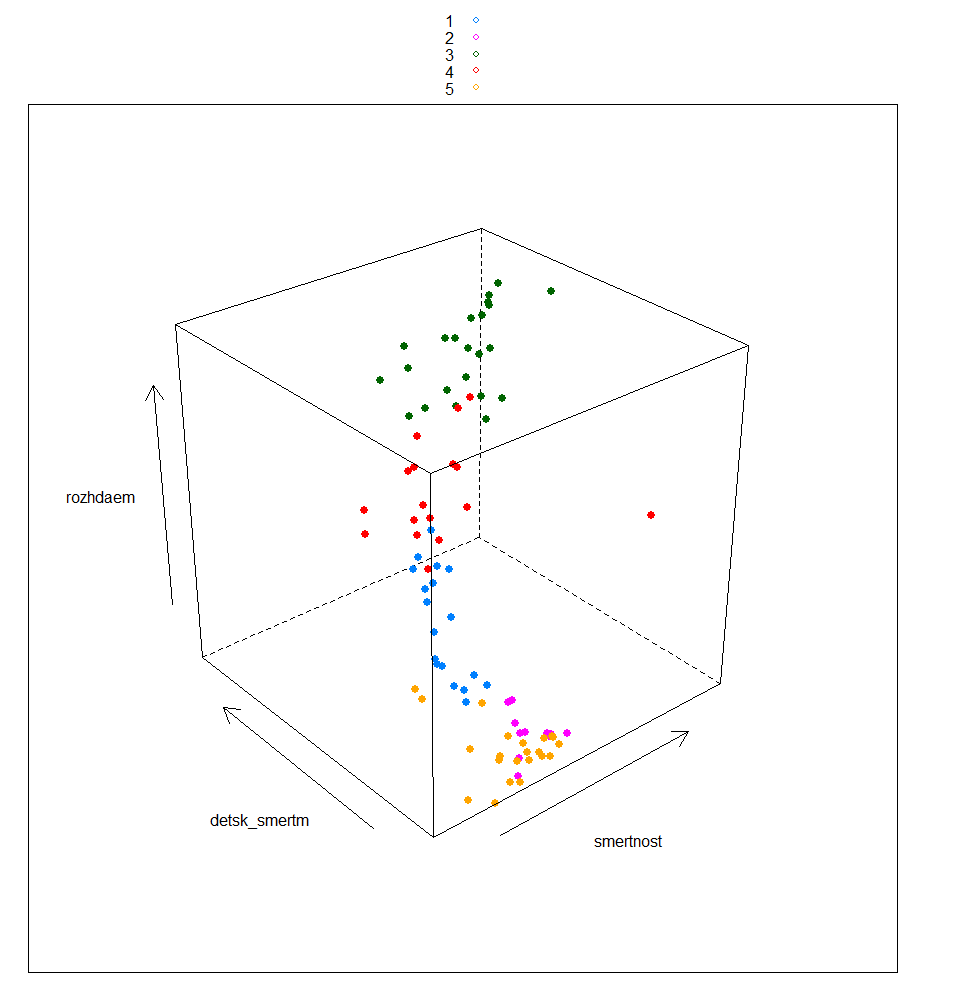


Рисунок 19 – Трехмерный график

Данный график классифицирует страны по индексу смертности, детской смертности и рождаемости.

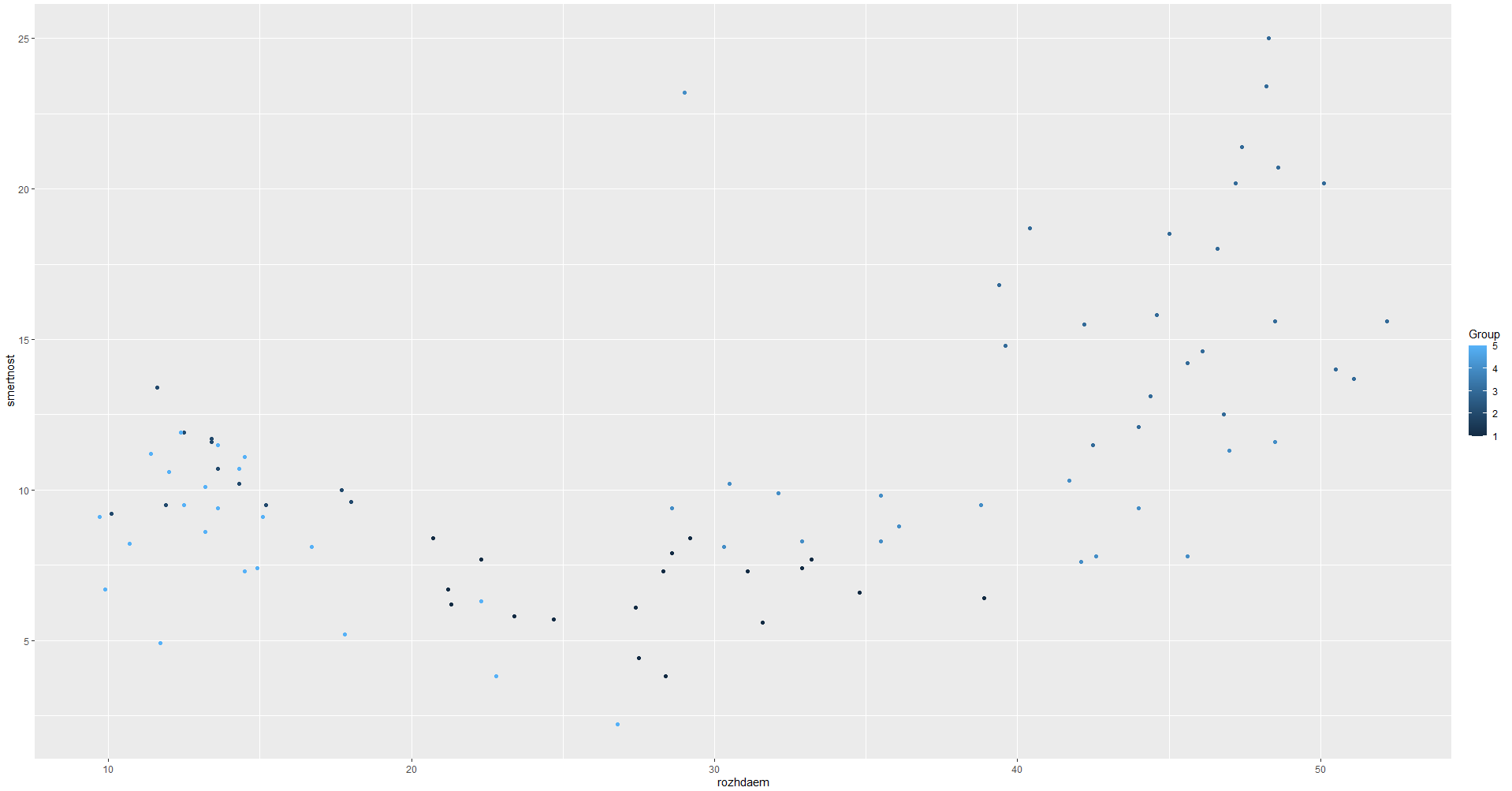


Рисунок 20 – Зависимость рождаемости от смертности в кластерах.

На данном шаге была произведена кластеризация k-means. Данная кластеризация позволяет отследить взаимосвязь двух непрерывных переменных. В данном случае – рождаемости и смертности в определенных кластерах.

**Вывод**: были закреплены знания об алгоритмах классификации и кластеризации данных, ознакомиться с некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Научился визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты и выполнять классификацию на основе формулы Байеса и деревьев решений.