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**

**Дисциплина: Обработка больших данных**

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

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Т.А.Приходько

**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.

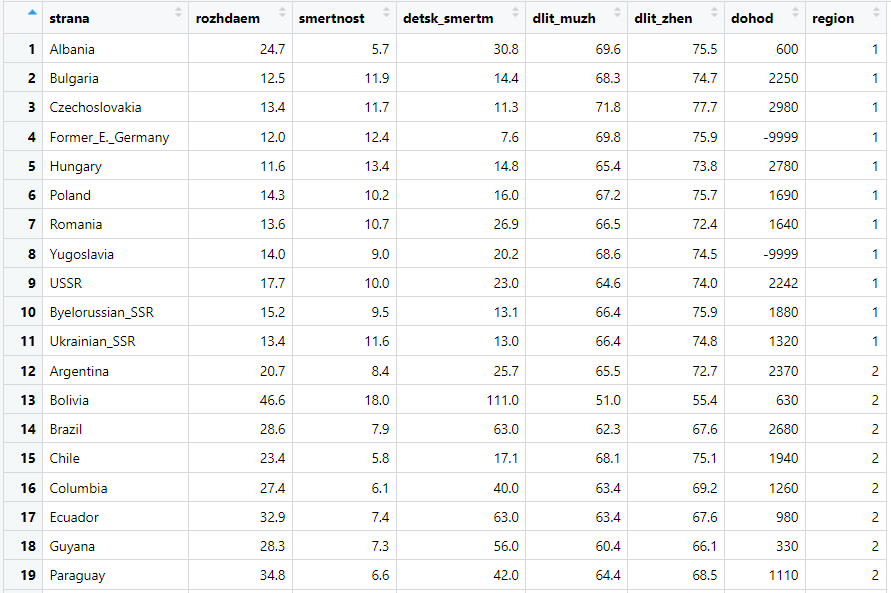


Рисунок 1 – Фрагмент исходной таблицы с данными

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.

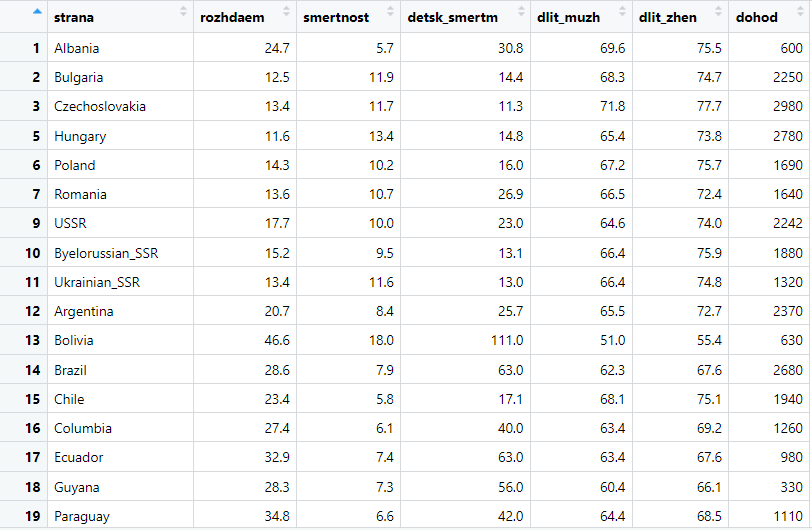


Рисунок 2 – Удалены пропущенные значения.

В данной задаче переменные существенно различны, поэтому необходимо выполнить нормализацию данных. Для стандартизации исключим колонку страна.

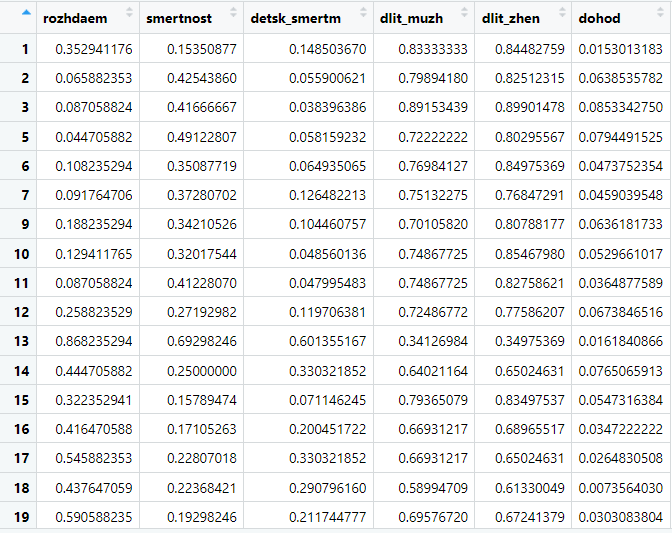


Рисунок 3 – Фрагмент стандартизированной таблицы с данными

По полученным данным построим дендрограмму.

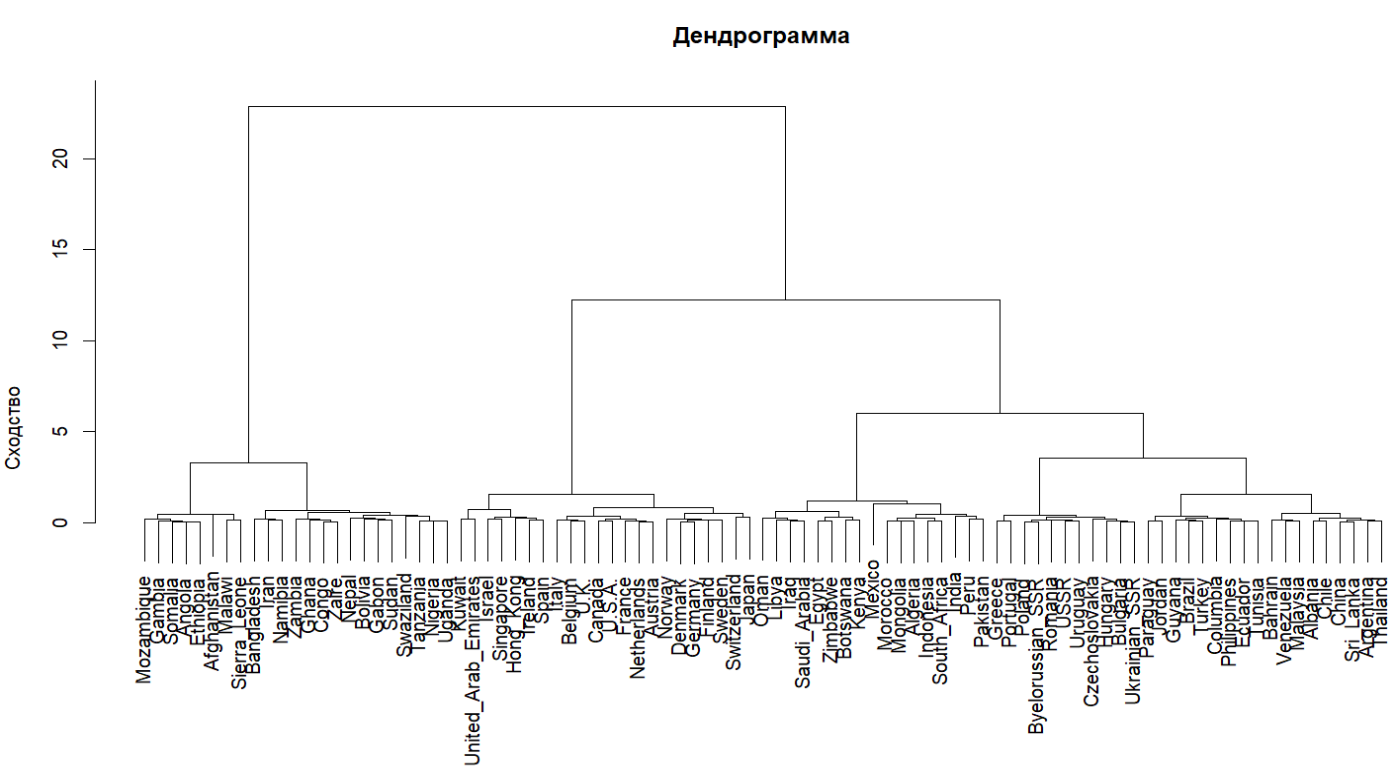


Рисунок 4 – Дендограмма, построенная по данным

Целесообразнее выделить 5 кластеров, выделим их красным цветом.

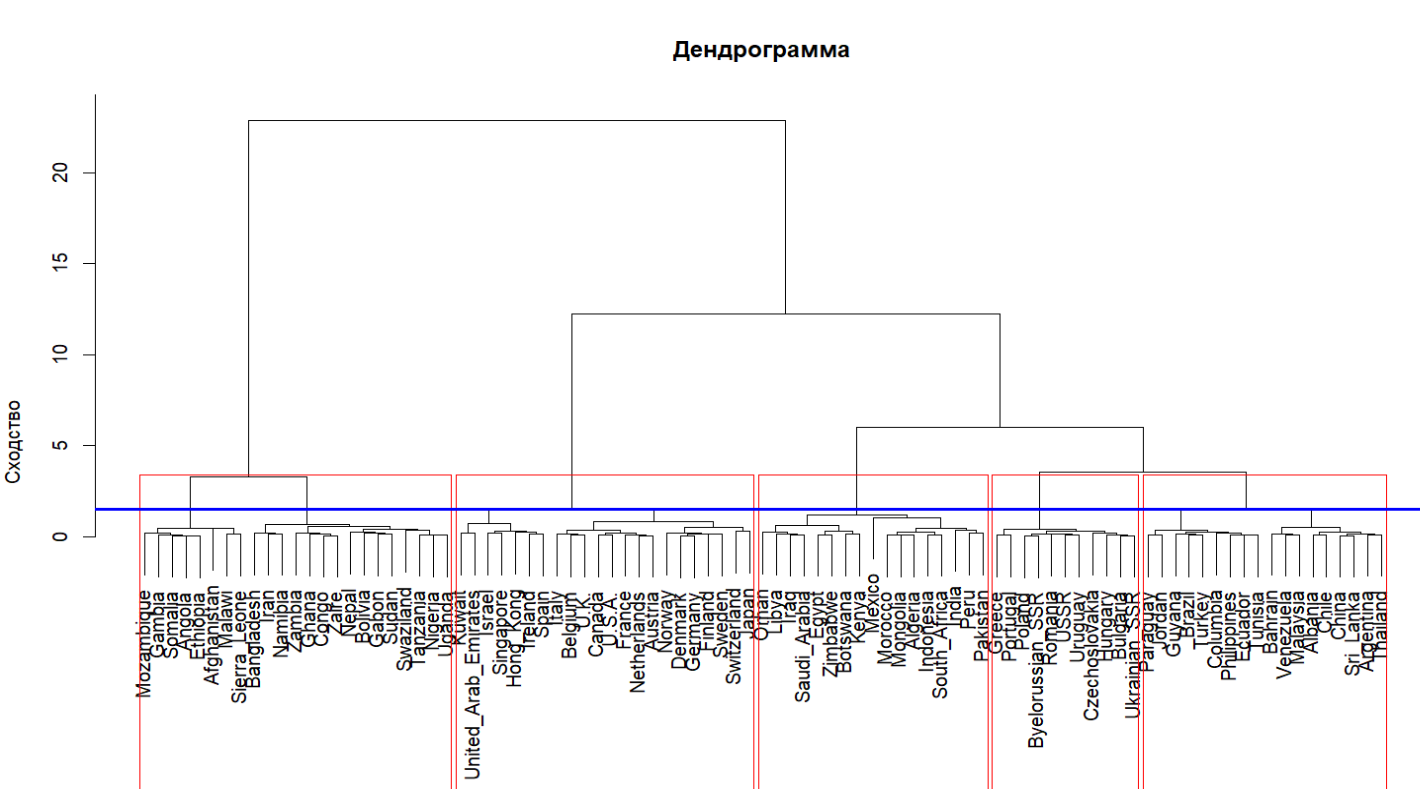


Рисунок 5 – Дендограмма с кластерами.

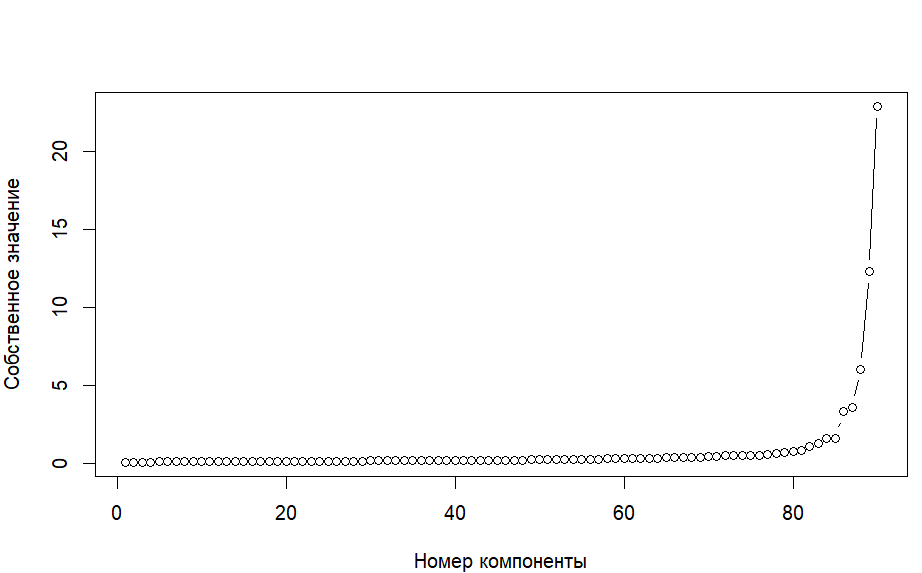


Рисунок 6 – Диаграмма «Каменная осыпь».

Далее представим списки городов, разбитых по кластерам.



Рисунок 7 – Города, разбитые по кластерам

Далее представим распределения характеристик по всем кластерам.

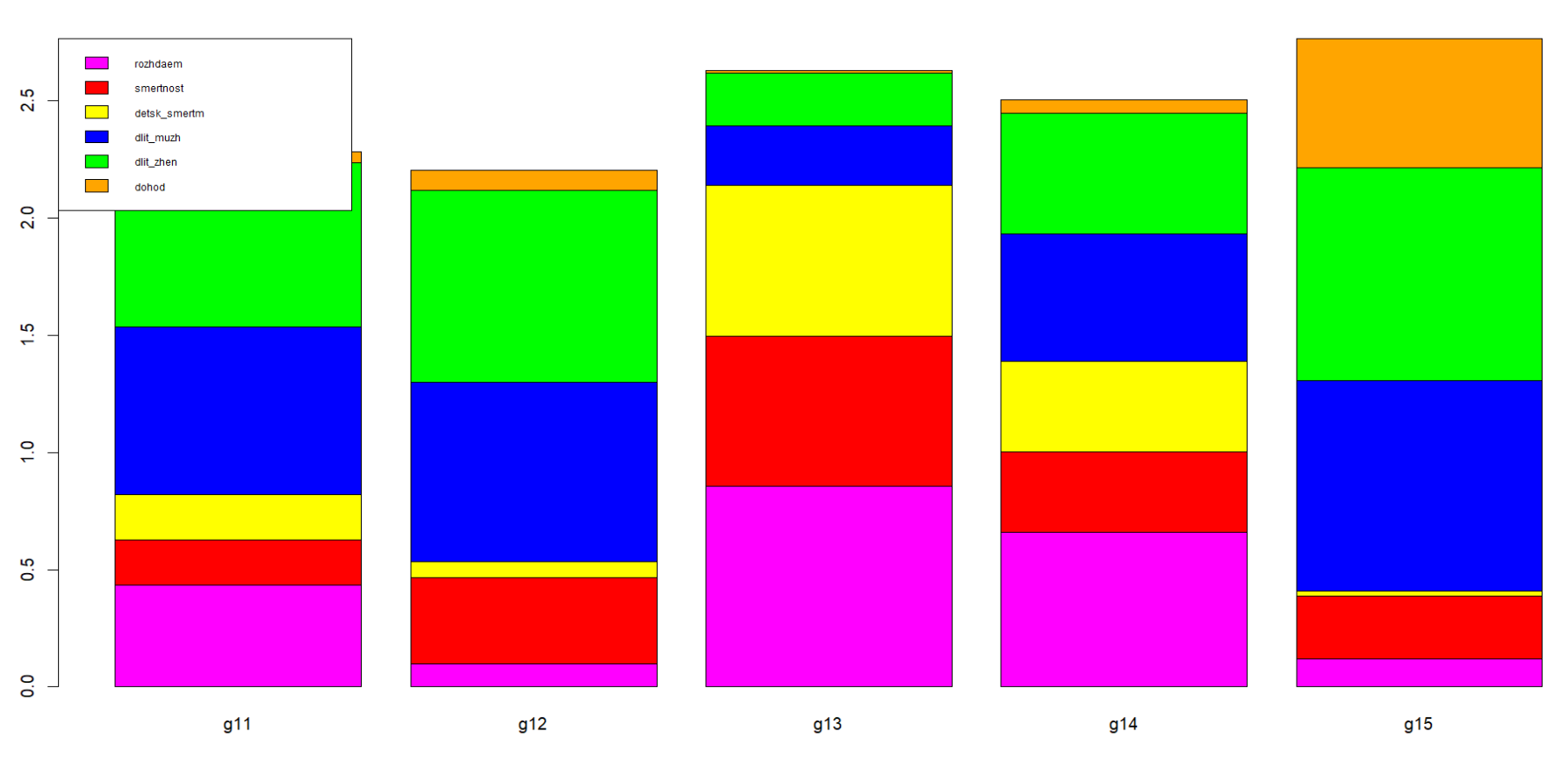


Рисунок 8 – Распределение характеристик в каждом кластере.

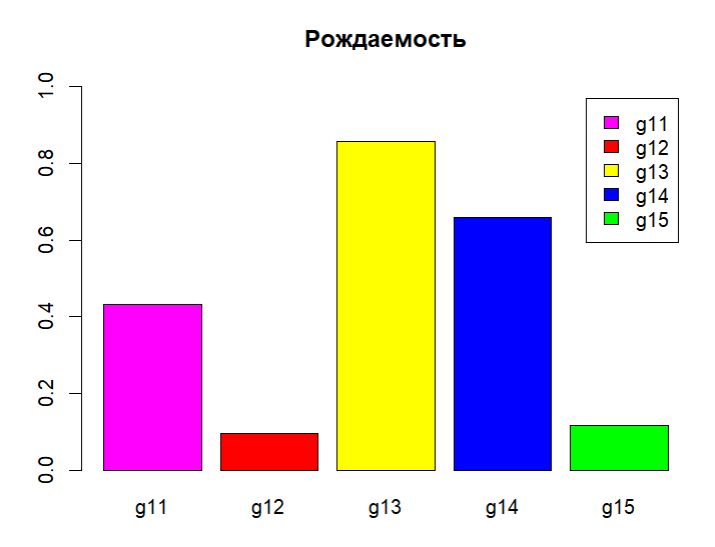


Рисунок 5 – Распределение характеристики Рождаемость.

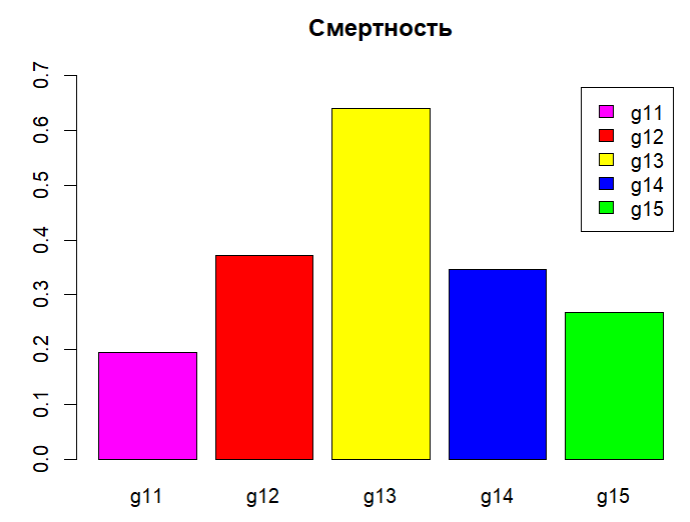


Рисунок 6 – Распределение характеристики Смертность.

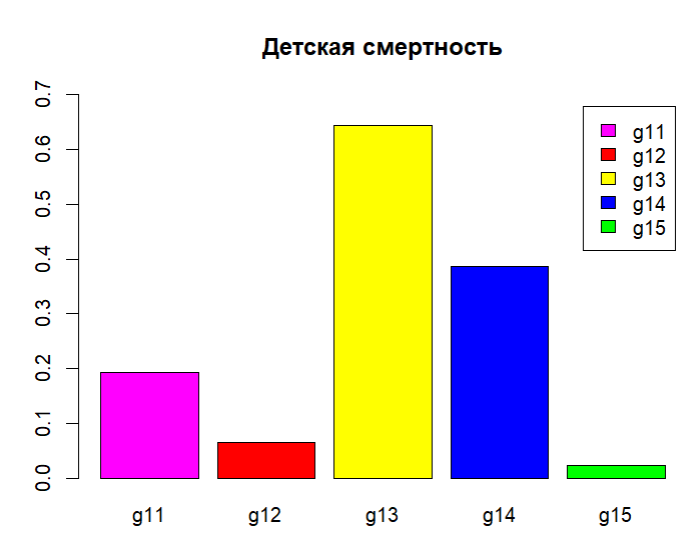


Рисунок 7 – Распределение характеристики Детская смертность.

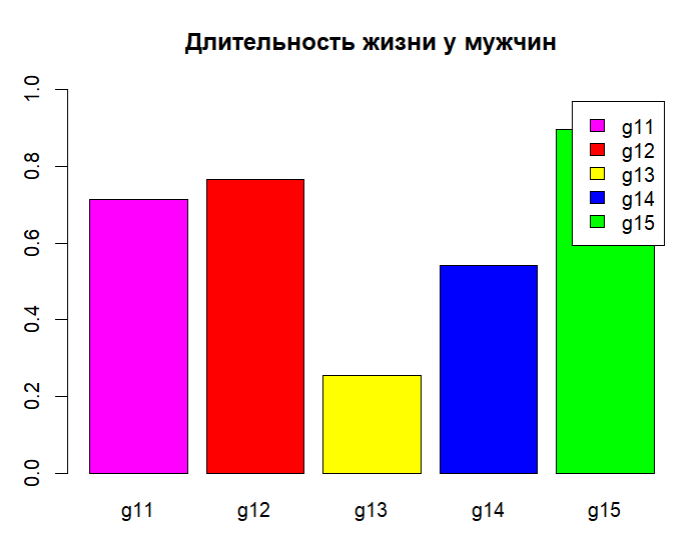


Рисунок 8 – Распределение характеристики Длительность жизни у мужчин.

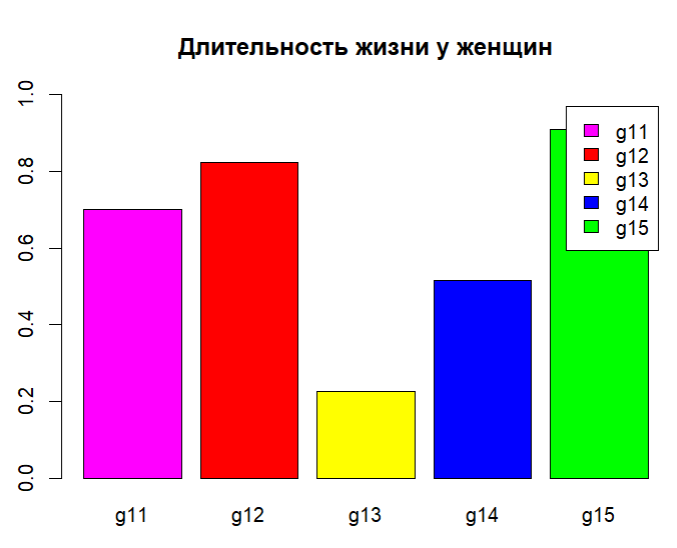


Рисунок 9 – Распределение характеристики Длительность жизни у женщин.

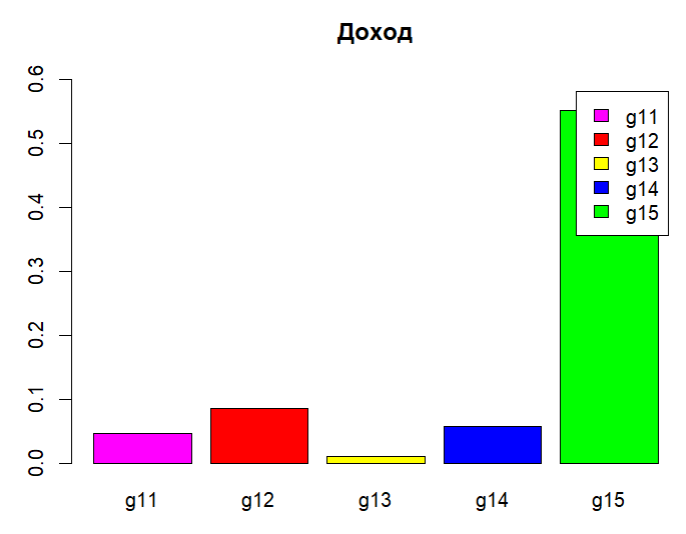


Рисунок 10 – Распределение характеристики Доход.

Рождаемость- Количество рождений на 1,000 населения.

Смертность - Количество смертей на 1,000 населения.

Детская смертность - Детская смертность (до 1 года ) на 1,000 населения.

Длительность мужской жизни - Ожидаемая длительность жизни на момент рождения для мужчин.

Длительность женской жизни - Ожидаемая длительность жизни на момент рождения для женщин.

Доход - Совокупный национальный доход на душу населения в долларах США.

На рисунке 10 изображена диаграмма рассеивания, выражающая зависимость между индексами Смертности и Рождаемости.

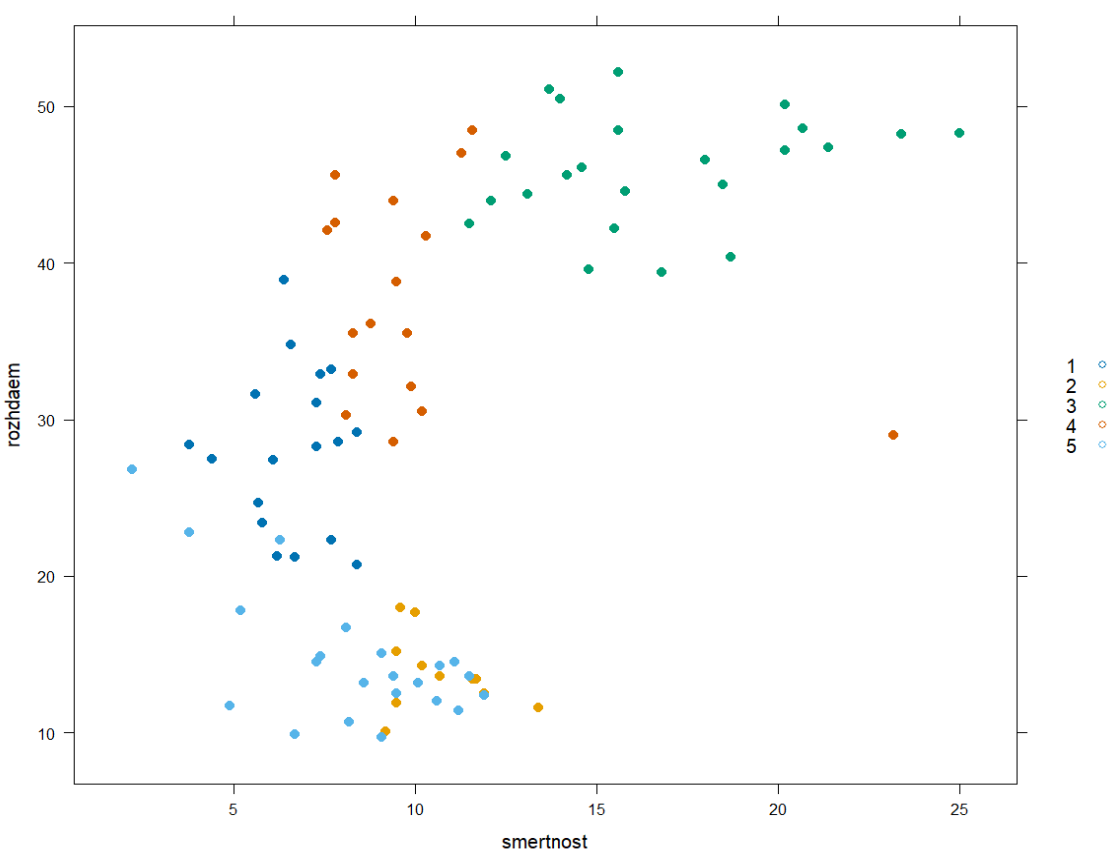


Рисунок 10 – Диаграмма рассеивания для рождаемости и смертности.

После были построены «ящики с усами», показанные на рисунке 11.

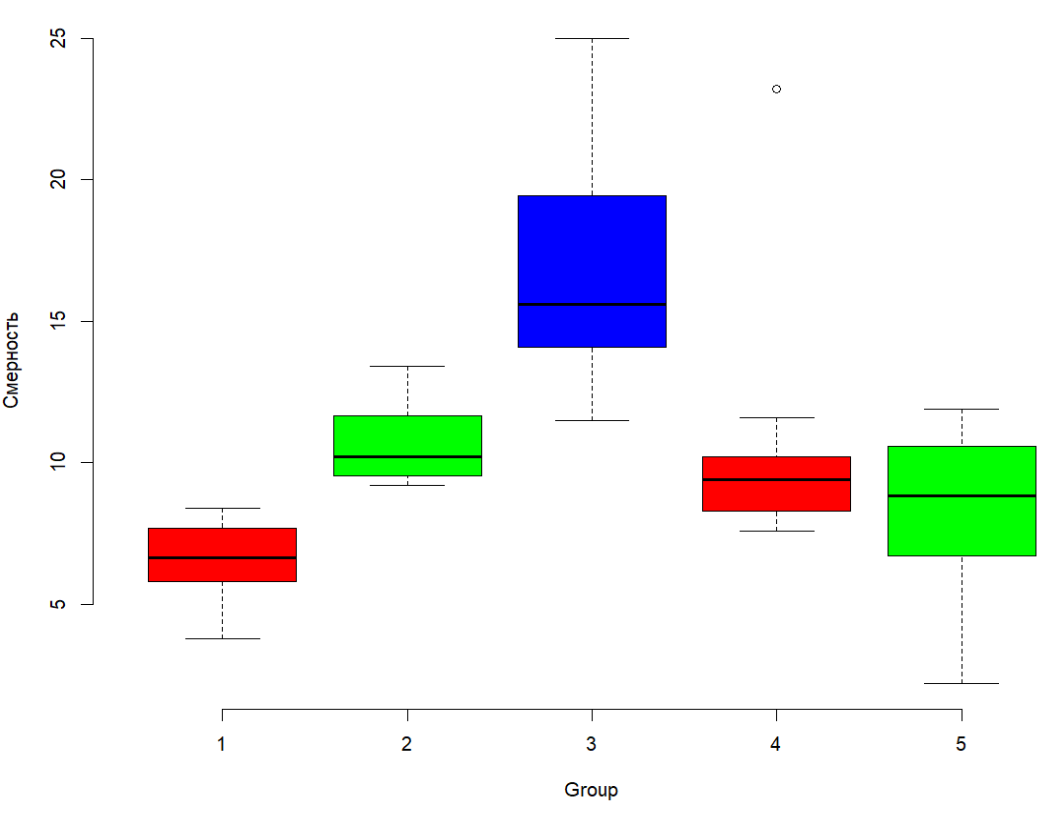


Рисунок 11 – Диаграмма ящик с усами.

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

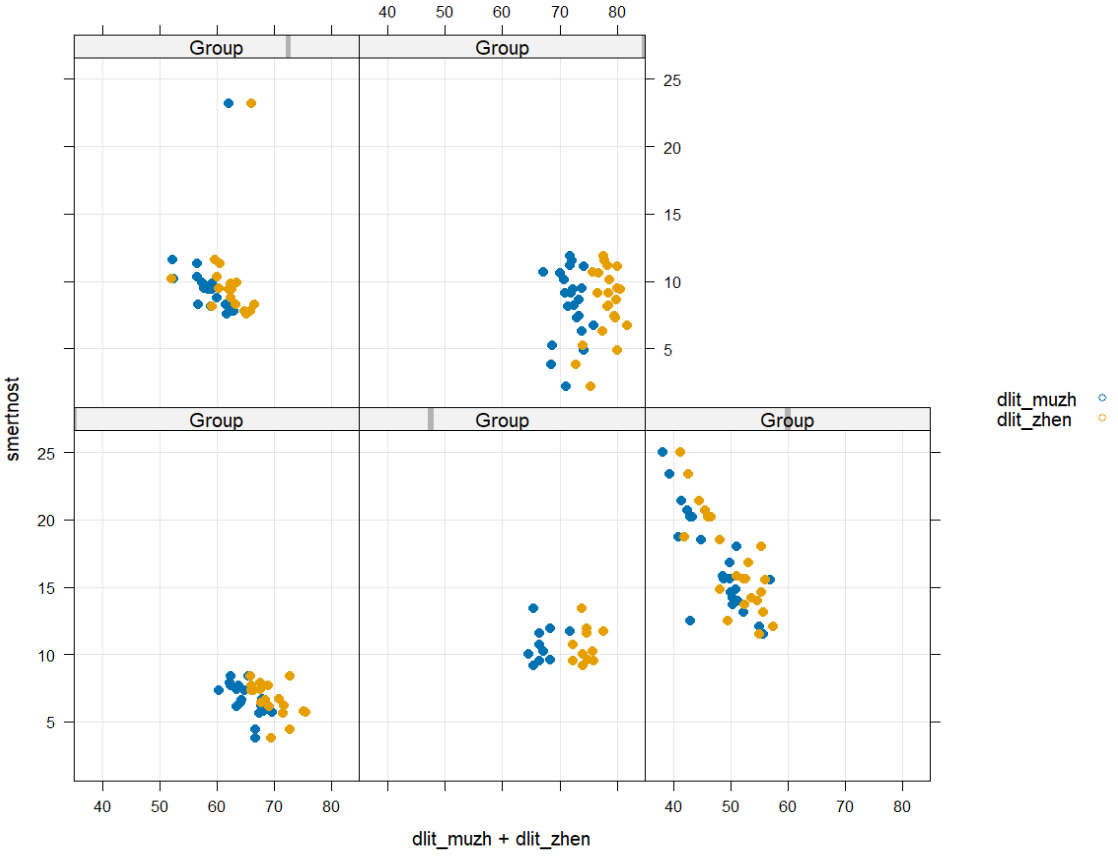


Рисунок 12 – График, показывающий распределение смертности и длительности мужской и женской жизни.

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

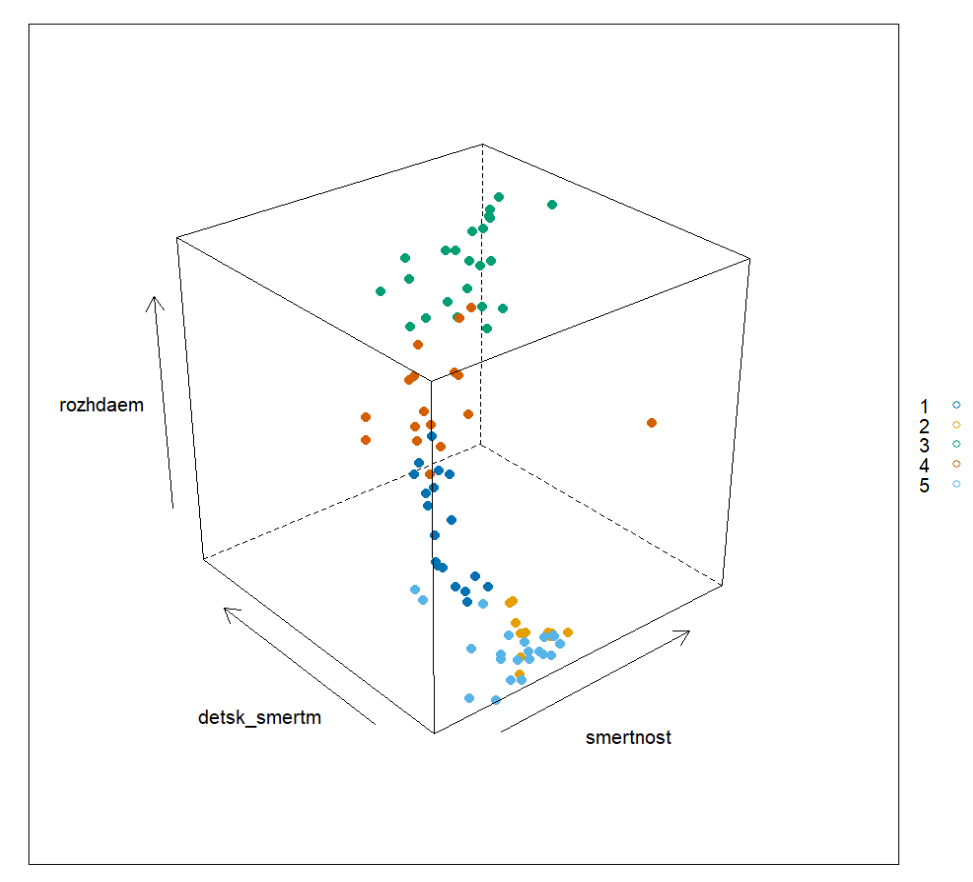


Рисунок 13 – Трехмерный график

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

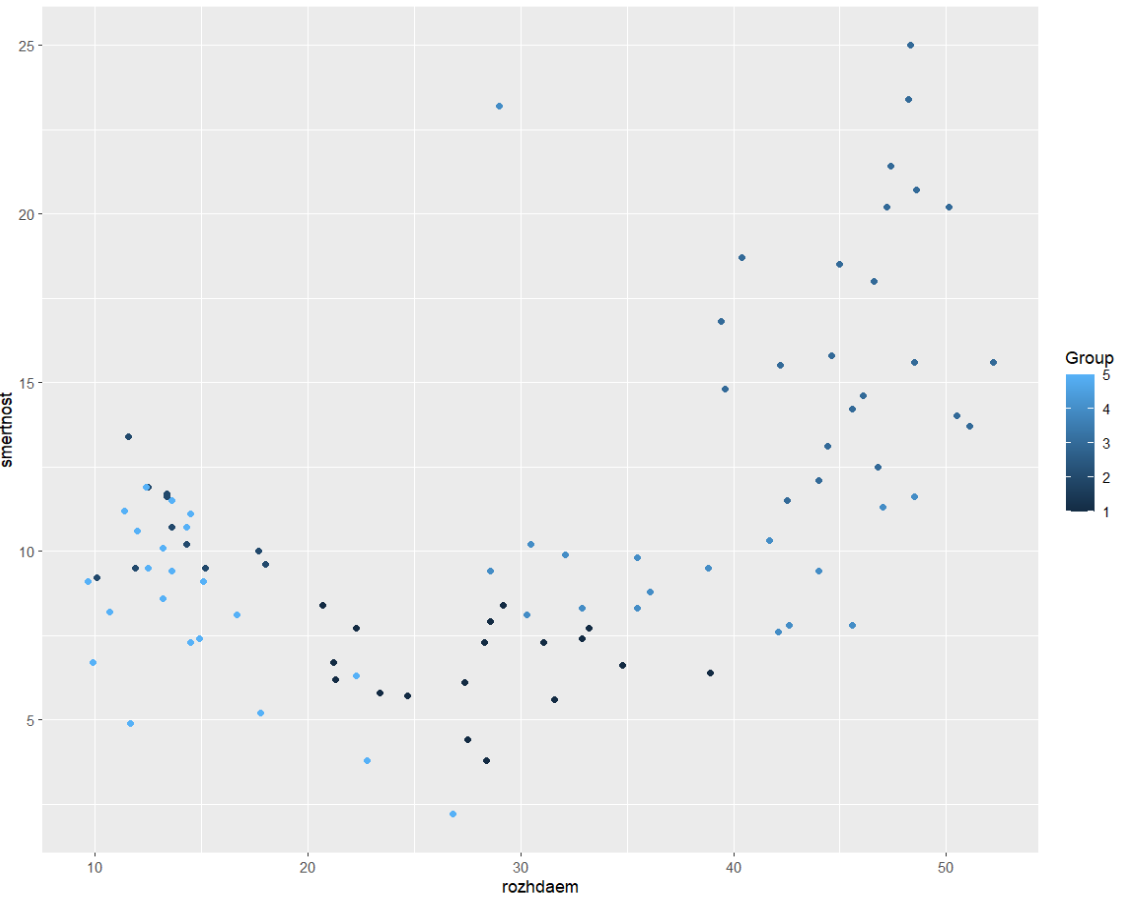


Рисунок 14 – Зависимость рождаемости от смертности в кластерах.

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.