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.Д. Воробьев

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. Также провести сравнение результатов кластеризации двух датасетов по одной тематике за различные периоды.

**Часть 1.** **Подготовка данных**

В качестве датасетов были выбраны уровни жизни стран за 2024 и 2012 годы, взятые с сайта <https://www.numbeo.com/> путём веб-скреппинга и далее парсинга данных.

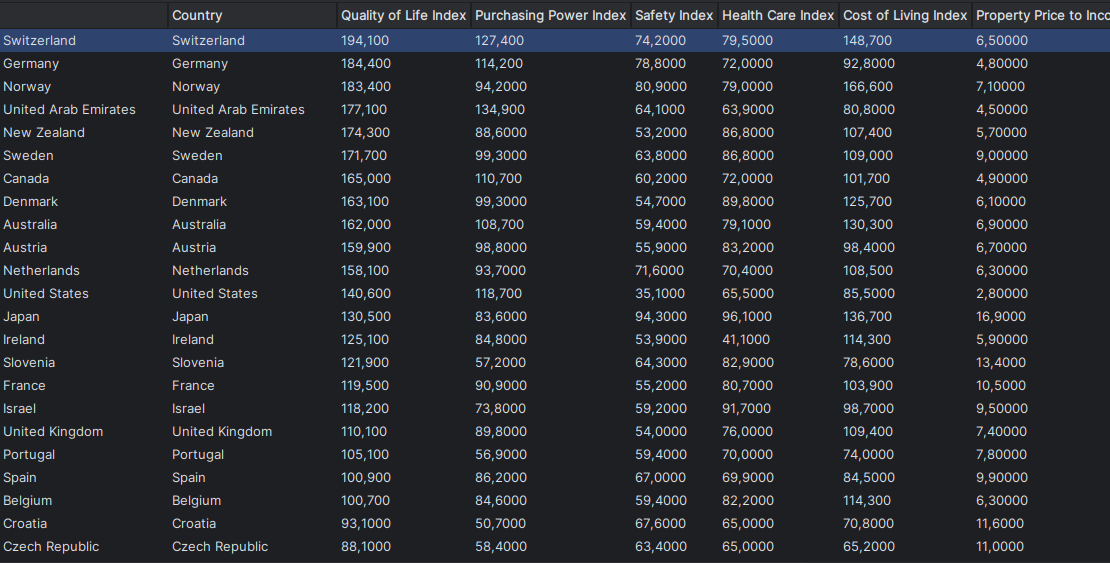


Рисунок 1 – Фрагмент исходной таблицы с данными за 2012 год

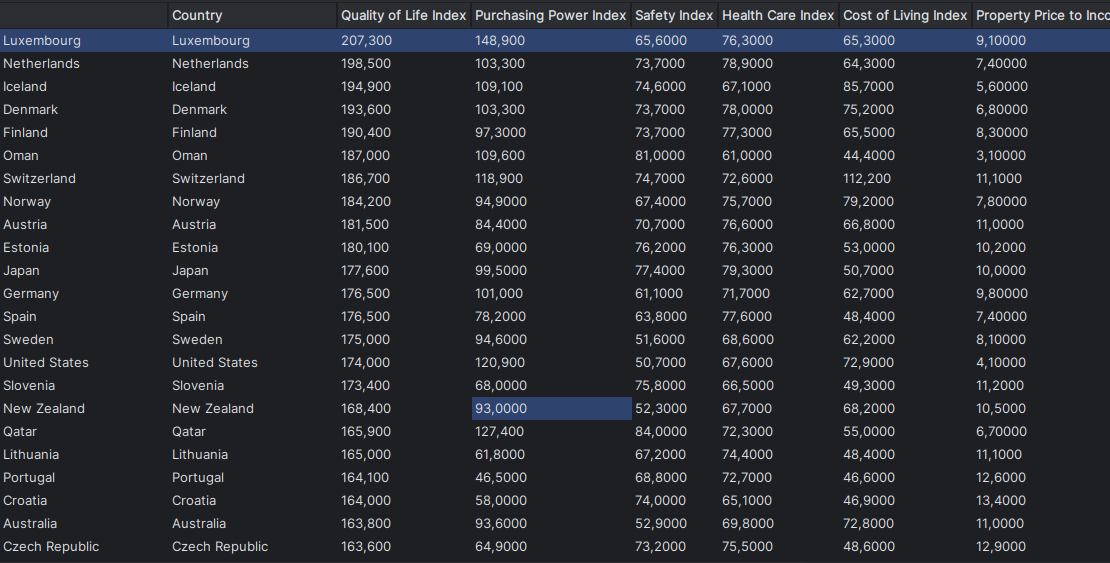
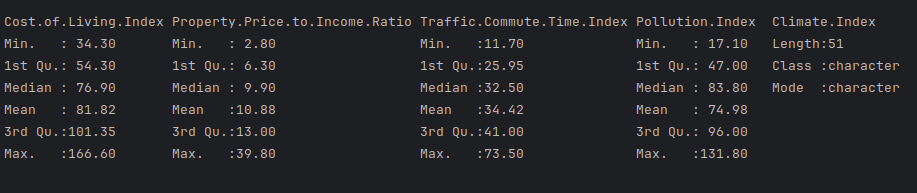
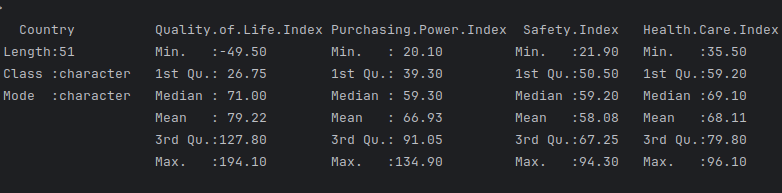
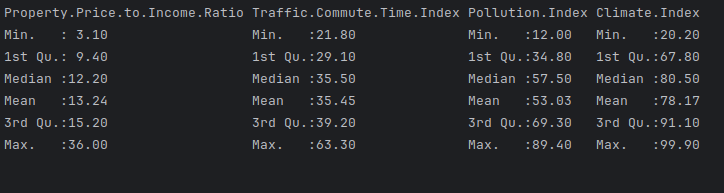
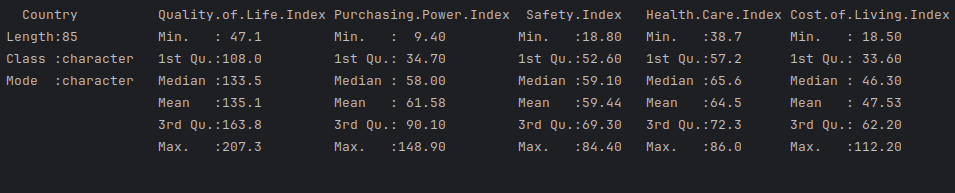


Рисунок 2 – Фрагмент исходной таблицы с данными за 2024 год

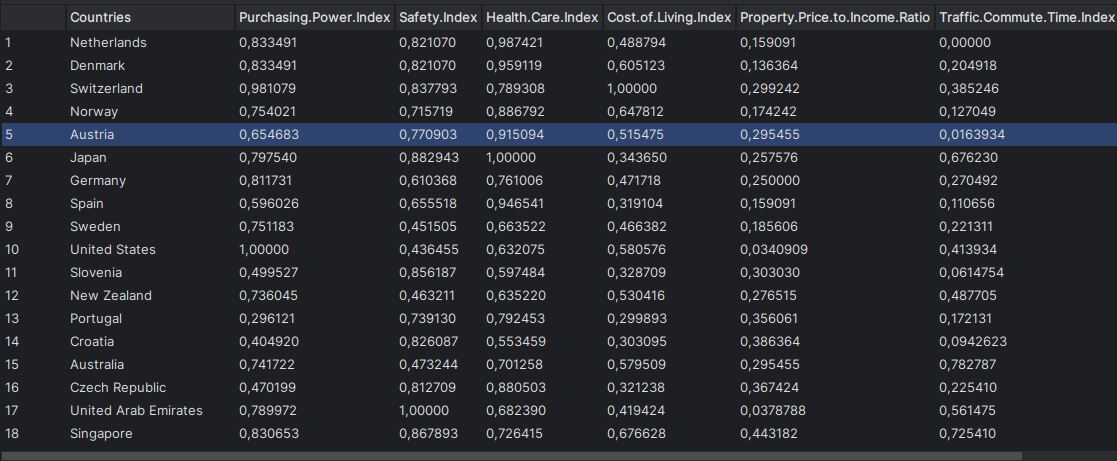
Далее был проведён стандартный описательный анализ для данных за 2012 и 2024 года

Рисунок 3 – Описательный анализ данных за 2012 год



Рисунок 4 – Описательный анализ данных за 2024 год

Исходя из результатов описательного анализа для датасетов очевидно несоответствие данных для сравнения ввиду разного количества стран и отсутствия данных по климатическому индексу в датасете 2012 года. Также необходимо провести нормализацию данных в обоих датасетах.

Рисунок 5 – Датасет за 2024 год с нормализованными данными

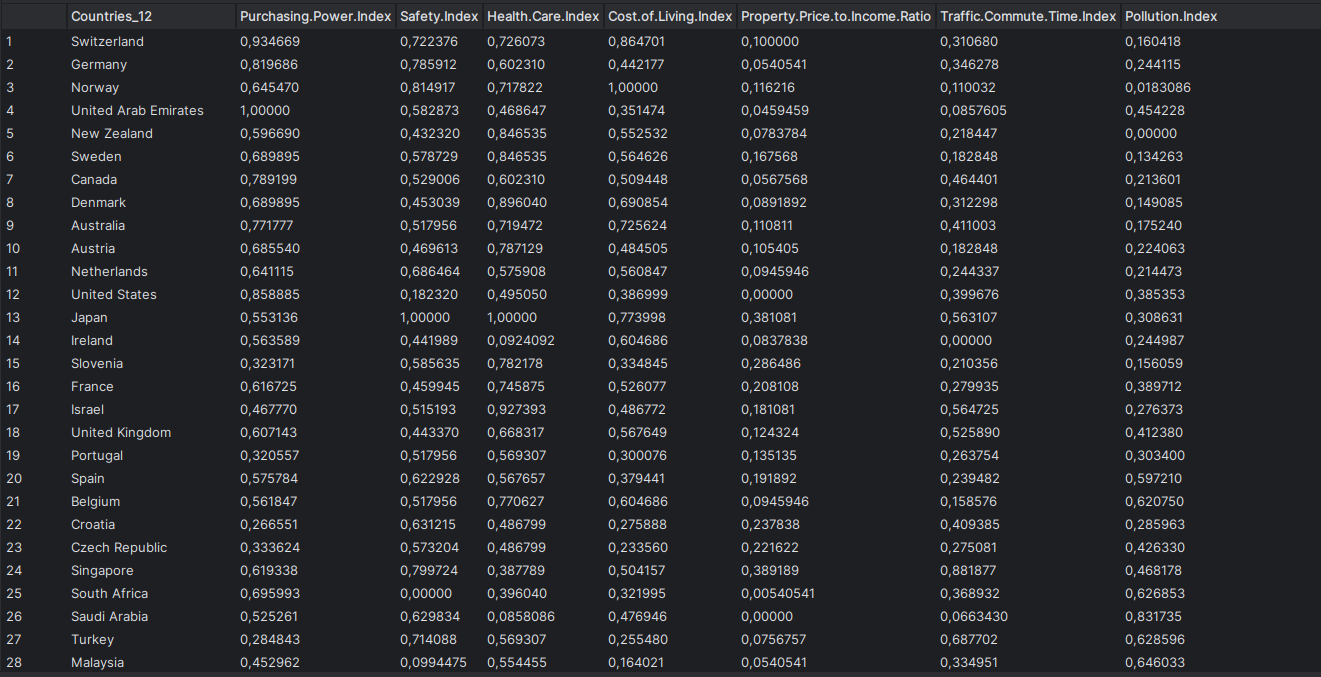
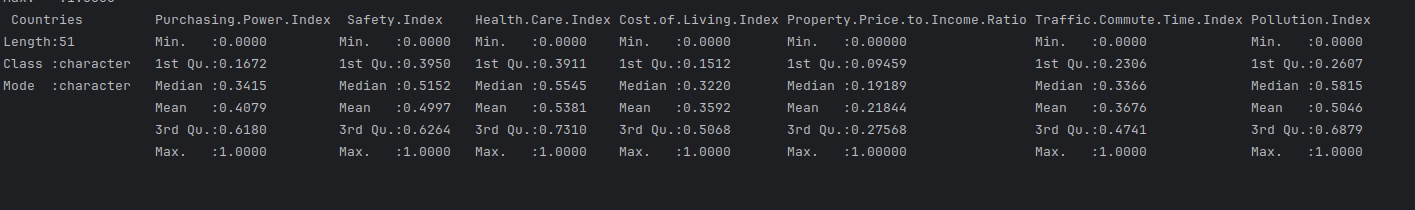
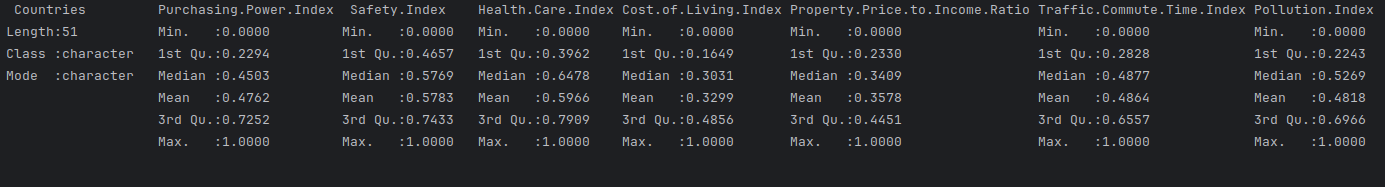


Рисунок 6 – Датасет за 2012 год с нормализованными данными

Вновь проведём дескриптивный анализ для подготовленных данных

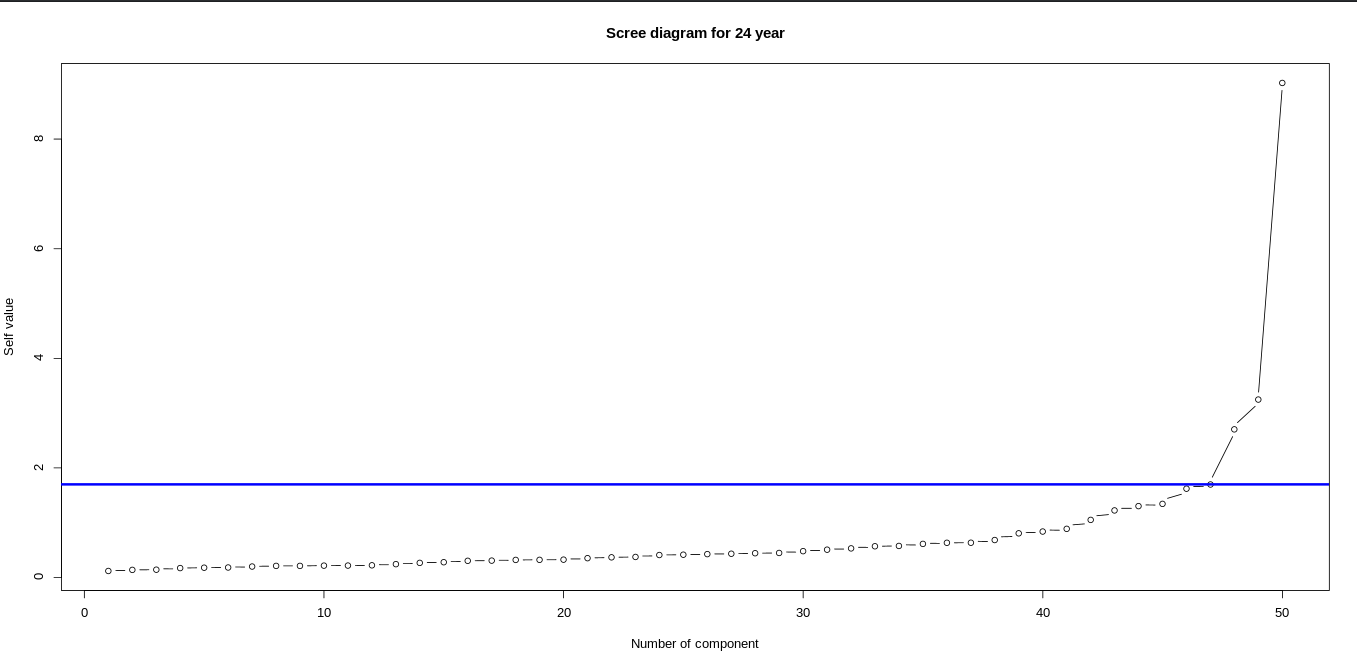
Рисунок 7 – Дескриптивный анализ датасета за 2012 год

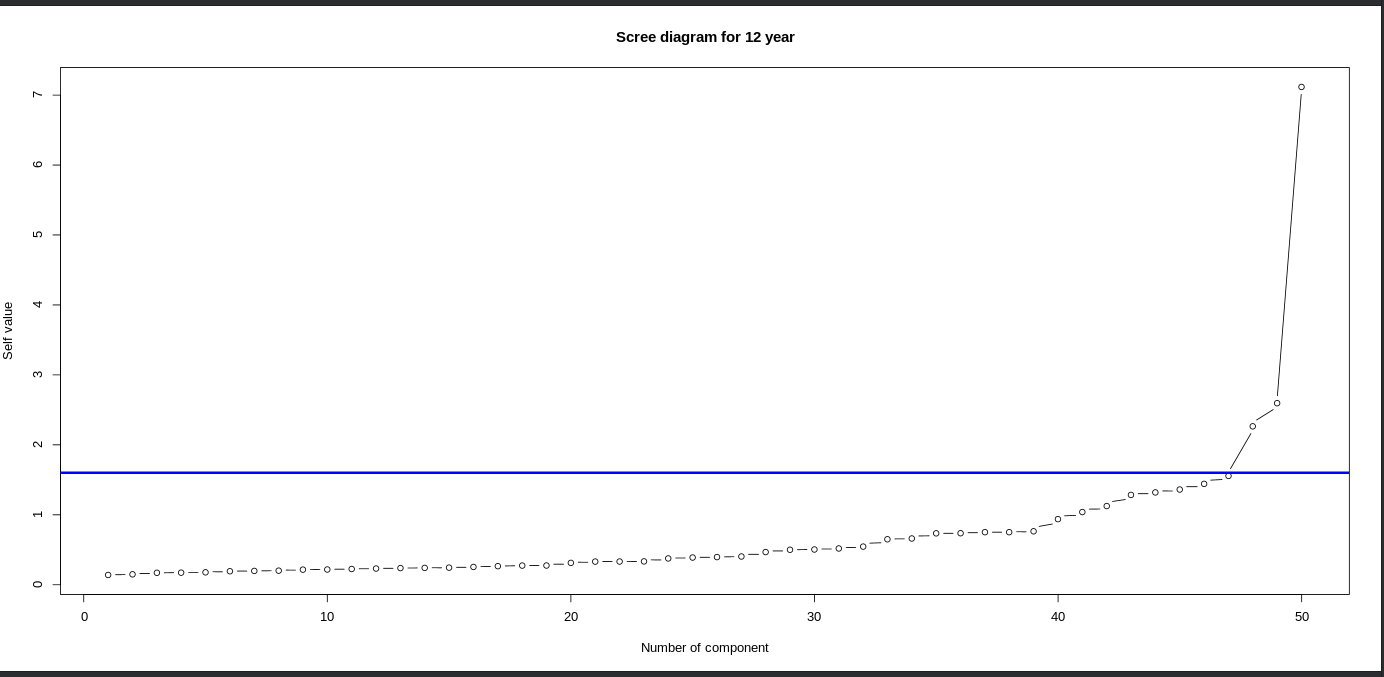
Рисунок 8 – Дескриптивный анализ датасета за 2024 год

Теперь данные приведены в форму, которую пригодно использовать для дальнейшей кластеризации и сравнения.

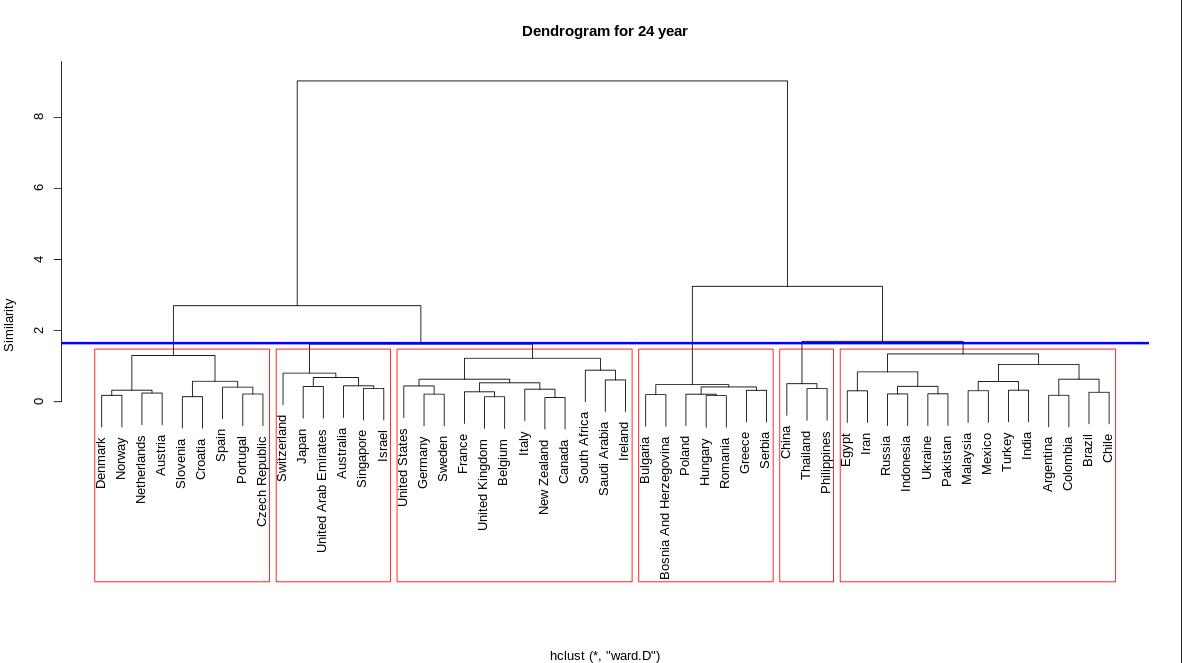
**Часть 2.** **Кластеризация**

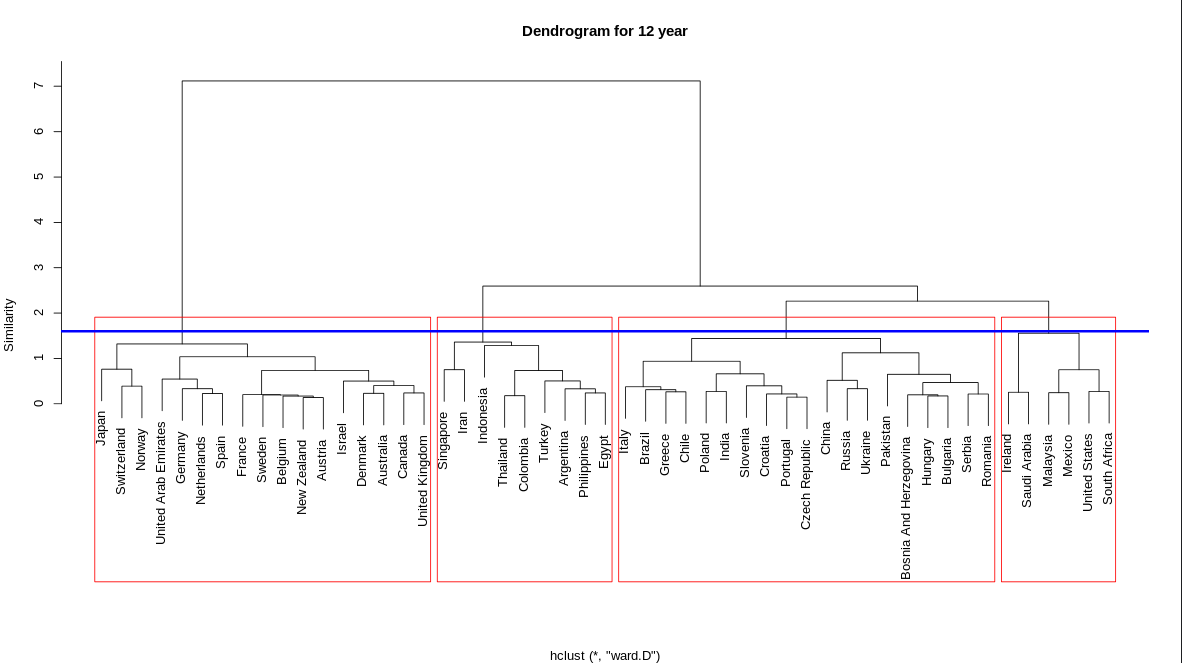
Для определения границы очерчивания кластеров взглянем на диаграмму каменная осыпь для двух датасетов.

Рисунок 9 – Диаграмма «Каменная осыпь» для 24 года.

Рисунок 10 – Диаграмма «Каменная осыпь» для 12 года.

Теперь применим полученный коэффициент схожести для отделения классов друг-от друга на дендрограмме

Рисунок 10 – Дендрограмма для 24 года.

Рисунок 10 – Дендрограмма для 12 года.

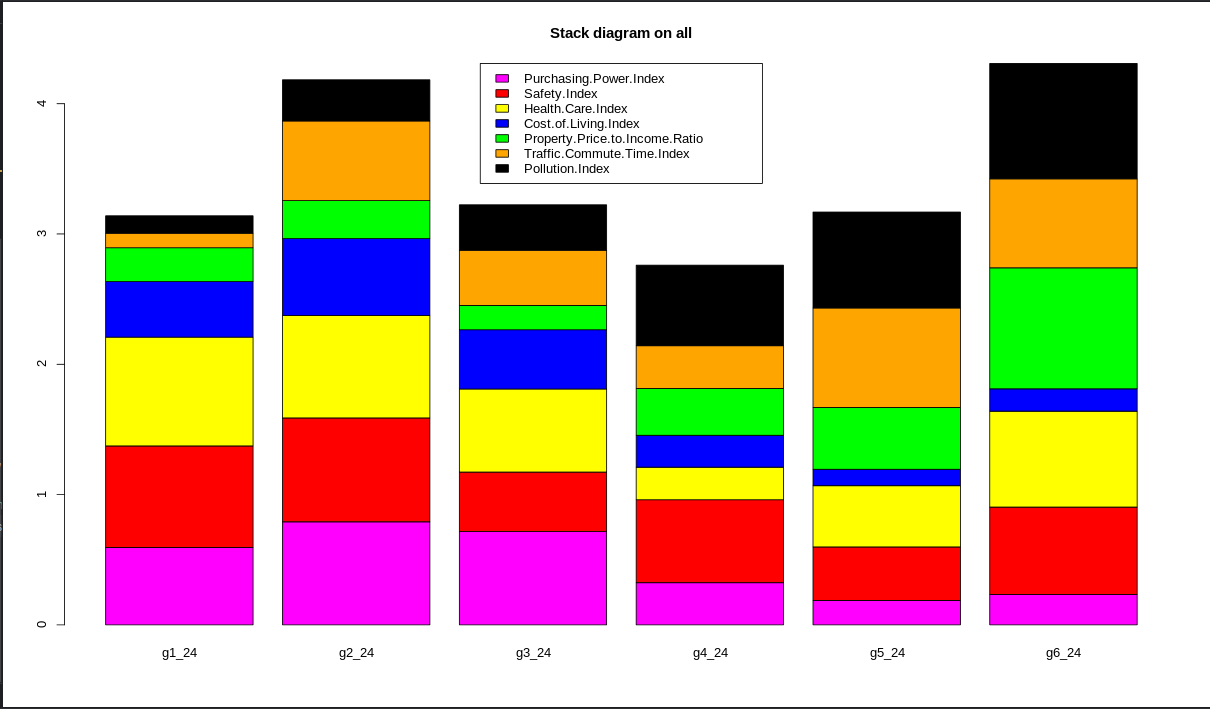
Смотря на выведенные дендрограммы видно, что в 24 году страны классифицируются более диверсифицированно. Выделяются 2 дополнительных класса.

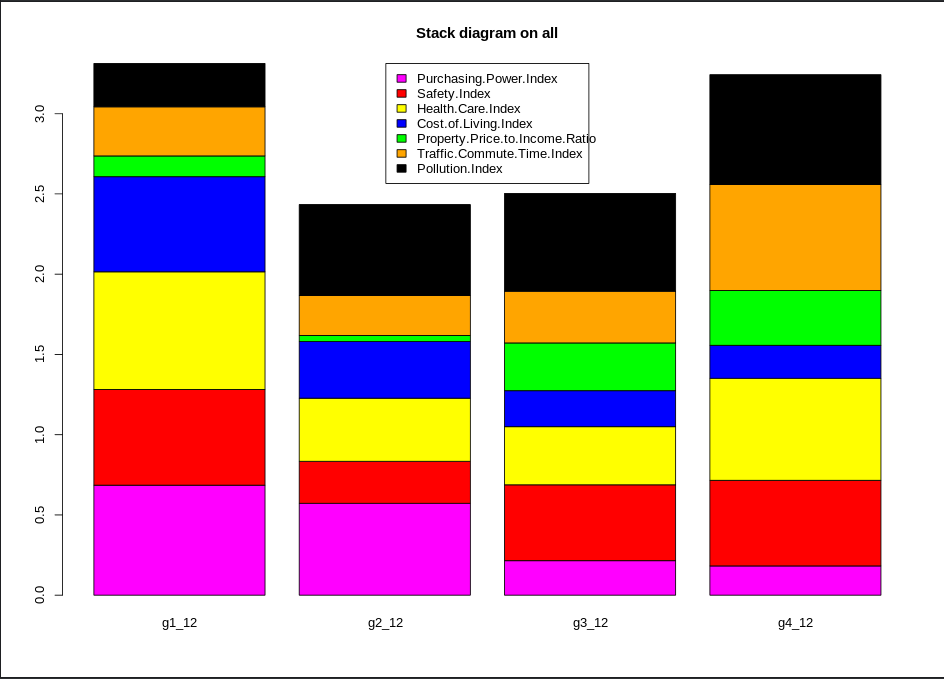
Далее представим списки стран, разбитых по кластерам.

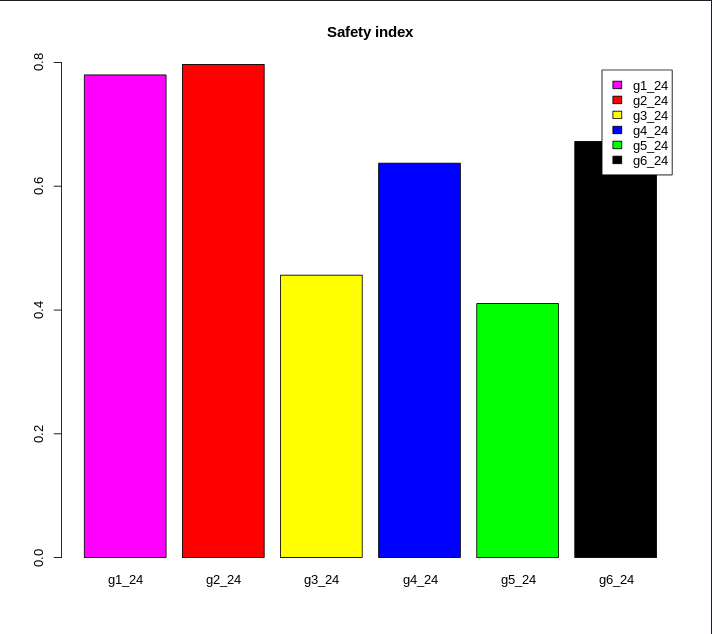
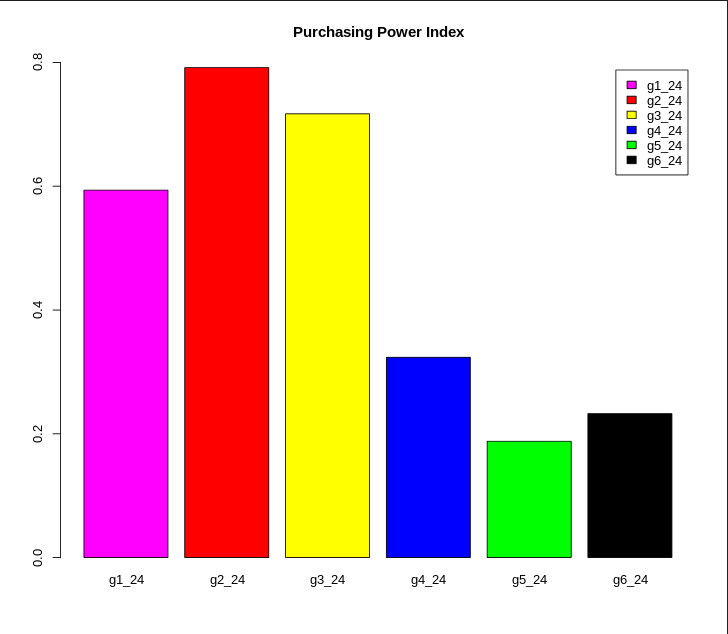
 Рисунок 11 – Страны, разбитые по кластерам в 2024 году

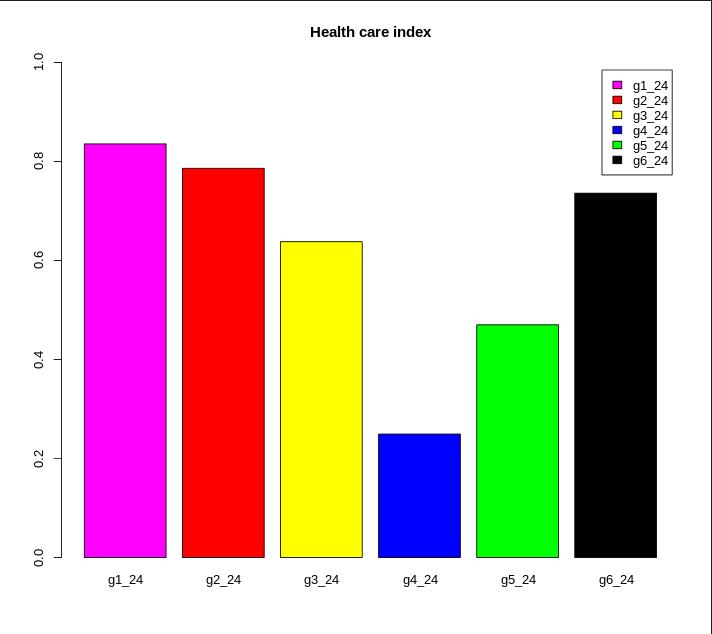
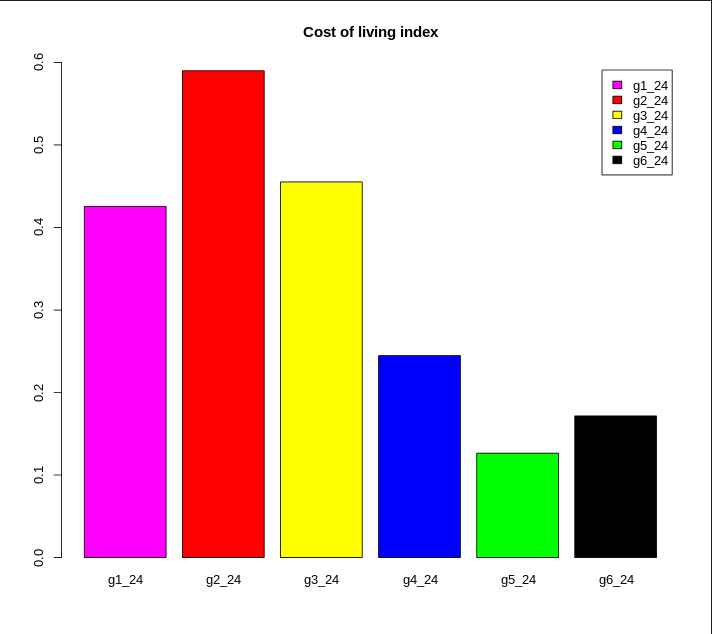
 Рисунок 12 – Страны, разбитые по кластерам в 2024 году

Далее представим распределения характеристик по всем кластерам.

Рисунок 13 – Распределение характеристик в каждом кластере за 2024 год.

Рисунок 14 – Распределение характеристик в каждом кластере за 2012 год.

Рисунок 14 – Распределение характеристик по группам по индексу покупательной способности и безопасности за 2024 год.

Рисунок 15 – Распределение характеристик по группам по индексу здравоохренения и уровня жизни за 2024 год.

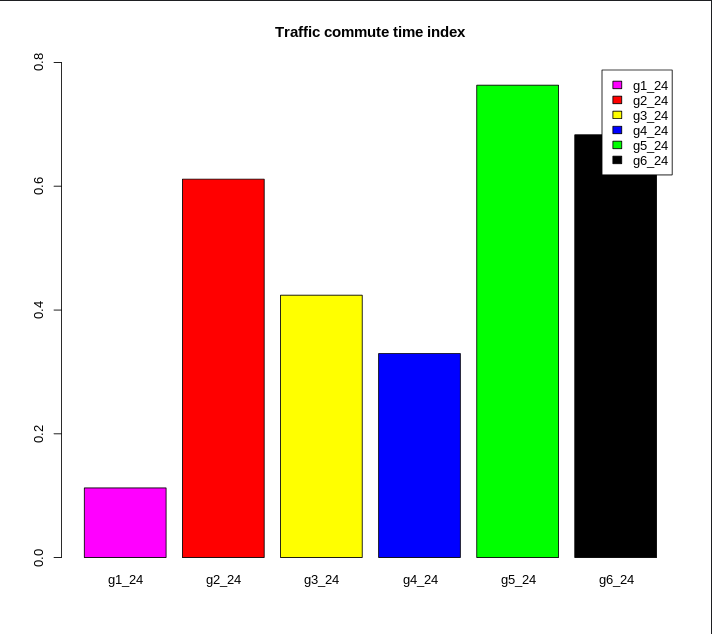
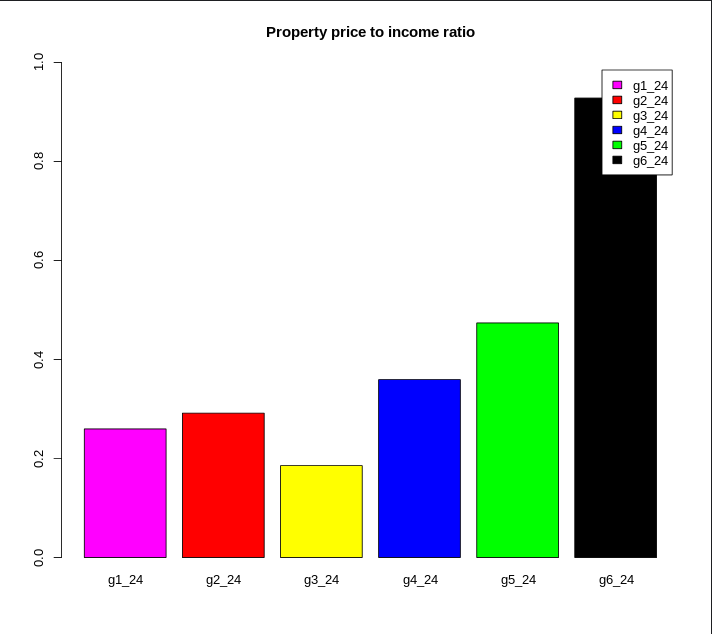
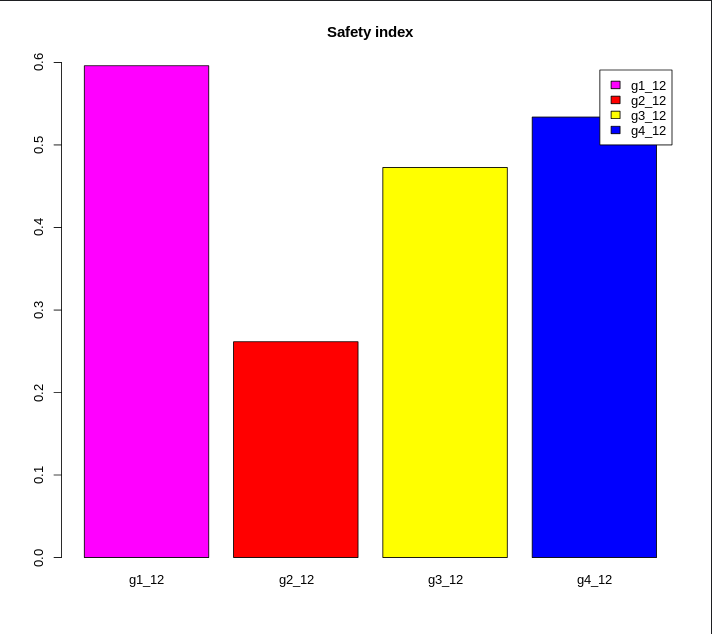
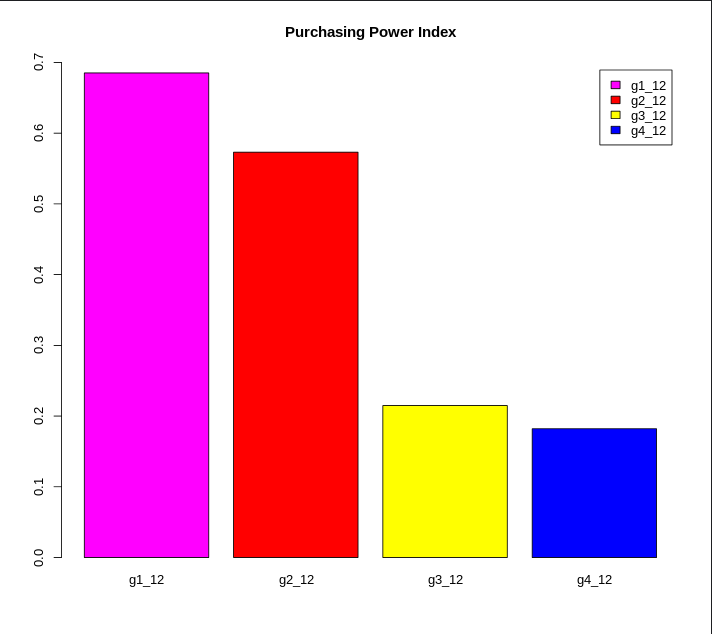
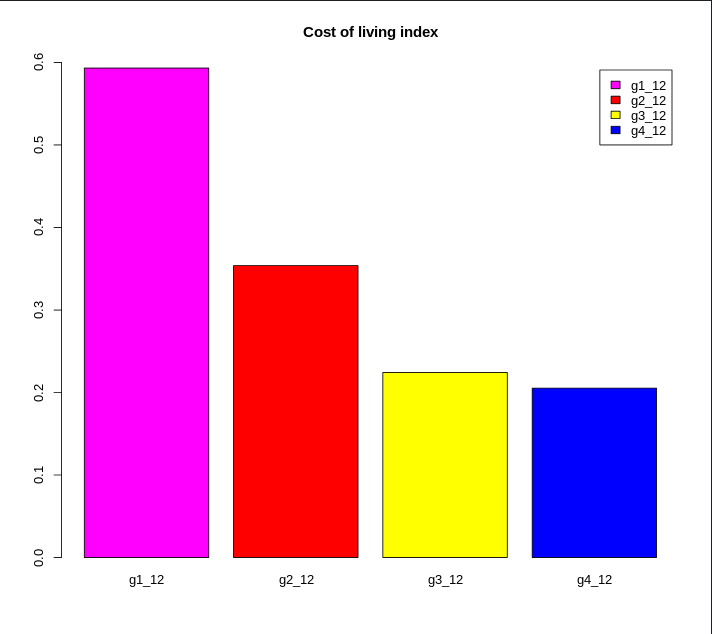
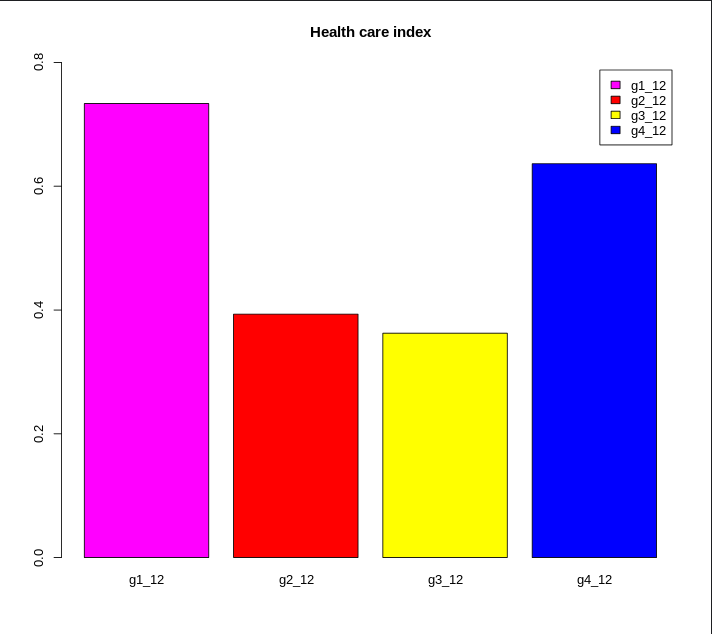


Рисунок 16– Распределение характеристик по группам по соотношению стоимости жилья к доходу и индексу пробок за 2024 год.

Рисунок 17 – Распределение характеристик по группам по индексу покупательной способности и безопасности за 2012 год.

Рисунок 18 – Распределение характеристик по группам по индексу здравоохренения и уровня жизни за 2012 год.

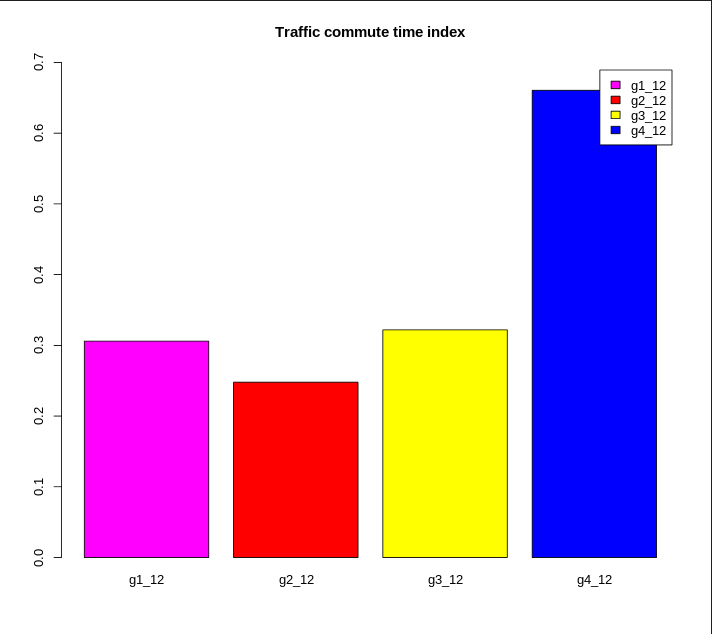
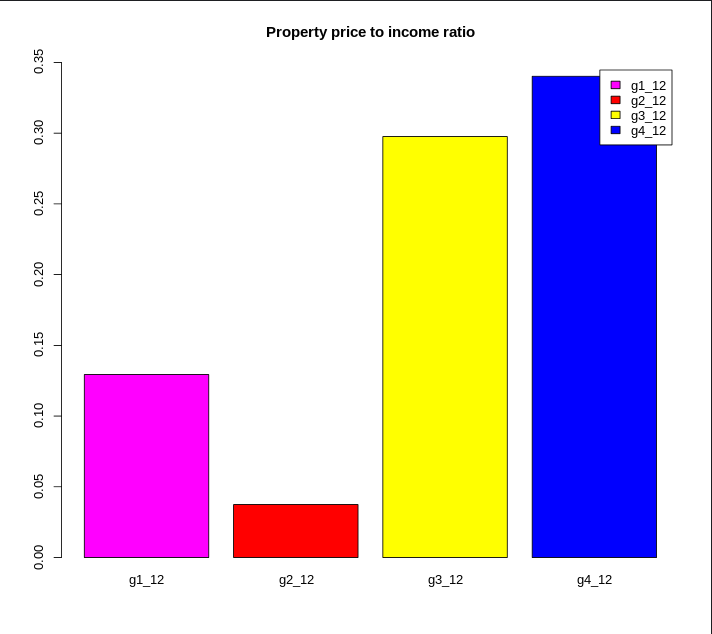
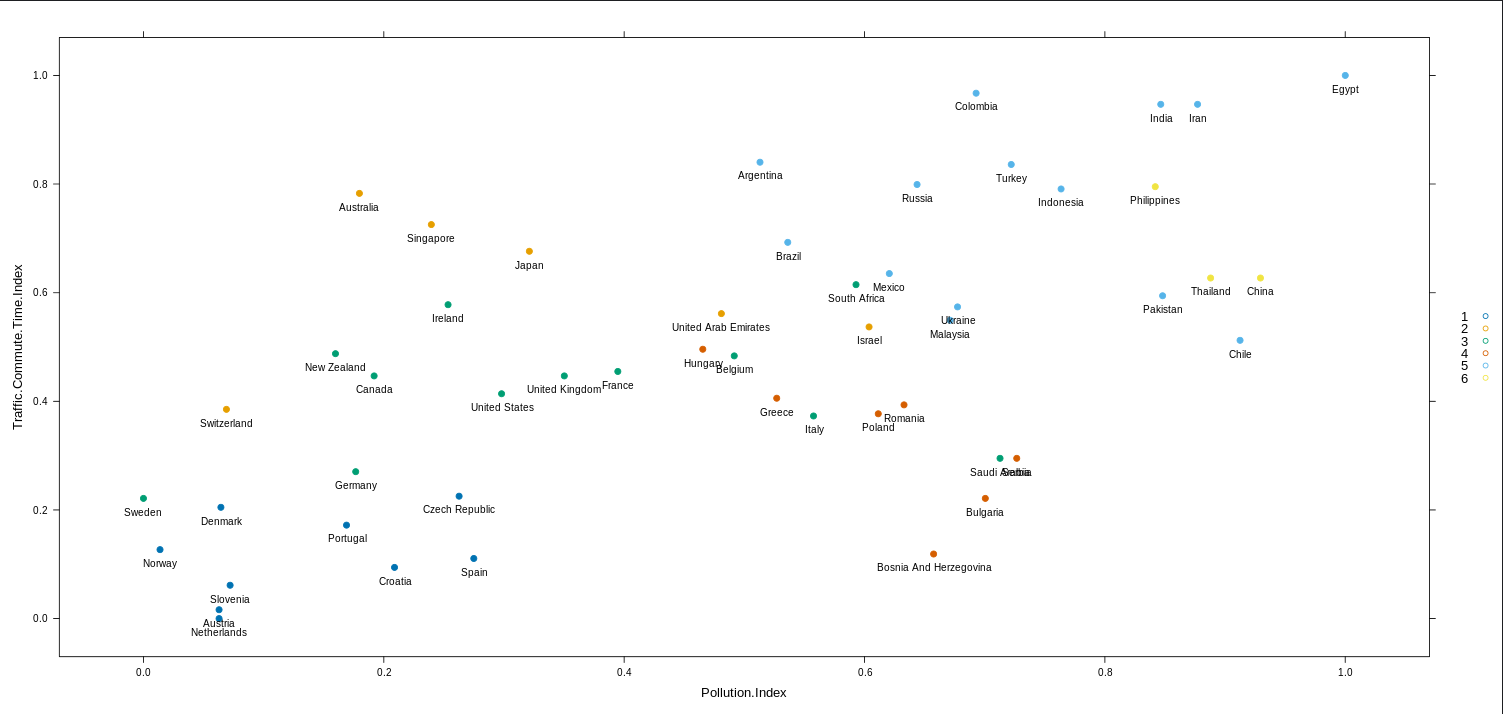


Рисунок 19– Распределение характеристик по группам по соотношению стоимости жилья к доходу и индексу пробок за 2012 год.

Рисунок 20 – Диаграмма рассеивания для пробок и индекса загрязнения за 2024 год.

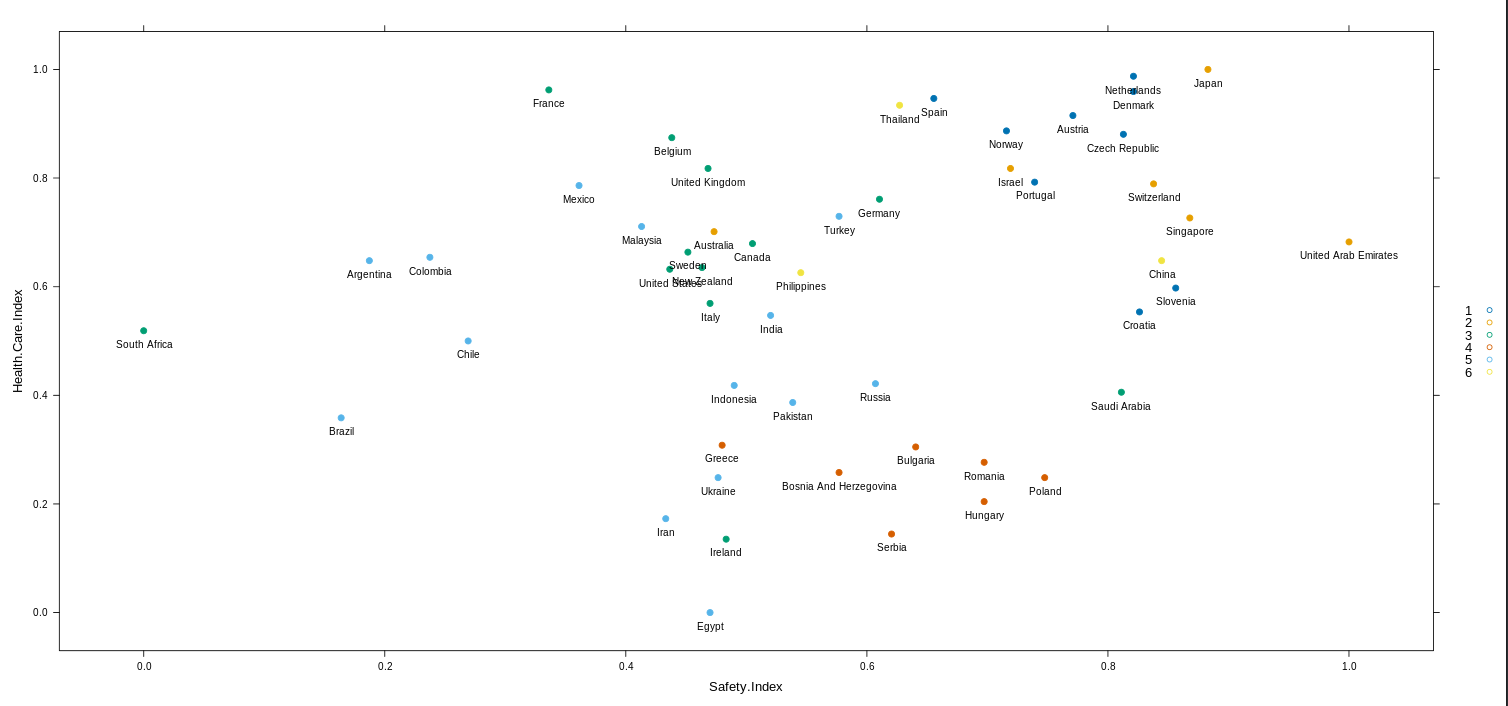


Рисунок 21 – Диаграмма рассеивания для уровня здравоохранения и индекса безопасности за 2024 год.

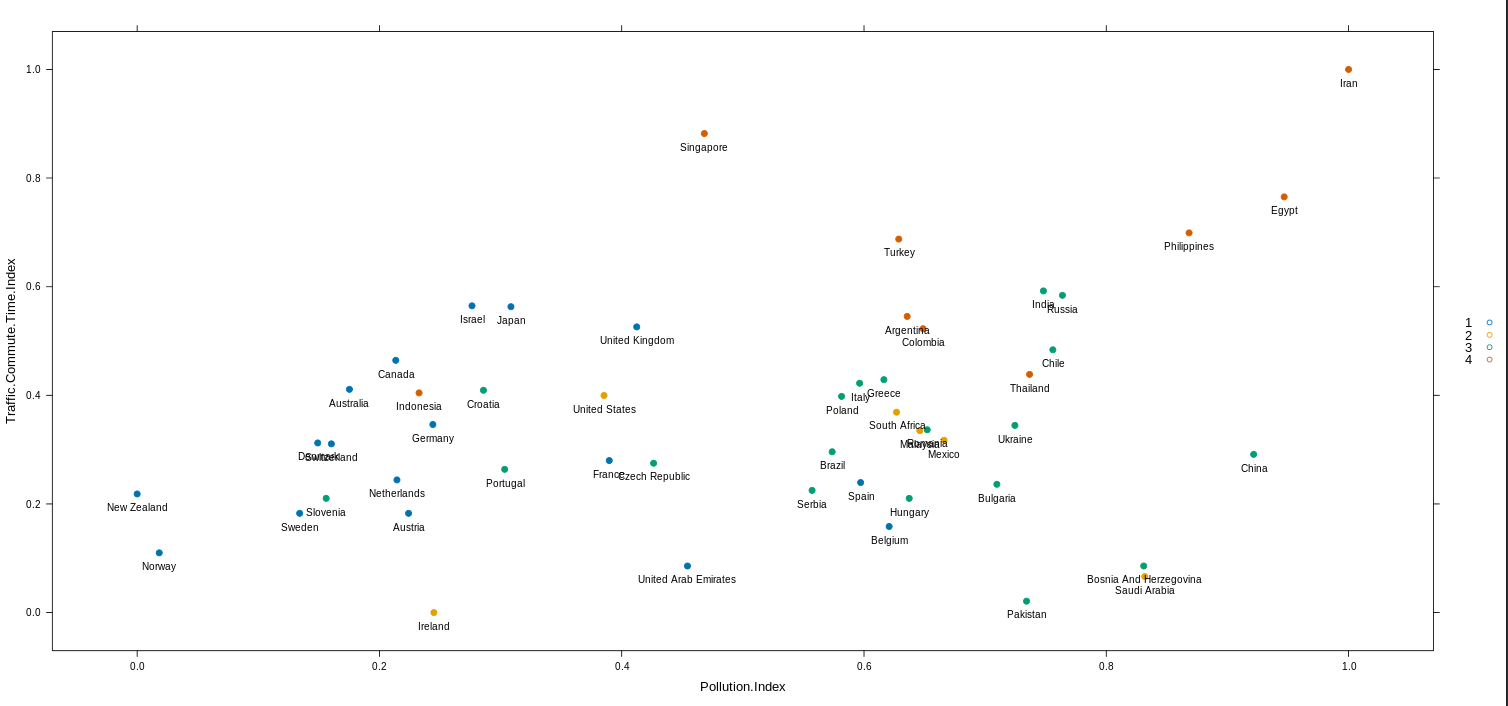
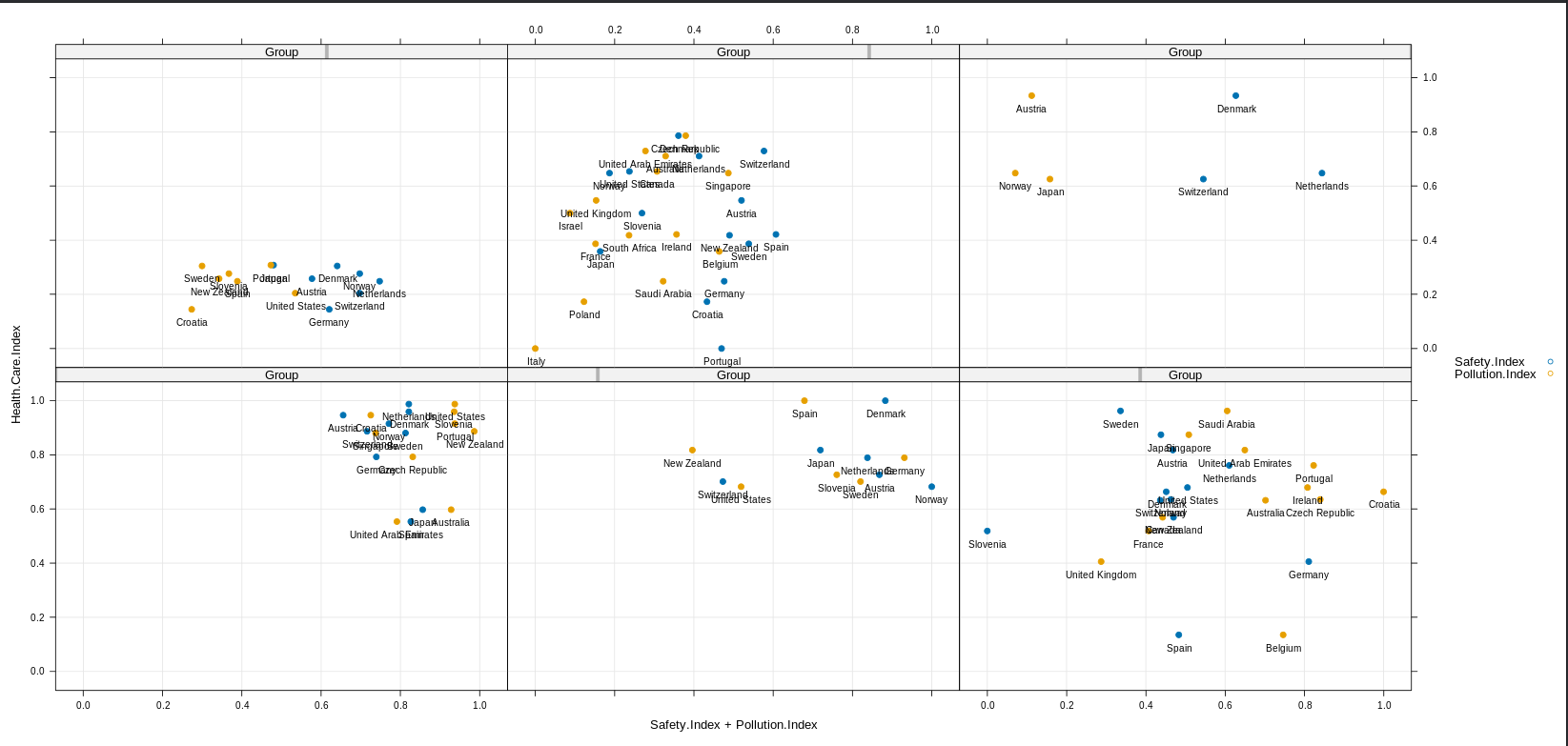
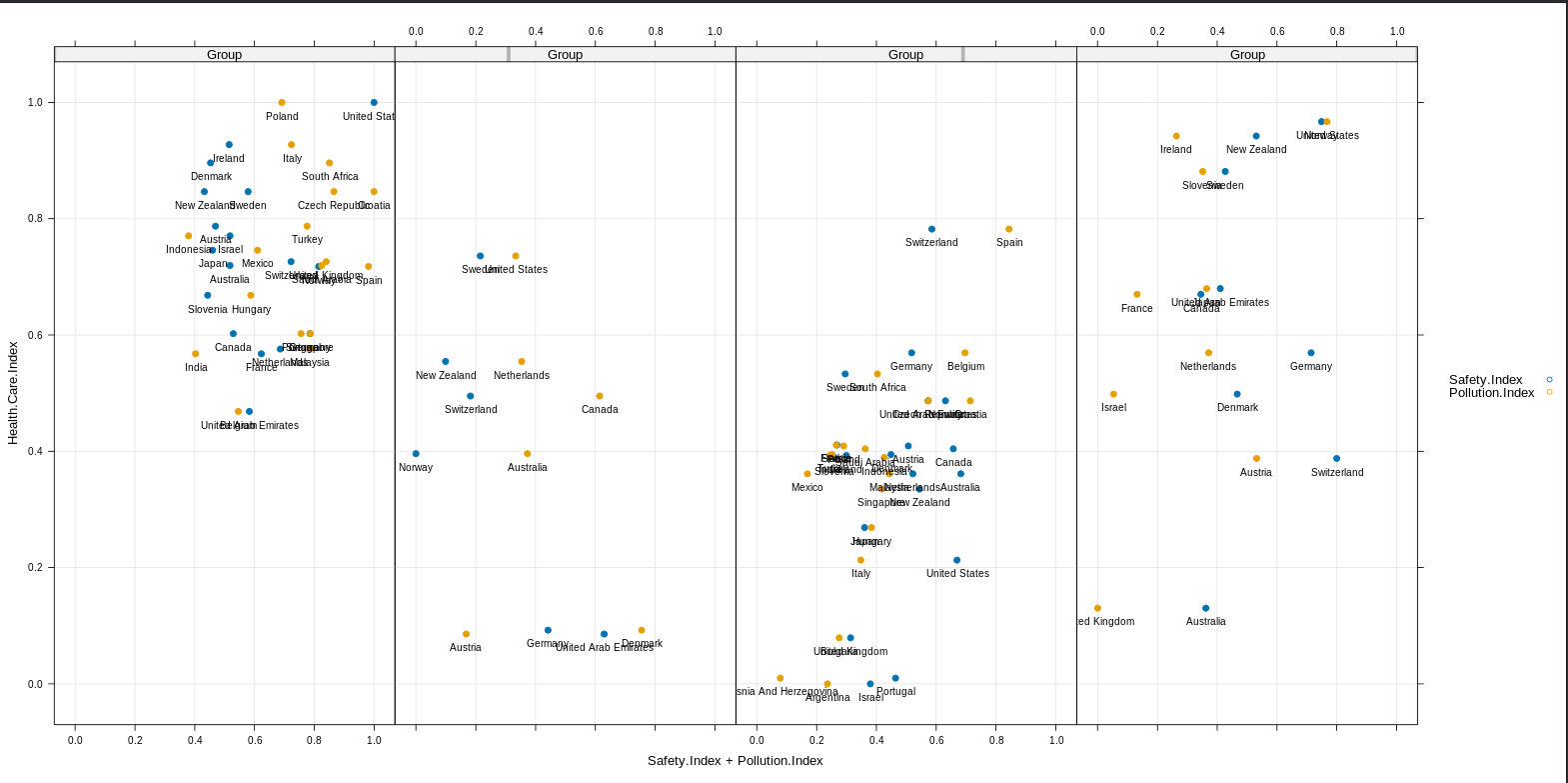
Рисунок 22 – Диаграмма рассеивания для пробок и индекса загрязнения за 2012 год.

Рисунок 23 – Диаграмма рассеивания для уровня здравоохранения и индекса безопасности за 2012 год.

Рисунок 24 – График, показывающий распределение индекса безопасности и инвертированного индекса загрязнения в соотношении с индексом здравоохранения за 24 год.

Можно сделать вывод, что прослеживается закономерность, чем больше индекс безопасности и «меньше» индекс загрязнения, тем выше уровень здравоохранения, в 12 году ситуация аналогичная.

Рисунок 25 – График, показывающий распределение индекса безопасности и инвертированного индекса загрязнения в соотношении с индексом здравоохранения за 12 год.

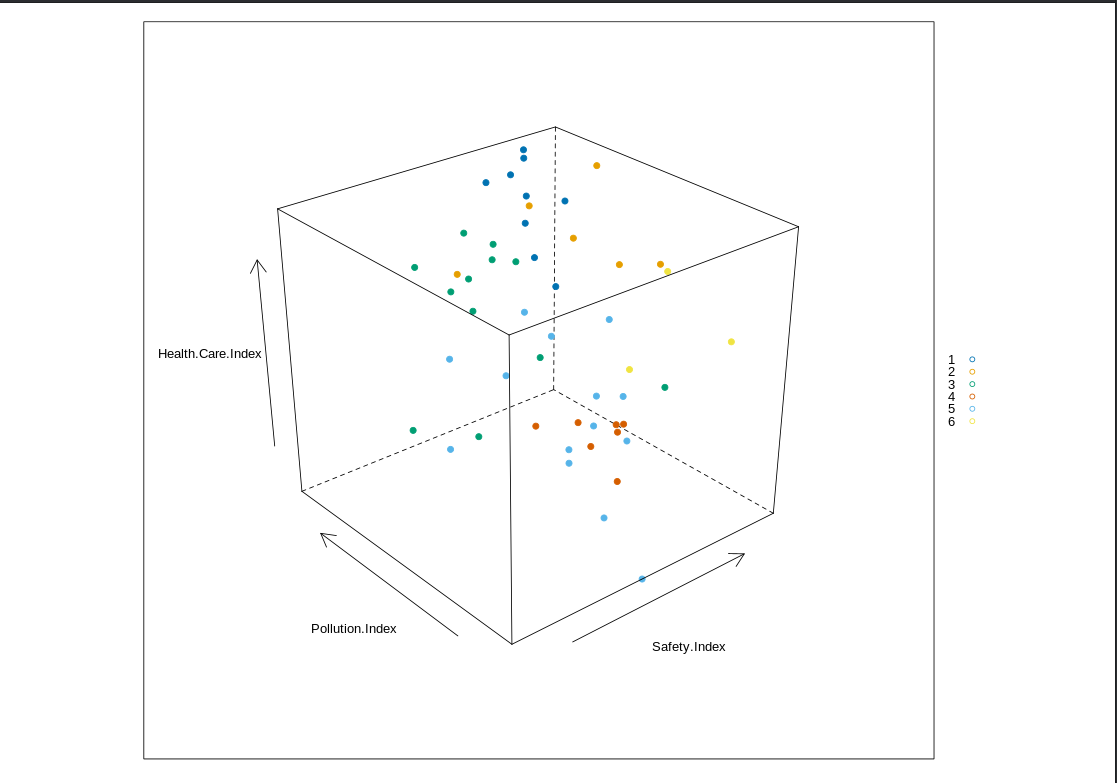


Рисунок 26 – 3Д график, показывающий распределение индекса безопасности и инвертированного индекса загрязнения в соотношении с индексом здравоохранения за 24 год.

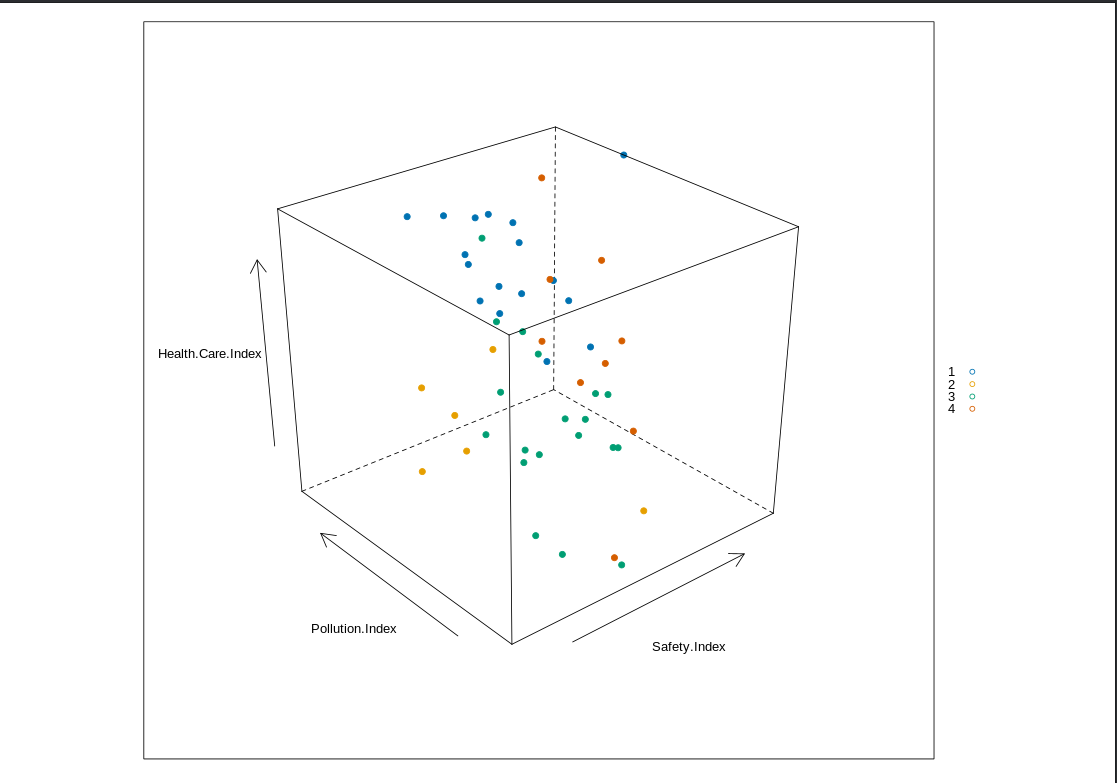


Рисунок 26 – 3Д график, показывающий распределение индекса безопасности и инвертированного индекса загрязнения в соотношении с индексом здравоохранения за 12 год.

**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.