Министерство образования и науки Российской Федерации

Севастопольский государственный университет

Институт информационных технологий

Кафедра ИС

# ОТЧЁТ

по лабораторной работе №4

КЛАСТЕРНЫЙ АНАЛИЗ. ОСНОВНЫЕ ЭТАПЫ И ЗАДАЧИ КЛАСТЕРНОГО АНАЛИЗА ДАННЫХ.

Выполнил:

ст. гр. ИС/б-21-2-о

Мовенко К. М.

Проверила:

Сырых О. А.

Севастополь

2024

# ЦЕЛЬ РАБОТЫ

Закрепить теоретические знания и приобрести практические навыки в проведении кластерного анализа. По экспериментальным данным исследовать возможности языка R для проведения кластерного анализа.

# ЗАДАНИЕ

* 1. Провести кластерный анализ экспериментальных данных;
  2. Проведя процедуру кластеризации (разбиение на классы или кластеры) несколько раз при различных значениях числа кластеров (2-10), выбрать лучшую группировку в смысле критерия минимума отношений средних внутри кластерных и межкластерных расстояний. Полученные результаты оформить в виде таблицы;
  3. Изобразить графически значения данного показателя качества классификации. Для этого построить диаграмму, на которой по оси Х – количество кластеров, по оси Y – значения показателя J;
  4. Провести иерархический кластерный анализ в среде Rcmdr;
  5. Провести анализ экспериментальных данных используя разные методы. Полученные результаты сравнить и сделать выводы.

# ХОД РАБОТЫ

В качестве экспериментальных данных был взят набор данных о показателях качества жизни по странам мира, использованный в ЛР №2. В среде R Commander был проведён многомерный кластерный анализ данных методом k-средних по всем параметрам, число кластеров взято за 2.

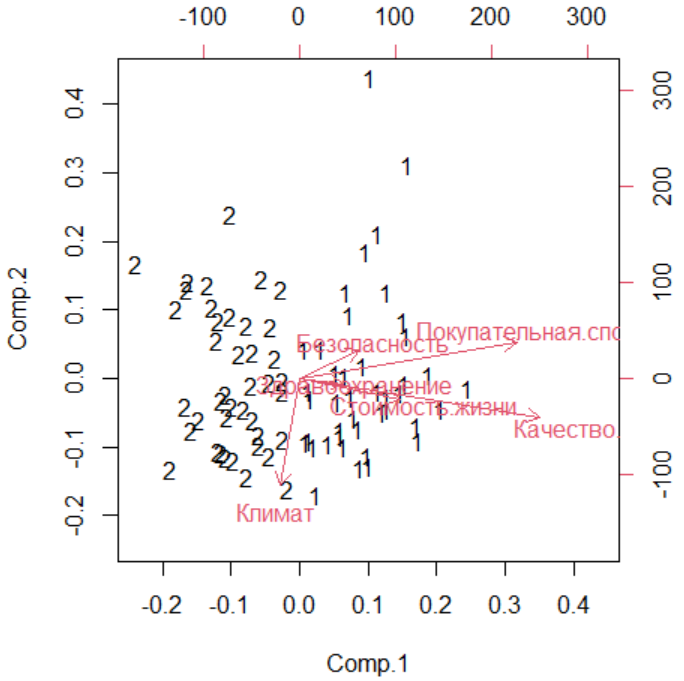


Рисунок 1 – Разбиение данных на 2 кластера методом k-средних

По результатам анализа можно заметить, что страны в наборе данных были разделены на два кластера: условно более и менее экономически развитые. Для оценки качества анализа требуются дополнительные исследования.

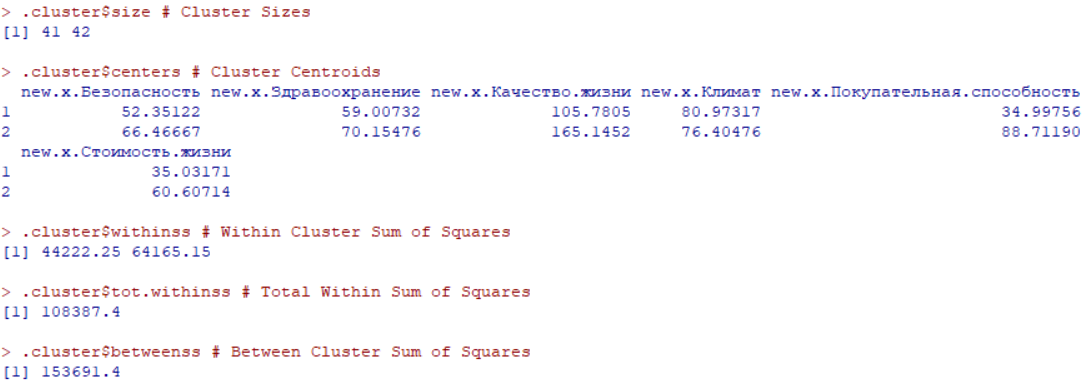


Рисунок 2 – Результат выполненного анализа

Была проведена кластеризация данных при различном числе кластеров (2-10). По значениям средних расстояний между точками внутри кластеров и между ними для каждой группировки была вычислена оценка качества кластеризации.

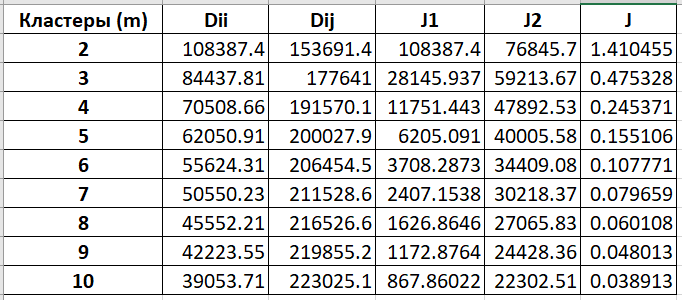


Рисунок 3 – Исследование качества кластеризации

Была построена диаграмма зависимости показателя качества классификации от числа кластеров (Рисунок 4). По критерию каменистой осыпи было определено оптимальное число кластеров – 4.

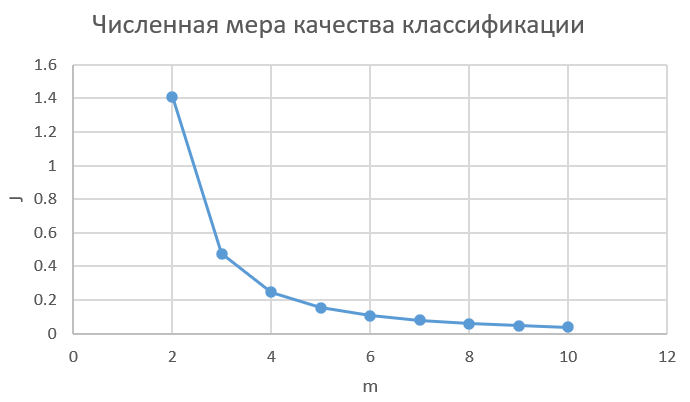


Рисунок 4 – Диаграмма численной меры качества классификации

Затем был проведён иерархический кластерный анализ данных тремя разными способами: методом Уорда (Рисунок 5), методом Мак-Квити (Рисунок 6) и методом полной связи (Рисунок 7).

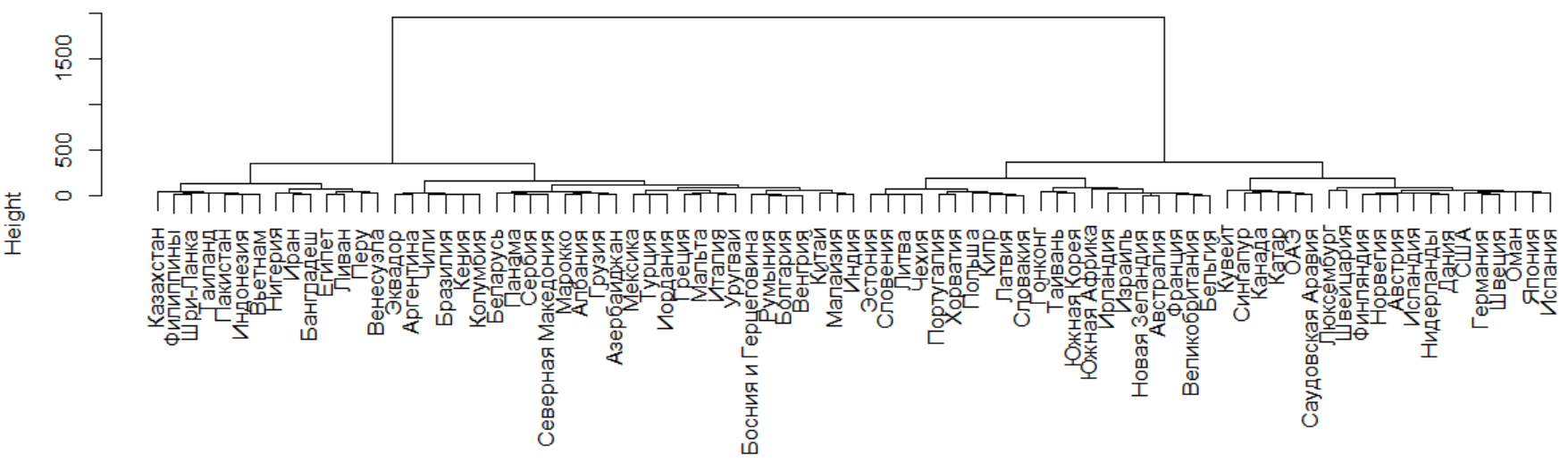


Рисунок 5 – Иерархический кластерный анализ методом Уорда

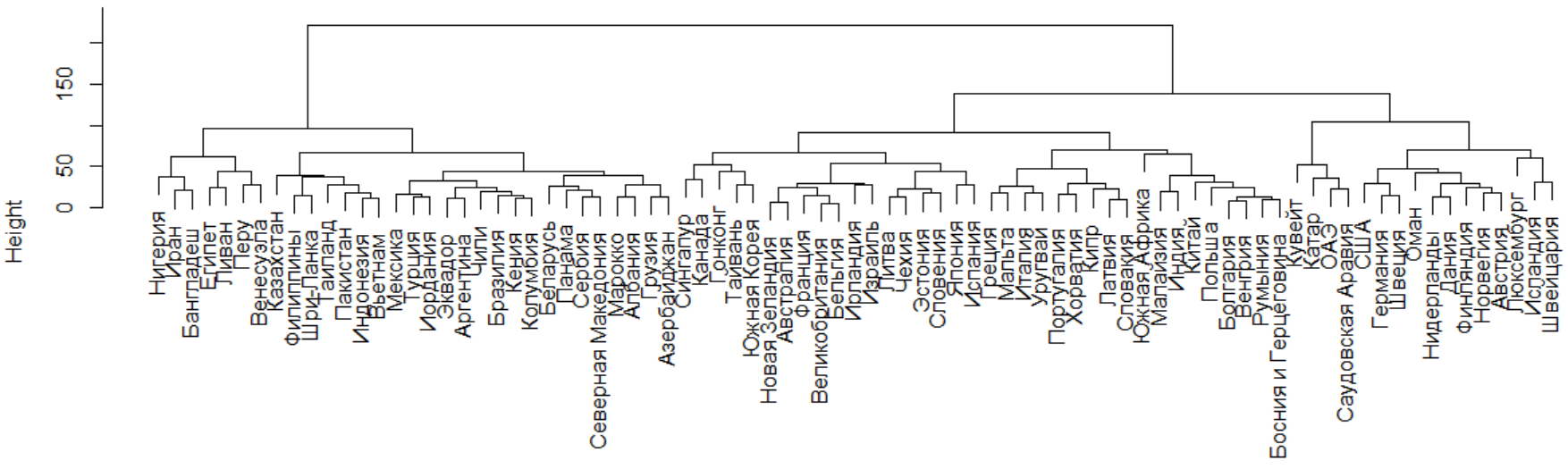


Рисунок 6 – Иерархический кластерный анализ методом полной связи

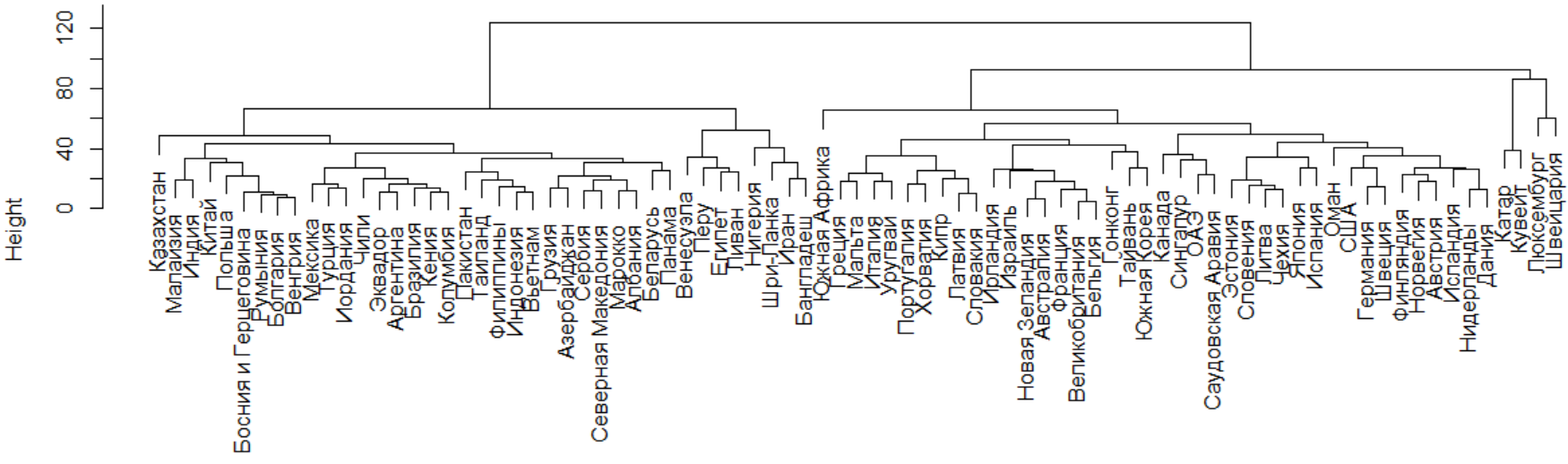


Рисунок 7 – Иерархический кластерный анализ методом Мак-Квитти

# ВЫВОД

В ходе работы были приобретены навыки проведения кластерного анализа экспериментальных данных, исследованы возможности языка R для проведения кластерного анализа.

Были опробованы два подхода: метод k-средних и иерархический анализ. Для рассматриваемого набора данных (коэффициент качества жизни по странам мира) при тестировании первого метода наилучший результат был получен при разделении данных на четыре кластера. Это значение является оптимальным для разбиения на кластеры.