Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования «Кубанский государственный университет»

Кафедра вычислительных технологий

**ОТЧЕТ**

о выполнении лабораторной работы №5.1

по дисциплине «Обработка больших данных»

Тема**:** Задачи классификации и кластеризации

Выполнил: ст. гр. 36/2

Решетка Д.В.

Проверил: преподаватель

Яхонтов А.А.

Краснодар

2025

**Цель:** 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.

### **Задание**

1. В начале отчета необходимо разместить формулировку задания и фрагмент исходного датасета.

2. Выполнить дескриптивный анализ данных (здесь приветствуются дополнительные исследования).

3. Оценить оптимальное число кластеров, для этого построить диаграмму "Метод силуэта", “Метод локтя”, "Статистику разрыва" и Алгоритм консенсуса.

4. Выполнить иерархическую кластеризацию вашего набора данных, построив дендрограмму. Подробно обосновать Ваш выбор числа групп.

5. Построить диаграмму со столбчатыми диаграммами (рис. 5.8) и боксплотами групп (рис. 5.12). Провести сравнительный анализ полученных групп.

6. Выполнить кластеризацию своего датасета по k-means (рис.5.9, 5.10).

7. Выполнить построение scatterplot (рис. 5.13) с помощью функций plot или pairs.

8. Построить трехмерную кластеризацию по scatterplot3d (5.16)

9. В целом: выполнить шаги 1-3,5 анализа для своего набора данных (если какие-то из шагов нерелевантны вашему набору данных, объяснить почему).

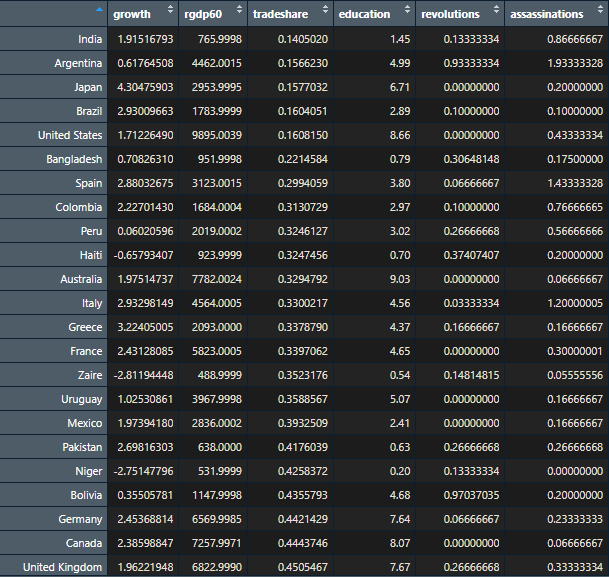
**Вариант**:



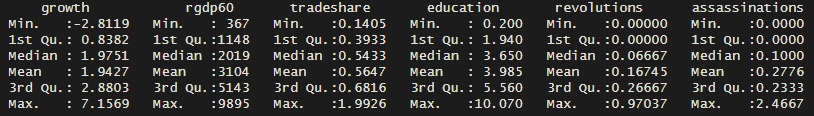
**Ход работы**

1. Исходные данные:

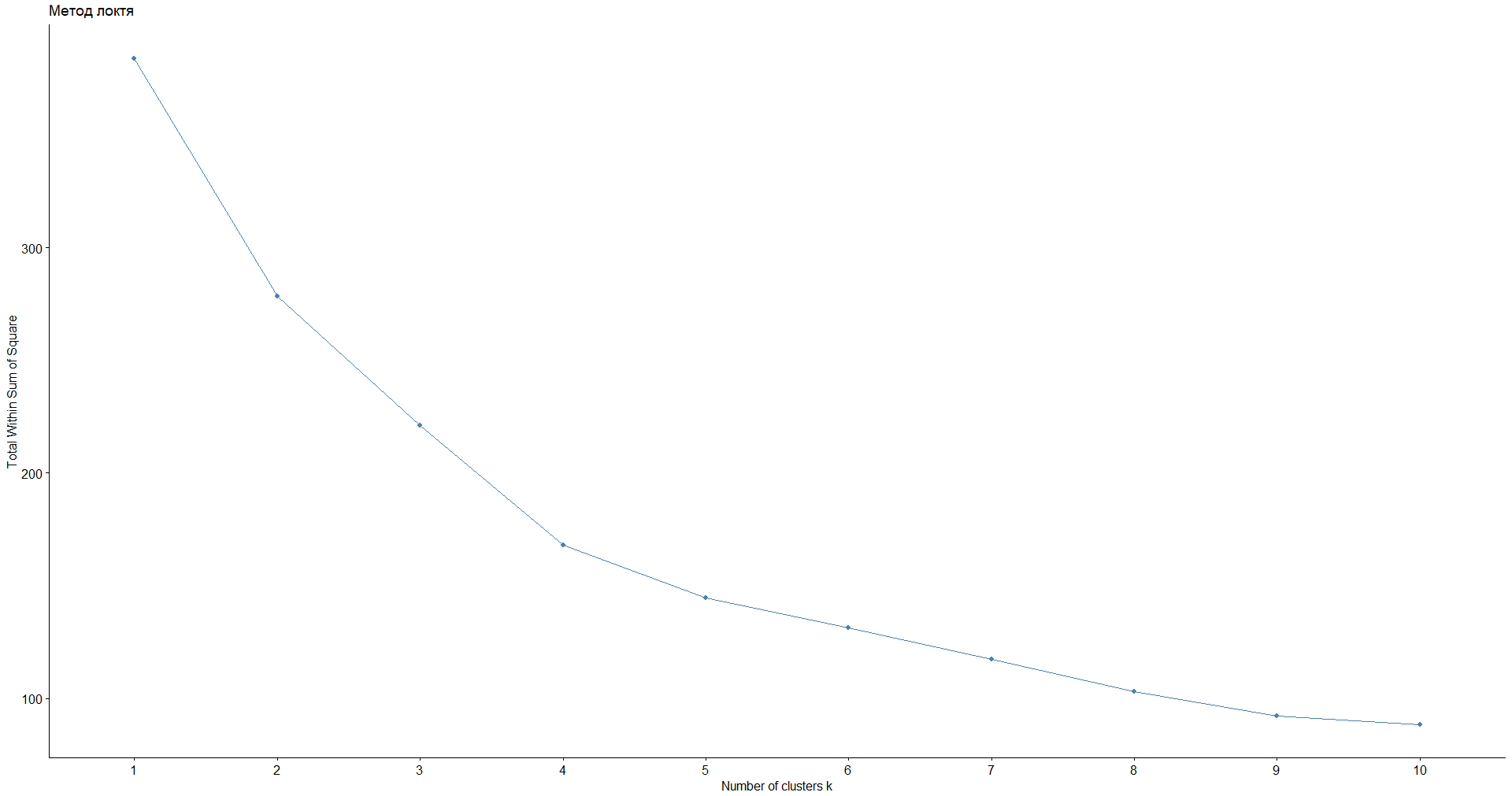
Таблица с метриками стран.



1. Дескриптивный анализ.

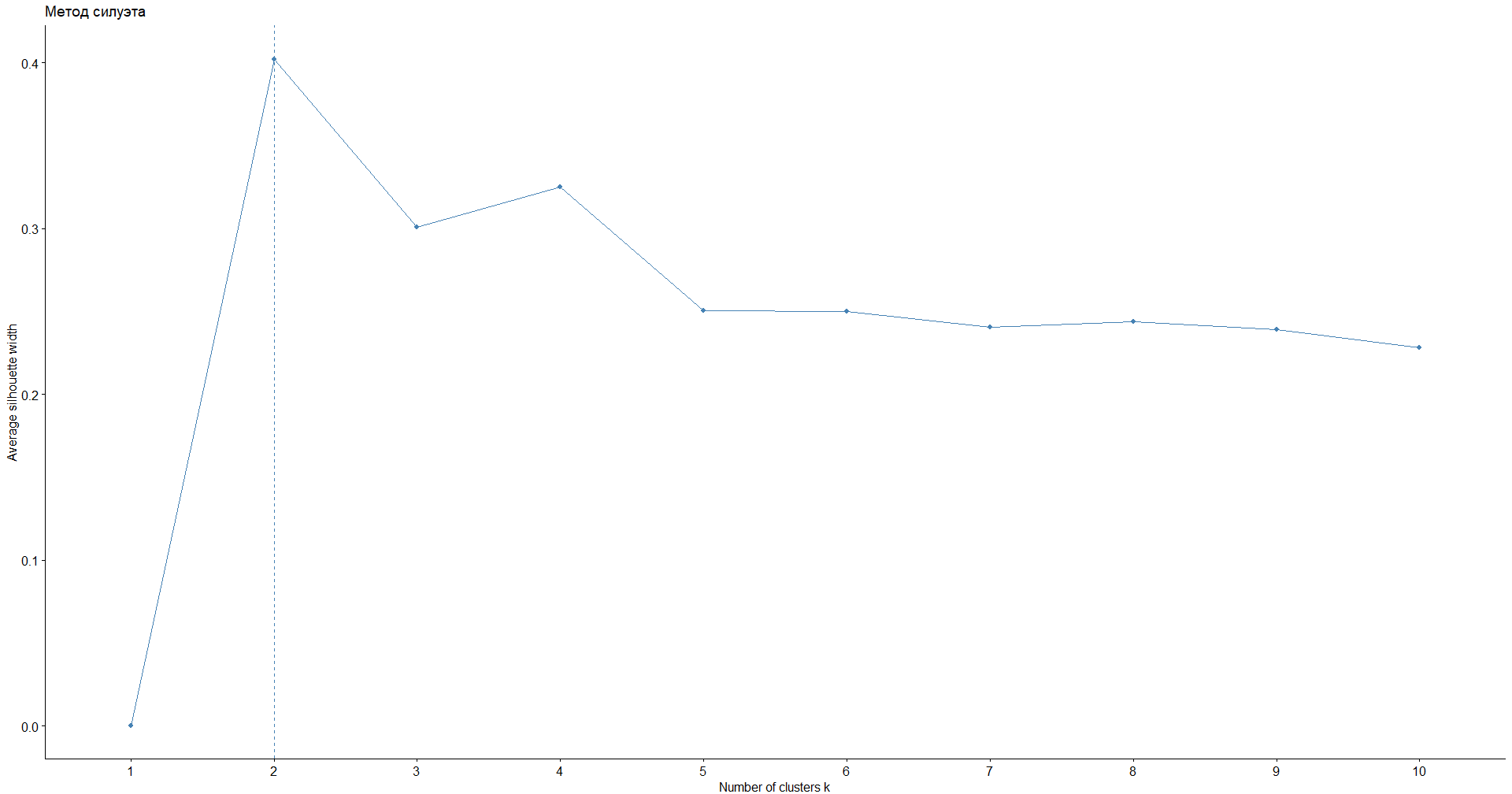


1. Метод локтя.



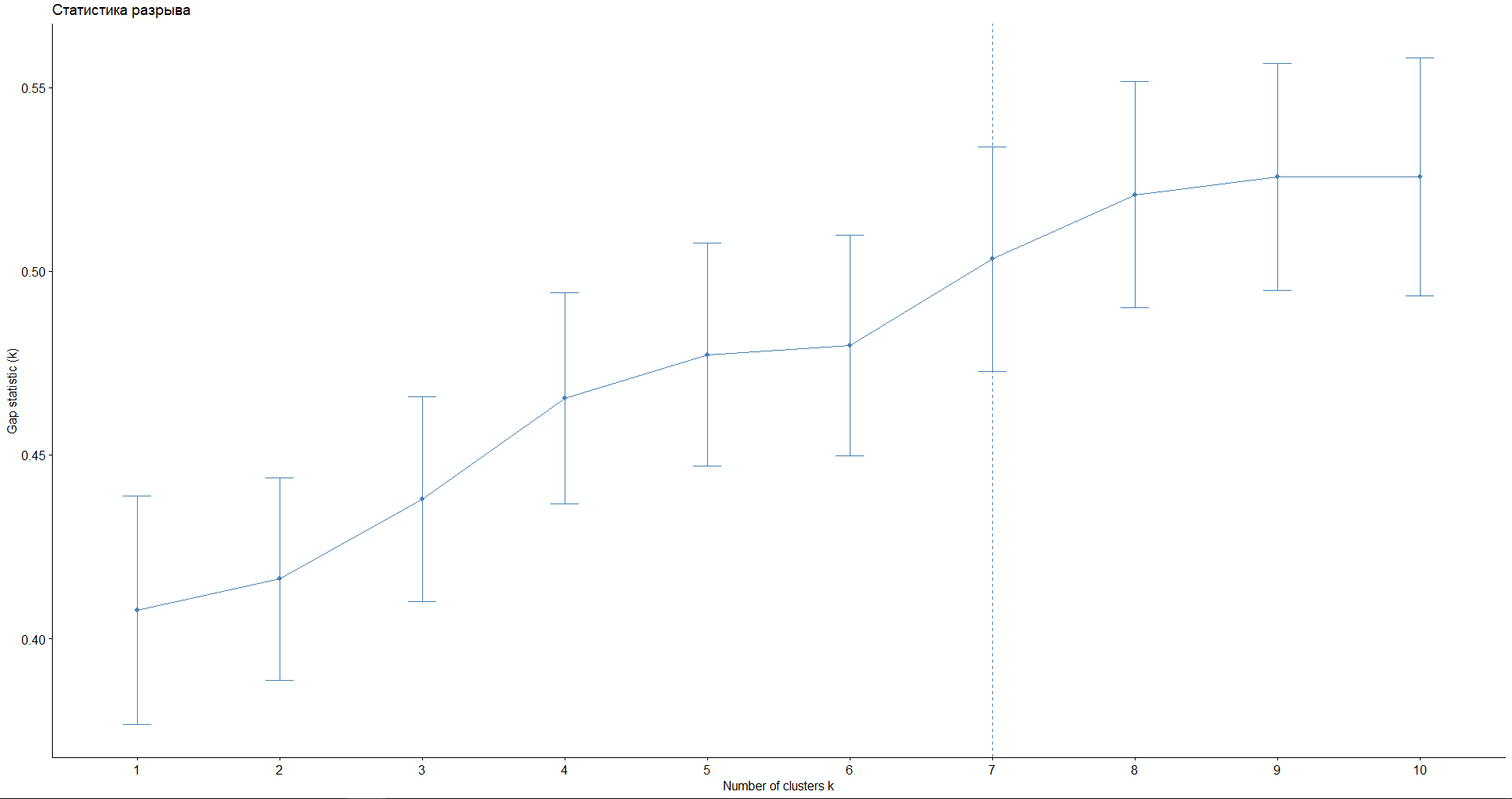
Метод Локтя рассматривает общую сумму квадратов внутри кластера (WSS) как функцию количества кластеров. Оптимальное количество кластеров будет в точке, где WSS перестаёт резко уменьшаться. В данном случае это 4 кластера.

1. Метод силуэта.



Метод Silhouette измеряет качество кластеризации и определяет, насколько хорошо каждая точка лежит в пределах своего кластера. Оптимальное количество кластеров будет в точке, где индекс силуэта имеет максимальное значение. В данном случае это 2 кластера.

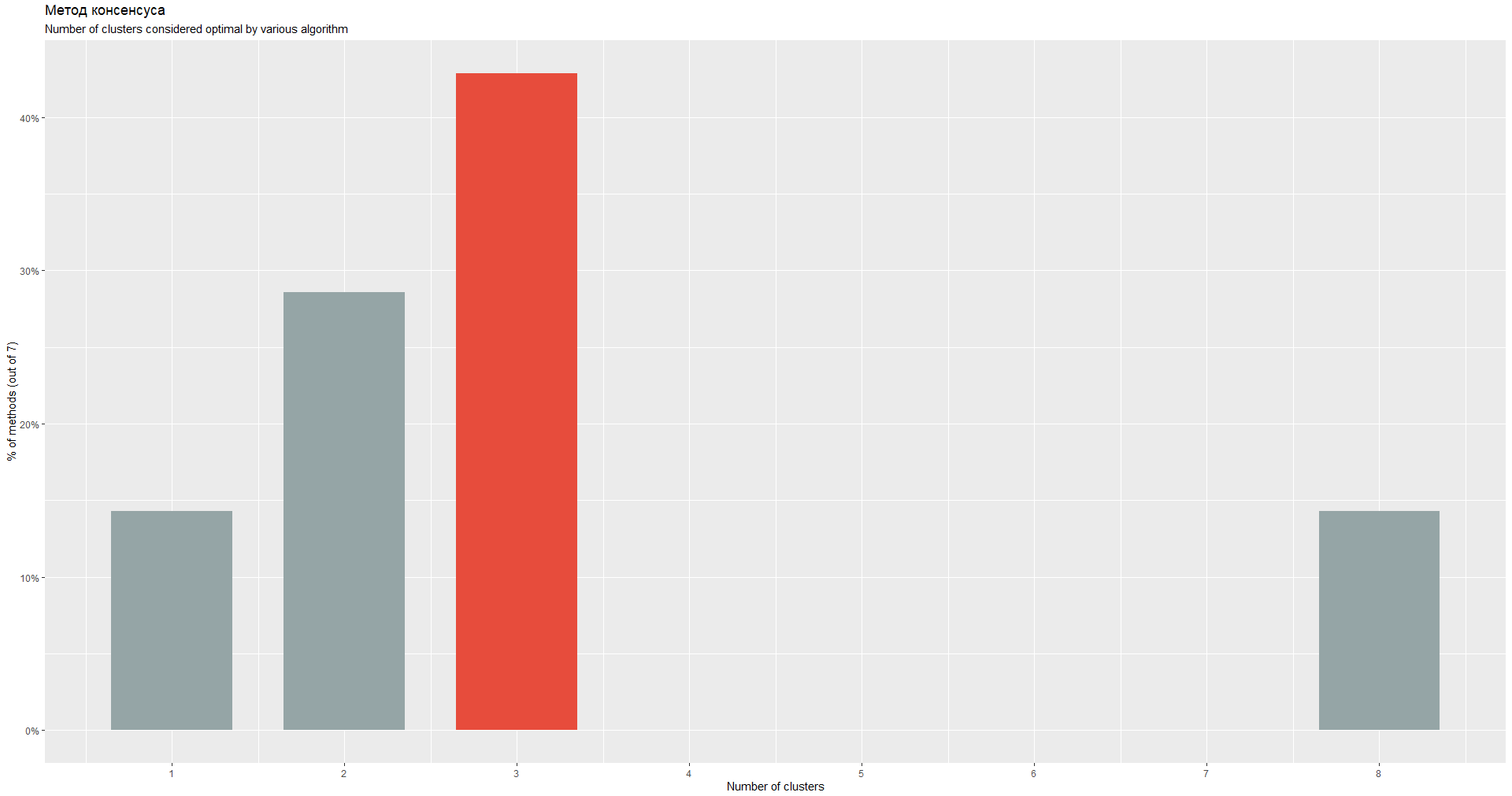
1. Статистика разрыва.



Данный метод сравнивает общую внутрикластерную дисперсию для разных значений k с их ожидаемыми значениями для распределения без

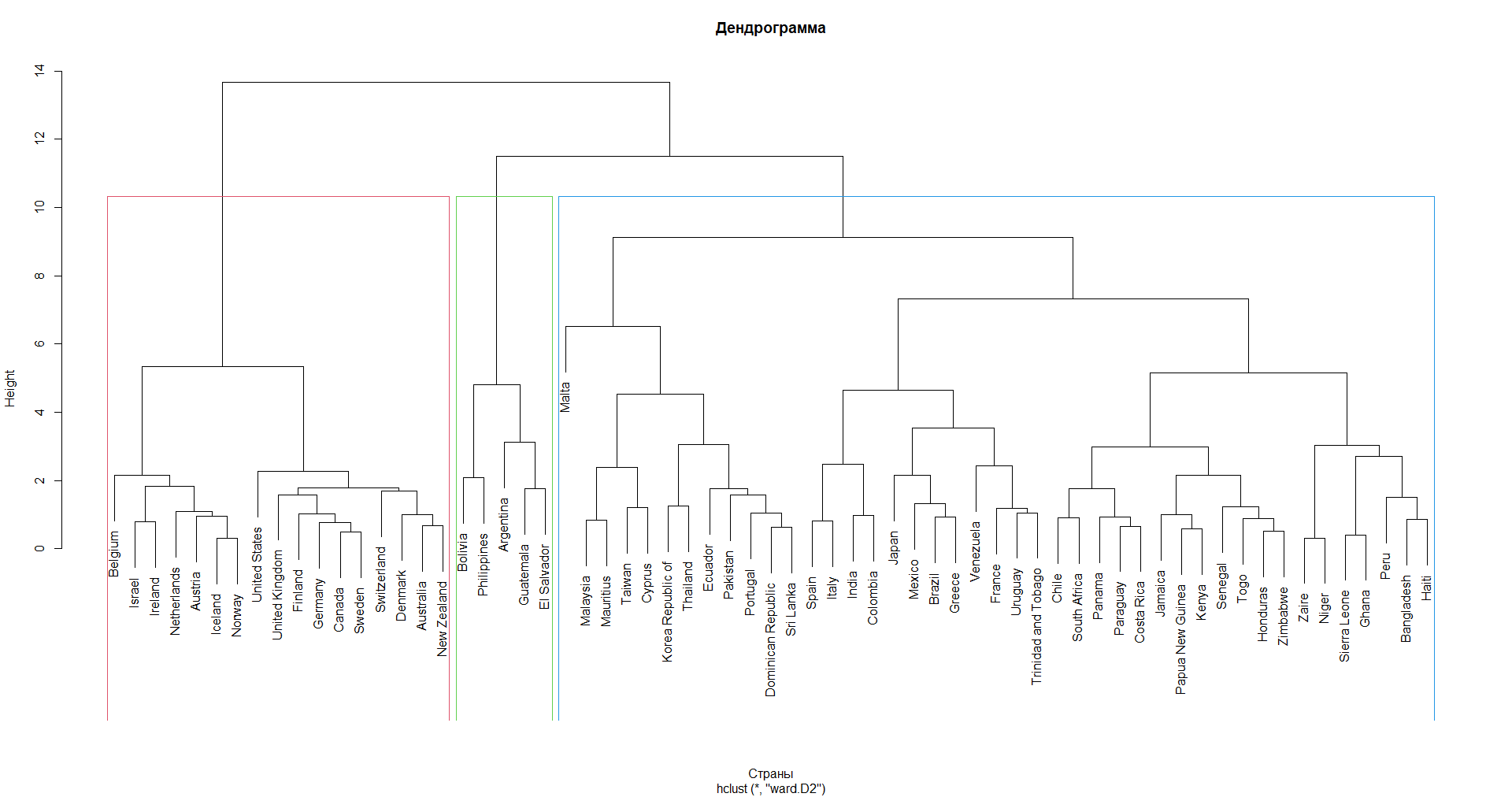
кластеризации. Из графика видно, что статистика зазора максимальна при семи кластерах.

1. Алгоритм консенсуса.



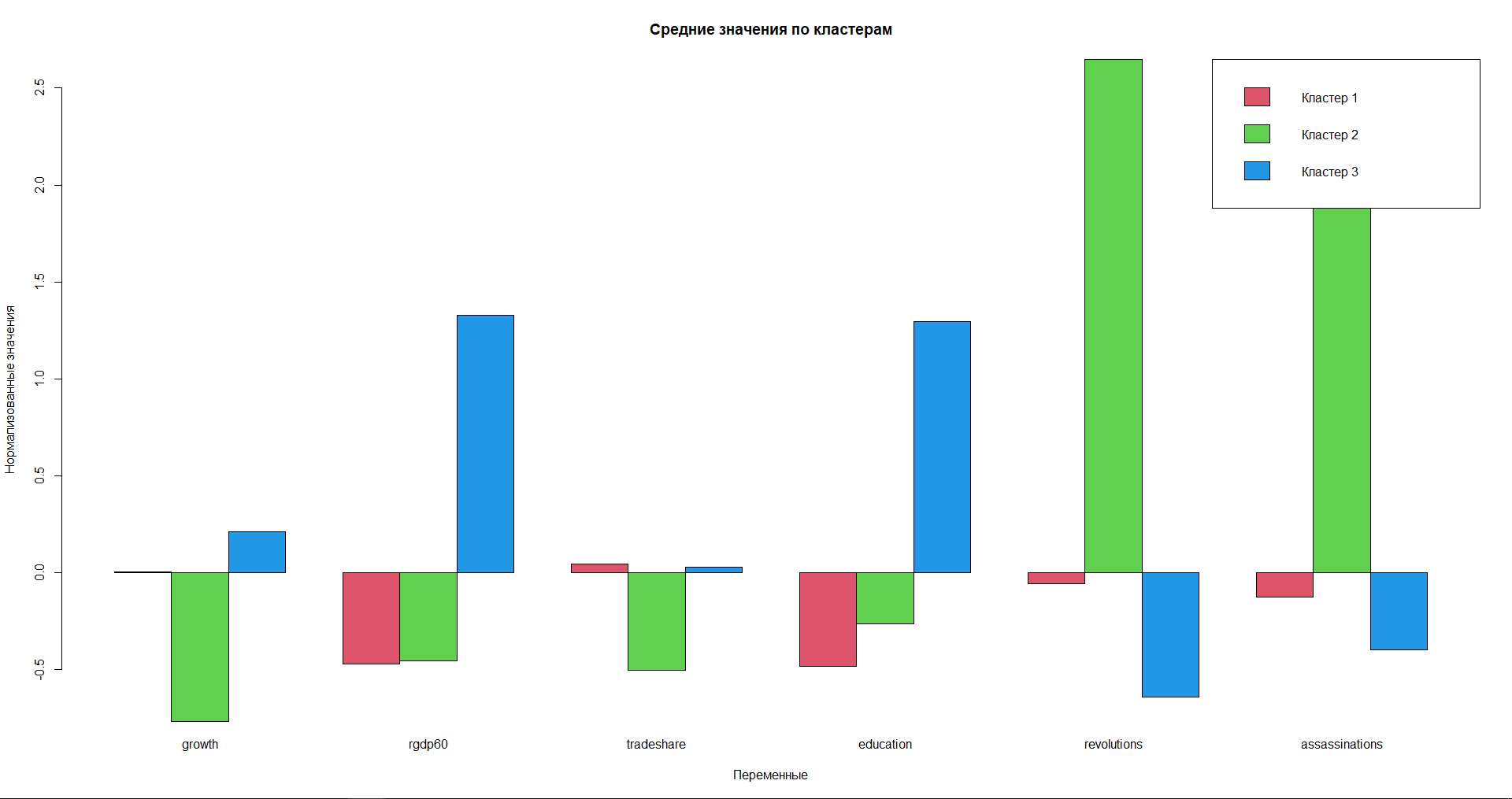
Данный график показывает оптимальное количество кластеров используя множество методов, которые согласуются между собой. В данном случае лучшее количество кластеров – 3.

1. Дендрограмма.

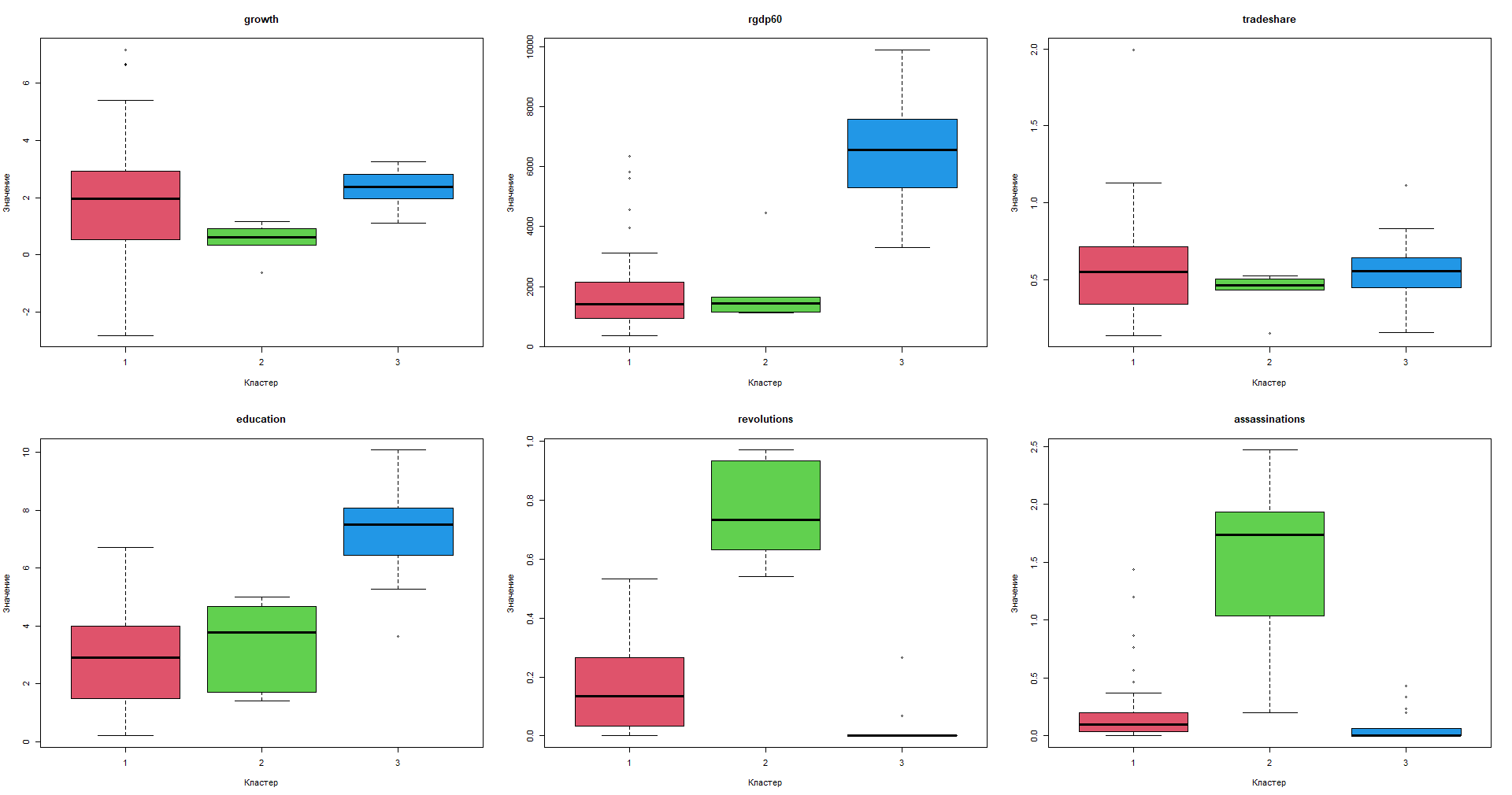


Дендрограмма показывает как группируются объекты (т.е. какие объекты похожи, а какие отличаются). Чем выше место объединения двух ветвей, тем больше различие между ними.

1. Столбчатые диаграммы и боксплоты.

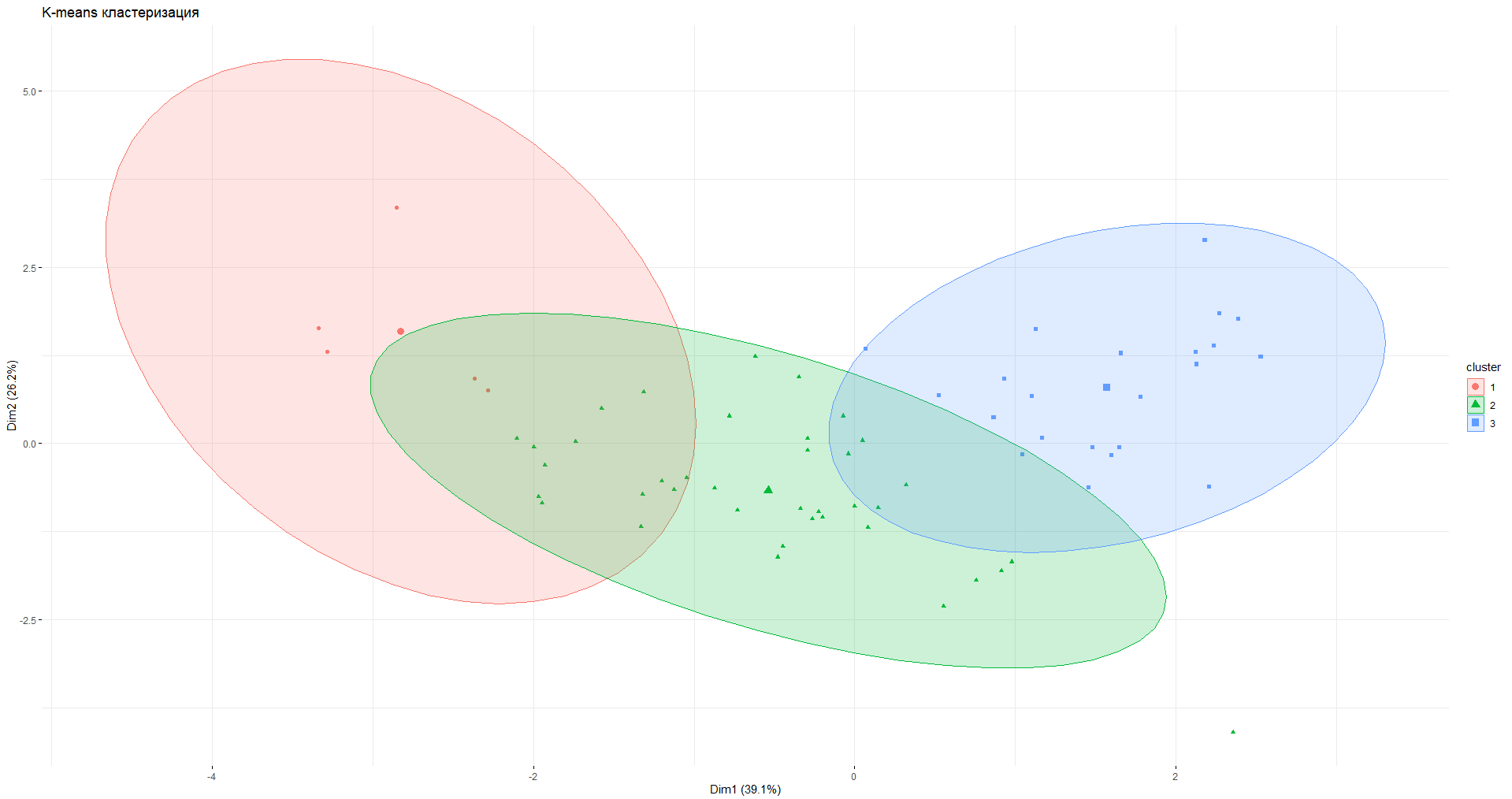


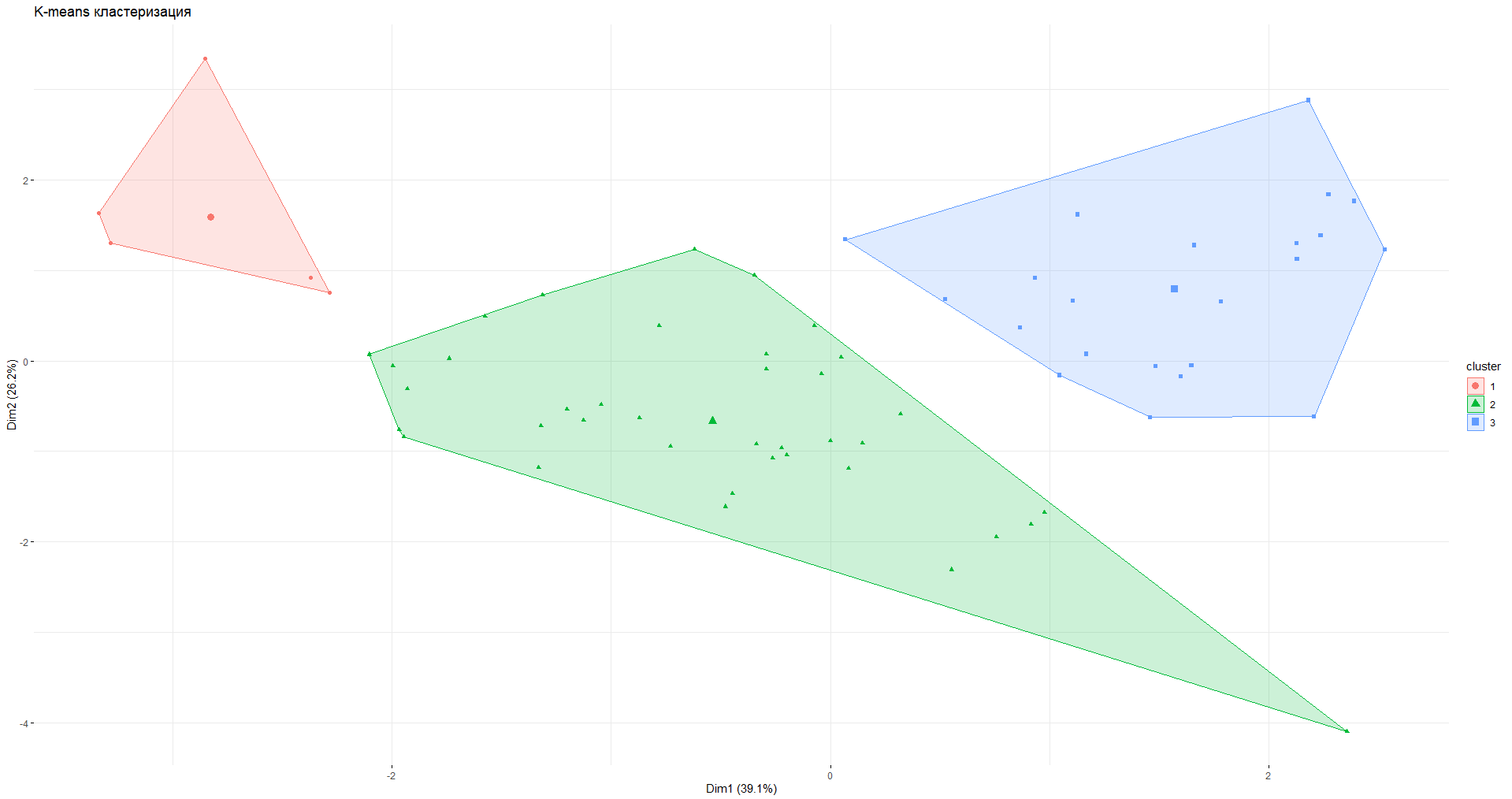
Данный график показывает средние значения столбцов для всех кластеров. В первом кластере выделены в отрицательную сторону rgdp60 и education. Во втором кластере очень сильно выделены в положительную сторону revolutions и assassinations, остальные параметры – отрицательные. В третьем кластере выделяются rgdp60 и education в положительную сторону, а revolutions и assassinations – в отрицательную.



Как можно увидеть из диаграммы, в третьем кластере самые высокие показатели growth, rgdp60, education и tradeshare. Второй кластер лидирует по revolutions и assassinations.

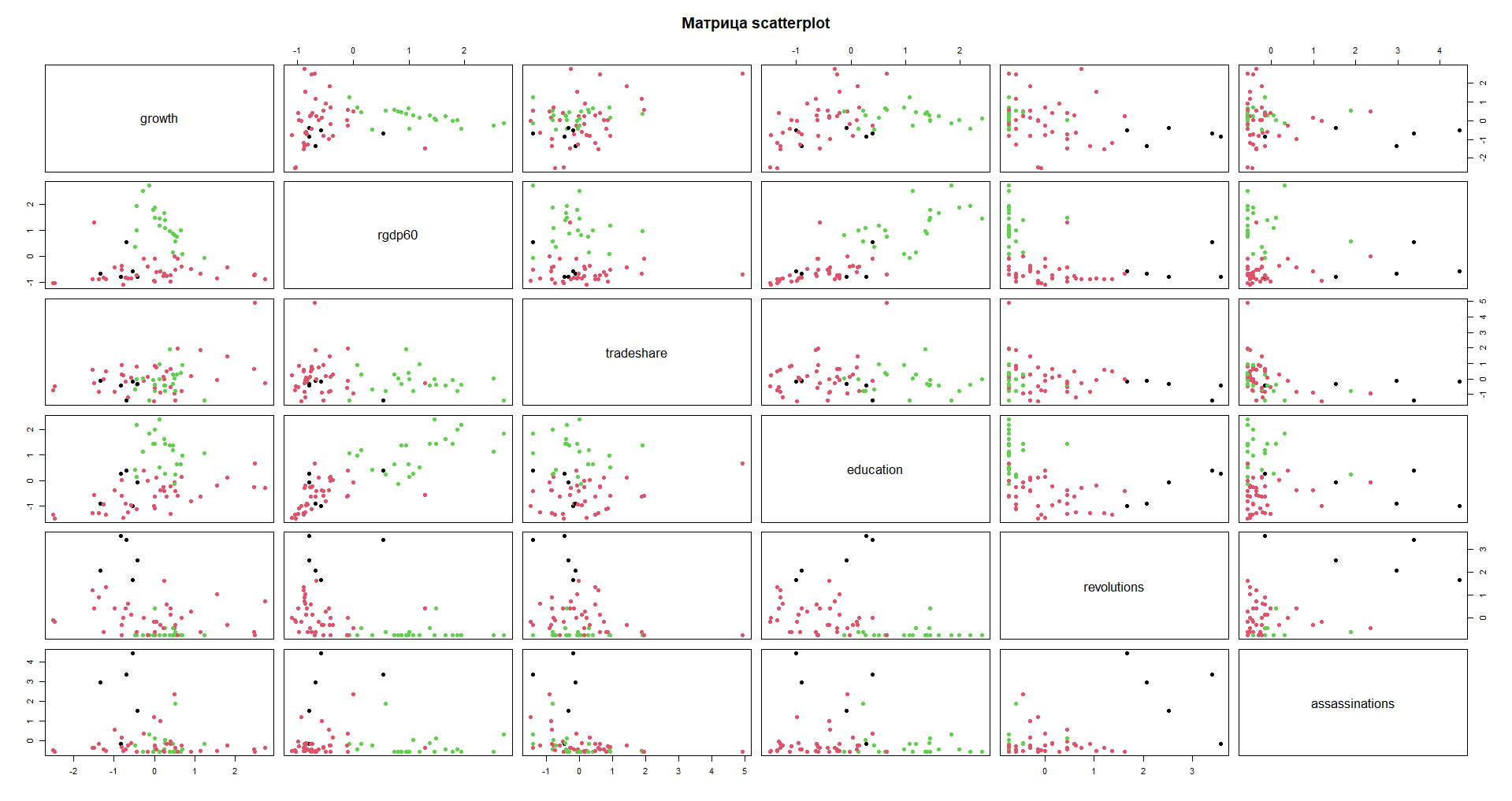
1. K-means кластеризация





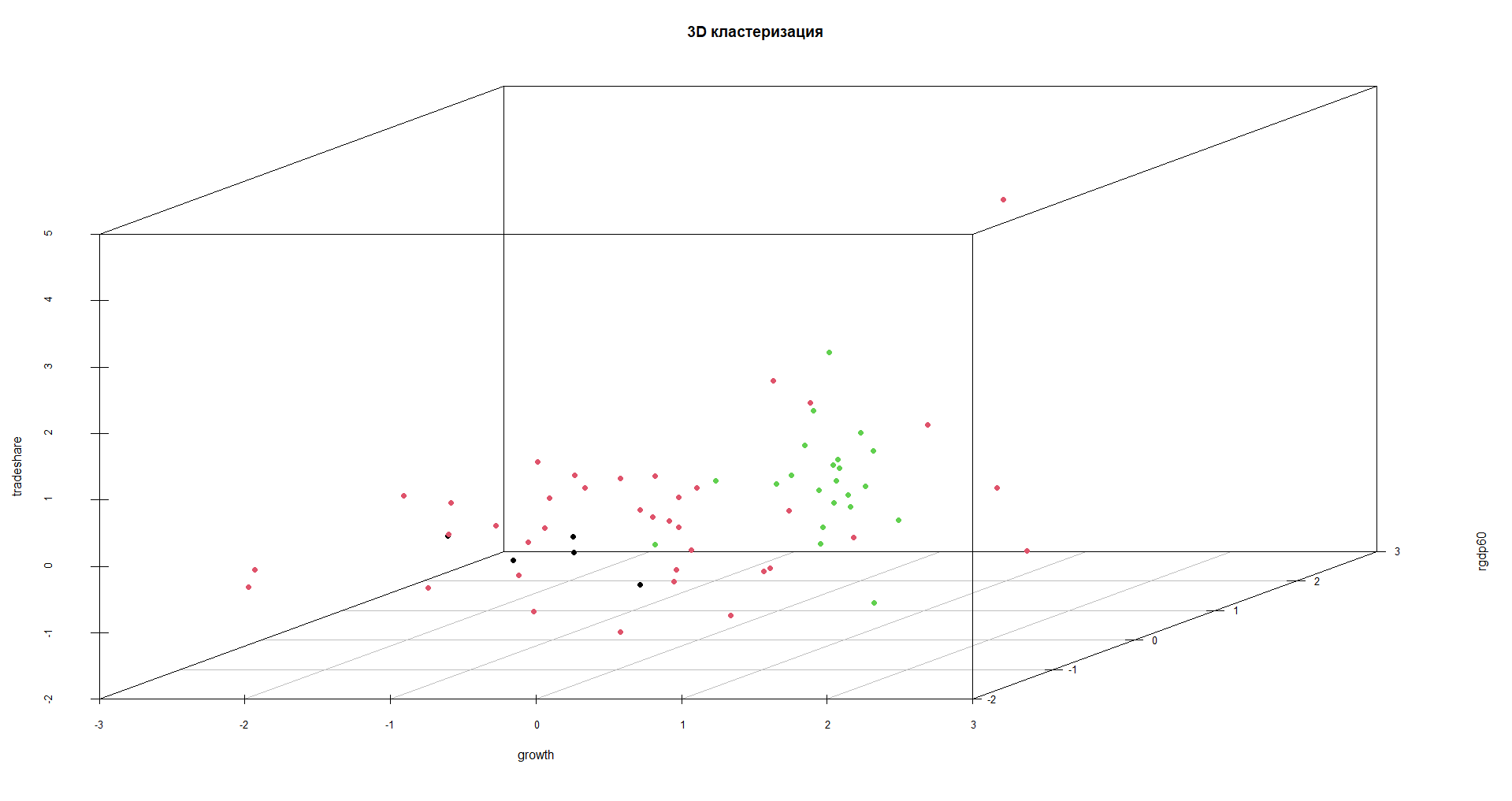
Эти диаграммы показывают влияние кластеров друг на друга. График без эллипсов показывает, что данные чётко разделены по кластерам, но на другом графике с эллипсами есть пересечения, это говорит о выбросах и о большой дисперсии.

1. Scatterplot.



Матрица диаграммы рассеяния, представляет собой сетку диаграмм рассеяния, которая отображает попарные связи между несколькими переменными в наборе данных. Пары признаков, на которых видно чёткое разделение по кластерам, можно считать наиболее информативными. На графике можно увидеть выбросы, они сильно отклоняются от основного кластера.

1. 3D визуализация.



Данная диаграмма позволяет увидеть насколько хорошо кластеры отделены в трёх признаках сразу.

**Вывод**

В ходе выполнения лабораторной работы была проведена работа с алгоритмами классификации и кластеризации данных и функциями, которые осуществляют этот анализ.