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

по дисциплине «Обработка больших данных»

Тема**:** Задачи классификации и кластеризации

Выполнил: ст. гр. 36/2

Бачурин И.А.

Проверил: преподаватель

Яхонтов А.А.

Краснодар

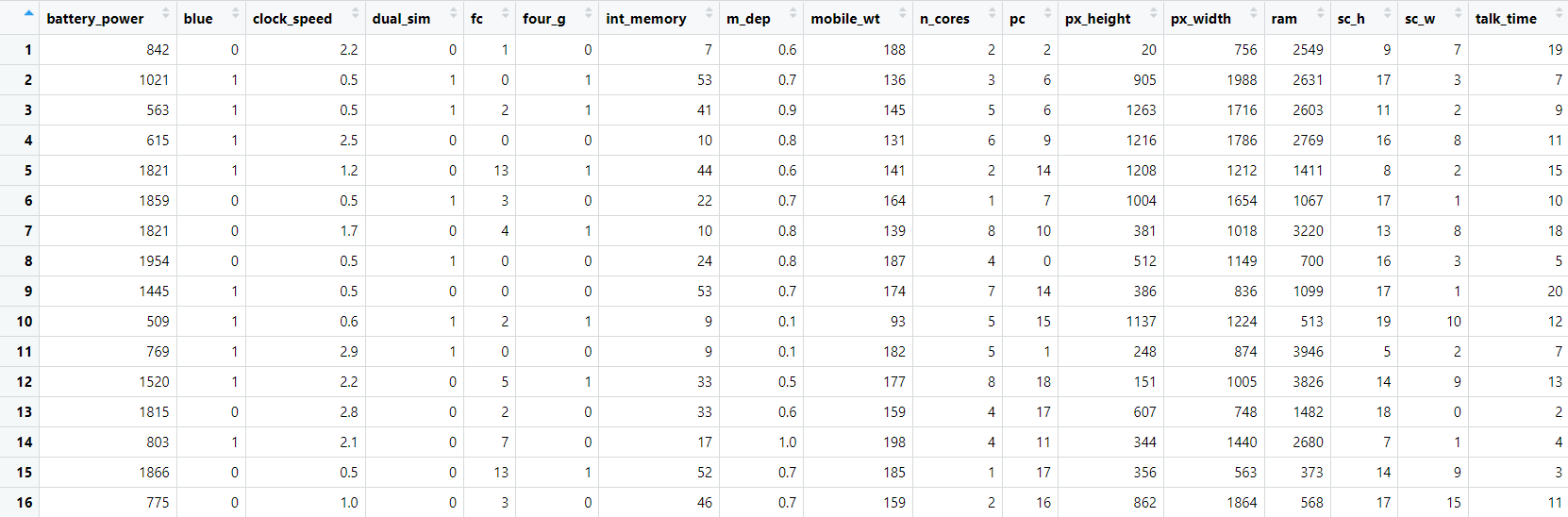
2025

**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.

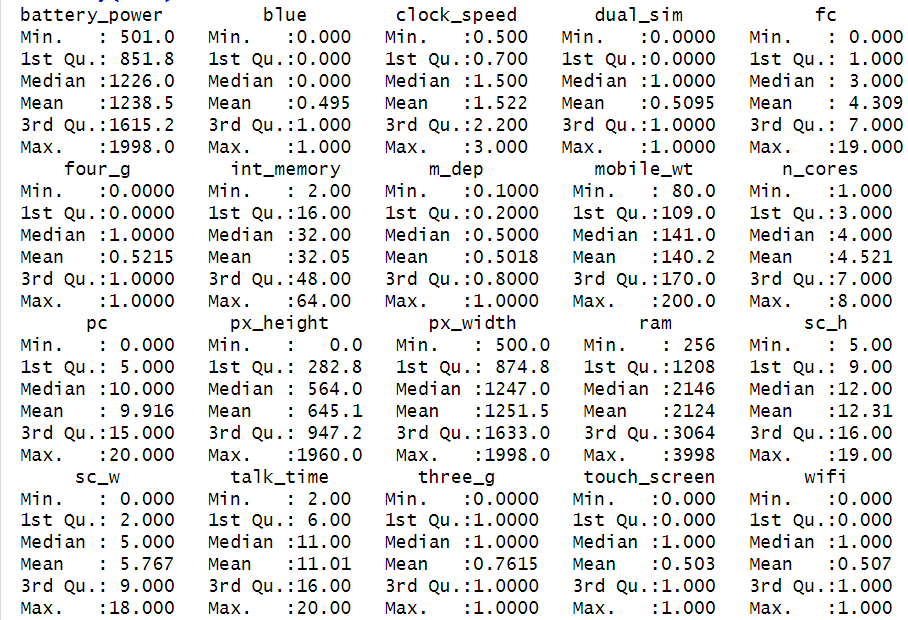
1. Выполнить дескриптивный анализ данных (здесь приветствуются дополнительные исследования).
2. Оценить оптимальное число кластеров, для этого построить диаграмму "Метод силуэта", “Метод локтя”, "Статистику разрыва" и Алгоритм консенсуса.
3. Выполнить иерархическую кластеризацию вашего набора данных, построив **дендрограмму**. Подробно обосновать Ваш выбор числа групп.
4. Построить диаграмму со столбчатыми диаграммами и боксплотами групп. Провести сравнительный анализ полученных групп.
5. Выполнить кластеризацию своего датасета по k-means.
6. Выполнить построение scatterplot с помощью функций plot или pairs.
7. Построить трехмерную кластеризацию по scatterplot3d.
8. В целом: выполнить шаги 1-3,5 анализа для своего набора данных (если какие-то из шагов нерелевантны вашему набору данных, объяснить почему).

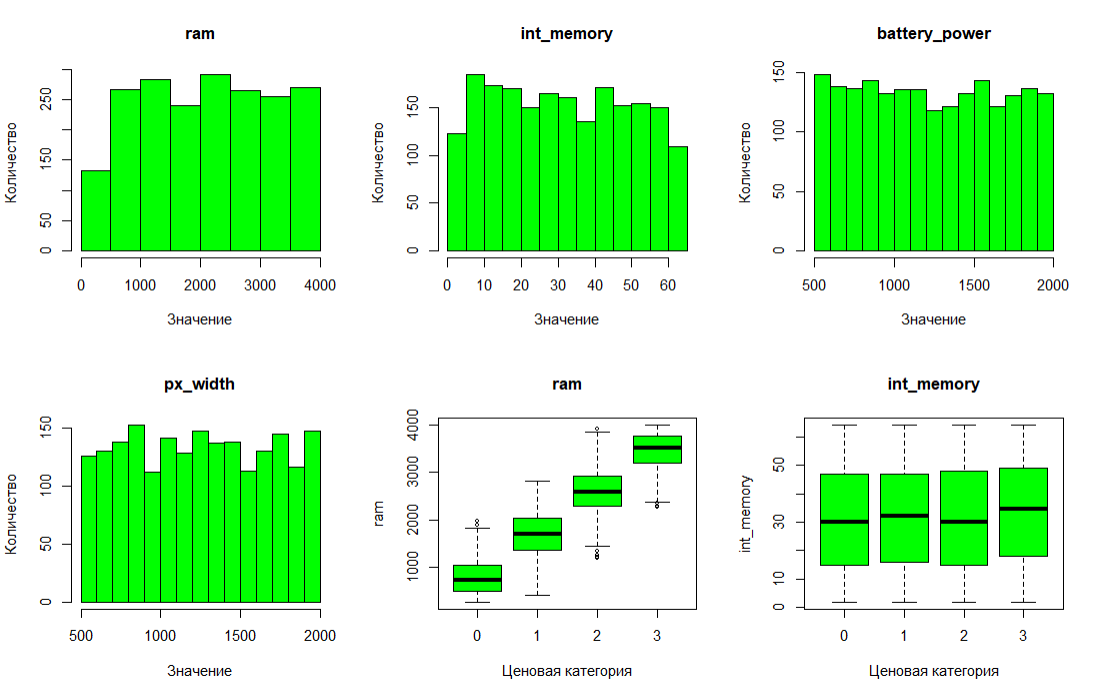
**Ход работы**

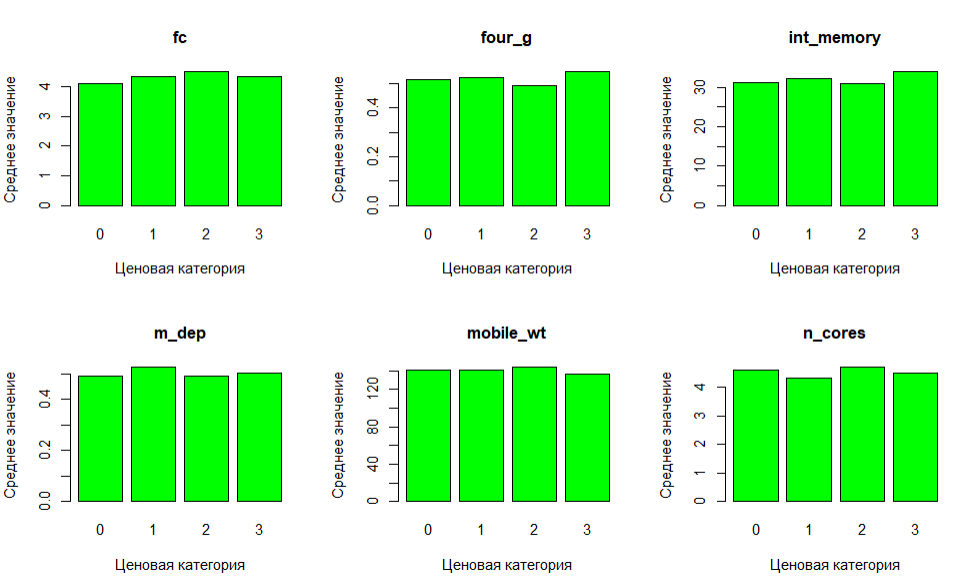
1. Набор данных при изначальном прочтении:



2. Выполним дескриптивный анализ данных:

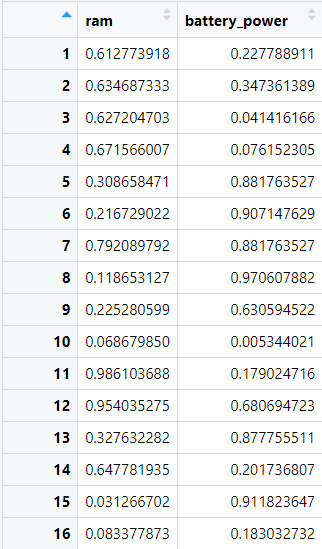




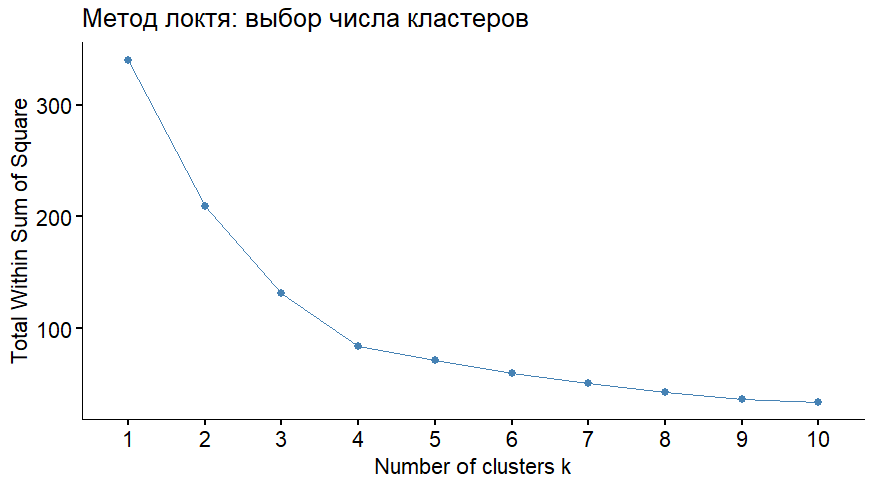


Здесь мы можем заметить какие именно параметры зависят от ценовой категории мобильного телефона. Большинство из них практически не отличаются в разных категориях. Уточним, что изначально датасет делят на 4 ценовые категории.

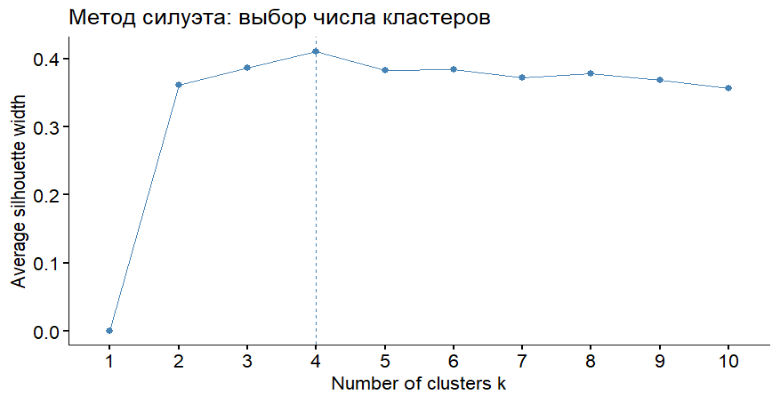
3. Далее была проведена нормализация (масштабирование) набора данных:



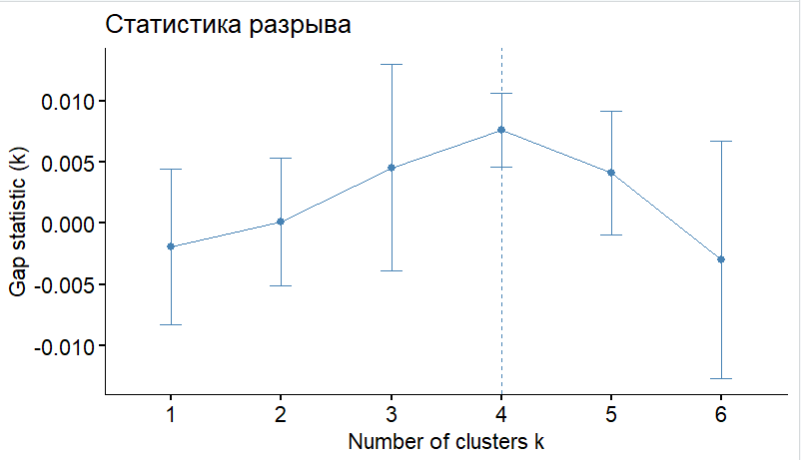
4. Оценим оптимальное число кластеров, для этого построим диаграмму "Метод силуэта", “Метод локтя”, "Статистику разрыва" и Алгоритм консенсуса:



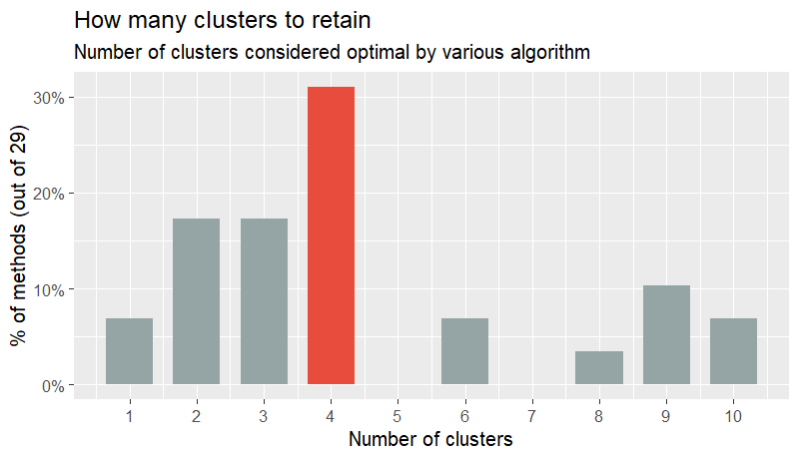
Метод локтя показывает оптимальное разбиение на 4 кластера



Метод силуэта также показал результат 4 кластера.

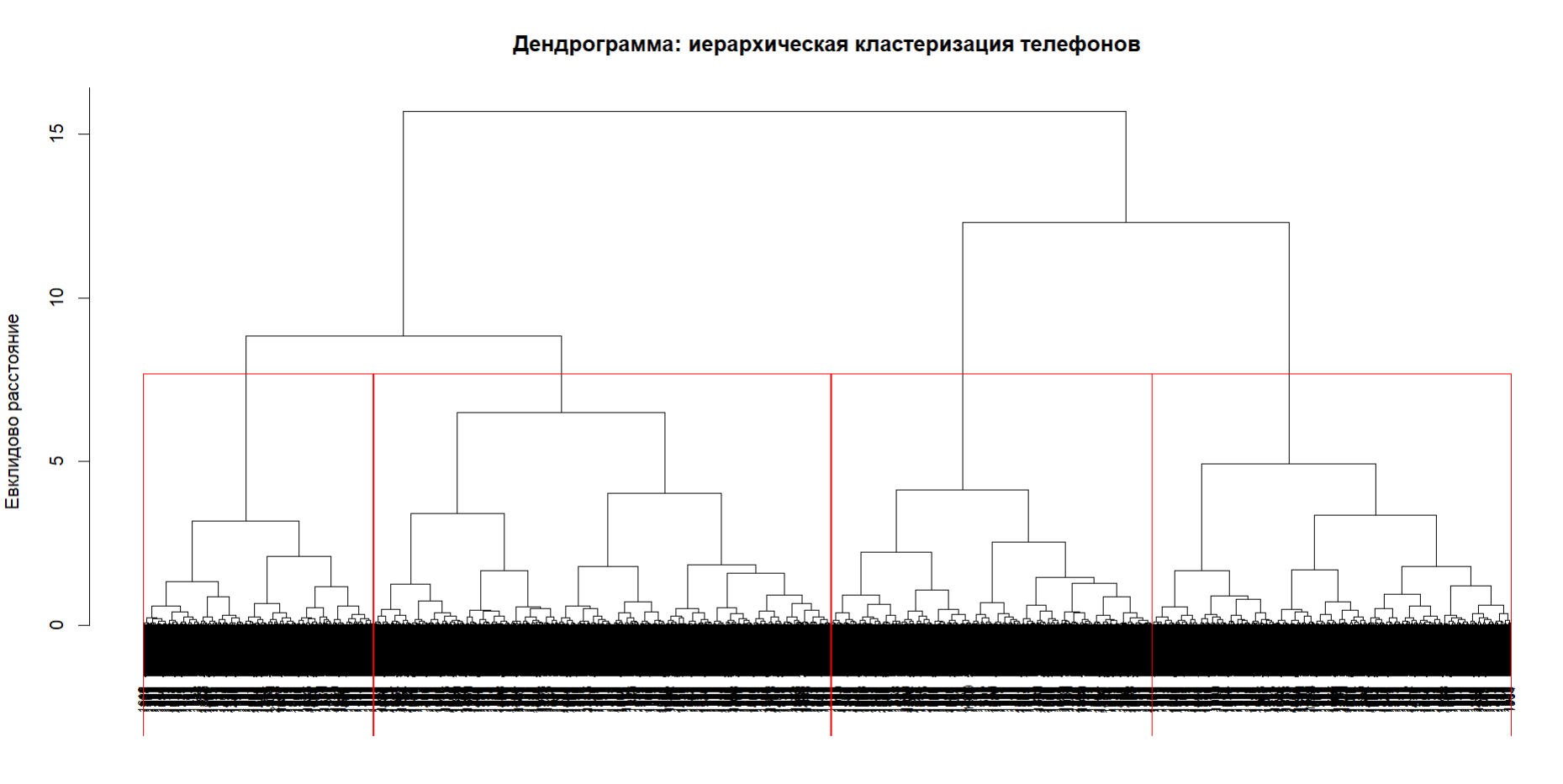


Тоже самое показал "Статистика разрыва"



Алгоритм консенсуса для уточнения результата.

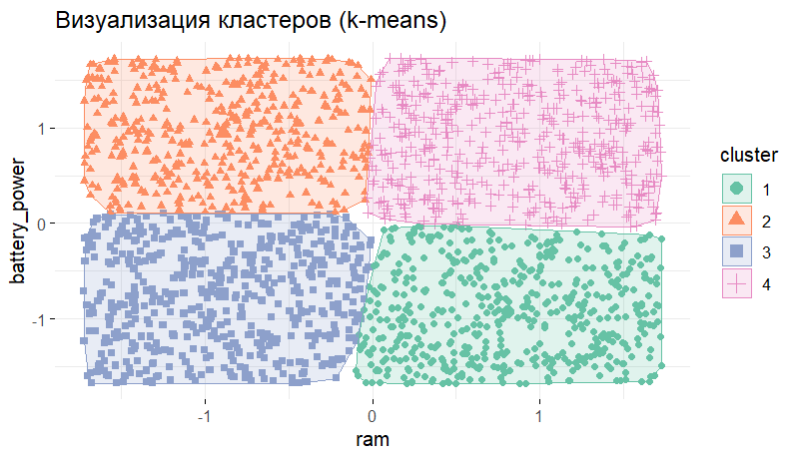
5. Разобьем набор данных на 5 кластеров и нарисуем дендрограмму:

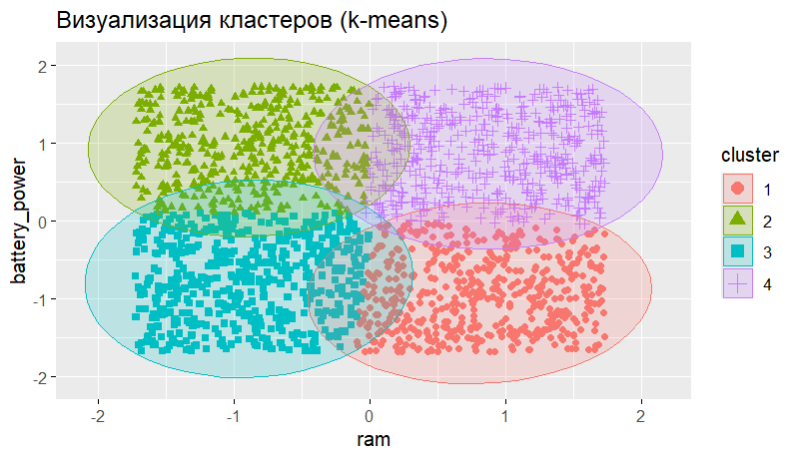


6. Построить диаграмму со столбчатыми диаграммами:

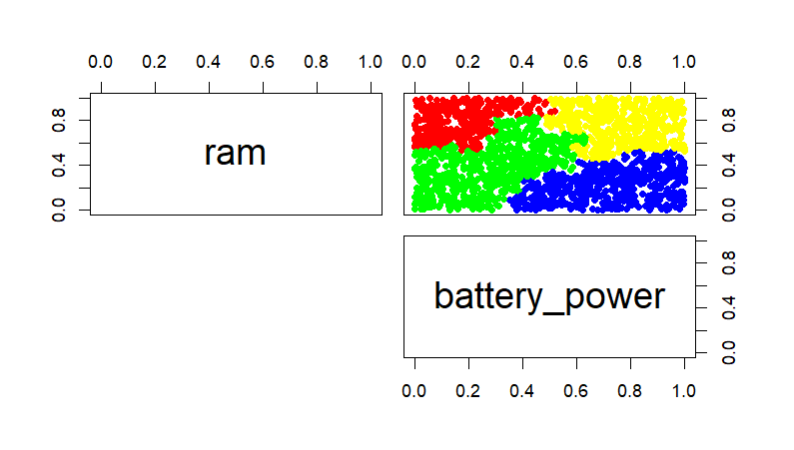


7. Выполним кластеризацию своего датасета по k-means:





8. Выполним построение scatterplot с помощью функций plot или pairs:



9. Построение трехмерной кластеризации по scatterplot3d:

