



Московский институт электроники и
математики им. А.Н. Тихонова

Кафедра информационной
безопасности киберфизических
систем

Москва 2025

Анализ данных с использованием кластерного анализа

Анализ данных с помощью кластерного анализа

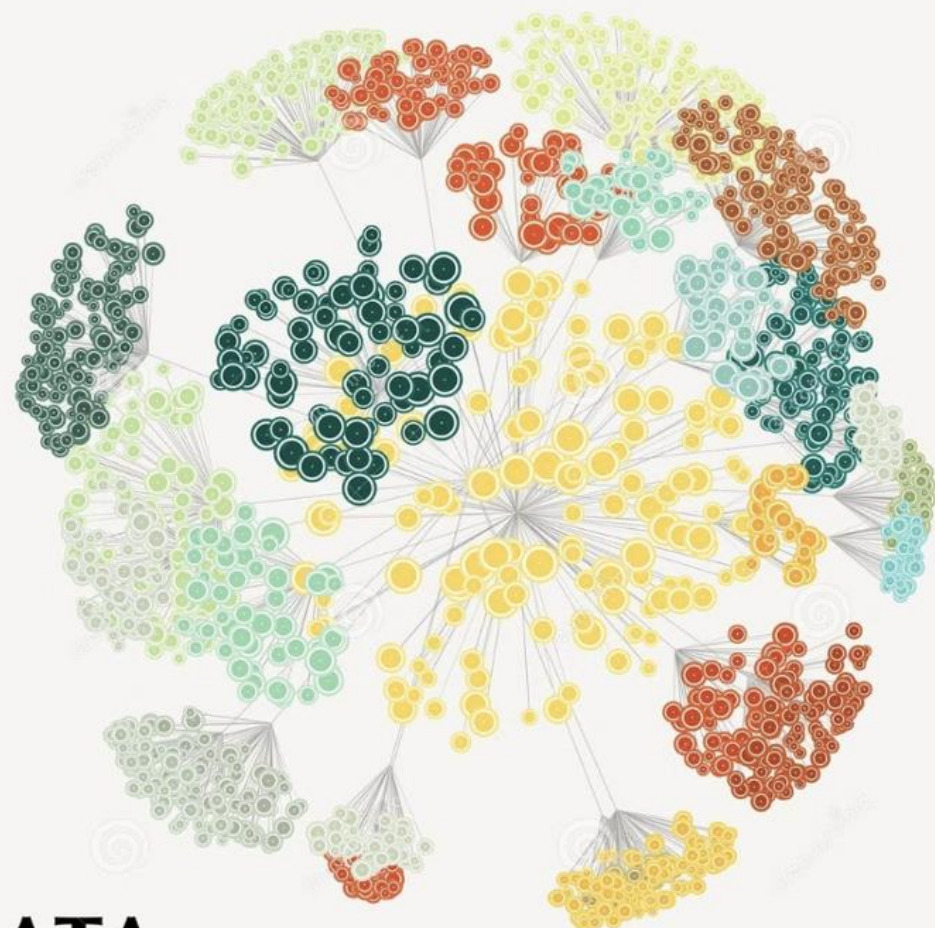
Цель: изучить методы кластерного анализа для выявления скрытых закономерностей и сегментации больших объёмов данных на примере анализа энергопотребления с использованием инструмента RapidMiner.





Задачи кластерного анализа

Кластерный анализ позволяет выявить группы похожих объектов в больших наборах данных, что важно для обнаружения закономерностей, сегментации рынка, оптимизации ресурсов и выявления аномалий.



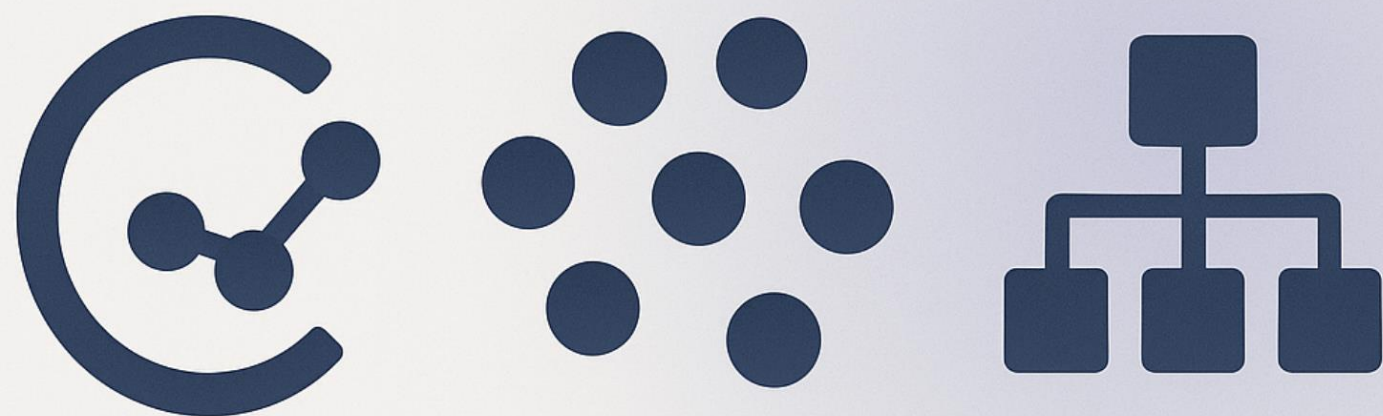
DATA
VISUALIZATION

Основные подходы кластеризации



Существуют различные подходы: иерархические методы, основанные на плотности (DBSCAN), и методы с заданным количеством кластеров, например, широко используемый метод K-Means.





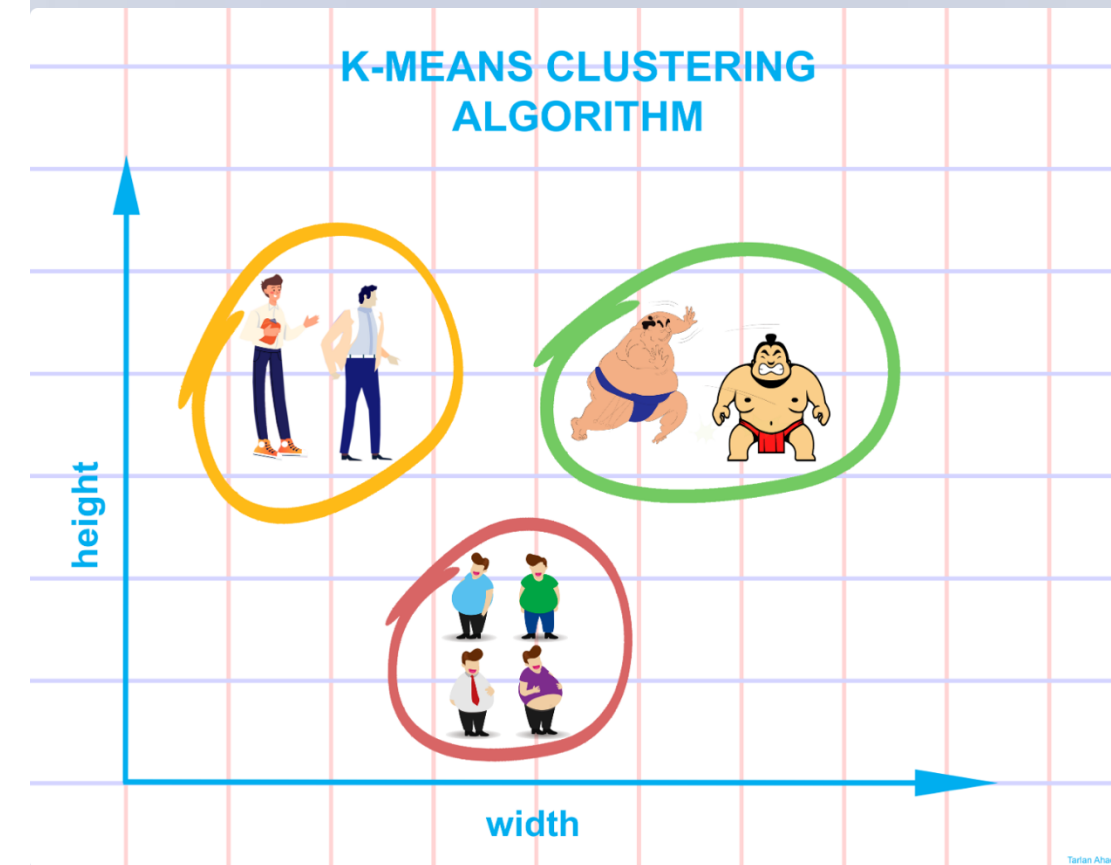
Выбор метода кластеризации

Выбор метода зависит от целей анализа: K-Means эффективен при больших объёмах данных, DBSCAN подходит для выявления аномалий и кластеров произвольной формы, иерархические — для небольших наборов.

Алгоритм K-Means: принцип работы

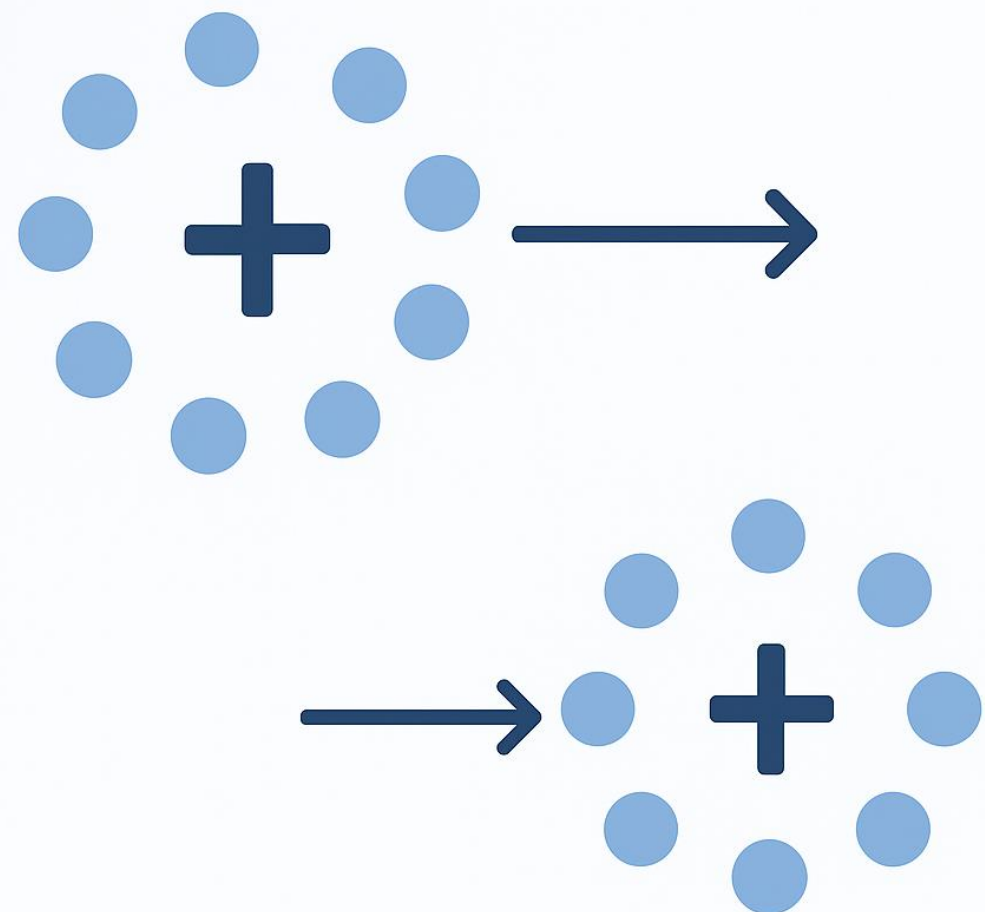
Алгоритм K-Means основан на минимизации внутрикластерной дисперсии.

Начальные центры кластеров выбираются случайно, затем объекты распределяются по ближайшим центрам.





Алгоритм K-Means: принцип работы



Центры кластеров
пересчитываются как средние
значения точек каждого
кластера.

Шаги повторяются, пока
центры кластеров перестают
существенно изменяться.

Преимущества и недостатки K-Means

Преимущества: высокая скорость работы, простота реализации, эффективность на больших выборках.

Недостатки: чувствительность к выбросам и необходимости предварительного задания числа кластеров.





Описание набора данных

Исследуется набор данных энергопотребления домохозяйств (более 2 млн записей).

Параметры: активная и реактивная мощность, напряжение, сила тока, распределение по подсистемам.

Community Samples (connected)

Import Data - Format your columns.

Format your columns.

Date format: HH:mm:ss 50% ☐ Replace errors with missing values

	Date <i>polynomial</i>	Time <i>time</i>	Global_acti... <i>real</i>	Global_reac... <i>real</i>	Voltage <i>real</i>	Global_inte... <i>real</i>	Sub_meteri... <i>real</i>	Sub_meteri.. <i>real</i>
1	16/12/2006	5:24:00 PM MSK	4.216	0.418	234.840	18.400	0.000	1.000
2	16/12/2006	5:25:00 PM MSK	5.360	0.436	233.630	23.000	0.000	1.000
3	16/12/2006	5:26:00 PM MSK	5.374	0.498	233.290	23.000	0.000	2.000
4	16/12/2006	5:27:00 PM MSK	5.388	0.502	233.740	23.000	0.000	1.000
5	16/12/2006	5:28:00 PM MSK	3.666	0.528	235.680	15.800	0.000	1.000
6	16/12/2006	5:29:00 PM MSK	3.520	0.522	235.020	15.000	0.000	2.000
7	16/12/2006	5:30:00 PM MSK	3.702	0.520	235.090	15.800	0.000	1.000
8	16/12/2006	5:31:00 PM MSK	3.700	0.520	235.220	15.800	0.000	1.000
9	16/12/2006	5:32:00 PM MSK	3.668	0.510	233.990	15.800	0.000	1.000
10	16/12/2006	5:33:00 PM MSK	3.662	0.510	233.860	15.800	0.000	2.000
11	16/12/2006	5:34:00 PM MSK	4.448	0.498	232.860	19.600	0.000	1.000
12	16/12/2006	5:35:00 PM MSK	5.412	0.470	232.780	23.200	0.000	1.000
13	16/12/2006	5:36:00 PM MSK	5.224	0.478	232.990	22.400	0.000	1.000
14	16/12/2006	5:37:00 PM MSK	5.268	0.398	232.910	22.600	0.000	2.000
15	16/12/2006	5:38:00 PM MSK	4.054	0.422	235.240	17.600	0.000	1.000
16	16/12/2006	5:39:00 PM MSK	3.384	0.282	237.140	14.200	0.000	0.000
17	16/12/2006	5:40:00 PM MSK	3.270	0.152	236.730	13.800	0.000	0.000
18	16/12/2006	5:41:00 PM MSK	3.120	0.156	237.060	14.400	0.000	0.000

no problems.

Previous Next Cancel

Этапы подготовки данных

Качество анализа зависит от предобработки: сначала удаляются или заполняются пропущенные значения, удаляются нечисловые признаки (дата и время), данные стандартизируются.



The screenshot displays the Orange3 software interface. The main workspace shows a workflow with three processes: 'Retrieve household_p...', 'Replace Missing Values', and 'Clustering'. The 'Replace Missing Values' process is selected, and its parameters are shown in the right-hand pane. The 'attribute filter' is set to 'all', 'invert selection' is unchecked, 'include special attributes' is unchecked, and the 'default' replacement method is set to 'average'. A list of columns is shown at the bottom of the parameters pane. An 'Edit Parameter List: columns' dialog box is open in the foreground, showing a list of attributes and their corresponding replacement functions. The 'replace with' dropdown is set to 'average'.

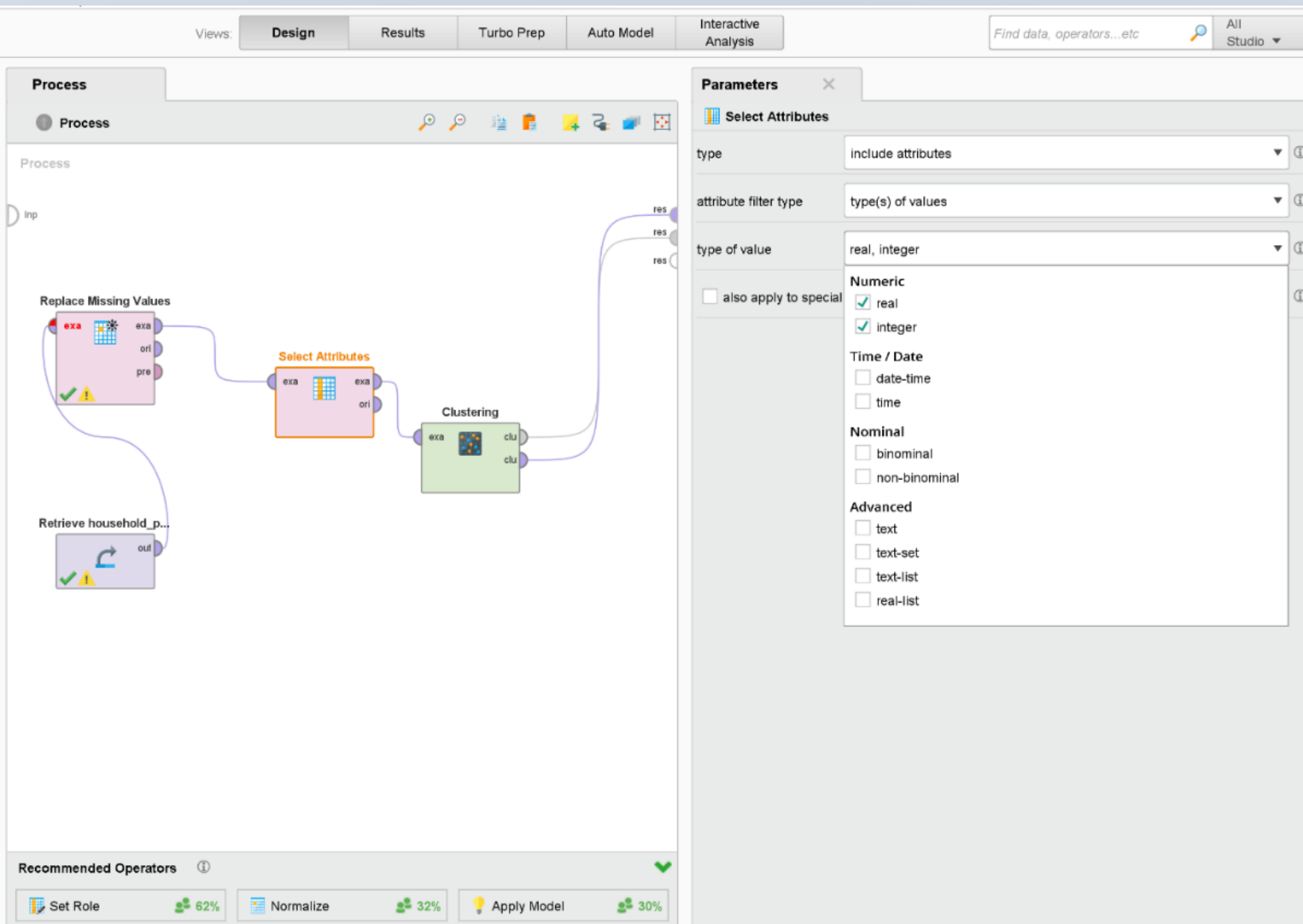
attribute	replace with
Enter value...	average
Date	
# Global_active_power	
# Global_intensity	
# Global_reactive_power	
# Sub_metering_1	
# Sub_metering_2	
# Sub_metering_3	
Time	



Этапы подготовки данных

Необходимо выбрать только числовые атрибуты.

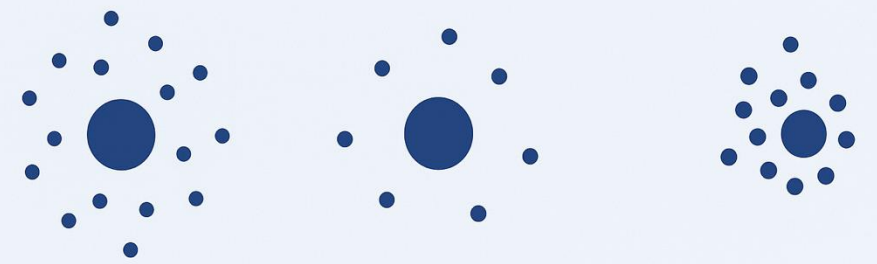
Это позволяет корректно использовать алгоритм K-Means, работающий исключительно с количественными данными.



Влияние предобработки на результаты кластеризации



Качество кластеризации напрямую связано с тщательностью предварительной подготовки данных: корректная обработка пропусков существенно улучшает разделение данных на кластеры.





Применение K-Means в RapidMiner

RapidMiner позволяет задать количество кластеров (параметр $k=3$). Полученные кластеры отражают различные режимы энергопотребления: основной, аномальный и пиковой нагрузки.

Parameters ✕

Clustering (k-Means)

☒ add cluster attribute ⓘ

☐ add as label ⓘ

☐ remove unlabeled ⓘ

k ⓘ

max runs ⓘ

☒ determine good start values ⓘ

measure types ⓘ


divergence ⓘ

max optimization steps ⓘ


☐ use local random seed ⓘ

Характеристика кластеров

Большинство записей попадает в основной кластер с низким энергопотреблением. Остальные два кластера отражают нетипичные ситуации и пиковые нагрузки, требующие дополнительного анализа.




Description



Folder View

Cluster Model

Cluster 0: 1969715 items
Cluster 1: 48996 items
Cluster 2: 56548 items
Total number of items: 2075259



Attribute	cluster_0	cluster_1	cluster_2
Global_active_power	0.945	3.657	3.983
Global_reactive_power	0.120	0.191	0.200
Voltage	241.020	237.793	237.190
Global_intensity	4.004	15.546	16.903
Sub_metering_1	0.111	0.816	36.598
Sub_metering_2	0.432	34.828	2.420
Sub_metering_3	6.212	10.796	11.277



Методы оценки качества кластеризации

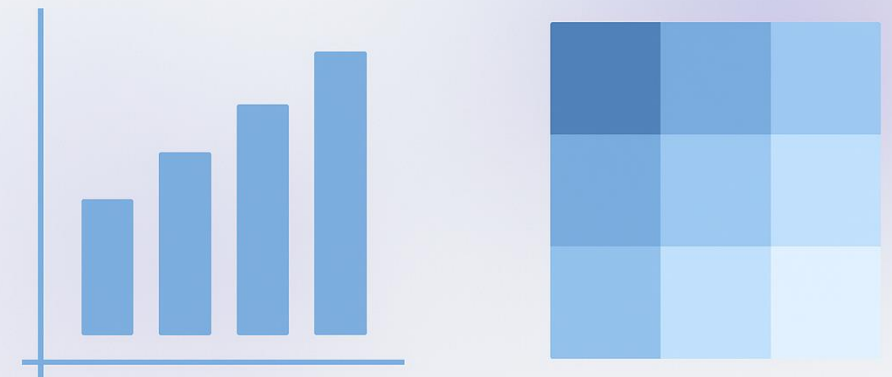


Оценка качества кластеров осуществляется через вычисление внутрикластерной и межкластерной дисперсии, силуэт-коэффициент (silhouette) и визуальный анализ распределений.

Визуализация результатов кластеризации

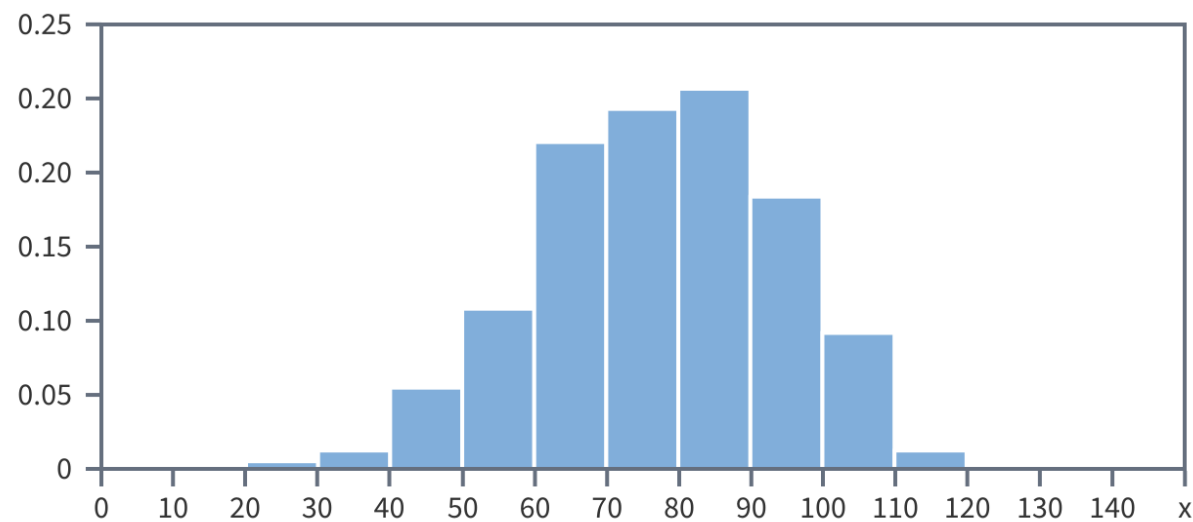
Визуализация помогает понять распределение объектов по кластерам.

Наиболее распространённые инструменты: гистограммы, диаграммы размаха, тепловые карты и корреляционные матрицы.



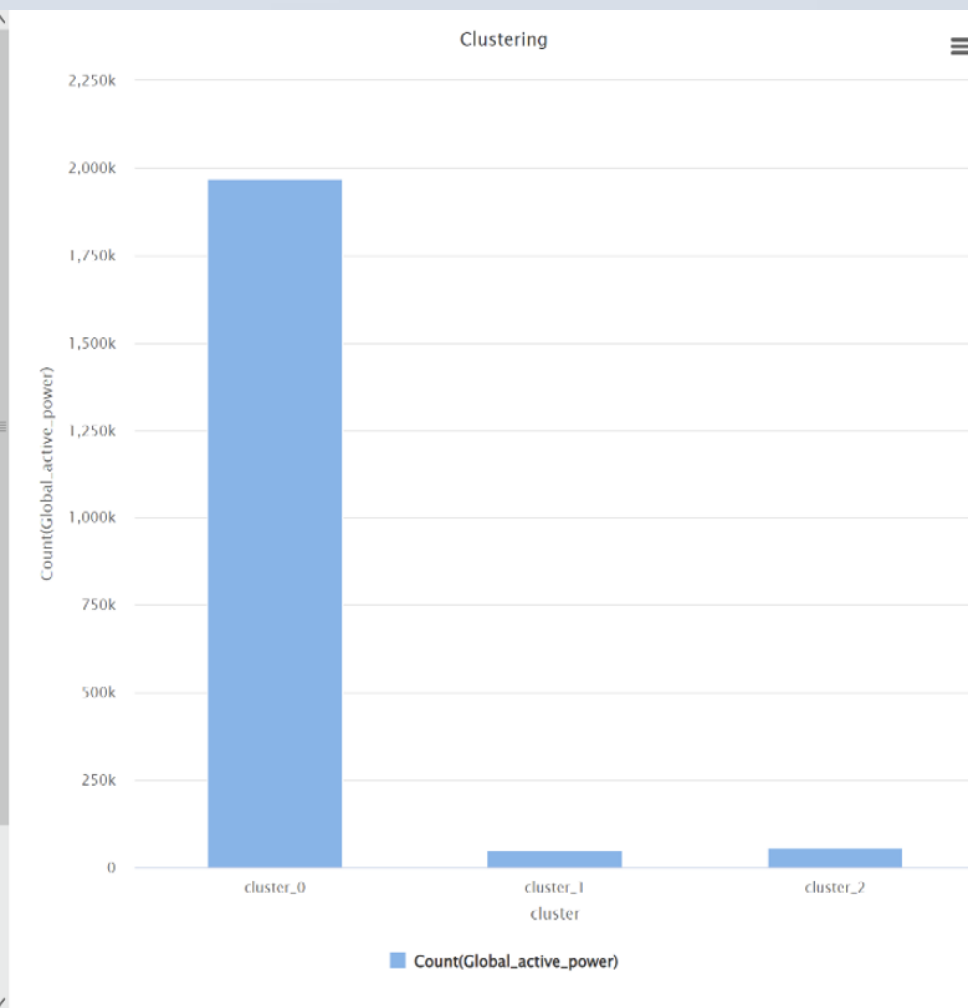


Роль гистограмм в кластерном анализе

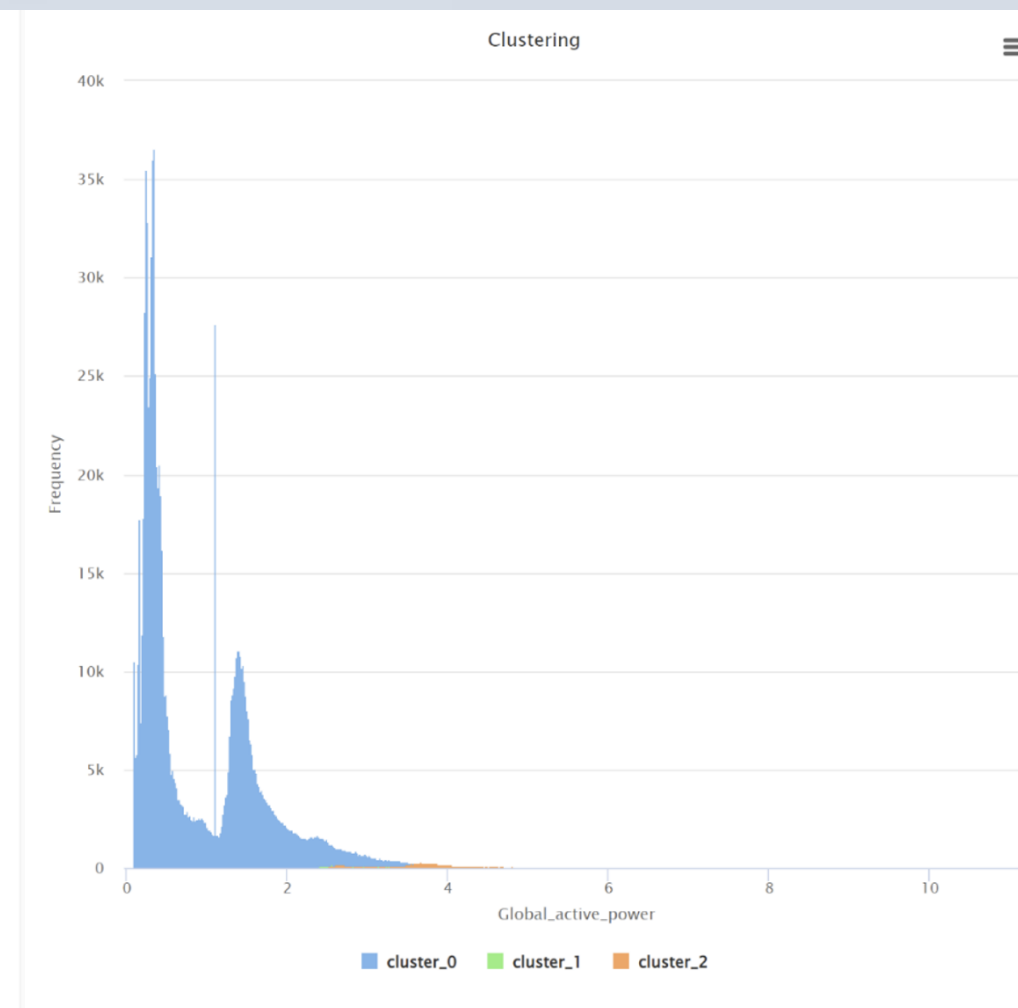


Гистограммы отображают распределение количественных признаков по кластерам, позволяя быстро оценить преобладание определённых значений, выявить тенденции и аномалии.

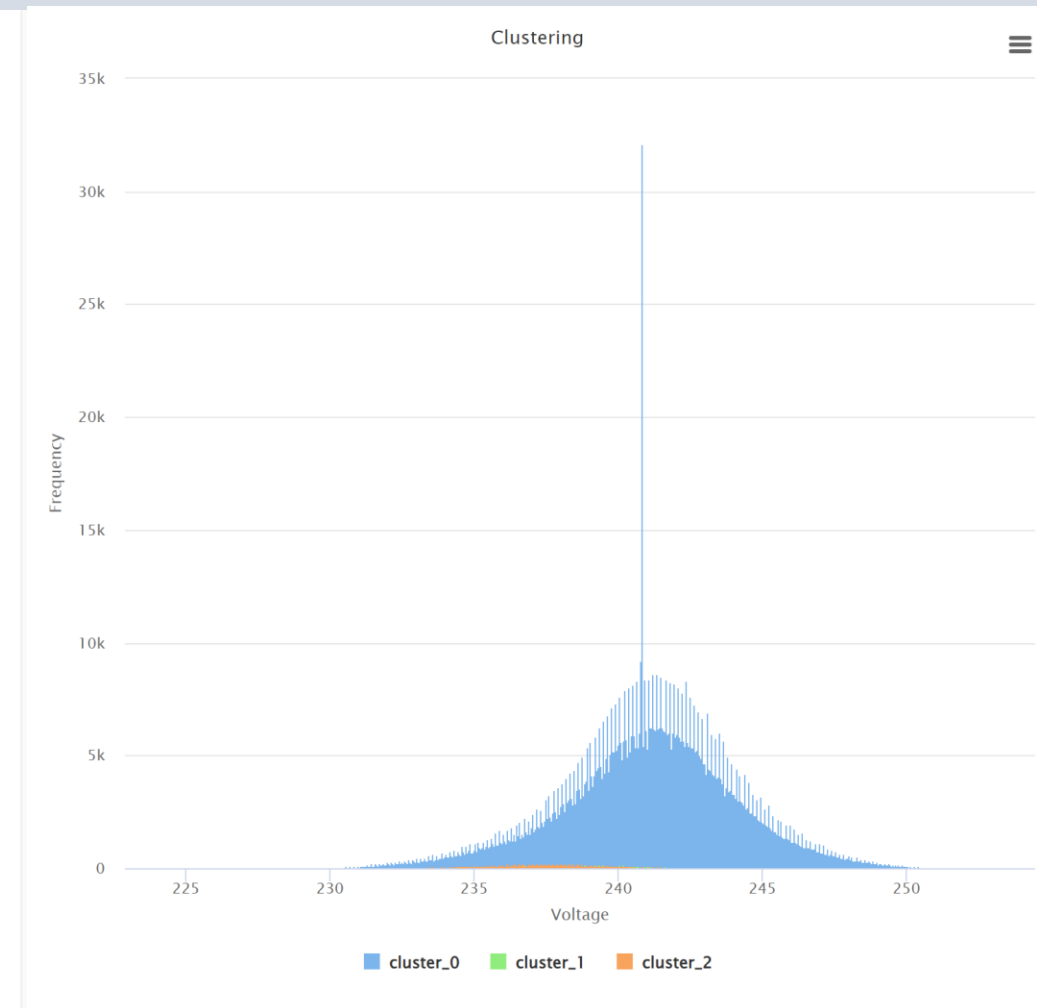
Гистограммы в рассматриваемом наборе данных



Распределение количества записей по кластерам



Гистограмма распределения мощности

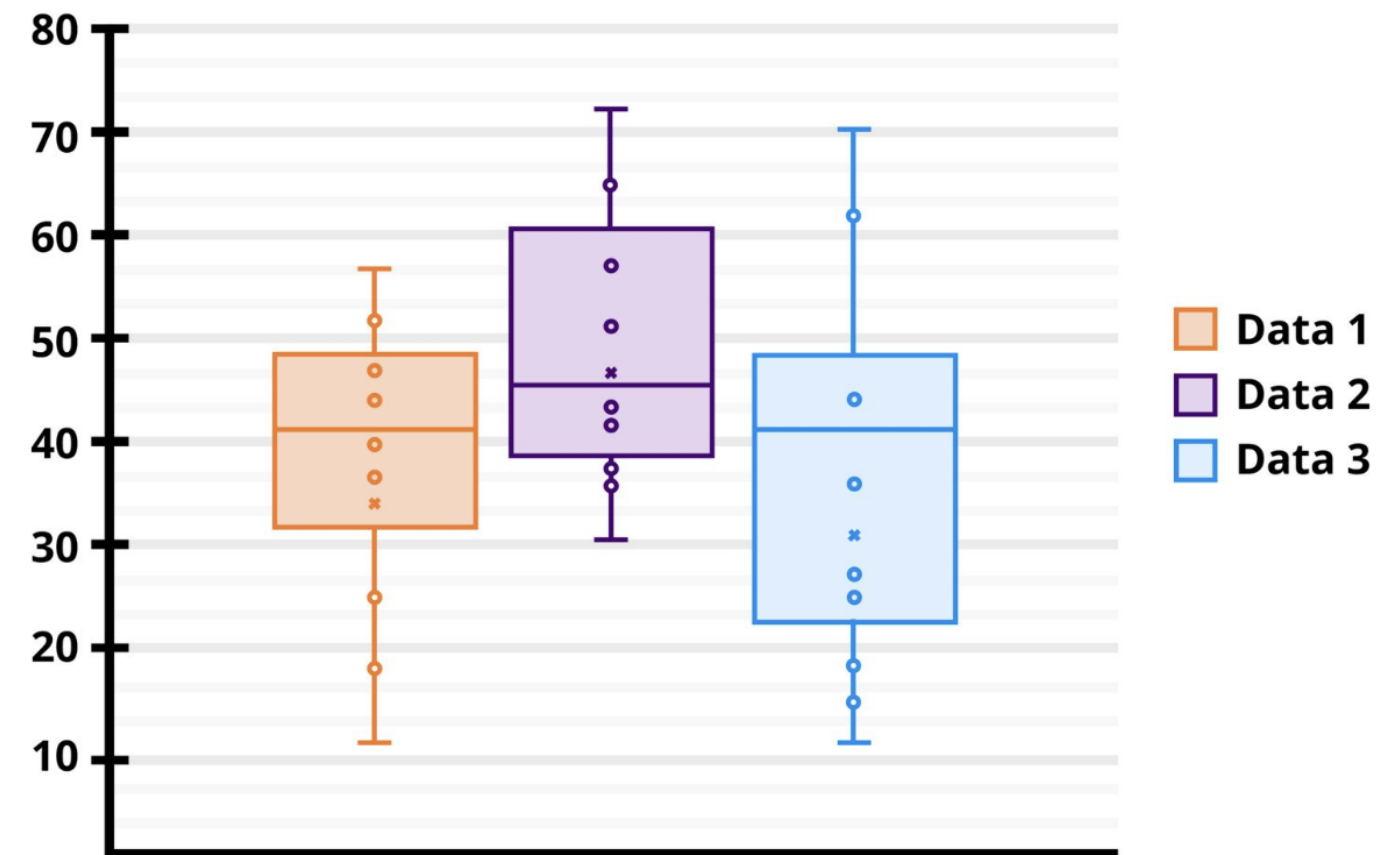


Гистограмма распределения напряжения

Диаграмма размаха (Box Plot)

Диаграмма размаха показывает медиану, минимальные и максимальные значения, квартили и выбросы, что позволяет оценить стабильность и вариативность параметров внутри кластеров.

Box plot



Диаграммы размаха в рассматриваемом наборе данных

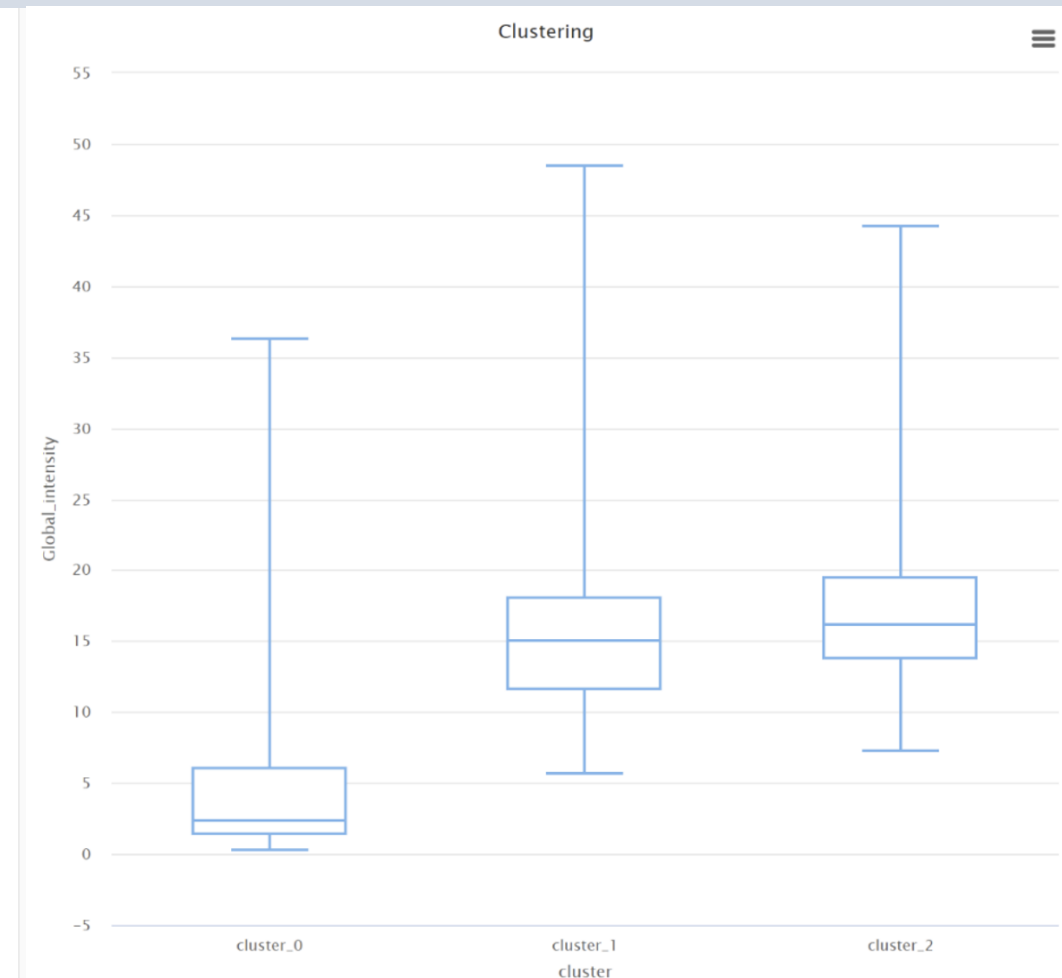
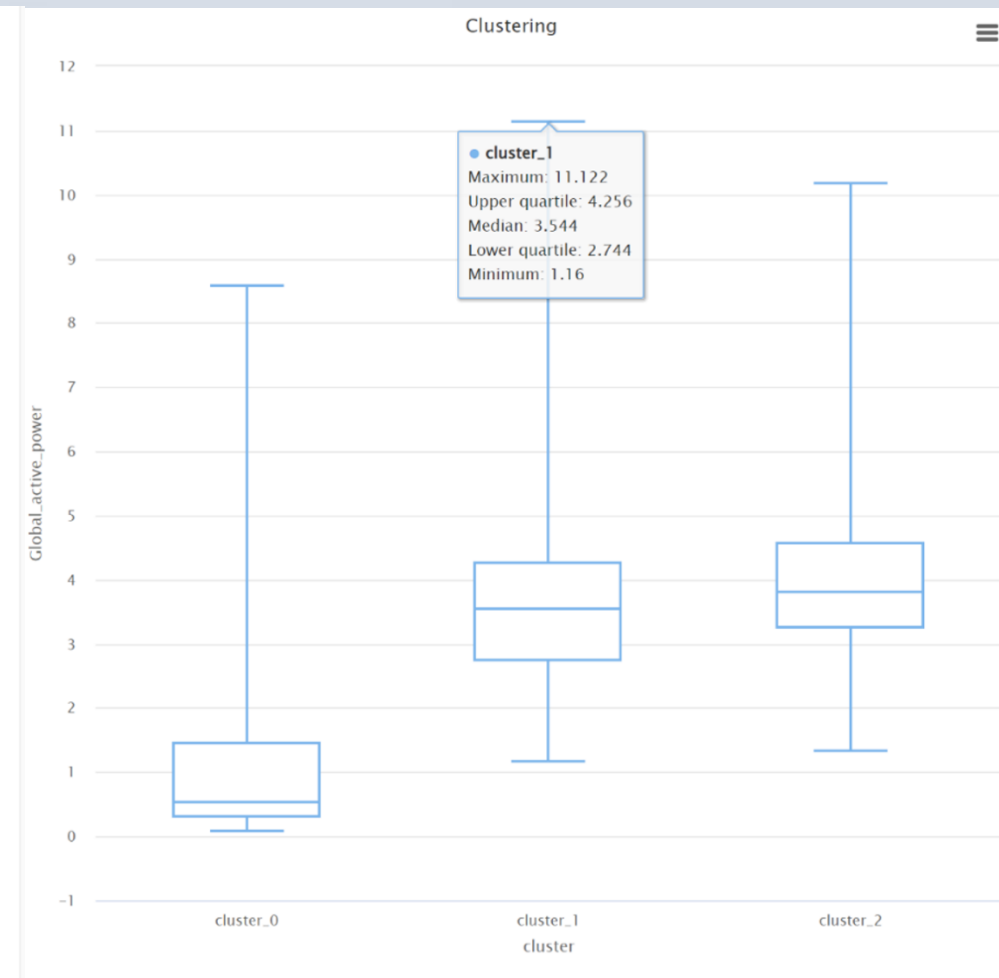
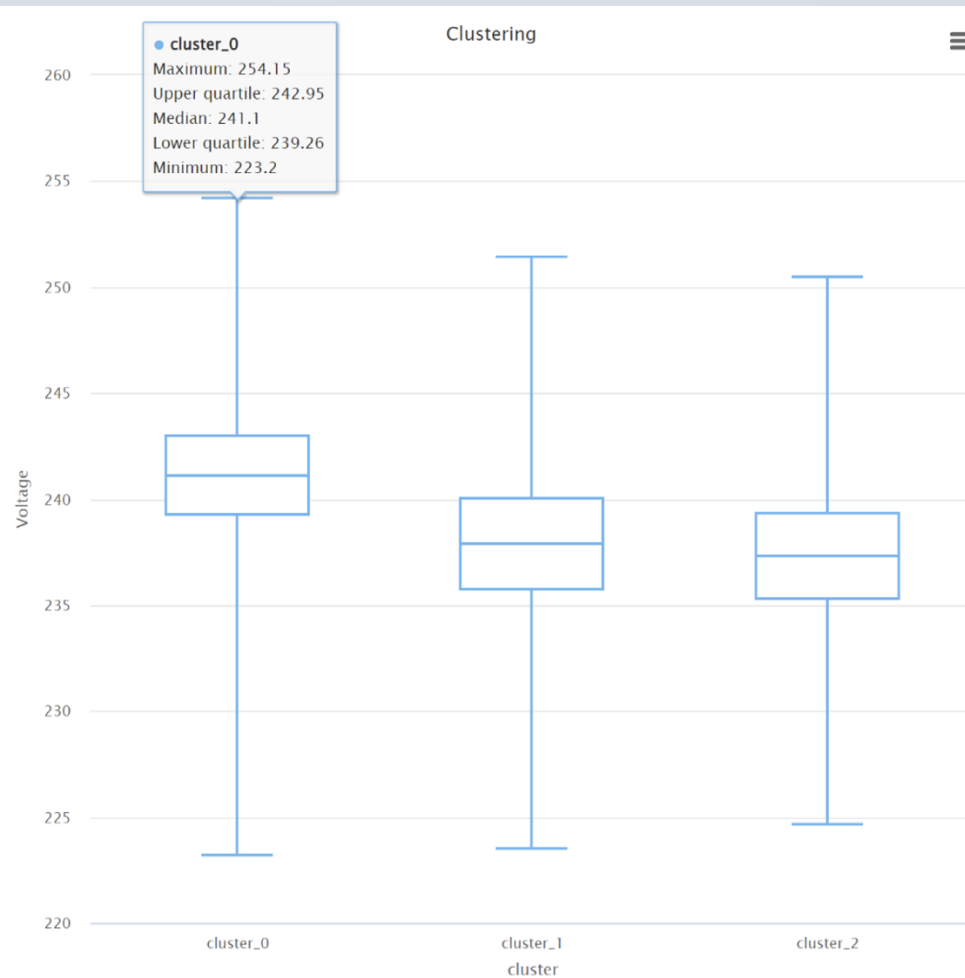


Диаграмма размаха
для напряжения

Распределения
значений активной
МОЩНОСТИ

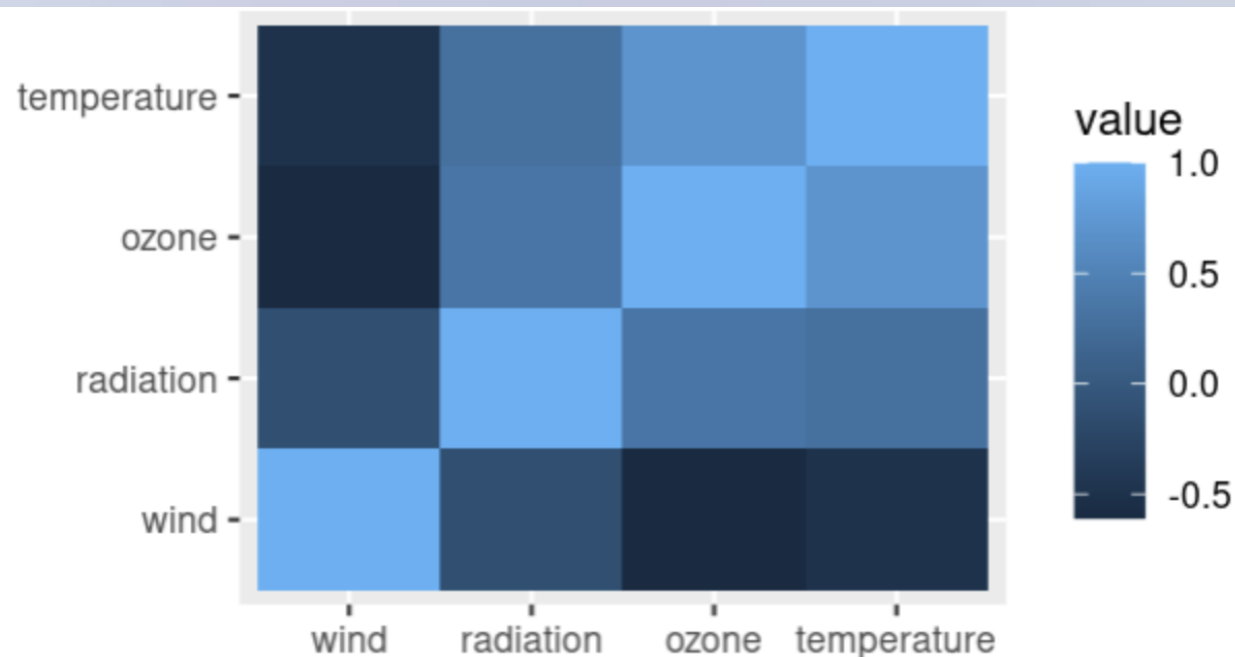
Распределения
интенсивности тока



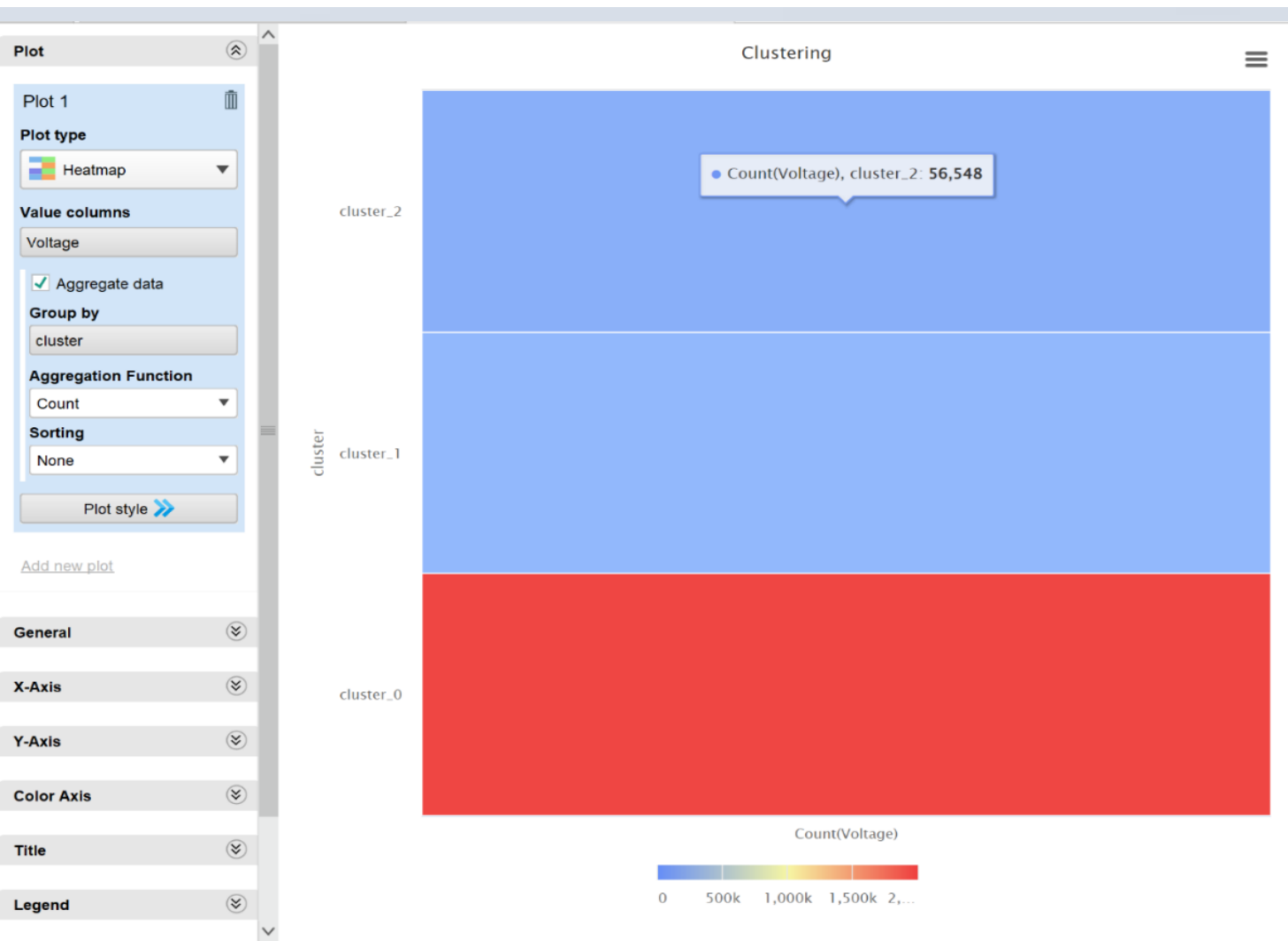
Тепловые карты для визуального анализа

Тепловые карты представляют корреляции между признаками через цветовую интенсивность.

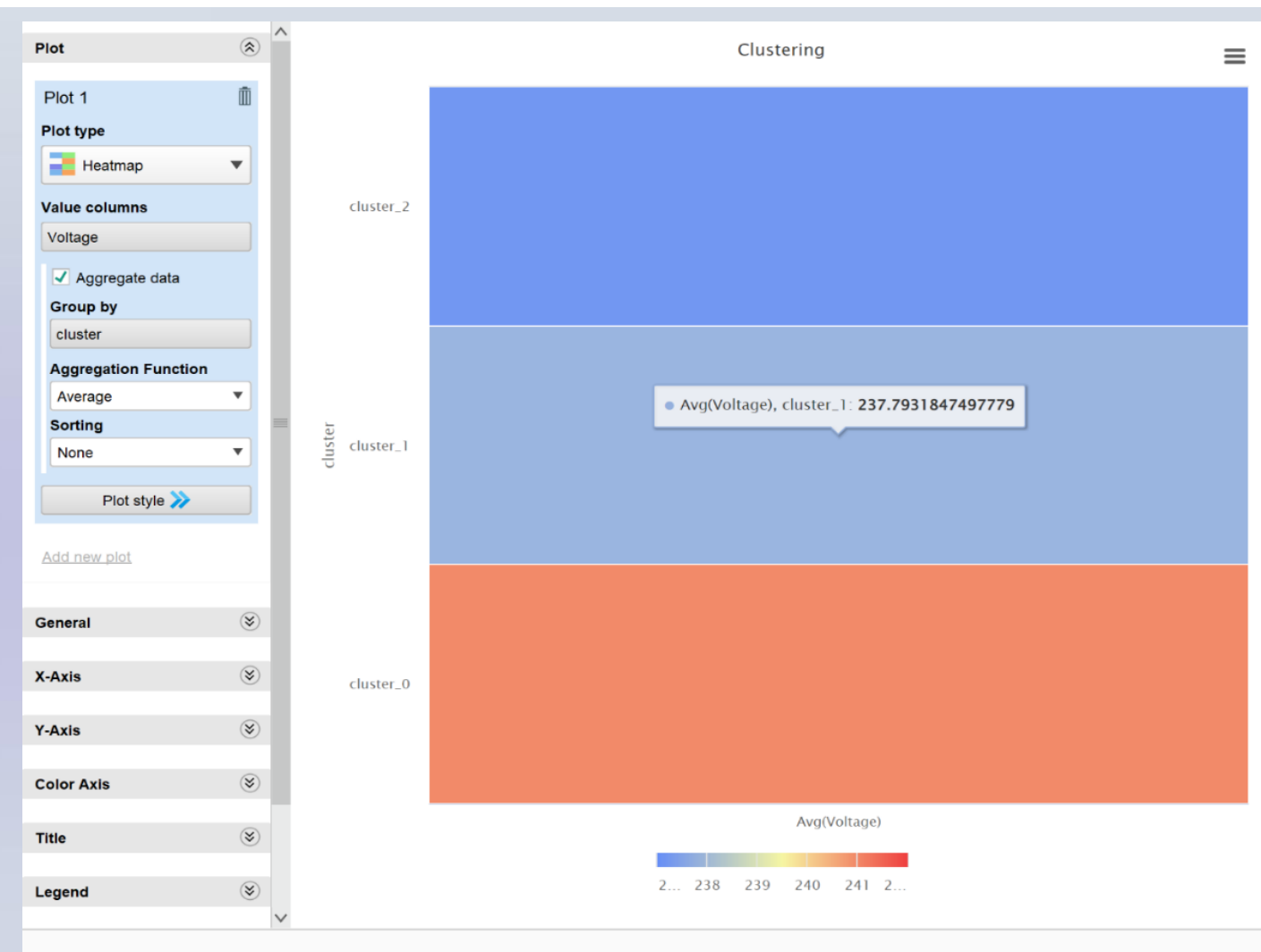
Чем насыщеннее цвет, тем сильнее связь между переменными, выявляя скрытые взаимосвязи.



Тепловые карты в рассматриваемом наборе данных



Тепловая карта распределения записей по кластерам



Тепловая карта среднего значения напряжения

Корреляционный анализ данных

Корреляционная матрица позволяет количественно оценить связи между признаками, выделить наиболее значимые переменные для анализа и интерпретировать полученные кластеры.

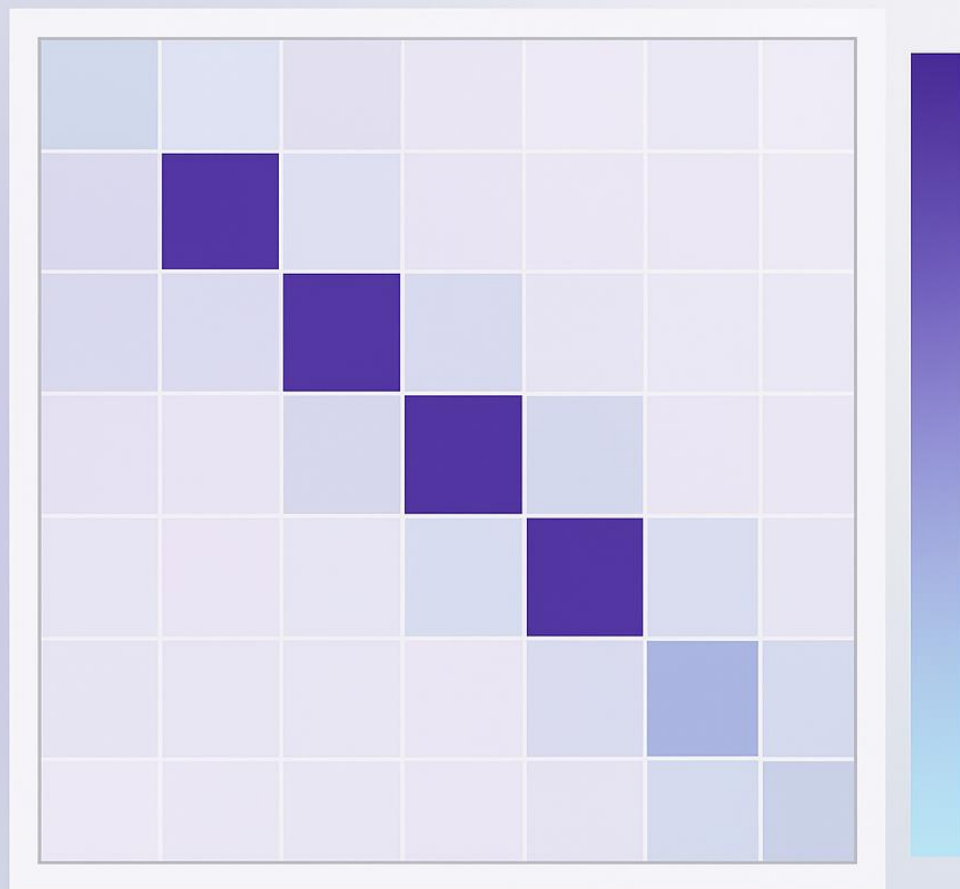
Attribu...	Global...	Global...	Voltage	Global...	Sub_m...	Sub_m...	Sub_m...
Global_...	1	0.247	-0.400	0.999	0.484	0.435	0.639
Global_r...	0.247	1	-0.112	0.266	0.123	0.139	0.090
Voltage	-0.400	-0.112	1	-0.411	-0.196	-0.167	-0.268
Global_i...	0.999	0.266	-0.411	1	0.489	0.440	0.627
Sub_me...	0.484	0.123	-0.196	0.489	1	0.055	0.103
Sub_me...	0.435	0.139	-0.167	0.440	0.055	1	0.081
Sub_me...	0.639	0.090	-0.268	0.627	0.103	0.081	1





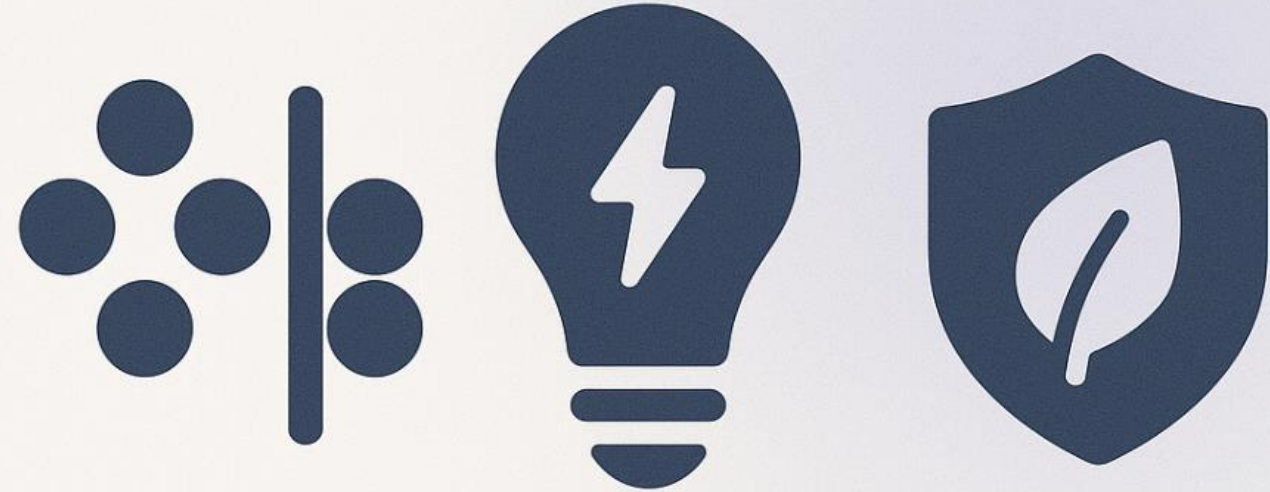
Особенности корреляций энергопотребления

Например, активная мощность и сила тока имеют практически идеальную корреляцию (0.999), что важно для выявления базовых закономерностей в энергетических данных.



Практическое значение кластерного анализа

Кластерный анализ данных энергопотребления позволяет оптимизировать расход электроэнергии, выявить неэффективные потребительские паттерны и разработать рекомендации по экономии ресурсов.

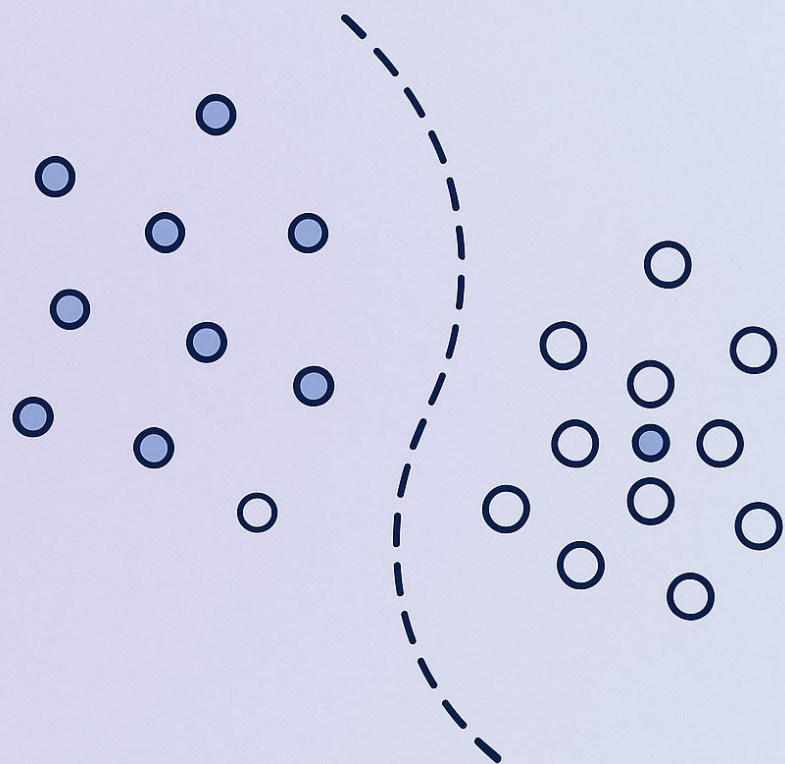




Проблемы и ограничения метода K-Means

Алгоритм чувствителен к выбросам и начальному выбору центров кластеров.

Для повышения точности анализа часто требуется многократный запуск с разными начальными условиями.



Заключение

В результате изучения кластерного анализа освоены методы выявления структур данных, проведён анализ энергопотребления, определены типичные и аномальные режимы работы электросетей.

