Правительство Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
(НИУ ВШЭ)

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

ПРАКТИЧЕСКАЯ РАБОТА № 2
по дисциплине «Информатика»

ТЕМА РАБОТЫ

«Анализ данных с использованием кластерного анализа в RapidMiner»

Москва, 2024

Оглавление

1. Введение

Целью данной практической работы является освоение методов анализа и визуализации данных на основе алгоритма кластеризации K-Means, который позволяет разделить объекты на группы по их сходству. В рамках работы студенты будут выполнять предобработку данных, применять алгоритм кластеризации, анализировать полученные результаты и визуализировать их в удобной форме.

Особое внимание уделено этапу предобработки данных, поскольку качество исходного набора данных напрямую влияет на результаты кластеризации. Студенты также изучат подходы к анализу полученных кластеров, такие как построение гистограмм, боксплотов и тепловых карт, а также выполнению корреляционного анализа для выявления взаимосвязей между признаками.

Данная работа не только поможет закрепить навыки работы с RapidMiner, но и углубить понимание методов обработки данных, что крайне важно для решения задач в реальных проектах.

2. Содержание практической работы

Описание работы:

В основе работы лежит набор данных о потреблении электроэнергии, который включает временные ряды с характеристиками энергопотребления в домохозяйствах.

Этапы выполнения работы:

1. Разделить записи в данных на три группы (кластера) на основе их сходства.
2. Провести анализ полученных кластеров для выявления закономерностей.
3. Построить визуализации, отражающие распределение данных и их ключевые характеристики.
4. Исследовать взаимосвязи между признаками с помощью корреляционного анализа.

О наборе данных:

Анализ проводится на наборе данных, содержащем следующие характеристики:

- **Global_active_power** – активная мощность (кВт).
- **Global_reactive_power** – реактивная мощность (кВт).
- **Voltage** – напряжение сети (В).
- **Global_intensity** – сила тока (А).
- **Sub_metering_1, Sub_metering_2, Sub_metering_3** – потребление энергии на отдельных группах устройств (Вт).

Ключевые особенности данных:

- Количество записей: более 2 миллионов.
- Формат данных – это временные ряды с числовыми характеристиками.
- Потенциальные проблемы - наличие пропущенных значений, высокая плотность данных.

3. Ход работы

Загрузка набора данных

1. Откройте RapidMiner Studio.
2. В главном меню выберите **"Create New Process"**.
3. На панели инструментов слева найдите оператор **"Read CSV"** и перетащите его на рабочее пространство.
4. Загрузите набор данных о потреблении электроэнергии, выбрав файл **"household_power_consumption"** в формате CSV.
5. Подключите оператор **"Read CSV"** к оператору **"Result"**, чтобы данные могли быть отображены.
6. Нажмите кнопку **"Run"** для выполнения процесса.
7. В результате вы увидите таблицу с данными, содержащими атрибуты (**Date; Time; Global_active_power; Global_reactive_power; Voltage; Global_intensity; Sub_metering_1; Sub_metering_2; Sub_metering_3**)

Данные успешно загружены, их структура показана на рисунке 3.2. При загрузке данных из "household_power_consumption" важно учитывать, что возможно наличие пропущенных значений, а поскольку нам предстоит кластеризация методом k-means, необходимо заполнить пропуски. Заполнение осуществляется автоматически на этапе создания таблицы. Для этого на этапе **"Format your columns"** необходимо поставить галочку в графе **"Replace errors with missing values"** рисунок 3.1.

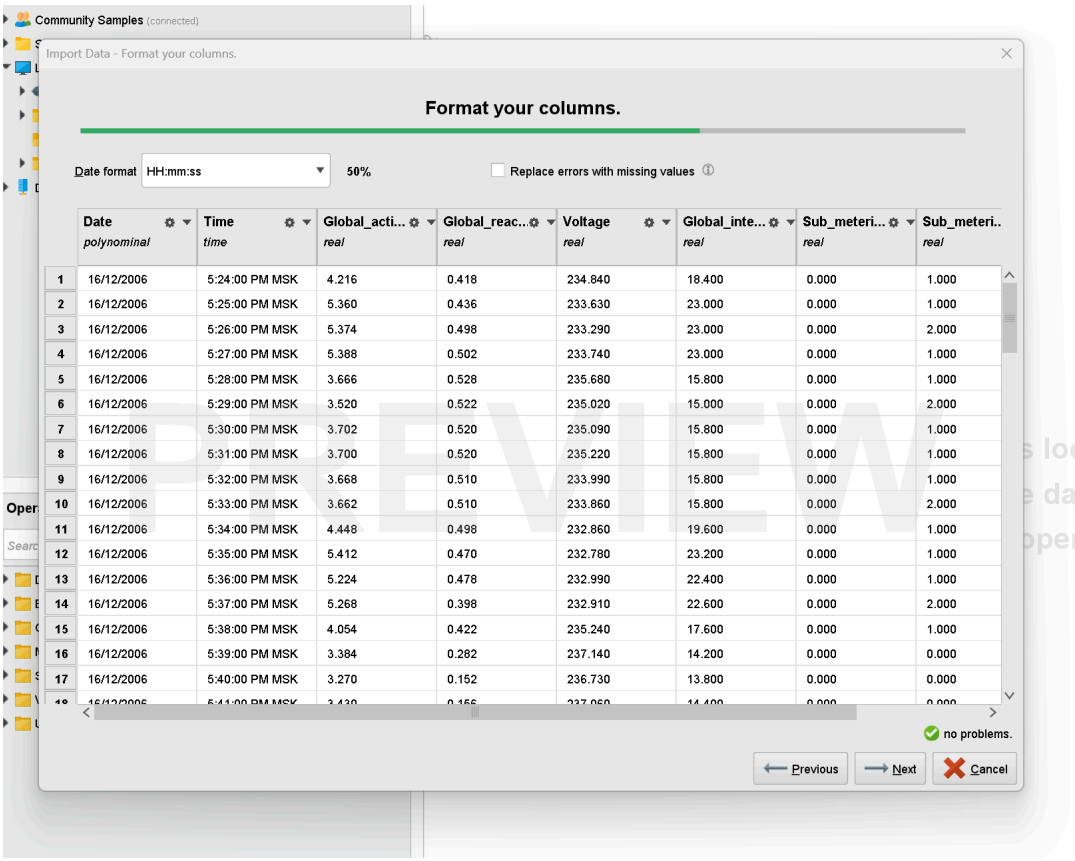


Рисунок 3.1 – подготовка данных к выгрузке

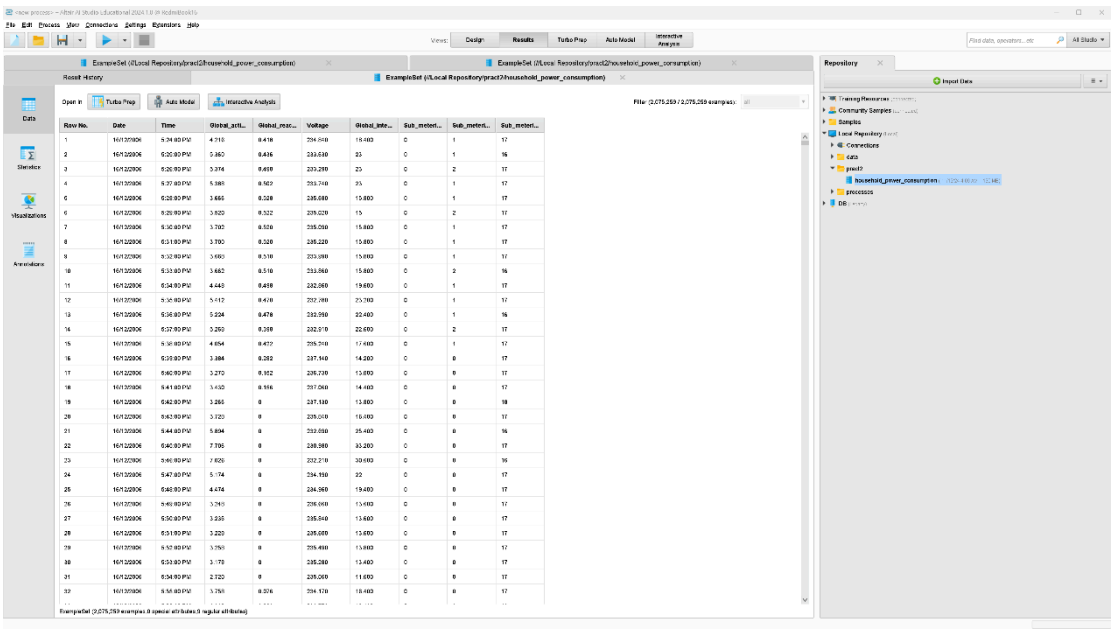


Рисунок 3.2 – выгруженные данные

Подготовка к кластеризации:

Для того чтобы применить алгоритм k-means необходимо убрать из нашей выборки данные, которые невозможно обработать – в их число входят дата и время. Для вернёмся к визуальному представлению на вкладку **"Design".** И добавим ряд новых процессов:

Clustering – необходимый для самого процесса кластеризации. Ключевой параметр k устанавливаем равным 3. Именно на такое количество кластеров разделятся наши данные. Настройка представлена на рисунке 3.3.

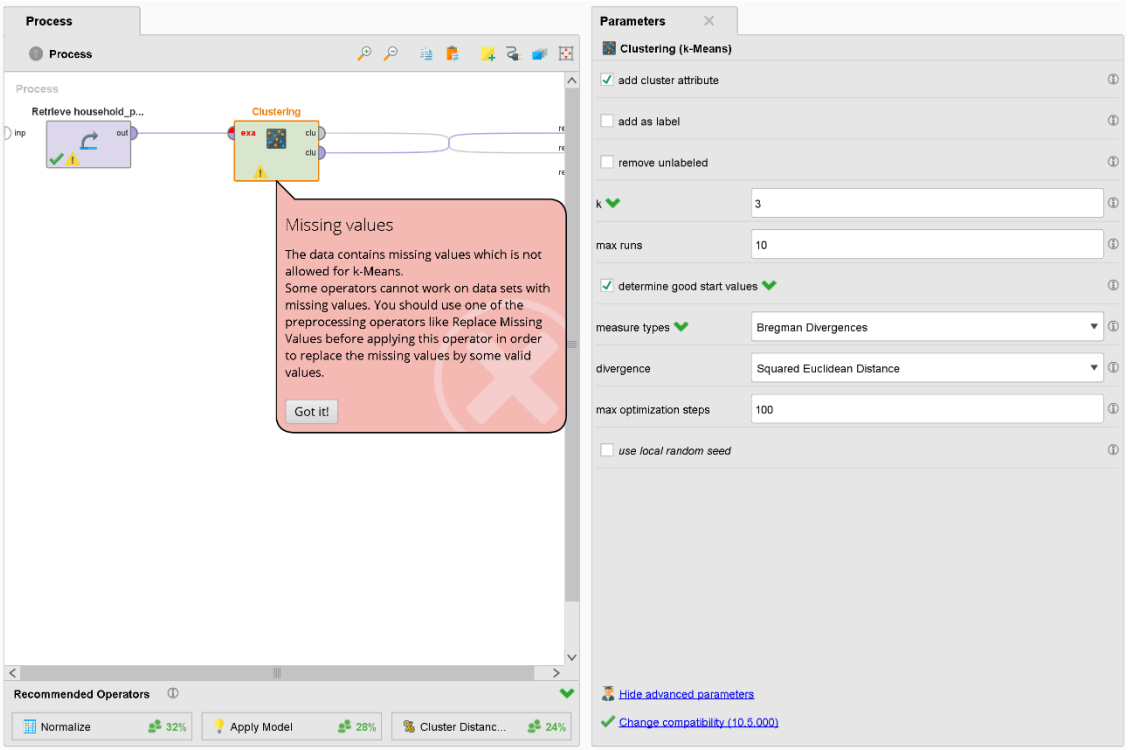


Рисунок 3.3 – настройки для Clustering

Замена пропущенных значений:

Replace Missing Values – необходим для заполнения пропущенных значений.

1. На панели инструментов найдите оператор **"Replace Missing Values"** и перетащите его на рабочее пространство.
2. Подключите выход оператора **"retrieve household power consumption"** ко входу оператора **"Replace Missing Values"**.
3. Выберите оператор **"Replace Missing Values"** и в параметрах укажите метод замены: **"average"** (среднее значение).

4. Подключите выход оператора **"Replace Missing Values"** ко входу оператора **"Result"**, чтобы отобразить обработанные данные.

Настройка представлена на рисунке 3.4.

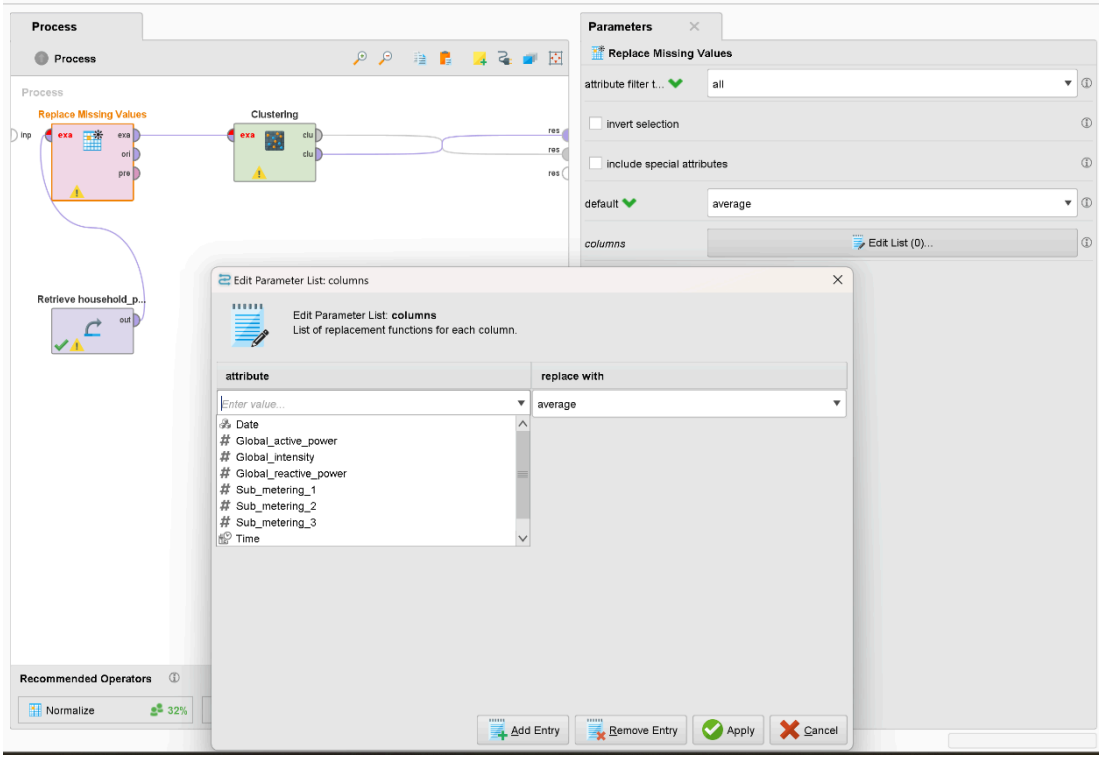


Рисунок 3.4 – настройки для Replace Missing Values

Выбор значимых признаков:

Select Attributes – необходим для выбора типа переменных подвергающихся процессу кластеризации.

1. Добавьте оператор **"Select Attributes"** из панели инструментов.
2. Подключите выход оператора **"Replace Missing Values"** ко входу оператора **"Select Attributes"**.
3. В настройках оператора **"Select Attributes"** выберите способ фильтрации: **"type(s) of values"**.
4. Укажите, что для анализа будут использоваться только числовые признаки (real и integer).
5. Подключите выход оператора **"Select Attributes"** ко входу оператора **"Result"**, чтобы отобразить отфильтрованные данные.
6. Нажмите кнопку **"Run"**, чтобы выполнить процесс.

Настройка представлена на рисунке 3.5.

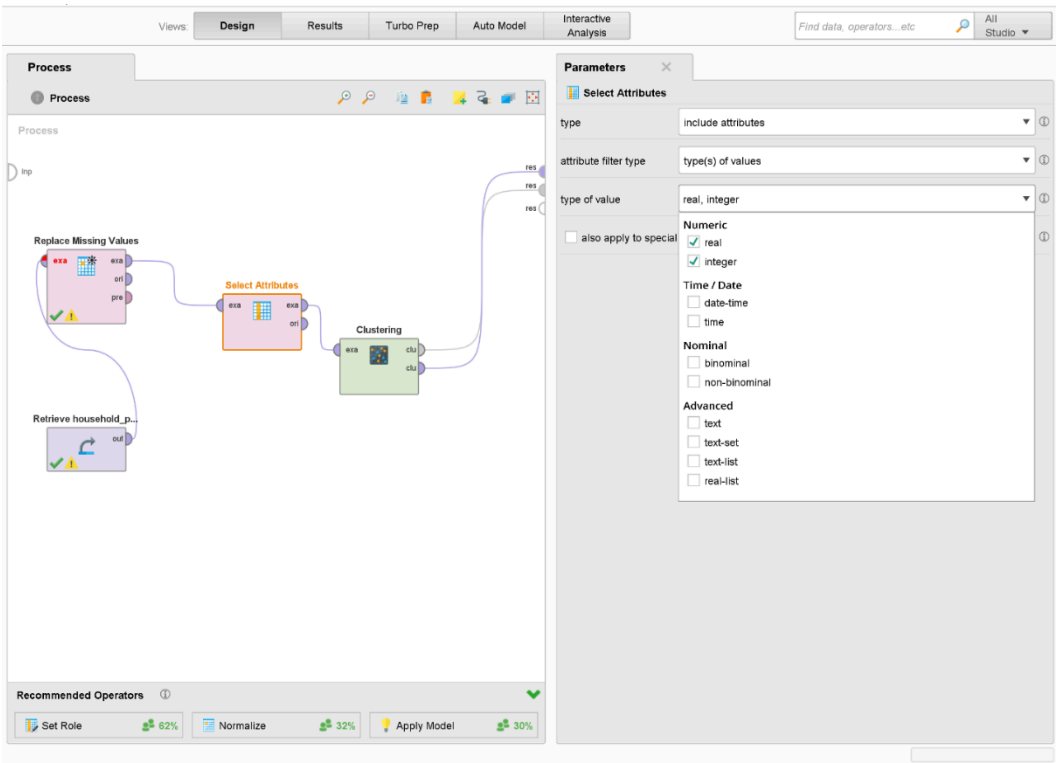


Рисунок 3.5 – настройки для Select Attributes

Результаты применения алгоритма:

На данном этапе работы были получены результаты кластеризации данных с использованием алгоритма **k-means**. Основные итоги применения алгоритма:

Распределение элементов по кластерам:

- Кластер 0: 1,969,715 элементов.
- Кластер 1: 48,996 элементов.
- Кластер 2: 56,548 элементов.
- Общее количество записей в наборе данных: 2,075,259.

Особенности кластеров:

Большинство данных (примерно 95%) находится в кластере 0, что может указывать на схожесть значений в основном массиве данных. Кластеры 1 и 2 содержат существенно меньше данных, что может сигнализировать о существовании подгрупп с уникальными характеристиками. Результаты представлены на рисунке 3.6

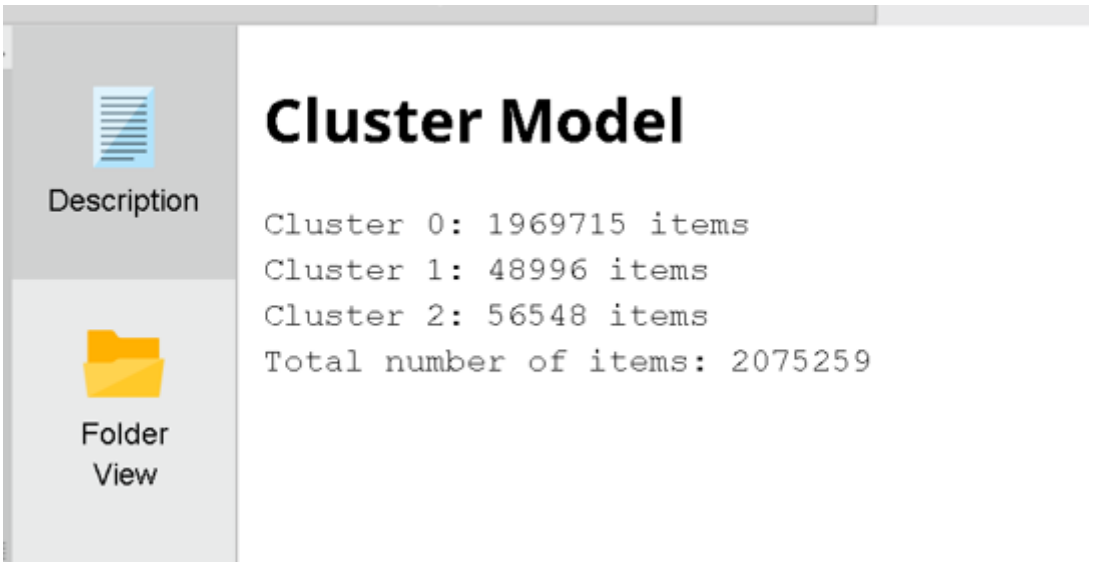


Рисунок 3.6 – Полученные кластеры

Детальный анализ кластеров:

На рисунке 3.7 представлен раздел **Centroid table** – это таблица, представляющая усреднённые значения характеристик для каждого кластера, определённого алгоритмом.

Рассмотрим представленные для каждого кластера результаты и сделаем выводы.

Кластер 0 можно считать фоновым или основным режимом энергопотребления, с минимальными затратами и стабильным уровнем напряжения.

Кластеры 1 и 2 выделяются повышенными значениями энергопотребления, что делает их интересными для дальнейшего анализа. Их различия в активной мощности могут быть связаны с определёнными устройствами или поведением пользователей.

Подобный анализ помогает выявить закономерности в данных и выделить подгруппы с особыми характеристиками. Это может использоваться для: оптимизации распределения электроэнергии, разработки рекомендаций по снижению потребления энергии, выявления возможных аномалий или интересных шаблонов в энергопотреблении.

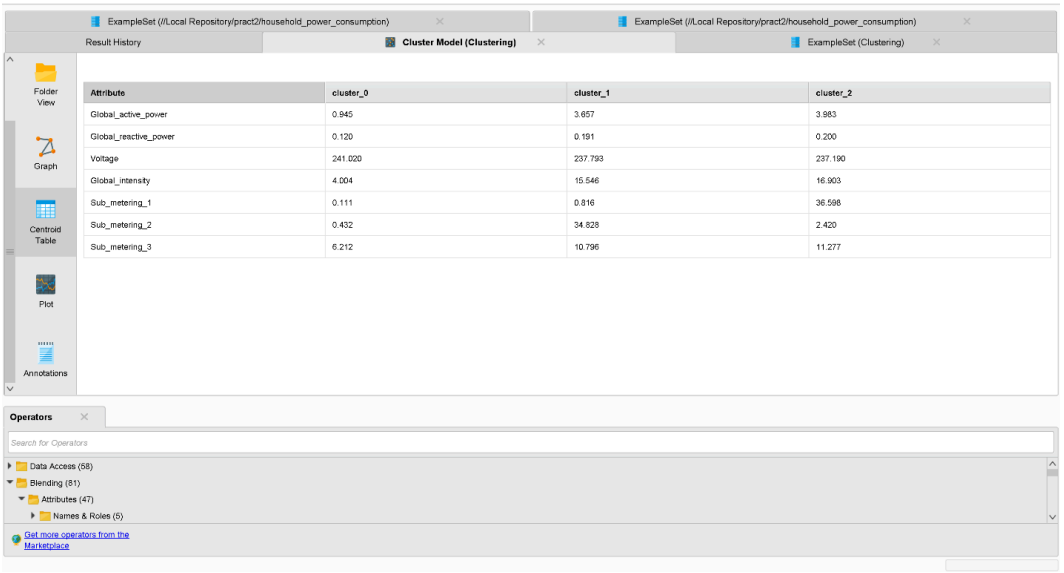


Рисунок 3.7 – Centroid table

Визуализация полученных результатов:

На рисунке 3.8 представлена гистограмма, отражающая распределение количества записей по каждому кластеру. График демонстрирует, как данные сегментировались в результате применения алгоритма K-means.

Анализ гистограммы показывает, что кластер 0 является доминирующим, так как в него попала большая часть записей. Это может свидетельствовать о том, что большинство наблюдений соответствует основным режимам энергопотребления, характеризующимся типичным или стандартным поведением.

Кластеры 1 и 2 содержат значительно меньше записей, что связано с их спецификой.

Кластер 1, включает редкие случаи более высокого энергопотребления, что может быть связано с использованием специфического оборудования или пиковыми нагрузками.

Кластер 2 отражает отдельные необычные события или низкую активность в домохозяйстве.

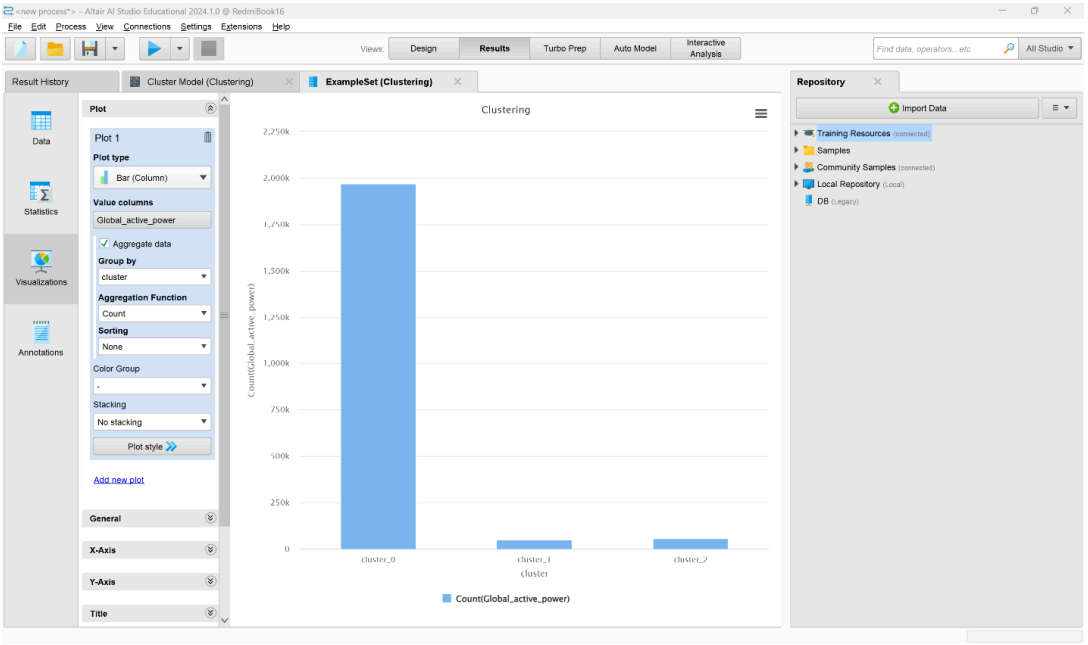


Рисунок 3.8 – распределение количества записей по кластерам

Построение гистограммы распределения мощности:

Продолжим строить графики для визуализации полученных результатов. Для построения гистограммы в разделе визуализации **"Plot type"** был выбран тип графика **"Histogram"**. В качестве переменной для анализа использовался атрибут **"Global_active_power"**, а в качестве группы — кластер. Данные автоматически разбились на интервалы, что позволило отобразить распределение активной мощности по каждому кластеру рисунок 3.9. Гистограмма распределения активной мощности показывает, что кластер 0 доминирует по численности и включает в себя значения с низким потреблением энергии, что соответствует фоновому режиму работы бытовых устройств. Кластеры 1 и 2 представлены значительно меньшим количеством примеров, но характеризуются более высокими значениями активной мощности, что может свидетельствовать о пиковых нагрузках или использовании энергоёмких устройств. Такое распределение подчёркивает различия в потребительских паттернах между кластерами

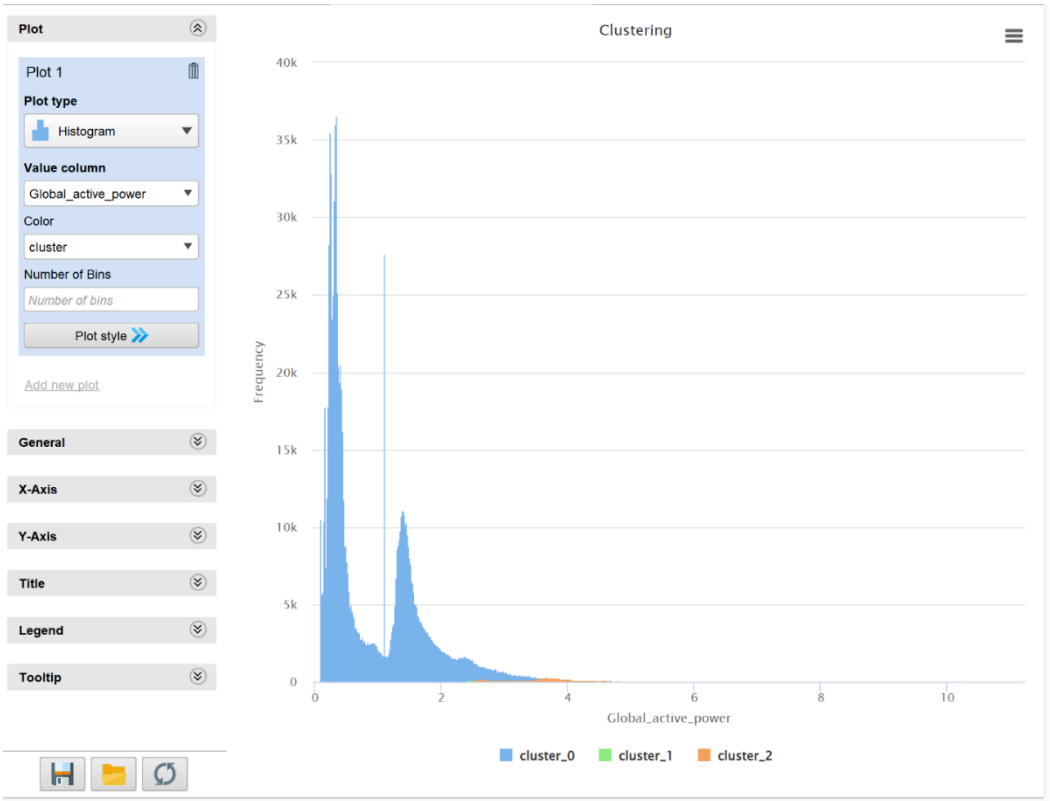


Рисунок 3.9 – гистограмма распределения мощности

Построение гистограммы распределения напряжения:

Для визуализации распределения напряжения по кластерам был построен график-гистограмма. В разделе визуализации **"Plot type"** был выбран тип графика **"Histogram"**. В качестве переменной для анализа использовался атрибут **"Voltage"**, а группировка осуществлялась по кластеру. Автоматическое разбиение данных на интервалы позволило выявить особенности распределения напряжения для каждого кластера, как показано на рисунке 3.10.

Гистограмма показывает, что кластер 0 содержит основную массу примеров и характеризуется нормальным распределением напряжения с пиком около среднего значения 240 В, что соответствует стандартным показателям бытовой электрической сети. Кластеры 1 и 2 представлены значительно меньшим количеством примеров и могут отражать редкие случаи отклонений от стандартного напряжения. Этот график подчёркивает устойчивость напряжения в подавляющем большинстве случаев, а также

выделяет меньшие группы с аномалиями, которые требуют дополнительного анализа.

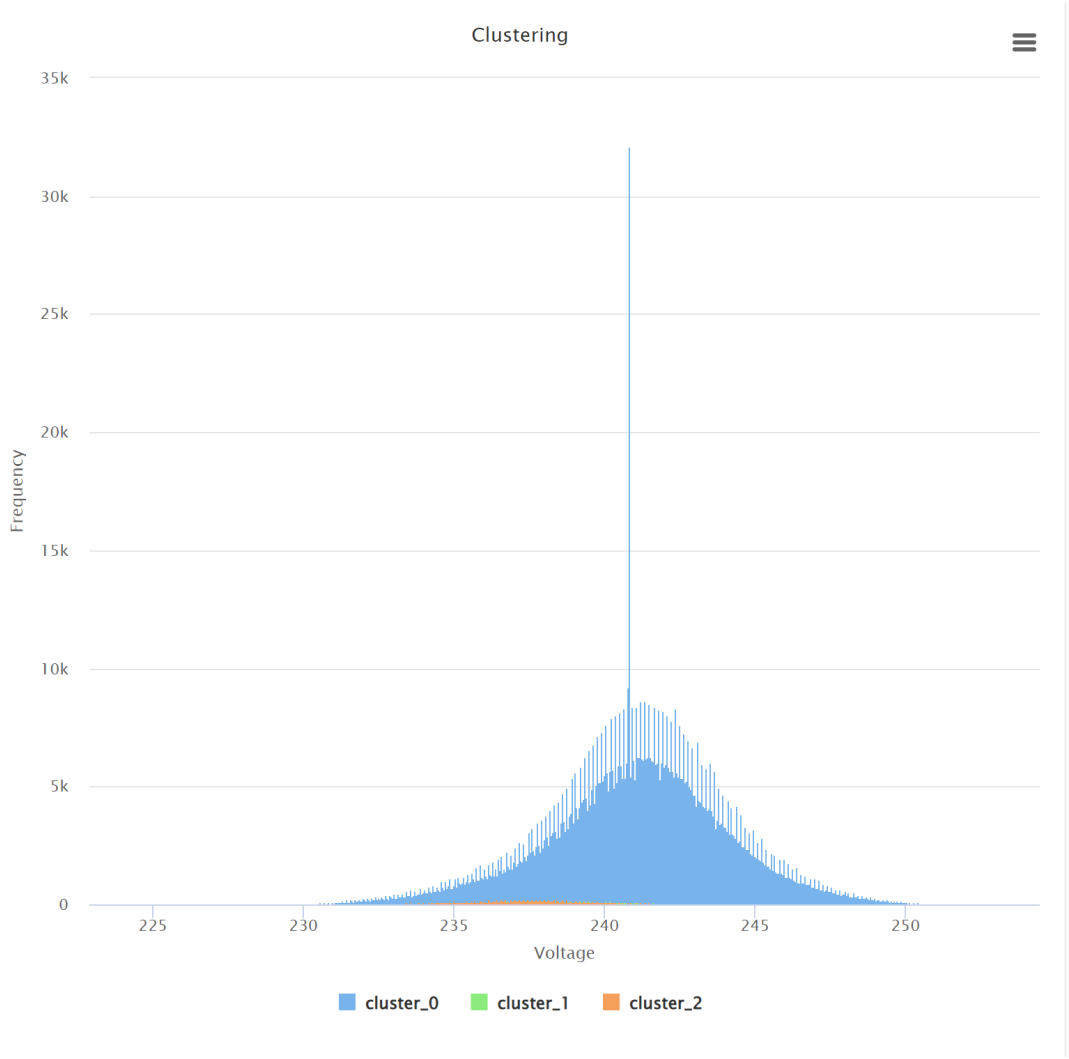


Рисунок 3.10 – гистограмма распределения напряжения

Построение диаграммы размаха для анализа напряжения:

Для визуализации статистических характеристик напряжения по каждому кластеру была построена диаграмма размаха **(boxplot)**. В разделе визуализации **"Plot type"** был выбран тип графика **"Boxplot"**, переменной для анализа стал атрибут **"Voltage"**, а группировка выполнялась по кластерам. Такой подход позволил отобразить минимальные, максимальные, медианные значения, а также границы квартилей напряжения в каждом из кластеров (рисунок 3.11).

Диаграмма показывает, что для кластера 0 напряжение варьируется в стандартных пределах с медианой около 241 В, что соответствует

нормальному рабочему диапазону электросети. Диапазон значений в этом кластере относительно узок, что свидетельствует о стабильности напряжения. Для кластеров 1 и 2 диапазоны шире, а медианные значения слегка отклоняются. Это может указывать на присутствие аномалий или внешних факторов, влияющих на напряжение.

Использование диаграммы размаха предоставляет визуальный способ оценить распределение и стабильность напряжения, помогая выявить закономерности и потенциальные аномалии в работе электросети.

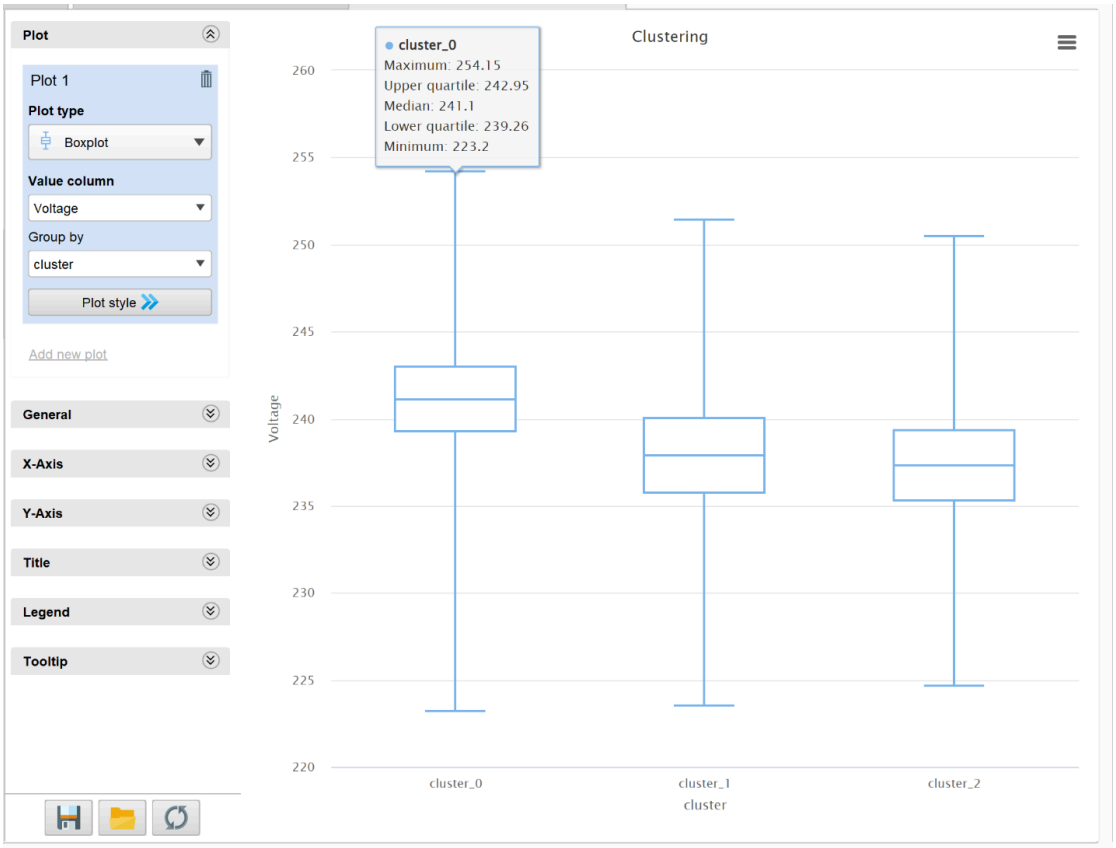


Рисунок 3.11 – диаграмма размаха для напряжения

Распределения активной мощности и интенсивности тока:

Построим аналогичные диаграммы размаха для анализа распределения значений активной мощности (**Global_active_power**) и интенсивности тока **(Global_intensity)** рисунки 3.12 и 3.13.

На графике активной мощности видно, что кластер 0 характеризуется низкими значениями мощности с медианой около 1, что подтверждает предположение о его фоновом режиме энергопотребления. Этот кластер

включает примеры с минимальными нагрузками, соответствующими стабильной работе бытовых приборов. В то же время, кластеры 1 и 2 демонстрируют более высокие значения мощности. Медиана в кластере 1 составляет около 3.5, с максимальными значениями, достигающими 11, что указывает на пиковые нагрузки или интенсивное использование энергоёмких устройств. Кластер 2 имеет медиану выше, чем кластер 1, и более широкий межквартильный размах, что говорит о повышенном и более разнообразном энергопотреблении.

Анализ распределения интенсивности тока также подтверждает эту закономерность. Кластер 0 имеет низкие значения с медианой около 5, что соответствует низкому уровню энергопотребления. Напротив, кластеры 1 и 2 демонстрируют значительно более высокие показатели. Медианы составляют около 15 и 17 соответственно, при этом кластер 2 отличается более широким диапазоном значений, что свидетельствует о большей вариативности потребления энергии.

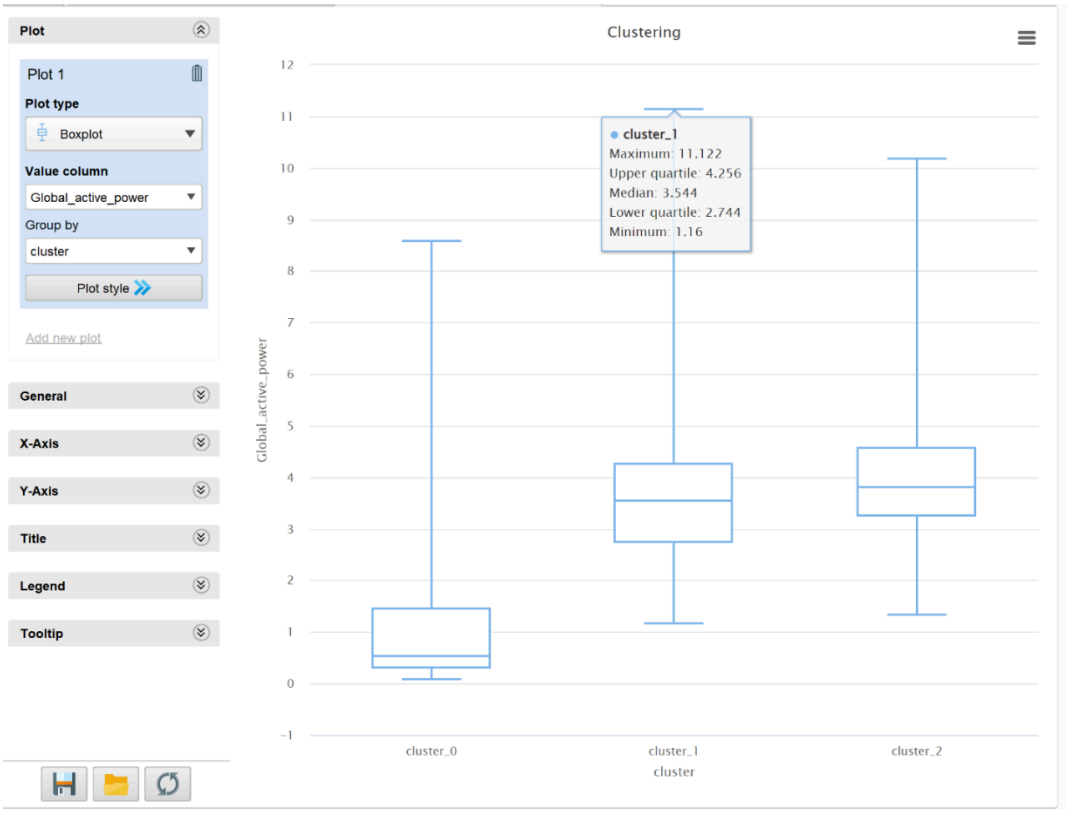


Рисунок 3.12 – распределения значений активной мощности

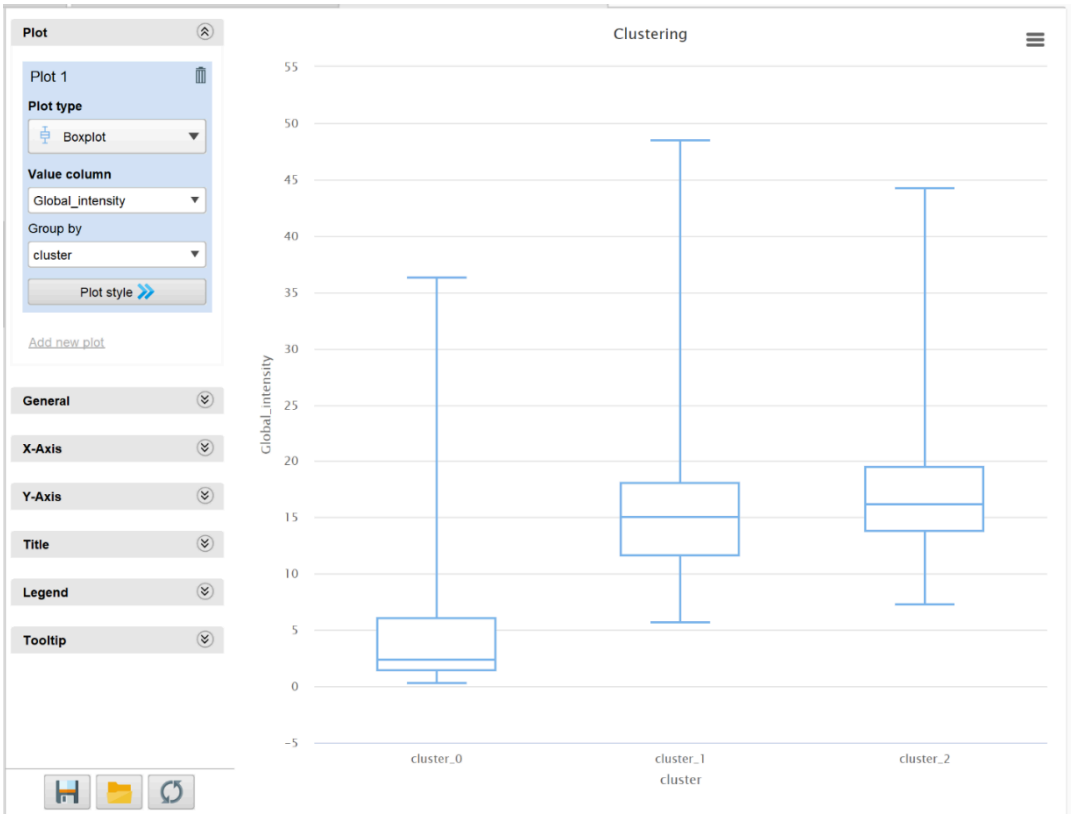


Рисунок 3.13 – распределения интенсивности тока

Построение тепловых карт:

Для визуализации распределения количества записей в кластерах были построены тепловые карты. В разделе визуализации **"Plot type"** был выбран тип графика **"Heatmap"**. В качестве переменной для анализа использовался атрибут **"Voltage"**, а группировка данных осуществлялась по кластеру. Результаты представлены на рисунке 3.14. Для второй тепловой карты, которая визуализирует средние значения напряжения для каждого кластера в разделе агрегации была выбрана функция **"Average"**, В настройках агрегации была выбрана функция подсчёта **"Count"**, что позволило отобразить количество записей для каждого кластера. Результаты представлены на рисунке 3.15.

Первая карта, использующая подсчёт записей, демонстрирует доминирующее положение кластера 0, который содержит подавляющее большинство данных, отражая стандартный режим работы сети. Кластеры 1 и 2 имеют значительно меньшее количество записей, что подчёркивает их специфичность и редкость. Вторая тепловая карта, основанная на среднем значении напряжения, показывает, что все кластеры имеют схожие средние значения напряжения около 240 В, соответствующие норме бытовой сети. Эти результаты подтверждают стабильность напряжения в основной части данных (кластер 0), а также выделяют меньшие кластеры как потенциальные области для дополнительного анализа из-за их специфических характеристик или возможных отклонений.

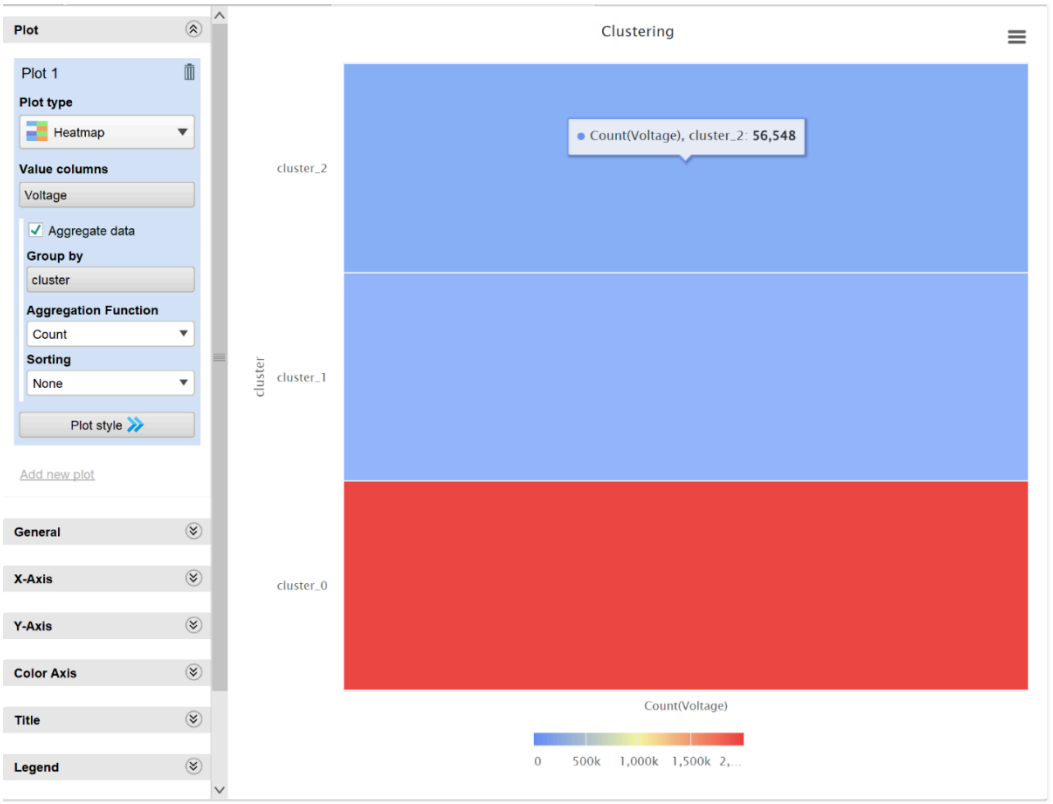


Рисунок 3.14 – тепловая карта распределения записей по кластерам

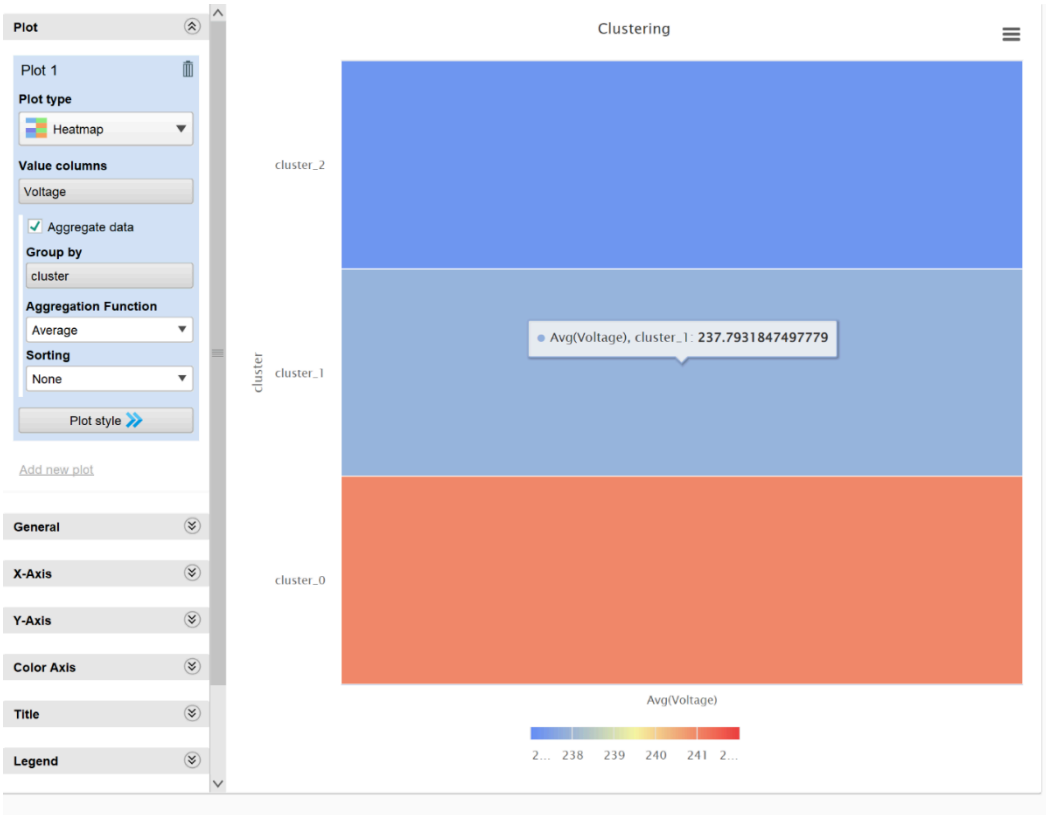


Рисунок 3.15 – тепловая карта среднего значения напряжения

Построение корреляционной матрицы:

Для анализа взаимосвязей между атрибутами данных был добавлен оператор построения корреляционной матрицы. В рабочем пространстве из библиотеки операторов выбран оператор **"Correlation Matrix"**. Этот оператор подключён к данным после этапа кластеризации, что позволяет изучить корреляции как в общем наборе данных, так и внутри выделенных кластеров.

В параметрах оператора были указаны следующие настройки:

- Тип атрибутов: включены все атрибуты, участвующие в анализе **"Include attributes"**.
- Нормализация весов: активирована опция **"Normalize weights"**, что обеспечивает равный вклад переменных в расчёты.
- Фильтрация: атрибуты, не влияющие на анализ (например, идентификаторы), исключены на этапе подготовки.

Работа с оператором представлена на рисунке 3.16.

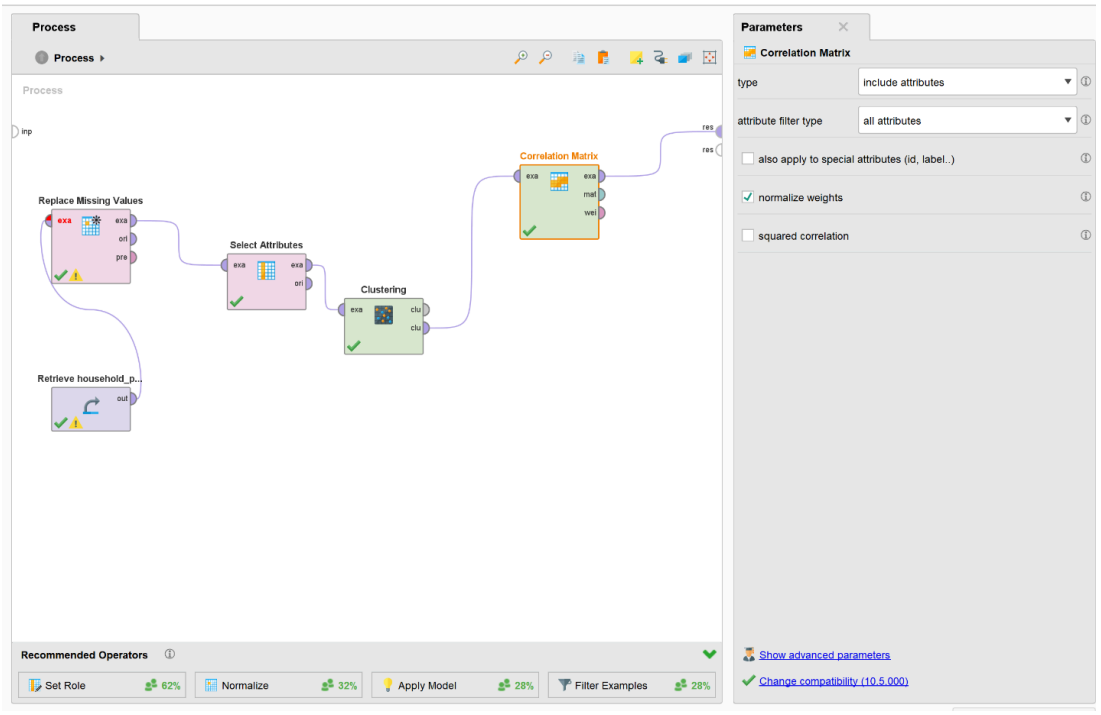


Рисунок 3.16 – настройка оператора для построения корреляционной матрицы

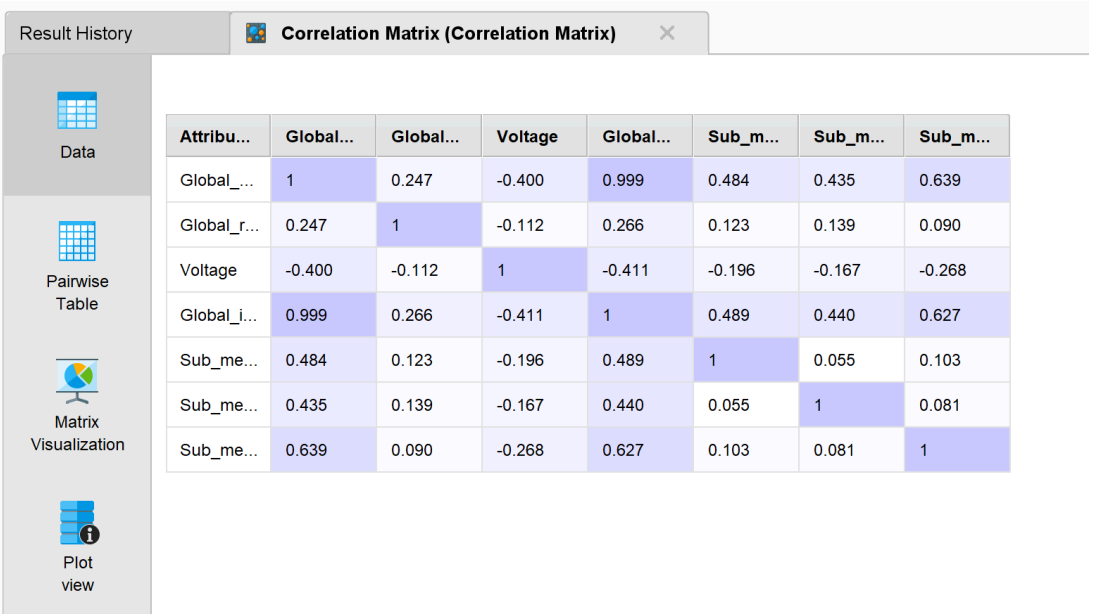
Анализ результата построения корреляционной матрицы:

После выполнения оператора Correlation Matrix был получен результат в виде таблицы рисунок 3.17, где указаны коэффициенты корреляции между каждой парой атрибутов. Эти значения помогают понять, насколько сильно связаны переменные между собой.

Основные наблюдения:

1. **Высокая корреляция между Global_active_power и Global_intensity**: Коэффициент 0.999 практически равен 1, что свидетельствует о прямой зависимости между этими показателями. Это логично, поскольку активная мощность напрямую связана с интенсивностью тока.
2. **Средняя корреляция между Global_active_power и Sub_metering_1, Sub_metering_2, Sub_metering_3**: Значения 0.484, 0.435 и 0.639 показывают умеренную положительную связь. Это говорит о том, что активная мощность частично объясняется энергопотреблением, распределённым по разным субсистемам.
3. **Отрицательная корреляция между Voltage и Global_active_power (-0.400)**: Указывает на слабую обратную связь. То есть при увеличении напряжения активная мощность, как правило, уменьшается, но эта связь не является строгой.
4. **Незначительная корреляция между Voltage и субметриками**: Значения около -0.2 и ниже говорят о слабой взаимосвязи, что можно интерпретировать как независимость этих параметров.
5. **Низкие значения корреляции между Sub_metering атрибутами**: Субметрики, такие как Sub_metering_1 и Sub_metering_2, имеют слабую положительную связь (0.055–0.103), что говорит о том, что они не сильно зависят друг от друга.

Также для визуализации взаимосвязей между переменными была построена тепловая карта, которая наглядно отображает значения коэффициентов корреляции между атрибутами. Построение осуществлено на основе данных корреляционной матрицы, где каждому сочетанию атрибутов соответствует цвет, отражающий степень их связи рисунок 3.18.



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

Рисунок 3.17 – корреляционная матрица

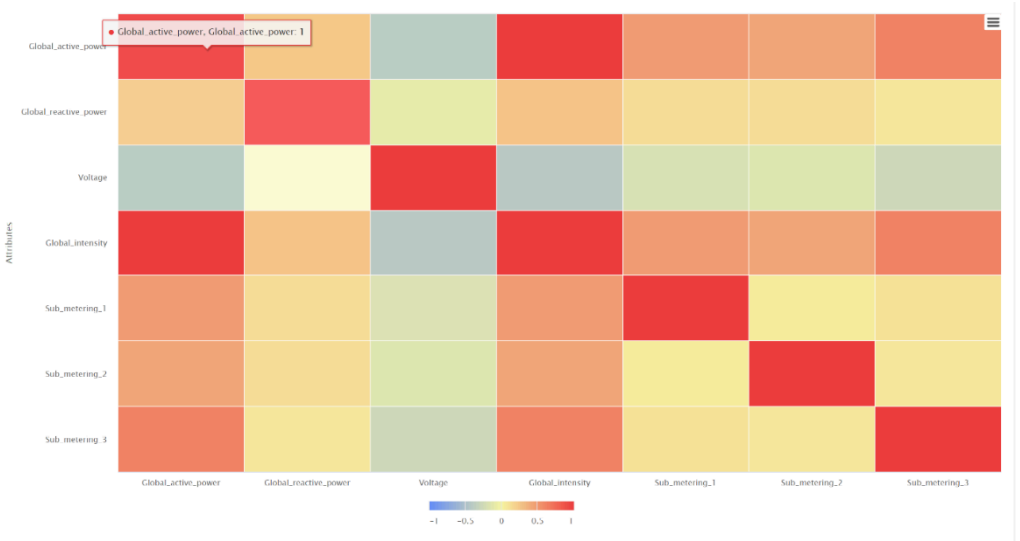


Рисунок 3.18 – тепловая матрица

4. Приобретаемые навыки

1. Работа с интерфейсом RapidMiner Studio для анализа данных и построения процессов.

2. Умение загружать, очищать и подготавливать данные для кластерного анализа.
3. Применение алгоритма кластеризации K-means и интерпретация его результатов.
4. Создание визуализаций для анализа распределений и взаимосвязей в данных.
5. Построение и интерпретация тепловых карт, гистограмм и box plot.
6. Анализ и использование корреляционной матрицы для изучения взаимосвязей между признаками.
7. Выявление ключевых характеристик кластеров и формулирование выводов по результатам анализа.
8. Освоение методов визуального представления и интерпретации данных для дальнейшего принятия решений.

5. Обобщенная задача для выполнения индивидуального варианта

Цель работы – провести кластерный анализ на предложенном в индивидуальном варианте табличном датасете с числовыми признаками. Постройте процесс в RapidMiner (Altair AI Studio), включающий следующие этапы:

1) Загрузка и предобработка данных

- Импортируйте CSV-файл с помощью оператора Read CSV.
- Обработайте пропуски (Replace Missing Values), приведите все числовые признаки к типу real.
- (При необходимости) выполните нормализацию или стандартизацию признаков оператором Normalize.

2) Выбор признаков для кластеризации

- С помощью Select Attributes отфильтруйте все нечисловые столбцы и исключите служебные поля (идентификаторы, временные метки и т.п.).
- Убедитесь, что в данных остались только те столбцы, по которым имеет смысл оценивать сходство объектов.

3) Проведение кластеризации

- Добавьте оператор K-Means и задайте число кластеров k.
- (Опционально) протестируйте несколько значений k, оценивая внутрикластерную инерцию или силу силуета.
- (Дополнительно) попробуйте иерархическую кластеризацию или алгоритм DBSCAN для сравнения.

4) Анализ и интерпретация кластеров

- Изучите распределение объектов по кластерам: выведите Centroid Table и подсчитайте доли объектов в каждом кластере.
- Проанализируйте средние и медианные значения признаков в кластерах, выявите характерные особенности.

5) Построение визуализаций:

- Гистограмма распределения ключевых признаков по кластерам.
- Box-plot для оценки разброса и медианы признаков в каждом кластере.
- Тепловая карта корреляций внутри кластеров или по всему набору данных.

6) Выводы

- Опишите, какие кластеры можно интерпретировать как «фоновое» состояние и какие — как аномальные или специализированные группы.

- Сформулируйте практические рекомендации по дальнейшему использованию результатов (целевые маркетинговые кампании, мониторинг аномалий, балансировка нагрузки и т.п.).

6. Распределение вариантов