# Тема 9. КЛАСТЕРИЗАЦИЯ

## 9.1. Введение в кластеризацию

Задача кластеризации применяется для автоматического разбиения элементов некоторого множества на группы в зависимости от схожести их свойств. Данная задача выполняет предварительную подготовку данных для дальнейшего анализа каждой выявленной группы в отдельности. Задача кластеризации применяется для анализа общей структуры множества данных, упрощения анализа за счет рассматривания каждой группы в отдельности, а также сокращения объема хранимых данных путем выбора наиболее индивидуальных представителей, выявления аномальных значений. Сформированные подгруппы в задаче кластеризации применяются дальше в задачах классификации и прогнозирования.

Для кластеризации применяют семейства алгоритмов *k-means* и *g-means*. Из них наиболее распространенных является алгоритм k-means.

В основе работы алгоритма k-means лежит принцип оптимального в определенном смысле разбиения множества данных на $k$ кластеров. Алгоритм пытается сгруппировать данные в кластеры таким образом, чтобы целевая функция алгоритма разбиения достигала экстремума. Выбор числа $k$ может базироваться на теоретических соображениях или интуиции. Центром кластера являются средние значения переменных объектов, входящих в кластер.

Алгоритм состоит из двух этапов:

- *первоначальное распределение объектов по кластерам*. Задается число $k$, и на первом шаге эти точки считаются *центрами* кластеров. Каждому кластеру соответствует один центр. Выбор начальных центров осуществляется случайным образом. В результате каждый объект назначен определенному кластеру;
- *итерационный процесс*. Вычисляются новые центры кластеров и объекты перераспределяются.

Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не стабилизируются центры кластеров, то есть все объекты будут принадлежать кластеру, которому они принадлежали до текущей итерации.

Если число кластеров назначить затруднительно, то можно использовать алгоритм g-means. Он определяет число кластеров в модели на основании последовательного выполнения статистического теста на то, что данные внутри каждого кластера подчиняются определенному гауссовскому закону распределения. Если тест дает отрицательный результат, кластер разбивается на два новых кластера (алгоритмом k-means) с центрами, расположенными на оси главных компонент.

Алгоритмы k-means и g-means ориентированы на гипотезу о компактности, которая предполагает, что данные выборки в виде многомерных векторов образуют в пространстве компактные сгустки сферической формы. В противном случае кластеры, найденные этими алгоритмами, будут малоинформативными. В этих случаях лучше использовать неметрические алгоритмы, например, на основе смеси распределений (*EM кластеризация*).

После того, как указаны входные поля, следует нормализация данных в обучающей выборке. Целью нормализации значений полей является преобразование данных к виду, наиболее подходящему для обработки алгоритмом. Для узлов, решающих задачи описательной или предсказательной аналитики, данные, поступающие на вход, должны иметь числовой тип, а их значения должны быть распределены в определенном диапазоне. Нормализатор может преобразовать дискретные данные к набору уникальных индексов или значения, лежащие в произвольном диапазоне, к диапазону [0...1].

Кроме того, существует настройка нормализации полей по умолчанию, то есть этап нормализации можно пропустить. В этом случае нормализация будет произведена автоматически в зависимости от вида данных полей:

- *дискретный* — нормализация битовой маской со способом кодирования — комбинация битов;
- *непрерывный* — линейная нормализация, для алгоритмов кластеризации без приведения в какой-либо диапазон.

В качестве функции расстояния k-means в Loginom используют:

- *евклидово расстояние* — для непрерывных числовых полей, а также упорядоченных категориальных признаков;
- *функцию отличия* — для неупорядоченных категориальных признаков.

## 9.2. Методические указания

В файле *Задача 9.1. Здравоохранение.xlsx* имеются данные о обеспеченности населения лечебными учреждениями регионов Российской Федерации в 2018 г. (рис. 9.1).

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Регион | Мощность амбулаторно-поликлинических учреждений на 10 тыс. населения, посещений в смену | Численность населения на одну больничную койку, чел. | Обеспеченность врачами на 10 тыс. человек населения, чел. | Обеспеченность средним медицинским персоналом на 10 тыс. человек населения, чел. |
| 2 | Белгородская область | 241,4 | 137,6 | 41,6 | 112,5 |
| 3 | Брянская область | 284,4 | 133,6 | 39,1 | 119,6 |
| 4 | Владимирская область | 345,9 | 119,1 | 33,9 | 98,9 |
| 87 | Еврейская автономная область | 264,2 | 77,4 | 36,8 | 119,8 |
| 88 | Чукотский автономный округ | 489,7 | 76,3 | 69,4 | 143,9 |

Рис. 9.1

Требуется провести кластерный анализ регионов, используя алгоритм k-means, и выяснить существуют ли заметные различия в их обеспеченности населения лечебными учреждениями.

### *Решение*

Создадим новый пакет *Кластеризация*. Выполним импорт исходных данных. Для этого создадим узел сценария, выполняющий действие импорта (рис. 9.2).

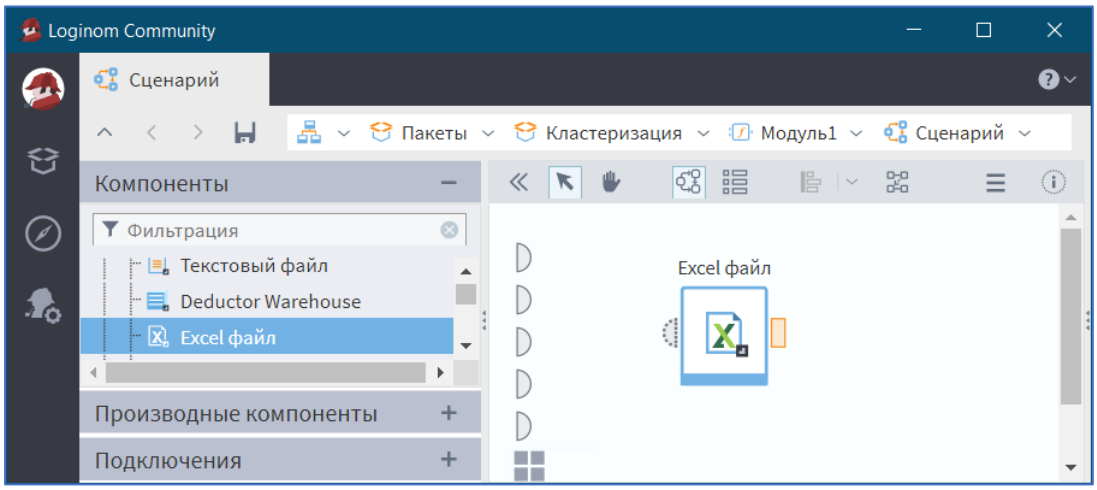


Рис. 9.2

Вызовем *Мастер настройки.* Пройдем шаги мастера, указав в описании узла метку *Задача 9.1. Здравоохранение.*

Добавим визуализаторы *Таблица* и *Статистика* к узлу сценария (рис. 9.3).

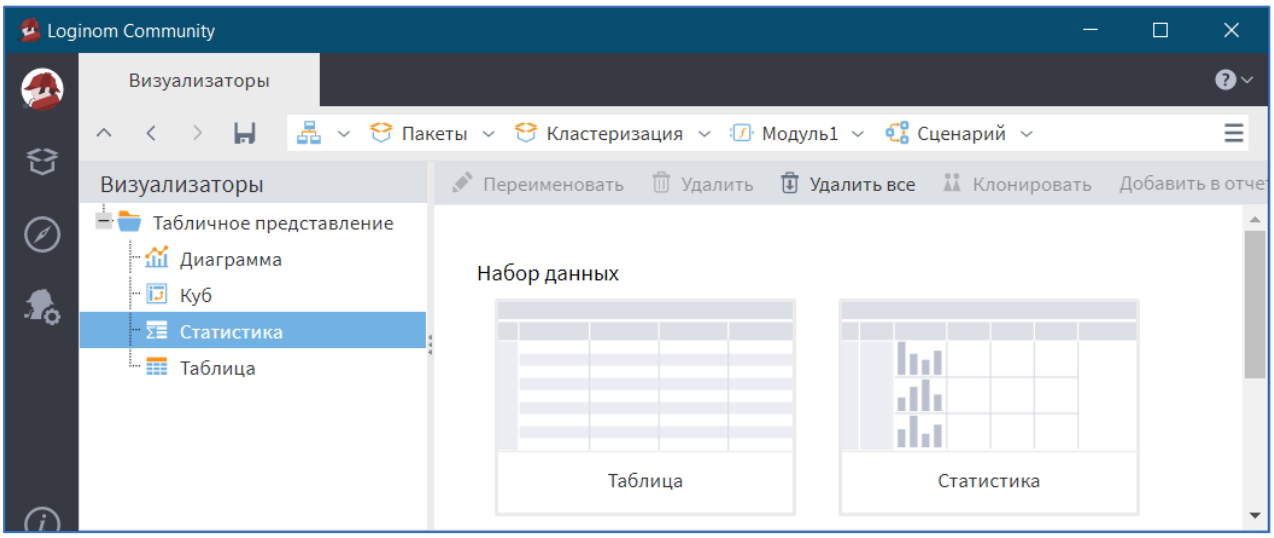


Рис. 9.3

Визуализаторы с исходными данными и статистическими показателями имеют вид (рис. 9.4–9.5).

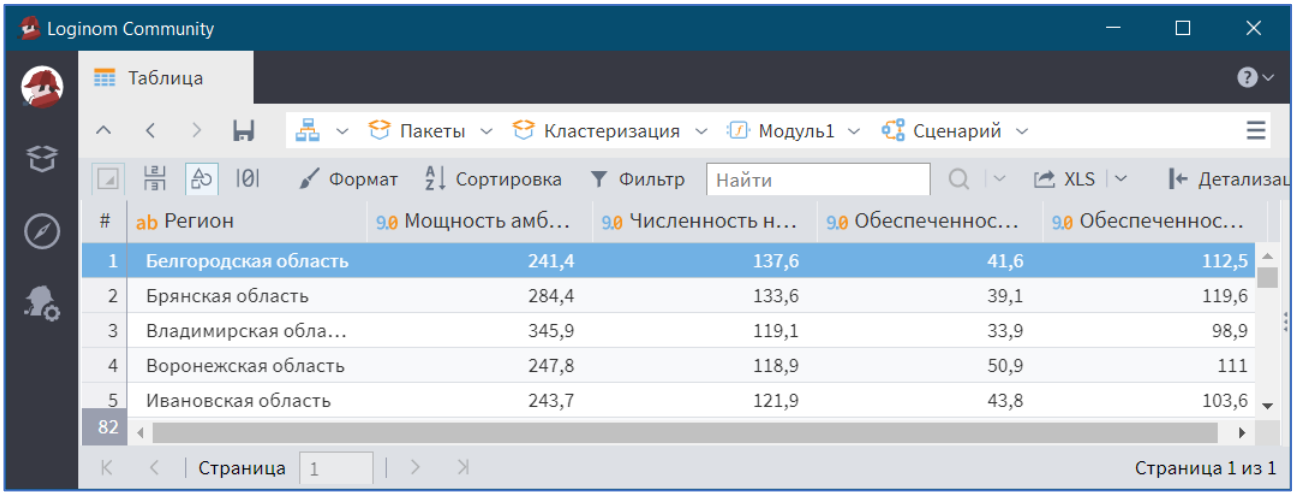


Рис. 9.4

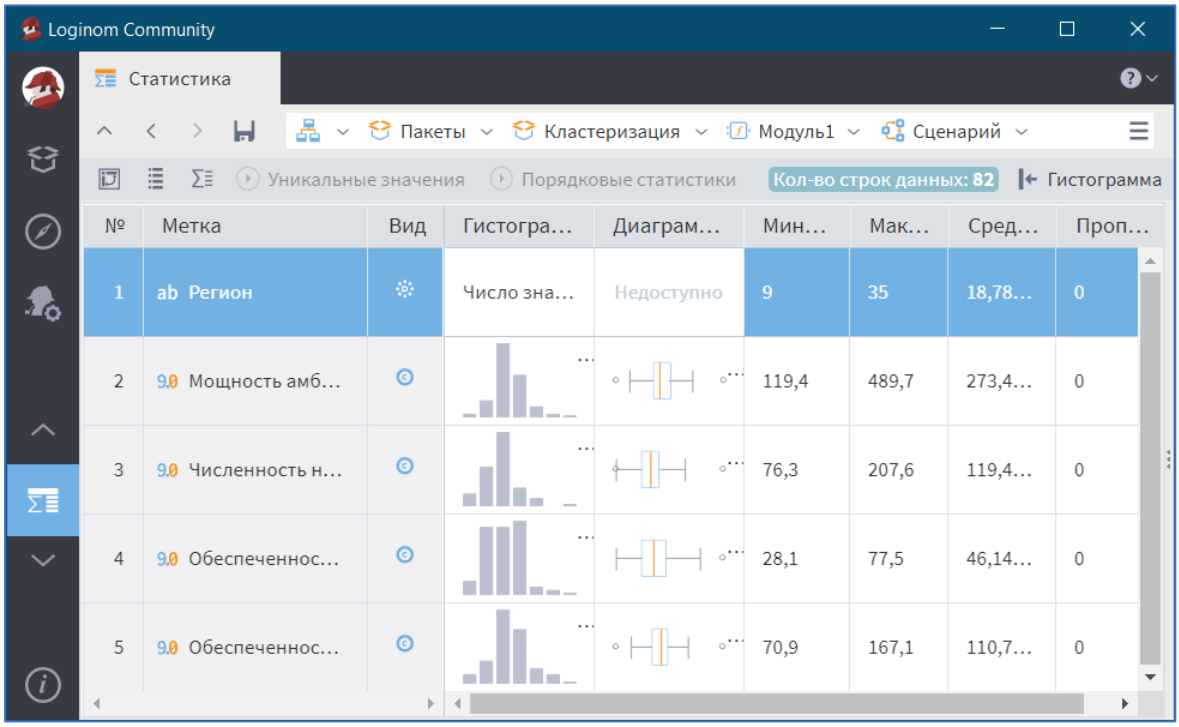


Рис. 9.5

Проведем кластеризацию регионов, используя алгоритм k-means. Для этого переместим компонент *Кластеризация* в рабочую область сценария. Последовательность обработки данных задается соединением выходного порта узла импорта с входным портом кластеризации (рис. 9.6).

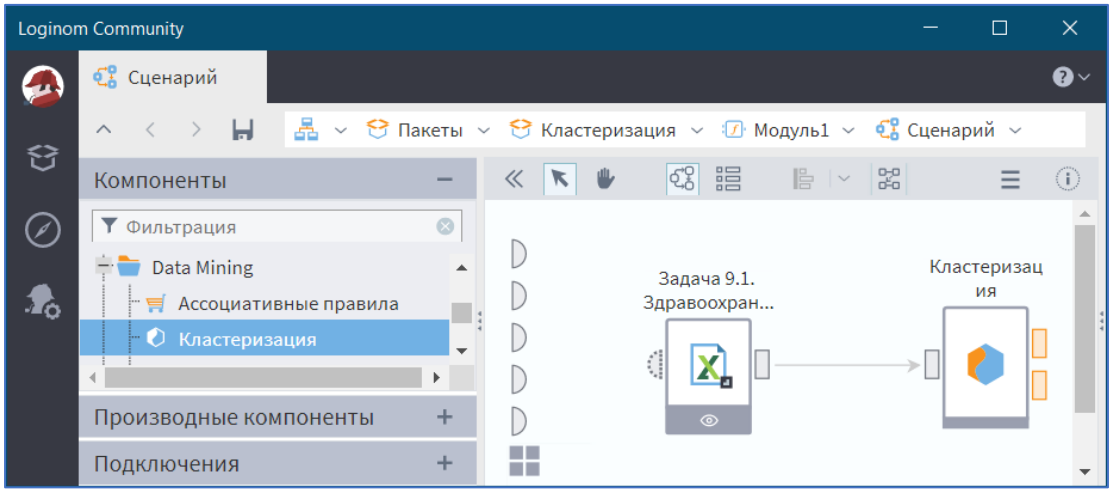


Рис. 9.6

Пройдем шаги *Мастера настройки*. На шаге *Настройка входных столбцов* настроим назначение исходных столбцов данных. Столбец *Регион* зададим как *Не задано*, остальные столбцы — как *Используемое* (рис. 9.7).

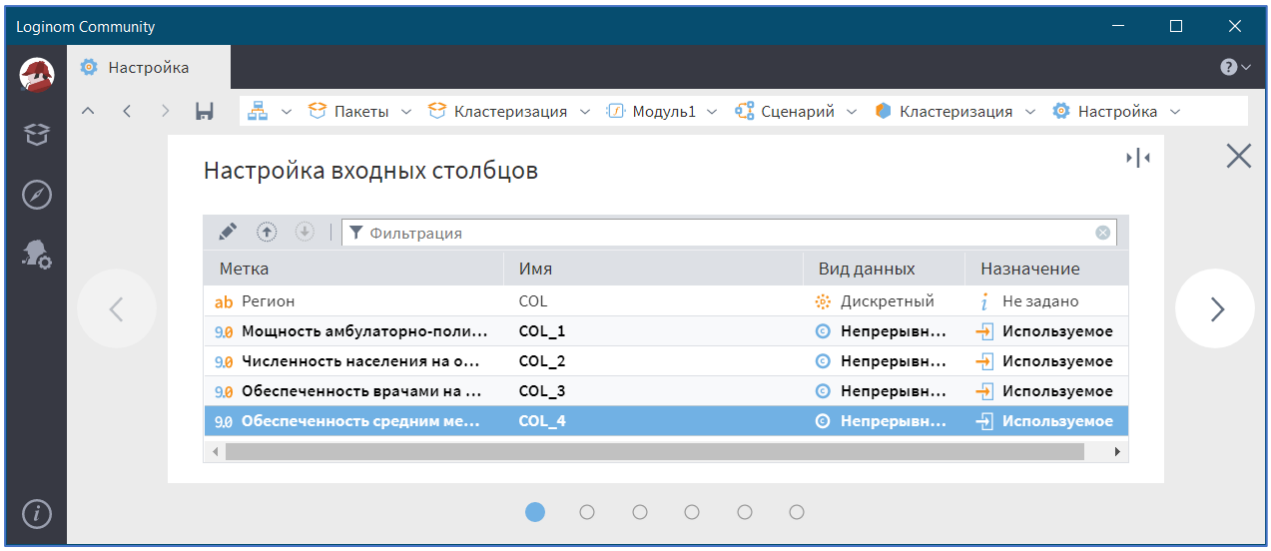


Рис. 9.7

На шаге *Настройка нормализации* оставим стандартные параметры по умолчанию (рис. 9.8).

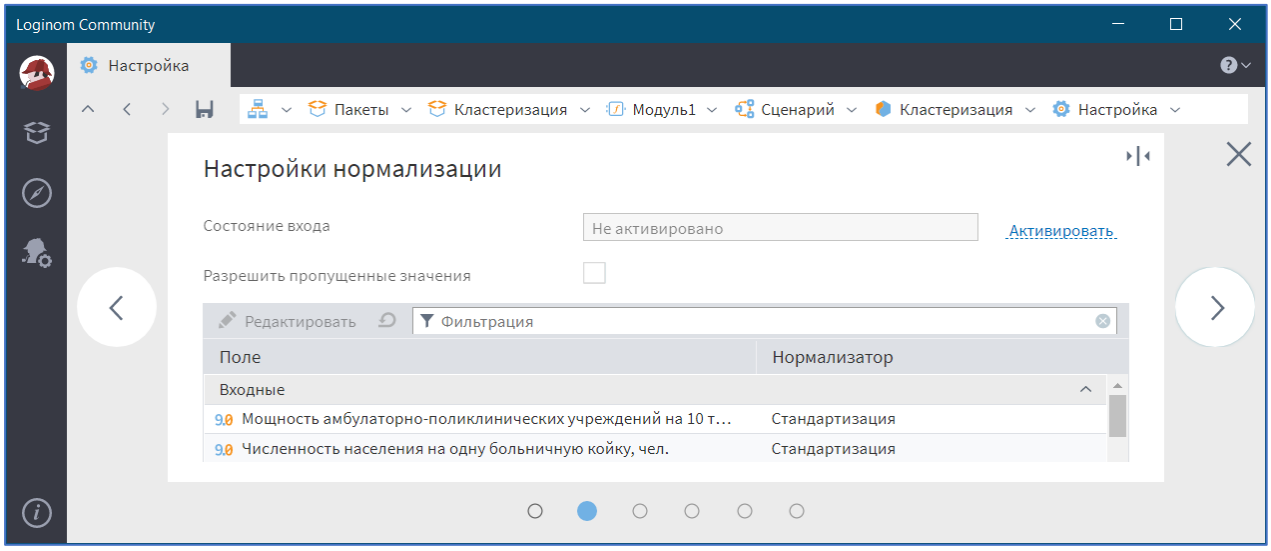


Рис. 9.8

На шаге *Кластеризация* снимем флажок с параметра *Автоматическая настройка* и зададим число кластеров равным 3, исходя из предположения, что регионы могут быть отнесены к группам с лучшей, средней и худшей обеспеченностью населения лечебными учреждениями (рис. 9.9).

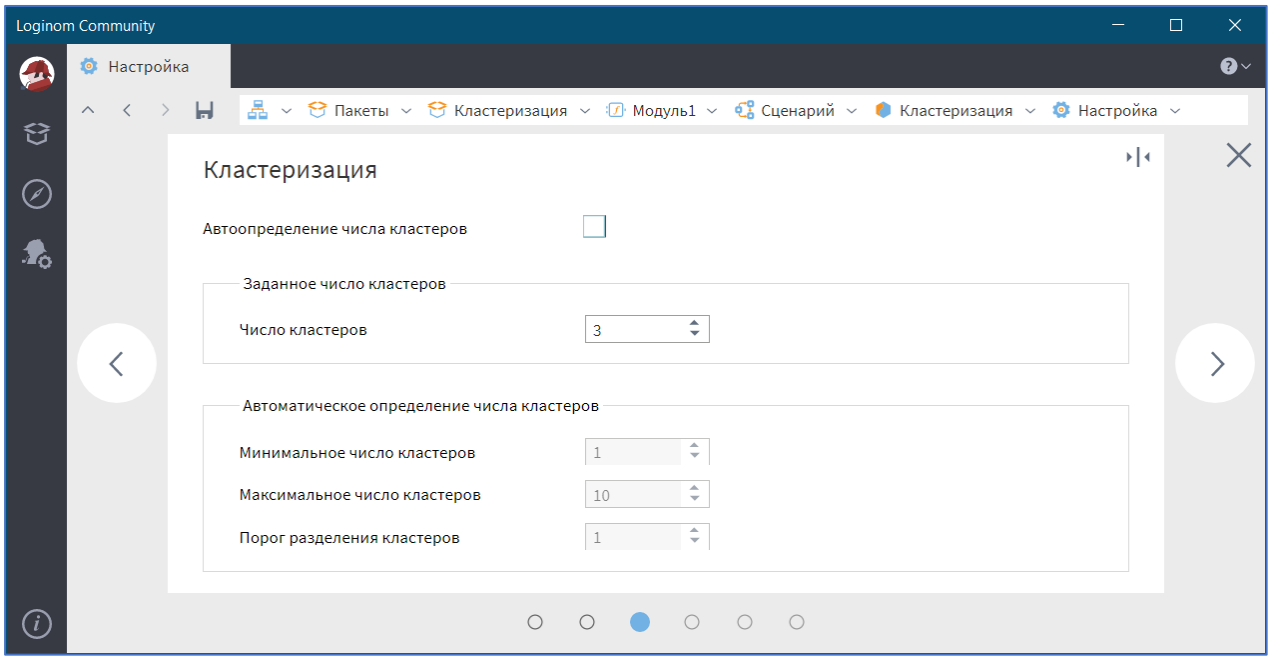


Рис. 9.9

Переобучим узел *Кластеризация* и перейдем к настройкам визуализаторов (рис. 9.10).

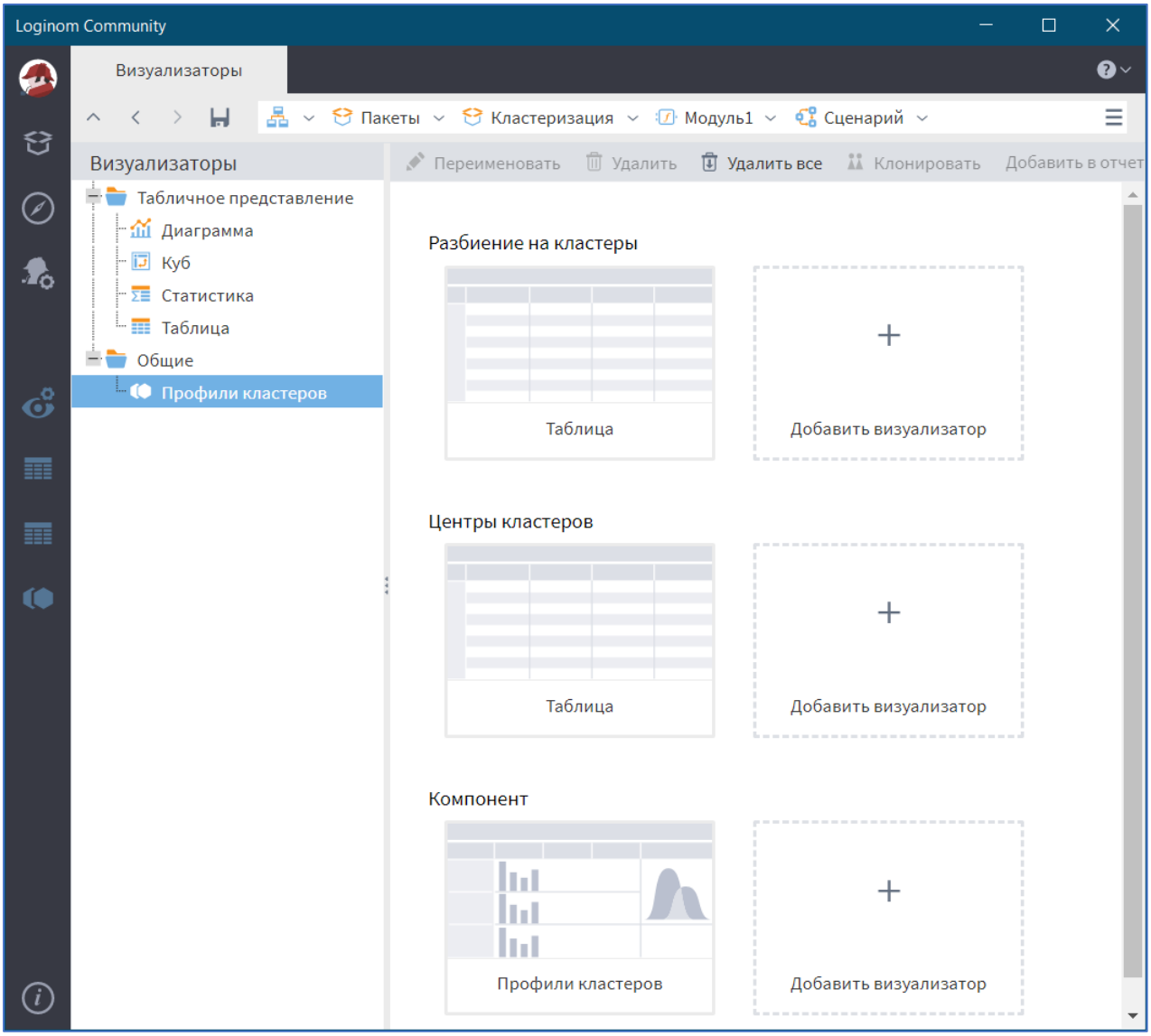


Рис. 9.10

Вначале откроем визуализатор *Центры кластеров* (рис. 9.11).

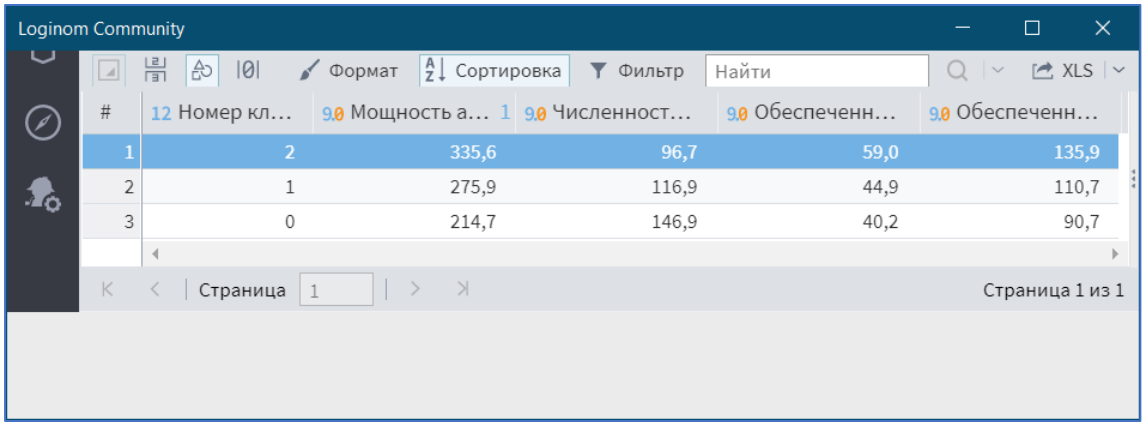


Рис. 9.11

Видно, что регионы, входящие в кластер 2, относятся к группе лучших по обеспеченности населения лечебными учреждениями, в кластер 1 — к группе

средних и в кластер 0 — к группе худших. Об этом можно судить по средним значениям мощности амбулаторно-поликлинических учреждений, численности населения на одну больничную койку, обеспеченность врачами и средним медицинским персоналом.

В визуализаторе *Профили кластеров* можно посмотреть общую структуру сформированных кластеров (рис. 9.12–9.14).

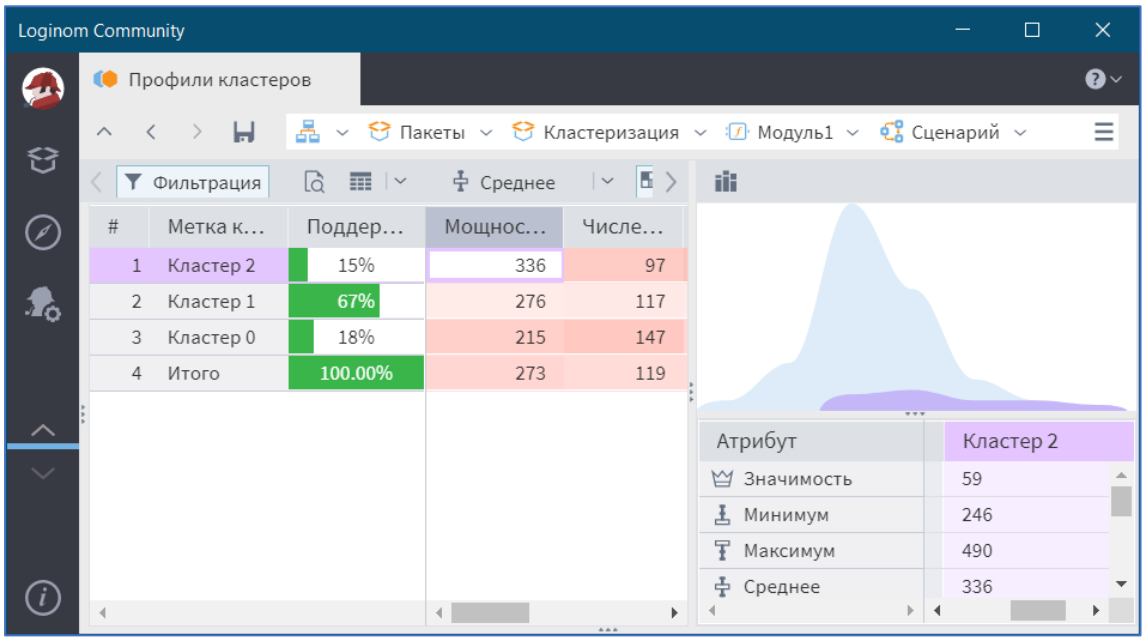


Рис. 9.12

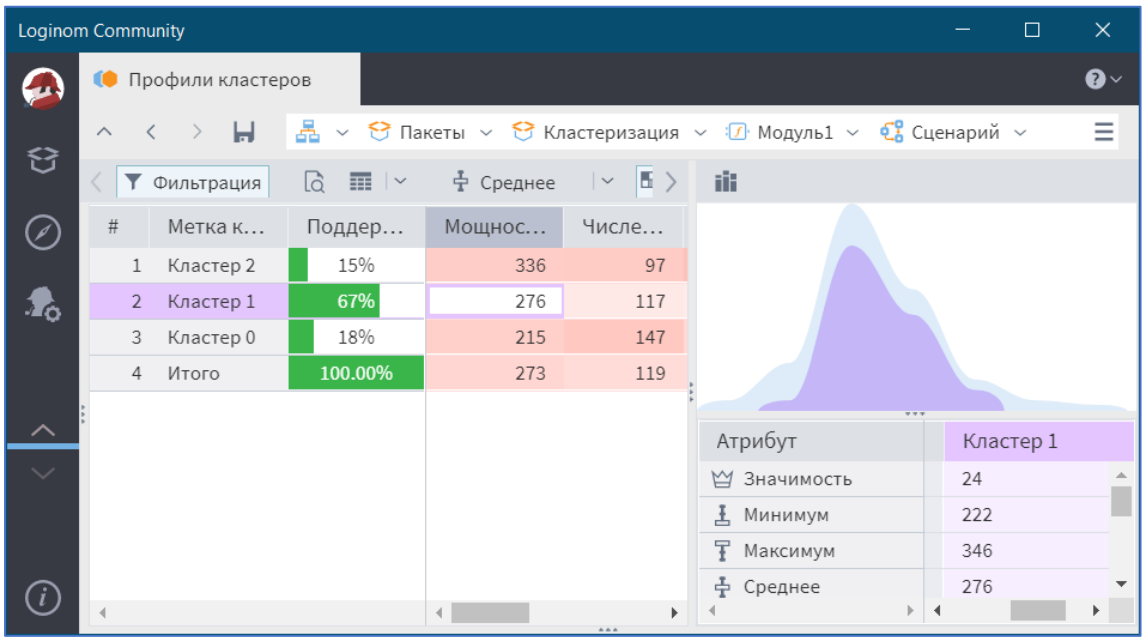


Рис. 9.13

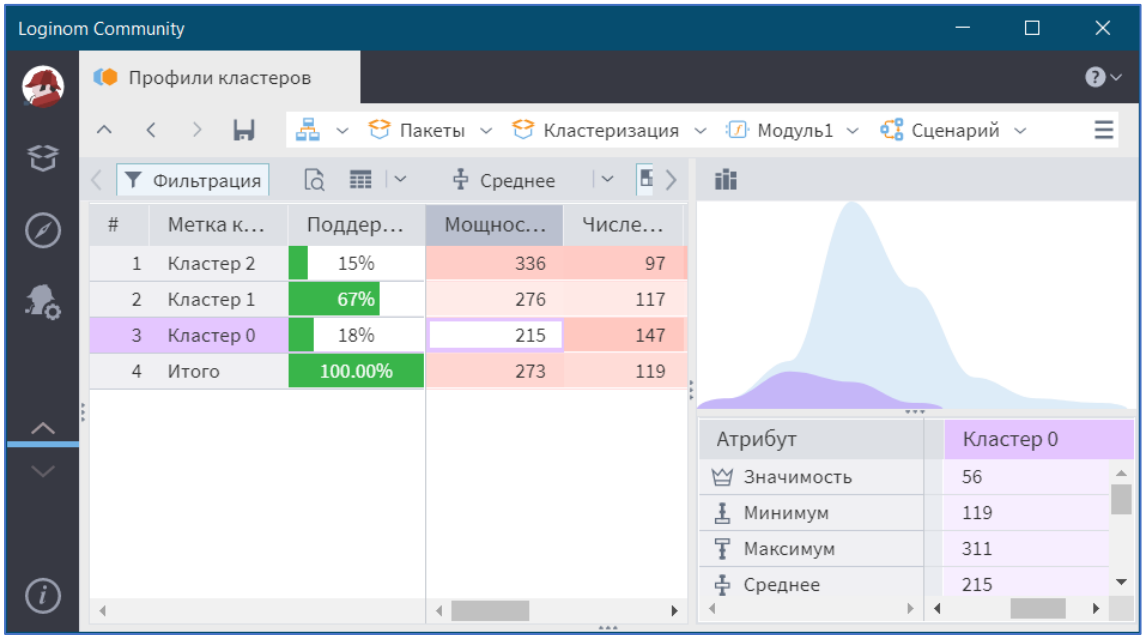


Рис. 9.14

В нем отображаются все рассматриваемые показатели вместе с характером влияния их на состав кластера. Основным определяющим состав кластера фактором является значимость свойств, выраженная в процентах. Общая значимость рассматриваемого поля определяется вариабельностью ее рассматриваемых параметров. Значимость для непрерывных и дискретных полей определяется поразному. Для непрерывных полей она устанавливается в зависимости от отклонения среднего значения рассматриваемой группы кластеров от общего среднего всей выборки. Чем больше выражено данное отклонение, тем больше его значимость. Значимость для дискретных полей определяется наличием индивидуальных различий, между рассматриваемыми группами, чем больше выражены различия, тем больше значимость. Для каждого рассматриваемого свойства в кластере вычисляется: среднее, стандартное отклонение, стандартная ошибка и др. Приводится накопительная диаграмма,

Визуализатор *Разбиение на кластеры* позволяет определить принадлежность регионов к соответствующим кластерам (рис. 9.15).

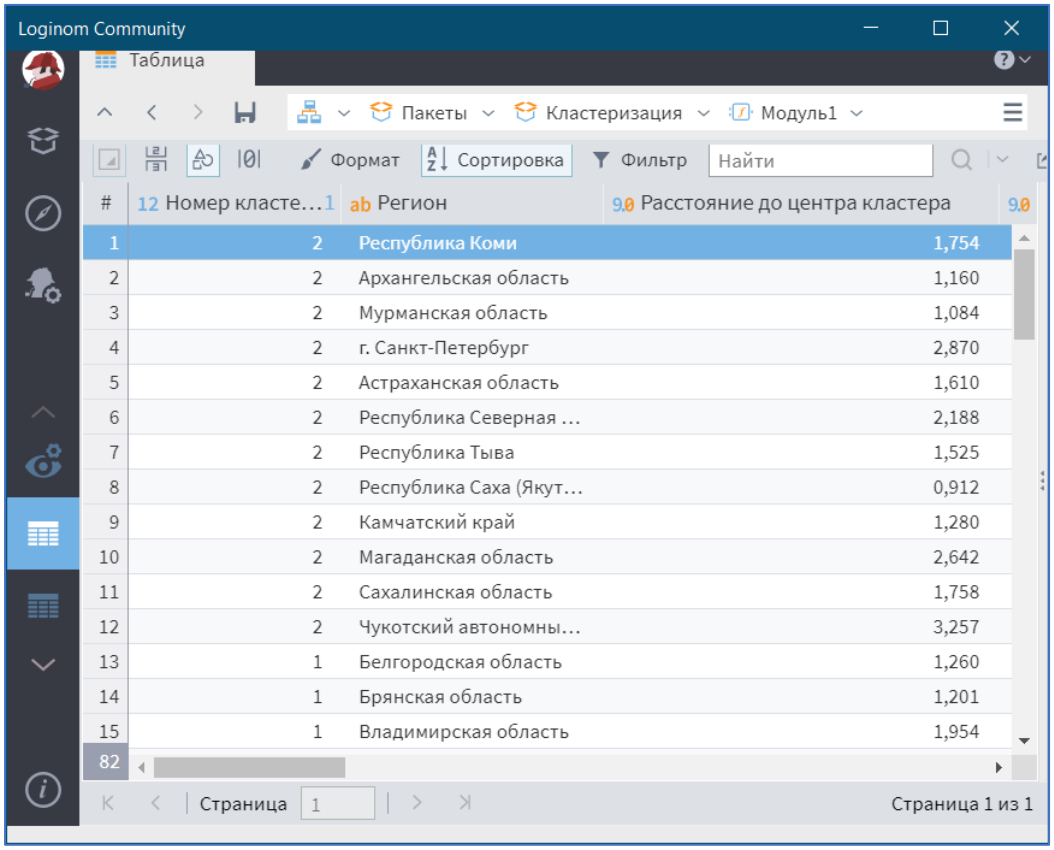


Рис. 9.15

## 9.3. Задание для самостоятельной работы

В файле *Задача 9.2. Высшее образование.xlsx* имеются данные распределения регионов Российской Федерации по обеспеченности высшими образовательными учреждениями. Для этого были отобраны пять показателей (рис. 9.16).

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Регион | 1. Число образовательных организаций высшего образования и филиалов на начало учебного года (2015-2016 гг.) на 1 млн. чел. населения | 2. Численность студентов, обучающихся по программам бакалавриата, специалитета, магистратуры на начало учебного года (2015-2016 гг.) на 10 тыс. чел. населения | 3. Прием на обучение по программам бакалавриата, специалитета, магистратуры (2015 г.) на 10 тыс. чел. населения | 4. Выпуск бакалавров, специалистов, магистров (2015 г.) на 10 тыс. чел. населения | 5. Численность профессорско-преподавательского персонала образовательных организаций высшего образования на 1000 студентов на начало учебного года (2015-2016 гг.) |
| 2 | Белгородская область | 9,7 | 342,6 | 81,3 | 95,5 | 50,7 |
| 3 | Брянская область | 15,5 | 281,4 | 59,5 | 83,2 | 41,7 |
| 4 | Владимирская область | 12,2 | 239,8 | 62,3 | 55,8 | 47,3 |
| 82 | Республика Крым | 5,8 | 238,1 | 65,5 | 38,3 | 67,0 |
| 83 | Севастополь | 16,8 | 365,4 | 98,6 | 55,3 | 71,3 |

Рис. 9.16

Требуется провести кластерный анализ регионов, используя алгоритм k-means, и выяснить существуют ли заметные различия в их обеспеченности

образовательными учреждениями. При этом необходимо задать число кластеров равным 5, исходя из предположения, что регионы могут быть отнесены к группам с лучшей, выше средней, средней, ниже средней и худшей обеспеченностью образовательными учреждениями.