Н. Ю. Прокопенко

АНАЛИТИЧЕСКИЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

на базе АП Loginom

Учебное пособие

Нижний Новгород
2020

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Нижегородский государственный архитектурно-строительный университет»

Н. Ю. Прокопенко

АНАЛИТИЧЕСКИЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ на базе АП Loginom

Утверждено редакционно-издательским советом университета
в качестве учебного пособия

Нижний Новгород
ННГАСУ
2020

ББК 32.813я73
П 78
УДК 004.89(075.8)

Печатается в авторской редакции

Рецензенты:

И.Н. Цветкова – канд. физ.- мат. наук, доцент, заведующая кафедрой информатики и информационных технологий Нижегородского института управления Российской академии народного хозяйства и государственной службы при Президенте РФ (НИУ РАНХиГС).

Е.М. Дмитриева – преподаватель кафедры информационных технологий и инструментальных методов в экономике института экономики и предпринимательства ФГАОУ ВО «Национальный исследовательский Нижегородский государственный ун-т им. Н. И. Лобачевского»

Прокопенко Н. Ю. Аналитические информационные системы поддержки принятия решений [Текст]: учеб. пособие / Н.Ю. Прокопенко; Нижегор. гос. архитектур.- строит. ун-т – Н. Новгород: ННГАСУ, 2020. – 142 с. ISBN 978-5-528-00395-5

В пособии раскрываются теоретические и практические основы использования аналитической платформы Loginom. Приведено много примеров, иллюстрирующих разработку и применение рассматриваемых методов и моделей. В заданиях представлены постановка задачи, исходные данные и последовательность их выполнения.

Предназначено для магистрантов направления 09.04.03 «Прикладная информатика» профиля «Прикладная информатика в аналитической экономике» и студентов бакалавриата направления 09.03.03 «Прикладная информатика» профиля «Прикладная информатика в экономике», изучающих в рамках дисциплин «Системы поддержки принятия решений», «Методы бизнес-аналитики» и «Бизнес-аналитика в практике предприятия» вопросы использования аналитических информационных систем поддержки принятия решений для решения бизнес-задач.

ISBN 978-5-528-00395-5

Содержание

Введение

Основной целью аналитических информационных систем является обеспечение быстрого доступа к данным, выполнение анализа данных и информационная поддержка процесса принятия решений.

Предназначение бизнес-аналитики (Business Intelligence, BI) – извлечь знания о бизнесе из данных с использованием различных аппаратно-программных технологий. Такие технологии дают возможность организациям превращать данные в информацию, а затем информацию в знания.

Настоящее учебное пособие предназначено для магистрантов и студентов, изучающих в рамках дисциплин «Системы поддержки принятия решений», «Методы бизнес-аналитики», «Бизнес аналитика в практике предприятий», вопросы использования современных корпоративных информационных систем, включающих системы обработки данных (СОД), информационные системы управления (ИСУ) и системы поддержки принятия управленческих решений (СППР).

В пособии раскрываются теоретические и практические основы использования свободно распространяемой аналитической Low-code платформы Loginom Academic 6.2.5. и Loginom Community 6.3.0-pre (https://loginom.ru/downloads). Аналитическая платформа Loginom, пришедшая на смену АП Deductor, унаследовала возможности системы предыдущего поколения и приобрела принципиально новый функционал, призванный изменить представление о доступности продвинутой аналитики.

1. Аналитически информационные системы поддержки принятия решений

В современном мире успех компании на рынке напрямую зависит от того, как быстро менеджмент компании может распознать изменения динамики рынка и насколько своевременно может отреагировать на них с целью увеличения прибыли, исходя из существующих реалий рынка. Менеджеры компании должны отслеживать тенденции рынка, идентифицировать конкурентов и угрозы, оценивать риски, оценивать свои ресурсы и т.д. Информация является необходимым производственным ресурсом для принятия эффективных управленческих решений. Компании накопили значительные объемы данных и имеют доступ к еще большим объемам внешних данных. Менеджерам необходимо, чтобы эта информация была преобразована, предварительно обработана и соответствующим образом организована для быстрого доступа, анализа и принятия решений.

Business Intelligence (сокращённо BI) – это методы и инструменты для поиска, анализа, моделирования и доставки информации, необходимой для принятия решений. Технологии BI обрабатывают большие объемы данных, чтобы найти стратегические возможности для бизнеса.

Рождение BI датируется 1958 годом, когда американский ученый Ханс Петер Лун (1896-1964) опубликовал в IBM System Journal статью «A Business Intelligence System». В ней он представил бизнес как набор различных видов деятельности в науке, технологиях, коммерции, индустрии и даже в законодательной сфере, а обеспечивающие его системы – системами, поддерживающими разумную деятельность (intelligence system).

Словом «intelligence» Лун обозначал способность устанавливать взаимосвязь между представлениями отдельных фактов и действиями в интересах решения поставленных задач и достижения намеченных целей. В 1989 году аналитик из Gartner Ховард Дреснер дал BI расширительную трактовку, предложив использовать BI в качестве общего термина для различных технологий, предназначенных для поддержки принятия решений.

Поддержка BI рассматривается как совокупность различных технологий. Среди них по-прежнему остается и классический инструмент – электронные таблицы, а также генераторы отчетов, технологии OLAP (OnLine Analytical Processing – оперативная аналитическая обработка данных), технологии разработки данных и текстов, а также многое другое.

Инструменты BI – программное обеспечение, которое позволяет бизнес-пользователям видеть и использовать большое количество сложных данных. Знания, основанные на данных, (data-based knowledge) получаются из данных с использованием инструментов business intelligence и процесса создания и ведения хранилища данных (data warehousing).

Аспекты проблемы анализа данных и необходимые для их разрешения функции нашли выражение в соответствующих программных продуктах. Соответственно средства автоматизации анализа представлены в различных видах. Имеются комплексные информационно-аналитические системы, выполняющие в той или иной степени функции в соответствии с рассмотренными аспектами. Представлены на рынке программные продукты и целевые программные системы, выполняющие в увеличенном объеме, расширенном составе и повышенной сложности какие-либо функции, например, оперативного или интеллектуального анализа.

В целом сложился рынок инструментальных средств создания и поддержки OLAP-систем, информационных хранилищ (DWH), СППР (DSS), интеллектуального анализа Data mining (DM), который получил обобщенное название – Business intelligence (BI).

Многообразие представленных на рынке решений, от мощных платформ до простых систем аналитики и отчетности, позволяет выбрать решение, доступное любой организации. Развитие средств визуального представления данных, мобильных и облачных технологий сделали BI-инструменты массовыми всего за последние несколько лет.

Крупнейшие поставщики предоставляют всевозможные решения для реализации аналитических систем, такие как SAP Business Objects (разработчик – компания SAP AG), Oracle OLAP (разработчик – Oracle Corporation), IBM Smart Analytics System, Statistical Analysis System, Microsoft, Prognoz Platform (разработчик – компания «Прогноз»), АП Loginom (разработчик – компания ООО «Аналитические технологии» Loginom Company) и др.

Компания SAP AG занимается разработкой и внедрением автоматизированных систем управления такими внутренними процессами предприятия, как: бухгалтерский учет, торговля, производство, финансы, управление персоналом, управление складами и т. д. Приложения обычно можно адаптировать под правовой контекст определённой страны. Аналитические приложения SAP работают с разнообразными источниками данных и ИТ-средами, в них предустановлены инструменты управления данными для той или иной отрасли или сектора.

Oracle (Oracle Corporation) – американская корпорация, крупнейший в мире разработчик программного обеспечения для организаций. Oracle Business Intelligence Suite – открытое, основанное на стандартах программное обеспечение, предоставляет единую, интегрированную инфраструктуру для бизнес-анализа, включающую комплексный набор продуктов для обработки запросов и проведения анализа, формирования корпоративных отчетов, доступа к средствам анализа с мобильных устройств, использования информационных панелей и порталов, интеграции с Microsoft Office и Excel, управления процессами бизнес-анализа, рассылки уведомлений в реальном времени, мониторинга бизнес-деятельности и множество других возможностей.

Система IBM Smart Analytics System предоставляет гибкий набор функциональных возможностей, включая бизнес-анализ, аналитическую отчетность, оценочные таблицы, инструментальные панели, извлечение информации из данных, сервисы Cubing Services (обеспечивающие многомерную визуализацию данных), текстовый анализ, управление хранилищем данных, а также серверную платформу и среду хранения данных.

SAS Institute Inc. (Statistical Analysis System) – американская частная компания, разработчик технологического программного обеспечения и приложений класса Business Intelligence, Data Quality и Business Analytics.

Компоненты аналитики SAS:

- Прогнозная аналитика и интеллектуальный анализ данных (Data Mining) – позволяет строить описательные и прогнозные модели, редактировать их и интегрировать результаты во все бизнес-процессы организации в единой среде.
- Визуализация данных – Повышает эффективность аналитики с помощью динамической интерактивной визуализации данных.
- Прогнозирование и эконометрика – дает возможность анализировать и предсказывать будущие результаты на основе исторических тенденций.
- Управление и мониторинг аналитических моделей – значительно упрощает процесс создания аналитических моделей, их управления и передачи на регламентное применение.
- Исследование операций и оптимизация – предлагает методы оптимизации и планирования для поиска наилучшей стратегии решения бизнес-задач.
- Контроль качества – наблюдение процессов и измерение качества процессов во времени.
- Статистика – позволяет использовать статистический анализ данных для принятия решений на основе исторических фактов.
- Текстовая аналитика – дает возможность максимально использовать информацию, скрытую в огромных объемах неструктурированных данных.

Microsoft поставляет BI-инструменты в составе трех групп продуктов – Excel, SharePoint и SQL Server. В каждом из них имеются свои функции аналитики и коллективной работы.

Аналитические платформы – это специализированные программные решения, которое содержат все инструменты, необходимые для осуществления про-

цесса извлечения скрытых зависимостей и закономерностей из больших массивов данных, предназначенные для формирования отчётности, управления бизнес-процессами, моделирования и прогнозирования показателей, визуализации и оперативного анализа данных, создания бизнес-приложений, в том числе веб-приложений и мобильных приложений.

В аналитических платформах присутствуют гибкие и развитые средства консолидации, включающие интеграционные механизмы с промышленными источниками данных, инструменты очистки и преобразования структурированных и неструктурированных данных, их последующее хранение в едином источнике – хранилище данных. Модели, описывающие выявленные зависимости и закономерности, правила, прогнозы также хранятся в специальном репозитарии моделей.

Аналитическая платформа обычно состоит из следующих компонентов:

- аналитический сервер
- клиентское приложение
- подсистема управления метаданными
- репозитарий моделей
- интеграционный сервер

Prognoz Platform – платформа бизнес-аналитики для создания информационных систем и применения в качестве самостоятельного решения. Prognoz Platform позволяет разрабатывать приложения (в том числе мобильные) для оперативного анализа данных, а также моделирования бизнес-процессов и прогнозирования. Платформа разработана компанией «Прогноз».

Loginom – аналитическая платформа, предоставляющая возможности глубокой аналитики и позволяющая принимать управленческие решения, основанные на точной и достоверной информации. Loginom пришла на смену платформе предыдущего поколения – Deductor. Платформа Loginom разработана компанией Loginom Company (ранее BaseGroup Labs ООО «Аналитические технологии»).

Ключевые возможности:

- Проведение сложных расчетов;
- Консолидация данных;
- Очистка данных;
- Прогнозирование и оптимизация данных;
- Интеграция с любыми хранилищами данных, включая базы данных, отдельные файлы, учетные системы, социальные сети, веб-сервисы.
- Комбинация из структурного и объектно-ориентированного подходов к моделированию.

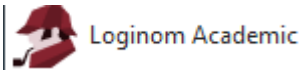
Loginom полностью построен на базе веб-технологий. Веб-интерфейс обладает рядом существенных преимуществ:

- браузеры функционируют на любых ОС (Windows, MacOS, Linux и др.) и множестве устройств;
- минимизируются усилия по администрированию, так как отсутствует необходимость инсталляции и настройки рабочих мест.

Платформа обладает пользовательским интерфейсом, не требующим для работы специальной подготовки. Loginom имеет поддержку технологий анализа: от простой логики до машинного обучения.

1.1. Интерфейс АП Loginom Studio

АП Loginom Academic – бесплатная версия для образовательных целей. В платформу встроены современные методы извлечения, визуализации и анализа данных. Данная версия платформы предназначена для установки на локальной машине, когда подразумевается работа одного пользователя.

Ярлык для запуска программы 

.

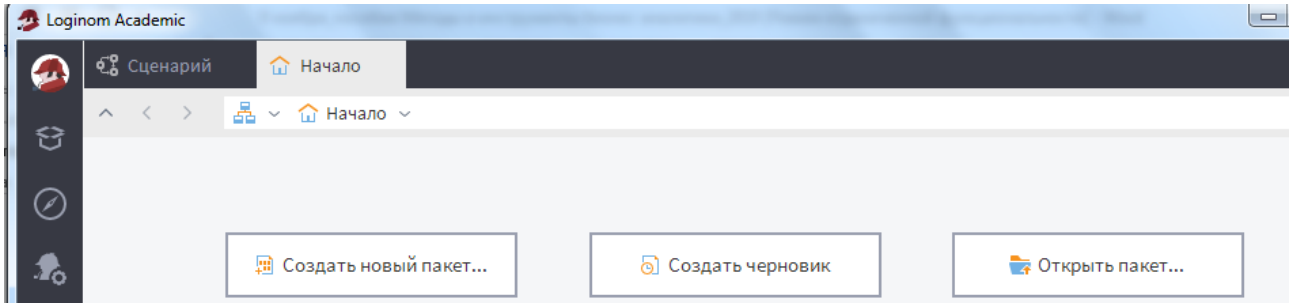


Рис. 1. Начало работы АП Loginom

После создания или открытия пакета на странице «Начало» появится главное рабочее окно программы, состоящее из четырех основных блоков (см. рис. 2):

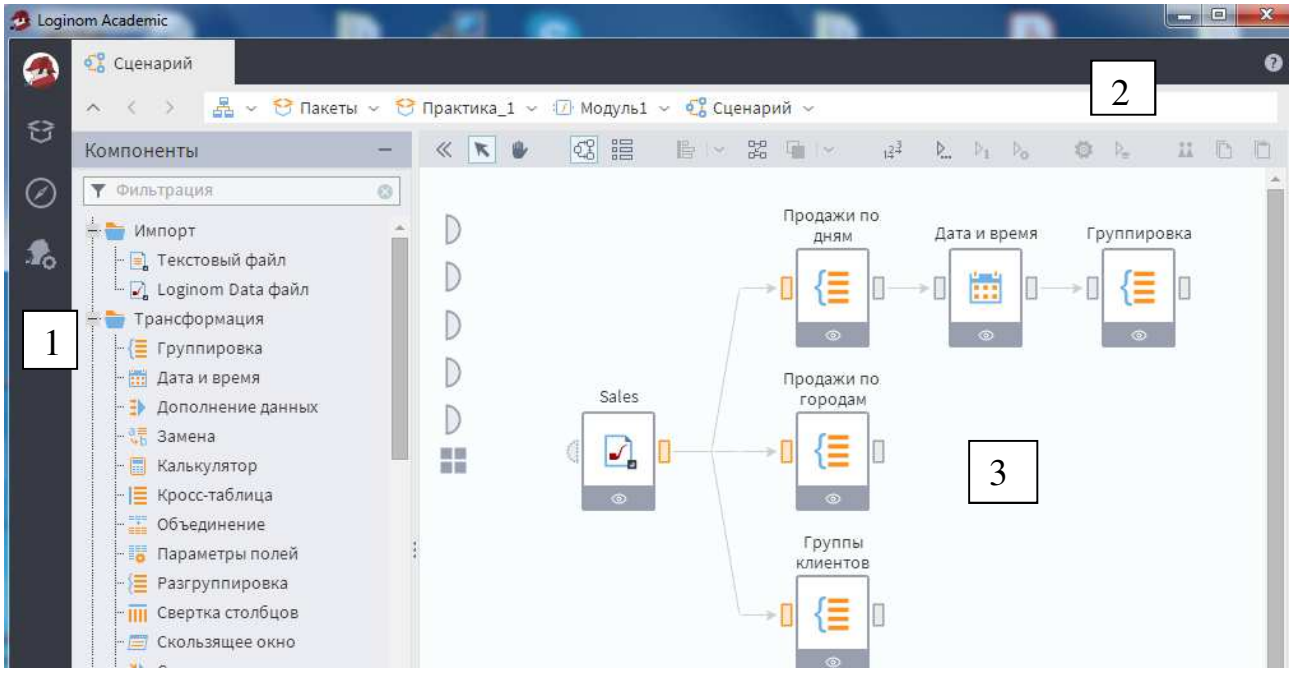


Рис. 2. Рабочее окно Loginom

1. Главное меню – панель с кнопками (Меню, Пакеты, Навигация, Администрирование) для манипуляции с различными настройками;

2. Адресная строка – строка, содержащая путь к открытому объекту;

3. Рабочее пространство – панель компонентов, панель инструментов и область построения Сценариев.

Нажатие на кнопку Пакеты, откроет список команд для работы с ними.

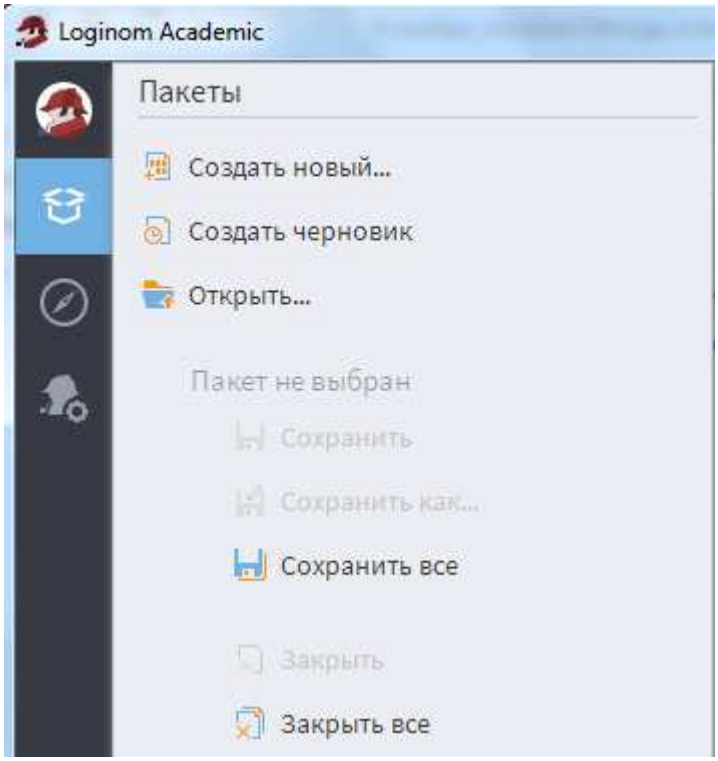


Рис. 3. Список команд для работы с пакетом

Важно: программа не поддерживает автосохранений, при закрытии окна программы (вкладки браузера) все изменения будут утеряны.

Пакет – важное понятие платформы Loginom. Все действия с проектом в Loginom Studio осуществляются в рамках Пакета, который является минимальной единицей поставки и представляет собой контейнер для компонентов, сценариев, подключений и т.д.

Пакеты сохраняются по-отдельности в виде файлов с расширением *.lgp, и включают в себя Ссылки и Модули.

Одновременно можно открыть любое количество пакетов и работать с ними параллельно.

Нажатие на кнопку Навигация откроет древовидную структуру объектов. Это так называемое дерево пакетов. Значок «+» указывает на наличие иерархии.

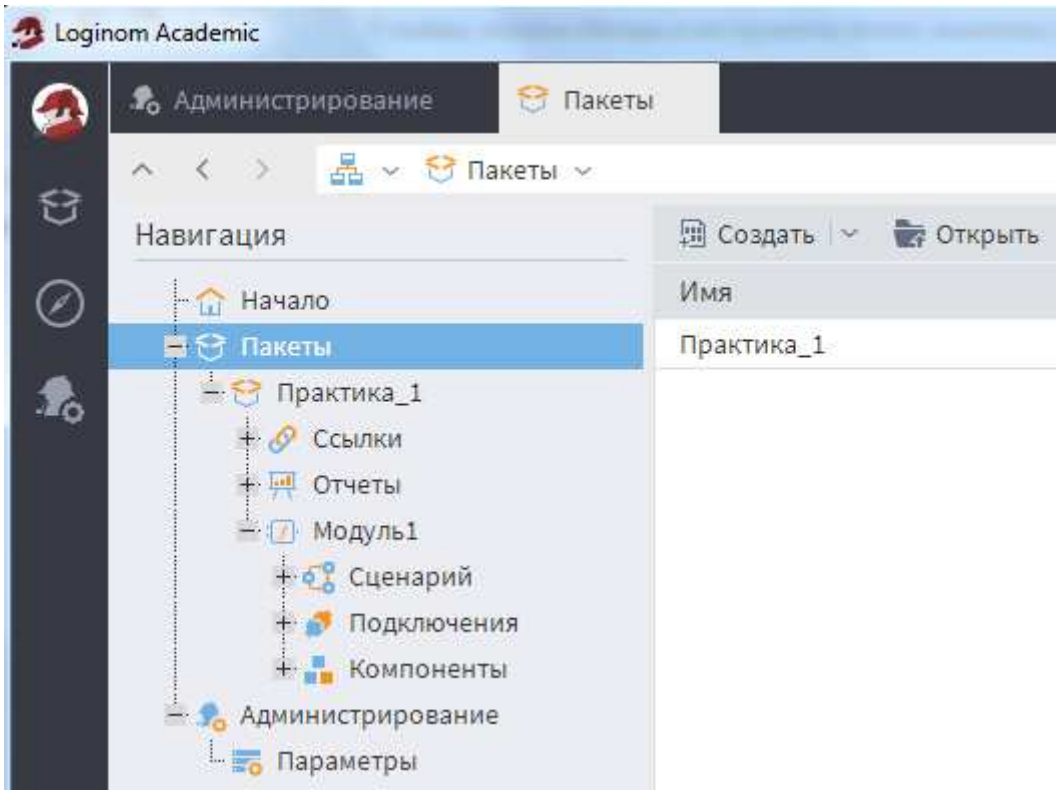


Рис. 4. Древовидная структура объектов на панели «Навигация»

Каждый пакет состоит из трех групп объектов: Ссылки, Отчеты и Модули (рис. 4.).

Ссылки применяются для подключения других пакетов с целью использования созданных в них производных компонентов и подключений в текущем проекте. Соответствующие объекты доступны только в том случае, когда они опубликованы для общего доступа.

Отчеты представляют собой визуализаторы, настроенные для набора данных.

Модуль включает в себя:

- Сценарий – содержит последовательность узлов обработки данных;
- Подключения – в них представлен список внешних источников и приемников данных, к которым можно подключиться;
- Компоненты – включают в себя доступные для работы подмодели, как созданные в рамках текущего пакета, так и заимствованные из других пакетов через ссылки.

- Содержимое объекта пакета отображается в рабочей области окна (рис. 5):

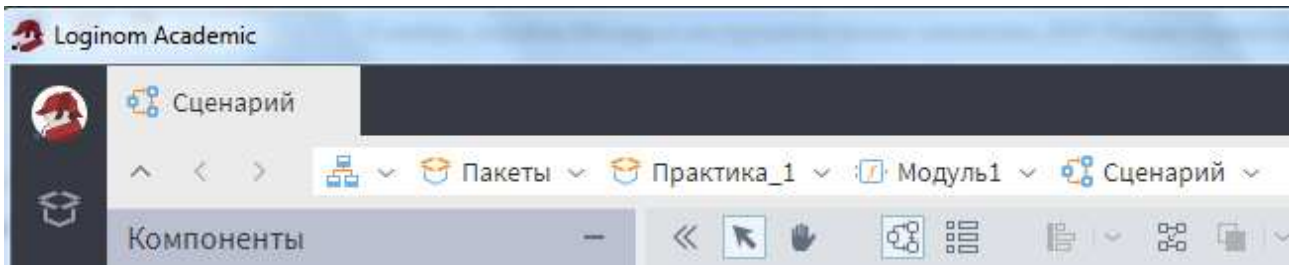


Рис. 5. Горизонтальная панель навигации

Горизонтальная панель навигации содержит путь к активному элементу, а также элементы навигации: пиктограммы «На уровень вверх», «Назад», «Вперед», которые позволяют с делать один шаг в прямом или обратном направлении, либо подняться на самый верхний уровень к списку пакетов.

Справа от названия объекта есть стрелка, щелкнув по ней кнопкой мыши, на экране появиться список объектов (сценарий, подключения, компоненты).

Модуль содержит три элемента: сам сценарий, который проектирует пользователь; подключения и компоненты.

Сценарий – главная часть модуля, представляет собой последовательность шагов по обработке данных, которые задаются узлами из компонентов стандартных или производных.

Область построения Сценария – полотно, содержащее узлы Сценария и связи между ними.

Сверху расположена панель инструментов, содержащая следующие операции для манипуляции с областью построения и ее составляющими:

 Показать/ Скрыть панель компонентов – позволяет открыть или закрыть панель компонентов;

 Режим выбора объекта – режим, использующийся для построения Сценария с помощью стандартных манипуляций;

 Режим навигации по сценарию – режим, использующийся для навигации по области построения Сценария с помощью мышки;

 Показать в виде сценария – отображает Сценарий в стандартном виде (в виде направленного графа);

 Показать в виде таблицы – компактное отображение Сценария в виде таблицы, содержащей используемые элементы;

 Вертикальное выравнивание – позволяет выравнить вертикально узлы Сценария на области построения. Имеются следующими виды вертикального выравнивания: по левому краю; по середине; по правому краю; по верхнему краю; по центру; по нижнему краю.

 Автоматическое упорядочивание узлов – автоматическое расположение узлов на области Сценария в соответствии с их последовательностью обработки данных;

 Переместить выделенные узлы – выставляет выделенные узлы и их подписи на передний план или задний план.

 Настроить порядок выполнения – позволяет задать собственный порядок выполнения узлов;

 Выполнить все – выполнить все узлы Сценария;

 Активировать/Деактивировать узел;

 Переобучить узел – переобучает выделенный узел;

 Настроить узел – заходит в настройки выделенного узла;

 Настроить режим активации узла – настройка режима активации выделенного узла;

 Клонировать узел – клонирование выделенного узла;

 Развернуть/Свернуть подмодель – позволяет свернуть выделенные узлы в Подмодель или развернуть выделенную Подмодель на составные узлы;

 Копировать узел/ вставить узел;

 Удалить выбранное – удаляет выделенные узлы/связи Сценария;

 Создать производный компонент – создает Производный компонент на основе выделенного узла;

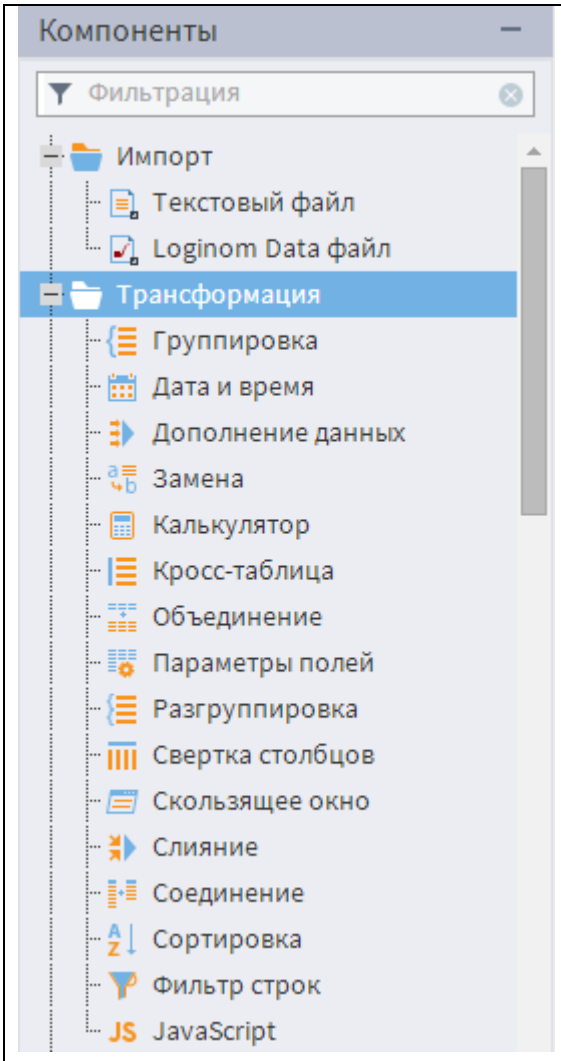
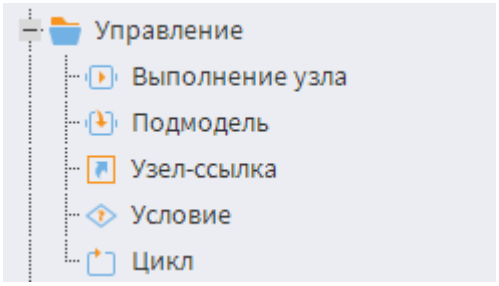
 Показать родительские узлы для производных – при наличии производных узлов показывает родительские узлы;

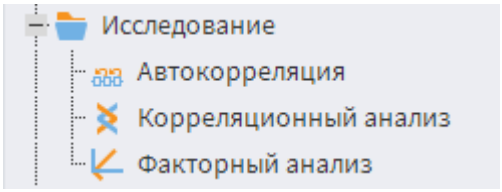
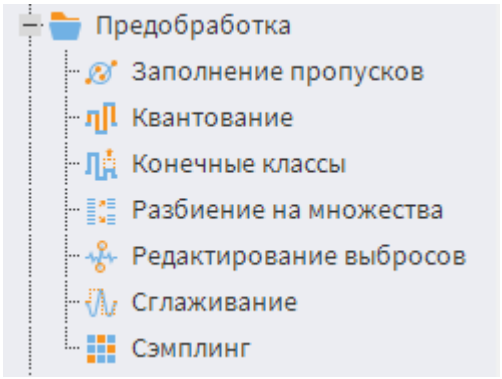
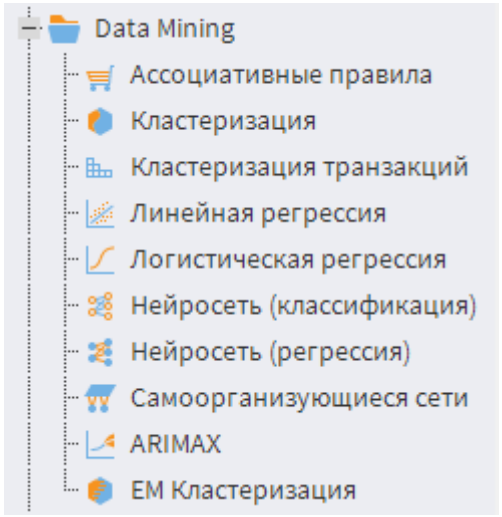
 Показать исходные узлы для Узлов-ссылок – при наличии Узлов-ссылок показывает узлы, на основе которых они создавались;

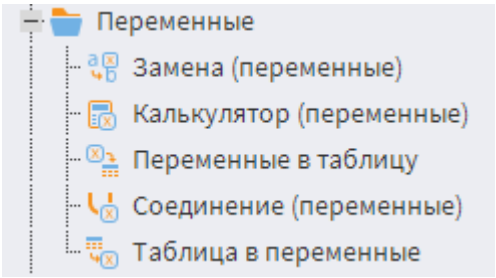
 Показать карту сценария – для навигации открывается уменьшенная копия области построения Сценария с возможностью масштабирования.

При переходе в сценарий открывается панель компонентов. Стандартные компоненты объединены в группы: Импорт, Трансформация, Управление, Исследование, Предобработка, Data Mining, Переменные, Интеграция, Экспорт.

Таблица 1. Панель стандартных компонентов

	Трансформация – набор компонентов для первоначальной подготовки и простой обработки исходных наборов данных. • Группировка • Дата и время • Дополнение данных • Замена • Калькулятор • Калькулятор • Кросс-таблица • Объединение • Параметры полей • Разгруппировка • Свёртка столбцов • Скользящее окно • Слияние • Соединение • Сортировка • Фильтр строк
	Управление Компоненты группы предназначены для оптимизации сценариев путем создания подмоделей и повторного использования узлов, а также формирования логики выполнения сценариев при помощи условий и циклов. • Выполнение узла • Подмодель • Узел-ссылка • Условие • Цикл

	<u>Исследование</u> С помощью этих компонентов можно оценить и/или визуализировать структуру и статистические характеристики данных. Также с их помощью проводятся разведочный и описательный анализы. Автокорреляция Корреляционный анализ Факторный анализ
	<u>Предобработка</u> Предварительная обработка данных для их дальнейшего использования в алгоритмах Data Mining. Применяются такие методы, как заполнение пропусков, сэмплинг, редактирование выбросов и другие. • Заполнение пропусков • Квантование • Конечные классы • Разбиение на множества • Редактирование выбросов • Сглаживание • Сэмплинг
	<u>Data Mining</u> Компоненти, выделенные в эту группу, являются инструментами для реализации различных методов Data Mining: кластеризация, ассоциативные правила и другие. • Ассоциативные правила • Кластеризация • Кластеризация транзакций • Линейная регрессия • Логистическая регрессия • Нейросеть (классификация) • Нейросеть (регрессия)

	• Самоорганизующиеся сети • ARIMAX • EM Кластеризация
	Переменные В Loginom имеется возможность создавать и использовать переменные. Компоненти этой группы позволяют проводить различные операции над ними: изменение, создание переменных из таблицы, расчет новых переменных с помощью различных функций. • Замена (переменные) • Калькулятор (переменные) • Переменные в таблицу • Соединение (переменные) • Таблица в переменные

1.2. Проектирование сценариев

Проектирование сценариев – описание режимов работы, создание собственных компонентов, работа с портами и переменными.

Аналитическая платформа Loginom позволяет разрабатывать сценарии как «снизу вверх», так и «сверху вниз», что позволяет выбирать тот подход, который позволит построить лучшее решение в каждом конкретном случае.

При проектировании «снизу вверх» любой сценарий анализа всегда будет начинаться с загрузки источников данных. На следующем уровне происходит первый шаг абстрагирования от данных, на котором производится их первичная обработка – очистка, группировка, агрегирование и т.д. На этом уровне аналитик оперирует уже не самими, конкретными данными, а метаданными (именами и свойствами полей данных, параметрами обработки). Следующий уровень – применение к результатам предыдущего уровня, тех или иных компонентов, которые, в результате своей работы порождают новые данные, ещё более абстраги-

рованные от первоначальных. И так продолжается до тех пор, пока не принимается решение, что поставленная задача решена. При этом оказывается, что на верхних уровнях сценария аналитик оперирует в основном условиями задачи, почти абстрагировавшись от исходных данных.

Рис. 6. Метод проектирования сценария «снизу вверх»

Преимущества проектирования «снизу вверх»:

- более прозрачную, простую и понятную для аналитика структуру сценария, а также его бизнес-логику;
- скорость и простоту разработки;
- проще искать ошибки в сценарии.

К недостаткам подхода могут быть отнесены:

- ориентированность на единичные задачи;
- сложность повторного использования в аналогичных задачах;

- необходимость при изменении структуры данных вносить изменения во всём сценарии.

По мере того, как технологии интеллектуального анализа данных получали всё более широкое распространение и проникали в различные сферы человеческой деятельности, в некоторых отраслях начала формироваться свои методологии и подходы к их использованию для поддержки принятия управленческих решений.

Рис.7. Метод проектирования сценария «сверху вниз»

При проектировании «сверху вниз», то есть в отсутствии наборов данных, сначала строятся уровни сценария, максимально абстрагированные от конкретных данных, а разработка производится по исходящей – к всё более конкретным действиям. Пройдя весь путь разработки, в конце ясно, какие именно данные потребуются и в каком виде они должны быть представлены.

Сценарий по умолчанию пустой и заполняется необходимыми компонентами в зависимости от решаемой задачи путем их добавления в область Сценария. Компонент при добавлении в область построения сценария превращается в «узел». Узел сценария выполняет отдельную операцию над данными.

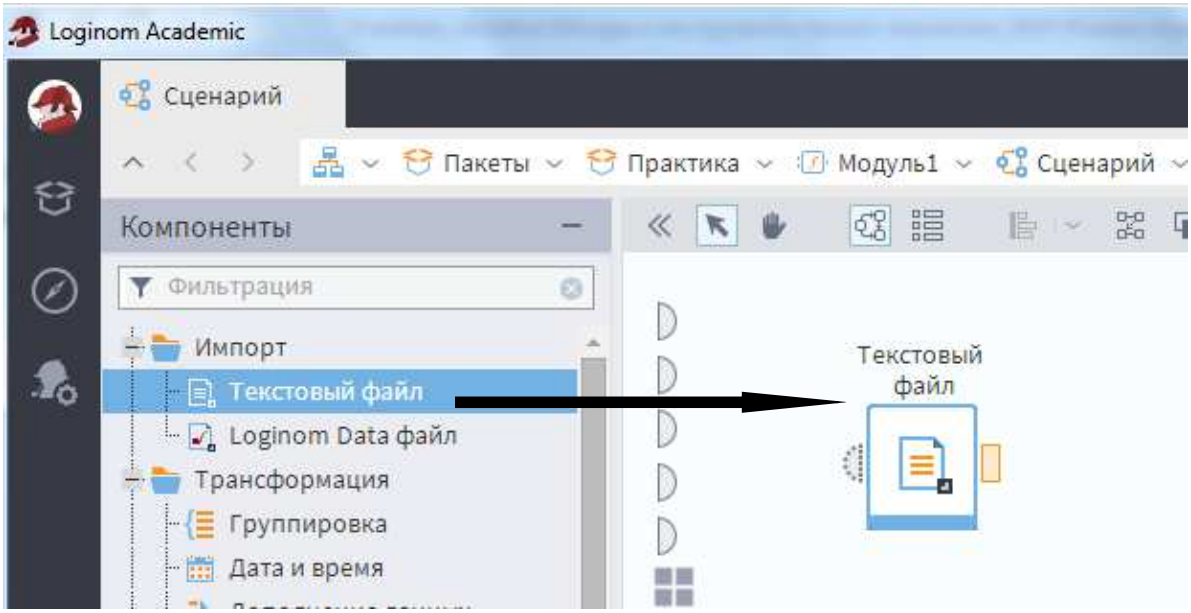


Рис. 8. Область построения сценария

Для того чтобы использовать в сценарии какой-либо компонент, его необходимо перенести мышью из панели компонентов в область построения сценария. Второй способ – вызвать контекстное меню на нужном компоненте и нажать команду *Добавить узел в сценарий*.

Входом и выходом узла являются *входные* и *выходные порты*.

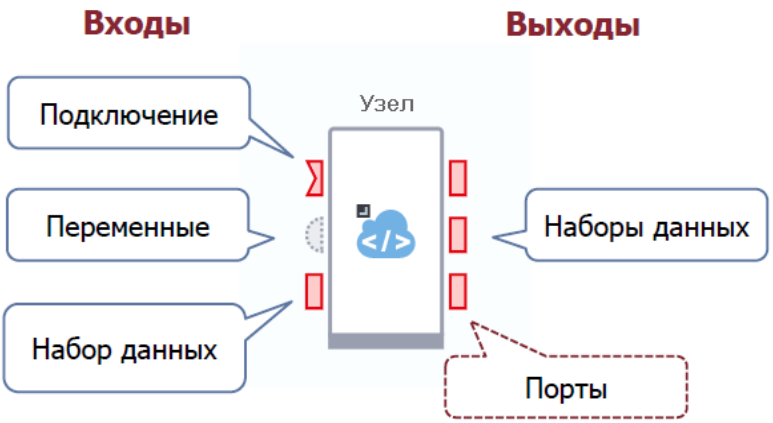


Рис. 9. Объект «Узел»

На вход узла могут подаваться таблицы, переменные, подключения. На выходе могут быть таблицы и переменные. Поскольку таблицы, переменные и подключения имеют разную структуру, то соответствующие им порты не могут быть соединены друг с другом и имеют разное обозначение.

Количество входов и выходов узла варьируется в зависимости от функционала. Входы узла могут настраиваться автоматически (при подключении связи), либо вручную.

Таблица 2. Виды портов узлов сценария

Порт	Описание
Таблица	Представляет собой структурированный набор данных, где все данные упорядочены в двумерную структуру, состоящую из столбцов и строк. В ячейках такой таблицы содержатся элементы данных: строки, числа, даты, логические значения.
Переменные	Представляют собой объекты, содержащие только одно значение. С помощью специальных компонентов имеется возможность преобразовать данные из таблиц в переменные и обратно.
Подключения	Определяют настройки для работы с внешними источниками и приемниками данных.

Последовательность обработки задается соединением выхода предыдущего узла сценария с входом последующего.

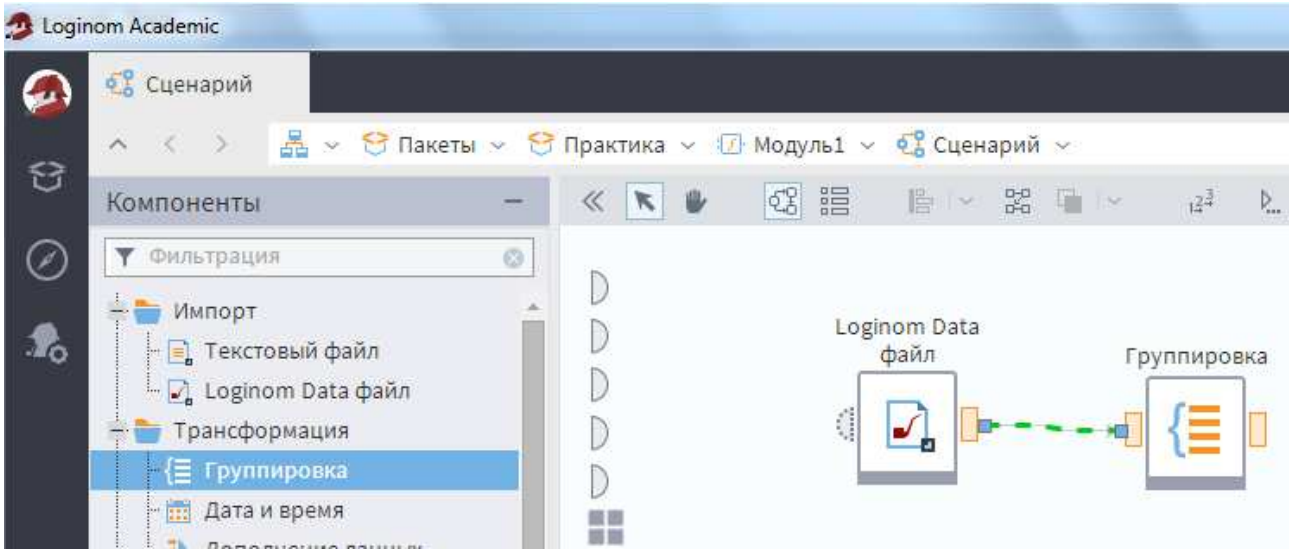


Рис. 10. Формирование связи между узлами

Удалить связь легко, для этого нужно ее выделить и нажать на кнопку  . Этой же кнопкой можно удалить узел, предварительно его выделив.

Основной функционал узла настраивается с помощью *мастера настройки.* Для этого сначала нужно узел выделить, щелкнув по нему мышкой. Его границы станут голубого цвета.

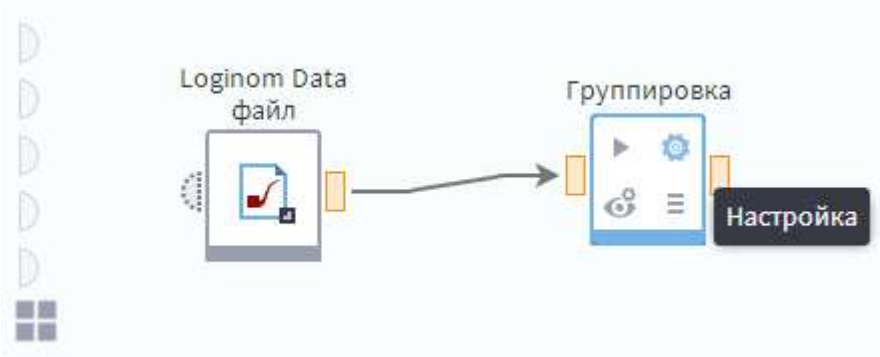


Рис. 11. Вызов мастера настройки

Обычно настройка узла включает несколько шагов:

- настройка функционала узла;
- настройка соответствия столбцов для выходного набора данных: слева отображаются поля внутри узла, справа – поля, подаваемые на выход узла. На этом этапе возможно изменять имена и метки полей (метка – то, что отображается на экране, имя – то, что используется в коде);
- настройка описания узла. На этом шаге можно задать имя узла и комментарий к узлу.

Выход из мастера осуществляется с сохранением или без сохранения внесенных изменений.

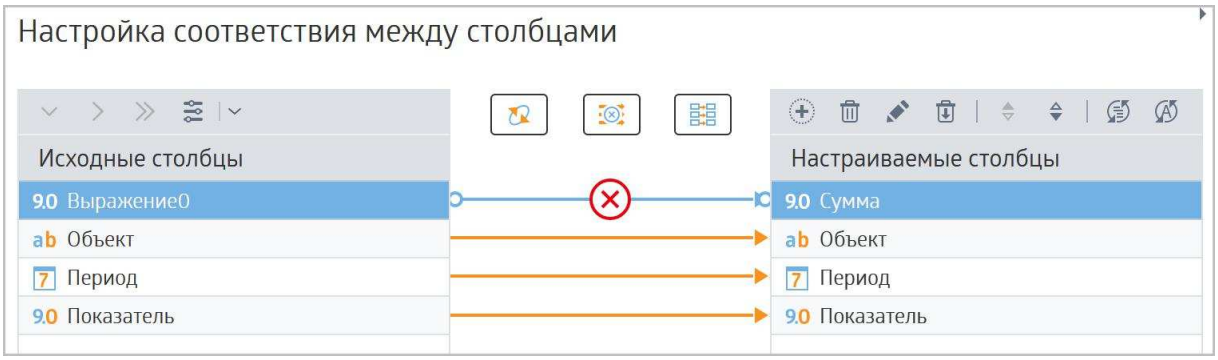


Рис. 12. Настройка выходных полей

Другие действия с узлом включают команды работы с узлом. Частично действия продублированы пиктограммами на панели инструментов сверху рабочей области.

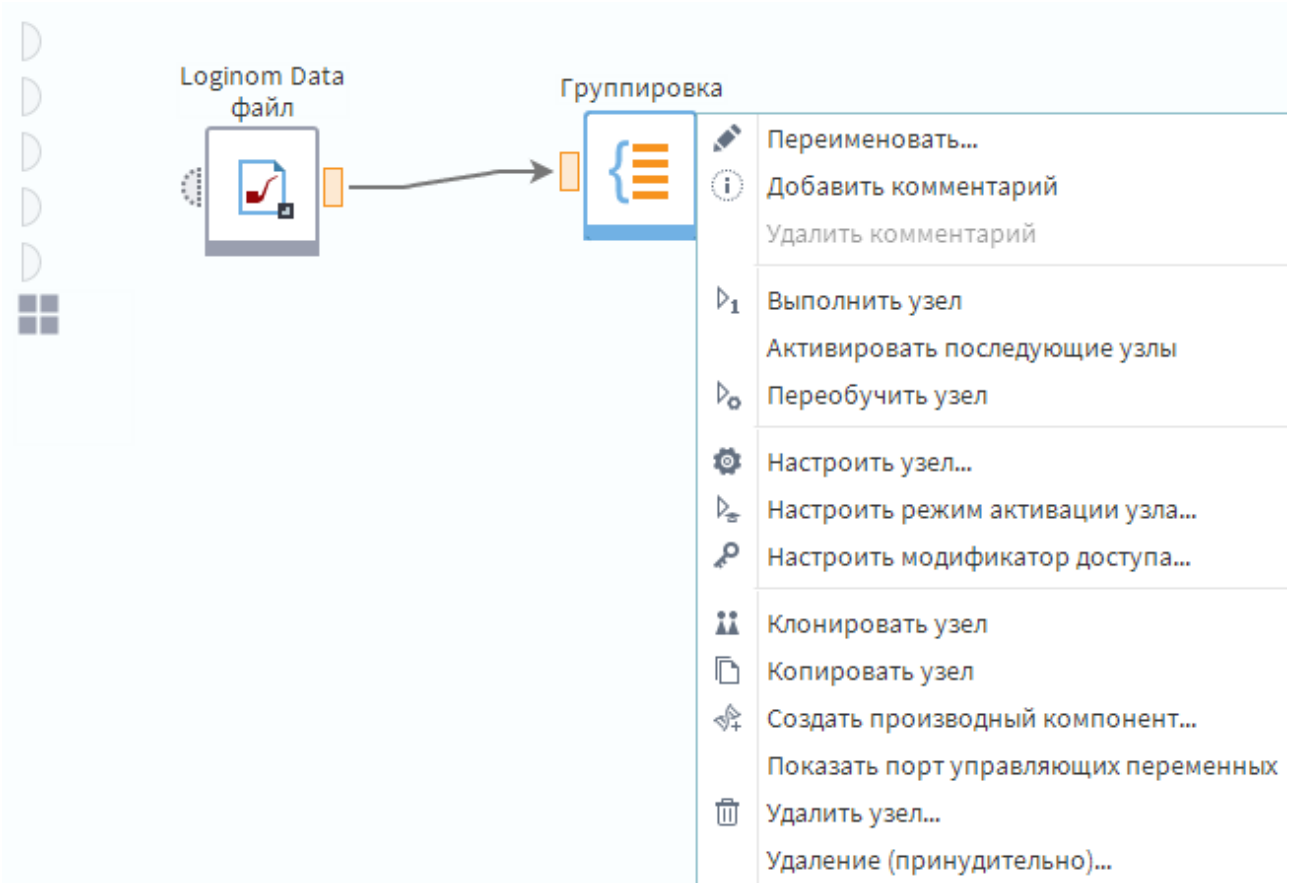


Рис. 13. Действия с узлом

Запуск узла обработки переводит его в активное состояние.

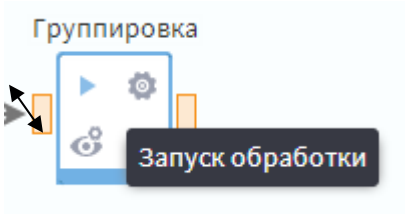


Рис. 14. Кнопка запуска узла

После запуска узла на обработку он может принять одно из двух состояний: успешное или неуспешное. На рис. 15. показано неуспешное состояние – контуры узла стали красными.

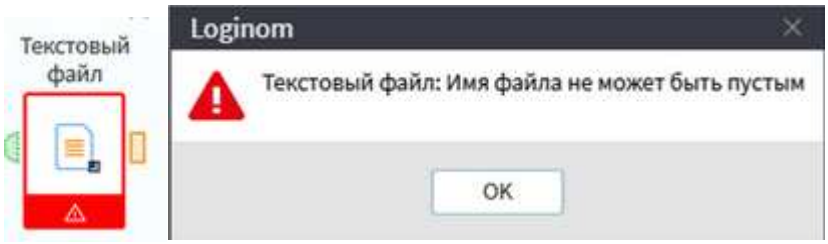


Рис. 15. Неуспешное состояние узла

Если узел отработал успешно, его контур станет зеленым. Контуры портов тоже станут зелеными, что означает, что они активированы и в них есть какие-то данные.

Для узла можно настроить визуализаторы (рис. 16) – диаграмму, таблицу или куб – путем перетаскивания элементов в рабочую область и дальнейшей их настройки.

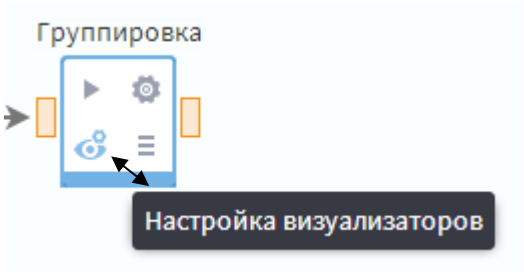


Рис. 16. Настройка визуализаторов для узла

Узлы сценария создаются из компонентов 2-х типов:

- Стандартные компоненты – предоставляются в рамках платформы;
- Производные компоненты – создаются и настраиваются пользователем.

Производный компонент можно создать из комбинации узлов сценария, реализующих произвольную логику обработки.

Чаще всего для создания производного компонента используется Подмодель.

Подмодель является специальным узлом, способным включать в себя другие узлы сценария. Удобным функционалом является *сворачивание в подмодель* – создание подсистем обработки данных, решающих конкретную задачу или часть задачи.

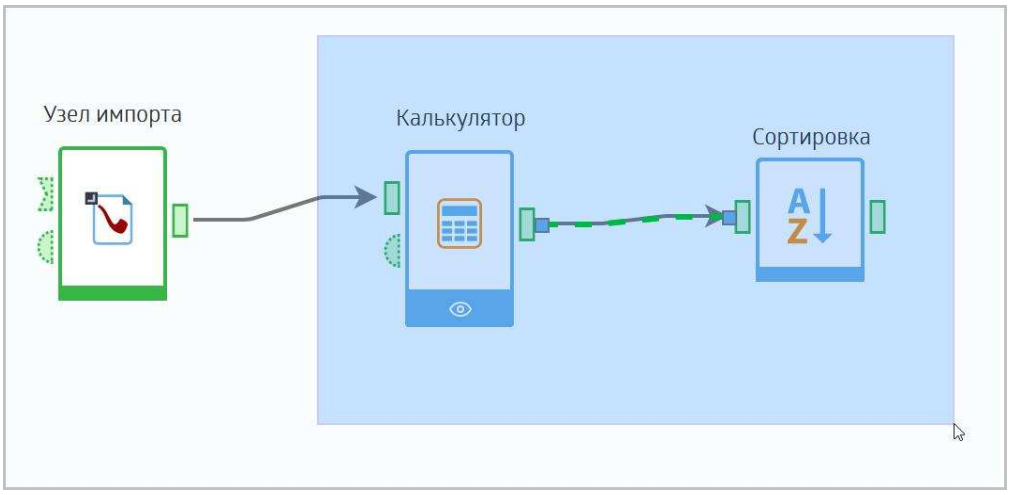


Рис. 17. Выделение узлов для сворачивания их в подмодель

При необходимости подмодель можно развернуть, восстановив исходную структуру сценария. Для разворачивания подмодели необходимо использовать соответствующую пиктограмму.

На рис. 18 «Пример сценария» узел «ABC-анализ» является производным компонентом – подмоделью.

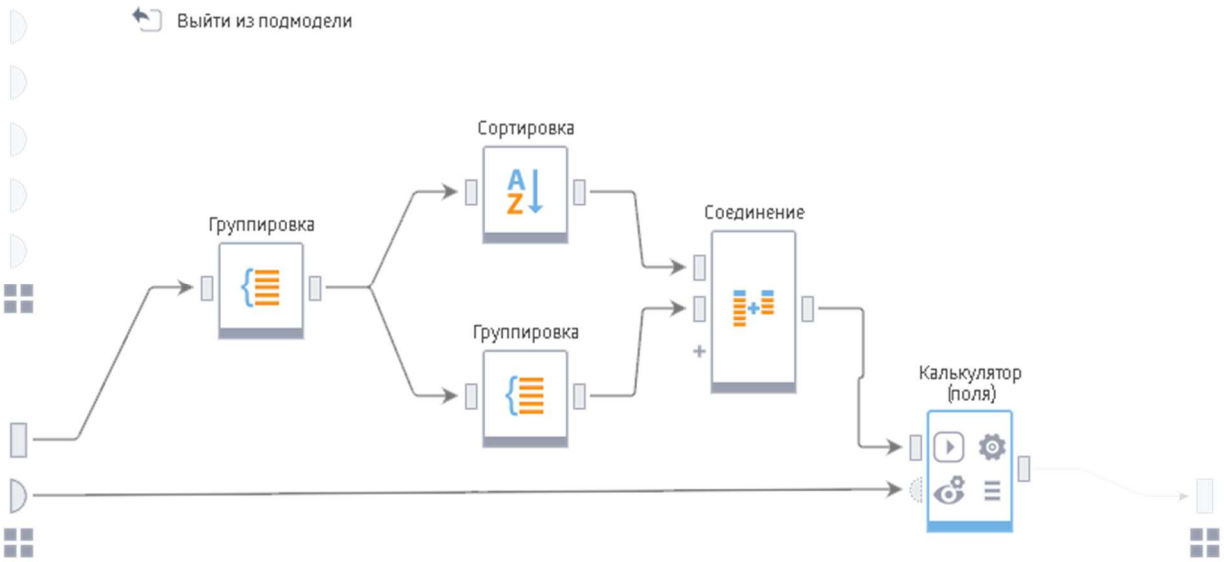


Рис. 18. Узлы подмодели «ABC-анализ»

Реализованная в подмодели логика может быть произвольной, при этом разработчик сценария может рассматривать её как «черный ящик». Подмодель принимает информацию через входные порты, производит обработку и выдает

результат на выходные порты. Входные и выходные порты задаются пользователем.

Произвольное количество входов и выходов подмодели обеспечивает высокую пластичность сценария.

Возможность параметрической настройки подмоделей позволяет гибко реализовывать логику анализа: задавать пороги, коэффициенты, пути к файлам и так далее.

В состав подмодели могут также включаться и другие подмодели. Вложенность подмоделей друг в друга не ограничена.

Создание подмоделей позволяет структурировать сценарий.

Правильно спроектированные и корректно названные подмодели помогают «читать» сценарий и понимать логику обработки.

Единожды спроектированную подмодель можно повторно использовать в других сценариях.

1.3. Лабораторная работа «Базовые навыки работы в АП Loginom»

Задание: Создадим простой сценарий, формирующий список 10 самых прибыльных групп товаров.

Сценарий выполнит действия:

- Импорт из файла Sales.lgd информации о продажах;
- Выделение 10 групп товаров с наибольшими суммами продаж;
- Экспорт полученных результатов.

1. Загрузите аналитическую платформу Loginom, создайте новый проект и сохраните его под именем лаб_раб_1.lgp.
2. Для того чтобы использовать в сценарии какой-либо компонент, его необходимо перенести мышью из панели компонентов в область построения сценария.

Выберите в разделе *Импорт* компонент *Loginom Data файл* и перенесите его в область построения. При этом создастся узел сценария, выполняющий действия импорта. При клике мышкой на узле отобразятся иконки возможных действий.

3. Вызовите *Мастер настройки* (рис. 20).

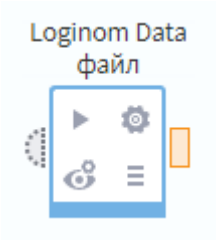


Рис. 20. Вызов мастера настройки узла

Укажите в параметре *Имя файла* местоположение файла Sales.lgd.

При выборе файла для импорта лучше использовать относительный путь, это означает, что файл с данными должен находиться в той же папке, что и файл проекта. Это позволит не перенастраивать узлы импорта при изменении местоположения папки на жестком диске и переносе сценариев с одного компьютера на другой.

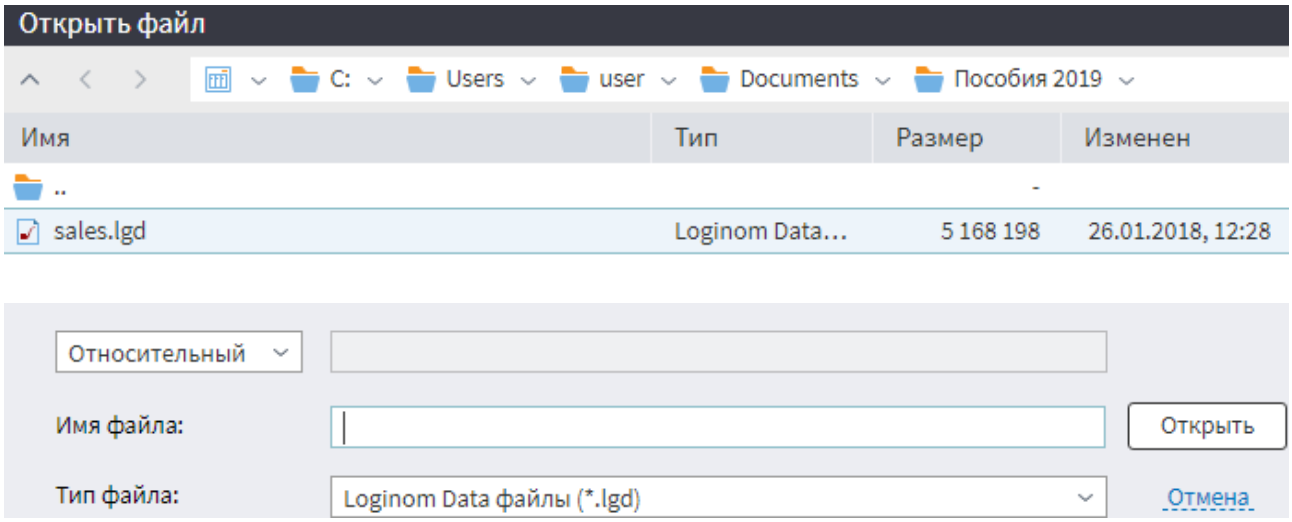


Рис. 20. Относительный путь для импорта данных

4. После настройки узла выполните его, используя меню возможных действий. Теперь в выходном порте узла присутствуют импортированные данные, которые можно увидеть, выбрав *Быстрый просмотр* в контекстном меню порта (рис. 21).

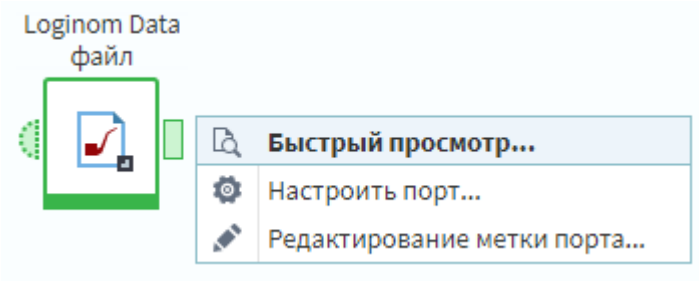


Рис. 21. Вызов быстрого просмотра

Если данные были успешно получены, то контур узла импорта будет зеленого цвета, иначе красного. Можно указать визуализаторы, которые будут использованы для отображения импортированных данных.

5. Настройте следующие визуализаторы к узлу импорта: Таблица, Статистика.

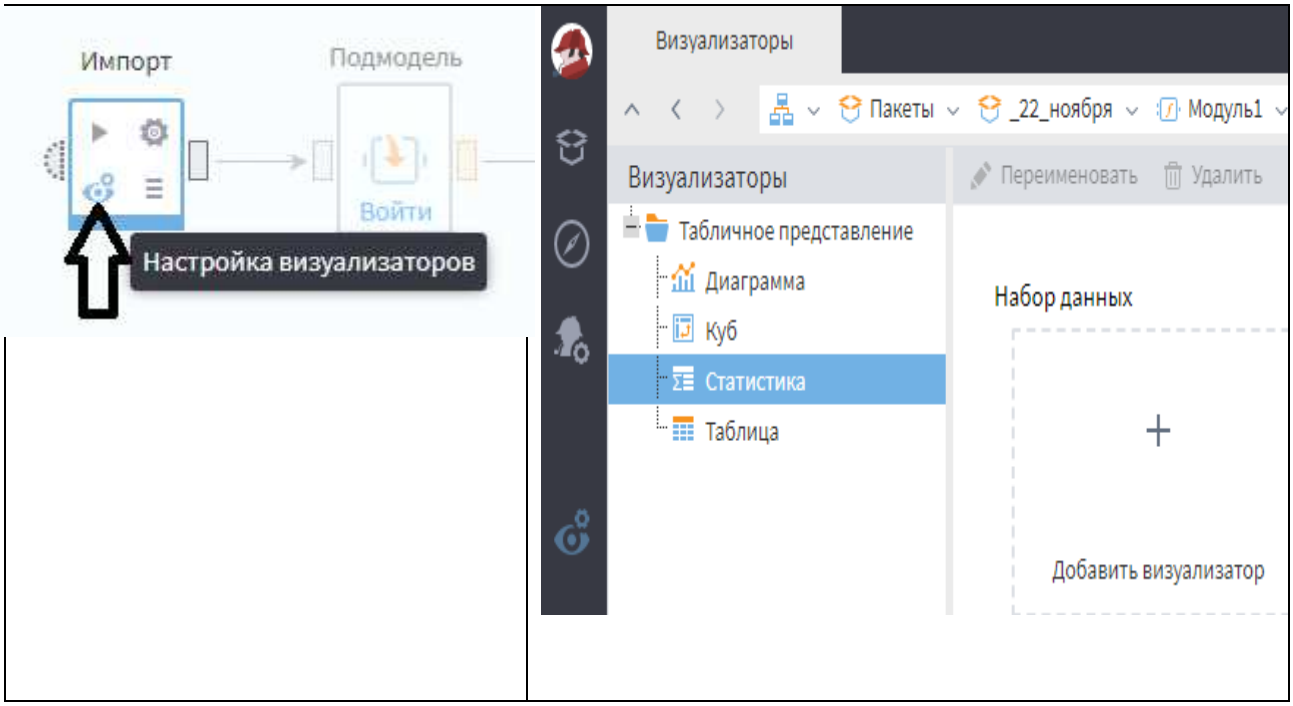
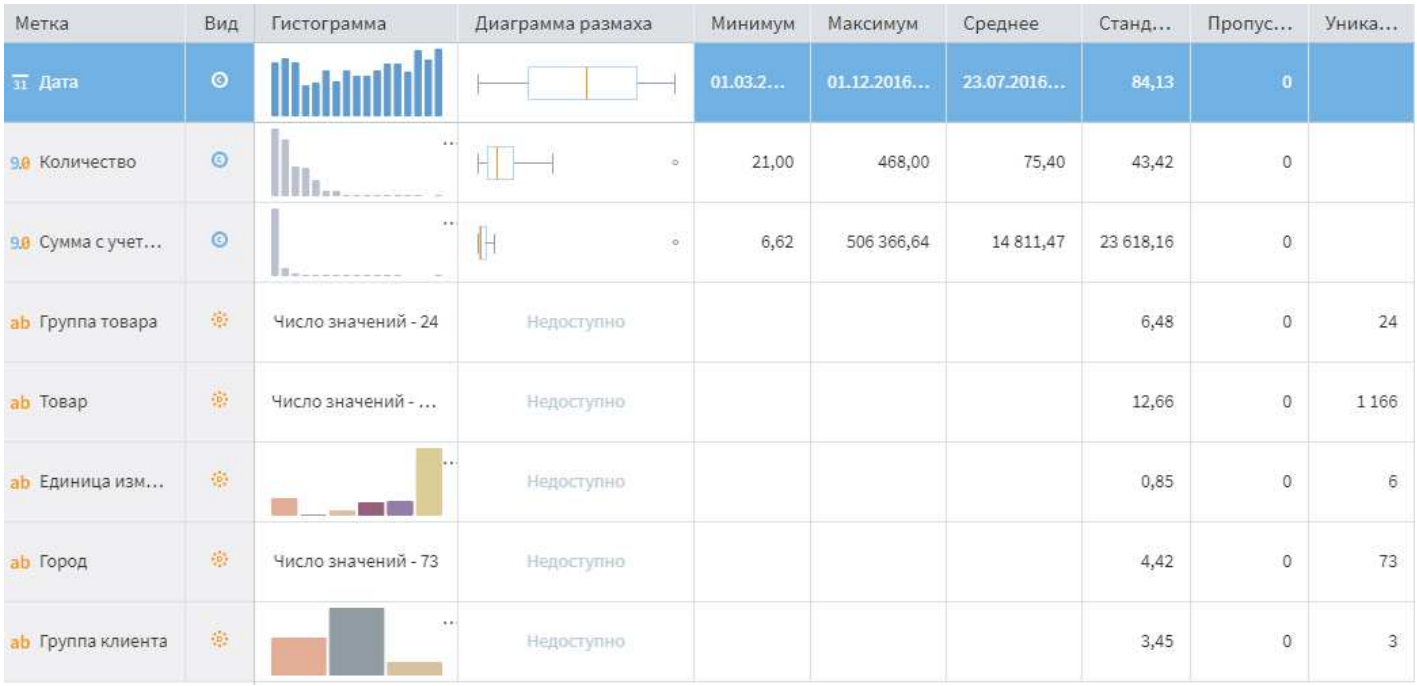


Рис. 22. Настройка визуализаторов

6. Изучите визуализатор Статистика.

С его помощью можно просмотреть различные статистические показатели по каждому полю набора данных.



Метка	Вид	Гистограмма	Диаграмма размаха	Минимум	Максимум	Среднее	Станд…	Пропус…	Уника…
Дата				01.03.2…	01.12.2016…	23.07.2016…	84,13	0	
Количество				21,00	468,00	75,40	43,42	0	
Сумма с учет…				6,62	506 366,64	14 811,47	23 618,16	0	
Группа товара		Число значений - 24	Недоступно				6,48	0	24
Товар		Число значений - …	Недоступно				12,66	0	1 166
Единица изм…			Недоступно				0,85	0	6
Город		Число значений - 73	Недоступно				4,42	0	73
Группа клиента			Недоступно				3,45	0	3

Рис. 23. Визуализатор *Статистика*

В верхней части окна визуализатора отображается общее количество записей в наборе данных. В окне статистики по каждому полю выборки отображается следующая информация: гистограмма, диаграмма размаха, минимальное, максимальное и среднее значения, стандартное отклонение, количество пропусков, количество уникальных значений.

Замечание: Диаграмма размаха доступна только для полей вещественного типа и типа Дата/время, уникальные значения рассматриваются только для полей с дискретным видом данных, гистограмма не отображается если в поле большое количество уникальных значений, как в поле *Группа товара* (где их 24) и *Товар* (где их 1166). Можно нажав на вкладку Гистограмма отобразить полный список интервалов/уникальных значений с количеством и процентом значений, относящихся к каждому интервалу.

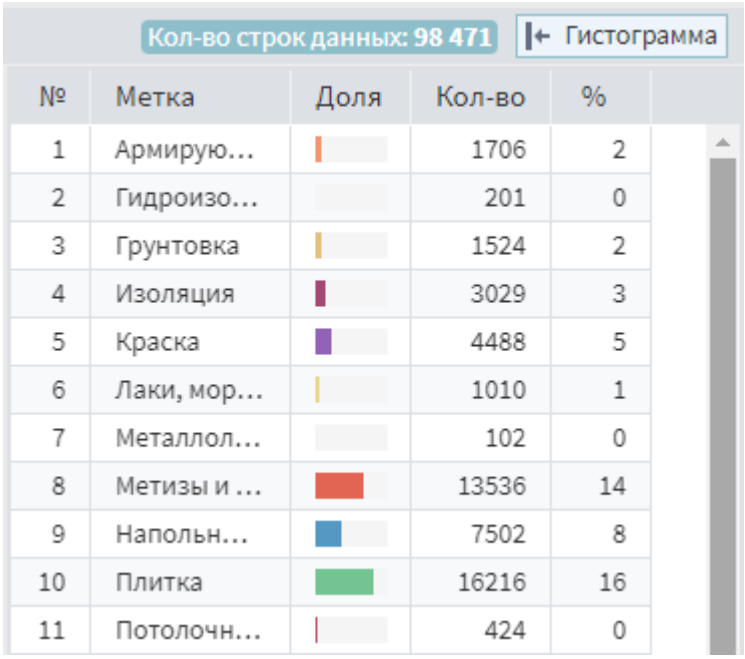


Рис. 24. Вкладка *Гистограмма*

7. Для того, чтобы изменить отображаемые статистики нужно нажать на кнопку Настройка показателей на панели инструментов визуализатора.

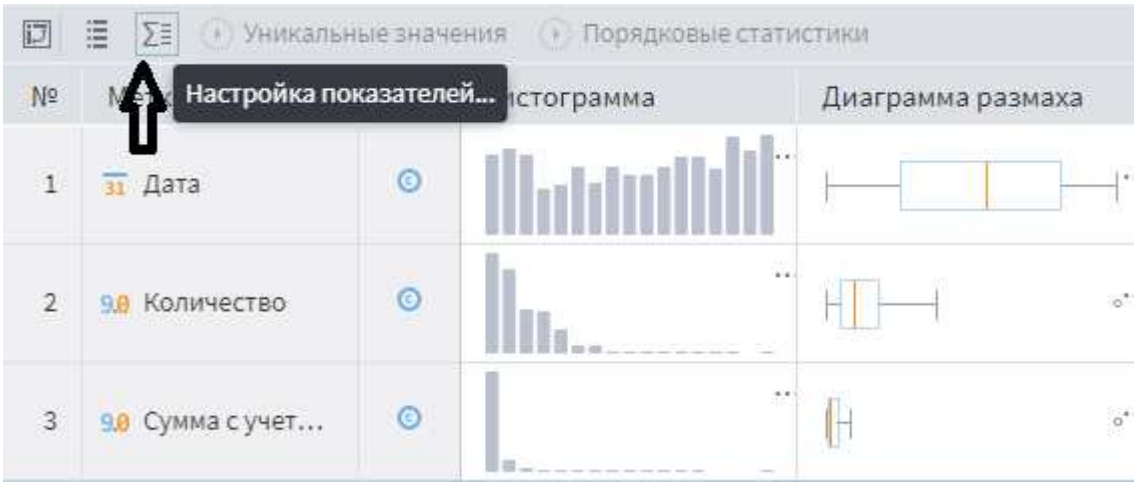


Рис. 25. Настройка показателей на панели инструментов визуализатора

Получим полный список доступных статистических показателей (рис. 26). большинство из них рассчитываются для полей вещественного, числового типов, а также типа дата/время. Но есть показатели статистики для строковых полей (минимальная, максимальная и средняя длина строки).

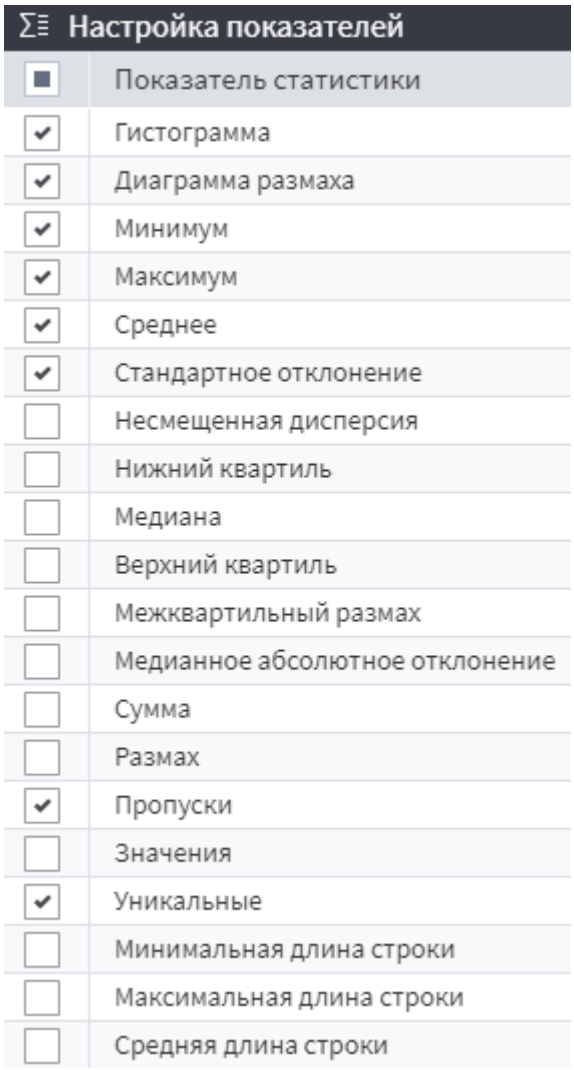


Рис. 26. Настройка показателей

Кнопка настройка полей (рис. 27) позволяет отключить отображение полей, которые нас не интересуют.

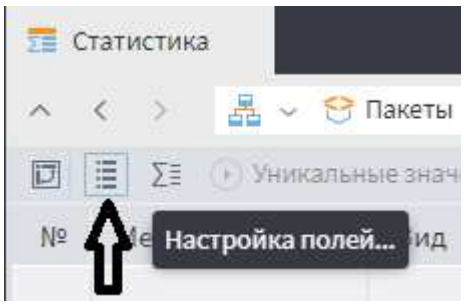


Рис. 27. Настройка показателей

Кроме того есть возможность поменять местами поля и показатели. Для этого нужно нажать на кнопку *Транспонировать* (рис. 28).

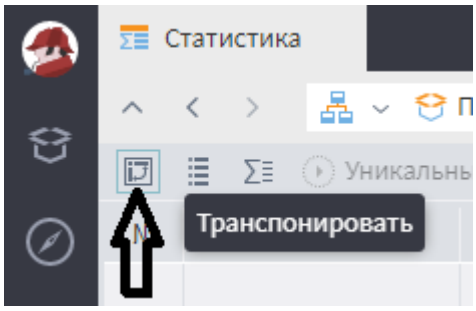


Рис. 28. Кнопка *Транспонировать* поля и показатели

8. Следующим шагом в алгоритме необходимо подсчитать общие суммы продаж по всем группам товаров. Для этого нужно переместить компонент *Группировка* в рабочую область сценария. Последовательность обработки данных задается соединением выходного порта узла импорта с входным портом группировки (рис. 29).

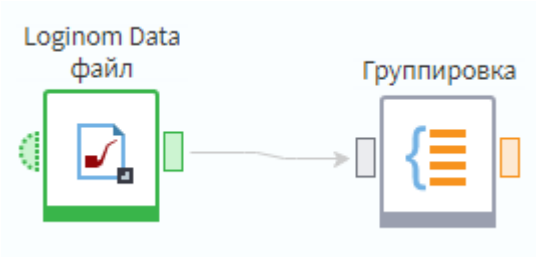


Рис. 29. Формирование связи

В Мастере настройки узла *Группировка* поле *Группа товара* задается как группа, а *Сумма* как показатель. После настройки и выполнения узла группировки в выходном порту содержатся данные об итоговых суммах покупок клиентов.

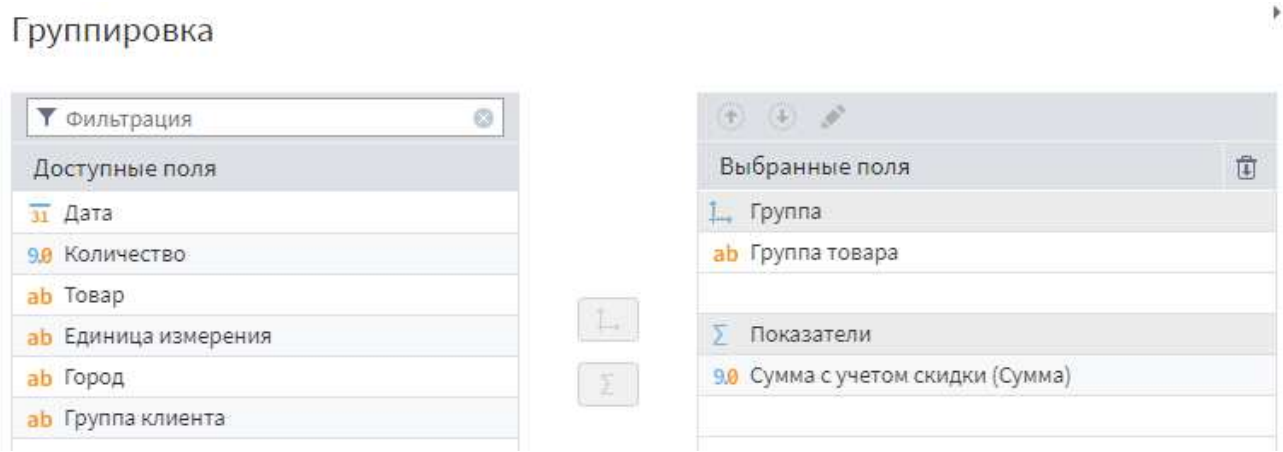


Рис. 30. настройка узла *Группировка*

9. Далее эти данные сортируются по убыванию суммы при помощи компонента *Сортировка* и затем передаются на узел выделения первых 10 строк таблицы. Для этого используется компонент *Фильтр строк*, в мастере которого задается условие: *"№ Номер строки <= 10*".
10. Добавьте узел экспорта и/или настройте *Визуализатор* результатов.

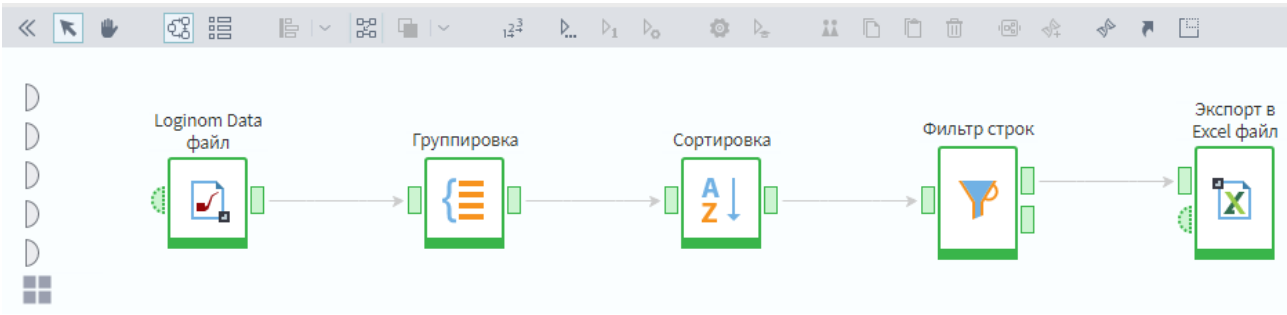


Рис. 31. Результирующий сценарий

11. Блок узлов, выполняющих формирование ТОП 10 групп товаров, возможно, сгруппировать в отдельную функцию, поместив их в *Подмодель*. Для этого необходимо выделить эти узлы и при помощи кнопки  создать подмодель.

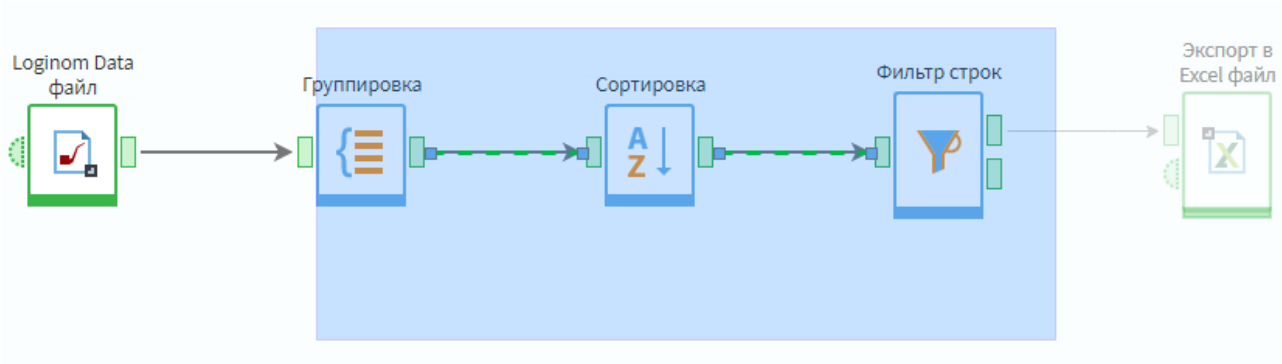


Рис. 32. Выделение блока узлов

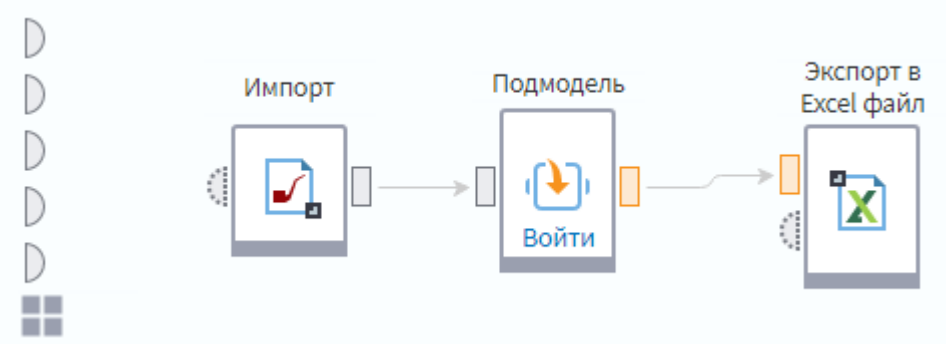


Рис. 33. Преобразование блока узлов в Подмодель

12. В дальнейшем подмодель, выполняющая заданную пользователем функцию, может быть опубликована как производный компонент и наравне со стандартными компонентами многократно использоваться в других сценариях.
13. Перед закрытием пакета его необходимо *сохранить*. Это можно сделать в меню Пакеты (см. рис. 34).

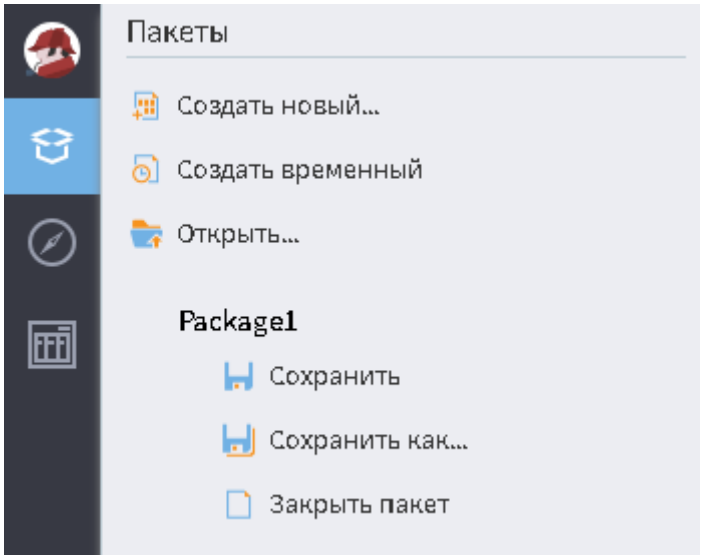


Рис. 34. Сохранение пакета

1.4 Вопросы для самопроверки

1. Что такое пакет в АП Loginom?
2. Как создать новый пакет и как сохранить текущий пакет под другим именем?
3. Сколько пакетов можно одновременно открыть в АП Loginom?
4. Чем отличаются абсолютный и относительный пути открытия файлов?
5. Принципы проектирования в АП Loginom?
6. Что включает модуль в АП Loginom ?
7. Что такое сценарий и узел сценария?
8. В какие группы объединены стандартные компоненты?
9. Для чего предназначены компоненты группы *Трансформация*?

10. Для чего предназначены компоненты группы *Управление*?
11. Для чего предназначены компоненты группы *Исследование*?
12. Для чего предназначены компоненты группы *Предобработка*?
13. Для чего предназначены компоненты группы *Data Mining*?
14. Для чего предназначены компоненты группы *Переменные*?
15. Вы импортировали текстовый файл, создав узел импорта. После чего обнаружили, что неправильно задали параметры импорта. Как легче всего исправить ошибку?
16. Что позволяет сделать компонент *Параметры полей*?
17. Можно ли настроить соответствия столбцов, которые имеют различный тип?
18. Какие характеристики набора данных показывает визуализатор Статистика?
19. Как обнаружить имеются ли в поле набора данных пропущенные значения?
20. К существующему в сценарии узлу импорта необходимо еще добавить один визуализатор. Что предпринять?
21. Для чего предназначен компонент *Группировка*?
22. Для чего предназначен компонент *Фильтр строк*?
23. Какие условия фильтрации существуют в АП Loginom?
24. Сколько записей будет отфильтровано в результате применения фильтра «([Размер ссуды, руб.] в интервале [2000..5000]) И ([Цель ссуды] = 'Покупка товара') И ([Цель ссуды] = 'Иное')»?

2. Компоненты обработки данных в АП Loginom

2.1 Предобработка данных

2.1.1 Заполнение пропусков

Данный обработчик предназначен для автоматического заполнения пропущенных значений в наборах данных. Для каждого столбца исходного набора данных пользователь может выбрать наиболее подходящий метод заполнения пропусков. Пропусками считаются Null-значения.

Доступны следующие методы обработки:

- оставить без изменения – выявленные пропуски заполняться не будут;
- удалять записи – строки с выявленными пропусками исключаются из набора данных;
- заменять случайными значениями – выявленные пропуски заменяются случайным значением столбца;
- заменять средним – выявленные пропуски заменяются средним значением столбца;
- заменять медианой – выявленные пропуски заменяются медианой, вычисленной по столбцу;
- заменять наиболее вероятным – выявленные пропуски заменяются наиболее вероятным значением по столбцу, замена производится на среднее значение из наиболее вероятного интервала, число интервалов варьируется в зависимости от объема выборки – чем она больше, тем больше интервалов;
- заменять значением "не задано" – выявленные пропуски заменяются значением "не задано";
- интерполировать – выявленные пропуски заменяются расчетным значением по столбцу.

2.1.2 Редактирование выбросов

Обработчик предназначен для автоматической корректировки аномальных значений (выбросов и экстремальных значений) в наборах данных.

Аномальными называются значения, которые не укладываются в общую модель поведения анализируемого процесса. Они сильно отличаются от окружающих данных и могут быть вызваны как ошибками измерений, так и некорректным вводом данных, или являться результатом их сильной изменчивости.

Аномальные значения являются следствием:

- ошибок в данных (погрешности измерений, неверная запись или считывание данных и т.п.);
- воздействие случайных, не поддающихся прогнозированию факторов (например, разовый наплыв клиентов из-за массового мероприятия);
- присутствия объектов «других» выборок (например, показаниями сломавшегося датчика).

Степень устойчивости алгоритма к наличию в данных аномальных значений называется *робастностью.*

В обработчике Редактирование выбросов для каждого поля исходного набора данных критерии определения выбросов и экстремальных значений задаются пользователем с помощью указания допустимого стандартного отклонения или интерквантильной ширины. Под ***выбросами*** при этом подразумеваются значения данных, существенно отклоняющиеся от средних, а под ***экстремальными*** – значения, которые настолько сильно отклоняются от типичных значений, что перестают соответствовать логике исследуемых процессов и явлений.

Для определения **выбросов и экстремальных значений** доступны два метода выявления:

- Стандартное отклонение – критерием является отклонение значения признака от среднего более, чем на заданное число стандартных отклонений. При этом данный параметр отдельно задается для выбросов и для экстремальных

значений. Данный метод следует применять, если известно, что распределение данных близко к нормальному.

- Интерквартильная ширина – критерием является расстояние между 1-м и 4-м квартилями распределения значений признака. Если значение признака отклоняется от медианы более, чем на заданное число интерквартильных ширин, то оно считается аномальным. Данный параметр задается отдельно для выбросов и экстремальных значений. Этот метод можно применять и в случае, когда распределение данных отличается от нормального.

Как для выбросов, так и для экстремальных значений доступны следующие методы редактирования:

- оставить без изменения;
- удалять записи – исключить строки с аномальными значениями из набора данных;
- заменять средним – заменять аномалии средним значением столбца;
- заменять медианой – заменять аномалии медианой, вычисленной по столбцу;
- заменять наиболее вероятным – замена аномалий наиболее вероятным значением по столбцу, замена производится на среднее значение из наиболее вероятного интервала, число интервалов варьируется в зависимости от объема выборки – чем она больше, тем больше интервалов;
- заменять заданным значением – замена аномалий на значение, прописанное вручную;
- ограничивать – аномалии будут заменены значением границы, с которой начинается определение аномалии.

2.1.3 Компонент «Параметры полей»

Узел позволяет изменить следующие параметры полей набора данных:

- Имя;
- Метку;
- Тип данных;
- Вид данных;
- Назначение.

Важно: Узел не накладывает ограничений при изменении типа данных поля. По возможности, при преобразовании типа сохраняется полнота информации, но в некоторых случаях изменение типа может привести к потере информации.

Так же узел позволяет кэшировать набор данных целиком, либо отдельные его поля.

- **Входной источник данных** – порт для подключения входного набора данных.
- **Выходной набор данных** – порт с измененными параметрами полей.

Мастер настройки

- **Кэшировать** – параметры кэширования выходного набора данных. Предоставляется выбор из следующих вариантов.
- **Отключено** – кэширование не будет производится. Используется по умолчанию.
- **При активации** – при активации узла будет производится кэширование всего набора данных.
- **При обращении** – будет производится кэширование тех полей выходного набора, данные которых запрошены последующими узлами сценария или визуализатором.
- **Выбранные поля** – параметры кэширования устанавливаются для каждого поля в отдельности.

- **Список параметров полей** – в табличном виде представлены параметры полей набора данных. Двойным кликом по выбранному полю вызывается диалог редактирования его параметров. В диалоге помимо редактирования стандартных параметров поля задается параметр кэширования. Данная опция доступна в случае установки для всего узла режима кэширования *Выбранные поля*. Предоставляется выбор из следующих вариантов.
- **Отключено** – кэширование не будет производится. Используется по умолчанию.
- **При активации** – при активации узла будет производится кэширование поля набора данных.
- **При обращении** – будет производится кэширование поля при первом запросе его данных последующими узлами сценария или визуализатором.

2.1.4 Лабораторная работа «Очистка и предобработка данных в АП Loginom»

Для расчетов стоимости и анализа рынка недвижимости, специалистам в этой области необходима информация об объектах недвижимости, которую можно найти на различных сайтах с предложениями, такими как avito, giper.nn, yandex и т.д.

Благодаря современным технологиям процесс сбора данных от ручного перешел к автоматизированному. Но даже автоматизация процесса сбора данных не поможет защититься от некачественно предоставленной информации об объектах. В большинстве своем многие параметры не указаны, либо указаны не точно, а иногда совсем неверно. Чтобы сделать полученную информацию пригодной для дальнейшего использования, ее необходимо обработать. Этот процесс называется предобработка данных. Он может представлять собой как оптимизацию, так и очистку. Под очисткой понимается процесс, направленный на исключение факторов, которые снижают качество данных и мешают корректной

работе аналитических алгоритмов, а оптимизация необходима для выявления и устранения незначащих факторов.

Начнем работа в аналитической программе Loginom с импорта данных.

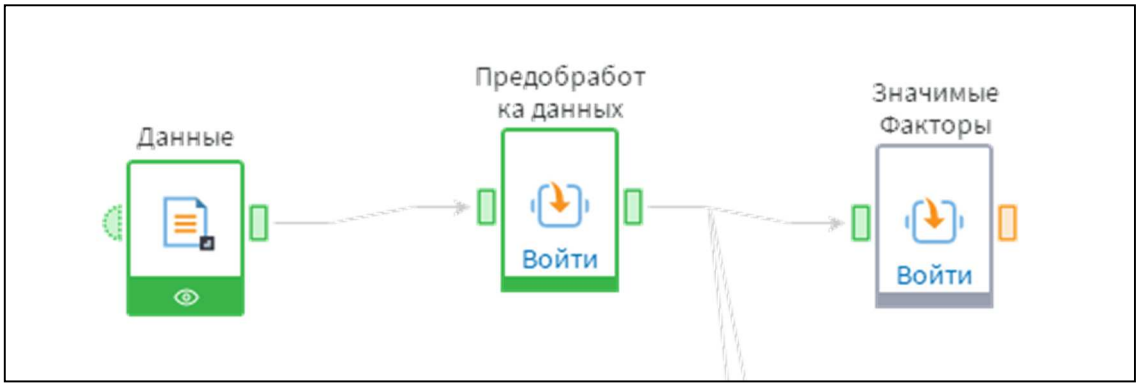
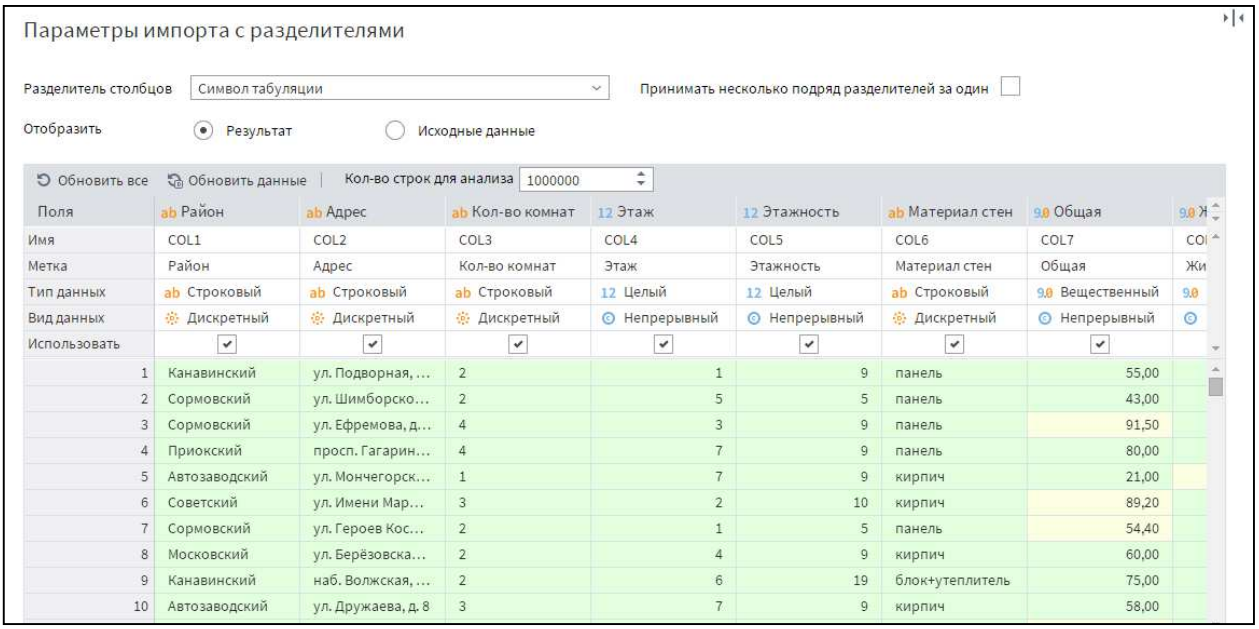


Рис. 35 Фрагмент сценарий загрузки и предобработки данных

Чтобы компоненты работали корректно, данные должны отвечать следующей обязательной структуре:

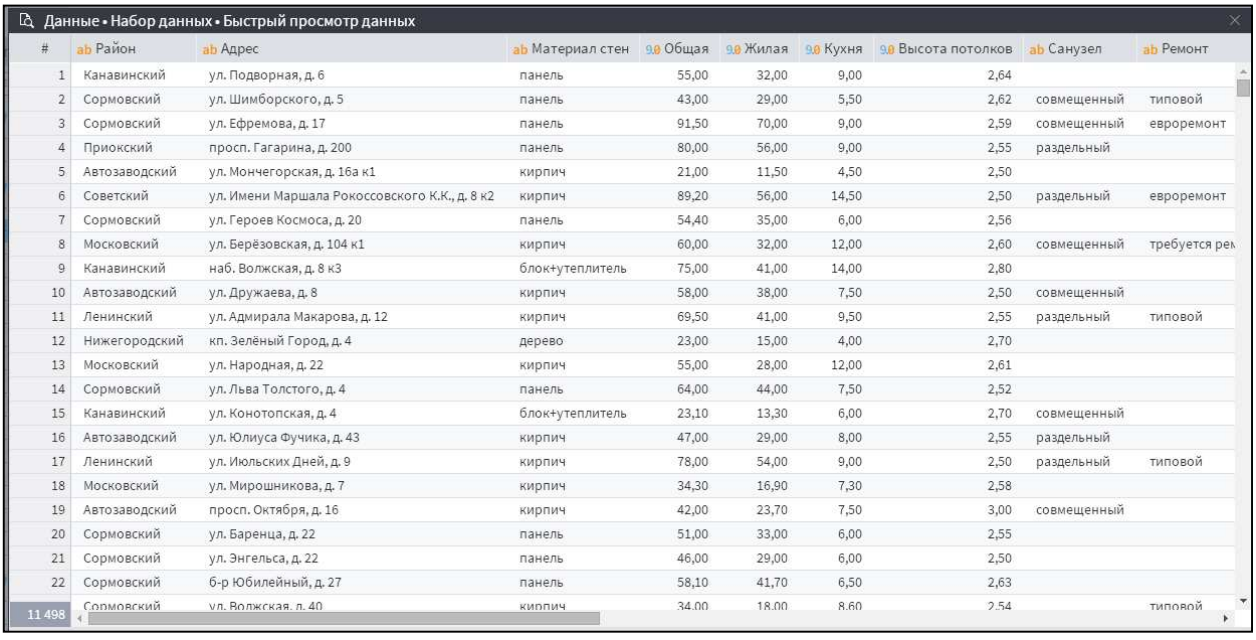
- Район (текст);
- Количество комнат (текст);
- Этаж (целое число);
- Этажность дома (целое число);
- Материал стен (текст);
- Общая площадь (вещественное число);
- Площадь жилая (вещественное число);
- Площадь кухни (вещественное число);
- Год постройки (целое число);
- Высота потолков (вещественное число);
- Балкон/Лоджия (текст);
- Статус объявления (текст);
- Дата размещения (дата/время);
- Цена предложения (вещественное число).

Необходимо убедиться правильно ли считала программа типы данных, иначе стоит установить их вручную (рис. Настройка импорта данных).



Поля	Район	Адрес	Кол-во комнат	Этаж	Этажность	Материал стен	Общая
Имя	COL1	COL2	COL3	COL4	COL5	COL6	COL7
Метка	Район	Адрес	Кол-во комнат	Этаж	Этажность	Материал стен	Общая
Тип данных	Строковый	Строковый	Строковый	Целый	Целый	Строковый	Вещественный
Вид данных	Дискретный	Дискретный	Дискретный	Непрерывный	Непрерывный	Дискретный	Непрерывный
Использовать	✓	✓	✓	✓	✓	✓	✓
1	Канавинский	ул. Подворная, …	2	1	9	панель	55,00
2	Сормовский	ул. Шимборско…	2	5	5	панель	43,00
3	Сормовский	ул. Ефремова, д…	4	3	9	панель	91,50
4	Приокский	просп. Гагарин…	4	7	9	панель	80,00
5	Автозаводский	ул. Мончегорск…	1	7	9	кирпич	21,00
6	Советский	ул. Имени Мар…	3	2	10	кирпич	89,20
7	Сормовский	ул. Героев Кос…	2	1	5	панель	54,40
8	Московский	ул. Берёзовска…	2	4	9	кирпич	60,00
9	Канавинский	наб. Волжская, …	2	6	19	блок+утеплитель	75,00
10	Автозаводский	ул. Дружаева, д. 8	3	7	9	кирпич	58,00

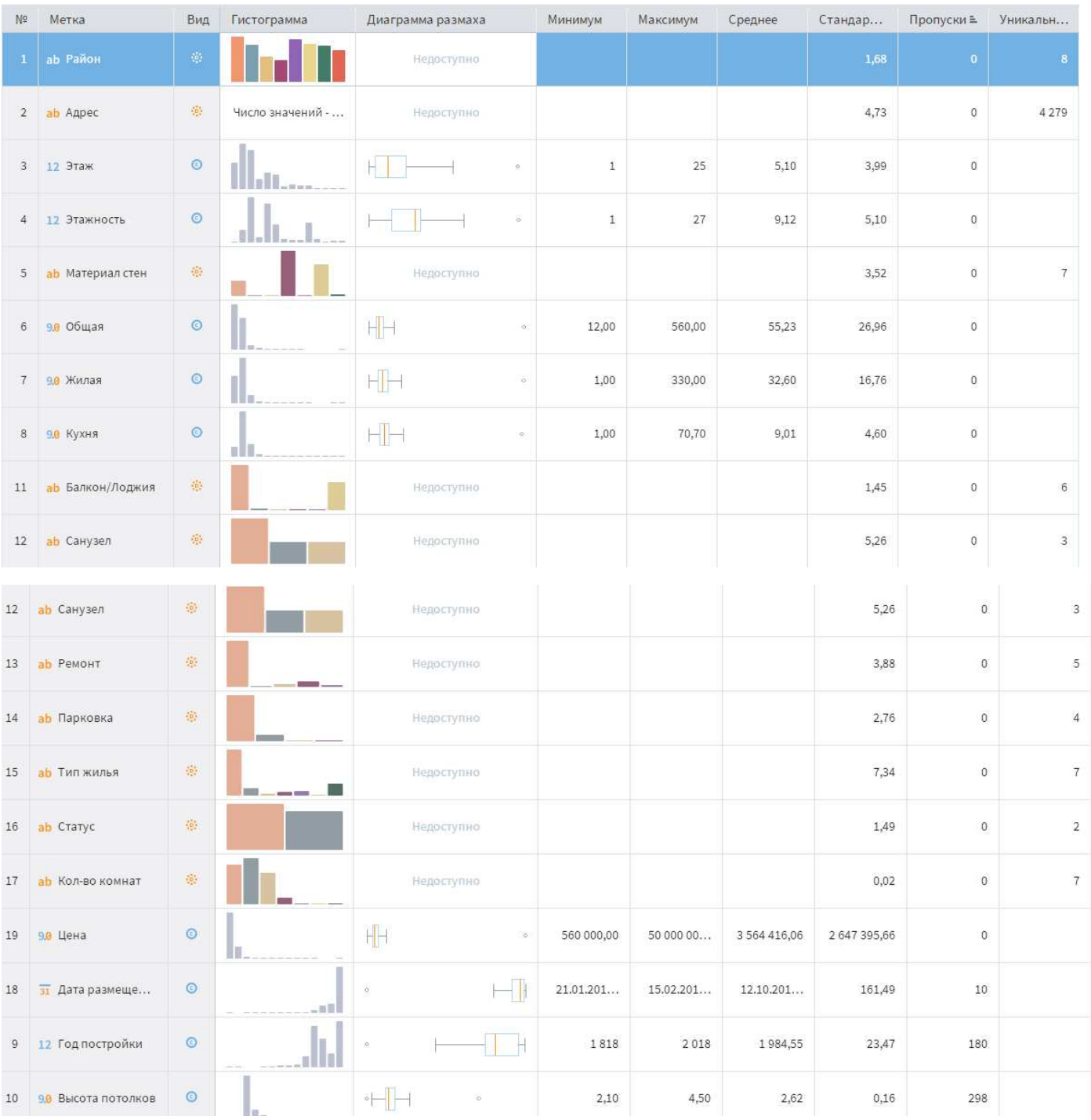
Рис. 36. Настройка импорта данных



#	Район	Адрес	Материал стен	Общая	Жилая	Кухня	Высота потолков	Санузел	Ремонт
1	Канавинский	ул. Подворная, д. 6	панель	55,00	32,00	9,00	2,64		
2	Сормовский	ул. Шимборского, д. 5	панель	43,00	29,00	5,50	2,62	совмещенный	типовой
3	Сормовский	ул. Ефремова, д. 17	панель	91,50	70,00	9,00	2,59	совмещенный	евроремонт
4	Приокский	просп. Гагарина, д. 200	панель	80,00	56,00	9,00	2,55	раздельный	
5	Автозаводский	ул. Мончегорская, д. 16а к1	кирпич	21,00	11,50	4,50	2,50		
6	Советский	ул. Имени Маршала Рокоссовского К.К., д. 8 к2	кирпич	89,20	56,00	14,50	2,50	раздельный	евроремонт
7	Сормовский	ул. Героев Космоса, д. 20	панель	54,40	35,00	6,00	2,56		
8	Московский	ул. Берёзовская, д. 104 к1	кирпич	60,00	32,00	12,00	2,60	совмещенный	требуется ре
9	Канавинский	наб. Волжская, д. 8 к3	блок+утеплитель	75,00	41,00	14,00	2,80		
10	Автозаводский	ул. Дружаева, д. 8	кирпич	58,00	38,00	7,50	2,50	совмещенный	
11	Ленинский	ул. Адмирала Макарова, д. 12	кирпич	69,50	41,00	9,50	2,55	раздельный	типовой
12	Нижегородский	кп. Зелёный Город, д. 4	дерево	23,00	15,00	4,00	2,70		
13	Московский	ул. Народная, д. 22	кирпич	55,00	28,00	12,00	2,61		
14	Сормовский	ул. Льва Толстого, д. 4	панель	64,00	44,00	7,50	2,52		
15	Канавинский	ул. Конотопская, д. 4	блок+утеплитель	23,10	13,30	6,00	2,70	совмещенный	
16	Автозаводский	ул. Юлиуса Фучика, д. 43	кирпич	47,00	29,00	8,00	2,55	раздельный	
17	Ленинский	ул. Июльских Дней, д. 9	кирпич	78,00	54,00	9,00	2,50	раздельный	типовой
18	Московский	ул. Мирошникова, д. 7	кирпич	34,30	16,90	7,30	2,58		
19	Автозаводский	просп. Октября, д. 16	кирпич	42,00	23,70	7,50	3,00	совмещенный	
20	Сормовский	ул. Баренца, д. 22	панель	51,00	33,00	6,00	2,55		
21	Сормовский	ул. Энгельса, д. 22	панель	46,00	29,00	6,00	2,50		
22	Сормовский	б-р Юбилейный, д. 27	панель	58,10	41,70	6,50	2,63		
11 498	Сормовский	ул. Волжская, д. 40	кирпич	34,00	18,00	8,60	2,54		типовой

Рис. 37. Просмотр загруженных данные

После настройки импорта, следует изучить основные статистические показатели данных. Для этого необходимо использовать инструмент визуализации данных *Статистика*, который помогает узнать, имеют ли данные пропуски, выбросы и экстремальные значения.



№	Метка	Вид	Гистограмма	Диаграмма размаха	Минимум	Максимум	Среднее	Стандар…	Пропуски	Уникальн…
1	ab Район			Недоступно				1,68	0	8
2	ab Адрес		Число значений - …	Недоступно				4,73	0	4 279
3	12 Этаж				1	25	5,10	3,99	0	
4	12 Этажность				1	27	9,12	5,10	0	
5	ab Материал стен			Недоступно				3,52	0	7
6	9,0 Общая				12,00	560,00	55,23	26,96	0	
7	9,0 Жилая				1,00	330,00	32,60	16,76	0	
8	9,0 Кухня				1,00	70,70	9,01	4,60	0	
11	ab Балкон/Лоджия			Недоступно				1,45	0	6
12	ab Санузел			Недоступно				5,26	0	3
12	ab Санузел			Недоступно				5,26	0	3
13	ab Ремонт			Недоступно				3,88	0	5
14	ab Парковка			Недоступно				2,76	0	4
15	ab Тип жилья			Недоступно				7,34	0	7
16	ab Статус			Недоступно				1,49	0	2
17	ab Кол-во комнат			Недоступно				0,02	0	7
19	9,0 Цена				560 000,00	50 000 00…	3 564 416,06	2 647 395,66	0	
18	31 Дата размеще…				21.01.201…	15.02.201…	12.10.201…	161,49	10	
9	12 Год постройки				1 818	2 018	1 984,55	23,47	180	
10	9,0 Высота потолков				2,10	4,50	2,62	0,16	298	

Рис. 38 Статистика данных

Видно, что есть пропуски в некоторых столбцах. В поле уникальные значения по параметру «адрес» видно, сколько уникальных объектов присутствует во всем множестве данных. Можно заметить, что их число, не равно общему количеству строк данных, следовательно, в выборке присутствуют дубликаты, которые необходимо устранить.

Описание подмодели «Предобработка данных»:

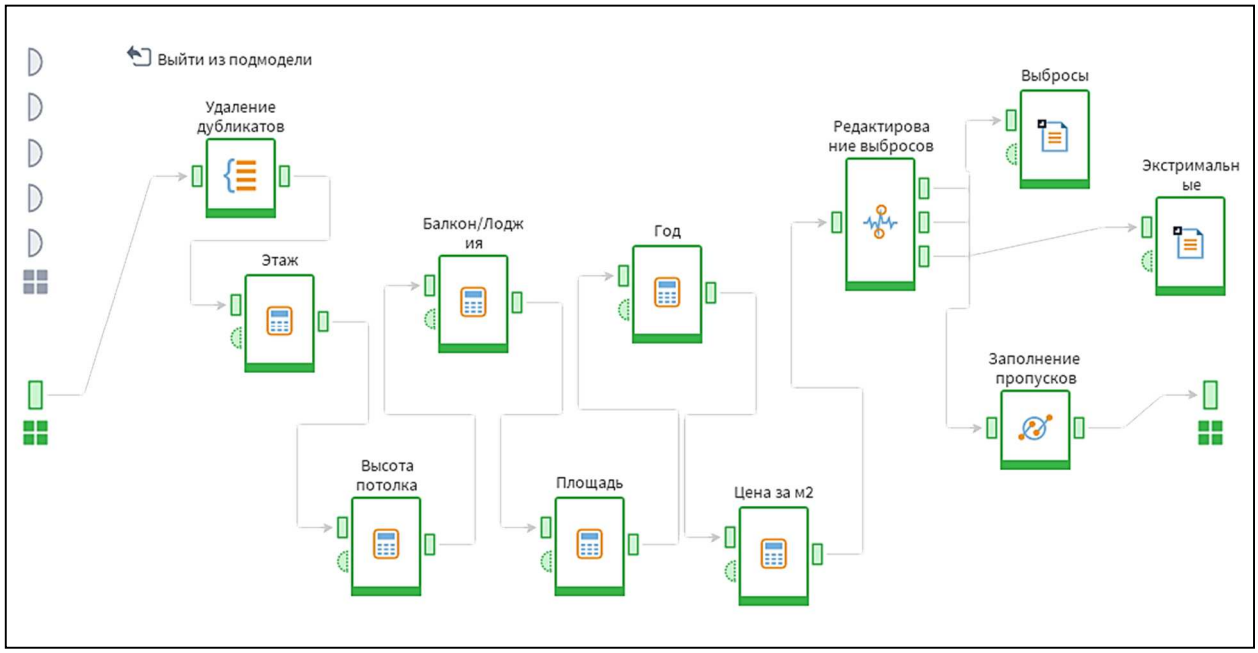


Рис. 39. Подмодель «Предобработка данных»

Для устранения дубликатов используется инструмент «Группировка», где в Группы попадут данные по которым будет одинаковая информация, а в Показателях все остальные параметры с агрегацией «первый».

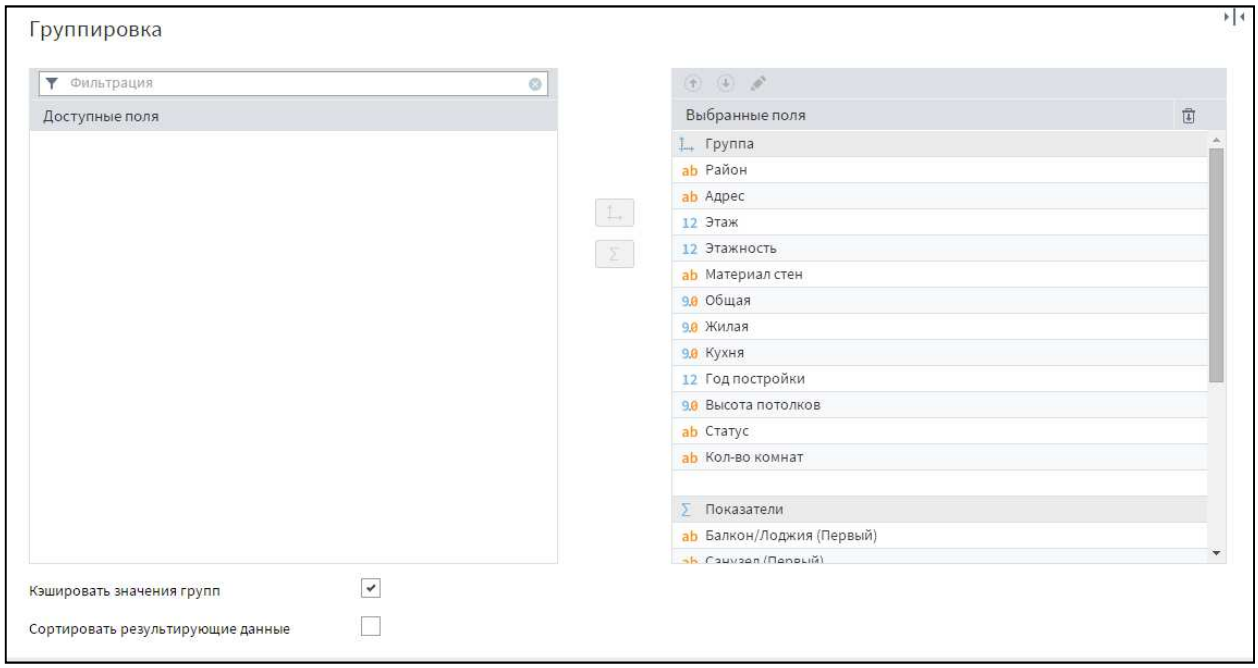


Рис. 40. Удаление дубликатов

Поле Этаж разделим на три группы: либо это первый этаж, то есть в поле этаж стоит 1, либо это последний, если этаж равен этажности дома, либо средний – все остальные.

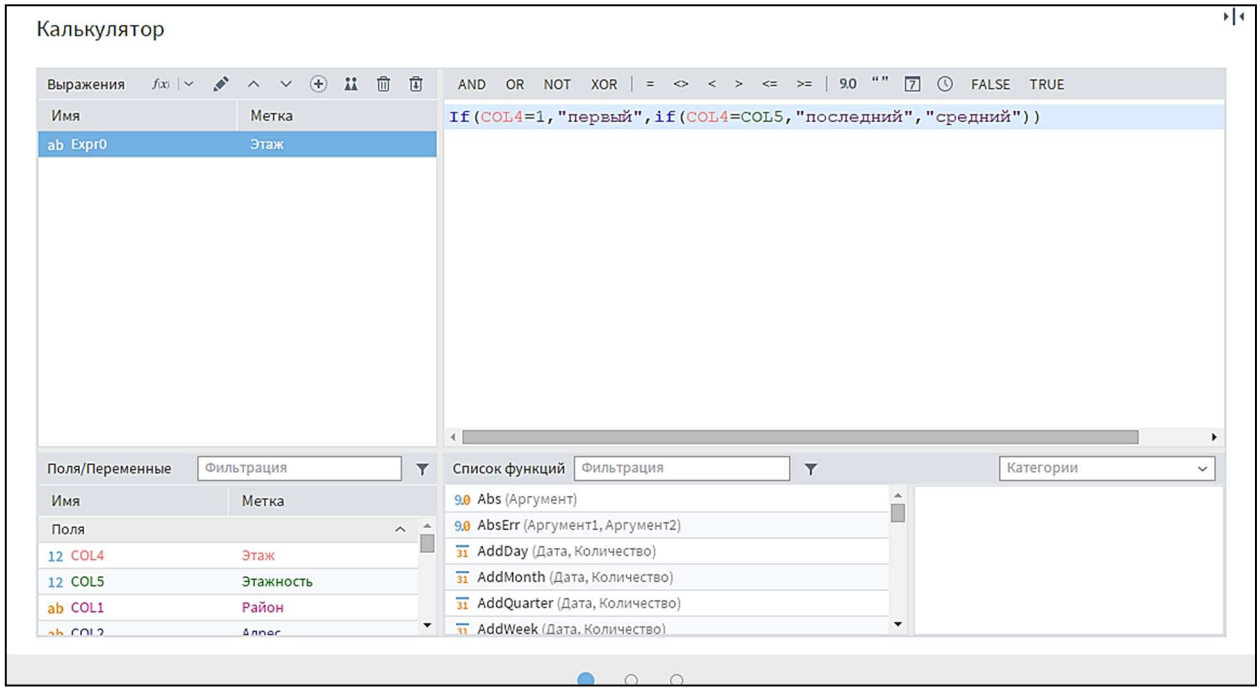


Рис. 41. Компонент калькулятор

По полю Высота потолков также разделим все объекты недвижимости по трем группам: менее 2.5 м, 2.5 м и более 2.5 м.

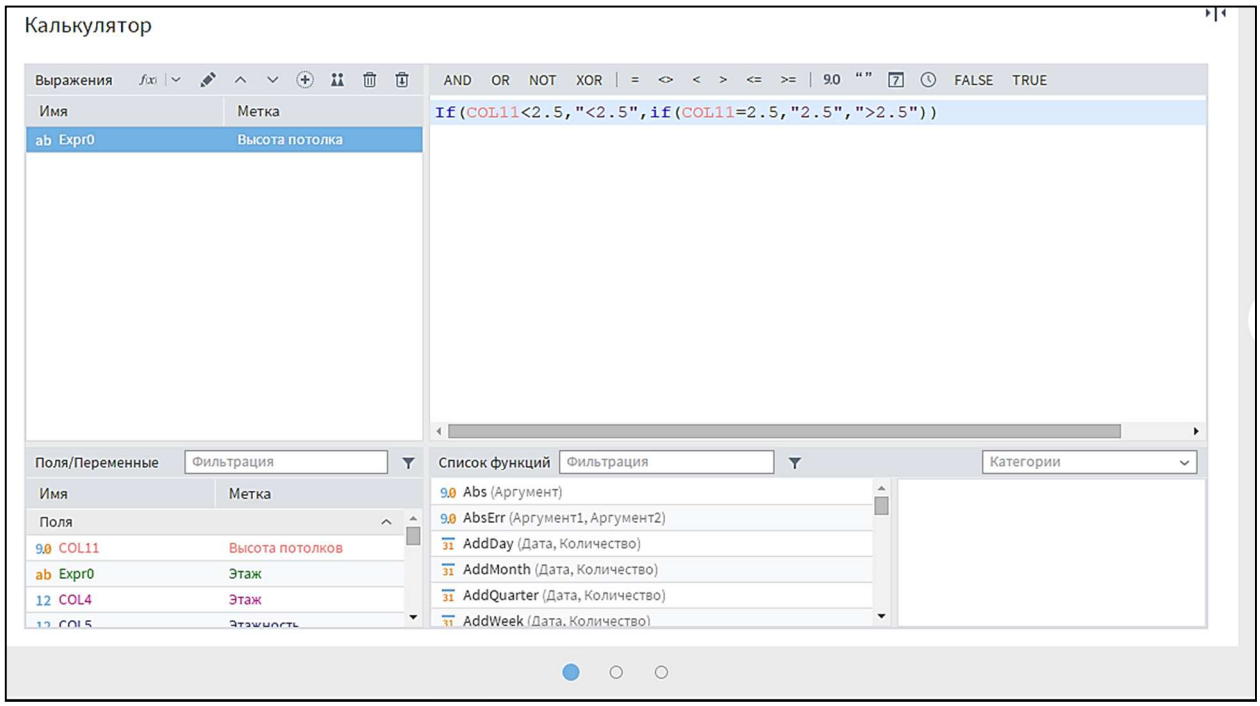


Рис. 42 . Компонент калькулятор «Высота потолка»

Заменим пустые значения в параметре Балкон на – *нет балкона*.

Везде где указано количество балконов/лоджий или «есть», поставим 1.

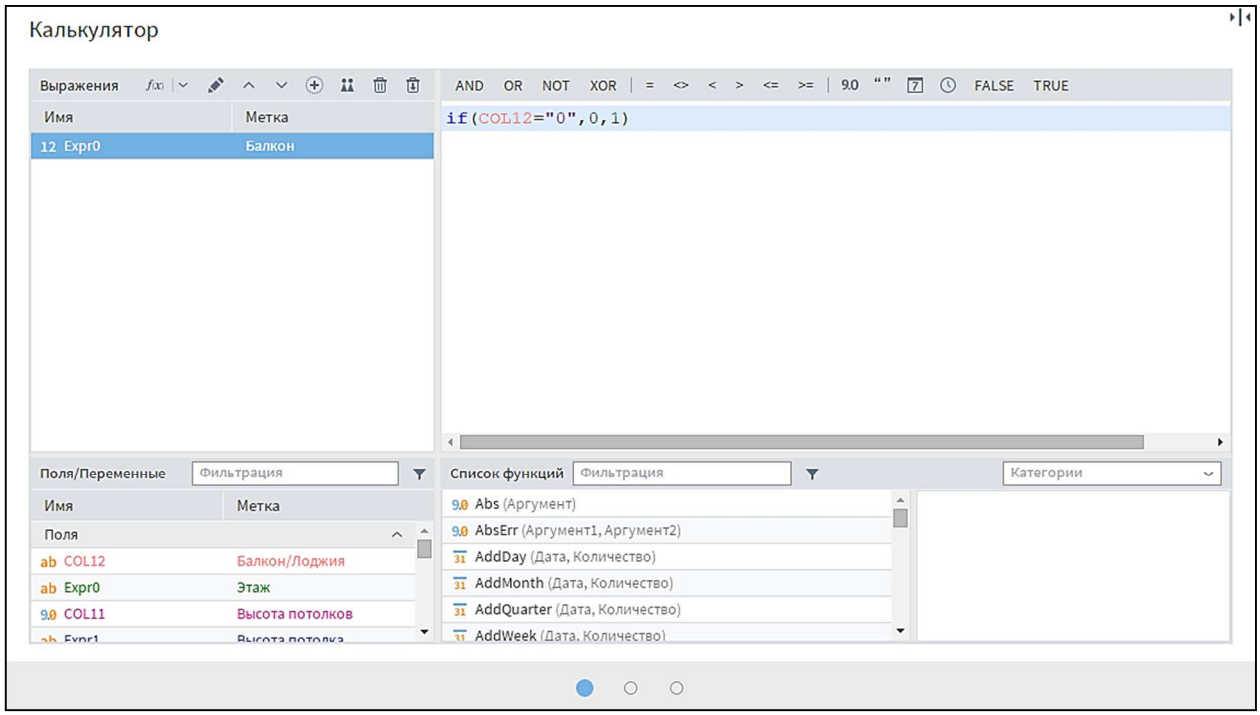


Рис. 43 . Компонент калькулятор «Балкон есть (1), нет (0)»

С помощью инструмента «Калькулятор» можно вычислить стоимость метра квадратного, поделив стоимость объекта на его площадь.

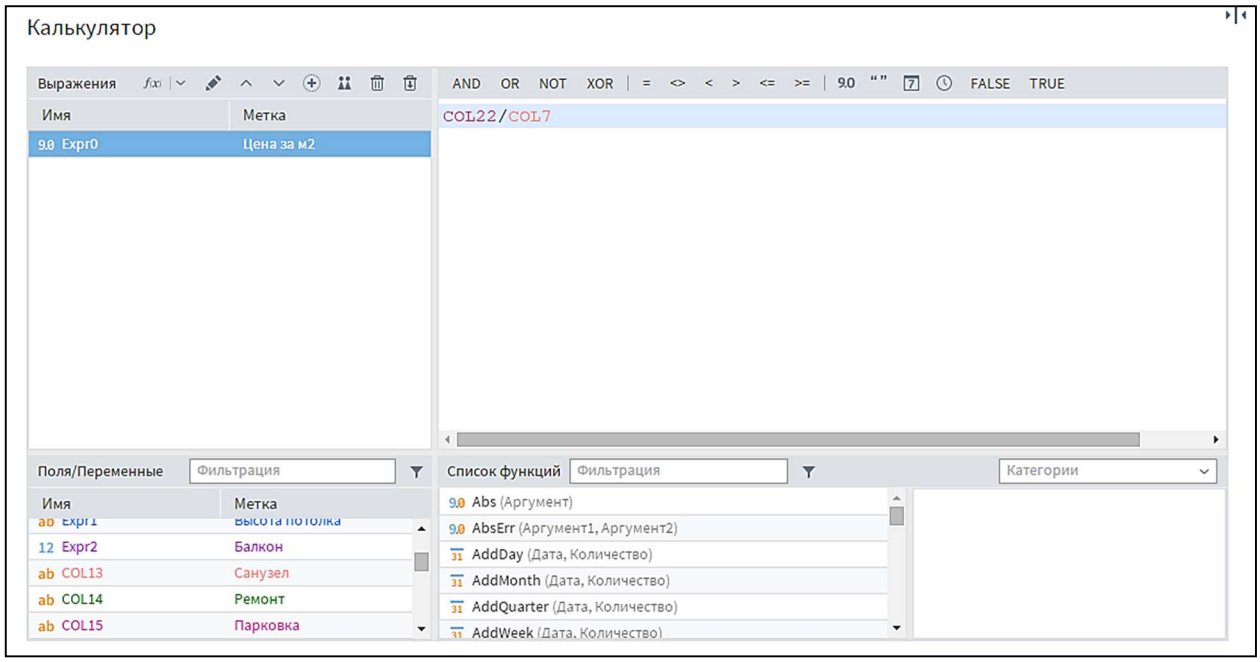


Рис. 44. Узел Калькулятор - Вычисление Цена за кв.м.

Для устранения аномальных значений следует использовать компонент *Редактирование выбросов*. На выходах у него очищенные данные, данные по выбросам и экстремальным значениям. Очищенные данные будут использоваться для дальнейших расчетов, а остальные следует экспортировать для дальнейшего анализа.

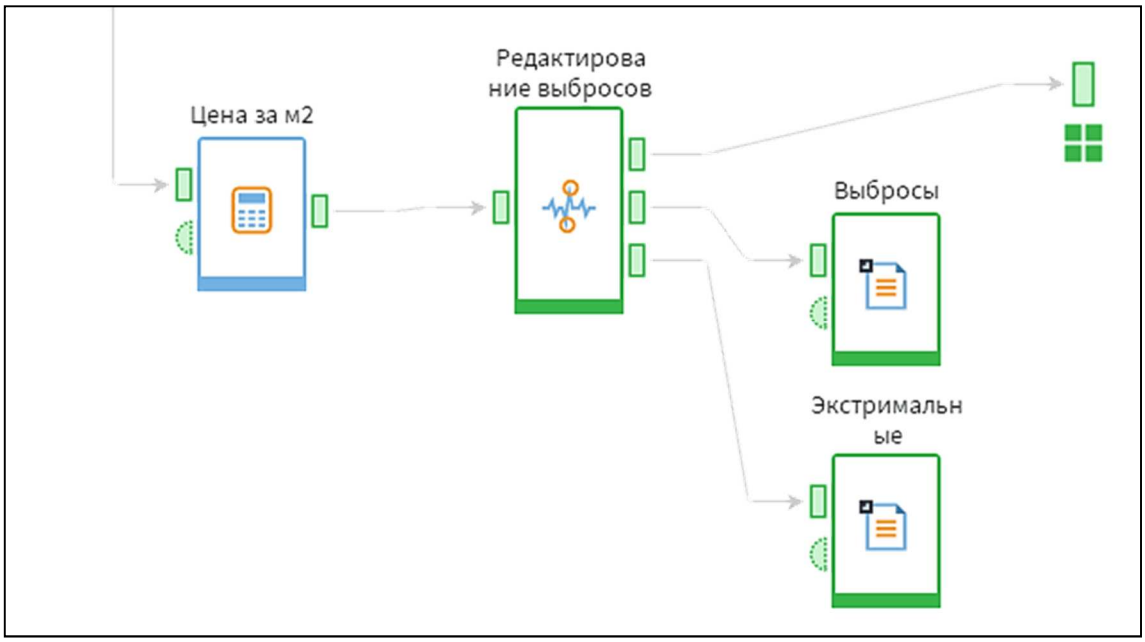


Рис. 45 Компонент «Редактирование выбросов»

Очищенные и обогащенные новой информацией данные можно экспортировать в новый файл.

2.2. Трансформация данных

2.2.1 Сортировка

Обработчик выполняет сортировку записей входного набора. Алгоритм позволяет сортировать последовательно по нескольким полям.

При сортировке учитывается:

- **Очередность полей сортировки** – в зависимости от позиции в списке *Поля сортировки* меняется очередность сортировки поля;
- **Порядок сортировки** – для каждого сортируемого поля задается порядок (*По убыванию* или *По возрастанию*), в котором оно сортируется;

- **Регистр данных** – у полей со *Строковым* или *Переменным* типом данных указывается их регистро-зависимость при сортировке.

Алгоритм сортирует записи по первому в очереди полю сортировки в соответствии с заданным порядком. Если существуют одинаковые значения, то содержащие их строки сортируются по второму в очереди полю сортировки и т.д. согласно очередности полей сортировки.

Вход

- **Входной источник данных** – порт для подключения входного набора данных.

Выход

- **Выходной набор данных** – на порт выводится таблица с набором данных, отсортированным по выбранным полям.

Мастер настройки

В левом списке отображаются поля, по которым можно производить сортировку. Список соответствует полям входного порта.

Список полей можно отфильтровать, введя имя или метку поля в области фильтрации.

Для настройки сортировки, необходимо переместить требуемые поля в список *Поля сортировки* при помощи:

- перетаскивания мышью (Drag-and-drop);
- двойного щелчка мышью по полю;
- нажатия кнопки *Добавить*.

Параметры настройки сортировки:

- **Порядок** – поле может принимать значения *По возрастанию* и *По убыванию*;
- **Регистр** – для сортировки строковых полей с учетом регистра нужно поставить флаг для этого поля.

Очередность полей сортировки можно изменить:

- **Переместить вверх** – перемещает выделенное *Поле сортировки* вверх по списку; **Переместить вниз** – перемещает выделенное *Поле сортировки* вниз по списку.

Для исключения сортировки по полю:

- перетащить запись из списка *Поля сортировки* в *Доступные поля*;
- дважды щелкнуть мышью по полю;
- *Удалить поле*. Для очистки всего списка нажать *Удалить все...*.

2.2.2 Фильтр строк

Обработчик *Фильтр строк* позволяет выделить записи, которые удовлетворяют одному или нескольким условиям. Несколько условий объединяются в сложное условие с помощью логических операторов И/ИЛИ.

Пример сложного условия:

(Город = Москва) **И** (Имя = Саша) **И** (Возраст >= 30) **И** (Пол = мужской) **ИЛИ** (Город = Тула)

В качестве параметра условий могут выступать управляющие переменные. В этом случае приведенный выше пример будет выглядеть следующим образом:

(Город = <VAR1>) **И** (Имя = <VAR2>) **И** (Возраст >= <VAR3>) **И** (Пол = <VAR4>) **ИЛИ** (Город = <VAR5>),

где: VAR1 ... VAR5 – имена управляющих переменных, принятые узлом в качестве входных параметров.

Таким образом, условие фильтра может задаваться динамически в ходе выполнения сценария.

Важно: при написании сложных условий приоритет будет у оператора **И**. Например, сложное условие вида: "A **ИЛИ** B **И** C **ИЛИ** D **И** E **И** F" будет выполняться следующим образом: "A **ИЛИ** (B **И** C) **ИЛИ** (D **И** E **И** F)".

Входной набор данных делится на два выходных набора (таблицы данных): записи, удовлетворяющие условию фильтрации, и записи, не удовлетворяющие условию.

Вход

- **Входной источник данных** (таблица данных).

Выход

- **Соответствуют условию** (таблица данных);
- **Не соответствуют условию** (таблица данных).

Мастер настройки

В верхней части мастера настройки находится *Состояния входа*.

Под строкой состояния располагается область настройки условий фильтрации. Новое условие добавляется нажатием на кнопку +. Далее выбирается имя поля, отношение сравнения (*Условие*) и значение сравнения. При создании нескольких условий между ними необходимо задать логические операторы И/ИЛИ. По умолчанию ставится оператор И. Чтобы изменить оператор, нужно кликнуть по нему левой кнопкой мыши.

В обработчике имеется возможность предпросмотра результата фильтрации (выводятся первые 25 строк результирующей таблицы). Чтобы запустить его, необходимо нажать кнопку *Применить фильтр*.

Примечание: Для вывода данных в окно предпросмотра обрабатывается только первая тысяча строк исходного набора. Если среди них не найдено записей, удовлетворяющих условию фильтрации, выводится предупреждение Достигнуто максимальное количество строк сканирования: 1000.

2.2.3 Замена

Обработчик заменяет данные исходного набора, используя таблицы замен. Таблицы замен содержат пары заменяемых и новых значений или вычисляющие

их регулярные выражения. Эти таблицы можно задать вручную в мастере настройки (внутренние) или подать на вход обработчика (внешние).

Последовательность действий алгоритма замены:

1. Вначале производится поиск и замена по точному совпадению со значениями, указанными в таблице замен;
2. Среди значений, не найденных по точному совпадению, производится поиск по регулярным выражениям. Такие выражения могут быть заданы во внутренних таблицах замен. Новые значения также вычисляются регулярными выражениями;
3. Выполняются правила замены для значений, не найденных на предыдущих шагах.

Вход

- **Входной источник данных** (таблица данных) – набор данных, подлежащий изменению;
- **Присоединяемая таблица [N]** (таблица данных) – набор данных, содержащий таблицу замен;
- **+ Добавление еще одного порта.** Поскольку таблиц замен может быть несколько, то необходимые порты для них могут быть добавлены пользователем.

Выход

- **Выходной набор данных** (таблица данных). При настройке соответствия между столбцами порта можно выбрать варианты замены или добавления столбцов.

Мастер настройки

Окно мастера настройки состоит из 3-х областей:

1. Способы замен;
2. Таблица замен;
3. Дополнительные параметры.

Способы замен

Для каждого поля исходного набора задается способ замены:

- **Не заменять** – выходной набор данных замены производиться не будут;
- **Ввод вручную** – используется внутренняя таблица замен;
- **Таблица замен** (Таблица замен N) – используется внешняя таблица замен. Данный способ присутствует в настройках, если на входной порт обработчика подаются данные внешней таблицы замен.

Таблица замен N – наименование принимающего таблицу порта.

Таблица замен

Отображает внешнюю или внутреннюю таблицу замен для выбранного поля исходного набора.

Настройка внутренней таблицы

Для ввода новой строки в таблице замены используется кнопка **+**. Таблица может содержать строки для поиска и замены:

- по точному совпадению;
- по регулярному выражению (применимо только к данным строкового типа).

При вводе вручную таблицы замены доступны ряд действий через панель инструментов области:

- **Импорт** – импорт таблицы замен из текстового файла;
- **Экспорт** – экспорт таблицы замен в текстовый файл;
- **Сортировка** – сортировка таблицы по полю исходного значения;
- **Изменить тип замены** – задает тип данных столбца с новыми значениями;
- **Редактировать текущую замену** – отображает область редактирования текущей строки таблицы подстановок;

- **Получить значения** – подгружает выпадающий список, вызываемый кнопкой ✓, в поле *Значение*, при ручном вводе таблицы замен.

Примечание: Замена по точному совпадению имеет приоритет перед заменой по регулярному выражению.

Настройка внешней таблицы

Для полей таблицы необходимо выбрать **Назначение** из следующих вариантов:

- **Не используемое** – поле таблицы замен не будет использоваться;
- **Значение** – поле содержит заменяемые значения;
- **Замена** – поле содержит новые значения;
- **Информационное** – поле содержит дополнительный вариант нового значения. Дополнительные варианты замены выводятся в результирующем наборе данных отдельным столбцом.

Примечание: Если замена одного значения может быть произведена по условиям нескольких строк таблицы замен, то приоритет имеет первая из них. В связи с этим *сортировка таблицы замен может влиять на результат обработки.*

Дополнительные параметры

- **Заменять остальные** – содержит варианты замены значений, не найденных при помощи таблицы замен.
  - **Не заменять** – замены производиться не будут.
  - **На пропущенное** – значения будут заменены на Null.
  - **На значение** – значения будут заменены на указанное.
  - **На регулярное выражение** – новое значение будет вычислено с использованием синтаксиса регулярного выражения.
- **Точность** – для полей целого и вещественного типов задает допустимый интервал от указанных в таблицах замен значений, при котором исходное значение будет заменено.

- **Регистрозависимые строки** – флаг устанавливает регистро-зависимый режим поиска значений в таблицах замен. По умолчанию данный режим не используется.

2.2.4 Дата и время

Обработчик *Дата и время* производит трансформацию данных формата *Дата/Время*.

Суть такого преобразования заключается в том, что на основе столбца с информацией о дате/времени формируется один или несколько дополнительных столбцов, в которых выделяется определенная информация о дате/времени.

Пример:

Дата (исходный столбец)	(Год + Месяц)	Дата (Год + День)	Дата (День месяца)	Дата (Минута)
03.04.2012 00:04	01.04.2012	03.04.2012	3	4
17.04.2012 19:39	01.04.2012	17.04.2012	17	39
09.05.2012 19:42	01.05.2012	09.05.2012	9	42
16.05.2012 12:13	01.05.2012	16.05.2012	16	13
28.05.2012 20:35	01.05.2012	28.05.2012	28	35

На основе исходного столбца формируются остальные. В каждом сформированном столбце выделена определенная информация из даты, например, день месяца, минута.

Возможны варианты преобразования в следующие типы данных:

- *Дата/Время*;
- *Число*;
- *Строка*.

2.2.5 Группировка и разгруппировка

Узел Группировка позволяет объединять записи избранных полей в группы, а для оставшихся полей вычислять статистические показатели (сумму,

среднее, минимум и т.д.). Для каждой группы возвращается одна строка. Статистические показатели (или функции агрегации) при этом вычисляются для каждой группы, а не для всего набора в целом.

Пример.

Исходная таблица:

Дата	**Товар**	**Вес, кг**
12.07.2015	Яблоки	20
12.07.2015	Яблоки	18
12.07.2015	Помидоры	24
13.07.2015	Помидоры	22
13.07.2015	Груши	12
13.07.2015	Груши	16

В качестве полей-групп выберем поля *Дата* и *Товар*, а поле-параметр (по которому будет проводиться агрегация) – *Вес, кг*. Для примера применим три функции агрегации: сумма, количество записей и среднее.

Результирующая таблица:

Дата	**Товар**	**Вес, кг (Количество)**	**Вес, кг (Сумма)**	**Вес, кг (Среднее)**
12.07.2015	Помидоры	1	24	24
12.07.2015	Яблоки	2	38	19
13.07.2015	Груши	2	28	14
13.07.2015	Помидоры	1	22	22

Как видно из примера, группа образуется уникальным сочетанием значений полей, выбранных в качестве группировочных.

Вход

- **Входной источник данных** – порт для подключения входного набора данных.

Выход

- **Выходной набор данных** – порт отдающий сгруппированную таблицу.

Мастер настройки

Окно мастера поделено на две области.

- **Доступные поля** – содержит список полей входного набора данных.
- **Выбранные поля** – делится на списки *Группа* и *Показатели*.
  - **Группа** – поля группировки.
  - **Показатели** – поля, по которым рассчитываются функции агрегации.

Для настройки требуется переместить поля входного набора в списки *Группа* или *Показатели*, перетаскивая их мышью. Так же это можно сделать с помощью кнопок : **Переместить в Группу** (комбинация горячих клавиш **Alt+G**) и **Переместить в Показатель** (комбинация горячих клавиш **Alt+S**). Над списком доступных полей расположено поле *Фильтрация*, оно позволяет найти поле по имени или его части.

Настройка метода агрегации для каждого показателя производится в отдельном окне. Чтобы его открыть, нужно дважды кликнуть по полю в списке *Параметры* или вызвать это окно из контекстного меню. Далее отметить галочками нужные методы агрегации. Результат для каждого метода будет записан в отдельный столбец.

В нижней части мастера расположены два параметра отмечаемые чекбоксами:

- *Кэшировать значения групп* – результирующие данные будет закэшированы для использования последующими узлами;
- *Сортировать результирующие данные* – данные в результирующей таблице будут отсортированы по полям группировки в зависимости от их последовательности расположения в списке *Группы*.

Компонент Разгруппировка выполняет действие обратное агрегации по сумме, применяемой в Группировке. Для определения групп, по которым будет производиться обратная агрегация численного поля, требуется опорная таблица, из которой заимствуются сами группы и учитывается процентное распределение значений численного поля между ними.

Область основного применения разгруппировки – это детализация спрогнозированных данных на основе уже имеющихся. Например, если имеются данные о суммах совершенных продаж конкретных товаров, эти данные можно использовать для разгруппировки по товарам таблицы прогноза, сделанного только для товарных групп.

Пример:

Допустим, у нас есть прогноз сумм продаж, составленный для двух товарных групп, и нам нужно выявить из этого прогноза суммы продаж для отдельных товаров.

Разгруппируемые данные:

Группа товаров	**Сумма продаж, тыс. руб.**
Товары для дачи	42,00
Товары для дома	5,00

В качестве опорных данных будем использовать данные о суммах продаж за предыдущий период.

Данные для расчета долей:

Группа товаров	**Название товара**	**Сумма продаж, тыс. руб.**
Товары для дачи	Кресло плетеное	16,00
Товары для дачи	Лопата совковая	23,50
Товары для дома	Сахарница расписная	5,70
Товары для дома	Графин стеклянный	4,20
Товары для дачи	Термос стальной	7,60
Товары для дачи	Семена тюльпанов красных	5,30
Товары для дачи	Жидкость для розжига	6,20
Товары для дома	Розовая шипучка	1,60

Группа товаров	Название товара	Сумма продаж, тыс. руб.
Товары для дома	Мыло детское	2,90

При настройке узла Разгруппировка выберем метод *С расчетом долей по всей выборке*, выставим округление до одного знака после запятой и *Пропорциональный* метод балансировки. В области настройки назначений полей свяжем поля Группа товаров обеих таблиц, полю Сумма продаж, тыс. руб. из разгруппируемой таблицы выставим назначение *Разгруппируемое*, а полям Название товара и Сумма продаж, тыс. руб. из опорной таблицы – назначения *Поле с наименованиями* и *Поле с долями* соответственно.

Выход азгруппировки:

Группа товаров	Название товара	Сумма продаж группы, тыс. руб.	Сумма продаж группы, тыс. руб. | Округлено	Разгруппированное значение
Товары для дачи	Кресло плетеное	42,00	42,00	11,50
Товары для дачи	Лопата совковая	42,00	42,00	16,90
Товары для дачи	Термос стальной	42,00	42,00	5,40
Товары для дачи	Семена тюльпанов красных	42,00	42,00	3,80
Товары для дачи	Жидкость для розжига	42,00	42,00	4,40
Товары для дома	Сахарница расписная	5,00	5,00	1,90
Товары для дома	Графин стеклянный	5,00	5,00	1,50
Товары для дома	Розовая шипучка	5,00	5,00	0,60
Товары для дома	Мыло детское	5,00	5,00	1,00

- **Сумма продаж группы, тыс. руб.** – сумма продаж для конкретной группы;
- **Сумма продаж группы, тыс. руб. | Округлено** – в этом поле выводятся значения, получаемые при применении округления;

- **Разгруппированное значение** – в текущем примере это детализация продаж товаров, измеряемая в тыс. руб.

2.2.6 Калькулятор

Создает новые поля, которые вычисляются по заданной формуле из значений переменных, других полей и функций или используя JavaScript код.

Вход

- Входной источник данных (таблица данных);
- Входные переменные (переменные), необязательный.

Выход

- Выходной набор данных (таблица данных).

Мастер настройки

Окно настроек содержит области:

1. Список выражений;
2. Область кода выражений;
3. Поля/переменные;
4. Список функций.

Список выражений

Область предназначена для ввода *Выражений* – вычисляемых полей, которыми в результате обработки будет дополнен входной набор данных. Значение выражения в каждой строке набора данных будет вычислено по формуле или JavaScript коду.

Новое выражение можно создать при помощи панели инструментов области или контекстного меню.

- **Синтаксис** (выпадающий список) – задание синтаксиса расчета выражений калькулятора;
- **Редактировать** – задание параметров выражения;

- **Переместить вверх** – поднять выражение на одну позицию вверх по списку;
- **Переместить вниз** – опустить выражение на одну позицию вниз по списку;
- **Добавить выражение** – добавляет новое выражение с параметрами по умолчанию;
- **Добавить выражение по образцу** – добавляет новое выражение с типом данных, описанием и формулой, как у текущего выражения;
- **Удалить выражение** – удаляет текущее выражение;
- **Удалить все выражения** – удаляет все имеющиеся выражения.

При добавлении и редактировании выражения отображается диалог редактирования параметров. Следующие параметры выражений доступны для изменений:

- **Имя** – вводится имя поля, присваемое столбцу в выходном наборе данных;

Важно: Имя должно быть уникальным, начинаться с заглавной или строчной латинской буквы или с символа подчеркивания. Последующие символы имени могут быть такими же, либо цифрами.

- **Метка** – вводится метка поля, присваемая столбцу в выходном наборе данных;
- **Тип данных** – выбирается тип данных поля в выходном наборе данных;
- **Промежуточное** – при установке этого флага выражение может использоваться в расчетах, не включается в список полей выходного набора данных;
- **Кэшировать** – сохранение однажды вычисленного значения выражения, целесообразно при неоднократном использовании значений выражения

последующими обработчиками и визуализаторами во избежание выполнения повторных вычислений;

Важно: *Кэширование* необходимо включать в выражениях *Калькулятора* в случае использования функции Data() при рекурсивном вычислении значений.

Рекомендуется применять *Кэширование* при использовании функций, результат которых зависит от момента времени, в которое происходит это вычисление, например: Random(), Today() и других.

- **Описание** – поясняющая информация.

Важно: Имя должно быть уникальным, начинаться с заглавной или строчной латинской буквы или с символа подчеркивания. Последующие символы имени могут быть такими же, либо цифрами.

При первом открытии мастера настройки список выражений содержит один элемент с именем Expr0 вещественного типа. По умолчанию для нового выражения назначается метка ВыражениеN и имя ExprN, где N – номер, обеспечивающий уникальность.

Область кода выражения

В области кода в зависимости от выбранного синтаксиса калькулятора задается формула расчета выражения или JavaScript код. Ссылки на поля/переменные и синтаксические конструкции функций можно вставлять в код выражения, выбрав их двойным кликом мыши в соответствующих областях или перетащив мышкой.

Изменения в области кода сохраняются при выходе из нее.

Поля/переменные

Область содержит список полей и переменных, передаваемых на вход обработчика. Перечень и параметры полей/переменных определяются при настройке входных портов обработчика.

Двойной клик мыши по позиции списка вводит имя поля/переменной в область кода выражения. То же самое можно сделать при помощи Drag-and-drop.

Категории функций

- Дата/время
- Модели
- Логические функции
- Математические функции
- Статистические функции
- Строковые функции
- Финансовые функции
- Прочие функции

Возможна фильтрация по категории и названию функции.

Двойной клик мыши по позиции выбранной функции вставляет ее синтаксис в область кода выражения. То же самое можно сделать при помощи Drag-and-drop.

Ссылки на поля/переменные и синтаксические конструкции функций можно вставлять в код выражения, выбрав их двойным кликом мыши в соответствующих областях или перетащив мышкой.

2.1.7 Квантование

Квантование применяется к данным с типами: целый, вещественный и дата/время.

Обработчик разбивает диапазон значений выбранных полей исходного набора на конечное число интервалов. Для разбиения возможно применение различных алгоритмов (см. далее методы квантования), а также использование *внешних таблиц* с заданными интервалами квантования (чтобы их подключить, нужно добавить еще один порт у обработчика, нажав кнопку **+**. В появившийся порт подключается таблица с данными о входящих диапазонах.

Вход

- Входной источник данных (таблица данных).

- **+** Добавить еще один порт. Внешние диапазоны квантования (таблица данных).

Выход

- Выходной набор данных (таблица данных). Описание структуры результирующего набора.
- Диапазоны для квантования (таблица данных). Описание структуры таблицы диапазонов.

Мастер настройки

Мастер настройки состоит из двух основных областей: область настройки параметров квантования и область отображения результатов квантования. Обе области организованы в виде таблицы. Над ними располагается строка состояния входа.

Область настройки параметров квантования

Область представлена в виде таблицы. Над полями расположены три кнопки:

- Редактировать – при нажатии позволяет редактировать параметры квантования для выбранного поля;
- Уменьшить разрядность – каждое нажатие кнопки уменьшит разрядность границ интервалов на один знак после запятой;
- Увеличить разрядность – каждое нажатие кнопки увеличит разрядность границ интервалов на один знак после запятой.

Таблица этой области состоит из нескольких столбцов:

1. **Поле** – содержит поля исходного набора данных, к которым применима процедура квантования. Это поля типа: целый, вещественный, дата/время.
2. **Метод** – поле представлено раскрывающимся списком, из которого необходимо выбрать метод квантования:

- Ширина – пользователь может выбирать ширину интервала, а количество интервалов рассчитывается автоматически, как отношение разности верхней и

нижней границ к заданной ширине. Выставив соответствующие флаги, можно задать:

- Верхнюю границу – верхняя граница самого высокого интервала;
- Нижнюю границу – нижняя граница самого высокого интервала.

• Количество – выбирается количество интервалов, а ширина рассчитывается автоматически, как отношение разности верхней и нижней границ к заданному количеству интервалов. Для этого метода так же можно задать верхнюю и нижнюю границы.

• Плитка – пользователь выбирает количество интервалов, а компонент задает диапазоны интервалов таким образом, чтобы в каждом интервале было примерно одинаковое количество значений. Имеется несколько способов обработки совпадающих значений:

– Добавлять в следующий – перенесет значения совпадающих наблюдений в следующий (более высокий) интервал разделения.

– Сохранять в текущем – сохраняет значения совпадающих наблюдений в текущем (более низком) интервале разделения. Этот метод может привести к тому, что всего будет создано меньше интервалов.

– Назначать случайно – типы границ интервалов будут определены случайно; возможно включение одинаковых значений в тот или иной интервал случайным образом.

– Оставить как есть – границы всех интервалов будут иметь тип **>=**, и возможна ситуация, когда совпадающие значения окажутся в разных интервалах.

– Одинаковые плитки – достижение равного количества значений в интервалах обеспечивается не только подбором диапазонов интервалов, но и подбором типов границ для каждого интервала (**>** или **>=**).

• Коэффициенты СКО – разбивает значения на интервалы в зависимости от выбранного диапазона, выраженного в количестве **σ** (СКО).

• Внешние диапазоны.

Для всех методов квантования можно установить флаг *Округлять границы*.

• **Автоматически** – установленная галочка в этом поле обеспечивает автоматическую настройку параметров квантования выбранного метода.

• **Интервалов** – количество интервалов, на которые будут разбиты значения поля.

• **Минимум** – отображается минимальное значение квантуемого поля.

• **Максимум** – отображается максимальное значение квантуемого поля.

Далее в каждой строке располагается кнопка "рассчитать интервалы" и в шапке таблицы "рассчитать все интервалы". При их нажатии пересчитываются параметры квантования (количество интервалов, минимум, максимум) в зависимости от изменения методов и/или настроек параметров. Этот функционал доступен только при состоянии "Вход активирован".

Область отображения результатов квантования

В этой области отображаются результаты квантования с возможностью их редактирования.

Над полями таблицы расположены несколько элементов управления:

- **Нижняя граница открыта** – убирает нижнюю границу;
- **Верхняя граница открыта** – убирает верхнюю границу;
- **Инвертировать тип** – меняет тип границ;
- – пересчитывает гистограмму согласно новым параметрам.
- **Шаблон** – в этом поле происходит настройка шаблона для отображения метки интервала, в нем можно составить пользовательский шаблон или при нажатии на выбрать один из готовых шаблонов. Чтобы применить шаблон необходимо нажать кнопку .
- **Образец** – при клике на эту кнопку открывается таблица обозначений, которые можно использовать при составлении шаблона.

Под элементами управления расположена таблица с результатами квантования выделенного поля, она содержит следующие поля:

- **№** – номер интервала;

- **Нижняя** – нижняя граница интервала;
- **Тип** – тип границы;
- **Верхняя** – верхняя граница интервала;
- **Метка** – метка интервала (ее можно задавать шаблоном);
- **Объём** – отображает объем значений, попавших в интервал (отображается в виде гистограммы).

2.2.8 Скользящее окно

Обработка данных методом скользящего окна применяется при предварительной обработке данных в задачах прогнозирования, когда на вход анализатора (например, нейронной сети) требуется подавать значения нескольких смежных отсчетов исходного набора данных. Термин скользящее окно отражает сущность обработки – выделяется некоторый непрерывный отрезок данных, называемый окном, а окно, в свою очередь, перемещается, "скользит" по всему исходному набору данных.

В результате будет получен набор данных, где в одном поле будет содержаться значение, соответствующее текущему отсчету (оно будет иметь то же имя, что и в исходном наборе), а слева и справа от него будут расположены поля со значениями, смещенными от текущего отсчета в прошлое и в будущее соответственно.

Следовательно, обработка методом скользящего окна имеет два параметра:

- **Глубина истории** – количество отсчетов в "прошлое";
- **Горизонт прогноза** – количество отсчетов в "будущее".

Необходимо отметить, что для граничных положений окна (конец и начало исходной выборки) будут формироваться неполные записи: в начале исходной выборки будут формироваться пустые значения для "прошлых" отсчетов, а в конце – для "будущих". В зависимости от конкретной ситуации пользователь может включать такие неполные записи в результирующую выборку или исключать их.

Порты

Вход

- **Входной источник данных** – порт для подключения входного набора данных.

Выход

- **Выходной набор данных** – на порт выводится таблица с набором данных дополненным смещенными полями.

Мастер настройки

Окно мастера настройки содержит список полей входной таблицы, для каждого поля имеются настраиваемые параметры:

- **Глубина истории** – количество значений из предыдущих записей, для которых создаются новые поля в выходном наборе данных;
- **Горизонт прогноза** – количество значений из последующих записей, для которых создаются новые поля в выходном наборе данных.

Параметр *Способ обработки неполных записей* предоставляет следующие методы:

- **Оставлять неполные записи** – сохраняет все добавленные узлом записи;
- **Удалять добавленные неполные записи** – удаляет записи, добавленные узлом, не трогая записи из изначального набора;
- **Удалять все неполные записи** – удаляет записи, добавленные узлом и записи с пустыми значениями в добавленных полях.

Пример. Варианты результирующей таблицы из примера с разными *Способами обработки неполных записей*.

Исходная таблица:

Дата	Продажи, шт.
01.01.2020	45
01.02.2020	82
01.03.2020	120
01.04.2020	192

Дата	Продажи, шт.
01.05.2020	229
01.06.2020	161

Для поля Продажи, шт. настроим параметр *Глубина истории* равным двум, а параметр *Горизонт прогноза* – равным единице. В зависимости от параметра *Оставлять неполные записи* получим разные результирующие таблицы.

Результирующая таблица при значении *Оставлять неполные записи*:

Дата	Продажи, шт.[-2]	Продажи, шт.[-1]	Продажи, шт.	Продажи, шт.[+1]
				45
01.01.2020			45	82
01.02.2020		45	82	120
01.03.2020	45	82	120	192
01.04.2020	82	120	192	229
01.05.2020	120	192	229	161
01.06.2020	192	229	161	
	229	161		
	161			

Результирующая таблица при значении *Удалять добавленные неполные записи*:

Дата	Продажи, шт.[-2]	Продажи, шт.[-1]	Продажи, шт.	Продажи, шт.[+1]
01.01.2020			45	82
01.02.2020		45	82	120
01.03.2020	45	82	120	192
01.04.2020	82	120	192	229
01.05.2020	120	192	229	161
01.06.2020	192	229	161	

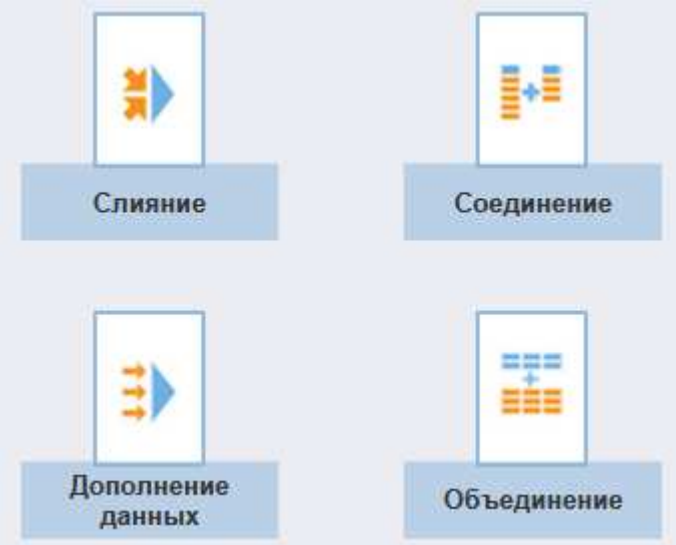
Результирующая таблица при значении *Удалять все неполные записи*:

Дата	Продажи, шт.[-2]	Продажи, шт.[-1]	Продажи, шт.	Продажи, шт.[+1]
01.04.2020	45	82	120	192
01.05.2020	82	120	192	229

Дата	Продажи, шт.[-2]	Продажи, шт.[-1]	Продажи, шт.	Продажи, шт.[+1]
01.06.2020	120	192	229	161

2.2.9. Компоненты связи для нескольких наборов данных

- Для связи нескольких наборов данных АП Loginom предлагает 4 компонента:



Слияние наборов данных, связанных по ключевым полям – аналог операции JOIN в SQL. Данные ключевых полей основного и присоединяемого наборов сравниваются между собой, результат операции сравнения используется алгоритмом слияния для формирования результирующего набора.

Варианты слияния наборов:

- Полное соединение;
- Внутреннее соединение;
- Левое соединение;
- Правое соединение;
- Разность.

Вход

- **Главная таблица** – в контексте понятий языка SQL запросов принимает данные левой таблицы для соединения;

- **Присоединяемая таблица** – в контексте понятий языка SQL запросов принимает данные правой таблицы для соединения.

Выход

- **Выходной набор данных** – результат слияния. Как правило, содержит поля основного и присоединяемого наборов.

Мастер настройки

- **Тип операции** – выбор способа слияния;
- **Область сопоставления полей** – настройка полей связи главного и присоединяемого наборов данных.

Поля связываются при помощи перетаскивания мыши (Drag-and-drop). Сопоставленные таким образом ключевые поля соединяются линиями связей. Связь между полями можно удалить, либо настроить на другое поле.

Связывание допускается только для полей с одинаковыми типами данных.

Пример полного соединения:

Аналогом данного способа слияния являются SQL-операторы CROSS JOIN и FULL JOIN.

При CROSS JOIN соединении производится перекрестное соединение (или декартово произведение) – каждая строка одной таблицы соединяется с каждой строкой второй таблицы, давая тем самым в результате все возможные сочетания строк двух таблиц. При таком соединении порядок таблиц (левая, правая) неважен, и отсутствует необходимость в сопоставлении ключевых полей.

Пример:

Для примера возьмем две таблицы. Персона – главная таблица и присоединяемою Город.

Главная таблица:

Имя	Id города
Андрей	1

Присоединяемая таблица:

Id	Город
1	Москва

Леонид	2
Сергей	1
Григорий	4

2	Санкт-Петер-бург
3	Казань

Результирующая таблица:

Имя	**Id города**	**Город**
Андрей	1	Москва
Андрей	1	Санкт-Петербург
Андрей	1	Казань
Леонид	2	Москва
Леонид	2	Санкт-Петербург
Леонид	2	Казань
Сергей	1	Москва
Сергей	1	Санкт-Петербург
Сергей	1	Казань
Григорий	4	Москва
Григорий	4	Санкт-Петербург
Григорий	4	Казань

При FULL JOIN соединении производится полное внешнее соединение двух наборов. В результирующий набор добавляются следующие записи:

1. Внутреннее соединение (INNER JOIN) первой и второй таблиц;
2. Записи первой таблицы, которые не вошли во внутреннее соединение на шаге 1. Для таких записей поля, соответствующие второй таблице, заполняются значениями NULL;
3. Записи второй таблицы, которые не вошли во внутреннее соединение на шаге 1. Для таких записей поля, соответствующие первой таблице, заполняются значениями NULL.

При таком соединении необходимо сопоставление ключевых полей, но порядок таблиц (левая, правая) неважен.

Пример:

Для примера возьмем две таблицы. Персона – главная таблица и присоединяемою Город.

Главная таблица:

Имя	Id города
Андрей	1
Леонид	2
Сергей	1
Григорий	4

Присоединяемая таблица:

Id	Город
1	Москва
2	Санкт-Петербург
3	Казань

Результирующая таблица:

Имя	Id города	Город
Андрей	1	Москва
Леонид	2	Санкт-Петербург
Сергей	1	Москва
<null>	<null>	Казань
Григорий	4	<null>

Важно: Для того, чтобы при способе слияния *Полное соединение* использовать FULL JOIN соединение, необходимо в мастере настройки сопоставить ключевые поля соединяемых наборов. Если сопоставление отсутствует, то задействуется алгоритм CROSS JOIN соединения. При данном способе слияния объем результирующей выборки может очень быстро расти.

Пример внутреннего соединения:

Аналогом такого соединения является SQL-оператор INNER JOIN.

При INNER JOIN происходит соединение двух таблиц. Для данного способа слияния порядок таблиц не важен. Результирующий набор данных содержит все строки, для которых найдено совпадение ключевых полей главной и присоединяемой таблицы.

Пример:

Для примера возьмем две таблицы. Персона – главная таблица и присоединяемою Город.

Главная таблица:

Присоединяемая таблица:

Имя	Id города
Андрей	1
Леонид	2
Сергей	1
Григорий	4

Id	Город
1	Москва
2	Санкт-Петер-бург
3	Казань

Результирующая таблица:

Имя	Id города	Город
Андрей	1	Москва
Леонид	2	Санкт-Петербург
Сергей	1	Москва

Пример левого соединения:

Аналогом данного вида слияния является SQL-оператор LEFT JOIN. Левой таблицей является основной набор данных.

При LEFT JOIN производится соединении двух таблиц – главной (левая таблица) и присоединяемой (правая таблица). В результирующий набор добавляются следующие записи:

1. Внутреннее соединение (INNER JOIN) левой и правой таблиц по ключевым полям;
2. Затем в результат добавляются те записи левой таблицы, которые не вошли во внутреннее соединение на шаге 1. Для таких записей поля, соответствующие правой таблице, заполняются значениями NULL.

Пример:

Для примера возьмем две таблицы. Персона – главная таблица и присоединяемою Город.

Главная таблица:

Имя	Id города
Андрей	1
Леонид	2
Сергей	1

Присоединяемая таблица:

Id	Город
1	Москва
2	Санкт-Петер-бург

Григорий	4

3	Казань

Результирующая таблица:

Имя	**Id города**	**Город**
Андрей	1	Москва
Леонид	2	Санкт-Петербург
Сергей	1	Москва
Григорий	4	<null>

Соединение

С помощью обработчика *Соединение* исходный набор данных дополняется полями присоединяемых наборов. При этом каждая запись исходного набора соединяется с записью такого же порядкового номера дополнительного набора.

Если соединяемые наборы данных имеют разное количество записей, то результирующий набор может содержать пустые значения. Мастер настройки предлагает различные варианты обработки данной ситуации. Соединяемые наборы могут обрезаться до количества записей меньшего набора или дополняться до наибольшего.

К исходному набору можно присоединять переменные. Каждая из присоединяемых переменных добавляет новый столбец к исходному набору. При этом в зависимости от настроек параметра *Дополнение до наибольшего набора* значение присоединенной переменной будет добавлено:

- во все строки нового столбца;
- только в первую строку, а в остальных строках для полей строкового и переменного типа будет выведено значение null, для полей остальных типов данных – пустая ячейка.

Вход

- **Главная таблица** – порт для входного набора данных.
- **Добавить еще один порт** – создает новые порты входа для присоединяемых таблиц и переменных. Новые порты могут быть двух типов:

- Присоединяемая таблица [N], где N порядковый номер таблицы;
- Присоединяемые переменные [N], где N порядковый номер порта переменных.

Выход

- **Выходной набор данных** – таблица с присоединенными столбцами.

Мастер настройки

Для настройки доступны следующие параметры:

- **Дополнение до наибольшего набора** – предлагается выбрать один из вариантов дополнения наименьших по количеству записей таблиц:
  - **Не дополнять** – дополнение записями, поля которых будут содержать пустые значения;
  - **Повторять набор** – таблицы дополнятся копиями своих же записей, начиная с первой;
  - **Дополнять последней строчкой** – дополнение копиями последней строки.
- **Количество строк соответствует** – предлагается выбрать один из вариантов определения количества записей результирующего набора данных:
  - по **Минимальному набору**;
  - по **Максимальному набору**;
  - **Определяется набором** – при выборе данного варианта становится доступен список *Набор данных, определяющий набор строк*, в нем необходимо выбрать набор, в соответствии с которым будет определяться количество строк результирующего набора.

Пример:

Для примера возьмем две таблицы.

Главная таблица:

Присоединяемая таблица:

ФИО	Год рождения
Абрамов	1972 г.
Авдеева	1956 г.
Агафонов	1978 г.
Аксёнова	1979 г.
Александров	1980 г.
Алексеев	1983 г.
Андреева	1982 г.
Анисимов	1963 г.
Антонов	1984 г.
Артемьев	1965 г.

КТУ	Кластер
> 0.8	1
> 0.8	1
0.5 - 0.8	2
0.5 - 0.8	2
0.2 - 0.5	3
< 0.2	4

- Параметр *Дополнение до наибольшего набора* выставлен в значение *Не дополнять*, параметр *Количество строк соответствует* – в значение *Максимальному набору*.

Результирующая таблица:

ФИО	Год рождения	КТУ	Кластер
Абрамов	1972 г.	> 0.8	1
Авдеева	1956 г.	> 0.8	1
Агафонов	1978 г.	0.5 - 0.8	2
Аксёнова	1979 г.	0.5 - 0.8	2
Александров	1980 г.	0.2 - 0.5	3
Алексеев	1983 г.	< 0.2	4
Андреева	1982 г.	<null>	<null>
Анисимов	1963 г.	<null>	<null>
Антонов	1984 г.	<null>	<null>
Артемьев	1965 г.	<null>	<null>

Дополнение данных

Соединение таблиц данных на основе связи по ключевым полям – аналог операции LEFT JOIN в SQL. Узел выполняет действие аналогичное Левому соединению узла Соединение, но количество присоединяемых таблиц произвольно.

Порты

Вход

- **Главная таблица** – в контексте понятий языка SQL-запросов является левой таблицей для соединения;
- **Присоединяемая таблица** – в контексте понятий языка SQL-запросов является правой таблицей для соединения;
- **Добавить еще один порт** – создает новые порты входа для последующих присоединяемых таблиц, которые будут автоматически пронумерованы.

Выход

- **Выходной набор данных** – таблица, содержащая поля всех таблиц, поданных на входные порты, кроме полей присоединяемых таблиц, выбранных в качестве ключевых. По желанию к меткам полей присоединяемых таблиц можно добавить префиксы.

Мастер настройки

- **Область настройки ключевых полей** – напротив поля главной таблицы, которое должно стать ключевым, следует выставить флаг в столбце присоединяемой таблицы. Из выпадающего списка необходимо выбрать поле, по которому таблицы будут связываться. При включенной фильтрации доступны совместимые поля, которые еще не связаны с ключевыми полями главной таблицы, при отключенной фильтрации можно выбрать любое из полей, совместимых по типу.
- **Использовать префиксы** – включение данного флага позволяет добавить в результирующей таблице префиксы к именам и меткам полей, взятых из присоединяемых таблиц.
  - **Префикс имени** – в это поле вводится префикс, добавляемый к имени присоединенных полей таблиц, состав именного префикса следует правилу Параметров полей набора данных.

- **Префикс метки** – в это поле вводится префикс, добавляемый к метке присоединенных полей таблиц, именуется согласно *Параметрам полей набора данных*.

Пример:

Для примера возьмем три таблицы. Персона – главная таблица, и две присоединяемых: Город и Регион.

Главная таблица:

Имя	Id города
Андрей	1
Леонид	2
Сергей	1
Григорий	4

Присоединяемая таблица 1:

Id	Город
1	Москва
2	Санкт-Петербург
3	Казань

Присоединяемая таблица 2:

Id города	Регион
1	Центральный
2	Северо-западный
3	Приволжский
4	Дальневосточный

Рис. 46. Порядок связей при присоединении

Результирующая таблица:

Имя	Id города	Город	Регион
Андрей	1	Москва	Центральный
Леонид	2	Санкт-Петербург	Северо-западный
Сергей	1	Москва	Центральный
Григорий	4	<null>	Дальневосточный

Объединение

С помощью обработчика Объединение исходный набор данных дополняется записями присоединяемых наборов. Объединение является аналогом операции UNION в SQL.

Пример:

Исходная таблица:

ФИО	Год
Андреева	1982 г.
Анисомов	1963 г.
Антонов	1984 г.
Артемьев	1965 г.

Присоединяемая таблица:

ФИО	Год	КТУ	Кластер
Абрамов	1972 г.	> 0.8	1
Авдеева	1956 г.	> 0.8	1
Агафонов	1978 г.	0.5 - 0.8	2
Аксёнова	1979 г.	0.5 - 0.8	2
Александров	1980 г.	0.2 - 0.5	3
Алексеев	1983 г.	< 0.2	4

Результирующая таблица:

ФИО	Год	КТУ	Кластер
Андреева	1982 г.	null	null
Анисомов	1963 г.	null	null
Антонов	1984 г.	null	null
Артемьев	1965 г.	null	null
Абрамов	1972 г.	> 0.8	1
Авдеева	1956 г.	> 0.8	1
Агафонов	1978 г.	0.5 - 0.8	2
Аксёнова	1979 г.	0.5 - 0.8	2
Александров	1980 г.	0.2 - 0.5	3
Алексеев	1983 г.	< 0.2	4

Вход

- **Главная таблица** – первая таблица, участвующая в объединении;
- **Присоединяемая таблица** – вторая таблица, участвующая в объединении, все последующие таблицы добавляются через *Добавить еще один порт*;
- **Добавить еще один порт** – создает новые порты входа для последующих присоединяемых таблиц, которые будут автоматически пронумерованы.

Выход

- **Выходной набор данных** – таблица, содержащая поля всех таблиц, поданных на входные порты, кроме полей присоединяемых таблиц, выбранных в качестве сопоставляемых. Выбранные поля объединяются и присоединяются к набору данных согласно проставленному сопоставлению. Поля без сопоставления дополняют набор данных. По желанию к меткам дополненных полей можно добавить префиксы.

Мастер настройки

Полям главной таблицы необходимо сопоставить поля присоединяемой таблицы. В результирующем наборе данные сопоставленных полей объединяются в одно поле. Это поле получает *Имя* и *Метку* поля главной таблицы. Данные не сопоставленных полей помещаются в отдельные столбцы результирующего набора, которые можно отметить префиксами.

Сопоставление возможно только для полей с одинаковым типом данных. При первоначальном открытии мастера поля с одинаковым именем и типом данных сопоставляются автоматически. Ручная настройка осуществляется с помощью элементов управления:

- **Область настройки сопоставления** – представляет таблицу где слева представлены поля *Главной таблицы*, а справа *Подключаемые таблицы*, отмечаемые чекбоксами и выпадающими списками. Обозначение *Подключаемых таблиц* идет таким образом: *Подключаемая таблица*, *Подключаемая таблица 2 ... Подключаемая таблица N*.
  - **Чекбокс** – в этих столбцах у *Подключаемых таблиц* флажком отмечаются поля участвующие в сопоставлении.
  - **Раскрывающиеся списки полей** – содержатся в каждой строке таблицы сопоставления. Список позволяет выбрать поле присоединяемой таблицы, которое будет сопоставлено полю главной таблицы.

- **Использовать префиксы** – применяется, если необходимо выделить не сопоставленные столбцы присоединяемых таблиц. Для таких столбцов в результирующем наборе данных можно задать.
  - **Префикс имени** – в это поле вводится префикс, добавляемый к имени не сопоставленных полей таблиц, состав именного префикса следует правилу Параметров полей набора данных.
  - **Префикс метки** – в это поле вводится префикс, добавляемый к метке не сопоставленных полей таблиц, именуется согласно *Параметрам полей набора данных*.

2.2.10 Компоненты Переменные в таблицу и Таблица в переменные

Переменные в таблицу

Обработчик преобразовывает список переменных в таблицу. Значения переменных можно расположить в строках, либо в столбцах результирующей таблицы.

Порты

Вход

- **Входные переменные** – список переменных, подлежащий преобразованию в таблицу.

Выход

- **Выходной набор данных** – таблица данных.

Мастер настройки

В окне мастера настройки предоставляется два варианта записи переменных:

- **в столбцы** – каждой переменной будет соответствовать отдельное поле таблицы;
- **в строки** – каждой переменной будет соответствовать отдельная запись таблицы.

Пример:

Имеется список переменных:

Имя	Значение
a	5
b	7
c	10
d	2

Результат преобразования списка переменных:

- В столбцы

a(сумма)	b(сумма)	c(сумма)	d(сумма)
5	7	10	2

- В строки

Имя	Метка	Значение
a	a(сумма)	5
b	b(сумма)	7
c	c(сумма)	10
d	d(сумма)	2

Таблица в переменные

Обработчик позволяет преобразовать табличные данные в переменные. Из данных каждого поля таблицы формируется переменная. У переменной может быть только одно значение, поэтому для его расчета к данным поля применяются различные функции агрегации.

Порты

Вход

- **Входной источник данных** – таблица данных.

Выход

- **Выходные переменные** – список переменных, полученный из входного источника данных с использованием функций агрегации по полю.

Мастер настройки

Окно мастера поделено на две области:

- **Доступные поля** – представляет список полей входного набора данных;
- **Выбранные поля** – представляет список результирующих переменных.

Перемещение полей между областями возможно с помощью Drag-and-Drop.

Доступные поля

Добавить выделенные поля в список *Выбранные поля* можно:

- Кнопкой  ;
- Через контекстное меню "Переместить в Переменные";
- Нажав Alt+S.

Выбранные поля

При добавлении поля в список *Выбранные поля* функция агрегации будет назначена по-умолчанию:

- для чисел будет выбрана функция *сумма*;
- для остальных – функция *количество*.

Действия над выбранными полями можно осуществлять мышью. Перетаскиванием внутри списка можно менять позицию в выделенных полей. Исключить можно с помощью переноса полей в список *Доступные поля*. Двойной клик по полю открывает окно "Выбора агрегации".

Эти же действия выполняют кнопки на панели быстрого доступа:

-  – перемещает выделенный элемент вверх по списку;
-  – перемещает выделенный элемент вниз по списку;
-  – для выделенных полей открывает окно выбора доступных вариантов агрегации;
-  – перемещает текущий элемент в список *Доступные поля*;

-  – перемещает все элементы в список *Доступные поля*.

Контекстное меню дублирует общие функции:

- **Переместить вверх** – перемещает выделенные поля вверх по списку;
- **Переместить вниз** – перемещает выделенные поля вниз по списку;
- **Редактировать** – для выделенных полей открывает окно выбора доступных вариантов агрегации;
- **Удалить выбранные** – перемещает выделенные поля обратно в список *Доступные поля*.

Доступны горячие клавиши, дублирующие перечисленные команды:

- Ctrl+Up – Переместить вверх;
- Ctrl+Down – Переместить вниз;
- F2 – Редактировать;
- Delete – Удалить выбранные.

Содержание выходного набора

В выходном наборе будут переменные, полученные из полей с помощью выбранных функций агрегации. Каждому варианту агрегации на выходе будет соответствовать отдельная переменная.

Имена, метки и значения переменных будут получены следующим образом:

- **Имя** – будет совпадать с именем поля, если была выбрана лишь одна функция. Если функций было больше одной, то к имени добавится окончание, соответствующее выбранной функции.
- **Метка** – будет состоять из имени поля и названии функции агрегации.
- **Значение** – формируется из всех значений поля, агрегированных выбранной функцией.

2.2.11 Компоненты Выполнение и Цикл

Компонент **Выполнение узла** позволяет повторно использовать уже имеющийся узел сценария для обработки новых данных.

Базовый узел – узел или Подмодель, настройки и алгоритм которого предполагается использовать повторно, может находиться за пределами текущей *Подмодели*, *Сценария*, *Модуля* или *Пакета*.

Порты

При создании узел не имеет портов. После настройки узел имеет порты, идентичные портам базового узла. Переопределить настройки портов возможно, однако, следует иметь в виду, что перенастройка входных портов может привести к несоответствию новых входных данных требованиям сценария, унаследованного от базового узла, и, как следствие, к ошибке выполнения.

Мастер настройки

При помощи радиокнопки необходимо выбрать узел сценария, который планируется повторно использовать для обработки новых данных. *Доступные* для выбора узлы представлены в виде дерева.

Дерево узлов имеет как минимум две корневые ветки:

- **Текущий модуль** – содержит перечень узлов модуля, в котором в данный момент создается узел *Выполнение узла*;
- **Текущий пакет** (наименование пакета) – содержит перечень узлов текущего пакета.

Если в текущем пакете настроены ссылки на внешние пакеты, то в дереве в отдельных ветках отобразятся узлы этих пакетов.

Важно: Узлы из внешних пакетов, других сценариев и подмоделей отобразятся в дереве только в том случае, если доступ к ним разрешен их модификатором доступа.

Интерфейс мастера предоставляет возможность осуществлять поиск узлов в дереве с помощью фильтров: по имени и комментарию узла.

Компонент *Выполнение узла* не может повторно использовать компоненты Цикл и Узел-ссылка.

Примечание:

- При выполнении узла базовый узел не выполняется.
- Возможно переобучение модели, унаследованной из базового узла. Однако, модель самого базового узла останется неизменной.

Компонент **Цикл** может применяться для циклического исполнения выбранного пользователем узла. В качестве такого узла, как правило, используется Подмодель, в которой задаются действия, выполняемые над данными в каждой итерации. Возможны следующие способы задать итерации цикла:

1. **Заданное количество раз** – аналог "FOR ... TO ...".
2. **Количество итераций задается условием выхода из цикла** – аналог "DO ... WHILE ...". На предмет соответствия этому условию анализируется значение выходной переменной узла, заключенного в цикл.
3. **Количество итераций задается количеством групп, на которые разделяются строки входного набора данных** – аналог "FOR EACH". В предельном случае количество итераций равно количеству строк входного набора.

В первом и втором случае применение входного набора данных не обязательно. Но, если таковой применяется, то в каждой итерации на вход узла, заключенного в цикл, подаются все строки этого набора (наборов).

В третьем случае строки входного набора разделяются по заданному признаку на группы строк, и в каждой итерации на вход узла, заключенного в цикл, подаются строки очередной группы. Если в качестве признака группы указываются уникальные идентификаторы строк входного набора, то такая группа будет содержать только одну строку. В этом случае цикл будет перебирать строки входного набора и передавать очередную строку на вход узла, заключенного в цикл.

Порты

При первоначальном создании узел не имеет портов. После задания параметров в мастере настройки узел цикла имеет набор портов узла, заключенного в цикл, кроме выходных портов для переменных.

Мастер настройки

Шаг 1. Выбор узла цикла

На данном этапе предоставляется список узлов текущего модуля. Для заключения узла в цикл следует отметить его кнопкой и перейти к следующему шагу.

Интерфейс мастера предоставляет возможность осуществлять поиск узлов в общем списке с помощью фильтров: по имени и комментарию узла.

Не допускается выполнение в цикле обработчиков, созданных на базе следующих компонентов: Выполнение узла, Узел-ссылка, Условие и *Цикл*.

Шаг 2. Настройка вида цикла

Исходный узел – информационное поле, отображающее узел, который заключается в цикл.

Вид цикла – определение логики работы цикла, задается радиокнопкой:

- **Заданные итерации** – данный вид цикла будет выполняться столько раз, сколько будет задано в параметре "Количество итераций".
- **Цикл с постусловием** – количество итераций такого цикла регулируется условием, на соответствие которому при каждой итерации проверяется переменная. На итерации, при которой значение переменной удовлетворяет условию, работа цикла заканчивается. Задаются следующие настройки:
  - **Переменная** – выбор переменной для условия выхода из цикла. Список выбора содержит переменные, передаваемые на выходные порты узла, заключенного в цикл.
  - **Условие завершения** – отношение сравнения переменной со Значением.

- **Значение** – поле для ввода значения, с которым будет сравниваться переменная. Следует учитывать, что в узле, заключенном в цикл, должна присутствовать операция над этой переменной, которая приводит к выполнению условия выхода из цикла, иначе цикл получится бесконечным.

- **Групповая обработка** – этот цикл разделяет исходные данные на группы по выбранному критерию, благодаря чему узел, заключенный в цикл, обрабатывает каждую группу данных отдельно. Критерий разделения определяется параметром "Вид групповой обработки":

  - **Фиксированный размер групп** – исходный набор данных будет последовательно разделен на группы, размер которых определяется параметром "количество строк в группе". Если количество строк в группе не кратно количеству строк в исходном наборе, размер последней группы окажется меньше.

  - **Фиксированное количество групп** – исходный набор данных будет разделен на заданное количество групп. Это количество задается параметром "Количество групп". В случае, если количество групп не кратно количеству строк в исходном наборе, в некоторых группах количество строк будет отличаться, и они будут равномерно распределены среди остальных групп.

  - **Разбиение по уникальным значениям полей** – следует выбрать в списке поля исходного набора, задающие уникальный идентификатор группы строк. Количество групп будет равно количеству уникальных идентификаторов.

Параллельная обработка – применяется для ускорения вычислений при работе цикла, количество потоков определяется параметром "максимальное количество потоков". Параллельная обработка не поддерживается циклом с постусловием.

Важно: В некоторых случаях применение параллельной обработки недопустимо. Например, если на каждой итерации происходит обращение к источнику данных, не допускающему параллельные запросы.

Добавлять идентификаторы итераций – флаг добавляет в выходную таблицу поле "Идентификатор итерации", где для каждой строки указан номер итерации, на которой строка была создана.

Игнорировать ошибки – флаг отключает прерывание выполнения цикла при обнаружении ошибок.

Переменная цикла – переменная, которой в ходе работы цикла присваивается номер текущей итерации. Нумерация итераций осуществляется с нуля. Переменная выбирается из списка переменных входных портов узла, заключенного в цикл.

Шаг 3. Сопоставление переменных

Данный этап становится доступен только при заключении в цикл *Подмоделей*, имеющих входные и выходные порты для переменных. На этом этапе настраивается передача значений выходных переменных соответствующим входным переменным на следующей итерации цикла.

Для создания соответствия входных и выходных переменных следует перетащить обозначение входной переменной на обозначение выходной. Данное соответствие графически отобразится линией связи. Такие связи можно удалить кнопкой  . Связывать переменные на этом этапе необязательно.

2.3 Лабораторная работа «Трансформация в АП Loginom»

Рассмотрим сценарий создания подмодели «RFM-анализ» в Loginom.

RFM-анализ является одним из методов базовой сегментации. Он широко применяется в клиентской аналитике. Это техника сегментации клиентов, опирающаяся на их поведение.

Основными являются показатели R (recency) и F (frequency) – давность и частота. Показатель M (monetary) – денежная стоимость клиента, обычно выражающаяся как сумма всех покупок клиента или среднее значение суммы покупок.

В АП Loginom существует возможность проектирования без данных, но лучше использовать небольшой тестовый набор при разработке сценария.

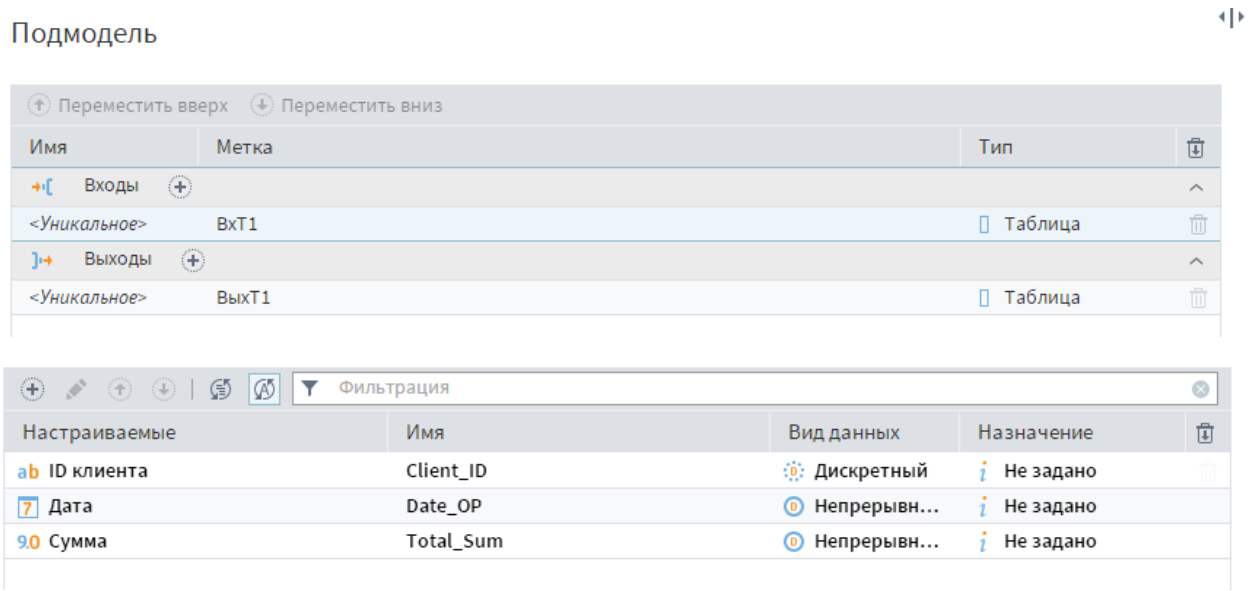
На входе подмодели будет набор данных со следующей структурой:

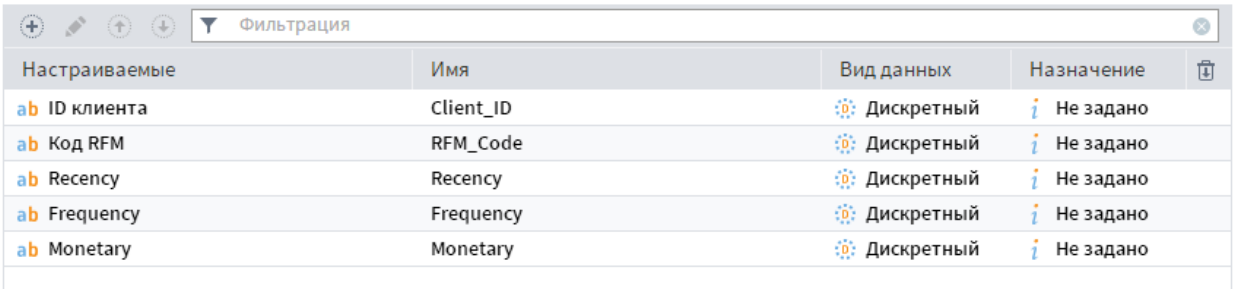
Поле	**Тип**
ID клиента	Строковый
Дата	Дата/Время
Сумма	Вещественный

На выходе должен получиться следующий набор данных:

Поле	**Тип**
ID клиента	Строковый
Recency	Строковый
Frequency	Строковый
Monetary	Строковый
Код RFM	Строковый

Первым делом необходимо добавить на полотно узел «Подмодель», добавить к нему входной и выходной порты, и задать в них описанную выше структуру данных:



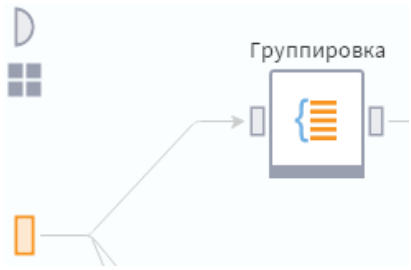


Далее все действия будут выполняться в созданной подмодели.

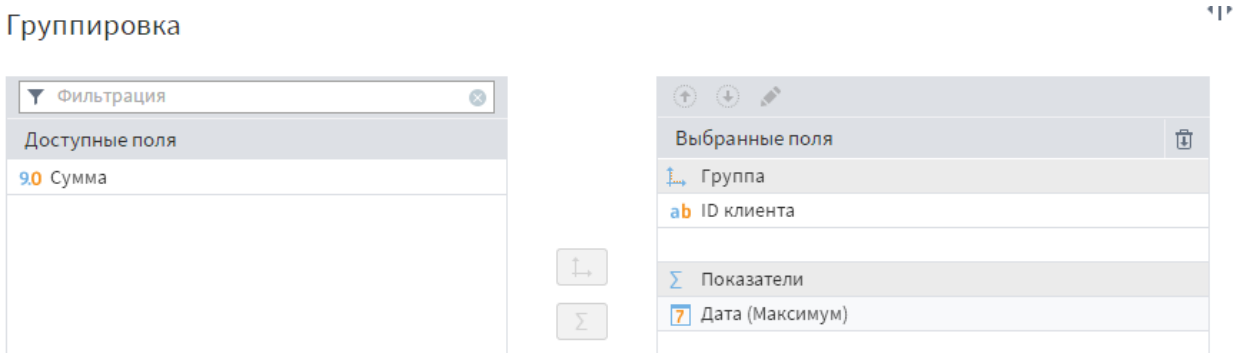
Алгоритм для параметра Recency:

- Для каждого клиента определить дату последней покупки;
- Для каждого клиента рассчитать давность покупки (Recency) как разность между текущей датой (в примере 10.01.2008) и датой последней покупки;
- Разбить полученные данные на 5 групп (квантилей). Каждый клиент при этом получит идентификатор от 1 до 5 в зависимости от его активности. Тем, кто недавно осуществлял покупку, будет присвоен код R=5. Те, кто дольше всех не покупал ничего, получат R=1.

В начале необходимо добавить узел «Группировка», и направить в него данные со входного порта.



Группировка должна проводиться по полю «ID клиента» с агрегацией по дате с типом «Максимум».

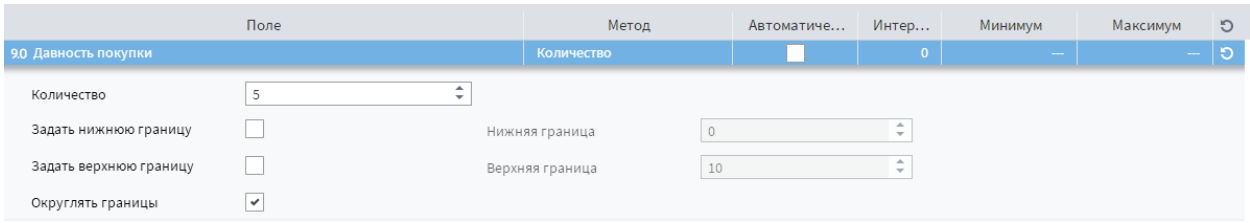


Далее, с помощью калькулятора рассчитывается давность покупки по следующей формуле:

DaysBetween(Date_OP,Today()), где

Date_OP – дата последней покупки.

На следующем шаге необходимо добавить узел «Квантование», с помощью которого и будет производиться расчёт параметра «Recency». Квантование должно проводиться по полю «Давность покупки». Так как данные должны разбиваться на 5 подгрупп, необходимо указать метод «Количество» и задать соответствующее число подгрупп.



В качестве шаблона метки нужно задать %N, что означает, что каждому интервалу присвоится номер от 0 до 4. Далее, с помощью узла «Замена» нужно переопределить номера интервалов, так как у той группы, где давность покупки наименьшая, должен быть код R=5, и наоборот, у группы с наибольшим значением давности покупки должен быть код R=1:

Замена

Поля	Способ замены
9.0 Давность покупки	Не заменять
ab ID клиента	Не заменять
Дата последней покупки	Не заменять
ab Давность покупки Интервальный Идентификатор интервалов	Не заменять
12 Давность покупки Интервальный Номер интервала	Не заменять
ab Давность покупки Интервальный Метка	**Ввод вручную**
9.0 Давность покупки Интервальный Нижняя граница	Не заменять
9.0 Давность покупки Интервальный Верхняя граница	Не заменять
0/1 Давность покупки Интервальный Нижний предел интервала включительно	Не заменять
0/1 Давность покупки Интервальный Верхний предел интервала включитель...	Не заменять

Импорт Экспорт Сортировать ab Изменить тип замены Редактировать Получить значения

№	ab Значение	ab Замена
Точное совпадение		
1	0	5
2	1	4
3	2	3
4	3	2
5	4	1

Алгоритм для параметра Frequency:

- Для каждого клиента определить количество покупок за определённый период;
- Разбить полученные данные на 5 групп (квантилей). Клиентам, совершившим наибольшее число покупок, будет присвоен код F=5, наименее активные покупатели получат F=1.

Узел «Группировка» и настраивается следующим образом:

Выбранные поля
Группа
ab ID клиента
Показатели
9.0 Сумма (Количество)

Узел «Квантование» настраивается по аналогии с параметром Recency, так же, как и узел «Замена». Однако, в случае с заменой необходимо, чтобы у групп с наибольшим числом покупок был код F=5, следовательно, нужно выставить номера по возрастанию от 1 до 5.

Замена

Поля	Способ замены
ID клиента	Не заменять
Количество действий	Не заменять
Количество действий Интервальный Идентификатор интервалов	Не заменять
Количество действий Интервальный Номер интервала	Не заменять
Количество действий Интервальный Метка	**Ввод вручную**
Количество действий Интервальный Нижняя граница	Не заменять
Количество действий Интервальный Верхняя граница	Не заменять
Количество действий Интервальный Нижний предел интервала включит…	Не заменять
Количество действий Интервальный Верхний предел интервала включит…	Не заменять
Количество действий Интервальный Нижняя граница диапазонов открыта	Не заменять

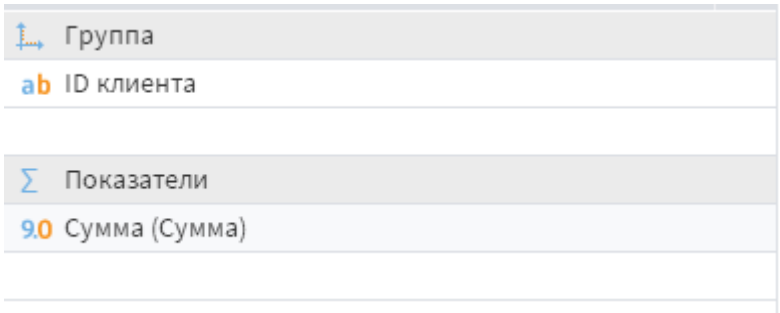
Импорт Экспорт Сортировать Изменить тип замены Редактировать Получить значения

№	Значение	Замена
Точное совпадение		
1	0	1
2	1	2
3	2	3
4	3	4
5	4	5

Алгоритм для параметра Monetary:

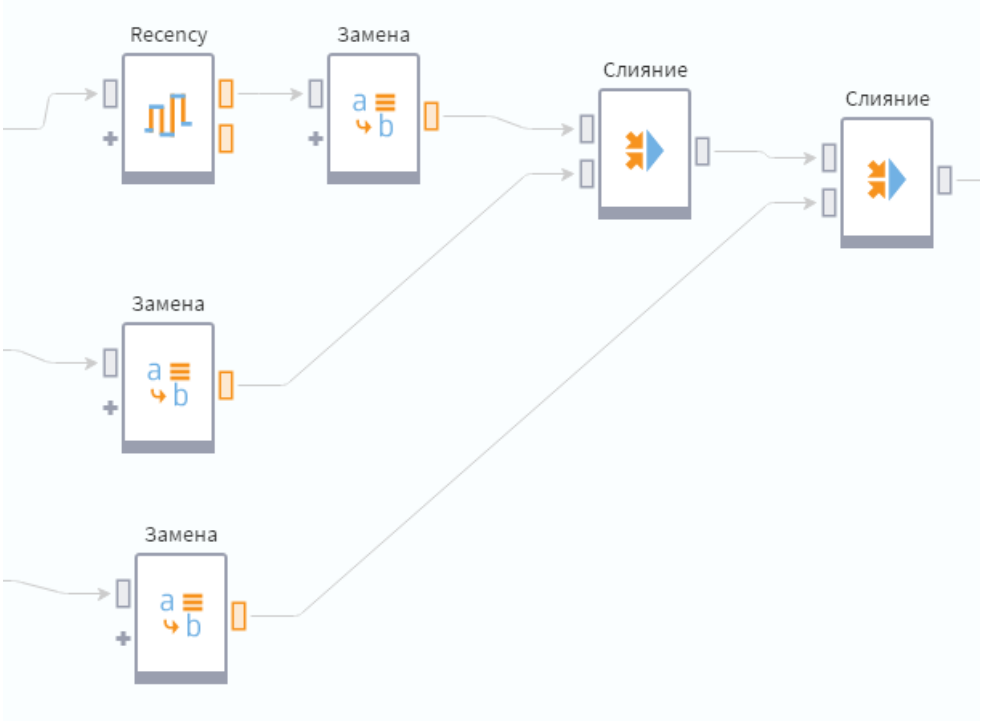
- Для каждого клиента определить сумму потраченных денег;
- Разбить полученные данные на 5 групп (квантилей). Клиентам, потратившим наибольшие суммы, будет присвоен код M=5, клиентам, потратившим наименьшие суммы – M=1.

Последовательность действий аналогична параметру Frequency, с единственным отличием в типе агрегации в узле «Группировка»

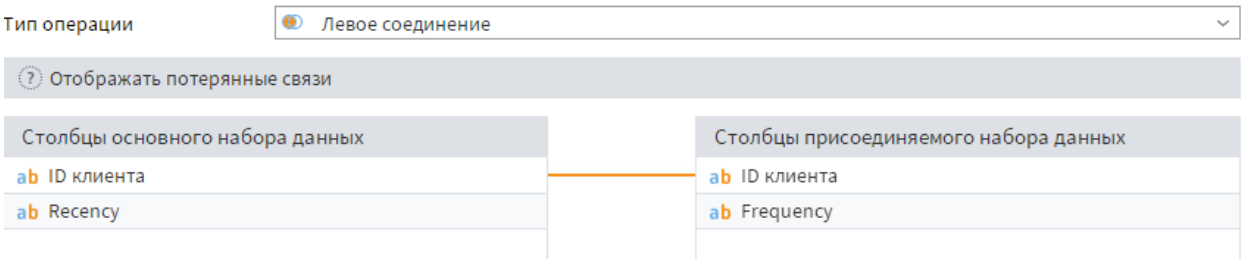


Объединение параметров

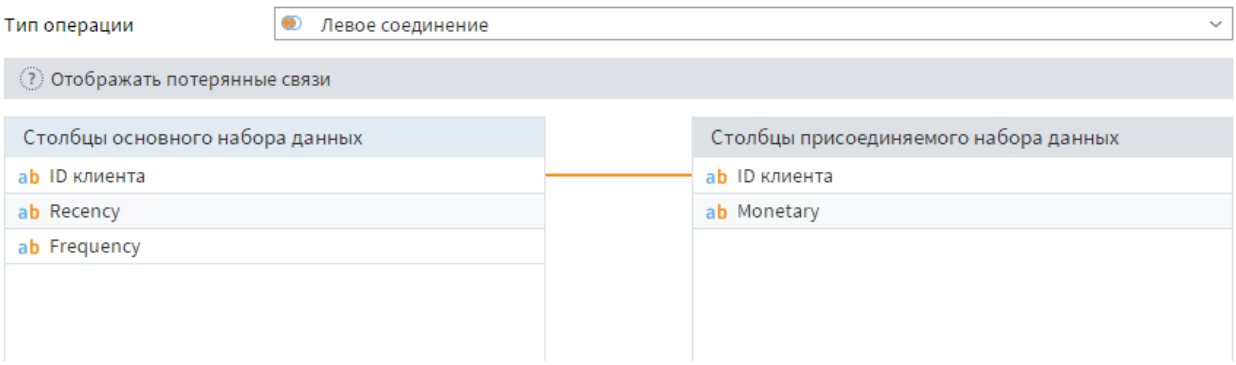
На следующем шаге необходимо объединить все получившиеся параметры в одну таблицу. Для этого используются узлы «Слияние»



В начале необходимо объединить параметры Recency и Frequency. Для этого производится слияние с типом «Левое соединение» и связью по полю «ID клиента». Таким образом, к набору данных с параметром Recency добавится параметр Frequency.

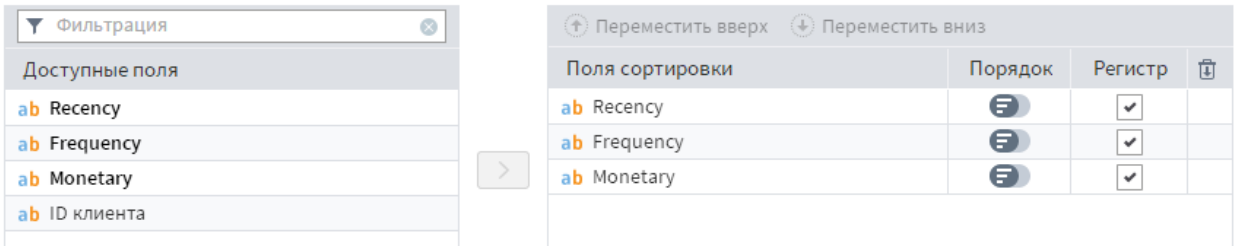


Аналогичным образом к набору данных добавляем параметр Monetary.



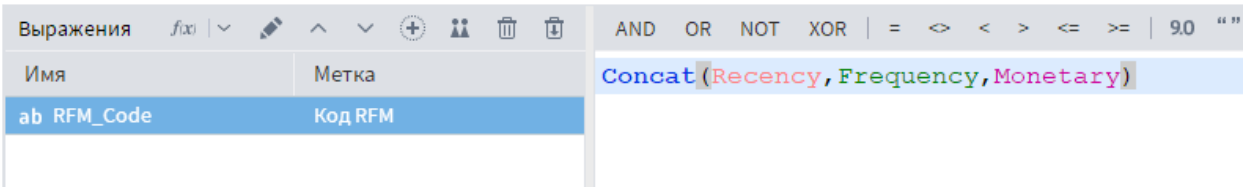
Далее нужно произвести сортировку всех параметров по убыванию.

Сортировка

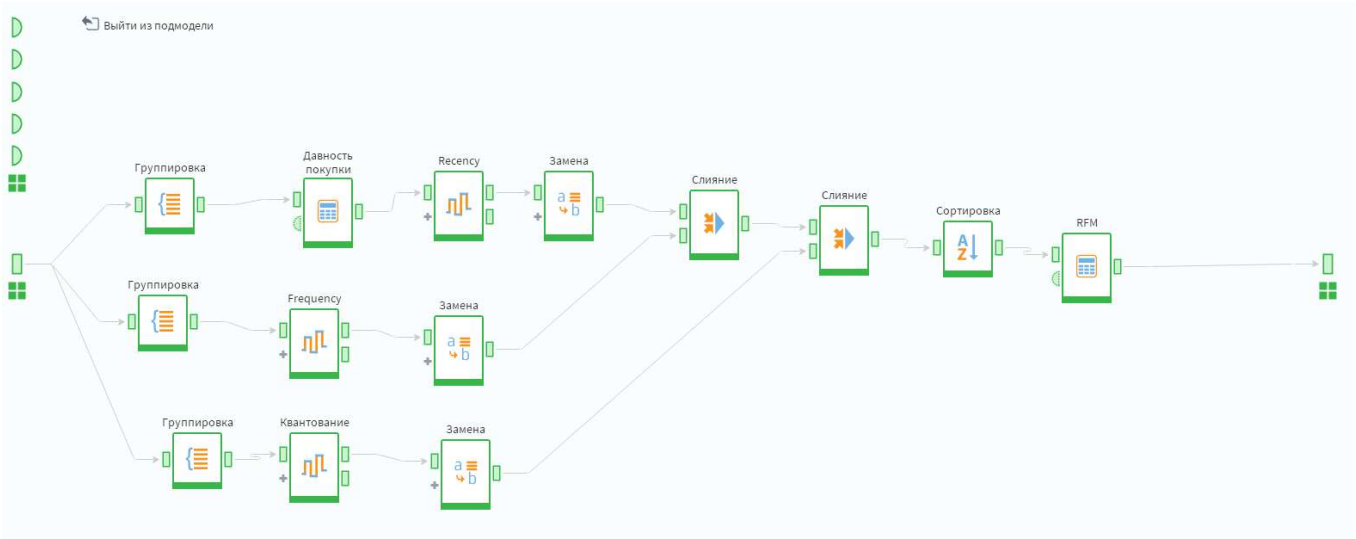


Совместить все параметры в одном столбце можно с помощью калькулятора.

Калькулятор



Общий вид подмодели:



3. Визуализация и аналитические отчеты

Для представления данных используются различные графические средства – отчеты, графики, диаграммы, настраиваемые при помощи параметров.

Общепринятым средством визуализации данных в современных BI-решениях являются информационные (контрольные, приборные) панели (dashboards), на которых результаты отображаются в виде шкал и индикаторов, позволяющих

контролировать текущие значения выбранных показателей, сравнивать их с критическими допустимыми) значениями и таким образом выявлять потенциальные угрозы для бизнеса.

Контрольные панели считаются одним из наиболее удобных способов представления информации о «состоянии здоровья» бизнеса. Они позволяют уместить на экране всю важнейшую информацию о текущих операциях, выявленных и потенциальных проблемах.

Контрольные панели, как и карты показателей (scorecards), основаны на анализе ключевых показателей эффективности (KPIs). Однако, как правило, контрольные панели отображают текущее состояние общих показателей, а карты показателей предназначены для сравнения текущих показателей с плановыми, целевыми, и отображают динамику изменения этих показателей во времени. Карты показателей обычно бывают более персонализированными, настраиваются в зависимости от ролей и задач конкретного пользователя (финансовое управление, снабжение, продажи и т.п.). При необходимости все эти показатели могут быть детализированы при помощи дополнительных отчетов, графиков и диаграмм.

В информационную среду бизнес-аналитики поступает первичный материал – данные, которые затем перерабатываются в автоматизированных системах и информационных продуктах.

В процессе переработки происходит переход от данных к информации. Хранилище данных извлекает данные из множества транзакционных или оперативных систем, а затем интегрирует и хранит данные в специализированной БД. Например, в хранилище данных могут приводиться в соответствие и объединяться пользовательские записи из четырех оперативных систем (приложений для обработки заказов, обслуживания, продаж и поставок). Такой процесс извлечения и интеграции преобразует данные в новый информационный продукт – информацию.

Затем пользователи, работающие с аналитическими инструментами (например, для создания запросов, отчетов, OLAP-анализа и выполнения операций интеллектуального анализа данных), обращаются к данным из хранилища данных и анализируют ее. Таким образом, определяются тенденции, структуры и исключения. Аналитические инструменты помогают пользователям преобразовать информацию в знания.

В BI-приложения часто встроены BI-инструменты (OLAP, генераторы запросов и отчетов, средства моделирования, статистического анализа, визуализации и data mining). Многие BI-приложения извлекают данные из ERP-приложений.

BI-приложения обычно ориентированы на конкретную функцию организации или задачу, такие как анализ и прогноз продаж, финансовое бюджетирование, прогнозирование, анализ рисков, анализ тенденций, «churn analysis» в телекоммуникациях и т.п. Они могут применяться и более широко как в случае приложений управления эффективностью предприятия (enterprise perfomance management) или системы сбалансированных показателей (balanced scorecard).

Для всестороннего анализа данных в современных BI используются OLAP-инструменты (online analytical processing). Они позволяют рассматривать различные срезы данных, в том числе временные, позволяющие выявлять различные тренды и зависимости (по регионам, продуктам, клиентам и т.п.).

Средства OLAP позволяют исследовать данные по различным измерениям. Пользователи могут выбрать, какие показатели анализировать, какие измерения и как отображать в кросс-таблице, обменять строки и столбцы «pivoting», затем сделать срезы и вырезки («slice&dice»), чтобы сконцентрироваться на определенной комбинации размерностей.

Можно изменять детальность данных, двигаясь по уровням с помощью детализации и укрупнения «drill down/ roll up», а также кросс-детализации «drill across» через другие измерения.

Отчет представляет собой документ, содержимое которого динамически формируется на основе информации, содержащейся в базе данных.

Генераторы запросов и отчетов – типично «настольные» инструменты, предоставляющие пользователям доступ к базам данных, выполняющие некоторый анализ и формирующие отчеты.

На рынке ПО сейчас представлено немало средств создания отчетов: как отдельных продуктов, так и входящих в состав средств разработки приложений СУБД, и реализованных в виде либо серверных служб, либо клиентских приложений.

Задача тиражирования знаний заключается в предоставлении возможности сотрудникам, не разбирающимся в методиках анализа и способа получения того или иного результата, получать ответ на аналитические запросы на основе моделей, подготовленных экспертом. Для эксперта предназначена панель сценариев, в которой он строит различные модели. Для конечного же пользователя предназначена панель отчетов.

Основные методы визуализации:

- **Табличные и графические.** Как правило, таблицы применяются в том случае, когда пользователю необходимо работать с отдельными значениями данных, вносить изменения, контролировать форматы данных, пропуски, противоречия и т. д. Графические методы позволяют лучше увидеть общий характер данных – закономерности, тенденции, периодические изменения. Кроме того, графические методы более эффективно сопоставляют данные: достаточно построить графики двух исследуемых процессов на одной системе координат, чтобы оценить степень их сходства и различия.
- **Одномерные и многомерные.** Одномерные визуализаторы представляют информацию только об одном измерении данных, в то время как многомерные – о двух или более. Если график показывает зависимость инфляции от даты, то он будет одномерным, поскольку на нем будет отображаться

только одно измерение – «Дата», значениям которого будет соответствовать факт «Значение показателя инфляция». Если же информация об инфляции приводится по датам и регионам, то появляется еще одно измерение – «Название региона», и тогда для корректного представления данных используется многомерный визуализатор. Популярные многомерные визуализаторы: OLAP-куб, многомерная диаграмма, карта Кохонена и др.

- **Общего назначения и специализированные.** Методы визуализации общего назначения не связаны с каким-либо определенным видом задач анализа или типом данных и могут использоваться на любом этапе аналитического процесса. Это своего рода типовые визуализаторы: графики и диаграммы, графы, гистограммы и их разновидности, статистические характеристики и др. В то же время существует ряд задач, специфика которых требует применения специализированных визуализаторов. Например, карты Кохонена специально разработаны для визуализации результатов кластеризации, матрицы классификации используются в основном для проверки состоятельности классификационных моделей, а с помощью диаграмм рассеяния оценивается корректность работы регрессионных моделей.

При изучении различных видов визуализации удобнее рассматривать их не по отдельности, а в контексте задач, для которых они наиболее часто применяются.

Можно выделить следующие группы методов визуализации:

1. Визуализаторы общего назначения – применяются для решения типовых задач анализа данных – визуальной оценки качества и характера данных, распределения значений признаков, статистических характеристик и т.д.

2. OLAP-анализ – комплекс методов для визуализации многомерных данных. Популярные многомерные визуализаторы: OLAP-куб, кросс-диаграмма, карты Кохонена и др.

3. Визуализаторы для оценки качества аналитических моделей позволяют оценивать различные характеристики, такие как точность, эффективность, достоверность результатов, интерпретируемость, устойчивость, корректность (регрессионных, классификационных, прогностических и т.д.) моделей, построенных в процессе анализа данных.

4. Визуализаторы для интерпретации результатов анализа служат для представления конечных результатов анализа в виде, наиболее удобном с точки зрения их интерпретации пользователем

Каждый из алгоритмов Data Mining использует определенный подход к визуализации:

1. Для деревьев решений это визуализатор дерева решений, список правил, таблица сопряженности.

2. Для нейронных сетей в зависимости от инструмента это может быть топология сети, график изменения величины ошибки, демонстрирующий процесс обучения.

3. Для карт Кохонена: карты входов, выходов, специфические карты.

4. Для линейной регрессии в качестве визуализатора выступает линия регрессии.

5. Для кластеризации: дендрограммы, диаграммы рассеивания.

Диаграмма рассеяния представляет собой график, по одной оси которого откладываются целевые значения выходной переменной (т.е. которые заданы в качестве эталона обучения), а по другой – реальные значения, полученные на выходе. Смысл диаграммы рассеяния состоит в следующем: если все или хотя бы основная масса точек, представляющих реальные выходные значения модели, сосредоточены вблизи линии идеальных значений, то модель работает хорошо.

Таблица сопряженности позволяет наиболее наглядно оценить результаты классификации, полученные с помощью той или иной модели. Она показывает результаты сравнения категориальных значений выходного поля исходной

(обучающей) выборки и категориальных значений выходного поля, рассчитанных с помощью модели.

Все эти способы визуального представления или отображения данных могут выполнять одну из функций:

- являются иллюстрацией построения модели (например, представление структуры (графа) нейронной сети);
- помогают интерпретировать полученный результат;
- являются средством оценки качества построенной модели;
- сочетают перечисленные выше функции.

Подсистемы визуализации данных содержатся не только в специализированных аналитических платформах, но и практически во всех программных средствах, которые связаны с обработкой данных, – от офисных приложений до систем компьютерной математики. Однако в аналитических платформах визуализации данных уделяется особое внимание, поскольку она является одной из составляющих аналитического процесса, без которой невозможно эффективно решить поставленные задачи.

Даже если для построения качественной модели данных недостаточно, визуализация позволяет выдвигать гипотезы, делать выводы на основе экспертных оценок, разрабатывать способы повышения информативности данных.

3.1 Визуализаторы Таблица и Диаграмма в АП Loginom

Визуализаторы представляют собой инструмент, позволяющие пользователю выбрать удобный вариант отображения данных.

В АП Loginom предусмотрены следующие способы визуализации данных:

- Диаграмма – графическое представление данных.
- Куб – многомерное представление данных.

- Таблица – табличное представление данных.
- Статистика – статистические показатели полей набора данных.
- Конечные классы – результаты процедуры оптимального квантования в виде начальных и конечных классов, а также WoE-диаграммы и значений информационных индексов IV.
- Отчет по регрессии – статистические параметры и результаты статистических тестов для анализа регрессионных моделей.
- Качество бинарной классификации – формирует наборы серий данных для построения диаграмм, определяются оптимальные пороги отсечения и вычисляются оценки классификации. Для получения точек серий строятся гистограммы распределения событий и не-событий в выборках.

Работа с визуализаторами

Для добавления визуализатора к узлу сценария требуется нажать кнопку *Настройка визуализаторов*. В открывшемся окне слева находится дерево доступных визуализаторов, справа расположен список выходных портов узла, данные которых можно визуализировать.

Для добавления визуализатора, надо выбрать в дереве необходимый визуализатор и нажать кнопку *Добавить визуализатор* у нужного выходного порта. Также это можно сделать, перетащив мышкой необходимый визуализатор в область кнопки *Добавить визуализатор* у нужного выходного порта.

Для удаления визуализатора необходимо нажать кнопку в правом верхнем углу визуализатора.

Важно: Данные не каждого порта возможно отобразить выбранным визуализатором. Если какой-либо визуализатор не поддерживается портом, то он добавлен не будет.

Таблица

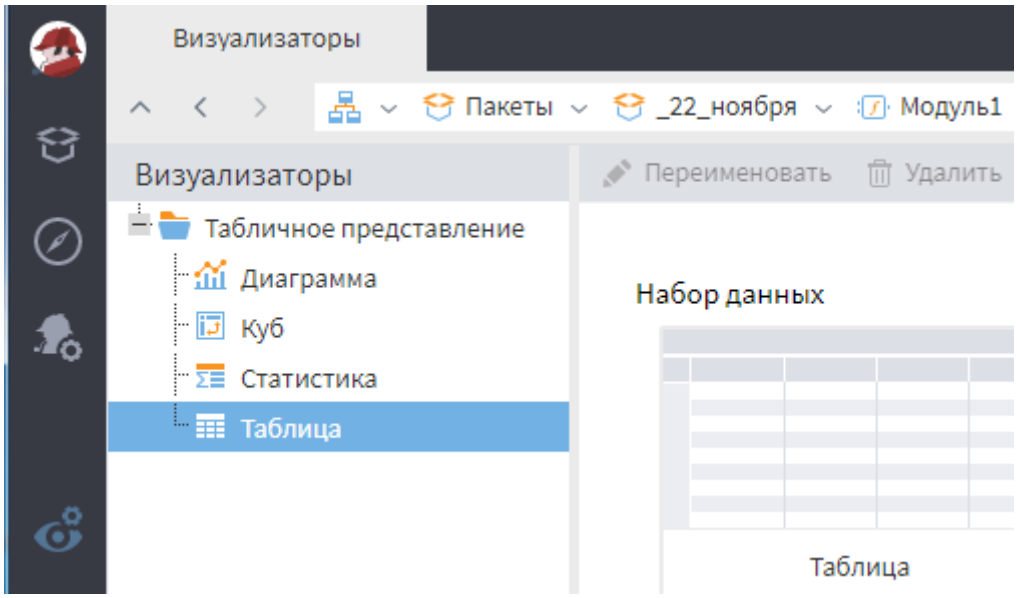


Рис. 47. Визуализатор «Таблица»

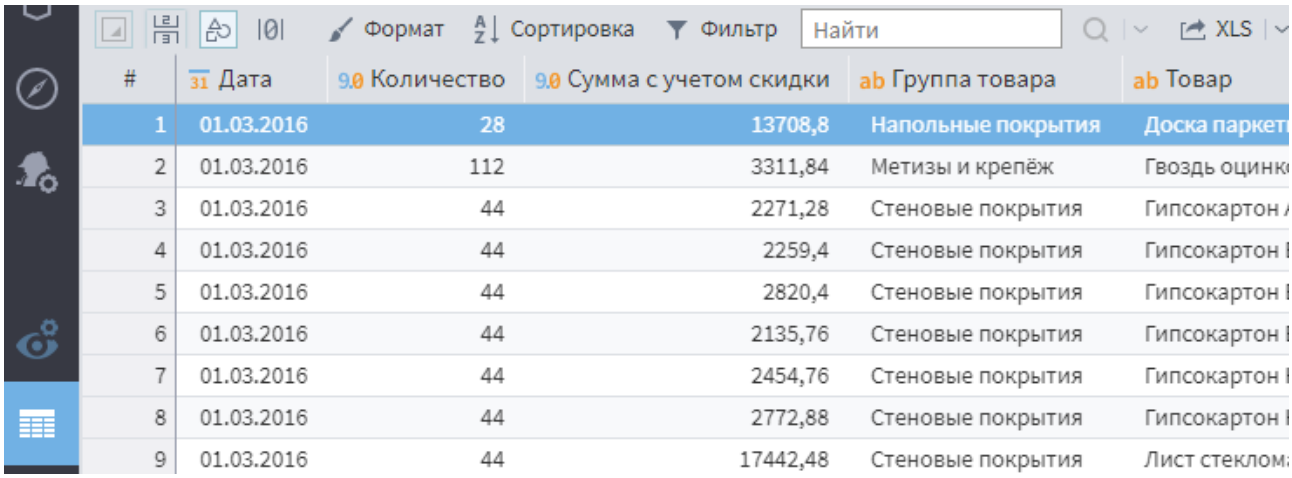


Рис. 48. Операции визуализатора «Таблица»

Таблица позволяет представить данные для пользователя в виде плоской двумерной таблицы с возможностью фильтрации, сортировки, изменения формата отображения данных и поиска.

Операции:

- **Номер строки** – показать/скрыть сквозной номер строки;
- **Типы данных** – показать/скрыть типы данных;
- **Показать значения null** – показать/скрыть null-значения;
- **Формат** — открыть окно настройки формата отображения данных;
- **Сортировка** – открыть окно настройки сортировки данных по столбцам;

- **Фильтр** – открыть окно настройки фильтрации;
- Найти **Поиск** – найти значение; открыть окно настройки поиска;
- Детализация **Детализация** – показать/скрыть детализацию по строке таблицы;
- **Перейти к строке** – перейти к строке с заданным номером.

Диаграмма.

Диаграмма – один из наиболее активно используемых визуализаторов.

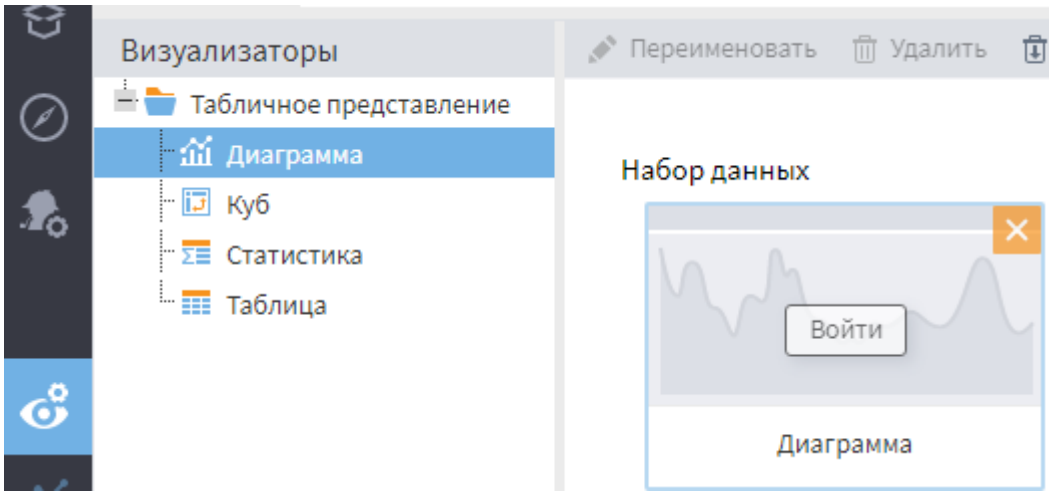


Рис. 49. Визуализатор «Диаграмма»

Диаграмма визуально отображает зависимость значений одного поля от другого. Наиболее часто используемый вид диаграмм – двумерный график. По горизонтальной его оси откладываются значения независимого столбца, а по вертикальной – соответствующие им значения зависимого столбца.

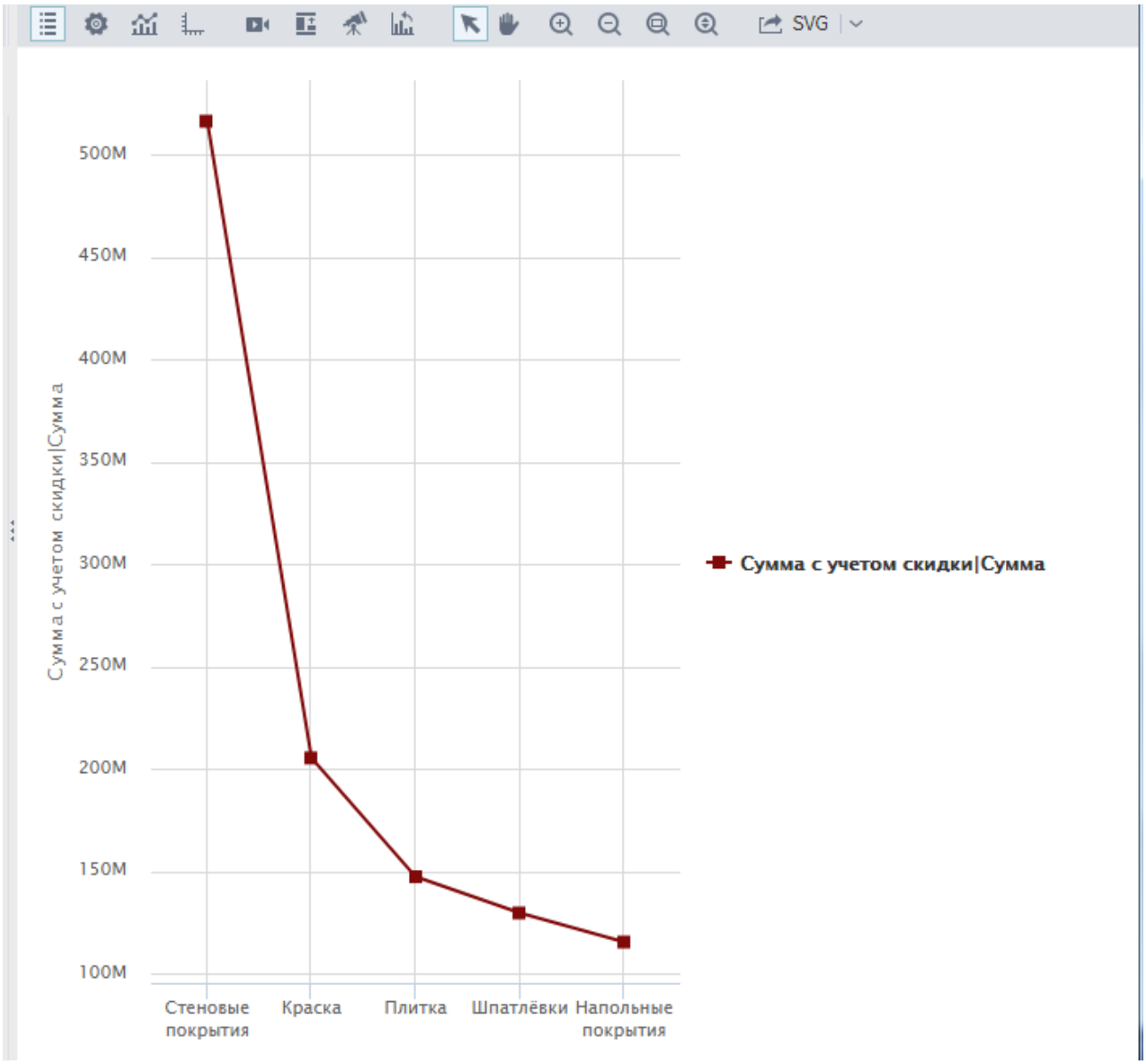


Рис. 50. Пример визуализатора «линейная диаграмма»

Над областью построения расположена панель инструментов, где можно изменить любые настройки диаграммы:

рис. 51. Панель инструментов визуализатора Диаграмма

Все настройки можно осуществить с помощью контекстного меню, дублирующего панель инструментов (вызов через щелчок правой кнопкой мыши в области построения диаграммы).

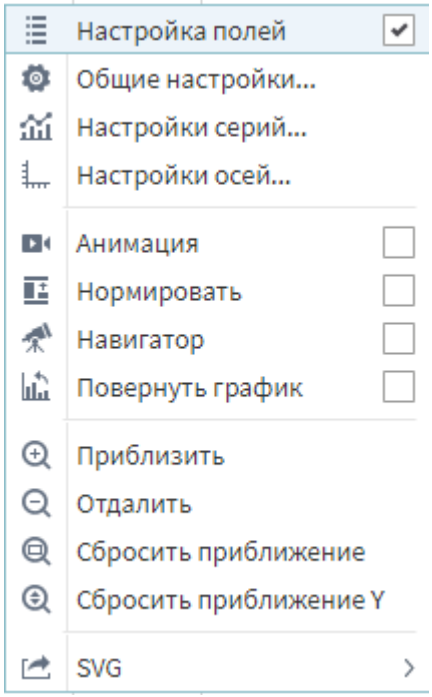


Рис. 52. Контекстное меню «Настройки полей диаграммы»

1. Настройка полей дает возможность показать/скрыть список полей набора данных.
2. Общие настройки отвечают за внешний вид диаграммы и области построения.
3. Окно «Настройки серий» серий имеет две вкладки: в основной можно выбрать тип линий, а дополнительные позволяют изменить цвет.

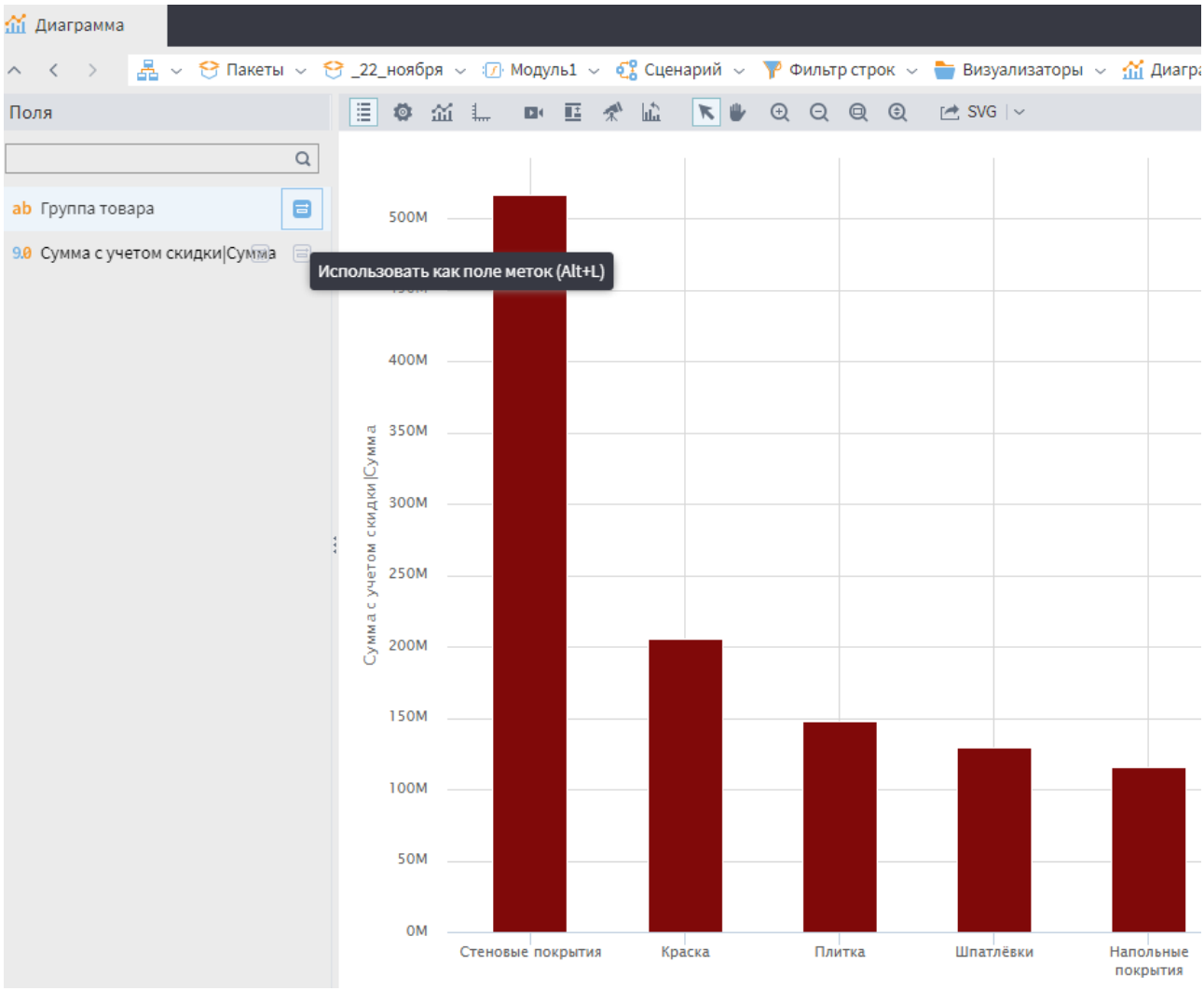


Рис. 53. Пример визуализатора «столбчатая диаграмма»

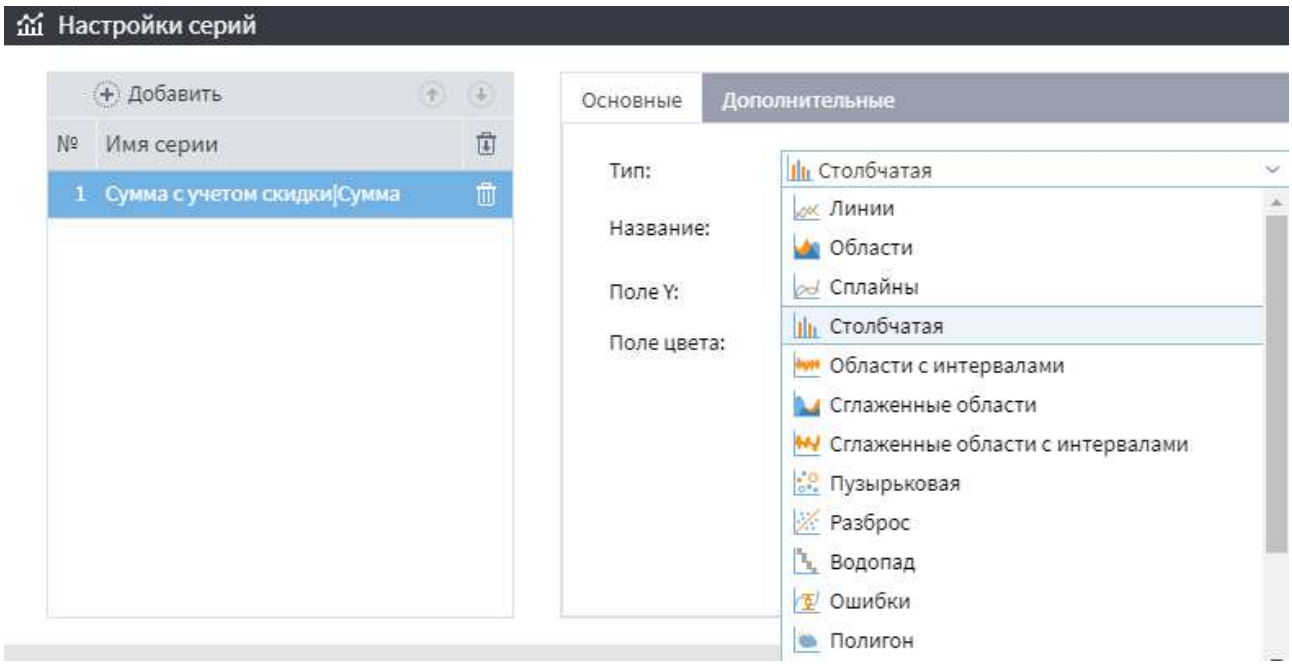


Рис. 54. Вкладка «Настройки серий»

4. Настройки осей позволяют задать параметры соответствующим осям диаграммы: отображение сетки, тип оси, заголовок, задать минимальное или максимальное значения по осям, выравнивание.

Настройки нижней оси позволяют задать поле оси Охи поле меток, а настройки левой/правой осей – отображение динамики по оси по значению или в процентах.

5. Анимация позволяет перемещение, приближение и отдаление диаграммы сделать более плавным, обновление данных отображается плавно.
6. Нормировать – приводит все графики к одному масштабу.
7. Навигатор позволяет детализировать по оси X какой-либо участок диаграммы, отображается снизу под осью. Передвигая край навигатора, можно выбрать необходимую область диаграммы.
8. Повернуть график – меняет местами оси, то есть поле по X перемещается на ось Y, и наоборот ось Y становится осью X.
9. Приблизить – приближает/увеличивает область диаграммы.
10. Отдалить – отдаляет/уменьшает область диаграммы.
11. Сбросить приближение – снимает все настройки приближения (по обеим осям).
12. Сбросить приближение Y – снимает приближение диаграммы по оси Y, но оставляет приближение по оси X.
13. Экспорт – позволяет сохранить диаграмму, как она выглядит в окне просмотра в файл с изображением. В данном пункте меню будет отображаться название того формата, который выбран для экспорта.

3.2 Визуализатор OLAP-куб в АП Loginom

Рассмотрим OLAP-кубы – визуализаторы, которые чаще всего используются в отчетах.

Куб является одним из распространенных методов комплексного многомерного анализа данных, получивших название OLAP (OnLine Analytical Processing). В его основе лежит представление данных в виде многомерных кубов, называемых также OLAP-кубами или гиперкубами.

Куб – это удобное средство визуализации многомерных данных и получения необходимых форм отчетов. Он содержит измерения и факты, определенные при построении. К основной особенности куба относится то, что его структура не является жестко определенной. Манипулируя с помощью мыши заголовками измерений, пользователь может добиться, чтобы куб выглядел наиболее информативно.

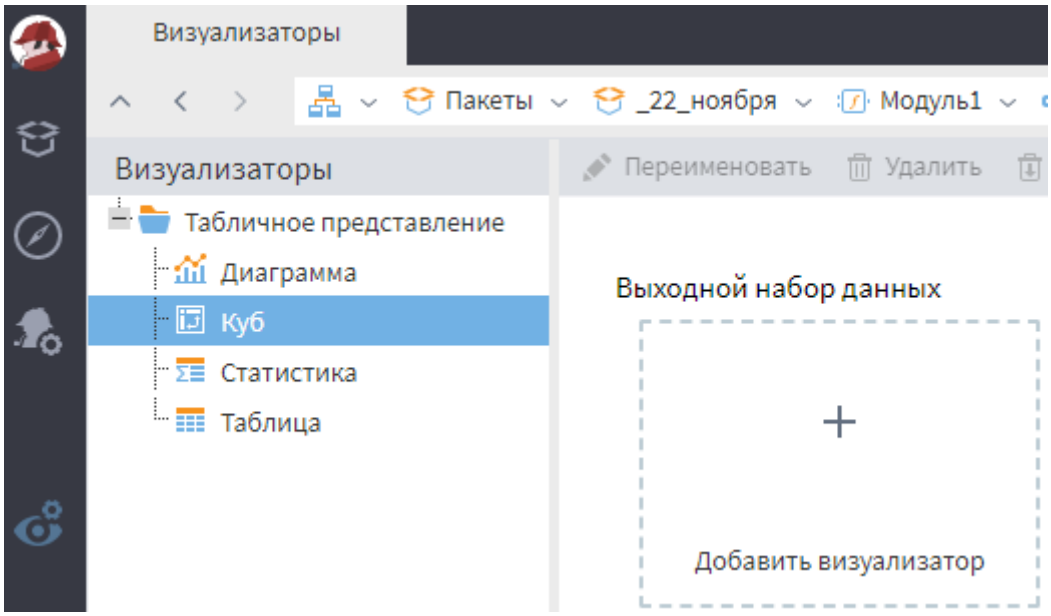


Рис. 55. Визуализатор «Куб»

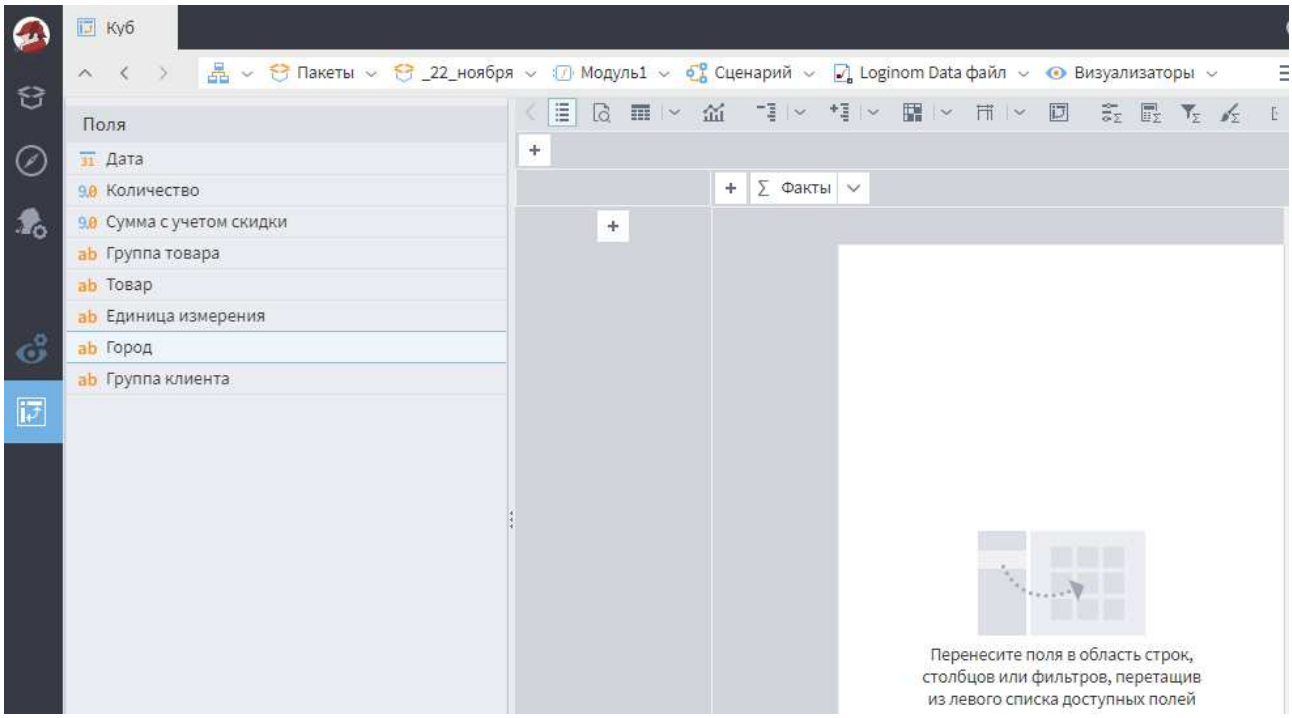


Рис. 56. Визуализатор «Куб»: слева – Область списка полей, справа – Область построения куба

Над областью построения куба расположена панель инструментов, где можно изменить любые настройки:

Рис. 57. Панель инструментов визуализатора Куб

Все настройки можно осуществить с помощью контекстного меню, дублирующего панель инструментов (вызов через щелчок правой кнопкой мыши в области построения куба).

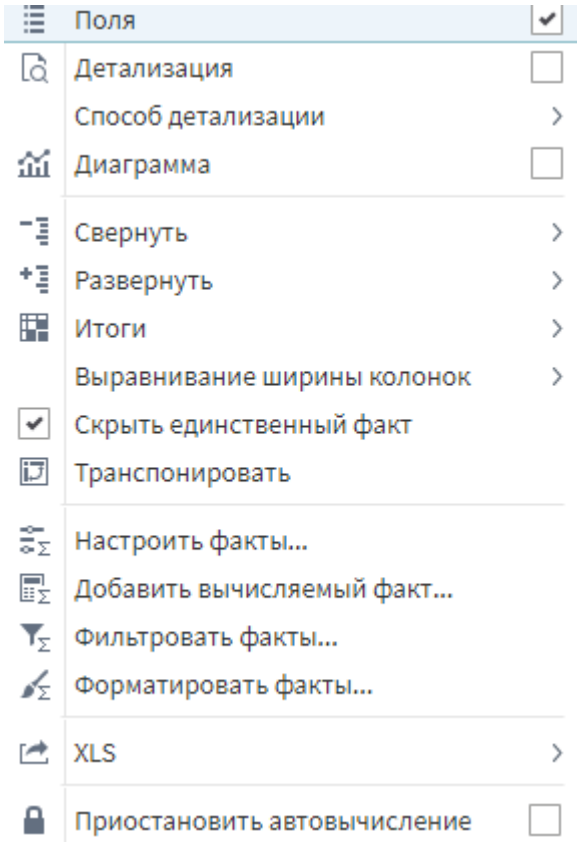


Рис. 58. Контекстное меню «Настройки полей куба»

Для построения информативного куба нужно перенести поля в область списков измерений. Количество измерений может варьироваться от нуля до количества доступных полей. Уже используемые измерения можно вернуть обратно в список доступных полей. Рекомендуется использовать не более 5-7 измерений, чтобы отчет был понятным и интерпретируемым.

Добавить измерения в строки или столбцы куба можно двумя способами:

- Перетащить (Drag-and-Drop) поле из левого списка в ту область, в которую его необходимо добавить: строки или столбцы;
- Нажать кнопку [+] нужной области и выбрать необходимое поле из списка.

Так как визуализатор Куб представляет собой плоскую двухмерную таблицу, то для отображения нескольких измерений на одной оси используется иерархическая система. Это позволяет детализировать содержание конкрет-

ного элемента измерения. Развернуть такой элемент можно кнопкой , находящейся в названии данного элемента, а свернуть – кнопкой . Также, свернуть или развернуть до нужного уровня все элементы можно из меню конкретного измерения и выбрав *Свернуть* или *Развернуть*.

Чтобы отсортировать значения измерения, необходимо нажать кнопку рядом с нужным измерением и выбрать необходимый вариант: по возрастанию, по убыванию, в исходном порядке.

В кубе можно настроить фильтры. Фильтрацию можно производить двумя способами:

- по значениям фактов
- по значениям измерений путем непосредственного выбора значений из списка.

Чтобы отфильтровать данные по измерению необходимо щёлкнуть по нужному измерению, выбрать значения из списка уникальных и нажать "Применить". При этом доступны операции: *Выбрать все, Отменить выбор, Инвертировать выбор*.

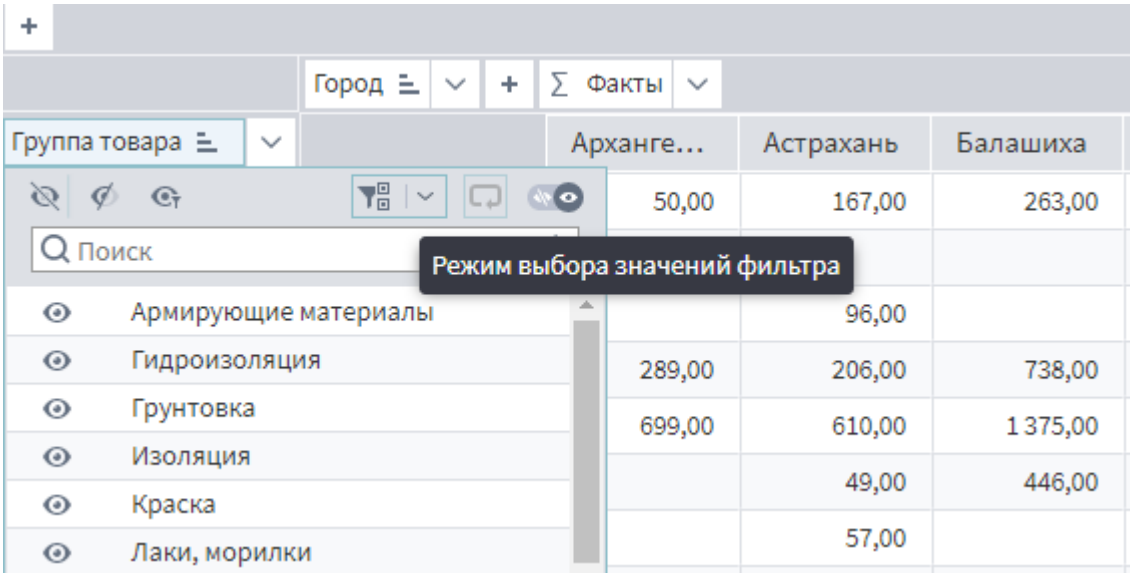


Рис. 59. Фильтрация в кубе

Удалить измерение можно, нажав кнопку Удалить рядом с нужным измерением и выбрав *Удалить*.

Добавить факты можно двумя способами:

- Перетащить (Drag-and-Drop) поле из левого списка в область фактов куба (область пересечения строк и столбцов);
- Нажать кнопку Σ Факты, выбрать необходимое поле и в появившемся окне выбрать нужный вариант агрегации и способ его отображения.

При добавлении фактов необходимо выбрать функцию агрегации (сумма, количество, среднее, минимум, максимум, первый, последний и др.)

Нажав на кнопку рядом с Σ Факты можно открыть одно из окон (рис. 60) :

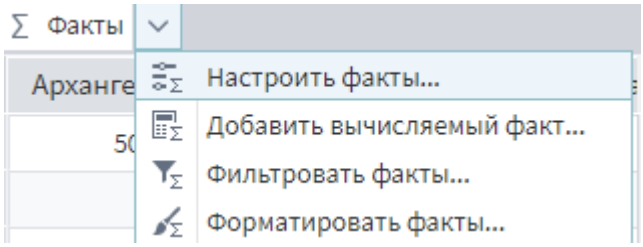


Рис. 60. Параметры настройки фактов в кубе

Удалить факт можно в окне *Настроить факты*.

Для задания фильтрации по значениям фактов необходимо нажать кнопку *Фильтровать факты* или выбрать аналогичный пункт в раскрывающемся меню рядом с кнопкой *Факты*.

Вычисляемый факт позволяет добавить свой факт в куб. Вычисляется на основе используемых в кубе и области фильтрации измерений и фактов.

Необходимо задать:

- **Имя** – строка, которая будет служить идентификатором в процедурах обработки;
- **Метка** – строка, под которой факт будет виден в кубе и диаграмме;
- **Тип данных** – тип данных вычисляемого выражения;
- **Название** – значок для вычисляемого выражения.

Область кода выражения

В области кода задается формула расчета выражения. Ссылки на измерения, факты и синтаксические конструкции функций можно вставлять в код выражения, выбрав их двойным кликом мыши в соответствующих областях либо написав вручную.

Формула выражения может содержать:

- Ссылки на используемые измерения и факты куба в виде наименования измерений и фактов;
- Варианты агрегации измерений и фактов;
- Скобки, определяющие порядок выполнения операций;
- Знаки математических операций и отношений;
- Логические операции (and, or, not, xor) и значения (true или 1, false или 0);
- Функции в соответствии с синтаксическим описанием (см. далее "Список функций");
- Строковые выражения в кавычках ("строковое выражение");
- Целые и вещественные числа;
- Однострочные и многострочные комментарии.

Однострочный комментарий начинается символами // (два слеша) и продолжается до конца строки. Многострочным комментарием считаются все символы, содержащиеся между /* (слеш-звездочка) и */ (звездочка-слеш).

Список измерений и фактов – область содержит список измерений и фактов, которые можно использовать в вычисляемых фактах.

Двойной клик мыши по позиции дерева измерений и фактов вводит имя измерения или факта с выбранной функцией агрегации в область кода выражения. То же самое можно сделать, написав в область кода вручную.

Примечание: в списке будут доступны только используемые факты и измерения. Измерения, которые не были перенесены в строки, столбцы или область фильтрации, в списке доступны не будут.

Список функций содержит наименование, входные аргументы и описание доступных для использования функций.

Возможна фильтрация по категории и названию функции.

Двойной клик мыши по позиции выбранной функции вставляет ее синтаксис в область кода выражения. То же самое можно сделать, перетащив функцию из списка в область кода (Drag-and-Drop).

3.2. Отчеты в АП Loginom

Для каждого визуализатора может быть добавлен отчет. Для добавления отчета к визуализатору необходимо выбрать нужный визуализатор и нажать кнопку *Добавить в отчеты*, создав при необходимости нужную группу или разместить отчет в существующую группу.

Для создания отчета необходимо добавить визуализатор, на основе которого будет построен отчет, выбрать данный визуализатор и нажать кнопку *Добавить в отчеты*. В этом случае отчет добавится в список "Без группы". Для добавления отчета в определенную группу необходимо выбрать визуализатор и открыть выпадающий список рядом с меню *Добавить в отчеты*. На выбор будет предложено поместить отчет в заранее созданную группу или создать новую группу.

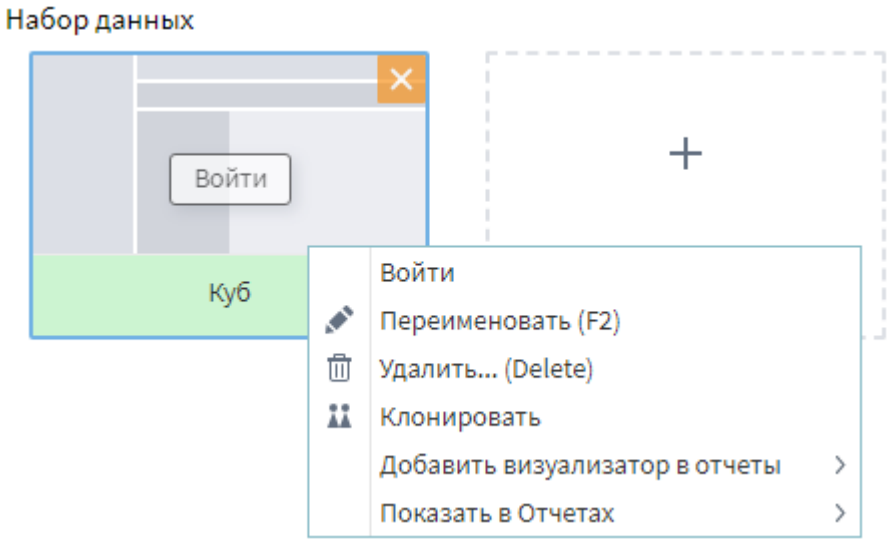


Рис. 61. Добавление визуализатора в отчеты

Так же добавить отчет можно нажав правой кнопкой мыши на необходимом визуализаторе и выбрать пункт меню "Добавить визуализатор в отчеты". Отчет можно добавить без группы, выбрать уже существующую группу или создать новую группу.

Отчеты желательно группировать в группы по их смысловому содержанию. Например, группа «Аналитические отчеты» может содержать различные кубы данных, группа «Прогнозы» может содержать диаграммы прогнозов каких-либо величин. Тогда конечный пользователь открывает панель отчетов, выбирает нужную группу и в этой группе активизирует нужный отчет. После такого выбора программа автоматически выполняет сценарий, соответствующий этому отчету, и выдает результат в зависимости от настроенного отображения отчета.

На практике часто встречаются ситуации, когда пользователю требуется получить отчет по некоторому подмножеству всех доступных данных, например, только по одному поставщику или клиенту, по нескольким группам товаров или регионам. В терминах многомерной модели данных такое подмножество называется срезом. Аналитик может создавать отчеты в предопределенных, наиболее востребованных разрезах, но не в силах предсказать все виды отчетов, которые могут потребоваться пользователю. Для решения данной задачи пользователь может самостоятельно настраивать необходимый ему вид отчетов. Сохранить данную настройку пользователь не может.

Для поиска отчета, созданного от текущего визуализатора необходимо щелкнуть на визуализаторе правой кнопкой мыши и выбрать пункт меню "Показать в отчетах". При этом произойдет переход на панель управления отчетами.

Для управления созданными отчетами необходимо перейти в группу *Отчеты* пакета. Панель управления отчетами предназначена для создания новых групп отчетов, переименования и удаления существующих групп отчетов, ре-

дактирования, переименования и удаления отчетов. При редактировании отчетов меняется так же исходный визуализатор, на основании которого построен отчет.

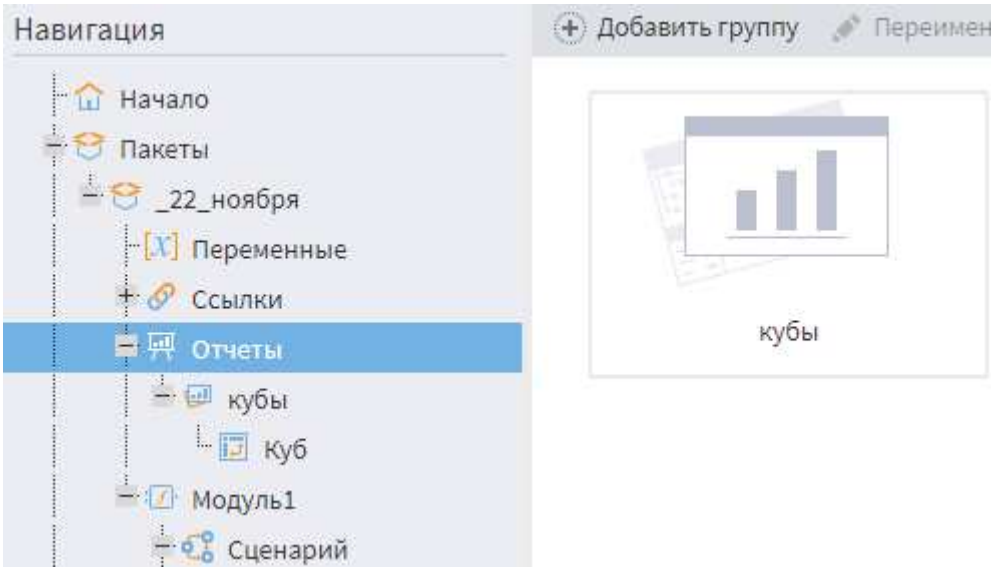


Рис. 62. Панель управления отчетами

Отчеты, которые находятся в открытом состоянии, подсвечиваются зеленым цветом. Такие отчеты можно обновить, нажав кнопку *Перечитать данные*. В этом случае будут деактивированы все узлы, находящиеся перед узлом с отчетом и активированы снова. Примечание: при этом также будут деактивированы все другие отчеты, которые находятся на деактивируемых узлах.

3.2. Лабораторная работа «Визуализаторы и отчеты в АП Loginom»

Задание 1. Создать отчеты по результатам анализа данных рынка недвижимости

1. Настройте визуализатор «Диаграмма»

На рис. 63 можно увидеть динамику средних цен на квартиры по месяцам, или же таблицу средних цен в разрезе количества комнат в квартире. По графику можно сказать, что относительно 2016 года, в 2017 и 2018 годах цены на квартиры снизились.

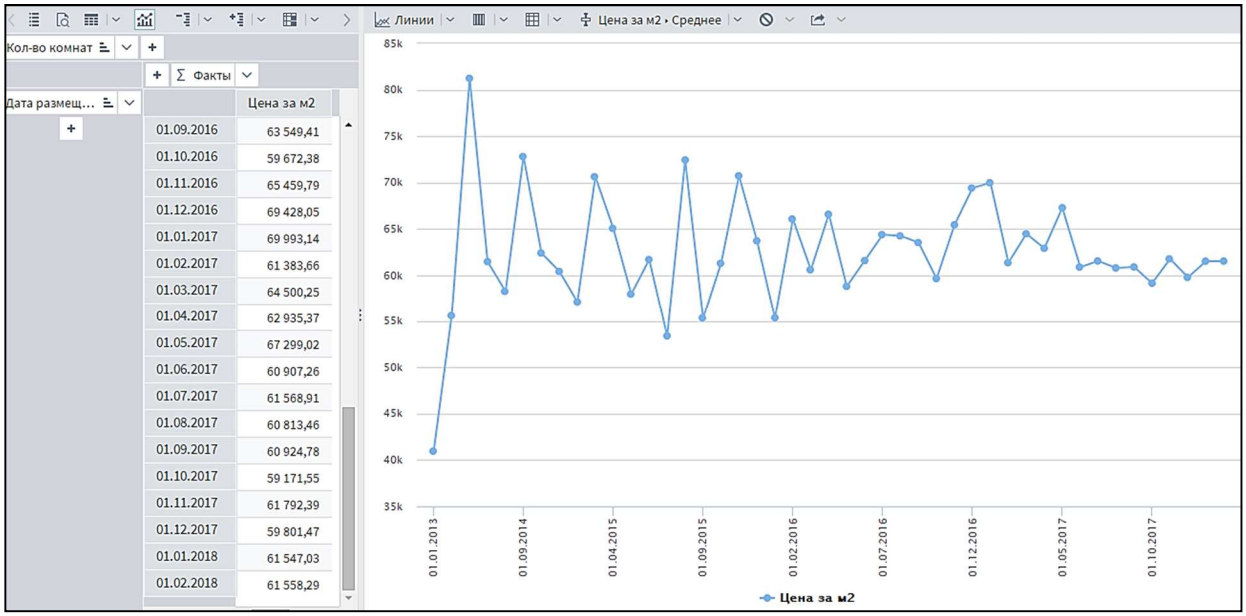


Рис. 63 «Динамика средних цен»

Кол-во комнат | Факты

Дата размещ...	1	2	3	4	5	6	7+	Итого:
01.09.2016	68 788,49	63 310,73	55 758,88	65 590,19		63 694,27		63 549,41
01.10.2016	62 581,83	58 349,36	58 817,88	58 092,33	56 701,03			59 672,38
01.11.2016	65 638,52	62 854,93	70 539,82	55 898,92	58 695,65			65 459,79
01.12.2016	62 962,80	70 740,15	72 579,72	65 744,13	58 095,24			69 428,05
01.01.2017	79 707,65	67 547,37	63 755,17	66 995,08				69 993,14
01.02.2017	67 543,82	56 713,69	58 514,70	61 544,14	66 350,71	88 750,00		61 383,66
01.03.2017	66 432,86	63 716,89	63 763,04	63 498,11	59 507,22			64 500,25
01.04.2017	65 708,61	62 110,59	60 461,64	64 418,18	79 686,95	52 173,91		62 935,37
01.05.2017	71 342,78	66 746,60	62 679,09	71 290,90	66 258,92			67 299,02
01.06.2017	62 898,16	61 478,59	58 427,14	59 158,25		92 024,54		60 907,26
01.07.2017	64 520,27	60 205,36	60 122,26	60 760,24	56 936,81			61 568,91
01.08.2017	65 227,55	59 366,84	56 670,83	60 141,21	100 977,20			60 813,46
01.09.2017	64 006,14	59 960,47	57 929,31	66 511,69	45 616,59			60 924,78
01.10.2017	61 658,42	58 687,96	57 649,34	54 407,70	63 767,20			59 171,55
01.11.2017	65 327,52	60 682,21	58 938,17	62 531,64	42 395,10			61 792,39
01.12.2017	63 158,70	58 932,04	56 974,84	57 374,11	68 100,82			59 801,47
01.01.2018	64 539,73	60 504,60	59 420,84	58 578,46	63 130,69	70 000,00	46 190,48	61 547,03
01.02.2018	64 754,40	60 378,42	59 382,04	57 856,14	59 336,98	67 302,30		61 558,29

Рис. 64 «Средние цены по видам квартир»

Можно наглядно увидеть динамику цен в зависимости от расположения объекта недвижимости. Сравнить цены во всех районах города, или рассмотреть каждый район отдельно. На графике видно, что самый высокий уровень цен в Нижегородском районе. Это объясняется тем, что он является центром города.

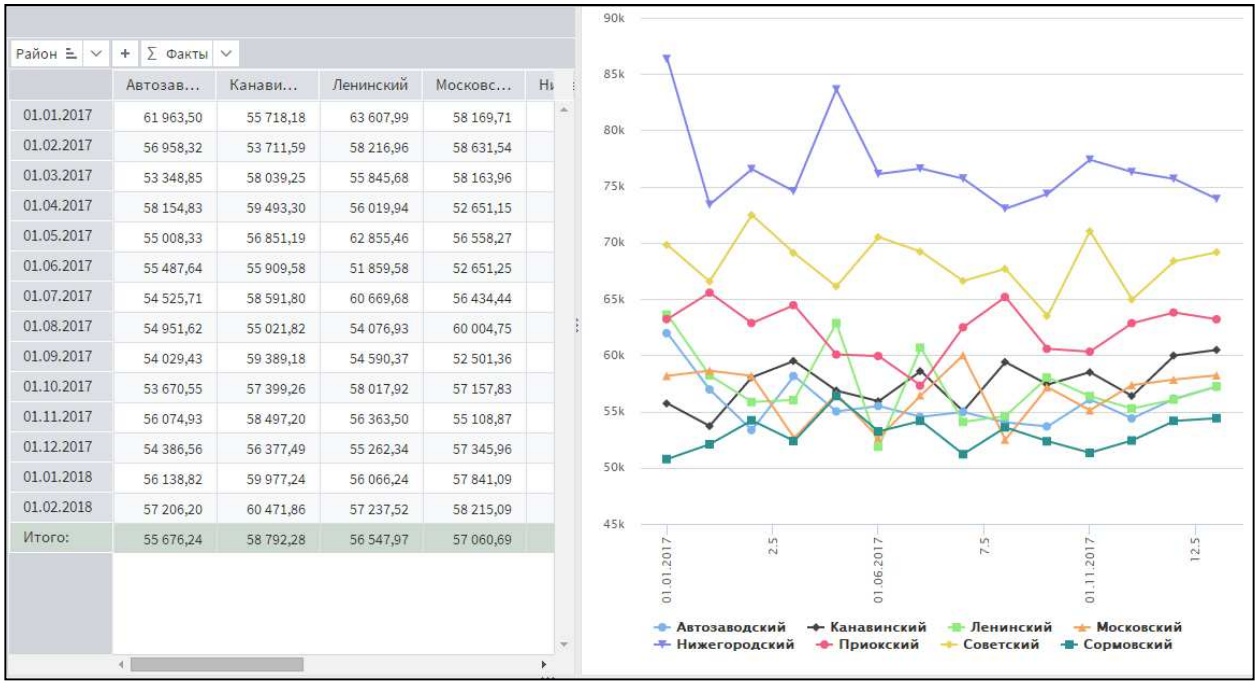


Рис. 65 «Динамика цен по районам»

Так же динамику средних цен можно рассмотреть не по отдельным периодам, а в разрезе лет, чтобы увидеть изменение общего уровня цен из года в год.

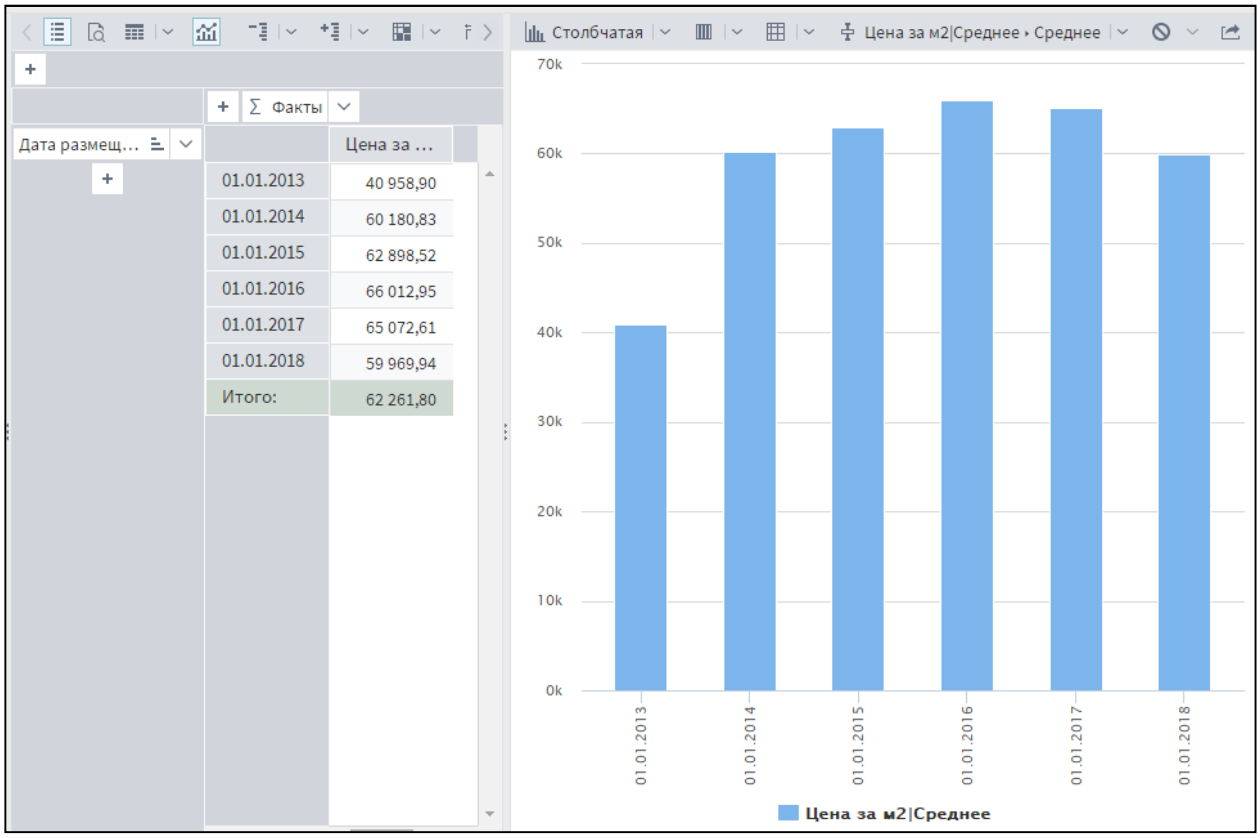


Рис. 66 «Динамика средних цен по годам»

Дата размещ…	Кол-во комнат	1	2	3	4	5	6	7+	Итого:
	01.01.2013				40 958,90				40 958,90
	01.01.2014	59 536,67	63 027,74	65 527,32	52 631,58				240 723,31
	01.01.2015	64 461,39	64 294,57	61 881,69	60 956,43				251 594,08
	01.01.2016	65 210,90	62 604,09	64 257,27	61 588,18	57 464,86	68 724,04	82 241,32	462 090,67
	01.01.2017	64 745,43	60 281,19	58 984,91	61 099,79	64 899,70	80 424,61		390 435,63
	01.01.2018	64 629,55	60 456,57	59 403,88	58 275,69	62 182,26	68 651,15	46 190,48	419 789,57
	Итого:	318 583,94	310 664,16	310 055,06	335 510,57	184 546,83	217 799,80	128 431,80	1 805 592…

Рис. 67 «Средние цены по видам квартир»

Можно так же увидеть статистику по продажам квартир по годам и видам квартир.

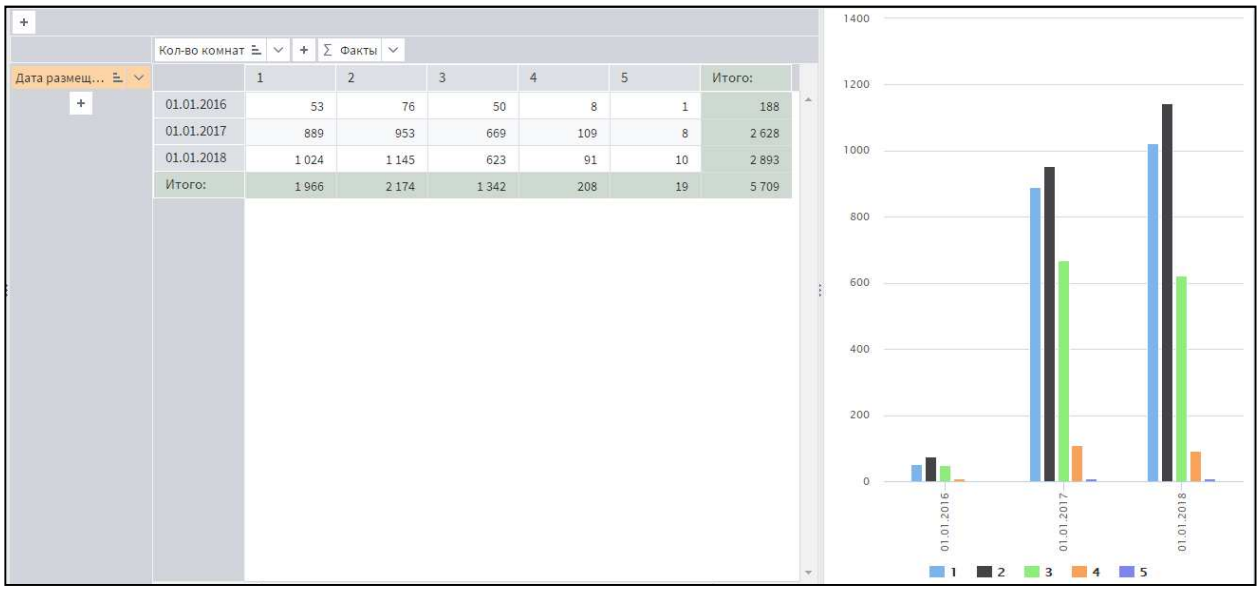


Рис. 68 «Количество проданных квартир по годам»

Постройте куб (и кросс-диаграмму)

Визуальные отчеты

Представление информации

Отчетность – возможность создания форматированных и интерактивных отчетов, с развитыми механизмами для их распространения и обновления.

Весь перечень визуализаторов был помещён в раздел пакета – Отчёты. Этот раздел необходим для мгновенного ознакомления с предложенными результатами работы компонентов без необходимости смотреть на их содержимое.

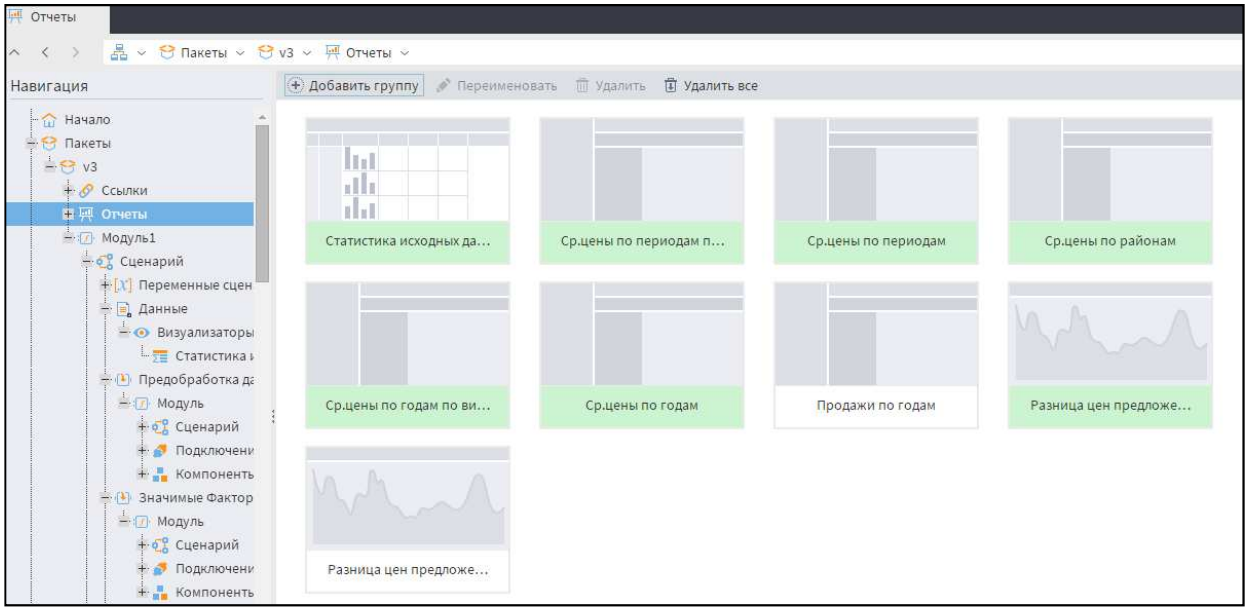


Рис. 69 «Раздел пакета – Отчеты»

Задание 2. Требуется разработать систему аналитической отчетности в АП Loginom. При помощи операций транспонирование измерений, фильтрация и агрегирование фактов сформировать отчеты и ответить на вопросы в заданиях.

Используемые обозначения основных социально-экономических показателей:

- Показатель *промышленность* – объем произведенной продукции (работ, услуг) в действующих ценах соответствующих лет по крупным и средним предприятиям, млн руб.;
- Показатель *с/х* – продукция сельского хозяйства в хозяйствах всех категорий (в фактически действовавших ценах), тыс.руб.;
- Показатель *инвестиции* – инвестиции в основной капитал крупных и средних предприятий, тыс.руб.;
- Показатель *зарплата* – среднемесячная заработная плата работающих на крупных и средних предприятиях, руб.;
- Показатель *безработица* – уровень официально зарегистрированной безработицы (в % от экономически активного населения);
- Показатель *доход* – доходы бюджета, млн руб.;

- Показатель *расход* – расходы бюджета, млн руб.

1. Построить OLAP-кубы и соответствующие им кросс-диаграммы для показателя *промышленность*, в которых будут указаны 5 лучших районов по объему произведенной продукции, 5 худших и районы, дающие 50% от общего объема произведенной продукции в Нижегородском регионе. Последовательность выполнения задания:
    1) Загрузить данные.
    2) Настроить визуализаторы в виде куба.
    3) Произвести настройку назначений полей куба: измерения – дата (размещаем в строки), код_региона-название (размещаем в колонки), код региона – информационный, остальные поля – факты (с агрегацией – сумма).
    4) В полученной кросс-таблице осуществить фильтрацию, выбрав факт промышленность, измерение код_региона-название и
        а) условие «Первые N», значение «5»;
        б) условие «Последние N», значение «5»;
        в) условие «Доля от общего», значение «50».
2. Сформировать многомерные отчеты и соответствующие им кросс-диаграммы для показателя *с/х*, в которых будут указаны 5 лучших районов по уровню с/х, 5 худших.
3. Определить первые 5 районов по инвестициям и 5 последних.
4. Определить районы с самым высоким и самым низким уровнем безработицы.
5. Определить районы с самым высоким и самым низким уровнем зарплаты.
6. Определить 5 самых доходных районов, 5 районов с самым низким уровнем доходов и районы, дающие 80% дохода бюджета всего Нижегородского региона.
7. Определить районы с наибольшим и наименьшим расходом бюджета.

3.5. Вопросы для самопроверки

1. Каковы цели и задачи визуализации данных в аналитических технологиях?
2. Зачем нужен визуальный контроль данных после их загрузки в аналитическое приложение?
3. Чем отличаются средства визуализации общего назначения от специализированных?
4. Какие средства визуализации данных относят к визуализаторам общего назначения и почему?
5. Какие виды диаграмм вам известны?
6. По какому принципу строится гистограмма, и какую информацию о поведении исследуемой величины из нее можно извлечь?
7. Для каких целей служит визуализатор *Статистика*?
8. Почему визуализация многомерных данных может оказаться проблематичной с точки зрения пользователя?
9. В чем заключается OLAP-анализ и каковы его цели?
10. Какова структура OLAP-куба?
11. Какие манипуляции с измерениями можно производить, чтобы сделать представление куба более информативным?
12. В чем заключаются операции транспонирования и детализации, каковы их цели?
13. Что такое кросс-диаграмма, и для каких целей она используется?

Глава 4 Разработка библиотеки компонентов на примере задачи оценки недвижимости

Структура *Проекта* может быть представлена в иерархическом виде:

- Проект может состоять из связанных между собой *Пакетов* – это возможно благодаря тому, что *каждый пакет может предоставлять свои объекты другим пакетам* через механизм ссылок.

- Пакет включает в себя *Модули* – декомпозиция пакета на уровне модулей.
- **Модуль** – сам по себе не содержит узлов обработки данных, но предоставляет отдельное пространство для Сценариев и Подключений к различным источникам данных.
- **Сценарий** – содержит последовательность узлов обработки данных. Сценарий может:
    - Включать в себя подпрограммы – *Подмодель*.
    - Получать данные от узлов из других сценариев и пакетов через механизм Узел-ссылка.
    - Использовать настройки и обученные модели узлов из других сценариев и пакетов через механизм *Выполнение узла*.
    - Использовать готовые алгоритмы обработки данных, созданные в других сценариях и пакетах через механизм *Производные компоненты*.

Подмодель – включает в себя другие узлы, предоставляя, таким образом, отдельное пространство для реализации произвольного алгоритма обработки данных. Подмодель в сценарии представлена в качестве узла, имеющего заданные пользователем входные и выходные порты. Может содержать в себе иерархию вложенных подмоделей. На базе подмодели может быть создан *Производный компонент*.

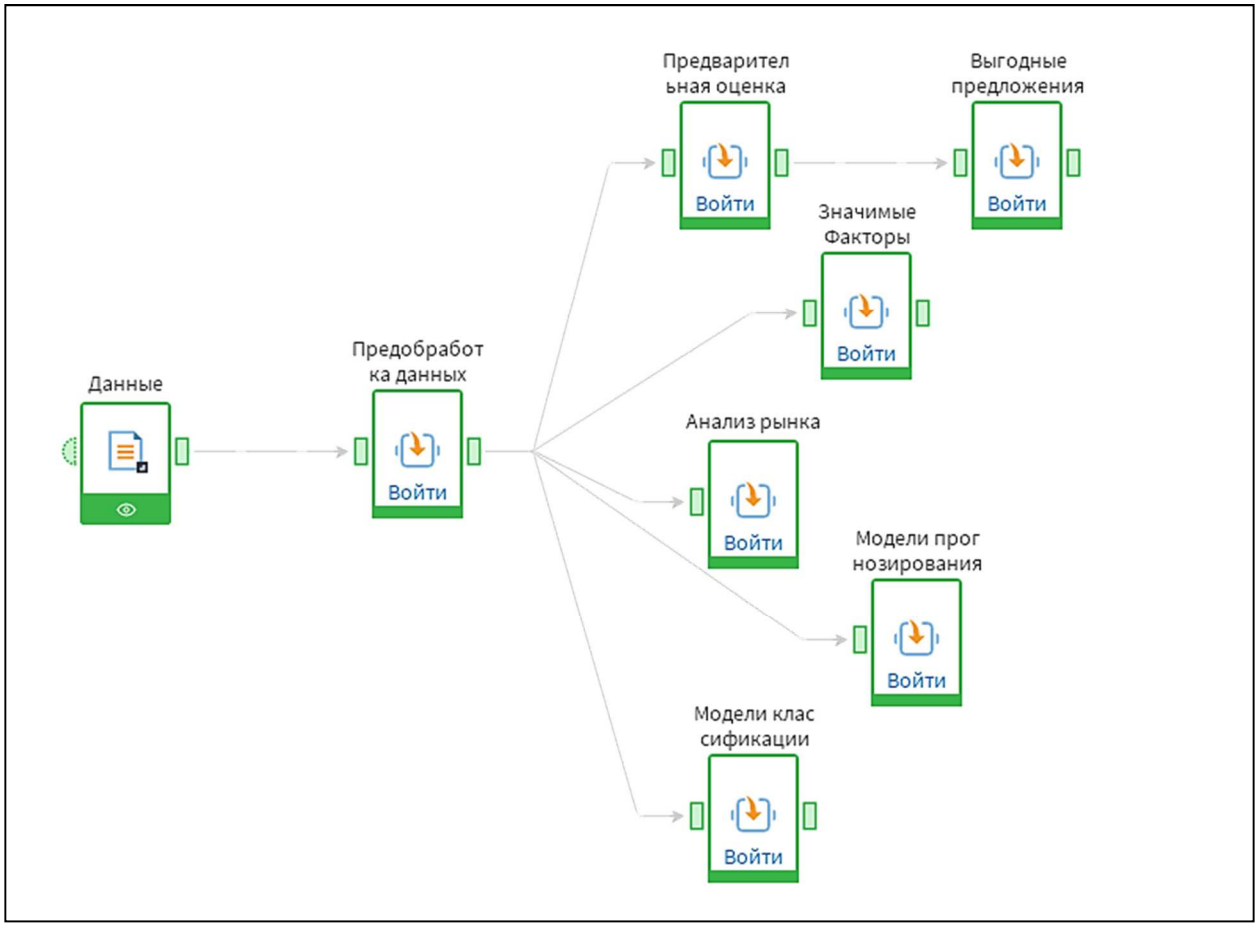


Рис. 70 «Готовая библиотека компонентов»

Подмодель «Предобработка данных» была описана в Лабораторной работе на стр. 43.

Дополнительно благодаря проведению корреляционного анализа можно получить информацию о том, какие факторы сильнее всего влияют на цену объекта недвижимости.

Обработчик Корреляционный анализ на основании вычисляемых коэффициентов корреляции выявляет взаимосвязь между рядами данных входного набора. Применяется для оценки предполагаемой зависимости факторов.

Проведем корреляционный анализ для данных об объектах недвижимости г. Нижнего Новгорода.

На вход компонента (рис. 71) подаются все ценообразующие факторы: район, число комнат, площадь, год постройки, высота потолков, материал стен,

этаж, балкон, тип санузла, состояние помещения, парковка, выходом является цена за м^2.

В таблице необходимо выбрать ряды для анализа взаимосвязи. Для каждого поля из "Набора 1", отмеченного флагом, будут вычислены коэффициенты корреляции с полями, отмеченными флагами в "Наборе 2".

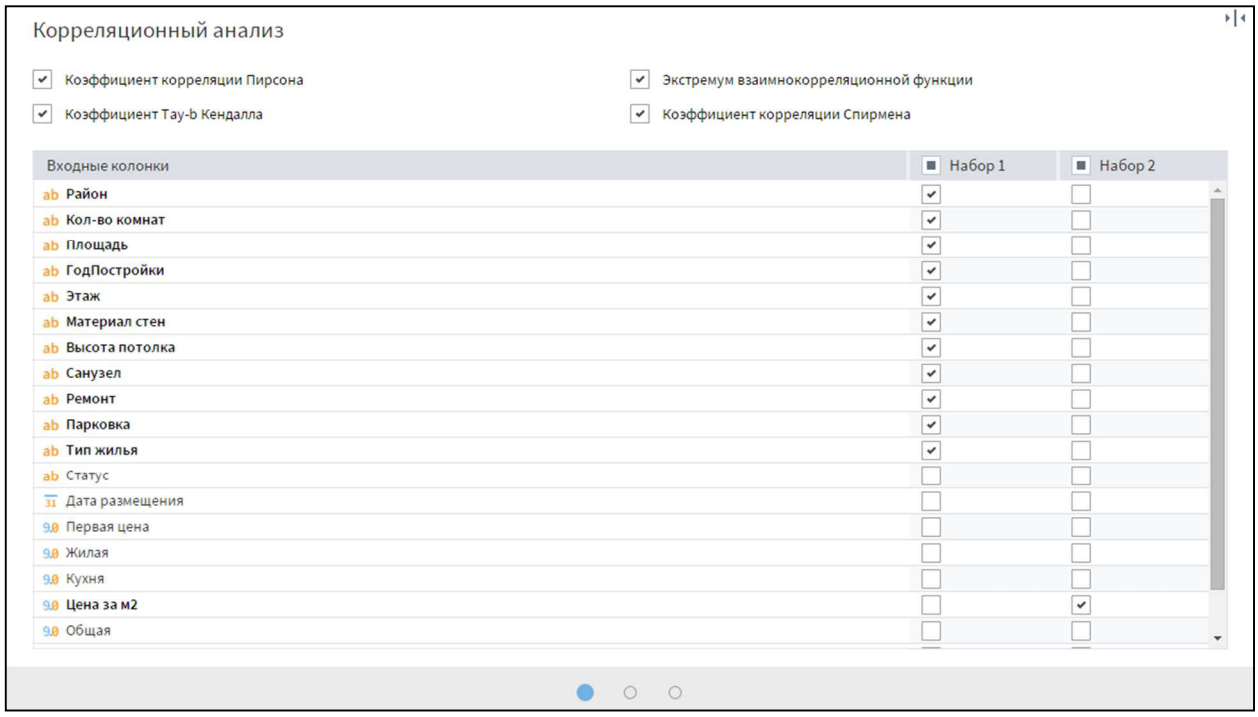


Рис. 71 «Настройка параметров узла Корреляционный анализ»

Выходной набор данных – таблица, содержащая данные о корреляции между полями. Имеет следующую структуру:

Обязательные поля:

- **Поля|Имя** – имя первого поля в корреляционной паре;
- **Поля|Метка** – метка первого поля в корреляционной паре;
- **Поля|Имя** – имя второго поля в корреляционной паре;
- **Поля|Метка** – метка второго поля в корреляционной паре.

Поля, наличие которых задается пользователем:

- **Пирсона** – коэффициенты корреляции Пирсона;
- **Экстремум К. Ф.** –экстремумы взаимно корреляционной функции;
- **Tay-b Кендалла** – коэффициенты ранговой корреляции Кендалла;
- **Спирмена** – коэффициенты корреляции Спирмена.

Корреляционный анализ-1 • Выходной набор данных • Быстрый просмотр данных

#	Поле1.Имя	Поле1.Метка	Поле2.Имя	Поле2.М...	Пирсона	Экстремум К.Ф.	Тау-b Кендалла	Спирмена
1	COL1	Район	Expr5_Label	Цена за м2 ...	0,28	0,28	0,46	0,28
2	COL3	Кол-во комнат	Expr5_Label	Цена за м2 ...	0,10	-0,31	0,25	0,12
3	Expr3	Площадь	Expr5_Label	Цена за м2 ...	-0,13	-0,24	-0,05	-0,13
4	Expr4	ГодПостройки	Expr5_Label	Цена за м2 ...	0,31	0,31	0,39	0,32
5	Expr0	Этаж	Expr5_Label	Цена за м2 ...	-0,39	-0,39	-0,16	-0,33
6	COL6	Материал стен	Expr5_Label	Цена за м2 ...	0,09	0,26	0,17	0,05
7	Expr1	Высота потолка	Expr5_Label	Цена за м2 ...	-0,04	-0,20	0,10	-0,03
8	COL13	Санузел	Expr5_Label	Цена за м2 ...	-0,00	-0,24	0,09	-0,02
9	COL14	Ремонт	Expr5_Label	Цена за м2 ...	-0,18	-0,33	0,00	-0,14
10	COL15	Парковка	Expr5_Label	Цена за м2 ...	-0,26	-0,31	-0,19	-0,26
11	COL16	Тип жилья	Expr5_Label	Цена за м2 ...	0,36	0,36	0,40	0,37
12	COL12	Балкон/Лоджия	Expr5_Label	Цена за м2 ...	-0,10	0,26	-0,06	-0,11

Рис. 72 «Значимость факторов»

Коэффициент корреляции Пирсона говорит о степени зависимости входного фактора от выходного. Принято считать, что, если коэффициент больше 0,6 – зависимость сильная, если меньше 0,3 – очень слабая.

Те факторы, которые меньше всего влияют на цену объекта можно в дальнейшем не учитывать.

4.1. Расчет электронной цены на основании коэффициентов и определение выгодных предложений

В оценке стоимости объектов недвижимости применяются 3 метода

1. Затратный метод

Метод отталкивается от идеи, что объект не может стоить больше, чем придется затратить для создания аналогичного. Для этого метода рассчитывают затраты за восстановление оцениваемого объекта, с учетом износа и отдельной оценке рыночной стоимости земли. Не применим к тому типу недвижимости, который рассматривается.

2. Доходный подход

Здесь опираются на перспективы будущих доходов от объекта оценки, а именно от перепродажи и эксплуатации. Метод не совсем подходит для оценки жилой недвижимости.

3. Сравнительный подход

Основан на сравнении цен на похожие объекты рынка. Покупатель выберет наиболее выгодный для него объект. Стоимость объекта получают через его сравнение с объектами, проданными на данном рынке в течение недавнего времени. Для применения данного подхода нужно обладать большим объемом информации. Этот метод большего всего подходит для расчета стоимости жилой недвижимости.

4. Современным и упрощенным методом считается нахождение электронной цены объекта, которую можно получить путем анализа рынка предложений.

Суть метода в том, что берется массив с данными об объектах, в программе MC Excel с помощью специально разработанного макроса рассчитываются весовые коэффициенты для каждого фактора, и рассчитывается стоимость квадратного метра «эталонного объекта», которая в последующем берется за основу для расчета. Каждому параметру оцениваемых объектов присваивается коэффициент. Потом, все коэффициенты перемножаются со стоимостью квадратного метра. И получается, так скажем, рыночная цена объекта. Таким образом можно увидеть, завышена стоимость на объект или наоборот является очень выгодной.

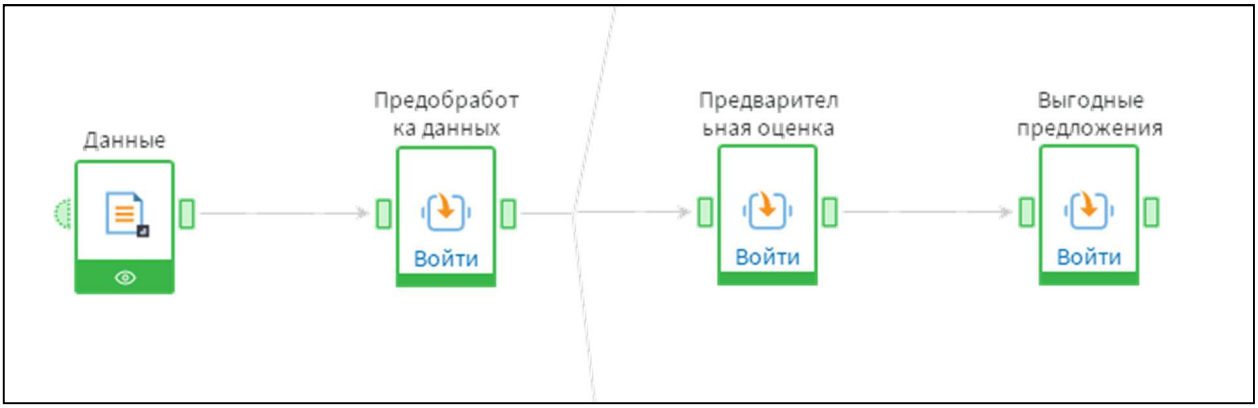


Рис. 73. Сценарий определения «Выгодных предложений»

Внутри компонента подаются файлы с коэффициентами по параметрам. Затем они преобразуются в переменные, чтобы в последующим с ними можно было проводить расчеты. Каждому объекту к соответствующему параметру проставляется коэффициент. На вход так же поступает переменная, в которой хранится стоимость квадратного метра «эталонного объекта». Перемножаем эту

стоимость с со всеми коэффициентами и получим электронную цену квадратного метра по каждому объекту.

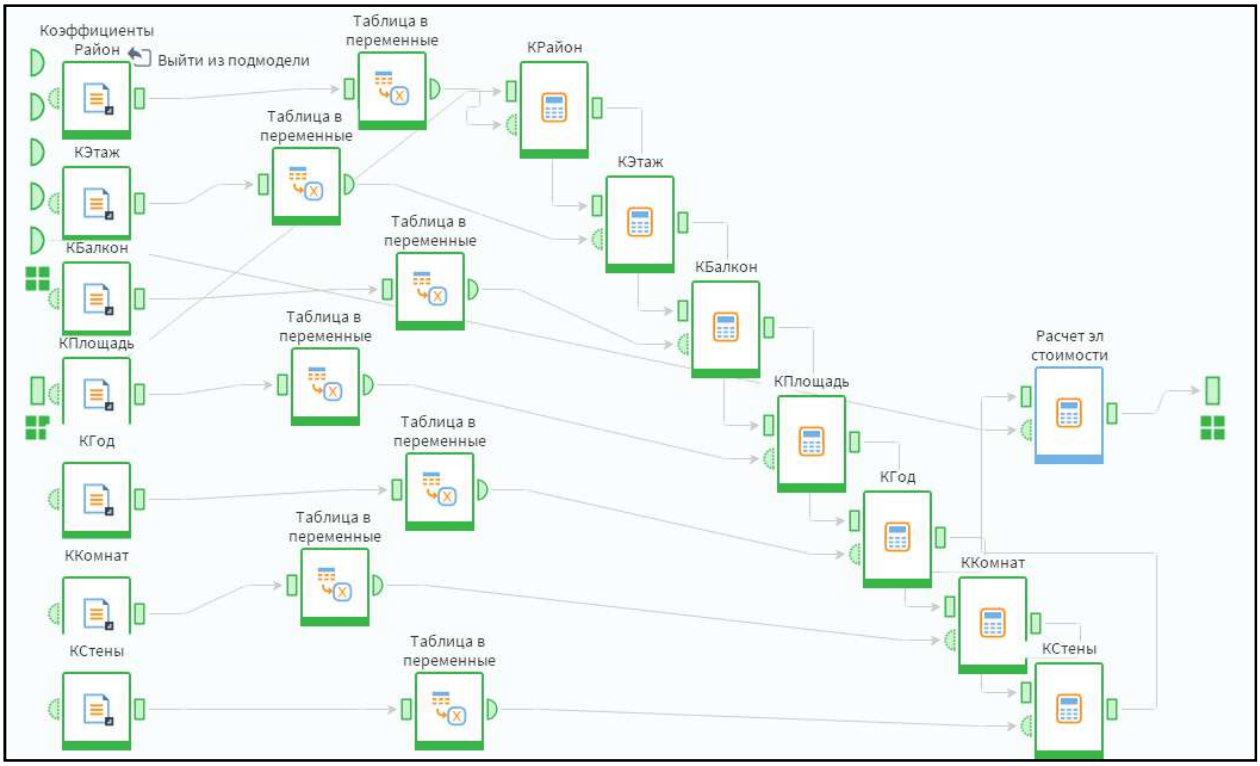


Рис. 74. Структура компонента «Предварительная оценка»

На основе полученных данных можно посмотреть какие предложения на рынке являются выгодными. Для этого нужно сравнить цену продавца и электронную цену.

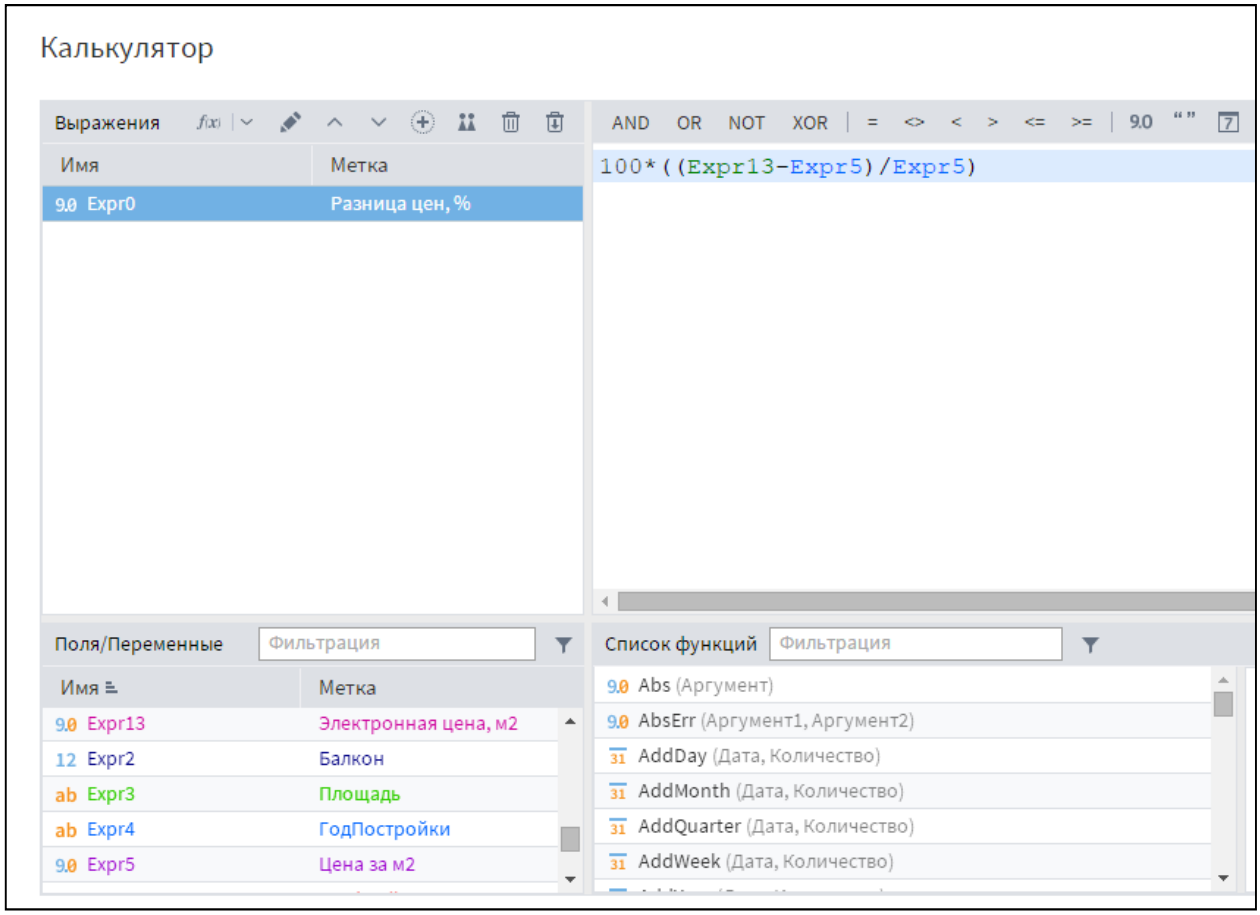


Рис. 75 «Нахождение разницы цен»

Необходимо определить разницу цен в процентах, расчёта используется формула: ((Электронная цена – цена предложения)/цена предложения) * 100%.

Если значения отрицательные, значит цена продавца завышена относительно похожих предложений на рынке при таких характеристиках объекта. Если значения больше 0, то у предложения относительно низкая цена. Отбираются только непроданные объекты, чтобы отметить на какие предложения стоит обратить внимание. Отсортируем их, в порядке убывания процента. В первых строках будут наиболее выгодные предложения к покупке.

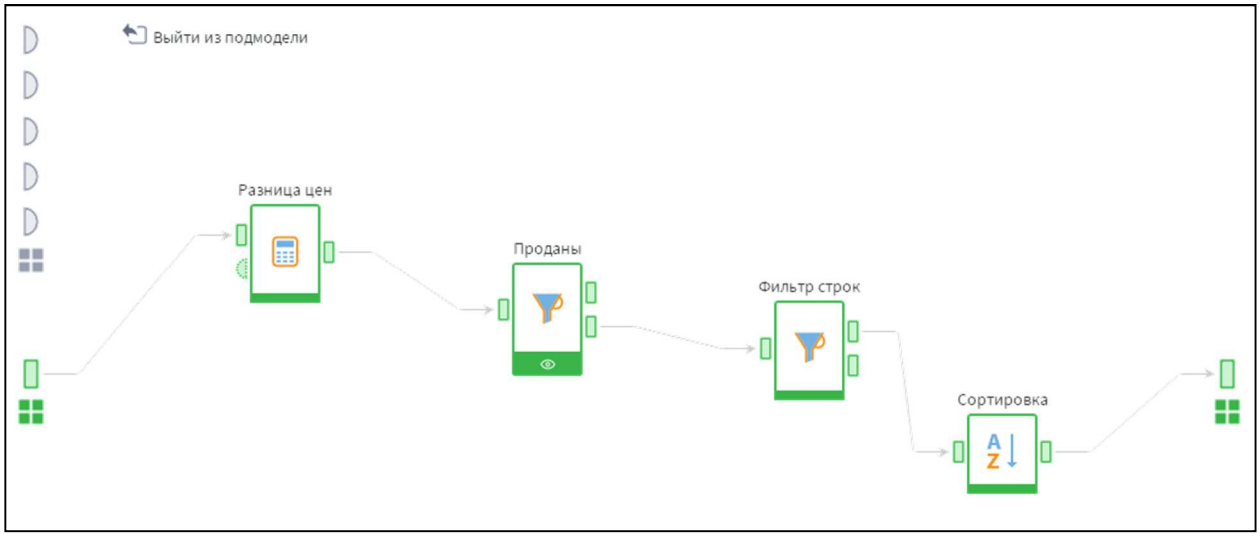


Рис. 76 Структура компонента «Выгодные предложения»

#	Разница цен, %	Электронная цена, м2	Цена за м2	Район	Площадь	Кол-во комнат	ГодПостройки	Материал стен	Этаж
1	136,32	64 003,85	27 083,33	Нижегородский	30-50	2	1930-1990	панель	первый
2	136,12	66 900,75	28 333,33	Нижегородский	>50	1	<1930	кирпич	первый
3	135,02	54 654,66	23 255,81	Канавинский	30-50	2	1930-1990	кирпич	последний
4	118,26	71 845,12	32 916,67	Нижегородский	>50	3	>2005	кирпич	последний
5	116,62	71 845,12	33 166,67	Нижегородский	>50	3	>2005	кирпич	последний
6	105,10	81 086,00	39 534,88	Нижегородский	30-50	1	>2005	кирпич	средний
7	104,06	43 080,21	21 111,11	Московский	30-50	2	1930-1990	дерево	последний
8	91,04	57 271,99	29 978,59	Канавинский	30-50	2	1930-1990	кирпич	средний
9	91,03	63 676,12	33 333,33	Канавинский	30-50	2	1990-2005	кирпич	средний
10	82,51	57 868,71	31 707,32	Ленинский	30-50	2	1930-1990	кирпич	средний
11	71,82	58 779,63	34 210,53	Ленинский	<30	1	1930-1990	каркасно-засыпной	первый
12	66,91	48 380,61	28 985,51	Автозаводский	>50	2	1930-1990	шлакоблок	средний
13	65,97	59 274,15	35 714,29	Советский	30-50	2	1930-1990	кирпич	первый
14	65,81	50 300,96	30 337,08	Сормовский	>50	3	1930-1990	кирпич	средний
15	64,90	46 826,49	28 397,21	Сормовский	>50	3	1930-1990	кирпич	первый
16	64,74	47 468,24	28 813,56	Сормовский	>50	4	1930-1990	панель	последний
17	64,34	56 667,40	34 482,76	Канавинский	<30	2	1930-1990	кирпич	первый
18	63,58	54 527,73	33 333,33	Канавинский	>50	2	1930-1990	кирпич	последний
19	61,93	58 883,83	36 363,64	Приокский	30-50	2	1930-1990	кирпич	средний
20	61,29	65 607,54	40 677,97	Нижегородский	>50	2	<1930	кирпич	последний

Рис. 77 «Выгодные предложения»

4.2. Прогнозирование продажи объекта недвижимости на основе логистической регрессии

Логистическая регрессия помогает определить принадлежность к одному из двух классов. В данном случае это проданные и непроданные объекты недвижимости. Необходимо понять по характеристикам квартиры будет ли она продана или нет. Если точность модели будет достаточно велика, у той квартиры у которой сейчас стоит статус «не продана», а модель предсказывает ей статус «продана», можно будет говорить о том, что в ближайшее время ее статус изменится.

Для корректной работы компонента на вход подается следующая информация:

- район;
- этаж;
- материал стен;
- высота потолков;
- год постройки;
- тип санузла;
- парковка;
- ремонт (состояние помещения);
- тип жилья;
- общая площадь;
- тип жилья;
- количество комнат.

На выходе будет статус объявления.

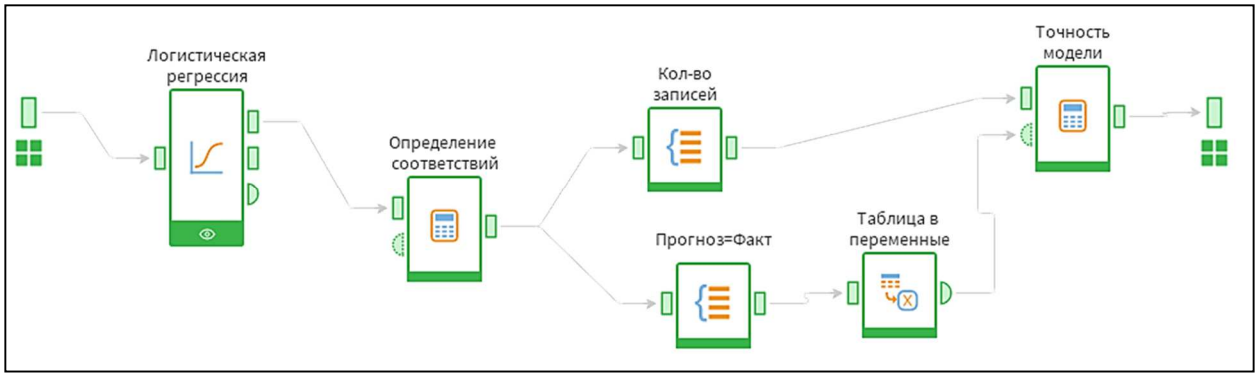


Рис. 78. Структура компонента «Логистическая регрессия»

Для обучения модели определено, что обучающее множество составляет 95% от общего количества данных, а тестовое 5%. Остальные настройки остаются стандартными.

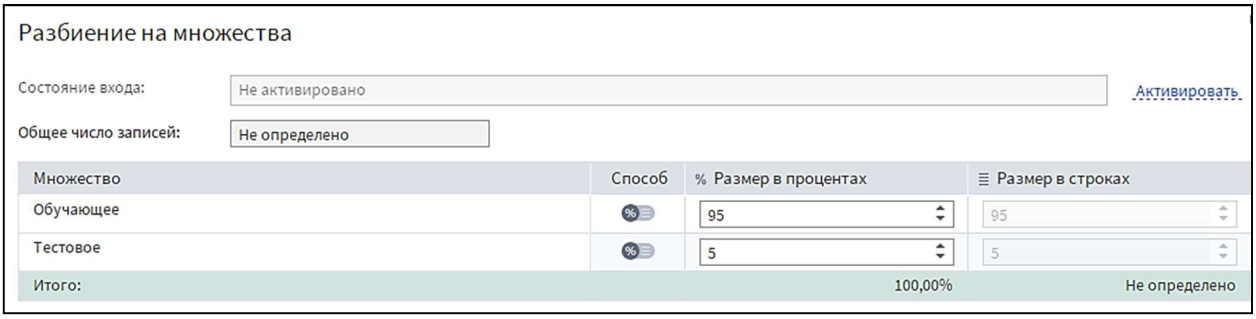


Рис. 79. «Разбиение на множества»

Результатом работы инструмента Логистическая регрессия становятся прогнозные значения статуса объявления с вероятностью этого события.

Логистическая регрессия • Выход регрессии • Быстрый просмотр данных

#	9.0 Вероятность события|Прогноз	ab Статус|Прогноз	ab Статус	ab Район	ab Площадь	ab Этаж	9.0 Цена за м2
1	0,87	Снято с продажи	Снято с продажи	Канавинский	>50	первый	42 727,27
2	0,62	Снято с продажи	Снято с продажи	Сормовский	30-50	последний	45 348,84
3	0,30	Не снято с продажи	Не снято с продажи	Сормовский	>50	средний	56 830,60
4	0,00	Не снято с продажи	Не снято с продажи	Приокский	>50	средний	60 000,00
5	1,00	Снято с продажи	Снято с продажи	Автозаводский	<30	средний	61 904,76
6	0,70	Снято с продажи	Снято с продажи	Советский	>50	средний	63 901,35
7	0,33	Не снято с продажи	Не снято с продажи	Сормовский	>50	первый	39 522,06
8	0,99	Снято с продажи	Снято с продажи	Московский	>50	средний	54 666,67
9	0,94	Снято с продажи	Снято с продажи	Канавинский	>50	средний	70 666,67
10	0,77	Снято с продажи	Не снято с продажи	Автозаводский	>50	средний	51 724,14
11	0,70	Снято с продажи	Не снято с продажи	Ленинский	>50	средний	69 064,75
12	0,00	Не снято с продажи	Не снято с продажи	Нижегородский	<30	первый	36 956,52
13	0,39	Не снято с продажи	Не снято с продажи	Московский	>50	последний	72 727,27
14	0,75	Снято с продажи	Снято с продажи	Сормовский	>50	последний	50 000,00
15	0,22	Не снято с продажи	Не снято с продажи	Канавинский	<30	средний	64 935,06
16	0,56	Снято с продажи	Снято с продажи	Автозаводский	30-50	средний	57 446,81
17	0,03	Не снято с продажи	Снято с продажи	Ленинский	>50	средний	58 846,15
18	0,05	Не снято с продажи	Не снято с продажи	Московский	30-50	последний	67 055,39
19	0,18	Не снято с продажи	Не снято с продажи	Автозаводский	30-50	средний	53 571,43
20	0,18	Не снято с продажи	Не снято с продажи	Сормовский	>50	первый	50 980,39
21	0,29	Не снято с продажи	Не снято с продажи	Сормовский	30-50	средний	54 347,83
22	0,17	Не снято с продажи	Не снято с продажи	Сормовский	>50	средний	45 611,02

Рис. 80. «Результат Логистической регрессии»

Для оценки точности модели необходимо произвести расчет на совпадение реальных показателей и прогнозных. Для этого находится их отношение, путем деления числа совпадений на общее число записей. Если отношение очень мало, то есть количество совпадений не большое, говорит о плохо построенной модели. Следует или увеличит обучающую выборку или произвести более тщательную предобработку (убрать экстремальные значения, шумы, изменить набор входных параметров). Если же отношение близко к 1, то есть совпадения показаний почти 100%, это означает что модель переобучена. В таком случае модель

на этих данных будет показывать отличные результаты, но будет бесполезна для других данных.

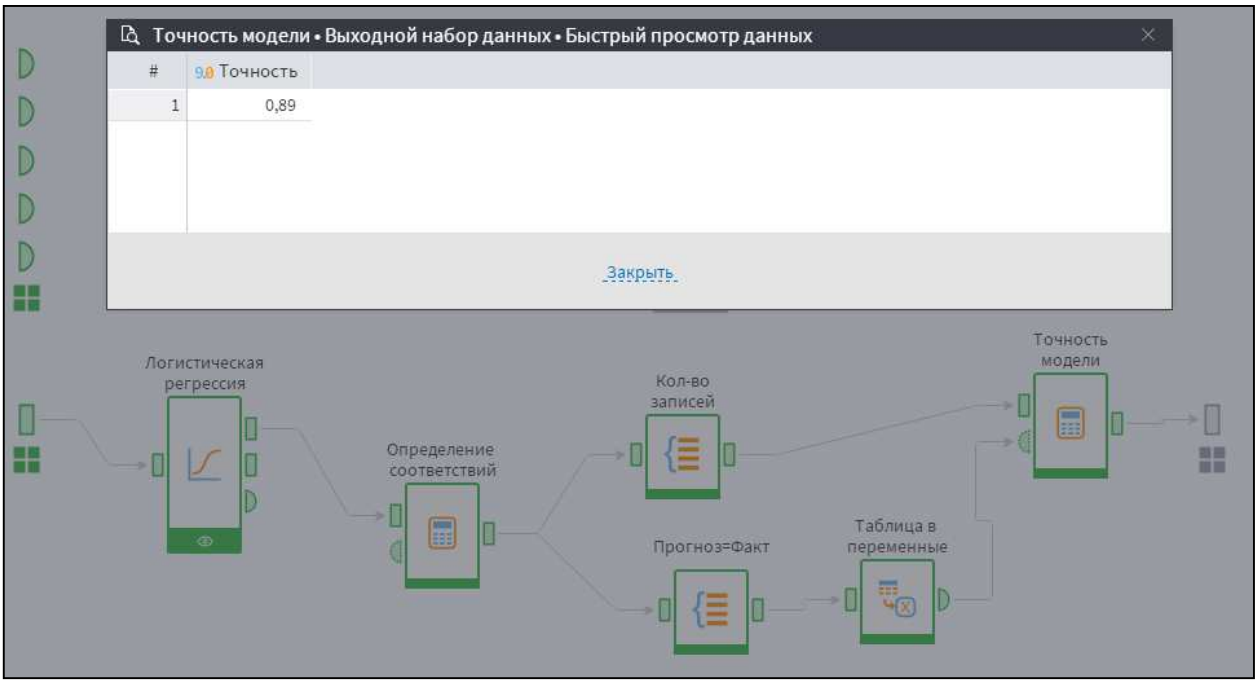


Рис. 81. «Точность модели Логистической регрессии»

Для оценки качества модели можно воспользоваться встроенным визуализатором – «качество бинарной регрессии», в котором показана ROC-кривая (часто применяется для представления результатов бинарной классификации). Она показывает зависимость количества верно классифицированных примеров от количества неверно классифицированных примеров. ROC-кривая получается следующим образом: для каждого значения порога отсечения, которое меняется от 0 до 1 с шагом dx (например, 0.01) рассчитываются значения чувствительности и специфичности. Строится график зависимости: по оси Y откладывается чувствительность, по оси X – (100% – специфичность).

Анализ этой кривой помогает оценить качество модели классификации. Чем меньше расстояние от диагональной прямой до кривой, тем менее эффективна модель. Диагональ соответствует бесполезной модели классификации. Чем кривая ближе к верхнему левому углу, тем больше предсказательная способность модели.

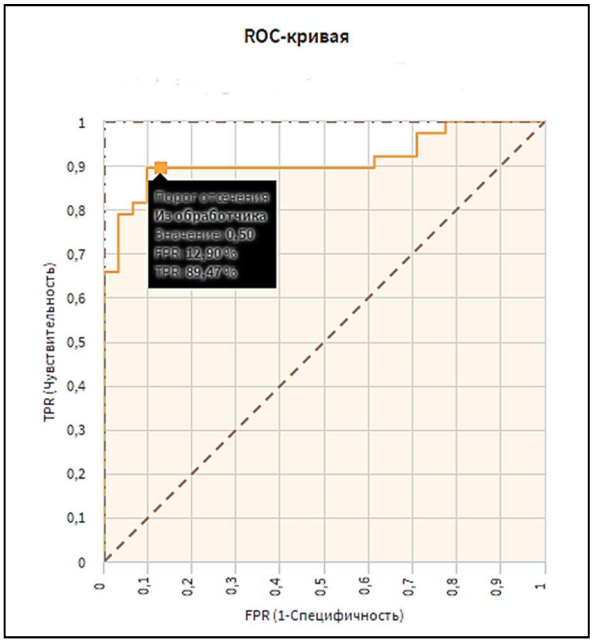


Рис. 82. «Качество бинарной классификации»

Как видно по графику площадь под кривой ROC большая, а показатель AUC близится к 1. Это говорит о том, что модель классифицирована хорошо.

Модель может предсказать по характеристикам объекта недвижимости, будет ли она популярна на рынке и есть ли большая вероятность того, что ее продадут. Если прогноз показывает, что квартира продана не будет, стоит пересмотреть характеристики, возможно указать те, что ранее были не описаны, либо пересмотреть цену своего предложения.

Атрибут	Значимость	Коэффициент	Стандартная ошибка	Коэффициент Вальда	Отношение шансов
ab Парковка (подземная)	1,000000	0,627921	0,000000	∞	1,873712
9,0 Общая	0,999999	0,009428	5 539,479322	0,000000	1,009473
12 Год постройки	0,999993	-0,000458	52,339632	0,000000	0,999543
ab Этаж (последний)	0,915814	-0,110338	1,043799	0,011174	0,895531
ab Этаж (первый)	0,884819	0,094849	0,654749	0,020985	1,099493
ab Район (Канавинский)	0,880494	-0,075954	0,505201	0,022603	0,926859
ab Кол-во комнат (3)	0,414216	-0,770577	0,943759	0,666668	0,462746
ab Район (Московский)	0,406930	-0,487734	0,588121	0,687753	0,614016
ab Район (Нижегородский)	0,335102	0,657555	0,682190	0,929079	1,930067
ab Район (Советский)	0,258986	0,608493	0,539066	1,274169	1,837660
ab Материал стен (панель)	0,090935	1,716468	1,015368	2,857750	5,564837
ab Парковка (во дворе)	0,046023	-1,819990	0,912193	3,980749	0,162027
ab Кол-во комнат (1)	0,007234	-2,342765	0,872261	7,213811	0,096062
ab Тип жилья (брежневка)	0,003144	-2,692061	0,911552	8,721820	0,067741
ab Тип жилья (хрущёвка)	0,002525	-1,645925	0,544938	9,122736	0,192834
ab Район (Приокский)	0,000253	2,019762	0,551949	13,390681	7,536531
12 Балкон	0,000094	5,419884	1,388115	15,245055	225,852979
ab Тип жилья (сталинка)	0,000076	2,093451	0,529028	15,659137	8,112863
ab Материал стен (блок+у…	0,000001	2,443983	0,505323	23,391508	11,518824
ab Ремонт ( )	0,000000	2,396099	0,454915	27,742772	10,980259
ab Район (Автозаводский)	0,000000	2,902771	0,550177	27,836858	18,224577
ab Парковка ( )	0,000000	-2,802922	0,503762	30,957903	0,060633
ab Район (Ленинский)	0,000000	3,265849	0,571439	32,662711	26,202343
ab Ремонт (типовой)	0,000000	5,022915	0,728041	47,599243	151,853317

Рис. 83. «Отчет по регрессии»

Отчет показывает значимость того или иного параметра при классификации с помощью логистической регрессии. Как видно из отчета, наиболее важными считается парковка, общая площадь объекта недвижимости, год постройки дома, этаж, район, количество комнат и материал стен. Данные для тестирования компонентов разработанной библиотеки, были взяты по объектам города Нижнего Новгорода. Анализ результатов показал, что наиболее предпочтительными считается последний этаж, из районов города – Канавинский, трехкомнатные квартиры и панельные дома. Все это достаточно объективно. Сейчас возрастает популярность на жилую недвижимость, находящуюся на верхних этажах домов, где открывается достаточно красивый вид на город. Канавинский район считается предпочтительным благодаря своему выгодному месторасположению, здесь и транспортная развязка хорошая, и близость к железнодорожному вокзалу, центру города, стадиону, паркам, наличие метро. Большую часть покупателей жилой недвижимости составляют или молодые семьи, или семьи, принявшие решение расширяться. Выбор трехкомнатных квартир обусловлен наиболее частыми вариантами, такими как, или детям, если их двое достается каждому по комнате, и одну для родителей, либо же комната родителей, детская и зал, а панельные дома давно завоевали доверие покупателей.

4.3. Прогнозирование цен на объекты жилой недвижимости на основе нейронных сетей

Прогнозирование цены объекта недвижимости можно сделать с помощью компонента Нейронная сеть. Для этого необходимо указать входные параметры: район, этаж, материал стен, высота потолков, год постройки, тип санузла, парковка, ремонт (состояние помещения), тип жилья, общая площадь, тип жилья, количество комнат, выходной параметр – цена м2. Далее указывается структура нейронной сети: входной слой состоит из 12 нейронов, соответствующих входным параметрам, один скрытый слой, состоящий из 6 нейронов, выходной слой содержит один нейрон, соответствующий полю цена.

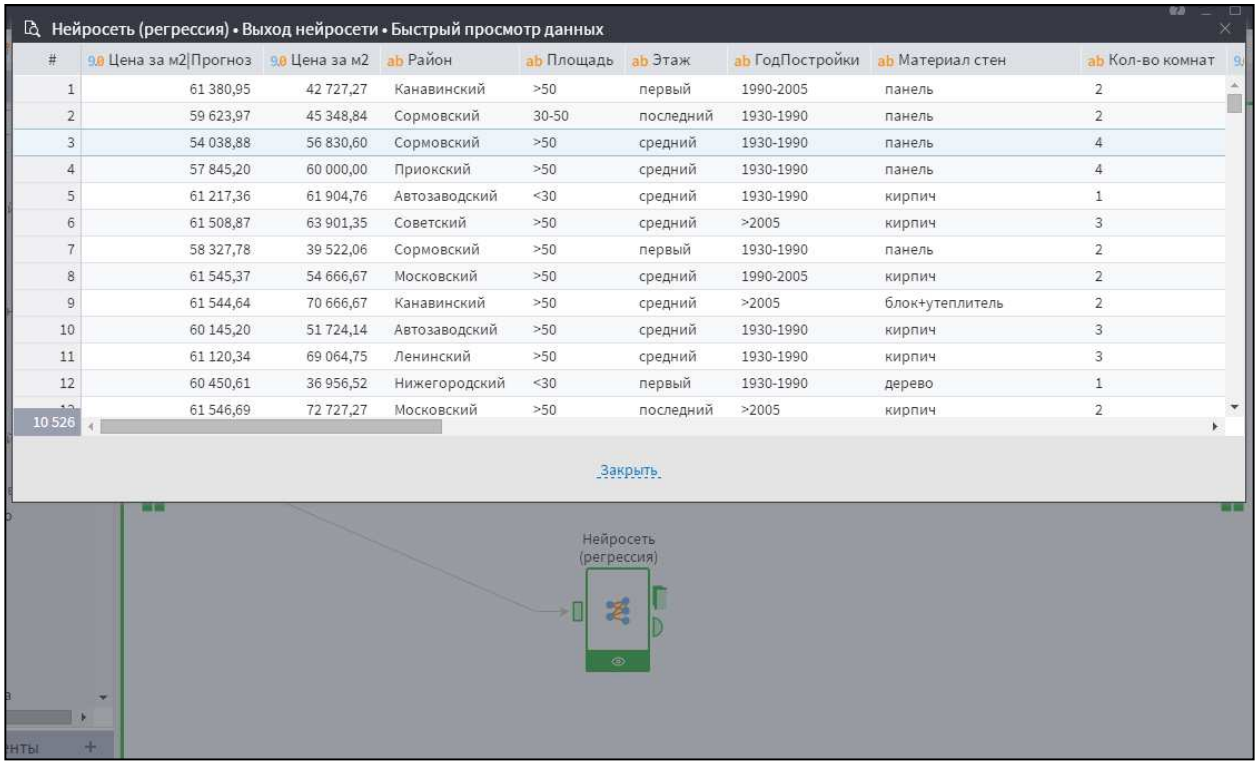


Рис. 84. «Прогнозирование с помощью нейросети»

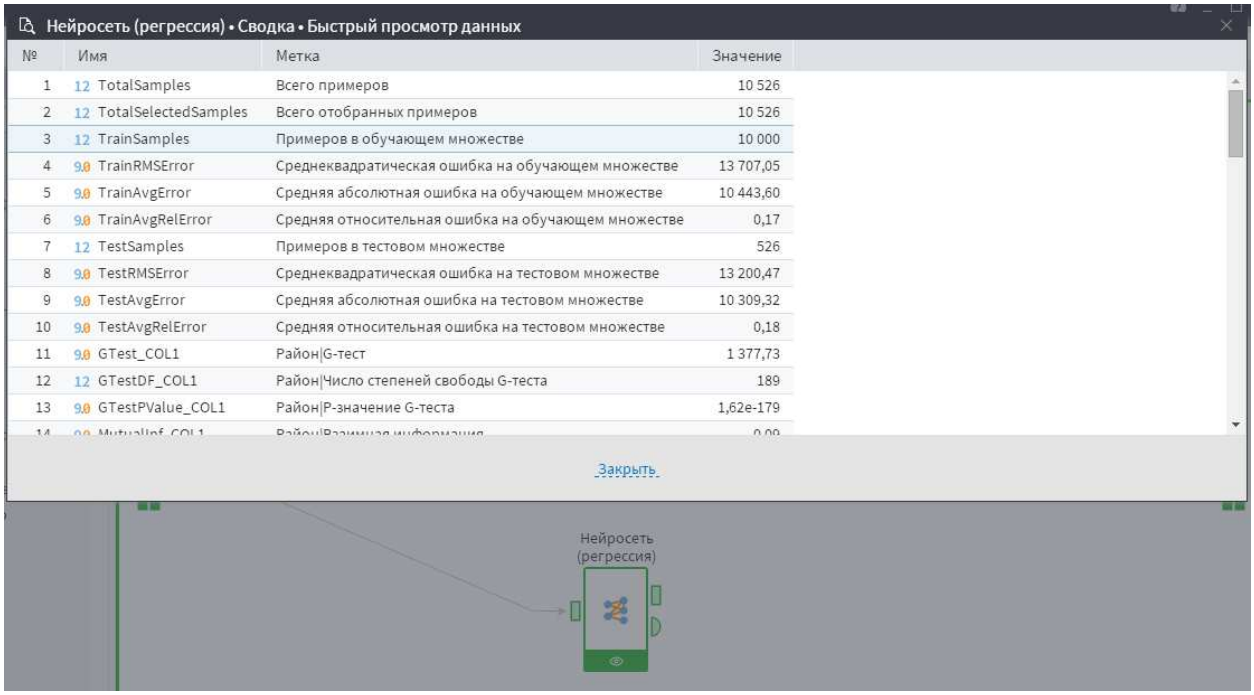


Рис. 85. «Сводная информация по прогнозированию с помощью нейросети»

Сводная информация говорит о том, что построенная модель имеет среднюю относительную ошибку на обучающем множестве 0,17, а на тестовом 0,18. Это говорит о том, что модель на этих данных выполняет хороший прогноз значений цены.

Список литературы

1. Барсегян, А.А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян, М.С. Куприянов, В.В. Степаненков, И.И. Холод. – Санкт-Петербург : БХВ-Петербург, 2012. – 335 с.
2. Паклин Н. Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям /Н.Б.Паклин, В.И. Орешков – СПб.: Питер, 2013. – 704 с.
3. Практикум по анализу данных на компьютере / И.А. Кацко, Н.Б. Паклин /под ред. Г.В. Гореловой – М.: КолосС, 2009. – 278 с.
4. Рафалович, В. Data Mining, или интеллектуальный анализ данных для занятых. Практический курс / В. Рафалович. – М.: SmartBook, 2018. – 352 с.
5. Флах, Петер Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. Учебник / Петер Флах. – М.: ДМК Пресс, 2015. – 400 с.
6. Официальный сайт компании Loginom Company [Электронный ресурс]. – Режим доступа: https://loginom.ru/.
7. Энциклопедия по Бизнес-анализу [Электронный ресурс] режим https://wiki.loginom.ru
8. Loginom – Руководство пользоавателя [Электронный ресурс]. – Режим доступа: https://help.loginom.ru/userguide/.

Прокопенко Наталья Юрьевна

АНАЛИТИЧЕСКИЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ на базе АП Loginom

Учебное пособие

Подписано в печать Формат 60х90 1/16 Бумага газетная. Печать трафаретная.
Уч. изд. л. 8,5. Усл. печ. л. 8,9. Тираж 300 экз. Заказ №

Федеральное государственное бюджетное образовательное учреждение высшего образования «Нижегородский государственный архитектурно-строительный университет»
603950, Нижний Новгород, ул. Ильинская, 65.
Полиграфический центр ННГАСУ, 603950, Н.Новгород, Ильинская, 65
http://www. nngasu.ru, srec@nngasu.ru

143