1

Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Автоматизированные системы обработки информации и управления»



**Отчет**

**Лабораторная работа № 6**

**«Знакомство с платформой для анализа данных Loginom"»**

**ИСПОЛНИТЕЛЬ:**

ФИО

Группа ИУ5-55Б

\_\_\_\_Финк Д.Д, \_\_\_\_\_

"\_18\_"\_\_\_декабря\_\_\_2021 г.

**ПРЕПОДАВАТЕЛЬ:**

Селиверстова А.В.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

"\_\_"\_\_\_\_\_\_\_\_\_\_\_2021 г.

Москва 2021

**1. Цель лабораторной работы**

Установить BI Loginom версии Community Edition.

Ознакомиться с интерфейсом подключения к внешним источникам данных и основными объектами платформы бизнес аналитики Loginom. Изучить пакеты и компоненты, предлагаемые платфомой.

**2. Теория**

**Loginom** — аналитическая платформа, позволяющая в единой среде

выполнить все этапы бизнес-анализа от консолидации данных и

построения моделей до визуализации и интеграции в бизнес-процесс.

Одной из основных концепций является Сценарий.

**Сценарий** — последовательность действий, которые необходимо

провести для анализа данных. Он представляет собой комбинацию узлов

обработки данных, настраиваемую пользователем для решения

конкретной задачи.

**Узел сценария** выполняет отдельную операцию над данными.

**Пакет** включает в себя Модули и декомпозирует анализ на уровне модулей.

**Подмодель** - специальный узел, который включает в

себя другие узлы сценария. Реализованная в подмодели логика может

быть произвольной, при этом разработчик сценария может рассматривать

её как «чёрный ящик». Подмодель принимает информацию через входные

порты, производит обработку и выдает результат на выходные порты.

Входные и выходные порты задаются пользователем.

**Проект** — комплекс сценариев, файлов, источников данных и

прочих элементов, предназначенных для решения отдельной

аналитической задачи.

Проект может объединять в себе несколько пакетов благодаря тому,

что каждый пакет имеет возможность предоставлять свои объекты другим пакетам через механизм ссылок.

Проект строится из визуальных блоков и не требует навыков программирования.

Loginom предоставляет инструменты для подготовки скоринговой выборки, анализа жизненных циклов, двумерного анализа, моделирования**.**

**3. Ход работы**

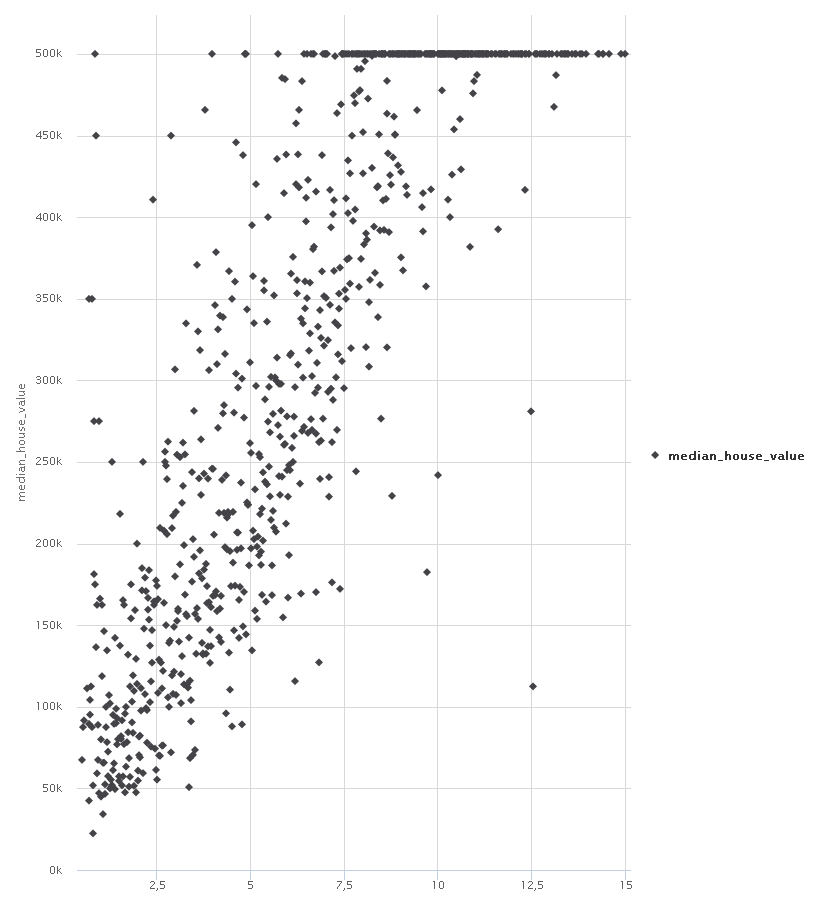
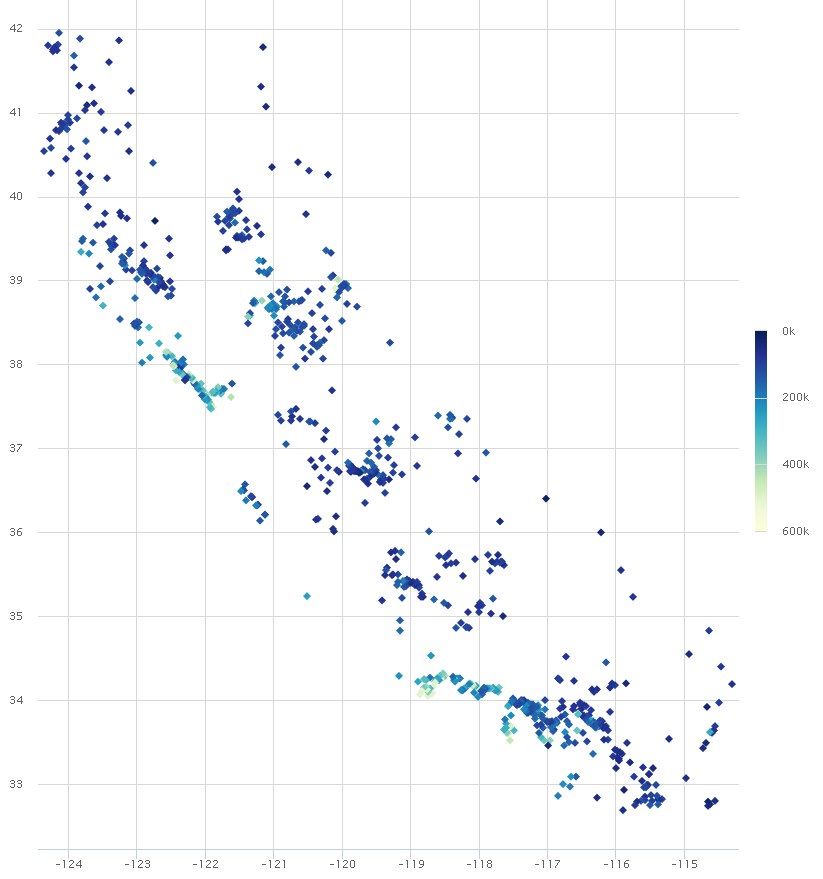
Ознакомился с теоретической частью и инструментарием программы.

Установил пробную версию ПО с базовыми пакетами.

Импортировал таблицу housing в формате csv, содержащую информацию о рынке арендной недвижимости, с помощью компонента «Импорт из текстового файла».

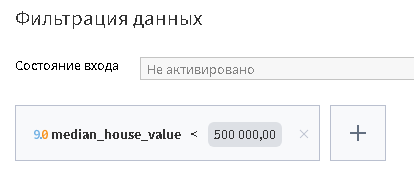


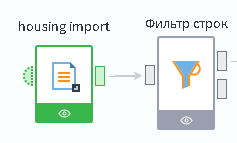
С помощью настройки визуализации компонента «Импорт из текстового файла» построил следующие диаграммы разброса

: 

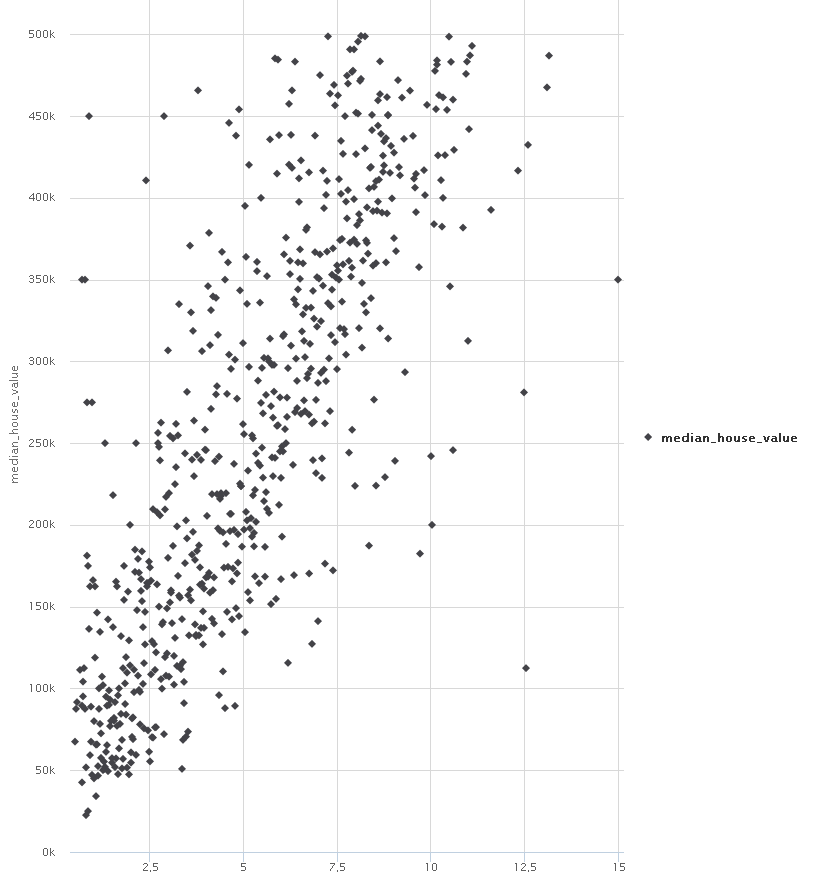
По второй видно, что есть неприятное множество точек на отметке $500 тыс, которое явно исказит дальнейший анализ, т.к. данных о таких домах вообще нет в исходной таблице.

С помощью компонента «Фильтр» убрал все дома стоимостью менее $500 тыс.

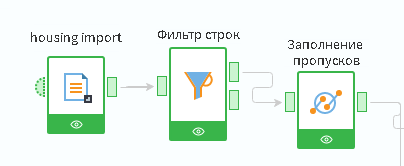


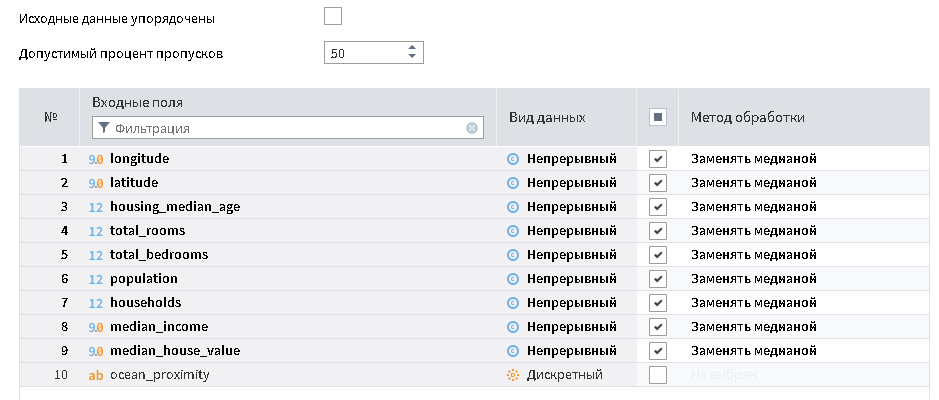


С помощью настройки визуализации компонента «Фильтр» перестроил диаграмму стоимости домов:



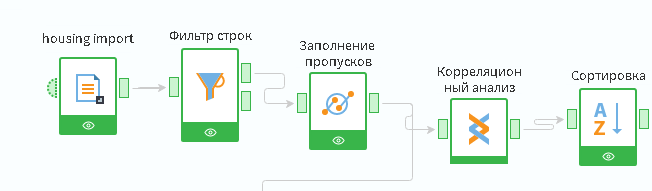
Теперь ложных значений нет.

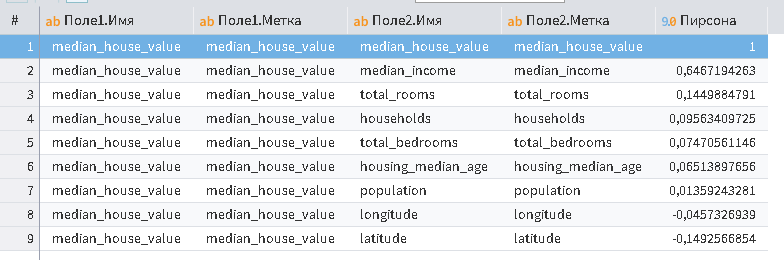
В исходной таблице были пропуски по значению числа спален. Всего более 400. Нужно очистить данные и заполнить пропуски. Для этого применил компонент «заполнение пропусков»  




Выполнил корреляционный анализ для восстановленных данных с помощью компонента «корреляционный анализ» методом Пирсона. С помощью компонента

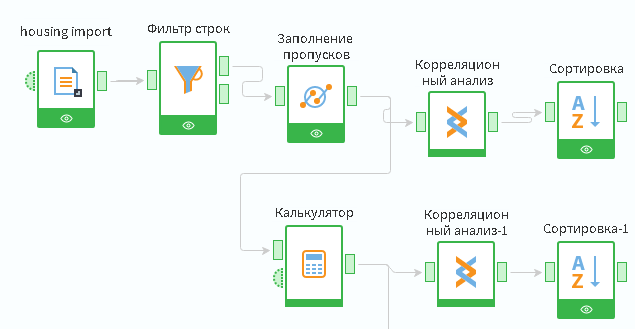
«Сортировка» сортировал имена полей по убыванию коэффициента корреляции. Визуализировал данные таблицей:

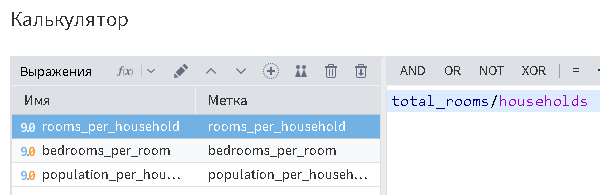


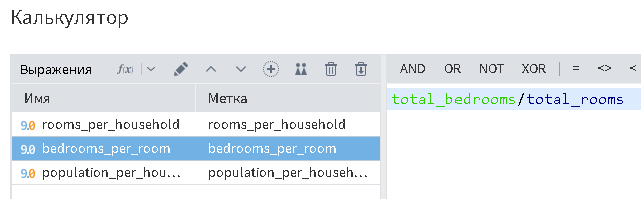


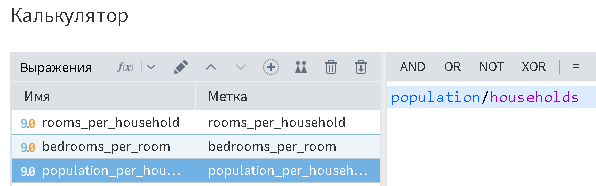
По ней видно, что наибольшее влияние на стоимость как дома, так и его арендной платы в порядке убывания важности влияют число комнат, число семей, проживающих в нем, число ванных комнат, возраст дома.

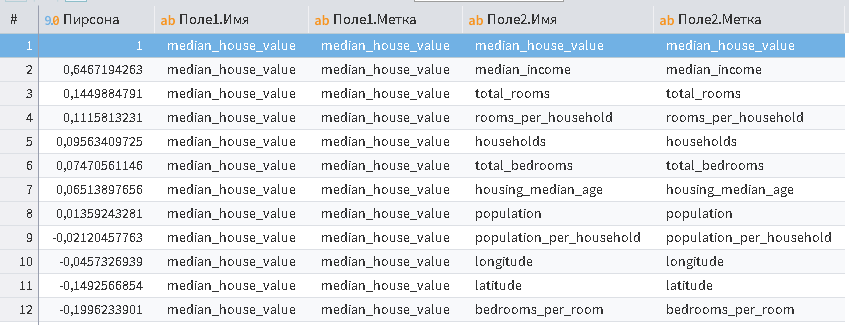
Добавил еще несколько полей для восстановленных данных с помощью компонента «калькулятор». Провел идентичный корреляционный анализ и получил следующие данные:







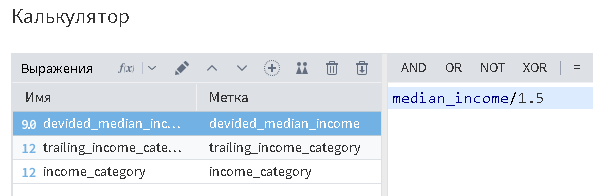


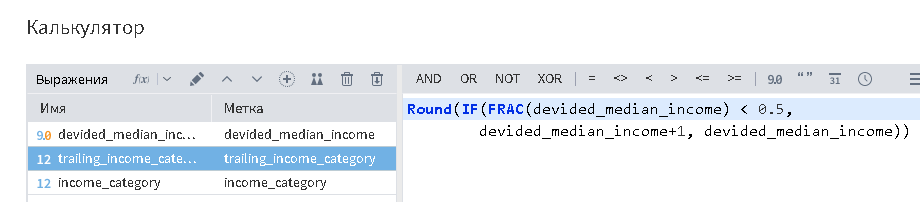


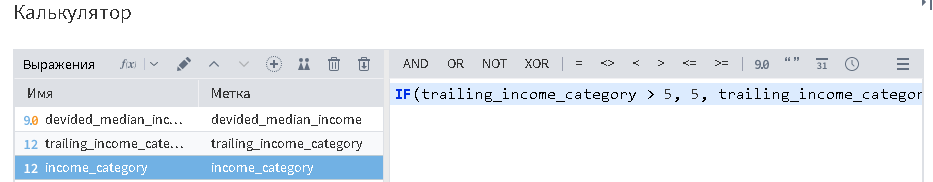
По данным ясно, что наибольшее влияние на стоимость недвижимости наибольшее влияние оказывают: число комнат, число комнат на семью, число семей, проживающих в доме, число ванных комнат, возраст дома, затем число жителей. Наименьшее влияние оказывает высотность и размер здания, а также число ванных комнат в расчете на число комнат.

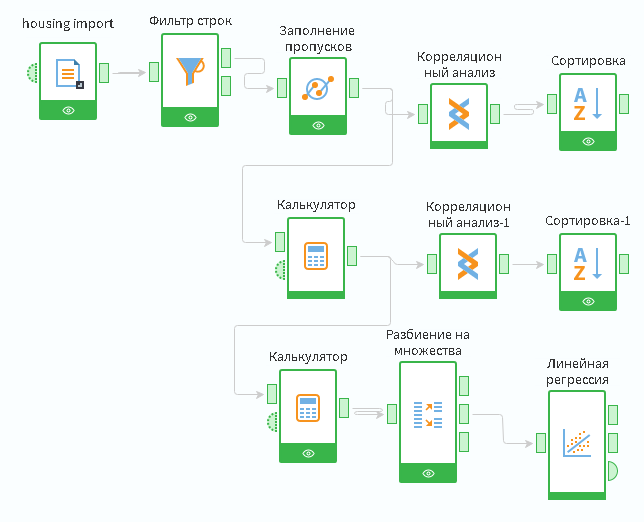
Выполнил еще одно исследование. Провел стратификационный анализ и построил модель методом линейной регрессии.

Для этого ввел поле income\_category, чтобы разделить дома на группы по доходности в условных единицах с помощью «калькулятора». Анализ проводится с помощью линейной регрессии (компонент «линейная регрессия») на основе данных, пригодных для нее, получаемых в компоненте «разбиения на множества».









Проанализировал полученную модель. Для этого:

Импортировал тестовые данные о кредитах – файл credits.txt, в который записана таблица.

***4.2***

Для дальнейшей работы произведите импорт текстового файла

Credit.txt, как делали это в предыдущих практических заданиях. Все

пункты выполнения импорта можно оставить по умолчанию, название

файла выбрать самостоятельно. После чего перейдите в панель

компонентов и используйте Кластеризацию. Сконфигурируйте входной

порт и перейдите в настройки параметров кластеризации.

В параметре настройки входных столбцов установите значение поля

“Назначение”

Используемым у следующих признаков:

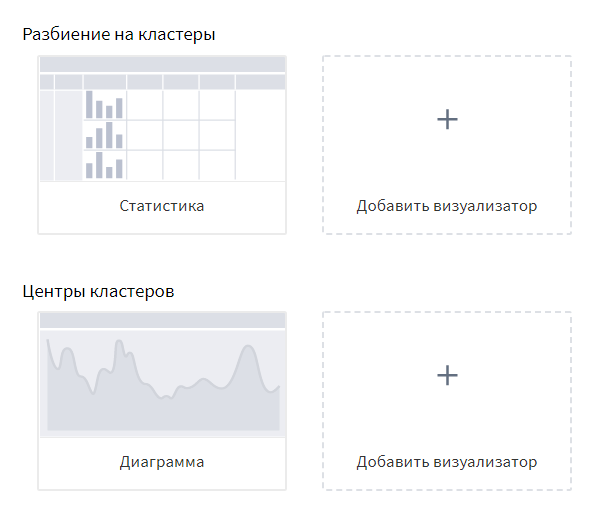
- Сумма кредита;

- Стоимость кредита;

- Срок кредита;

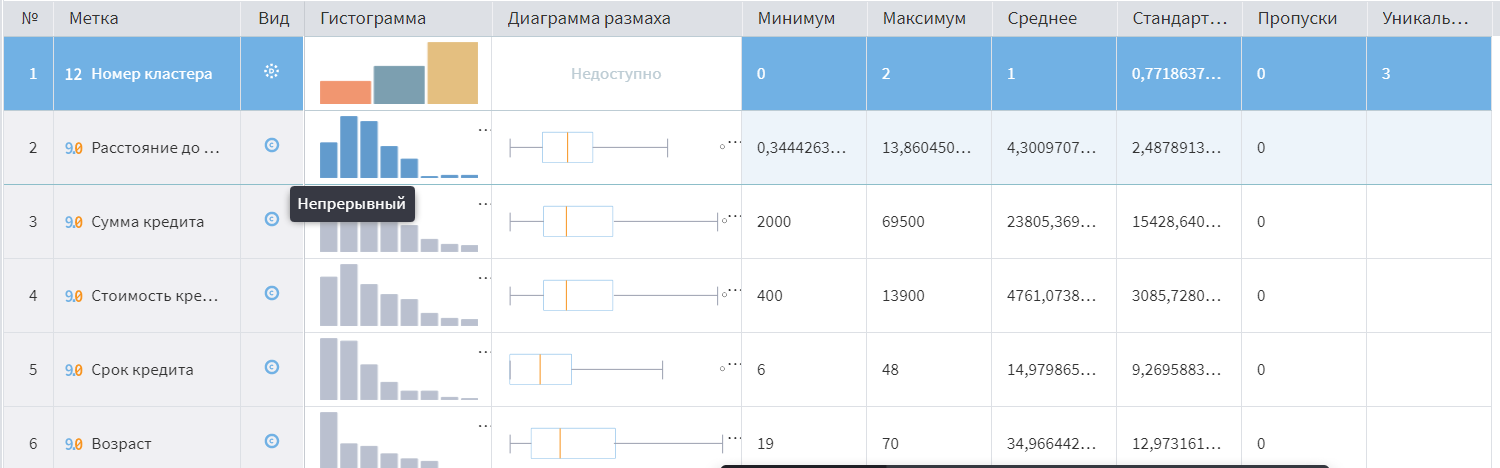
- Возраст;

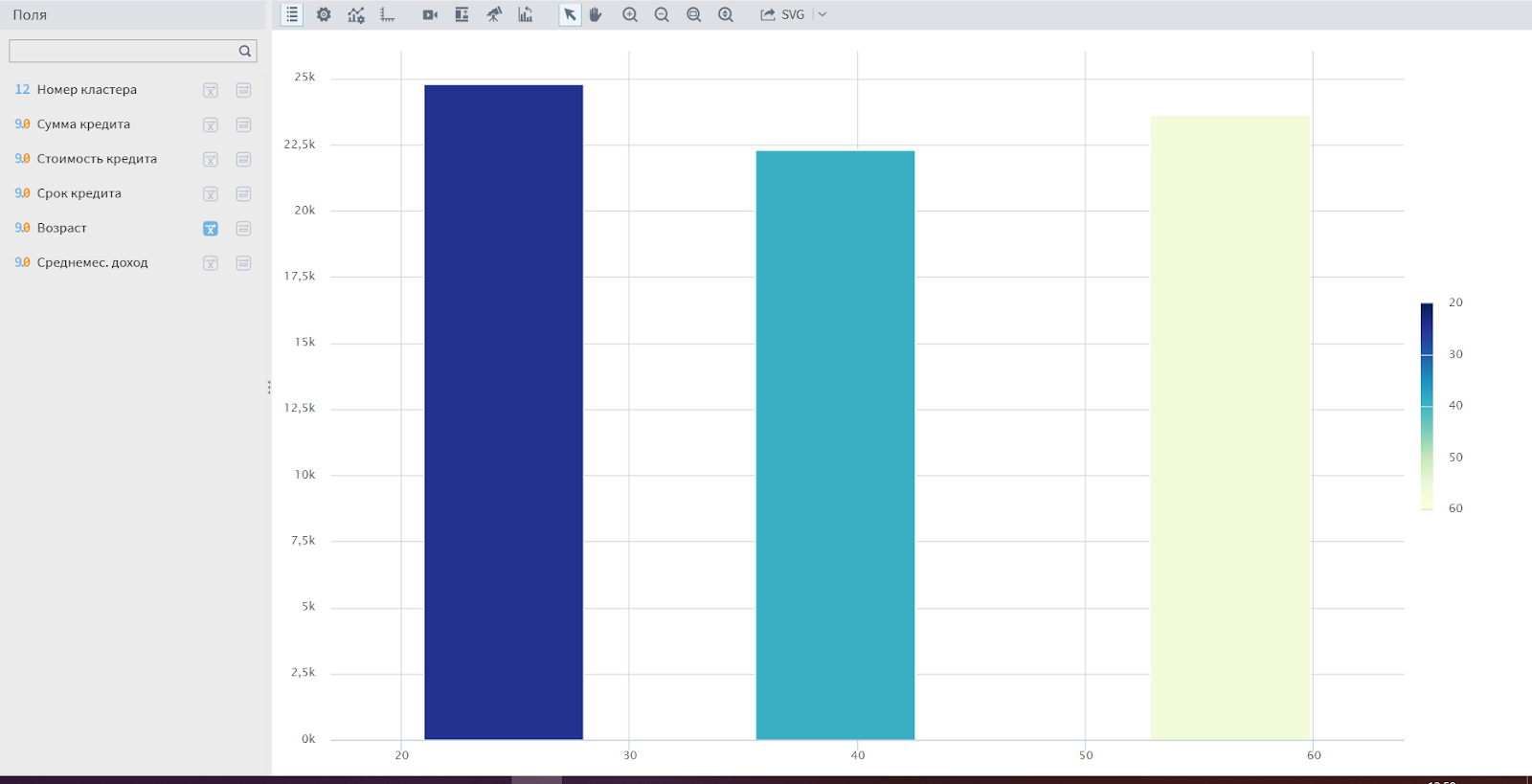
- Среднемесячный доход;

****

В настройках нормализации постарайтесь самостоятельно настроить

нормализации (здесь пригодится теоретический материал) таким образом,чтобы кластеры были распределены по возрасту наиболее логичным образом. Необходимо получить предположительно 3 кластера.

****

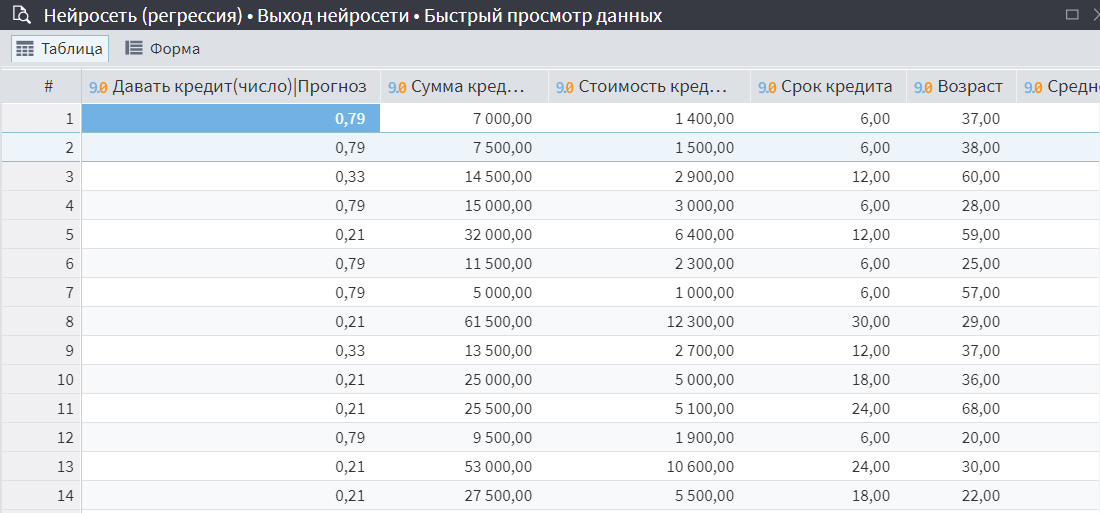
****

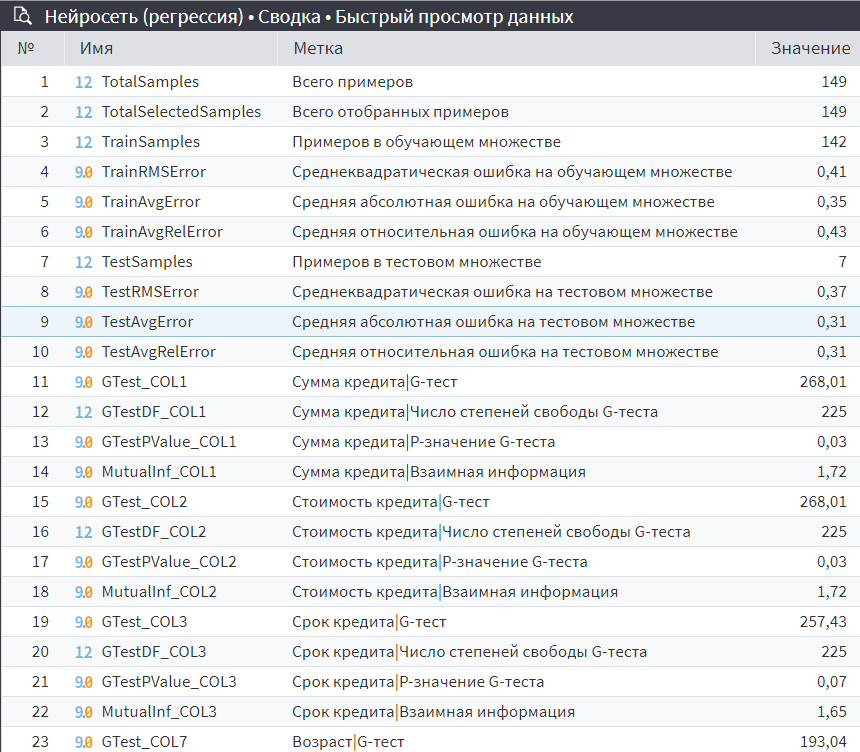
К ранее используемому источнику данных примените компонент

“Нейросеть (регрессия). Задайте назначение входного параметра для тех

же меток, что использовались в кластеризации.

Выходным- Давать кредит (число). Нормализацией выходного параметра выберите Масштабирование [min; max]. Остальные оставьте по умолчанию. Разбиение на множества установите 95% к 5%. Количество скрытых слоёв -1;  Количество нейронов в 1 скрытом слое- 5. Название каждого узла всегда остается на ваше усмотрение. Правой кнопкой мыши нажмите на узел и произведите его переобучение перед выполнением во избежание последующих ошибок. Нажмите на выход нейросети, выберите быстрый просмотр, изучите полученный прогноз.





**4. Выводы.**

Мною были изучены инструментарий программы Loginom, структурный принцип построение анализа в ней, а также применены методы математической статистики и ML для анализа бизнес-задач.

**Контрольные вопросы.**

**1. Декомпозиция проекта в Loginom. Что такое пакет? Модуль? Сценарий? Узел?**

**Сценарий** — последовательность действий, которые необходимо

провести для анализа данных. Он представляет собой комбинацию узлов

обработки данных, настраиваемую пользователем для решения

конкретной задачи.

**Пакет** включает в себя Модули и декомпозирует анализ на уровне модулей.

**Подмодель** - специальный узел, который включает в

себя другие узлы сценария. Реализованная в подмодели логика может

быть произвольной, при этом разработчик сценария может рассматривать

её как «чёрный ящик». Подмодель принимает информацию через входные

порты, производит обработку и выдает результат на выходные порты.

Входные и выходные порты задаются пользователем.

**2. Способы нормализации непрерывных данных.**

Варианты нормализации данных в Loginom:

**Нет** — отсутствие нормализации. В таком случае данные поступают в основной алгоритм без предварительной обработки.

**Масштабирование** [min;max] — приведение данных линейным преобразованием к заданному пользователем диапазону [min;max]:

Минимум — минимальное значение;

Максимум — максимальное значение.

**Масштабирование [-1;1]** — приведение данных линейным преобразованием к диапазону [-1;1].

**Масштабирование [0;1**] — приведение данных линейным преобразованием к диапазону [0;1].

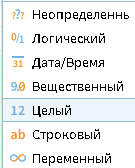
**Абсолютное масштабирование** — каждое значение делится на максимальное абсолютное значение.

**Стандартизация** — из каждого значения вычитается среднее значение и делится на стандартное отклонение.

**Отношение** — каждое значение делится на статистический показатель, либо на заданное пользователем значение:

**3. Типы дискретных и непрерывных данных.**

Все типы:



Из них дискретными являются: логический, дата, строковый, а непрерывными являются: вещественный и целый.

**4. Какие визуализаторы определены для всех базовых компонентов Loginom?**

Диаграмма, OLAP-куб, статистика, таблица.

**5. С какой целью проводится очистка данных от пропущенных и критических значений?**

Приведение данных к пригодному для анализа виду. Устранение значений, заведомо приводящих к недостоверным результатам анализа. Также улучшается качество анализа.

**6. С какой целью проводится стратификация исходных данных?**

Для равномерного и непредвзятого отбора данных для обучения моделей, которые с точки зрения имитации бизнес-процесса должны учитывать разные категории входных данных.