минобрнауки россии

федеральное государственное бюджетное

образовательное учреждение высшего образования

«ЧЕРЕПОВЕЦКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

|  |
| --- |
| Иформационных технологий |
| наименование института (факультета) |
| Математическое и программное обеспечение ЭВМ |
| наименование кафедры |

ОТЧЁТ

по учебной: технологической практике 1

Листов 26

Студента Кринкина Олег Алексеевича группы 1ПИб-02-2оп-23

Место прохождения практики

ФГБОУ ВО «Череповецкий государственный университет» кафедра математического и программного обеспечения ЭВМ, компьютерный класс

Студент \_\_\_\_\_\_\_\_\_\_\_\_

(подпись)

Руководитель практики

от кафедры МПО ЭВМ доцент к.т.н. Юдина О. В.

(должность) (подпись, Ф.И.О.)

2024 год

Содержание

[Введение 3](#_Toc185380130)

[Раздел 1. Работа в пакете Orange 10](#_Toc185380131)

[1.1. Задание 1. Работа с выбранным датасетом 10](#_Toc185380132)

[1.2. Задание 2. Регрессия из готового датасета 13](#_Toc185380133)

[Раздел 2. Описание решения оптимизационной задачи в Excel 17](#_Toc185380134)

[2.1. Задание 1. Максимальный выпуск продукции по стоимости 17](#_Toc185380135)

[2.2. Задание 2. Минимальная стоимость сырья 20](#_Toc185380136)

[Заключение 23](#_Toc185380137)

[Список литературы 24](#_Toc185380138)

# Введение

Целью технологической практики является обзорное знакомство с некоторыми пакетами для обработки больших данных. Некоторые из этих пакетов будут использоваться в дальнейшем процессе обучения.

Инструменты для работы с большими данными:

1. KNIME

Knime Analytics Platform – система для анализа данных, позволяющая реализовывать полный цикл анализа данных включающий чтение данных из различных источников, преобразование и фильтрацию, собственно анализ, визуализацию и экспорт [1]. Эта платформа позволяет обрабатывать данные без необходимости в навыках программирования, при этом имея в своём наборе большую библиотеку алгоритмов. Инструмент является свободно распространяемым, при этом компания-разработчик представляет облачные услуги хранения, обработки и автоматизации обработки данных по подписочной модели. Аналитическая платформа имеет открытый исходный код, который доступен на сервисе GitHub [2]. Пакет может работать с различными форматами баз данных, в том числе с различными табличными данными. Результатом работы является логическая цепочка из узлов для обработки данных, на конце которой строится какое-либо из визуальных представлений статистических данных.

1. IBM SPSS

IMB Statistical Package for the Social Sciences это пакет для обработки статистических данных, разработанный в 1968 году внутри чикагского университета [3]. Пакет позволяет анализировать большие объёмы статистических данных, моделировать результат обработки, анализировать различные финансовые ситуации, проводить исследования для принятия оптимальных социальных и организационных решений в больших компаниях и многое другое [4]. Пакет распространяется по подписочной модели с различными уровнями доступа к возможностям пакета. При этом также имеется академический режим, но он также является одним из уровней подписочной модели. Инструмент предлагает различные инструменты для визуализации данных, построения таблиц, деревьев, нахождения мод, регрессии, а также имеется возможность обработки данных при помощи искусственного интеллекта [5]. Для работы с пакетом не требуется умение программировать, но требуются обширные знания в области статистической обработки. Пакет может работать с форматом Excel, .tab, .csv и другими представлениями табличных данных в виде текста [6]. Результатом может являться график, диаграмма, гистограмма и другие виды визуализации данных.

1. SAS

Пакет для обработки статистических данных разработанный компанией SAS Institute в 1972 году. Пакет распространяется по корпоративной подписочной модели, а с 2014 года имеется некоммерческая версия для учебных заведений. В зависимости от уровня подписки в программе доступно свыше 200 дополнений, которые реализуют возможности статистической обработки, визуализации данных, операций исследования, эконометрики, контроля качества, сбора данных, создания скриптов обработки данных и многие другие [7]. Для успешной работы с пакетом потребуются обширные знания во многих прикладных сферах статистики, при этом для базовой обработки данных не требуются навыки программирования. SAS может обрабатывать данные из текста с разделителями (.csv), Excel-таблиц и собственного проприетарного формата [8]. Результатом может являться различные виды визуализации статистических данных или заключение по ним.

1. STATISTICA

Statistica – программный пакет для статистического анализа, разработанный компанией StatSoft в 1984, реализующий функции анализа данных, управления данными, добычи данных, визуализации данных с привлечением статистических методов. Пакет обладает широкими графическими возможностями, позволяет выводить информацию в виде различных типов графиков (включая научные, деловые, трёхмерные и двухмерные графики в различных системах координат, специализированные статистические графики — гистограммы, матричные, категорированные графики и др.), все компоненты графиков настраиваются [9]. Для обработки данных существует возможность написания скриптов на языке R, но людям, что не обладают необходимыми для этого навыками, скорее всего будет достаточно крайне богатого набора инструментов пакета. При этом для успешной работы необходимы глубокие знания и навыки с сфере статистической обработки. Statistica распространяется по подписочной модели длиной в год и выдаётся на одно рабочее место [10], при этом существует лицензия для ВУЗов по заказу и пробная версия на 30 дней [11]. Пакет может работать с данными из .csv, Excel файлов таблиц, а также из множества проприетарных форматов графического представления данных и базами данных SQL. Результатом может являться различные виды визуализации статистических данных, сохранённые в собственных бинарных форматах [12].

1. Minitab

Minitab – программный пакет для обработки статистических данных. Он разработан в Университете штата Пенсильвания в 1972 учёными Барбарой Райан, Томасом Райаном-младшим и Брайаном Джорнером. Продукт разрабатывался как облегчённая версия другого пакета – OMNITAB [13]. Программа распространяется по модели единовременного платежа, при этом имеется бесплатная пробная версия. Также предусмотрена лицензия для образовательных учреждений, но только по запросу [14]. Сегодня Minitab часто используется в сочетании с применением методик «Шесть сигм», CMMI и другими статистическими процедурами, основанными на методах развития. В пакете имеется инструментарий для работы для управления данными и файлами, регрессионного анализа, выборки, многомерного анализа и других. Также отдельно могут быт докуплены модули, расширяющие эти возможности. Для работы с пакетом от пользователя требуются как минимум базовые статистические знания – требование глубины знания зависят от выполняемой статистической задачи. Программа может работать с некоторыми форматами баз данных: ODBC (Open Database Connectivity), DDE (Dynamic Data Exchange) [15]. Результатом работы является некоторое статистическое заключение или прогноз.

1. R + R Studio

R – интерпретируемый язык программирования для статистической обработки данных. Имеет свободную модель распространения и имеет открытый исходный код. Язык был создан в качестве открытой реализации языка S от Bell Labs в 1993 году. Написание программ преимущественно происходит в среде разработки R Studio. Широко используется как статистическое программное обеспечение для анализа данных и фактически стал стандартом для статистических программ. [16] Имеет огромное количество встроенных функций для расчёта. Различного рода значений и решения различного рода задач. Для работы с языком от пользователя потребуется умение программировать, хоть и язык является весьма примитивным [16], а также некоторые статистические знания, в зависимости от задачи. Может работать с любыми данными, которые можно представить в виде чисел, логических значений, комплексных значений [16]. Результатом работы может стать практически что угодно – от сценария для составления статистического прогноза, до более сложной с программы.

1. Python

Python – высокоуровневый мультипарадигменный интерпретируемый язык программирования общего назначения, разработанный 20 февраля 1991 года Гвидо ван Россумом [17]. Интерпретатор распространяется по свободной модели и имеет открытый исходный код [18]. Может использоваться для разработки огромного количества программ, используя при этом практически любую парадигму программирования. Имеет огромное количество встроенных библиотек для решения максимального круга задач, также существует огромное множество сторонних библиотек. Для статистической обработки зачастую используют библиотеки NumPy, SciPy, RPython (интеграция с языком R) или встроенную библиотеку statistics [19]. Сторонние библиотеки предоставляют обширны функции для моделирования практически любой статистической задачи, при этом количество обрабатываемых данных не ограничено. Не смотря на то что базовый синтаксис языка прост, для успешной работы требуются обширные знания в области программирования и статистические знания для решения задач. Данные для работы могут быть представлены любом формате – всё зависит от того, как программистом (пользователем) будет реализовано чтение этих данных. Результат работы также зависит только от самого программиста (пользователя) – в области статистики можно получить от конкретного значения или графика, до полноценного пакета для статистической обработки (прим. Orange).

1. Orange

Orange – программный пакет для визуализации данных, машинного обучения, глубокого анализа данных с открытым исходным кодом. Предоставляет графический интерфейс для визуального программирования работы с данными [20] . Имеет на борту большое количество встроенных инструментов, которые можно комбинировать между собой для решения задачи. Предназначен для обработки большого количества данных для разнообразных целей. Для работы с пакетом потребуются некоторые знания в статистической области (уровнем в зависимости от задачи). Как таковые навыки программирования для работы с пакетом не требуются. Может работать с распространёнными форматами представления табличных данных - .tab, .csv, .xlsx, .dat и др. Результатом работы может стать статистическое заключение, график, таблица.

1. Loginom

Loginom – программный пакет анализа данных за разработкой ООО «Аналитические технологии». Пакет распространяется по модели единоразового платежа. Также имеется бесплатная версия, бесплатная серверная пробная версия и лицензирование для учебных заведений (по договорённости). [21] Пакет предназначен для аналитической обработки большого количества данных, включая обширные функции для статистической обработки. В списке функций пакета имеются множественные алгоритмы и возможности для работы с данными различного бизнес содержания. При этом пакет может быть использован и в различных научных целях. В пример задач можно привести: сегментация рынка, скоринг, стресс-тестирование, таргетинг, управление запасами, управление моделями машинного обучения, прогнозирование и многие другие. [22] Для успешной работы с пакетом не требуются навыки программирования, на что разработчик акцентирует своё внимание. Loginom может работать с различными форматами баз данных: колоночная база данных, концентратор данных, ФИАС, SQL базы и др. [23] Результатом работы может являться любое статистическое заключение: от графика, до прогноза или полноценной модели бизнес решений.

# Раздел 1. Работа в пакете Orange

## 1.1. Задание 1. Работа с выбранным датасетом

1. Создан новый рабочий процесс в Orange (рис. 1).

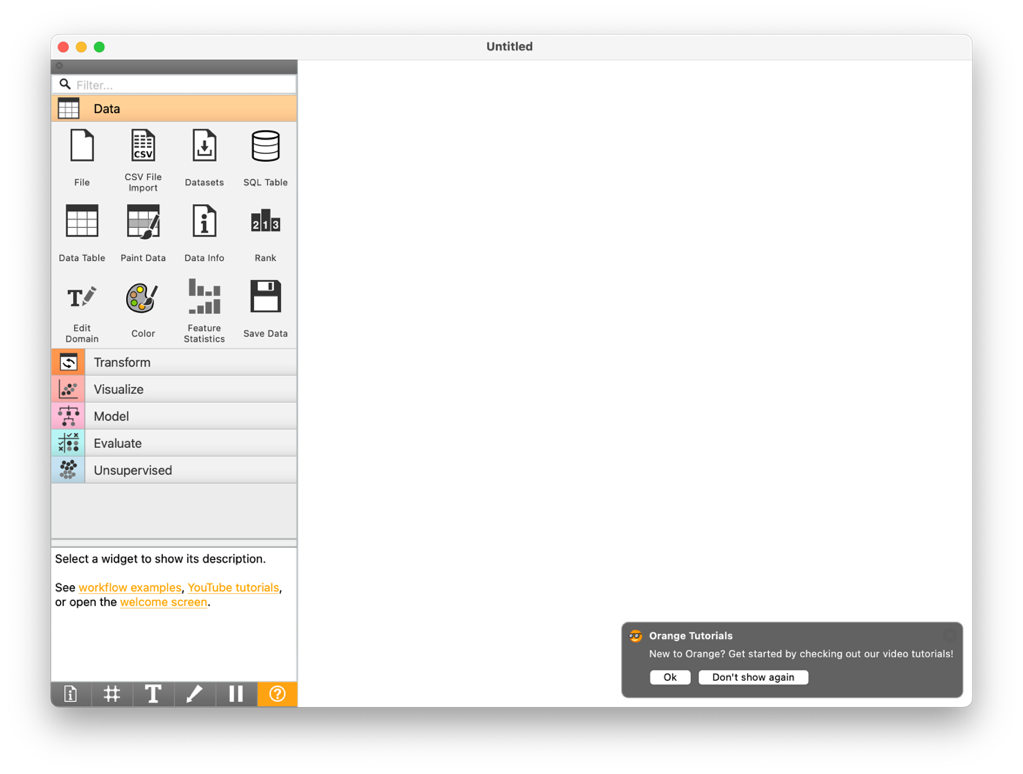


Рис. 1. Пустой рабочий процесс

1. Создан и выбран новый виджет Dataset. Выбранный набор данных – Philadelphia Crime (Преступления в г. Филадельфия) (рис. 2).

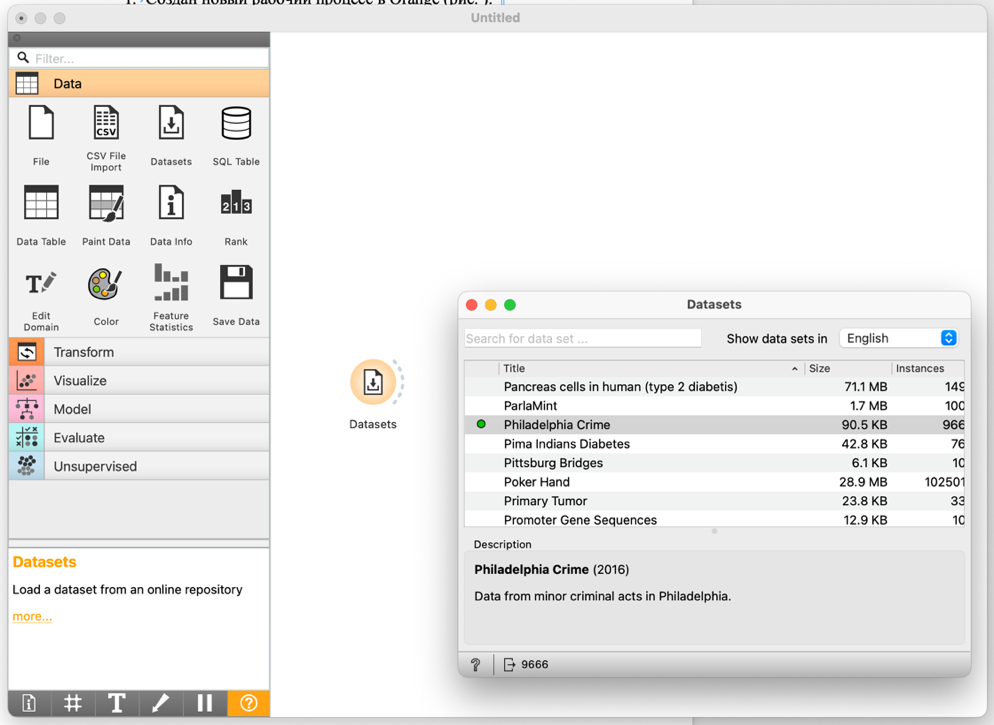


Рис. 2. Окно выбора набора данных

1. Из созданного виджета выведены ещё два виджета – Data Info и Data Table (рис. 3).

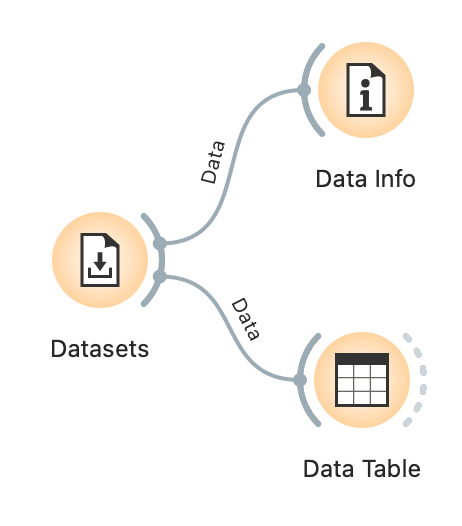


Рис. 3. Виджеты для просмотра данных

В таблице из этого набора данных имеется четыре столбца: дата, преступление, широта и долгота (рис. 4).

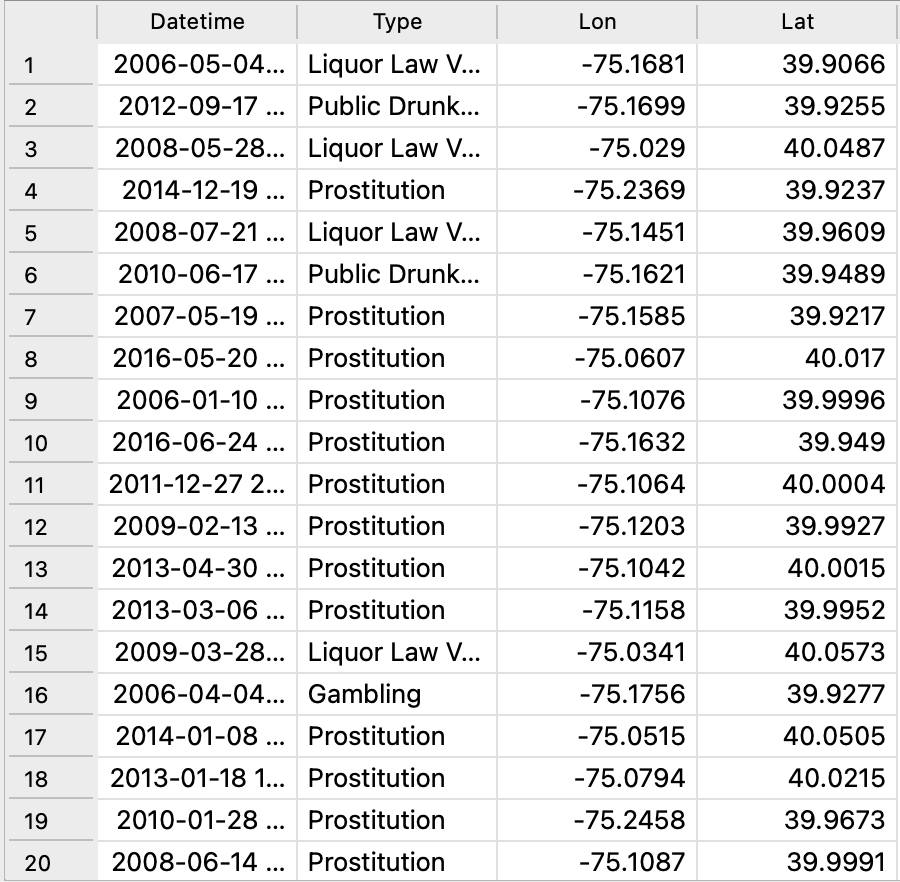


Рис. 4. Часть таблицы с данными

1. Из виджета Datasets создан новый виджет Scatter Plot (рис. 5).

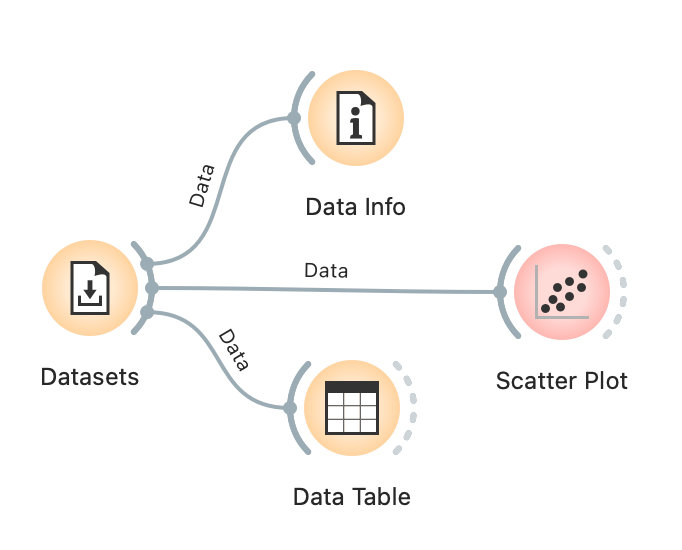


Рис. 5. Добавленный виджет Scatter Plot

Внутри себя виджет предоставляет визуализацию данных в виде строчки точечных единиц по критерию (рис. 6). Для текущих данных в виджете был включен цвет, показывающий местоположение преступления, и был сделан некоторый разброс точек для более удобного считывания данных. Такой вид представления является очень удобным, например из рис. можно выяснить что в восточной части города в период с 2008 по 2010 были нарушения закона о спиртных напитках, что может указывать, например, на нахождение там преступной группировки, занимающейся незаконной продажей алкоголя. Также можно сделать вывод и о том, что в центральной части города распространена проституция.

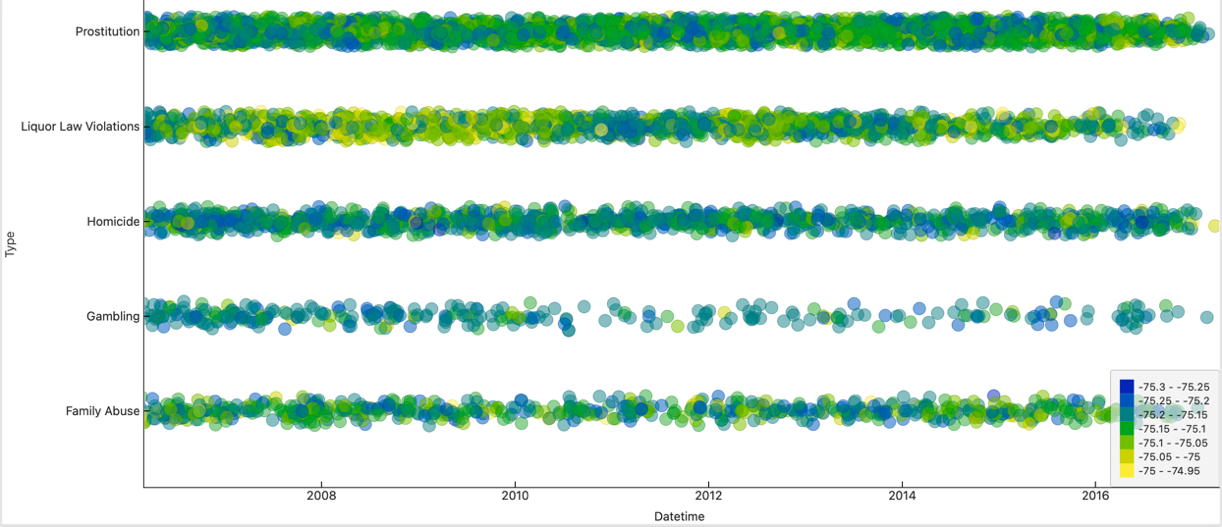


Рис. 6. Визуализация Scatter Plot

## 1.2. Задание 2. Регрессия из готового датасета

1. Из виджета Datasets создан новый виджет Box Plot (рис. 7). Внутри себя он построил процентное соотношение всех мелких преступлений в городе (рис. 8).

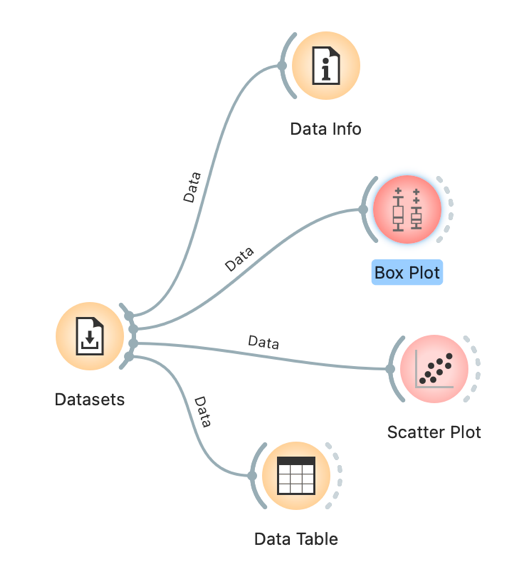


Рис. 7. Добавленный Box Plot

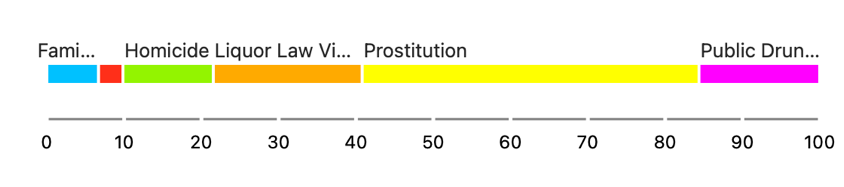


Рис. 8. Распределение в Box Plot

1. Согласно заданию создан новый рабочий процесс и выбран набор данных Housing (жильё в Бостоне). (рис. 9).

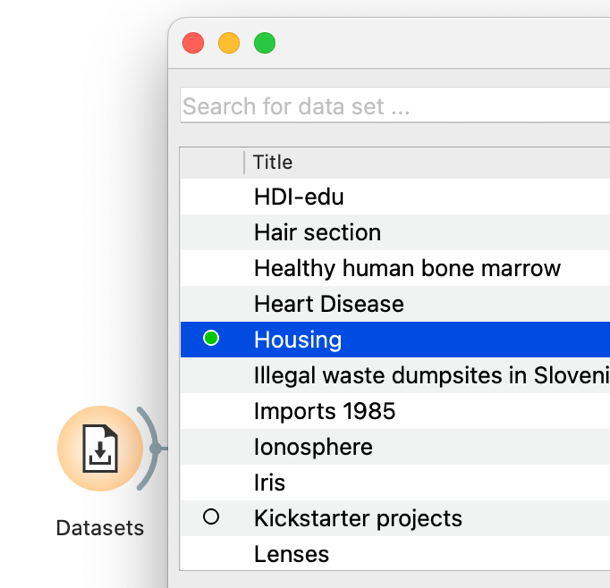


Рис. 9. Установка нового набора данных

1. К набору данных добавлен виджет Data Table к которому подключены виджеты Linear Regression и Test and Score для построения и оценки степени регрессии (рис. 10). Полученные значения представлены на рис. 11.

* MSE – средний квадрат ошибки;
* RMSE – cреднеквадратическое отклонение (квадратный корень из MSE);
* MAE – cредняя абсолютная ошибка;
* MAPE – средняя абсолютная ошибка в процентах;
* R2 – коэффициент детерминации;

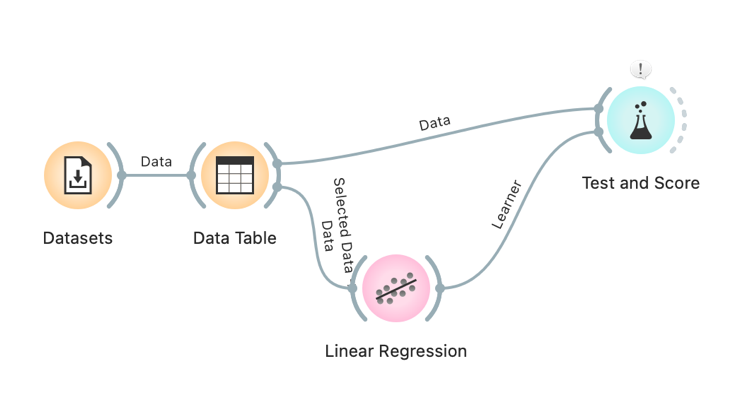


Рис. 10. Виджеты для построения регрессии

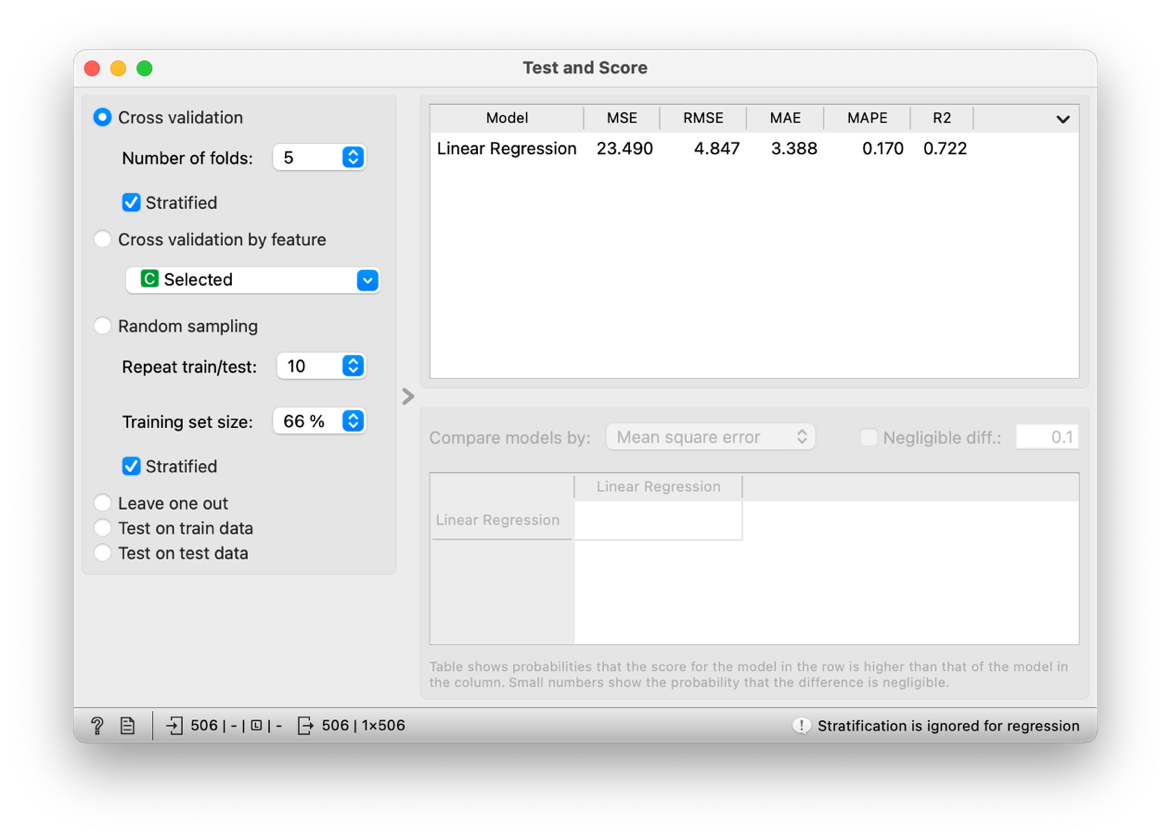


Рис. 11. Результат построения регрессии

1. Для улучшения расчёта результатов к данным был добавлен виджет Preprocessor, который был соединён с выходом Data Table -> Data и со входом Test and Score -> Preprocessor (рис. 12). В настройках препроцессора выбран алгоритм Normalize Features.

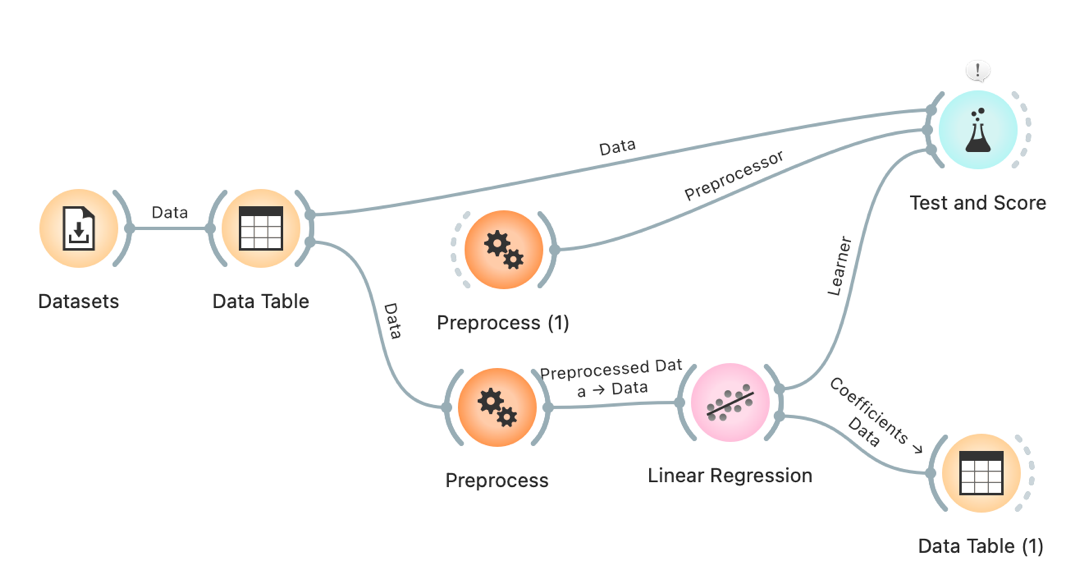


Рис. 12. Процесс с доп. обработками

1. К данным добавлены виджеты Tree, Random Tree и Tree Viewer для просмотра деревьев (рис. 13). Виджет визуализирует данные в виде дерева. В настройках визуализации установлено ограничение значения глубины дерева в 5 поддеревьев (рис. 14). Конечное дерево представлено на рис. 15.

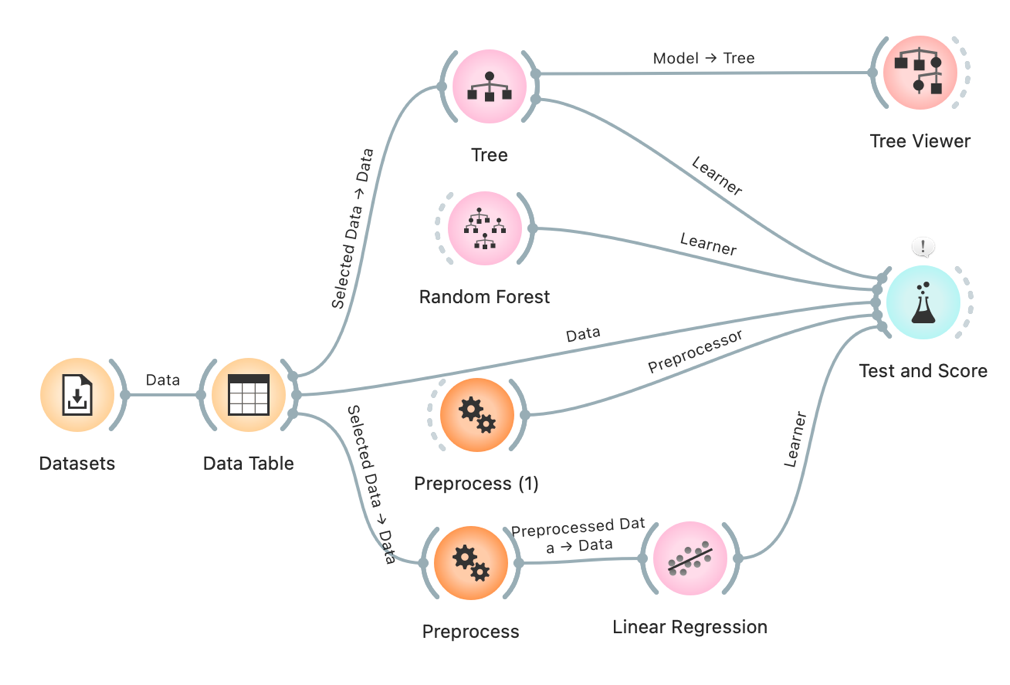


Рис. 13. Процесс с добавленными деревьями

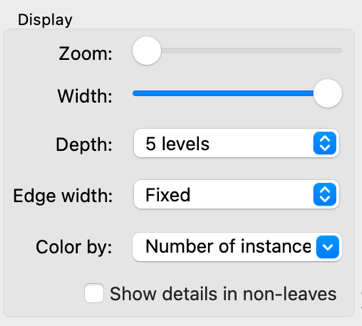


Рис. 14. Настройки отображения дерева

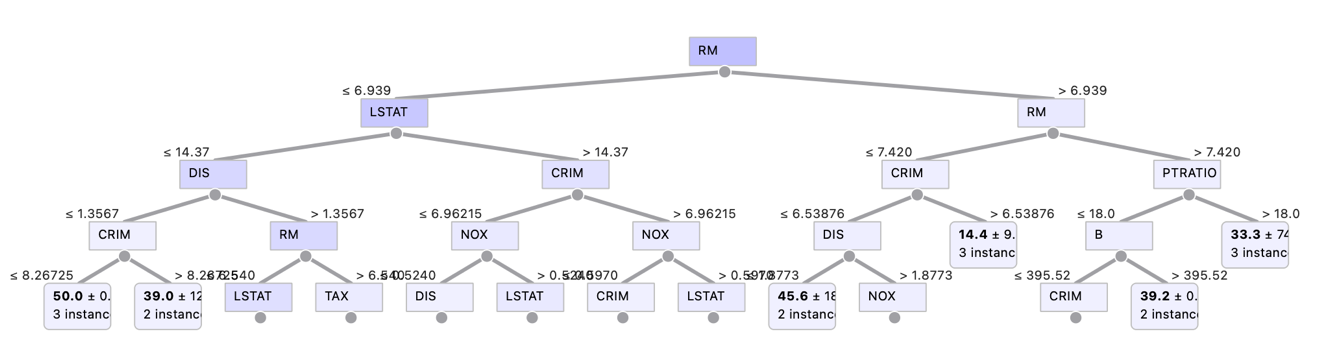


Рис. 15. Визуализация дерева

1. Конечные значения в виджете Test and Score после применения обработки и построения деревьев выглядят следующим образом (рис. 16):

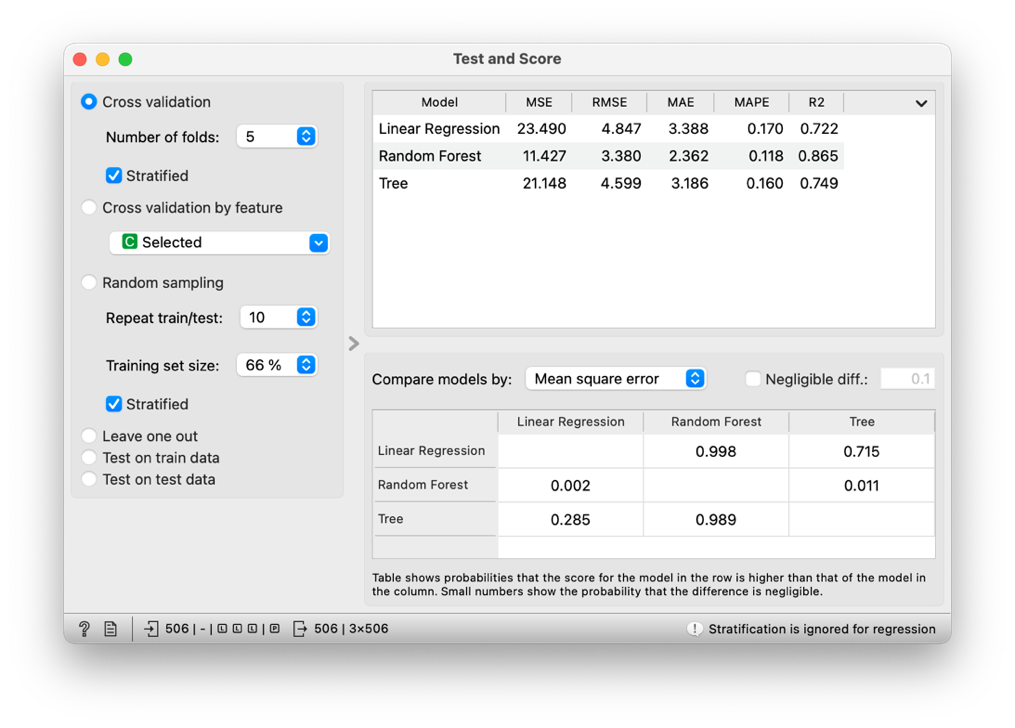


Рис. 16. Конечные результаты регрессии

# Раздел 2. Описание решения оптимизационной задачи в Excel

## 2.1. Задание 1. Максимальный выпуск продукции по стоимости

Для изготовления трёх видов изделий P1, P2 и P3 используют три вида материалов: S1, S2, S3. Запасы материалов, технологические нормы расхода материалов на каждое изделие и цена единицы изделия приведены в таблице 1. Составить план выпуска изделий, обеспечивающих их максимальный выпуск по стоимости.

Таблица 1. Задание 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Вид материала** | **Норма расхода материала на одно изделие, кг** | | | **Запас материала** |
| **P1** | **P2** | **P3** |
| S1 | 18 | 9 | 6 | 540 |
| S2 | 4 | 2 | 4 | 340 |
| S3 | 3 | 3 | 1 | 120 |
| Цена одного изделия (у. е.) | 3 | 4 | 3 | max |

1. Составить математическую модель задачи
2. Решить задачу в Excel
3. Сделать вывод

Для решения задачи составлена следующая математическая модель:

Неизвестные: x1 – количество продукции P1;

x2 – количество продукции P2;

x3 – количество продукции P3.

Ограничения задачи:

Целевая функция:

Для решения задачи в Excel построены следующие таблицы:

* Таблица с исходными данными (рис. 17)

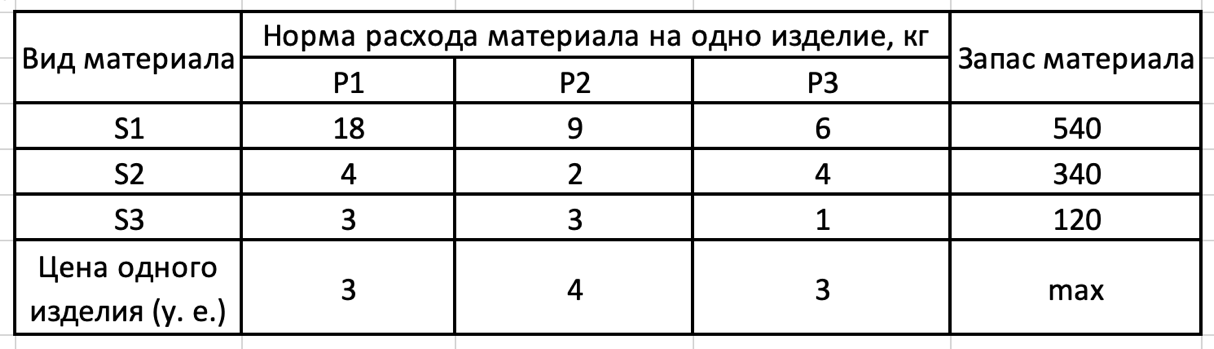


Рис. 17. Таблица с исходными данными

* Рабочая таблица с переменными (рис. 18), в которую были внесены: изначальные значения для x1, x2, x3, в ячейку C12 вписана целевая функция =, где – цена изделия 1, – цена изделия 2, – цена изделия 3, – , – , – , ограничения задачи: в левую часть записаны формулы, в правую часть записано значение ограничения.

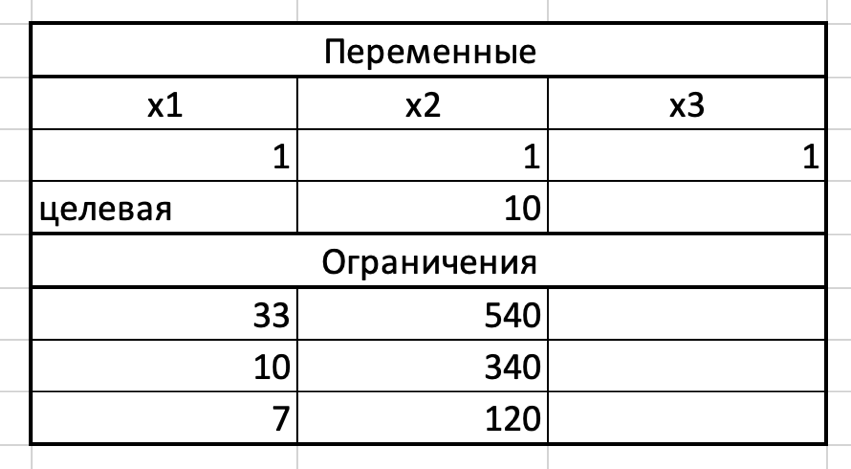


Рис. . Рабочая таблица

* При помощи пункта “Tools->Excel Add-ins...” добавлен инструмент Solver (Поиск решения).
* Для поиска оптимальных значений переменных вызван добавленный инструмент Solver, в окно которого внесены следующие значения (рис. 19):

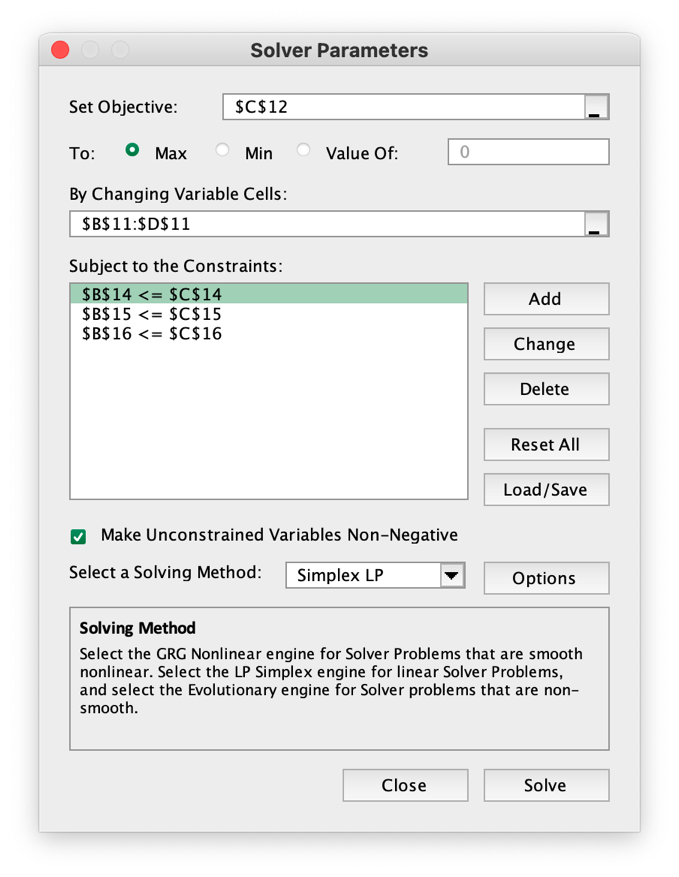


Рис. . Настройки инструмента Solver

Где $C$12 – целевая функция, $B$11:$D$11 – диапазон ячеек с переменными, $B$14 <= $C$14, $B$15 <= $C$15, $B$16 <= $C$16 – пределы задачи.

* В окне инструмента Solver выбран режим решения Simplex LP (Поиск решения линейных задач симплекс-методом) и нажата клавиша Solve (Найти решение)
* В открывшемся окне результатов (рис. 20) выбран пункт Keep Solver Solution (Сохранить найденное решение) и нажата клавиша OK.

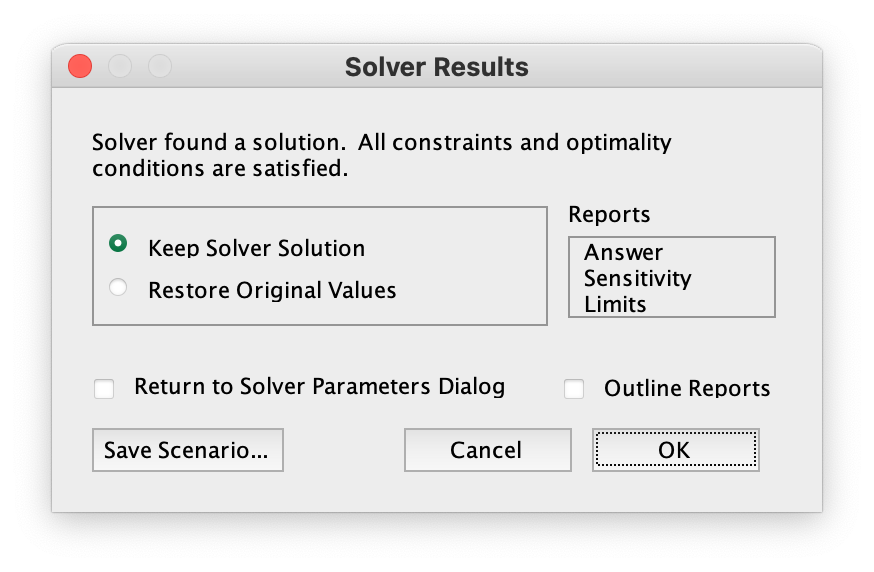


Рис. . Окно результата

* Рабочая таблица приняла следующий вид (рис. 21):

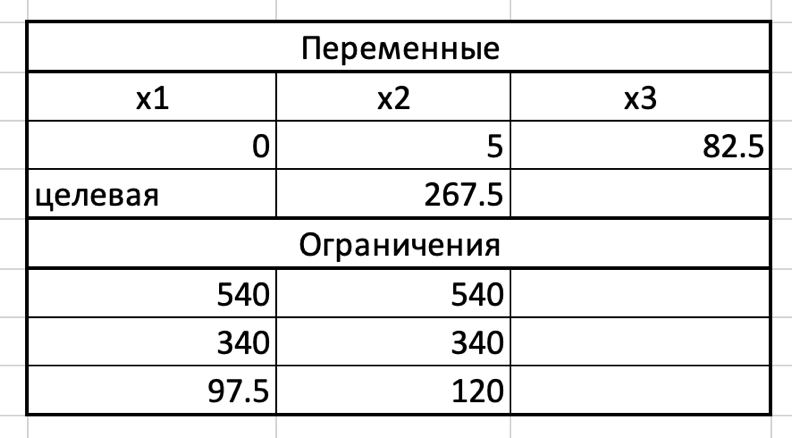


Рис. . Рабочая таблица после Solve

* Среди оптимальных решений x3 – дробное число, для того чтобы решить задачу необходимо округлить значение в меньшую сторону (если округлить в большую сторону – случится нехватка ресурсов). После округления полученного значения до , получено значение целевой функции .

Исходя из полученных данных можно сделать вывод о том, что для получения максимальной прибыли в размере у.е. необходимо изготовить изделий вида , изделия вида , а изделия вида производить не выгодно.

## 2.2. Задание 2. Минимальная стоимость сырья

Требуется определить минимальную по стоимости смесь сырья для изготовления пищевых концентратов, которые должны содержать питательные вещества (П). Эти вещества содержатся в сырье (М) в различных сочетаниях. Содержание питательных веществ в сырье и готовом продукте, а также цена на каждый вид сырья показаны в таблице 2.

Таблица 2. Задание 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Питательные вещества | Виды сырья | | | Минимальное содержание питательных веществ в готовом продукте |
| M1 | M2 | M3 |
| П1 | 4 | 1 | 1 | 130 |
| П2 | 1 | 4 | 3 | 140 |
| П3 | 1 | 1 | 0 | 50 |
| П4 | 3 | 0 | 2 | 80 |
| Цена за единицу сырья | 10 | 8 | 12 | min |

1. Составить математическую модель задачи;
2. Решить задачу в Excel;
3. Сделать вывод.

Задача была тем же методом что и предыдущая (рис. 22), за исключением что в настройках инструмента Solver необходимо установить поиск минимального значения (рис. 23) и изменить знак <= на >=.



Рис. . Построенные для выполнения задания таблицы

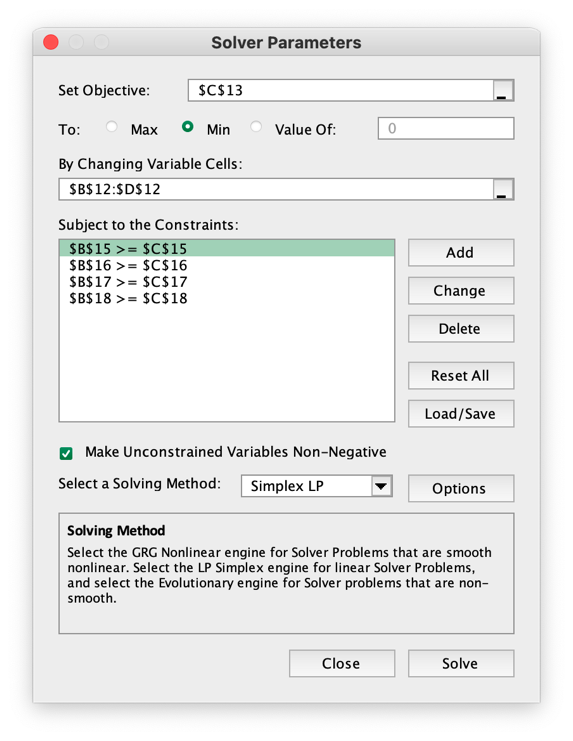


Рис. . Настройки инструмента Solver для задачи

После нажатия кнопки Solve переменные приняли дробные значения. Для решения задачи необходимо округлить значения переменных в большую сторону (рис. 24).



Рис. . Таблица с найденными значениями

По итогу самыми выгодными вариантами сырья являются M1 и M2, а производить пищевые концентраты из сырья M3 не выгодно.

# Заключение

В процессе выполнения работы в программном пакете Orange была выполнена обработка массива данных преступности в городе Филадельфия, США. К обработанным данным построены различные модели их визуализации. К сожалению, большая часть встроенных моделей в пакет была не применима к выбранному набору данных.

После аналогична была выполнена обработка заданного массива данных с вычислением статистического значения регрессии с применением алгоритма корректировки и построение дерева по полученным данным.

Программный пакет Orange показал себя весьма мощным средством для по обработке данных, при этом не требуя от пользователя навыков программирования и имея крайне интуитивный интерфейс.

Также в процессе работы были выполнены задачи по математической оптимизации в программном обеспечении Excel. В результате одна из задач не имела решения, что означает что существуют ситуации, в которых нельзя однозначно найти оптимальный вариант действий.

В процессе выполнения заданий по ознакомительной практике частично были освоены компетенции:

Способен использовать современные информационные технологии и программные средства, в том числе отечественного производства, при решении задач профессиональной деятельности (ОПК-2);

Способен решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности (ОПК-3).

# Список литературы

1. “Обзор Knime Analytics Platform — open source системы для анализа данных” [Электронный ресурс] – Habr. Режим доступа: https://habr.com/ru/articles/320500, свободный – (03.11.2024);
2. Исходный код KNIME [Электронный ресурс] – GitHub. Режим доступа: https://github.com/knime, свободный – (03.11.2024);
3. SPSS [Электронный ресурс] – Википедия. Режим доступа: https://ru.wikipedia.org/wiki/SPSS, свободный – (03.11.2024);
4. «IBM SPSS Statistics» [Электронный ресурс] – официальный сайт IMB. Режим доступа: https://www.ibm.com/products/spss-statistics, свободный – (03.11.2024);
5. «IBM SPSS Statistics GradPack and Faculty Packs» [Электронный ресурс] - официальный сайт IBM. Режим доступа: https://www.ibm.com/products/spss-statistics/gradpack, свободный – (03.11.2024);
6. «Файлы данных» [Электронный ресурс] - документация IBM. Режим доступа: https://www.ibm.com/docs/ru/spss-statistics/beta?topic=system-data-files, свободный – (03.11.2024);
7. «Statistical Analysis System» [Электронный ресурс] - Википедия. Режим доступа: https://en.wikipedia.org/wiki/SAS\_(software) , свободный – (03.11.2024);
8. «Basic Tasks for Data in Reports» [Электронный ресурс] - документация SAS. Режим доступа: https://documentation.sas.com/doc/en/vacdc/v\_029/vareportdata/n0hk5nlzlwj7don19ziq4946pwdn.htm, свободный – (03.11.2024);
9. «Statistica» [Электронный ресурс] - Википедия. Режим доступа: https://ru.wikipedia.org/wiki/Statistica, свободный – (03.11.2024);
10. «Цены на программные продукты компании СтатСофт» [Электронный ресурс] - документация СтатСофт. Режим доступа: https://statsoftai.ru/uploads/prices/prices\_StatSoft.pdf, свободный – (03.11.2024);
11. «СтатСофт» [Электронный ресурс] – официальный сайт СтатСофт. Режим доступа: https://statsoftai.ru, свободный – (03.11.2024);
12. «Краткое руководство ПО СтатСофт» [Электронный ресурс] - документация СтатСофт. Режим доступа: https://statsoftai.ru/uploads/docs/Краткое руководство ПО СтатСофт.pdf, свободный – (03.11.2024);
13. «Minitab» [Электронный ресурс] - Википедия. Режим доступа: https://ru.wikipedia.org/wiki/Minitab, свободный – (20.11.2024);
14. “The Minitab Store” [Электронный ресурс] – официальный цифровой магазин компании. Режим доступа: https://www.minitab.com/en-us/try-buy, свободный – (20.11.2024);
15. “Overview for importing data” [Электронный ресурс] - документация к Minitab. Режим доступа: https://support.minitab.com/en-us/minitab/help-and-how-to/data-input-and-output/open-files-and-import-data/overview, свободный – (20.11.2024);
16. «R (язык программирования)» [Электронный ресурс] - Википедия. Режим доступа: https://ru.wikipedia.org/wiki/R\_(язык\_программирования), свободный – (20.11.2024);
17. “Python” [Электронный ресурс] – Википедия. Режим доступа: https://ru.wikipedia.org/wiki/Python, свободный – (20.11.2024);
18. Репозиторий интерпретатора CPython [Электронный ресурс] – GitHub. Режим доступа: https://github.com/python/cpython, свободный – (20.11.2024);
19. “Mathematical statistics functions” [Электронный ресурс] – Python Docs. Режим доступа: https://docs.python.org/3/library/statistics.html, свободный – (20.11.2024);
20. «Orange (software)» [Электронный ресурс] - Википедия (En). Режим доступа: https://en.wikipedia.org/wiki/Orange\_(software) , свободный – (20.11.2024);
21. Страница загрузки Loginom. [Электронный ресурс] Режим доступа: https://loginom.ru/download, свободный – (20.11.2024);
22. «Бизнес-задачи» [Электронный ресурс] - Loginom Вики. Режим доступа: https://wiki.loginom.ru/business-tasks.html, свободный – (20.11.2024);
23. “Источники данных” [Электронный ресурс] – Loginom Вики. Режим доступа: https://wiki.loginom.ru/data-sources.html, свободный – (20.11.2024);
24. «Load the Solver Add-in in Excel» [Электронный ресурс] - Microsoft Docs. Режим доступа: https://support.microsoft.com/en-us/office/load-the-solver-add-in-in-excel-612926fc-d53b-46b4-872c-e24772f078ca, свободный – (15.12.2024).