

Автоматизация предобработки открытых региональных данных средствами Loginom (на примере рынка труда Еврейской автономной области)

Афанасьева Екатерина Анатольевна

Приамурский государственный университет имени Шолом-Алейхема

Студент

Аннотация

Цель исследования — на примере данных о рынке труда Еврейской автономной области визуально доказать непригодность сырых открытых статистических данных для прямого анализа и разработать автоматизированный сценарий их первичной обработки. Методология включает визуальную диагностику качества данных и построение ETL-сценария (импорт, фильтрация, преобразование типов, очистка) в российской low-code платформе Loginom. Результаты. Выявлены системные артефакты исходного файла (нечисловые значения в числовых полях, критический процент пропусков), которые делают невозможным корректный математический анализ. Применение автоматизированного сценария сокращает время подготовки данных с 45–60 минут до 3–5 минут, обеспечивая воспроизводимость и документированность процесса. Вывод: внедрение автоматизированной первичной обработки является не опцией, а обязательным условием использования открытых государственных данных в управленческой практике.

Ключевые слова: Loginom, ETL, качество открытых данных, визуальная диагностика, предобработка данных, рынок труда, Еврейская автономная область.

Automation of open regional data preprocessing using Loginom (on the example of the labor market of the Jewish Autonomous Region)

Afanasyeva Ekaterina Anatolyevna

Sholom-Aleichem Priamursky State University

Student

Abstract

The study aims to visually prove the unsuitability of raw open statistical data for direct analysis, using the example of labor market data in the Jewish Autonomous Region, and to develop an automated primary processing scenario. Methods. The methodology includes visual data quality diagnostics and building an ETL scenario (import, filtering, data type conversion, cleaning) in the Russian low-code platform Loginom. Results. Systemic artifacts in the source file (non-numeric values in numeric fields, a critical percentage of missing values) were identified, making

correct mathematical analysis impossible. The use of an automated scenario reduces data preparation time from 45–60 minutes to 3–5 minutes, while ensuring reproducibility and process documentation. Conclusions. The implementation of automated primary processing is a mandatory condition for using open government data in management practice.

Keywords: Loginom, ETL, data preprocessing, open data quality, visual diagnostics, labor market, Jewish Autonomous Region.

Введение

Публикация статистических данных в машиночитаемых форматах декларируется как основа открытого государства. На практике файлы, размещаемые на официальных порталах, часто содержат структурные артефакты: объединённые ячейки, служебные строки, текстовое представление чисел, пропуски (обозначенные как «—» или «...») и неконсистентные наименования категорий. Эти дефекты делают невозможной непосредственную загрузку данных в системы бизнес-аналитики (BI) или инструменты машинного обучения.

Для региональных аналитиков эта проблема трансформируется в рутинную ручную обработку, которая снижает оперативность мониторинга и повышает риск человеческих ошибок. Ситуация усугубляется при агрегации данных из нескольких источников: неочищенные пропуски и нестандартизированные ключи гарантируют потерю строк при соединении таблиц и формирование искажённой картины по региону.

Цель данного исследования — на примере открытого файла о рынке труда Еврейской автономной области (ЕАО) продемонстрировать критические проблемы качества исходных данных и предложить автоматизированный подход к их устранению. В качестве инструмента выбрана российская low-code платформа Loginom, предоставляющая встроенные средства визуальной диагностики, что позволяет объективно зафиксировать дефекты, заложенные на этапе публикации данных источником, а не возникшие по вине исследователя.

Вопросам автоматизации обработки данных в государственном управлении посвящено значительное число работ. В работе В. Ф. Турыгиной рассматривается архитектура системы мониторинга региональной экономической динамики, подчёркивается важность потоковой обработки и машинного обучения для оперативного анализа социально-экономических показателей [1]. Отмечается, что базовым этапом любой подобной системы является надёжный конвейер предобработки данных.

С. М. Бекетов, Д. Э. Федяевская, А. Е. Схведиани, С. Г. Редько и Ж.В.Бурлуцкая [2] разрабатывают цифровой инструмент для автоматизации сбора, хранения и обработки данных, особое внимание уделяя проблемам качества и непостоянства структур данных из открытых источников. Сделан вывод о необходимости адаптивных средств синхронизации, что напрямую соотносится с задачами настоящего исследования.

В работе М. А. Травкина, Н. М. Шубича, Н. А. Иванова и Д.С.Мещанинова [3] демонстрируется применение low-code решений для автоматизации обработки и визуализации табличных данных Банка России. Показано, что использование встроенных инструментов позволяет сократить трудоёмкость типового цикла обработки данных с 45 до 7 минут (экономия 84%). Этот результат подтверждает перспективность выбранного подхода.

Отдельного внимания заслуживает исследование Н. Г. Ярушкиной и А.С. Желепова [4], в котором описывается прототип автоматизированной системы поиска и подбора ИТ-команд на основе данных открытых репозиторий. Несмотря на другую предметную область, в работе подчёркивается значимость структурирования и очистки данных для получения достоверных метрик.

В работе Э. Е. Дадашева и М. Д. Самакаевой [5] рассматриваются методы выравнивания динамических рядов при анализе больших данных в цифровой экономике. Авторы акцентируют внимание на необходимости предварительной обработки временных рядов для построения точных прогнозов, что также указывает на важность ETL-этапа.

Таким образом, научное сообщество единодушно признаёт критическую важность этапа первичной обработки данных, однако в публикациях редко приводятся наглядные примеры того, насколько дефектными являются исходные открытые данные государственных служб и как именно визуальная диагностика позволяет это зафиксировать. Настоящее исследование направлено на восполнение этого пробела.

Методология исследования

Объект исследования: открытый статистический файл о рынке труда Еврейской автономной области, опубликованный территориальным органом государственной статистики.

Предмет исследования: процесс визуальной диагностики качества данных и их первичной автоматизированной обработки.

Исходные данные: файл XLSX Фонд начисленной заработной платы работников списочного состава и внешних совместителей, загруженный с официального открытого портала [6].

Инструментарий: российская low-code платформа Loginom Community (версия 7.x), предоставляющая визуальный конструктор ETL-сценариев и встроенные средства диагностики качества данных.

Этапы исследования:

- Импорт файла в Loginom и первичная визуальная диагностика качества.
- Детальный анализ качества полей с использованием встроенных средств платформы.
- Построение автоматизированного сценария очистки данных.
- Выполнение сценария и оценка эффективности.

Результаты исследования

Этап 1. Импорт и первичная диагностика: фиксация проблем качества

В интерфейсе Loginom Community создан новый сценарий. Добавлен узел «Excel файл» из категории «Импорт». В настройках указан путь к файлу XLSX, выбран лист с данными, настроен пропуск первых двух служебных строк, которые источник публикации разместил в качестве заголовка (рисунок 1).

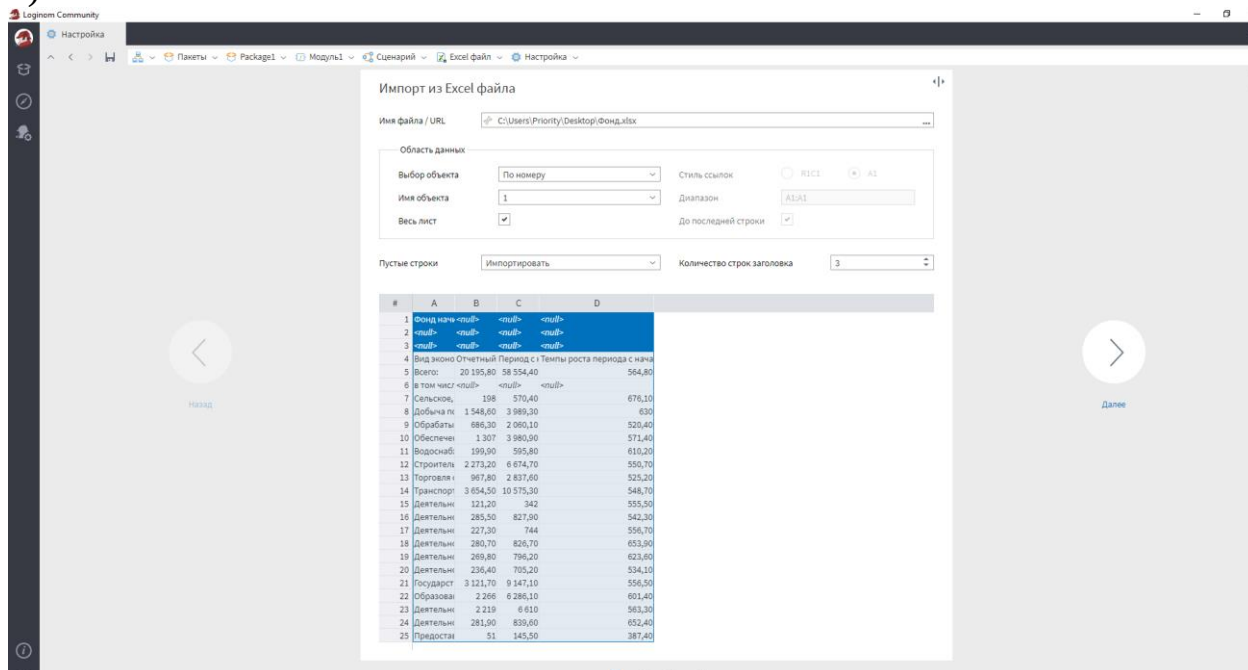


Рисунок 1 – Настройка узла импорта данных из файла XLSX

После выполнения узла импорта Loginom автоматически формирует краткую характеристику набора данных, которая является первым индикатором его пригодности для анализа. В этой характеристике отражены:

- общее количество строк и столбцов;
- доля заполненных полей (percentage of filled fields);
- индекс EPV (Expected Predictive Value) — интегральная метрика качества данных (рисунок 2).

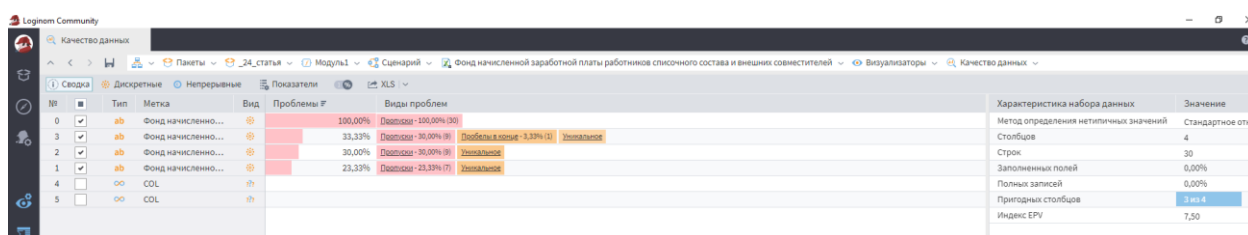


Рисунок 2 – Краткая характеристика набора данных: количество строк, столбцов, доля заполненных полей и индекс EPV

Уже на этом этапе визуальная диагностика объективно фиксирует наличие критических проблем качества: в числовых столбцах присутствуют нечисловые значения (прочерки «—», троеточия «...»), числа сохранены как текст с пробелами-разделителями разрядов, в наименованиях отраслей встречаются лишние пробелы и неконсистентный регистр. Эти артефакты

были заложены источником публикации и не являются следствием действий исследователя.

Этап 2. Детальный анализ: доказательство непригодности данных для анализа

Для углублённой диагностики качества данных выполнен переход к просмотру обзора качества дискретных полей (кнопка «Дискретные» в интерфейсе Loginom). Данный режим позволяет визуализировать распределение значений, выявить аномалии и пропуски в категориальных и числовых переменных (рисунок 3).

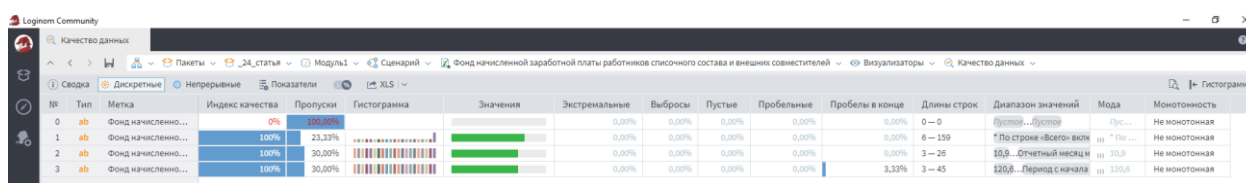


Рисунок 3 – Обзор качества дискретных полей

При нажатии кнопки «Детализация» для конкретного поля выявлено, что столбец «Фонд начисленной заработной платы работников списочного состава и внешних совместителей» содержит критически высокий процент пустых значений (рисунок 4).

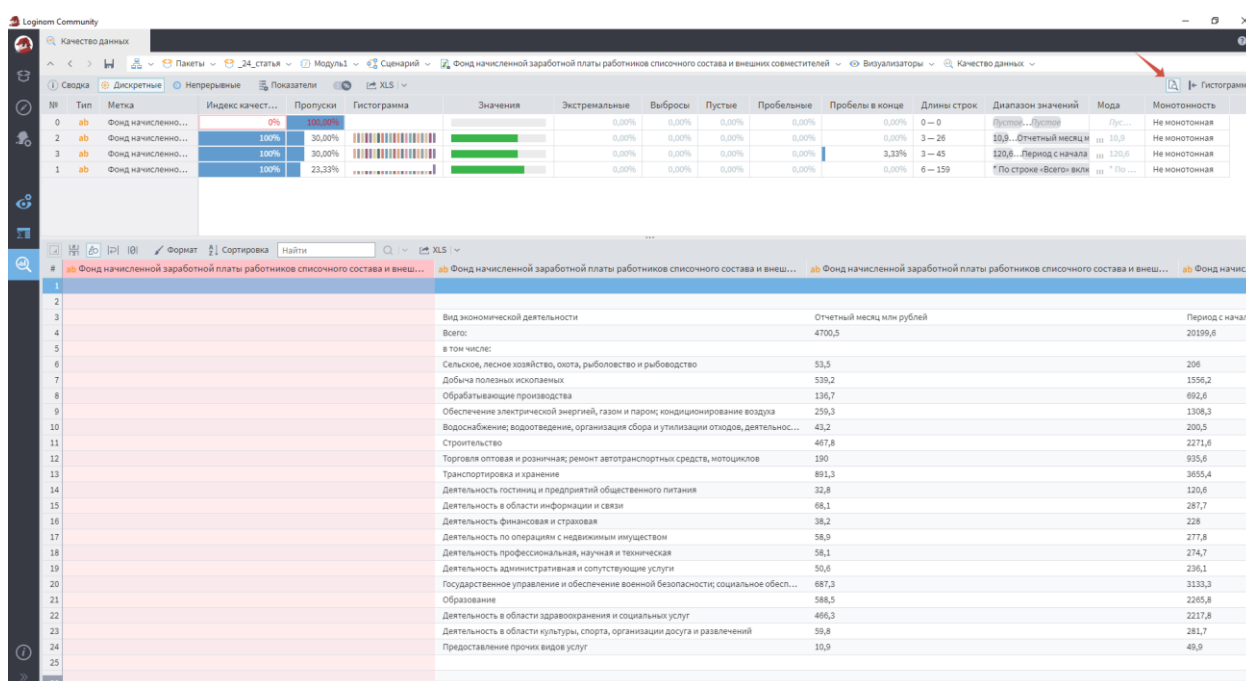


Рисунок 4 – Детализация пустых значений в столбце «Фонд начисленной заработной платы»

Этот факт объективно свидетельствует о непригодности исходных открытых данных для непосредственной обработки в аналитических программах. Наличие таких артефактов создаёт следующие проблемы:

- Невозможность корректного математического анализа. Функции агрегации (сумма, среднее, максимум, минимум) дают ошибочные результаты или полностью отказывают при наличии нечисловых значений и пропусков.
- Усложнение и удлинение процесса обработки. Требуется дополнительная рутинная ручная работа по заполнению или удалению пропусков, преобразованию типов и стандартизации наименований.
- Риск получения недостоверных выводов. Если бы обработке подвергался только этот один файл, проблемы могли бы остаться незамеченными до этапа интерпретации. Однако при условии последующей выборки данных для всей области и их объединения с другими файлами по ключу «Вид экономической деятельности», наличие пропусков приводит к потере строк при внутреннем соединении и формированию неполной и ложной картины рынка труда региона.

Дополнительно обратим внимание на поле «Значение»: здесь построены диаграммы, где цветом обозначены уникальные единичные значения по строкам. Как видно, значения уникальны, что подтверждает отсутствие скрытых дубликатов на данном этапе, но наличие структурных проблем качества остаётся критическим (рисунок 5).

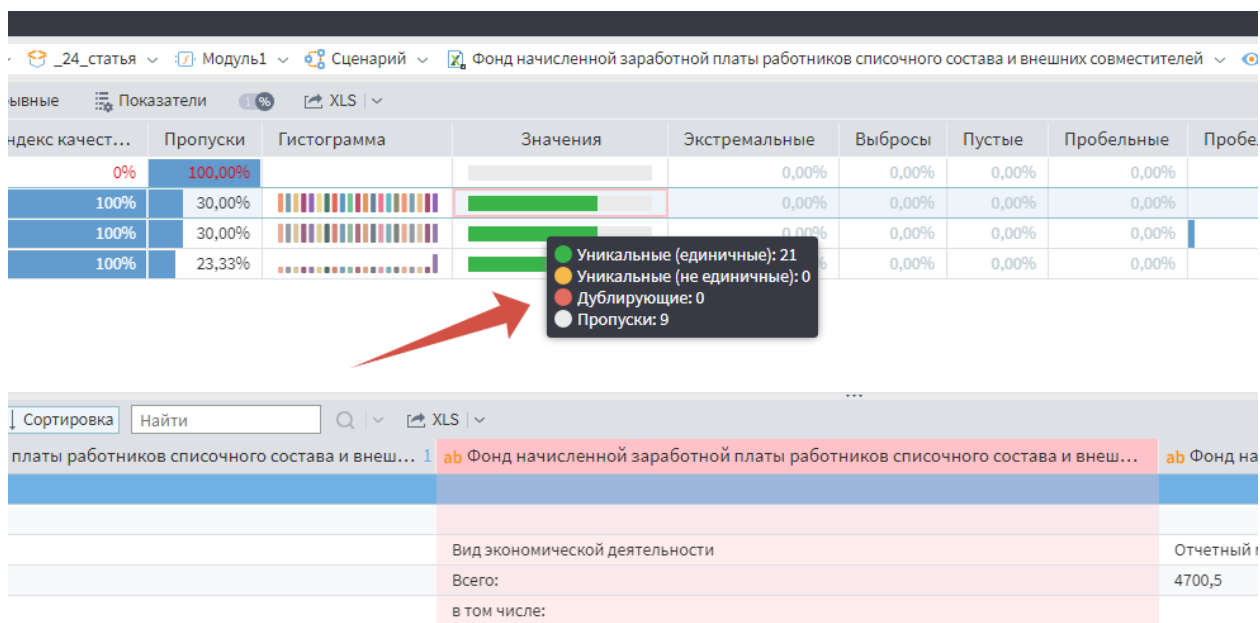


Рисунок 5 – Визуализация распределения уникальных значений по строкам

Таким образом, визуальная диагностика в Logiот объективно доказывает: исходные открытые данные государственной службы непригодны для анализа без первичной обработки. Это не субъективное мнение исследователя, а зафиксированный инструментально факт.

Этап 3. Автоматизированная первичная обработка

На основе выявленных проблем построен автоматизированный ETL-сценарий, который устраняет все зафиксированные артефакты. Сценарий включает следующие узлы:

3.1. Фильтрация строк (Row Filter). Удалены строки, где значение столбца «Вид экономической деятельности» равно «Итого» или является пустым. Это позволило исключить служебные строки, не участвующие в поотраслевом анализе (рисунок 6).

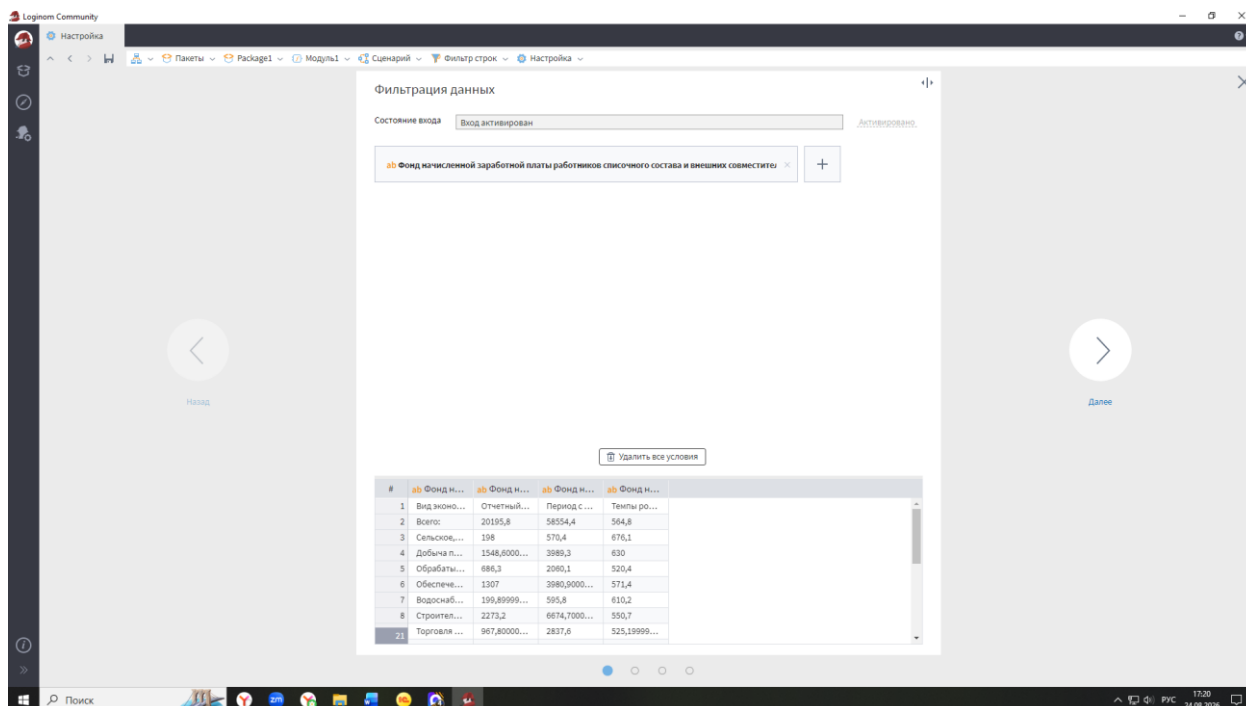


Рисунок 6 – Настройка узла «Фильтр строк»

3.2. Преобразование типов и обработка пропусков (Параметры полей). Для всех столбцов с числовыми данными тип изменён с «Дискретный» на «Непрерывный». При преобразовании LogiDom автоматически распознал нечисловые значения (прочерки «—», троеточия «...») и заменил их на пропуски (NULL), что обеспечило корректную обработку без необходимости создания отдельного узла замены (рисунок 7).

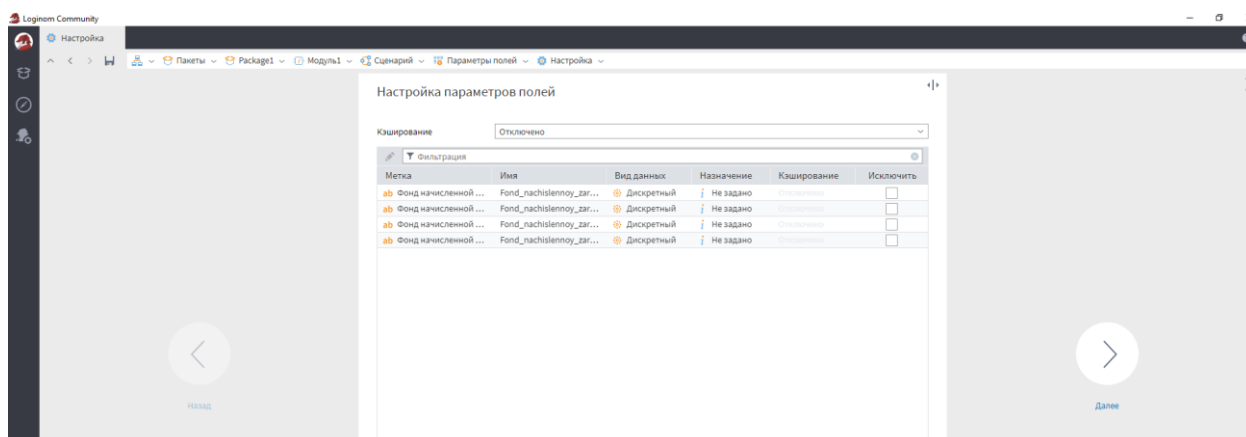


Рисунок 7 – Настройка узла «Параметры полей»

3.4. Экспорт результатов и общий вид сценария

Финальным узлом сценария является узел экспорта (на скриншоте обозначен как «Excel файл», также может использоваться узел «Текстовый файл» для формата CSV). В настройках узла указан формат выходного файла, кодировка UTF-8 и путь сохранения.

На рисунке 8 представлен общий вид построенного автоматизированного сценария предобработки данных в Loginom, включающего все рассмотренные выше узлы: импорт исходного файла, фильтрацию строк, преобразование типов данных с автоматической обработкой пропусков и экспорт очищенных данных (рисунок 8).

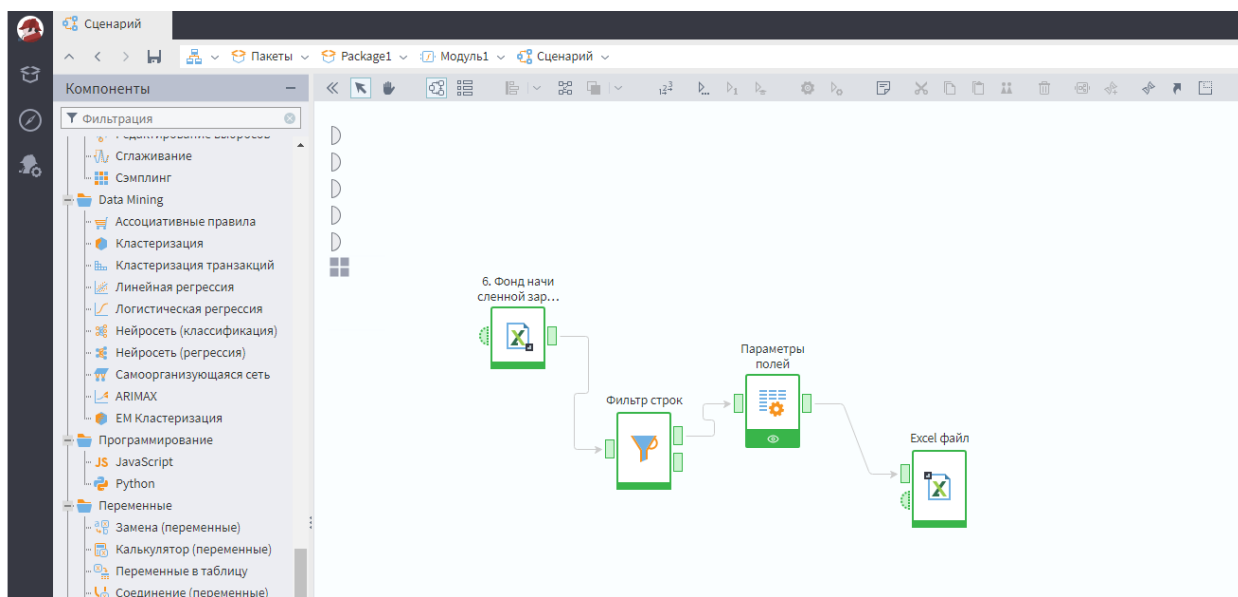


Рисунок 8 – Общий вид автоматизированного сценария предобработки данных в Loginom (импорт → фильтрация → преобразование → экспорт)

Этап 4. Оценка эффективности

Для оценки эффективности разработанного сценария был проведен экспериментальный хронометраж: время ручной обработки (45–60 минут) зафиксировано при типовой очистке аналогичного файла в MS Excel аналитиком, а время автоматизированной обработки (3–5 минут) включает запуск сценария в Loginom и экспорт результатов.

Таблица 1 - Количественные показатели эффективности первичной обработки

Показатель	Ручная обработка в Excel	Автоматизированная обработка в Loginom
Время обработки одного файла	45–60 минут	3–5 минут
Воспроизводимость	Низкая (зависит от исполнителя)	Высокая (сценарий выполняется идентично)
Риск человеческой ошибки	Высокий	Минимальный
Документированность процесса	Отсутствует	Полная (визуальный сценарий)
Сокращение трудозатрат	—	в 12 раз (экономия 92%)

Полученный результат сопоставим с данными работы М.А. Травкина и соавторов [3], где экономия при автоматизации обработки данных составила 84%, что подтверждает высокую эффективность low-code подхода к первичной обработке открытых данных

Обсуждение результатов

Проведённое исследование позволило сделать несколько важных выводов о состоянии открытых региональных данных и подходах к их обработке.

Во-первых, проблема низкого качества открытых данных носит системный характер. Артефакты, выявленные в файле территориального органа статистики ЕАО (служебные строки, нечисловые значения в числовых полях, текстовое представление чисел, избыточные пробелы, высокий процент пропусков), типичны для открытых данных государственных служб. Это означает, что любой аналитик, работающий с подобными данными, неизбежно столкнётся с необходимостью их первичной обработки.

Во-вторых, визуальная диагностика качества данных средствами Loginom позволяет объективно зафиксировать проблемы исходного файла. Это важно не только для самого исследователя, но и для коммуникации с заказчиком: наглядные отчёты о качестве данных доказывают, что проблемы заложены на этапе публикации источником, а не являются следствием некорректных действий аналитика.

В-третьих, автоматизация первичной обработки обеспечивает не только сокращение трудозатрат, но и воспроизводимость, документированность процесса, а также минимизацию риска человеческой ошибки. Разработанный сценарий может быть многократно применён к аналогичным файлам за другие периоды или к файлам других регионов с минимальной адаптацией.

Ограничения исследования:

1. Сценарий ориентирован на конкретную структуру файла Росстата; при изменении формата публикаций потребуется адаптация узлов.
2. Loginom Community имеет ограничения по объёму обрабатываемых данных (до 1 млн строк), что, однако, достаточно для задач региональной статистики.
3. Отсутствует возможность прямой публикации результатов в веб-интерфейс (требуется экспорт в файлы или интеграция с BI-системами).

Заключение

В ходе исследования на примере открытого файла о рынке труда Еврейской автономной области продемонстрировано, что открытые статистические данные, публикуемые государственными и территориальными органами статистики, в большинстве случаев непригодны для прямого аналитического использования. Визуальная диагностика средствами платформы Loginom объективно зафиксировала критические проблемы качества исходного файла: наличие служебных строк, нечисловых значений в числовых полях, текстового представления чисел, избыточных пробелов и

высокого процента пустых значений. Эти артефакты заложены на этапе публикации данных источником и не являются следствием действий исследователя.

Разработанный автоматизированный ETL-сценарий первичной обработки позволил устранить все выявленные артефакты и привести данные в вид, пригодный для дальнейшего анализа. Применение сценария сократило трудозатраты на подготовку данных в 12 раз (с 45–60 до 3–5 минут) при одновременном повышении воспроизводимости и достоверности результата.

Главный вывод исследования: внедрение автоматизированных сценариев первичной обработки открытых данных является не опциональным улучшением, а обязательным условием их использования в аналитической и управленческой практике. Без такого этапа любые аналитические выводы, построенные на открытых государственных данных, рискуют быть недостоверными.

Направления дальнейших исследований

- Разработка адаптивных механизмов обработки изменяющихся форматов данных Росстата.
- Создание репозитория типовых ETL-сценариев первичной обработки для различных видов открытых региональных данных.
- Исследование возможностей применения машинного обучения для автоматического выявления и исправления артефактов в открытых данных.

Библиографический список

1. Турыгина В. Ф. Архитектура системы мониторинга региональной экономической динамики на основе потоковой обработки данных и машинного обучения //Новое в экономической кибернетике. 2026. №. 1. С. 112-127. URL: <https://dongu-nec.ru/index.php/nec/article/view/506>
2. Бекетов С. М. и др. Цифровой инструмент автоматизации процессов сбора, хранения и обработки данных об инновационном развитии регионов //Экономика. Информатика. 2024. Т. 51. №. 3. С. 735-748. URL: <https://cyberleninka.ru/article/n/tsifrovoy-instrument-avtomatizatsii-protsessov-sbora-hraneniya-i-obrabotki-dannyh-ob-innovatsionnom-razviti-regionov>
3. Травкин М. А. . Применение low-code решений для автоматизации обработки и визуализации табличных данных Банка России //Теория и практика общественного развития. 2026. №. 5. С. 249-258. URL: <https://cyberleninka.ru/article/n/primenenie-low-code-resheniy-dlya-avtomatizatsii-obrabotki-i-vizualizatsii-tablichnyh-dannyh-banka-rossii>
4. Ярушкина Н. Г., Желепов А. С. Прототип системы поиска и выбора "сформированных" команд ИТ-специалистов на основе данных проектных репозиториях //Автоматизация процессов управления. 2020. №. 1. С. 97-105. URL: https://www.researchgate.net/profile/Nadejda-Yarushkina/publication/341483193_A_PROTOTYPE_OF_A_SYSTEM_FOR_SEARCHING_AND_SELECTING_THE_'STAFFED'_TEAMS_OF_IT-

PROFESSIONALS_BASED_ON_DATA_FROM_PROJECT_REPOSITORIE
S/links/5ed4db0c299bf1c67d3229a0/A-PROTOTYPE-OF-A-SYSTEM-FOR-
SEARCHNING-AND-SELECTING-THE-STAFFED-TEAMS-OF-IT-
PROFESSIONALS-BASED-ON-DATA-FROM-PROJECT-
REPOSITORIES.pdf

5. Дадашев Э. Е., Самакаева М. Д. Выравнивание динамических рядов в аналитике больших данных в цифровой экономике // Big Data и анализ высокого уровня: сборник научных статей XII Международной научно-практической конференции, Республика Беларусь, Минск, 23 апреля 2026 года: в 2 ч. Ч. 1 / Белорусский государственный университет информатики и радиоэлектроники [и др.] ; редкол.: В. А. Богуш [и др.]. Минск, 2026. С. 107–113. URL: <https://libeldoc.bsuir.by/handle/123456789/63554>
6. Фонд начисленной заработной платы работников списочного состава и внешних совместителей. URL: [https://docs.yandex.ru/docs/view?url=ya-browser%3A%2F%2F4DT1uXEPRrJRXlUFoewruMmfsvD_MPzB-lrkwwJTqT3wvmbiYzz22IbMD89SGfSqVEsCKklppR1CXRqNOt50uNy26wen-ZmcZ-jXaWXaurfyYxidG7-GF4j2ulETUY29z1IlCJACUkZwO1z2U3QurQ%3D%3D%3Fsign%3DWXpn7a-1fza6XZb2CUXv2guOzA2-L3rs1PaBAW82WBQ%3D&name=Фонд%20начисленной%20заработной%20платы%20работников%20списочного%20состава%20и%20внешних%20совместителей\(33\).xlsx&nosw=1](https://docs.yandex.ru/docs/view?url=ya-browser%3A%2F%2F4DT1uXEPRrJRXlUFoewruMmfsvD_MPzB-lrkwwJTqT3wvmbiYzz22IbMD89SGfSqVEsCKklppR1CXRqNOt50uNy26wen-ZmcZ-jXaWXaurfyYxidG7-GF4j2ulETUY29z1IlCJACUkZwO1z2U3QurQ%3D%3D%3Fsign%3DWXpn7a-1fza6XZb2CUXv2guOzA2-L3rs1PaBAW82WBQ%3D&name=Фонд%20начисленной%20заработной%20платы%20работников%20списочного%20состава%20и%20внешних%20совместителей(33).xlsx&nosw=1)