Департамент образования и науки города Москвы

Государственное автономное образовательное учреждение высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики, управления и технологий

ДИСЦИПЛИНА:

Проектный практикум по разработке ETL-решений

**Лабораторная работа 1.1**

**Установка и настройка ETL-инструмента. Создание конвейеров данных**

Выполнила: Сергеева А. И., группа: АДЭУ-211

Преподаватель: Босенко Т.М.

Москва

2025

**Цель работы:** изучение основных принципов работы с ETL-инструментами на примере Pentaho Data Integration (PDI), настройка конвейера обработки данных, фильтрация и замена значений в Excel-файле, а также выгрузка обработанных данных в базу данных MySQL/PostgreSQL.

**Ход работы:**

**Вариант 12. Анализ рынка недвижимости: очистка и трансформация данных.** Для начала данные были загружены через объект CSV Input. CSV Input в Pentaho — это шаг для чтения данных из файла с разделителями, как показано на рисунке 1.

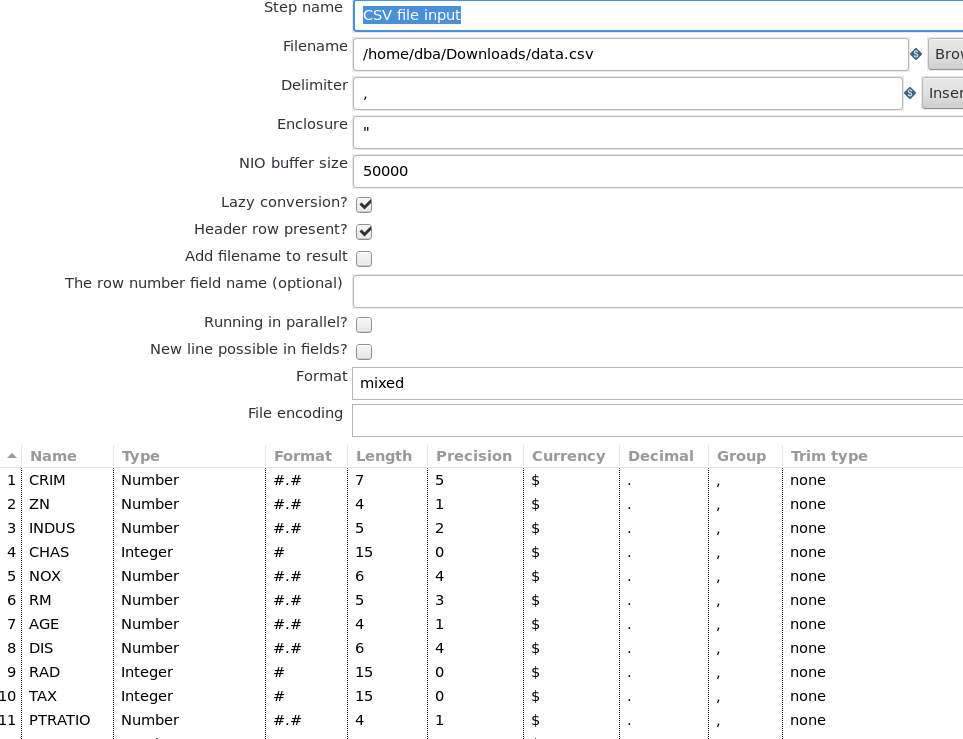


Рисунок 1 – Настройка объекта CSV Input

Следующим шагом является преобразование типов данных для столбцов, для этого используется объект Select Values на рисунке 2. В процессе работы над файлом было выявлено, что в столбце RM содержатся значения в формате 6.76Е+00, такое число одно, такие форматы не читаются, как нужно, поэтому было принято решение почистить число с помощью регулярных выражений до целого, хоть число и одно, но все равно был создан объект, очищающий подобные случаи. Данный столбец характеризует среднее число комнат, комнат 6.76, но почему-то было подставлено Е+00, которое по итогу все равно умножает на 1 и не меняет значения.

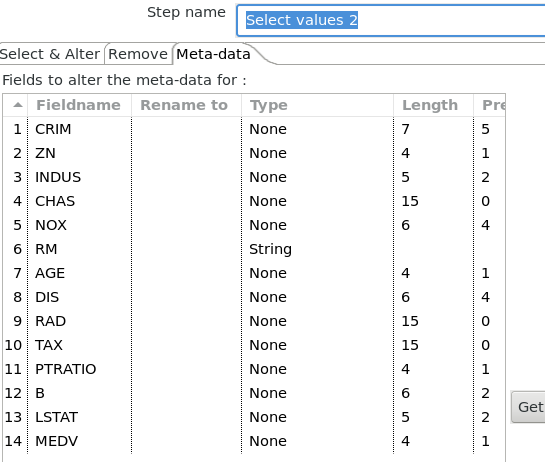


Рисунок 2 – Преобразование столбца RM в строковый тип данных

Select Values в Pentaho — это шаг для выбора, удаления, переименования, изменения типов данных и настройки длины и точности полей в потоке. После изменения типа данных можно работать уже с регулярными выражениями, как показано на рисунке 3.

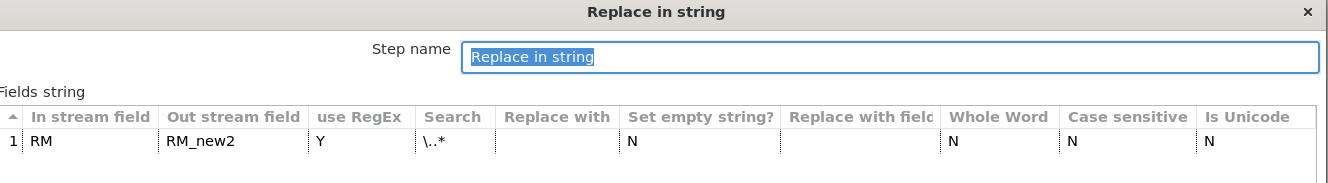


Рисунок 3 – Настройка объекта Replace in string

Replace in string позволяет изменять значения в столбцах строкового типа данных, задавать изменения с помощью регулярных выражений, данное выражение убирает знаки после запятой, в том числе и случаи, когда умножение происходит на 10 в первой степени, а именно (E+00), что итак не меняет цифру. Теперь после изменения значений в столбце на среднее число комнат необходимо было изменить типы других столбцов, а столбцу RM задать тип integer. Налог (TAX) целочисленный тип, потому что налог всегда исчисляется в полных рублях, а нахождение у реки Чарльз (CHAS) также, ведь принимает только два значения: 0 и 1. Другие столбцы настроены аналогично на рисунке 4.

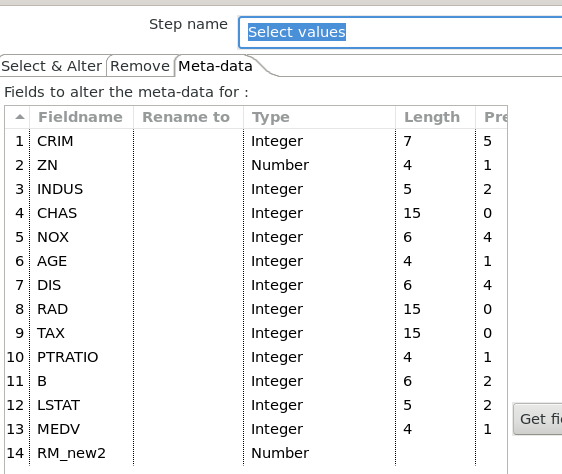


Рисунок 4 – Настройка типов данных для остальных столбцов

Следующим шагом была выполнена фильтрация, которая отсеивает нулевые значения там, где их можно отбросить, если нет, то выполняются другие способы, например, замена средним значением, как продемонстрировано на рисунке 4.

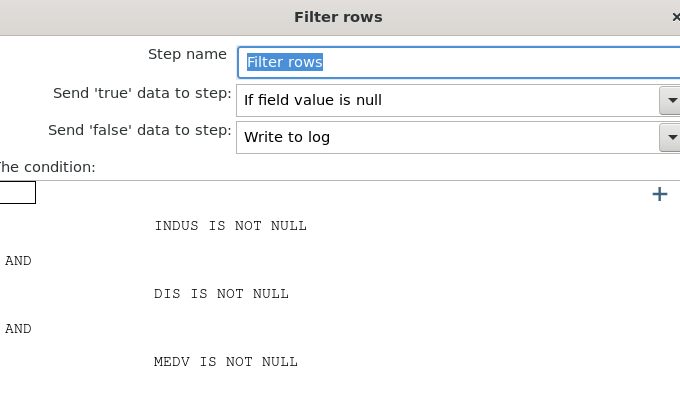


Рисунок 4 – Выполнение фильтрации для отсеивания нулевых значений

Доля неторговых бизнес-точек может колебаться и заменять ее на среднее тоже не совсем корректно, хотя этот показатель важен для оценки развития территории в сфере услуг, производства и технологий, поэтому строки пустые по этому полю лучше убрать. Расстояние не может быть нулевым до трудовых точек, тоже было принято решение убрать такие строки, стоимость домов также важна и при заполнении значений средними может сильно ухудшиться анализ, лучше убрать. Для других столбцов предложены другие методы. Уровень преступности может быть заменен средним, это не исказит выборку, а может быть наоборот хорошим приближением. Доля земли под участки может быть заменена медианой, которая устойчива к выбросам, ведь площади могут быть и чрезмерно маленькими, и чрезмерно большими, расположение у реки имеет только значение 0 и 1, поэтому может быть заменено модой, концентрация оксидов азота заменена средними, комнаты заменены медианой, т. к. тоже могут быть выбросы, доля единиц, занимаемых владельцами, заменена средними, индекс доступности радиальных шоссе может быть заменен модой, ставка налога средним, отношение учеников к учителю заменено средними, доля чернокожего населения средними, процент населения с низкими доходами средним, все видно на рисунках 5-8, а именно вычисления среднего и медианы с помощью Group by и расчеты отдельно для моды. В рамках данного объекта необязательно указывать поля для группировки, но можно выполнять операции агрегирования.

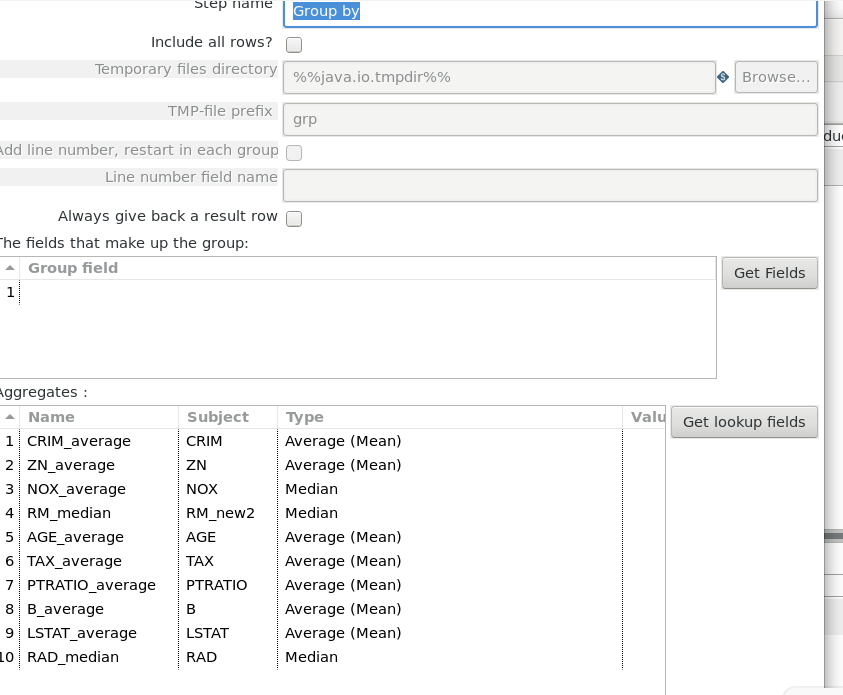


Рисунок 5 – Расчет средних значений и медианы

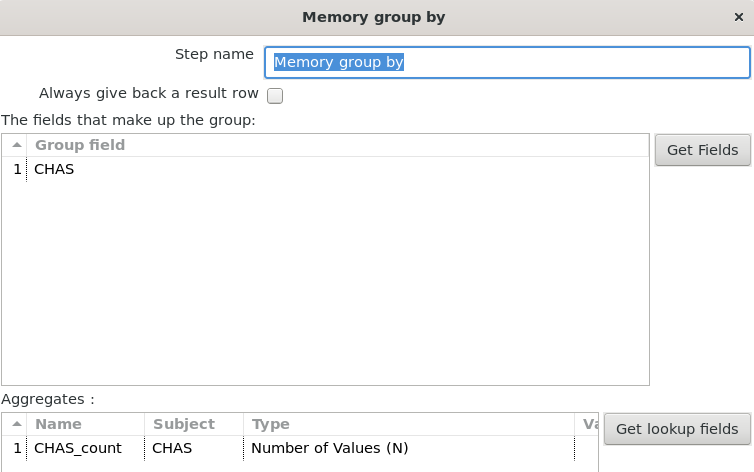


Рисунок 6 – Группировка по полю нахождения у реки и подсчет количества

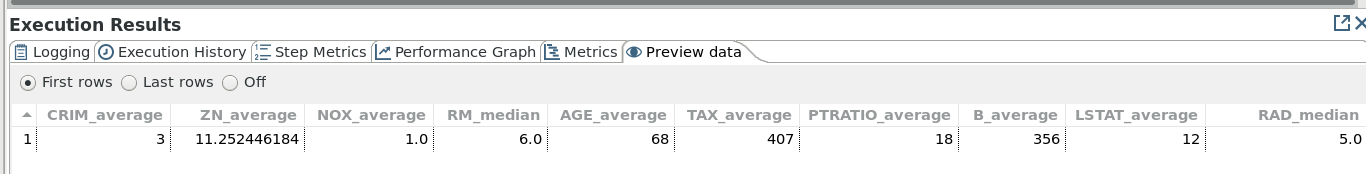


Рисунок 7 – Рассчитанные средние и медианы

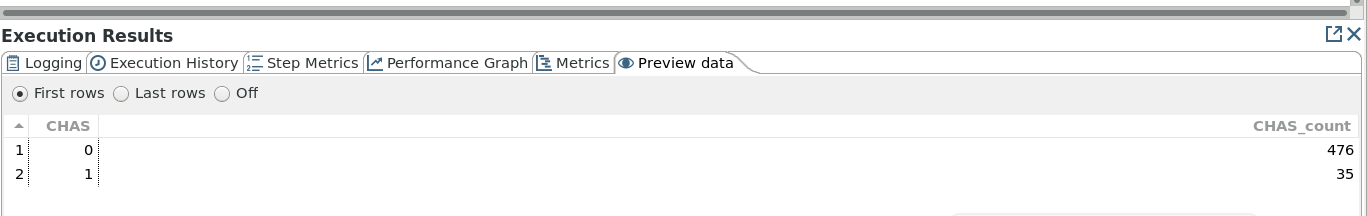


Рисунок 8 – Подсчитанные поля по вхождению расположения у реки

Больше всего домов не у реки, поэтому пустые строки можно заменить 0, а в других столбцах пропущенные значения средними и медианой, которые уже рассчитаны. Далее был вставлен объект If field value is null, который позволяет менять нулевые значения на нужные, как показано на рисунке 9.

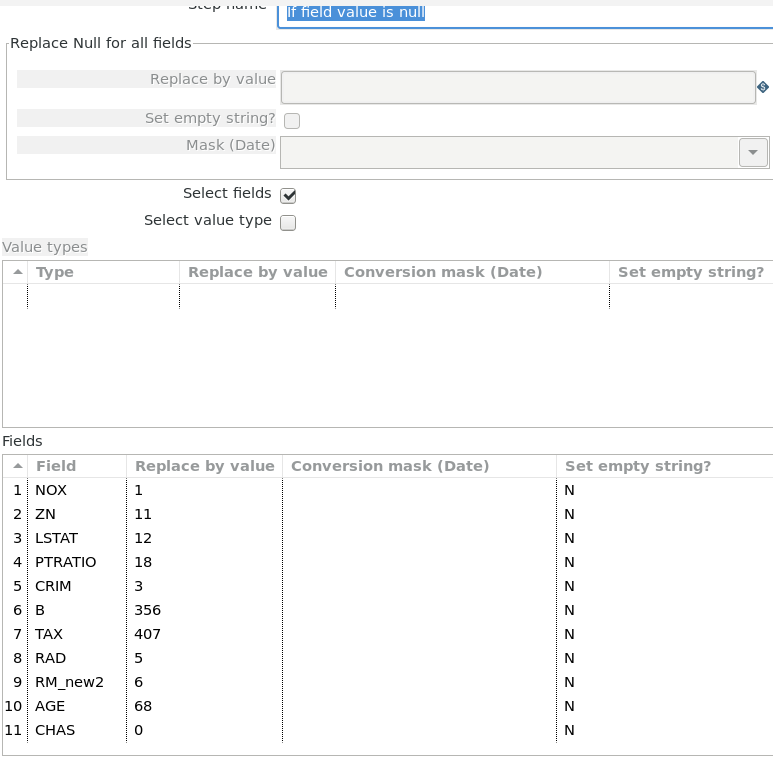


Рисунок 9 – Замена нулевых значений нужными

В данном объекте можно выбрать, менять ли нулевые значения по типам данных или по конкретным столбцам, для этого была поставлена галочка по столбцам и указаны столбцы со значениями. После проведенных изменений был выполнен анализ данных для выявления, как показатели влияют на целевую переменную, в данном случае на стоимость недвижимости. Первый показатель – уровень преступности (CRIM). По логике при его увеличении должна уменьшаться средняя стоимость. Для данной проверки была выполнена группировка по уровню преступности и для каждого рассчитана средняя стоимость жилья, все продемонстрировано на рисунках 10-11.

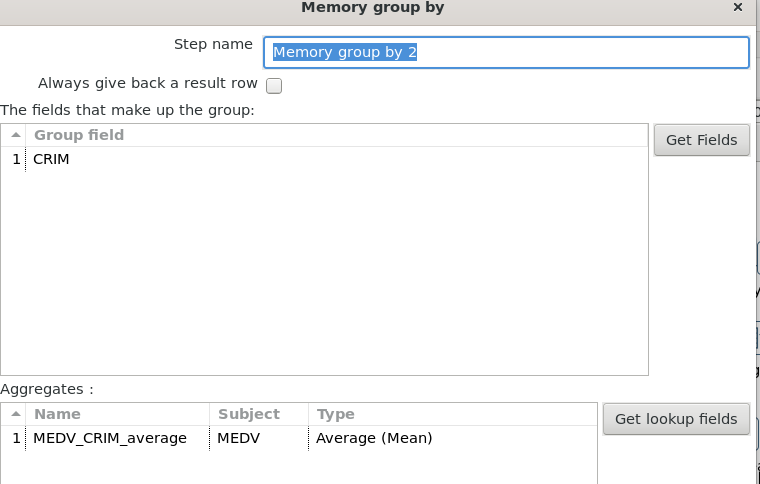


Рисунок 10 – Расчет средней цены по уровню индекса преступности

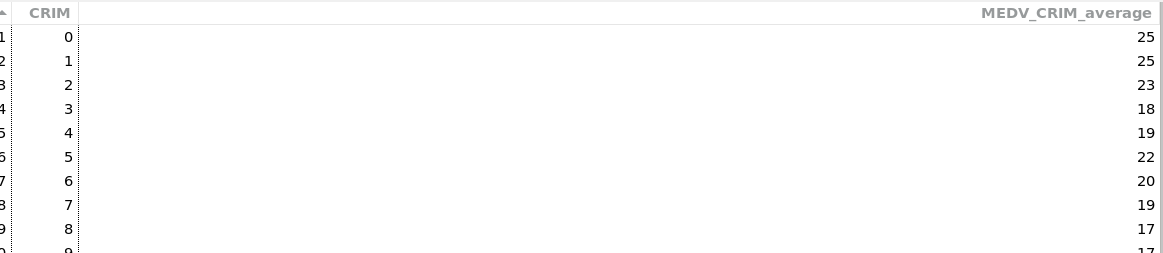


Рисунок 11 – Фрагмент результирующей таблицы по уровню преступности

Данная гипотеза оказалась верна. Теперь другая о том, что при увеличении числа комнат растет и цена на жилье на рисунках 12-13.

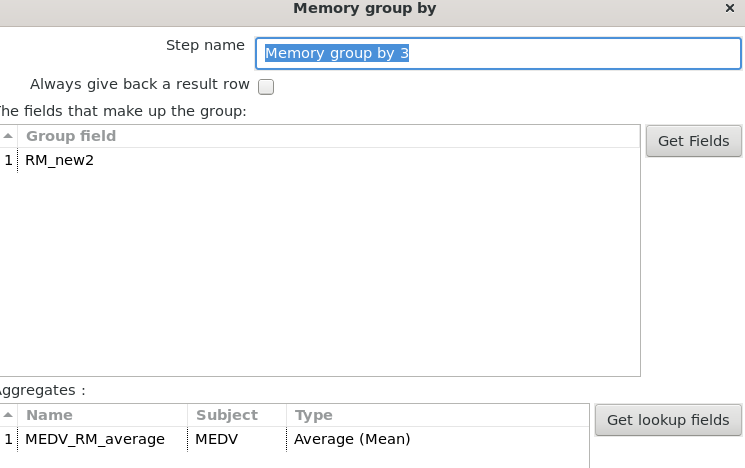


Рисунок 12 – Расчет средней цены на жилье по числу комнат

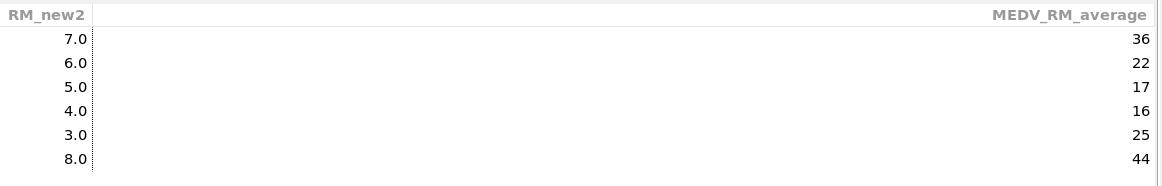


Рисунок 13 – Фрагмент результирующей таблицы по комнатам

После второй проверенной гипотезы третья: чем у населения меньше зарплата, тем и дешевле жилье, ведь жители не могут себе позволить дорогие дома, что видно на рисунках 14-15.

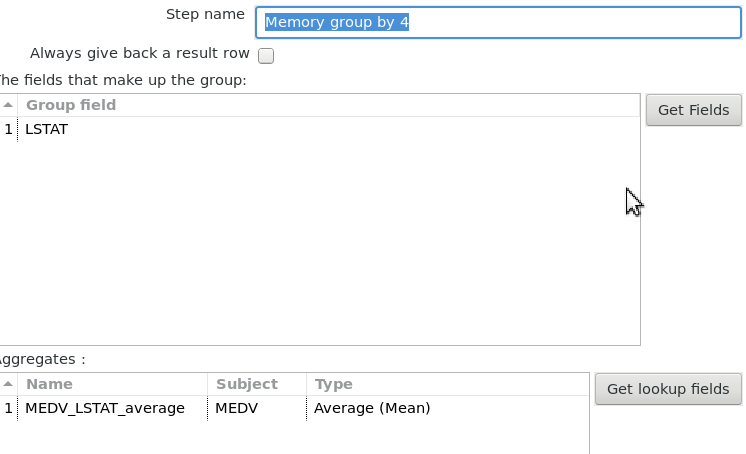


Рисунок 14 – Расчет средней цены жилья для доли низкооплачиваемого населения

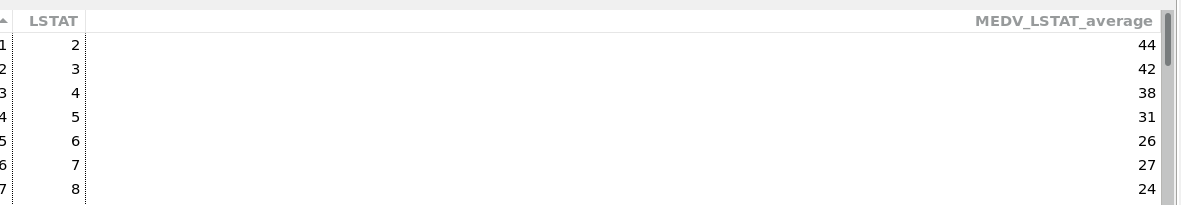


Рисунок 15 – Фрагмент результирующей таблицы по доле низкооплачиваемого населения

После проведенного анализа данные были выгружены в базу данных и выполнено к ней подключение, как показано на рисунках 16-17.

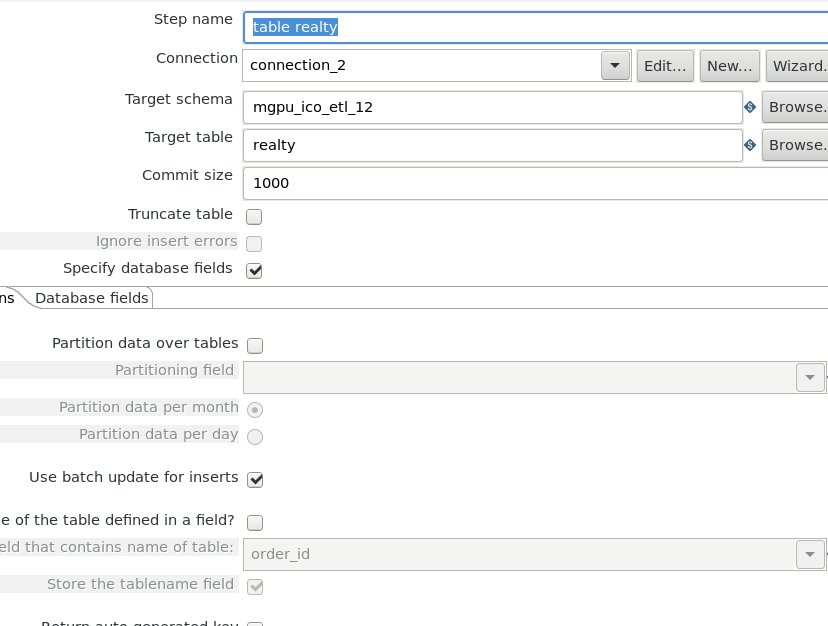


Рисунок 16 – Настройка подключения к базе данных

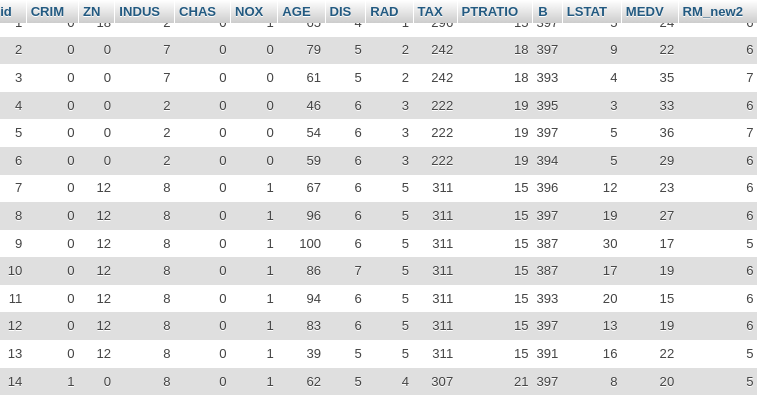


Рисунок 17 – Фрагмент загруженной таблицы

Конечная трансформация показана на рисунке 18.

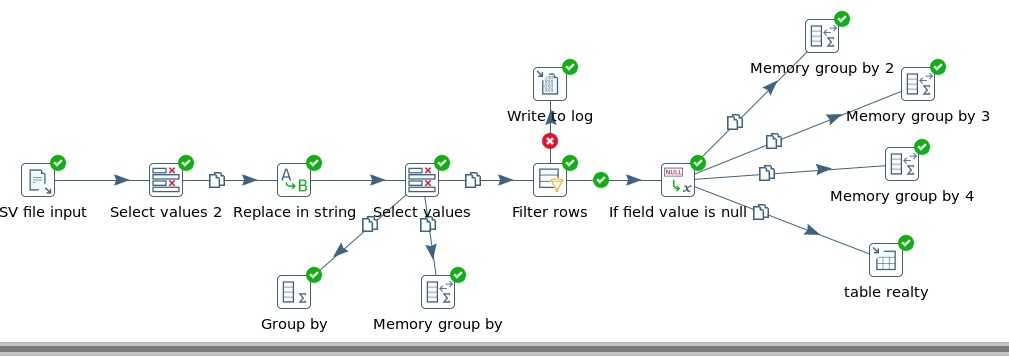


Рисунок 18 – Трансформация для данных по недвижимости

**Общий вывод**:

В ходе выполнения работы были изучены основные принципы работы с ETL-инструментами на примере Pentaho Data Integration (PDI), благодаря продвинутым возможностям Pentaho Data Integration удалось настроить динамические соединения к различным источникам данных, таким как база данных MySQL, а также к внешнему файлу в формате CSV. Были изучены объекты, позволяющие работать с регулярными выражениями, менять типы данных, группировать и фильтровать данные. Промежуточные результаты можно мониторить при просмотре данных, которые получились при выполнении трансформации. Группировка позволяет рассчитывать важные статистические показатели (медианы, средние, количество значений и др.), при этом для расчета моды приходится выполнять группировку и выявлять вручную наиболее распространённое значение. Для группировки необязательно указывать поле, по которому будет проводиться группировка, достаточно указать агрегатную функцию для нужного столбца, в данном случае расчет будет выполняться для всей таблицы без группы. Все цель и задачи были выполнены, получилась трансформация, в рамках которой данные загружаются, обрабатываются, анализируются и выгружаются уже обработанные и очищенные в базу данных.