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:

Инструменты для хранения и обработки больших данных

**Лабораторная работа №6.1.**

**Тема:**

**«**Обработка данных с использованием Apache Spark**»**

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

Преподаватель: Босенко Т. М.

Москва

2024

**Цель и задачи работы:**

* Познакомиться с понятием «большие данные» и способами их обработки.
* Познакомиться с инструментом Apache Spark и возможностями, которые он предоставляет для обработки больших данных.
* Получить навыки выполнения разведочного анализа данных использованием pyspark.

**Ход работы**

Apache Spark — это фреймворк для обработки и анализа больших объёмов информации. Он позволяет быстро выполнять операции с данными в вычислительных кластерах и поддерживает такие языки программирования, как Scala, Java, Python, R и SQL. В данной работе используется PySpark (Python API для Apache Spark). Он позволяет выполнять в реальном времени масштабную обработку данных в распределённой среде с помощью Python.

Некоторые возможности PySpark:

* поддерживает различные операции, такие как фильтрация, объединение, агрегация и сортировка данных;
* интегрируется с различными источниками данных, включая HDFS, Cassandra, HBase и S3;
* обрабатывает данные в реальном времени благодаря распределённой вычислительной модели;
* поддерживает машинное обучение через библиотеку MLlib.

Для работы с PySpark использовался Google Colaboratory и интеграция с библиотекой для анализа данных Pandas. С платформы Kaggle были выгружены данные о продаже недвижимости в Англии и Уэльсе, полученные из земельного реестра правительства Великобритании HM. Он предлагает ценную информацию о сделках с недвижимостью, включая цены продажи, местоположение и типы продаваемых объектов [1]. Файл содержит 22 миллиона строк (рисунок 1-3).

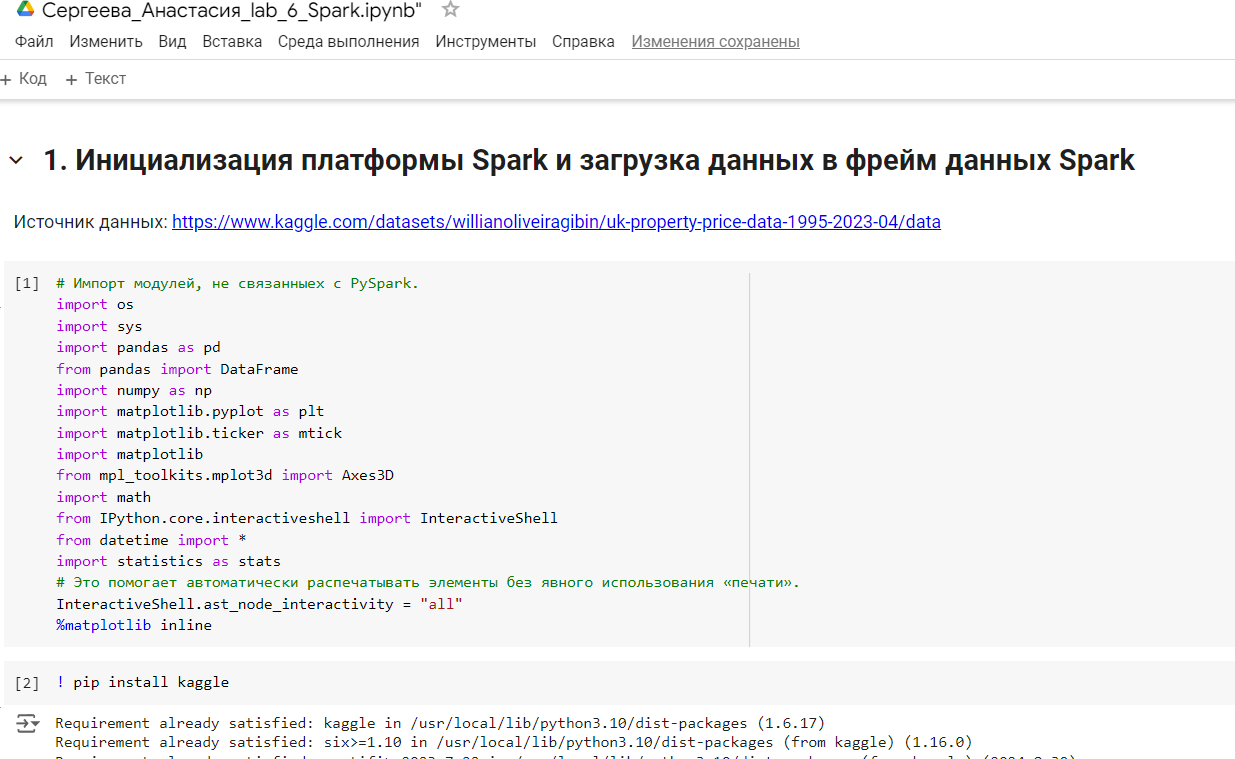


Рисунок 1 – Загрузка необходимых библиотек

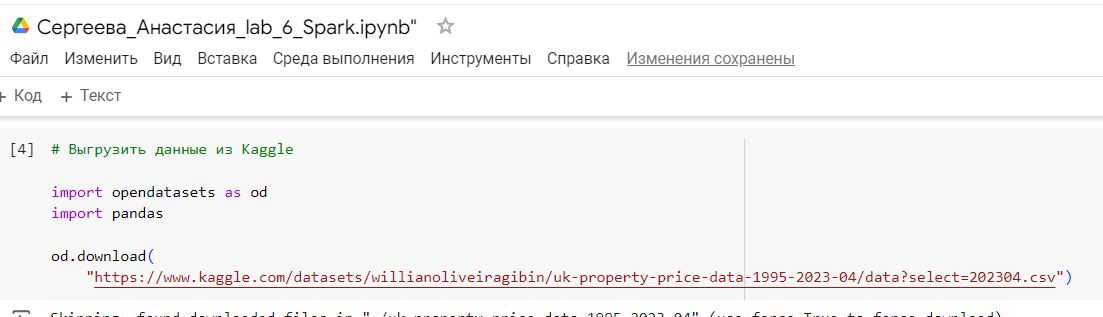


Рисунок 2 – Выгрузка данных из Kaggle

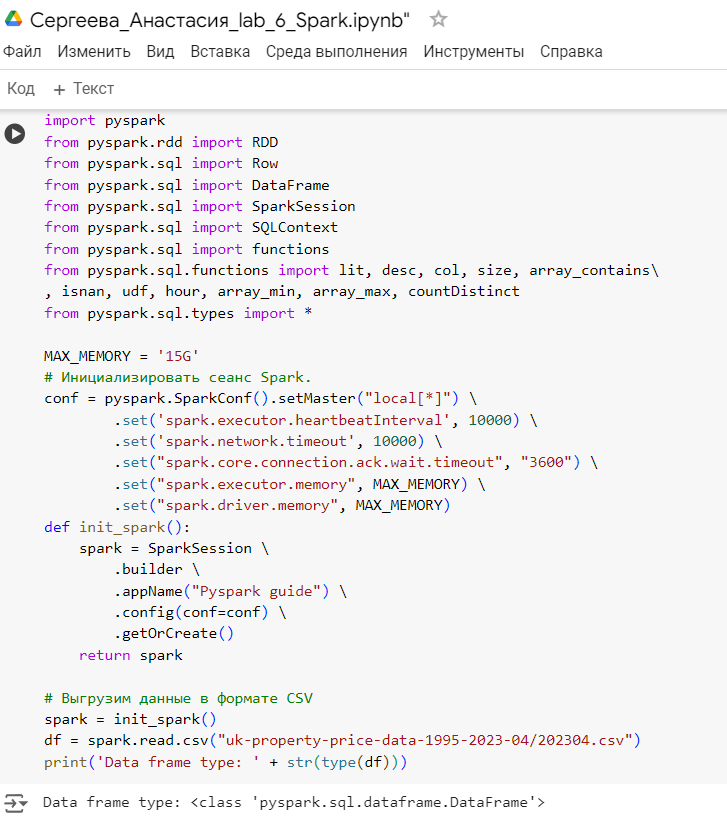


Рисунок 3 – Инициализация сеанса Spark и выгрузка файла в формате CSV

Таблица состоит из следующих столбцов: Postcode (почтовый индекс, по которому расположена недвижимость), PAON (основное адресное название объекта), SAON (вторичное адресное название объекта), Street (название улицы, на которой расположена недвижимость), Locality (дополнительная информация о населенном пункте), Town/City (населенный пункт, в котором находится недвижимость), District (округ, в котором находится недвижимость), County (округ, в котором расположена недвижимость), Price Paid (цена, по которой была продана недвижимость), Property\_Type (тип собственности).

Была просмотрена схема данных, все данные строковые, поэтому было принято решение поменять дату покупки на тип даты с учетом времени, а цену продажи на целочисленный тип. А также были переименованы столбцы, как должны быть (рисунок 4-6).

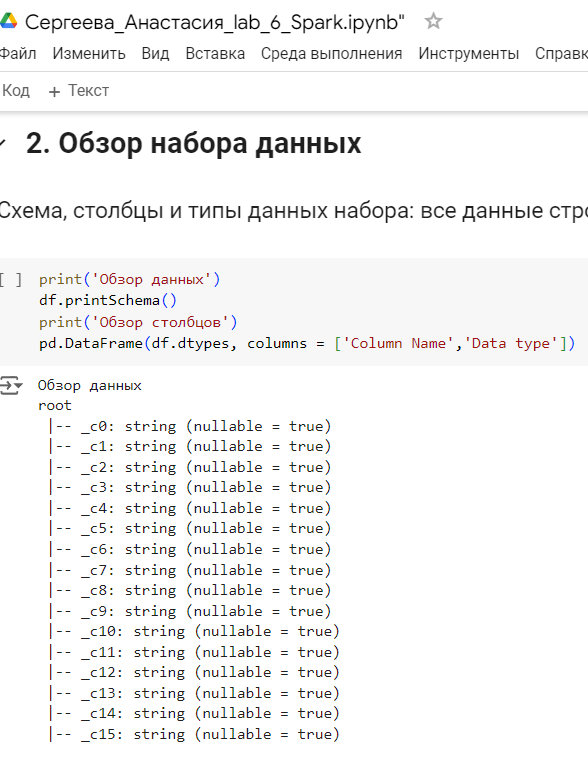


Рисунок 4 – Просмотр схемы данных, построение датафрейма с типами столбцов

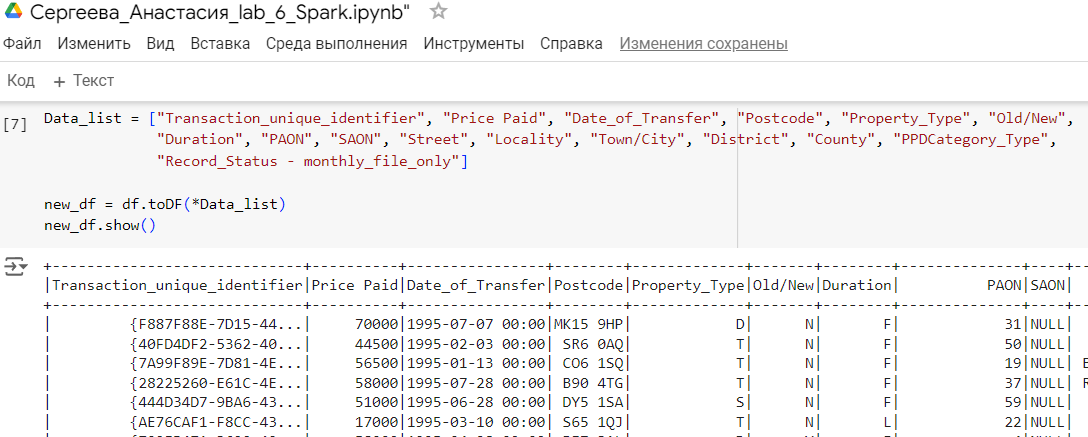


Рисунок 5 – Переименование столбцов

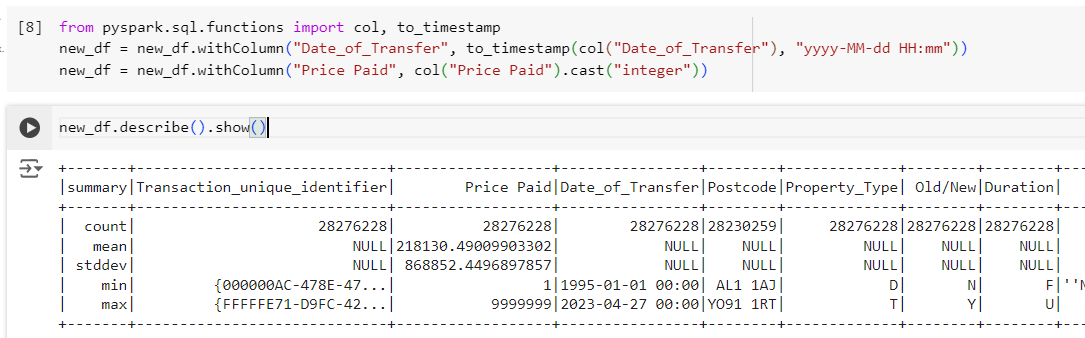


Рисунок 6 – Изменение типов данных столбцов «Date\_of\_Transfer» и «Price Paid» и описательная статистика

Далее были удалены дубликаты, если они присутствовали (рисунок 7).

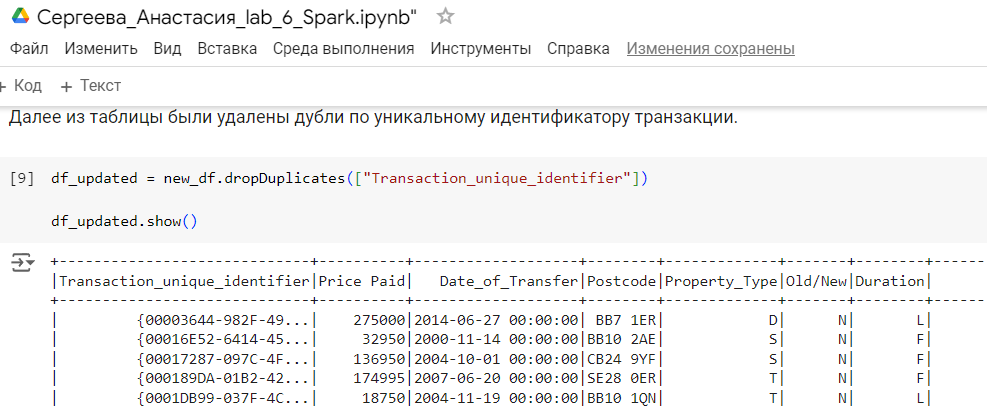


Рисунок 7 – Удаление дубликатов по номеру транзакции

После предварительного просмотра столбцов первое, что должны проверить, — это наличие в наборе данных какого-либо пропущенного значения (рисунок 8).



Рисунок 8 – Поиск пропущенных значений

Наблюдается отсутствие некоторых данных по расположению домов, а именно PAON, Street, SAON, Locality. Почтовые индексы Великобритании не всегда соответствуют конкретным улицам, поскольку они предназначены для обозначения более широких районов. Одна улица может иметь несколько почтовых индексов, в зависимости от ее протяженности и расположения, поэтому даже наличие почтового индекса не поможет вычислить улицу. При этом поля Locality и SAON необязательны для запонения, т.к. отвечают за дополнительную информацию, которая необязательно могла быть указана. Поэтому было принято решение удалить данные столбцы, т. к. они не будут использоваться для анализа, также, как и столбец, фиксирующий актуальность данных («Record\_Status - monthly\_file\_only»). Все данные актуальны (рисунок 9-11).

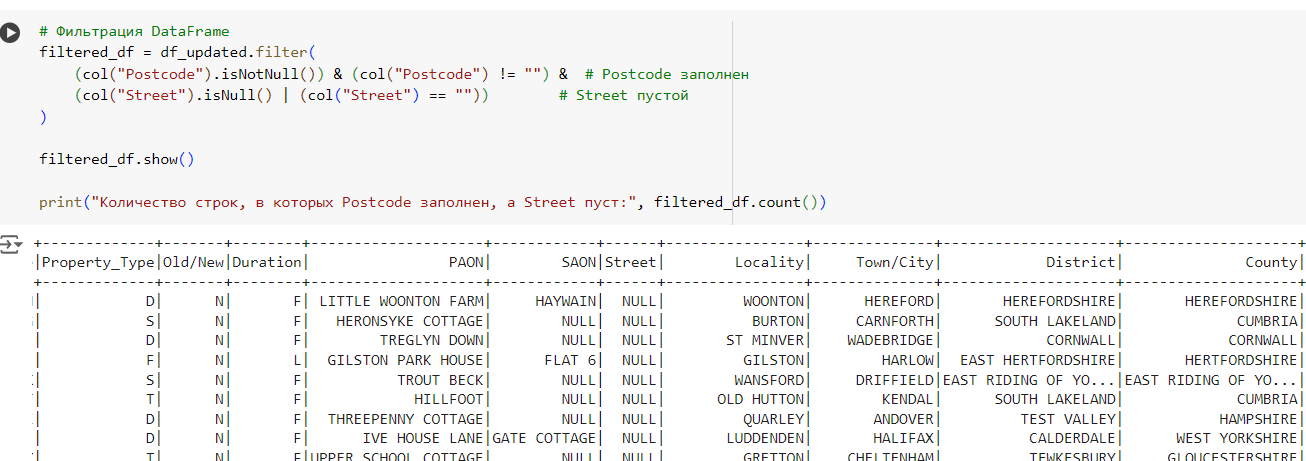


Рисунок 9 – Транзакции с почтовым индексом и без указания улицы

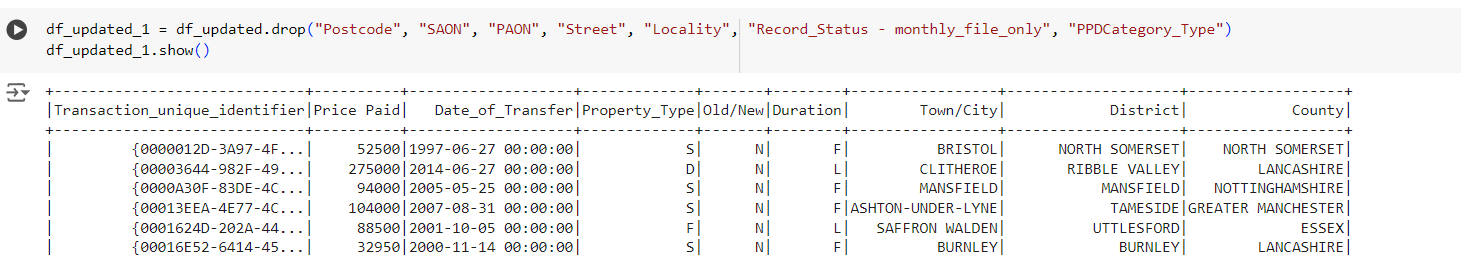


Рисунок 10 – Удаление ненужных столбцов

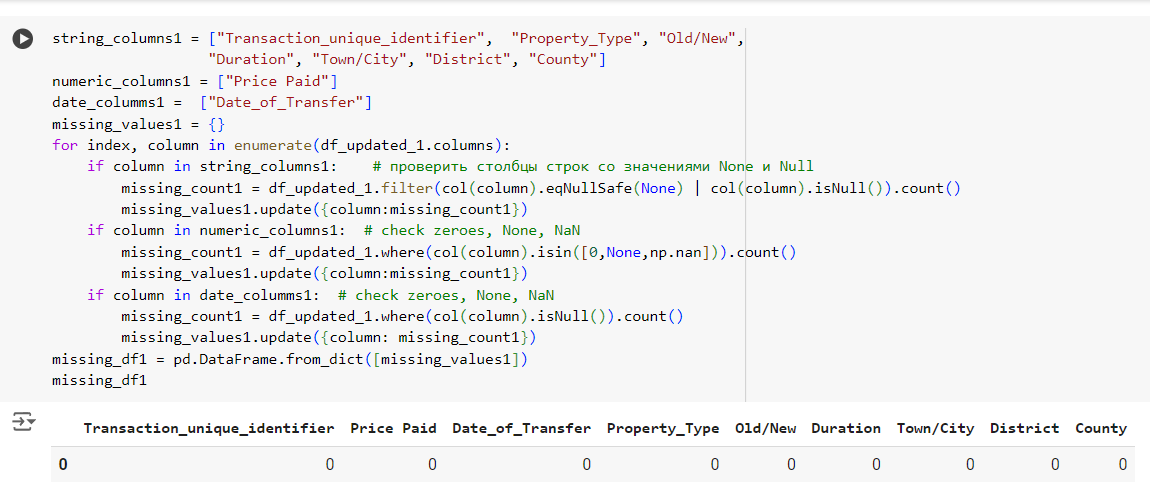


Рисунок 11 – Повторный поиск пропущенных значений

Была создана UDF функция для извлечения года и месяца из даты (рисунок 12).

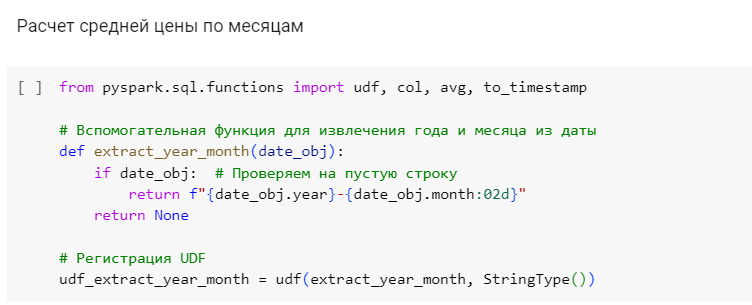


Рисунок 12 - UDF функция для извлечения года и месяца из даты

Данная функция помогла в анализе средней продажи недвижимости по месяцам. Для визуализации результата использовалась библиотека Altair (рисунок 13-14).

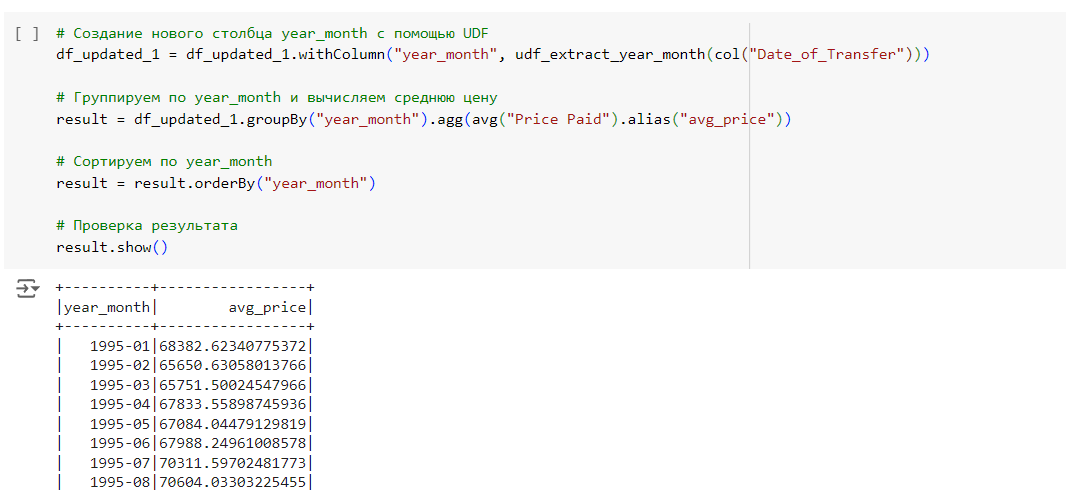


Рисунок 13 – Группировка по месяцам средней продажи недвижимости

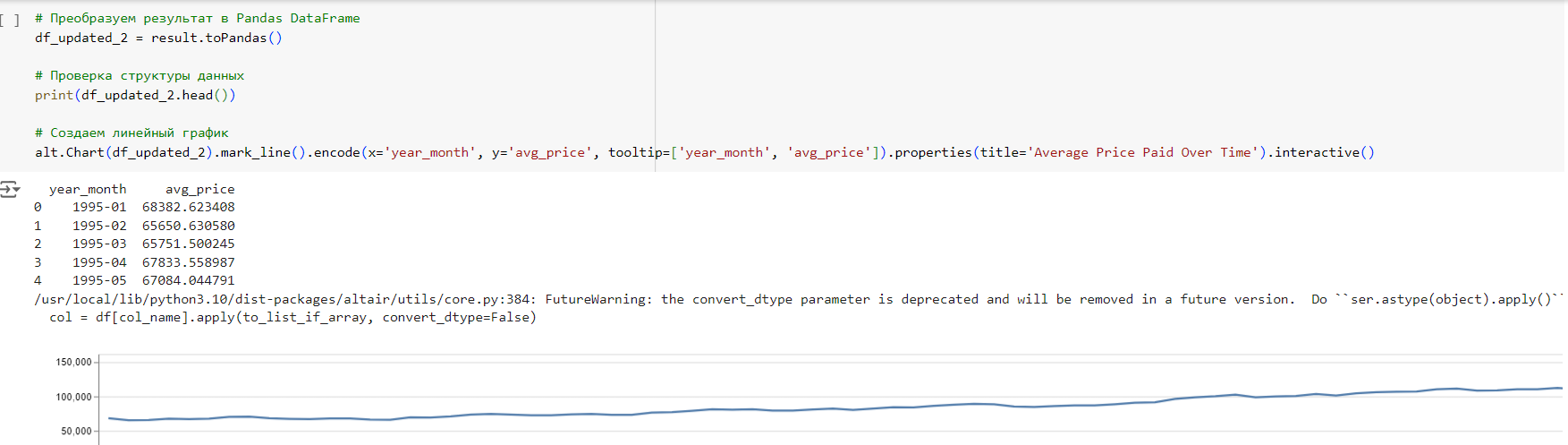


Рисунок 14 – Визуализация средней продажи по месяцам

Наблюдается положительный тренд в росте средней цены на недвижимость, при этом с 2020 года цены стали меняться более резко. В целом это можно объяснить несколькими причинами:

1. Это был период пандемии, все сидели дома, что побудило людей больше ценить личный комфорт и покупать жилье.
2. В июле 2020 года правительство ввело приостановку гербового сбора — налога на покупку домов.
3. При этом в 2021 году отменили льготы, поэтому спрос на жилье ниже, чем в 2020.
4. Цены с каждым годом меняются, как и прожиточный минимум.

Похожие графики были построены за последние 10 лет по городам, лидирующим по продажам недвижимости, по типам недвижимости, округам, регионам, типам владения недвижимостью с возможностью фильтрации. Данные были отфильтрованы по году, создан столбец с годом и осуществлена группировка. Был создан список с возможностью выбрать все пункты или один из перечисленных для визуализации (рисунки 15-25).



Рисунок 15 – Загрузка необходимых библиотек

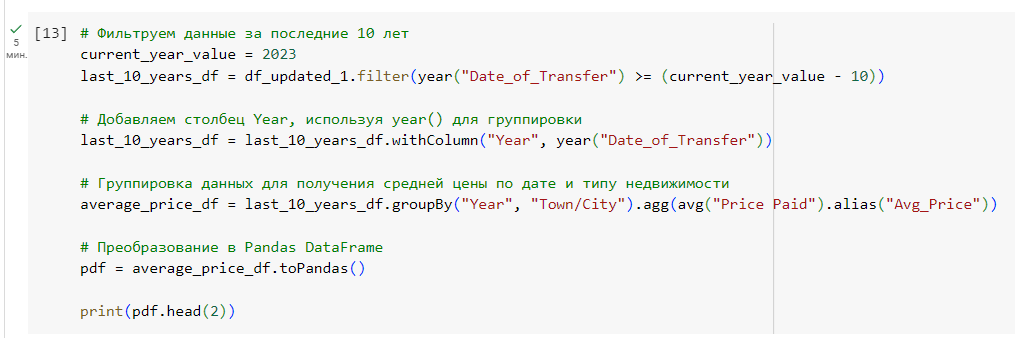


Рисунок 16 – Расчет средней цены за последние 10 лет



Рисунок 17 – Создание функции для создания интерактивного графика

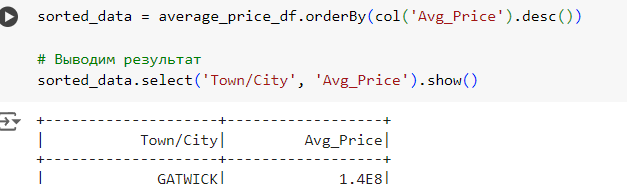


Рисунок 18 – Поиск самых дорогих городов

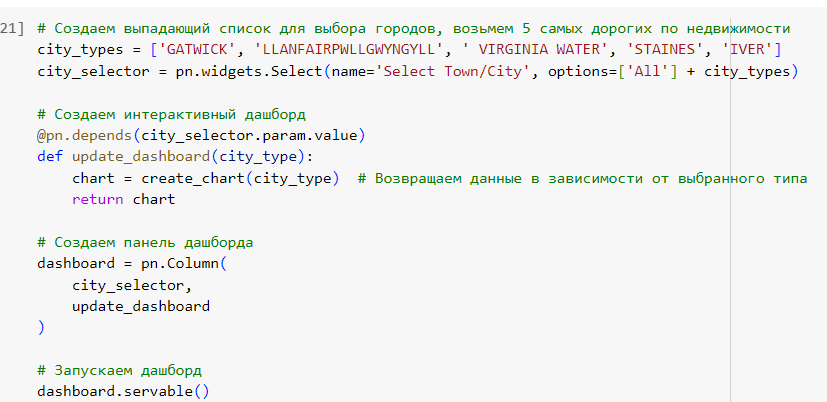


Рисунок 19 – Построение графика для средней цены по 5 городам

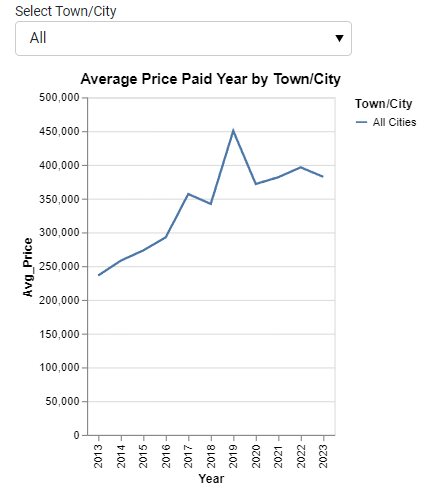


Рисунок 20 – Средняя цена за 10 лет для всех 5 городов

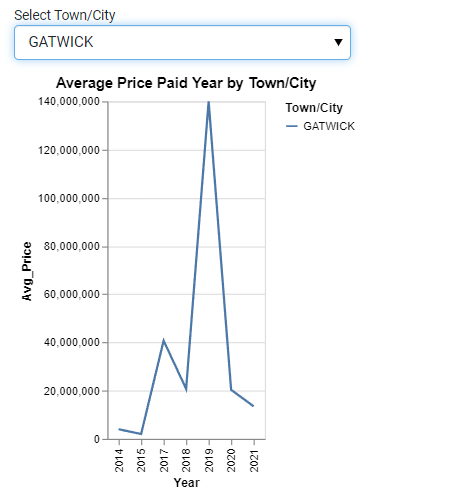


Рисунок 21 - Средняя цена за 10 лет для Gatwick

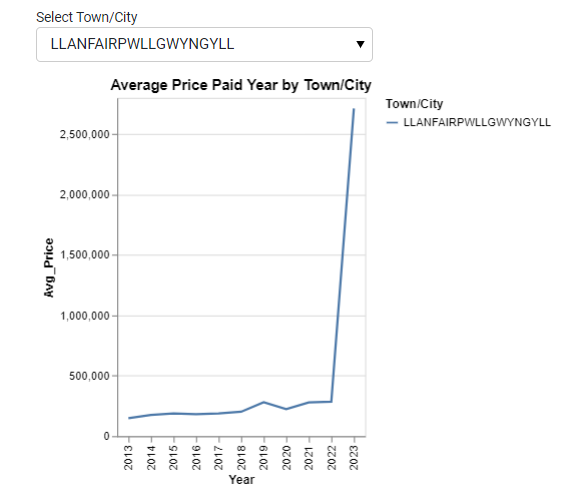


Рисунок 22 - Средняя цена за 10 лет для Llanfairpwllgwyngyll

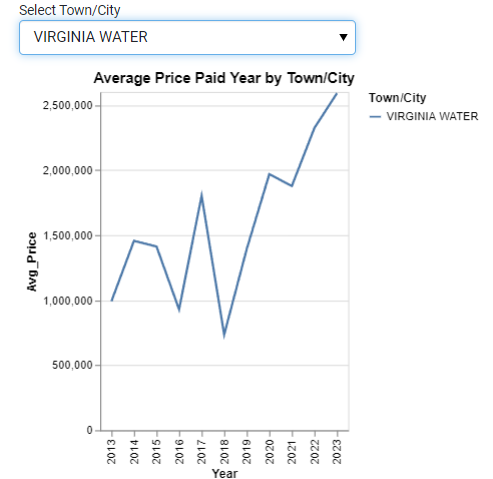


Рисунок 23 - Средняя цена за 10 лет для Virginia Water

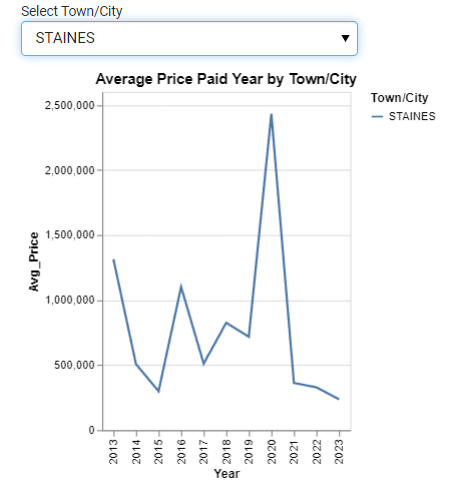


Рисунок 24 - Средняя цена за 10 лет для Staines

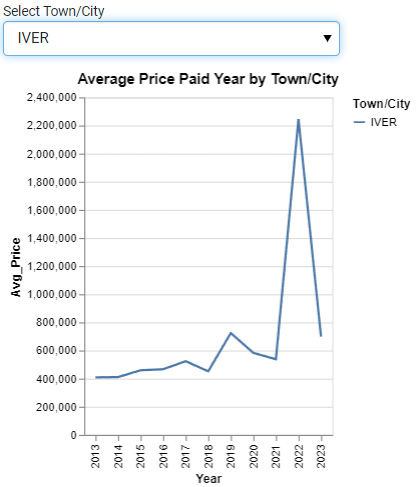


Рисунок 25 - Средняя цена за 10 лет для Iver

Средняя цена у Gatwick, схожа с общим трендом у городов, а именно наблюдается резкий скачок цен в промежутке 2018-2020 годов. У Llanfairpwllgwyngyll с 2020 года резкий рост цен. У Iver скачок наблюдается с 2021 по 2023. У остальных городов цены колеблются, но не так заметно.

Далее был проанализирован рынок по типам недвижимости (рисунки 26-33).

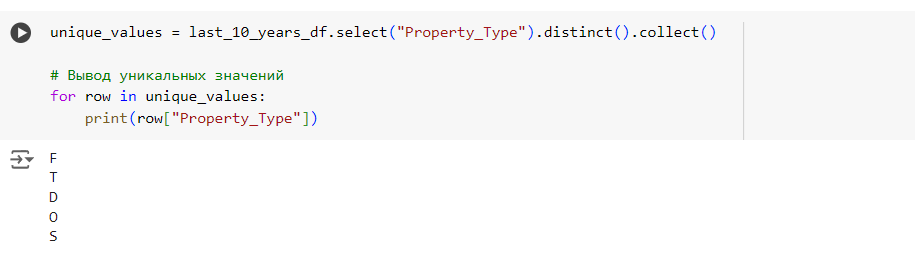


Рисунок 26 – Вывод типов недвижимости



Рисунок 27 – Построение графика для типов недвижимости также за 10 лет

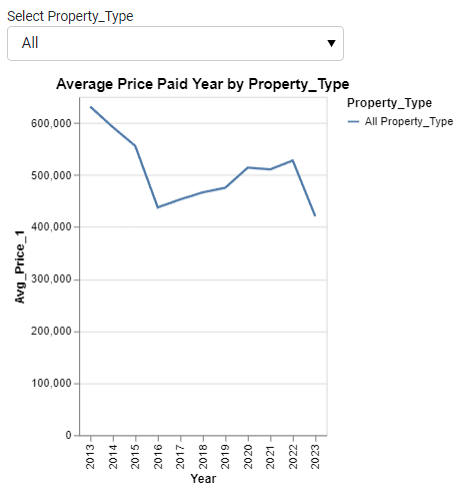


Рисунок 28 – График изменения средней цены недвижимости по всем типам недвижимости

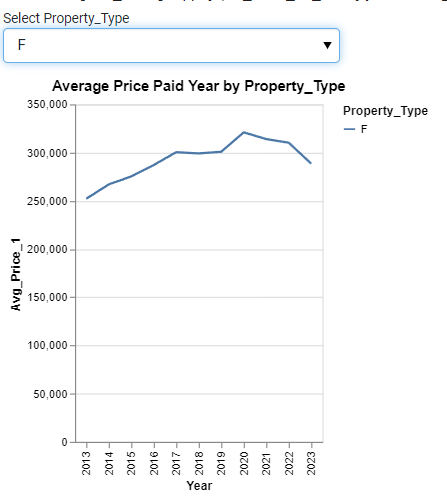


Рисунок 29 – График изменения средней цены недвижимости по типу F

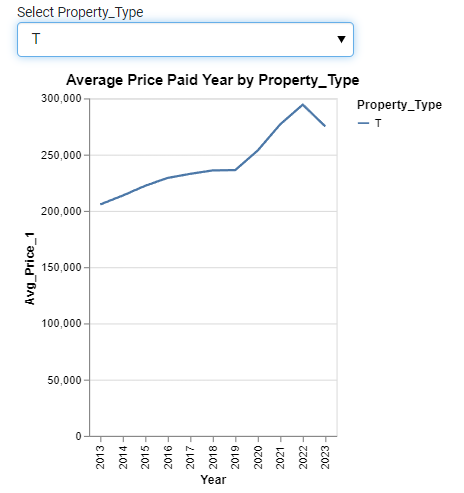


Рисунок 30 – График изменения средней цены недвижимости по типу T

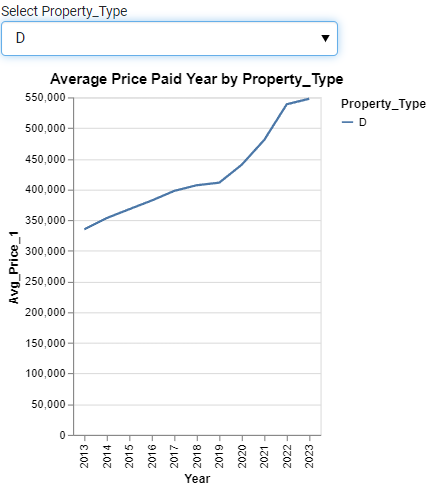


Рисунок 31 – График изменения средней цены недвижимости по типу D

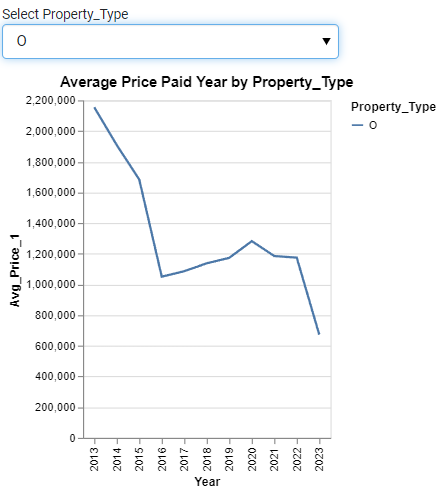


Рисунок 32 – График изменения средней цены недвижимости по типу O

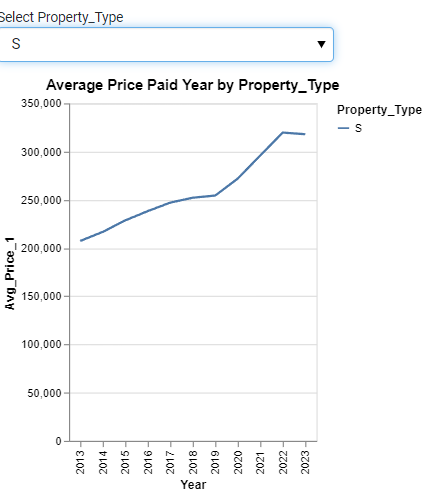


Рисунок 33 – График изменения средней цены недвижимости по типу S

На обычное владение недвижимостью freehold цена в среднем стабильная, на Terraced house с 2020 по 2022 был небольшой рост, на Detached house с 2019 года наблюдается рост цены, на тип недвижимости O до 2015 был спад, до 2022 года цена сохранялась примерно на одном уровне, на Semi-detached house цена растет.

Далее был построен график для старых и новых цен (рисунки 34-37).

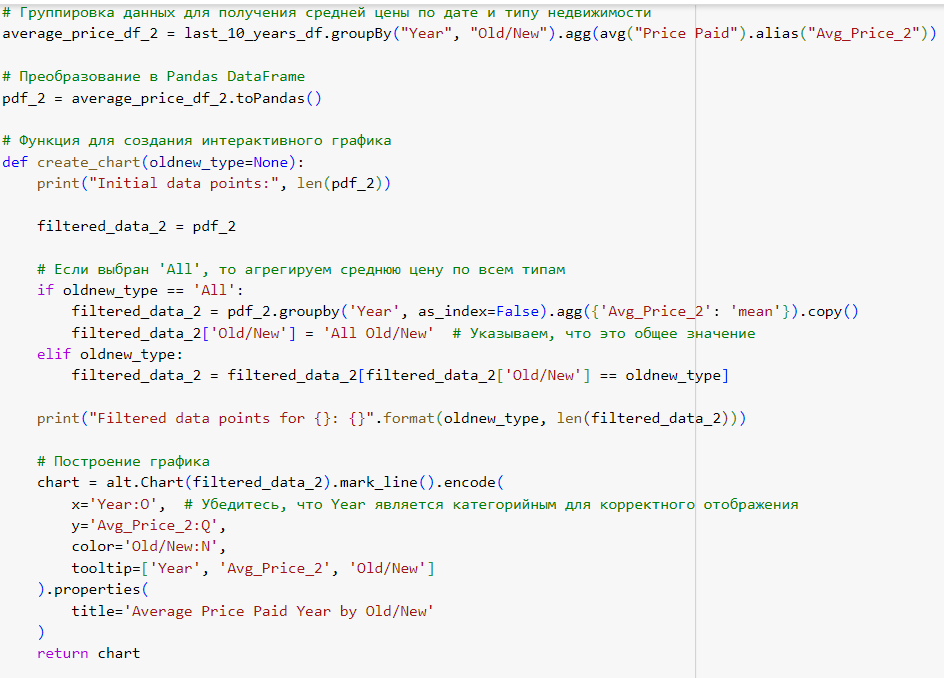


Рисунок 34 - Построение графика для старых и новых домов

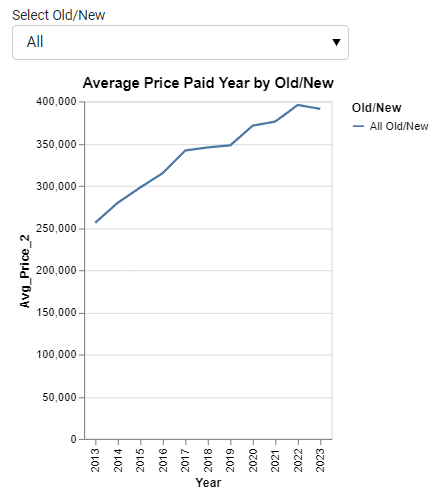


Рисунок 35 - График изменения средней цены недвижимости по новизне дома

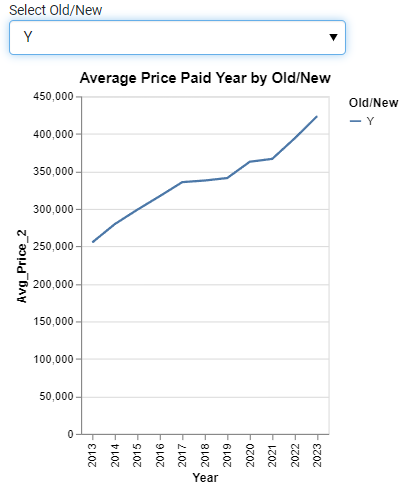


Рисунок 36 - График изменения средней цены недвижимости для старых домов

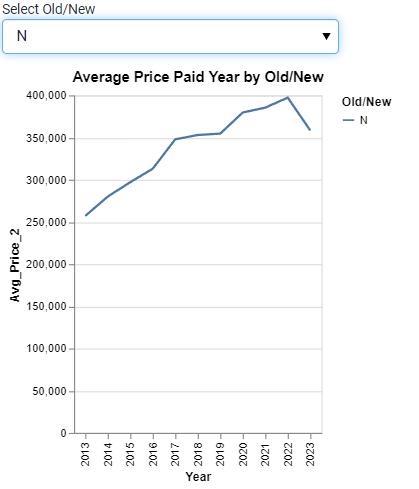


Рисунок 37 - График изменения средней цены недвижимости для новых домов

Цены на недвижимость растут, но у новых домов спад цены с 2022 года. Теперь регионы (рисунки 38-).

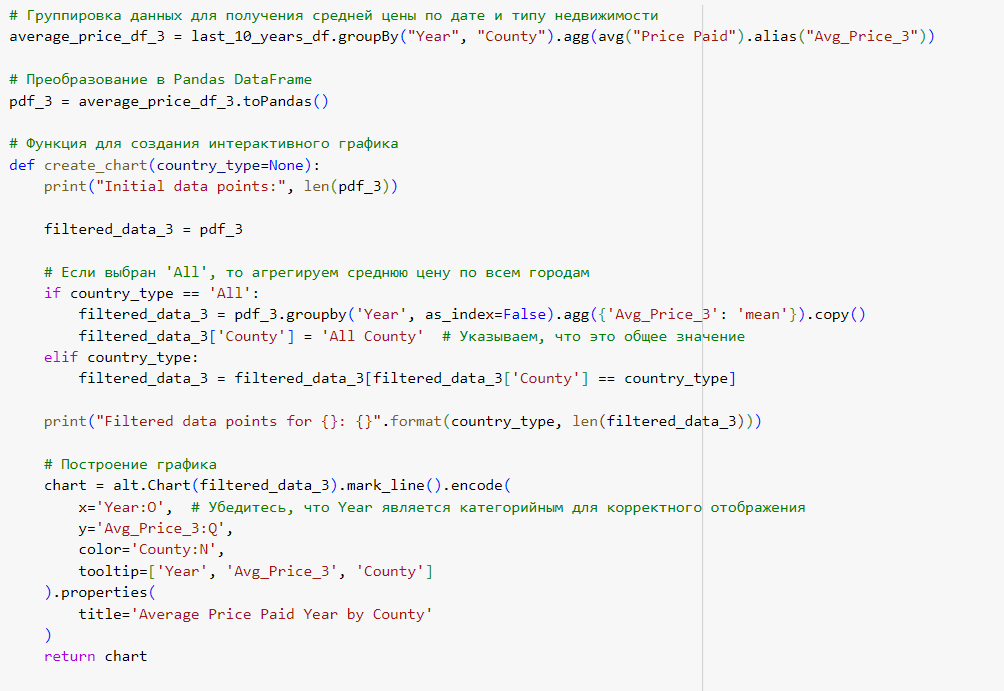


Рисунок 38 - Построение графика средней цены по регионам

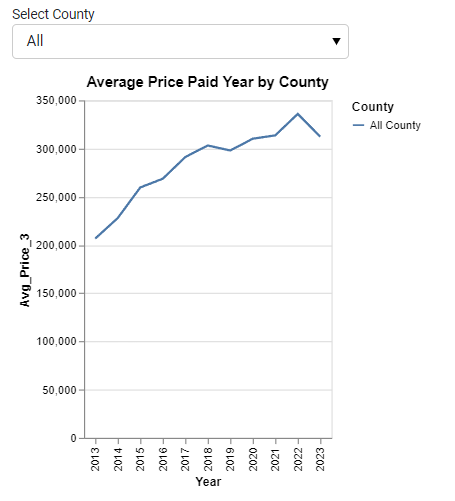


Рисунок 39 - График изменения средней цены недвижимости по регионам

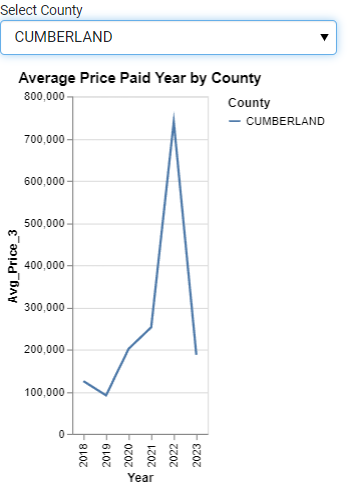


Рисунок 40 - График изменения средней цены недвижимости для Cumberland

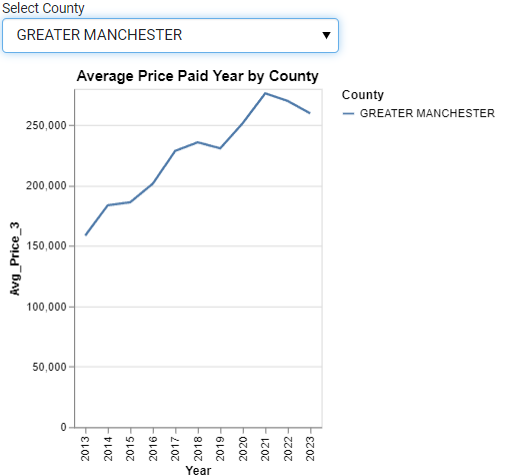


Рисунок 41 - График изменения средней цены недвижимости для Greater Manchester

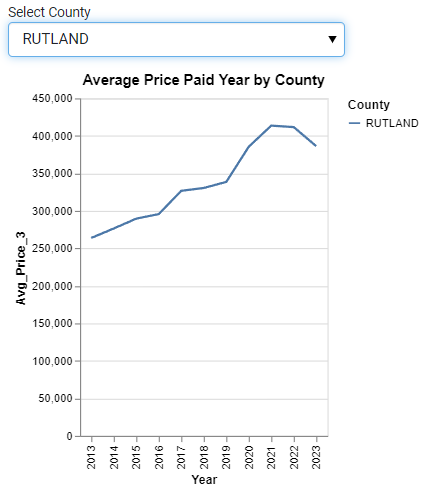


Рисунок 42 - График изменения средней цены недвижимости для Rutland

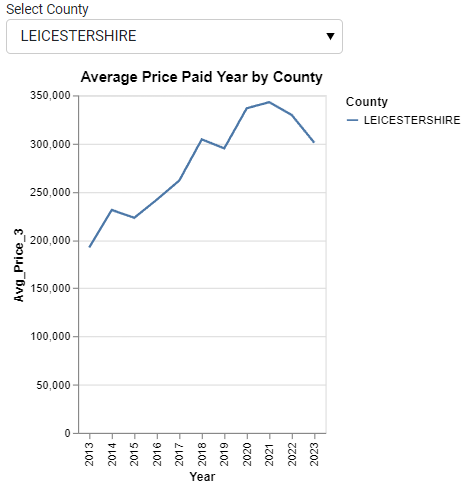


Рисунок 43 - График изменения средней цены недвижимости для Leicestershire

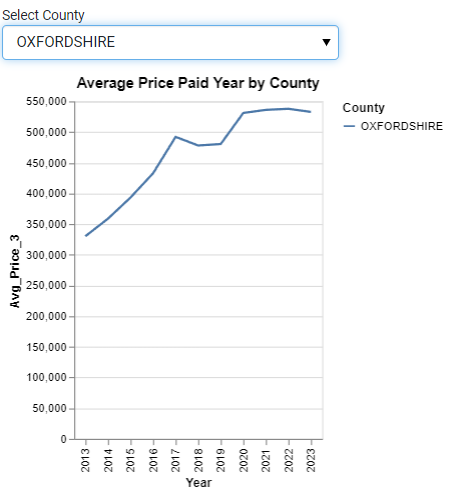


Рисунок 44 - График изменения средней цены недвижимости для Oxfordshire

У Cumberland с 2021 по 2023 наблюдается резкий рост, у остальных городов стабильный рост.

Далее была проведена описательная статистика с группировкой по данным столбцам (рисунки 45-48). Было выявлено, что новые дома стоят чуть дешевле старых по средней цене, при этом минимальная цена нового дома меньше, чем старого. Это говорит о том, что цены на новую недвижимость более вариативны, то есть у них больше разброс, чем у старой недвижимости.



Рисунок 45 - Описательная статистика по новизне дома

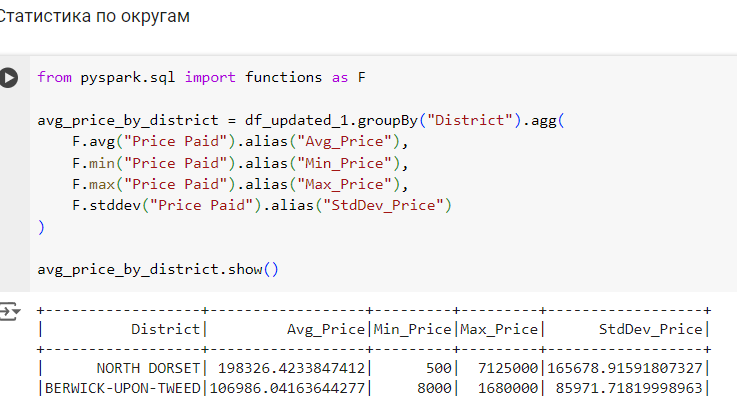


Рисунок 46 – Описательная статистика по округам

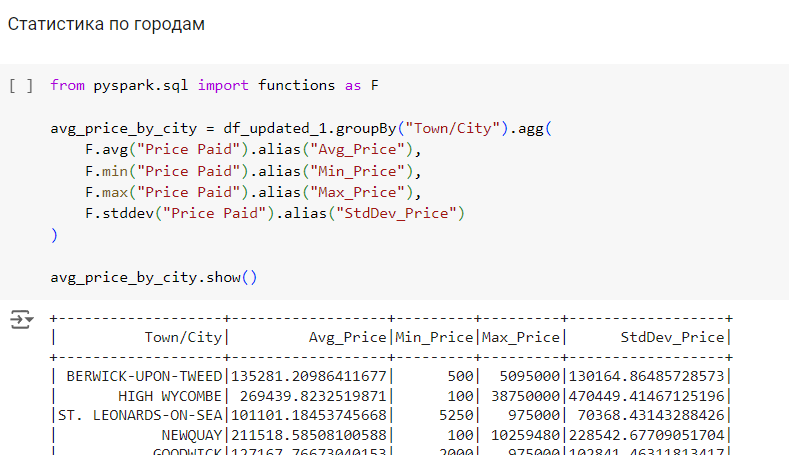


Рисунок 47 – Описательная статистика по городам

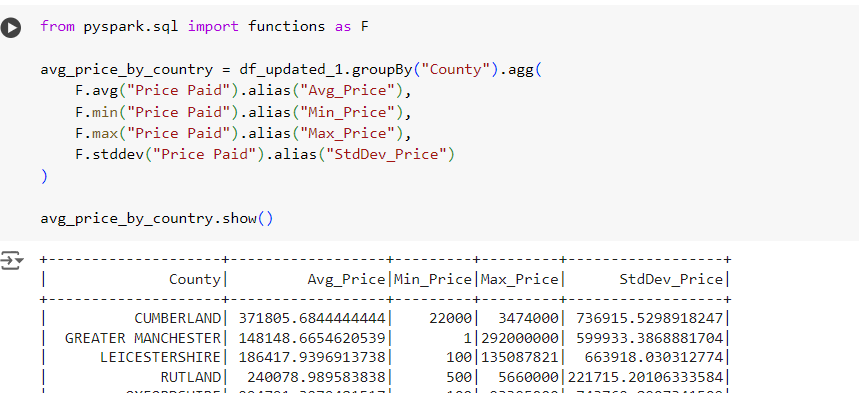


Рисунок 48 – Описательная статистика по регионам

Далее было принято решение поделить покупки недвижимости на категории высоких, средних и низких цен (рисунок 49).



Рисунок 49 – Подсчет количества продаж по каждой из категорий цен

Как видно, высоких цен больше, чем низких. В сегменте высоких цен зафиксировано 7.072.890, в категории низких 7.055.266, а в категории средних 14.148.072 продаж.

**Выводы**

1. PySpark позволяет обрабатывать огромные объемы данных за счет распределенной вычислительной мощности. В ходе лабораторной работы было наблюдено, что выполнение задач на кластере Spark значительно быстрее, чем на локальных машинах, даже при использовании эффективных библиотек для анализа данных, таких как Pandas.
2. В лабораторной работе активно использовался язык SQL-запросов на Spark, что позволяло удобно выполнять операции агрегации, группировки, сортировки и обращения к частям данных.
3. Интеграция Pandas и Spark при работе с данными помогала от обработки данных на Spark переходить на библиотеки для визуализаций через Pandas.
4. Построены с помощью библиотеки Altair визуализации с возможностью фильтрации.
5. Проанализирована динамика средних цен на недвижимость по типам недвижимости, новизне дома, региону, городу.
6. Применены UDF функции, например, для извлечения года и месяца из даты.
7. Средние цены были поделены на категории высоких, средних и низких цен, подсчитано количество продаж по каждой из категорий.