Целью данной работы было познакомиться с понятием «большие данные» и способами их обработки с помощью инструмента Apache Spark.

Произвести разведочный анализ датасета и определить:

* типы признаков в датасете;
* пропущенные значения и устранить их;
* выбросы и устранить их;
* рассчитать статистические показатели признаков (средних, квартилей и т.д.);
* визуализовать распределения наиболее важных признаков;
* корреляции между признаками.

Разведочный анализ данных был проведён с помощью фреймворка Apache Spark. Apache Spark – фреймворк для реализации распределённой обработки неструктурированных и слабоструктурированных данных. Spark работает в парадигме резидентных вычислений – хранит промежуточные данные в оперативной памяти, что приводит как к выигрышу в скорости работы для определенных задач, так и к возможности многократного доступа, что позволяет выполнять на нём различные алгоритмы машинного обучения.

Spark содержит в себе инструменты для работы с данными, хранящимися в виде устойчивого распределенного датасета (RDD), используя SQL-подобные конструкции и алгоритмы машинного обучения.

Для исполнения программы использовался Docker контейнер, содержащий jupyter-notebook с установленной библиотекой pyspark.

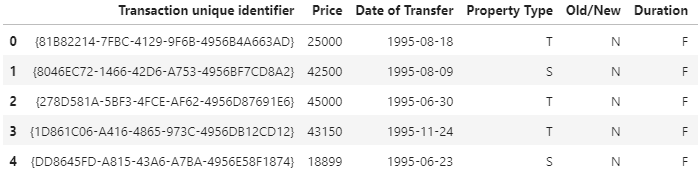
В качестве датасета был выбран следующий набор данных - <https://www.kaggle.com/datasets/hm-land-registry/uk-housing-prices-paid>. В датасете содержатся данные о покупке/продаже недвижимости в Великобритании с 1995 по 2017 гг.

Рассмотрим представленный датасет с Kaggle и отберём интересующие нас колонки для дальнейшего анализа.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Признак** | **Описание** | **Тип данных** | **Тип признака** | **Комментарий** |
| Transaction unique identifier | Уникальный идентификатор транзакции | Строковый | Порядковый | - |
| Price | Цена транзакции | Целочисленный | Количественный | - |
| Date of Transfer | Дата совершения транзакции | Дата | Порядковый | - |
| Property Type | Тип собственности | Строковый | Номинальный | Преобразован посредством one-hot кодировки |
| Old/New | Является ли строение новым или старым | Строковый | Номинальный | Преобразован в бинарный признак |
| Duration | Длительность владения недвижимостью (Freehold, Leasehold) | Строковый | Номинальный | Преобразован в бинарный признак |
| Town/City | Город, где находится собственность | Строковый | Номинальный | - |
| District | Район города | Строковый | Номинальный | - |
| County | Графство | Строковый | Номинальный | - |
| PPDCategory Type | Тип записи в реестре PPD | Строковый | Номинальный | Преобразован в бинарный признак |
| Record Status - monthly file only | Статус записи | Строковый | Номинальный | Не несёт информации т.к. все строки содержат одно и то же значение |

Выбранный датасет в формате CSV был загружен в среду исполнения Apache Spark. После загрузки файл CSV был преобразован в spark dataframe, были убраны неиспользуемые признаки.

Пример первых 5 записей из датафрейма представлен на Рисунке 1.



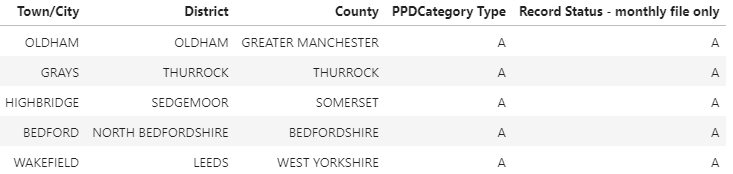


Рисунок 1 – Начало датасета

Общее количество записей в датасете – 22489348.

Определим наличие пропущенных значений в датасете. 

Рисунок 2 – Количество пропущенных значений

Удалим выбросы из датасета используя межквартильное расстояние. Для этого определим первую и третью квартиль для колонки *Price*, и вычислим расстояние между ними, после чего удалим из датасета все строки отстоящие от первой квартили вниз и от третьей квартили вверх более чем на 1,5 расстояния.

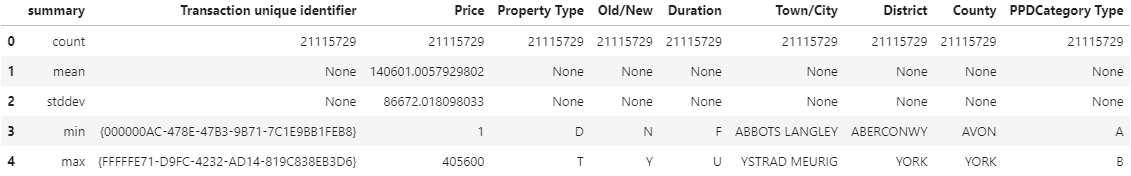


Рисунок 3 – Результат удаления выбросов

Колонка *PPDCategory Type* была приведена к бинарному типу – преобразована в колонку *additional\_entry*, означающую, является ли запись дополнительной или основной:

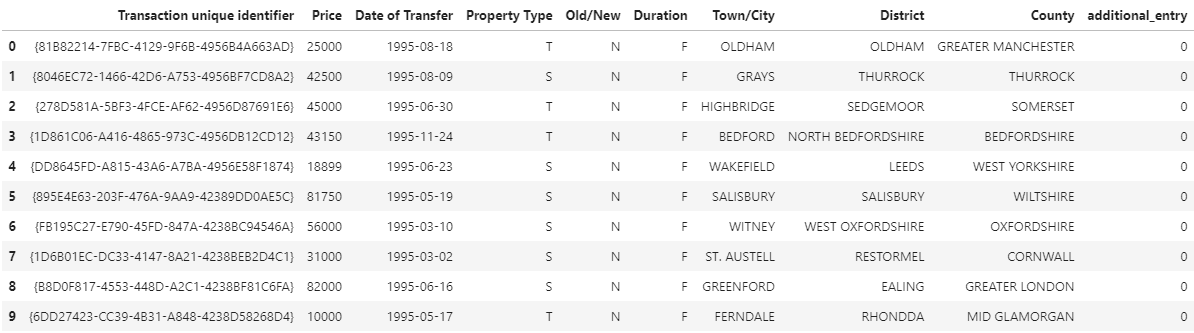


Рисунок 4 – Преобразованная колонка *PPDCategory Type*

Для преобразования была объявлена UDF (функция, определяемая пользователем), которая конвертировала строковое значение в числовое. Затем эта функция была применена к выбранной колонке и её результат был помещён в новую колонку, а старая колонка была удалена. Для последующих колонок были выполнены подобные преобразования.

Из колонки *Duration* были удалены значения “U” – неизвестная длительность владения, т.к. таких значений в датасете было около 500, и они не влияют на общую картину. Затем колонка была приведена к бинарному типу – преобразована в колонку *freehold*, означающую, владеет ли покупатель землей, на которой находится собственность (freehold – человек владеет и собственностью, и землей; leasehold – человек владеет собственностью, но арендует землю):

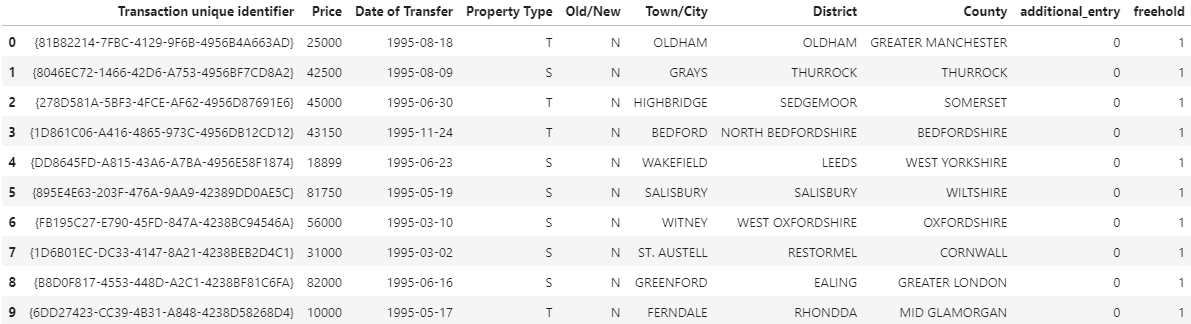


Рисунок 5 - Преобразованная колонка *Duration*

Колонка *Old/New* была приведена к бинарному типу – преобразована в колонку *new*, означающую, является ли собственность новой или подержанной:

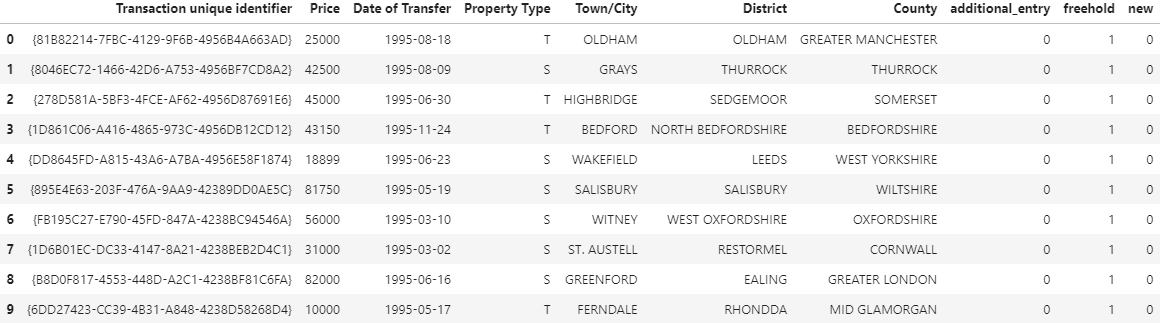


Рисунок 6 - Преобразованная колонка *Old/New*

Колонка *Property Type* содержит 5 различных значений:

* D – detached,
* S – semi\_detached,
* T – terraced,
* F – flats,
* O – other.

Эта колонка была преобразована в 5 новых колонок используя кодировку one-hot. Для этого был сначала использован StringIndexer, чтобы получить колонку, содержащую числа от 0 до 4 соответствующие значениям исходной колонки. Затем к новой колонке был применен OneHotEncoder, создавший колонку векторов, в которых единица на i-той позиции означает, что в предыдущей колонке было значение i. После этого, используя UDF, получающую элемент вектора в заданной позиции, были созданы 5 колонок, соответствующих вышеуказанным типам собственности. Промежуточные колонки были удалены из датасета.



Рисунок 7 - Преобразованная колонка *Property Type*

Для построения графиков датасет был сгруппирован по значениям некоторых колонок (property\_type, freehold, new, additional\_price), после чего было определено количество записей в каждой группе. На основе этих данных были построены круговые диаграммы.

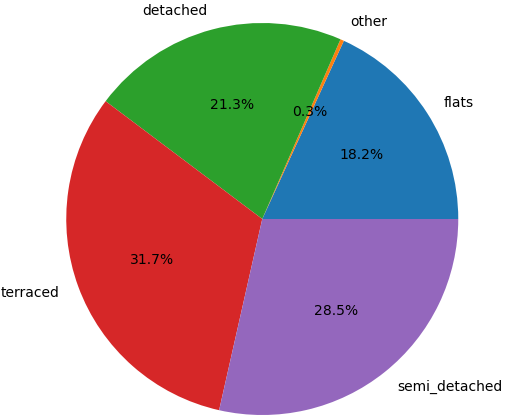


Рисунок 8 – Доля записей по типу собственности

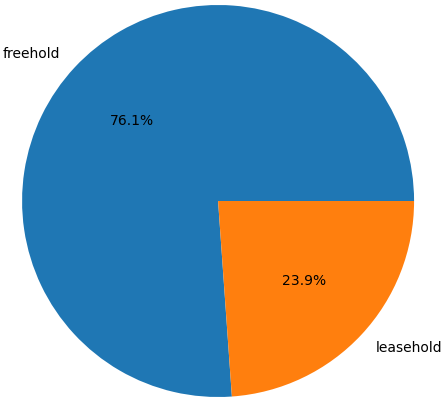


Рисунок 9 – Доля записей по типу владения собственностью

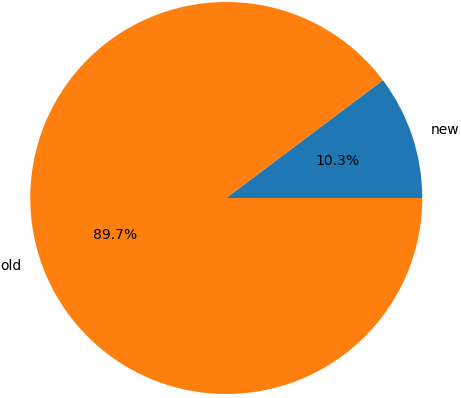


Рисунок 10 – Доля записей по новизне собственности

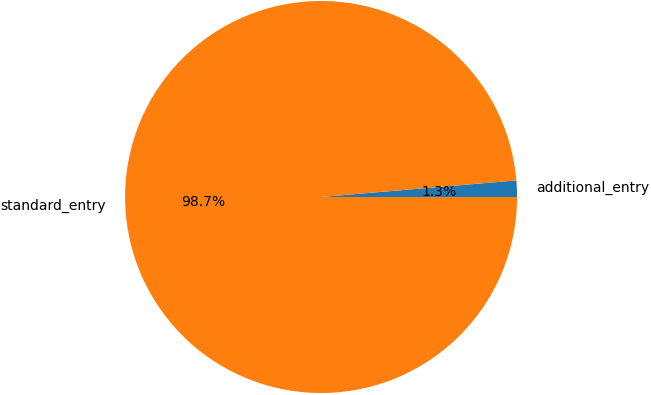


Рисунок 11 – Доля записей по типу записи

Таким же образом была построена гистограмма, показывающая количество проданной собственности каждого из типов собственности.

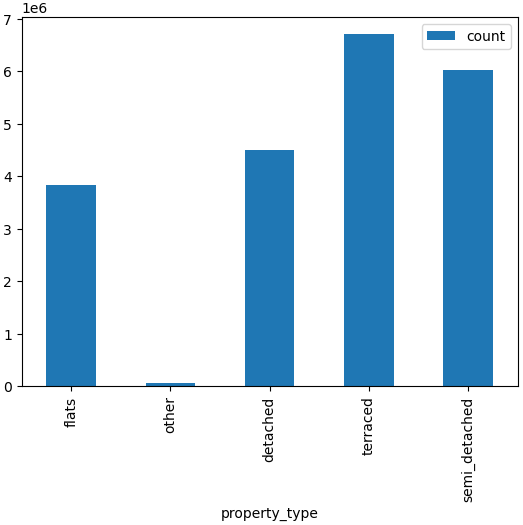


Рисунок 12 – Гистограмма количества проданной собственности по типам

Для построения гистограмм были определены размеры интервала (разделив максимальное значение цены на количество интервалов), после чего датасет был разбит на эти интервалы и вычислено количество записей в них.

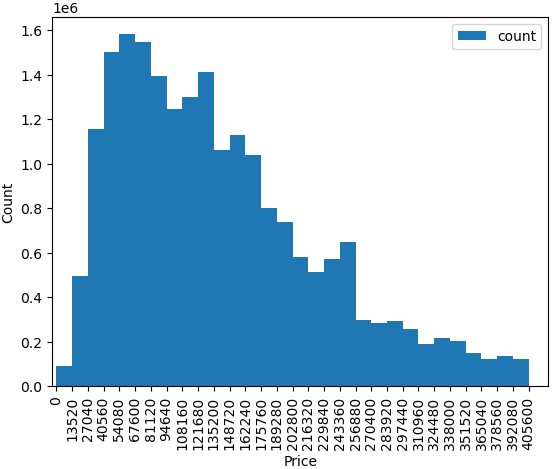


Рисунок 13 – Гистограмма распределения количества записей по цене

Для построения гистограммы, сгруппированной по типу собственности, датасет был так же разбит на интервалы, после чего был разбит на 5 датасетов по типу собственности. Каждый из этих датасетов был сгруппирован по интервалам и определено количество записей в каждом из них, после чего датасеты были объединены с использованием операции join.

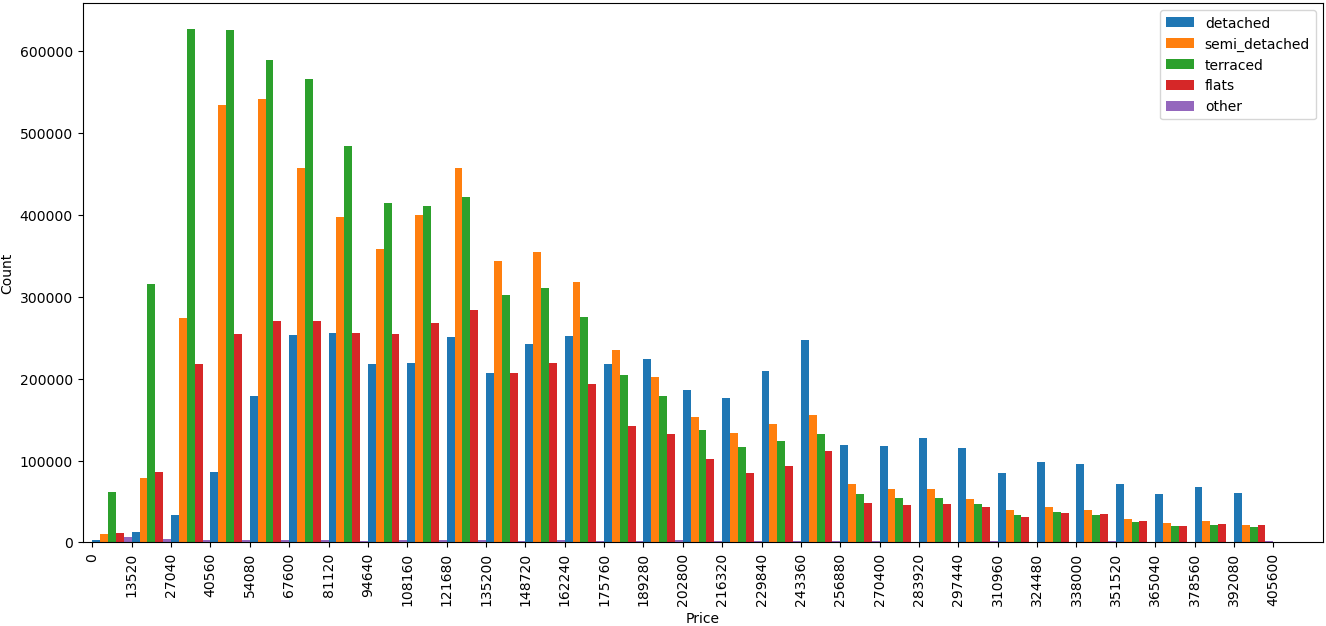


Рисунок 14 - Гистограмма распределения количества записей по цене (сгруппированная по типу собственности)

Для построения гистограммы распределения максимальной цены по графствам датасет был сгруппирован по графствам, после чего для каждой группы была определена максимальная цена. Эти группы были упорядочены по убыванию и оставлены первые 10 из них.

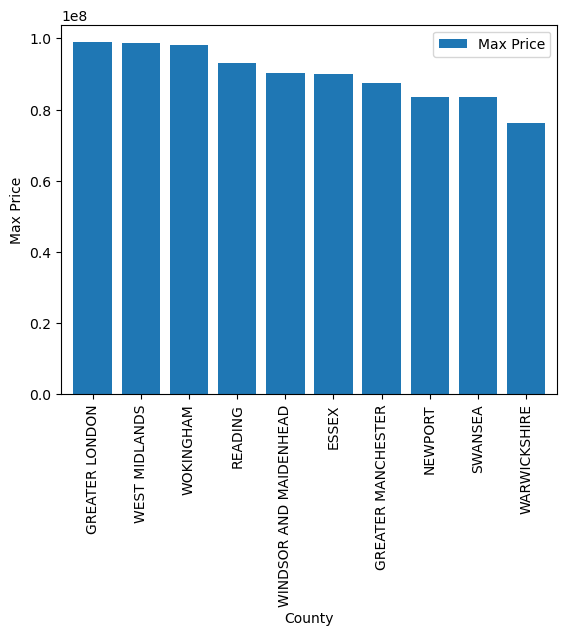


Рисунок 15 – 10 графств с наибольшей максимальной ценой

Для построения графиков количества проданной недвижимости и суммы произведенных транзакций по годам в датасет была добавлена колонка *year*, по которой датасет затем был сгруппирован, после чего были вычислены сумма цены и количества транзакций.

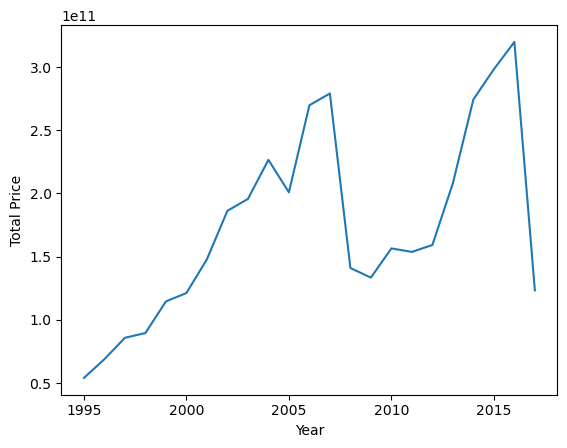


Рисунок 16– График зависимости суммы транзакций от года



Рисунок 17 – График зависимости числа транзакций от года

Выводы: был произведён разведочный анализ датасета, в котором содержатся данные о продаже недвижимости в Великобритании с 1995 по 2017 г., для этого был использован инструмент Apache Spark. Пропущенные значения в датасете отсутствуют, выбросы по цене были устранены за счёт ограничения ценового диапазона.

На диаграмме типов собственности мы наблюдаем преобладание типа *terraced*, что составляет примерно треть от общего числа покупаемых собственностей.

На диаграмме типов владения собственности преобладает *freehold* – полное владение собственностью и землей, что составляет ¾ от общего числа собственностей.

На диаграмме новизны собственности наибольшую часть составляет старая, перепродаваемая собственность.

При этом большая часть записей является основными и лишь 1,5% - дополнительными.

По гистограмме можно судить, что максимальная цена распределена по показательному закону. Также можно заметить, что в минимальном ценовом диапазоне преобладают здания типа *terraced*, в то время как в среднем преобладает тип *semi\_detached*, а в высоком – *detached*.

По последним графикам видно влияние кризиса 2008 года на продажу собственности, а так же то, что к 2016 году рынок так и не восстановился до конца.