МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
Национальный исследовательский ядерный университет «МИФИ»

Институт
интеллектуальных кибернетических систем

Кафедра кибернетики (№ 22)

Направление подготовки 09.04.04 Программная инженерия

БДЗ

по дисциплине «Анализ данных и машинное обучение Часть 2»

на тему:

«Анализ истории продаж объектов недвижимости»

Группа M23-524

Студент _____ Леонов В.В.
(подпись) (ФИО)

Преподаватель _____ Киреев В.С.
(подпись) (ФИО)

Москва 2024

Содержание

Введение

С каждым годом объем данных в различных сферах жизни продолжает расти, что приводит к значительному увеличению интереса к методам их анализа и интерпретации. Одной из ключевых задач в области анализа данных является изучение факторов, влияющих на рыночные процессы, в частности, на рынок недвижимости. Данные о сделках с недвижимостью содержат важную информацию, которая может помочь в оценке стоимости объектов, прогнозировании трендов на рынке и принятии стратегических решений для инвесторов и аналитиков.

В рамках настоящего исследования был использован набор данных, содержащий информацию о сделках с недвижимостью в США за период с 2001 по 2022 год. Данный набор данных включает информацию о ценах, различной характеристиках объектов недвижимости, что позволяет провести комплексный анализ рыночных тенденций. Однако, для эффективного извлечения полезных инсайтов из этого набора данных необходимо провести предварительную обработку и анализ, включающий как факторный анализ признаков, так и кластеризацию объектов.

Цель данного исследования — провести углубленный анализ истории продаж объектов недвижимости с использованием методов факторного и кластерного анализа для выявления закономерностей, связанных с ценовыми тенденциями на рынке недвижимости. В процессе работы необходимо решить следующие задачи:

1. Предобработка данных: очистка набора данных от пропущенных значений (NaN), а также добавление нескольких новых признаков для улучшения качества анализа.
2. Факторный анализ признаков: использование методов факторного анализа для уменьшения размерности и выявления скрытых факторов, влияющих на цену недвижимости.
3. Кластеризация данных: применение методов кластерного анализа для выделения групп объектов недвижимости с схожими характеристиками, что позволяет лучше понять структуру рынка.
4. Визуализация результатов: создание визуальных представлений для лучшего понимания полученных результатов и выявления ключевых трендов.

1 Набор данных

Предобработка данных является важнейшим этапом в любом исследовательском проекте, так как от качества обработки зависит дальнейший успех анализа. В рамках данного исследования были выполнены несколько ключевых шагов по очистке и подготовке данных из открытого набора о сделках с недвижимостью США за период с 2001 по 2022 год. Начальный набор данных содержал 1,097,629 строк и 14 столбцов, однако, после предварительной очистки и добавления новых признаков его размер был сокращен до 59,042 строк и 19 столбцов.

1.1 Удаление избыточных и нерелевантных данных

Первоначально в наборе данных присутствовало несколько столбцов, которые не представляли ценности для дальнейшего анализа и могли бы привести к излишним вычислительным затратам. В частности:

- Удаление столбца «Residential Type»: этот столбец оказался почти полным дубликатом столбца «Property Type», в котором уже содержалась необходимая информация о типе недвижимости. Удаление данного столбца позволило устранить избыточность в данных.
- Удаление строк с пропущенными значениями: были удалены все строки, в которых отсутствовали важные для анализа значения в столбцах «Property Type», «Non Use Code» и «Location». Эти столбцы критичны для определения типа недвижимости, использования и местоположения объектов, поэтому их отсутствие в строках делает такие записи бесполезными для анализа.
- Удаление столбцов «Assessor Remarks» и «OPM remarks»**:** данные столбцы содержали текстовую информацию, которая либо была слишком разрозненной, либо не несла существенной ценности для дальнейшего анализа, что делало их избыточными.

После выполнения этих шагов набор данных был значительно очищен, что позволило сосредоточиться на действительно полезных признаках.

1.2 Преобразование категориальных признаков

Множество столбцов в исходном наборе данных являлись категориальными и требовали числового представления для использования в дальнейшем анализе, особенно в алгоритмах машинного обучения. Заменены категориальные признаки «Property Type» и «Non Use Code» на числовые значения: для этого был использован метод замены категориальных значений на

среднее значение «Sale Amount» для каждой категории. Это позволило сохранить полезную информацию, связывающую тип недвижимости и её цену, что может быть полезным для анализа закономерностей.

1.3 Добавление новых признаков

Для повышения информативности набора данных были добавлены новые признаки, извлеченные из уже существующих данных:

1. Извлечение дня недели и месяца из столбца с датой сделки: визуальные и временные тренды, связанные с днями недели и месяцами, могут значительно влиять на рыночные ценовые колебания. Для каждого объекта недвижимости были добавлены признаки, отражающие день недели и месяц, в который была совершена сделка. Это позволяет анализировать сезонные и временные колебания цен на недвижимость.

2. Определение штата по координатам: на основе координат местоположения недвижимости был добавлен новый признак — «State», который указывает на штат, в котором находится объект недвижимости.

3. Замена категориального признака «State» на числовое значение: для того, чтобы включить этот признак в дальнейшие модели, категориальный столбец «State» был заменен числовыми значениями, аналогично тому, как это было сделано для других категориальных признаков.

1.4 Итоговый результат

После всех выполненных шагов предобработки данные стали значительно более чистыми и пригодными для дальнейшего анализа. Признаки были приведены к единому формату, устранены все строки с пропущенными значениями в ключевых столбцах, а также добавлены новые признаки для улучшения качества анализа. Итоговый размер набора данных после очистки и добавления новых признаков составил 59,042 строки и 19 столбцов, что делает его гораздо более управляемым для последующих этапов анализа, таких как факторный анализ и кластеризация.

2 Факторный анализ

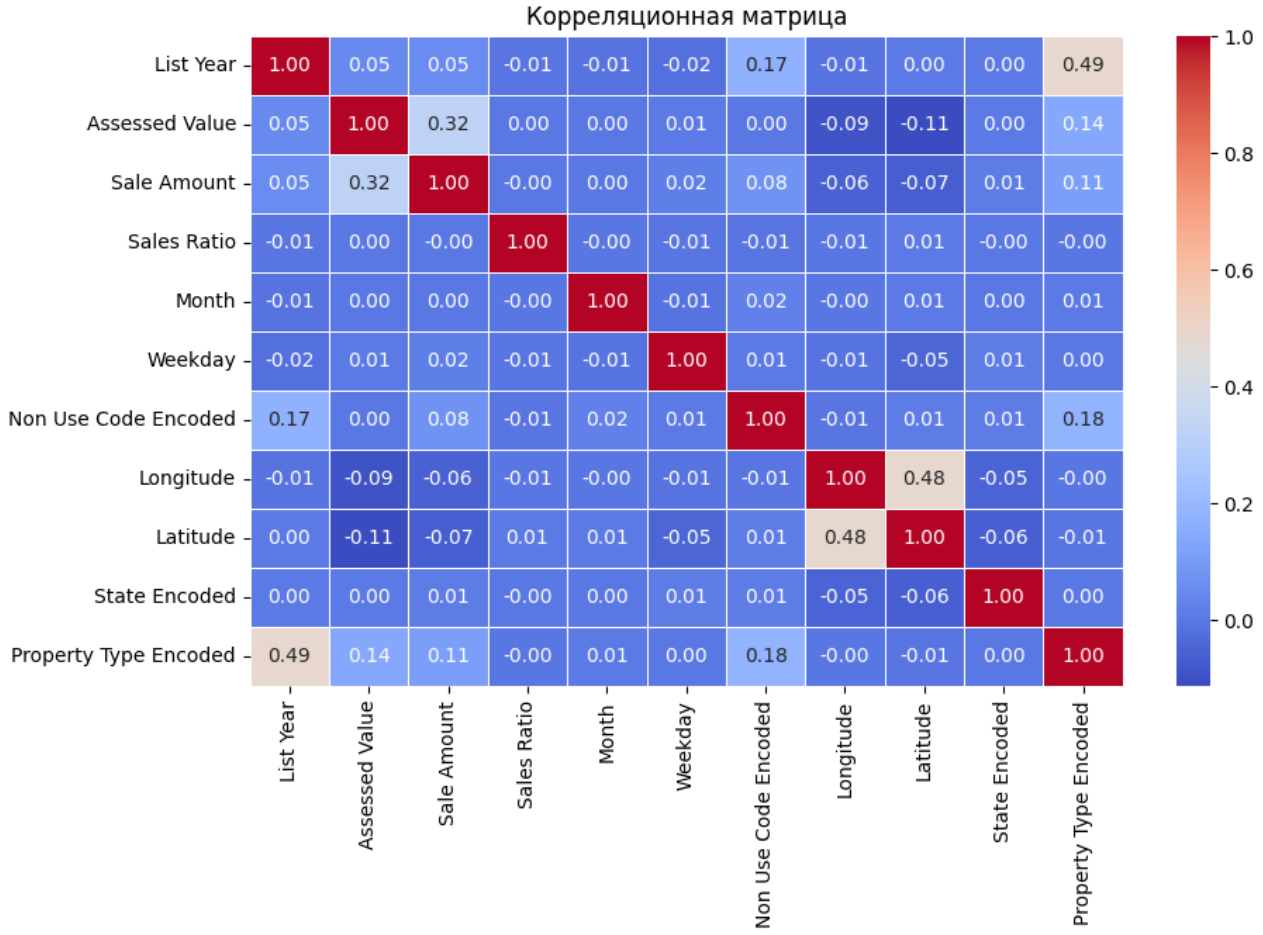
2.1 Преобразование данных

Для дальнейшего анализа данных, в том числе для построения различных корреляционных матриц и применения методов машинного обучения, важно привести численные признаки к единой шкале. Это необходимо, чтобы избежать доминирования признаков с большими величинами и улучшить работу алгоритмов, которые чувствительны к масштабу данных.

Для нормализации использовалась функция **StandardScaler** из библиотеки **sklearn.preprocessing**, которая преобразует данные так, чтобы они имели нулевое среднее значение и стандартное отклонение, равное 1. Это гарантирует, что все числовые признаки будут находиться в одинаковых масштабах и не будут иметь чрезмерного влияния на анализ.

2.2 Корреляционная матрица Пирсона

Для исследования линейных зависимостей между различными числовыми признаками была построена корреляционная матрица Пирсона.

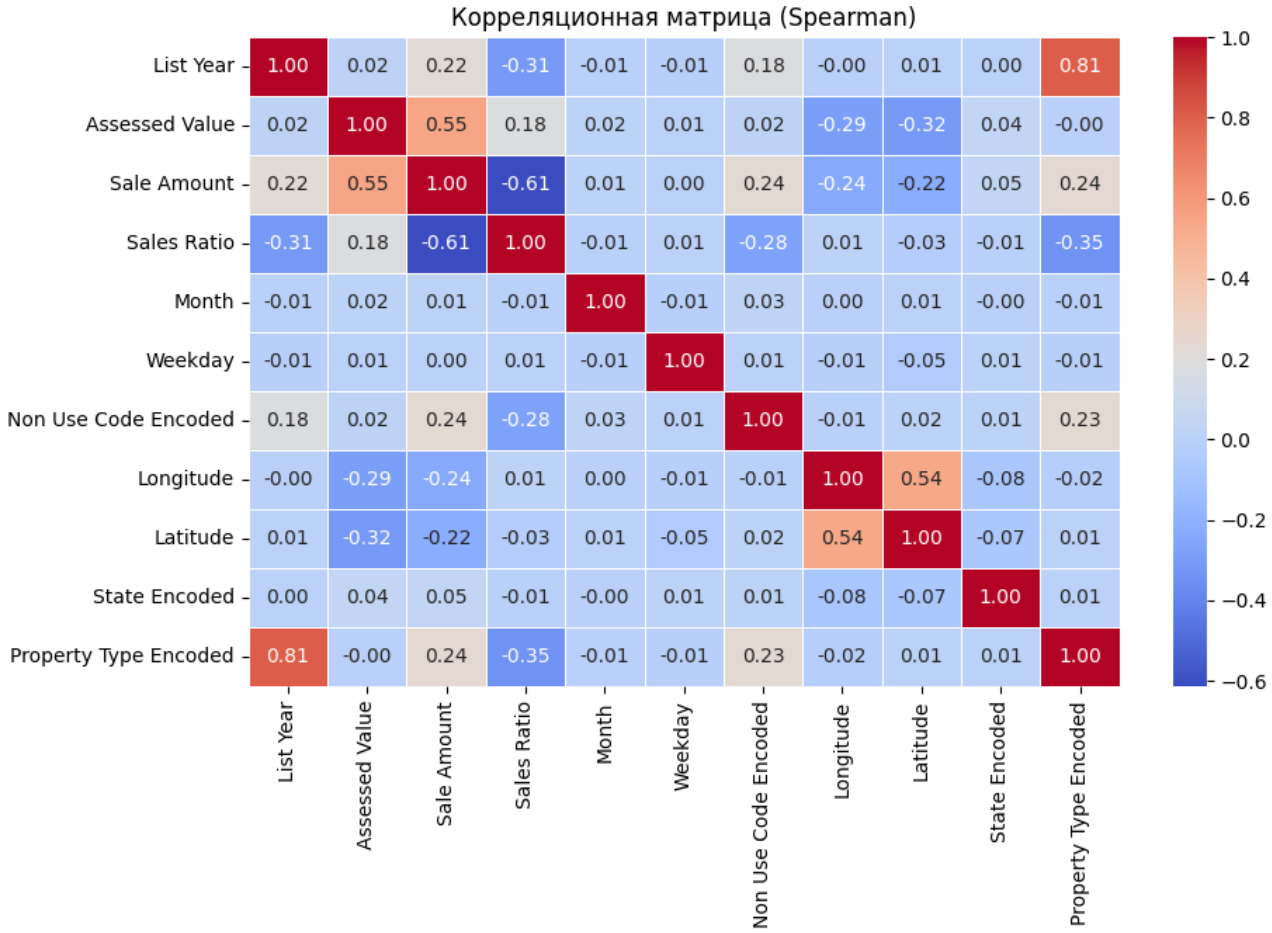


На основе этой матрицы можно сделать следующие выводы:

1. **«List Year» и «Property Type Encoded» (корреляция = 0.49):**
    a. Это самая высокая положительная корреляция в матрице.
    b. Это может означать, что коды типов недвижимости («Property Type Encoded») имеют тенденцию изменяться с течением времени (в зависимости от года). Например, типы недвижимости, популярные в более ранние годы, отличаются от тех, что преобладают в более поздние периоды.
2. **«Assessed Value» и «Sale Amount» (корреляция = 0.32):**
    a. Умеренная положительная корреляция указывает на то, что оценочная стоимость объекта («Assessed Value») связана с его продажной ценой («Sale Amount»). Это ожидаемая связь, так как оценочная стоимость часто служит индикатором реальной рыночной цены, но не является идеальным предсказателем.
3. **«Non Use Code Encoded» и «Property Type Encoded» (корреляция = 0.18):**
    a. Слабая положительная корреляция говорит о том, что тип недвижимости («Property Type Encoded») имеет небольшую связь с кодом неиспользования («Non Use Code Encoded»). Это может означать, что некоторые типы недвижимости более склонны к определённым категориям использования.
4. **«Assessed Value» и «Property Type Encoded» (корреляция = 0.14):**
    a. Слабая положительная корреляция предполагает, что тип недвижимости оказывает минимальное влияние на её оценочную стоимость.
5. **Большинство признаков имеют корреляции близкие к 0**, что указывает на отсутствие значимой линейной зависимости.

2.3 Корреляционная матрица Спирмана

Корреляционная матрица Спирмана оценивает степень монотонной зависимости между признаками. В отличие от Пирсона, метод Спирмана лучше работает с нелинейными зависимостями, измеряя, как хорошо значения одной переменной можно представить в виде монотонной функции от другой.



На основе этой матрицы можно сделать следующие выводы:

1. **«List Year» и «Property Type Encoded» (корреляция = 0.81):**
    a. Это сильная положительная зависимость. Она указывает на то, что закодированные типы недвижимости («Property Type Encoded») тесно связаны с годом («List Year»). Возможно, разные типы недвижимости становятся популярными или более актуальными в зависимости от временного периода.
2. **«Assessed Value» и «Sale Amount» (корреляция = 0.55):**
    a. Умеренная положительная зависимость между оценочной стоимостью и продажной ценой. Это ожидаемая связь, так как оценочная стоимость является индикатором рыночной цены объекта недвижимости.
3. **«Longitude» и «Latitude» (корреляция = 0.54):**
    a. Умеренная положительная зависимость между координатами недвижимости. Это связано с географическим расположением, где некоторые районы или направления имеют тенденцию быть связанными друг с другом.
4. **«Sales Ratio» и «Sale Amount» (корреляция = -0.61):**

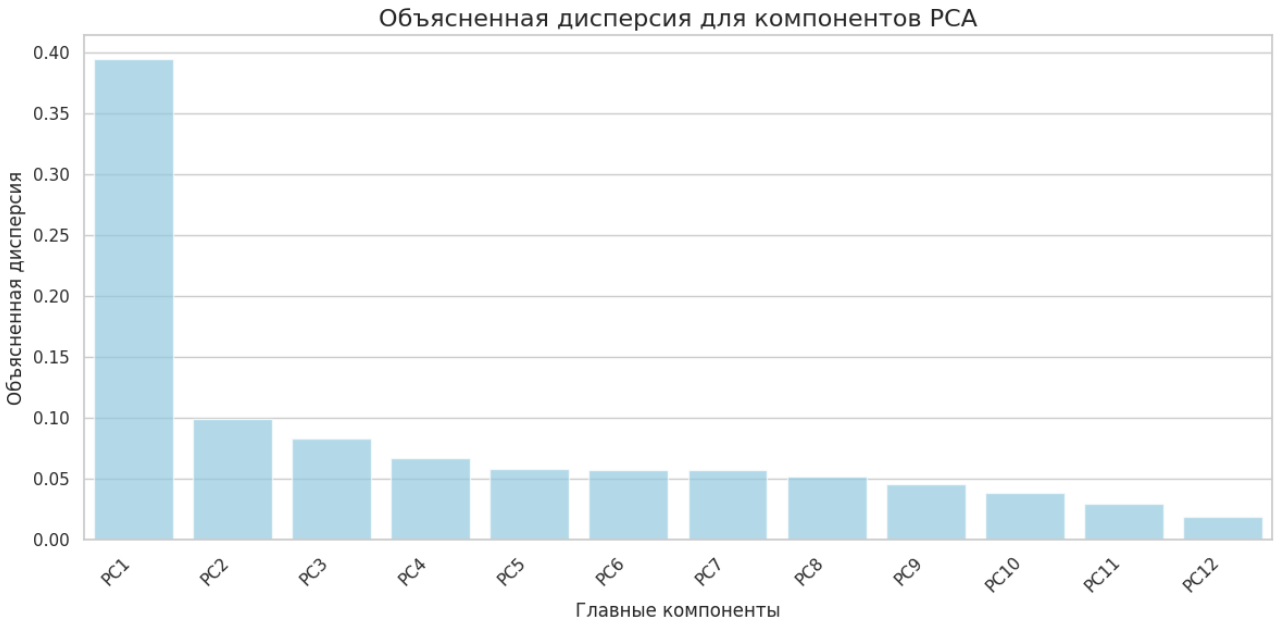
a. Сильная отрицательная зависимость указывает на то, что с увеличением суммы продажи («Sale Amount») коэффициент продаж («Sales Ratio») уменьшается. Это может быть связано с различиями в методике оценки для объектов с разной ценой.

5. **«Sales Ratio» и «Property Type Encoded» (корреляция = -0.35):**
    a. Умеренная отрицательная зависимость говорит о том, что разные типы недвижимости могут по-разному влиять на коэффициент продаж.
6. **«List Year» и «Sales Ratio» (корреляция = -0.31):**
    a. Умеренная отрицательная зависимость предполагает, что коэффициент продаж («Sales Ratio») имеет тенденцию снижаться с годами. Возможно, это связано с изменением рыночных условий или методик оценки.
7. **«Month» и другие признаки:**
    a. Месяц сделки имеет практически нулевую корреляцию с другими переменными, что указывает на отсутствие значимых монотонных связей. Это может говорить о том, что временные колебания (например, сезонность) не оказывают сильного влияния на основные характеристики данных.
8. **«Weekday» и другие признаки:**
    a. День недели сделки также практически не связан с другими характеристиками, что подтверждает гипотезу о его низкой значимости для анализа.

2.4 Анализ главных компонент

Для уменьшения размерности данных был проведён анализ главных компонент (PCA). Главной целью PCA стало выявление направлений с наибольшей вариативностью данных и сохранение основной информации при снижении числа признаков.

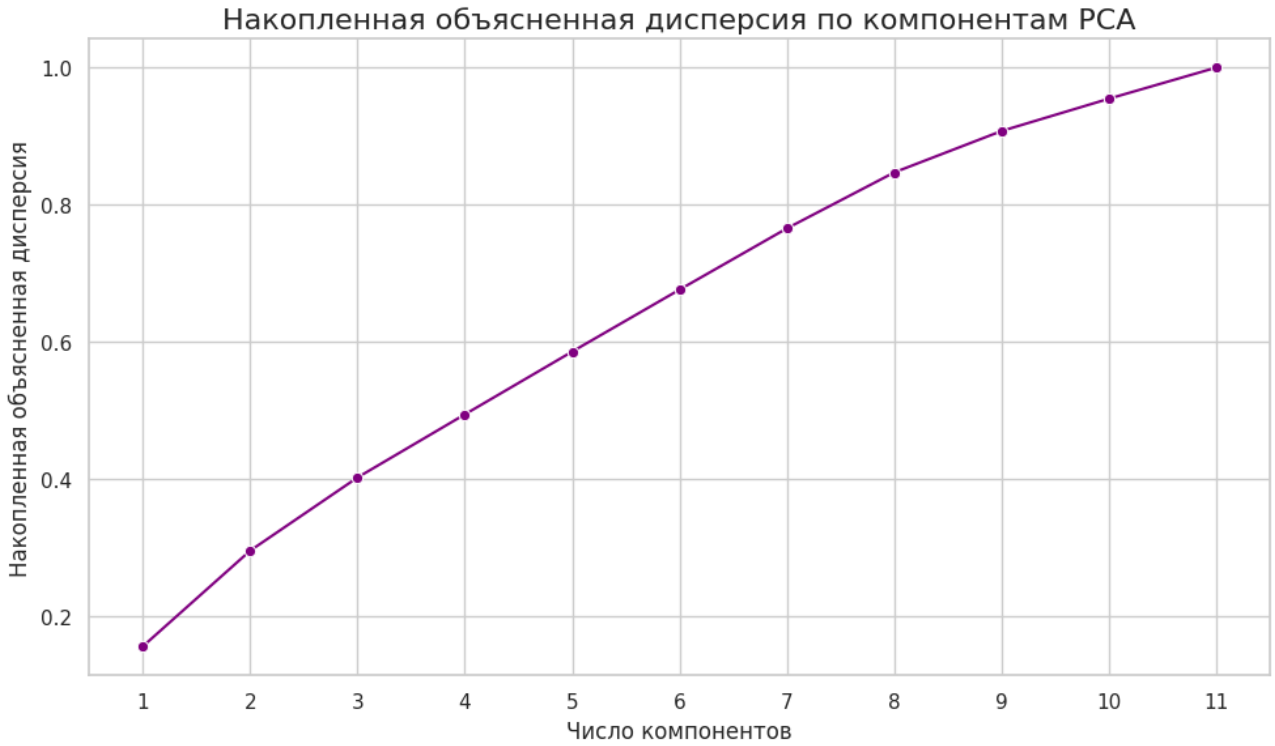
1. Первая главная компонента (PC1) объясняет около 40% общей дисперсии, что указывает на её высокую значимость в структуре данных.
2. Вторая компонента (PC2) добавляет ещё 10-12% дисперсии, в сумме с PC1 они покрывают 50-55% дисперсии.
3. Начиная с третьей компоненты, вклад резко снижается, и компоненты с PC5 и далее вносят минимальный вклад, что позволяет считать их слабо информативными.



На представленном графике показано распределение данных после применения анализа главных компонент (PCA) в пространстве первых двух компонент (PC1 и PC2). Вот что можно из него понять. На графике видно несколько кластеров точек, которые могут представлять группы данных со схожими характеристиками. Точки на графике отображают объекты данных, спроецированные в новое пространство, сформированное первой и второй главными компонентами.

На графике представлена **накопленная объяснённая дисперсия** в зависимости от числа компонентов PCA. Этот график помогает определить, какое количество компонент достаточно для сохранения основной информации в данных. Вот что можно из него понять:

1. **Первая компонента (PC1):** Объясняет примерно 20-25% дисперсии данных.
2. **Первые две компоненты (PC1 + PC2):** В сумме объясняют около 40-50% дисперсии.
3. **Первые четыре компоненты (PC1 + PC2 + PC3 + PC4):** Покрывают около 70% дисперсии данных.
4. **Первые шесть компонент:** Объясняют уже ~85% дисперсии.
5. **Полное покрытие (100%):** Требуется все 11 компонент, но накопленная дисперсия стремится к 1 уже на 8-9 компонентах.



3 Кластерный анализ

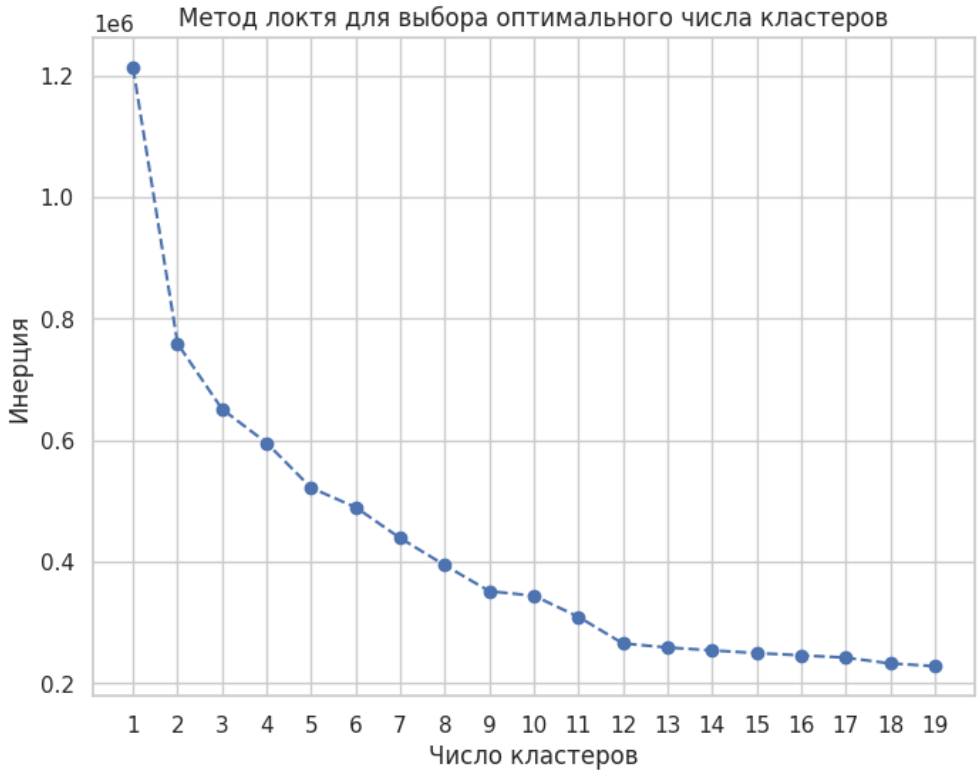
Для кластерного анализа было принято решение использовать **K-Means** — это алгоритм кластеризации, который группирует данные в k кластеров, минимизируя расстояние между объектами и центроидами кластеров.

3.1 Выбор оптимального количества кластеров

Метод локтя использовался для определения оптимального числа кластеров в рамках кластерного анализа. На графике отображается зависимость значения инерции (внутрикластерной суммы квадратов) от количества кластеров. Инерция уменьшается по мере увеличения числа кластеров, так как объекты распределяются по меньшим и более плотным кластерам. Однако после определенного момента снижение инерции замедляется, и дальнейшее увеличение числа кластеров приводит к незначительному улучшению.

На данном графике точка локтя наблюдается при 10 кластерах. Это означает, что выбранное количество кластеров обеспечивает баланс между качеством кластеризации и сложностью модели. Использование большего числа кластеров не приводит к значительному снижению инерции и может быть неоправданным с точки зрения интерпретируемости и вычислительных затрат.

Таким образом, оптимальное число кластеров для данного набора данных составляет 10. Это значение будет использоваться для последующего анализа и визуализации кластеров.



3.2 Результаты кластеризации

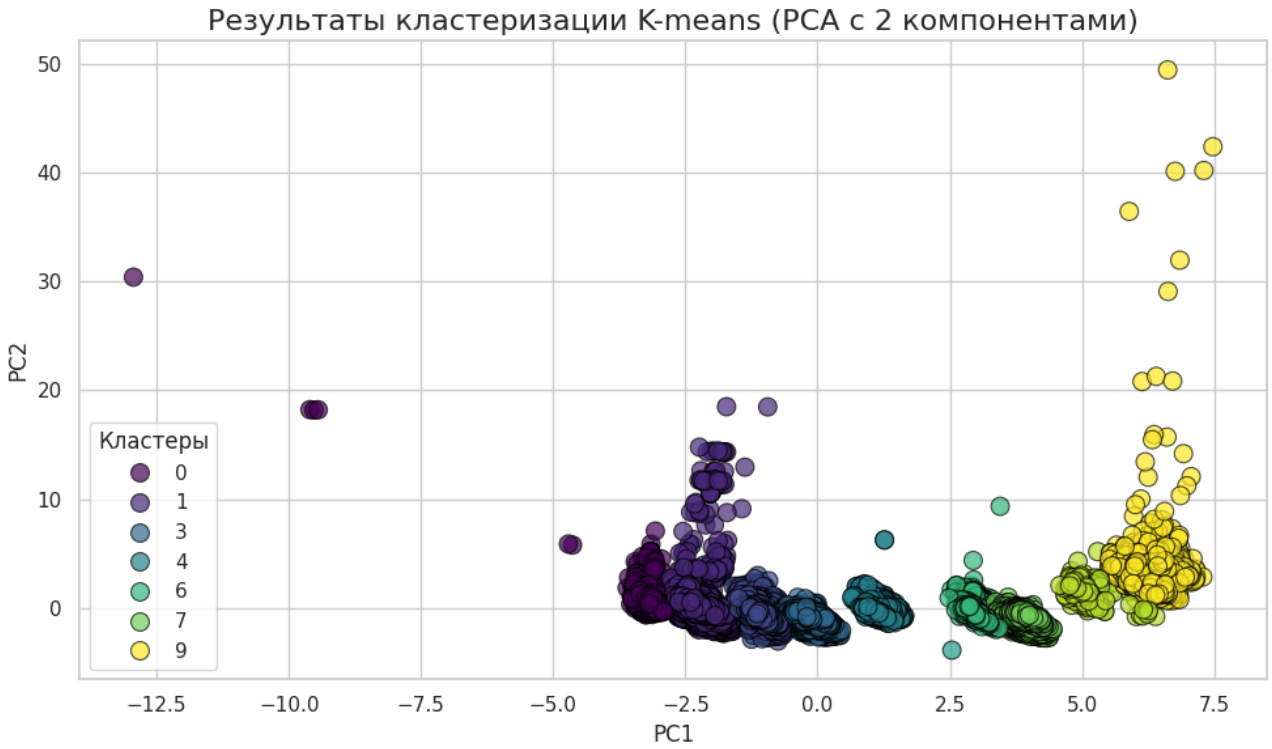
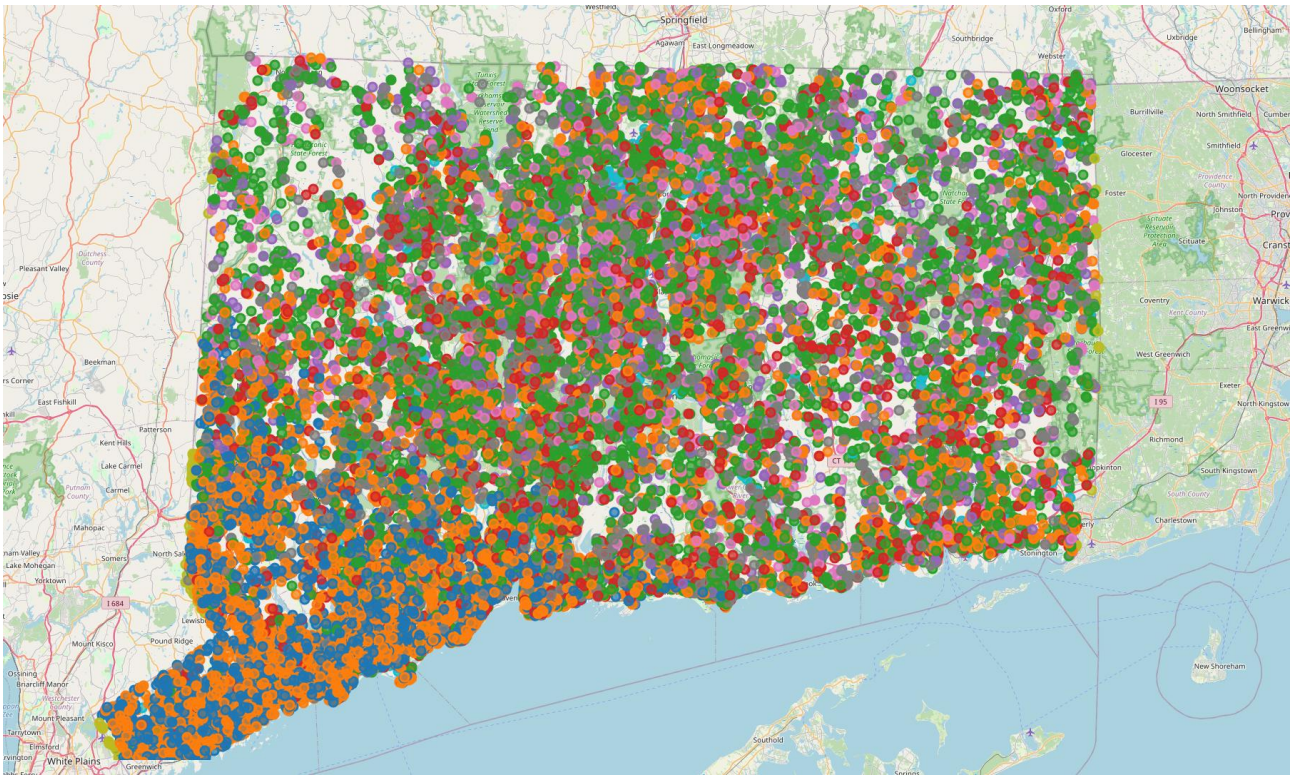


График выше иллюстрирует результаты кластеризации методом K-means с использованием снижения размерности с помощью метода главных компонент (PCA) до двух компонент (PC1 и PC2). На графике каждая точка представляет собой объект из исходного набора данных, а цвет точки соответствует кластеру, к которому данный объект был отнесен алгоритмом K-means.

1. Полученные результаты согласуются с гипотезой описанной в п. 2.4.
2. Кластеры имеют четкие границы, хотя некоторые из них слегка перекрываются, что свидетельствует о возможной корреляции между компонентами.
3. Кластеры имеют разное распределение по двум компонентам, что подтверждает их различие.
4. Большинство кластеров сгруппированы компактно, что говорит о том, что объекты внутри кластеров имеют сходство.
5. Некоторые кластеры (например, кластер 9) занимают больше пространства, что может указывать на внутреннюю неоднородность.

3.3 Визуализация кластеризации



Визуализация кластеров объектов недвижимости на карте также выполнена.

1. Распределение точек показывает высокую плотность объектов в южной части региона (вблизи побережья), в то время как северные области характеризуются меньшей концентрацией недвижимости.

2. Южный регион (вдоль побережья) доминирует кластерами, представленными оранжевым, синим и зеленым цветами. Это может свидетельствовать о различии в типах недвижимости (например, дорогие дома на побережье, районы с высокой плотностью застройки).

3. Центральная часть карты характеризуется более равномерным распределением точек различных кластеров. Это может отражать разнообразие типов объектов недвижимости в этом районе.

4. Северные регионы карты включают преимущественно объекты, относящиеся к фиолетовому и зеленому кластерам. Это может быть связано с меньшей плотностью населения и типами недвижимости (например, сельские дома или земельные участки).

5. Высокая плотность точек в южной части может указывать на более развитую инфраструктуру, близость к морю, высокую стоимость недвижимости.

6. Разнообразие кластеров может отражать смешанную застройку: от жилых многоквартирных домов до коммерческих объектов.
7. Меньшая плотность недвижимости и преобладание нескольких кластеров могут быть связаны с доминированием сельских или пригородных территорий.

Выводы

В рамках анализа истории продаж объектов недвижимости были выполнены следующие этапы:

1. Предобработка данных:
    a. Данные очищены от пропущенных значений и нерелевантных признаков.
    b. Проведена трансформация категориальных данных и добавлены новые признаки для улучшения качества анализа.
2. Факторный анализ:
    a. Построены корреляционные матрицы (Пирсона и Спирмана) для выявления взаимосвязей между признаками.
    b. Применен метод главных компонент (PCA), который позволил сократить размерность данных и выделить ключевые скрытые факторы, влияющие на цену недвижимости.
3. Кластерный анализ:
    a. Метод локтя использовался для выбора оптимального числа кластеров.
    b. Объекты недвижимости были разделены на группы с помощью K-means. Это позволило выявить сегменты рынка с различными характеристиками, такими как стоимость, расположение и типы объектов.
4. Визуализация результатов:
    a. Построены графики и карты, отражающие структуру кластеров.
    b. Визуализация помогла наглядно интерпретировать результаты анализа и выявить географические и ценовые закономерности.

Анализ показал, что факторные признаки и кластеризация позволяют лучше понимать структуру рынка недвижимости, сегментировать данные и выявлять ключевые тренды. Результаты могут быть использованы для прогнозирования цен и принятия стратегических решений в области недвижимости.